Introduction to Machine Learning

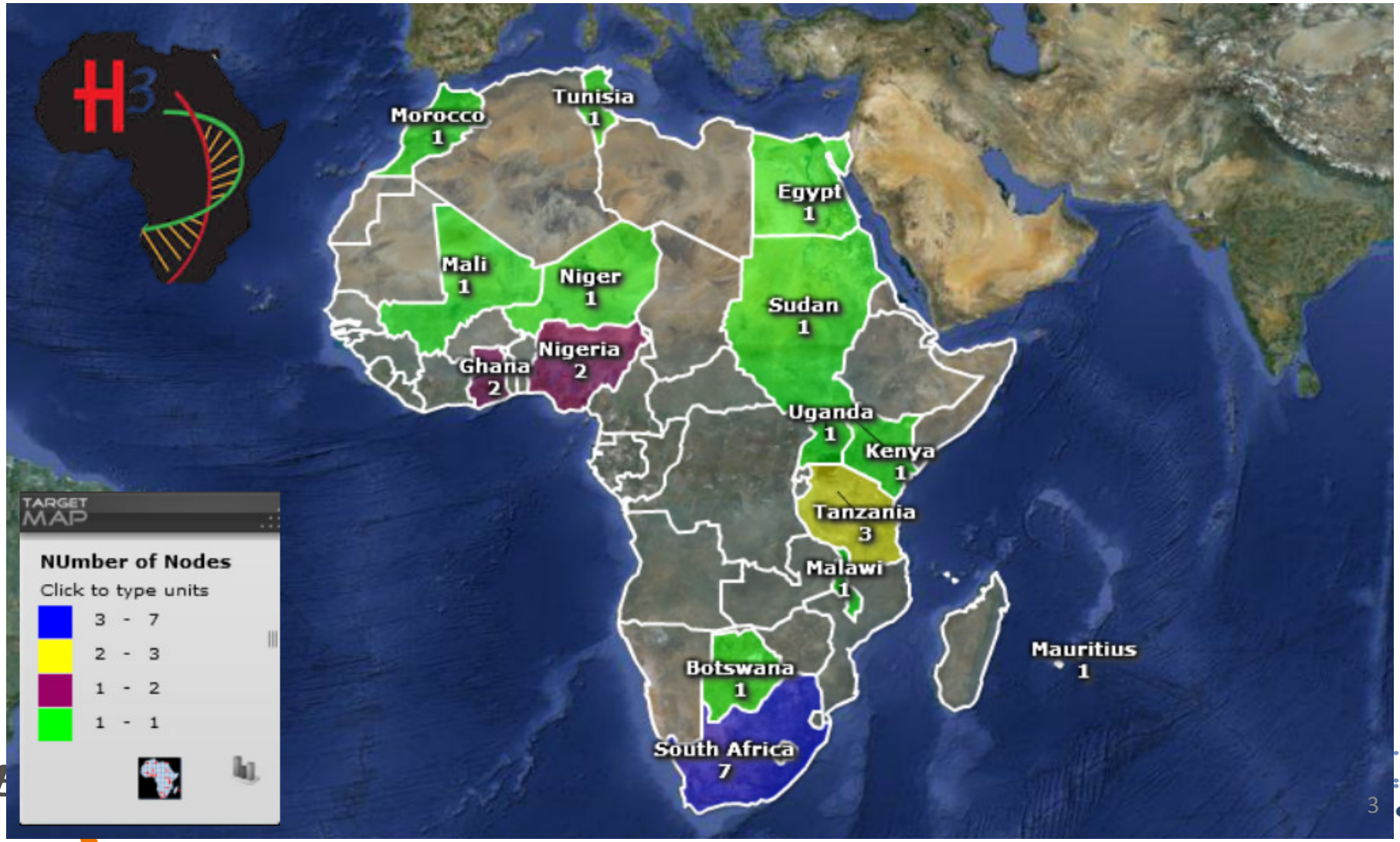
Amel Ghouila
amel.ghouila@pasteur.tn

@AmelGhouila

Institut Pasteur de Tunis

H3ABioNet

Pan African Bioinformatics Network for H3Africa



Session overview

01 Introduction to basic concepts of Data mining and Machine learning

02 Machine learning taxonomy

03 Supervised classification vs unsupervised classification

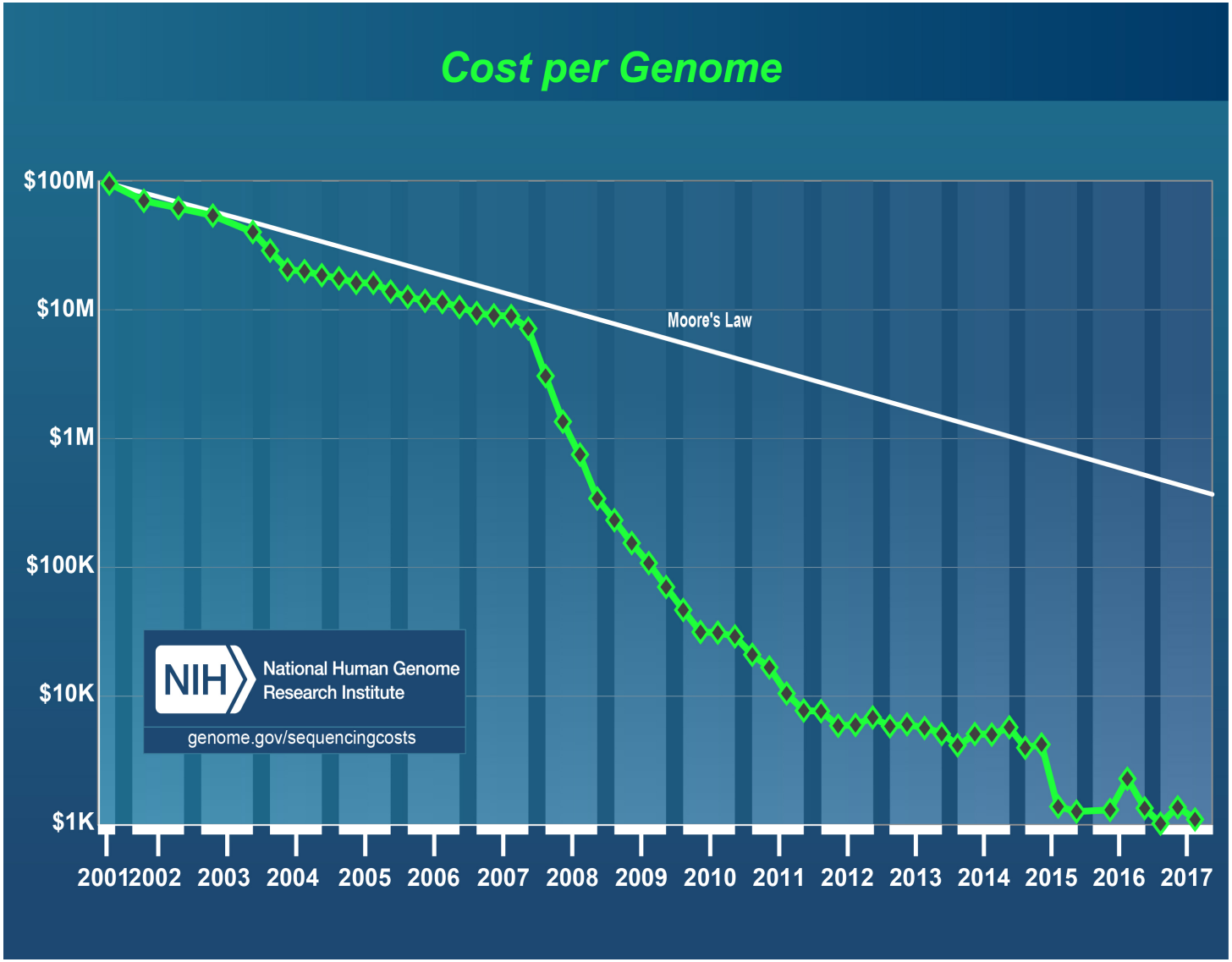
04 Algorithms examples

05 Examples of applications in Bioinformatics

Technology Timeline

1980	1990	2000	2010	2020
The age of processing	Hardware increase and World Wide Web	Portable devices and connectivity	Apps and Personalisation	Hyper Convergence & Internet of Things

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


Cost per Genome
$100M
$10M
$1M
$100K
$10K
$1K
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017

From Data to knowledge

AI & ML

- AI is a broader concept than ML which adresses the use of computers to mimic the congnitive 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 relashionship 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

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 sens

Compare to expectations

Analyze differences

Creativity, Limited memory

How does a computer work?

