THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

Baker
HEART & DIABETES INSTITUTE

Quantum Machine Learning

Alex de Sá Alex.deSa@baker.edu.au
Ashar Malik Ashar.Malik@baker.edu.au

https://github.com/alexgcsa/incob2023

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

Baker
HEART & DIABETES INSTITUTE

~~Quantum~~ Machine Learning

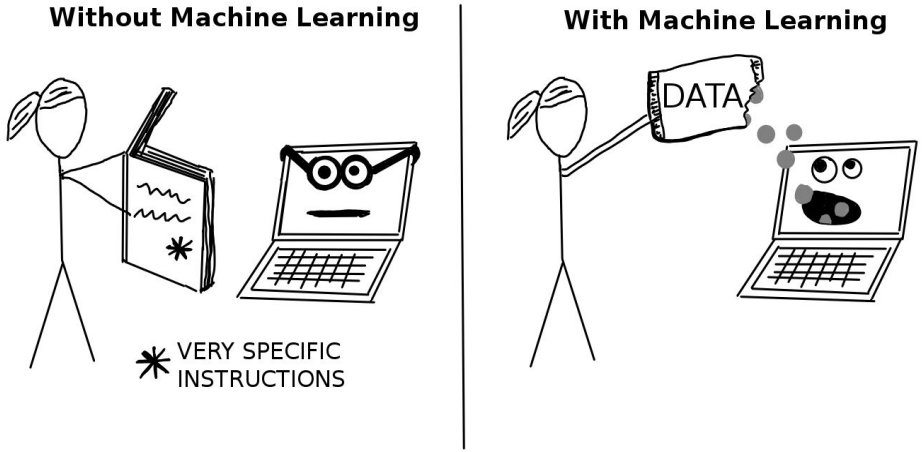
Alex de Sá — Alex.deSa@baker.edu.au
Ashar Malik — Ashar.Malik@baker.edu.au

https://github.com/alexgcsa/incob2023

Machine Learning

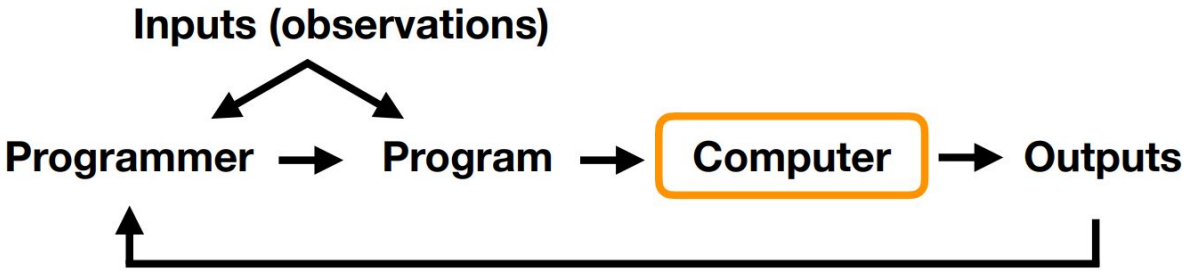
Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed.

Arthur L. Samuel, AI pioneer, 1959



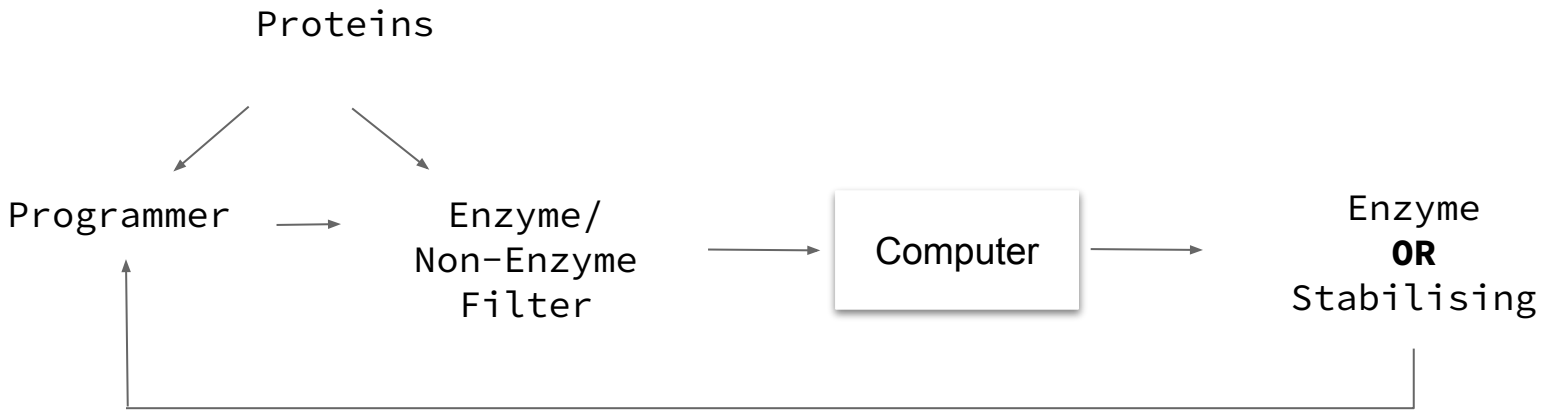
Molnar, 2021

Traditional Programming versus Machine Learning



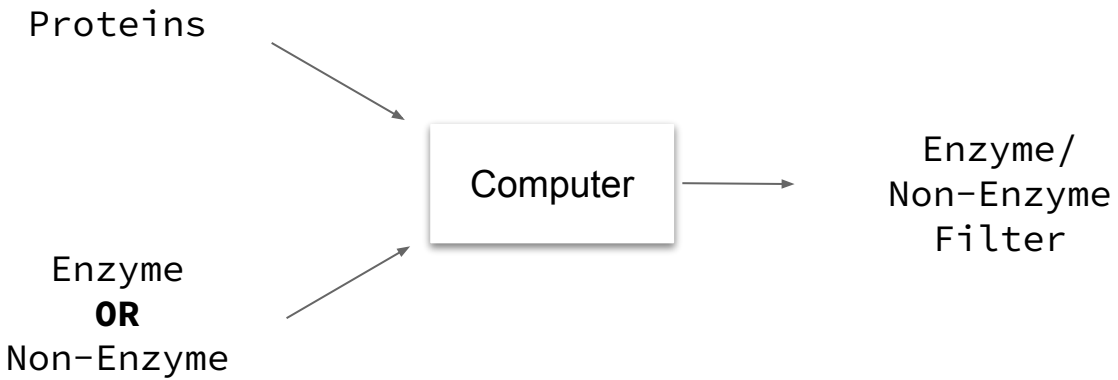
Raschka, 2021

Traditional Programming



Raschka, 2021

Machine Learning

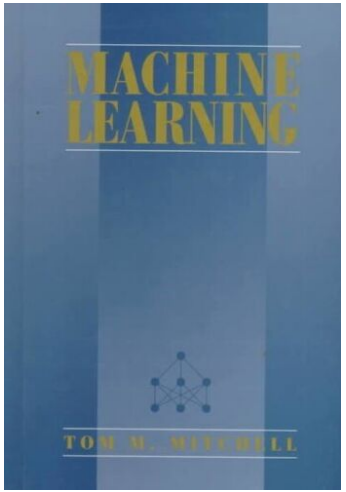


Raschka, 2021

Machine Learning Definition

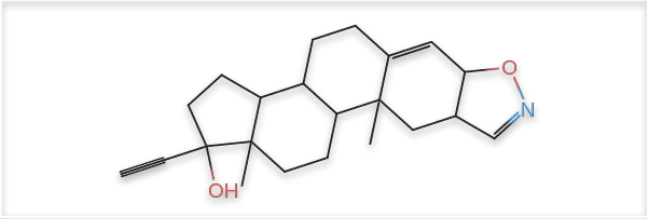
A computer program is said to **learn** from experience **E** with respect to some class of tasks **T** and performance measure **P**, if its performance at tasks in **T**, as measured by **P**, improves with experience **E**.”

