



THE UNIVERSITY
OF QUEENSLAND
AUSTRALIA



QUANTUM MACHINE LEARNING

Alex de Sá
Ashar Malik

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

<https://github.com/alexgcsa/incob2023>



THE UNIVERSITY
OF QUEENSLAND
AUSTRALIA



~~QUANTUM~~ MACHINE LEARNING

Alex de Sá
Ashar Malik

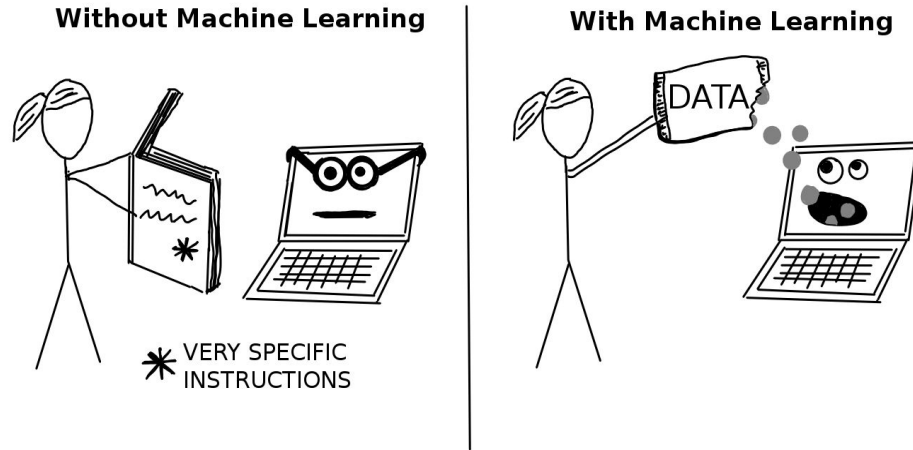
Alex.deSa@baker.edu.au
Ashar.Malik@baker.edu.au

<https://github.com/alexgcsa/incob2023>

MACHINE LEARNING

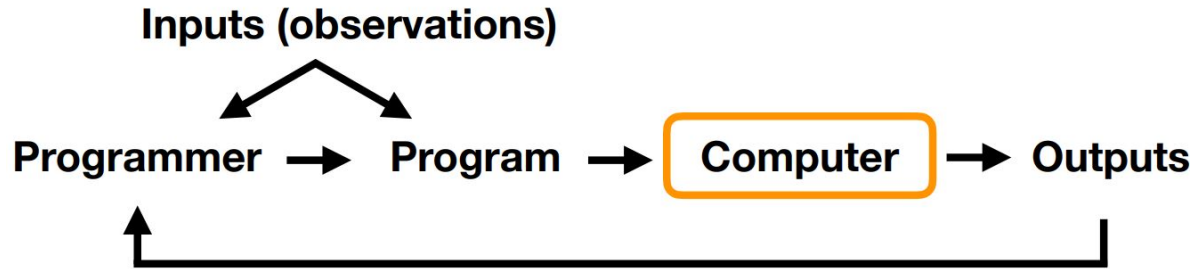
Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed.

Arthur L. Samuel, AI pioneer, 1959

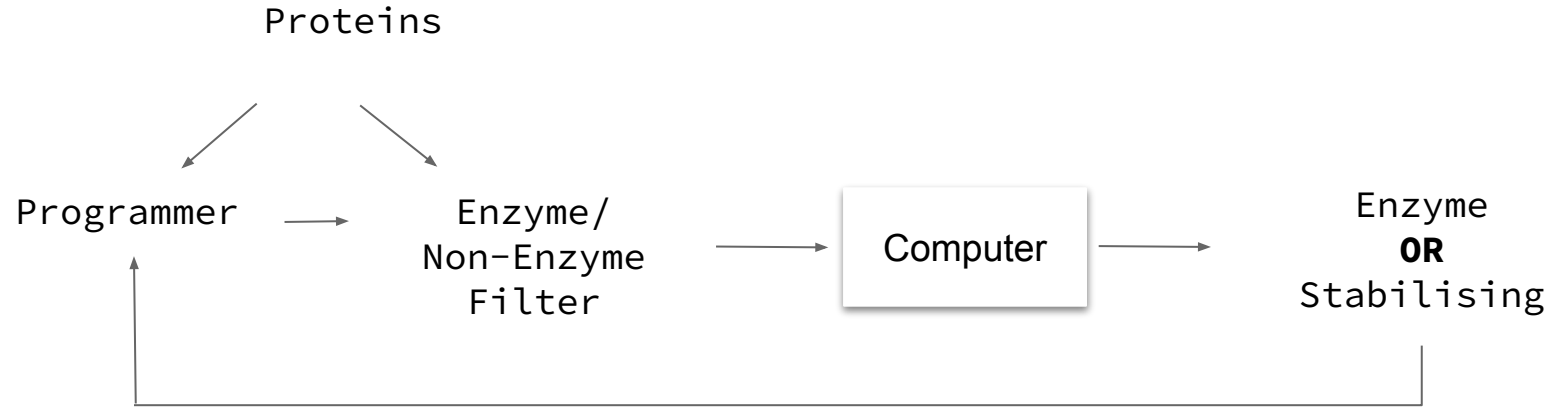


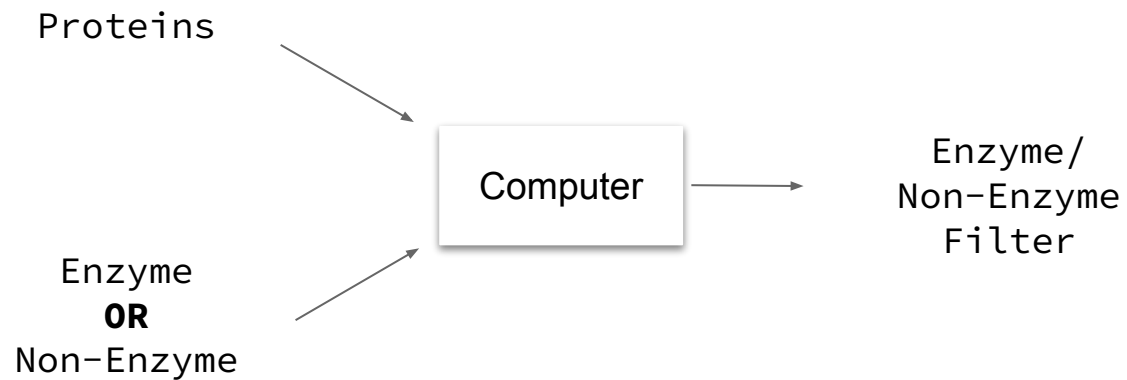
[Molnar, 2021](#)