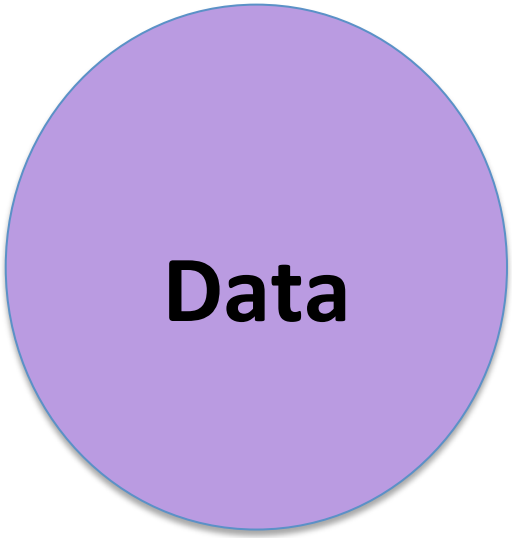
Follow instructions given by human

Artificial intelligence

Stimulate human behavior and cognitive process

Capture and preseve human expertise

Fast response
Ability to memorize big amounts of data

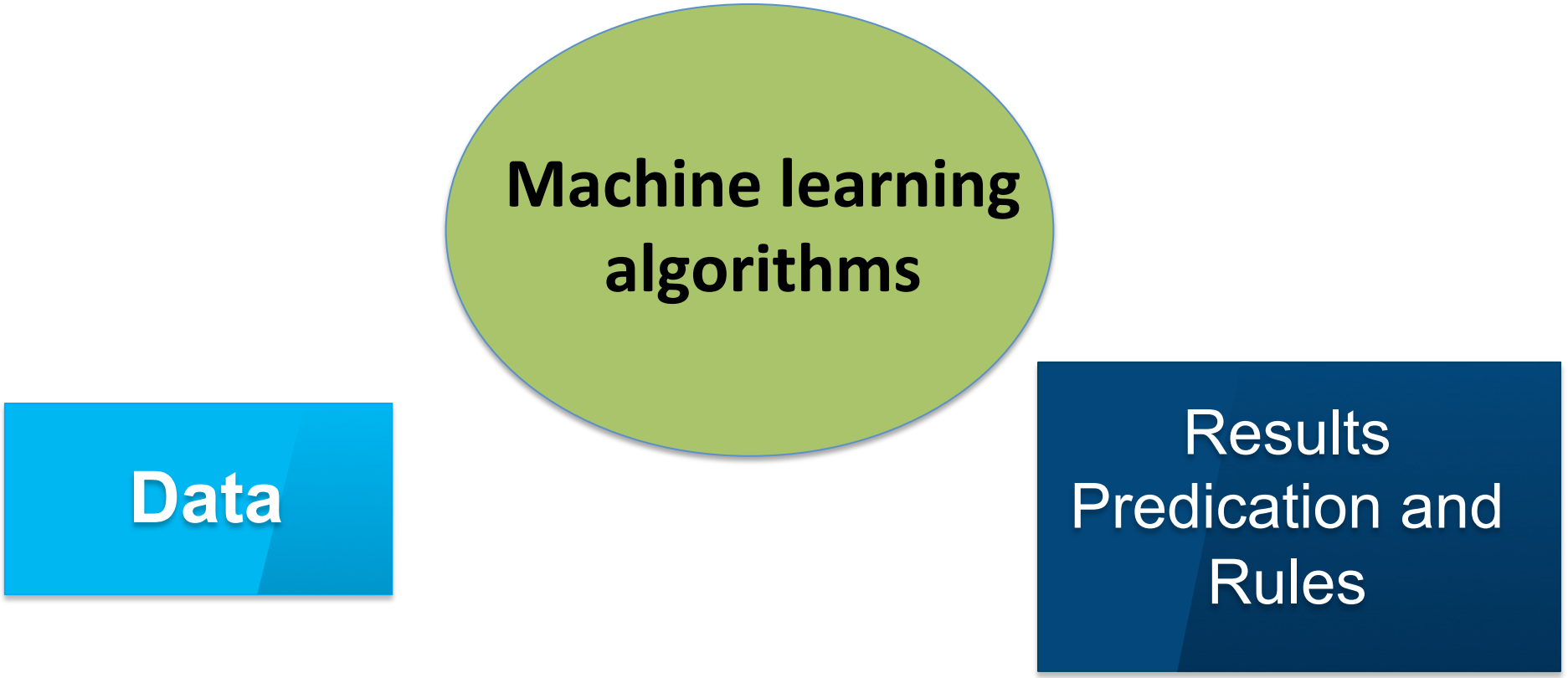


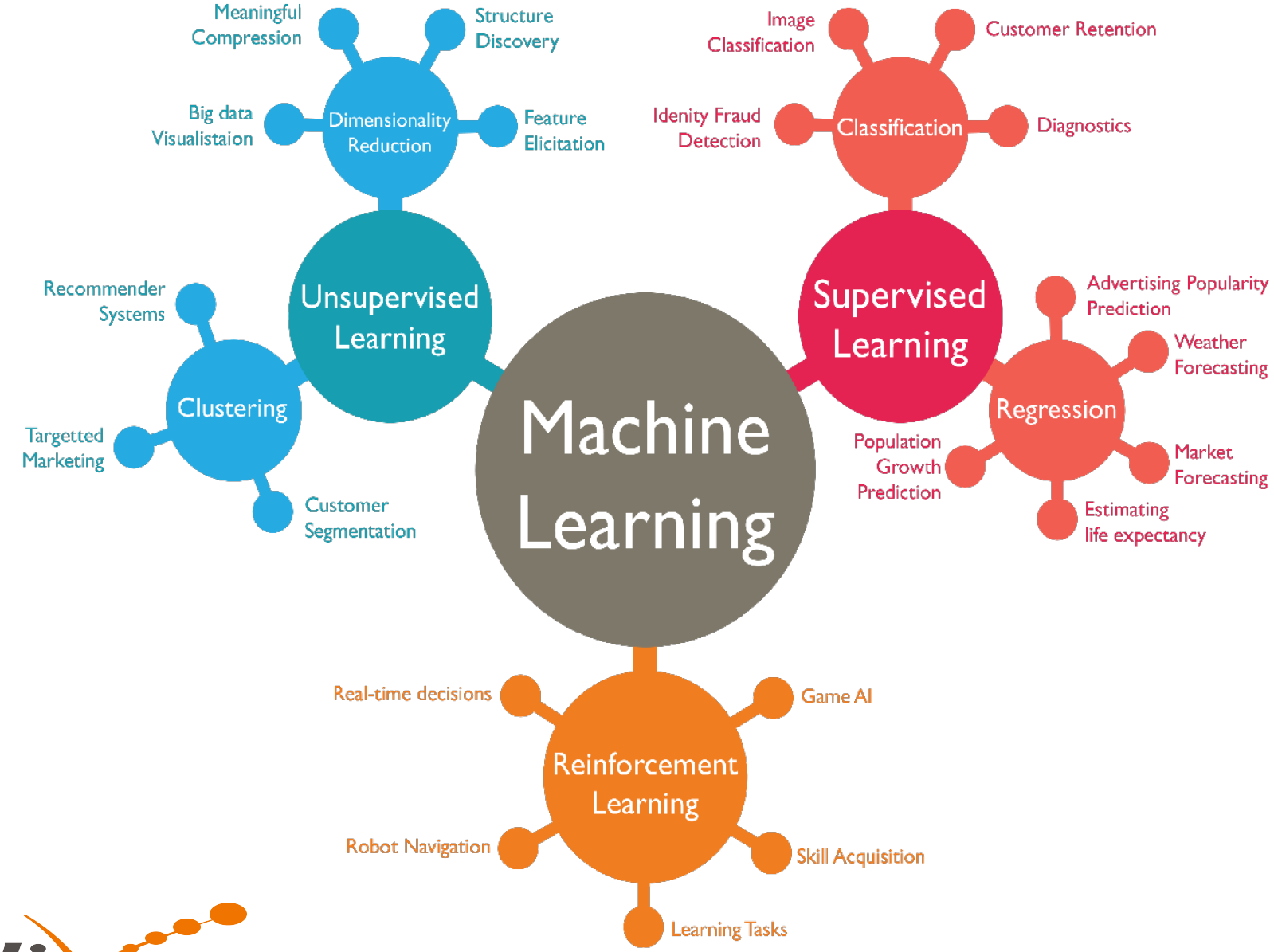
Computing

+

Storage

Artificial intelligence




Machine Learning
Unsupervised Learning
Meaningful Compression
Structure Discovery
Dimensionality Reduction
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Image Classification
Customer Retention
Classification
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