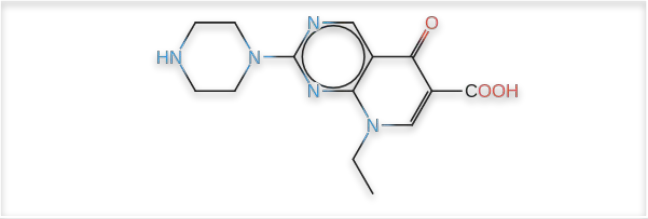
Tom Mitchell, Professor at Carnegie Mellon University

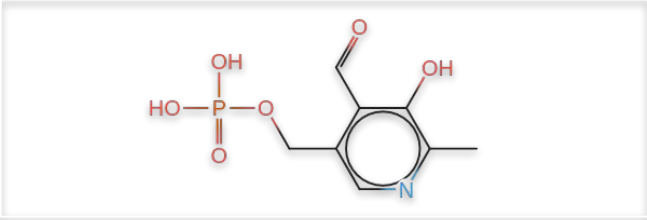


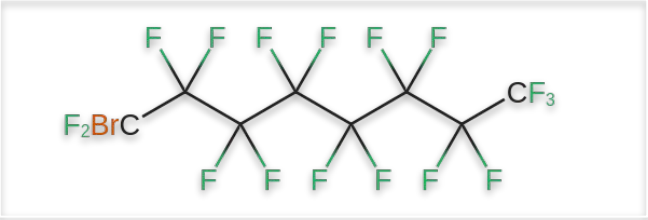
https://www.cs.cmu.edu/~tom/mlbook.html

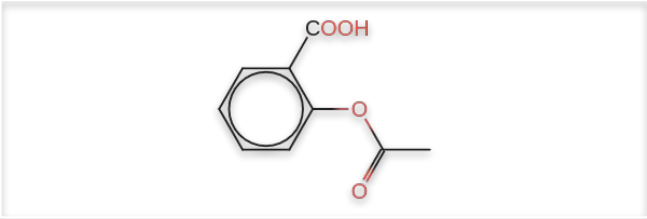
Identification of Hepatotoxicity in Small Molecules

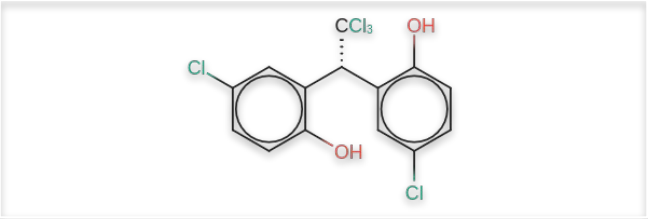




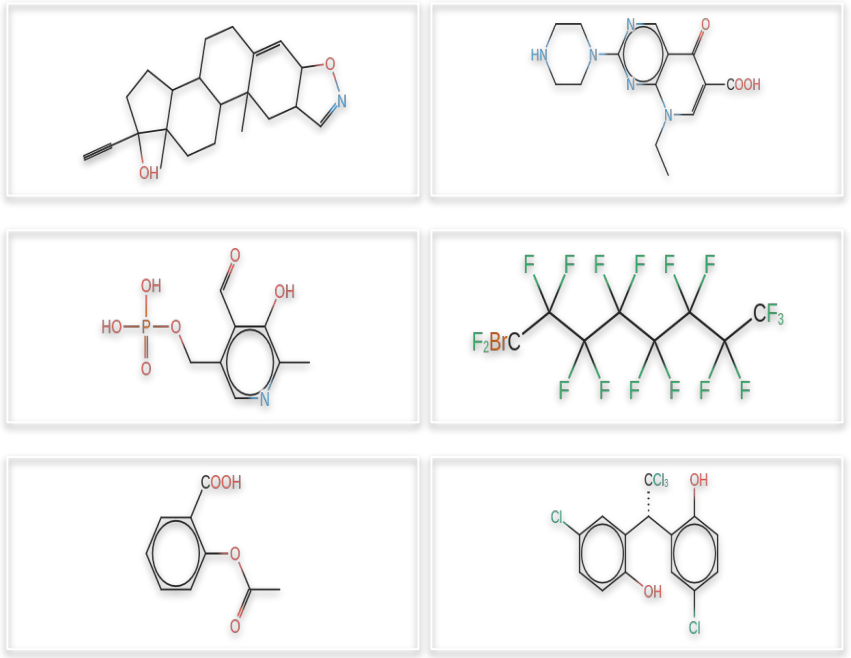






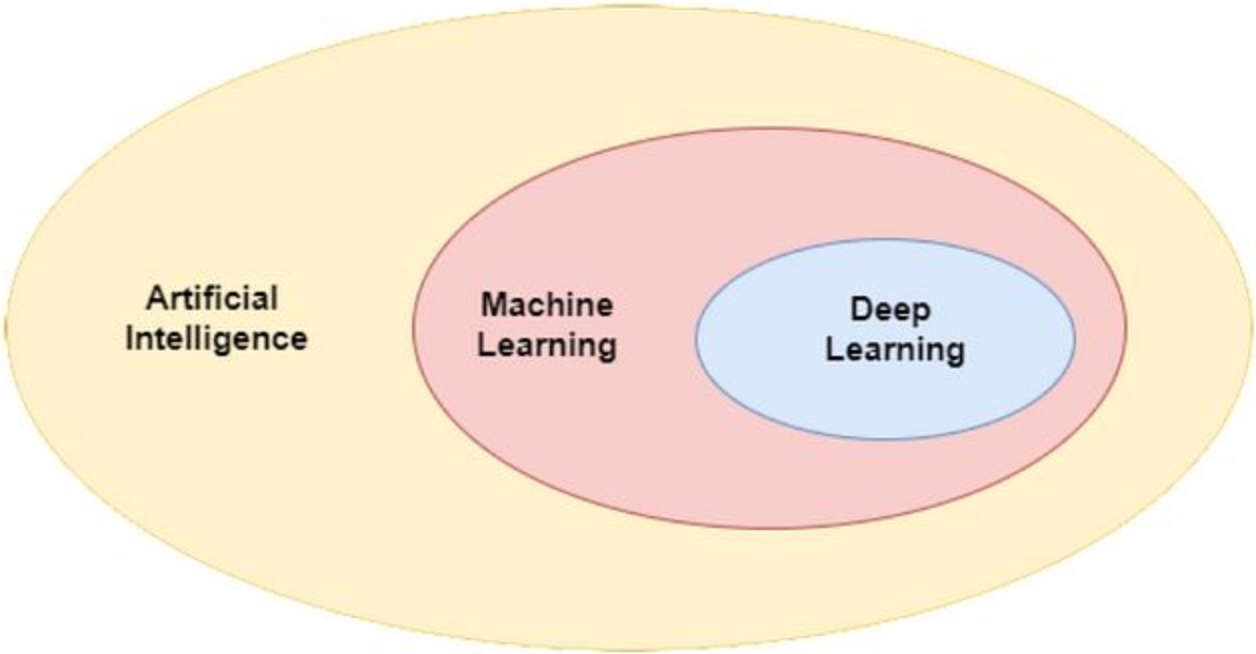


Identification of Hepatotoxicity in Small Molecules



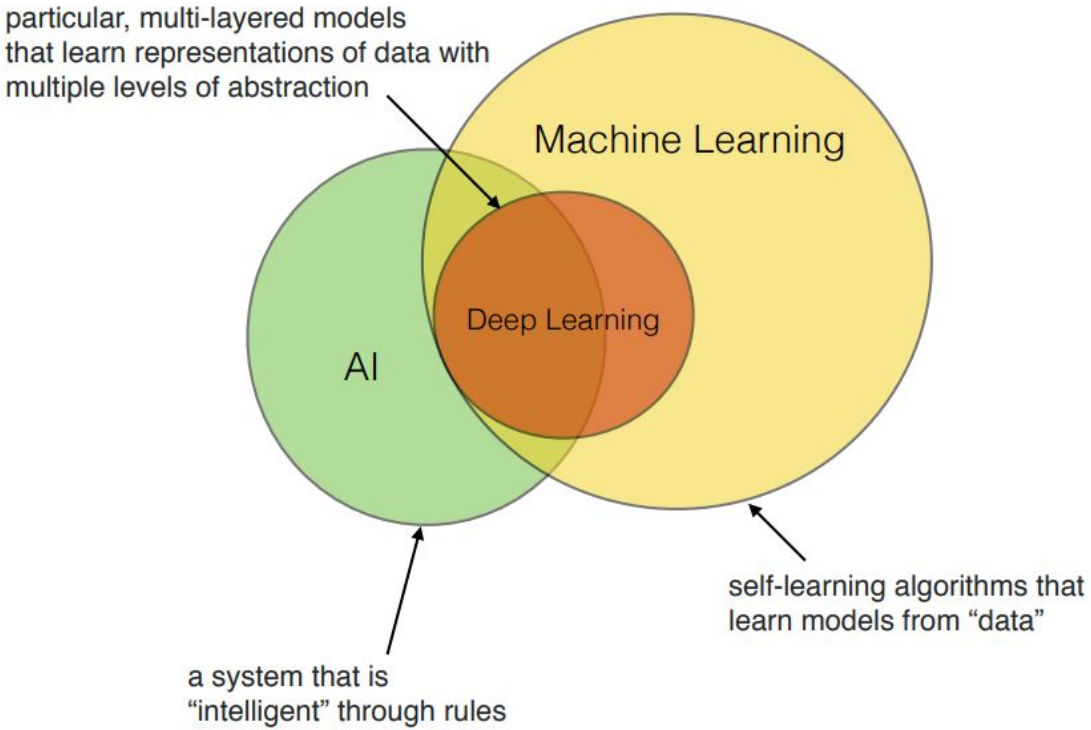
- **Task**: predicting hepatotoxicity from small molecules.
- **Performance**: percentage of molecules classified correctly as toxic.
- **Experience**: dataset of small molecules experimentally distinguishing them between toxic and non-toxic for the liver.

