CSC_51054_EP (ex. INF 554)
Machine Learning and Deep Learning

Prof. M. Vazirgiannis

Labs/TDs:
Prof. D. Buscaldi, Prof. J. Lutzyer, Dr. C. Molero del Río, J. Dentan

Data Science and Mining Team, LIX
http://www.lix.polytechnique.fr/dascim/

Sept 24, 2024

Course Syllabus

- **General Introduction to Machine Learning**
  - Machine Learning paradigms
  - The Machine Learning Pipeline
- **Supervised Learning**
  - Generative and non generative methods
  - Naive Bayes, KNN and regressions
  - Tree based methods

  Unsupervised Learning
  - Dimensionality reduction
  - Clustering
- **Advanced Machine Learning Concepts**
  - Regularization
  - Model selection
  - Feature selection
  - Ensemble Methods – ramndom forests

Course Syllabus

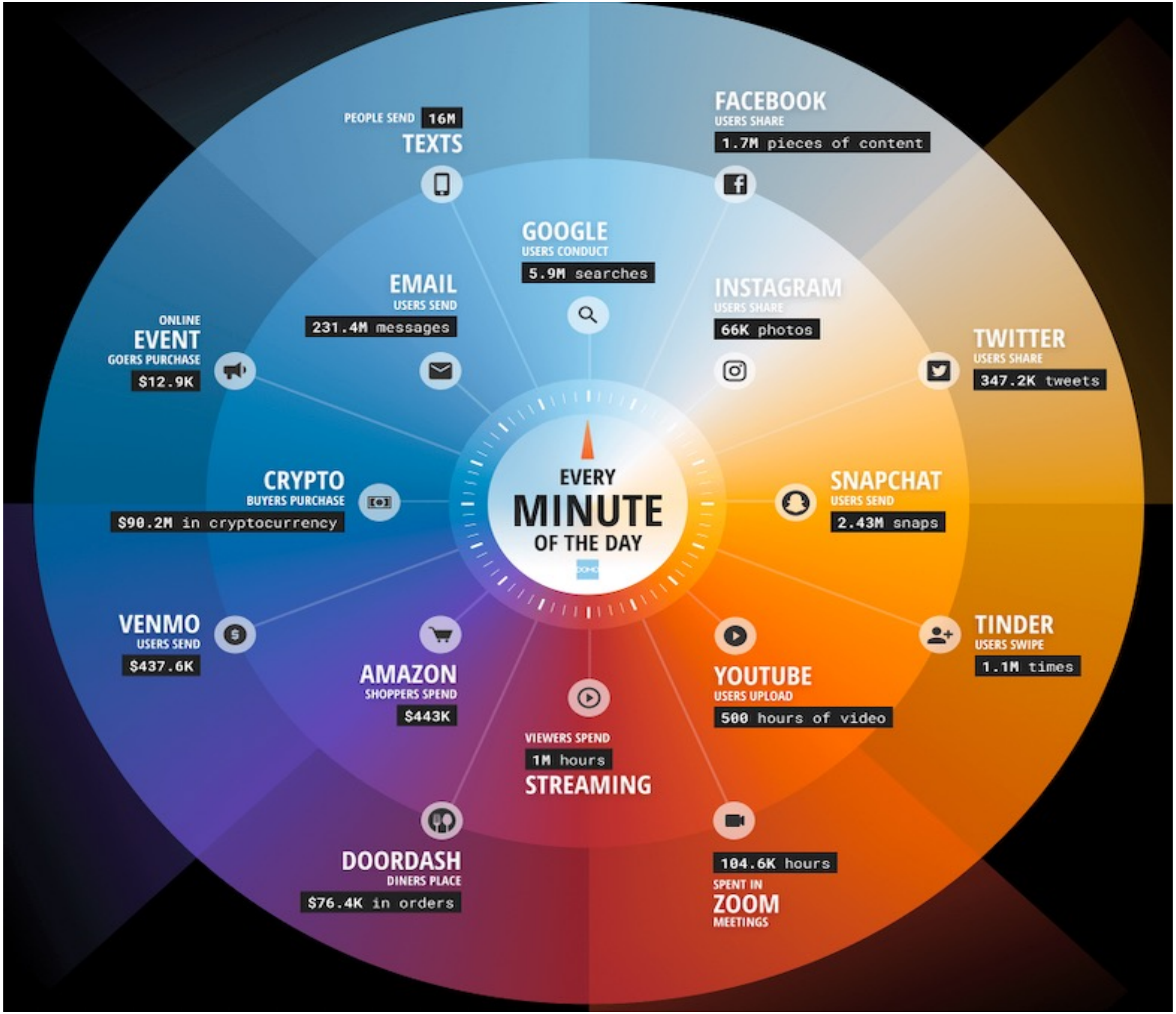
- **Kernels**
  - Introduction to kernels, Support Vector Machines
- **Neural Networks**
  - Introduction to Neural Networks
  - Perceptrons and back-propagation
- **Deep Learning I**
  - Convolutional Neural Networks
  - Recurrent Neural Networks
  - Applications
- **Deep Learning II**
  - Modern Natural Language Processing
  - Unsupervised Deep Learning
  - Embeddings, Auto-Encoders
- **Graph based ML**
  - Kernels, Node/Graph embeddings

Course Logistics

- Class: 14:00 – 16:00
- Labs: 16:15 – 18:30
- Interaction/Q&As
    - Different channels within will be set up for course/labs/other-admin issues
- **Install the software requested and the runtime setup instructions !**
- Read carefully our announcements
- For **IPP master students & others** – please fill in the registration form: https://tinyurl.com/353ft47k
    - Get access to the material and course communication
    - to communicate your results to your home institution

Course Logistics

- **Magistral classes** 14:00-16:00 - Amphi Faure
- **Labs/TDs: :**
- Monday (16:15 - 18:15)
  - PC03 (Gr1), PC05 (Gr2), PC06 (Gr3).
- Wednesday (14:00 - 16:00)
  - PC04 (Gr4), PC05 (Gr5), PC06 (Gr6).
  - **Distribution based on your surname initial letter:**
    - **1: A-G, 2: H -M, 3:N·Z**
- Evaluation
  - Assignment (A) - an individual take-home assessment handed out October 14, submission deadline November 4.
  - Course project (CP) – Kaggle data challenge – dates (tentative) handed out November 8, submission deadline: December 12. oral assessments: last week before December holidays
  - Grading scheme
    - Final Grade = A*20% + CP*80%
- Course/Lab Material: @Moodle CSC_51054_EP –2024: https://moodle.polytechnique.fr/course/view.php?id=19308



https://www.domo.com/data-never-sleeps#

Data are heterogeneous

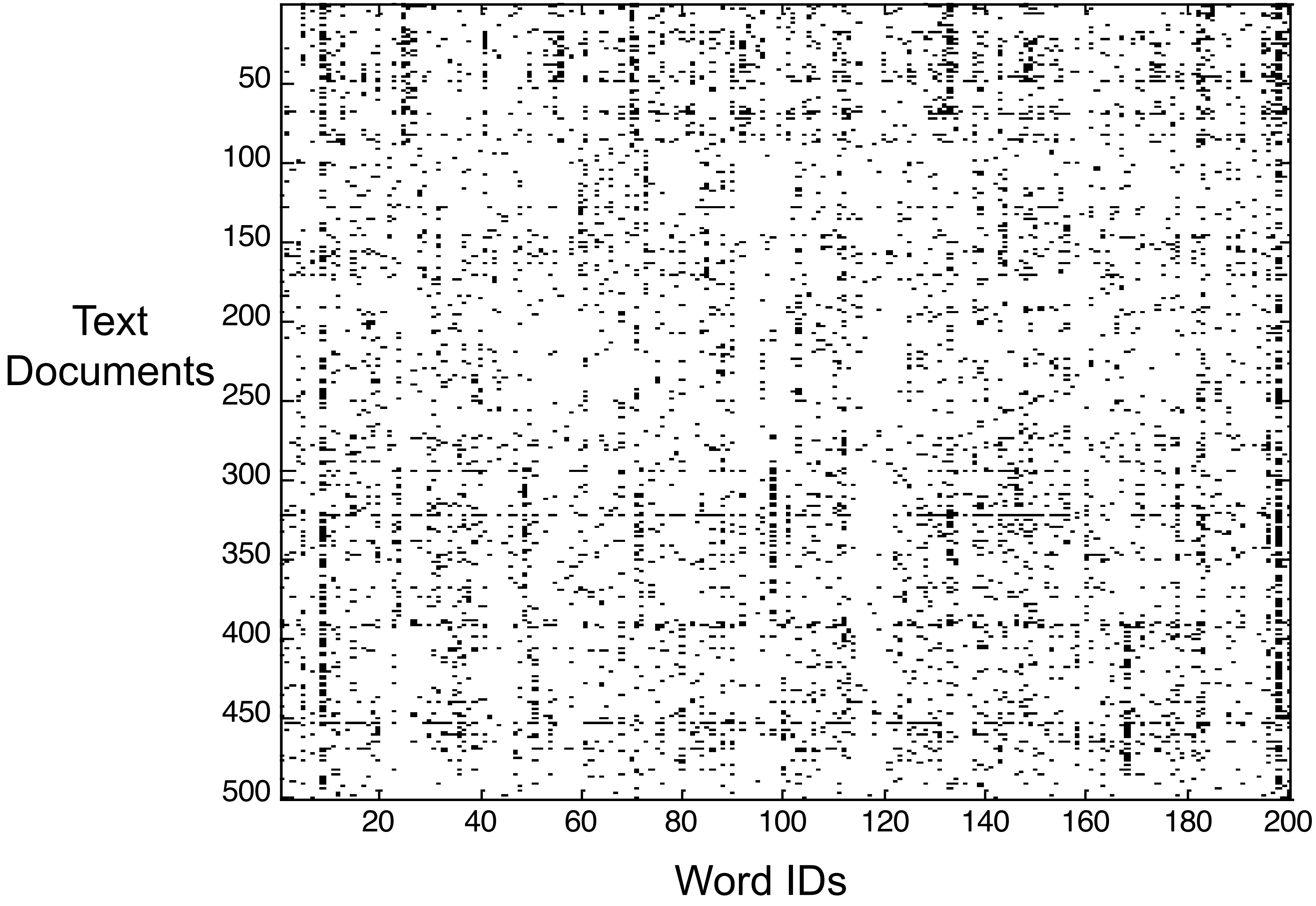
- Data are considered as valuable resource
  - User behavior, Queries
  - User generated content in Social networks
  - Experimental/Scientific data – genome, …
- Traditional: numerical, categorical, or binary
- Text: emails, tweets, news articles
- Geo-based location data
- Network, Sensor data
- Images, Video
- *Synthetic data (LLMs, image/speech/sound generators) are already dominating*