How do Machines learn?

Data to model

Decision

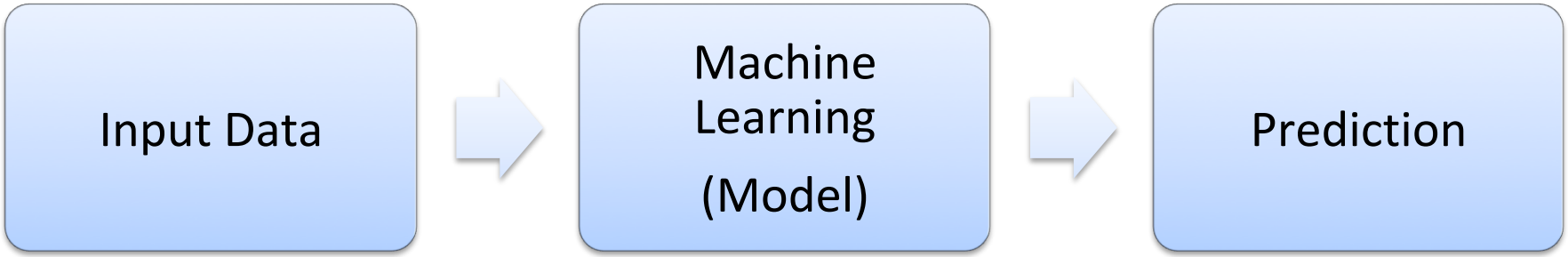
Create models

Evaluate models

Refine models

Prediction, categorization

Introduction Machine Learning[1]



- 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.

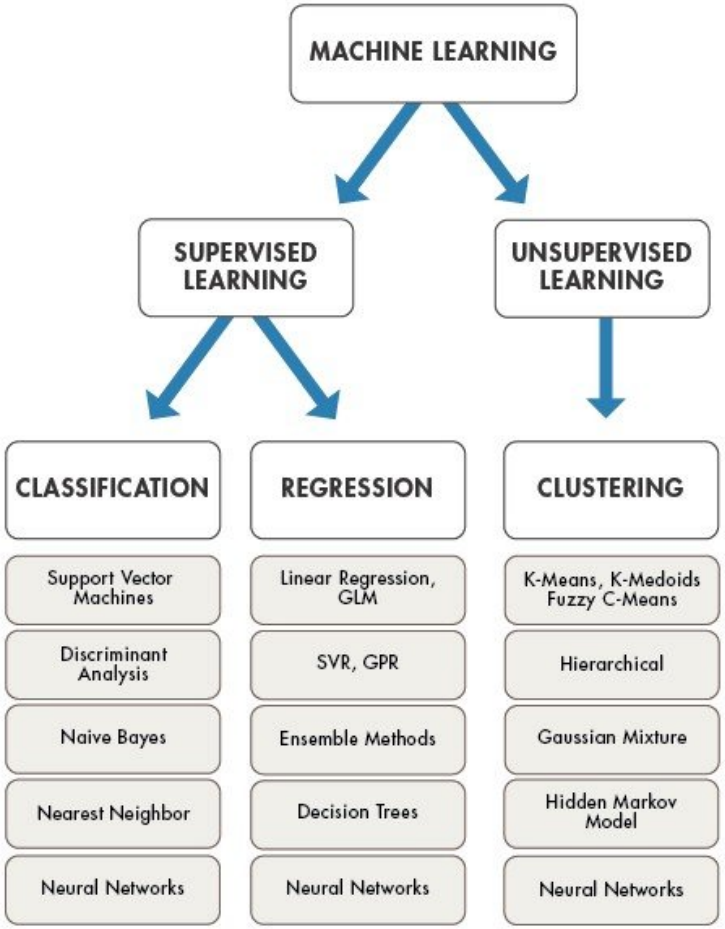
Introduction Machine Learning[2]

- For example 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.

Introduction to Machine Learning[3]

- 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 sequences whose genes have been previously discovered and their locations annotated.
  - 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



