



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>



- **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**

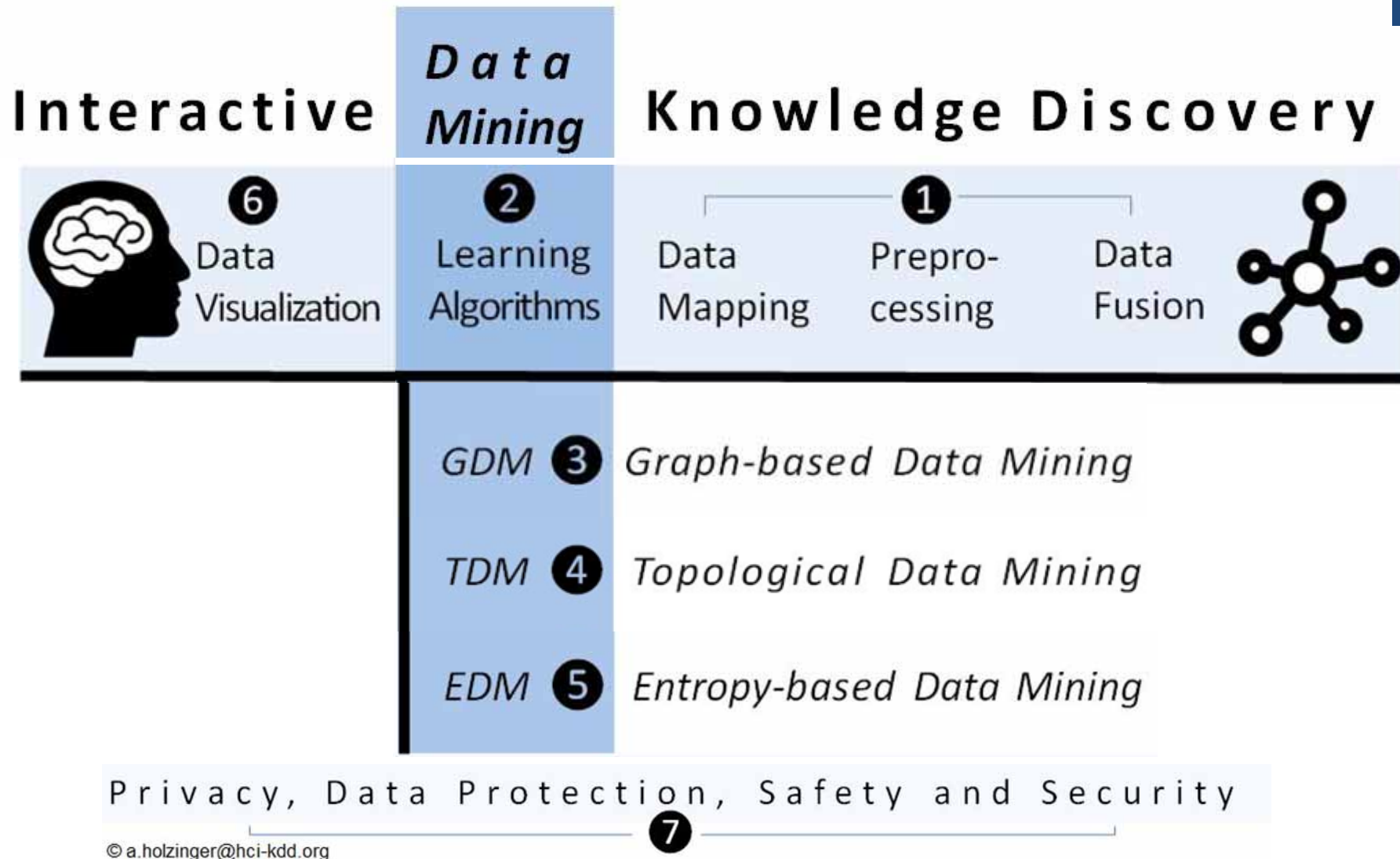
01 What is the



approach?

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





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

Features are key to learning and understanding



concerted effort
international
without boundaries ...

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





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

02 Solve Intelligence then solve everything else

“Solve intelligence – then solve everything else”



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

Demis Hassabis, 22 May 2015

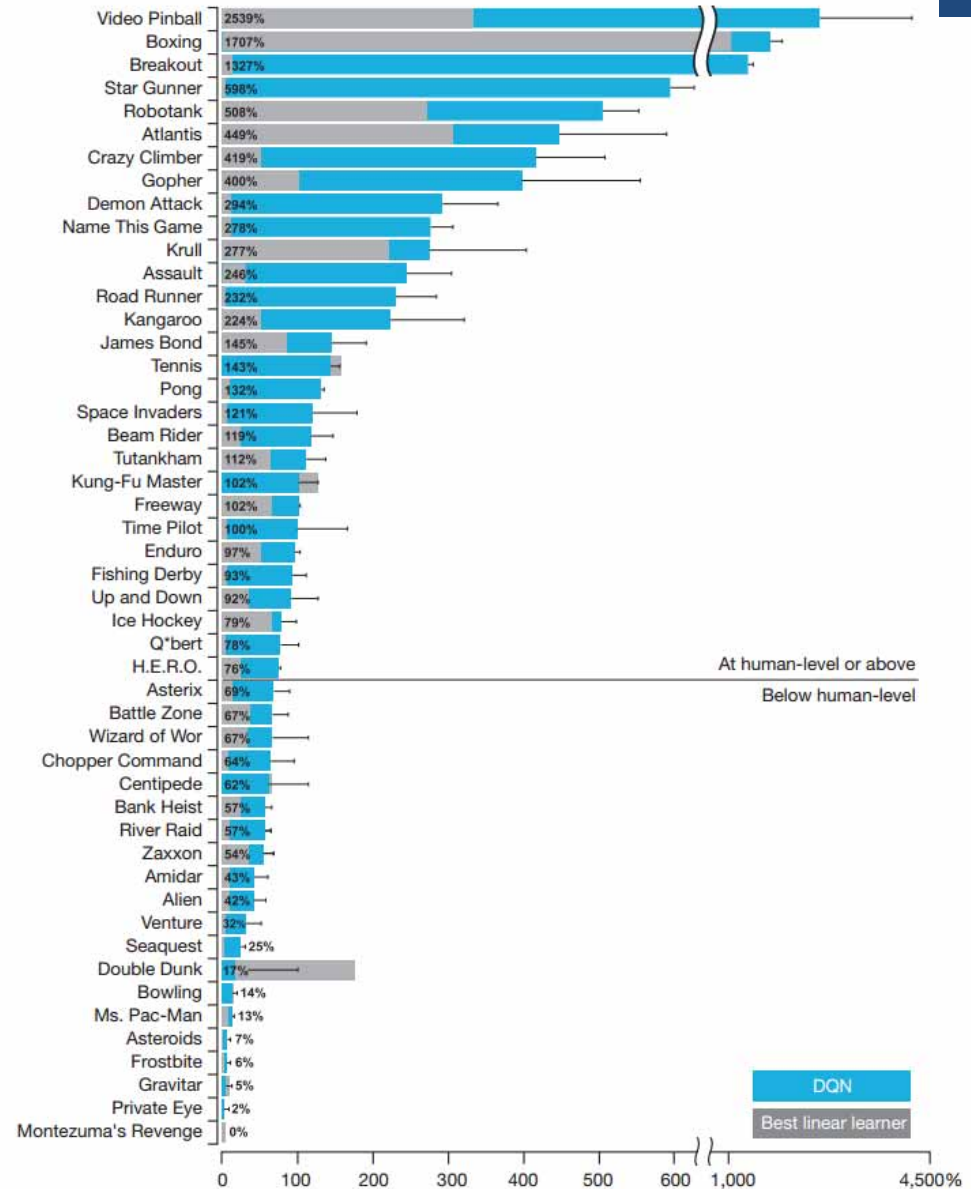
The Royal Society,
Future Directions of Machine Learning Part 2



- 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

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

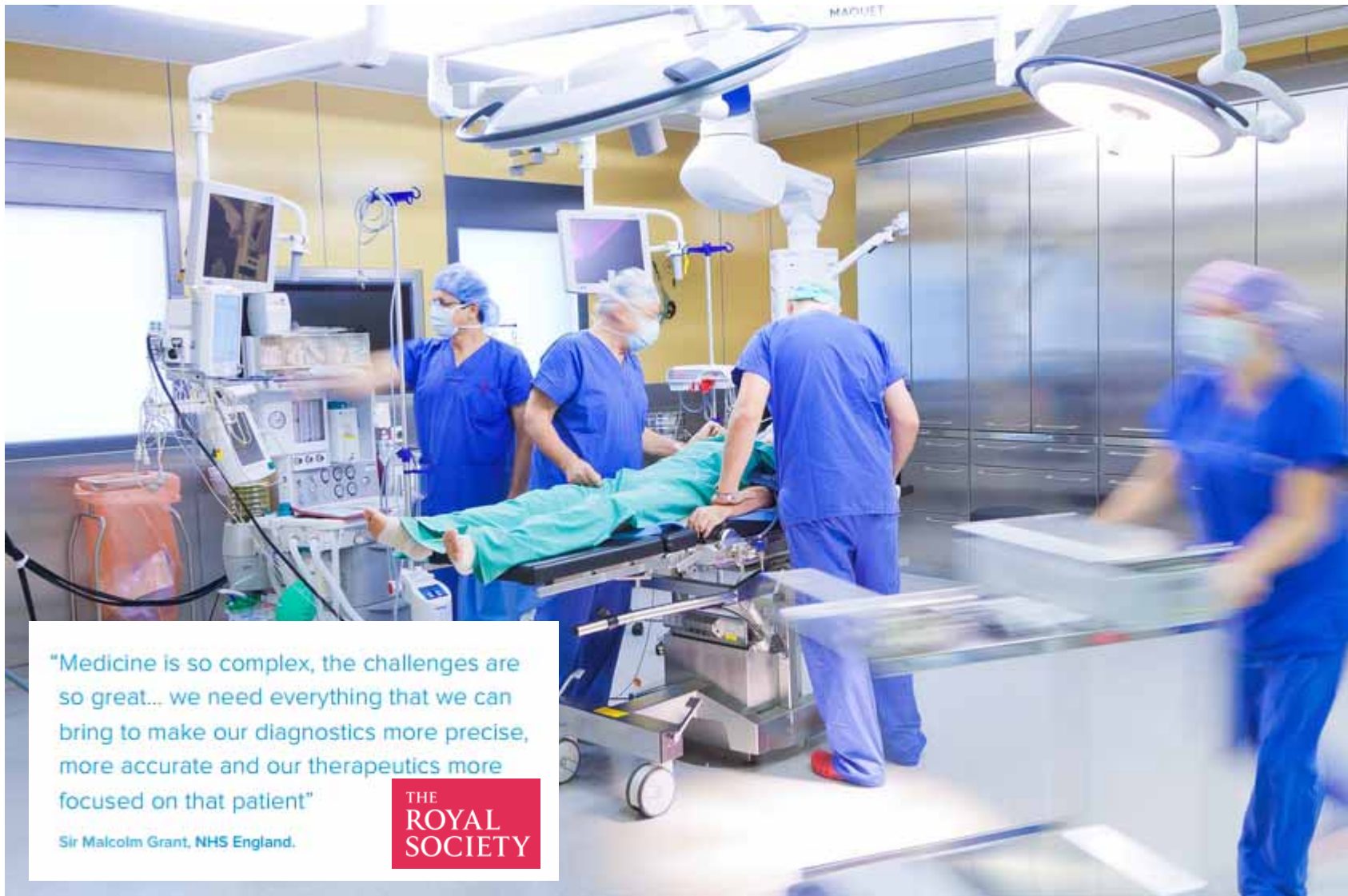


- Alan Turing (1912 – 1954)
- Herbert Simon (1916 – 2001)
- John McCarthy (1927 – 2011)
- Marvin Minsky (1927 – 2016)
- Allen Newell (1927 – 1992)
- ... pleaded for building machines that can learn similar to humans, e.g. like children
- **None of them knew what they were talking about ...**
(Josh Tenenbaum)



03 Application Area Health Informatics

Why is this application area complex ?




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

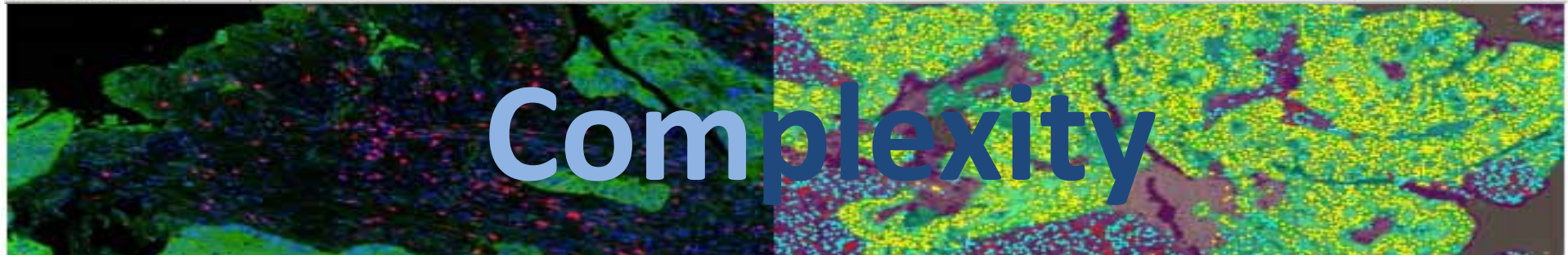
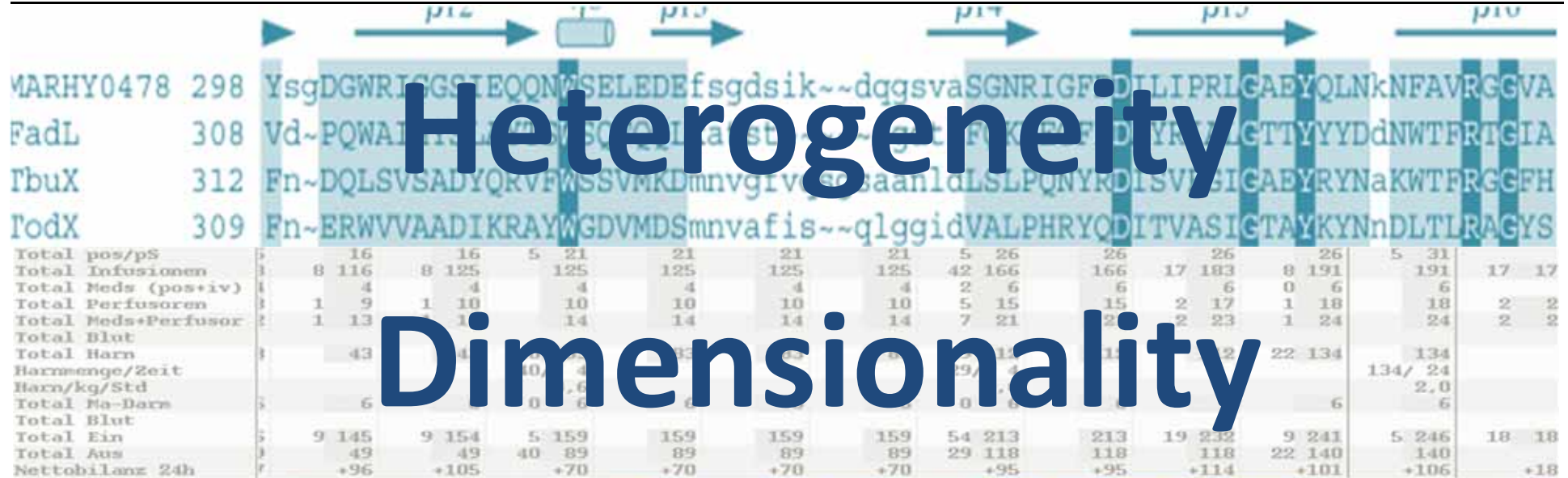


Our central hypothesis: Information may bridge this gap

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



Where is the
problem in building
this bridge?

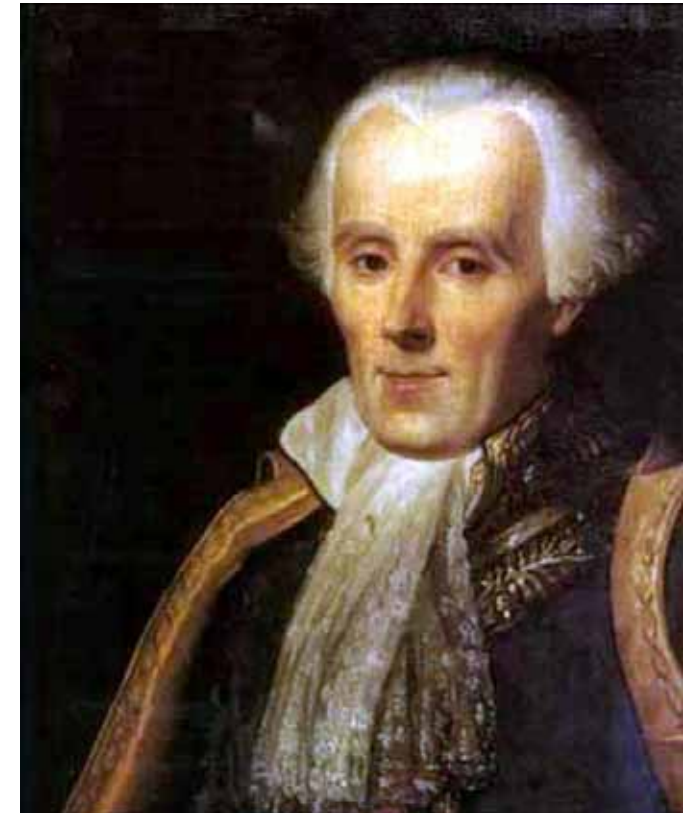


Uncertainty

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.

04 Probabilistic Information $p(x)$

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



Pierre Simon de Laplace (1749-1827), 1812



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:

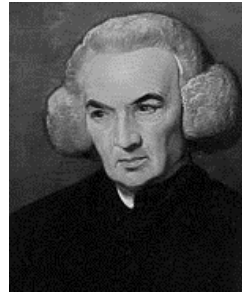
$$\frac{p(x, y) * p(y)}{p(y)} = \frac{p(y|x) * p(x)}{p(y)} \quad (4)$$

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

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



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.

