Optimization for ML: introduction

University of Rennes - IRISA - INRIA

April 2024

Outline

- Introduction
  - General introduction
  - Technical introduction (notations, empirical cost)
- Definition of optimization problem
- Properties of optimization problems
- Convexity

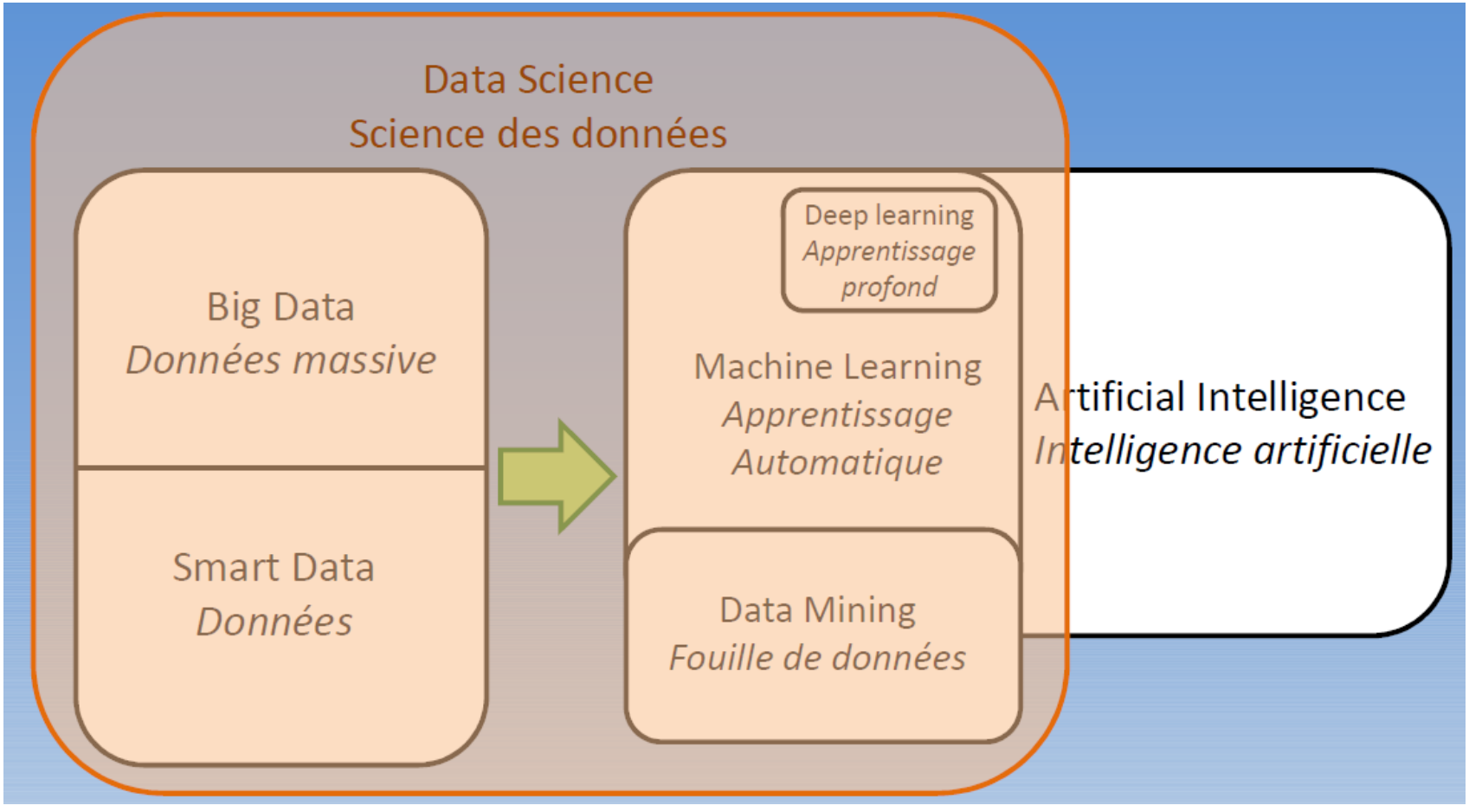
Introduction

- Definitions
- Why the explosion of deep learning
- Successes and failures
- Mathematical notations

Artificial intelligence(AI) versus Machine learning (ML)

'Artificial intelligence is Intelligence displayed by machines, in contrast with the natural intelligence displayed by humans and other animals. In computer science AI research is defined as the study of "intelligent agents": any device that perceives its environment and takes actions that maximize its chance of success at some goal.' wikipedia

'Machine learning is a field of computer science that gives computers the ability to learn without being explicitly programmed.' wikipedia

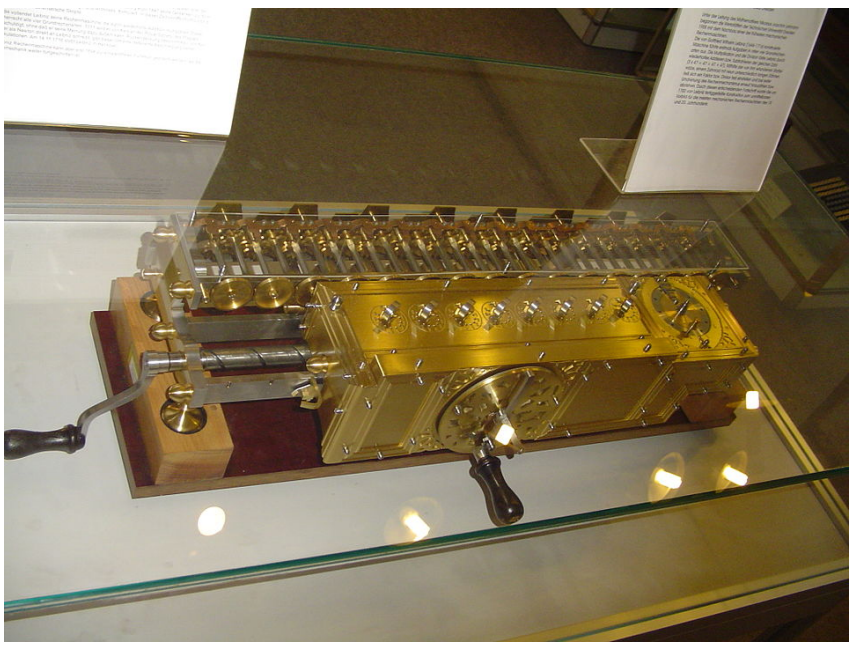

Data Science
Science des données
Big Data
Données massive
Smart Data
Données
Deep learning
Apprentissage profond
Machine Learning
Apprentissage Automatique
Data Mining
Fouille de données
Artificial Intelligence
Intelligence artificielle

Historical perspective at 17's century

- Descartes, 1637: *Ces longues chaînes de raisons, toutes simples et faciles, dont les géomètres ont coutume de se servir pour parvenir à leurs plus difficiles démonstrations, m'avaient donné occasion de m'imaginer que toutes les choses qui peuvent tomber sous la connaissance des hommes s'entresuivent en même façon*
- Thomas Hobbes, 1651: *Car la raison (...) n'est rien d'autre que le fait de calculer, c'est-à-dire additionner et soustraire, les consécutions des dénominations générales admises pour marquer et signifier nos pensées*
- Leibniz, 1684: *Alors, il ne sera plus besoin entre deux philosophes de discussions plus longues qu'entre deux mathématiciens, puisqu'il suffira qu'ils saisissent leur plume, qu'ils s'asseyent à leur table de calcul (en faisant appel, s'ils le souhaitent, à un ami) et qu'ils se disent l'un à l'autre : Calculons !*

Calculus Ratiocinator

- Leibniz, De arte combinatoria, 1666
- Universal characteristics: universal philosophical language used to determine truth in any discussion



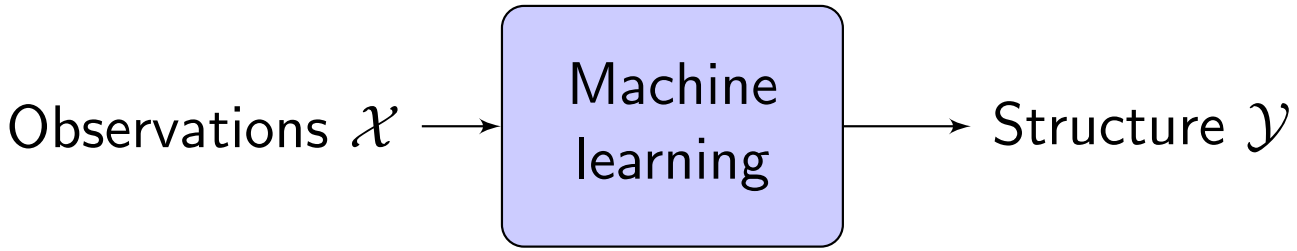
Definition of intelligence by psychologists

- Psychologists were at the beginning of AI (Rosenblatt)
- Louis Leon Thurstone (1887-1955) was an engineer and psychologist, pioneering psychometrics.
- Definition: "intelligence is to put in practice adaptive capabilities resulting from aptitudes"
- Aptitude = functional efficiency of various cognitive processes like perception, attention, sensation, memory, representation, language, reasoning, decision-making, recognition, learning, emotion, forgetting, behaviour
- To be debated: are AI/ML systems intelligent? Sub-question: do ML systems recite or reason?

Definition of intelligence by psychologists (2)

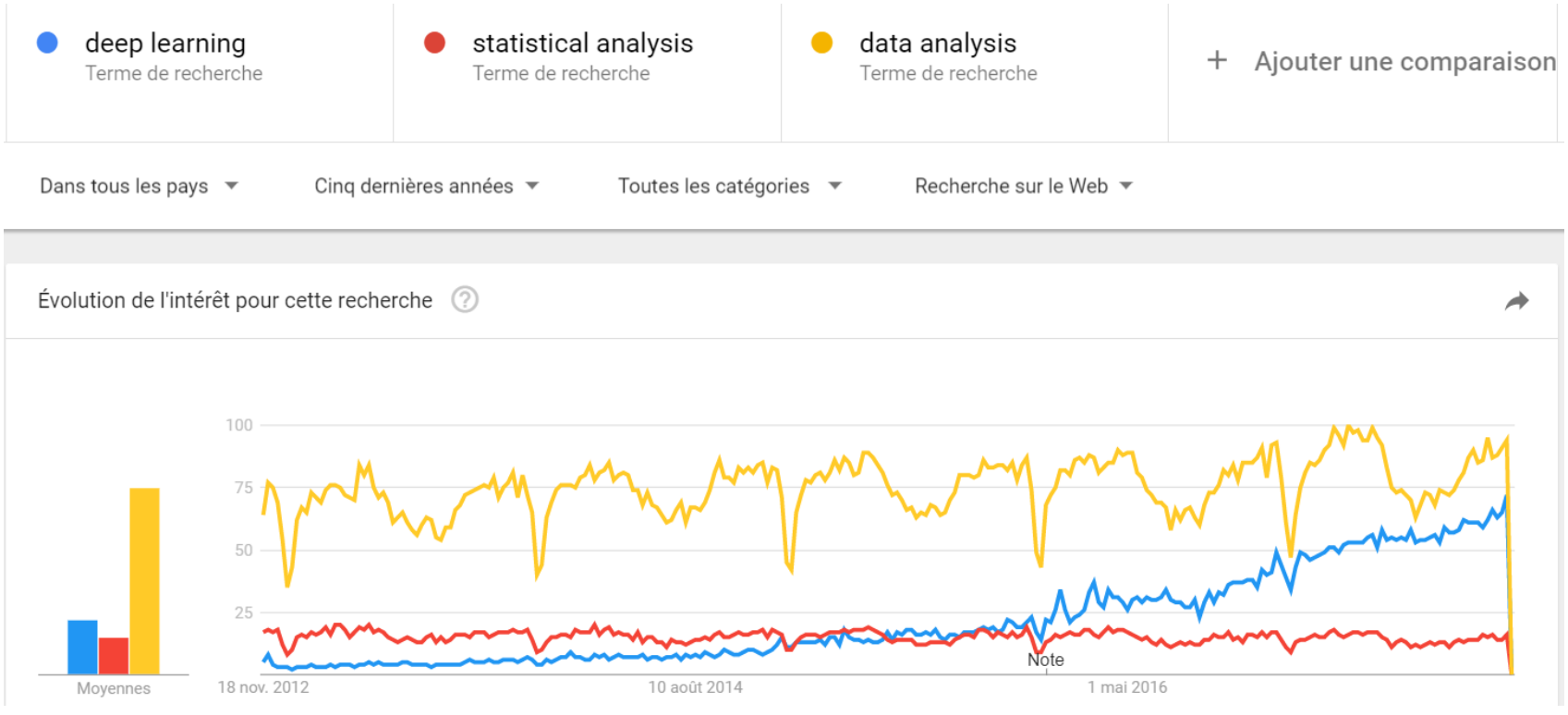
- Psychologists were at the beginning of AI (Rosenblatt)
- Robert Sternberg (1949-) is a professor of cognitive psychology.
- Defines three types of intelligence:
  - Analytic intelligence: problem solving.
  - Creative intelligence. See things differently and tackle new and unusual situations.
  - Practical intelligence. Adapt in day-to-day live by leveraging knowledge and aptitudes.

General view of machine learning



where $\mathcal{X}$ can be any kind of data (numbers, measurements, images, text, audio, *etc.*)
and $\mathcal{Y}$ can be of various type (signal, numbers, images, labels, *etc.*)

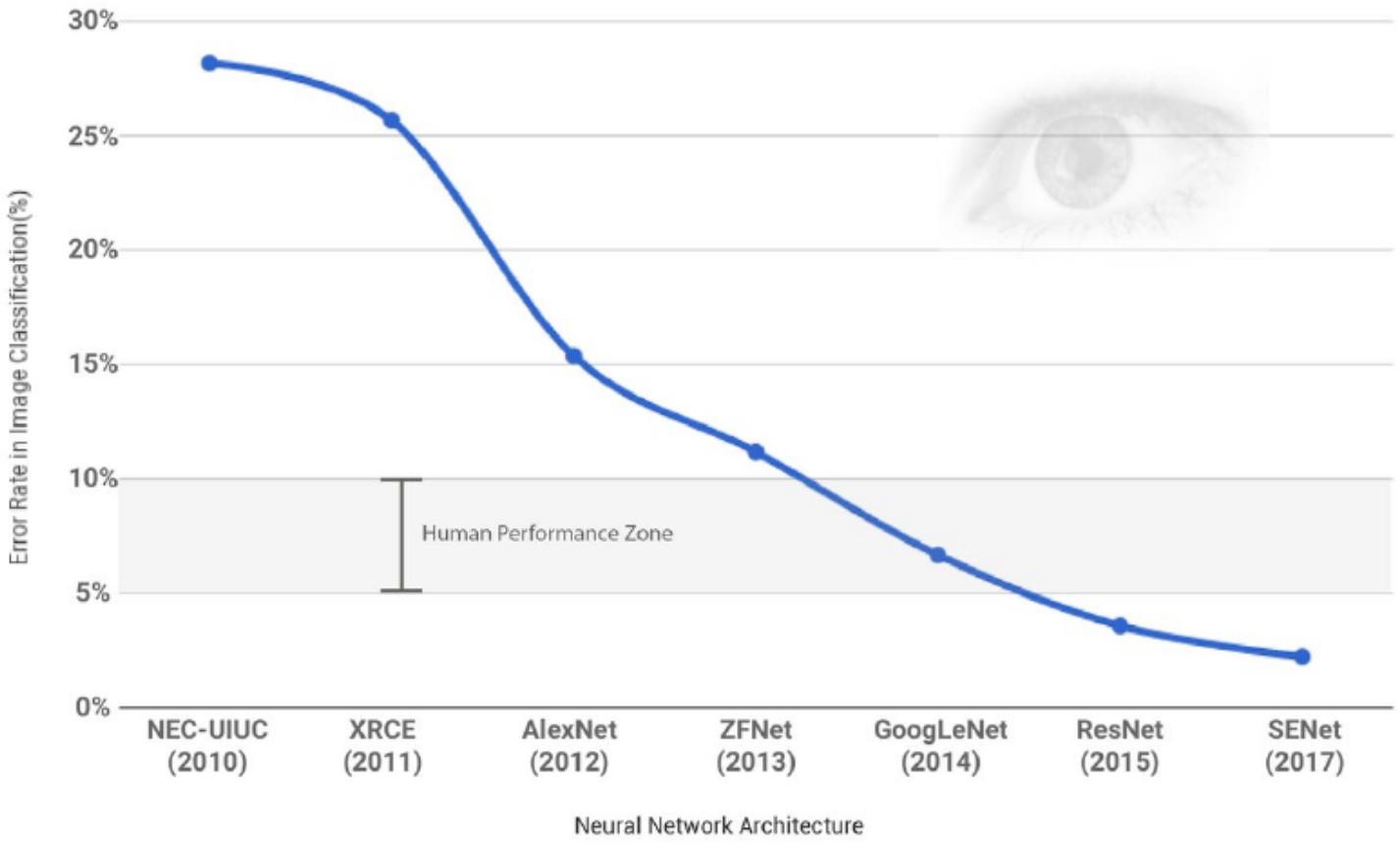
Explosion of deep learning



(Short) history of ML

- 1992: TD-Gammon is able to rival, but not consistently surpass, the abilities of top human backgammon players
- 1997: IBM's Deep Blue beats the world champion at chess
- 2011: IBM's Watson beats two human champions in a Jeopardy! competition
- 2012: AlexNet beats all classical methods on ImageNet challenge
- 2014: Facebook DeepFace uses neural networks that identifies faces with 97.35% accuracy
- 2016: Google's AlphaGo beats a professional player
- 2018: alphafold placed forst in CASP challenge (Protein Structure Prediction)
- Post 2020 genAI methods and LLM
- 2024: AI methods embarked massively on smartphones (gemini, iphones)

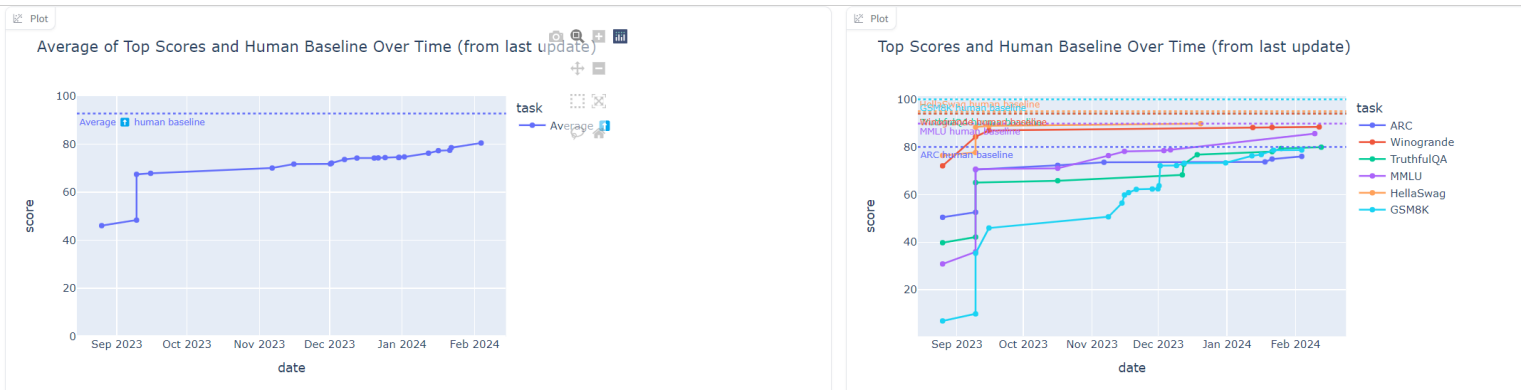
Image recognition



source: https://www.linkedin.com/pulse/must-read-path-breaking-papers-image-classification-muktabh-mayank

