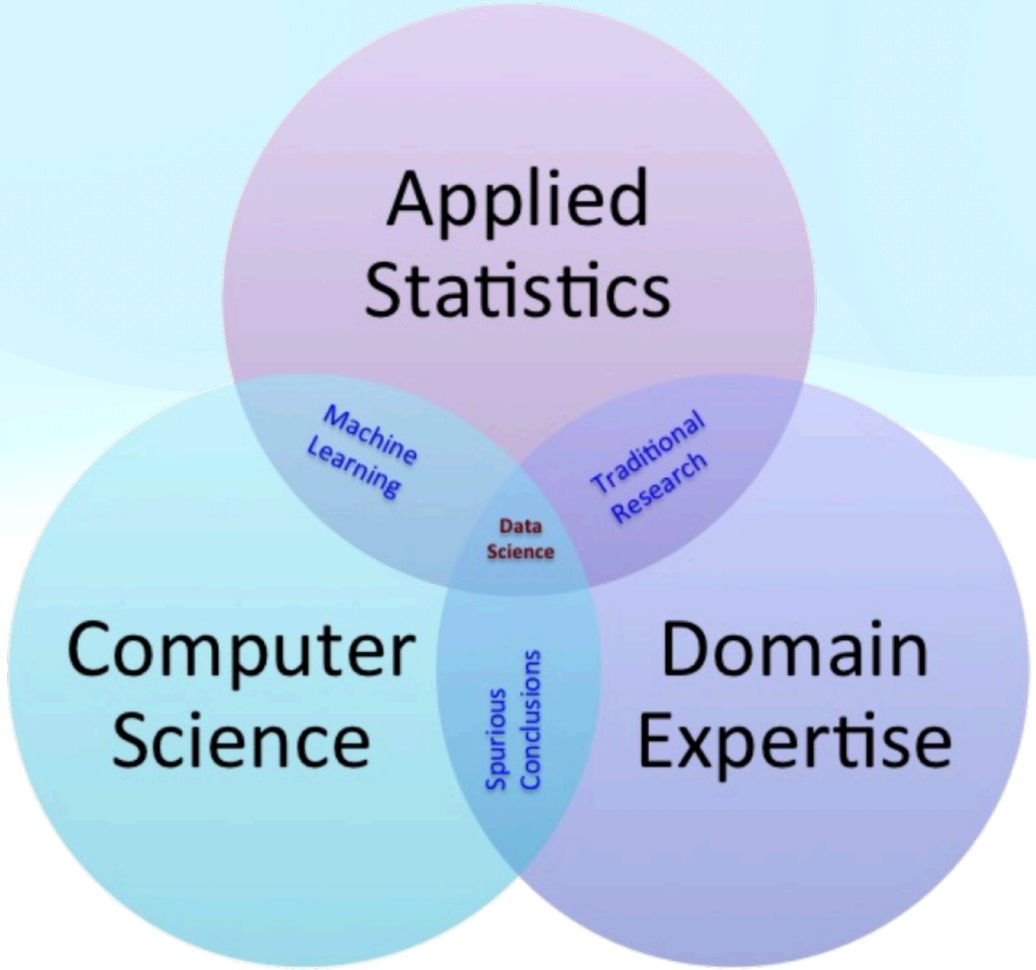
Data Science & Machine Learning in 2h

Data Science

Is ***Data Science*** a Science?

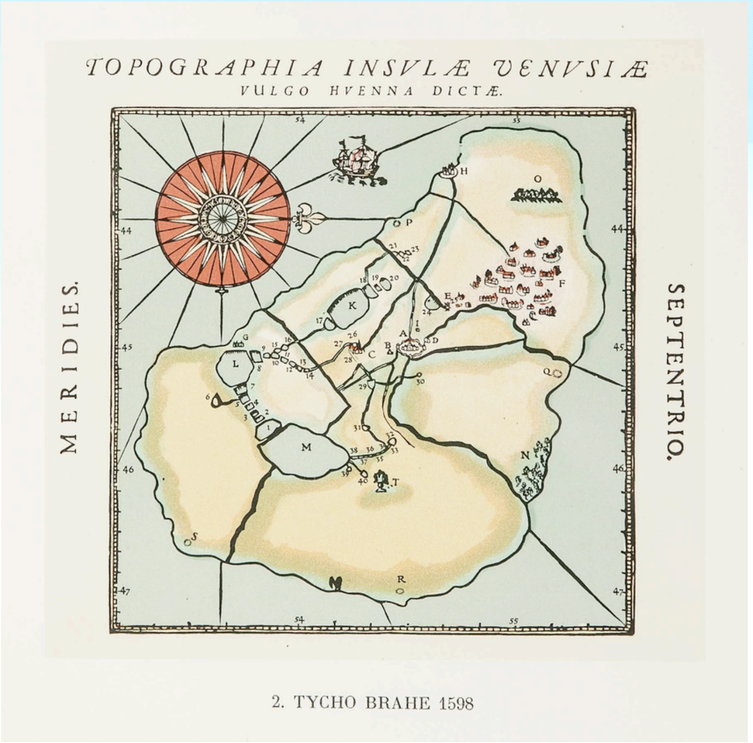
- Data Science as a discipline
- Data (intensive) Sciences



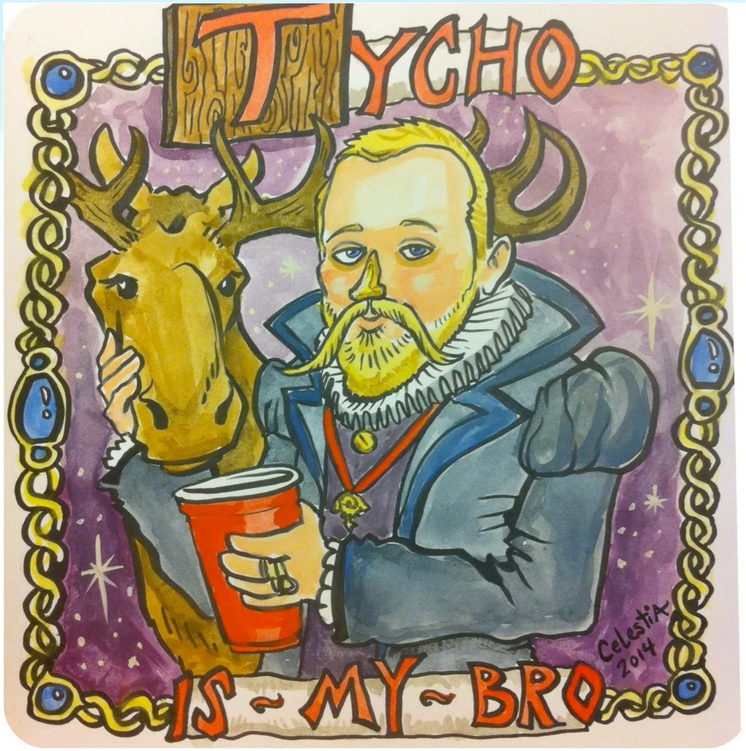
Data Science *is not about the data.*
It's about the questions.

Data Science *is not new…*

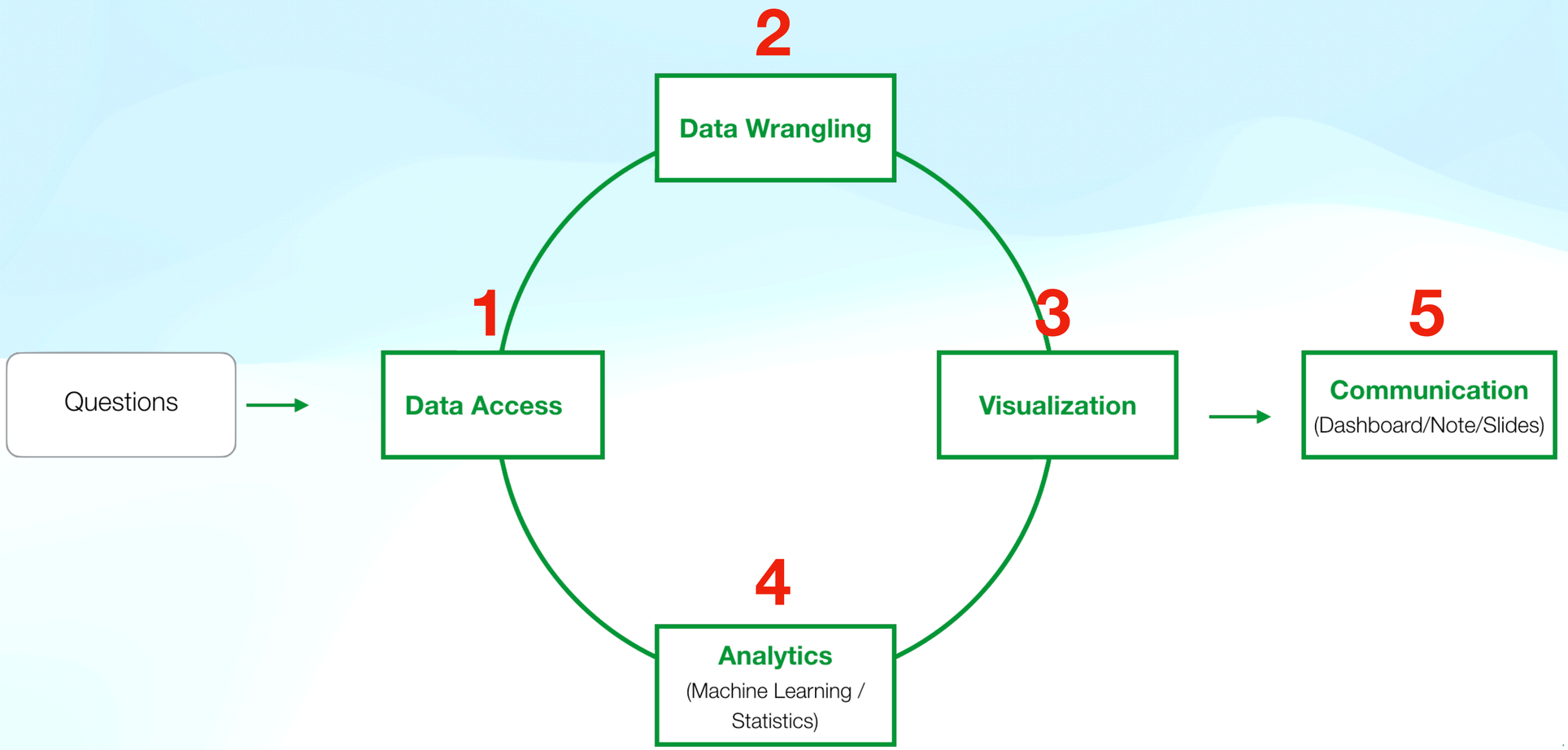
Small detour: *Tycho Brahe*



2. TYCHO BRAHE 1598



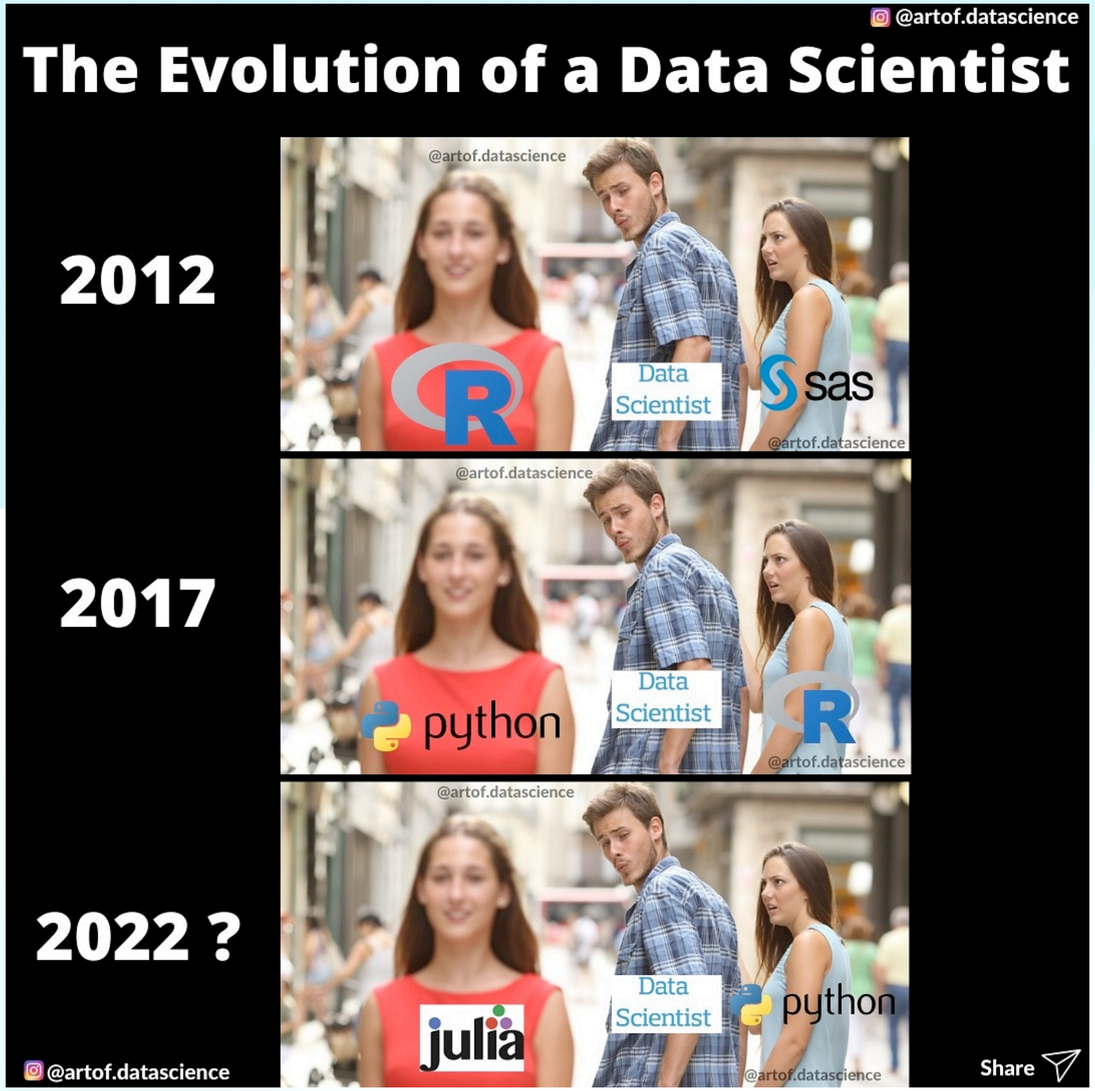
The 5 components of **Data science** practice



