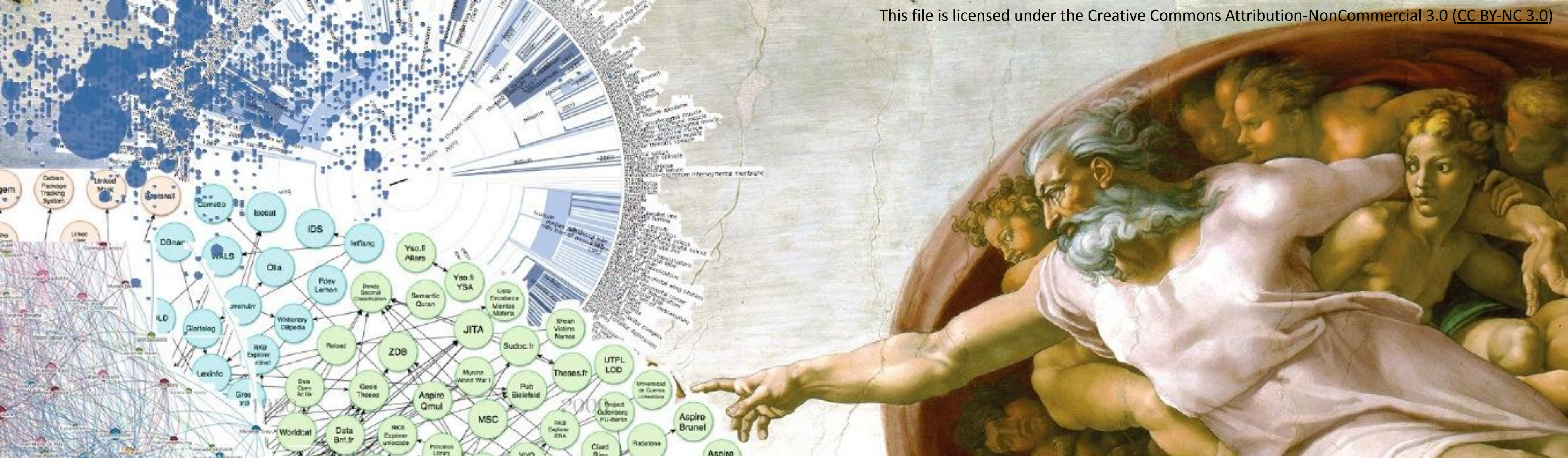


# Information Service Engineering

## Lecture 10: Basic Machine Learning - 1

Karlsruher Institut für Technologie

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

Prof. Dr. Harald Sack
FIZ Karlsruhe - Leibniz Institute for Information Infrastructure
AIFB - Karlsruhe Institute of Technology

**Summer Semester 2021**

# Information Service Engineering

## Last Lecture: Knowledge Graphs - 4



- OWL Building Blocks
- OWL Complex Classes
- OWL Strict & Loose Bindings
- OWL Property Paths
- Knowledge Graph Programming
- Graph Theory for Knowledge Graphs

## Lecture Overview

1. Information, Natural Language and the Web
2. Natural Language Processing
3. Knowledge Graphs
4. **Basic Machine Learning**
5. ISE Applications

https://money.cnn.com/2017/03/03/technology/google-breast-cancer-ai/

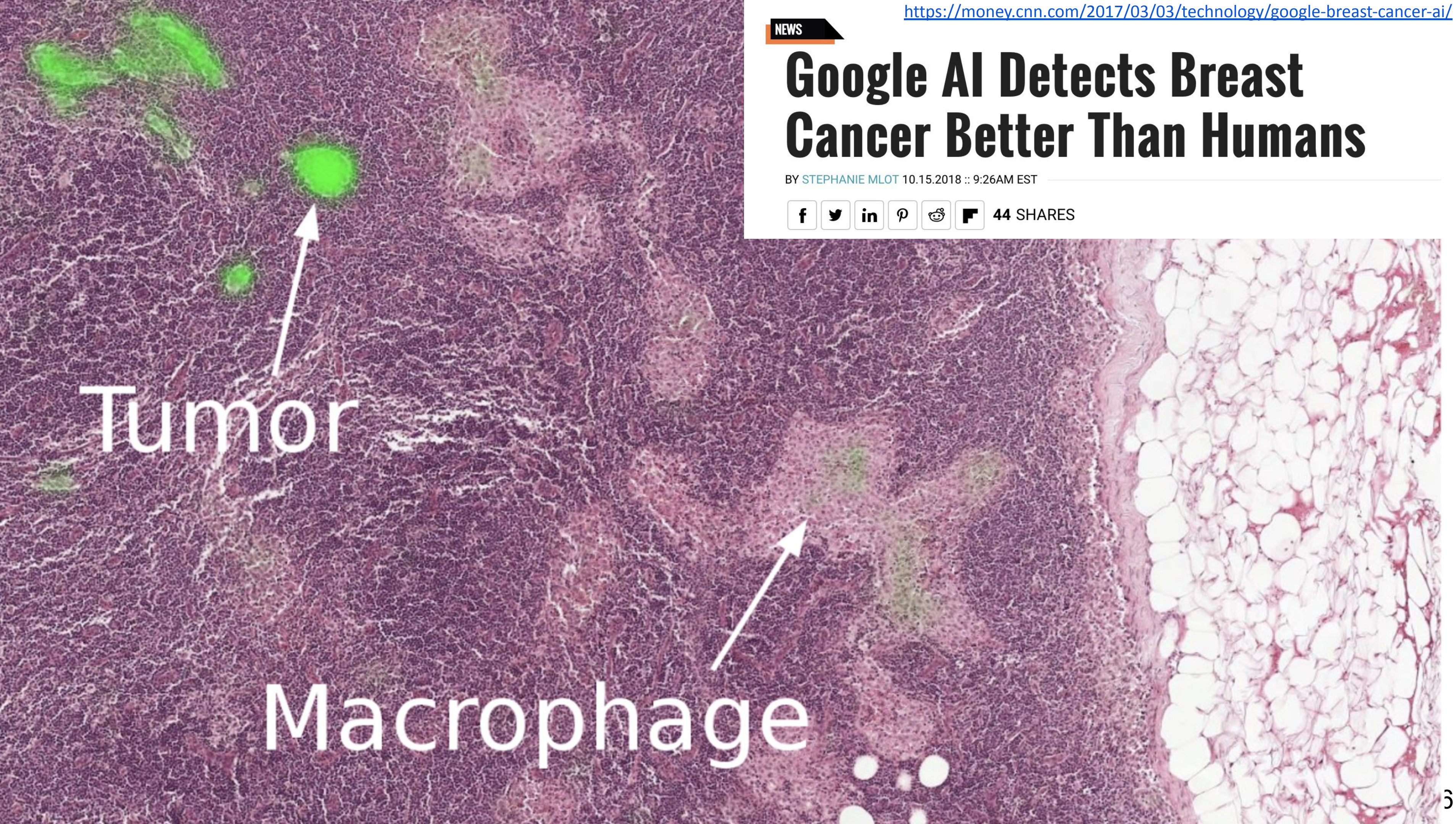
https://money.cnn.com/2017/03/03/technology/google-breast-cancer-ai/

NEWS

# Google AI Detects Breast Cancer Better Than Humans

BY STEPHANIE MLOT 10.15.2018 :: 9:26AM EST

44 SHARES



Science

# AlphaGo Zero: Google DeepMind supercomputer learns 3,000 years of human knowledge in 40 days

share

17

AlphaGo


http://www.telegraph.co.uk/science/2017/10/18/alphago-zero-google-deepmind-supercomputer-learns-3000-years/

https://www.christies.com/features/A-collaboration-between-two-artists-one-human-one-a-machine-9332-1.aspx

# Is artificial intelligence set to become art's next medium?

16 October 2018

PHOTOGRAPHS & PRINTS |

AI artwork sells for $432,500 — nearly 45 times its high estimate — as Christie's becomes the first auction house to offer

Humans and Technology

# Machine learning has been used to automatically translate long-lost languages

Some languages that have never been deciphered could be the next ones to get the machine translation treatment.

https://www.technologyreview.com/s/613899/machine-learning-has-been-used-to-automatically-translate-long-lost-languages/

by **Emerging Technology from the arXiv** Jul 1, 2019

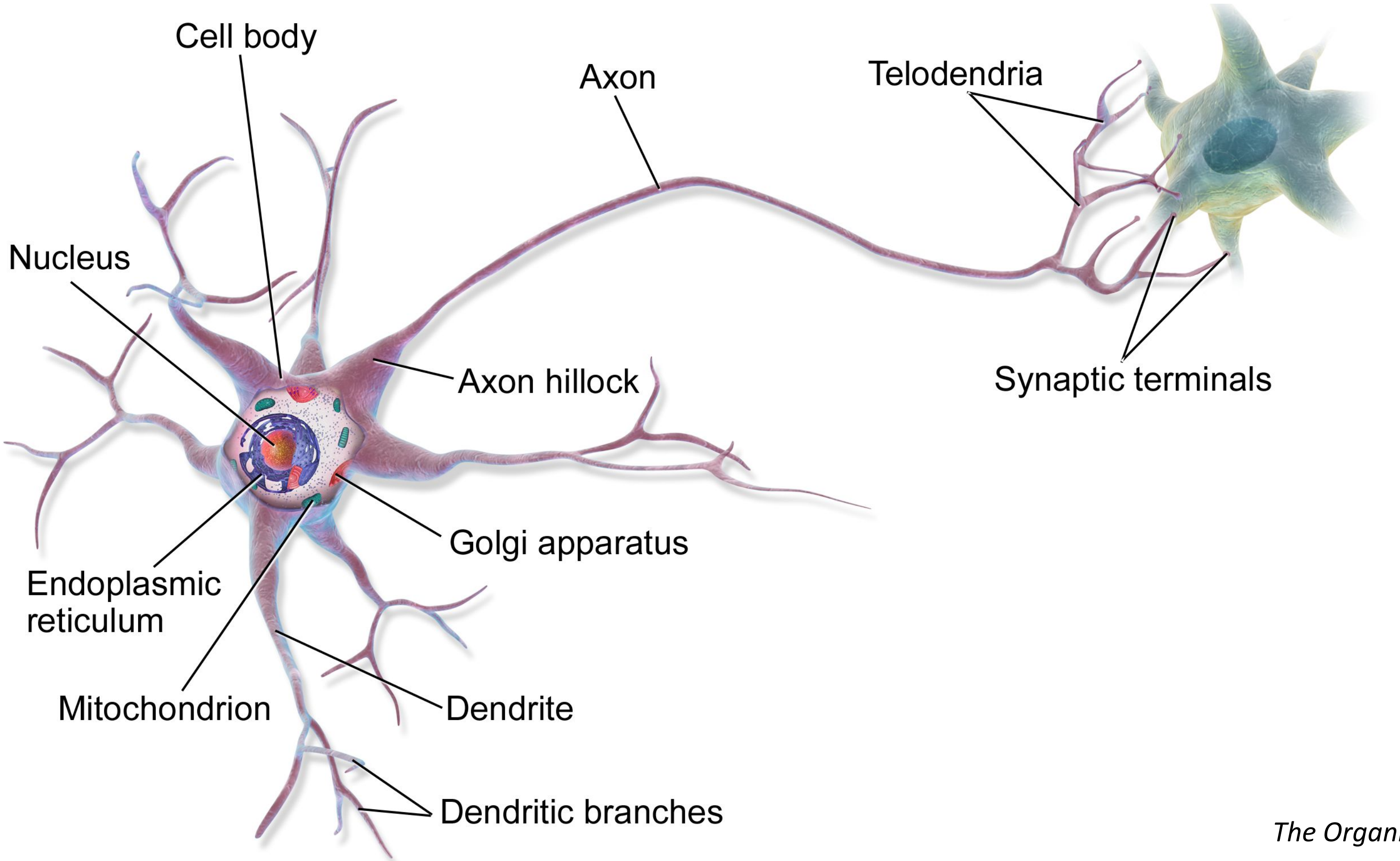
*"...in from three to eight years we will have a machine with the general intelligence of an average human being", Marvin Minsky (1970)*

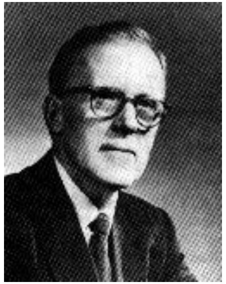
# Are we all doomed…?

...or do we simply have a tendency to overestimate technology?

# From Biological Neuron to the Artificial Neuron Model



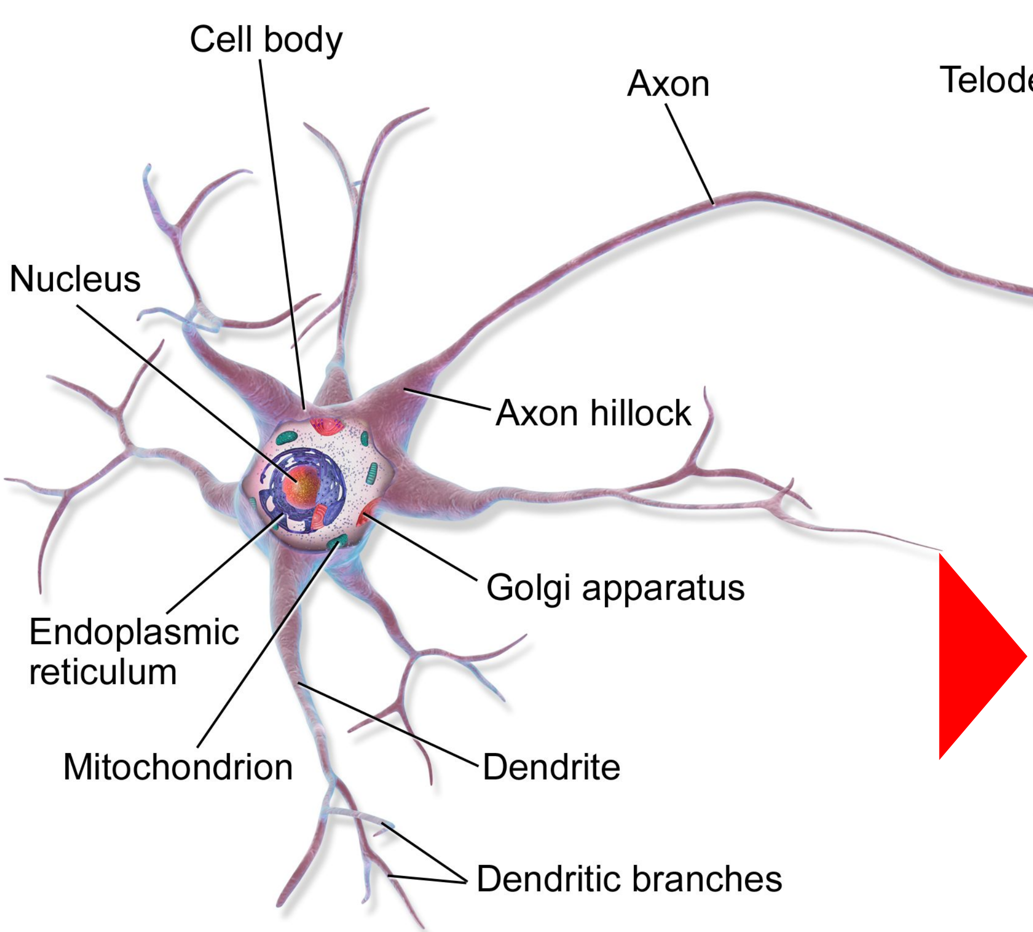


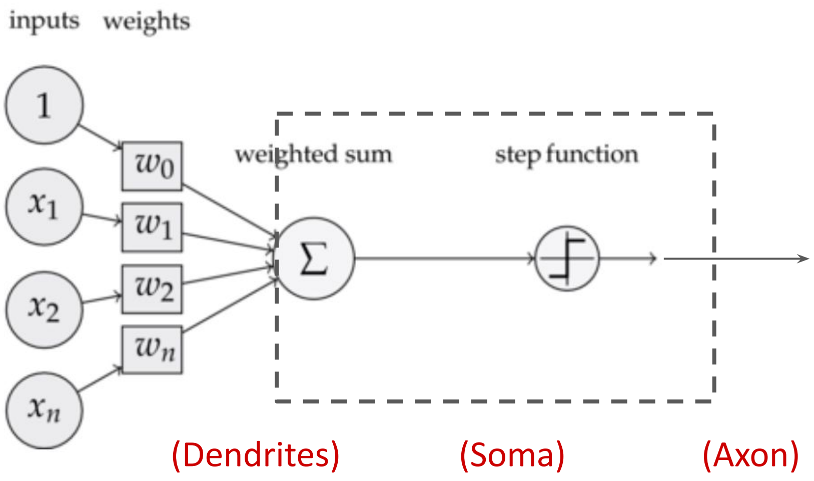
Donald Hebb
*The Organization of Behaviour* (1949)

