Outline of this lecture

1. General introduction

2. Short selection of [simple] applications of supervised learning to chemistry

3. Tutorials / Live demonstrations ← Jupyter notebooks

github repository

Python in the Physical Chemistry Lab

[pyPhysChem]

https://github.com/rpoteau/PytChem

General context

Context

Artificial Intelligence (AI)

intelligence demonstrated by machines, as opposed to the natural intelligence displayed by humans or animals

Goals

reasoning & (basic) problem solving

knowledge representation

planning: making choices and hierarchy of events

learning (*i.e.* machine learning)

natural language processing

perception of the world from sensors

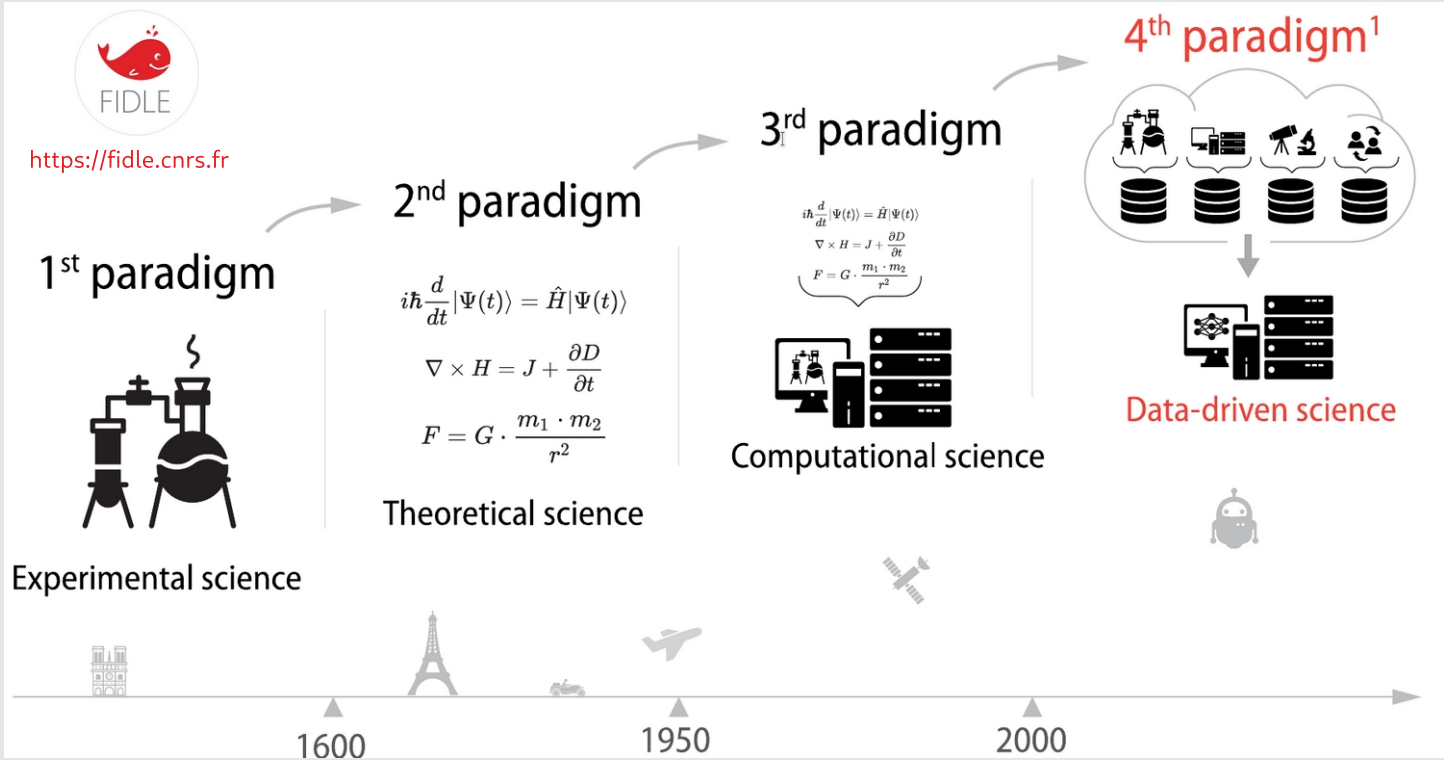
ability to move and manipulate objects

simulating human affects

long-term goal: ability to solve an arbitrary problem

N.B. in some fields "artificial intelligence" means "machine learning with neural networks"

