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

Assoc.Prof. Dr. Andreas Holzinger
185.A83 Machine Learning for Health Informatics
2019S, VU, 2.0 h, 3.0 ECTS
Lecture 01 - Dienstag, 12.03.2019

MAKE Health
Machine Learning & Knowledge Extraction in health informatics: challenges & directions

andreas.holzinger AT tuwien.ac.at
https://hci-kdd.org/machine-learning-for-health-informatics-class-2019

Organizational Details: ECTS Breakdown

LV 185.A83 Machine Learning for Health Informatics (Class of 2019)

Study Code: 066 936 Master program Medical Informatics

https://tiss.tuwien.ac.at/curriculum/public/curriculum.xhtml?dswid=9468&dsrid=253&key=56089&semester=NEXT

Semester hours: 2.0 h; ECTS-Credits: 3.0; Type: VU Lecture and Exercise

ECTS-Breakdown (sum=75 h, corresponds with 3 ECTS, where 1 ECTS = 25 h workload):

Presence during lecture	8 * 3 h	24 h
Preparation before and after lecture	8 * 1 h	08 h
Preparation of assignments and presentation	28 h + 2	30 h
Written exam including preparation	1 h + 12 h	13 h
TOTAL students' workload		**75 h**

https://hci-kdd.org/machine-learning-for-health-informatics-class-2019

All Slides will be put on-line AFTER each class!

Organizational Details: Schedule

Class Schedule for 2019 (subject to change: please check class URL for any changes):

Nr	*Day, Date*	*Time*	*h*	*Topic*
1	Dienstag 12.3.2019	17:30-20:30	3 h	Machine learning for health informatics: Introduction, challenges and future directions
2	Dienstag 19.3.2019	17:30-20:30	3h	From clinical decision making to explainable AI: selected methods of transparent machine learning
3	Dienstag 26.3.2019	17:30-20:30	3 h	Tutorial Augmentation and Explainability And FIRST ASSIGNMENT
4	Dienstag 02.4.2019	17:30-20:30	3 h	Probabilistic Graphical Models: from knowledge representation to graph model learning
5	Dienstag 09.4.2019	17:30-20:30	3 h	Tutorial: Probabilistic Programming with Python and SECOND ASSIGNMENT
Easter Break and Time for working on the assignments				
6	Dienstag 30.4.2019	17:30-20:30	3 h	Data for machine learning: quality, fusion, integration, probabilistic information and entropy
7	Dienstag 07.5.2019	17:30-20:30	3 h	Causality and causal machine learning for decision support, ethical, legal and social issues of AI in health
Finalization of assignments				
8	Dienstag 28.5.2019	17:30-20:30	3 h	Final exam (written test, 40 %) and presentations of the assignments (orally, 10 %) quality of the assignments 25 % each (coding, 50 %)

Transparent Procedure how to get grades: sample exam will be made openly available

Let's start with a statement …

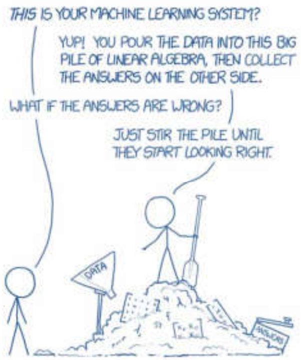


Image Source: Randall Munroe https://xkcd.com

This lecture is only the overview and motivation part ☺

- 01 The HCI-KDD approach: integrative ML
- 02 Application Area Health
- 03 Probabilistic Learning
- 04 Automatic Machine Learning (aML)
- 05 Interactive Machine Learning (iML)
- 06 Causality vs. Causability
- 07 Explainable AI
- Conclusion and Future Outlook

01 What is the HCI-KDD approach?

ML is a very practical field …

- algorithm development is at the core – however, successful ML needs a concerted effort of various topics …

Andreas Holzinger 2017. Introduction to Machine Learning and Knowledge Extraction (MAKE). *Machine Learning and Knowledge Extraction,* 1, (1), 1-20, doi:10.3390/make1010001.

… successful ML needs …

http://www.bach-cantatas.com

See you in Canterbury (Kent), UK, End of August

Image with friendly permission of Michael D. Beckwith

Grand Goal: Understanding Intelligence ?

"Solve intelligence – then solve everything else"

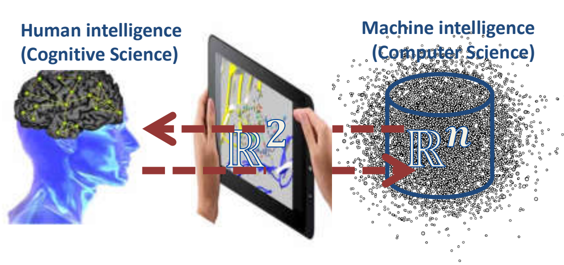
https://youtu.be/XAbLn66iHcQ?t=1h28m54s

Demis Hassabis, 22 May 2015

The Royal Society, Future Directions of Machine Learning Part 2

Google DeepMind

Grand challenge: Transfer results from $\mathbb{R}^n$ to $\mathbb{R}^2$



Andreas Holzinger 2013. Human–Computer Interaction and Knowledge Discovery (HCI-KDD): What is the benefit of bringing those two fields to work together? In: Lecture Notes in Computer Science LNCS 8127. pp. 319-328, doi:10.1007/978-3-642-40511-2_22.

To reach a level of usable intelligence we need to …

04

- 1) **learn** from prior data
- 2) **extract** knowledge
- 2) **generalize,** i.e. guessing where a probability mass function 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

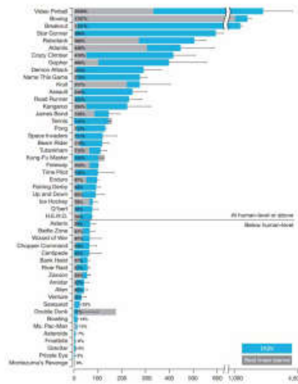
Our goal: Understanding Context !

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



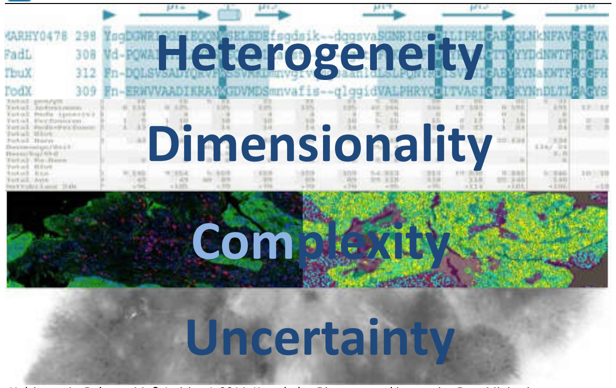
Not our Goal: Humanoid AI

Health is a complex area

Why is this application area complex ?

Our central hypothesis: Information may bridge this gap

Andreas Holzinger & Klaus-Martin Simonic (eds.) 2011. Information Quality in e-Health. Lecture Notes in Computer Science LNCS 7058, Heidelberg, Berlin, New York: Springer, doi:10.1007/978-3-642-25364-5.



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.

