# INTRODUCTION TO MACHINE LEARNING

## Challenges, Trends and Solutions in Life Sciences

Wandrille Duchemin
*Swiss Institute of bioinformatics*
University of Basel CH

Crhistian Cardona
*University of Tübingen DE*
*Dundee University UK*

# 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

**Pedro L. Fernandes**

- Bioinformatics Training Coordinator
- Instituto Gulbenkian de Ciência

**Wandrille Duchemin**

- Bioinformatics Trainer & Computational Biologist
- SIB Swiss Institute of Bioinformatics
- University of Basel

**Crhistian Cardona**

- Ph.D. Researcher
- Universität Tübingen/ University of Dundee

# COMMUNICATION

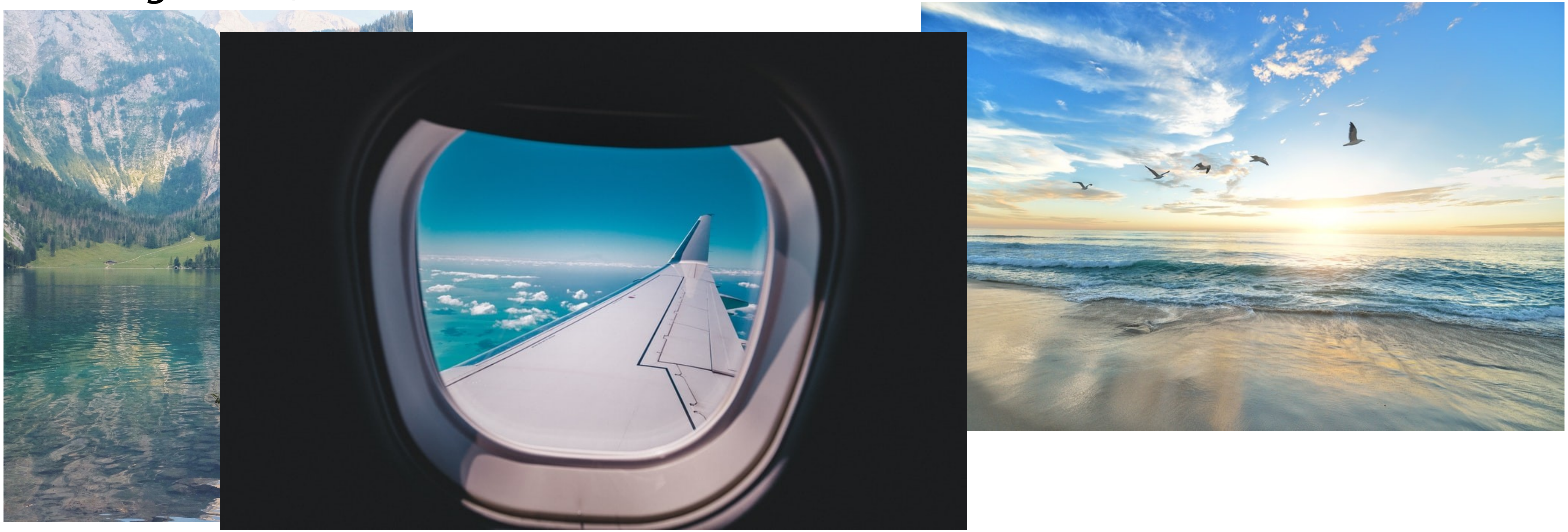
We will be using GDoc to exchange information

Get the material from the github repo:

