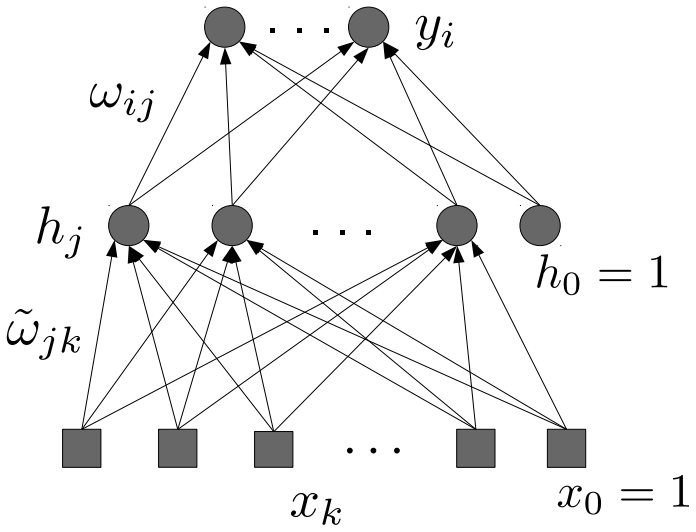
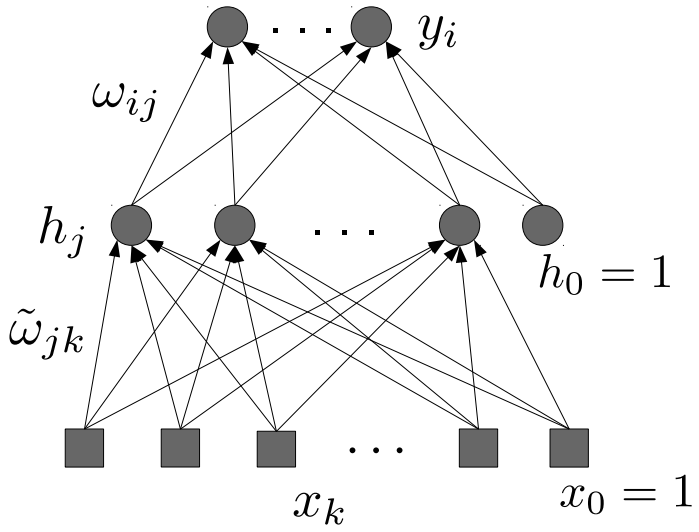
Interesting
Deep Learning / Machine Learning
Topics

Topic 1: Understanding a deep learning model



How can we understand what is happening inside the “Black Box”?

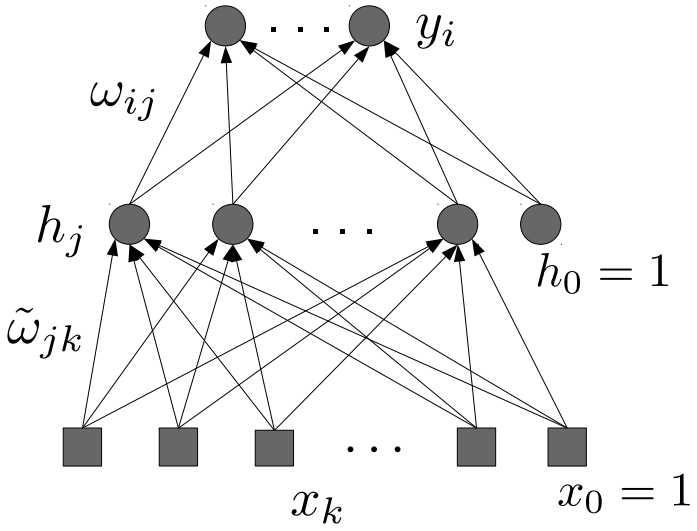
What do we mean by ”understand”?



Task!

Make a ranking of the importance of the different inputs?

How?



Task!

For a given input x, how can I understand the “decision”?

Can I understand what are the important features that make a cat a cat and a dog a dog, from the network point of view?

