



DATA SCIENCE
STRATA LONDON 2017

@ASIDataScience

Practical Machine Learning with Python

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

- 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

SHERLOCK ML

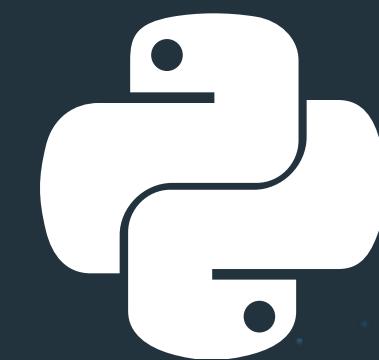
Meet your teacher



Andrew Brookes

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

SHERLOCK ML



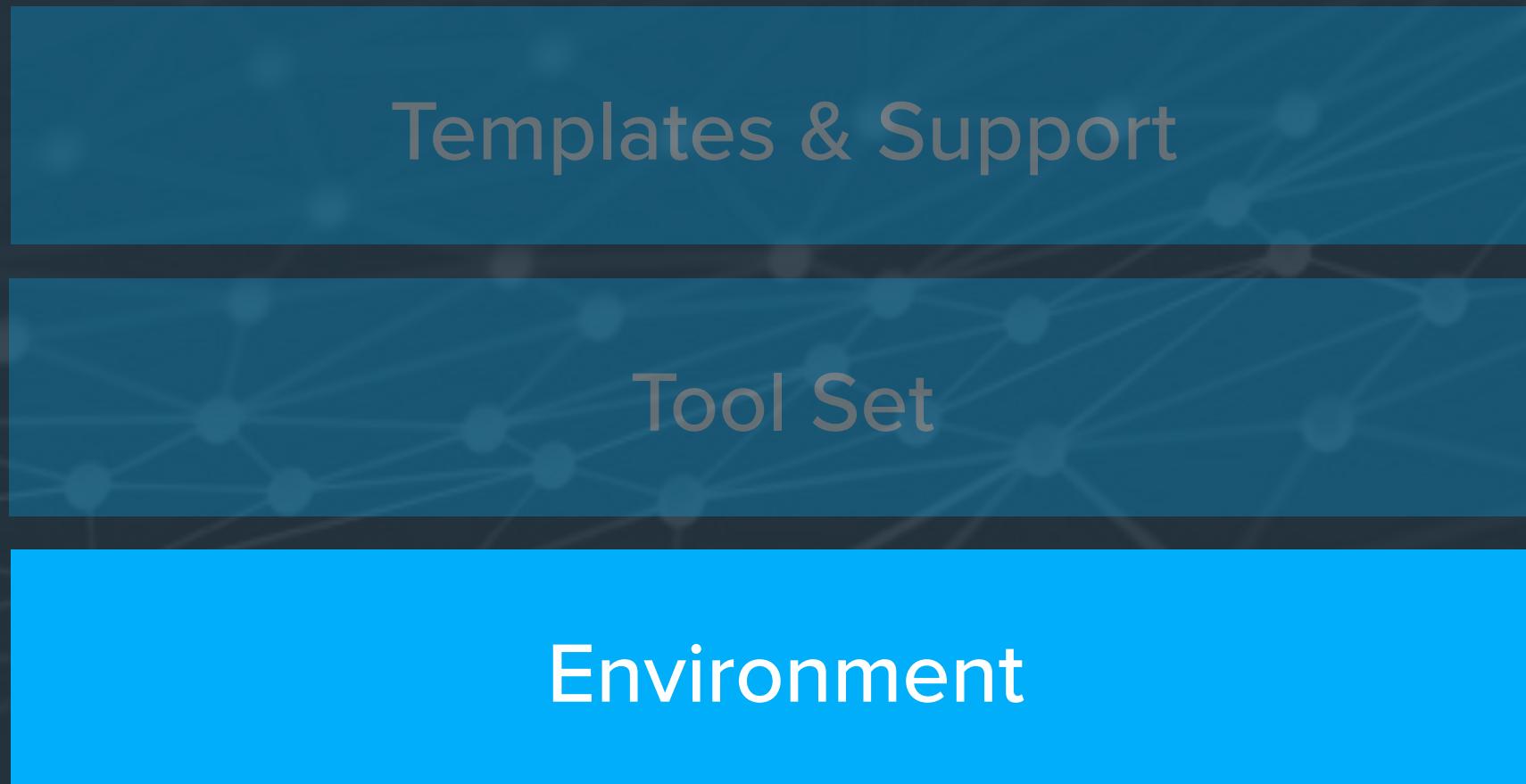
Sherlock has three functional layers

Templates & Support

Tool Set

Environment

Sherlock has three functional layers



Secure

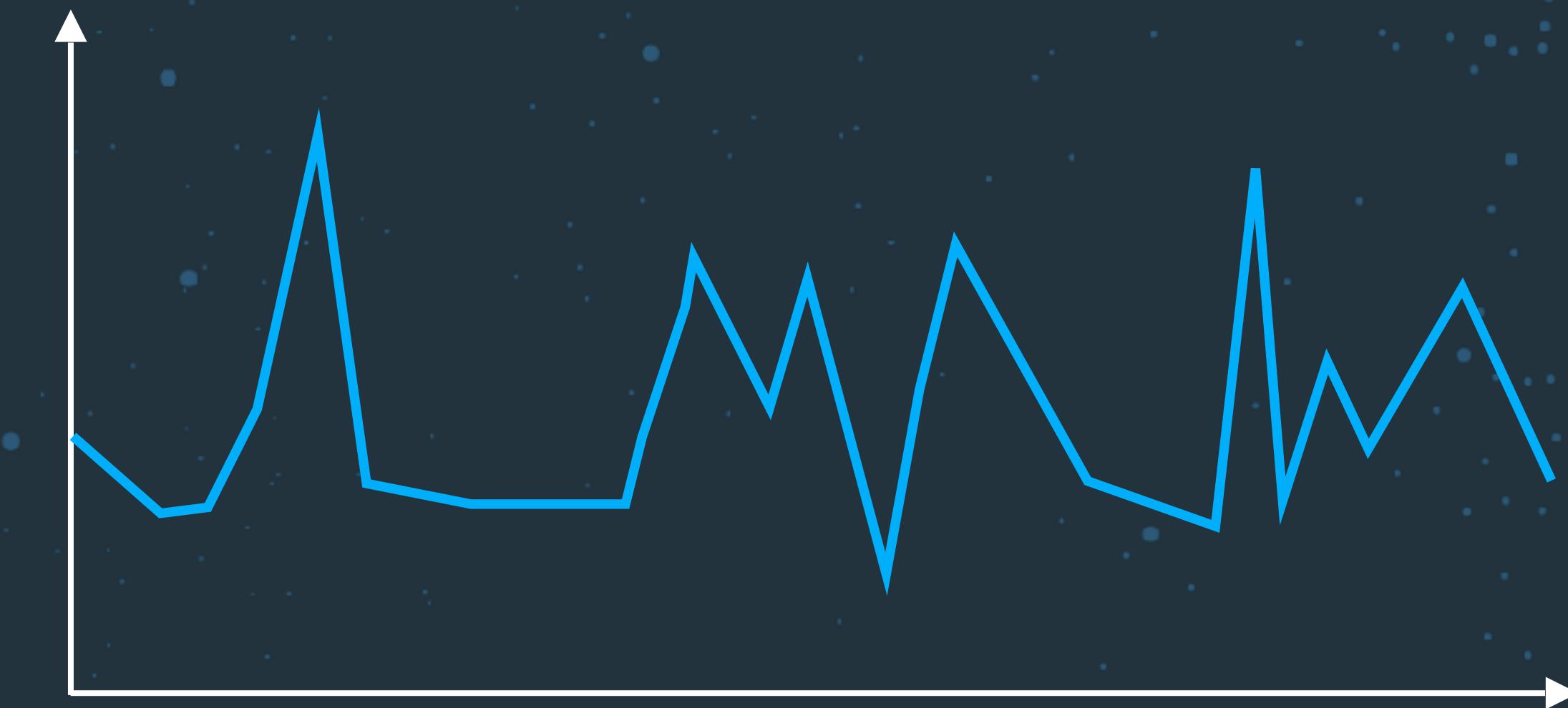


Collaborative

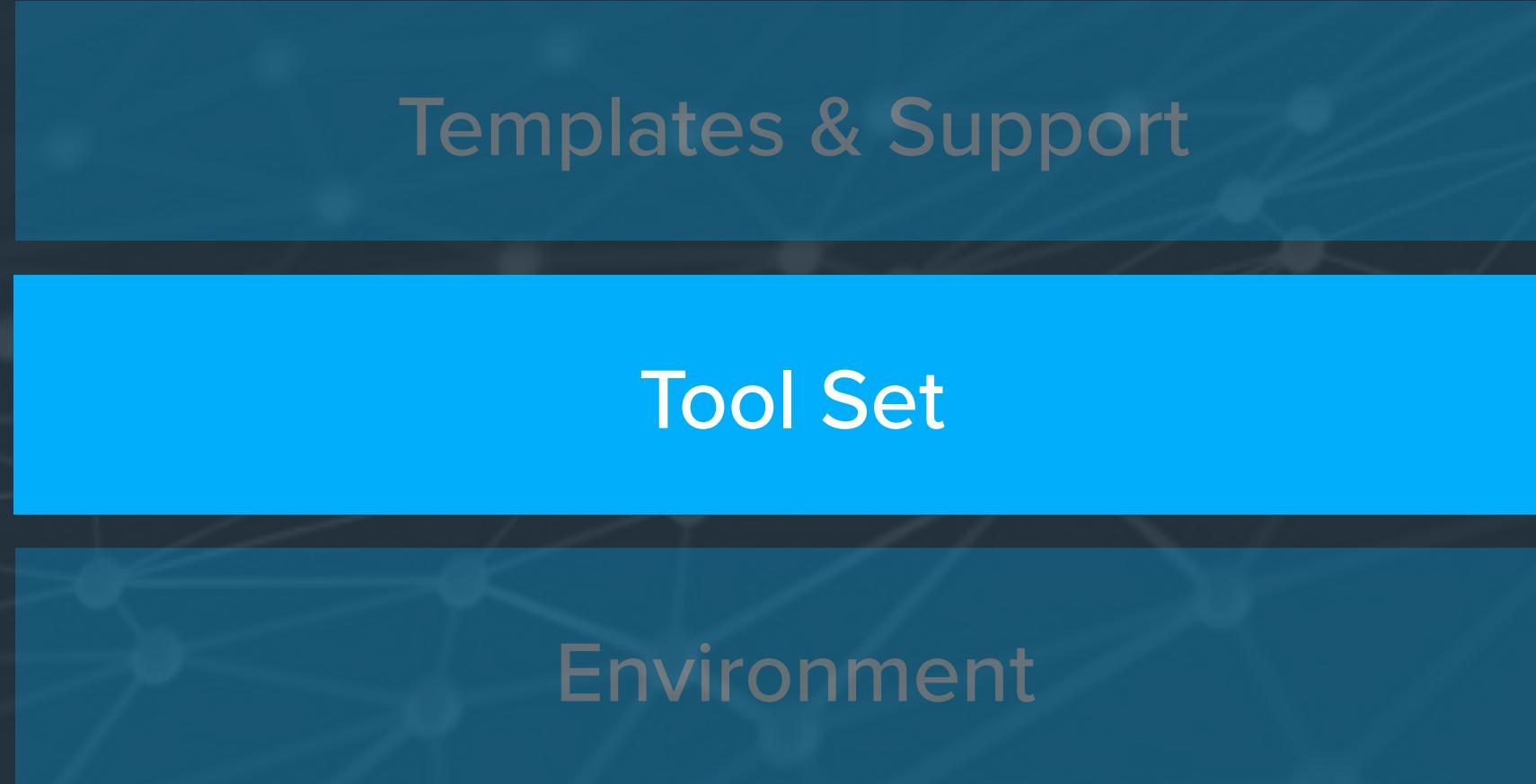


Auditable

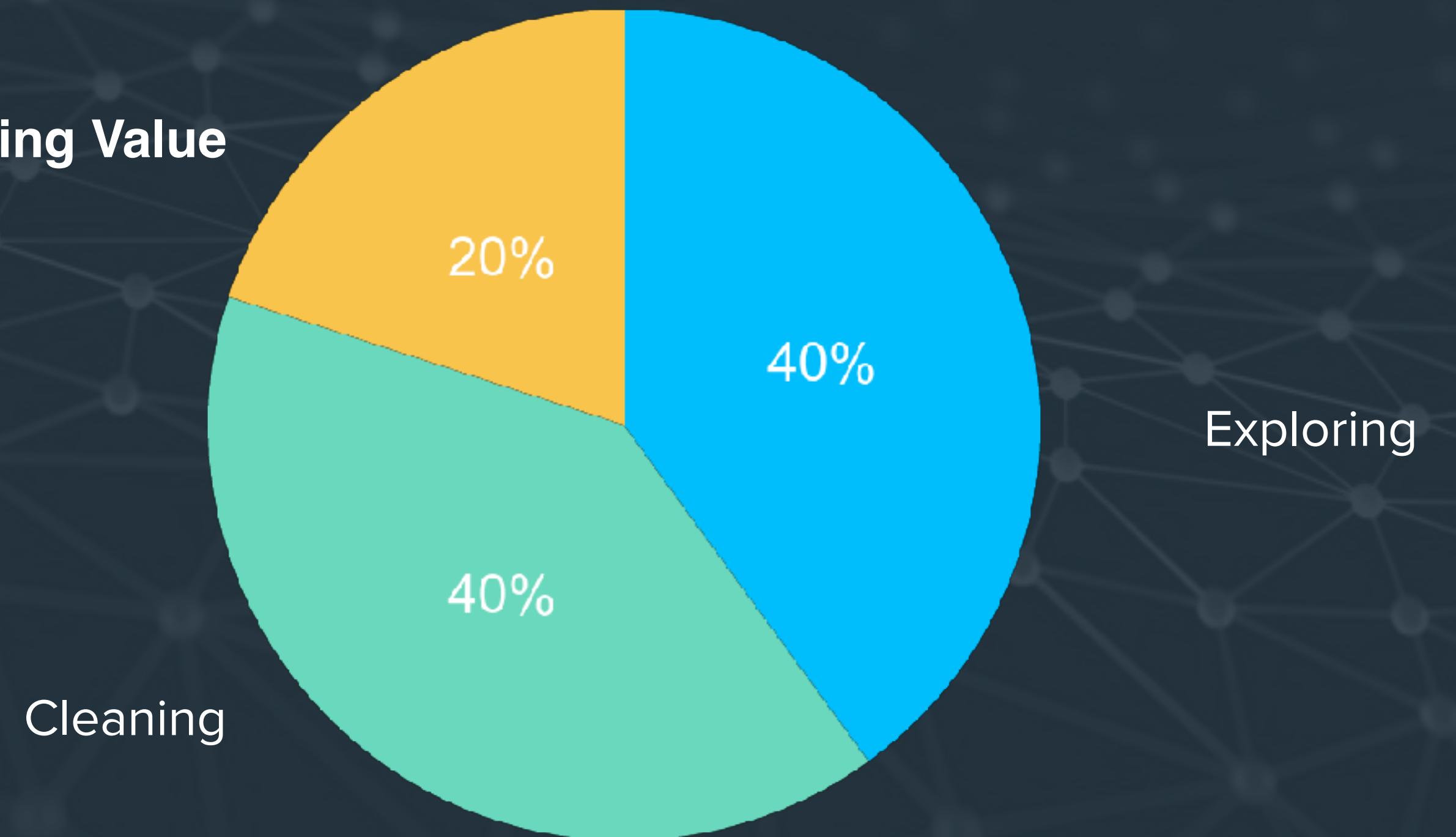
Secure, Scalable Compute



Sherlock has three functional layers



Adding Value



Sherlock has three functional layers

Templates & Support

Tool Set

Environment

ASI

The screenshot shows a Jupyter Notebook interface with the title "Machine Image Recognition". The notebook contains text about the human visual system and machine learning systems. It also includes a section titled "Label custom images" with instructions for uploading photos and running a machine learning model. A red arrow points to the "Upload" button in the top right corner of the Jupyter interface.

jupyter ImageRecognition Last Checkpoint: 03/14/2017 (autosaved)

File Edit View Insert Cell Kernel Widgets Help

Python [conda env:Python3]

Machine Image Recognition

The human visual system is one of the wonders of the world.

We carry in our heads a supercomputer. Our brain has billions of neurons, with even more connections between them, all tuned by evolution over hundreds of millions of years to be superbly adapted to understand the visual world.

This machine learning system knows its Chihuahuas from its Pomeranians. Using pixels alone.

Over the past four years you have doubtlessly noticed quantum leaps in the quality of everyday technologies. Take your photo collection - you can now search or automatically organize collections of photos with no identifying tags.

Think about that. To gather up dog pictures, the app must identify anything from a Chihuahua (or Pomeranian) to a German shepherd and not be tripped up if the pup is depicted upside down or partially obscured, at the right or the left of the frame, in fog or snow, sun or shade.

Artificial neural networks are the fundamental programmes making it possible for machines to replicate the performance of the human visual system. Here we give you access to a neural network, which has many millions of neurons, for you to put to the task of recognising any image. Your move.

Text inspired by Nielsen (2016) and Parilloff (2016).

Label custom images

1. Upload your JPEG photo in the same directory as this notebook
 - Use the Upload button at the top-right corner of the Jupyter page.
2. Run the machine learning model by selecting Cell and Run All
3. View results for different images using the drop-down menu

We have provided photos of our beautiful office Chihuahuas (`toby.JPG` & `chibi.JPG`)

jupyter

Files Running Clusters Conda

Select items to perform actions on them.

Upload New

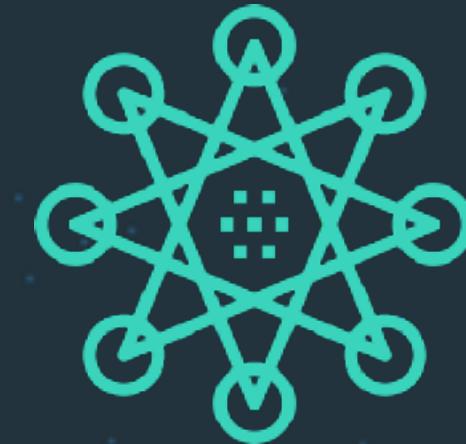


Accountants » Excel

Designers » Photoshop

Data Scientists » ?

SHERLOCK ML



Powerful



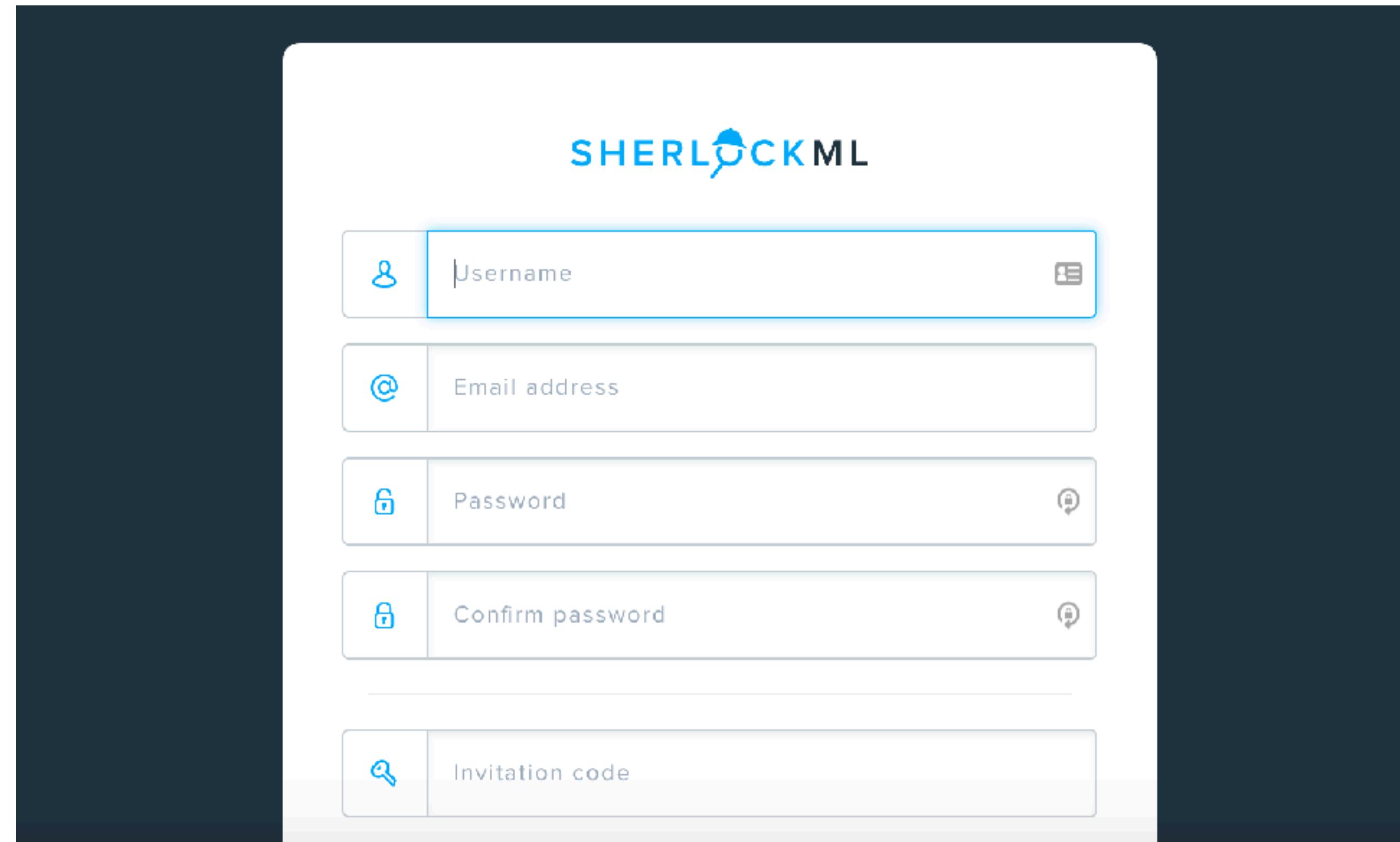
Simple



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

The screenshot shows the SherlockML web interface. At the top, there is a navigation bar with the logo "SHERLOCKML", a "PROJECTS" dropdown, "DOCS", "CHARLOTTE", and a user profile icon. Below the navigation bar is a "CREATE PROJECT" button with a plus sign. A dashed blue rectangle highlights this button. Underneath, there are "QUICK FILTERS" for "ACTIVE" and "MY PROJECTS", and a search bar with a magnifying glass icon. The main area displays three project cards:

- basic_sherlock_data_handling**: Status: 1 SERVER RUNNING, 2 Cores, 4 GB. Owner: ben.
- Business analyst to Data Scientist Course April 2017**: Status: 1 SERVER RUNNING, 2 Cores, 4 GB. Owner: charlotte.
- databot-analysis**: Status: 1 SERVER RUNNING, 2 Cores, 4 GB. Owner: robbie.