Large Langage models

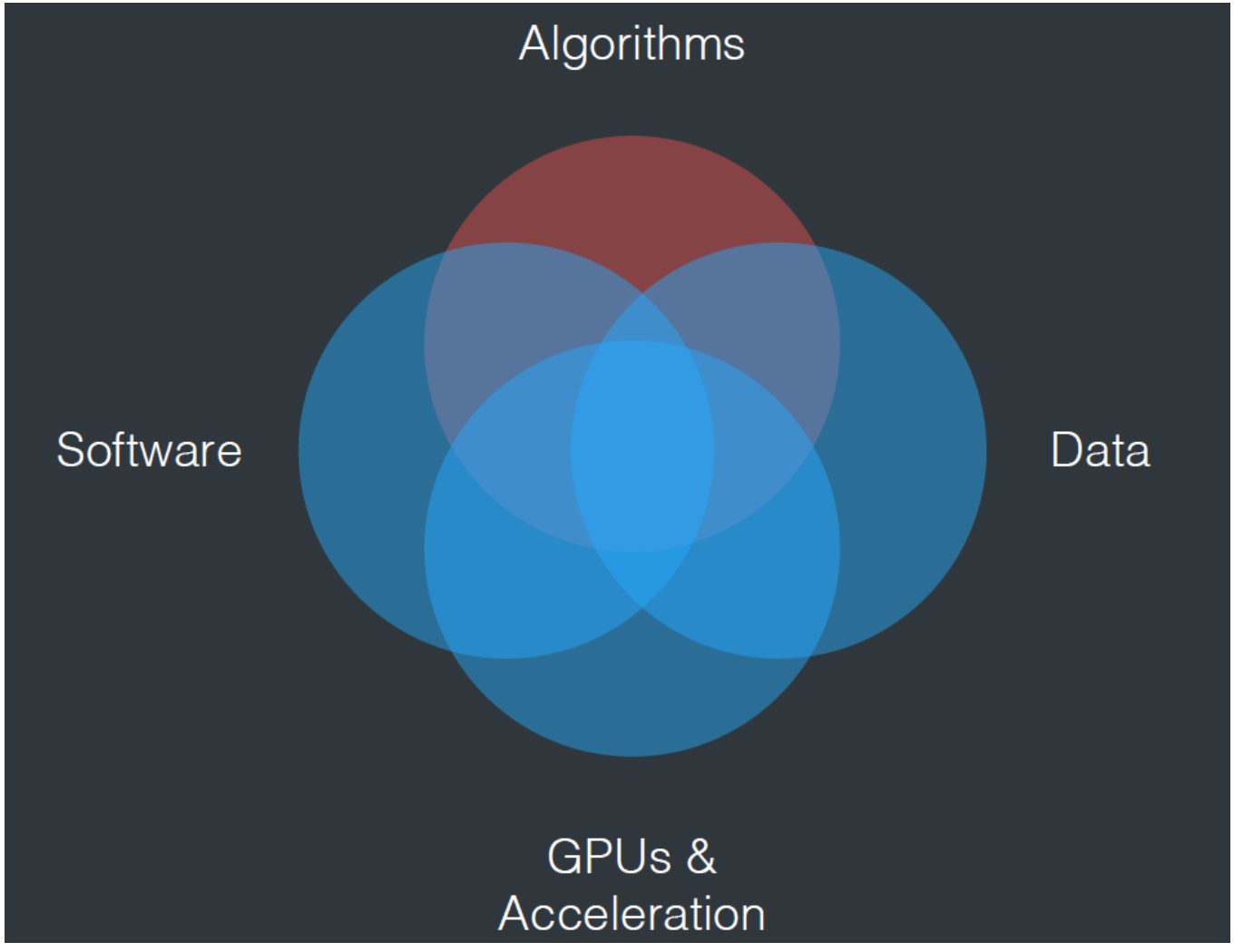
Source `https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard`



Trends [1]

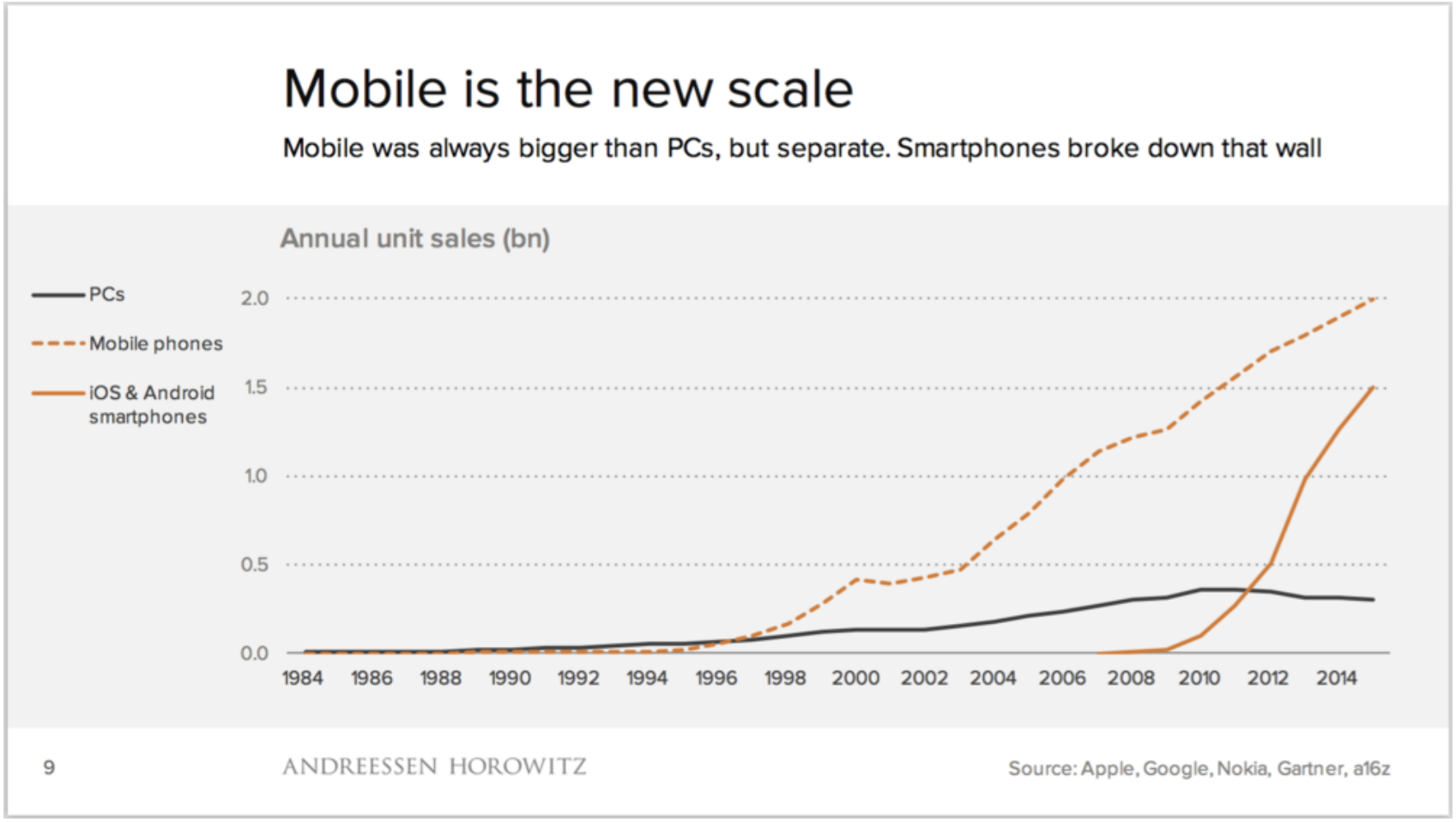
[1] Source: Nvidia

Maturity of machine (deep) learning

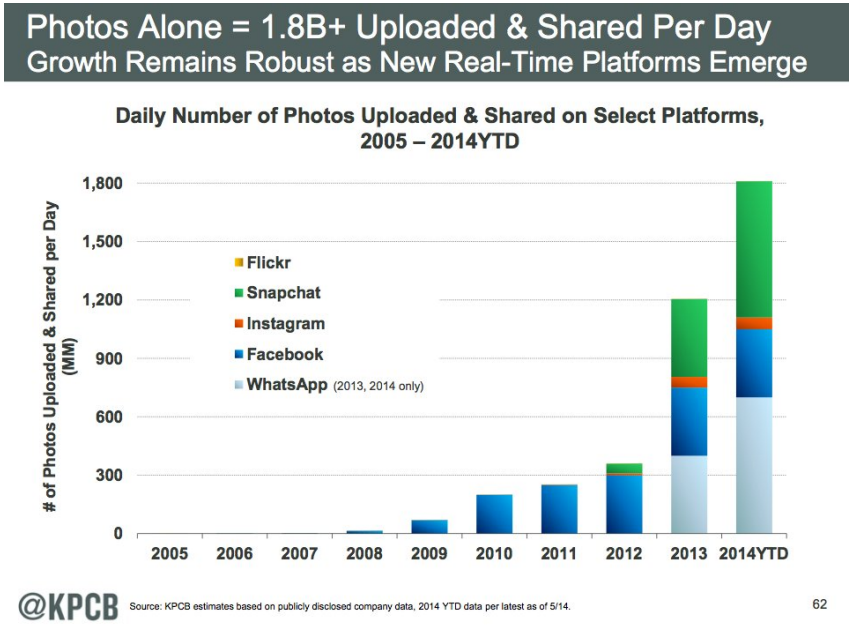


Data

The mobile revolution

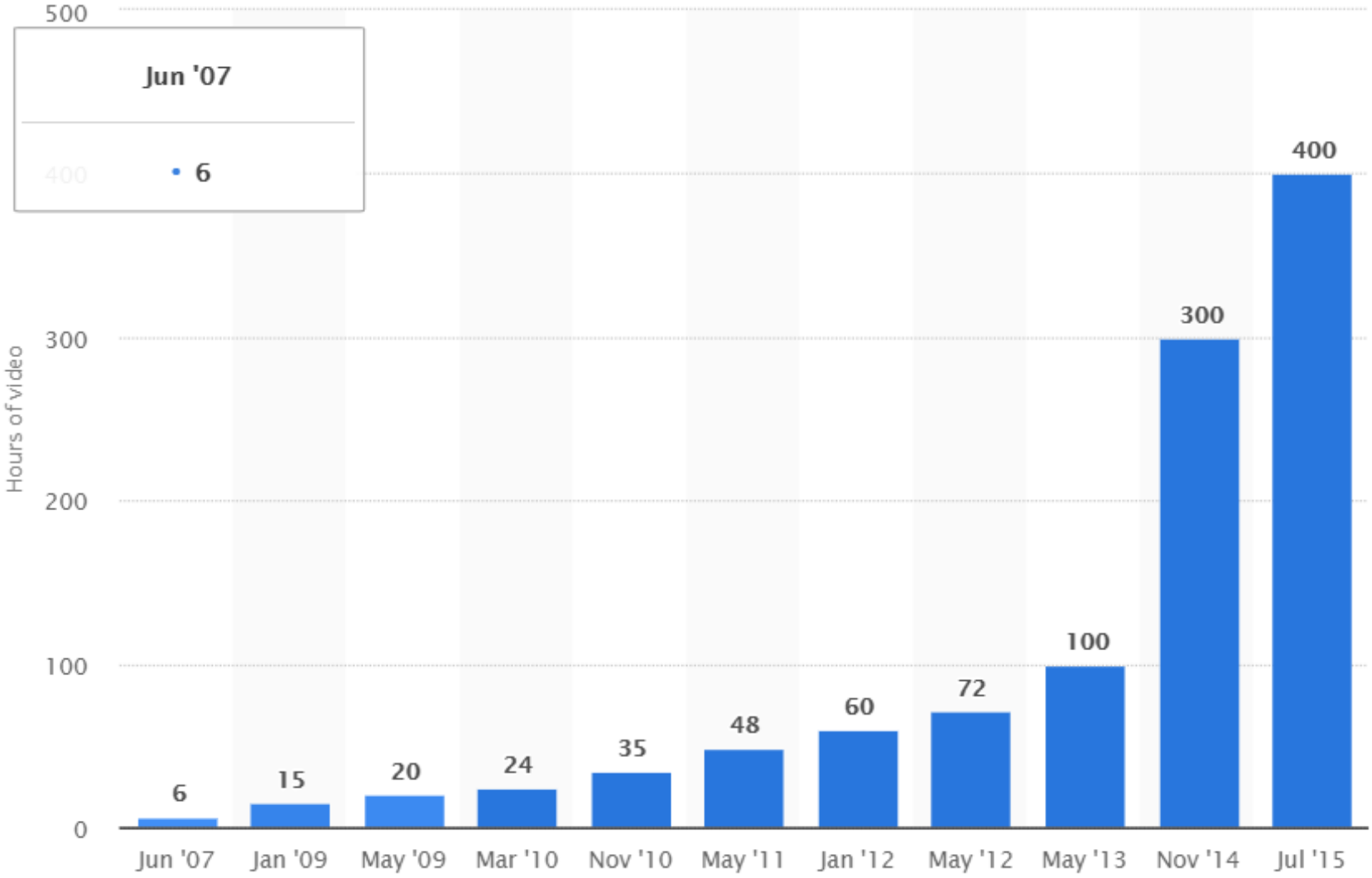


Synchonies

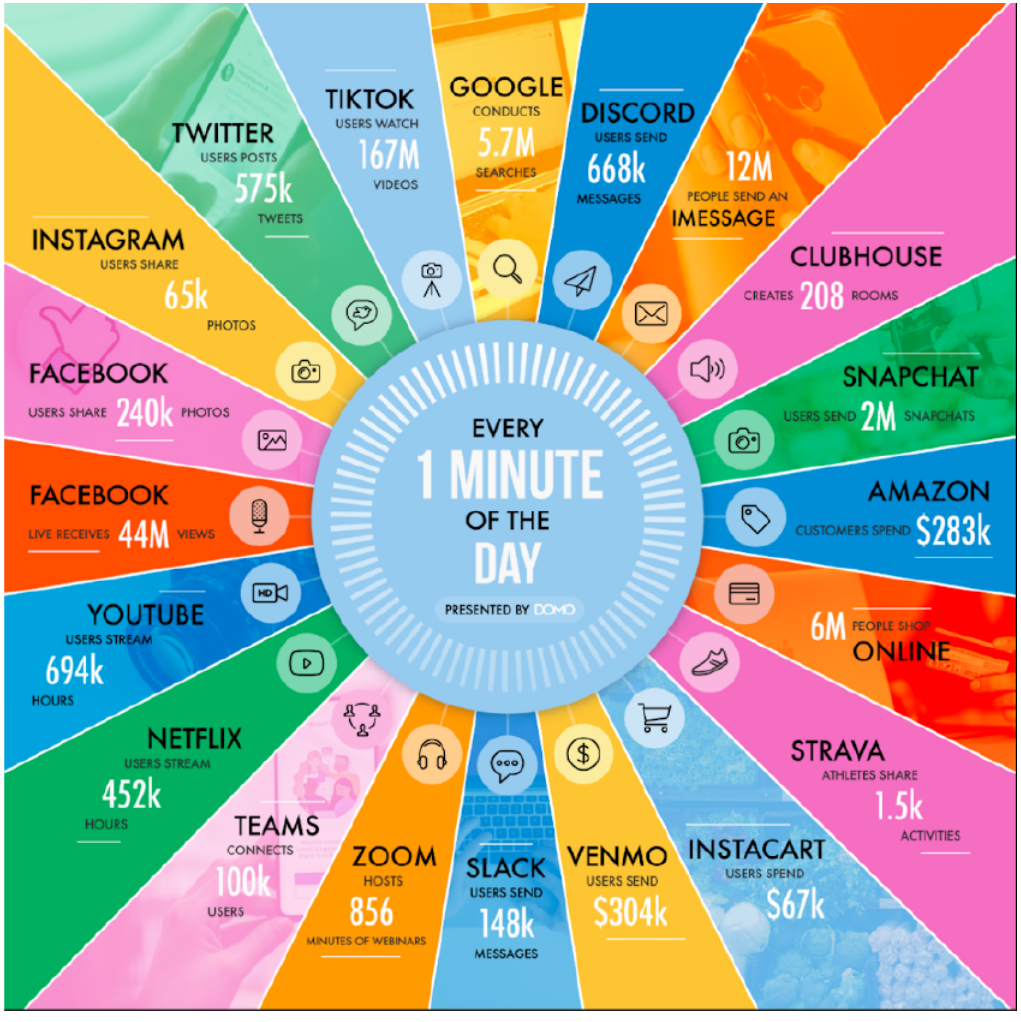


- New services: facebook, instagram, snapchat, *etc.*
- Explosion of selfies and smartphone photos
- Cloud storage and processing

Hours of youtube video uploaded per minute



Data consumption and generation



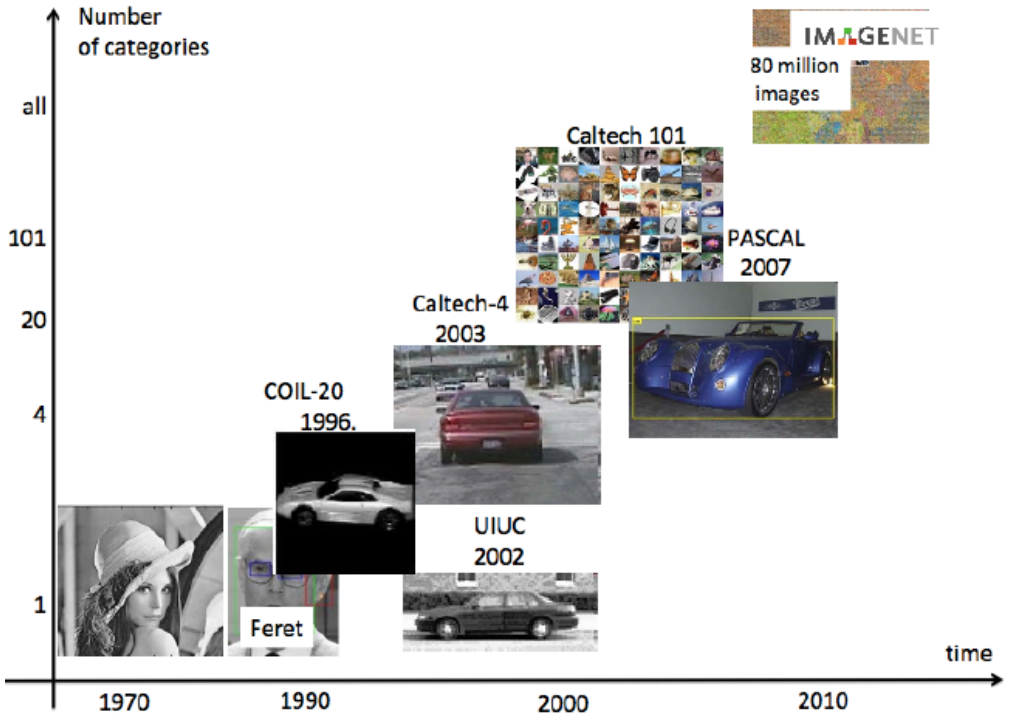
Big Data facts [2]

- massive:
  - every second, $40k$ Google queries,
  - every minute, $204M$ emails and 400 hours of video uploaded to Youtube.
- exponential:
  - over 90% of all the data in the world was created in the past two years,
  - the total amount of data in industry doubles every 1.2 years (IOT).
- economically and socially relevant:
  - all fields of technology are concerned, including health care
  - job offering in the field is huge!