Taxonomy of Machine Learning



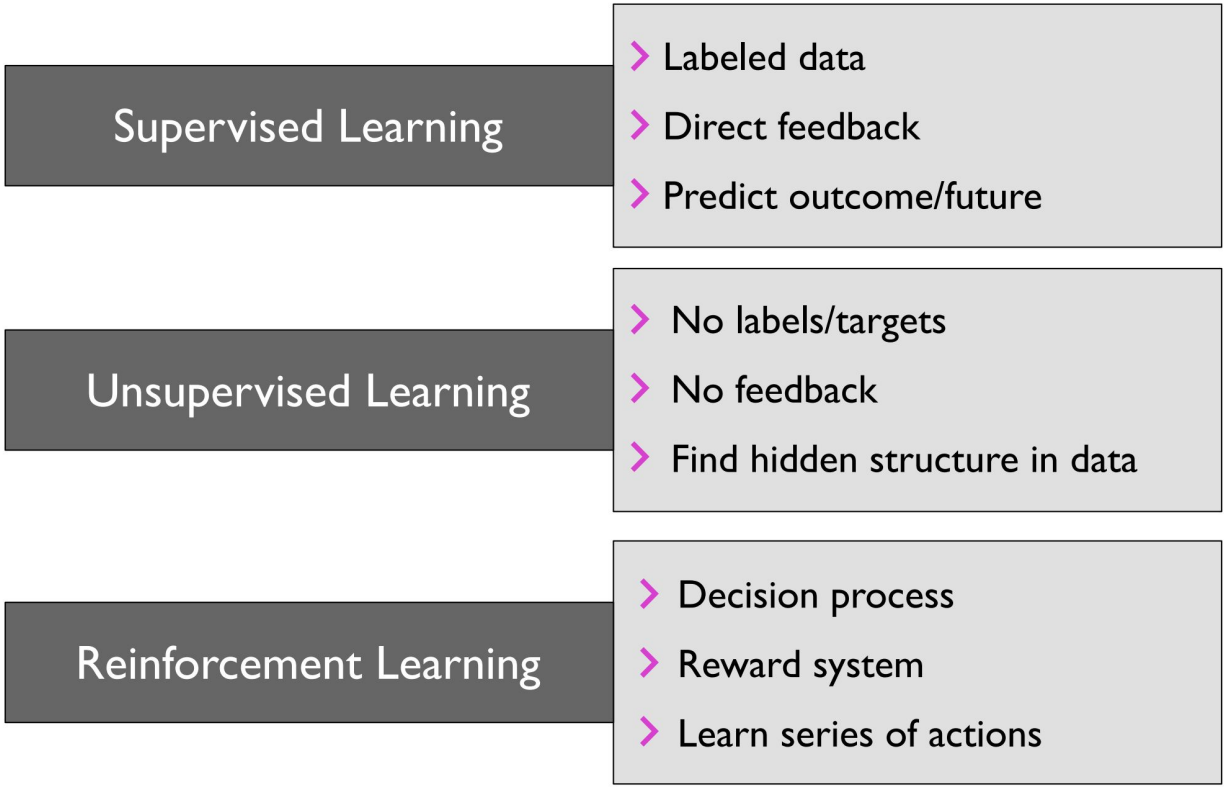
Towards Data Science, 2020

Taxonomy of Machine Learning



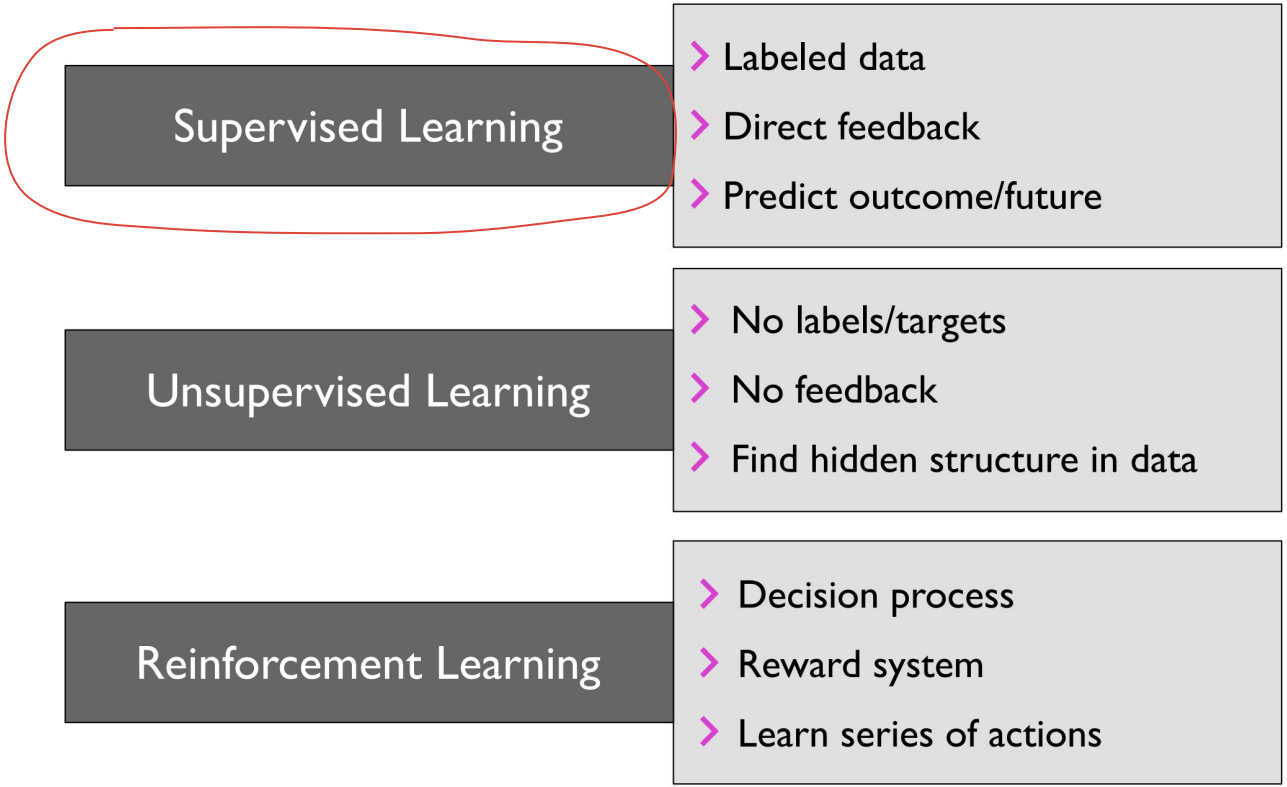
Raschka, 2021

Taxonomy of Machine Learning



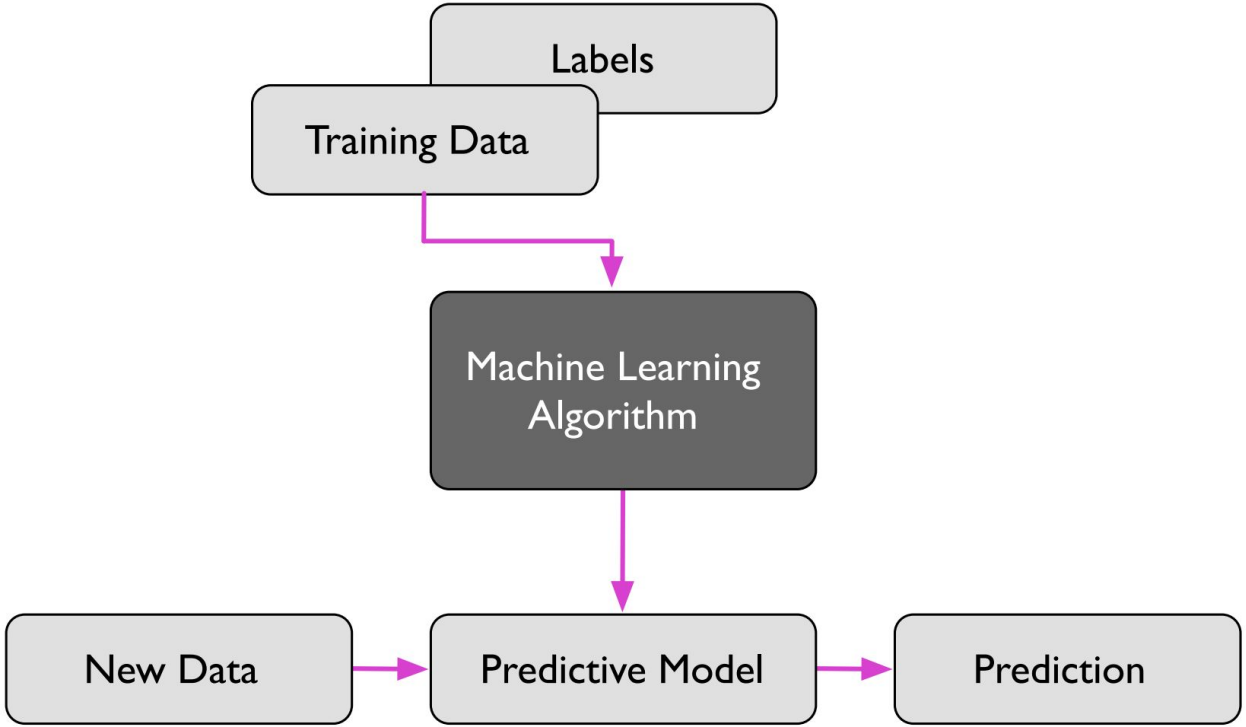
Raschka, 2021

Taxonomy of Machine Learning



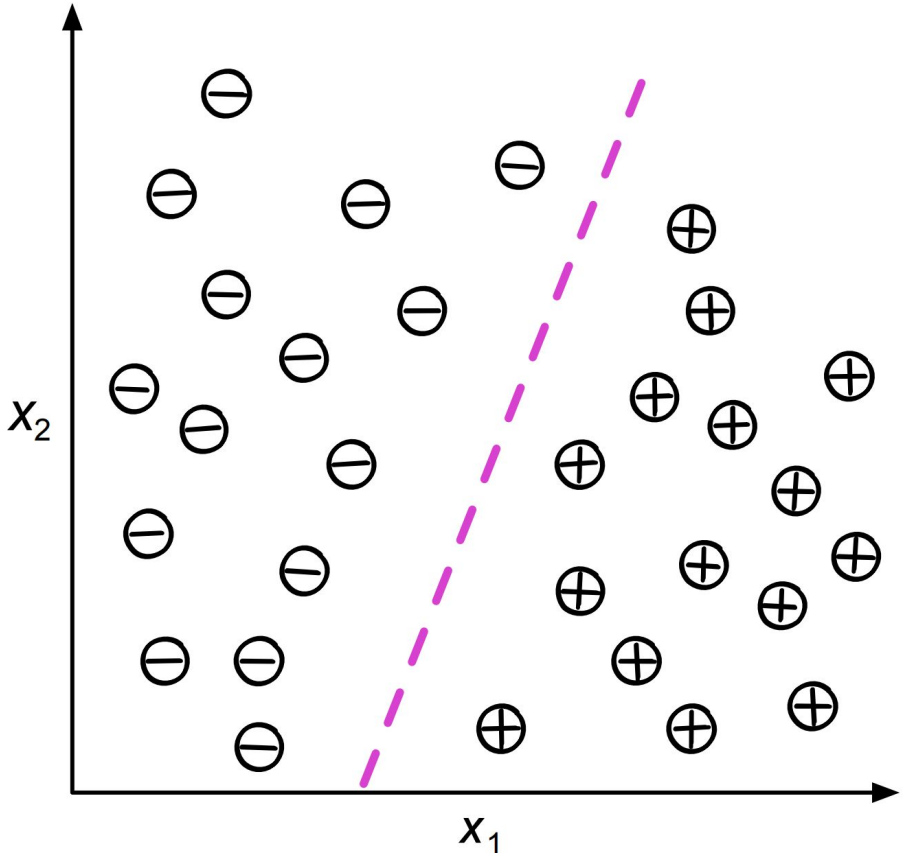
Raschka, 2021

Supervised Learning