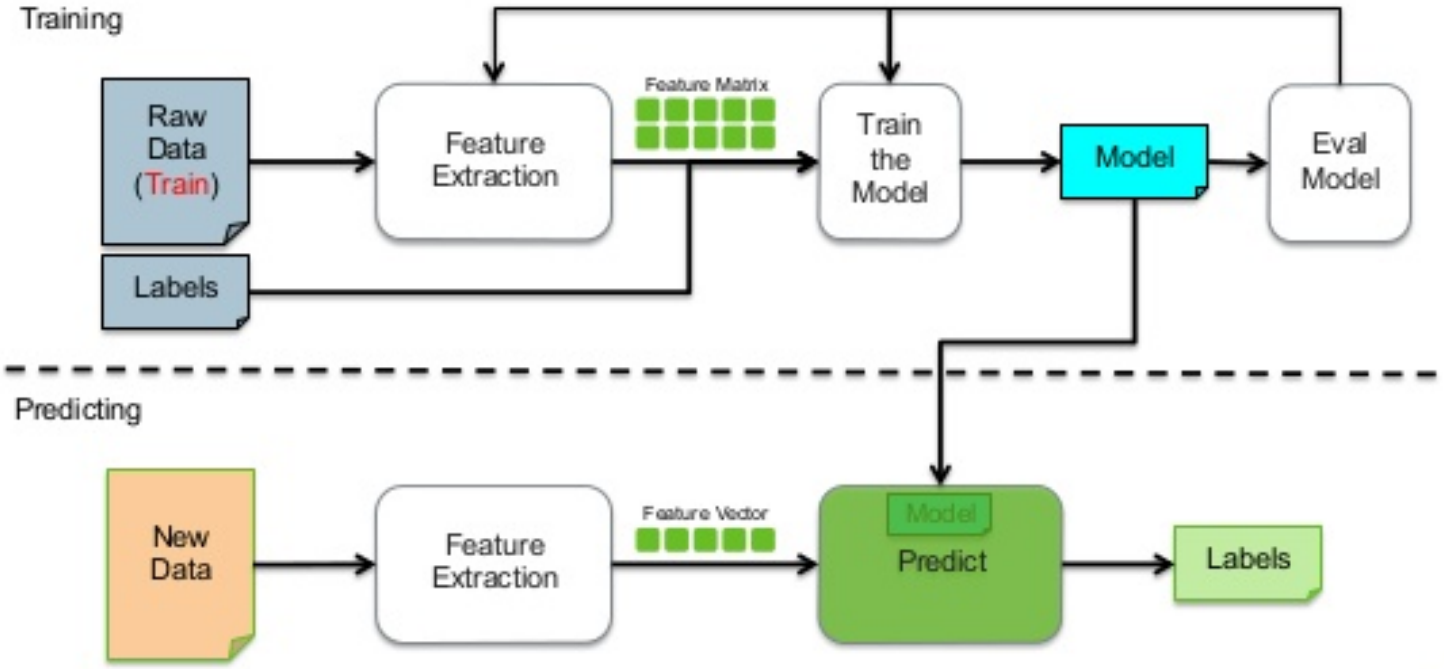
The ML taxonomy

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

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

Supervised Machine Learning

Supervised Learning Workflow



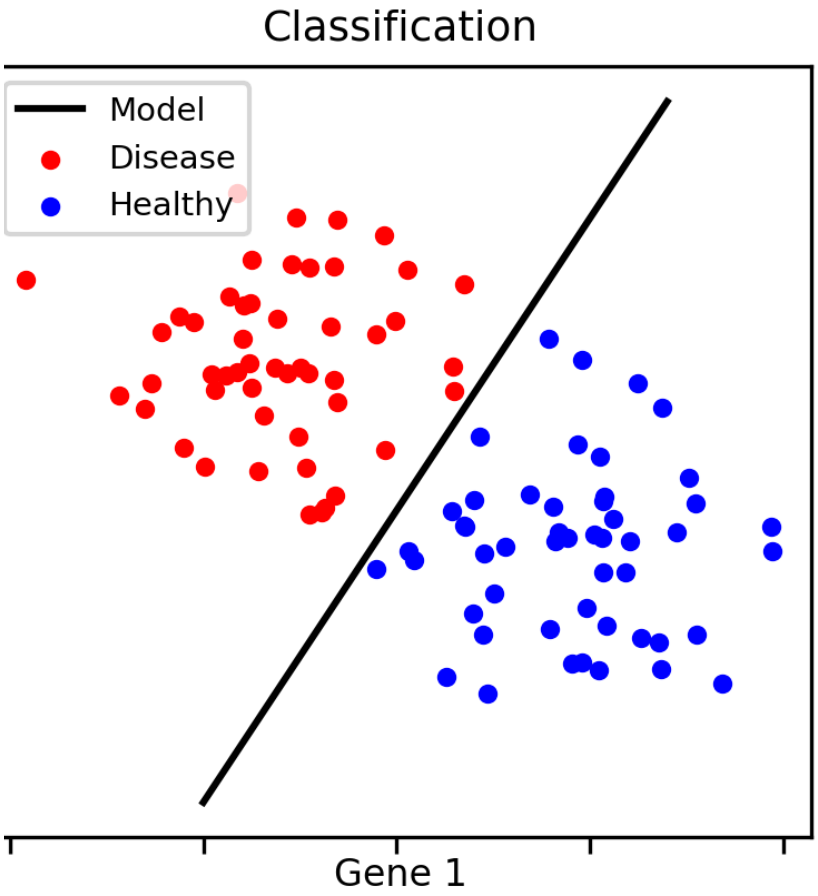
Hortonworks

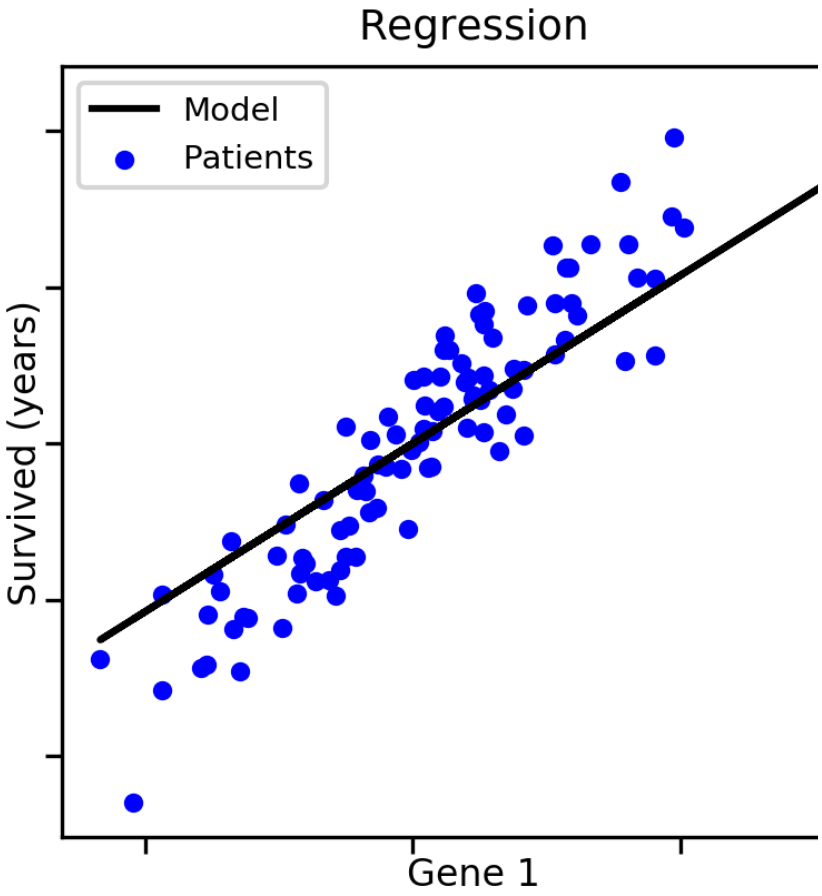
H3ABioNet

Supervised Machine Learning Algorithms[1]

- Apply what has been learned in the past to new data using labeled examples to predict future events.
- Starting from the analysis of a known training dataset, the learning algorithm produces a prediction model that can provide targets for any new input (after sufficient training).
- The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify and improve the prediction model accordingly.

Classification vs regression





Classification vs regression

Classification	Regression
Discreate, categorical variable	Continous (real number range)
Supervised classification problem	Supervised classification problem
Assign the output to a class (a label)	Predict the output value using training data
Predict the type of tumor (harmful vs not harmful)	Predict a house price, predict survival time

