

Machine Learning with Python

Wenbin Guo
Bioinformatics IDP, UCLA
wbguo@ucla.edu
2023 Spring

Notation of the slides

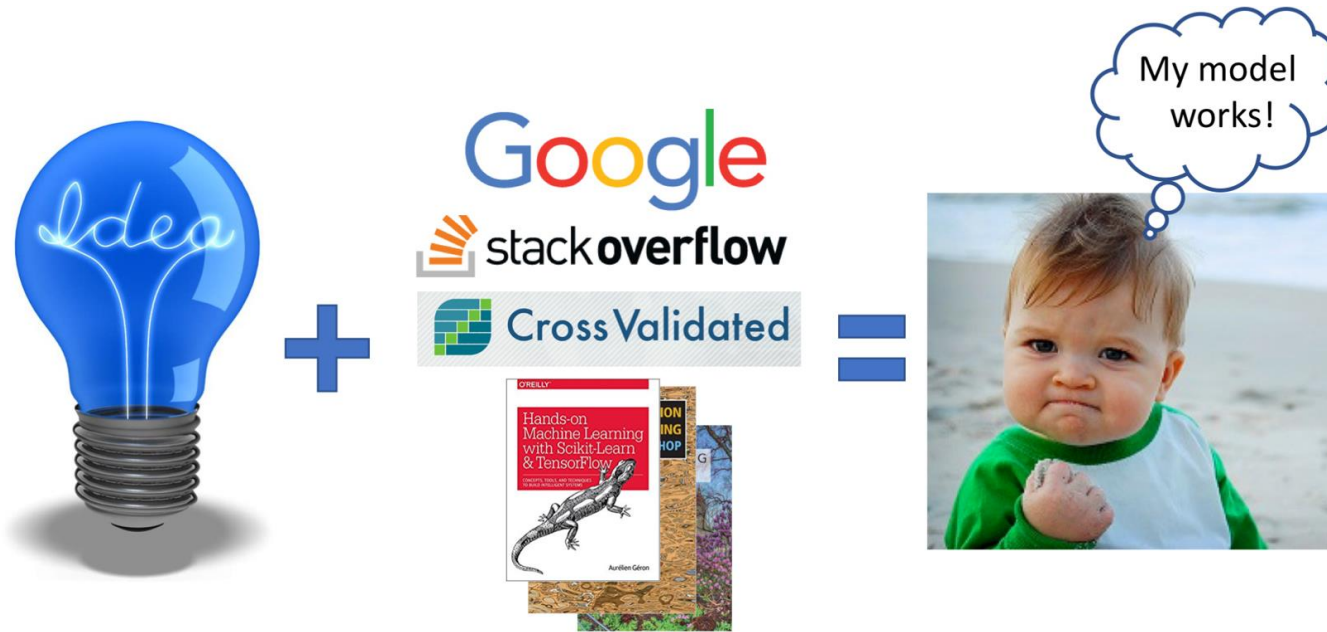
- Code or Pseudo-Code chunk starts with "➤", e.g.
➤ `print("Hello world!")`
- [Link](#) is underlined
- Important terminology is in **bold** font

Workshop goals

- Understand the rationales behind machine learning algorithms
- Know the advantages/limitations of some machine learning algorithms
- Apply machine learning algorithms to solve problems, and make the models perform better

Workshop goals

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

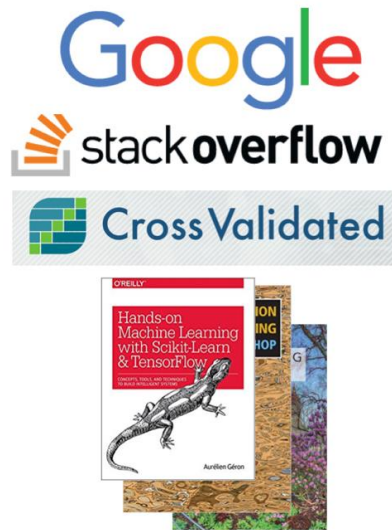


Workshop goals

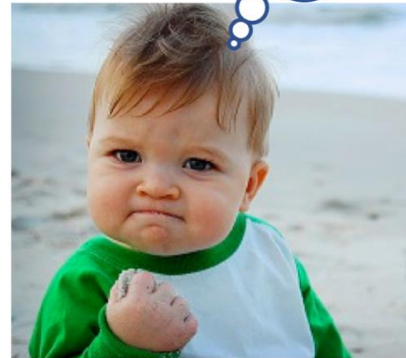
- Use online resources to transform your idea into a model!



+



=

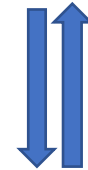


My model works!

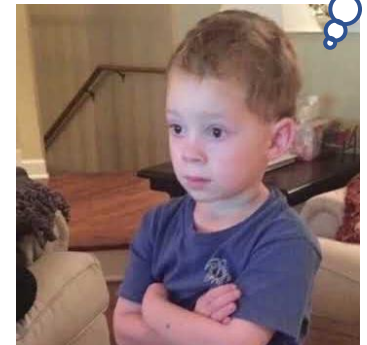
Tune parameters



It works better!



Ouch, why it's getting worse...

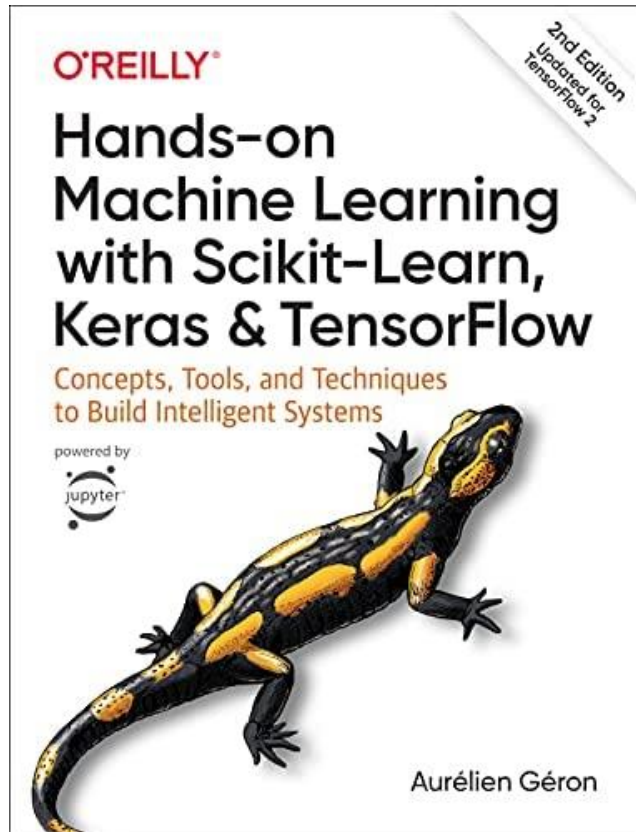


Agenda

- 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](#)

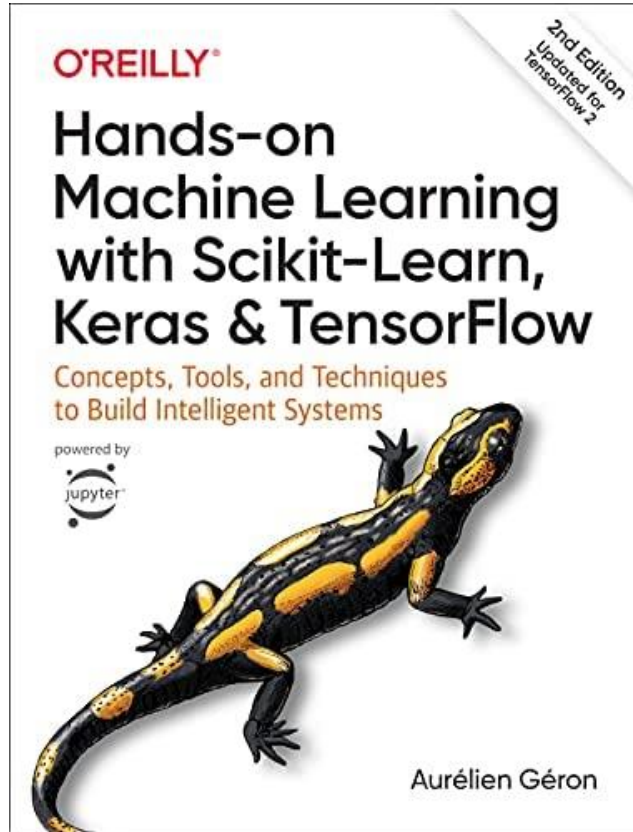
- Other useful reference

When a beginner asks for recommendations for studying machine learning



- Write down questions to this [Google doc](#)

