

Science is to test crazy ideas – Engineering is to put these ideas into Business HCI-KDD

Andreas Holzinger
185.A83 Machine Learning for Health Informatics
2017S, VU, 2.0 h, 3.0 ECTS
Module 00 - 07.03.2016

MAKE Health
Machine Learning & Knowledge Extraction
in health informatics: challenges & directions
a.holzinger@hci-kdd.org
<http://hci-kdd.org/machine-learning-for-health-informatics-course>



Holzinger Group, HCI-KDD.org 1 MAKE Health 00

This lecture is only the overview and motivation part HCI-KDD

- The HCI-KDD approach: integrative ML
- Understanding Intelligence
- Complexity of the health domain
- Probabilistic information
- Automatic Machine Learning (aML)
- Interactive Machine Learning (iML)
- Active Representation Learning
- Multi-Task Learning
- Generalization and Transfer Learning

Holzinger Group, HCI-KDD.org 2 MAKE Health 00

01 What is the
HCI-KDD
approach?

Holzinger Group, HCI-KDD.org 3 MAKE Health 00



ML is a very practical field –
algorithm development is at the core –
however,
successful ML needs a concerted effort of
various topics ...



Holzinger Group, HCI-KDD.org 4 MAKE Health 00

Machine Learning and Knowledge Extraction Pipeline HCI-KDD

01

Interactive	Data Mining	Knowledge Discovery
 <p>6 Data Visualization</p>	<p>2 Learning Algorithms</p>	<p>1 Data Mapping Pre-processing Data Fusion</p> 
	<p>GDM 3 Graph-based Data Mining</p> <p>TDM 4 Topological Data Mining</p> <p>EDM 5 Entropy-based Data Mining</p>	
<p>Privacy, Data Protection, Safety and Security</p> <p>7</p>		

© a.holzinger@hci-kdd.org

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

Holzinger Group, HCI-KDD.org 5 MAKE Health 00

Knowledge Extraction is necessary as first step ... HCI-KDD

Features are key to
learning and
understanding

Holzinger Group, HCI-KDD.org 6 MAKE Health 00

... successful ML needs ... HCI-KDD

concerted effort

international
without boundaries ...

<http://www.bach-cantatas.com>



Holzinger Group, HCI-KDD.org 7 MAKE Health 00




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

Holzinger Group, HCI-KDD.org 8 MAKE Health 00



Cognitive Science AND Computer Science HCI-KDD

02





- Cognitive Science → human intelligence
- Computer Science → computational intelligence
- Human-Computer Interaction → the bridge

Holzinger Group, HCI-KDD.org 9 MAKE Health 00

02 Solve Intelligence then solve everything else




Grand Goal: Understanding Intelligence
03

“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>





To reach a level of usable intelligence we need to ...
04

- 1) extract knowledge
- 2) learn from prior data
- 3) generalize, i.e. guessing where a probability measure concentrates
- 4) fight the curse of dimensionality
- 5) disentangle underlying explanatory factors of data, i.e.
- 6) understand the data in the context of an application domain

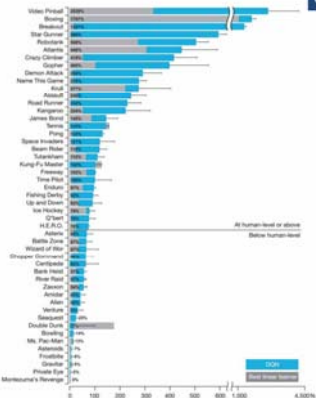









How far are we already ?
05

Compare your best ML algorithm with a seven year old child ...

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S. & Hassabis, D. 2015. Human-level control through deep reinforcement learning. Nature, 518, (7540), 529-533, doi:10.1038/nature14236





Scientists who pleaded for “humanoid AI”

- Alan Turing (1912 – 1954)
- Herbert Simon (1916 – 2001)
- John McCarthy (1927 – 2011)
- Marvin Minsky (1927 – 2016)
- Allen Newell (1927 – 1992)
- ... pleaded for building machines that can learn similar to humans, e.g. like children
- None of them knew what they were talking about ... (Josh Tenenbaum)**




Not our Goal: Humanoid AI












03 Application Area Health Informatics







Health is a complex area

Why is this application area complex ?







Machine Learning and Health Informatics!



“Medicine is so complex, the challenges are so great... we need everything that we can bring to make our diagnostics more precise, more accurate and our therapeutics more focused on that patient!”
Sir Maudslayi House, North England

<https://royalsociety.org/events/2015/05/breakthrough-science-technologies-machine-learning>

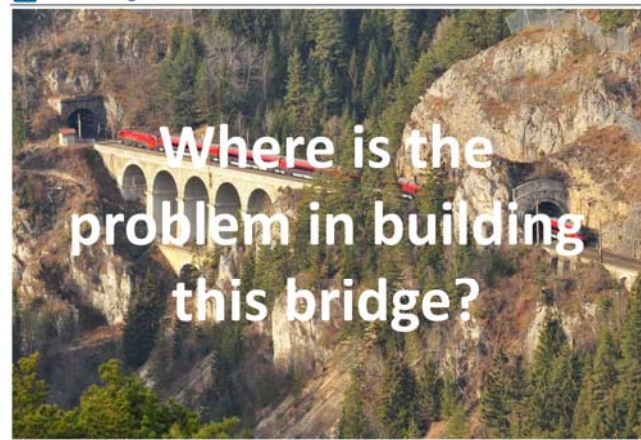




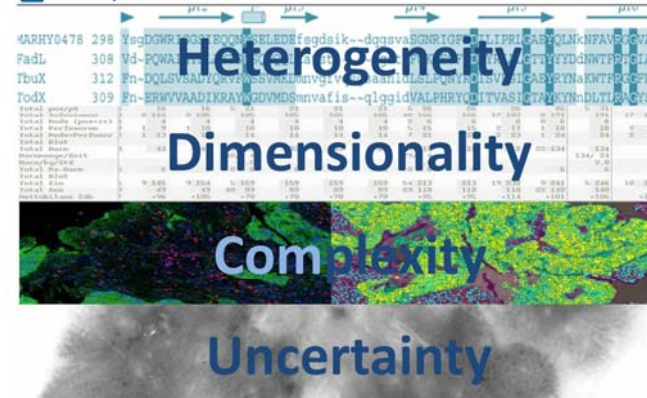

Our central hypothesis: Information may bridge this gap

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

Holzinger Group, HCI-KDD.org 19 MAKE Health 00



Holzinger Group, HCI-KDD.org 20 MAKE Health 00



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), 11.

Holzinger Group, HCI-KDD.org 21 MAKE Health 00

04 Probabilistic Information $p(x)$

Holzinger Group, HCI-KDD.org 22 MAKE Health 00

Probability theory
is nothing but
common sense
reduced to
calculation ...



