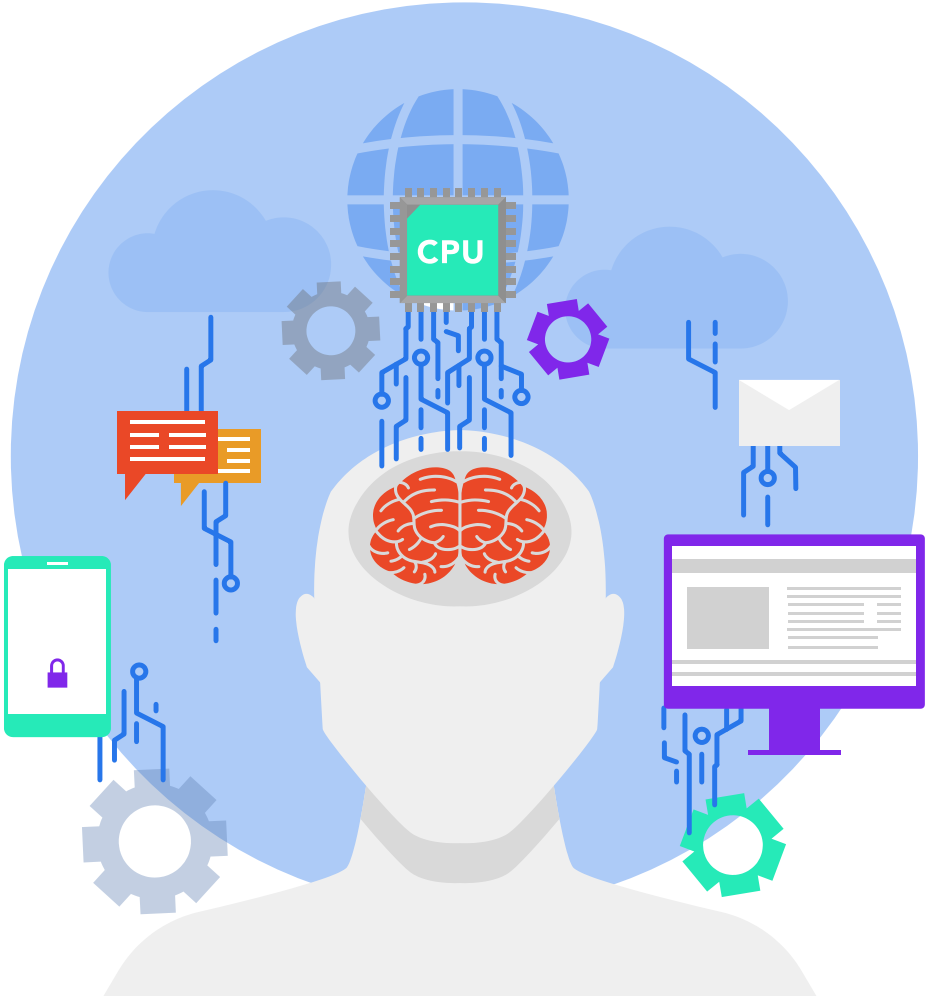



UNIVERSITÄT BONN MATTER

RHEINISCHE FRIEDRICH-WILHELMS-UNIVERSITÄT BONN

TRANSDISCIPLINARY RESEARCH AREA

# Introduction to Machine Learning

Bhoomika Basu Mallik

Journal Club

7 Oct 2022

# Timeline

01
**1950s**
Simple algorithms

02
**1960s**
Bayesian methods

03
**1970s**
AI winter caused by pessimism

04
**1980s**
Resurgence in ML research

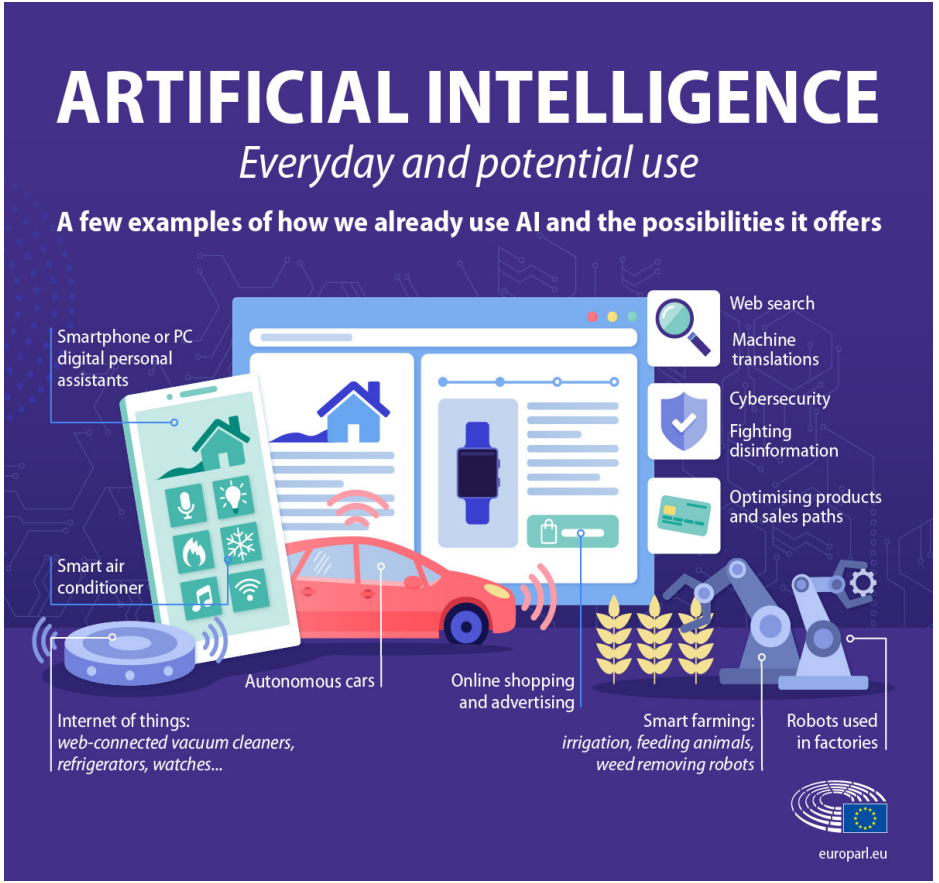
05
**1990s**
Data-driven approach

06
**2000s**
Support-Vector Clustering

07
**2010s**
Deep learning popularity

# Artificial Intelligence (AI)

- Simulation of human intelligence in machines
- Mimicking their actions
- Learning or problem solving



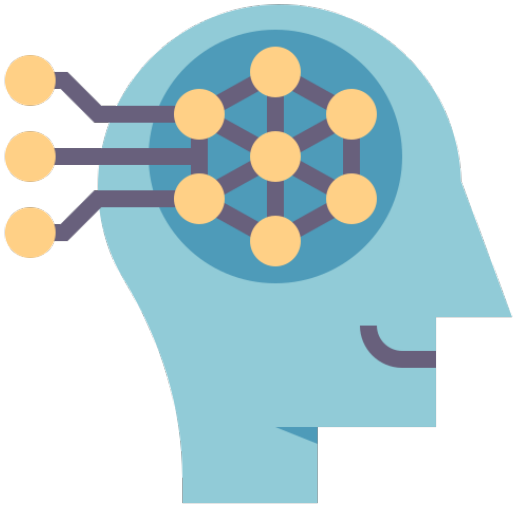
# What is Machine Learning ?