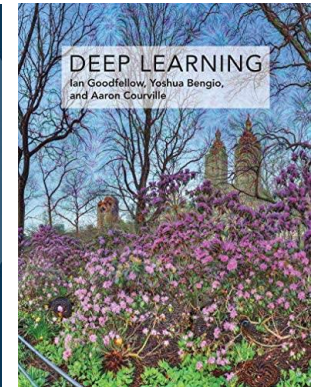
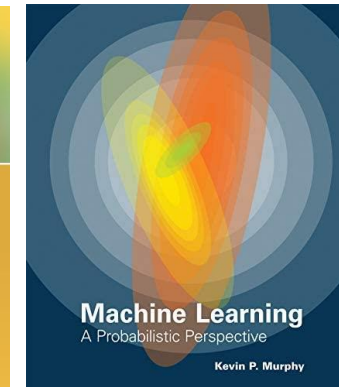
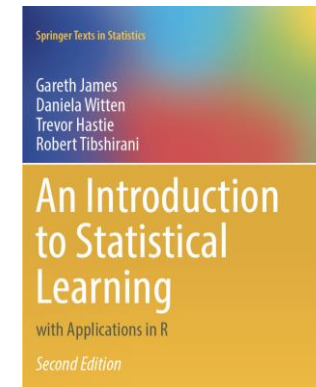
References



[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
2023 Spring

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?



Arthur Lee Samuel (1959)

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

What is machine learning?

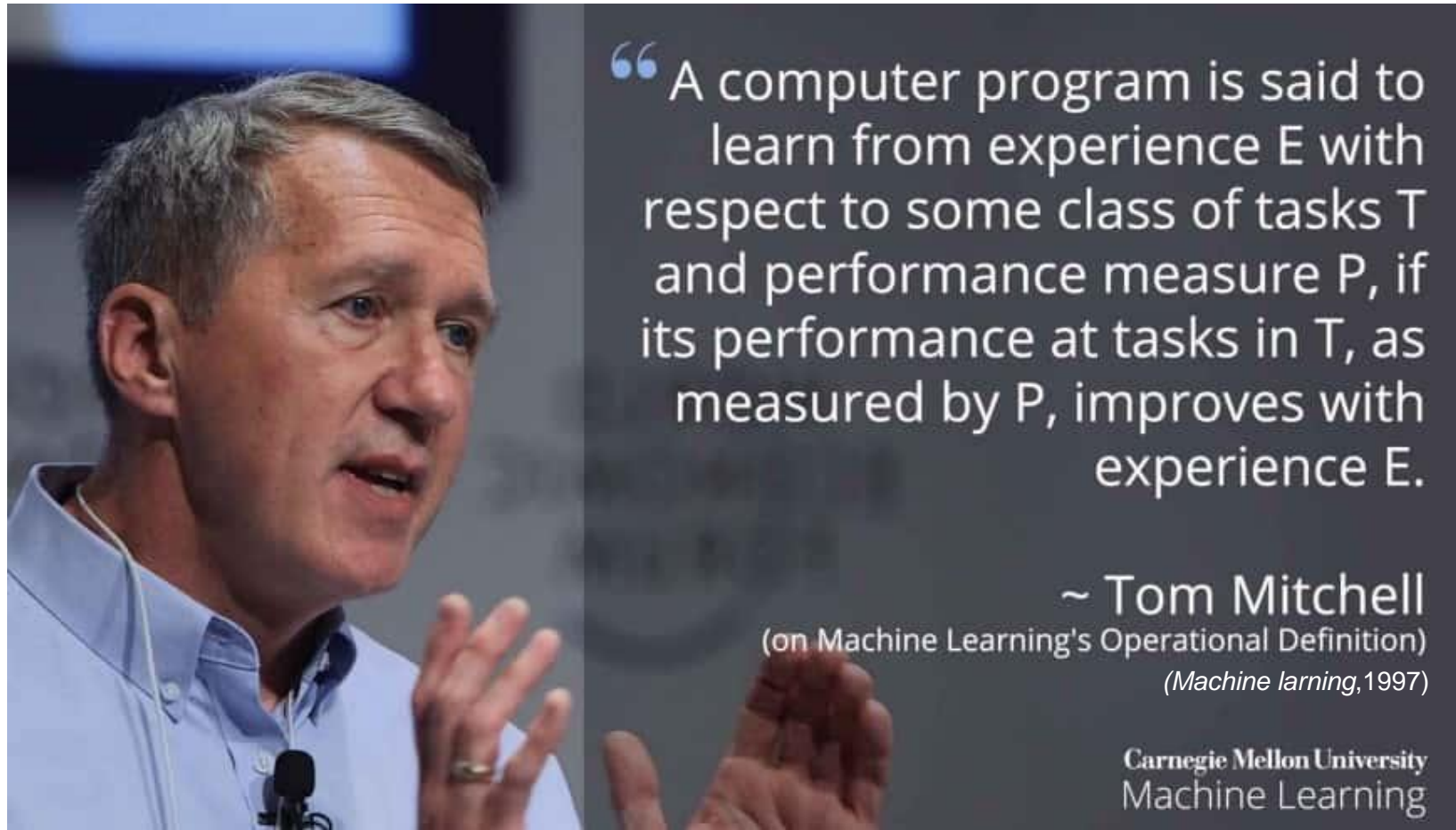
A photograph of Tom Mitchell, a man with grey hair wearing a light blue button-down shirt, speaking and gesturing with his hands. He is positioned on the left side of the slide.

“A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P , if its performance at tasks in T , as measured by P , improves with experience E .

~ Tom Mitchell
(on Machine Learning's Operational Definition)
(Machine learning, 1997)

Carnegie Mellon University
Machine Learning

What is machine learning?



Experience

- Data

Task

- Classification
- Regression
- Clustering
- Dimension reduction

...

Performance

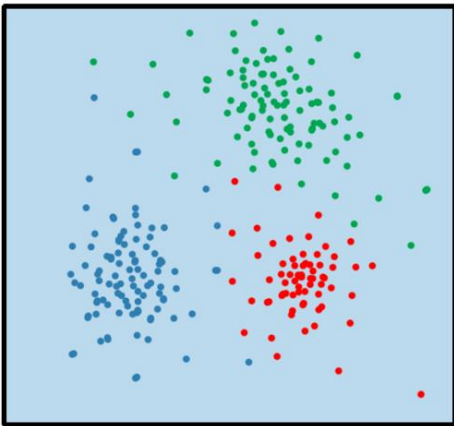
- Entropy loss
- Mean squared error
- Reconstruction error

...