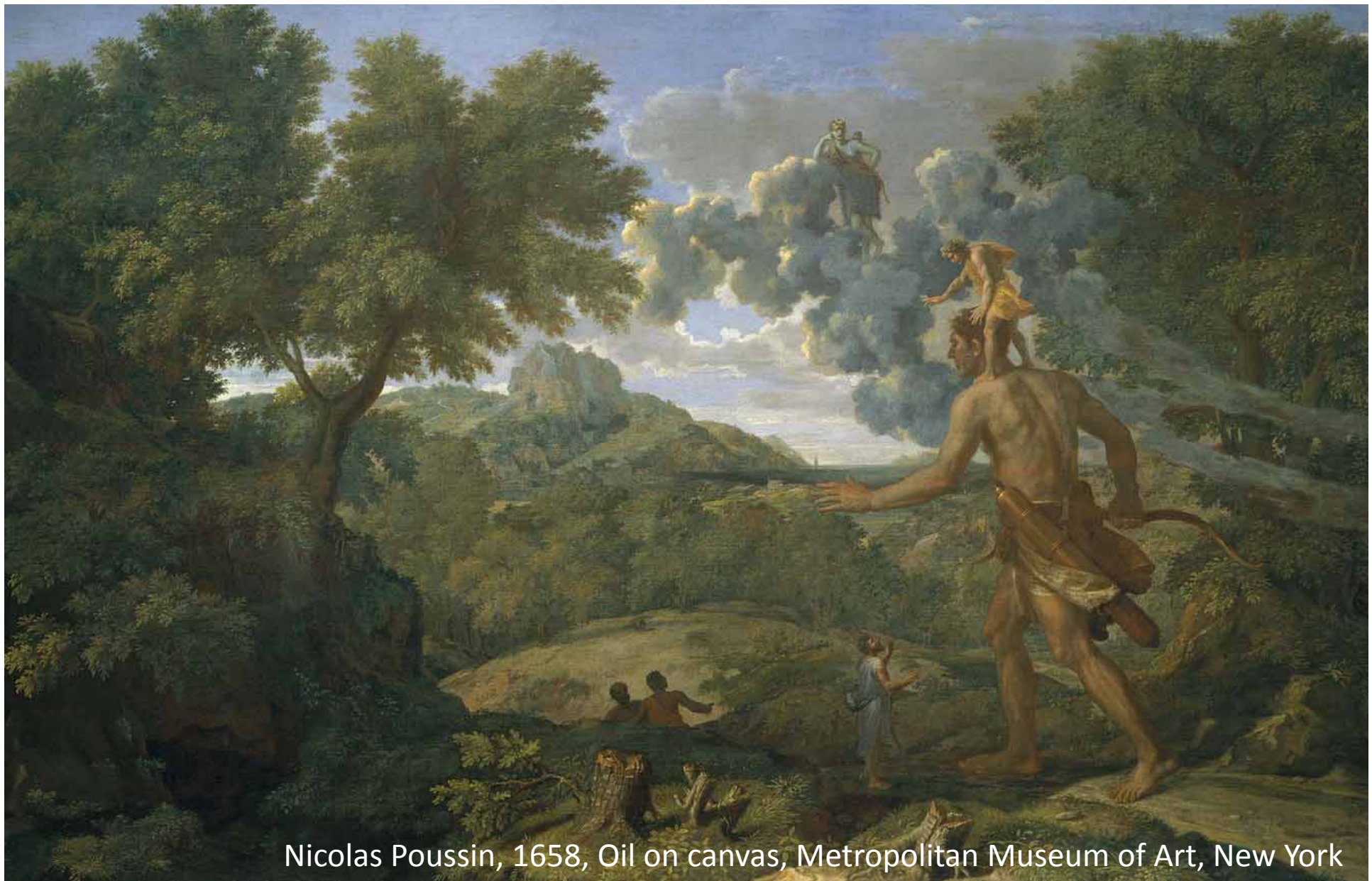
$$p(x_i) = \sum P(x_i, y_j)$$

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

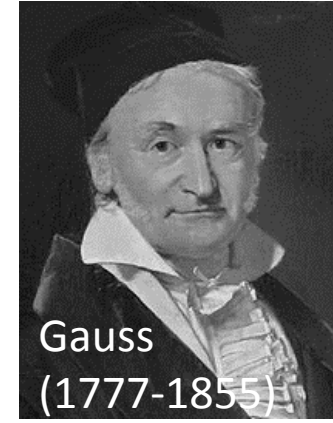
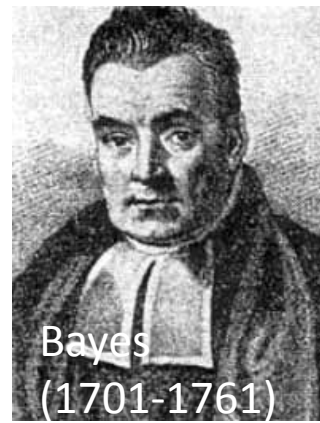
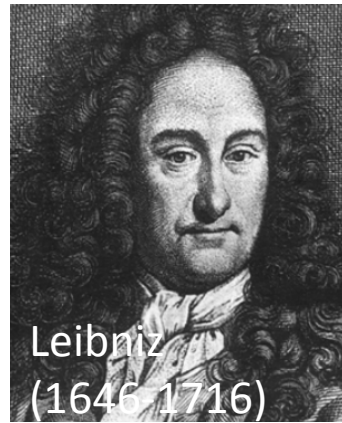
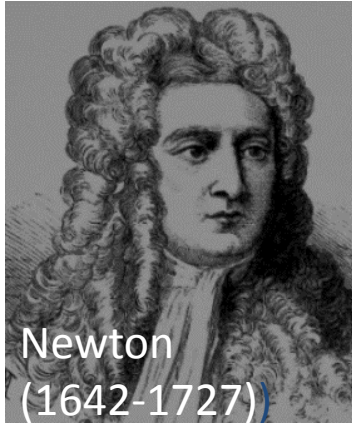
Bayes' Rule is a corollary of the Sum Rule and Product Rule:

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

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.



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



- **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**

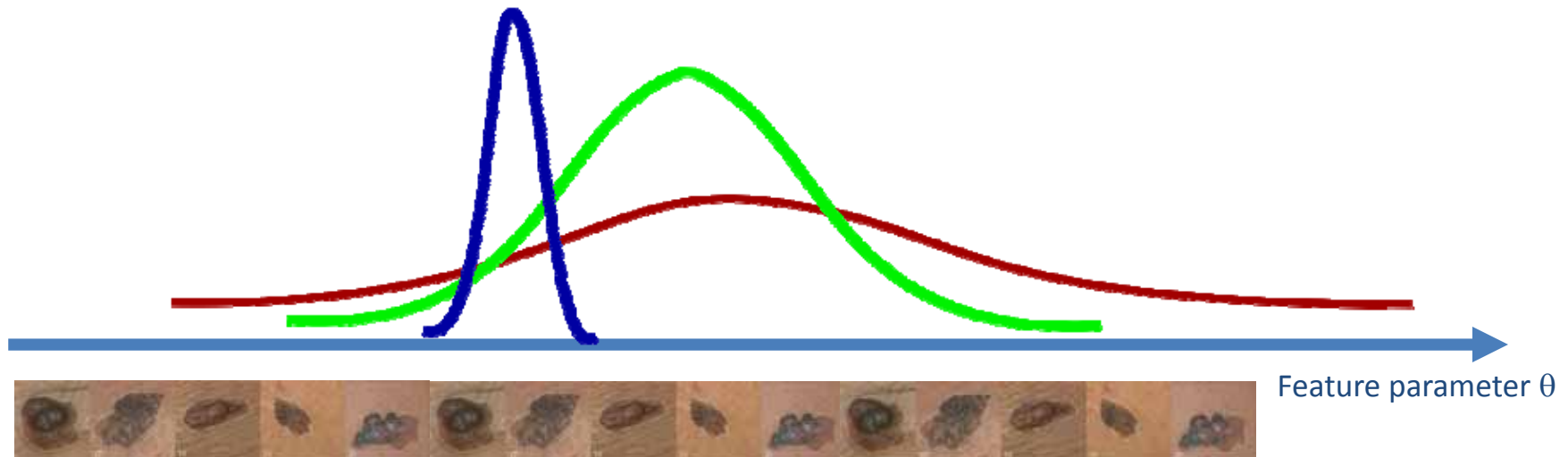
d ... data $\mathcal{H} \dots \{H_1, H_2, \dots, H_n\}$ $\forall h, d \dots$ h ... hypotheses

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

Likelihood

Prior Probability

Posterior Probability

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

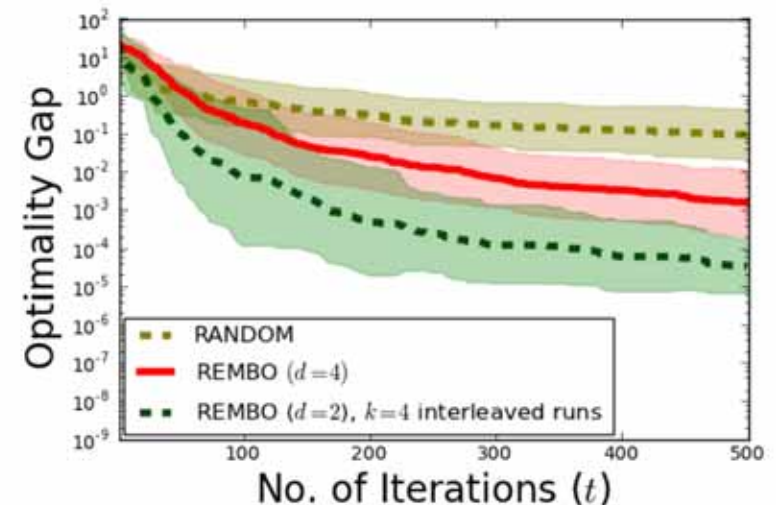
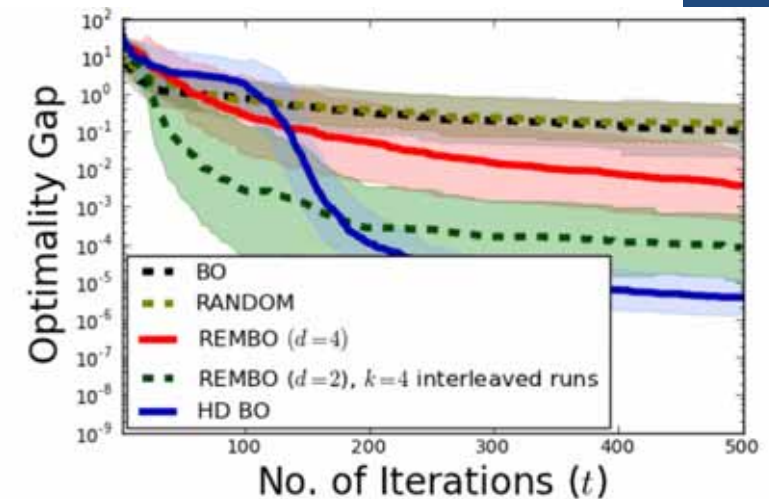
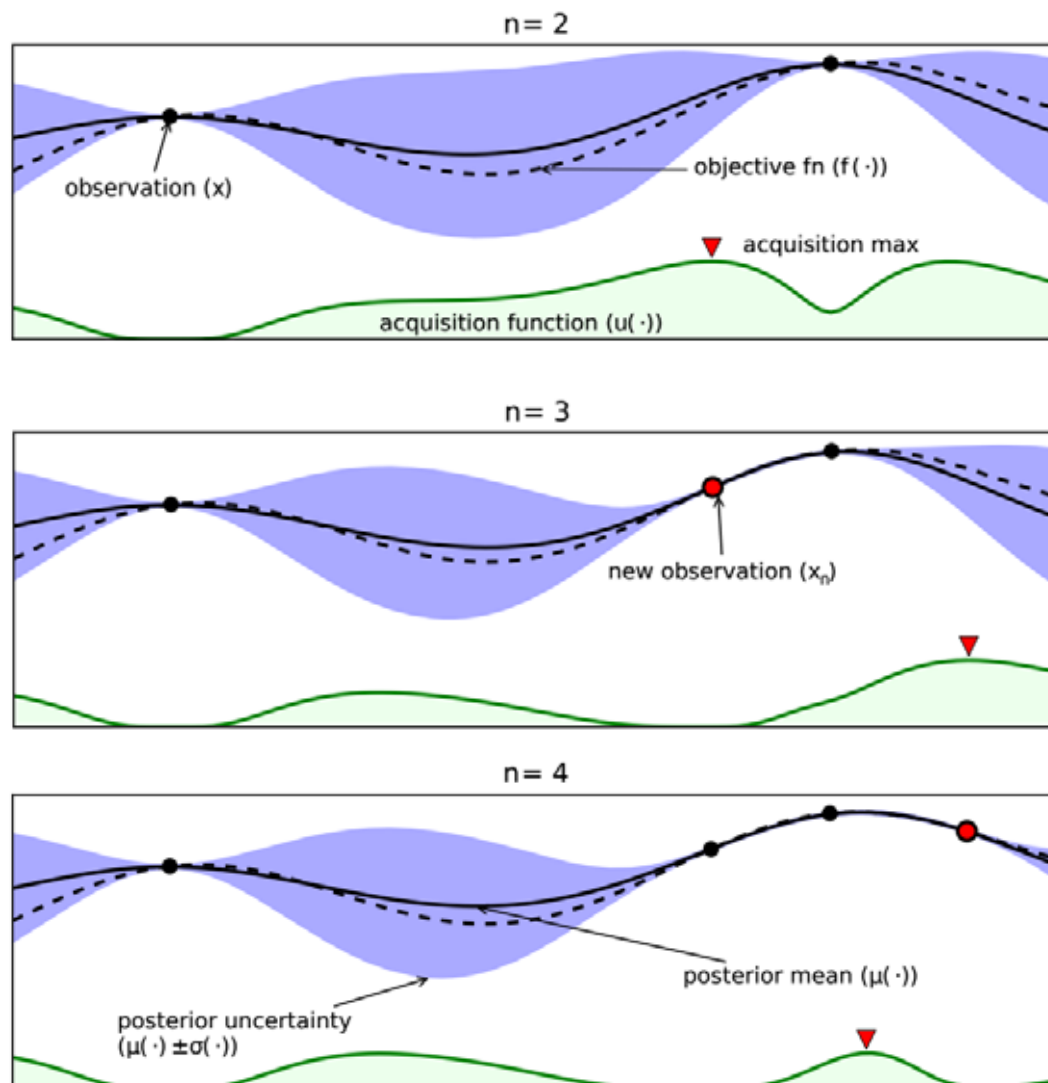
$$\mathcal{D} = x_{1:n} = \{x_1, x_2, \dots, x_n\} \quad p(\mathcal{D}|\theta)$$



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

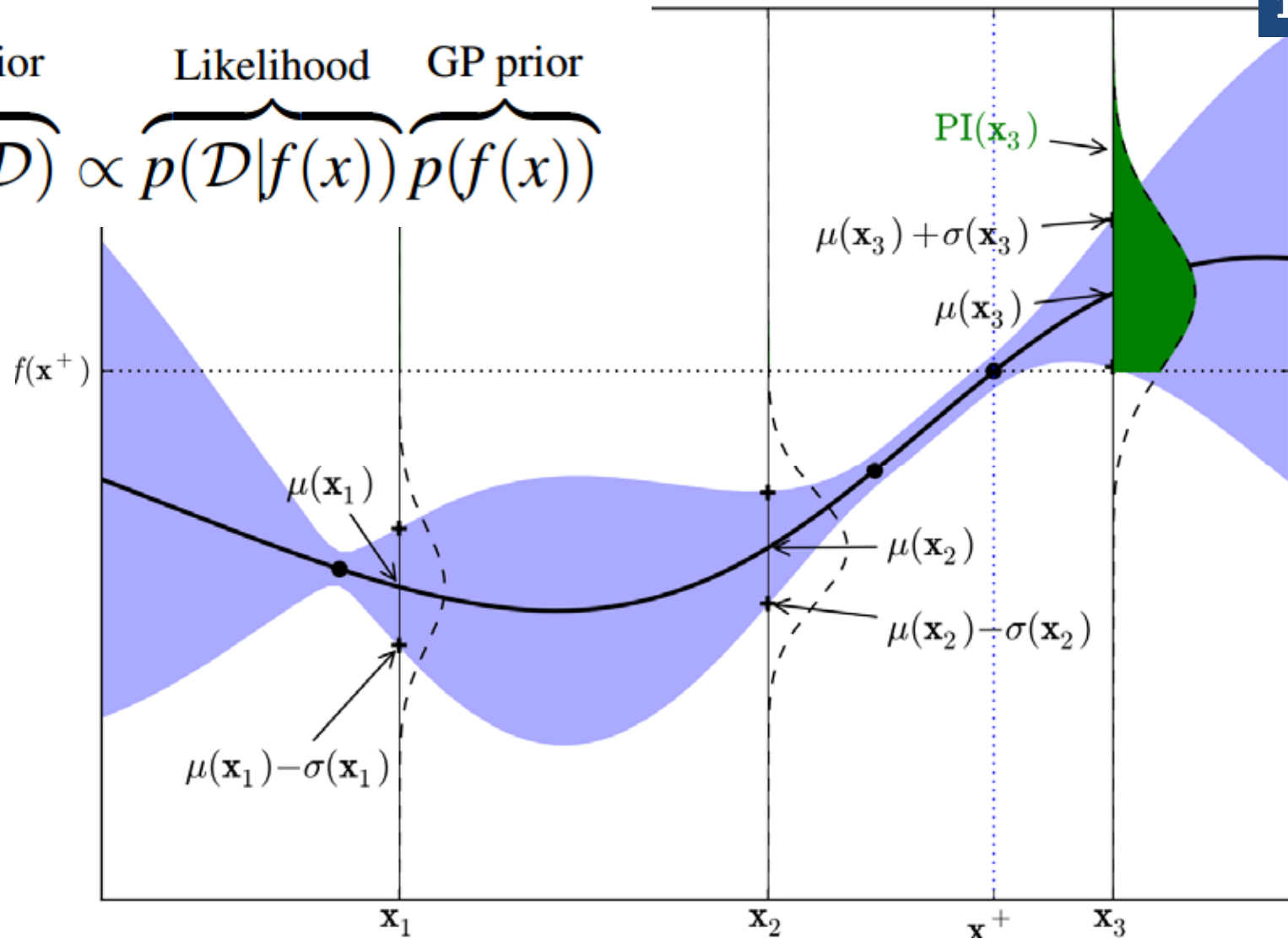
$$\text{posterior} = \frac{\text{likelihood} * \text{prior}}{\text{evidence}}$$

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

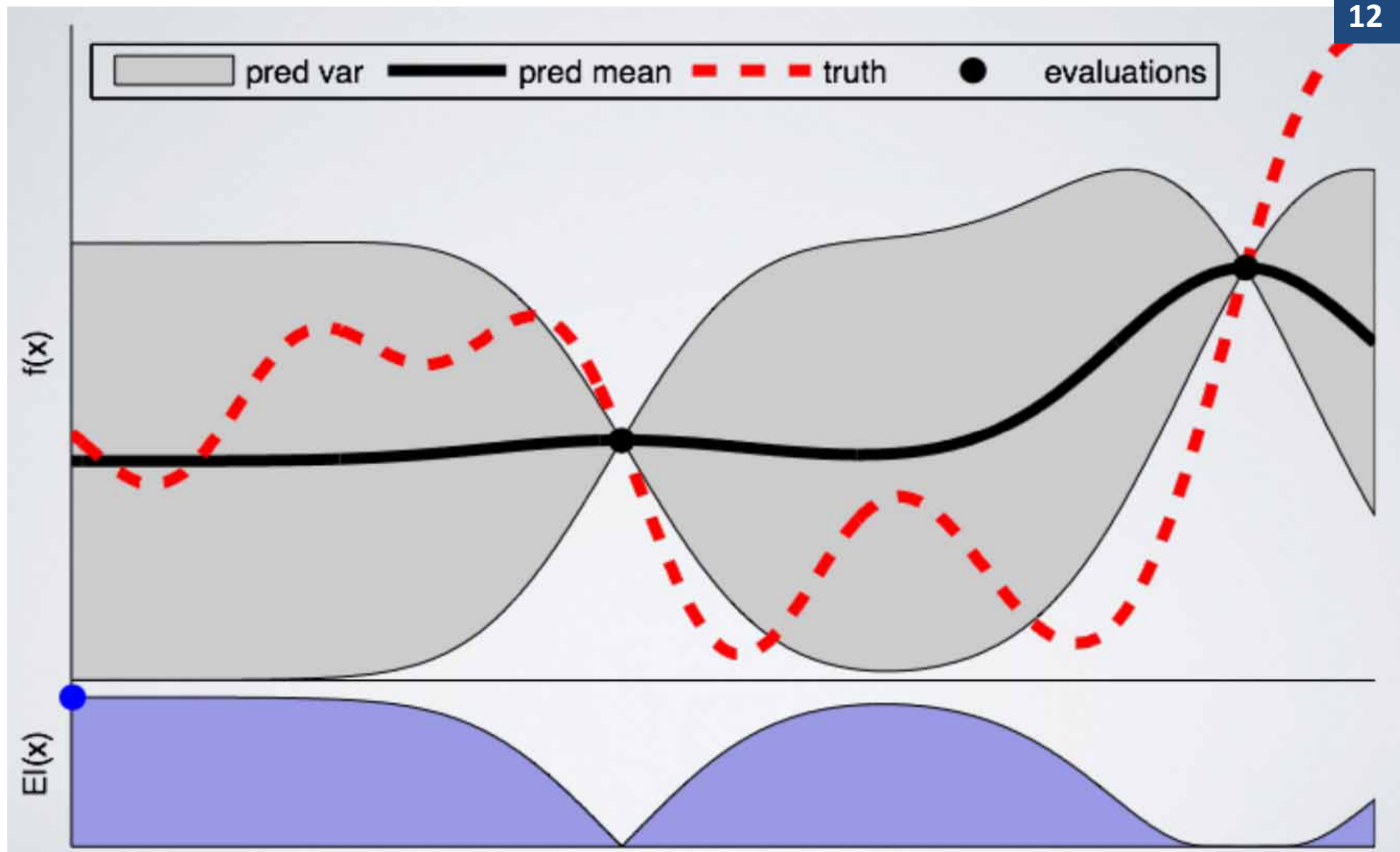


Wang, Z., Hutter, F., Zoghi, M., Matheson, D. & De Feitas, N. 2016. Bayesian optimization in a billion dimensions via random embeddings. Journal of Artificial Intelligence Research, 55, 361-387, doi:10.1613/jair.4806.

