Machine Learning with Python

Wenbin Guo
Bioinformatics IDP, UCLA
wbguo@ucla.edu

2022 Fall

Thanks Dr. Seyoon Ko for some code materials

Notations of the slides

- `Code or Pseudo-Code chunk` starts with "➢", e.g.
  - ➢ `print("Hello world!")`

- Link is underlined

Workshop goals

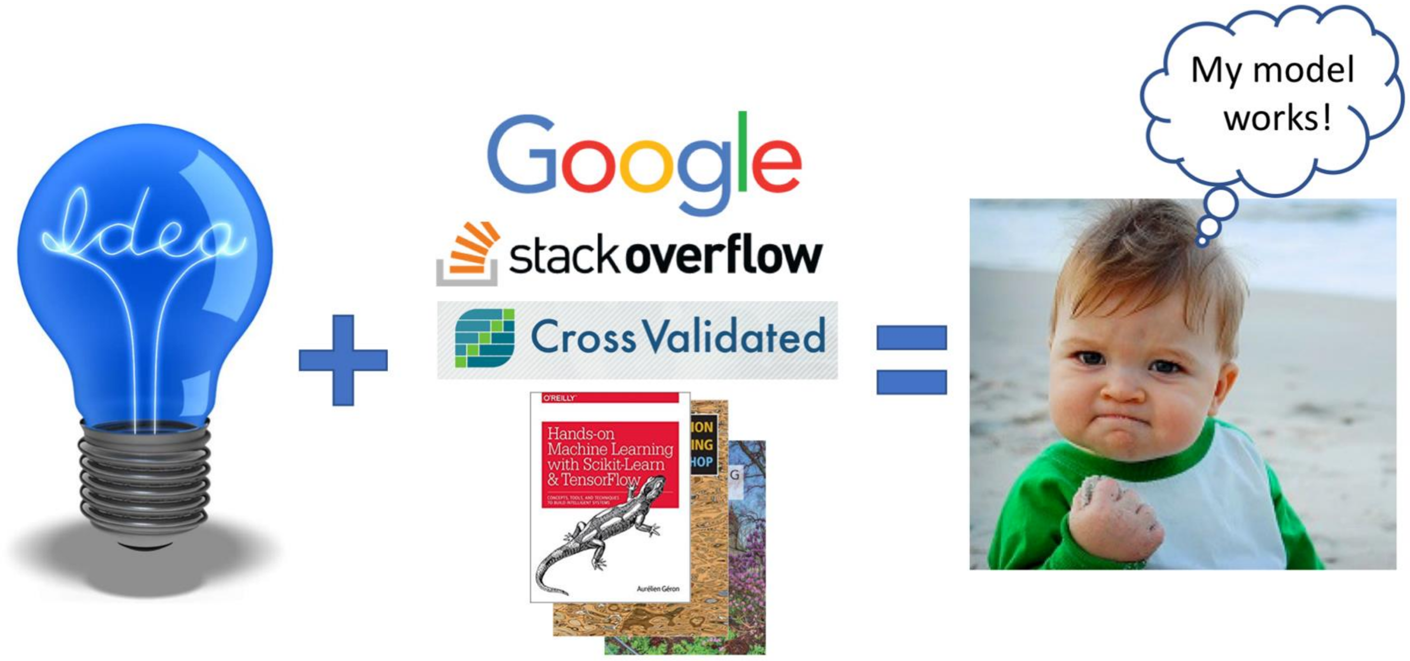
- Understand the rationales of machine learning algorithms

- Know the advantages/limitations of some machine learning algorithms

- Apply machine learning models to solve a problem, and make the model perform better

Workshop goals

- Use online resources to make your idea into a model!



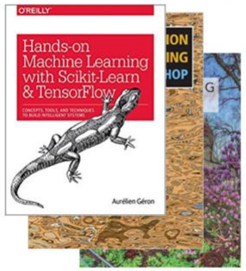
Workshop goals

- Use online resources to make your idea into a model!

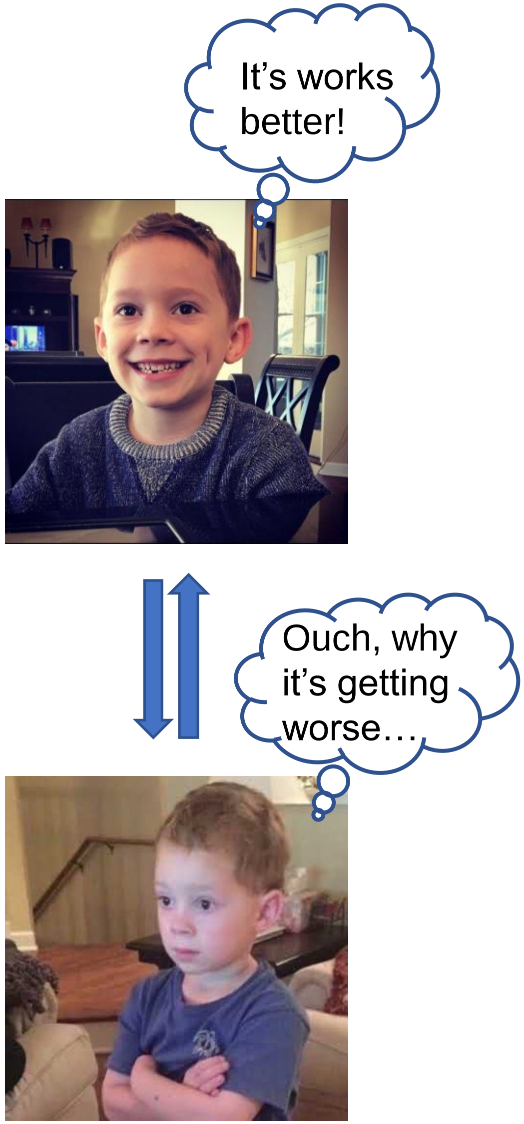
Google

stackoverflow

Cross Validated



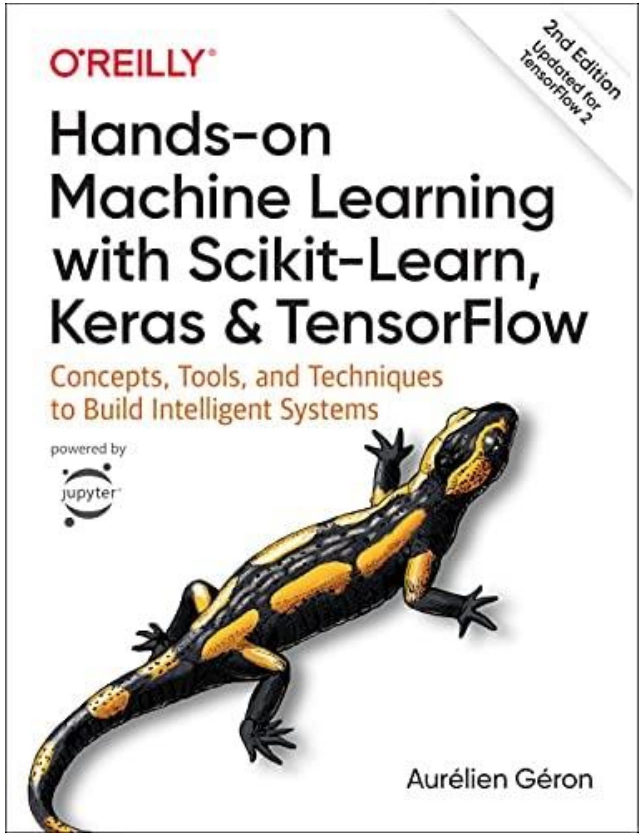
Tune parameters



Agenda

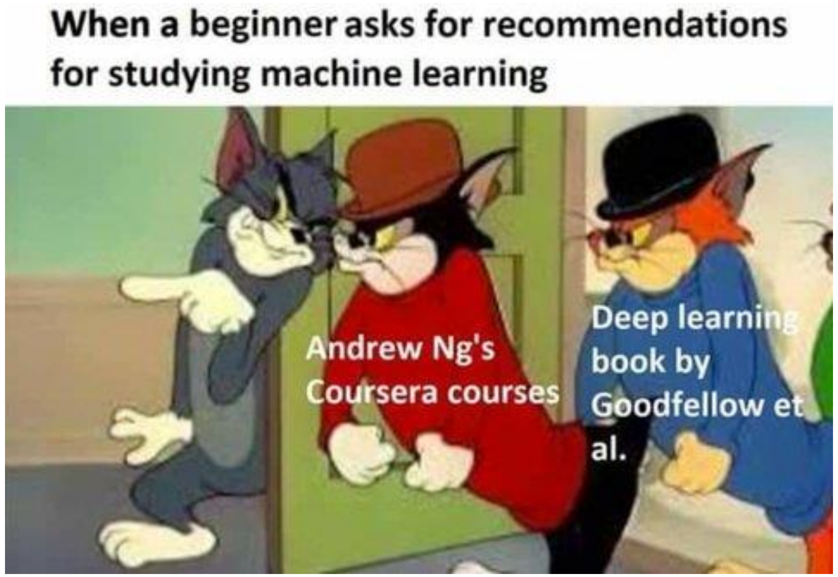
- Day 1: Introduction to machine learning
  - Some key concepts in machine learning
  - Jupyter notebook and some packages usage
- Day 2: Supervised learning
  - Classification
  - Regression
  - Regularization
- Day 3: Unsupervised learning
  - Dimension reduction
  - Clustering

References



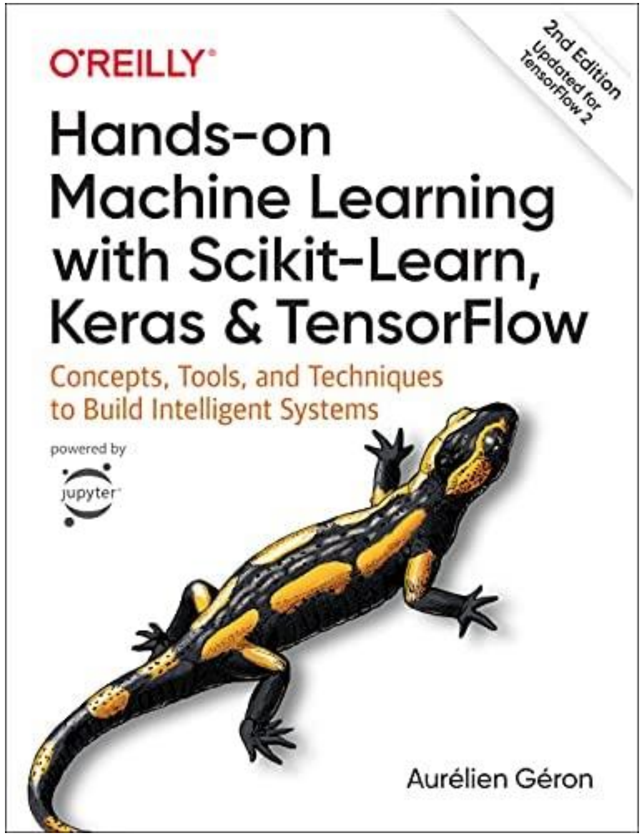
[link]