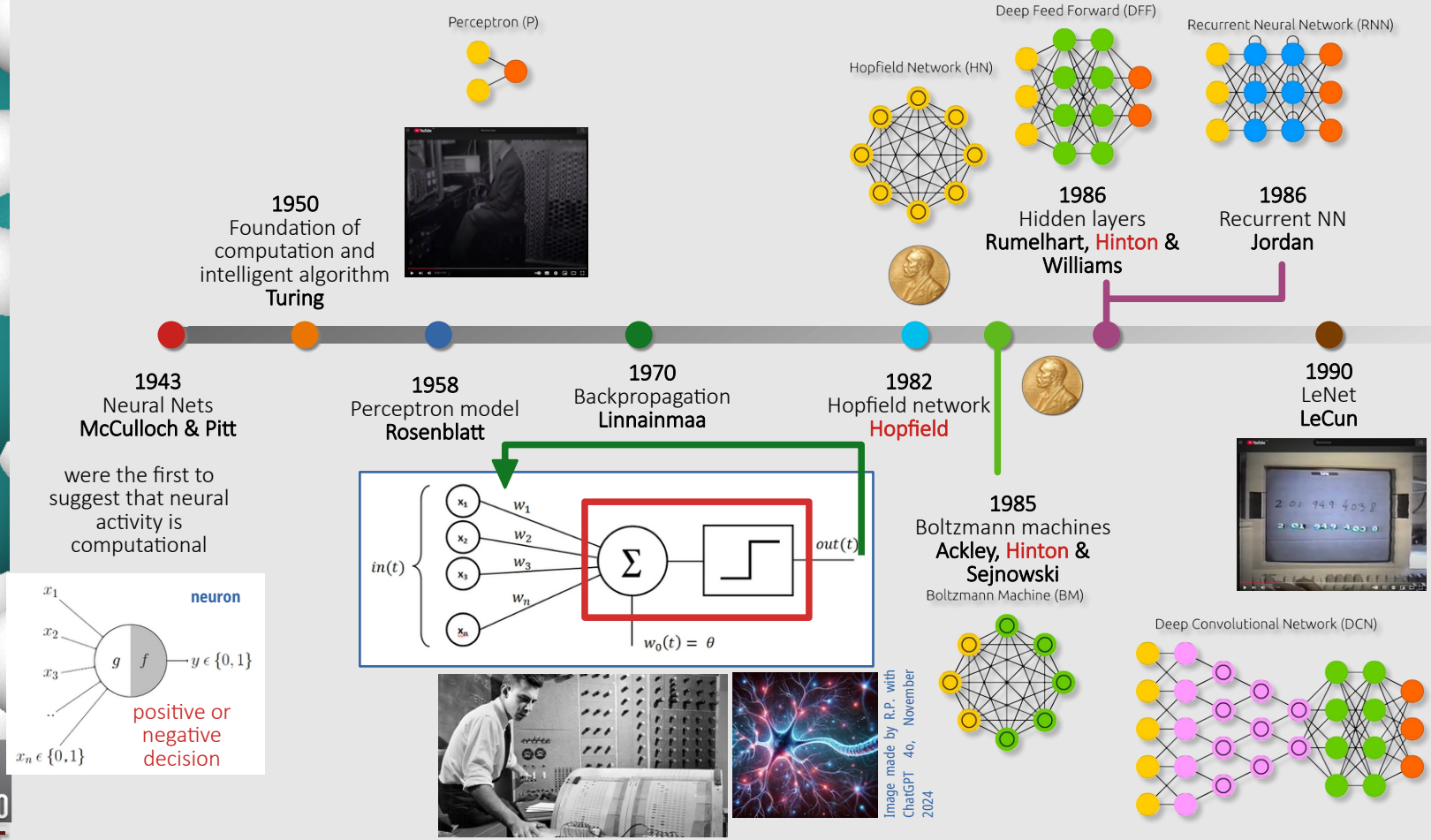
Artificial Intelligence and Scientific Research



(1) Hey, T.; Tansley, S.; Tolle, K. The Fourth Paradigm: Data-Intensive Scientific Discovery; The Fourth Paradigm: Data-Intensive Scientific Discovery; Microsoft Research, 2009

Outlook: data science and machine learning

Artificial Intelligence and deep learning timeline



W. S. McCulloch & W. Pitts (**1943**)
A logical calculus of the ideas immanent in nervous activity
Bull. Math. Biophys. **5**: 115-133

F. Rosenblatt (**1958**)
The perceptron: A probabilistic model for information storage and organization in the brain
Psychol. Rev. **65**: 386-408

S. Linnainmaa (**1970**)
The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors (Masters thesis, in Finnish), University of Helsinki. pp. 6–7

J. J. Hopfield (**1982**)
Neural networks and physical systems with emergent collective computational abilities
Proc. Natl. Acad. Sci. U S A. **79**:2554