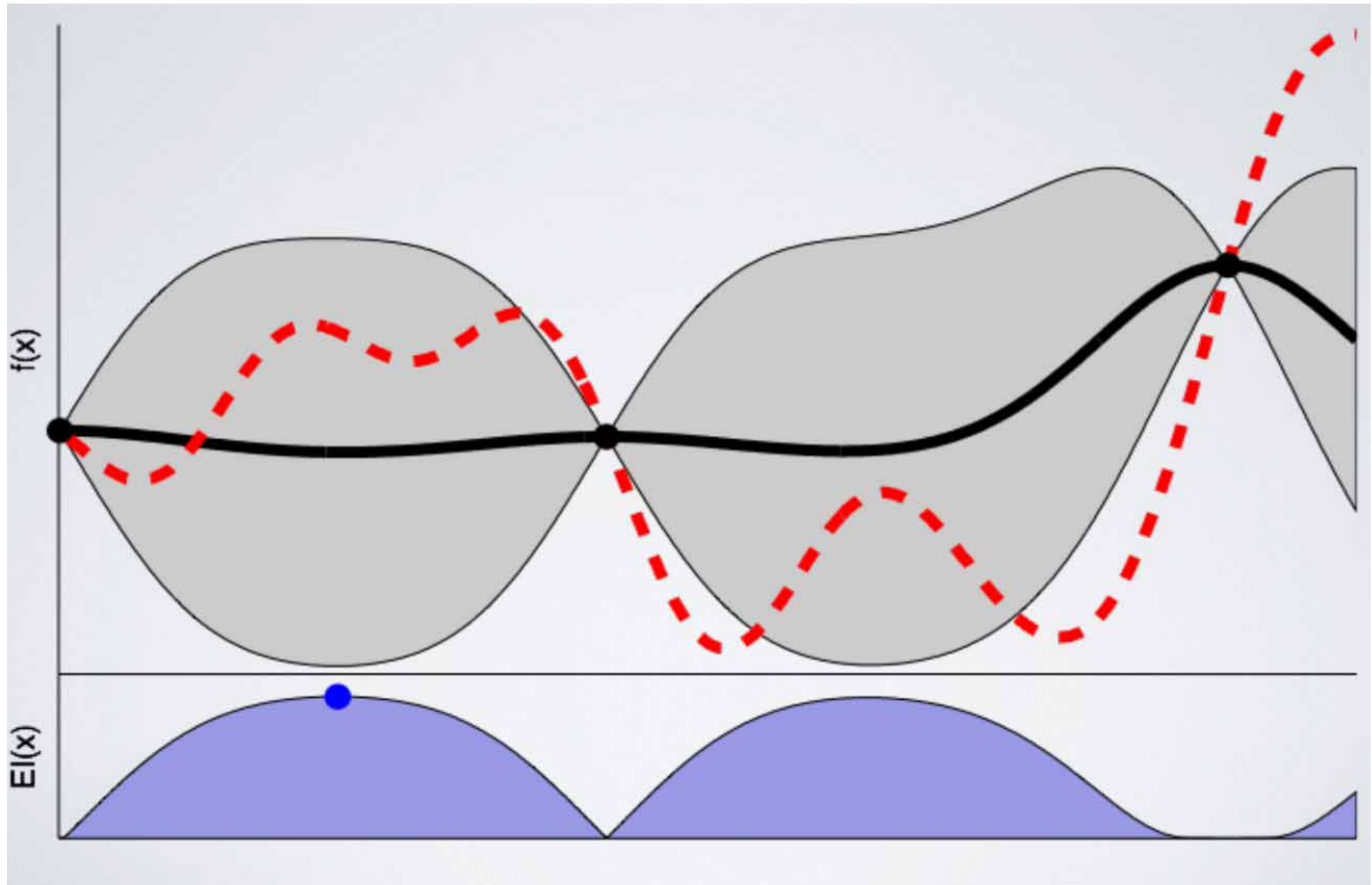
$$\overbrace{p(f(x)|\mathcal{D})}^{\text{GP posterior}} \propto \overbrace{p(\mathcal{D}|f(x))}^{\text{Likelihood}} \overbrace{p(f(x))}^{\text{GP prior}}$$

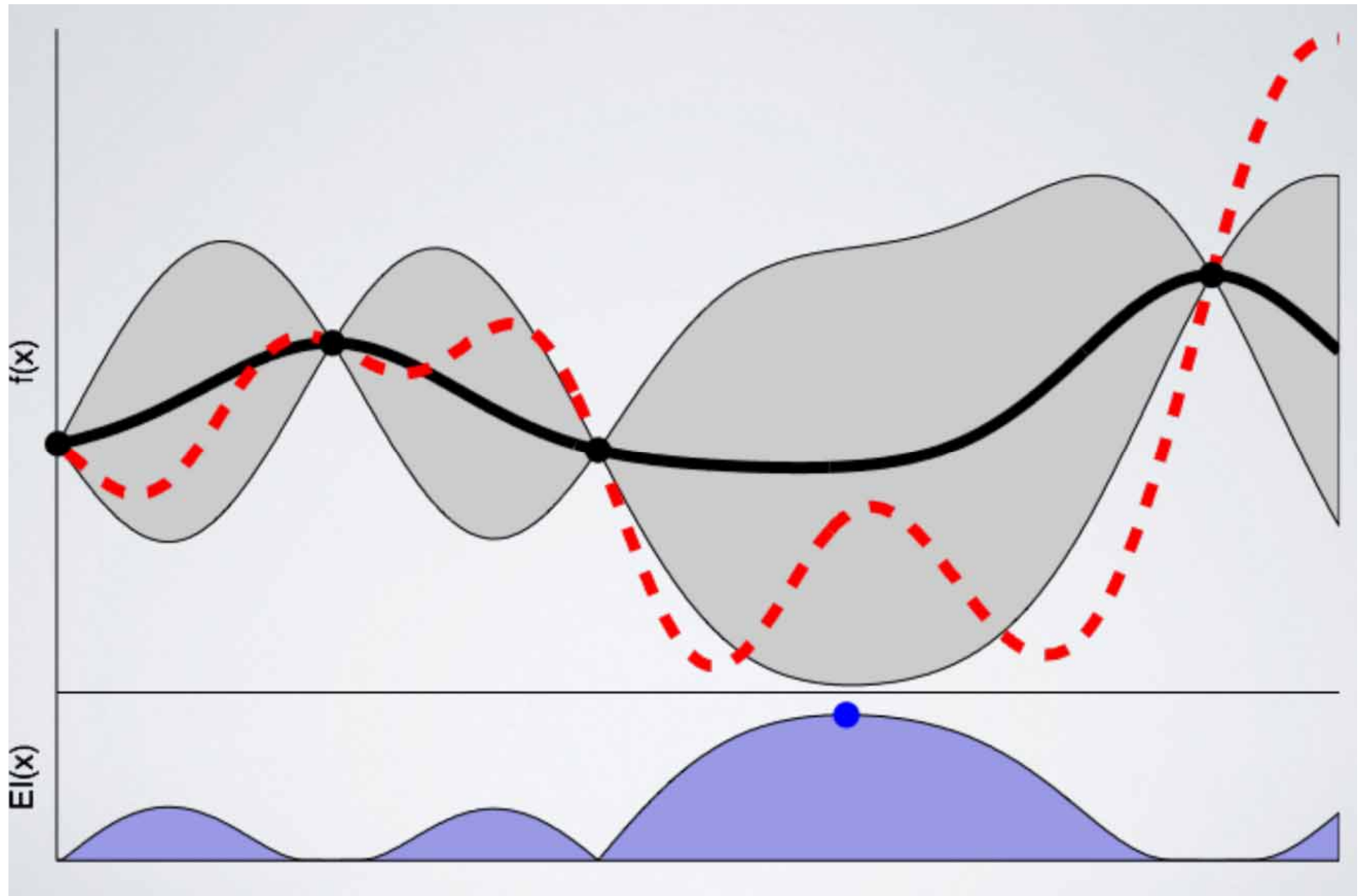


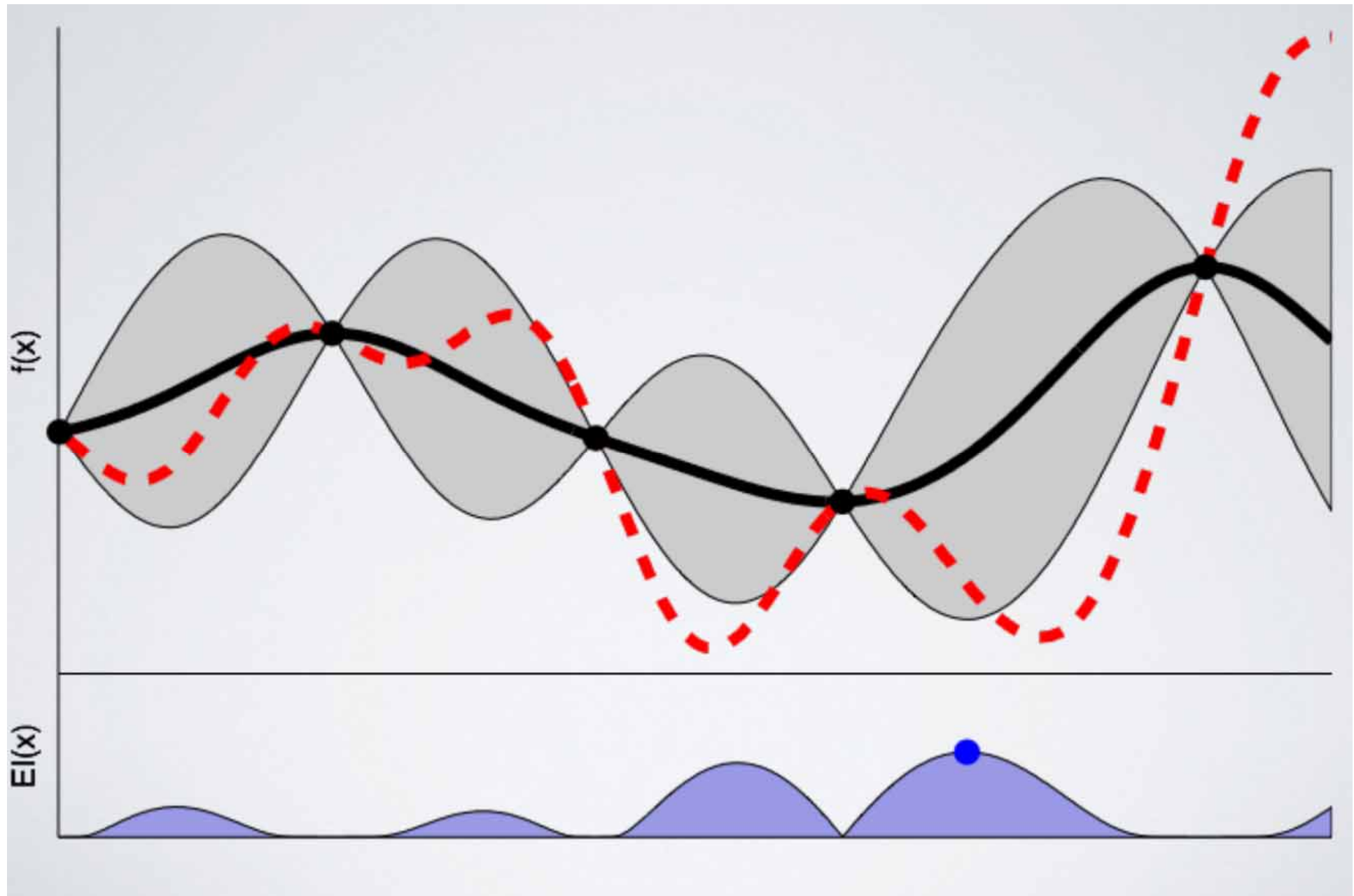
Brochu, E., Cora, V. M. & De Freitas, N. 2010. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. arXiv:1012.2599.

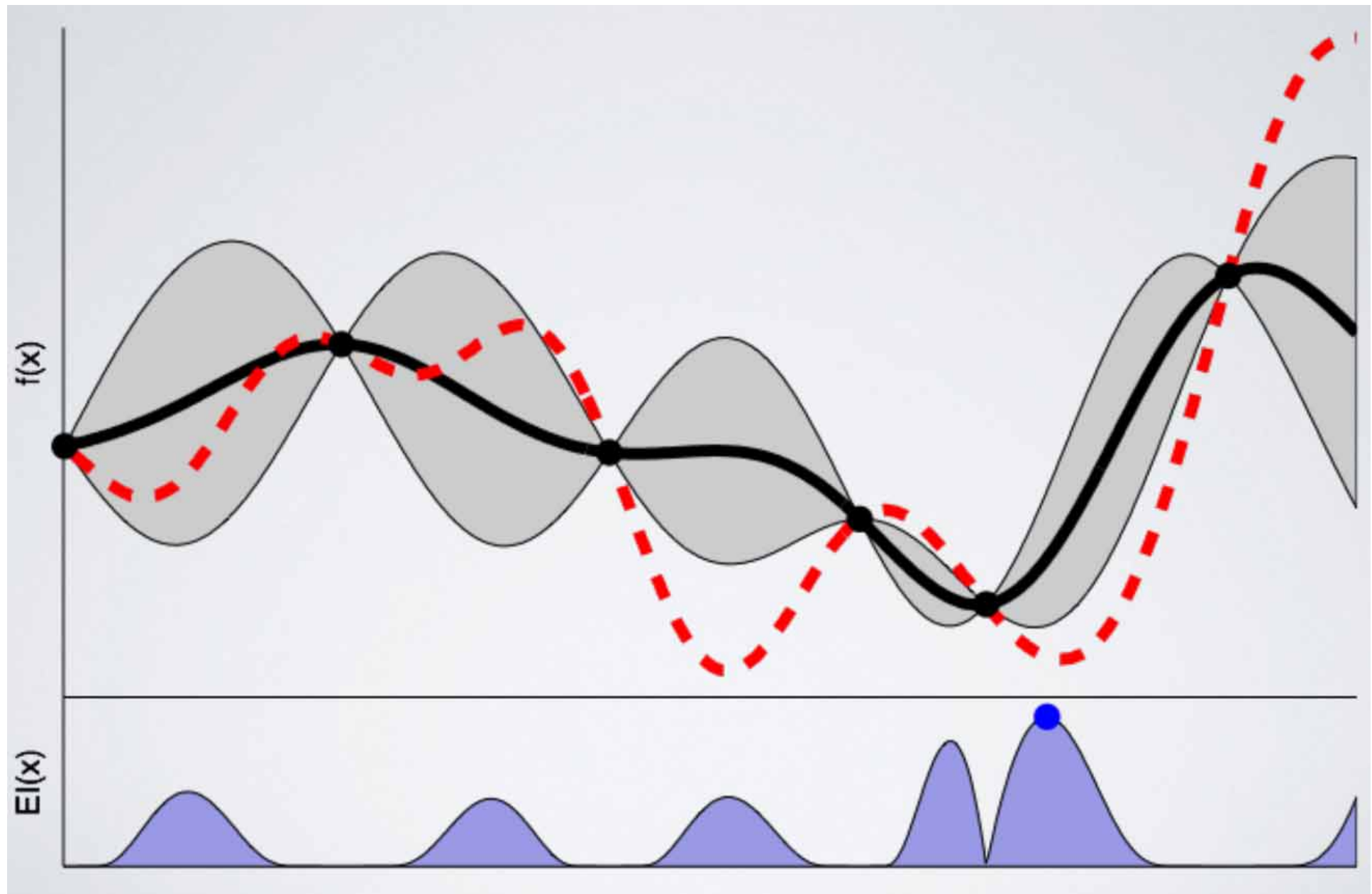


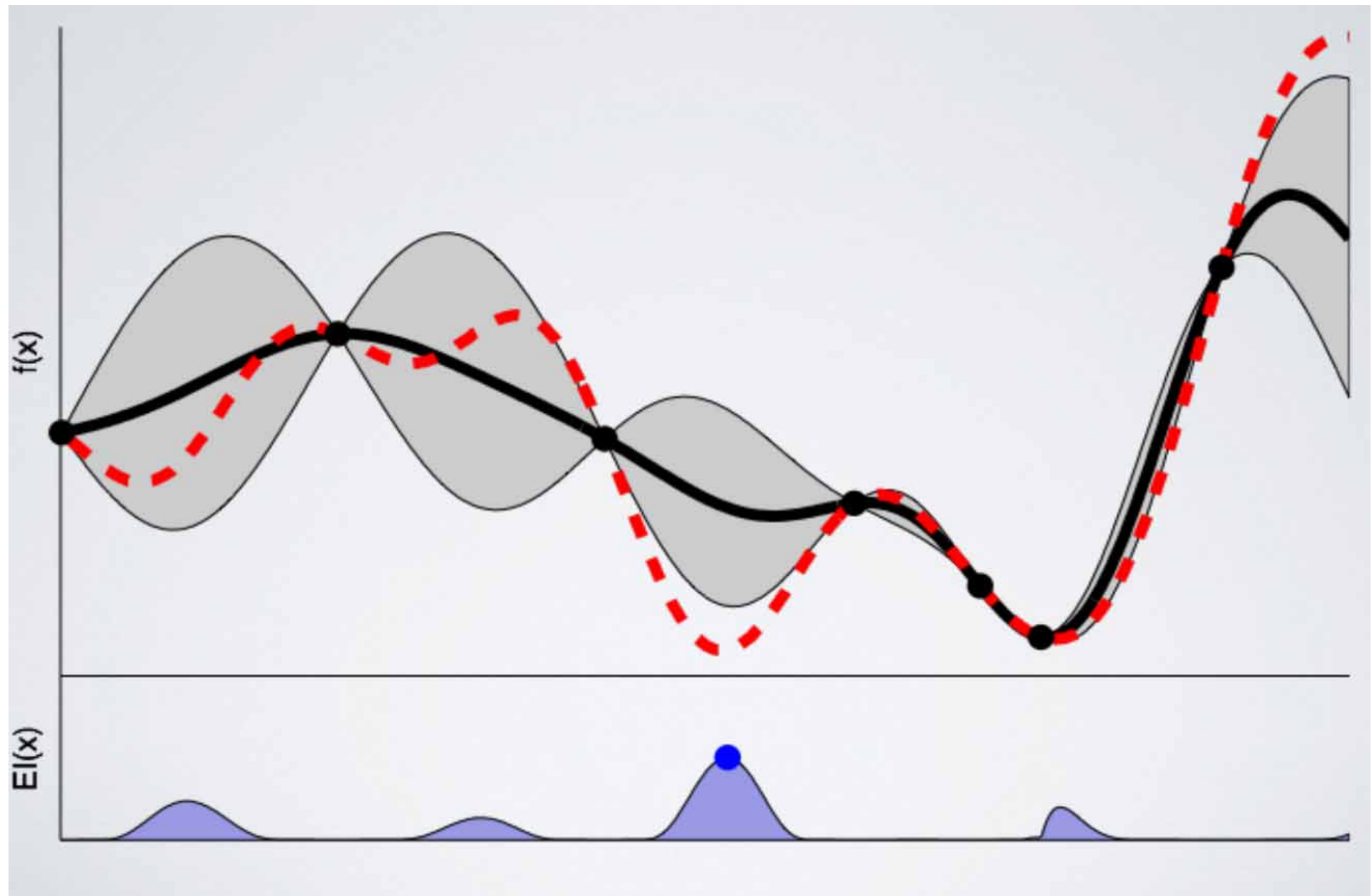
Snoek, J., Larochelle, H. & Adams, R. P. Practical Bayesian optimization of machine learning algorithms. Advances in neural information processing systems, 2012. 2951-2959.