Pierre Simon de Laplace (1749-1827), 1812

Holzinger Group, HCI-KDD.org 23 MAKE Health 00

What is the simplest mathematical operation for us?

$$p(x) = \sum_y p(x, y) \quad (1)$$

How do we call repeated adding?

$$p(x, y) = p(y|x) * p(x) \quad (2)$$

Laplace (1773) showed that we can write:

$$p(x, y) * p(y) = p(y|x) * p(x) \quad (3)$$

Now we introduce a third, more complicated operation:

$$\frac{p(x, y) * p(y)}{p(y)} = \frac{p(y|x) * p(x)}{p(y)} \quad (4)$$

We can reduce this fraction by $p(y)$ and we receive what is called Bayes rule:

$$p(x, y) = \frac{p(y|x) * p(x)}{p(y)} \quad p(h|d) = \frac{p(d|h)p(h)}{p(d)} \quad (5)$$

Holzinger Group, HCI-KDD.org 24 MAKE Health 00



Thomas Bayes
1701 - 1761



Richard Price
1723-1791

Bayes, T. (1763). An Essay towards solving a Problem in the Doctrine of Chances (Postum communicated by Richard Price). *Philosophical Transactions*, 53, 370-418.

$$p(x_i) = \sum_j P(x_i, y_j) \quad p(x_i, y_j) = p(y_j|x_i)P(x_i)$$

Bayes' Rule is a corollary of the Sum Rule and Product Rule:

$$p(x_i|y_j) = \frac{p(y_j|x_i)p(x_i)}{\sum_i p(x_i, y_j)p(x_i)}$$

Barnard, G. A., & Bayes, T. (1958). Studies in the history of probability and statistics: IX. Thomas Bayes's essay towards solving a problem in the doctrine of chances. *Biometrika*, 45(3/4), 293-315.

Holzinger Group, HCI-KDD.org 25 MAKE Health 00



Nicolas Poussin, 1658, Oil on canvas, Metropolitan Museum of Art, New York

Holzinger Group, HCI-KDD.org 26 MAKE Health 00



- Newton, Leibniz, ... developed calculus – mathematical language for describing and dealing with rates of change
- Bayes, Laplace, ... developed probability theory - the mathematical language for describing and dealing with uncertainty
- Gauss generalized those ideas

Holzinger Group, HCI-KDD.org 27 MAKE Health 00

04

Posterior Probability

$$p(\mathcal{D}|\theta)$$

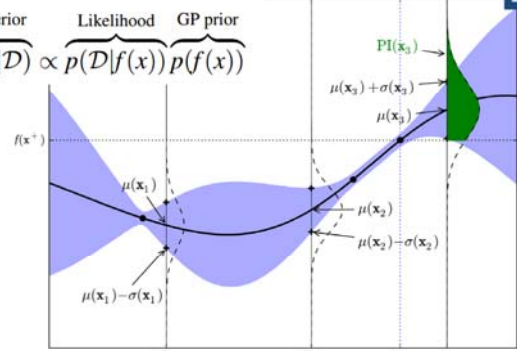
$$posterior = \frac{likelihood * prior}{evidence}$$

The inverse probability allows to learn from data, infer unknowns, and make predictions

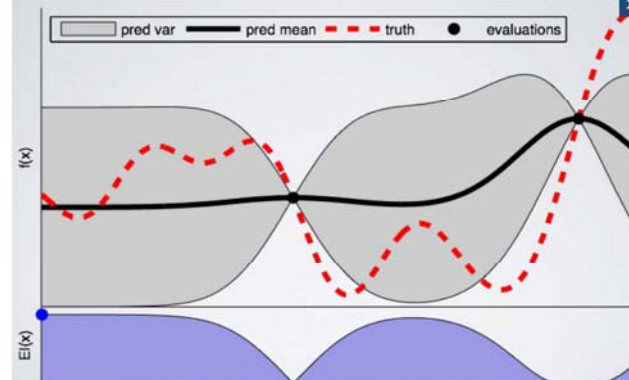
Figure 1 illustrates the REMBO algorithm. The left side shows three panels for $n=2$, $n=3$, and $n=4$, depicting the iterative process of observing points x_1, x_2, x_3 and updating the acquisition function $f(x; I)$ and posterior uncertainty. The right side shows two plots of the Optimal Gap versus the No. of Iterations (t). The top plot compares BO (black dotted line), REMBO ($d=1$) (red solid line), and REMBO ($d=2, k=1$) interleaved runs (green dashed line). The bottom plot compares BO (black dotted line), REMBO ($d=1$) (red solid line), and REMBO ($d=2, k=1$) interleaved runs (green dashed line). The plots show that REMBO methods achieve a faster reduction in the optimal gap compared to BO.

Wang, Z., Hutter, F., Zoghi, M., Matheson, D. & De Freitas, N. 2016. Bayesian optimization in a billion dimensions via random embeddings. *Journal of Artificial Intelligence Research*, 55, 361-387, doi:10.1613/jair.4806.

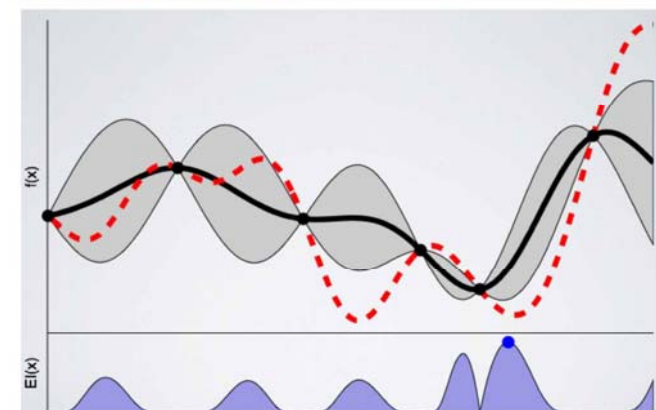
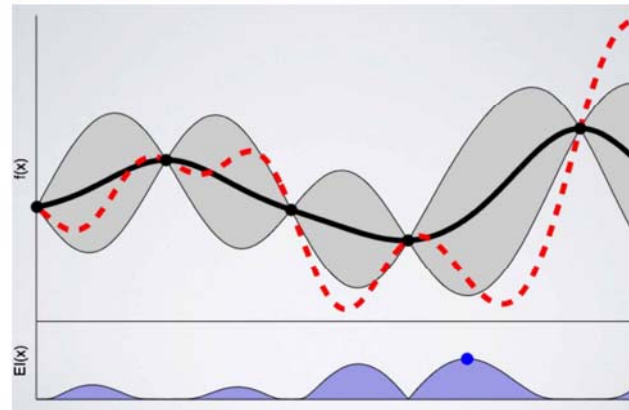
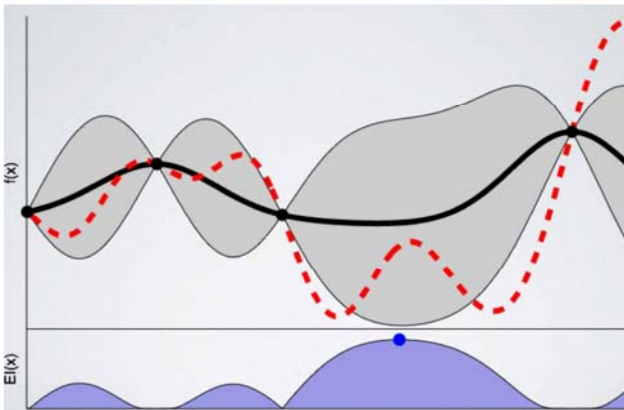
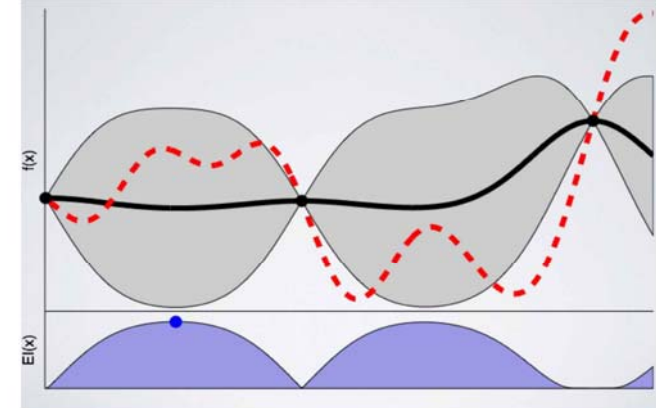
11