- Other useful reference



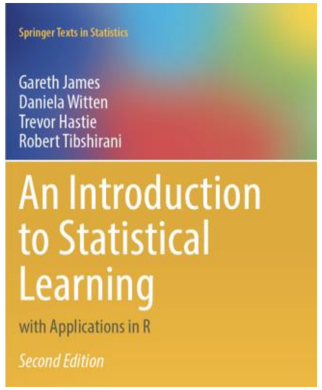
- Write down questions to this [Google doc]

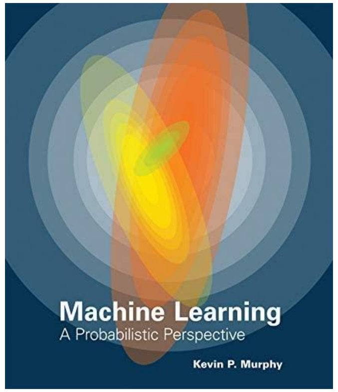
References

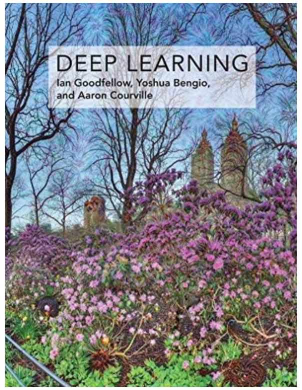


link

- Other useful reference
  - An introduction to statistical learning
  - Machine learning: a probabilistic perspective
  - Deep learning







- Write down questions to this Google doc

Day 1: Introduction to Machine Learning

Wenbin Guo
Bioinformatics IDP, UCLA
wbguo@ucla.edu
2022 Fall

Overview

Time

- 3-hour workshop (45min + 45min + 30min + practice/Q&A)

Topics

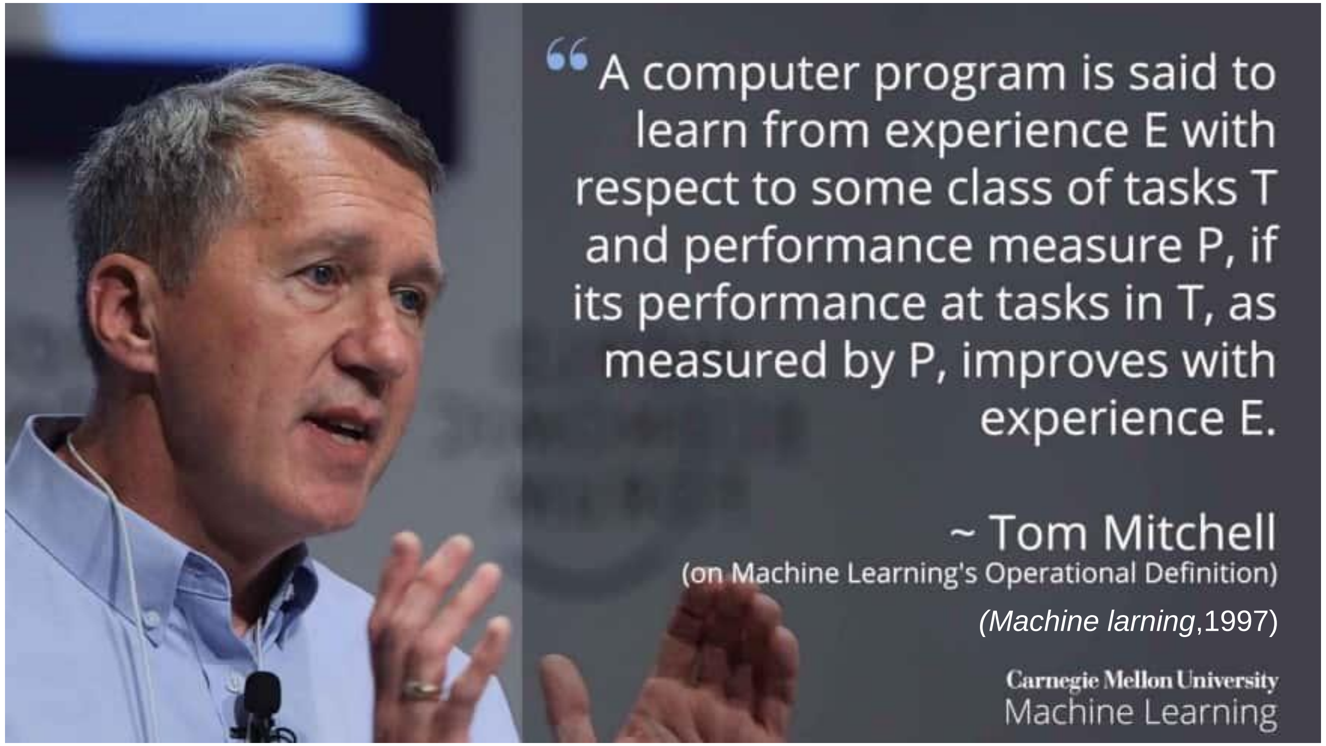
- ❑ Introduction to machine learning
    - What's machine learning?
    - Types of machine learning
    - Machine learning applications
- ❑ Some key concepts in machine learning
- ❑ Recap of useful tools and packages
    - Jupyter notebook
    - NumPy, Matplotlib
- ❑ Examples and practices

What is machine learning?

Machine Learning *the "field of study that gives computers the ability to learn without being explicitly programmed".*

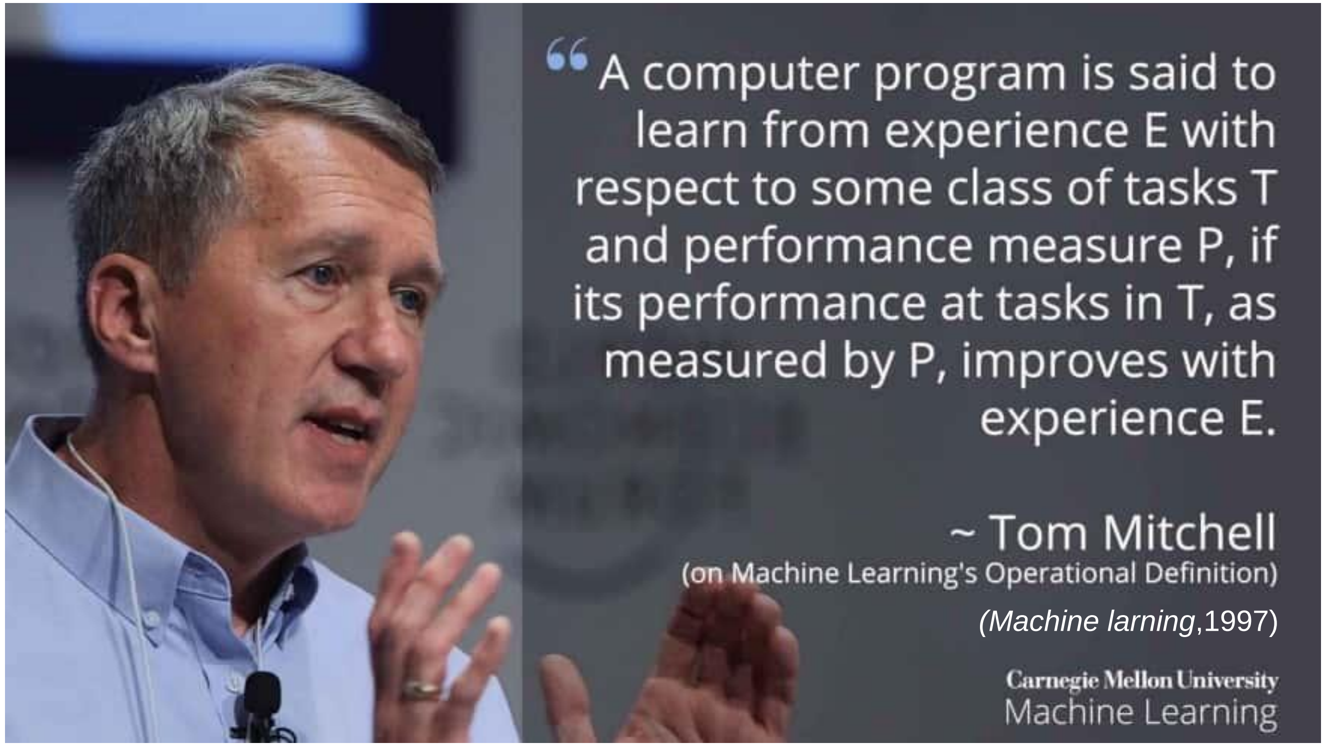
Arthur Lee Samuel (1959)

What is machine learning?



(Machine larning,1997)

What is machine learning?



- Experience
  - Data
- Task
  - Classification
  - Regression
  - Clustering
  - Dimension reduction
  - …
- Performance
  - Entropy loss
  - Mean squared error
  - Reconstruction error
  - …

Why machine learning?