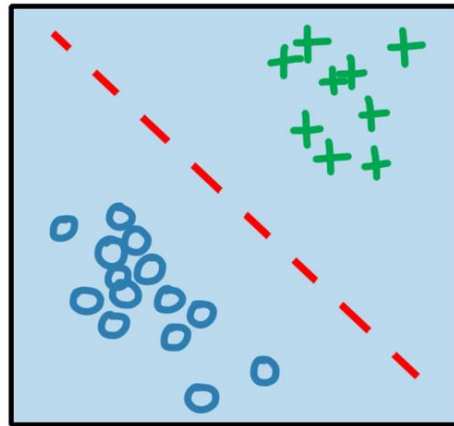
Types of machine learning

machine learning

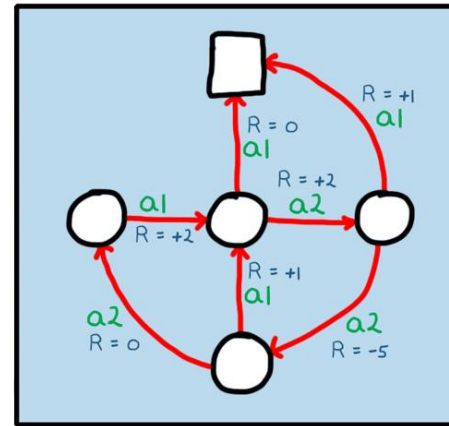
unsupervised
learning



supervised
learning



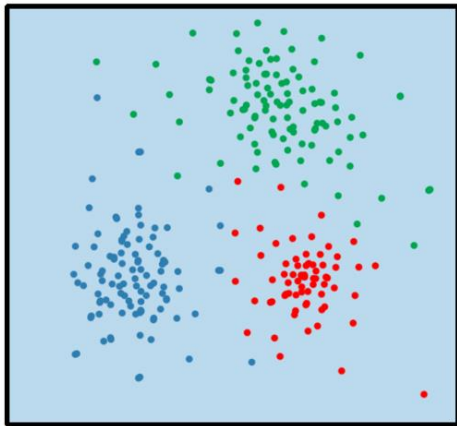
reinforcement
learning



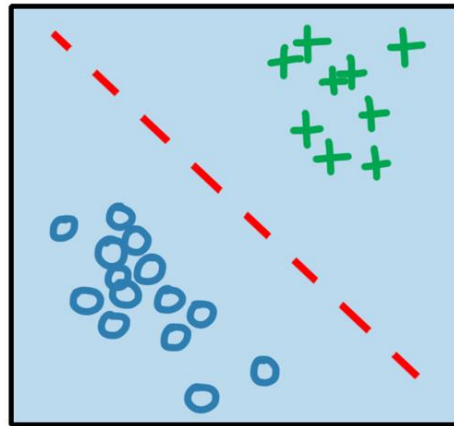
Types of machine learning

machine learning

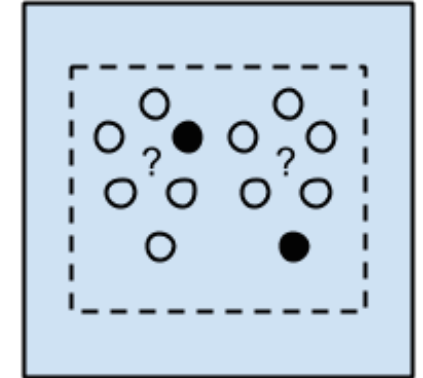
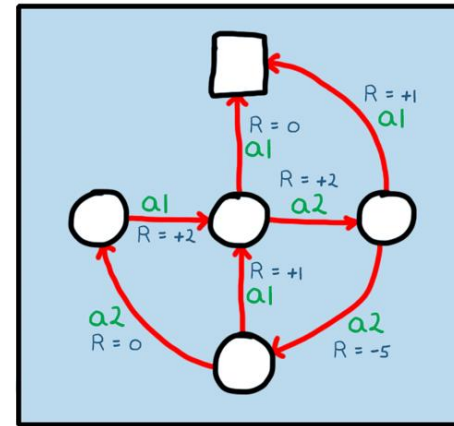
unsupervised
learning



supervised
learning



reinforcement
learning



Semi-supervised
Learning Algorithms

Self-supervised Learning



Other category types

- Batch learning vs Online-learning

- **Batch learning:** system is trained using all available data
- **Online learning:** system is trained incrementally by feeding data sequentially, data can be on the fly

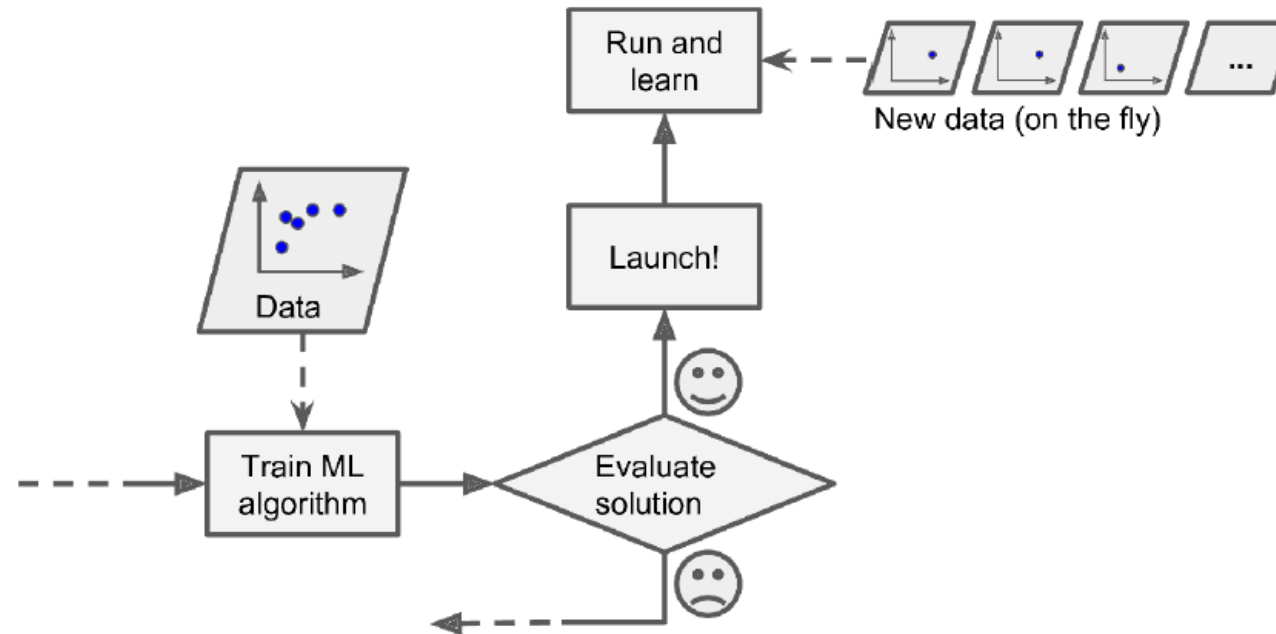


Figure 1-13. In online learning, a model is trained and launched into production, and then it keeps learning as new data comes in

Other category types

- Batch learning vs Online-learning

- **Batch learning:** system is trained using all available data
- **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

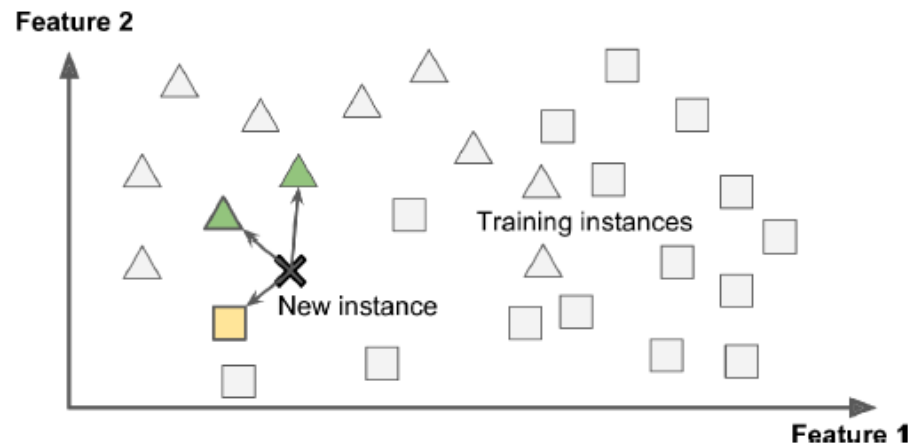


Figure 1-15. Instance-based learning

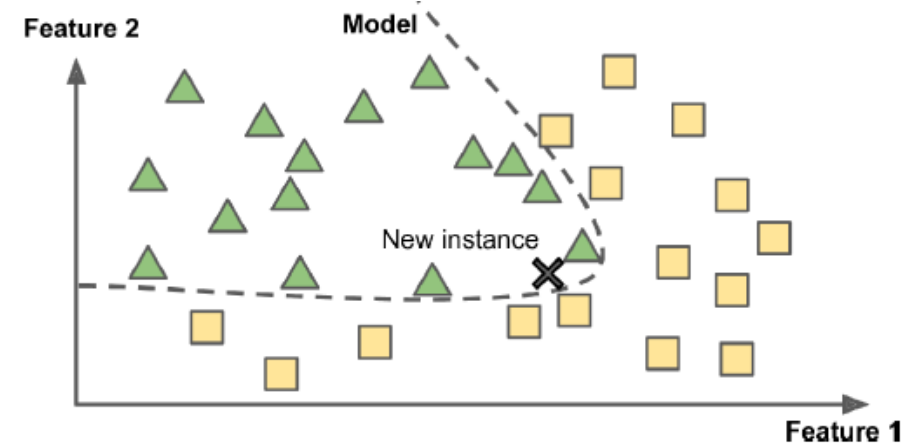
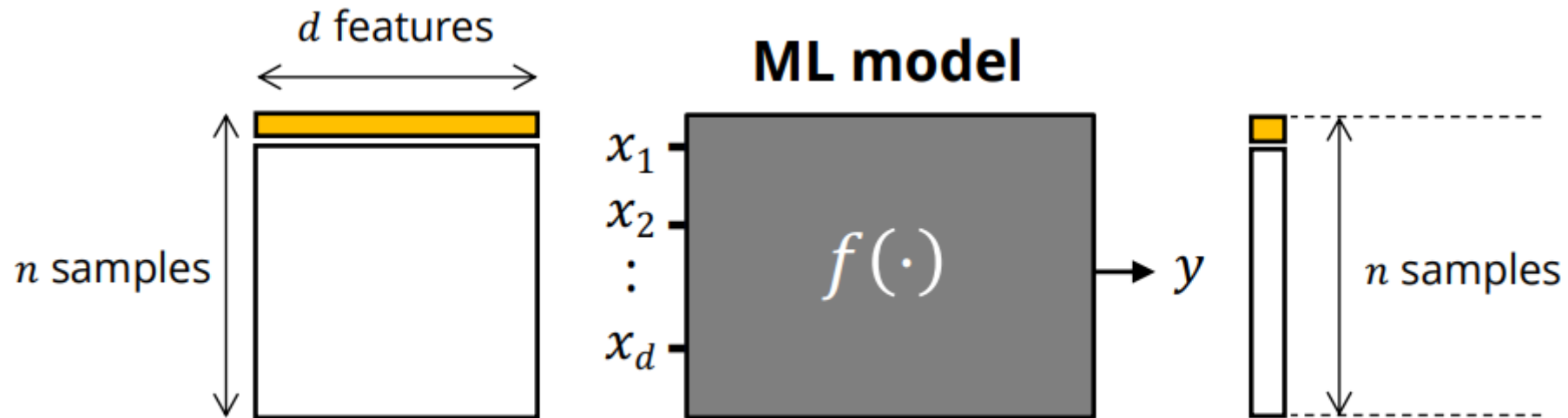


