Machine Learning

MA DTA, 2023-2024

walter.daelemans@uantwerpen.be
nicolae.banari@uantwerpen.be
Extra practice: ine.gevers@uantwerpen.be


Predicting the future?


Or predicting the past / present?
Or automating complex tasks?
Or getting rid of programming?
Or finding explanations?
Or simulating human intelligence?
By finding patterns in data

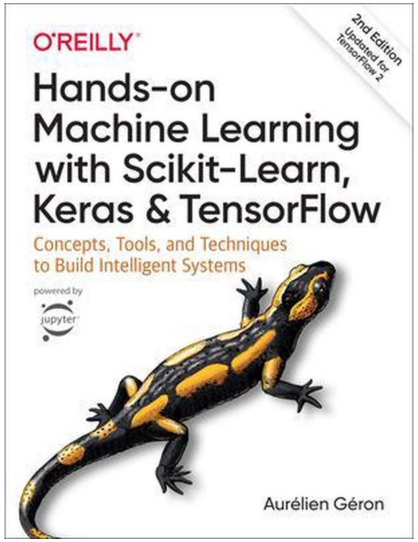
Arthur Samuel (1901-1990)

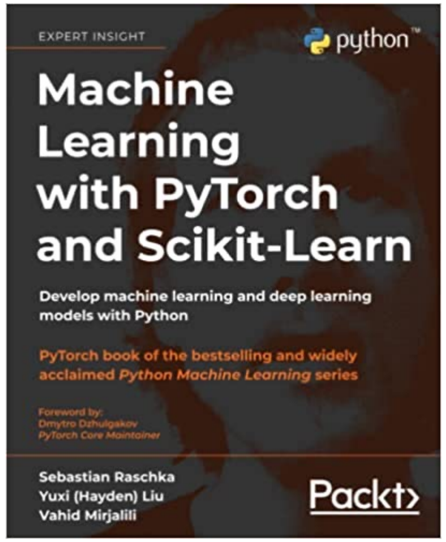
- Game of checkers (1959)
- Search: minimax, + evaluation function
- “rote learning” (position + outcome) / eval function update
- “self play”
- Machine beats its maker

- AlphaGo (Google Deepmind)
- Defeats champion in Go, Lee Sedol (2016)
  - Also learns *chess* from scratch with deep RL

Machine Learning

- Contents
  - Part I: Classical Machine Learning Algorithms
  - Part II: Neural Networks
- Course Material
  - Notebooks + Slides
  - Background material: textbooks, links to sites, …





Evaluation (changed due to ChatGPT)

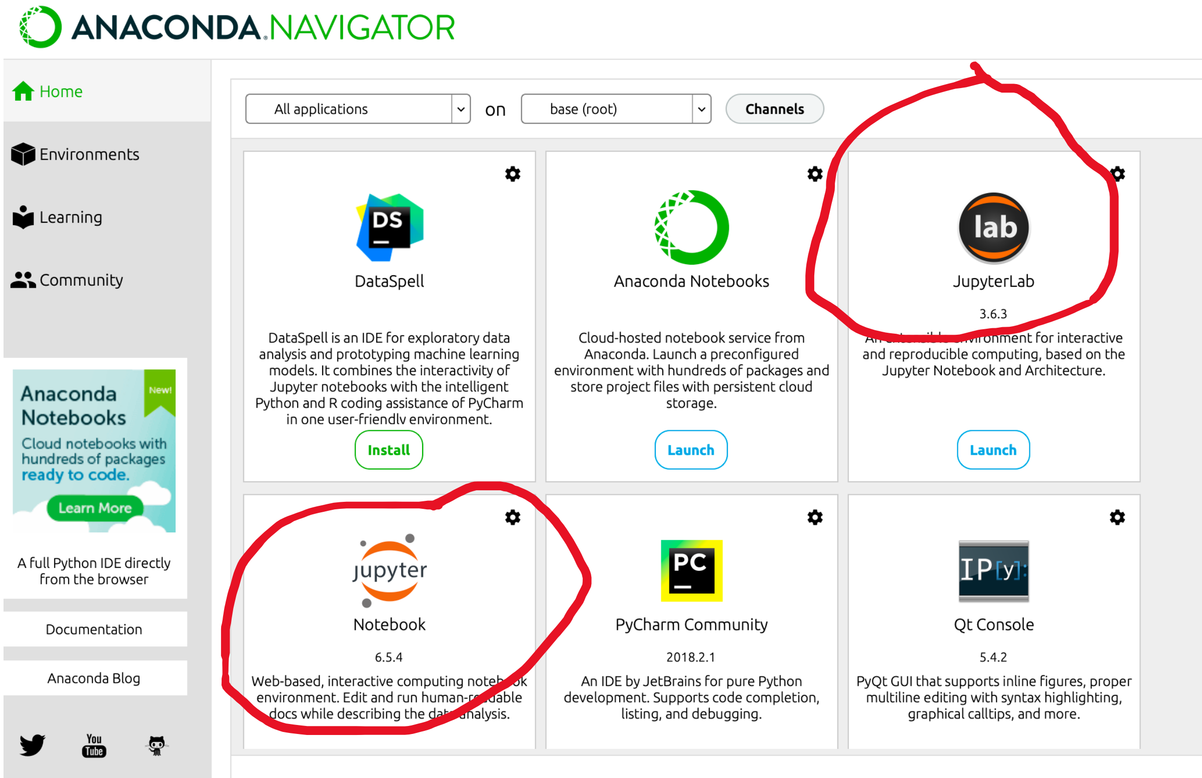
- No more individual project
- No more graded homework
  - Homework will be provided and is very much encouraged, discussion in class and in extra practice sessions
- Written exam
  - Questions about ML concepts / methods / code snippets

Program (preliminary!)

