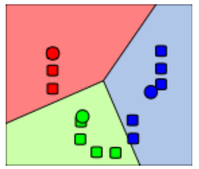
From (chemical) data to information

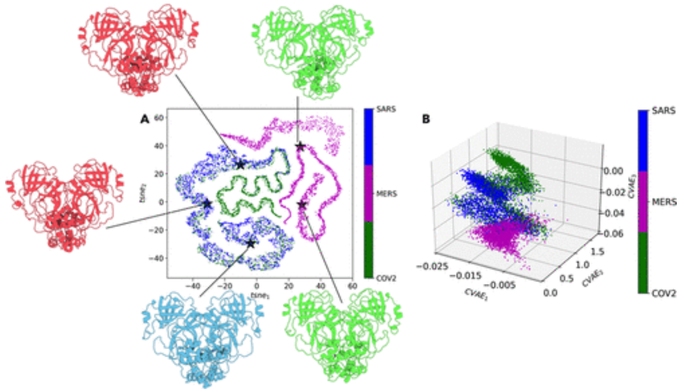
Introduction to Machine learning for Chemistry Lecture 1

Dr Antonia Mey

antonia.mey@ed.ac.uk

Room 214 JBB





What is machine learning?

Artificial intelligence

Design an intelligent agent that perceives its environment and makes decisions to maximise chances of achieving its goal.

Machine learning

Gives computers the ability to learn without specifically being programmed (Arthur Samuel 1959)

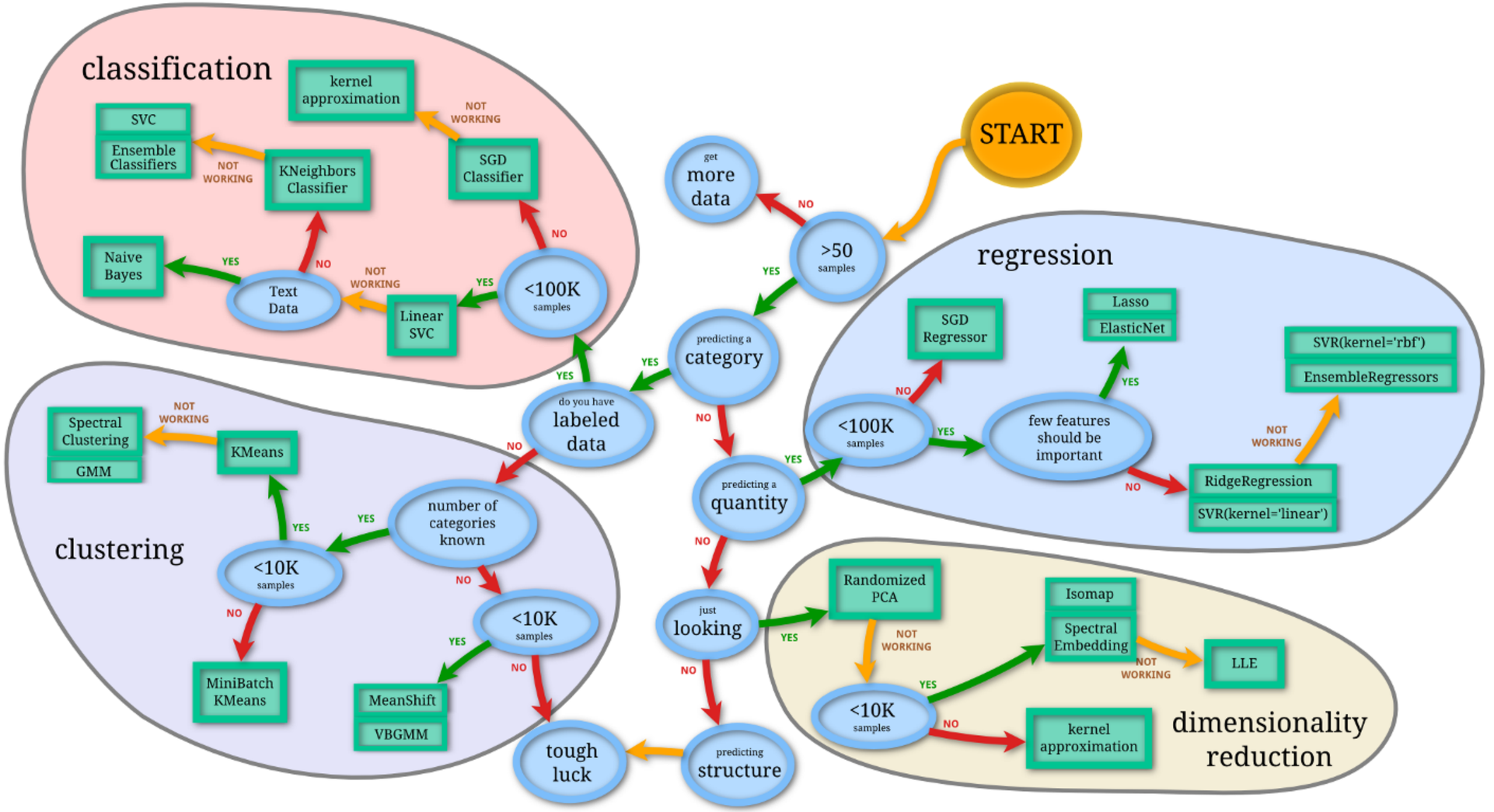
Supervised learning

Unsupervised learning

reinforcement learning

https://medium.com/machine-learning-for-humans/why-machine-learning-matters-6164faf1df12

The Data Mining World



From **scikit-learn.org**

Topics overview

Lecture 1

Lecture 2

Learning outcomes:

Topics overview

Lecture 1

- What is machine learning?

Lecture 2

Learning outcomes:

Topics overview

Lecture 1

- What is machine learning?
- Examples of machine learning (in Chemistry)

Lecture 2

Learning outcomes:

Topics overview

Lecture 1

- What is machine learning?
- Examples of machine learning (in Chemistry)
- Introduction to **unsupervised learning**:
  - Clustering (k-means and others)
  - How does actual input data look like?

Lecture 2

Learning outcomes:

Topics overview

Lecture 1

- What is machine learning?
- Examples of machine learning (in Chemistry)
- Introduction to **unsupervised learning**:
  - Clustering (k-means and others)
  - How does actual input data look like?
- Molecular fingerprints and nomenclature

Lecture 2

Learning outcomes:

Topics overview

Lecture 1

- What is machine learning?
- Examples of machine learning (in Chemistry)
- Introduction to **unsupervised learning**:
  - Clustering (k-means and others)
  - How does actual input data look like?
- Molecular fingerprints and nomenclature
- Introduction to **supervised learning:**
  - What is a classification problem?

Lecture 2

Learning outcomes:

Topics overview

Lecture 1

- What is machine learning?
- Examples of machine learning (in Chemistry)
- Introduction to **unsupervised learning**:
  - Clustering (k-means and others)
  - How does actual input data look like?
- Molecular fingerprints and nomenclature
- Introduction to **supervised learning:**
  - What is a classification problem?

Lecture 2

- Unsupervised learning continued:
  - Dimensionality reduction (PCA)
  - t-SNE

Learning outcomes:

Topics overview

Lecture 1

- What is machine learning?
- Examples of machine learning (in Chemistry)
- Introduction to **unsupervised learning**:
  - Clustering (k-means and others)
  - How does actual input data look like?
- Molecular fingerprints and nomenclature
- Introduction to **supervised learning:**
  - What is a classification problem?

Lecture 2

- Unsupervised learning continued:
  - Dimensionality reduction (PCA)
  - t-SNE
- Regressions

Learning outcomes:

Topics overview

Lecture 1

- What is machine learning?
- Examples of machine learning (in Chemistry)
- Introduction to **unsupervised learning**:
  - Clustering (k-means and others)
  - How does actual input data look like?
- Molecular fingerprints and nomenclature
- Introduction to **supervised learning:**
  - What is a classification problem?

Lecture 2

- Unsupervised learning continued:
  - Dimensionality reduction (PCA)
  - t-SNE
- Regressions
- Classifications in practice:
  - Random Forests
  - Multilayer perceptrons

Learning outcomes:

Topics overview

Lecture 1

- What is machine learning?
- Examples of machine learning (in Chemistry)
- Introduction to **unsupervised learning**:
  - Clustering (k-means and others)
  - How does actual input data look like?
- Molecular fingerprints and nomenclature
- Introduction to **supervised learning:**
  - What is a classification problem?

Lecture 2

- Unsupervised learning continued:
  - Dimensionality reduction (PCA)
  - t-SNE
- Regressions
- Classifications in practice:
  - Random Forests
  - Multilayer perceptrons

Learning outcomes:

- Understand the main pillars of machine learning

Topics overview

Lecture 1

- What is machine learning?
- Examples of machine learning (in Chemistry)
- Introduction to **unsupervised learning**:
  - Clustering (k-means and others)
  - How does actual input data look like?
- Molecular fingerprints and nomenclature
- Introduction to **supervised learning:**
  - What is a classification problem?

Lecture 2

- Unsupervised learning continued:
  - Dimensionality reduction (PCA)
  - t-SNE
- Regressions
- Classifications in practice:
  - Random Forests
  - Multilayer perceptrons

Learning outcomes:

- Understand the main pillars of machine learning
- Know about different clustering techniques as part of unsupervised learning

Topics overview

Lecture 1

- What is machine learning?
- Examples of machine learning (in Chemistry)
- Introduction to **unsupervised learning**:
  - Clustering (k-means and others)
  - How does actual input data look like?
- Molecular fingerprints and nomenclature
- Introduction to **supervised learning:**
  - What is a classification problem?

Lecture 2

- Unsupervised learning continued:
  - Dimensionality reduction (PCA)
  - t-SNE
- Regressions
- Classifications in practice:
  - Random Forests
  - Multilayer perceptrons

Learning outcomes:

- Understand the main pillars of machine learning
- Know about different clustering techniques as part of unsupervised learning
- Be able to use common nomenclature used in machine learning

Topics overview

Lecture 1

- What is machine learning?
- Examples of machine learning (in Chemistry)
- Introduction to **unsupervised learning**:
  - Clustering (k-means and others)
  - How does actual input data look like?
- Molecular fingerprints and nomenclature
- Introduction to **supervised learning:**
  - What is a classification problem?

Lecture 2

- Unsupervised learning continued:
  - Dimensionality reduction (PCA)
  - t-SNE
- Regressions
- Classifications in practice:
  - Random Forests
  - Multilayer perceptrons

Learning outcomes:

- Understand the main pillars of machine learning
- Know about different clustering techniques as part of unsupervised learning
- Be able to use common nomenclature used in machine learning
- Use PCA to reduce the dimensions of your data set

Topics overview

Lecture 1

- What is machine learning?
- Examples of machine learning (in Chemistry)
- Introduction to **unsupervised learning**:
  - Clustering (k-means and others)
  - How does actual input data look like?
- Molecular fingerprints and nomenclature
- Introduction to **supervised learning:**
  - What is a classification problem?

Lecture 2

- Unsupervised learning continued:
  - Dimensionality reduction (PCA)
  - t-SNE
- Regressions
- Classifications in practice:
  - Random Forests
  - Multilayer perceptrons

Learning outcomes:

- Understand the main pillars of machine learning
- Know about different clustering techniques as part of unsupervised learning
- Be able to use common nomenclature used in machine learning
- Use PCA to reduce the dimensions of your data set
- Understand how a regression problem can be cast as a machine learning problem

Topics overview

Lecture 1

- What is machine learning?
- Examples of machine learning (in Chemistry)
- Introduction to **unsupervised learning**:
  - Clustering (k-means and others)
  - How does actual input data look like?
- Molecular fingerprints and nomenclature
- Introduction to **supervised learning:**
  - What is a classification problem?

Lecture 2

- Unsupervised learning continued:
  - Dimensionality reduction (PCA)
  - t-SNE
- Regressions
- Classifications in practice:
  - Random Forests
  - Multilayer perceptrons

Learning outcomes:

- Understand the main pillars of machine learning
- Know about different clustering techniques as part of unsupervised learning
- Be able to use common nomenclature used in machine learning
- Use PCA to reduce the dimensions of your data set
- Understand how a regression problem can be cast as a machine learning problem
- Be aware of how random forests and multilayer perceptrons can be used in a classification problem

Topics overview

Lecture 1

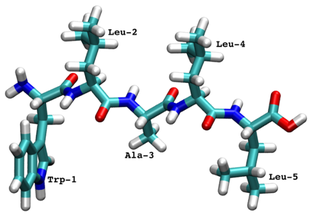
- What is machine learning?
- Examples of machine learning (in Chemistry)
- Introduction to **unsupervised learning**:
  - Clustering (k-means and others)
  - How does actual input data look like?
- Molecular fingerprints and nomenclature
- Introduction to **supervised learning:**
  - What is a classification problem?

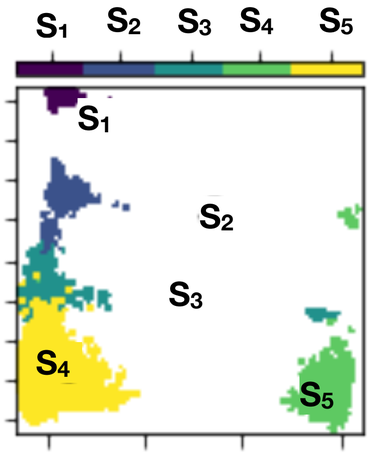
Lecture 2

- Unsupervised learning continued:
  - Dimensionality reduction (PCA)
  - t-SNE
- Regressions
- Classifications in practice:
  - Random Forests
  - Multilayer perceptrons

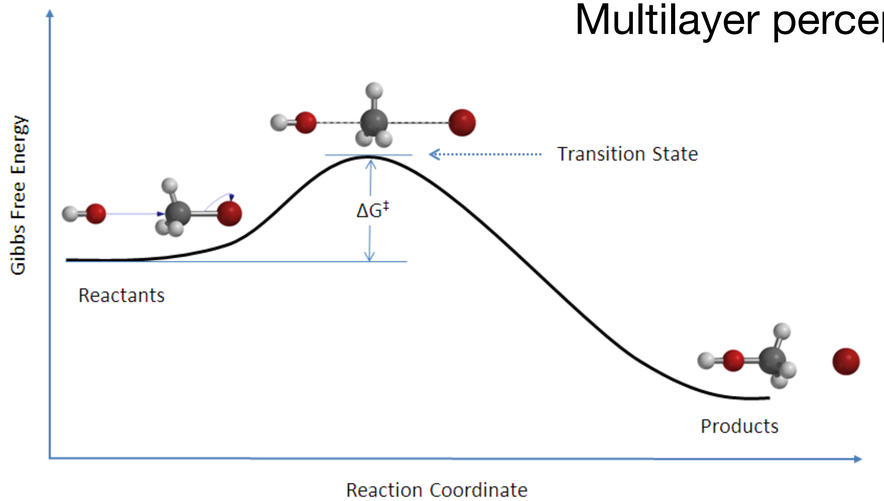
Planned Workshop topics for Semester 2

Dimensionality Reduction
Clustering





Classification problems with:
Random Forests
Multilayer perceptrons

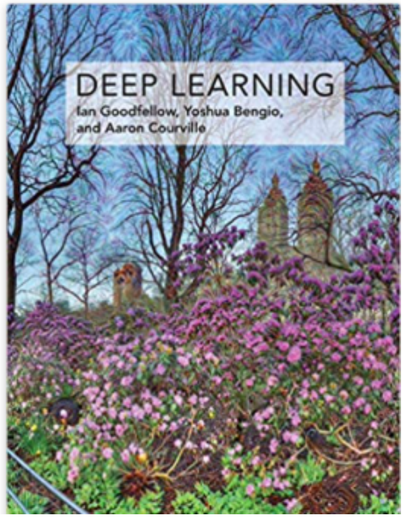


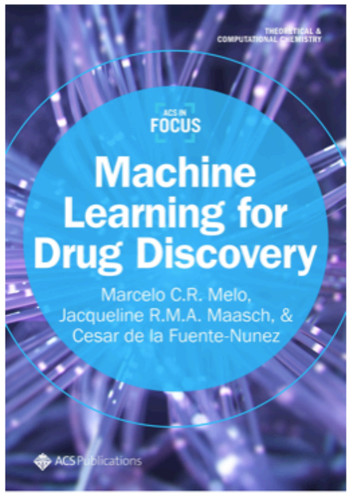
Further resources

Best practices in machine learning for chemistry

Statistical tools based on machine learning are becoming integrated into chemistry research workflows. We discuss the elements necessary to train reliable, repeatable and reproducible models, and recommend a set of guidelines for machine learning reports.

Nongnuch Artrith, Keith T. Butler, François-Xavier Coudert, Seungwu Han, Olexandr Isayev, Anubhav Jain and Aron Walsh





Read Now

Machine Learning for Drug Discovery

Author(s): Marcelo C.R. Melo, Jacqueline R. M. A. Maasch, Cesar de la Fuente Nunez
Publication Date: March 11, 2022
Copyright © 2022 American Chemical Society

Subscribed

Cite This: *Machine Learning for Drug Discovery*, American Chemical Society, 2022. DOI: 10.1021/acsinfocus.7e5017

eISBN:	9780841299238
DOI:	10.1021/acsinfocus.7e5017
Read Time:	eight hours
Collection:	1
Publisher:	American Chemical Society

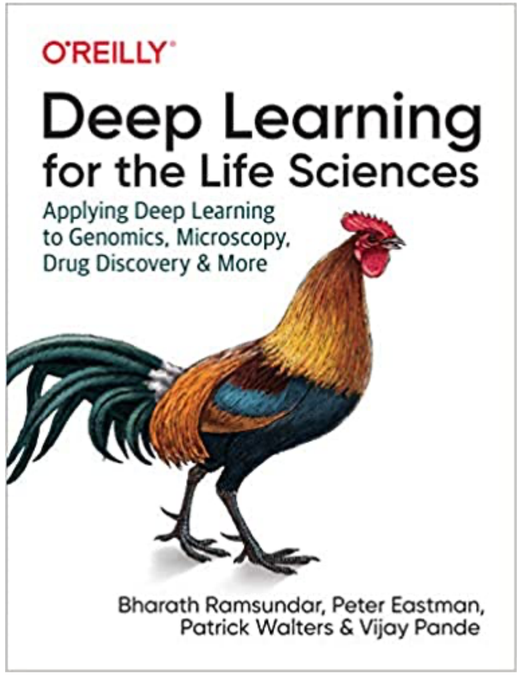
Get it on Google Play

Artificial Intelligence in Drug Discovery

Editor: Nathan Brown

Machine Learning in Chemistry: The Impact of Artificial Intelligence

Editor: Hugh M Cartwright

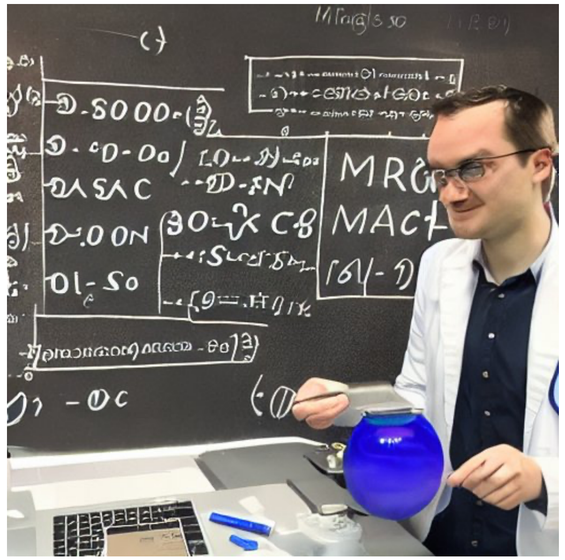


Examples of Machine Learning (in Chemistry)

Dall-E 2

Examples of Machine Learning (in Chemistry)

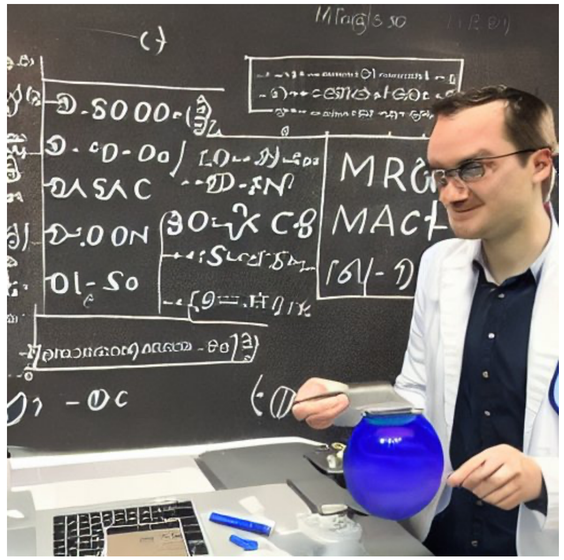
Dall-E 2



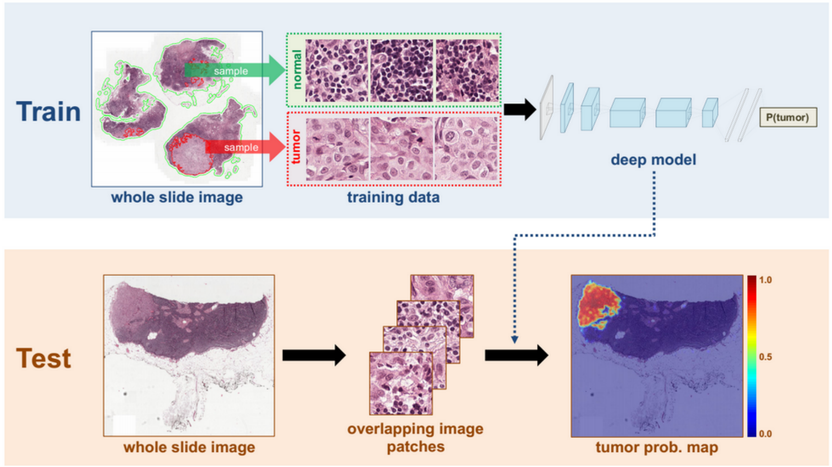
MSc Student in Chemistry learning about ML
Stable Diffusion

Examples of Machine Learning (in Chemistry)

Dall-E 2



MSc Student in Chemistry learning about ML
Stable Diffusion



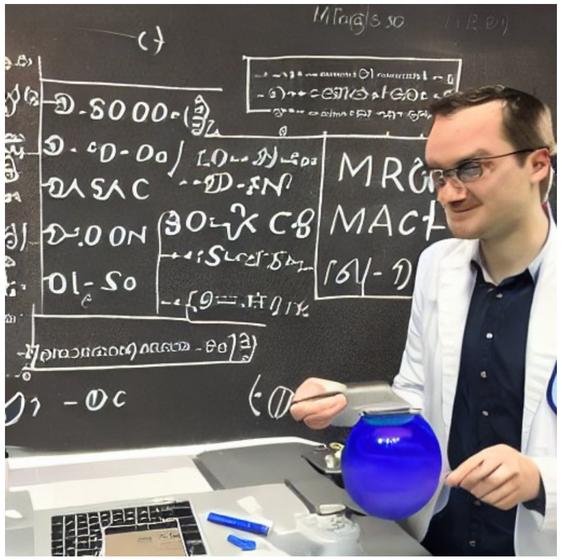
Identifying cancerous cells

Examples of Machine Learning (in Chemistry)