A new field has emerged

XAI

(Explainable Artificial Intelligence)

arXiv:1910.10045v2 [cs.AI] 26 Dec 2019

Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI

Alejandro Barredo Arrieta[a], Natalia Díaz-Rodríguez[b], Javier Del Ser[c,d], Adrien Bennetot[b,e,f], Siham Tabik[g], Alberto Barbado[h], Salvador Garcia[g], Sergio Gil-Lopez[a], Daniel Molina[g], Richard Benjamins[h], Raja Chatila[f], and Francisco Herrera[g]

[a]TECNALIA, 48160 Derio, Spain
[b]ENSTA, Institute Polytechnique Paris and INRIA Flowers Team, Palaiseau, France
[c]University of the Basque Country (UPV/EHU), 48013 Bilbao, Spain
[d]Basque Center for Applied Mathematics (BCAM), 48009 Bilbao, Bizkaia, Spain
[e]Segula Technologies, Parc d'activité de Pissaloup, Trappes, France
[f]Institut des Systèmes Intelligents et de Robotique, Sorbonne Université, France
[g]DaSCI Andalusian Institute of Data Science and Computational Intelligence, University of Granada, 18071 Granada, Spain
[h]Telefonica, 28050 Madrid, Spain

Abstract

In the last few years, Artificial Intelligence (AI) has achieved a notable momentum that, if harnessed appropriately, may deliver the best of expectations over many application sectors across the field. For this to occur shortly in Machine Learning, the entire community stands in front of the barrier of explainability, an inherent problem of the latest techniques brought by sub-symbolism (e.g. ensembles or Deep Neural Networks) that were not present in the last hype of AI (namely, expert systems and rule based models). Paradigms underlying this problem fall within the so-called *eXplainable* AI (XAI) field, which is widely acknowledged as a crucial feature for the practical deployment of AI models. The overview presented in this article examines the existing literature and contributions already done in the field of XAI, including a prospect toward what is yet to be reached. For this purpose we summarize previous efforts made to define explainability in Machine Learning, establishing a novel definition of explainable Machine Learning that covers such prior conceptual propositions with a major focus on the audience for which the explainability is sought. Departing from this definition, we propose and discuss about a taxonomy of recent contributions related to the explainability of different Machine Learning models, including those aimed at explaining Deep Learning methods for which a second dedicated taxonomy is built and examined in detail. This critical literature analysis serves as the motivating background for a series of challenges faced by XAI, such as the interesting crossroads of data fusion and explainability. Our prospects lead toward the concept of *Responsible Artificial Intelligence*, namely, a methodology for the large-scale implementation of AI methods in real organizations with fairness, model explainability and accountability at its core. Our ultimate goal is to provide newcomers to the field of XAI with a thorough taxonomy that can serve as reference material in order to stimulate future research advances, but also to encourage experts and professionals from other disciplines to embrace the benefits of AI in their activity sectors, without any prior bias for its lack of interpretability.

Keywords: Explainable Artificial Intelligence, Machine Learning, Deep Learning, Data Fusion, Interpretability, Comprehensibility, Transparency, Privacy, Fairness, Accountability, Responsible Artificial Intelligence.

*Corresponding author. TECNALIA. P. Tecnologico, Ed. 700. 48170 Derio (Bizkaia), Spain. E-mail: javier.delser@tecnalia.com

Recommended reading:

“

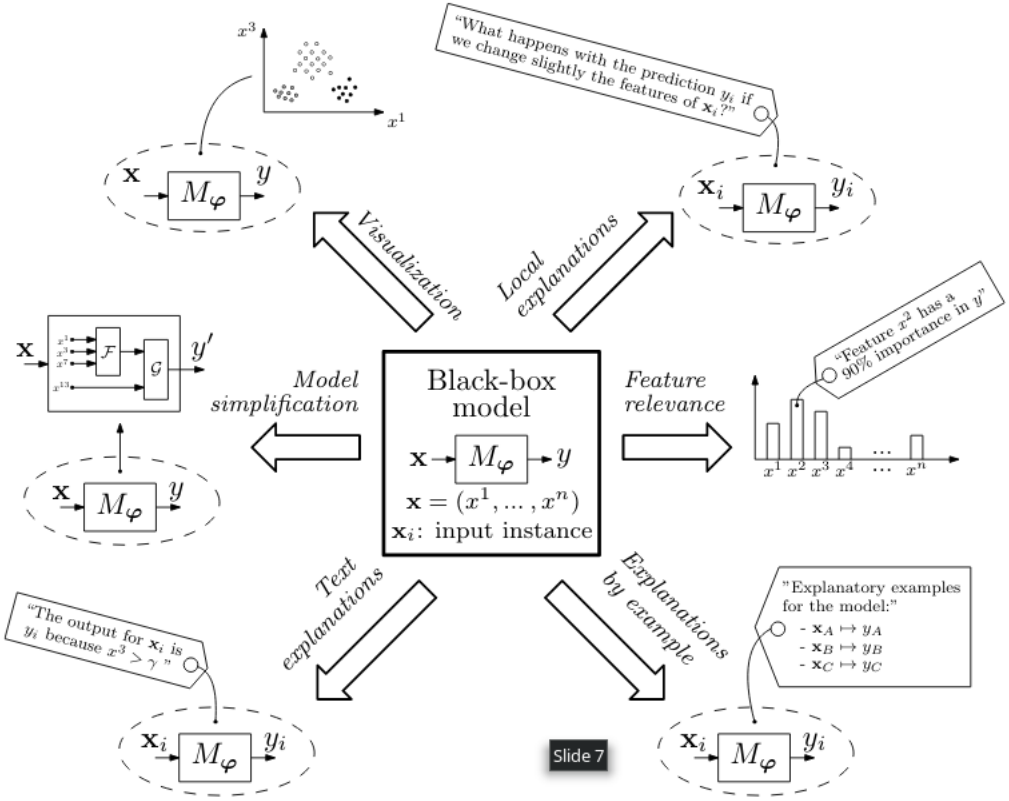
Explainable Artificial Intelligence (XAI): Concepts, Taxonomies,Opportunities and Challenges toward Responsible AI

”

Definition:

*Given an audience, an **explainable** Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand.*

Examples of approaches



Topic 2: Providing uncertainty estimations

Why do we need uncertainty estimations for DNNs?

An MNIST example (from the paper:

Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images

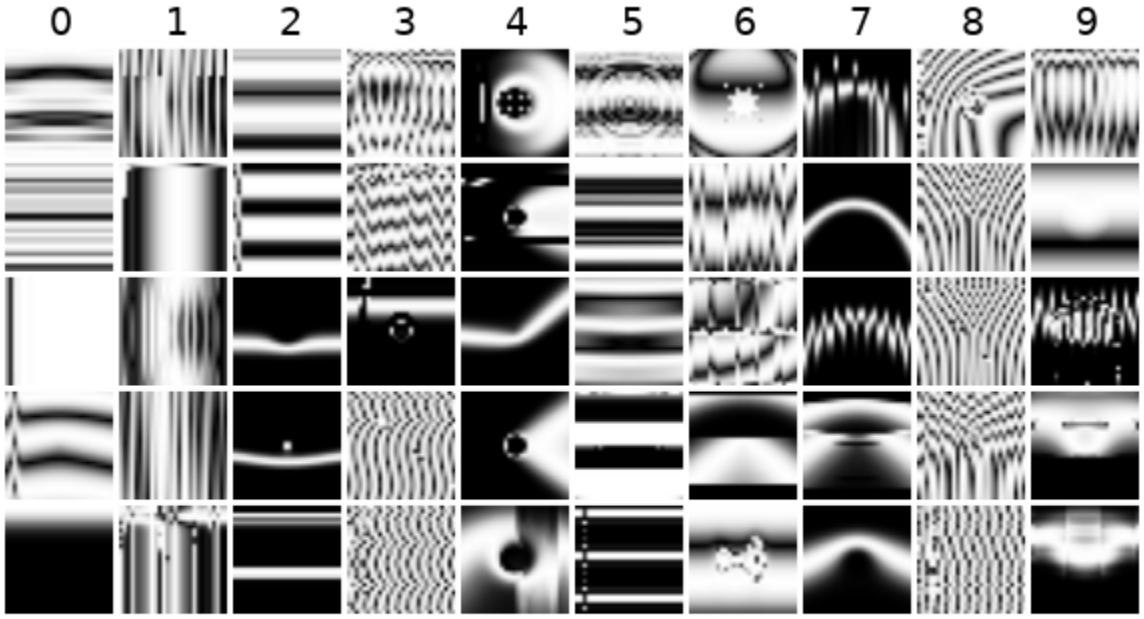


Figure 5. Indirectly encoded, thus regular, images that MNIST DNNs believe with 99.99% confidence are digits 0-9. The column and row descriptions are the same as for Fig. 4.

$$p(d|\mathbf{x}, \mathcal{D}) = \int p(d|\mathbf{x}, \boldsymbol{\omega}) p(\boldsymbol{\omega}|\mathcal{D}) d\boldsymbol{\omega}$$

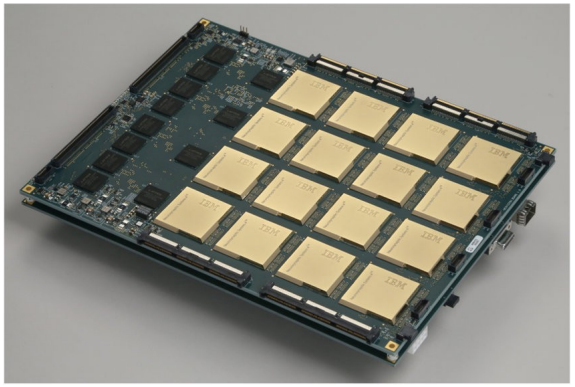
Bayesian learning of neural networks

- Markov Chain Monte Carlo methods
- Variational Inference methods

Alternative approaches

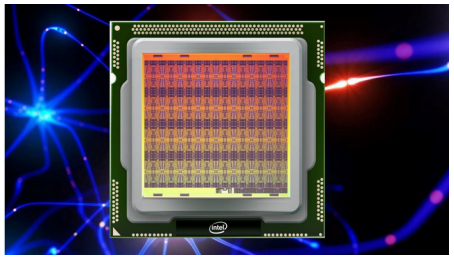
- Monte-Carlo Dropout
- Deep Ensembles

Topic 3: Neuromorphic computing / engineering



Neuromorphic Hardware: “It encompasses any electrical device which mimics the natural biological structures of our nervous system. The goal is to impart cognitive abilities to a machine by implementing neurons in silicon. Due to its much better energy efficiency and parallelism it is being considered as an alternative over conventional architectures and energy hungry GPUs.”

- IBM Truenorth: 4096 cores, each core 256 neurons with each 256 synapses

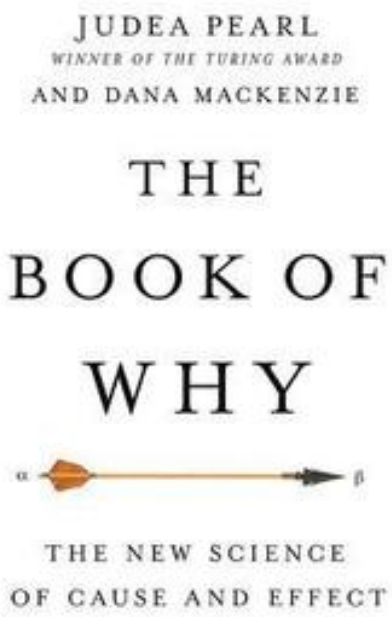


- Intel Loihi: 128 cores with 1096 neurons in each core.

- SpiNNaker: “Neuromorphic supercomputer” in connection with the Human Brain Project.

