INTRODUCTION TO MACHINE LEARNING

Challenges, Trends and Solutions in Life Sciences

Fotis E. Psomopoulos
INAB|CERTH

Van Du T. Tran
Vital-IT, SIB Swiss Institute of Bioinformatics

elixir

SIB Swiss Institute of Bioinformatics

CERTH
CENTRE FOR RESEARCH & TECHNOLOGY HELLAS

INAB
INSTITUTE OF APPLIED BIOSCIENCES
ΙΝΣΤΙΤΟΥΤΟ ΕΦΑΡΜΟΣΜΕΝΩΝ ΒΙΟΕΠΙΣΤΗΜΩΝ
CERTH

SOME HOUSEKEEPING

Please remain muted throughout the course, unless you are invited to speak by the Instructors

Cameras are optional but might lead to bandwidth issues

Please use the “Chat” or the Gdoc to raise questions for further discussions

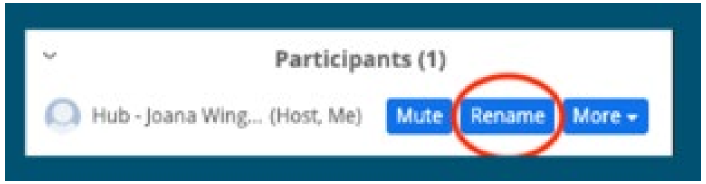
Please use the “hand-raising function” to indicate you would like to contribute directly

This meeting will be run in line with the ELIXIR Code of Conduct. If you have any concerns, please refer to the Code of Conduct, found on the ELIXIR website

If you have any questions during the course, please use the Gdoc document

It would be helpful to include the Country abbreviation after your name using the “rename” function.

Example: Johann Schmidt (DE)



CODE OF CONDUCT

Our values: a place to feel respected, a place to feel safe!

This course falls under the **ELIXIR Hub Code of Conduct** (full document here)

As defined in the ELIXIR Hub Code of Conduct, we encourage the following kinds of behaviours:

- Use welcoming and inclusive language
- Be respectful of different viewpoints and experiences
- Foster scientific and technical rigour and curiosity with constructive and facts-based critique
- Gracefully accept constructive criticism
- Show courtesy and respect towards other participants
- Be mindful of your own biases and do not let them get in the way of respectful interaction
- Speak up if you believe the spirit of the Code has not been upheld. Ideally, where feasible, directly address the issue with the person who committed the transgression
- Adjust the behaviour where it was seen to be short of the requirements indicated in this Code.

A QUICK ROUND OF INTRODUCTIONS

Fotis Psomopoulos

- Researcher (Assistant Professor level)
- Institute of Applied Biosciences
  Centre for Research and Technology Hellas

Van Du T. Tran

- Senior Computational Biologist
- Vital-IT, SIB Swiss Institute of Bioinformatics

COMMUNICATION

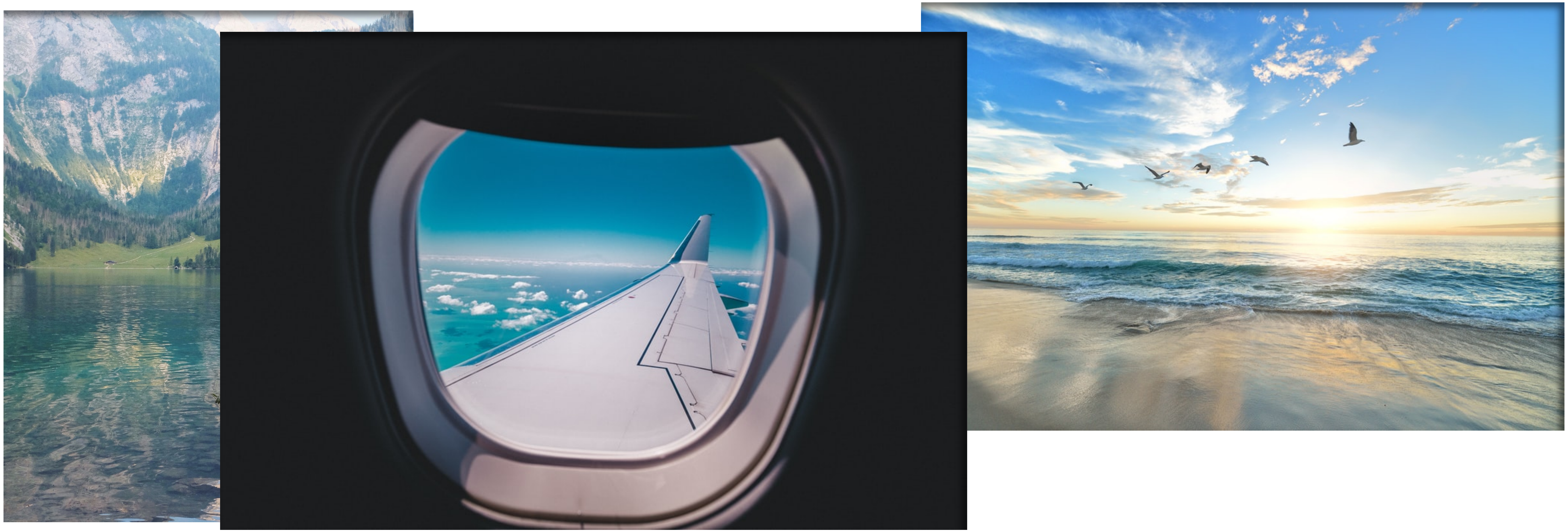
We will be using GDoc to exchange information:

https://tinyurl.com/elixir-fr-ml-course-2021

Access the Gdoc and sign in! ☺

AN ICEBREAKER

“If I could be on vacation anywhere right now (pandemic-free 😷), I’d go to…, because…”



COURSE AGENDA

09:00 - 09:30	**Day 1: Course Introduction.** - Welcome. - Introduction and CoC. - Way to interact - Practicalities (agenda, breaks, etc). - Setup
09:30 - 10:00	**Introduction to Machine Learning** (*theory*)
10:00 - 11:30	**What is Exploratory Data Analysis (EDA) and why is it useful?** (*hands-on*) - Loading omics data - PCA
11:30 - 11:45	*Coffee Break*
11:45 - 12:15	**Introduction to Unsupervised Learning** (*theory*)
12:15 - 13:00	**Agglomerative Clustering: k-means** (*practical*)
13:00 - 14:00	*Lunch break*
14:00 - 14:45	**Agglomerative Clustering: k-means** (*practical*) (cont’d)
14:45 - 15:30	**Divisive Clustering: hierarchical clustering** (*practical*)
15:30 - 15:45	*Coffee Break*
15:45 - 16:30	**Divisive Clustering: hierarchical clustering** (*practical*) (cont’d)
16:30	*Closing of Day 1*

09:00 - 09:30	**Welcome Day 2.** - Questions from Day 1 - Agenda
09:30 - 10:00	**Introduction to Supervised Learning** (*theory*) - Overview of multiple algorithms - Advantages and Disadvantages
10:00 - 10:30	**Classification Metrics** (*theory*) - F1 Score, Precision, Recall - Confusion Matrix, ROC-AUC
10:30 - 11:30	**Classification** (*practical*) - Decision trees - Random Forests
11:30 - 11:45	*Coffee Break*
11:45 - 12:30	**Classification** (*practical*) (cont’d)
12:30 - 13:30	*Lunch break*
13:30 - 14:00	**Regression** (*theory*)
14:00 - 15:15	**Regression** (*practical*) - Linear regression - Generalized Linear Model (GLM)
15:15 - 15:30	*Coffee Break*
15:30 - 16:00	**Regression** (*practical*) (cont’d)
16:00 - 16:30	*Closing questions, Discussion*

SESSION OVERVIEW

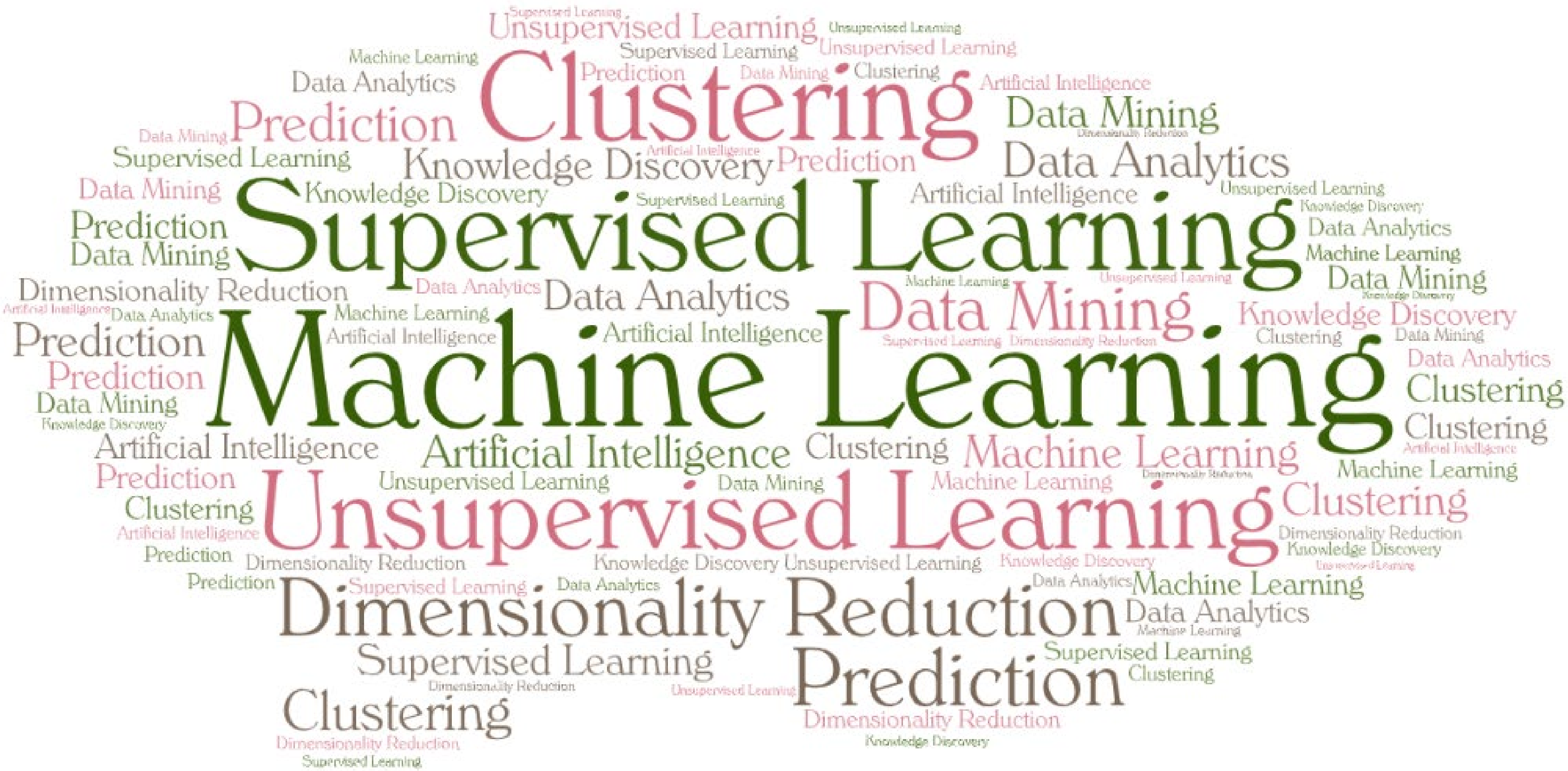
01
- Introduction to basic concepts of Data mining and Machine learning

02
- Machine learning taxonomy

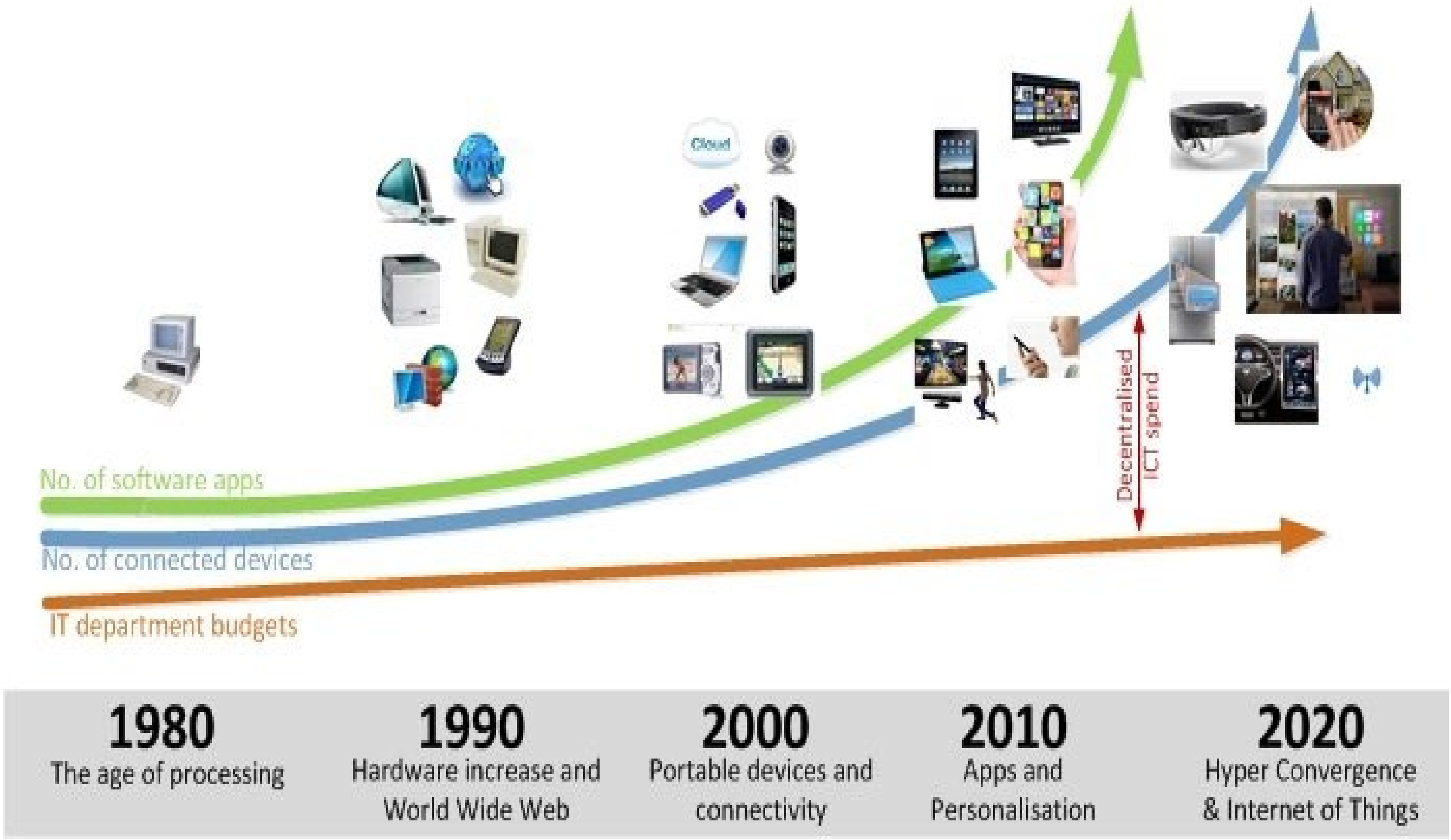
03
- Supervised classification vs unsupervised classification

04
- Algorithms examples

05
- Examples of applications in Bioinformatics

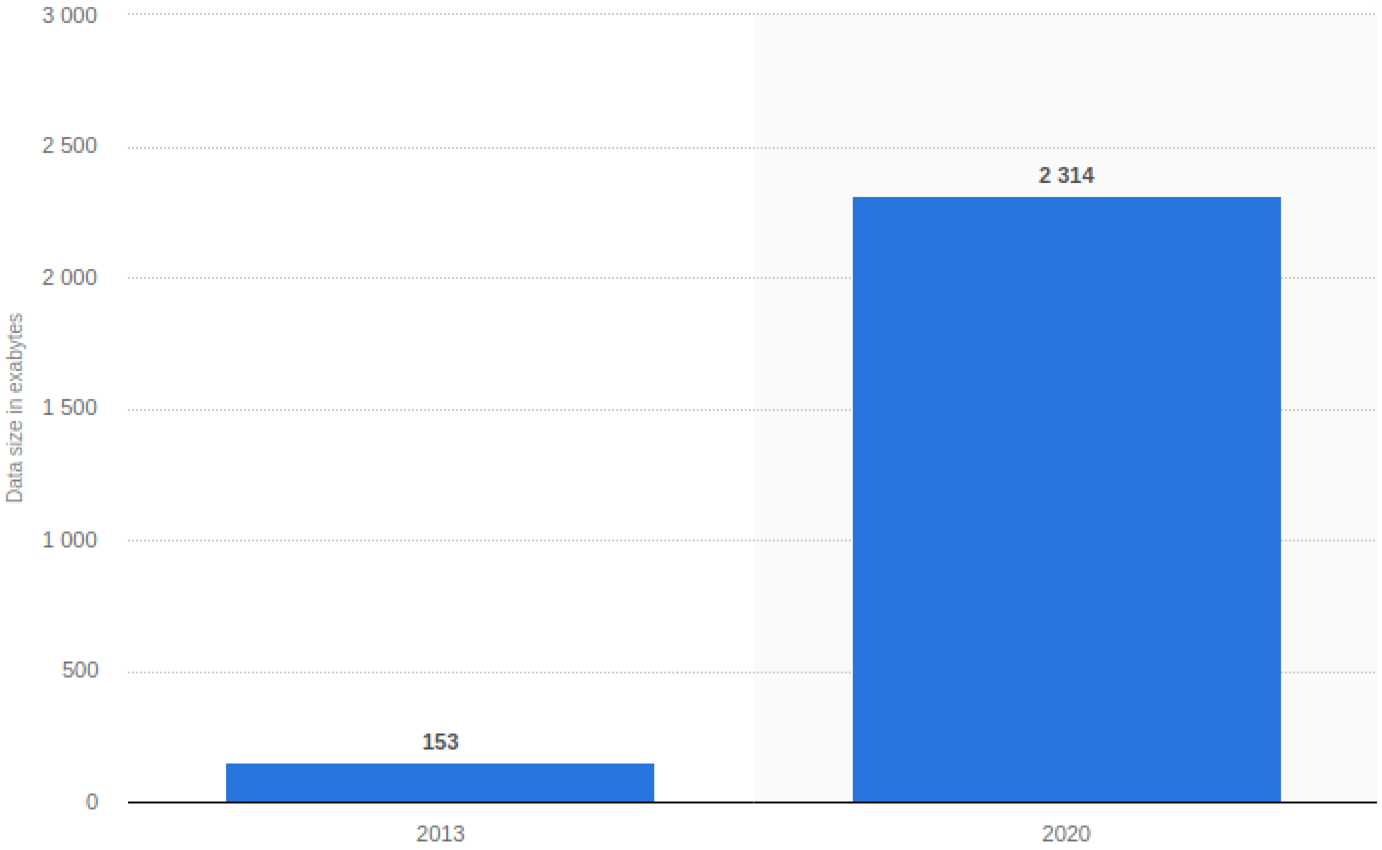

Unsupervised Learning
Supervised Learning
Machine Learning
Data Analytics
Clustering
Prediction
Data Mining
Artificial Intelligence
Data Mining
Supervised Learning
Knowledge Discovery
Prediction
Data Analytics
Knowledge Discovery
Supervised Learning
Artificial Intelligence
Prediction
Data Mining
Data Analytics
Machine Learning
Dimensionality Reduction
Data Analytics
Data Mining
Machine Learning
Artificial Intelligence
Knowledge Discovery
Prediction
Machine Learning
Clustering
Data Analytics
Prediction
Clustering
Data Mining
Clustering
Artificial Intelligence
Artificial Intelligence
Machine Learning
Machine Learning
Prediction
Unsupervised Learning
Data Mining
Machine Learning
Clustering
Unsupervised Learning
Dimensionality Reduction
Knowledge Discovery
Prediction
Dimensionality Reduction
Knowledge Discovery
Unsupervised Learning
Knowledge Discovery
Supervised Learning
Data Analytics
Machine Learning
Dimensionality Reduction
Data Analytics
Supervised Learning
Prediction
Clustering
Clustering
Dimensionality Reduction
Knowledge Discovery
Supervised Learning