Basic Data / features representation

$$n\left\{\underbrace{\begin{array}{|c|c|c|c|}\hline 2.3 & -1.5 & \ldots & -1.3 \\ \hline 1.1 & 0.1 & \ldots & -0.1 \\ \hline \ldots & \ldots & \ldots & \ldots \\ \hline \end{array}}_{p}\right.$$

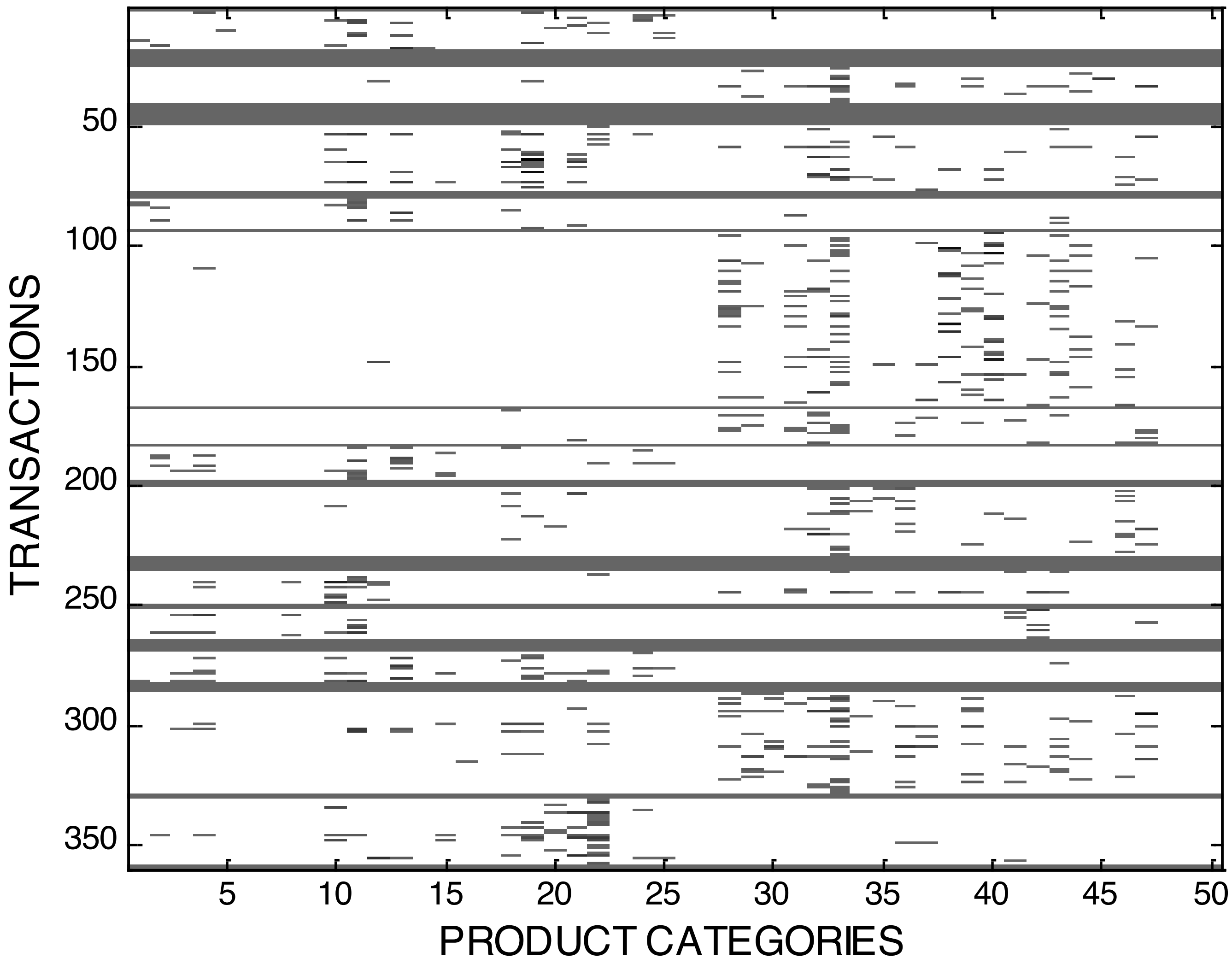
- Rows = objects
- Columns = measurements on objects
  - Represent each row as a *p*-dimensional vector, where p is the dimensionality
    - In efffect, embed our objects in a *p*-dimensional vector space
    - Often useful, but always appropriate
- Both *n* and *p* can be very large in certain ML applications
- VC-dimension - a theoretical understanding of how model complexity affects generalization and how many data points are needed for effective learning.

Sparse Matrix (Text) Data



- *Data Sparsity* – significant challenge…

“Market Basket” Data



Sequence (Web) Data

```
128.195.36.195, -, 3/22/00, 10:35:11, W3SVC, SRVR1, 128.200.39.181, 781, 363, 875, 200, 0, GET, /top.html, -,
128.195.36.195, -, 3/22/00, 10:35:16, W3SVC, SRVR1, 128.200.39.181, 5288, 524, 414, 200, 0, POST, /spt/main.html, -,
128.195.36.195, -, 3/22/00, 10:35:17, W3SVC, SRVR1, 128.200.39.181, 30, 280, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.195.36.101, -, 3/22/00, 16:18:50, W3SVC, SRVR1, 128.200.39.181, 60, 425, 72, 304, 0, GET, /top.html, -,
128.195.36.101, -, 3/22/00, 16:18:58, W3SVC, SRVR1, 128.200.39.181, 8322, 527, 414, 200, 0, POST, /spt/main.html, -,
128.195.36.101, -, 3/22/00, 16:18:59, W3SVC, SRVR1, 128.200.39.181, 0, 280, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:54:37, W3SVC, SRVR1, 128.200.39.181, 140, 199, 875, 200, 0, GET, /top.html, -,
128.200.39.17, -, 3/22/00, 20:54:55, W3SVC, SRVR1, 128.200.39.181, 17766, 365, 414, 200, 0, POST, /spt/main.html, -,
128.200.39.17, -, 3/22/00, 20:54:55, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:55:07, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:55:36, W3SVC, SRVR1, 128.200.39.181, 1061, 382, 414, 200, 0, POST, /spt/main.html, -,
128.200.39.17, -, 3/22/00, 20:55:36, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:55:39, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:56:03, W3SVC, SRVR1, 128.200.39.181, 1081, 382, 414, 200, 0, POST, /spt/main.html, -,
128.200.39.17, -, 3/22/00, 20:56:04, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:56:33, W3SVC, SRVR1, 128.200.39.181, 0, 262, 72, 304, 0, GET, /top.html, -,
128.200.39.17, -, 3/22/00, 20:56:52, W3SVC, SRVR1, 128.200.39.181, 19598, 382, 414, 200, 0, POST, /spt/main.html, -,
```

User 1	2 3 2 2 3 3 3 1 1 1 3 1 3 3 3 3
User 2	3 3 3 1 1 1
User 3	7 7 7 7 7 7 7 7
User 4	1 5 1 1 1 5 1 5 1 1 1 1 1 1
User 5	5 1 1 5
...	...