Step 3

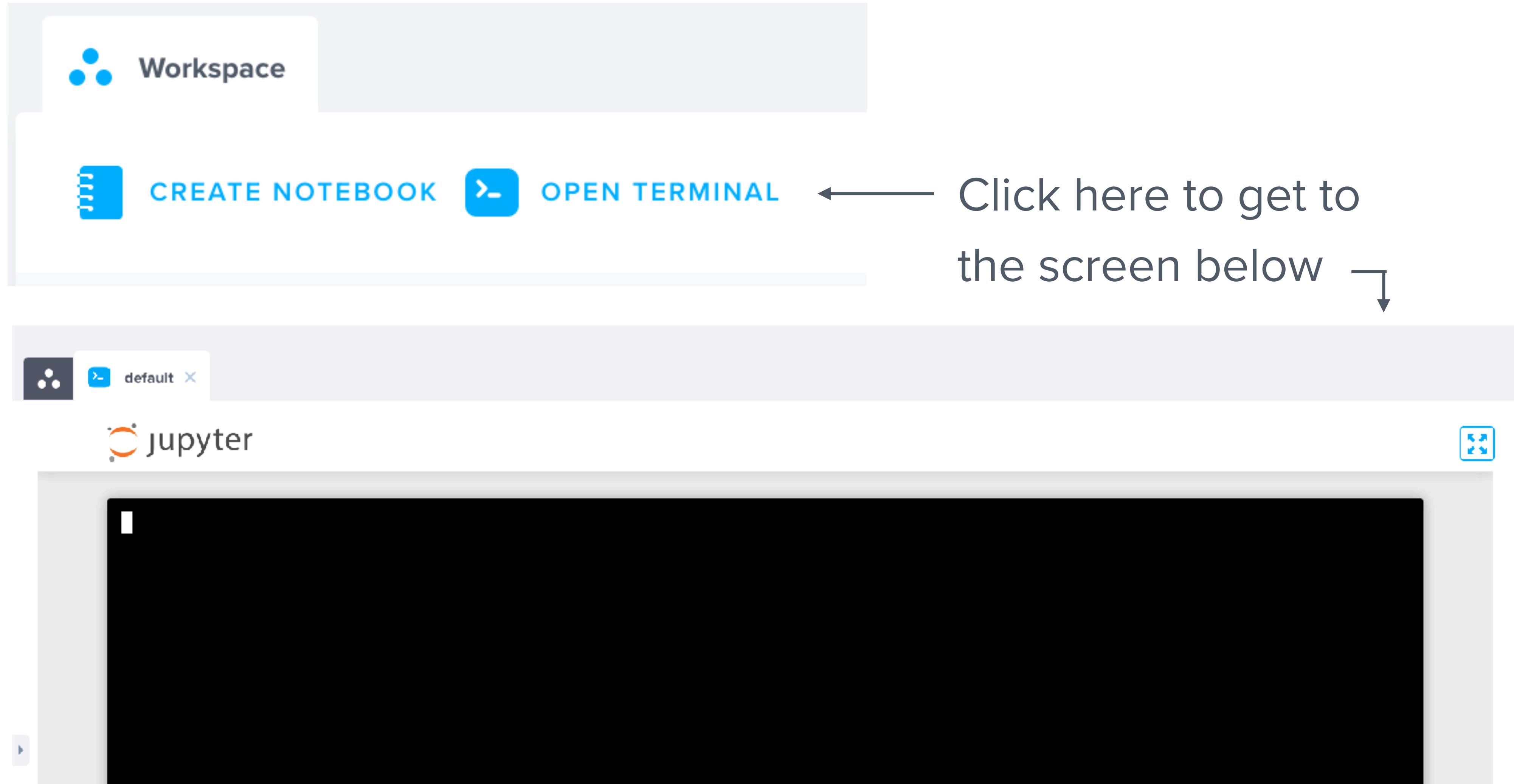
Wait for a server to be created



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



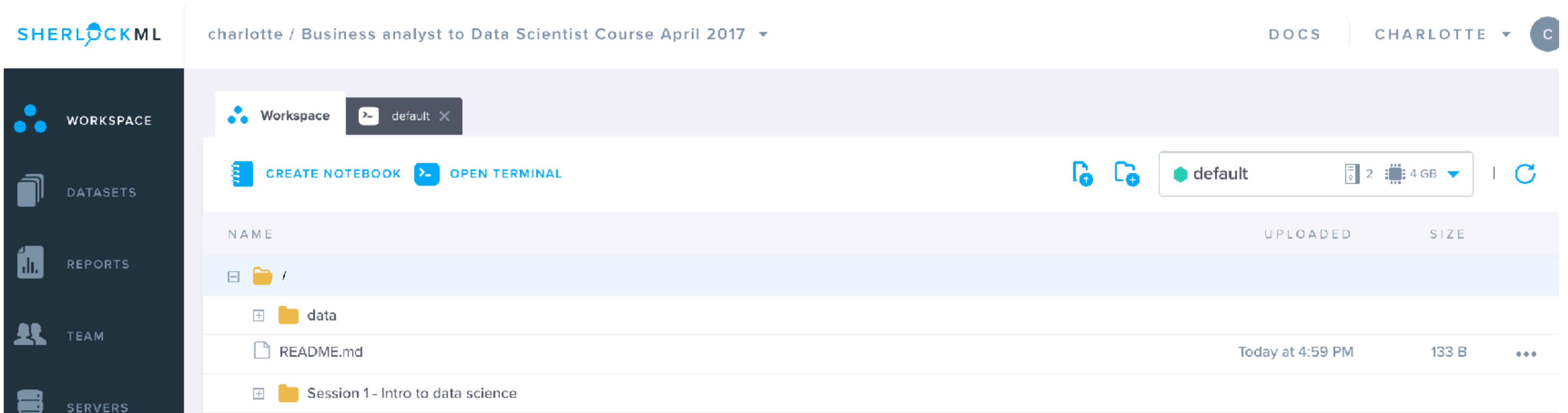
The screenshot shows a Jupyter Notebook interface. At the top, there's a toolbar with a file icon, a dropdown menu labeled "default", and a close button. Below the toolbar, the word "jupyter" is displayed in a large, orange font. On the right side of the interface, there's a small blue square icon with a grid pattern. The main area is a terminal window with a black background and white text. The text in the terminal is as follows:

```
(Python3) sherlock@jupyter:~/workspace$ git clone https://github.com/ASIDataScience/bi2ai-session-1.git
```

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

Step 6

Go back to your workspace and refresh



The screenshot shows the Sherlock ML workspace interface. On the left, there's a sidebar with icons for WORKSPACE, DATASETS, REPORTS, TEAM, and SERVERS. The main area shows a workspace named 'default' with options to CREATE NOTEBOOK or OPEN TERMINAL. Below this is a file list table:

NAME	UPLOADED	SIZE
data	Today at 4:59 PM	133 B
README.md		...
Session 1 - Intro to data science		

The 'data' folder and 'README.md' file are visible, indicating the files have been successfully uploaded.