The ecosystem of **Data science: R vs Python**

The ecosystem of **Data science: R vs Python**

The ecosystem of **Data science: R vs Python**



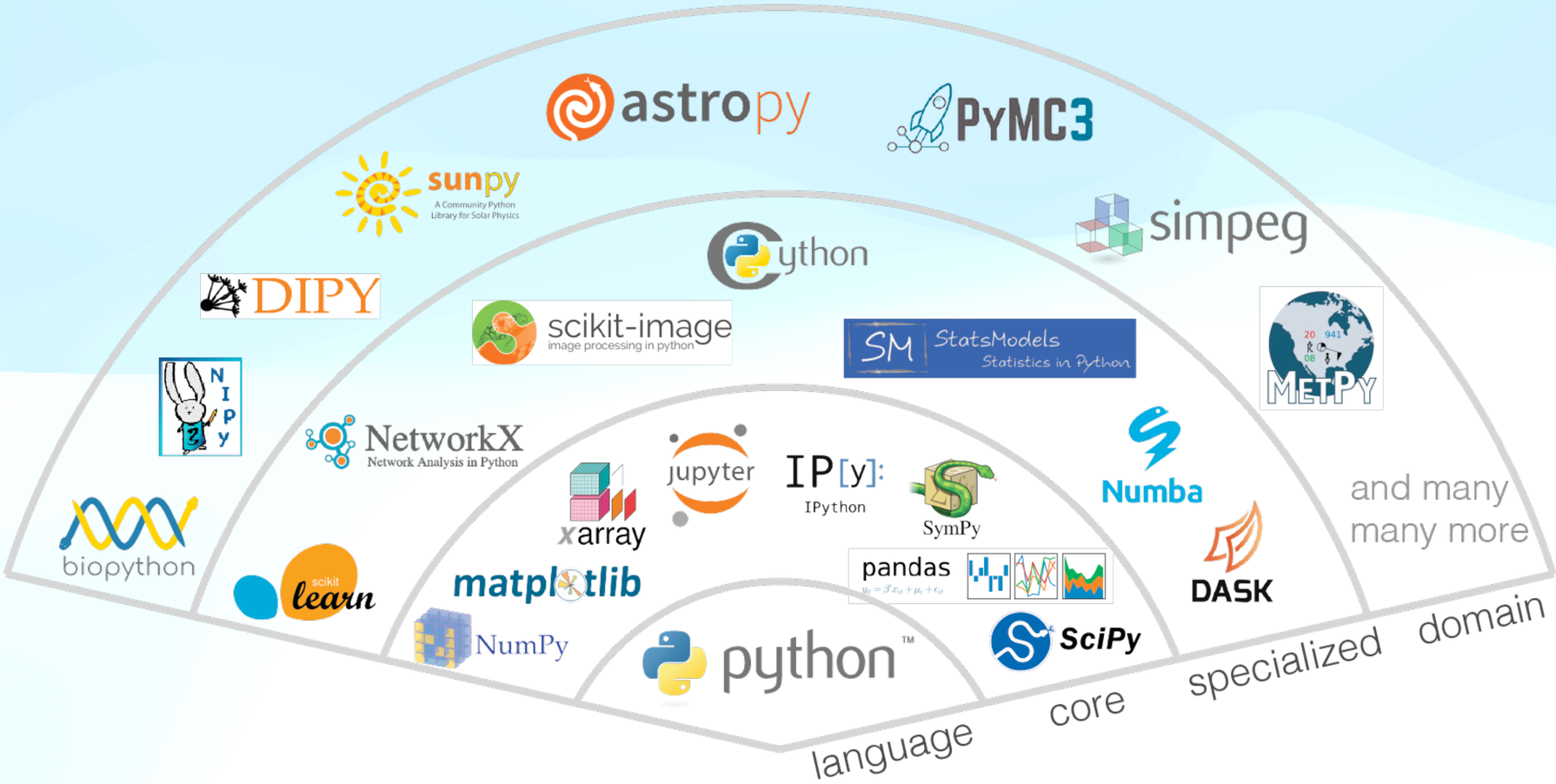
- Languages undergo fashion trends

- Coding is a tool, not a goal.
If it does the job, it’s a good tool.

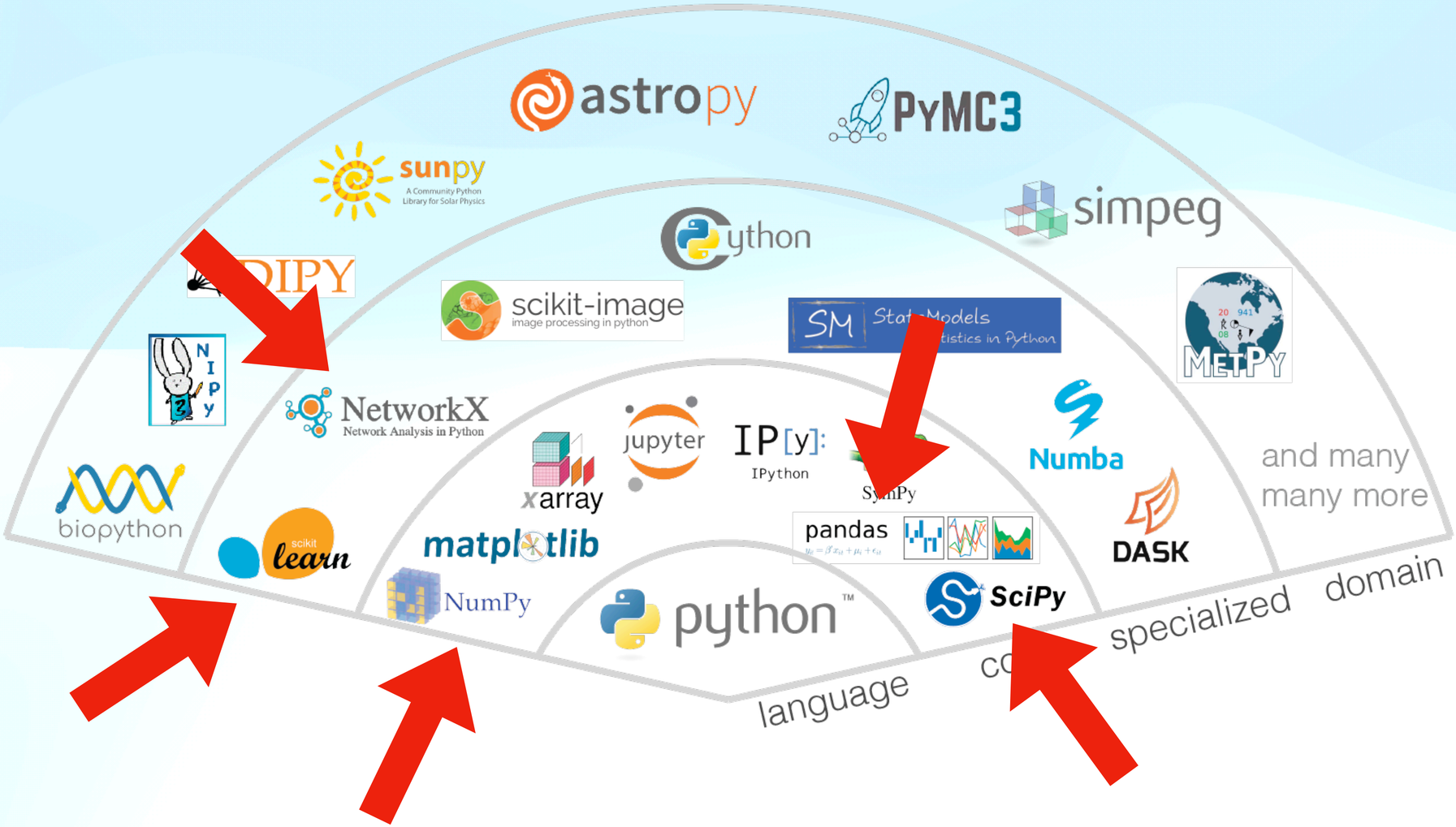
- Unless it’s Matlab, never use Matlab 🤢

(Or any proprietary software when there are better open-source community-driven alternatives)

It’s all about the **python ecosystem**



It’s all about the **python ecosystem**



It’s all about the **python ecosystem**

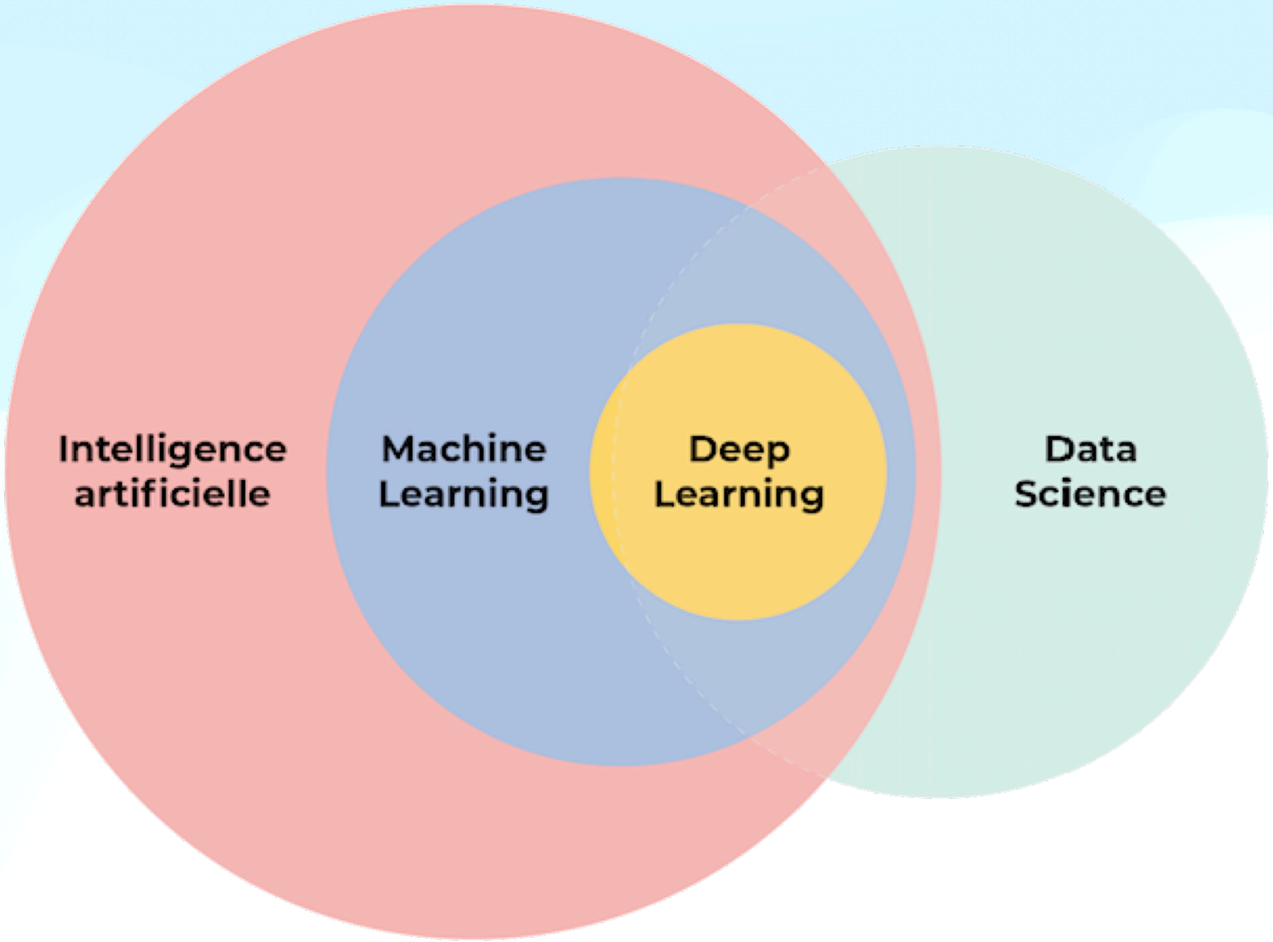
Python

- Nengo - Library for creating and simulating large-scale brain models.
- Nitime - Timeseries analysis for neuroscience data.
- Nilearn - Module for performing statistical learning/machine learning on NeuroImaging data.
- DIPY - Toolbox for analysis of MR diffusion imaging.
- MNE-Python - Community-driven software for processing time-resolved neural signals including electroencephalography (EEG) and magnetoencephalography (MEG).
- NiBabel - Provides read and write access to some common medical and neuroimaging file formats.
- PsychoPy - Package for running psychology and neuroscience experiments. It allows for creating psychology stimuli in Python.
- NuPic - Numenta Platform for Intelligent Computing is an implementation of Hierarchical Temporal Memory (HTM), a theory of intelligence based strictly on the neuroscience of the neocortex.
- Brian2 - Free, open source simulator for spiking neural networks.
- expyriment - Platform-independent lightweight Python library for designing and conducting timing-critical behavioural and neuroimaging experiments.
- BindsNET - Package for simulating spiking neural networks for reinforcement & machine learning.
- SpikeInterface - Framework designed to unify spike-sorting technologies
- NiMARE - NiMARE is a Python package for neuroimaging meta-analyses

