Andreas Holzinger

340.300 Principles of Interaction
Summer Term 2017

Selected Topics of interactive Machine Learning (iML): Interaction with Agents Part 1: Top-Level Overview

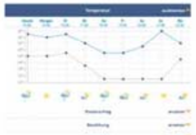
a.holzinger@hci-kdd.org
http://hci-kdd.org/interactive-machine-learning

Overview

- 01 The HCI-KDD approach: integrative ML
- 02 Understanding Intelligence
- 03 Example for Complexity
- 04 Probabilistic information
- 05 Automatic Machine Learning (aML)
- 06 Interactive Machine Learning (iML)
- 07 Active Representation Learning
- 08 Multi-Task Learning
- 09 Generalization and Transfer Learning
- 10 Federated Learning

Warm-up Quiz: Which tasks are based on ML?

1



2

3

4

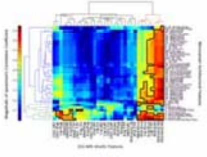
5

6

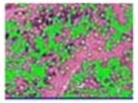
7

8

9



10



11

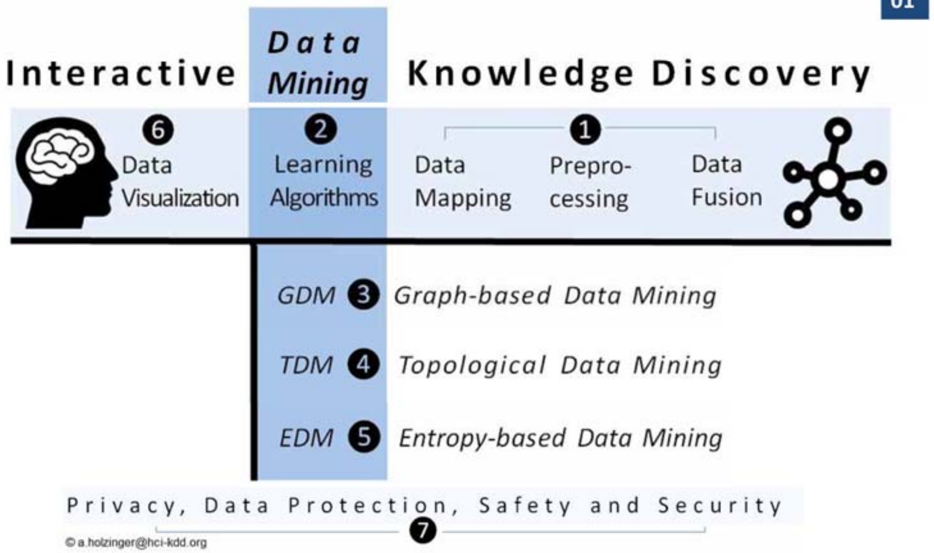
12

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 …

Machine Learning and Knowledge Extraction Pipeline

01



Interactive Data Mining Knowledge Discovery

❻ Data Visualization

❷ Learning Algorithms

❶ Data Mapping, Preprocessing, Data Fusion

GDM ❸ Graph-based Data Mining

TDM ❹ Topological Data Mining

EDM ❺ Entropy-based Data Mining

Privacy, Data Protection, Safety and Security ❼

© a.holzinger@hci-kdd.org

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

… successful ML needs …

http://www.bach-cantatas.com

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

Cognitive Science AND Computer Science

02

$\mathbb{R}^2$
$\mathbb{R}^n$

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

02 Solve Intelligence then solve everything else

Grand Goal: Understanding Intelligence

03

“Solve intelligence – then solve everything else”

THE ROYAL SOCIETY

Our Mission

1 · Solve intelligence

2 · Use it to solve everything else

Future directions of machine learning: Part 2

The Royal Society

10,797 views

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

Demis Hassabis, 22 May 2015

The Royal Society,
Future Directions of Machine Learning Part 2

Google DeepMind

To reach a level of usable intelligence we need to …

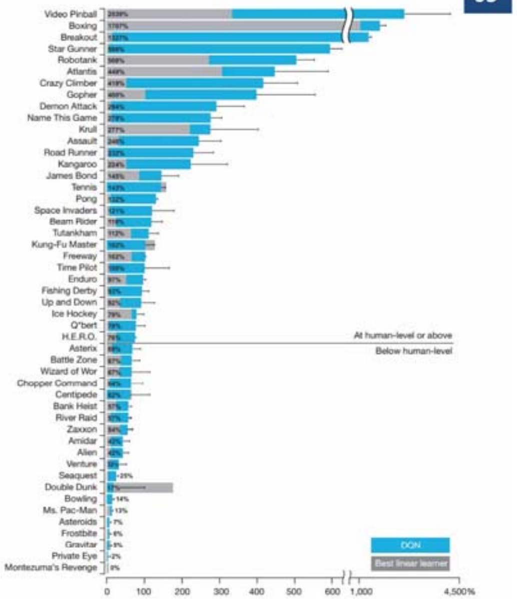
04

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

How far are we already ?

05

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



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

Scientists who pleaded for "humanoid AI"

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

Not our Goal: Humanoid AI

Humanoid AI

≠

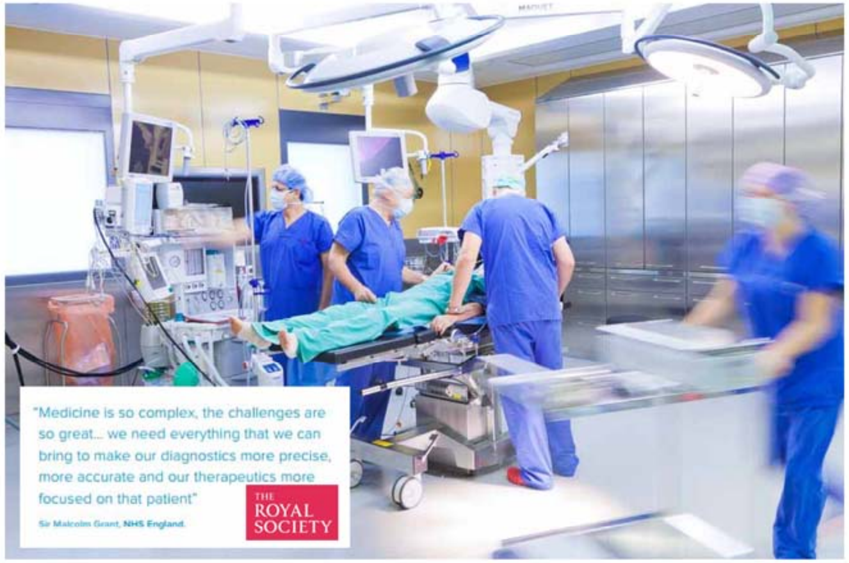
Human-level AI

03 Application Area Health Informatics

Health is a complex area

Why is this application area complex ?

Machine Learning and Health Informatics!



"Medicine is so complex, the challenges are so great... we need everything that we can bring to make our diagnostics more precise, more accurate and our therapeutics more focused on that patient"

Sir Malcolm Grant, NHS England

THE ROYAL SOCIETY

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

In medicine we have two different worlds …

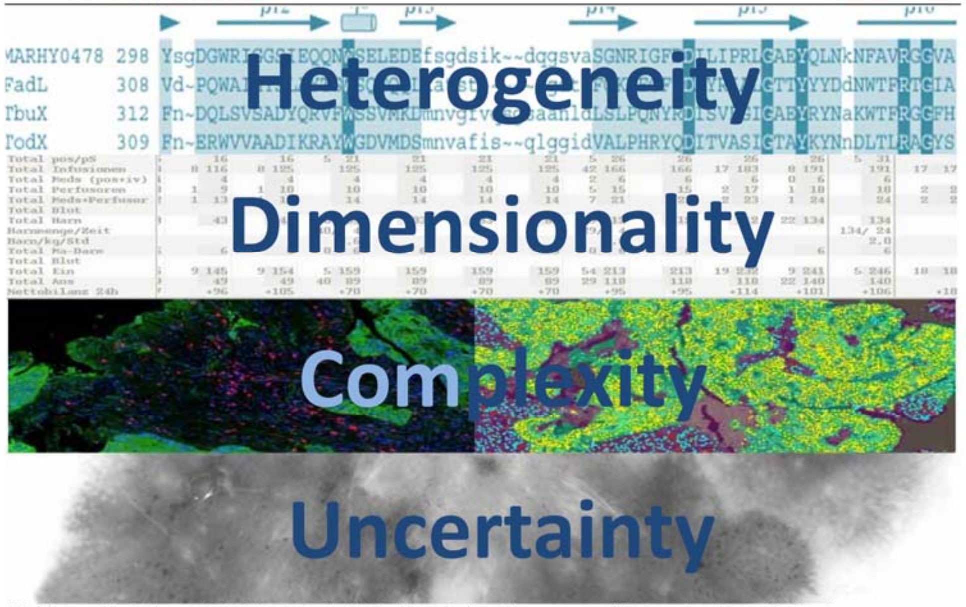
Our central hypothesis: Information may bridge this gap

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

The bridge …

Where is the problem in building this bridge?

Main problems …



Heterogeneity

Dimensionality

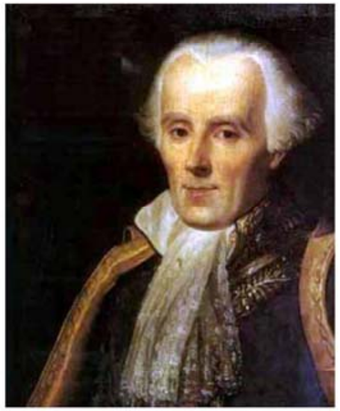
Complexity

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

Repetition of Bayes - on the work of Laplace

06

What is the simplest mathematical operation for us?

$$p(x) = \sum_y (p(x,y)) \tag{1}$$

How do we call repeated adding?

$$p(x,y) = p(y|x) * p(y) \tag{2}$$

Laplace (1773) showed that we can write:

$$p(x,y) * p(y) = p(y|x) * p(x) \tag{3}$$

Now we introduce a third, more complicated operation:

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

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

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

The foundation for machine learning was laid in 1763

07

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.

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

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

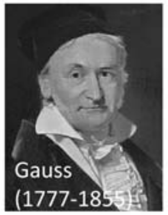
$$p(x_i|y_j) = \frac{p(y_j|x_i)p(x_i)}{\sum p(x_i, y_j)p(x_i)}$$

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.