Files should be there: you're good to go

Machine Learning Models

Machine Learning Models

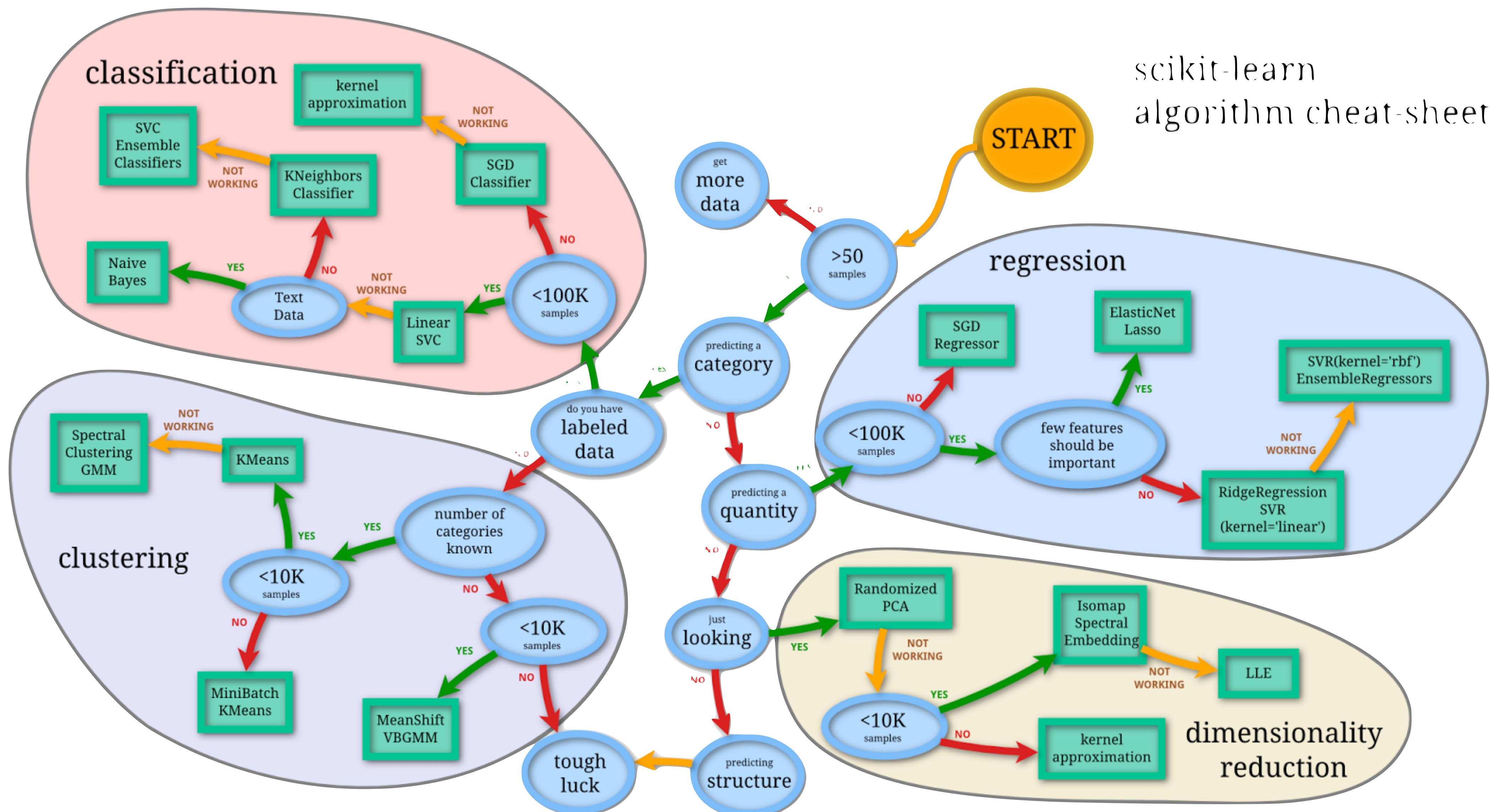
SUPERVISED

- 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

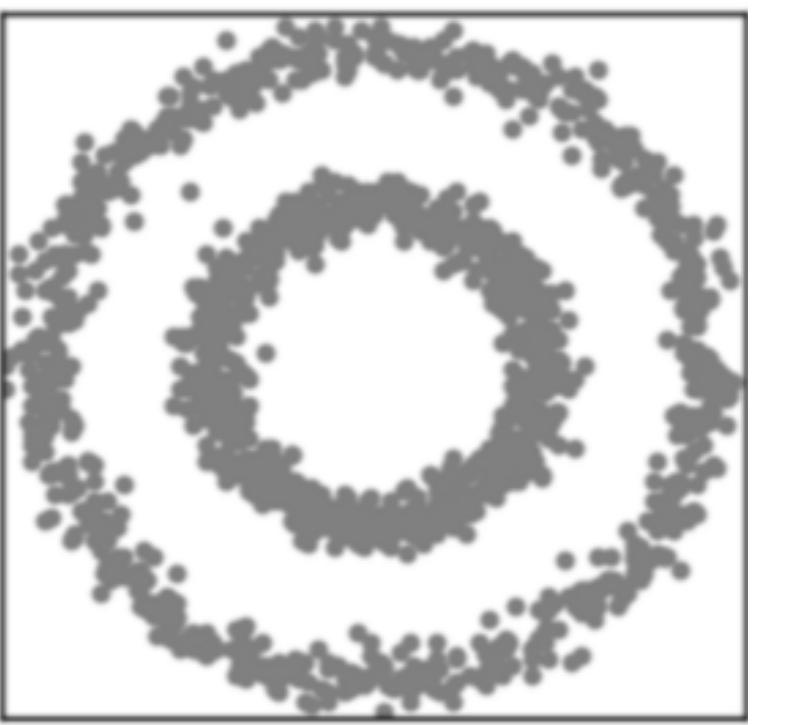
Clustering

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

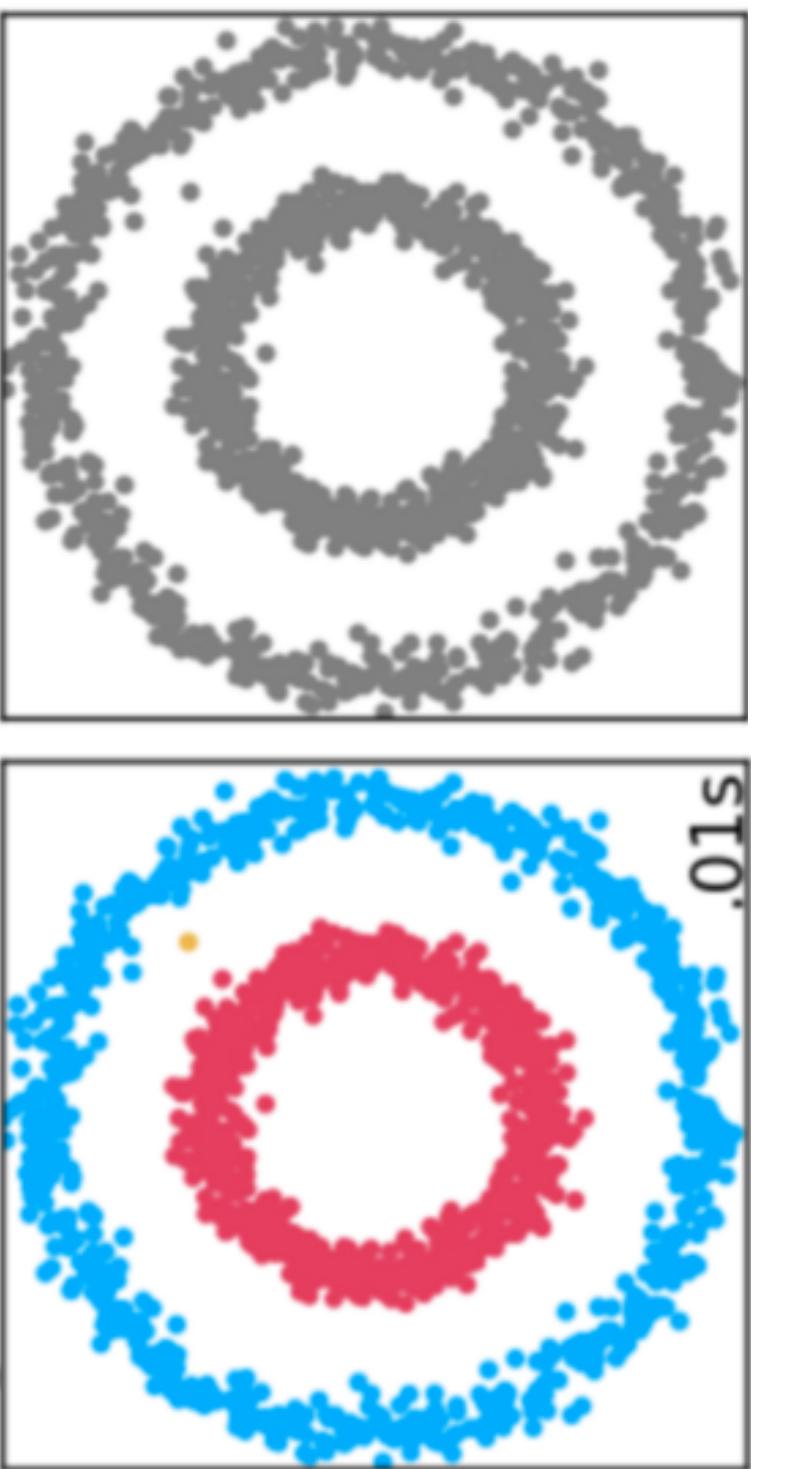
Clustering

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

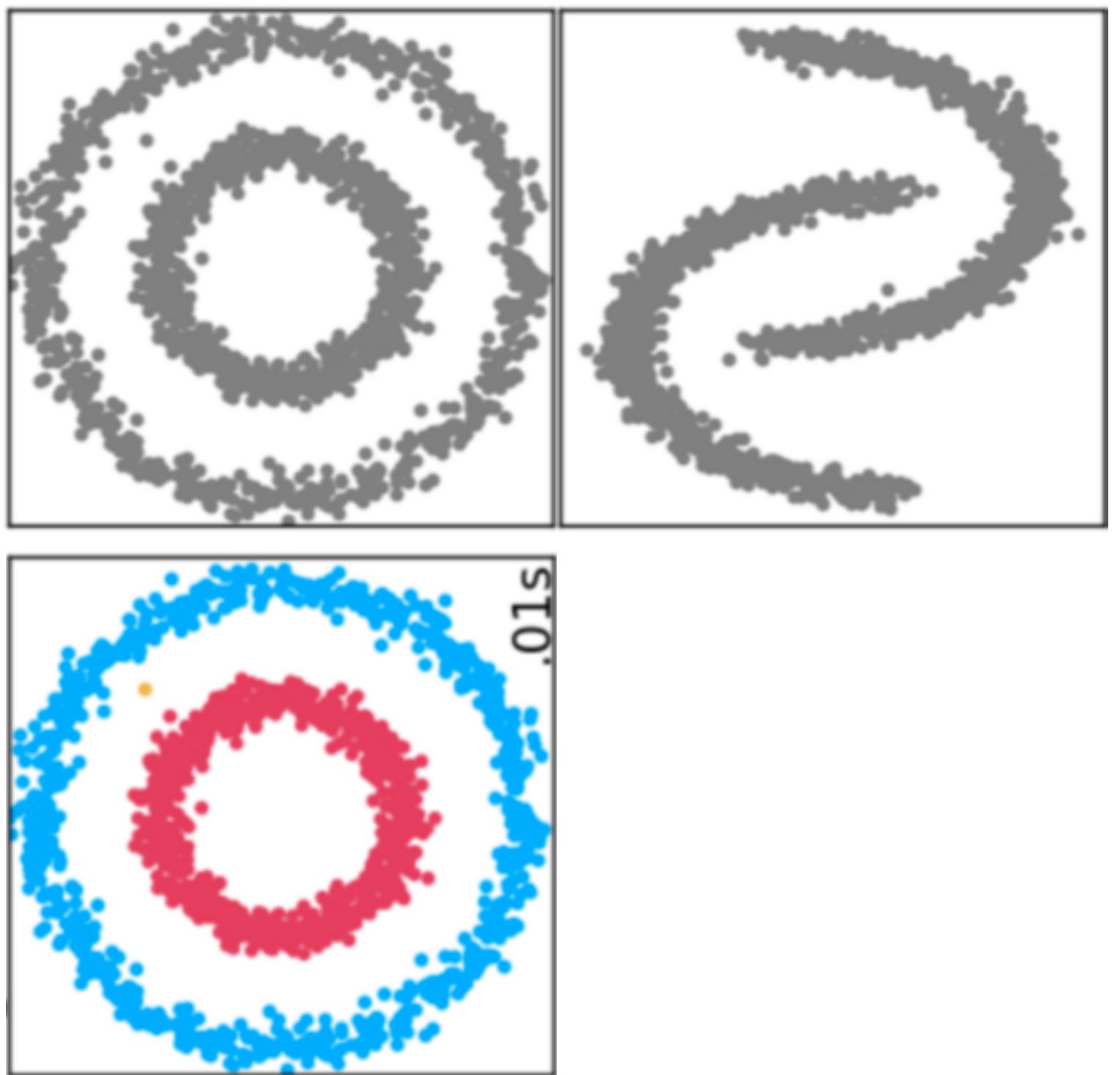
Example datasets



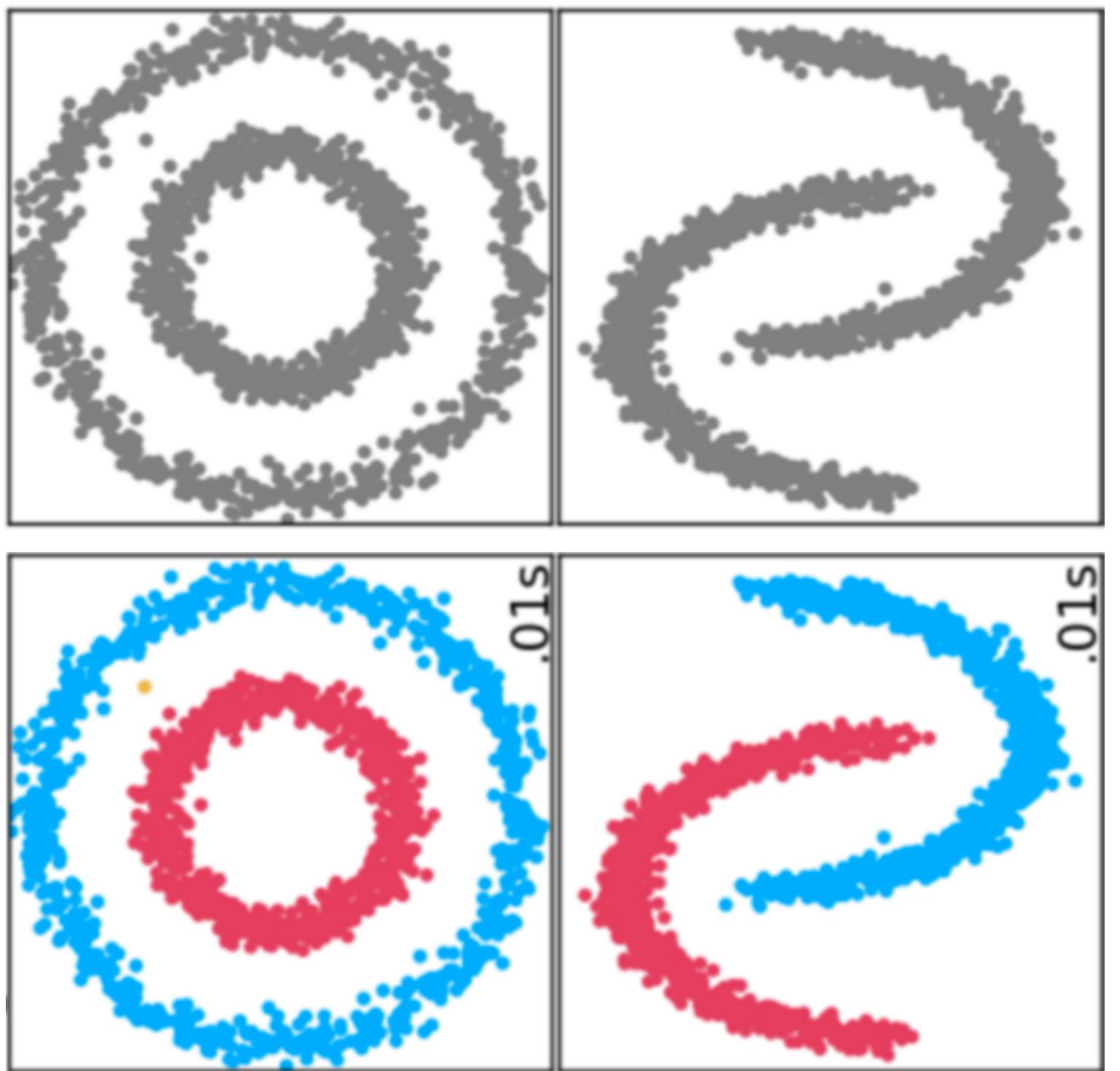
Example datasets



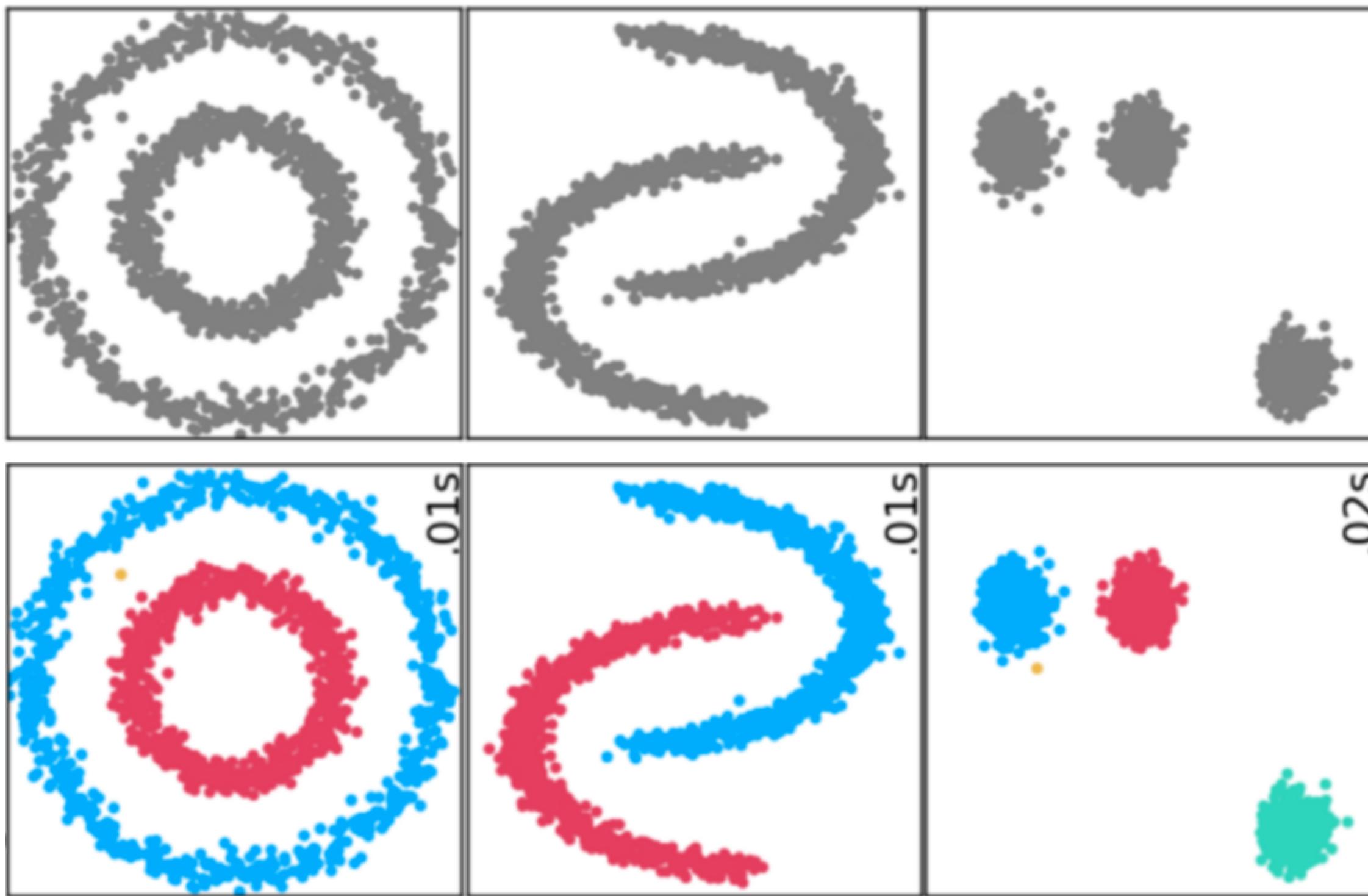
Example datasets



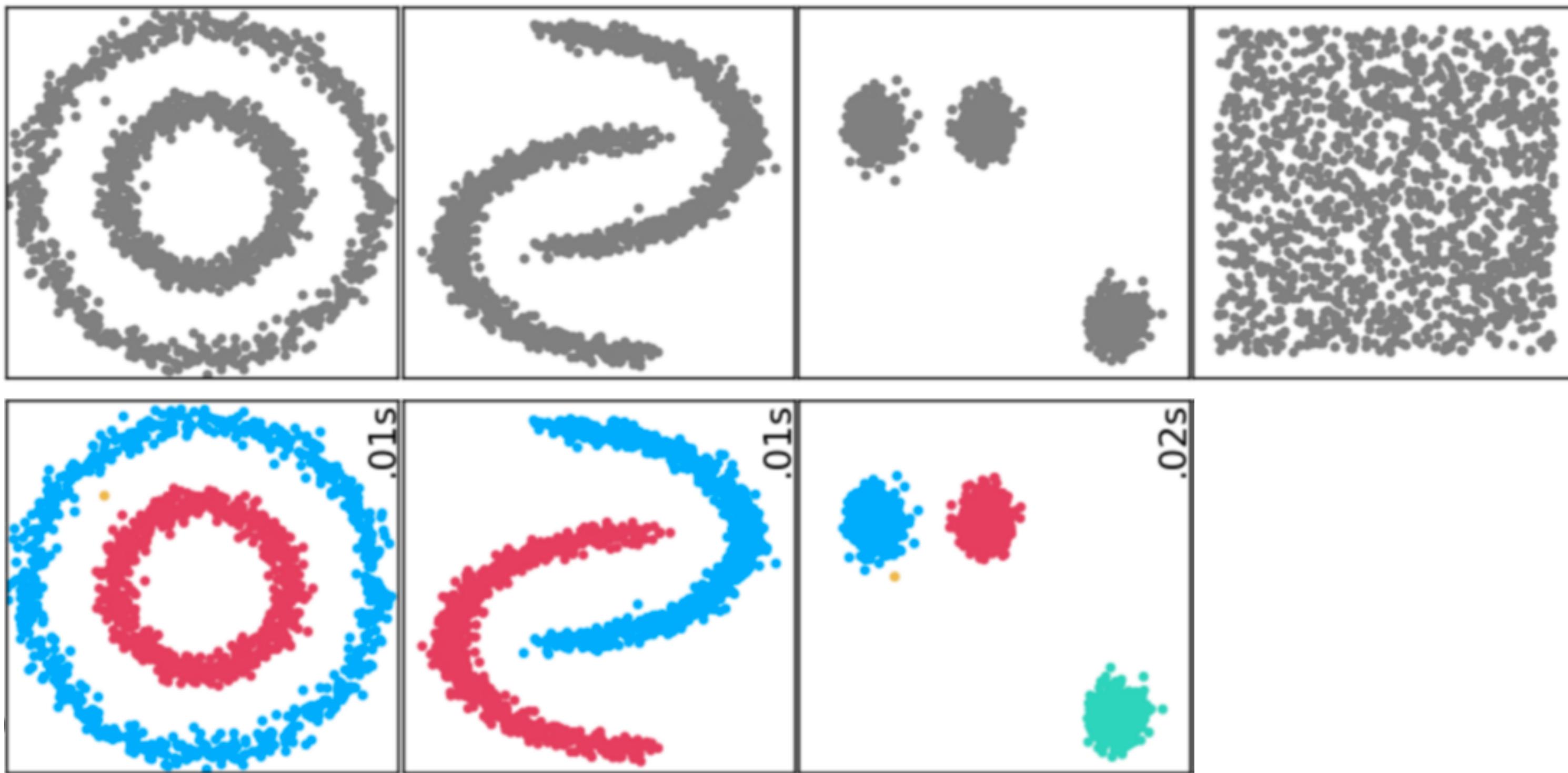
Example datasets



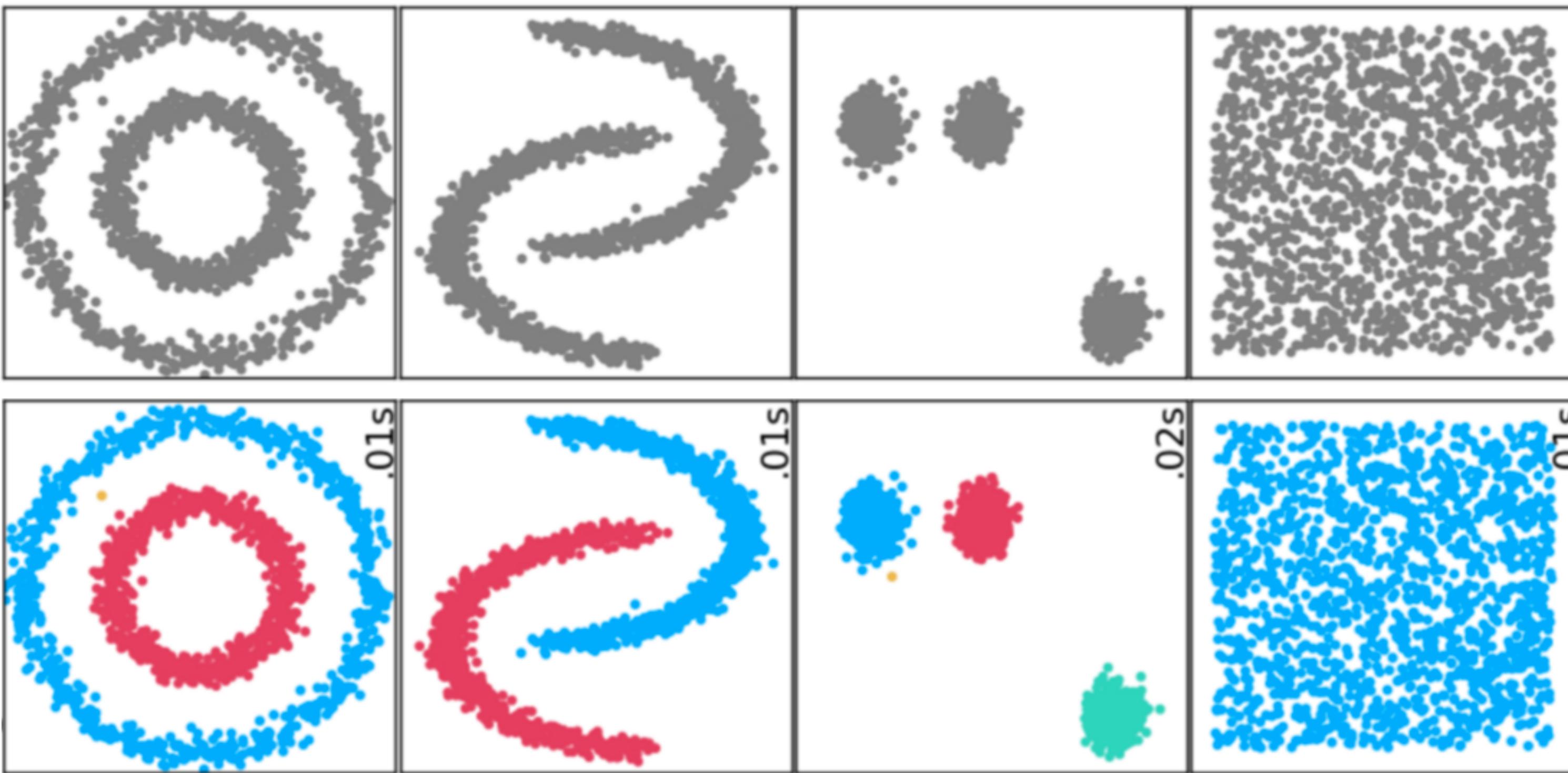
Example datasets



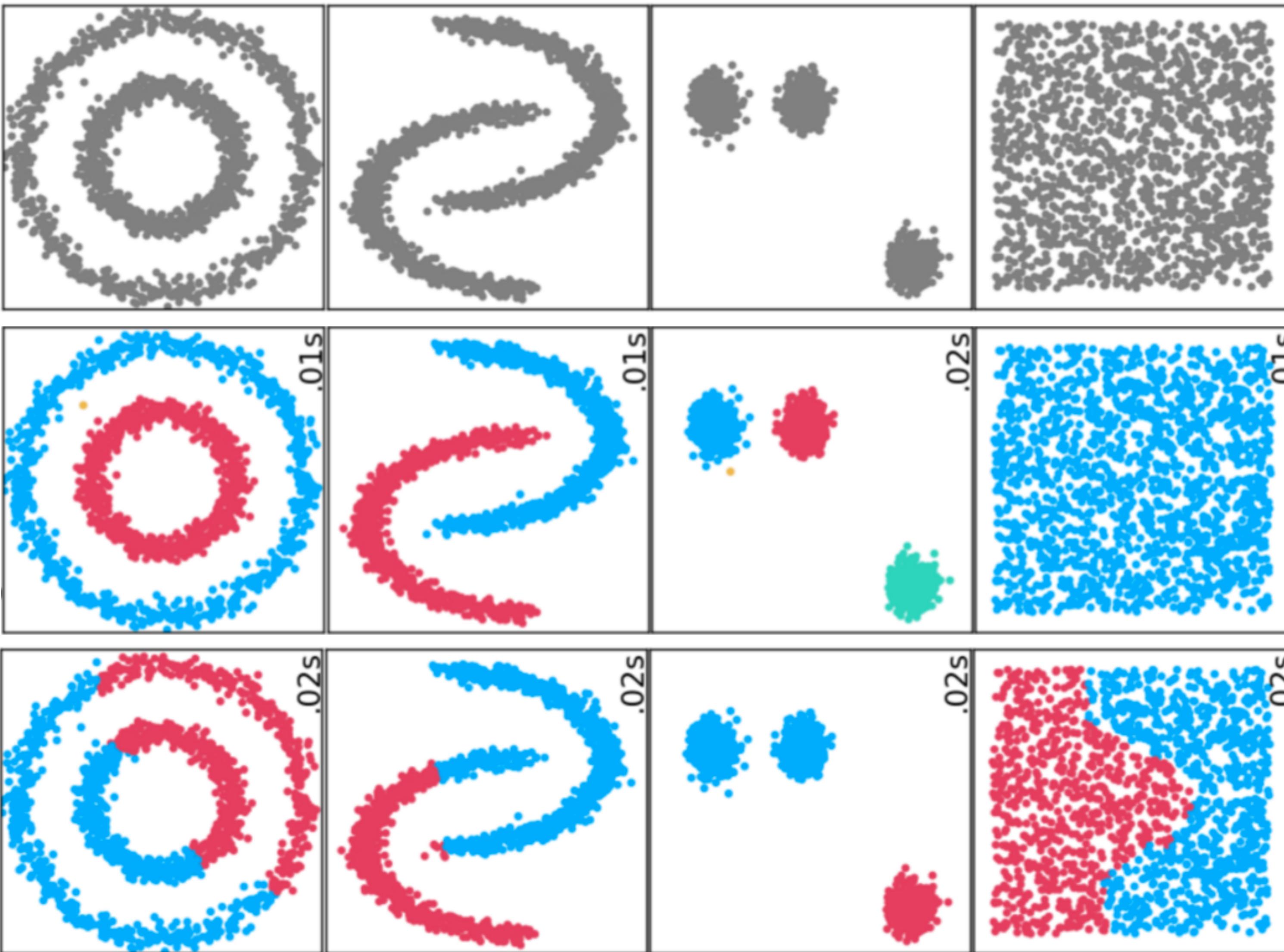
Example datasets



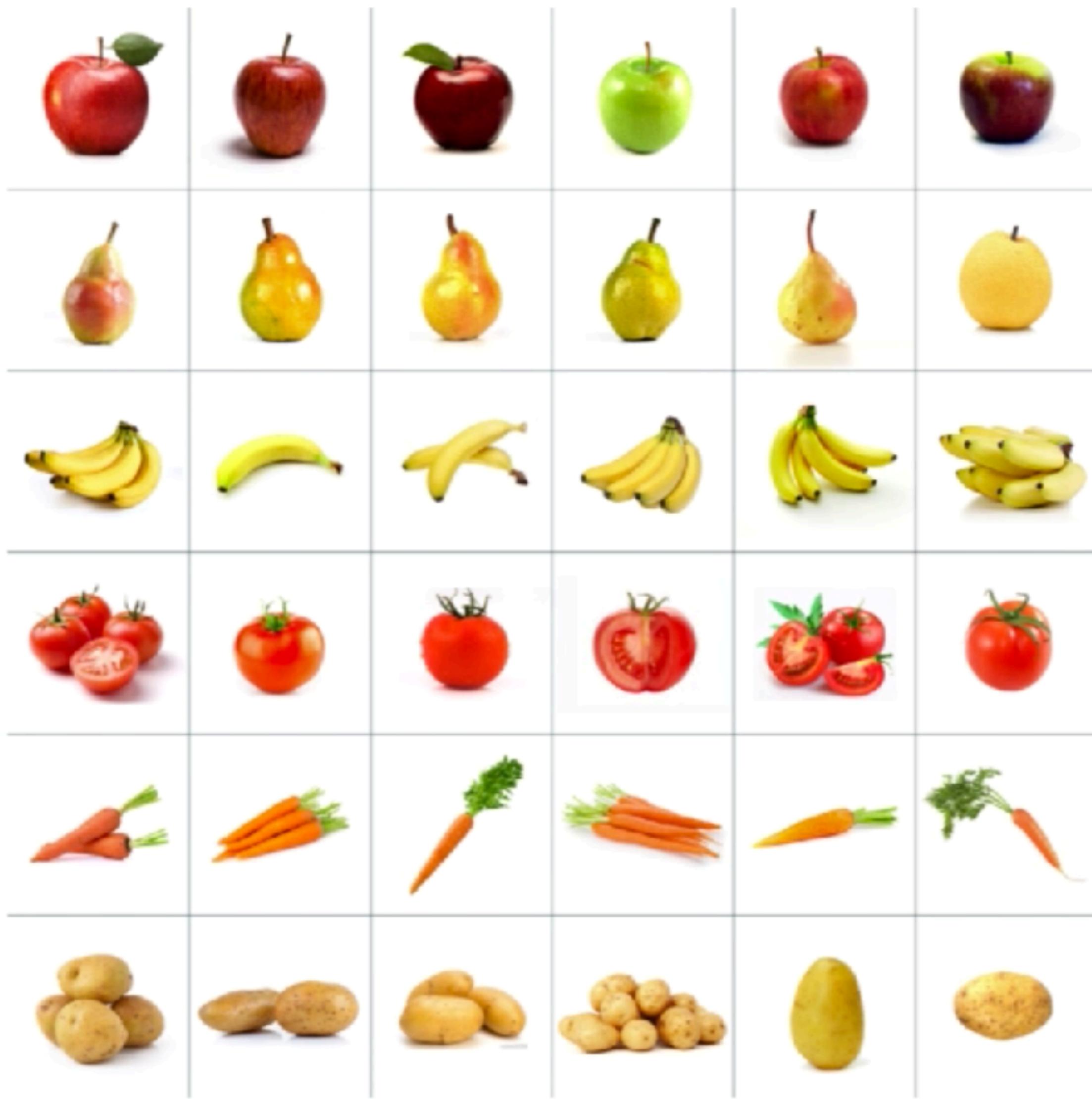
Example datasets



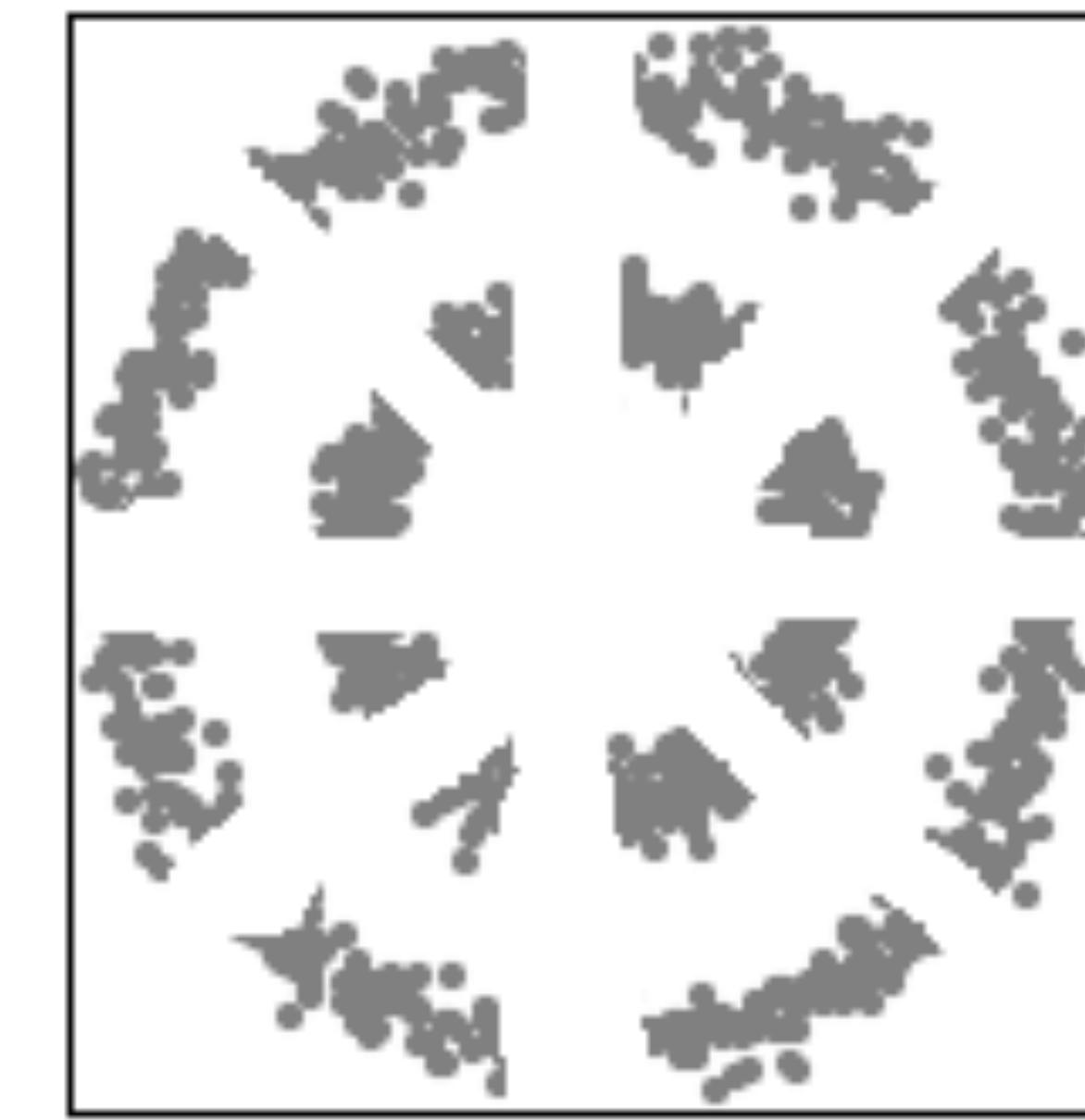
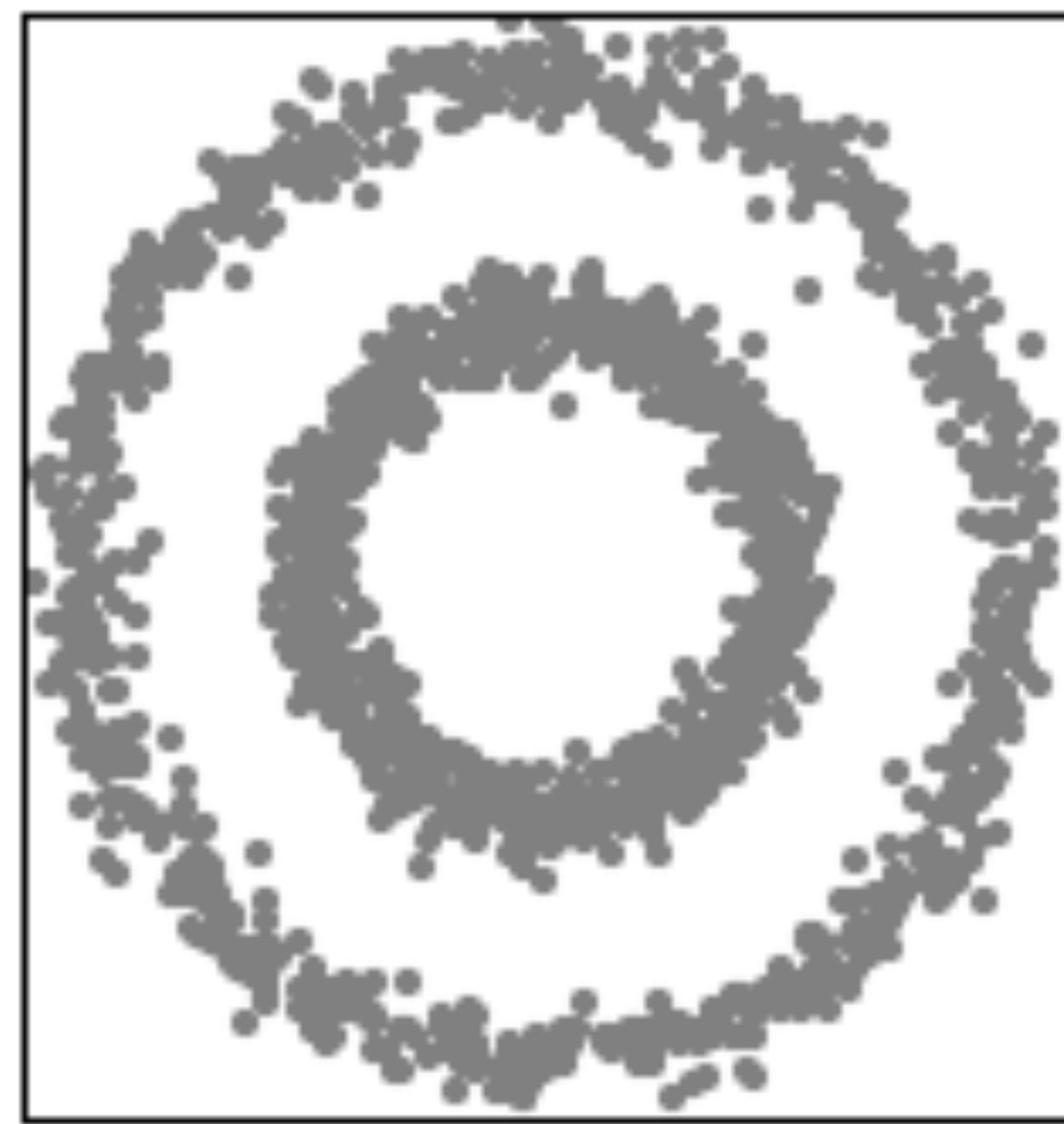
Example datasets