TRADITIONAL PROGRAMMING VERSUS MACHINE LEARNING



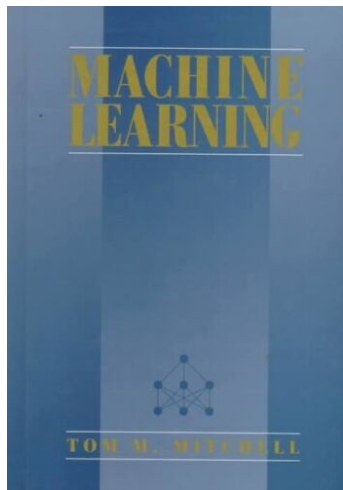
TRADITIONAL PROGRAMMING





MACHINE LEARNING DEFINITION

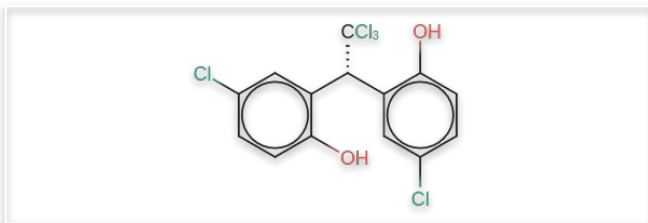
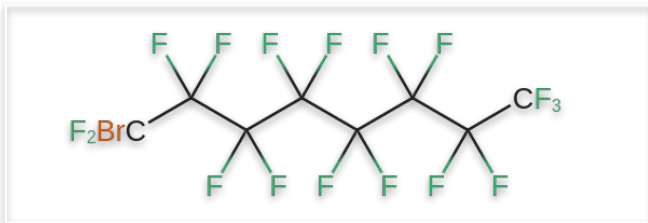
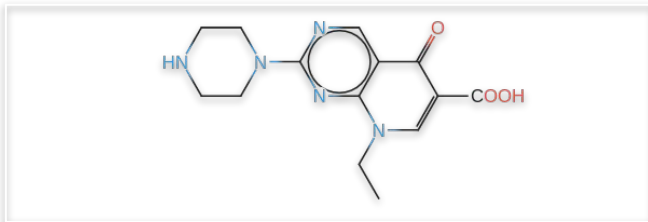
A computer program is said to **learn** from experience **E** with respect to some class of tasks **T** and performance measure **P**, if its performance at tasks in **T**, as measured by **P**, improves with experience **E**.”



Tom Mitchell, Professor at Carnegie Mellon University

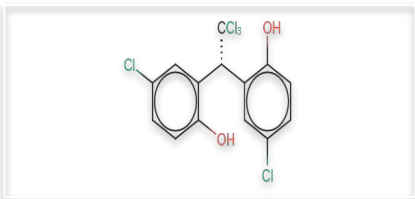
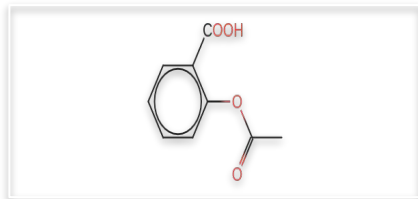
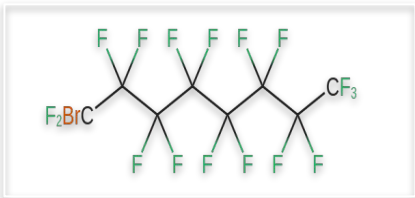
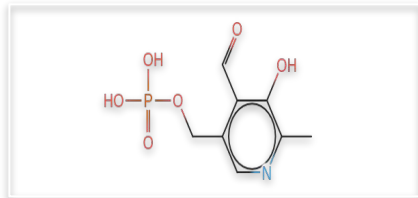
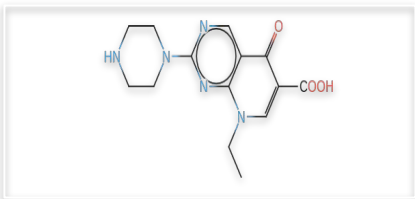
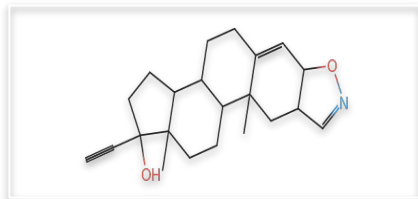
<https://www.cs.cmu.edu/~tom/mlbook.html>

