ASI DATA SCIENCE
STRATA LONDON 2017

Practical Machine Learning with Python

@ASIDataScience

Charlotte Werger
charlotte@asidatascience.com

23-05-2017

Meet your teacher

Charlotte Werger, PhD

- Now: Training at ASI
- Before: Quant hedge fund manager/researcher
- Econometrics, Machine Learning
- Data science in Finance

“45% of the activities individuals are paid to perform can be automated with currently available technologies. These activities represent about $2 trillion in annual wages”

McKinsey Report, Nov 2015

Black Box Artificial Intelligence

- Computer beats humans at Jeopardy (2011)
- Computer beats professional at Go (2015)
- Computer beats humans at poker (2017)

Artificial Intelligence for everyone

About ASI Data Science

PEOPLE
TRAINING

EXPERTISE
CONSULTING

TECHNOLOGY
SHERLOCKML

Data Science Training at ASI

Training by practitioners and experts

Emphasis on practical implementation

Designed for continued learning

Outline for today

- Intro to SherlockML & Machine learning
- Clustering with K-means
- Classification with K-NN

Meet your teacher

Andrew Brookes

- CTO of ASI
- Computer Scientist
- 10+ years professional experience in data engineering
- Expert in technology implementation and deployment
- Leads SherlockML team

SHERLOCKML

Sherlock has three functional layers

Templates & Support

Tool Set

Environment

Sherlock has three functional layers

Templates & Support

Tool Set

Environment

Secure

Collaborative

Auditable

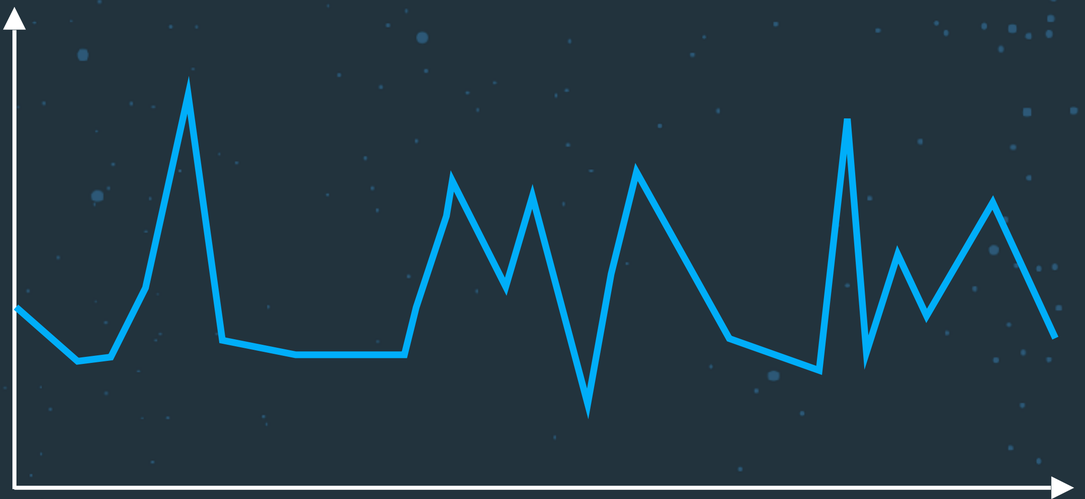
Secure, Scalable Compute

2017 to date

3500
CPUs

10 TB
Memory

1280
Servers

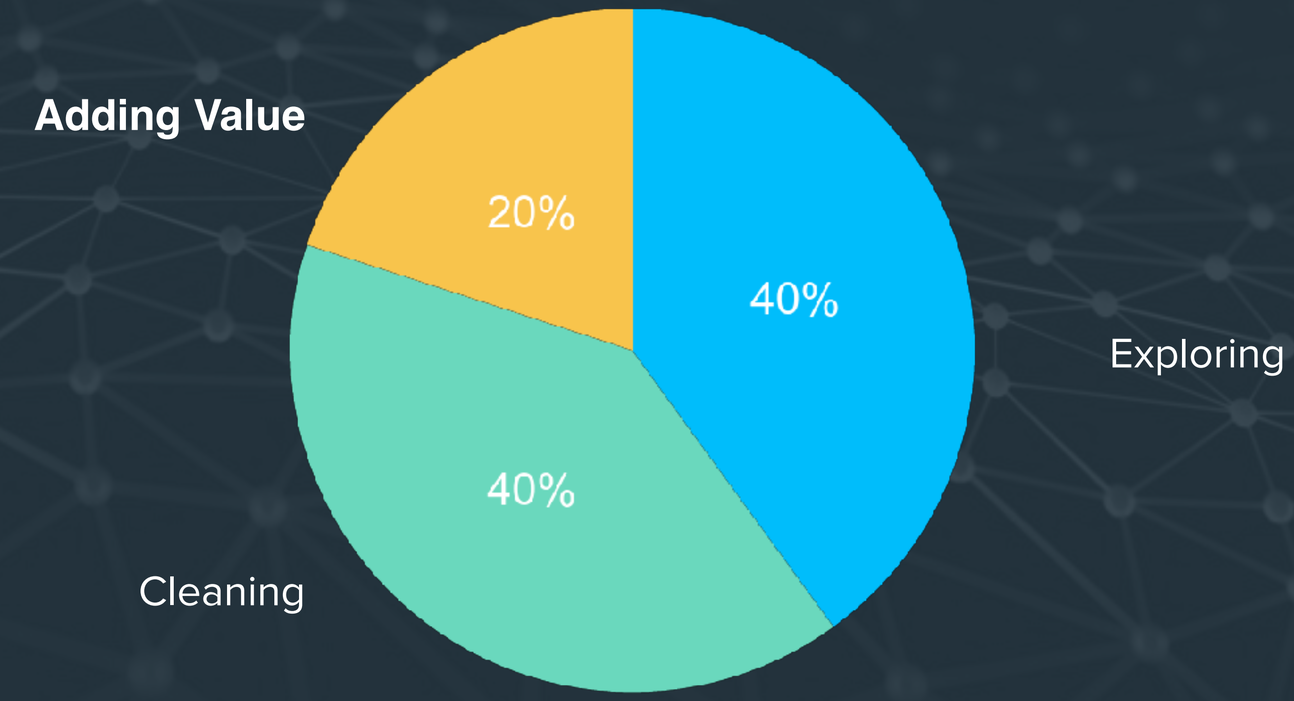


Sherlock has three functional layers

Templates & Support

Tool Set

Environment

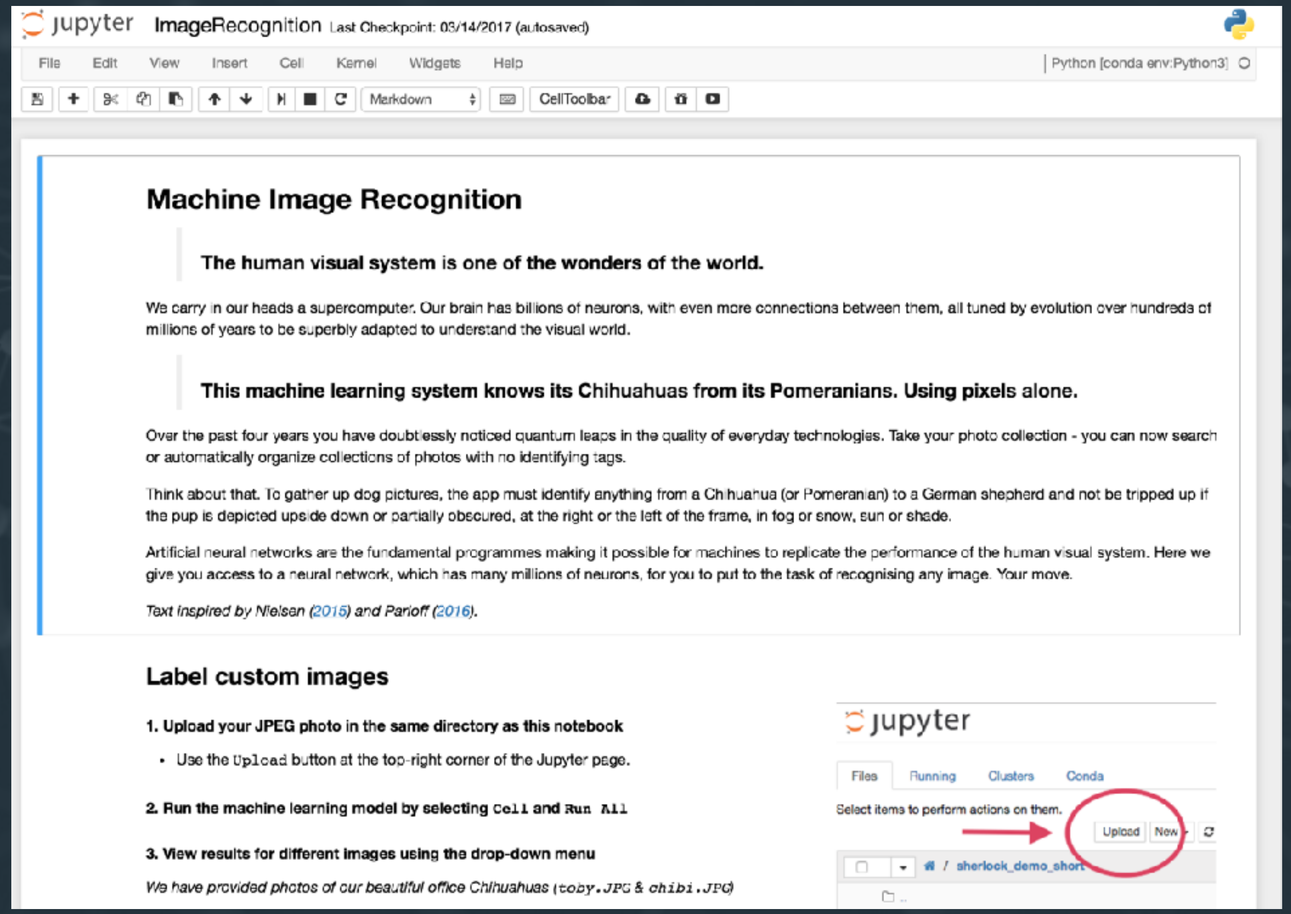


Sherlock has three functional layers