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:

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

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

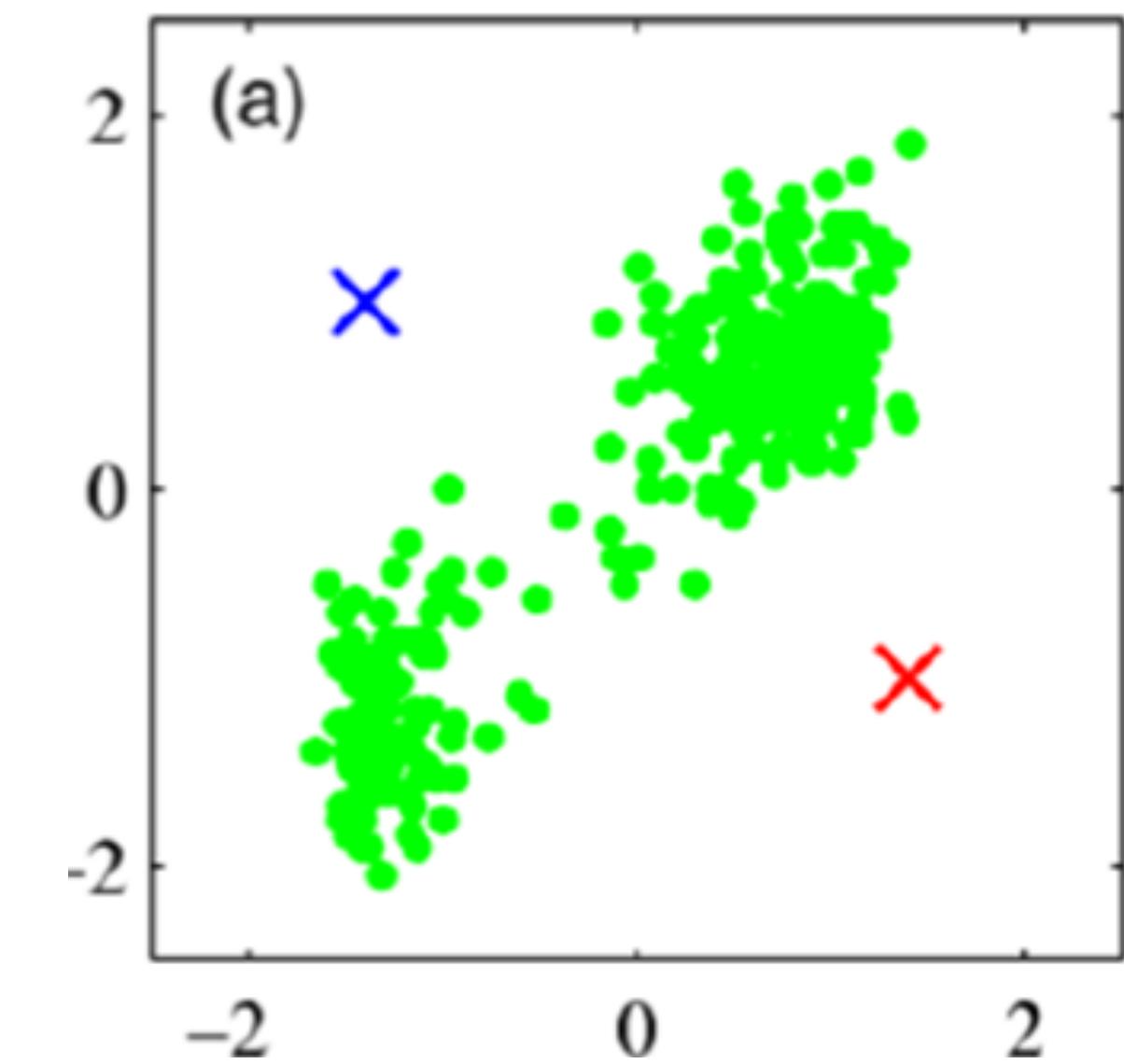
$$C_1, C_2, \dots, C_k$$

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 C_i defined as $\boldsymbol{\mu}_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$



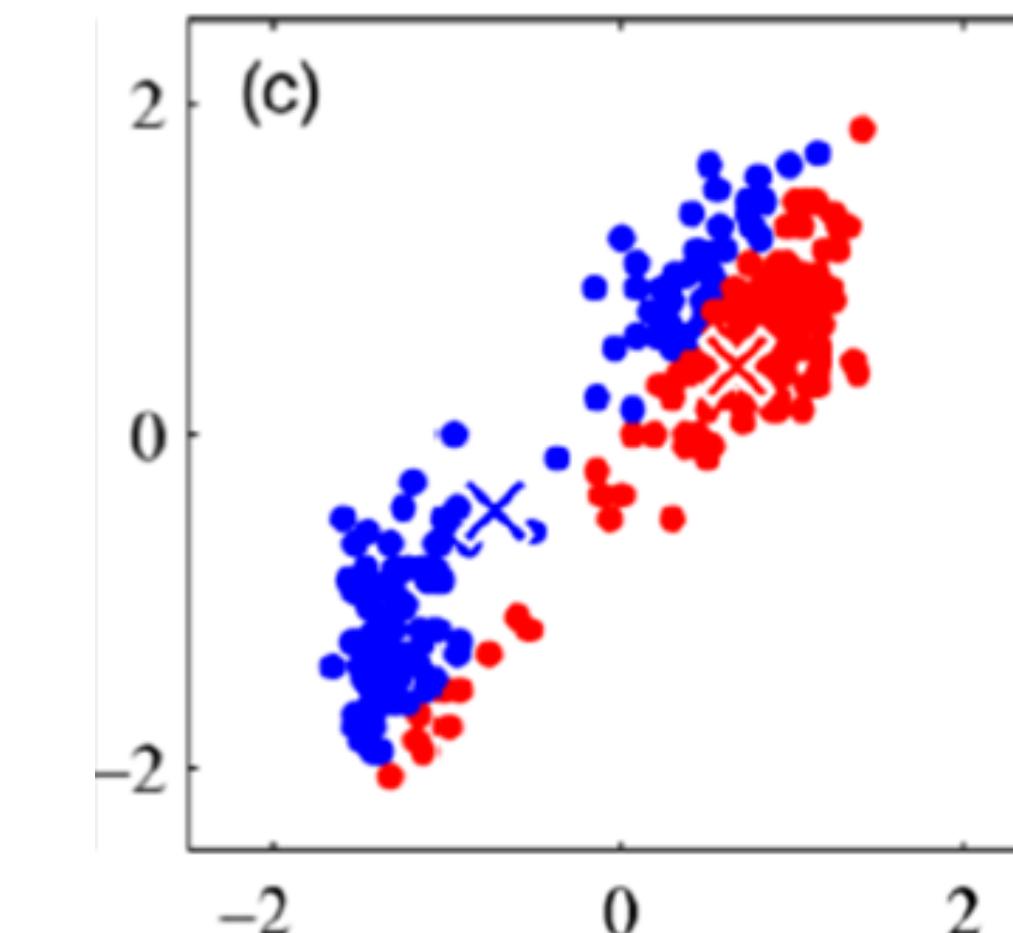
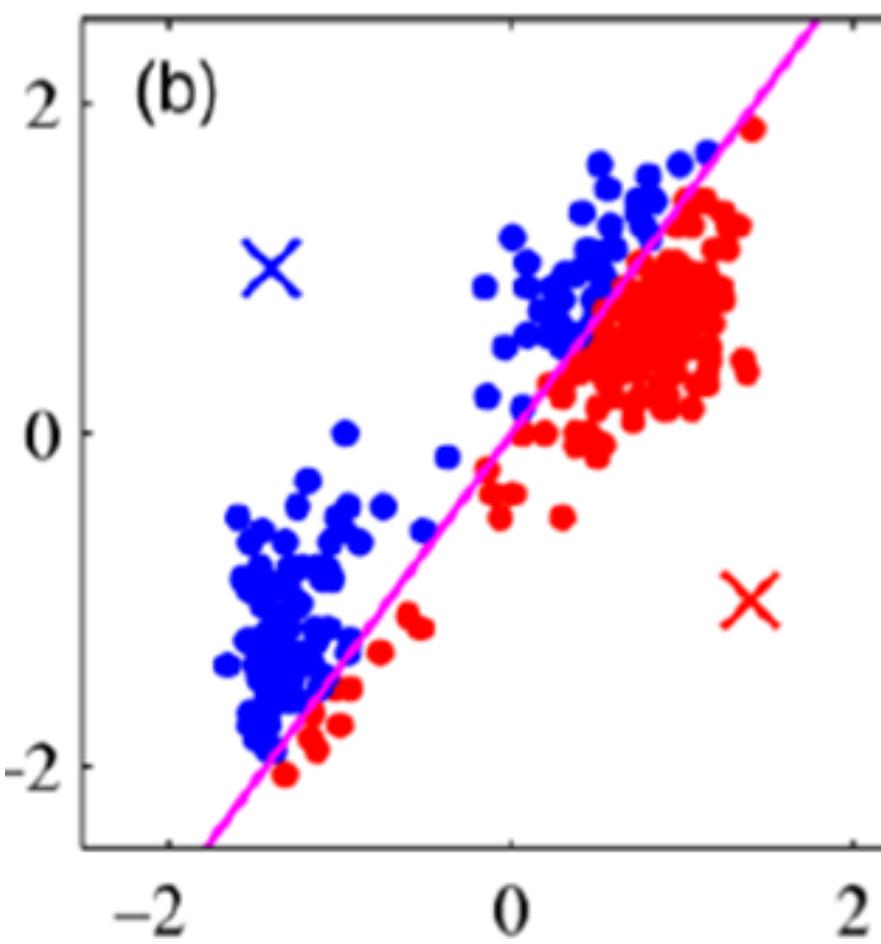
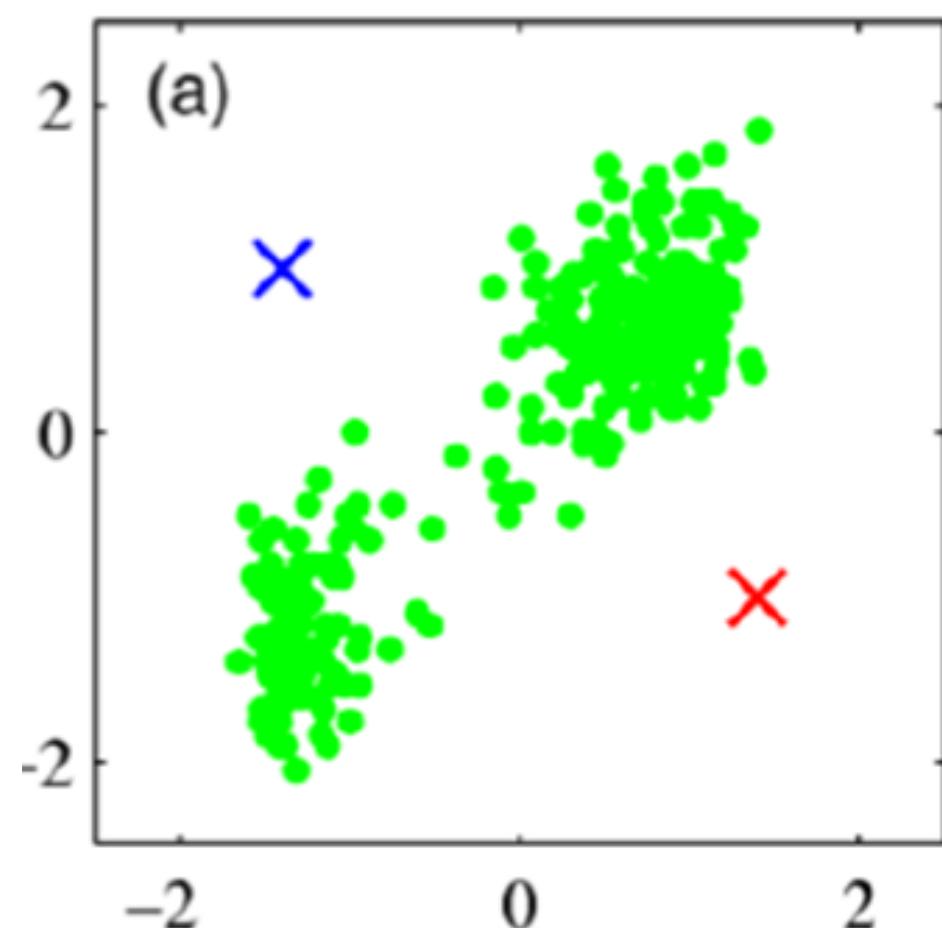
The algorithm's steps

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

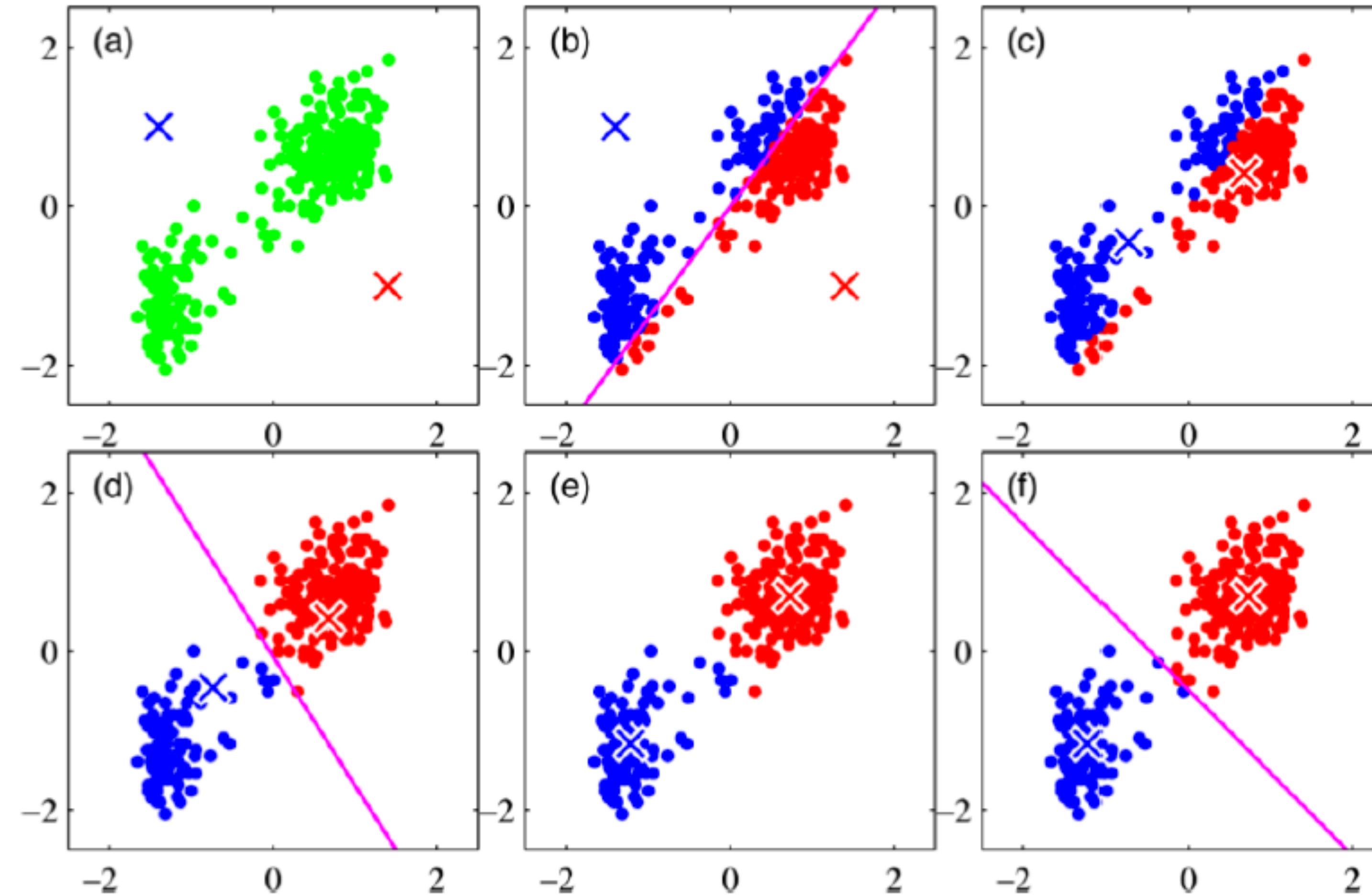
2. Assign each data point to its nearest centroid.

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

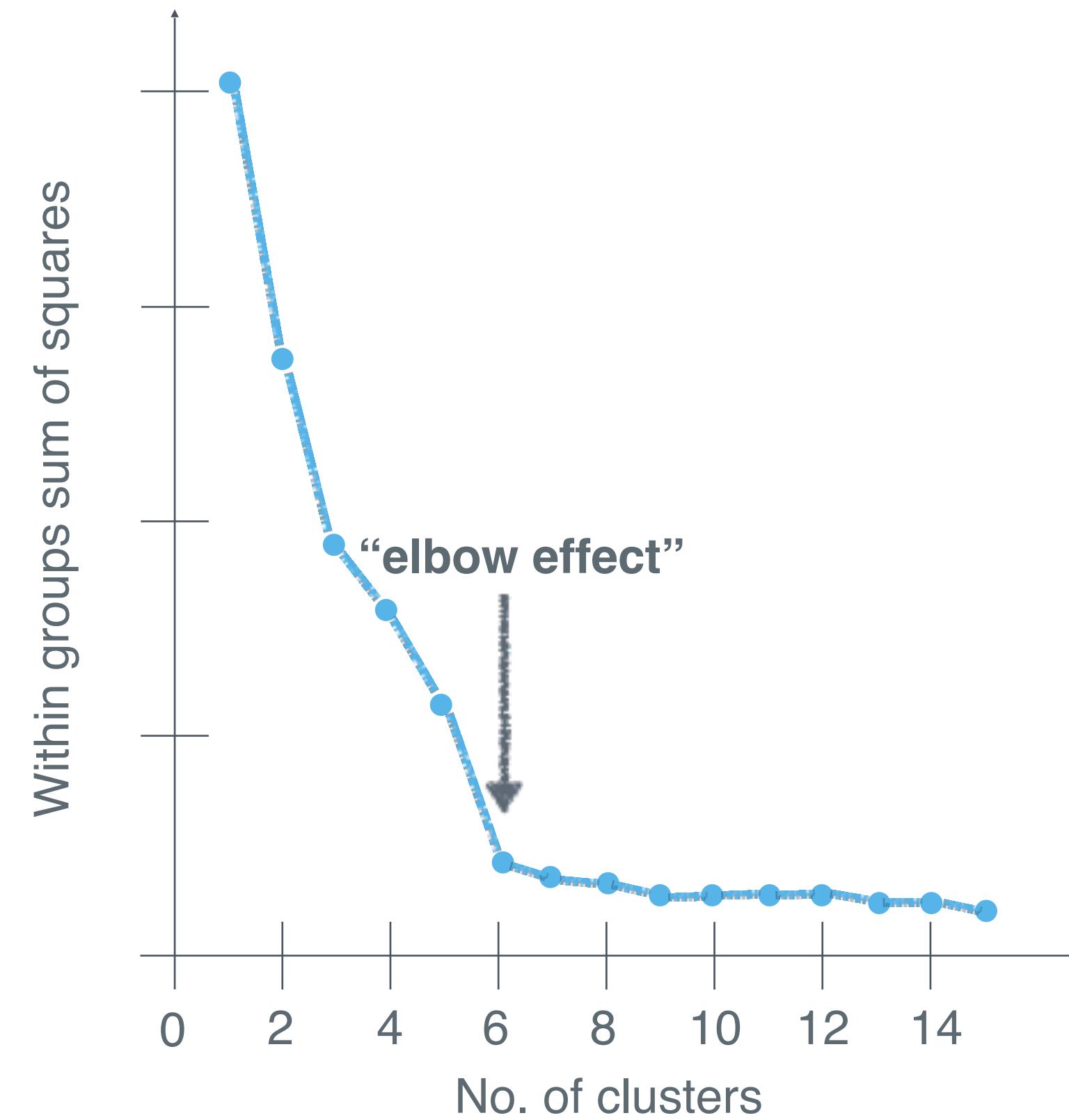
$$\sum_{i=1}^k \sum_{x \in C_i} d(\mathbf{x}, \boldsymbol{\mu}_i)$$

How do we select the number of clusters?