“Learning is any process by which a system improves performance from experience.”
-Herbert Simon

**Definition by Tom Mitchell (1998)**

Machine Learning is the study of algorithms that :
**improve their performance P**
**at some task T**
**with experience E**

**A well-defined learning task is given by <P, T, E>**



# Machine Learning is not equal to Artificial Intelligence



## Artificial Intelligence

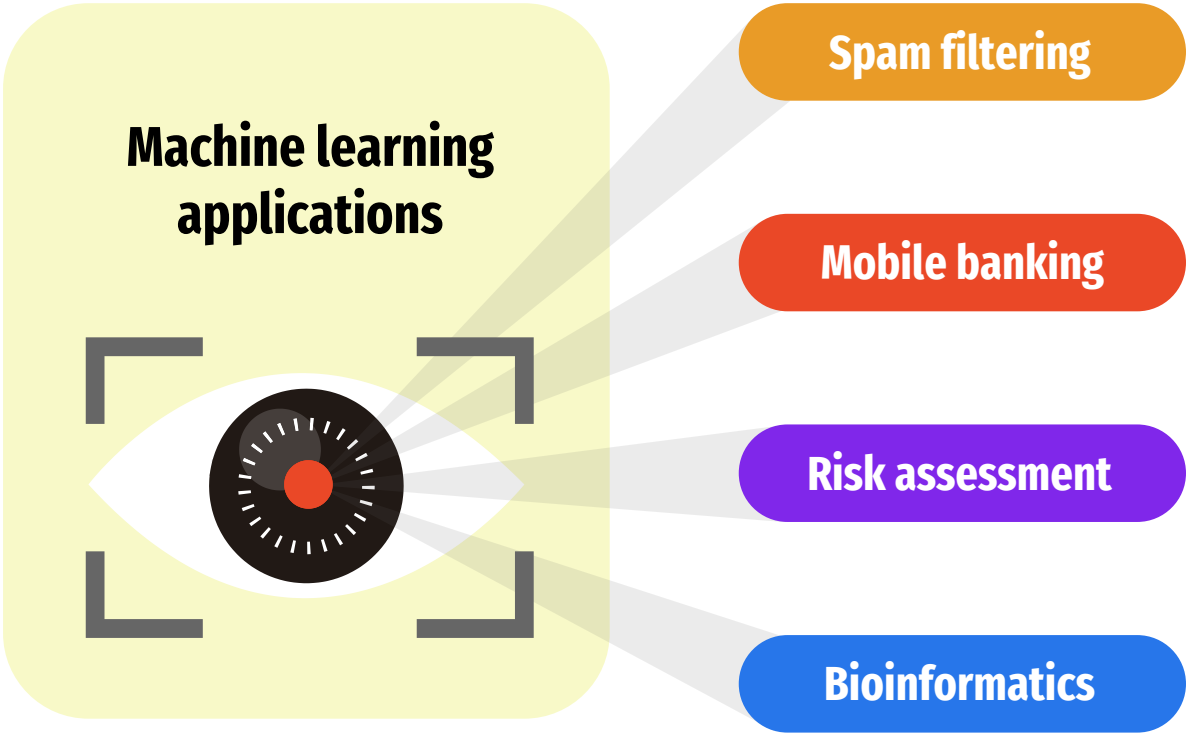
- Computers act on their own
- They act according to environment
- Systems display cognitive ability
- Computers make decisions

Vs

## Machine learning

- It’s an application of AI
- Computers observe and analyze
- Predict based on previous patterns
- Pre-programmed algorithms