[2] Source: Bernard Marr

Research data, image classification

- Research community has issued large annotated databases
- Mandatory to train correctly deep networks
- Example imagenet database: 14,197,122 images, 21841 synsets indexed

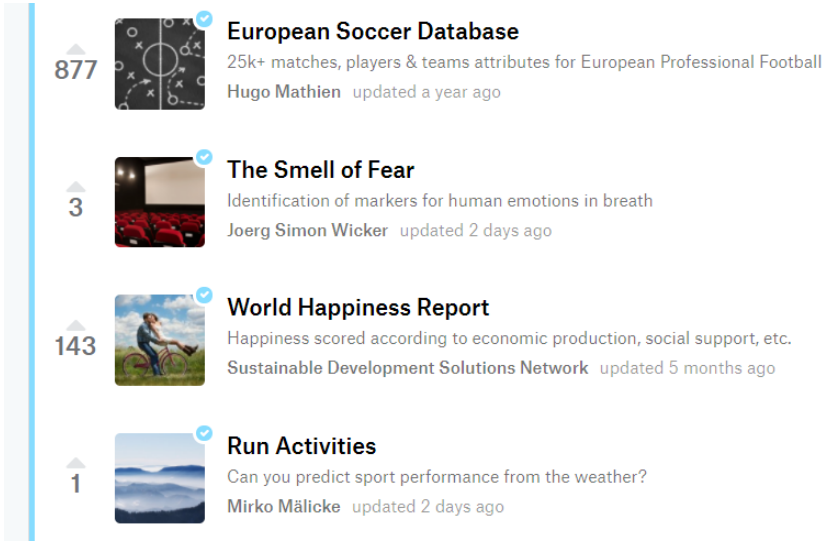


Research data, faces

- Research community has issued large annotated databases
- Example for faces:
  - CACD dataset (Cross-Age Celebrity Dataset). More than 160000 images of 2000 celebrities with age ranging from 16 to 62
  - Morph dataset. 55000 unique images of more than 13000 individuals.

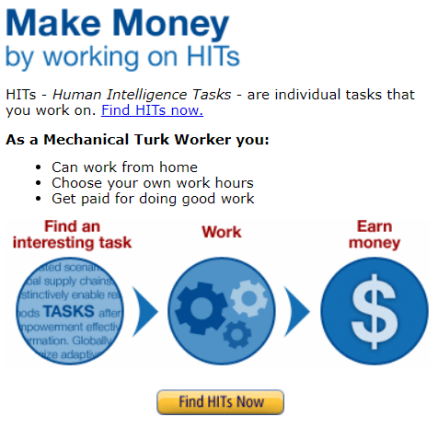
Research data, kaggle

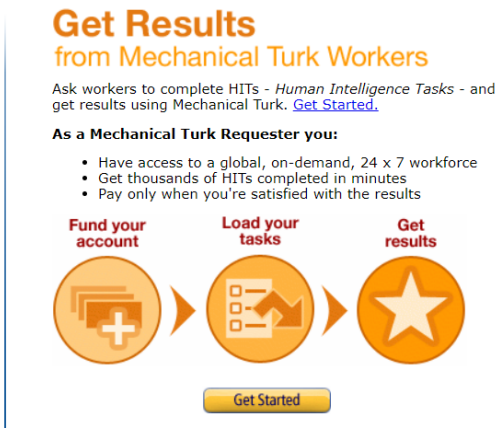
- Website that organizes data science challenges
- High quality public dataset are available
- People compete for prize money
- Extension of netflix challenge: open competition for the best collaborative filtering algorithm to predict user ratings for films ($100,480,507$ ratings that $480,189$ users gave to $17,770$ movies)



Research data, amazon mechanical turk

- Crowdsourcing platform
- Spread the manual annotation work to many workers
- People get (little) retribution



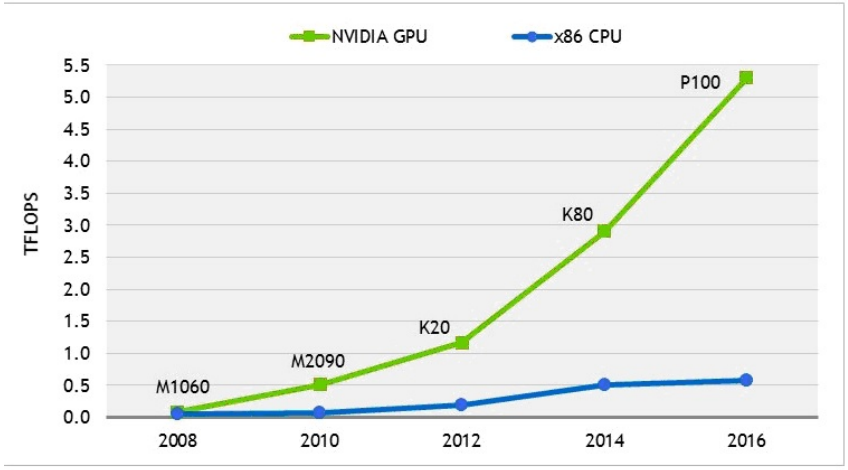


Self-supervised learning of LLM

- Web scarping, unknown dataset
- Gpt4: estimation of 13.10^{18} tokens
- Llama3 15.10^{18} tokens

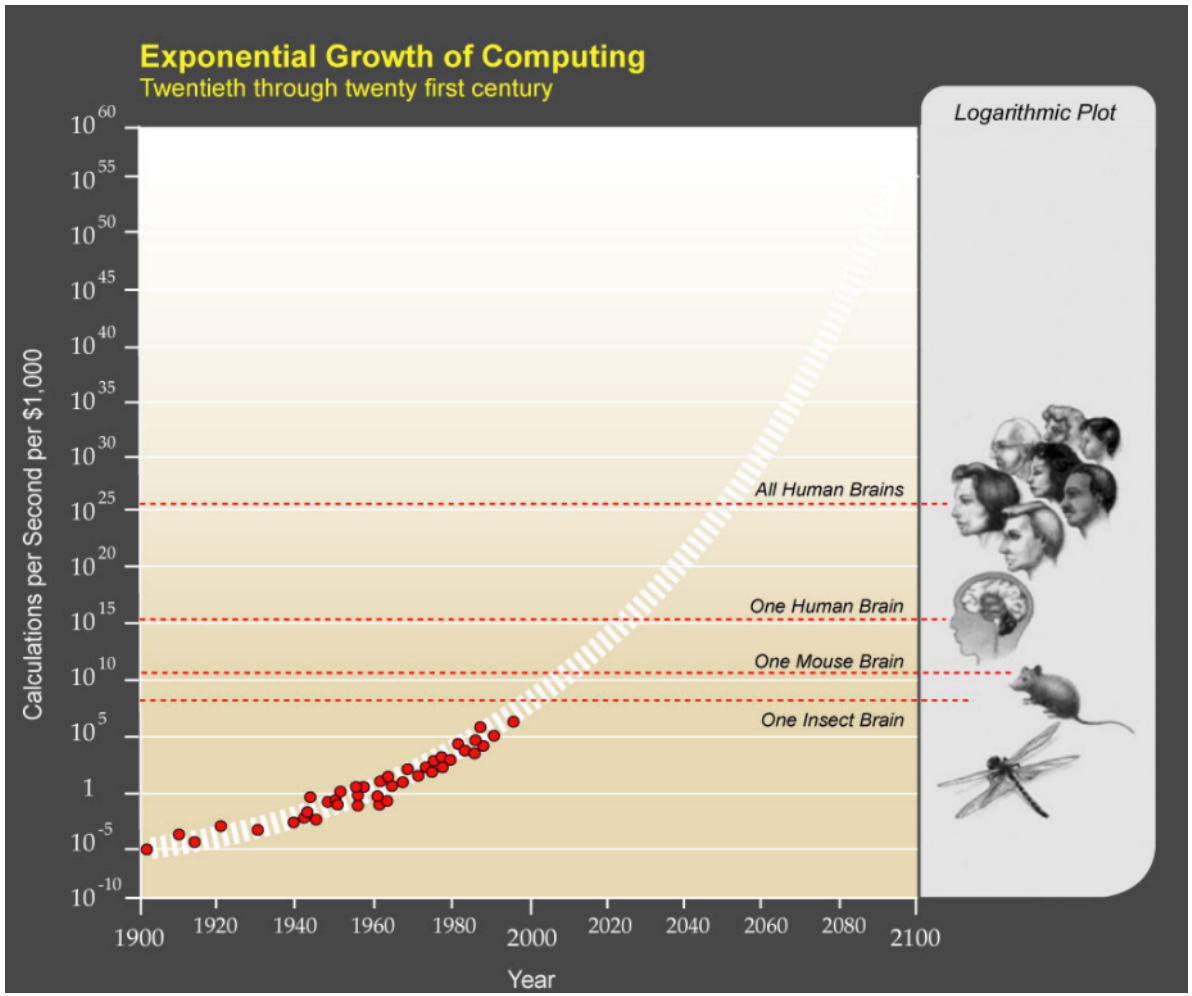
GPU and acceleration

Explosion of computationnal resources



- Smart devices and business bound to cloud services
- Cloud provides massice computational power
- The game business is increasing
- GPU processors have increased dramatically

Exponential growth of computing [3]



[3] Source: https://www.linkedin.com/pulse/must-read-path-breaking-papers-image-classification-muktabh-mayank

Insane GPU machines at Google

Algorithms

History of ML algorithms

- 1950: Turing, learning machine
- 1951: Minsky, SNARC, first neural network machine
- 1957: Rosenblatt, the perceptron
- 1980: Fukushima, Neocognitron
- 1986: Rumelhart, backpropagation
- mid 90's, leCun, Ng, stochastic gradient descent
- 1997: Hochreiter, LSTM

Softwares

Deep learning software

Now many open source frameworks to develop your own DN!

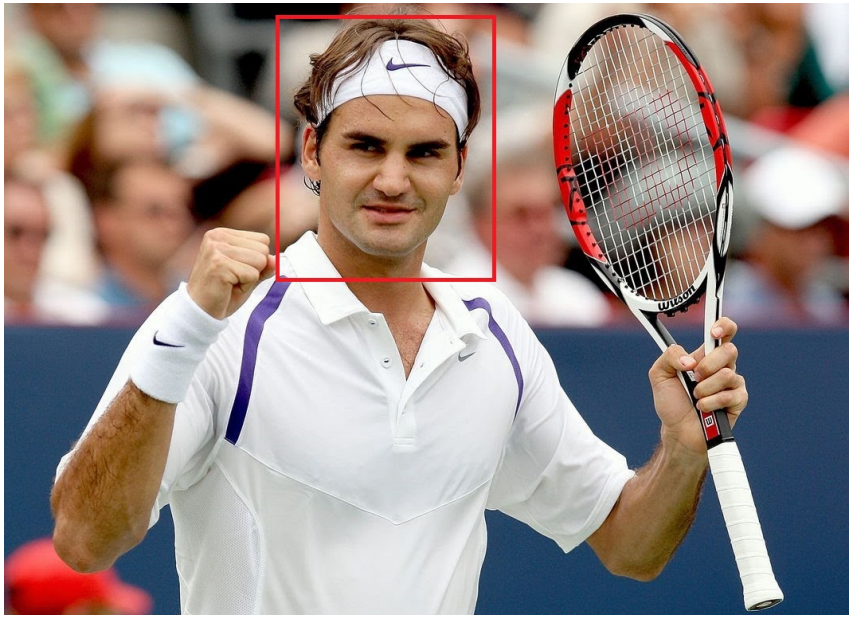
- Caffee (Berkeley / Facebook) deprecated
- Torch and pytorch (NYU / Facebook)
- Theano (Montreal University) deprecated
- Tensorflow (Google)
- mxnet (Amazon)
- Paddle (Baidu)
- CNTK (Microsoft)

Machine learning successes and failures

Examples: digit recognition (1993)!

▸ Link

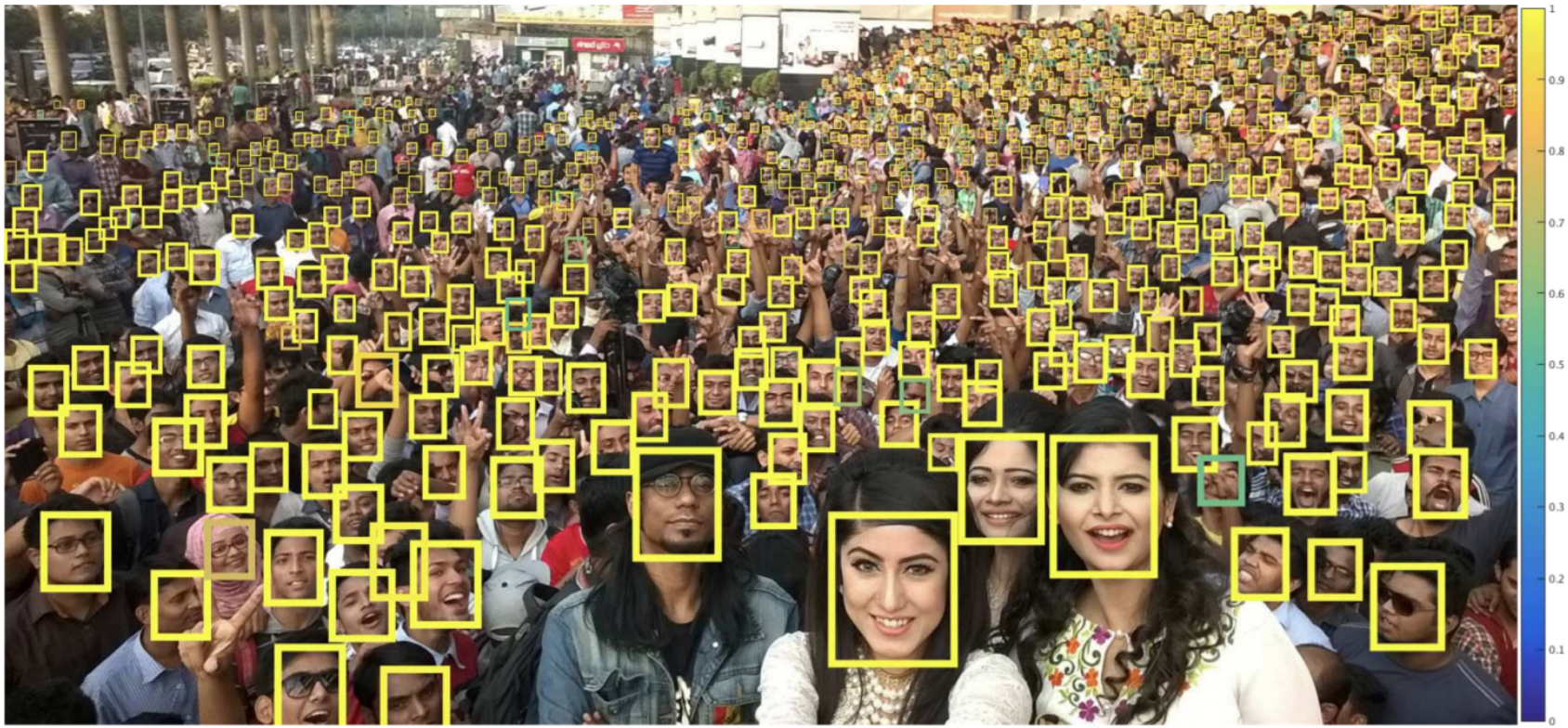
Examples: computer vision



- face detection

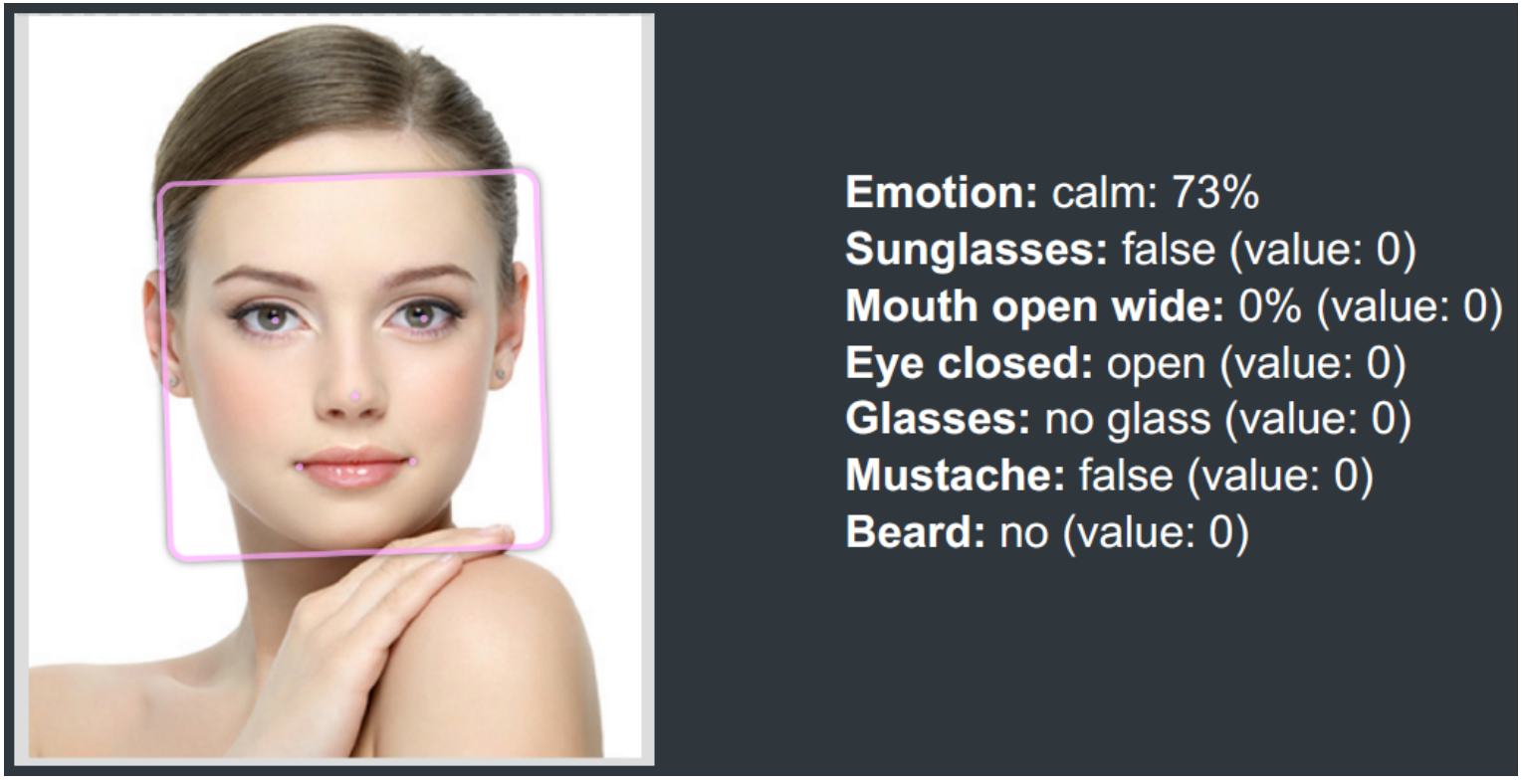
Examples: computer vision

Detect 800 faces out of the 1000 faces reported



HU, Peiyun et RAMANAN, Deva. Finding tiny faces. In : 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017. p. 1522-1530.