Figure 1-16. Model-based learning

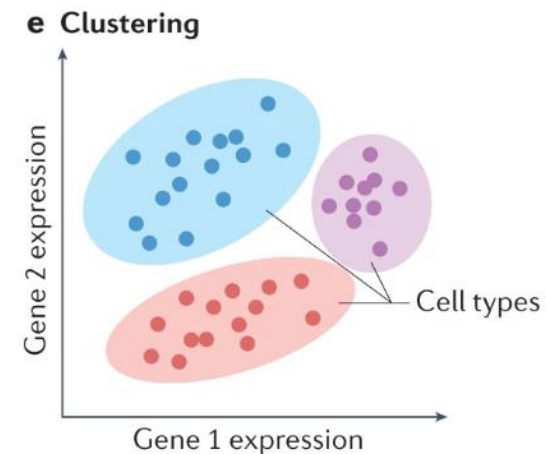
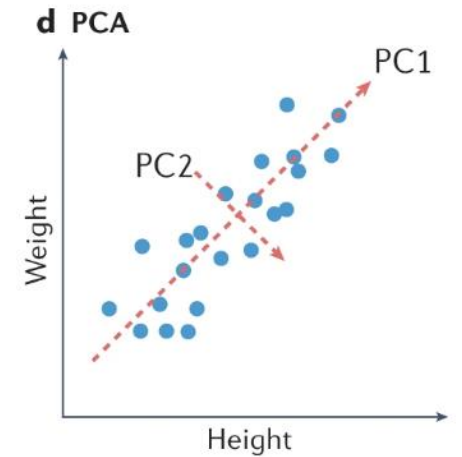
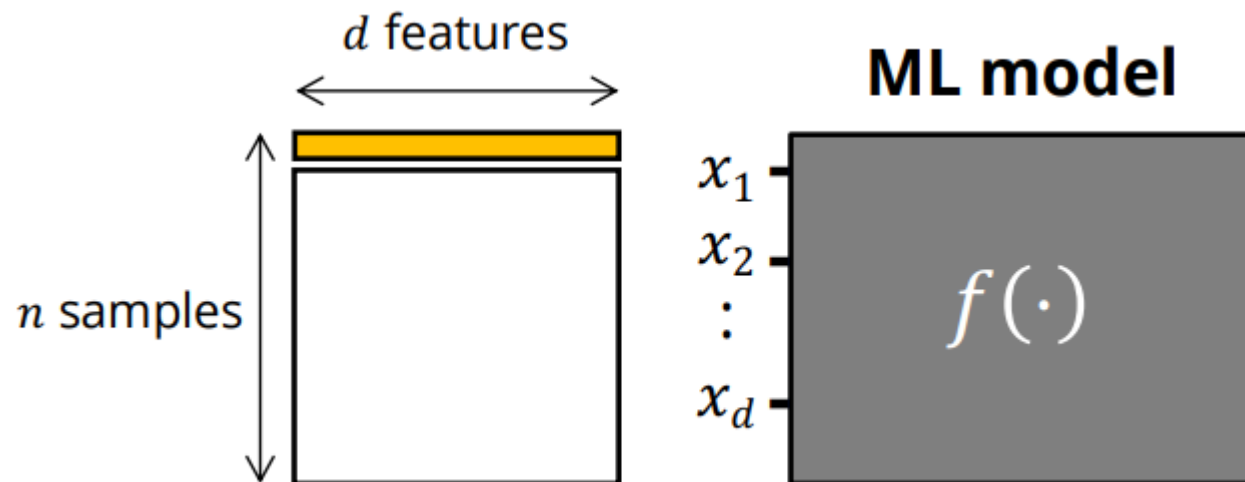
Supervised learning

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



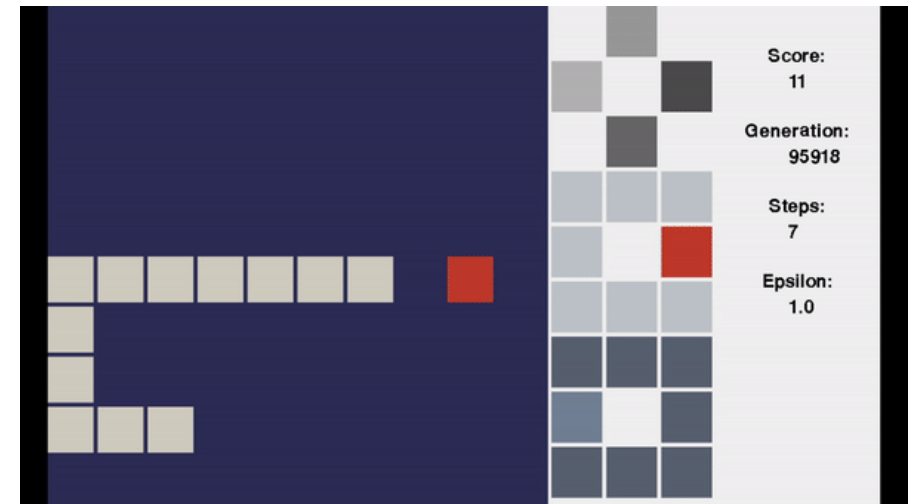
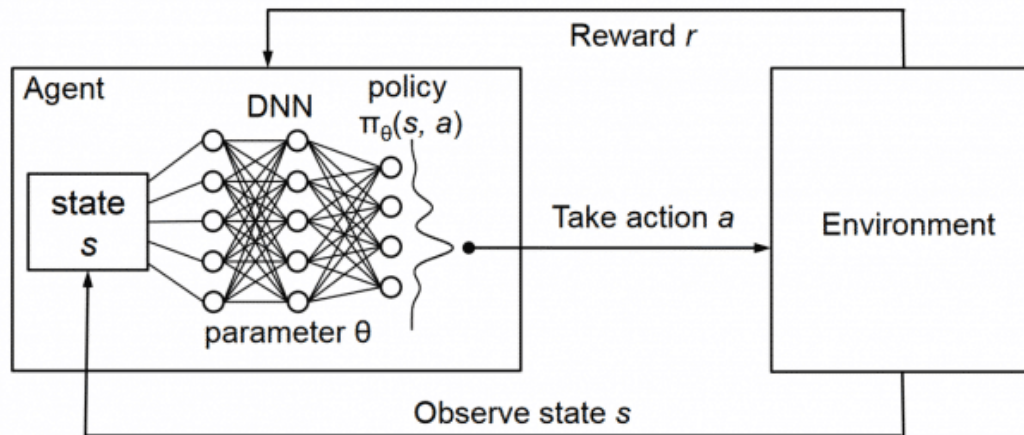
Unsupervised learning

- Training data with n ***samples*** of ***features*** x , **no label**
- Identify underlying patterns in unlabeled data



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

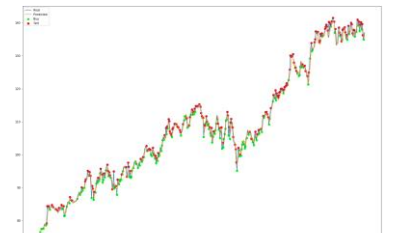


What can machine learning do?

- Face/Speech recognition
- Recommendation system
- Machine translation
- Self-driving system
- Stock market prediction
- Create images/songs/paintings/stories
- ...