- Some problems with existing traditional solutions require a long list of rules, machine learning can simplify the code and probably perform better

- Some problems either are too complex for traditional approach or have no known algorithms

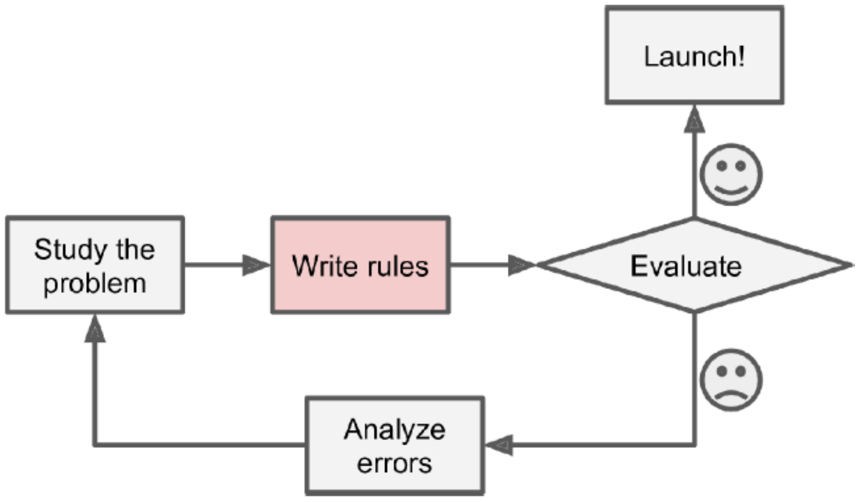


Figure 1-1. The traditional approach

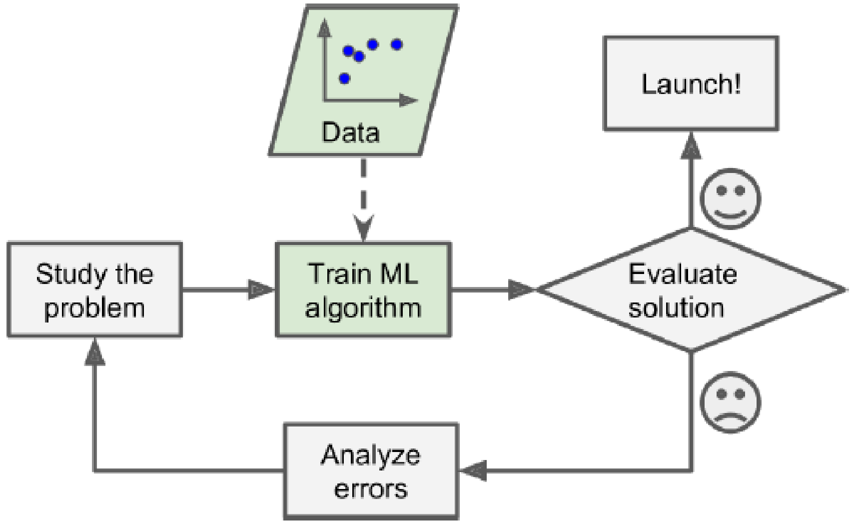


Figure 1-2. The Machine Learning approach

Why machine learning?

- Some problems with existing traditional solutions require a long list of rules, machine learning can simplify the code and probably perform better

- Some problems either are too complex for traditional approach or have no known algorithms

- A machine learning system can adapt to new data

- Machine learning system can help us to get insights about complex problems

…

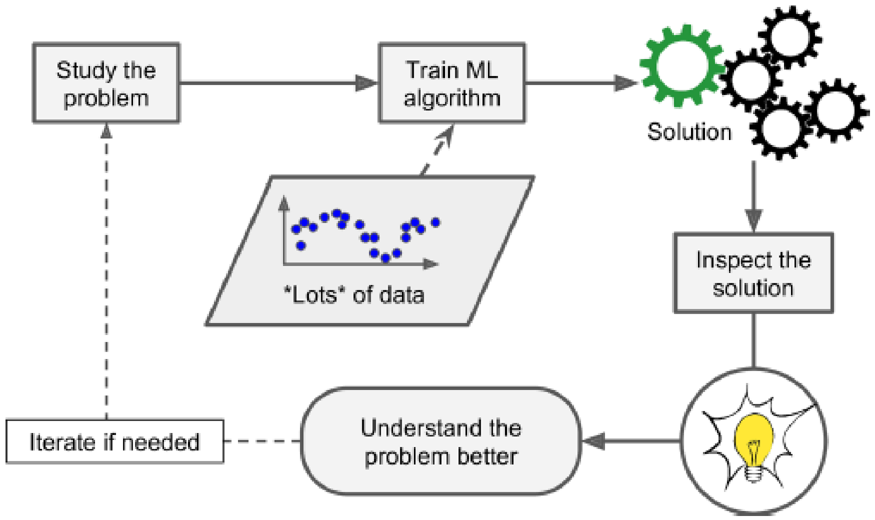
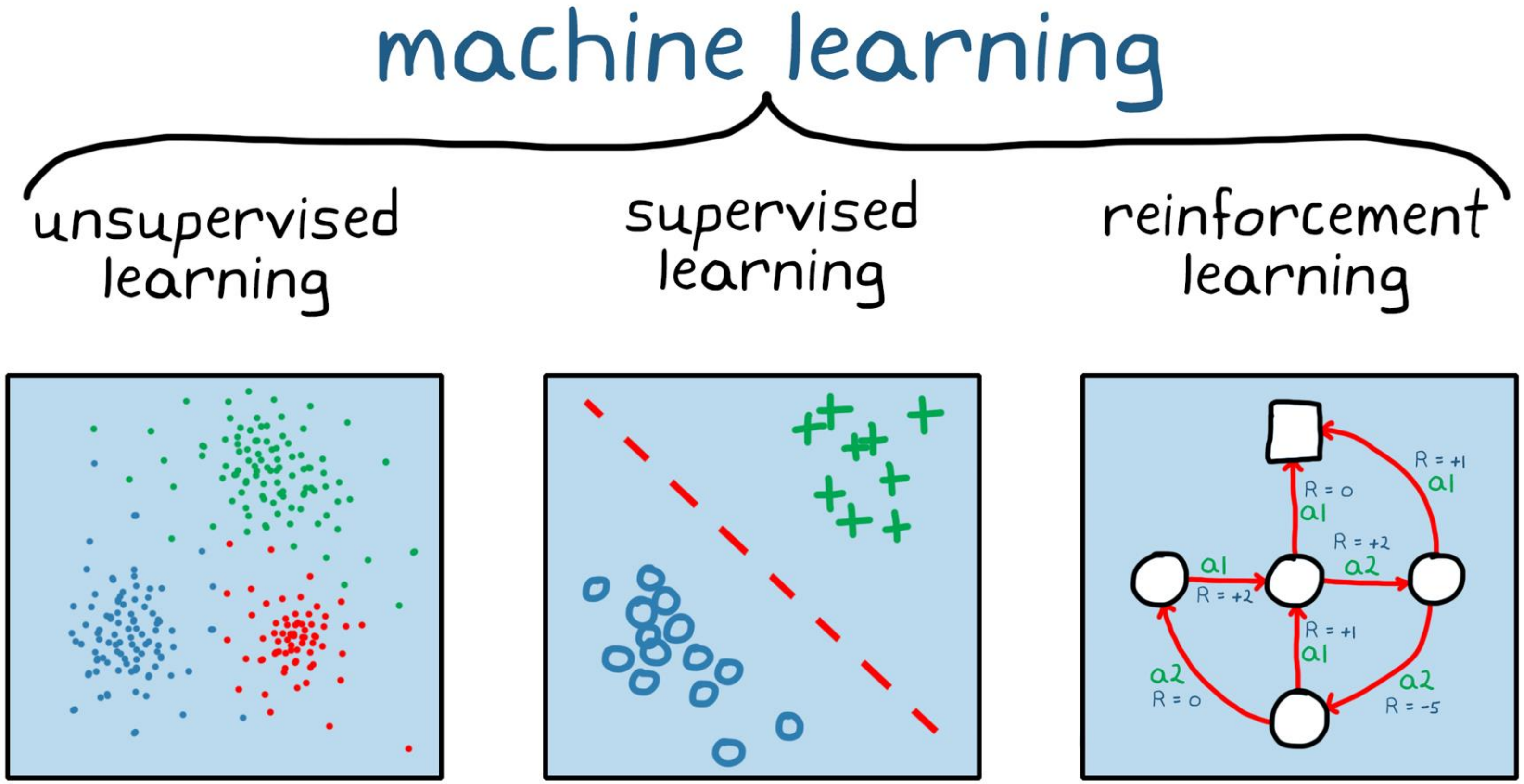


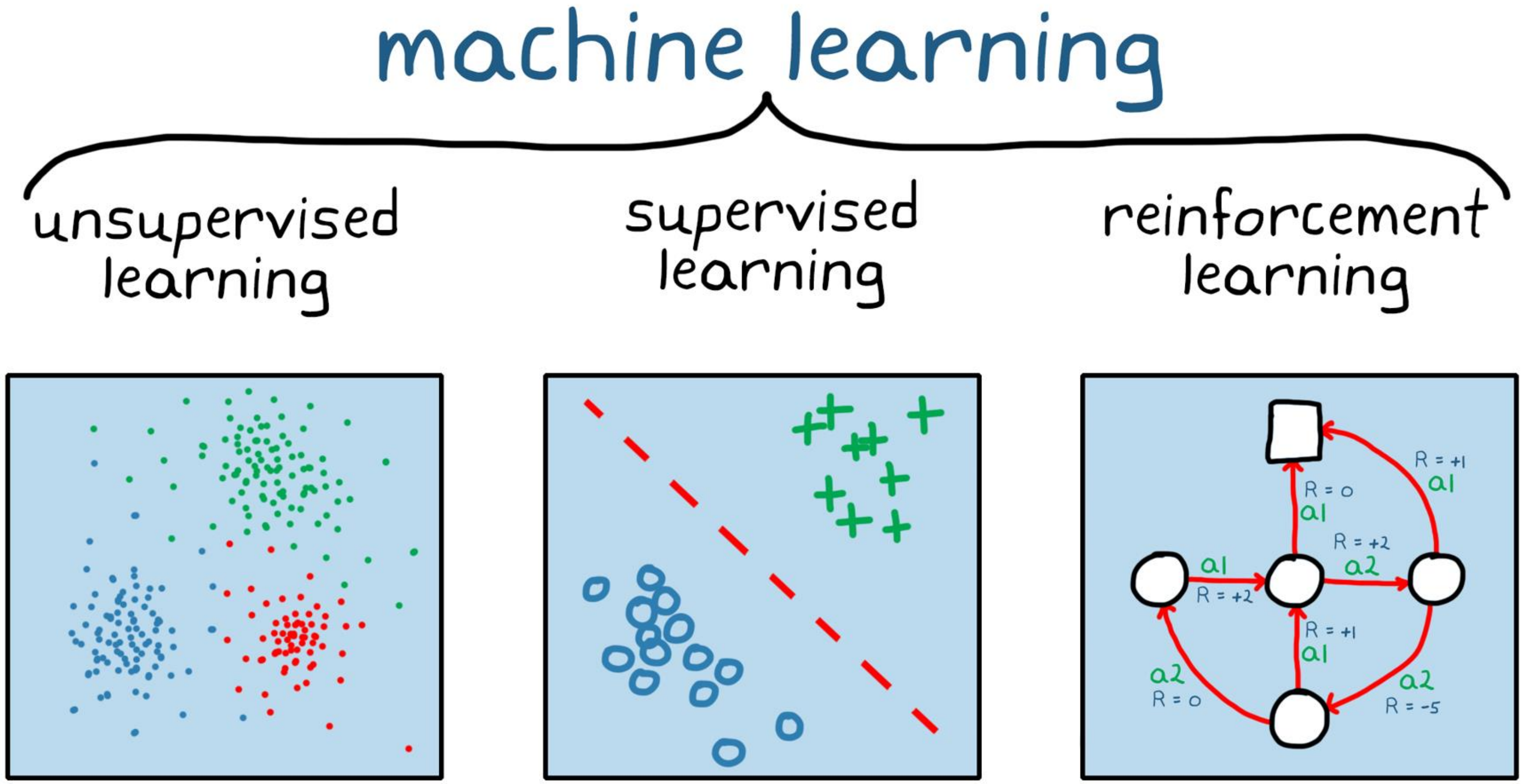
Figure 1-4. Machine Learning can help humans learn