# Few Applications

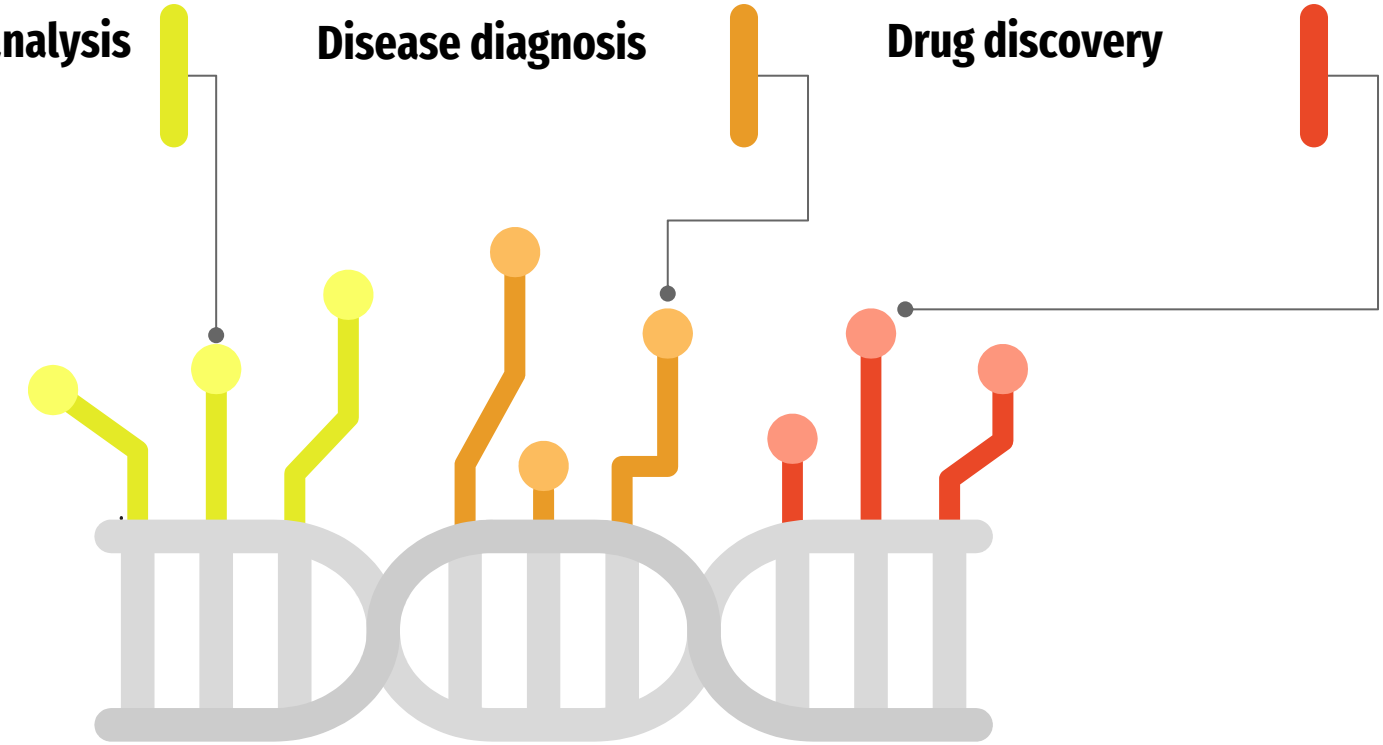


# Application in biology

Biochemical analysis

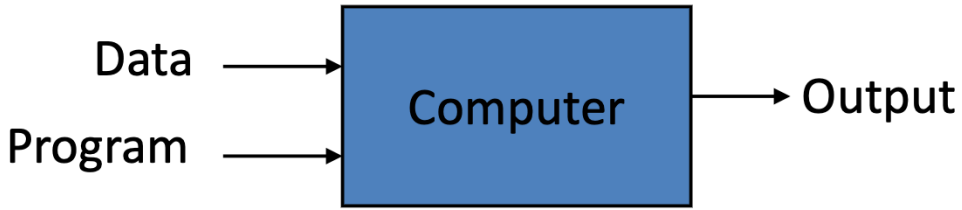
Disease diagnosis

Drug discovery

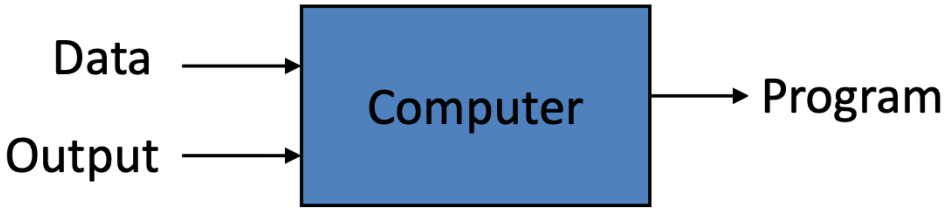


# Programming vs ML

## Traditional Programming

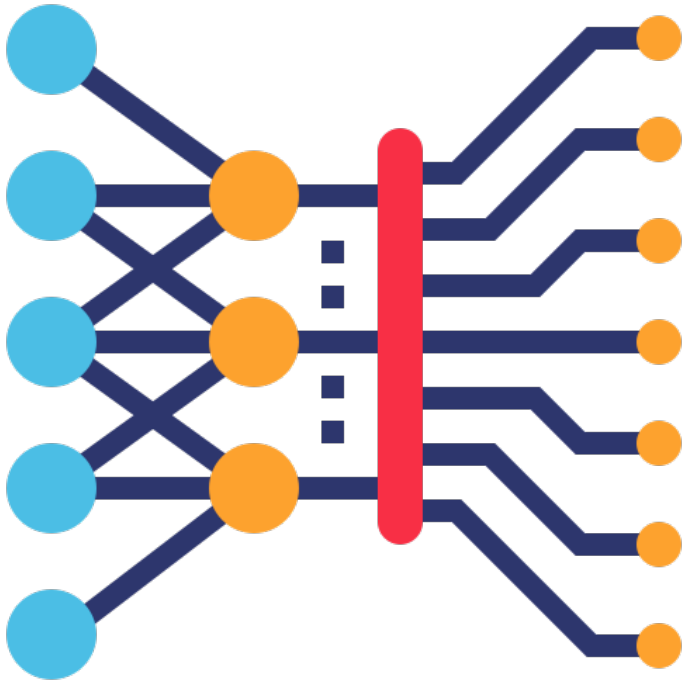



## Machine Learning

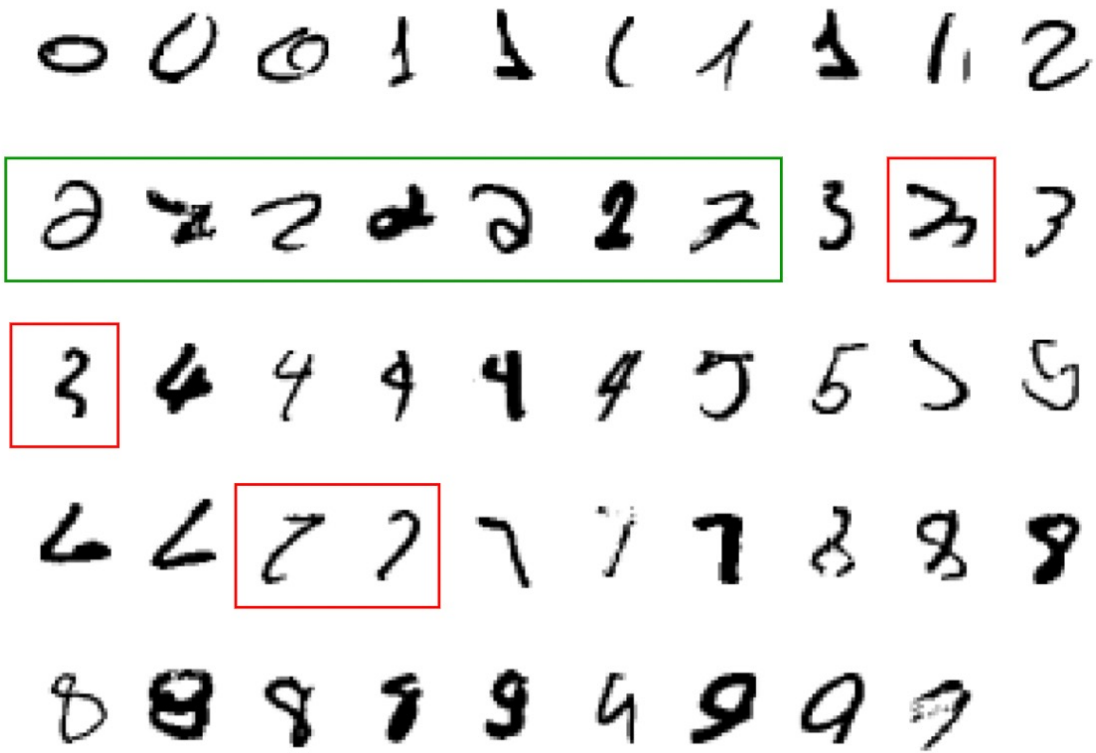


# The need to use ML

- Human expertise does not exist
- Humans can’t explain their expertise
- Models need to be customized
- Models are based on huge amounts of data



# Example: Identification of handwritten numbers

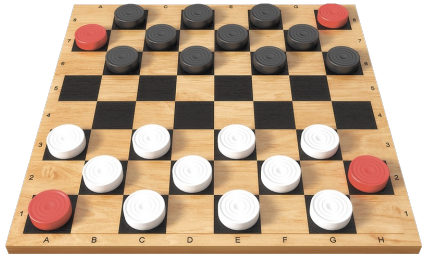




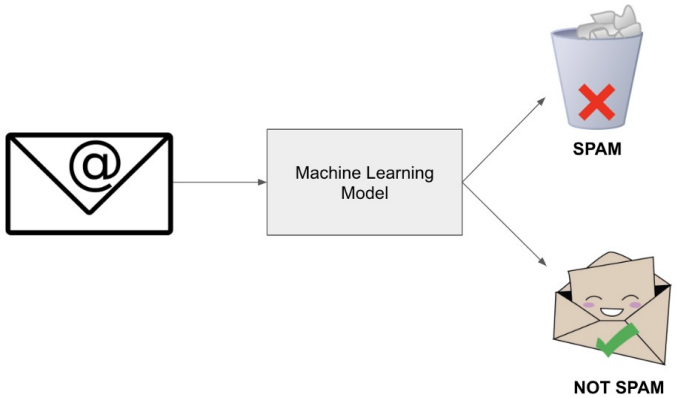
MNIST Handwritten Digit Classification Dataset

# Goal of ML

Improve on task T, with respect to performance metric P, based on experience E



T : Playing checkers
P : ?
E : ?




T: Categorize email messages as spam or legitimate
P : ?
E : ?