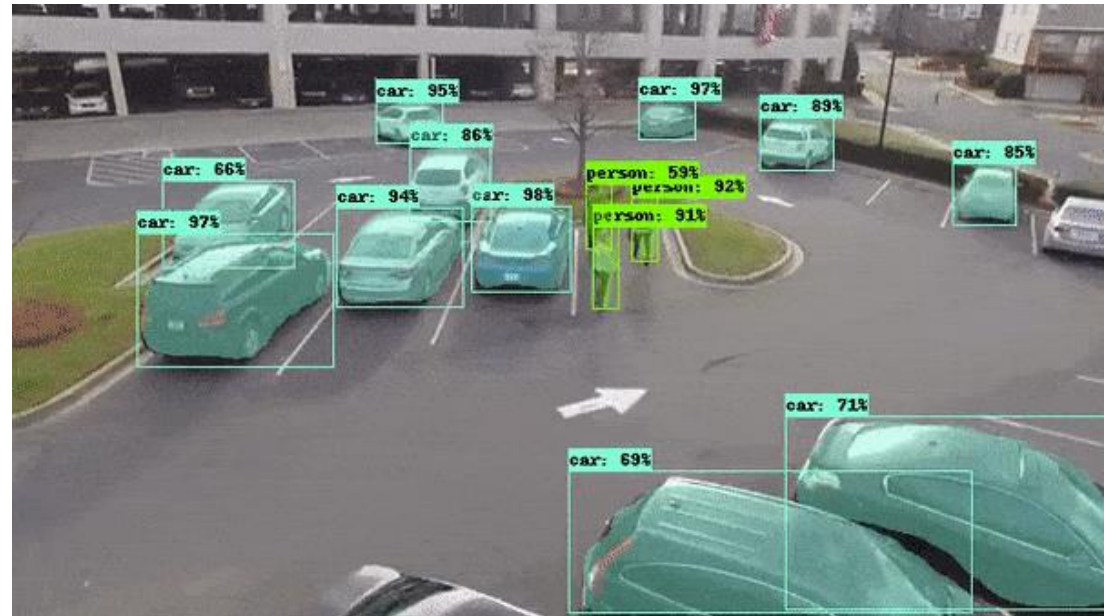
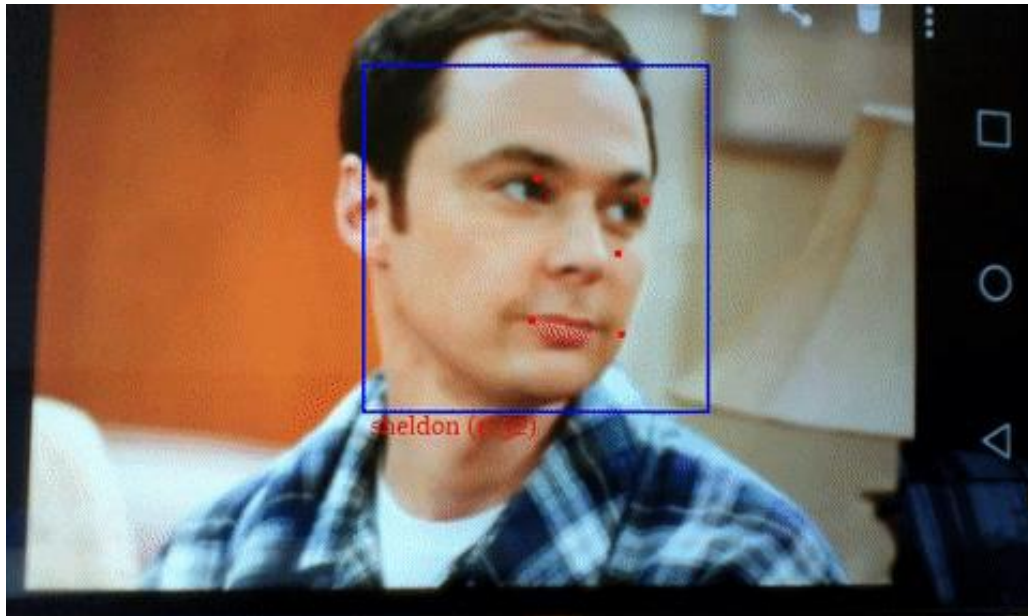


Google Translate



More examples

- Object segmentation and recognition



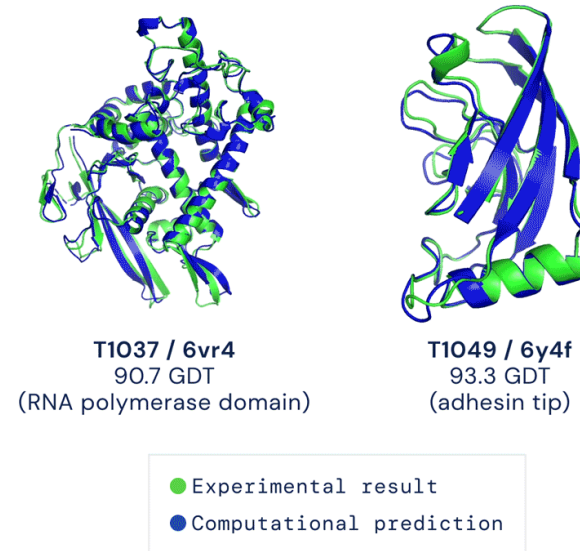
More examples

AlphaGo (2016)



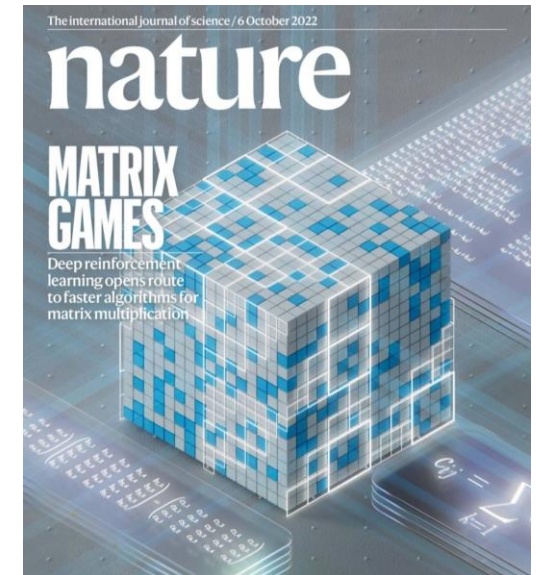
Monte Carlo tree search

AlphaFold (2021)



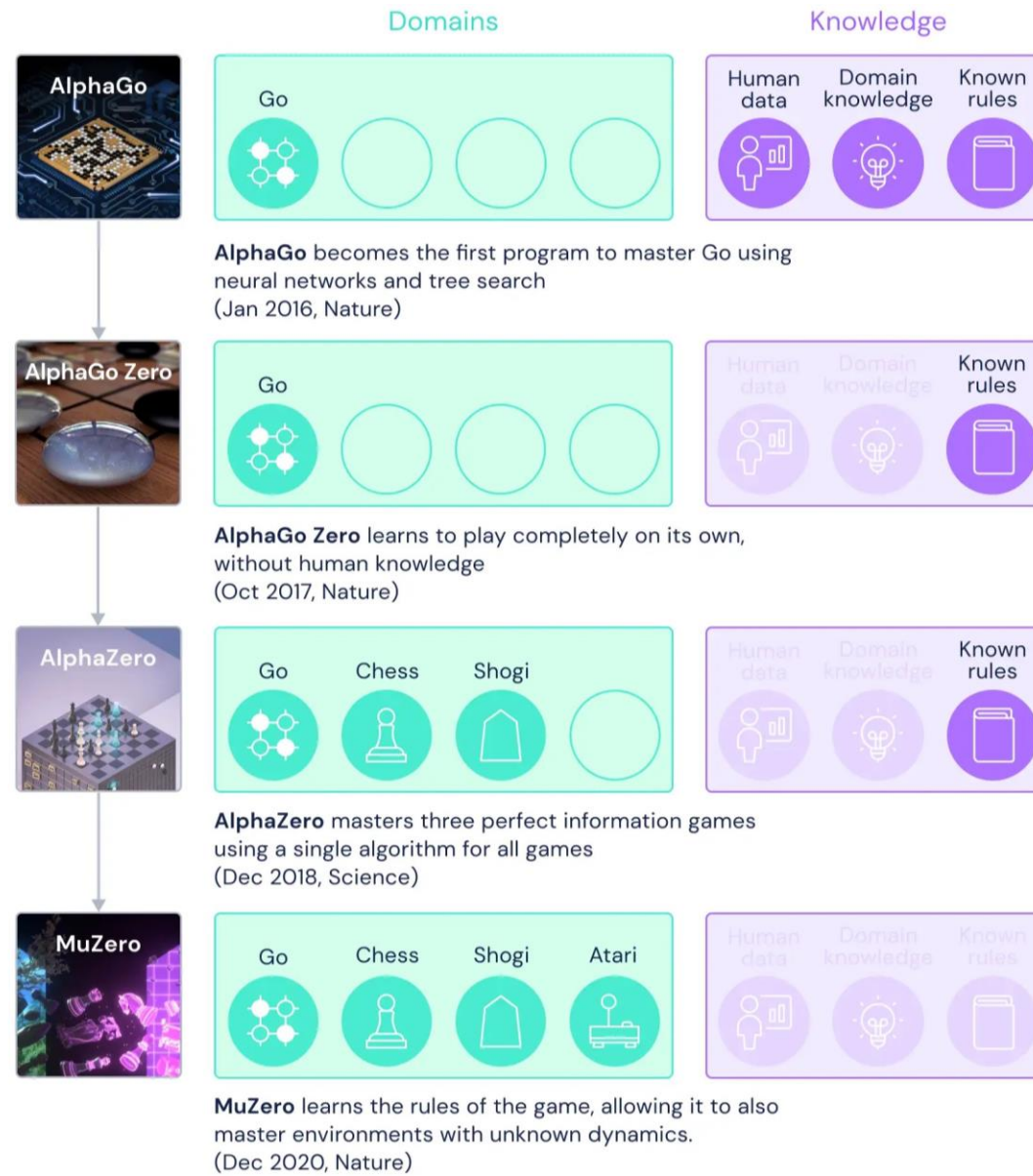
Transformer (self-attention)