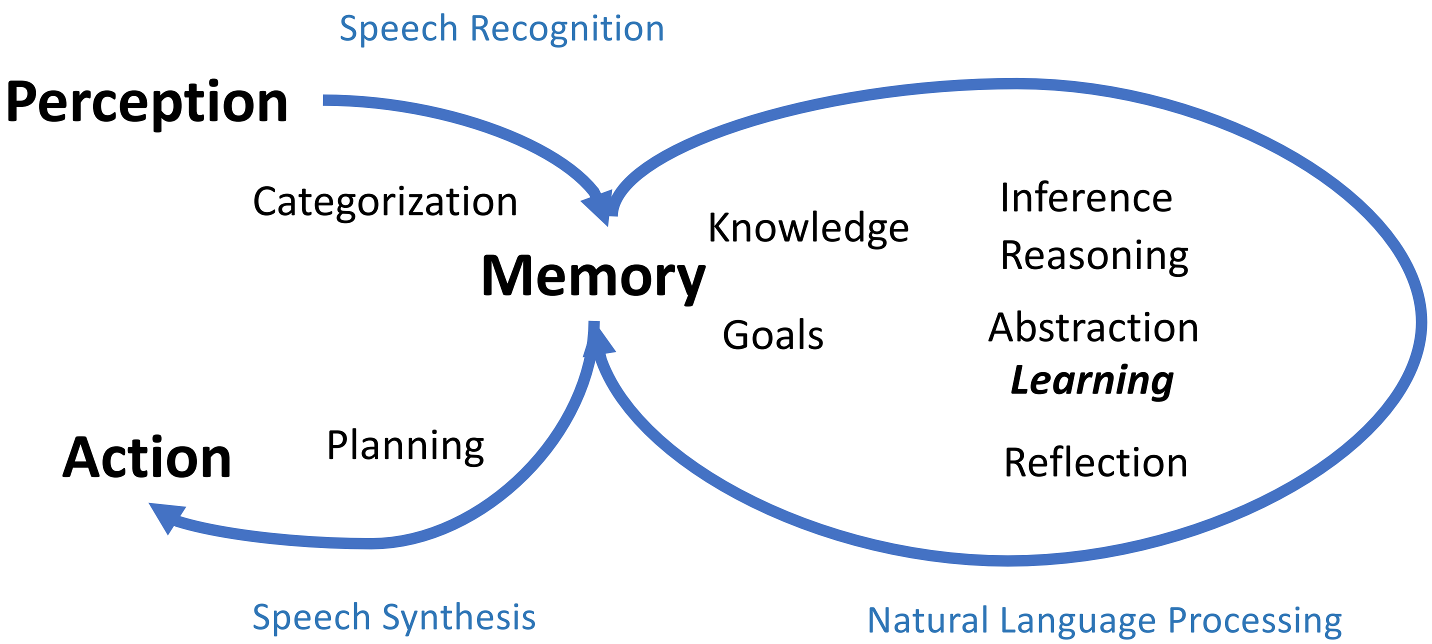
- Oct 15: Introduction to the course and to ML, Linear Regression intro
- Oct 22: Linear Regression , Gradient Descent, Logistic Regression
- Oct 29: Pipelines and Gridsearch
- Nov 5: Evaluation, Bias & Variance, Comparing ML Algorithms
- Nov 12: ML Pipelines for Text, more ML Algorithms
- Nov 19: Ensemble methods, Unsupervised Learning
- Nov 26: Introduction to neural networks, Google Colab, PyTorch, basic linear models in PyTorch
- Dec 3: Classification models, first deep models, early stopping
- Dec 10: Underfitting, overfitting, regularization, dropout, CNNs
- Dec 17: RNNs, LSTMs (transformers: NLP course)

From Next Week …

- Bring your laptop (full battery)
- Install Anaconda on it (free)
  - https://www.anaconda.com/
- Get used to Jupyter notebooks if you aren't already



Artificial Intelligence / Intelligent Agents

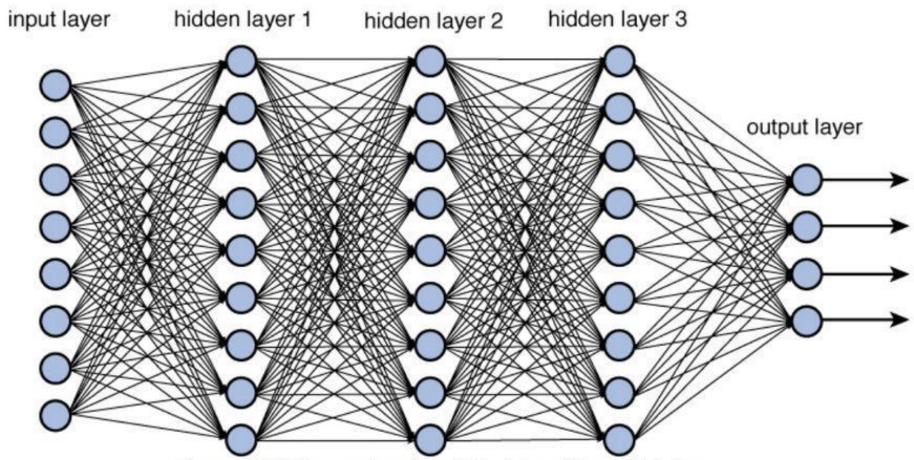


What is Machine Learning?

- Finding patterns in data related to some task
  - To improve the understanding of that task
  - To automate the task or predict the (possible) future
- These days: Transformers = DL = ML = NLP = AI
- The success of ML is due to
  - Exponential growth in available data
  - Exponential growth in computing power
  - Clever engineering and optimization algorithms
  - Creation of special hardware (GPU, TPU)
  - Commercially useful applications