03 Probabilistic Learning

The true logic of this world is in the calculus of probabilities.
James Clerk Maxwell

Probability theory is nothing but common sense reduced to calculation …

$$\hat{y} = \hat{f}(\mathbf{x}) = \operatorname*{argmax}_{c=1}^{C} p(y = c|\mathbf{x}, \mathcal{D})$$

Pierre Simon de Laplace (1749-1827)

06

What is the simplest mathematical operation for us?

$$p(x) = \sum_{y}(p(x, y)) \quad (1)$$

How do we call repeated adding?

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

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:

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

- 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

$$p(x_i) = \sum P(x_i, y_j)$$

$$p(x_i, y_j) = p(y_j|x_i)P(x_i)$$

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

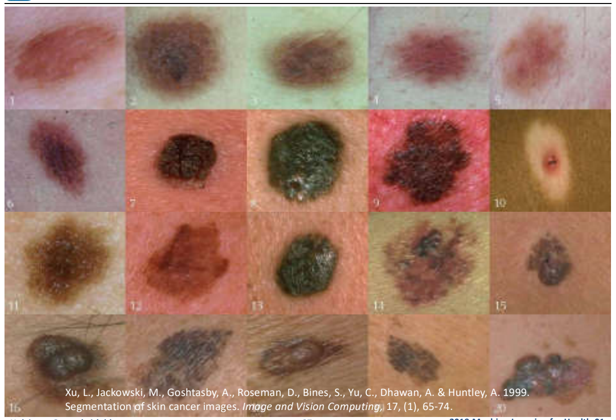
Bayes' Rule is a corollary of the Sum Rule and Product Rule:

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

$$P(\text{hypothesis}|\text{data}) = \frac{P(\text{hypothesis})P(\text{data}|\text{hypothesis})}{\sum_h P(h)P(\text{data}|h)} \qquad P(\theta|\mathcal{D}, m) = \frac{P(\mathcal{D}|\theta, m)P(\theta|m)}{P(\mathcal{D}|m)}$$

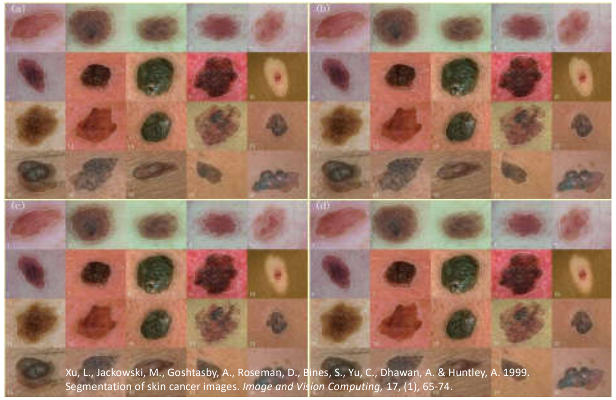
$P(D|\theta, m)$ likelihood of parameters θ in model m
$P(\theta|m)$ prior probability of θ
$P(\theta|D, m)$ posterior of θ given data D

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.



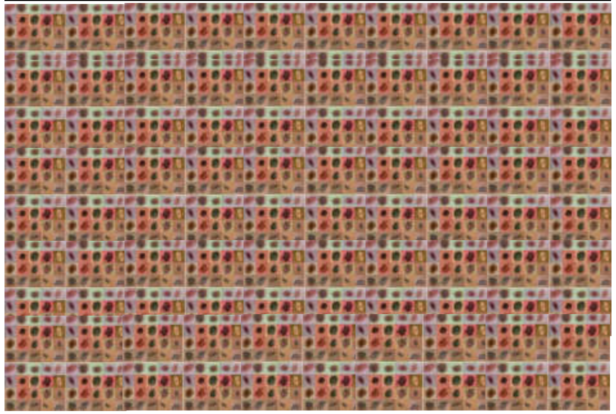
Xu, L., Jackowski, M., Goshtasby, A., Roseman, D., Bines, S., Yu, C., Dhawan, A. & Huntley, A. 1999. Segmentation of skin cancer images. *Image and Vision Computing*, 17, (1), 65-74.

The more the better …



Xu, L., Jackowski, M., Goshtasby, A., Roseman, D., Bines, S., Yu, C., Dhawan, A. & Huntley, A. 1999. Segmentation of skin cancer images. *Image and Vision Computing*, 17, (1), 65-74.

More, more, more, …



Bayesian Learning from data

05

$\mathcal{D} = x_{1:n} = \{x_1, x_2, ..., x_n\}$ $\quad p(\mathcal{D}|\theta)$

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta) * p(\theta)}{p(\mathcal{D})}$$

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

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

Learning and Inference

04

d … *data* $\quad \mathcal{H}$ … $\{H_1, H_2, ..., H_n\}$ $\quad \forall\, h, d$ …

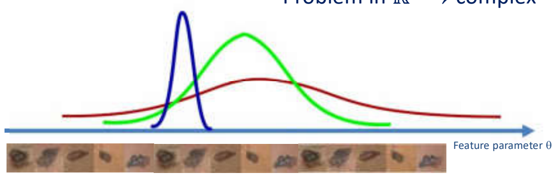
h … *hypotheses*

Likelihood — Prior Probability

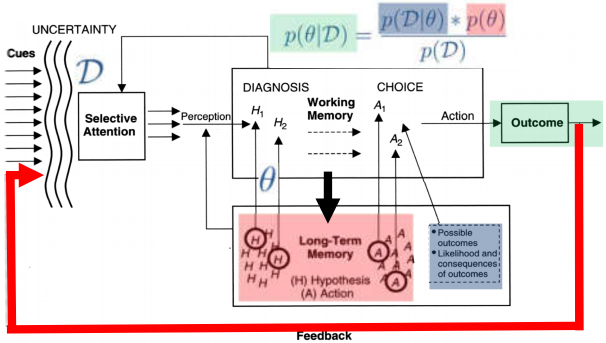
$$p(h|d) = \frac{p(d|h) * p(h)}{\sum_{h \in H} p(d|h')\, p(h')}$$

Posterior Probability

Problem in $\mathbb{R}^n \rightarrow$ complex



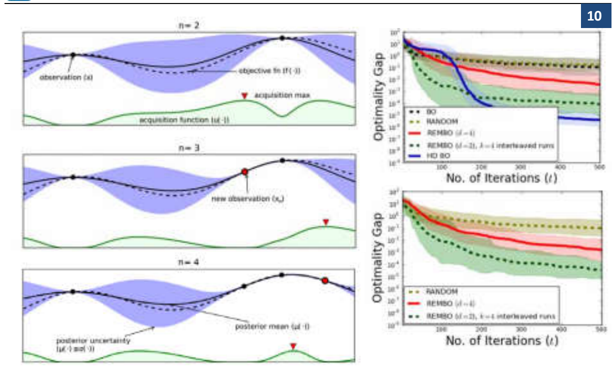
Connection to Cognitive Science: Decision Making



Wickens, C. D. (1984) *Engineering psychology and human performance.* *Columbus (OH), Charles Merrill, modified by Holzinger, A.*

Scaling to high-dimensions is the holy grail in ML

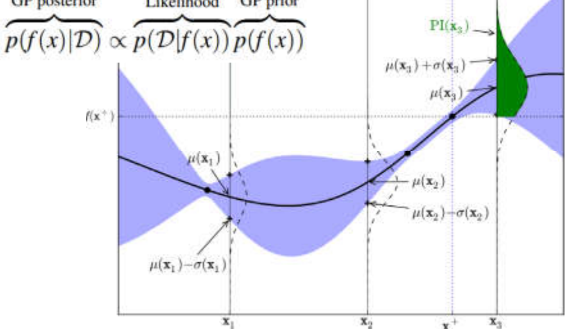
10



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

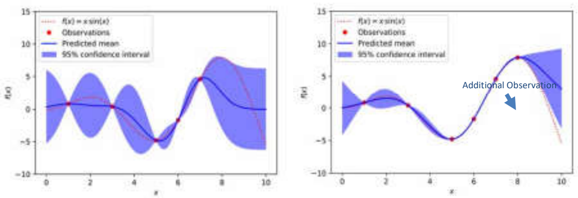
GP = distribution, observations occur in a cont. domain, e.g. t or space

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.