Technology Timeline



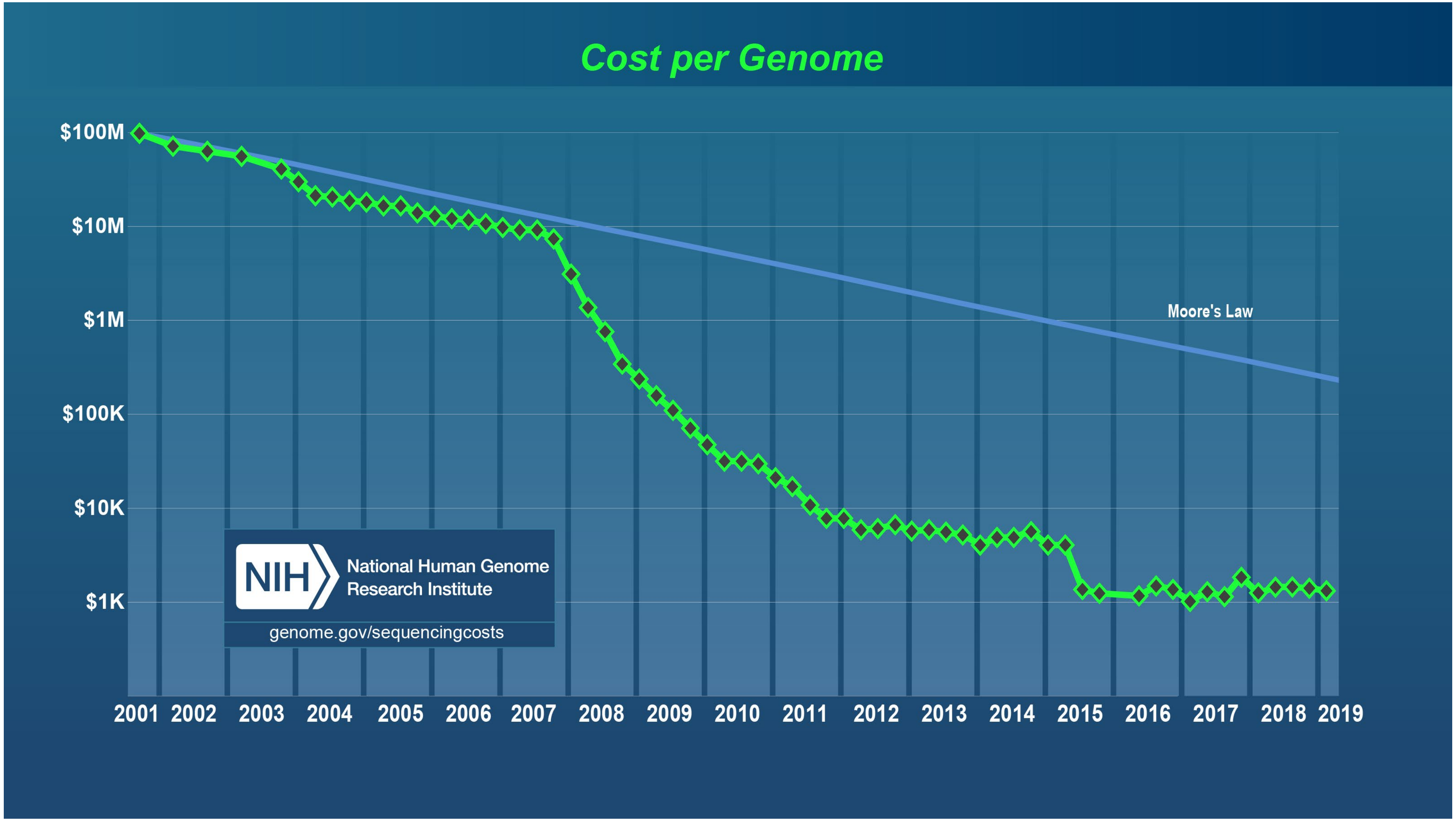
https://www.linkedin.com/pulse/technology-increase-vs-department-budgets-sam-errington/

TOTAL AMOUNT OF GLOBAL HEALTHCARE DATA GENERATED AND PROJECTIONS FOR END 2020 (IN EXABYTES)



1 exabyte = 1 billion gigabytes

Source: https://www.statista.com/statistics/1037970/global-healthcare-data-volume/



https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

FROM DATA TO KNOWLEDGE

AI & ML

AI is a broader concept than ML which addresses the use of computers to mimic the cognitive functions of humans.

When machines carry out tasks based on algorithms in an intelligent manner, that is AI

ML is a subset of AI and focuses on the ability of machines to receive a set of data and learn from it, improve algorithms as they learn more about information being processed

ML & DATA MINING

ML embodies the principles of DM

DM and ML have the same foundation but in different ways

- DM requires human interaction
- DM can’t see the relationship between different data aspects with the same depth as ML
- ML learns from the data and allows the machine to teach itself

DM is typically used as an information source for ML to pull from

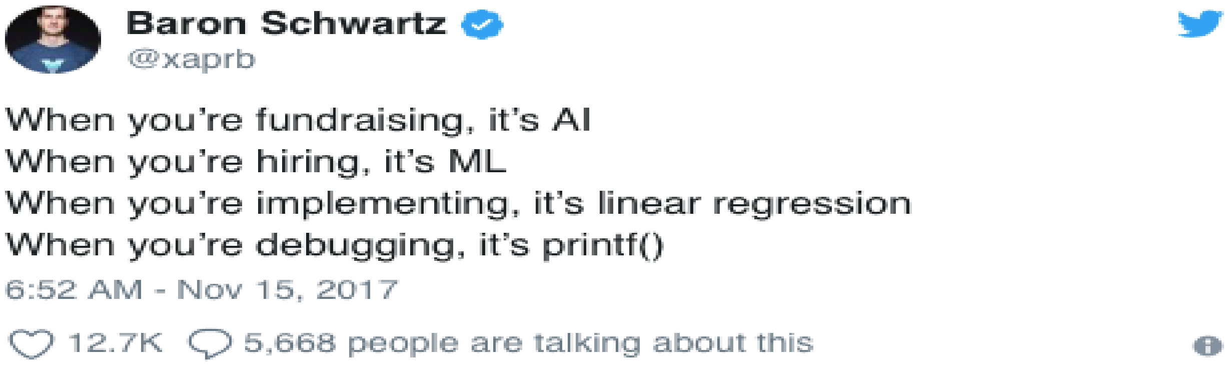
ML is more about building the prediction model

AI, ML & DM

Data mining produces insights

ML produces predictions

AI produces actions



Baron Schwartz
@xaprb

When you're fundraising, it's AI
When you're hiring, it's ML
When you're implementing, it's linear regression
When you're debugging, it's printf()

6:52 AM - Nov 15, 2017

12.7K 5,668 people are talking about this

https://medium.freecodecamp.org/using-machine-learning-to-predict-the-quality-of-wines-9e2e13d7480d

DEEP LEARNING

Deep learning is a subset of ML

Deep learning algorithms go a level deeper than classical ML involving many layers

Layers: set of nested hierarchy of related concepts

The answer to a question is obtained by answering other related deeper questions

DATA IS AT THE HEART OF ML

Machine learning algorithms are driven by the data used

Data quality is very important!

Identifying incomplete, incorrect and irrelevant parts of the data is an important step

Preprocessing data before applying ML is crucial step

HOW DO WE HUMAN MAKE DECISIONS? DO WE ALL MAKE THE SAME DECISIONS?

Observations

Experiences

External information

Beliefs, creativity, common sense

Compare to expectations

Analyze differences

Creativity, Limited memory

HOW DOES A COMPUTER WORK?

Follow instructions given by human

ARTIFICIAL INTELLIGENCE

Stimulate human behavior and cognitive process

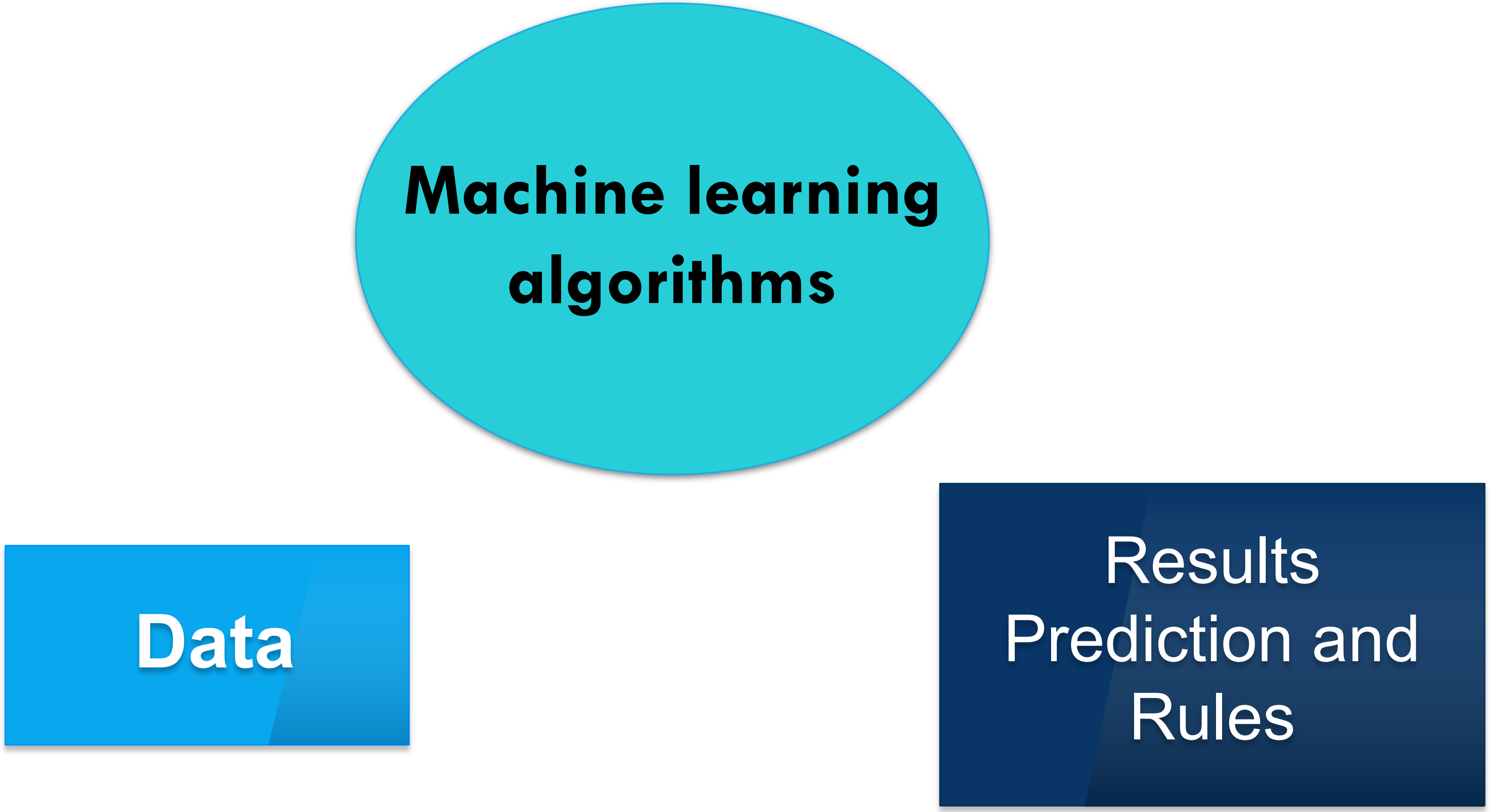
Capture and preserve human expertise

Fast response
Ability to memorize big amounts of data

Data

Computing
+
Storage

ARTIFICIAL INTELLIGENCE



Machine learning algorithms

Data

Results Prediction and Rules

HOW DO MACHINES LEARN?

Data to model

Decision

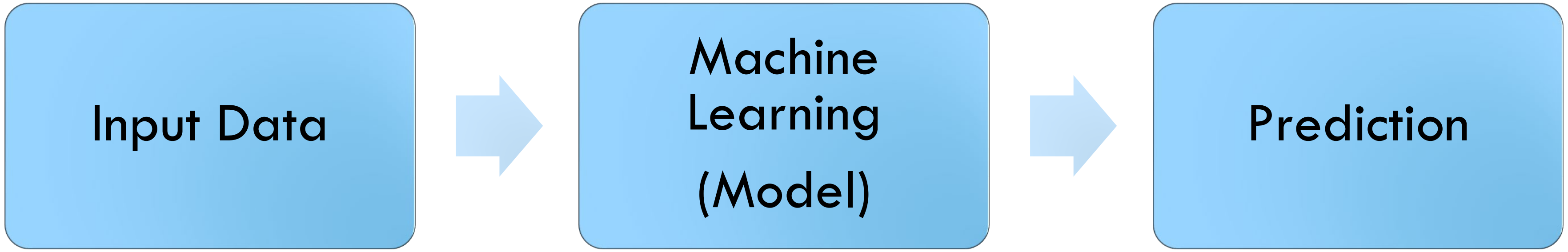
Create models

Evaluate models

Refine models

Prediction,
categorization

WHAT IS MACHINE LEARNING?



Learning begins with observations or data

- Examples: direct experience, or instruction

The system looks for patterns in data and makes better decisions in the future based on the examples that we provide

The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly.

MACHINE LEARNING AND GENOMICS

In the context of genome annotation, a machine learning system can be used to:

- ‘learn’ how to recognize the locations of transcription start sites (TSSs) in a genome sequence
- identify splice sites and promoters

In general, if one can compile a list of sequence elements of a given type, then a machine learning method can probably be trained to recognize those elements.

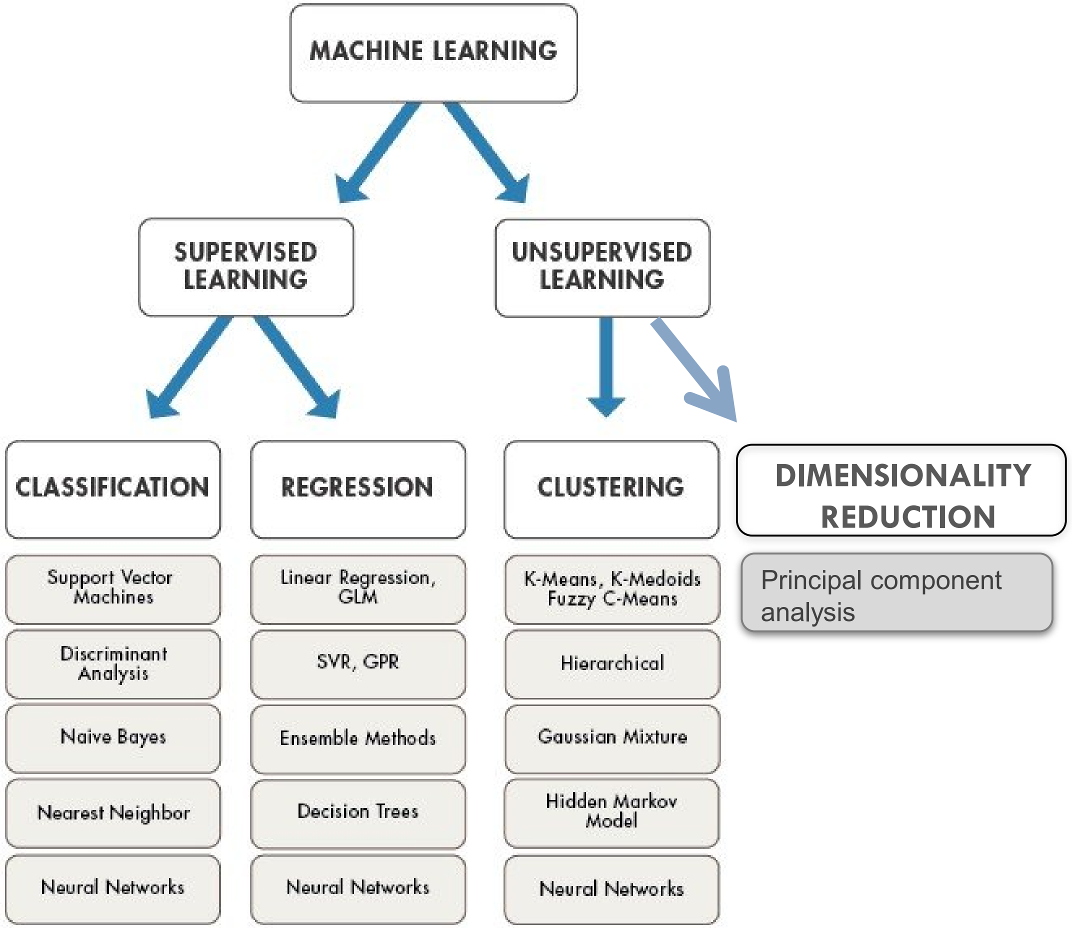
More info about this task can be obtained from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5204302/)

MACHINE LEARNING CONCEPTS

Any machine learning problem can be represented with the following three concepts:

- We will have to learn to solve a task T.
  - For example, perform genome annotation.
- We will need some experience E to learn to perform the task. Usually, experience is represented through a dataset.
  - For the gene prediction, experience comes as a set of DNA sequences provided as input to a learning procedure, along with binary labels indicating whether each sequence is centered on a TSS or not. The learning algorithm produces a model which can then be subsequently used, in conjunction with a prediction algorithm, to assign predicted labels to unlabeled test sequences.
- We will need a measure of performance P to know how well we are solving the task and also to know whether after doing some modifications, our results are improving or getting worse.
  - The percentage of genes that our gene prediction model is correctly classifying as genes could be P for our gene prediction task.