Validation of supervised ML algorithms results

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

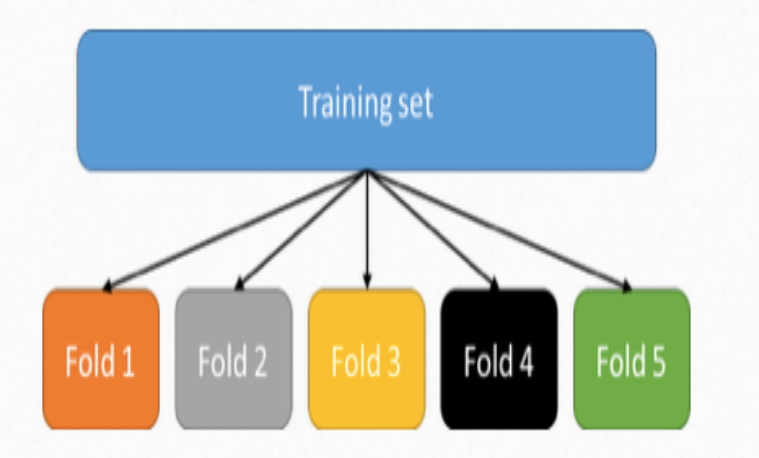
Training set and test set

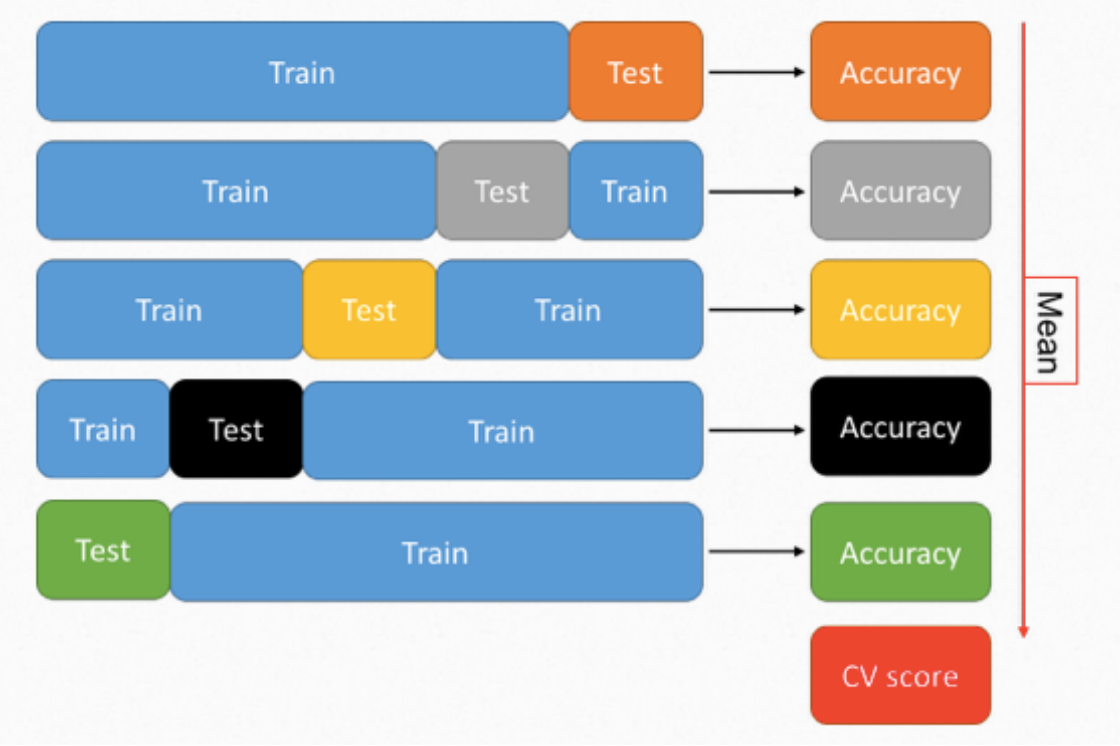
Used to train the algorithm

Estimate the accuracy of the model

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

K-fold cross validation





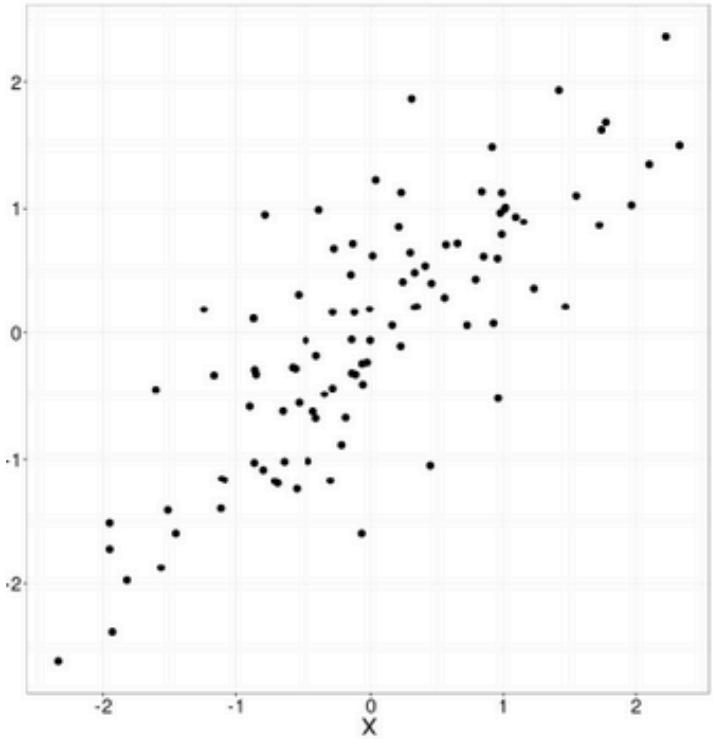
https://aldro61.github.io/microbiome-summer-school-2017/sections/basics/#type-of-learning-problems

Examples of supervised learning algorithms

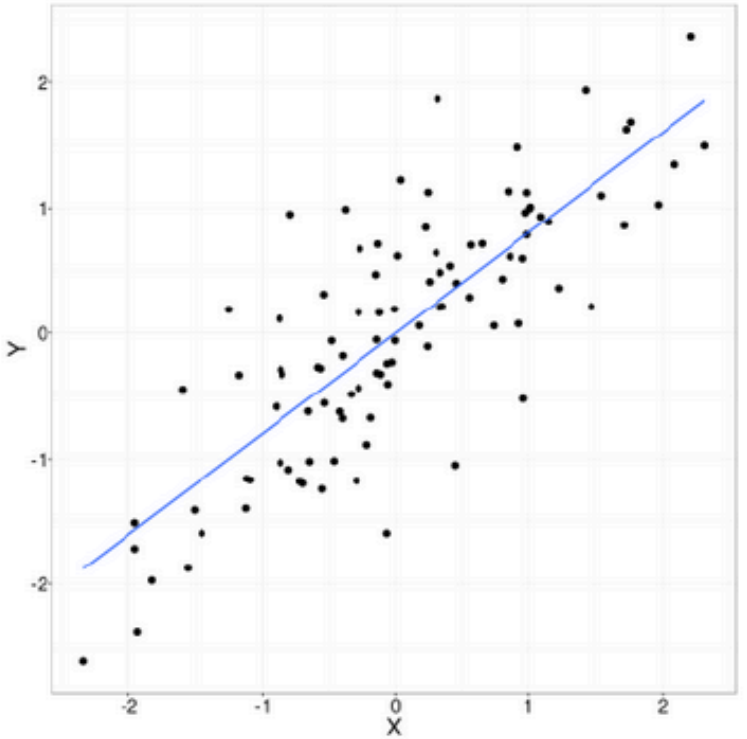
Linear regression

- Regression algorithms can be used for example when some continuous value needs to be computed as compared to classification where the output is categoical.
- 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 are:
  - Using gene expression data to classify (or predict) tumor types using gene expression data

Applying linear regression



Scatterplot of our dataset.



Fitting of the regression line (blue).