Identification of Hepatotoxicity in Small Molecules



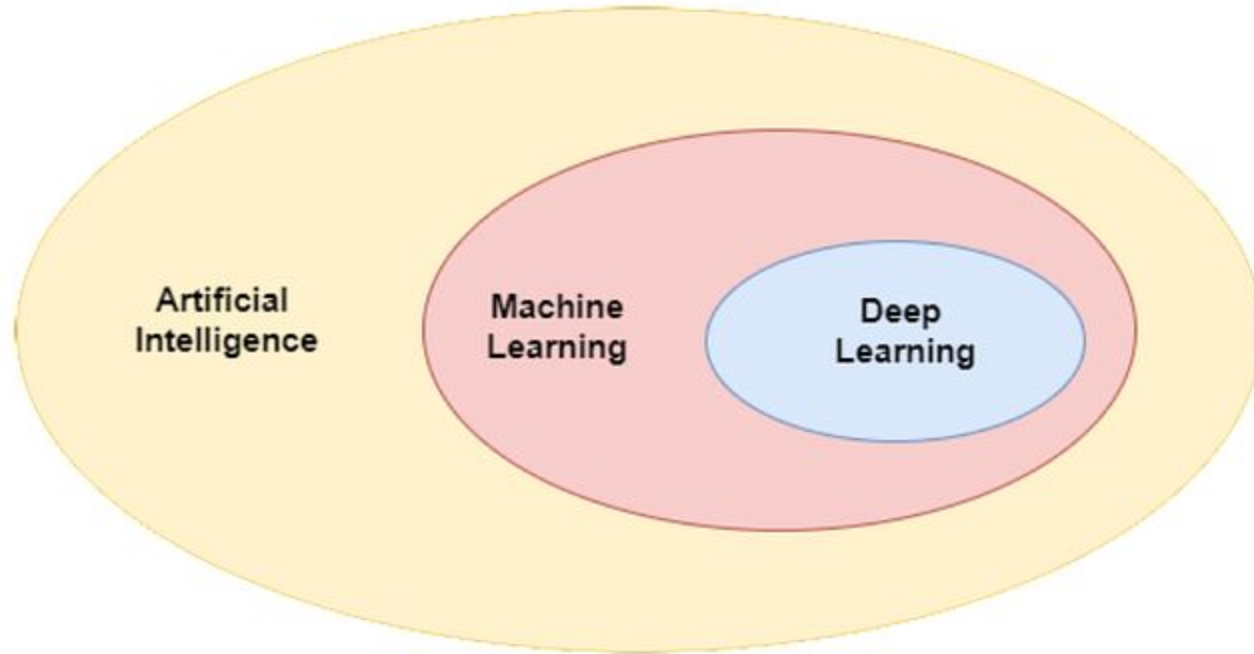
EXAMPLE - MACHINE LEARNING DEFINITION

Identification of Hepatotoxicity in Small Molecules



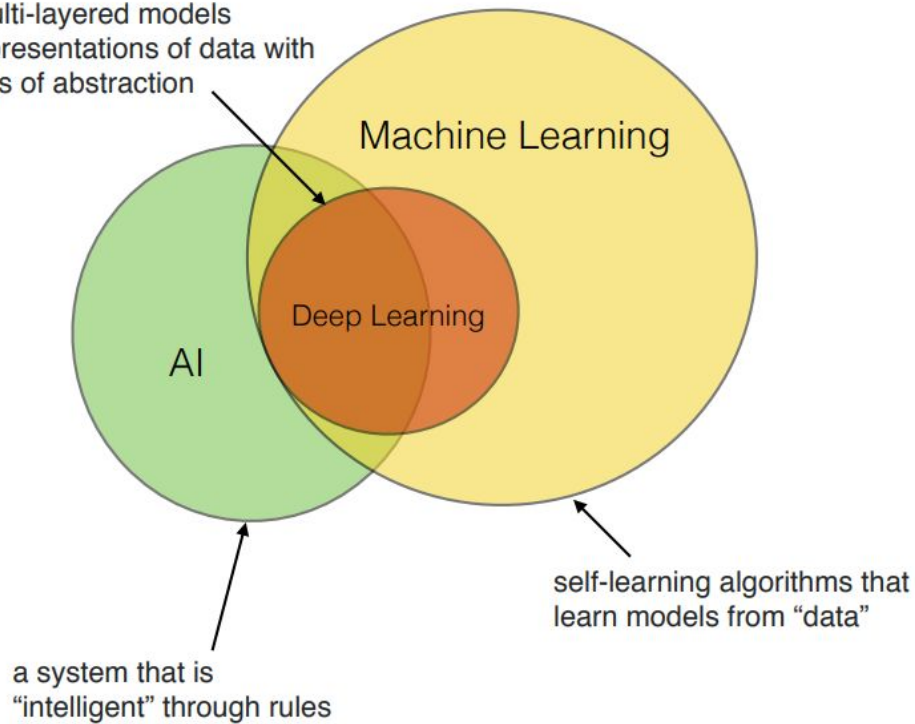
- **Task:** predicting hepatotoxicity from small molecules.
- **Performance:** percentage of molecules classified correctly as toxic.
- **Experience:** dataset of small molecules experimentally distinguishing them between toxic and non-toxic for the liver.

TAXONOMY OF MACHINE LEARNING



TAXONOMY OF MACHINE LEARNING

particular, multi-layered models
that learn representations of data with
multiple levels of abstraction