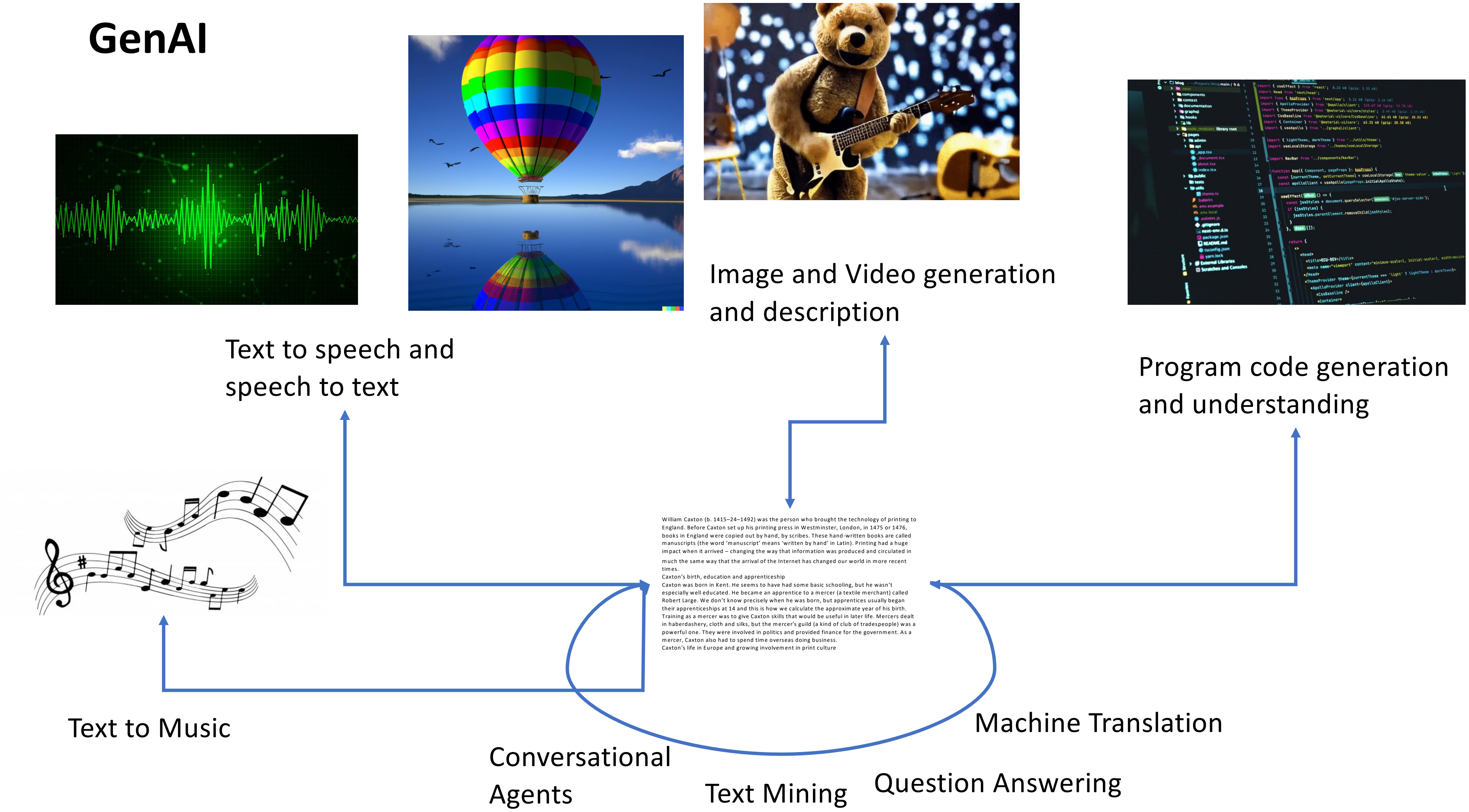
ML in 2024: Deep Neural Networks

Perception

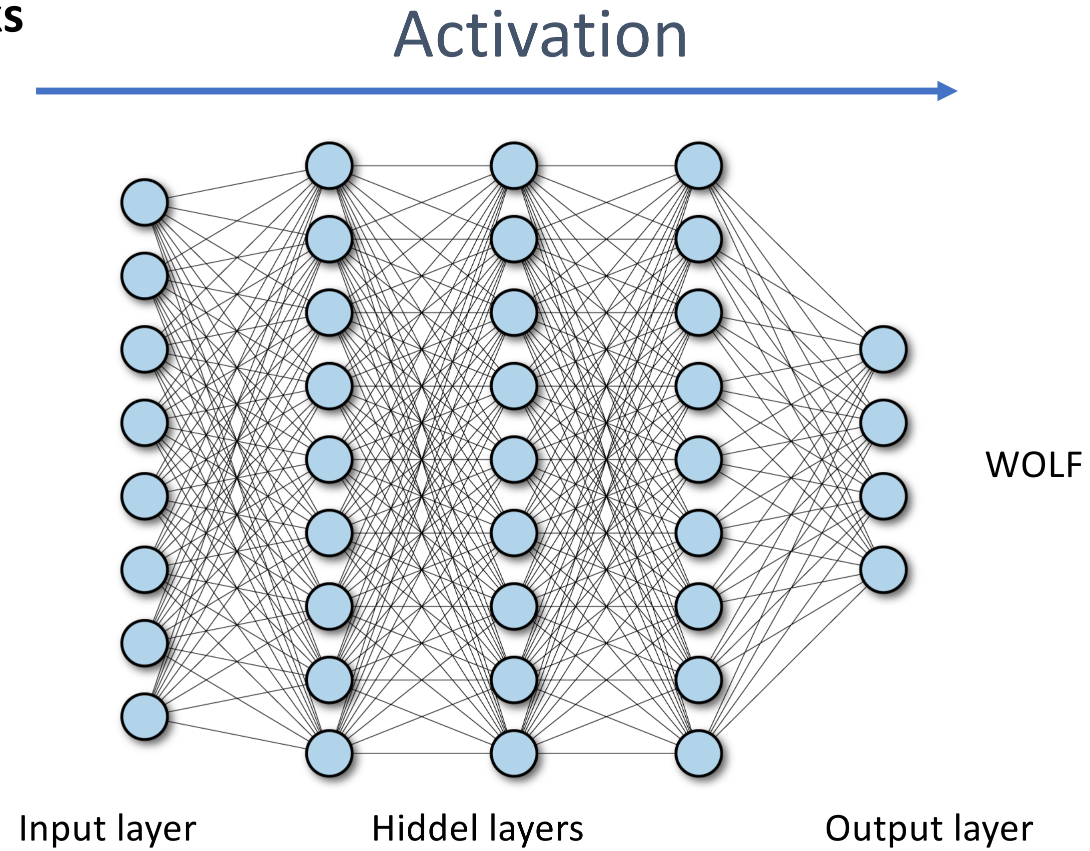


Action

End-to-End Deep Neural Network

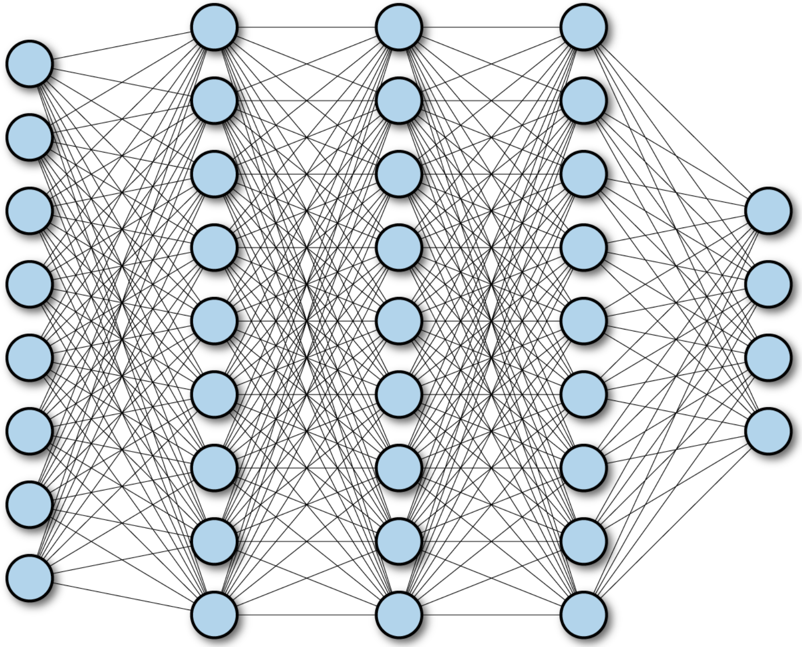

GenAI
Text to speech and speech to text
Image and Video generation and description
Program code generation and understanding
Text to Music
Conversational Agents
Text Mining
Question Answering
Machine Translation
William Caxton (b. 1415–24–1492) was the person who brought the technology of printing to England. Before Caxton set up his printing press in Westminster, London, in 1475 or 1476, books in England were copied out by hand, by scribes. These hand-written books are called manuscripts (the word 'manuscript' means 'written by hand' in Latin). Printing had a huge impact when it arrived – changing the way that information was produced and circulated in much the same way that the arrival of the Internet has changed our world in more recent times.
Caxton's birth, education and apprenticeship
Caxton was born in Kent. He seems to have had some basic schooling, but he wasn't especially well educated. He became an apprentice to a mercer (a textile merchant) called Robert Large. We don't know precisely when he was born, but apprentices usually began their apprenticeships at 14 and this is how we calculate the approximate year of his birth. Training as a mercer was to give Caxton skills that would be useful in later life. Mercers dealt in haberdashery, cloth and silks, but the mercer's guild (a kind of club of tradespeople) was a powerful one. They were involved in politics and provided finance for the government. As a mercer, Caxton also had to spend time overseas doing business.
Caxton's life in Europe and growing involvement in print culture