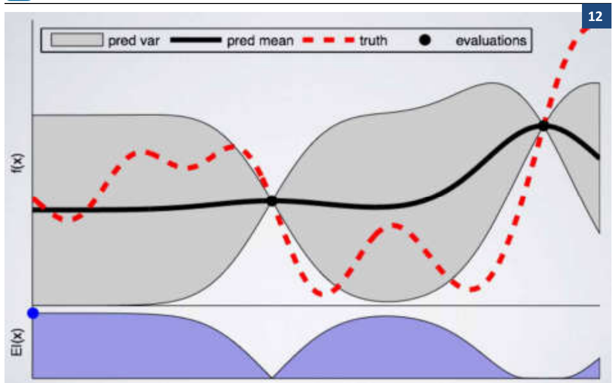
Goal: Reduce uncertainty by learning representations



$$\mathbb{E}[f] = \int p(x) f(x)\, dx \qquad \mathbb{E}[f] \simeq \frac{1}{N}\sum_{n=1}^{N} f(x_n)$$

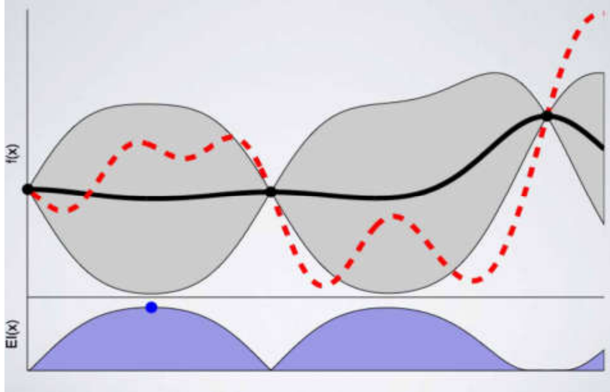
Holzinger, A. 2017. Introduction to Machine Learning and Knowledge Extraction (MAKE). Machine Learning and Knowledge Extraction, 1, (1), 1-20, doi:10.3390/make1010001.

12

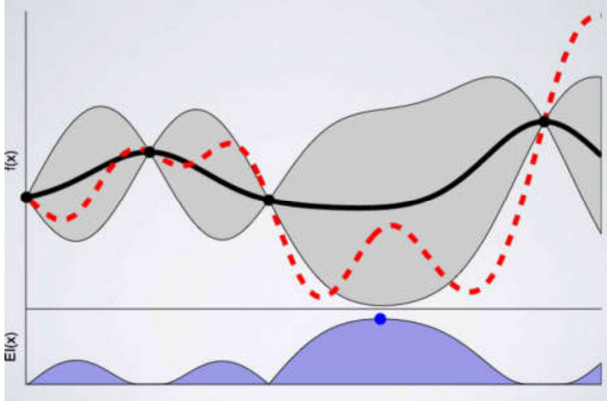


Snoek, J., Larochelle, H. & Adams, R. P. Practical Bayesian optimization of machine learning algorithms. Advances in neural information processing systems, 2012. 2951-2959.