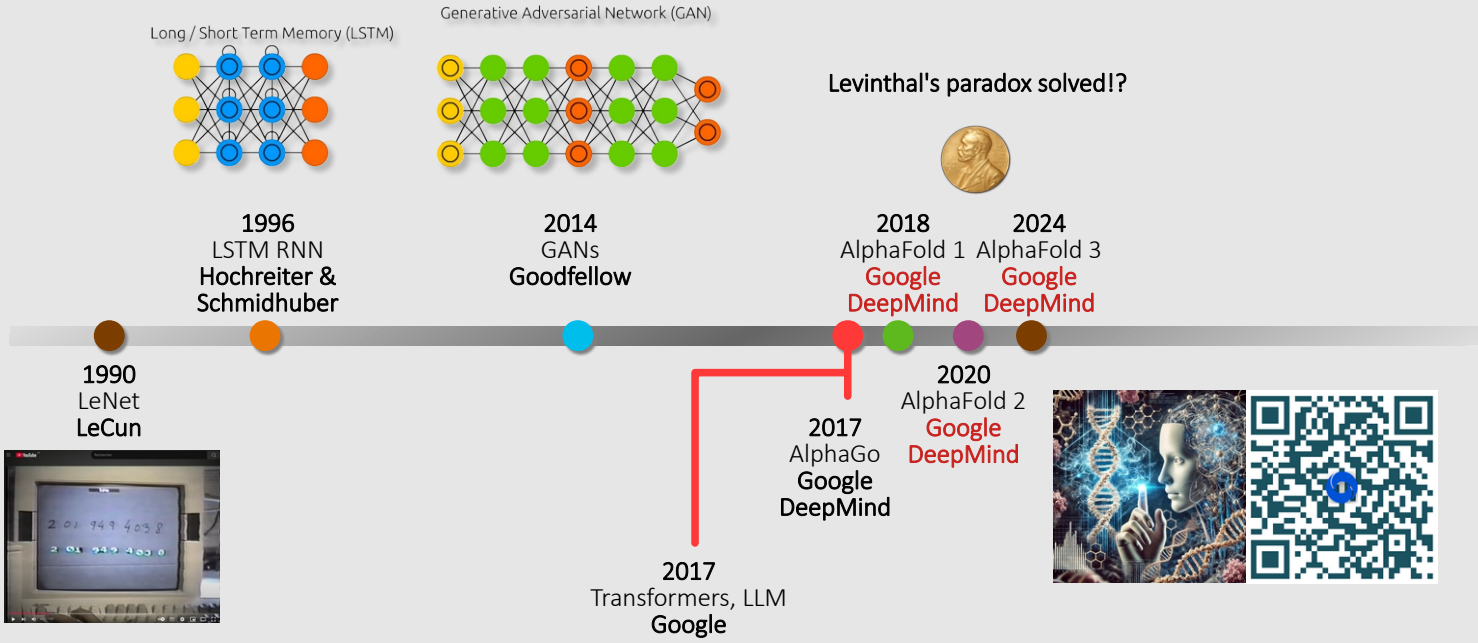
D. H. Ackley, G. E. Hinton, G. E., T. J. Sejnowski (**1985**)
A learning algorithm for Boltzmann Machines
Cog. Sci. **9**: 147

D. Rumelhart, G. Hinton & R. Williams (**1986**)
Learning representations by back-propagating errors
Nature **323**: 533

M. I. Jordan (**1986**)
Serial order: A parallel distributed processing approach
ICS Report 8604. Institute for Cognitive Science, UCSD, La Jolla

Outlook: data science and machine learning

Artificial Intelligence and deep learning timeline



S. Hochreiter & J. Schmidhuber (**1996**)
LSTM can solve hard long time lag problems
Advances in Neural Information Processing Systems 9: Proceedings of The 1996 Conference, MIT Press 1997, p. 473

I. Goodfellow, J. Pouget-Abade, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio (**2014**)
Generative Adversarial Nets. Proceedings of the International Conference on Neural Information Processing Systems (NIPS 2014), p. 2672

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin (**2017**)
Attention is All you Need
in Advances in Neural Information Processing Systems. 30. Curran Associates, Inc.