Deep Neural Networks



Deep Neural Network

Activation



DOG

Cost Function

Input layer Hidden layers Output layer

SUPERVISED LEARNING

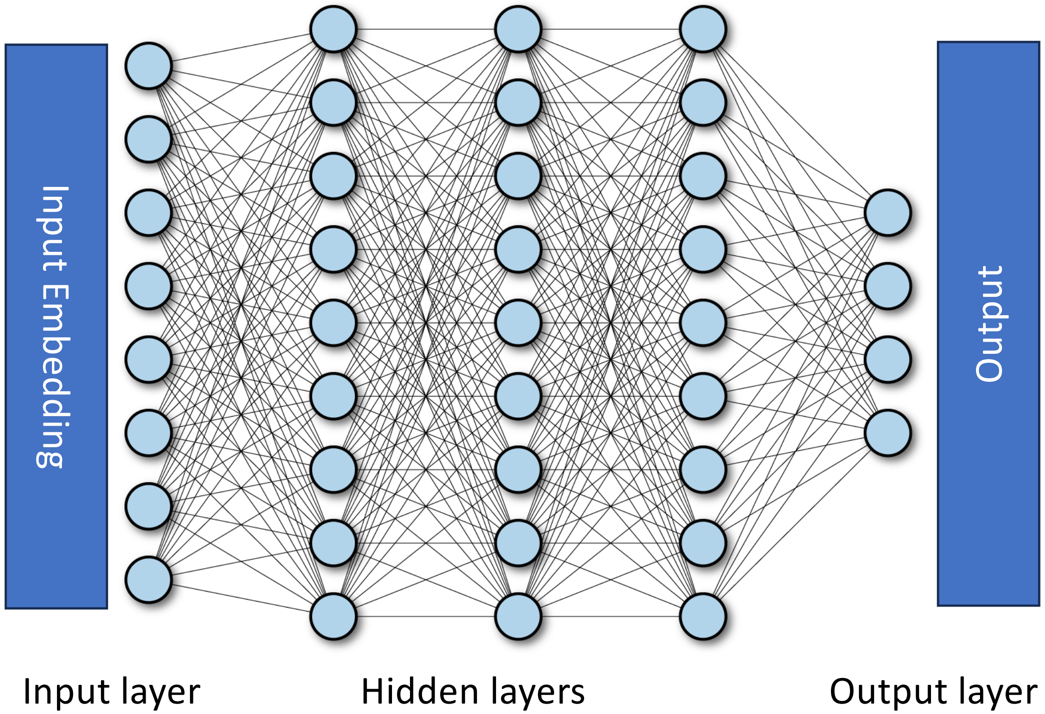
(Training Data)

Backpropagation of errors

LLMs: “Completion on steroids”

How do we apply this to Language?

Nee, polarisering in de samenleving kun je niet oplossen met enkele simpele aanpassingen aan de algoritmes van Facebook en Instagram. Als er al een conclusie te trekken valt uit vier uitgebreide studies die donderdag verschenen in de tijdschriften *Science* en



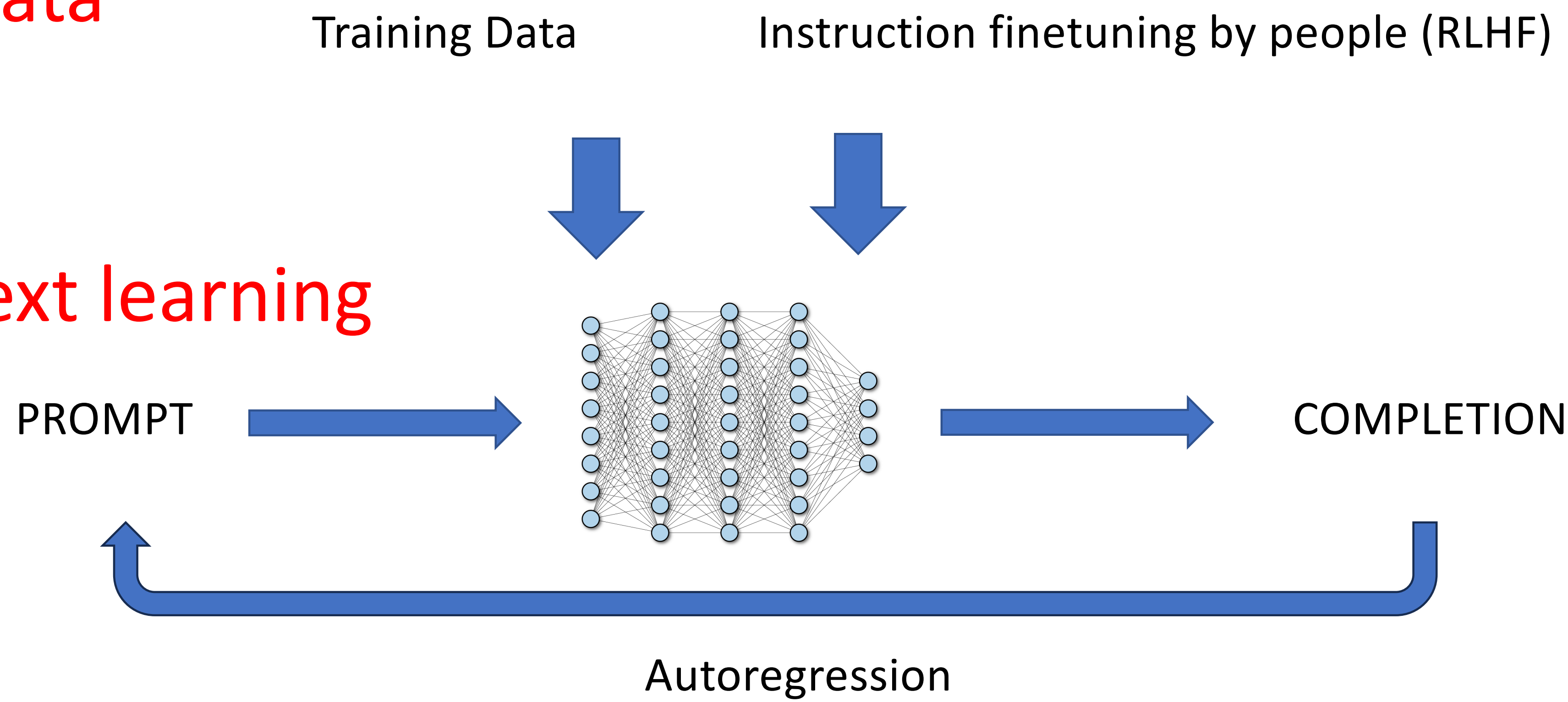
Nature 0.7
gelijkaardige 0.1
Technology 0.05
cocktails 0.0002
...