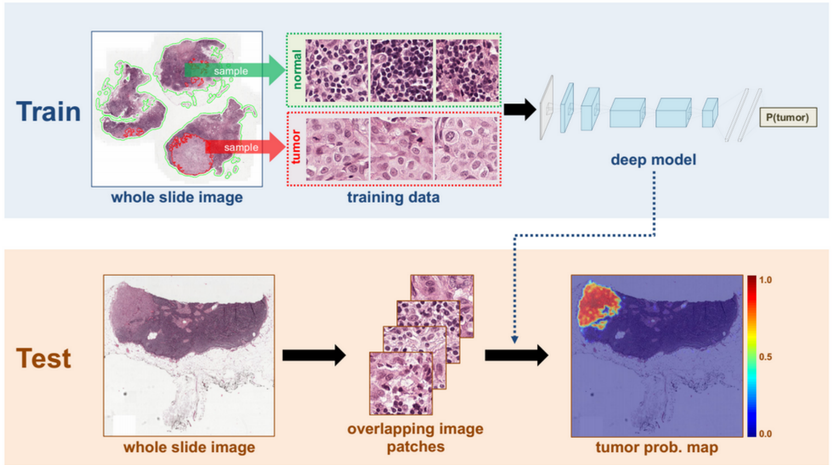
Dall-E 2

Alpha - Go

Alexa



MSc Student in Chemistry learning about ML
Stable Diffusion



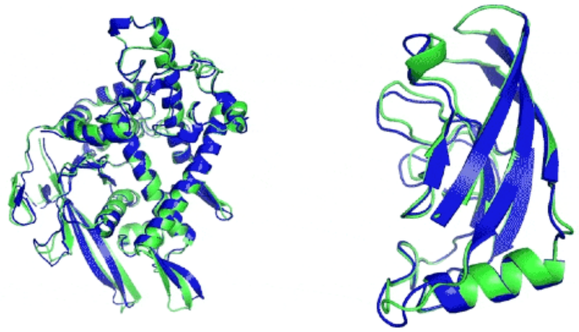
Identifying cancerous cells

Examples of Machine Learning (in Chemistry)

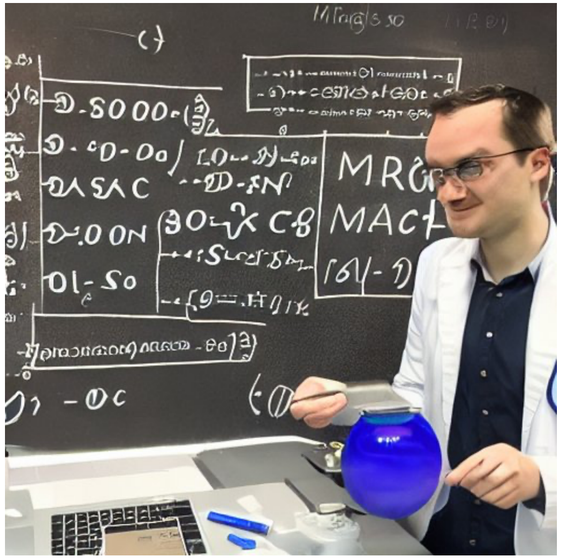
Dall-E 2

Alpha - Go

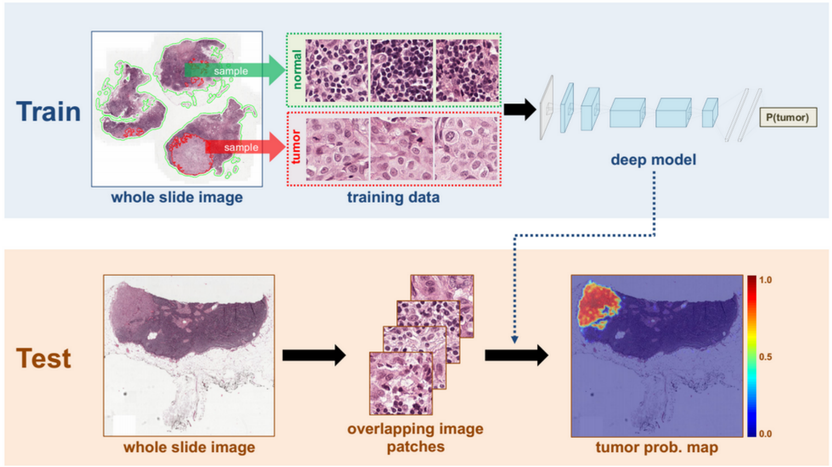
AlphaFold 2



Alexa



MSc Student in Chemistry learning about ML
Stable Diffusion



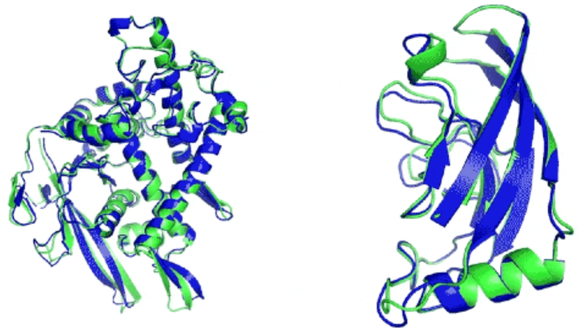
Identifying cancerous cells

Examples of Machine Learning (in Chemistry)

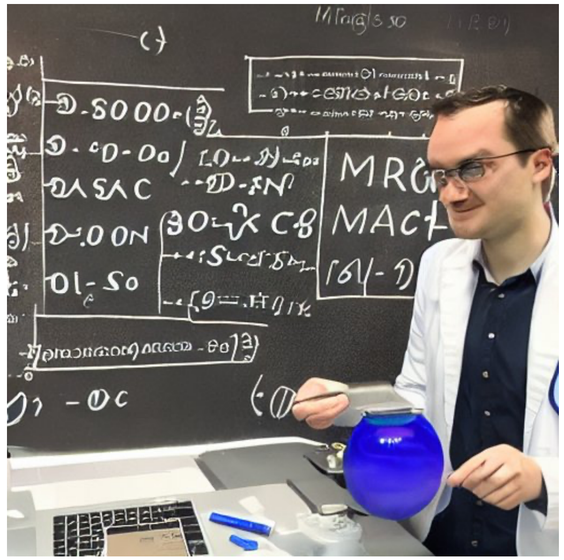
Dall-E 2

Alpha - Go

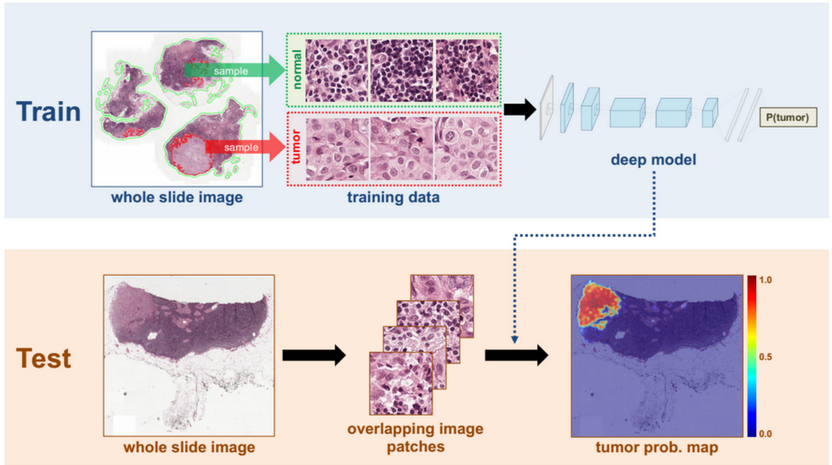
AlphaFold 2



Alexa

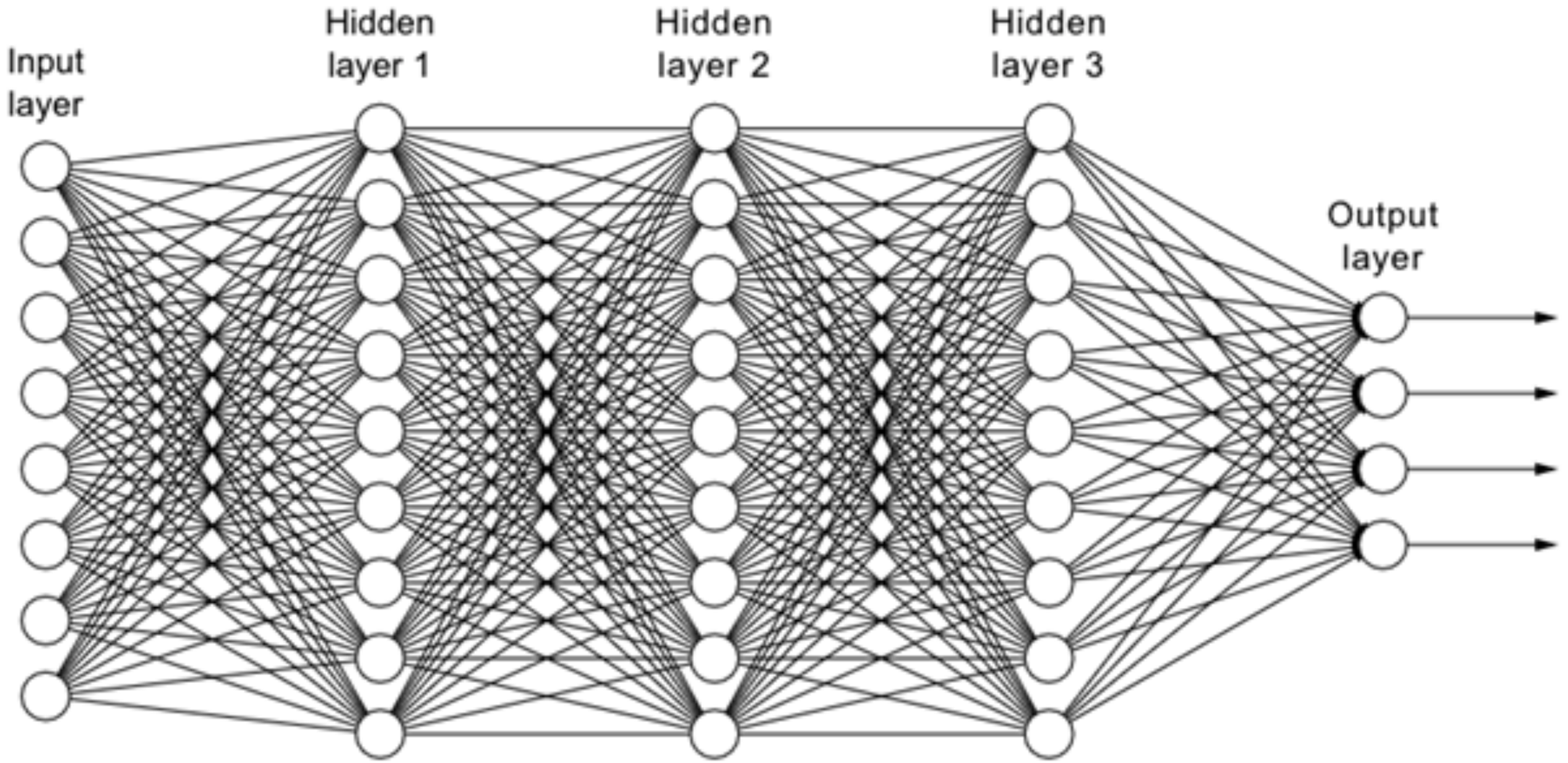


MSc Student in Chemistry learning about ML
Stable Diffusion

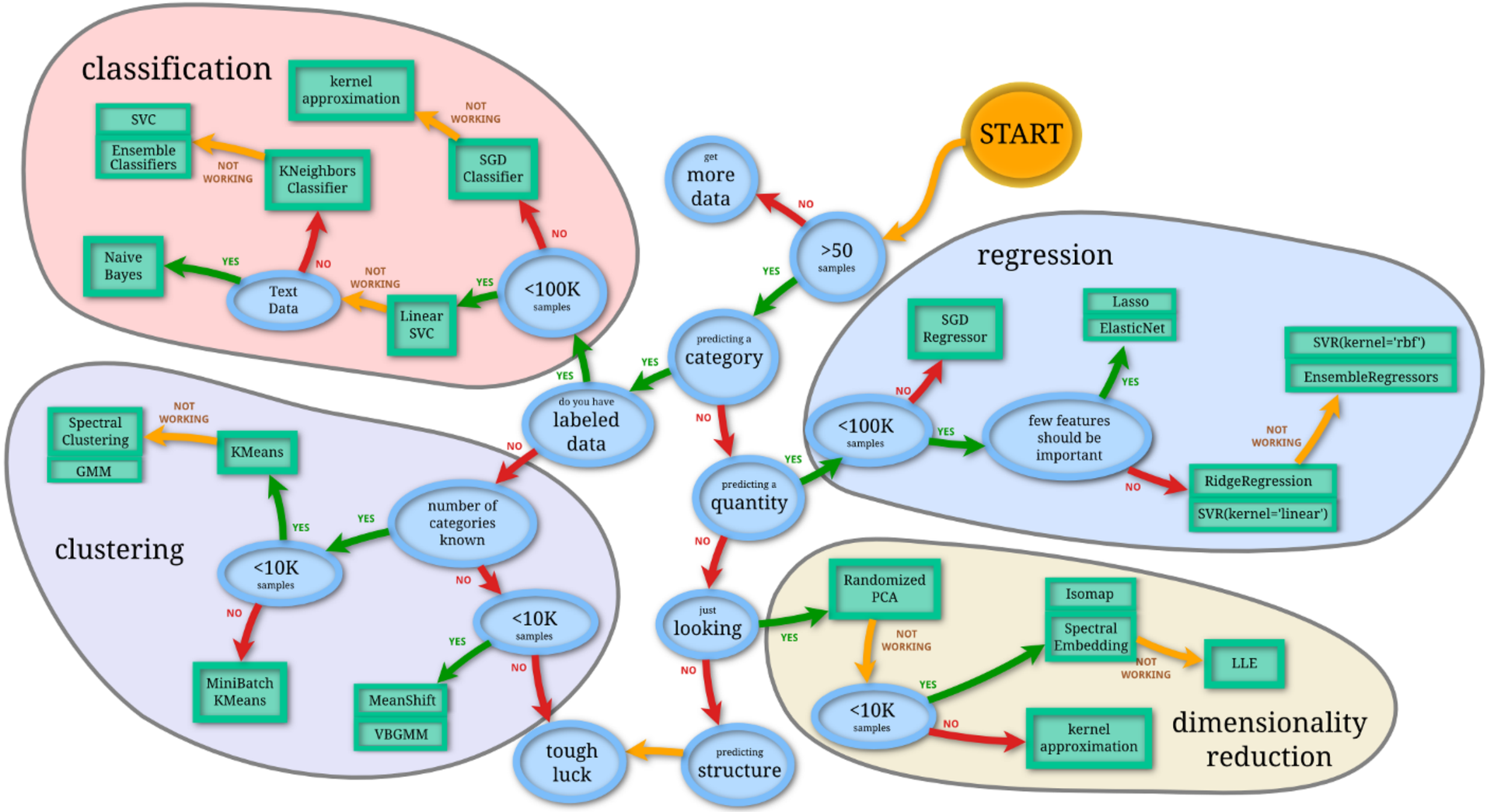


Identifying cancerous cells

What you might think of as machine learning

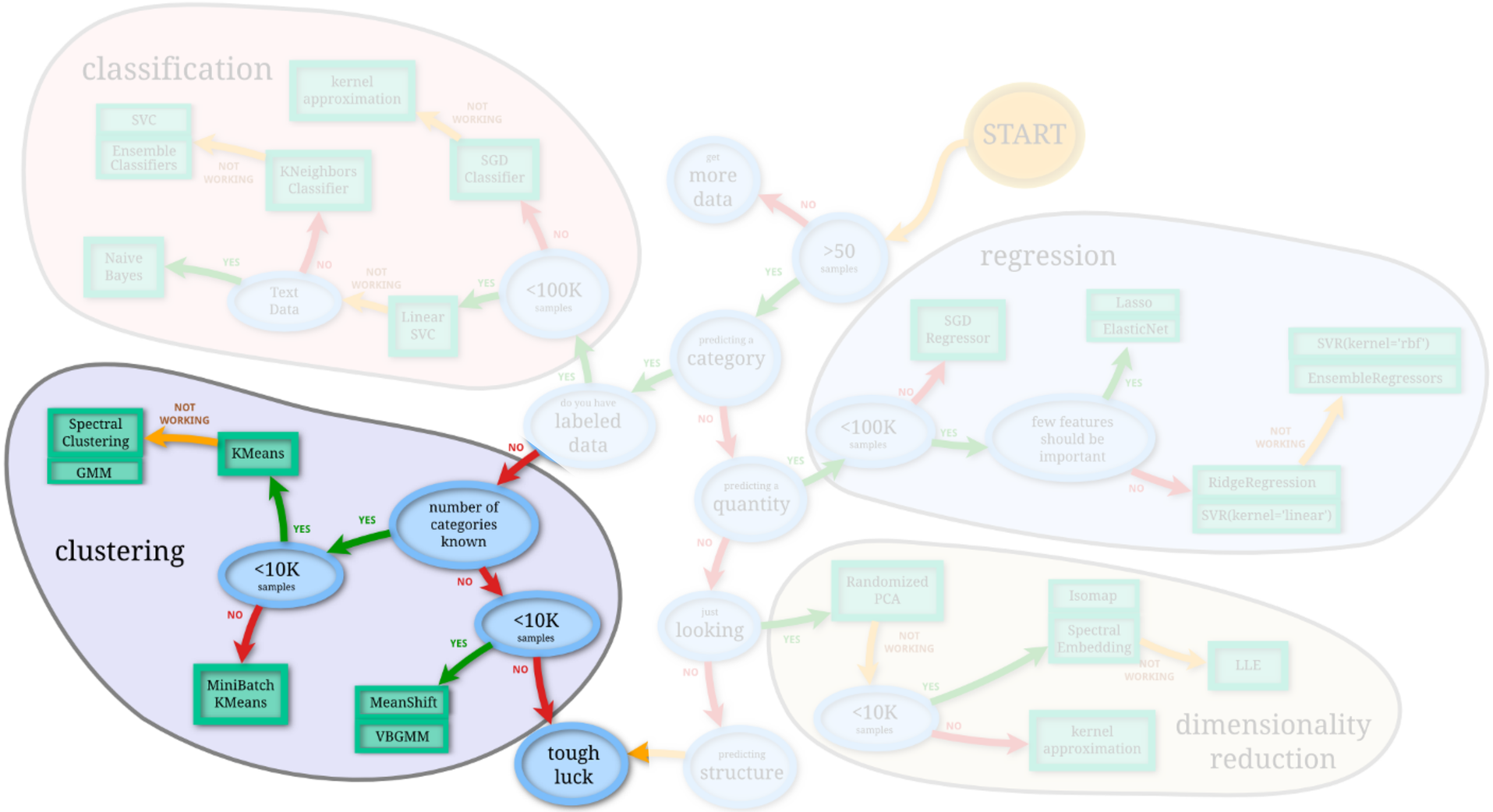


The Data Mining World



From **scikit-learn.org**

The Data Mining World



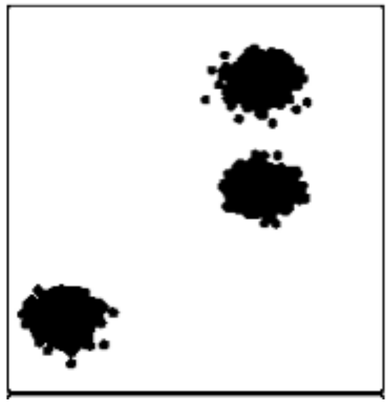
From **scikit-learn.org**

A large ecosystem of Python-based tools for ML

And many more….

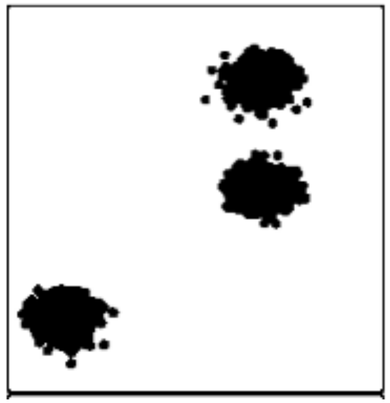
Clustering is an unsupervised learning technique

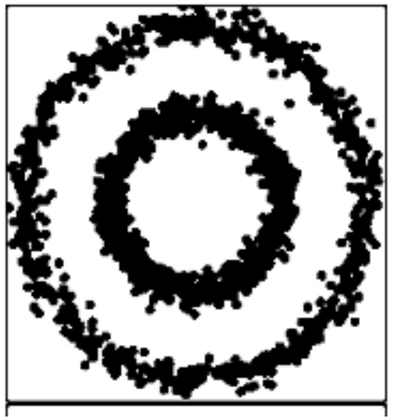
How many clusters are there?



Clustering is an unsupervised learning technique

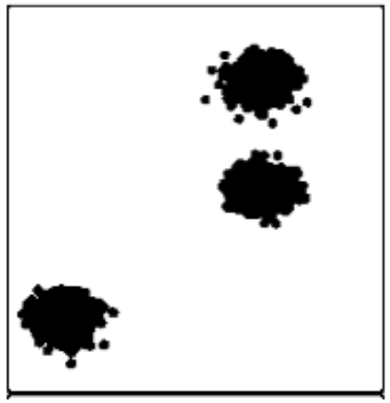
How many clusters are there?

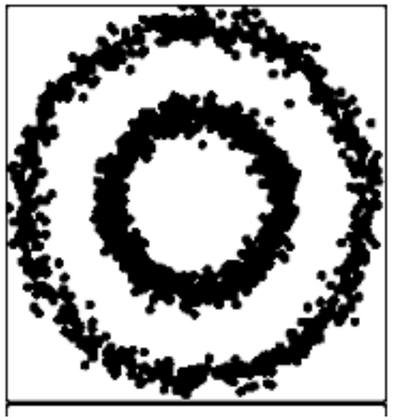




Clustering is an unsupervised learning technique

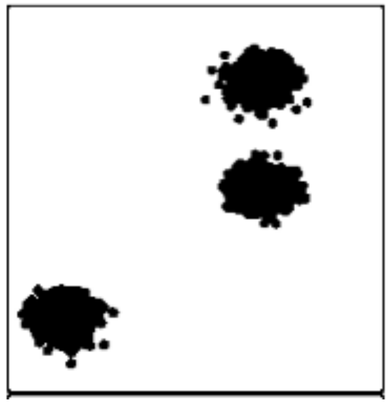
How many clusters are there?