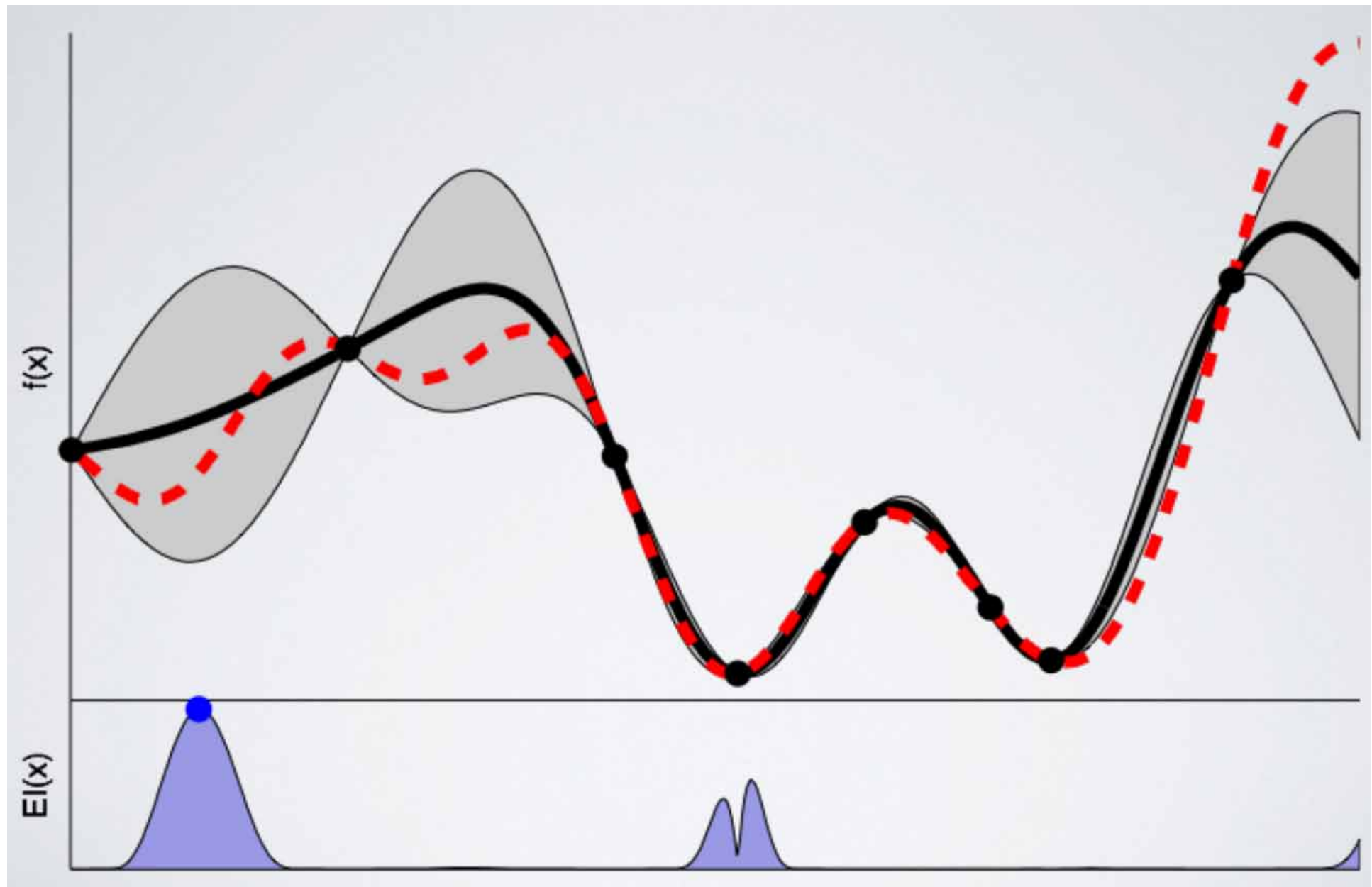


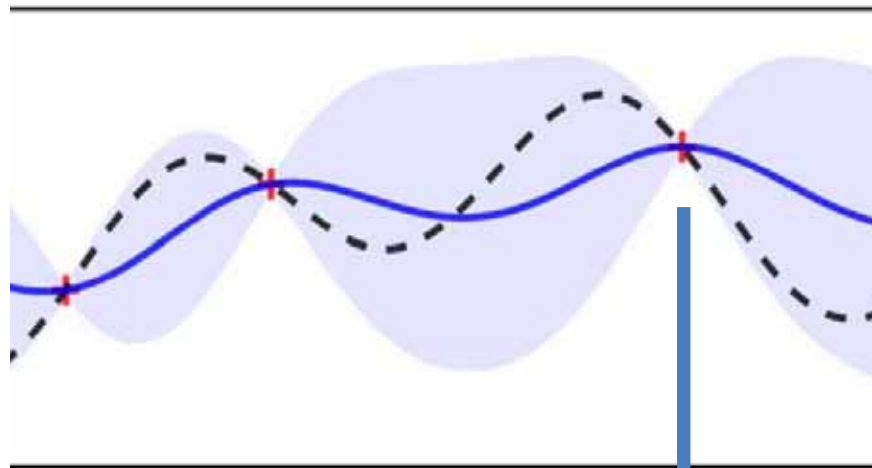




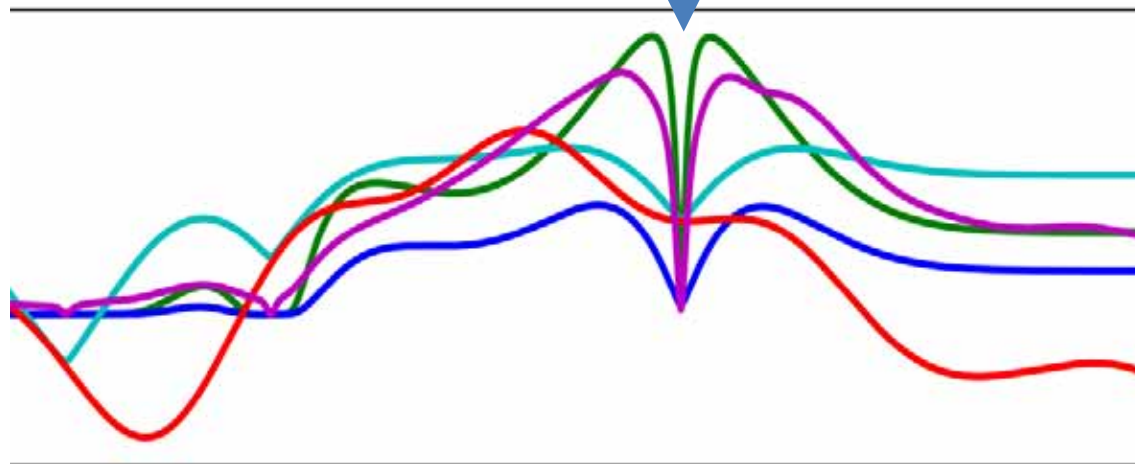






**Algorithm 1** Bayesian optimization

- 1: **for** $n = 1, 2, \dots$ **do**
- 2: select new \mathbf{x}_{n+1} by optimizing acquisition function α
$$\mathbf{x}_{n+1} = \arg \max_{\mathbf{x}} \alpha(\mathbf{x}; \mathcal{D}_n)$$
- 3: query objective function to obtain y_{n+1}
- 4: augment data $\mathcal{D}_{n+1} = \{\mathcal{D}_n, (\mathbf{x}_{n+1}, y_{n+1})\}$
- 5: update statistical model
- 6: **end for**



- PI Probability of Improvement
- EI Expected Improvement
- UCB Upper Confidence Bound
- TS Thompson Sampling
- PES Predictive Entropy Search

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.

05 aML

- 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**

Best practice examples of aML ...

amazon.co.uk
Try Prime

All glass cutter circular

Shop by
Department

Your Amazon.co.uk

Today's Deals

Gift Cards & Top Up

Sell

Help

Hello
You

Amazon.co.uk

Today's Deals

Warehouse Deals

Outlet

Subscribe & Save

Vouchers

Amazon Family

Amazon Prime

Amazon Video

Amazon Student

Mobile Apps

An

Showing results for "glass cutter circular"

Show results for

DIY & Tools >

Glass Cutters

Cold Chisels

Power Tools

Sports & Outdoors >

Compasses

+ See All 131

Refine by

Delivery Opt

☒ Prime☐ Free UK Del

Brand

☐ sourcingm☐ SODIAL(R

Silverline 101228 Circular Glass Cutter with 65-300 mm Diameter 10 Oct 2014

by Silverline

£7.81 £10.02 Prime

Get it by Tomorrow, Sep 5

Eligible for FREE UK Delivery

More buying choices

£6.40 new (22 offers)

★★★★☆ 42

DIY & Tools: See all 162 items



Highlander 3 Hole Thinsulate Balaclava

by Highlander

£1.99 - £7.00 Prime

More buying choices

£1.99 new (5 offers)

★★★★☆ 163

Sports & Outdoors: See all 5,918 items



Sanwood® Outdoor Motorcycle Cycling Ski Neck Protecting Lycra Balaclava Full Face Mask

by Phoenix B2C UK

£1.74 - £3.57

More buying choices

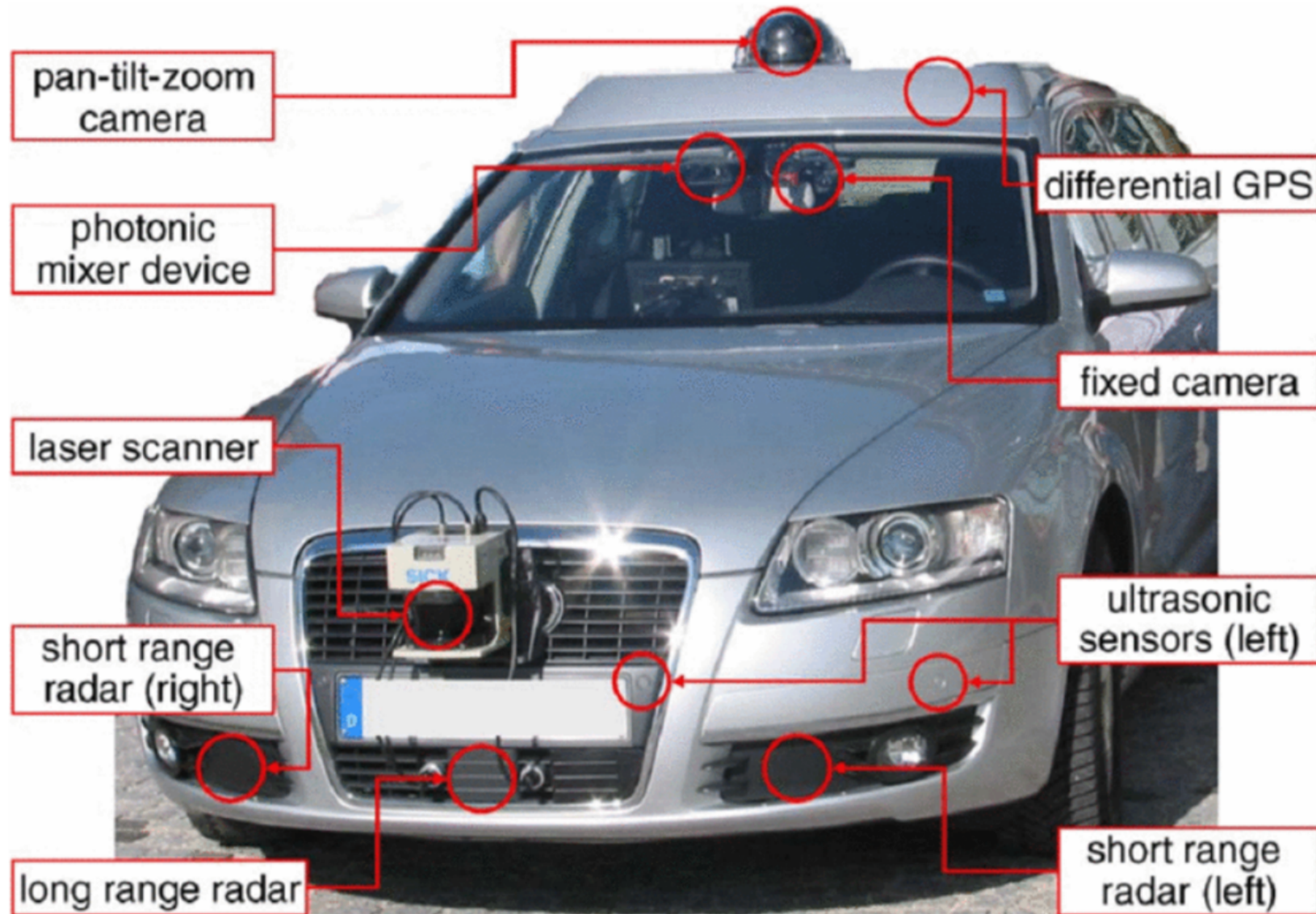
£0.01 new (4 offers)

★★★★☆ 73

Sports & Outdoors: See all 5,918 items



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



Mukhtar, A., Xia, L. & Tang, T. B. 2015. Vehicle Detection Techniques for Collision Avoidance Systems: A Review. IEEE Transactions on Intelligent Transportation Systems, 16, (5), 2318-2338, doi:10.1109/TITS.2015.2409109.



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

1960s Citroën DS driverless car test



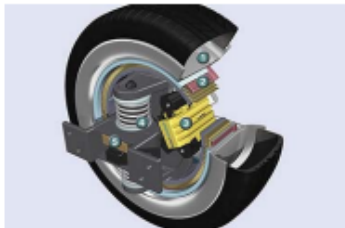
Sunday Times Driving



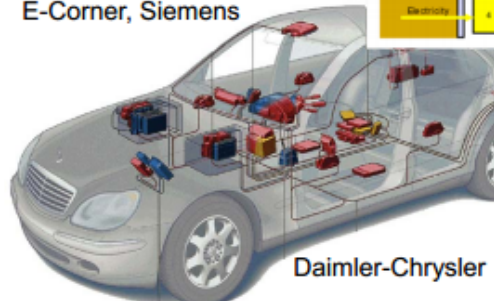
8,605 views

Cyber-Physical Systems (CPS):
*Tight integration of networked computation
 with physical systems*

Automotive

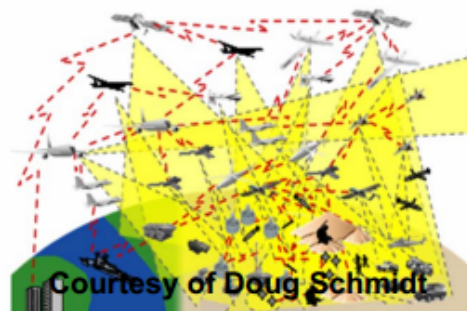


E-Corner, Siemens



Daimler-Chrysler

Military systems:

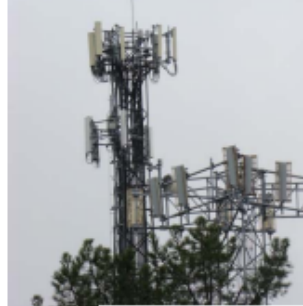


Courtesy of Doug Schmidt

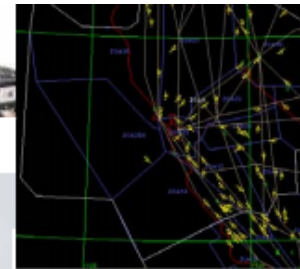
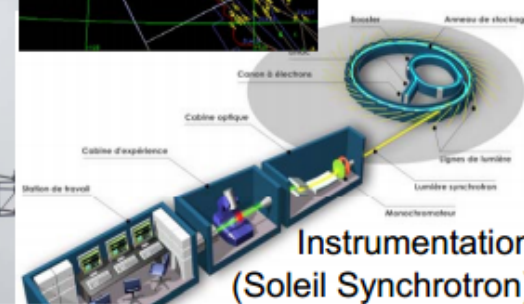
Building Systems



Telecommunications



Avionics

Transportation
(Air traffic control at SFO)Instrumentation
(Soleil Synchrotron)

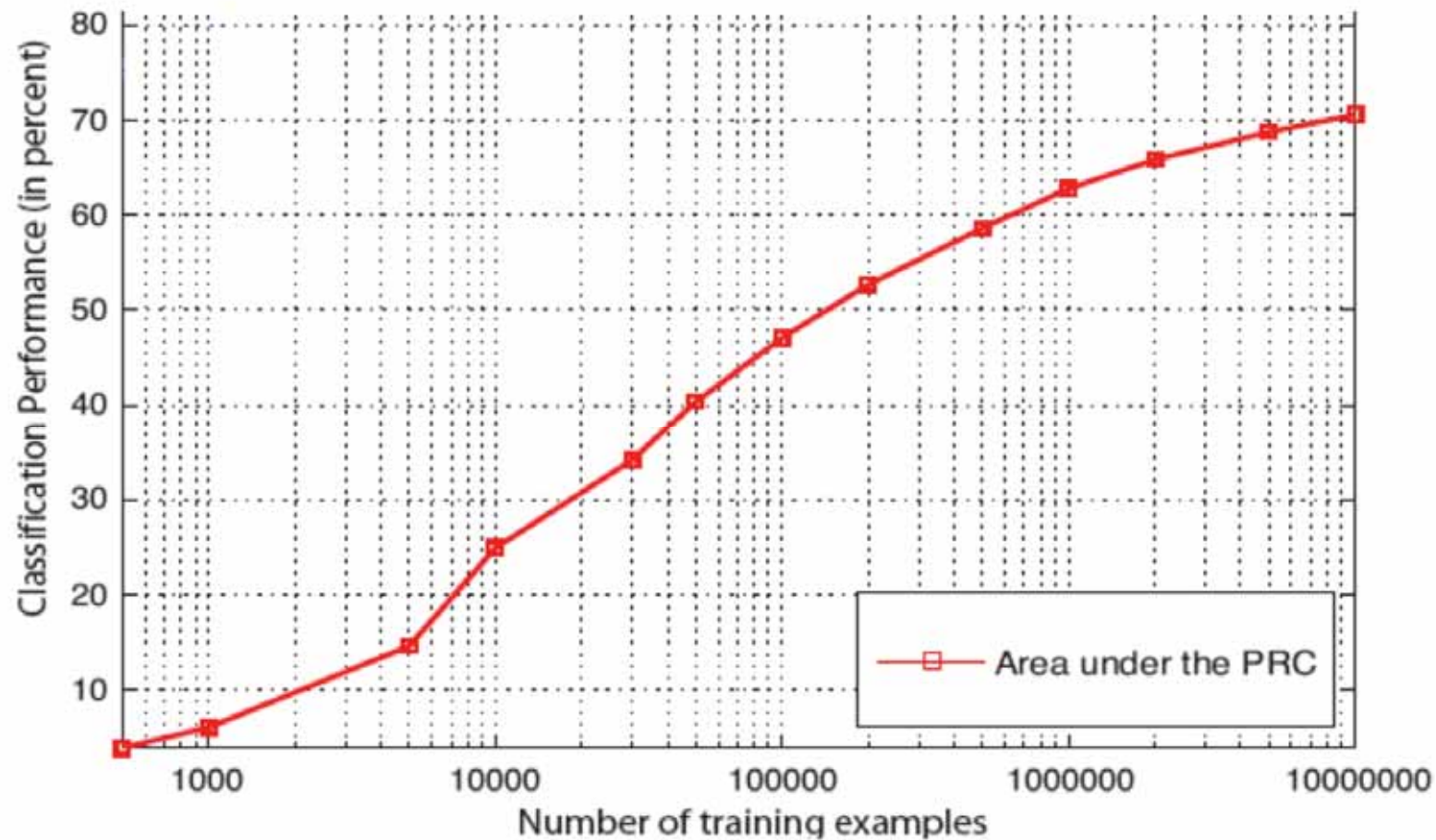
Factory automation



Courtesy of Kuka Robotics Corp.