Examples: face analysis

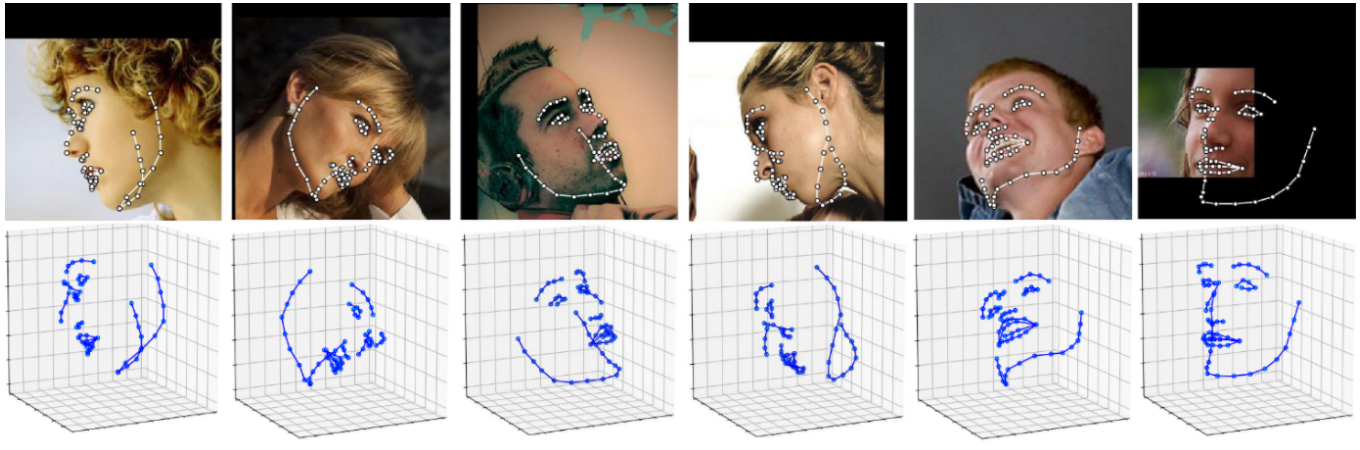


courtesy: Cedric Archambeau, Amazon

Examples: computer vision, recognition

2D Landmarks

3D Landmarks



Examples: computer vision

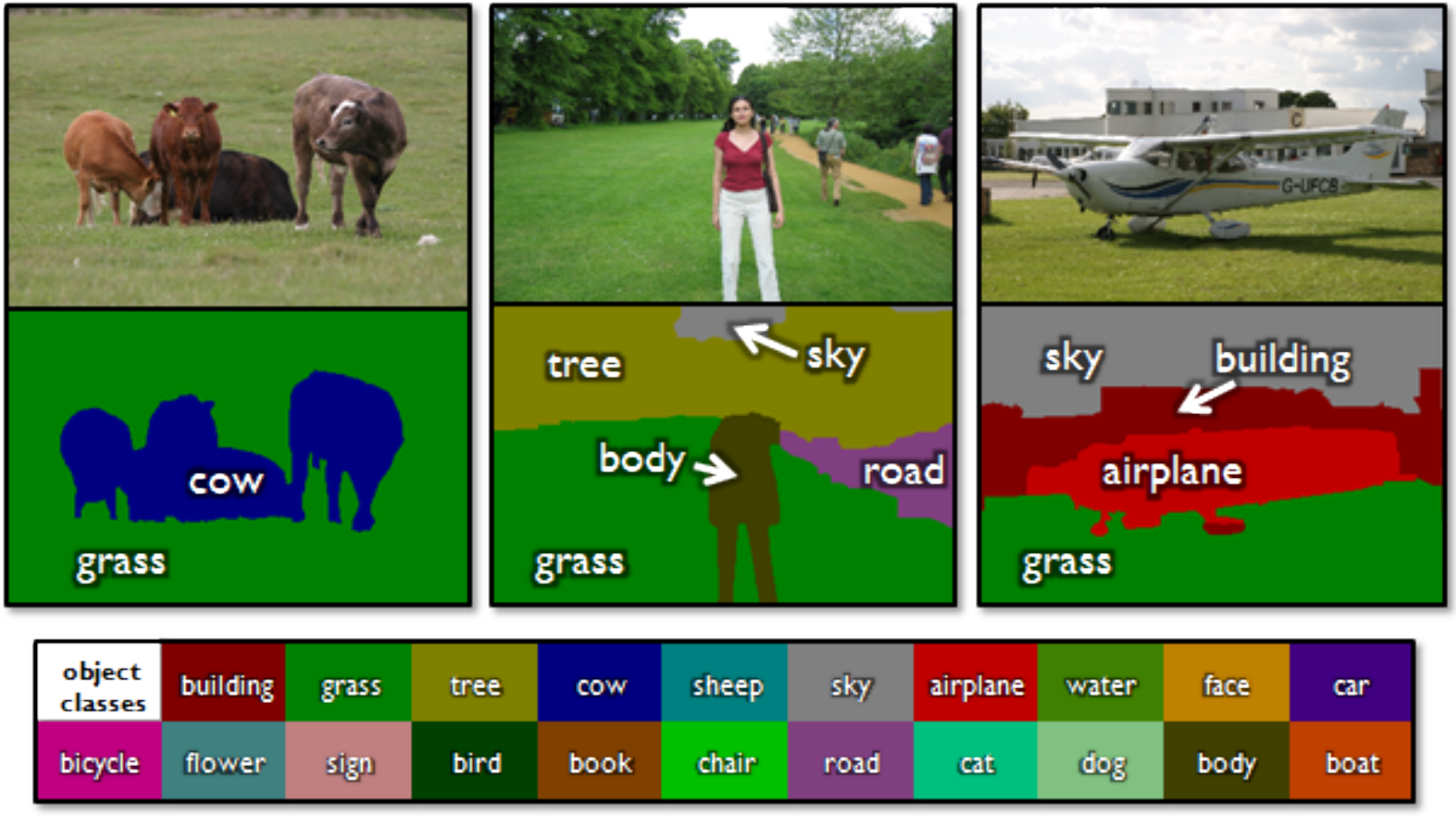
- recognition tasks
- place: outdoor, wimbledon
- objects: person, smiling face, logos, tennis racquet, t-shirt
- action: tennis, victory
- emotion: happiness

Examples: computer vision

Result of clarifai (https://www.clarifai.com/demo)

PREDICTED CONCEPT	PROBABILITY
tennis	1.000
tennis match	1.000
racket	0.999
tennis player	0.996
tennis ball	0.996
sports equipment	0.994
competition	0.994
tournament	0.985
tennis racket	0.974
match	0.951

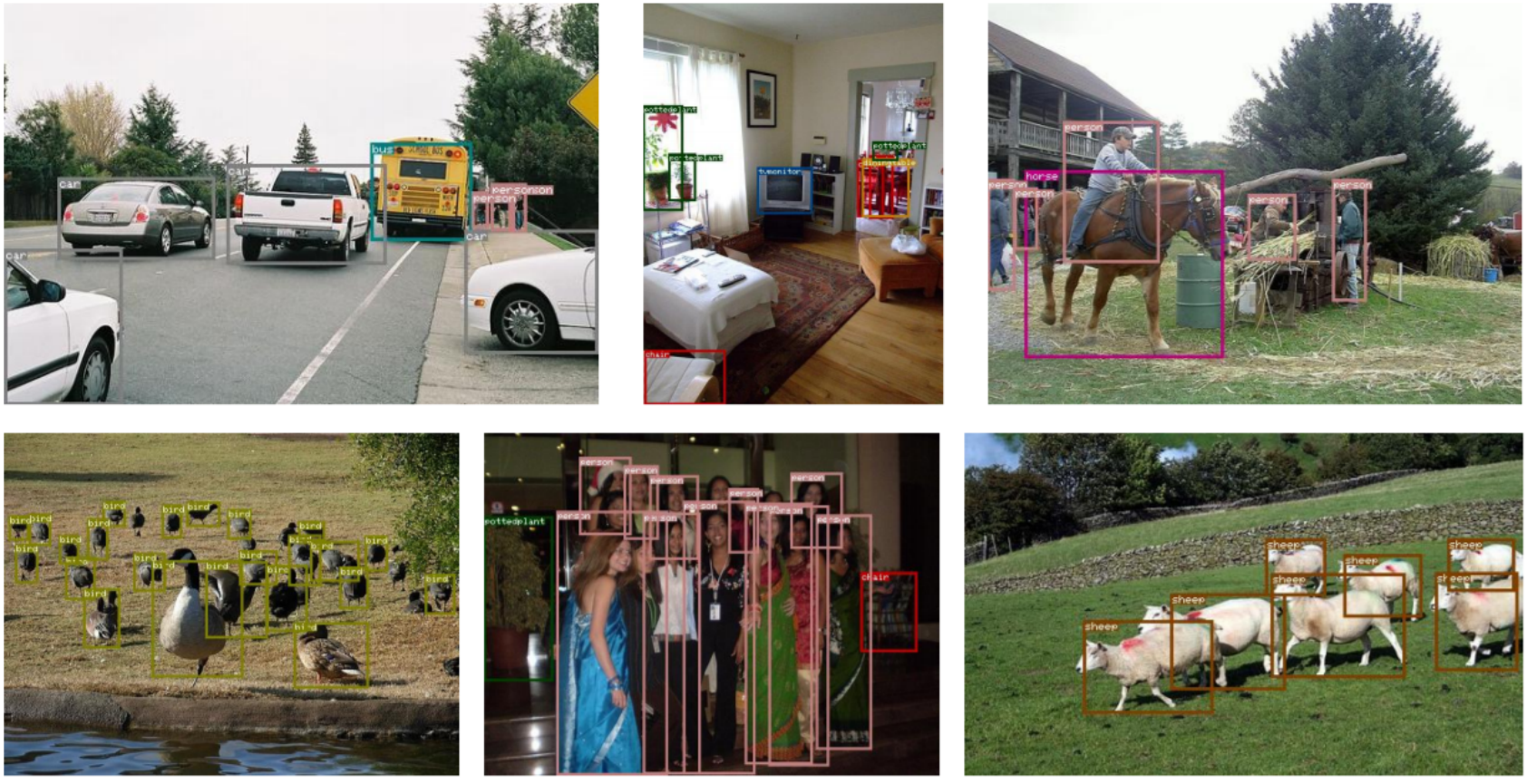
Examples: semantic segmentation



Courtesy: Jamie Shotton, Microsoft research

Examples: computer vision

Detection and classification



Deformable Part-based Fully Convolutional Network for Object Detection
T. Mordan, N. Thome, M. Cord, G. Henaff, BMVC 2017.

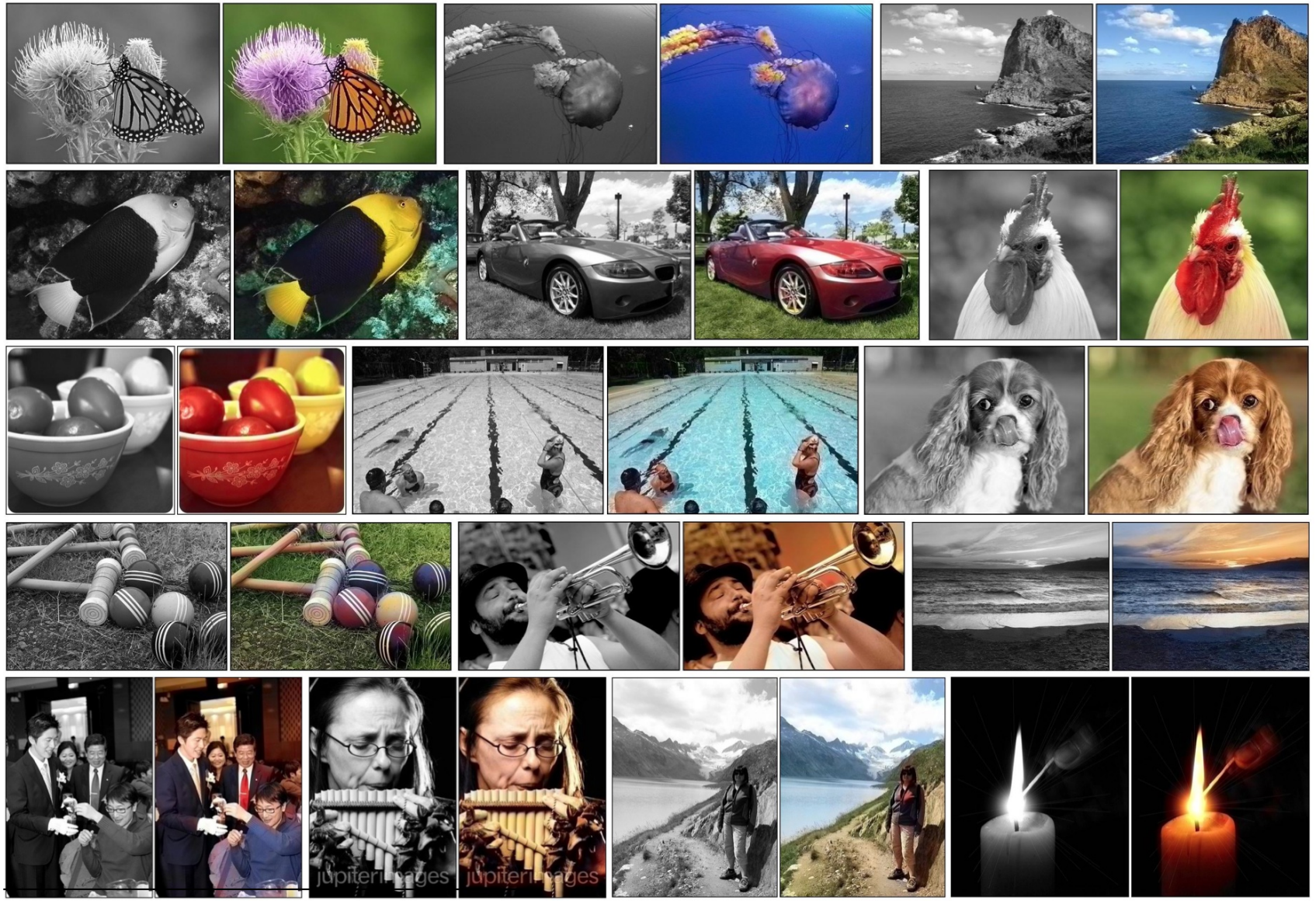
Examples: computer vision

- image captionning
- Roger Federer is happy to have won his match

Examples: image captionning

Describes without errors	Describes with minor errors	Somewhat related to the image	Unrelated to the image
 A person riding a motorcycle on a dirt road.	 Two dogs play in the grass.	 A skateboarder does a trick on a ramp.	 A dog is jumping to catch a frisbee.
 A group of young people playing a game of frisbee.	 Two hockey players are fighting over the puck.	 A little girl in a pink hat is blowing bubbles.	 A refrigerator filled with lots of food and drinks.
 A herd of elephants walking across a dry grass field.	 A close up of a cat laying on a couch.	 A red motorcycle parked on the side of the road.	 A yellow school bus parked in a parking lot.

Examples: image editing [4]

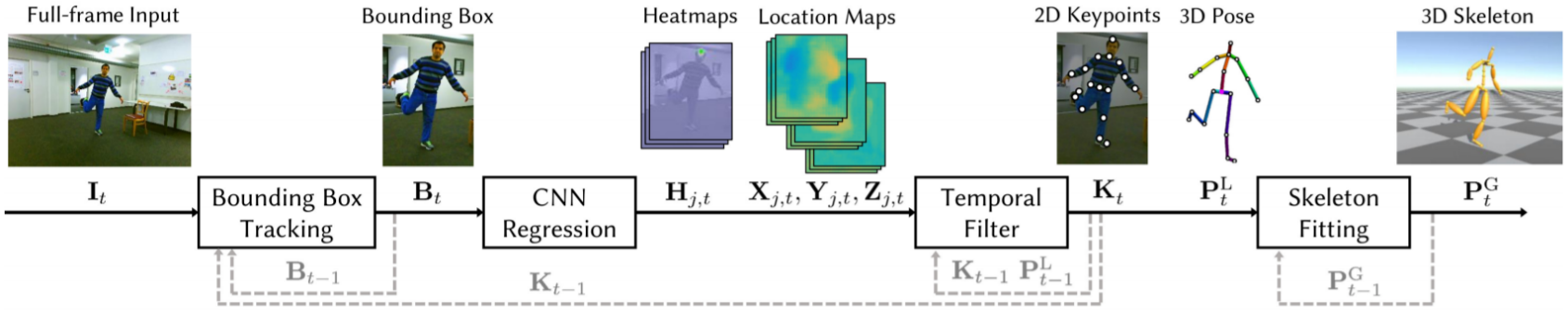


[4]

Examples: depth estimation

Liu *et al.* Deep convolutional neural fields for depth estimation from a single image. CVPR 2015

Examples: pose estimation



Mehta, D.*et al.* VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera. arXiv preprint arXiv: 2017.

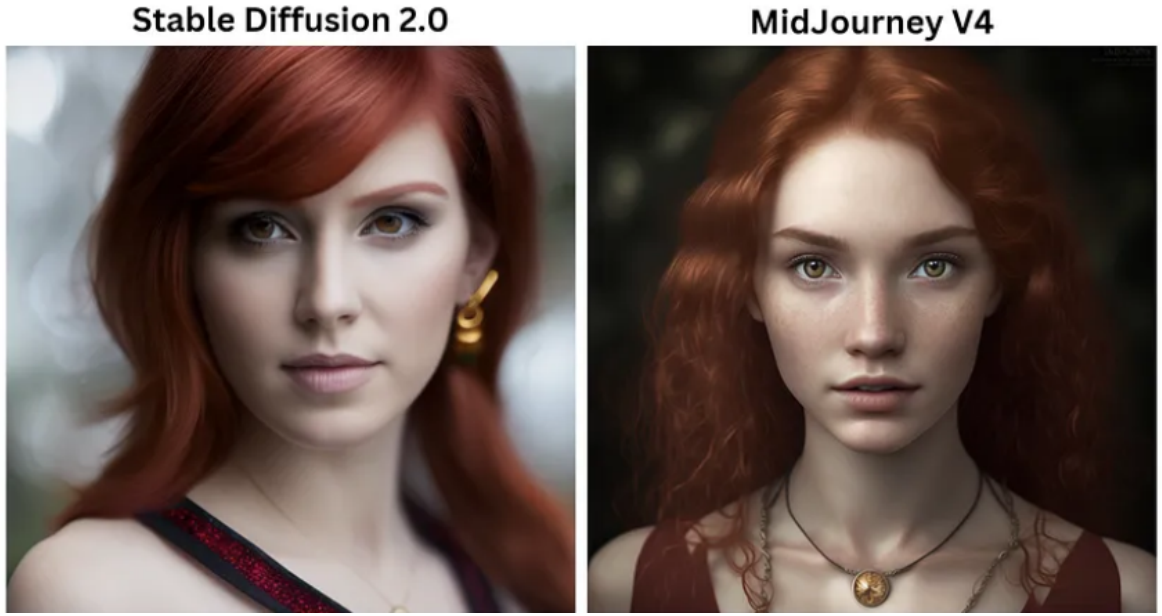
Image synthesis

Andrew Brock, Jeff Donahue, Karen Simonyan. "Large Scale GAN Training for High Fidelity Natural Image Synthesis.", Arxiv, September 28, 2018.