AlphaTensor (2022)



Reinforcement learning

AlphaGo keeps evolving



Machine learning in biomedical research

- Imaging/sequencing processing
- Protein 3D structure prediction
- Genetic risk assessment
- Disease diagnosis
- Drug design
- ...

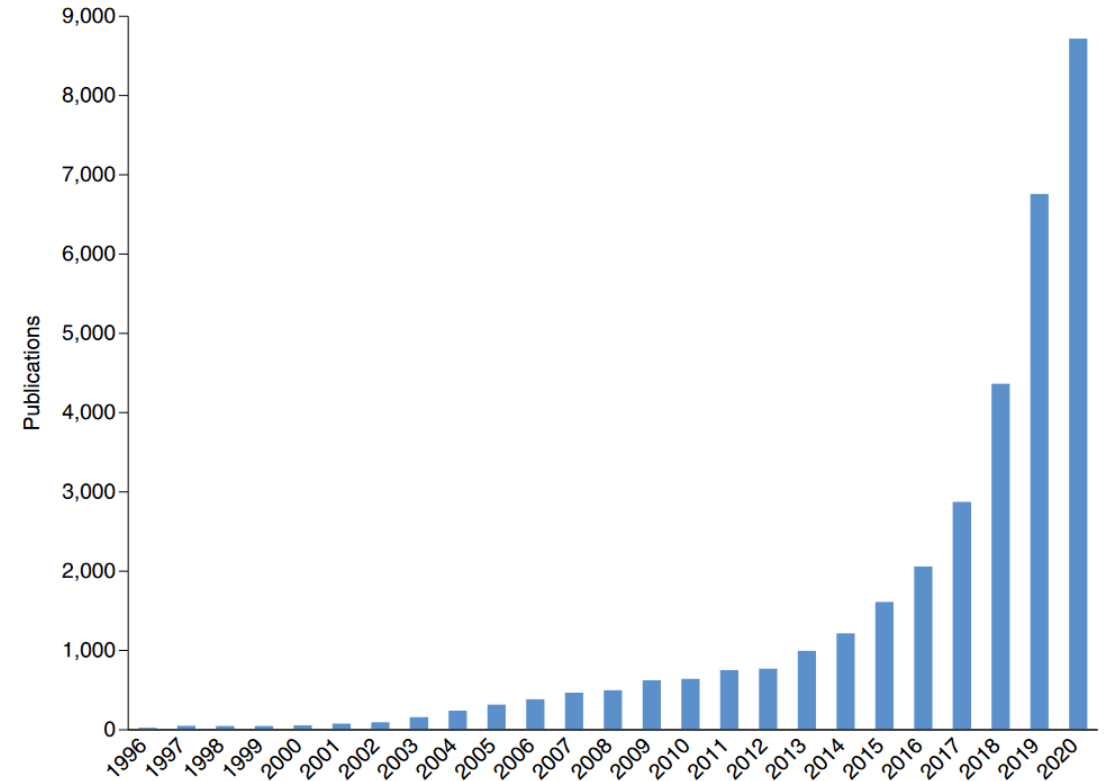


Fig. 1 | Exponential increase of ML publications in biology. The number of ML publications per year is based on Web of Science from 1996 onwards using the topic category for “machine learning” in combination with each of the following terms: “biolog*”, “medicine”, “genom*”, “prote*”, “cell*”, “post translational”,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

Train a machine learning model: a big picture



Train a machine learning model: a big picture



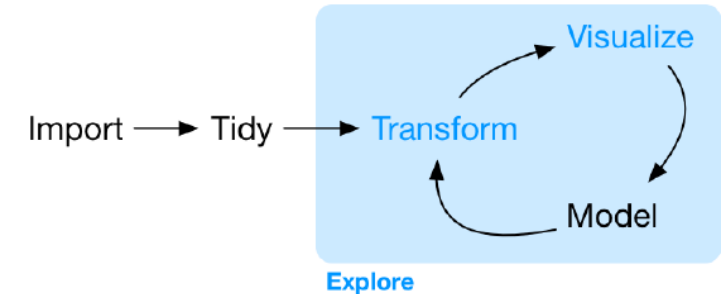
- Frame the problem
 - Supervised or unsupervised?
 - Classification or regression?
 - ...
- Select the performance measure (**loss function**):
 - Regression: RMSE, MAE
 - Classification: cross entropy
 - Dimension reduction: reconstruction error
 - ...
- **Cost / loss function** measures how *bad* your model fits the data

Train a machine learning model: a big picture



- Prepare the data

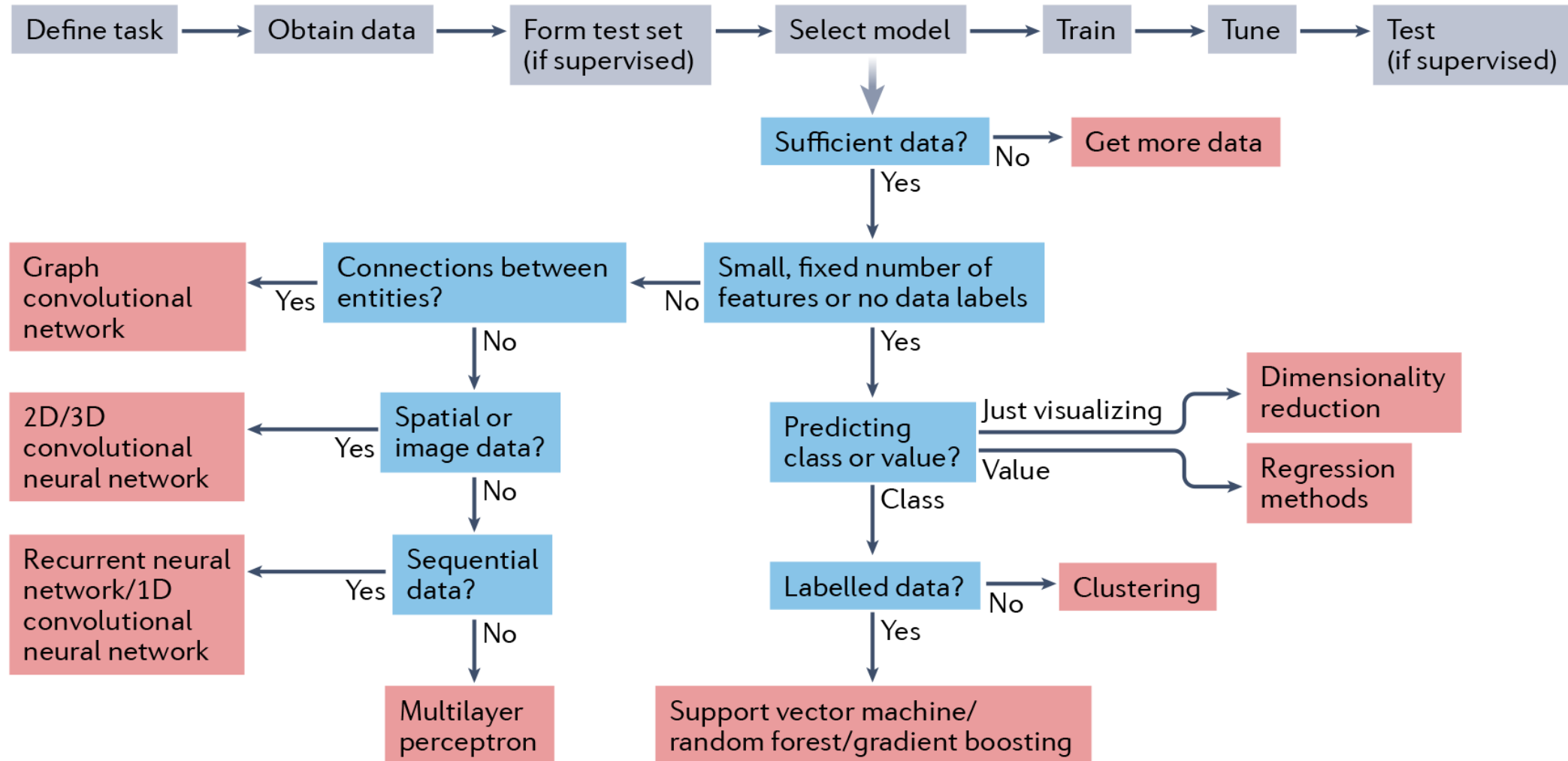
- Download / gather data
- Explore the data structure (descriptive analysis)
- Perform data cleansing, processing, feature engineering
- ...