Power
generation and
distributionCourtesy of
General Electric

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**

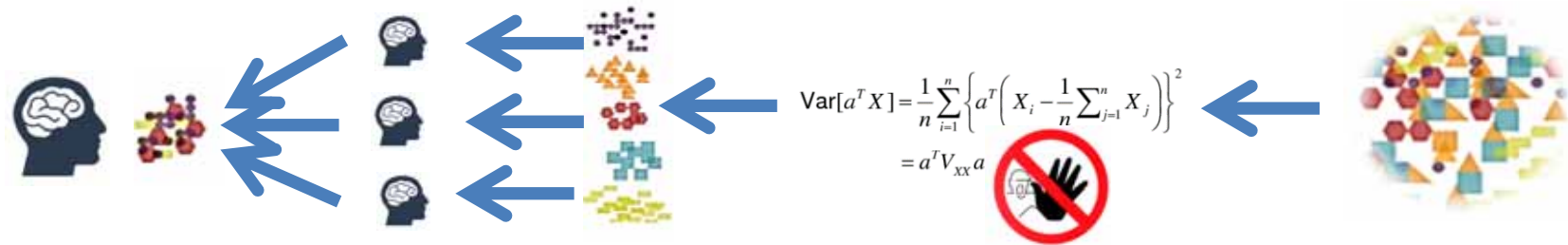
Holzinger, A. 2016. Interactive Machine Learning (iML). Informatik Spektrum, 39, (1), 64-68, doi:10.1007/s00287-015-0941-6.



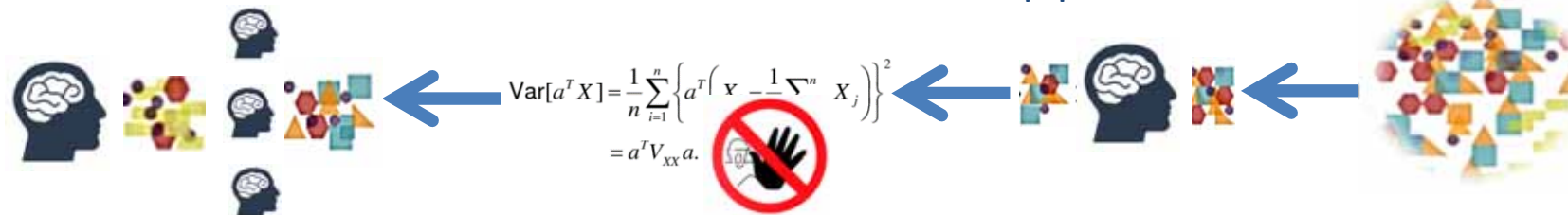




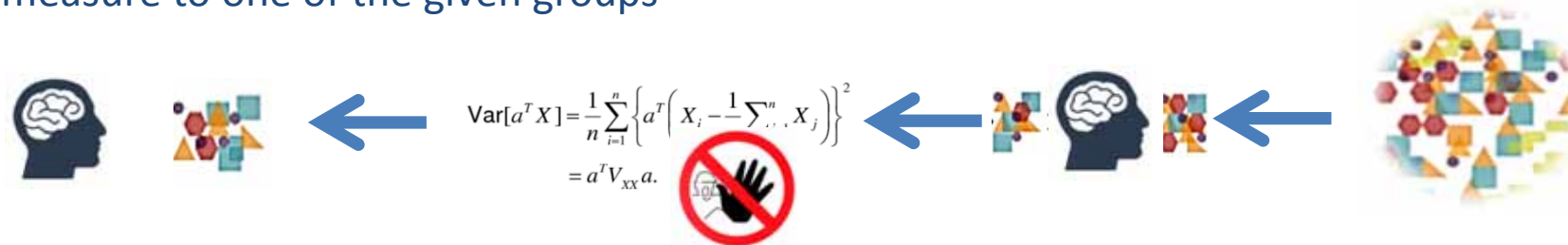
A) Unsupervised ML: Algorithm is applied on the raw data and learns fully automatic – Human can check results at the end of the ML-pipeline



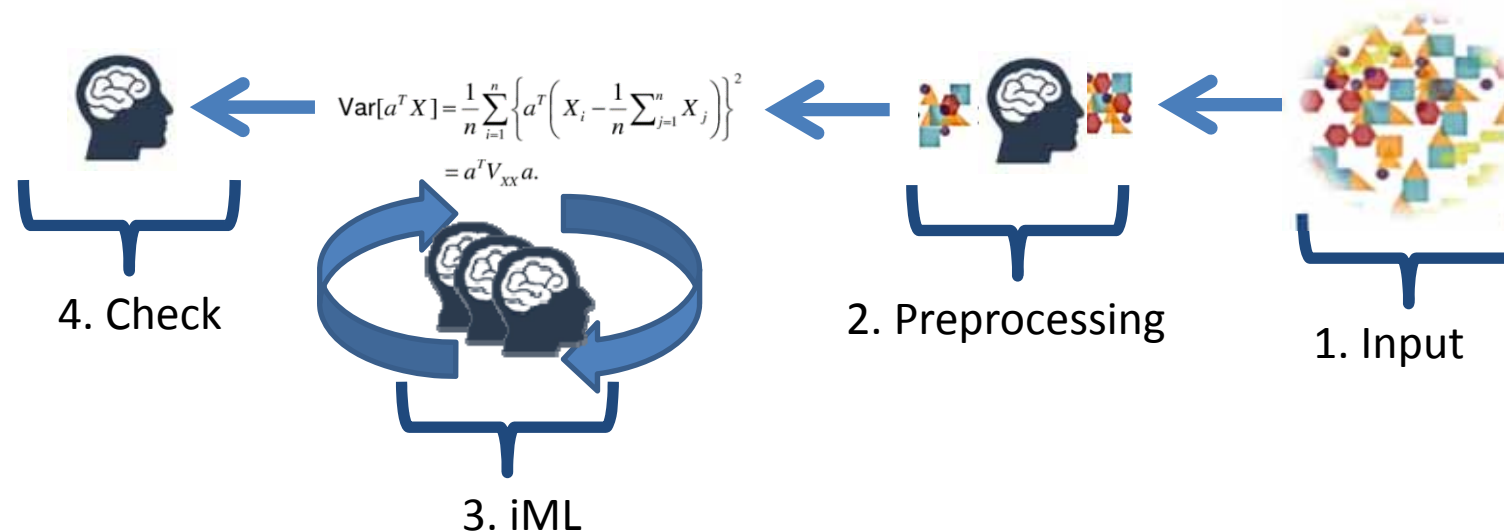
B) Supervised ML: Humans are providing the labels for the training data and/or select features to feed the algorithm to learn – the more samples the better – Human can check results at the end of the ML-pipeline



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



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 ...



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

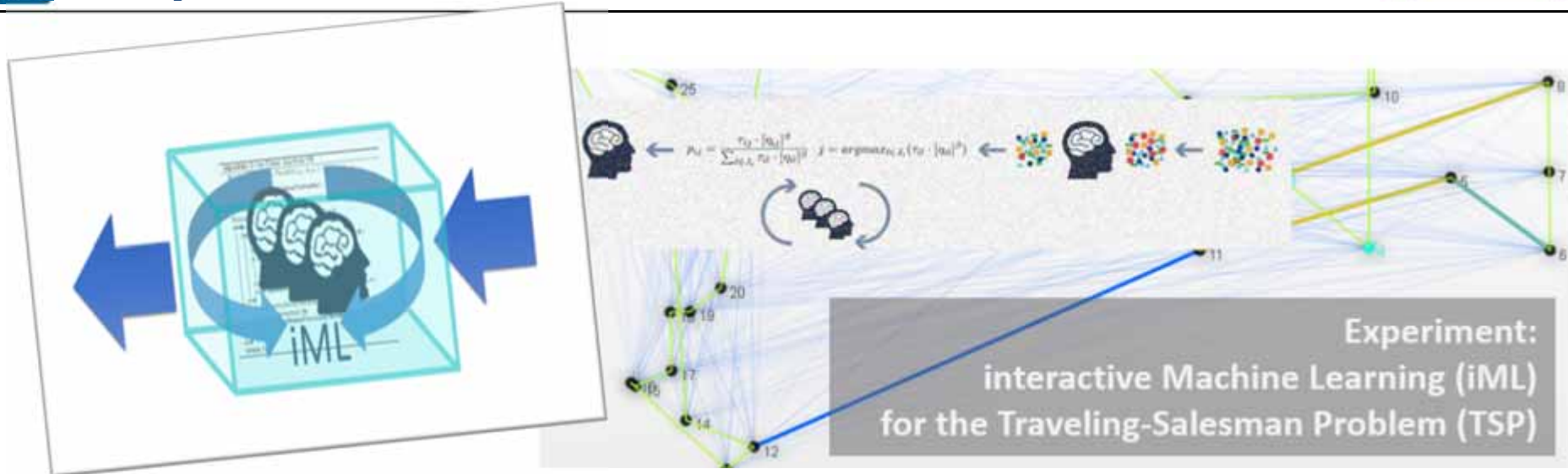
Holzinger, A. 2016. Interactive Machine Learning for Health Informatics: When do we need the human-in-the-loop? Brain Informatics (BRIN), 3, (2), 119-131, doi:10.1007/s40708-016-0042-6.

- **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.