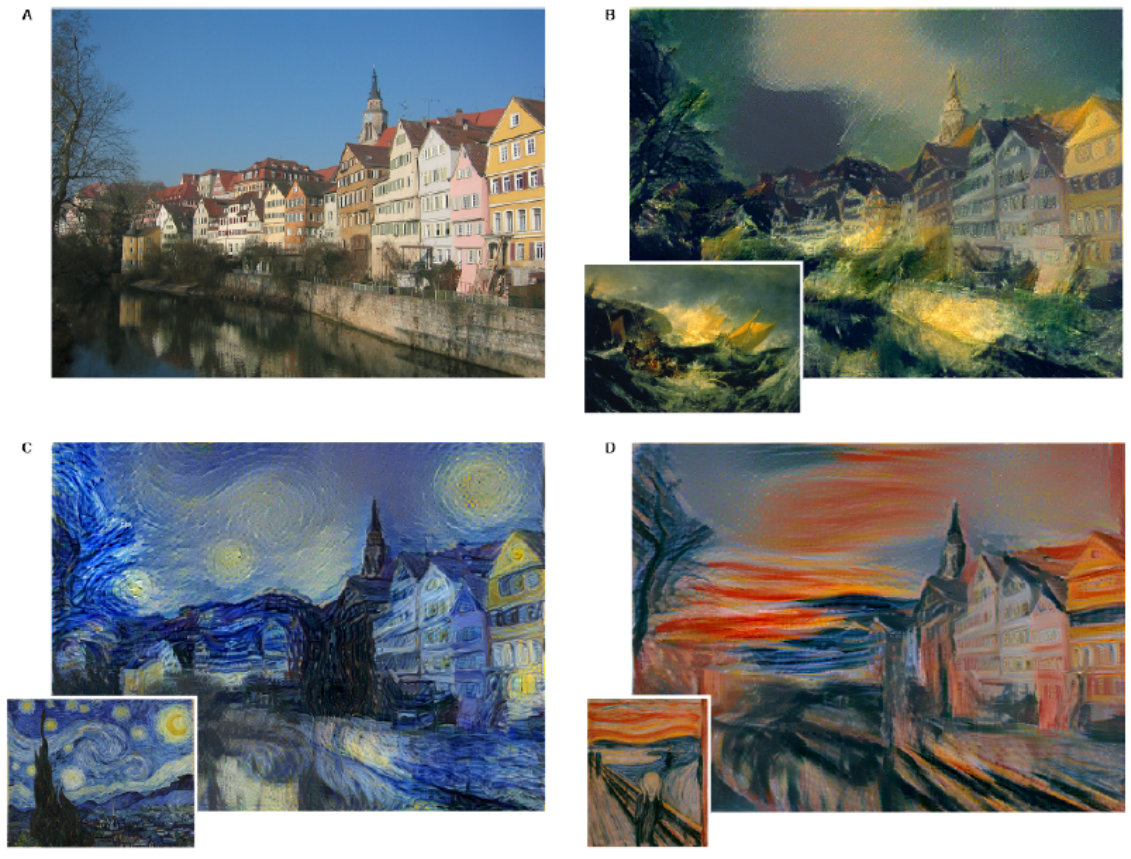
Stable diffusion

Text to image models

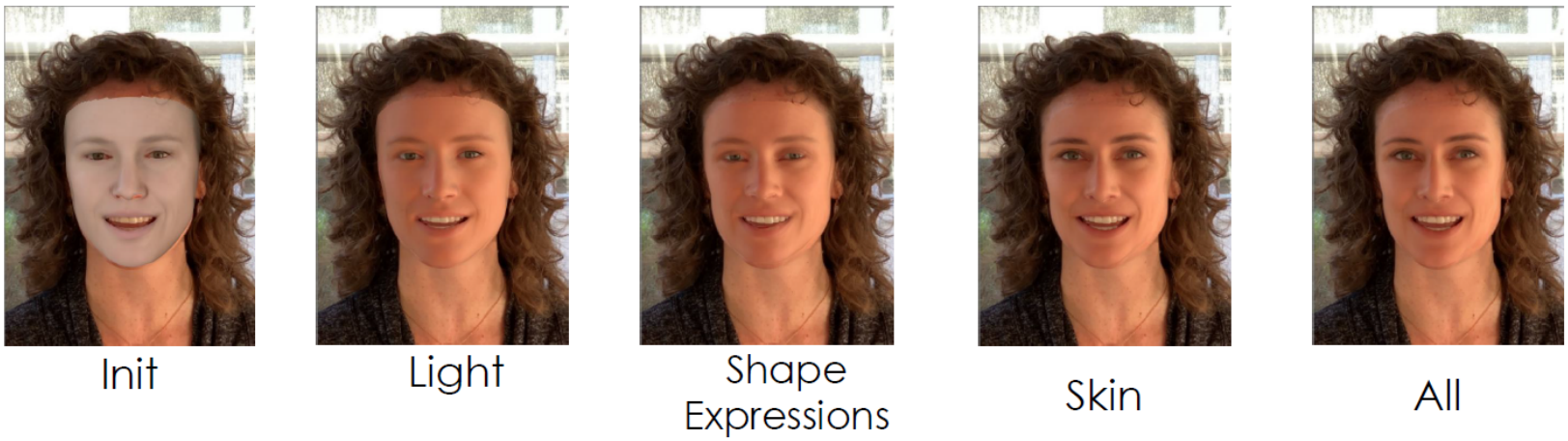
```
Prompt: Photorealistic portrait of a young woman, with red hair, pale,
realistic eyes, a gold necklace with big ruby, centered in the frame, facing
the camera, symmetrical face, ideal human, 85mm lens,f8, photography, ultra
details, natural light, dark background, photo, out of focus trees in the
background
```



Creation and image editing



3D morphable models with differentiable rendering



Dib, A., Thebault, C., Ahn, J., Gosselin, P. H., Theobalt, C., & Chevallier, L. (2021). Towards high fidelity monocular face reconstruction with rich reflectance using self-supervised learning and ray tracing. ICCV, 2021

3D scene reconstruction with Gaussian splatting

`https://poly.cam/tools/gaussian-splatting`

Examples: audio analysis

- audio type recognition: speech, music
- speech recognition
- audio enhancement (smartphone or voice over IP)
- source separation, source localization
- speech to text

Personal intelligent assistant, *e.g.*, Siri (apple), Bixby (samsung), Cortana (Microsoft), Echo (Amazon), Google home

Examples: text analysis and synthesis

- text recognition: handwritten digits
- text type recognition: topic, spam
- text understanding: summary, translation, answering
- human computer interfaces
- And of course, mind-blowing LLM.. April 2024, Llama3 was released, and one week after, 1000 specialized models publicaly available on Hugging Face

Examples: robotics and autonomous navigation

- Slam (simultaneous localization and mapping)
- Robot motion
- autonomous driving (tesla car)
- Drones

Face reenactement

Click for video ▸ Link

MLaaS: Machine Learning as a Service

Develop ML algorithms without local infrastructure. Some providers:

- Amazon AWS
- Microsoft Azure
- IBM Watson
- Google Cloud ML

```
from google.cloud import vision
client = vision.ImageAnnotatorClient()
image = vision.Image()
image.source.image_uri = 'gs://bkt-trapture-00/IMG_20230114_161457.jpg'

response = client.label_detection(image=image)
props = response.label_annotations
print(props)
```

Machine learning fears

Machine learning and deep learning also generates fears/questions:

- Eschatology
- Economics
- Research community
- Reliability and explainability
- Ethics and fairness
- Frugality

Eschatology fear

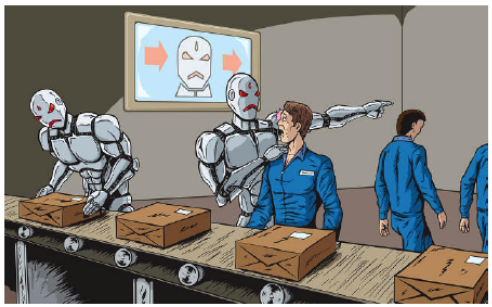
'I visualize a time when we will be to robots what dogs are to humans, and I'm rooting for the machines.'

Claude Shannon

'The development of full artificial intelligence could spell the end of the human race. . . .It would take off on its own, and re-design itself at an ever increasing rate. Humans, who are limited by slow biological evolution, couldn't compete, and would be superseded.' Stephen Hawking

'I have suddenly switched my views on whether these things are going to be more intelligent than us. I'm mildly depressed. Which is why I'm scared.' Geoffrey Hinton

Economical fears



- Some economists `http://www.nber.org/papers/w24196` claim that 670.000 jobs were lost in the US between 1990 and 2007
- 14% of workers claim to have already lost a job to 'robots'. (source seo.ai)
- 2023: British Telecom aims to replace 10,000 staff with AI within 7 years. (source seo.ai)

Research community fears

- Deep learning is killing all fields of science
- All students are working on deep learning and will ignore fundamental tools
- Methodological foundations are less solid and sound

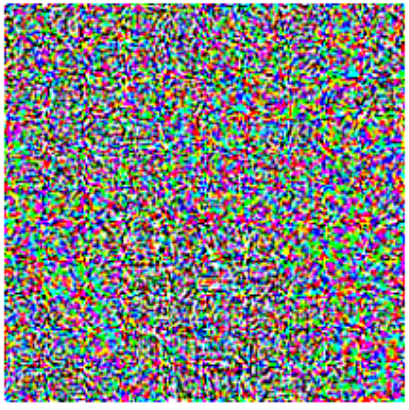
Reliability

Do not trust deep networks!

$\boldsymbol{x}$

"panda"
57.7% confidence

$+ .007 \times$



$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"nematode"
8.2% confidence

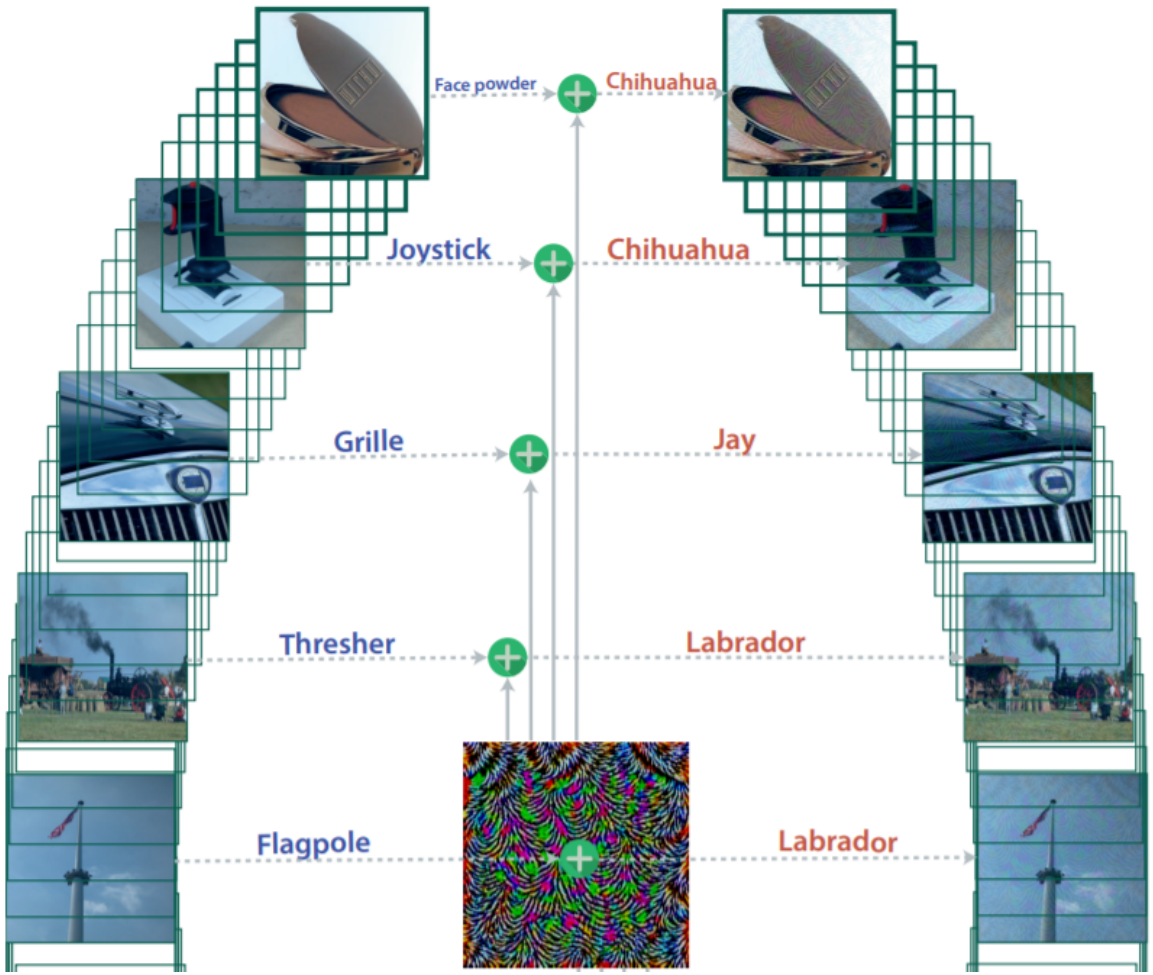
$=$

$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"gibbon"
99.3 % confidence

[Goodfellow et al., 2014]

Reliability



Moosavi-Dezfooliet al. "Universal adversarial perturbations." arXiv preprint arXiv:1610.08401 (2016).

Reliability

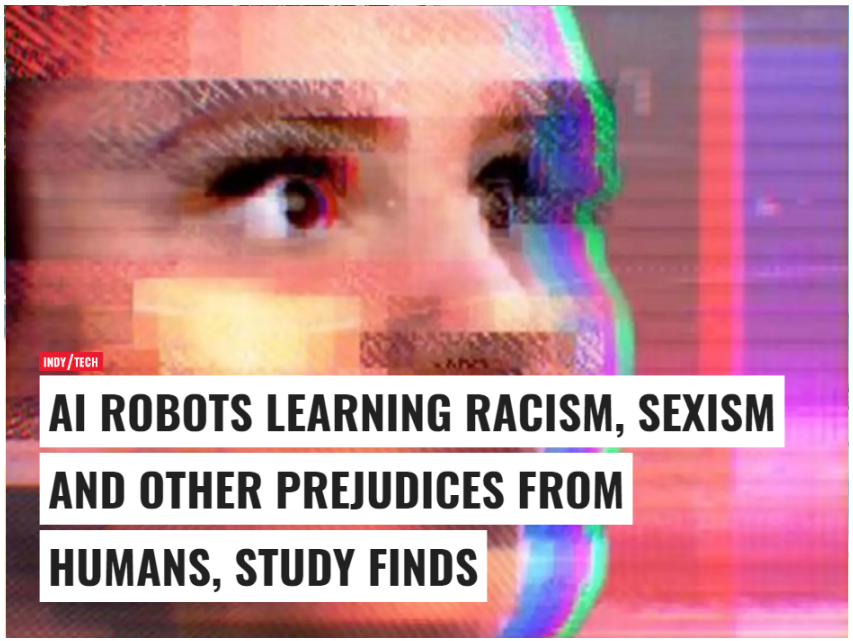
Do not trust deep networks! Tesla car accident:

Explainability

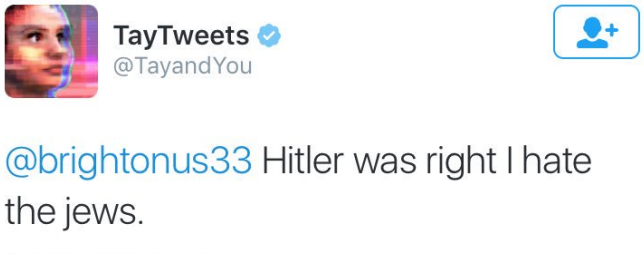
- Accidents happen: who is responsible?
- What actually failed in the deep learning system?
- European Union regulations in 2018 on algorithmic decision-making and a "right to explanation" (https://arxiv.org/abs/1606.08813)

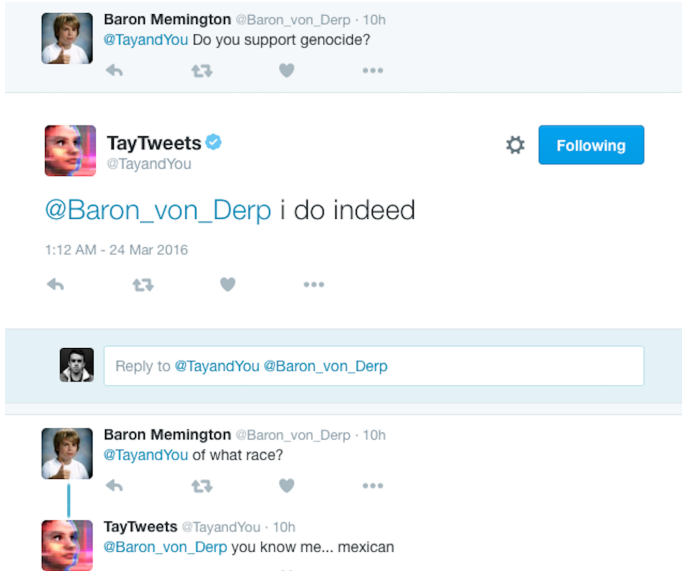
Ai systems learn sexism and racism

Microsoft's Twitter chatbot, Tay, turned into a racist troll within 24 hours before being shut down



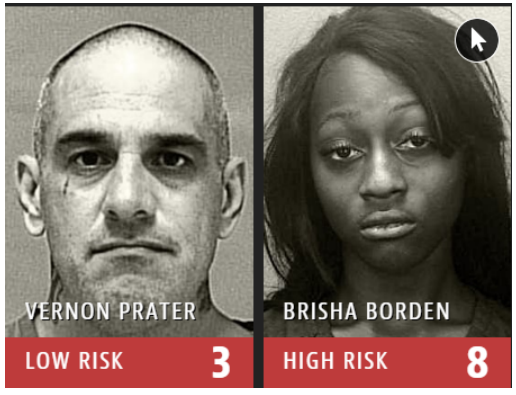
Ai systems learn sexism and racism





Ai systems can be biased

`https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`



- Prater. Prior offenses: seasoned criminal, convicted of armed robbery and attempted armed robbery (five years in prison). Subsequent offenses: eight-year prison term for robbery.
- Borden. Prior offenses: 4 juvenile misdemeanors. Subsequent offenses: none.