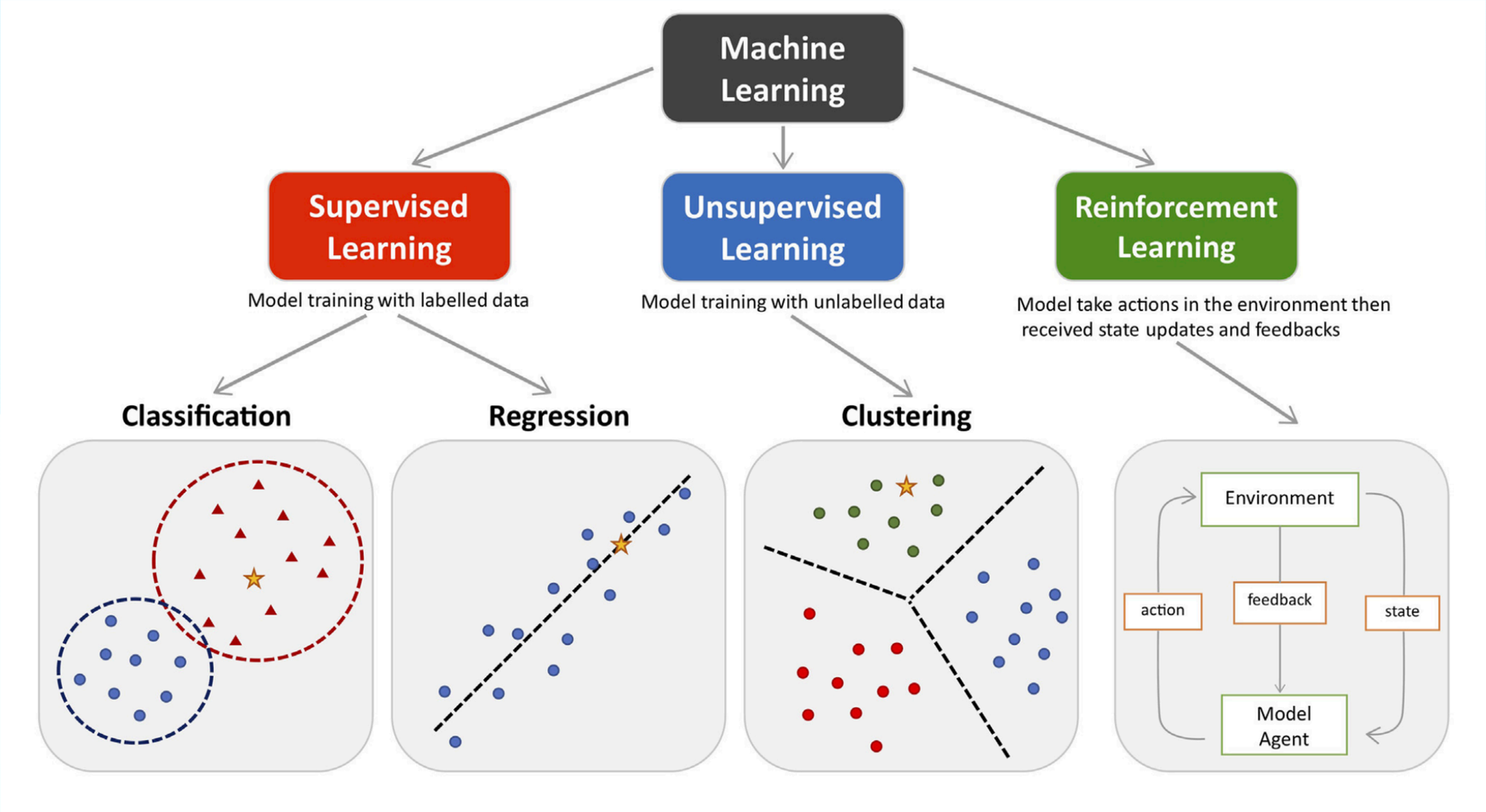
https://github.com/analyticalmonk/awesome-neuroscience#python

Machine Learning

What is Machine Learning



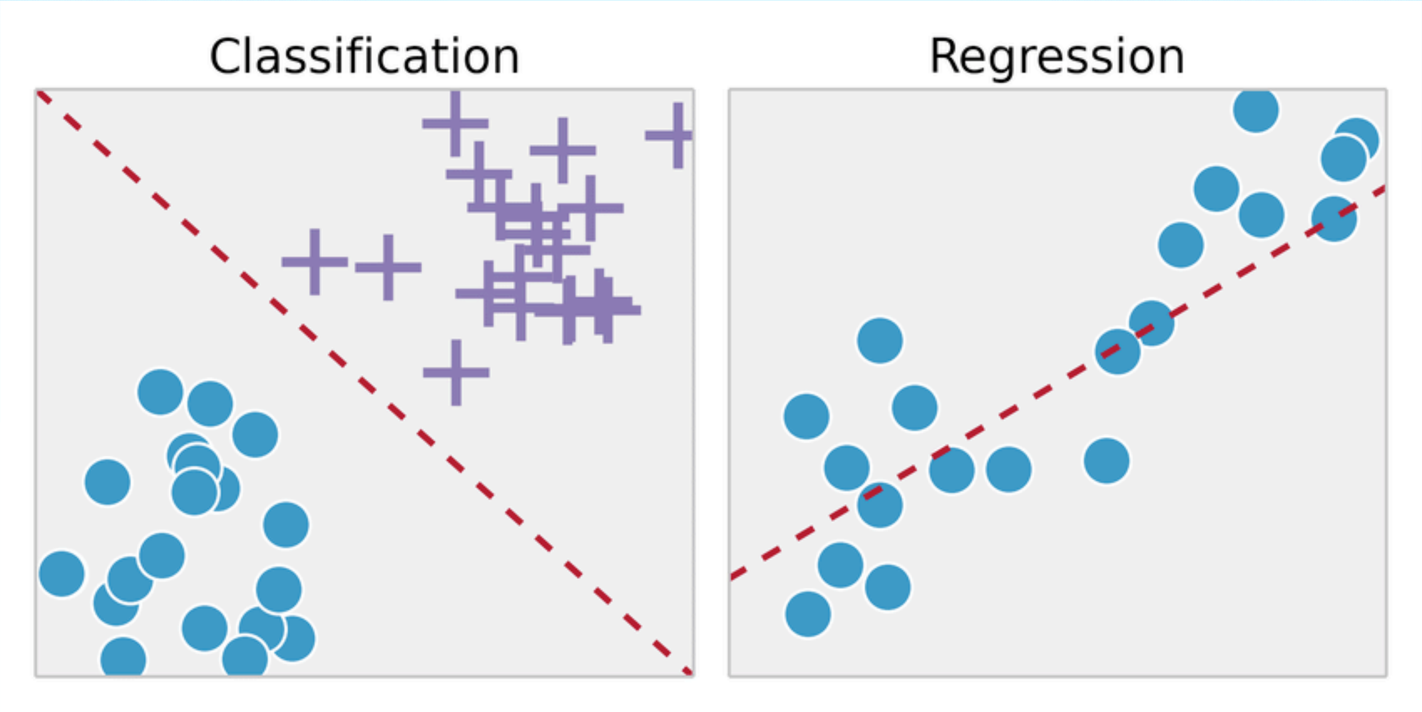
Types of **Machine Learning** ~~models~~ **problems**



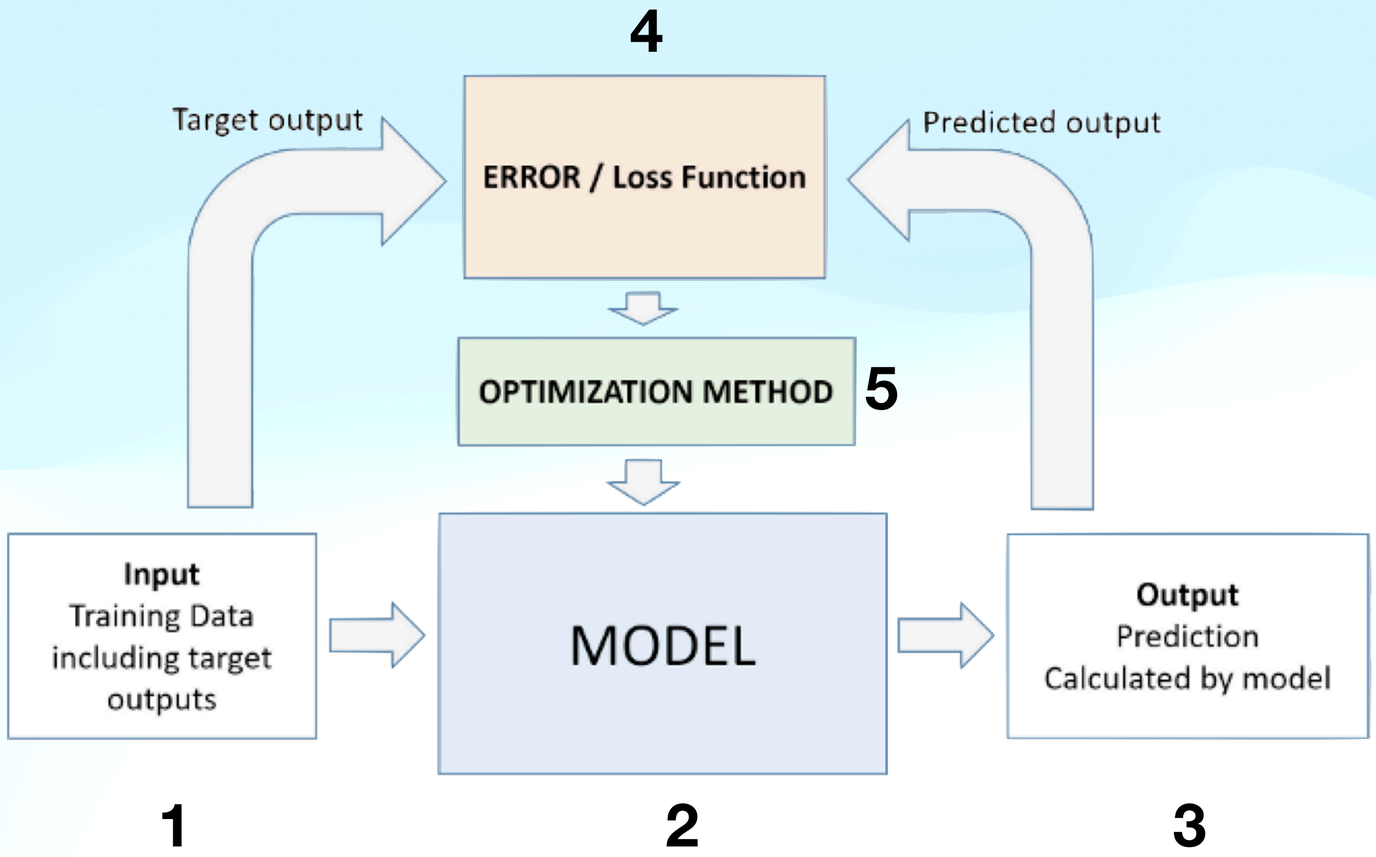
Machine Learning:

Supervised Learning (SL) tasks

Supervised Learning: regression vs classification

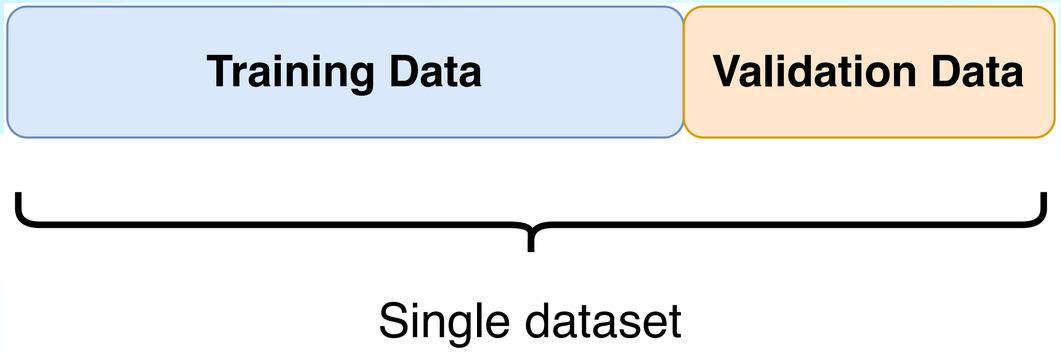


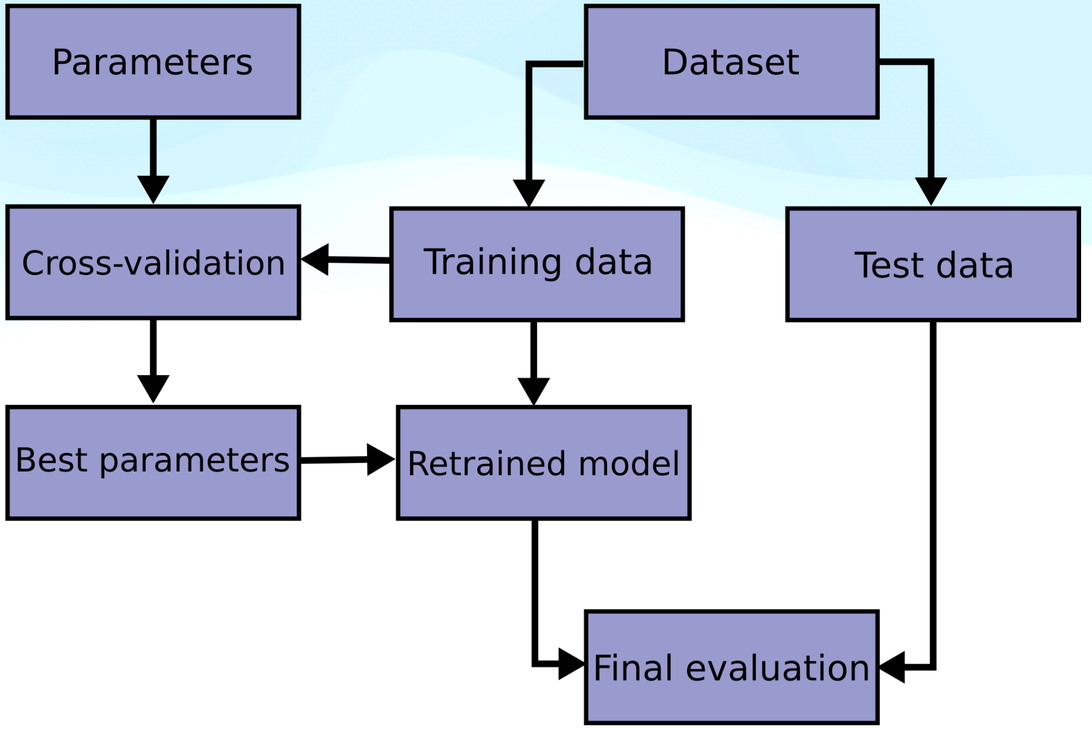
Supervised Learning: training



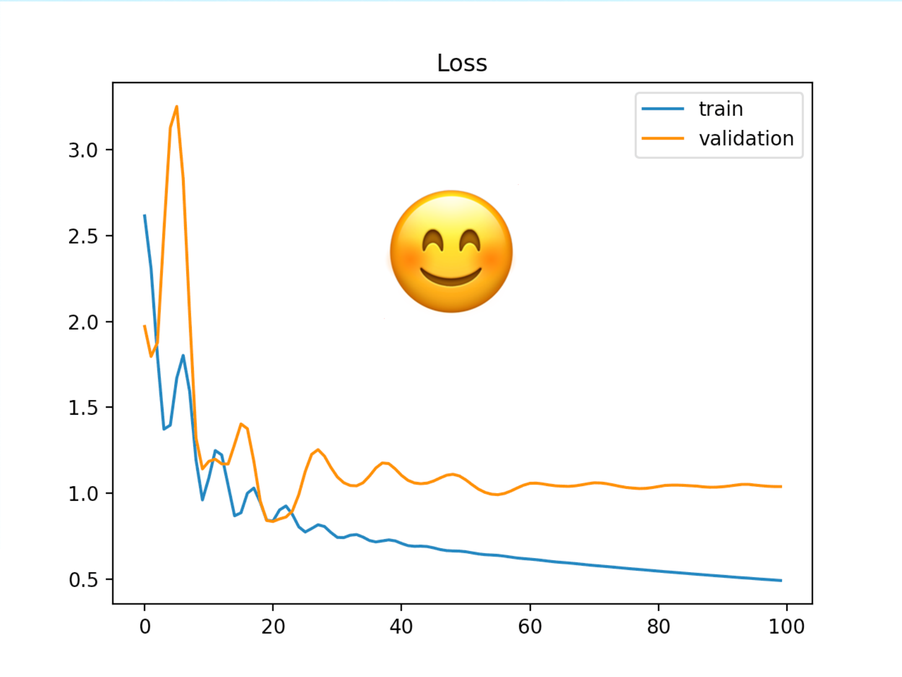
Training loop

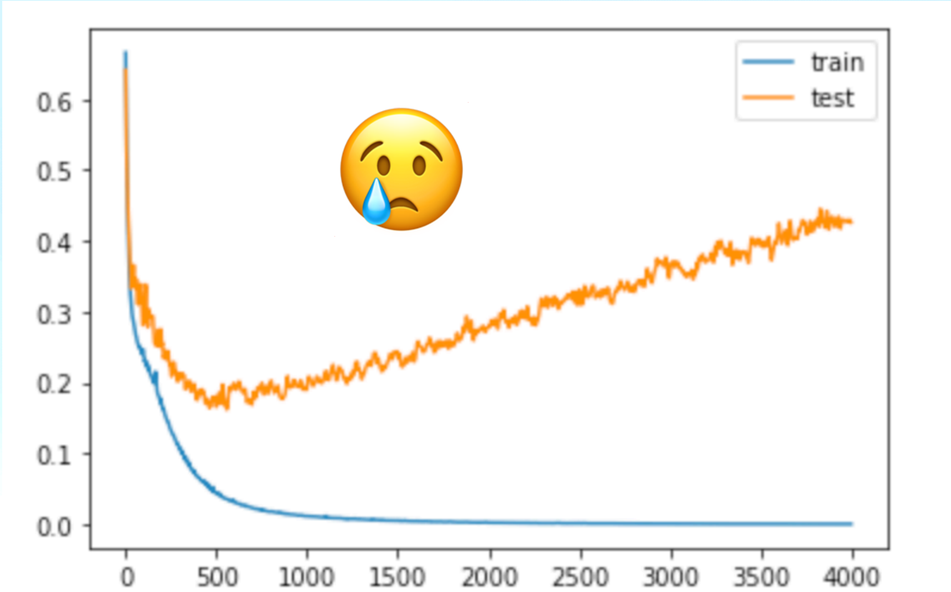
Supervised Learning: cross-validation

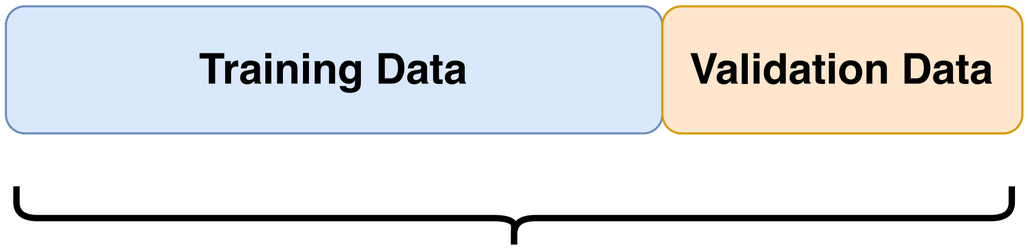




Supervised Learning: cross-validation







Types of **Machine Learning** ~~problems~~ **models**

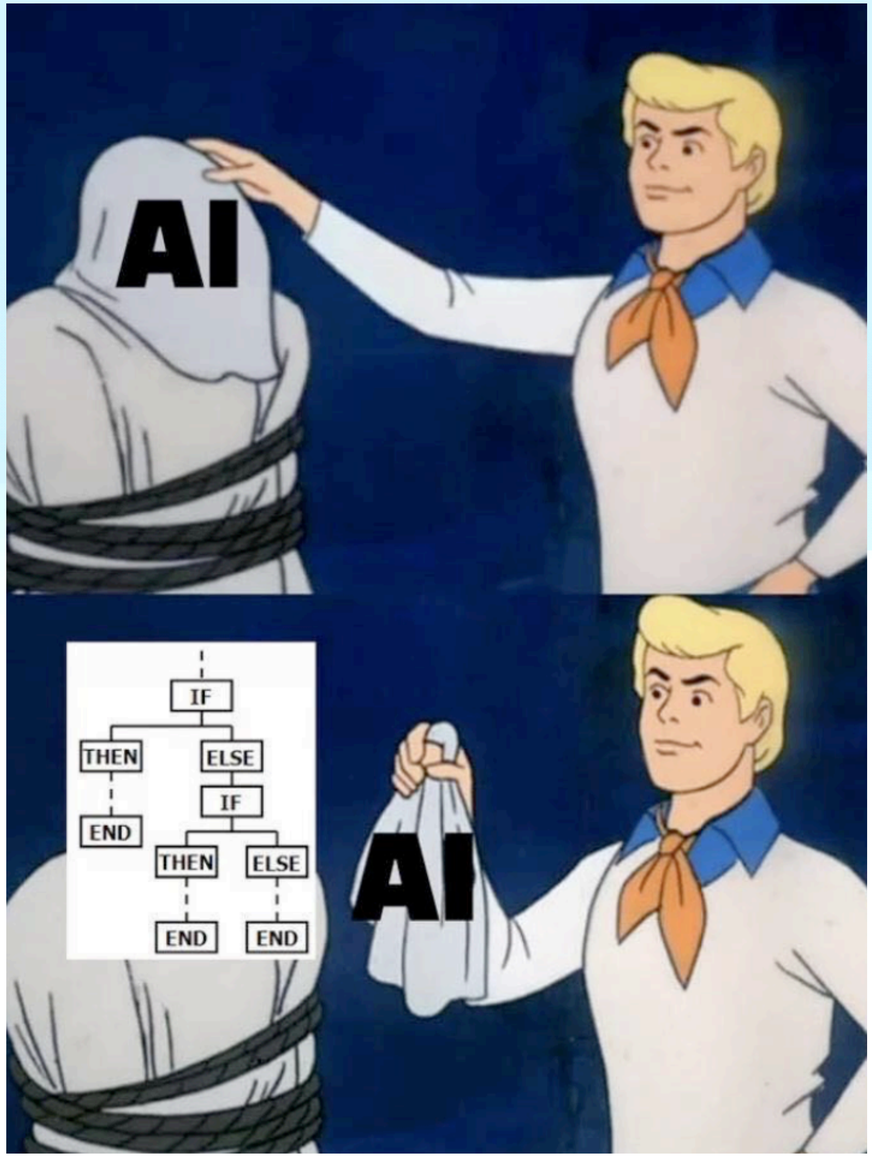
1. Decision trees & ensemble methods
2. Neural networks

Decision trees

Decision trees - Definition

Decision trees are a **supervised learning** algorithm used to predict categories or values —classification or regression tasks— based on a set of recursive binary splits of the feature variables.

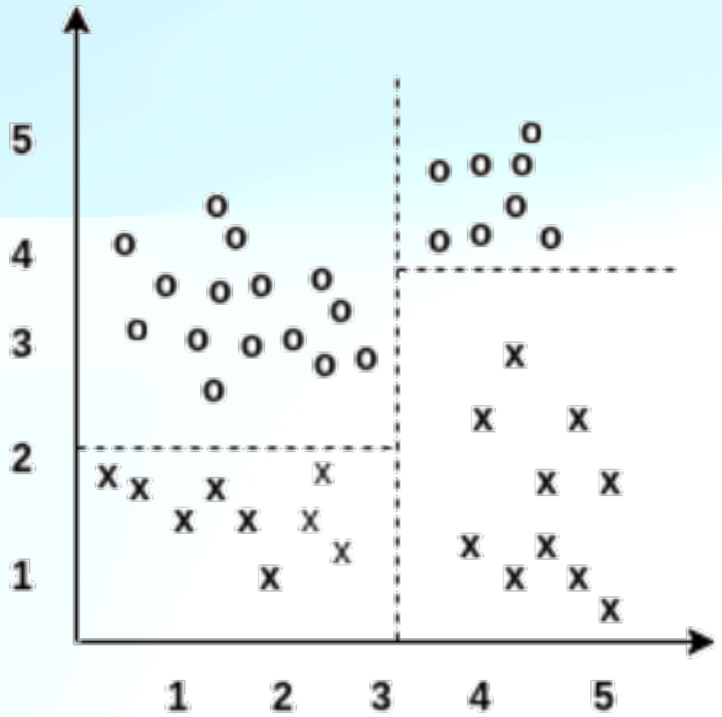
Like all SL tasks we are trying to find $f : X \rightarrow Y$ based on a given dataset {X,Y}

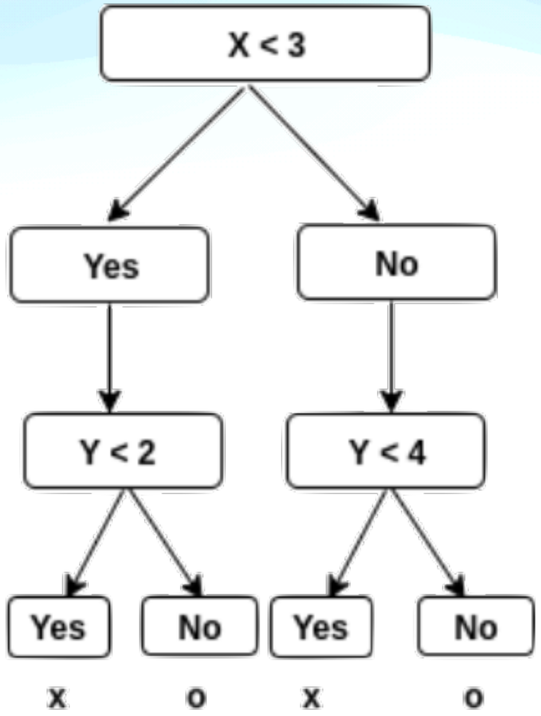


Decision trees

Decision trees - Classification

Classification : Split intervals of feature variables up to a certain depth

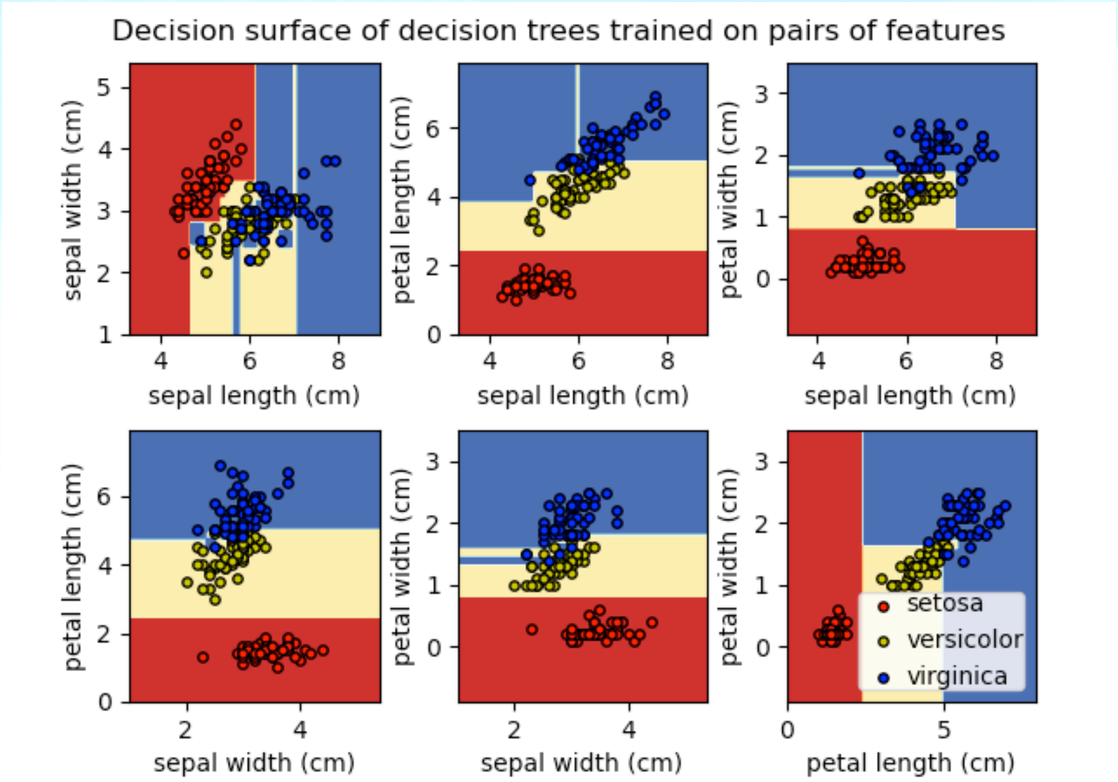




Decision trees

Decision trees - Classification

Trees : Not just binary! Any feature space dimension, and any amount of categories