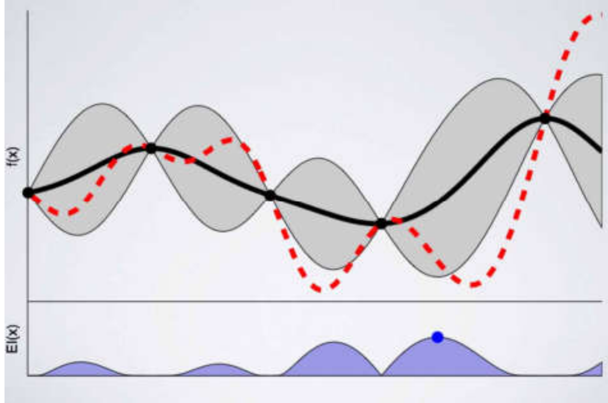
Bayesian Optimization 1



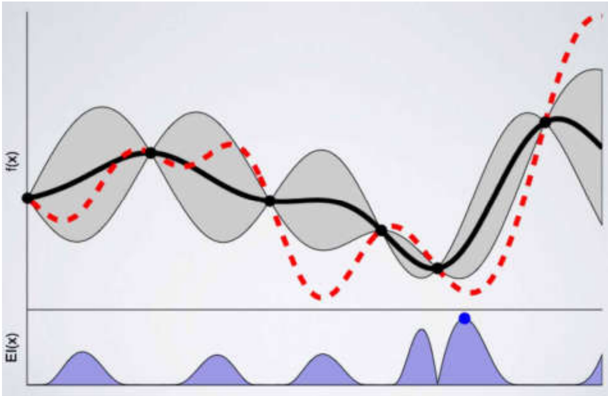
Bayesian Optimization 2



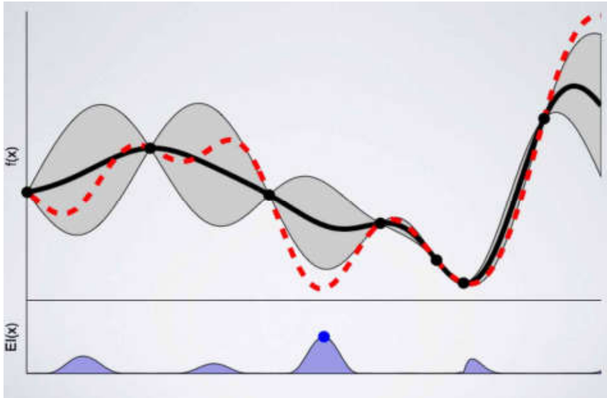
Bayesian Optimization 3



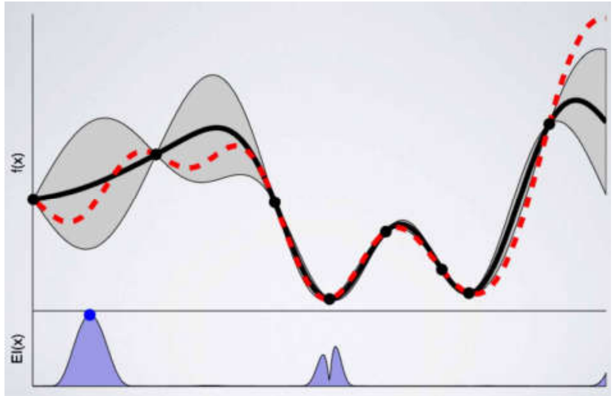
Bayesian Optimization 4



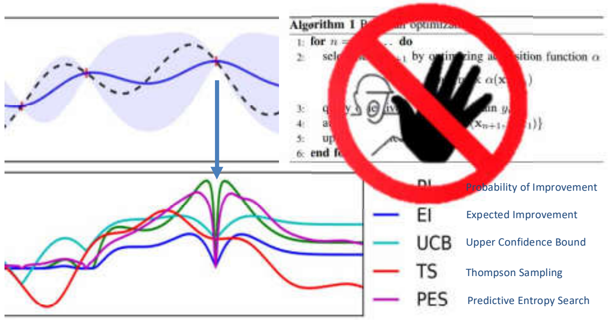
Bayesian Optimization 5



Bayesian Optimization 6



Goal: Taking the human out of the loop

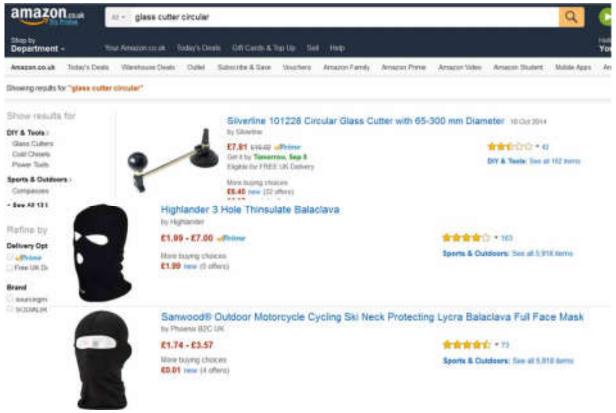


Shahriari, B., Swersky, K., Wang, Z., Adams, R. P. & De Freitas, N. 2016. **Taking the human out of the loop:** A review of Bayesian optimization. *Proceedings of the IEEE*, 104, (1), 148-175, doi:10.1109/JPROC.2015.2494218.

04 aML

Best practice examples of aML …

Recommender Systems



Fully automatic autonomous vehicles ("Google car")

Guizzo, E. 2011. How google's self-driving car works. IEEE Spectrum Online, 10, 18.

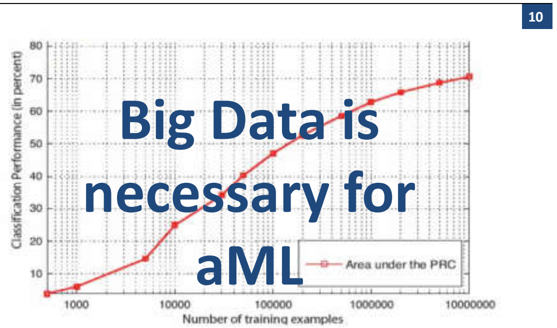
Why did the car do that? Who is responsible?

Image Source: http://www.businessinsider.de/who-is-responsible-when-a-driverless-car-crashes-2016-2?r=US&IR=T

… and thousands of industrial aML applications …

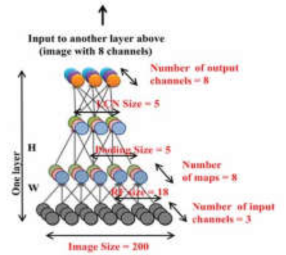
Seshia, S. A., Juniwal, G., Sadigh, D., Donze, A., Li, W., Jensen, J. C., Jin, X., Deshmukh, J., Lee, E. & Sastry, S. 2015. Verification by, for, and of Humans: Formal Methods for Cyber-Physical Systems and Beyond. Illinois ECE Colloquium.

Big Data is necessary for aML !



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

10 million 200 x 200 px images downloaded from Web



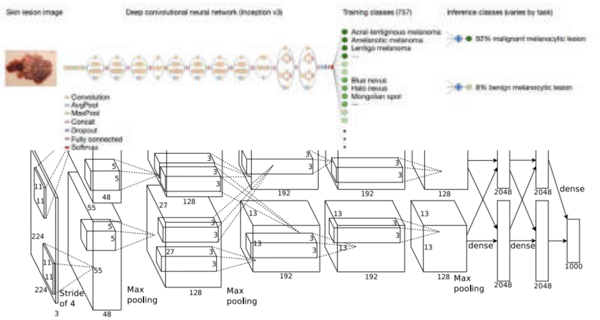
$$x^* = \arg\min_x f(x; W, H), \quad \text{subject to } ||x||_2 = 1.$$

Le, Q. V., Ranzato, M. A., Monga, R., Devin, M., Chen, K., Corrado, G. S., Dean, J. & Ng, A. Y. 2011. Building high-level features using large scale unsupervised learning. arXiv preprint arXiv:1112.6209.

Le, Q. V. 2013. Building high-level features using large scale unsupervised learning. *IEEE Intl. Conference on Acoustics, Speech and Signal Processing ICASSP.* IEEE. 8595-8598, doi:10.1109/ICASSP.2013.6639343.

Deep Convolutional Neural Network Pipeline

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M. & Thrun, S. 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542, (7639), 115-118, doi:10.1038/nature21056.



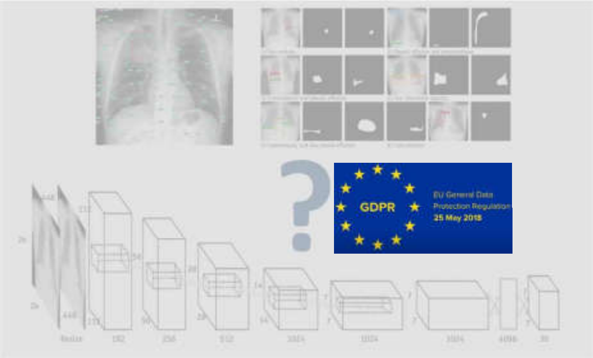
Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C. J. C., Bottou, L. & Weinberger, K. Q., eds. Advances in neural information processing systems (NIPS 2012), 2012 Lake Tahoe. 1097-1105.

When does aML fail …

- 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, …

Houston, we have a problem …

Source: NASA, Image is in the public domain



June-Goo Lee, Sanghoon Jun, Young-Won Cho, Hyunna Lee, Guk Bae Kim, Joon Beom Seo & Namkug Kim 2017. Deep learning in medical imaging: general overview. Korean journal of radiology, 18, (4), 570-584, doi:10.3348/kjr.2017.18.4.570.

Consequently …

There is an urgent need for “explainability”

This brings us now naturally to …

05 iML

Definition of iML (Holzinger – 2016)

16

- 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.

Sometimes we need a doctor-in-the-loop

Image Source: 10 Ways Technology is Changing Healthcare http://newhealthypost.com Posted online on April 22, 2018

A group of experts-in-the-loop

A crowd of people-in-the-loop

Why using human intuition?

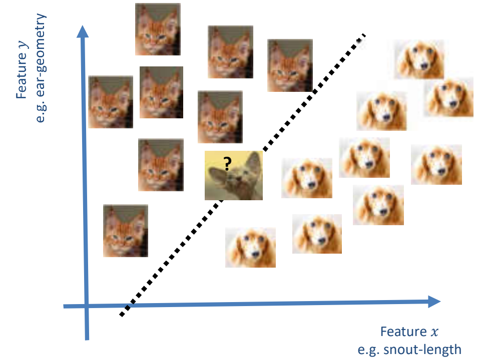
Humans can understand the context

- **Humans can generalize even from few examples …**
  - They learn relevant representations
  - Can disentangle the explanatory factors
  - Find the 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.