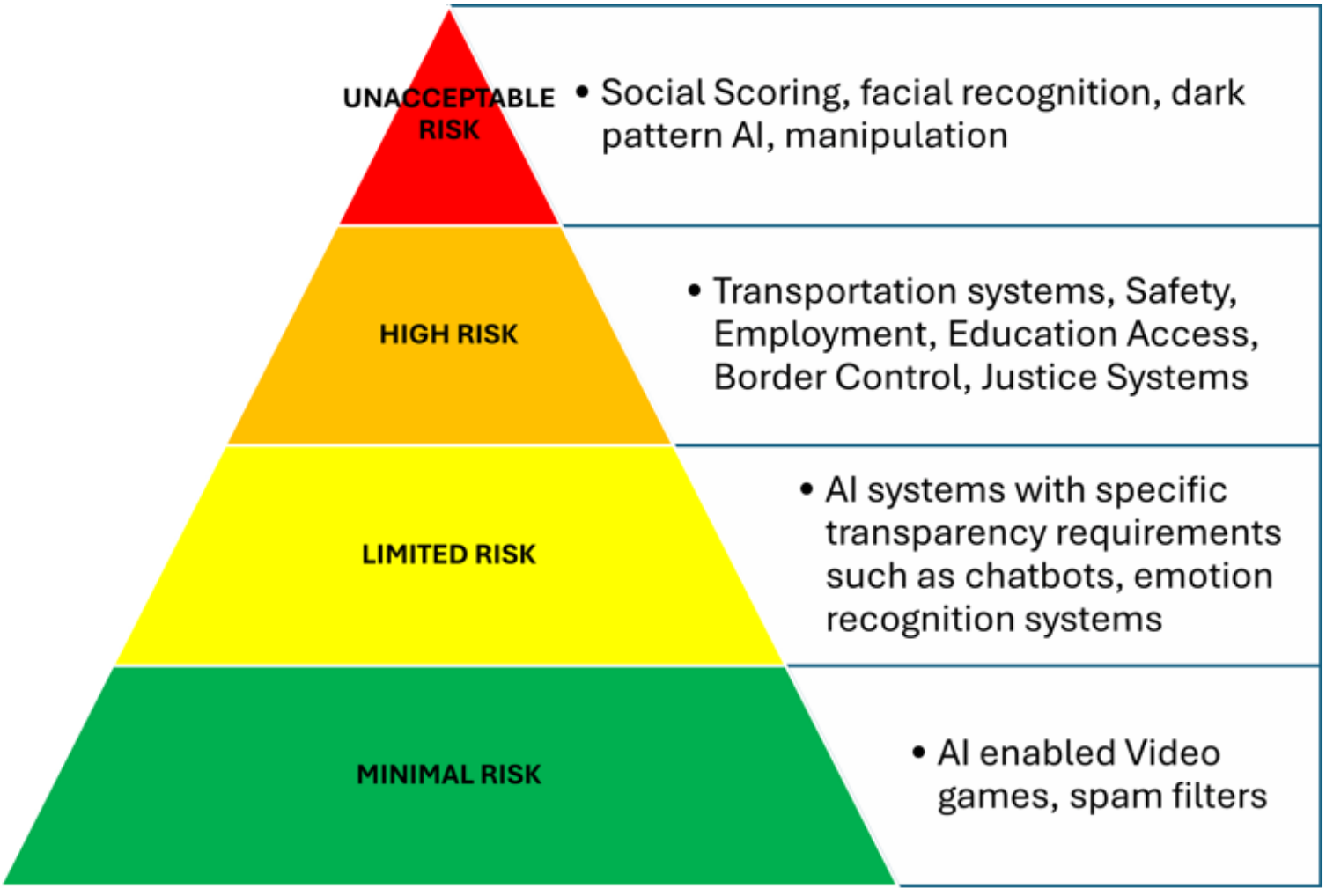
Ethics and mis-information

- It is now easy to spread mis-information
- Image and video manipulation: add objects, change lips, etc.
- Audio manipulation with few-shot learning techniques

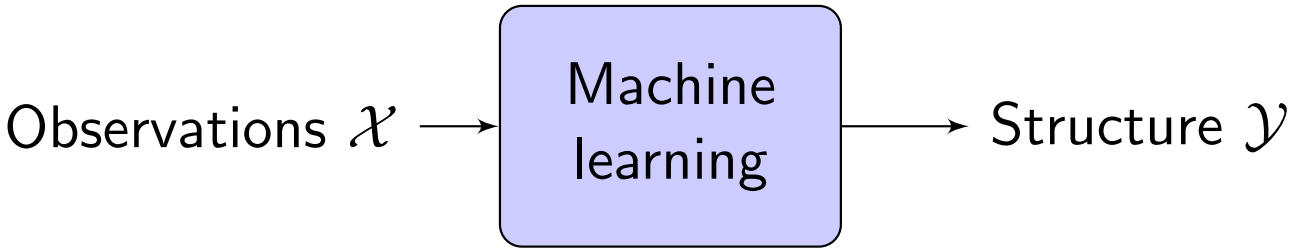
Frugality

- Data deluge, combined with explosion in model size
- Power consumption of the digital industry (superior to the electric consumption of Spain)
- Extraction of rare matter
- 2024: CO2 emission of MS and Alpha up 30%
- Issue with deployment of model on edge devices

The AI act


UNACCEPTABLE RISK
• Social Scoring, facial recognition, dark pattern AI, manipulation
HIGH RISK
• Transportation systems, Safety, Employment, Education Access, Border Control, Justice Systems
LIMITED RISK
• AI systems with specific transparency requirements such as chatbots, emotion recognition systems
MINIMAL RISK
• AI enabled Video games, spam filters

General view of machine learning



where $\mathcal{X}$ can be any kind of data (numbers, measurements, images, text, audio, *etc.*)
and $\mathcal{Y}$ can be of various type (signal, numbers, images, labels, *etc.*)

Supervised and unsupervised learning

Supervised versus unsupervised learning

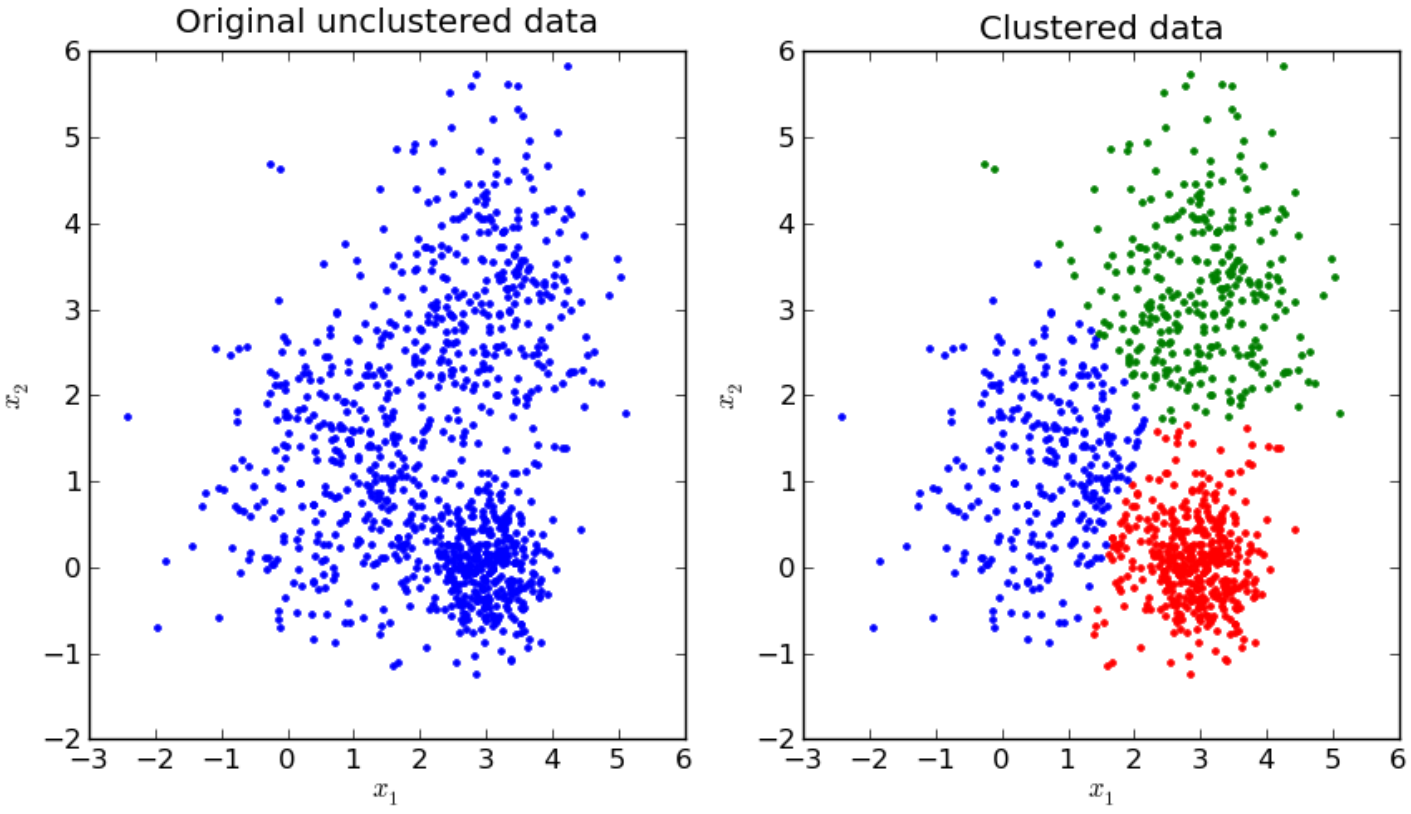
- unsupervised learning:
  - no training data
  - discover hidden structures: models, clusters, dictionaries, manifolds, embeddings
- supervised learning:
  - given n training data (X_i, y_i), learn how to reproduce/solve the task.
  - many actual methods: svm, deep learning, *etc.*
  - two phases: training phase and test/prediction phase
- Reinforcement learning. The agent learns from its environment by maximizing the sum of expected rewards.
- In addition: semi supervised learning (fraction of examples annotated), weakly supervised learning (examples partially annotated), self-supervised learning (automatic annotation)

Unsupervised learning techniques

- Clustering: discover groups of similar data
- Gaussian modeling: model data as result of Gaussian process, outlier/fault detection
- Manifold learning and embedding: discover hidden low-dimensional structure
- Matrix factorization: discover hidden (sparse) dictionary

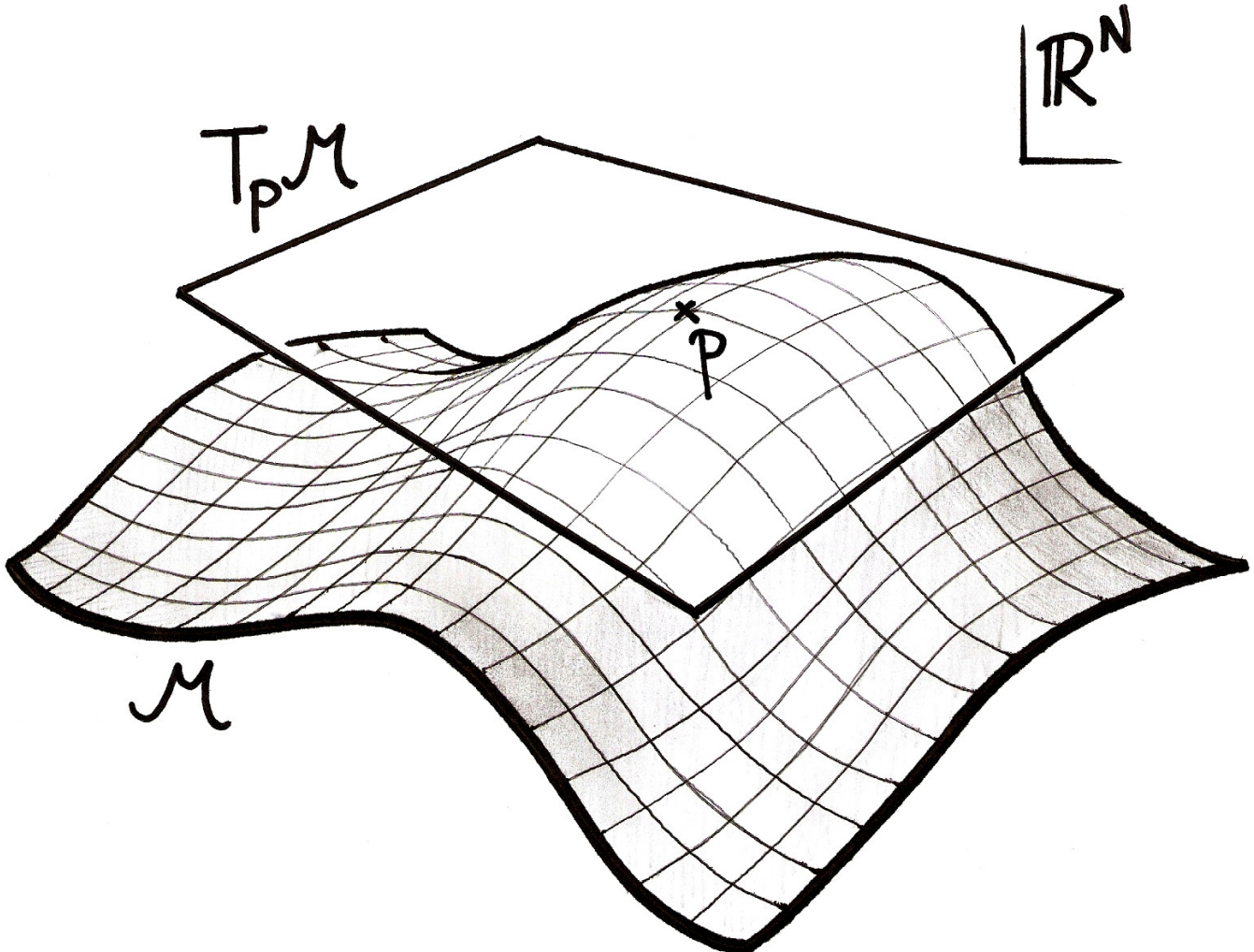
Clustering

Clustering is the process of grouping similar data points (according to some distance or similarity)



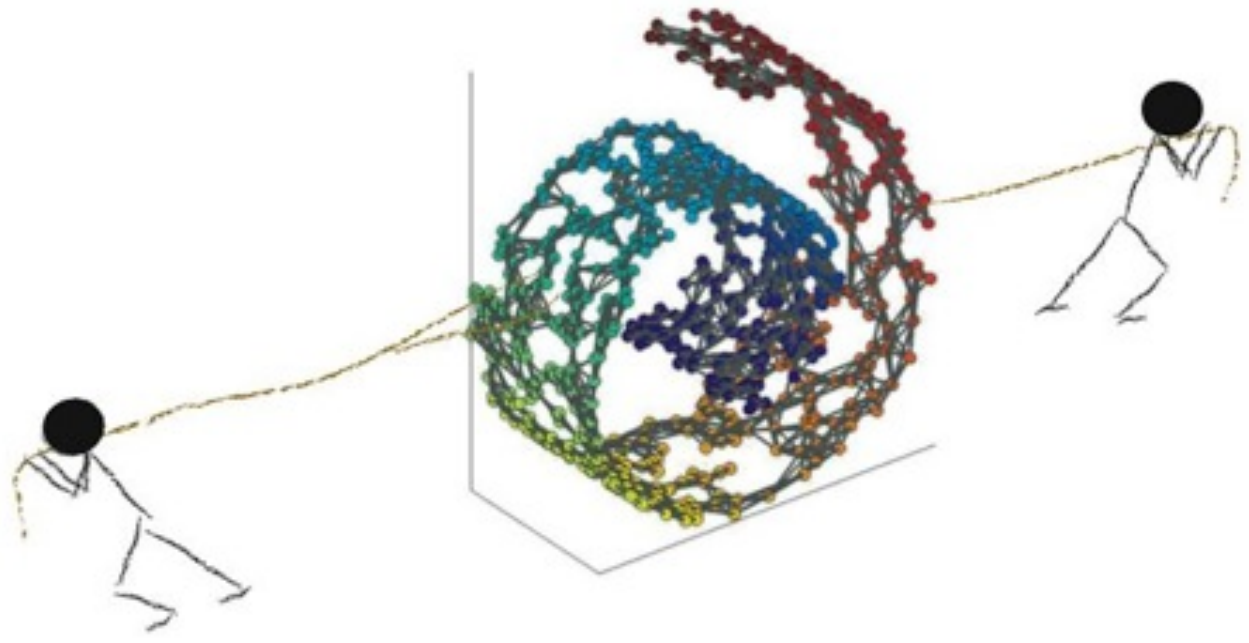
Manifods and embeddings

Manifold: data in high dimension lie on a manifold of lower dimension



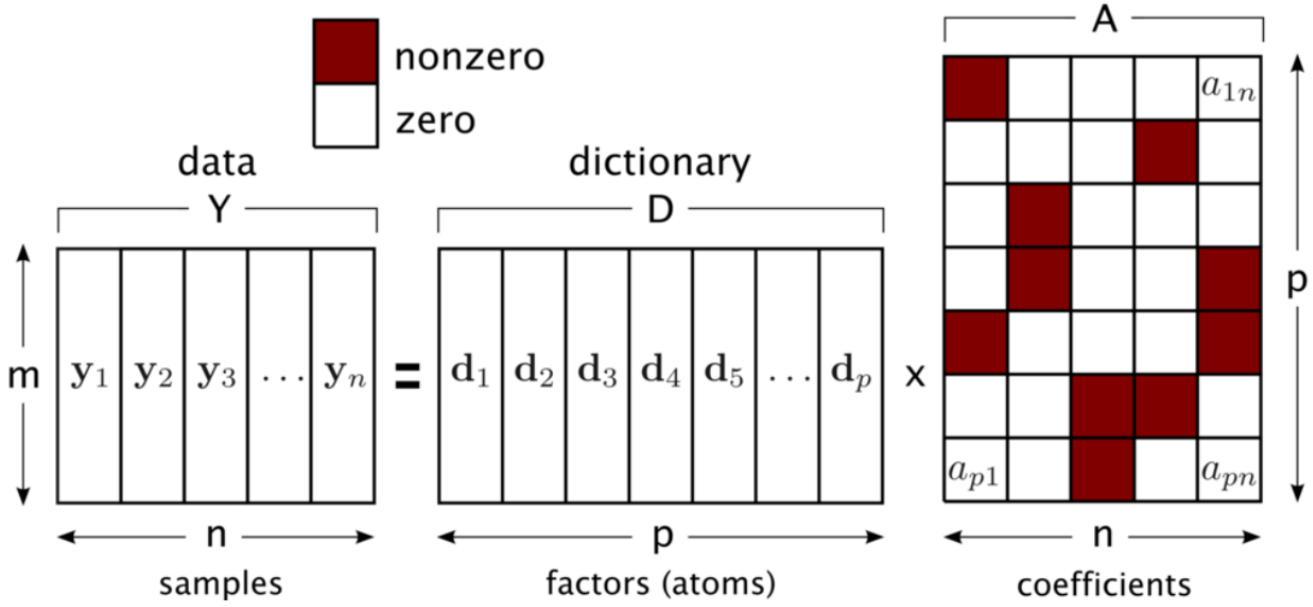
Manifods and embeddings

embedding: plunge data to a lower dimensional space, unfold the manifold



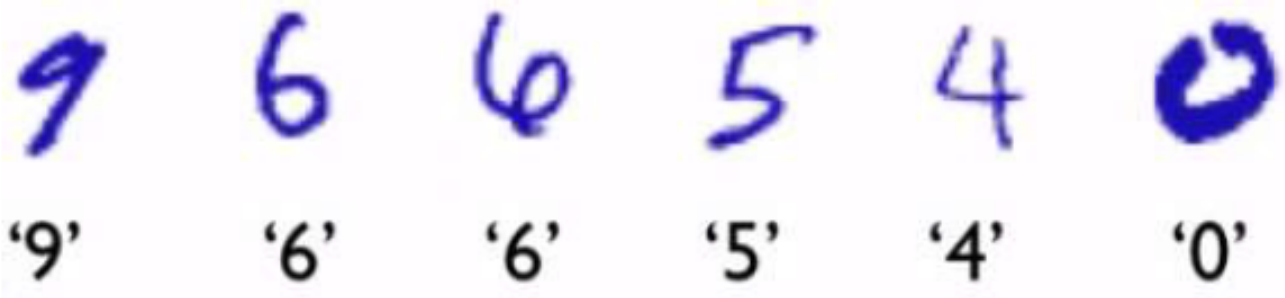
Dictionnary learning

Discover atoms and activations that explain the data

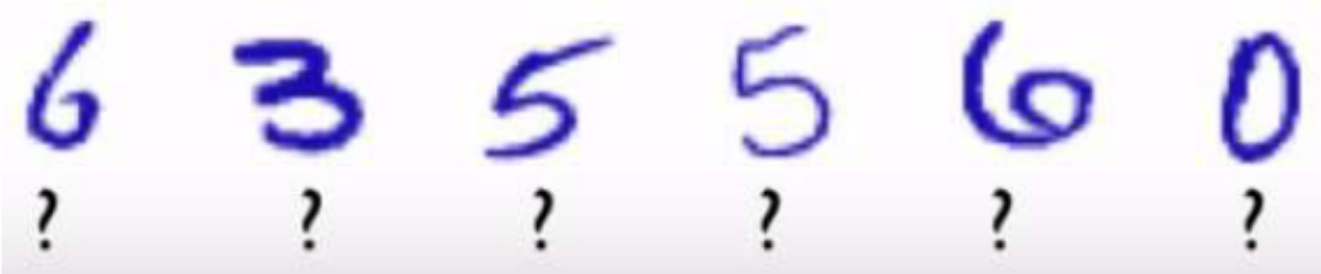


Supervised learning with training data (X_i, y_i)