https://github.com/BiodataAnalysisGroup/2021-11-ml-elixir-pt

# AN ICEBREAKER

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



# COURSE AGENDA

**Hours:**

- 9:30 to 18:30
- Breaks : 10:30-11:00 & 15:00-15:30
- Lunch : 12:30-14:00

**Day 1 :**

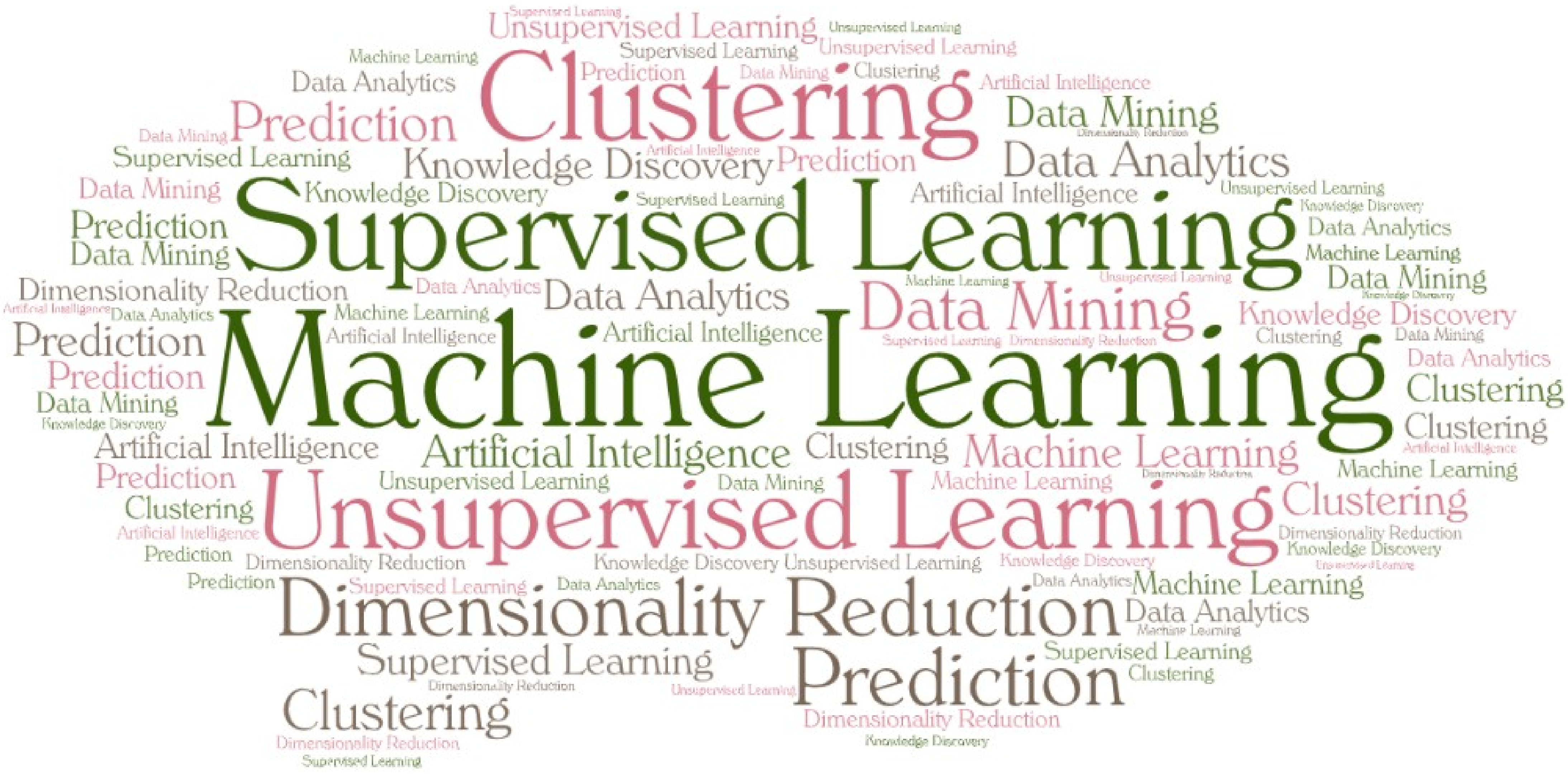
- Welcome
- Exploratory Data Analysis
- Unsupervised Learning