… can infer from little data …

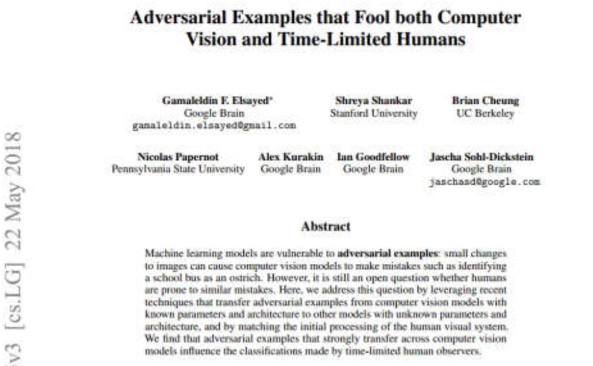
Tenenbaum, J. B., Kemp, C., Griffiths, T. L. & Goodman, N. D. 2011. How to grow a mind: Statistics, structure, and abstraction. Science, 331, (6022), 1279-1285, doi:10.1126/science.1192788.



Examples: https://imgur.com/a/K4RWn

See also: Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy 2014. Explaining and harnessing adversarial examples. arXiv:1412.6572.

Adversarial > Deep Learning > Deep Fake



Adversarial Examples that Fool both Computer Vision and Time-Limited Humans

Gamaleldin F. Elsayed* Google Brain gamaleldin.elsayed@gmail.com
Shreya Shankar Stanford University
Brian Cheung UC Berkeley
Nicolas Papernot Pennsylvania State University
Alex Kurakin Google Brain
Ian Goodfellow Google Brain
Jascha Sohl-Dickstein Google Brain jaschasd@google.com

arXiv:...v3 [cs.LG] 22 May 2018

Abstract

Machine learning models are vulnerable to **adversarial examples**: small changes to images can cause computer vision models to make mistakes such as identifying a school bus as an ostrich. However, it is still an open question whether humans are prone to similar mistakes. Here, we address this question by leveraging recent techniques that transfer adversarial examples from computer vision models with known parameters and architecture to other models with unknown parameters and architecture, and by matching the initial processing of the human visual system. We find that adversarial examples that strongly transfer across computer vision models influence the classifications made by time-limited human observers.

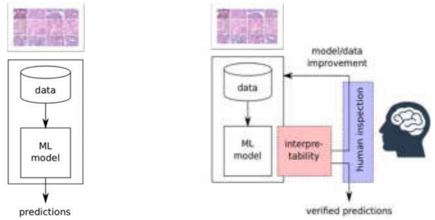
Gamaleldin F Elsayed, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alex Kurakin, Ian Goodfellow & Jascha Sohl-Dickstein 2018. Adversarial Examples that Fool both Human and Computer Vision. arXiv:1802.08195.

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.

Know your errors !!!

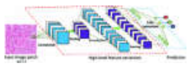


Generalization Error

Generalization Error + Human Experience

Why?

Verify that algorithms/classifiers work as expected
Wrong decisions can be costly and dangerous



Understanding the weaknesses and errors of the ML-Model - Detection of bias in both directions

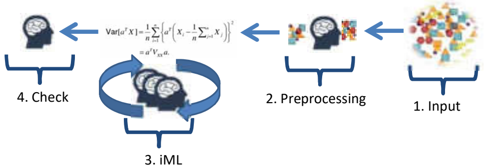
Scientific interpretability, replicability, causality
The "why" is often more important than the prediction

Enable re-traceability, re-enactivity
Compliance to legislation "right for explanation", retain human reliability, fosters trust and acceptance

iML: bringing the human-in-the-loop

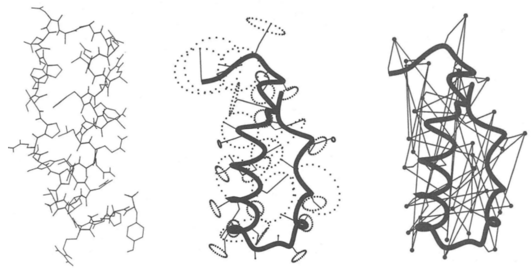
17

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 …



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.

Example: Protein Folding is a TSP



Bohr, H. & Brunak, S. 1989. A travelling salesman approach to protein conformation. Complex Systems, 3, 9-28

19

```
Input   : ProblemSize, m, β, ρ, σ, q0
Output: Pbest
Pbest ← CreateHeuristicSolution(ProblemSize);
Pbest_cost ← Cost(Pbest);
Pheromone_init ← 1.0 / (ProblemSize × Pbest_cost);
Pheromone ← InitializePheromone(Pheromone_init);
while ¬StopCondition() do
    for i = 1 to m do
        S_i ← ConstructSolution(Pheromone, ProblemSize, β, q0);
        Si_cost ← Cost(S_i);
        if Si_cost ≤ Pbest_cost then
            Pbest_cost ← Si_cost;
            Pbest ← S_i;
        end
        LocalUpdateAndDecayPheromone(Pheromone, S_i, Si_cost, ρ);
    end
    GlobalUpdateAndDecayPheromone(Pheromone, Pbest, Pbest_cost, ρ);
    while isUserInteraction() do
        GlobalAddAndRemovePheromone(Pheromone, Pbest, Pbest_cost, ρ);
    end
end
return P_best;
```

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.

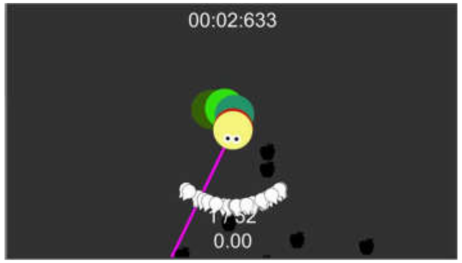
Influence the Pheromone Level by human intuition

$$p_{ij} = \frac{[\tau_{ij}]^{\alpha} \cdot [\eta_{ij}]^{\beta}}{\sum_{l \in J_i^k} [\tau(t)]^{\alpha} \cdot [\eta]^{\beta}}$$

- p_{ij} ... **probability** of ants that they, at a particular node i, select the route from node $i \rightarrow j$ (**"heuristic desirability"**)
- $\alpha > 0$ and $\beta > 0$ ... the **influence parameters** (α ... history coefficient, β ...heuristic coefficient) usually $\alpha \approx \beta \approx 2 < 5$
- τ_{ij} ... the **pheromone value** for the components, i.e. the amount of pheromone on edge (i, j)
- k ... the set of usable components
- J_i ... the set of nodes that ant k can reach from v_i (tabu list)
- $\eta_{ij} = \frac{1}{dij}$ ... attractiveness computed by a heuristic, indicating the "a-priori **desirability"** of the move

Experiment 1: The travelling Snakesman

http://hci-kdd.org/gamification-interactive-machine-learning/



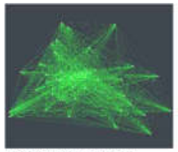
LIVE DEMO

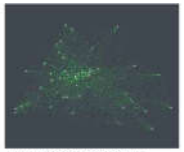
(https://iml.hci-kdd.org/imlTspSolver/)

ANDROID:
https://play.google.com/store/apps/details?id=com.hcikdd.imlacosolver

Explainable Model - the visualization

- The pheromones are showing "the state" (high or low frequented paths of ants) of the algorithm.