Topic 4: Causal Inference and machine learning

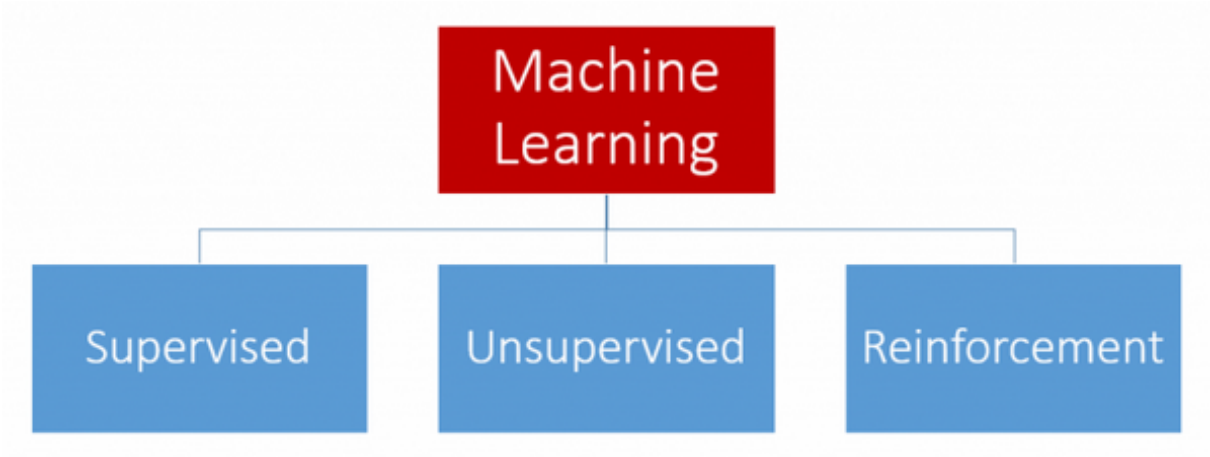
Recommended reading



To simplify: Deep learning is good at finding relationships between X and X, but will not easily reveal causal relations between X and Y.

Causal inference is used mostly to reach a prescription in the form of do X so that Y happens.

Topic 5: Reinforcement learning

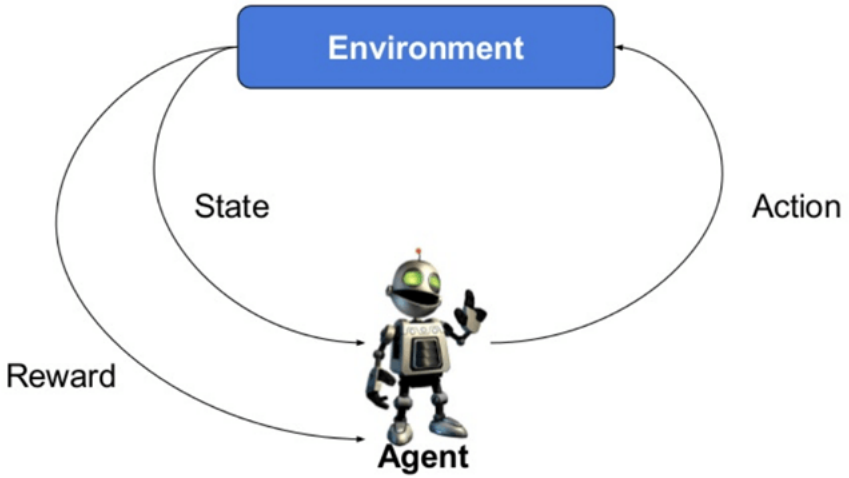


Task driven (Regression/ Classification)

Data driven (Clustering)

Algorithms learns to react to an environment

Reinforcement scenario



- **Agent:** It is an assumed entity which performs actions in an environment to gain some reward.
- **Environment**: A scenario that an agent has to face.
- **Reward**: Return given to an agent when he or she performs specific action or task.
- **State**: State refers to the current situation returned by the environment.