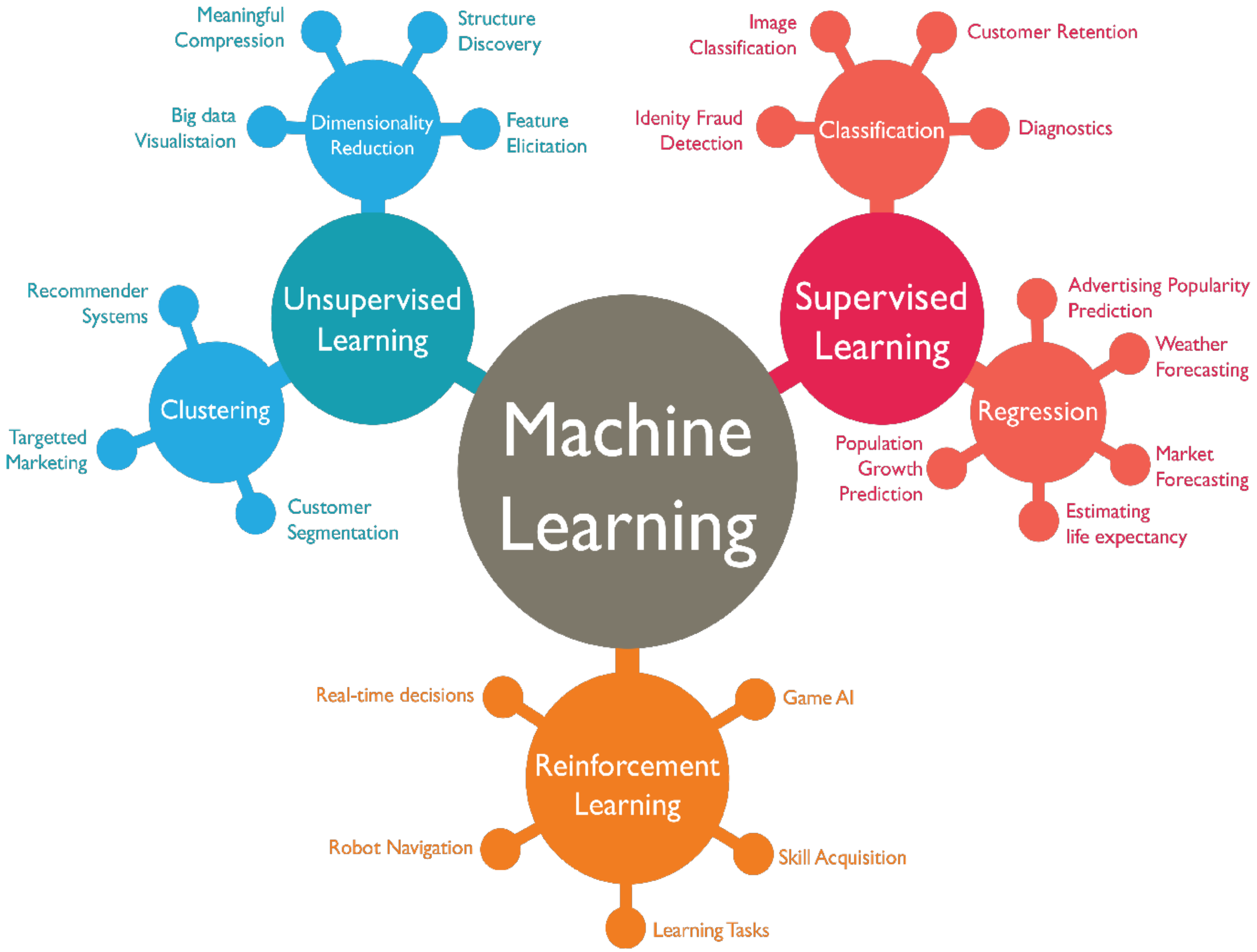
THE ML TAXONOMY



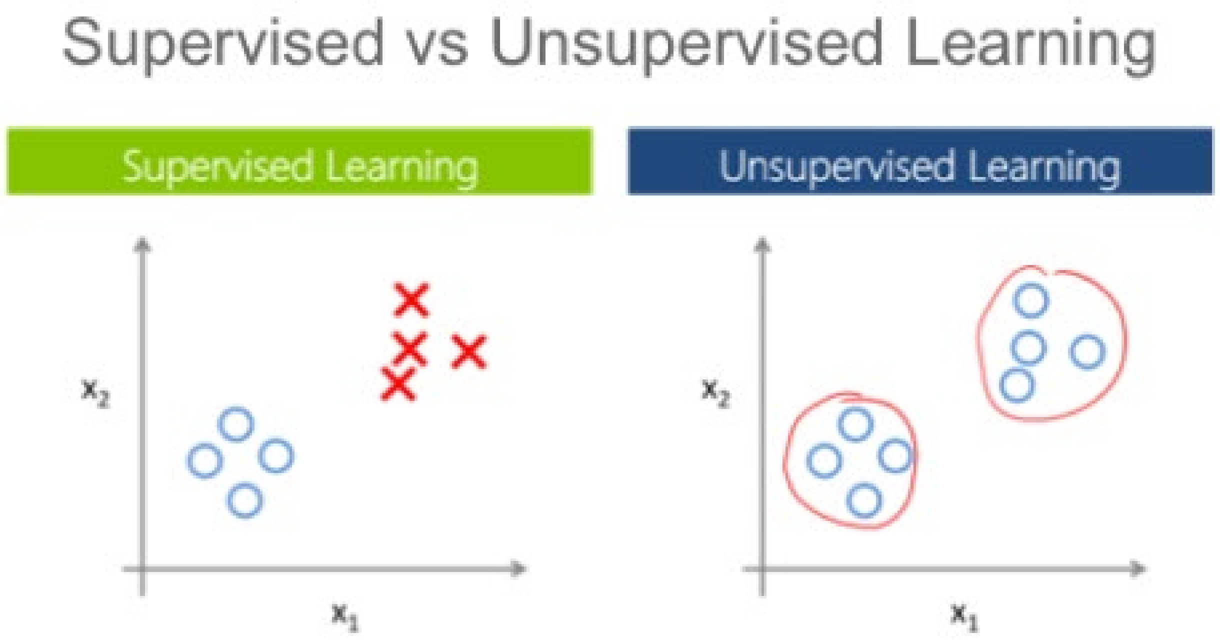
THE ML TAXONOMY

Machine learning algorithms are often categorized as **supervised** or **unsupervised**.

We also have **semi-supervised** machine learning and **reinforcement** machine learning.

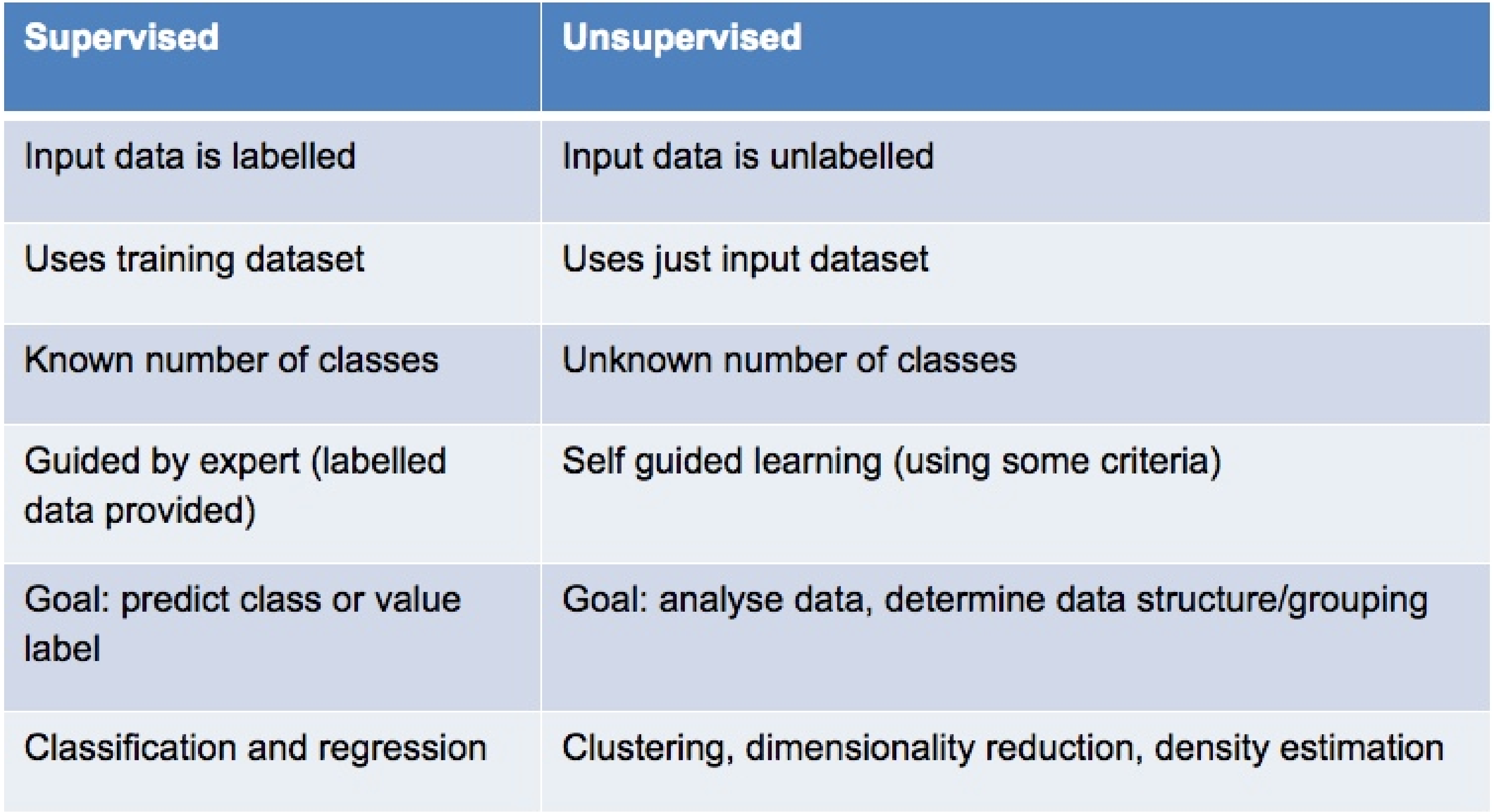

Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

SUPERVISED VS UNSUPERVISED LEARNING



https://www.cisco.com/c/m/en_us/network-intelligence/service-provider/digital-transformation/get-to-know-machine-learning.html

SUPERVISED VS UNSUPERVISED



Supervised	Unsupervised
Input data is labelled	Input data is unlabelled
Uses training dataset	Uses just input dataset
Known number of classes	Unknown number of classes
Guided by expert (labelled data provided)	Self guided learning (using some criteria)
Goal: predict class or value label	Goal: analyse data, determine data structure/grouping
Classification and regression	Clustering, dimensionality reduction, density estimation

Unsupervised Learning

UNSUPERVISED MACHINE LEARNING ALGORITHMS[1]

- Are applied when given data are ***neither classified nor labeled***.
- No desired output, rather find difference among data
- Draw inferences to describe hidden structures from unlabeled data.

Goal: model the underlying structure of or distribution in the data, group data according to similarities, represent data in a compressed format

- Algorithms are left to their own devising to discover and present the interesting structure in the input data.

UNSUPERVISED MACHINE LEARNING ALGORITHMS[2]

Clustering: discover the inherent groupings in the data

- e.g. clustering DNA sequences into functional groups.

Association: discover rules that describe large portions of your data

- e.g. association analysis-based techniques for pre-processing protein interaction networks for the task of protein function prediction.

Dimensionality reduction: reduce the variable space of high dimensionality before the subsequent analysis is carried out.

- e.g. in a gene-expression analysis, for finding a list of candidate genes with a more operable length ideally including all the relevant genes.

EXAMPLES OF UNSUPERVISED LEARNING ALGORITHMS

PRINCIPAL COMPONENT ANALYSIS (PCA)

PCA can be applied for dimensionality reduction.

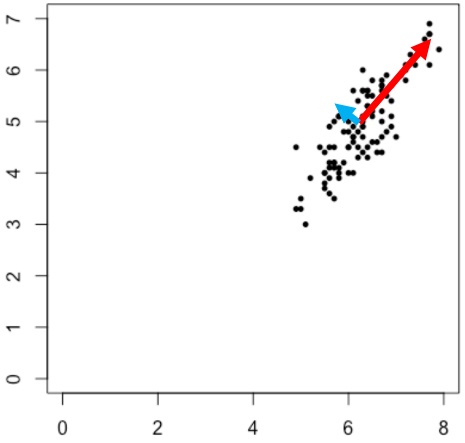
- Data with a wide range of features (e.g. omics), probably highly correlated between each other => overfitting models

Algorithm: find linear combinations of variables having maximum variances

- Standardization
- Covariance matrix computation
- Compute eigenvalues (amounts of variance) and eigenvectors (PCs) of the covariance matrix

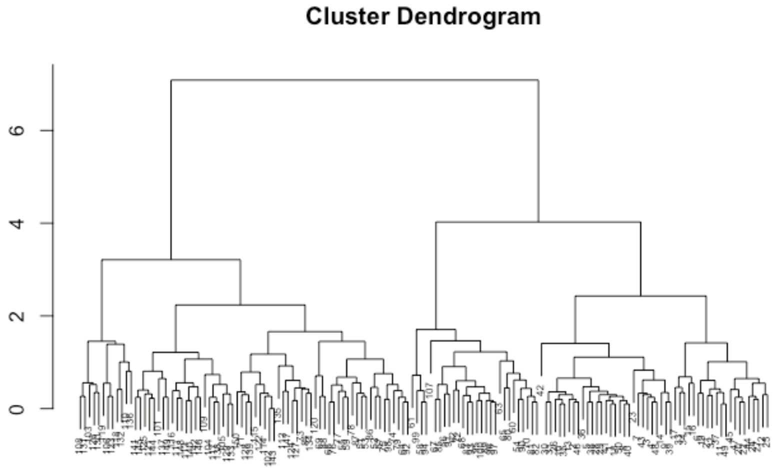
Advantage:

- low-dimensional sample representation
- synchronized low-dimensional representation of the variables
- visually find variables that are characteristic of a group of samples.



HIERARCHICAL CLUSTERING

Group similar objects together into *clusters* (unknown number of clusters *a priori*):
Bottom-up (Agglomerative) & Top-down (Divisive)



Agglomerative

- Algorithm:

  It starts by treating each observation as a separate cluster.

  Then, it repeatedly executes the following two steps:

  (1) identify the two clusters that are closest together, and

  (2) merge the two most similar clusters.

  This iterative process continues until all the clusters are merged together.

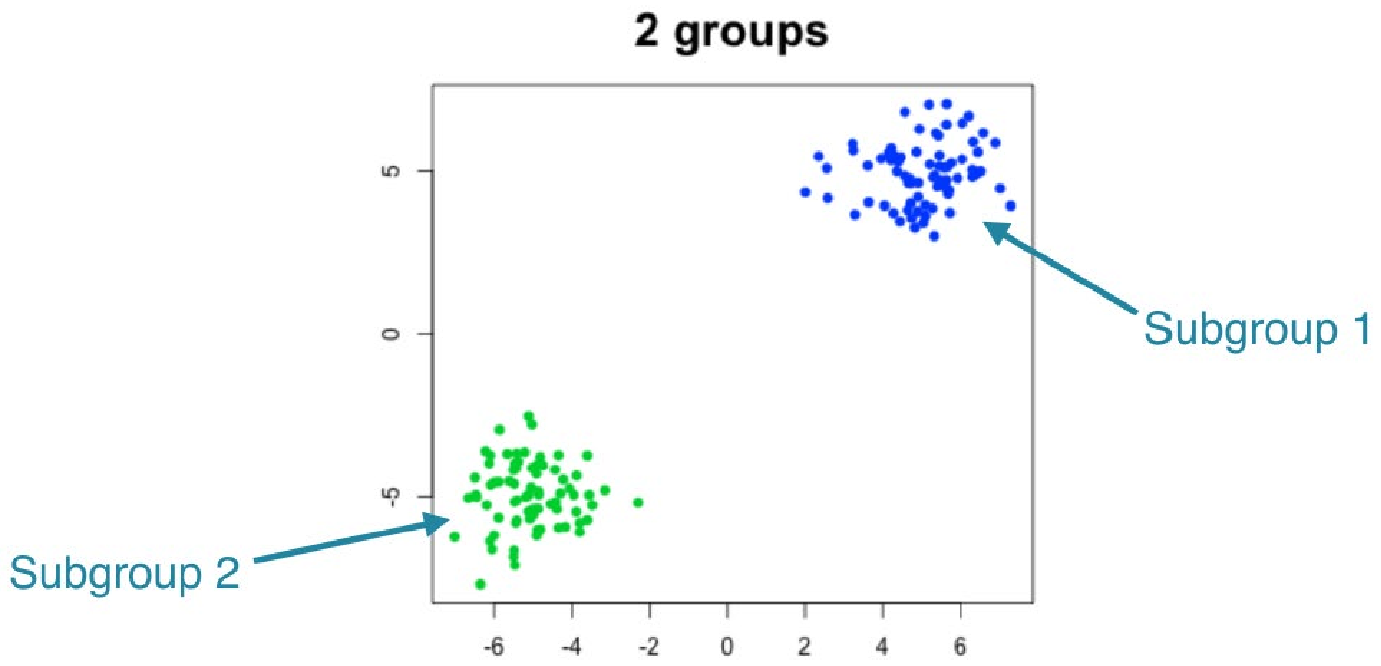
- The main output of Hierarchical Clustering is a dendrogram, which shows the hierarchical relationship between the clusters

E.g. gene expression data analysis - genes with similar expression patterns are grouped together and are connected by a series of branches.

K-MEANS[1]

Find homogeneous subgroups in a population

- Break observations into a pre-defined number of clusters



K-MEANS[2]

Algorithm

1. Divide the data into K clusters

   Initialize the centroids with the mean of the clusters

2. Assign each item to the cluster with closest centroid

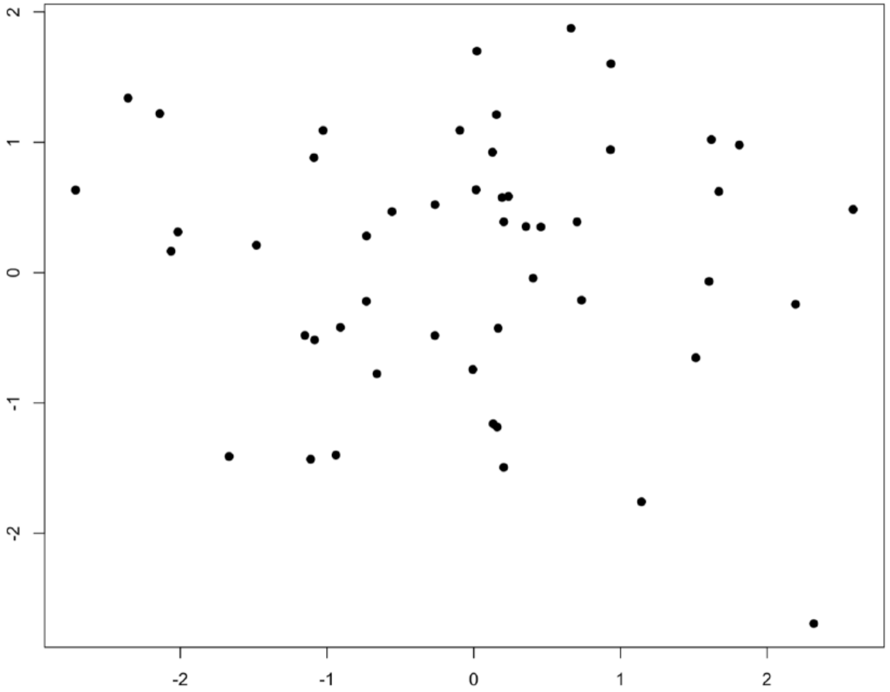
3. When all objects have been assigned, recalculate the centroids (mean)

4. Repeat 2-3 until the centroids no longer move

K-MEANS[3]

The Algorithm in action

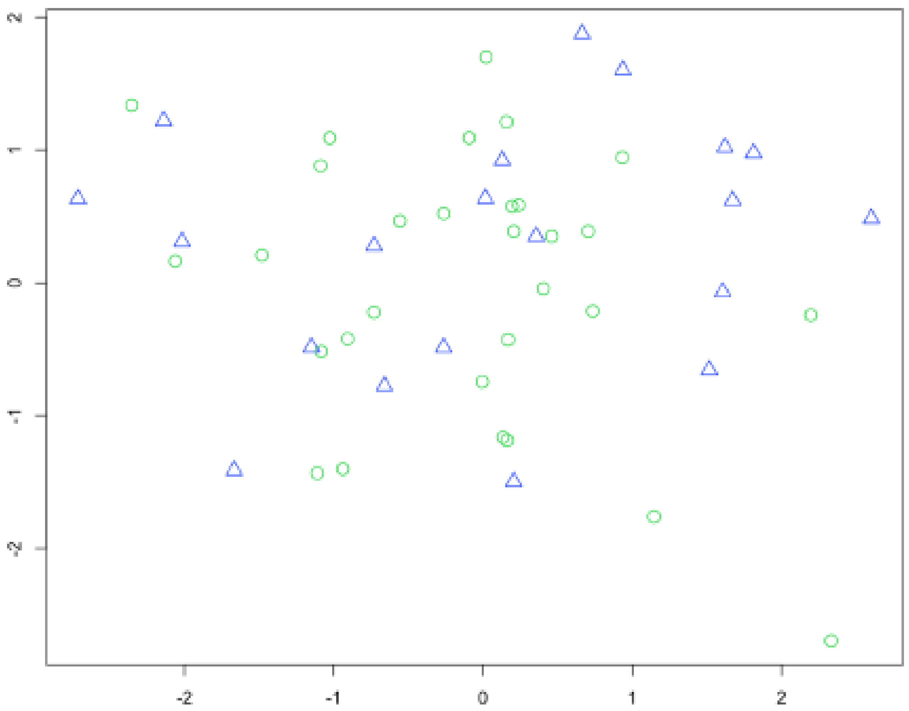
Observations



K-MEANS[4]