Cedalion standing on the shoulders of Orion

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

Analogies



- **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**

Learning and Inference

04

$\boldsymbol{d}$ … $\boldsymbol{data}$

$\boldsymbol{h}$ … $\boldsymbol{hypotheses}$

$\mathcal{H}$ … $\{H_1, H_2, \ldots, H_n\}$ $\quad \forall\, h, d$ …

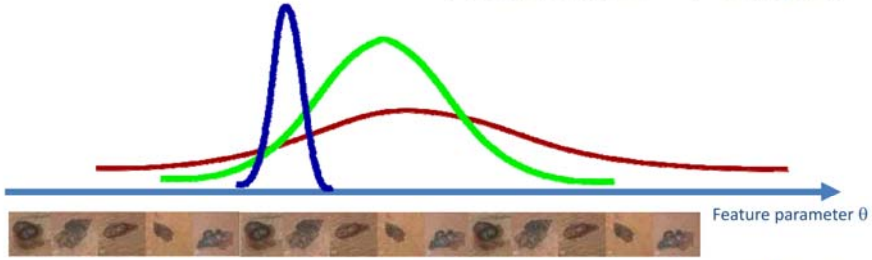
Likelihood

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



Bayesian Learning from data → Generalize

09

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

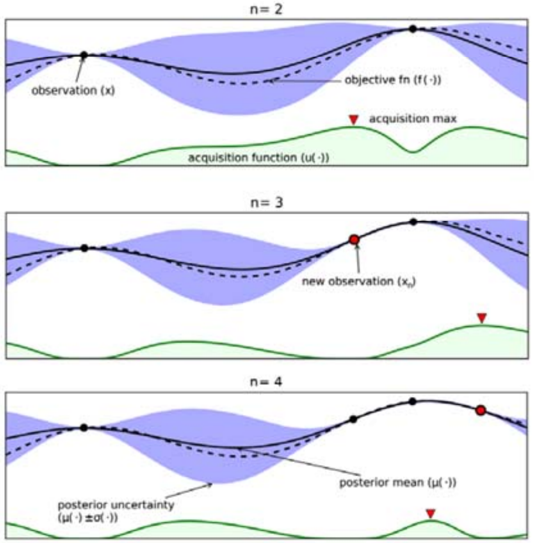
$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta) * p(\theta)}{p(\mathcal{D})}$$

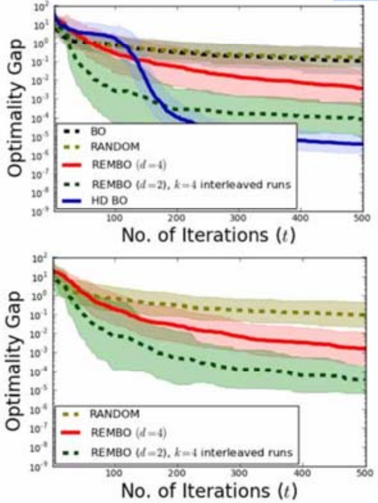
$$posterior = \frac{likelihood * prior}{evidence}$$

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

Scaling to high-dimensions is the holy grail in ML

10



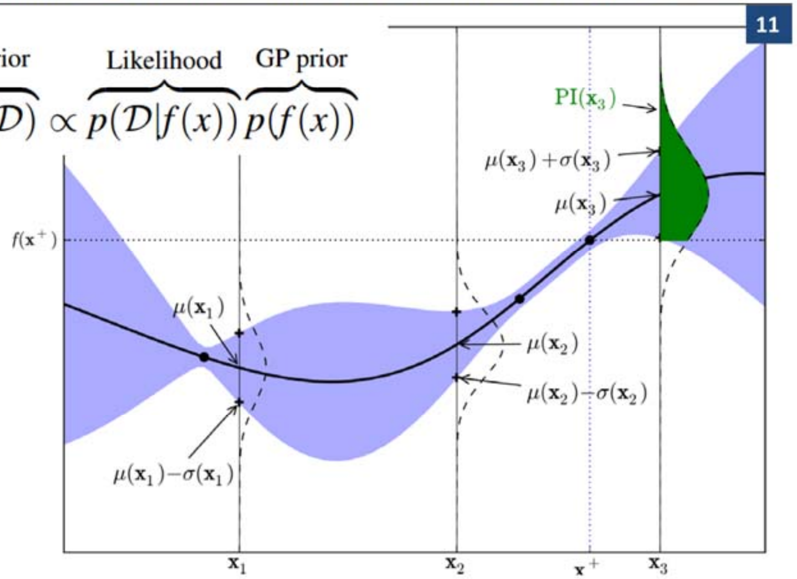


Wang, Z., Hutter, F., Zoghi, M., Matheson, D. & De Feitas, N. 2016. Bayesian optimization in a billion dimensions via random embeddings. Journal of Artificial Intelligence Research, 55, 361-387, doi:10.1613/jair.4806.

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

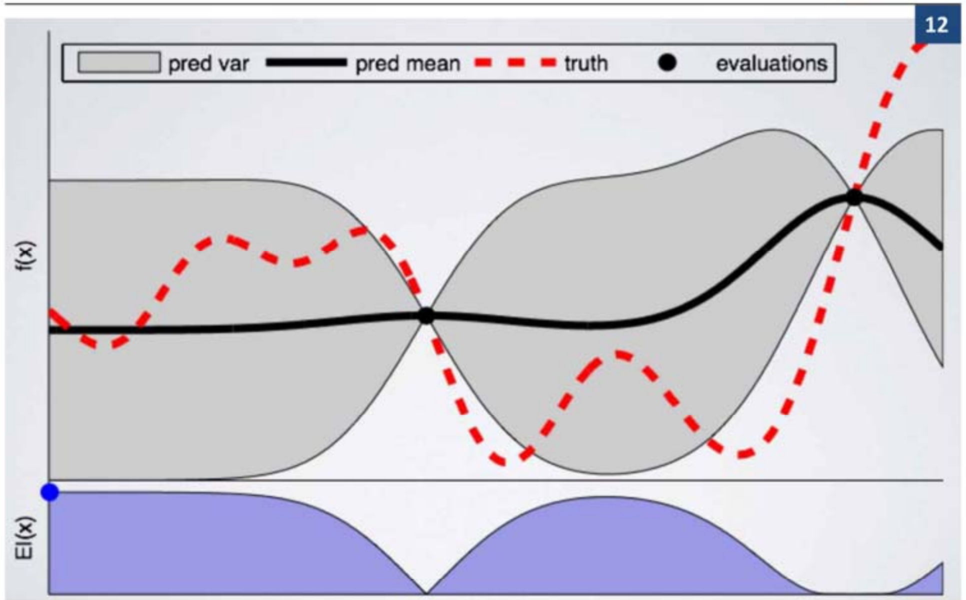
11

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



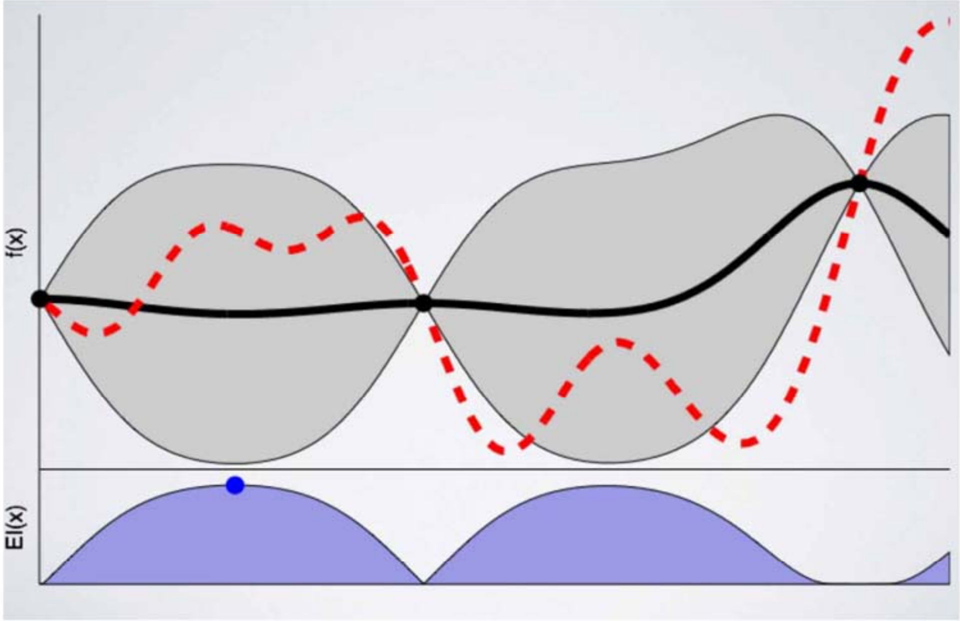
Brochu, E., Cora, V. M. & De Freitas, N. 2010. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. arXiv:1012.2599.

12

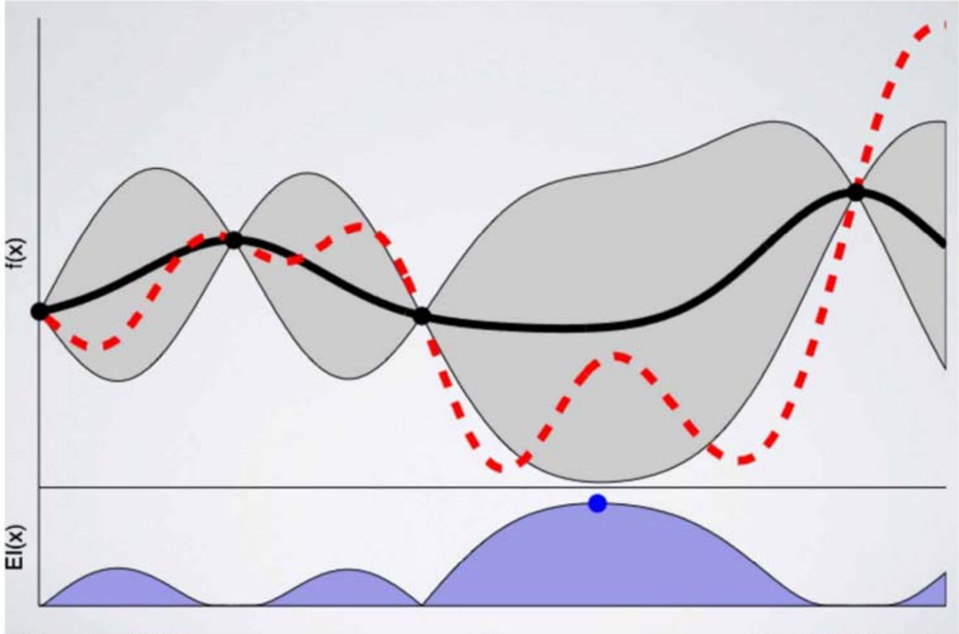


Snoek, J., Larochelle, H. & Adams, R. P. Practical Bayesian optimization of machine learning algorithms. Advances in neural information processing systems, 2012. 2951-2959.