Time Series Data

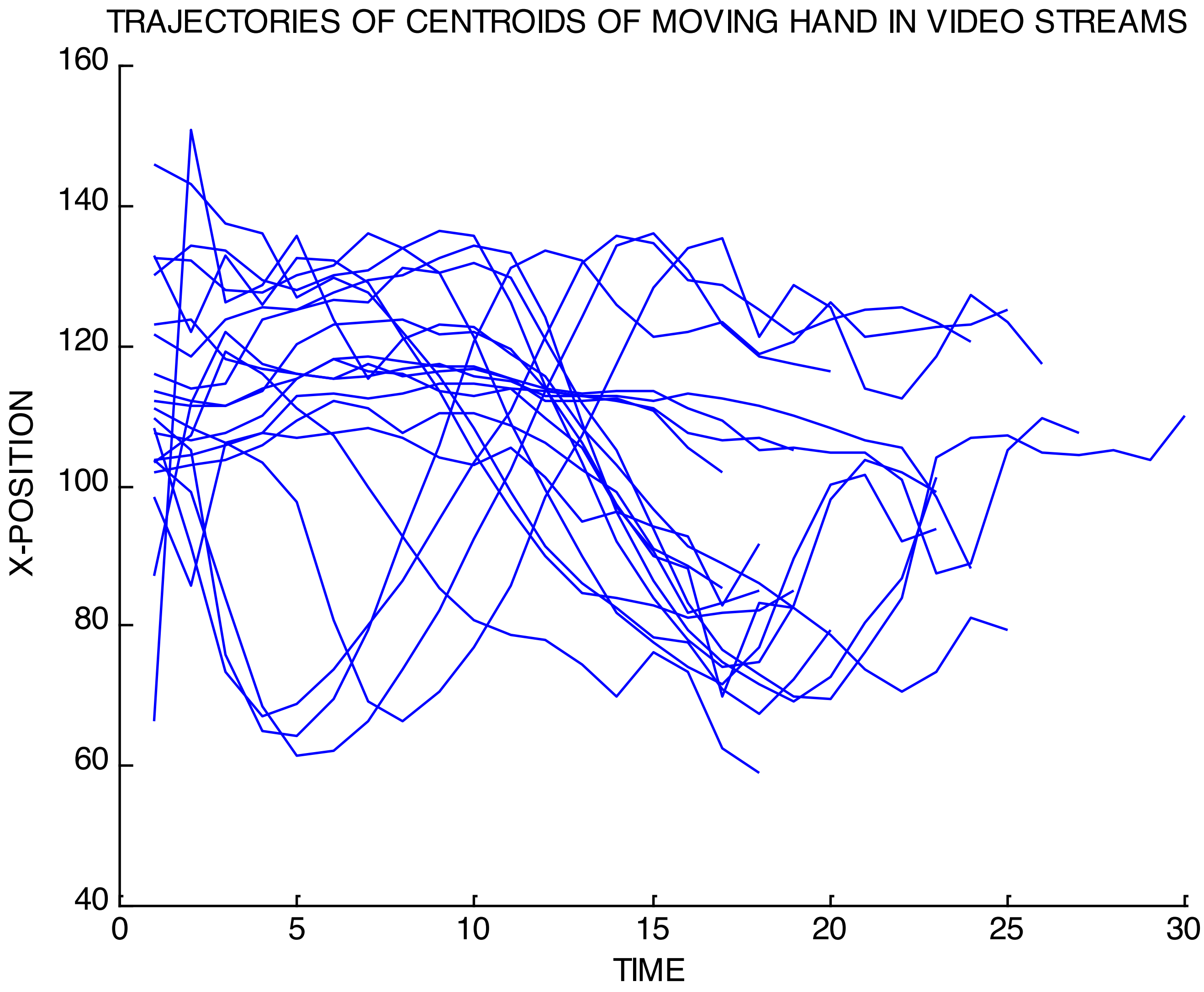
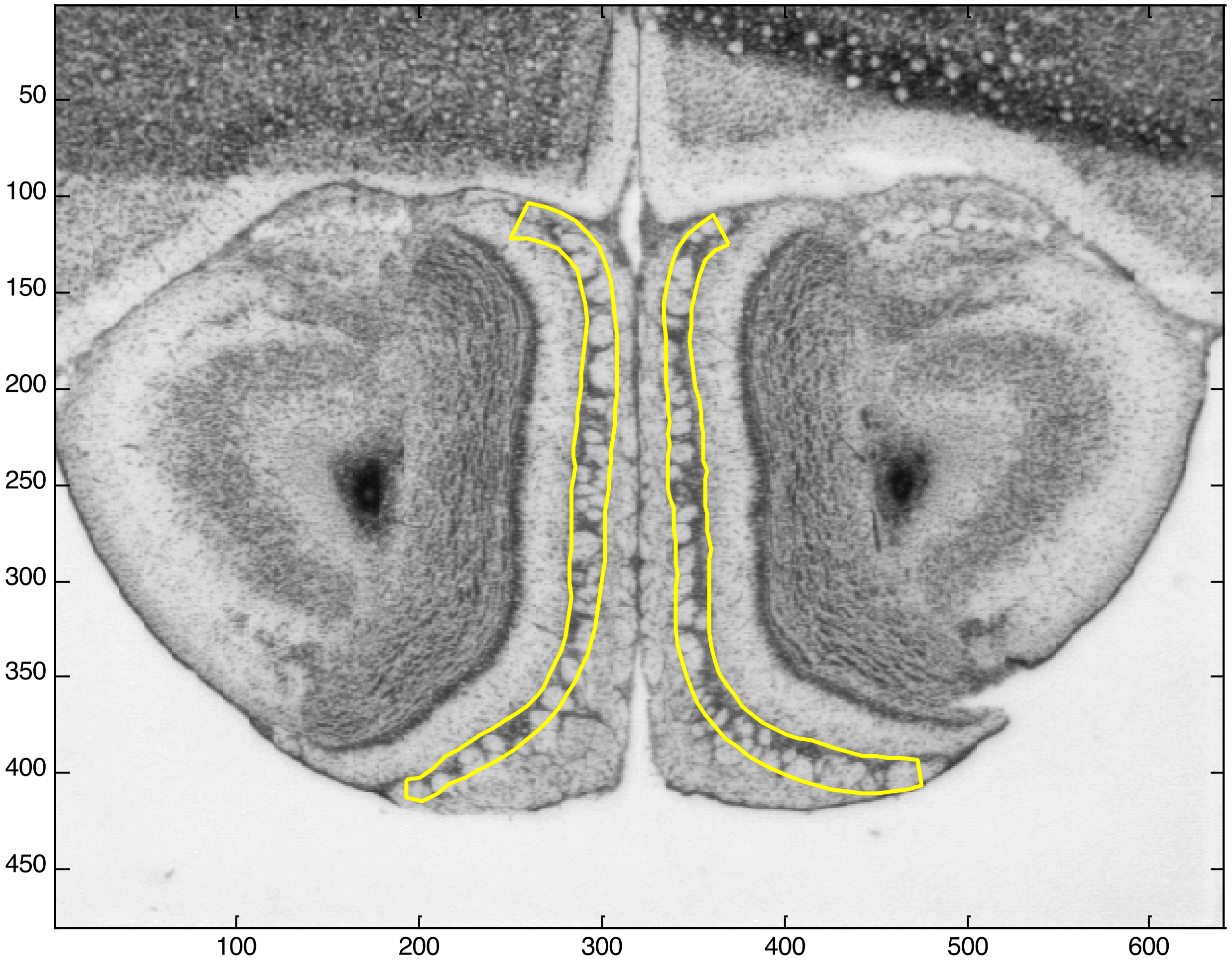
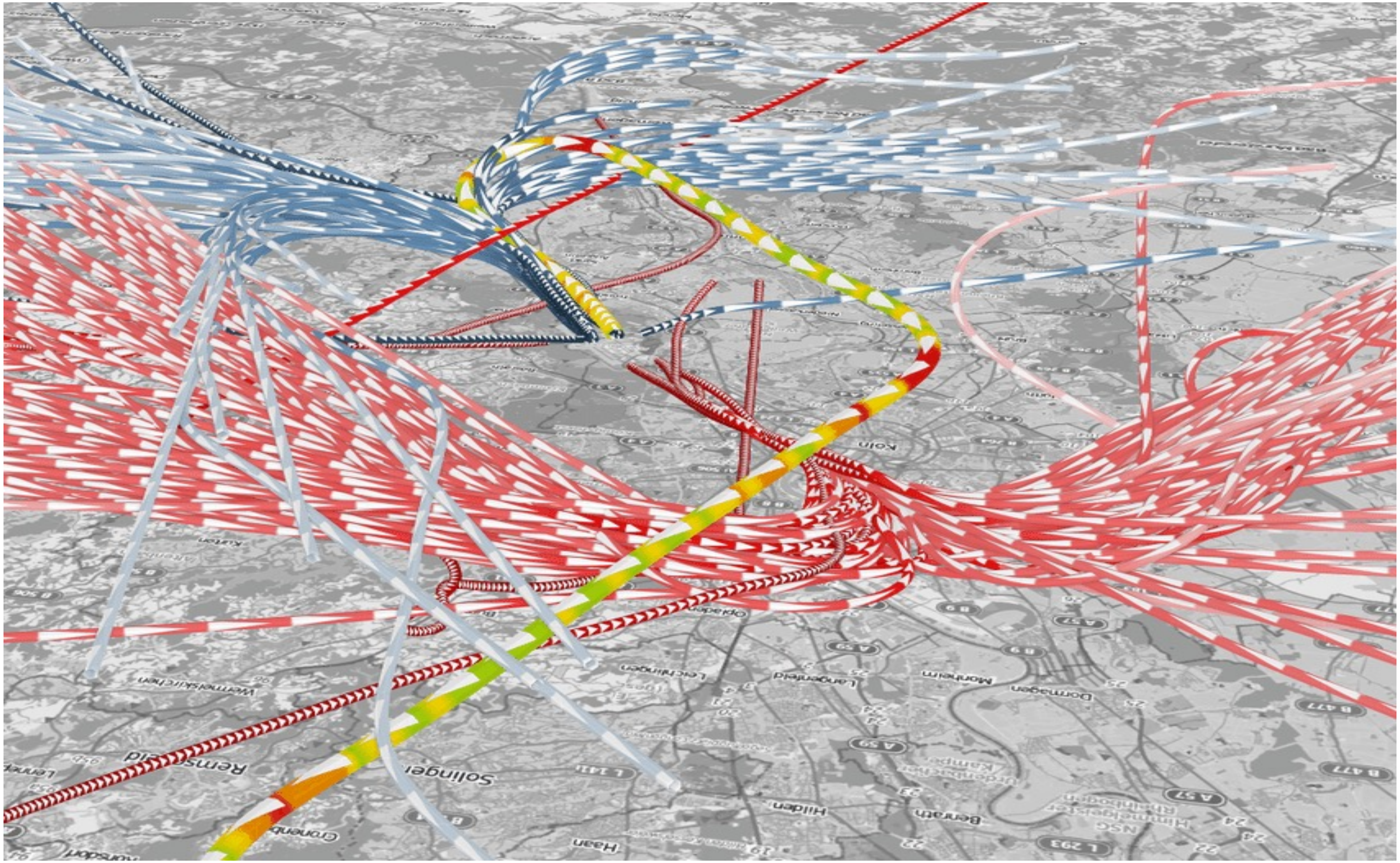