Images taken from: https://www.guru99.com/reinforcement-learning-tutorial.html

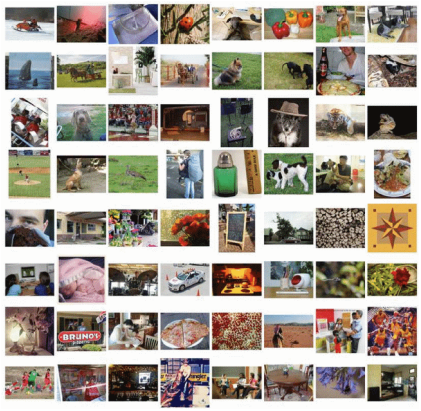
https://www.youtube.com/watch?v=gn4nRCC9TwQ

Topic 6: Machine learning in medical applications

Why using machine learning on medical data?

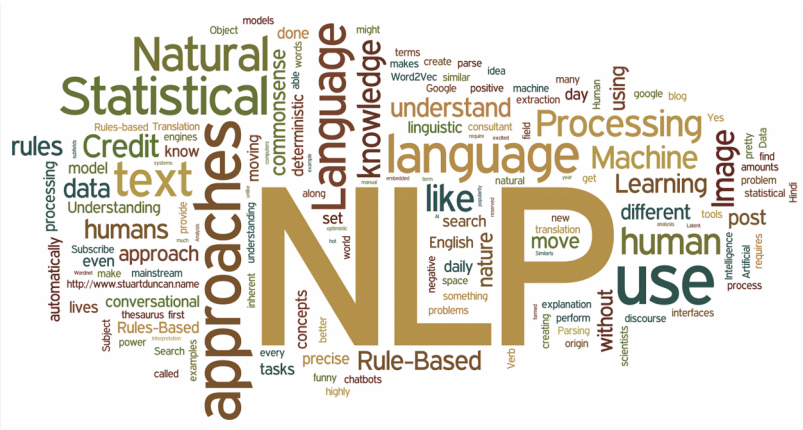
Two strong ML areas of relevance for healthcare

