ECML PKDD 2013

PRAGUE
23–27 SEPTEMBER
2013

EUROPEAN CONFERENCE ON MACHINE LEARNING AND PRINCIPLES AND PRACTICE OF KNOWLEDGE DISCOVERY IN DATABASES

PROGRAMME

PARTNERS AND SUPPORTERS

We would like to thank the following sponsors and supporters

Gold sponsor

WINTON

Silver sponsors

CISCO

Deloitte.

KNIME

YAHOO! LABS

Bronze sponsors

ČSKI

definity SYSTEMS

DIKW academy

Google

xerox

zalando

Supporter

Office of Naval Research GLOBAL Science & Technology

Prize sponsors

Springer

Google

YAHOO! LABS

Deloitte.

CONTENT

WELCOME ADDRESS

Dear colleagues,

It is our great pleasure to welcome you to ECMLPKDD 2013 in Prague, Czech Republic.

We have worked hard - together with a local team and a large number of dedicated chairs who are listed in the conference proceedings - to bring you a high quality and varied scientific programme. This booklet has been put together to help you find your way through the conference and make the most of your stay in Prague.

Enjoy ECMLPKDD 2013 and Prague !

Hendrik Blockeel
Kristian Kersting
Siegfried Nijssen
Filip Železný
programme chairs

Jiří Kléma
local chair

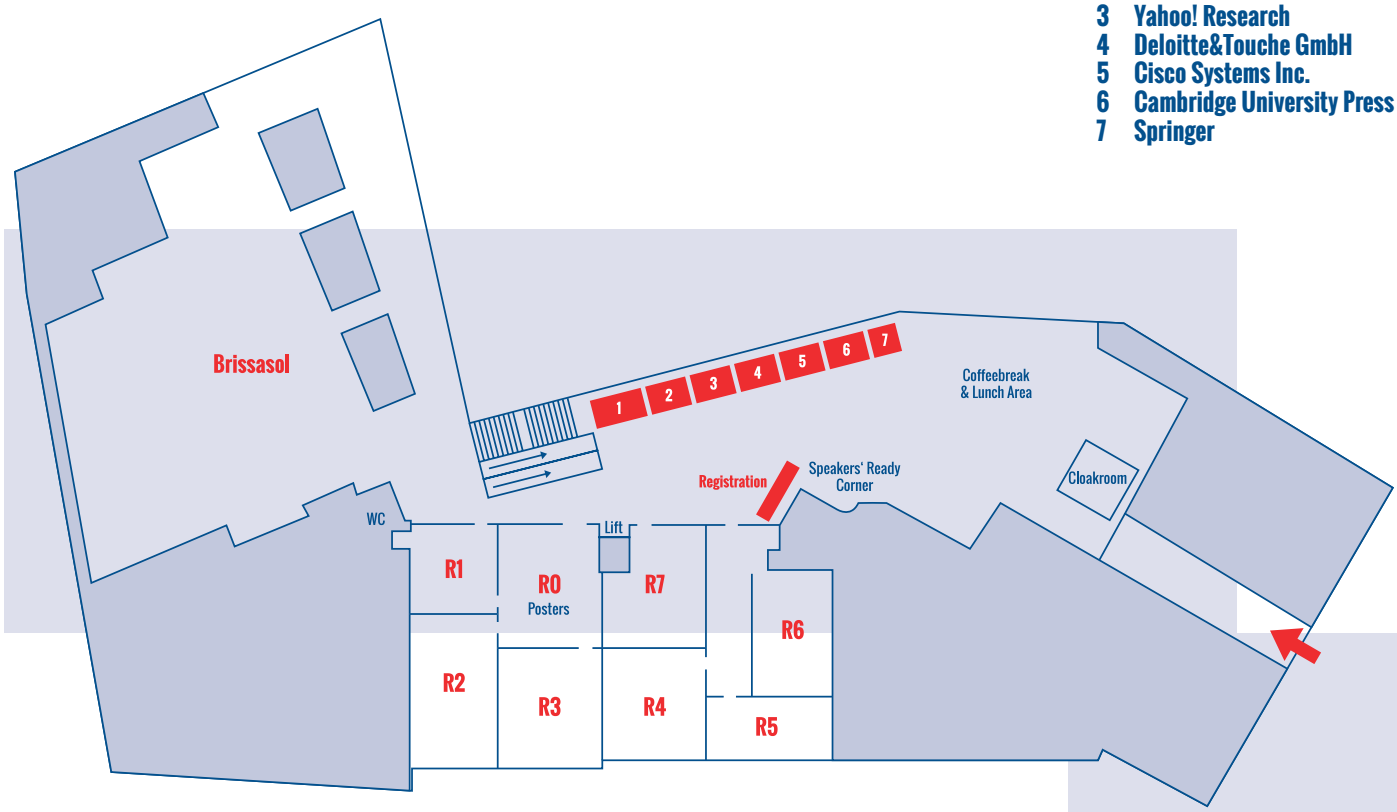
CONFERENCE SECRETARIAT

GUARANT International spol. s r.o.
Mrs. Marcela Rajtorová
Na Pankráci 17
140 21, Praha 4
Czech Republic
Tel.: +420 284 001 444
Fax: +420 284 001 448
E-mail: rajtorova@guarant.cz

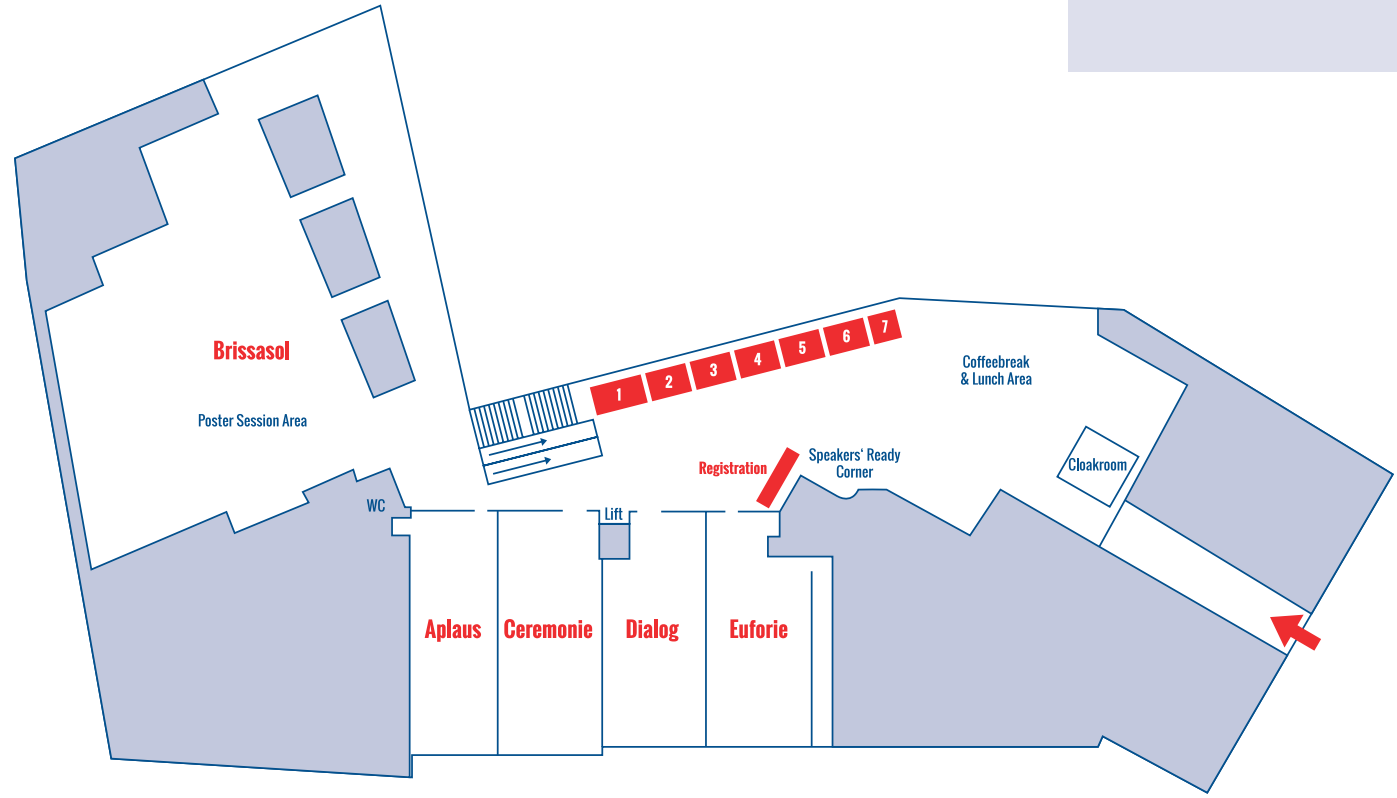
FLOOR PLANS

Floor Plan for September 23 - 27, 2013

Exhibition (legend)
1 Winton Capital Management
2 KNIME.com AG
3 Yahoo! Research
4 Deloitte&Touche GmbH
5 Cisco Systems Inc.
6 Cambridge University Press
7 Springer

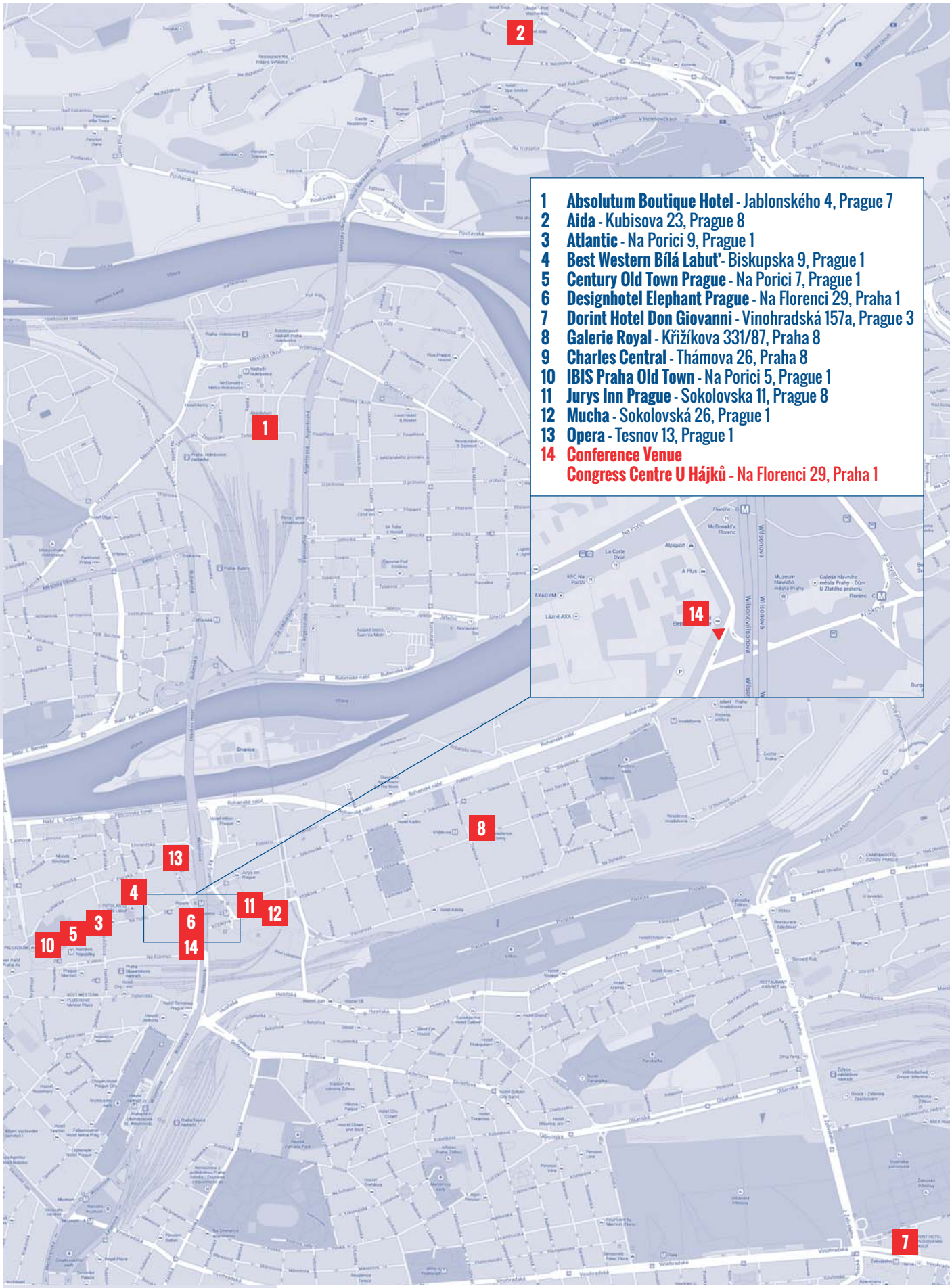


Floor Plan for September 24-26, 2013



USEFUL MAPS

Hotels available for ECML PKDD 2013



REGISTRATION

The Registration Desk is located in the foyer of the Congress Centre U Hájků

Opening hours

Monday, Sept. 23, 2013	08:00-20:00
Tuesday, Sept 24, 2013	08:00-18:00
Wednesday, Sept. 25, 2013	08:00-18:00
Thursday, Sept. 26, 2013	08:30-18:00
Friday, Sept. 27, 2013	08:00-19:00

Registration Fees

Regular	EUR 470,-
Student	EUR 440,-
Accompanying Person	EUR 200,-

Registration Fee Includes:
- Admission to all scientific sessions, workshops and tutorials on Monday and Friday
- Admission to the technical programme on Tuesday, Wednesday and Thursday
- Conference materials
- Welcome Cocktail: September 23, 2013
- Conference Dinner: September 25, 2013
- Free transport ticket valid for five days

Accompanyning person´s Fee Includes:
- Welcome Cocktail: September 23, 2013
- Sightseeing Tour: September 24, 2013
- Conference Dinner: September 25, 2013
- Free transport ticket valid for five days

BADGES

Name badges have been colour-coded as follows:

Organizers - red	
Delegates - blue	
Accompanying persons - yellow	
Staff - orange	

PRESENTATION INSTRUCTIONS

Who Presents How
- **Main technical track** (both proceedings and journal papers): oral presentation and a poster
- **Nectar and industrial tracks, plenary talks, tutorials**: only oral presentation
- **Demo track**: oral spotlight presentation and live demo
- **Workshops**: oral presentation OR poster + oral spotlight presentation (The authors have been informed about their mode of presentation.)

Oral Presentations
- Presentation time:
  - **Journal** papers in the main technical track: **25 min**. incl. 5 min. for discussion
  - **Proceedings** papers in the main technical track: 20 min. incl. 5 min. for discussion
  - **Demo** spotlights: **9 min**
  - Workshop papers: per workshop instructions
- A **computer** and a data **projector** will be available. Please bring your presentation on a **USB memory stick** or on a **CD**. Accepted file formats are MS-Power Point **PPT** and Adobe **PDF.**
- Please **upload** your presentation in the lecture room where your talk will take place **10 minutes** before the start of your session. **Exception**: for **Friday 10:45 workshops**, please submit your presentation at the Speakers Ready Corner (see below) during the coffee break preceding these workshops.
- When your session is over, your presentation will be deleted from all computers, no copies or backups will be made.
- Speakers can use their **own computers** for the presentation. This option is not recommended if the presentation is a standard PPT or PDF file. A VGA cable and a 220V power outlet will be available. The setup should be tested in the session room in one of the breaks before the session as early as possible.
- The **Speakers Ready Corner** will be available for speakers requiring assistance with format conversions, file upload etc. The operation hours are from 8:00 AM to 8:00 PM (Mon) / 6:00 PM (Tue, Wed, Thu) / 7:00 PM (Fri).
- Special **PowerPoint** considerations:
  - Please use one of these versions: PP 97-2003 and 95 or 2007, 2010 and save your presentation as a PPT rather than PPS. All videos or animations in the presentation must run automatically.
  - **Fonts:** Only fonts that are included in the basic installation of MS-Windows will be available (English version of Windows). Other fonts can cause a wrong layout of your presentation. The suggested fonts are Arial, Times New Roman, Tahoma. If you insist on using different fonts, these must be embedded into your presentation by choosing the right option when saving your presentation: Click on “File”, then “Save As”, check the “Tools” menu and select “Embed True Type Fonts.”

- **Pictures and Videos**
  - JPG is the preferred file format for images. GIF, TIF or BMP formats will be accepted as well.
  - Images inserted into PowerPoint are embedded into the presentations. Images that are created at a dpi setting higher than 200 dpi are not necessary and will only increase the file size of your presentation.
  - We cannot provide support for embedded videos in your presentation; please test your presentation with the on-site PC several hours before your presentation. Generally, the WMV format should work with no difficulties.
  - In case your video is not embedded in the presentation it is possible to have it in other formats: MPEG 2,4, AVI (code: DivX, XviD, h264) or WMV. Suggested bitrate for all mpeg4 based code is about 1Mbps with SD PAL resolution (1024x576pix with square pixels, AR: 16/9)
  - In case of Full HD videos, please let us know before the meeting so that they can be tested.
  - Videos that require additional reading or projection equipment (e.g., VHS cassettes) will be not accepted.

POSTERS

Poster Area

- Posters of the **main technical track** will be displayed in the special **Poster Room** accessible from the foyer of the congress centre. From Tuesday to Thursday, it will be accessible all day long and besides poster visits, it can be used for meetings and discussions.
- **Workshop** posters will be displayed in Room R0, which is one of the sections of the main conference hall on Monday and Friday.

Mounting and Removal Times

- Monday's workshop papers: mount after Monday 8:30 AM, remove before Monday 5:30 PM.
- Tuesday's main track papers: mount after Tuesday 8:30 AM, remove after Tuesday's poster session
- Wednesday's and Thursday's main track papers: mount after Wednesday 8:30 AM, remove after Thursday's poster session
- Friday's workshops papers: mount after Friday 8:30 AM, remove before Friday 7:00 PM.

Format and Mounting

- All posters of the **main technical track** will be mounted on poster **stands** in the poster room. Workshop posters will be mounted on the **walls** of room R0 or on poster **stands** in that room.
- The **maximum size** of the poster is 180 cm (height, 71 inches) x 97 cm (width, 38 inches). The **recommended size** is 100 cm (39 inches) x 95 cm (37 inches).
- Fixing material (pins and stickers) will be available. For wall mounting, only stickers can be used.
- Congress staff will be available to assist you during the time of poster mounting. The poster positions will be numbered. The number of your poster can be found in the final programme.

LIVE DEMOS

- Live demos will be shown during the Tuesday's poster session using the authors' own computers. Tables and 220V power outlets will be available. Please contact the local chair as soon as possible if you need additional assistance.
- A demo may be accompanied by a poster on a poster stand next to the demo table. Demo presenters interested in displaying a poster should write a request to the demo chair as soon as possible. The poster conditions described above apply here as well, with mounting and removal times same as for Tuesday's main track papers.

GENERAL INFORMATION

The Conference4me smartphone application:
The Conference4me smartphone app is a comfortable tool for planning your participation at ECML PKDD 2013. Browse the complete programme directly from your phone or tablet and create your very own agenda on the fly. To download Conference4me mobile app, please visit http://conference4me.eu/download or type Conference4me in Google Play or iTunes App Store. You can also use the QR codes:

Within the application you will be able to download the ECMLPKDD´13 data file.

Currency and exchange rate:
The official currency of the Czech Republic is the Czech Crown = ČESKÁ KORUNA (CZK = Kč). The exchange rate is approx. 26,- CZK= 1 EUR (September 2013). International credit cards are accepted for payments in most hotels, restaurants and shops. Exchange offices and ATM machines are easily available throughout the city and at the Václav Havel Airport.

Wi-Fi
Wi-Fi will be available at the Congress Centre U Hájků during the whole conference.

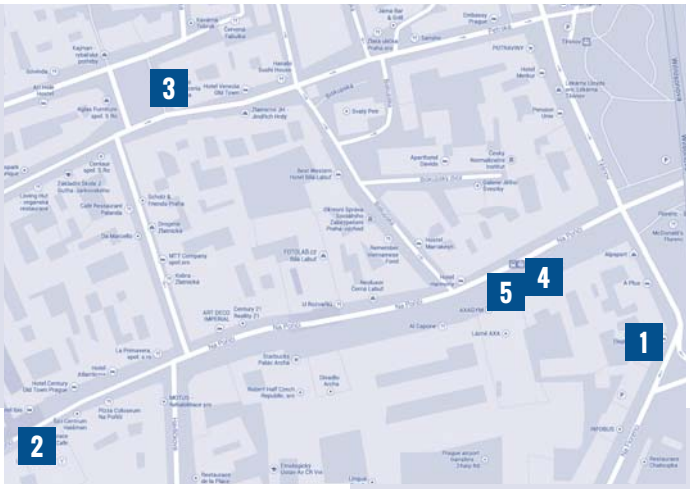
Cloakroom
A cloakroom is located in the foyer - see the floor plan. Opening hours correspond with the opening hours of the Registration Desk.

Insurance
The Organisers of the Conference do not accept liability for any injury, loss or damage, arising from accidents or other situations during, or as a consequence of the Conference. Participants are therefore advised to arrange insurance for health and accident prior to travelling to the Conference.

Language
The official language of the Conference is English. Simultaneous interpretation is not provided.

Lunches
ECML PKDD 2013 does not organize lunches for the participants. However there are several places suitable for lunches:

1 **Congress Centre U Hájků** - foyer (limited capacity)
2 **Restaurant La Republica** - Na poříčí 12, Prague 1, www.larepublica.cz
3 **Restaurant Alforno (pizzeria)** - Petrské náměstí 4, Prague 1, www.al-forno.cz
4 **Bageteria Boulevard** - Na poříčí 42, Prague 1, www.bb.cz
5 **KFC** - Na poříčí 46, Prague 1, www.kfc.cz



Message System
The Message Board is located in the Registration Area in the foyer of the Congress centre U Hájků. There you may leave a message for your friends or colleagues.

Mobile Phones
Delegates are kindly requested to switch off their mobile phones during the sessions.

Programme Changes
The organisers cannot assume liability for any changes in the programme due to external or unforeseen circumstances.

Refreshment
Complimentary coffee refreshments will be served to all registered participants in the foyer of the Congress centre U Hájků.

Smoking Policy
For the comfort and health of all participants, smoking is not permitted at any time.

Staff
Conference staff will be happy to assist to participants during the Conference.

Ticket for Public Transportation
A free 5-day ticket will be distributed during registration to all registered participants, accompanying persons and exhibitors.

GOOD TO KNOW

Electricity
The Czech Republic uses a 220 volt 50 Hz system.

Shopping
Most shops in Prague are open from 9:00 to 18:00, Monday through Saturday. Shops in the city centre are usually open from 10:00 to 20:00, Monday through Sunday.

Taxi Service
In the city centre, taxis are easy to hail from the street but we strongly recommend to use hotel taxis or to call taxi by phone, through the radio taxi service of the following reliable companies:

AAA Radiotaxi: 14014
ProfiTaxi: 844 700 800

Tipping
Service is usually included in the bill in bars and restaurants but tips are well accepted. If you consider the service good enough to warrant a tip, we suggest about 5-10%.

SOCIAL EVENTS

Welcome Drink
Sept. 23, 2013 at the Congress Centre U Hájků - 20:30

Conference Dinner
Sept. 25, 2013 at the Strahov Monastery - Klášterní Restaurant - 20:00
Meeting point at the Registration Desk at 19:20. Those who will go to the restaurant on thein own will receive at the Registration Desk the map „How to get to the Klašterní Restaurant„

Poster Sessions
Wine, beer and snacks will be served during both poster sessions

SIGHTSEEING TOURS

Informative Tour of the Old Town - walking tour
Sept. 24, 2013 - 14:00-17:00
Meeting point at the Registration Desk at 13:55

Due to the lack of participants other tours have been cancelled.

KEYNOTE SPEAKERS

Plenary Invited Speakers

Rayid Ghani
Using Machine Learning Powers for Good

Abstract:
The past few years have seen increasing demand for machine learning and data mining - both for tools as well as experts. This has been mostly motivated by a variety of factors including better and cheaper data collection, realization that using data is a good thing, and the ability for a lot of organizations to take action based on data analysis. Despite this flood of demand, most applications we hear about in machine learning involve search, advertising, and financial areas. This talk will talk about examples on how the same approaches can be used to help governments and non-prpofits make social impact. I'll talk about a summer fellowship program we ran at University of Chicago on social good and show examples from projects in areas such as education, healthcare, energy, transportation and public safety done in conjunction with governments and non-profits.

Bio:
Rayid Ghani was the Chief Scientist at the Obama for America 2012 campaign focusing on analytics, technology, and data. His work focused on improving different functions of the campaign including fundraising, volunteer, and voter mobilization using analytics, social media, and machine learning; his innovative use of machine learning and data mining in Obama's reelection campaign received broad attention in the media such as the New York Times, CNN, and others. Before joining the campaign, Rayid was a Senior Research Scientist and Director of Analytics research at Accenture Labs where he led a technology research team focused on applied R&D in analytics, machine learning, and data mining for large-scale & emerging business problems in various industries including healthcare, retail & CPG, manufacturing, intelligence, and financial services. In addition, Rayid serves as an adviser to several start-ups in Analytics, is an active organizer of and participant in academic and industry analytics conferences, and publishes regularly in machine learning and data mining conferences and journals.

Thorsten Joachims
Learning with Humans in the Loop

Abstract:
Machine Learning is increasingly becoming a technology that directly interacts with human users. Search engines, recommender systems, and electronic commerce already heavily rely on adapting the user experience through machine learning, and other applications are likely to follow in the near future (e.g., autonomous robotics, smart homes, gaming). In this talk, I argue that learning with humans in the loop requires learning algorithms that explicitly account for human behavior, their motivations, and their judgment of performance. Towards this goal, the talk explores how integrating microeconomic models of human behavior into the learning process leads to new learning models that no longer reduce the user to a "labeling subroutine". This motivates an interesting area for theoretical, algorithmic, and applied machine learning research with connections to rational choice theory, econometrics, and behavioral economics.

Bio:
Thorsten Joachims is a Professor of Computer Science at Cornell University. His research interests center on a synthesis of theory and system building in machine learning, with applications in language technology, information retrieval, and recommendation. His past research focused on support vector machines, text classification, structured output prediction, convex optimization, learning to rank, learning with preferences, and learning from implicit feedback. In 2001, he finished his dissertation advised by Prof. Katharina Morik at the University of Dortmund. From there he also received his Diplom in Computer Science in 1997. Between 2000 and 2001 he worked as a PostDoc at the GMD Institute for Autonomous Intelligent Systems. From 1994 to 1996 he was a visiting scholar with Prof. Tom Mitchell at Carnegie Mellon University.

Ulrike von Luxburg
Unsupervised learning with graphs: a theoretical perspective

Abstract:
Applying a graph-based learning algorithm usually requires a large amount of data preprocessing. As always, such preprocessing can be harmful or helpful. In my talk I am going to discuss statistical and theoretical properties of various preprocessing steps. We consider questions such as: Given data that does not have the form of a graph yet, what do we loose when transforming it to a graph? Given a graph, what might be a meaningful distance function? We will also see that graph-based techniques can lead to surprising solutions to preprocessing problems that a priori don't involve graphs at all.

Bio:
Ulrike von Luxburg is a professor for computer science/machine learning at the University of Hamburg. Her research focus is the theoretical analysis of machine learning algorithms, in particular for unsupervised learning and graph algorithms. She is (co)-winner of several best student paper awards (NIPS 2004 and 2008, COLT 2003, 2005 and 2006, ALT 2007). She did her PhD in the Max Planck Institute for Biological Cybernetics in 2004, then moved to Fraunhofer IPSI in Darmstadt, before returning to the Max Planck Institute in 2007 as a research group leader for learning theory. Since 2012 she is a professor for computer science at the University of Hamburg.

Christopher Re
Making Systems That use Statistical Reasoning Easier to Build and Maintain over Time

Abstract:
The question driving my work is, how should one deploy statistical data-analysis tools to enhance data-driven systems? Even partial answers to this question may have a large impact on cience, government, and industry-each of whom are increasingly turning to statistical techniques to get value from their data.

To understand this question, my group has built or contributed to a diverse set of data-processing systems: a system, called GeoDeepDive, that reads and helps answer questions about the geology literature; a muon filter that is used in the IceCube neutrino telescope to process over 250 million events each day in the hunt for the origins of the universe; and enterprise applications with Oracle and Pivotal. This talk will give an overview of the lessons that we learned in these systems, will argue that data systems research may play a larger role in the next generation of these systems, and will speculate on the future challenges that such systems may face.

Bio:
Christopher (Chris) Re is an assistant professor in the Department of Computer Science at Stanford University. The goal of his work is to enable users and developers to build applications that more deeply understand and exploit data. Chris received his PhD from the University of Washington in Seattle under the supervision of Dan Suciu. For his PhD work in probabilistic data management, Chris received the SIGMOD

2010 Jim Gray Dissertation Award. Chris's papers have received four best-paper or best-of-conference citations, including best paper in PODS 2012, best-of-conference in PODS 2010 twice, and one best-of-conference in ICDE 2009). Chris received an NSF CAREER Award in 2011 and an Alfred P. Sloan fellowship in 2013.

John Shawe-Taylor
Deep-er Kernels

Abstract:
Kernels can be viewed as shallow in that learning is only applied in a single (output) layer. Recent successes with deep learning highlight the need to consider learning richer function classes. The talk will review and discuss methods that have been developed to enable richer kernel classes to be learned. While some of these methods rely on greedy procedures many are supported by statistical learning analyses and/or convergence bounds. The talk will highlight the trade-offs involved and the potential for further research on this topic.

Bio:
John Shawe-Taylor obtained a PhD in Mathematics at Royal Holloway, University of London in 1986 and joined the Department of Computer Science in the same year. He was promoted to Professor of Computing Science in 1996. He moved to the University of Southampton in 2003 to lead the ISIS research group. He was Director of the Centre for Computational Statistics and Machine Learning at University College, London between July 2006 and September 2010. He has coordinated a number of European wide projects investigating the theory and practice of Machine Learning, including the PASCAL projects. He has published over 300 research papers with more than 25000 citations. He has co-authored with Nello Cristianini two books on kernel approaches to machine learning: 'An Introduction to Support Vector Machines' and 'Kernel Methods for Pattern Analysis'.

Industrial Track Invited Speakers

Andreas Antrup
ML and Business: A Love-Hate Relationship

Abstract:
Based on real world examples. the talk explores common gaps in the mutual understanding of the business and the analytical side; particular focus shall be on misconceptions of the needs and expectations of business people and the resulting problems. It also touches on some approaches to bridge these gaps and build trust. At the end we shall discuss possibly under-researched areas that may open the doors to a yet wider usage of ML principles and thus unlock more of its value and beauty.

Bio:
Andreas Antrup heads Data Intelligence at Zalando - Europe's leading online shop for shoes and fashion. After brief stints in entrepreneurship and banking he joined Zalando in 2011 to build analytics and data-driven automation. Having received a Diplom in business administration from WHU - Otto Beisheim School of Management in 2007, Andreas changed to economics at the University of Edinburgh to graduate with an MSc in 2008 and a PhD in 2011. Together with a team of machine learners, physists, econometricians and others he now drives predictive analytics across the value chain of Zalando.

Ralf Herbrich
Bayesian Learning in Online Service: Statistics Meets Systems

Abstract:
Over the past few years, we have entered the world of big and structured data - a trend largely driven by the exponential growth of Internet-based online services such as Search, eCommerce and Social Networking as well as the ubiquity of smart devices with sensors in everyday life. This poses new challenges for statistical inference and decision-making as some of the basic assumptions are shifting:

- The ability to optimize both the likelihood and loss functions
- The ability to store the parameters of (data) models
- The level of granularity and 'building blocks' in the data modeling phase
- The interplay of computation, storage, communication and inference and decision-making techniques

In this talk, I will discuss the implications of big and structured data for Statistics and the convergence of statistical model and distributed systems. I will present one of the most versatile modeling techniques that combines systems and statistical properties - factor graphs - and review a series of approximate inference techniques such as distributed message passing. The talk will be concluded with an overview of real-world problems at Amazon.

Bio:
Ralf is Director of Machine Learning Science at Amazon Berlin, Germany. In 2011, he worked at Facebook leading the Unified Ranking and Allocation team. This team is focused on building horizontal large-scale machine learning infrastructure for learning user-action-rate predictors that enabled unified value experiences across the products. Ralf joined Microsoft Research in 2000 as a Postdoctoral researcher and Research Fellow of the Darwin College Cambridge. From 2006 - 2010, together with Thore Graepel, he was leading the Applied Games and Online Services and Advertising group which engaged in research at the intersection of machine learning and computer games and in the areas of online services, search and online advertising combining insights from machine learning, information retrieval, game theory, artificial intelligence and social network analysis. From 2009 to 2011, he was Director of Microsoft's Future Social Experiences (FUSE) Lab UK working on the development of computational intelligence technologies on large online data collections.