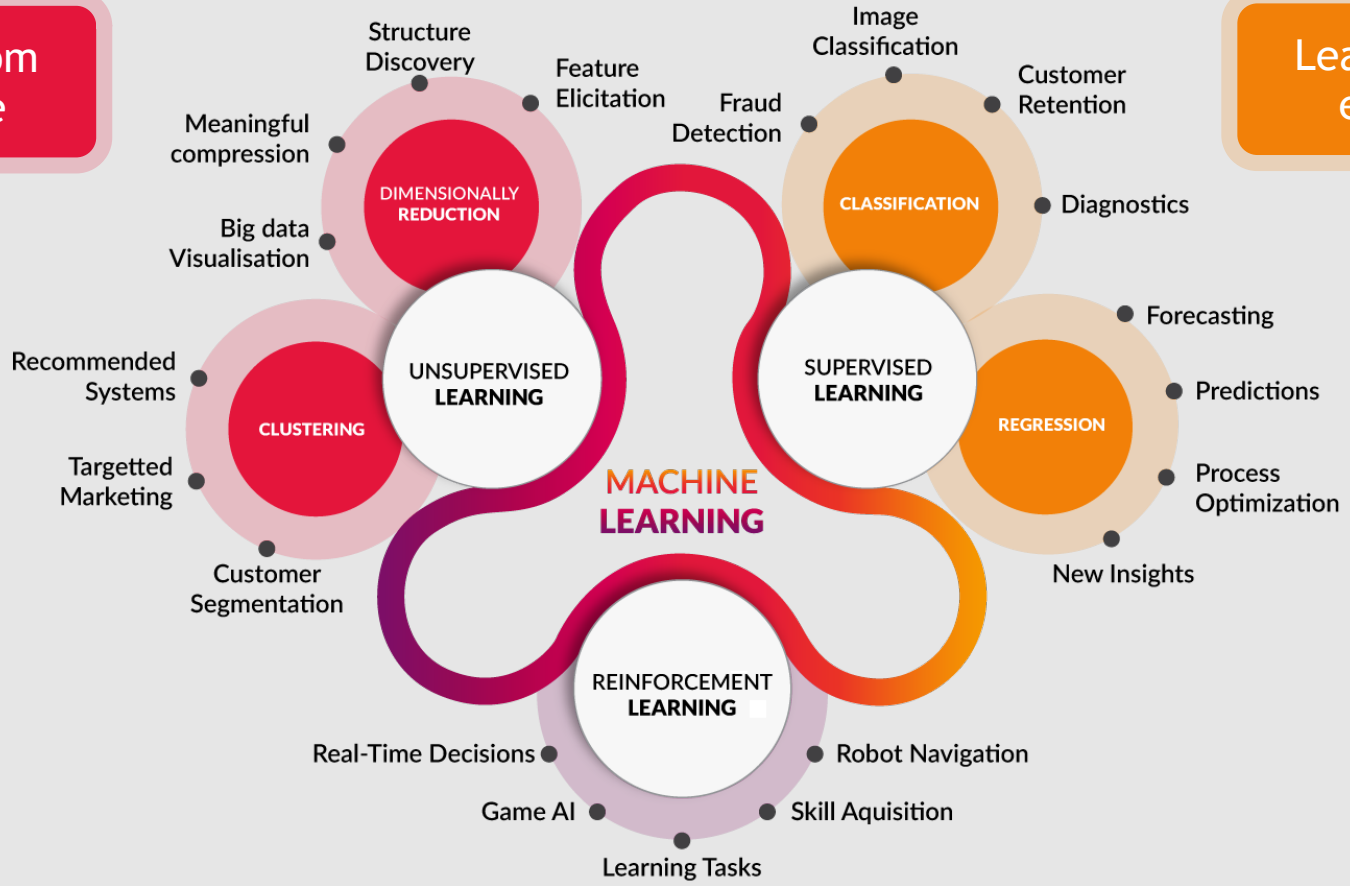
J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, S. W. Bodenstein, D. A. Evans, C.-C. Hung, M. O'Neill, D. Reiman, K. Tunyasuvunakool, Z. Wu, A. Žemgulytė, E. Arvaniti, C. Beattie, O. Bertolli, A. Bridgland, A. Cherepanov, M. Congreve, A. I. Cowen-Rivers, A. Cowie, M. Figurnov, F. B. Fuchs, H. Gladman, R. Jain, Y. A. Khan, C. M. R. Low, K. Perlin, A. Potapenko, P. Savy, S. Singh, A. Stecula, A. Thillaisundaram, C. Tong, S. Yakneen, E. D. Zhong, M. Zielinski, A. Žídek, V. Bapst, P. Kohli, M. Jaderberg, D. Hassabis & J. M. Jumper (**2024**)
Accurate structure prediction of biomolecular interactions with AlphaFold 3
Nature **630**: 493

Machine Learning

Machine learning

Learning from data alone

Learning from examples



https://towardsdatascience.com/coding-deep-learning-for-beginners-types-of-machine-learning-b9e651e1ed9d

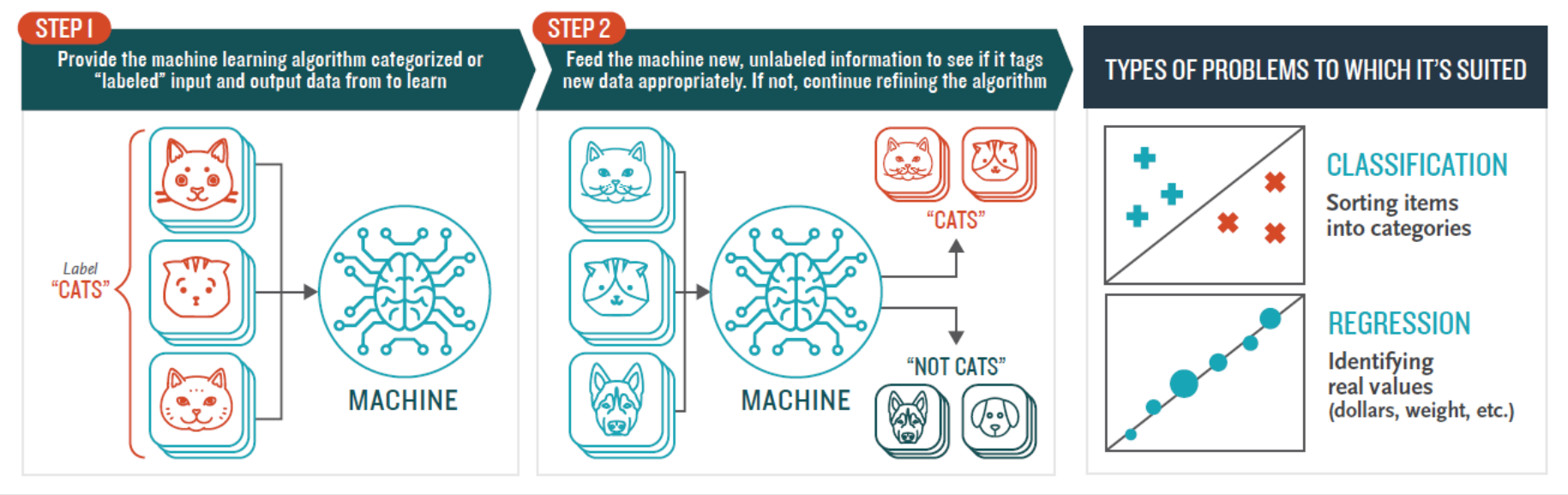
Supervised learning

Data are provided along with the desired output (*i.e.* labelled data)