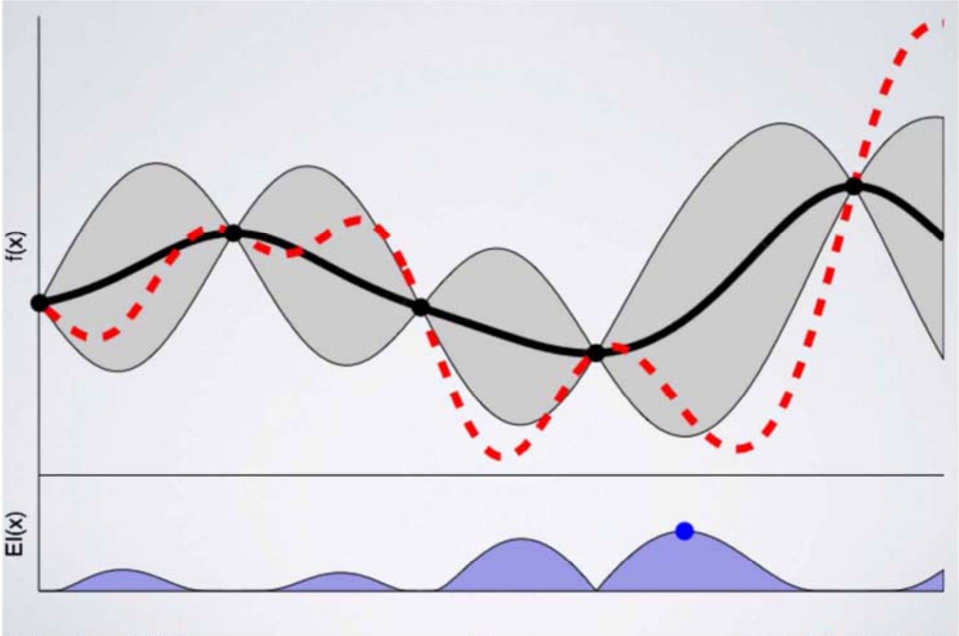
Bayesian Optimization 1



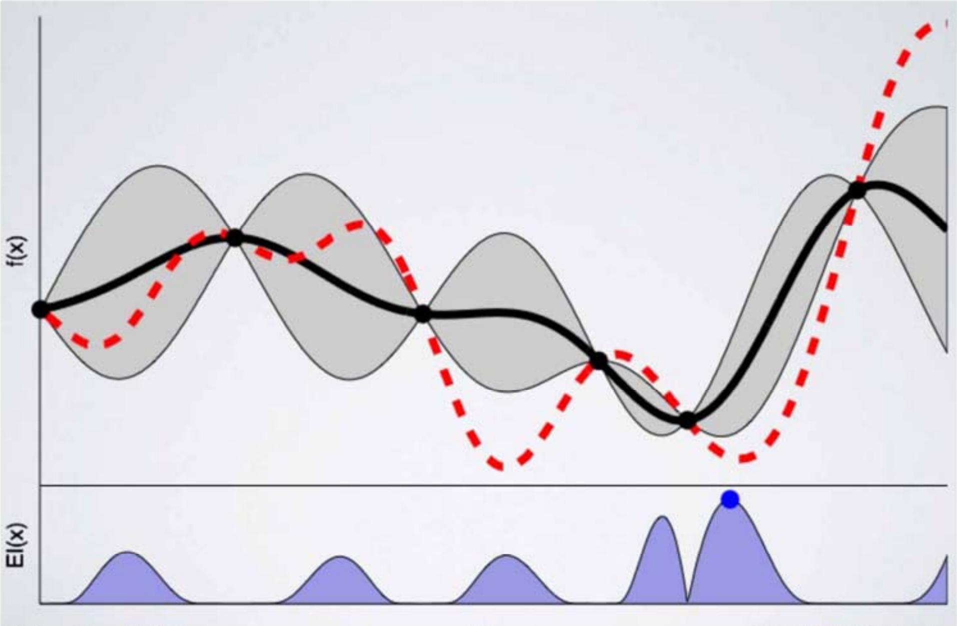
Bayesian Optimization 2



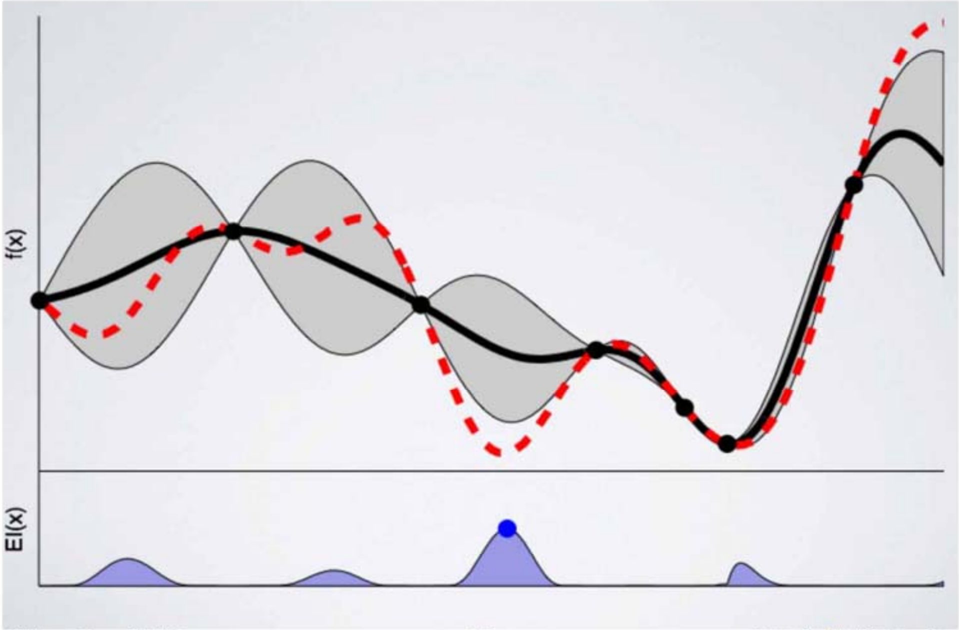
Bayesian Optimization 3



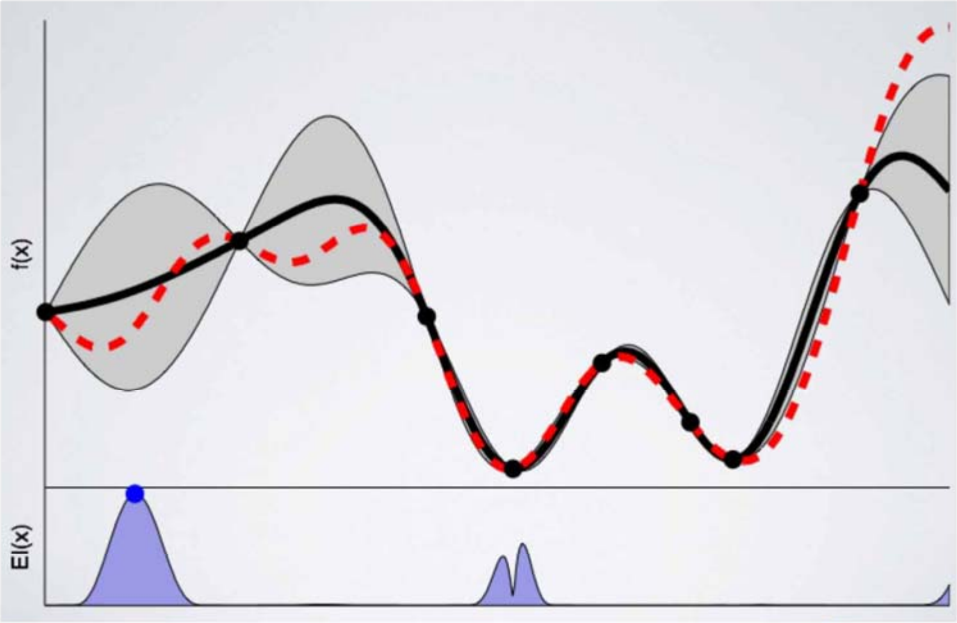
Bayesian Optimization 4



Bayesian Optimization 5

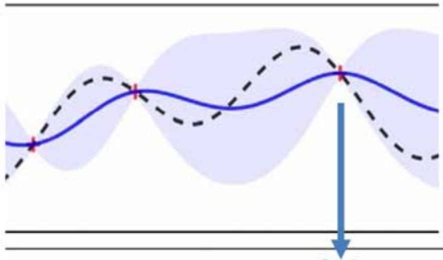


Bayesian Optimization 6

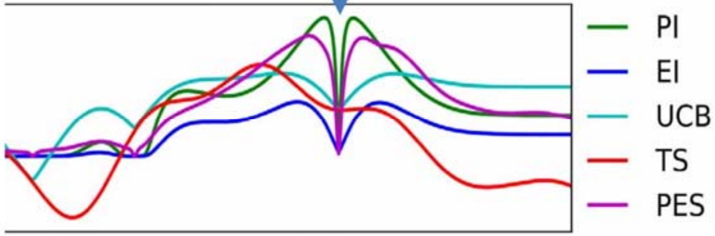


Fully automatic → Goal: Taking the human out of the loop

13



```
Algorithm 1 Bayesian optimization
for $n = 1, 2, \dots$ do
    select new $\mathbf{x}_{n+1}$ by optimizing acquisition function $\alpha$
        $\mathbf{x}_{n+1} = \arg\max_{\mathbf{x}} \alpha(\mathbf{x}; \mathcal{D}_n)$
    query objective function to obtain $y_{n+1}$
    augment data $\mathcal{D}_{n+1} = \{\mathcal{D}_n, (\mathbf{x}_{n+1}, y_{n+1})\}$
    update statistical model
end for
```



Probability of Improvement

Expected Improvement

Upper Confidence Bound

Thompson Sampling

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

Everything is machine learning …

- Today most ML-applications are using automatic Machine Learning (aML) approaches

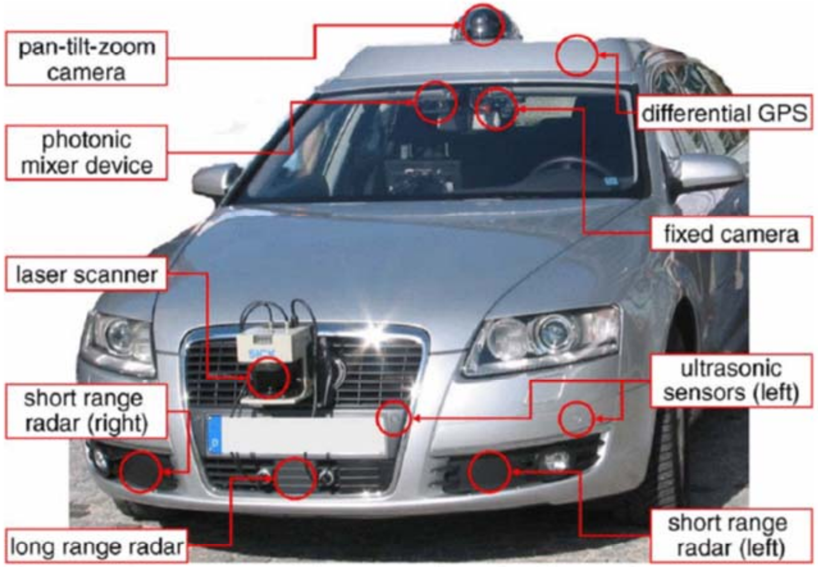
- **aML := algorithms which interact with agents and can optimize their learning behaviour trough this interaction**

Best practice examples of aML …

Fully automatic autonomous vehicles (“Google car”)

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

Good example for Learning from big data



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.

JKU ... an old dream to make it automatic

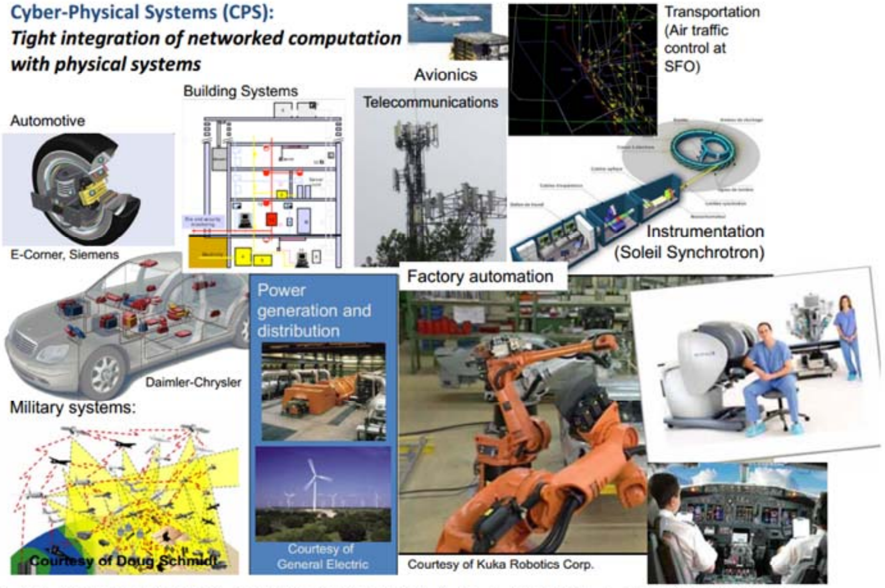
This Citroen DS with "automated steering" was tested in the early 1960s...

1960s Citroën DS driverless car test

Sunday Times Driving

8,605 views

JKU ... and thousands of industrial aML applications ...



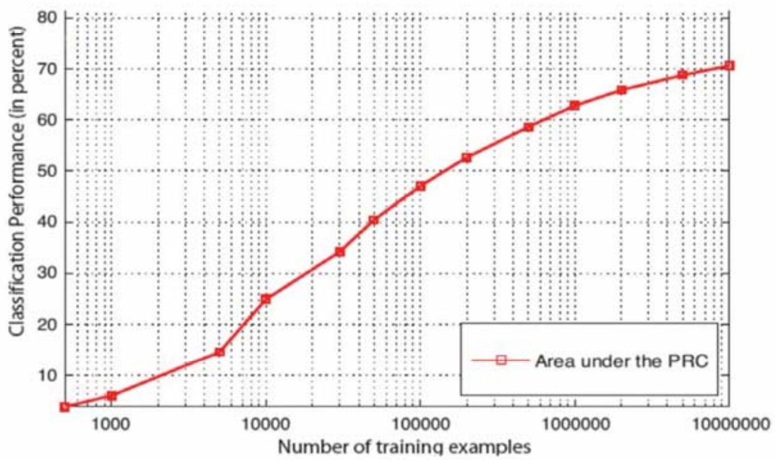
Cyber-Physical Systems (CPS):
Tight integration of networked computation with physical systems

Avionics
Transportation (Air traffic control at SFO)
Building Systems
Telecommunications
Automotive
E-Corner, Siemens
Instrumentation (Soleil Synchrotron)
Factory automation
Power generation and distribution
Daimler-Chrysler
Military systems:
Courtesy of Doug Schmidt
Courtesy of General Electric
Courtesy of Kuka Robotics Corp.