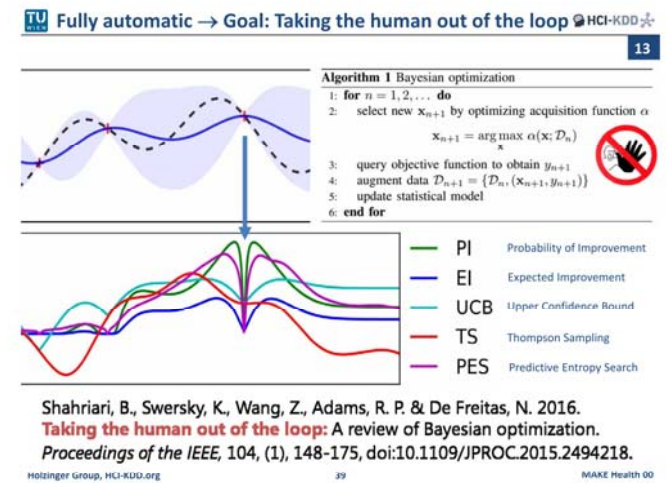
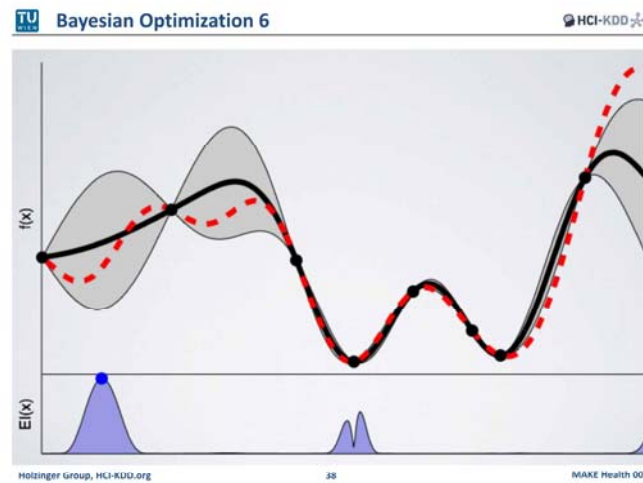
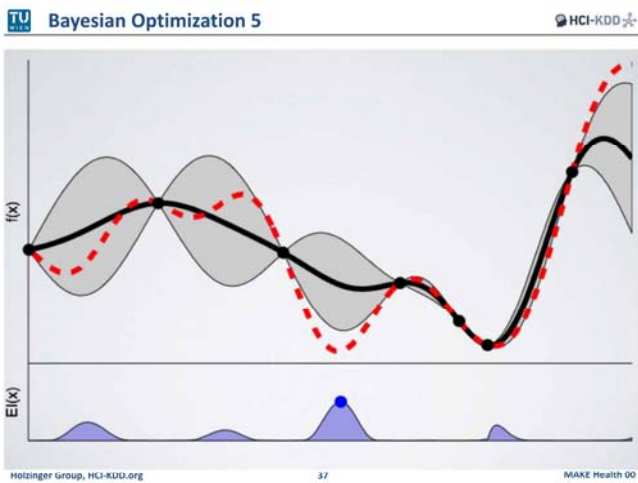


Brochu, E., Cora, V. M. & De Freitas, N. 2010. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. arXiv:1012.2599.



Snoek, J., Larochelle, H. & Adams, R. P. Practical Bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, 2012. 2951-2959.





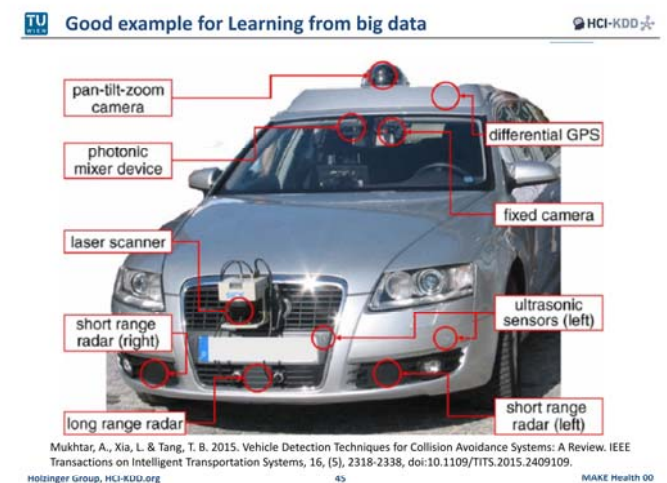
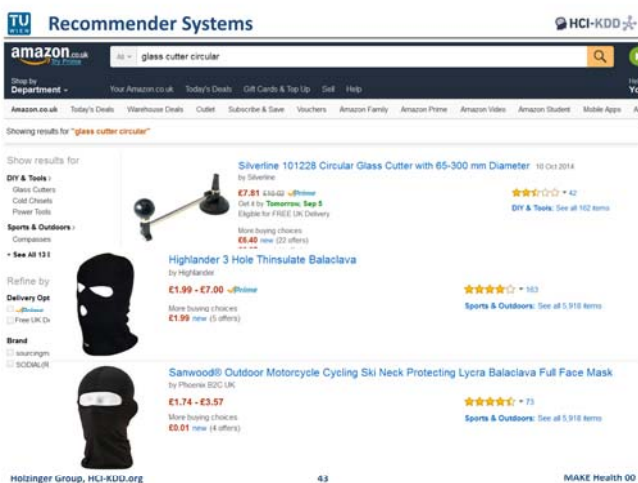
05 aML

Holzinger Group, HCI-KDD.org 40 MAKE Health 00

- Everything is machine learning ...** HCI-KDD
- Today most ML-applications are using automatic Machine Learning (aML) approaches
 - aML := algorithms which interact with agents and can optimize their learning behaviour through this interaction
- Holzinger Group, HCI-KDD.org 41 MAKE Health 00

Best practice examples of aML ...