The Algorithm in action

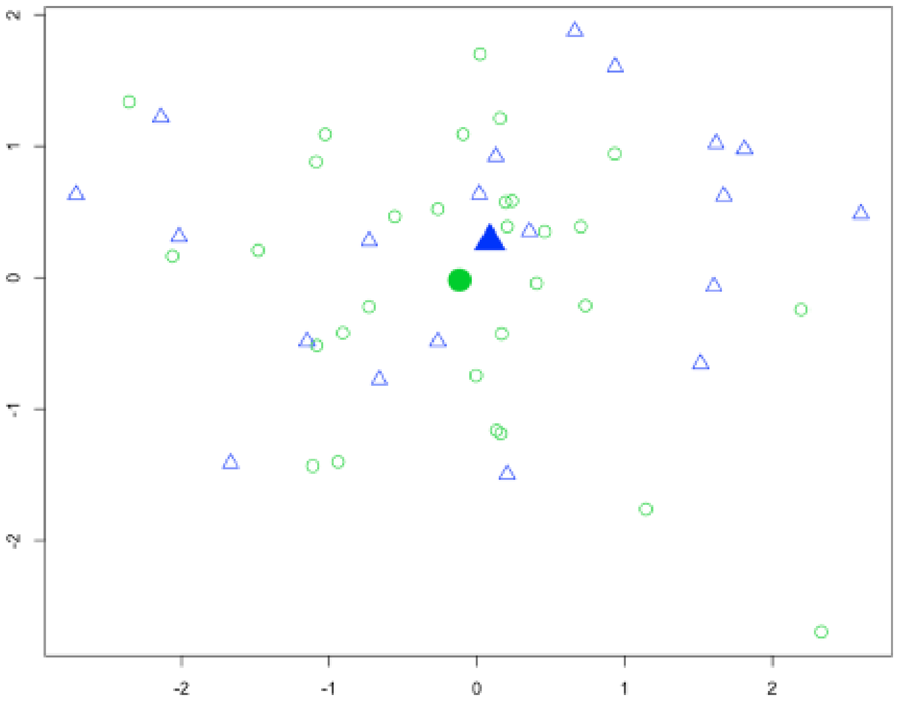
Random Cluster Assignment



K-MEANS[5]

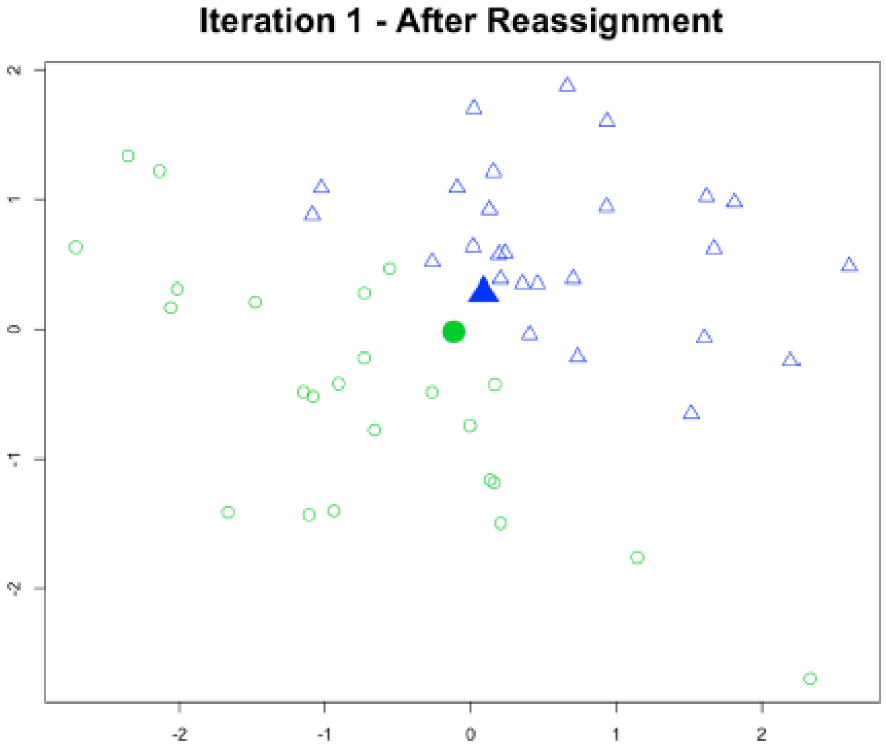
The Algorithm in action

Cluster Centers Calculated



K-MEANS[6]

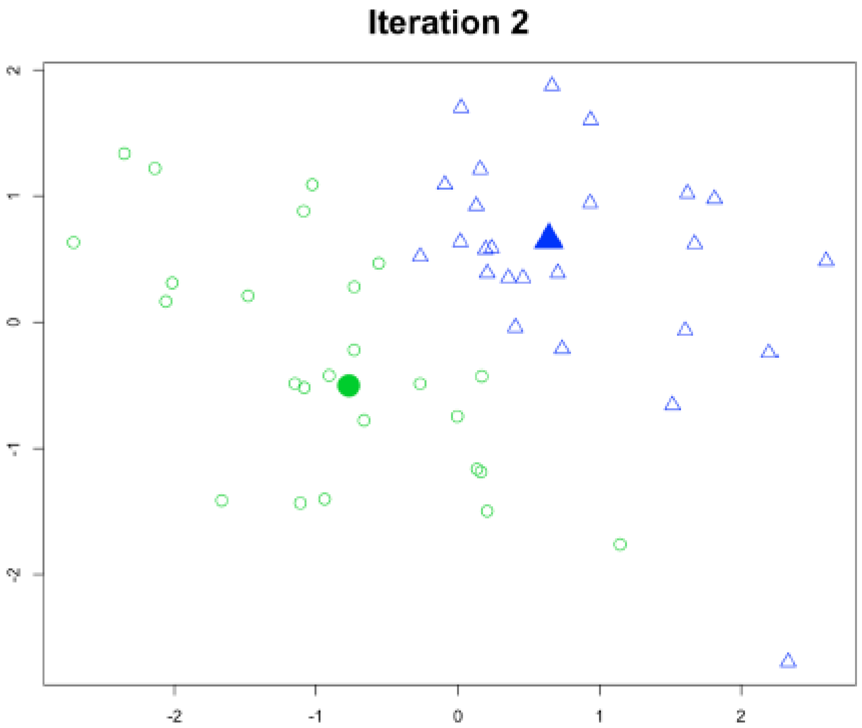
The Algorithm in action



16-17 June 2021

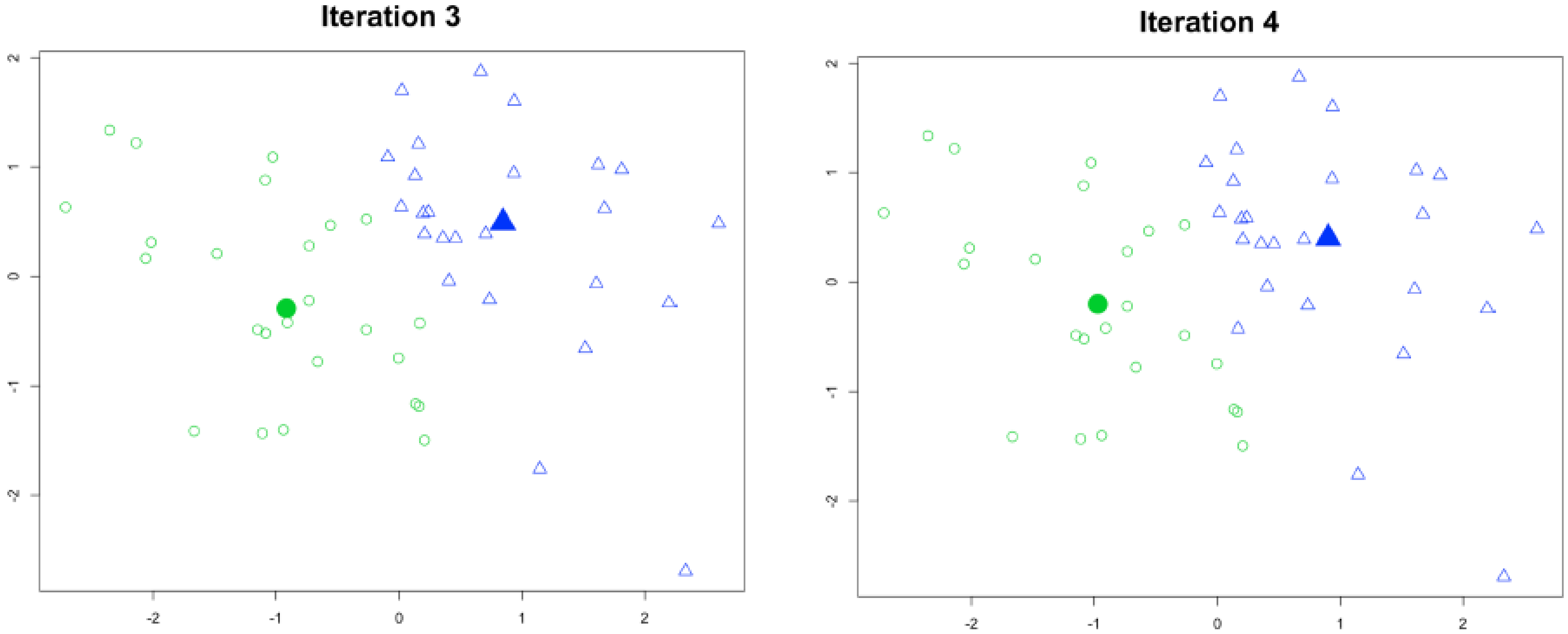
K-MEANS[7]

The Algorithm in action



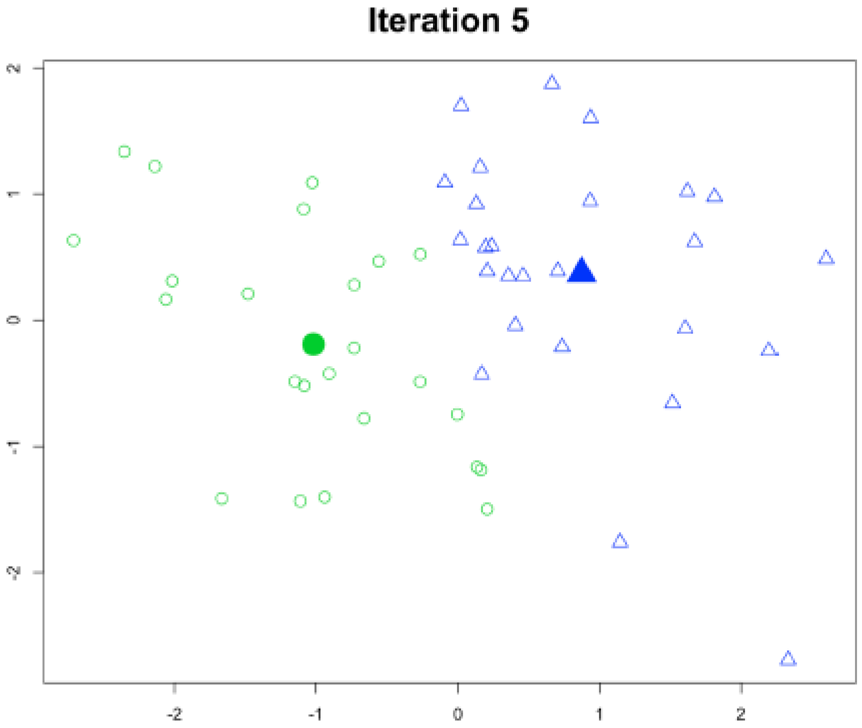
K-MEANS[8]

The Algorithm in action



K-MEANS[9]

The Algorithm in action



Remember k-means has a random component!

K-MEANS[10]

Goal: find groups in the data, with a pre-defined number of groups *K*.

The algorithm works iteratively to assign each data point to one of *K* groups based on the features that are provided. Data points are clustered based on feature similarity.

Advantage: Easy to implement and fast and efficient in terms of computational cost

Disadvantage include:

- Initial seeds have a strong impact on the final results
- The order of the data has an impact on the final results
- K-Means needs to know in advance how many clusters there will be in your data, so this may require a lot of trials to "guess" the best K number of clusters to define.

E.g. clustering COVID regions, virus sequences, lockdown measures

Supervised Learning

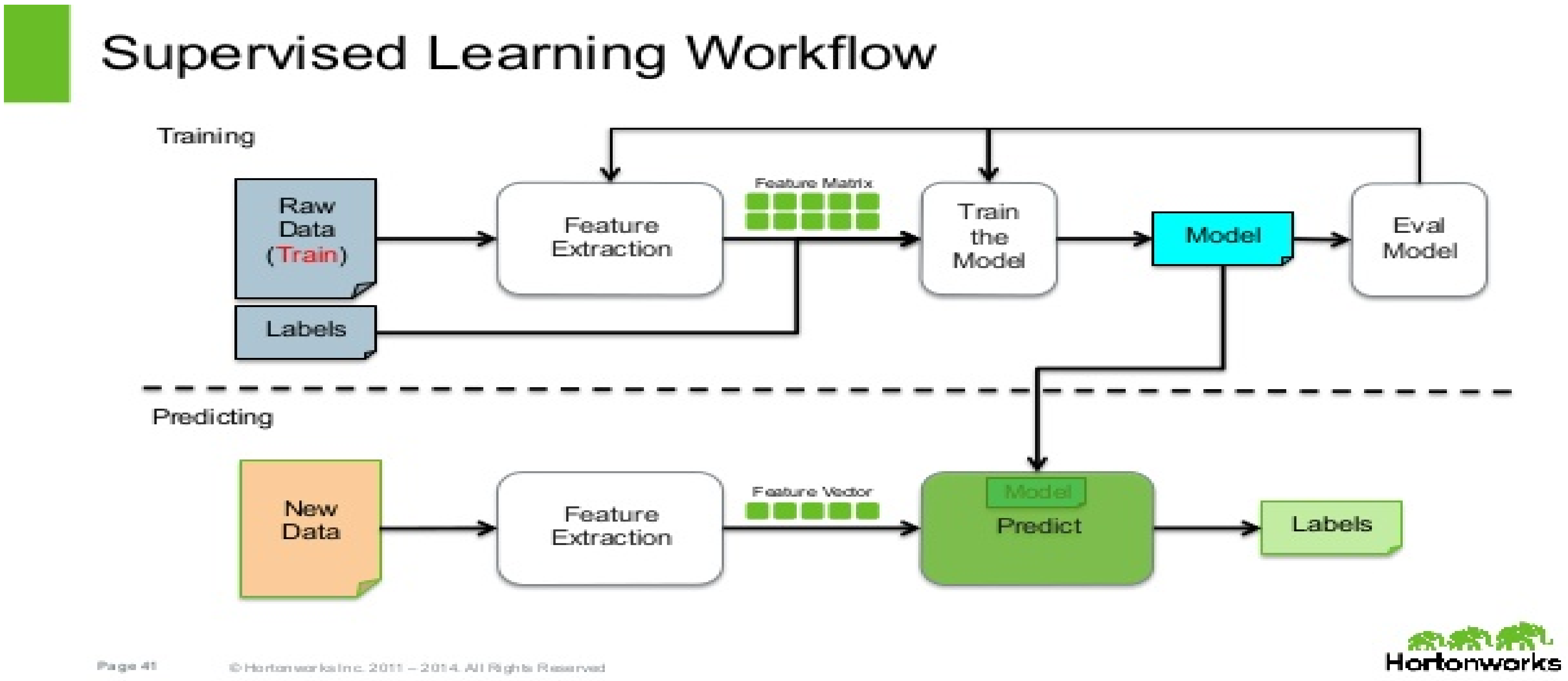
SUPERVISED MACHINE LEARNING ALGORITHMS[1]

- Are applied when given data are ***classified or labeled.***
- Train models with labelled data then predict the output (known output values)
- Learn the mapping function from the input ***x*** to the output ***y***: ***y = h(x)***

Goal: approximate the mapping function so well that it can be used to predict the output ***y*** of new input data ***x***

- Algorithms learn to make predictions on the training data, while supervised by labels
- Learning stops when achieving an acceptable level of performance

SUPERVISED MACHINE LEARNING ALGORITHMS[2]



Supervised Learning Workflow

Training

Raw Data (Train)

Labels

Feature Extraction

Feature Matrix

Train the Model

Model

Eval Model

Predicting

New Data

Feature Extraction

Feature Vector

Model

Predict

Labels

Page 41

© Hortonworks Inc. 2011 – 2014. All Rights Reserved

Hortonworks

SUPERVISED MACHINE LEARNING ALGORITHMS[3]

Let's assume our simple predictor has this form: $h(x) = \vartheta_0 + \vartheta_1 x$

- Goal: find the values of ϑ_0 and ϑ_1 to make our predictor work as well as possible.

Optimizing the predictor ***h(x)*** is done using training examples.

- For each training example, we have an input value ***x_train***, for which a corresponding output, ***y***, is known in advance.
- For each example, we find the difference between the known, correct value ***y***, and our predicted value ***h(x_train)***.
- With enough training examples, these differences give us a useful way to measure the "wrongness" of ***h(x)***.
- We can then tweak ***h(x)*** by tweaking the values of ϑ_0 and ϑ_1 to make it "less wrong".
- This process is repeated over and over until the system has converged on the best values for ϑ_0 and ϑ_1
- In this way, the predictor becomes trained, and is ready to do some real-world predicting.