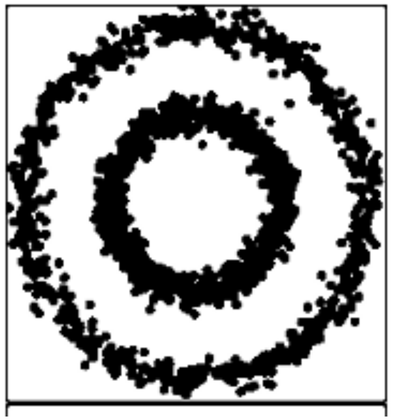


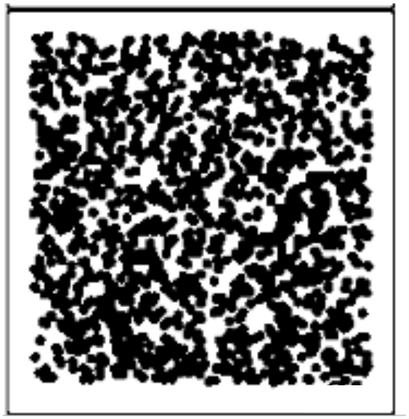


Clustering is an unsupervised learning technique

How many clusters are there?

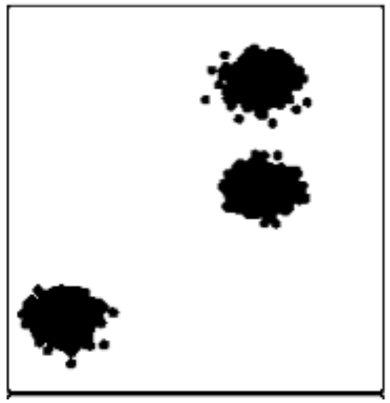
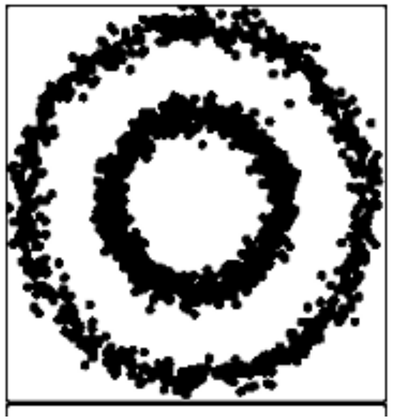


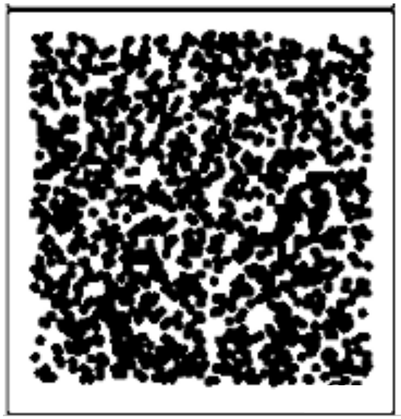




Clustering is an unsupervised learning technique

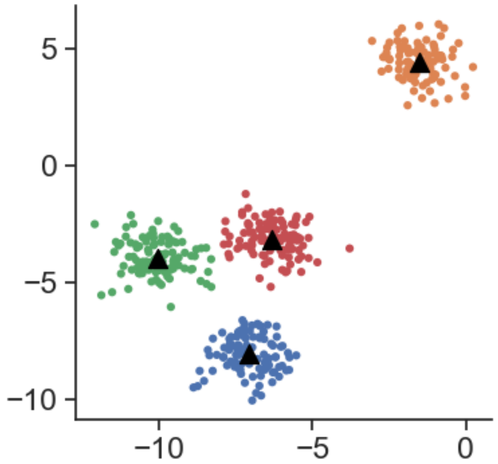
How many clusters are there?



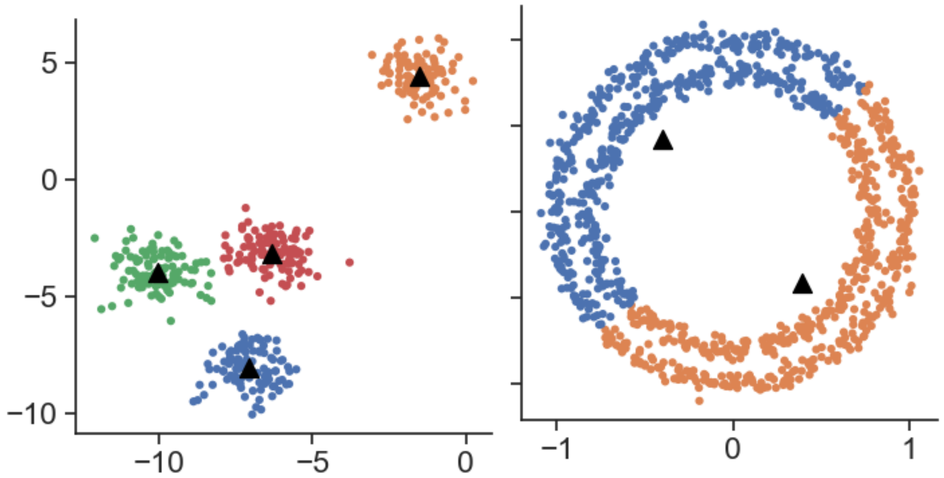

Can the computer do better than the human?

This is how the k-means algorithm performs



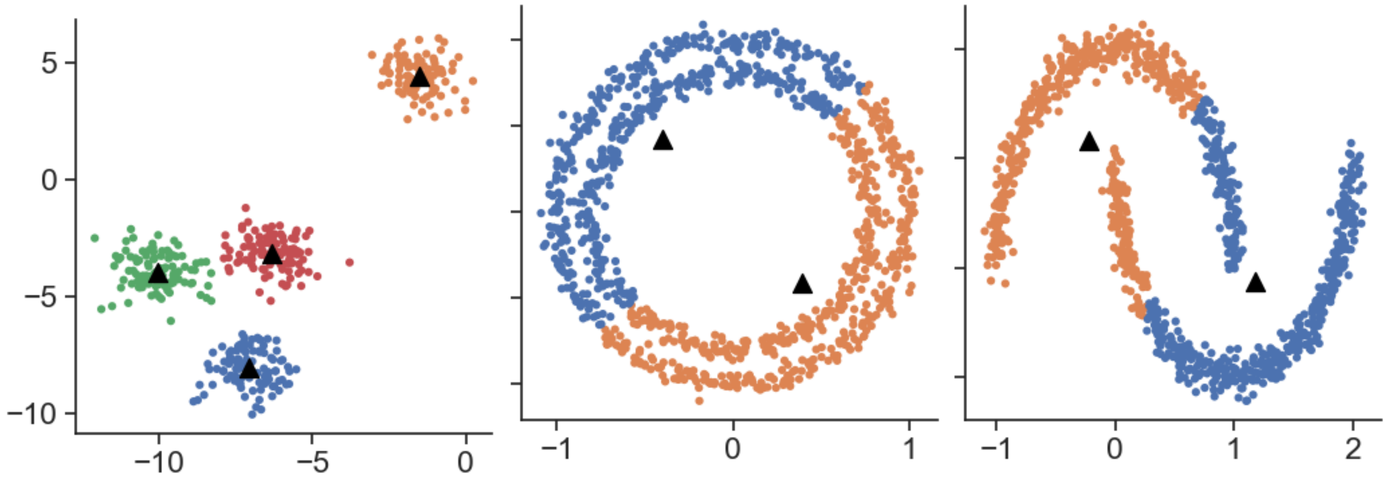
Can the computer do better than the human?

This is how the k-means algorithm performs



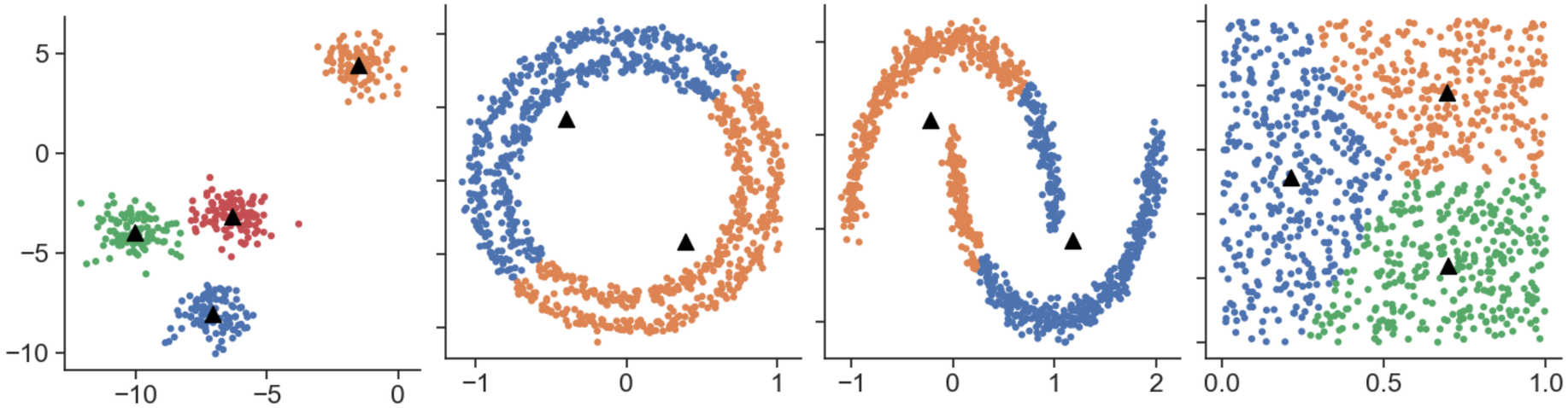
Can the computer do better than the human?

This is how the k-means algorithm performs



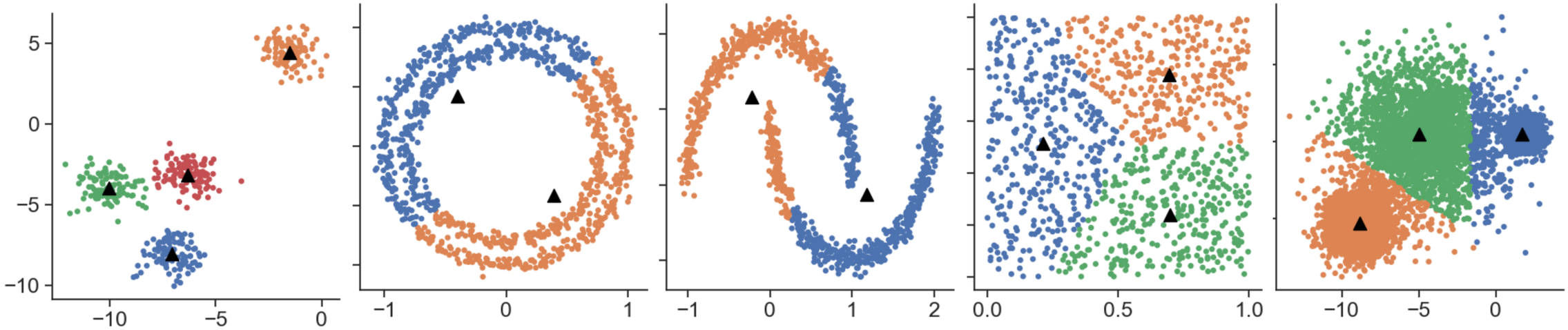
Can the computer do better than the human?

This is how the k-means algorithm performs



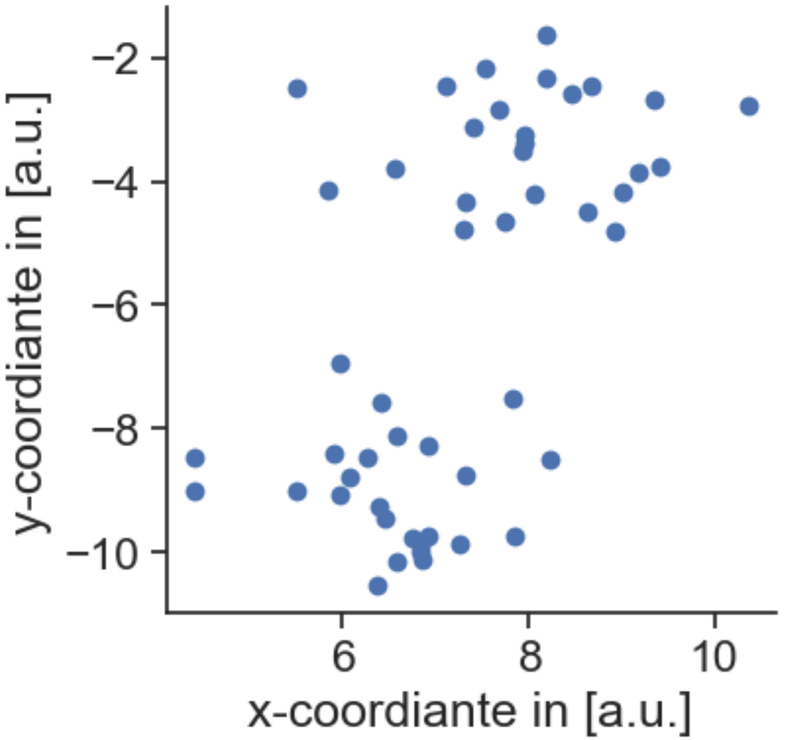
Can the computer do better than the human?

This is how the k-means algorithm performs



How does k-means work?

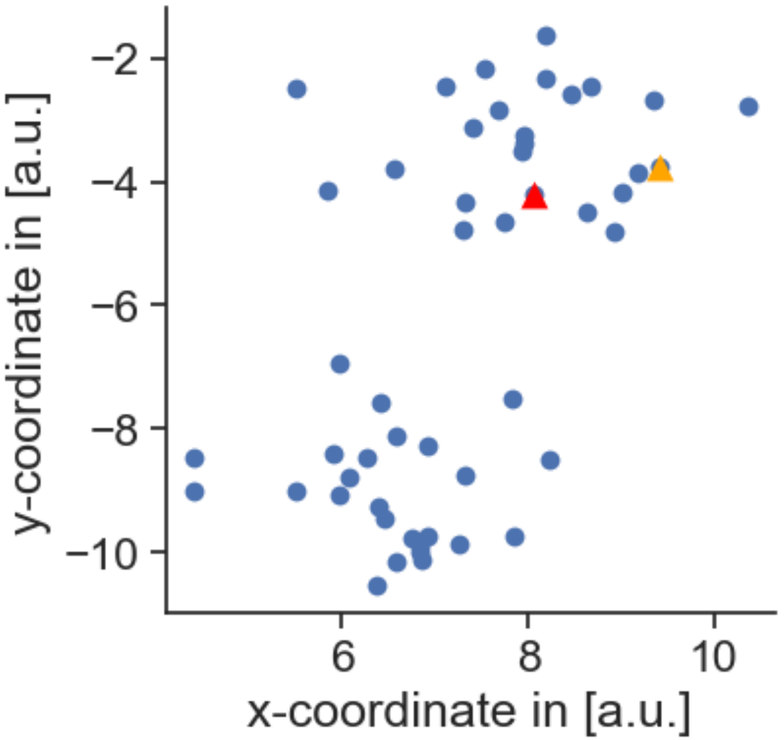
Input: K, set of points $x_1 \ldots x_n$ this can be in N-dimensions



How does k-means work?

Input: K, set of points $x_1 \ldots x_n$ this can be in N-dimensions

Place centroids, $c_1 \ldots, c_n$ at random locations



How does k-means work?

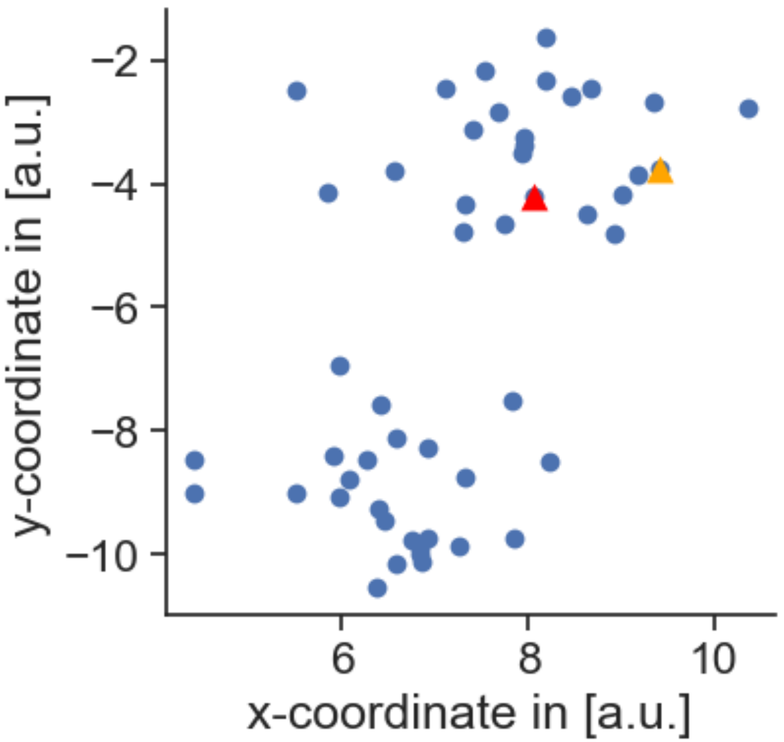
Input: K, set of points $x_1 \ldots x_n$ this can be in N-dimensions

Place centroids, $c_1 \ldots, c_n$ at random locations

Repeat until convergence:

- For each point x_i:
  - Find nearest centroid c_j
  - Assign the point x_i to cluster j

$$\arg\min_j D(x_i, c_j)$$



How does k-means work?

Input: K, set of points $x_1 \ldots x_n$ this can be in N-dimensions

Place centroids, $c_1 \ldots, c_n$ at random locations

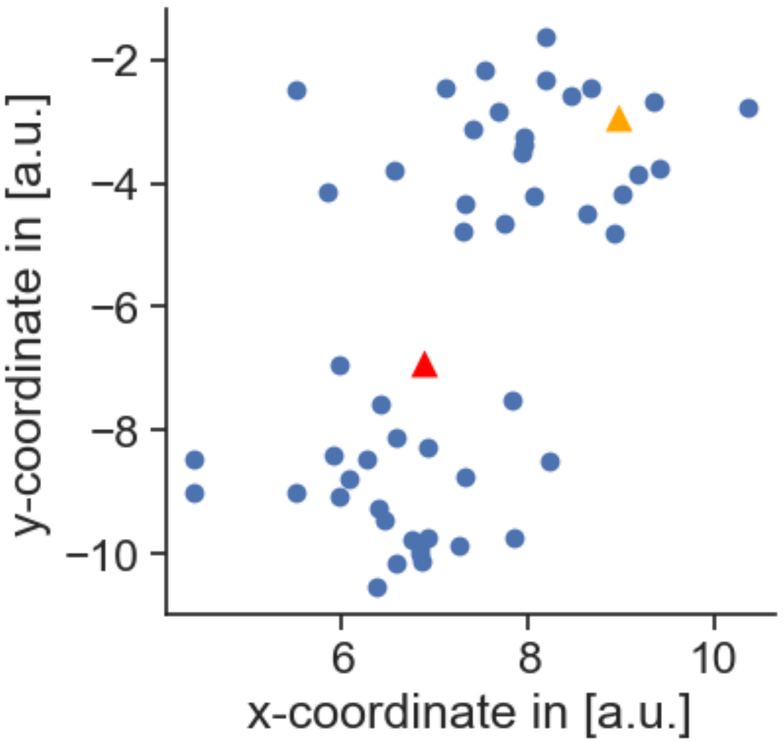
Repeat until convergence:

- For each point x_i:
  - Find nearest centroid c_j
  - Assign the point x_i to cluster j

$$\arg\min_j D(x_i, c_j)$$

- Fore each cluster j = 1…K:
  - Compute the centroid mean for all points in one cluster and update the centroid

$$c_j(a) = \frac{1}{n_j} \sum_{x_i \to c_j} x_i(a)$$



How does k-means work?

Input: K, set of points $x_1 \ldots x_n$ this can be in N-dimensions

Place centroids, $c_1 \ldots, c_n$ at random locations

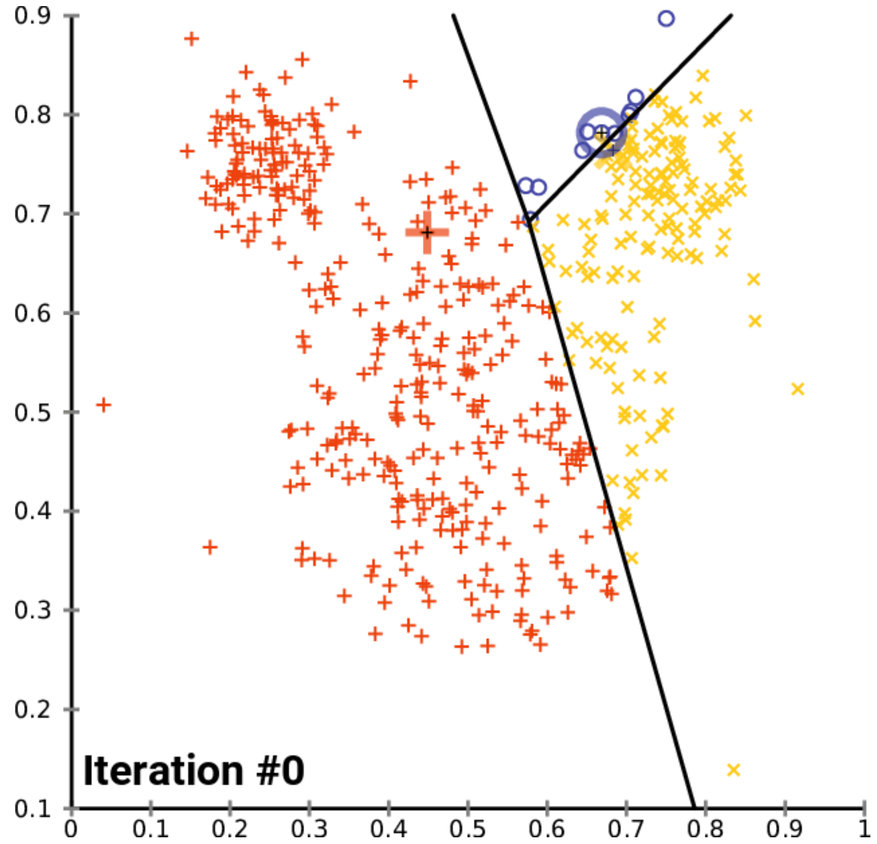
Repeat until convergence:

- For each point x_i:
  - Find nearest centroid c_j
  - Assign the point x_i to cluster j

$$\arg\min_j D(x_i, c_j)$$

- Fore each cluster j = 1…K:
  - Compute the centroid mean for all points in one cluster and update the centroid

$$c_j(a) = \frac{1}{n_j} \sum_{x_i \to c_j} x_i(a)$$



How does k-means work?