Decision Trees (Supervised)

- 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.
- Decision trees easily handle feature interactions and they are non-parametric, so there is no need to worry about outliers or whether the data is linearly separable.
- Disadvantages are:
  - Often the tree needs to be rebuilt when new examples come on.
  - Decision trees easily overfit, but ensemble methods like random forests (or boosted trees) take care of this problem.
  - They can also take a lot of memory (the more features you have, the deeper and larger your decision tree is likely to be)
- Trees are excellent tools for helping to choose between several courses of action.
- Example: Classification of genomic islands using decision trees and ensemble algorithms

Random Forest (Supervised)

- Random Forest is an ensemble of decision trees.
- It can solve both regression and classification problems with large data sets.
- It also helps identify most significant variables from thousands of input variables.
- Random Forest is highly scalable to any number of dimensions and has generally quite acceptable performances.
- However with Random Forest, learning may be slow (depending on the parameterization) and it is not possible to iteratively improve the generated models
- Random Forest can be used in real-world applications such as:
  - Predict patients for high risks for certain diseases

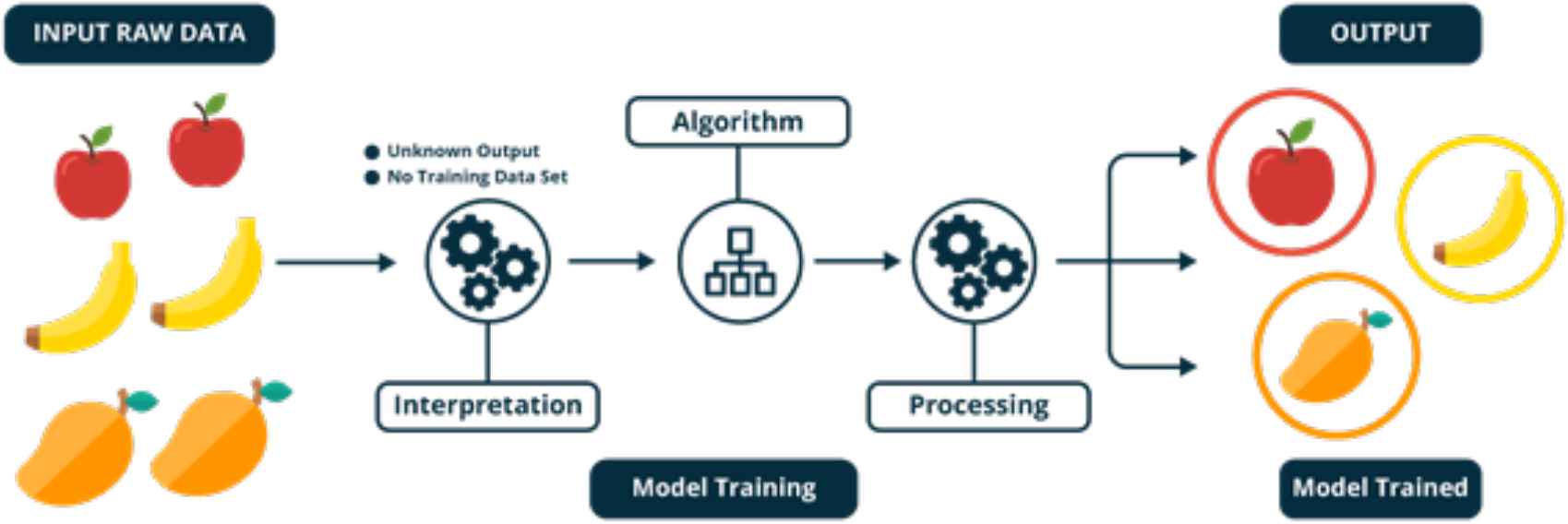
Support Vector Machines (Supervised)

- Support Vector Machine (SVM) is a supervised machine learning technique that is widely used in pattern recognition and classification problems — when your data has exactly two classes.
- Advantages include high accuracy and even if the data is not linearly separable in the base feature space, SVM can work well with an appropriate kernel.
- However SVMs are memory-intensive, hard to interpret, and difficult to tune.
- SVM is especially popular in text classification problems where very high-dimensional spaces are the norm.
- SVM can be used in real-world bioinformatics applications such as:
  - detecting persons with common diseases such as diabetes
  - Classification of genomic islands

Naive Bayes (Supervised)

- It is a classification technique based on Bayes' theorem.
- Advantages include:
  - very easy to build and particularly useful for very large data sets.
  - outperform even highly sophisticated classification methods.
  - a good choice when CPU and memory resources are a limiting factor.
  - A good method if something fast and easy that performs pretty well is needed.
- Its main disadvantage is that it does not consider the interactions between features.
- Naive Bayes can be used in real-world applications such as:
  - Mining housekeeping genes
  - genetic association studies
  - discovering Alzheimer genetic biomarkers from whole genome sequencing (WGS) data

Unsupervised Learning



https://www.quora.com/What-is-the-difference-between-supervised-and-unsupervised-learning-algorithms

Unsupervised Machine Learning Algorithms[1]

- In contrast to supervised machine learning algorithms, they:
  - are applied when the information used to train is ***neither classified nor labeled***.
  - can infer a function to describe a hidden structure from unlabeled data.
  - do not figure out the right output, but explore the data and can draw inferences from datasets to describe hidden structures from unlabeled data.
- The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data.
  - Algorithms are left to their own devises to discover and present the interesting structure in the input data.

Unsupervised Machine Learning Algorithms[2]

- Unsupervised learning problems can be further grouped into clustering, association and dimensionality reduction problems:
  - **Clustering**: A clustering problem is where you want to discover the inherent groupings in the data, such as clustering DNA sequences into functional groups.
  - **Association**: An association rule learning problem is where you want to discover rules that describe large portions of your data, such as using association analysis-based techniques for pre-processing protein interaction networks for the task of protein function prediction.
  - **Dimensionality Reduction**: Often we are working with data of high dimensionality—each observation comes with a high number of measurements—a dimension reduction procedure is usually conducted to reduce the variable space before the subsequent analysis is carried out..
    - For example in a gene-expression analysis, dimension reduction can be used to find a list of candidate genes with a more operable length ideally including all the relevant genes.
    - Leaving many uninformative genes in the analysis can lead to biased estimates and reduced power.

Unsupervised Machine Learning Algorithms[3]

- **Clustering** is an exploratory data analysis technique that allows us to organize a pile of information into meaningful subgroups (clusters) without having any prior knowledge of their group memberships.
- Each cluster that arises during the analysis
  - defines a group of objects that share a certain degree of similarity but are more dissimilar to objects in other clusters, which is why clustering is also sometimes called **unsupervised classification.**

Unsupervised Machine Learning Algorithms[5]

- **Taking the example** of the gene-finding model, when a labeled training set is not available, unsupervised learning is required.
- Consider the interpretation of a heterogeneous collection of epigenomic data sets, such as those generated by the Encyclopedia of DNA Elements (ENCODE) Consortium and the Roadmap Epigenomics Project.

Unsupervised Machine Learning Algorithms[6]

- A priori, we expect that the patterns of chromatin accessibility, histone modifications and transcription factor binding along the genome should be able to provide a detailed picture of the biochemical and functional activity of the genome.
  - We may also expect that these activities could be accurately summarized using a fairly small set of labels.
- To discover what types of label best explain the data, rather than imposing a pre-determined set of labels on the data, unsupervised learning method can be applied.