Templates & Support

Tool Set

Environment



Accountants » **Excel**

Designers » **Photoshop**

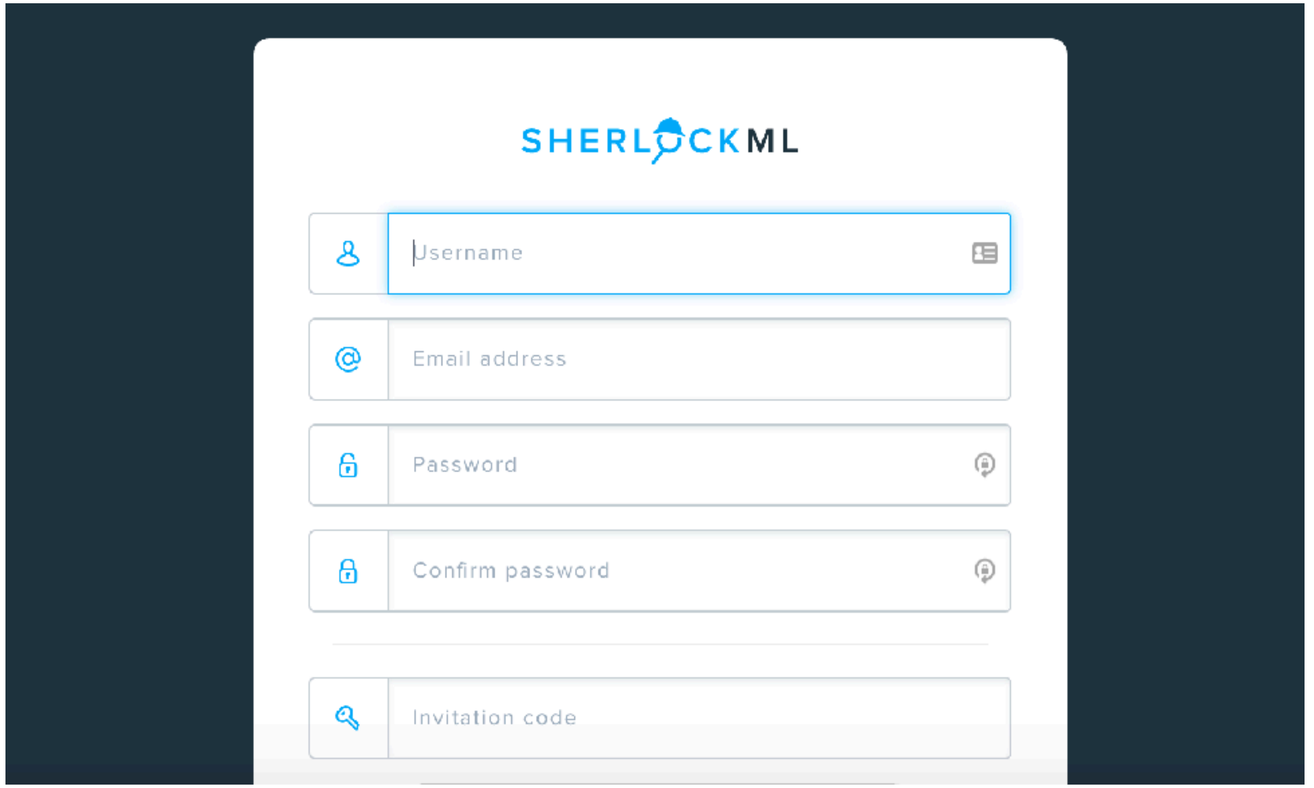
Data Scientists » **?**

SHERLOCKML

Powerful Simple Secure

Step 1

Login to SherlockML or sign up if you haven't done so yet

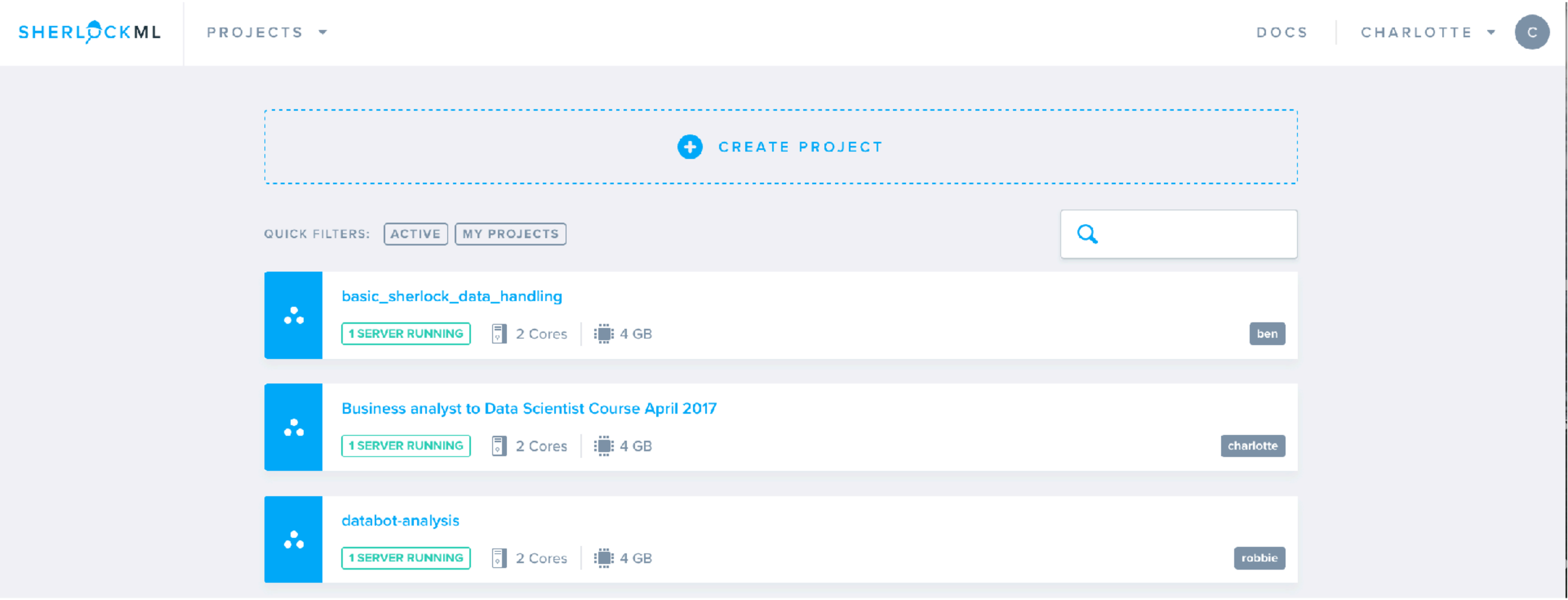


Go to **https://sherlockml.com**

And use invite code **Strata2017**

Step 2

Click on your “My sandbox” project



Step 3

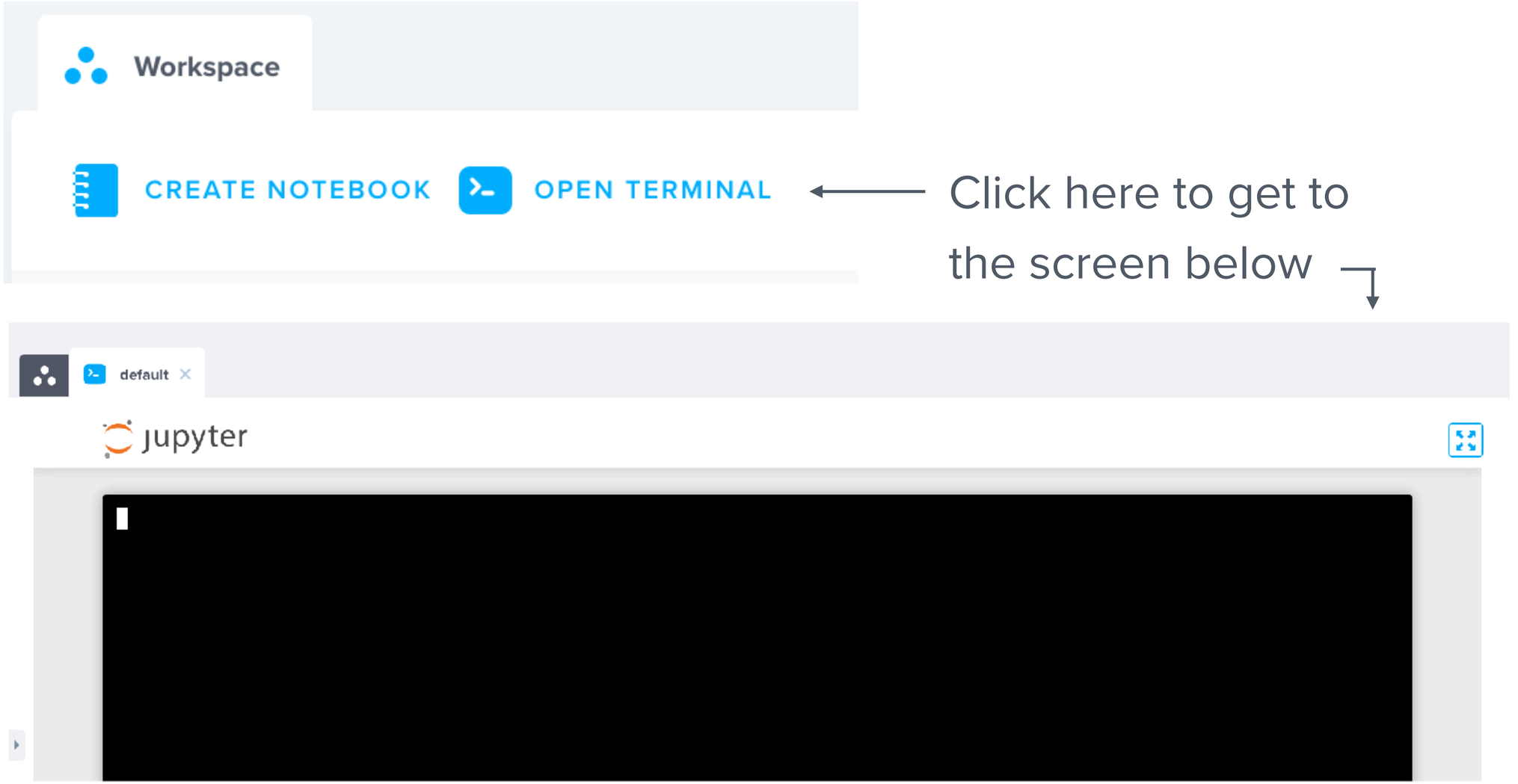
Wait for a server to be created

Click the refresh button to check if it’s ready.

Once it looks like this, you’re ready to start.

Step 4

Click on the “Open Terminal” tab



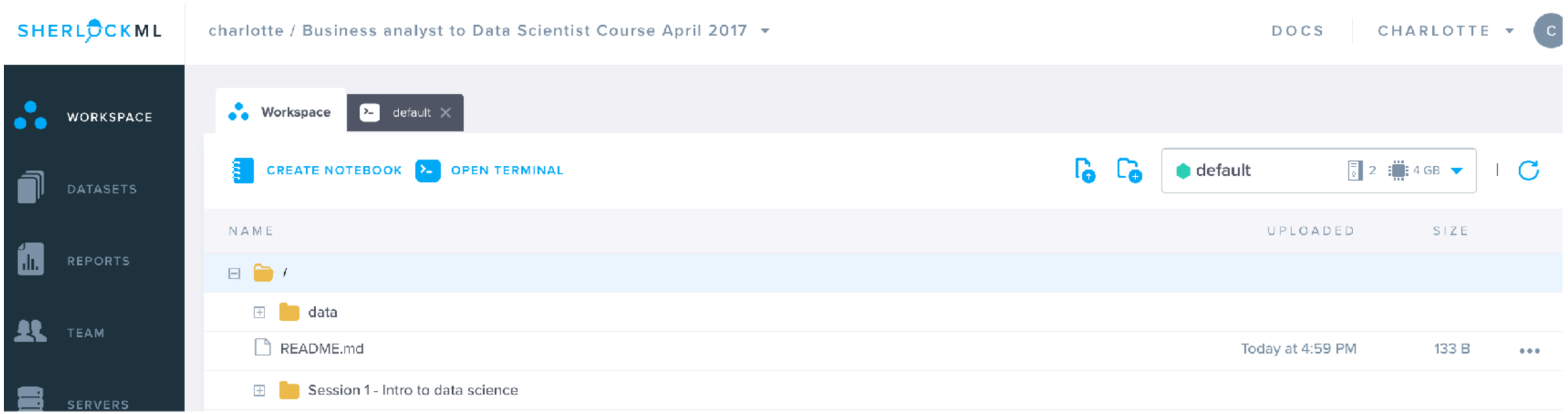
Step 5

Write a command to copy the training materials into your workspace

Write: git clone https://github.com/ASIDataScience/stratapracticalmachinelearning.git

Step 6

Go back to your workspace and refresh



Files should be there: you're good to go

Machine Learning Models

Machine Learning Models

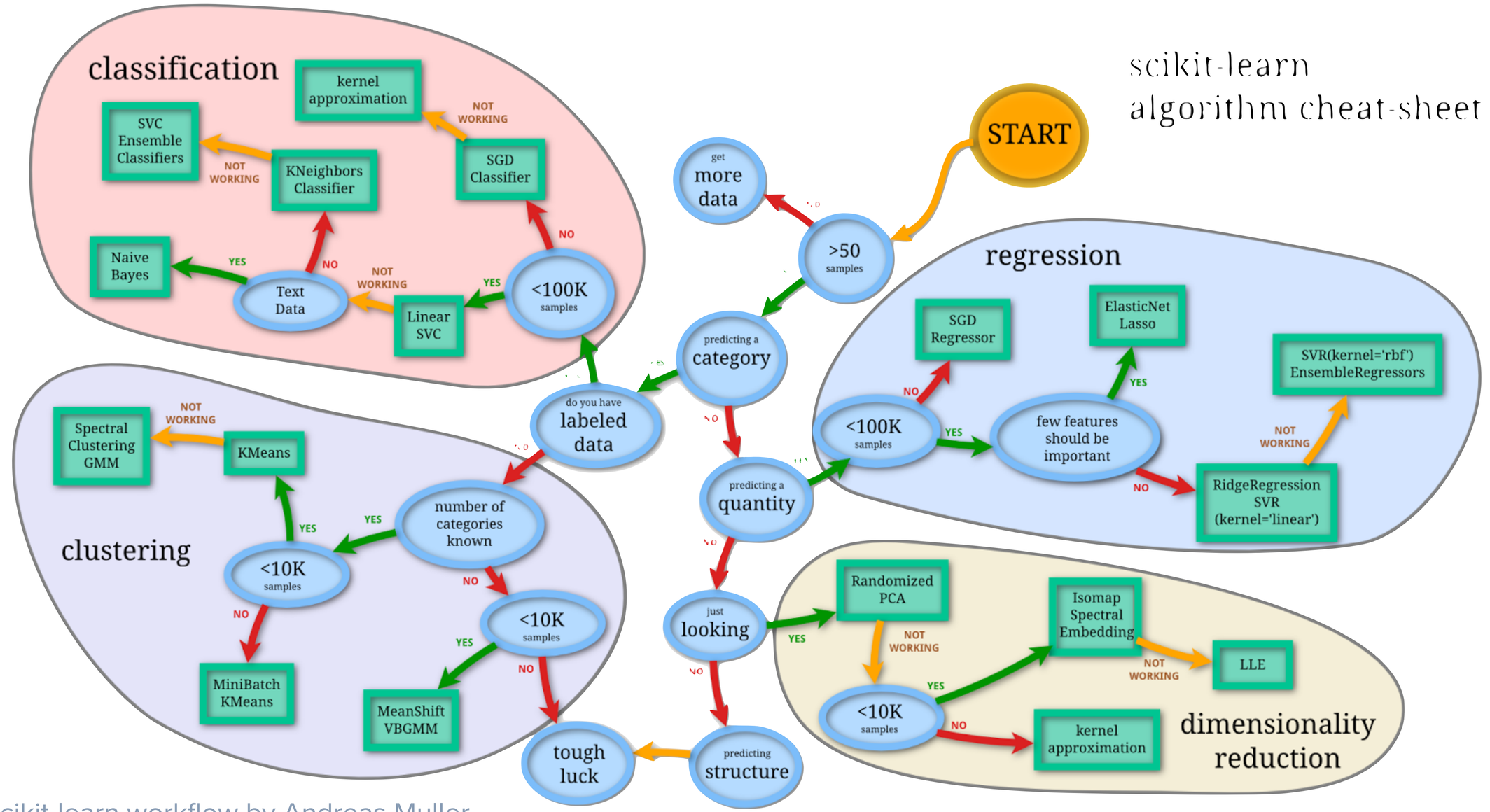
SUPERVISED

- Access to target variable
- e.g. linear regression or classification

UNSUPERVISED

- No access to target variable
- e.g. clustering or dimensionality reduction

Machine Learning Models



Scikit-learn workflow by Andreas Muller

Clustering with K-means

Clustering

- Overview
- Example data sets
- Applications
- K-means clustering
- Other clustering algorithms

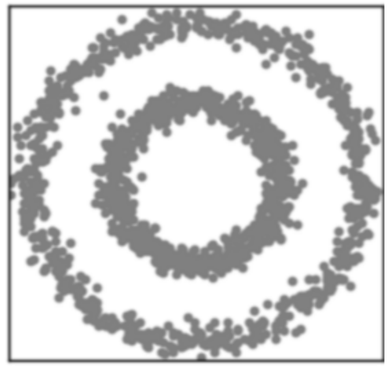
Clustering

- Unsupervised learning task
- Given unlabelled data
- Group data into “clusters” of similar data points

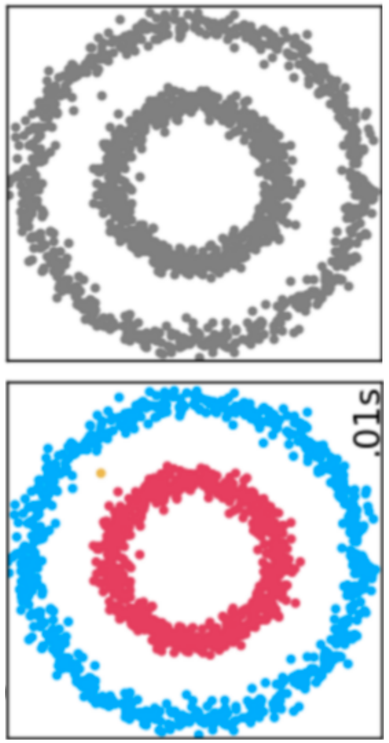
Clustering

- Aim to find patterns in the data
- Not trying to make a correct prediction/classification
- Important component of exploratory data analysis
- Hard to evaluate

