Andreas Holzinger

VO 706.996 W DiplomandInnen Seminar
VO 706.997 W DissertantInnen Seminar

Graz University of Technology

Welcome to the Holzinger Group HCI-KDD

Part 2: What are some of our research topics, goals, questions and projects ?

a.holzinger@hci-kdd.org
Web: http://hci-kdd.org/scientific-working-for-students

Please do not forget:

- The "best" is the enemy of the "good" – whenever you try to be "perfect" – there is the danger that you finalize nothing*) …"

*) zero, nada, null

François-Marie Arouet (1694 – 1778)
known as "Voltaire"

Motto of the Holzinger Group

<u>Science</u> is to test crazy ideas – <u>Engineering</u> is put these ideas into Business!

Learning Goals

- At the end of this seminar you should
- … be familiar with the formal requirements
- … be aware of the requirements for a PhD
- … know the HCI-KDD approach
- … have an overview on our research topics
- … understand what research is
- … getting started with your work
- … understand how to write a paper

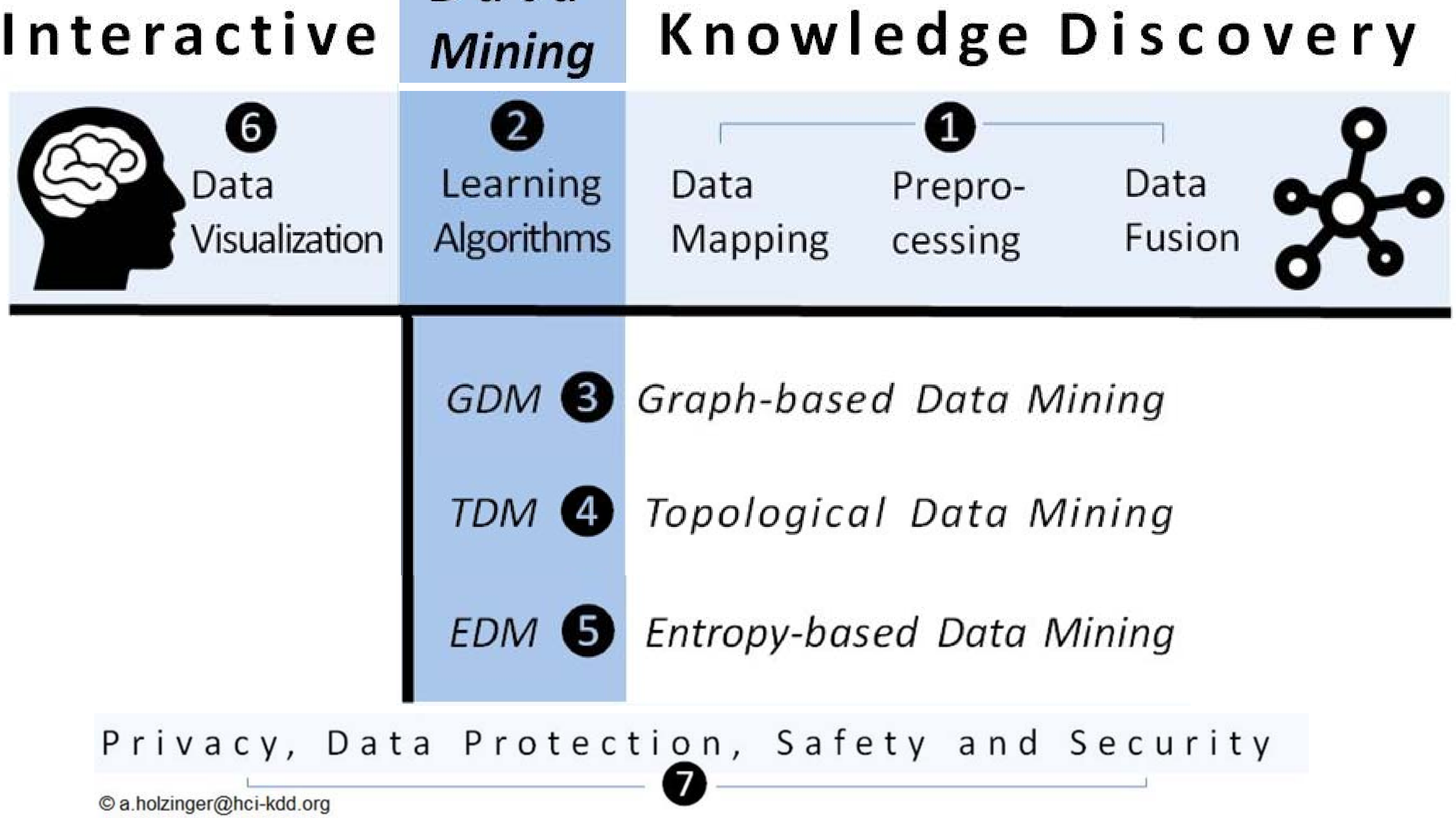
Agenda for the second lecture

- 01 Research Topics/Methods (samples)
- 02 Research Goals (samples)
- 03 Research Questions (samples)
- 04 Research Projects (samples)