Example of cats detection:

- **collect thousands of images of cats**
- **draw a bounding box around each cat**
- **feed the entire dataset to the machine so it can learn all by itself**

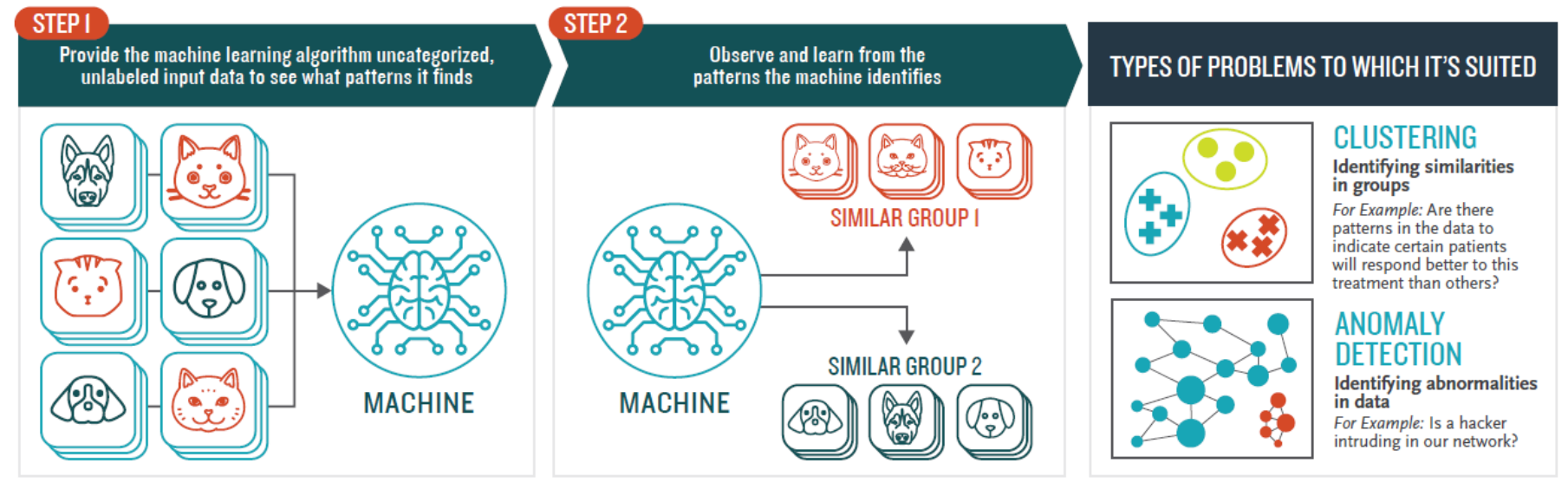
Learning from examples



Unsupervised learning

Learning from data alone

- Just provide data
- Let the machine find out (or cluster) the patterns in the dataset



Transformative role of AI in sciences of matter research

Image made by R.P. with ChatGPT 4o, November 2024

Acceleration of Material Discovery

- Property prediction prior to laboratory synthesis
- Exploration of a broad parameter space
- Inverse design of materials
- Catalyst design

Understanding mechanisms and multi-factorial processes through explainable AI (XAI)

Advanced Modeling

- Predictions comparable to *ab initio* calculations
- Simulations of physical and chemical processes at challenging scales

Design of Experiments

- Automated processing with robotic systems guided by AI and real-time analysis of experimental data
- Optimizing experimental setups through Design of Experiments (DoE)

Data Exploration

- Detecting trends and anomalies using unsupervised methods
- Derive insights from diverse datasets like crystal structures or spectroscopy

How to develop home-made ML tools?

mathematica

The ML galaxy within python™ is an exceptional ecosystem

Python is a high-level, interpreted, object-oriented, general-purpose programming language