TAXONOMY OF MACHINE LEARNING

Supervised Learning

- Labeled data
- Direct feedback
- Predict outcome/future

Unsupervised Learning

- No labels/targets
- No feedback
- Find hidden structure in data

Reinforcement Learning

- Decision process
- Reward system
- Learn series of actions

TAXONOMY OF MACHINE LEARNING

Supervised Learning

- Labeled data
- Direct feedback
- Predict outcome/future

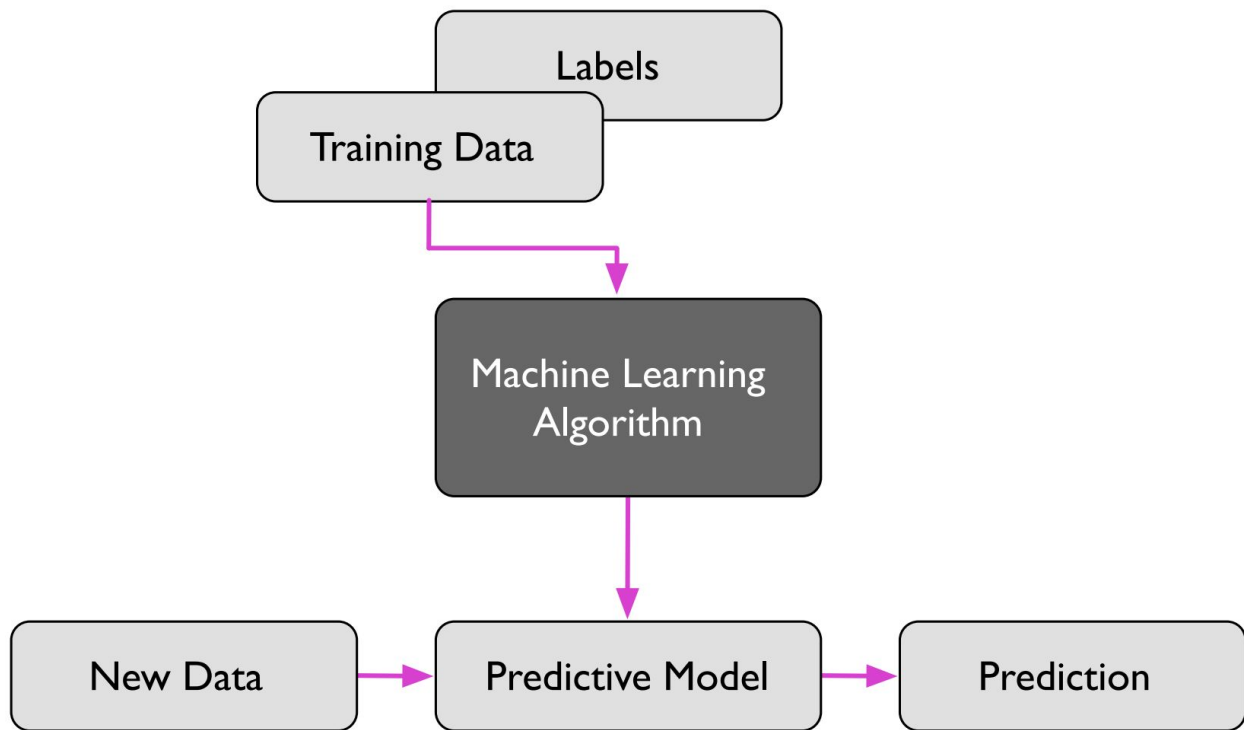
Unsupervised Learning

- No labels/targets
- No feedback
- Find hidden structure in data

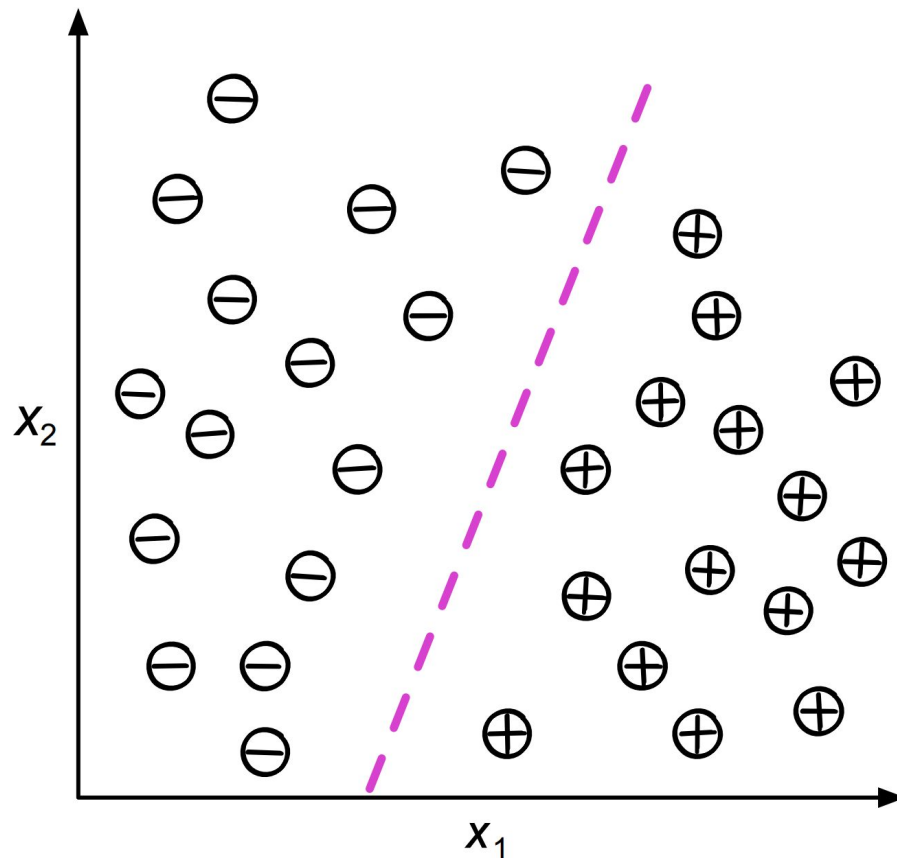
Reinforcement Learning

- Decision process
- Reward system
- Learn series of actions

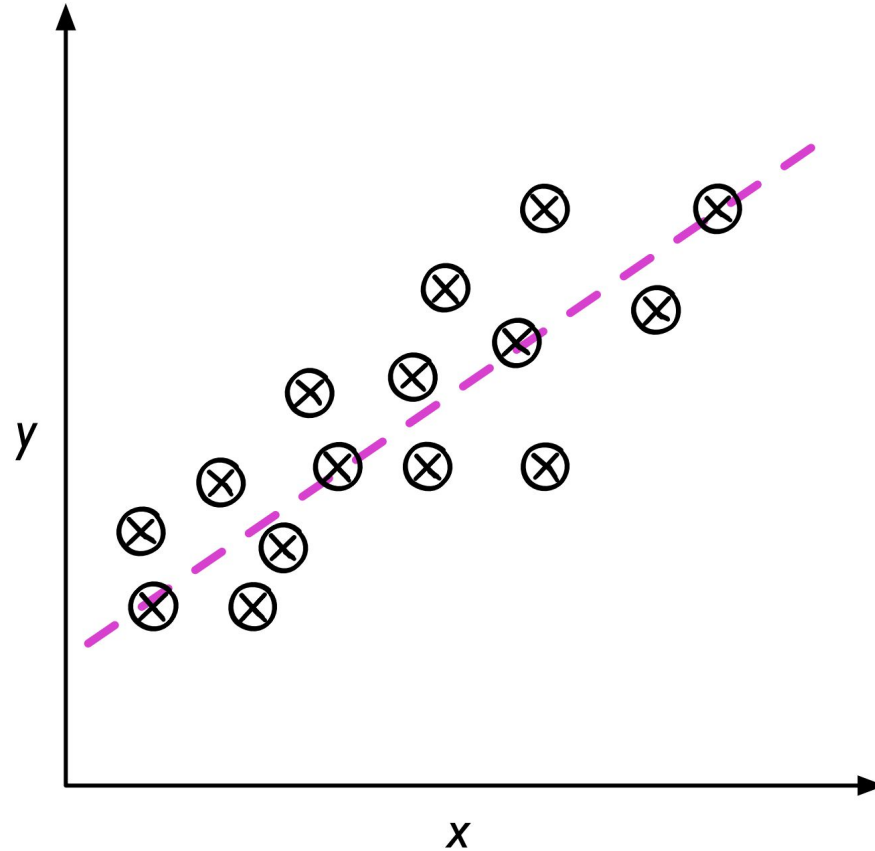
SUPERVISED LEARNING



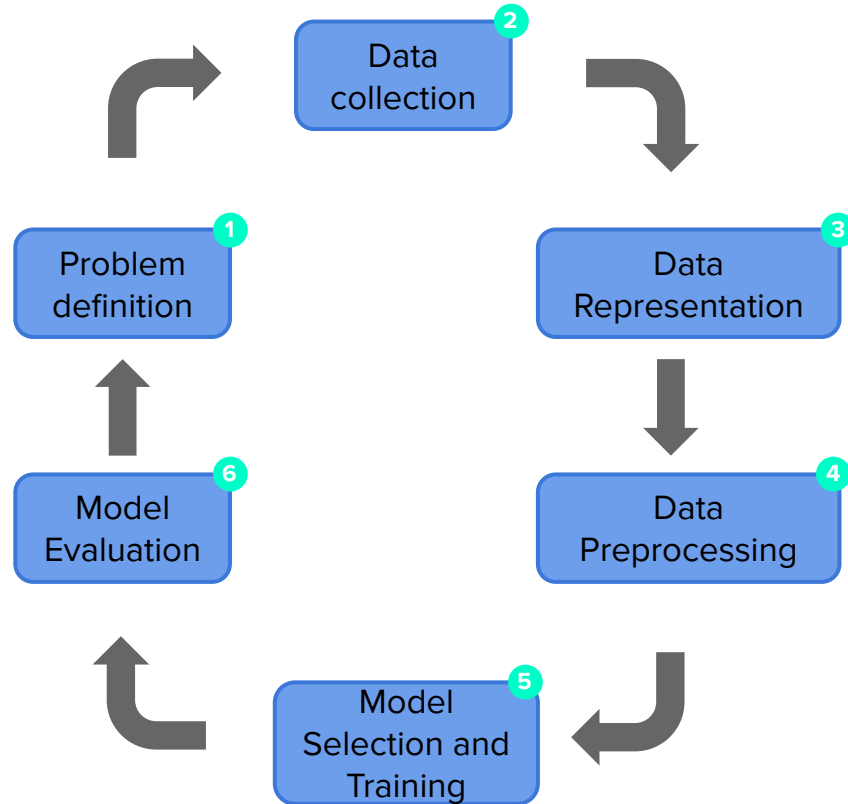
SUPERVISED LEARNING - CLASSIFICATION



SUPERVISED LEARNING - REGRESSION



MACHINE LEARNING WORKFLOW

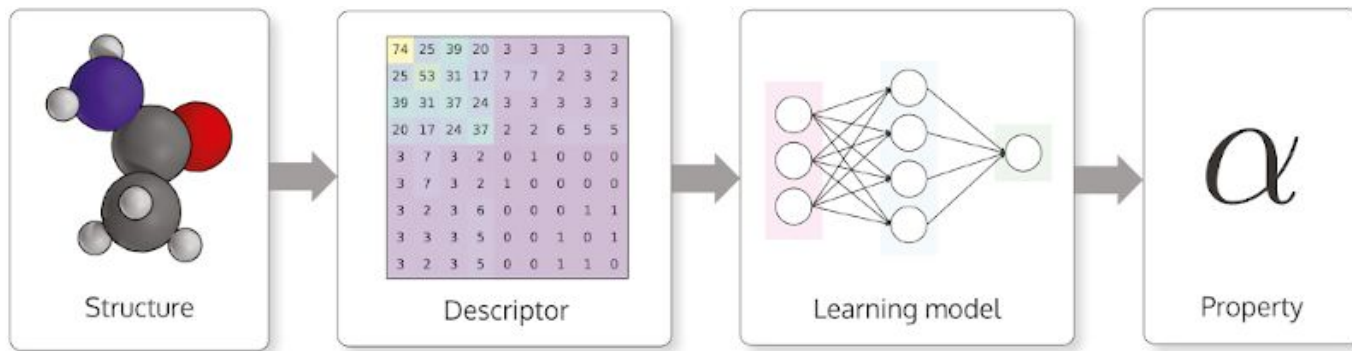


DATA REPRESENTATION