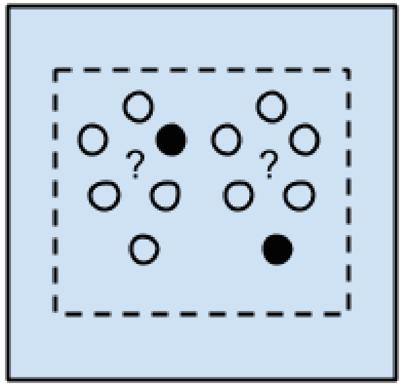
Types of machine learning



https://www.mathworks.com/discovery/reinforcement-learning/_jcr_content/mainParsys3/discoverysubsection/mainParsys/image.adapt.full.medium.png/1663235165820.png

Types of machine learning





Semi-supervised Learning Algorithms

https://www.mathworks.com/discovery/reinforcement-learning/_jcr_content/mainParsys3/discoverysubsection/mainParsys/image.adapt.full.medium.png/1663235165820.png

Other category types

- Batch learning vs Online-learning
  - Batch learning: system is trained using all available data (usually take long, thus is done offline)
  - Online learning: system is trained incrementally by feeding data sequentially, data can be on the fly

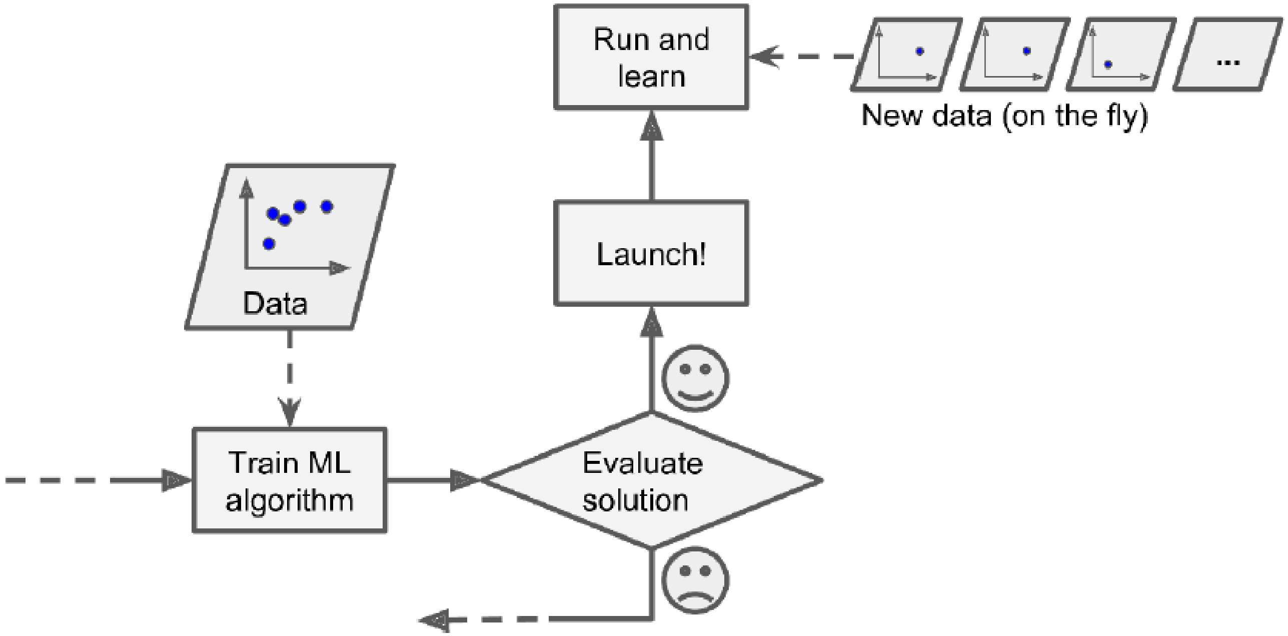


Figure 1-13. In online learning, a model is trained and launched into production, and then it keeps learning as new data comes in

Other category types

- Batch learning vs Online-learning
  - Batch learning: system is trained using all available data (usually take long, thus is done offline)
  - Online learning: system is trained incrementally by feeding data sequentially, data can be on the fly

- Instance-based learning vs model-based learning
  - Instance-based learning: the system learns the example by heart, then generalizes to new cases by using a similarity measure
  - Model-based learning: build a model from existing examples, and use the model to make predictions

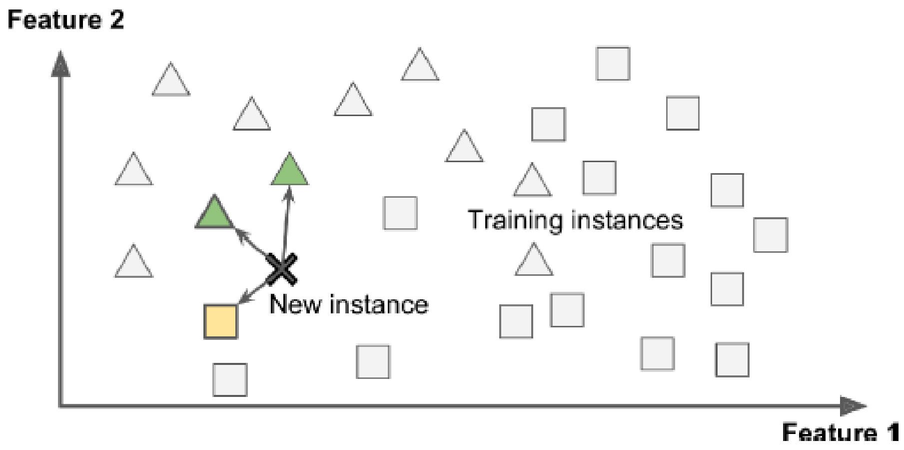


Figure 1-15. Instance-based learning

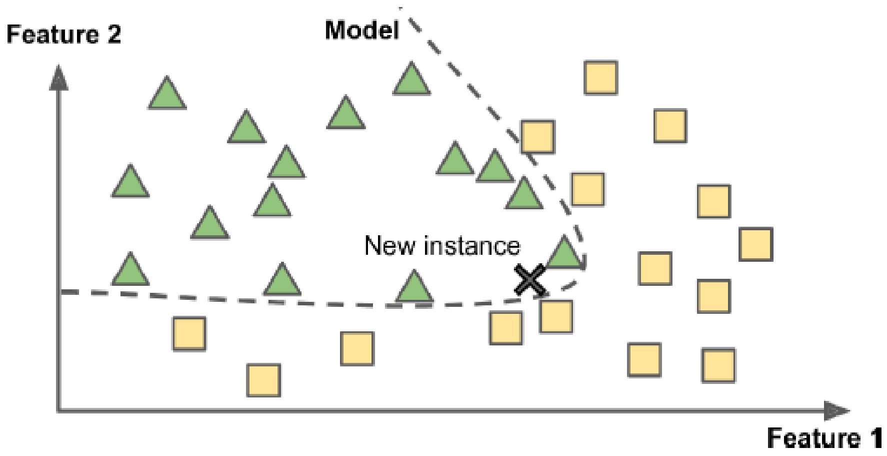
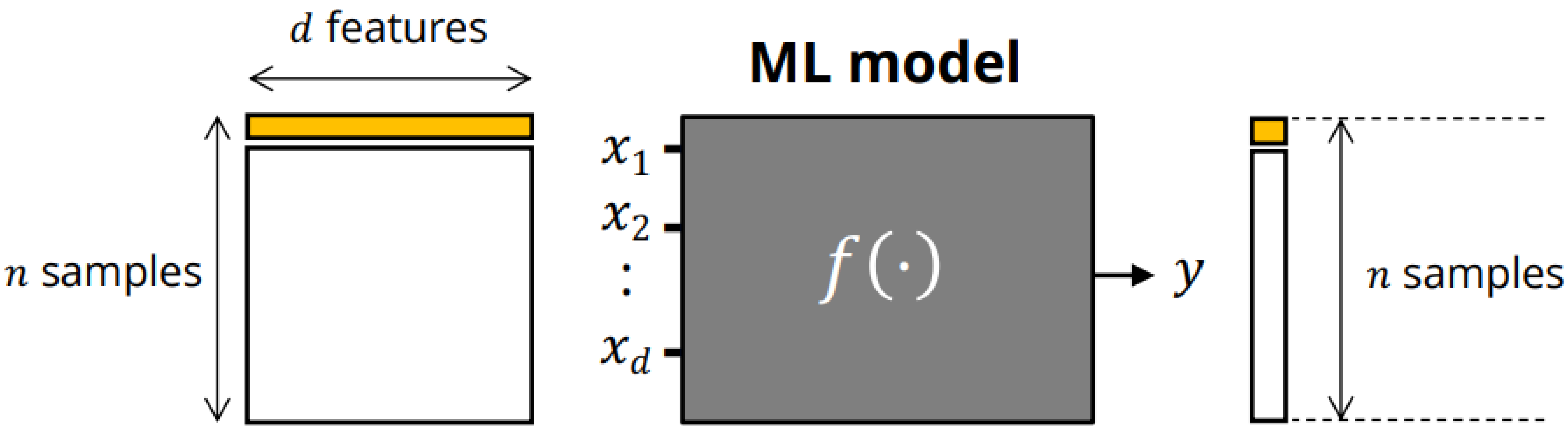