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.

JKU Big Data is necessary for aML !

14



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

JKU 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, ...

06 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.

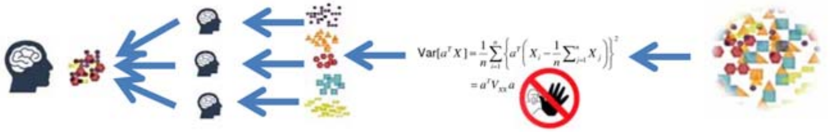
Sometimes we need a doctor-in-the-loop

A group of experts-in-the-loop

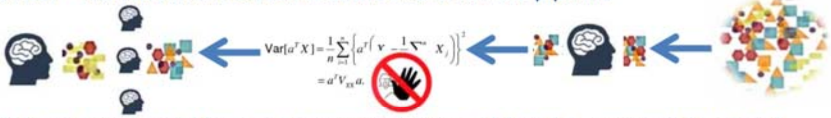
A crowd of people-in-the-loop

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

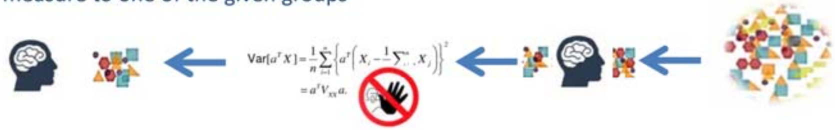
A) Unsupervised ML: Algorithm is applied on the raw data and learns fully automatic – Human can check results at the end of the ML-pipeline


$\mathrm{Var}[a^T X] = \frac{1}{n}\sum_{i=1}^{n}\left\{a^T\left(X_i - \frac{1}{n}\sum_{j=1}^{n} X_j\right)\right\}^2 = a^T V_{XX} a$

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


$\mathrm{Var}[a^T X] = \frac{1}{n}\sum_{i=1}^{n}\left\{a^T\left(X_i - \frac{1}{n}\sum_{j=1}^{n} X_j\right)\right\}^2 = a^T V_{XX} a.$

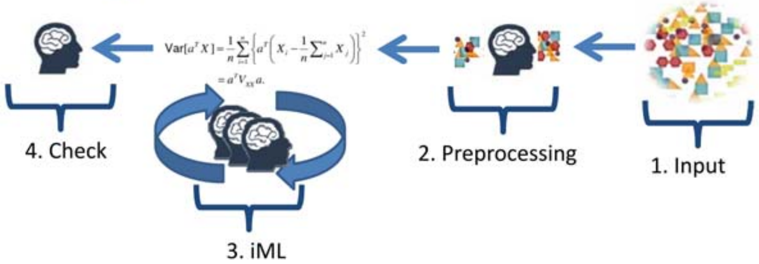
C) Semi-Supervised Machine Learning: A mixture of A and B – mixing labeled and unlabeled data so that the algorithm can find labels according to a similarity measure to one of the given groups


$\mathrm{Var}[a^T X] = \frac{1}{n}\sum_{i=1}^{n}\left\{a^T\left(X_i - \frac{1}{n}\sum_{j=1}^{n} X_j\right)\right\}^2 = a^T V_{XX} a.$

iML: bringing the human-in-the-loop

17

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 …


$\mathrm{Var}[a^T X] = \frac{1}{n}\sum_{i=1}^{n}\left\{a^T\left(X_i - \frac{1}{n}\sum_{j=1}^{n} X_j\right)\right\}^2 = a^T V_{XX} a.$
4. Check
3. iML
2. Preprocessing
1. Input

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.

Three examples for the usefulness of the iML approach

18

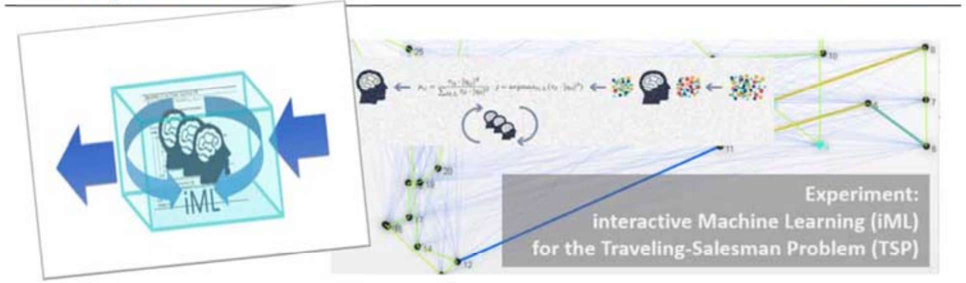
- **Example 1: Subspace Clustering**
- **Example 2: k-Anonymization**
- **Example 3: Protein Design**

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: iML



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

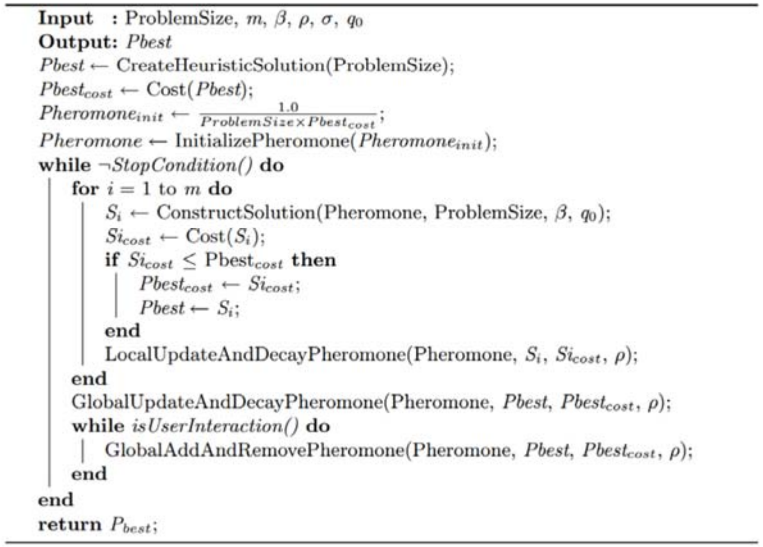
- From black-box to glass-box ML
- Exploit human intelligence for solving hard problems (e.g. Subspace Clustering, k-Anonymization, Protein-Design)
- Towards multi-agent systems with humans-in-the-loop

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

Natural Multi-Agent Systems

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

19



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

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

The grand question of cognitive science

15a

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

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

07 Active Representation Learning

The grand question of cognitive science

15b

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

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

What is this …

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

Active Learning – study of ML that improve by asking …

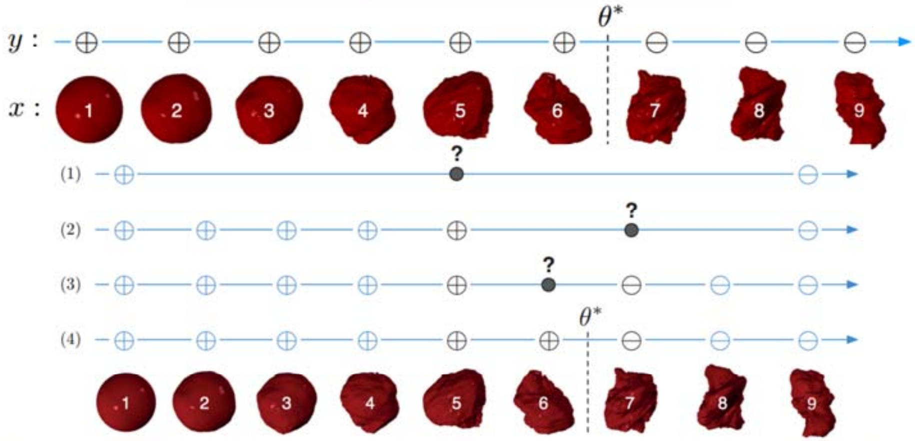
- := ML algorithm can perform better with less training if it is allowed to choose the data from which it learns.
- "Active learner" may pose queries, usually in the form of unlabeled data instances to be labeled by an "oracle" (e.g., a human annotator) that **understands** the **context** of the problem.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

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

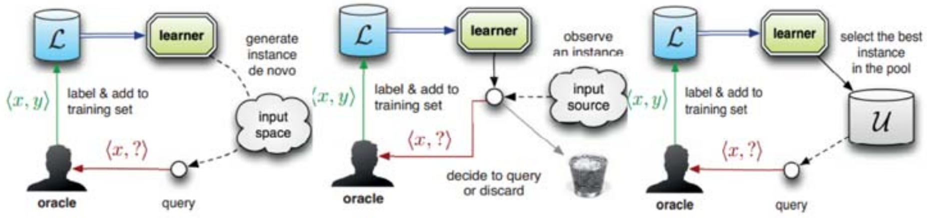
Goal: Automating Inquiries (Settles: alien fruits)

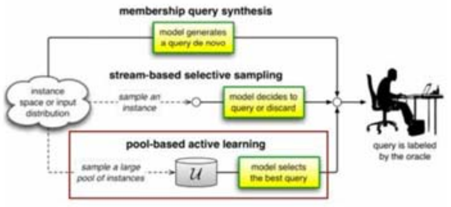
- A classifier to determine objects as a function mapping $h\colon X \rightarrow Y$, parameterized by a threshold θ:

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



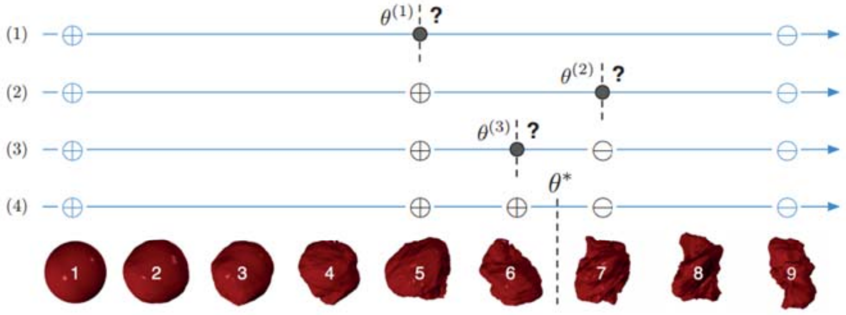
Scenarios for active learning




membership query synthesis
model generates a query de novo
stream-based selective sampling
instance space or input distribution
sample an instance
model decides to query or discard
pool-based active learning
sample a large pool of instances
$\mathcal{U}$
model selects the best query
query is labeled by the oracle

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

Uncertainty Sampling



```
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 $t = 1, 2, \ldots$ do
4:     $\theta = \textbf{train}(\mathcal{L})$
5:     select $x^* \in \mathcal{U}$, the most uncertain instance according to model $\theta$
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
```

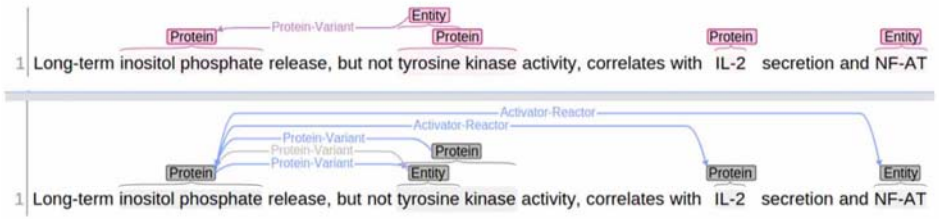
Settles, B. 2012. Active Learning, San Rafael (CA), Morgan & Claypool, doi:10.2200/S00429ED1V01Y201

From Active learning to Multi-Task Active learning

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

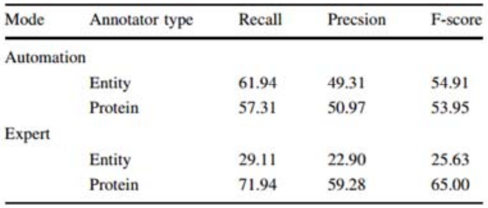
Settles, B. 2012. Active Learning, San Rafael (CA), Morgan & Claypool, doi:10.2200/S00429ED1V01Y201207AIM018.