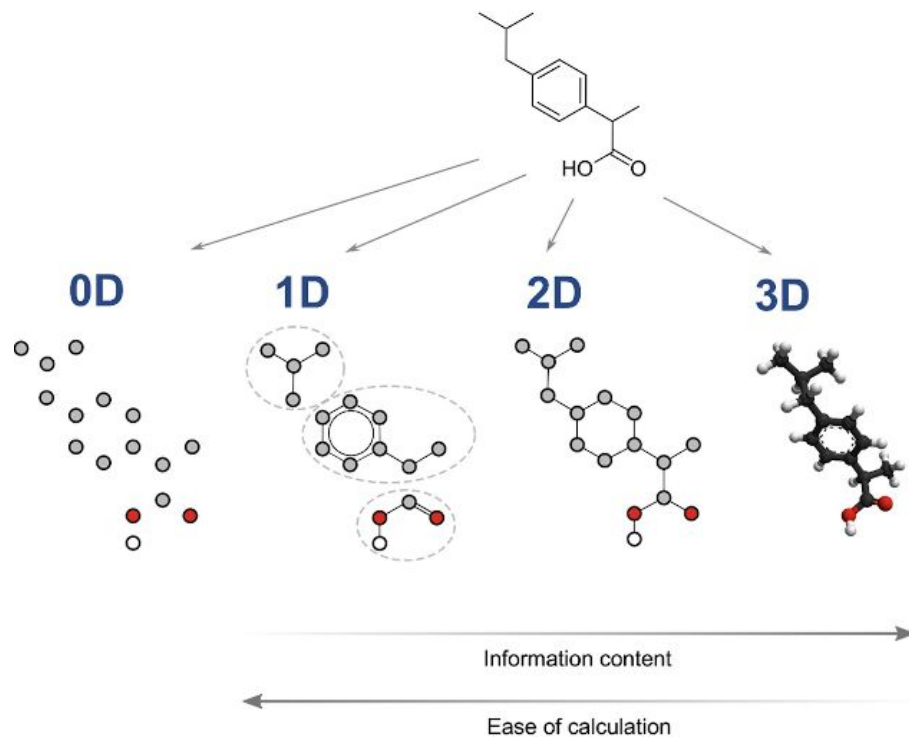
- Most of the recent Machine Learning tools (e.g., scikit-learn, etc) only accept numerical matrices (or dataframes) as inputs.
- We need to find ways to represent our biological, chemical, human data, etc in a numerical way.

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



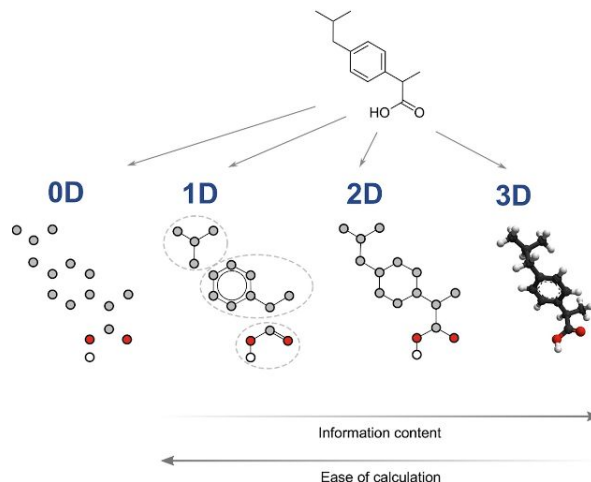
DATA REPRESENTATION - SMALL MOLECULES



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



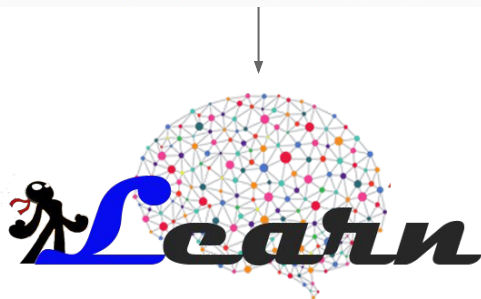
Open-Source Cheminformatics
and Machine Learning

3D Representation:

Axen et al., 2017

Sequence-based descriptors

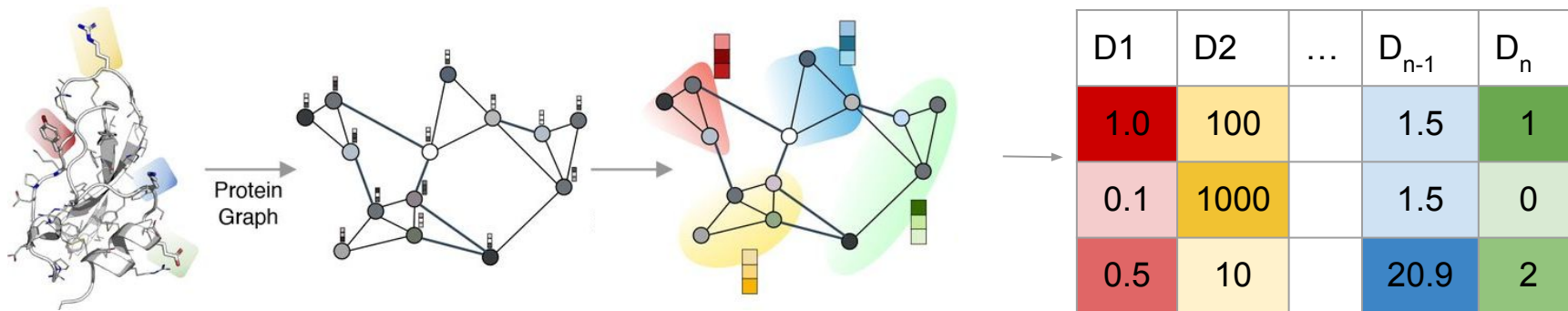
MAALSGGGGGAEPGQALFNGDMEPEAGAGAGAAAASSAADPAIPEEVWNIQMILKTQEH
IEALLDKFGGEHNPPSIYLEAYEYTSKLDALQQREQQLLESLGNGTDFSVSSSASMDTV
TSSSSSSLSVLPSSLSVFNPTDVARSNPKSPQKPIVRVFLPNKQRTVVPARCGVTVRDS
LKKALMMRGLIPECCAVYRIQDGEKKPIGWDTDISWLTEELHVEVLENVPLTTHNFVRK