Figure 1-16. Model-based learning

Supervised learning

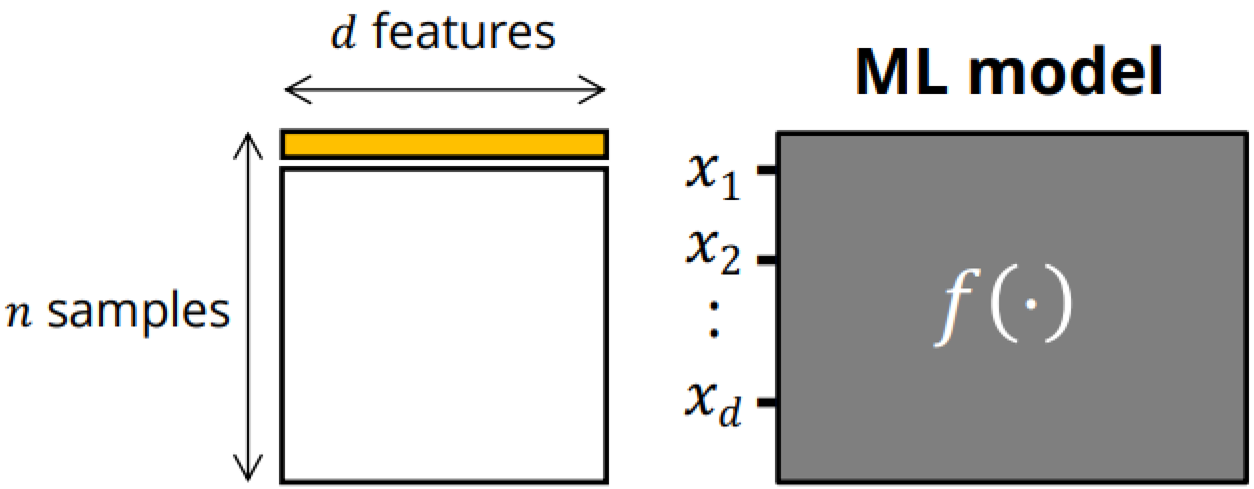
- Training data with n ***samples*** of ***features*** x and ***label*** y
- Learn a function class f(x) to describe y based on x

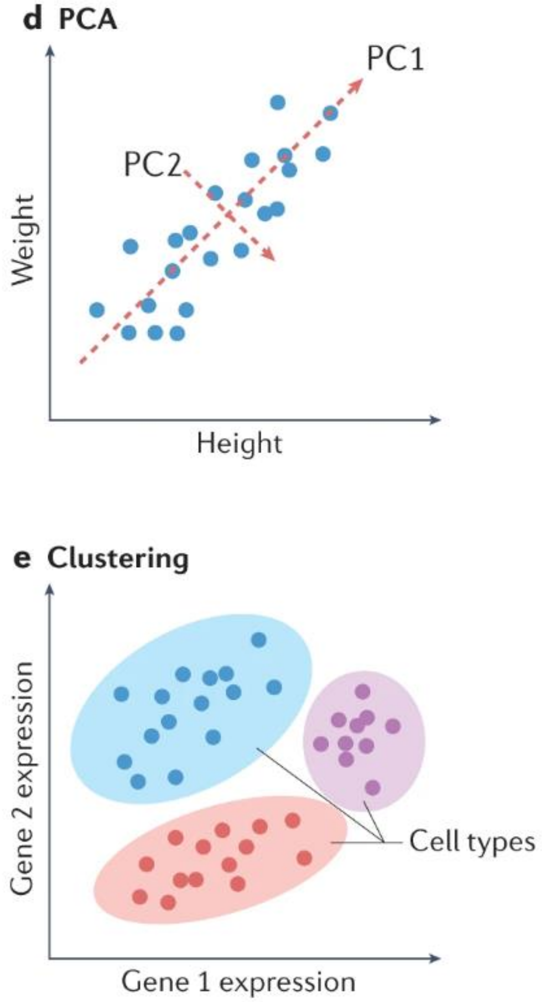


Figures is adopted from CSEP 590B Explainable AI

Unsupervised learning

- Training data with n ***samples*** of ***features*,** **no label**
- Identify patterns in unlabelled data

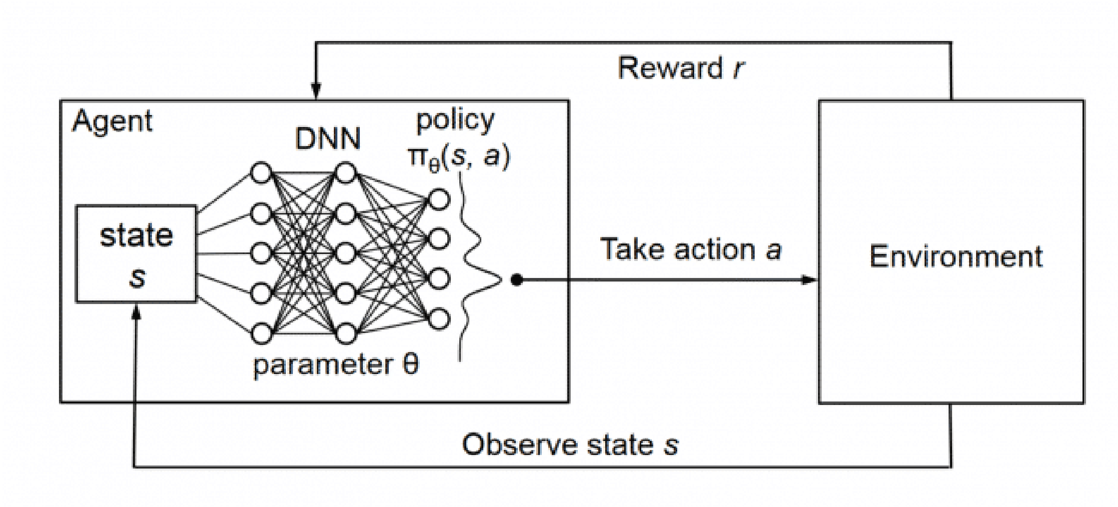


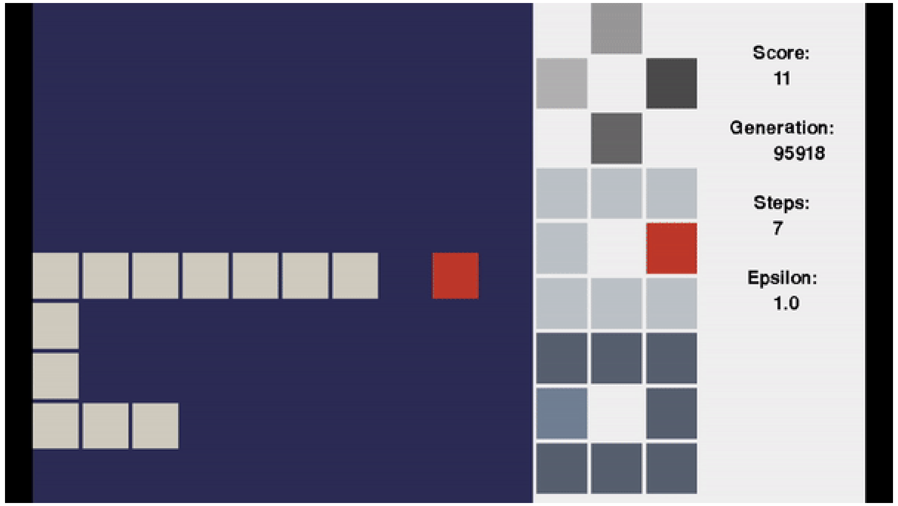


Greener et.al, Nature reviews molecular cell biology (2021)

Reinforcement learning

- Learning system (agent) observe the *environment*, select and perform *actions*, and get *rewards* in return
- Learn by itself what is the best strategy (*policy*), to get the most reward over time





https://www.novatec-gmbh.de/en/blog/deep-q-networks/
https://github.com/ludius0/SnakeQ

What can machine learning do?

- Face/Speech recognition
- Recommendation system
- Machine translation
- Self-driving system

- Stock market prediction

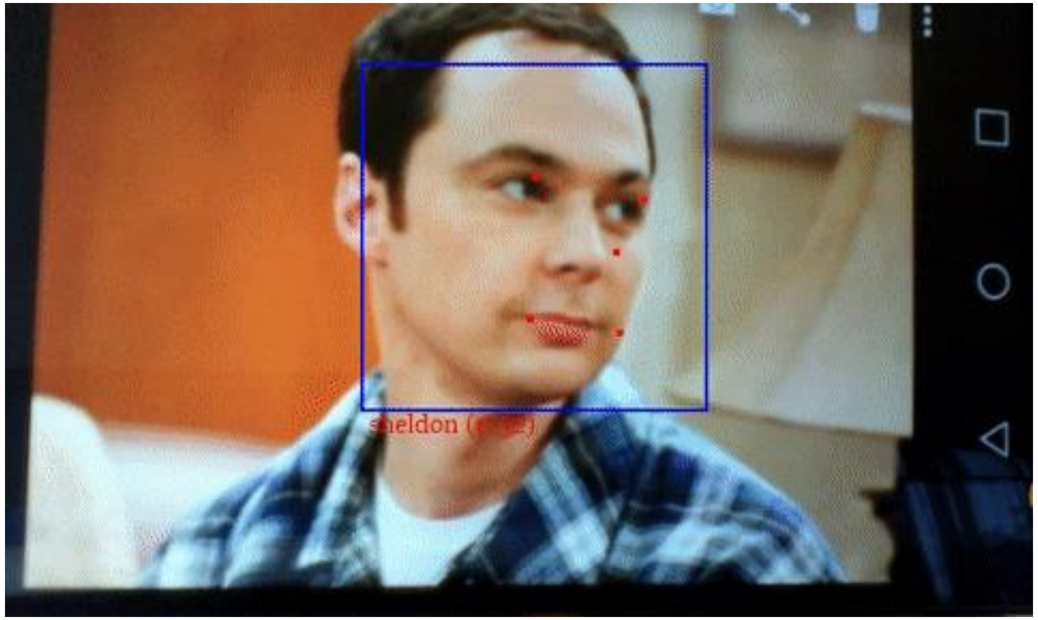
- Create images/songs/paintings/stories

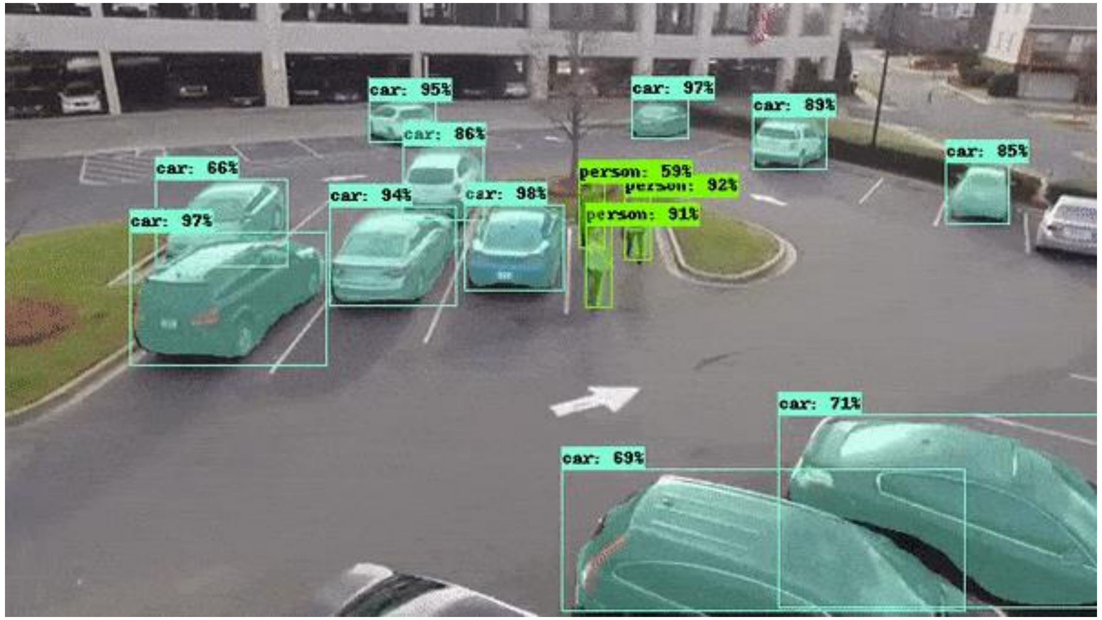
…

Images from internet and [STATS231A]