Example for the Human-in-the-Loop



Long-term inositol phosphate release, but not tyrosine kinase activity, correlates with IL-2 secretion and NF-AT

Long-term inositol phosphate release, but not tyrosine kinase activity, correlates with IL-2 secretion and NF-AT

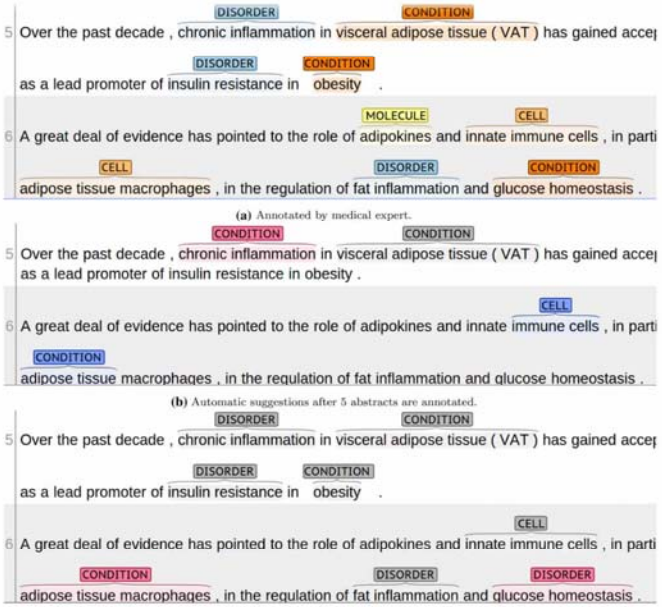


Mode	Annotator type	Recall	Precsion	F-score
Automation				
	Entity	61.94	49.31	54.91
	Protein	57.31	50.97	53.95
Expert				
	Entity	29.11	22.90	25.63
	Protein	71.94	59.28	65.00

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.

Example for the Human-in-the-Loop

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.



Over the past decade , chronic inflammation in visceral adipose tissue (VAT) has gained accep

as a lead promoter of insulin resistance in obesity .

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.

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

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.

Over the past decade , chronic inflammation in visceral adipose tissue (VAT) has gained accep

as a lead promoter of insulin resistance in obesity .

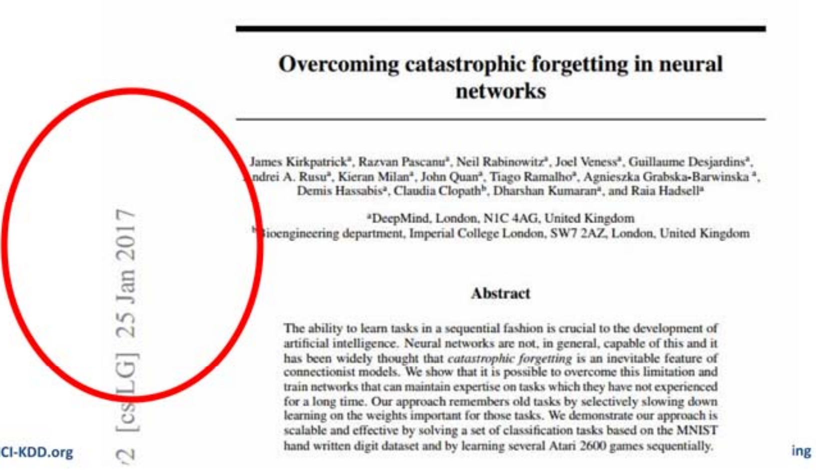
A great deal of evidence has pointed to the role of adipokines and innate immune cells , in parti

adipose tissue macrophages , in the regulation of fat inflammation and glucose homeostasis .

08 Multi-Task Learning

Catastrophic Forgetting

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



2 [cs.LG] 25 Jan 2017

Overcoming catastrophic forgetting in neural networks

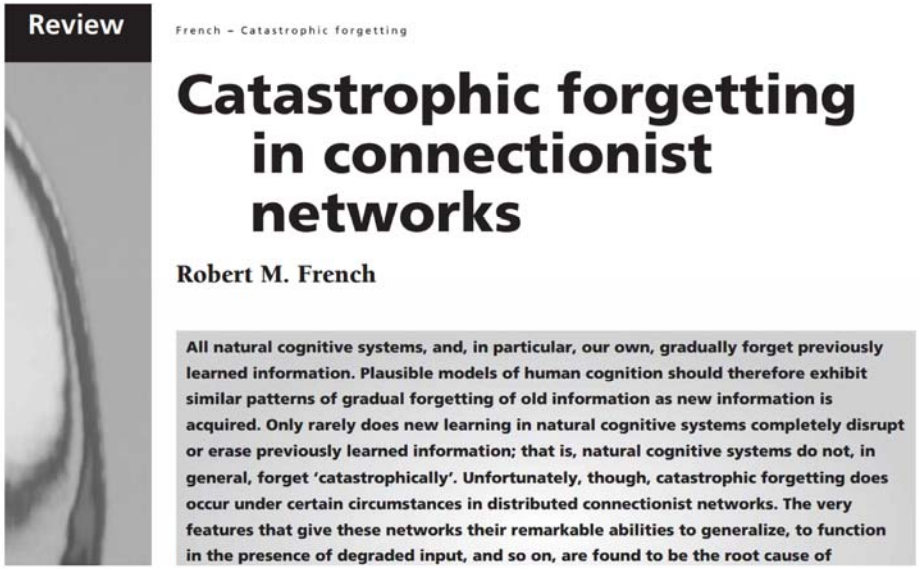
James Kirkpatrick[a], Razvan Pascanu[a], Neil Rabinowitz[a], Joel Veness[a], Guillaume Desjardins[a], Andrei A. Rusu[a], Kieran Milan[a], John Quan[a], Tiago Ramalho[a], Agnieszka Grabska-Barwinska [a], Demis Hassabis[a], Claudia Clopath[b], Dharshan Kumaran[a], and Raia Hadsell[a]

[a]DeepMind, London, N1C 4AG, United Kingdom
[b]Bioengineering 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.

"Old" Phenomenon



Review

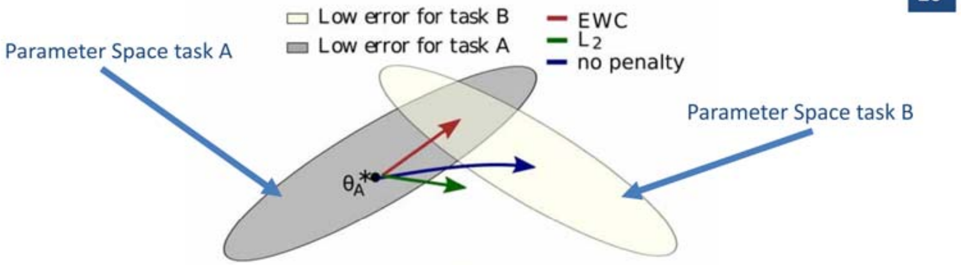
French – Catastrophic forgetting

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

Overcoming Catastrophic Forgetting: Deep Learning Bayes



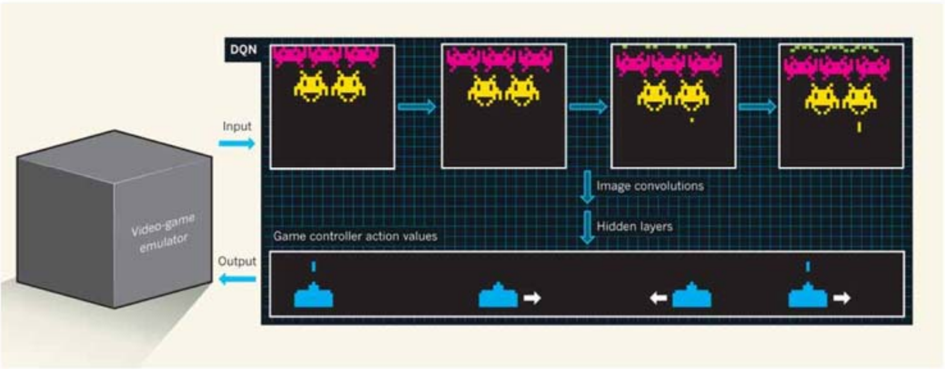
$$\log p(\theta|\mathcal{D}) = \log p(\mathcal{D}|\theta) + \log p(\theta) - \log p(\mathcal{D})$$

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

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

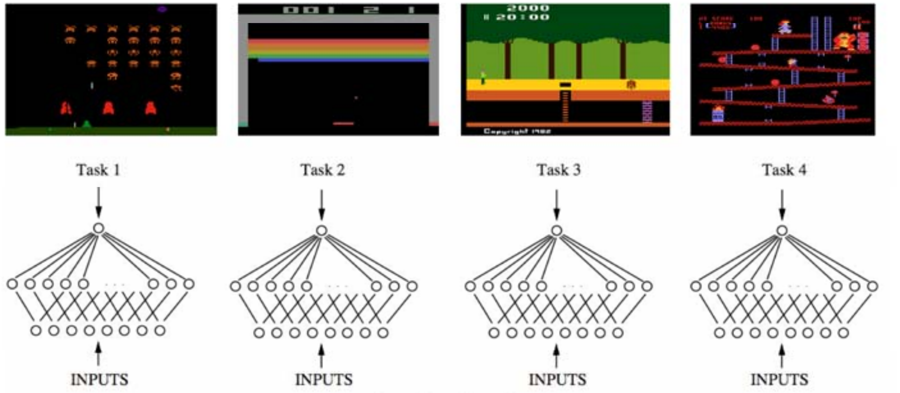
Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D. & Hadsell, R. 2016. Overcoming catastrophic forgetting in neural networks. arXiv preprint arXiv:1612.00796.

This experiment (2016) was done with Atari games ...

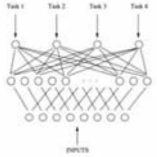


Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S. & Hassabis, D. 2015. Human-level control through deep reinforcement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

Example for Multi-Task Learning

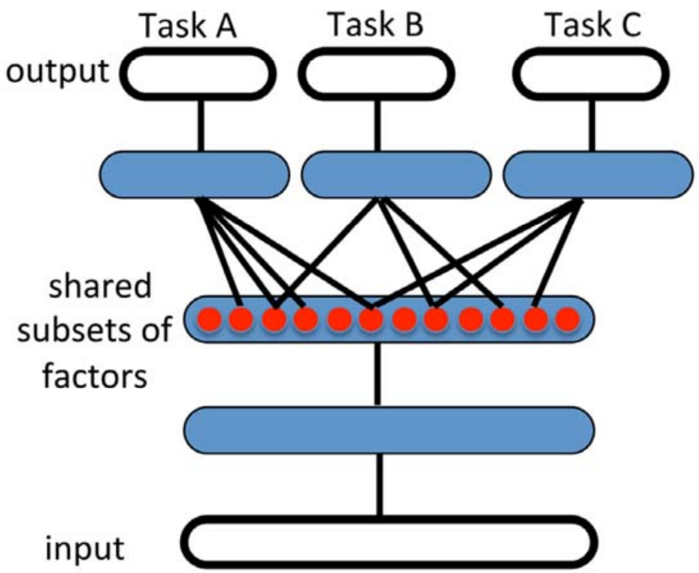


V. Mnih *et al.*, "Playing Atari with Deep Reinforcement Learning", Nature (2015)
Rich Caruana, "Multi-task Learning", MLJ (1998)



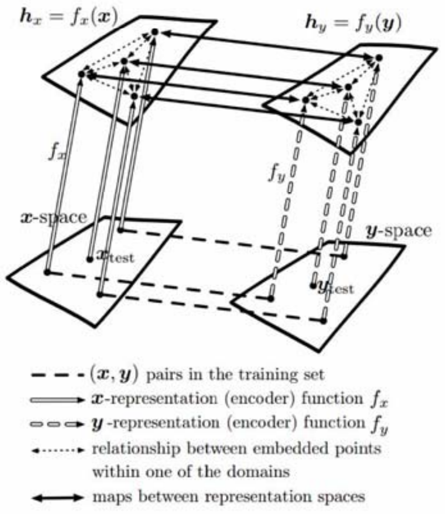
Representation Learning discovering explanatory factors

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.