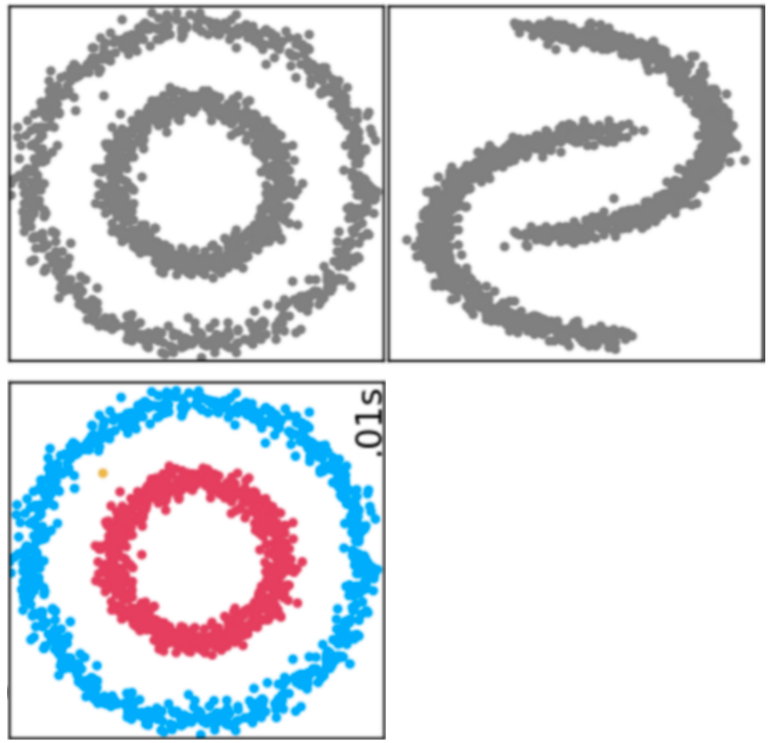
Example datasets



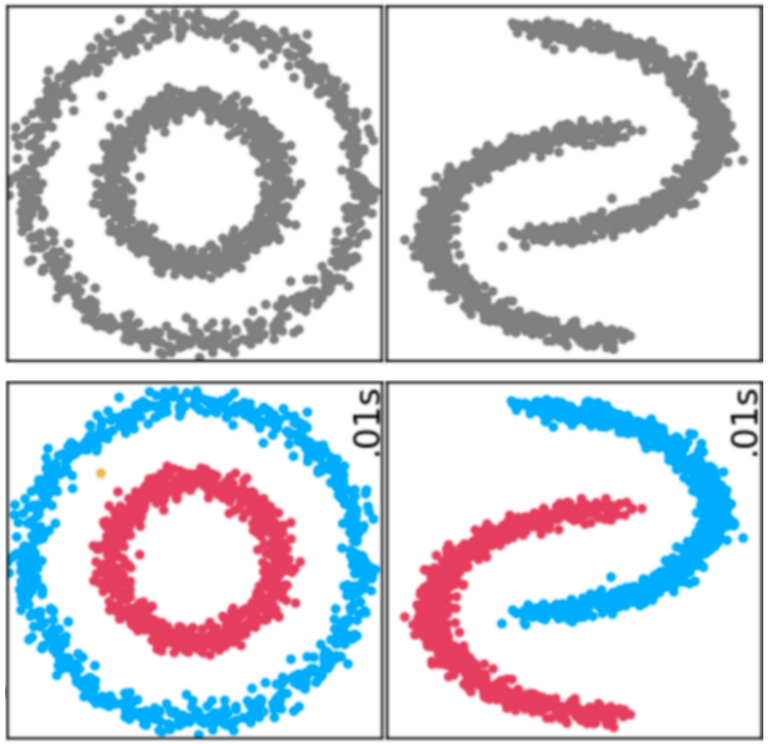
Example datasets



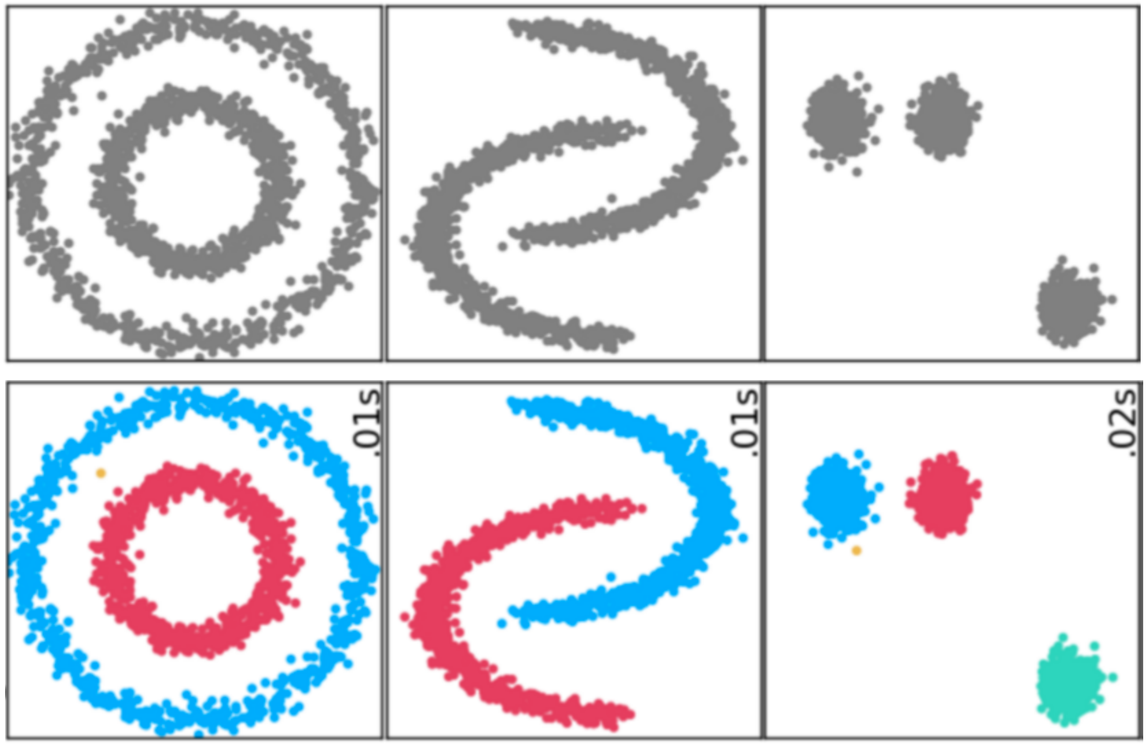
Example datasets



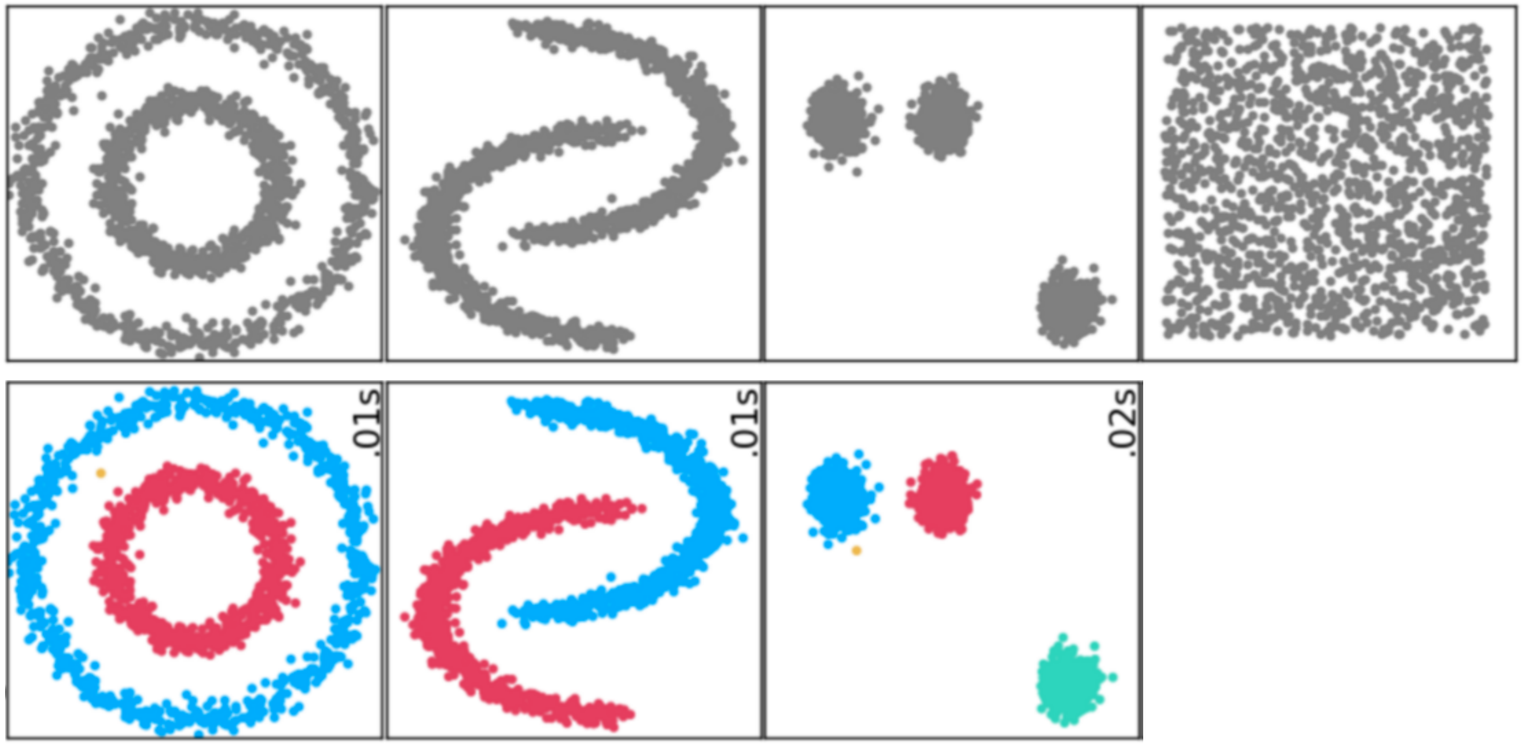
Example datasets



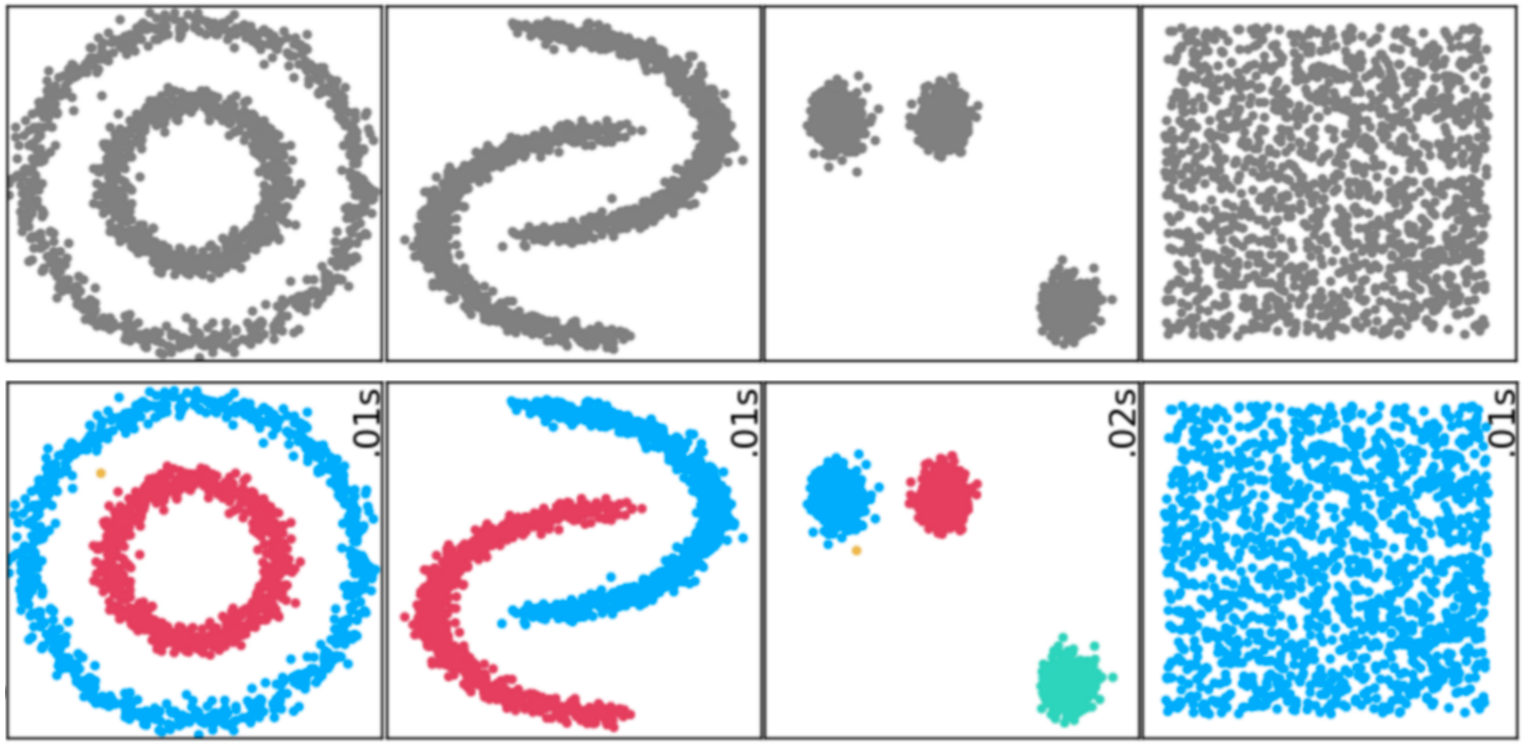
Example datasets



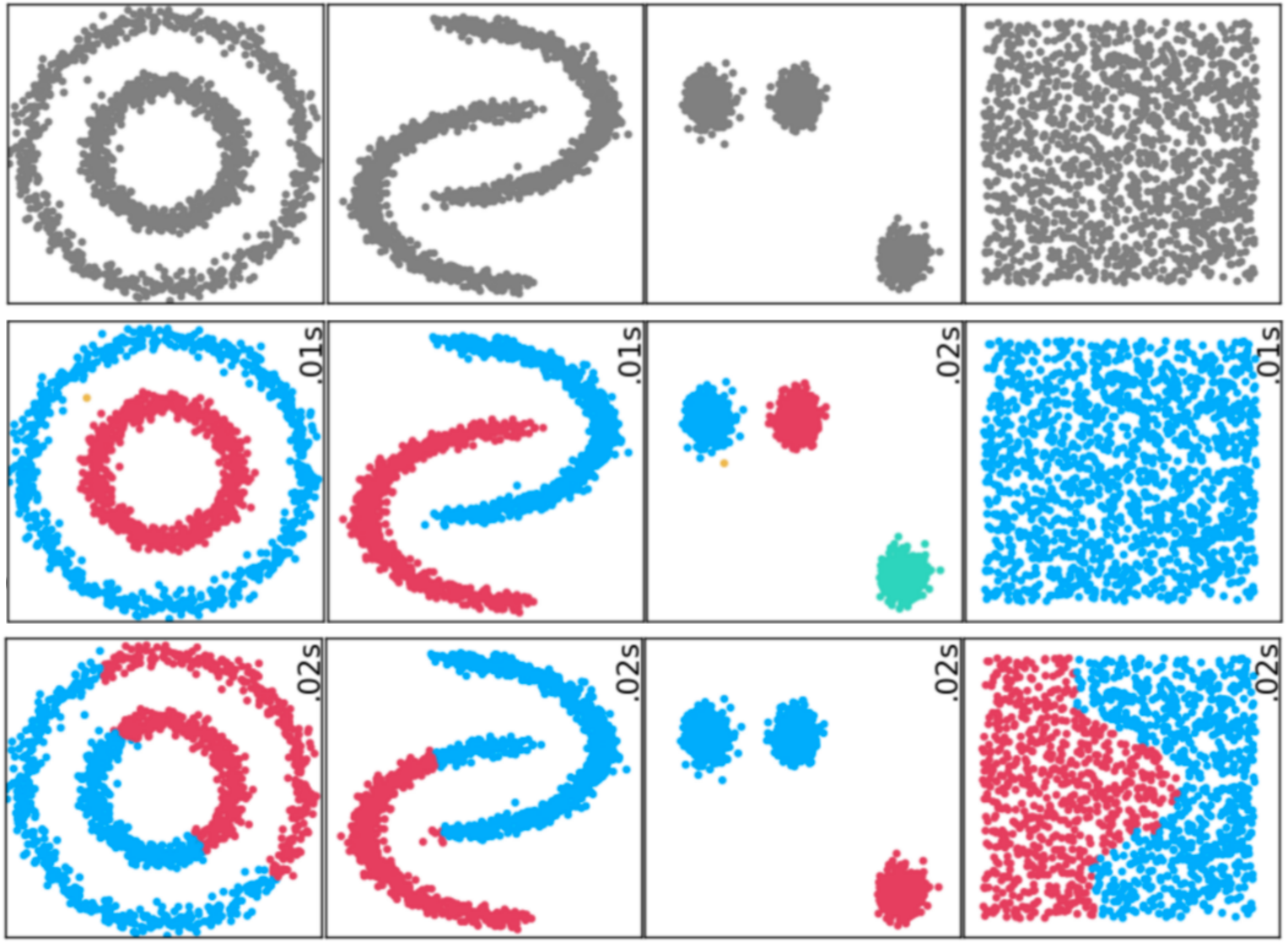
Example datasets



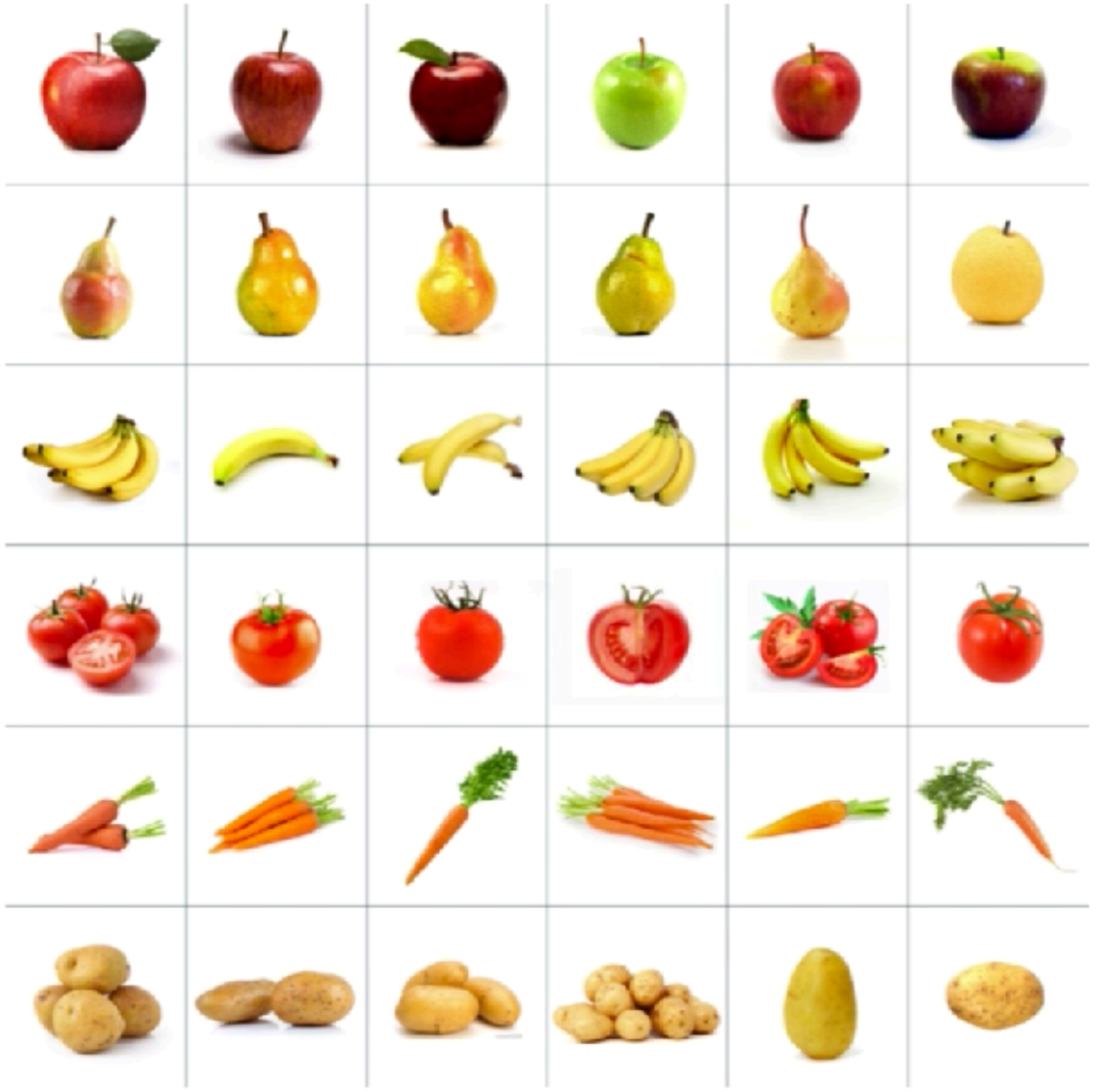
Example datasets



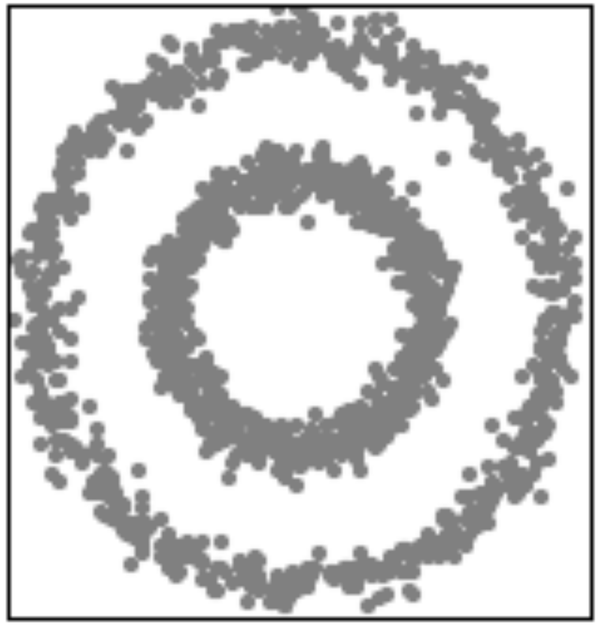
Example datasets

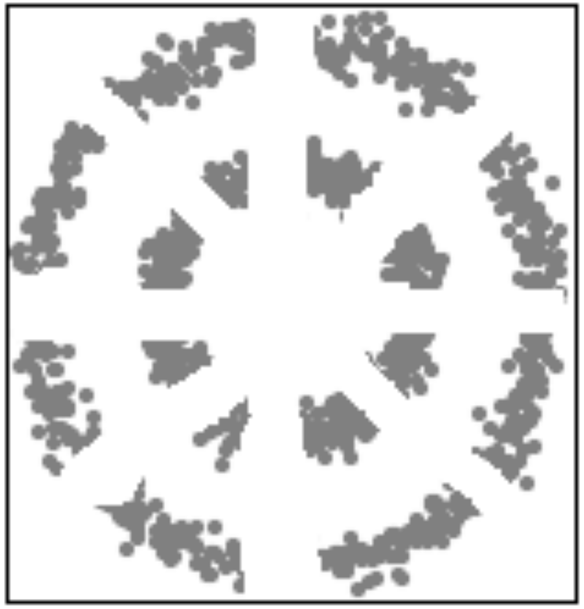


Example datasets



Example datasets





Applications

- **Marketing:** Find groups of customers with similar buying behaviour Biology: Clustering of genes that have similar functions
- **Media:** Summarise news stories.
- **Image segmentation:** Digital artists can edit individual objects and neuroscientists can find the boundary between tumor and healthy tissue
- **Law enforcement:** Detect dangerous crime zones.
- **Information retrieval:** Topic discovery, document and web page clustering for recommender systems e.g. Amazon book recommendation, refined search engine queries.