More examples

- Object segmentation and recognition



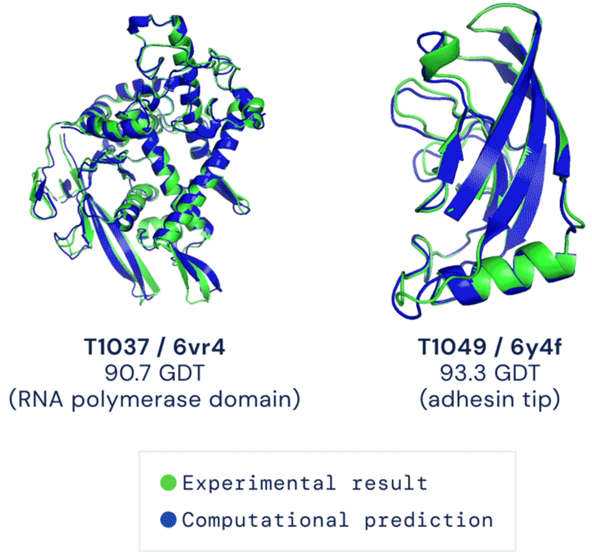


More examples

AlphaGo (2016)

Monte Carlo tree search

AlphaFold (2021)



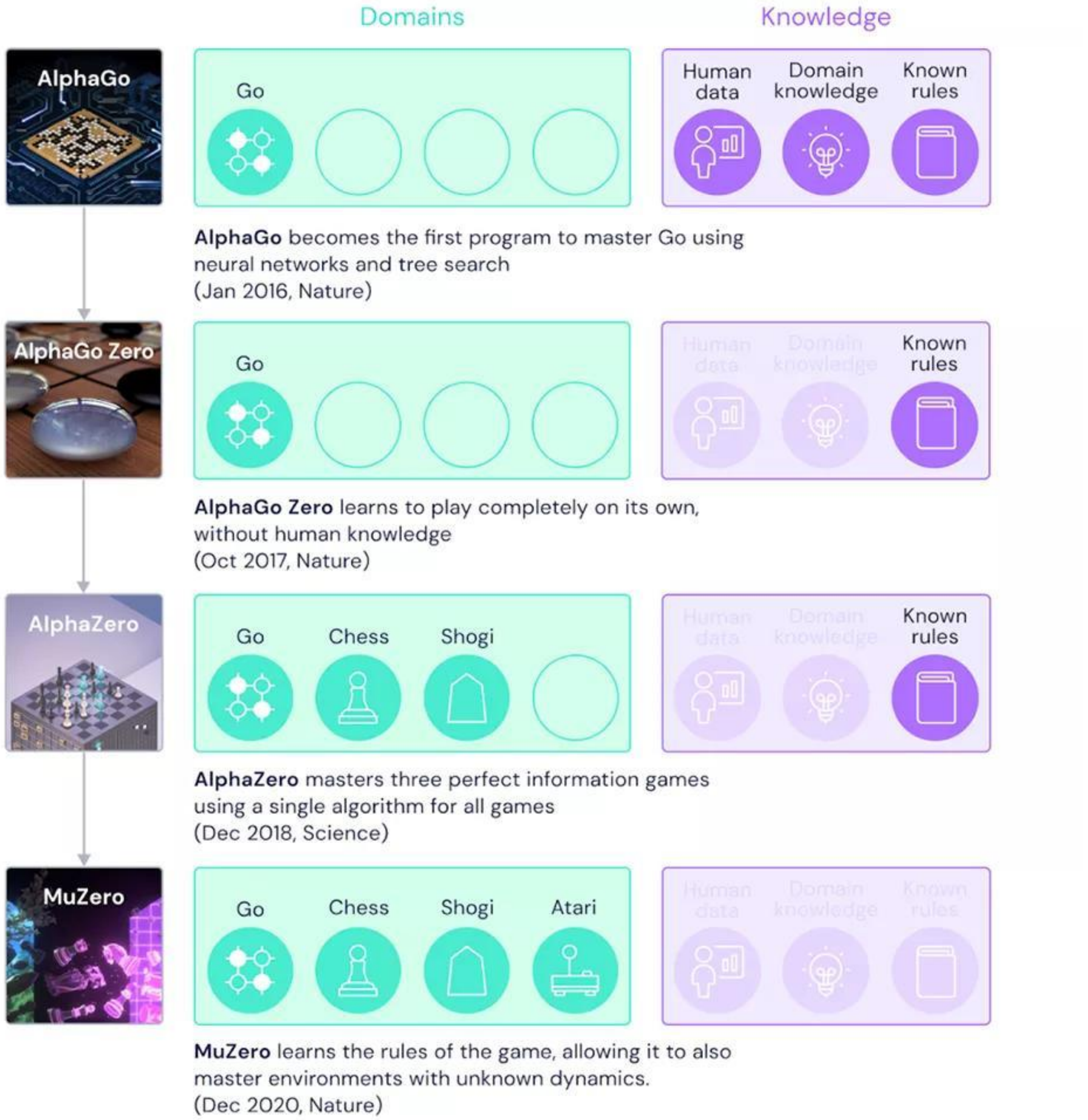
Transformer (self-attention)

AlphaTensor (2022)

Reinforcement learning

Images from internet

AlphaGo keeps evolving



Machine learning in biomedical research

- Imaging/Sequencing enhancement
- Protein 3D structure prediction
- Genetic risk assessment
- Disease diagnosis
- Drug design

…

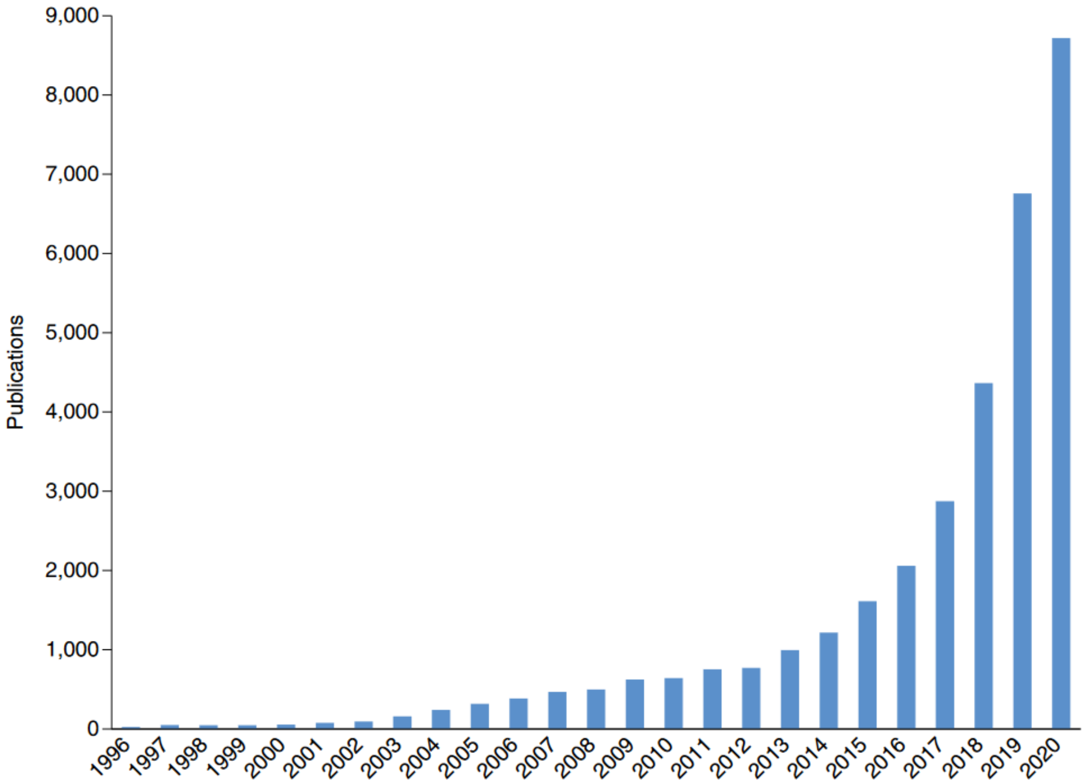


Fig. 1 | Exponential increase of ML publications in biology. The number of ML publications per year is based on Web of Science from 1996 onwards using the topic category for "machine learning" in combination with each of the following terms: "biolog*", "medicine", "genom*", "prote*", "cell*", "post translational", "metabolic" and "clinical".

Walsh et.al, Nat Methods (2021)

Train a machine learning model: a big picture

Greener et.al, Nature reviews molecular cell biology (2021)

Train a machine learning model: a big picture

- Frame the problem
  - Supervised or unsupervised?
  - Classification or regression?