# Types of Machine Learning

## Supervised learning

### Classification

- Fraud detection
- Email spam detection
- Diagnostics
- Image classification

### Regression

- Risk assessment
- Score prediction

## Unsupervised learning

### Reduction

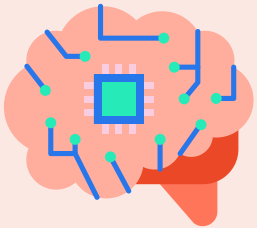
- Text mining
- Data visualization
- Face detection
- Voice detection

### Regression

- City planning
- Targeted marketing

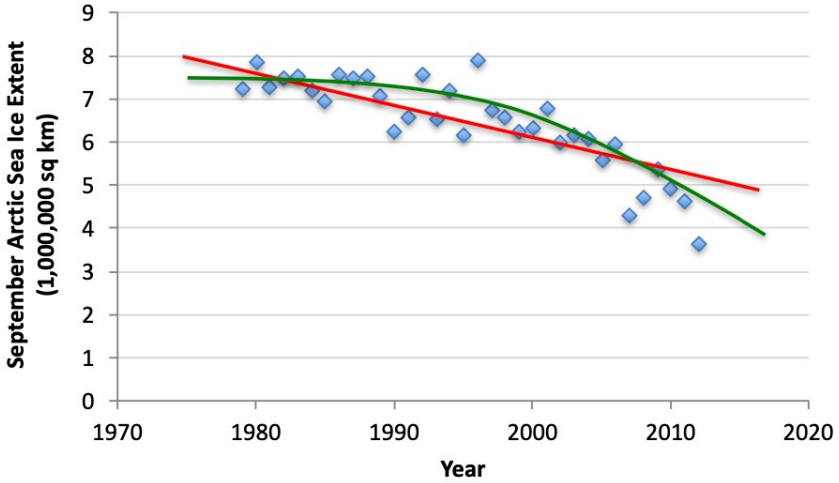
## Reinforcement learning

- Finances
- Manufacturing
- Stock  management
- Autonomous cars



# Supervised Learning : Regression

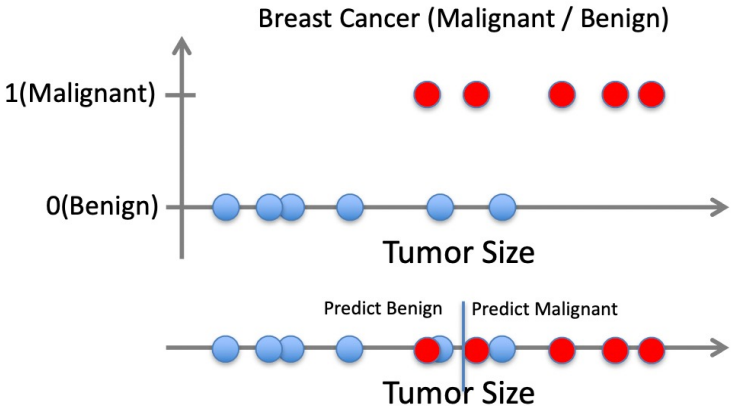
- Given $(x_1, y_1)$, $(x_2, y_2)$, …, $(x_n, y_n)$
- Learn a function $f(x)$ to predict $y$ given $x$
  - $y$ is real-valued == regression



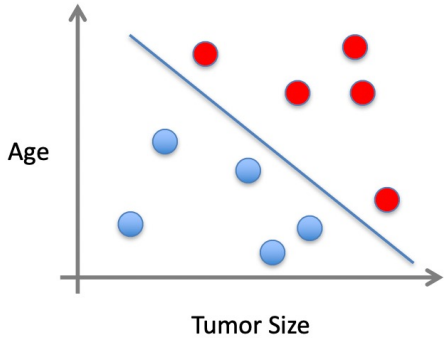

Data from G. Witt. Journal of Statistics Education, Volume 21, Number 1 (2013)

# Supervised Learning : Classification

- Given $(x_1, y_1)$, $(x_2, y_2)$, …, $(x_n, y_n)$
- Learn a function $f(x)$ to predict $y$ given $x$
  - $y$ is categorical == classification




- $x$ can be multi-dimensional
  - Each dimension corresponds to an attribute

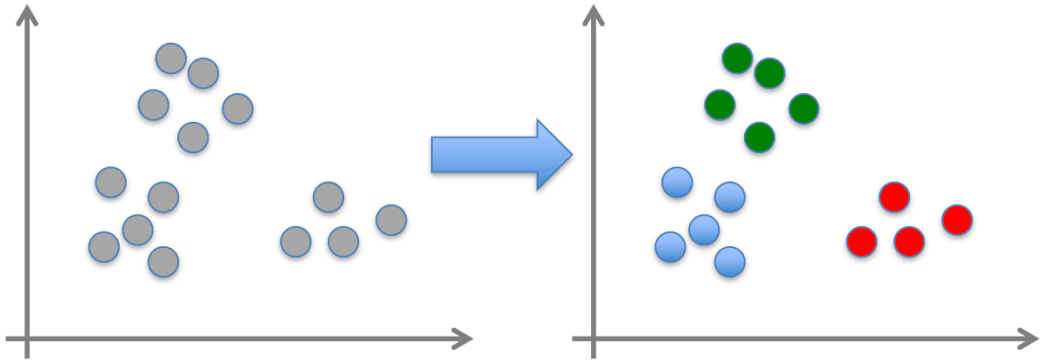



- Clump Thickness
- Uniformity of Cell Size
- Uniformity of Cell Shape

…

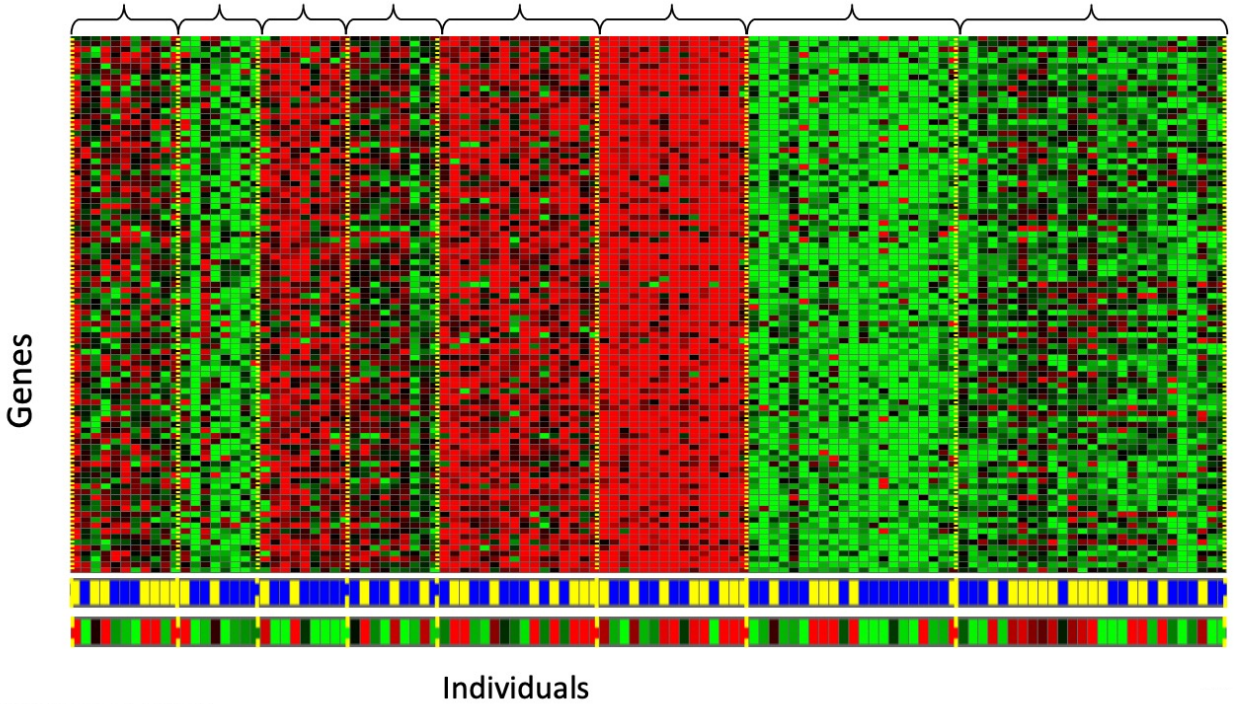
# Unsupervised Learning

- Given $x_1$, $x_2$, ..., $x_n$ (without labels)
- Output hidden structure behind the $x$'s
  - E.g., clustering