Prior to joining Microsoft, Ralf worked at the Technical University Berlin as a teaching assistant where I obtained both a diploma degree in Computer Science in 1997 and a Ph.D. degree in Statistics in 2000. Ralf's research interests include Bayesian inference and decision making, computer games, kernel methods and statistical learning theory. Ralf is one of the inventors of the Drivatars™ system in the Forza Motorsport series as well as the TrueSkill™ ranking and matchmaking system in Xbox 360 Live. He also co-invented the adPredictor click-prediction technology launched in 2009 in Bing's online advertising system.

Jean-Paul Schmetz
Machine Learning in a large diversified Internet Group

Abstract:
The talk covers a wide survey of the use of machine learning techniques across a large number of subsidiaries (40+) of an Internet group (Burda Digital) with special attention to issues regarding (1) personnel training in state of the art techniques, (2) management buy-in of complex non interpretable results and (3) practical and measurable bottom line results/solutions.

Bio:
Jean-Paul is the Chief Scientist of Hubert Burda Media - a global media company. He is also the founder and CEO of 10betterpages GmbH and serves on different board most notably on the Board of Directors of XING AG (a leading online professional network) and as a director at Hackfwd. He was the CTO and CEO of Burda Digital from 1996 to 2003. Jean-Paul

received a Master's Degree in Philosophy (magna cum laude) from the University of Louvain (Belgium) and a B.A. in Economics from the University of Louvain. He is currently enrolled in a post-graduate program in Computer Science at Stanford University. Jean-Paul is fluent in French, English, German and Dutch. He is an avid mathematician, hardware hacker and is fluent with all the latest software technologies.

Hugo Zaragoza
Some of the Problems and Applications of Opinion Analysis

Abstract:
Websays strives to provide the best possible analysis of online conversation to marketing and social media analysts. One of the obsessions of Websays is to provide "near-man-made" data quality at marginal costs. I will discuss how we approach this problem using innovative machine learning and UI approaches.

Bio:
Hugo Zaragoza is the founding CEO of Websays, a company dedicated to the analysis of conversations and opinions online. Hugo has been a researcher at the frontier of Natural Language Processing, Machine Learning and Search (or Information Retrieval) for over ten years. At Yahoo! Research (Barcelona) Hugo led the Natural Language Retrieval group from 2006 to 2011. From 2001 to 2006 Hugo worked at Microsoft Research (Cambridge, UK). Before this Hugo obtained a Ph.D. in probabilistic inference models for text analysis at the U. Paris 6.

PROGRAMME AT A GLANCE

	Monday 23/9		Tuesday 24/9		Wednesday 25/9		Thursday 26/9		Friday 27/9
09:00	Workshops & Tutorials	09:00	Invited talk by Christopher Re (plenary)	09:00	Invited talk by Rayid Ghani (plenary)			09:00	Invited talk by John Shawe-Taylor (plenary)
						09:30	Invited talk by Thorsten Joachims (plenary)		
		10:00	Test-of-time presentation (plenary)	10:00	Coffee break			10:00	Coffee break
10:30	Coffee break	10:30	Coffee break	10:30	Nectar (1) (E) Active Learning and Optimization (D) Networks (2) (C) Structured Output, Multi-task (A)	10:30	Coffee break		
11:00	Workshops & Tutorials	11:00	Reinforcement Learning (E) Networks (1) (D) Privacy and Security (C) Ranking and Recommender Systems (A)			11:00	Industrial track (1) (E) Sequential Pattern Mining (D) Graphical Models (C) Unsupervised Learning (A)	10:45	Workshops & Tutorials
12:30	Lunch break	12:25	Lunch break	12:10	Lunch break	12:25	Lunch break	12:15	Lunch break
14:00	Workshops & Tutorials	14:15	Markov Decision Processes (E) Tensor Analysis & Dimensionality Reduction (D) Biomedical Applications (C) Demo spotlights (A)	14:00	Nectar (2) (E) Models for Sequential Data (D) Graph Mining (C) Natural Language Processing & Probabilistic Models (A)	14:15	Industrial track (2) (E) Dynamic Graphs (D) Statistical Learning (1) (C) Evaluation & kNN (A)	13:45	Workshops & Tutorials
15:30	Coffee break							15:15	Coffee break
16:00 ↓ 17:30	Workshops & Tutorials	15:55	Coffee break	15:40	Coffee break	15:40	Coffee break	15:45	Workshops & Tutorials
		16:25 ↓ 18:05	Inverse RL & RL Applications (E) Matrix Analysis (D) Applications (C) Semi-supervised Learning (A)	16:10 ↓ 17:50	Subgroup Discovery & Streams (E) Multi-label Classification & Outlier Detection (D) Ensembles (C) Bayesian Learning (A)	16:10 ↓ 17:35	Sequence & Time Series Analysis (E) Declarative Data Mining & Meta Learning (D) Topic Models (C) Statistical Learning (2) (A)	17:15	Coffee break
								17:30 ↓ 19:00	Workshops & Tutorials
						18:00 ↓ 19:00	Community meeting		
19:00	Openings & awards								
19:30	Invited Talk by Ulrike von Luxburg (plenary)	19:30	Poster session & demos (Brissasol)	19:30	Conference dinner (Klášterní Restaurant)	19:30	Poster session (Brissasol)		
20:30	Welcome Reception								

Room name abbreviations: A - Aplaus | **C** - Ceremonie | **D** - Dialog | **E** - Euforie

Monday 23/9

	Brissasol	R1	R2	R3	R4	R5	R6	R7
09:00	T: Performance Evaluation of Machine Learning Algorithms	W: Mining Ubiquitous and Social Environments	T: Second Order Learning	W: Scalable Decision Making: Uncertainty, Imperfection, Deliberation	W: Data Analytics for Renewable Energy Integration	W: Reinforcement Learning with Generalized Feedback	W: Languages for Data Mining and Machine Learning	W: Music and Machine Learning
10:30	Coffee Break							
11:00	T: Performance Evaluation of Machine Learning Algorithms	W: Mining Ubiquitous and Social Environments	T: Second Order Learning	W: Scalable Decision Making: Uncertainty, Imperfection, Deliberation	W: Data Analytics for Renewable Energy Integration	W: Reinforcement Learning with Generalized Feedback	W: Languages for Data Mining and Machine Learning	W: Music and Machine Learning
12:30	Lunch break							
14:00	W: Data Mining on Linked Data	W: Mining Ubiquitous and Social Environments	T: Mining and Learning with Network-Structured Data	W: Scalable Decision Making: Uncertainty, Imperfection, Deliberation	T: Discovering Roles and Anomalies in Graphs: Theory and Applications	W: Reinforcement Learning with Generalized Feedback	W: Languages for Data Mining and Machine Learning	W: Music and Machine Learning
15:30	Coffee Break							
16:00 ↓ 17:30	W: Data Mining on Linked Data	W: Mining Ubiquitous and Social Environments	T: Mining and Learning with Network-Structured Data	W: Scalable Decision Making: Uncertainty, Imperfection, Deliberation	T: Discovering Roles and Anomalies in Graphs: Theory and Applications	W: Reinforcement Learning with Generalized Feedback	W: Languages for Data Mining and Machine Learning	W: Music and Machine Learning
19:00	Opening & Awards Ceremony (plenary)							
19:30	Invited Talk by Ulrike von Luxburg (plenary)							
20:30	Welcome Reception (U Hájků)							

Friday 27/9

	Brissasol	R1	R2	R3	R4	R5	R6	R7
09:00	Invited Talk by John Shawe-Taylor (plenary)							
10:00	Coffee Break							
10:45	T: Statistically Sound Pattern Discovery	Nameling Discovery Challenge 2013	W: Solving Complex Machine Learning Problems with Ensemble Methods	T: Web Scale Information Extraction	W: Real-World Challenges for Data Stream Mining	W: Sports Analytics	W: Tensor Methods in Machine Learning	W: New Frontiers in Mining Complex Patterns
12:15	Lunch break							
13:45	T: Statistically Sound Pattern Discovery	Nameling Discovery Challenge 2013	W: Solving Complex Machine Learning Problems with Ensemble Methods	T: Web Scale Information Extraction	W: Real-World Challenges for Data Stream Mining	W: Sports Analytics	W: Tensor Methods in Machine Learning	W: New Frontiers in Mining Complex Patterns
15:15	Coffee Break							
15:45	T: Algorithmic Techniques for Modeling and Mining Large Graphs	Nameling Discovery Challenge 2013	W: Solving Complex Machine Learning Problems with Ensemble Methods	T: Multi-Agent Reinforcement Learning	W: Real-World Challenges for Data Stream Mining	W: Sports Analytics	W: Tensor Methods in Machine Learning	W: New Frontiers in Mining Complex Patterns
17:15	Coffee Break							
17:30 ↓ 19:00	T: Algorithmic Techniques for Modeling and Mining Large Graphs	Nameling Discovery Challenge 2013	W: Solving Complex Machine Learning Problems with Ensemble Methods	T: Multi-Agent Reinforcement Learning	W: Real-World Challenges for Data Stream Mining	W: Sports Analytics	W: Tensor Methods in Machine Learning	W: New Frontiers in Mining Complex Patterns

MONDAY 23 SEPTEMBER 2013

MONDAY MORNING TUTORIALS

Performance Evaluation of Machine Learning Algorithms

Mohak Shah and Nathalie Japkowicz
Time: 09:00-12:30
Room: Brissasol

Machine learning and Data mining are increasingly being applied to a wide range of domains and are quickly reaching maturity to a point that various instances of technologies are commoditized. Due to their inherent interdisciplinary nature the fields draw researchers from varying backgrounds. Not only that, data scientists and other experts practicing these techniques on a regular basis are more broad in terms of their backgrounds. It is of critical importance in such a case that these researchers and practitioners are aware of both the proper methodologies and the respective issues that arise in terms of evaluating novel learning approaches. This tutorial aims at educating as well as getting the broad machine-learning and data science community to discuss these critical issues in performance evaluation. The tutorial will cover various aspects of the evaluation process with a focus on classification algorithms. We will discuss various choice decision vis-a-vis the issues, assumptions and constraints involved at various steps of this process. It will also present R and WEKA tools that can be utilized to apply them. The tutorial will span four areas of classifier evaluation:

- Performance Measures (evaluation metrics and graphical methods)
- Error Estimation/Re-sampling Techniques
- Statistical Significance Testing
- Issues in data Set Selection and evaluation benchmarks design

The tutorial will, in part, be based on the book by the presenters on the subject: Japkowicz, Nathalie and Shah, Mohak, Evaluating Learning Algorithms: A Classification Perspective, Cambridge University Press (2011).

The purpose of the tutorial is to promote an appreciation of the need for rigorous and objective evaluation and an understanding of the available alternatives along with their assumptions, constraints and context of application. Machine learning researchers and practitioners alike will all benefit from the contents of the tutorial, which discusses the need for sound evaluation strategies, practical approaches and tools for evaluation, going well beyond those described in existing machine learning and data mining textbooks, so far.

Second Order Learning

Koby Crammer
Time: 09:00-12:30
Room: R2

Second order algorithms are included both in optimization (e.g. Newton method) and online learning, where they are motivated both geometrically (i.e. second order perceptron) and from statistical properties of natural language (i.e. confidence weighted learning). These methods converge fast and achieve state-of-the-art results in many natural language processing tasks. Indeed, some of these methods were in fact motivated from natural language processing (NLP), where feature-occurrence statistics follows an heavy tail distribution, yet first-order algorithms design nor their analysis are directly related to these NLP properties. For example, in some sentiment classification tasks, some predictive features are very common, yet most of them are relatively rare, indicating that modeling even infrequent features may be useful for learning.

In this tutorial, we will focus on old and recent techniques for learning that exploit such statistics and their analysis. We will introduce systematically second order learning, covering both basic principles and specific algorithms and tools, such as second order perceptron, adaptive gradient methods, and confidence-weighted learning, as well as usages in various tasks and settings: active learning, domain adaptation, confidence estimation and, bandit prediction and selective sampling.

MONDAY AFTERNOON TUTORIALS

Discovering Roles and Anomalies in Graphs: Theory and Applications

Tina Eliassi-Rad and Christos Faloutsos
Time: 14:00-17:30
Room: R4

Given a graph, how can we find suspicious nodes? How can we automatically discover roles for nodes? Roles are compact summaries of a node's behavior that generalize across networks. For example, one role could be 'star'-with a star node being both influential and having a low neighborhood overlap. Are there good features that we can extract for nodes that indicate role- membership? What are the different applications in which these discovered roles can be effectively used? The objective of this tutorial is to provide a concise and intuitive overview of the most important concepts and tools, which can detect roles (or functions) for nodes in both static and dynamic graphs. We review the state of the art in three related fields: (a) community discovery, (b) equivalences (from sociology), and (c) propositionalisation (from multi-relational data mining). The emphasis of this tutorial is to give the intuition behind these powerful mathematical concepts and tools, which are usually lost in the various technical literatures, as well as to give case studies that illustrate their practical use.

Mining and Learning with Network-Structured Data

Jan Ramon
Time: 14:00-17:30
Room: R2

In recent years, several relevant relations between the mining of large data networks and different branches of computer science and mathematics have been revealed, e.g. graph theory, statistical physics, complex systems, algorithmics and spectral graph theory. Exploiting these relations and combining aspects of several of these fields is an exciting direction of research often leading to surprising results. This tutorial aims at giving an introduction to and an overview of the several fields related to mining networks and their basic principles, and at providing examples of results linking these fields to data mining tasks. The overall goal is give a basic introduction and point to interesting links, triggering discussions and new ideas for future research.

The tutorial will consist of three parts. The first part will present an overview of several related fields, their essential concepts and questions, and the relations between them. These include algorithmics, graph theory, pattern matching, pattern mining, statistical physics, complex systems, link prediction and statistical relational learning and spectral graph theory. A second part will consider patterns in data networks and will consist of a discussion on the emergence of (local) patterns, 0-1 laws, (graph) algorithmic and complexity theoretic aspects of pattern matching and pattern mining, sampling approaches, criteria for measuring statistical significance or other qualities of patterns, relations to random graph models. A final part will concern learning and will survey machine learning settings in graphs, basic algorithmic ideas, relations to the pattern mining approaches of the previous part, and relations to statistical physics based generative graph models

MONDAY WORKSHOP

Data Mining on Linked Data

Claudia d'Amato, Petr Berka, Vojtech Svatek and Krzysztof Wecel
Time: 14:00-17:30
Room: Brissasol

Linked data (LD), published on the web in RDF format, represents a novel type of data source that has been so far nearly untouched by advanced data mining methods. It breaks down many traditional assumptions on source data and thus represents a number of challenges. While the individual published datasets typically follow a relatively regular structure, the presence of semantic links among them makes the resulting „hyper-dataset" akin to general graph datasets. On the other hand, compared to graphs such as social networks, there is a larger variety of link types in the graph. The datasets have also been published for entirely different purposes, such as statistical data publishing based on legal commitment of government bodies vs. publishing of encyclopedic data by internet volunteers vs. data sharing within a researcher community. This introduces further data modeling heterogeneity and uneven degree of completeness and reliability. Finally, the amount and diversity of resources as well as their link sets is steadily growing, which allows for inclusion of new linked datasets into the mining dataset nearly on the fly, at the same time, however, making the feature selection problem extremely hard.

DMoLD'13 is envisaged to bring the semantic web and data mining communities closer together and to foster further research in their intersection. It features both an Open Track and a Linked Data Mining Challenge (LDMC), the latter relying upon a benchmark dataset from the highly topical public procurement domain.

14:00-14:10 **DMoLD'13 Opening**

14:10-15:05 **Invited talk: Exploiting Linked Open Data as Background Knowledge in Data Mining**
Heiko Paulheim

15:05-15:30 **Lattice Based Data Access (LBDA): An Approach for Relating Data and Linked Open Data in Biology**
Mehwish Alam, Melisachew Wudage Chekol, Adrien Coulet, Amedeo Napoli and Malika Smail-Tabbone

15:30-16:00 **Coffee break**

16:00-16:25 **A Fast and Simple Graph Kernel for RDF**
Gerben Klaas Dirk de Vries and Steven de Rooij

16:25-16:40 **Challenge Track Intro: Linked Data Mining Challenge (LDMC) 2013 Summary**
Vojtech Svatek, Jindrich Mynarz and Petr Berka

16:40-16:55 **Graph Kernels for Task 1 and 2 of the Linked Data Data Mining Challenge 2013**
Gerben Klaas Dirk de Vries

16:55-17:10 **A Machine Learning approach to the Linked Data Mining Challenge 2013**
Eneldo Loza Mencia, Simon Holthausen, Axel Schulz and Frederik Janssen

17:10-17:30 **Open Discussion: Data Mining Supported by Linked Data: Opportunities and Challenges**

MONDAY WORKSHOP

MUSE: Mining Ubiquitous and Social Environments

Martin Atzmueller and Christoph Scholz
Time: 09:00-15:30
Room: R1

Mining in ubiquitous and social environments is an emerging area of research focusing on advanced systems for data mining in such distributed and network-organized systems. It also integrates some related technologies such as activity recognition, social web mining, privacy issues and privacy-preserving mining, predicting user behavior, etc.

MUSE aims to promote an interdisciplinary forum for researchers working in the fields of ubiquitous computing, mobile sensing, social web, Web 2.0, and social networks which are interested in utilizing data mining in a ubiquitous setting. In short, we want to accelerate the process of identifying the power of advanced data mining operating on data collected in ubiquitous and social environments, as well as the process of advancing data mining through lessons learned in analyzing these new data.

09:00-09:15 **MUSE 2013 Welcome and Introduction**

09:15-10:30 **Invited talk: Mining Information Propagation Traces in Social Networks**
Francesco Bonchi

10:30-11:00 **Coffee break**

11:00-11:40 **APP: Aperiodic and Periodic Model for Long-Term Human Mobility Prediction Using Ambient Simple Sensors**
Danaipat Sodkomkham

11:40-12:00 **Subgroup Analytics on Ubiquitous Data–Onto Perceptions and Semantics**
Martin Atzmueller and Juergen Mueller

12:00-12:20 **Open Smartphone Data for Mobility and Utilization Analysis in Ubiquitous Environments**
Jochen Streicher, Nico Piatkowski, Katharina Morik and Olaf Spinczyk

12:20-14:00 **Lunch break**

14:00-14:20 **Content-Based Geo-Location Detection for Placing Tweets Pertaining To Trending News on Map**
Saurabh Khanwalkar, Marc Seldin, Amit Srivastava, Anoop Kumar and Sean Colbath

14:20-15:00 **Learning the Shortest Path for Text Summarisation**
Emmanouil Tzouridis and Ulf Brefeld

15:00-15:30 **Discussion + Closing**

MONDAY WORKSHOP

SCALE: Scalable Decision Making: Uncertainty, Imperfection, Deliberation

Tatiana V.Guy and Miroslav Kárný
Time: 08:55-17:30
Room: R3

Machine learning (ML) and knowledge discovery both use and serve to decision making (DM), which has to cope with uncertainty, incomplete knowledge, problem and data complexity as well as limited cognitive and evaluating capabilities (imperfection) of the involved heterogeneous multiple participants (aka agents, decision makers, components, controllers, classifiers, etc.). Contemporary DM deals with complex systems characterised by heterogeneous components and their goal-motivated dynamic interactions. The individual participants are selfish, i.e. follow their individual goals. There is no well-justified way to influence or describe the resulting collective behaviour of such a system via a well-proved combination of the selfish components. Economic and natural sciences describe concepts governing the functioning of systems of selfish participants as well as ways influencing their behaviour. However, the majority of solutions rely on the human moderator/manager controlling such a system. Sophisticated ML and AI solutions developed consider artificial moderators (for instance, automatic traders used in markets, e-democracy support) as well.

The SCALE workshop aims to exploit the knowledge and experience of multi-disciplinary scientific community and to extract a set of fundamental concepts describing a phenomenon of dynamic decision making with interacting imperfect selfish participants.

08:55-09:00 **Opening Session**

09:00-09:35 **Scalable Information Aggregation from Weak Information Sources**
Stephen Roberts

09:35-10:10 **Designing Societies of Robots**
David Rios Insua and Pablo G. Esteban

10:10-10:30 **Cooperative Dimensionality Reduction for Intelligent Feature Selection in Individualised Medicine**
Dietlind Zühlke, Gernoth Grunst and Kerstin Röser

10:30-11:00 **Coffee Break**

11:00-11:35 **Preference Elicitation for Social Choice: A Study in Stable Matching and Voting**
Craig Boutilier, Joanna Drummond and Tyler Lu

11:35-12:00 **Poster spotlights**

12:00-12:30 **Posters and Demonstrations:**

Granger Lasso Causal Models in High Dimensions–Application to Gene Expression Regulatory Networks
Katerina Hlavackova-Schindler and Hamed Bouzari

On Approximate Fully Probabilistic Design of Decision Making Strategies
Miroslav Kárný

Preliminaries of Probabilistic Hierarchical Fault Detection
Ladislav Jirsa, Lenka Pavelková and Kamil Dedecius

What Lies Beneath Players' Non-Rationality in Ultimatum Game?
Zuzana Knejflová, Galina Avanesyan, Tatiana V. Guy and Miroslav Kárný

A Note on Weighted Combination Methods for Probability Estimation
Vladimíra Sečkárová

Estimating Efficiency Offset between Two Groups of Decision-Making Units
Karel Macek

12:30-14:00 **Lunch Break**

14:00-14:20 **Economic Prediction Using Heterogeneous Data Streams from the World Wide Web**
Abby Levenberg, Edwin Simpson, Stephen Roberts and Georg Gottlob

14:20-14:55 **A Unified View on Roots of Imperfection**
Miroslav Kárný and Tatiana V. Guy

14:55-15:30 **Predictive Information and the Brain's Internal Time**
Naftali Tishby

15:30-16:00 **Coffee break. Posters & Demos (cont.)**

16:00-16:20 **Belief CSP: A New CSP Framework Under Uncertainty**
Aouatef Rouahi, Kais Ben Salah and Khaled Ghédira

16:20-16:55 **Stimulus Evaluation and Response Systems Studied by Reaction Times in Decision Making Tasks**
Alessandro E.P. Villa

16:55-17:30 **Panel Discussion, Closing Remarks**

MONDAY WORKSHOP

Data Analytics for Renewable Energy Integration

Wei Lee Woon, Stuart Madnick and Zeyar Aung
Time: 09:00-12:00
Room: R4

Climate change, the depletion of natural resources and rising energy costs have led to an increasing focus on renewable sources of energy. A lot of research has been devoted to the technologies used to extract energy from these sources; however, equally important is the storage and distribution of this energy in a way that is efficient and cost effective. Achieving this would generally require integration with existing energy infrastructure. The challenge of renewable energy integration is inherently multidisciplinary and is particularly dependent on the use of techniques from the domains of data analytics, pattern recognition and machine learning. Examples of relevant research topics include the forecasting of electricity supply and demand, the detection of faults, demand response applications and many others. This workshop will provides a forum where interested researchers from the various related domains will be able to present and discuss their findings.

09:00-09:15 **Welcome Address**

09:20-09:40 **Statistical Learning for Short-Term Photovoltaic Power Predictions**
Björn Wolff, Elke Lorenz and Oliver Kramer

09:40-10:00 **First Steps Towards a Systematical Optimized Strategy for Solar Energy Supply Forecasting**
Robert Ulbricht, Ulrike Fischer, Wolfgang Lehner and Hilko Donker

10:00-10:20 **Aggregation of Features for Wind Energy Prediction with Support Vector Regression and Nearest Neighbors**
Nils André Treiber, Justin Heinermann and Oliver Kramer

10:30-11:00 **Coffee Break**

11:00-11:20 **Fault Detection of Large Amounts of Photovoltaic Systems**
Patrick Traxler

11:20-11:40 **Drivers of variability in energy consumption**
Adrian Albert, Timnit Gebru, Jerome Ku, Jungsuk Kwac, Jure Leskovec and Ram Rajagopal

11:40-12:00 **Load Decomposition and Profiling for "Smart Grid" Demand-Side Management Applications**
Sasa Djokic and Andreas Paisios

MONDAY WORKSHOP

Reinforcement Learning with Generalized Feedback

Johannes Fürnkranz and Eyke Hüllermeier
Time: 09:30-17:30
Room: R5

In recent years, different generalizations of the standard setting of reinforcement learning have emerged; in particular, several attempts have been made to relax the quite restrictive requirement for numeric feedback and to learn from different types of more flexible training information. Examples of generalized settings of that kind include apprenticeship learning, inverse reinforcement learning, multi-objective reinforcement learning, and preference-based reinforcement learning. Learning in these generalized frameworks can be considerably harder than learning in MDPs because rewards cannot be easily aggregated over different states. The main goal of this workshop is to help in unifying and streamlining research on generalizations of standard reinforcement learning, which, for the time being, seem to be pursued in a rather uncoordinated manner.

09:30-09:40 **Opening Remarks**
Eyke Hüllermeier and Johannes Fürnkranz

09:40-10:30 **Invited Talk by Michele Sebag**

10:30-11:00 **Coffee Break**

Human Interaction for Effective Reinforcement Learning
L. Adrian Leon, Ana C. Tenorio and Eduardo F. Morales

11:25-11:50 **Interactive Robot Education**
Riad Akrour, Marc Schoenauer and Michele Sebag

11:50-12:15 **Interactive Q-Learning with Ordinal Rewards and Unreliable Tutor**
Paul Weng, Robert Busa-Fekete and Eyke Hüllermeier

12:15-12:40 **Iterative Model Refinement of Recommender MDPs based on Expert Feedback**
Omar Zia Khan, Pascal Poupart and John Mark Agosta

12:40-14:00 **Lunch Break**

RL with Non-numerical Feedback

14:00-14:25 **Preference-Based Reinforcement Learning: A Preliminary Survey**
Christian Wirth and Johannes Fürnkranz

14:25-14:50 **Preference-based Evolutionary Direct Policy Search**
Robert Busa-Fekete, Balazs Szörenyi, Paul Weng, Weiwei Cheng and Eyke Hüllermeier

14:50-15:15 **A Learning Agent for Parameter Estimation in Speeded Tests**
Daniel Bengs and Ulf Brefeld

15.15-15.30 **Discussion**

15:30-16:00 **Coffee Break**

Inverse RL and Multi-Dimensional Feedback

16:00-16:25 **Applying Inverse Reinforcement Learning to Medical Records of Diabetes**
Hideki Asoh, Masanori Shiro, Shotaro Akaho, Toshihiro Kamishima,Koiti Hasida, Eiji Aramaki and Takahide Kohro

16:25-16:50 **Multiobjective Reinforcement Learning Using Adaptive Dynamic Programming And Reservoir Computing**
Mohamed Oubbati, Timo Oess, Christian Fischer and Günther Palm

16:50-17:15 **Comparative Evaluation of Reinforcement Learning with Scalar Rewards and Linear Regression with Multidimensional Feedback**
Petar Kormushev and Darwin G. Caldwell

17:15-17.30 **Discussion**

MONDAY WORKSHOP

Languages for Data Mining and Machine Learning

Bruno Cremilleux, Luc De Raedt, Paolo Frasconi and Tias Guns
Time: 09:00-17:30
Room: R6

Research in Data Mining and Machine Learning has progressed significantly in the last decades, through the development of advanced algorithms and techniques. In the past few years there has been a growing attention to the development of languages for use in data mining and machine learning. Such languages provide common buildings blocks and abstractions, and can provide an alternative interface to advanced algorithms and systems that can greatly increase the utility of such systems.

The workshop aims to bring together researchers and stimulate discussions on languages for data mining and machine learning. Its main motivation is the believe that designing generic and declarative modeling languages for data mining and machine learning, together with efficient solving techniques, is an attractive direction that can boost scientific progress.

09:00-09:15 **LML'13 Introduction by the organizers**

09:15-10:15 **Invited talk by Dino Pedreschi**

10:15-10:30 **A query language for constraint-based clustering [published]**
Antoine Adam and Hendrik Blockeel

10:30-11:00 **Coffee Break**

11:00-11:30 **Declarative In-Network Sensor Data Analysis**
George Valkanas, Ixent Galpin, Alasdair J.G. Gray, Alvaro A. A. Fernandes, Norman W. Paton and Dimitrios Gunopulos

11:30-12:00 **Mining (Soft-) Skypatterns using Constraint Programming**
Willy Ugarte Rojas, Patrice Boizumault, Samir Loudni, Bruno Cremilleux and Alban Lepailleur

12:00-12:30 **Query Rewriting for Rule Mining in Databases**
Brice Chardin, Emmanuel Coquery, Benjamin Gouriou, Marie Pailloux and Jean-Marc Petit

12:30-14:00 **Lunch Break**

14:00-14:30 **A Constraint Programming Approach for Mining Sequential Patterns in a Sequence Database**
Jean-Philippe Metivier, Samir Loudni and Thierry Charnois

14:30-15:00 **The representation of sequential patterns and their projections within Formal Concept Analysis**
Aleksey Buzmakov, Elias Egho, Nicolas Jay, Sergei O. Kuznetsov, Amedeo Napoli and Chedy Raissi

15:00-15:15 **Language of Conclusions and Formal Frame for Data Mining with Association Rules**
Rauch Jan

15:15-15:30 **Lower and upper queries for graph-mining**
Amina Kemmar, Yahia Lebbah, Samir Loudni and Mohammed Ouali

15:30-16:00 **Coffee Break**

16:00-16:30 **API design for machine learning software: experiences from the scikit-learn project**
Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Muller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake Vanderplas, Arnaud Joly, Brian Holt and Gael Varoquaux

16:30-16:45 **ParaMiner: A generic pattern mining algorithm for multi-core architectures [published]**
Benjamin Negrevergne, Alexandre Termier, Marie-Christine Rousset and Jean-Francois Mehaut

16:45-17:00 **A declarative query language for statistical inference [Extended Abstract]**
Gitte Vanwinckelen and Hendrik Blockeel

17:00-17:30 **Discussion session**

MONDAY WORKSHOP
Music and Machine Learning

Rafael Ramirez, Darrell Conklin and José Manuel Iñesta
Time: 9:00-17:00
Room: R7

With the current explosion and quick expansion of music in digital formats, and the computational power of modern systems, the research on machine learning and music has gained increasing popularity. As complexity of the problems investigated by researchers on machine learning and music increases, there is a need to develop new algorithms and methods to solve these problems. Machine learning has proved to provide efficient solutions to many music-related problems both of academic and commercial interest. MML 2013 concentrates around the topic of 'Intelligent Content-Based Music Processing' and will welcome contributions describing machine learning approaches in this context, e.g. automatic classification of music (audio and MIDI), style-based interpreter recognition, automatic composition and improvisation, music recommender systems, and expressive performance modeling.

09:00-09:10 **MML'13 Opening**

09:10-09:30 **Dance Hit Song Science**
Dorien Herremans, David Martens, and Kenneth Sorensen

09:30-09:50 **A Hierarchical Cluster Analysis to Identify Principles of Popular Music Composition Used by The Beatles**
Douglas J. Mason

09:50-10:10 **A Deep Learning Approach to Rhythm Modelling with Applications**
Aggelos Pikrakis

10:10-10:30 **Music Emotion Recognition: The Importance of Melodic Features**
Bruno Rocha, Renato Panda and Rui Pedro Paiva

10:30-11:00 **Coffee break**

11:00-11:20 **Discovery of mediating association rules for folk music analysis**
Kerstin Neubarth, Colin G. Johnson, Darrell Conklin

11:20-11:40 **Chord Estimation using Compositional Hierarchical Model**
Matevz Pesek and Matija Marolt

11:40-12:00 **A Neural Probabilistic Model for Predicting Melodic Sequences**
Srikanth Cherla, Artur d'Avila Garcez and Tillman Weyde

12:00-12:30 **Poster Craze**

12:30-14.00 **Lunch**

14:00-15:30 **Poster Session**

An Efficient Shift-Invariant Model for Polyphonic Music Transcription
Emmanouil Benetos, Srikanth Cherla and Tillman Weyde

Timbre-based Drum Pattern Classification using Hidden Markov Models
Michael Blass

Using Mutual Proximity for Novelty Detection in Audio Music Similarity
Arthur Flexer and Dominik Schnitzer

Music Visualization Using Audio Features and Tags
Tzu-Chun Lin, Wei Lee Woon and Jimmy C. Peng

Music Emotion Recognition from Lyrics: A Comparative Study
Ricardo Malheiro, Renato Panda, Paulo Gomes and Rui Pedro Paiva

A Machine Learning Approach for Clustering Western and Non-Western Folk Music Using Low-level and Mid-level Features
Neocleous Andreas, Panteli Maria, Rafaela Ioannou, Nicolai Petkov and Christos N. Schizas

Percussive Beat tracking using real-time median filtering
Andrew Robertson, Adam Stark and Matthew E. P. Davies

Musical Onset Detection with Convolutional Neural Networks
Jan Schlüter and Sebastian Böck

Real-Time Modeling of Emotions by Linear Regression
Sergio Giraldo and Rafael Ramirez

Fusion Functions for Multiple Viewpoints
Darrell Conklin

15:30-16:00 **Coffee break**

16:00-16:20 **The Power of Less: Exemplar-based Automatic Transcription of Polyphonic Piano Music**
Ismail Ari, Ali Taylan Cemgil and Lale Akarun

16:20-16:40 **Enhanced peak picking for onset detection with recurrent neural networks**
Sebastian Böck, Jan Schlüter, Arthur Flexer and Gerhard Widmer

16:40-17:00 **Recognition of Online Handwritten Music Symbols**
Jorge Calvo-Zaragoza, Jose Oncina and Jose M. Iñesta

MONDAY INVITED TALK

Speaker: Ulrike von Luxburg
Title: Unsupervised Learning with Graphs: A Theoretical Perspective
Time: 19:30-20:30
Room: plenary

Abstract
Applying a graph-based learning algorithm usually requires a large amount of data preprocessing. As always, such preprocessing can be harmful or helpful. In my talk I am going to discuss statistical and theoretical properties of various preprocessing steps. We consider questions such as: Given data that does not have the form of a graph yet, what do we lose when transforming it to a graph? Given a graph, what might be a meaningful distance function? We will also see that graph-based techniques can lead to surprising solutions to preprocessing problems that a priori don't involve graphs at all.