K-means clustering

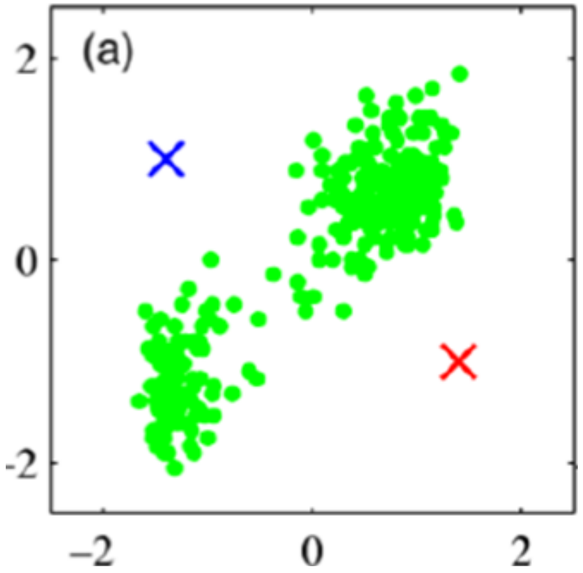
Given a set of observations:

$$X = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n), \mathbf{x}_i \in \Re^m \forall i$$

k-means clustering will attempt to partition the n observations into k disjoint subsets (or clusters):

$$\boldsymbol{C}_1, \boldsymbol{C}_2, ..., \boldsymbol{C}_k$$

where k is a parameter provided to the algorithm



Objective function = $$\sum_{i=1}^{k} \sum_{x \in C_i} \|\mathbf{x} - \boldsymbol{\mu_i}\|^2$$

where $\boldsymbol{\mu}_i$ is the centroid of cluster $\boldsymbol{C}_i$ defined as $$\boldsymbol{\mu}_i = \frac{1}{|\boldsymbol{C}_i|} \sum_{x \in C_i} x$$

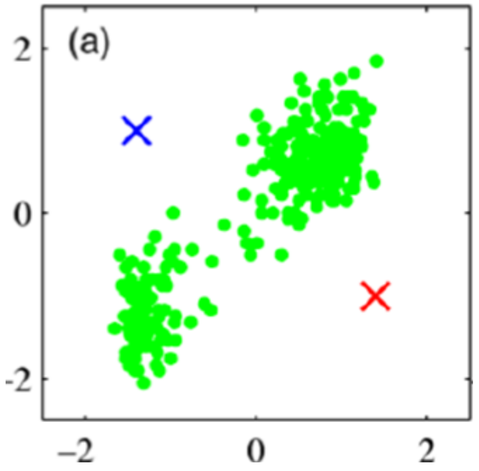
The algorithm's steps

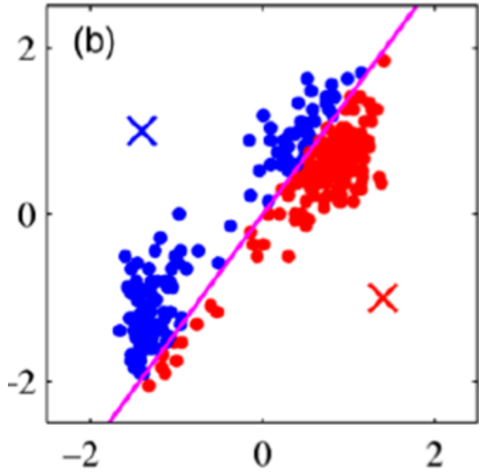
1. Define the number of clusters, k, (e.g. k = 2) and assign random points as the initial centroids.

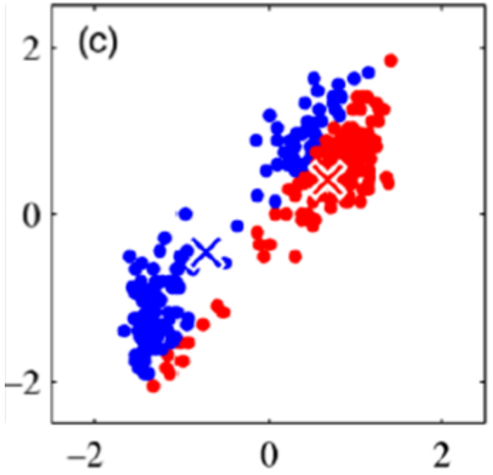
2. Assign each data point to its nearest centroid.

3. Determine the new centroid for each cluster.

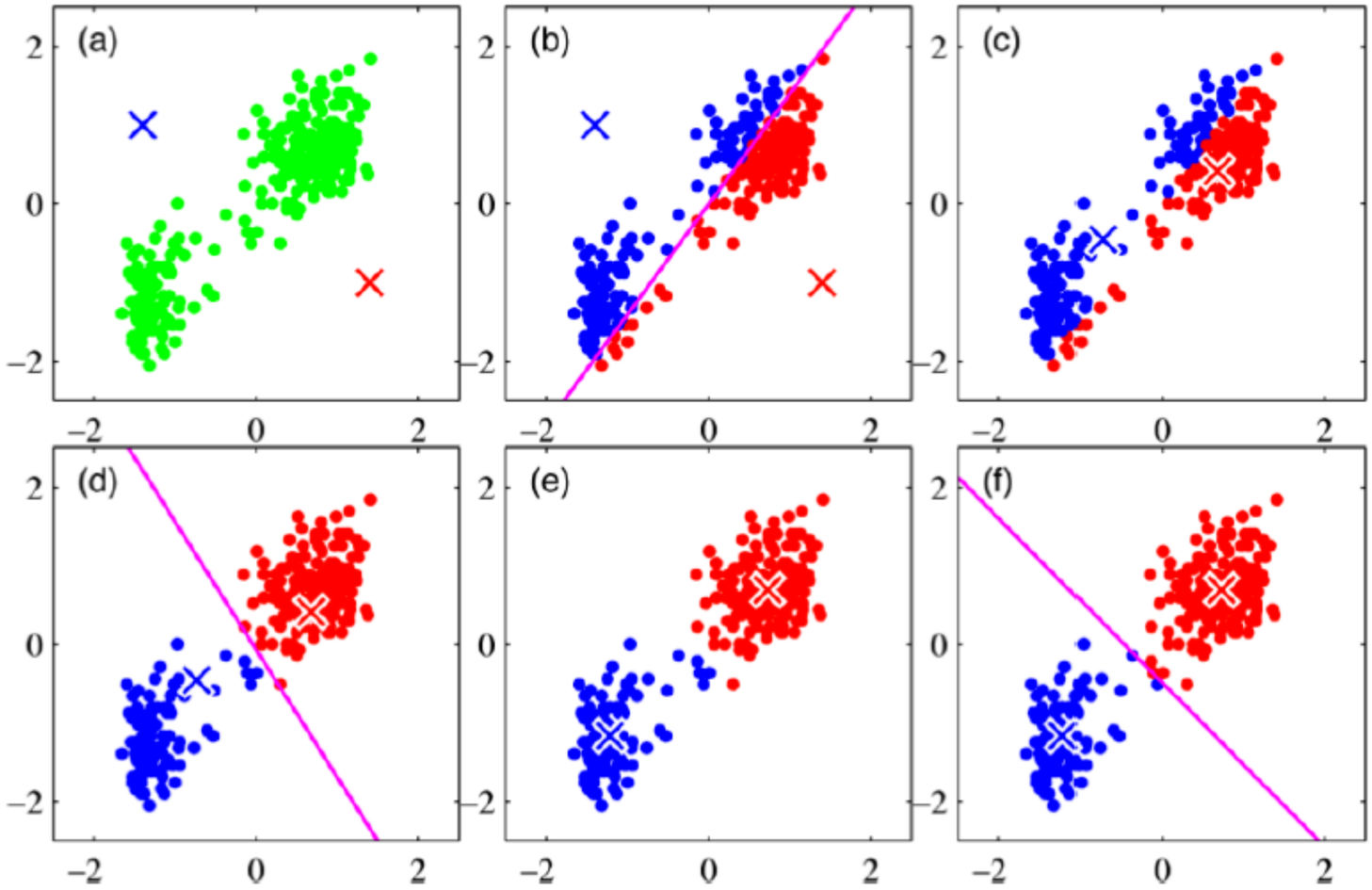
Iterate until convergence of objective function:







Step by step in 2D space



K-mens Distance Functions

Objective function is a sum of squared Euclidean distances

$$\sum_{i=1}^{k} \sum_{x \in C_i} \|\mathbf{x} - \boldsymbol{\mu_i}\|^2$$

Generalise to use an arbitrary distance function d

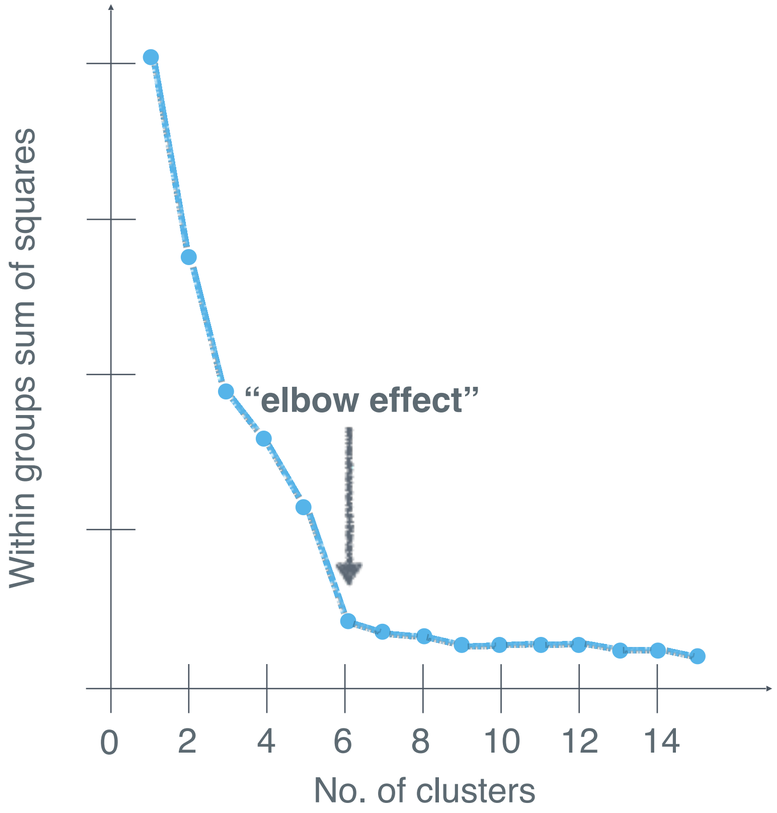
$$\sum_{i=1}^{k} \sum_{x \in C_i} d(\mathbf{x}, \boldsymbol{\mu_i})$$

How do we select the number of clusters?

- “By eye” (simple cases)
- Elbow method

Potential problems:

- More than one elbow
- No elbow



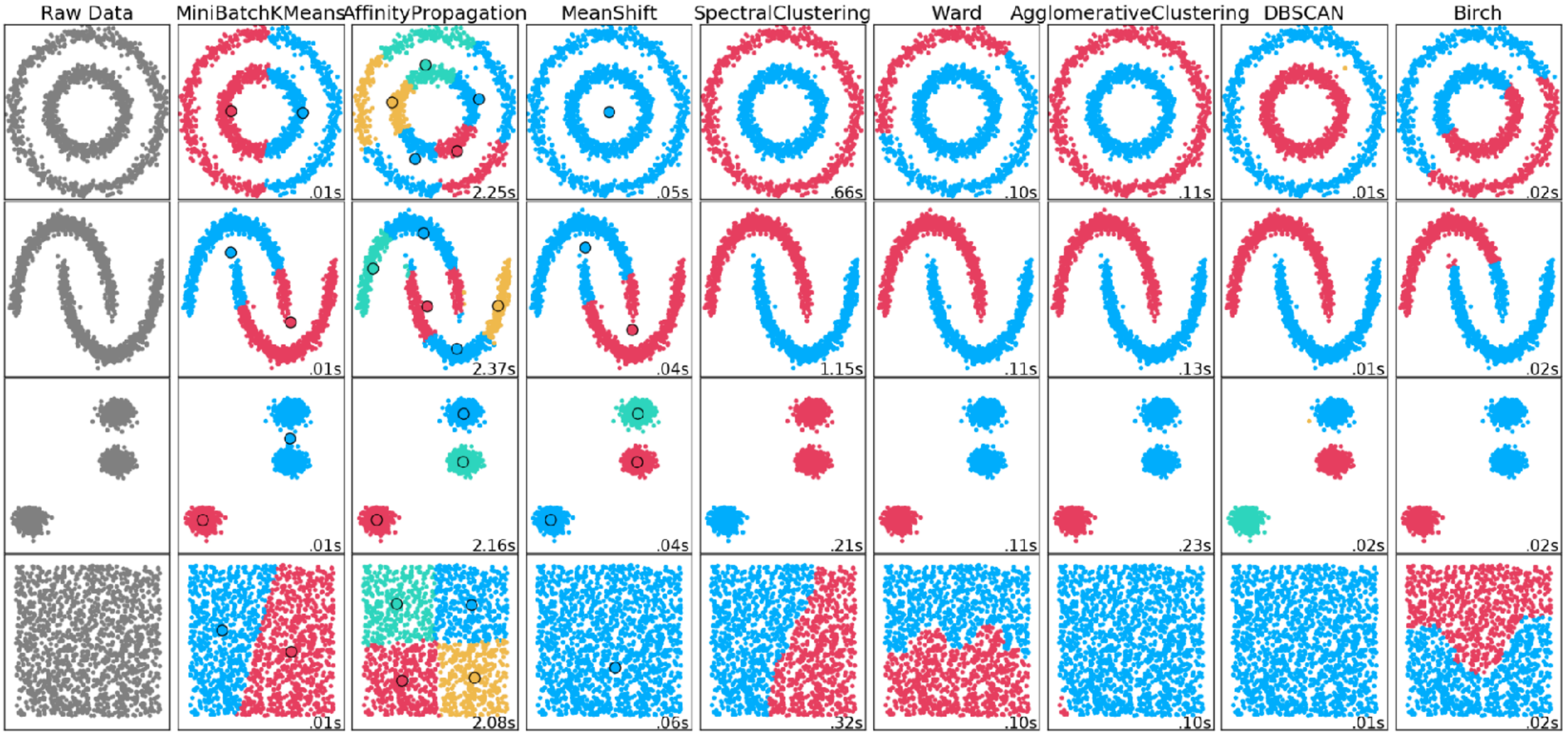
K-means drawbacks

- K-Means
  - Problem: sensitive to initial partition
  - Can’t handle categorical data
  - Number of clusters (k) must be set a priori
  - Can’t deal with weirdly shaped clusters
- Solutions: Use a different type of clustering algorithm or a different distance function

Types of Clustering Algorithms

- Hierarchical clustering
  - Agglomerative
  - Divisive
- Partitional
  - K-means and K-medoids
  - Spectral clustering (K-means on spectrum of distance matrix)
  - Density-based clustering (e.g. DBSCAN)
- Distribution-based clustering: (e.g. EM for Mixture of Gaussians)
  - Can also provide a fuzzy clustering in addition to both hard clustering
  - Various distances between points (Euclidean, Mahalanobis, Minkowski, etc.)
  - Various distances between clusters (Single-linkage, complete-linkage, etc.)