SUPERVISED MACHINE LEARNING ALGORITHMS[4]

Classification: identify category of new observations on the basis of training data

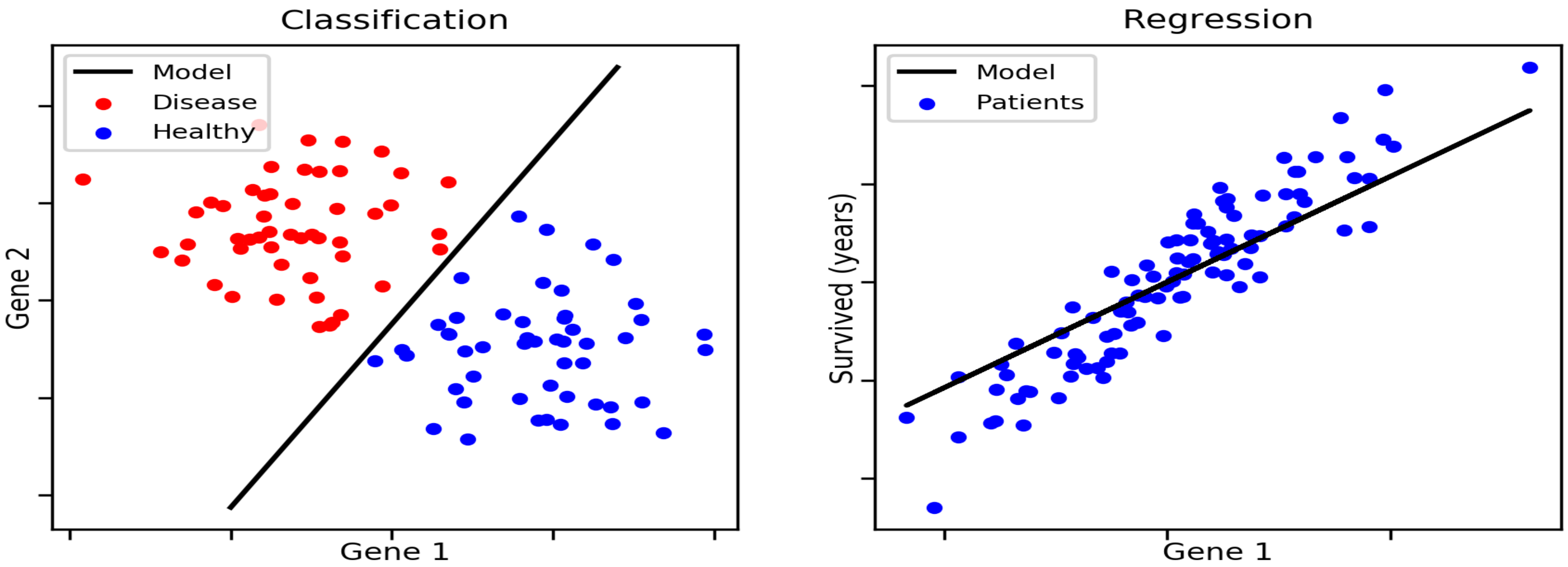
- e.g. binary classifier: is this tumor cancerous?, is this email a spam?; multi-class classifier: classification of types of music, virus variants

Regression: model the relationship between a dependent (target) and independent (predictor) variables

- e.g. salary of employees ~ year of experience, gene expression ~ genetic variants (eQTL)

SUPERVISED MACHINE LEARNING ALGORITHMS [5]

CLASSIFICATION VS REGRESSION



https://aldro61.github.io/microbiome-summer-school-2017/sections/basics/

EXAMPLES OF SUPERVISED LEARNING ALGORITHMS

DECISION TREES (SUPERVISED)

A decision tree is a tree-like graph with

- Nodes: places for an attribute
- Edges: rules
- Leaves: actual outputs or class labels

Single trees are used very rarely, but in composition with many others they build very efficient algorithms such as Random Forest or Gradient Tree Boosting.

Used for both classification and regression tasks.

DECISION TREES (SUPERVISED)

Advantages:

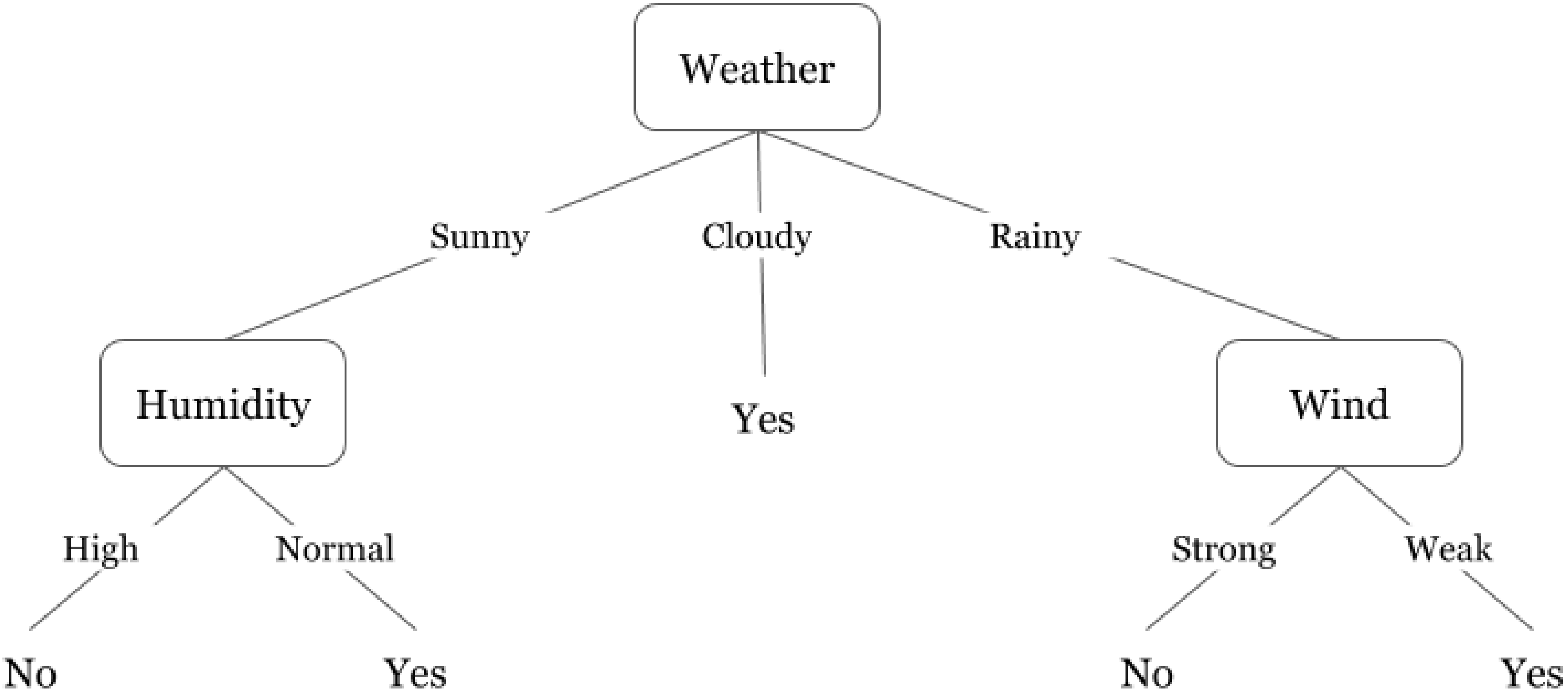
- Simple linear decision surface for non-linear decision making
- Easily handle feature interactions
- Non-parametric
- Dealing with outliers
- Solve **both regression and classification** problems

Disadvantages:

- Often the tree needs to be rebuilt when new examples come on.
- Easily overfit, but ensemble methods like random forests (or boosted trees) take care of this problem.
- Take a lot of memory (the more features you have, the deeper and larger your decision tree is likely to be)

E.g. Classification of genomic islands using decision trees and ensemble algorithms

DECISION TREES (SUPERVISED)



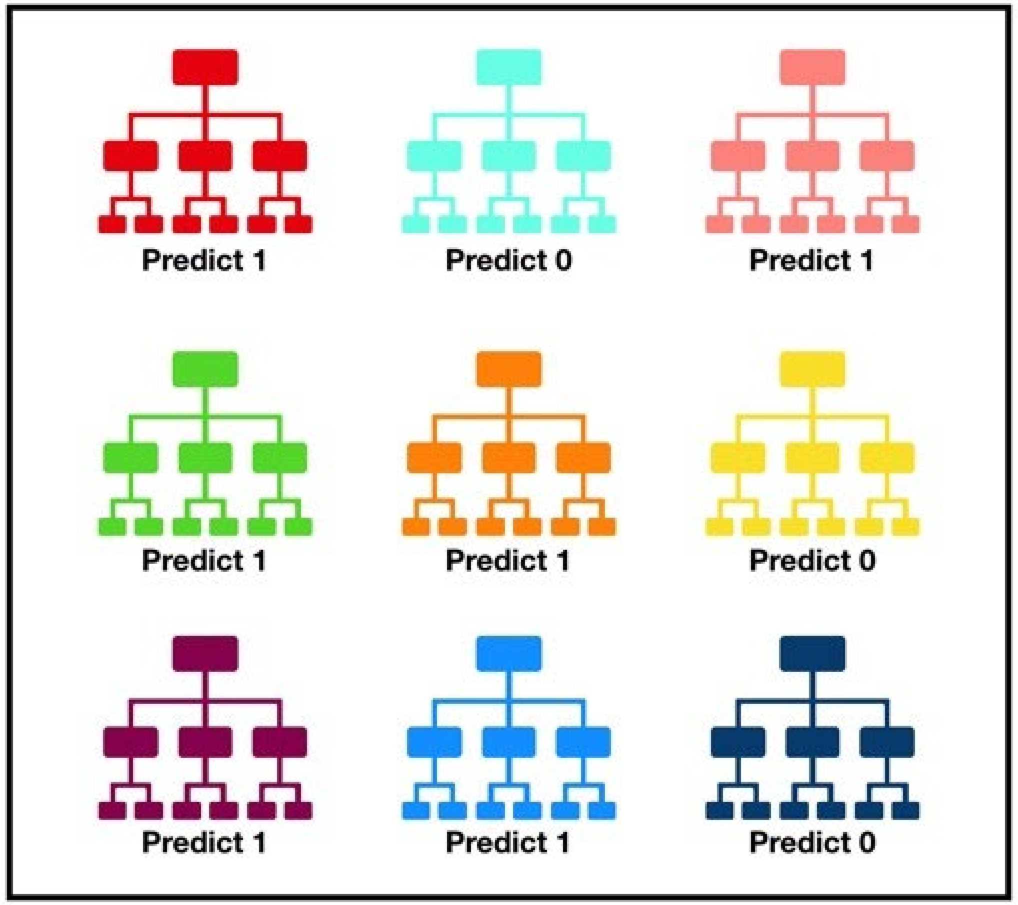
RANDOM FOREST (SUPERVISED)

Random forest: multiple individual decision trees operating as an ensemble.

One class prediction from each individual tree

=> model's prediction = class with most votes

=> more accurate and stable prediction



Tally: Six 1s and Three 0s
Prediction: 1

RANDOM FOREST (SUPERVISED)

Advantages:

- Solve **both regression and classification** problems with large data sets.
- Help identify most significant variables from thousands of input variables.
- Highly scalable to any number of dimensions with generally quite acceptable performances.

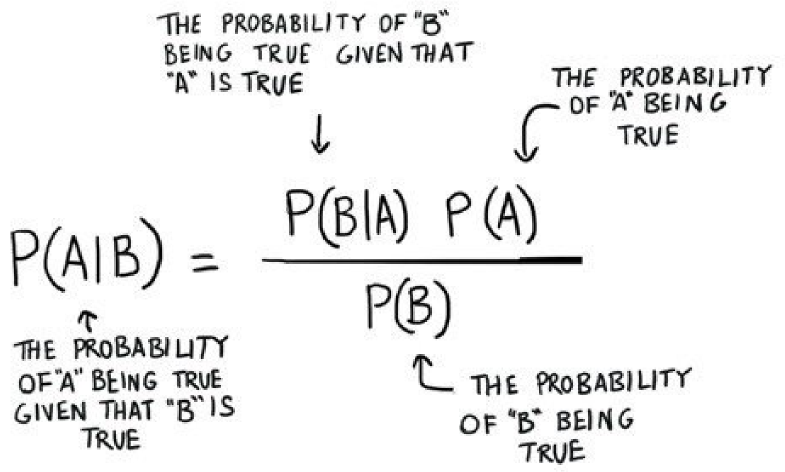
Disadvantages:

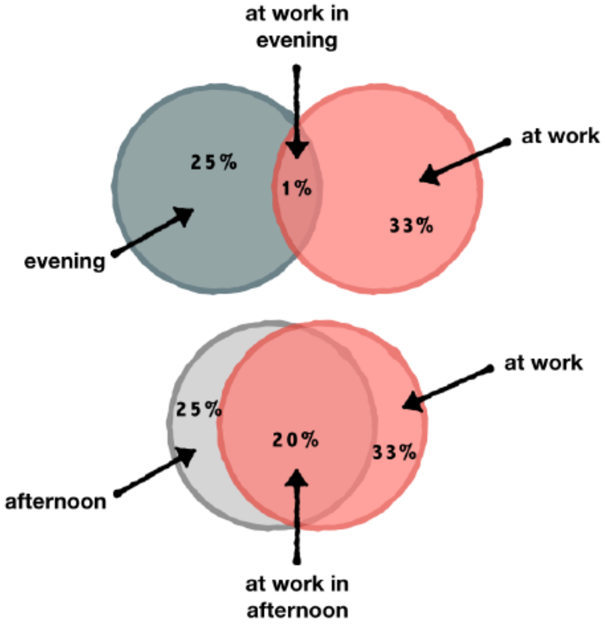
- ***Learning may be slow*** (depending on the parameterization)
- It is not possible to iteratively improve the generated models

E.g. Predict patients for high risks for certain diseases

NAIVE BAYES (SUPERVISED)

Classification technique based on Bayes' theorem (conditional probability and dependent events).


THE PROBABILITY OF "B" BEING TRUE GIVEN THAT "A" IS TRUE
THE PROBABILITY OF "A" BEING TRUE
P(A|B) = P(B|A) P(A) / P(B)
THE PROBABILITY OF "A" BEING TRUE GIVEN THAT "B" IS TRUE
THE PROBABILITY OF "B" BEING TRUE



The **conditional probability** of events A and B is denoted P(A | B)

- P(A | B) = P(A and B) / P(B)
- P(work | evening) = 1 / 25 = 4%
- P(work | afternoon) = 20 / 25 = 80%

NAIVE BAYES (SUPERVISED)

Advantages:

- very easy to build and particularly useful for very large data sets.
- perform well for both binary and multi-class classifications.
- a good choice when CPU and memory resources are a limiting factor or if something fast and easy that performs pretty well is needed.

Disadvantages:

- Assume all the features are independent/unrelated, then cannot learn the interactions between features.

E.g.

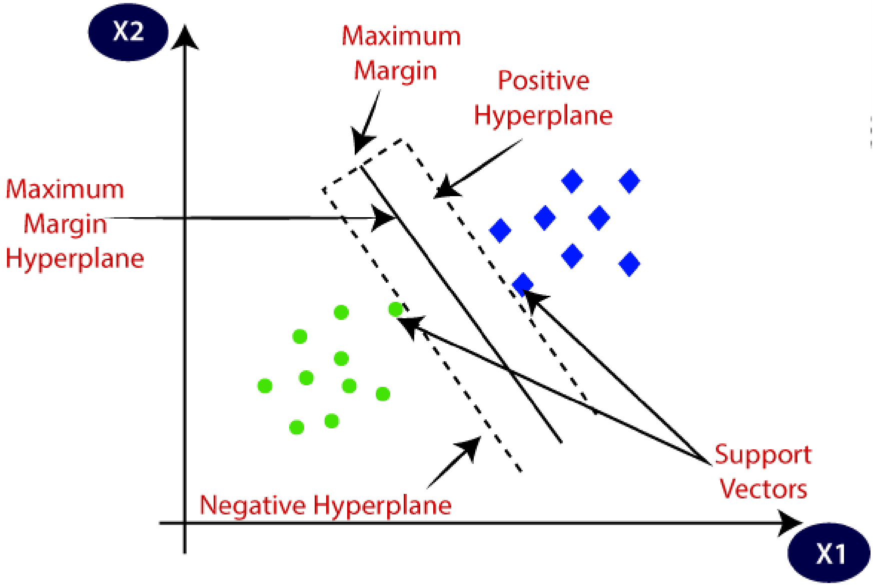
- mining housekeeping genes
- genetic association studies
- discovering Alzheimer genetic biomarkers from whole genome sequencing (WGS) data

SUPPORT VECTOR MACHINES (SUPERVISED)

Used for both classification and regression problems, but primarily for classification

Classification: when the data has exactly two classes.

Goal: find the best decision boundary (hyperplan) that differentiates the two classes in n-dimensional space (n features)



https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm

SUPPORT VECTOR MACHINES (SUPERVISED)

Advantages:

- high accuracy
- high dimensional data
- work with both linearly and non-linearly separable data, with an appropriate kernel

Disadvantages:

- memory-intensive
- hard to interpret
- and difficult to tune.

E.g.

- Detecting common diseases such as diabetes
- Classification of genomic islands
- Classification of genes

SUPERVISED LEARNING - DATASETS

Training dataset

A subset of the dataset provided to the algorithm for learning

Validation dataset

A subset used to tune the trained model parameters

Test dataset

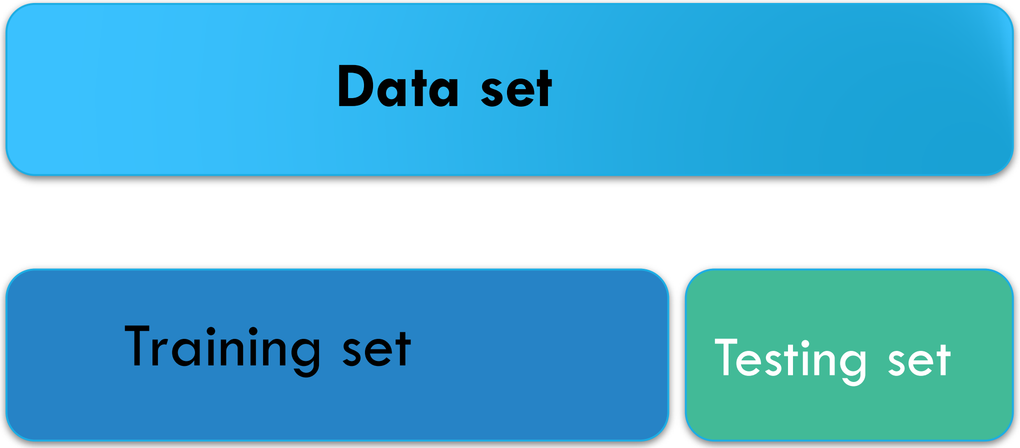
A dataset used only to assess the performance of a fully-specified model (classifier/regressor)

VALIDATION OF SUPERVISED ML ALGORITHMS RESULTS

To test the performance of the learning system:

- The trained model can be tested with objects with known labels (and were excluded from the training set because they were intended to be used for this purpose).
- Based on the results on the test data, the performance of the learning system can be assessed.

TRAINING SET AND TEST SET



Data set

Training set

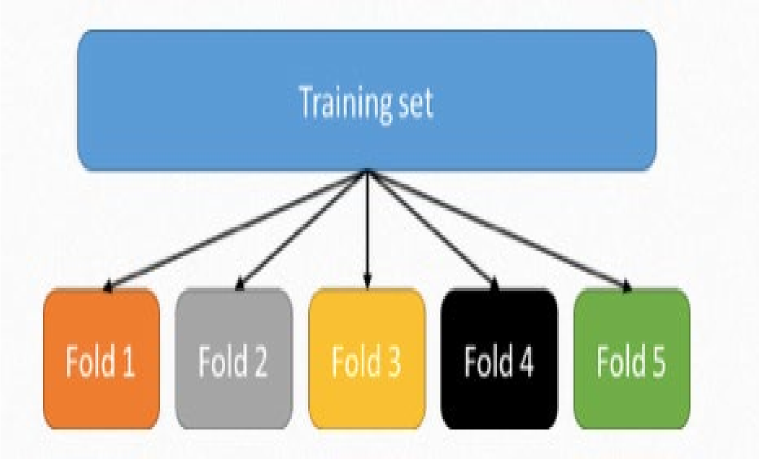
Testing set

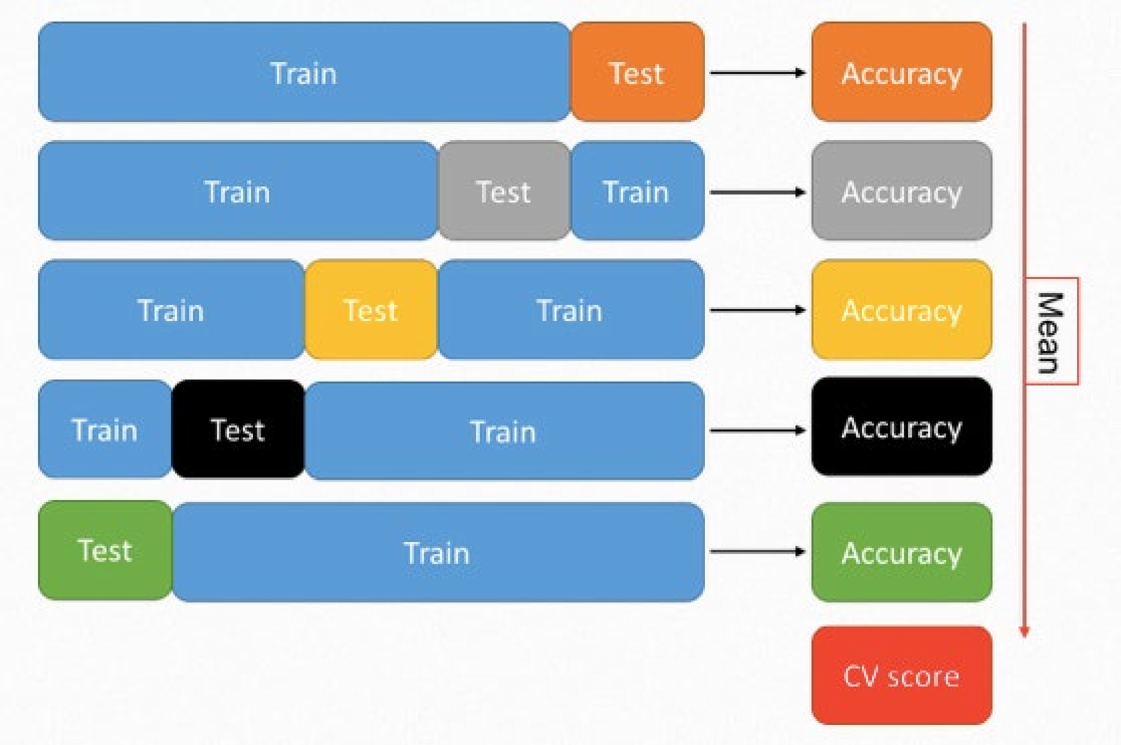
Used to train the algorithm

Estimate the accuracy of the model

Split the dataset randomly!
Use cross-validation
Underfitting and over fitting problems

K-FOLD CROSS VALIDATION





https://aldro61.github.io/microbiome-summer-school-2017/sections/basics/#type-of-learning-problems

Classification metrics

WHY THE NEED TO EVALUATE?