Holzinger Group, HCI-KDD.org 42 MAKE Health 00





This Citroën DS with "automated steering" was tested in the early 1960s...

1960s Citroën DS driverless car test

Sunday Times Driving 10

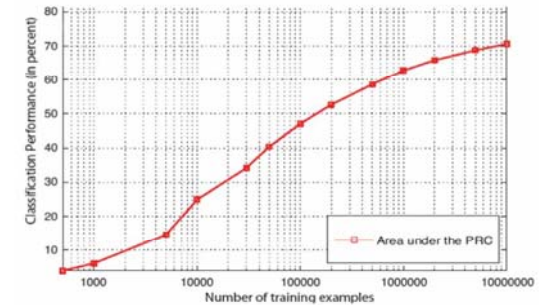
8,605 views

Cyber-Physical Systems (CPS):

Tight integration of networked computation with physical systems



Seshia, S. A., Juniwal, G., Sadigh, D., Donze, A., Li, W., Jensen, J. C., Jin, X., Deshmukh, J., Lee, E. & Sastry, S. 2015. Verification by, for, and of Humans: Formal Methods for Cyber-Physical Systems and Beyond. Illinois ECE Colloquium.



Sonnenburg, S., Rätsch, G., Schäfer, C. & Schölkopf, B. 2006. Large scale multiple kernel learning. Journal of Machine Learning Research, 7, (7), 1531-1565.

- Sometimes we **do not have "big data"**, where aML-algorithms benefit.
- Sometimes we have
 - Small amount of data sets**
 - Rare Events – no training samples**
 - NP-hard problems, e.g.**
 - Subspace Clustering,
 - k-Anonymization,
 - Protein-Folding, ...

06 iML

- iML := algorithms which interact with agents*) and can optimize their learning behaviour through this interaction
- *) where the agents can be human**

Holzinger, A. 2016. Interactive Machine Learning (iML). Informatik Spektrum, 39, (1), 64-68, doi:10.1007/s00287-015-0941-6.



aML: taking the human-out-of-the-loop HCI-KDD

A) Unsupervised ML: Algorithm is applied on the raw data and learns fully automatic – Human can check results at the end of the ML-pipeline

B) Supervised ML: Humans are providing the labels for the training data and/or select features to feed the algorithm to learn – the more samples the better – Human can check results at the end of the ML-pipeline

C) Semi-Supervised Machine Learning: A mixture of A and B – mixing labeled and unlabeled data so that the algorithm can find labels according to a similarity measure to one of the given groups

Holzinger Group, HCI-KDU.org 55 MAKE Health 00

iML: bringing the human-in-the-loop HCI-KDD

D) **Interactive Machine Learning:** Human is seen as an agent involved in the actual learning phase, step-by-step influencing measures such as distance, cost functions ...

4. Check 2. Preprocessing 1. Input 3. iML

Constraints of humans: Robustness, subjectivity, transfer?
Open Questions: Evaluation, replicability, ...

Holzinger, A. 2016. Interactive Machine Learning for Health Informatics: When do we need the human-in-the-loop? Brain Informatics (BRIN), 3, (2), 119-131, doi:10.1007/s40708-016-0042-6.

Holzinger Group, HCI-KDU.org 56 MAKE Health 00

Three examples for the usefulness of the iML approach HCI-KDD

- Example 1: Subspace Clustering**
- Example 2: k-Anonymization**
- Example 3: Protein Design**

Hund, M., Böhm, D., Sturm, W., Sedlmair, M., Schreck, T., Ullrich, T., Keim, D. A., Majnarić, L. & Holzinger, A. 2016. Visual analytics for concept exploration in subspaces of patient groups: Making sense of complex datasets with the Doctor-in-the-loop. Brain Informatics, 1-15, doi:10.1007/s40708-016-0043-5.

Kieseberg, P., Malle, B., Fruehwirt, P., Weippl, E. & Holzinger, A. 2016. A tamper-proof audit and control system for the doctor in the loop. Brain Informatics, 3, (4), 269-279, doi:10.1007/s40708-016-0046-2.

Lee, S. & Holzinger, A. 2016. Knowledge Discovery from Complex High Dimensional Data. In: Michaelis, S., Piatkowski, N. & Stolpe, M. (eds.) Solving Large Scale Learning Tasks. Challenges and Algorithms, Lecture Notes in Artificial Intelligence LNAI 9580. Springer, pp. 148-167, doi:10.1007/978-3-319-41706-6_7.

Holzinger Group, HCI-KDU.org 57 MAKE Health 00

Project: iML HCI-KDD

Experiment: Interactive Machine Learning (iML) for the Traveling-Salesman Problem (TSP)

- 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.

Holzinger Group, HCI-KDU.org 58 MAKE Health 00

http://hci-kdd.org/projects/iml-experiment HCI-KDD

```

Input : ProblemSize, m, β, ρ, σ, q0
Output: Pbest
Pbest ← CreateHeuristicSolution(ProblemSize);
PbestCost ← Cost(Pbest);
Pheromoneinit ←  $\frac{1.0}{ProblemSize \times P_{bestCost}}$ ;
Pheromone ← InitializePheromone(Pheromoneinit);
while ¬StopCondition() do
  for i = 1 to m do
    Si ← ConstructSolution(Pheromone, ProblemSize, β, q0);
    SiCost ← Cost(Si);
    if SiCost ≤ PbestCost then
      PbestCost ← SiCost;
      Pbest ← Si;
    end
    LocalUpdateAndDecayPheromone(Pheromone, Si, SiCost, ρ);
  end
  GlobalUpdateAndDecayPheromone(Pheromone, Pbest, PbestCost, ρ);
  while isUserInteraction() do
    GlobalAddAndRemovePheromone(Pheromone, Pbest, PbestCost, ρ);
  end
end
return Pbest;

```

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. 81-95, doi:10.1007/978-3-319-45507-56.

Holzinger Group, HCI-KDU.org 59 MAKE Health 00

Example: Discovery of causal relationships from data ... HCI-KDD

Hans Holbein d.J., 1533, The Ambassadors, London: National Gallery

Lopez-Paz, D., Muandet, K., Schölkopf, B. & Tolstikhin, I. 2015. Towards a learning theory of cause-effect inference. Proceedings of the 32nd International Conference on Machine Learning, JMLR, Lille, France.