https://scikit-learn.org/

https://keras.io/

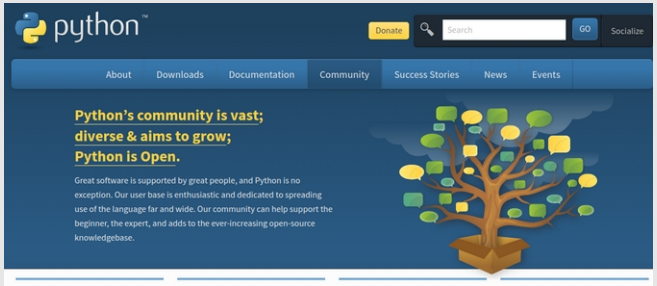
https://www.tensorflow.org/

Using Python as an everyday tool for scientific calculation and computation

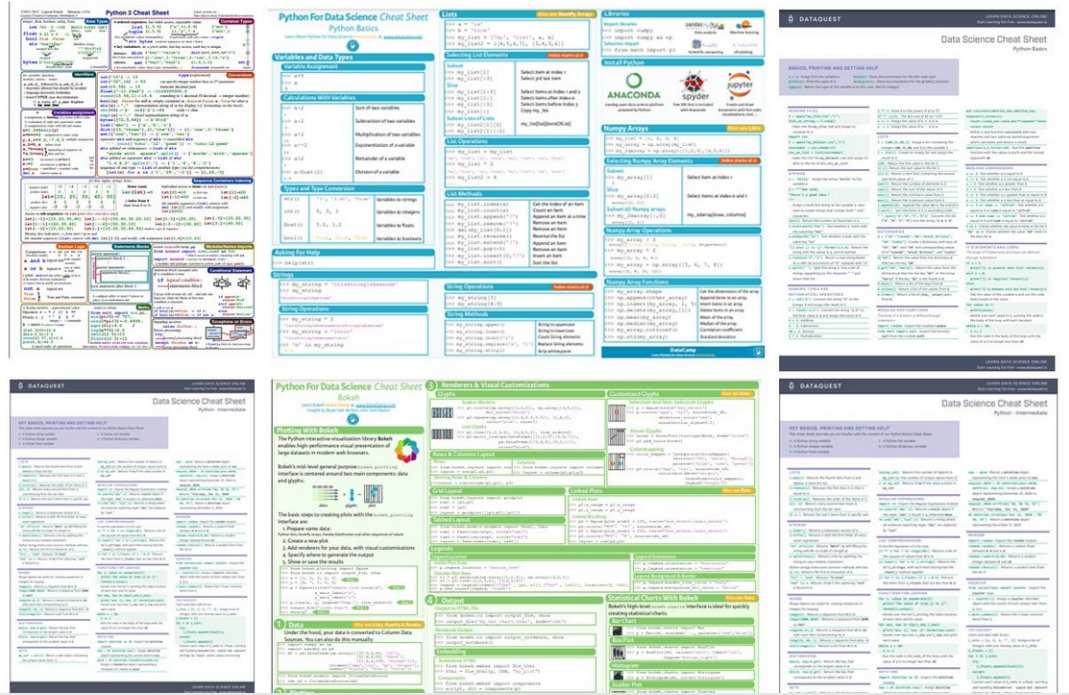
(it can even replace excel… by far)

→ basic knowledge of programming languages (variables, arrays, loops, conditional tests...)

→ enthusiastic and vast community



→ cheatsheets & cut/paste



Uneasy?
- yes and no
- important initial investment
- worth the effort

ML using the Python programming language

https://scikit-learn.org/

Machine learning in Python with scikit-learn

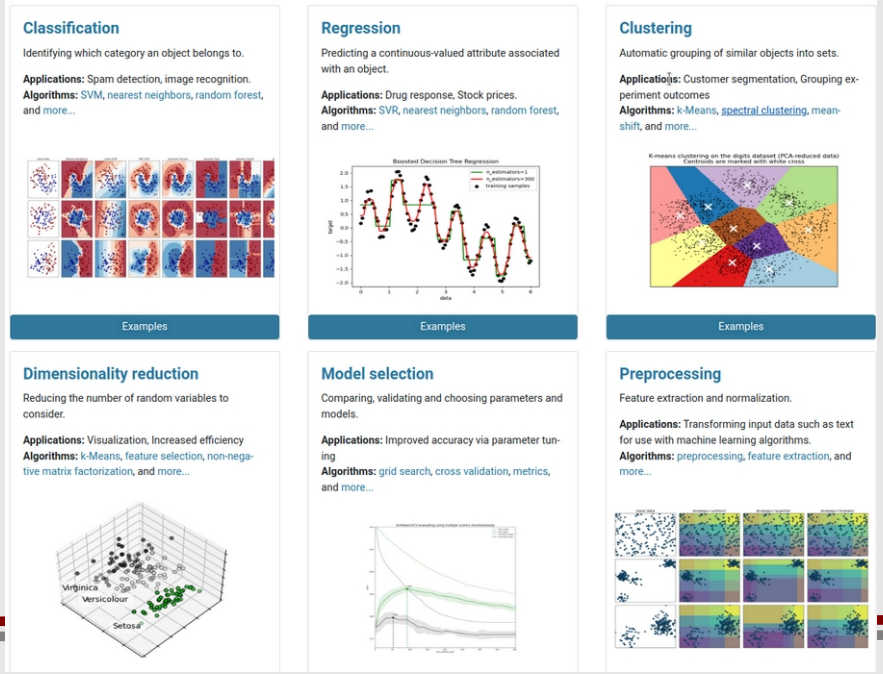
Simple and efficient tools for predictive data analysis

Accessible to everybody, and reusable in various contexts

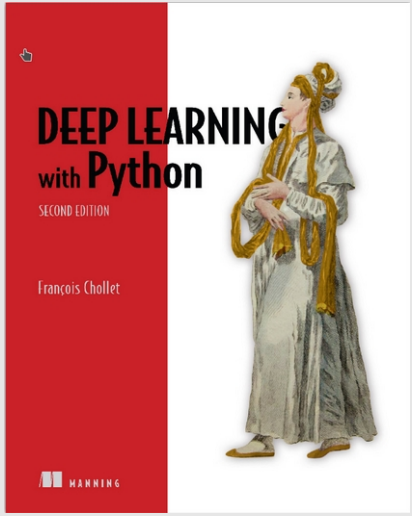
Built on NumPy, SciPy, and matplotlib

INRIA took leadership of the project and made the first public release on February 2010