01 Research Topics and Methods (samples)

ML needs a concerted effort fostering integrated research

http://hci-kdd.org/international-expert-network



Holzinger, A. 2014. Trends in Interactive Knowledge Discovery for Personalized Medicine: **Cognitive Science meets Machine Learning.** IEEE Intelligent Informatics Bulletin, 15, (1), 6-14.

ML-Jungle Top Level View

Maths

Challenges

Cognition | Visualization | Data structure

Perception | Preprocessing

Decision | Interaction | Integration

Always with a focus/application in health informatics

CONCEPTS	THEORIES	PARADIGMS	MODELS	METHODS	TOOLS
Curse of Dim	Bayesian p(x)	unsupervised	Gaussian P.	Regularization	Python
Dim. Reduction					
NfL-Theorem	Complexity	supervised	Graphical M.	Validation	Julia
Overfitting	KL-Divergence	Semi-supv.	NN	Aggregation	Etc.
			DL		
Non-Parametric	Info Theory	online	SVM	Input Processes	Azure
		iML	Linear Models		
		RL, PL, AL	D. Trees		

Exp. & Eval.

Privacy ML

02 Research Goals (samples)

"Solve intelligence – then solve everything else"

https://youtu.be/XAbLn66iHcQ?t=1h28m54s

Demis Hassabis, 22 May 2015

The Royal Society,
Future Directions of Machine Learning Part 2

Google DeepMind

In medicine we have two different worlds …

Our central hypothesis: Information may bridge this gap

Holzinger, A. & Simonic, K.-M. (eds.) 2011. *Information Quality in e-Health. Lecture Notes in Computer Science LNCS 7058, Heidelberg, Berlin, New York: Springer.*

Multi-Task Learning (MTL)

for improving prediction performance, help to reduce **catastrophic forgetting**

Transfer learning (TL)

is not easy: learning to perform a task by exploiting knowledge acquired when solving previous tasks:

a solution to this problem would have major impact to AI research generally and ML specifically.

Multi-Agent-Hybrid Systems (MAHS)

To include collective intelligence and crowdsourcing and making use of **discrete** models – avoiding to seek perfect solutions – better have a good solution < 5 min.

Key Problems with data in health informatics

- Heterogeneous, distributed, inconsistent data sources (need for **data integration** & fusion) [1]
- **Complex data** (high-dimensionality – challenge of dimensionality reduction and visualization) [2]
- Noisy, uncertain, missing, dirty, and imprecise, imbalanced data (challenge of **pre-processing**)
- The discrepancy between data-information-knowledge (**various definitions**)
- **Big data** sets (manual handling of the data is awkward, and often impossible) [3]

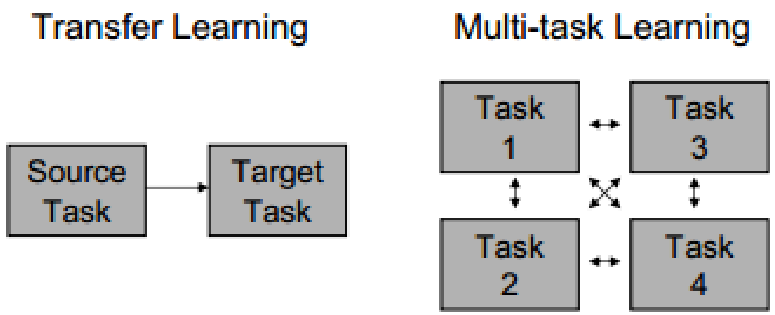
1. Holzinger A, Dehmer M, & Jurisica I (2014) Knowledge Discovery and interactive Data Mining in Bioinformatics - State-of-the-Art, future challenges and research directions. BMC Bioinformatics 15(S6):I1.
2. Hund, M., Sturm, W., Schreck, T., Ullrich, T., Keim, D., Majnaric, L. & Holzinger, A. 2015. Analysis of Patient Groups and Immunization Results Based on Subspace Clustering. In: LNAI 9250, 358-368.
3. Holzinger, A., Stocker, C. & Dehmer, M. 2014. Big Complex Biomedical Data: Towards a Taxonomy of Data. in CCIS 455. Springer 3-18.