SELF-SUPERVISED LEARNING
Result: PRE-TRAINED MODEL

Finetuning on user data

Alignment

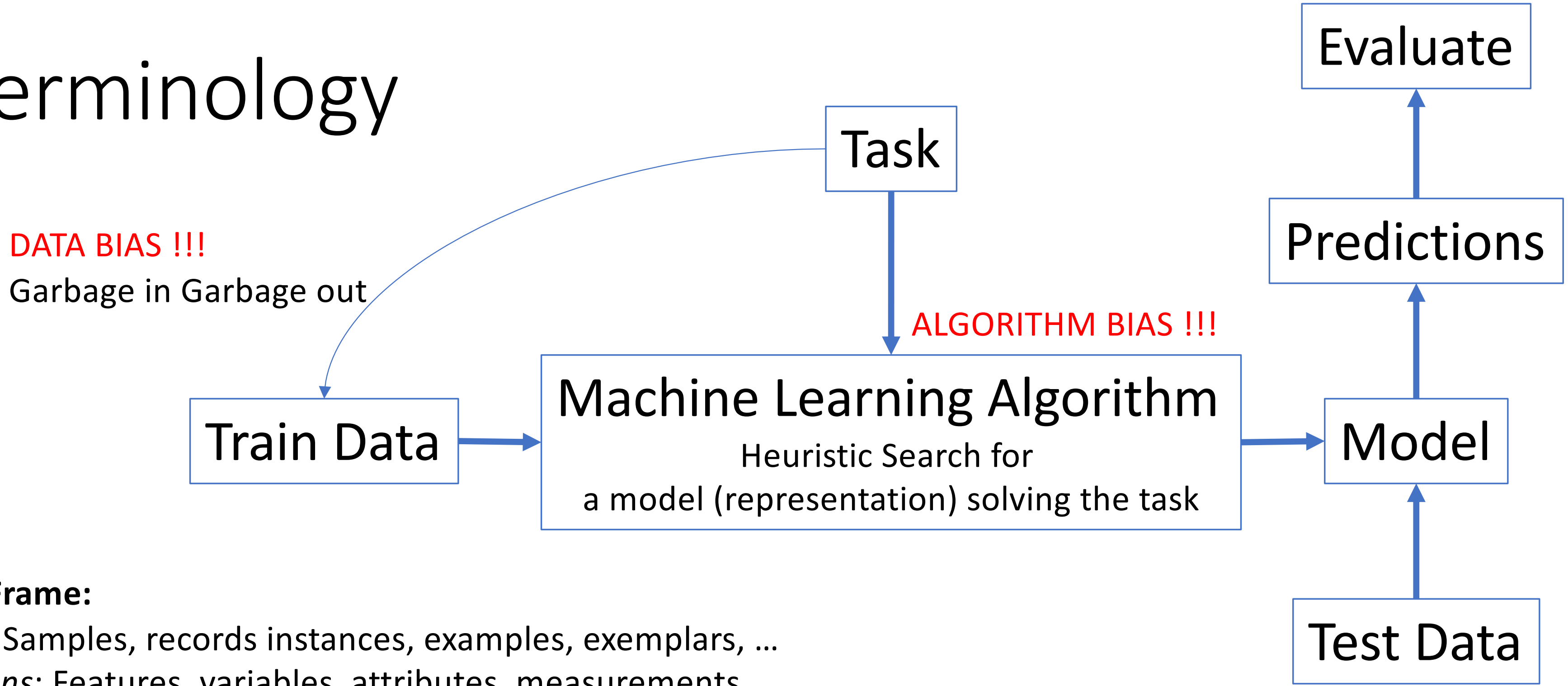
In context learning



ML in NLP

Knowledge-Based	Machine Learning	Deep neural networks	Pre-trained models (embeddings, foundation models)
Handcrafted	Supervised	Supervised	Self-supervised
Modular / pipelines	Feature engineering (Modular / pipelines)	End-to-end	Finetuning
			“Emergence”
-1995	1995-2010	2010-2015	2015-
Interpretable	Interpretable-ish	Not Interpretable	Not Interpretable

Terminology



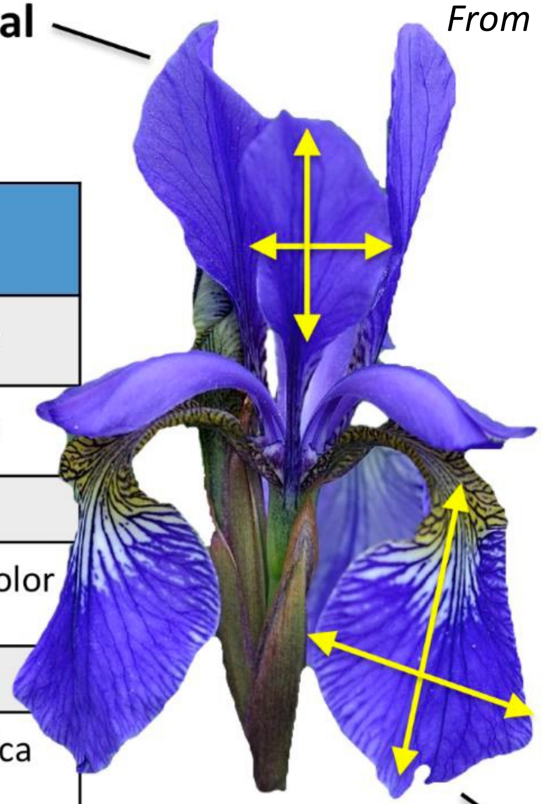
Data Frame:
Rows: Samples, records instances, examples, exemplars, …
Columns: Features, variables, attributes, measurements, …

Supervised Learning: input and output features (independent and dependent variables)
Unsupervised Learning: no output feature(s)

From Python ML, 3rd edition

Samples
(instances, observations)

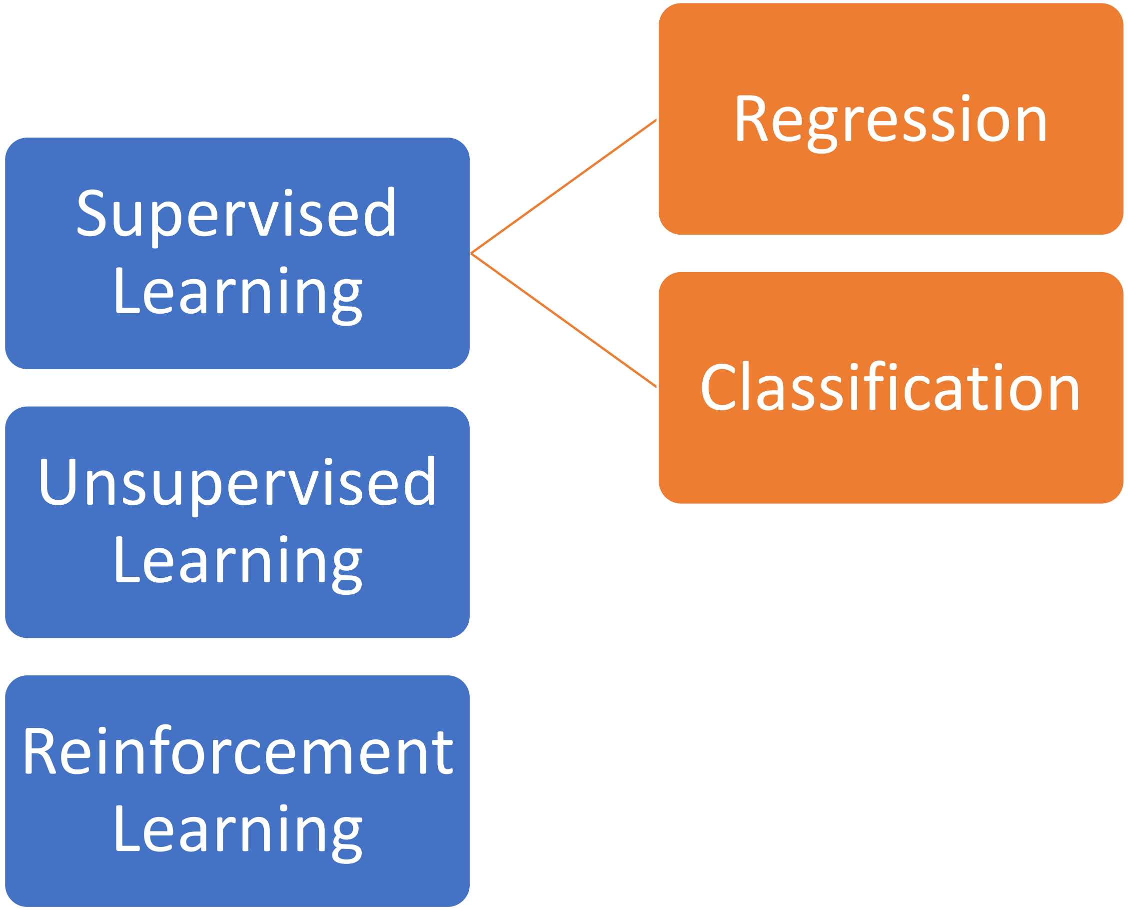
	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
			…		
50	6.4	3.5	4.5	1.2	Versicolor
			…		
150	5.9	3.0	5.0	1.8	Virginica



Features
(attributes, measurements, dimensions)