Image Data

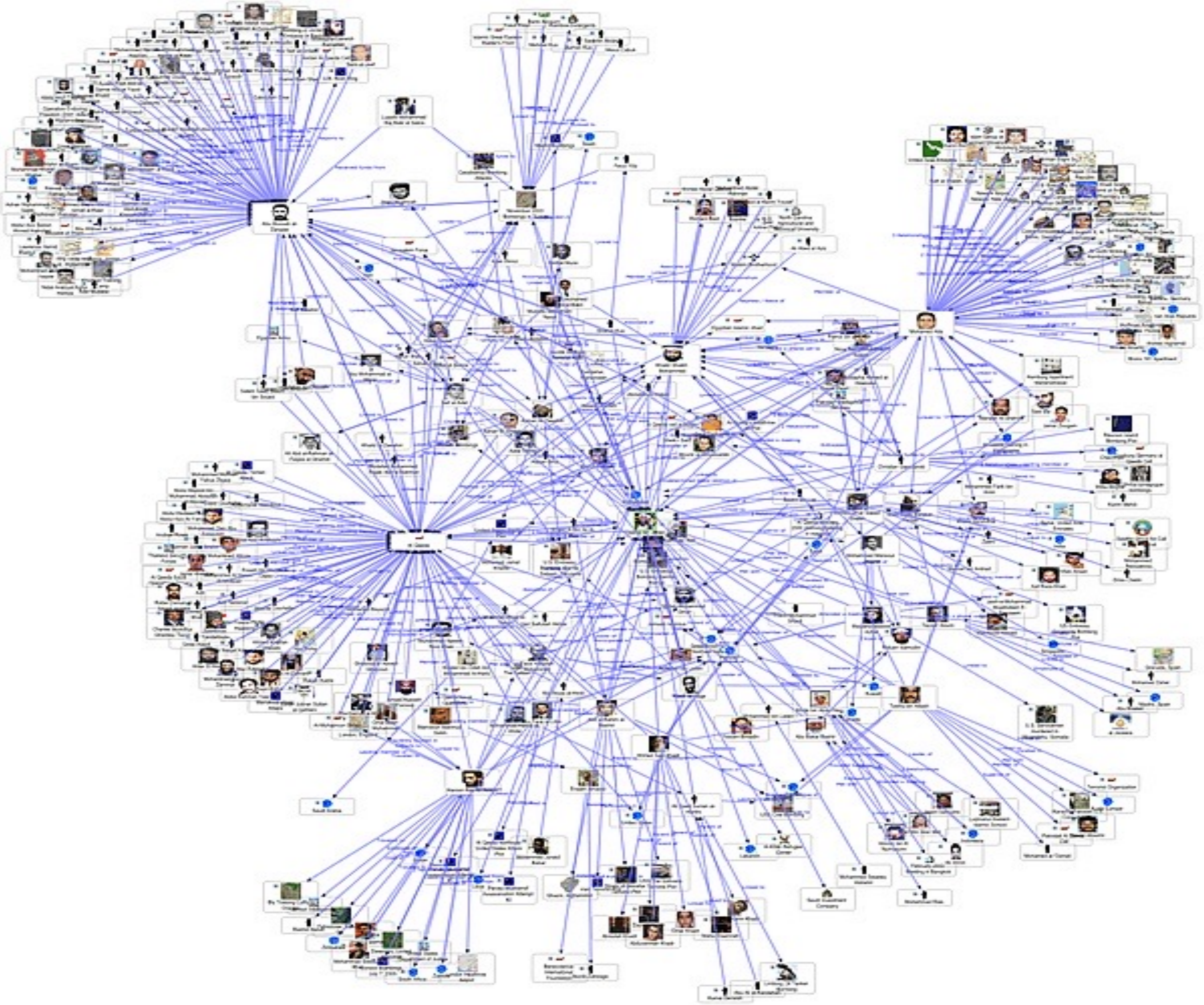


Spatio-temporal data



Buschmann, Stefan & Trapp, Matthias & Döllner, Jürgen. (2014). Real-Time Animated Visualization of Massive Air-Traffic Trajectories. Proceedings - 2014 International Conference on Cyberworlds, CW 2014. 10.1109/CW.2014.32.

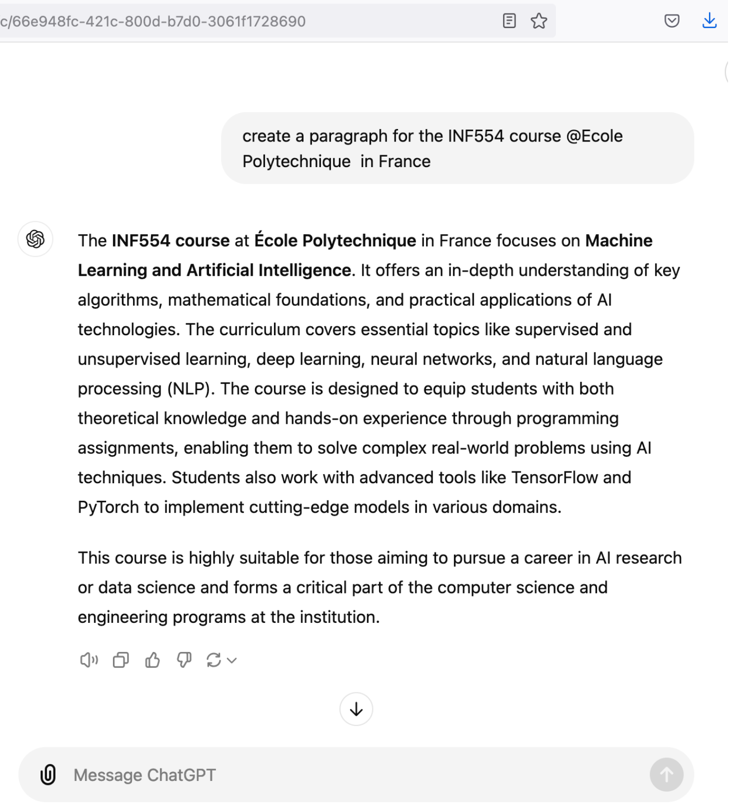
Social Networks – Graphs



https://stackoverflow.com/questions/4594962/social-network-directed-graph-library-for-net

Synthetic Data

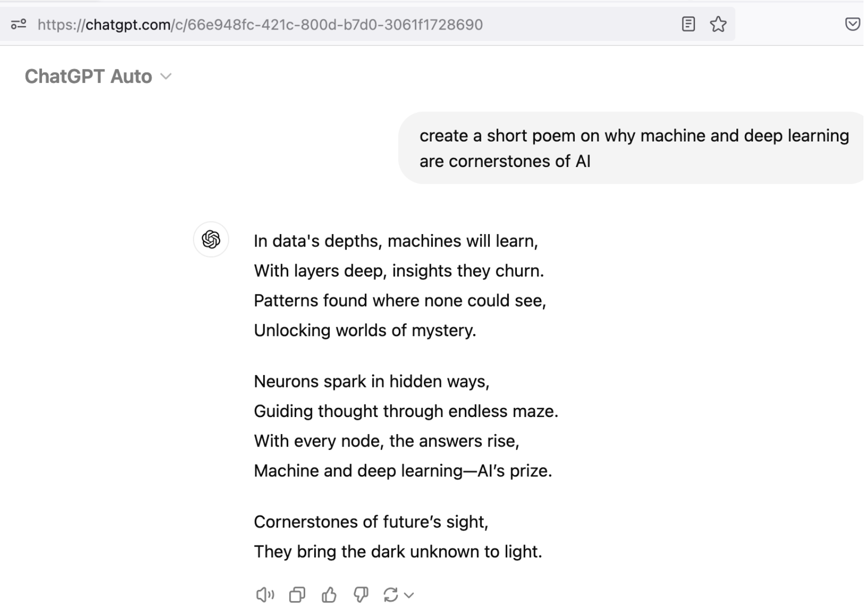
https://this-person-does-not-exist.com/en



create a paragraph for the INF554 course @Ecole Polytechnique in France

The **INF554 course** at **École Polytechnique** in France focuses on **Machine Learning and Artificial Intelligence**. It offers an in-depth understanding of key algorithms, mathematical foundations, and practical applications of AI technologies. The curriculum covers essential topics like supervised and unsupervised learning, deep learning, neural networks, and natural language processing (NLP). The course is designed to equip students with both theoretical knowledge and hands-on experience through programming assignments, enabling them to solve complex real-world problems using AI techniques. Students also work with advanced tools like TensorFlow and PyTorch to implement cutting-edge models in various domains.

This course is highly suitable for those aiming to pursue a career in AI research or data science and forms a critical part of the computer science and engineering programs at the institution.



create a short poem on why machine and deep learning are cornerstones of AI

In data's depths, machines will learn,
With layers deep, insights they churn.
Patterns found where none could see,
Unlocking worlds of mystery.

Neurons spark in hidden ways,
Guiding thought through endless maze.
With every node, the answers rise,
Machine and deep learning—AI's prize.

Cornerstones of future's sight,
They bring the dark unknown to light.

Exploratory Data Analysis

- Getting an overall sense of the data set
  - Computing summary statistics:
    - Number of distinct values, max, min, mean, median, variance, skewness,..

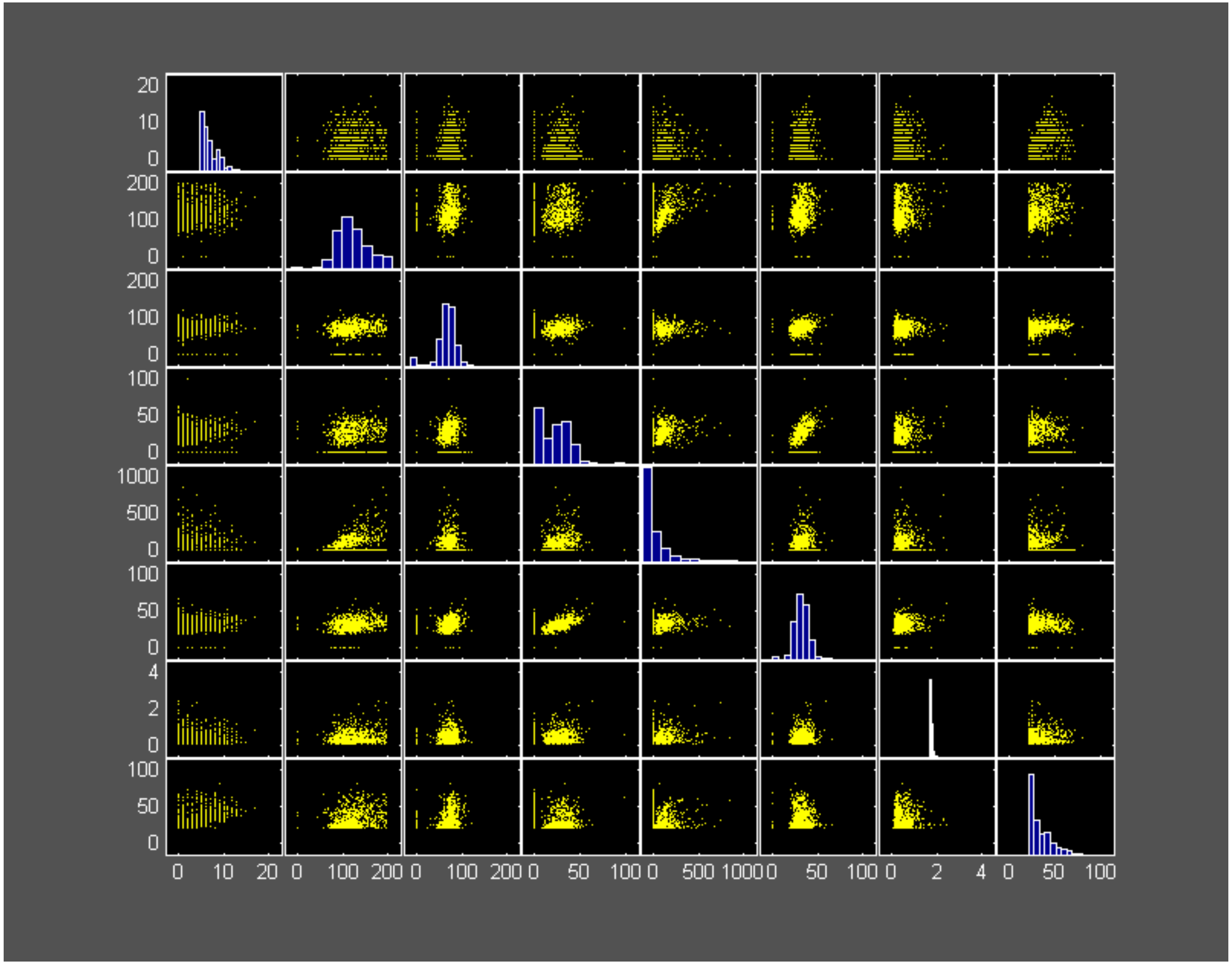
- Visualization is widely used
  - 1d histograms
  - 2d scatter plots
  - Higher-dimensional methods

