SICSA Data Science Meeting: Well Sorted Materials

3/7/2014

Dear participant,

Thank you for submitting your research interest in response to the following request:

"Simply try to encapsulate your main data science research interest(s) in the 50-character 'Title' below. Optionally, you can also provide a supporting description of up to 255 characters."

This document contains your answers, grouped by the average of your online sorts.

Contents

For an interactive, online version of the Tree Map and a PDF version of this document, go here:

http://www.well-sorted.org/explore/SICSADataScience/

Tree Map - titles of all 64 research interests grouped by 'averaging' all participants' groupings

PURPLE

- Sinanovic: Information-theoretic anomaly detection
- Bayesian inference, probabilistic machine learning
- i) Transfer learning ii) Representation learning
- Diakonikolas: Sublinear Learning Algorithms
- Inference & learning in recurrent neural networks
- large-scale feature selection and extraction
- Understanding causes, making predictions
- similarity search, probabilistic distance metrics
- Deep learning for perceptual data
- probabilistic machine learning from time series
- Probabilistic Machine Learning
- Probabilistic Machine Learning
- Probabilistic Machine Learning
- Automated machine learning
- Probabilistic Machine Learning

GREEN

- Large-scale heterogeneous systems data processing
- Distributed Optimization and Learning of Big Data
- Programming languages for big data
- Big data system architectures, complex queries
- Petrie: Applied Constraint Programming
- Learning, linking, structuring of unlabelled data
- Cleaning, linking and interpreting data
- Maintaining consistency between data sources

ORANGE

- Privacy-preserving data archiving and sharing
- Privacy preserving big data processing and sharing
- differential privacy and probabilistic programming

RED

- statistical machine learning and bioinformatics
- Disease management from images of the retina
- Analysis of biological data, Machine Learning
- Big data for neurophysiologist-s
- Bioinformatics to computational neuroscience
- understanding epigenetic data

BLUE

- Complex networks, complex processes
- Visualisation, combined with stats/formal models
- Linked open data, data visualization
- Information Visualisation

YELLOW

- Social Media Analysis and Retrieval: IR approach
- Computer Supported Collaborative Work
- social computation
- Compositional Distributional Semantics
- Semantic Web Discovery

PINK

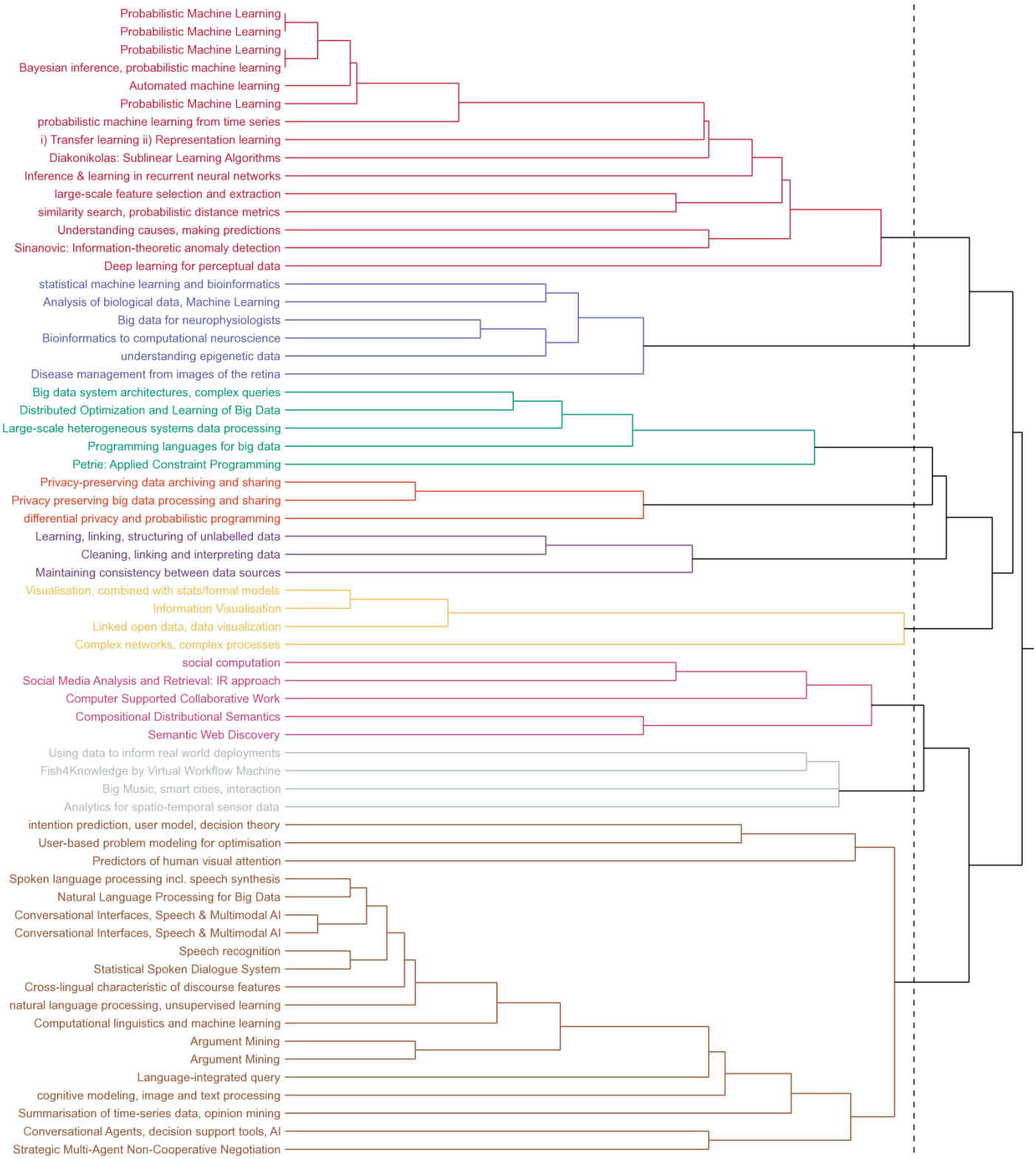
- intention prediction, user model, decision theory
- natural language processing, unsupervised learning
- Summarisation of time-series data, opinion mining
- Strategic Multi-Agent Non-Cooperative Negotiation
- Conversational Agents, decision support tools, AI
- Language-integrated query
- User-based problem modeling for optimisation
- Natural Language Processing for Big Data
- Statistical Spoken Dialogue System
- Argument Mining
- Argument Mining