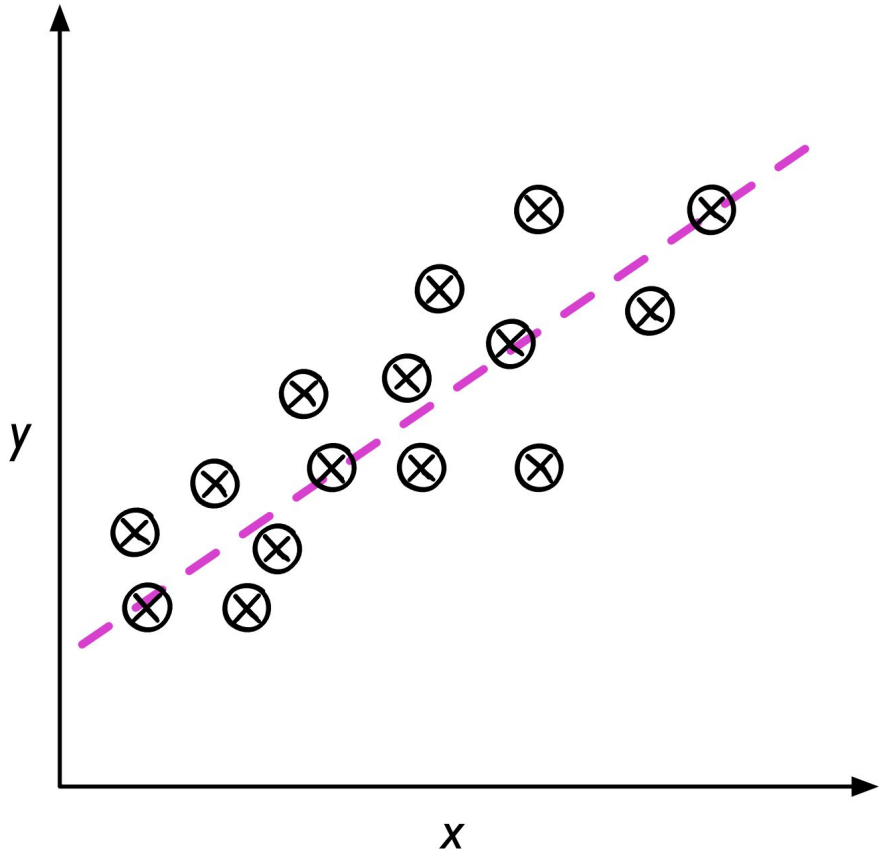
Raschka, 2021

Supervised Learning - Classification



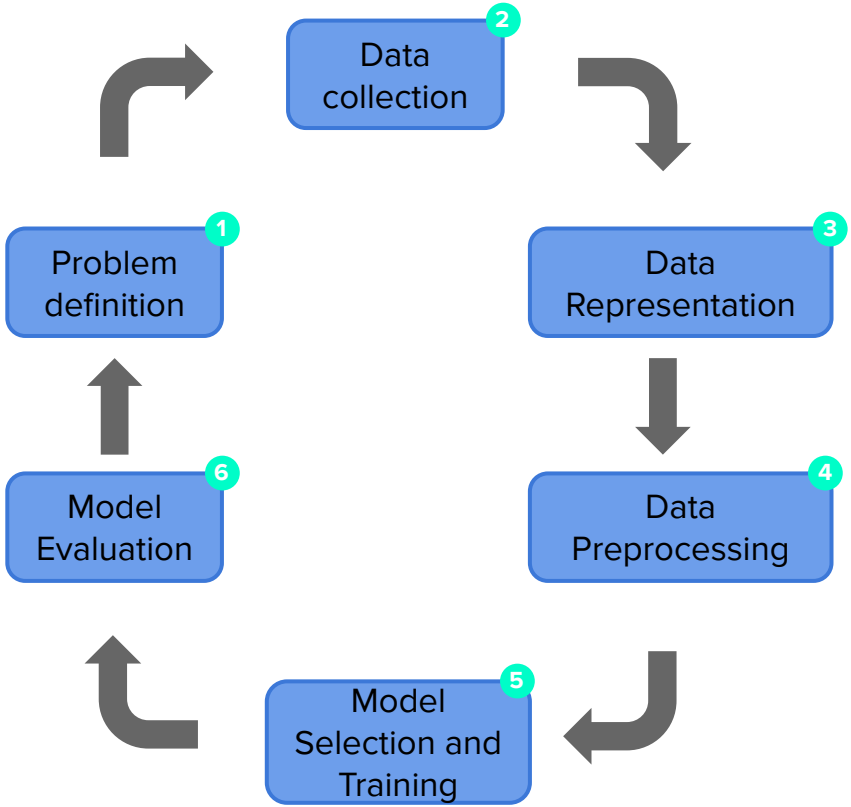
Raschka, 2021

Supervised Learning - Regression



Raschka, 2021

Machine Learning Workflow



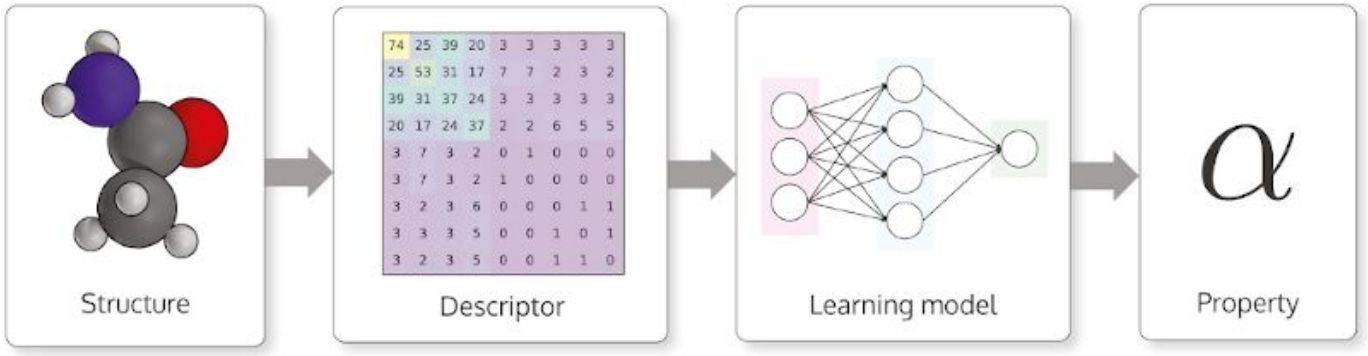
Data Representation

- Most of the recent Machine Learning tools (e.g., scikit-learn, etc) only accept numerical matrices (or dataframes) as inputs.