# From Biological Neuron to the Artificial Neuron Model

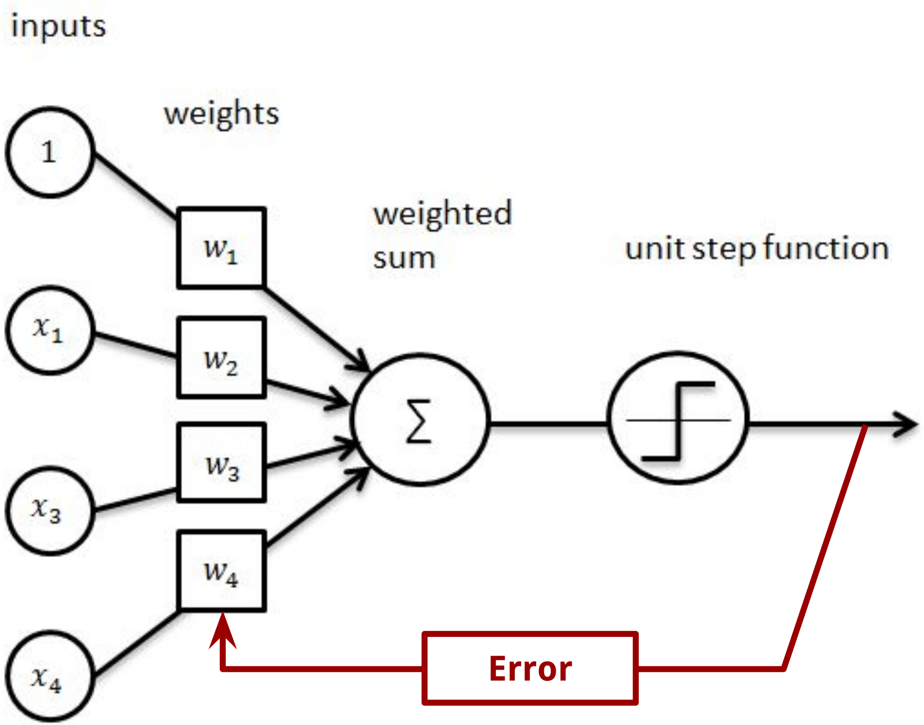





Donald Hebb (1940)
McCulloch & Pitts (1943)



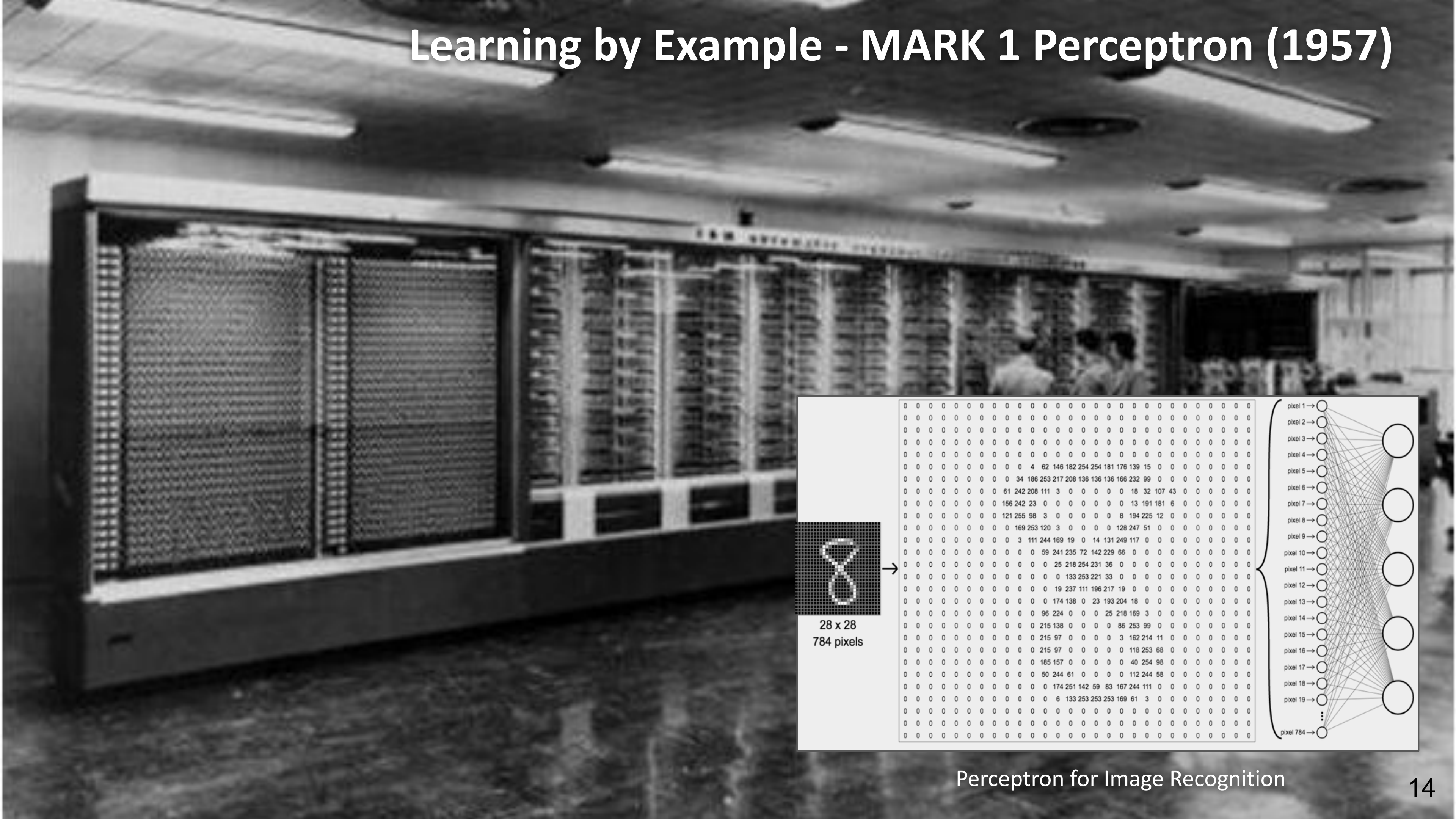
# Perceptron Algorithm




Frank Rosenblatt
*The perceptron: a probabilistic model for information storage and organization in the brain. (1958)*

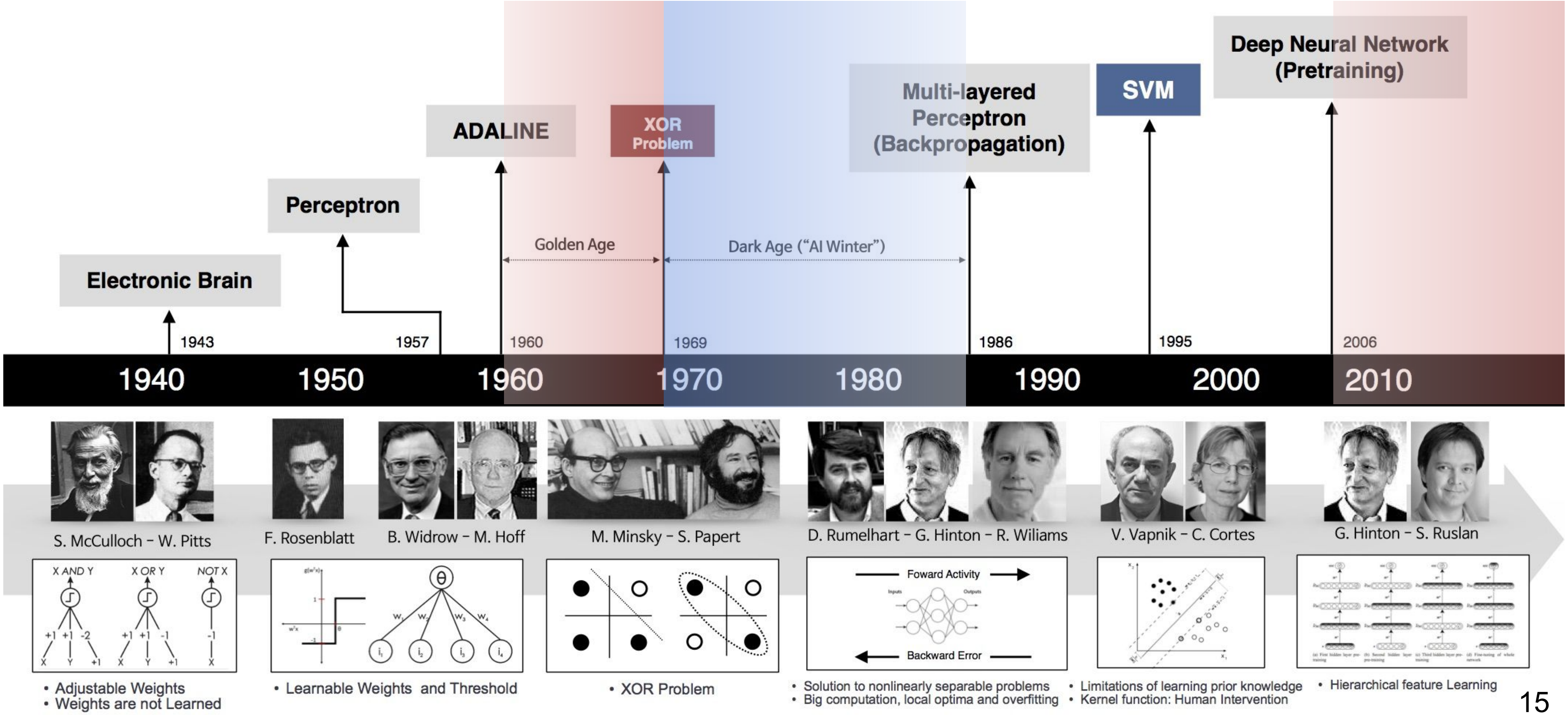
$$w_{ij}^{new} = w_{ij}^{old} + \Delta w_{ij}$$
$$\Delta w_{ij} = \alpha \cdot (t_j - o_j) \cdot x_i \,.$$

# Learning by Example - MARK 1 Perceptron (1957)

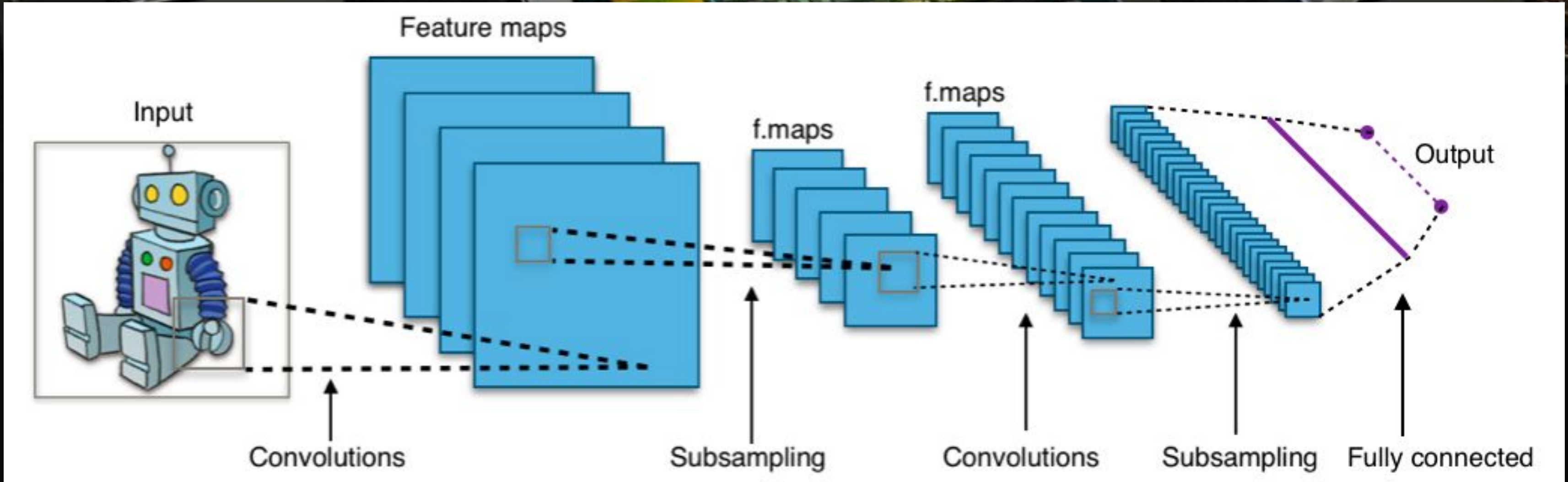



Perceptron for Image Recognition

# Machine Learning Timeline

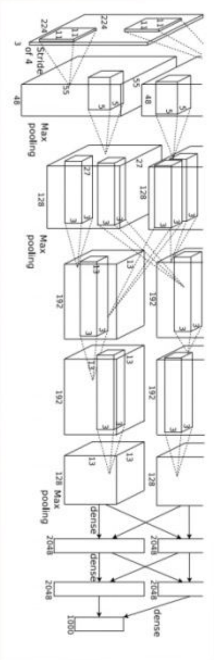


# Deep Convolutional Neural Networks on GPU Supercomputers



# Reusable Highly Complex Pre-Trained and Re-Usable Models

"AlexNet"



[Krizhevsky et al. NIPS 2012]

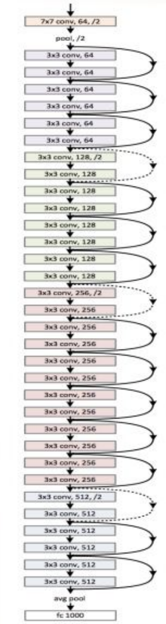
"GoogLeNet"

[Szegedy et al. CVPR 2015]

"VGG Net"