BROWN

- Using data to inform real world deployments
- Big Music, smart cities, interaction
- Fish4Knowledge by Virtual Workflow Machine
- Analytics for spatio-temporal sensor data
- cognitive modeling, image and text processing
- Speech recognition
- Conversational Interfaces, Speech & Multimodal AI
- Cross-lingual characteristic of discourse features
- Conversational Interfaces, Speech & Multimodal AI
- Predictors of human visual attention
- Spoken language processing incl. speech synthesis
- Computational linguistics and machine learning

SILVER

Dendrogram - showing similarity between research interests



Research interest titles and descriptions ordered by group

Colour	#	Group Members	Description
Red	1	Probabilistic Machine Learning	machine learning models and methods, probabilistic modelling, Bayesian methods, deep learning, sampling, machine learning for neuroscience, applied probability, inference, time series, machine learning markets, agents, mechanism design
	2	Probabilistic Machine Learning	My group builds new probabilistic methodologies for machine learning motivated by a diverse range of applications, include natural language processing, sustainable energy, and software engineering.
	3	Probabilistic Machine Learning	Interests: theoretical and practical issues in machine learning, statistical pattern recognition, probabilistic graphical models and computer vision. This includes theoretical foundations, the development of new models and algorithms, and applications.
	4	Bayesian inference, probabilistic machine learning	- Bayesian inference and Markov chain Monte Carlo methods - probabilistic machine learning and statistical pattern recognition - Gaussian processes - applications in neuroimaging, time series analysis, disclosure risk estimation, and signal processing
	5	Automated machine learning	Machine learning requires many human decisions to be made. Automatically allowing the machine to make the right decision would improve the quality of the result.
	6	Probabilistic Machine Learning	Including inference, with statements of uncertainty, from indirect and messy observations.
	7	probabilistic machine learning from time series	Learning from medical signals
	8	i) Transfer learning ii) Representation learning	i) In many data science scenarios, the test set distribution differs from the training one. Transfer learning tries to address this problem. ii) The unsupervised extraction of good features is often crucial when analysing large collections of data.
	9	Diakonikolas: Sublinear Learning Algorithms	* What questions about a dataset can be answered extremely efficiently, that is, only require computational resources (time, or memory) that are sub-linear in the size of the dataset? * What structural properties of a dataset can be efficiently deduced?
	10	Inference & learning in recurrent neural networks	Sampling based inference and learning in recurrent neural networks with irreversible stochastic dynamics, from both the perspective of improving sampling efficiency and relevance to biological neural networks
	11	large-scale feature selection and extraction	For many very high-profile prediction tasks, multiple data sources and types are increasingly available. I am interested in new and fast ways to find the right features to build the best predictive models.
	12	similarity search, probabilistic distance metrics	extraction of similar objects from very large data collections where only available information is a similarity relation among objects
	13	Understanding causes, making predictions	As they say, "data scientist" = "statistician who lives in California". I have previously worked on medical and other challenges, now play with financial data on the minutes-to-milliseconds scale: interesting because no time-invariant underlying causes.
	14	Sinanovic: Information-theoretic anomaly detection	I am interested in using the tools from information theory in spotting anomalous behaviour in data. I have experience in developing algorithms and implementing them in hardware in the analysis of wireless communication data.
	15	Deep learning for perceptual data	Deep learning with interest in sparsity, multimodality, attention, and temporal data on spatiotemporal data modalities such as vision and audition