Performance of various algorithms



EXERCISE: K-MEANS CLUSTERING

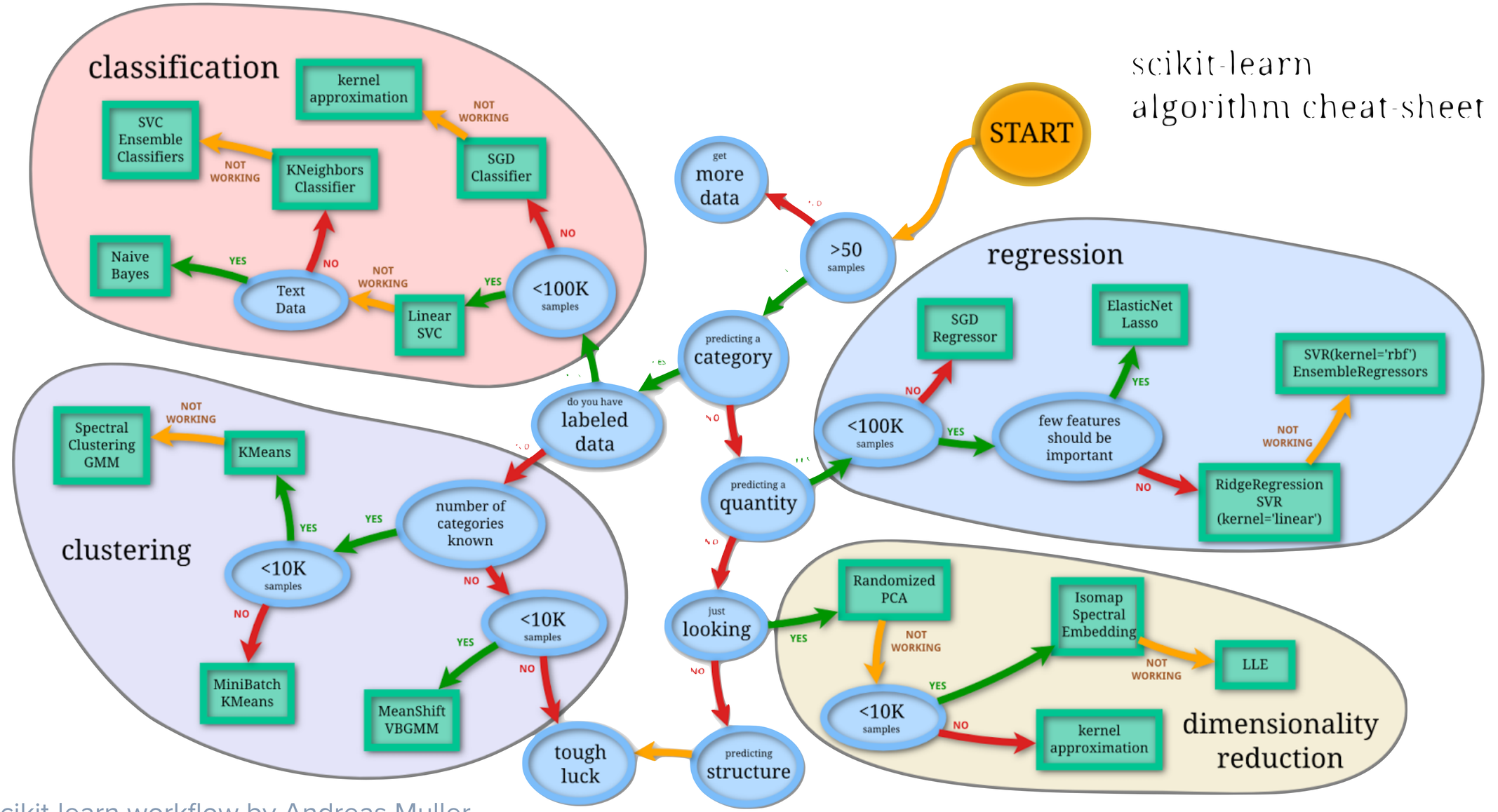
Classification with K-Nearest Neighbours

Meet your teacher

Alessandra Staglianò, PhD

- Senior Data Scientist at ASI
- PhD in Computer Science
- Expert in Machine Learning & Machine Vision
- Experienced working with unstructured data
- Natural Language Processing