[Simonyan & Zisserman, ICLR 2015]

"ResNet"



[He et al. CVPR 2016]

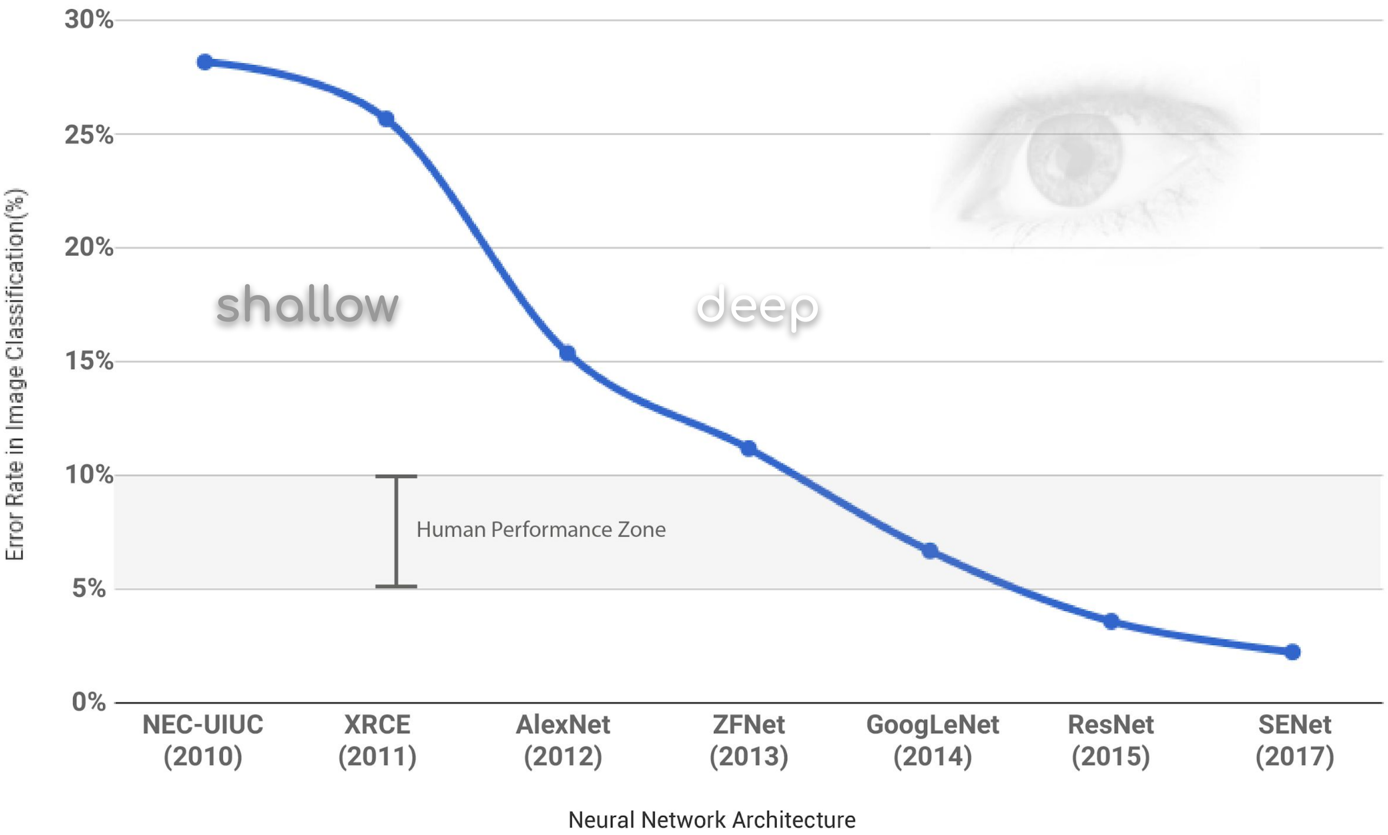
*"First, we find that the performance on vision tasks still increases linearly with orders of magnitude of training data size."*

*C. Sun at al, Revisiting Unreasonable Effectiveness of Data in Deep Learning Era, 2017*

# Availability of Large Annotated Training Data Sets

# The ImageNet Effect






IMAGENET

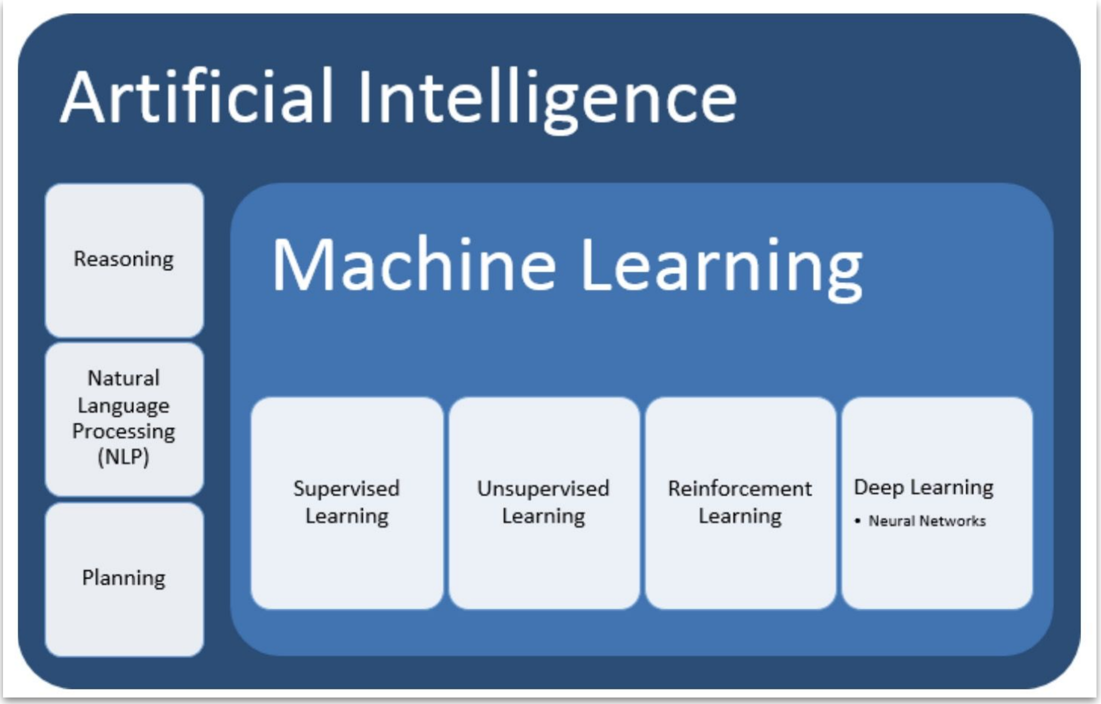
Large Scale Visual Recognition Challenge (ILSVRC)

http://image-net.org/challenges/LSVRC/

# What Deep Learning has achieved so far

- Near-human to superhuman level **image classification**
- Near-human level **speech recognition**
- Near-human level **handwriting transcription**
- Improved **machine translation**
- Improved **text-to-speech conversion**
- **Digital assistants**
- Near-human level **autonomous driving**
- Superhuman Go playing

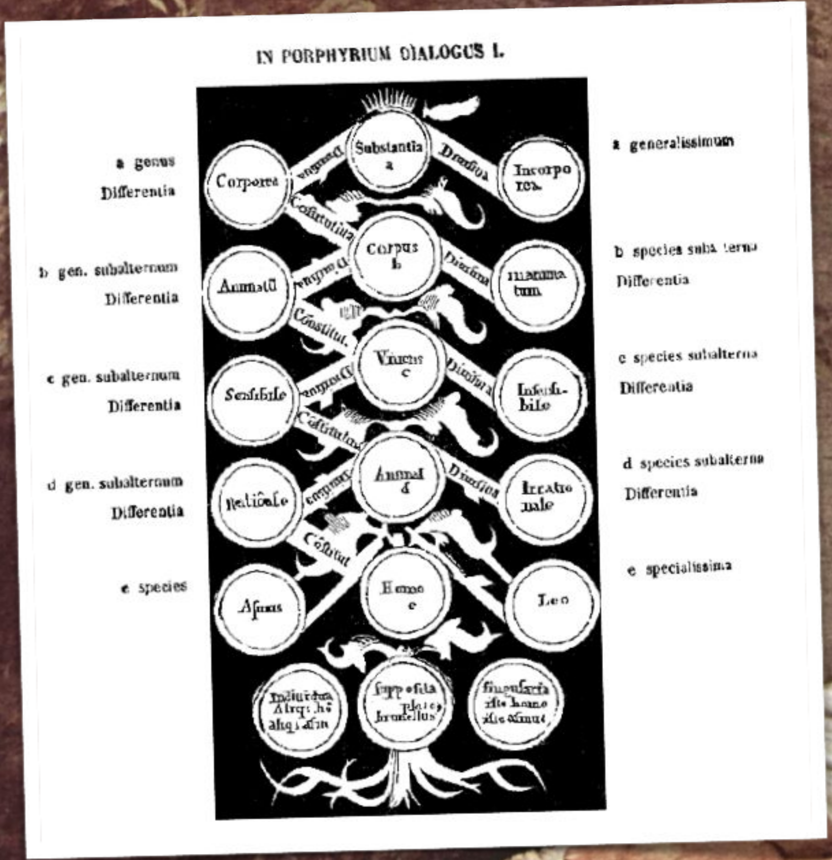
# Artificial Intelligence and Machine Learning




“The Goal of AI is to develop machines that behave as though they were intelligent.”

- John McCarthy (1955)

# The Universal Categories - Aristotle (384–322 BC)

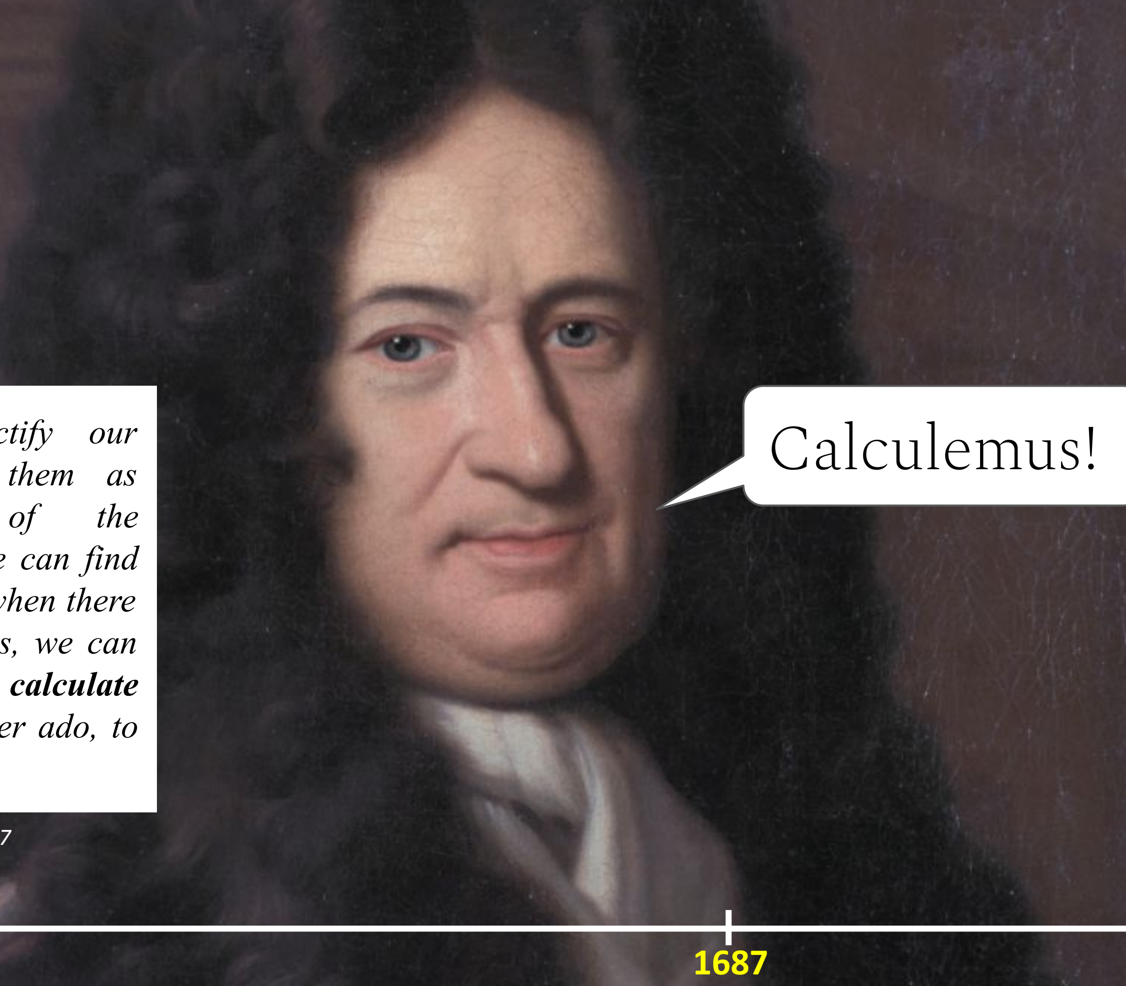



350 BCE

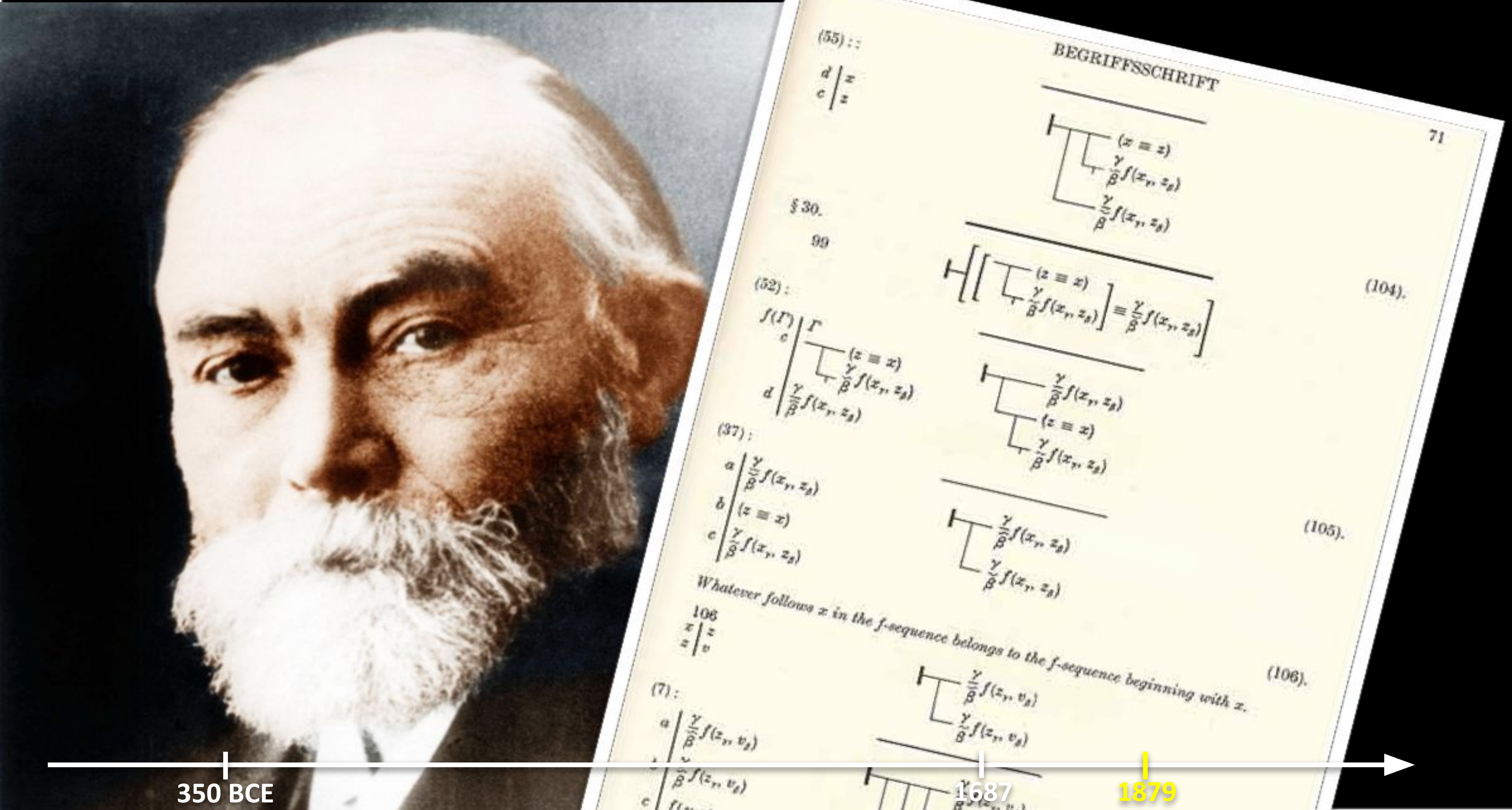
# Calculus Ratiocinator - Gottfried Wilhelm Leibniz (1646-1716)