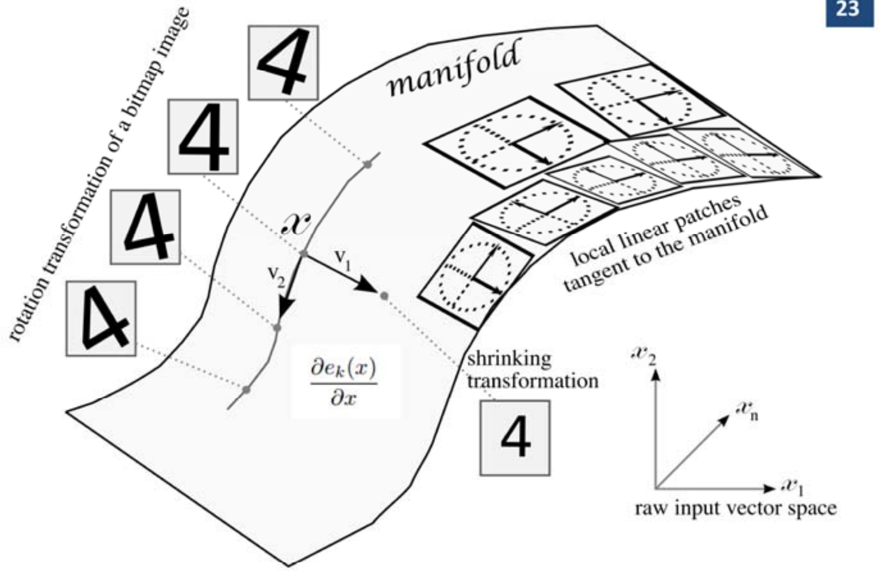
Maps between shared representations

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


$h_x = f_x(x)$
$h_y = f_y(y)$
f_x
f_y
x-space
y-space
x_{test}
y_{test}
(x, y) pairs in the training set
x-representation (encoder) function f_x
y-representation (encoder) function f_y
relationship between embedded points within one of the domains
maps between representation spaces

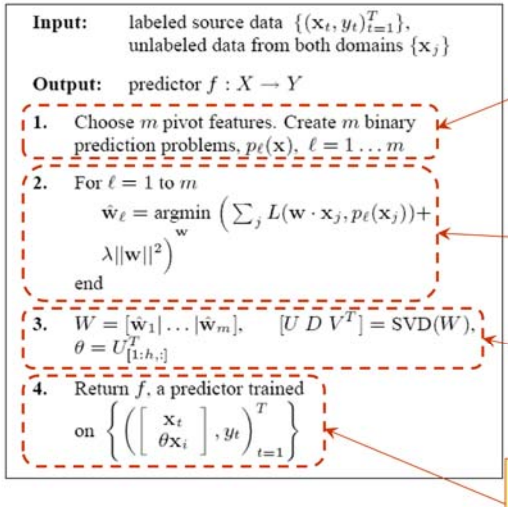
Goodfellow, I., Bengio, Y. & Courville, A. 2016. Deep Learning, Cambridge: MIT Press, p.542

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


rotation transformation of a bitmap image
4
manifold
x
v_1
v_2
$\frac{\partial e_k(x)}{\partial x}$
local linear patches tangent to the manifold
shrinking transformation
x_2
x_n
x_1
raw input vector space

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.

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



```
Input:   labeled source data $\{(\mathbf{x}_t, y_t)_{t=1}^T\}$,
         unlabeled data from both domains $\{\mathbf{x}_j\}$
Output:  predictor $f : X \to Y$
1. Choose $m$ pivot features. Create $m$ binary
   prediction problems, $p_\ell(\mathbf{x})$, $\ell = 1 \dots m$
2. For $\ell = 1$ to $m$
       $\hat{\mathbf{w}}_\ell = \underset{\mathbf{w}}{\mathrm{argmin}} \left( \sum_j L(\mathbf{w} \cdot \mathbf{x}_j, p_\ell(\mathbf{x}_j)) + \lambda ||\mathbf{w}||^2 \right)$
   end
3. $W = [\hat{\mathbf{w}}_1 | \dots | \hat{\mathbf{w}}_m]$,   $[U\ D\ V^T] = \mathrm{SVD}(W)$,
   $\theta = U^T_{[1:h,:]}$
4. Return $f$, a predictor trained
   on $\left\{ \left( \begin{bmatrix} \mathbf{x}_t \\ \theta \mathbf{x}_i \end{bmatrix}, y_t \right)_{t=1}^T \right\}$
```

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.

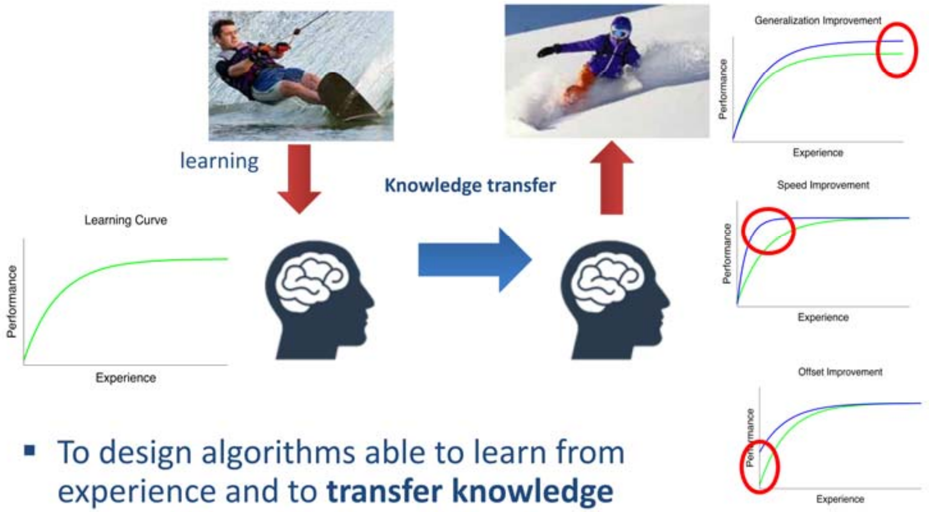
Use the learnt mapping function to construct new features and train classifiers onto the new representations.

09 Generalization & Transfer Learning

Transfer Learning is studied for more than 100 years

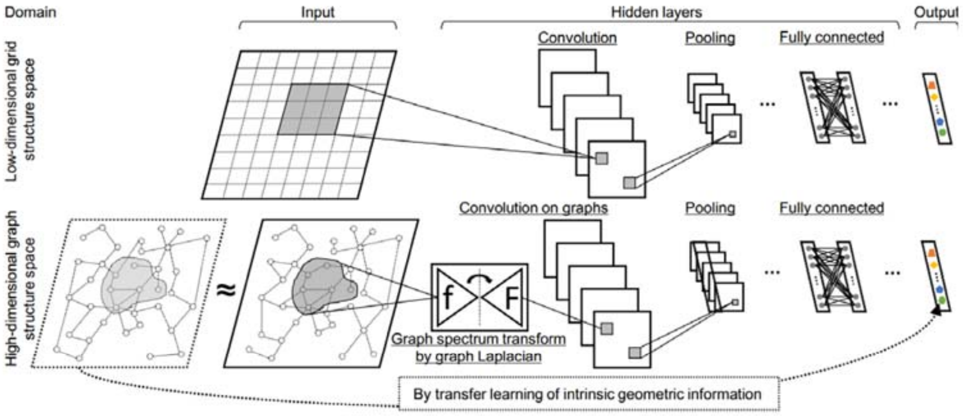
- Thorndike & Woodworth (1901) explored how individuals would transfer in one context to another context that share similar characteristics:
- or how "improvement in one mental function" could influence a related one
- Their theory implied that transfer of learning depends on how similar the learning 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

Grand Challenge: Transfer Learning



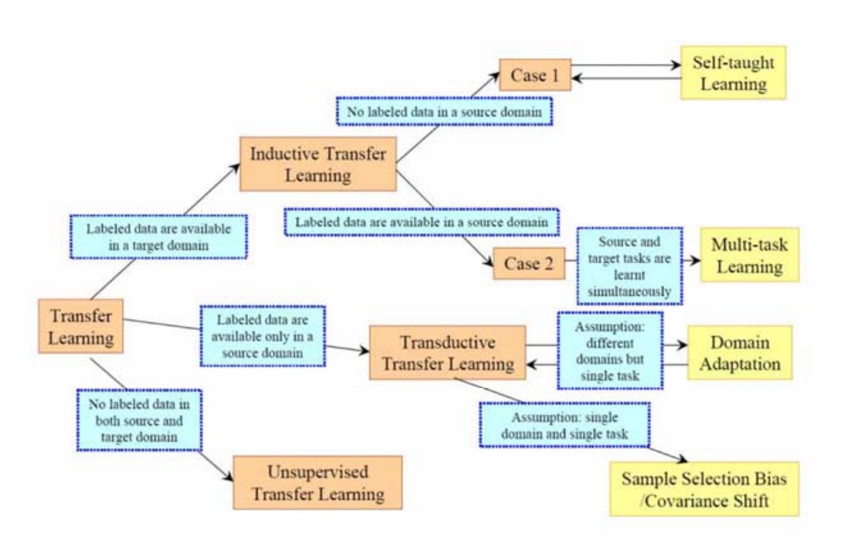
- To design algorithms able to learn from experience and to **transfer knowledge across different tasks and domains** to improve their learning performance

Transfer Learning for Deep Learning on Graphs


Domain
Input
Hidden layers
Output
Low-dimensional grid structure space
Convolution
Pooling
Fully connected
...
...
High-dimensional graph structure space
≈
Convolution on graphs
Pooling
Fully connected
f
F
Graph spectrum transform by graph Laplacian
By transfer learning of intrinsic geometric information

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

Overview



Transfer Learning

Labeled data are available in a target domain

Inductive Transfer Learning

No labeled data in a source domain

Case 1

Self-taught Learning

Labeled data are available in a source domain

Case 2

Source and target tasks are learnt simultaneously

Multi-task Learning

Labeled data are available only in a source domain

Transductive Transfer Learning

Assumption: different domains but single task

Domain Adaptation

Assumption: single domain and single task

Sample Selection Bias /Covariance Shift

No labeled data in both source and target domain

Unsupervised Transfer Learning

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.

Domain and Task

- Feature space $\mathcal{X}$;
- $P(x)$, where $x \in \mathcal{X}$.

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

Two domains are different $\Rightarrow$

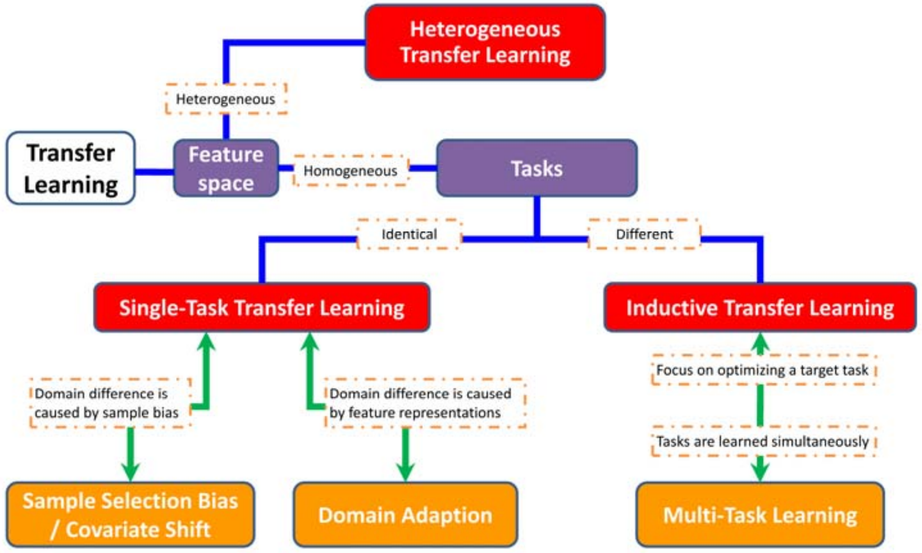
$\mathcal{X}_S \neq \mathcal{X}_T$, or $P_S(x) \neq P_T(x)$.