- 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 | 0.089 | 0.114 | 0.028 |
| P61981 | 0.093 | 0.012 | 0.089 | 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 |

- Structure-based descriptors:



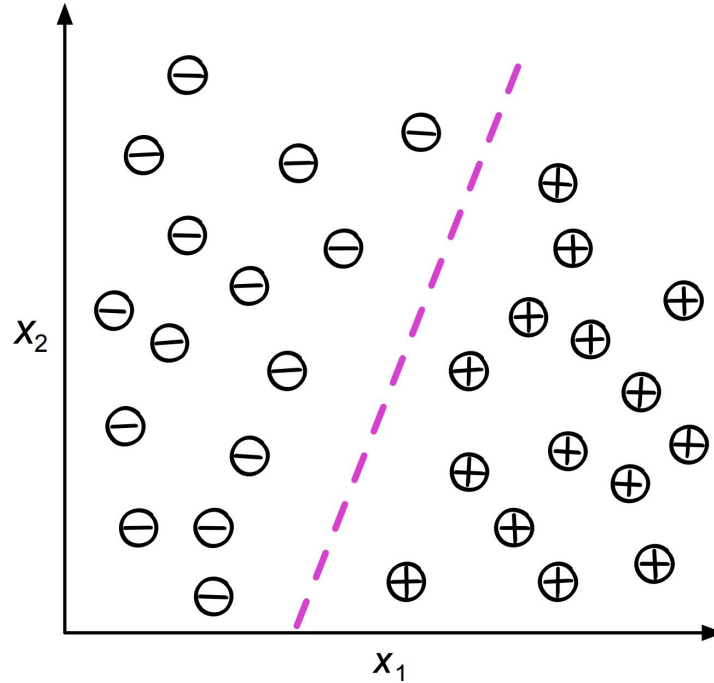
Sanyal et al., 2021

- 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

Pink

Purple



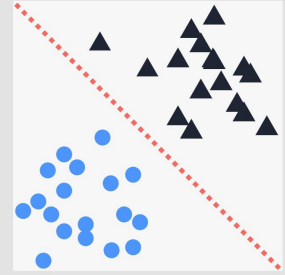
DECISION BOUNDARY

Pink

Many things that are
pink will now be
classified as purple



Classification



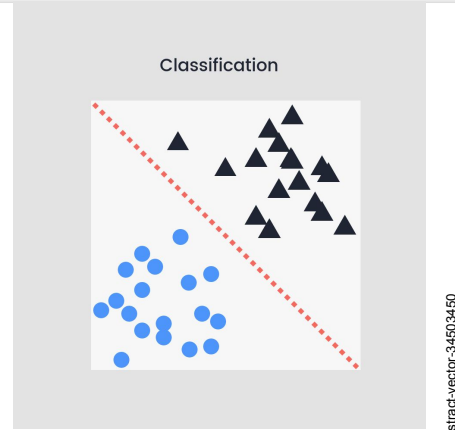
DECISION BOUNDARY

Pink

Many things that are
pink will now be
classified as purple



Purple



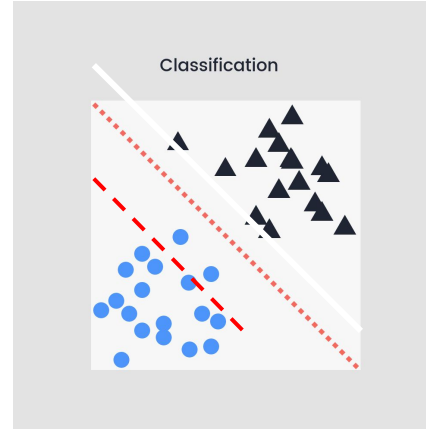
DECISION BOUNDARY

Pink

Many things that are purple will now be classified as pink



Purple

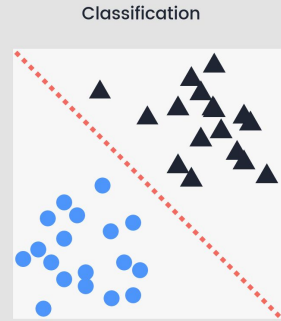


DECISION BOUNDARY

Pink

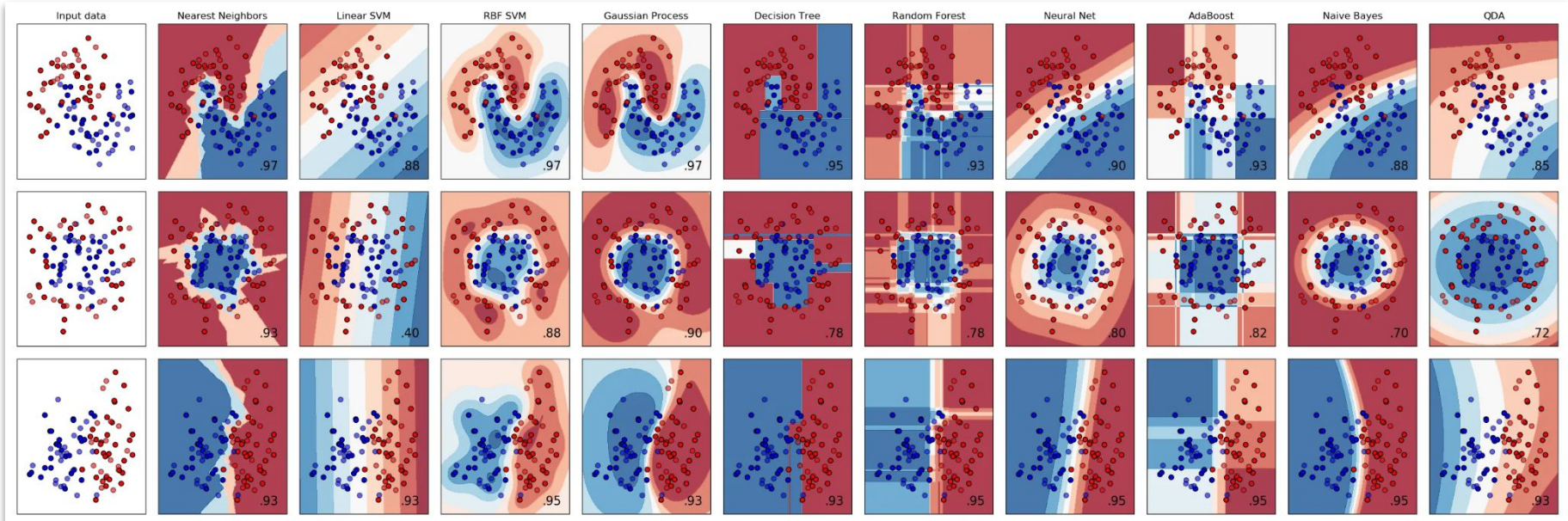
Ideal boundary?