<https://www.youtube.com/watch?v=9KiVNIUMmCc>

Holzinger Group, HCI-KDU.org 60 MAKE Health 00

The grand question of cognitive science HCI-KDD

15a

- How get our mind so much out of so little?**
 - Our minds build rich models of the world
 - make strong generalizations
 - from input data that is sparse, noisy, and ambiguous – in many ways far too limited to support the inferences we make
- How do we do it?
- ... we do not know yet ...

Tenenbaum, J. B., Kemp, C., Griffiths, T. L. & Goodman, N. D. 2011. How to grow a mind: Statistics, structure, and abstraction. Science, 331, (6022), 1279-1285, doi:10.1126/science.1192788.

Holzinger Group, HCI-KDU.org 61 MAKE Health 00

The grand question of cognitive science HCI-KDD

07 Active Representation Learning

Holzinger Group, HCI-KDU.org 62 MAKE Health 00

The grand question of cognitive science HCI-KDD

15b

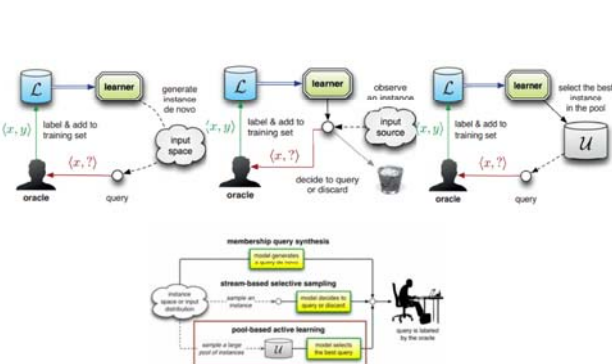
- "How do humans generalize from very few examples?"**
- They transfer knowledge from previous learning:
 - Representation learning (features!)
 - Explanatory factors
 - Previous learning from unlabeled data and labels for other tasks
- Prior: shared underlying explanatory factors, in particular between $P(x)$ and $P(Y|X)$, with a causal link between $Y \rightarrow X$

Bengio, Y., Courville, A. & Vincent, P. 2013. Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence, 35, (8), 1798-1828, doi:10.1109/TPAMI.2013.50.

Holzinger Group, HCI-KDU.org 63 MAKE Health 00

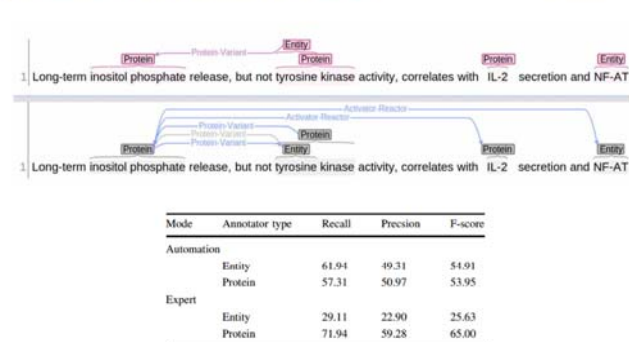


Scenarios for active learning



Settles, B. 2012. Active Learning, San Rafael (CA), Morgan & Claypool, doi:10.2200/S00429ED1V01Y201207AIM018.

Example for the Human-in-the-Loop



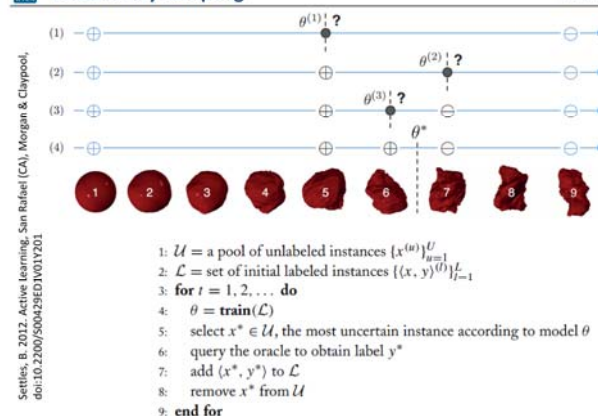
Yimam, S. M., Biemann, C., Majnarić, L., Šabanović, Š. & Holzinger, A. 2016. An adaptive annotation approach for biomedical entity and relation recognition. Brain Informatics, 1-12, doi:10.1007/s40708-016-0036-4.

Active Learning – study of ML that improve by asking ...

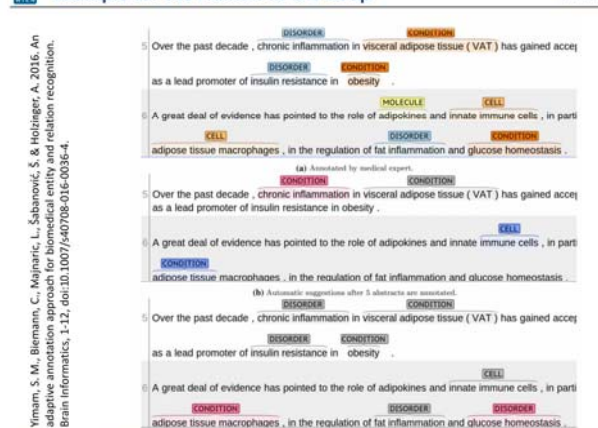
- ML algorithm can perform better with less training if it is allowed to choose the data from which it learns.
- “Active learner” may pose queries, usually in the form of unlabeled data instances to be labeled by an “oracle” (e.g., a human annotator) that **understands the context** of the problem.
- It is useful, where unlabeled data is abundant or easy to obtain, but training labels are difficult, time-consuming, or expensive to obtain ...

Settles, B. 2012. Active Learning, San Rafael (CA), Morgan & Claypool, doi:10.2200/S00429ED1V01Y201207AIM018.

Uncertainty Sampling



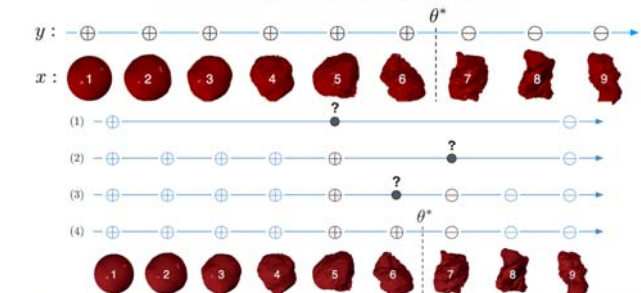
Example for the Human-in-the-Loop



Goal: Automating Inquiries (Settles: alien fruits)

- A classifier to determine objects as a function mapping $h: X \rightarrow Y$, parameterized by a threshold θ :

$$h(x; \theta) = \begin{cases} \oplus \text{ safe} & \text{if } x < \theta, \text{ and} \\ \ominus \text{ noxious} & \text{otherwise.} \end{cases}$$



From Active learning to Multi-Task Active learning

- The typical active learning setting assumes a single machine learner trying to solve a single task.
- In real-world problems, however, the same data might be labeled in multiple ways for several different subtasks.
- In such cases, it is more economical to label a single instance for all subtasks simultaneously, or to choose instance-task query pairs that provide as much information as possible to all tasks.