Bio
Ulrike von Luxburg is a professor for computer science/machine learning at the University of Hamburg. Her research focus is the theoretical analysis of machine learning algorithms, in particular for unsupervised learning and graph algorithms. She is (co)-winner of several best student paper awards (NIPS 2004 and 2008, COLT 2003, 2005 and 2006, ALT 2007). She did her PhD in the Max Planck Institute for Biological Cybernetics in 2004, then moved to Fraunhofer IPSI in Darmstadt, before returning to the Max Planck Institute in 2007 as a research group leader for learning theory. Since 2012 she is a professor for computer science at the University of Hamburg.

TUESDAY 24 SEPTEMBER 2013

TUESDAY INVITED TALK

Making Systems That use Statistical Reasoning Easier to Build and Maintain over Time

Speaker: Christopher Re
Time: 09:00-10:00
Room: plenary

Abstract
The question driving my work is, how should one deploy statistical data-analysis tools to enhance data-driven systems? Even partial answers to this question may have a large impact on cience, government, and industry-each of whom are increasingly turning to statistical techniques to get value from their data.

To understand this question, my group has built or contributed to a diverse set of data-processing systems: a system, called GeoDeepDive, that reads and helps answer questions about the geology literature: a muon filter that is used in the IceCube neutrino telescope to process over 250 million events each day in the hunt for the origins of the universe; and enterprise applications with Oracle and Pivotal. This talk will give an overview of the lessons that we learned in these systems, will argue that data systems research may play a larger role in the next generation of these systems, and will speculate on the future challenges that such systems may face.

Bio
Christopher (Chris) Re is an assistant professor in the Department of Computer Science at Stanford University. The goal of his work is to enable users and developers to build applications that more deeply understand and exploit data. Chris received his PhD from the University of Washington in Seattle under the supervision of Dan Suciu. For his PhD work in probabilistic data management, Chris received the SIGMOD 2010 Jim Gray Dissertation Award. Chris's papers have received four best-paper or best-of-conference citations, including best paper in PODS 2012, best-of-conference in PODS 2010 twice, and one best-of-conference in ICDE 2009). Chris received an NSF CAREER Award in 2011 and an Alfred P. Sloan fellowship in 2013.

TEST OF TIME PRESENTATION

Logistic Model Trees

Authors: Niels Landwehr, Mark Hall, Eibe Frank
Time: 10:00-10:30
Room: plenary

At ECML PKDD 2003, Niels Landwehr, Mark Hall, and Eibe Frank presented a paper on „Logistic Model Trees". This paper has now been selected as the paper from ECML PKDD 2003 with the highest impact, measured after ten years. The talk will motivate Logistic Model Trees in the context of earlier work, summarize the main results of the paper, and discuss how it has influenced further work in the field of machine learning and knowledge discovery.

SESSIONS AT A GLANCE

Tue1A: Reinforcement Learning

Room: Euforie

11:00-11:20 **Learning Graph-Based Representations for Continuous Reinforcement Learning Domains**
Jan Hendrik Metzen

11:20-11:40 **Greedy Confidence Pursuit: A Pragmatic Approach to Multi-bandit Optimization**
Philip Bachman and Doina Precup

11:40-12:00 **Exploiting Multi-step Sample Trajectories for Approximate Value Iteration**
Robert Wright, Lei Yu, Steven Loscalzo and Philip Dexter

12:00-12:20 **Automatically Mapped Transfer Between Reinforcement Learning Tasks via Three-Way Restricted Boltzmann Machines**
Haitham Bou Ammar, Decebal Constantin Mocanu, Matthew Taylor, Kurt Driessens, Gerhard Weiss and Karl Tuyls

Tue1B: Networks (1)

Room: Dialog

11:00-11:25 **What Distinguish One from Its Peers in Social Networks?**
Yi-Chen Lo, Jhao-Yin Li, Mi-Yen Yeh, Shou-De Lin and Jian Pei

11:25-11:45 **Detecting Bicliques in GF[q]**
Jan Ramon, Pauli Miettinen and Jilles Vreeken

11:45-12:05 **As Strong as the Weakest Link: Mining Diverse Cliques in Weighted Graphs**
Petko Bogdanov, Ben Baumer, Prithwish Basu, Amotz Bar-Noy and Ambuj Singh

12:05-12:25 **How Robust is the Core of a Network?**
Abhijin Adiga and Anil Vullikanti

Tue1C: Privacy and Security

Room: Ceremonie

11:00-11:25 **Differential Privacy Based on Importance Weighting**
Zhanglong Ji and Charles Elkan

11:25-11:45 **Anonymizing Data with Relational and Transaction Attributes**
Giorgos Poulis, Grigorios Loukides, Aris Gkoulalas-Divanis and Spiros Skiadopoulos

11:45-12:05 **Privacy-Preserving Mobility Monitoring using Sketches of Stationary Sensor Readings**
Michael Kamp, Christine Kopp, Michael Mock, Mario Boley and Michael May

12:05-12:25 **Evasion Attacks Against Machine Learning at Test Time**
Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Srndic, Pavel Laskov, Giorgio Giacinto and Fabio Roli

Tue1D: Ranking and Recommender Systems

Room: Aplaus

11:00-11:25 **Growing a List**
Benjamin Letham, Cynthia Rudin and Katherine A Heller

11:25-11:45 **A Pairwise Label Ranking Method with Imprecise Scores and Partial Predictions**
Sebastien Destercke

11:45-12:05 **Learning Socially Optimal Information Systems from Egoistic Users**
Karthik Raman and Thorsten Joachims

12:05-12:25 **Socially Enabled Preference Learning from Implicit feedback data**
Julien Delporte, Alexandros Karatzoglou, Tomasz Matuszczyk and Stephane Canu

Tue2A: Markov Decision Processes

Room: Euforie

14:15-14:35 **Expectation Maximization for Average Reward Decentralized POMDPs**
Joni Pajarinen and Jaakko Peltonen

14:35-14:55 **Properly Acting under Partial Observability with Action Feasibility Constraints**
Caroline Carvalho Chanel and Florent Teichteil-Konigsbuch

14:55-15:15 **Iterative Model Refinement of Recommender MDPs Based on Expert Feedback**
Omar Khan, Pascal Poupart and John Mark Agosta

15:15-15:35 **Solving Relational MDPs with Exogenous Events and Additive Rewards**
Saket Joshi, Roni Khardon, Prasad Tadepalli, Aswin Raghavan and Alan Fern

15:35-15:55 **Continuous Upper Confidence Trees with Polynomial Exploration–Consistency**
Adrien Couetoux, David Auger and Olivier Teytaud

Tue2B: Tensor Analysis & Dimensionality Reduction

Room: Dialog

14:15-14:35 **An Analysis of Tensor Models for Learning on Structured Data**
Maximilian Nickel and Volker Tresp

14:35-14:55 **Learning Modewise Independent Components from Tensor Data Using Multilinear Mixing Model**
Haiping Lu

14:55-15:15 **Embedding with Autoencoder Regularization**
Wenchao Yu, Guangxiang Zeng, Ping Luo, Fuzhen Zhuang, Qing He and Zhongzhi Shi

15:15-15:35 **Learning Exemplar-Represented Manifolds in Latent Space for Classification**
Shu Kong and Donghui Wang

15:35-15:55 **Locally Linear Landmarks for Large-Scale Manifold Learning**
Max Vladymyrov and Miguel Carreira-Perpinan

Tue2C: Biomedical Applications

Room: Ceremonie

14:15-14:35 **Forest-Based Point Process for Event Prediction from Electronic Health Records**
Jeremy Weiss, Michael Caldwell and David Page

14:35-14:55 **On Discovering the Correlated Relationship between Static and Dynamic Data in Clinical Gait Analysis**
Yin Song, Jian Zhang, Longbing Cao and Morgan Sangeux

14:55-15:15 **Computational Drug Repositioning by Ranking and Integrating Multiple Data Sources**
Ping Zhang, Pankaj Agarwal and Zoran Obradovic

15:15-15:35 **Score As You Lift (SAYL): A Statistical Relational Learning Approach to Uplift Modeling**
Houssam Nassif, Finn Kuusisto, Elizabeth Burnside, David Page, Jude Shavlik and Vitor Santos Costa

15:35-15:55 **Protein Function Prediction using Dependence Maximization**
Guoxian Yu, Carlotta Demoniconi, Huzefa Rangwala and Guoji Zhang

Tue2D: Demo spotlights

Room: Aplaus

14:15-14:24 **Image Hub Explorer: Evaluating Representations and Metrics for Content-Based Image Recognition**
Nenad Tomasev and Dunja Mladenic

14:24-14:33 **Ipseity–A Laboratory for Synthesizing and Validating Artificial Cognitive Systems in Multi-agent Systems**
Fabrice Lauri, Nicolas Gaud, Stephane Galland and Vincent Hilaire

14:33-14:42 **OpenML: A Collaborative Science Platform**
Jan Van Rijn, Bernd Bischl, Luis Torgo, Bo Gao, Venkatesh Umaashankar, Simon Fischer, Patrick Winter,Bernd Wiswedel, Michael Berthold and Joaquin Vanschoren

14:42-14:51 **ViperCharts: Visual Performance Evaluation Platform**
Borut Sluban and Nada Lavrac

14:51-15:00 **Targeted Linked Hypernym Discovery: Real-Time Classification of Entities in Text with Wikipedia**
Milan Dojchinovski and Tomas Kliegr

15:00-15:09 **Hermoupolis: A Trajectory Generator for Simulating Generalized Mobility Patterns**
Nikos Pelekis, Christos Ntrigkogias, Panagiotis Tampakis, Stylianos Sideridis and Yannis Theodoridis

15:09-15:18 **AllAboard: A System for Exploring Urban Mobility and Optimizing Public Transport Using Cellphone Data**
Michele Berlingerio, Francesco Calabrese, Giusy Di Lorenzo, Rahul Nair, Fabio Pinelli and Marco Sbodio

15:18-15:27 **ScienScan–Efficient Visualization and Browsing Tool for Academic Search**
Daniil Mirylenka and Andrea Passerini

15:27-15:36 **InVis: A Tool for Interactive Visual Data Analysis**
Daniel Paurat and Thomas Gaertner

15:36-15:45 **Kanopy: Analysing the Semantic Network around Document Topics**
Ioana Hulpus, Conor Hayes, Marcel Karnstedt, Derek Greene and Marek Jozwowicz

15:45-15:54 **SCCQL: A Constraint-Based Clustering System**
Antoine Adam, Hendrik Blockeel, Sander Govers and Abram Aertsen

Tue3A: Inverse RL & RL Applications

Room: Euforie

16:25-16:45 **A Cascaded Supervised Learning Approach to Inverse Reinforcement Learning**
Edouard Klein, Bilal Piot, Matthieu Geist and Olivier Pietquin

16:45-17:05 **Learning From Demonstrations: Is it Worth Estimating a Reward Function?**
Bilal Piot, Matthieu Geist and Olivier Pietquin

17:05-17:25 **Recognition of Agents based on Observation of Their Sequential Behavior**
Qifeng Qiao and Peter Beling

17:25-17:45 **Learning Throttle Valve Control Using Policy Search**
Bastian Bischoff, Duy Nguyen-Tuong, Torsten Koller, Heiner Markert and Alois Knoll

17:45-18:05 **Model-Selection for Non-Parametric Function Approximation in Continuous Control Problems: A Case Study in a Smart Energy System**
Daniel Urieli and Peter Stone

Tue3B: Matrix Analysis

Room: Dialog

16:25-16:45 **Noisy Matrix Completion Using Alternating Minimization**
Suriya Gunasekar, Ayan Acharya, Neeraj Gaur and Joydeep Ghosh

16:45-17:05 **A Nearly Unbiased Matrix Completion Approach**
Dehua Liu, Tengfei Zhou, Hui Qian, Congfu Xu and Zhihua Zhang

17:05-17:25 **A Counterexample for the Validity of Using Nuclear Norm as a Convex Surrogate of Rank**
Hongyang Zhang, Zhouchen Lin and Chao Zhang

17:25-17:45 **Efficient Rank-one Residue Approximation Method for Graph Regularized Non-negative Matrix Factorization**
Qing LIAO and Qian Zhang

17:45-18:05 **Maximum Entropy Models for Iteratively Identifying Subjectively Interesting Structure in Real-Valued Data**
Kleanthis-Nikolaos Kontonasios, Jilles Vreeken and Tijl De Bie

Tue3C: Applications

Room: Ceremonie

16:25-16:45 **Incremental Sensor Placement Optimization on Water Network**
Xiaomin Xu, Yiqi Lu, Sheng Huang, Yanghua Xiao and Wei Wang

16:45-17:05 **Detecting Marionette Microblog Users for Improved Information Credibility**
Xian Wu, Ziming Feng, Wei Fan, Jing Gao and Yong Yu

17:05-17:25 **Will my Question be Answered? Predicting "Question Answerability" in Community Question-Answering Sites**
Gideon Dror, Yoelle Maarek and Idan Szpektor

17:25-17:45 **Learning to Detect Patterns of Crime**
Tong Wang, Cynthia Rudin, Dan Wagner and Rich Sevieri

17:45-18:05 **Space Allocation in the Retail Industry: A Decision Support System Integrating Evolutionary Algorithms and Regression Models**
Fabio Pinto and Carlos Soares

Tue3D: Semi-supervised Learning

Room: Aplaus

16:25-16:45 **Exploratory Learning**
Bhavana Dalvi, William Cohen and Jamie Callan

16:45-17:05 **Semi-supervised Gaussian Process Ordinal Regression**
Srijith P. K., Shirish Shevade and Sundararajan S.

17:05-17:25 **Influence of Graph Construction on Semi-supervised Learning**
Celso Andre De Sousa, Gustavo Batista and Solange Rezende

17:25-17:45 **Tractable Semi-Supervised Learning of Complex Structured Prediction Models**
Kai-Wei Chang, Sundararajan S. and Sathiya Keerthi

17:45-18:05 **PSSDL: Probabilistic Semi-Supervised Dictionary Learning**
Behnam Babagholami-Mohamadabadi, Ali Zarghami, Mohammadreza Zolfaghari and Mahdieh Soleymani Baghshah

TUESDAY SESSIONS, WITH ABSTRACTS

Tue1A: Reinforcement Learning

Room: Euforie

11:00-11:20 **Learning Graph-Based Representations for Continuous Reinforcement Learning Domains**
Jan Hendrik Metzen

Graph-based domain representations have been used in discrete reinforcement learning domains as basis for, e.g., autonomous skill discovery and representation learning. These abilities are also highly relevant for learning in domains which have structured, continuous state spaces as they allow to decompose complex problems into simpler ones and reduce the burden of hand-engineering features. However, since graphs are inherently discrete structures, the extension of these approaches to continuous domains is not straightforward. We argue that graphs should be seen as discrete, generative models of continuous domains. Based on this intuition, we define the likelihood of a graph for a given set of observed state transitions and derive a heuristic method entitled FIGE that allows to learn graph-based representations of continuous domains with large likelihood. Based on FIGE, we present a new skill discovery approach for continuous domains. Furthermore, we show that the learning of representations can be considerably improved by using FIGE.

11:20-11:40 **Greedy Confidence Pursuit: A Pragmatic Approach to Multi-bandit Optimization**
Philip Bachman and Doina Precup

In some reinforcement learning problems an agent may be provided with a set of input policies, perhaps learned from prior experience or provided by advisors. We present a reinforcement learning with policy advice (RLPA) algorithm which leverages this input set and learns to use the best policy in the set for the reinforcement learning task at hand. We prove that RLPA has a sub-linear regret of $\tilde{O}(\sqrt{T})$ relative to the best input policy, and that both this regret and its computational complexity are independent of the size of the state and action space. Our empirical simulations support our theoretical analysis. This suggests RLPA may offer significant advantages in large domains where some prior good policies are provided.

11:40-12:00 **Exploiting Multi-step Sample Trajectories for Approximate Value Iteration**
Robert Wright, Lei Yu, Steven Loscalzo and Philip Dexter

Approximate value iteration methods for reinforcement learning (RL) generalize experience from limited samples across large state-action spaces. The function approximators used in such methods typically introduce errors in value estimation which can harm the quality of the learned value functions. We present a new batch-mode, off-policy, approximate value iteration algorithm called Trajectory Fitted Q-Iteration (TFQI). This approach uses the sequential relationship between samples within a trajectory, a set of samples gathered sequentially from the problem domain, to lessen the adverse influence of approximation errors while deriving long-term value. We provide a detailed description of the TFQI approach and an empirical study that analyzes the impact of our method on two well-known RL benchmarks. Our experiments demonstrate this approach has significant benefits including: better learned policy performance, improved convergence, and some decreased sensitivity to the choice of function approximation.

12:00-12:20 **Automatically Mapped Transfer Between Reinforcement Learning Tasks via Three-Way Restricted Boltzmann Machines**
Haitham Bou Ammar, Decebal Constantin Mocanu, Matthew Taylor, Kurt Driessens, Gerhard Weiss and Karl Tuyls

Existing reinforcement learning approaches are often hampered by learning tabula rasa. Transfer for reinforcement learning tackles this problem by enabling the reuse of previously learned results, but may require an inter-task mapping to encode how the previously learned task and the new task are related. This paper presents an autonomous framework for learning inter-task mappings based on an adaptation of Restricted Boltzmann Machines. Both a full model and a computationally eficient factored model are introduced and shown to be effective in multiple transfer learning scenarios.

Tue1B: Networks (1)

Room: Dialog

11:00-11:25 **What Distinguish One from Its Peers in Social Networks?**
Yi-Chen Lo, Jhao-Yin Li, Mi-Yen Yeh, Shou-De Lin and Jian Pei

Being able to discover the uniqueness of an individual is a meaningful task in social network analysis. This paper proposes two novel problems in social network analysis: how to identify the uniqueness of a given query vertex, and how to identify a group of vertices that can mutually identify each other. We further propose three intuitive yet effective methods to identify the uniqueness identification sets and the mutual identification groups of different properties. We further conduct an extensive experiment on both real and synthetic datasets to demonstrate the effectiveness of our model.

11:25-11:45 **Detecting Bicliques in GF[q]**
Jan Ramon, Pauli Miettinen and Jilles Vreeken

We consider the problem of finding planted bicliques in random matrices over GF[q]. More in particular, we study several models of random bipartite graphs and their adjacency matrices, and show quasi-polynomial time approximations for finding planted bicliques, i.e. matrices that are a sum of a , while it is NP-hard to find the largest such clique. Real graphs, however, are typically extremely sparand seldom contain such large bicliques. We investigate the practical problem of how small a biclique can be such that we can still approximately correctly identify it in quasi-polynomial time. Our derivations show that with high probability planted bicliques of size logarithmic in the network size can be detected in data following the Erdos-Renyi model and two bipartite variants of the Barabasi-Albert model.

11:45-12:05 **As Strong as the Weakest Link: Mining Diverse Cliques in Weighted Graphs**
Petko Bogdanov, Ben Baumer, Prithwish Basu, Amotz Bar-Noy and Ambuj Singh

Mining for cliques in networks provides an essential tool for the discovery of strong associations among entities. Applications vary, from extracting core subgroups in team performance data arising in sports, entertainment, research and business: to the discovery of functional complexes in high-throughput gene interaction data. A challenge in all of these scenarios is the large size of real-world networks and the computational complexity associated with clique enumeration. Furthermore, when mining for multiple cliques within the same network, the results need to be diversified in order to extract meaningful information that is both comprehensive and representative of the whole dataset. We formalize the problem of weighted diverse clique mining (mDkC) in large networks, incorporating both individual clique strength (measured by its weakest link) and diversity of the cliques in the result set. We show that the problem is NP-hard due to the diversity requirement. However, our formulation is sub-modular, and hence can be approximated within a constant factor from the optimal. We propose algorithms for mDkC that exploit the edge weight distribution in the input network and produce performance gains of more than 3 orders of magnitude compared to an exhaustive solution. One of our algorithms, Diverse Cliques (DiCliQ), guarantees a constant factor approximation while the other, Bottom Up Diverse Cliques (BUDiC), scales to large and dense networks without compromising the solution quality. We evaluate both algorithms on 5 real-world networks of different genres and demonstrate their utility for discovery of gene complexes and effective collaboration subgroups in sports and entertainment.

12:05-12:25 **How Robust is the Core of a Network?**
Abhijin Adiga and Anil Vullikanti

The k-core is commonly used as a measure of importance and well connectedness for nodes in diverse applications in social networks and bioinformatics. Since network data is commonly noisy and incomplete, a fundamental issue is to understand how robust the core decomposition is to noise. Further, in many settings, such as online social media networks, usually only a sample of the network is available. Therefore, a related question is: how robust is the top core set under such sampling? We find that, in general, the top core is quite sensitive to both noise and sampling; we quantify this in terms of the Jaccard similarity of the set of top core nodes between the original and perturbed/sampled graphs. Most importantly, we find that the overlap with the top core set varies non-monotonically with the extent of perturbations/sampling. We explain some of these empirical observations by rigorous analysis in simple network models. Our work has important implications for the use of the core decomposition and nodes in the top cores in network analysis applications, and suggests the need for a more careful characterization of the missing data and sensitivity to it.

Tue1C: Privacy and Security

Room: Ceremonie

11:00-11:25 **Differential Privacy Based on Importance Weighting**
Zhanglong Ji and Charles Elkan

This paper analyzes a novel method for publishing data while still protecting privacy. The method is based on computing weights that make an existing dataset, for which there are no confidentiality issues, analogous to the dataset that must be kept private. The existing dataset may be genuine but public already, or it may be synthetic. The only necessary requirement to use the method is that there be overlap between features in the two datasets. The weights are importance sampling weights, but to protect privacy, they are regularized and have noise added. The weights allow statistical queries to be answered approximately while provably guaranteeing differential privacy. We derive an expression for the asymptotic variance of the approximate answers. Experiments show that the new mechanism performs well even when the privacy budget is small, and when the public dataset is quite different from the private dataset.

11:25-11:45 **Anonymizing Data with Relational and Transaction Attributes**
Giorgos Poulis, Grigorios Loukides, Aris Gkoulalas-Divanis and Spiros Skiadopoulos

Publishing datasets about individuals that contain both relational and transaction (i.e., set-valued) attributes is essential to support many applications, ranging from healthcare to marketing. However, preserving the privacy and utility of these datasets is challenging, as it requires (i) guarding against attackers, whose knowledge spans both attribute types, and (ii) minimizing the overall information loss. Existing anonymization techniques are not applicable to such datasets, and the problem cannot be tackled based on popular, multi-objective optimization strategies. This work proposes the first approach to address this problem. Based on this approach, we develop two frameworks to offer privacy, with bounded information loss in one attribute type and minimal information loss in the other. To realize each framework, we propose privacy algorithms that effectively preserve data utility, as verified by extensive experiments.

11:45-12:05 **Privacy-Preserving Mobility Monitoring using Sketches of Stationary Sensor Readings**
Michael Kamp, Christine Kopp, Michael Mock, Mario Boley and Michael May

Two fundamental tasks of mobility modeling are (1) to track the number of distinct persons that are present at a location of interest and (2) to reconstruct flows of persons between two or more different locations. Stationary sensors, such as Bluetooth scanners, have been applied to both tasks with remarkable success. However, this approach has privacy problems. For instance, Bluetooth scanners store the MAC address of a device that can in principle be linked to a single person. Unique hashing of the address only partially solves the problem because such a pseudonym is still vulnerable to various linking attacks. In this paper we propose a solution to both tasks using an extension of linear counting sketches. The idea is to map several individuals to the same position in a sketch, while at the same time the inaccuracies introduced by this overloading are compensated by using several independent sketches. This idea provides, for the first time, a general set of primitives for privacy preserving mobility modeling from Bluetooth and similar address-based devices.

12:05-12:25 **Evasion Attacks Against Machine Learning at Test Time**
Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Srndic, Pavel Laskov, Giorgio Giacinto and Fabio Roli

In security-sensitive applications, the success of machine learning depends on a thorough vetting of their resistance to adversarial data. In one pertinent, well-motivated attack scenario, an adversary may attempt to evade a deployed system at test time by carefully manipulating her attack samples. In this work, we present a simple but effective gradient-based approach that can be exploited to systematically assess the security of several, widely-used classification algorithms against evasion attacks. Following a recently proposed framework for security evaluation, we simulate attack scenarios that exhibit different risk levels for the classifier by increasing the attacker's knowledge of the system and her ability to manipulate attack samples. This gives the classifier designer a better picture of the classifier performance under evasion attacks, and allows him to perform a more informed model selection (or parameter setting). We evaluate our approach on the relevant security task of malware detection in PDF files, and show that such systems can be easily evaded. We also sketch some countermeasures suggested by our analysis.

Tue1D: Ranking and Recommender Systems

Room: Aplaus

11:00-11:25 **Growing a List**
Benjamin Letham, Cynthia Rudin and Katherine A Heller

It is easy to find expert knowledge on the Internet on almost any topic, but obtaining a complete overview of a given topic is not always easy: Information can be scattered across many sources and must be aggregated to be useful. We introduce a method for intelligently growing a list of relevant items, starting from a small seed of examples. Our algorithm takes advantage of the wisdom of the crowd, in the sense that there are many experts who post lists of things on the Internet. We use a collection of simple machine learning components to find these experts and aggregate their lists to produce a single complete and meaningful list. We use experiments with gold standards and open-ended experiments without gold standards to show that our method significantly outperforms the state of the art. Our method uses the ranking algorithm Bayesian Sets even when its underlying independence assumption is violated, and we provide a theoretical generalization bound to motivate its use.

11:25-11:45 **A Pairwise Label Ranking Method with Imprecise Scores and Partial Predictions**
Sebastien Destercke

In this paper, we are interested in the label ranking problem. We are more specifically interested in the recent trend consisting in predicting partial but robust rankings rather than complete ones. To do so, we propose a ranking method based on pairwise imprecise scores obtained from likelihood functions. We discuss how such imprecise scores can be aggregated to produce interval orders, which are specific types of partial orders. We then analyse the performances of the method as well as its sensitivity to missing data and parameters values.

11:45-12:05 **Learning Socially Optimal Information Systems from Egoistic Users**
Karthik Raman and Thorsten Joachims

Many information systems aim to present results that maximize the collective satisfaction of the user population. The product search of an online store, for example, needs to present an appropriately diverse set of products to best satisfy the different tastes and needs of its user population. To address this problem, we propose two algorithms that can exploit observable user actions (e.g. clicks) to learn how to compose diverse sets (and rankings) that optimize expected utility over a distribution of utility functions. A key challenge is that individual users evaluate and act according to their own utility function, but that the system aims to optimize collective satisfaction. We characterize the behavior of our algorithms by providing upper bounds on the social regret for a class of submodular utility functions in the coactive learning model. Furthermore, we empirically demonstrate the efficacy and robustness of the proposed algorithms for the problem of search result diversification.

12:05-12:25 **Socially Enabled Preference Learning from Implicit feedback data**
Julien Delporte, Alexandros Karatzoglou, Tomasz Matuszczyk and Stephane Canu

In the age of information overload, collaborative filtering and recommender systems have become essential tools for content discovery. The advent of online social networks has added another approach to recommendation whereby the social network itself is used as a source for recommendations i.e. users are recommended items that are preferred by their friends. In this paper we develop a new model-based recommendation method that merges collaborative and social approaches by utilizing implicit feedback and social graph data. Employing factor models, we represent each user profile as a mixture of his own and his friends' profiles. This assumes and exploits ``homophily'' in the social network, a phenomenon that has been studied in the social sciences. The interaction between the preferences of a user and his friends is directly modeled as an egde-weighted graph that can be seen as a measure of trust (or influence) between a given user and his friends. In this model both the factors and the influence weights are learned. We extensively test our model on the Epinions data and on the Tuenti Places Recommendation data, a large-scale industry dataset, where it outperforms several state-of-the-art methods.

Tue2A: Markov Decision Processes

Room: Euforie

14:15-14:35 **Expectation Maximization for Average Reward Decentralized POMDPs**
Joni Pajarinen and Jaakko Peltonen

Planning for multiple agents under uncertainty is often based on decentralized partially observable Markov decision processes (Dec-POMDPs), but current methods must de-emphasize long-term effects of actions by a discount factor. In tasks like wireless networking, agents are evaluated by average performance over time, both short and long-term effects of actions are crucial, and discounting based solutions can perform poorly. We show that under a common set of conditions expectation maximization (EM) for average reward Dec-POMDPs is stuck in a local optimum. We introduce a new average reward EM method: it outperforms a state of the art discounted-reward Dec-POMDP method in experiments.

14:35-14:55 **Properly Acting under Partial Observability with Action Feasibility Constraints**
Caroline Carvalho Chanel and Florent Teichteil-Konigsbuch

We propose a sequential decision-making model named Action-Constrained Partially Observable Markov Decision Process (AC-POMDP), which arose from studying critical robotic applications with damaging actions. AC-POMDPs restrict the optimized policy to apply only feasible actions at execution: each action is feasible in a subset of the state space, and the agent can observe the set of applicable actions in the current hidden state, in addition to standard observations. We present optimality equations for AC-POMDPs, which imply to operate on alpha-vectors that are defined over many different small belief subspaces. Finally, we present an algorithm named PreCondition Value Iteration (PCVI), which fully exploits the specific structure of AC-POMDPs by implementing alpha-vector operations on different belief subspaces. We also designed a relaxed version of PCVI whose complexity is exponentially smaller than PCVI. Experimental results on POMDP benchmarks with action feasibility constraints and on real robotic problems, exhibit the benefits of explicitly exploiting the semantic richness of action-feasibility observations in AC-POMDPs over equivalent but unstructured POMDPs.

14:55-15:15 **Iterative Model Refinement of Recommender MDPs Based on Expert Feedback**
Omar Khan, Pascal Poupart and John Mark Agosta

In this paper, we present a method to iteratively refine the parameters of a Markov Decision Process by leveraging constraints implied from an expert's review of the policy. We impose a constraint on the parameters of the model for every case where the expert's recommendation differs from the recommendation of the policy. We demonstrate that consistency with an expert's feedback leads to non-convex constraints on the model parameters. We refine the parameters of the model, under these constraints, by partitioning the parameter space and iteratively applying alternating optimization. We demonstrate how the approach can be applied to both flat and factored MDPs and present results based on diagnostic sessions from a manufacturing scenario.

15:15-15:35 **Solving Relational MDPs with Exogenous Events and Additive Rewards**
Saket Joshi, Roni Khardon, Prasad Tadepalli, Aswin Raghavan and Alan Fern

We formalize a simple but natural subclass of service domains for relational planning problems with object-centered, independent exogenous events and additive rewards capturing, for example, problems in inventory control. Focusing on this subclass, we present a new symbolic planning algorithm which is the first algorithm that has explicit performance guarantees for relational MDPs with exogenous events. In particular, our planning algorithm provides a monotonic lower bound approximation on the optimal value function. To support this algorithm we present novel evaluation and reduction techniques for generalized first order decision diagrams, a knowledge representation for real-valued functions over relational world states. Our planning algorithm uses a set of focus states, which serves as a training set, to simplify and approximate the symbolic solution, and can thus be seen to perform learning for planning. A preliminary experimental evaluation demonstrates the validity of our approach.

15:35-15:55 **Continuous Upper Confidence Trees with Polynomial Exploration–Consistency**
Adrien Couetoux, David Auger and Olivier Teytaud

Upper Confidence Trees (UCT) are now a well known algorithm for sequential decision making; it is a provably consistent variant of Monte-Carlo Tree Search. However, the consistency is only proved in a the case where the action space is finite. We here propose a proof in the case of fully observable Markov Decision Processes with bounded horizon, possibly including infinitely many states, infinite action space and arbitrary stochastic transition kernels. We illustrate the consistency on two benchmark problems, one being a legacy toy problem, the other a more challenging one, the famous energy unit commitment problem.