- We need to find ways to represent our biological, chemical, human data, etc in a numerical way.

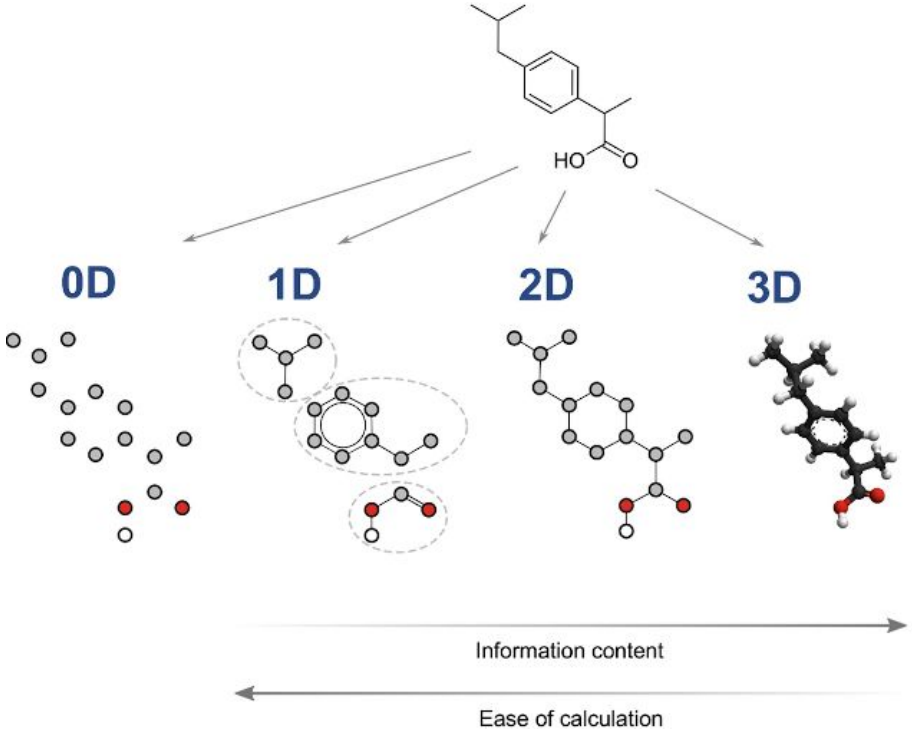
Data Representation - Small Molecules

Given the structure of the molecule, we are able to derive a list of descriptors aiming to characterise this input molecule to predict a given property



Chem Intelligence, 2021

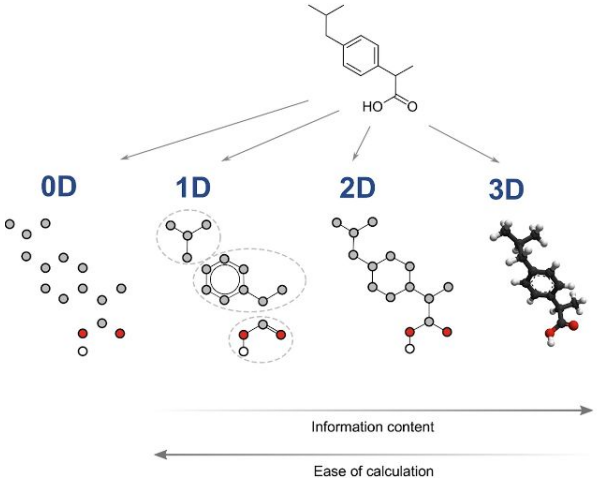
Data Representation - Small Molecules



Chem Intelligence, 2021

Data Representation - Small Molecules

- **0D Descriptors:** No information about structure and connectivity.
  - Atom counts, or molecular weights
- **1D Descriptors:** Partial information about the structure and connectivity.
  - Fingerprints.
- **2D Descriptors:** Information on molecular topology based on the graph representation.
  - Atom distance matrix.
- **3D Descriptors:** Information about the spatial coordinates of atoms of a molecule
  - 3D fingerprints.



Chem Intelligence, 2021

RDKit
Open-Source Cheminformatics and Machine Learning

3D Representation:

Axen et al., 2017

Sequence-based descriptors

```
MAALSGGGGGGAEPGQALFNGDMEPEAGAGAGAAASSAADPAIPEEVWNIKQMIKLTQEH
IEALLDKFGGEHNPPSIYLEAYEEYTSKLDALQQREQQLLESLGNGTDFSVSSSASMDTV
TSSSSSSLSVLPSSLSVFQNPTDVARSNPKSPQKPIVRVFLPNKQRTVVPARCGVTVRDS
LKKALMMRGLIPECCAVYRIQDGEKKPIGWDTDISWLTGEELHVEVLENVPLTTHNFVRK
```

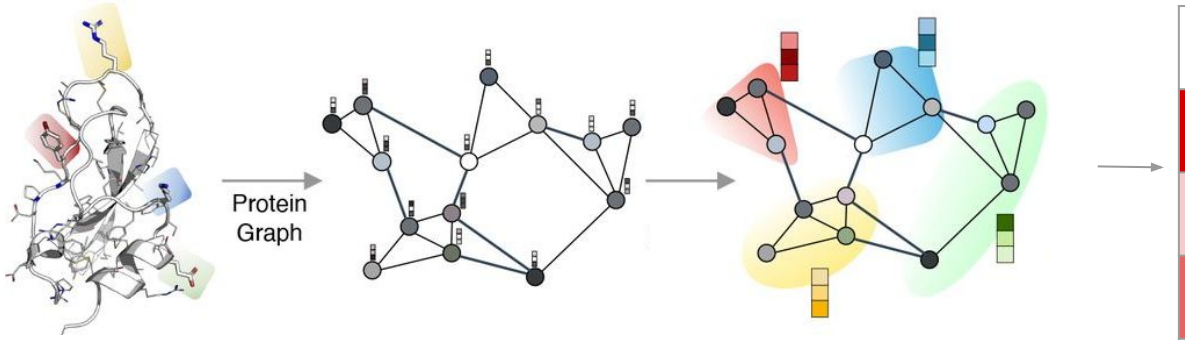
- Amino acid composition
- Physicochemical properties
- Disorder propensity scores
- …

#	A	C	D	E	F
P31946	0.085	0.008	0.057	0.122	0.024
P62258	0.098	0.012	0.086	0.114	0.020
Q04917	0.098	0.012	0.069	0.114	0.028
P61981	0.093	0.012	0.069	0.113	0.012
P31947	0.105	0.008	0.060	0.133	0.020
P27348	0.098	0.020	0.073	0.106	0.024

iFeature, 2018

Data Representation - Proteins

- Structure-based descriptors:



Sanyal et al., 2021

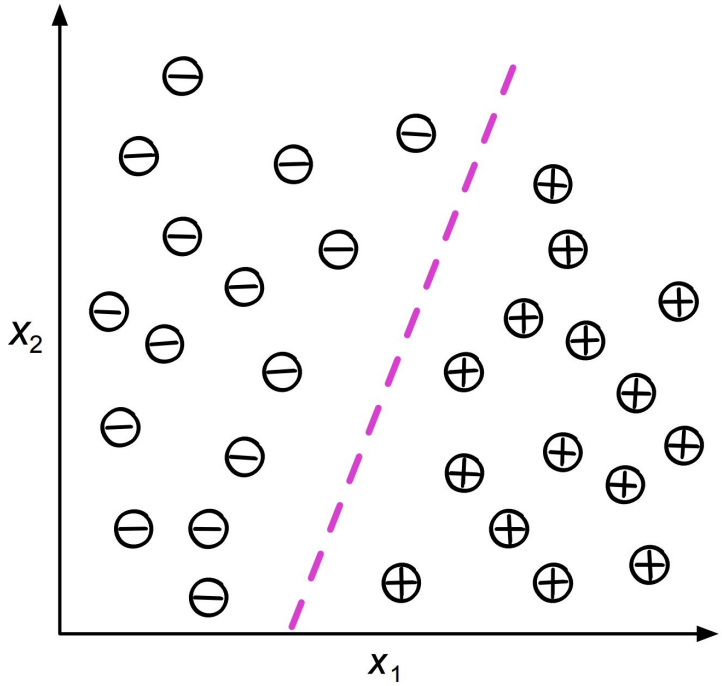
D1	D2	…	D_{n-1}	D_n
1.0	100		1.5	1
0.1	1000		1.5	0
0.5	10		20.9	2

- residue depth
- solvent accessible surface area
- secondary structure distribution
- torsion angles
- Cumulative pair distances between pharmacophore atom groups

Focusing on Classification

Decision Boundary

Definition: A decision boundary, is a surface that separates data points belonging to different class labels. (Sahu, 2021)