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

Potential problems:

- More than one elbow
- No elbow



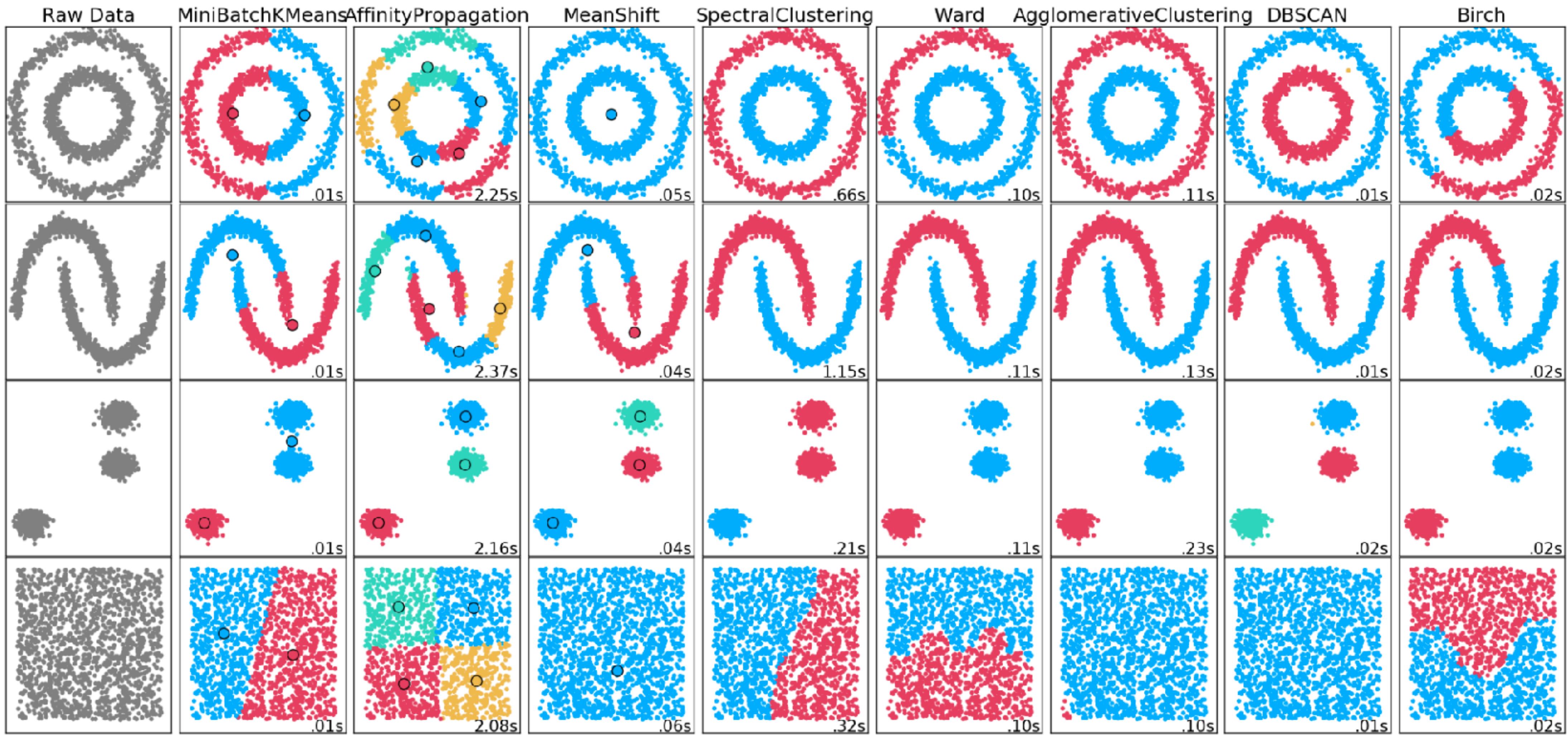
K-means drawbacks

- 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

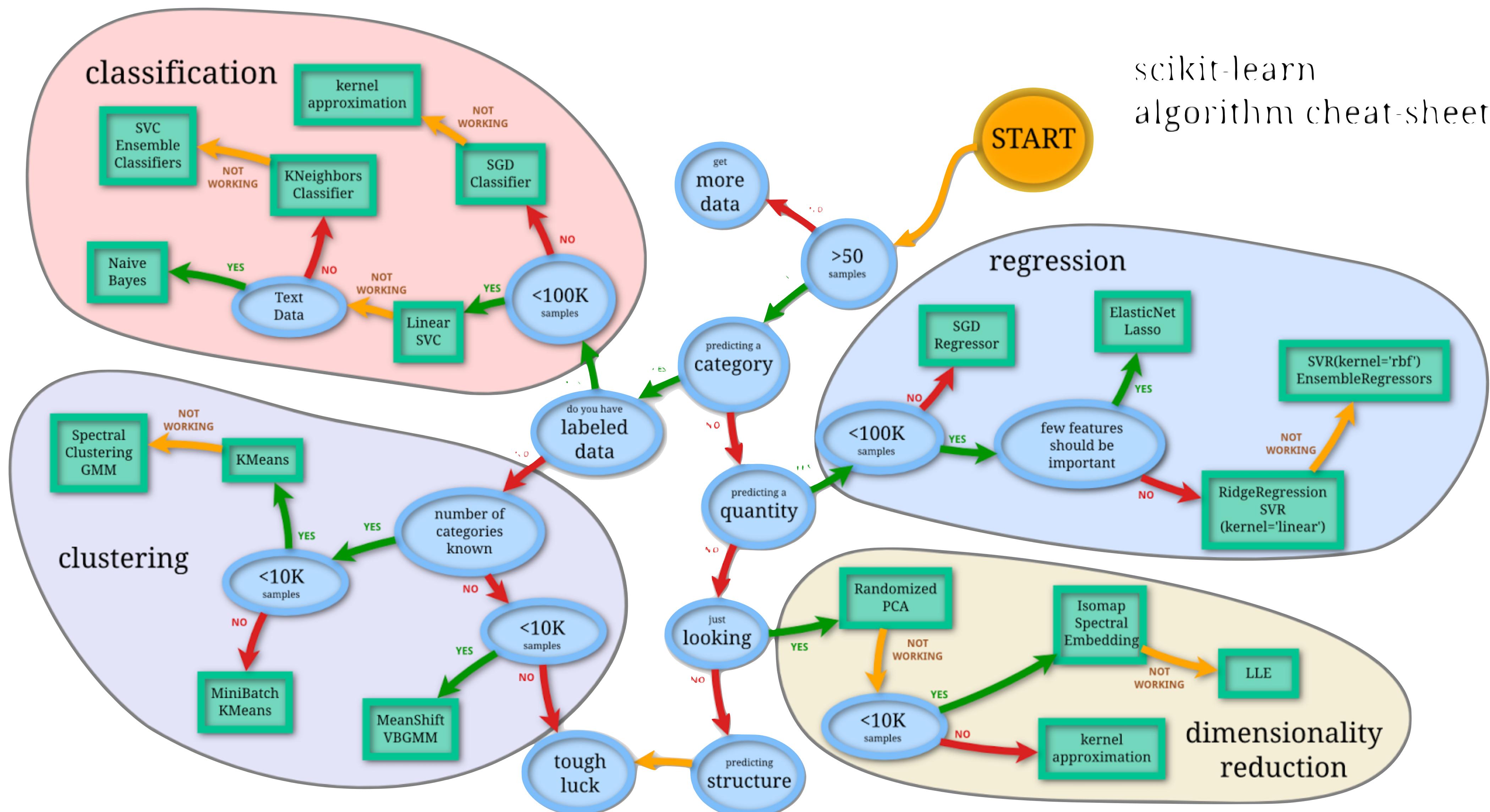
Meet your teacher



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

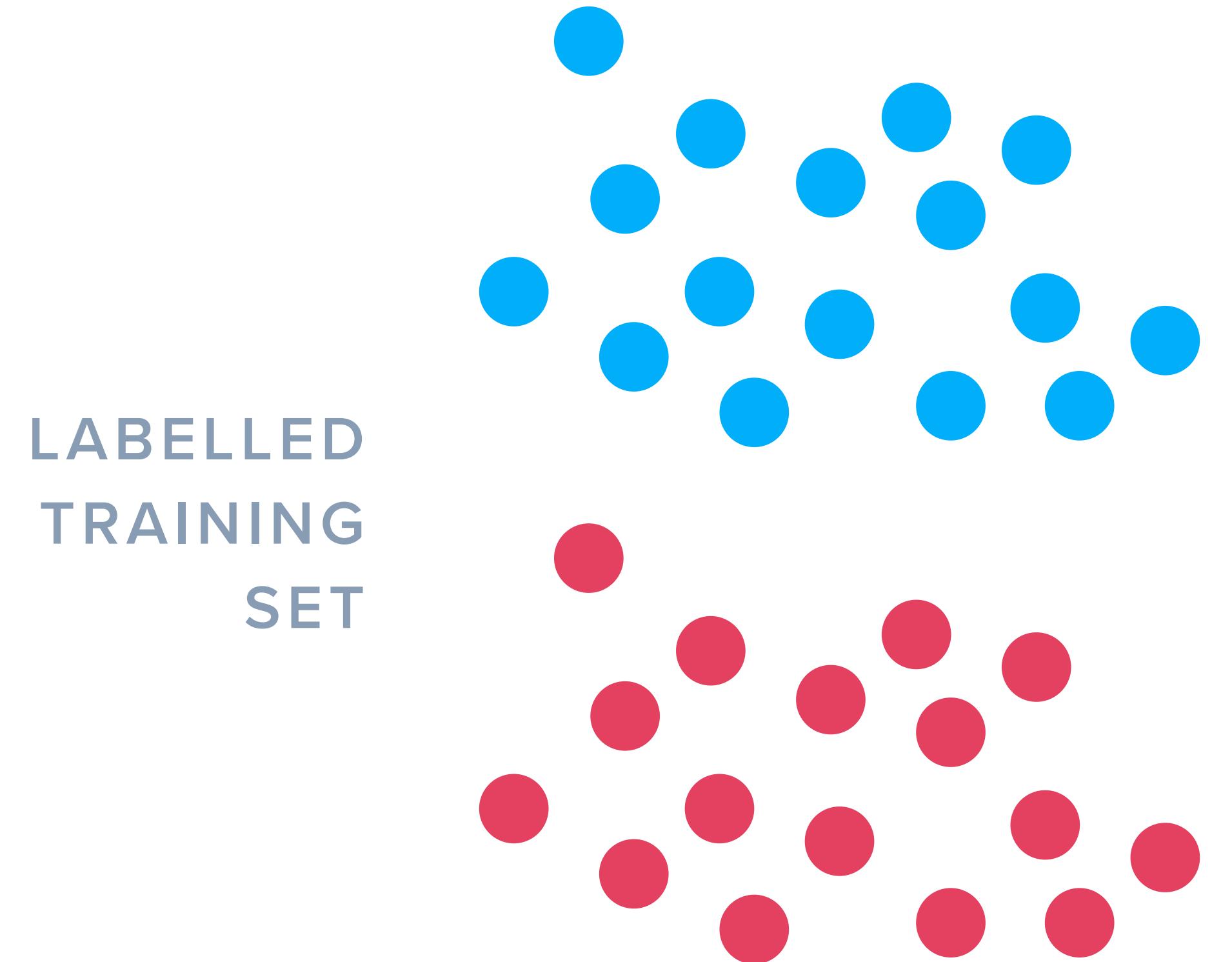


Scikit-learn workflow by Andreas Muller

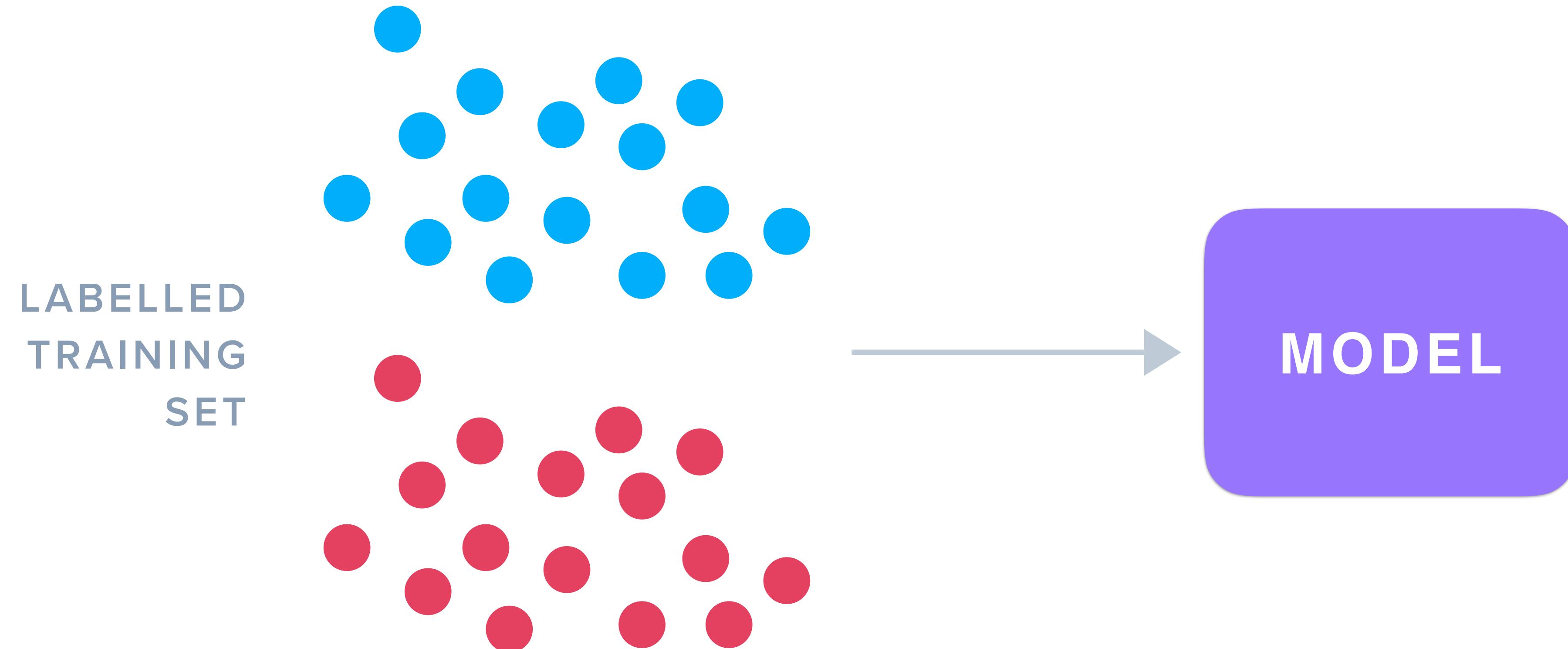
Classification



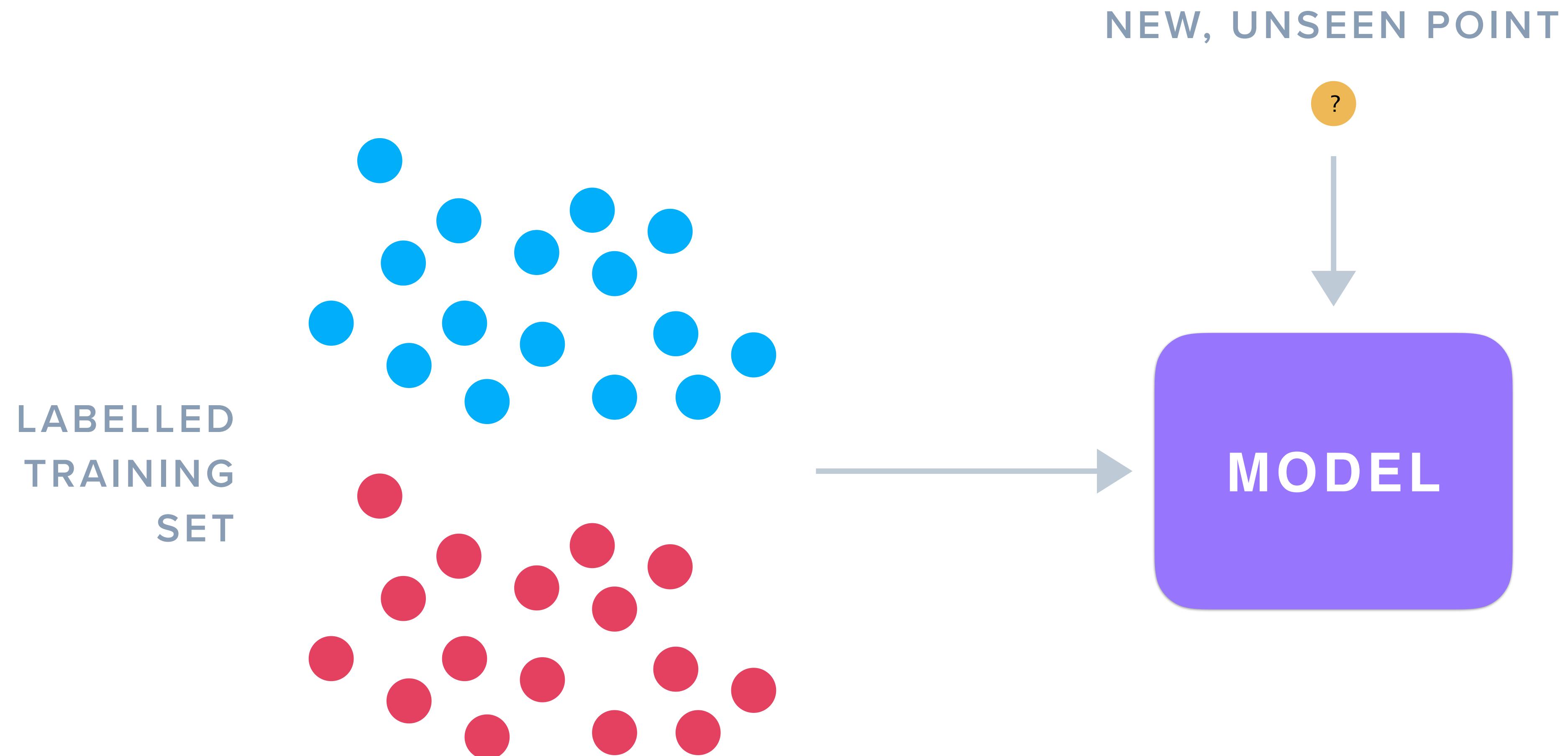
Classification



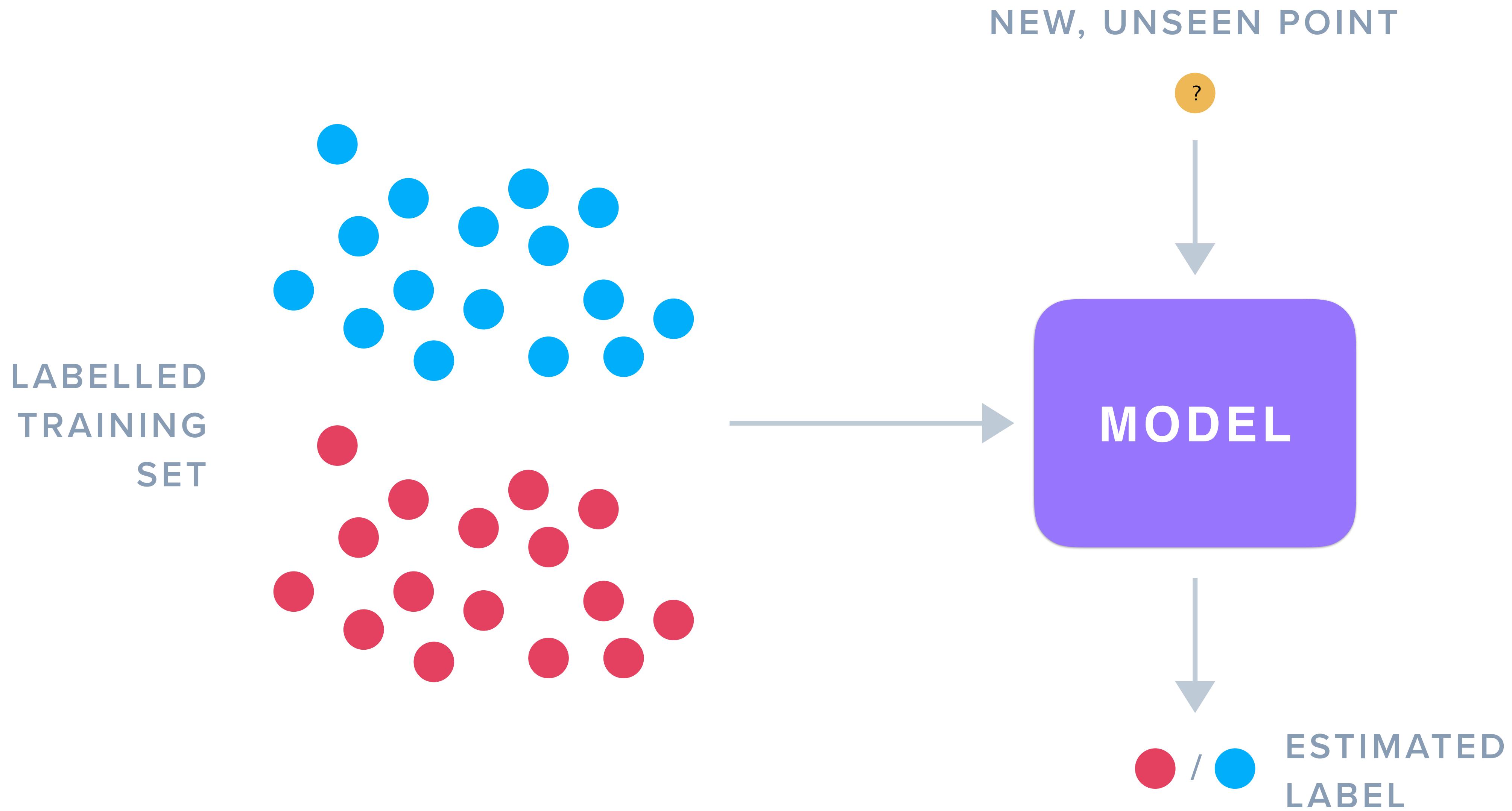
Classification



Classification

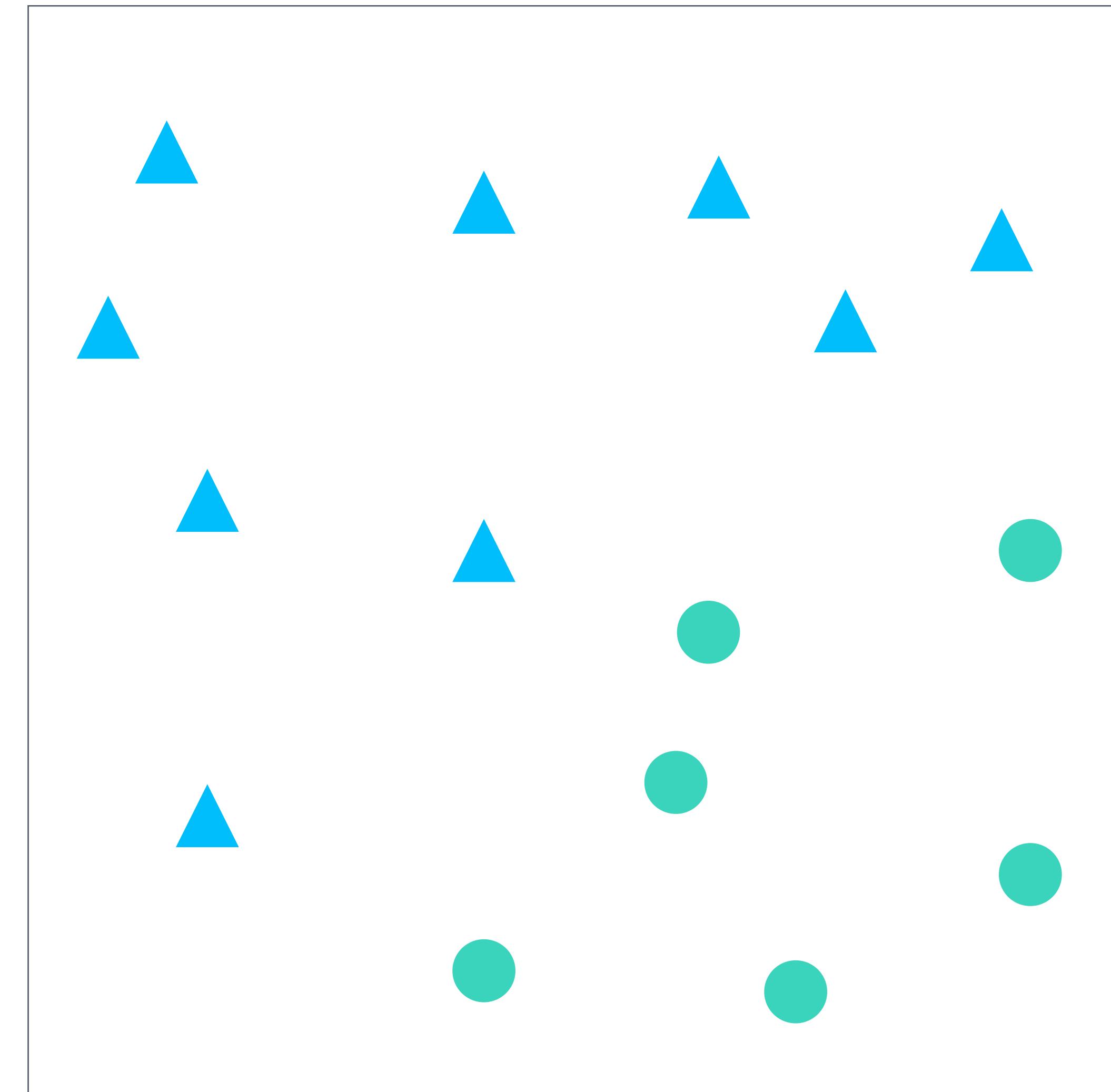


Classification



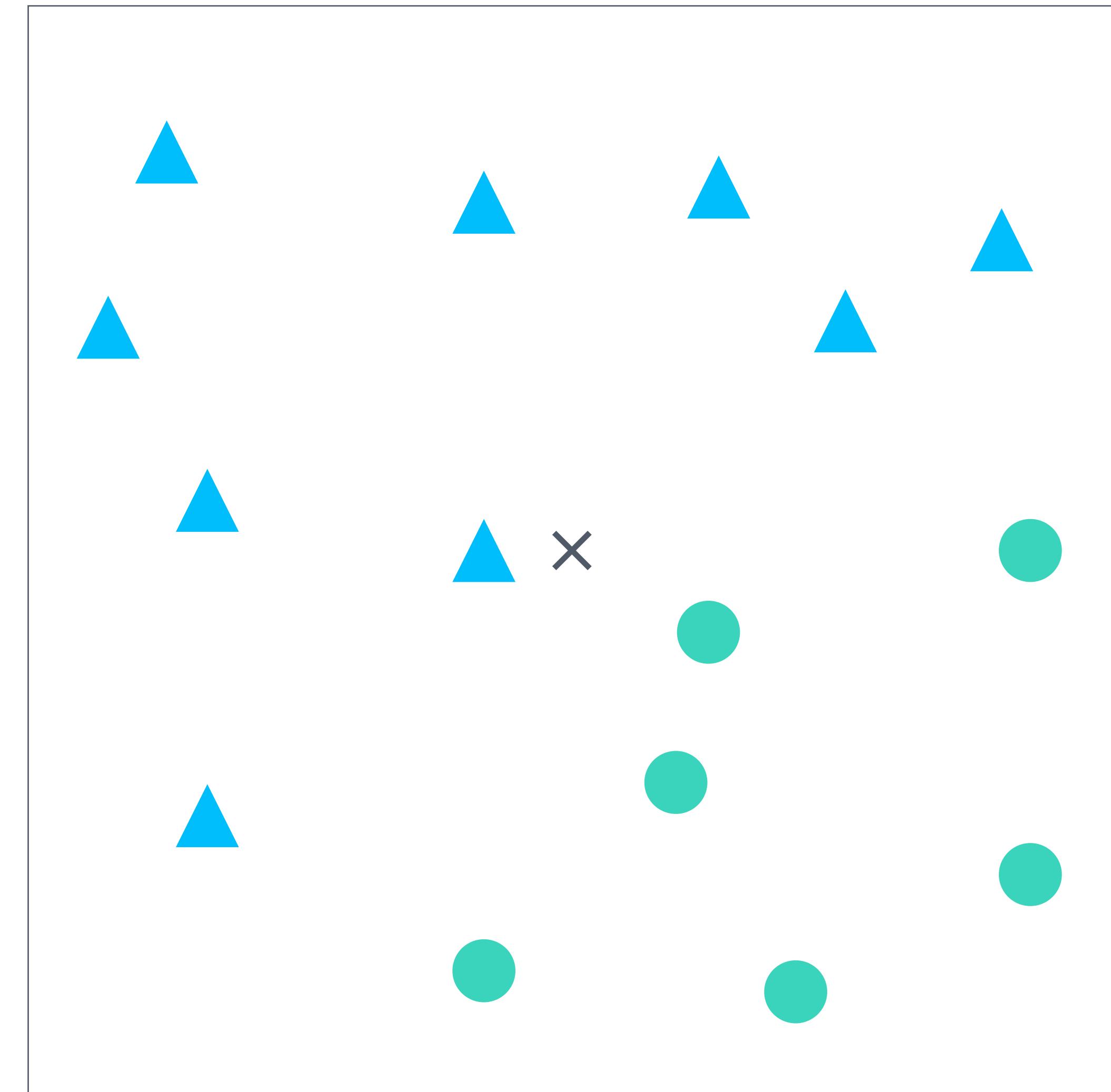
K-Nearest Neighbours

K-Nearest Neighbours (KNN)