Blue	1	statistical machine learning and bioinformatics	Probabilistic methods for large-scale, non-i.i.d data generated from biological experiments, with particular focus on dynamical systems and time series data.
	2	Analysis of biological data, Machine Learning	na
	3	Big data for neurophysiologists	Neurophysiological datasets are large, difficult to collect, and need specialised analysis software. Enabling data sharing and cross-analysis through international organisations (INCF), as well as through portals, encourages real scientific collaboration.
	4	Bioinformatics to computational neuroscience	The aim is to produce mechanistic models of synaptic function constrained by databases of protein-protein interactions.
	5	understanding epigenetic data	I am looking at epigenetic data (protein binding sites, methylation loci, gene expression) and exploring interactions, or the lack thereof, between them.
	6	Disease management from images of the retina	Signal and image processing combined with machine learning to characterise pathologies on image data obtained from the human retina. Image data is acquired in clinical studies.
Green	1	Big data system architectures, complex queries	Big data systems, complex analytics query processing, massively-parallel data access, graph/structured/unstructured data systems, scalability, efficiency, availability, NoSQL systems, Machine Learning techniques
	2	Distributed Optimization and Learning of Big Data	Large datasets must be handled using clusters. My interest is developing data processing algorithms that can utilize the power of multiple distributed computers. With additional interest in streaming algorithms, geometric techniques and spatial data.
	3	Large-scale heterogeneous systems data processing	Large-scale heterogeneous systems offer tremendous data processing power at the cost of extreme programming difficulties. This research aims at simplifying the task of programming these systems while achieving high performance.
	4	Programming languages for big data	Languages which make it natural to write big data parallel programs (with MapReduce a paradigmatic example).
	5	Petrie: Applied Constraint Programming	Interested in how Constraint Programming and related techniques can be applied to solve problems in Data Science.
Orange	1	Privacy-preserving data archiving and sharing	I am abusing this box to just note that I won't be able to make the launch event. Have fun!
	2	Privacy preserving big data processing and sharing	How to protect data in use and shared with untrusted parties. How to make the protection mechanisms efficient and can scale to big datasets.
	3	differential privacy and probabilistic programming	differential privacy and probabilistic programming
Purple	1	Learning, linking, structuring of unlabelled data	The majority of data nowadays comes without any labels and can be of any type and content. Focus should be put on unsupervised methods for learning from and structuring such data, without any knowledge given about the type/content of data.
	2	Cleaning, linking and interpreting data	Data coming from many sources is messy and difficult to align with other sources. Even if it does align nicely, the alignment/mapping/linking of data is contextual. Users need to interpret the links to enable the use of the integrated data.
	3	Maintaining consistency between data sources	Motivated especially from cases where the data sources in question are models used in large-scale software development, each describing an aspect of the system or its use. Mathematical properties of consistency maintainers (bx), e.g. least change.

Yellow	1	Visualisation, combined with stats/formal models	Data visualisation, linking out to colleagues in inference and formal modelling in new exploratory data analysis. Focus now mostly on log data from phone apps, but keen to expand to more general 'future cities' areas
	2	Information Visualisation	Data analytics, Information Visualisation, making data analysis algorithms visible to researchers and the public.
	3	Linked open data, data visualization	This should be optional.
	4	Complex networks, complex processes	Complex networks seem to be a promising way to model a whole range of phenomena. Processes over them can be analysed mathematically or simulated, so they sit at an interesting nexus between mathematics and computer science.
Pink	1	social computation	Deriving and managing data obtained voluntarily at large scale from human populations.
	2	Social Media Analysis and Retrieval: IR approach	Analysing the streams of social media information (Twitter, Flickr etc.) and detecting events, building contextual models, and using them to develop robust retrieval algorithms. These include the analysis of both textual and multimedia documents.
	3	Computer Supported Collaborative Work	What future tools and methodologies do we need to aid the Working Together Initiative and other group collaborations, e.g. Networks of academics
	4	Compositional Distributional Semantics	.
	5	Semantic Web Discovery	Using inference over semanticised descriptions of web resources to discover, connect and represent web resources to their users and potential users.
Silver	1	Using data to inform real world deployments	How can data be used to improve and redesign games and mobile applications? What are the ethical issues?
	2	Fish4Knowledge by Virtual Workflow Machine	Given 5 years of continuously marine life video recording in the open sea, the Knowledge based Virtual Workflow Machine (lead by Dr. Yun-Heh Chen-Burger) composes workflow and execute image processing modules at the rune-time to provide meta-data labels.
	3	Big Music, smart cities, interaction	The big music is a combination of automatic analysis of music content, but also millions of users interacting with that content and each other. Interaction, visualisation and machine learning)
	4	Analytics for spatio-temporal sensor data	Analytics and predictive models for the "sense-learn-act" cycle in wireless sensor/actuator networks. Applications include healthcare (model deterioration in long-term conditions), environmental monitoring and motion capture using body-worn sensors.