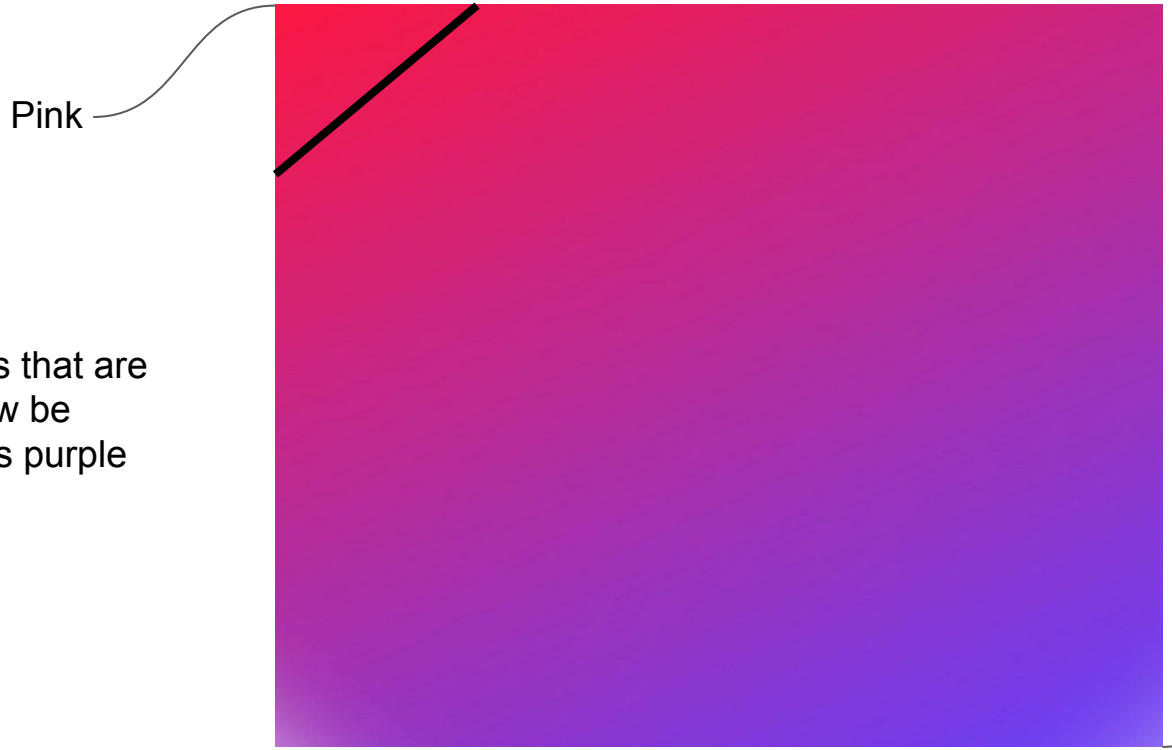


Decision Boundary

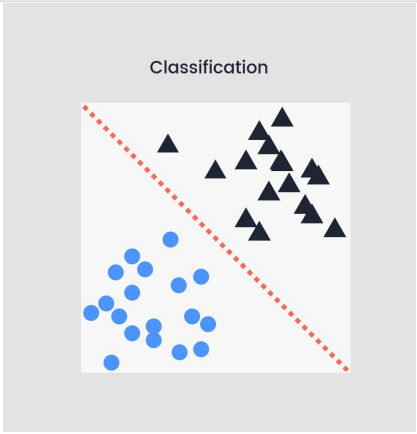
Binary
Classification

https://www.vectorstock.com/royalty-free-vector/smooth-transition-from-pink-to-purple-abstract-vector-34503450

Decision Boundary



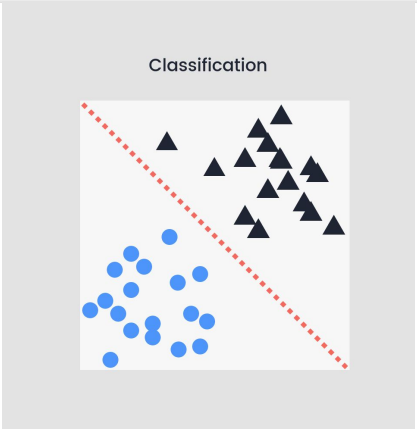
Many things that are pink will now be classified as purple



https://www.vectorstock.com/royalty-free-vector/smooth-transition-from-pink-to-purple-abstract-vector-34503450

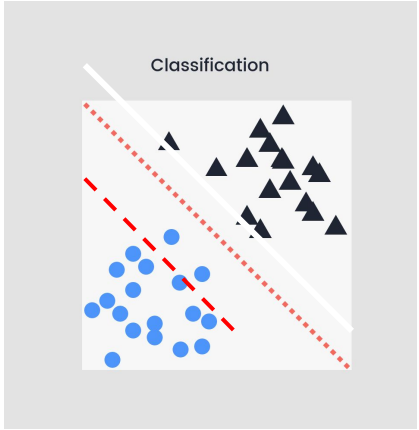
Decision Boundary

Many things that are pink will now be classified as purple



https://www.vectorstock.com/royalty-free-vector/smooth-transition-from-pink-to-purple-abstract-vector-34503450

Decision Boundary



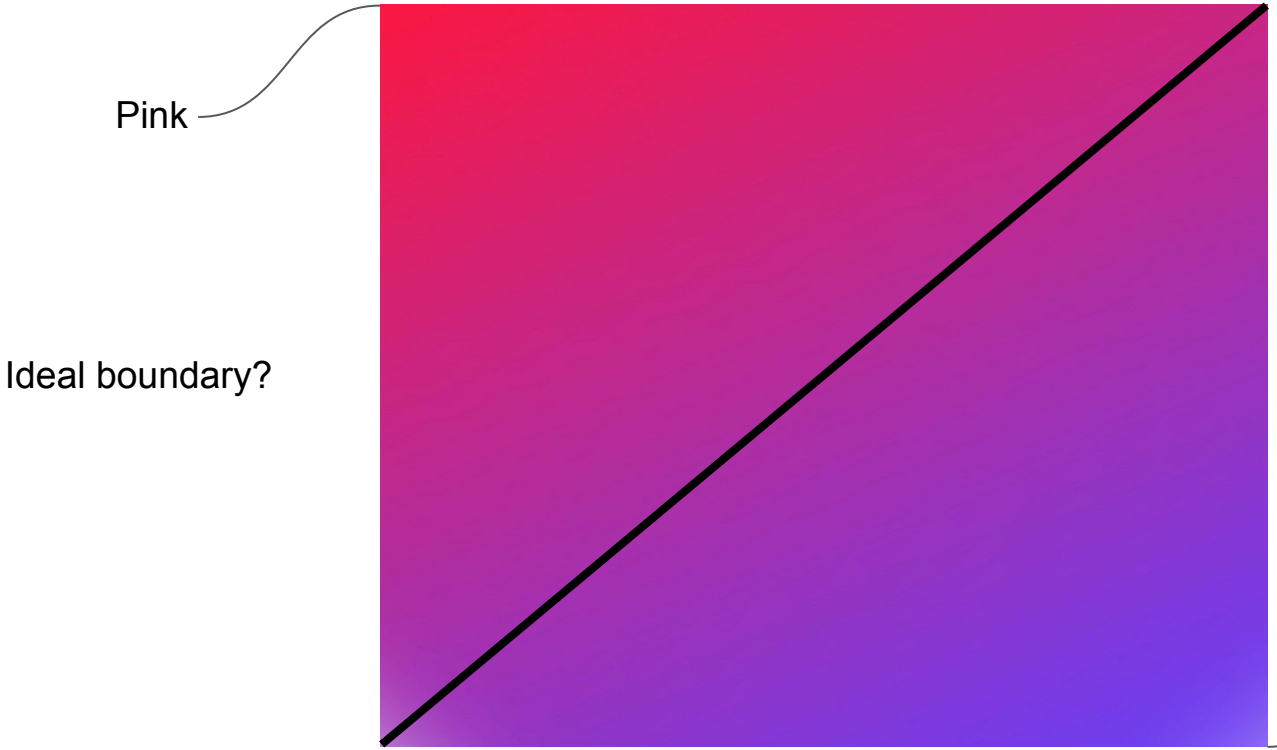
Pink

Many things that are purple will now be classified as pink

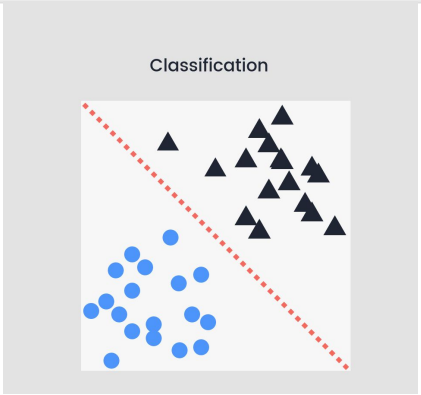
Purple

https://www.vectorstock.com/royalty-free-vector/smooth-transition-from-pink-to-purple-abstract-vector-34503450