Purple



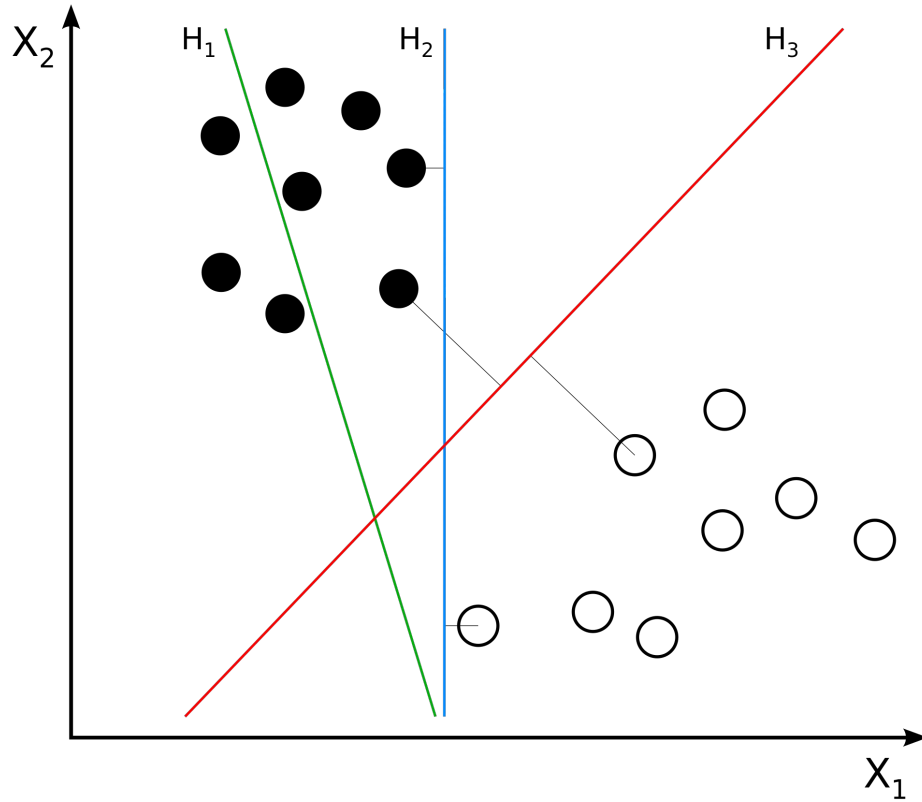
DECISION BOUNDARY

Comparison of the decision boundaries of 10 machine learning models:



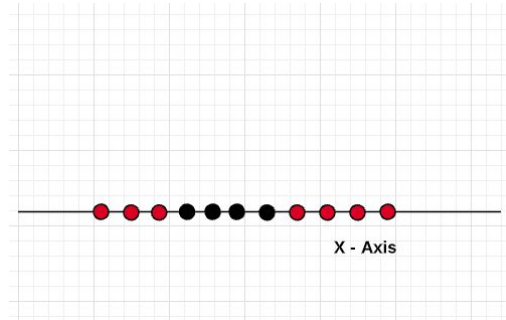
Varoquaux and Müller

CLASSICAL SUPPORT VECTOR CLASSIFIER (SVC)



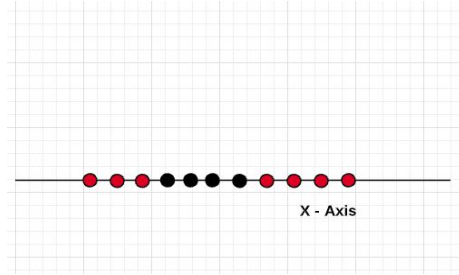
CLASSICAL SUPPORT VECTOR CLASSIFIER (SVC)

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



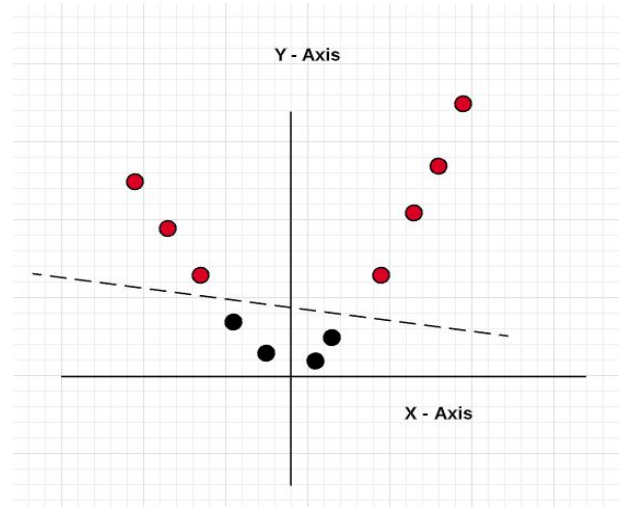
Non-linearly separable data

CLASSICAL SUPPORT VECTOR CLASSIFIER (SVC)