Input: K, set of points $x_1 \ldots x_n$ this can be in N-dimensions

Place centroids, $c_1 \ldots, c_n$ at random locations

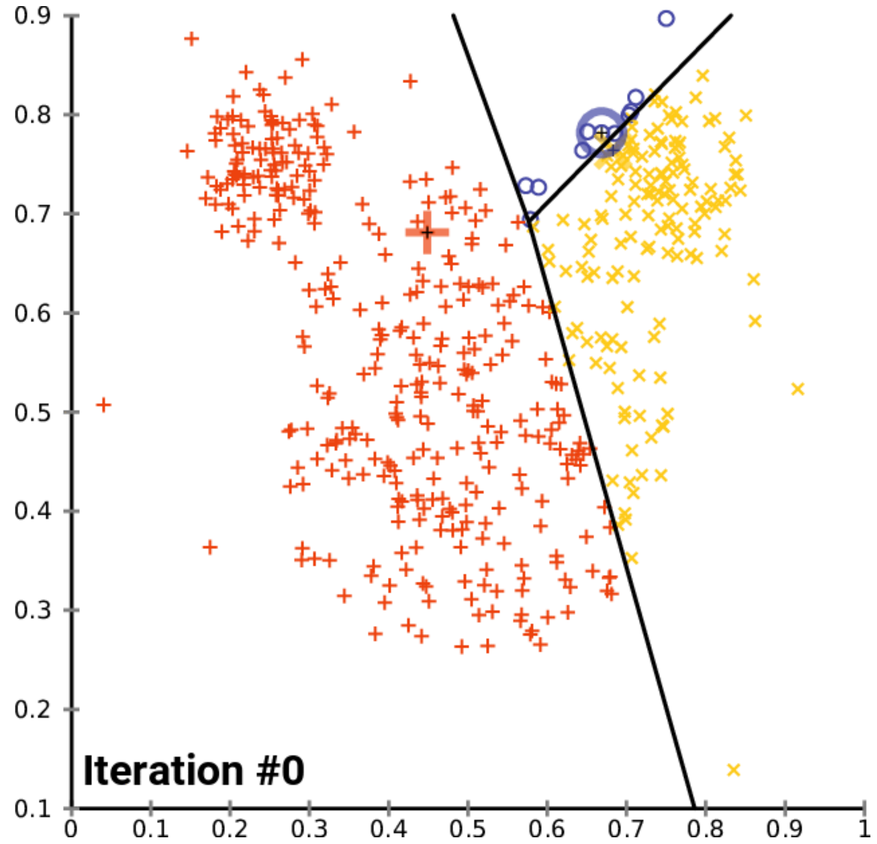
Repeat until convergence:

- For each point x_i:
  - Find nearest centroid c_j
  - Assign the point x_i to cluster j

$$\arg\min_j D(x_i, c_j)$$

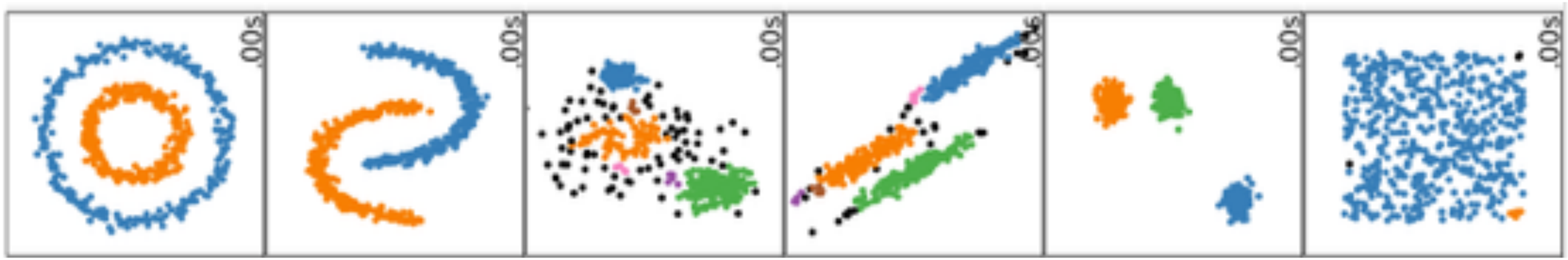
- Fore each cluster j = 1…K:
  - Compute the centroid mean for all points in one cluster and update the centroid

$$c_j(a) = \frac{1}{n_j} \sum_{x_i \to c_j} x_i(a)$$

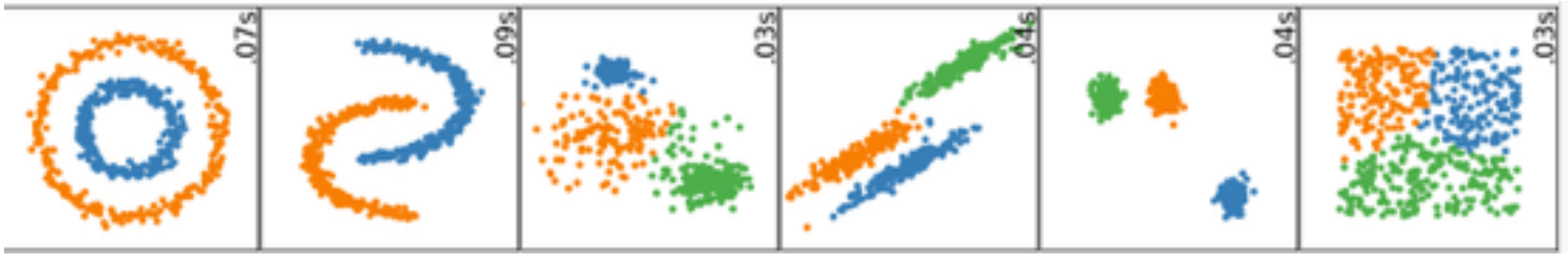


Not all clustering methods perform the same

DBSCAN

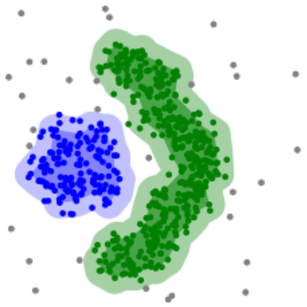


Spectral clustering



Density based clustering and spectral clustering

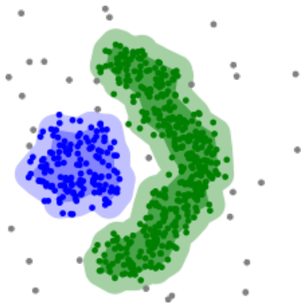
DBSCAN



1. Find the points in the ε (eps) neighbourhood of every point, and identify the core points with more than *minPts* neighbours.
2. Find the connected components of core points on the neighbour graph, ignoring all non-core points.
3. Assign each non-core point to a nearby cluster if the cluster is an ε (eps) neighbour, otherwise assign it to noise.

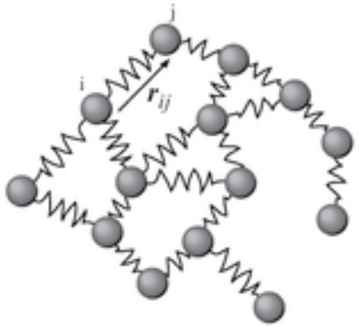
Density based clustering and spectral clustering

DBSCAN



1. Find the points in the ε (eps) neighbourhood of every point, and identify the core points with more than *minPts* neighbours.
2. Find the connected components of core points on the neighbour graph, ignoring all non-core points.
3. Assign each non-core point to a nearby cluster if the cluster is an ε (eps) neighbour, otherwise assign it to noise.

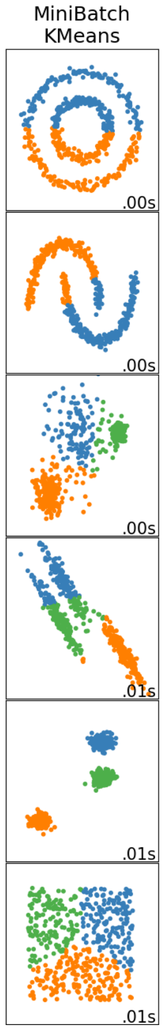
Spectral clustering



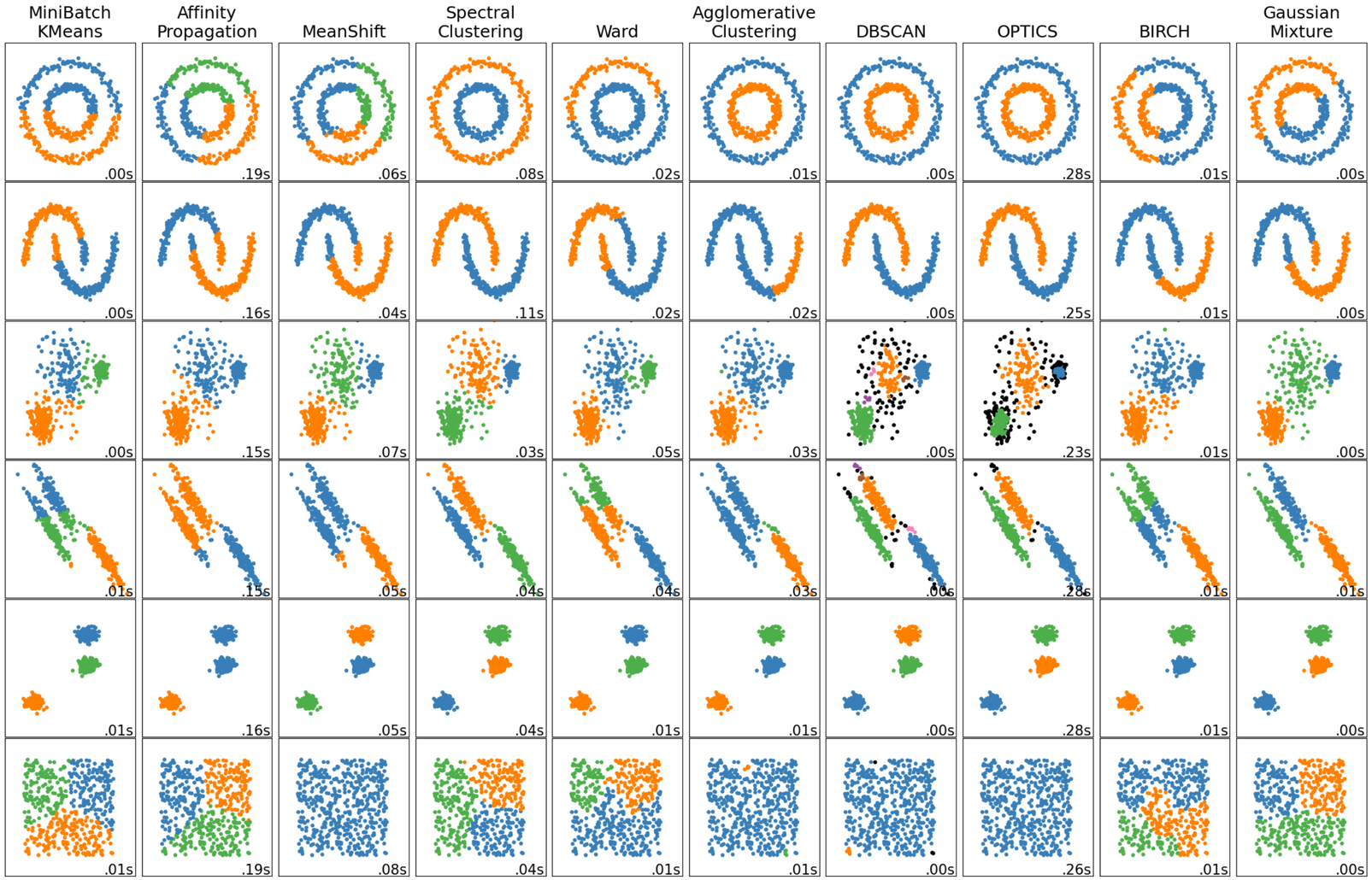
$$\mathbf{L} = \mathbf{D} - \mathbf{A}$$

1. Calculate the Laplacian
2. Calculate the first k eigenvectors
3. Consider the matrix forms by the first k-eigenvectors
4. Cluster the graph nodes based on these features (e.g. k-means)

There are many different clustering algorithms

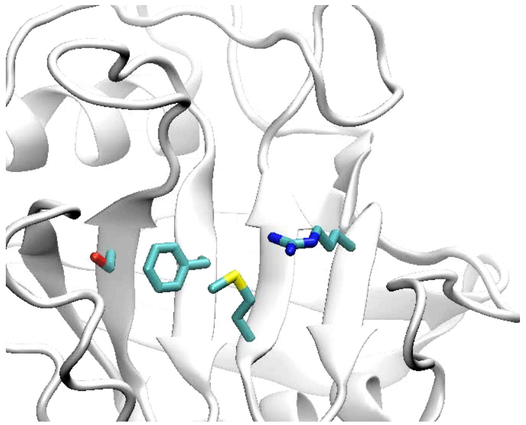


There are many different clustering algorithms



Nature is not 2-dimensional

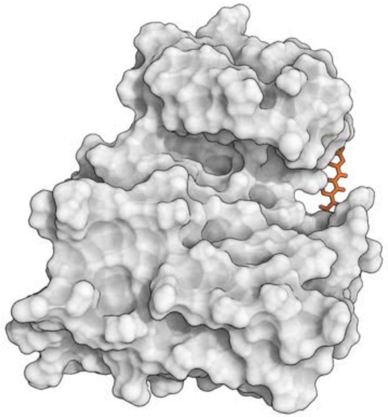
Cyclophilin



Catalysis

Wapeesittipan, **Mey**, et al., *Comms. Chem.* **2**, 41 (2019)

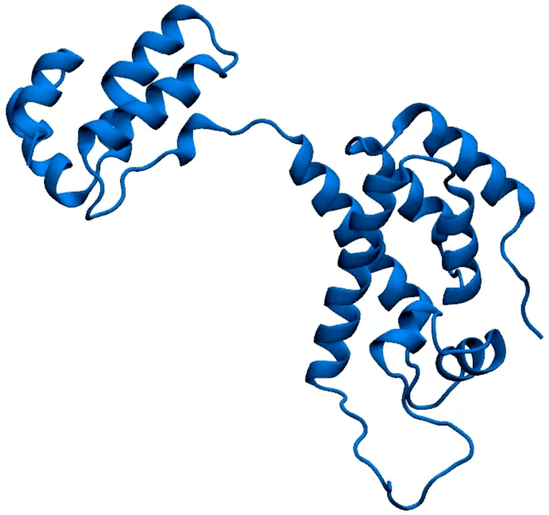
Tyrosine kinase —dasatanib



Binding affinities

Shan, et al. *JACS* **133,** 9181 (2011)

HIV Capsomer

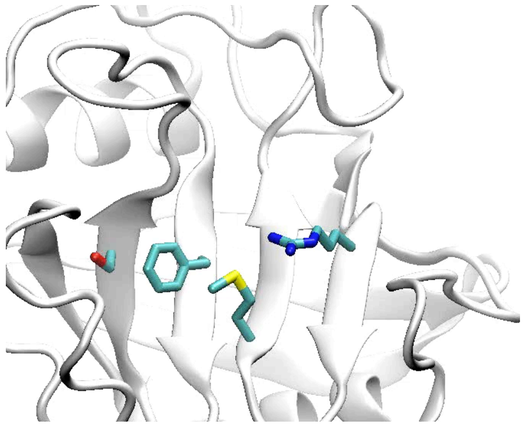


Assembly

Degiacomi, *Structure* **27**, 1034 (2019)

Nature is not 2-dimensional

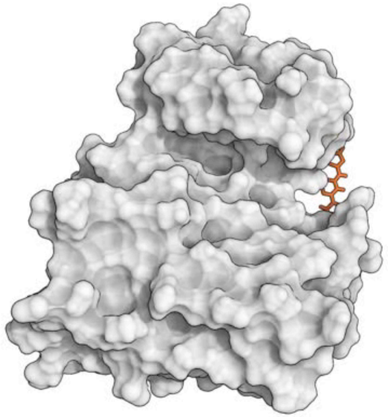
Cyclophilin



Catalysis

Wapeesittipan, **Mey**, et al., *Comms. Chem.* **2**, 41 (2019)

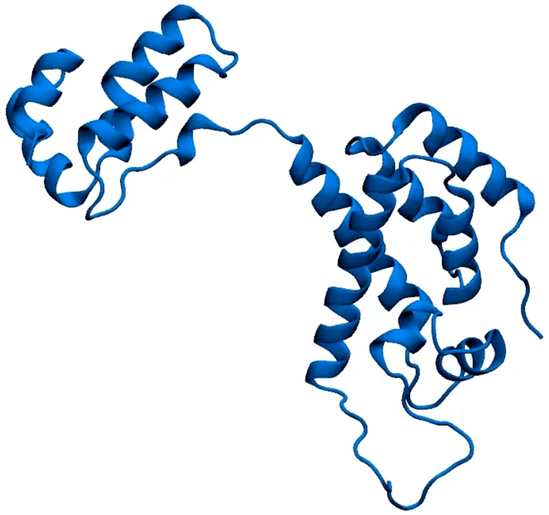
Tyrosine kinase —dasatanib



Binding affinities

Shan, et al. *JACS* **133,** 9181 (2011)

HIV Capsomer

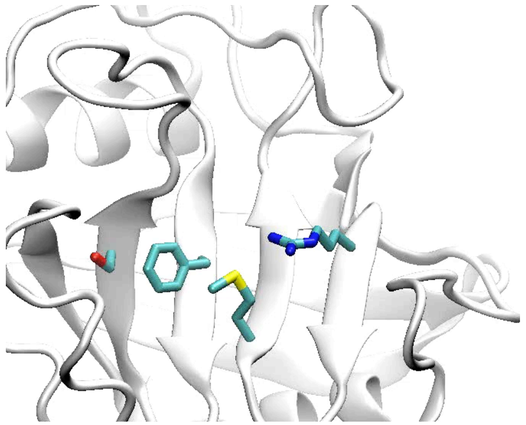


Assembly

Degiacomi, *Structure* **27**, 1034 (2019)

Nature is not 2-dimensional

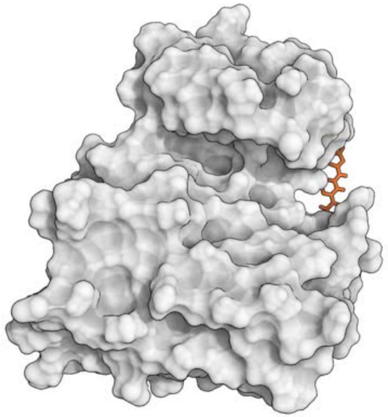
Cyclophilin



Catalysis

Wapeesittipan, **Mey**, et al., *Comms. Chem.* **2**, 41 (2019)

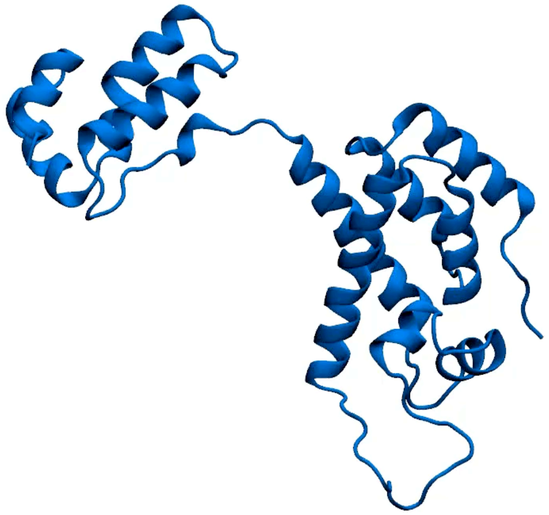
Tyrosine kinase —dasatanib



Binding affinities

Shan, et al. *JACS* **133,** 9181 (2011)

HIV Capsomer

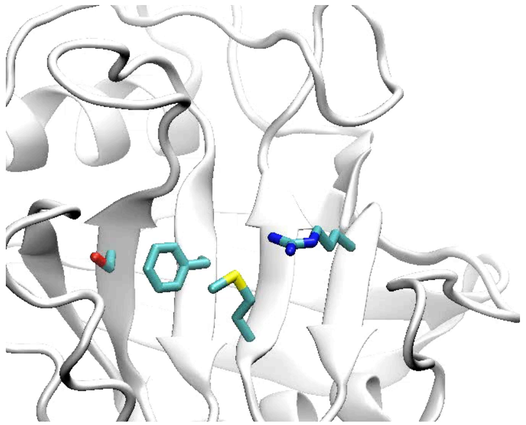


Assembly

Degiacomi, *Structure* **27**, 1034 (2019)

Nature is not 2-dimensional

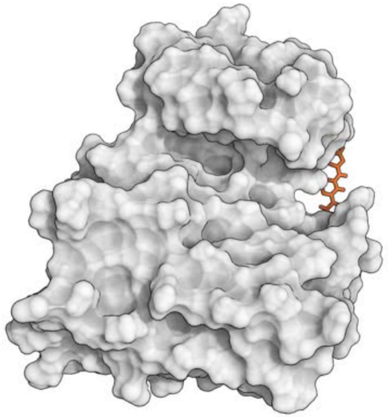
Cyclophilin



Catalysis

Wapeesittipan, **Mey**, et al., *Comms. Chem.* **2**, 41 (2019)

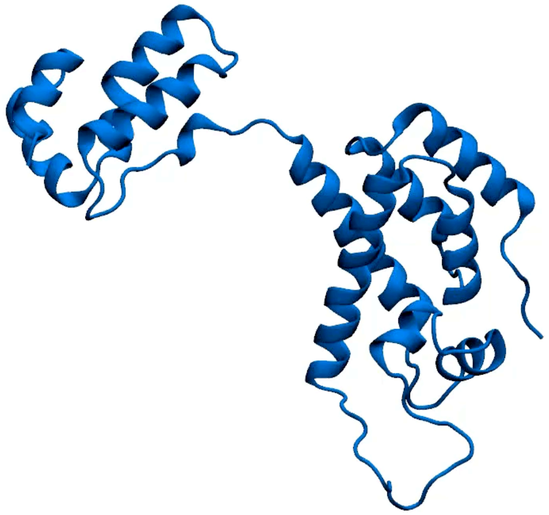
Tyrosine kinase —dasatanib



Binding affinities

Shan, et al. *JACS* **133,** 9181 (2011)

HIV Capsomer

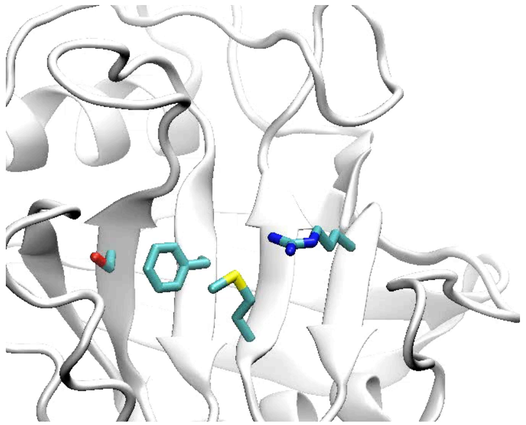


Assembly

Degiacomi, *Structure* **27**, 1034 (2019)

Nature is not 2-dimensional

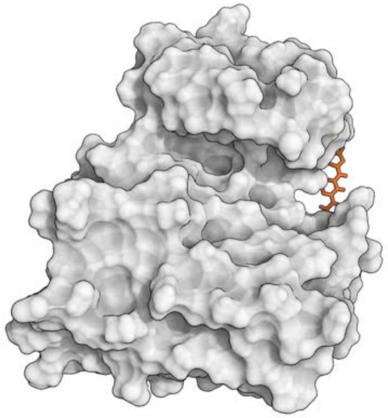
Cyclophilin



Catalysis

Wapeesittipan, **Mey**, et al., *Comms. Chem.* **2,** 41 (2019)