Image analysis



Diagnostic support

Natural language processing

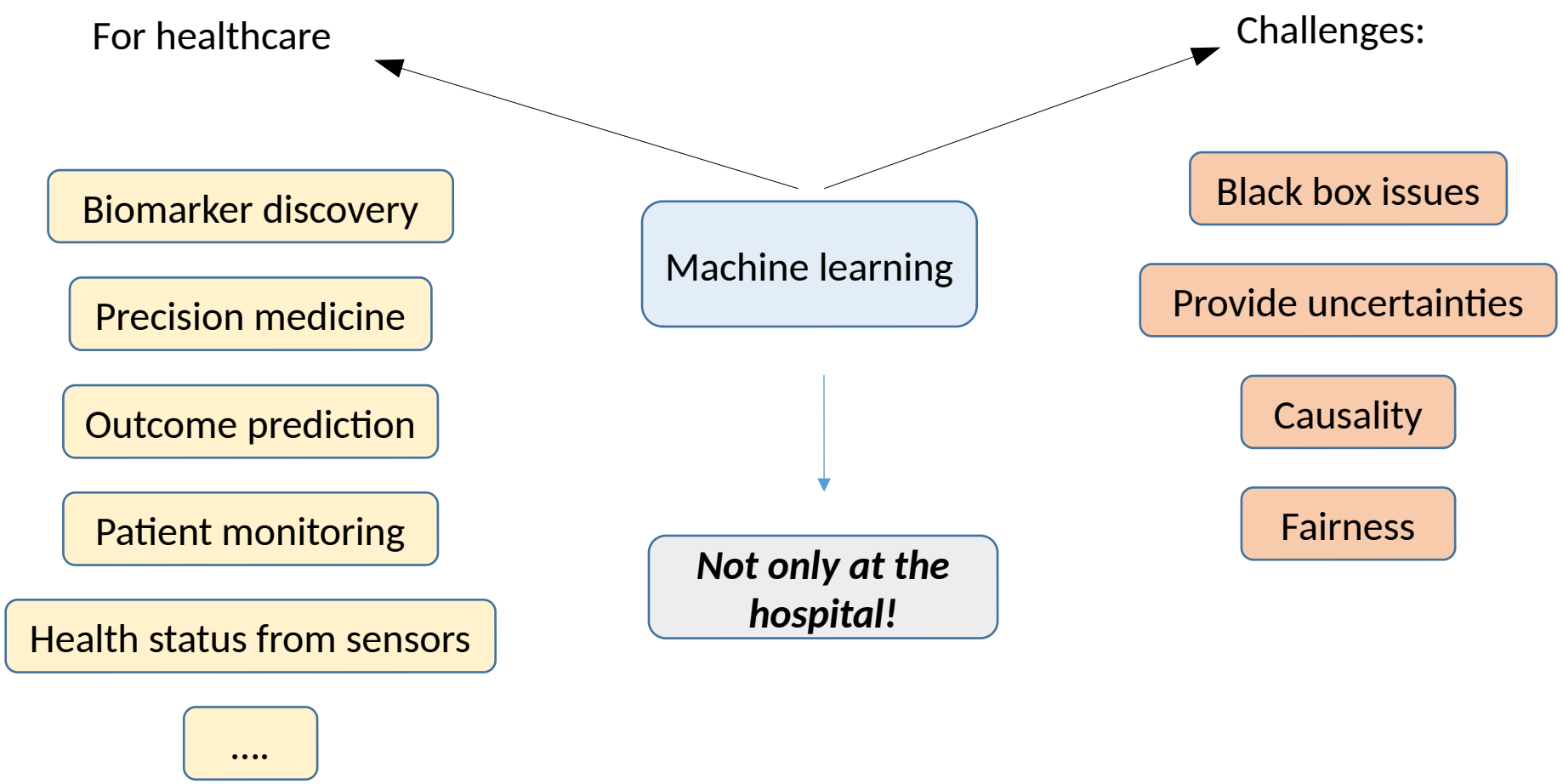


Textmining for scientific publications

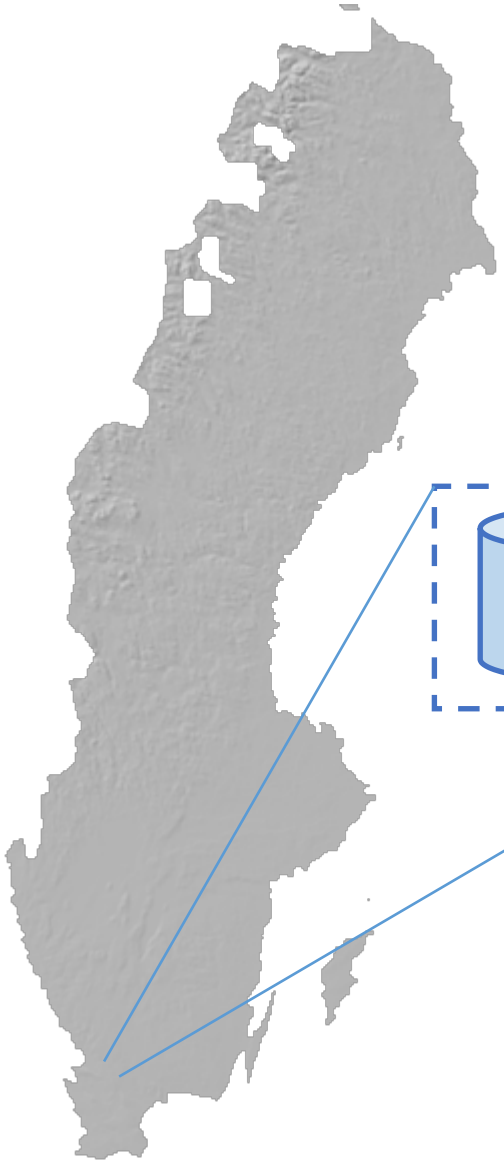
Analyzing patient records

Digitalization

Machine learning is good for finding patterns in data!