Unsolved Problem: Data Integration and Data Fusion in the Life Sciences

How to combine these different data types together to obtain a unified view of the activity in the cell is one of the major challenges of systems biology

Navlakha, S. & Bar-Joseph, Z. 2014. Distributed information processing in biological and computational systems. *Commun. ACM*, 58, (1), 94-102, doi:10.1145/2678280.

- Big data with many training sets (this is good for ML!)
- **Small number of data sets, rare events**
- **Very-high-dimensional problems**
- **Complex data – NP-hard problems**
- **Missing, dirty, wrong, noisy, …, data**
- **GENERALISATION**
- **TRANSFER**

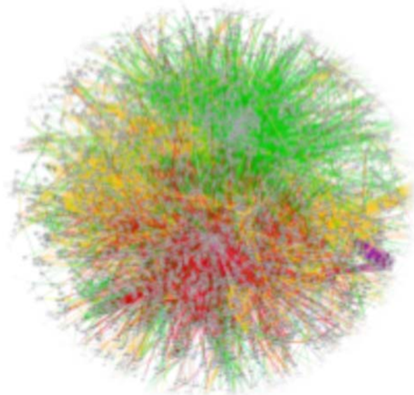


Torrey, L. & Shavlik, J. 2009. Transfer learning. Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques, 242-264, doi:10.4018/978-1-60566-766-9.ch011.

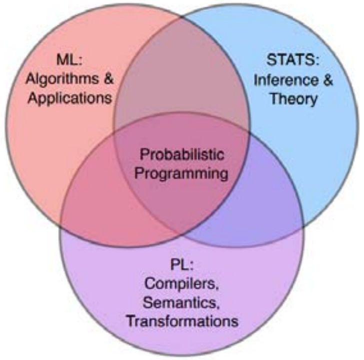
Transfer Learning is studied for more than 100 years

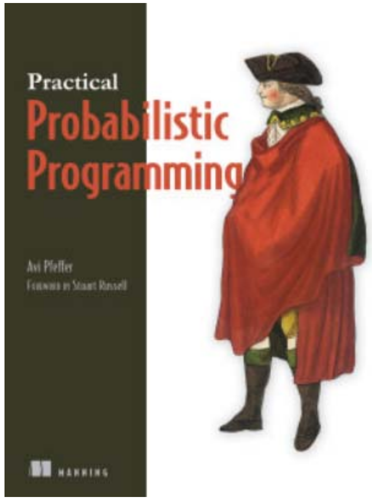
- Thorndike & Woodworth (1901) explored how individuals would transfer in one context to another context that share similar characteristics:
- They explored how individuals would transfer learning in one context to another, similar context
- or how "improvement in one mental function" could influence a related one.
- Their theory implied that transfer of learning depends on how similar the learning task and transfer tasks are,
- or where "identical elements are concerned in the influencing and influenced function", now known as the identical element theory.
- Today example: C++ -> Java; Python -> Julia
- Mathematics -> Computer Science
- Physics -> Economics

- PGM can be seen as a combination between
- Graph Theory + Probability Theory + Machine Learning
- One of the most exciting advancements in AI in the last decades
- Compact representation for exponentially-large probability distributions
- Example Question:
  "Is there a path connecting two proteins?"
- $Path\ (X,Y) :=\ edge\ (X,Y)$
- $Path\ (X,Y) :=\ edge\ (X,Y), path(Z,Y)$
- This can NOT be expressed in first-order logic
- Need a Turing-complete fully-fledged language



Practical Issue: Probabilistic Programming





https://www.manning.com/books/practical-probabilistic-programming

- Probabilistic programs -> functional or imperative programs with two added constructs:
- (1) the ability to draw values at random from distributions, and
- (2) the ability to condition values of variables via observations.
- Models from diverse application areas such as computer vision, coding theory, cryptographic protocols, biology and reliability analysis can be written as probabilistic program …

Gordon, A. D., Henzinger, T. A., Nori, A. V. & Rajamani, S. K. Probabilistic programming. Proceedings of the on Future of Software Engineering, 2014. ACM, 167-181.