Unsupervised Machine Learning Algorithms[7]

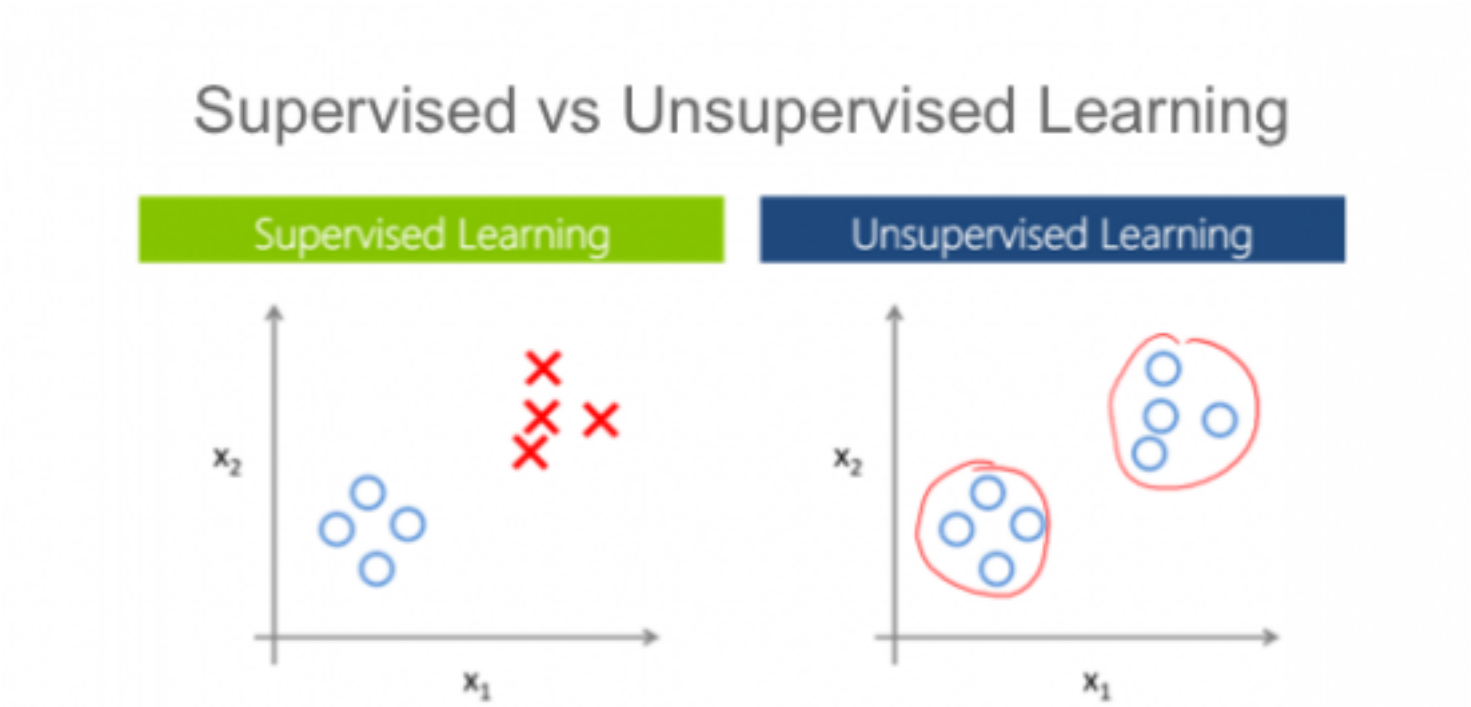
  - It will use only the unlabeled data and the desired number of different labels to assign as input to automatically partition the genome into segments and assign a label to each segment, with the goal of assigning the same label to segments that have similar data.

- The unsupervised approach requires an additional step in which semantics must be manually assigned to each label, but it provides the benefits of enabling training when labeled examples are unavailable and has the ability to identify potentially novel types of genomic elements.

Supervised vs unsupervised learning

Supervised	Unsupervised
Input data is labelled	Input data is unlabelled
Uses training dataset	Uses just input dataset
Known number of classes	Unkown 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

Supervised vs unsupervised Learning



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

Clustering validation

Assess the quality of a clustering algorithm results is important to avoid finding random patterns in data.

- Internal cluster validation: uses only internal information to the clustering process without reference to external information (clusters separability, clusters homogenity, etc.)
- Clusters should be well-separated and intra-cluster distance should be small
  - Silhouette coefficient: estimates the average distance between clusters
  - Dunn index: estimates distances between objects in the same cluster vs objects in different clusters (should be maximized)
- External cluster validation: compares results to an externally known results (example: class labels), compare different clustering methods results, etc.
- Relative cluster validation: evaluates the clustering methods by varying different parameters values for the same algorithm (example:number of clusters)

Examples of unsupervised learning algorithms

Neural Networks (Can be supervised or unsupervised)

- Neural Networks take in the weights of connections between neurons. When all weights are trained, the neural network can be utilized to predict the class or a quantity.
- With Neural networks, extremely complex models can be trained and they can be utilized as a kind of black box.
- Disadvantages:
  - parameterization is extremely difficult in neural networks.
  - They are also very resource and memory intensive.
- NN can be joined with the “deep approach” to build models that can pick previously unpredictable cases.
- They may be applied for classification, predictive modelling and biomarker identification within data sets of high complexity such as transcript or gene expression data generated from DNA microarray analysis, or peptide/protein level data generated by mass spectrometry.

Principal Component Analysis (PCA) (Unsupervised)

- PCA provides dimensionality reduction.
- Sometimes you have a wide range of features, probably highly correlated between each other, and models can easily overfit on a huge amount of data. Then PCA can be applied.
- Advantage:
  - in addition to the low-dimensional sample representation, it provides a synchronized low-dimensional representation of the variables. The synchronized sample and variable representations provide a way to visually find variables that are characteristic of a group of samples.
- PCA can be used in bioinformatics to:
  - Analyse gene expression data

K-Means (Unsupervised)

- The goal of k-means is to find groups in the data, with the number of groups represented by the variable *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.
- Example: popular and simple partition computational models for clustering microarray data

Semi-supervised Machine Learning Algorithms[1]

- 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.

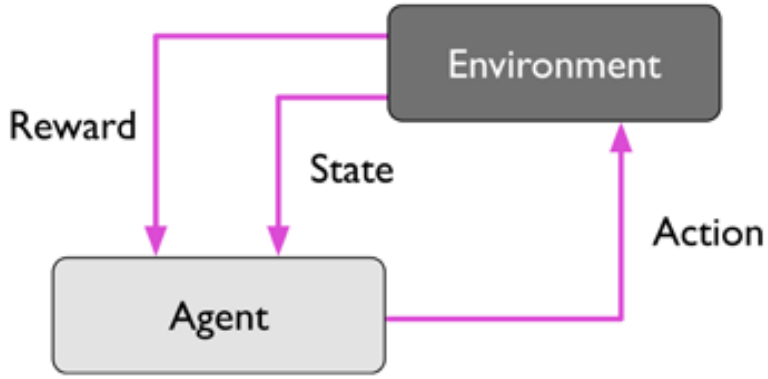
Semi-supervised Machine Learning Algorithms[2]

- 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[3]

- 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[1]



- 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[2]

- 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[1]

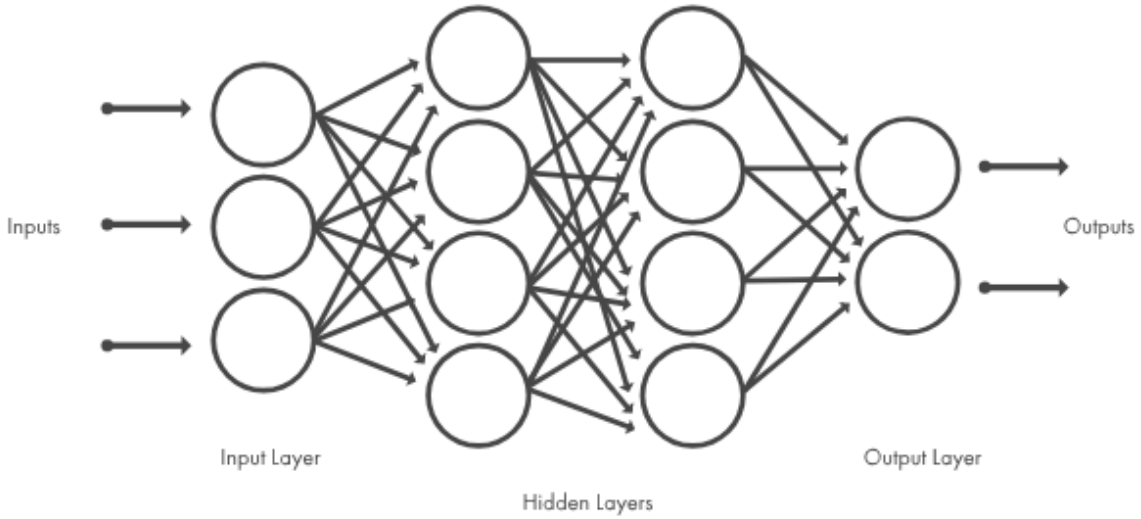
- Also known as deep structured learning or hierarchical learning
- It is a subfield of machine learning concerned with algorithms inspired by the structure and function of the brain called artificial neural networks.
- Can perform learning in supervised and/or unsupervised manners.
- Teach computers 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.
  - **Used 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[2]

- 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[3]

- 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.