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.

Transfer Learning Settings



Heterogeneous Transfer Learning

Heterogeneous

Transfer Learning

Feature space

Homogeneous

Tasks

Identical

Different

Single-Task Transfer Learning

Inductive Transfer Learning

Domain difference is caused by sample bias

Domain difference is caused by feature representations

Focus on optimizing a target task

Tasks are learned simultaneously

Sample Selection Bias / Covariate Shift

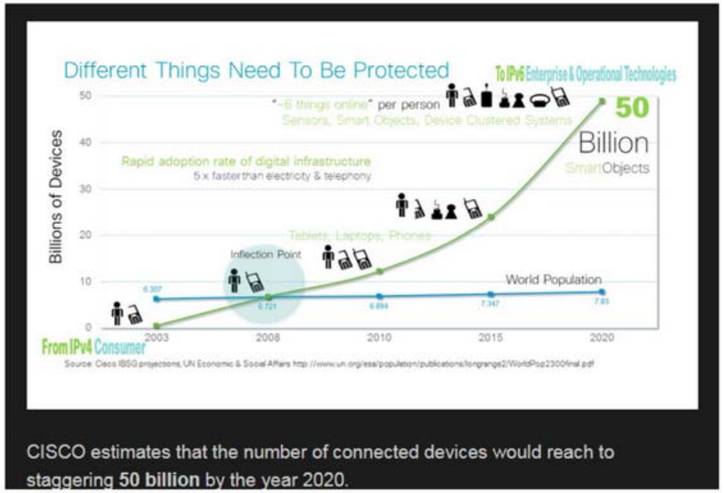
Domain Adaption

Multi-Task Learning

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.

10 Federated Machine Learning

More mobile devices on this planet than humans …

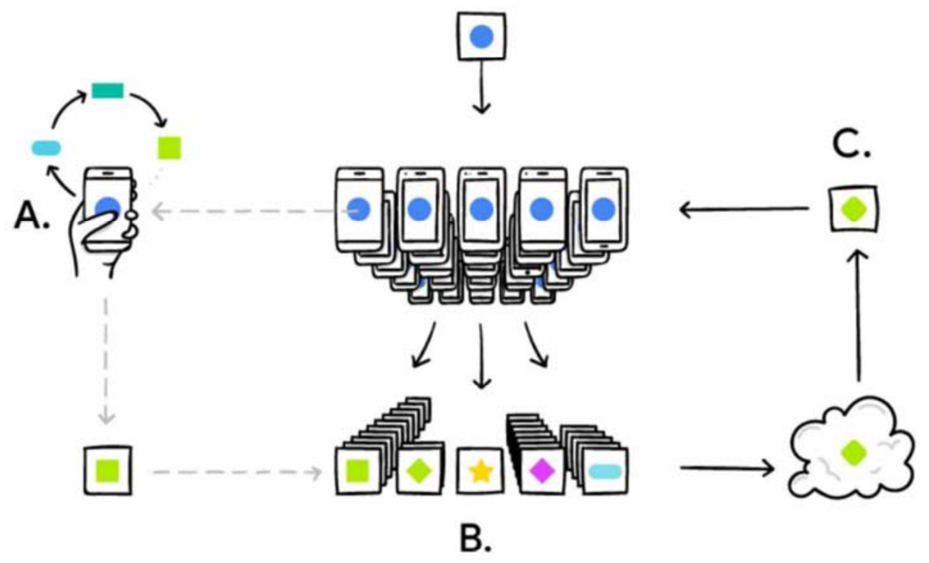


CISCO estimates that the number of connected devices would reach to staggering **50 billion** by the year 2020.

https://iotblogg.wordpress.com/

http://www.independent.co.uk/life-style/gadgets-and-tech/news/there-are-officially-more-mobile-devices-than-people-in-the-world-9780518.html

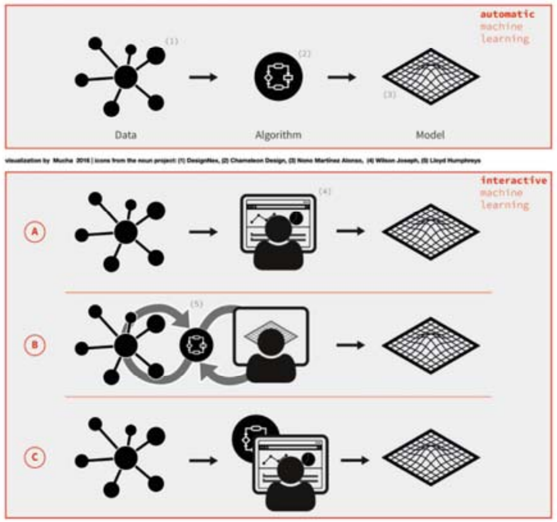
Federated Collaborative Optimization and Learning



https://research.googleblog.com/2017/04/federated-learning-collaborative.html

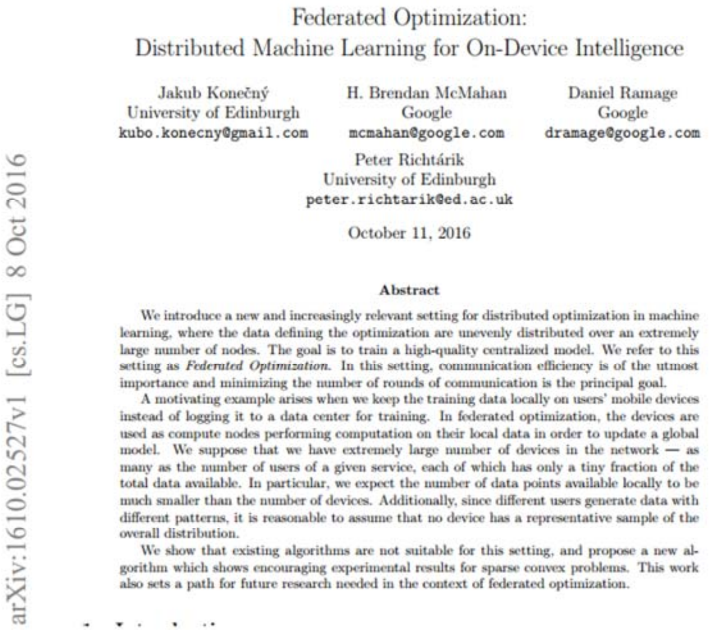
Collaborative interactive Machine Learning (ciML)

https://rd.springer.com/chapter/10.1007/978-3-319-50478-0_18


automatic
machine
learning
(1)
(2)
(3)
Data
Algorithm
Model
visualization by Mucha 2016 | icons from the noun project: (1) DesignNex, (2) Chameleon Design, (3) Nono Martínez Alonso, (4) Wilson Joseph, (5) Lloyd Humphreys
interactive
machine
learning
(4)
A
(5)
B
C

Robert, S., Büttner, S., Röcker, C. & Holzinger, A. 2016. Reasoning Under Uncertainty: Towards Collaborative Interactive Machine Learning. In: Machine Learning for Health Informatics: Lecture Notes in Artifical Intelligence LNAI 9605. Springer, pp. 357-376

On-Device Intelligence



Federated Optimization: Distributed Machine Learning for On-Device Intelligence

Jakub Konečný
University of Edinburgh
kubo.konecny@gmail.com

H. Brendan McMahan
Google
mcmahan@google.com

Daniel Ramage
Google
dramage@google.com

Peter Richtárik
University of Edinburgh
peter.richtarik@ed.ac.uk

October 11, 2016

arXiv:1610.02527v1 [cs.LG] 8 Oct 2016

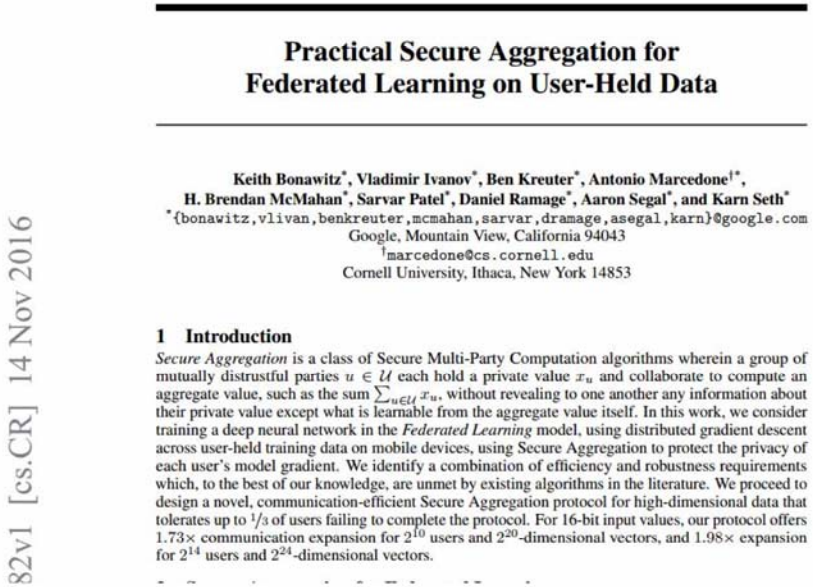
Abstract

We introduce a new and increasingly relevant setting for distributed optimization in machine learning, where the data defining the optimization are unevenly distributed over an extremely large number of nodes. The goal is to train a high-quality centralized model. We refer to this setting as *Federated Optimization*. In this setting, communication efficiency is of the utmost importance and minimizing the number of rounds of communication is the principal goal.

A motivating example arises when we keep the training data locally on users' mobile devices instead of logging it to a data center for training. In federated optimization, the devices are used as compute nodes performing computation on their local data in order to update a global model. We suppose that we have extremely large number of devices in the network — as many as the number of users of a given service, each of which has only a tiny fraction of the total data available. In particular, we expect the number of data points available locally to be much smaller than the number of devices. Additionally, since different users generate data with different patterns, it is reasonable to assume that no device has a representative sample of the overall distribution.

We show that existing algorithms are not suitable for this setting, and propose a new algorithm which shows encouraging experimental results for sparse convex problems. This work also sets a path for future research needed in the context of federated optimization.