Settles, B. 2012. Active Learning, San Rafael (CA), Morgan & Claypool, doi:10.2200/S00429ED1V01Y201207AIM018.

08 Multi-Task Learning

Catastrophic Forgetting HCI-KDD

20

- When trained on one task, then trained on a 2nd task, many machine learning models ("deep learning"!) forget how to perform the first task.

Overcoming catastrophic forgetting in neural networks

James Kirkpatrick¹, Razvan Pascanu¹, Neil Rabinowitz¹, Joel Veness¹, Guillaume Desjardins¹, Andre A. Rusu¹, Kieran Milan¹, John Quan¹, Tiago Ramalho¹, Agnieszka Grzibka-Barwinska¹, Dennis Hassabis¹, Claudia Clopath¹, Dharshan Kumaran¹, and Raia Hadsell¹

¹DeepMind, London, N1C 4AG, United Kingdom
Engineering department, Imperial College London, SW7 2AZ, London, United Kingdom

Abstract

The ability to learn tasks in a sequential fashion is crucial to the development of artificial intelligence. Neural networks are not, in general, capable of this and it has been widely thought that catastrophic forgetting is an inevitable feature of connectionist models. We show that it is possible to overcome this limitation and train networks that can maintain expertise on tasks which they have not experienced for a long time. Our approach remembers old tasks by selectively slowing down learning on the weights important for those tasks. We demonstrate our approach is scalable and effective by solving a set of classification tasks based on the MNIST hand-written digit dataset and by learning several Atari 2600 games sequentially.

2 [cs.LG] 25 Jan 2017

Holzinger Group, HCI-KUO.org

"Old" Phenomenon HCI-KDD

Review

French - Catastrophic forgetting

Catastrophic forgetting in connectionist networks

Robert M. French

All natural cognitive systems, and, in particular, our own, gradually forget previously learned information. Plausible models of human cognition should therefore exhibit similar patterns of gradual forgetting of old information as new information is acquired. Only rarely does new learning in natural cognitive systems completely disrupt or erase previously learned information; that is, natural cognitive systems do not, in general, forget 'catastrophically'. Unfortunately, though, catastrophic forgetting does occur under certain circumstances in distributed connectionist networks. The very features that give these networks their remarkable abilities to generalize, to function in the presence of degraded input, and so on, are found to be the root cause of

Holzinger Group, HCI-KUO.org

Overcoming Catastrophic Forgetting: Deep Learning Bayes HCI-KDD

20

Parameter Space task A

Parameter Space task B

Low error for task B
Low error for task A
EWC
L2
no penalty

$\log p(\theta|\mathcal{D}) = \log p(\mathcal{D}|\theta) + \log p(\theta) - \log p(\mathcal{D})$

$\log p(\theta|\mathcal{D}) = \log p(\mathcal{D}_B|\theta) + \log p(\theta|\mathcal{D}_A) - \log p(\mathcal{D}_B)$

$\mathcal{L}(\theta) = \mathcal{L}_B(\theta) + \sum_i \frac{\lambda}{2} F_i(\theta_i - \theta_{A,i}^*)^2$

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D. & Hadsell, R. 2016. Overcoming catastrophic forgetting in neural networks. arXiv preprint arXiv:1612.00796.

Holzinger Group, HCI-KUO.org

This experiment (2016) was done with Atari games ... HCI-KDD

21

Input

Video-game environment

Game controller action values

Hidden layers

Image convolutions

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S. & Hassabis, D. 2015. Human-level control with deep reinforcement learning. Nature, 518, (7540), 529-533, doi:10.1038/nature14236

Holzinger Group, HCI-KUO.org

Example for Multi-Task Learning HCI-KDD

Task 1

Task 2

Task 3

Task 4

INPUTS

V. Mnih et al., "Playing Atari with Deep Reinforcement Learning", Nature (2015)

Rich Caruana, "Multi-task Learning", ML (1998)

Holzinger Group, HCI-KUO.org

Representation Learning discovering explanatory factors HCI-KDD

21

output

Task A

Task B

Task C

shared subsets of factors

input

Bengio, Y., Courville, A. & Vincent, P. 2013. Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence, 35, (8), 1798-1828, doi:10.1109/TPAMI.2013.50.

Holzinger Group, HCI-KUO.org

Maps between shared representations HCI-KDD

22

- x and y represent different modalities, e.g. text, sound, images, ...
- Generalization to new categories
- Larochelle et al. (2008) AAAI

$h_x = f_x(x)$

$h_y = f_y(y)$

x -space

y -space

f_x

f_y

$\mathbf{x}_{\text{train}}$

$\mathbf{y}_{\text{train}}$

\mathbf{x}_{test}

\mathbf{y}_{test}

--- (x, y) pairs in the training set

→ x -representation (encoder) function f_x

→ y -representation (encoder) function f_y

→ relationship between embedded points within one of the domains

→ maps between representation spaces

Goodfellow, I., Bengio, Y. & Courville, A. 2016. Deep Learning, Cambridge: MIT Press, p.542

Holzinger Group, HCI-KUO.org

Big Problem: Real-world data is on Curved Manifolds ! HCI-KDD

23

manifold

rotation transformation of a binary image

4

4

4

4

local linear patches tangent to the manifold

shrinking transformation

$\partial e_k(x) / \partial x$

raw input vector space

x_1

x_2

x_n

Bengio, Y., Monperrus, M. & Larochelle, H. 2006. Nonlocal estimation of manifold structure. Neural Computation, 18, (10), 2509-2528, doi:10.1162/neco.2006.18.10.2509.

Holzinger Group, HCI-KUO.org

SCL Blitzer et al. (2006) of the Weinberger Group HCI-KDD

24

Input: labeled source data $\{(x_i, y_i)\}_{i=1}^n$, unlabeled data from both domains $\{x_j\}$

Output: predictor $f: X \rightarrow Y$

1. Choose m pivot features. Create m binary prediction problems, $p_\ell(x)$, $\ell = 1, \dots, m$
2. For $\ell = 1$ to m
 - $\hat{w}_\ell = \argmin_w \left(\sum_j L(w \cdot x_j, p_\ell(x_j)) + \lambda \|w\|^2 \right)$
 - end
3. $W = [\hat{w}_1, \dots, \hat{w}_m]$, $[U, D, V^T] = \text{SVD}(W)$, $\theta = U_{1:m, :}$
4. Return f , a predictor trained on $\left\{ \left(\begin{bmatrix} x_i \\ \theta x_i \end{bmatrix}, y_i \right) \right\}_{i=1}^n$

a) Heuristically choose m pivot features, which is task specific.

b) Transform each vector of the pivot feature to a vector of binary values and then create the corresponding prediction problem.

Learn the parameters of each prediction problem

Do Eigen Decomposition on the matrix of parameters and learn the linear mapping function.