Decision Boundary



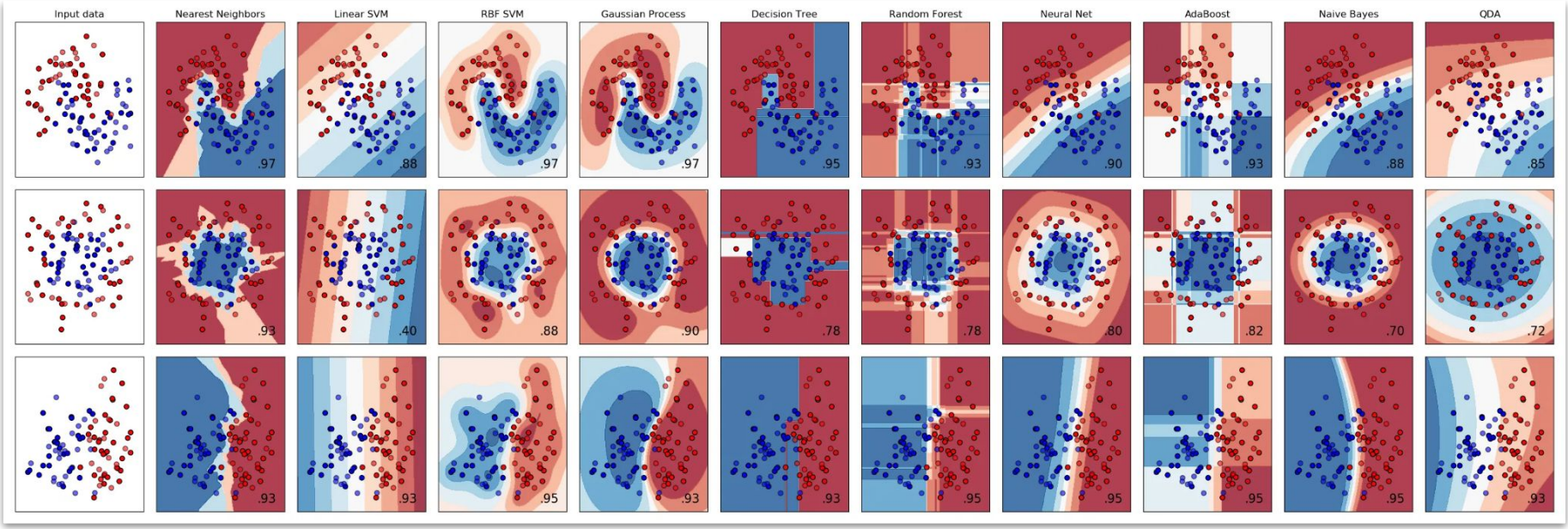
Ideal boundary?



https://www.vectorstock.com/royalty-free-vector/smooth-transition-from-pink-to-purple-abstract-vector-34503450

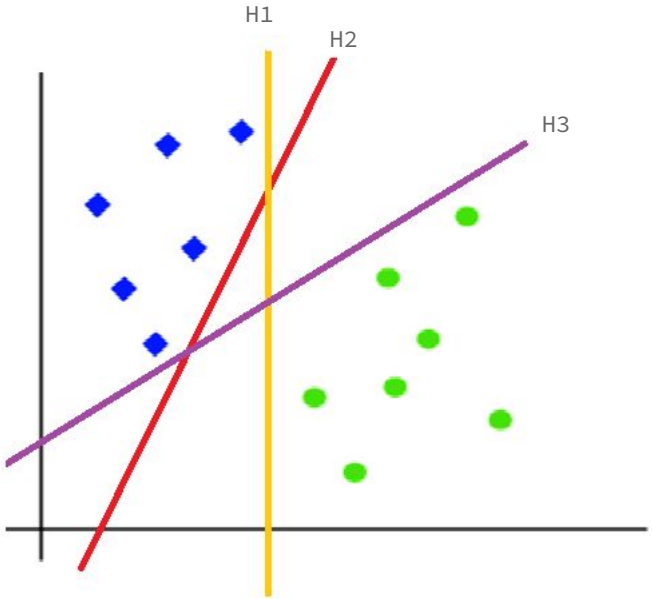
Decision Boundary

Comparison of the decision boundaries of 10 machine learning models:



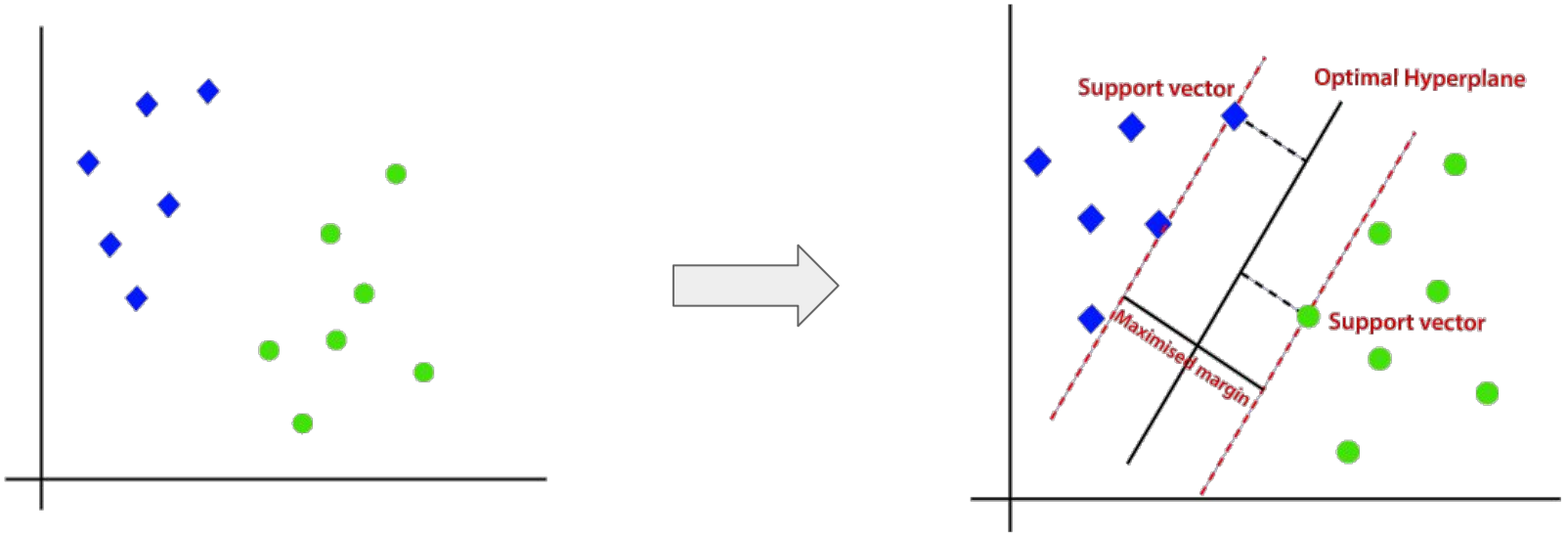
Varoquaux and Müller

Classical Support Vector Classifier (SVC)



Saini,2023

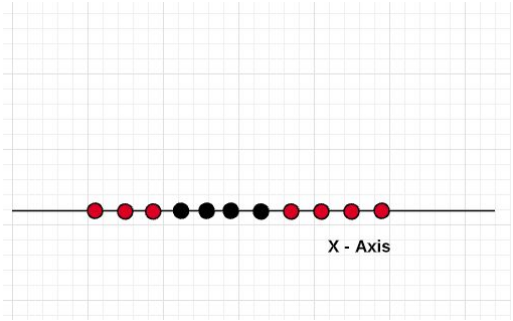
Classical Support Vector Classifier (SVC)



Saini,2023

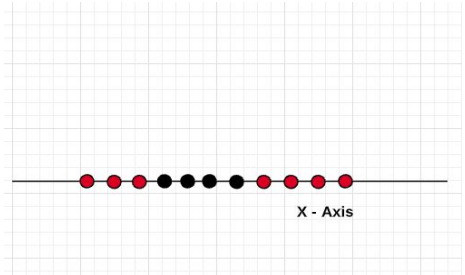
Classical Support Vector Classifier (SVC)

Support Vector Machines were developed to deal with linear data.
What happens when we take data like:



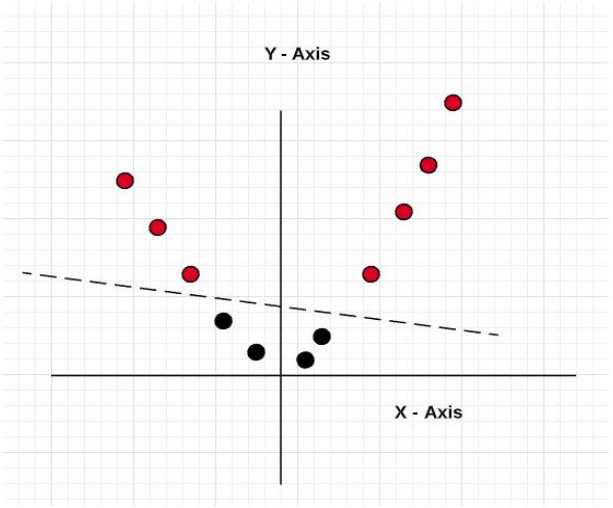
Non-linearly separable data

Classical Support Vector Classifier (SVC)