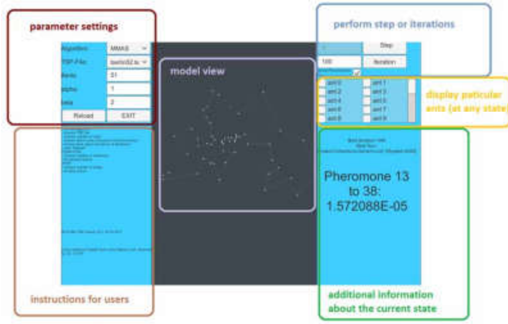

initial pheromone distibution


pheromones after 100 iterations


pheromones after 500 iterations

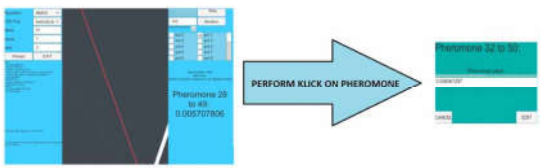
Explainable Interface - overview

http://iml.hci-kdd.org/imlTspSolver/



Explainable Interface - interactive(iML) parts

- iteration vs. step: look inside the iteration
- make the ant algorithm interactive
  - change pheromones at any time
  - *change routes of certain ants in the current iteration (future work)*



06
Causality vs. Causability

Example: Discovery of causal relationships from data ...

15

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

Causality and Explanations

David Hume (1711-1776)

Causation is a matter of perception

We remember seeing the flame, and feeling a sensation called heat; without further ceremony, we call the one cause and the other effect

Statistical ML

Forget causation! Correlation is all you should ask for.

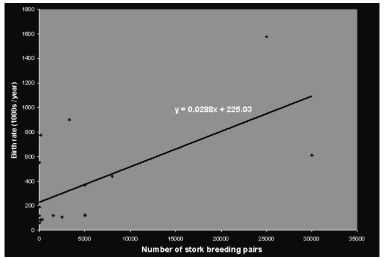
Karl Pearson (1857-1936)

Judea Pearl (1936-)

A mathematical definition of causality

Forget empirical observations! Define causality based on a network of known, physical, causal relationships

Example: Correlation vs. Causality



Storks Deliver Babies ($p = 0.008$)

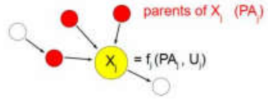
KEYWORDS:
Teaching;
Correlation;
Significance;
p-values.

Robert Matthews
Aston University, Birmingham, England.
e-mail: rajm@compuserve.com

Summary
This article shows that a highly statistically significant correlation exists between stork populations and human birth rates across Europe. While storks may not deliver babies, unthinking interpretation of correlation and *p*-values can certainly deliver unreliable conclusions.

Functional Causal Model *(Pearl et al.)*

- Set of observables $X_1, \ldots, X_n$
- directed acyclic graph G with vertices $X_1, \ldots, X_n$
- Semantics: parents = direct causes
- $X_i = f_i(\mathrm{ParentsOf}_i, \mathrm{Noise}_i)$, with independent $\mathrm{Noise}_1, \ldots, \mathrm{Noise}_n$.
- "Noise" means "unexplained" (or "exogenous"), we use U_i
- Can add requirement that $f_1, \ldots, f_n, \mathrm{Noise}_1, \ldots, \mathrm{Noise}_n$ "independent" (cf. *Lemeire & Dirkx 2006, Janzing & Schölkopf 2010* — more below)



This requires …

Why is it a cat? What makes a cat a cat?

Image in the public domain, Credit to Kevin Dooley

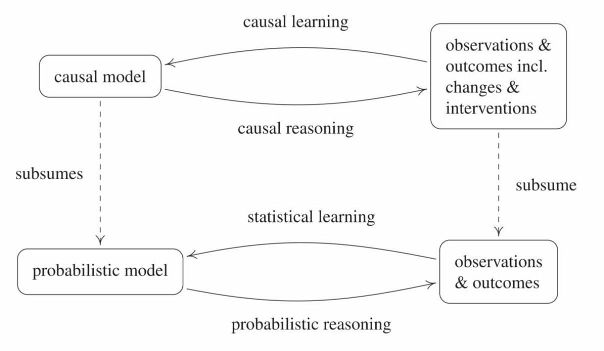
Definition:

Explainability	in a technical sense highlights decision-relevant parts of the used representations of the algorithms and active parts in the algorithmic model, that either contribute to the model accuracy on the training set, or to a specific prediction for one particular observation. It does not refer to an explicit human model.
Causability	as the extent to which an explanation of a statement to a human expert achieves a specified level of causal understanding with effectiveness, efficiency and satisfaction in a specified context of use.

- **Causability := a property of a person, while**
- **Explainability := a property of a system**

07
explainable AI

Causality is going beyond explainability



Remember:
Context !!!

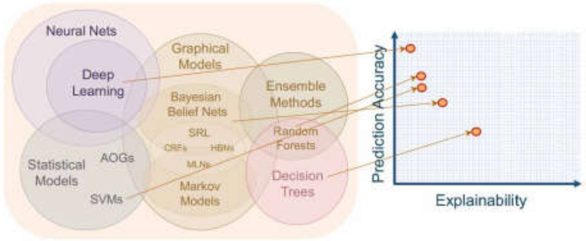
The state-of-the-art is terrible …

Andrej Karpathy & Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015. 3128-3137.

Image Captions by dee learning : github.com/karpathy/neuraltalk2

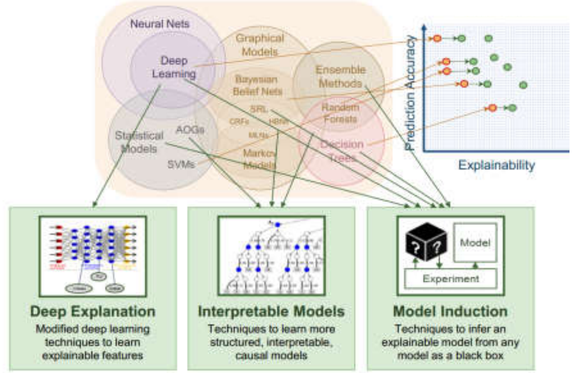
Image Source: Gabriel Villena Fernandez; Agence France-Press, Dave Martin (left to right)

Problem: Interpretability vs. Performance



David Gunning 2016. Explainable artificial intelligence (XAI): Technical Report Defense Advanced Research Projects Agency DARPA-BAA-16-53, Arlington, USA, DARPA.

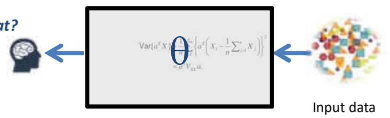
Making Models Interpretable, Explainable, Retraceable



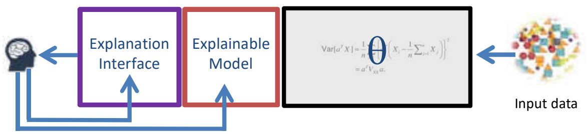
David Gunning 2016. Explainable artificial intelligence (XAI): Technical Report Defense Advanced Research Projects Agency DARPA-BAA-16-53, Arlington, USA, DARPA.

The challenge

Why did the algorithm do that?
Can I trust these results?
How can I correct an error?