Tue2B: Tensor Analysis & Dimensionality Reduction

Room: Dialog

14:15-14:35 **An Analysis of Tensor Models for Learning on Structured Data**
Maximilian Nickel and Volker Tresp

While tensor factorizations have become increasingly popular for learning on various forms of structured data, only very few theoretical results exist on the generalization abilities of these methods. Here, we discuss the tensor product as a principled way to represent structured data in vector spaces for machine learning tasks. By extending known bounds for matrix factorizations, we are able to derive generalization error bounds for the tensor case. Furthermore, we analyze analytically and experimentally how tensor factorization behaves when applied to over- and understructured representations, for instance, when two-way tensor factorization, i.e. matrix factorization, is applied to three-way tensor data.

14:35-14:55 **Learning Modewise Independent Components from Tensor Data Using Multilinear Mixing Model**
Haiping Lu

Independent component analysis (ICA) is a popular unsupervised learning method. This paper extends it to multilinear modewise ICA (MMICA) for tensors and explores two architectures in learning and recognition. MMICA models tensor data as mixtures generated from modewise source matrices that encode statistically independent information. Its sources have more compact representations than the sources in ICA. We embed ICA into the multilinear principal component analysis framework to solve for each source matrix alternatively with a few iterations. Then we obtain mixing tensors through regularized inverses of the source matrices. Simulations on synthetic data show that MMICA can estimate hidden sources accurately from structured tensor data. Moreover, in face recognition experiments, it outperforms competing solutions with both architectures.

14:55-15:15 **Embedding with Autoencoder Regularization**
Wenchao Yu, Guangxiang Zeng, Ping Luo, Fuzhen Zhuang, Qing He and Zhongzhi Shi

The problem of embedding arises in many machine learning applications with the assumption that there may exist a small number of variabilities which can guarantee the ‚semantics“ of the original high-dimensional data. Most of the existing embedding algorithms perform to maintain the locality-preserving property. In this study, inspired by the remarkable success of representation learning and deep learning, we propose a framework of embedding with autoencoder regularization (EAER for short), which incorporates embedding and autoencoding techniques naturally. In this framework, the original data are embedded into the lower dimension, represented by the output of the hidden layer of the autoencoder, thus the resulting data can not only maintain the locality-preserving property but also easily revert to their original forms. This is guaranteed by the joint minimization of the embedding loss and the autoencoder reconstruction error. It is worth mentioning that instead of operating in a batch mode as most of the previous embedding algorithms conduct, the proposed framework actually generates an inductive embedding model and thus supports incremental embedding efficiently. To show the effectiveness of EAER, we adapt this joint learning framework to three canonical embedding algorithms, and apply them to both synthetic and real-world data sets. The experimental results show that the adaption of EAER outperforms its original counterpart. Besides, compared with the existing incremental embedding algorithms, the results demonstrate that EAER performs incremental embedding with more competitive efficiency and effectiveness.

15:15-15:35 **Learning Exemplar-Represented Manifolds in Latent Space for Classification**
Shu Kong and Donghui Wang

Intrinsic manifold structure of a data collection is valuable information for classification task. By considering the manifold structure in the data set for classification and with the sparse coding framework, we propose an algorithm to: (1) find exemplars from each class to represent the class-specific manifold structure, in which way the object-space dimensionality is reduced; (2) simultaneously learn a latent feature space to make the mapped data more discriminative according to the class-specific manifold measurement. We call the proposed algorithm Exemplar-represented Manifold in Latent Space for Classification (EMLSC). We also present the nonlinear extension of EMLSC based on kernel tricks to deal with highly nonlinear situations. Experiments on synthetic and real-world datasets demonstrate the merit of the proposed method.

15:35-15:55 **Locally Linear Landmarks for Large-Scale Manifold Learning**
Max Vladymyrov and Miguel Carreira-Perpinan

Spectral methods for manifold learning and clustering typically construct a graph weighted with affinities from a dataset and compute eigenvectors of a graph Laplacian. With large datasets, the eigendecomposition is too expensive, and is usually approximated by solving for a smaller graph defined on a subset of the points (landmarks) and then applying the Nystrom formula to estimate the eigenvectors over all points. This has the problem that the affinities between landmarks do not benefit from the remaining points and may poorly represent the data if using few landmarks. We introduce a modified spectral problem that uses all data points by constraining the latent projection of each point to be a local linear function of the landmarks' latent projections. This constructs a new affinity matrix between landmarks that preserves manifold structure even with few landmarks, allows one to reduce the eigenproblem size, and defines a fast, nonlinear out-of-sample mapping.

Tue2C: Biomedical Applications

Room: Ceremonie

14:15-14:35 **Forest-Based Point Process for Event Prediction from Electronic Health Records**
Jeremy Weiss, Michael Caldwell and David Page

Accurate prediction of future onset of disease from Electronic Health Records (EHRs) has important clinical and economic implications. In this domain the arrival of data comes at semi-irregular intervals and makes the prediction task challenging. We elect to model events with a proposed method called multiplicative-forest point processes (MFPPs) that learns the rate of future events based on an event history. MFPPs join previous theory in multiplicative forest continuous-time Bayesian networks and piecewise-continuous conditional intensity models. We analyze the advantages of using MFPPs over previous methods and show that on synthetic and real EHR forecasting of heart attacks, MFPPs outperform earlier methods and augment off-the-shelf machine learning algorithms.

14:35-14:55 **On Discovering the Correlated Relationship between Static and Dynamic Data in Clinical Gait Analysis**
Yin Song, Jian Zhang, Longbing Cao and Morgan Sangeux

'Gait' is a person's manner of walking. Patients may have an abnormal gait due to a range of physical impairment or brain damage. Clinical gait analysis (CGA) is a technique for identifying the underlying impairments that affect a patient's gait pattern. The CGA is critical for treatment planning. Essentially, CGA tries to use patients' physical examination results, known as static data, to interpret the dynamic characteristics in an abnormal gait, known as dynamic data. This process is carried out by gait analysis experts, mainly based on their experience which may lead to subjective diagnoses. To facilitate the automation of this process and form a relatively objective diagnosis, this paper proposes a new probabilistic correlated static-dynamic model (CSDM) to discover correlated relationships between the dynamic characteristics of gait and their root cause in the static data space. We propose an EM-based algorithm to learn the parameters of the CSDM. One of the main advantages of the CSDM is its ability to provide intuitive knowledge. For example, the CSDM can describe what kinds of static data will lead to what kinds of hidden gait patterns in the form of a decision tree, which helps us to infer dynamic characteristics based on static data. Our initial experiments indicate that the CSDM is promising for discovering the correlated relationship between physical examination (static) and gait (dynamic) data.

14:55-15:15 **Computational Drug Repositioning by Ranking and Integrating Multiple Data Sources**
Ping Zhang, Pankaj Agarwal and Zoran Obradovic

Drug repositioning helps identify new indications for marketed drugs and clinical candidates. In this study, we proposed an integrative computational framework to predict novel drug indications for both approved drugs and clinical molecules by integrating chemical, biological and phenotypic data sources. We defined different similarity measures for each of these data sources and utilized a weighted k-nearest neighbor algorithm to transfer similarities of nearest neighbors to prediction scores for a given compound. A large margin method was used to combine individual metrics from multiple sources into a global metric. A large-scale study was conducted to repurpose 1007 drugs against 719 diseases. Experimental results showed that the proposed algorithm outperformed similar previously developed computational drug repositioning approaches. Moreover, the new algorithm also ranked drug information sources based on their contributions to the prediction, thus paving the way for prioritizing multiple data sources and building more reliable drug repositioning models.

15:15-15:35 **Score As You Lift (SAYL): A Statistical Relational Learning Approach to Uplift Modeling**
Houssam Nassif, Finn Kuusisto, Elizabeth Burnside, David Page, Jude Shavlik and Vitor Santos Costa

We introduce Score As You Lift (SAYL), a novel Statistical Relational Learning (SRL) algorithm, and apply it to an important task in the diagnosis of breast cancer. SAYL combines SRL with the marketing concept of uplift modeling, uses the area under the uplift curve to direct clause construction and final theory evaluation, integrates rule learning and probability assignment, and conditions the addition of each new theory rule to existing ones. Breast cancer, the most common type of cancer among women, is categorized into two subtypes: an earlier in situ stage where cancer cells are still confined, and a subsequent invasive stage. Currently older women with in situ cancer are treated to prevent cancer progression, regardless of the fact that treatment may generate undesirable side-effects, and the woman may die of other causes. Younger women tend to have more aggressive cancers, while older women tend to have more indolent tumors. Therefore older women whose in situ tumors show significant dissimilarity with in situ cancer in younger women are less likely to progress, and can thus be considered for watchful waiting. Motivated by this important problem, this work makes two main contributions. First, we present the first multi-relational uplift modeling system, and introduce, implement and evaluate a novel method to guide search in an SRL framework. Second, we compare our algorithm to previous approaches, and demonstrate that the system can indeed obtain differential rules of interest to an expert on real data, while significantly improving the data uplift.

15:35-15:55 **Protein Function Prediction using Dependence Maximization**
Guoxian Yu, Carlotta Demoniconi, Huzefa Rangwala and Guoji Zhang

Protein function prediction is one of the fundamental tasks in the post genomic era. The vast amount of available proteomic data makes it possible to computationally annotate proteins. Most computational approaches predict protein functions by using the labeled proteins and assuming that the annotation of labeled proteins is complete, and without any missing functions. However, partially annotated proteins are common in real-world scenarios, that is a protein may have some confirmed functions, and whether it has other functions is unknown. In this paper, we make use of partially annotated proteomic data, and propose an approach called Protein Function Prediction using Dependence Maximization (ProDM). ProDM works by leveraging the correlation between different function labels, the ‚guilt by association' rule between proteins, and maximizes the dependency between function labels and feature expression of proteins. ProDM can replenish the missing functions of partially annotated proteins (a seldom studied problem), and can predict functions for completely unlabeled proteins using the partially annotated ones. An empirical study on publicly available protein-protein interaction (PPI) networks shows that, when the number of missing functions is large, ProDM performs significantly better than other related methods with respect to various evaluation criteria.

Tue2D: Demo spotlights

Room: Aplaus

14:15-14:24 **Image Hub Explorer: Evaluating Representations and Metrics for Content-Based Image Retrieval and Object Recognition**
Nenad Tomasev and Dunja Mladenic

Large quantities of image data are generated daily and visualizing large image datasets is an important task. We present a novel tool for image data visualization and analysis, Image Hub Explorer. The integrated analytic functionality is centered around dealing with the recently described phenomenon of hubness and evaluating its impact on the image retrieval, recognition and recommendation process. Hubness is reflected in that some images (hubs) end up being very frequently retrieved in 'top k' result sets, regardless of their labels and target semantics. Image Hub Explorer offers many methods that help in visualizing the influence of major image hubs, as well as state-of-the-art metric learning and hubness-aware classification methods that help in reducing the overall impact of extremely frequent neighbor points. The system also helps in visualizing both beneficial and detrimental visual words in individual images. Search functionality is supported, along with the recently developed hubness-aware result set re-ranking procedure.

14:24-14:33 **Ipseity–A Laboratory for Synthesizing and Validating Artificial Cognitive Systems in Multi-agent Systems**
Fabrice Lauri, Nicolas Gaud, Stephane Galland and Vincent Hilaire

This article presents an overview on Ipseity, an open-source rich-client platform developed in C++ with the Qt framework. Ipseity facilitates the synthesis of artificial cognitive systems in multi-agent systems. The current version of the platform includes a set of plugins based on the classical reinforcement learning techniques like Q-Learning and Sarsa. Ipseity is targeted at a broad range of users interested in artificial intelligence in general, including industrial practitioners, as well as machine learning researchers, students and teachers. It is daily used as a course support in Artificial Intelligence and Reinforcement Learning and it has been used successfully to manage power flows in simulated microgrids using multi-agent reinforcement learning.

14:33-14:42 **OpenML: A Collaborative Science Platform**
Jan Van Rijn, Bernd Bischl, Luis Torgo, Bo Gao, Venkatesh Umaashankar, Simon Fischer, Patrick Winter,Bernd Wiswedel, Michael Berthold and Joaquin Vanschoren

We present OpenML, a novel open science platform that provides easy access to machine learning data, software and results to encourage further study and application. It organizes all submitted results online so they can be easily found and reused, and features a web API which is being integrated in popular machine learning tools such as Weka, KNIME, RapidMiner and R packages, so that experiments can be shared easily.

14:42-14:51 **ViperCharts: Visual Performance Evaluation Platform**
Borut Sluban and Nada Lavrac

The paper presents the ViperCharts web-based platform for visual performance evaluation of classification, prediction, and information retrieval algorithms. The platform enables to create interactive charts for easy and intuitive evaluation of performance results. It includes standard visualizations and extends them by offering alternative evaluation methods like F-isolines, and by establishing relations between corresponding presentations like Precision-Recall and ROC curves. Additionally, the interactive performance charts can be saved, exported to several formats, and shared via unique web addresses. A web API to the service is also available.

14:51-15:00 **Targeted Linked Hypernym Discovery: Real-Time Classification of Entities in Text with Wikipedia**
Milan Dojchinovski and Tomas Kliegr

Targeted Hypernym Discovery (THD) performs unsupervised classification of entities appearing in text. A hypernym mined from the free-text of the Wikipedia article describing the entity is used as a class. The type as well as the entity are cross-linked with their representation in DBpedia, and enriched with additional types from DBpedia and YAGO knowledge bases providing a semantic web interoperability. The system, available as a web application and web service at entityclassifier.eu, currently supports English, German and Dutch.

15:00-15:09 **Hermoupolis: A Trajectory Generator for Simulating Generalized Mobility Patterns**
Nikos Pelekis, Christos Ntrigkogias, Panagiotis Tampakis, Stylianos Sideridis and Yannis Theodoridis

During the last decade, the domain of mobility data mining has emerged providing many effective methods for the discovery of intuitive patterns representing collective behavior of trajectories of moving objects. Although a few real-world trajectory datasets have been made available recently, these are not sufficient for experimentally evaluating the various proposals, therefore, researchers look to synthetic trajectory generators. This case is problematic because, on the one hand, real datasets are usually small, which compromises scalability experiments, and, on the other hand, synthetic dataset generators have not been designed to produce mobility pattern driven trajectories. Motivated by this observation, we present Hermoupolis, an effective generator of synthetic trajectories of moving objects that has the main objective that the resulting datasets support various types of mobility patterns (clusters, flocks, convoys, etc.), as such producing datasets with available ground truth information.

15:09-15:18 **AllAboard: A System for Exploring Urban Mobility and Optimizing Public Transport Using Cellphone Data**
Michele Berlingerio, Francesco Calabrese, Giusy Di Lorenzo, Rahul Nair, Fabio Pinelli and Marco Sbodio

The deep penetration of mobile phones offers cities the ability to opportunistically monitor citizens' interactions and use data-driven insights to better plan and manage services. In this context, transit operators can leverage pervasive mobile sensing to better match observed demand for travel with their service offerings. With large scale data on mobility patterns, operators can move away from the costly and resource intensive four-step transportation planning processes prevalent in the West, to a more data-centric view, that places the instrumented user at the center of development. In this framework, using mobile phone data to perform transit analysis and optimization represents a new frontier with significant societal impact, especially in developing countries. In this demo, we present AllAboard, a system for optimizing public transport using cellphone data. Our system uses mobile phone location data to infer origin-destination flows in the city, which is then converted to ridership on the existing transit network. Sequential travel patterns extracted from individual call location data are used to propose new candidate transit routes. An optimization model evaluates which new routes would best improve the existing transit network to increase ridership and user satisfaction, both in terms of reduced travel and wait time. The system provides also a User Interface that allows the interaction with results and the data themselves. The system in its whole is intended to be used by city authorities for improving their public transport systems, using cell phone data, which have a large penetration even in developing countries, and provide a cheaper, faster, alternative to costly surveys.

15:18-15:27 **ScienScan–Efficient Visualization and Browsing Tool for Academic Search**
Daniil Mirylenka and Andrea Passerini

In this paper we present ScienScan-a browsing and visualization tool for academic search. The tool operates in real time by post-processing the query results returned by an academic search engine. ScienScan discovers topics in the search results and summarizes them in the form of a concise hierarchical topic map. The produced topical summary informatively represents the results in a visual way and provides an additional filtering control. We demonstrate the operation of ScienScan deploying it on top of the search API of Microsoft Academic Search.

15:27-15:36 **InVis: A Tool for Interactive Visual Data Analysis**
Daniel Paurat and Thomas Gaertner

We present InVis, a tool to visually analyse data by interactively shaping a two dimensional embedding of it. Traditionally, embedding techniques focus on finding one fixed embedding, which emphasizes a single aspects of the data. In contrast, our application enables the user to explore the structures of a dataset by observing and controlling a projection of it. Ultimately it provides a way to search and find an embedding, emphasizing aspects that the user desires to be highlighted.

15:36-15:45 **Kanopy: Analysing the Semantic Network around Document Topics**
Ioana Hulpus, Conor Hayes, Marcel Karnstedt, Derek Greene and Marek Jozwowicz

External knowledge bases, both generic and domain specific, available on the Web of Data have the potential of enriching the content of text documents with structured information. We present the Kanopy system that makes explicit use of this potential. Besides the common task of semantic annotation of documents, Kanopy analyses the semantic network that resides in DBpedia around extracted concepts. The system's main novelty lies in the translation of social network analysis measures to semantic networks in order to find suitable topic labels. Moreover, Kanopy extracts advanced knolwedge in the form of subgraphs that capture the relationships between the concepts.

15:45-15:54 **SCCQL: A Constraint-Based Clustering System**
Antoine Adam, Hendrik Blockeel, Sander Govers and Abram Aertsen

This paper presents the first version of a new inductive data-base system called SCCQL. The system performs constraint-based clustering on a relational database. Clustering problems are formulated with a query language, an extension of SQL for clustering that includes must-link and cannot-link constraints. The functioning of the system is explained. As an example of use of this system, an application in the context of microbiology has been developed that is presented here.

Tue3A: Inverse RL & RL Applications

Room: Euforie

16:25-16:45 **A Cascaded Supervised Learning Approach to Inverse Reinforcement Learning**
Edouard Klein, Bilal Piot, Matthieu Geist and Olivier Pietquin

This paper considers the Inverse Reinforcement Learning (IRL) problem, that is inferring a reward function for which a demonstrated expert policy is optimal. We propose to break the IRL problem down into two generic Supervised Learning steps: this is the Cascaded Supervised IRL (CSI) approach. A classification step that defines a score function is followed by a regression step providing a reward function. A theoretical analysis shows that the demonstrated expert policy is near-optimal for the computed reward function. Not needing to repeatedly solve a Markov Decision Process (MDP) and the ability to leverage existing techniques for classification and regression are two important advantages of the CSI approach. It is furthermore empirically demonstrated to compare positively to state-of-the-art approaches when using only transitions sampled according to the expert policy, up to the use of some heuristics. This is exemplified on two classical benchmarks (the mountain car problem and a highway driving simulator).

16:45-17:05 **Learning From Demonstrations: Is it Worth Estimating a Reward Function?**
Bilal Piot, Matthieu Geist and Olivier Pietquin

This paper provides a comparative study between Inverse Reinforcement Learning (IRL) and Apprenticeship Learning (AL). IRL and AL are two frameworks, using Markov Decision Processes (MDP), which are used for the imitation learning problem where an agent tries to learn from demonstrations of an expert. In the AL framework, the agent tries to learn the expert policy whereas in the IRL framework, the agent tries to learn a reward which can explain the behavior of the expert. This reward is then optimized to imitate the expert. One can wonder if it is worth estimating such a reward, or if estimating a policy is sufficient. This quite natural question has not really been addressed in the literature right now. We provide partial answers, both from a theoretical and empirical point of view.

17:05-17:25 **Recognition of Agents based on Observation of Their Sequential Behavior**
Qifeng Qiao and Peter Beling

We study the use of inverse reinforcement learning (IRL) as a tool for recognition of agents on the basis of observation of their sequential decision behavior. We model the problem faced by the agents as a Markov decision process (MDP) and model the observed behavior of an agent in terms of forward planning for the MDP. The reality of the agent's decision problem and process may not be expressed by the MDP and its policy, but we interpret the observation as optimal actions in the MDP. We use IRL to learn reward functions for the MDP and then use these reward functions as the basis for clustering or classification models. Experimental studies with GridWorld, a navigation problem, and the secretary problem, an optimal stopping problem, show algorithms' performance in different learning scenarios for agent recognition where the agents' underlying decision strategy may be expressed by the MDP policy or not. Empirical comparisons of our method with several existing IRL algorithms and with direct methods that use feature statistics observed in state-action space suggest it may be superior for agent recognition problems, particularly when the state space is large but the length of the observed decision trajectory is small.

17:25-17:45 **Learning Throttle Valve Control Using Policy Search**
Bastian Bischoff, Duy Nguyen-Tuong, Torsten Koller, Heiner Markert and Alois Knoll

The throttle valve is a technical device used for regulating a fluid or a gas flow. Throttle valve control is a challenging task, due to its complex dynamics and demanding constraints for the controller. Using state-of-the-art throttle valve control, such as model-free PID controllers, time-consuming and manual adjusting of the controller is necessary. In this paper, we investigate how reinforcement learning (RL) can help to alleviate the effort of manual controller design, by automatically learning a control policy from experiences. In order to obtain a valid control policy for the throttle valve, several constraints need to be addressed, such as no-overshoot. Furthermore, the learned controller must be able to follow given desired trajectories, while moving the valve from any start to any goal position and, thus, multi-targets policy learning needs to be considered for RL. In this study, we employ a policy-search RL approach, e.g. Pilco, to learn a throttle valve control policy. We adapt the Pilco algorithm, while taking in account the practical requirements and constraints on the controller. For evaluation, we employ the resulting algorithm to solve several control tasks in simulation, as well as on a physical throttle valve system. The results show that policy-search RL is able to learn a consistent control policy for complex, real-world systems.

17:45-18:05 **Model-Selection for Non-Parametric Function Approximation in Continuous Control Problems: A Case Study in a Smart Energy System**
Daniel Urieli and Peter Stone

This paper investigates the application of value-function-based reinforcement learning to a smart energy control system, specifically the task of controlling an HVAC system to minimize energy while satisfying residents' comfort requirements. In theory, value-function-based reinforcement learning methods can solve control problems such as this one optimally. However, since choosing an appropriate parametric representation of the value function turns out to be difficult, we develop an alternative method, which results in a practical algorithm for value function approximation in continuous state-spaces. To avoid the need to carefully design a parametric representation for the value function, we use a smooth non-parametric function approximator, specifically Locally Weighted Linear Regression (LWR). LWR is used within Fitted Value Iteration (FVI), which has met with several practical successes. However, for efficiency reasons, LWR is used with a limited sample-size, which leads to poor performance without careful tuning of LWR's parameters. We therefore develop an efficient meta-learning procedure that performs online model-selection and tunes LWR's parameters based on the Bellman error. Our algorithm is fully implemented and tested in a realistic simulation of the HVAC control domain, and results in significant energy savings.

Tue3B: Matrix Analysis

Room: Dialog

16:25-16:45 **Noisy Matrix Completion Using Alternating Minimization**
Suriya Gunasekar, Ayan Acharya, Neeraj Gaur and Joydeep Ghosh

The task of matrix completion involves estimating the entries of a matrix, $M \in R^{m \times n}$, when a subset, $\Omega \subset \{(i,j): 1 \leq i \leq m, 1 \leq j \leq n\}$ of the entries are observed. A particular set of low rank models for this task approximate the matrix as a product of two low rank matrices, $\hat{M} = U V^T$, where $U \in R^{m \times k}$ and $V \in R^{n \times k}$ and $k \ll min\{m,n\}$. A popular algorithm of choice in practice for recovering M from the partially observed matrix using the low rank assumption is alternating least square (ALS) minimization, which involves optimizing over U and V in an alternating manner to minimize the squared error over observed entries while keeping the other factor fixed. Despite being widely experimented in practice, only recently were theoretical guarantees established bounding the error of the matrix estimated from ALS to that of the original matrix M. In this work we extend the results for a noiseless setting and provide the first guarantees for recovery under noise for alternating minimization. We specifically show that for well conditioned matrices corrupted by random noise of bounded Frobenius norm, if the number of observed entries is $O(k^7 n \log n)$, then the ALS algorithm recovers the original matrix within an error bound that depends on the norm of the noise matrix. The sample complexity is the same for the noise-free matrix completion using ALS.

16:45-17:05 **A Nearly Unbiased Matrix Completion Approach**
Dehua Liu, Tengfei Zhou, Hui Qian, Congfu Xu and Zhihua Zhang

Low-rank matrix completion is an important theme both theoretically and practically. However, the state-of-the-art methods based on convex optimization usually lead to a certain amount of deviation from the original matrix. To perfectly recover a data matrix from a sampling of its entries, we consider a non-convex alternative to approximate the matrix rank. In particular, we minimize a matrix γ-norm under a set of linear constraints. Accordingly, we derive a shrinkage operator, which is nearly unbiased in comparison with the well-known soft shrinkage operator. Furthermore, we devise two algorithms, non-convex soft imputation (NCSI) and non-convex alternative direction method of multipliers (NCADMM), to fulfil the numerical estimation. Experimental results show that these algorithms outperform existing matrix completion methods in accuracy. Moreover, the NCADMM is as efficient as the current state-of-the-art algorithms.

17:05-17:25 **A Counterexample for the Validity of Using Nuclear Norm as a Convex Surrogate of Rank**
Hongyang Zhang, Zhouchen Lin and Chao Zhang

Rank minimization has attracted a lot of attention due to its robustness in data recovery. To overcome the computational difficulty, rank is often replaced with nuclear norm. For several rank minimization problems, such a replacement has been theoretically proven to be valid, i.e., the solution to nuclear norm minimization problem is also the solution to rank minimization problem. Although it is easy to believe that such a replacement may not always be valid, no concrete example has ever been found. We argue that such a validity checking cannot be done by numerical computation and show, by analyzing the noiseless latent low rank representation (LatLRR) model, that even for very simple rank minimization problems the validity may still break down. As a by-product, we find that the solution to the nuclear norm minimization formulation of LatLRR is non-unique. Hence the results of LatLRR reported in the literature may be questionable.

17:25-17:45 **Efficient Rank-one Residue Approximation Method for Graph Regularized Non-negative Matrix Factorization**
Qing LIAO and Qian Zhang

Nonnegative matrix factorization (GNMF) aims to decompose a given data matrix X into the product of two lower-rank nonnegative factor matrices UV^T. Graph regularized NMF (GNMF) is a recently proposed NMF method that and preserves its the geometric structure of X during such decomposition. Although It GNMF has been widely used in computer vision and data mining. However, the its multiplicative update rule (MUR) based algorithm solver suffers from both slow convergence and non-stationarity problems. In this paper, we propose a new efficient GNMF solver called rank-one residue approximation (RRA). Different from MUR, which updates both factor matrices (U and V) in a whole in each iteration round, RRA updates each column of them by approximating the residue matrix by their outer product. Since each column of both factor matrices is updated optimally in an analytic formulation, RRA is theoretical and empirically proved to converge rapidly to a stationary point both in theoretical analysis and empirical experiments. Moreover, since RRA need neither extra computational cost nor parameter tuning, it enjoys the simplicity with MUR but performs much faster than MUR. Experimental results on both dense and sparse datasetsreal-world datasets show that RRA is much more efficient than MUR for GNMF. To confirm the stationarity of the solution obtained by RRA, we conduct clustering experiments on real-world image datasets by comparing with the representative solvers for GNMF. The experimental results confirm the effectiveness of RRA.

17:45-18:05 **Maximum Entropy Models for Iteratively Identifying Subjectively Interesting Structure in Real-Valued Data**
Kleanthis-Nikolaos Kontonasios, Jilles Vreeken and Tijl De Bie

In exploratory data mining it is important to assess the significance of results. Given that analysts have only limited time, it is important that we can measure this with regard to what we already know. That is, we want to be able to measure whether a result is interesting from a subjective point of view. With this as our goal, we formalise how to probabilistically model real-valued data by the Maximum Entropy principle, where we allow statistics on arbitrary sets of cells as background knowledge. As statistics, we consider means and variances, as well as histograms. The resulting models allow us to assess the likelihood of values, and can be used to verify the significance of (possibly overlapping) structures discovered in the data. As we can feed those structures back in, our model enables iterative identification of subjectively interesting structures. To show the flexibility of our model, we propose a subjective informativeness measure for tiles, i.e. rectangular sub-matrices, in real-valued data. The Information Ratio quantifies how strongly the knowledge of a structure reduces our uncertainty about the data with the amount of effort it would cost to consider it. Empirical evaluation shows that iterative scoring effectively reduces redundancy in ranking candidate tiles-showing the applicability of our model for a range of data mining fields aimed at discovering structure in real-valued data.

Tue3C: Applications

Room: Ceremonie

16:25-16:45 **Incremental Sensor Placement Optimization on Water Network**
Xiaomin Xu, Yiqi Lu, Sheng Huang, Yanghua Xiao and Wei Wang

Sensor placement on water networks is critical for the detection of accidental or intentional contamination event. With the development and expansion of cities, the public water distribution systems in cities are continuously growing. As a result, the current sensor placement will lose its effectiveness in detecting contamination event. Hence, in many real applications, we need to solve the incremental sensor placement (ISP) problem.We expect to find a sensor placement solution that reuses existing sensor deployments as much as possible to reduce cost, while ensuring the effectiveness of contamination detection. In this paper, we propose scenario-cover model to formalize ISP and prove that ISP is NP-hard and propose our greedy approaches with provable quality bound. Extensive experiments show the effectiveness, robustness and efficiency of the proposed solutions.

16:45-17:05 **Detecting Marionette Microblog Users for Improved Information Credibility**
Xian Wu, Ziming Feng, Wei Fan, Jing Gao and Yong Yu

In this paper, we mine a special group of microblog users: the marionette users, who are created or employed by backstage puppeteers, either through programs or manually. Unlike normal users that access microblogs for information sharing or social communication, the marionette users perform specific tasks to earn financial profits. For example, they follow certain users to increase their statistical popularity, or retweet some tweets to amplify their statistical impact. The fabricated follower or retweet counts not only mislead normal users to wrong information, but also seriously impair microblog-based applications, such as popular tweets selection and expert finding. In this paper, we study the important problem of detecting marionette users on microblog platforms. This problem is challenging because puppeteers are employing complicated strategies to generate marionette users that present similar behaviors as normal ones. To tackle this challenge, we propose to take into account two types of discriminative information: (1) individual user tweeting behaviors and (2) the social interactions among users. By integrating both information into a semi-supervised probabilistic model, we can effectively distinguish marionette users from normal ones. By applying the proposed model to one of the most popular microblog platform (Sina Weibo) in China, we find that the model can detect marionette users with f-measure close to 0.9. In addition, we propose an application to measure the credibility of retweet counts.

17:05-17:25 **Will my Question be Answered? Predicting "Question Answerability" in Community Question-Answering Sites**
Gideon Dror, Yoelle Maarek and Idan Szpektor

All askers who post questions in Community-based Question Answering (CQA) sites such as Yahoo! Answers, Quora or Baidu's Zhidao, expect to receive an answer, and are frustrated when their questions remain unanswered. We propose to provide a type of "heads up" to askers by predicting how many answers, if at all, they will get. Giving a preemptive warning to the asker at posting time should reduce the frustration effect and hopefully allow askers to rephrase their questions if needed. To the best of our knowledge, this is the first attempt to predict the actual number of answers, in addition to predicting whether the question will be answered or not. To this effect, we introduce a new prediction model, specifically tailored to hierarchically structured CQA sites.We conducted extensive experiments on a large corpus comprising 1 year of answering activity on Yahoo! Answers, as opposed to a single day in previous studies. These experiments show that the F1 we achieved is 24% better than in previous work, mostly due the structure built into the novel model.

17:25-17:45 **Learning to Detect Patterns of Crime**
Tong Wang, Cynthia Rudin, Dan Wagner and Rich Sevieri

Our goal is to automatically detect patterns of crime. Among a large set of crimes that happen every year in a major city, it is challenging, time-consuming, and labor-intensive for crime analysts to determine which ones may have been committed by the same individual(s). If automated, data-driven tools for crime pattern detection are made available to assist analysts, these tools could help police to better understand patterns of crime, leading to more precise attribution of past crimes, and the apprehension of suspects. To do this, we propose a pattern detection algorithm called Series Finder, that grows a pattern of discovered crimes from within a database, starting from a seed of a few crimes. Series Finder incorporates both the common characteristics of all patterns and the unique aspects of each specic pattern, and has had promising results on a decade's worth of crime pattern data collected by the Crime Analysis Unit of the Cambridge Police Department.