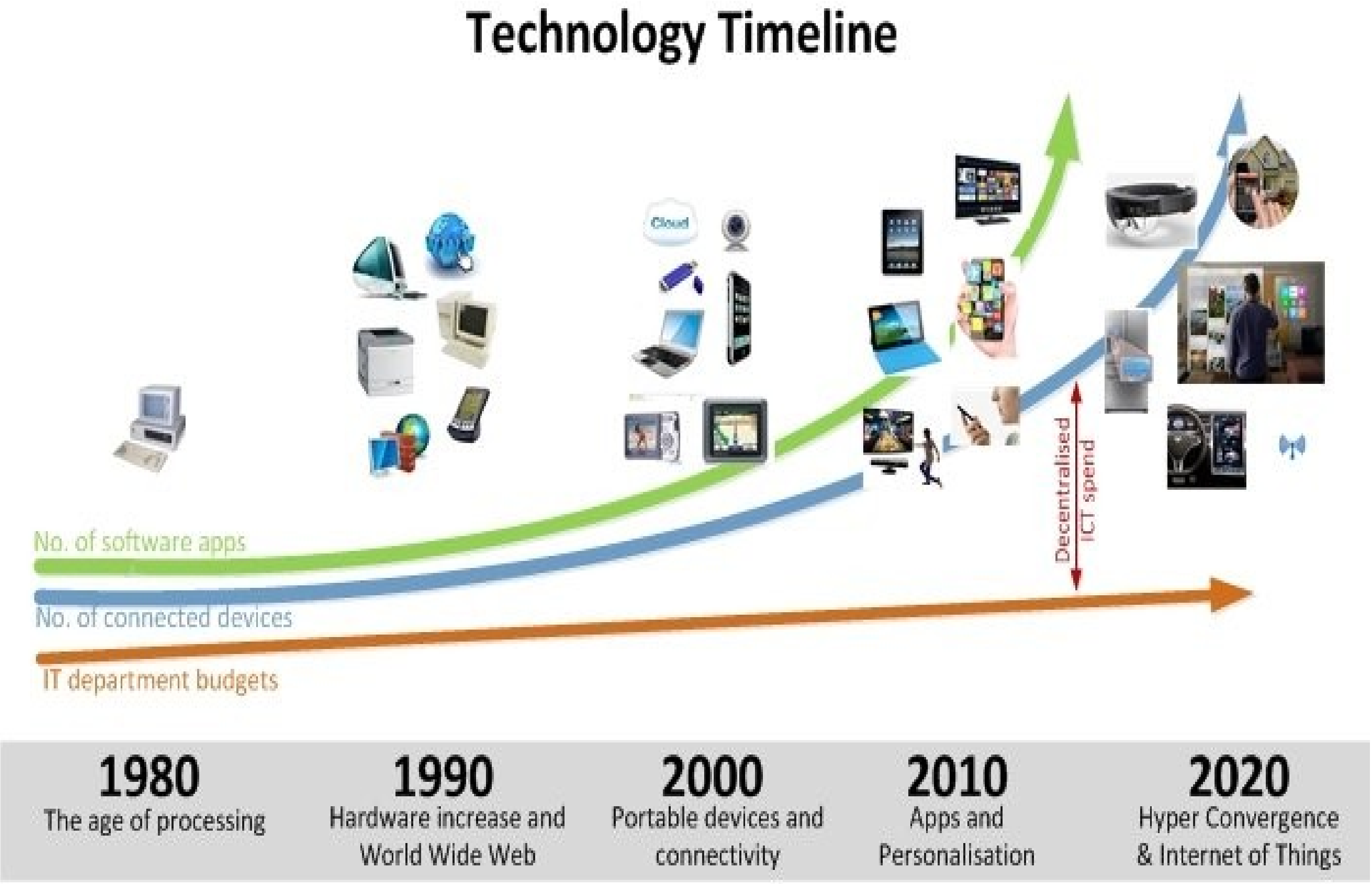
**Day 2 :**

- Unsupervised Learning (continued)
- Supervised Learning – classification

**Day 3 :**

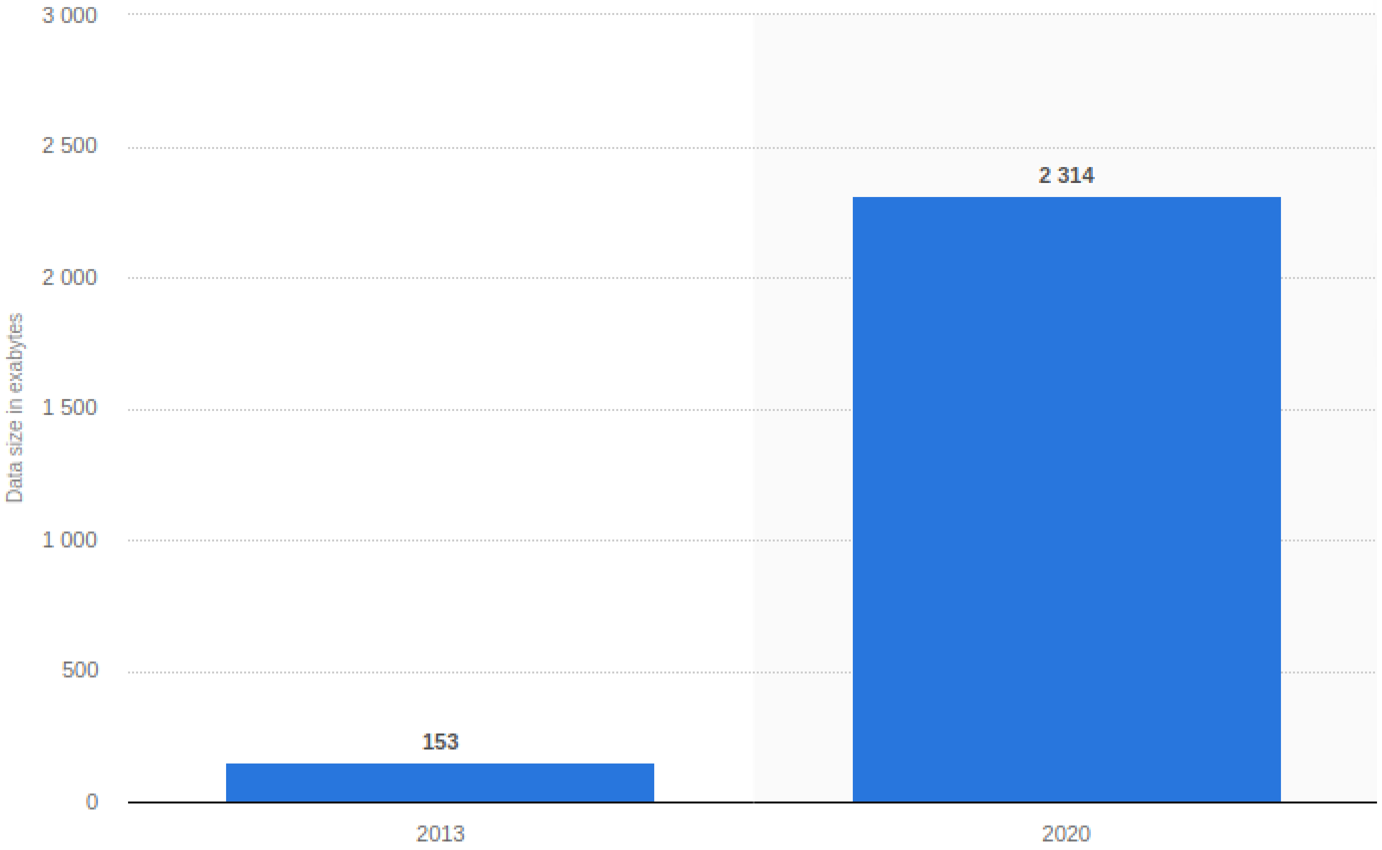
- Classification (continued)
- Supervised Learning -regression
- Discussion