03 Research Questions (samples)

Key Challenges

- Medicine is an extremely complex application domain – dealing most of the time with uncertainties -> **probable information!**
- Key: Structure learning and prediction in large-scale biomedical networks with probabilistic graphical models
- Causal and Probabilistic Inference:
- Uncertainties are present at all levels in health related systems
- Data sets from which ML learns are noisy, mislabeled, atypical, etc. etc.
- Even with data of high quality, gauging and combining a multitude of data sources and constraints in usually imperfect models of the world requires us to represent and process uncertain knowledge in order to make viable decisions.
- In the increasingly complicated settings of modern science, model structure or causal relationships may not be known a-priori [1].
- Approximating probabilistic inference in Bayesian belief networks is NP-hard [2] -> here we need the "human-in-the-loop" [3]

[1] Sun, X., Janzing, D. & Schölkopf, B. Causal Inference by Choosing Graphs with Most Plausible Markov Kernels. ISAIM, 2006.
[2] Dagum, P. & Luby, M. 1993. Approximating probabilistic inference in Bayesian belief networks is NP-hard. Artificial intelligence, 60, (1), 141-153.
[3] Holzinger, A. 2016. Interactive Machine Learning for Health Informatics: When do we need the human-in-the-loop? Springer Brain Informatics (BRIN), 3, 1-13, doi:10.1007/s40708-016-0042-6.

Key Challenges

- Uncertainty, Validation, Curse of Dimensionality
- Large spaces gets sparse
- Distance Measures get useless
- Patterns occur in different subspaces
- Central question Nr. 1 “What is interesting?”
- Additional question “What is relevant?”

Why is Reinforcement Learning interesting?

- Reinforcement Learning is the **oldest approach,** with the longest history and can provide insight into understanding human learning [1]
- RL is the **"AI problem in the microcosm"** [2]
- Future opportunities are in Multi-Agent RL (MARL), Multi-Task Learning (MTL), Generalization and **Transfer-Learning** [3], [4].

[1] Turing, A. M. 1950. Computing machinery and intelligence. Mind, 59, (236), 433-460.

[2] Littman, M. L. 2015. Reinforcement learning improves behaviour from evaluative feedback. Nature, 521, (7553), 445-451, doi:10.1038/nature14540.

[3] Taylor, M. E. & Stone, P. 2009. Transfer learning for reinforcement learning domains: A survey. The Journal of Machine Learning Research, 10, 1633-1685.
[4] Pan, S. J. & Yang, Q. A. 2010. A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering, 22, (10), 1345-1359, doi:10.1109/tkde.2009.191.

Key Challenges

- Medicine is an extremely complex application domain – dealing most of the time with uncertainties -> **probable information!**
- When we have big data but little knowledge automatic ML can help to gain insight:
- **Structure learning and prediction in large-scale biomedical networks with probabilistic graphical models**
- If we have little data and deal with NP-hard problems we still need the human-in-the-loop

- Study of the design of **intelligent agents**
- Set of *nature-inspired* methodologies to solve complex real-world problems, when traditional methods might be useless, because:
- 1) the processes might be too **complex** for mathematical reasoning within the given time,
- 2) the problem contains a lot of **uncertainties**
- 3) the problem/process is **stochastic** in nature

Kruse, R., Borgelt, C., Klawonn, F., Moewes, C., Steinbrecher, M. & Held, P. 2013. Computational Intelligence: A Methodological Introduction, Heidelberg, New York, Springer.
Online in both German and English: http://www.computational-intelligence.eu/

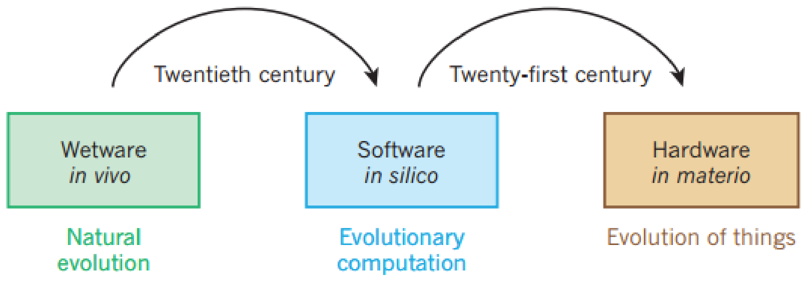
IFIP WG 12.9 http://www.ifip.org/bulletin/bulltcs/memtc12.htm

Key challenges

- Many applications in medical imaging, image segmentation, medical data mining, modelling and simulating medical processes, diagnosis, treatment.
- Whenever a **decision** is required, it is possible to find a niche for evolutionary techniques [1, 2]
- Two relevant (and difficult!) questions:
- 1) For a given problem: what is the best algorithm?
- 2) For a given algorithm: what is the problem to solve?