Use the learnt mapping function to construct new features and train classifiers onto the new representations.

Holzinger Group, HCI-KUO.org

- 1) Challenges include –omics data analysis, where KL divergence and related concepts could provide important measures for discovering biomarkers.
- 2) Hot topics are new entropy measures suitable for computations in the context of complex/uncertain data for ML algorithms.
- Inspiring is the abstract geometrical setting underlying ML main problems, e.g. Kernel functions can be completely understood in this perspective. Future work may include **entropic concepts and geometrical settings**

- 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**



Thank you!

Questions

- What is the HCI-KDD approach?
- What is meant by “integrative ML”?
- Why is a direct integration of AI-solutions into the workflow important?
- What are features?
- Why is understanding intelligence important?
- What are currently (state-of-the-art) the best algorithms?
- What is the difference between Humanoid AI and Human-Level AI?
- Why is the health domain probably the most complex application domain for machine learning?

- Why are we speaking about “two different worlds” in the medical domain?
- Where is the problem in building the bridge between those two worlds?
- Why is the work of Bayes so important for machine learning?
- Why are Newton/Leibniz, Bayes/Laplace and Gauss so important for machine learning?
- What is learning and inference?
- What is the inverse probability?
- How does Bayesian optimization in principle work?

- What is the definition of aML?
- What is the best practice of aML?
- Why is “big data” necessary for aML?
- Provide examples for rare events!
- Give examples for NP-hard problems relevant for health informatics!
- Give the definition of iML?
- What is the benefit of a “human-in-the-loop”?
- Explain the differences of iML in contrast to supervised and semi-supervised learning!

- What is causal relationship from purely observational data and why is it important?
- What is generalization?
- Why is understanding the context so important?
- What does the oracle in Active learning do?
- Explain catastrophic forgetting!
- Give an example for multi-task learning!
- What is the goal of transfer learning and why is this important for machine learning?
- Why would a contribution to a solution to transfer learning be a major breakthrough for artificial intelligence in general – and machine learning specifically?

Appendix

- Active Learning
- Bayesian inference, Bayesian Learning
- Gaussian Processes
- Graphical Models
- Multi-Task Learning
- Reinforcement Learning
- Statistical Learning
- Transfer Learning
- Multi-Agent Hybrid Systems

- *"The most interesting facts are those which can be used several times, those which have a chance of recurring ..."*
- *which, then, are the facts that have a chance of recurring?*
- *In the first place, **simple** facts."*



Jules Henri Poincaré (1854–1912)

Henri Poincare, Sciences et Methods (1908)

Humanoid AI

≠

Human-level AI

- Bernhard Schölkopf (MPI Tübingen)
<https://is.tuebingen.mpg.de/person/bs>
- Leslie Valiant (Harvard)
<https://people.seas.harvard.edu/~valiant>
- Joshua Tenenbaum (MIT)
<http://web.mit.edu/cocosci/josh.html>
- Andrew G. Wilson Cornell (Eric P. Xing, CMU)
<https://people.ori.cornell.edu/andrew>
- Nando de Freitas (Oxford)
<https://www.cs.ox.ac.uk/people/nando.defreitas>
- Yoshua Bengio (Montreal)
http://www.iro.umontreal.ca/~bengioy/yoshua_en
- David Blei (Columbia)
<http://www.cs.columbia.edu/~blei>
- Zoubin Ghahramani (Cambridge)
<http://mlg.eng.cam.ac.uk/zoubin>
- Noah Goodman (Stanford)
<http://cocolab.stanford.edu/ndg.html>

April 24–26, 2014
SIAM SDM14

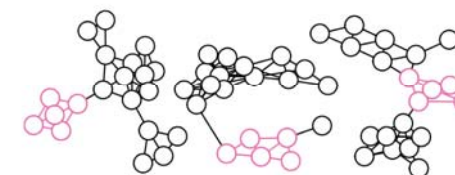
Multi-Task Feature Selection on Multiple Networks via Maximum Flows

Mahito Sugiyama^{1,2}, Chloé Agathe Azencott³, Dominik Grimm^{2,4}, Yoshinobu Kawahara¹, Karsten Borgwardt^{2,4}

¹Osaka University, ²Max Planck Institutes Tübingen, ³Mines ParisTech, Institut Curie, INSERM, ⁴Eberhard Karls Universität Tübingen

Sugiyama, M., Azencott, C.-A., Grimm, D., Kawahara, Y. & Borgwardt, K. M. Multi-Task Feature Selection on Multiple Networks via Maximum Flows. SDM, 2014. 199-207.

- Given multiple graphs
- Find features (=vertices), which are associated with the target response and tend to be connected to each other



$$\underset{\substack{S_1, \dots, S_K \subseteq V \\ K \text{ tasks}}}{\operatorname{argmax}} \sum_{i=1}^K \left(\underbrace{f_i(S_i)}_{\text{association}} - g_i(S_i) \right) - \underbrace{\sum_{i < j} h(S_i, S_j)}_{\text{penalty}}$$

$$f_i(S_i) := \sum_{v \in S_i} q_i(v), \quad g_i(S_i) := \lambda \sum_{\substack{e \in E_i \\ \text{connectivity}}} w_i(e) + \underbrace{\eta |S_i|}_{\text{sparsity}},$$

$$h(S_i, S_j) := \mu |S_i \Delta S_j| = \mu |(S \cup S') \setminus (S \cap S')|$$

- efficiently solved by max-flow algorithms
- performance is superior to Lasso-based methods

- Networks (graphs) are everywhere in health informatics
- Biological pathways (KEGG), chemical compounds, (PubChem), social networks, ...
- Question often: Which part of the network is responsible for performing a particular function?
- → Feature selection on networks
- – Features = vertices (nodes)
- – Network topology = a priori knowledge of relationships between features
- **Multi-task feature selection should be considered for more effectiveness**

- Single task feature selection on a network
- Given a weighted graph $G = (V, E)$
- – Each $v \in V$ has a relevance score $q(v)$
- – If you have a design matrix $\mathbf{X} \in \mathbb{R}^{N \times |V|}$
- and a response vector $\mathbf{y} \in \mathbb{R}^N$, $q(v)$ is the association of \mathbf{y} and each feature of \mathbf{X}

Goal: Find a subset $S \subseteq V$ which maximizes

$$f(S) := \sum_{v \in S} q(v)$$

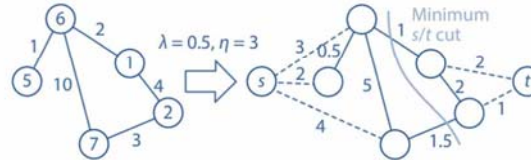
while S is small and vertices are connected

- $\text{argmax}_{S \subseteq V} f(S) - g(S)$
- $f(S) := \sum_{v \in S} q(v), \quad g(S) := \underbrace{\lambda \sum_{e \in E} w(e)}_{\text{connectivity}} + \underbrace{\eta |S|}_{\text{sparsity}}$
- $B = \{\{v, u\} \in E \mid v \in V \setminus S, u \in S\}$ (boundary)
- $w : E \rightarrow \mathbb{R}^+$ is a weighting function

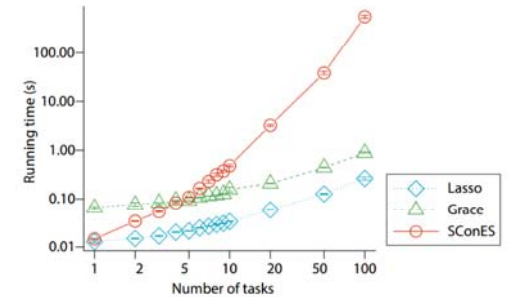


Azencott, C.-A., Grimm, D., Sugiyama, M., Kawahara, Y. & Borgwardt, K. M. 2013. Efficient network-guided multi-locus association mapping with graph cuts. *Bioinformatics*, 29, (13), i171-i179.