Observe Data and labels



Compute the mapping function



Supervised learning

- Different tasks: regression, classification
- Different techniques: support vector machines, deep learning

Data and objects are vectors

Math conventions

In the following:

- $\mathbf{x}$ denotes a column vector, $\mathbf{x}^t$ its transpose
- $\mathbf{1}$ denotes a vector of ones
- X denotes a matrix, $X = [\mathbf{x}_1, \ldots, \mathbf{x}_m]$
- I_n denotes the identity matrix in $\mathbb{R}^{nn}$
- $< \mathbf{x}, \mathbf{y} >$ denotes the dot product; $< \mathbf{x}, \mathbf{y} >= \sum_k x_k y_k = \mathbf{x}^t \mathbf{y}$
- Norm of $\mathbf{x}$: $||\mathbf{x}||^2 =< \mathbf{x}, \mathbf{x} >$
- Distance between $\mathbf{x}$ and $\mathbf{y}$: $d(\mathbf{x}, \mathbf{y}) = ||\mathbf{x} - \mathbf{y}||^2 = ||\mathbf{x}||^2 + ||\mathbf{y}||^2 - 2* < \mathbf{x}, \mathbf{y} >$
- Frobenius norm of Matrix A: $||A||_F^2 = \sum_{ij} a_{ij}^2 = tr(A^t A)$

Notations

In the following, let us assume that we have a collection of m points $X_1, \ldots, X_m$ in $\mathbb{R}^n$

$$X = [\mathbf{x}_1, \ldots, \mathbf{x}_m] \in \mathbb{R}^{nm}$$

$$X = \begin{bmatrix} x_{11} & \cdots & x_{m1} \\ \vdots & \ddots & \vdots \\ x_{1n} & \cdots & x_{mn} \end{bmatrix} \in \mathbb{R}^{nm}$$

Covariance and Gram matrices

- Mean vector: $\bar{x} = \frac{1}{m}\sum_i x_i = \frac{1}{m}X.1$
- In the following, for sake of simplicity, data are centered, i.e., $\bar{x} = 0$
- Covariance matrix $\Sigma_X = \frac{1}{m}XX^t \in \mathbb{R}^{nn}$. A Covariance matrix is positive semidefinite.
- Gram matrix G_X of scalar products: $g_{ij} =< x_i, x_j >$.
  $G = X^tX \in \mathbb{R}^{mm}$. A Gram matrix is positive semidefinite.
- Note the dimension difference!
- Singular value decomposition of X: $X = USV^t$:
  - $\Sigma_X = \frac{1}{m}US^2U^t$
  - $G_X = VS^2V^t$

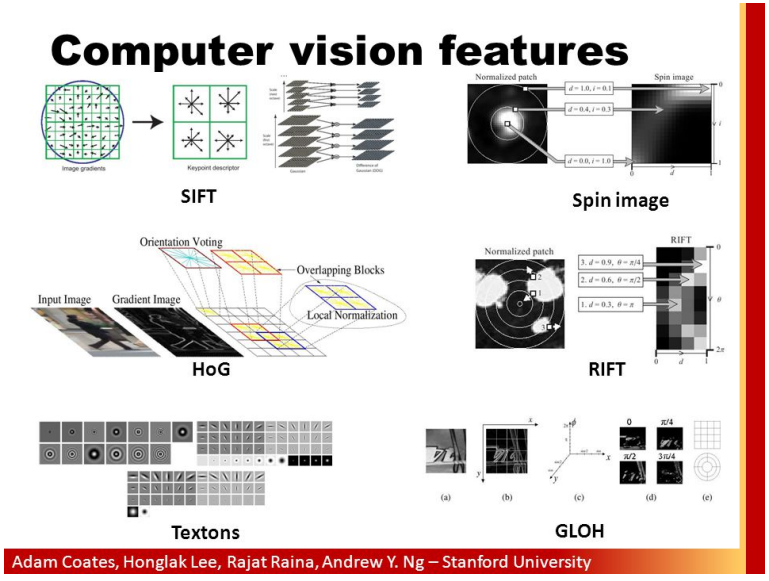
From data to vectors

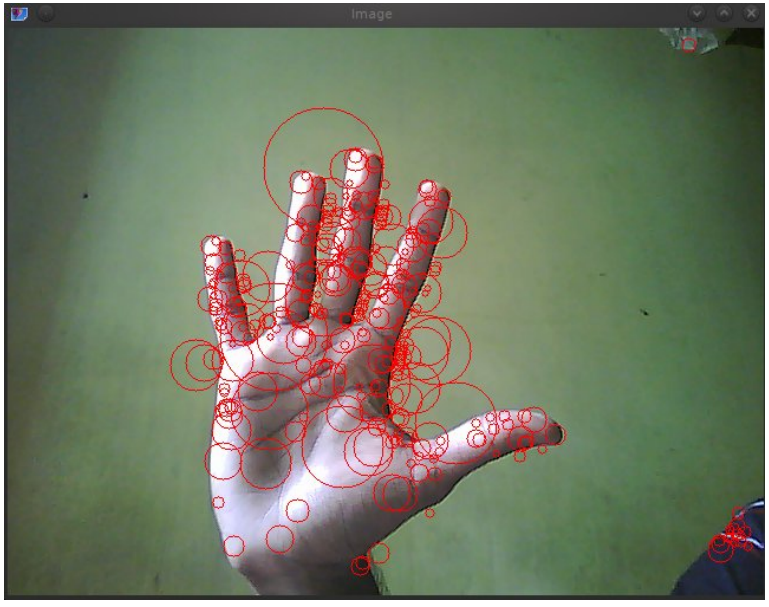
Any data (continuous/discrete, static/temporal)can be represented by vectors:

- obvious case: data are numerical measurements
- *ad hoc* construction of vector representation
- embedding

From data to vectors: *ad hoc* representation

- Signal can be represented by vectors:
  - text: quantization using dictionary of letters
  - audio: shape of the envelope and MFCC (mel-frequency cepstrum) coefficients on a audio frame
  - image: many representations, like HOG (histogram of oriented gradients), or sift.





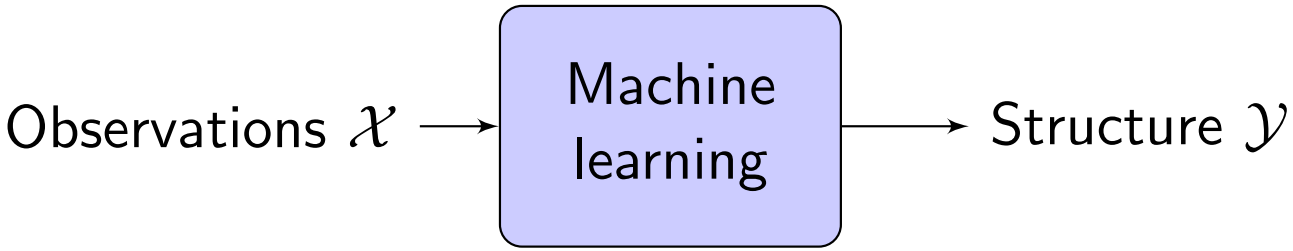
From data to vectors: embedding

- In the most general case, suppose that we have a collection of n objects $o_1, \ldots, o_n$
- we can define the similarity between the objects $k_{ij} = o_i \sim o_j$
- Matrix $K = [k_{ij}]$ is a PSD matrix
- Factorization: $K = U \Lambda U^t$, where U is orthogonal
- $K = Y^t Y$, with $Y = \Lambda^{\frac{1}{2}} U^t$
- K is the Gram matrix of n vectors $\mathbf{y}_1, \ldots, \mathbf{y}_n$
- In other words, the rows of U define the columns of Y, and an embedding in $\mathbb{R}^n$ of the n objects

From data to vectors: learn the embedding

- Word2vec proposed by google in 2013 (Mikolov *et al.* Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781)
- Learn mapping from word to $\mathbb{R}^n$ with typically $n = 1000$
- Deep network that analyzes a corpus an produces the representation
- words that share common contexts in the corpus have close embedding
- "king - man + female = queen"

General view of machine learning



where $\mathcal{X}$ can be any kind of data (numbers, measurements, images, text, audio, *etc.*)
and $\mathcal{Y}$ can be of various type (signal, numbers, images, labels, *etc.*)

Training data

A collection of m annotated examples $\{(X_i, Y_i), i \in [1 \dots m]\}$

( , "car"), . . . , ( , "car")

( , "bicycle"), . . . , ( , "bicycle")

Training data

Two steps:

- Training step: compute mapping function f such that $f(X) = Y$
- Prediction step: use f for new input samples X_{new}. $\hat{y} = f(x_{new})$

The empirical risk

Goal: compute f such that $f(X) \sim Y$

The empirical risk

f is parametric and chosen in a library of models indexed by θ
Goal: compute best θ such that $f_\theta(X) \sim Y$

The empirical risk

This shall be true for all training data

Goal: compute best θ such that $\forall i, f_\theta(X_i) \sim Y_i$

The empirical risk

Define a distance, or loss $\mathcal{L}$ to measure similarity

Goal: compute best θ such that $\forall i, \mathcal{L}(f_\theta(X_i), Y_i)$ is minimal

The empirical risk

Consider the average loss

Goal: compute best θ such that $\frac{1}{m}\sum_{i=1}^{m}\mathcal{L}(f_\theta(X_i), Y_i)$ is minimal

General formulation: the empirical risk

Available data: m training samples $\{(X_i, Y_i), i \in [1 \dots m]\}$
Compute best parameters θ:

$$\hat{\theta} = arg \min_{\theta} \mathcal{J}(\theta; X, Y) = arg \min_{\theta} \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}(f_\theta(X_i), Y_i)$$

where

- $\mathcal{L}$ is the loss function
- f_θ is the learned function parametrized by θ

General formulation: the empirical risk

Available data: m training samples $\{(X_i, Y_i), i \in [1 \dots m]\}$
Compute best parameters θ:

$$\hat{\theta} = arg \min_{\theta} \mathcal{J}(\theta; X, Y) = arg \min_{\theta} \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}(f_\theta(X_i), Y_i)$$

where

- $\mathcal{L}$ is the loss function
- f_θ is the learned function parametrized by θ

Example:

- X is an image 224×224
- Y is a classification vector (1000 classes)

General formulation: the empirical risk

$$\hat{\theta} = arg\min_{\theta} \mathcal{J}(\theta; X, Y) = arg\min_{\theta} \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}(f_\theta(X_i), Y_i)$$

Problems:

- Define $\mathcal{L}$ according to the problem
- Define f_θ
- Find the optimum

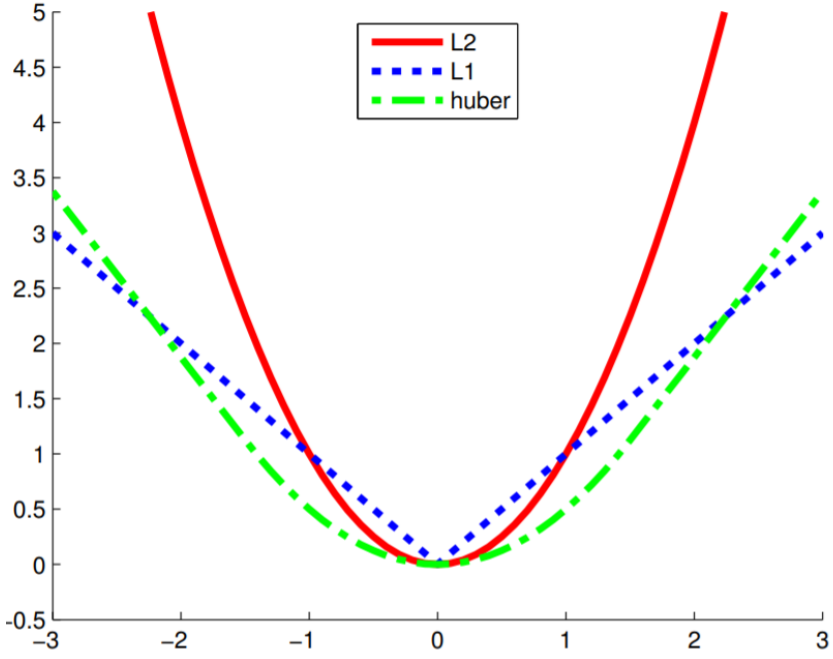
Loss functions $\mathcal{L}$

The loss quantifies the discrepancy between the predicted value and the true value, and depends on the task, or in other words the price to pay for inaccuracy. Two main losses:

- Prediction. Usually $\mathcal{L}(f_\theta(X_i), Y_i) = ||f_\theta(X_i) - Y_i||^2$. Any p-norm (with $p > 1$ will be OK).
- Classification. Different possible choices (hinge loss, logistic loss).

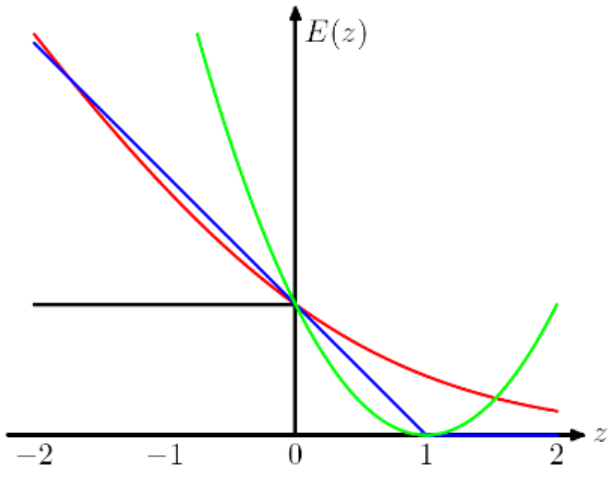
Loss functions $\mathcal{L}$ for regression

- L_2 norm. Not robust
- L_1 norm. Median regression
- Huber loss. Quadratic when $|f_\theta(X_i) - Y_i| < \delta$ and linear otherwise. Robust and differentiable.



Loss functions $\mathcal{L}$ for binary classification

- Binary classification, classes $\{-1, 1\}$. $\hat{y} = 1$ if $f_\theta(X) > 0$
- $\mathcal{L}(\hat{y}, y) = \mathbb{I}_{\hat{y}y<0}$
- Indicator function is non-convex, discontinuous and non differentiable
- Hinge loss (convex) $\mathcal{L}_{hinge}(\hat{y}, y) = \max(0, 1 - \hat{y}y)$
- L_2 norm $\mathcal{L}_2(\hat{y}, y) = (1 - \hat{y}y)^2$
- Logistic loss (convex and differentiable) $\mathcal{L}_{logistic}(\hat{y}, y) = log(1 + e^{-\hat{y}y})$



Optimization

$$\hat{\theta} = arg\min_{\theta} \mathcal{J}(\theta; X, Y) = arg\min_{\theta} \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}(f_\theta(X_i), Y_i)$$

- Every machine learning problem boils down to a (tough) optimization problem.
- Search space can be extremely large (millions of variables, aka ChatGPT)
- Various techniques according o the problem and its constraints

Evaluation

Because optimization is tough and is uncertain, we need to monitor that it performs well!

- Split data into three sets: training, validation and test data
- Train on training data (empirical loss) and evaluate on test data (same loss)
- Avoid underfitting and overfitting

Optimization at large

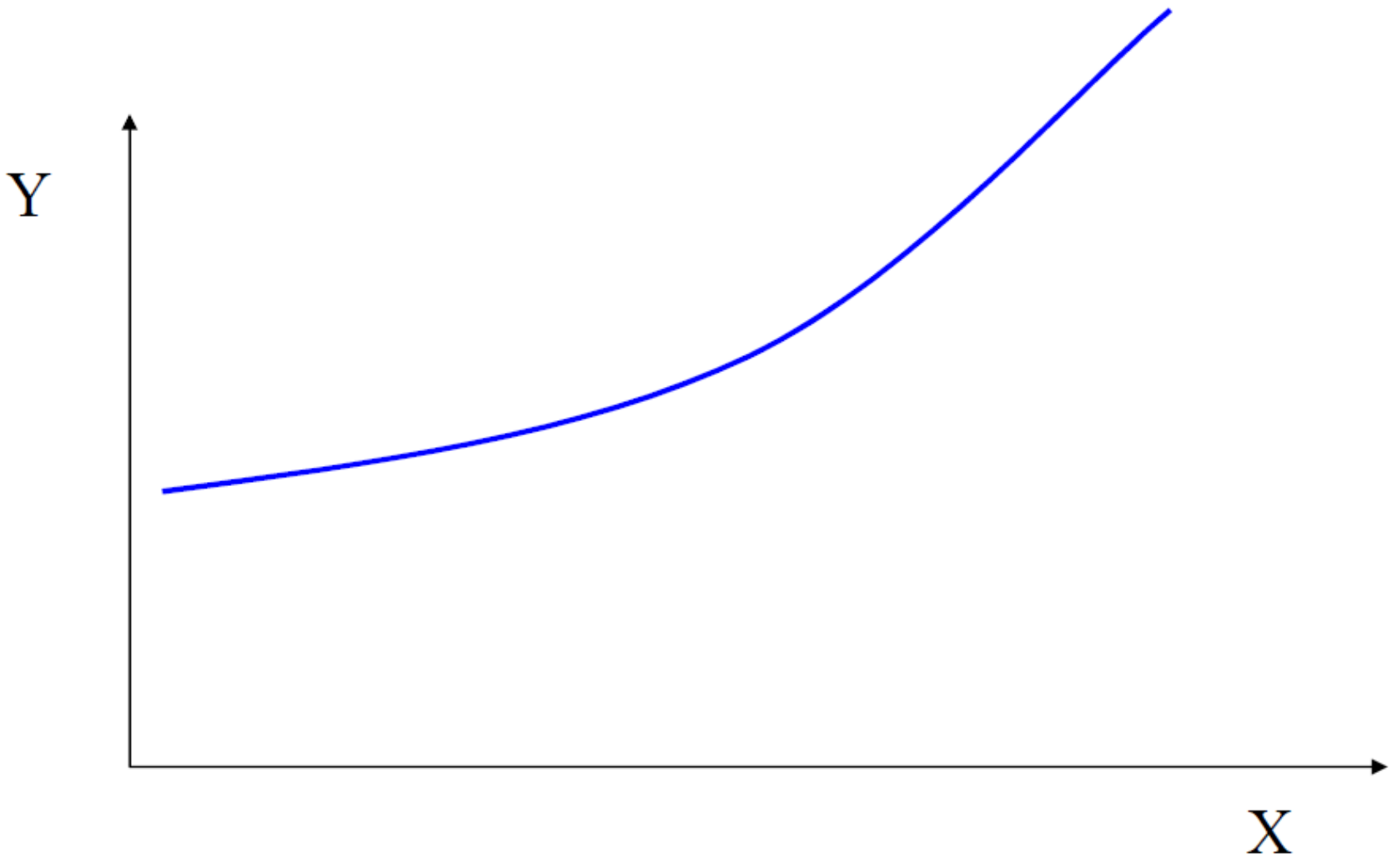
Optimizing an ML system is larger than minimizing the cost:

- Data gathering and preparation
- Embed the data into smaller dimension (curse of dimensionality)
- Use the appropriate model size (Occam's razor)
- Use the appropriate functions for deployment (hardware constraints)
- Beware the paradigm *bigger is better*