- Multiple methods are available to classify or predict
- For each method, multiple choices are available for settings
- To choose best model, need to assess each model's performance

MISCLASSIFICATION ERROR

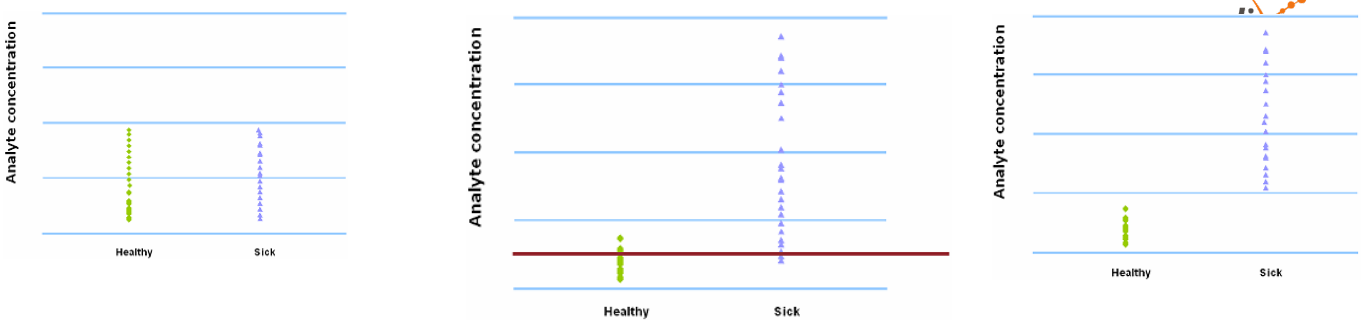
- **Error** = classifying a record as belonging to one class when it belongs to another class.

- **Error rate** = percent of misclassified records out of the total records in the validation data

DIFFERENT SCORING METRICS

1. Confusion Matrix
   - True positives
   - False negatives
   - False positives
   - True negatives
2. Sensitivity and Specificty
3. Precision and Recall
4. F-measure
5. Overall accuracy and Cohen's kappa

MAIN DEFINITIONS



- Confusion matrix

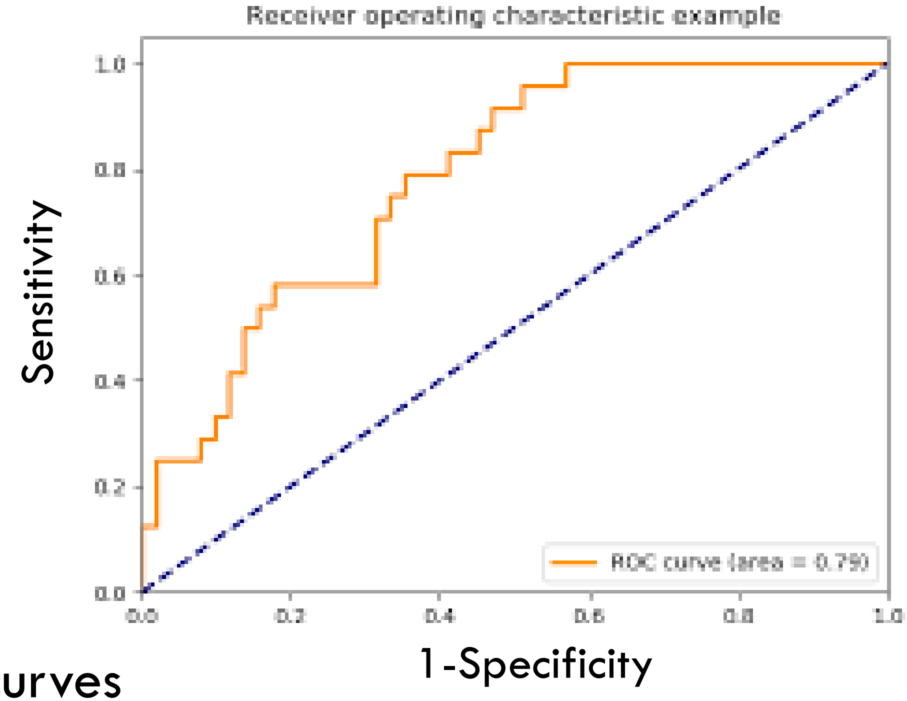
n=165	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

- Precision $\frac{tp}{tp + fp}$

- Specificity $\frac{TN}{FP+TN}$

- Recall / Sensitivity $\frac{tp}{tp + fn}$

- Receiver Operating Characteristic (ROC) and AUC curves



https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html

F-MEASURE

F-measure $= \; 2 * \frac{precision * recall}{precision + recall}$

Harmonic mean of precision and recall

Are ALL and ONLY positive class events found by the model?

OVERALL ACCURACY

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

Target class distribution must be balanced!

Probability of classifying a positive OR negative class event correctly.

WHY DIFFERENT METRICS?

1. What is your objectives?

2. What is the target class distribution?

3. Is the target binomial or multinomial?

Supervised Learning - Regression

WHAT IS REGRESSION?

Regression: Predict a numerical outcome ("dependent variable") from a set of inputs ("independent variables").

Statistical Sense: Predicting the expected value of the outcome.

Casual Sense: Predicting a numerical outcome, rather than a discrete one.

WHAT IS REGRESSION?

How many patients will come to the emergency unit on a Sunday evening? (**Regression**)

Is this histopathological image classified as "cancer" or "non-cancer" type? (**Classification**)

How many days will this patient spend in the hospital? (**Regression**)

LINEAR REGRESSION (SUPERVISED)

Regression algorithms can be used for example when some continuous value needs to be computed as compared to classification where the output is categorical.

So whenever there is a need to predict some future value of a process which is currently running, regression algorithm can be used.

Operating on a two dimensional set of observations (two continuous variables), simple linear regression attempts to fit, as best as possible, a line through the data points.

The regression line (our model) becomes a tool that can help uncover underlying trends in our dataset.

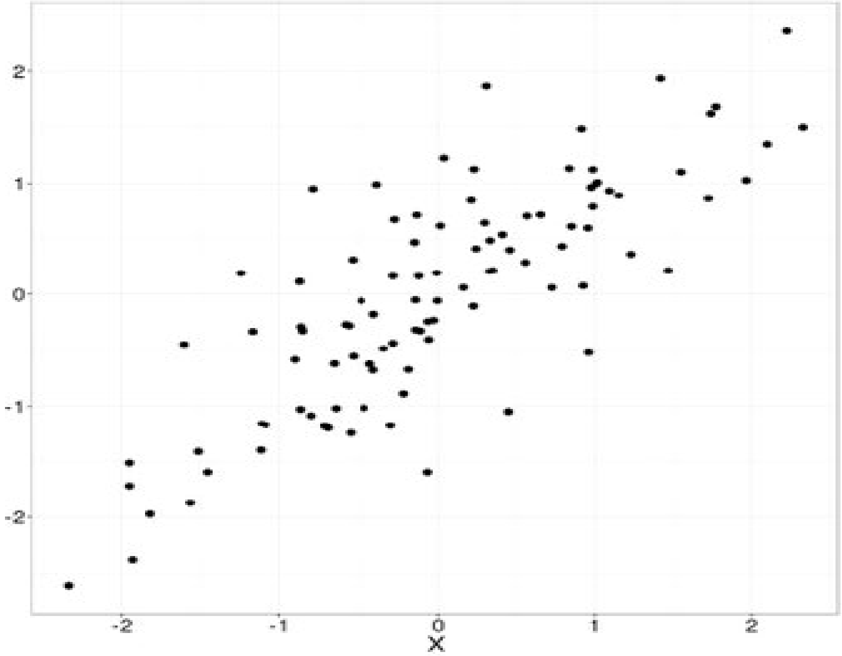
The regression line, when properly fitted, can serve as a predictive model for new events.

Linear Regressions are however unstable in case features are redundant, i.e. if there is multicollinearity

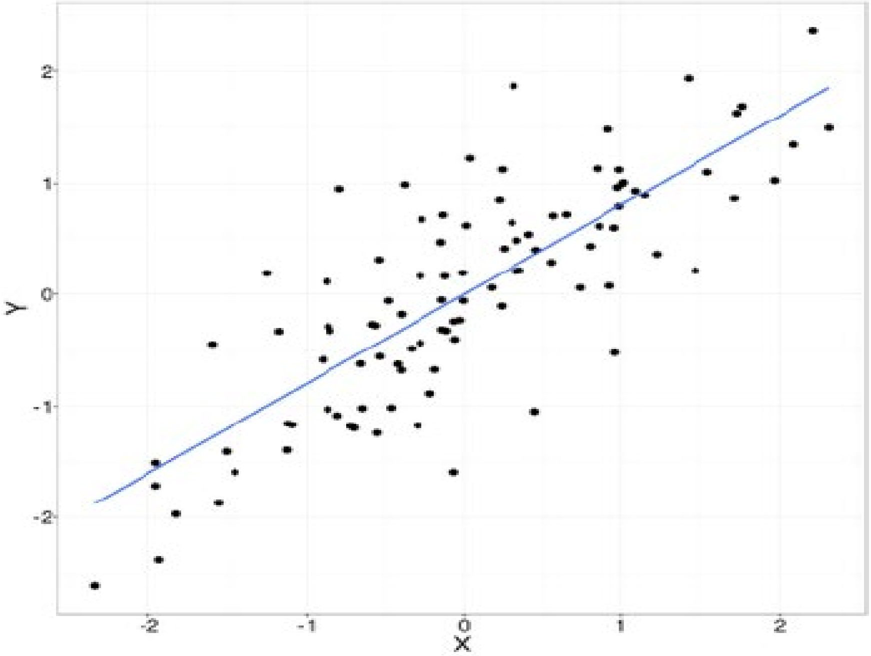
Example where linear regression can be used is:

- predicting drug resistance by correlating genotypic information with phenotypic profiles

APPLYING LINEAR REGRESSION



Scatterplot of our dataset.

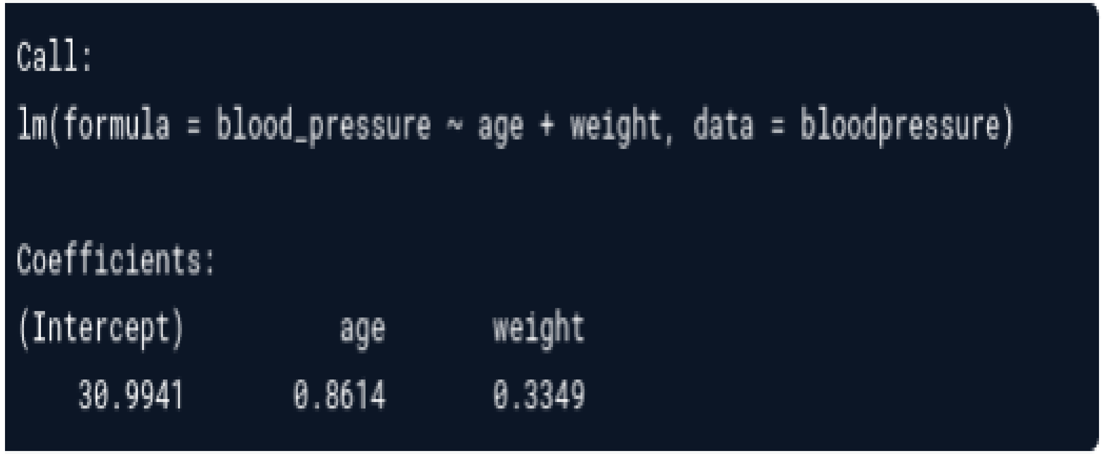


Fitting of the regression line (blue).

LINEAR REGRESSION – PROS AND CONS

Pros

- Easy to fit and apply
- Concise
- Less prone to over-fitting
- Interpretable



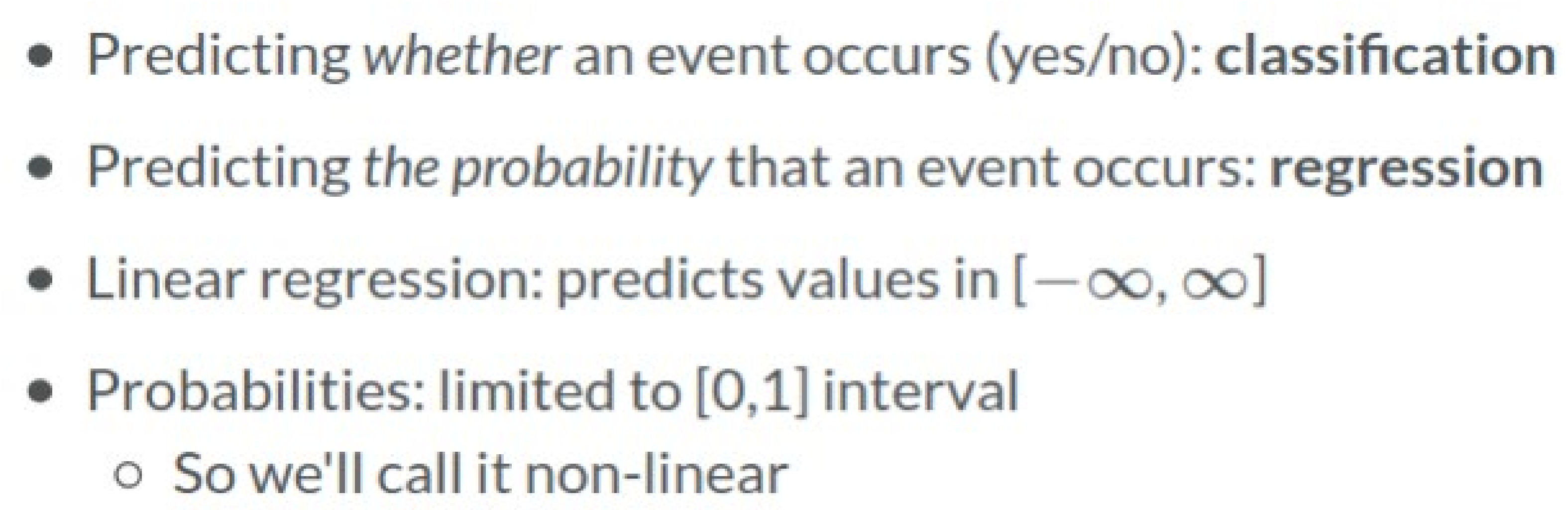
```
Call:
lm(formula = blood_pressure ~ age + weight, data = bloodpressure)

Coefficients:
(Intercept)          age       weight
    30.9941       0.8614       0.3349
```

Cons

- Can only express linear and additive relationships

LOGISTIC REGRESSION (SUPERVISED)

It is a regression model that predicts probabilities



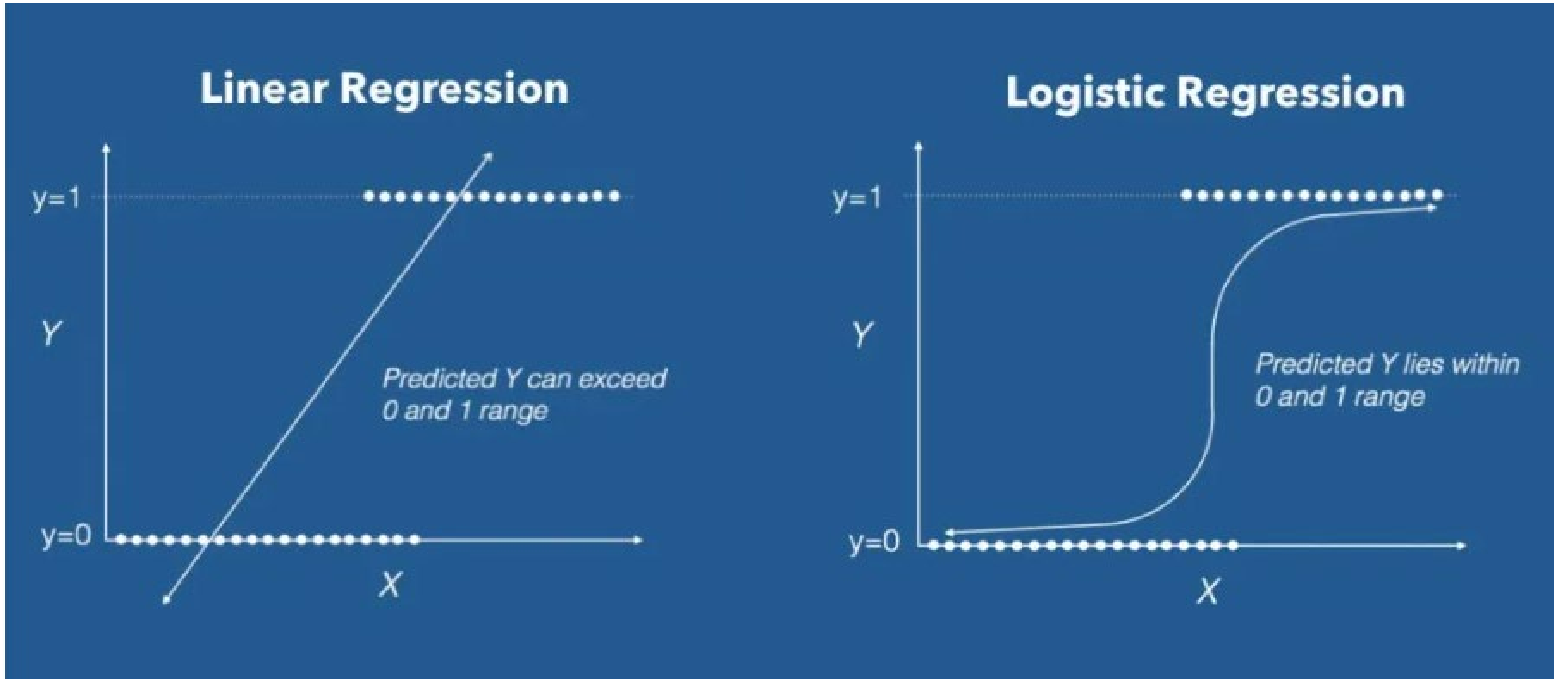
- Predicting *whether* an event occurs (yes/no): **classification**
- Predicting *the probability* that an event occurs: **regression**
- Linear regression: predicts values in $[-\infty, \infty]$
- Probabilities: limited to [0,1] interval
  - So we'll call it non-linear

Note: Classification refers to predicting whether an event will occur (Yes/No). While **regression** refers to the probability that an event will occur.