*"The only way to rectify our reasonings is to make them as tangible as those of the Mathematicians, so that we can find our error at a glance, and when there are disputes among persons, we can simply say:* ***Let us calculate*** *[calculemus], without further ado, to see who is right.*

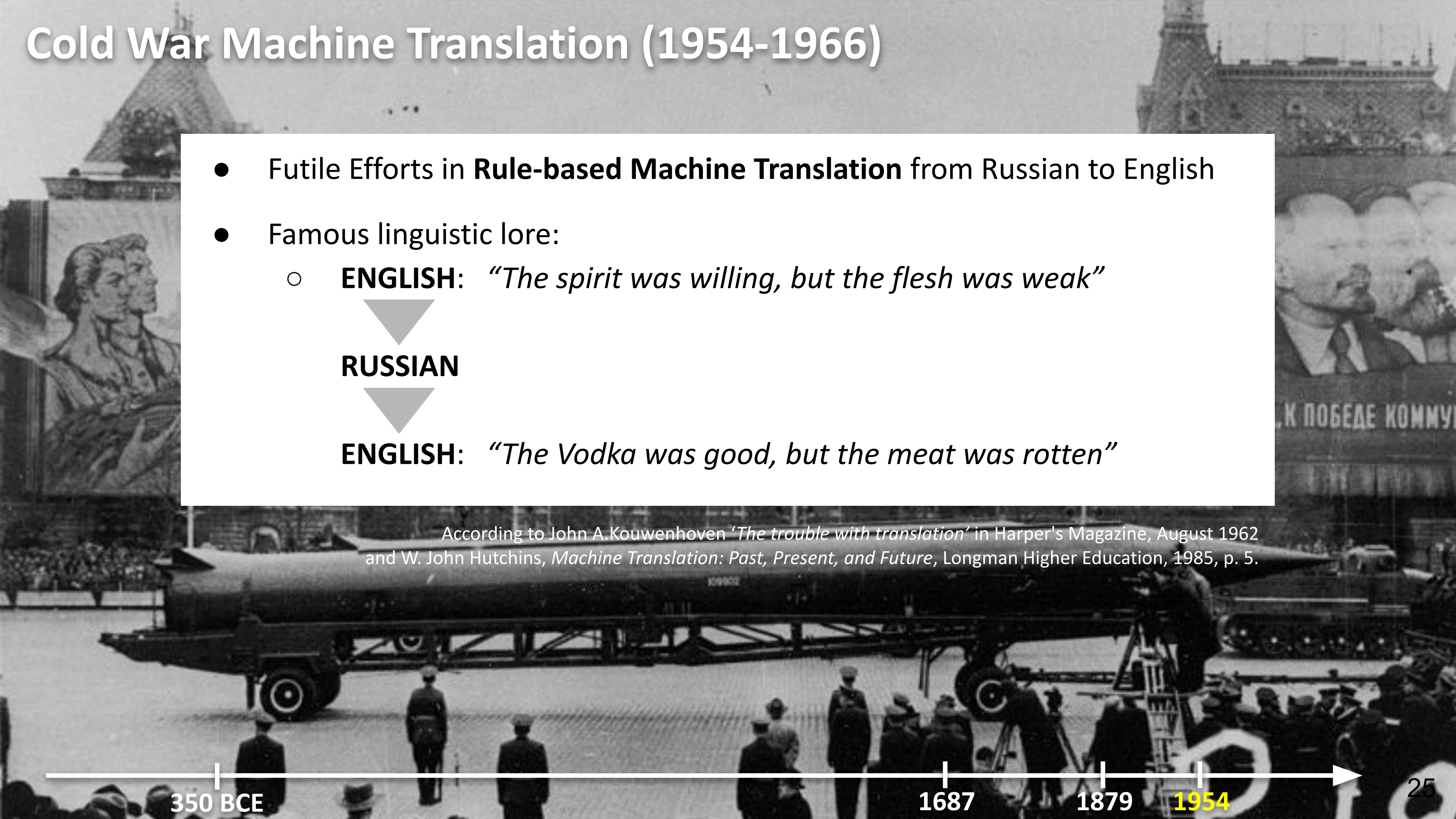
*Leibniz in a letter to Ph. J. Spener, Juli 1687*



# Begriffsschrift - Gottlob Frege (1848-1925)



# Cold War Machine Translation (1954-1966)




- Futile Efforts in **Rule-based Machine Translation** from Russian to English
- Famous linguistic lore:
  - **ENGLISH**: *"The spirit was willing, but the flesh was weak"*

    ▼

    **RUSSIAN**

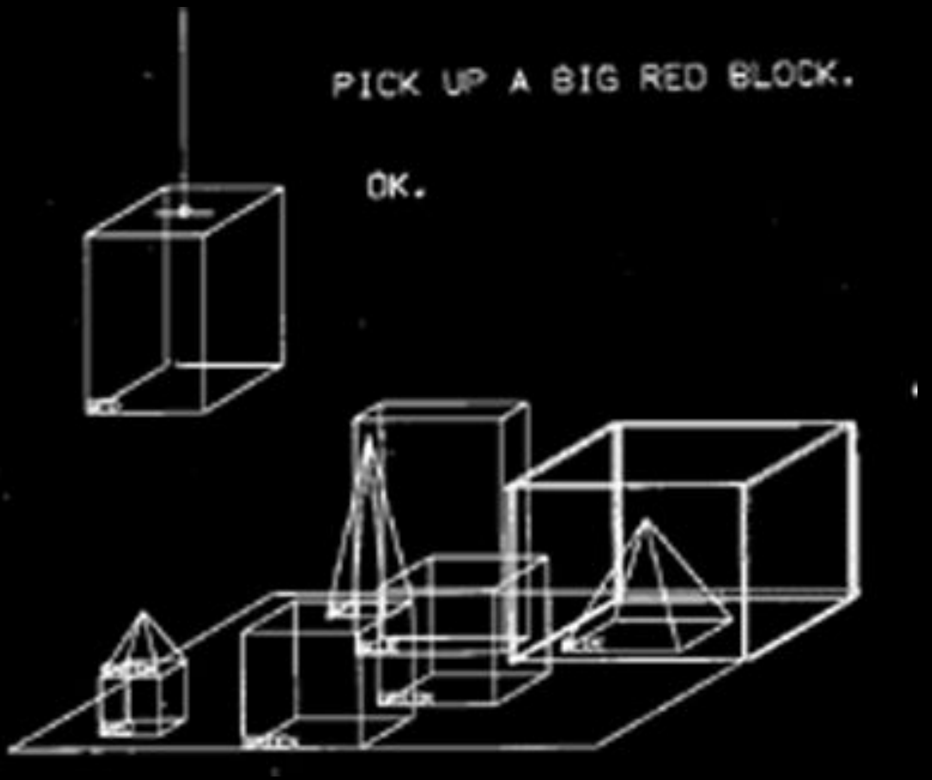
    ▼

    **ENGLISH**: *"The Vodka was good, but the meat was rotten"*

According to John A.Kouwenhoven *'The trouble with translation'* in Harper's Magazine, August 1962
and W. John Hutchins, *Machine Translation: Past, Present, and Future*, Longman Higher Education, 1985, p. 5.

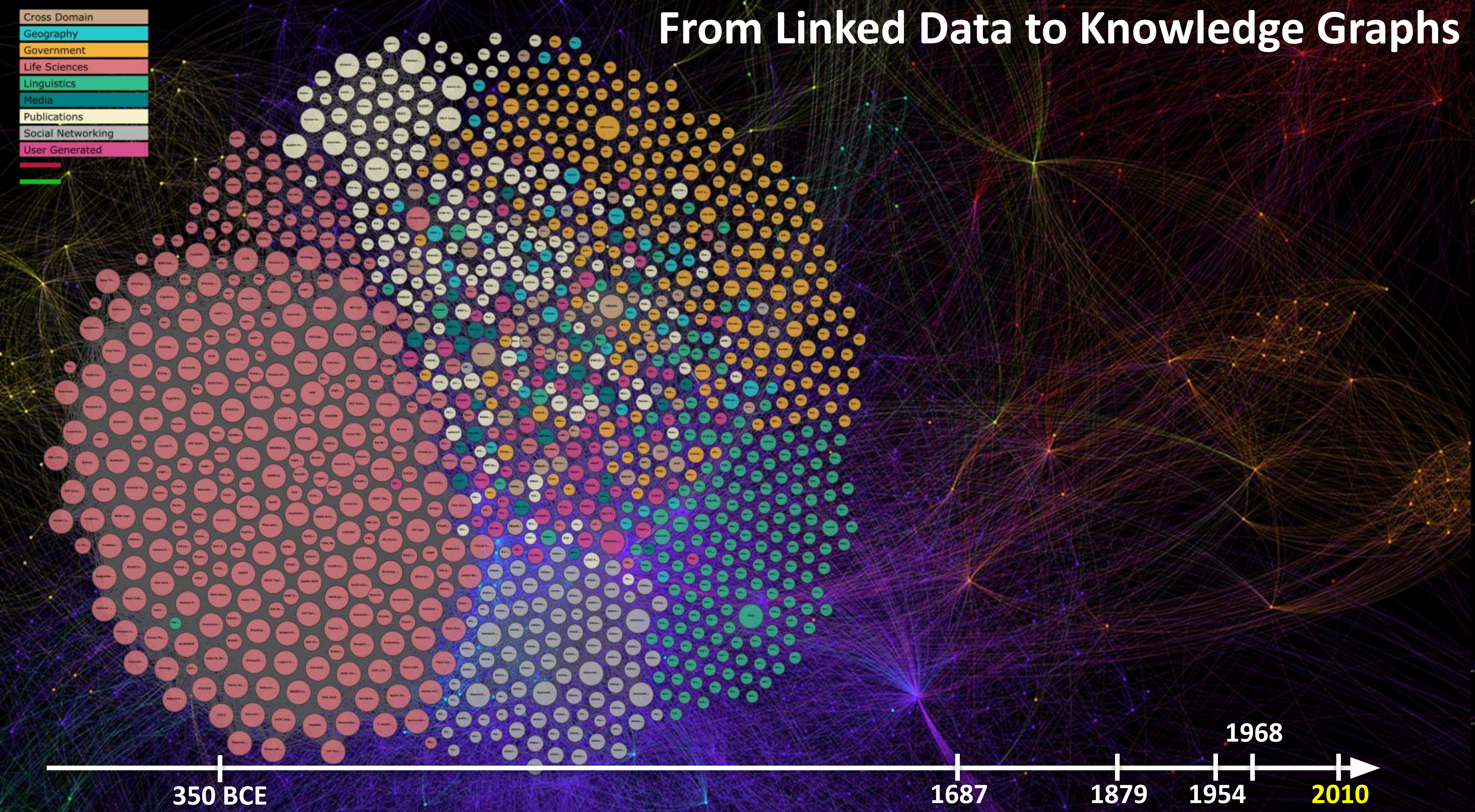
350 BCE — 1687 — 1879 — 1954

# Symbolic Manipulation to Rule the World

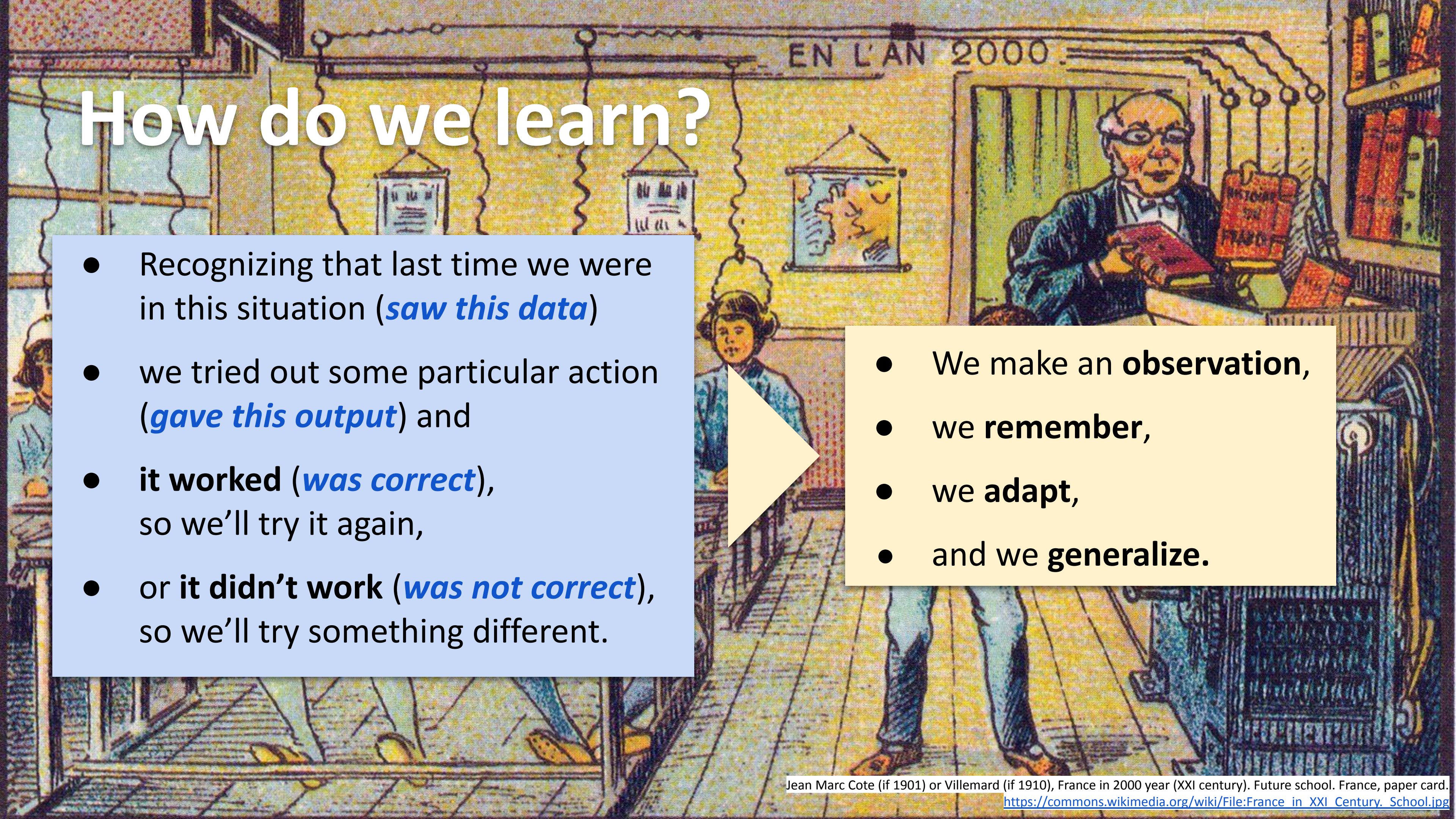



- **SHRDLU** by Terry Winograd (1968-1970)