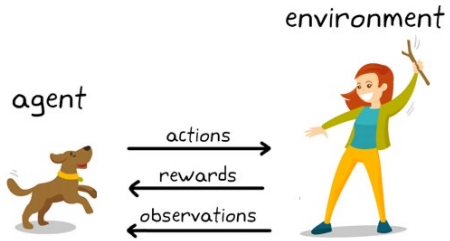
# RL Example : Unsupervised Learning

Genomics application: group individuals by genetic similarity




Source: Daphne Koller]

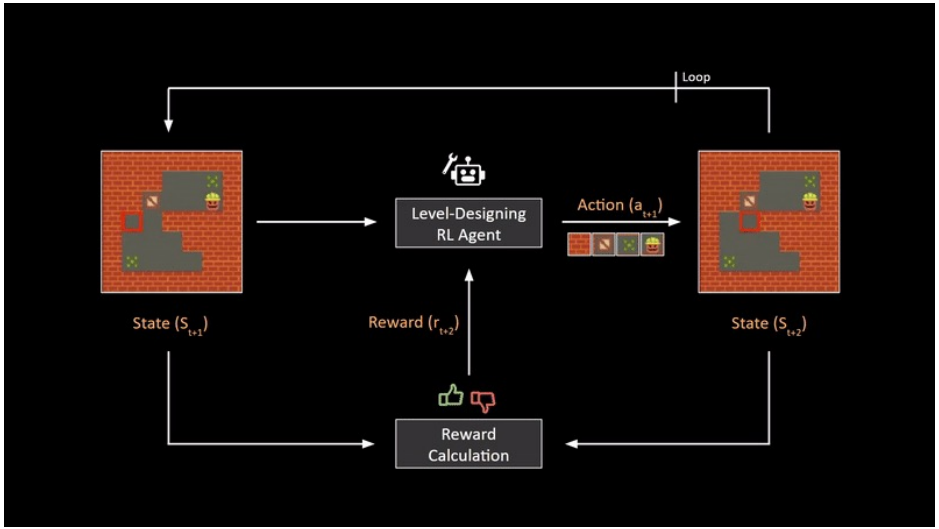
# Reinforcement learning




sequence of decisions

complex environment

rewards or penalties



# The Machine Learning Toolbox

## Functions

- Numerical functions
  - Linear regression
  - Neural networks
  - Support vector machines
- Symbolic functions
  - Decision trees
  - Rules in propositional logic
  - Rules in first-order predicate logic
- Instance-based functions
  - Nearest-neighbor
  - Case-based
- Probabilistic Graphical Models
  - Naïve Bayes
  - Bayesian networks
  - Hidden-Markov Models (HMMs)
  - Probabilistic Context Free Grammars (PCFGs)
  - Markov networks

## Algorithms

- Gradient descent
  - Perceptron
  - Backpropagation
- Dynamic Programming
  - HMM Learning
  - PCFG Learning
- Divide and Conquer
  - Decision tree induction
  - Rule learning
- Evolutionary Computation
  - Genetic Algorithms (GAs)
  - Genetic Programming (GP)
  - Neuro-evolution

# Commonly used platforms

**Programming languages**
Python
R
C++
...

**Many libraries**
scikit-learn

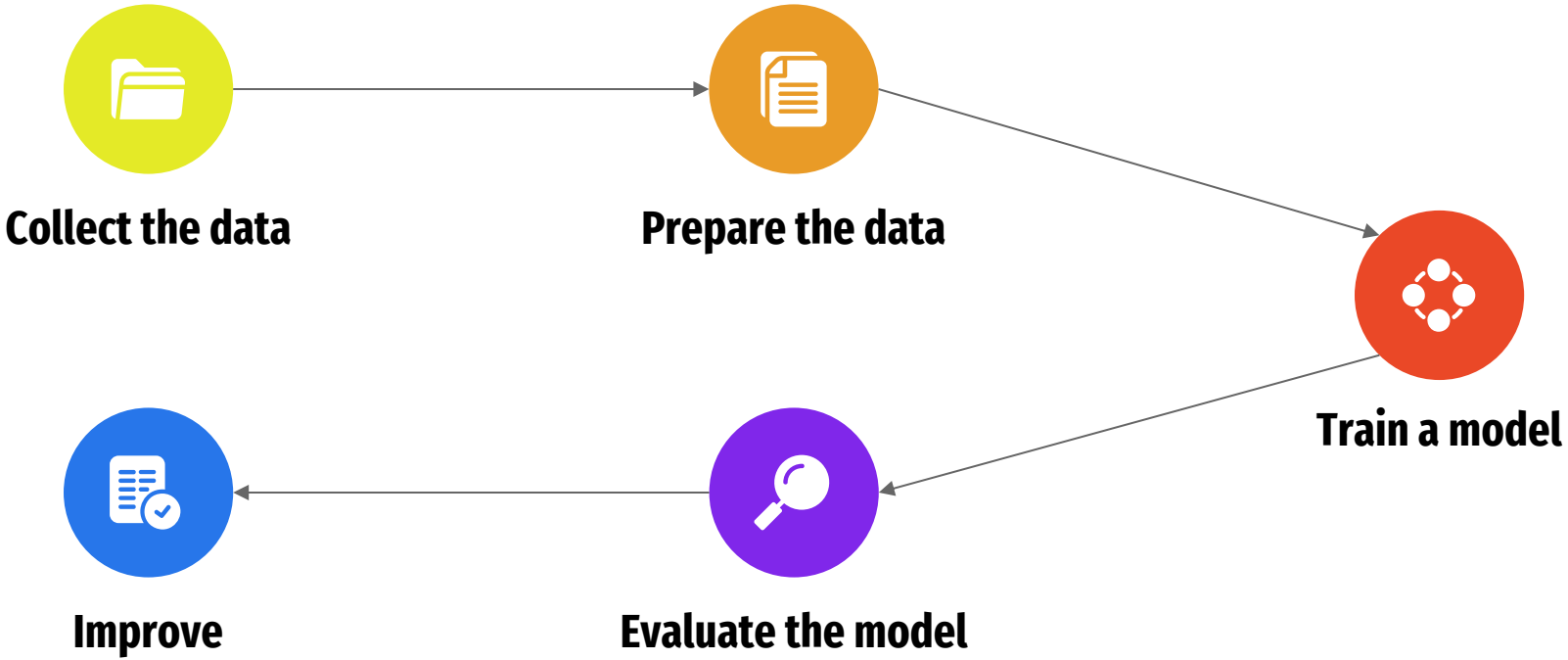
PyTorch
TensorFlow
Keras
...

classic machine learning

deep learning frameworks

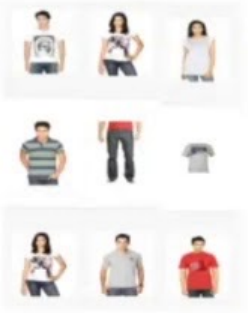


# General Workflow



# What is Data Labeling?

> **Data labeling**, also called **data annotation/tagging**, is the process of preparing labeled datasets for machine learning.