  …

- Select the performance measure (loss function):
  - Regression: RMSE, MAE
  - Classification: cross entropy
  - Dimension reduction: reconstruction error

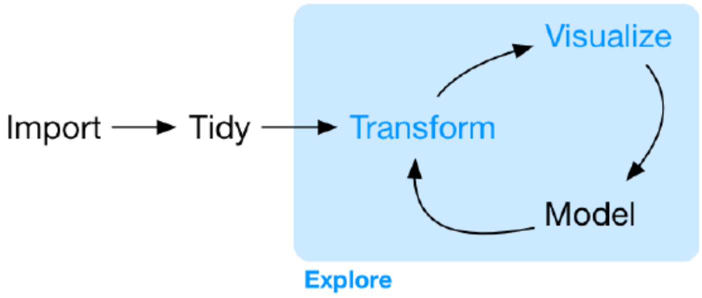
  …

- Cost / loss function measures how *bad* your model fits the data

Train a machine learning model: a big picture

- Prepare the data
  - Download / Gather samples
  - Explore the data structure (descriptive analysis)
  - Perform Data processing, feature engineering

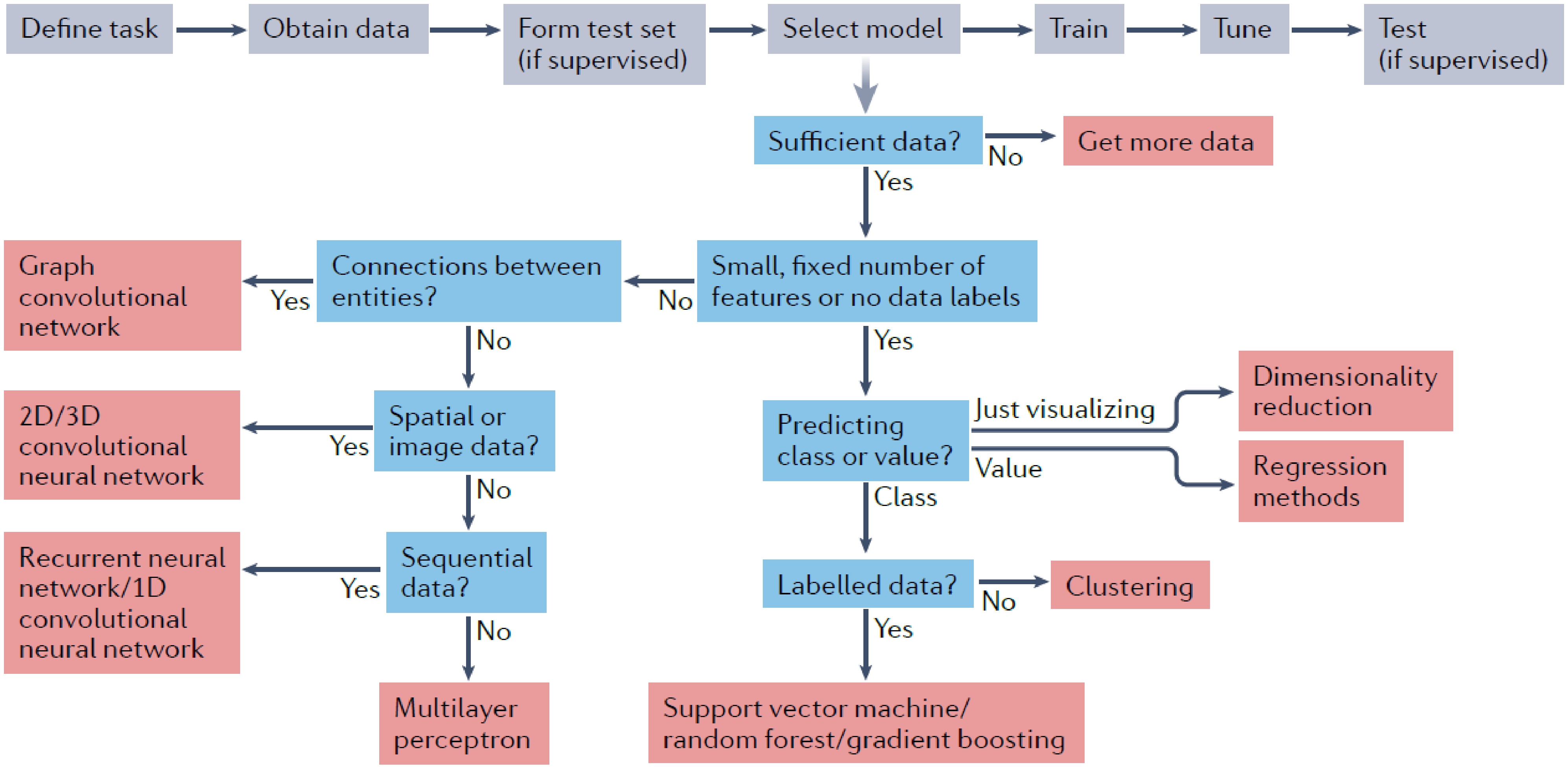
    …



- Create a test set (if supervised)
  - The only way to know how well a model will generalize to new cases is to actually try it out on new cases
  - Split the data into training and test set
    - Training set: train your models
    - Test set: evaluate the model (estimate the generalization error)
    - usually use 80% samples for training, holdout 20% samples for test

R for data science, 2nd edition

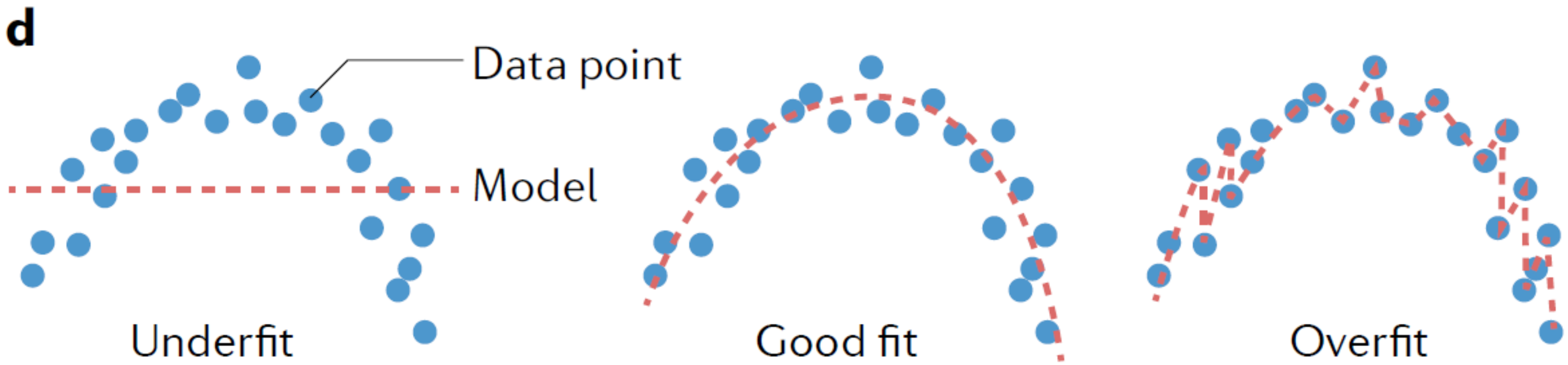
Train a machine learning model: a big picture



Greener et.al, *Nature reviews molecular cell biology* (2021)

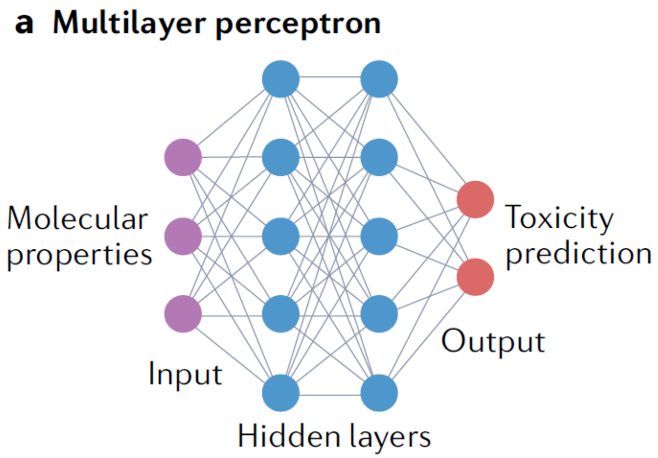
Train a machine learning model: a big picture

- Train a model:
  - run an algorithm to find the model parameters that will make it best fit the training data (and hopefully make good predictions on new data)



Train a machine learning model: a big picture

- Tune hyperparameters:
  - A hyperparameter is a parameter of a learning algorithm (not of the model)
  - Unlike model parameters, hyperparameters are not are not updated during training (although they are adjustable)



Example of hyperparameters:
- Number of neurons per layer
- Number of hidden layers

…

Train a machine learning model: a big picture

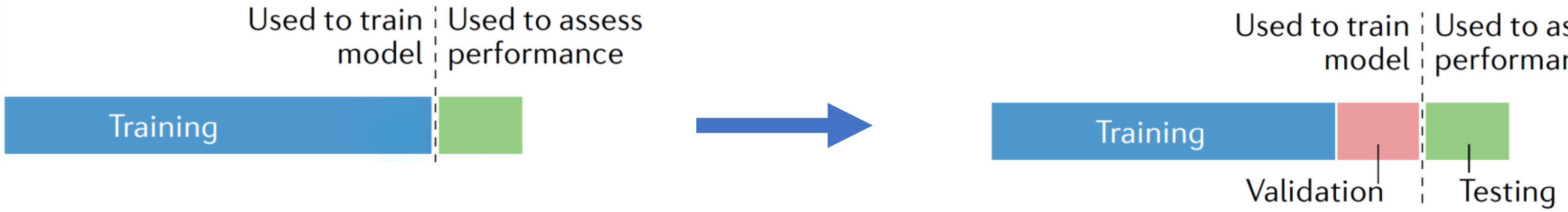
- Consider you have a series hyperparameters/models to try
  - For each of them, you trained the model on training dataset, and then evaluate it on the test dataset, then you pick out the model with the best performance
  - Any problem?