- Create a test set (for supervised learning)

- 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

Train a machine learning model: a big picture

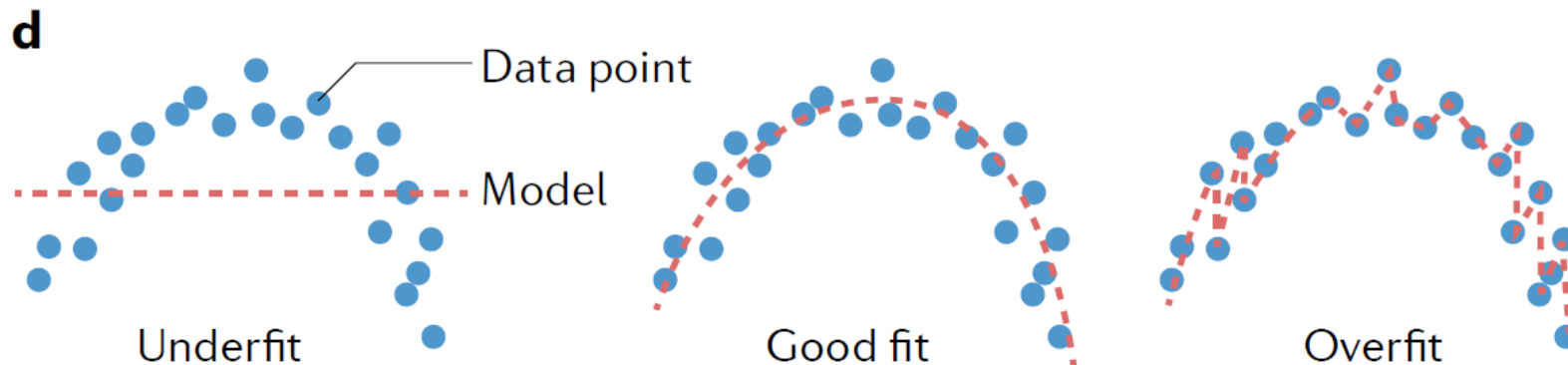


Train a machine learning model: a big picture



- **Train a model:**

- Run an algorithm to **find the model parameters that fits** the training data the best (and hopefully make good predictions on new data)



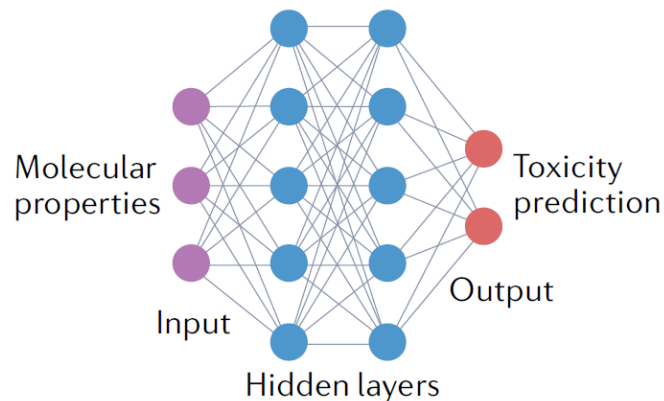
Train a machine learning model: a big picture



- Tune **hyperparameters**:

- **Hyperparameters** are **parameters of a learning algorithm (not of the model)**
- Unlike model parameters, hyperparameters are not updated during training (although they are adjustable)

a Multilayer perceptron



Example of hyperparameters:

- Number of neurons per layer
- Number of hidden layers
- ...

Train a machine learning model: a big picture



- Consider you have a set of 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

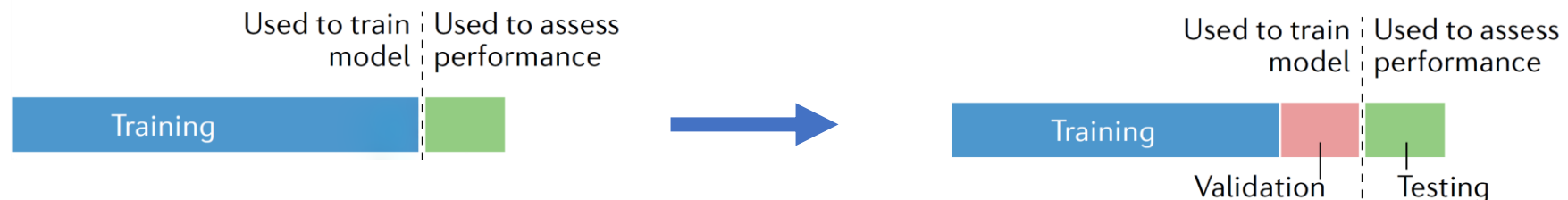


- Consider you have a set of hyperparameters/models to try
 - For each of them, you trained the model on training dataset, and then evaluate it on the test dataset, then you pick out the model with the best performance
 - Any problem?
 - You measured the **generalization error** multiple times on the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



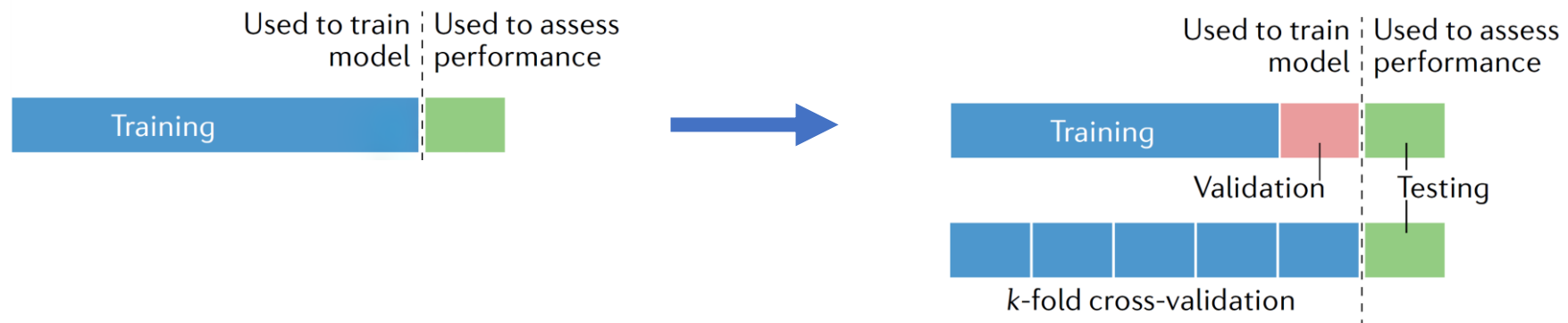
- Consider you have a set of hyperparameters/models to try
 - For each of them, you trained the model on training dataset, and then evaluate it on the test dataset, then you pick out the model with the best performance
 - Any problem?
 - You measured the **generalization error** multiple times on the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

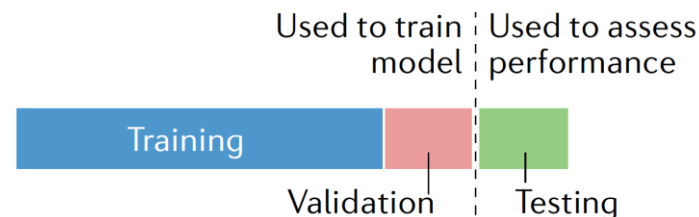


Train a machine learning model: a big picture



- 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**
 - 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**
 - Use a training set that is representative of the cases you want to generalize to
- **Poor quality data (garbage in, garbage out)**
 - Spend time cleaning up your training data
 - Remove clear outliers
- **Irrelevant features**
 - Feature selection: selecting the most useful features
 - Feature extraction: combining existing features to produce a more useful one
 - Creating new features

Summary: what we have learned so far

Key concepts in machine learning:

- ☐ Definition of 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 tuning (model selection)
- ☐ Cross validation
- ☐ Challenges in machine learning

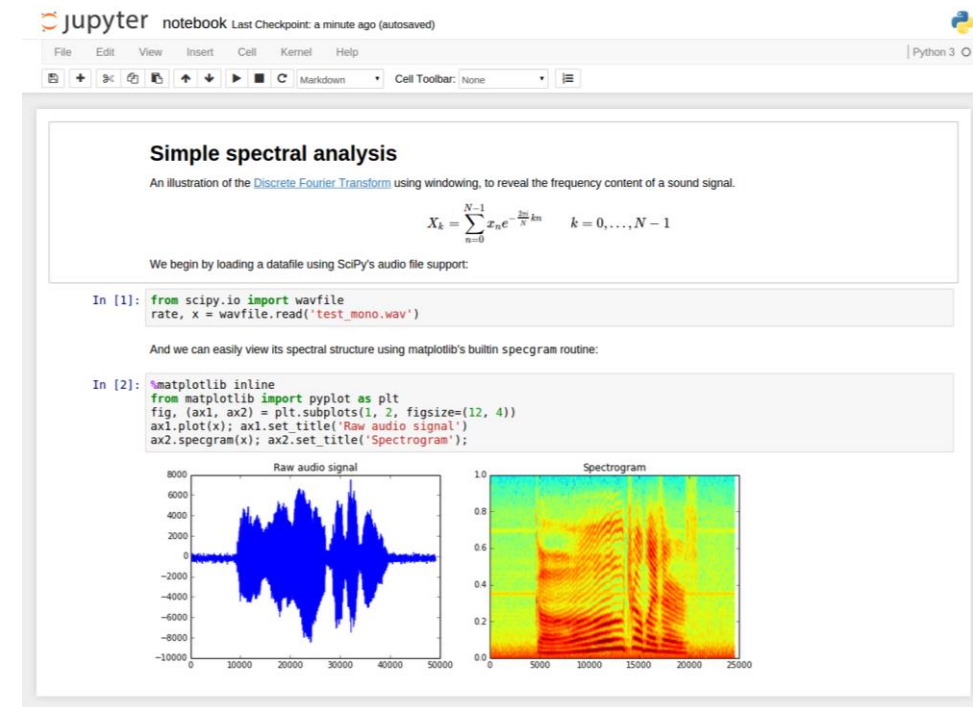
Enough theory, let's play!

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

Jupyter notebook

- 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 **J**ulia, **P**ython, **R**.



Preparation: Install Jupyter notebook

- Install on your local computer: [link](#)

Install the classic Jupyter Notebook with:

```
pip install notebook
```

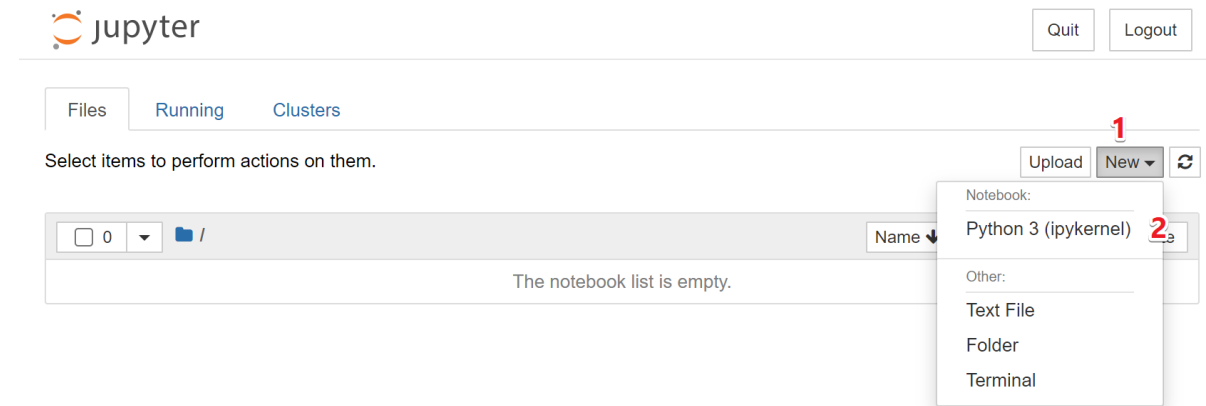
To run the notebook:

```
jupyter notebook
```

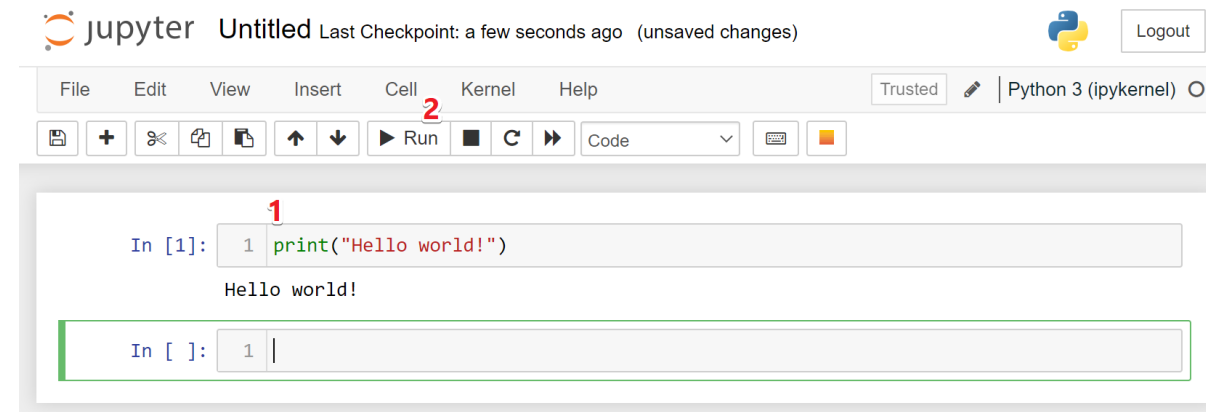
- If you are on Hoffman2: [link](#)
 - `wget https://raw.githubusercontent.com/rdauria/jupyter-notebook/main/h2jupynb`
 - `chmod u+x h2jupynb`
 - `./h2jupynb --help`
 - `./h2jupynb -u your_username -t 4 -m 8 -v 3.9.6`
- Alternatively, you can use [Google colab](#)

Preparation: Create a new notebook

- Create a new notebook



- Say “Hello world!”



- Now let's move to the Jupyter notebook tutorial

➤ `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/>
- <https://openai.com/blog/chatgpt/>



ChatGPT

The screenshot displays the ChatGPT web interface. On the left is a dark sidebar with navigation options: '+ New chat', 'Machine Learning Defin' (with edit and delete icons), 'LOOCV and Overfitting Preven', 'New chat', 'Clear conversations', 'Dark mode', 'OpenAI Discord', 'Updates & FAQ', and 'Log out'. The main area has a dark background and shows two chat messages. The first message is from a user (purple 'W' icon) asking 'what is machine learning'. The response is from ChatGPT (green OpenAI logo icon), explaining that machine learning is a method of data analysis that automates analytical model building, based on artificial intelligence, learning from data, and focusing on developing computer programs that learn from data. The second message is from the user asking to 'explain the difference between supervised learning and reinforcement learning'. The response explains that supervised learning involves training on labeled data for tasks like classification and regression, while reinforcement learning involves an agent learning to make decisions through trial and error based on rewards and penalties. A 'Regenerate response' button is visible below the second response. At the bottom, a footer states: 'ChatGPT Jan 9 Version. Free Research Preview. Our goal is to make AI systems more natural and safe to interact with. Your feedback will help us improve.'

+ New chat

Machine Learning Defin

LOOCV and Overfitting Preven

New chat

Clear conversations

Dark mode

OpenAI Discord

Updates & FAQ

Log out

W what is machine learning

Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.

W explain the difference between supervised learning and reinforcement learning

Supervised learning and reinforcement learning are two main categories of machine learning.

Supervised learning is a type of machine learning where the model is trained on labeled data, meaning the data used to train the model includes both the input features and the desired output. The goal of supervised learning is to learn a mapping from input features to output labels, such that the model can make accurate predictions on new, unseen data. Examples of supervised learning tasks include classification and regression.

Reinforcement learning, on the other hand, is a type of machine learning where an agent learns to make decisions by interacting with its environment. The agent receives feedback in the form of rewards or penalties based on its actions, and the goal is to learn a policy that maximizes the cumulative reward over time. The agent does not have a set of predefined actions, and instead must learn to map states of the environment to actions through trial and error.

Regenerate response

ChatGPT Jan 9 Version. Free Research Preview. Our goal is to make AI systems more natural and safe to interact with. Your feedback will help us improve.

Q&A

Google docs