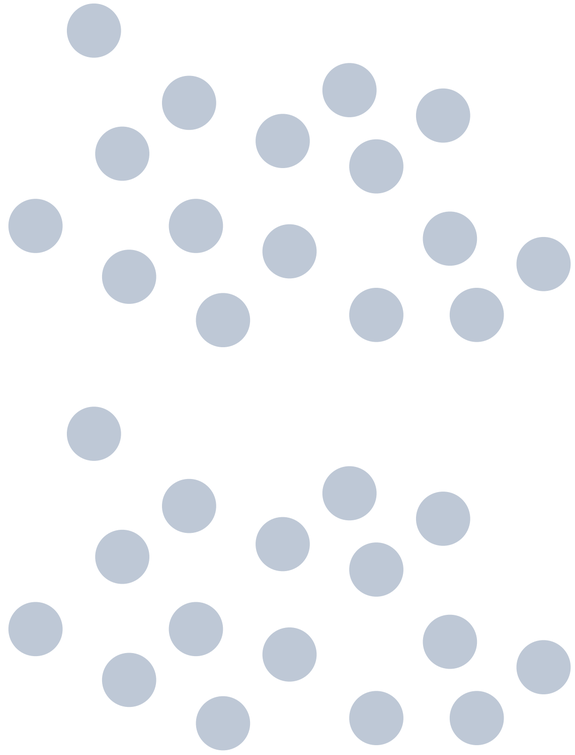
Machine Learning Models



Scikit-learn workflow by Andreas Muller

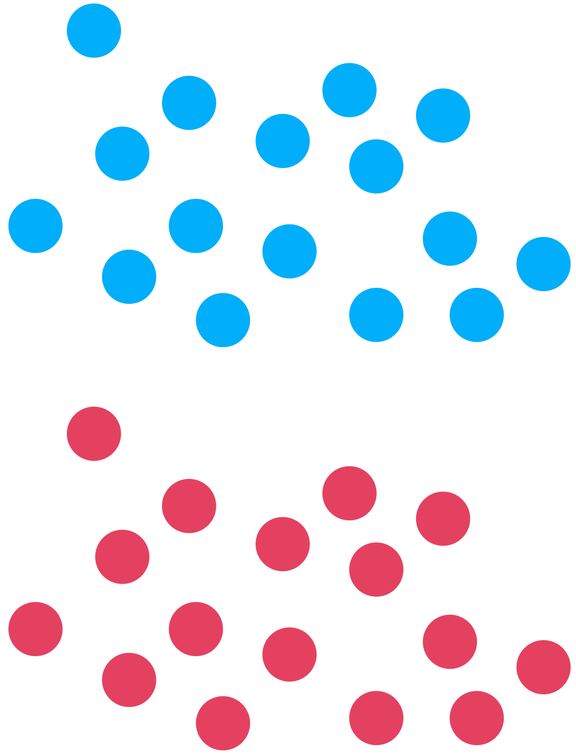
Classification

TRAINING SET

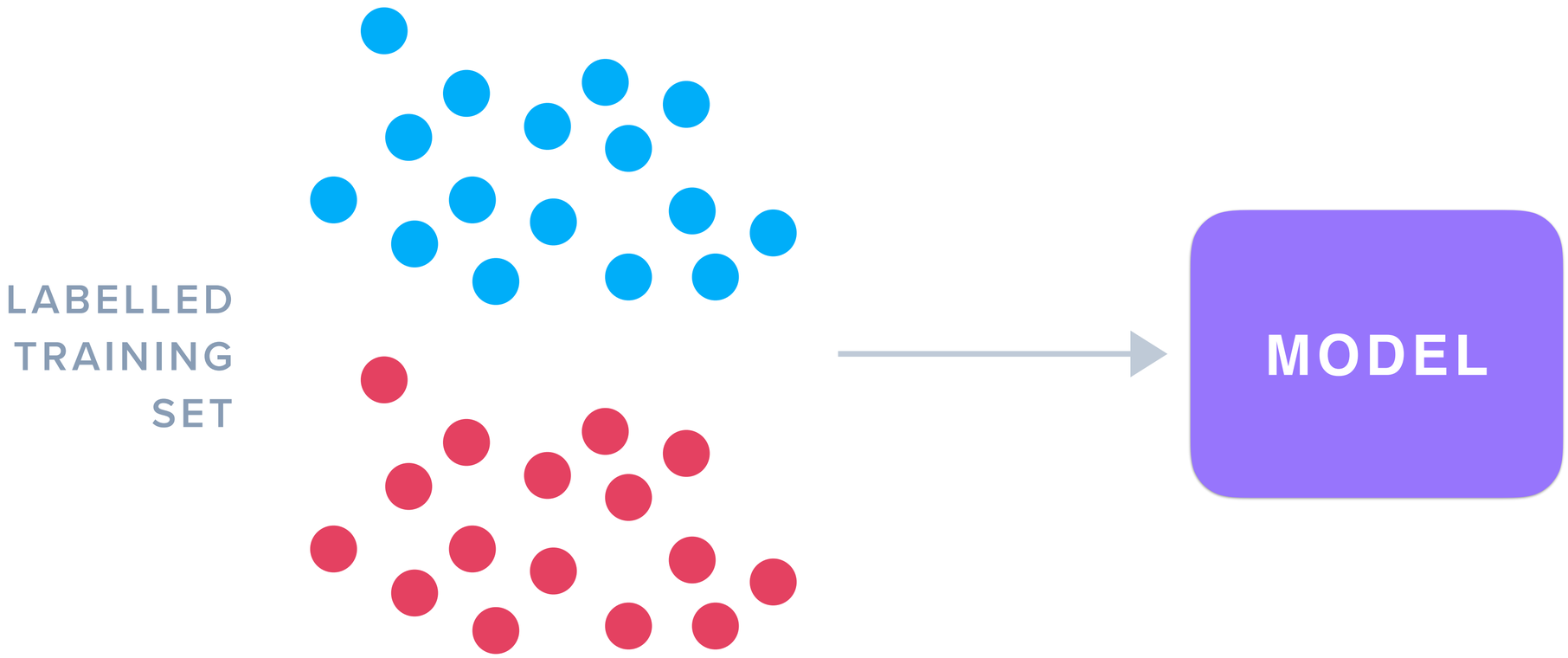


Classification

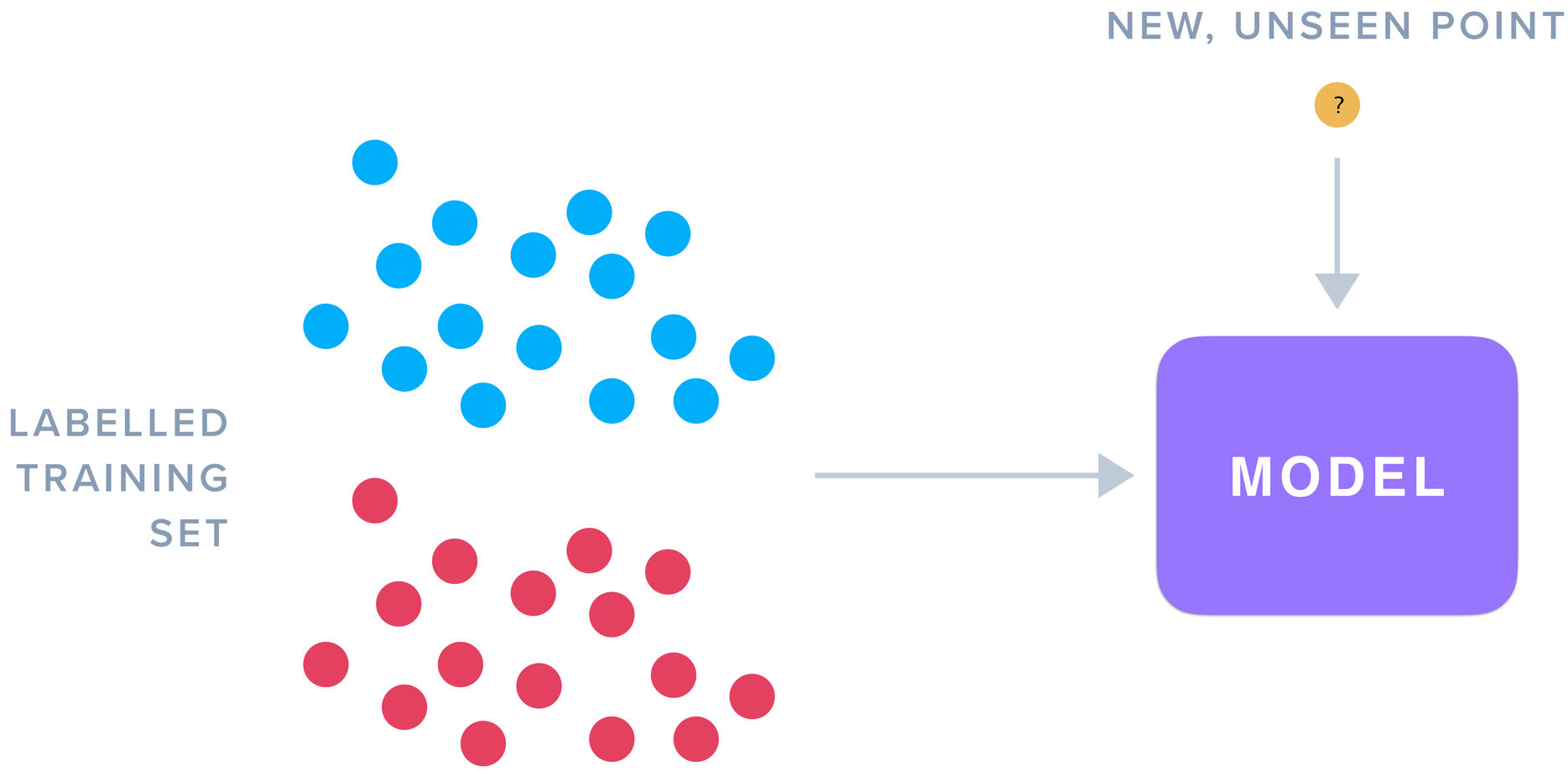
LABELLED TRAINING SET



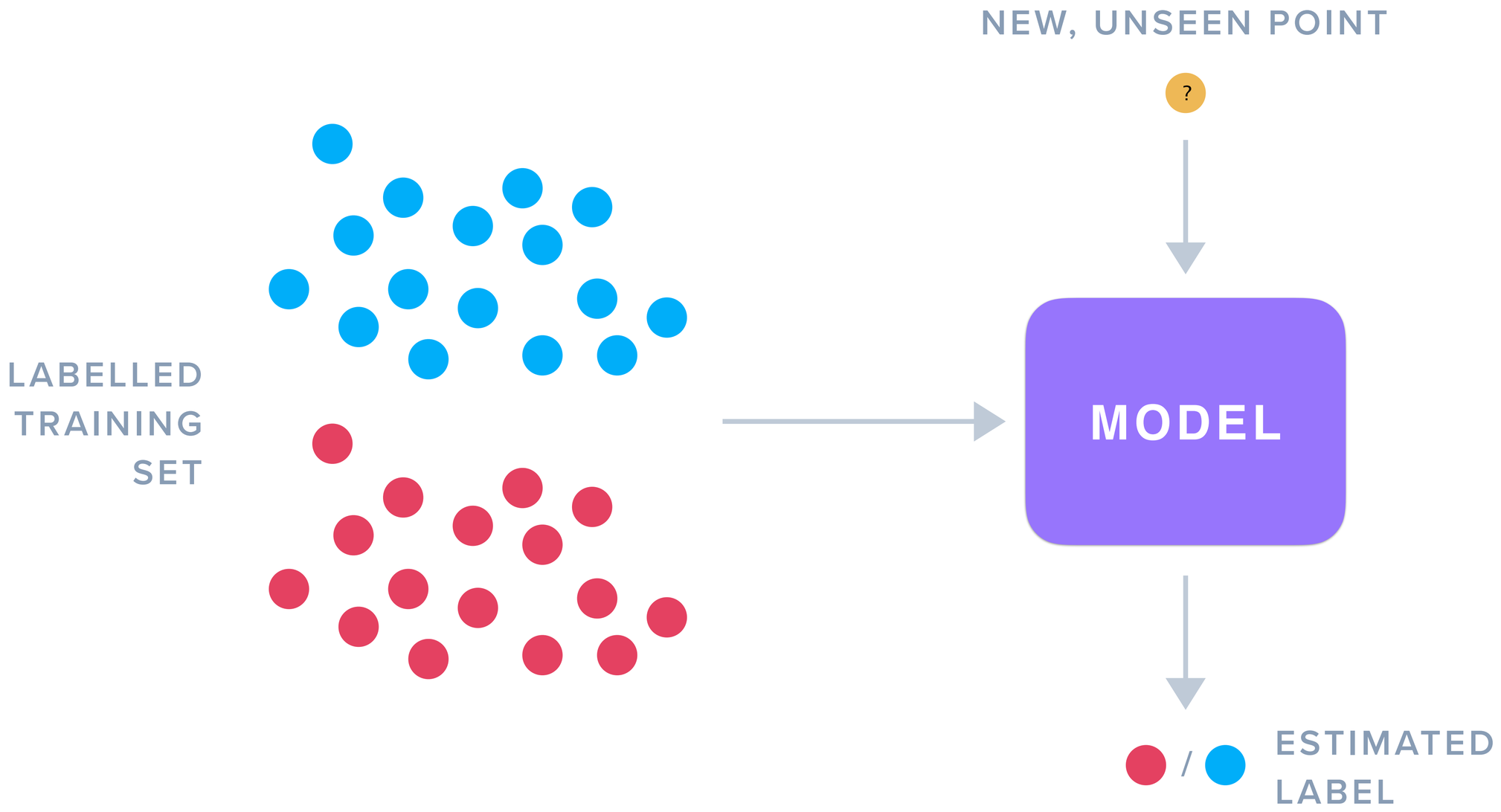
Classification



Classification

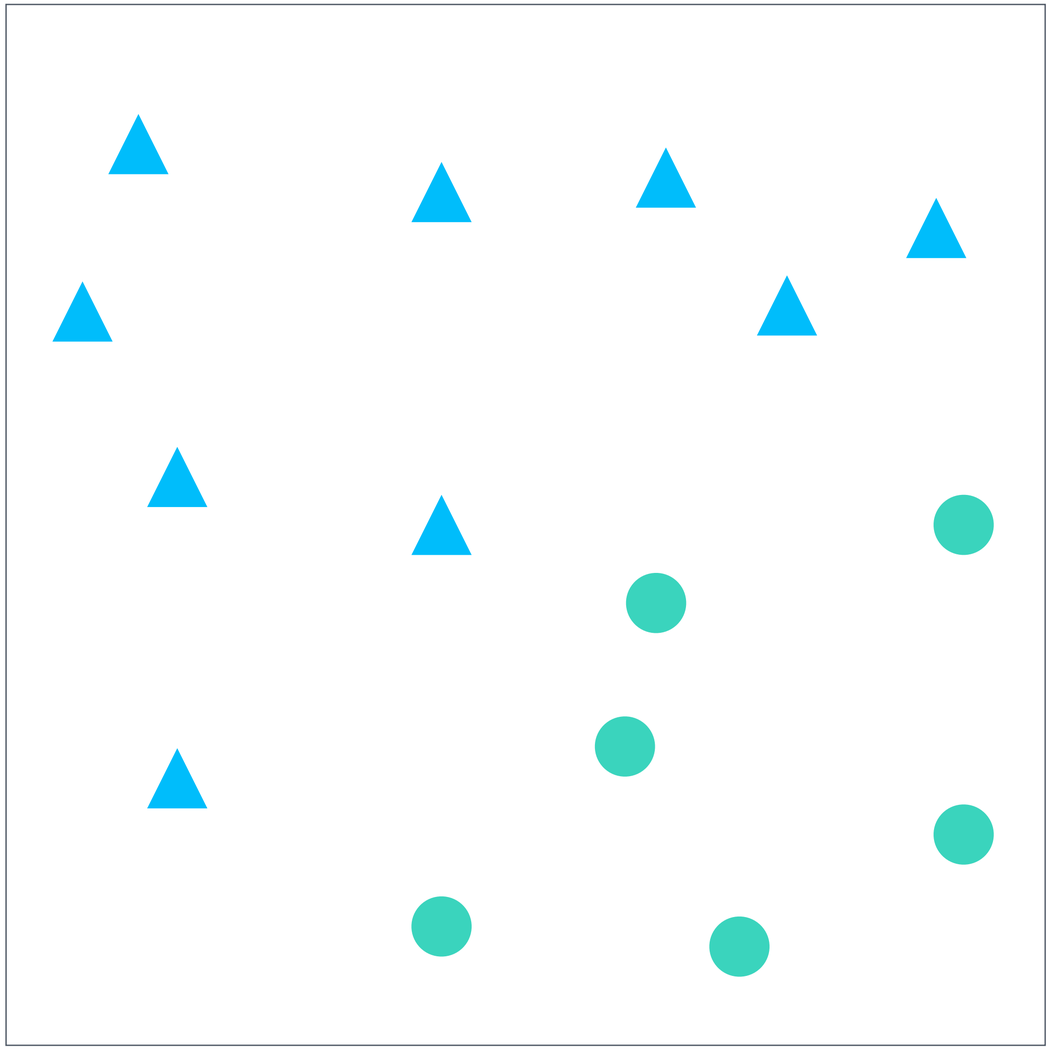


Classification



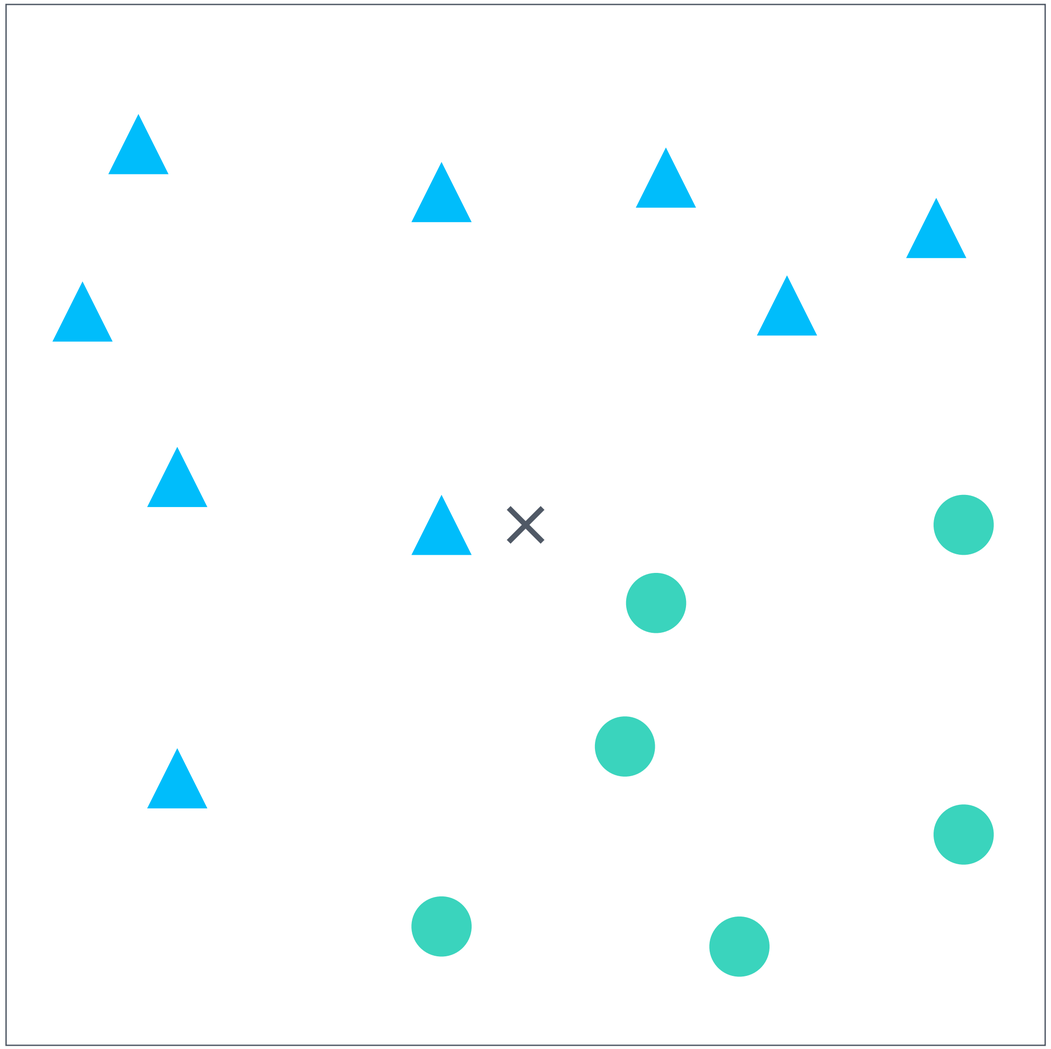
K-Nearest Neighbours

K-Nearest Neighbours (KNN)



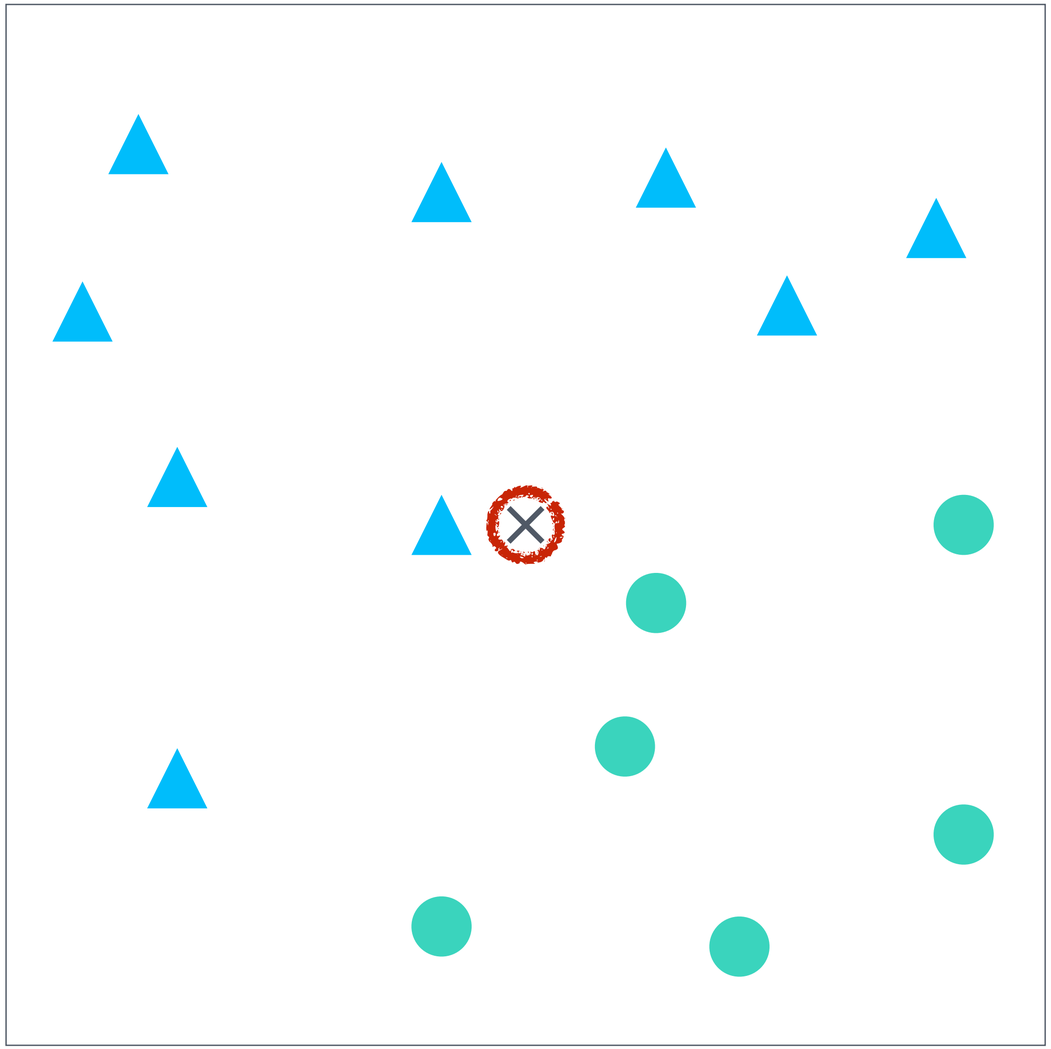
K-Nearest Neighbours

K-Nearest Neighbours (KNN)



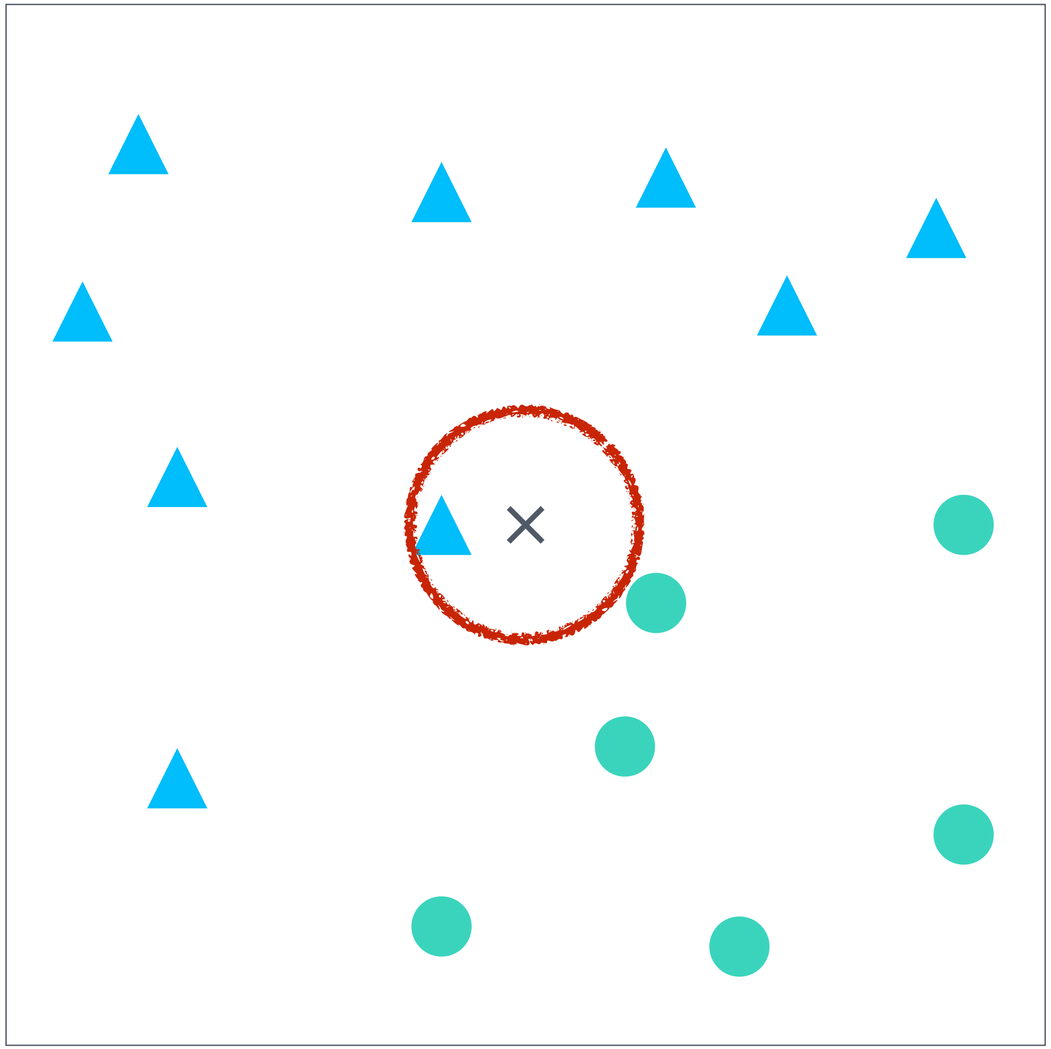
K-Nearest Neighbours

K-Nearest Neighbours (KNN)