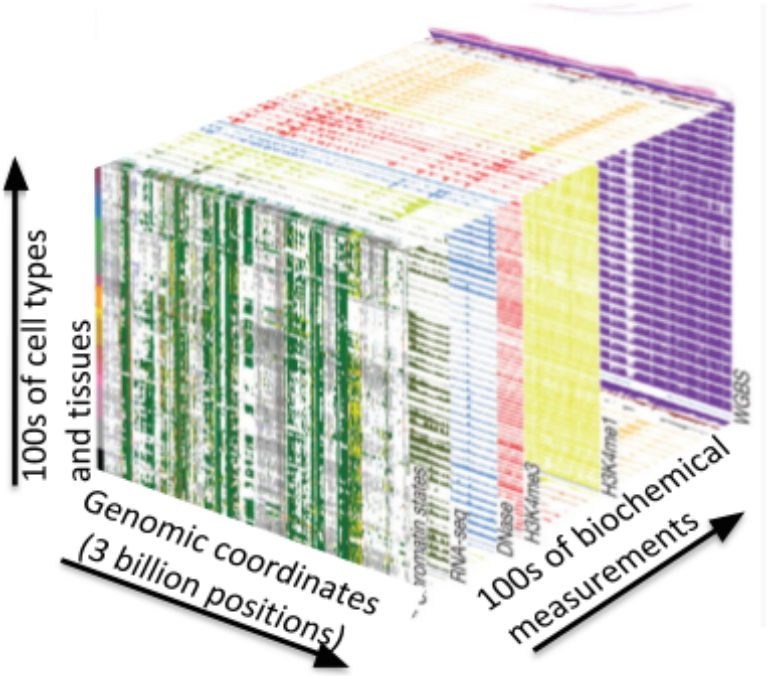
Deep Learning Algorithms[4]

- 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[5]

- 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

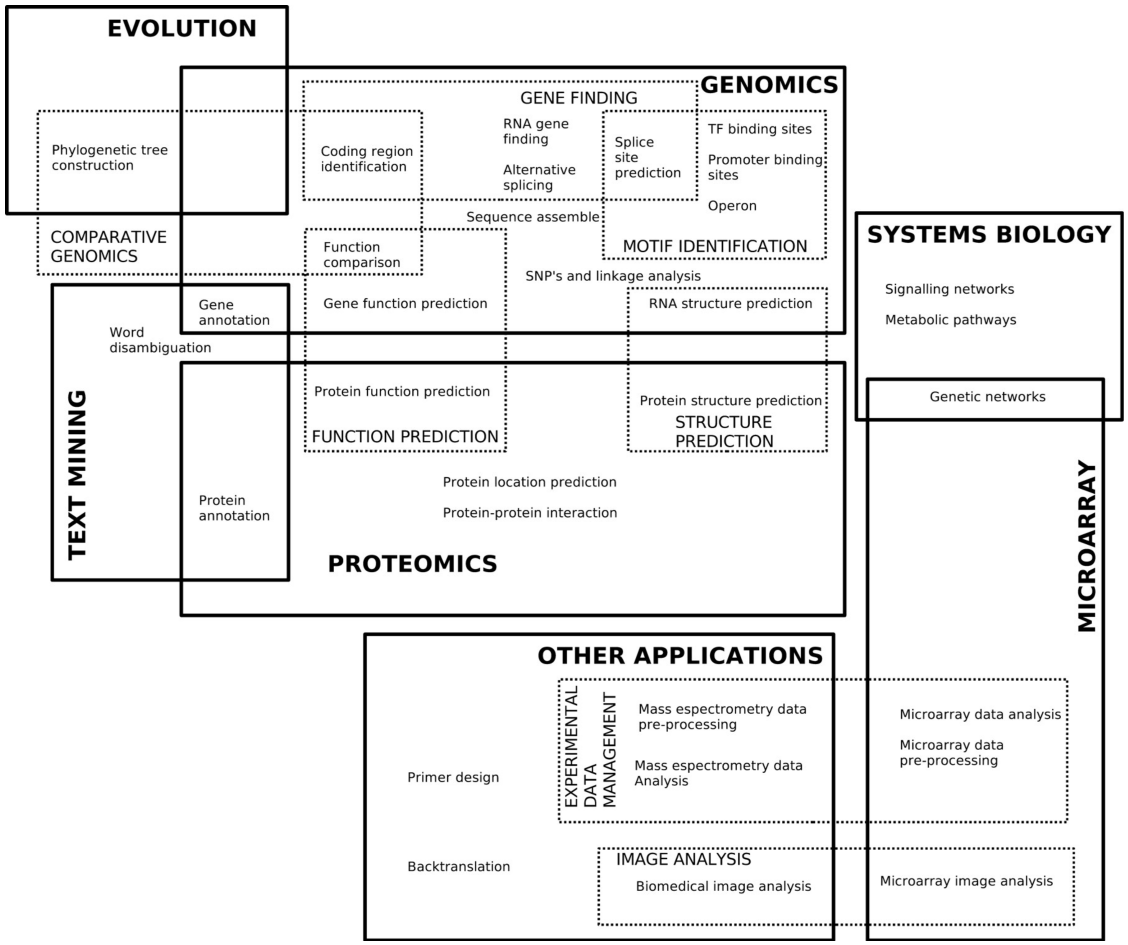


Deep Learning

- Functional annotation of the human genome → Acceleration of Drug Discovery
- Prediction of disease-associated mutations → Disease Diagnosis
- Liquid biopsy decomposition → Early cancer Detection

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

Applications of ML in Bioinformatics



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

Some examples of use tools used for ML in bioinformatics

Problem at hand	Data	Method/s
Identification of biomarkers	Proteomics datasets Transcriptomics datasets	BioHEL – rule-based learning method (Swan et al, 2015)
Cancer Classification from Microarray Gene Expression Data	Transcriptomics data	J4.8 decision tree from Weka Naïve Bayes, SVM (Peng et al, 2007)
Inference of demographic history and recombination rates in population genetics	Population genomic datasets	Artificial Neural Network (ANN) (Blum et al, 2010)
Showing how relationships among individuals sampled from Europe largely mirrored geography	Population genetics	PCA (November et al, 2008)
Uncover differences in evolutionary rates along a chromosome	Phylogenetic Data	Hidden Markov Model(HMM) (Schrider et al, 2018)
Quantify the ability of TF-binding signals to statistically predict the expression levels of promoters.	Cell-line-specific TRF binding data	Random Forest, Support Vector Regression(SVR), multivariate adaptive regression splines (MARS)
Genomic Selection in Breeding Wheat for Rust Resistance	Genomic Selection data	Reproducing kernel Hilbert space, Bayesian LASSO, random forest regression, Support vector

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 labelled, it is a supervised learning problem.
- However, even when labels are available, it is not always the case that taking a supervised approach is a good idea (size and quality of training and test sets).
- Ingeneral, supervised learning should be employed only when the training set and test set are expected to exhibit similar statistical properties

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 excpected 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

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)