# From Linked Data to Knowledge Graphs



# How do we learn?

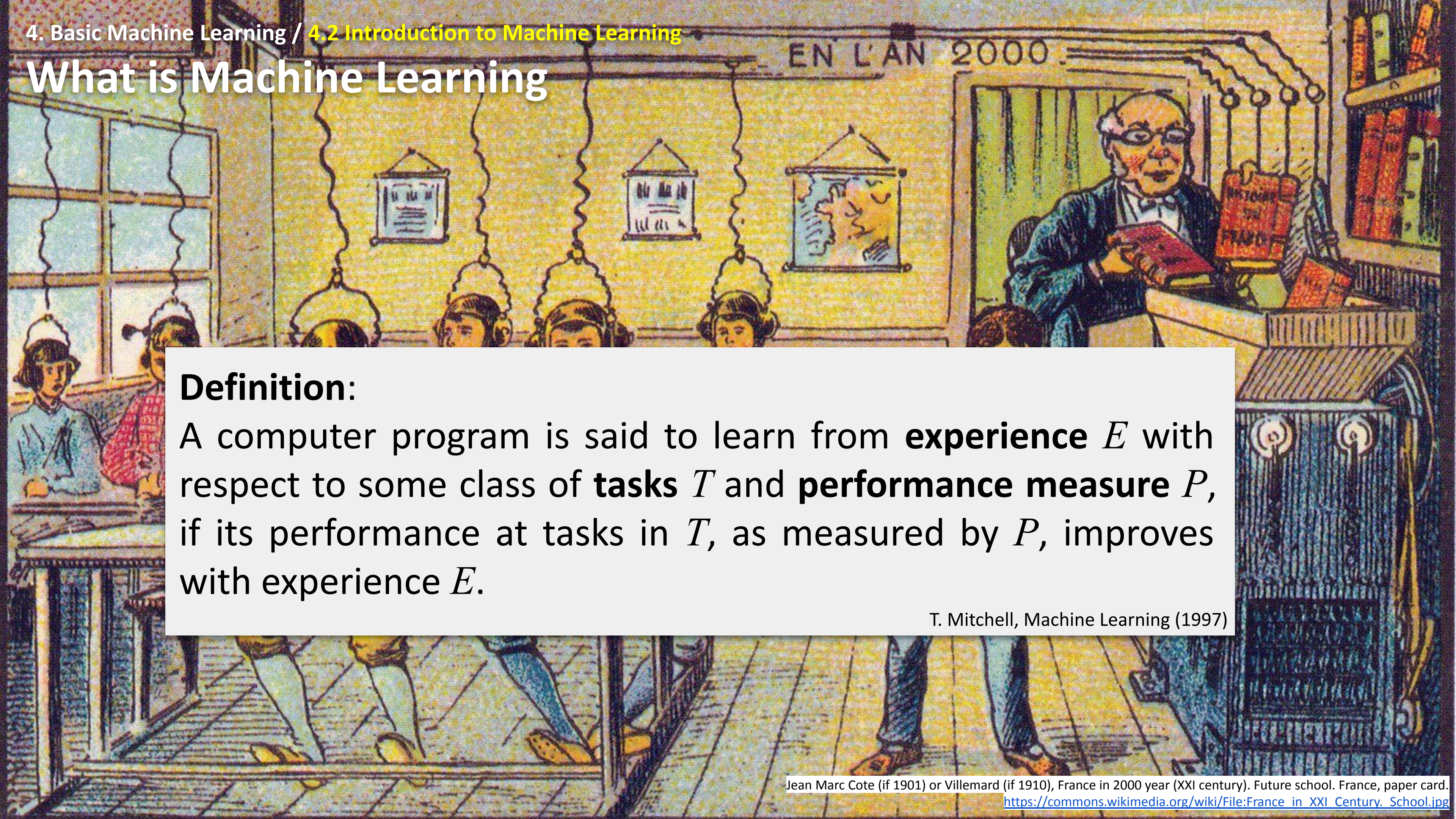



- Recognizing that last time we were in this situation (***saw this data***)
- we tried out some particular action (***gave this output***) and
- **it worked** (***was correct***), so we'll try it again,
- or **it didn't work** (***was not correct***), so we'll try something different.

- We make an **observation**,
- we **remember**,
- we **adapt**,
- and we **generalize.**

Jean Marc Cote (if 1901) or Villemard (if 1910), France in 2000 year (XXI century). Future school. France, paper card.
https://commons.wikimedia.org/wiki/File:France_in_XXI_Century._School.jpg

# What is Machine Learning

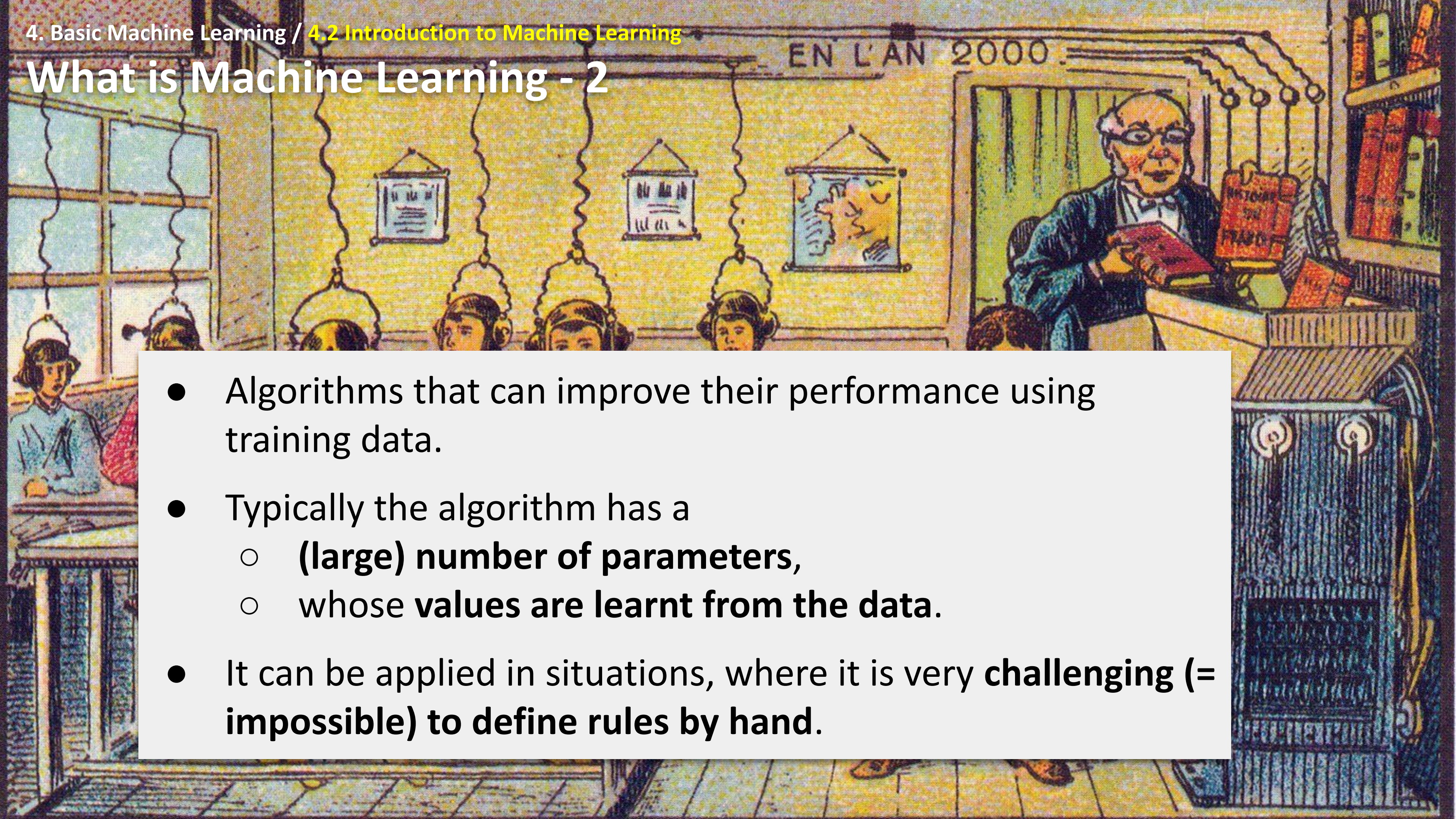



**Definition**:
A computer program is said to learn from **experience** $E$ with respect to some class of **tasks** $T$ and **performance measure** $P$, if its performance at tasks in $T$, as measured by $P$, improves with experience $E$.

T. Mitchell, Machine Learning (1997)

Jean Marc Cote (if 1901) or Villemard (if 1910), France in 2000 year (XXI century). Future school. France, paper card.
https://commons.wikimedia.org/wiki/File:France_in_XXI_Century._School.jpg

# What is Machine Learning - 2



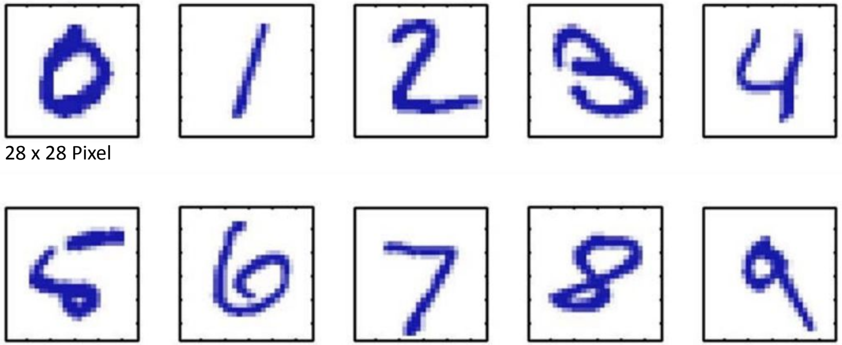

- Algorithms that can improve their performance using training data.
- Typically the algorithm has a
  - **(large) number of parameters**,
  - whose **values are learnt from the data**.
- It can be applied in situations, where it is very **challenging (= impossible) to define rules by hand**.

# Example Problem Formulation

## Handwritten Digit Recognition

Assign the correct value to a handwritten digit.

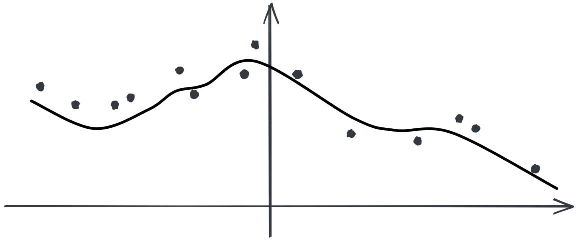
- Represent the **input image as a vector** $x \in R^{784}$
- Learn a **classifier *y=f(x)*** such that, $f : x \rightarrow \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$



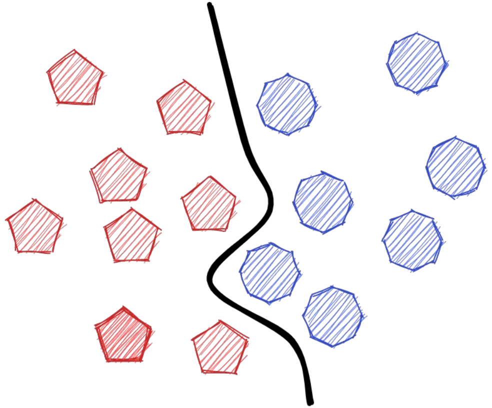
# Regression vs. Classification Problems



- Problems with a **quantitative response** most times are considered as **regression problems**.

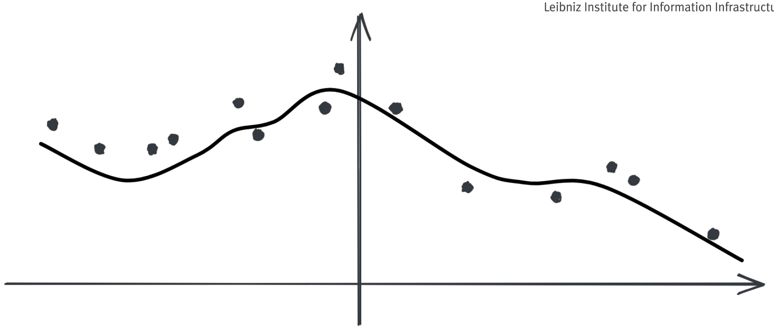


- Problems involving a **qualitative response** are often referred to as **classification problems.**

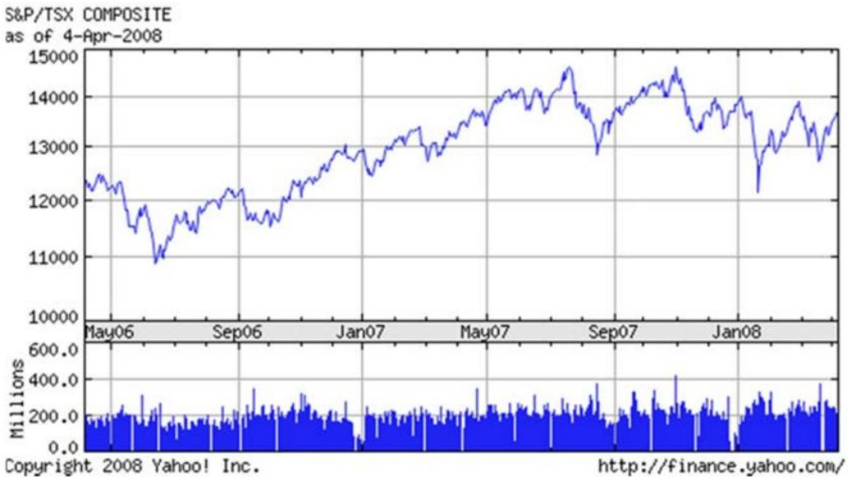


# Regression Problems

- The goal of **Regression** is to estimate a real-valued variable $y \in \mathbb{R}$ given a pattern $x$.