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



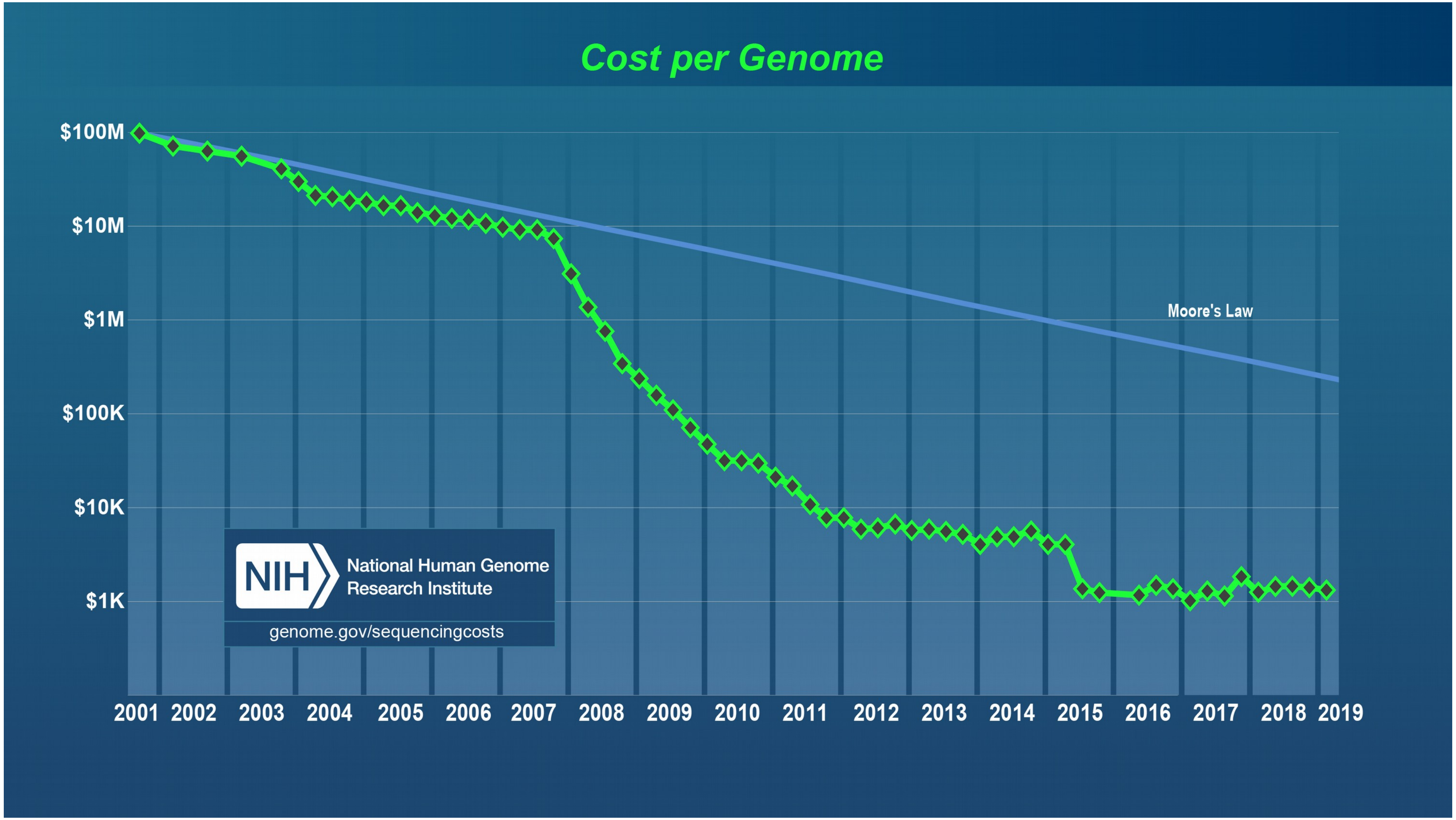
https://www.linkedin.com/pulse/technology-increase-vs-department-budgets-sam-errington/

# 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

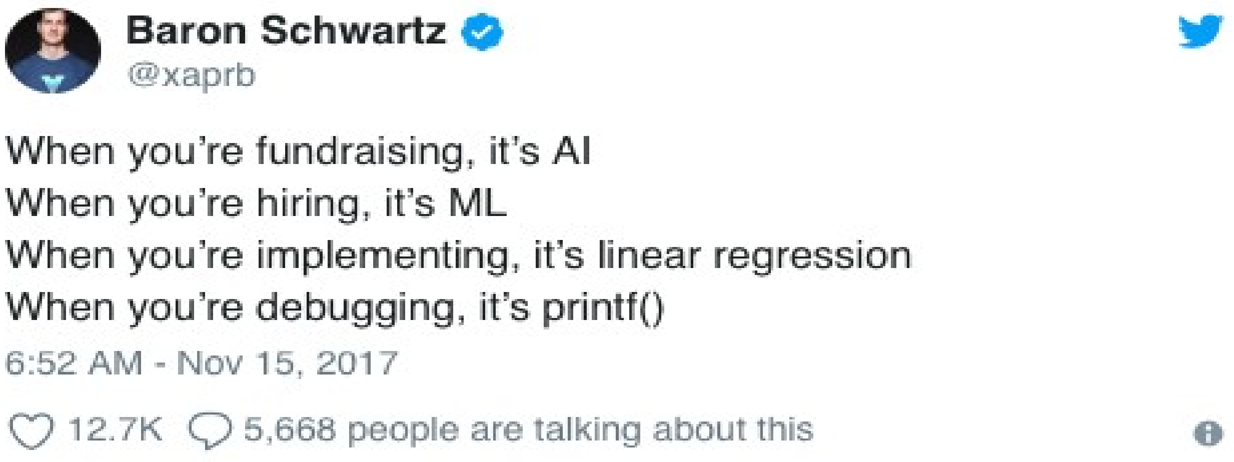
ML is more about building the prediction model

