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
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”
Science is to test crazy ideas – Engineering 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

Interactive

Data Mining

Knowledge Discovery

Privacy, Data Protection, Safety and Security


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

© a.holzinger@hci-kdd.org
Always with a focus/application in health informatics

**CONCEPTS**
- Curse of Dim
- Dim. Reduction
- NfL-Theorem
- Overfitting
- Non-Parametric

**THEORIES**
- Bayesian p(x)
- Complexity
- KL-Divergence
- Info Theory

**PARADIGMS**
- unsupervised
- supervised
- Semi-supv.
- online

**MODELS**
- Gaussian P.
- Graphical M.
- NN
- SVM
- Linear Models

**METHODS**
- Regularization
- Validation
- Aggregation
- Input Processes

**TOOLS**
- Python
- Julia
- Etc.
- Azure

**Challenges**
- Exp. & Eval.
- Privacy ML
- Maths
- Cognition
- Visualization
- Data structure
- Perception
- Preprocessing
- Decision
- Interaction
- Integration

Holzinger Group, HCI-KDD.org
02 Research Goals (samples)
“Solve intelligence – then solve everything else”

Demis Hassabis, 22 May 2015
The Royal Society, Future Directions of Machine Learning Part 2

https://youtu.be/XAbLn66iHcQ?t=1h28m54s
In medicine we have two different worlds ... 

Our central hypothesis:
Information may bridge this gap

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]

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

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

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
Probabilistic Graphical Models

- 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?”

- $\text{Path} (X, Y) := \text{edge} (X, Y)$
- $\text{Path} (X, Y) := \text{edge} (X, Y), \text{path}(Z, Y)$
- This can NOT be expressed in first-order logic
- Need a Turing-complete fully-fledged language
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 ...
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].

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].

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
Key Challenges

- 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

Online in both German and English: http://www.computational-intelligence.eu/

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

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?


- **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

---

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 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

- 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

- 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

This is only possible in a concerted effort international without boundaries ...
ML-Algorithms are key but needs also concerted effort

Interactive

Data Mining

Knowledge Discovery

1. Data Mapping
2. Learning Algorithms
3. Graph-based Data Mining
4. Topological Data Mining
5. Entropy-based Data Mining

Privacy, Data Protection, Safety and Security


Holzinger Group, HCI-KDD.org

Student Seminar, Winter 2016
Thank you!