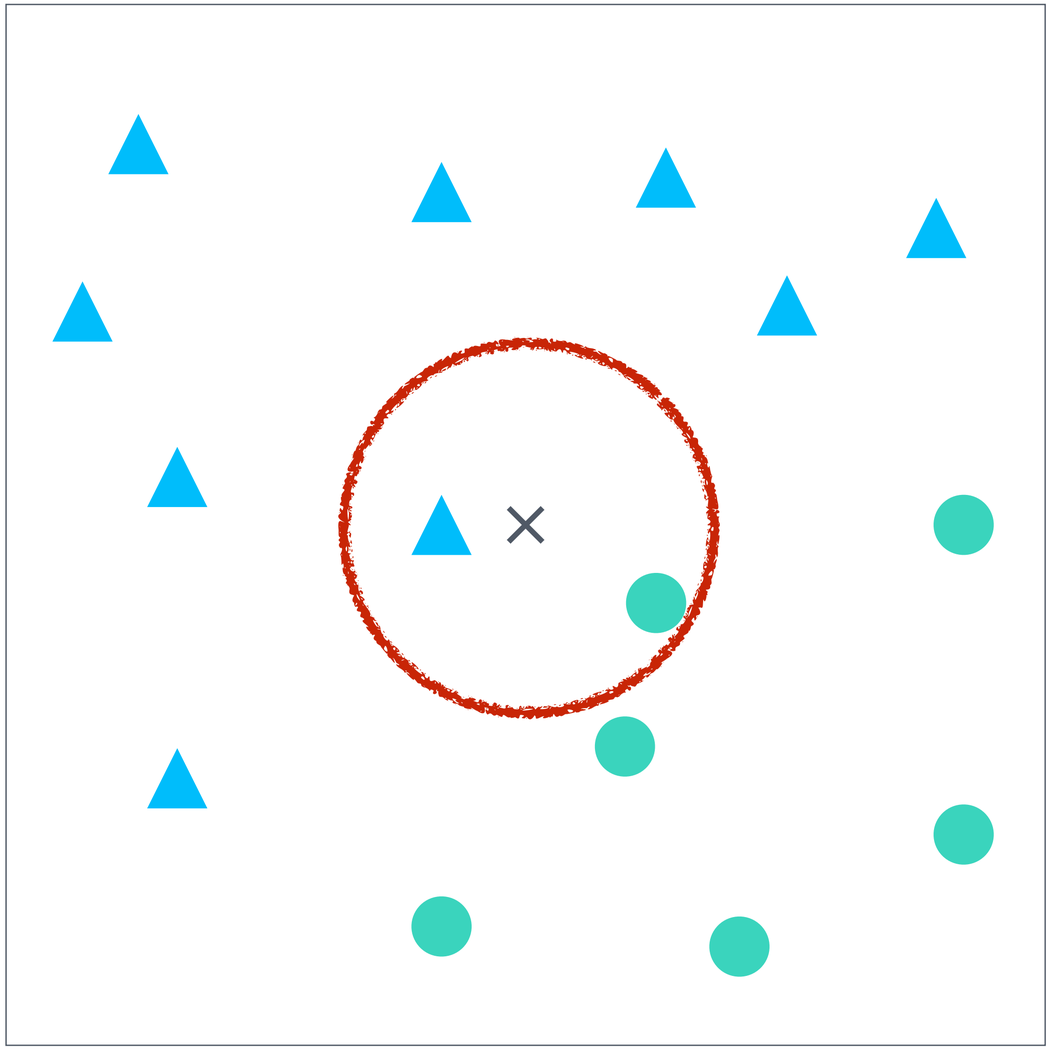
K-Nearest Neighbours

K = 1



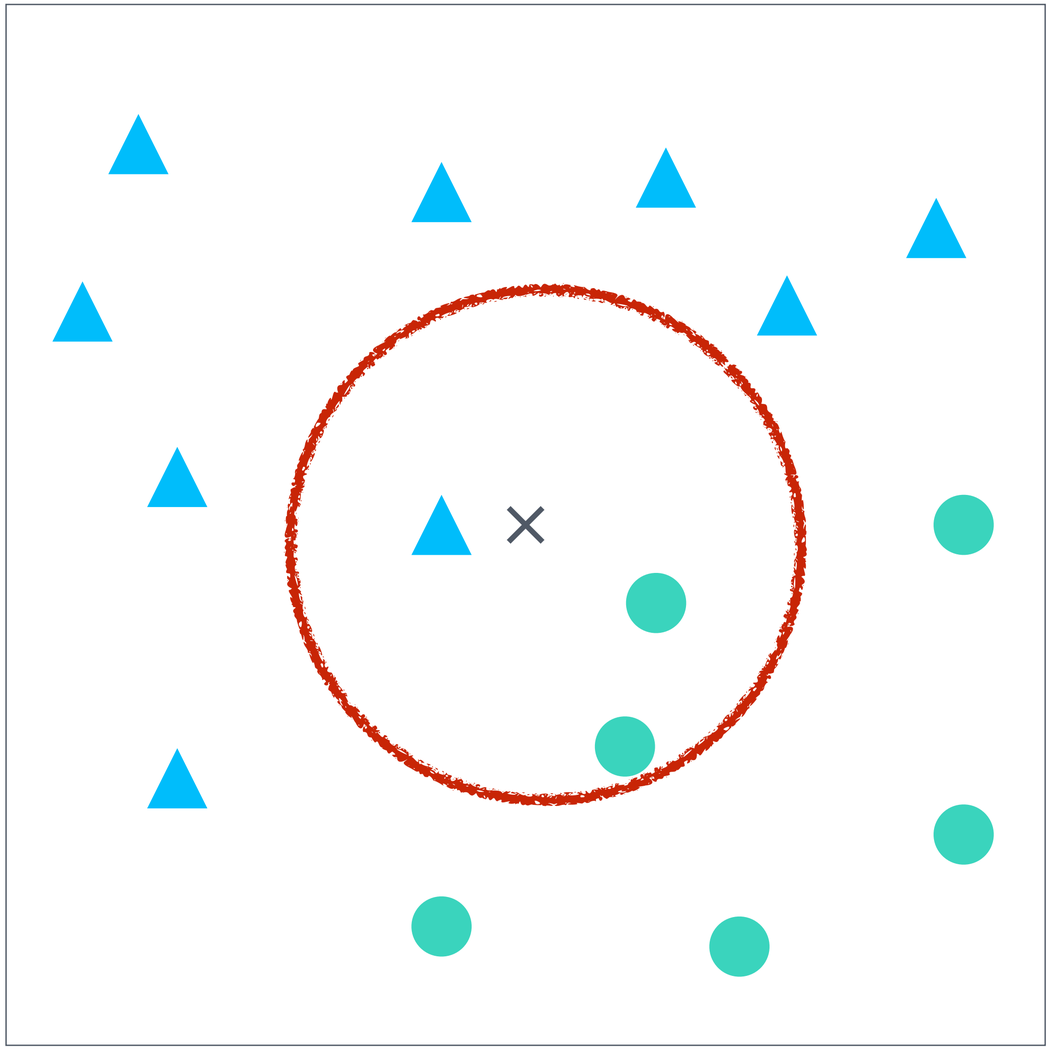
K-Nearest Neighbours

K = 2



K-Nearest Neighbours

K = 3



Characteristics

- Non-parametric method for classification and regression
- Lazy learning: generalisation beyond the training data is delayed until a query is made to the system
- Instance-based model: no actual model is built

Ingredients

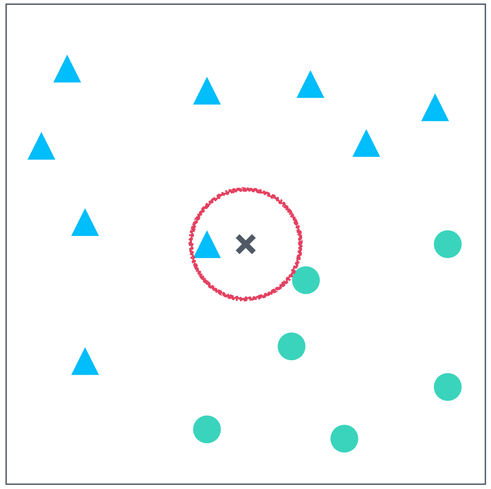
- The input training data
- The number k of nearest neighbours
- A metric to compute the distance between instances

Algorithm's Steps

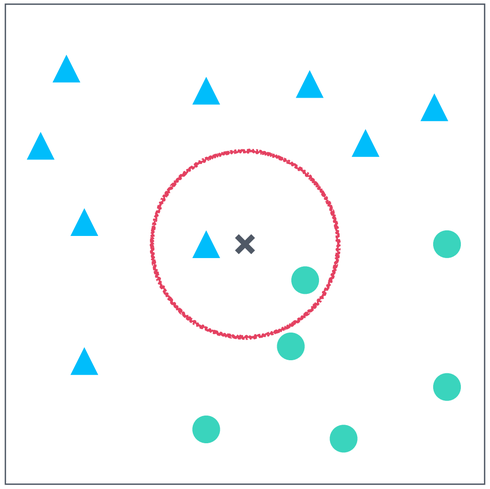
Classification of a new unknown, unlabelled test example:

- Compute distance to **all** training instances
- Identify K nearest neighbours
- Determine the classes of these nearest neighbours
- Take the majority vote of class labels among the K-nearest neighbours

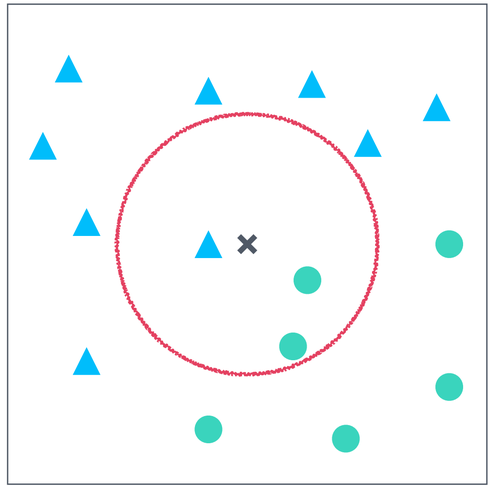
K-NN: 1-NN, 2-NN, 3-NN



1-nearest neighbour



2-nearest neighbour



3-nearest neighbour

k=1	▲
k=2	?
k=3	●

Which K?

A couple of details:

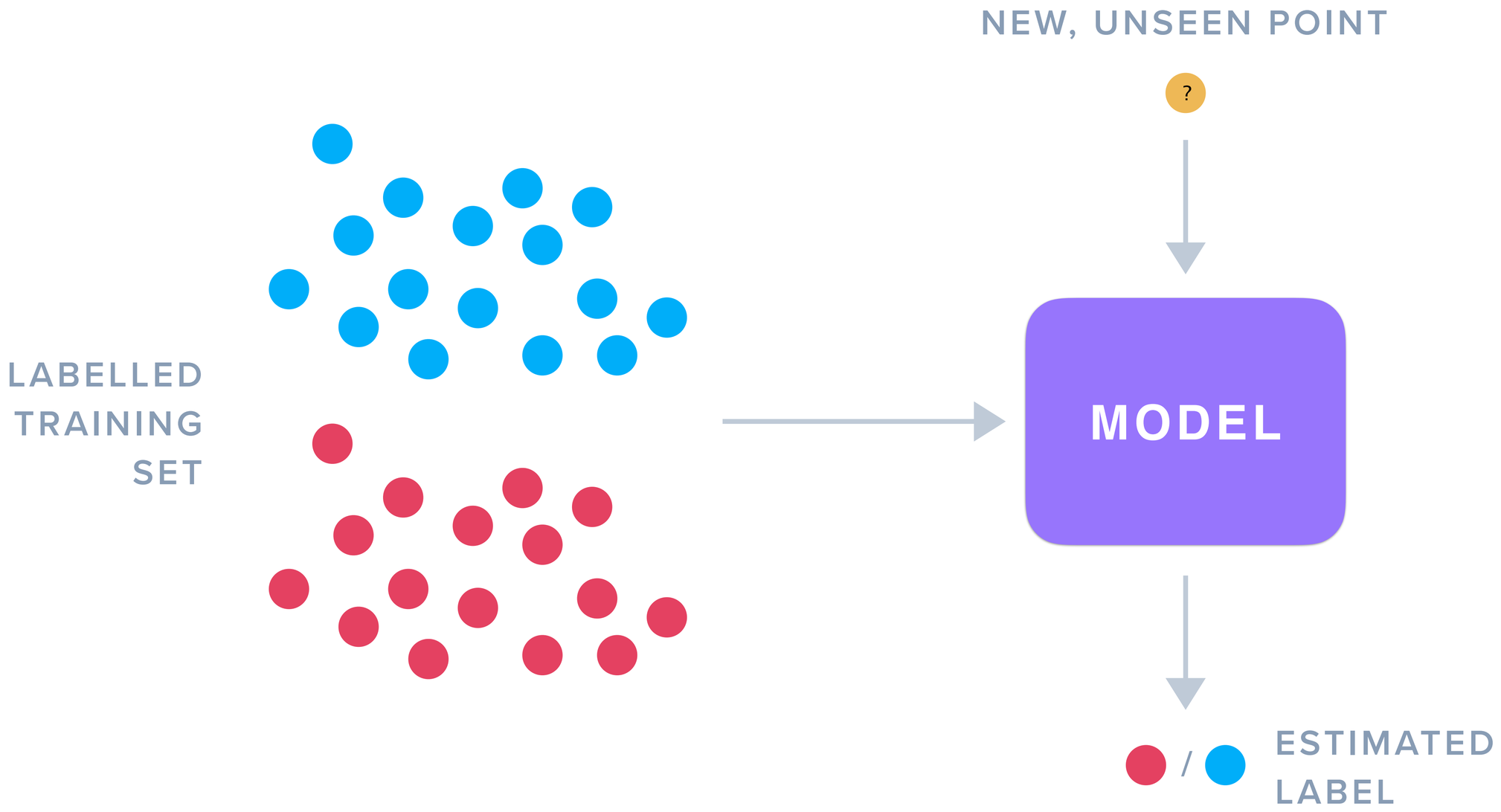
- In binary classification, it is advisable to choose an odd number of K
- During voting, neighbours can have equal influence or different weights according to distance

Applications

- **Content Retrieval & Computer Vision**
  Handwriting detection, Video detection
- **Gene Expression**
- **Protein-Protein interaction**
- **3D structure prediction**
- **Geographical Information Systems (GIS)**
  Detect cities close to current location

Model Evaluation

Remember?



Model Evaluation

Assessment of the “generalisation performance”: the prediction of performance on future unseen data

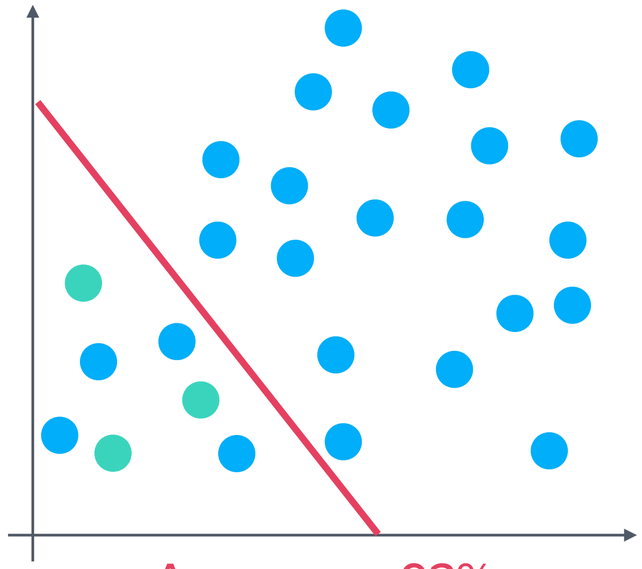
Confusion matrix: number of correct and incorrect predictions

Model Evaluation

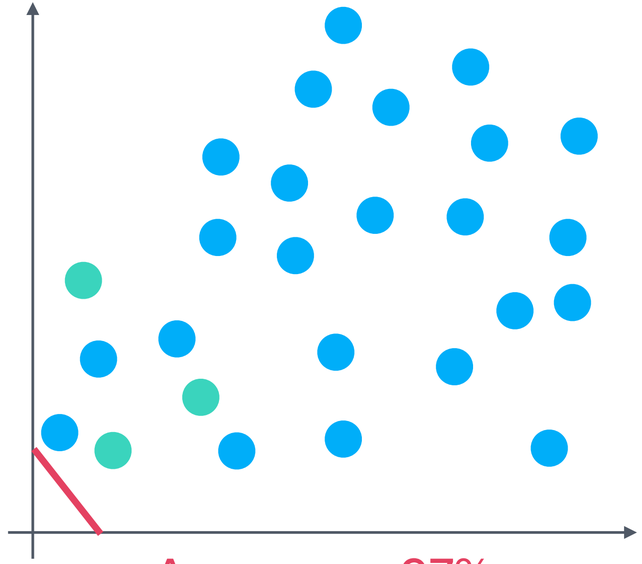
- **Accuracy**: the proportion of the total number of predictions that were correct

$$\frac{TP + TN}{P + N}$$

Sometimes we can’t rely on accuracy



Accuracy = 93%



Accuracy = 97%

Model Evaluation

- **Recall or sensitivity**: the proportion of actual positive cases which are correctly identified

$$\frac{TP}{P} = \frac{TP}{TP + FN}$$

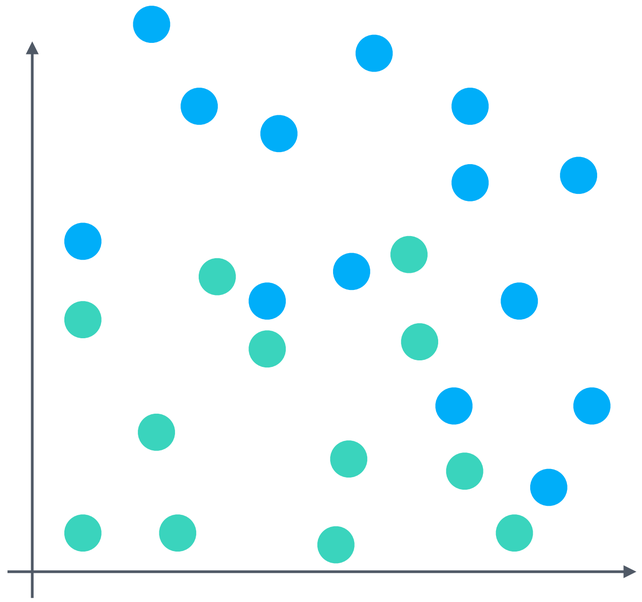
- **Precision**: the proportion of positive cases that were correctly identified