3-clause BSD License (permissive free software license, compatible with the GNU GPL)



https://keras.io/



High Level Deep Learning Application Programming Interface (API)

By François Cholet (Google)

Part on TensorFlow since 2017

MIT license (permissive free software license)

how to start?

```
import numpy as np
import tensorflow as tf
from tensorflow import keras
```

https://www.tensorflow.org/

Google Brain's second-generation system

Supported by Google

Low level API

Apache license (yet another permissive free software license)

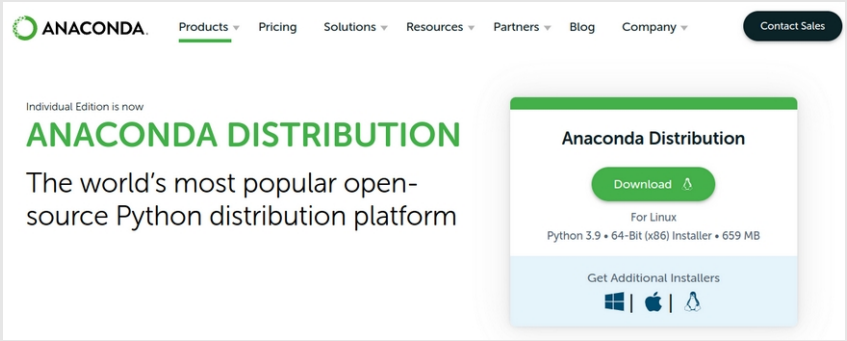
https://pytorch.org/

From Torch library
Supported by Facebook
BSD licence
(permissive free software license)

An open source machine learning framework that accelerates the path from research prototyping to production deployment

Widely used in the field of AI research

ANACONDA



Anaconda is a distribution of the Python and R programming languages for scientific computing (data science, machine learning applications, large-scale data processing, predictive analytics, etc.), that aims to simplify package management and deployment

jupyter

JupyterLab is a web-based interactive development environment for notebooks, code, and data

Users can arrange workflows in **data science**, **scientific computing**, and **machine learning**

you can save your everyday data manipulation / visualization as you do in your chemistry laboratory notebooks

Not familiar with python… yet?

"Python in the Physical Chemistry Lab (PPCL)" in a nutshell

This Python computer lab assumes a very basic knowledge of a programming language and algorithm development.

To run the content of a Python cell: click on a cell to select it. Then press SHIFT+ENTER on your keyboard, or press the play button in the top left corner of this window

If you click on a text cell by accident, you will see the so-called markdown coding of this cell (it is closely related to the HTML language). The corresponding formatted text/images/tables will be rendered by running the cell (SHIT-ENTER or play button).

Ready? Put down your mobile phone 📵, please, and let's enter into the Python realm. 👑🐍👑

PPCL.ipynb

Simple calculations

Basic mathematical operations

```
# Every line that starts with a # character is a comment

# addition
3 + 2 # it a also possible to add a comment after a command
```

```
#This is a new cell. It is possible to define several operations or commands in a cell
# multiplication
3*2

#division
7/2
```

github repository

https://github.com/rpoteau/pyPhysChem

README

Université de Toulouse

Python in the Physical Chemistry Lab

[pyPhysChem]

Ce dépôt GitHub propose une collection de notebooks Jupyter conçus pour intégrer la programmation en Python dans l'enseignement de la chimie physique. Ces notebooks fournissent des exemples commentés et illustrés, couvrant des sujets tels que les dérivées et les intégrales, l'atome d'hydrogène et les représentations moléculaires. Ce dépôt inclut également des ressources pour des applications d'apprentissage automatique en chimie, comme les réseaux de neurones artificiels et les autoencodeurs. Pour utiliser ces outils, les utilisateurs sont invités à installer Jupyter ainsi qu'une distribution Python, Anaconda étant recommandée. Des instructions détaillées pour cloner le dépôt et exécuter les notebooks sont disponibles dans ce fichier README.md

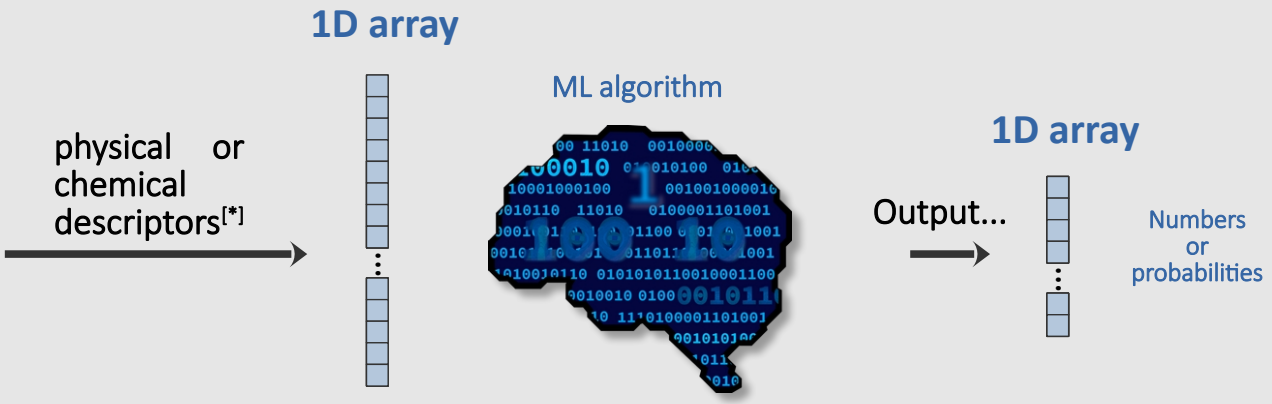
🇬🇧 This GitHub repository offers a collection of Jupyter Notebooks designed to integrate Python programming into physical chemistry education. These notebooks provide commented and illustrated examples, covering topics such as derivatives and integrals, the hydrogen atom, and molecular representations. The repository also includes resources for machine learning applications in chemistry, like artificial neural networks and autoencoders. To utilize these materials, users are advised to install Jupyter and a Python distribution, with Anaconda being a recommended option. Detailed instructions for cloning the repository and running the notebooks are provided in the present README.md document