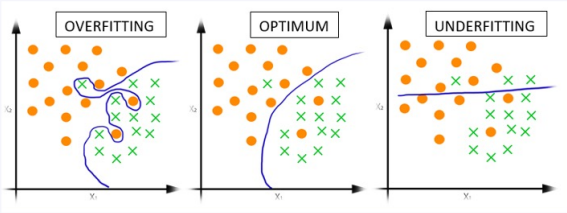
Images

Data labeling

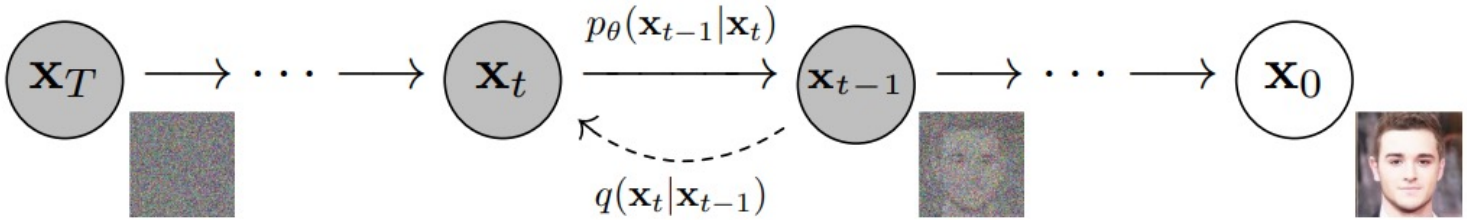

Image Classification
ML Model

# Supervised Learning : Method

- Select model, e.g., random forest, (deep) neural network, …
- Train model, i.e., determine parameters
  - Data: input + output
    - training data → determine model parameters
    - validation data → yardstick to avoid overfitting
- Test model
  - Data: input + output
    - testing data → final scoring of the model
- Production
  - Data: input → predict output

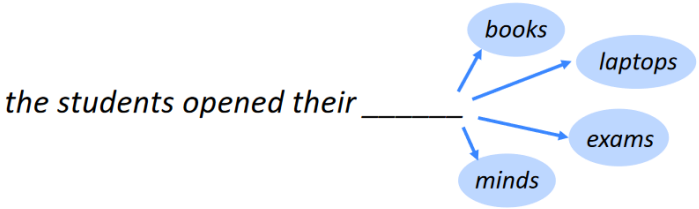


# Genarative Probabalistic: Diffusion Model

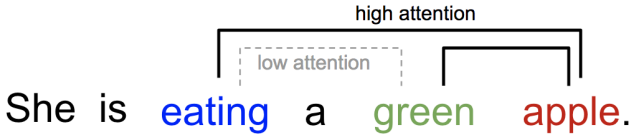



# Other models

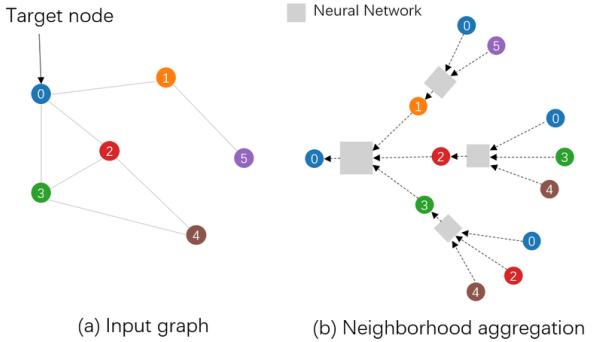
**Language model** (LM) **uses of various statistical and probabilistic techniques to determine the probability of a given sequence of words occurring in a sentence**.



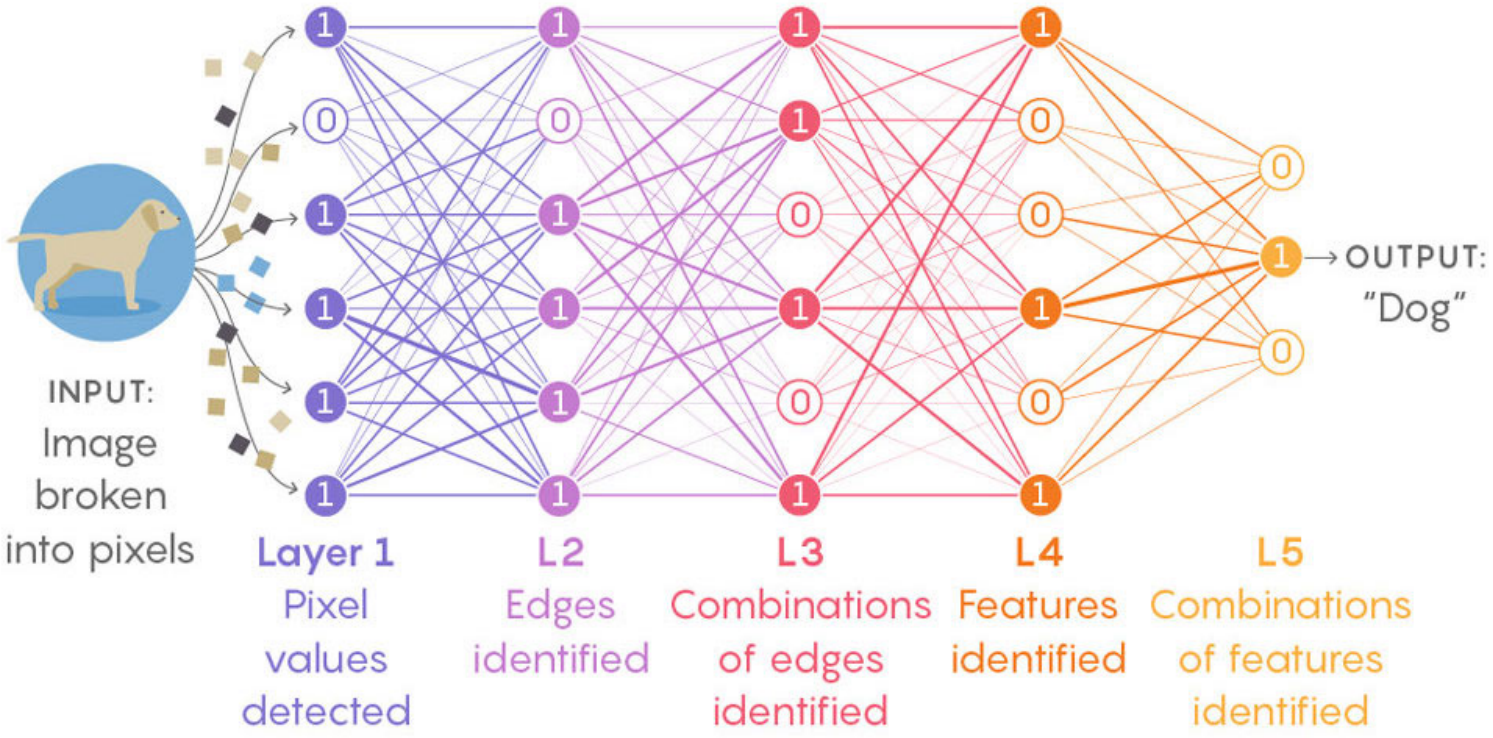

**Attention models**, also called attention mechanisms, are **deep learning techniques used to provide an additional focus on a specific component**.