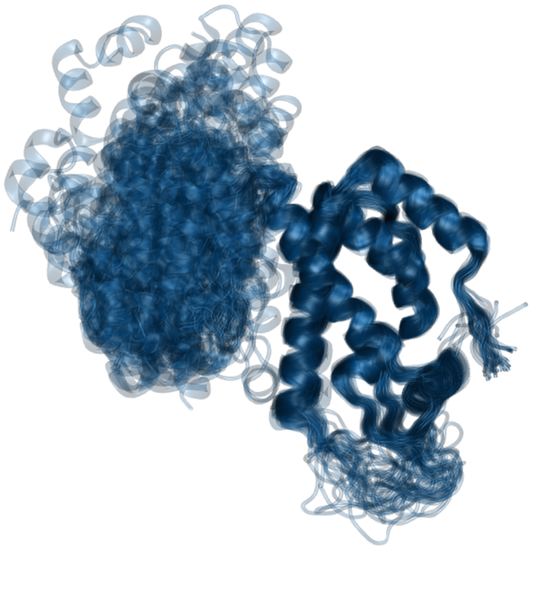
Tyrosine kinase —dasatanib



Binding affinities

Shan, et al. *JACS* **133,** 9181 (2011)

HIV Capsomer

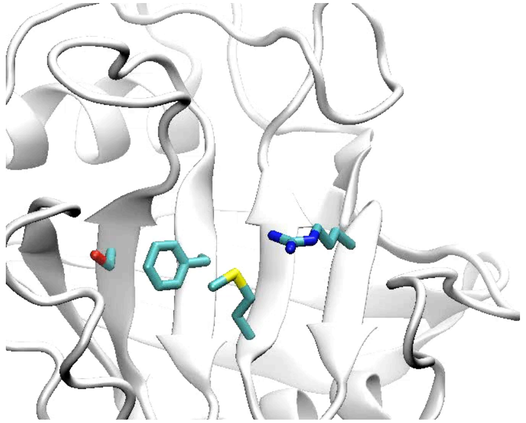


Assembly

Degiacomi, *Structure* **27**, 1034 (2019)

Nature is not 2-dimensional

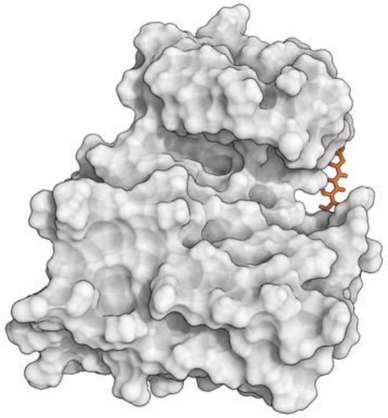
Cyclophilin



Catalysis

Wapeesittipan, **Mey**, et al., *Comms. Chem.* **2,** 41 (2019)

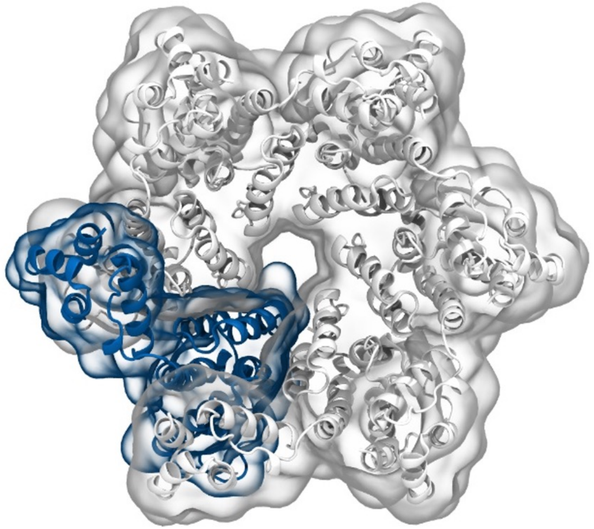
Tyrosine kinase —dasatanib



Binding affinities

Shan, et al. *JACS* **133,** 9181 (2011)

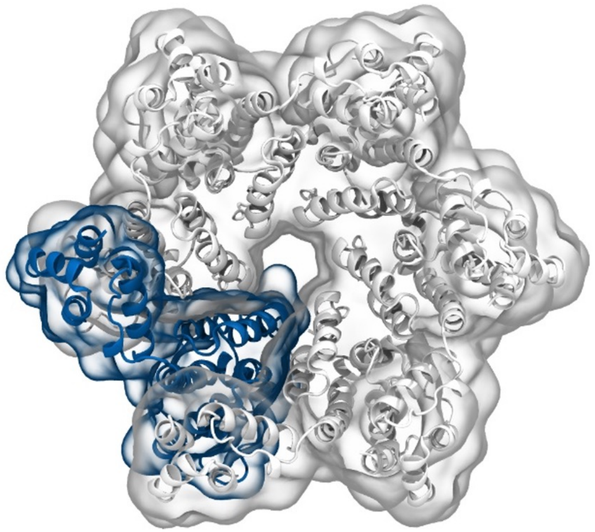
HIV Capsomer



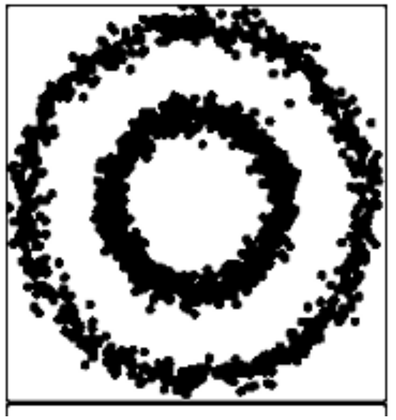
Assembly

Degiacomi, *Structure* **27**, 1034 (2019)

The many dimensions of chemistry



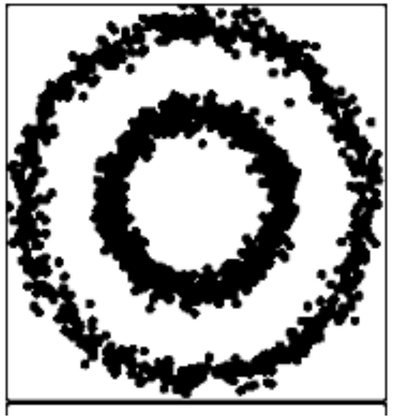
≠



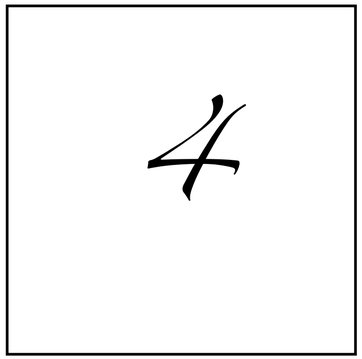
Looking at dimensions of an image

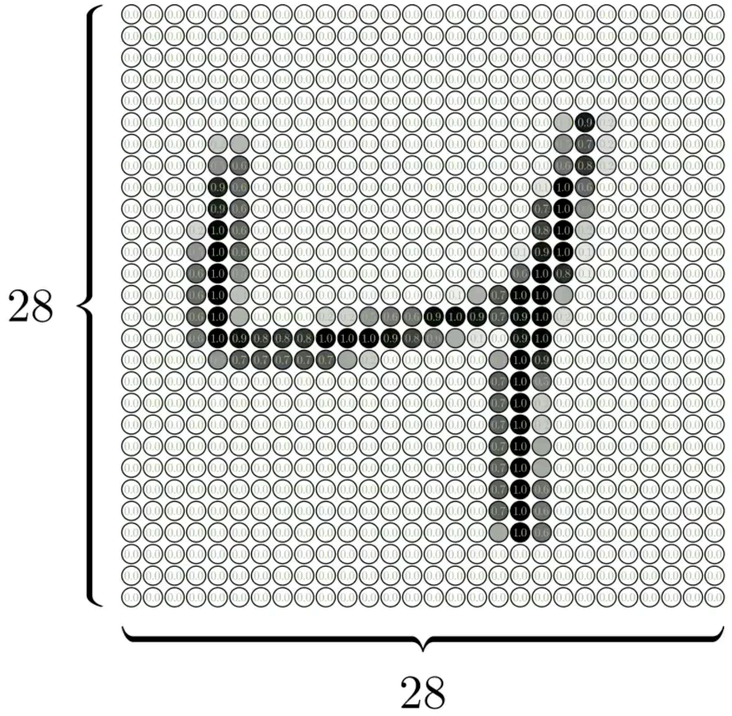
Features provide input

2D coordinates



N-D vector

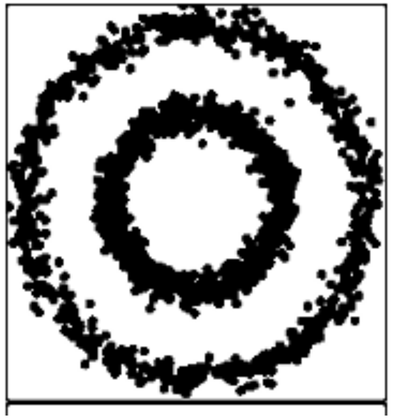




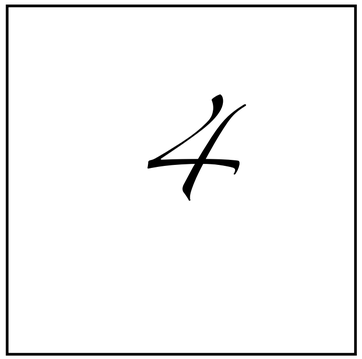
Looking at dimensions of an image

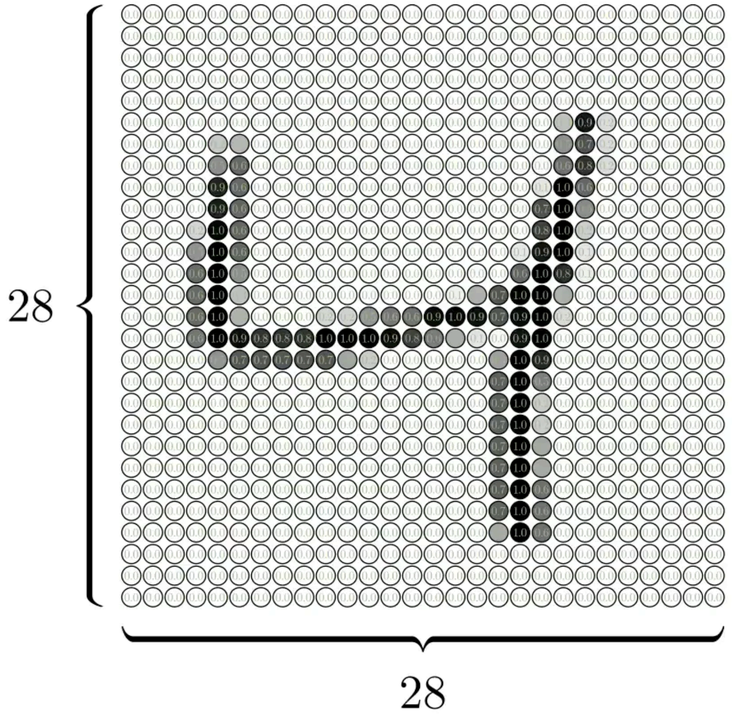
Features provide input

2D coordinates



N-D vector



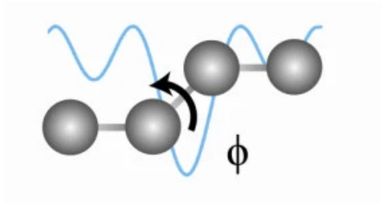


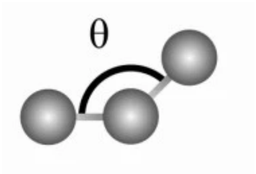
Features are possible representation of data as input

Features from coordinates

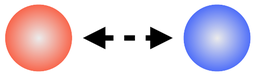
Charge

Dihedral





Angle



Distances

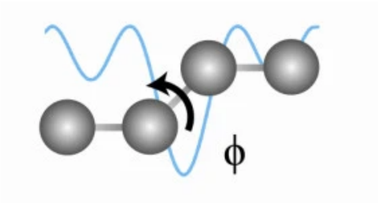
Bonds

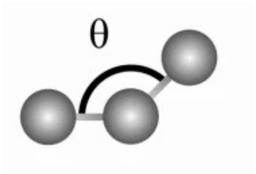
Features are possible representation of data as input

Features from coordinates

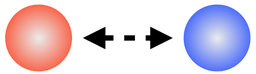
Charge

Dihedral





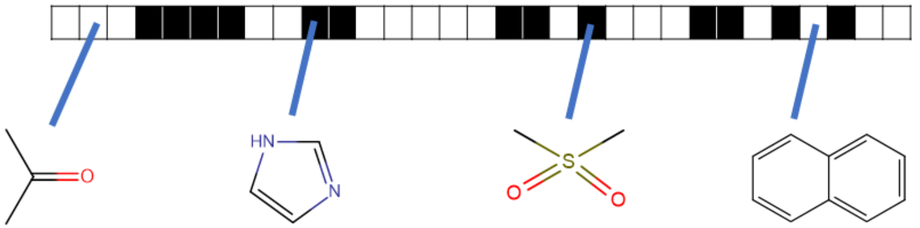
Angle

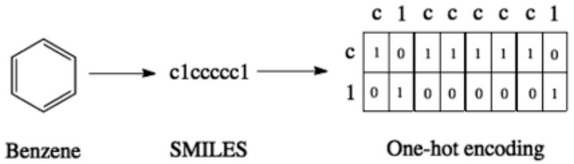


Distances

Bonds

Other features



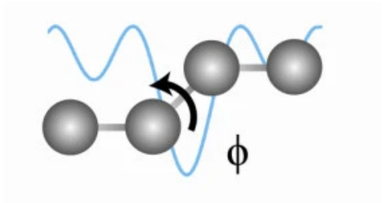


Smiles string

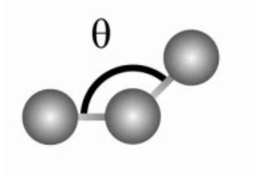
Features are possible representation of data as input

Features from coordinates

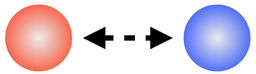
Charge



Dihedral



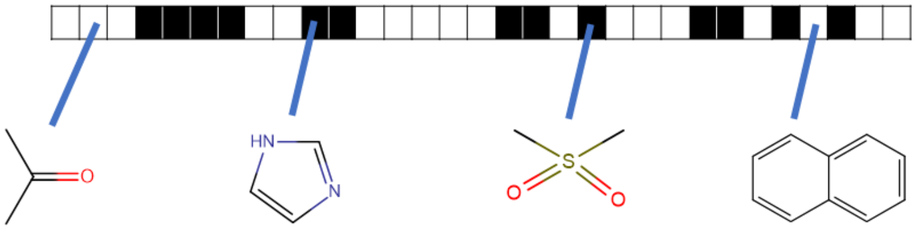
Angle

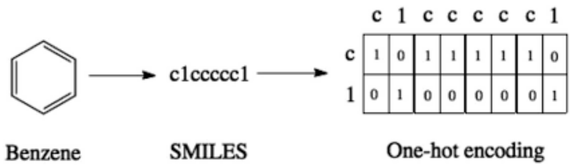


Distances

Bonds

Other features



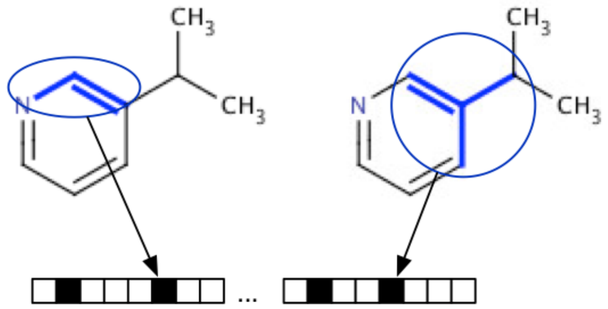


Smiles string

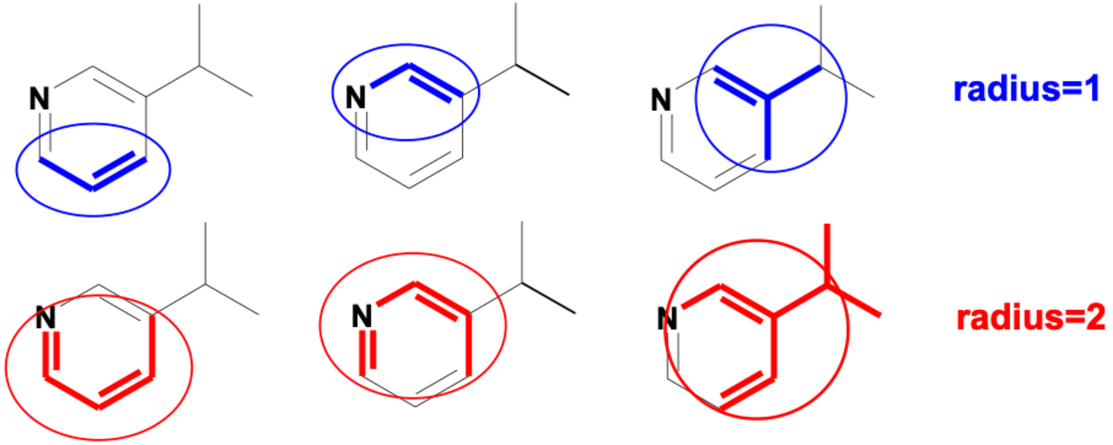
Feature vectors

```
['ATOM:ACE 1 CH3 1 x',
 'ATOM:ACE 1 CH3 1 y',
 'ATOM:ACE 1 CH3 1 z',
 'ATOM:ACE 1 C 4 x',
 'ATOM:ACE 1 C 4 y',
 'ATOM:ACE 1 C 4 z',
 'ATOM:ACE 1 O 5 x',
 'ATOM:ACE 1 O 5 y',
 'ATOM:ACE 1 O 5 z',
 'ATOM:ALA 2 N 6 x',
 'ATOM:ALA 2 N 6 y',
 'ATOM:ALA 2 N 6 z',
 'ATOM:ALA 2 CA 8 x',
 'ATOM:ALA 2 CA 8 y',
 'ATOM:ALA 2 CA 8 z',
 'ATOM:ALA 2 CB 10 x',
 'ATOM:ALA 2 CB 10 y',
 'ATOM:ALA 2 CB 10 z',
 'ATOM:ALA 2 C 14 x',
 'ATOM:ALA 2 C 14 y',
 'ATOM:ALA 2 C 14 z',
 'ATOM:ALA 2 O 15 x',
 'ATOM:ALA 2 O 15 y',
 'ATOM:ALA 2 O 15 z',
 'ATOM:NME 3 N 16 x',
 'ATOM:NME 3 N 16 y',
 'ATOM:NME 3 N 16 z',
 'ATOM:NME 3 C 18 x',
 'ATOM:NME 3 C 18 y',
 'ATOM:NME 3 C 18 z']
```

Molecular fingerprints are a typical way of encoding molecule information



- Similar fingerprints mean similar molecules
- There are many different ways of defining a fingerprint

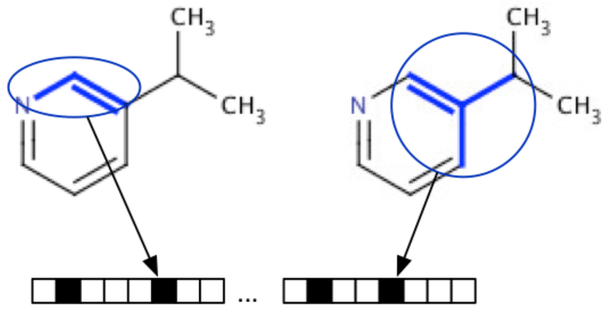


Atom types :

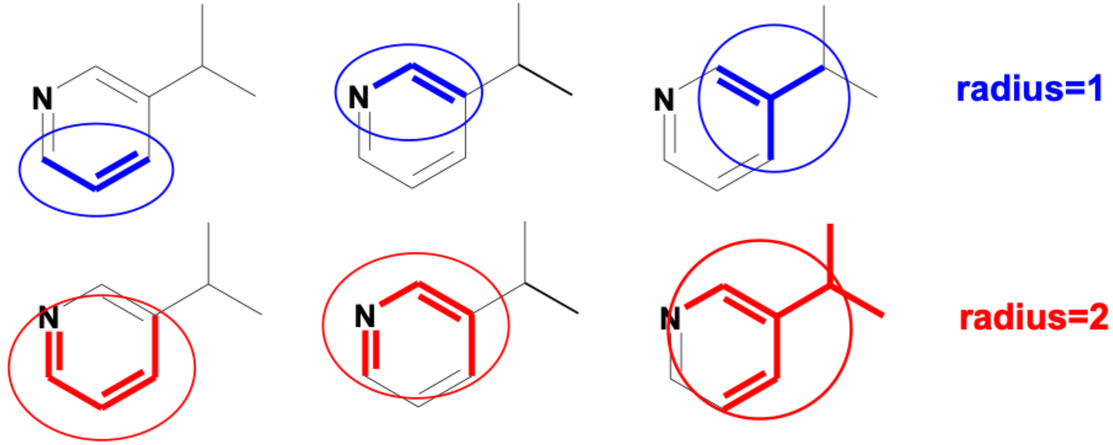
- Connectivity: (Element, #heavy neighbors, #Hs, charge, isotope, inRing)
- Chemical features: Donor, Acceptor, Aromatic, Halogen, Basic, Acidic
- Fingerprint takes into account the neighborhood of each atom (typical radii 0-3 bonds)

RDKit
Open-Source Cheminformatics and Machine Learning

Molecular fingerprints are a typical way of encoding molecule information

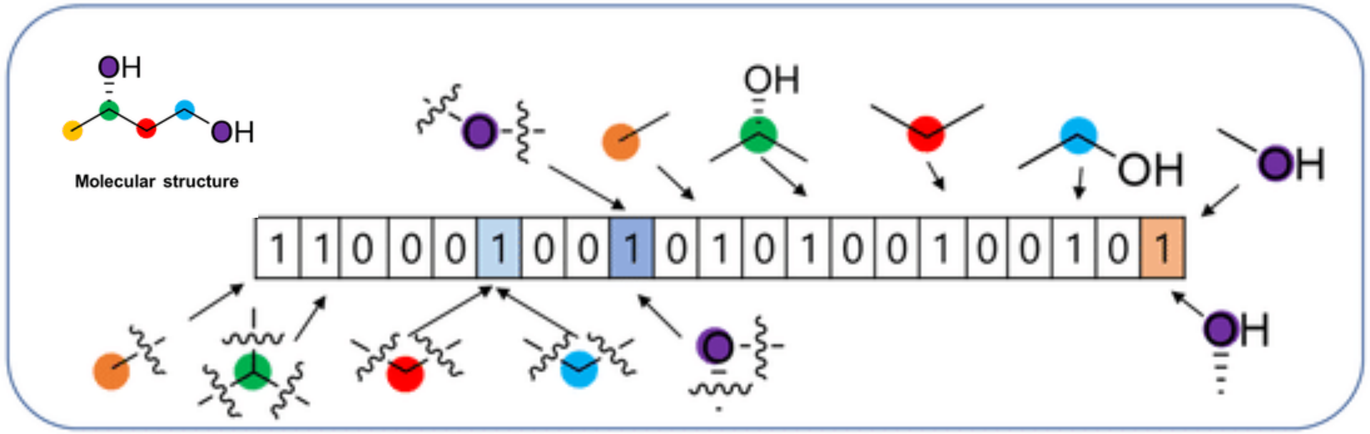


- Similar fingerprints mean similar molecules
- There are many different ways of defining a fingerprint