[1] Pena-Reyes, C. A. & Sipper, M. 2000. Evolutionary computation in medicine: an overview. Artificial Intelligence in Medicine, 19, (1), 1-23, doi:10.1016/S0933-3657(99)00047-0.
[2] Holzinger, K., Palade, V., Rabadan, R. & Holzinger, A. 2014. Darwin or Lamarck? Future Challenges in Evolutionary Algorithms for Knowledge Discovery and Data Mining. In: Lecture Notes in Computer Science LNCS 8401. Heidelberg, Berlin: Springer, pp. 35-56, doi:10.1007/978-3-662-43968-5_3.

Open scientific issues and important research trends



- **Automated design** and tuning of EA for customizing an initial algorithm set-up for a given problem offline (before the run) or online (during the run) and automated parameter tuning
- **Surrogate models:** EA for problems in which evaluating each population member over many generations would take too long to permit effective evolution
- **Multi-objectives** handling at the same time
- **Interactive Evolutionary Algorithms,** bringing in user-preferences, expert knowledge -> human-in-the-loop

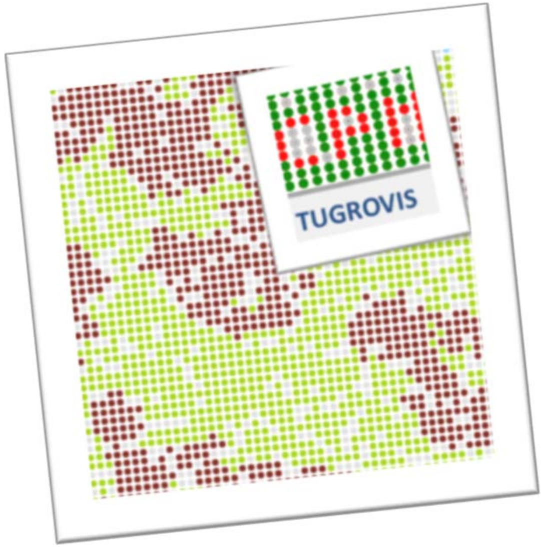
Eiben, A. E. & Smith, J. 2015. From evolutionary computation to the evolution of things. Nature, 521, (7553), 476-482, doi:10.1038/nature14544.

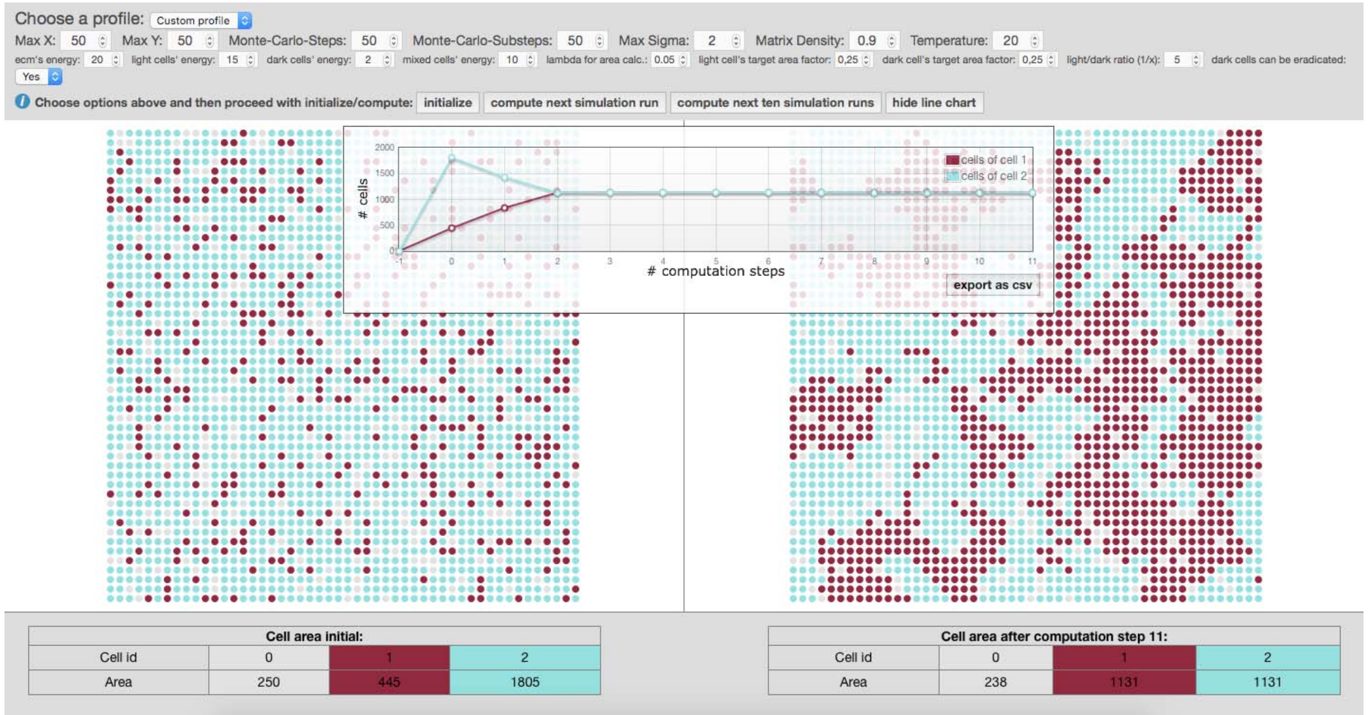
What makes a machine intelligent? Cross-cutting issues