**GNN and MPNN**, for processing data that can be represented as **graphs. Well known example: Protein MPNN**

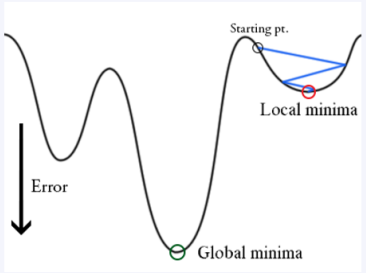


# Neural Networks



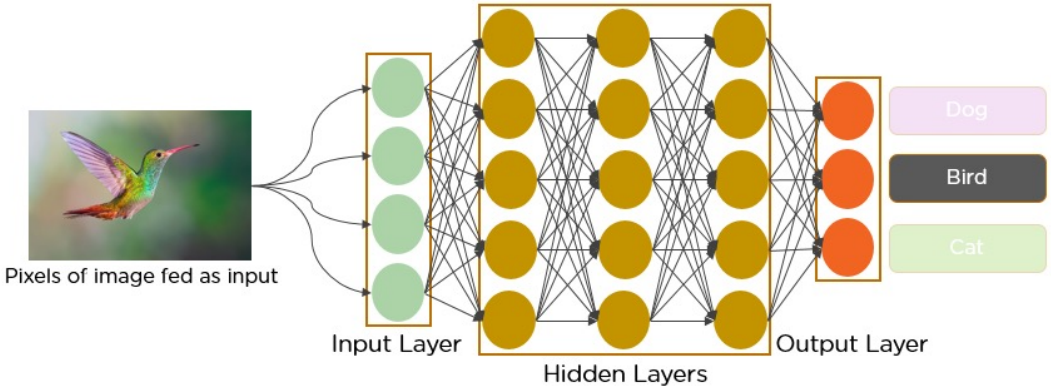
# How do we determine weights and biases ?

## Training: backpropagation




- Initialize weights "randomly"
- For all training epochs
  - for all input-output in training set
    - using input, compute output (forward)
    - compare computed output with training output
    - adapt weights (backward) to improve output
  - if accuracy is good enough, stop

# Convolutional Neural Network

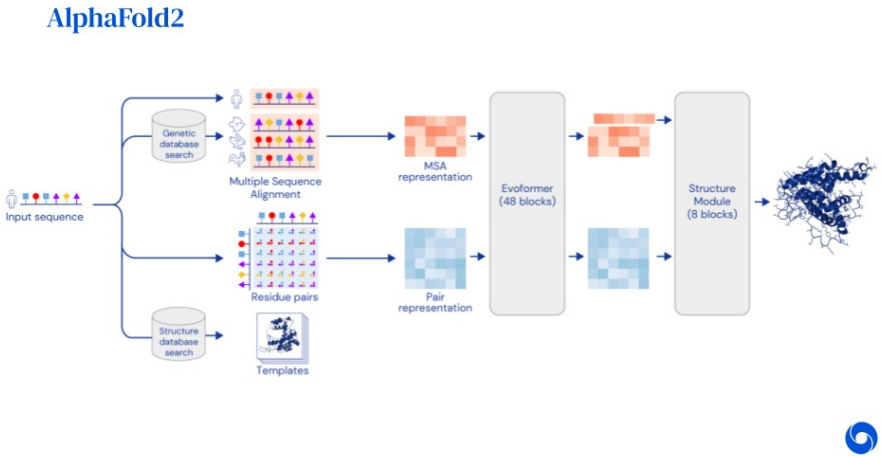



DALL·E 2

# Deep Neural Networks

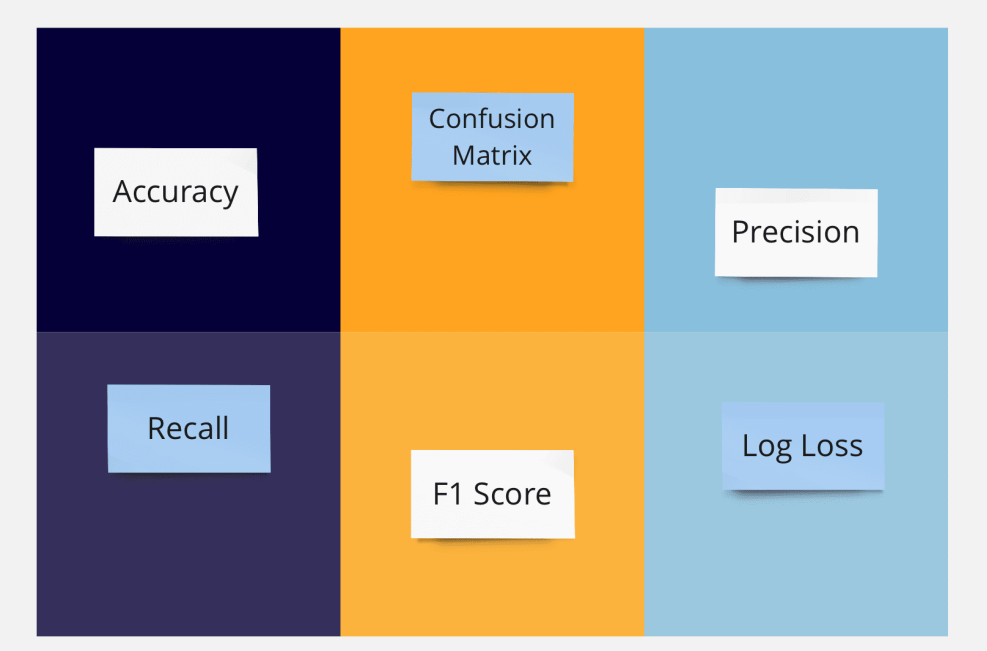
- Many layers
- Features are *learned*, not given
- Low-level features combined into high-level features