- Useful for data checking
  - Finding the some variables are highly skewed

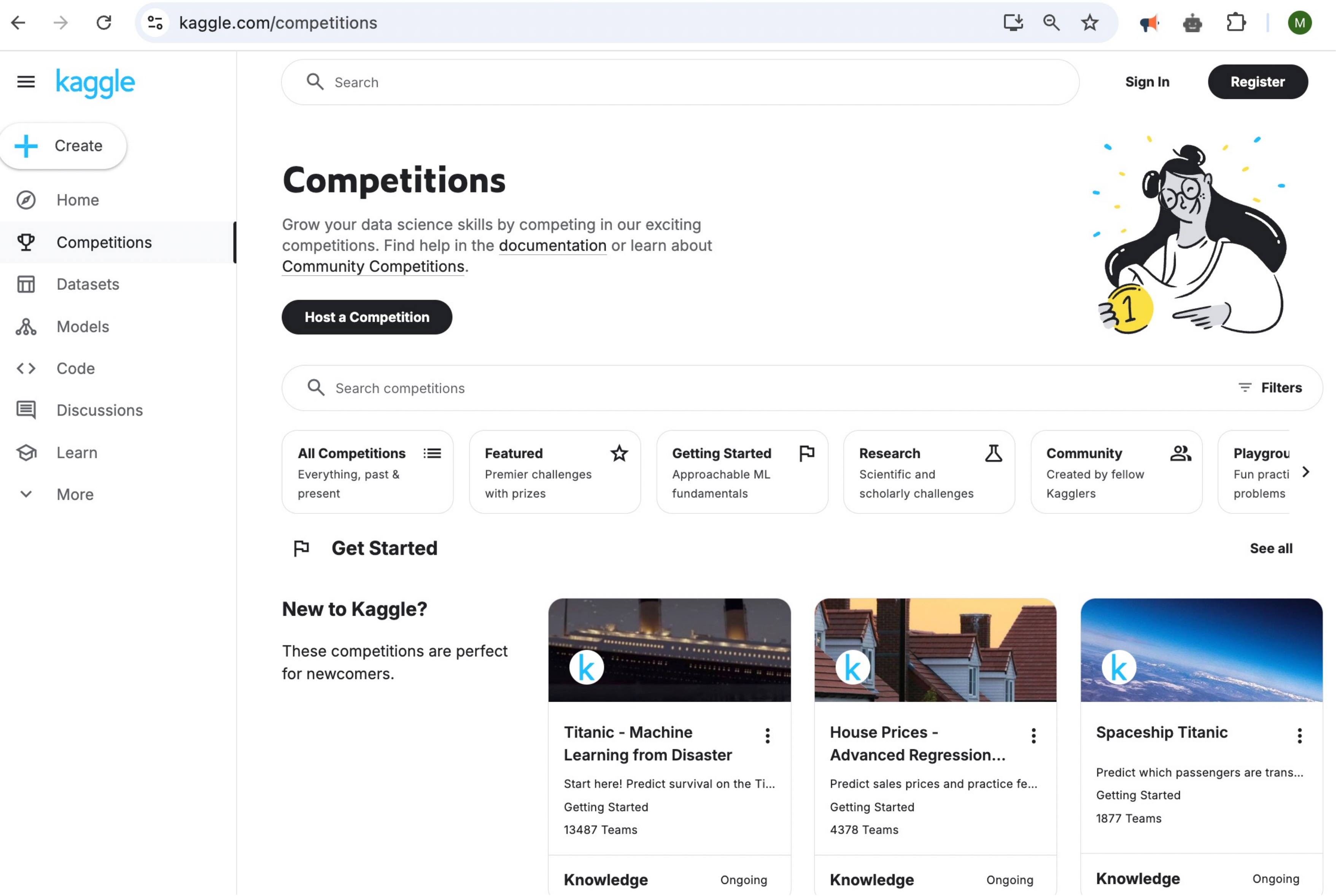
- Simple exploratory analysis is extremely valuable
  - You should always "look" at your data before applying any machine learning

Example of Exploratory Data Analysis

(Pima Indians data, scatter plot matrix)



Machine Learning for solving real problems



Competitions

Grow your data science skills by competing in our exciting competitions. Find help in the documentation or learn about Community Competitions.

All Competitions Everything, past & present

Featured Premier challenges with prizes

Getting Started Approachable ML fundamentals

Research Scientific and scholarly challenges

Community Created by fellow Kagglers

Playgrou Fun practi problems

Get Started

New to Kaggle?

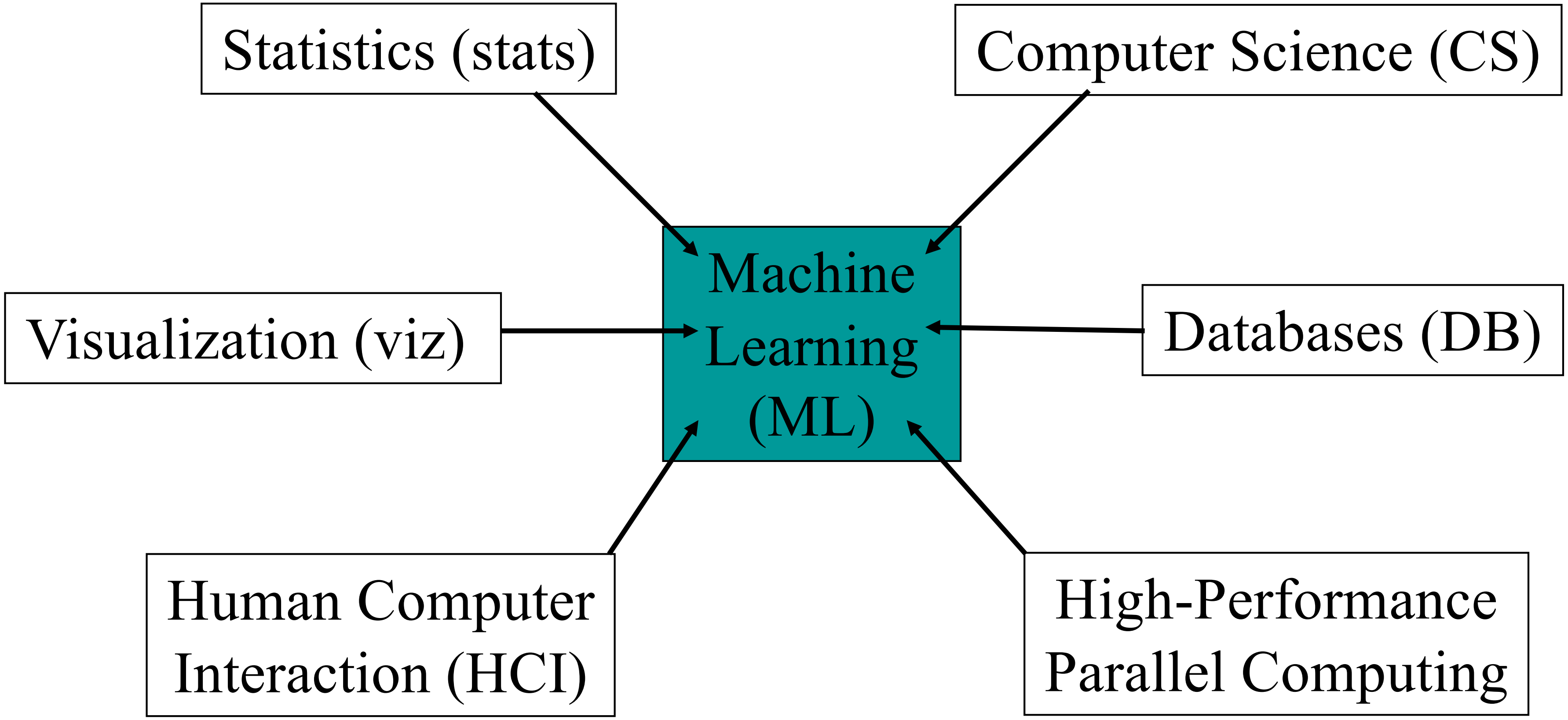
These competitions are perfect for newcomers.

Titanic - Machine Learning from Disaster
Start here! Predict survival on the Ti...
Getting Started
13487 Teams
Knowledge Ongoing

House Prices - Advanced Regression...
Predict sales prices and practice fe...
Getting Started
4378 Teams
Knowledge Ongoing

Spaceship Titanic
Predict which passengers are trans...
Getting Started
1877 Teams
Knowledge Ongoing

ML/DL: Intersection of Many Fields



Statistics (stats)

Computer Science (CS)

Visualization (viz)

Machine Learning (ML)

Databases (DB)

Human Computer Interaction (HCI)

High-Performance Parallel Computing

Machine learning

- Tom Mitchell(1998): A computer program is said to learn from experience **E** with respect to some task **T** and some performance measure **P**, if its performance on T, as measured by P, *improves with experience* E.

email spam prediction

- **Task**: email classification to spam/no-spam
- **Experience**: the user's action to characterize emails
- **Performance**: # of emails characterized as spam correctly.

What is Machine Learning

- "…*computational methods using experience to improve performance or to make accurate predictions*" **Mohri et. al**. (2012)
- experience: past information available to the learner, in the form of electronic data collected and made available for analysis.
- Data quality and size are crucial to the success of the predictions made by the learner.
- Machine learning consists of
  - designing efficient & accurate prediction *algorithms* - time and space complexity.
  - Additionally - sample complexity to evaluate the sample size required for the algorithm to learn a family of concepts.

  More generally, theoretical learning guarantees for an algorithm depend on the complexity of the concept classes and the size of the training sample.
- learning techniques are data-driven methods combining fundamental concepts in computer science with ideas from statistics, probability and optimization.