- To hear, to see, to talk
  - Speech recognition, computer vision, natural language processing,
- To store, to represent, to access
  - Knowledge representation, semantic networks, ontologies, information retrieval
- To reason, to understand, to reflect
  - Logic, Bayesian inference, **contextual understanding,**
  - **language understanding**
- To learn from data
  - Improve with experience from previous events, to forecast, to **predict the future**

04 Research Projects (samples)

Project: Tumor-Growth Simulation

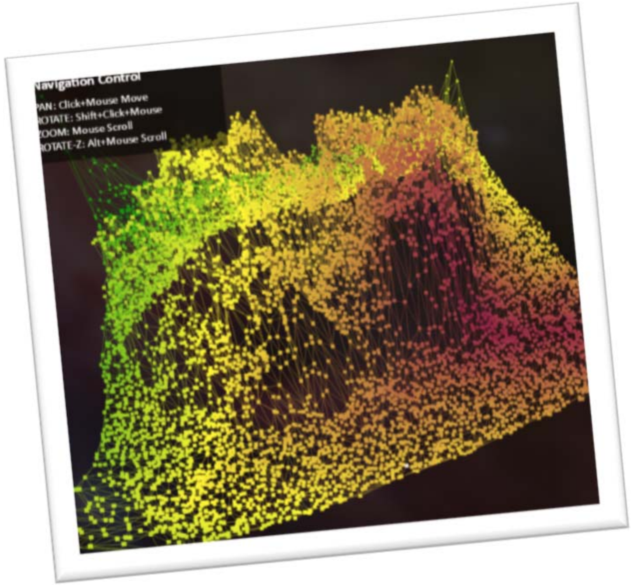


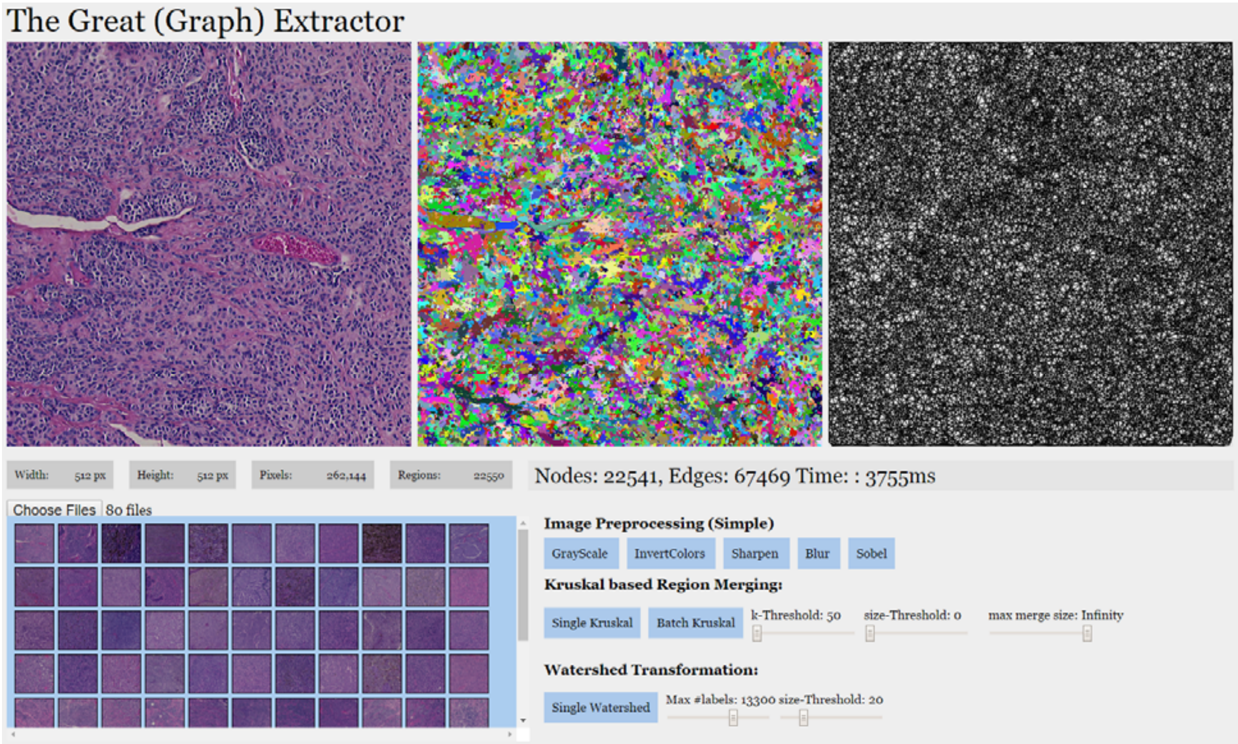


- Contribute to understanding tumor growth
- Goal: Help to Refine → Reduce → Replace
- Towards discrete Multi-Agent Hybrid Systems

Jeanquartier, F., Jean-Quartier, C., Cemernek, D. & Holzinger, A. 2016. In silico modeling for tumor growth visualization. BMC Systems Biology, 10, (1), 1-15, doi:10.1186/s12918-016-0318-8.

Project: Graphinius

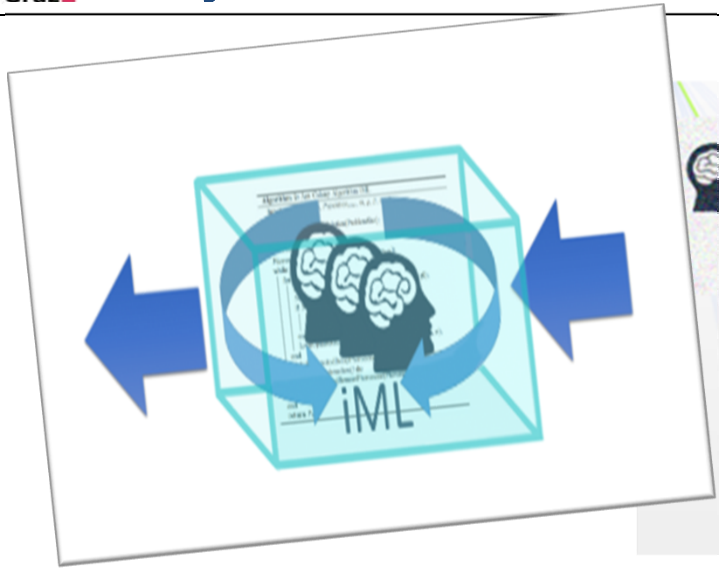




- Contribute to graph understanding and algorithm prototyping by real-time visualization, interaction and manipulation
- Goal: Help to foster ML-on-graphs research replication
- Towards an online graph exploration and analysis platform

Malle, B., Kieseberg, P., Weippl, E. & Holzinger, A. 2016. The right to be forgotten: Towards Machine Learning on perturbed knowledge bases. Springer Lecture Notes in Computer Science LNCS 9817. Heidelberg, Berlin, New York: Springer, pp. 251-256, doi:10.1007/978-3-319-45507-5_17.

Project: iML



- From black-box to glass-box ML
- Exploit human intelligence for solving hard problems (e.g. Subspace Clustering, k-Anonymization, Protein-Design)
- Towards multi-agent systems with humans-in-the-loop

Holzinger, A., Plass, M., Holzinger, K., Crisan, G., Pintea, C. & Palade, V. 2016. Towards interactive Machine Learning (iML): Applying Ant Colony Algorithms to solve the Traveling Salesman Problem with the Human-in-the-Loop approach. Springer Lecture Notes in Computer Science LNCS 9817. Heidelberg, Berlin, New York: Springer, pp. 81-95, doi:10.1007/978-3-319-45507-56.