EXAMPLE OF LOGISTIC REGRESSION – PREDICTING DUCHENNE MUSCULAR DYSTROPHY (DMD)

We want to develop a test to detect the gene for DMD in women.

- The test uses the measurements of 2 enzymes in the blood (CK and H).
- What is the probability that a woman is a DMD carrier based on her CK and H levels?
- We cannot use linear regression (where the outcome is 0:False and 1:True), because the linear model will predict probabilities outside the range of 0 and 1.


Linear Regression
y=1
Y
y=0
X
Predicted Y can exceed 0 and 1 range
Logistic Regression
y=1
Y
y=0
X
Predicted Y lies within 0 and 1 range

GENERALIZED LINEAR MODELS (GLM)

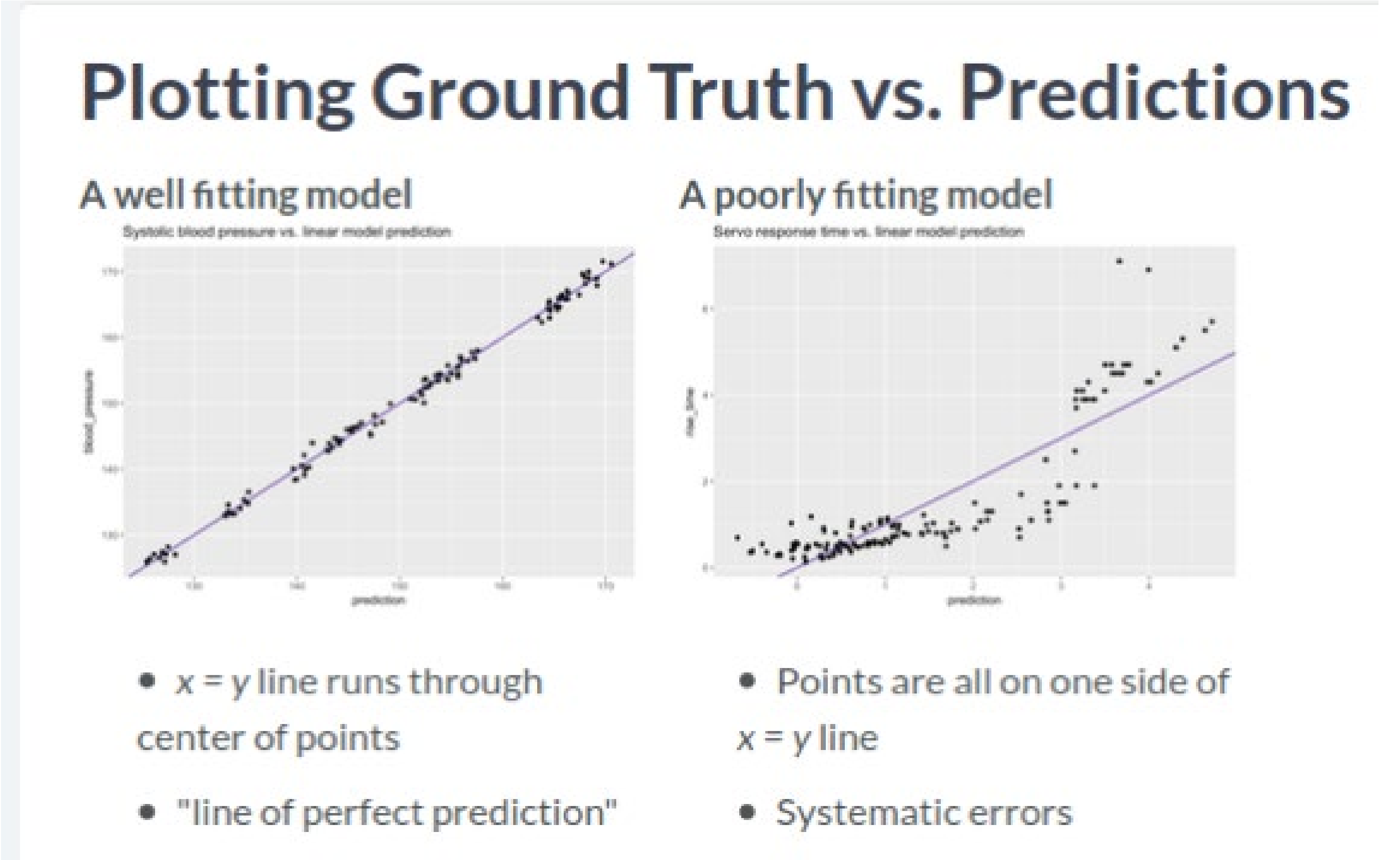
- The term ***generalized linear model*** (GLIM or GLM) refers to a larger class of models
- In these models, the response variable y_i is assumed to follow an exponential family distribution with mean μ_i, which is assumed to be some (often nonlinear) function.
- GLMs are a broad class of models that include linear regression, ANOVA, Poisson regression, log-linear models etc.
- Some of the models are:

Model	Probability Distribution
Linear Regression	Normal
Logistic Regression	Binomial
Poisson Regression	Poisson

Evaluating a Regression Model

EVALUATING OUR REGRESSION MODEL GRAPHICALLY

First of all we can visualize our ground truths vs the predicted values to see how well our model has performed the predictions.



Plotting Ground Truth vs. Predictions

A well fitting model

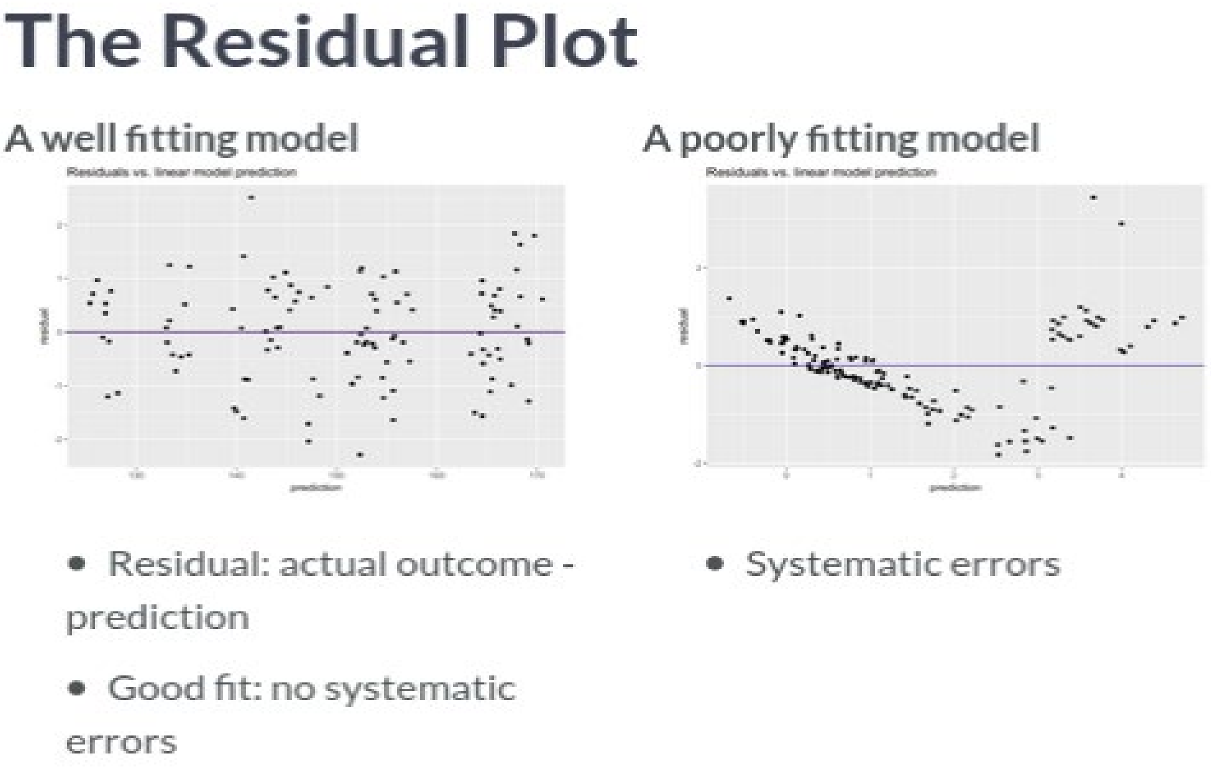
- x = y line runs through center of points
- "line of perfect prediction"

A poorly fitting model

- Points are all on one side of x = y line
- Systematic errors

EVALUATING OUR REGRESSION MODEL GRAPHICALLY

Secondly we can also visualize the residuals against the predictions



The Residual Plot

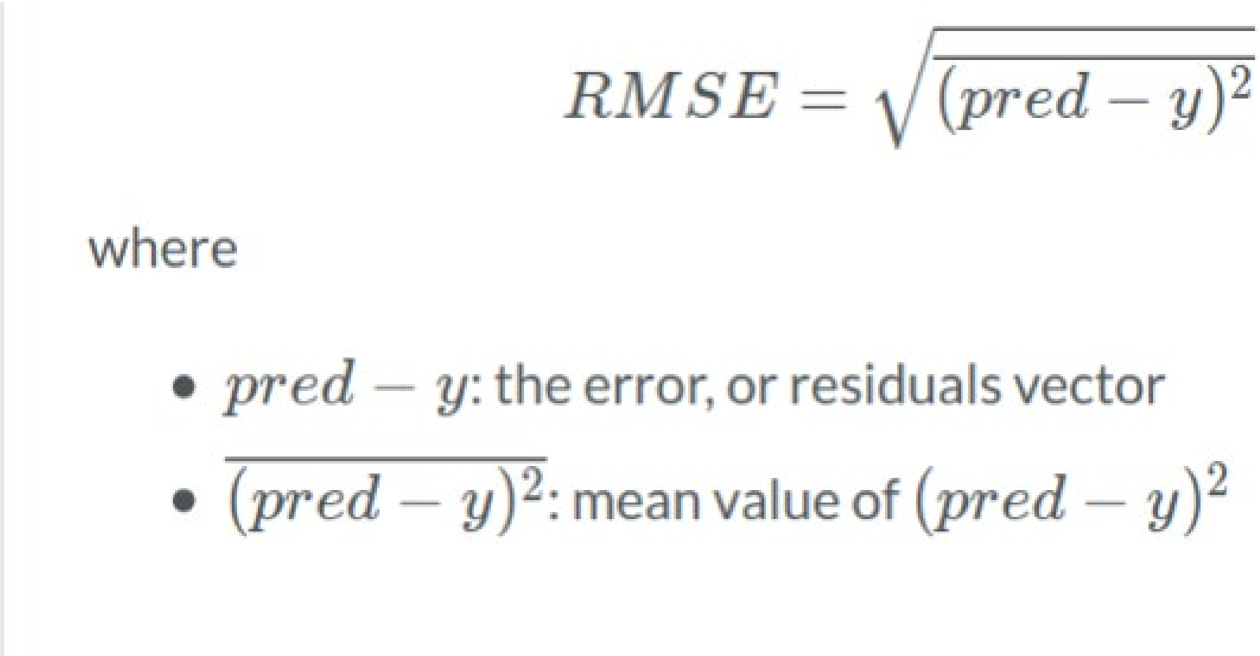
A well fitting model

- Residual: actual outcome - prediction
- Good fit: no systematic errors

A poorly fitting model

- Systematic errors

EVALUATION OF OUR REGRESSION MODEL – USING RMSE (ROOT MEAN SQUARE ERROR)



$$RMSE = \sqrt{\overline{(pred - y)^2}}$$

where

- $pred - y$: the error, or residuals vector
- $\overline{(pred - y)^2}$: mean value of $(pred - y)^2$

EVALUATION OF OUR REGRESSION MODEL – USING R^2

Coefficient of Determination or R^2 is another metric used for evaluating the performance of a regression model.

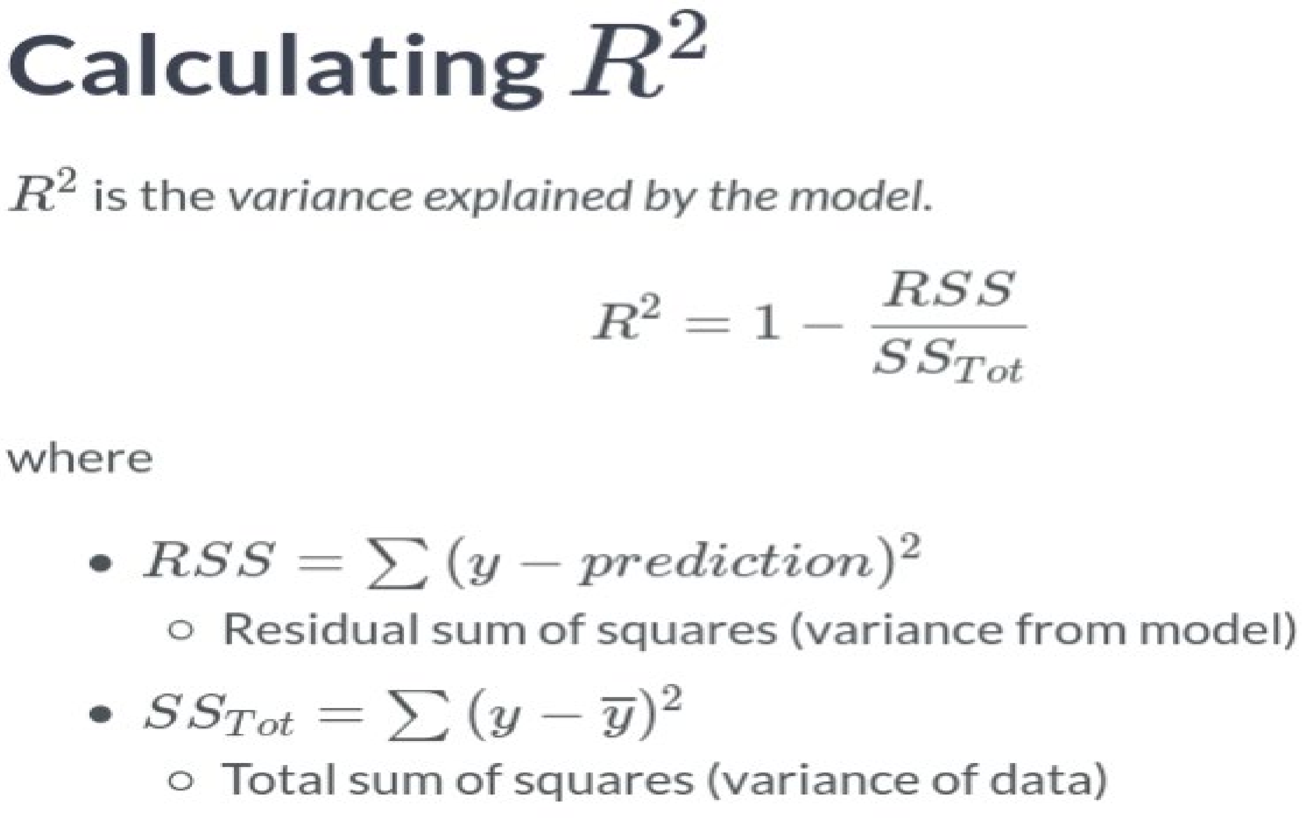
It helps us to compare our current model with a constant baseline and tells us how much our model is better.

The constant baseline is chosen by taking the mean of the data and drawing a line at the mean.

R^2 is a scale-free score that implies it doesn't matter whether the values are too large or too small, the R^2 will always be less than or equal to 1.

The closer the value of R^2 to 1, the better is our model

EVALUATION OF OUR REGRESSION MODEL – USING R^2



Calculating R^2

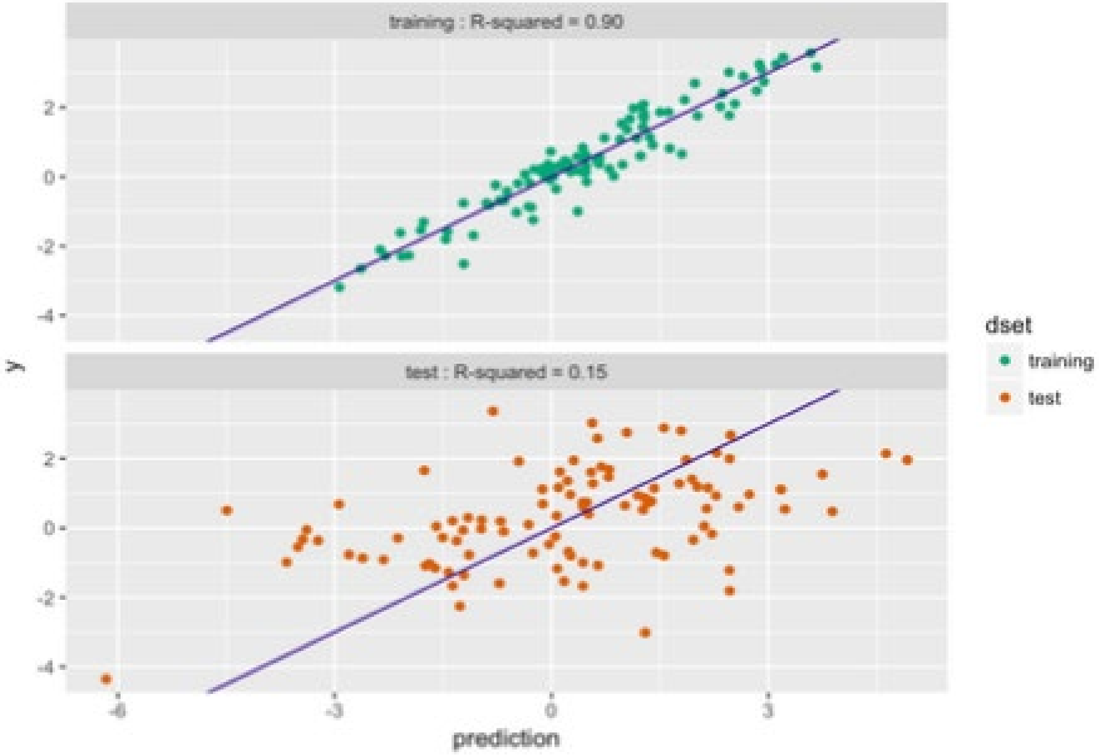
R^2 is the *variance explained by the model.*

$$R^2 = 1 - \frac{RSS}{SS_{Tot}}$$

where

- $RSS = \sum(y - prediction)^2$
  - Residual sum of squares (variance from model)
- $SS_{Tot} = \sum(y - \overline{y})^2$
  - Total sum of squares (variance of data)

REGRESSION – PROPERLY TRAINING A MODEL



- Training R^2: 0.9; Test R^2: 0.15 -- **Overfit**

REGRESSION – PROPERLY TRAINING A MODEL

In general models can perform much better on training than on data they have not yet seen.

For simple models, this difference between training data and test data results is often not severe.

But for more complex models or even for linear model with too many variables, using only the training data to evaluate the model can produce misleading results.

In the previous slide example we get the value of R^2 as 0.9 on training data but 0.15 on new data.

- It means this model was overfit.

When we have a lot of data, the best thing to do is to split your data into 2, one set to train the model and another set to test it.

When we don't have enough data we must do cross-validation

Going into the “grey” area

SEMI-SUPERVISED MACHINE LEARNING ALGORITHMS

In supervised learning, the algorithm receives as input a collection of data points, each with an associated label, whereas in unsupervised learning the algorithm receives the data but no labels.

- The semi-supervised setting is a mixture of these two approaches: the algorithm receives a collection of data points, but only a subset of these data points have associated labels.