ϕ

$y = x^2$

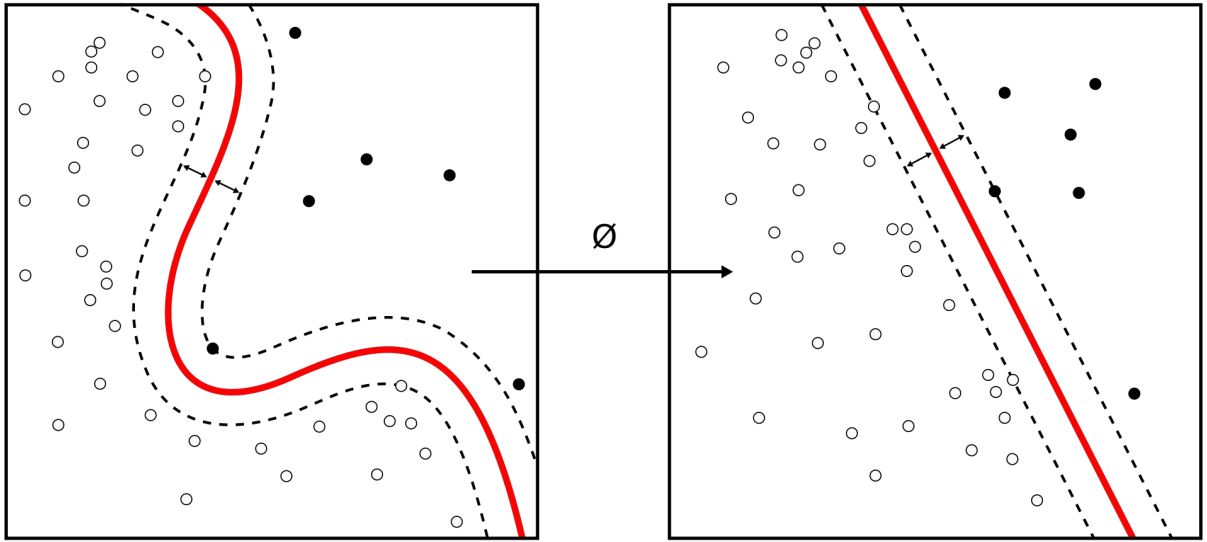


Non-linearly separable data

Support Vector Machines have a key component called **kernel machine.**

Classical Support Vector Classifier (SVC)

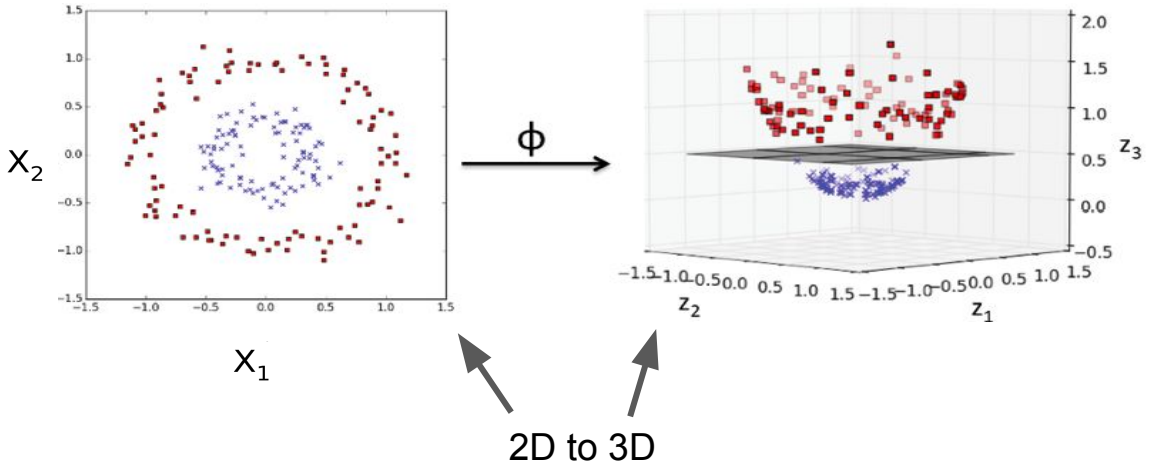
A **Kernel Function** manipulates the training data to transform a non-linear lower dimension space into a higher dimension space, which we can get a linear decision boundary



SVM - Wikipedia

Classical Support Vector Classifier (SVC)

A **Kernel Function** manipulates the training data to transform a non-linear lower dimension space into a higher dimension space, which we can get a linear decision boundary



SVM - Wikipedia

Github Repository

https://github.com/alexgcsa/incob2023

ML Resources

- Tom Mitchell's Book and Youtube Course:
  - https://www.cs.cmu.edu/~tom/mlbook.html
  - https://www.youtube.com/watch?v=m4NlfvrRCdg&list=PLl-BBnDxtUt1hLXmIwu27P22bTi6VwMkN
- Sebastian Raschka's Course:
  - https://sebastianraschka.com/blog/2021/ml-course.html
- Andrew Ng's Course:
  - https://www.coursera.org/specializations/machine-learning-introduction

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

Baker
HEART & DIABETES INSTITUTE

~~Quantum~~ Machine Learning

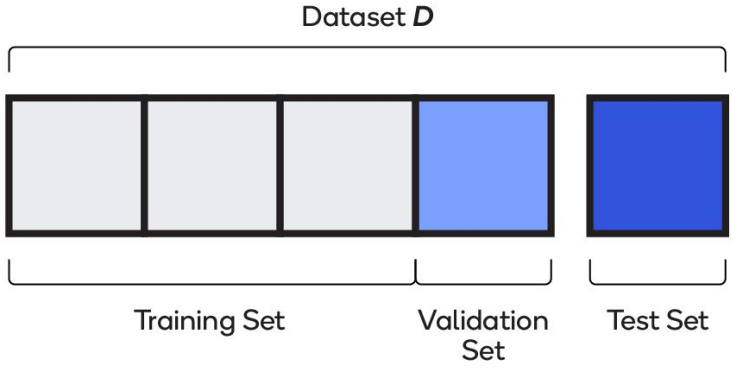
Alex de Sá — Alex.deSa@baker.edu.au
Ashar Malik — Ashar.Malik@baker.edu.au

https://github.com/alexgcsa/incob2023

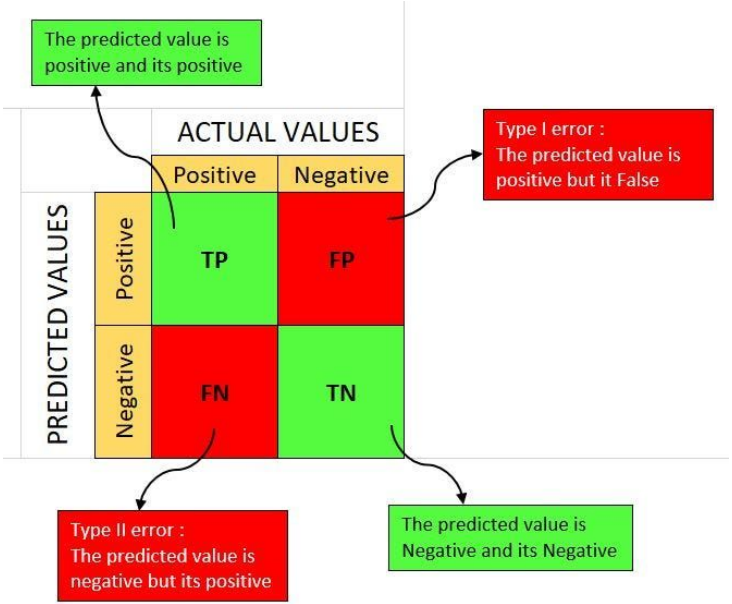
K-Fold Cross-Validation

Train/Test

Test the model on new data, assessing its generalisation



Classification Metrics

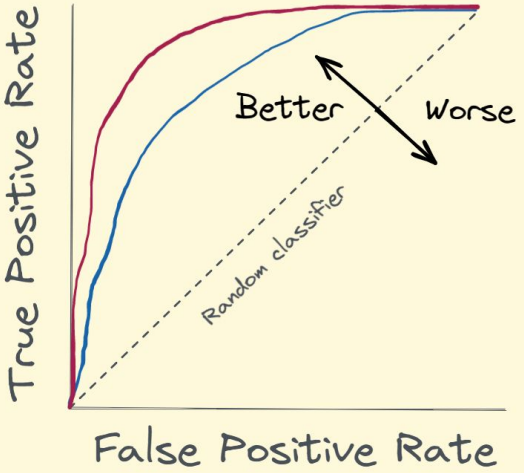


$$\text{Accuracy} = \frac{TP + TN}{\text{Total Samples}}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$



[Towards Data Science, 2023]

[Medium, 2020]