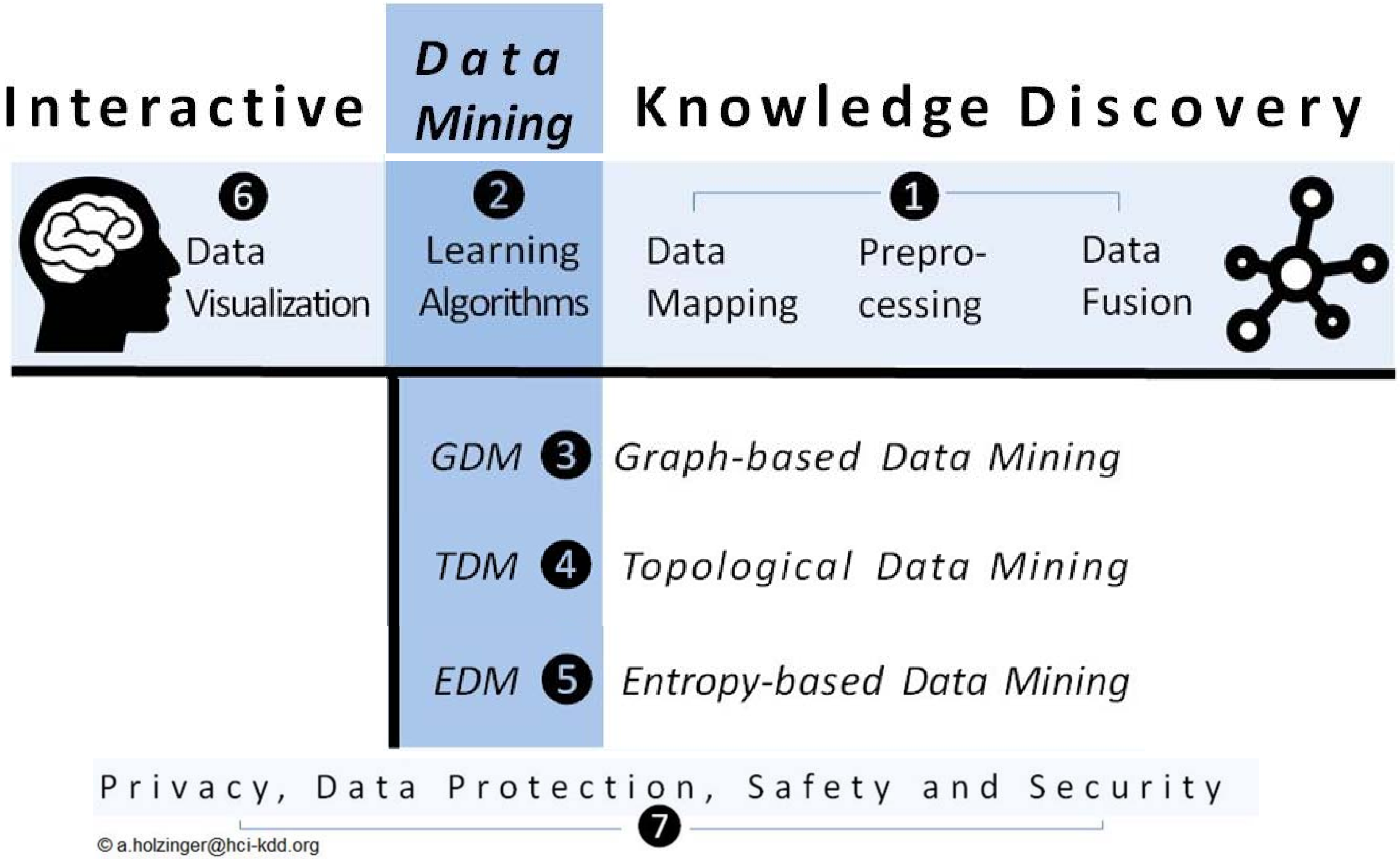
This is only possible in a …

concerted effort

international

without boundaries …

ML-Algorithms are key but needs also concerted effort



Holzinger, A. 2014. Trends in Interactive Knowledge Discovery for Personalized Medicine: Cognitive Science meets Machine Learning. IEEE Intelligent Informatics Bulletin, 15, (1), 6-14.

Thank you!