“Using machine learning on medical data” – Today!



Typical approach

Anonymized data

Machine learning

Medical data

Alternative approach

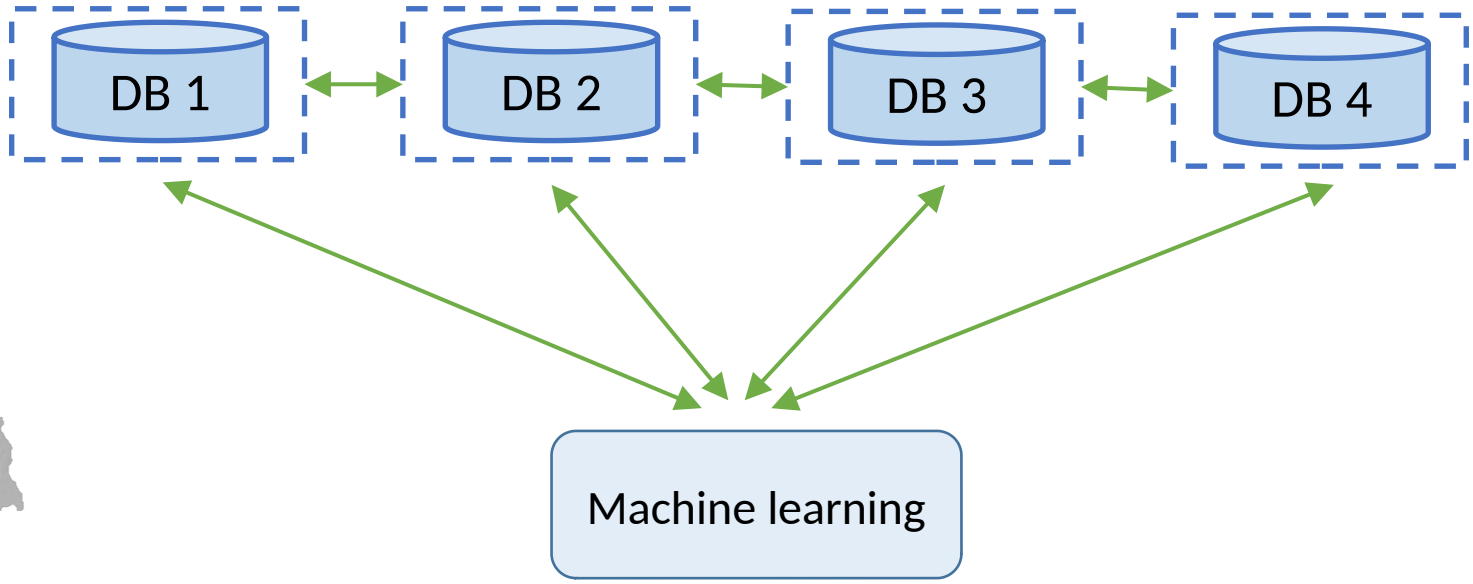
Medical data

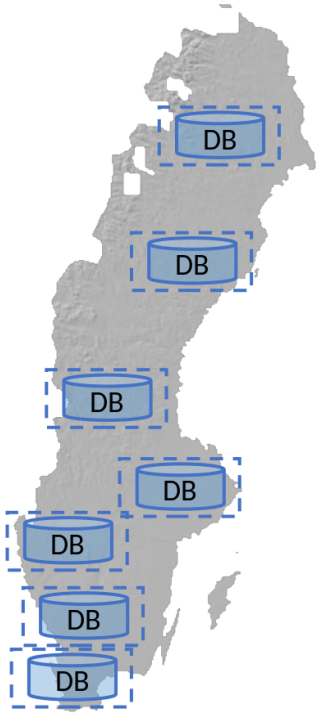
Machine learning

“Region Halland”

Privacy wall

“Using machine learning on medical data” – Tomorrow!





Privacy preserving communication

Privacy preserving distributed machine learning for medical data