Brown	1	intention prediction, user model, decision theory	My work aims at interactively intelligent systems that can make decisions in dynamic environments, involving intention and plan recognition, as well as modelling human choice behaviour incorporating behavioural and long-term variations from optimal models
	2	User-based problem modeling for optimisation	Real-world problems are difficult to formulate by domain experts. Explicitly describing all constraints is often impossible, which impacts problem optimisation and recommendations. Using user feedback to improve the problem model could improve the optimis
	3	Predictors of human visual attention	Encompasses Human-Computer Interaction, information visualisation, accessibility. Covers multimodal media: combinations of text + audio + images (static and dynamic) and infographics (although I'm not particularly fond of that term...).
	4	Spoken language processing incl. speech synthesis	Speech synthesis, speech recognition, speech signal processing, speaker diarization, natural language processing applied to speech synthesis, speech science
	5	Natural Language Processing for Big Data	Most content data is either unstructured text or semizstructured data (comments fields in medical databases). In the latter case, the content is the structured data of the future, if we can only make it so. We apply robust NLP techniques to this problem.
	6	Conversational Interfaces, Speech & Multimodal AI	Using machine learning for conversational natural language processing. Supervised learning, reinforcement learning, and unsupervised methods for language understanding, dialogue management, and natural language generation. Real spoken and multimodal data
	7	Conversational Interfaces, Speech & Multimodal AI	Using machine learning for conversational natural language processing. Supervised learning, reinforcement learning, and unsupervised methods for language understanding, dialogue management, and natural language generation. Real spoken and multimodal data
	8	Speech recognition	learning speech recognition models from large amounts of diverse data
	9	Statistical Spoken Dialogue System	Dialogue State Tracking, Dialogue Policy Optimisation, POMDP, Reinforcement Learning, Multi-Domain Dialogue, Open-Domain Dialogue
	10	Cross-lingual characteristic of discourse features	Multi-sentence texts in every language use linguistic features that support efficient reference to the same thing more than once or that efficiently convey how situations and events are meant to relate to one another. Modelling this requires alot of data.
	11	natural language processing, unsupervised learning	probabilistic models for NLP and modelling of human language acquisition, also analysis of linguistic corpus data (including text and speech)
	12	Computational linguistics and machine learning	My broad interests are in the intersection of computational linguistics and statistical learning. I am interested in developing ways to computationally reason about structured objects that appear in language or otherwise.
	13	Argument Mining	Using structural and statistical techniques to identify arguments -- linguistic expressions of inferences -- in diverse source material including both monologue and dialogue.
	14	Argument Mining	Argument mining exploits the techniques and methods of NLP (text and opinion mining) for semi-automatic and automatic extraction of structured argument data from unstructured natural language text. (See SICSA workshop: www.arg-tech.org/swam2014)
	15	Language-integrated query	Integrating database query into a programming language can reduce impedance mismatch. Example languages include Buneman et al's Kleisli and Microsoft's LINQ.
	16	cognitive modeling, image and text processing	Probabilistic models of cognition, parsing, language production, language acquisition, language vision interface, eyetracking
	17	Summarisation of time-series data, opinion mining	My primary research interest is the automatic textual summarisation from time-series data using machine learning techniques. Recently I have been interested in opinion mining as well.
	18	Conversational Agents, decision support tools, AI	Use data-driven statistical techniques to optimise interaction between agents. reinforcement learning and semi-supervised methods. Combine Natural Language Processing techniques with decision support. Sentiment analysis and distant supervision.
	19	Strategic Multi-Agent Non-Cooperative Negotiation	With the use of Reinforcement Learning on a dialogue agent, in a multi-agent non-cooperative environment, we exploit the adversarial decisions through manipulating actions (deception) which are based on implicatures.

Heat Map - showing a similarity matrix between research interests