A possible solution

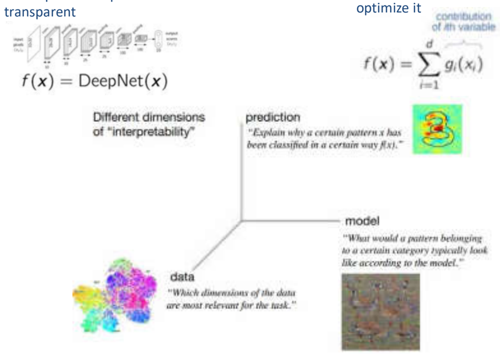


The domain expert can understand why …
The domain expert can learn and correct errors …
The domain expert can re-enact on demand …

Post-hoc vs. Ante-hoc

Post-hoc: Select a model and develop a technique to make it transparent

Ante-hoc: Select a model that is already transparent and optimize it



Methods of ex-AI

- 1) Gradients
- 2) Sensitivity Analysis
- **3) Decomposition Relevance Propagation**
  (Pixel-RP, Layer-RP, Deep Taylor Decomposition, …)
- 4) Optimization (Local-IME – model agnostic, BETA transparent approximation, …)
- 5) Deconvolution and Guided Backpropagation
- 6) Model Understanding
  - Feature visualization, Inverting CNN
  - Qualitative Testing with Concept Activation Vectors TCAV
  - Network Dissection

Andreas Holzinger LV 706.315 From explainable AI to Causability, 3 ECTS course at Graz University of Technology
https://hci-kdd.org/explainable-ai-causability-2019 (course given since 2016)

Conclusion and Future Outlook

Three Main future challenges

13

Multi-Task Learning (MUTL)

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

Transfer learning (TRAL)

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.

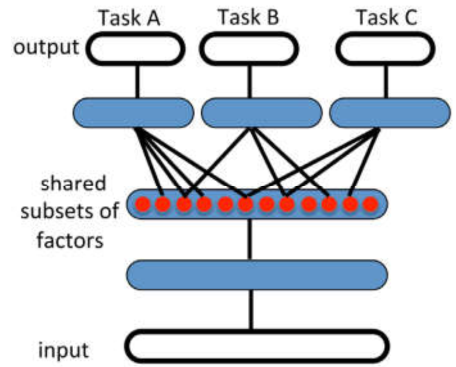
Multi-Agent-Hybrid Systems (MAHS)

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

Representation Learning discovering explanatory factors

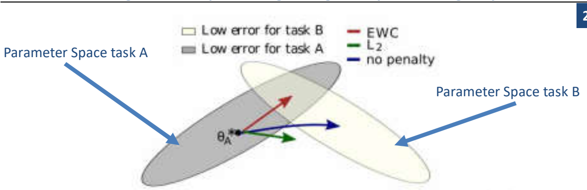
21

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.



Overcoming Catastrophic Forgetting: Deep Learning Bayes

20



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

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

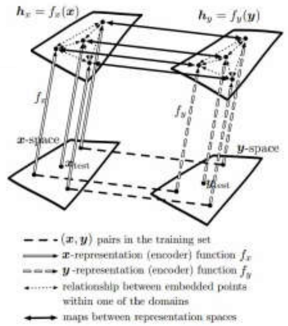
$$\mathcal{L}(\theta) = \mathcal{L}_B(\theta) + \sum_i \frac{\lambda}{2} F_i(\theta_i - \theta^*_{A,i})^2$$

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.

Maps between shared representations

22

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

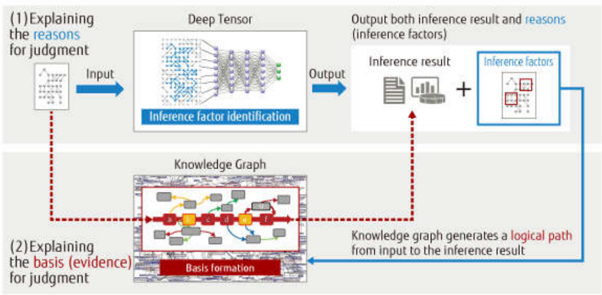


Goodfellow, I., Bengio, Y. & Courville, A. 2016. Deep Learning, Cambridge: MIT Press, p.542

Conclusion

- **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**

Combination Probabilistic + Logic approaches



Randy Goebel, Ajay Chander, Katharina Holzinger, Freddy Lecue, Zeynep Akata, Simone Stumpf, Peter Kieseberg & Andreas Holzinger 2018. Explainable AI: the new 42? Springer Lecture Notes in Computer Science LNCS 11015

A final citation attributed to Albert Einstein …

- Computers are fast, accurate and stupid,
- humans are slow, inaccurate and brilliant,
- **together** they are powerful beyond imagination

(Einstein never said that)

https://www.benshoemate.com/2008/11/30/einstein-never-said-that

Thank you!

Questions

Questions (1/4)

- What is the HCI-KDD approach?
- What is meat by "integrative ML"?
- Why is a direct integration of AI-solutions into the workflow important?
- What are features?
- Why is understanding intelligence important?
- Why is understanding context even more important?
- What are currently the "best" ML-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?

Questions (2/4)

- 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?

Questions (3/4)

- 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!

Questions (4/4)

- 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

Keywords

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

Henri Poincare in Sciences et Methods (1908)

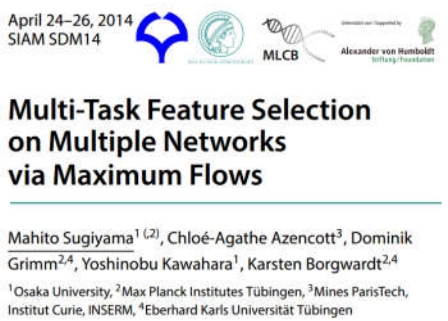
- *"The most interesting facts are*
- *those which can be used several times, those which have a chance of recurring …*
- *which, then, are the facts that have a chance of recurring?*
- *In the first place, **simple** facts."*


Jules Henri Poincaré (1854–1912)

Henri Poincare, Sciences et Methods (1908)

Scientists recognizing this … (totally incomplete list!)

- 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.orie.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

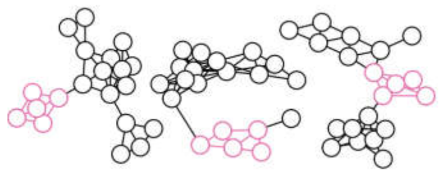
Mahito Sugiyama[1,(2)], Chloé-Agathe Azencott[3], Dominik Grimm[2,4], Yoshinobu Kawahara[1], Karsten Borgwardt[2,4]

[1]Osaka University, [2]Max Planck Institutes Tübingen, [3]Mines ParisTech, Institut Curie, INSERM, [4]Eberhard 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.

Goal

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



Result: New formulation of MTF-Selection

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

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

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

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

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.

Remember: Graphs are everywhere!

- 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**

Selecting Connected Explanatory SNPs SConES

- 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 $S \subset V$ which maximizes

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

while S is small and vertices are connected

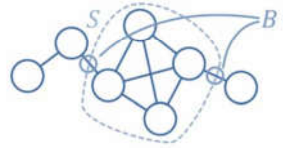
Azencott, C.-A., Grimm, D., Sugiyama, M., Kawahara, Y. & Borgwardt, K. M. 2013. Efficient network-guided multi-locus association mapping with graph cuts. Bioinformatics, 29, (13), i171-i179.