So, they fall somewhere in between supervised and unsupervised learning, since they use both labeled and unlabeled data for training – **typically a small amount of labeled data and a large amount of unlabeled data**.

The systems that use this method are able to considerably improve learning accuracy.

SEMI-SUPERVISED MACHINE LEARNING ALGORITHMS

Consider the gene finding model where the system is provided with labeled data and unlabeled data.

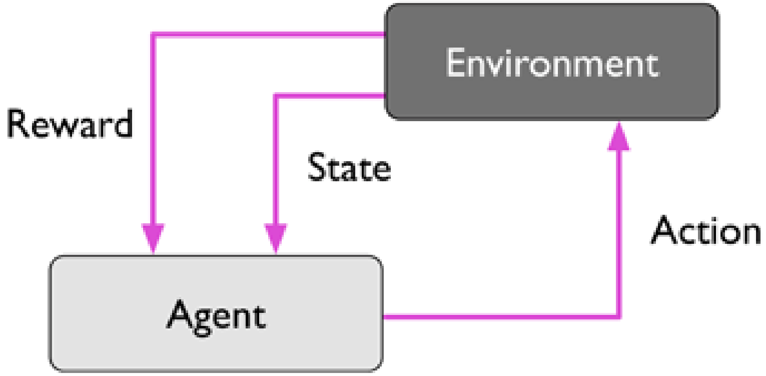
- The learning procedure begins by constructing an initial gene-finding model on the basis of the labeled subset of the training data alone.
- Next, the model is used to scan the genome, and tentative labels are assigned throughout the genome.
- These tentative labels can then be used to improve the learned model, and the procedure iterates until no new genes are found.

SEMI-SUPERVISED MACHINE LEARNING ALGORITHMS

In practice, gene-finding systems are often trained using a semi-supervised approach, in which the input is a collection of annotated genes and an unlabeled whole-genome sequence.

The semi-supervised approach can work much better than a fully supervised approach because the model is able to learn from a much larger set of genes — all of the genes in the genome — rather than only the subset of genes that have been identified with high confidence.

REINFORCEMENT MACHINE LEARNING ALGORITHMS



The learning system interacts with the environment by producing actions and discovers errors or rewards.

- The goal is to develop a system (agent) that improves its performance based on interactions with its environment.

Through its interaction with the environment, an agent can then use reinforcement learning to learn a series of actions that maximizes this reward via an exploratory trial-and-error approach or deliberative planning.

REINFORCEMENT MACHINE LEARNING ALGORITHMS

The idea behind ***Reinforcement Learning*** is that an agent will learn from the environment by interacting with it and receiving rewards for performing actions.

Learning from interaction with the environment comes from our natural experiences.

- Consider a child in a living room who sees a fireplace and approaches it.
- It’s warm, it’s positive, the child feels good (*Positive Reward* +1) and understands that fire is a positive thing.
- Next he tries to touch the fire and it burns his hand (Negative reward -1). He then understands that fire is positive when he is a sufficient distance away, because it produces warmth. But getting too close to it, he will be burned.

DEEP LEARNING ALGORITHMS

Also known as deep structured learning or hierarchical learning

It is a subfield of machine learning concerned with algorithms inspired by the structure and function of the brain called artificial neural networks.

Can perform learning in supervised and/or unsupervised manners.

Teach computers to do what comes naturally to humans: **learn by example**

- key technology behind driverless cars, enabling them to recognize a stop sign, or to distinguish a pedestrian from a lamppost.
- **In medical Research**
  - Cancer researchers are using deep learning to automatically detect cancer cells.
  - Teams at UCLA built an advanced microscope that yields a high-dimensional data set used to train a deep learning application to accurately identify cancer cells.

DEEP LEARNING ALGORITHMS

While deep learning was first theorized in the 1980s, there are two main reasons it has only recently become useful:

- Deep learning requires large amounts of labeled data.
  - For example, driverless car development requires millions of images and thousands of hours of video.
- Deep learning requires substantial computing power.
  - High-performance GPUs have a parallel architecture that is efficient for deep learning.
  - When combined with clusters or cloud computing, this enables development teams to reduce training time for a deep learning network from weeks to hours or less.

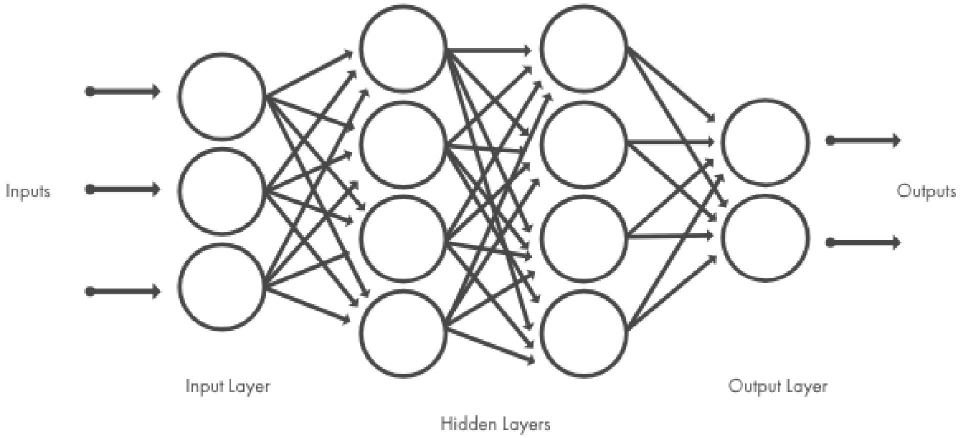
DEEP LEARNING ALGORITHMS

Most deep learning methods use neural network architectures, which is why **deep learning models** are often referred to as **deep neural networks**.

The term “**deep**” usually refers to the number of hidden layers in the neural network.

- Traditional neural networks only contain 2-3 hidden layers, while deep networks can have as many as 150.

Deep learning models are trained by using large sets of labeled data and neural network architectures that learn features directly from the data without the need for manual feature extraction.



DEEP LEARNING ALGORITHMS

Deep learning is now one of the most active fields in machine learning and has been shown to improve performance in image and speech recognition.

The potential of deep learning in high-throughput biology is clear

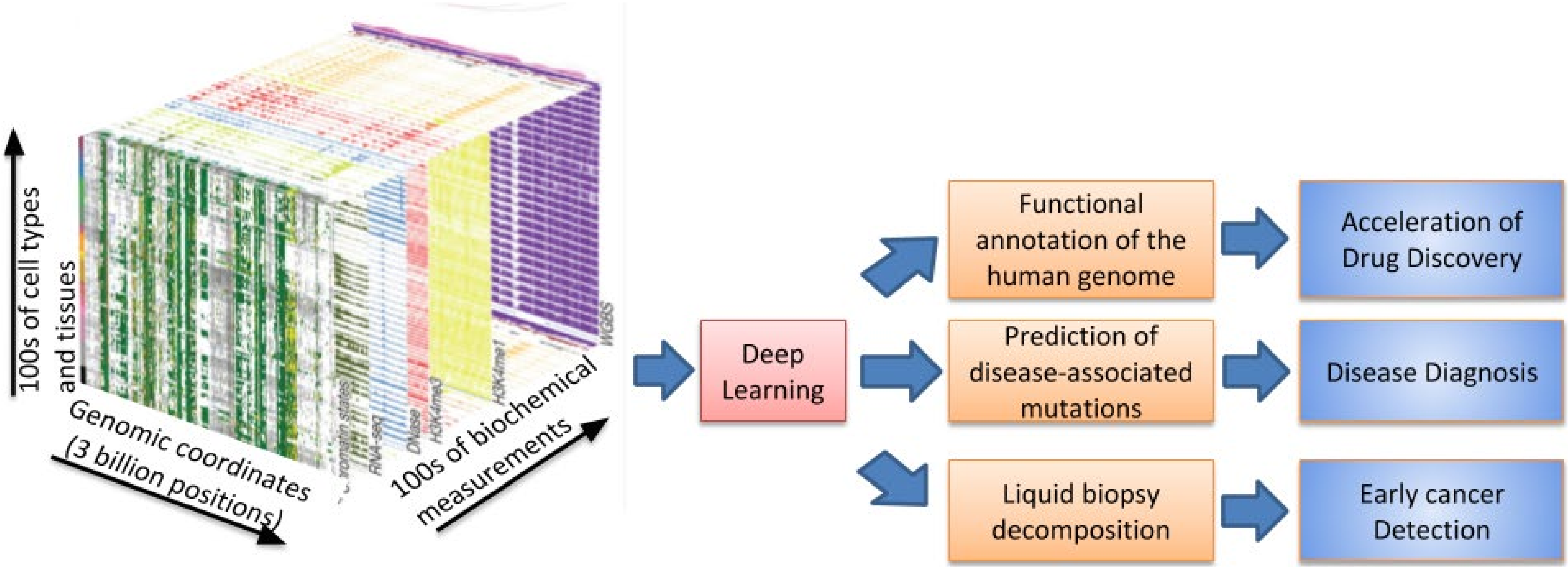
- it allows to better exploit the availability of increasingly large and high-dimensional data sets (e.g. from DNA sequencing, RNA measurements, flow cytometry or automated microscopy) by training complex networks with multiple layers that capture their internal structure

DEEP LEARNING ALGORITHMS

Example

- **Multi-label Deep Learning for Gene Function Annotation in Cancer Pathways** [Renchu Guan, Xu Wang, Mary Qu Yang, Yu Zhang, Fengfeng Zhou, Chen Yang & Yanchun Liang Scientific Reports volume 8, (2018)]
- Applied deep learning to explore full texts of biomedical articles containing detailed methodologies, experimental results, critical discussions and interpretations can be found, for the analysis of gene multi-functions relevant to cancer pathways derived from full-text biomedical publications.
  - Without the involvement of a biologist to do a feature study about the data.
- Experimental results on eight KEGG cancer pathways revealed that this new system is not only superior to classical multi-label learning models, but it can also achieve numerous gene functions related to important cancer pathways.

OPPORTUNITIES FOR DEEP LEARNING IN GENOMICS


100s of cell types and tissues
Genomic coordinates (3 billion positions)
100s of biochemical measurements
Chromatin states
RNA-seq
DNase
H3K4me3
H3K4me1
WGBS

Deep Learning

Functional annotation of the human genome

Prediction of disease-associated mutations

Liquid biopsy decomposition

Acceleration of Drug Discovery

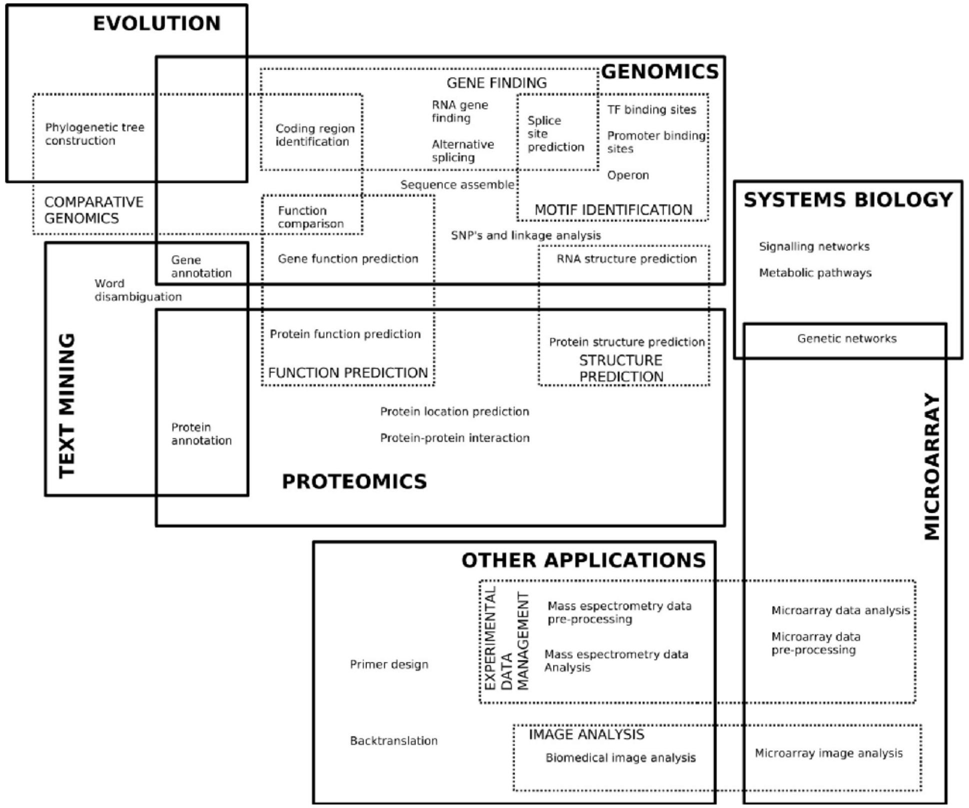
Disease Diagnosis

Early cancer Detection

https://towardsdatascience.com/opportunities-and-obstacles-for-deep-learning-in-biology-and-medicine-6ec914fe18c2

In closing

APPLICATIONS OF ML IN BIOINFORMATICS



From: Machine learning in bioinformatics
Brief Bioinform. 2006;7(1):86-112. doi:10.1093/bib/bbk007

IS THERE A PERFECT ML TECHNIQUE?

There is not one solution (one machine learning algorithm) or one approach that fits all problems.

For each problem, there is not one single solution.

WHICH TECHNIQUE TO USE?

Size, quality and nature of the data to be analysed.

The question, the answer expected, and also expected accuracy.

How the result will be used

Time and computing resources available.

Always good to check performance of different algorithms and compare results.

WHAT KIND OF DATA DO YOU HAVE?

If the data to be analysed is unlabelled and the aim is to find structure, it is an unsupervised learning problem.

If the aim is to optimize an objective function by interacting with an environment, it is a reinforcement learning problem.

When supervised learning is feasible, it is often the case that additional, unlabelled data points are easy to obtain.

How do you decide whether it’s a supervised or semi-supervised approach?

A good rule of thumb is to use semi-supervised learning if you do not have very much labelled data and you have a very large amount of unlabelled data

WHAT IS THE EXPECTED OUTPUT?

If the output of your model is a number, it is a regression problem.

- Two-class classification of gene expression data

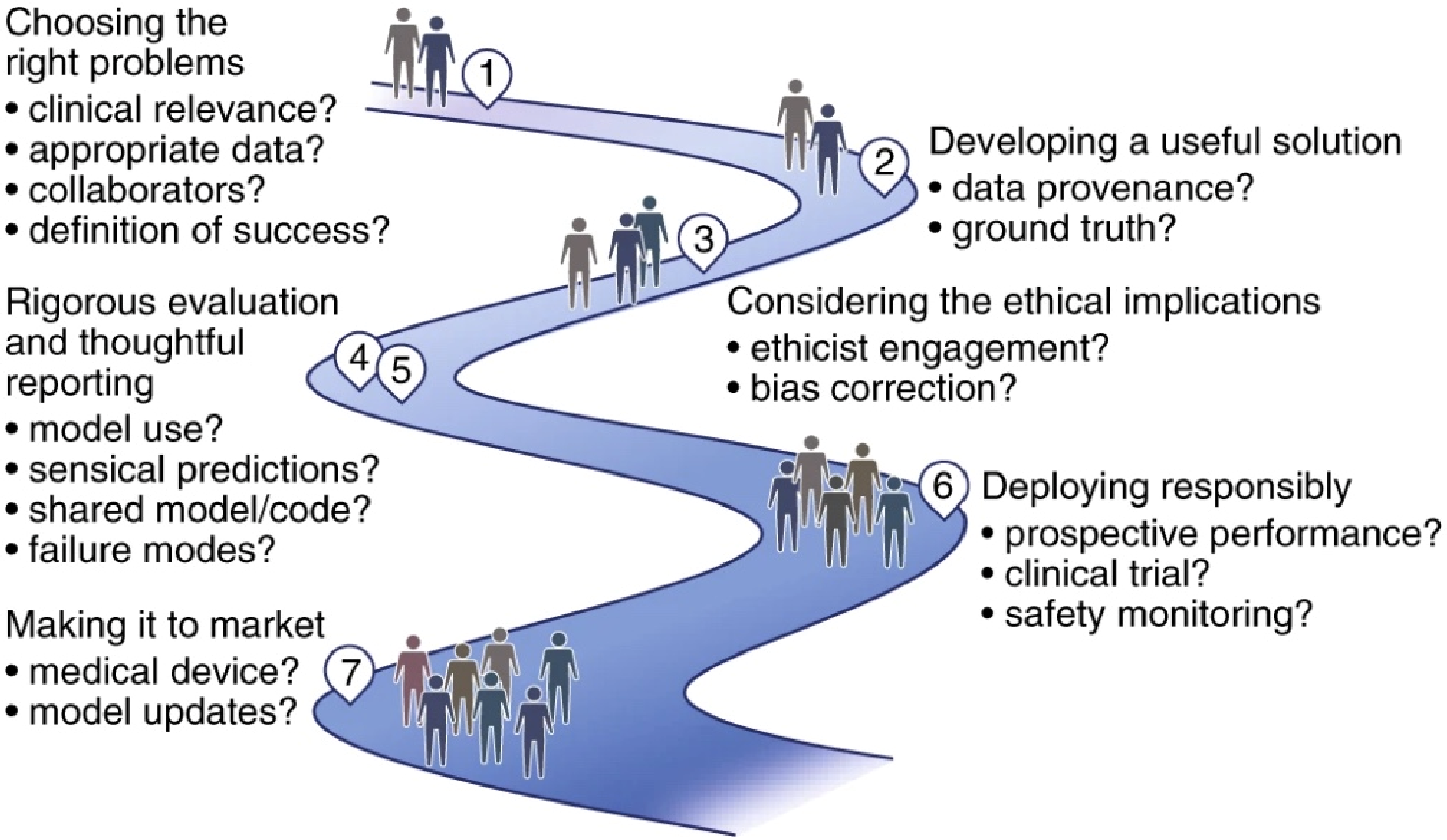
If the output of your model is a class, it is a classification problem.

- Genomic classification of AML

If the output of your model is a set of input groups, it is a clustering problem.

- Patterns in gene expression at different developmental stages of zebrafish

DO NO HARM: A ROADMAP FOR RESPONSIBLE MACHINE LEARNING FOR HEALTH CARE



Choosing the right problems
- clinical relevance?
- appropriate data?
- collaborators?
- definition of success?

Developing a useful solution
- data provenance?
- ground truth?

Considering the ethical implications
- ethicist engagement?
- bias correction?

Rigorous evaluation and thoughtful reporting
- model use?
- sensical predictions?
- shared model/code?
- failure modes?

Deploying responsibly
- prospective performance?
- clinical trial?
- safety monitoring?

Making it to market
- medical device?
- model updates?

https://www.nature.com/articles/s41591-019-0548-6/figures/1

TOOLS

All the methods listed above are already available either in Python, R (https://www.r-project.org/about.html) or Matlab using existing packages. Some basic code needs to be written.

If you are not used to writing code, you may use a tool like WEKA (https://www.cs.waikato.ac.nz/ml/weka/) or RapidMiner (https://rapidminer.com/) – the methods are already implemented and you simply need to load your data in either csv, arff,… format and run the selected methods.

Some useful R packages R implementing many ML techniques: https://cran.r-project.org/web/views/MachineLearning.html

SOME ONLINE RESOURCES

https://machinelearningmastery.com/start-here/

https://www.datascience.com/blog

https://www.mathworks.com/discovery/machine-learning.html

https://www.coursera.org/browse/data-science

SOURCES

http://www.sthda.com/english/articles/29-cluster-validation-essentials/97-cluster-validation-statistics-must-know-methods/#data-preparation

https://medium.mybridge.co/30-amazing-machine-learning-projects-for-the-past-year-v-2018-b853b8621ac7

Shakuntala Baichoo and Zahra Mungloo slides (H3ABionet, ML group)

NOW GO FORTH AND ML! ☺