# 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

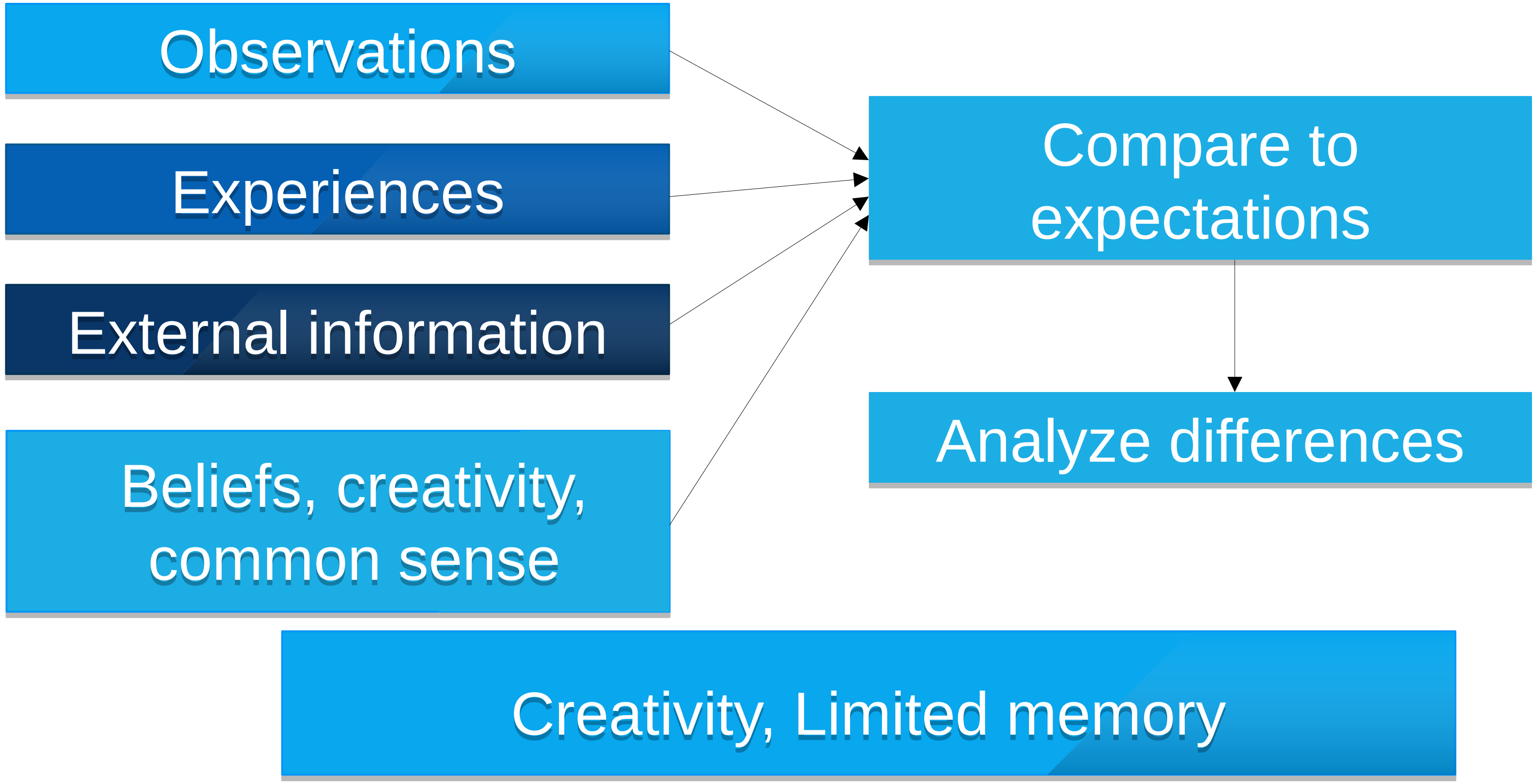
Machine learning algorithms are driven by the data used