Decision trees

Decision trees - Classification

Scikit-Learn tree classifier

- **Split criterion** or Loss function deciding where to split
  - Gini split criterium: Gini Index is 0 when the node is pure: all the contained elements in the node are of one unique class. Therefore, this node will not be split again. Splits chosen to minimise Gini of all splits over features.
  - Cross-entropy —or log-loss—: same as the one used for binary classification that you say in previous week: the sum of the log of each category prediction probabilities [0,1]. Splits chosen to minimise cross-entropy.
- See documentation for all options

$$H(Q_m) = \sum_k p_{mk}(1 - p_{mk})$$

$$H(Q_m) = -\sum_k p_{mk} \log(p_{mk})$$

k - number of classes

p_{mk} - fraction of samples correctly classified to their category k below split m

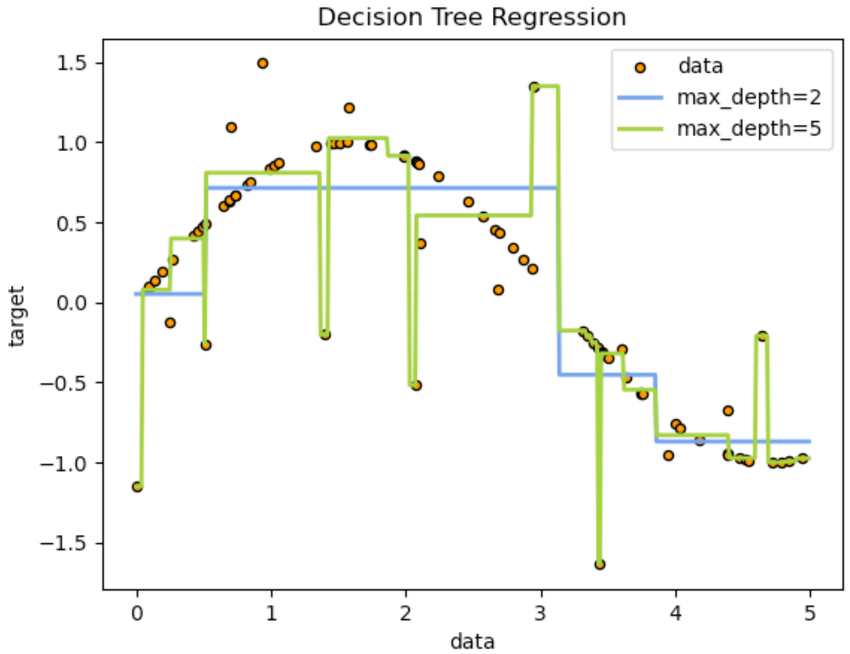
sklearn.tree.DecisionTreeClassifier

class sklearn.tree.**DecisionTreeClassifier**(*, *criterion='gini'*, *splitter='best'*, *max_depth=None*, *min_samples_split=2*, *min_samples_leaf=1*, *min_weight_fraction_leaf=0.0*, *max_features=None*, *random_state=None*, *max_leaf_nodes=None*, *min_impurity_decrease=0.0*, *class_weight=None*, *ccp_alpha=0.0*) [source]

Decision trees

Decision trees - Regression

Regression : Same as for classification but the output of the decision tree is not a categorical label but a **scalar value** —or a set of them—.



Click here to visualise the role of tree depth in fitting the data in a regression task

Decision trees

Decision trees - Regression

Scikit-Learn tree regressor

- Criterion: Loss function, e.g. square error, abs error, etc.
- max_depth : maximal depth of the tree
- …
- See documentation for all options

sklearn.tree.DecisionTreeRegressor

class sklearn.tree.**DecisionTreeRegressor**(*, *criterion='squared_error'*, *splitter='best'*, *max_depth=None*, *min_samples_split=2*, *min_samples_leaf=1*, *min_weight_fraction_leaf=0.0*, *max_features=None*, *random_state=None*, *max_leaf_nodes=None*, *min_impurity_decrease=0.0*, *ccp_alpha=0.0*) ¶ [source]

Decision trees

Decision trees

Advantages over ANNs

- Easier to train
- Explainable-ish
- No data normalisation needed (scale invariant of feature values)
- Less architecture design —we still have some hyper parameters.

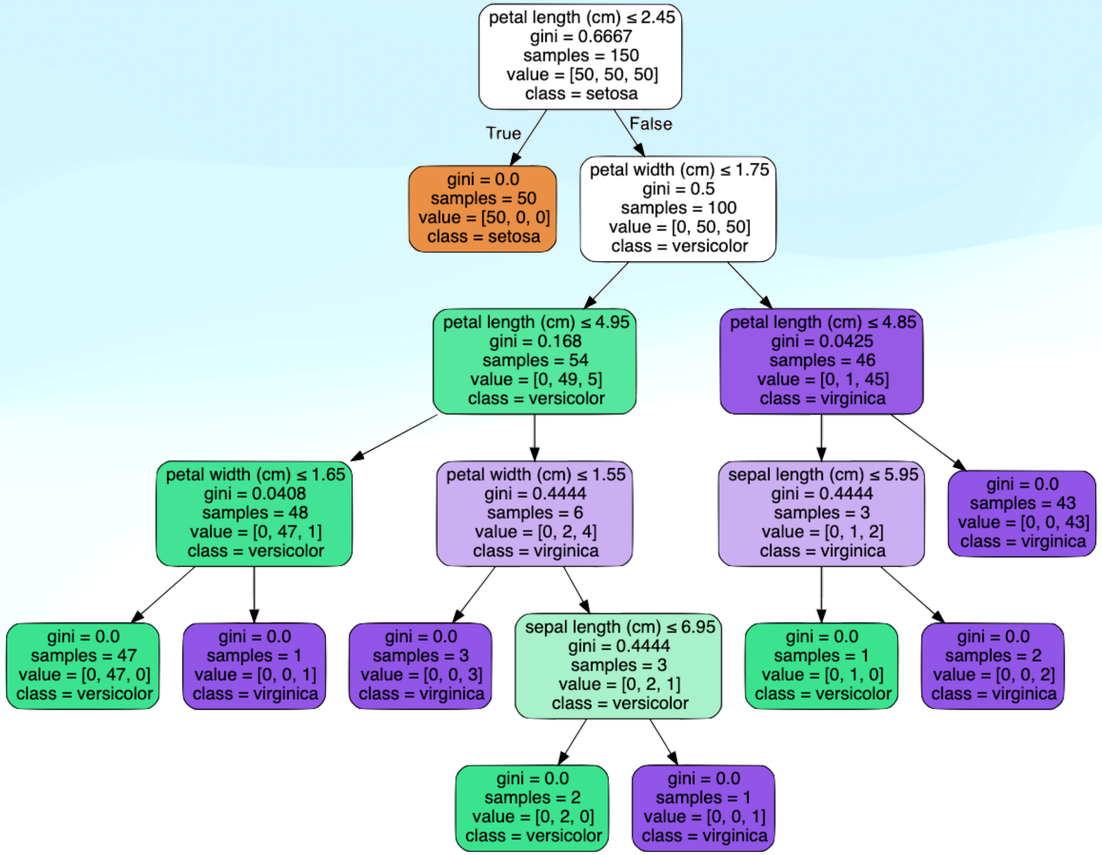
Ultimate criterium:

Empirical performance

More on advantage disadvantages: https://scikit-learn.org/stable/modules/tree.html

Decision trees

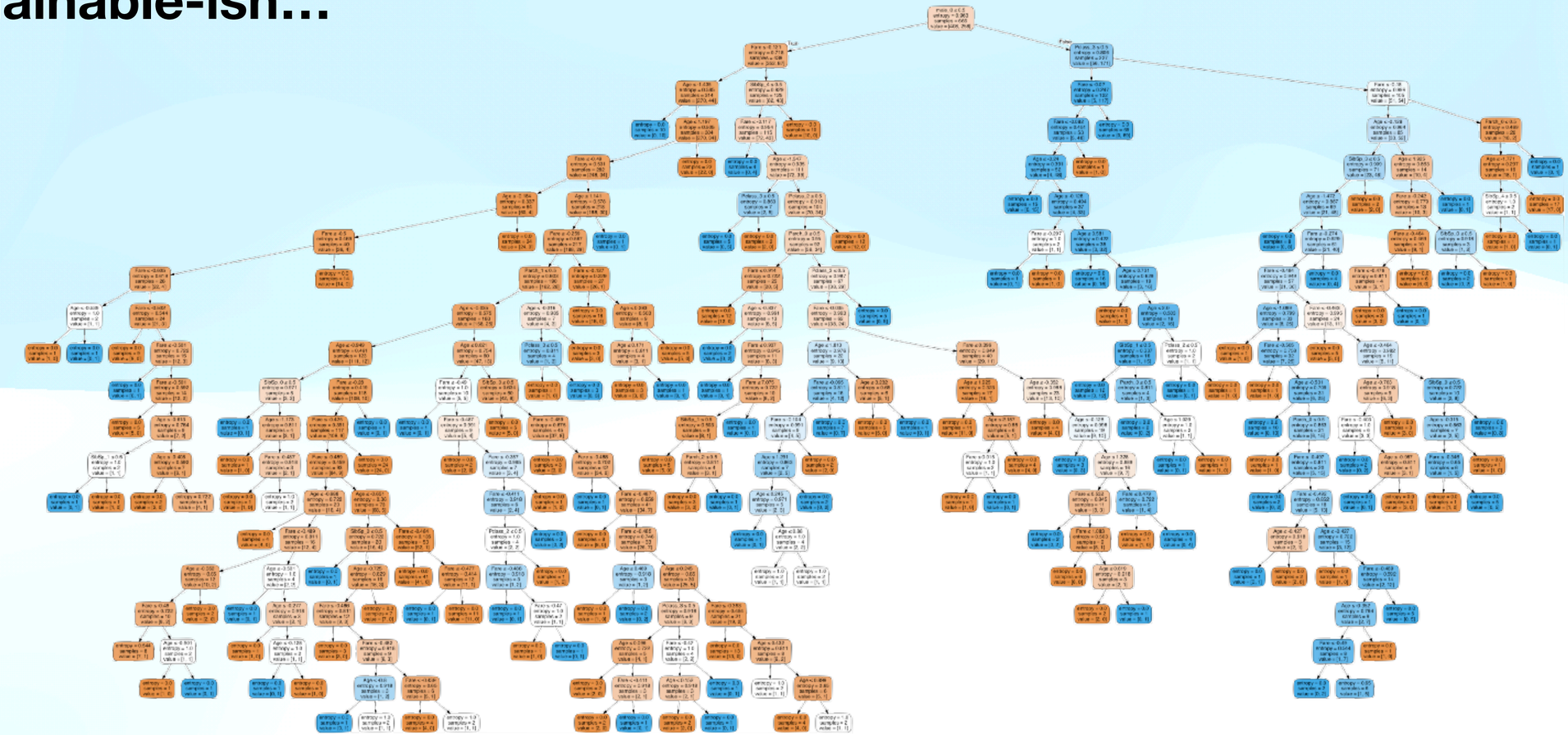
Explainable-ish…



Flower types as function of petal and sepal sizes

Decision trees

Explainable-ish…



Survival on the titanic based on age, sex, ticket fare…

Decision trees

Decision trees

More information on decision trees, see:

- Aurélien Géron's textbook **chapter 6**
- Scikit-Learn **page on decision trees**

Decision trees

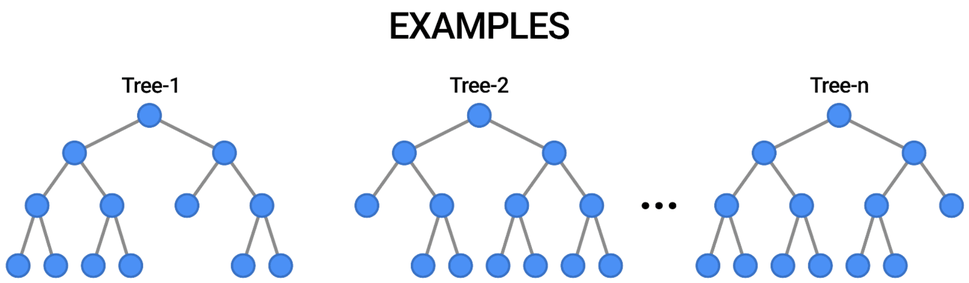
Decision trees

Ensemble methods

Simplest ensemble method: **voting** 🗳️

Boosting: use many trees and add weights (boost) to focus on difficult parts of data. Extremely performing, ie. **XGBoost** is often the winner algorithm at Kaggle competitions involving tabular data