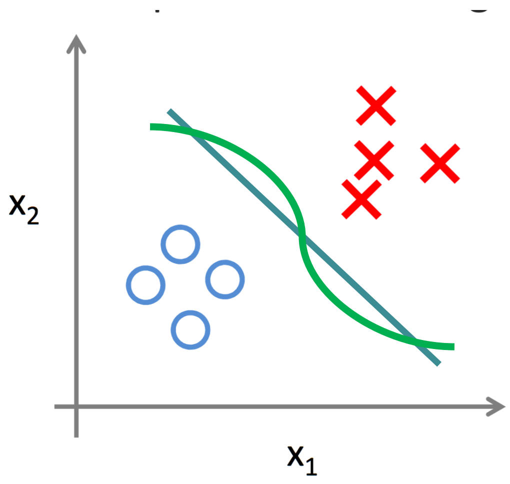
Applications of Machine/Deep Learning

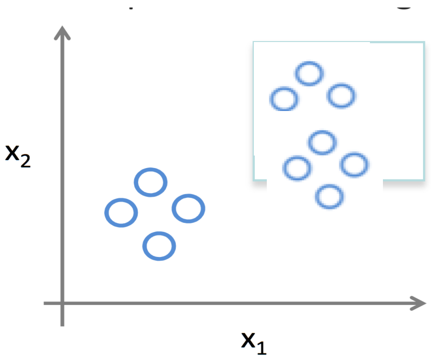
- Text or document classification, e.g., spam detection;
- Natural language processing, e.g., morphological analysis, part-of-speech tagging, statistical parsing, named-entity recognition
- Recommendation systems, search engines, information extraction systems
- Fraud detection (credit card, telephone) and network intrusion
- Speech recognition, speech synthesis, speaker verification;
- Optical character recognition (OCR);
- Computational biology applications, e.g., protein function or structured prediction, Medical diagnosis;
- Computer vision tasks, e.g., image recognition, face detection;
- Games, e.g., chess, backgammon;
- Unassisted vehicle control (robots, navigation);
- …

ML Tasks

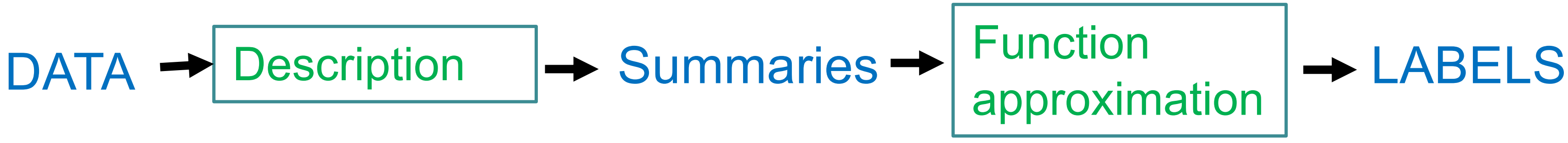
Main Tasks

- Supervised Learning - Approximation
- Unsupervised Learning – Description





- Supervised & unsupervised learning synergy



DATA → Description → Summaries → Function approximation → LABELS

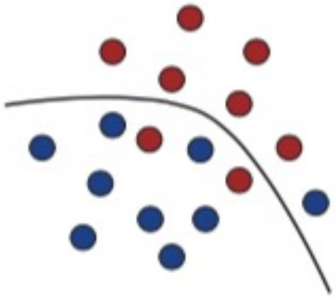
Unsupervised learning

Supervised learning

More ML/DL Tasks

- Structured output/generative models
  - Transcription (optical character, speech recognition – out put is characters)
  - Machine translation, Summarization
- Anomaly detection
  - flags data as unusual/atypical – i.e. credit card fraud detection.
- Synthesis and sampling
  - generate new examples that are similar to those in the training data.
  - useful for media applications: video games automatically generate textures for landscapes
- Imputation of missing values
  - prediction of the values of the missing entries.
- De-noising

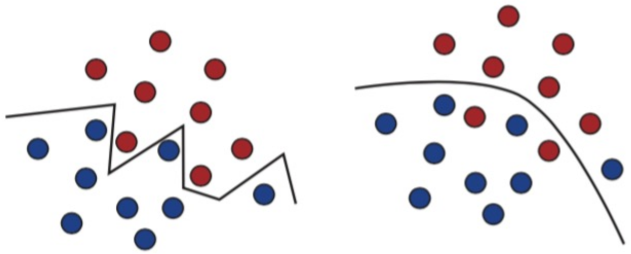
Machine Learning example



- Red and blue dots - training set
- Red/Blue - labels/classes
- Features: the space in which the training set is embedded (i.e. the (x,y) coordinates for this example)
- Objective: Learn a model (a function) f that based on the position of a sample decides the class of the point.
- Test sample: Examples to evaluate the performance of a learning algorithm - separate from the training and not made available in the learning stage

Foundations of Machine Learning - Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar, MIT Press, 2012

Machine Learning example



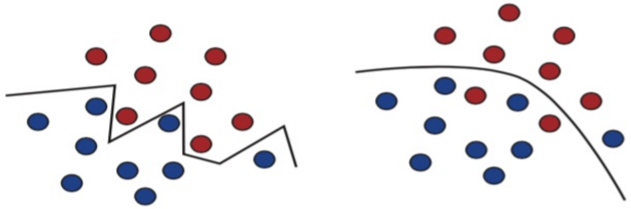
- Loss function: function *L* measures error/loss between a predicted and a true label. Let *Y/Y'* true/predicted labels:

$$L = YxY' \rightarrow R_+$$

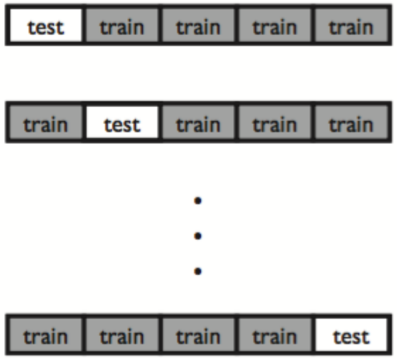
  - Square loss: $E = \sum_{i=1}^{k} (y(i) - y'(i))^2$
  - Other loss functions: Hinge, Logistic, Cross entropy…
- Hypothesis set: set of functions mapping features to labels (i.e. points to blue/red)
- Over fitting vs generalization: a function may be consistent (i.e. zero training error) but not generalize well…

Foundations of Machine Learning - Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar, MIT Press, 2012

Machine Learning example



- Cross-validation: in many cases not enough training data.
  - Split the *m* data into *n* subsets(folds) and let θ the model parameters
  - Train the algorithm for *n-1* folds and test on the *n-th*
  - Compute the cross validation error
  - Choose parameters θ that minimize the CV error



test train train train train

train test train train train

⋮

train train train train test

$$\widehat{R}_{\mathrm{CV}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\frac{1}{m_i}\sum_{j=1}^{m_i} L(h_i(x_{ij}), y_{ij})}_{\text{error of } h_i \text{ on the } i\text{th fold}} \, .$$

Error Optimization – gradient descent

- Learning & Optimization: Assume ***J(θ)*** the objective error function, ***θ*** hypothesis parameters.
- Objective: find θ that minimizes ***J(θ)***:
  - update the parameters in the opposite direction of the gradient of the objective function: $\boldsymbol{\nabla_\theta J(\theta)}$ w.r.t. to the parameters
  - Batch gradient descent

$$\theta = \theta - \eta \bigtriangledown_\theta J(\theta)$$

  - ***η*** the learning rate
  - *Redundant computations:* re-computes gradients for similar examples before each parameter update.

Error Optimization – gradient descent

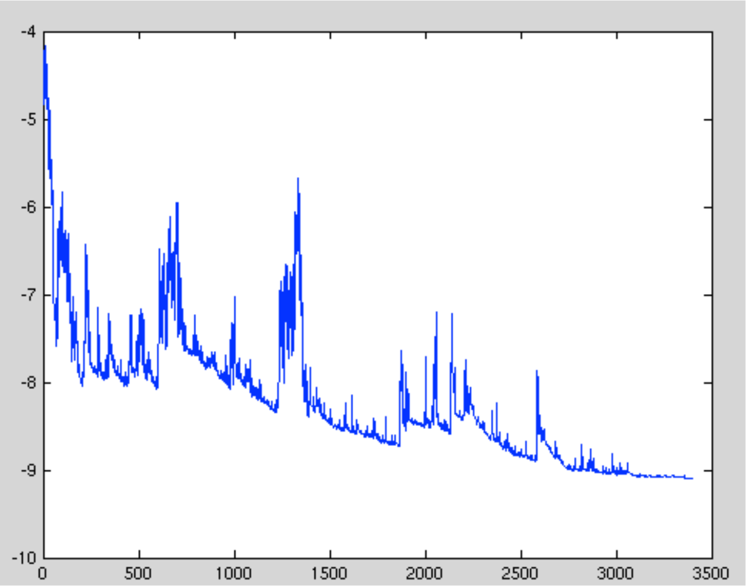
– Stochastic gradient descent: update the parameters for each training data point *(x(i), y(i))*