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

Simulate human behavior and cognitive process

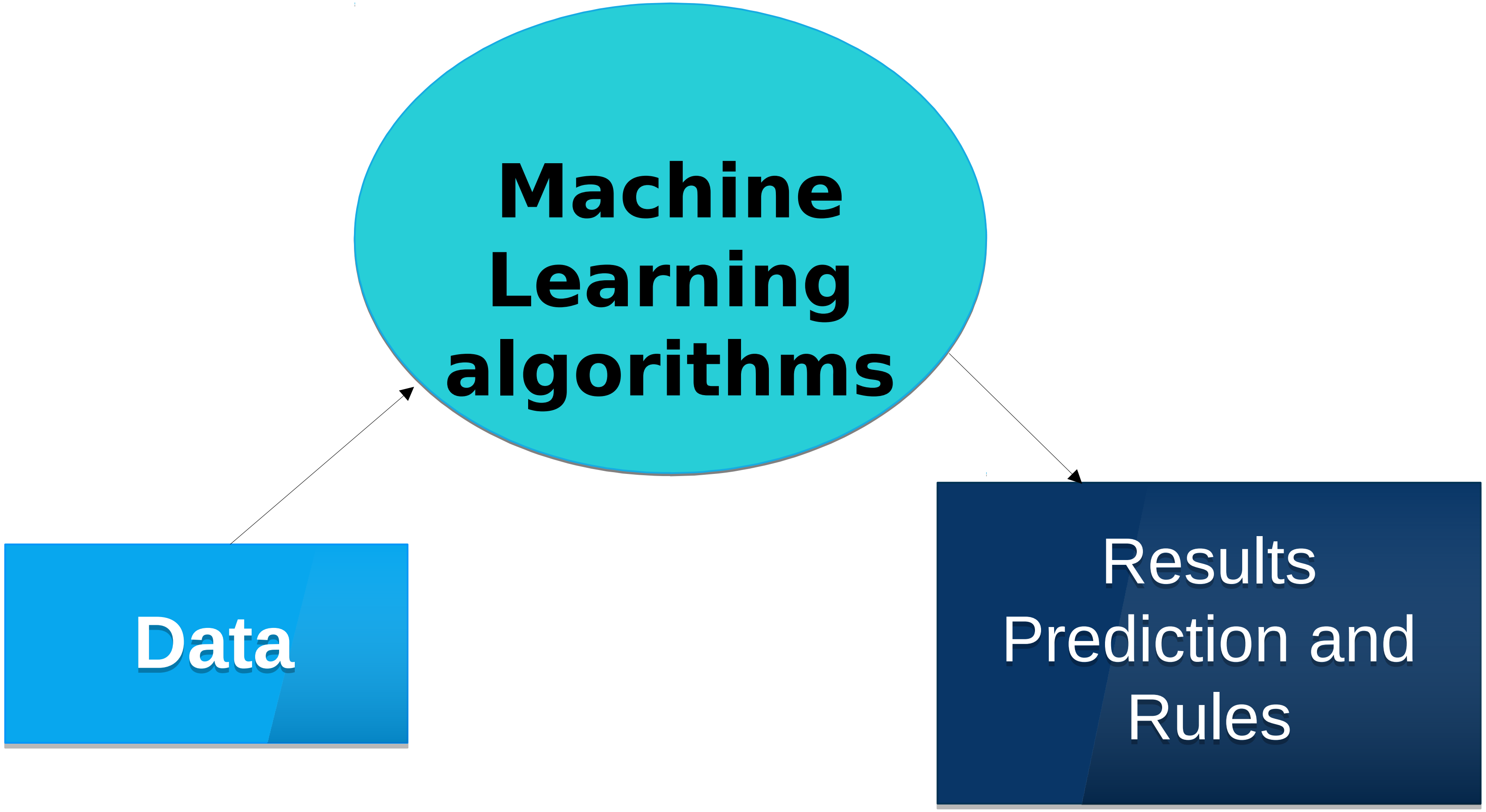
Capture and preserve human expertise

Fast response
Ability to memorize big amounts of data

**Data**

**Computing**
**+**
**Storage**

# ARTIFICIAL INTELLIGENCE

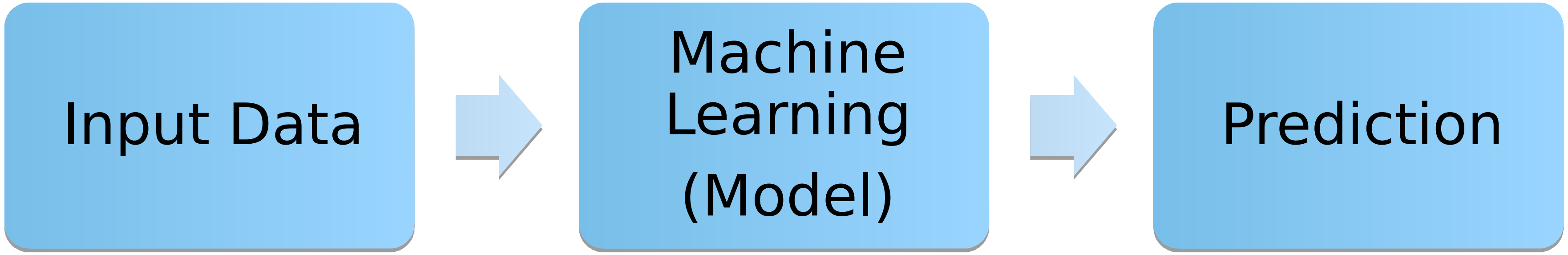


Machine Learning algorithms

Data

Results Prediction and Rules

# 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.

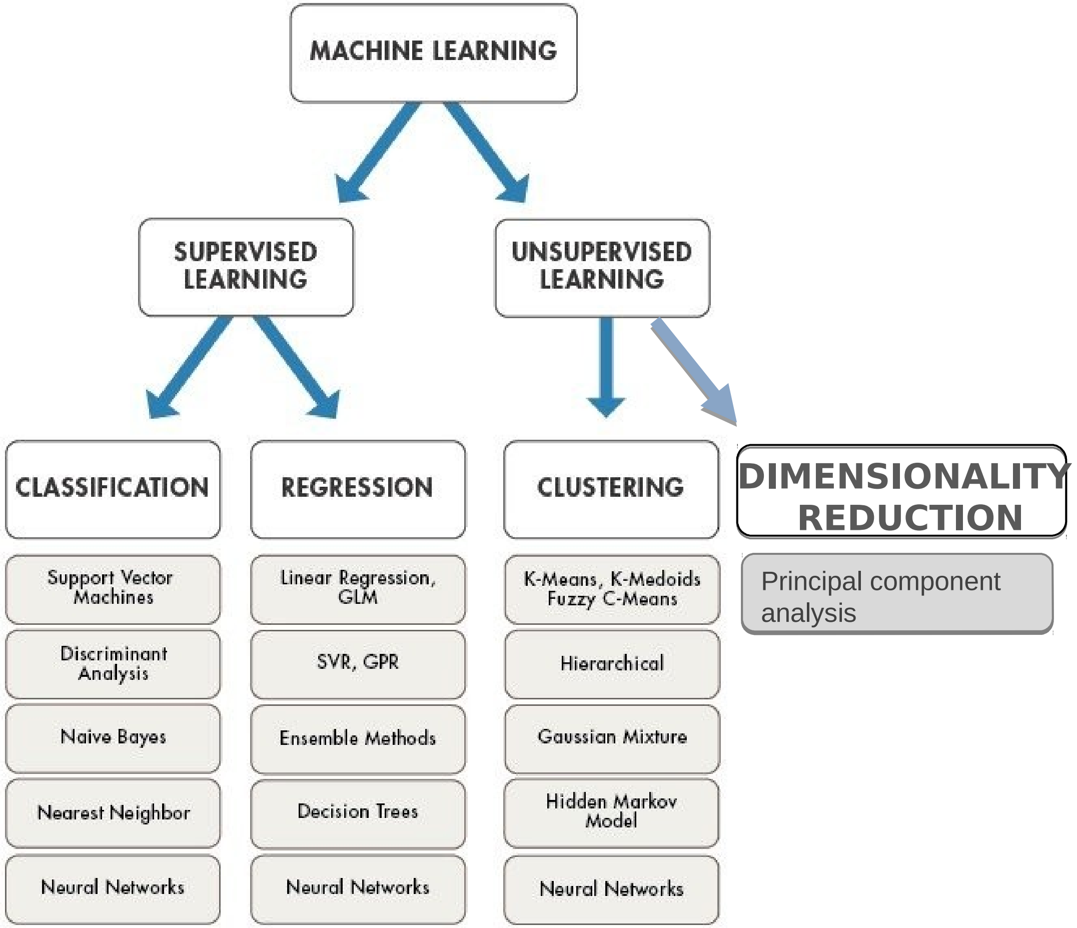
More info about this task can be obtained from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5204302/)