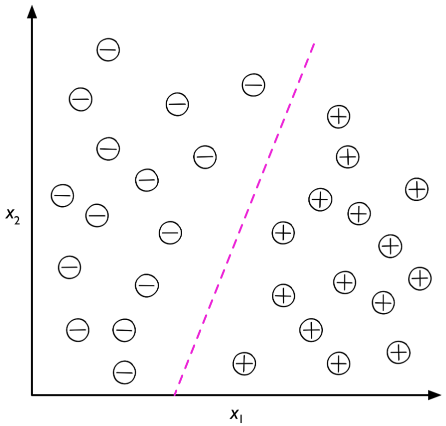
Class labels
(targets)

Types of Machine Learning

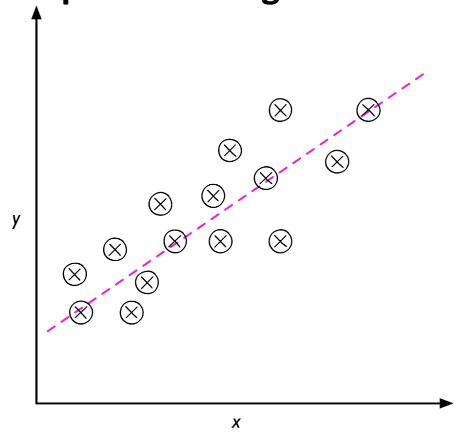


From Python ML, 3rd edition

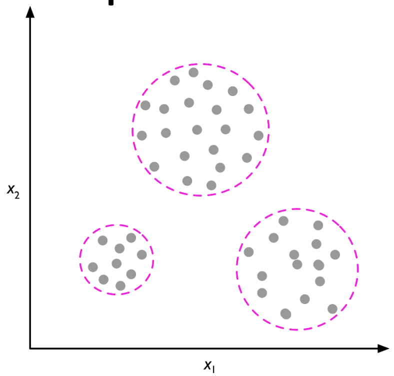
Supervised: classification



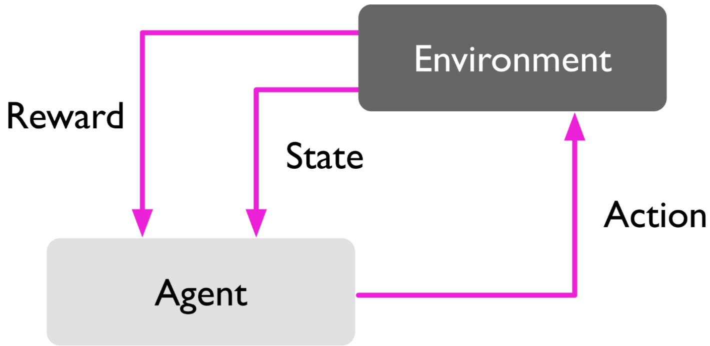
Supervised: regression



Unsupervised: clustering



Reinforcement

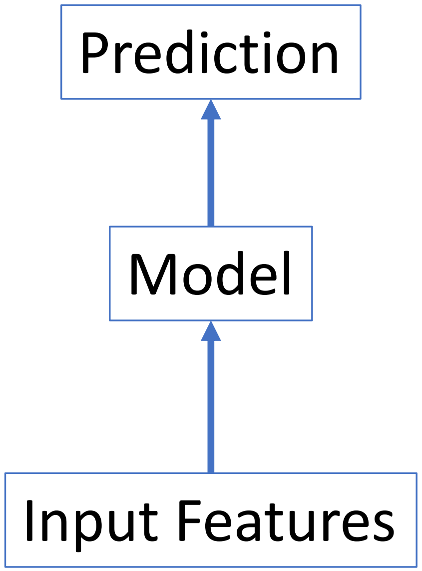


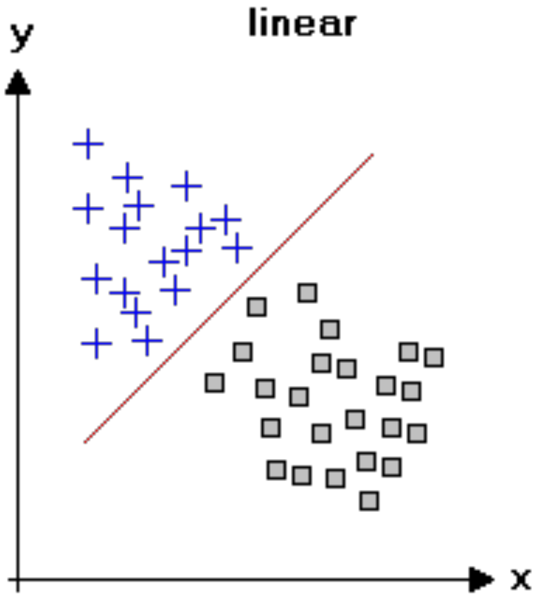
Batch and Online

- Batch learning (offline):
  - Train the model on all available train data
  - In case of new train data: **merge old + new** and retrain from scratch
- Incremental learning (online):
  - Train the model on all available train data
  - In case of new train data (samples or mini-batches): **keep existing** model and update by **training on new** data only

Common Learning Tasks

- Multi-class classification
  - Output is one of a set of labels
  - Examples: Face recognition, gender detection, …
- Binary classification
  - Output is either 1 or 0
  - Examples: spam detection, gender detection, fraud detection, …
- Multi-class multi-label classification
  - Output is subset of the set of labels
  - Examples: topic detection, medical codes detection, …
- Regression
  - Output is a number
  - Examples: age prediction, house price prediction, predicting next years' revenue, …
- Clustering
  - Output is a number of groups of similar samples
  - Examples: grouping customers, recommendation, …





From Python ML, 3rd edition

Train Data

Enough Data

Representative Data

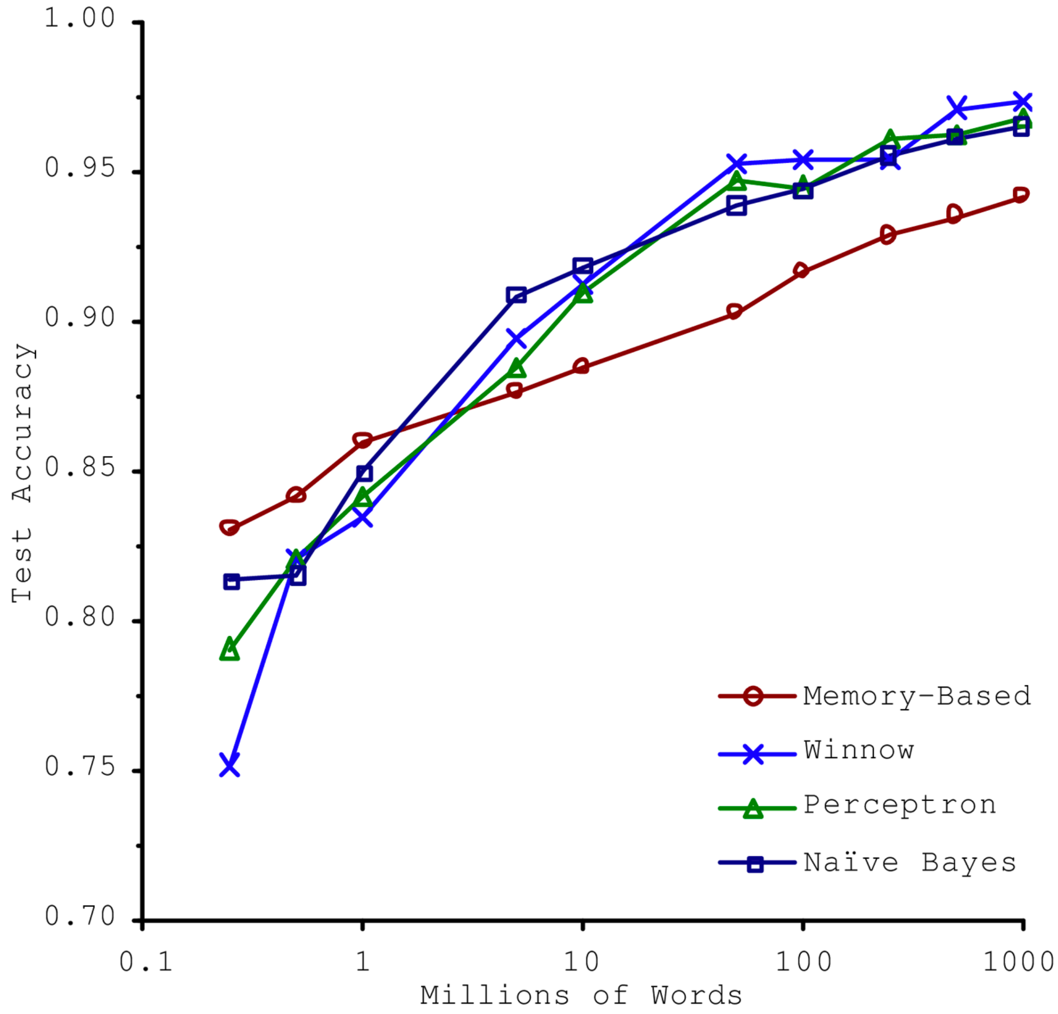
- Unbiased
- Similar distribution to test data