$$\theta = \theta - \eta \bigtriangledown_\theta J(\theta, x(i), y(i))$$

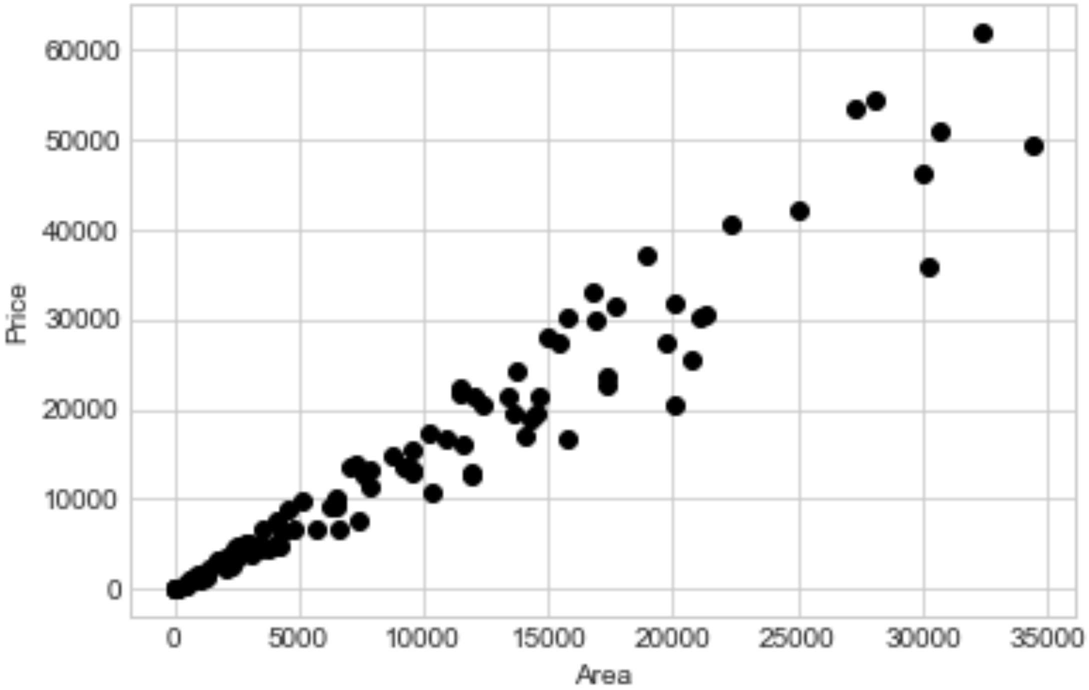
– One update at a time, faster

– High variance - fluctuation

- *Other optimization methods*
    - *Expectation Maximization*

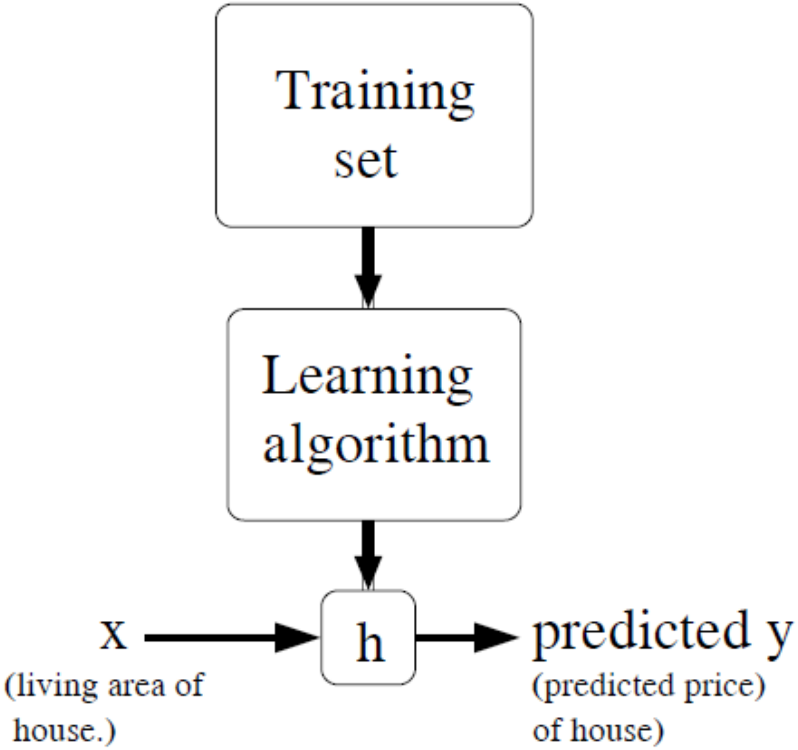


Supervised learning - prediction



- Can we predict the price of a house based on its size (surface in m^2) ?

Supervised learning - prediction



- y continuous value: ***prediction***
- y discrete value: ***classification***

Prediction

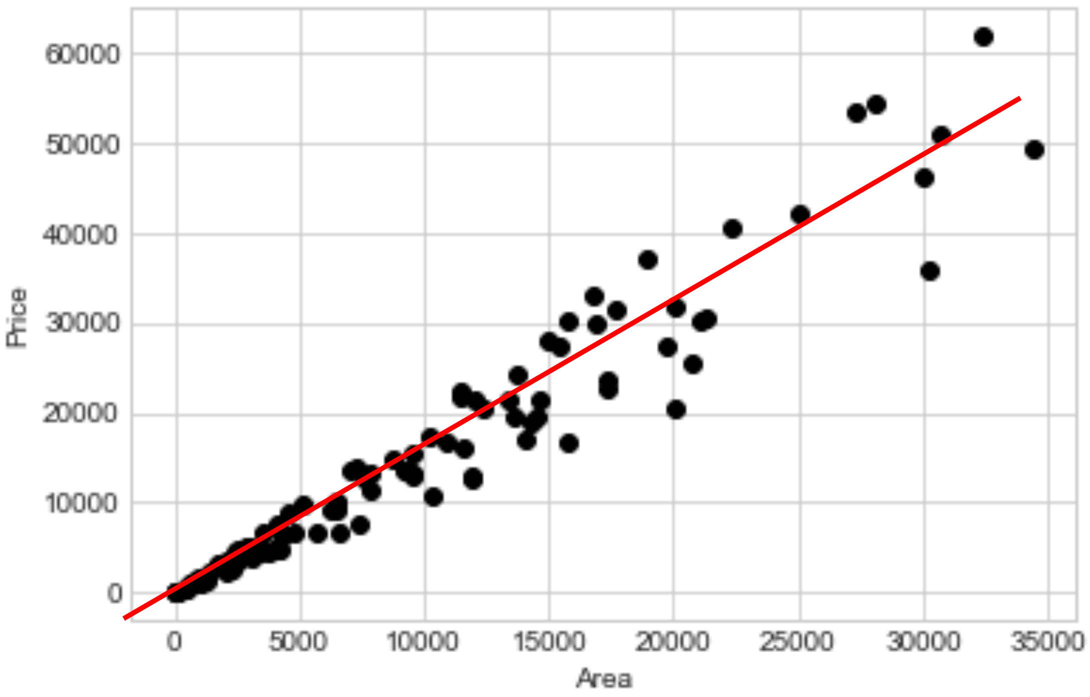
- *x(i)*: "input" variables (input features)
- *y(i)*: "output" or target variable - trying to predict
- pair *(x(i), y(i))*: training example
- *training set* - set of m training examples *{(x(i), y(i)); i =1, . . . ,m}*.
- *X*: space of input values, *Y*: output values.
- The supervised learning problem:
  - given a training set,
  - learn a function $h : X \rightarrow Y$: *h(x)* "good" predictor for the corresponding value of *y* - *h* also called *hypothesis*.

Prediction with Regression

- Aims at fitting a line to a set of observations

$\{(x_1, y_1),\ \ldots,\ (x_N, y_N)\},$ there is a straight line

$y = ax + b.$



Regression

- the individual point error is: $y - (ax + b)$
- thus the error set is: $\{y_1 - (ax_1 + b), ...y_n - (ax_n + b)\}$
- the total error is: $E(a,b) = \sum_{n=1}^{N} (y_n - (ax_n + b))^2$

Based on notes at: http://www.williams.edu/go/math/sjmiller/public_html/BrownClasses/54/handouts/MethodLeastSquares.pdf

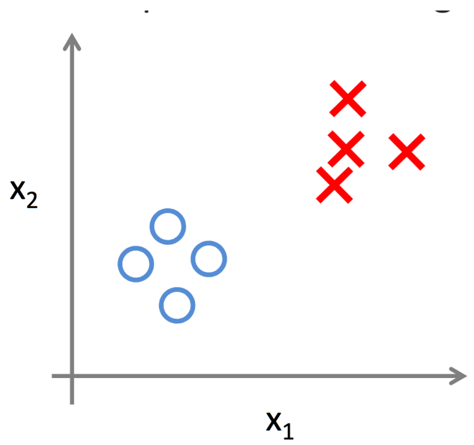
Least Squares method

- The objective is to minimize $E(a,b) = \sum_{n=1}^{N} (y_n - (ax_n + b))^2$
- Thus to find values α, b such that: $\frac{\partial E}{\partial a} = 0, \quad \frac{\partial E}{\partial b} = 0.$
- Differentiation leads to:

$$\begin{aligned} \frac{\partial E}{\partial a} &= \sum_{n=1}^{N} 2\left(y_n - (ax_n + b)\right) \cdot (-x_n) \\ \frac{\partial E}{\partial b} &= \sum_{n=1}^{N} 2\left(y_n - (ax_n + b)\right) \cdot 1. \end{aligned}$$

Supervised Learning - Classification

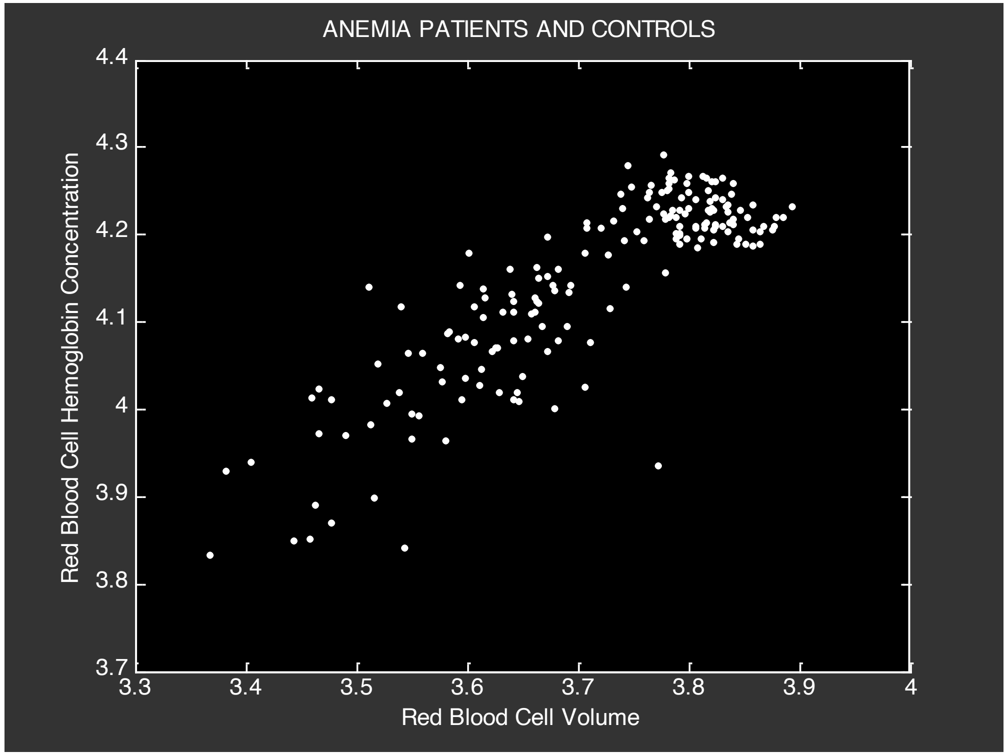
- Class-conditional/probabilistic, based on p($\underline{x}$ | c_k),
  - Naïve Bayes (simple, but often effective in high dimensions)
  - Parametric generative models, e.g., Gaussian (can be effective in low-dimensional problems)
- Discriminative models, focus on locating optimal decision boundaries
  - Linear discriminants, perceptron: simple, sometimes effective
  - Support vector machines: generalization of linear discriminants, can be quite effective, computational complexity is an issue
  - Nearest neighbor: simple, can scale poorly in high dimensions
  - Decision trees: learning splitting of the data that maximize information learning, effective in high dimensions