Atom types :

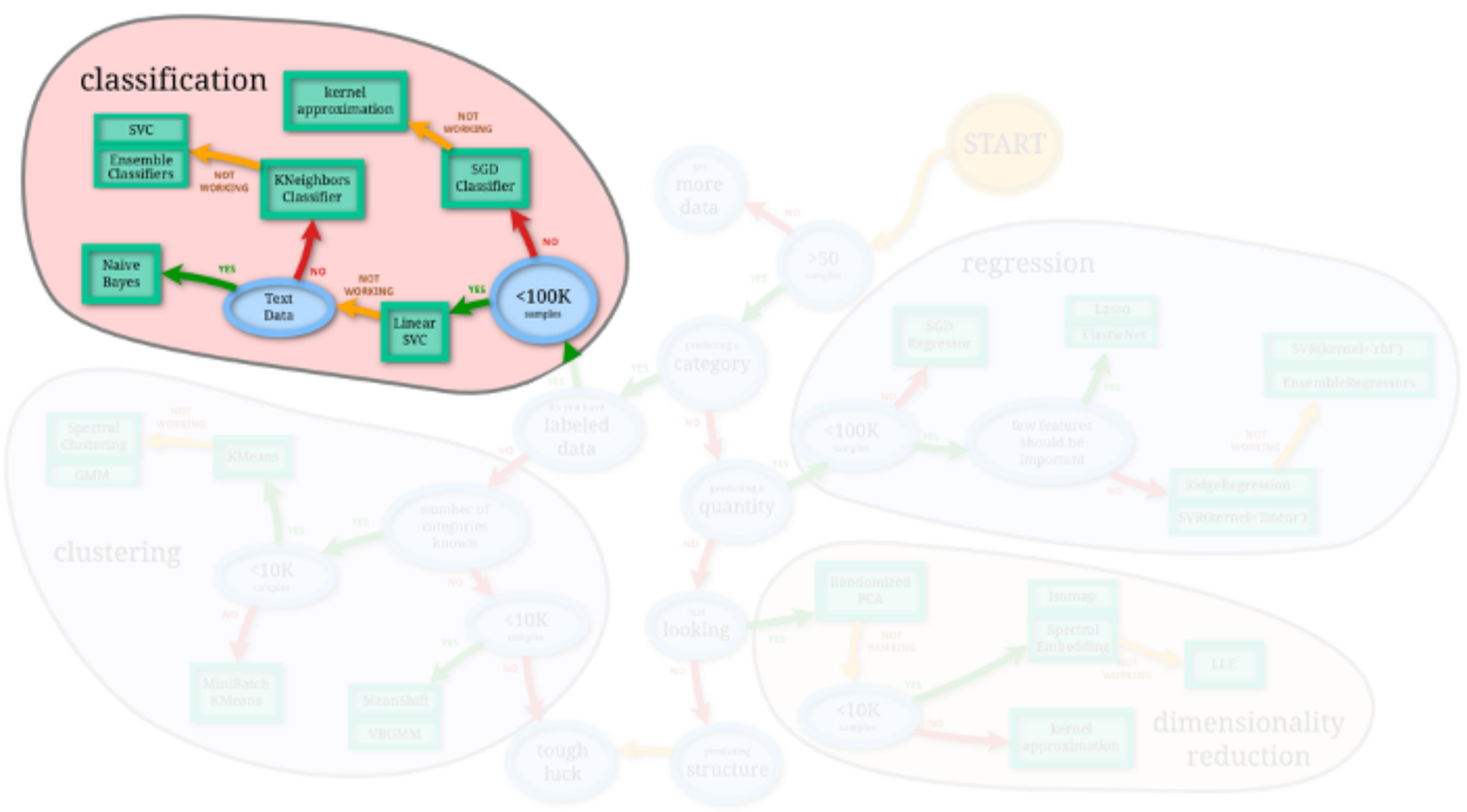
- Connectivity: (Element, #heavy neighbors, #Hs, charge, isotope, inRing)
- Chemical features: Donor, Acceptor, Aromatic, Halogen, Basic, Acidic
- Fingerprint takes into account the neighborhood of each atom (typical radii 0-3 bonds)



Original Morgan fingerprints

RDKit
Open-Source Cheminformatics and Machine Learning

The Data Mining World — Classification problem



From **scikit-learn.org**

What is machine learning?

Artificial intelligence

Design an intelligent agent that perceives its environment and makes decisions to maximise chances of achieving its goal.

Machine learning

Gives computers the ability to learn without specifically being programmed (Arthur Samuel 1959)

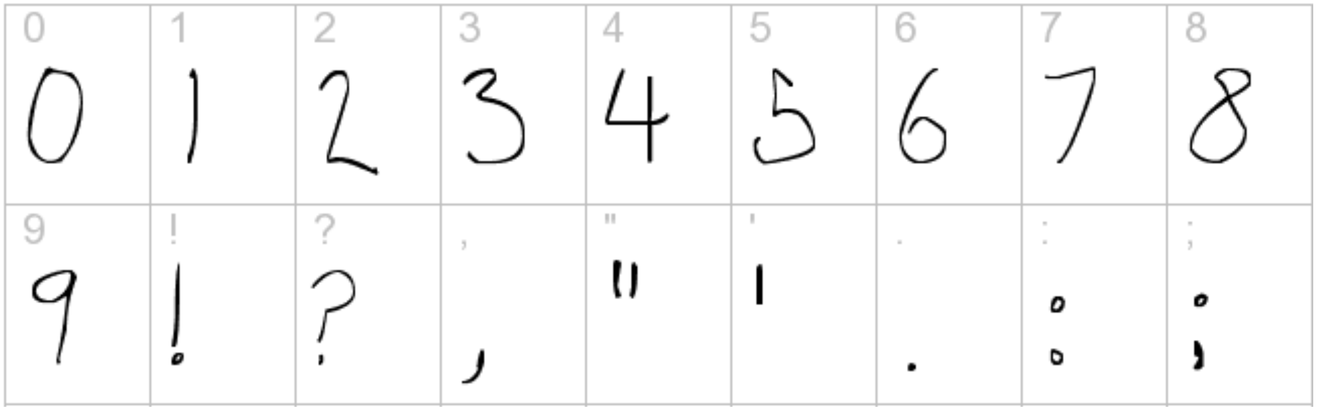
Supervised learning

Unsupervised learning

reinforcement learning

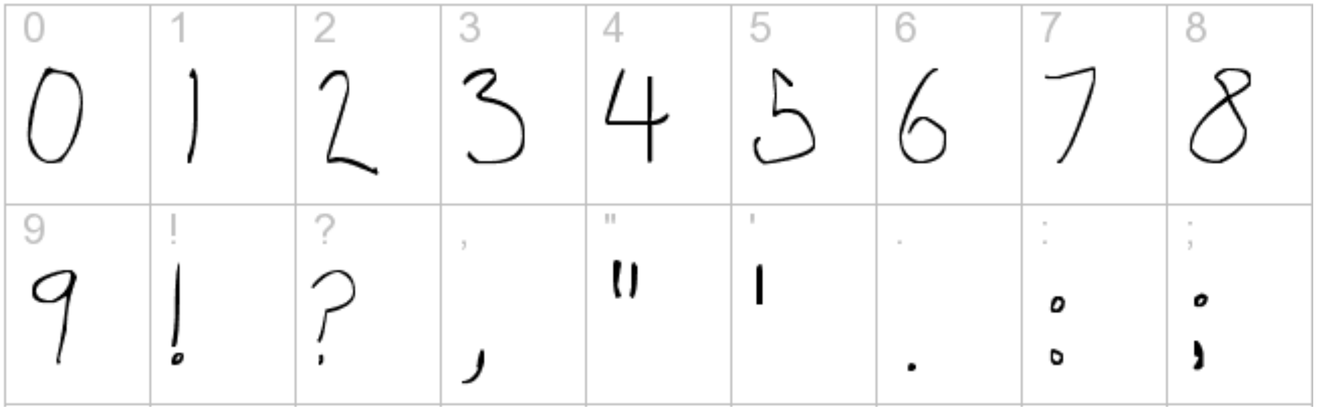
https://medium.com/machine-learning-for-humans/why-machine-learning-matters-6164faf1df12