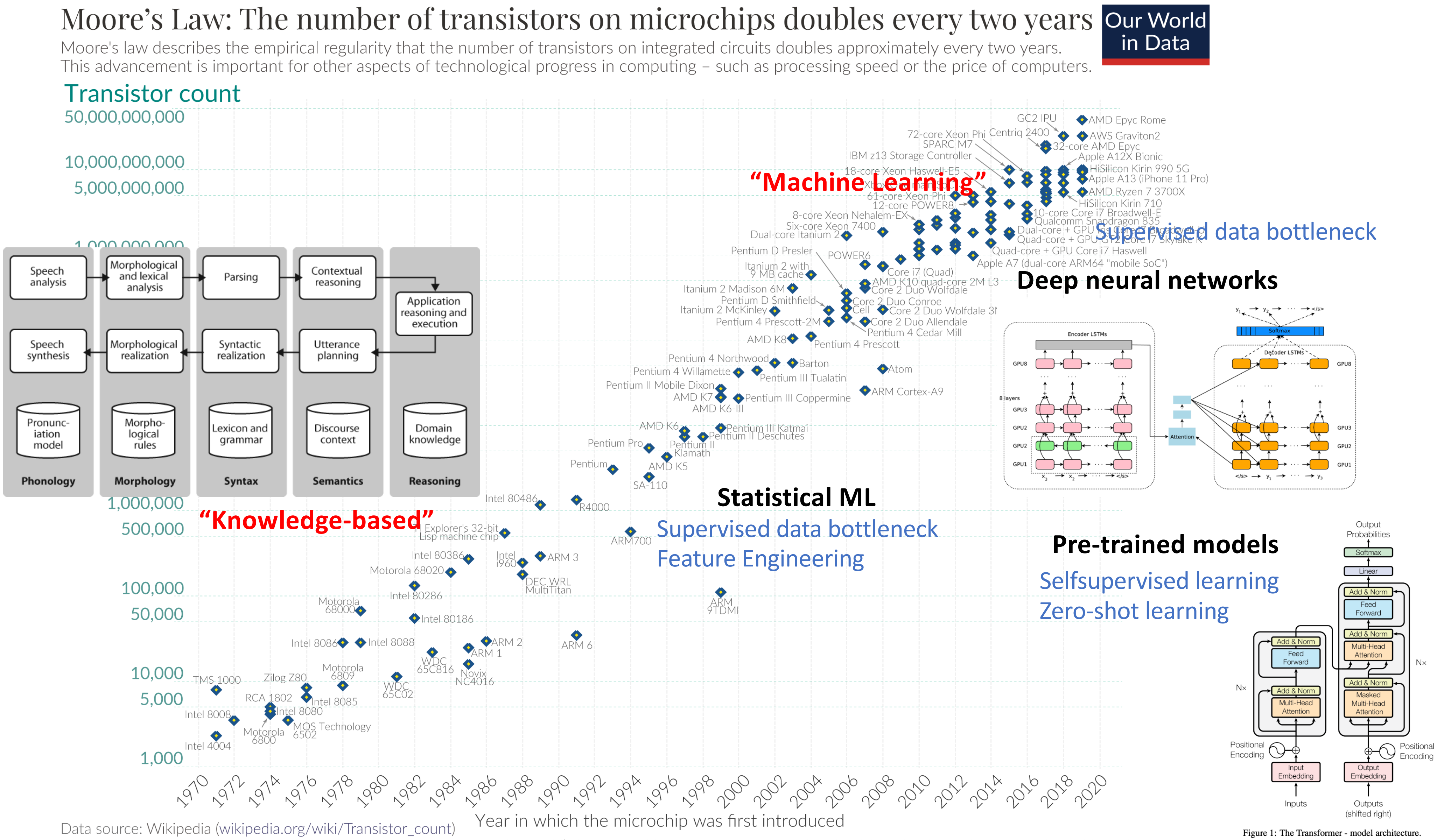
Qualitative Data

- Informative Features
- Learnable Classes

Scaling: Size matters!

Banko & Brill, 2001





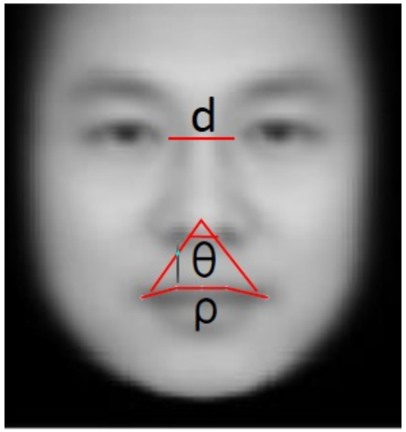
Moore's Law: The number of transistors on microchips doubles every two years
Our World in Data
Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years.
This advancement is important for other aspects of technological progress in computing – such as processing speed or the price of computers.
Transistor count
50,000,000,000
10,000,000,000
5,000,000,000
1,000,000
500,000
100,000
50,000
10,000
5,000
1,000
"Machine Learning"
Supervised data bottleneck
Deep neural networks
Speech analysis
Morphological and lexical analysis
Parsing
Contextual reasoning
Application reasoning and execution
Speech synthesis
Morphological realization
Syntactic realization
Utterance planning
Pronunciation model
Morphological rules
Lexicon and grammar
Discourse context
Domain knowledge
Phonology
Morphology
Syntax
Semantics
Reasoning
"Knowledge-based"
Statistical ML
Supervised data bottleneck
Feature Engineering
Pre-trained models
Selfsupervised learning
Zero-shot learning
Encoder LSTMs
Decoder LSTMs
Softmax
Attention
Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Multi-Head Attention
Masked Multi-Head Attention
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)
Figure 1: The Transformer - model architecture.
1970 1972 1974 1976 1978 1980 1982 1984 1986 1988 1990 1992 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012 2014 2016 2018 2020
Year in which the microchip was first introduced
Data source: Wikipedia (wikipedia.org/wiki/Transistor_count)
OurWorldinData.org – Research and data to make progress against the world's largest problems.
Licensed under CC-BY by the authors Hannah Ritchie and Max Roser.