Train a machine learning model: a big picture

- Consider you have a series hyperparameters/models to try
  - For each of them, you trained the model on training dataset, and then evaluate it on the test dataset, then you pick out the model with the best performance
  - Any problem?
  - You measured the generalization error multiple times on the test set, and you adapted the model and hyperparameters to produce the best model *for that particular set*

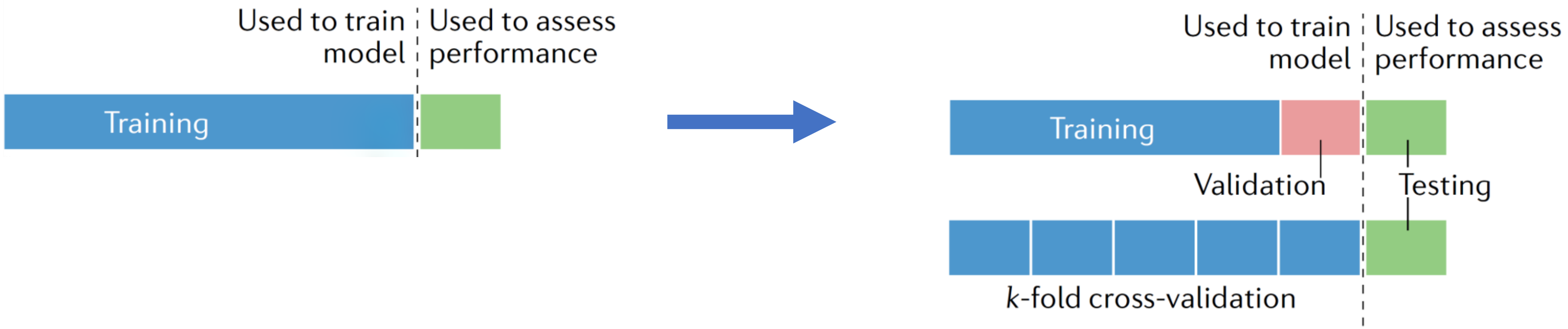
Train a machine learning model: a big picture

- Consider you have a series hyperparameters/models to try
  - For each of them, you trained the model on training dataset, and then evaluate it on the test dataset, then you pick out the model with the best performance
  - Any problem?
  - You measured the generalization error multiple times on the test set, and you adapted the model and hyperparameters to produce the best model *for that particular set*
  - Solution: *holdout validation*



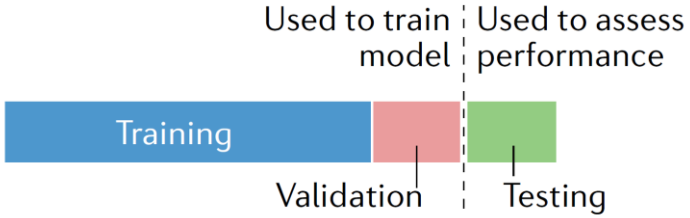
Train a machine learning model: a big picture

- Consider you have a series hyperparameters/models to try
  - For each of them, you trained the model on training dataset, and then evaluate it on the test dataset, then you pick out the model with the best performance
  - Any problem?
  - You measured the generalization error multiple times on the test set, and you adapted the model and hyperparameters to produce the best model *for that particular set*
  - Solution: *holdout validation*, typically use cross validation when validation set is small



Train a machine learning model: a big picture

- Tune hyperparameters
  - Train multiple models with various hyperparameters on the reduced training set
  - Select the model that performs best on the validation set (or repeated cross validation)
  - Train the best model on the full training set (including the validation set), this gives you the final model
  - Lastly, evaluate this final model on the test set to get an estimate of the generalization error



Main challenges for machine learning

The system will not perform well if your training set is too small, or if the data is not representative, is noisy, or is polluted with irrelevant features

- Insufficient data
  - It takes a lot of data for most Machine Learning algorithms to work properly
  - but small- and medium-sized datasets are still very common in biological research

- Nonrepresentative data (sampling bias)
  - Use a training set that is representative of the cases you want to generalize to

Main challenges for machine learning

The system will not perform well if your training set is too small, or if the data is not representative, is noisy, or is polluted with irrelevant features

- Insufficient data
  - It takes a lot of data for most Machine Learning algorithms to work properly
  - but small- and medium-sized datasets are still very common in biological research

- Nonrepresentative data (sampling bias)
  - Use a training set that is representative of the cases you want to generalize to

- Poor quality data (hard to detect patterns under noise)
  - Spend time cleaning up your training data
  - Remove clear outliers instance or try to fix the errors manually

- Irrelevant features (garbage in, garbage out)
  - Feature selection: selecting the most useful features to train on among existing features
  - Feature extraction: combing existing features to produce a more useful one
  - Creating new features

Summary: What have we covered so far

Key concepts in machine learning:

- ❑ What’s machine learning
- ❑ 3 types of machine learning
- ❑ The big picture of training a machine learning model

More details about:

- ❑ Training/test set
- ❑ Loss function
- ❑ Overfitting/underfitting
- ❑ Hyperparameters tunning (model selection)
- ❑ Cross validation
- ❑ Challenges in machine learning

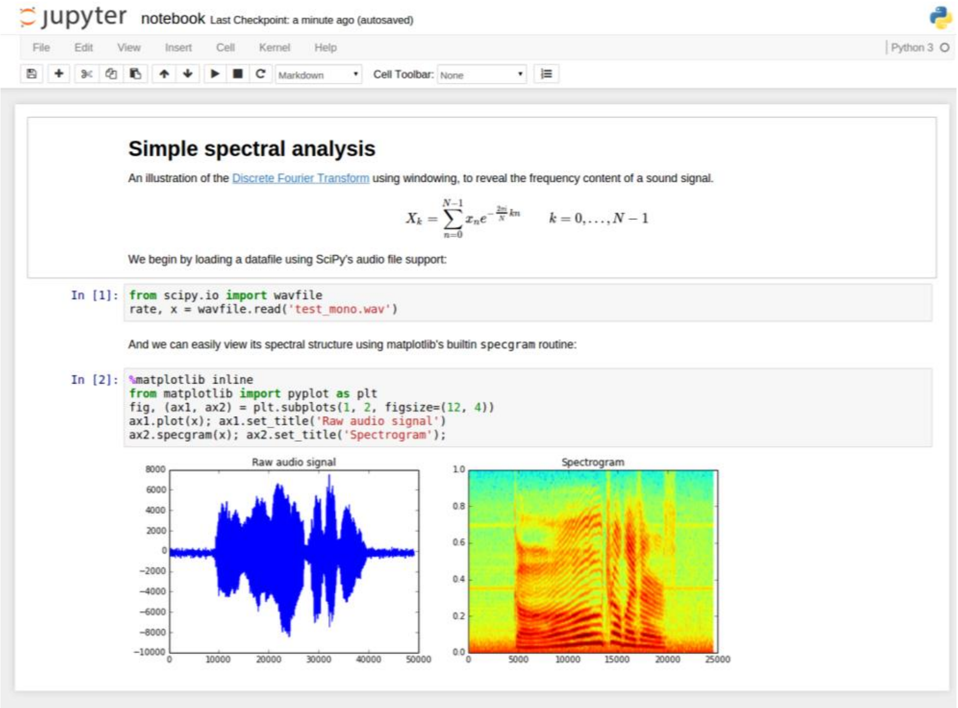
Enough theory for today, let’s do some practice!

- Jupyter notebook
- Usage of some popular packages
- A toy example in machine learning

Jupyter notebook

- Jupyter notebook is an open-source, web-based interactive computing platform
  - Suitable for exploration, and prototyping
  - Convenient in sharing, documentation and making plots, powerful for data analysis

- Similar applications:
  - JupyterLab (an extension of Jupyter notebook)
  - R markdown Notebook (R)
  - Matlab Live Script (Matlab)



- Fun fact: Jupyter stands for **Ju**lia, **Pyth**on, **R**. The native language is Python, but you can install kernels for other languages

Preparation: Install Jupyter notebook

- Install on your local computer: [link]

Install the classic Jupyter Notebook with:

```
pip install notebook
```

To run the notebook:

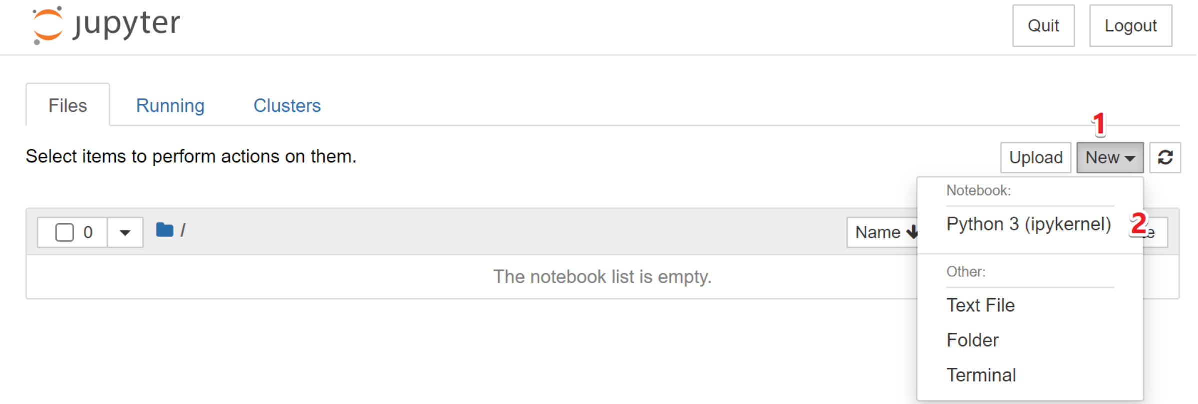
```
jupyter notebook
```

- If you are on Hoffman2: [link]
  - wget https://raw.githubusercontent.com/rdauria/jupyter-notebook/main/h2jupynb
  - chmod u+x h2jupynb
  - ./h2jupynb --help
  - ./h2jupynb -u wbguo -t 4 -m 8 –v 3.9.6

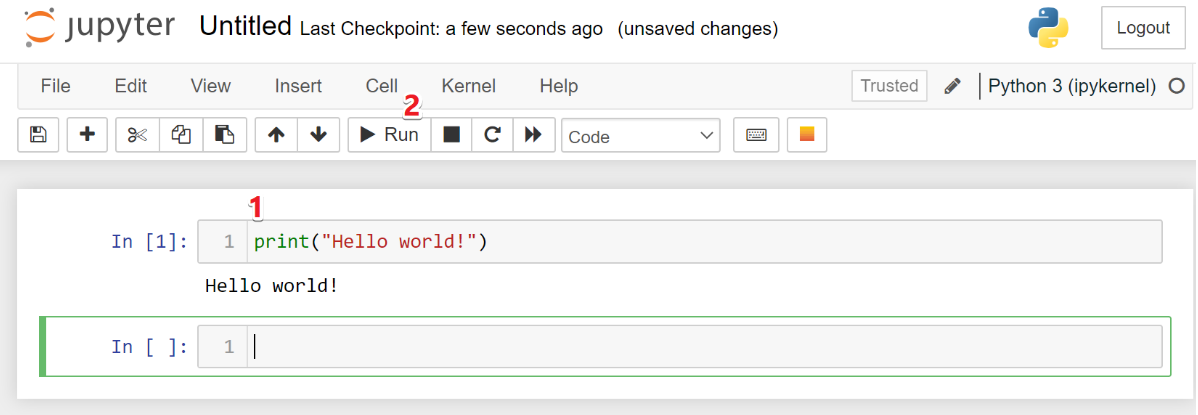
- Alternatively, you can use Google colab

Preparation: Create a new notebook

- Create a new notebook



- Say “Hello world!”



- Now let’s move to the Jupyter notebook
  - ➢ `git clone https://github.com/wbvguo/qcbio-ML_w_Python.git`

Where to get help?

- https://www.google.com
- https://stackoverflow.com
- https://stats.stackexchange.com/
- https://towardsdatascience.com/

Q&A

Google docs