- The s/t -network $M(G) = (V \cup \{s, t\}, E \cup S \cup T)$ with $S = \{\{s, v\} \mid v \in V, q(v) > \eta\}, T = \{\{t, v\} \mid v \in V, q(v) < \eta\}$ and set the capacity $c : E' \rightarrow \mathbb{R}^+$ to
- $c(\{v, u\}) = \begin{cases} |q(u) - \eta| & \text{if } u \in \{s, t\} \text{ and } v \in V, \\ \lambda w(\{v, u\}) & \text{otherwise} \end{cases}$
- The minimum s/t cut of $M(G)$ = the solution of SConES



Azencott, C.-A., Grimm, D., Sugiyama, M., Kawahara, Y. & Borgwardt, K. M. 2013. Efficient network-guided multi-locus association mapping with graph cuts. *Bioinformatics*, 29, (13), i171-i179.



Azencott, C.-A., Grimm, D., Sugiyama, M., Kawahara, Y. & Borgwardt, K. M. 2013. Efficient network-guided multi-locus association mapping with graph cuts. *Bioinformatics*, 29, (13), i171-i179.

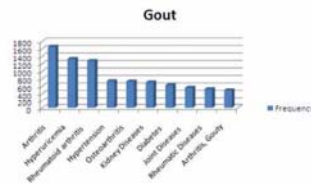
Let two words, w_i and w_j , have probabilities $P(w_i)$ and $P(w_j)$. Then their mutual information $PMI(w_i, w_j)$ is defined as:

$$PMI(w_i, w_j) = \log \left(\frac{P(w_i, w_j)}{P(w_i)P(w_j)} \right)$$

For w_i denoting *rheumatoid arthritis* and w_j representing *diffuse scleritis* the following simple calculation yields:

$$P(w_i) = \frac{94,834}{20,033,079}, \quad P(w_j) = \frac{74}{20,033,079}$$

$$P(w_i, w_j) = \frac{13}{94,834}, \quad PMI(w_i, w_j) = 7.7.$$



Holzinger, A., Simon, K. M. & Yildirim, P. Disease-Disease Relationships for Rheumatic Diseases: Web-Based Biomedical Textmining and Knowledge Discovery to Assist Medical Decision Making. 36th Annual IEEE Computer Software and Applications Conference (COMPSAC), 16-20 July 2012 1212 Izmir, IEEE, 573-580, doi:10.1109/COMPSAC.2012.77.

156

A. Holzinger et al.

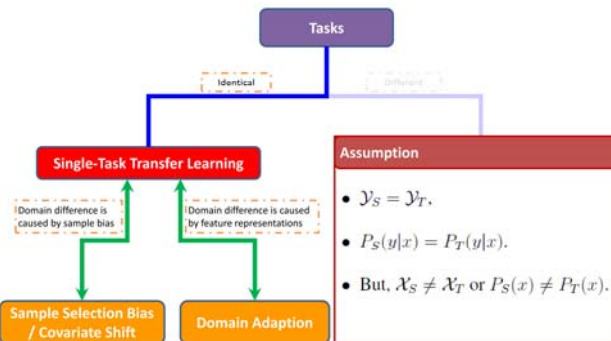
Table 4 Comparison of FACTAs ranking of related concepts from the category Symptom for the query "rheumatoid arthritis" created by the methods co-occurrence frequency, PMI, and SCP

Frequency		PMI		SCP	
joint	5667	impaired body balance	7.8	swollen joints	0.00
Arthritis	661	ASPIRIN INTOLERANCE	7.8	joint	0.00
fatigue	429	Epileptoclear lymphadenopathy	7.8	Arthritis	0.00
diabetes	301	swollen joints	7.8	fatigue	0.00
swollen joints	299	Joint tenderness	7.8	erythema	0.00
erythema	255	Occipital headache	6.2	splenomegaly	0.00
Back Pain	254	Neuromuscular excitation	6.2	Back Pain	0.00
tendache	230	Restless sleep	5.8	polymyalgia	0.00
splenomegaly	228	joint crepitus	5.7	joint stiffness	0.00
Anaesthesia	221	joint symptom	5.5	joint tenderness	0.00
dyspnea	218	Painful feet	5.5	hip pain	0.00
weakness	210	feeling of malaise	5.5	metatarsalgia	0.00
nausea	199	Human's sign	5.4	Skin Manifestations	0.00
Recovery of Function	193	Diffuse pain	5.2	sick pain	0.00
low back pain	167	Palmar erythema	5.2	Eye Manifestations	0.00
abdominal pain	141	Abnormal sensation	5.2	low back pain	0.00

Holzinger, A., Yildirim, P., Geier, M. & Simon, K.-M. 2013. Quality-Based Knowledge Discovery from Medical Text on the Web. In: Pasi, G., Bordogna, G. & Jain, L. C. (eds.) Quality Issues in the Management of Web Information, Intelligent Systems Reference Library, ISRL 50. Berlin Heidelberg: Springer, pp. 145-158, doi:10.1007/978-3-642-37688-7_7.

- Motivation: If two domains are related to each other, then there may exist some "pivot" features across both domain.
- Pivot features are features that behave in the same way for discriminative learning in both domains.
- Main Idea: To identify correspondences among features from different domains by modeling their correlations with pivot features.
- Non-pivot features form different domains that are correlated with many of the same pivot features are assumed to correspond, and they are treated similarly in a discriminative learner.
- Blitzer, J., McDonald, R. & Pereira, F. Domain adaptation with structural correspondence learning. *Proceedings of the 2006 conference on empirical methods in natural language processing*, 2006. Association for Computational Linguistics, 120-128.

Blitzer, J., McDonald, R. & Pereira, F. Domain adaptation with structural correspondence learning. *Proceedings of the 2006 conference on empirical methods in natural language processing*, 2006. Association for Computational Linguistics, 120-128.



**Open Problem:
How to avoid
negative transfer?**