Non-linearly
separable data

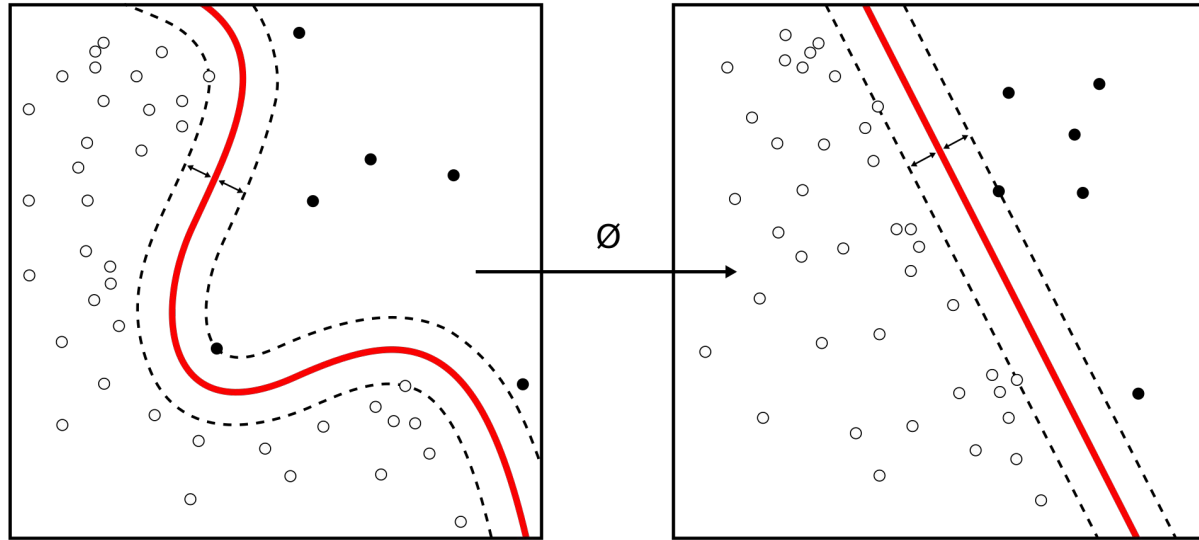
$$\phi \rightarrow$$
$$y = x^2$$



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

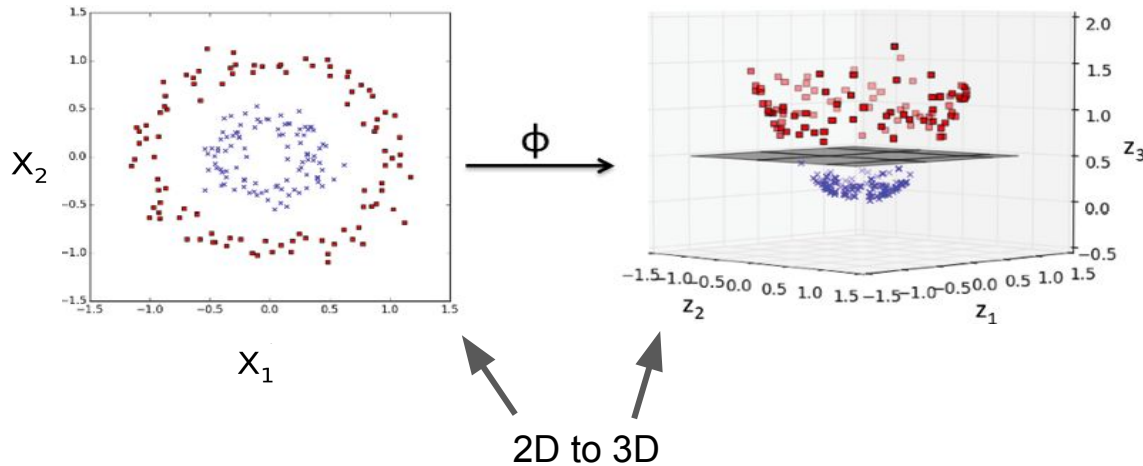
CLASSICAL SUPPORT VECTOR CLASSIFIER (SVC)

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



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



<https://github.com/alexgcsa/incob2023>

- 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



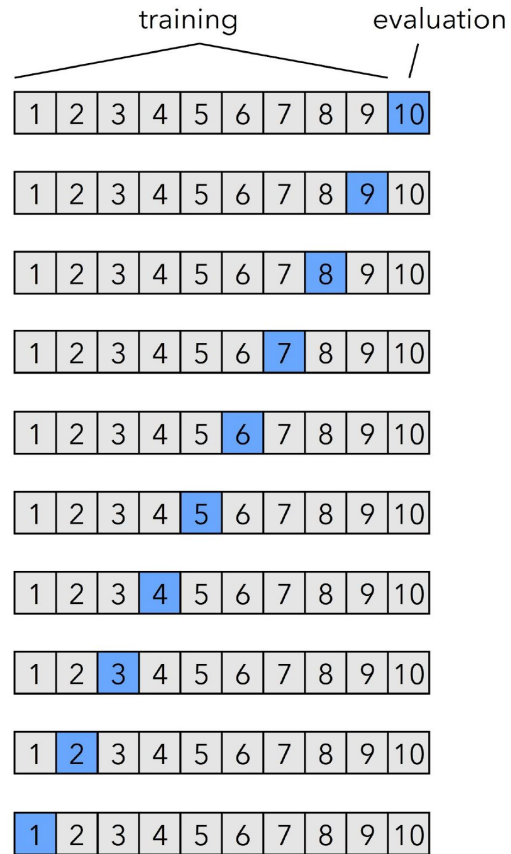
~~QUANTUM~~ MACHINE LEARNING

Alex de Sá
Ashar Malik

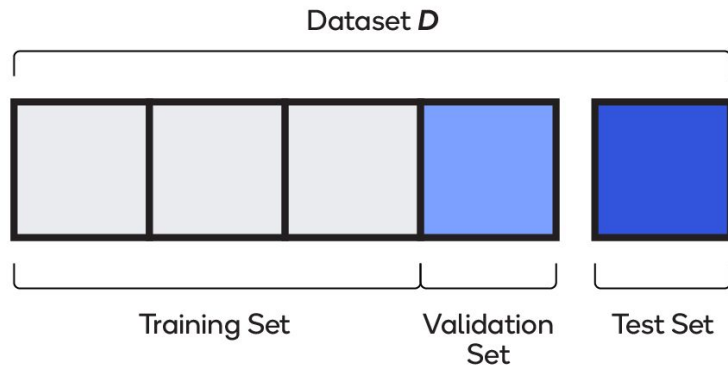
Alex.deSa@baker.edu.au
Ashar.Malik@baker.edu.au

<https://github.com/alexgcsa/incob2023>

K-FOLD CROSS-VALIDATION



Test the model on new data, assessing its generalisation



CLASSIFICATION METRICS

| | | ACTUAL VALUES | |
|------------------|----------|---------------|----------|
| | | Positive | Negative |
| PREDICTED VALUES | Positive | TP | FP |
| | Negative | FN | TN |

The predicted value is positive and its positive

Type I error :
The predicted value is positive but it False

Type II error :
The predicted value is negative but its positive

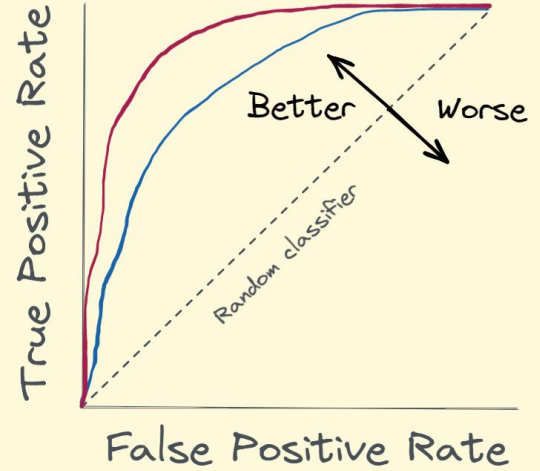
The predicted value is Negative and its Negative

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

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

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

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



[Towards Data Science, 2023](#)

[Medium, 2020](#)