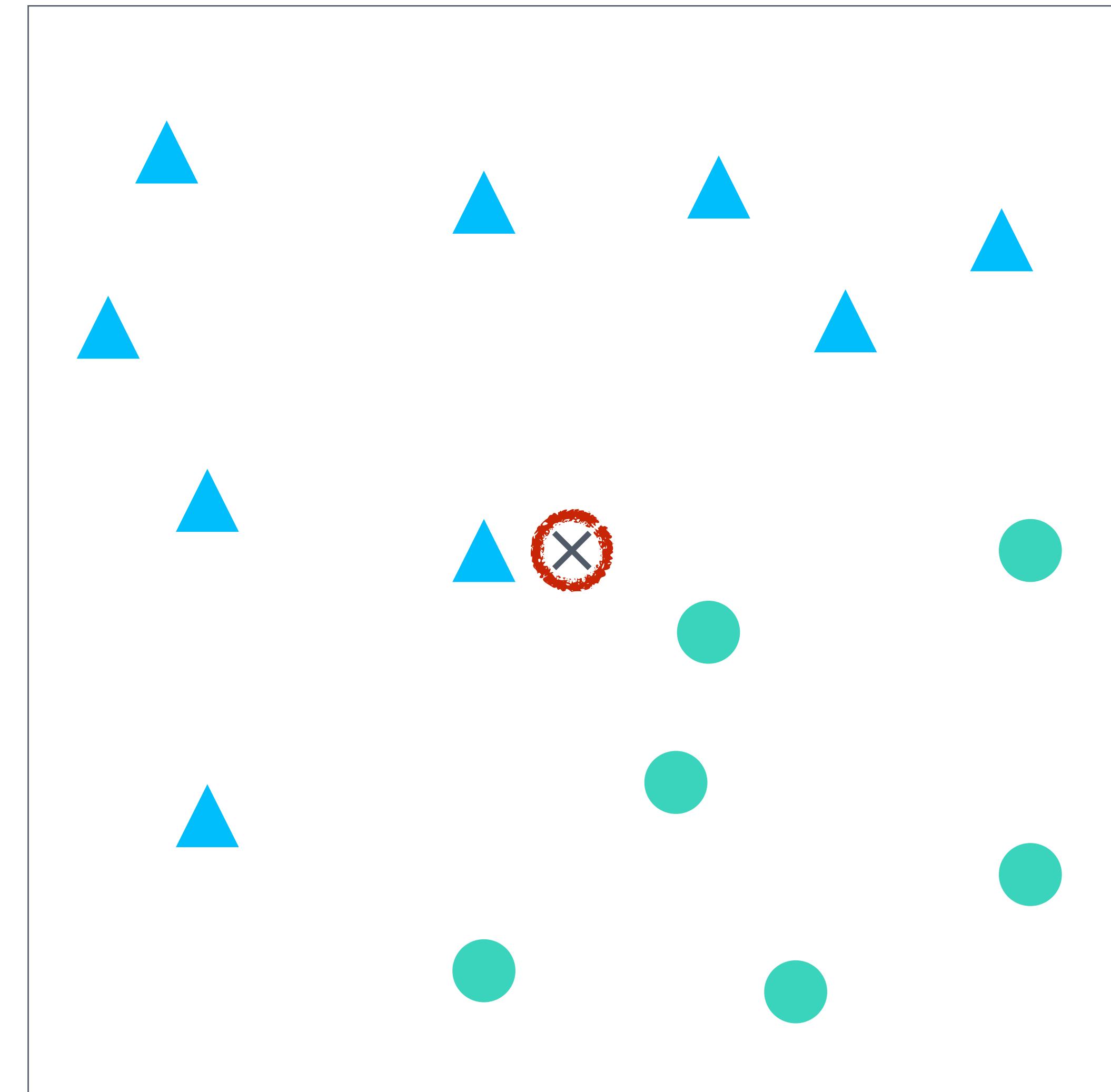
K-Nearest Neighbours

K-Nearest Neighbours (KNN)



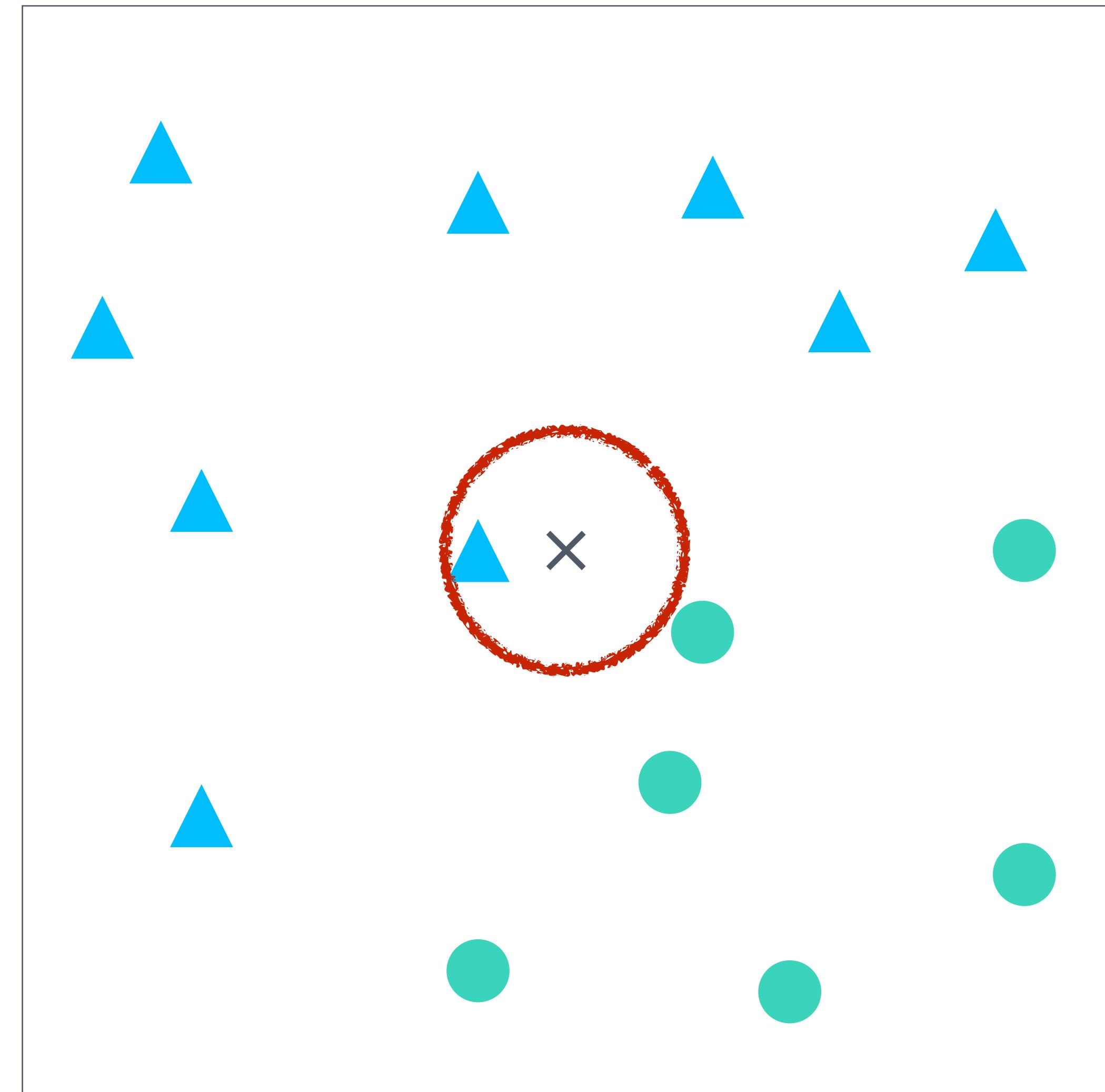
K-Nearest Neighbours

K-Nearest Neighbours (KNN)



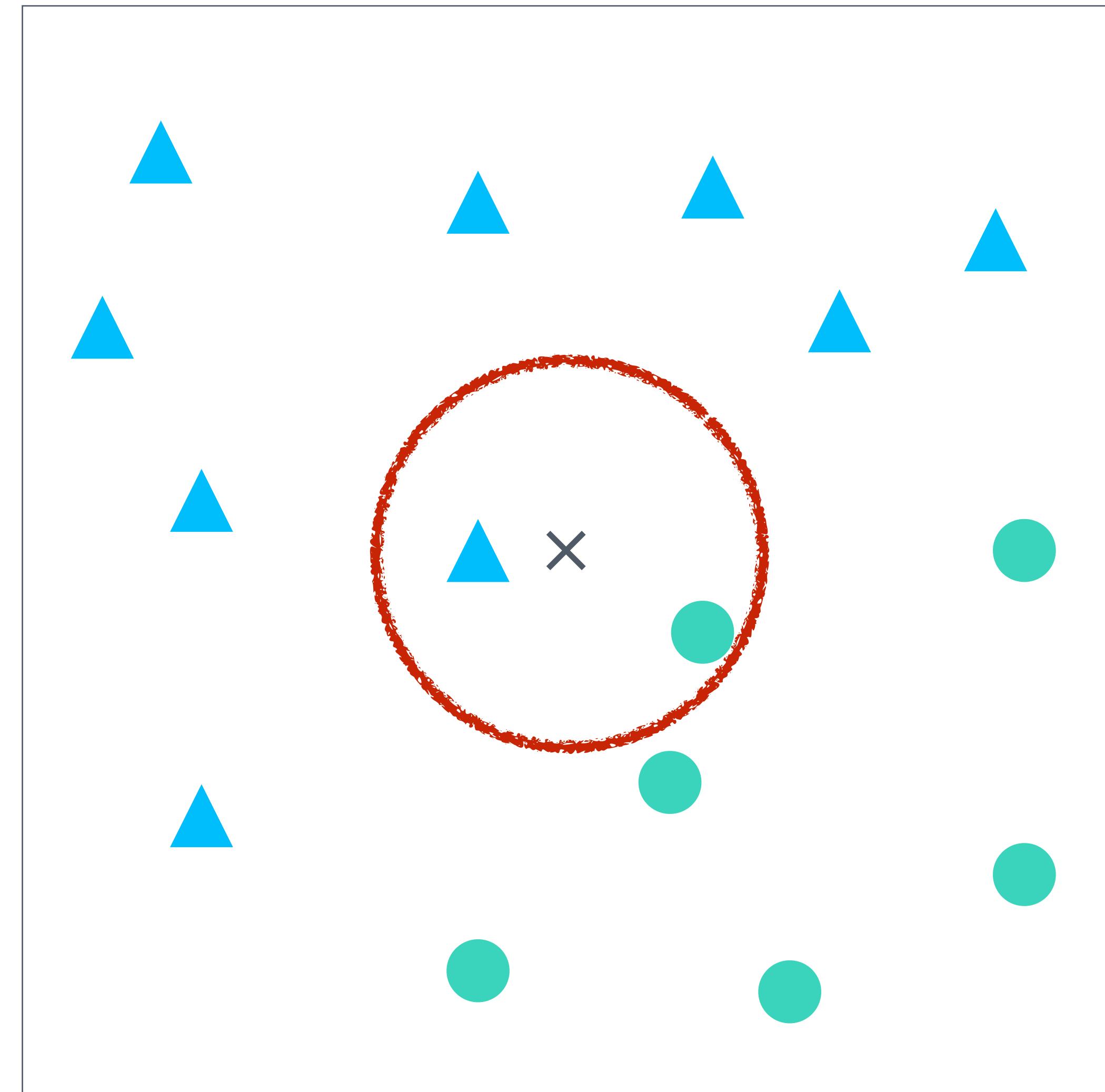
K-Nearest Neighbours

$K = 1$



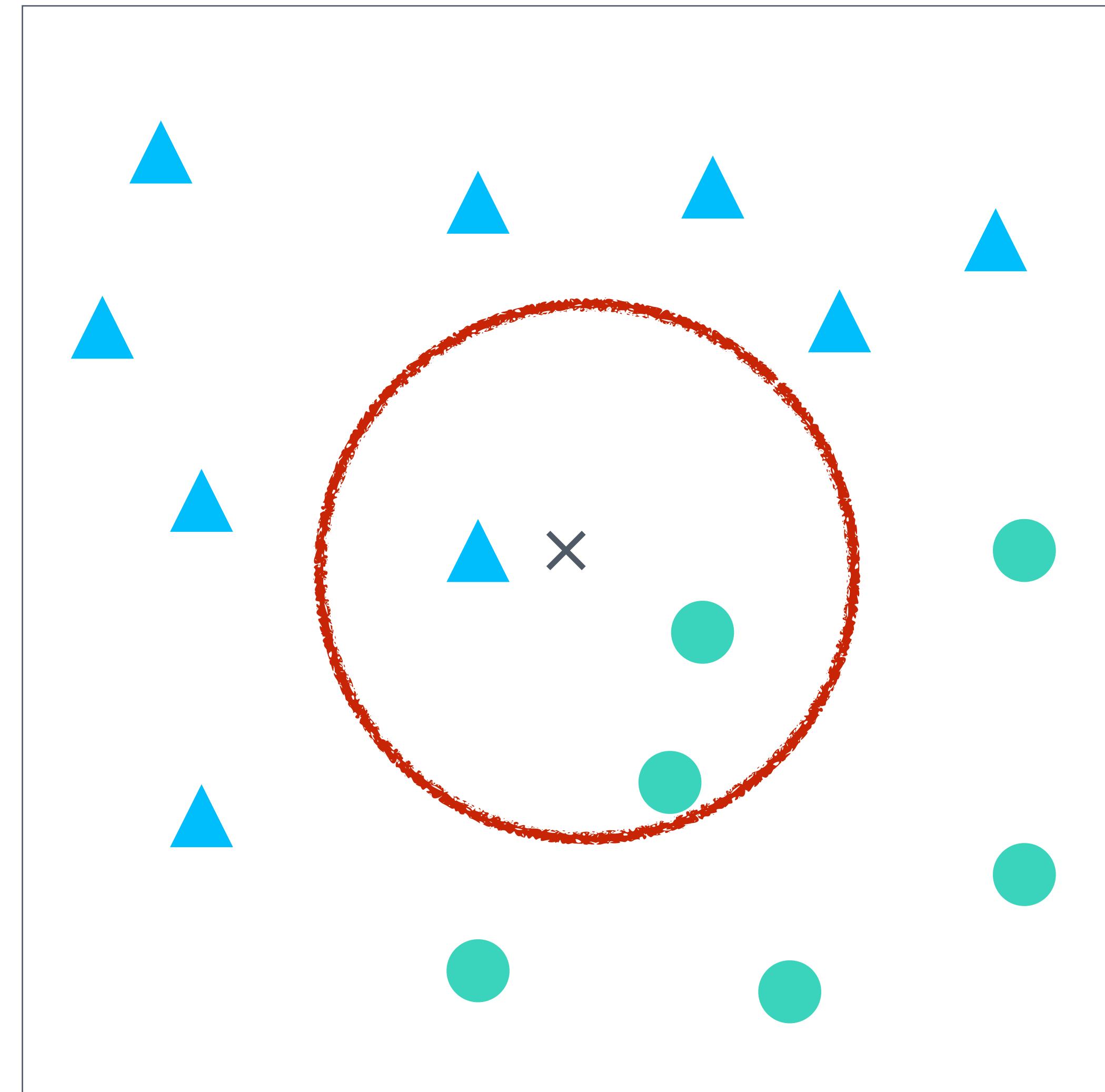
K-Nearest Neighbours

$K = 2$



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

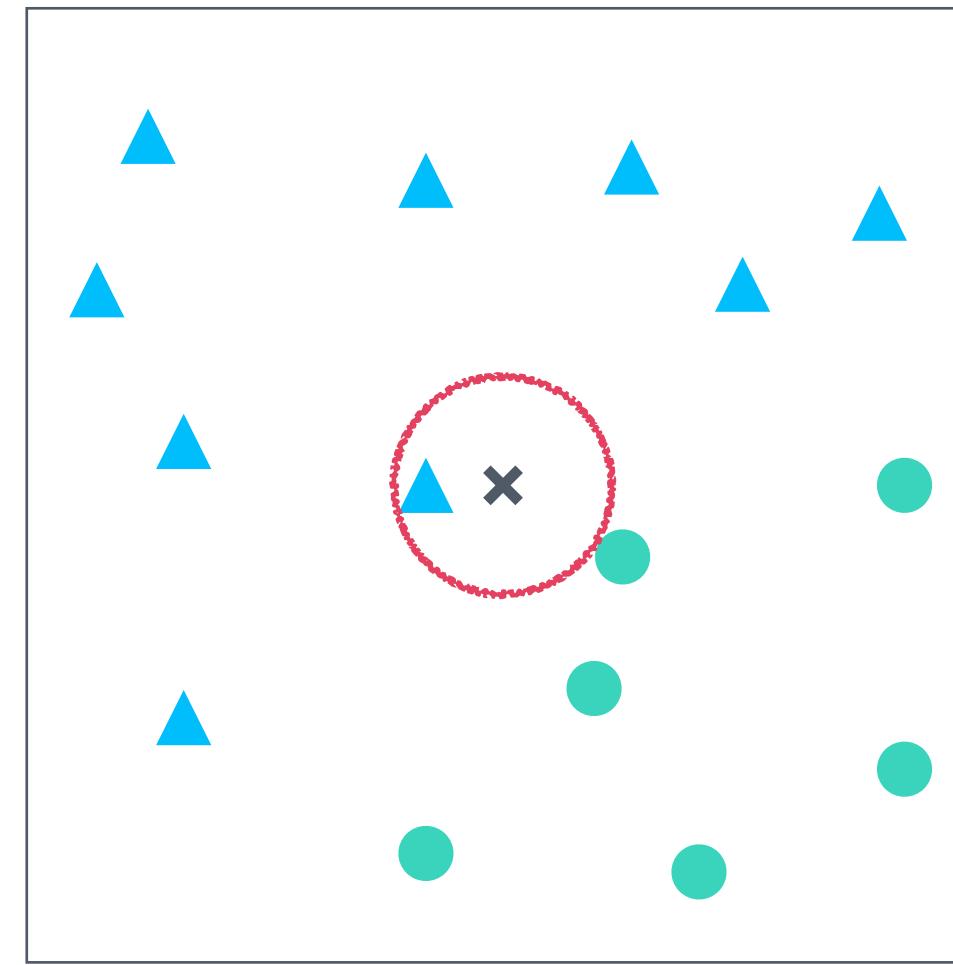
- The input training data
- The number k of nearest neighbours
- A metric to compute the distance between instances

Algorithm's Steps

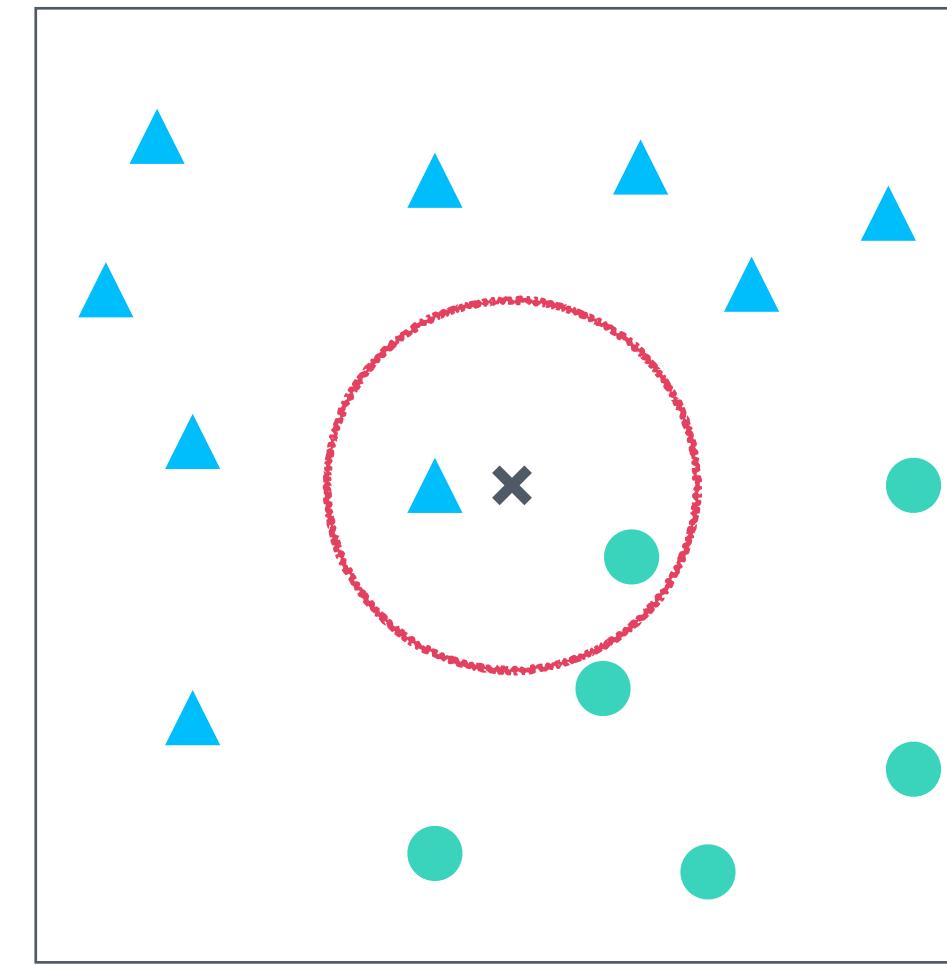
Classification of a new unknown, unlabelled test example:

- 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

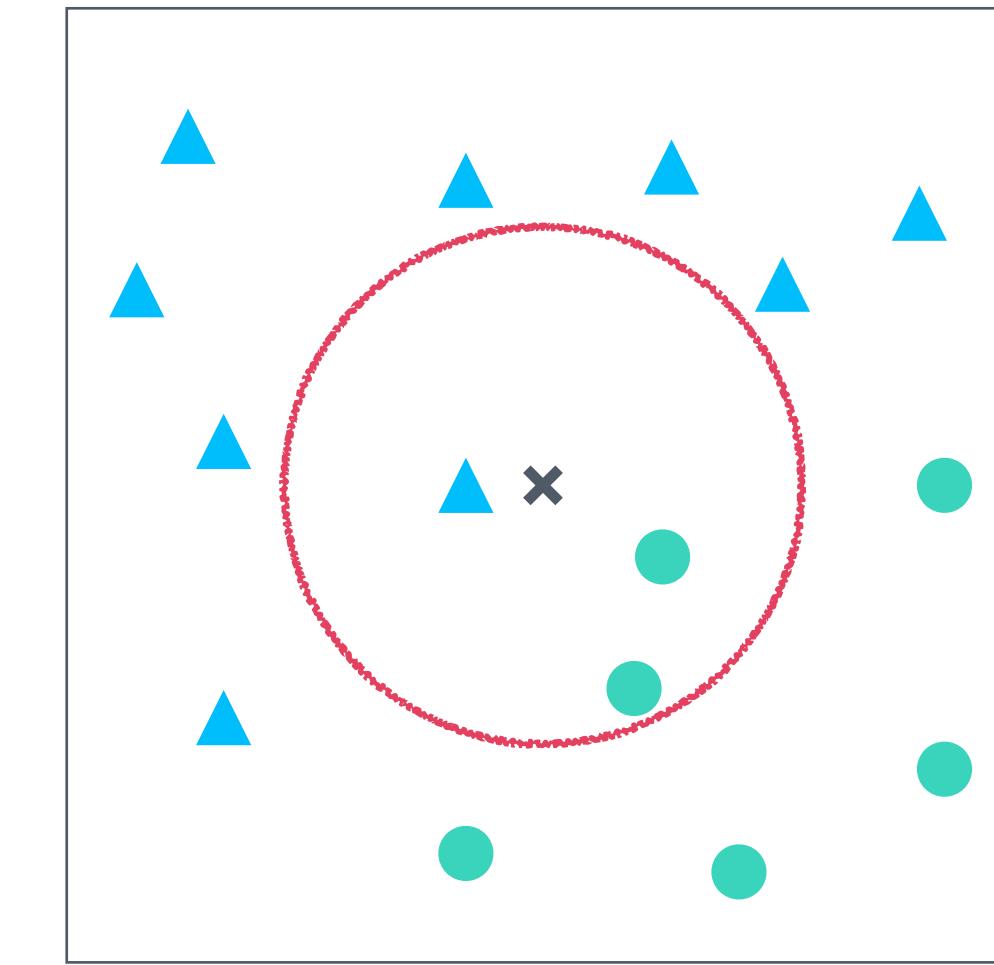
K-NN: 1-NN, 2-NN, 3-NN



1-nearest neighbour



2-nearest neighbour



3-nearest neighbour

$k=1$
 $k=2$
 $k=3$



Which K?