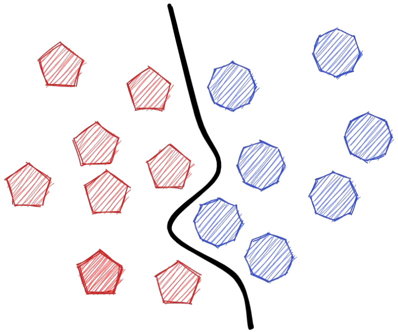


- Example:
  **Prediction of stock prices** for a future date
  or
  **Prediction of population numbers**
  for a future date

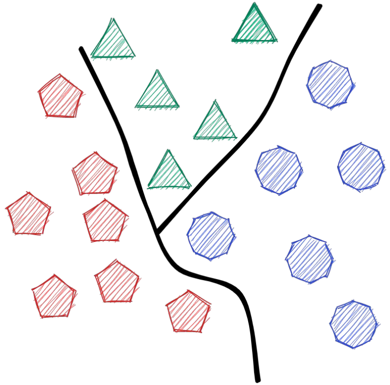


# Classification Problems

- **Binary Classification**
  - given a pattern $x$ drawn from a domain $X$, estimate which value an associated binary random variable $y \in \{\pm 1\}$ will assume.

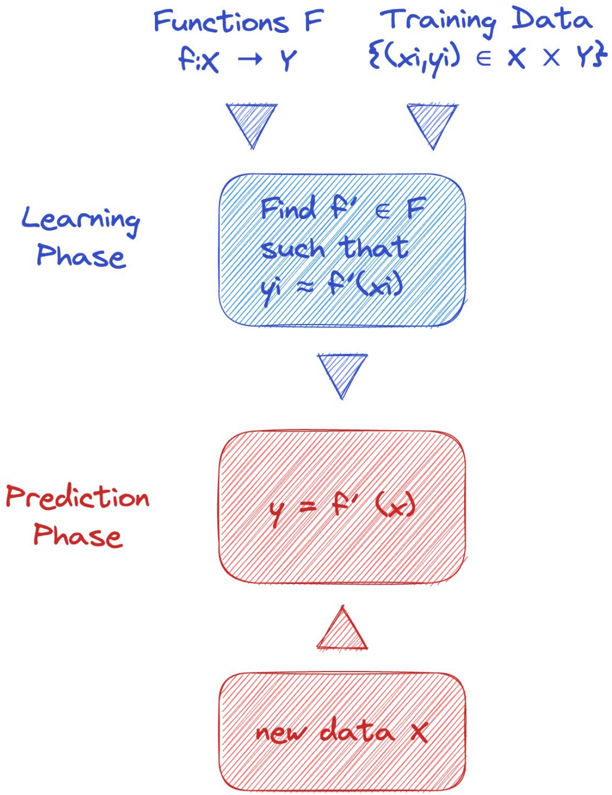


- **Multi Class Classification**
  - given a pattern $x$ drawn from a domain $X$, estimate which value an associated binary random variable $y \in \{1,...,n\}$ will assume.

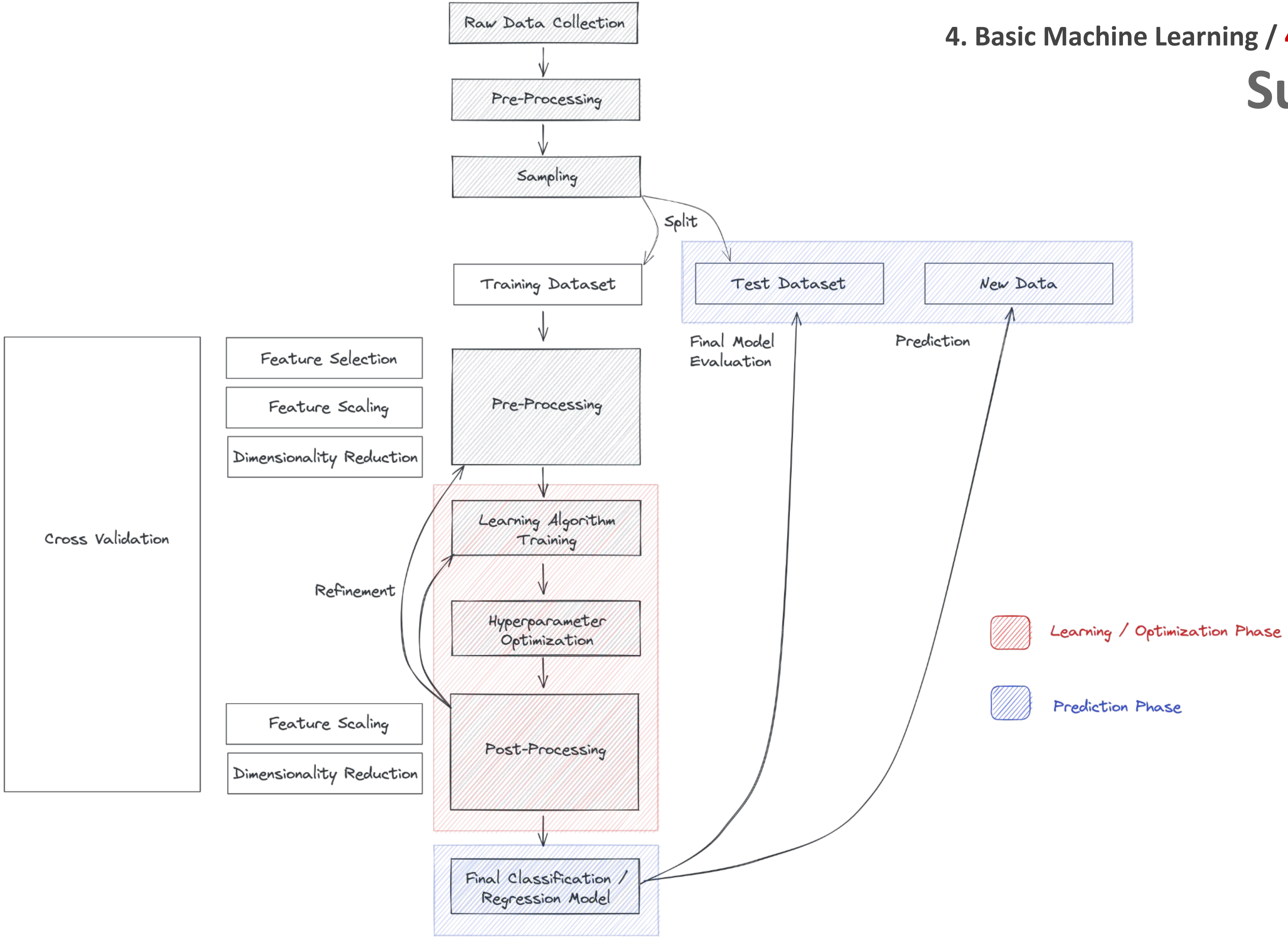


# Supervised Learning

- A **training set** of examples with the **correct responses** (targets) is provided and, based on this training set,
- the **algorithm generalizes to respond correctly to all possible inputs**.
- Typically, each **example** is a pair consisting of
  - an **input object** (typically a vector) and
  - a desired **output value**
    (also called the supervisory signal).

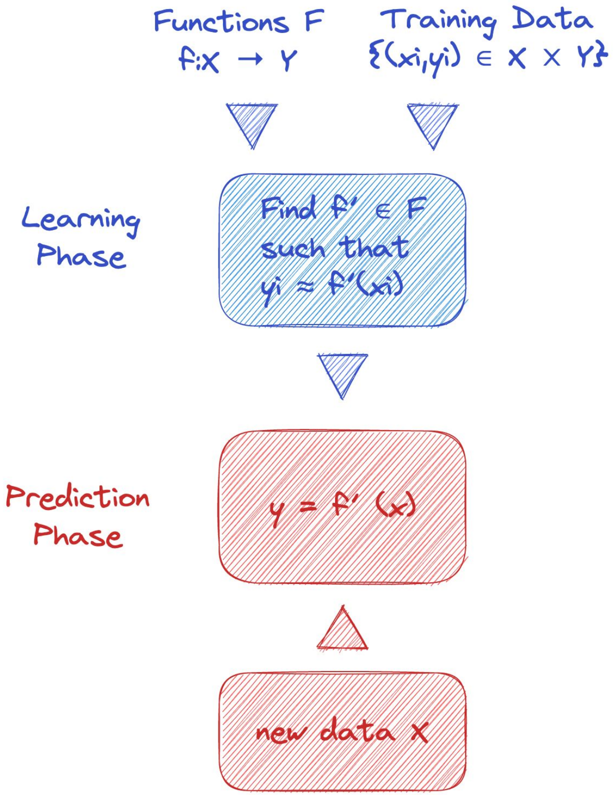


# Supervised Learning



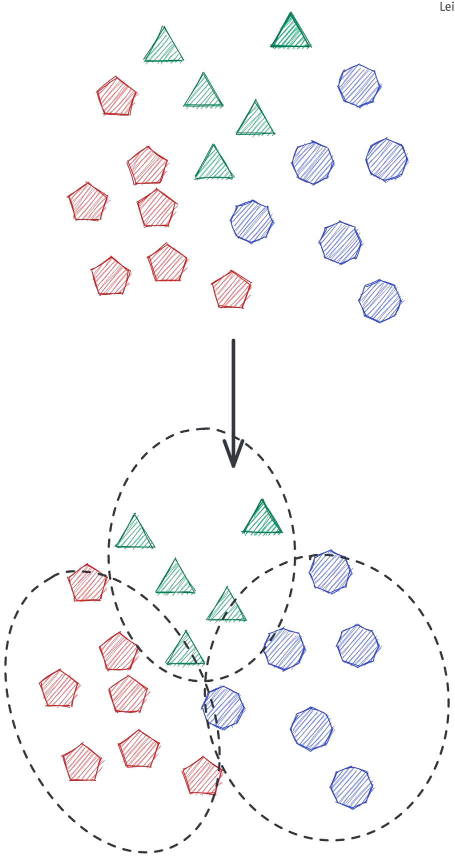
# Supervised Learning - Examples

- k-Nearest Neighbors
- **Linear Regression**
- Logistic Regression
- Support Vector Machines (SVM)
- **Decision Trees**
- **Neural Networks**



# Unsupervised Learning

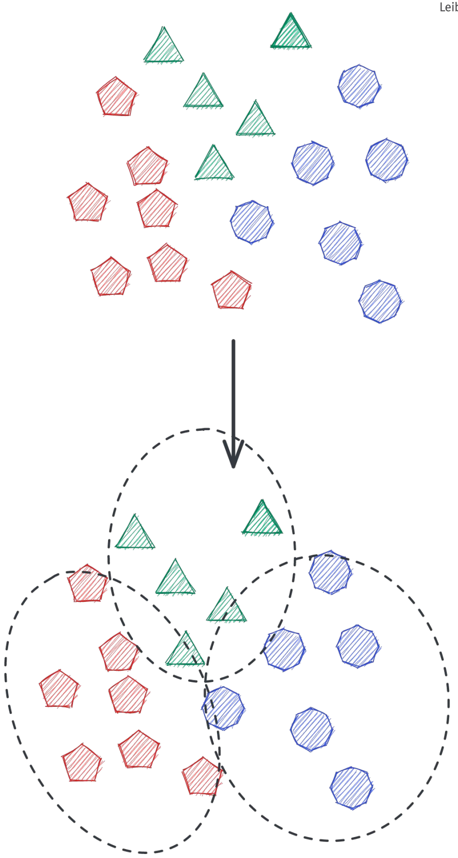
- Inferring a function to describe hidden structure from "unlabeled" data.
- **Correct responses are not provided**, but instead
- the algorithm tries to **identify similarities between the inputs** so that inputs that have something in common are categorized together.
- The statistical approach to unsupervised learning is known as **density estimation**.



# Unsupervised Learning - Examples

## Clustering Algorithms

- **k-Means**
- Hierarchical Cluster Analysis (HCA)
- Expectation Maximization



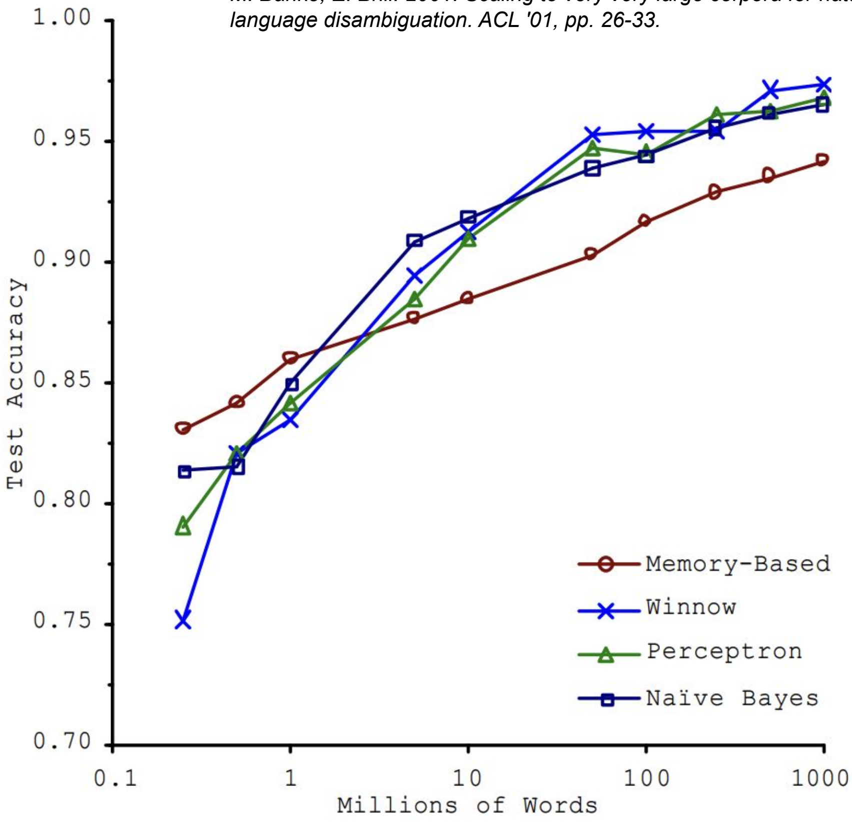
Karlsruher Institut für Technologie

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

# Main Challenges of Machine Learning

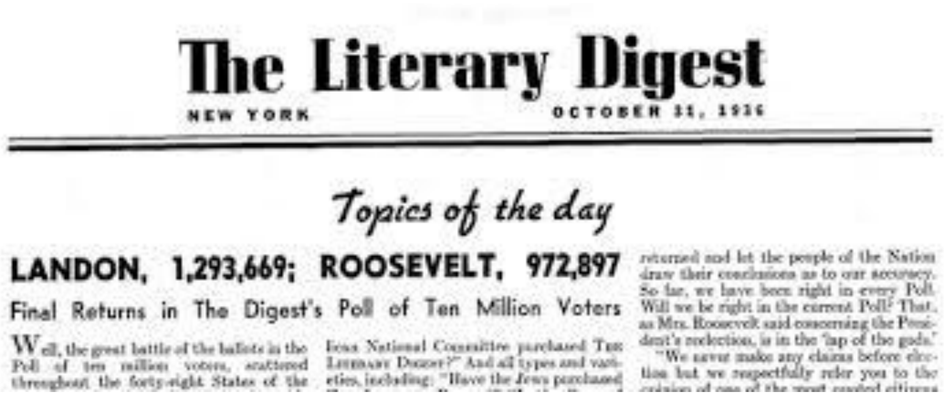
- **Insufficient Quantity of Training Data**

*M. Banko, E. Brill. 2001. Scaling to very very large corpora for natural language disambiguation. ACL '01, pp. 26-33.*



# Main Challenges of Machine Learning

- Insufficient Quantity of Training Data
- **Nonrepresentative Training Data**


The Literary Digest

NEW YORK OCTOBER 31, 1936

Topics of the day

LANDON, 1,293,669; ROOSEVELT, 972,897

Final Returns in The Digest's Poll of Ten Million Voters

**In 1936** The *Literary Digest* **wrongly predicted** the win of Landon with 57% of the votes against Roosevelt. In the poll they used telephone directories, list of magazine subscribers and club membership lists. **Roosevelt won** with 62% of the votes.