- 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.

```

Input : ProblemSize,  $m$ ,  $\beta$ ,  $\rho$ ,  $\sigma$ ,  $q_0$ 
Output:  $P_{best}$ 
 $P_{best} \leftarrow \text{CreateHeuristicSolution}(\text{ProblemSize});$ 
 $P_{best\_cost} \leftarrow \text{Cost}(P_{best});$ 
 $\text{Pheromone}_{init} \leftarrow \frac{1.0}{\text{ProblemSize} \times P_{best\_cost}};$ 
 $\text{Pheromone} \leftarrow \text{InitializePheromone}(\text{Pheromone}_{init});$ 
while  $\neg \text{StopCondition}()$  do
  for  $i = 1$  to  $m$  do
     $S_i \leftarrow \text{ConstructSolution}(\text{Pheromone}, \text{ProblemSize}, \beta, q_0);$ 
     $S_{i\_cost} \leftarrow \text{Cost}(S_i);$ 
    if  $S_{i\_cost} \leq P_{best\_cost}$  then
       $P_{best\_cost} \leftarrow S_{i\_cost};$ 
       $P_{best} \leftarrow S_i;$ 
    end
     $\text{LocalUpdateAndDecayPheromone}(\text{Pheromone}, S_i, S_{i\_cost}, \rho);$ 
  end
   $\text{GlobalUpdateAndDecayPheromone}(\text{Pheromone}, P_{best}, P_{best\_cost}, \rho);$ 
  while  $\text{isUserInteraction}()$  do
     $\text{GlobalAddAndRemovePheromone}(\text{Pheromone}, P_{best}, P_{best\_cost}, \rho);$ 
  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.

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>

- *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.

07 Active Representation Learning

- “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.

Δελφοί



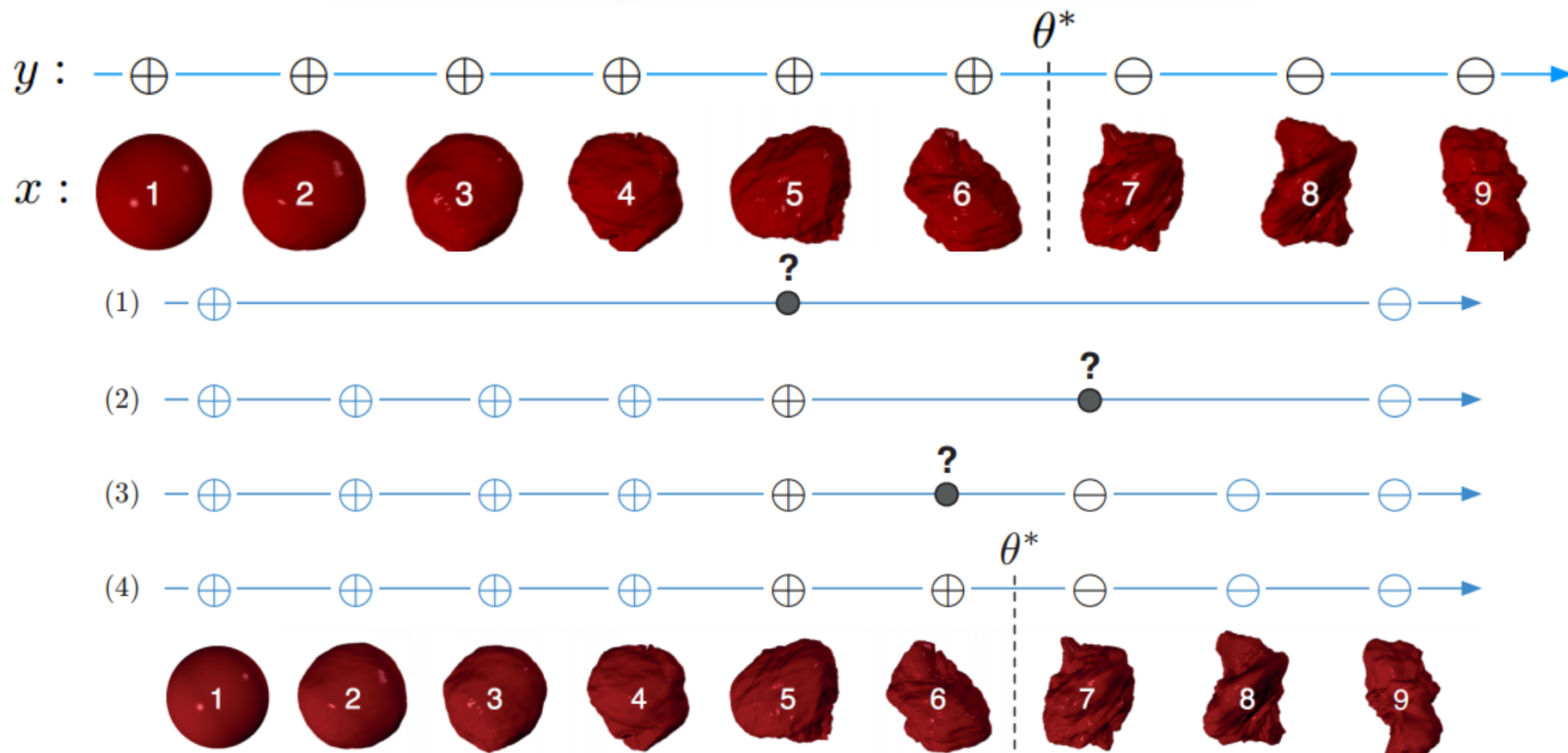
<https://en.wikipedia.org/wiki/Delphi>

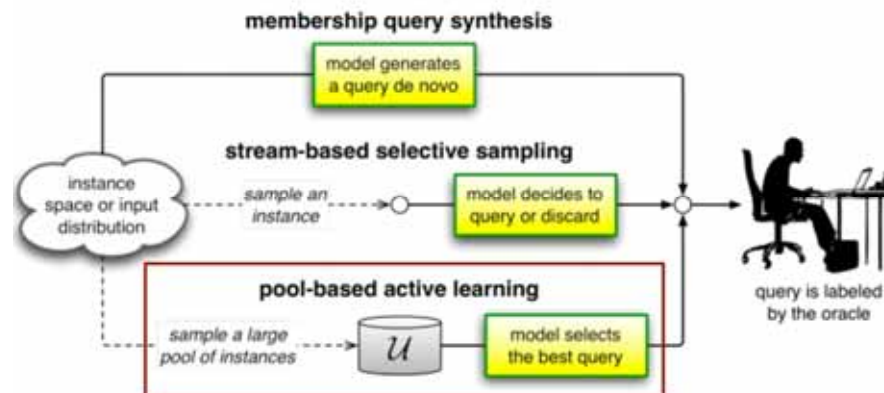
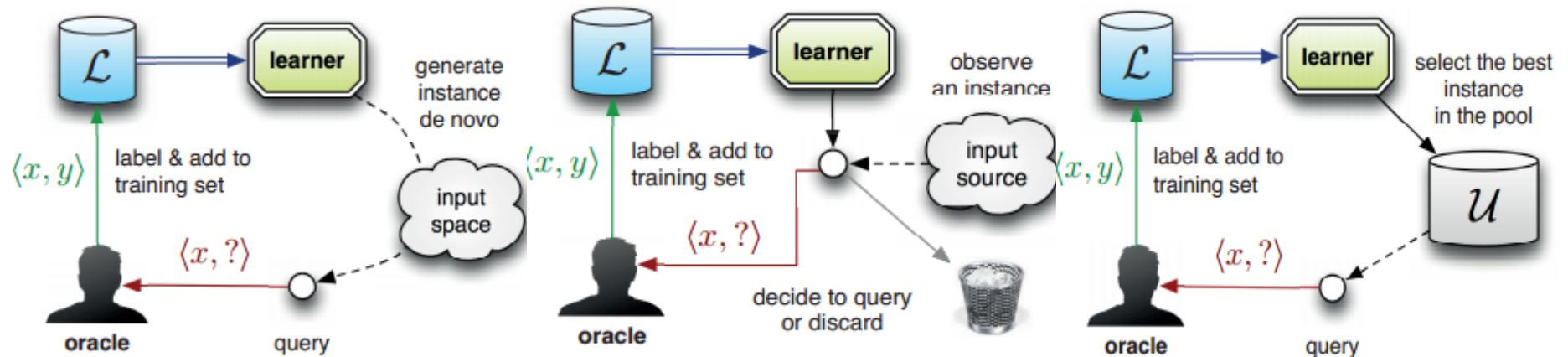
- := ML algorithm can perform better with less training if it is allowed to choose the data from which it learns.
- “Active learner” may pose queries, usually in the form of unlabeled data instances to be labeled by an “oracle” (e.g., a human annotator) that **understands** the **context** of the problem.
- 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.

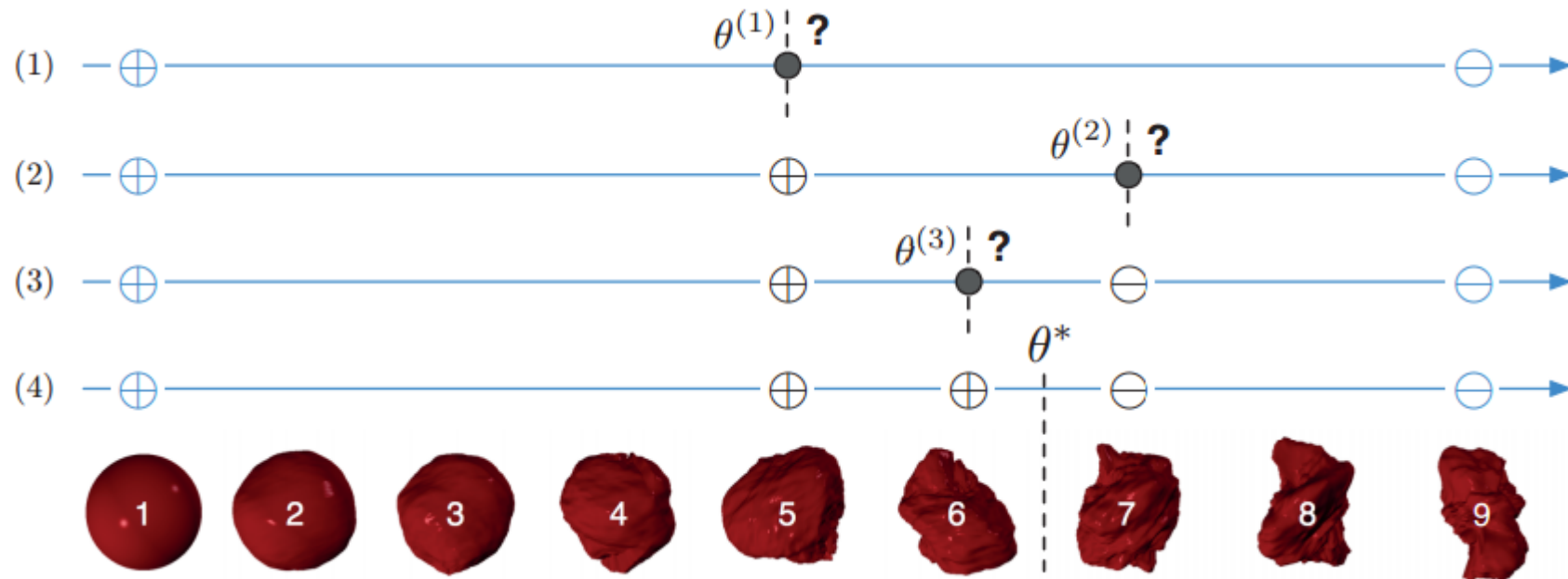
- A classifier to determine objects as a function mapping $h: X \rightarrow Y$, parameterized by a threshold θ :

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



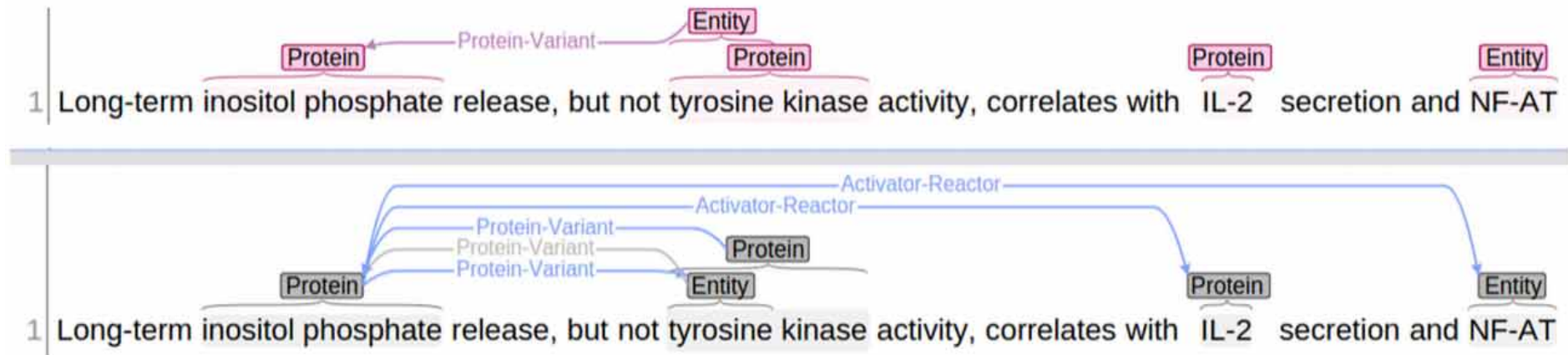


Settles, B. 2012. Active Learning, San Rafael (CA), Morgan & Claypool, doi:10.2200/S00429ED1V01Y201207AIM018.



- 1: \mathcal{U} = a pool of unlabeled instances $\{x^{(u)}\}_{u=1}^U$
- 2: \mathcal{L} = set of initial labeled instances $\{\langle x, y \rangle^{(l)}\}_{l=1}^L$
- 3: **for** $t = 1, 2, \dots$ **do**
- 4: $\theta = \text{train}(\mathcal{L})$
- 5: select $x^* \in \mathcal{U}$, the most uncertain instance according to model θ
- 6: query the oracle to obtain label y^*
- 7: add $\langle x^*, y^* \rangle$ to \mathcal{L}
- 8: remove x^* from \mathcal{U}
- 9: **end for**

- 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.



Mode	Annotator type	Recall	Precision	F-score
Automation				
	Entity	61.94	49.31	54.91
	Protein	57.31	50.97	53.95
Expert				
	Entity	29.11	22.90	25.63
	Protein	71.94	59.28	65.00

Yimam, S. M., Biemann, C., Majnaric, 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.

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.

5 Over the past decade , chronic inflammation in visceral adipose tissue (VAT) has gained accep
as a lead promoter of insulin resistance in obesity .

6 A great deal of evidence has pointed to the role of adipokines and innate immune cells , in parti
adipose tissue macrophages , in the regulation of fat inflammation and glucose homeostasis .