Classification problems are ubiquitous



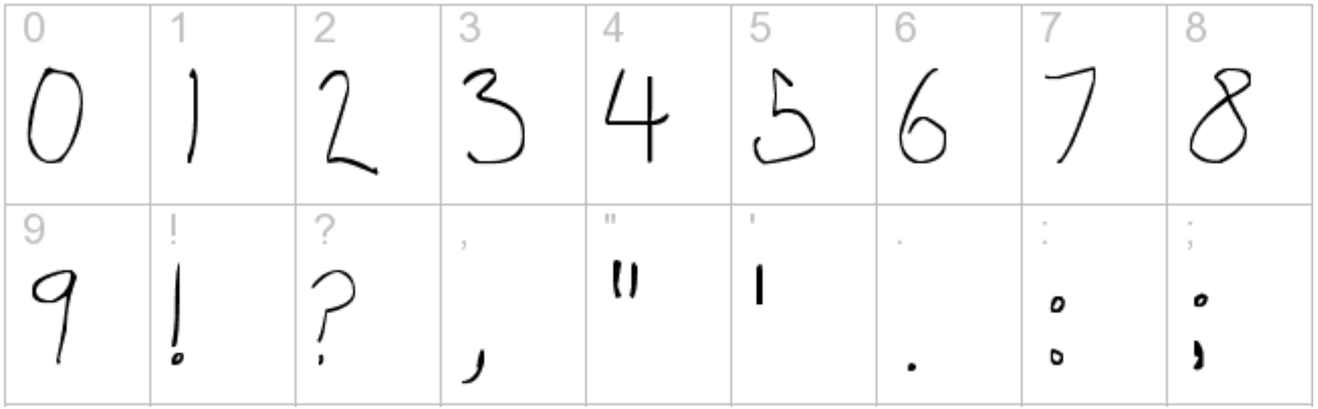
http://yann.lecun.com/exdb/mnist/

Classification problems are ubiquitous

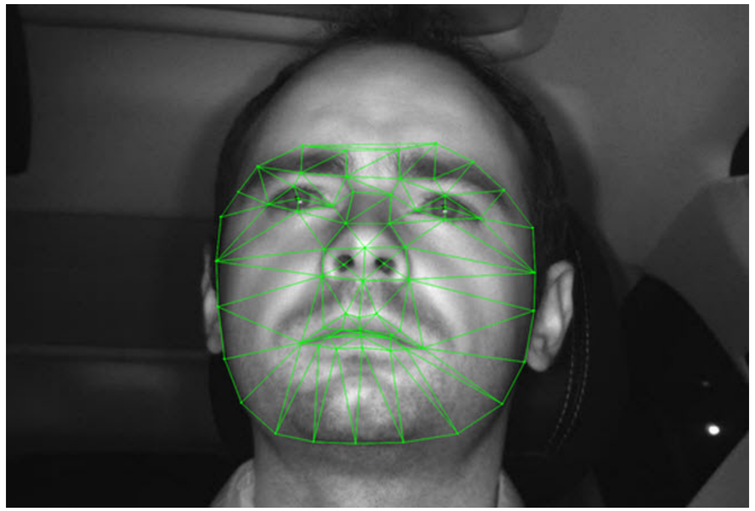


http://yann.lecun.com/exdb/mnist/

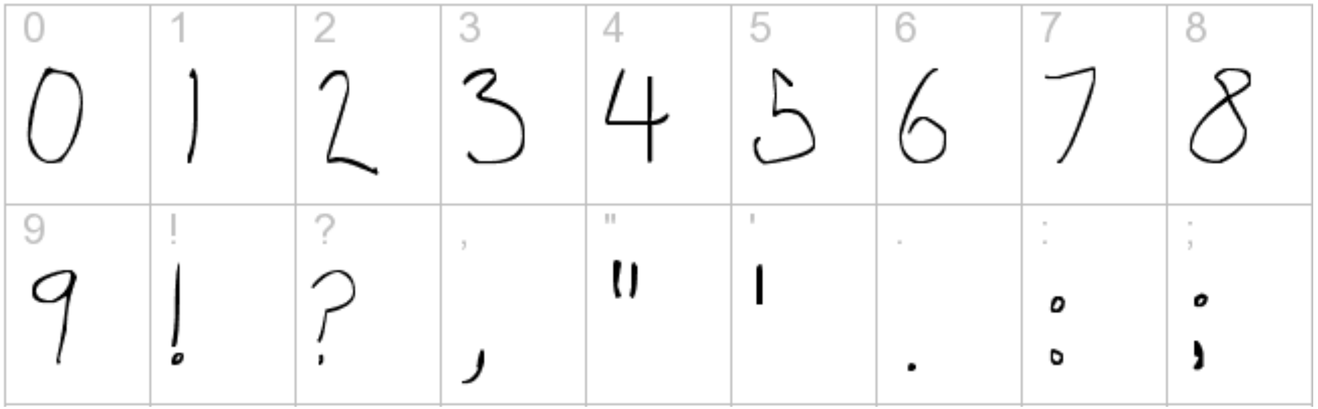
Classification problems are ubiquitous



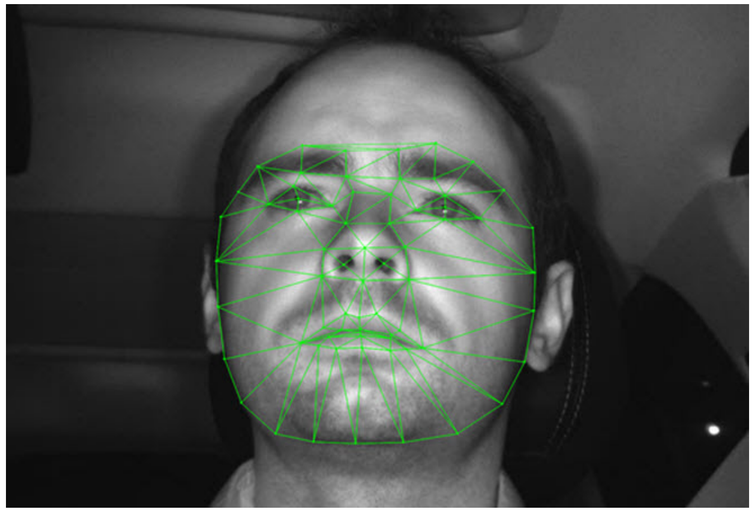
http://yann.lecun.com/exdb/mnist/

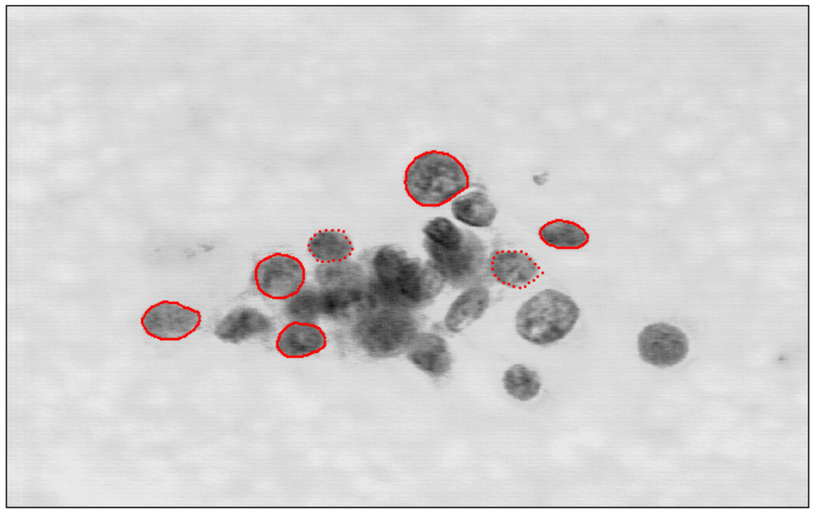


Classification problems are ubiquitous

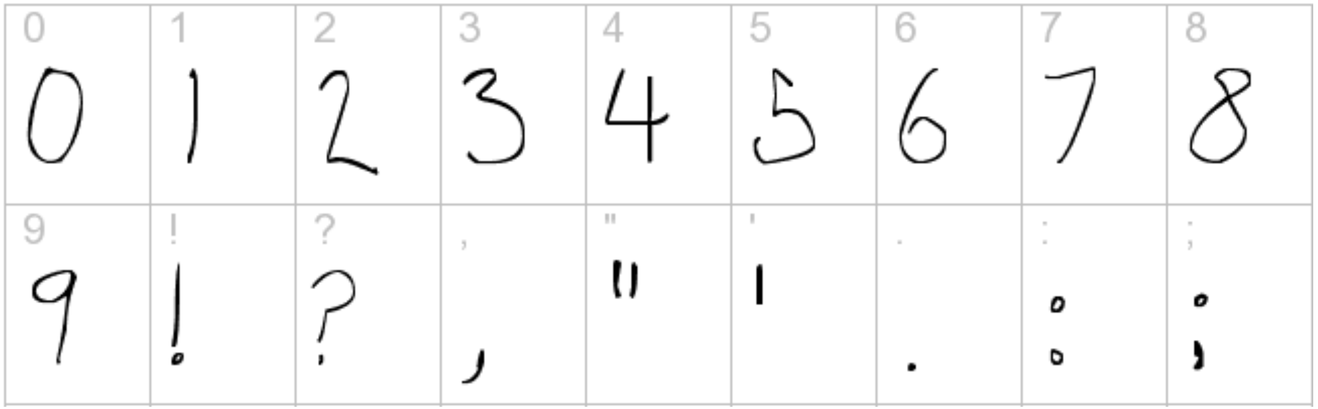


http://yann.lecun.com/exdb/mnist/

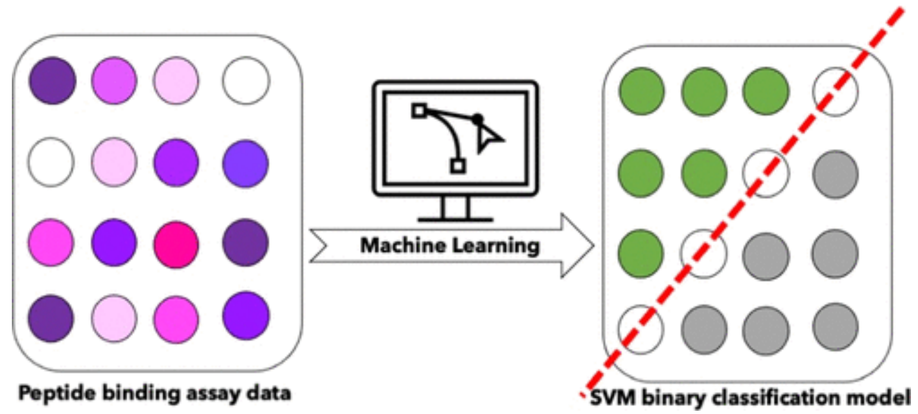




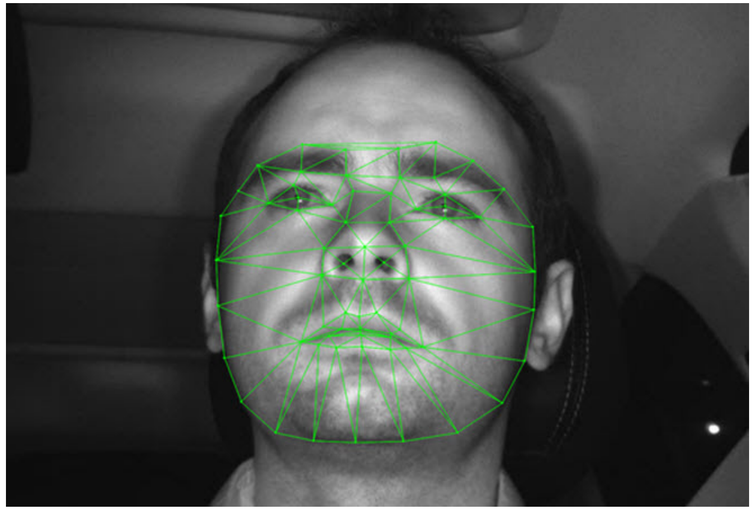
Classification problems are ubiquitous

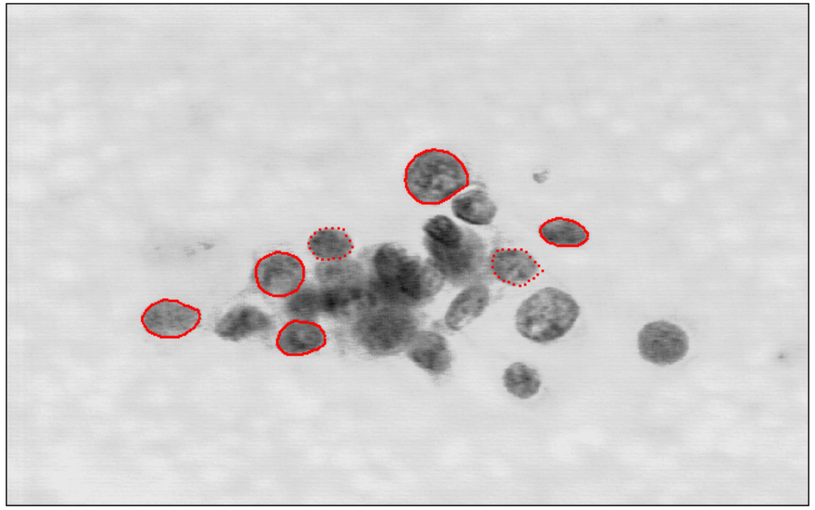


http://yann.lecun.com/exdb/mnist/

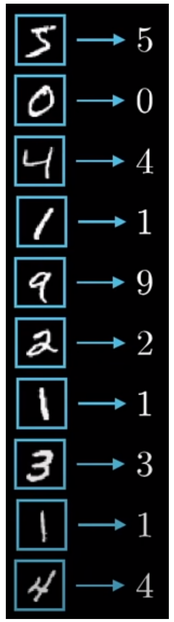


ACS Omega 2022, 7, 16, 14069–14073



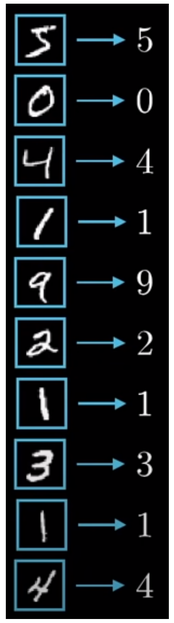


Labels are needed for supervised learning tasks



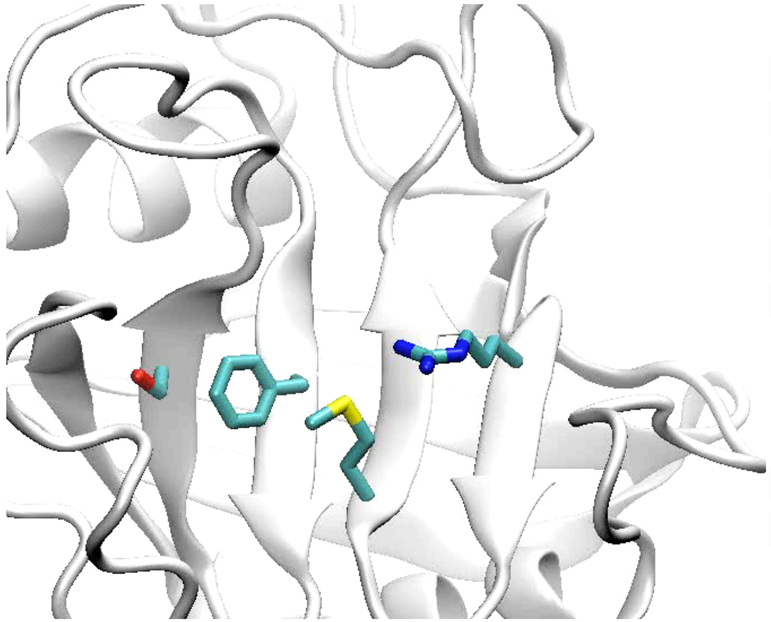
Label

Labels are needed for supervised learning tasks



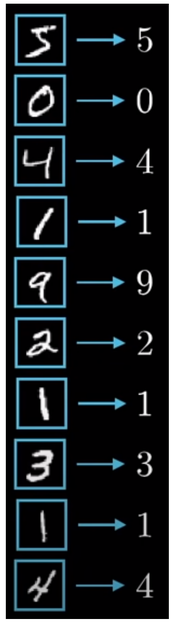
Label

Labelled energies



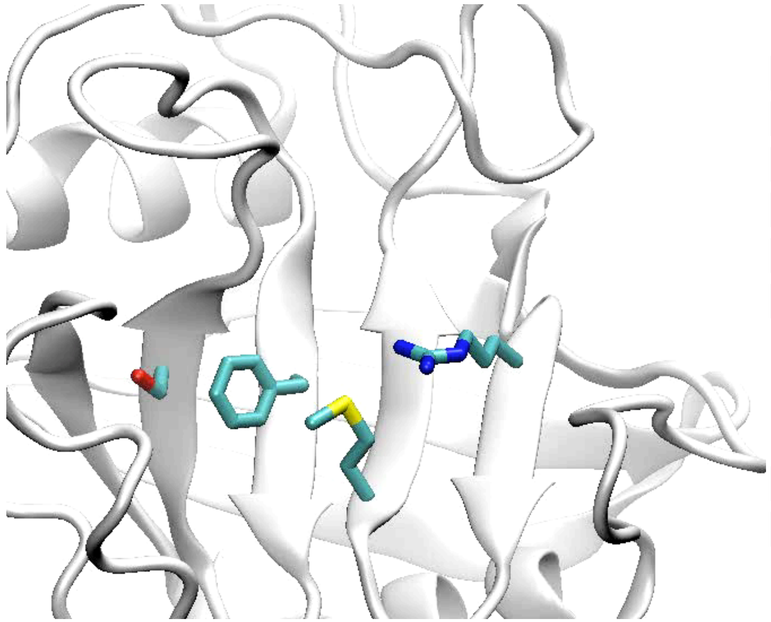
[-12008 kJ]
[-12078 kJ]
[-12045 kJ]
[-12083 kJ]
[-12062 kJ]
[-12058 kJ]
.
.
.
[-12099 kJ]
[-12093 kJ]

Labels are needed for supervised learning tasks



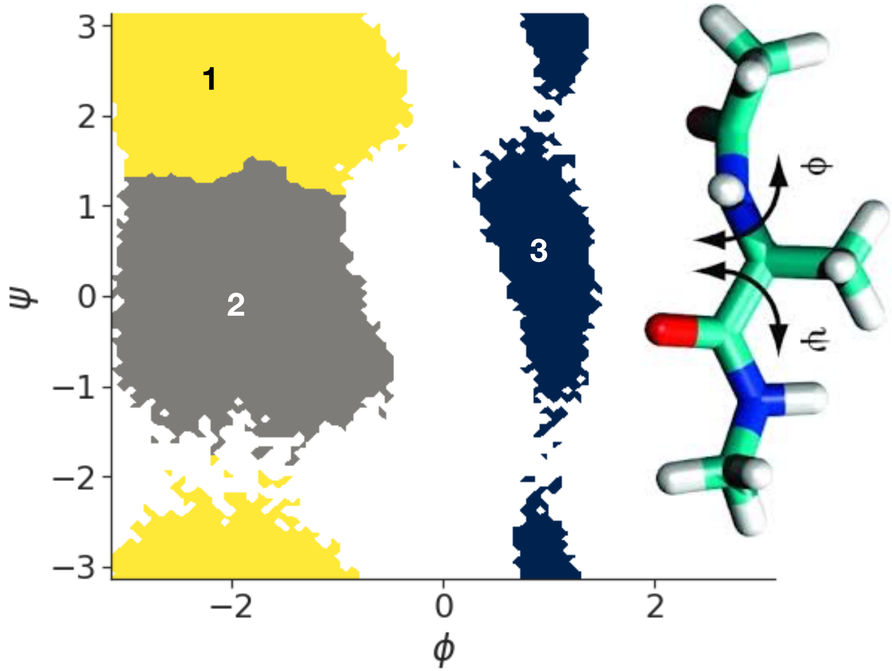
Label

Labelled energies



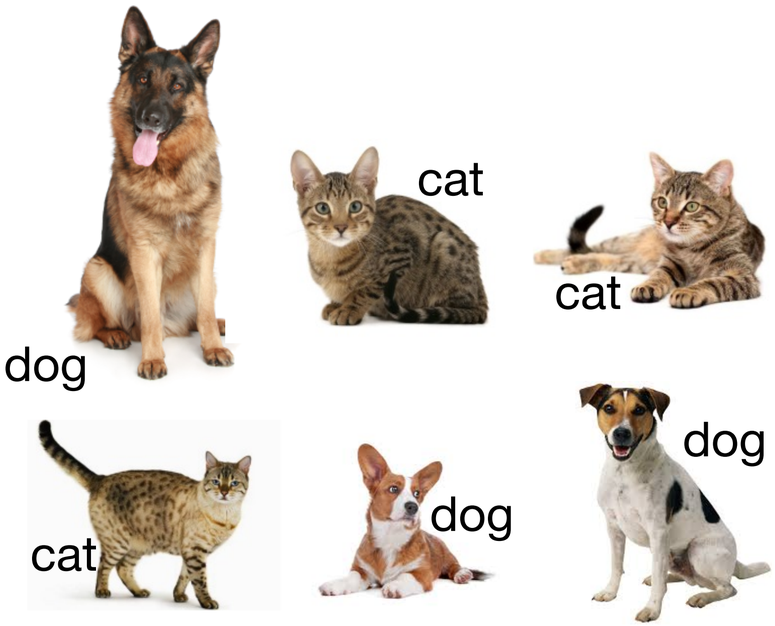
[-12008 kJ]
[-12078 kJ]
[-12045 kJ]
[-12083 kJ]
[-12062 kJ]
[-12058 kJ]
.
.
.
[-12099 kJ]
[-12093 kJ]

Labelled states



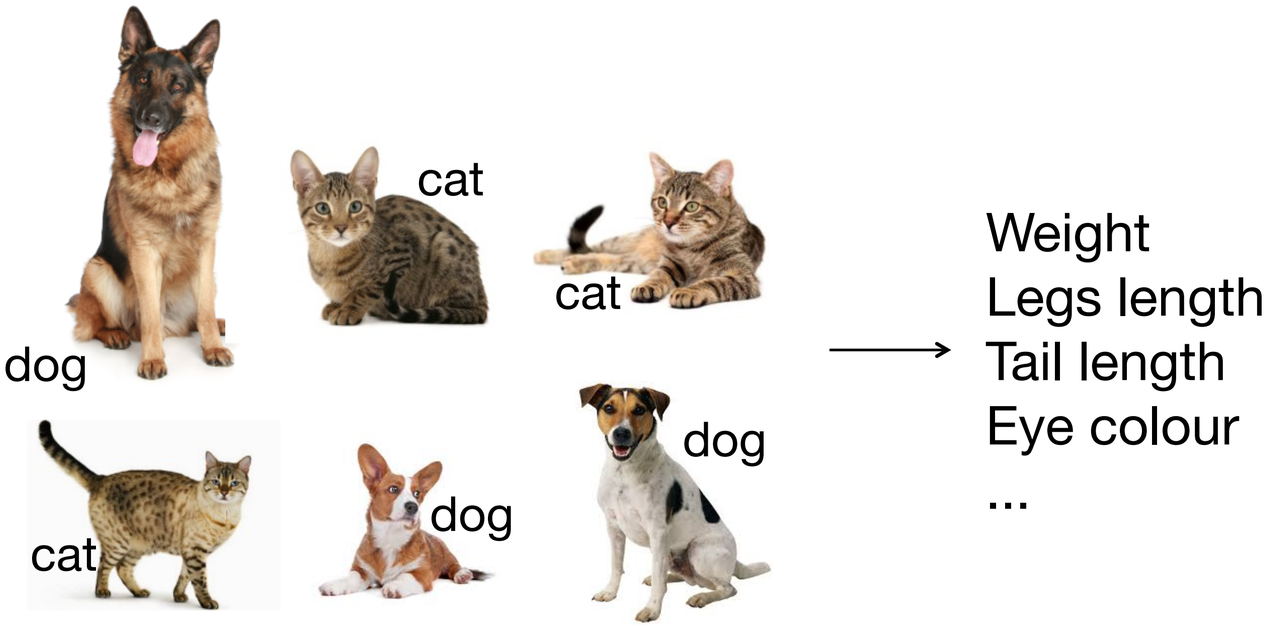
Data Classification via Supervised Learning

- Take **labelled data**



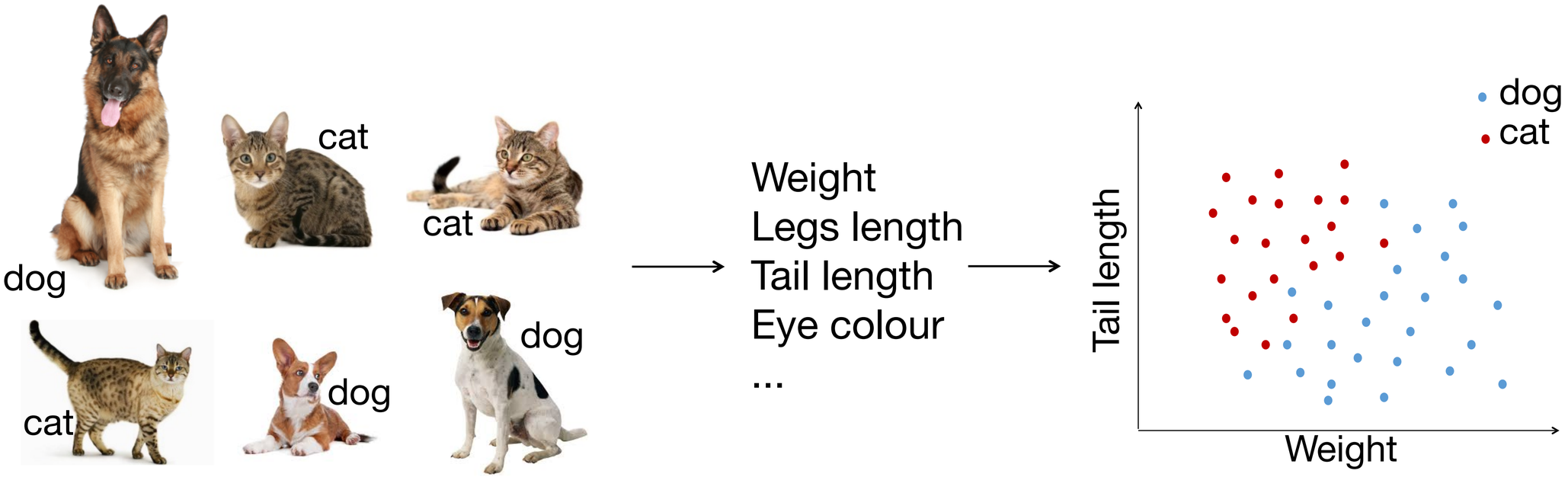
Data Classification via Supervised Learning

- Take **labelled data**
- Create an n-dimensional **feature vector** from data



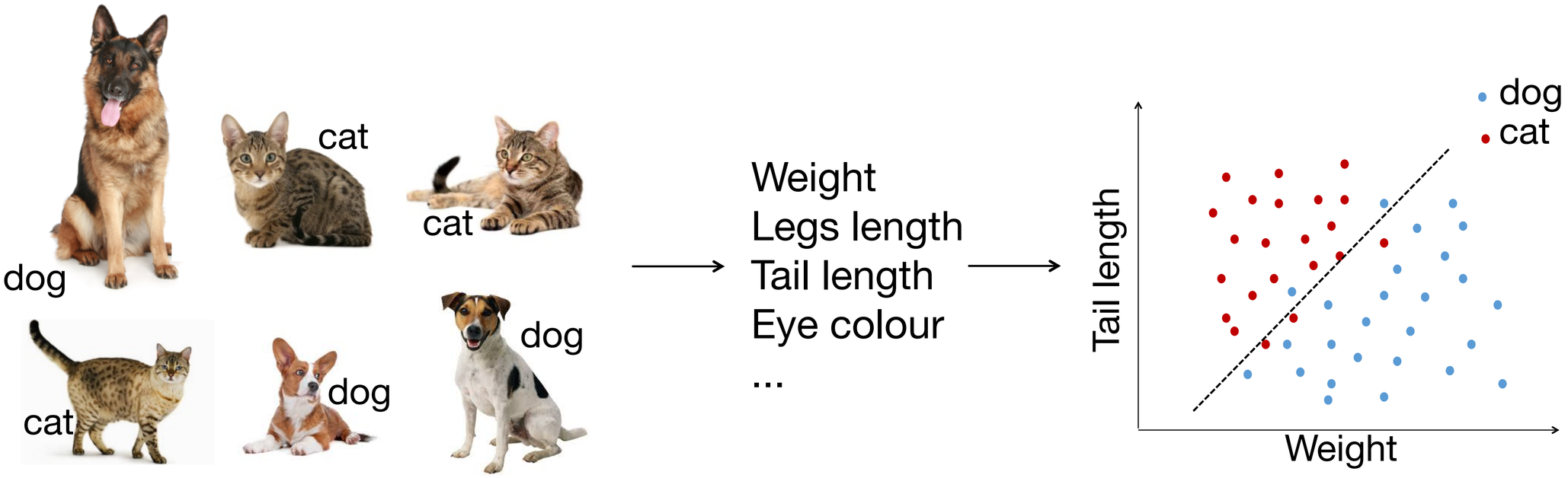
Data Classification via Supervised Learning

- Take **labelled data**
- Create an n-dimensional **feature vector** from data



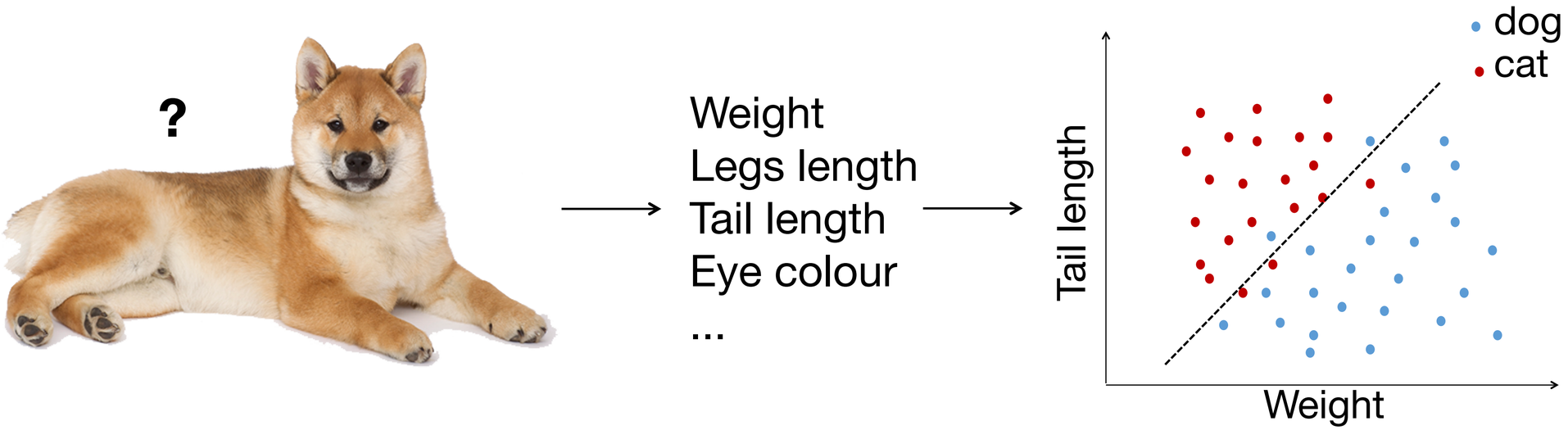
Data Classification via Supervised Learning

- Take **labelled data**
- Create an n-dimensional **feature vector** from data
- Separate “feature space” in different regions



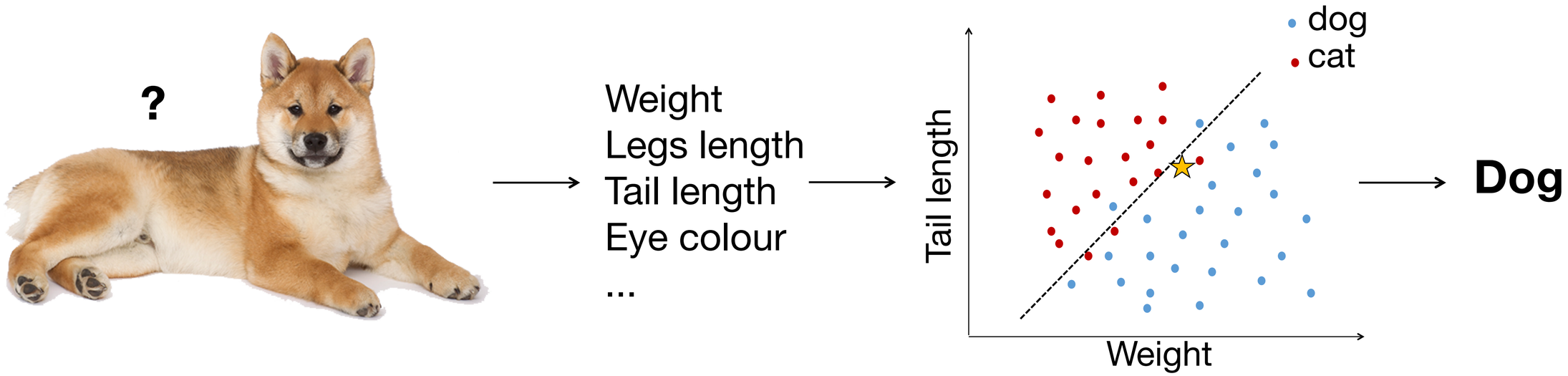
Data Classification via Supervised Learning

- Take **labelled data**
- Create an n-dimensional **feature vector** from data
- Separate “feature space” in different regions



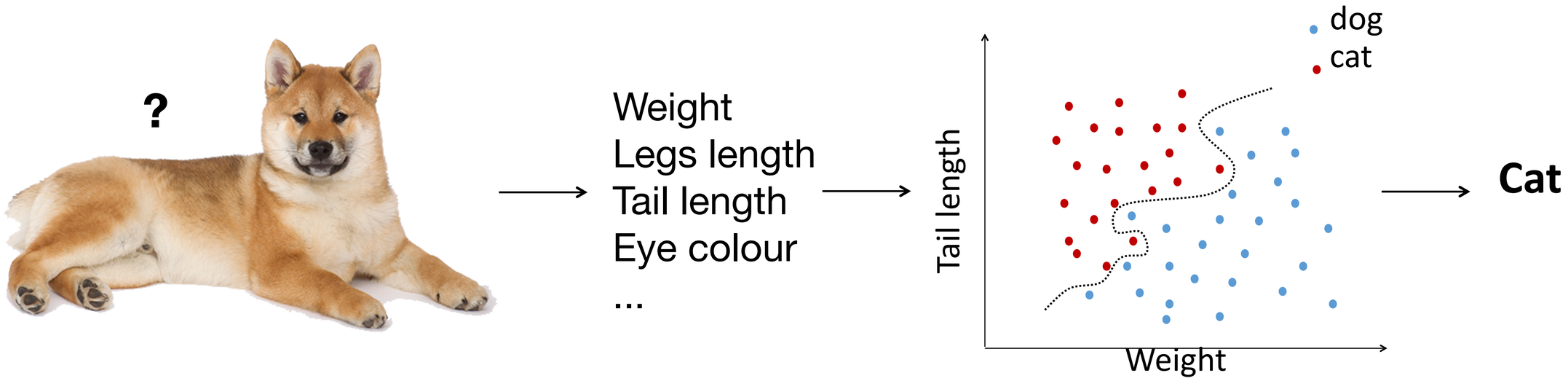
Data Classification via Supervised Learning

- Take **labelled data**
- Create an n-dimensional **feature vector** from data
- Separate “feature space” in different regions



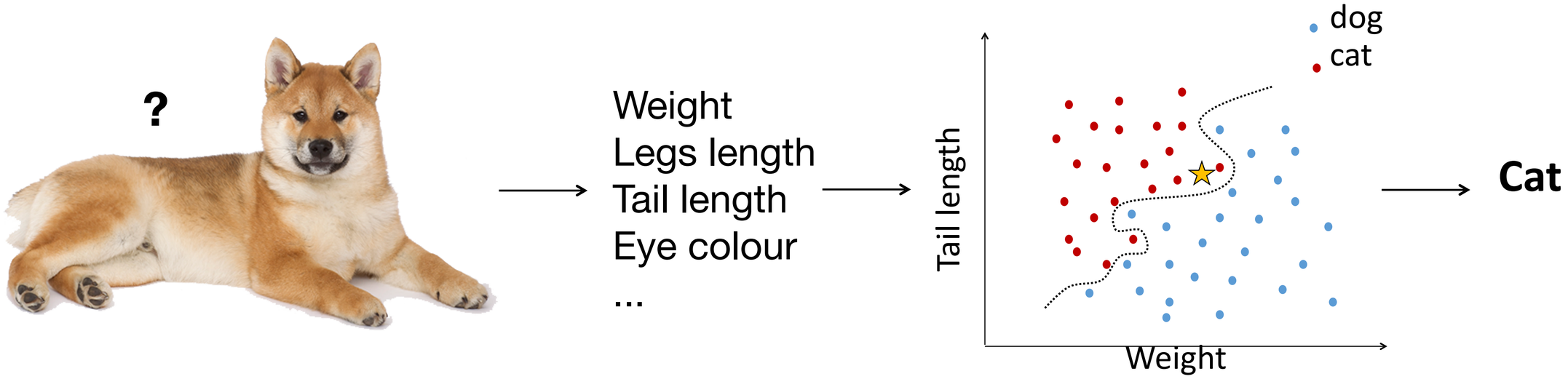
Data Classification via Supervised Learning