Ensembles sacrifice explainability ✝️



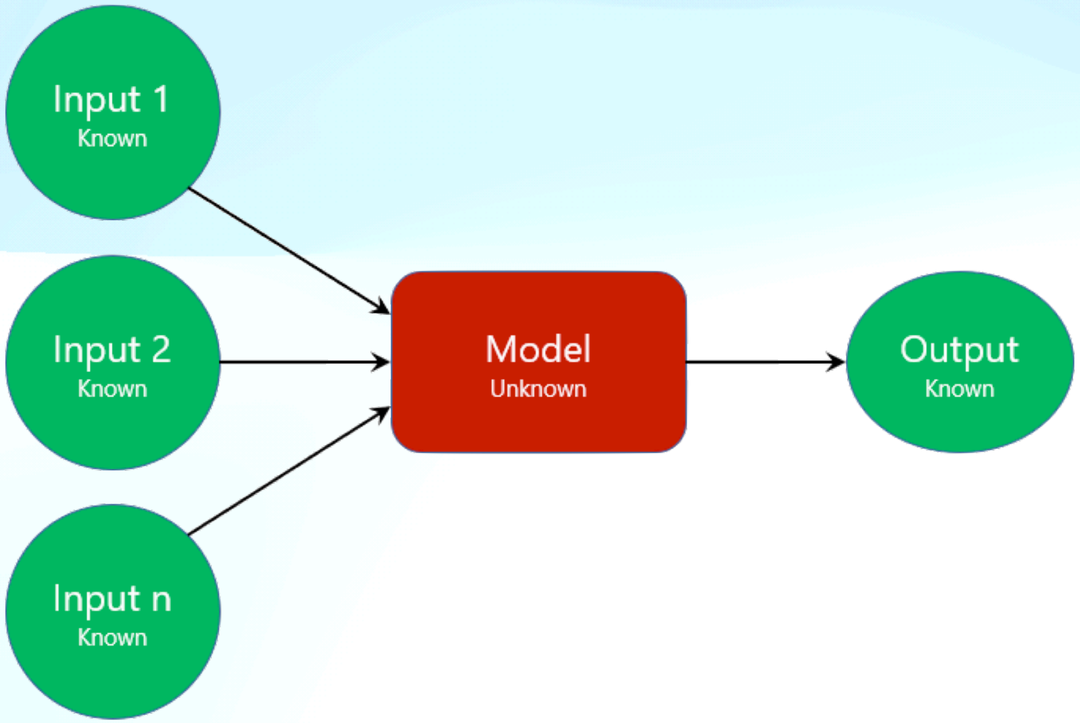
Machine Learning:

Supervised Learning (SL):

Deep learning

ML models - **Deep Learning**

Deep learning ≈ a **black box** that maps one space onto another



- RNA-seq to neuron type
- fMRI activity to visual perception
- Protein concentration to healthy/ pathological state

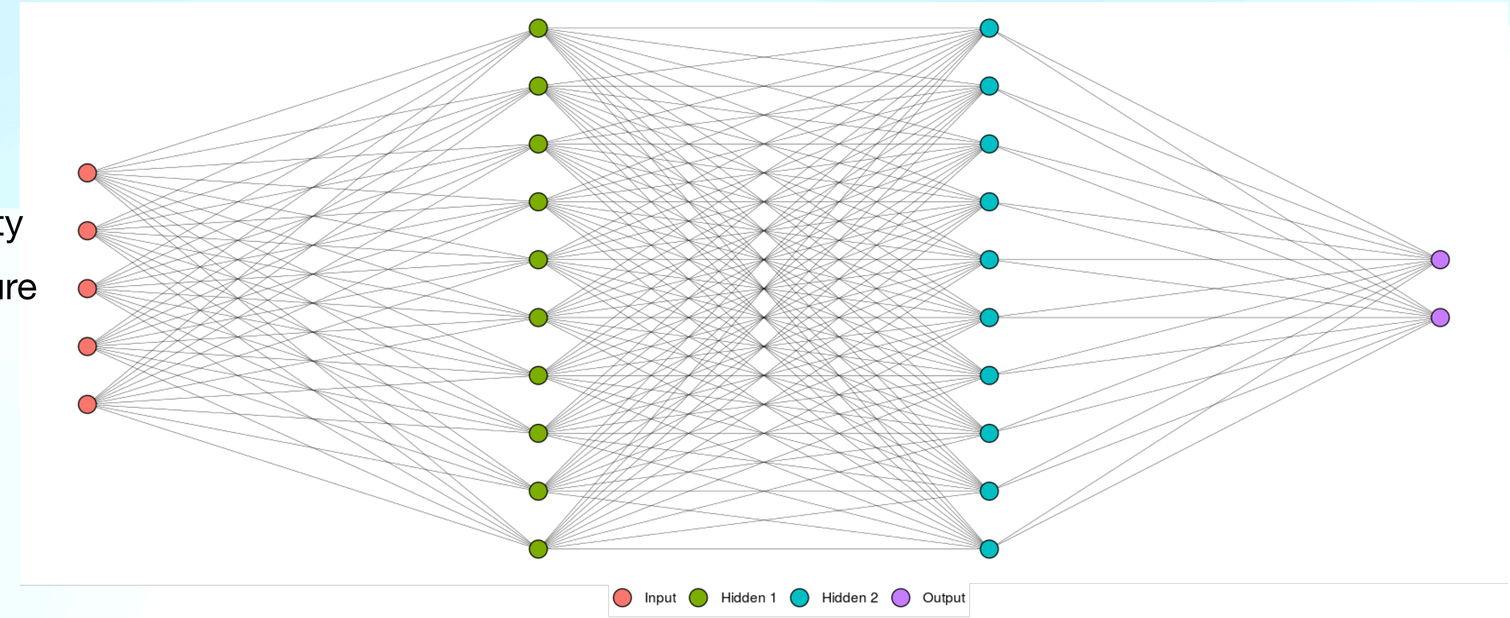
Supervised Learning: how to

- Deep learning ≈ function approximation, map one space onto another
- Effective and modular

Star Luminosity

Star Temperature

...

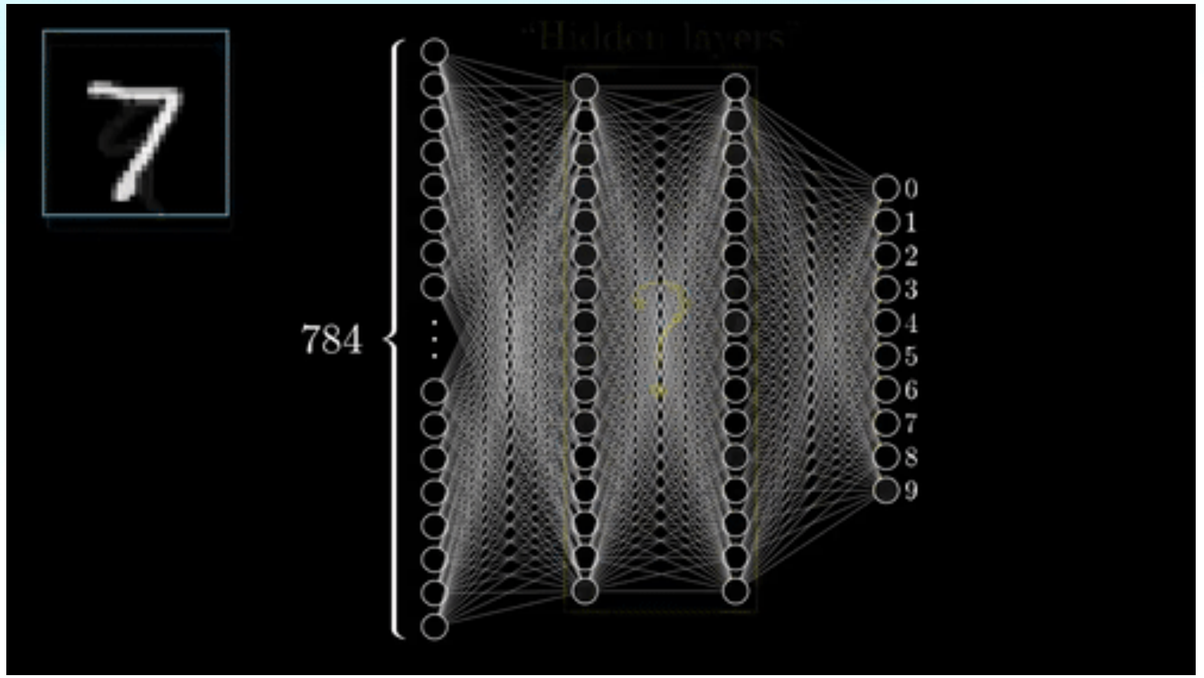


Red Giant

White Dwarf

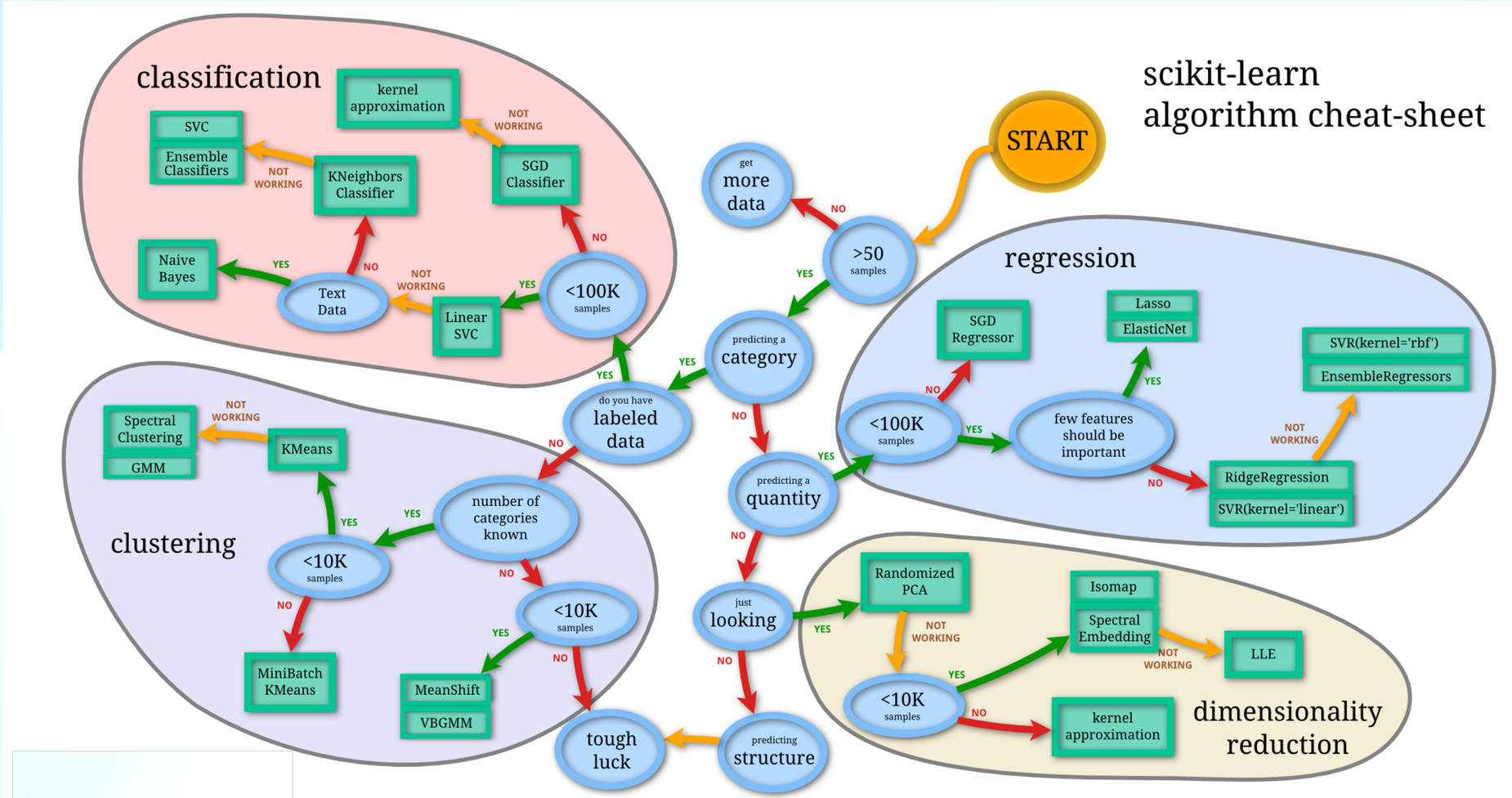
ML Inverse models - **Deep Learning**

- Deep learning ≈ function approximation, map one space onto another
- Effective and modular



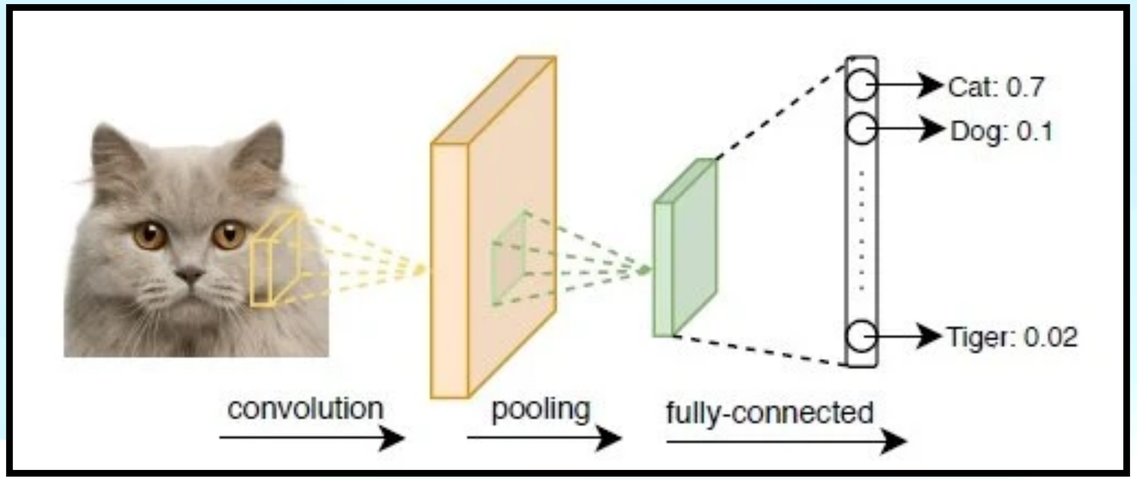
Supervised Learning: choosing your model

Do I need neural networks, trees or linear models for my data?