(a) Annotated by medical expert.

5 Over the past decade , chronic inflammation in visceral adipose tissue (VAT) has gained accep
as a lead promoter of insulin resistance in obesity .

6 A great deal of evidence has pointed to the role of adipokines and innate immune cells , in parti
adipose tissue macrophages , in the regulation of fat inflammation and glucose homeostasis .

(b) Automatic suggestions after 5 abstracts are annotated.

5 Over the past decade , chronic inflammation in visceral adipose tissue (VAT) has gained accep
as a lead promoter of insulin resistance in obesity .

6 A great deal of evidence has pointed to the role of adipokines and innate immune cells , in parti
adipose tissue macrophages , in the regulation of fat inflammation and glucose homeostasis .

08 Multi-Task Learning

- 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 Kirkpatrick^a, Razvan Pascanu^a, Neil Rabinowitz^a, Joel Veness^a, Guillaume Desjardins^a, Andrei A. Rusu^a, Kieran Milan^a, John Quan^a, Tiago Ramalho^a, Agnieszka Grabska-Barwinska^a, Demis Hassabis^a, Claudia Clopath^b, Dharshan Kumaran^a, and Raia Hadsell^a

^aDeepMind, London, NIC 4AG, United Kingdom

^bBioengineering 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.

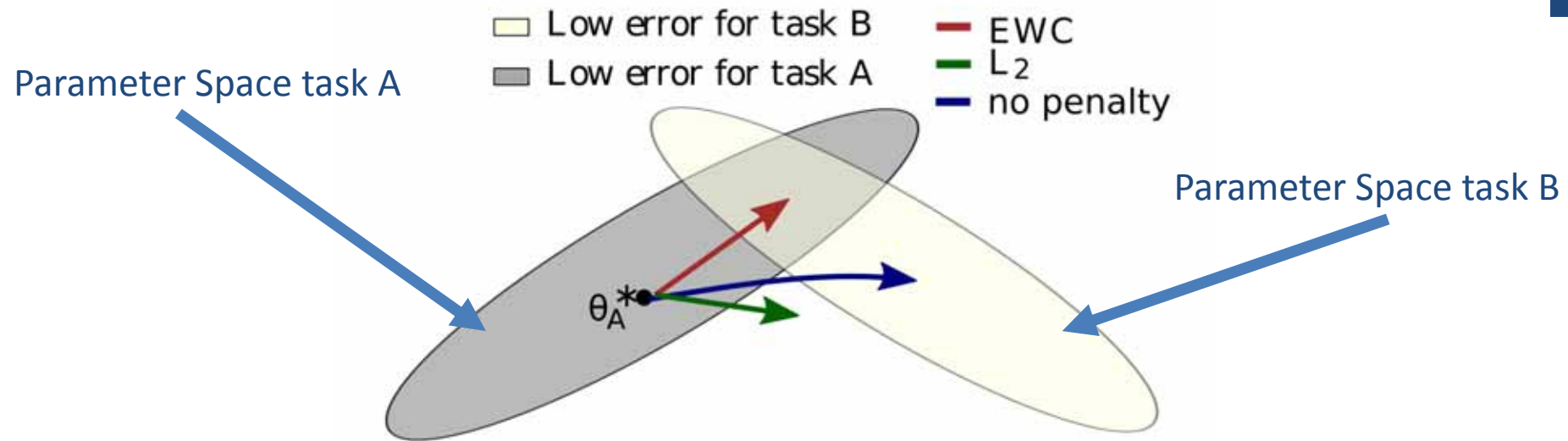
Review

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

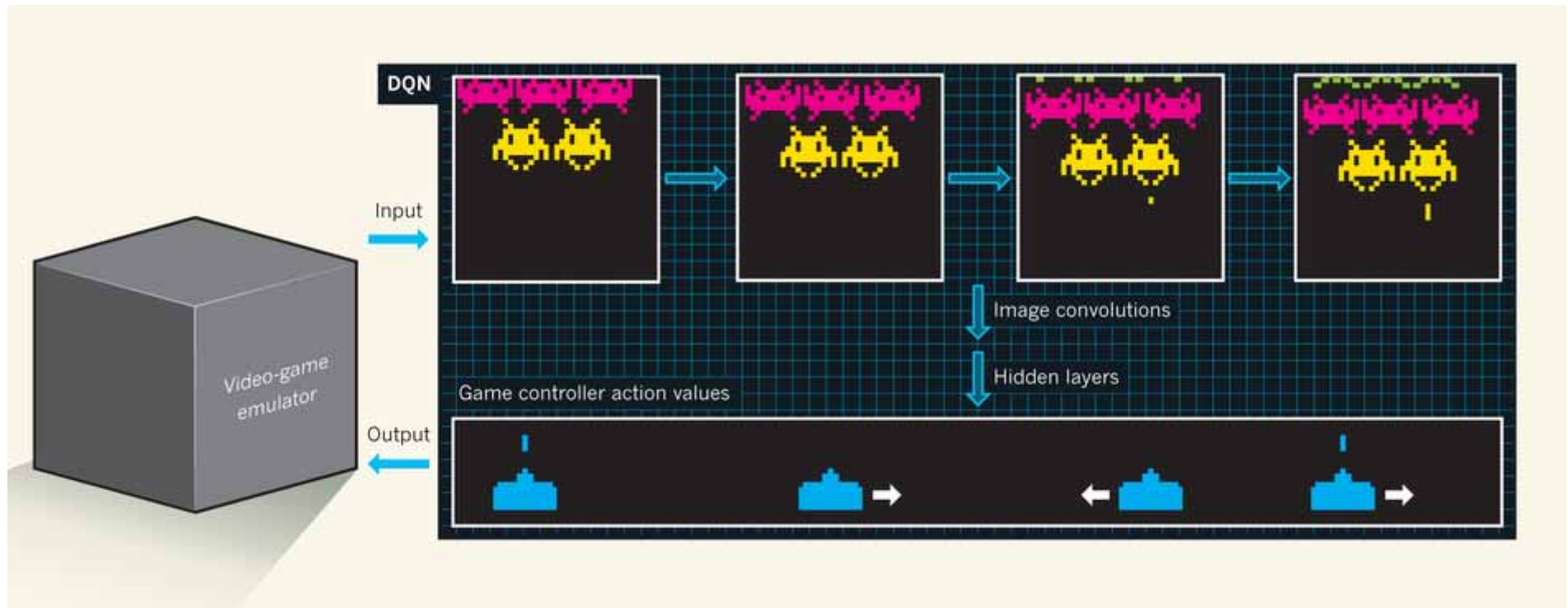


$$\log p(\theta|\mathcal{D}) = \log p(\mathcal{D}|\theta) + \log p(\theta) - \log p(\mathcal{D})$$

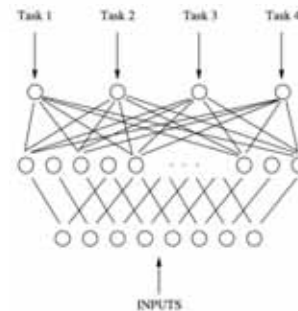
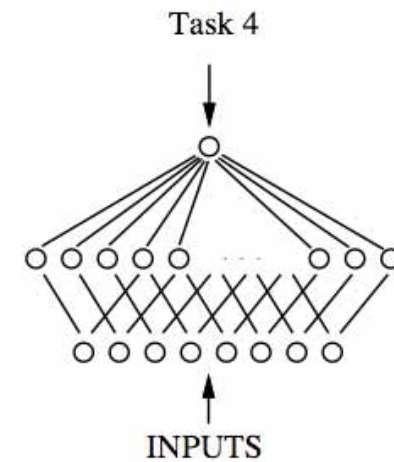
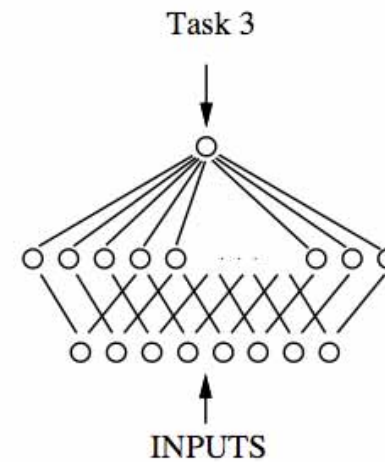
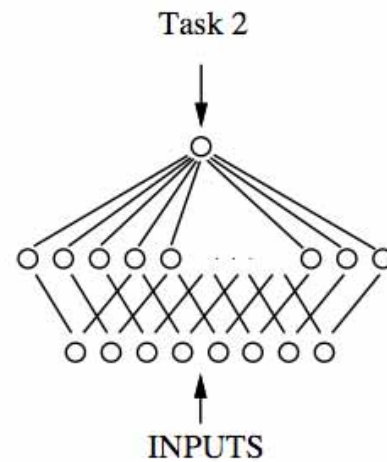
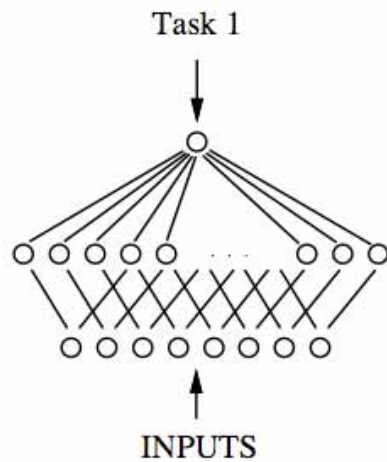
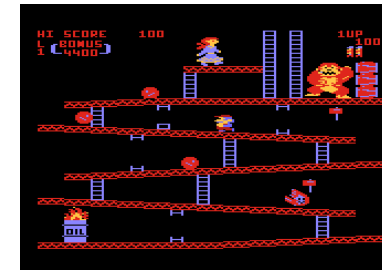
$$\log p(\theta|\mathcal{D}) = \log p(\mathcal{D}_B|\theta) + \log p(\theta|\mathcal{D}_A) - \log p(\mathcal{D}_B)$$

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

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.

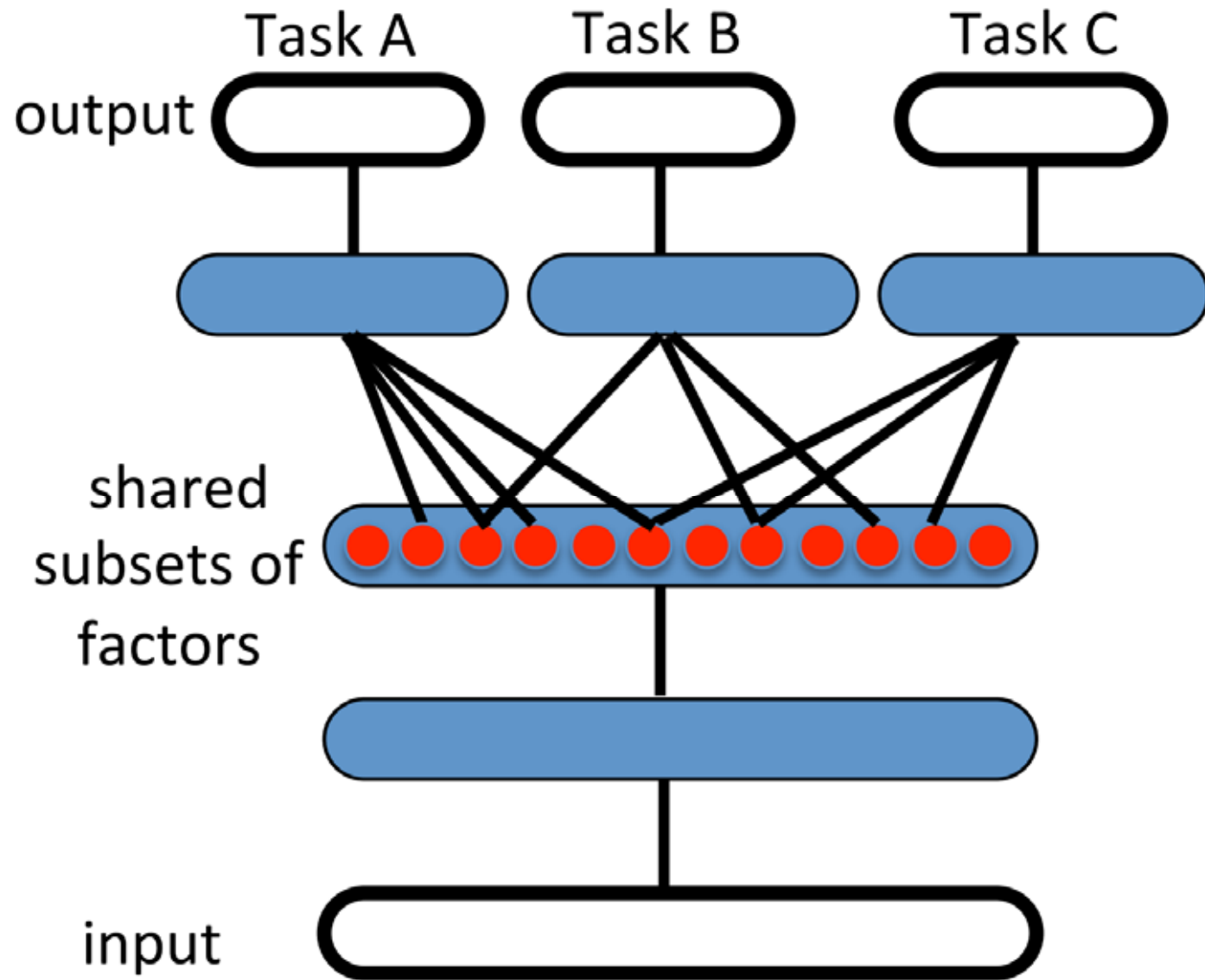


Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S. & Hassabis, D. 2015. Human-level control through deep reinforcement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

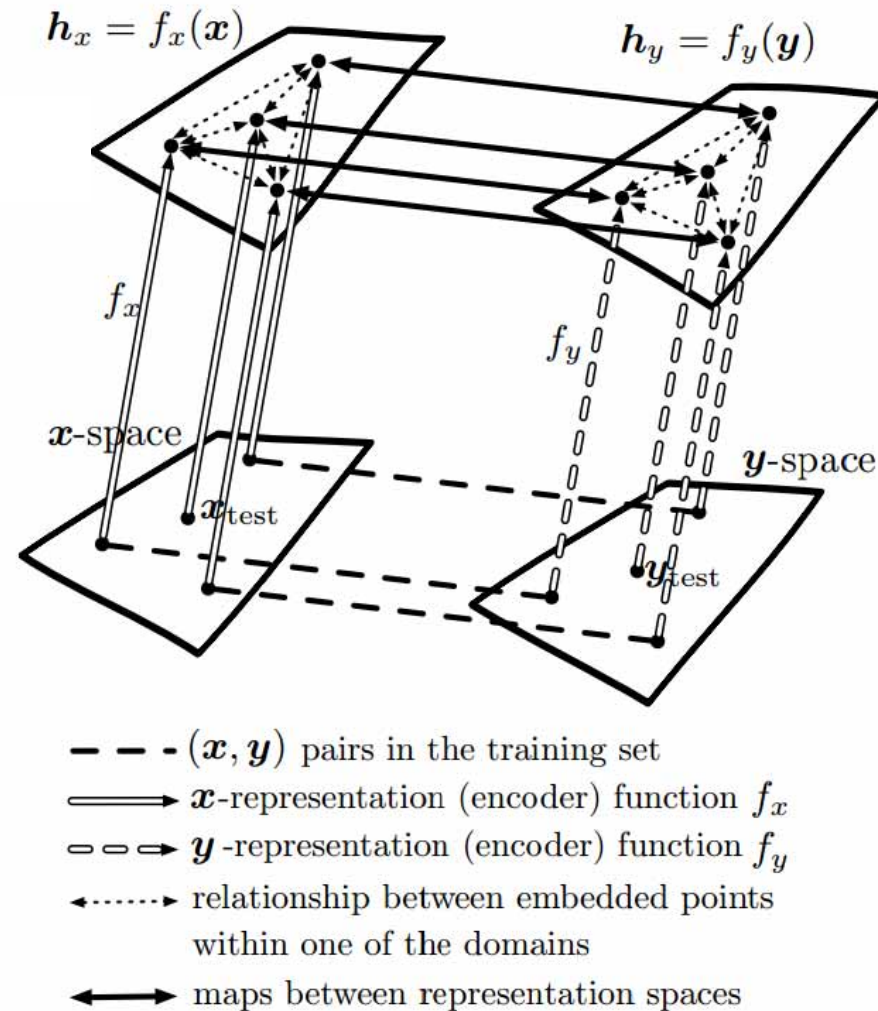


V. Mnih *et al.*, "Playing Atari with Deep Reinforcement Learning", Nature (2015)
 Rich Caruana, "Multi-task Learning", MLJ (1998)

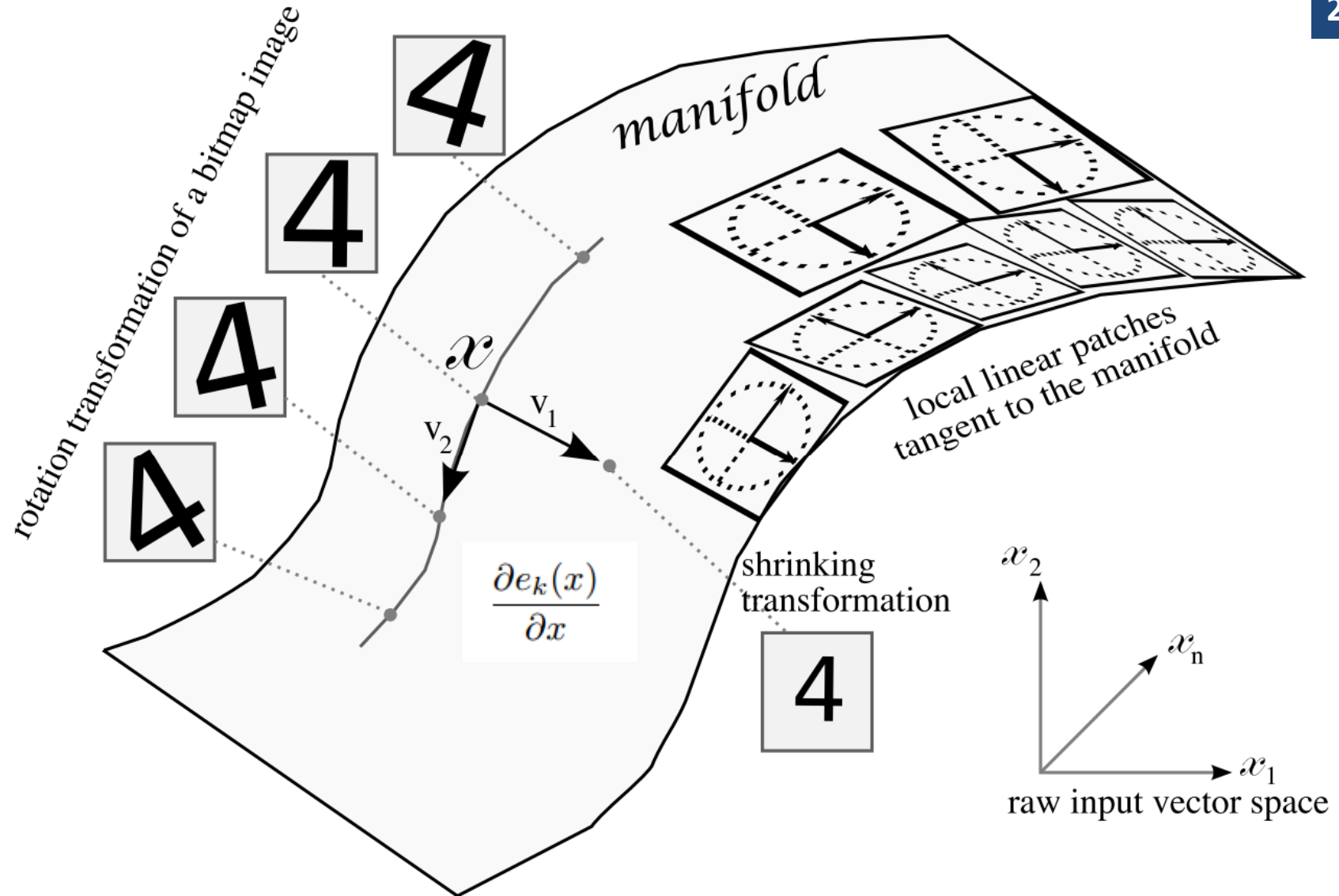
Bengio, Y., Courville, A. & Vincent, P. 2013. Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence, 35, (8), 1798-1828, doi:10.1109/TPAMI.2013.50.



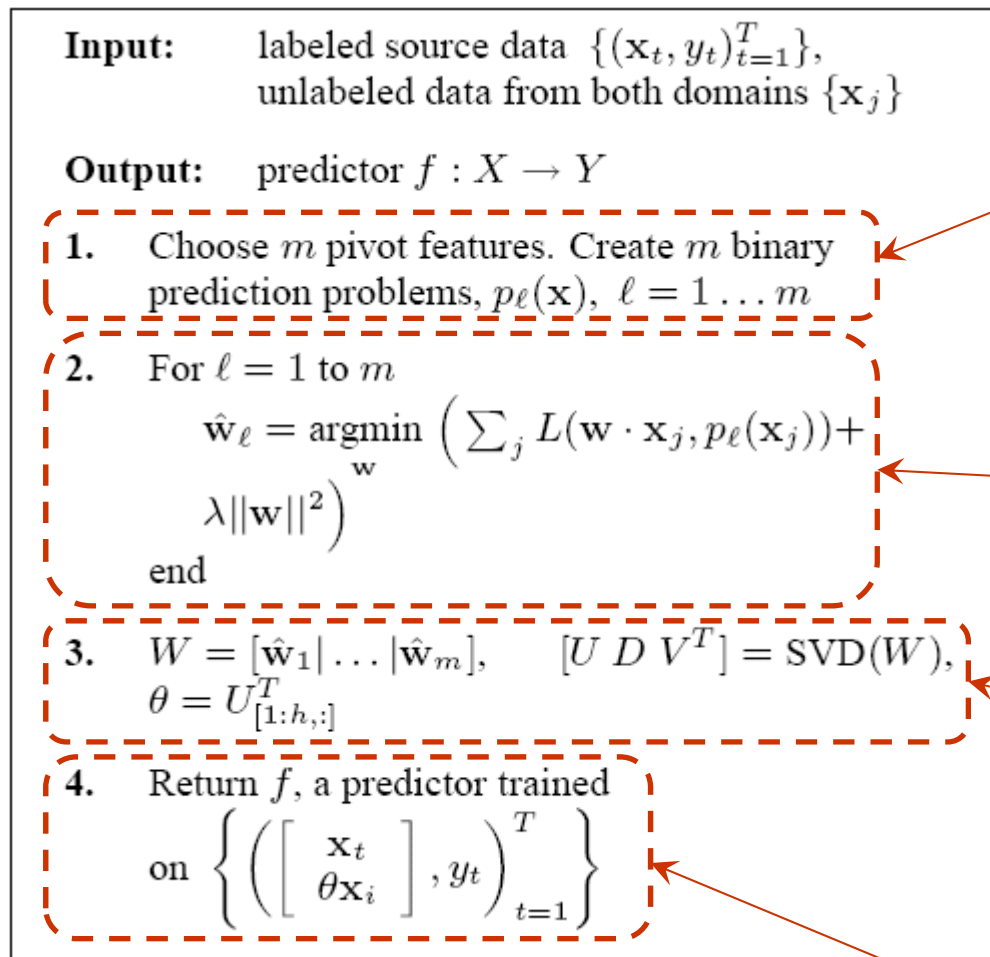
- x and y represent different modalities, e.g. text, sound, images, ...
- Generalization to new categories
- Larochelle et al. (2008) AAAI



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



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.



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.

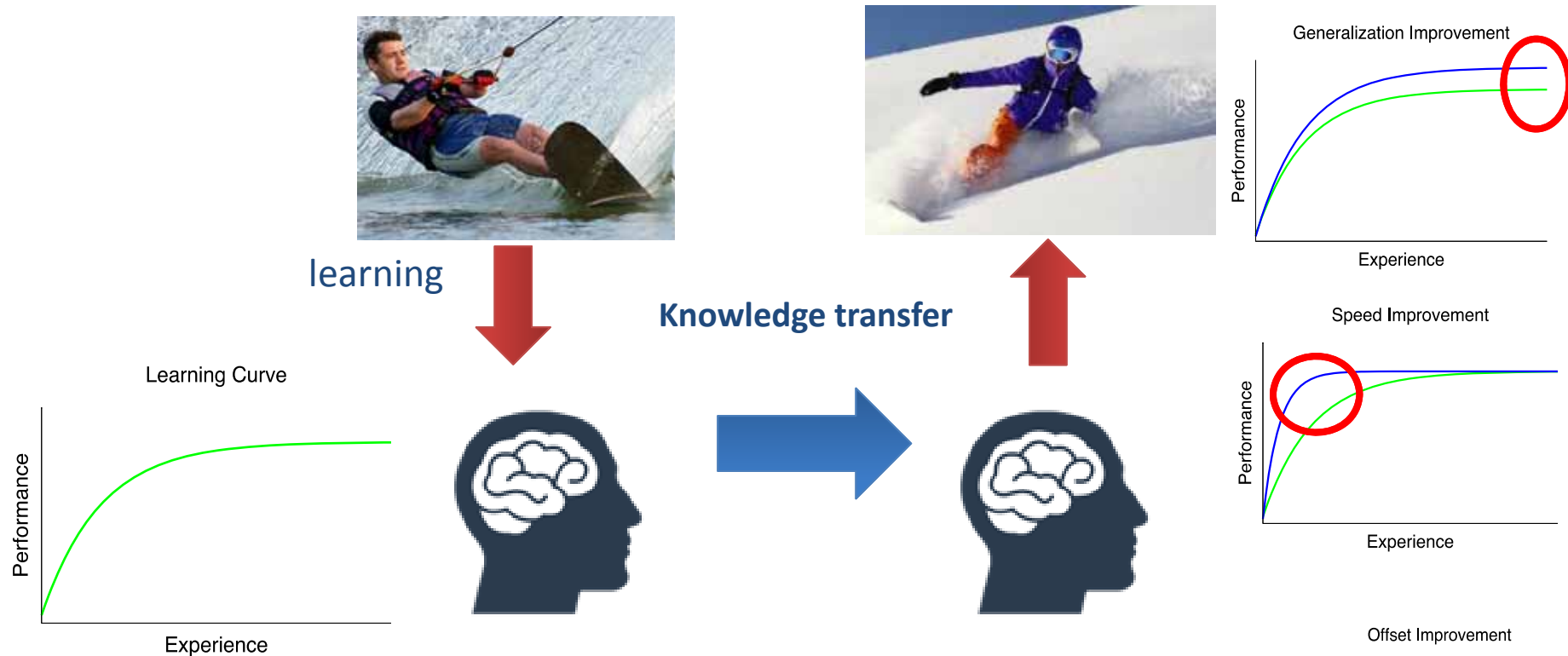
Learn the parameters of each prediction problem

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

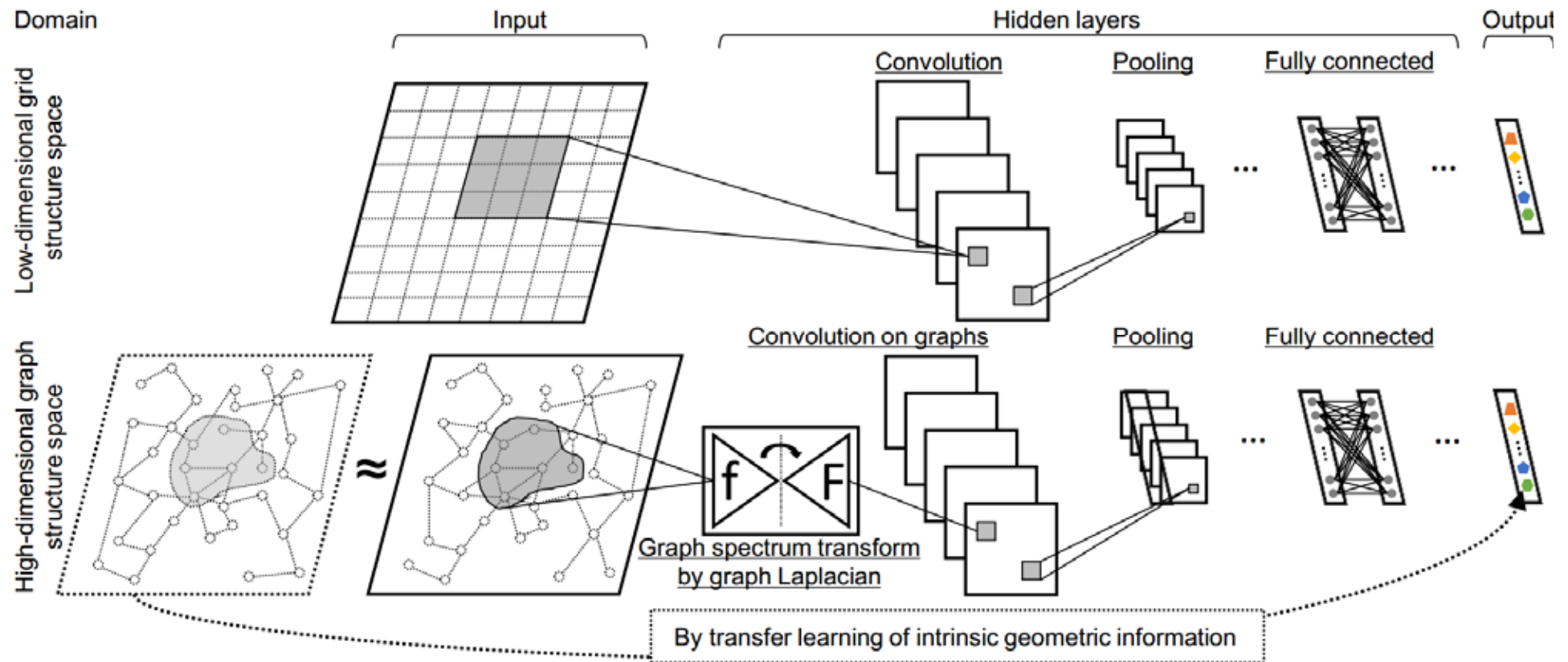
Use the learnt mapping function to construct new features and train classifiers onto the new representations.

09 Generalization & Transfer Learning

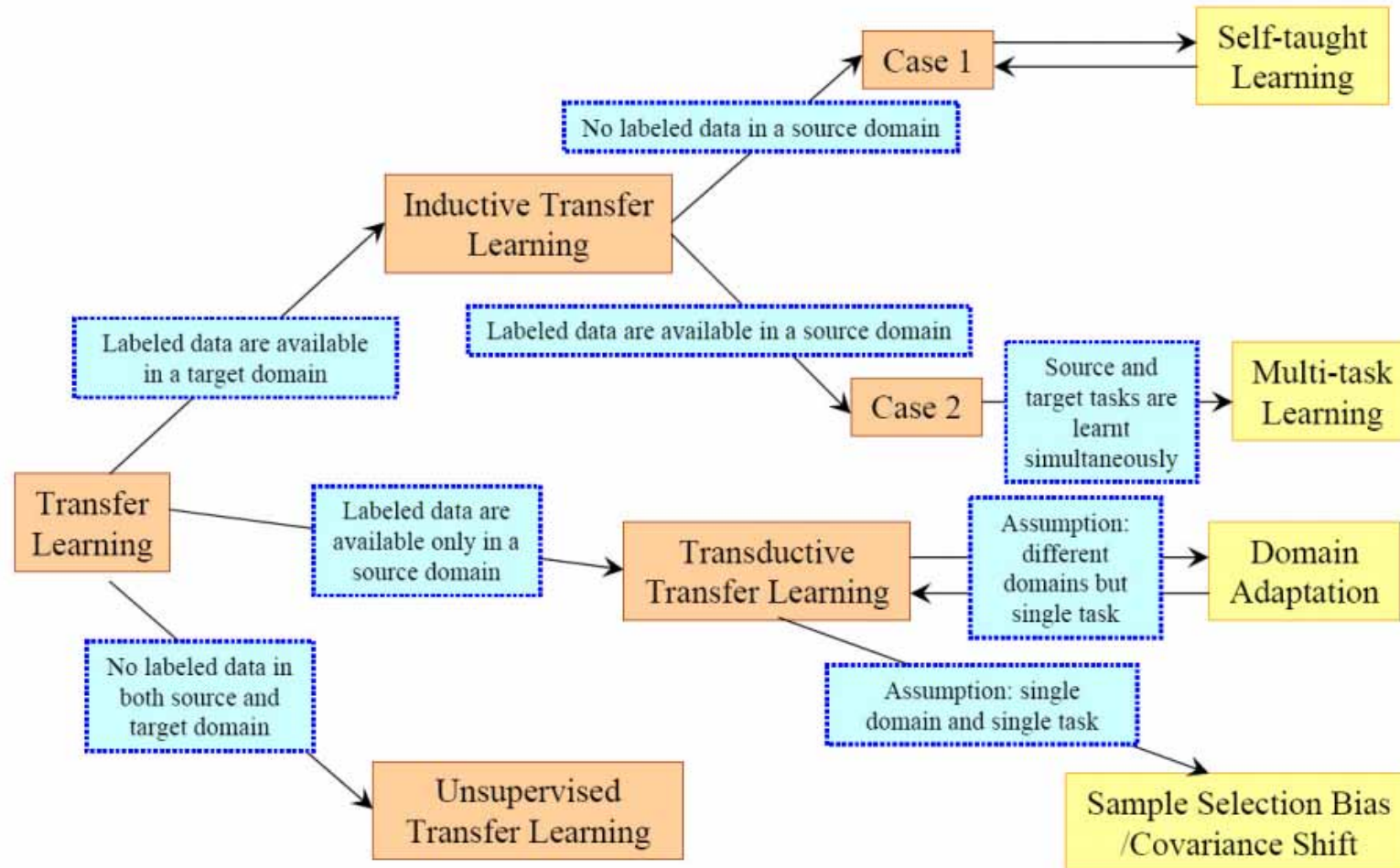
- Thorndike & Woodworth (1901) explored how individuals would transfer in one context to another context that share similar characteristics:
- or how "improvement in one mental function" could influence a related one