17:45-18:05 **Space Allocation in the Retail Industry: A Decision Support System Integrating Evolutionary Algorithms and Regression Models**
Fabio Pinto and Carlos Soares

One of the hardest resources to manage in retail is space. Retailers need to assign limited store space to a growing number of product categories such that sales and other performance metrics are maximized. Although this seems to be an ideal task for a data mining approach, there is one important barrier: the representativeness of the available data. In fact, changes to the layout of retail stores are infrequent. This means that very few values of the space variable are represented in the data, which makes it hard to generalize. In this paper, we describe a Decision Support System to assist retailers in this task. The system uses an Evolutionary Algorithm to optimize space allocation based on the estimated impact on sales caused by changes in the space assigned to product categories. We assess the quality of the system on a real case study, using different regression algorithms to generate the estimates. The system obtained very good results when compared with the recommendations made by the business experts. We also investigated the effect of the representativeness of the sample on the accuracy of the regression models. We selected a few product categories based on a heuristic assessment of their representativeness. The results indicate that the best regression models were obtained on products for which the sample was not the best. The reason for this unexpected results remains to be explained.

Tue3D: Semi-supervised Learning

Room: Aplaus

16:25-16:45 **Exploratory Learning**
Bhavana Dalvi, William Cohen and Jamie Callan

In multiclass semi-supervised learning (SSL), it is sometimes the case that the number of classes present in the data is not known, and hence no labeled examples are provided for some classes. In this paper we present variants of well-known semi-supervised multiclass learning methods that are robust when the data contains an unknown number of classes. In particular, we present an "exploratory" extension of expectation-maximization (EM) that explores different numbers of classes while learning. "Exploratory" SSL greatly improves performance on three datasets in terms of F1 on the classes with seed examples-i.e., the classes which are expected to be in the data. Our Exploratory EM algorithm also outperforms a SSL method based non-parametric Bayesian clustering.

16:45-17:05 **Semi-supervised Gaussian Process Ordinal Regression**
Srijith P. K., Shirish Shevade and Sundararajan S.

Ordinal regression problem arises in situations where examples are rated in an ordinal scale. In practice, labeled ordinal data are difficult to obtain while unlabeled ordinal data are available in abundance. Designing a probabilistic semi-supervised classifier to perform ordinal regression is challenging. In this work, we propose a novel approach for semi-supervised ordinal regression using Gaussian Processes(GP). It uses the expectation-propagation approximation idea, widely used for GP ordinal regression problem. The proposed approach makes use of the unlabeled data in addition to the labeled data to learn a model by matching ordinal label distributions approximately between labeled and unlabeled data. The resulting mixed integer programming problem, involving model parameters (real-valued) and ordinal labels (integers) as variables, is solved efficiently using a sequence of alternating optimization steps. Experimental results on synthetic, bench-mark and real-world data sets demonstrate that the proposed GP based approach makes effective use of the unlabeled data to give better generalization performance (on the absolute error metric, in particular) than the supervised approach. Thus, it is a useful approach for probabilistic semi-supervised ordinal regression problem.

17:05-17:25 **Influence of Graph Construction on Semi-supervised Learning**
Celso Andre De Sousa, Gustavo Batista and Solange Rezende

A variety of graph-based semi-supervised learning (SSL) algorithms and graph construction methods have been proposed in the last few years. Despite their apparent empirical success, the field of SSL lacks a detailed study that empirically evaluates the influence of graph construction on SSL. In this paper we provide such an experimental study. We combine a variety of graph construction methods as well as a variety of graph-based SSL algorithms and empirically compare them on a number of benchmark data sets widely used in the SSL literature. The empirical evaluation proposed in this paper is subdivided into four parts: (1) best case analysis; (2) classifiers' stability evaluation; (3) influence of graph construction; and (4) influence of regularization parameters. The purpose of our experiments is to evaluate the trade-off between classification performance and stability of the SSL algorithms on a variety of graph construction methods and parameter values. The obtained results show that the mutual k-nearest neighbors (mutKNN) graph may be the best choice for adjacency graph construction while the RBF kernel may be the best choice for weighted matrix generation. In addition, mutKNN tends to generate smoother error surfaces than other adjacency graph construction methods. However, mutKNN is unstable for a relatively small value of k. Our results indicate that the classification performance of the graph-based SSL algorithms are heavily influenced by the parameters setting and we found no evident explorable pattern to relay to future practitioners. We discuss the consequences of such instability in research and practice.

17:25-17:45 **Tractable Semi-Supervised Learning of Complex Structured Prediction Models**
Kai-Wei Chang, Sundararajan S. and Sathiya Keerthi

Semi-supervised learning has been widely studied in the literature. However, most previous works assume that the output structure is simple enough to allow the direct use of tractable inference/learning algorithms (e.g., binary la-bel or linear chain). Therefore, these methods cannot be applied to problems with complex structure. In this paper, we propose an approximate semi-supervised learning method that uses piecewise training for estimating the model weights and a dual decomposition approach for solving the inference problem of finding the labels of unlabeled data subject to domain specific constraints. This allows us to extend semi-supervised learning to general structured prediction problems. As an example, we apply this approach to the problem of multi-label classification (a fully connected pairwise Markov random field). Experimental results on benchmark data show that, in spite of using approximations, the approach is effective and yields good improvements in generalization performance over the plain supervised method. In addition, we demonstrate that our inference engine can be applied to other semi-supervised learning frameworks, and extends them to solve problems with complex structure.

17:45-18:05 **PSSDL: Probabilistic Semi-Supervised Dictionary Learning**
Behnam Babagholami-Mohamadabadi, Ali Zarghami, Mohammadreza Zolfaghari and Mahdieh Soleymani Baghshah

While recent supervised dictionary learning methods have attained promising results on the classification tasks, their performance depends on the availability of the large labeled datasets. However, in many real world applications, accessing to sufficient labeled data may be expensive and/or time consuming, but its relatively easy to acquire a large amount of unlabeled data. In this paper, we propose a probabilistic framework for discriminative dictionary learning which uses both the labeled and unlabeled data. Experimental results demonstrate that the performance of the proposed method is significantly better than the state of the art dictionary based classification methods.

WEDNESDAY 25 SEPTEMBER 2013

WEDNESDAY INVITED TALK

Using Machine Learning Powers for Good

Speaker: Rayid Ghani
Time: 09:00-10:00
Room: plenary

Abstract
The past few years have seen increasing demand for machine learning and data mining-both for tools as well as experts. This has been mostly motivated by a variety of factors including better and cheaper data collection, realization that using data is a good thing, and the ability for a lot of organizations to take action based on data analysis. Despite this flood of demand, most applications we hear about in machine learning involve search, advertising, and financial areas. This talk will talk about examples on how the same approaches can be used to help governments and non-profits make social impact. I'll talk about a summer fellowship program we ran at University of Chicago on social good and show examples from projects in areas such as education, healthcare, energy, transportation and public safety done in conjunction with governments and non-profits.

Bio
Rayid Ghani was the Chief Scientist at the Obama for America 2012 campaign focusing on analytics, technology, and data. His work focused on improving different functions of the campaign including fundraising, volunteer, and voter mobilization using analytics, social media, and machine learning: his innovative use of machine learning and data mining in Obama's reelection campaign received broad attention in the media such as the New York Times, CNN, and others. Before joining the campaign, Rayid was a Senior Research Scientist and Director of Analytics research at Accenture Labs where he led a technology research team focused on applied R&D in analytics, machine learning, and data mining for large-scale & emerging business problems in various industries including healthcare, retail & CPG, manufacturing, intelligence, and financial services. In addition, Rayid serves as an adviser to several start-ups in Analytics, is an active organizer of and participant in academic and industry analytics conferences, and publishes regularly in machine learning and data mining conferences and journals.

WEDNESDAY SESSIONS AT A GLANCE

Wed1A: Nectar (1)

Room: Euforie

10:30-11:20 **Towards Robot Skill Learning: From Simple Skills to Table Tennis**
Jan Peters, Katharina Muelling, Jens Kober, Oliver Kroemer and Gerhard Neumann

11:20-12:10 **Functional MRI Analysis with Sparse Models**
Irina Rish

Wed1B: Active Learning and Optimization

Room: Dialog

10:30-10:50 **A Lipschitz Exploration-Exploitation Scheme for Bayesian Optimization**
Ali Jalali, Javad Azimi, Xiaoli Fern and Ruofei Zhang

10:50-11:10 **Parallel Gaussian Process Optimization with Upper Confidence Bound and Pure Exploration**
Emile Contal, David Buffoni, Alexandre Robicquet and Nicolas Vayatis

11:10-11:30 **Regret Bounds for Reinforcement Learning with Policy Advice**
Mohammad Azar, Alessandro Lazaric and Emma Brunskill

11:30-11:50 **A Time and Space Efficient Algorithm for Contextual Linear Bandits**
Jose Bento, Stratis Ioannidis, S Muthukrishnan and Jinyun Yan

11:50-12:10 **Knowledge Transfer for Multi-labeler Active Learning**
Meng Fang, Jie Yin and Xingquan Zhu

Wed1C: Networks (2)

Room: Ceremonie

10:30-10:55 **ABACUS: Frequent Pattern Mining Based Community Discovery in Multidimensional Networks**
Michele Berlingerio, Fabio Pinelli and Francesco Calabrese

10:55-11:15 **Discovering Nested Communities**
Nikolaj Tatti and Aristides Gionis

11:15-11:35 **CSI: Community-Level Social Influence analysis**
Yasir Mehmood, Nicola Barbieri, Francesco Bonchi and Antti Ukkonen

11:35-11:55 **Community Distribution Outlier Detection in Heterogeneous Information Networks**
Manish Gupta, Jing Gao and Jiawei Han

Wed1D: Structured Output, Multi-task

Room: Aplaus

10:30-10:50 **Taxonomic Prediction with Tree-Structured Covariances**
Matthew Blaschko, Wojciech Zaremba and Arthur Gretton

10:50-11:10 **Position Preserving Multi-output Prediction**
Zubin Abraham and Pang-Ning Tan

11:10-11:30 **Structured Output Learning with Candidate Labels for Local Parts**
Chengtao Li, Jianwen Zhang and Zheng Chen

11:30-11:50 **Shared Structure Learning for Multiple Tasks with Multiple Views**
Xin Jin, Fuzhen Zhuang, Shuhui Wang, Qing He and Zhongzhi Shi

11:50-12:10 **Using Both Latent and Supervised Shared Topics for Multitask Learning**
Ayan Acharya, Aditya Rawal, Eduardo Hruschka and Raymond Mooney

Wed2A: Nectar (2)

Room: Euforie

14:00-14:33 **A Theoretical Framework for Exploratory Data Mining: Recent Insights and Challenges Ahead**
Tijl De Bie and Eirini Spyropoulou

14:33-15:06 **Tensor Factorization for Multi-relational Learning**
Maximilian Nickel and Volker Tresp

15:06-15:40 **MONIC–Modeling and Monitoring Cluster Transitions**
Myra Spiliopoulou, Eirini Ntoutsi and Yannis Theodoridis

Wed2B: Models for Sequential Data

Room: Dialog

14:00-14:20 **Spectral Learning of Sequence Taggers over Continuos Sequences**
Ariadna Quattoni and Adria Recasens

14:20-14:40 **Fast Variational Bayesian Linear State-Space Model**
Jaakko Luttinen

14:40-15:00 **Inhomogeneous Parsimonious Markov Models**
Ralf Eggeling, Andre Gohr, Pierre-Yves Bourguignon, Edgar Wingender and Ivo Grosse

15:00-15:20 **Future Locations Prediction with Uncertain Data**
Disheng Qiu, Paolo Papotti and Blanco Lorenzo

15:20-15:40 **Modeling Short-Term Energy Load with Continuous Conditional Random Fields**
Hongyu Guo

Wed2C: Graph Mining

Room: Ceremonie

14:00-14:25 **Activity Preserving Graph Simplification**
Francesco Bonchi, Gianmarco De Francisci Morales, Aristides Gionis and Antti Ukkonen

14:25-14:45 **A Fast Approximation of the Weisfeiler-Lehman Graph Kernel for RDF Data**
Gerben De Vries

14:45-15:05 **Improving Relational Classification Using Link Prediction Techniques**
Cristina Perez-Sola and Jordi Herrera-Joancomarti

15:05-15:25 **Efficient Frequent Connected Induced Subgraph Mining in Graphs of Bounded Treewidth**
Tamas Horvath, Keisuke Otaki and Jan Ramon

Wed2D: Natural Language Processing & Probabilistic Models

Room: Aplaus

14:00-14:20 **Supervised Learning of Syntactic Contexts for Uncovering Definitions and Extracting Hypernym Relations in Text Databases**
Guido Boella and Luigi Di Caro

14:20-14:40 **Error Prediction with Partial Feedback**
William Darling, Guillaume Bouchard, Shachar Mirkin and Cedric Archambeau

14:40-15:00 **Boot-Strapping Language Identifiers for Short Colloquial Postings**
Moises Goldszmidt, Marc Najork and Stelios Paparizos

15:00-15:20 **A Bayesian Classifier for Learning from Tensorial Data**
Wei Liu, Jeffrey Chan, James Bailey, Christopher Leckie, Fang Chen and Rao Kotagiri

15:20-15:40 **Prediction with Model-Based Neutrality**
Kazuto Fukuchi, Jun Sakuma and Toshihiro Kamishima

Wed3A: Subgroup Discovery & Streams

Room: Euforie

16:10-16:30 **Discovering Skylines of Subgroup Sets**
Matthijs van Leeuwen and Antti Ukkonen

16:30-16:50 **Difference-Based Estimates for Generalization-Aware Subgroup Discovery**
Florian Lemmerich, Martin Becker and Frank Puppe

16:50-17:10 **Fast and Exact Mining of Probabilistic Data Streams**
Reza Akbarinia and Florent Masseglia

17:10-17:30 **Pitfalls in benchmarking data stream classification and how to avoid them**
Albert Bifet, Jesse Read, Indre Zliobaite, Bernhard Pfahringer and Geoff Holmes

17:30-17:50 **Adaptive Model Rules from High-Speed Data Streams**
Joao Gama, Ezilda Almeida, Carlos Ferreira and Petr Kosina

Wed3B: Multi-label Classification & Outlier Detection

Room: Dialog

16:10-16:30 **Multi-label Classification with Output Kernels**
Yuhong Guo and Dale Schuurmans

16:30-16:50 **Probabilistic Clustering for Hierarchical Multi-label Classification of Protein Functions**
Rodrigo Barros, Ricardo Cerri, Alex Freitas and Andre Carvalho

16:50-17:10 **Mining Outlier Participants: Insights Using Directional Distributions in Latent Models**
Didi Surian and Sanjay Chawla

17:10-17:30 **Anomaly Detection in Vertically Partitioned Data by Distributed Core Vector Machines**
Marco Stolpe, Kanishka Bhaduri, Kamalika Das and Katharina Morik

17:30-17:50 **Local Outlier Detection with Interpretation**
Xuan-Hong Dang, Barbora Micenkova, Ira Assent and Raymond T. Ng

Wed3C: Ensembles

Room: Ceremonie

16:10-16:30 **Boosting for Unsupervised Domain Adaptation**
Amaury Habrard, Jean-Philippe Peyrache and Marc Sebban

16:30-16:50 **AR-Boost: Reducing Overfitting by a Robust Data-Driven Regularization Strategy**
Baidya Nath Saha, Gautam Kunapuli, Nilanjan Ray, Joseph Maldjian and Sriraam Natarajan

16:50-17:10 **Parallel Boosting with Momentum**
Indraneel Mukherjee, Yoram Singer, Rafael Frongillo and Kevin Canini

17:10-17:30 **Inner Ensembles: Using Ensemble Methods in the Learning Phase**
Houman Abbasian, Chris Drummond, Nathalie Japkowicz and Stan Matwin

17:30-17:50 **Mixtures of Large Margin Nearest Neighbor Classifiers**
Murat Semerci and Ethem Alpaydin

Wed3D: Bayesian Learning

Room: Aplaus

16:10-16:35 **A Comparative Evaluation of Stochastic-Based Inference Methods for Gaussian Process**
Maurizio Filippone, Mingjun Zhong and Mark Girolami

16:35-16:55 **Decision-Theoretic Sparsification for Gaussian Process Preference Learning**
Ehsan Abbasnejad, Edwin Bonilla and Scott Sanner

16:55-17:15 **Variational Hidden Conditional Random Fields with Coupled Dirichlet Process Mixtures**
Konstantinos Bousmalis, Stefanos Zafeiriou, Louis-Philippe Morency, Maja Pantic and Zoubin Ghahramani

17:15-17:35 **Sparsity in Bayesian Blind Source Separation and Deconvolution**
Vaclav Smidl and Ondrej Tichy

WEDNESDAY SESSIONS, WITH ABSTRACTS

Wed1A: Nectar (1)

Room: Euforie

10:30-11:20 **Towards Robot Skill Learning: From Simple Skills to Table Tennis**
Jan Peters, Katharina Muelling, Jens Kober, Oliver Kroemer and Gerhard Neumann

Learning robots that can acquire new motor skills and re- fine existing ones have been a long standing vision of both robotics, and machine learning. However, off-the-shelf machine learning approaches ap- pear not to be adequate for robot skill learning as they neither scale into anthropomorphic robotics nor do they fulfill the crucial real-time require- ments. As an alternative, we propose to divide the generic skill learning problem into parts that can be well-understood from a robotics point of view. In this context, we have developed machine learning methods that scale into robot skill learning. This paper discusses our recent progress ranging from simple skill learning problems to robot table tennis.

11:20-12:10 **Functional MRI Analysis with Sparse Models**
Irina Rish

Sparse models embed variable selection into model learning (e.g., by using l_1 norm regularizer). In small-sample high-dimensional problems, this leads to improved generalization accuracy combined with interpretability, which is important in scientific applications such as biology. In this paper, we summarize our recent work on sparse models, including both sparse regression and sparse Gaussian Markov Random Fields (GMRF), in neuroimaging applications, such as functional MRI data analysis, where the central objective is to gain a better insight into brain functioning, besides just learning predictive models of mental states from imaging data.

Wed1B: Active Learning and Optimization

Room: Dialog

10:30-10:50 **A Lipschitz Exploration-Exploitation Scheme for Bayesian Optimization**
Ali Jalali, Javad Azimi, Xiaoli Fern and Ruofei Zhang

The problem of optimizing unknown costly-to-evaluate functions has been studied extensively in the context of Bayesian optimization. Algorithms in this field aim to find the optimizer of the function by requesting only a few function evaluations at carefully selected locations. An ideal algorithm should maintain a perfect balance between exploration (probing unexplored areas) and exploitation (focusing on promising areas) within the given evaluation budget. In this paper, we assume the unknown function is Lipschitz continuous. Leveraging the Lipschitz property, we propose an algorithm with a distinct exploration phase followed by an exploitation phase. The exploration phase aims to select samples that shrink the search space as much as possible, while the exploitation phase focuses on the reduced search space and selects samples closest to the optimizer. We empirically show that the proposed algorithm significantly outperforms the baseline algorithms.

10:50-11:10 **Parallel Gaussian Process Optimization with Upper Confidence Bound and Pure Exploration**
Emile Contal, David Buffoni, Alexandre Robicquet and Nicolas Vayatis

In this paper, we consider the challenge of maximizing an unknown function f for which evaluations are noisy and are acquired with high cost. An iterative procedure uses the previous measures to actively select the next estimation of f which is predicted to be the most useful. We focus on the case where the function can be evaluated in parallel with batches of fixed size and analyze the benefit compared to the purely sequential procedure in terms of cumulative regret. We introduce the Gaussian Process Upper Confidence Bound and Pure Exploration algorithm (GP-UCB-PE) which combines the UCB strategy and Pure Exploration in the same batch of evaluations along the parallel iterations. We prove theoretical upper bounds on the regret with batches of size K for this procedure which show the improvement of the order of $\sqrt{K}$ for fixed iteration cost over purely sequential versions. Moreover, the multiplicative constants involved have the property of being dimension-free. We also confirm empirically the efficiency of GP-UCB-PE on real and synthetic problems compared to state-of-the-art competitors.

11:10-11:30 **Regret Bounds for Reinforcement Learning with Policy Advice**
Mohammad Azar, Alessandro Lazaric and Emma Brunskill

We address the practical problem of maximizing the number of high-confidence results produced among multiple experiments sharing an exhaustible pool of fungible resources. We formalize this problem in the framework of bandit optimization as follows: given a set of multiple multi-armed bandits and a budget on the total number of trials allocated among them, select the top-m arms (with high confidence) for as many of the bandits as possible. To solve this problem, which we call greedy confidence pursuit, we develop a method based on posterior sampling. We show empirically that our method outperforms existing methods for the single bandit top-m selection problem, which has been studied previously, and improves performance over baseline methods for the full greedy confidence pursuit problem, which has not been studied previously.

11:30-11:50 **A Time and Space Efficient Algorithm for Contextual Linear Bandits**
Jose Bento, Stratis Ioannidis, S Muthukrishnan and Jinyun Yan

We consider a multi-armed bandits problem where payoffs are a linear function of an observed stochastic contextual variable. In the scenario where there exists a gap between optimal and suboptimal rewards, several algorithms have been proposed that achieve $O(\log T)$ regret after T time steps. However, proposed methods either have a computation complexity per iteration that scales linearly with T or achieve regrets that grow linearly with the number of contexts, $|\{X\}|$. We propose an εgreedy type of algorithm that solves both limitations. In particular, when contexts are variables in R^d, we prove that our algorithm has a constant computation complexity per iteration of $O(poly(d))$ and can achieve a regret of $O(poly(d)\log T)$ even when $|\{X\}| = \Omega(2^d)$. In addition, unlike previous algorithms, its space complexity scales like $O(K d^2)$ and does not grow with T.

11:50-12:10 **Knowledge Transfer for Multi-labeler Active Learning**
Meng Fang, Jie Yin and Xingquan Zhu

In this paper, we address multi-labeler active learning, where data labels can be acquired from multiple labelers with various levels of expertise. Because obtaining labels for data instances can be very costly and time-consuming, it is highly desirable to model each labeler's expertise and only to query an instance's label from the labeler with the best expertise. However, in an active learning scenario, it is very difficult to accurately model labelers' expertise, because the quantity of instances labeled by all participating labelers is rather small. To solve this problem, we propose a new probabilistic model that transfers knowledge from a rich set of labeled instances in some auxiliary domains to help model labelers' expertise for active learning. Based on this model, we present an active learning algorithm to simultaneously select the most informative instance and its most reliable labeler to query. Experiments demonstrate that transferring knowledge across related domains can help select the labeler with the best expertise and thus significantly boost the active learning performance.

Wed1C: Networks (2)

Room: Ceremonie

10:30-10:55 **ABACUS: Frequent Pattern Mining Based Community Discovery in Multidimensional Networks**
Michele Berlingerio, Fabio Pinelli and Francesco Calabrese

Community Discovery in complex networks is the problem of detecting, for each node of the network, its membership to one of more groups of nodes, the communities, that are densely connected, or highly interactive, or, more in general, similar, according to a similarity function. So far, the problem has been widely studied in monodimensional networks, i.e. networks where only one connection between two entities can exist. However, real networks are often multidimensional, i.e., multiple connections between any two nodes can exist, either reflecting different kinds of relationships, or representing different values of the same type of tie. In this context, the problem of Community Discovery has to be redefined, taking into account multidimensional structure of the graph. We define a new concept of community that groups together nodes sharing memberships to the same monodimensional communities in the different single dimensions. As we show, such communities are meaningful and able to group highly correlated nodes, even if they might not be connected in any of the monodimensional networks. We devise ABACUS (Apriori-BAsed Community discoverer in mUltidimensional networkS), an algorithm that is able to extract multidimensional communities based on the apriori itemset miner applied to monodimensional community memberships. Experiments on two different real multidimensional networks confirm the meaningfulness of the introduced concepts, and open the way for a new class of algorithms for community discovery that do not rely on the dense connections among nodes.

10:55-11:15 **Discovering Nested Communities**
Nikolaj Tatti and Aristides Gionis

Discovering communities is one of the most well-studied problems in graph mining. Whenever the graph does not have a clear clique structure, selecting a single community is a difficult choice as the optimization criterion has to make a difficult choice between selecting a tight but small community or a more inclusive but a more sparse community.
In order to avoid the problem of selecting only a single community we propose discovering a sequence of nested communities. More formally, given a graph and a starting set, our goal is to discover a sequence of communities all containing the starting set and each community is a more dense subgraph of the next community. Discovering optimal sequence is a complex optimization problem, and hence we divide this problem into two subproblems: 1) discover the optimal sequence for a fixed order of vertices, a subproblem that we can solve efficiently, and 2) find a good order. We employ simple heuristic for discovering order and provide empirical and theoretical evidence that our order is good.

11:15-11:35 **CSI: Community-Level Social Influence analysis**
Yasir Mehmood, Nicola Barbieri, Francesco Bonchi and Antti Ukkonen

Modeling how information propagates in social networks driven by peer influence, is a fundamental research question towards understanding the structure and dynamics of these complex networks, as well as developing viral marketing applications. Existing literature studies influence at the level of individuals, mostly ignoring the existence of a community structure in which multiple nodes may exhibit a common influence pattern. In this paper we introduce CSI, a model for analyzing information propagation and social influence at the granularity of communities. CSI builds over a novel propagation model that generalizes the classic Independent Cascade model to deal with groups of nodes (instead of single nodes) influence. Given a social network and a database of past information propagation, we propose a hierarchical approach to detect a set of communities and their reciprocal influence strength. CSI provides a higher level and more intuitive description of the influence dynamics, thus representing a powerful tool to summarize and investigate patterns of influence in large social networks. The evaluation on various datasets suggests the effectiveness of the proposed approach in modeling information propagation at the level of communities. It further enables to detect interesting patterns of influence, such as the communities that play a key role in the overall diffusion process, or that are likely to start information cascades.

11:35-11:55 **Community Distribution Outlier Detection in Heterogeneous Information Networks**
Manish Gupta, Jing Gao and Jiawei Han

Heterogeneous networks are ubiquitous. For example, bibliographic data, social data, medical records, movie data and many more can be modeled as heterogeneous networks. Rich information associated with multi-typed nodes in heterogeneous networks motivates us to propose a new definition of outliers, which is different from those defined for homogeneous networks. In this paper, we propose the novel concept of Community Distribution Outliers (CDOutliers) for heterogeneous information networks, which are defined as objects whose community distribution does not follow any of the popular community distribution patterns.We extract such outliers using a type-aware joint analysis of multiple types of objects. Given community membership matrices for all types of objects, we follow an iterative two-stage approach which performs pattern discovery and outlier detection in a tightly integrated manner. We first propose a novel outlier-aware approach based on joint non-negative matrix factorization to discover popular community distribution patterns for all the object types in a holistic manner, and then detect outliers based on such patterns. Experimental results on both synthetic and real datasets show that the proposed approach is highly effective in discovering interesting community distribution outliers.

Wed1D: Structured Output, Multi-task

Room: Aplaus

10:30-10:50 **Taxonomic Prediction with Tree-Structured Covariances**
Matthew Blaschko, Wojciech Zaremba and Arthur Gretton

Taxonomies have been proposed numerous times in the literature in order to encode semantic relationships between classes. Such taxonomies have been used to improve classification results by increasing the statistical efficiency of learning, as similarities between classes can be used to increase the amount of relevant data during training. In this paper, we show how data-derived taxonomies may be used in a structured prediction framework, and compare the performance of learned and semantically constructed taxonomies. Structured prediction in this case is multi-class categorization with the assumption that categories are taxonomically related. We make three main contributions: (i) We prove the equivalence between tree-structured covariance matrices and taxonomies; (ii) We use this covariance representation to develop a highly computationally efficient optimization algorithm for structured prediction with taxonomies; (iii) We show that the taxonomies learned from data using the Hilbert-Schmidt Independence Criterion (HSIC) often perform better than imputed semantic taxonomies. Source code of this implementation, as well as machine readable learned taxonomies are available for download from https://github.com/blaschko/tree-structured-covariance/.

10:50-11:10 **Position Preserving Multi-output Prediction**
Zubin Abraham and Pang-Ning Tan

There is a growing demand for multiple output prediction methods capable of both minimizing residual errors and capturing the joint distribution of the response variables in a realistic and consistent fashion. Unfortunately, current methods are designed to optimize one of the two criteria, but not both. This paper presents a framework for multiple output regression that preserves the relationships among the response variables (including possible non-linear associations) while minimizing the residual errors of prediction by coupling regression methods with geometric quantile mapping. We demonstrate the effectiveness of the framework in modeling daily temperature and precipitation for climate stations in the Great Lakes region. We showed that, in all climate stations evaluated, the proposed framework achieves low residual errors comparable to standard regression methods while preserving the joint distribution of the response variables.

11:10-11:30 **Structured Output Learning with Candidate Labels for Local Parts**
Chengtao Li, Jianwen Zhang and Zheng Chen

This paper introduces a special setting of weakly supervised structured output learning, where the training data is a set of structured instances and supervision involves candidate labels for some local parts of the structure. We show that the learning problem with this weak supervision setting can be efficiently handled and then propose a large margin formulation. To solve the non-convex optimization problem, we propose a proper approximation of the objective to utilize the Constraint Concave Convex Procedure (CCCP). To accelerate each iteration of CCCP, a 2-slack cutting plane algorithm is proposed. Experiments on some sequence labeling tasks show the effectiveness of the proposed method.

11:30-11:50 **Shared Structure Learning for Multiple Tasks with Multiple Views**
Xin Jin, Fuzhen Zhuang, Shuhui Wang, Qing He and Zhongzhi Shi

Real-world problems usually exhibit dual-heterogeneity, i.e., every task in the problem has features from multiple views, and multiple tasks are related with each other through one or more shared views. To solve these multi-task problems with multiple views, we propose a shared structure learning framework, which can learn shared predictive structures on common views from multiple related tasks, and use the consistency among different views to improve the performance. An alternating optimization algorithm is derived to solve the proposed framework. Moreover, the computation load can be dealt with locally in each task during the optimization, through only sharing some statistics, which significantly reduces the time complexity and space complexity. Experimental studies on four real-world data sets demonstrate that our framework significantly outperforms the state-of-the-art baselines.

11:50-12:10 **Using Both Latent and Supervised Shared Topics for Multitask Learning**
Ayan Acharya, Aditya Rawal, Eduardo Hruschka and Raymond Mooney

This paper introduces two new frameworks, Doubly Supervised Latent Dirichlet Allocation (DSLDA) and its non-parametric variation (NP-DSLDA), that integrate two different types of supervision: topic labels and category labels. This approach is particularly useful for multitask learning, in which both latent and supervised topics are shared between multiple categories. Experimental results on both document and image classification show that both types of supervision improve the performance of both DSLDA and NP-DSLDA and that sharing both latent and supervised topics allows for better multitask learning.

Wed2A: Nectar (2)

Room: Euforie

14:00-14:33 **A Theoretical Framework for Exploratory Data Mining: Recent Insights and Challenges Ahead**
Tijl De Bie and Eirini Spyropoulou

Exploratory Data Mining (EDM), the contemporary heir of Exploratory Data Analysis (EDA) pioneered by Tukey in the seventies, is the task of facilitating the extraction of interesting nuggets of information from possibly large and complexly structured data. Major conceptual challenges in EDM research are the understanding of how one can formalise a nugget of information (given the diversity of types of data of interest), and how one can formalise how interesting such a nugget of information is to a particular user (given the diversity of types of users and intended purposes). In this Nectar paper we briefly survey a number of recent contributions made by us and collaborators towards a theoretically motivated and practically usable resolution of these challenges.

14:33-15:06 **Tensor Factorization for Multi-relational Learning**
Maximilian Nickel and Volker Tresp

Tensor factorization has emerged as a promising approach for solving relational learning tasks. Here we review recent results on a particular tensor factorization approach, i.e. RESCAL, which has demonstrated state-of-the-art relational learning results, while scaling to knowledge bases with millions of entities and billions of known facts.

15:06-15:40 **MONIC–Modeling and Monitoring Cluster Transitions**
Myra Spiliopoulou, Eirini Ntoutsi and Yannis Theodoridis

There is much recent discussion on data streams and big data, which except of their volume and velocity are also characterized by volatility. Next to detecting change, it is also important to interpret it. Consider customer profiling as an example: Is a cluster corresponding to a group of customers simply disappearing or are its members migrating to other clusters? Does a new cluster reflect a new type of customers or does it rather consist of old customers whose preferences shift? To answer such questions, we have proposed the framework MONIC for modeling and tracking cluster transitions. MONIC has been re-discovered some years after publication and is enjoying a large citation record from papers on community evolution, cluster evolution and change prediction.

Wed2B: Models for Sequential Data

Room: Dialog

14:00-14:20 **Spectral Learning of Sequence Taggers over Continuos Sequences**
Ariadna Quattoni and Adria Recasens

In this paper we present a spectral algorithm for learning weighted finite-state sequence taggers (WFSTs) over paired input-output sequences, where the input is continuous and the output discrete. WFSTs are an important tool for modeling paired input-output sequences and have numerous applications in real-world problems. Our approach is based on generalizing the class of weighted finite-state sequence taggers over discrete input-output sequences to a class where transitions are linear combinations of elementary transitions and the weights of the linear combination are determined by dynamic features of the continuous input sequence.The resulting learning algorithm is efficient and accurate.