- Special types of layers
  - convolutional
  - drop-out
  - recurrent
  - …



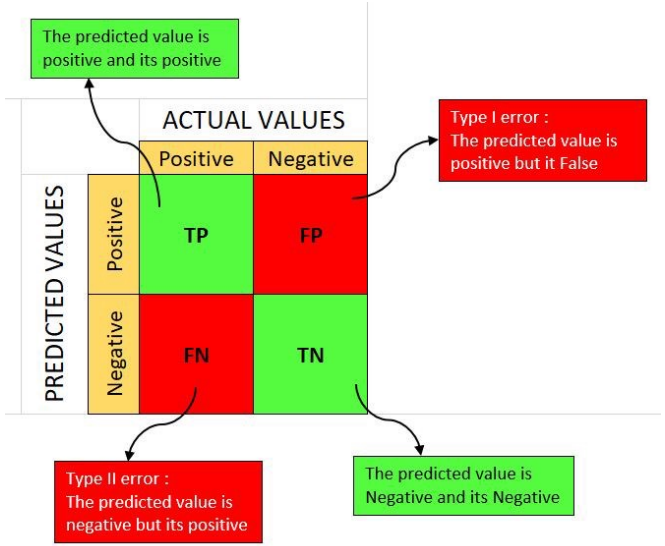
# Model Evaluation

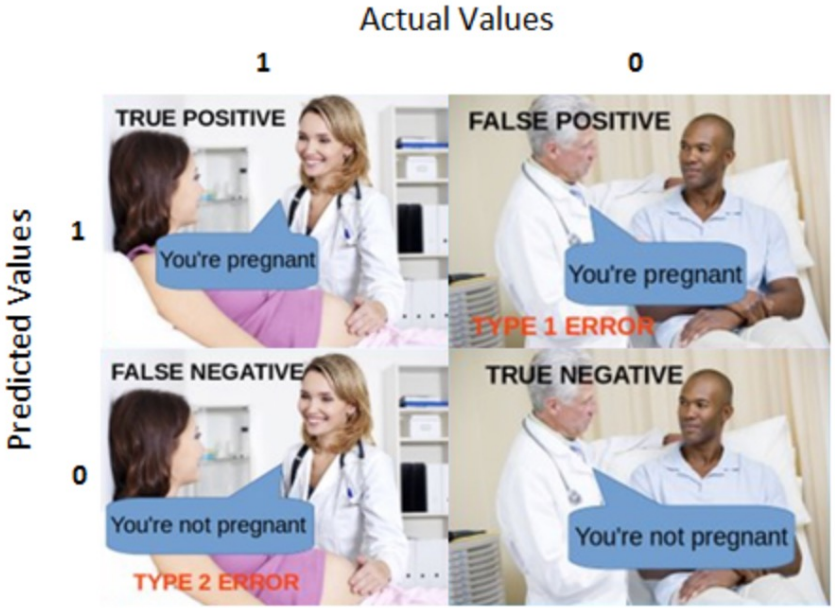
- Accuracy
- Precision and recall
- Squared error
- Likelihood
- Posterior probability
- Cost / Utility
- Margin
- Entropy
- K-L divergence
- etc.



# Confusion Matrix

**Recall & Precision**



Asking the wrong question

Trying to solve the wrong problem

Not having enough data

Not having the right data

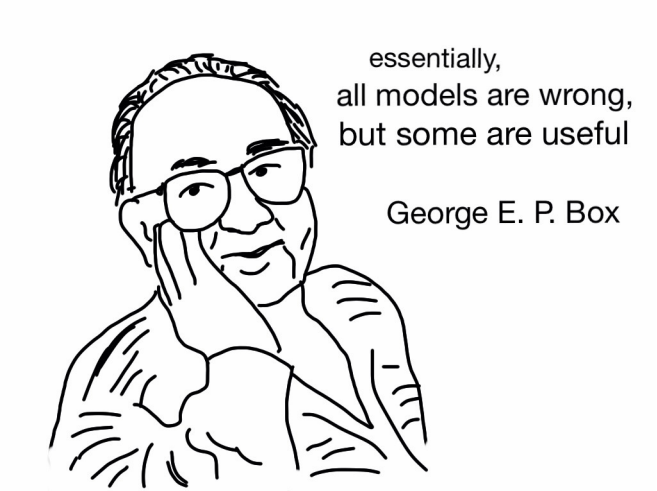
Having too much data

Hiring the wrong people

Using the wrong tools

Not having the right model

Not having the right yardstick

# Can you explain the prediction ?

Predicted: Wolf
True: Wolf

Predicted: Husky
True: Husky

Predicted: Husky
True: Husky

Predicted: Wolf
True: Wolf

Predicted: Wolf
True: Wolf

Predicted: Wolf
True: Wolf

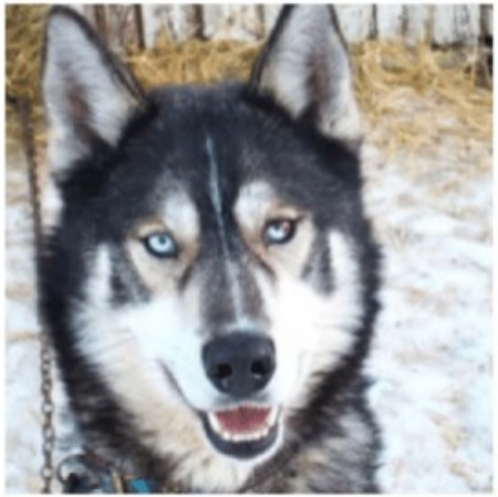
Predicted: Husky
True: Wolf

Predicted: Wolf
True: Wolf

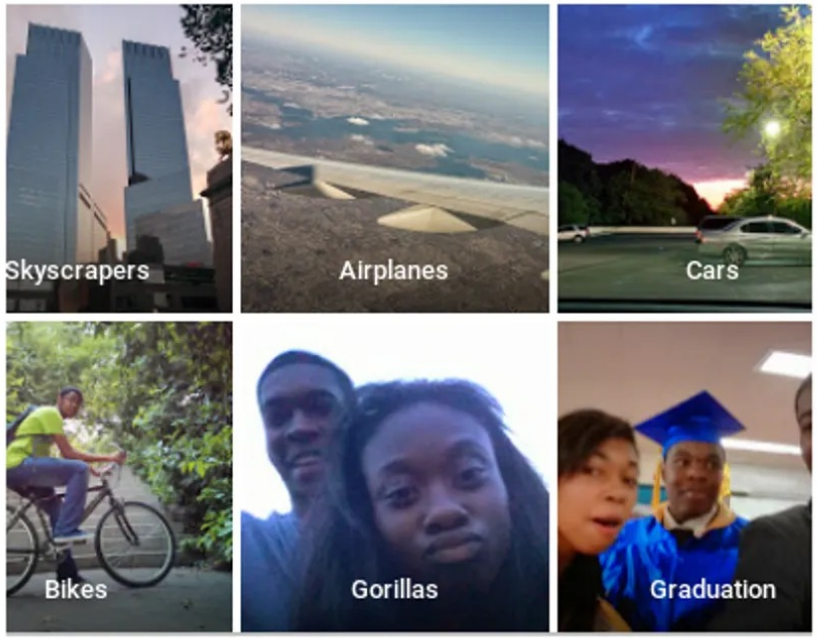
Predicted: Wolf
True: Husky

Predicted: Husky
True: Husky

# Biased Data leads to biased algorithms



# ML Hall of shame




RETWEETS 1,518 FAVORITES 748

8:22 PM - 28 Jun 2015

**Jacky Alciné**
@jackyalcine

Google Photos, y'all fucked up. My friend's not a gorilla.

**Two years later, Google solves 'racist algorithm' problem by purging 'gorilla' label from image classifier**

Cory Doctorow 1 year ago

# Take home messages:

- Effective machine learning is an extension of statistics not an alternative
- Simplest model first
- Training data ≠ Test Data
- ML is more of art than science
- Experienced data analysts are the most valuable tools
- Understanding what is predictable is as important that the prediction
- Your analysis is as good as your data
- All models are wrong, but some are useful !

“Garbage in, garbage out”

# Thank You