- Their theory implied that transfer of learning depends on how similar the learning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



- To design algorithms able to learn from experience and to **transfer knowledge across different tasks and domains** to improve their learning performance



Lee, J., Kim, H., Lee, J. & Yoon, S. 2016. Intrinsic Geometric Information Transfer Learning on Multiple Graph-Structured Datasets. arXiv:1611.04687.

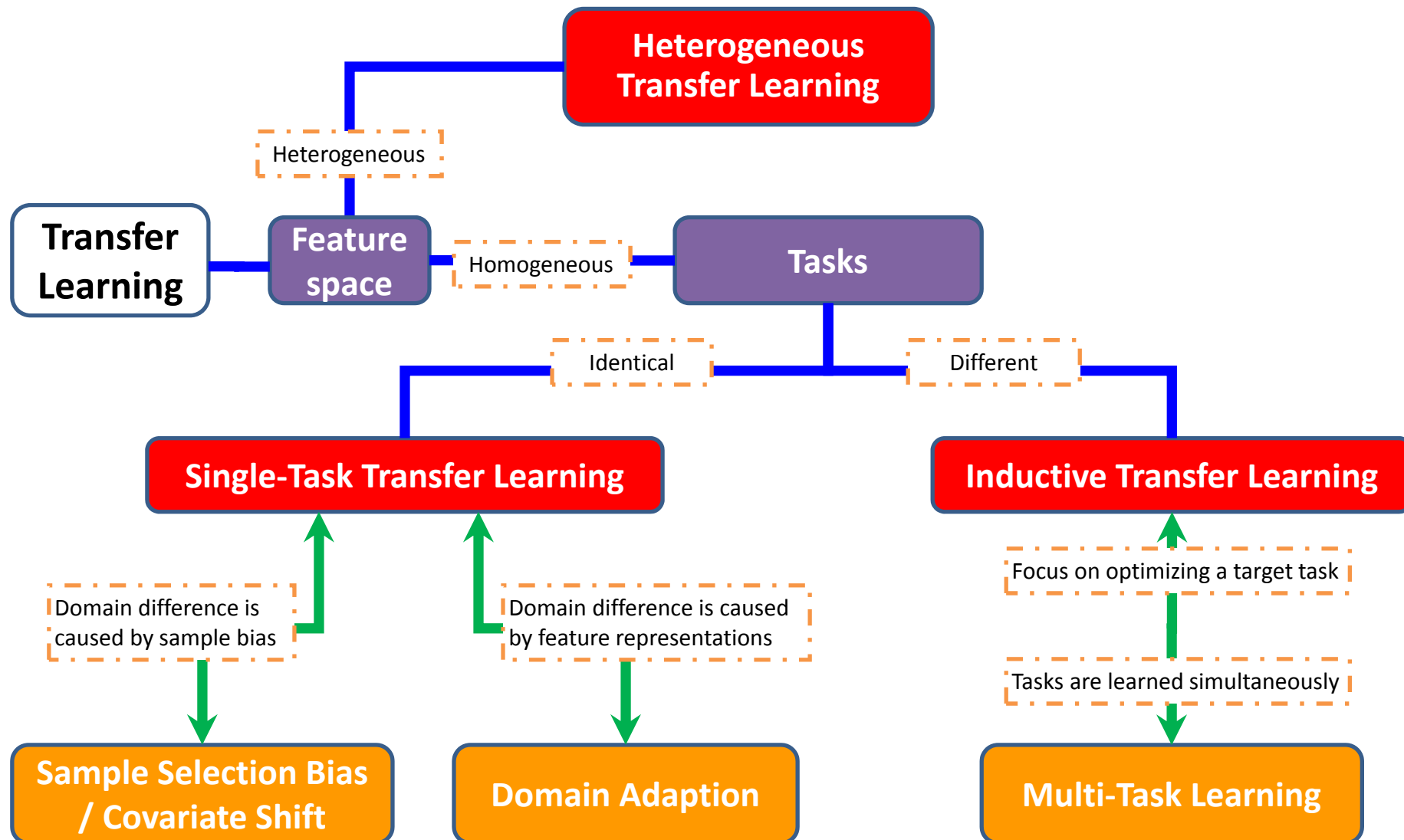


Pan, S. J. & Yang, Q. A. 2010. A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering, 22, (10), 1345-1359, doi:10.1109/tkde.2009.191.

- Feature space \mathcal{X} ;
- $P(x)$, where $x \in \mathcal{X}$.
- Given \mathcal{X} and label space \mathcal{Y} ;
- To learn $f : x \rightarrow y$, or estimate $P(y|x)$, where $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.

Two domains are different $\Rightarrow \mathcal{X}_S \neq \mathcal{X}_T$, or $P_S(x) \neq P_T(x)$.
Two tasks are different $\Rightarrow \mathcal{Y}_S \neq \mathcal{Y}_T$, or $f_S \neq f_T$ ($P_S(y|x) \neq P_T(y|x)$).

Pan, S. J. & Yang, Q. A. 2010. A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering, 22, (10), 1345-1359, doi:10.1109/tkde.2009.191.



Conclusion and Future Outlook

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.

- 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.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



Unterstützt von / Supported by



Alexander von Humboldt
Stiftung / Foundation

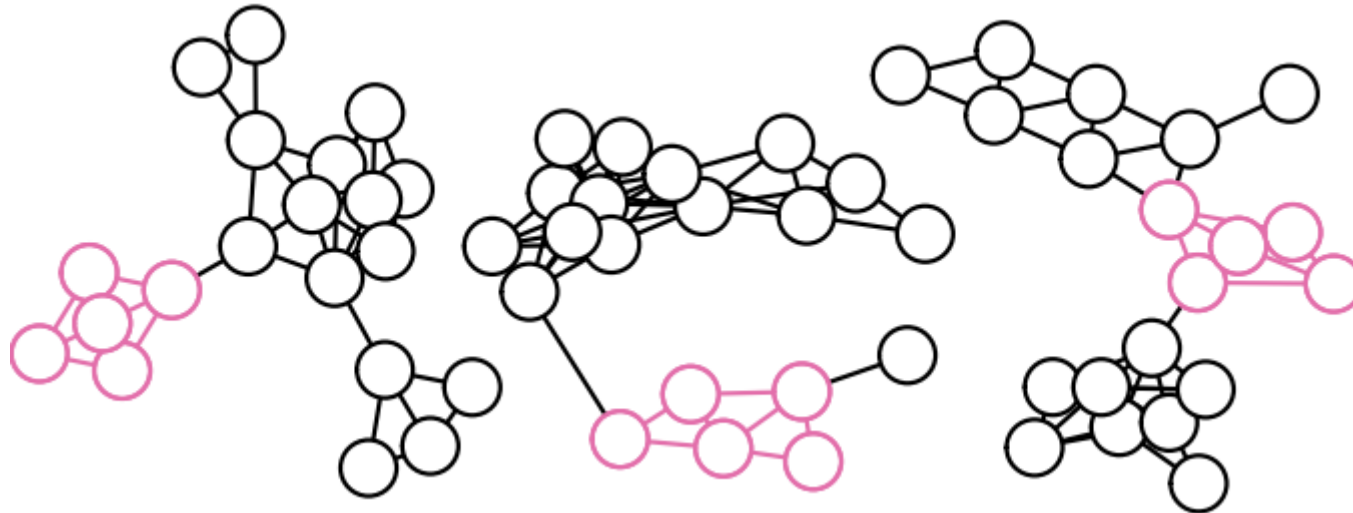
Multi-Task Feature Selection on Multiple Networks via Maximum Flows

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



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

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

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

- 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.

- 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 $S \subset V$ 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.

$$\bullet \operatorname{argmax}_{S \subset V} f(S) - g(S)$$

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

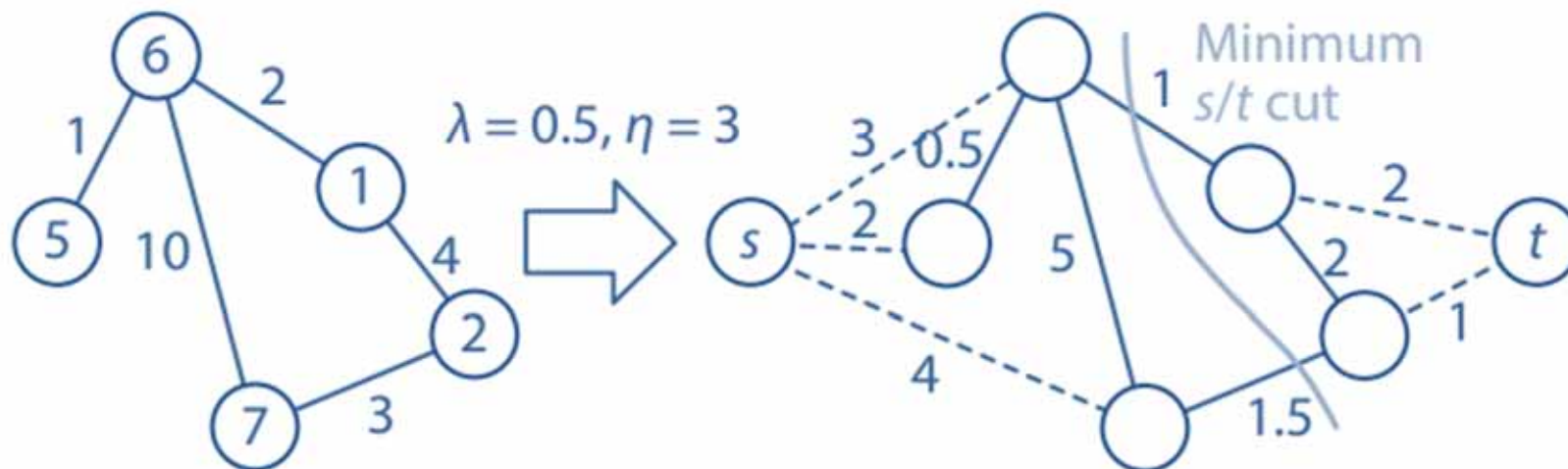
- $B = \{ \{v, u\} \in E \mid v \in V \setminus S, u \in S \}$ (boundary)
- $w : E \rightarrow \mathbb{R}^+$ is a weighting function



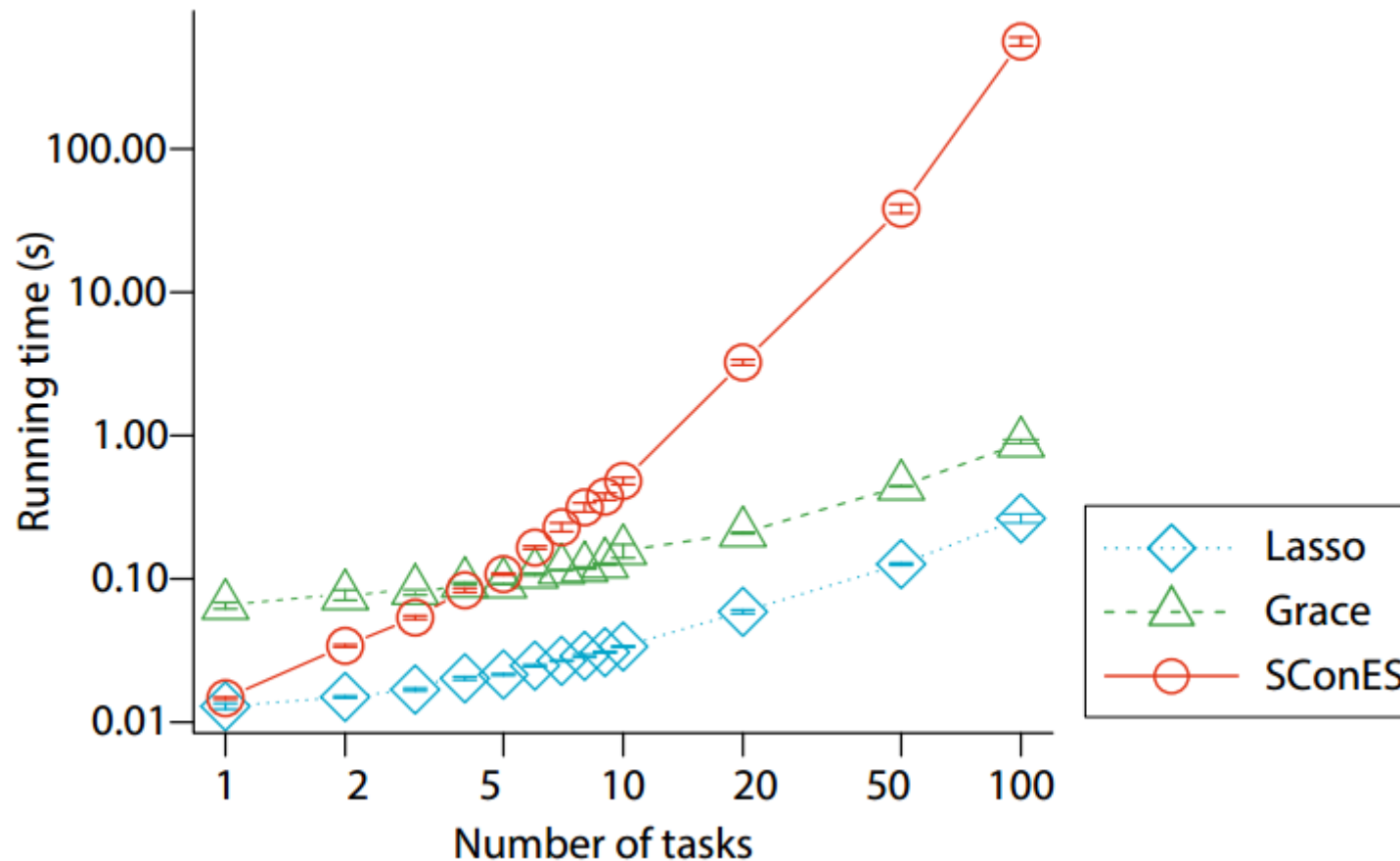
Azencott, C.-A., Grimm, D., Sugiyama, M., Kawahara, Y. & Borgwardt, K. M. 2013. Efficient network-guided multi-locus association mapping with graph cuts. *Bioinformatics*, 29, (13), i171-i179.

- 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.

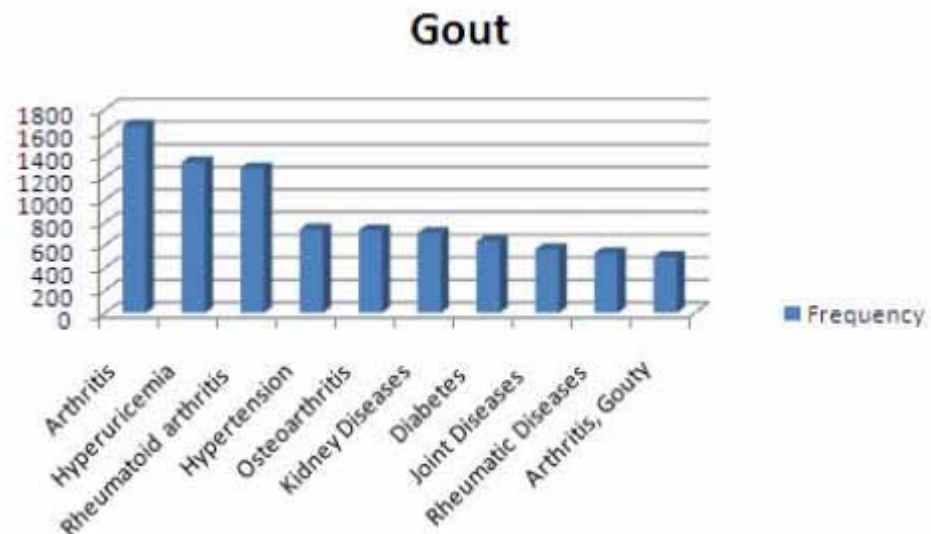
Let two words, w_i and w_j , have probabilities $P(w_i)$ and $P(w_j)$.
Then their mutual information $PMI(w_i, w_j)$ is defined as:

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

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., Simonic, 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 2012 Izmir. IEEE, 573-580, doi:10.1109/COMPSAC.2012.77.

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

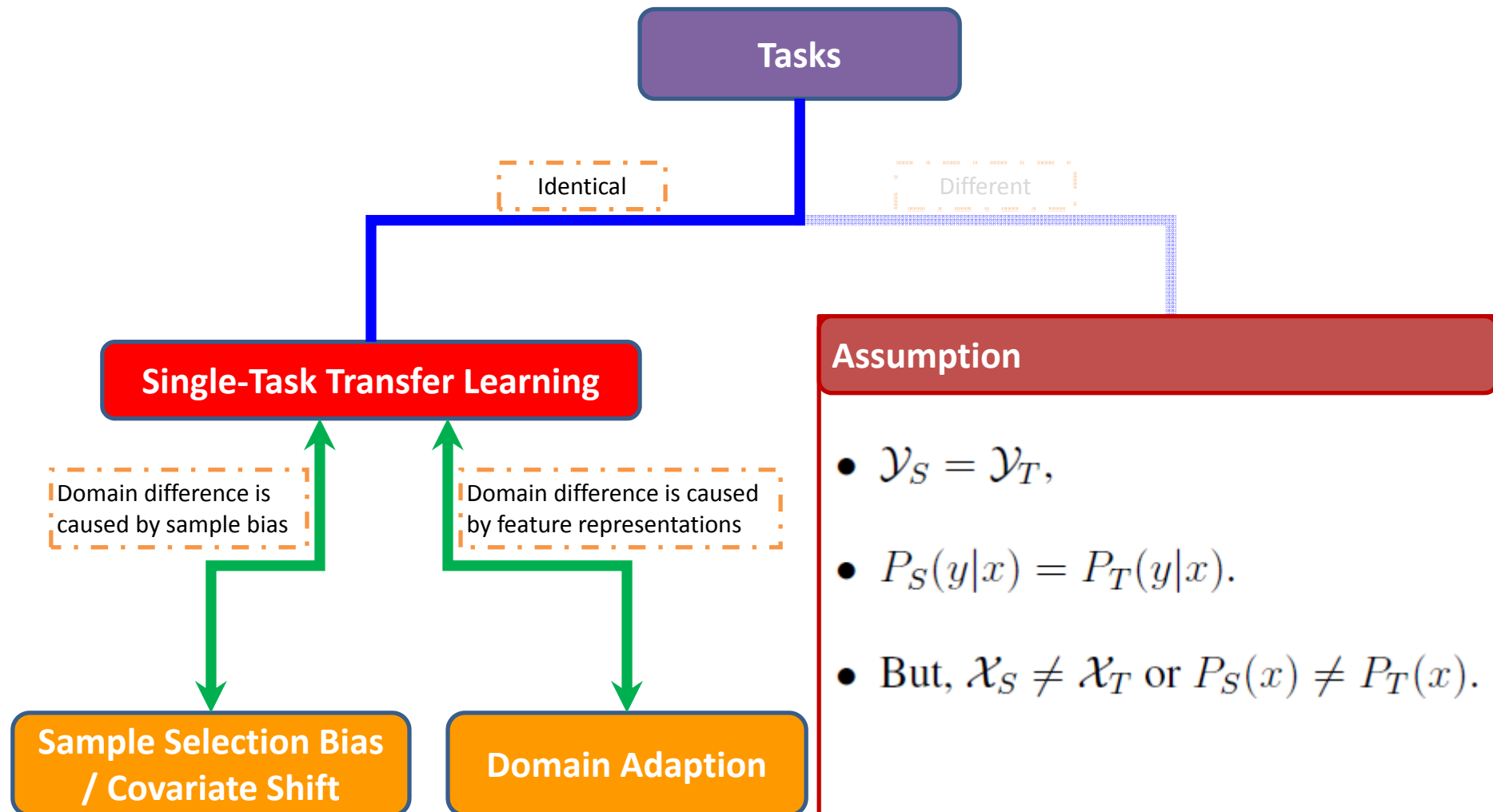
$$SCP(x, y) = p(x|y) \cdot p(y|x) = \frac{p(x, y)}{p(y)} \cdot \frac{p(x, y)}{p(x)} = \frac{p(x, y)^2}{p(x) \cdot p(y)}$$

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	5,5	hip pain	0.000
weakness	210	feeling of malaise	5,5	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.

- 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 from 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?