A couple of details:

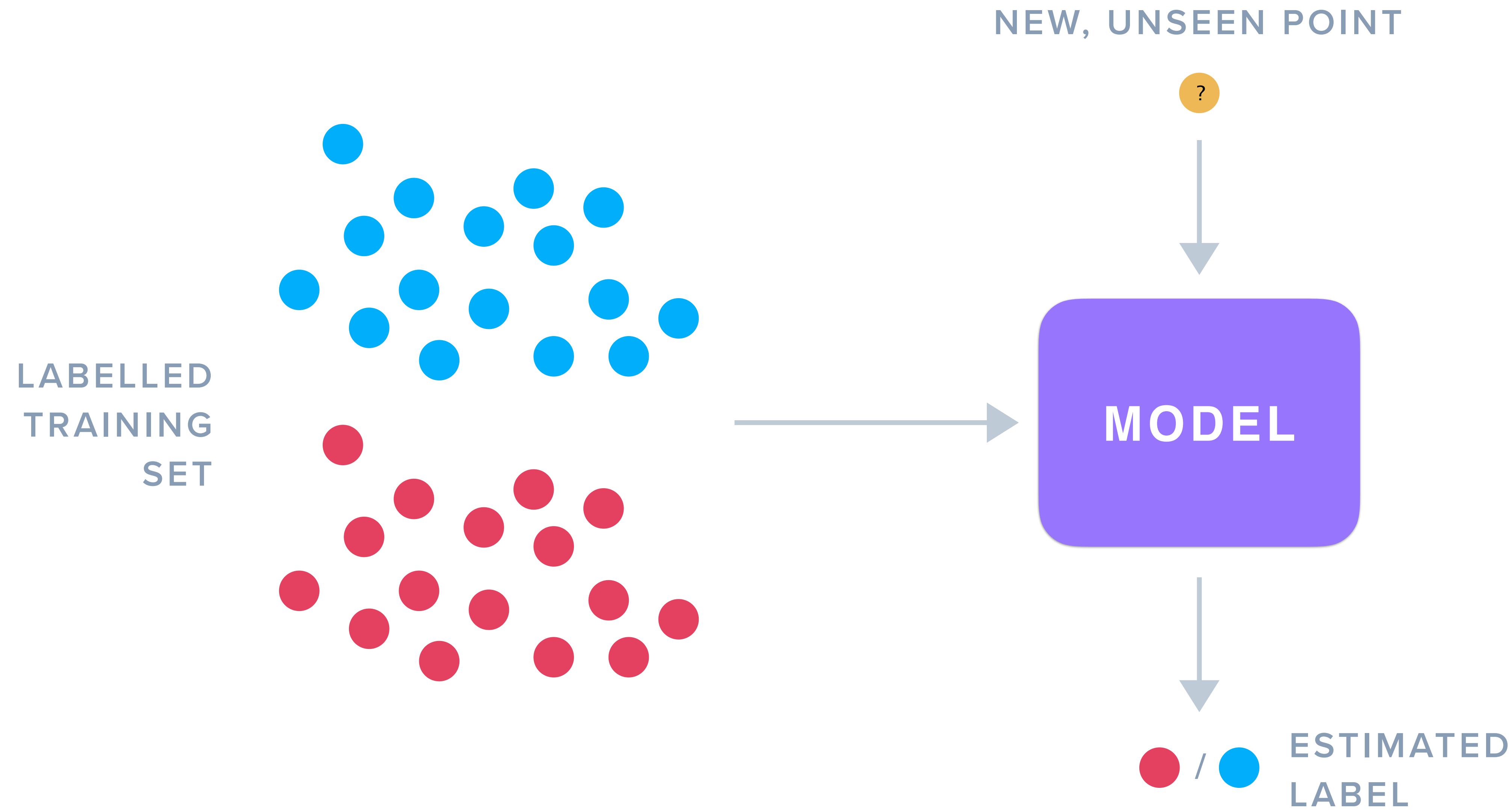
- In binary classification, it is advisable to choose an odd number of K
- During voting, neighbours can have equal influence or different weights according to distance

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

	POSITIVE	NEGATIVE	REALLY IS
POSITIVE	TRUE POSITIVE TP	FALSE NEGATIVE FN	POSITIVE
NEGATIVE	FALSE POSITIVE FP	TRUE NEGATIVE TN	NEGATIVE

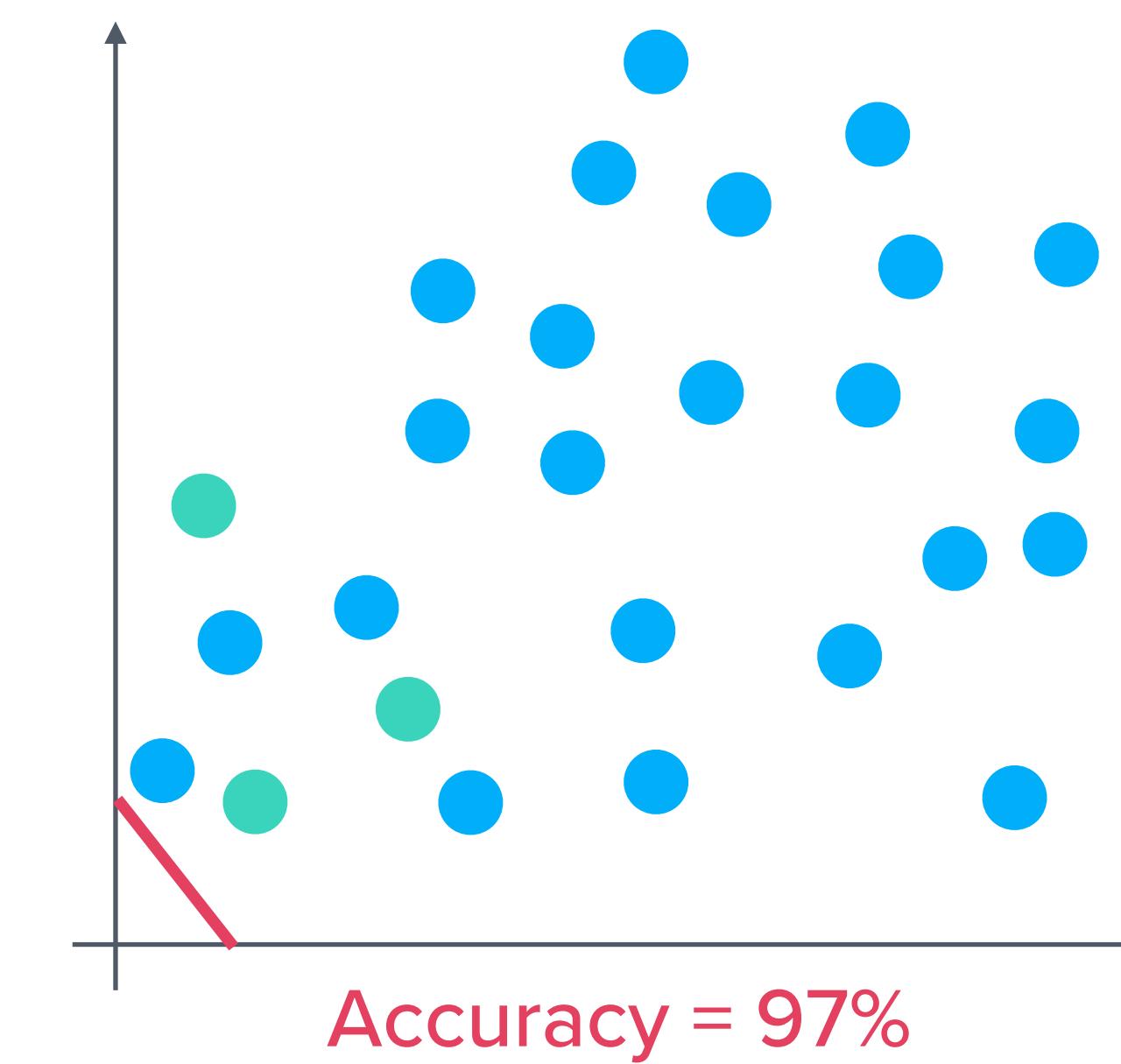
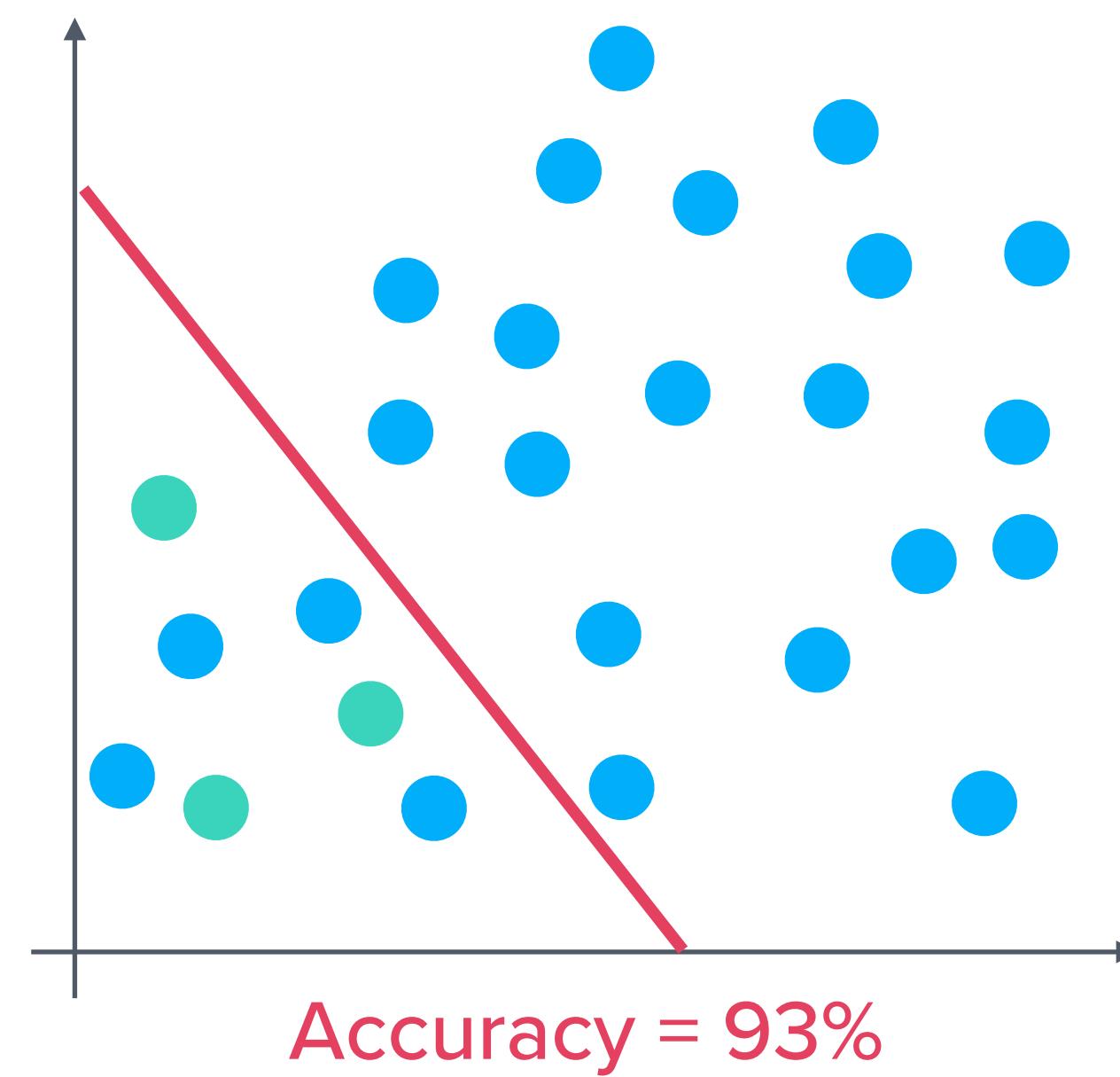
Confusion matrix: number of correct and
incorrect predictions

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



Model Evaluation

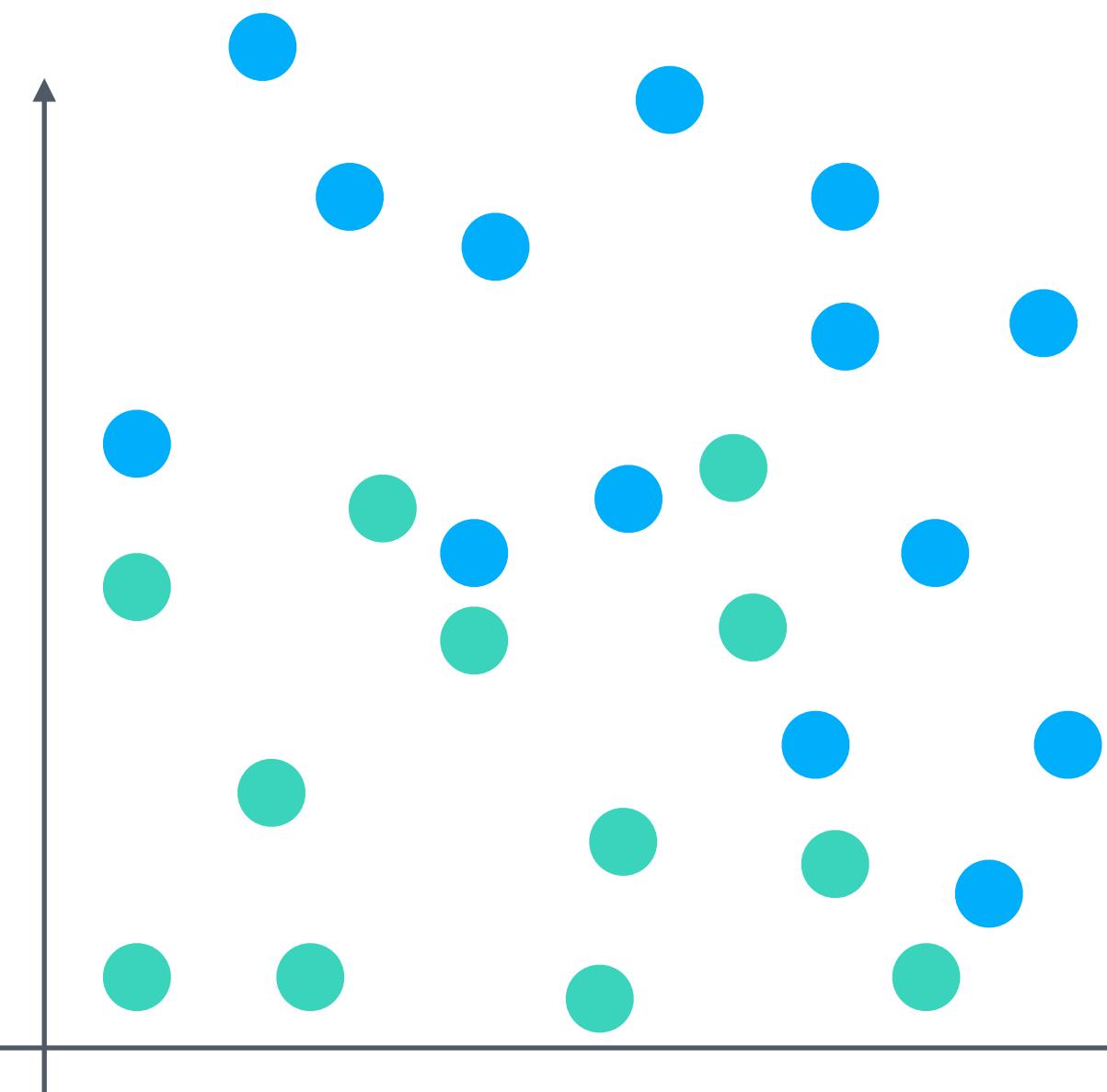
- **Recall or sensitivity:** the proportion of actual positive cases which are correctly identified
- **Precision:** the proportion of positive cases that were correctly identified
- **F-score:** combination of the two