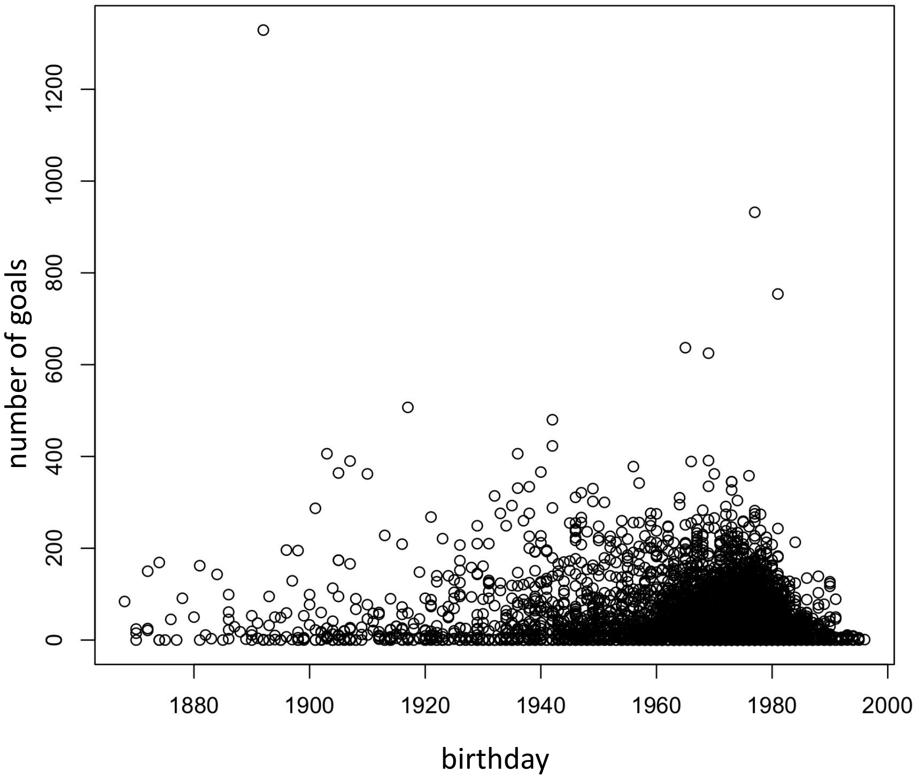
# Main Challenges of Machine Learning

- Insufficient Quantity of Training Data
- Nonrepresentative Training Data
- **Poor-Quality Data**

White Noise, https://commons.wikimedia.org/wiki/File:White-noise-mv255-240x180.png

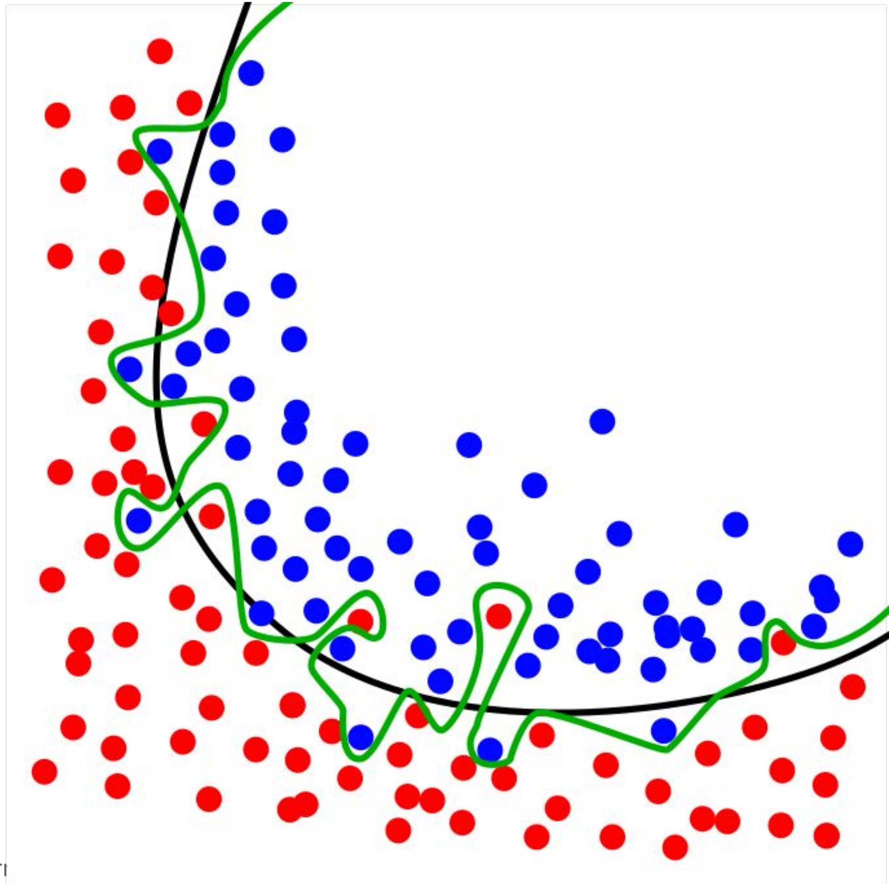
# Main Challenges of Machine Learning

- Insufficient Quantity of Training Data
- Nonrepresentative Training Data
- Poor-Quality Data
- **Irrelevant Features**
    - Problem can be addressed via **Feature Engineering**:
        - Feature selection
        - Feature extraction
        - Creating new features



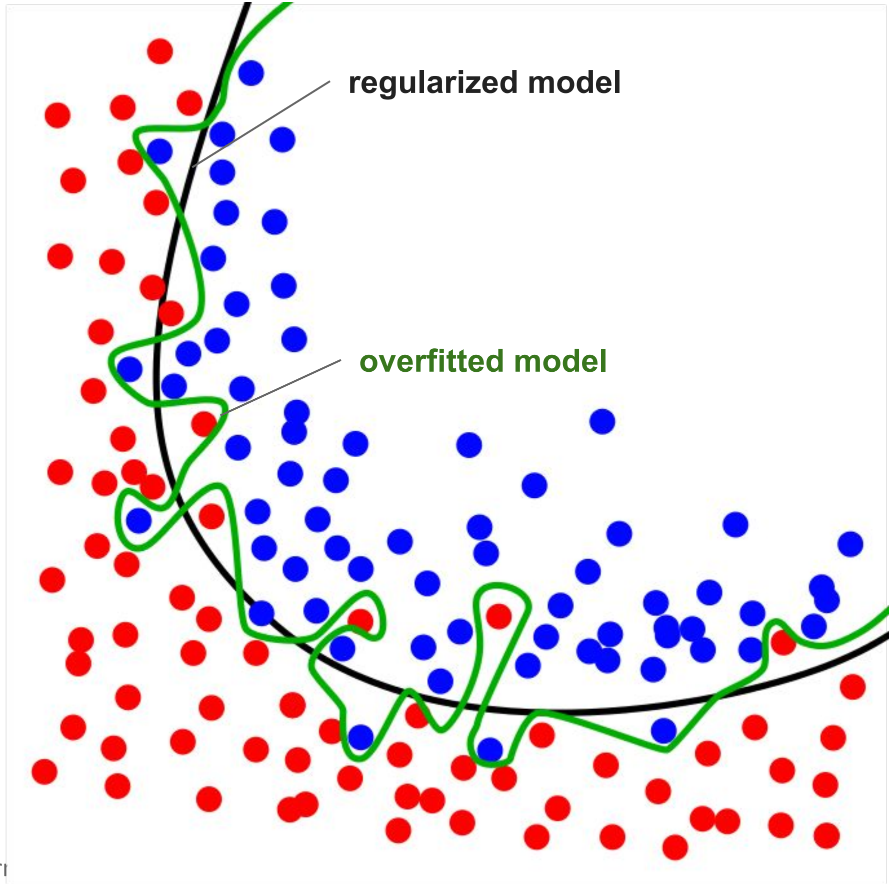
# Main Challenges of Machine Learning

- Insufficient Quantity of Training Data
- Nonrepresentative Training Data
- Poor-Quality Data
- Irrelevant Features
- **Overfitting the Training Data**



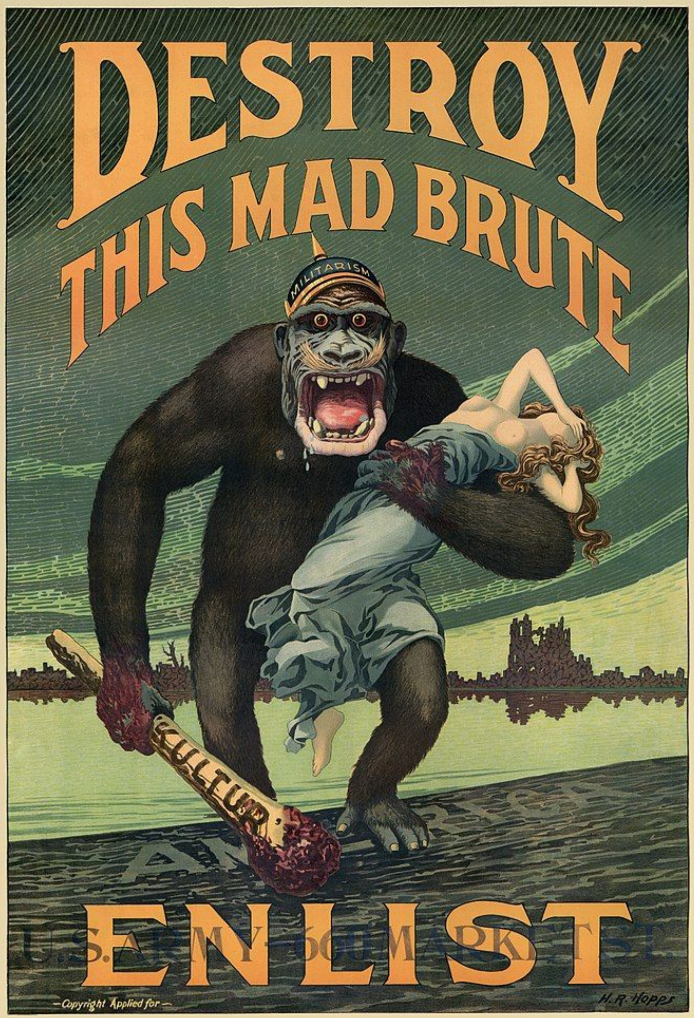
# Overfitting

- The **overfitted model** follows exactly the training data.
  - Too high dependency on potential noise
  - Lack of generalization
- The **overfitted model** is likely to have a higher error rate on new unseen data, compared to the **regularized model**.
  - **Better generalization** to the underlying classification function required.
- **Consequence:**
  Stop training the model before the algorithm overfits.



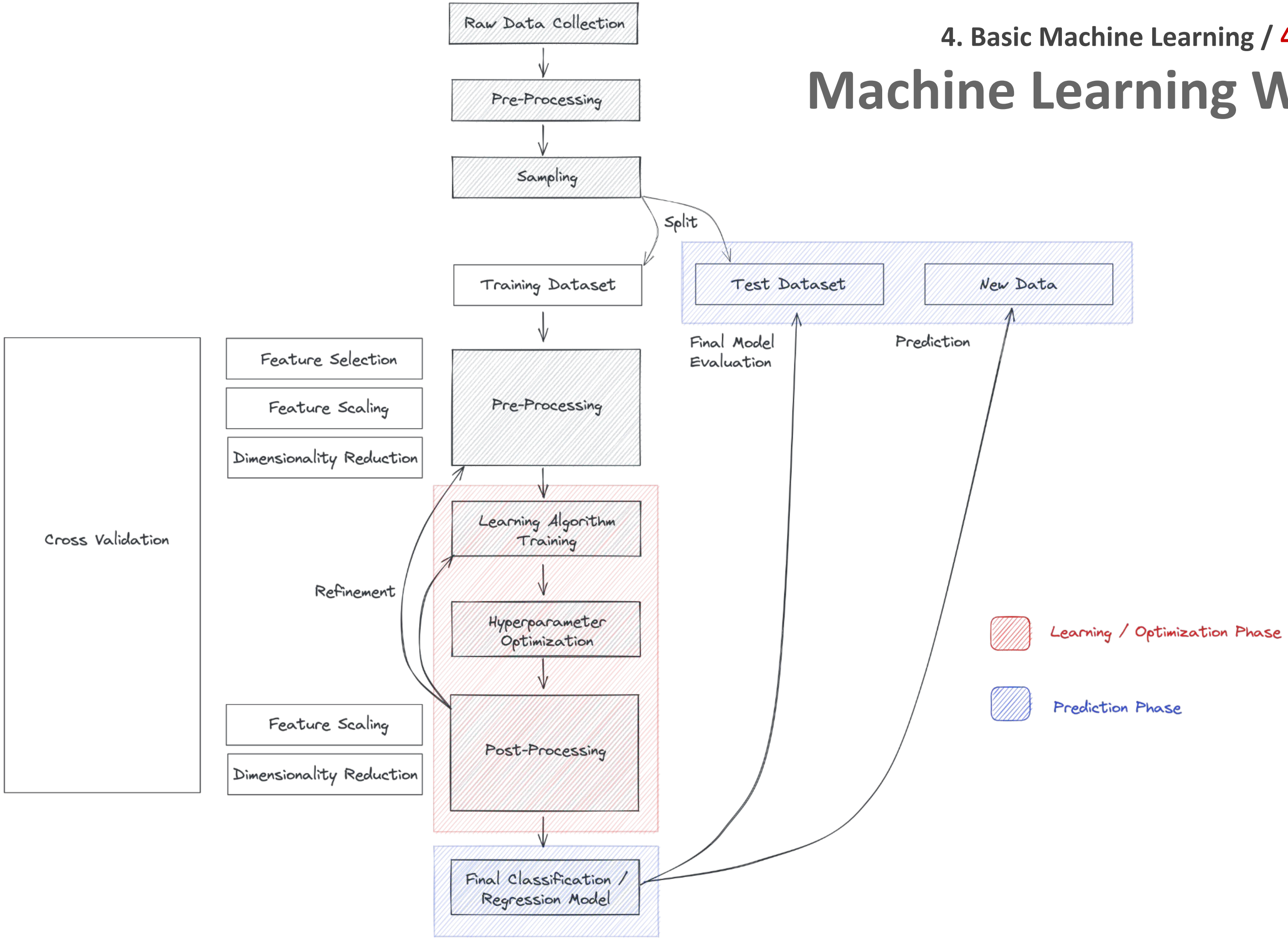
# There is no Bias Free Learning

- **Inductive bias:** Generalization learning is only possible, if the learning system has an inductive bias.
- **Restriction/language bias**: Not every model can be expressed by the given hypothesis language.
- **Preference/search bias**: Typically learning algorithms are based on a greedy search strategy in the hypothesis space. The bias directs search and influences which model is learned.
- **Sampling bias**: Independent of ML algorithm. How representative are the training data for the (infinite) set of all possible instances.