- Take **labelled data**
- Create an n-dimensional **feature vector** from data
- Separate “feature space” in different regions



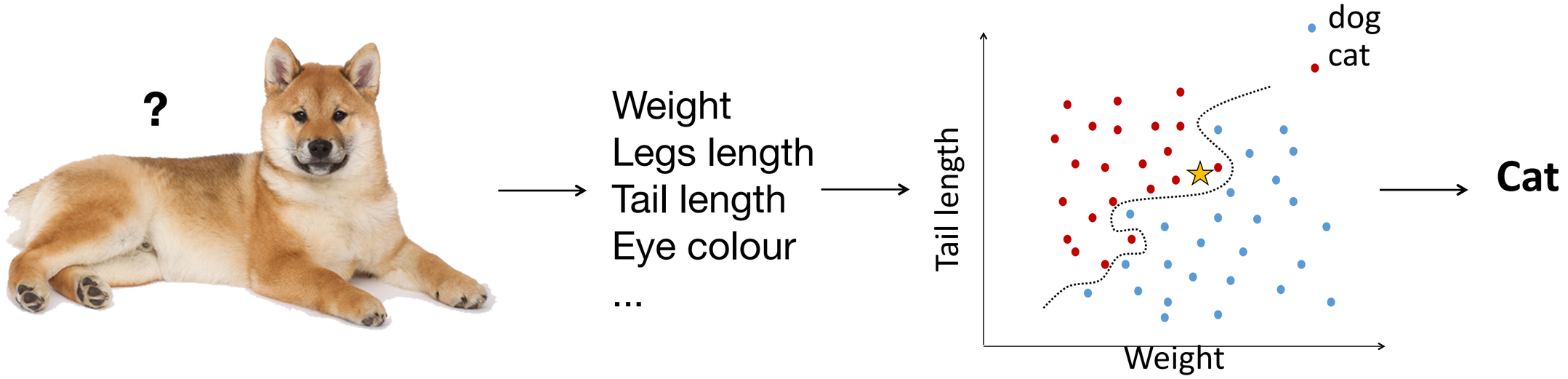
Data Classification via Supervised Learning

- Take **labelled data**
- Create an n-dimensional **feature vector** from data
- Separate “feature space” in different regions



Data Classification via Supervised Learning

- Take **labelled data**
- Create an n-dimensional **feature vector** from data
- Separate “feature space” in different regions
- Warning: a too precise classification of examples might sacrifice generality (**overfitting**)

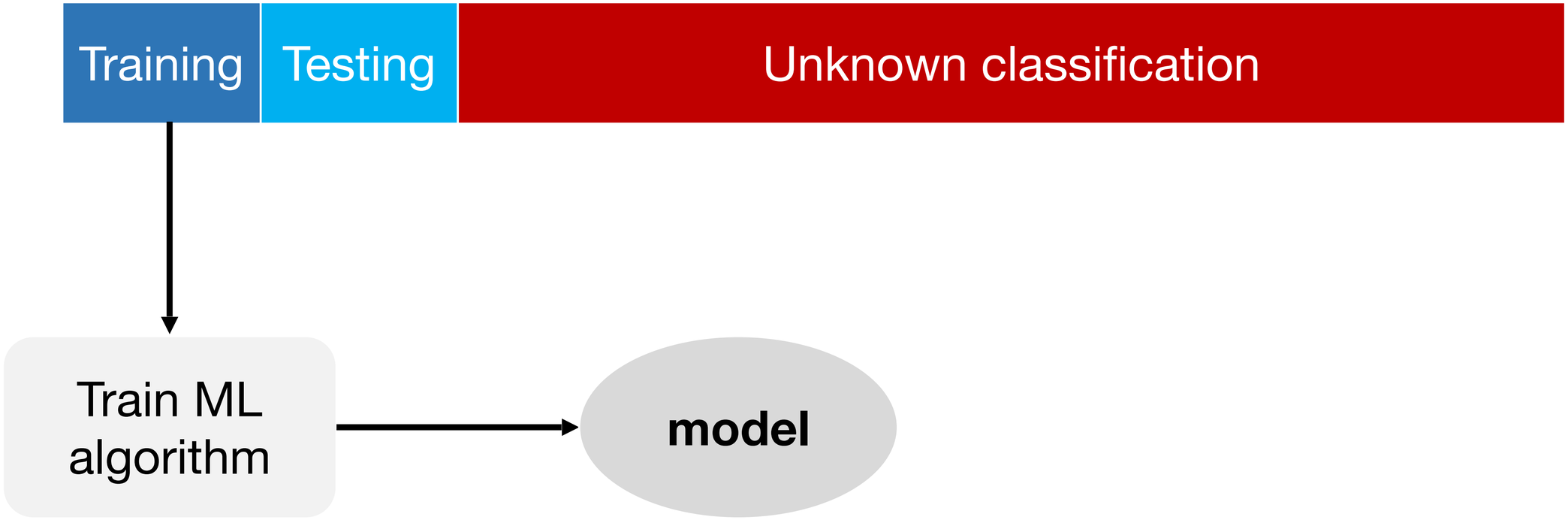


How to handle the data for classification

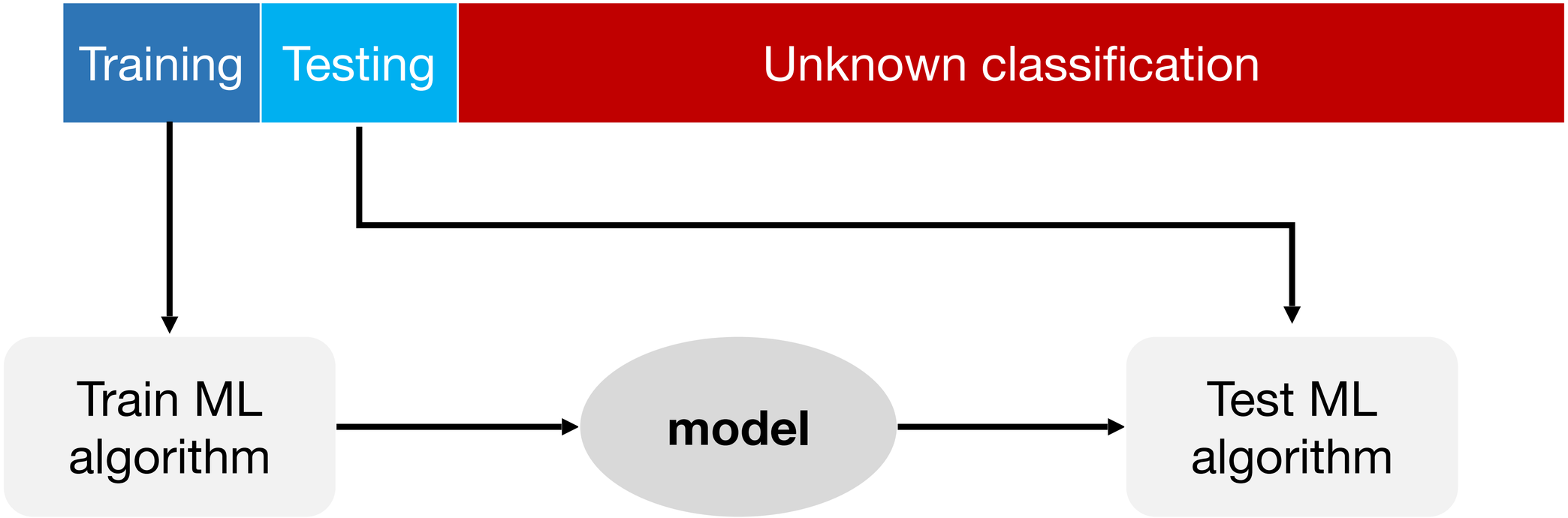
Data

How to handle the data for classification

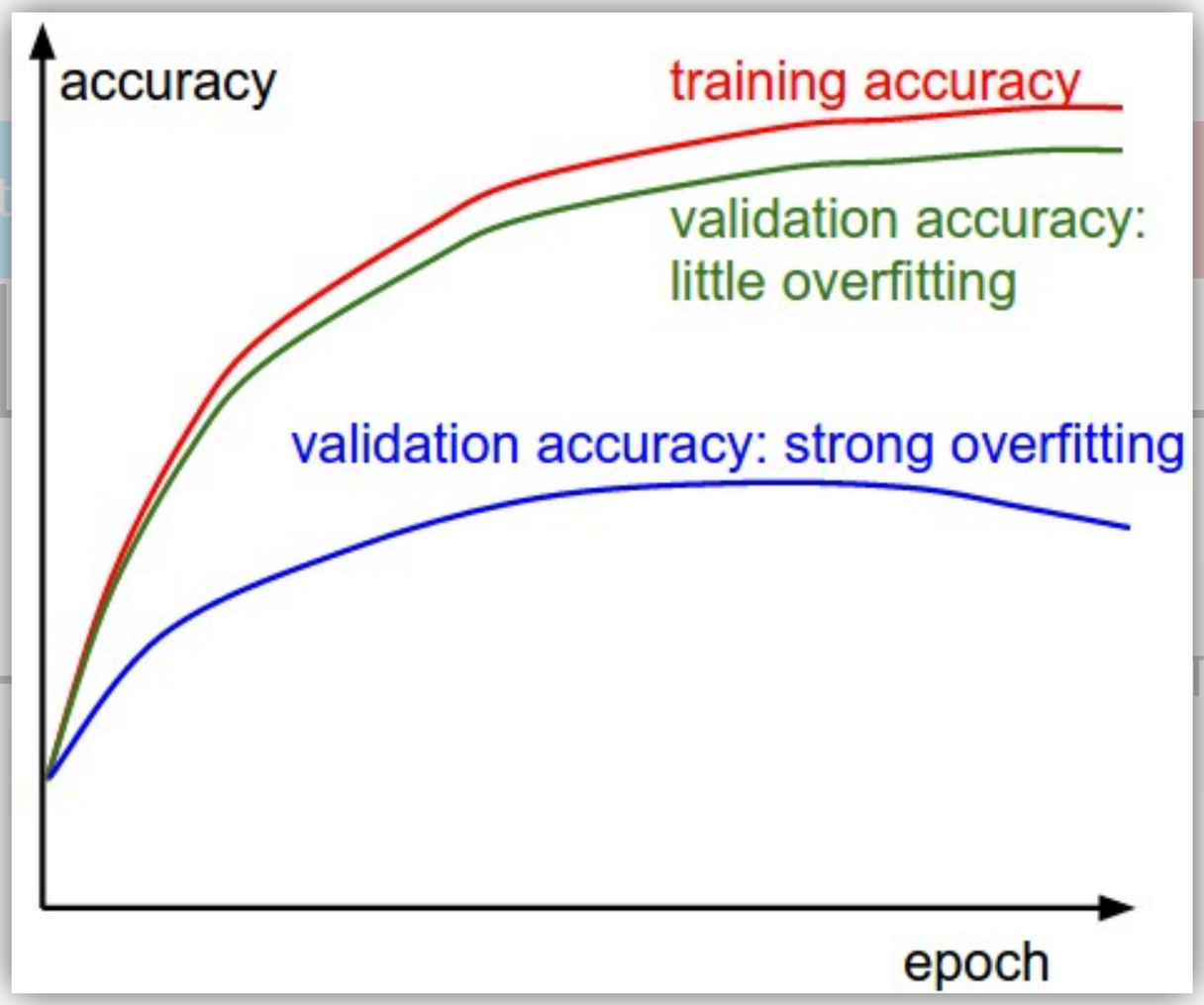
How to handle the data for classification



How to handle the data for classification



Data Classification using supervised learning



More nomenclature

en.wikipedia.org/wiki/Precision_and_recall

More nomenclature

result \ real	Dog	Cat
Dog	90	10
Cat	12	88

en.wikipedia.org/wiki/Precision_and_recall

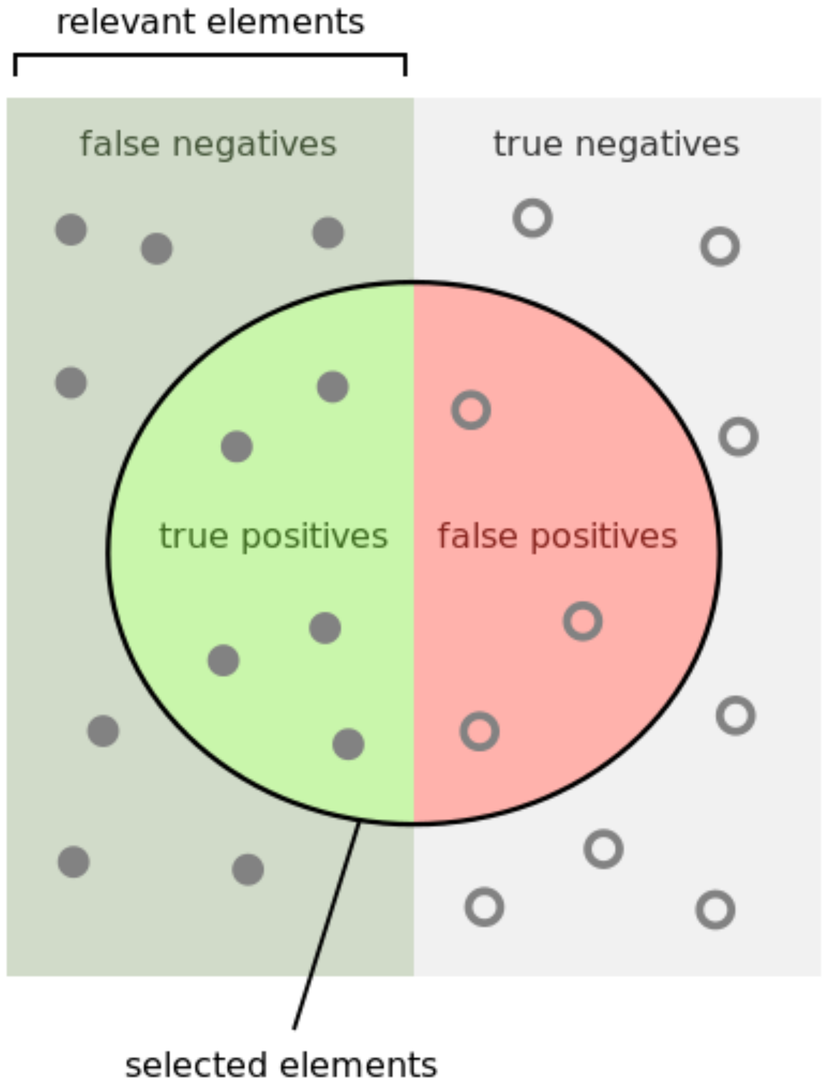
More nomenclature

- **Confusion Matrix:**
  describes classification results
  can also describe n classes

result \ real	Dog	Cat
Dog	90	10
Cat	12	88

en.wikipedia.org/wiki/Precision_and_recall

More nomenclature

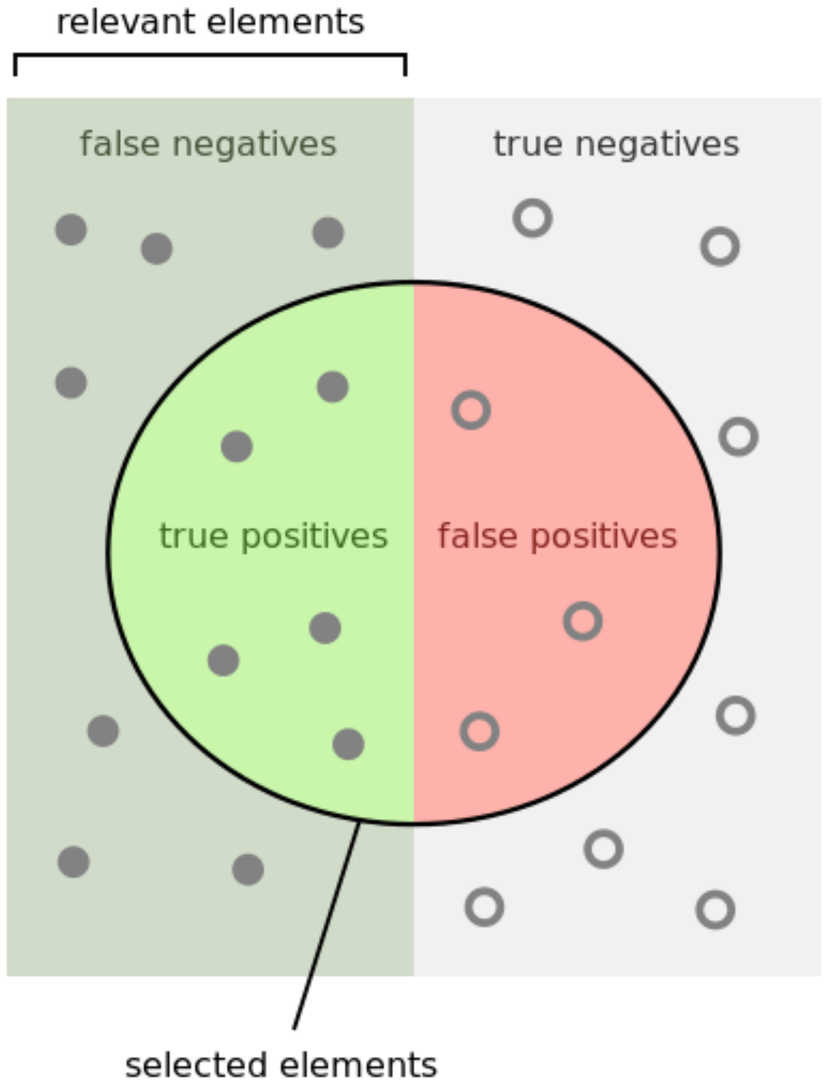


- **Confusion Matrix:**
  describes classification results
  can also describe n classes

result \ real	Dog	Cat
Dog	90	10
Cat	12	88

en.wikipedia.org/wiki/Precision_and_recall

More nomenclature



- **Confusion Matrix:**
  describes classification results
  can also describe n classes

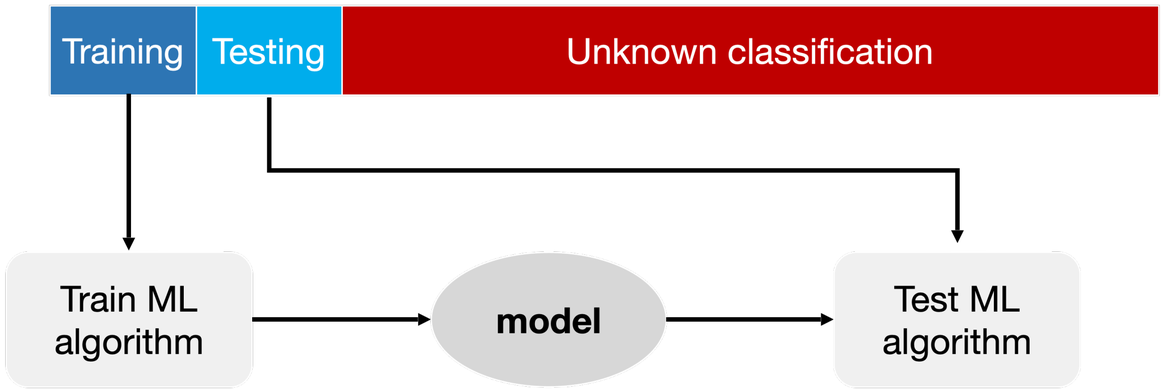
result \ real	Dog	Cat
Dog	90	10
Cat	12	88

- $\textbf{precision} = \dfrac{\text{true positives}}{\text{selected elements}} =$ 
- $\textbf{sensitivity} = \textbf{recall} = \dfrac{\text{true positives}}{\text{relevant elements}} =$ 
- $\textbf{accuracy} = \dfrac{\text{true positives} + \text{true negatives}}{\text{total population}}$

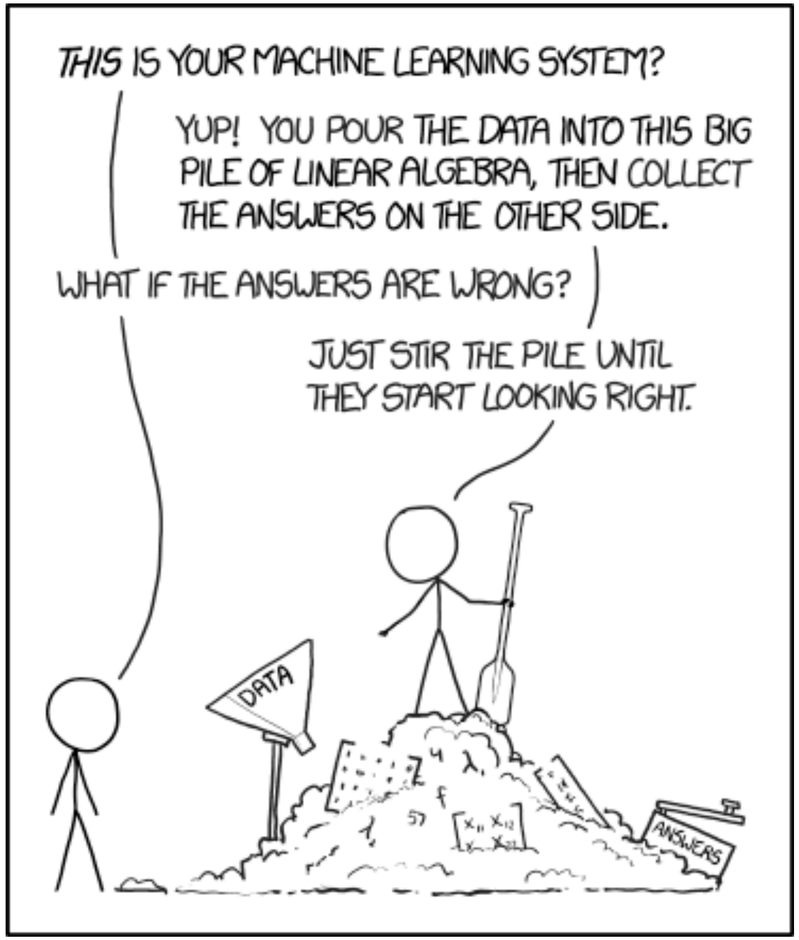
en.wikipedia.org/wiki/Precision_and_recall

Learning Algorithms

- Decision Tree (DT)
- Random Forests (RF)
- Artificial Neural Network (ANN)
- Support Vector Machine (SVM)
- Logistic Regression (LOGRES)
- Naïve Bayes (NB)
- K Nearest Neighbor (KNN)
- ...



Valuable lessons on machine learning



https://xkcd.com/1838/

Something you liked 😎👍

Something you think could be improved 😬👎