14:20-14:40 **Fast Variational Bayesian Linear State-Space Model**
Jaakko Luttinen

This paper presents a fast variational Bayesian method for linear state-space models. The standard variational Bayesian expectation-maximization (VB-EM) algorithm is improved by a parameter expansion which optimizes the rotation of the latent space. With this approach, the inference is orders of magnitude faster than the standard method. The speed of the proposed method is demonstrated on an artificial dataset and a large real-world dataset, which shows that the standard VB-EM algorithm is not suitable for large datasets because it converges extremely slowly. In addition, the paper estimates the temporal state variables using a smoothing algorithm based on the block LDL decomposition. This smoothing algorithm reduces the number of required matrix inversions and avoids the model augmentation compared to previous approaches.

14:40-15:00 **Inhomogeneous Parsimonious Markov Models**
Ralf Eggeling, Andre Gohr, Pierre-Yves Bourguignon, Edgar Wingender and Ivo Grosse

We introduce inhomogeneous parsimonious Markov models for modeling statistical patterns in discrete sequences. These models are based on parsimonious context trees, which are a generalization of context trees, and thus generalize variable order Markov models. We follow a Bayesian approach, consisting of structure and parameter learning. Structure learning is a challenging problem due to an overexponential number of possible tree structures, so we describe an exact and efficient dynamic programming algorithm for finding the optimal tree structures. We apply model and learning algorithm to the problem of modeling binding sites of the human transcription factor C/EBP, and find an increased prediction performance compared to fixed order and variable order Markov models. We investigate the reason for this improvement and find several instances of context-specific dependences that can be captured by parsimonious context trees but not by traditional context trees.

15:00-15:20 **Future Locations Prediction with Uncertain Data**
Disheng Qiu, Paolo Papotti and Blanco Lorenzo

The ability to predict future movements for moving objects enables better decisions in terms of time, cost, and impact on the environment. Unfortunately, future location prediction is a challenging task. Existing methods exploit techniques to predict a trip destination, but they are effective only when location data are precise (e.g., GPS data) and movements are observed over long periods of time (e.g., weeks). We introduce a data mining approach based on a Hidden Markov Model (HMM) that overcomes these limits and improves existing results in terms of precision of the prediction, for both the route (i.e., trajectory) and the final destination. The model is resistant to uncertain location data, as it works with data collected by using cell-towers to localize the users instead of GPS devices, and reaches good prediction results in shorter times (days instead of weeks in a representative real-world application). Finally, we introduce an enhanced version of the model that is orders of magnitude faster than the standard HMM implementation.

15:20-15:40 **Modeling Short-Term Energy Load with Continuous Conditional Random Fields**
Hongyu Guo

Short-term energy load forecasting, such as hourly predictions for the next n (n >= 2) hours, will benefit from exploiting the relationships among the n estimated outputs. This paper treats such multi-steps ahead regression task as a sequence labeling (regression) problem, and adopts a Continuous Conditional Random Fields (CCRF) strategy. This discriminative approach intuitively integrates two layers: the first layer aims at the prior knowledge for the multiple outputs, and the second layer employs edge potential features to implicitly model the interplays of the n interconnected outputs. Consequently, the proposed CCRF makes predictions not only basing on observed features, but also considering the estimated values of related outputs, thus improving the overall predictive accuracy. In particular, we boost the CCRF's predictive performance with a multi-target function as its edge feature. These functions convert the relationship of related outputs with continuous values into a set of sub-relationships, each providing more specific feature constraints for the interplays of the related outputs. We applied the proposed approach to two real-world energy load prediction systems: one for electricity demand and another for gas usage. Our experimental results show that the proposed strategy can meaningfully reduce the predictive error for the two systems, in terms of mean absolute percentage error and root mean square error, when compared with three benchmarking methods. Promisingly, the relative error reduction achieved by our CCRF model was up to 50%.

Wed2C: Graph Mining

Room: Ceremonie

14:00-14:25 **Activity Preserving Graph Simplification**
Francesco Bonchi, Gianmarco De Francisci Morales, Aristides Gionis and Antti Ukkonen

We study the problem of simplifying a given directed graph by keeping a small subset of edges. Our goal is to maintain the connectivity required to explain a given set of observed traces of information propagation across the graph. Unlike previous work, we do not make any assumption about a model of information propagation. Instead, we approach the task as a combinatorial problem. We prove that the resulting optimization problem is NP-hard. We show that a standard greedy algorithm performs very well in practice, even though it does not have theoretical guarantees. Additionally, if the activity traces have a tree structure, we show that the objective function is supermodular, and experimentally verify that the approach for size-constrained submodular minimization recently proposed by Nagano et al. produces very good results. Moreover, when applied to the task of reconstructing an unobserved graph, our methods perform comparably to a state-of-the-art algorithm devised specifically for this task.

14:25-14:45 **A Fast Approximation of the Weisfeiler-Lehman Graph Kernel for RDF Data**
Gerben De Vries

In this paper we introduce an approximation of the Weisfeiler-Lehman graph kernel algorithm aimed at improving the computation time of the kernel when applied to Resource Description Framework (RDF) data. Typically, applying graph kernels to RDF is done by extracting subgraphs from a large RDF graph and computing the kernel on this set of subgraphs. In contrast, our algorithm computes the Weisfeiler-Lehman kernel directly on the large RDF graph, but still retains the subgraph information. We show that this algorithm is faster than the regular Weisfeiler-Lehman kernel for RDF data and has at least the same performance. Furthermore, we show that our method has similar or better performance, and is faster, than other recently introduced graph kernels for RDF.

14:45-15:05 **Improving Relational Classification Using Link Prediction Techniques**
Cristina Perez-Sola and Jordi Herrera-Joancomarti

In this paper, we address the problem of classifying entities belonging to networked datasets. We show that assortativity is positively correlated with classification performance and how we are able to improve the classification accuracy by increasing the assortativity of the network. Our method to increase assortativity is based on modifying the weights of the edges in a graph using a scoring function. We evaluate the ability of different scoring functions to serve for this purpose. Experimental results show that, for the appropriated scoring functions, classification over networks with edges weights modified outperforms the classification using the original edge weight.

15:05-15:25 **Efficient Frequent Connected Induced Subgraph Mining in Graphs of Bounded Treewidth**
Tamas Horvath, Keisuke Otaki and Jan Ramon

We study the frequent connected induced subgraph mining problem, i.e., the problem of listing all connected graphs that are induced subgraph isomorphic to a given number of transaction graphs. We first show that this problem cannot be solved for arbitrary transaction graphs in output polynomial time (if $P \neq NP$) and then prove that for graphs of bounded tree-width, frequent connected induced subgraph mining is possible in incremental polynomial time by levelwise search. Our algorithm is an adaptation of the technique developed for frequent connected subgraph mining in bounded tree-width graphs. While the adaptation is relatively natural for many steps of the original algorithm, we need entirely different combinatorial arguments to show the correctness and efficiency of the new algorithm. Since induced subgraph isomorphism between bounded tree-width graphs is NP-complete, the positive result of this paper provides another example of efficient pattern mining with respect to computationally intractable pattern matching operators.

Wed2D: Natural Language Processing & Probabilistic Models

Room: Aplaus

14:00-14:20 **Supervised Learning of Syntactic Contexts for Uncovering Definitions and Extracting Hypernym Relations in Text Databases**
Guido Boella and Luigi Di Caro

In this paper we address the problem of automatically constructing structured knowledge from plain texts. In particular, we present a supervised learning technique to first identify definitions in text data, while then finding hypernym relations within them making use of extracted syntactic structures. Instead of using pattern matching methods that rely on lexico-syntactic patterns, we propose a method which only uses syntactic dependencies between terms extracted with a syntactic parser. Our assumption is that syntax is more robust than patterns when coping with the length and the complexity of the texts. Then, we transform the syntactic contexts of each noun in a coarse-grained textual representation, that is later fed into hyponym/hypernym-centered Support Vector Machine classifiers. The results on an annotated dataset of definitional sentences demonstrate the validity of our approach overtaking the current state of the art.

14:20-14:40 **Error Prediction with Partial Feedback**
William Darling, Guillaume Bouchard, Shachar Mirkin and Cedric Archambeau

In this paper, we propose a probabilistic framework for predicting the root causes of errors in data processing pipelines made up of several components when we only have access to partial feedback: that is, we are aware when some error has occurred in one or more of the components, but we do not know which one. The proposed error model enables us to direct the user feedback to the correct components in the pipeline to either automatically correct errors as they occur, retrain the component with assimilated training examples, or take other corrective action. We present the model and describe an Expectation Maximization (EM)-based algorithm to learn the model parameters and predict the error configuration. We demonstrate the accuracy and usefulness of our method first on synthetic data, and then on two distinct tasks: error correction in a 2-component opinion summarization system, and phrase error detection in statistical machine translation.

14:40-15:00 **Boot-Strapping Language Identifiers for Short Colloquial Postings**
Moises Goldszmidt, Marc Najork and Stelios Paparizos

There is tremendous interest in mining the abundant user generated content on the web. Many analysis techniques are language dependent and rely on accurate language identification as a building block. Even though there is already research on language identification, it focused on very `clean' editorially managed corpora, on a limited number of languages, and on relatively large-sized documents. These are not the characteristics of the content to be found in say, Twitter or Facebook postings, which are short and riddled with vernacular. In this paper, we propose an automated, unsupervised, scalable solution based on publicly available data. To this end we thoroughly evaluate the use of Wikipedia to build language identifiers for a large number of languages (52) and a large corpus and conduct a large scale study of the best-known algorithms for automated language identification, quantifying how accuracy varies in correlation to document size, language (model) profile size and number of languages tested. Then, we show the value in using Wikipedia to train a language identifier directly applicable to Twitter. Finally, we augment the language models and customize them to Twitter by combining our Wikipedia models with location information from tweets. This method provides massive amount of automatically labeled data that act as a bootstrapping mechanism which we empirically show boosts the accuracy of the models. With this work we provide a guide and a publicly available tool to the mining community for language identification on web and social data.

15:00-15:20 **A Bayesian Classifier for Learning from Tensorial Data**
Wei Liu, Jeffrey Chan, James Bailey, Christopher Leckie, Fang Chen and Rao Kotagiri

Traditional machine learning methods characterize data observations by feature vectors, where an entry of a vector denotes a scalar feature value of a data instance. While this data representation facilitates the application of conventional machine learning algorithms, in many cases it is not the best way of extracting all useful information from the data observations. In this paper we relax the (often unstated) assumption of vectorizing features of data instances, and allow a more natural representation of the data in a tensor format. Tensors are multi-mode (aka multi-way) arrays, of whom vectors (i.e., one-mode tensors) and matrices (i.e., two-mode tensors) are special cases. We show that the tensor representation captures useful information that is difficult to provide in the conventional vector format. More importantly, to effectively utilize the rich information contained in tensors, we propose a novel semi-naive Bayesian tensor classification method (which we call Bat) that builds predictive models directly on data in tensor form (instead of on their vectorizations). We apply Bat to supervised learning problems, and perform comprehensive experiments on classifying text documents and graphs, which demonstrate (1) the advantage of the tensor representation over conventional feature-vectorization approaches, and (2) the superiority of the proposed Bat tensor classifier over other existing learners.

15:20-15:40 **Prediction with Model-Based Neutrality**
Kazuto Fukuchi, Jun Sakuma and Toshihiro Kamishima

With recent developments in machine learning technology, the resulting predictions can now have a significant impact on individual lives and activities. In some cases, predictions made by machine learning can cause unfair treatments to individuals unexpectedly. For example, if the hiring decisions are highly dependent on some personal attributes, such as gender or ethnicity, such hiring might be deemed discriminatory. This paper investigates neutralization of a probabilistic model with respect to another probabilistic model, referred to as a viewpoint. We present a novel neutrality definition for probabilistic models, η neutrality, and introduce a model neutralization method. Following the definition of η neutrality, we introduce a systematic method to enforce neutrality to prediction models obtained by maximum likelihood estimation. Our methods can be applied to various machine learning algorithms, as demonstrated by η neutral logistic regression and η neutral linear regression.

Wed3A: Subgroup Discovery & Streams

Room: Euforie

16:10-16:30 **Discovering Skylines of Subgroup Sets**
Matthijs van Leeuwen and Antti Ukkonen

Many tasks in exploratory data mining aim to discover the top-k results with respect to a certain interestingness measure. Unfortunately, in practice top-k solution sets are hardly satisfactory, if only because redundancy in such results is a severe problem. To address this, a recent trend is to find diverse sets of high-quality patterns. However, a `perfect' diverse top-k cannot possibly exist, since there is an inherent trade-off between quality and diversity. We argue that the best way to deal with the quality-diversity trade-off is to explicitly consider the Pareto front, or skyline, of non-dominated solutions, i.e. those solutions for which neither quality nor diversity can be improved without degrading the other quantity. In particular, we focus on k-pattern set mining in the context of Subgroup Discovery. For this setting, we present two algorithms for the discovery of skylines: an exact algorithm and a levelwise heuristic. We evaluate the performance of the two proposed skyline algorithms, and the accuracy of the levelwise method. Furthermore, we show that the skylines can be used for the objective evaluation of subgroup set heuristics. Finally, we show characteristics of the obtained skylines, which reveal that different quality-diversity trade-offs result in clearly different subgroup sets. Hence, the discovery of skylines is an important step towards a better understanding of `diverse top-k's'.

16:30-16:50 **Difference-Based Estimates for Generalization-Aware Subgroup Discovery**
Florian Lemmerich, Martin Becker and Frank Puppe

For the task of subgroup discovery, generalization-aware interesting measures that are based not only on the statistics of the patterns itself, but also on the statistics of their generalizations have recently been shown to be essential. A key technique to increase runtime performance of subgroup discovery algorithms is the application of optimistic estimates to limit the search space size. These are upper bounds for the interestingness that any specialization of the currently evaluated pattern may have. Until now these estimates are based on the anti-monotonicity of instances, which are covered by the current pattern. This neglects important properties of generalizations. Therefore, we present in this paper a new scheme of deriving optimistic estimates for generalization aware subgroup discovery, which is based on the instances by which patterns differ in comparison to their generalizations. We show, how this technique can be applied for the most popular interestingness measures for binary as well as for numeric target concepts. The novel bounds are incorporated in an efficient algorithm, which outperforms previous methods by up to an order of magnitude.

16:50-17:10 **Fast and Exact Mining of Probabilistic Data Streams**
Reza Akbarinia and Florent Masseglia

Discovering Probabilistic Frequent Itemsets (PFI) is very challenging since algorithms designed for deterministic data are not applicable in probabilistic data. The problem is even more difficult for probabilistic data streams where massive frequent updates need to be taken into account while respecting data stream constraints. In this paper, we propose FEMP (Fast and Exact Mining of Probabilistic data streams), the first solution for exact PFI mining in data streams with sliding windows. FEMP allows updating the frequentness probability of an itemset whenever a transaction is added or removed from the observation window. Using these update operations, we are able to extract PFI in sliding windows with very low response times. Furthermore, our method is exact, meaning that we are able to discover the exact probabilistic frequentness distribution function for any monitored itemset, at any time. We implemented FEMP and conducted an extensive experimental evaluation over synthetic and real-world data sets; the results illustrate its very good performance.

17:10-17:30 **Pitfalls in benchmarking data stream classification and how to avoid them**
Albert Bifet, Jesse Read, Indre Zliobaite, Bernhard Pfahringer and Geoff Holmes

Pitfalls in benchmarking data stream classification and how to avoid them Data stream classification plays an important role in modern data analysis, where data arrives in a stream and needs to be mined in real time. In the data stream setting the underlying distribution from which this data comes may be changing andevolving, and so classifiers that can update themselves during operation are becoming the state-of-the-art. In this paper we show that data streams may have an important temporal component, which currently is not considered in the evaluationand benchmarking of data stream classifiers. We demonstrate how a naive classifier considering the temporal component only outperforms a lot of current state-of-the-art classifiers on real data streams that have temporal dependence, i.e.data is autocorrelated. We propose to evaluate data stream classifiers taking into account temporal dependence, and introduce a new evaluation measure, which provides a more accurate gauge of data stream classifier performance. In response tothe temporal dependence issue we propose a generic wrapper for data stream classifiers, which incorporates the temporal component into the attribute space.

17:30-17:50 **Adaptive Model Rules from High-Speed Data Streams**
Joao Gama, Ezilda Almeida, Carlos Ferreira and Petr Kosina

Decision rules are one of the most expressive languages for machine learning. In this paper we present Adaptive Model Rules (AMRules), the first streaming rule learning algorithm for regression problems. In AMRules the antecedent of a rule is a conjunction of conditions on the attribute values, and the consequent is a linear combination of attribute values. Each rule uses a Page-Hinkley test to detect changes in the process generating data and react to changes by pruning the rule set. In the experimental section we report the results of AMRules on benchmark regression problems, and compare the performance of our system with other streaming regression algorithms.

Wed3B: Multi-label Classification & Outlier Detection

Room: Dialog

16:10-16:30 **Multi-label Classification with Output Kernels**
Yuhong Guo and Dale Schuurmans

Although multi-label classification has become an increasingly important problem in machine learning, current approaches remain restricted to learning in the original label space (or in a simple linear projection of the original label space). Instead, we propose to use kernels on output label vectors to significantly expand the forms of label dependence that can be captured. The main challenge is to reformulate standard multi-label losses to handle kernels between output vectors. We first demonstrate how a state-of-the-art large margin loss for multi-label classification can be reformulated, exactly, to handle output kernels as well as input kernels. Importantly, the pre-image problem for multi-label classification can be easily solved at test time, while the training procedure can still be simply expressed as a quadratic program in a dual parameter space. We then develop a projected gradient descent training procedure for this new formulation. Our empirical results demonstrate the efficacy of the proposed approach on complex image labeling tasks.

16:30-16:50 **Probabilistic Clustering for Hierarchical Multi-label Classification of Protein Functions**
Rodrigo Barros, Ricardo Cerri, Alex Freitas and Andre Carvalho

Hierarchical Multi-Label Classification is a complex classification problem where the classes are hierarchically structured. This task is very common in protein function prediction, where each protein can have more than one function, which in turn can have more than one sub-function. In this paper, we propose a novel hierarchical multi-label classification algorithm for protein function prediction, namely HMC-PC. It is based on probabilistic clustering, and it makes use of cluster membership probabilities in order to generate the predicted class vector. We perform an extensive empirical analysis in which we compare our new approach to four different hierarchical multi-label classification algorithms, in protein function datasets structured both as trees and directed acyclic graphs. We show that HMC-PC achieves superior or comparable results compared to the state-of-the-art method for hierarchical multi-label classification.

16:50-17:10 **Mining Outlier Participants: Insights Using Directional Distributions in Latent Models**
Didi Surian and Sanjay Chawla

In this paper we will propose a new probabilistic topic model to score the expertise of participants on the projects that they contribute to based on their previous experience. Based on each participant's score, we rank participants and define those who have the lowest scores as outlier participants. Since the focus of our study is on outliers, we name the model as (M)ining (O)utlier (P)articipants from (P)rojects (MOPP) model. MOPP is a topic model that is based on directional distributions which are particularly suitable for outlier detection in high-dimensional spaces. Extensive experiments on both synthetic and real data sets have shown that MOPP gives better results on both topic modeling and outlier detection tasks.

17:10-17:30 **Anomaly Detection in Vertically Partitioned Data by Distributed Core Vector Machines**
Marco Stolpe, Kanishka Bhaduri, Kamalika Das and Katharina Morik

Observations of physical processes suffer from instrument malfunction and noise and demand data cleansing. However, rare events are not to be excluded from modeling, since they can be the most interesting findings. Often, sensors collect features at different sites, so that only a subset is present (vertically distributed data). Transferring all data or a sample to a single location is impossible in many real-world applications due to restricted bandwidth of communication. Finding interesting abnormalities thus requires efficient methods of distributed anomaly detection. We propose a new algorithm for anomaly detection on vertically distributed data. It aggregates the data directly at the local storage nodes using RBF kernels. Only a fraction of the data is communicated to a central node. Through extensive empirical evaluation on controlled datasets, we demonstrate that our method is an order of magnitude more communication efficient than state of the art methods, achieving a comparable accuracy.

17:30-17:50 **Local Outlier Detection with Interpretation**
Xuan-Hong Dang, Barbora Micenkova, Ira Assent and Raymond T. Ng

Outlier detection aims at searching for a small set of objects that are inconsistent or considerably deviating from other objects in a dataset. Existing research focuses on outlier identification while omitting the equally important problem of outlier interpretation. This paper presents a novel method named LODI to address both problems at the same time. In LODI, we develop an approach that explores the quadratic entropy to adaptively select a set of neighboring instances and a learning method to seek an optimal subspace in which an outlier is maximally separated from its neighbors. We show that this learning task can be optimally solved via the matrix eigen-decomposition and its solution contains all essential information to reveal which features in the original data space are most important to interpret the exceptional property of the outliers. We demonstrate the appealing performance of LODI via a number of synthetic and real world datasets and compare its outlier detection rates against state-of-the-art algorithms.

Wed3C: Ensembles

Room: Ceremonie

16:10-16:30 **Boosting for Unsupervised Domain Adaptation**
Amaury Habrard, Jean-Philippe Peyrache and Marc Sebban

To cope with machine learning problems where the learner receives data from different source and target distributions, a new learning framework named domain adaptation (DA) has emerged, opening the door for designing theoretically well-founded algorithms. In this paper, we present SLDAB, a self-labeling DA algorithm, which takes its origin from both the theory of boosting and the theory of DA. SLDAB works in the difficult unsupervised DA setting where source and target training data are available, but only the former are labeled. To deal with the absence of labeled target information, SLDAB jointly minimizes the classification error over the source domain and the proportion of margin violations over the target domain. To prevent the algorithm from inducing degenerate models, we introduce a measure of divergence whose goal is to penalize hypotheses that are not able to decrease the discrepancy between the two domains. We present a theoretical analysis of our algorithm and show practical evidences of its efficiency compared to two widely used DA approaches.

16:30-16:50 **AR-Boost: Reducing Overfitting by a Robust Data-Driven Regularization Strategy**
Baidya Nath Saha, Gautam Kunapuli, Nilanjan Ray, Joseph Maldjian and Sriraam Natarajan

We introduce a novel and robust data-driven regularization strategy called Adaptive Regularized Boosting (AR-Boost), motivated by a desire to reduce overfitting. We replace AdaBoost's hard margin with a regularized soft margin that trades-off between a larger margin, at the expense of some misclassification errors. Minimizing this regularized exponential loss results in a boosting algorithm that relaxes the weak learning assumption further: it can use classifiers with error greater than 1/2 . This enables a natural extension to multiclass boosting, and further reduces overfitting, in both the binary and multi-class cases. We derive bounds for training and generalization error, and relate them to AdaBoost. Finally, we show empirical results on benchmark data that establish the robustness of our approach, and improved performance overall.

16:50-17:10 **Parallel Boosting with Momentum**
Indraneel Mukherjee, Yoram Singer, Rafael Frongillo and Kevin Canini

We describe a new, simplified, and general analysis of a fusion of Nesterov's accelerated gradient with parallel coordinate descent. The resulting algorithm, which we call boom, for boosting with momentum, enjoys the merits of both techniques. Namely, boom retains the momentum and convergence properties of the accelerated gradient method while taking into account the curvature of the objective function. We describe an distributed implementation of boom which is suitable for massive high dimensional datasets. We show experimentally that boom is especially effective in large scale learning problems with rare yet informative features.

17:10-17:30 **Inner Ensembles: Using Ensemble Methods in the Learning Phase**
Houman Abbasian, Chris Drummond, Nathalie Japkowicz and Stan Matwin

Ensemble Methods represent an important research area within machine learning. Here, we argue that the use of such methods can be generalized and applied in many more situations than they have been previously. Instead of using them only to combine the output of an algorithm, we can apply them to the decisions made inside the learning algorithm, itself. We call this approach Inner Ensembles. The main contribution of this work is to demonstrate how broadly this idea can applied. Specifically, we show that the idea can be applied to different classes of learner such as Bayesian networks and K-means clustering.

17:30-17:50 **Mixtures of Large Margin Nearest Neighbor Classifiers**
Murat Semerci and Ethem Alpaydin

The accuracy of the k-nearest neighbor algorithm depends on the distance function used to measure similarity between instances. Methods have been proposed in the literature to learn a good distance function from a labelled training set. One such method is the large margin nearest neighbor classifier that learns a global Mahalanobis distance. We propose a mixture of such classifiers where a gating function divides the input space into regions and a separate distance function is learned in each region in a lower dimensional manifold. We show that such an extension improves accuracy and allows visualization.

Wed3D: Bayesian Learning

Room: Aplaus

16:10-16:35 **A Comparative Evaluation of Stochastic-Based Inference Methods for Gaussian Process Models**
Maurizio Filippone, Mingjun Zhong and Mark Girolami

Gaussian Process models are extensively used in data analysis given their flexible modeling capabilities and interpretability. The fully Bayesian treatment of GP models is analytically intractable, and therefore it is necessary to resort to either deterministic or stochastic approximations. This paper focuses on stochastic-based inference for Gaussian Process models. First, challenges associated with the fully Bayesian treatment of GP models are discussed, and then a number of inference strategies based on Markov chain Monte Carlo methods are presented and rigorously assessed. In particular, strategies based on efficient parametrizations and efficient proposal mechanisms are extensively compared based on speed of convergence, sampling efficiency, and computational cost.

16:35-16:55 **Decision-Theoretic Sparsification for Gaussian Process Preference Learning**
Ehsan Abbasnejad, Edwin Bonilla and Scott Sanner

We propose a decision-theoretic sparsification method for Gaussian process preference learning. This method overcomes the loss-insensitive nature of popular sparsification approaches such as the Informative Vector Machine (IVM). Instead of selecting a subset of users and items as inducing points based on uncertainty-reduction principles, our sparsification approach is underpinned by decision theory and directly incorporates the {loss} function inherent to the underlying preference learning problem. We show that by selecting different specifications of the loss function, the IVM's differential entropy criterion, a value of information criterion, and an upper confidence bound (UCB) criterion used in the bandit setting can all be recovered from our decision-theoretic framework. We refer to our method as the Valuable Vector Machine (VVM) as it selects the most useful items during sparsification to minimize the corresponding loss. We evaluate our approach on one synthetic and two real-world preference datasets, including one generated via Amazon Mechanical Turk and another collected from Facebook. Experiments show that variants of the VVM significantly outperform the IVM on all datasets under similar computational constraints.

16:55-17:15 **Variational Hidden Conditional Random Fields with Coupled Dirichlet Process Mixtures**
Konstantinos Bousmalis, Stefanos Zafeiriou, Louis-Philippe Morency, Maja Pantic and Zoubin Ghahramani

Hidden Conditional Random Fields (HCRFs) are discriminative latent variable models which have been shown to successfully learn the hidden structure of a given classification problem. An infinite HCRF is an HCRF with a countably infinite number of hidden states, which rids us not only of the necessity to specify a priori a fixed number of hidden states available but also of the problem of overfitting. Markov chain Monte Carlo (MCMC) sampling algorithms are often employed for inference in such models. However, convergence of such algorithms is rather difficult to verify, and as the complexity of the task at hand increases, the computational cost of such algorithms often becomes prohibitive. These limitations can be overcome by variational techniques. In this paper, we present a generalized framework for infinite HCRF models, and a novel variational inference approach on a model based on coupled Dirichlet Process Mixtures, the HCRF-DPM. We show that the variational HCRF-DPM is able to converge to a correct number of represented hidden states, and performs as well as the best parametric HCRFs-chosen via cross-validation-for the difficult tasks of recognizing instances of agreement, disagreement, and pain in audiovisual sequences.

17:15-17:35 **Sparsity in Bayesian Blind Source Separation and Deconvolution**
Vaclav Smidl and Ondrej Tichy

Blind source separation algorithms are based on various separation criteria. Differences in convolution kernels of the sources are common assumptions in audio and image processing. Since it is still an ill posed problem, any additional information is beneficial. In this contribution, we investigate the use of sparsity criteria for both the source signal and the convolution kernels. A probabilistic model of the problem is introduced and its Variational Bayesian solution derived. The sparsity of the solution is achieved by introduction of unknown variance of the prior on all elements of the convolution kernel and the mixing matrix. Properties of the model are analyzed on simulated data and compared with state of the art methods. Performance of the algorithm is demonstrated on the problem of decomposition of a sequence of medical data. Specifically, the assumption of sparseness is shown to suppress artifacts of unconstrained separation method.

THURSDAY 26 SEPTEMBER 2013

THURSDAY INVITED TALK

Learning with Humans in the Loop

Speaker: Thorsten Joachims
Time: 09:30-10:30
Room: plenary

Abstract
Machine Learning is increasingly becoming a technology that directly interacts with human users. Search engines, recommender systems, and electronic commerce already heavily rely on adapting the user experience through machine learning, and other applications are likely to follow in the near future (e.g., autonomous robotics, smart homes, gaming). In this talk, I argue that learning with humans in the loop requires learning algorithms that explicitly account for human behavior, their motivations, and their judgment of performance. Towards this goal, the talk explores how integrating microeconomic models of human behavior into the learning process leads to new learning models that no longer reduce the user to a "labeling subroutine". This motivates an interesting area for theoretical, algorithmic, and applied machine learning research with connections to rational choice theory, econometrics, and behavioral economics.

Bio
Thorsten Joachims is a Professor of Computer Science at Cornell University. His research interests center on a synthesis of theory and system building in machine learning, with applications in language technology, information retrieval, and recommendation. His past research focused on support vector machines, text classification, structured output prediction, convex optimization, learning to rank, learning with preferences, and learning from implicit feedback. In 2001, he finished his dissertation advised by Prof. Katharina Morik at the University of Dortmund. From there he also received his Diplom in Computer Science in 1997. Between 2000 and 2001 he worked as a PostDoc at the GMD Institute for Autonomous Intelligent Systems. From 1994 to 1996 he was a visiting scholar with Prof. Tom Mitchell at Carnegie Mellon University.