# 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



MACHINE LEARNING

SUPERVISED LEARNING

UNSUPERVISED LEARNING

CLASSIFICATION

REGRESSION

CLUSTERING

DIMENSIONALITY REDUCTION

Support Vector Machines

Discriminant Analysis

Naive Bayes

Nearest Neighbor

Neural Networks

Linear Regression, GLM

SVR, GPR

Ensemble Methods

Decision Trees

Neural Networks

K-Means, K-Medoids Fuzzy C-Means

Hierarchical

Gaussian Mixture

Hidden Markov Model

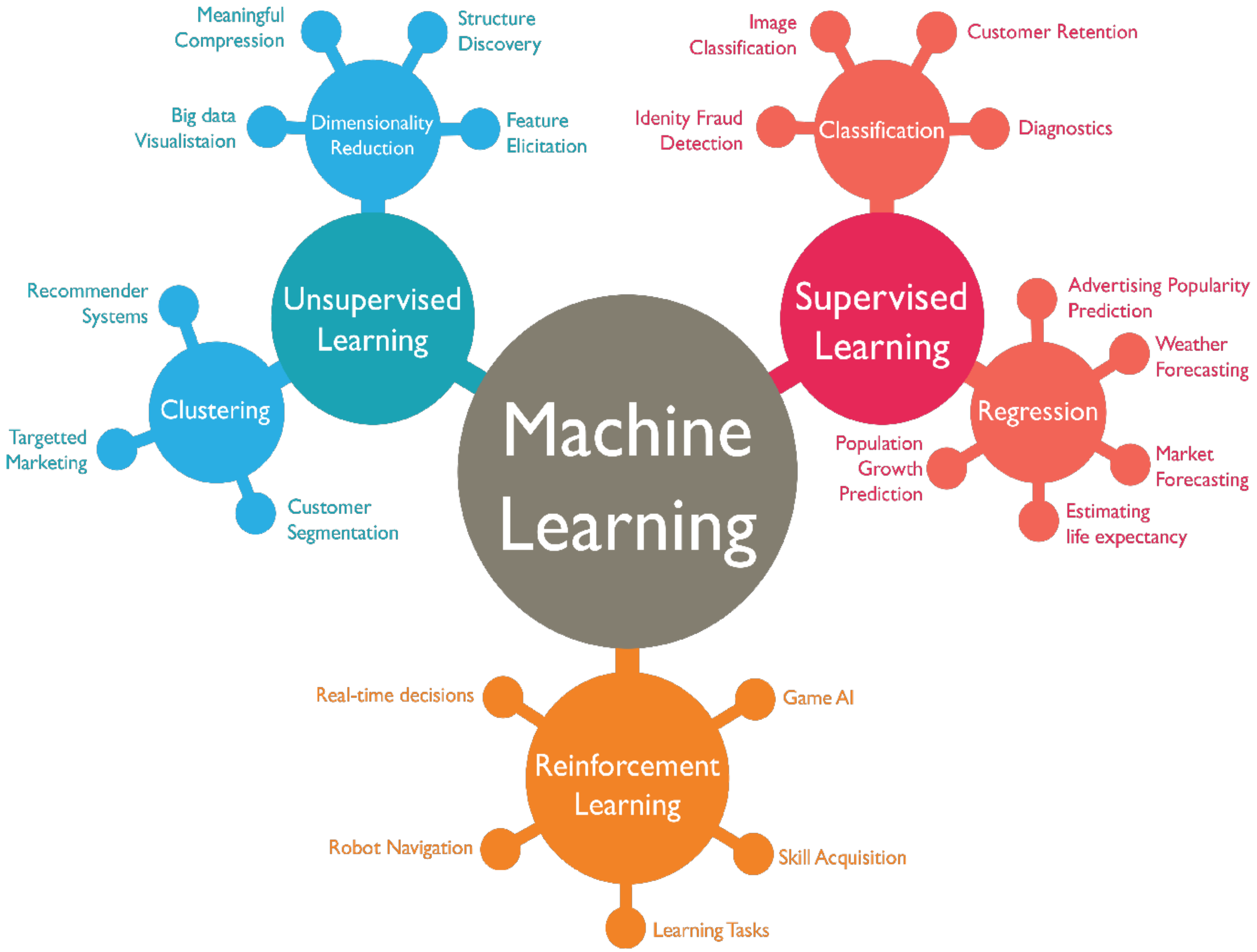
Neural Networks

Principal component analysis

# THE ML TAXONOMY

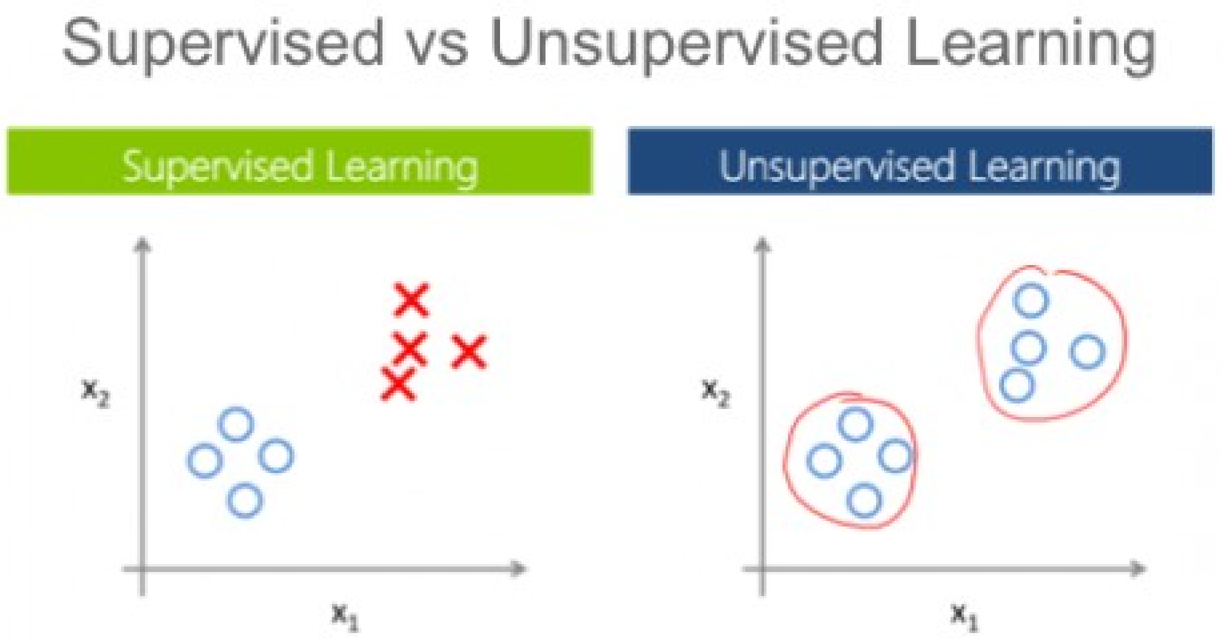
Machine learning algorithms are often categorized as **supervised** or **unsupervised**.

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



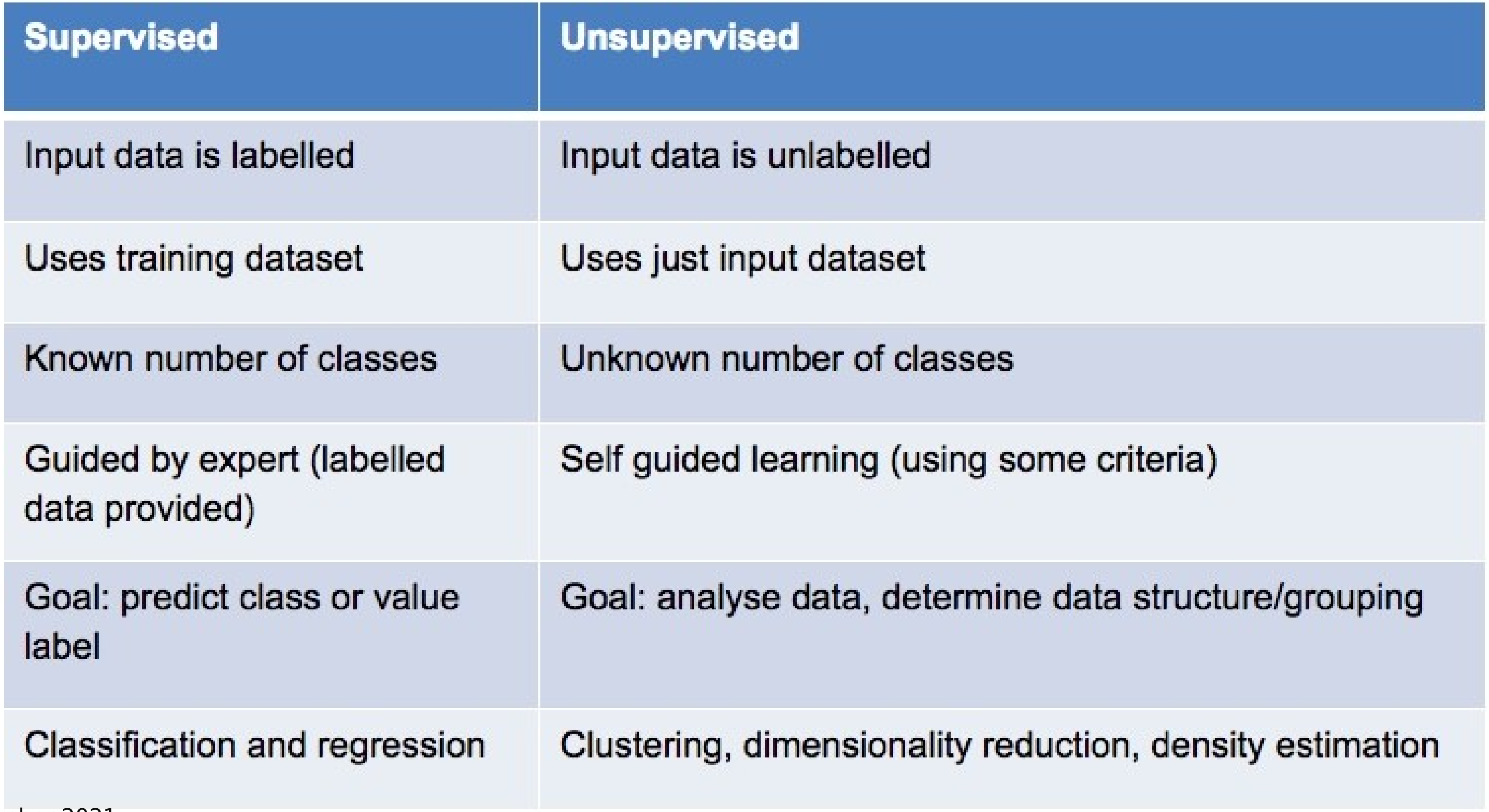
15-17 November 2021

# 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 |