Table des Matières

🇬🇧 Table of Contents

Machine Learning in Physical Chemistry

General context

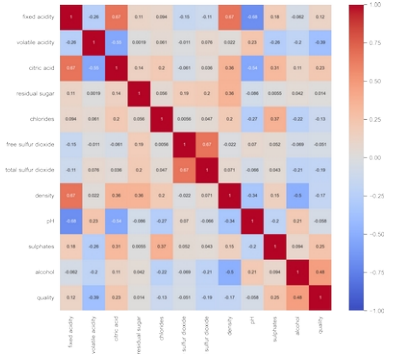


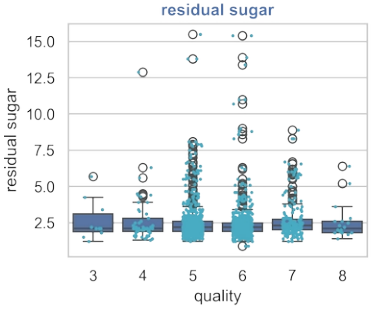
[*] electron-counting rules, local aromaticity, HOMO-LUMO gaps, d-band centers, Fermi energy, global or local structural parameters, atomic charges, electron density, BDE, experimental spectrum, kinetic curve, thermal conductivity, heat capacity, resistivity, ductility,....

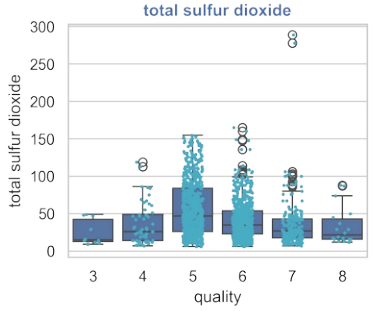
Simplified – but recommended - workflow

Exploratory Data Analysis

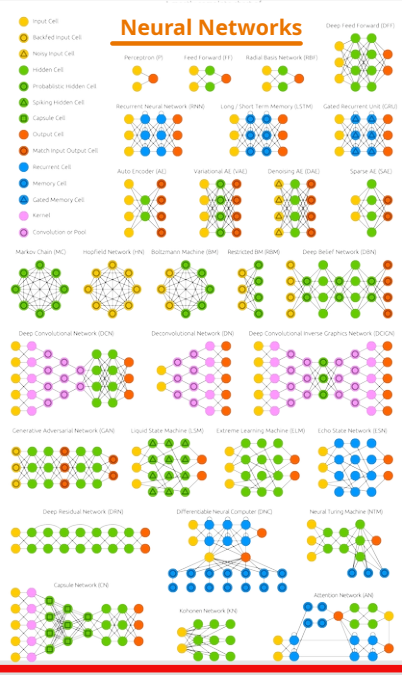
Understanding the Dataset
Data Quality Assessment
Descriptors (feature) Engineering
Training accuracy



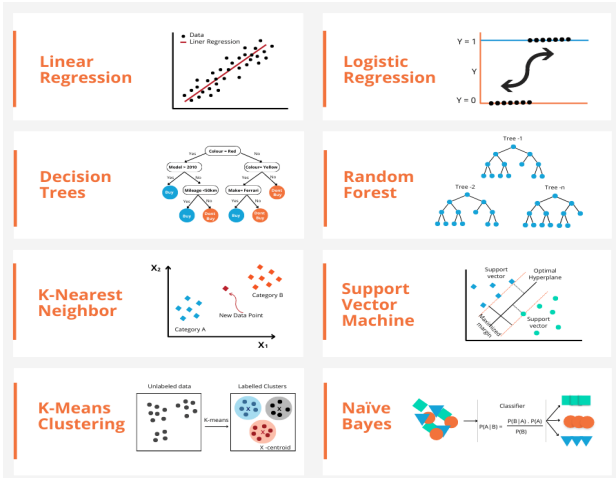




Neural Networks



Choice of an ML algorithm



find/create bare data

20% test set

80% training set

input data manipulation = data analysis and filtering, extraction of features (**descriptors**), standardization...

(all samples of the test & training sets)

flattening of each sample

choice of an **ML model** + **hyperparameters** tuning

training, **evaluation** and optimization of the **hyperparameters** of the model

K-fold cross validation

application of the best ML model to new data

Model Explainability (XAI)

Shap