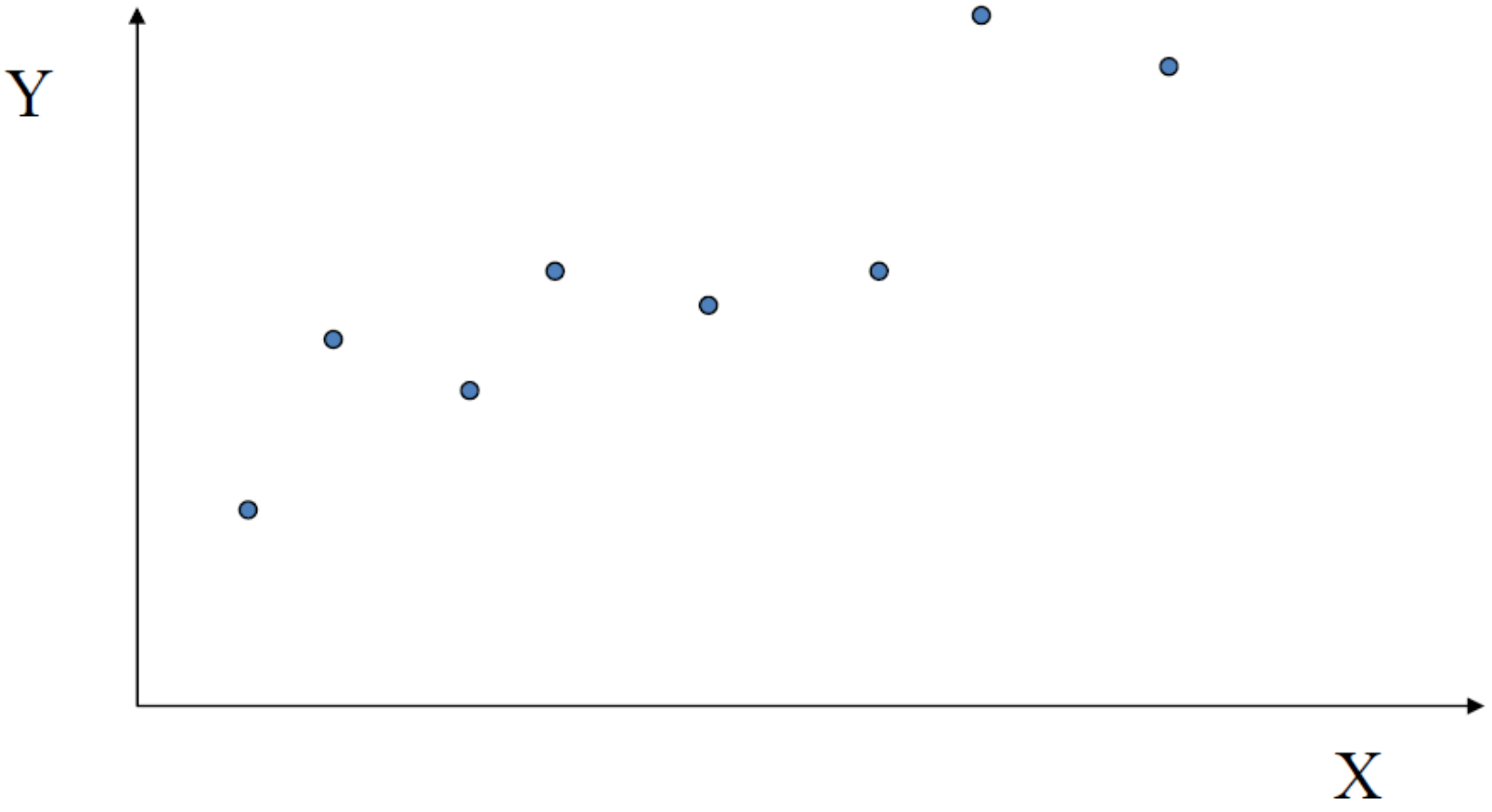
Overfitting and underfitting

- The complexity of the true model is unknown
- We observe points in high dimension
- Observations are noisy
- How to choose the capacity of the model f_θ
- capacity = degrees of freedom = number of parameters

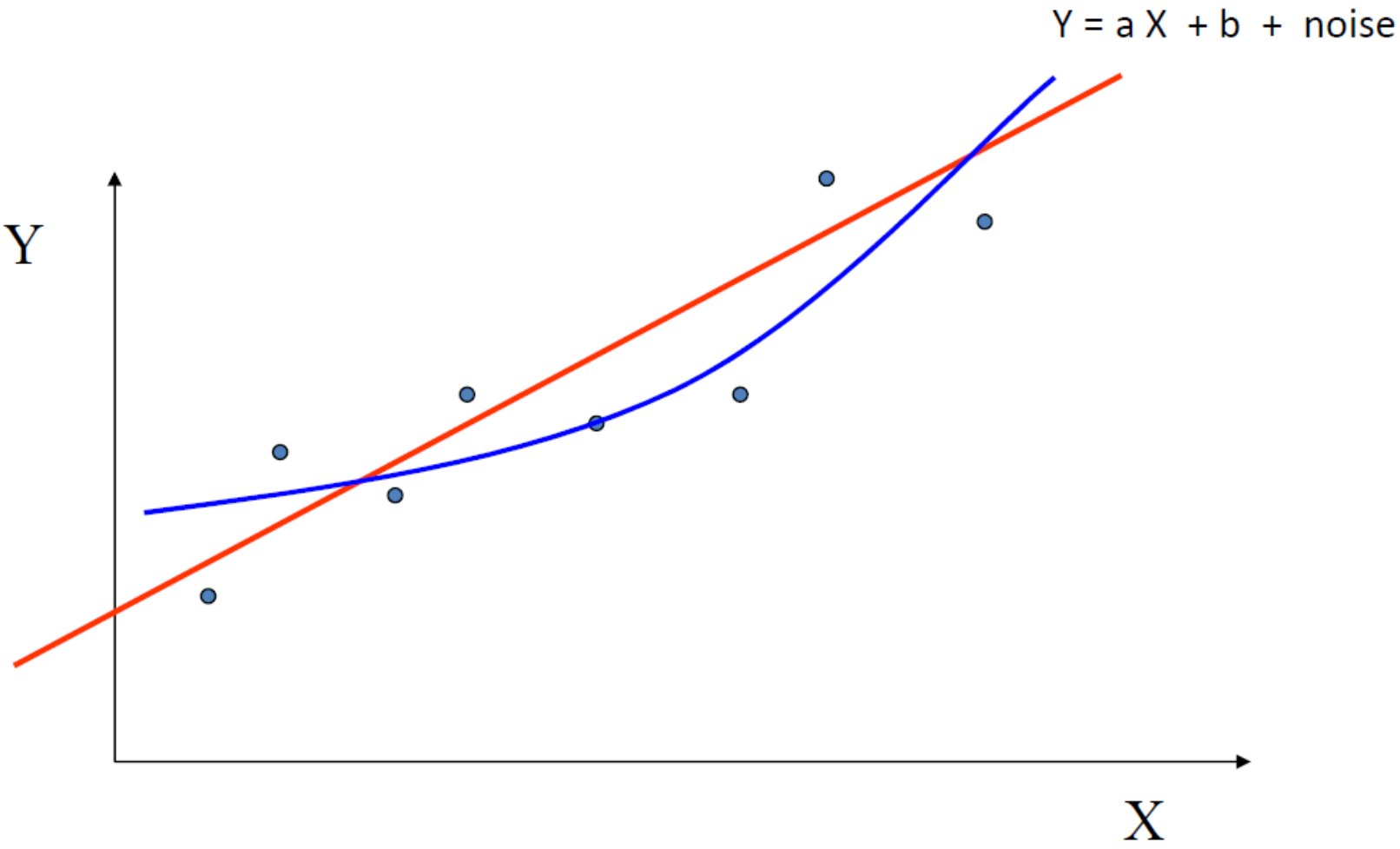
Hidden true function



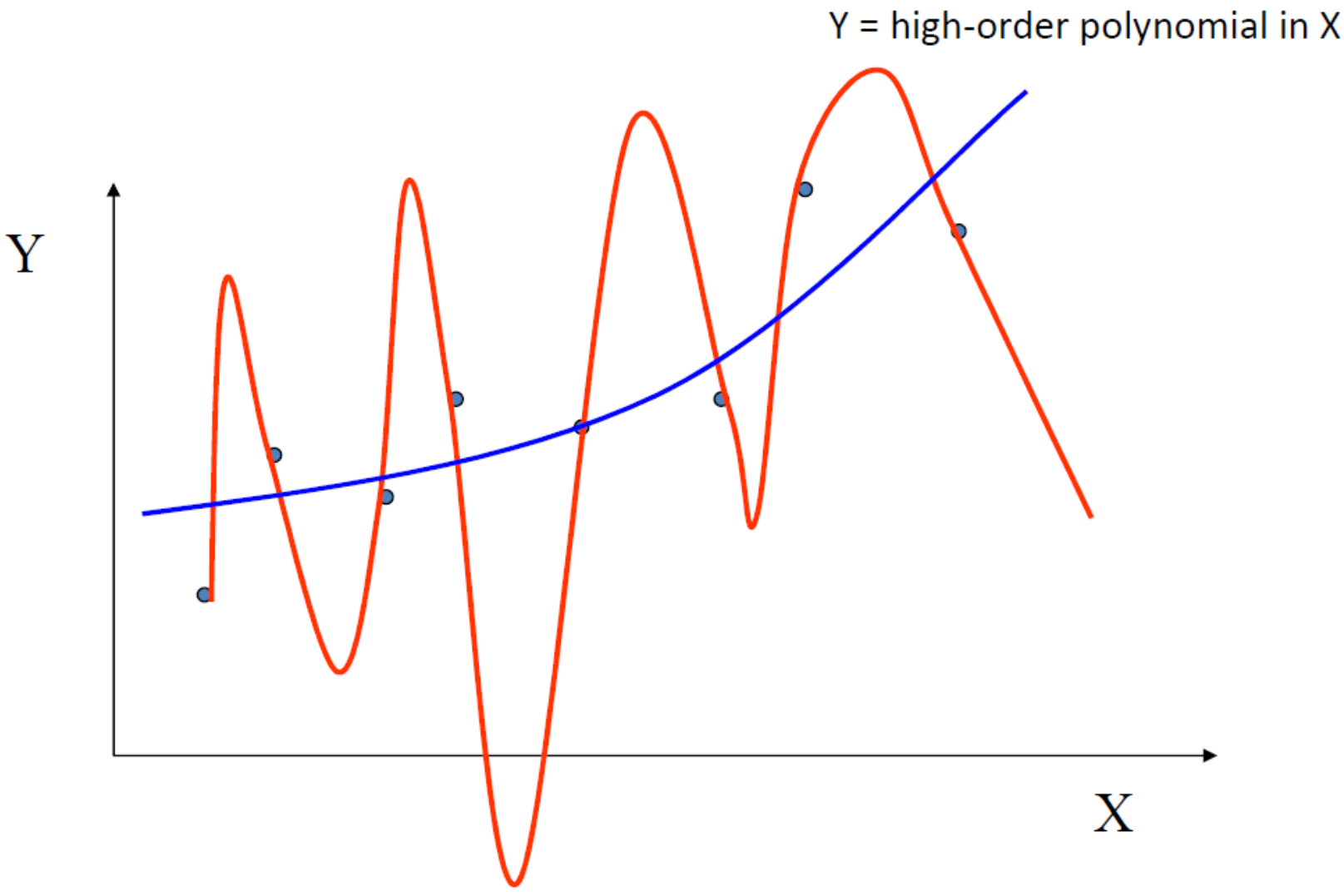
Observed points



Underfitting



Overfitting

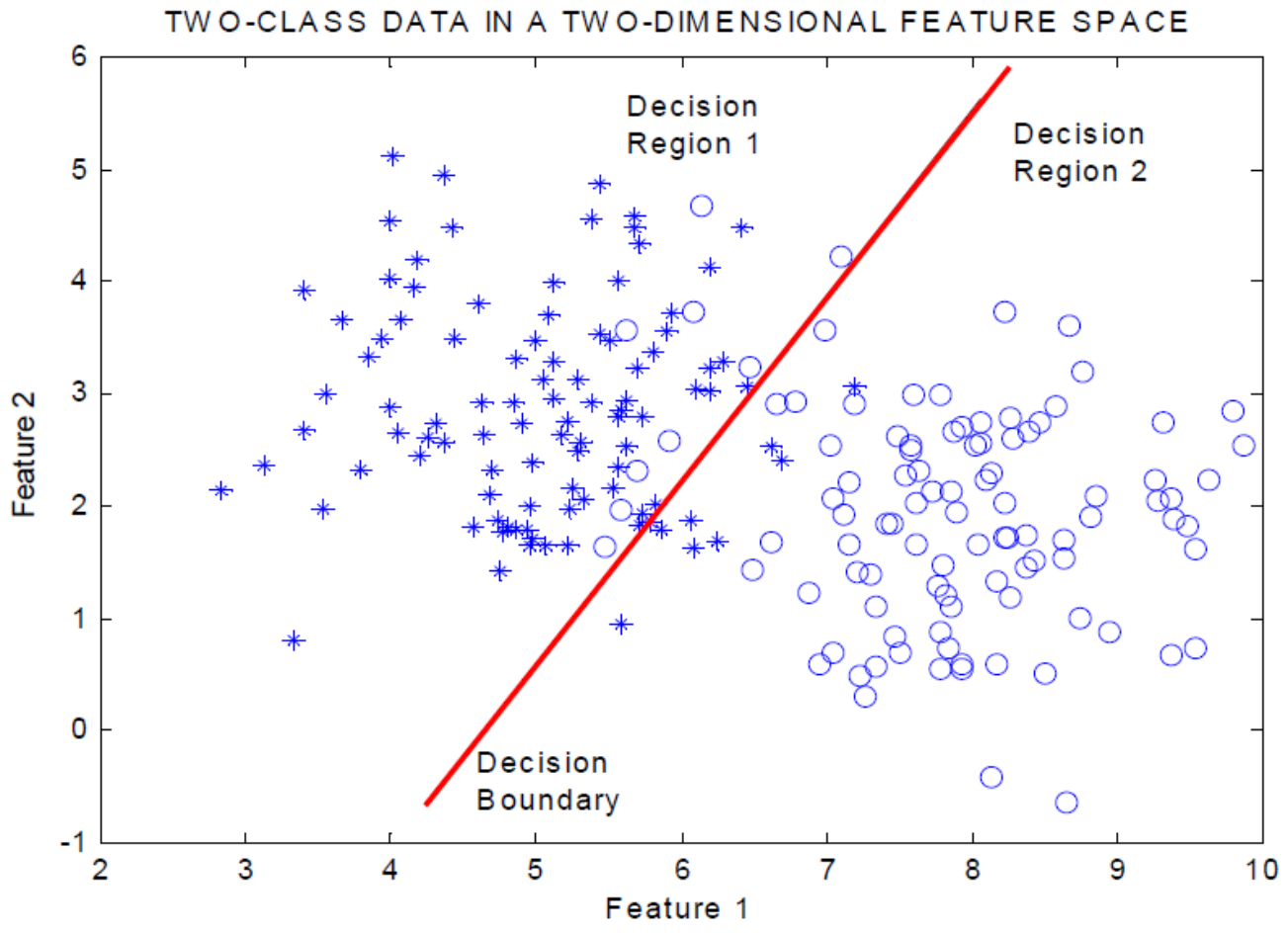


Overfitting and underfitting

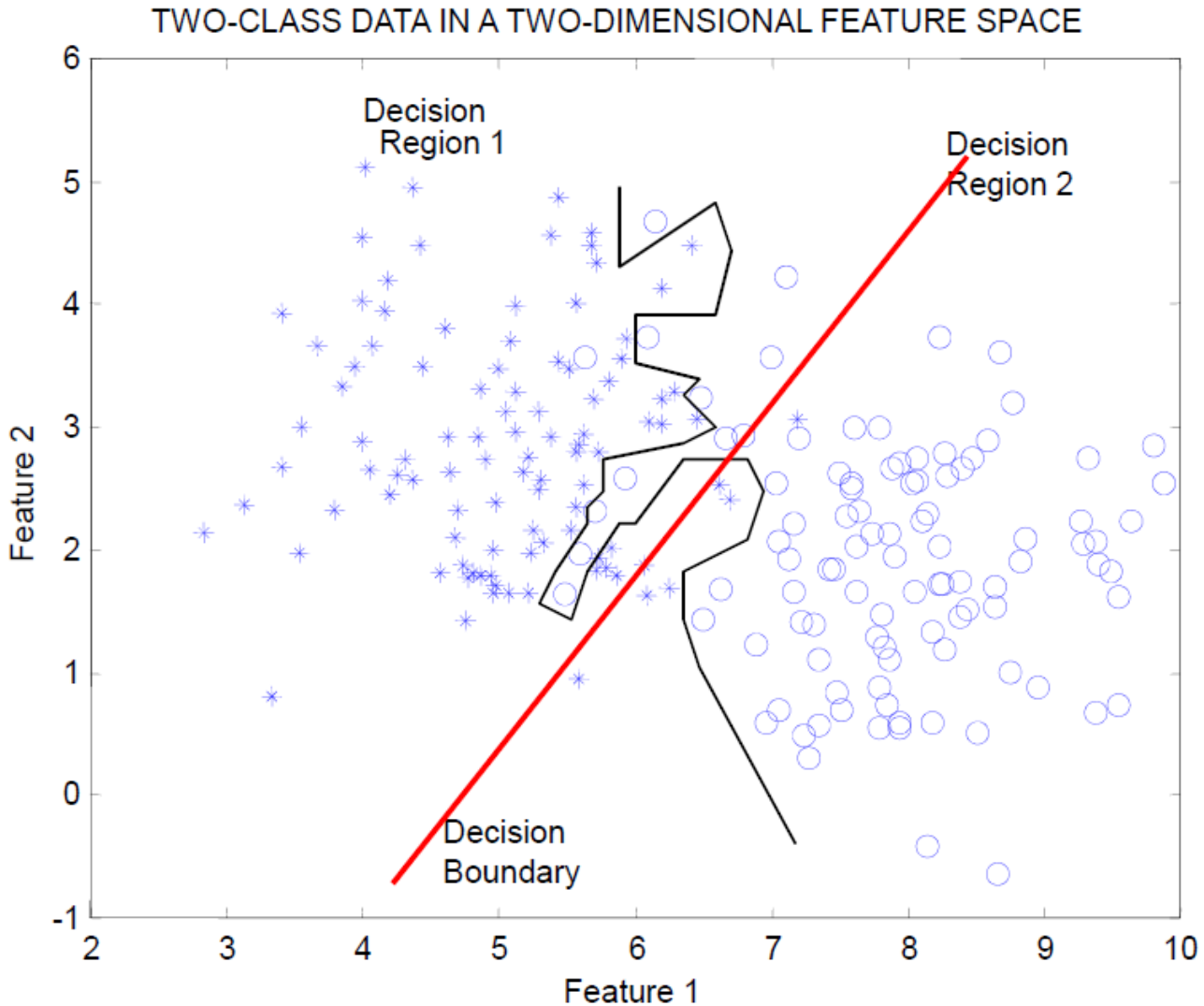
Bias/variance tradeoff

- Underfitting
  - Does not fit the training data perfectly
  - Training cost is high
  - Generalizes to other data with limited accuracy and error
- Overfitting
  - Perfectly fits the data
  - Generalizes with potentially large errors
  - Testing cost is high

Overfitting



Overfitting



Strategies to avoid overfitting

- AIC and BIC criteria (linear regression section)
- Regularization (non-smooth optimization section)

Regularization trick to avoid overfitting

New cost function (Structural risk minimization Vapnik 2013)

$$\hat{\theta} = arg\min_{\theta} \mathcal{J}(\theta; X, Y) = arg\min_{\theta} \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}(f_\theta(X_i), Y_i) + \lambda \mathcal{R}(\theta)$$

where

- $\mathcal{R}$ penalizes large values of θ (ridge) or enforces prior/sparse knowledge (lasso)
- λ is a weighting parameter to be adjusted