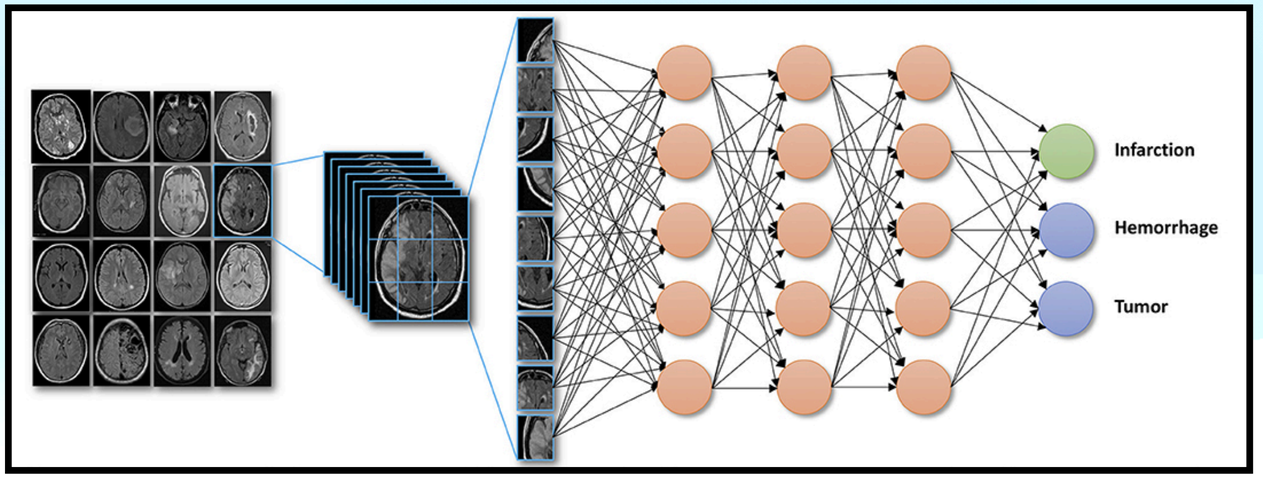


ML models - **Deep Learning**

Useful applications



Even more useful applications



Bad applications

Deep Neural Networks Are More Accurate Than Humans at Detecting Sexual Orientation From Facial Images

By Michal Kosinski, Yilun Wang

Journal of Personality and Social Psychology. February 2018, Vol. 114, Issue 2, Pages 246–257.

Dystopian applications

Article | Open Access | Published: 11 January 2021

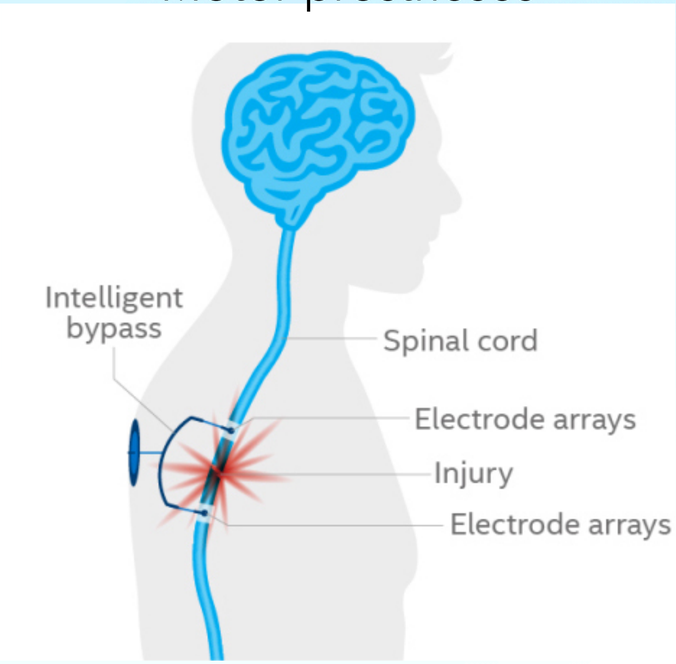
Facial recognition technology can expose political orientation from naturalistic facial images

Michal Kosinski

Scientific Reports **11**, Article number: 100 (2021) | Cite this article

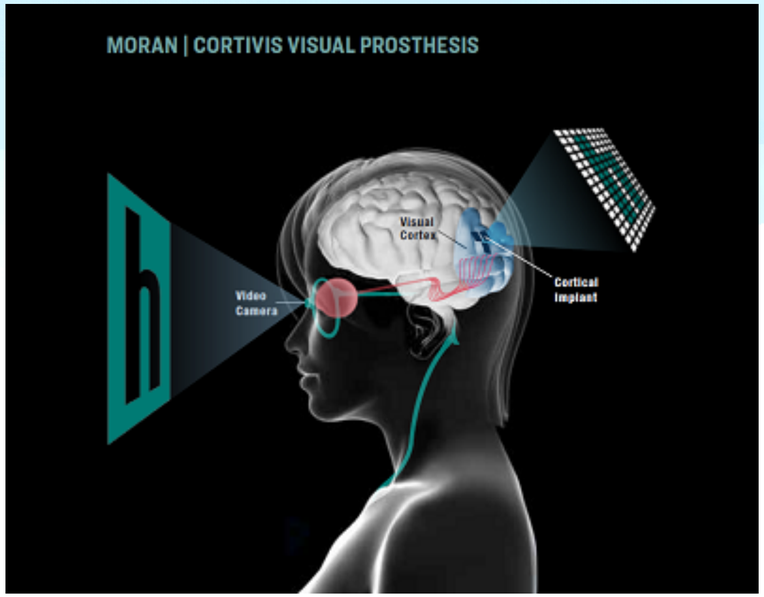
ML models applications - **Neural prosthetics**

Motor prostheses



Auditory prostheses

Visual prostheses



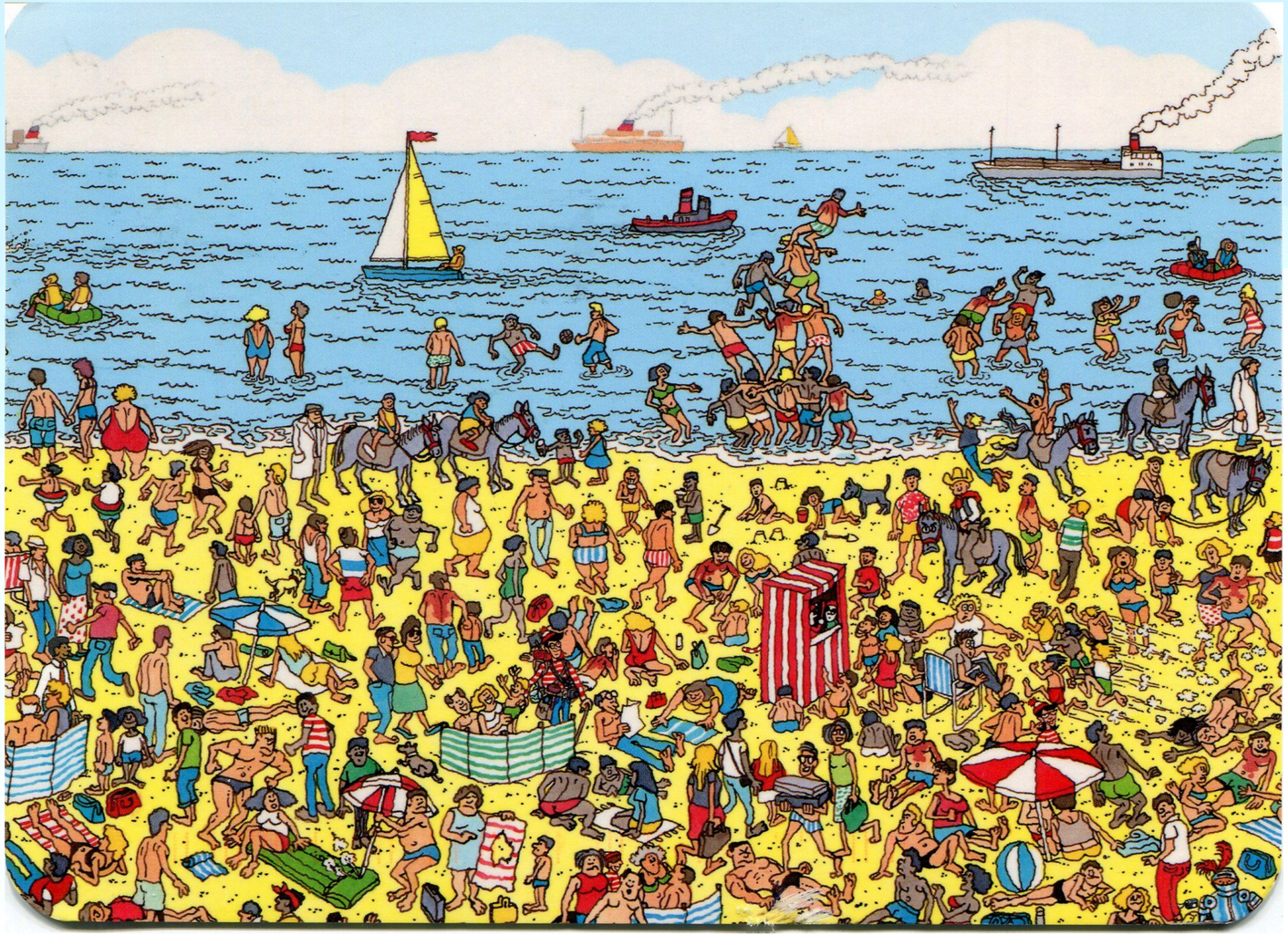
Machine Learning:

Unsupervised Learning

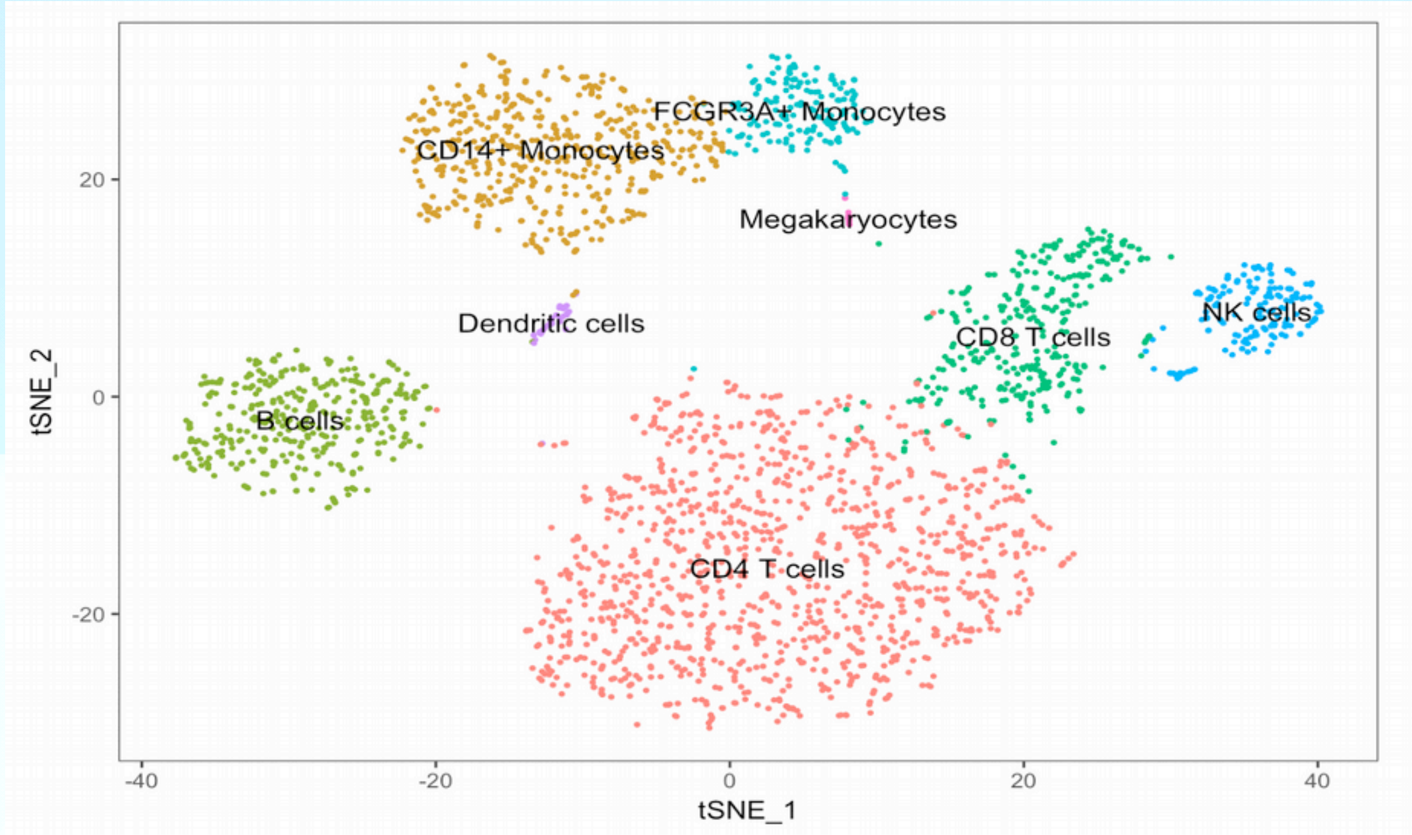
Machine Learning:

Unsupervised Learning : Dimensionality reduction

Dimensionality reduction: *Where is Wally?*



Dimensionality reduction: *Clustering*

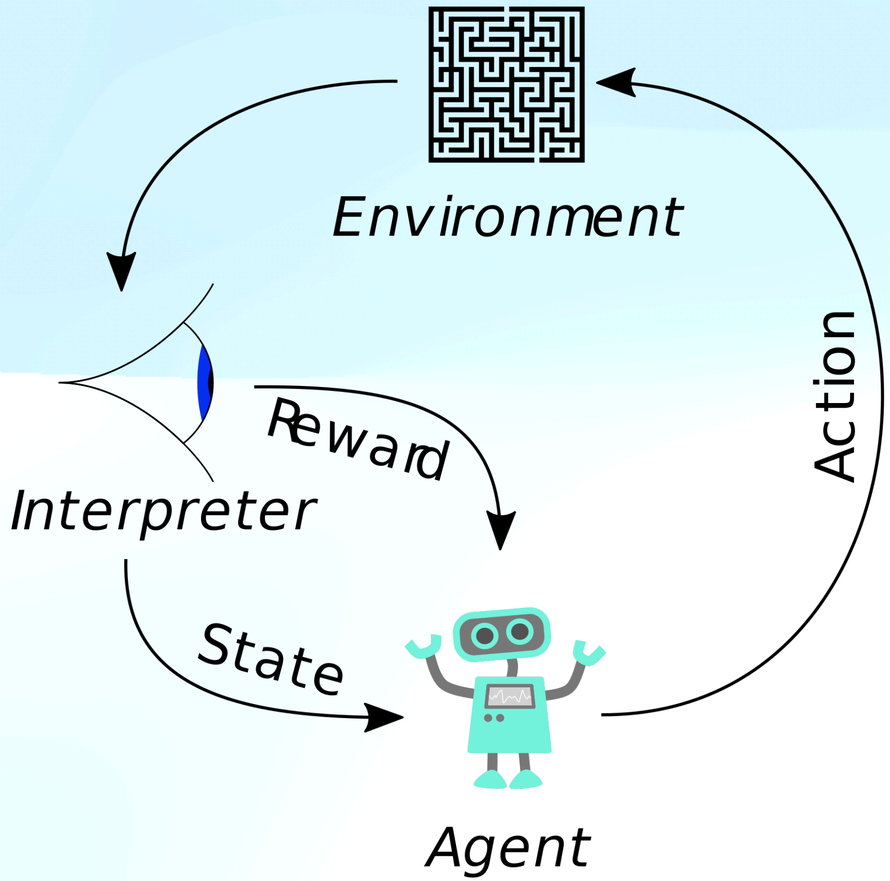


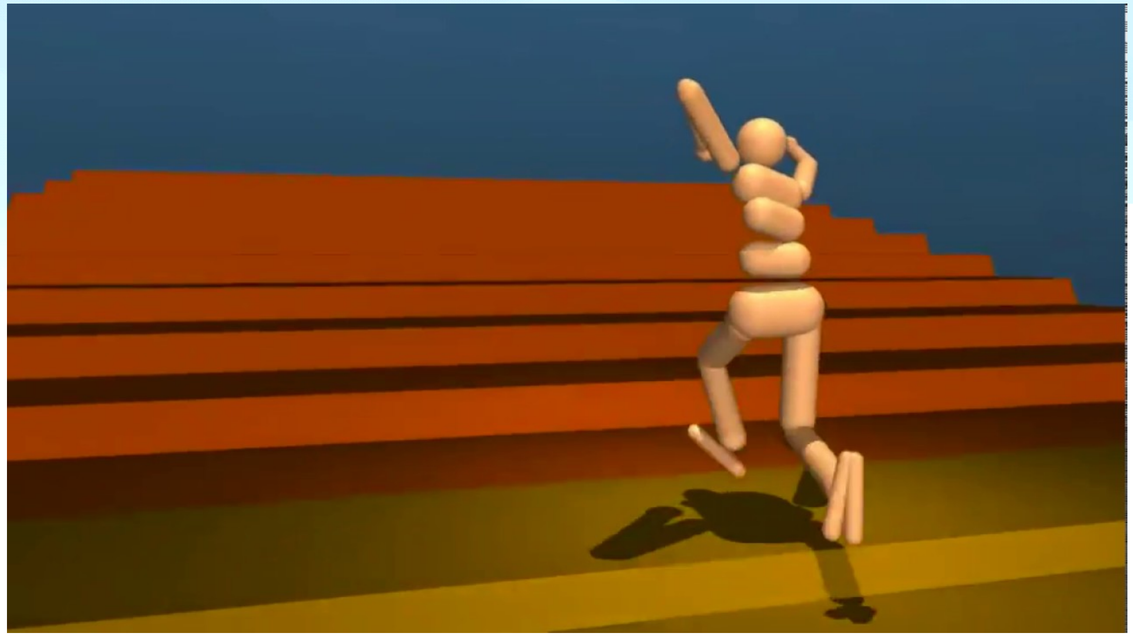
Single cell transcriptomics

Machine Learning:

Reinforcement Learning (RL)

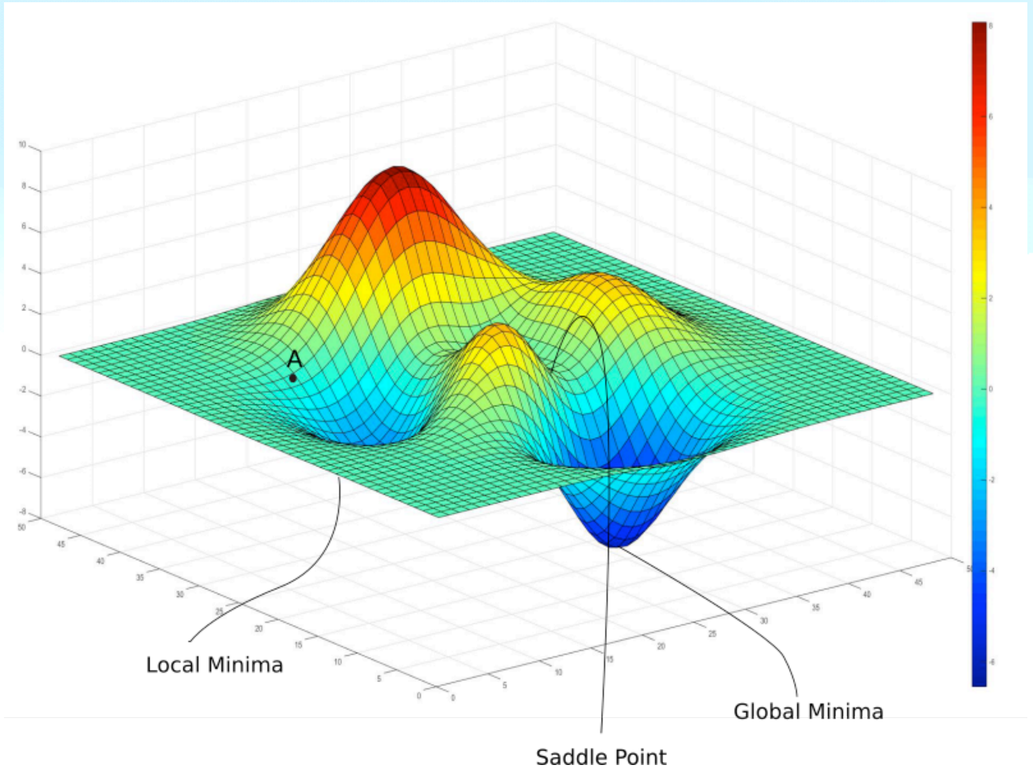
Reinforcement Learning





Reinforcement Learning is just fancy optimisation of a **objective function**

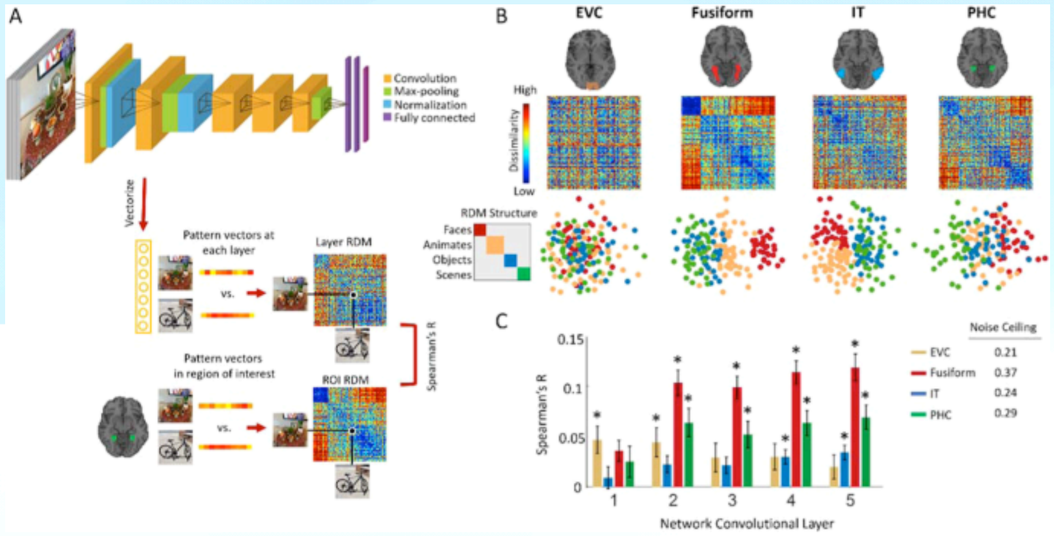
- Also called **fitness function** or **loss function**
- **Objective function** is a metric —a distance —that indicates how close a given solution is to produce our objective.
- Set of aims may include:
  - **minimising** the objective function (expressing a loss or cost)
  - **maximising** the objective function (expressing a reward)

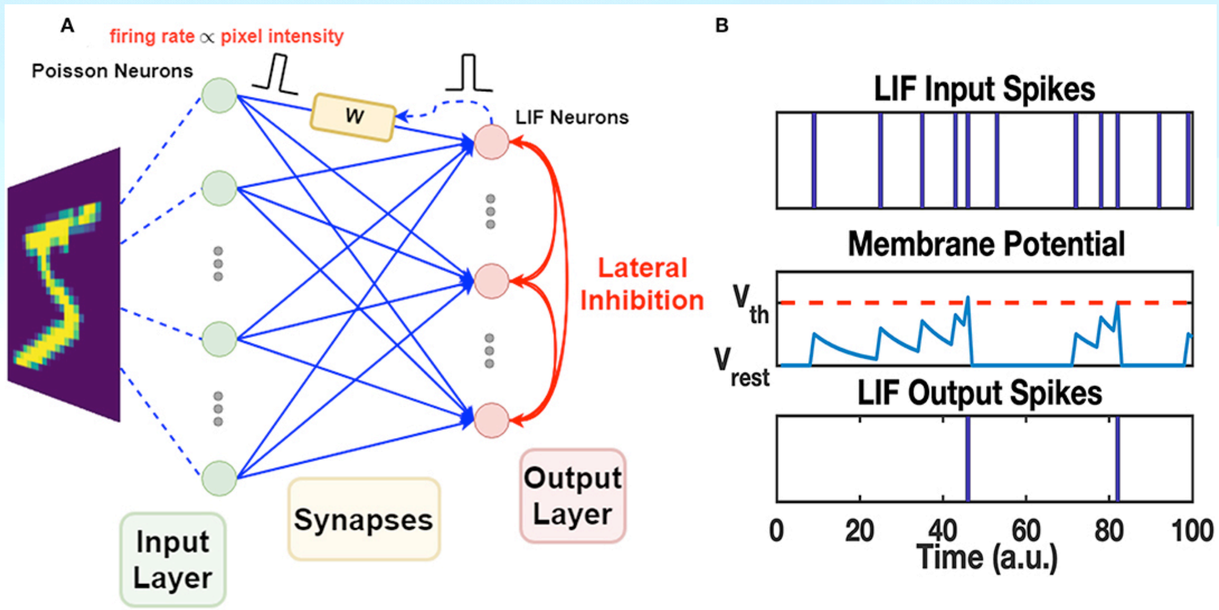


Geometric view on the objective function

Deep Learning <-> Neuroscience

Machine Learning <-> Neuroscience

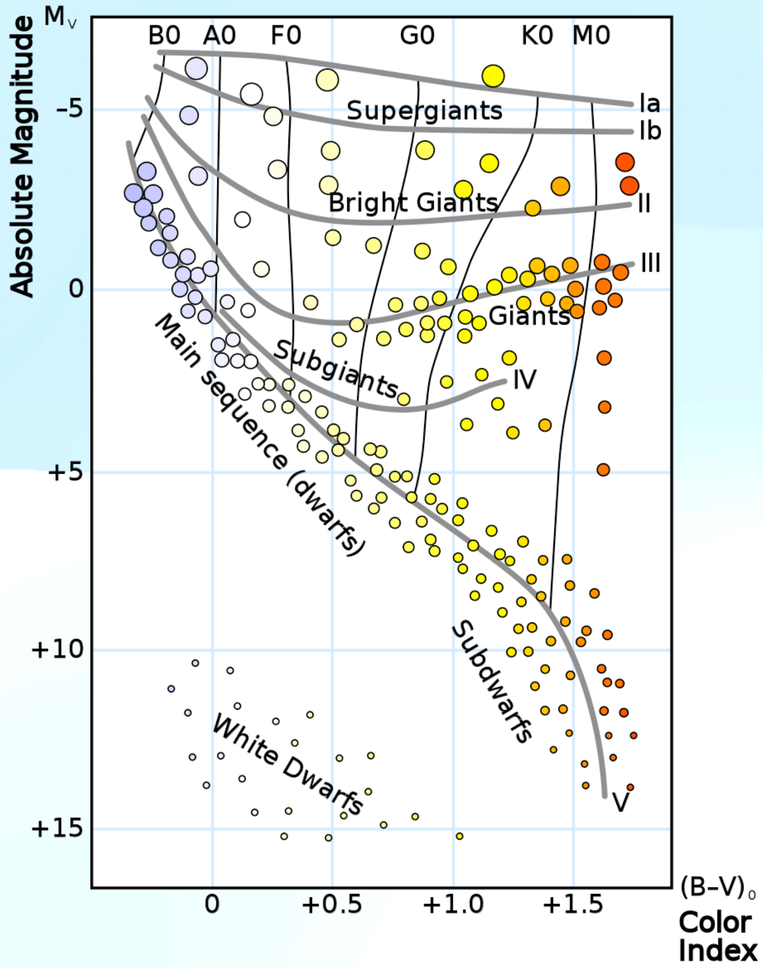




Machine Learning -> Neuroscience

Neuroscience -> Machine Learning

This week’s notebooks data ⭐



	Temperature	Luminosity	Size	A_M	Color	Spectral_Class	Type
0	3068	0.002400	0.1700	16.12	Red	M	Red Dwarf
1	3042	0.000500	0.1542	16.60	Red	M	Red Dwarf
2	2600	0.000300	0.1020	18.70	Red	M	Red Dwarf
3	2800	0.000200	0.1600	16.65	Red	M	Red Dwarf
4	1939	0.000138	0.1030	20.06	Red	M	Red Dwarf

Bonus

Using chatGPT (as your personal TA) for coding