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

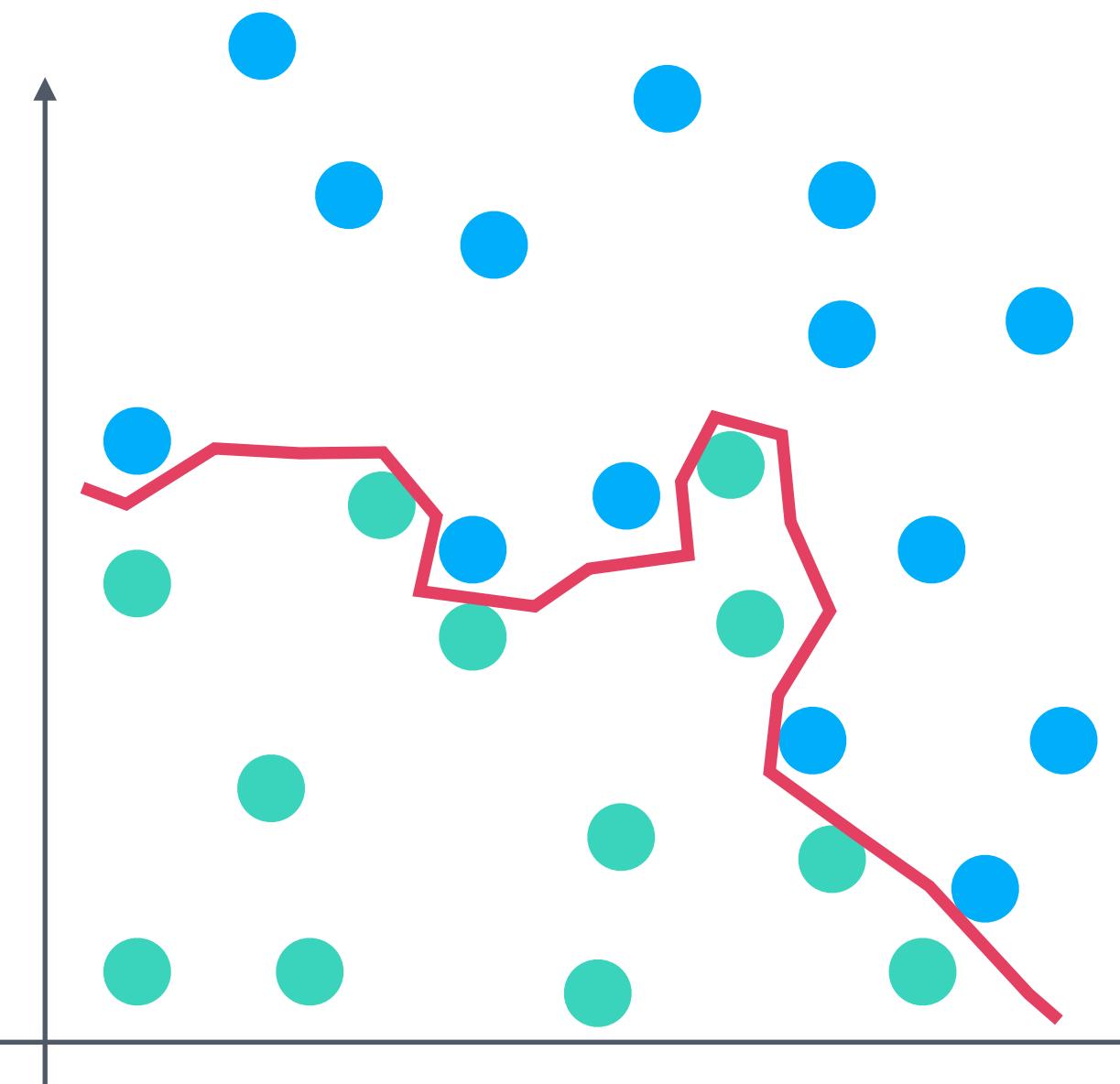
$$\frac{TP}{TP + FP}$$

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

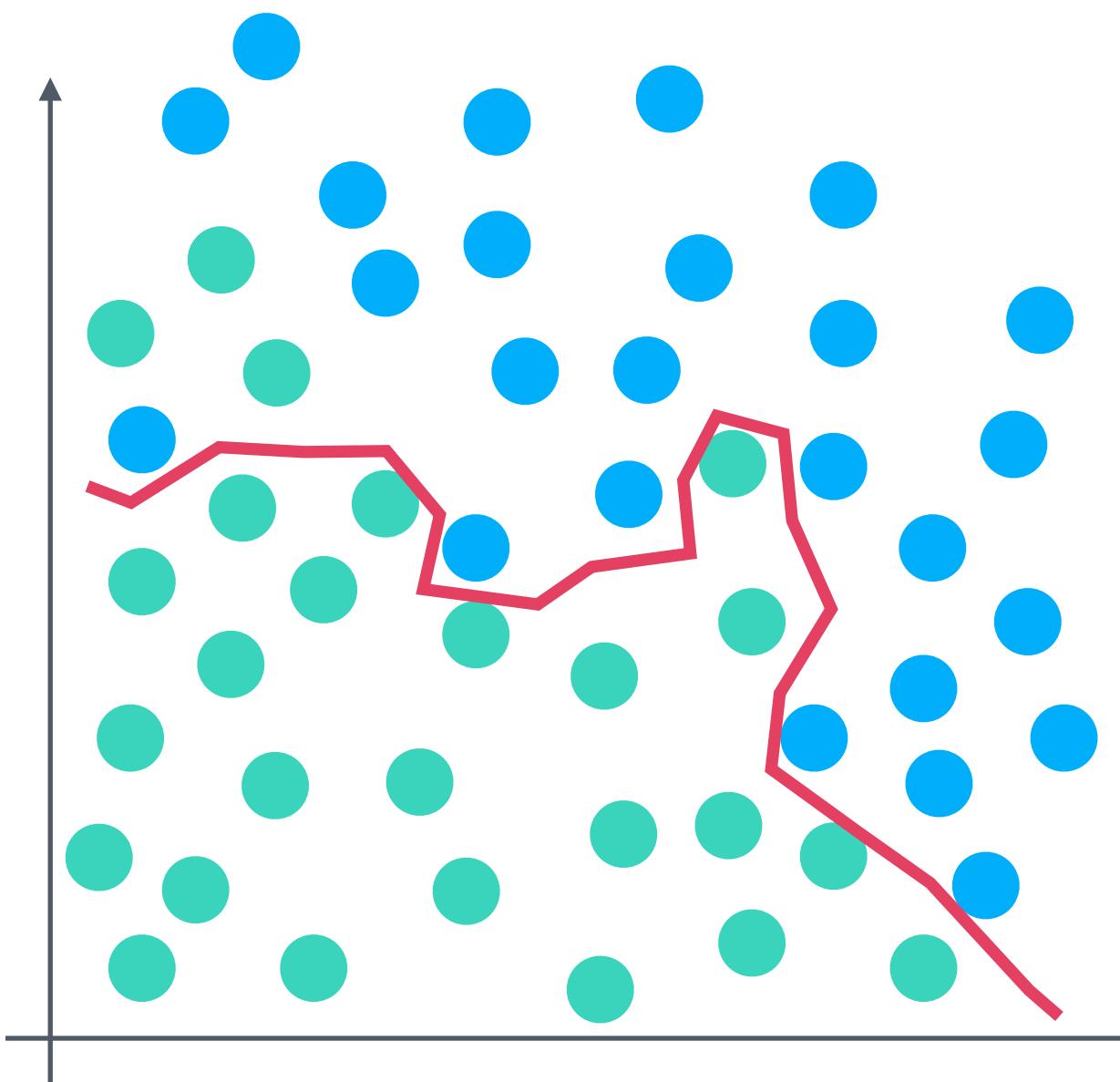
Sometimes we fall in love with the data



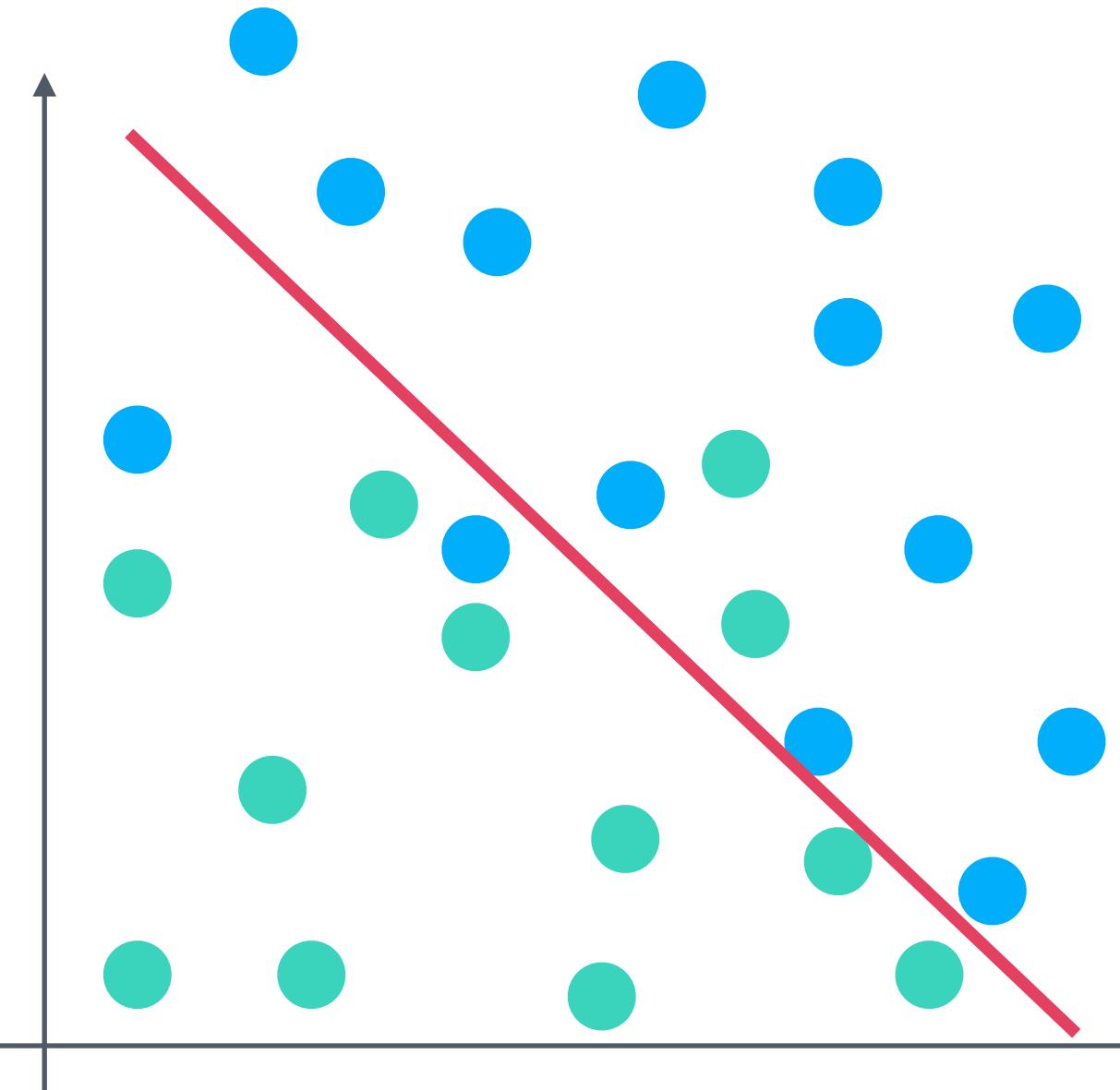
Sometimes we fall in love with the data



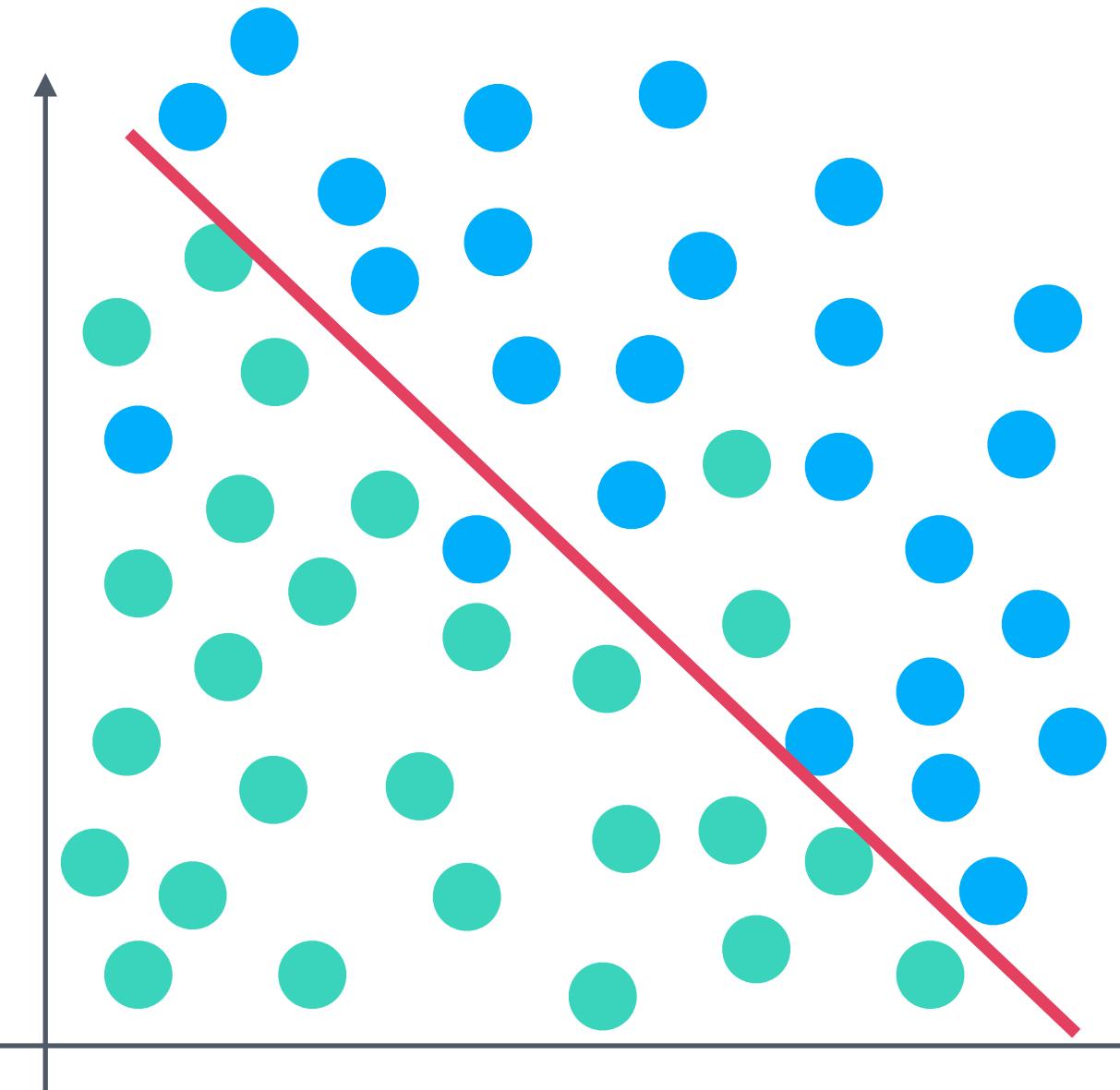
Sometimes we fall in love with the data



Sometimes we fall in love with the data

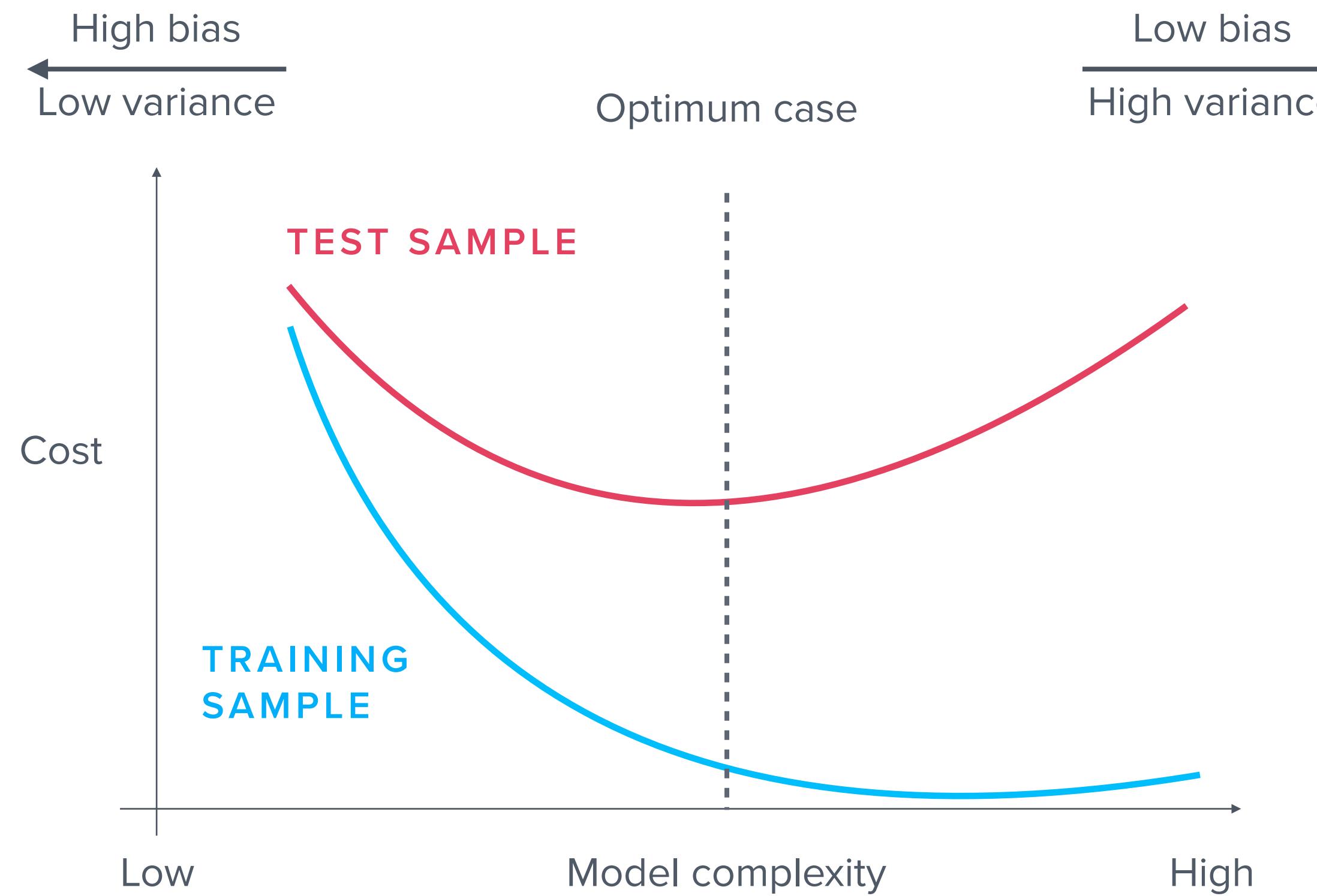


Sometimes we fall in love with the data

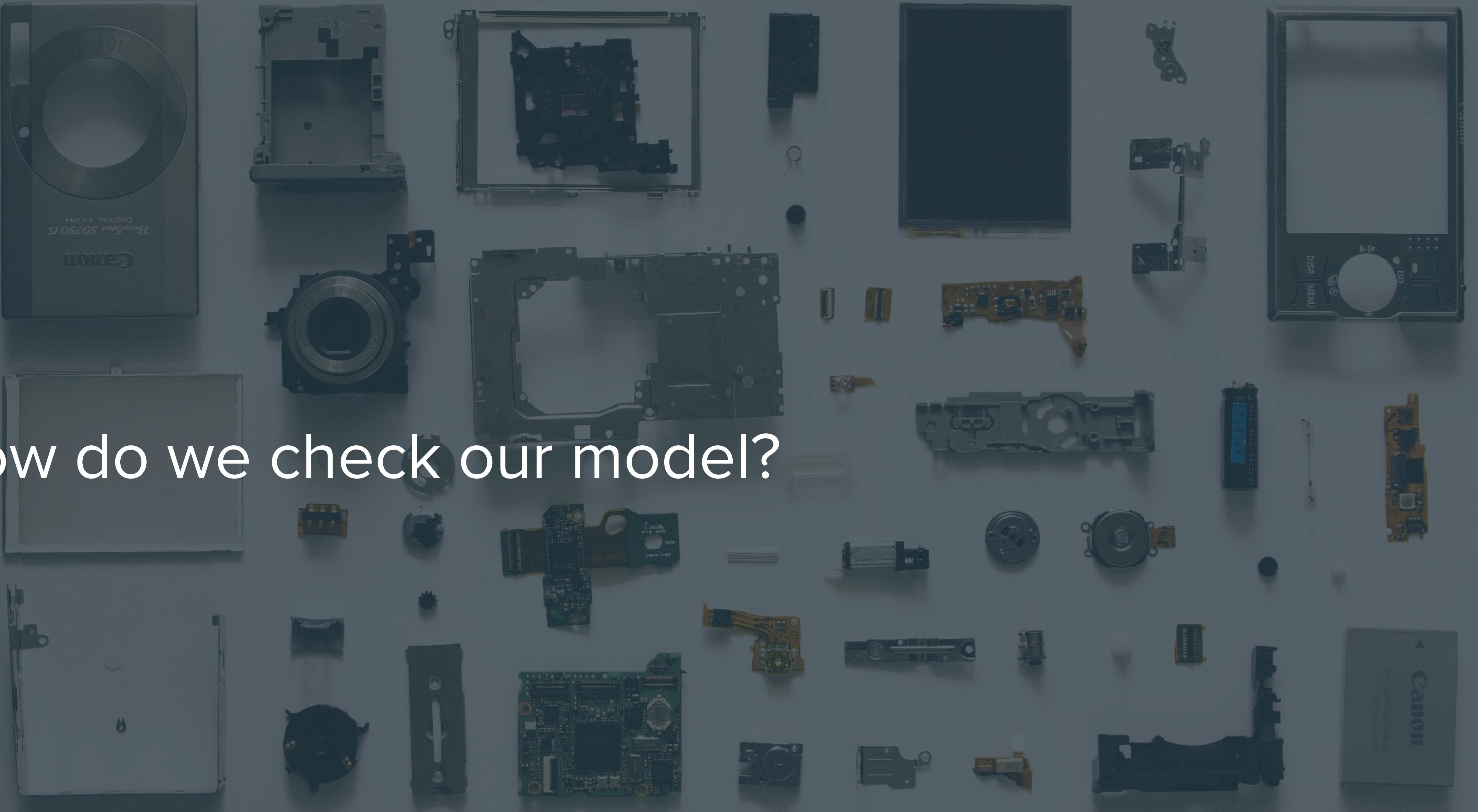


Model Evaluation

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

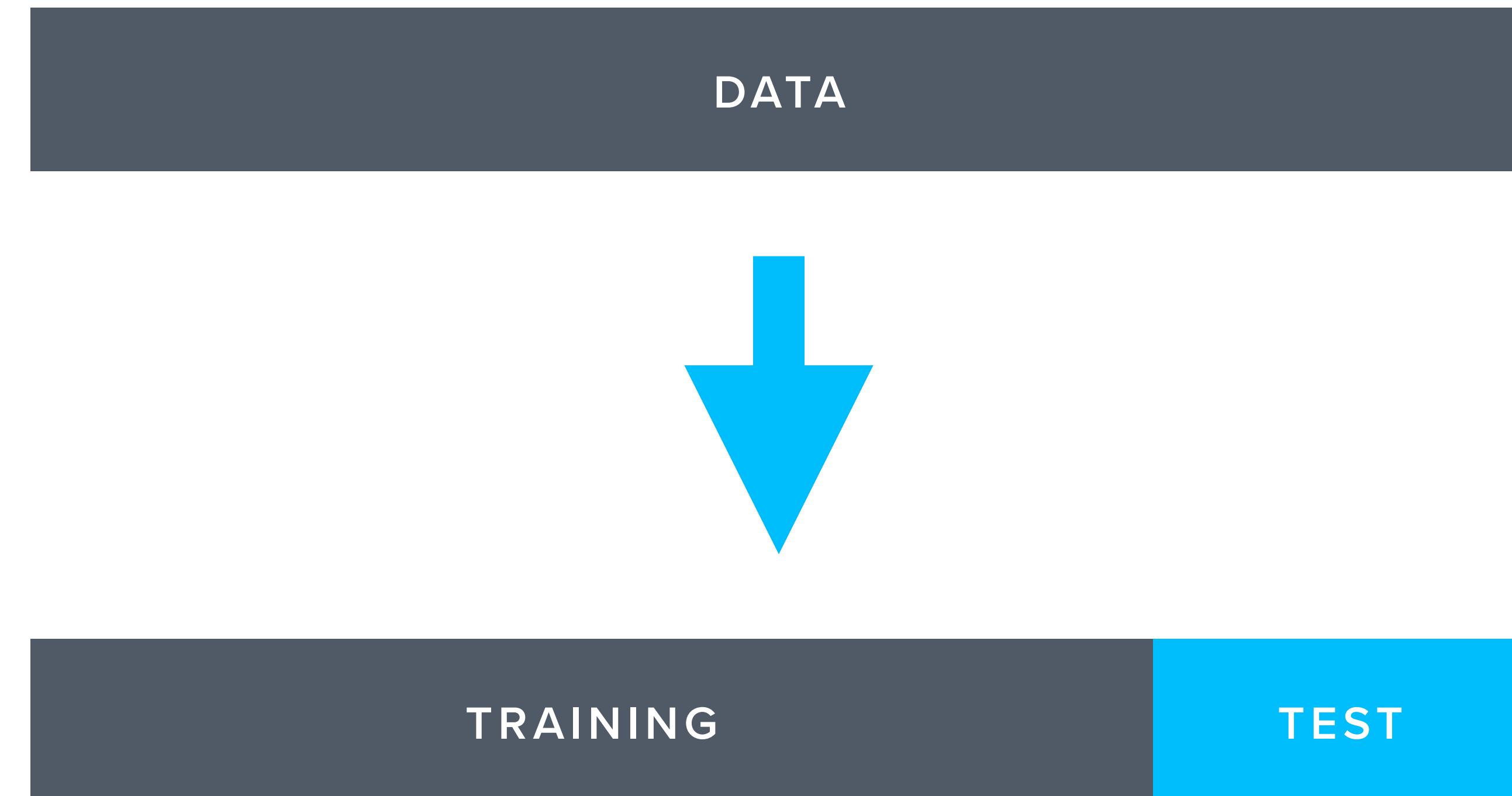


An optimal stopping point has to be found before overfitting



How do we check our model?

Splitting data into a training and test set



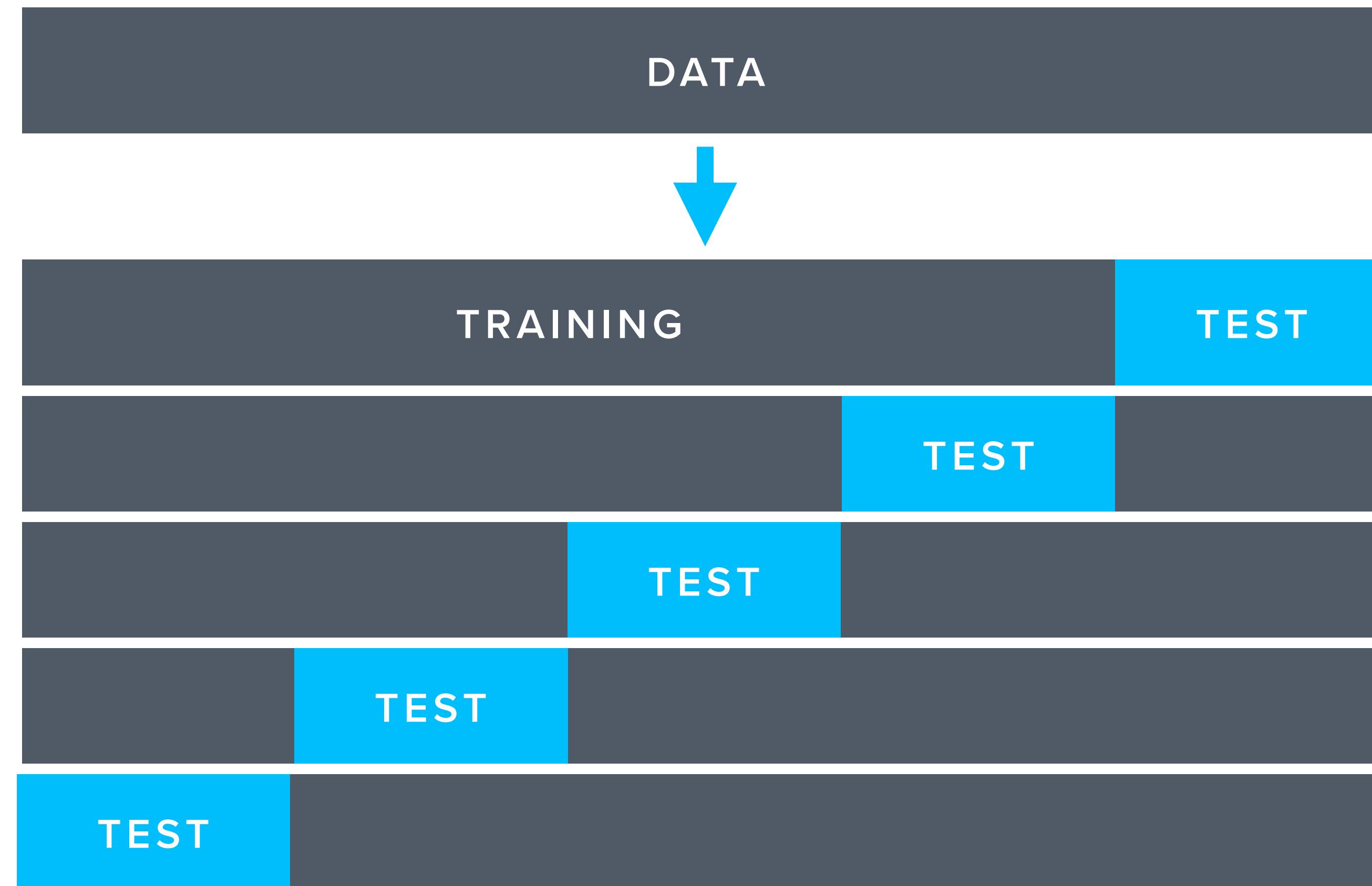
Validation

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



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



Please leave your feedback!

Click here for feedback:

<https://goo.gl/forms/2ueNRaOfcrTgOKsv1>

Don't hesitate to get in touch!

training@asidatascience.com

andrew@asidatascience.com