Federated Machine Learning



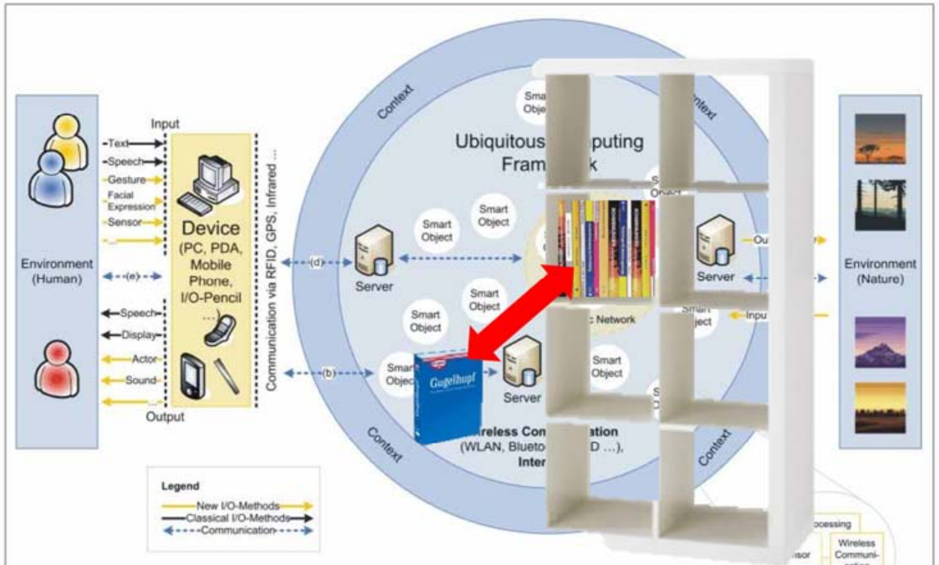
Practical Secure Aggregation for Federated Learning on User-Held Data

Keith Bonawitz[*], Vladimir Ivanov[*], Ben Kreuter[*], Antonio Marcedone[†*],
H. Brendan McMahan[*], Sarvar Patel[*], Daniel Ramage[*], Aaron Segal[*], and Karn Seth[*]
[*]{bonawitz,vlivan,benkreuter,mcmahan,sarvar,dramage,asegal,karn}@google.com
Google, Mountain View, California 94043
[†]marcedone@cs.cornell.edu
Cornell University, Ithaca, New York 14853

82v1 [cs.CR] 14 Nov 2016

1 Introduction

Secure Aggregation is a class of Secure Multi-Party Computation algorithms wherein a group of mutually distrustful parties $u \in \mathcal{U}$ each hold a private value x_u and collaborate to compute an aggregate value, such as the sum $\sum_{u \in \mathcal{U}} x_u$, without revealing to one another any information about their private value except what is learnable from the aggregate value itself. In this work, we consider training a deep neural network in the *Federated Learning* model, using distributed gradient descent across user-held training data on mobile devices, using Secure Aggregation to protect the privacy of each user's model gradient. We identify a combination of efficiency and robustness requirements which, to the best of our knowledge, are unmet by existing algorithms in the literature. We proceed to design a novel, communication-efficient Secure Aggregation protocol for high-dimensional data that tolerates up to $^1/_3$ of users failing to complete the protocol. For 16-bit input values, our protocol offers 1.73× communication expansion for 2^{10} users and 2^{20}-dimensional vectors, and 1.98× expansion for 2^{14} users and 2^{24}-dimensional vectors.



Holzinger, A., Nischelwitzer, A., Friedl, S. & Hu, B. 2010. Towards life long learning: three models for ubiquitous applications. Wireless Communications and Mobile Computing, 10, (10), 1350-1365, doi:10.1002/wcm.715.

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.

Thank you!

Questions

Questions (1/4)

- 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?

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

Open Questions, future outlook, hot topics, challenges

- 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**

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**

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)

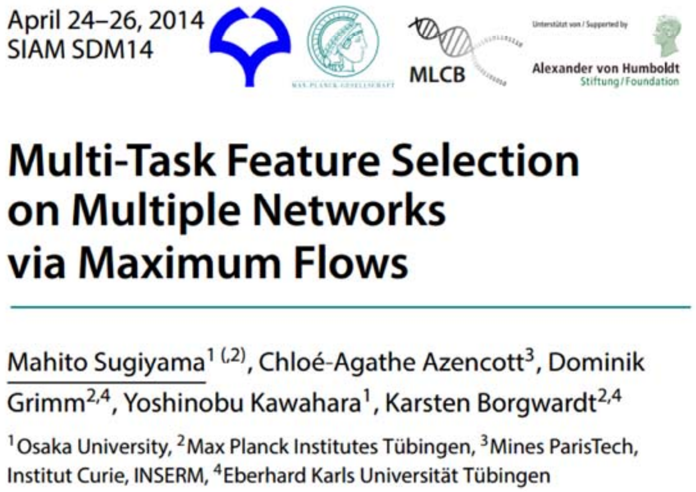
Humanoid AI

≠

Human-level AI

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

MLCB
Unterstützt von / Supported by
Alexander von Humboldt
Stiftung / Foundation

Multi-Task Feature Selection on Multiple Networks via Maximum Flows

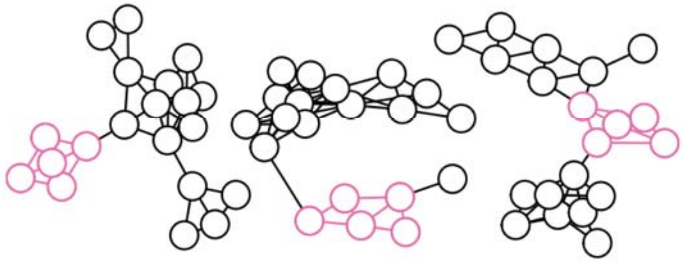
Mahito Sugiyama[1 (,2)], Chloé-Agathe Azencott[3], Dominik Grimm[2,4], Yoshinobu Kawahara[1], Karsten Borgwardt[2,4]

[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

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

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

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

- 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

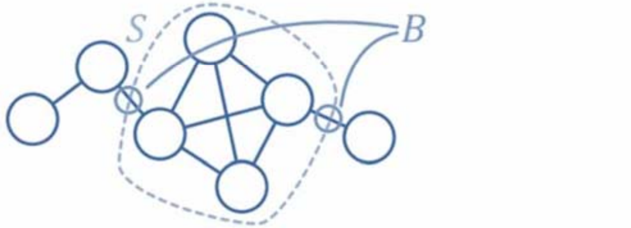
Azencott, C.-A., Grimm, D., Sugiyama, M., Kawahara, Y. & Borgwardt, K. M. 2013. Efficient network-guided multi-locus association mapping with graph cuts. Bioinformatics, 29, (13), i171-i179.

Formulation of SConES

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

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

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



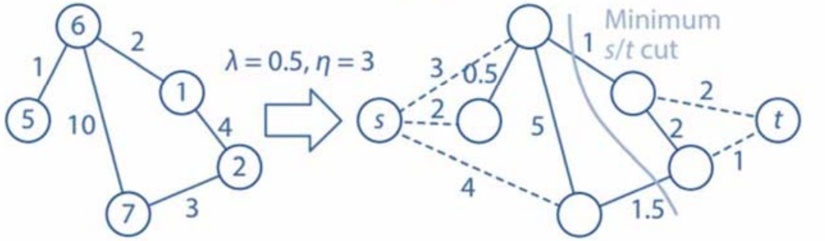
Azencott, C.-A., Grimm, D., Sugiyama, M., Kawahara, Y. & Borgwardt, K. M. 2013. Efficient network-guided multi-locus association mapping with graph cuts. Bioinformatics, 29, (13), i171-i179.

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

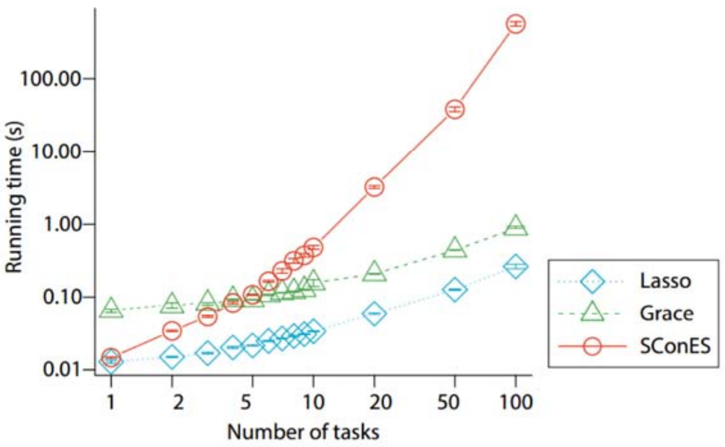
$$c(\{v,u\}) = \begin{cases} |\, q(u) - \eta \,| & \text{if } u \in \{s,t\} \text{ and } v \in V, \\ \lambda w(\{v,u\}) & \text{otherwise} \end{cases}$$

- 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

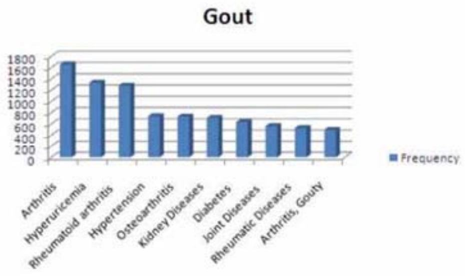
Let two words, w_i and w_j, have probabilities $P(w_i)$ and $P(w_j)$. Then their mutual information $PMI\ (w_i, w_i)$ is defined as:

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

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

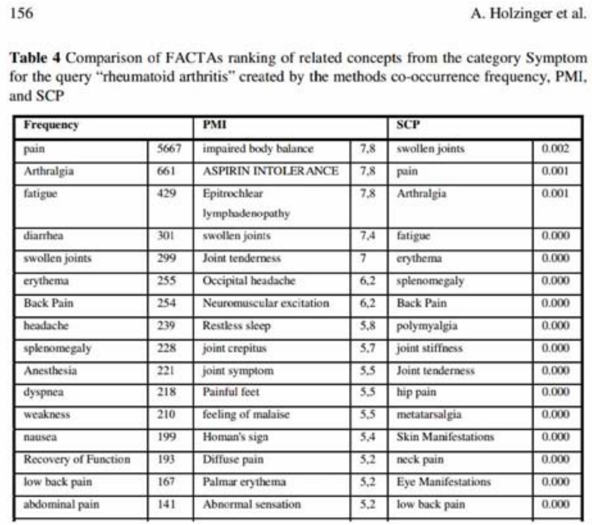
$$P(w_i) = \frac{94{,}834}{20{,}033{,}079},\quad P(w_j) = \frac{74}{20{,}033{,}079}$$

$$P(w_i, w_j) = \frac{13}{94{,}834},\quad PMI(w_i, w_j) = 7{,}7.$$



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.

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



156 A. Holzinger et al.

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

Frequency		PMI		SCP	
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

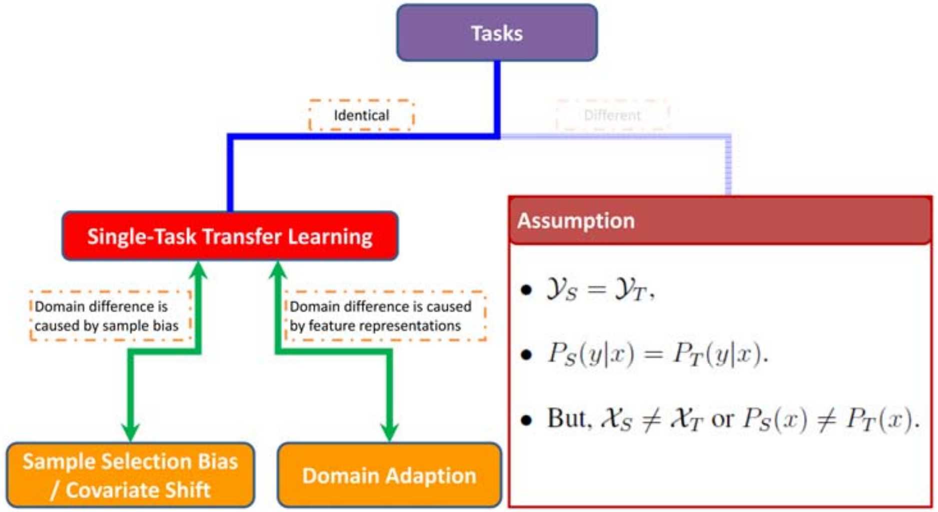
Holzinger, A., Yildirim, P., Geier, M. & Simonic, K.-M. 2013. Quality-Based Knowledge Discovery from Medical Text on the Web. In: Pasi, G., Bordogna, G. & Jain, L. C. (eds.) Quality Issues in the Management of Web Information, Intelligent Systems Reference Library, ISRL 50. Berlin Heidelberg: Springer,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

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



Tasks

Identical

Different

Single-Task Transfer Learning

Domain difference is caused by sample bias

Domain difference is caused by feature representations

Sample Selection Bias / Covariate Shift

Domain Adaption

Assumption

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

Egerstedt, M. 2011. Complex networks: Degrees of control. *Nature, 473*, ***158-159.***

Open Problem: How to avoid negative transfer?