ML project steps

- Data collection and preprocessing
  - Feature extraction, construction and selection
  - Feature scaling
  - Dimensionality reduction
- Machine Learning experiments
  - Model selection
  - Cross-validation
  - Hyper-parameter optimization
  - Evaluation metrics
  - Error analysis

ML and Ethics
Criminal Faces (?!?)

- “Automated Inference on Criminality using Face Images” (Wu & Zhang), 2016
- 1856 real persons, half of them convicted
- Accuracy ~ 90% (cNN)
- Predictive features: lip curvature, distance corners of the eye, nose-mouth angle



(a) Three samples in criminal ID photo set S_c.

(b) Three samples in non-criminal ID photo set S_n

Figure 1. Sample ID photos in our data set.

Stereotypes // Cluster-centroïds

- Types of criminals
  - “criminals have a higher degree of dissimilarity in facial appearance than non-criminals” (*Law of normality)*
- Correlated with the opinion of 500 students
- Points at a rather disturbing alternative explanation: people with faces that ‘look criminal’ are sentenced more often?


IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 36, NO. 12, DECEMBER 2014 2483

Learning Race from Face: A Survey

Siyao Fu, Member, IEEE, Haibo He, Senior Member, IEEE, and Zeng-Guang Hou, Senior Member, IEEE

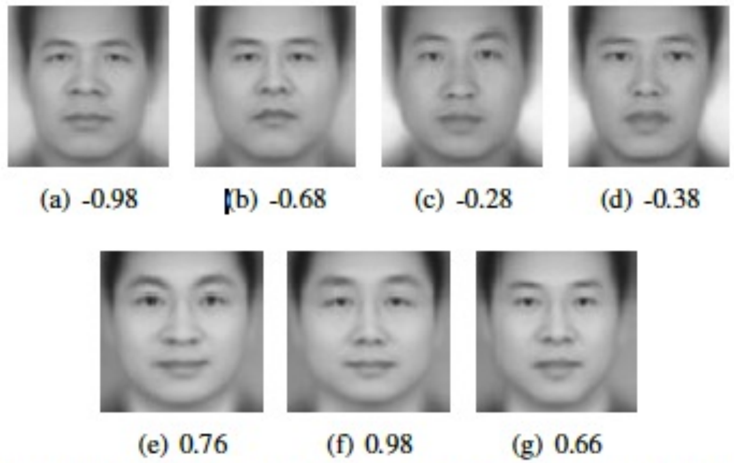


Figure 13. (a), (b), (c) and (d) are the four subtypes of criminal faces corresponding to four cluster centroids on the manifold of S_c; (e), (f) and (g) are the three subtypes of non-criminal faces corresponding to three cluster centroids on the manifold of S_n. The number associated with each face is the average score of human judges (-1 for criminals; 1 for non-criminals).

Background + Solution

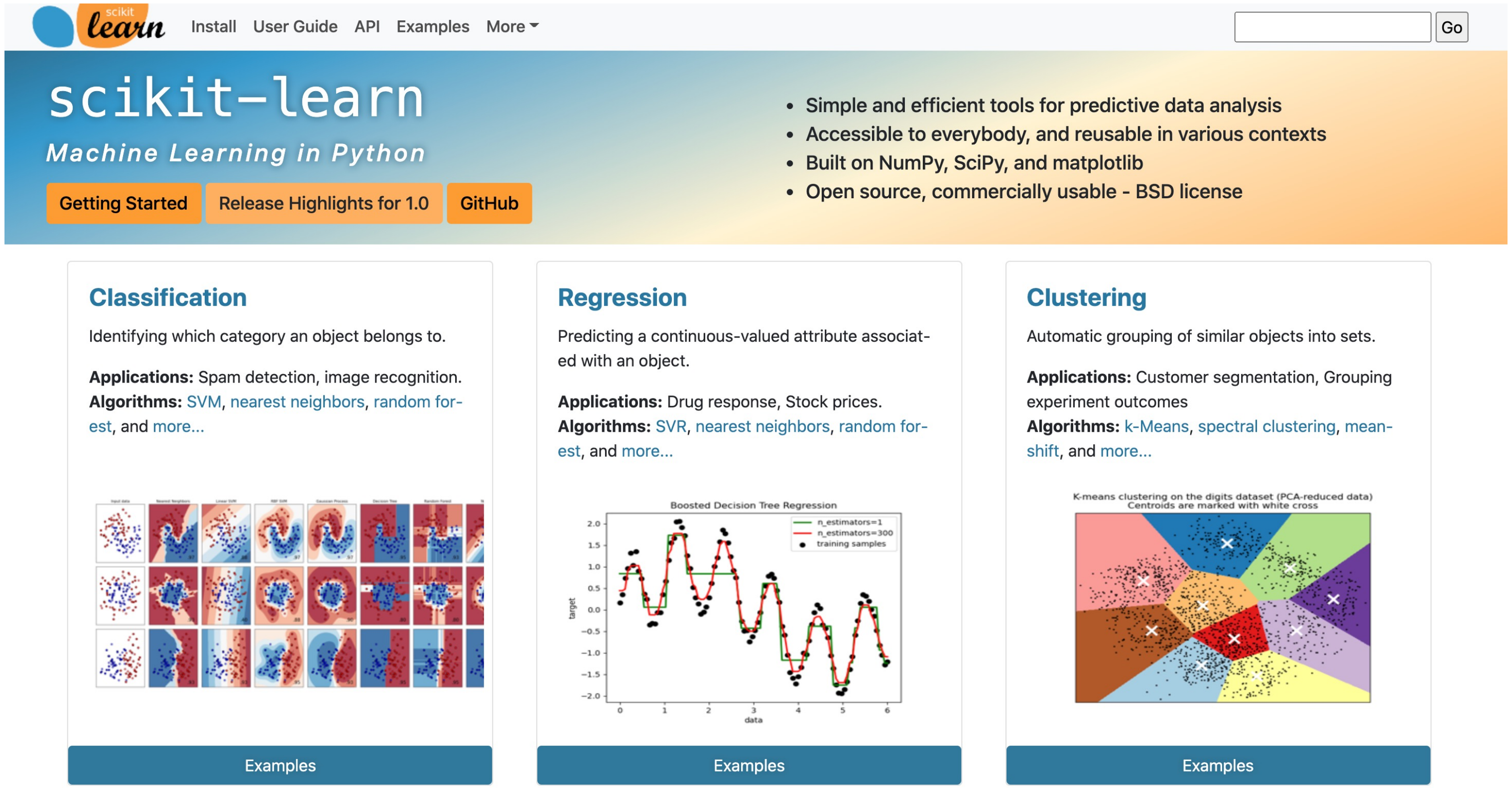
- https://www.callingbullshit.org/case_studies/case_study_criminal_machine_learning.html

https://www.hownormalami.eu/

- What can companies and governments infer from your face using ML?
- Privacy safe (I think)
- Developed in the framework of a European project on ethical aspects of AI
    - https://www.project-sherpa.eu/

scikit-learn

https://scikit-learn.org/stable/



linear regression

- It is regression (output is a number)
- Can be univariate or multiple (one or more input features)
  - $x \rightarrow y$ $f(x) = y$
  - X (= x_1, x_2, … x_i)$\rightarrow y$ $F(X) = y$
- Model is a straight line, hence *linear,* on the data
  - $f(x)$ = predicted $y = wx + b$
    - A line is defined by two values: the intercept (b, where it crosses y-axis) and the coefficient or weight (w) determining the slope and direction of the line