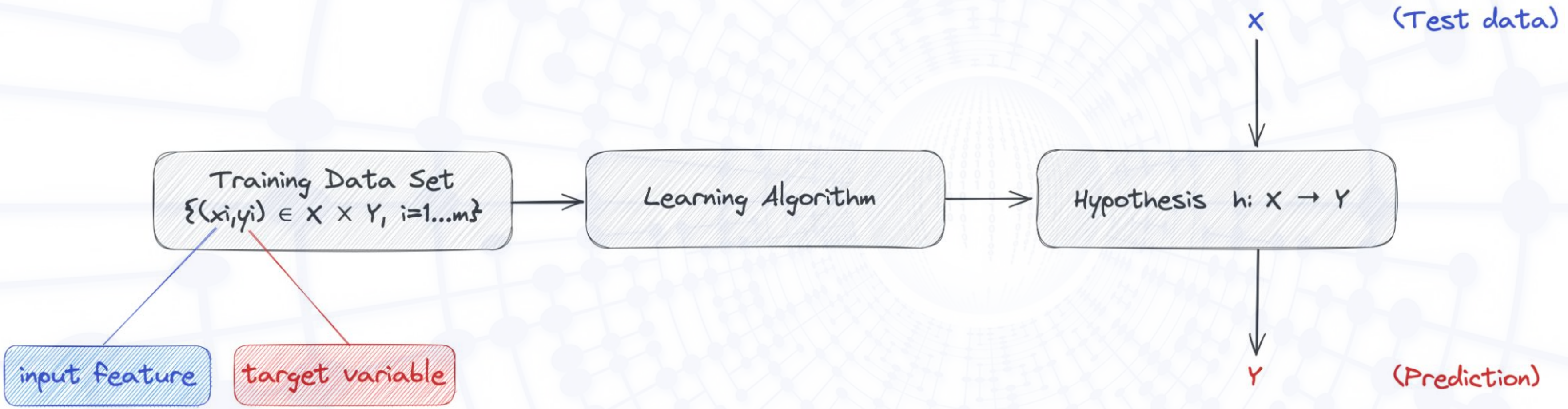

https://upload.wikimedia.org/wikipedia/commons/6/6e/Harry_R._Hopps%2C_Destroy_this_mad_brute_Enlist_-_U.S._Army%2C_03216u_edit.jpg

# Machine Learning Workflow Overview



# Data Collection

- **Raw Data Collection**
  - The **larger** and the **more diverse** the collected data, the better the learning task can be performed.

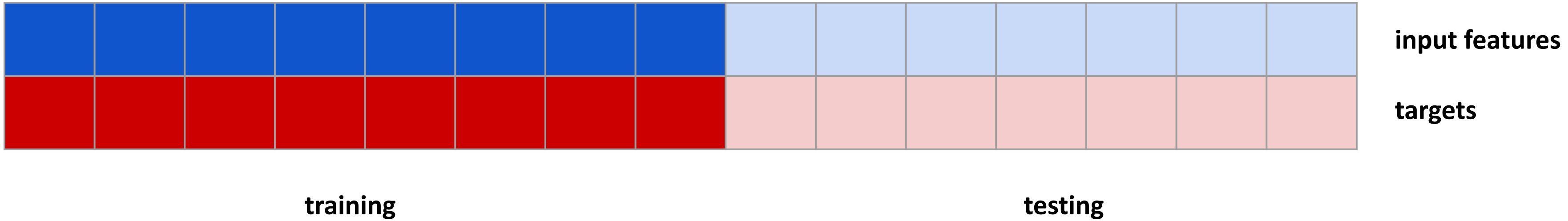


# Data Preprocessing and Data Cleaning

- **Data Preprocessing**
  - Typically to **create suitable training data**, the collected raw data has to be preprocessed (cleaned up) to remove errors, as e.g., dummy values, absence of data, contradicting data, etc.
- **Data Cleaning Steps**
  - **Parsing**: locates and identifies individual data elements in raw data.
  - **Correcting**: corrects parsed individual data components using sophisticated data algorithms.
  - **Normalization**: applies conversion routines to transform data into standard formats.
  - **Matching**: searching and matching records within and across data based on predefined rules.
  - **Consolidating**: merges data into one representation.

# Training Data and Test Data

- The easiest way to obtain training data is to split up the original dataset
  - Training Data (used to train the algorithm)
  - Test Data (used to evaluate the performance of the readily trained algorithm)

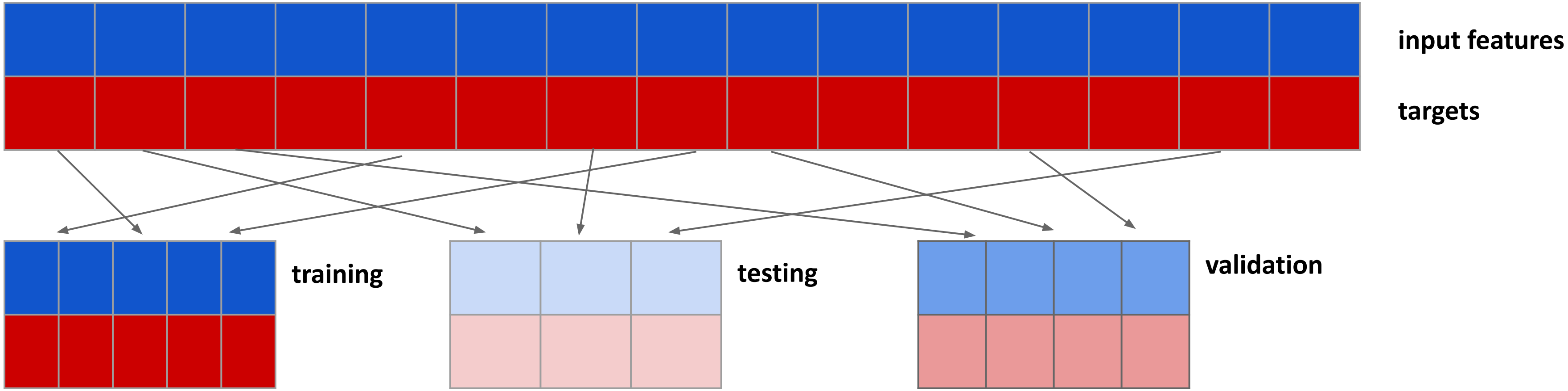



- BUT:
  - Evaluations obtained tend to reflect the particular way the data are divided up.
- SOLUTION:
  - **Statistical sampling** to get more accurate measurements.

# Sampling

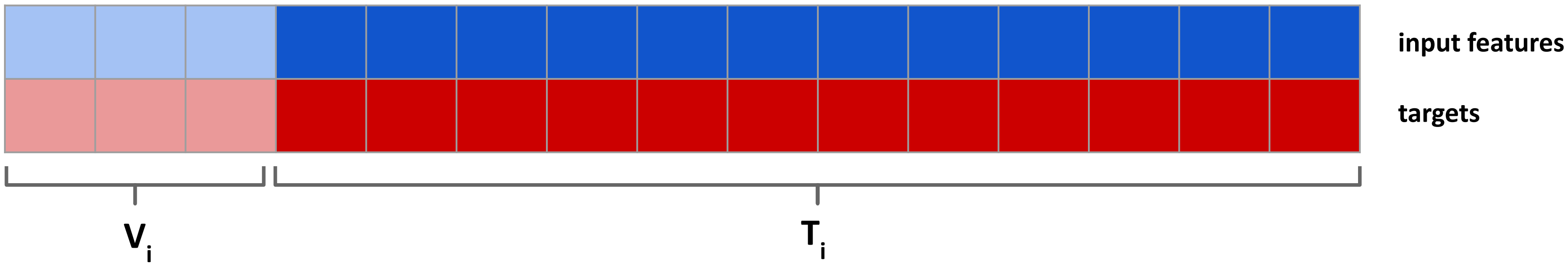


- The collected data should be divided into
  - Training Data (used to train the algorithm)
  - Validation Data (to keep track about the performance of the algorithm while it learns)
  - Test Data (used to evaluate the performance of the readily trained algorithm)



# K-fold Cross Validation

- The aim of **cross-validation** is to ensure that every example from the original dataset has the same chance of appearing in the training and testing set.
- In **K-fold cross-validation**, the dataset X is divided randomly into K equal-sized parts, $X_i$, i = 1,...,K.
- To generate each pair,
  - we keep **one of the K** parts out as the **validation set** $\mathbf{V_i = X_i}$ and
  - combine the **remaining K − 1** parts to form the training set $\mathbf{T_i = X_1 \cup \ldots \cup X_{i-1} \cup X_{i+1} \cup \ldots \cup X_K}$



# Feature Selection

- Select **attributes** of/from the available data that are **relevant to determine the projected outcome**.

- Simple Example: **SPAM Detection**
  - Input: emails x
  - Feature Vector:

$$f(x) = \begin{bmatrix} f(x_1) \\ f(x_2) \\ \dots \\ f(x_n) \end{bmatrix}, \text{ e.g., } f(x_i) = \begin{cases} 1 & \text{if the email contains "viagra"} \\ 0 & \text{otherwise} \end{cases}$$

**Training Data Set**
$\{(\mathbf{x}_i, \mathbf{y}_i) \in X \times Y, i=1,\dots,m\}$

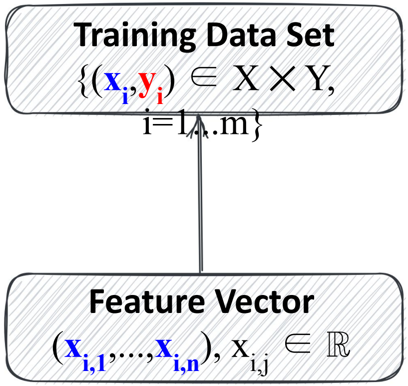
**Feature Vector**
$(\mathbf{x}_{i,1},\dots,\mathbf{x}_{i,n}),\ x_{i,j} \in \mathbb{R}$

# Feature Selection

- Select **attributes** of/from the available data that are **relevant to determine the projected outcome**.

- **Why?**
  - Avoid **overfitting** and achieve better **generalization ability**
  - Reduce the **storage requirement** and **training time**
  - **Interpretability**

- **Potential Difficulties:**
  - Irrelevant Attributes
  - Missing Attributes
  - Missing Attribute Values
  - Redundant Attributes
  - Attribute Value Noise



# Evaluation - Accuracy, Recall, Precision

- To evaluate the performance of a ML model, the following **Metrics** can be applied:

$$Accuracy = \frac{\#TP + \#TN}{\#TP + \#FP + \#TN + \#FN}$$

$$Recall = \frac{\#TP}{\#TP + \#FN}$$

$$Precision = \frac{\#TP}{\#TP + \#FP}$$

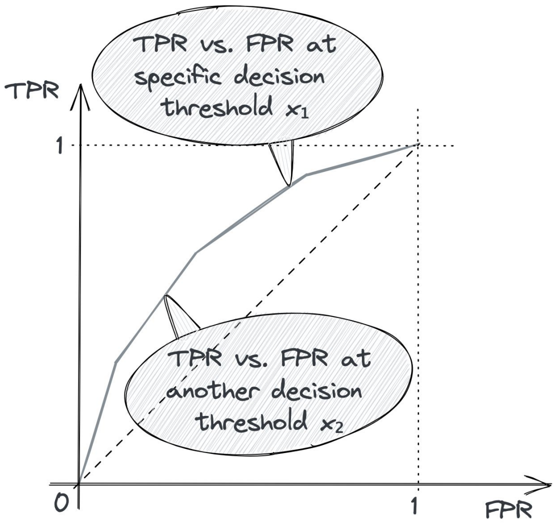
$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

| | | experiment | |
|---|---|---|---|
| | | **true** | **false** |
| **ground truth** | **true** | **true positive** | **false negative** |
| | **false** | **false positive** | **true negative** |

Confusion Matrix

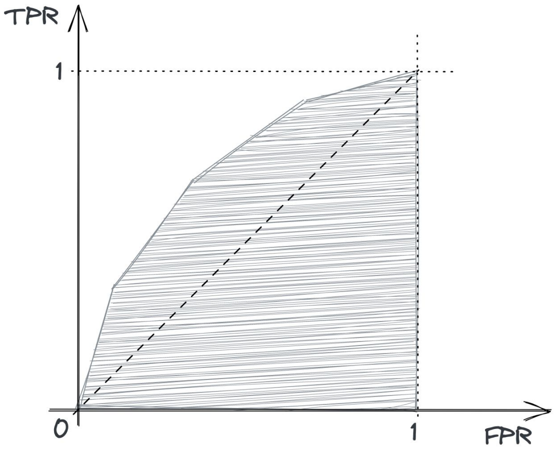
# Evaluation - ROC Curve

- How do we compare the performance of different models or models using different parameters?
- **ROC Curve (Receiver-Operator Characteristic)**
  - Y-axis: True Positive Rate $$TPR = \frac{\#TP}{\#TP + \#FN}$$
  - X-axis: False Positive Rate $$FPR = \frac{\#FP}{\#FP + \#TN}$$
  - An ROC curve plots **TPR vs. FPR at different classification thresholds**. Lowering the classification threshold classifies more items as positive, thus increasing both FP and TP.
    - Classifiers that give curves closer to the top-left corner indicate a better performance.



# Evaluation - AUC (Area under the ROC Curve)

- The **Area under the ROC Curve (AUC)** measures the entire two-dimensional area underneath the entire ROC curve.
- **AUC** represents the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one.
- AUC provides an aggregate measure of performance across all possible classification thresholds.
  - **AUC is 0** if predictions are 100% wrong.
  - **AUC is 1** if all predictions are correct.
  - AUC is scale-invariant and classification-threshold-invariant.



# 4. Machine Learning - 1

## Bibliography


- S. Marsland, ***Machine Learning, An Algorithmic Perspective***, 2nd. ed., Chapman & Hall / CRC Press, 2015
  - Chap. 1 (Types of Machine Learning, Supervised Learning)
  - Chap. 2 (Terminology, Machine Learning Challenges, Statistics)

  *(The book should also be available on the Web as pdf, just keep looking…)*

- E. Kochi, *How to Prevent Discriminatory Outcomes in Machine Learning*, medium.com
- *Machine Learning and Human Bias*, Google @ YouTube

# 4. Machine Learning - 1
## Syllabus Questions

- Explain the two fundamental approaches of Artificial Intelligence.
- What was the reason for the “AI Winter”?
- What is AI and what is the goal of AI?
- What is Machine Learning?
- Explain, how humans learn.
- What is the difference between Classification and Regression?
- Explain, how Supervised Learning works.
- Explain, how Unsupervised Learning works and for what kind of application it is useful.
- Explain the main challenges of Machine Learning.
- Explain the term Overfitting.
- What tasks are included in Data Cleaning?
- Explain K-fold Cross Validation.
- What is the difference between recall/precision and the ROC curve?