THURSDAY SESSIONS AT A GLANCE

Thu1A: Industrial track (1)

Room: Euforie

11:00-11:42 **Some of the Problems and Applications of Opinion Analysis**
Hugo Zaragoza

11:42-12:25 **Machine Learning in a large diversified Internet Group**
Jean-Paul Schmetz

Thu1B: Sequential Pattern Mining

Room: Dialog

11:00-11:20 **Itemset Based Sequence Classification**
Cheng Zhou, Boris Cule and Bart Goethals

11:20-11:40 **A Relevance Criterion for Sequential Patterns**
Henrik Grosskreutz, Bastian Lang and Daniel Trabold

11:40-12:00 **A Fast and Simple Method for Mining Subsequences with Surprising Event Counts**
Jefrey Lijffijt

12:00-12:20 **Relevant Subsequence Detection with Sparse Dictionary Learning**
Sam Blasiak, Huzefa Rangwala and Kathryn Laskey

Thu1C: Graphical Models

Room: Ceremonie

11:00-11:25 **Spatio-Temporal Random Fields: Compressible Representation and Distributed Estimation**
Nico Piatkowski, Sangkyun Lee and Katharina Morik

11:25-11:45 **Knowledge Intensive Learning: Combining Qualitative Constraints with Causal Independence for Parameter Learning in Probabilistic Models**
Shuo Yang and Sriraam Natarajan

11:45-12:05 **Direct Learning of Sparse Changes in Markov Networks by Density Ratio Estimation**
Song Liu, John Quinn, Michael Gutmann and Masashi Sugiyama

12:05-12:25 **Greedy Part-Wise Learning of Sum-Product Networks**
Robert Peharz, Bernhard Geiger and Franz Pernkopf

Thu1D: Unsupervised Learning

Room: Aplaus

11:00-11:25 **A Framework for Semi-Supervised and Unsupervised Optimal Extraction of Clusters from Hierarchies**
Ricardo J.G.B. Campello, Davoud Moulavi, Arthur Zimek and Jorg Sander

11:25-11:45 **Minimal Shrinkage for Noisy Data Recovery**
Deguang Kong and Chris Ding

11:45-12:05 **Reduced-Rank Local Distance Metric Learning**
Yinjie Huang, Cong Li, Michael Georgiopoulos and Georgios Anagnostopoulos

12:05-12:25 **Cross-Domain Recommendation via Cluster-Level Latent Factor Model**
Sheng Gao

Thu2A: Industrial track (2)

Room: Euforie

14:15-14:57 **Bayesian Learning in Online Service: Statistics Meets Systems**
Ralf Herbrich

14:57-15:40 **ML and Business: A Love-Hate Relationship**
Andreas Antrup

Thu2B: Dynamic Graphs

Room: Dialog

14:15-14:35 **Incremental Local Evolutionary Outlier Detection for Dynamic Social Networks**
Tengfei Ji, Jun Gao and Dongqing Yang

14:35-14:55 **Continuous Similarity Computation over Streaming Graphs**
Elena Valari and Apostolos Papadopoulos

14:55-15:15 **Trend Mining in Dynamic Attributed Graphs**
Elise Desmier, Marc Plantevit, Celine Robardet and Jean-Francois Boulicaut

15:15-15:35 **How Long Will She Call Me? Distribution, Social Theory and Duration Prediction**
Yuxiao Dong, Jie Tang, Tiancheng Lou, Bin Wu and Nitesh Chawla

Thu2C: Statistical Learning (1)

Room: Ceremonie

14:15-14:40 **Block Coordinate Descent Algorithms for Large-Scale Sparse Multiclass Classification**
Mathieu Blondel, Kazuhiro Seki and Kuniaki Uehara

14:40-15:00 **MORD: Multi-class Classifier for Ordinal Regression**
Konstiantyn Antoniuk, Vojtech Franc and Vaclav HIavac

15:00-15:20 **Identifiability of Model Properties in Over-Parameterized Model Classes**
Manfred Jaeger

15:20-15:40 **Multi-core Structural SVM Training**
Kai-Wei Chang, Vivek srikumar and Dan Roth

Thu2D: Evaluation & kNN

Room: Aplaus

14:15-14:40 **ROC Curves in Cost Space**
Cesar Ferri, Jose Hernandez-Orallo and Peter Flach

14:40-15:00 **Area Under the Precision-Recall Curve: Point Estimates and Confidence Intervals**
Kendrick Boyd, Kevin Eng and David Page

15:00-15:20 **Hub Co-occurrence Modeling for Robust High-Dimensional kNN Classification**
Nenad Tomasev and Dunja Mladenic

15:20-15:40 **Fast kNN Graph Construction with Locality Sensitive Hashing**
Yan-Ming Zhang, Kaizhu Huang, Guanggang Geng and Cheng-Lin Liu

Thu3A: Sequence & Time Series Analysis

Room: Euforie

16:10-16:35 **Fast Sequence Segmentation Using Log-Linear Models**
Nikolaj Tatti

16:35-16:55 **Explaining Interval Sequences by Randomization**
Andreas Henelius, Jussi Korpela and Kai Puolamaki

16:55-17:15 **A Layered Dirichlet Process for Hierarchical Segmentation of Sequential Grouped Data**
Adway Mitra, Ranganath B.N. and Indrajit Bhattacharya

17:15-17:35 **Fault Tolerant Regression for Sensor Data**
Indre Zliobaite and Jaakko Hollmen

Thu3B: Declarative Data Mining & Meta Learning

Room: Dialog

16:10-16:35 **Pairwise Meta-rules for Better Meta-learning-Based Algorithm Ranking**
Quan Sun and Bernhard Pfahringer

16:35-16:55 **SNNAP: Solver-Based Nearest Neighbor for Algorithm Portfolios**
Marco Collautti, Yuri Malitsky and Barry O'Sullivan

16:55-17:15 **Top-k Frequent Closed Itemset Mining Using Top-k SAT Problem**
Said Jabbour, Lakhdar Sais and Yakoub Salhi

17:15-17:35 **A declarative framework for Constrained Clustering**
Thi-Bich-Hanh Dao, Khanh-Chuong Duong and Christel Vrain

Thu3C: Topic Models

Room: Ceremonie

16:10-16:35 **Probabilistic Topic Models for Sequence Data**
Nicola Barbieri, Antonio Bevacqua, Marco Carnuccio, Giuseppe Manco and Ettore Ritacco

16:35-16:55 **Nested Hierarchical Dirichlet Process for Nonparametric Entity-Topic Analysis**
Priyanka Agrawal, Lavanya Tekumalla and Indrajit Bhattacharya

16:55-17:15 **From Topic Models to Semi-supervised Learning: Biasing Mixed-Membership Models to Exploit Topic-Indicative Features in Entity Clustering**
Ramnath Balasubramanyan, William Cohen and Bhavana Dalvi

17:15-17:35 **Sparse Relational Topic Models for Document Networks**
Aonan Zhang, Jun Zhu and Bo Zhang

Thu3D: Statistical Learning (2)

Room: Aplaus

16:10-16:35 **The Flip-the-State Transition Operator for Restricted Boltzmann Machines**
Kai Brugge, Asja Fischer and Christian Igel

16:35-16:55 **Learning Discriminative Sufficient Statistics Score Space**
Xiong Li, Bin Wang, Yuncai Liu and Tai Sing Lee

16:55-17:15 **The Stochastic Gradient Descent for the Primal L1-SVM Optimization Revisited**
Constantinos Panagiotakopoulos and Petroula Tsampouka

17:15-17:35 **Bundle CDN: A Highly Parallelized Approach for Large-Scale L1-regularized Logistic**
Yatao Bian, Xiong Li, mingqi Cao and Yuncai Liu

THURSDAY SESSIONS, WITH ABSTRACTS

Thu1A: Industrial track (1)

Room: Euforie

11:00-11:42 **Some of the Problems and Applications of Opinion Analysis**
Hugo Zaragoza

Websays strives to provide the best possible analysis of online conversation to marketing and social media analysts. One of the obsessions of Websays is to provide "near-man-made" data quality at marginal costs. I will discuss how we approach this problem using innovative machine learning and UI approaches.

11:42-12:25 **Machine Learning in a large diversified Internet Group**
Jean-Paul Schmetz

The talk covers a wide survey of the use of machine learning techniques across a large number of subsidiaries (40+) of an Internet group (Burda Digital) with special attention to issues regarding (1) personnel training in state of the art techniques, (2) management buy-in of complex non interpretable results and (3) practical and measurable bottom line results/solutions.

Thu1B: Sequential Pattern Mining

Room: Dialog

11:00-11:20 **Itemset Based Sequence Classification**
Cheng Zhou, Boris Cule and Bart Goethals

Sequence classification is an important task in data mining. We address the problem of sequence classification using rules composed of interesting itemsets found in a dataset of labelled sequences and accompanying class labels. We measure the interestingness of an itemset in a given class of sequences by combining the cohesion and the support of the itemset. We use the discovered itemsets to generate confident classification rules, and present two different ways of building a classifier. The first classifier is based on the CBA (Classification based on associations) method, but we use a new ranking strategy for the generated rules, achieving better results. The second classifier ranks the rules by first measuring their value specific to the new data object. Experimental results show that our classifiers outperform existing comparable classifiers in terms of accuracy and stability, while maintaining a computational advantage over sequential pattern based classification.

11:20-11:40 **A Relevance Criterion for Sequential Patterns**
Henrik Grosskreutz, Bastian Lang and Daniel Trabold

The theory of relevance is an approach for redundancy avoidance in labeled itemset mining. In this paper, we adapt this theory to the setting of sequential patterns. While in the itemset setting it is suggestive to use the closed patterns as representatives for the relevant patterns, we argue that due to the different properties of the space of sequential patterns, it is preferable to use the minimal generator sequences as representatives, instead of the closed sequences. Thereafter, we show that we can efficiently compute the relevant sequences via the minimal generators in the negatives. Unlike existing iterative or post-processing approaches for pattern subset selection, our approach thus results both in a reduction of the set of patterns and in a reduction of the search space – and hence in lower computational costs.

11:40-12:00 **A Fast and Simple Method for Mining Subsequences with Surprising Event Counts**
Jefrey Lijffijt

We consider the problem of mining subsequences with surprising event counts. When mining patterns, we often test a very large number of potentially present patterns, leading to a high likelihood of finding spurious results. Typically, this problem grows as the size of the data increases. Existing methods for statistical testing are not usable for mining patterns in big data, because they are either computationally too demanding, or fail to take into account the dependency structure between patterns, leading to true findings going unnoticed. We propose a new method to compute the significance of event frequencies in subsequences of a long data sequence. The method is based on analyzing the joint distribution of the patterns, omitting the need for randomization. We argue that computing the p-values exactly is computationally costly, but that an upper bound is easy to compute. We investigate the tightness of the upper bound and compare the power of the test with the alternative of post-hoc correction. We demonstrate the utility of the method on two types of data: text and DNA. We show that the proposed method is easy to implement and can be computed quickly. Moreover, we conclude that the upper bound is sufficiently tight and that meaningful results can be obtained in practice.

12:00-12:20 **Relevant Subsequence Detection with Sparse Dictionary Learning**
Sam Blasiak, Huzefa Rangwala and Kathryn Laskey

Sparse Dictionary Learning has recently become popular for discovering latent components that can be used to reconstruct elements in a dataset. Analysis of sequence data could also benefit from this type of decomposition, but sequence datasets are not natively accepted by the Sparse Dictionary Learning model. A strategy for making sequence data more manageable is to extract all subsequences of a fixed length from the original sequence dataset. This subsequence representation can then be input to a Sparse Dictionary Learner. This strategy can be problematic because self-similar patterns within sequences are over-represented. In this work, we propose an alternative for applying Sparse Dictionary Learning to sequence datasets. We call this alternative Relevant Subsequence Dictionary Learning (RS-DL). Our method involves constructing separate dictionaries for each sequence in a dataset from shared sets of relevant subsequence patterns. Through experiments, we show that decompositions of sequence data induced by our RS-DL model can be effective both for discovering repeated patterns meaningful to humans and for extracting features useful for sequence classification.

Thu1C: Graphical Models

Room: Ceremonie

11:00-11:25 **Spatio-Temporal Random Fields: Compressible Representation and Distributed Estimation**
Nico Piatkowski, Sangkyun Lee and Katharina Morik

Modern sensing technology allows us enhanced monitoring of dynamic activities in business, traffic, and home, just to name a few. The increasing amount of sensor measurements, however, brings us the challenge for efficient data analysis. This is especially true when sensing targets can interoperate - in such cases we need learning models that can capture the relations of sensors, possibly without collecting or exchanging all data. Generative graphical models namely the Markov random fields (MRFs) fit this purpose, which can represent complex spatial and temporal relations among sensors, producing interpretable answers in terms of probability. The only drawback will be the cost for inference, storing and optimizing a very large number of parameters - not uncommon when we apply them for real-world applications. In this paper, we investigate how we can make discrete probabilistic graphical models practical for predicting sensor states in a spatio-temporal setting. A set of new ideas allows keeping the advantages of such models while achieving scalability. We first introduce a novel alternative to represent model parameters, which enables us to compress the parameter storage by removing uninformative parameters in a systematic way. For finding the best parameters via maximal likelihood estimation, we provide a separable optimization algorithm that can be performed independently in parallel in each graph node. We illustrate that the prediction quality of our suggested methods is comparable to those of the standard MRFs and a spatio-temporal k-nearest neighbor method, while using much less computational resources.

11:25-11:45 **Knowledge Intensive Learning: Combining Qualitative Constraints with Causal Independence for Parameter Learning in Probabilistic Models**
Shuo Yang and Sriraam Natarajan

In Bayesian networks, prior knowledge has been used in the form of causal independencies between random variables or employing qualitative constraints such as monotonicities. In this work, we extend and combine the two different ways of providing domain knowledge. We derive an algorithm based on gradient descent for estimating the parameters of a Bayesian network in the presence of causal independencies in the form of Noisy-Or and qualitative constraints such as monotonicities and synergies. Noisy-Or structure can reduce the data requirements by separating the influence of each parent thereby reducing greatly the number of parameters. Qualitative constraints on the other hand, allow for imposing constraints on the parameter space making it possible to learn more accurate parameters from a very small number of data points. Our exhaustive empirical validation conclusively proves that the synergy constrained Noisy-OR leads to more accurate models in the presence of smaller amount of data.

11:45-12:05 **Direct Learning of Sparse Changes in Markov Networks by Density Ratio Estimation**
Song Liu, John Quinn, Michael Gutmann and Masashi Sugiyama

We propose a new method for detecting changes in Markov network structure between two sets of samples. Instead of naively fitting two Markov network models separately to the two data sets and figuring out their difference, we directly learn the network structure change by estimating the ratio of Markov network models. This density-ratio formulation naturally allows us to introduce sparsity in the network structure change, which highly contributes to enhancing interpretability. Furthermore, computation of the normalization term, which is a critical computational bottleneck of the naive approach, can be remarkably mitigated. Through experiments on gene expression and Twitter data analysis, we demonstrate the usefulness of our method.

12:05-12:25 **Greedy Part-Wise Learning of Sum-Product Networks**
Robert Peharz, Bernhard Geiger and Franz Pernkopf

Sum-product networks allow to model complex variable interactions while still granting efficient inference. However, most learning algorithms proposed so far are explicitly or implicitly restricted to the image domain, either by assuming variable neighborhood or by assuming that dependent variables are related by their magnitudes over the training set. In this paper, we introduce a novel algorithm, learning the structure and parameters of sum-product networks in a greedy bottom-up manner. Our algorithm iteratively merges probabilistic models of small variable scope to larger and more complex models. These merges are guided by statistical dependence test, and parameters are learned using a maximum mutual information principle. In experiments our method competes well with the existing learning algorithms for sum-product networks on the task of reconstructing covered image regions, and outperforms these when neither neighborhood nor correlations by magnitude can be assumed.

Thu1D: Unsupervised Learning

Room: Aplaus

11:00-11:25 **A Framework for Semi-Supervised and Unsupervised Optimal Extraction of Clusters from Hierarchies**
Ricardo J.G.B. Campello, Davoud Moulavi, Arthur Zimek and Jorg Sander

We introduce a framework for the optimal extraction of flat clusterings from local cuts through cluster hierarchies. The extraction of a flat clustering from a cluster tree is formulated as an optimization problem and a linear complexity algorithm is presented that provides the globally optimal solution to this problem in semi-supervised as well as in unsupervised scenarios. A collection of experiments is presented involving clustering hierarchies of different natures, a variety of real data sets, and comparisons with specialized methods from the literature.

11:25-11:45 **Minimal Shrinkage for Noisy Data Recovery**
Deguang Kong and Chris Ding

Noisy data recovery is an important problem in machine learning field, which has widely applications for collaborative prediction, recommendation systems, etc. One popular model is to use trace norm model for noisy data recovery. However, it is ignored that the reconstructed data could be shrank (i.e., singular values could be greatly suppressed). In this paper, we present novel noisy data recovery models, which replaces the standard rank constraint (i.e., trace norm). The proposed model is attractive due to its suppression on the shrinkage of singular values at smaller parameter p. We analyze the optimal solution of proposed models, and characterize the rank of optimal solution. Efficient algorithms are presented, the convergences of which are rigorously proved. Extensive experiment results on 6 noisy datasets demonstrate the good performance of proposed minimum shrinkage models.

11:45-12:05 **Reduced-Rank Local Distance Metric Learning**
Yinjie Huang, Cong Li, Michael Georgiopoulos and Georgios Anagnostopoulos

We propose a new method for local metric learning based on a conical combination of Mahalanobis metrics and pair-wise similarities between the data. Its formulation allows for controlling the rank of the metrics' weight matrices. We also offer a convergent algorithm for training the associated model. Experimental results on a collection of classification problems imply that the new method may offer notable performance advantages over alternative metric learning approaches that have recently appeared in the literature.

12:05-12:25 **Cross-Domain Recommendation via Cluster-Level Latent Factor Model**
Sheng Gao

Recommender systems always aim to provide recommendations for a user based on historical ratings collected from a single domain (e.g., movies or books) only, which may suffer the data sparsity problem. Recently, several recommendation models have been proposed to transfer knowledge by pooling together the rating data from multiple domains to alleviate the sparsity problem, which typically assume that multiple domains share a latent common rating pattern based on the user-item co-clustering. In practice, however, the related domains do not necessarily share such a common rating pattern, and diversity among the related domains might outweigh the advantages of such common pattern, which may result in performance degradations. In this paper, we propose a novel cluster-level based latent factor model to enhance the cross-domain recommendation, which can not only learn the common rating pattern shared across domains with the flexibility in controlling the optimal level of sharing, but also learn the domain-specific rating patterns of users in each domain that involve the discriminative information propitious to performance improvement. To this end, the proposed model is formulated as an optimization problem based on joint nonnegative matrix tri-factorization and an efficient alternating minimization algorithm is developed with convergence guarantee. Extensive experiments on several real world datasets suggest that our proposed model outperforms the state-of-the-art methods for the cross-domain recommendation task.

Thu2A: Industrial track (2)

Room: Euforie

14:15-14:57 **Bayesian Learning in Online Service: Statistics Meets Systems**
Ralf Herbrich

Over the past few years, we have entered the world of big and structured data-a trend largely driven by the exponential growth of Internet-based online services such as Search, eCommerce and Social Networking as well as the ubiquity of smart devices with sensors in everyday life. This poses new challenges for statistical inference and decision-making as some of the basic assumptions are shifting:
The ability to optimize both the likelihood and loss functions
The ability to store the parameters of (data) models
The level of granularity and 'building blocks' in the data modeling phase
The interplay of computation, storage, communication and inference and decision-making techniques
In this talk, I will discuss the implications of big and structured data for Statistics and the convergence of statistical model and distributed systems. I will present one of the most versatile modeling techniques that combines systems and statistical properties-factor graphs-and review a series of approximate inference techniques such as distributed message passing. The talk will be concluded with an overview of real-world problems at Amazon.

14:57-15:40 **ML and Business: A Love-Hate Relationship**
Andreas Antrup

Based on real world examples. the talk explores common gaps in the mutual understanding of the business and the analytical side; particular focus shall be on misconceptions of the needs and expectations of business people and the resulting problems. It also touches on some approaches to bridge these gaps and build trust. At the end we shall discuss possibly under-researched areas that may open the doors to a yet wider usage of ML principles and thus unlock more of its value and beauty.

Thu2B: Dynamic Graphs

Room: Dialog

14:15-14:35 **Incremental Local Evolutionary Outlier Detection for Dynamic Social Networks**
Tengfei Ji, Jun Gao and Dongqing Yang

Numerous applications in dynamic social networks, ranging from telecommunications to financial transactions, create evolving datasets. Detecting outliers in such dynamic networks is inherently challenging, because the arbitrary linkage structure with massive information is changing over time. Little research has been done on detecting outliers for dynamic social networks, even then, they represent networks as un-weighted graphs and identify outliers from a relatively global perspective. Thus, existing approaches fail to identify the objects with abnormal evolutionary behavior only with respect to their local neighborhood. We define such objects as local evolutionary outliers, LEOutliers. This paper proposes a novel incremental algorithm IcLEOD to detect LEOutliers in weighted graphs. By focusing only on the time-varying components (e.g., node, edge and edge weight), IcLEOD algorithm is highly efficient in large and gradually evolving networks. Experimental results on both real and synthetic datasets illustrate that our approach of finding local evolutionary outliers can be practical.

14:35-14:55 **Continuous Similarity Computation over Streaming Graphs**
Elena Valari and Apostolos Papadopoulos

Large network analysis is a very important topic in data mining. A significant body of work in the area studies the problem of node similarity. One way to express node similarity is to associate with each node the set of 1-hop neighbors and compute the Jaccard similarity between these sets. This information can be used subsequently for more complex operations like link prediction, clustering or dense subgraph discovery. In this work, we study algorithms to monitor the result of a similarity join between nodes continuously, assuming a sliding window accommodating graph edges. Since the arrival of a new edge or the expiration of an existing one may change the similarity between several node pairs, the challenge is to maintain the similarity join result as efficiently as possible. Our theoretical study is validated by a thorough experimental evaluation, based on real-world as well as synthetically generated graphs, demonstrating the superiority of the proposed technique in comparison to baseline approaches.

14:55-15:15 **Trend Mining in Dynamic Attributed Graphs**
Elise Desmier, Marc Plantevit, Celine Robardet and Jean-Francois Boulicaut

Many applications see huge demands of discovering important patterns in dynamic attributed graph. In this paper, we introduce the problem of discovering trend sub-graphs in dynamic attributed graphs. This new kind of pattern relies on the graph structure and the temporal evolution of the attribute values. Several interestingness measures are introduced to focus on the most relevant patterns with regard to the graph structure, the vertex attributes, and the time. We design an efficient algorithm that benefits from various constraint properties and provide an extensive empirical study from several real-world dynamic attributed graphs.

15:15-15:35 **How Long Will She Call Me? Distribution, Social Theory and Duration Prediction**
Yuxiao Dong, Jie Tang, Tiancheng Lou, Bin Wu and Nitesh Chawla

Call duration analysis is a key issue for understanding underlying patterns of (mobile) phone users. In this paper, we study to which extent the duration of a call between users can be predicted in a dynamic mobile network. We have collected a mobile phone call data from a mobile operating company, which results in a network of 272,345 users and 3.9 million call records during two months. We first examine the dynamic distribution properties of the mobile network including periodicity and demographics. Then we study several important social theories in the call network including strong/weak ties, link homophily, opinion leader and social balance. The study reveals several interesting phenomena such as people with strong ties tend to make shorter calls and young females tend to make long calls, in particular in the evening. Finally, we present a time-dependent factor graph model to model and infer the call duration between users, by incorporating our observations in the distribution analysis and the social theory analysis. Experiments show that the presented model can achieve much better predictive performance compared to several baseline methods. Our study offers evidences for social theories and also unveils several different patterns in the call network from online social networks.

Thu2C: Statistical Learning (1)

Room: Ceremonie

14:15-14:40 **Block Coordinate Descent Algorithms for Large-Scale Sparse Multiclass Classification**
Mathieu Blondel, Kazuhiro Seki and Kuniaki Uehara

Over the past decade, L1 regularization has emerged as a powerful way to learn classifiers with implicit feature selection. More recently, mixed-norm (e.g., L1/L2) regularization has been utilized as a way to select entire groups of features. In this paper, we propose a novel direct multiclass formulation specifically designed for large-scale and high-dimensional problems such as document classification. Based on a multiclass extension of the squared hinge loss, our formulation employs L1/L2 regularization so as to force weights corresponding to the same features to be zero across all classes, resulting in compact and fast-to-evaluate multiclass models. For optimization, we employ two globally-convergent variants of block coordinate descent, one with line search (Tseng and Yun, 2009) and the other without (Richtarik and Takac, 2012). We present the two variants in a unified manner and develop the core components needed to efficiently solve our formulation. The end result is a couple of block coordinate descent algorithms specifically tailored to our multiclass formulation. Experimentally, we show that block coordinate descent performs favorably to other solvers such as FOBOS, FISTA and SpaRSA. Furthermore, we show that our formulation obtains very compact multiclass models and outperforms L1/L2-regularized multiclass logistic regression in terms of training speed, while achieving comparable test accuracy.

14:40-15:00 **MORD: Multi-class Classifier for Ordinal Regression**
Konstiantyn Antoniuk, Vojtech Franc and Vaclav Hlavac

We show that classification rules used in ordinal regression are equivalent to a certain class of linear multi-class classifiers. This observation not only allows to design new learning algorithms for ordinal regression using existing methods for multi-class classification but it also allows to derive new models for ordinal regression. For example, one can convert learning of ordinal classifier with (almost) arbitrary loss function to a convex unconstrained risk minimization problem for which many efficient solvers exist. The established equivalence also allows to increase discriminative power of the ordinal classifier without need to use kernels by introducing a piece-wise ordinal classifier. We demonstrate advantages of the proposed models on standard benchmarks as well as in solving a real-life problem. In particular, we show that the proposed piece-wise ordinal classifier applied to visual age estimation outperforms other standard prediction models.

15:00-15:20 **Identifiability of Model Properties in Over-Parameterized Model Classes**
Manfred Jaeger

Classical learning theory is based on a tight linkage between hypothesis space (a class of function on a domain X), data space (function-value examples $(x, f(x))$), and the space of queries for the learned model (predicting function values for new examples x). However, in many learning scenarios the 3-way association between hypotheses, data, and queries can really be much looser. Model classes can be over-parameterized, i.e., different hypotheses may be equivalent with respect to the data observations. Queries may relate to model properties that do not directly correspond to the observations in the data. In this paper we make some initial steps to extend and adapt basic concepts of computational learnability and statistical identifiability to provide a foundation for investigating learnability in such broader contexts. We exemplify the use of the framework in three different applications: the identification of temporal logic properties of probabilistic automata learned from sequence data, the identification of causal dependencies in probabilistic graphical models, and the transfer of probabilistic relational models to new domains.

15:20-15:40 **Multi-core Structural SVM Training**
Kai-Wei Chang, Vivek srikumar and Dan Roth

Many problems in natural language processing and computer vision can be framed as structured prediction problems. Structural support vector machines (SVM) is a popular approach for training structured predictors, where learning is framed as an optimization problem. Most structural SVM solvers alternate between a model update phase and an inference phase (which predicts structures for all training examples). As structures become more complex, inference becomes a bottleneck and thus slows down learning considerably. In this paper, we propose a new learning algorithm for structural SVMs called DEMI-DCD that extends the dual coordinate descent approach by decoupling the model update and inference phases into different threads. We take advantage of multi-core hardware to parallelize learning with minimal synchronization between the model update and the inference phases. We prove that our algorithm not only converges but also fully utilizes all available processors to speed up learning, and validate our approach on two real-world NLP problems: part-of-speech tagging and relation extraction. In both cases, we show that our algorithm utilizes all available processors to speed up learning and achieves competitive performance. For example, it achieves a relative duality gap of 1% on a POS tagging problem in 192 seconds using 16 threads, while a standard implementation of a multi-threaded dual coordinate descent algorithm with the same number of threads requires more than 600 seconds to reach a solution of the same quality.

Thu2D: Evaluation & kNN

Room: Aplaus

14:15-14:40 **ROC Curves in Cost Space**
Cesar Ferri, Jose Hernandez-Orallo and Peter Flach

ROC curves and cost curves are two popular ways of visualising classifier performance, finding appropriate thresholds according to the operating condition, and deriving useful aggregated measures such as the area under the ROC curve (AUC) or the area under the optimal cost curve. In this paper we present new findings and connections between ROC space and cost space. In particular, we show that ROC curves can be transferred to cost space by means of a very natural threshold choice method, which sets the decision threshold such that the proportion of positive predictions equals the operating condition. We call these new curves rate-driven curves, and we demonstrate that the expected loss as measured by the area under these curves is linearly related to AUC. We show that the rate-driven curves are the genuine equivalent of ROC curves in cost space, establishing a point-point rather than a point-line correspondence. Furthermore, a decomposition of the rate-driven curves is introduced which separates the loss due to the threshold choice method from the ranking loss (Kendall tau distance). We also derive the corresponding curve to the ROC convex hull in cost space: this curve is different from the lower envelope of the cost lines, as the latter assumes only optimal thresholds are chosen.

14:40-15:00 **Area Under the Precision-Recall Curve: Point Estimates and Confidence Intervals**
Kendrick Boyd, Kevin Eng and David Page

The area under the precision-recall curve (AUCPR) is a single number summary of the information in the precision-recall (PR) curve. Similar to the receiver operating characteristic curve, the PR curve has its own unique properties that make estimating its enclosed area challenging. Besides a point estimate of the area, an interval estimate is often required to express magnitude and uncertainty. In this paper we perform a computational analysis of common AUCPR estimators and their confidence intervals. We find both satisfactory estimates and invalid procedures and we recommend two simple intervals that are robust to a variety of assumptions.

15:00-15:20 **Hub Co-occurrence Modeling for Robust High-Dimensional kNN Classification**
Nenad Tomasev and Dunja Mladenic

The emergence of hubs in k-nearest neighbor (kNN) topologies of intrinsically high-dimensional data has recently been shown to be quite detrimental to many standard machine learning tasks, including classification. Robust hubness-aware learning methods are required in order to overcome the impact of the highly uneven distribution of influence. In this paper, we have adapted the Hidden Naive Bayes (HNB) model to the problem of modeling neighbor occurrences and co-occurrences in high-dimensional data. Hidden nodes are used to aggregate all pairwise occurrence dependencies. The result is a novel kNN classification method tailored specifically for intrinsically high-dimensional data, the Augmented Naive Hubness Bayesian k-nearest Neighbor (ANHBNN). Neighbor co-occurrence information forms an important part of the model and our analysis reveals some surprising results regarding the influence of hubness on the shape of the co-occurrence distribution in high-dimensional data. The proposed approach was tested in the context of object recognition from images in class imbalanced data and the results show that it offers clear benefits when compared to the other hubness-aware kNN baselines.

15:20-15:40 **Fast kNN Graph Construction with Locality Sensitive Hashing**
Yan-Ming Zhang, Kaizhu Huang, Guanggang Geng and Cheng-Lin Liu

The k nearest neighbors (kNN) graph, perhaps the most popular graph in machine learning, plays an essential role for graph-based learning methods. Despite its many elegant properties, the brute force kNN graph construction method has computational complexity of $O(n^2)$, which is prohibitive for large scale data sets. In this paper, based on the divide-and-conquer strategy, we propose an efficient algorithm for approximating kNN graphs, which has the time complexity of $O(l(d+\log n)n)$ only (d is the dimensionality and l is usually a small number). This is much faster than most existing fast methods. Specifically, we engage the locality sensitive hashing technique to divide items into small subsets with equal size, and then build one kNN graph on each subset using the brute force method. To enhance the approximation quality, we repeat this procedure for several times to generate multiple basic approximate graphs, and combine them to yield a high quality graph. Compared with existing methods, the proposed approach has features that: (1) much more efficient in speed (2) applicable to generic similarity measures; (3) easy to parallelize. Finally, on three benchmark large-scale data sets, our method beats existing fast methods with obvious advantages.

Thu3A: Sequence & Time Series Analysis

Room: Euforie

16:10-16:35 **Fast Sequence Segmentation Using Log-Linear Models**
Nikolaj Tatti

Segmentation, dividing a given sequence into homogeneous segments, is a well-studied problem that has many practical applications. A standard approach to find the optimal solution is by using a dynamic program. Unfortunately, the execution time of the program is quadratic w.r.t the length of the input sequence. This makes the algorithm slow for a sequence of non-trivial length. In this paper we study segmentation that use log-linear models, a rich family that contains many of the standard distributions. We give a theoretical result allowing us to prune some segments. Using this result, we modify the standard dynamic program and develop an efficient pruning technique for one-dimensional log-linear models. We demonstrate empirically, that this approach can significantly reduce the computational burden of finding the optimal segmentation.

16:35-16:55 **Explaining Interval Sequences by Randomization**
Andreas Henelius, Jussi Korpela and Kai Puolamaki

Sequences of events are an ubiquitous form of data. In this paper, we show that it is feasible to present an event sequence as an interval sequence. We show how sequences can be efficiently randomized, how to choose a correct null model and how to use randomizations to derive confidence intervals. Using these techniques, we gain knowledge of the temporal structure of the sequence. Time and Fourier space representations, autocorrelations and arbitrary features can be used as constraints in investigating the data. The methods presented are applied to two real-life datasets: a medical heart interbeat interval dataset and a word dataset from a book. We find that the interval sequence representation and randomization methods provide a powerful way to explore interval sequences and explain their structure.

16:55-17:15 **A Layered Dirichlet Process for Hierarchical Segmentation of Sequential Grouped Data**
Adway Mitra, Ranganath B.N. and Indrajit Bhattacharya

We address the problem of hierarchical segmentation of sequential grouped data, such as a collection of textual documents, and propose a Bayesian nonparametric approach for this problem. Existing Bayesian nonparametric models such as the sticky HDP-HMM are suitable only for single-layer segmentation. We propose the Layered Dirichlet Process (LaDP), where each layer has a countable set of Dirichlet Processes, draws from which define a distribution over the countable set of Dirichlet Processes at the next layer. Each data item gets assigned to a distribution (index) from each layer of the hierarchy, leading to hierarchical segmentation of the sequence. The complexity of inference depends upon the exchangeability assumptions for the measures at different layers. We propose a new notion of exchangeability called Block Exchangeability, which lies between Markov Exchangeability (used in HDP-HMM) and Complete Group Exchangeability (used in HDP), and allows for faster inference than Markov Exchangeability. Using experiments on a news transcript dataset and a product review dataset, we show that LaDP generalizes better than existing non-parametric models for sequential data, and by simultaneously segmenting at multiple levels, outperforms existing models in terms of single-layer segmentation. We also show empirically that using Block Exchangeability greatly speeds up inference and allows trading off accuracy for execution time.

17:15-17:35 **Fault Tolerant Regression for Sensor Data**
Indre Zliobaite and Jaakko Hollmen

Many systems rely on predictive models using sensor data, with sensors being prone to occasional failures. From the operational point of view predictions need to be tolerant to sensor failures such that the loss in accuracy due to temporary missing sensor readings would be minimal. In this paper, we theoretically and empirically analyze robustness of linear predictive models to temporary missing data. We demonstrate that if the input sensors are correlated the mean imputation of missing values may lead to a very rapid deterioration of the prediction accuracy. Based on the theoretical results we introduce a quantitative measure that allows to assess how robust is a given linear regression model to sensor failures. We propose a practical strategy for building and operating robust linear models in situations when temporal sensor failures are expected. Experiments on six sensory datasets and a case study in environmental monitoring with streaming data validate the theoretical results and confirm the effectiveness of the proposed strategy.