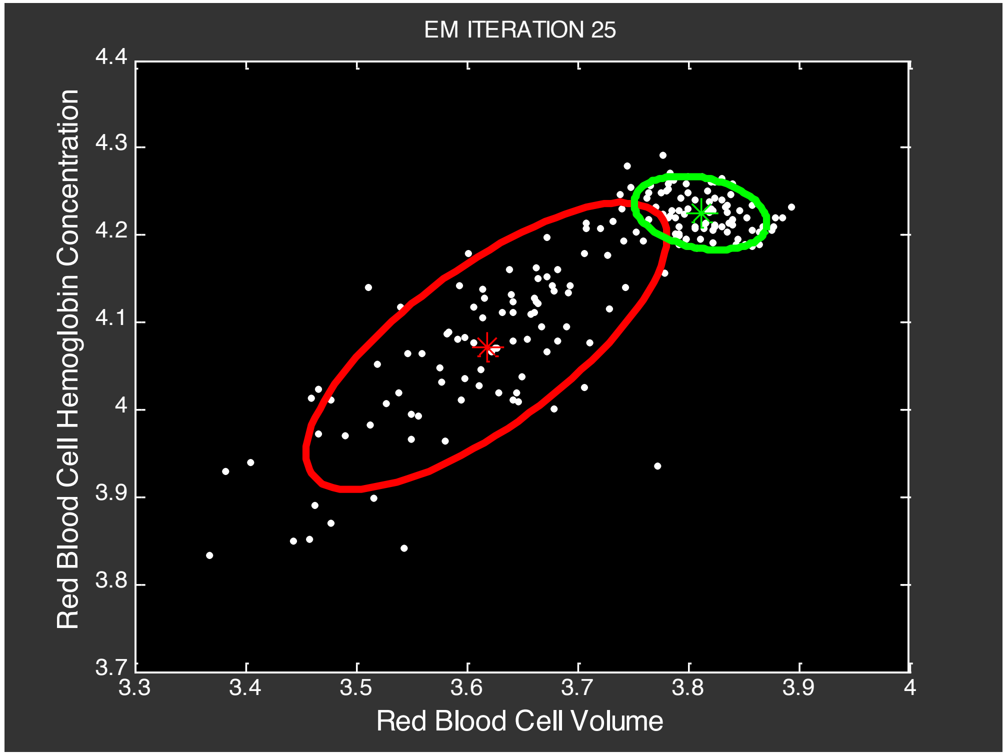


Unsupervised learning

- Learn a “generative” or “descriptive” model,
  - E.g., a model to simulate/generate the data if needed
  - Model underlying processes
- Examples:
  - Density estimation:
    - estimate the joint data distribution P(x_1,……x_p)
  - Cluster analysis:
    - Find natural groups in the data
  - Dimensionality reduction
    - Learn latent spaces (SVD, PCA,)
    - Word/document embeddings, auto-encoders….

Unsupervised learning - clustering





Images from P. Smyth, ICS278, slides

Unsupervised learning - Pattern Discovery

```
ADACABDABAABBDDBCADDDDBCDDBCCBBCCDADADAADABDBBDABABBCDD
DCDDABDCBBDBDBCBBABBBCBBABCBBACBBDBAACCADDADBDBBCBBCCBB
BDCABDDBBADDBBBBCCACDABBABDDCDDBBABDBDDBDDBCACDBBCCBBAC
DCADCBACCADCCCACCDDADCBCADADBAACCDDDCBDBDCCCCACACACCDAB
DDBCADADBCBDDADABCCABDAACABCABACBDDDCBADCBDADDDDCDDCADC
CBBADABBAAADAAABCCBCABDBAADCBCDACBCABABCCBACBDABDDDADAA
BADCDCCDBBCDBDADDCCBBCDBAADADBCAAAADBDCADBDBBBCDCCBCCCD
CCADAADACABDABAABBDDBCADDDDBCDDBCCBBCCDADADACCCDABAABBC
BDBDBADBBBBCDADABABBDACDCDDDBBCDBBCBBCCDABCADDADBACBBBC
CDBAAADDDBDDCABACBCADCDCBAAADCADDADAABBACCBB
```

Unsupervised learning - Pattern Discovery

ADACABDABAABBDDBCADDDDBCDDBCCBBCCDADADAADABDBBDABABBCDD

DCDDABDCBBDBDBCBBABBBCBBABCBBACBBDBAACCADDADBDBBCBBCCBB

BDCABDDBBADDBBBBCCACDABBABDDCDDBBABDBDDBDDBCACDBBCCBBAC

DCADCBACCADCCCACCDDADCBCADADBAACCDDDCBDBDCCCCACACACCDAB

DDBCADADBCBDDADABCCABDAACABCABACBDDDCBADCBDADDDDCDDCADC

CBBADABBAAADAAABCCBCABDBAADCBCDACBCABABCCBACBDABDDDADAA

BADCDCCDBBCDBDADDCCBBCDBAADADBCAAAADBDCADBDBBBCDCCBCCCD

CCADAADACABDABAABBDDBCADDDDBCDDBCCBBCCDADADACCCDABAABBC

BDBDBADBBBBCDADABABBDACDCDDDBBCDBBCBBCCDABCADDADBACBBBC

CDBAAADDDBDDCABACBCADCDCBAAADCADDADAABBACCBB

Machine Learning – quiz..

Of the following examples, which would you address with *unsupervised/supervised* learning?

1. Given email labeled as spam/not spam, learn a spam filter.

2. Given data of patients diagnosed with cancer, learn to classify new patients for this disease.

3. Given a set of news articles found on the web, group them into set of articles about the same story.

4.Given a database of customer data, discover market segments and group customers into different market segments.

Machine Learning – quiz..

- Classification or regression?
  - Credit history -> offer a loan?
  - Human face picture -> {kid, adolescent, adult}
  - Human face picture -> age

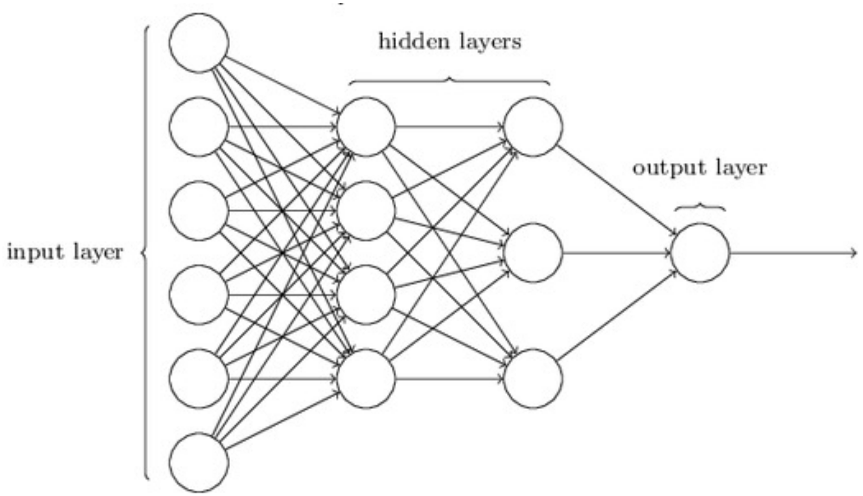
Discrete vs continuous output

More recent ML types of algorithms

- **Deep Learning**
- Reinforcement learning
  - Target, states, actions rewards, policy
  - Search for optimal policy..
- Adversarial learning
- …

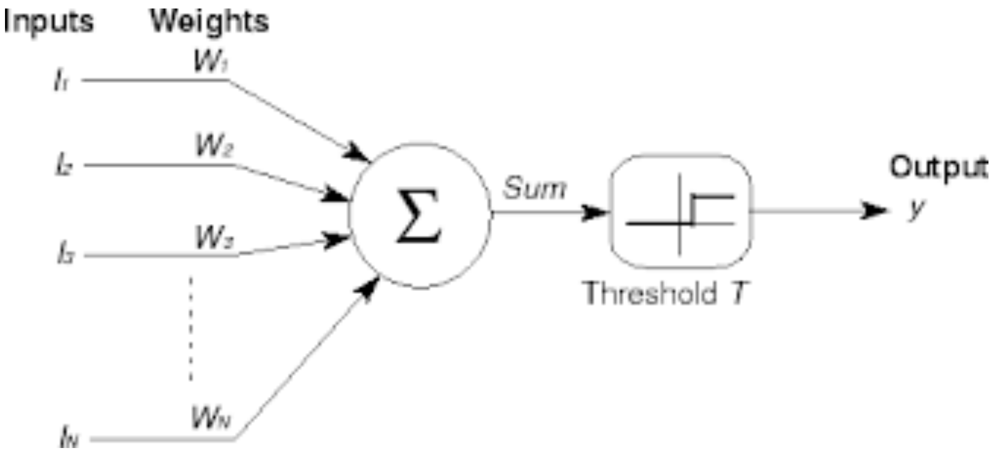
Deep Learning

- Based on perceptron – basic learning unit
- Layers of perceptrons learning a complex function *y=f(X)*



- *Learning features/embeddings*
- *Used for predictions*

Neural networks - perceptron



$$y = f(\sum_{j=1}^{n} w_j X_j + b)$$

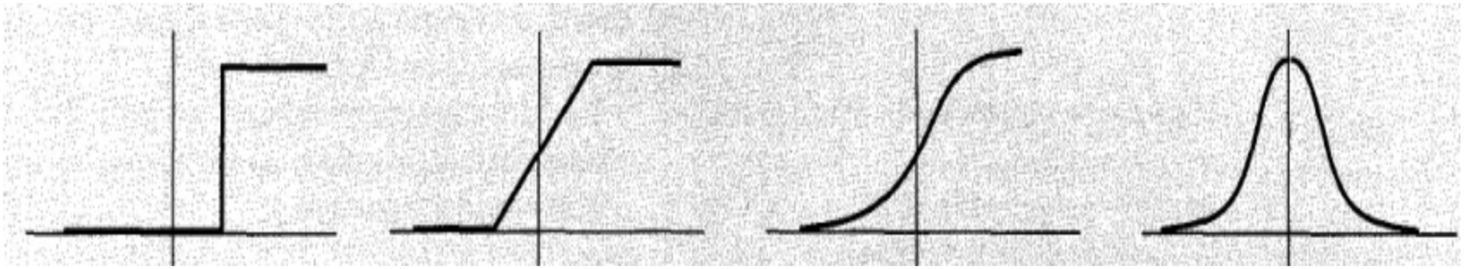
McCulloch-Pitts Neuron Model

-f: activation function

-w_j: weight of the j-th input Xj

-b : bias

- activation functions: piecewise linear, sigmoid, or Gaussian



Deep Learning

Dominant in recent years – AI based on DL architectures

Different architectures – non exhaustive

- Multilayer perceptrons (MLPs)
- Convolutional Neural Networks (CNNs)
- Recurrent Neural Networks (RNNs) – for sequential data
  - GRUs
  - LSTMs
- Attention based Architectures
  - Self Attention, Transformer (Bert, Barthez), ...
- Autoencoders
- **Graph Neural Networks**