$$\frac{TP}{TP + FP}$$

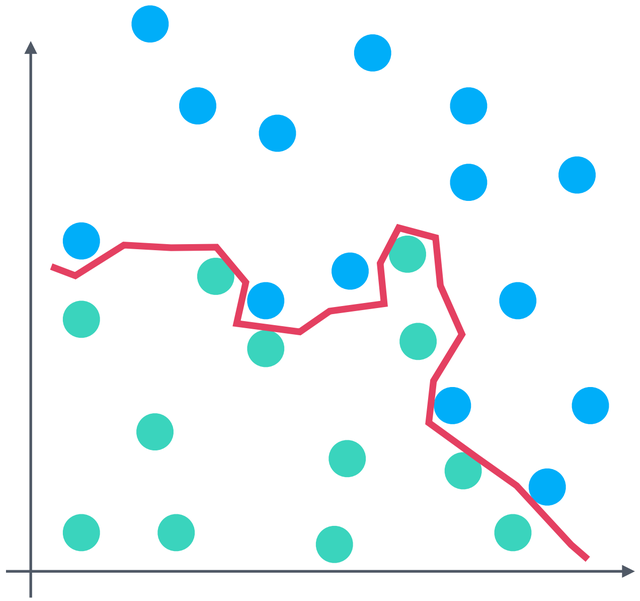
- **F-score**: combination of the two

$$\frac{2TP}{2TP + FP + FN}$$

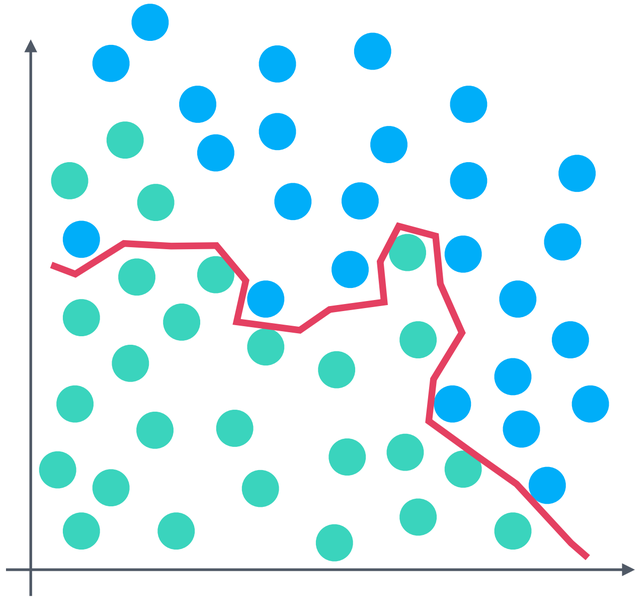
Sometimes we fall in love with the data



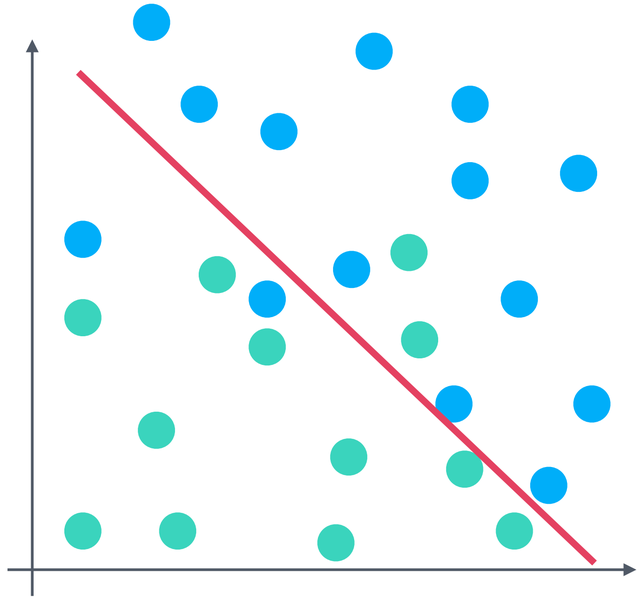
Sometimes we fall in love with the data



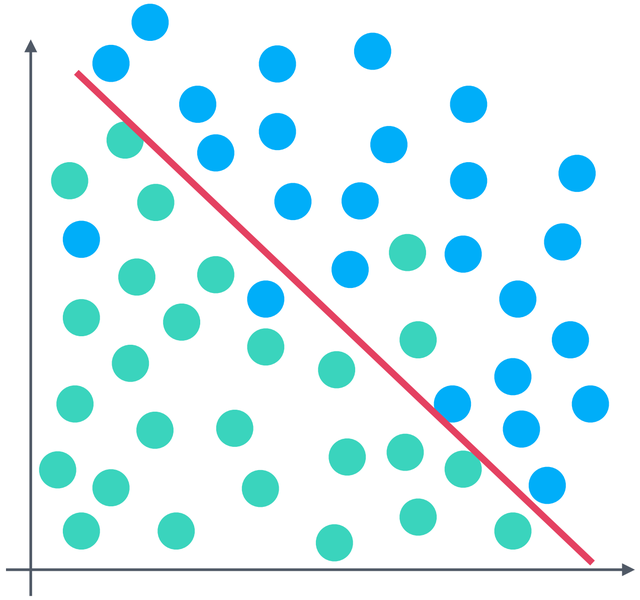
Sometimes we fall in love with the data



Sometimes we fall in love with the data

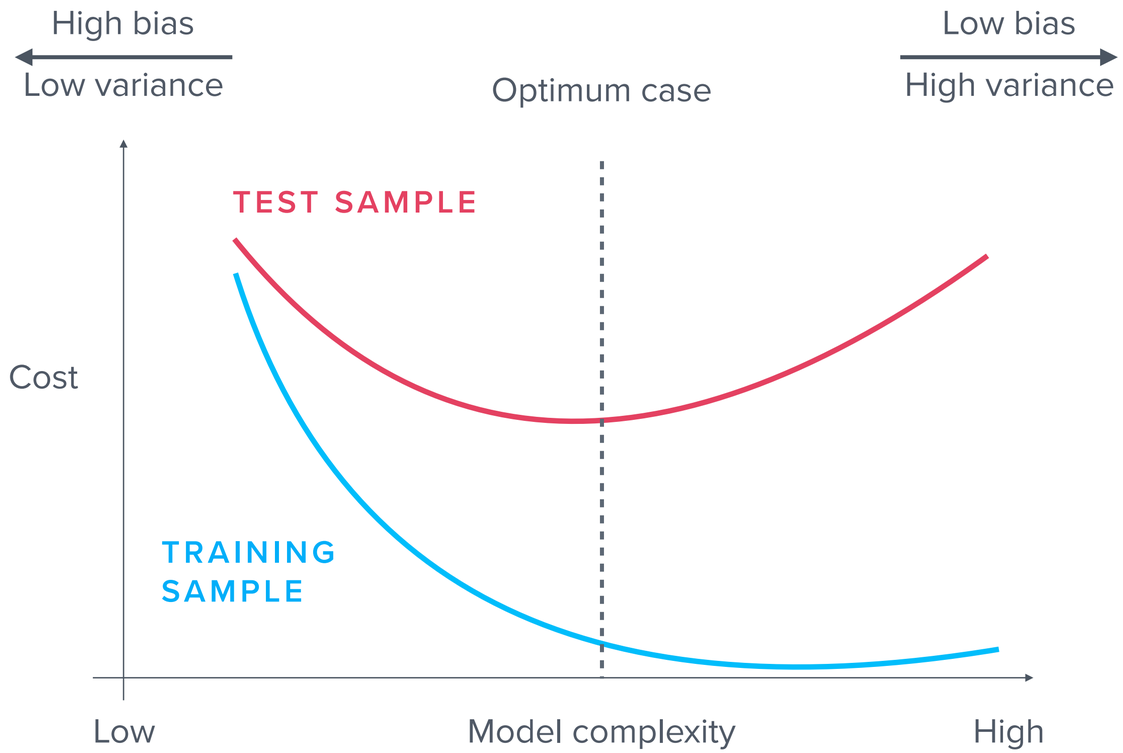


Sometimes we fall in love with the data



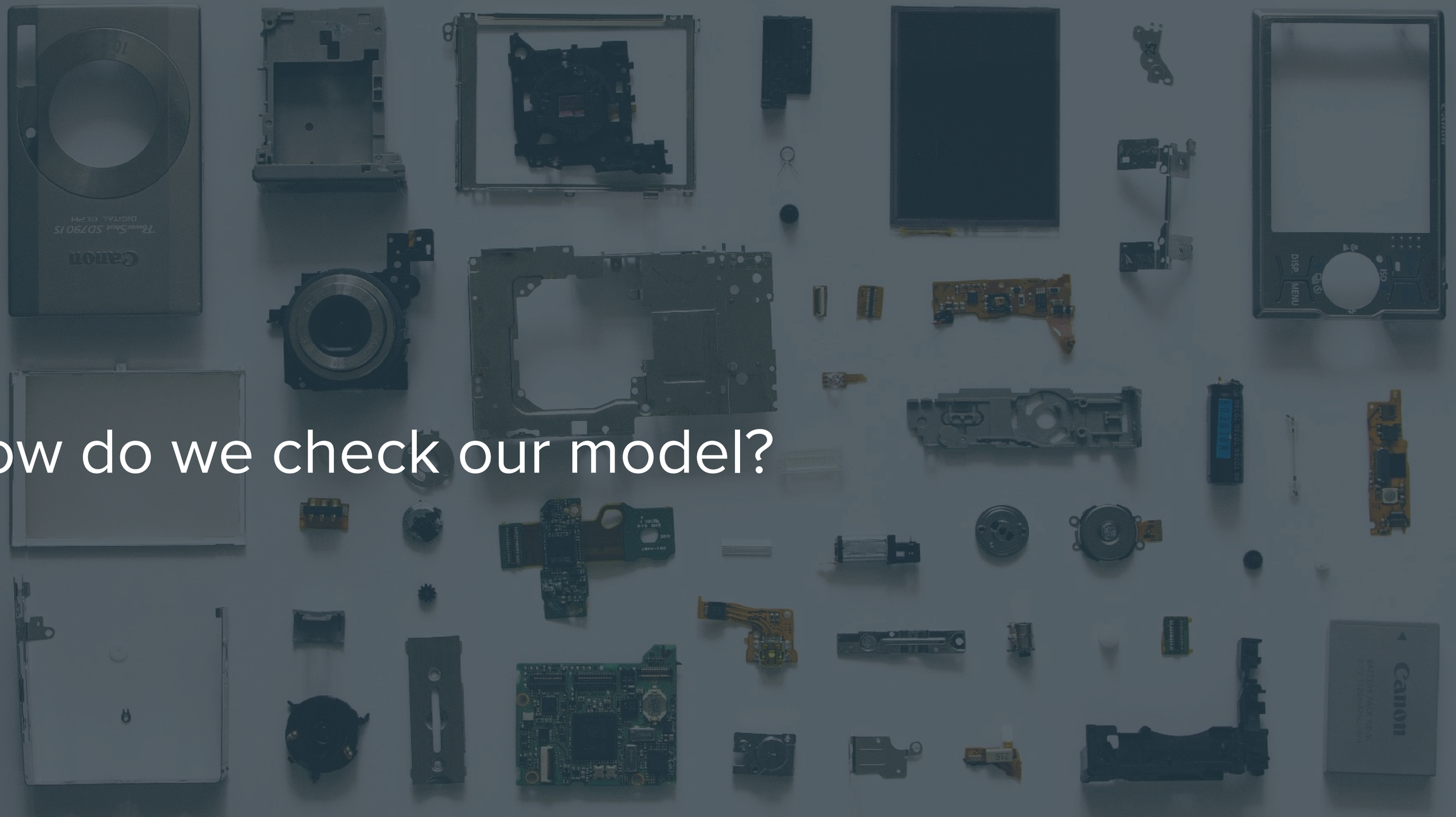
Model Evaluation

Model complexity and over-fitting; the bias-variance trade-off



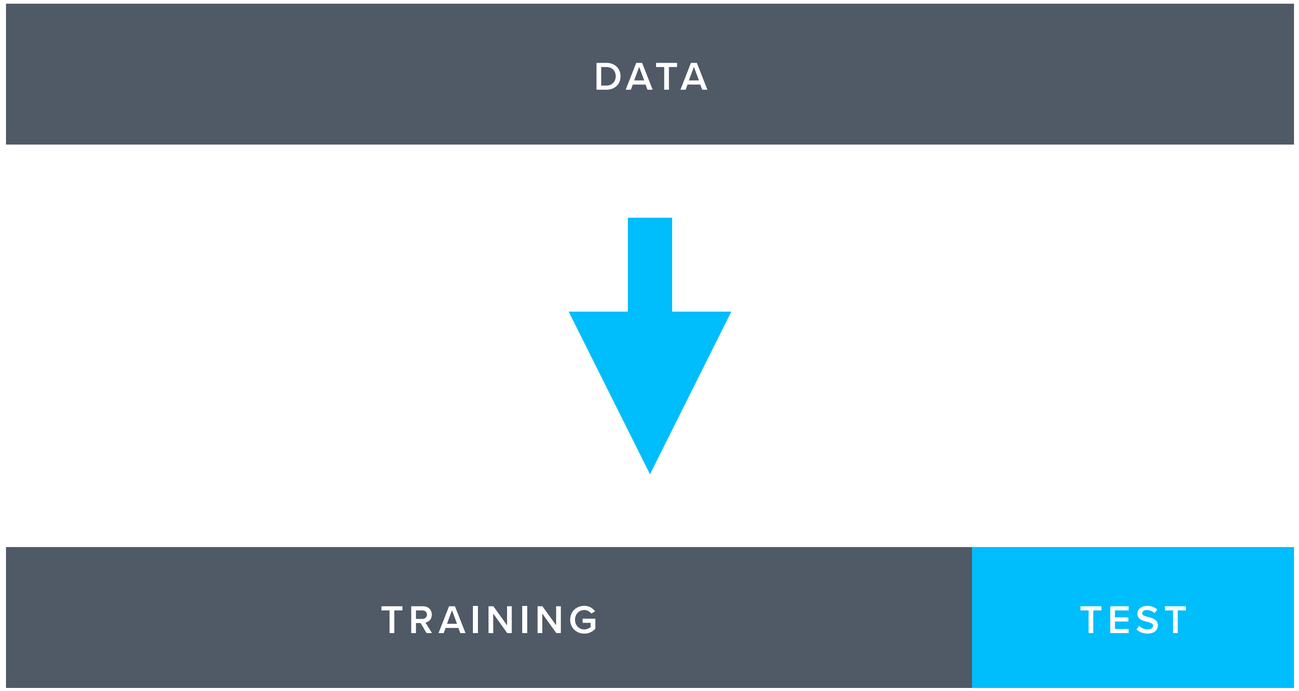
An optimal stopping point has to be found before overfitting

How do we check our model?



Validation

Splitting data into a training and test set



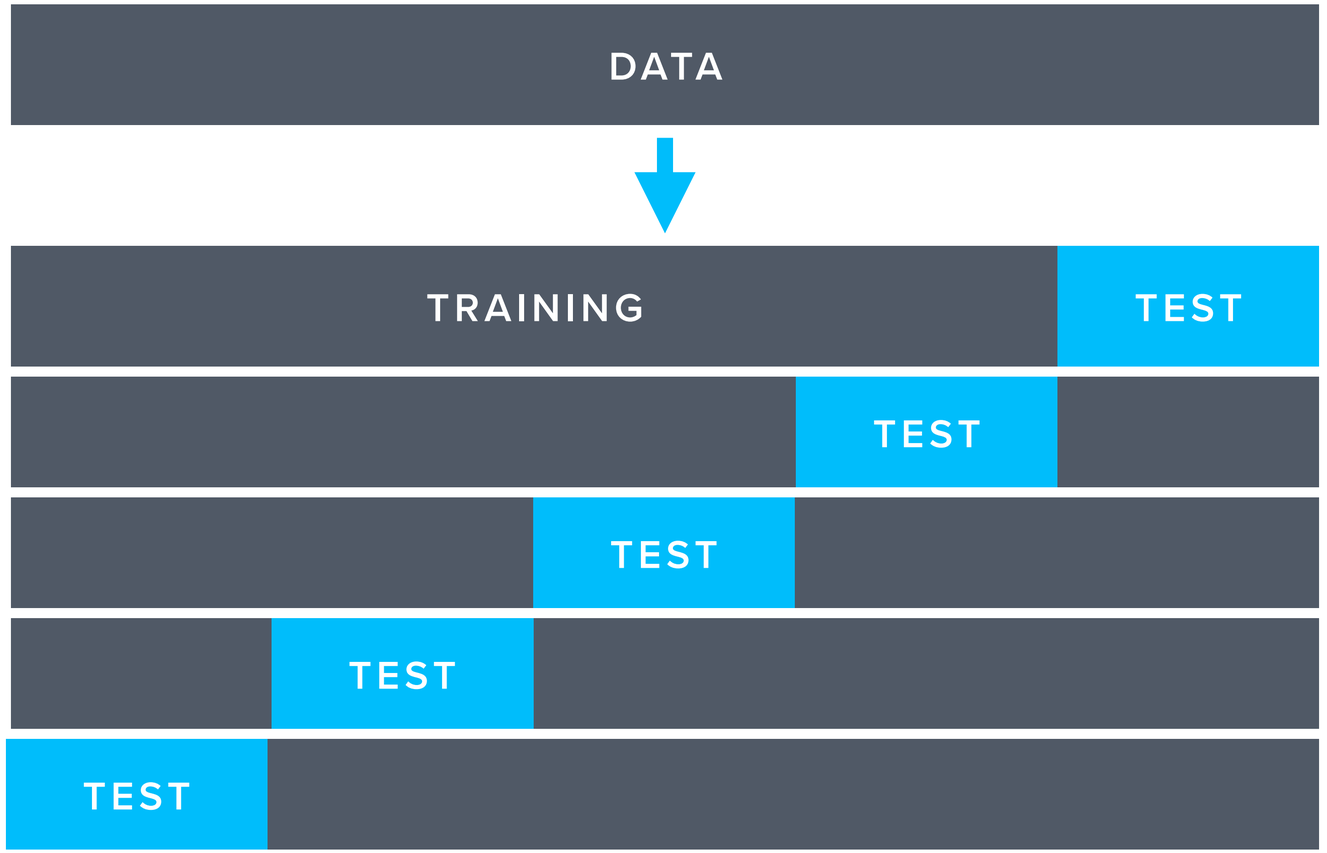
Validation

Three-way split:

- Training set used for tuning model parameters
- Validation set used for evaluation and model selection
- Testing set used to assess the generalisation performance

Validation

k-fold cross validation and leave-one-out cross-validation (LOOCV)



EXERCISE: K-NN

More ASI Training!

- **AI for Executives:**
  - 18-20 July
- **From Data Analyst to Data Scientist:**
  - 24-28 July
- **Data Science Masterclass - Neural Nets**:
  - 1st of September
- Sign up via:

www.asidatascience.com/training

Get in touch

training@asidatascience.com

andrew@asidatascience.com