Formulation of SConES

- $\text{argmax}_{S \subseteq V} f(S) - g(S)$

$$f(S) := \sum_{v \in S} q(v), \quad g(S) := \underbrace{\lambda \textstyle\sum_{e \in B} w(e)}_{\text{connectivity}} + \underbrace{\eta |S|}_{\text{sparsity}}$$

- $B = \{\{v, u\} \in E \mid v \in V \setminus S,\, u \in S\}$ (boundary)
- $w : E \to \mathbb{R}^{+}$ is a weighting function



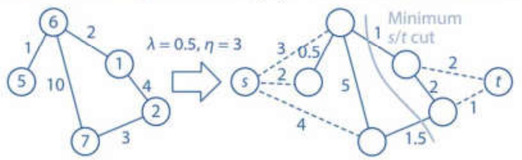
Azencott, C.-A., Grimm, D., Sugiyama, M., Kawahara, Y. & Borgwardt, K. M. 2013. Efficient network-guided multi-locus association mapping with graph cuts. Bioinformatics, 29, (13), i171-i179.

Solution of SConES via Maximum Flow

- 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 $c : E' \to \mathbb{R}^{+}$ to

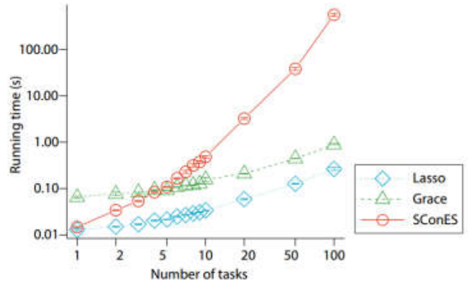
$$c(\{v, u\}) = \begin{cases} |q(u) - \eta| & \text{if } u \in \{s, t\} \text{ and } v \in V, \\ \lambda w(\{v, u\}) & \text{otherwise} \end{cases}$$

- 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.

Better performance is always convincing!



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.

Example: Disease-Disease Relationship

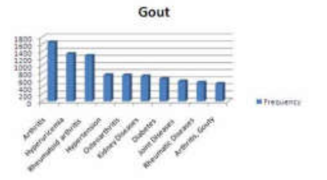
Let two words, w_i and w_j, have probabilities $P(w_i)$ and $P(w_j)$. Then their mutual information $PMI(w_i, w_j)$ is defined as:

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

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

$$SCP(x, y) = p(x|y) \cdot p(y|x) = \frac{p(x, y)}{p(y)} \cdot \frac{p(x, y)}{p(x)} = \frac{p(x, y)^2}{p(x) \cdot p(y)}$$

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	
pain	5667	impaired body balance	7,8	swollen joints	0,002
Arthralgia	661	ASPIRIN INTOLERANCE	7,8	pain	0,001
fatigue	429	Epitrochlear lymphadenopathy	7,8	Arthralgia	0,001
diarrhea	301	swollen joints	7,4	fatigue	0,000
swollen joints	299	Joint tenderness	7	erythema	0,000
erythema	255	Occipital headache	6,2	splenomegaly	0,000
Back Pain	254	Neuromuscular excitation	6,2	Back Pain	0,000
headache	239	Restless sleep	5,8	polymyalgia	0,000
splenomegaly	228	joint crepitus	5,7	joint stiffness	0,000
Anesthesia	221	joint symptom	5,5	Joint tenderness	0,000
dyspnea	218	Painful feet	5,3	hip pain	0,000
weakness	210	feeling of malaise	5,3	metatarsalgia	0,000
nausea	199	Homan's sign	5,4	Skin Manifestations	0,000
Recovery of Function	193	Diffuse pain	5,2	neck pain	0,000
low back pain	167	Palmar erythema	5,2	Eye Manifestations	0,000
abdominal pain	141	Abnormal sensation	5,2	low back pain	0,000

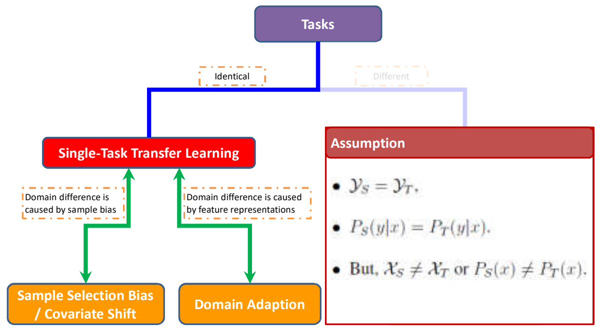
Holzinger, A., Yildirim, P., Geier, M. & Simonic, K.-M. 2013. Quality-Based Knowledge Discovery from Medical Text on the Web. In: Pasi, G., Bordogna, G. & Jain, L. C. (eds.) Quality Issues in the Management of Web Information, Intelligent Systems Reference Library, ISRL 50. Berlin Heidelberg: Springer,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

Domain Adaptation: Structural Correspondence Learning

- 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.

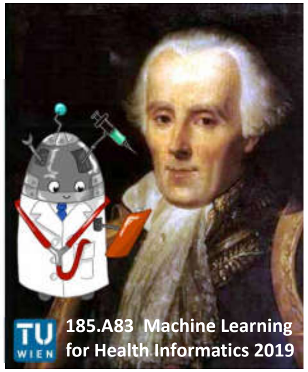
Identical Tasks



- $\mathcal{Y}_S = \mathcal{Y}_T$.
- $P_S(y|x) = P_T(y|x)$.
- But, $\mathcal{X}_S \neq \mathcal{X}_T$ or $P_S(x) \neq P_T(x)$.

Open Problem: How to avoid negative transfer?

The foundation for modern machine learning …



Limitations of Deep Learning approaches

- Computational resource intensive (supercomps, cloud CPUs, **federated learning,** …)
- Black-Box approaches – lack **transparency,** do not foster trust and acceptance among end-user, legal aspects make "black box" difficult!
- **Non-convex:** difficult to set up, to train, to optimize, needs a lot of expertise, error prone
- Very bad in dealing with **uncertainty**
- **Data intensive, needs often millions of training samples …**

Three examples for the usefulness of the iML approach

18

- **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., Majnaric, 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.

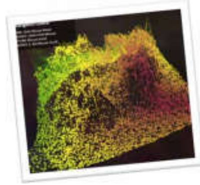
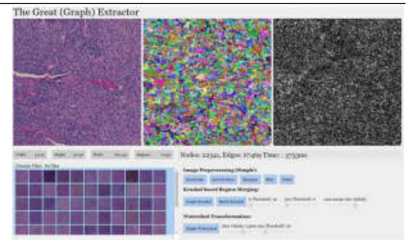
Project: Tumor-Growth Simulation

- 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.

Jeanquartier, F., Jean-Quartier, C., Kotlyar, M., Tokar, T., Hauschild, A.-C., Jurisica, I. & Holzinger, A. 2016. Machine Learning for In Silico Modeling of Tumor Growth. In: Springer Lecture Notes in Artificial Intelligence LNAI 9605. Cham: Springer International Publishing, pp. 415-434, doi:10.1007/978-3-319-50478-0_21.

Project: Graphinius

- Contribute to graph understanding and algorithm prototyping by real-time visualization, interaction and manipulation
- Supports client-based federated learning
- 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.