Thu3B: Declarative Data Mining & Meta Learning

Room: Dialog

16:10-16:35 **Pairwise Meta-rules for Better Meta-learning-Based Algorithm Ranking**
Quan Sun and Bernhard Pfahringer

We present a novel meta-feature generation method for meta-learning, which is based on rules that compare the performance of individual base learners in a one-against-one manner. In addition to these new meta-features, we also introduce a new meta-learner called Approximate Ranking Tree Forests (ART forests) that performs very competitively when compared with several state-of-the-art meta-learners. Our experimental results are based on a large collection of datasets and show that the proposed new techniques can improve the overall performance of meta-learning for algorithm ranking significantly. A key point in our approach is that each performance figure of any base learner for any specific dataset is generated by optimising the parameters of the base learner separately for each dataset.

16:35-16:55 **SNNAP: Solver-Based Nearest Neighbor for Algorithm Portfolios**
Marco Collautti, Yuri Malitsky and Barry O'Sullivan

The success of portfolio algorithms in competitions in the area of combinatorial problem solving, as well as in practice, has motivated interest in the development of new approaches to determine the best solver for the problem at hand. Yet, although there are a number of ways in which this decision can be made, it always relies on a rich set of features to identify and distinguish the structure of the problem instances. In this paper, we show how one of the more successful portfolio approaches, ISAC, can be augmented by taking into account the past performance of solvers as part of the feature vector. Testing on a variety of SAT datasets, we show how our new formulation continuously outperforms an unmodified/standard version of ISAC.

16:55-17:15 **Top-k Frequent Closed Itemset Mining Using Top-k SAT Problem**
Said Jabbour, Lakhdar Sais and Yakoub Salhi

In this paper, we introduce a new problem, called Top-k SAT, that consists in enumerating the top-k models of a propositional formula. A top-k model is defined as a model with less than k-1 models preferred to it w.r.t. a preference relation. We show that Top-k SAT generalizes the two well known problems: the partial Max-SAT problem and the problem of computing minimal models. Moreover, we propose a general algorithm for Top-k SAT. Then, we give the first application of our declarative framework in data mining, namely, the problem of enumerating top-k frequent closed itemsets of length at least min. Finally, to show the nice declarative aspects of our framework, we encode several other variants of this data mining problem as Top-k SAT problems.

17:15-17:35 **A Feclarative Framework for Constrained Clustering**
Thi-Bich-Hanh Dao, Khanh-Chuong Duong and Christel Vrain

In recent years, clustering has been extended to constrained clustering, so as to integrate knowledge on objects or on clusters, but adding such constraints generally requires to develop new algorithms. We propose a declarative and generic framework, based on Constraint Programming, which enables to design clustering tasks by specifying an optimization criterion and some constraints either on the clusters or on pairs of objects. In our framework, several classical optimization criteria are considered and they can be coupled with different kinds of constraints. Relying on Constraint Programming has two main advantages: the declarativity, which enables to easily add new constraints and the ability to find an optimal solution satisfying all the constraints (when there exists one). On the other hand, computation time depends on the constraints and on their ability to reduce the domain of variables, thus avoiding an exhaustive search.

Thu3C: Topic Models

Room: Ceremonie

16:10-16:35 **Probabilistic Topic Models for Sequence Data**
Nicola Barbieri, Antonio Bevacqua, Marco Carnuccio, Giuseppe Manco and Ettore Ritacco

Probabilistic topic models are widely used in different contexts to uncover the hidden structure in large text corpora. One of the main (and perhaps strong) assumption of these models is that generative process follows a bag-of-words assumption, i.e each token is independent from the previous one. We extend the popular Latent Dirichlet Allocation model by exploiting three different conditional markovian assumptions: (i) the token generation depends on the current topic and on the previous token; (ii) the topic associated with each observation depends on topic associated with the previous one; (iii) the token generation depends on the current and previous topic. For each of these modeling assumptions we present a fast Gibbs Sampling procedure for the parameters estimation. A thorough experimental evaluation over real-word data shows the performance advantages, in terms of recall and precision, of the sequence-modeling approaches.

16:35-16:55 **Nested Hierarchical Dirichlet Process for Nonparametric Entity-Topic Analysis**
Priyanka Agrawal, Lavanya Tekumalla and Indrajit Bhattacharya

The Hierarchical Dirichlet Process (HDP) is a Bayesian nonparametric prior for grouped data, such as collections of documents, where each group is a mixture of a set of shared mixture densities, or topics, where the number of topics is not fixed, but grows with data size. The Nested Dirichlet Process (NDP) builds on the HDP to cluster the documents, but allowing them to choose only from a set of specific topic mixtures. In many applications, such a set of topic mixtures may be identified with the set of entities for the collection. However, in many applications, multiple entities are associated with documents, and often the set of entities may also not be known completely in advance. In this paper, we address this problem using a nested HDP (nHDP), where the base distribution of an outer HDP is itself an HDP. The inner HDP creates a countably infinite set of topic mixtures and associates them with entities, while the outer HDP associates documents with these entities or topic mixtures. Making use of a nested Chinese Restaurant Franchise (nCRF) representation for the nested HDP, we propose a collapsed Gibbs sampling based inference algorithm for the model. Because of couplings between two HDP levels, scaling up is naturally a challenge for the inference algorithm. We propose an inference algorithm by extending the direct sampling scheme of the HDP to two levels. In our experiments on two real world research corpora, we show that, even when large fractions of author entities are hidden, the nHDP is able to generalize significantly better than existing models. More importantly, we are able to detect missing authors at a reasonable level of accuracy.

16:55-17:15 **From Topic Models to Semi-supervised Learning: Biasing Mixed-Membership Models to Exploit Topic-Indicative Features in Entity Clustering**
Ramnath Balasubramanyan, William Cohen and Bhavana Dalvi

We present methods to introduce different forms of supervision into mixed-membership latent variable models. Firstly, we introduce a technique to bias the models to exploit topic-indicative features, i.e. features which are apriori known to be good indicators of the latent topics that generated them. Next, we present methods to modify the Gibbs sampler used for approximate inference in such models to permit injection of stronger forms of supervision in the form of labels for features and documents, along with a description of the corresponding change in the underlying generative process. This ability allows us to span the range from unsupervised topic models to semi-supervised learning in the same mixed membership model. Experimental results from an entity-clustering task demonstrate that the biasing technique and the introduction of feature and document labels provide a significant increase in clustering performance over baseline mixed-membership methods.

17:15-17:35 **Sparse Relational Topic Models for Document Networks**
Aonan Zhang, Jun Zhu and Bo Zhang

Learning latent representations is playing a pivotal role in machine learning and many application areas. Previous work on relational topic models (RTM) has shown promise on learning latent topical representations for describing relational document networks and predicting pairwise links. However under a probabilistic formulation with normalization constraints, RTM could be ineffective in controlling the sparsity of the topical representations, and may often need to make strict mean-field assumptions for approximate inference. This paper presents sparse relational topic models (SRTM) under a non-probabilistic formulation that can effectively control the sparsity via a sparsity-inducing regularizer. Our model can also handle imbalance issues in real networks via introducing various cost parameters for positive and negative links. The deterministic optimization problem of SRTM admits efficient coordinate descent algorithms. We also present a generalization to consider all pairwise topic interactions. Our empirical results on several real network datasets demonstrate better performance on link prediction, sparser latent representations, and faster running time than the competitors under a probabilistic formulation.

Thu3D: Statistical Learning (2)

Room: Aplaus

16:10-16:35 **The Flip-the-State Transition Operator for Restricted Boltzmann Machines**
Kai Brugge, Asja Fischer and Christian Igel

Most learning and sampling algorithms for restricted Boltzmann machines (RMBs) rely on Markov chain Monte Carlo (MCMC) methods using Gibbs sampling. The most prominent examples are Contrastive Divergence learning (CD) and its variants as well as Parallel Tempering (PT).The performance of these methods strongly depends on the mixing properties of the Gibbs chain. We propose a Metropolis-type MCMC algorithm relying on a transition operator maximizing the probability of state changes. It is shown that the operator induces an irreducible and aperiodic and, thus, a properly converging Markov chain also for the typically used periodic update schemes. The transition operator can replace Gibbs sampling in RBM learning algorithms without producing computational overhead. It is shown empirically that this leads to faster mixing and in turn to more accurate learning.

16:35-16:55 **Learning Discriminative Sufficient Statistics Score Space**
Xiong Li, Bin Wang, Yuncai Liu and Tai Sing Lee

Generative score spaces provide a principled method to exploit generative information, e.g., data distribution and hidden variables, in discriminative classifiers. The underlying methodology is to derive measures or score functions from generative models. The derived score functions, spanning the so-called score space, provide features of a fixed dimension for discriminative classification. In this paper, we propose a simple yet effective score space which is essentially the sufficient statistics of the adopted generative models and does not involve the parameters of generative models. We further propose a discriminative learning method for the score space that seeks to utilize label information by constraining the classification margin over the score space. The form of score function allows the formulation of simple learning rules, which are essentially the same learning rules for a generative model with an extra posterior imposed over its hidden variables. Experimental evaluation of this approach over two generative models shows that performance of the score space approach coupled with the proposed discriminative learning method is competitive with state-of-the-art classification methods.

16:55-17:15 **The Stochastic Gradient Descent for the Primal L1-SVM Optimization Revisited**
Constantinos Panagiotakopoulos and Petroula Tsampouka

We reconsider the stochastic (sub)gradient approach to the unconstrained primal L1-SVM optimization. We observe that if the learning rate is inversely proportional to the number of steps, i.e., the number of times any training pattern is presented to the algorithm, the update rule may be transformed into the one of the classical perceptron with margin in which the margin threshold increases linearly with the number of steps. Moreover, if we cycle repeatedly through the possibly randomly permuted training set the dual variables defined naturally via the expansion of the weight vector as a linear combination of the patterns on which margin errors were made are shown to obey at the end of each complete cycle automatically the box constraints arising in dual optimization. This renders the dual Lagrangian a running lower bound on the primal objective tending to it at the optimum and makes available an upper bound on the relative accuracy achieved which provides a meaningful stopping criterion. In addition, we propose a mechanism of presenting the same pattern repeatedly to the algorithm which maintains the above properties. Finally, we give experimental evidence that algorithms constructed along these lines exhibit a considerably improved performance.

17:15-17:35 **Bundle CDN: A Highly Parallelized Approach for Large-Scale L1-regularized Logistic Regression**
Yatao Bian, Xiong Li, mingqi Cao and Yuncai Liu

Parallel coordinate descent algorithms emerge with the growing demand of large-scale optimization. In general, previous algorithms are usually limited by their divergence under high degree of parallelism (DOP), or need data pre-process to avoid divergence. To better exploit parallelism, we propose a coordinate descent based parallel algorithm without needing of data pre-process, termed as Bundle Coordinate Descent Newton (BCDN), and apply it to large-scale l_1 regularized logistic regression. BCDN first randomly partitions the feature set into Q non-overlapping subsets/bundles in a Gauss-Seidel manner, where each bundle contains P features. For each bundle, it finds the descent directions for the P features in parallel, and performs P dimensional Armijo line search to obtain the stepsize. By theoretical analysis on global convergence, we show that BCDN is guaranteed to converge with a high DOP. Experimental evaluations over five public datasets show that BCDN can better exploit parallelism and outperforms state-of-the-art algorithms in speed, without losing testing accuracy.

FRIDAY 27 SEPTEMBER 2013

FRIDAY INVITED TALK

Deep-er Kernels

Speaker: John Shawe-Taylor
Time: 09:00-10:00
Room: plenary

Abstract
Kernels can be viewed as shallow in that learning is only applied in a single (output) layer. Recent successes with deep learning highlight the need to consider learning richer function classes. The talk will review and discuss methods that have been developed to enable richer kernel classes to be learned. While some of these methods rely on greedy procedures many are supported by statistical learning analyses and/or convergence bounds. The talk will highlight the trade-offs involved and the potential for further research on this topic.

Bio
John Shawe-Taylor obtained a PhD in Mathematics at Royal Holloway, University of London in 1986 and joined the Department of Computer Science in the same year. He was promoted to Professor of Computing Science in 1996. He moved to the University of Southampton in 2003 to lead the ISIS research group. He was Director of the Centre for Computational Statistics and Machine Learning at University College, London between July 2006 and September 2010. He has coordinated a number of European wide projects investigating the theory and practice of Machine Learning, including the PASCAL projects. He has published over 300 research papers with more than 25000 citations. He has co-authored with Nello Cristianini two books on kernel approaches to machine learning: 'An Introduction to Support Vector Machines' and 'Kernel Methods for Pattern Analysis'.

FRIDAY MORNING TUTORIALS

Statistically Sound Pattern Discovery

Wilhelmiina Hämäläinen and Geoff Webb
Time: 10:45-15:15
Room: Brissasol

Pattern discovery is a core data mining activity. Initial approaches were dominated by the frequent pattern discovery paradigm-only frequent patterns were explored. Having been thoroughly explored and its limitations now well understood, this paradigm is giving way to two emerging alternatives-the information theoretic minimum message length paradigm and statistically sound paradigm. This tutorial presents the foundations of the latter. In this paradigm, patterns are required to pass statistical tests with respect to user defined null-hypotheses, providing great flexibility about the properties that are sought, and strict control over the risk of false discoveries and overfitting.

Web Scale Information Extraction

Ziqi Zhang and Anna Lisa Gentile
Time: 10:45-15:15
Room: R3

This tutorial analyses open challenges for Web-scale Information Extraction (IE) and introduces the usage of Linked Data as a ground-breaking solution for the field. Training data is an essential resource to machine learning. However, it is expensive to create. The limited availability of such data has so far prevented the study of the generalised use of large-scale resources to port to specific user information needs on Information Extraction tasks. For the last few years Linked Data has grown to a gigantic knowledge base, which, as of 2013, comprised 31 billion triples in 295 data sets (http://lod-cloud.net/state/). Such resources can become invaluable training data for Web-scale Information Extraction and natural language tasks because they are: (i) very large scale, (ii) constantly growing, (iii) covering multiple domains and (iv) being used to annotate a growing number of pages that can be exploited for training.

This tutorial will show how to exploit Linked Data for IE and will explore Information Extraction techniques able to scale at web level and adapt to user information need. We will particularly focus on the tasks of Wrapper Induction and Table Interpretation. As an example of linked data driven IE, we will present and discuss a multi-strategy learning method and framework designed to train Web-scale IE using Linked Data, while coping with noise in the training data. The approach uses multiple strategies: (i) it wraps very regular web sites generated by backing databases; (ii) extracts from regular structures such as tables and lists and (iii) learns lexical-syntactic extraction patterns for information extraction from natural language.

FRIDAY AFTERNOON TUTORIALS

Algorithmic Techniques for Modeling and Mining Large Graphs

Alan Frieze, Aristides Gionis and Charalampos Tsourakakis
Time: 15:45-19:00
Room: Brissasol

Network science has emerged over the last years as an interdisciplinary area spanning traditional domains including mathematics, computer science, sociology, biology and economics. Since complexity in social, biological and economical systems, and more generally in complex systems, arises through pairwise inter - actions there exists a surging interest in understanding networks.

In this tutorial, we will provide an in-depth presentation of the most popular random-graph models used for modeling real-world networks. We will then discuss efficient algorithmic techniques for mining large graphs, with emphasis on the problems of extracting graph sparsifiers, partitioning graphs into densely connected components, and finding dense subgraphs. We will motivate the problems we will discuss and the algorithms we will present with real-world applications.

Our aim is to survey important results in the areas of modeling and mining large graphs, to uncover the intuition behind the key ideas, and to present future research directions.

Multi-agent Reinforcement Learning

Daan Bloembergen, Daniel Hennes, Michael Kaisers, Karl Tuyls and Peter Vrancx
Time: 15:45-19:00
Room: R3

Multi-agent reinforcement learning (MARL) is an important and fundamental topic within agent-based research. After giving successful tutorials on this topic at EASSS 2004 (the European Agent Systems Summer School), ECML 2005, ICML 2006, EWRL 2008 and AAMAS 2009-2012, with different collaborators, we now propose a revised and updated tutorial, covering both theoretical as well as practical aspects of MARL.

Participants will be taught the basics of single-agent reinforcement learning and the associated theoretical convergence guarantees related to Markov Decision Processes. We will then outline why these convergence guarantees no longer hold in a setting where multiple agents learn. We will explain practical approaches on how to scale single agent reinforcement learning to these situations where multiple agents influence each other and introduce a framework, based on game theory and evolutionary game theory, that allows thorough analysis of the dynamics of multi-agent learning. Finally, we will discuss multi-agent learning in continuous state and action spaces.

FRIDAY WORKSHOP

Nameling Discovery Challenge 2013

Stephan Doerfel, Andreas Hotho, Robert Jäschke, Folke Mitzlaff and Juergen Mueller
Time: 10:45-15:15
Room: R1

All over the world, future parents are facing the task of finding a suitable given name for their children. This choice is influenced by different factors, such as the social context, language, cultural background and especially personal taste. Although this task is omnipresent, little research has been conducted on the analysis and application of interrelations among given names from a data mining perspective.

This year's ECML PKDD Discovery Challenge tackles the task of recommending given names. The challenge comprises two phases: First, an offline competition, where participants predict future search activities based on a training data set which is derived from the name search website nameling. Secondly, an online competition, where participants integrate their recommender systems into the nameling website.

10:45-11:15 **20DC13 Opening**

11:15-11:35 **Improving the Recommendation of Given Names by Using Contextual Information**
Marcos Aurélio Domingues, Ricardo Marcondes Marcacini, Solange Oliveira Rezende and Gustavo E. A. P. A. Batista

11:35-11:55 **Nameling Discovery Challenge–Collaborative Neighborhoods**
Dirk Schäfer and Robin Senge

11:55-12:15 **Collaborative Filtering Ensemble for Personalized Name Recommendation**
Bernat Coma-Puig, Ernesto Diaz-Aviles and Wolfgang Nejdl

12:15-13:45 **Lunch break**

13:45-14:05 **A Mixed Hybrid Recommender System for Given Names**
Rafael Glauber, Angelo Loula and João Rocha-Junior

14:05-14:25 **Factor Models for Recommending Given Names**
Imannuel Bayer and Steffen Rendle

14:25-14:45 **Similarity-Weighted Association Rules for a Name Recommender System**
Benjamin Letham

14:45-15:15 **Discussion**

FRIDAY WORKSHOP

Solving Complex Machine Learning Problems with Ensemble Methods

Ioannis Katakis, Daniel Hernández-Lobato, Gonzalo Martínez-Muñoz and Ioannis Partalas
Time: 10:45-19:00
Room: R2

Ensemble methods are widely utilized within the machine learning community due to their accuracy-improving and robustness attributes. Since even elementary ensemble approaches outperform single learners, multiple classifier systems are the go-to solution in applications where higher predictive performance is required. The emphasis in COPEM is to discuss ensemble strategies that solve difficult machine learning tasks. This workshop will bring together the ensemble method community and researchers that are not ensemble-experts but could benefit from utilizing such techniques to confront interesting research challenges. The goals of COPEM are: a) to discuss state-of-the-art approaches that exploit ensembles to solve complex machine learning problems and, b) to bring the community together and discuss interesting future applications. The ultimate objective of COPEM is not only to present high quality research papers but, even most importantly, to dynamically initiate new collaborations that will work towards new challenges.

10:45-11:00 **COPEM: Machine Learning and Ensemble Methods–Overview**
Ioannis Katakis, Daniel Hernández-Lobato, Gonzalo Martínez-Muñoz and Ioannis Partalas

11:00-11:50 **Invited Speaker (Pierre Dupont, Université catholique de Louvain)**

11.50-12.15 **Local Neighbourhood in Generalizing Bagging for Imbalanced Data**
Jerzy Błaszczyński, Jerzy Stefanowski and Marcin Szajek

12:15-13.45 **Lunch Break**

13:45-14:07 **Anomaly Detection by Bagging**
Tomas Pevny

14:07-14:30 **Efficient semi-supervised feature selection by an ensemble approach**
Mohammed Hindawi, Haytham Elghazel and Khalid Benabdeslem

14:30-14:52 **Feature ranking for multi-label classification using predictive clustering trees**
Dragi Kocev, Ivica Slavkov and Sašo Džeroski

14:52-15:15 **Identification of Statistically Significant Features from Random Forests**
Jérôme Paul, Michel Verleysen and Pierre Dupont

15.15-15.45 **Coffee Break**

15:45-16:07 **Prototype Support Vector Machines: Supervised Classification in Complex Datasets**
April Shen and Andrea Danyluk

16:07-16:30 **Software Reliability prediction via two different implementations of Bayesian model averaging**
Alex Sarishvili and Gerrit Hanselmann

16:30-16:52 **Multi-Space Learning for Image Classification Using AdaBoost and Markov Random Fields**
Wenrong Zeng, Xue-Wen Chen, Hong Cheng and Jing Hua

16:52-17:15 **An Empirical Comparison of Supervised Ensemble Learning Approaches**
Mohamed Bibimoune, Haytham Elghazel and Alex Aussem

17.15-17.30 **Coffee Break**

17.30-17.55 **Clustering Ensemble on Reduced Search Spaces**
Sandro Vega-Pons and Paolo Avesani

17.55-18.20 **An Ensemble Approach to Combining Expert Opinions**
Hua Zhang, Evgueni Smirnov, Nikolay Nikolaev, Georgi Nalbantov and Ralf Peeters

18.20-19.00 **Discussion and Conclusions**

FRIDAY WORKSHOP

Real-World Challenges for Data Stream Mining

Georg Krempl, Indrė Žliobaitė, Yin Wang and George Forman
Time: 10:45-16:41
Room: R4

This workshop will provide a forum for researchers and practitioners to discuss real-world challenges for data stream mining, identify gaps between data streams research and meaningful applications, and define new application-relevant research directions for data stream mining. The workshop will focus on oral presentations and discussions.

10:45-11:05 **High-Throughput Continuous Clustering of Message Streams**
Oisin Boydell, Marek Landowski, Guangyu Wu and Padraig Cunningham

11:08-11:28 **Early Classification of Individual Electricity Consumptions**
Asma Dachraoui, Alexis Bondu and Antoine Cornuejols

11:31-11:51 **Streaming Virtual Patient Records**
Pedro Pereira Rodrigues and Ricardo Cruz-Correia

11:54-12:14 **A Discussion on ISS Columbus Data Streams**
Tino Noack, Ingo Schmitt and Enrico Noack

12:14-13:45 **Lunch Break**

13:45-14:15 **Real World Issues in Stream Classification**
Invited talk by Vincent Lemaire

14:18-14:38 **Analysis of Videos using Tile Mining**
Toon Calders, Elisa Fromont, Baptiste Jeudy and Lam Hoang Thanh

14:41-15:01 **Event History Analysis on Data Streams: An Application to Earthquake Occurrence**
Ammar Shaker and Eyke Huellermeier

15:04-15:14 **Real Time Event Monitoring with Trident**
Igor Brigadir, Derek Greene, Padraig Cunningham and Gavin Sheridan

15:14-15:45 **Coffee Break**

15:45-16:05 **Survival Analysis Meets Data Stream Mining**
Mark Last and Hezi Halpert

16:08-16:28 **Anticipative and Dynamic Adaptation to Concept Changes**
Ghazal Jaber, Antoine Cornuejols and Philippe Tarroux

16:31-16:41 **Classifiers for Concept-drifting Data Streams: Evaluating Things That Really Matter**
Dariusz Brzezinski and Jerzy Stefanowski

FRIDAY WORKSHOP
Sports Analytics

Albrecht Zimmermann, Jesse Davis and Jan van Haaren
Time: 10:45-18:30
Room: R5

Sports Analytics has been a steadily growing area in the last decade, especially in the context of US professional sports leagues but also in connection with European football leagues. The recent implementation of strict financial fair-play regulations in European football will definitely render sports analytics even more important in the coming years. In addition, there is of course the always popular sports betting. Developed approaches have been used for decision support in all aspects of professional sports:

- Player acquisition and team spending
- Training regimens and focus
- Match strategy
- Injury prediction and prevention
- Predicting match outcomes
- Betting odds calculation
- Text analysis of match reports
- Descriptive modeling

The majority of techniques used in the field so far are statistical and while there has been some interest in the Machine Learning and Data Mining community, it has been somewhat muted so far. We intend to change this by hosting a workshop on Sports Analytics at ECML/PKDD 2013. Not only do we believe that the setting is interesting and challenging, and can potentially be a source of new data, but also that this offers a great opportunity to bring people from outside of the ML community into contact with typical ECML/PKDD contributors, and to highlight what the community has done and can do in the field.

10:45-10:55 **Introduction**
Albrecht Zimmermann

10:55-11:45 **Invited talk**
Thomas Gärtner (University of Bonn and Fraunhofer IAIS)

11:45-12:05 **Why Do Sports Officials Dropout?**
Fabrice Dosseville, François Rioult and Sylvain Laborde

12:05-12:25 **Strategic Patterns Discovery in RTS-games for E-Sport with Sequential Pattern Mining**
Guillaume Bosc, Mehdi Kaytoue, Chedy Raïssi and Jean-François Boulicaut

12:25-13:55 **Lunch break**

13:55-14:15 **Maps for Reasoning in Ultimate**
Jeremy Weiss and Sean Childers

14:15-14:35 **Predicting the NFL using Twitter**
Shiladitya Sinha, Chris Dyer, Kevin Gimpel and Noah A. Smith

14:35-14:55 **Use of Performance Metrics to Forecast Success in the National Hockey League**
Joshua Weissbock, Herna Viktor and Diana Inkpen

14:55-15:15 **Finding Similar Movements in Positional Data Streams**
Jens Haase and Ulf Brefeld

15:15-15:45 **Coffee break**

15:45-16:35 **Invited talk**
Steven Probst (TopSportsLab)

16:35-16:55 **Comparison of Machine Learning Methods for Predicting the Recovery Time of Professional Football Players After an Undiagnosed Injury**
Stylianos Kampakis

16:55-17:15 **Predicting NCAAB Match Outcomes Using ML Techniques–Some Results and Lessons Learned**
Albrecht Zimmermann, Sruthi Moorthy and Zifan Shi

17:15-17:35 **Coffee break**

17:35-17:55 **Inverse Reinforcement Learning for Strategy Extraction**
Katharina Muelling, Abdeslam Boularias, Betty Mohler, Bernhard Schoelkopf and Jan Peters

17:55-18:15 **Key Point Selection and Clustering of Swimmer Coordination Through Sparse Fisher-EM**
John Komar, Romain Herault and Ludovic Seifert

18:15-18:30 **Wrap up and discussion**
Albrecht Zimmermann

FRIDAY WORKSHOP

Tensor Methods for Machine Learning

Maximilian Nickel and Volker Tresp
Time: 10:45-17:15
Room: R6

Tensors, as generalizations of vectors and matrices, have become increasingly popular in different areas of machine learning and data mining, where they are employed to approach a diverse number of difficult learning and analysis tasks. Prominent examples include learning on multi-relational data and large-scale knowledge bases, recommendation systems, computer vision, mining Boolean data, neuroimaging or the analysis of time-varying networks. The success of tensors methods is strongly related to their ability to efficiently model, analyze and predict data with multiple modalities. To address specific challenges and problems, a variety of methods has been developed in different fields of application. This workshop should serve as a basis for an interdisciplinary exchange of methods, ideas and techniques, with the goal to develop a deeper understanding of tensor methods in machine learning, further advance existing approaches and enable new approaches to important problems. The workshop is intended for researchers in the machine learning, data mining and tensor communities to discuss novel methods and applications as well as theoretical advances.

Tentative schedule, please check workshop site for updates.

10:45-10:55 **Welcome Address**

10:55-11:35 **Tensor Decompositions for Machine Learning and the Modelling of Neuroimaging Data**
Morten Mørup

11:35-12:15 **Advances in (Numerical) Multilinear Algebra**
Lieven de Lathauwer

12:15-13:45 **Lunch Break**

13:45-14:25 **Factorization Machines**
Steffen Rendle

14:25-15:05 **Probabilistic Latent Tensor Factorization with Applications to Audio Processing and Source Separation**
Ali Taylan Cemgil

15:05-15:15 **Spotlight Talks**

15:15-15:45 **Poster Session and Coffee Break**

15:45-16:25 **Boolean Tensor and Matrix Factorization**
Pauli Miettinen

16:25-16:45 **Non-Negative Tensor Factorization with RESCAL**
Denis Krompass

16:45-17:15 **Discussion and Wrap-Up**

FRIDAY WORKSHOP

New Fronteirs in Mining Complex Patterns

Annalisa Appice, Michelangelo Ceci, Corrado Loglisci, Giuseppe Manco, Elio Masciari, Zbigniew W. Ras
Time: 10:45-19:10
Room: R7

Data mining and knowledge discovery can be considered today as mature research fields with numerous algorithms and studies to extract knowledge from data in different forms. Although, most existing data mining approaches look for patterns in tabular data, there are also numerous studies which already look for patterns in complex data. The recent developments in technologies and life sciences have paved the way to the proliferation of data collections representing new complex interactions between entities in distributed and heterogeneous sources. These interactions may be spanned at multiple levels of granularity as well as at spatial and temporal dimensions. The purpose of this workshop is to bring together researchers and practitioners of data mining interested in exploring emerging technologies and applications where complex patterns in expressive languages are principally extracted from new prominent data sources like blogs, event or log data, biological data, spatio-temporal data, social networks, mobility data, sensor data and streams, and so on. We are interested in advanced techniques which preserve the informative richness of data and allow us to efficiently and efficaciously identify complex information units present in such data.

10:45-10:50 **Welcome**

10:50-11:35 **Invited Talk : Evolving Social Networks: trajectories of communities**
Joao Gama

Data Streams and Time Series Analysis 1

11:35-11:55 **A Study on Parameter Estimation for a Mining Flock Algorithm**
Chiara Renso, Rebecca Ong, Mirco Nanni, Monica Wachowicz and Dino Pedreschi

11:55-12:05 **Online Batch Weighted Ensemble for Mining Data Streams with Concept Drift**
Magdalena Deckert

12:05-12:15 **Feature extraction over multiple representations for time series classification**
Dominique Gay, Romain Guigourès, Marc Boullé and Fabrice Clérot

12:15-13:45 **Lunch**
Classification and pattern discovery

13:45-14:05 **A Hybrid Distance-based Method and Support Vector Machines for Emotional Speech Detection**
Vladimer Kobayashi

14:05-14:25 **IndexSpan: Efficient Discovery of Item-Indexable Sequential Patterns**
Rui Henriques, Claudia Antunes and Sara Madeira

14:25-14:45 **F2G: Efficient Discovery of Full-Patterns**
Rui Henriques, Claudia Antunes and Sara Madeira

14:45-15:05 **Mining Frequent Partite Episodes with Partwise Constraints**
Takashi Kato, Shin-Ichiro Tago, Tatsuya Asai, Hiroaki Morikawa, Junichi Shigezumi and Hiroya Inakoshi

15:05-15:15 **Structure Determination and Estimation of Hierarchical Archimedean Copulas Based on Kendall Correlation Matrix**
Jan Górecki and Martin Holeňa

15:15-15:45 **Coffee break**

Machine Learning and Music

15:45-16: 05 **Developing Personalized Classiifiers for Retrieving Music by Mood**
Amanda Mostafavi, Zbigniew Ras and Alicja Wieczorkowska

16:05-16:15 **Mining Audio Data for Multiple Instrument Recognition in Classical Music**
Elzbiera Kubera and Alicja Wieczorkowska

Networks and Graphs

16:15-16:35 **AGWAN: A Generative Model for Labelled, Weighted Graphs**
Michael Davis, Weiru Liu and Paul Miller

16:35-16:45 **Thresholding of Semantic Similarity Networks using a Spectral Graph Based Technique**
Pietro Hiram Guzzi, Simone Truglia, Pierangelo Veltri and Mario Cannataro

Mining Relational Data

16:45-17:05 **A Relational Unsupervised Approach to Author Identification**
Fabio Leuzzi, Stefano Ferilli and Fulvio Rotella

17:05-17:15 **Towards extracting relations from unstructured data through natural language semantics**
Diana Trandabat

17:15-17:30 **Coffe Break**

Classification and clustering

17:30-17:50 **Extending ReliefF for Hierarchical Multi-label Classification**
Jana Karcheska, Ivica Slavkov, Dragi Kocev, Slobodan Kalajdziski and Saso Dzeroski

17:50-18:00 **The use of the label hierarchy in HMC improves performance: A case study in predicting community structure in ecology**
Jurica Levatić, Dragi Kocev and Sašo Džeroski

18:00-18:20 **XML Document Partitioning using Ensemble Clustering**
Gianni Costa and Riccardo Ortale

Data Streams and Time Series Analysis 2

18:20-18:40 **Conditional Log-Likelihood for Continuous Time Bayesian Network Classifiers**
Daniele Codecasa and Fabio Stella

18:40-18:50 **A Sliding Window Approach for Discovering Dense Areas in Trajectory Streams**
Corrado Loglisci and Donato Malerba

18:50-19:00 **Sequential Pattern Mining from Trajectory Data**
Elio Masciari, Gao Shi and Carlo Zaniolo

19:00-19:10 **Process Mining to Forecast Future of Running Cases**
Annalisa Appice, Sonja Pravilovic and Donato Malerba

NOTES

WWW.ECMLPKDD2013.ORG