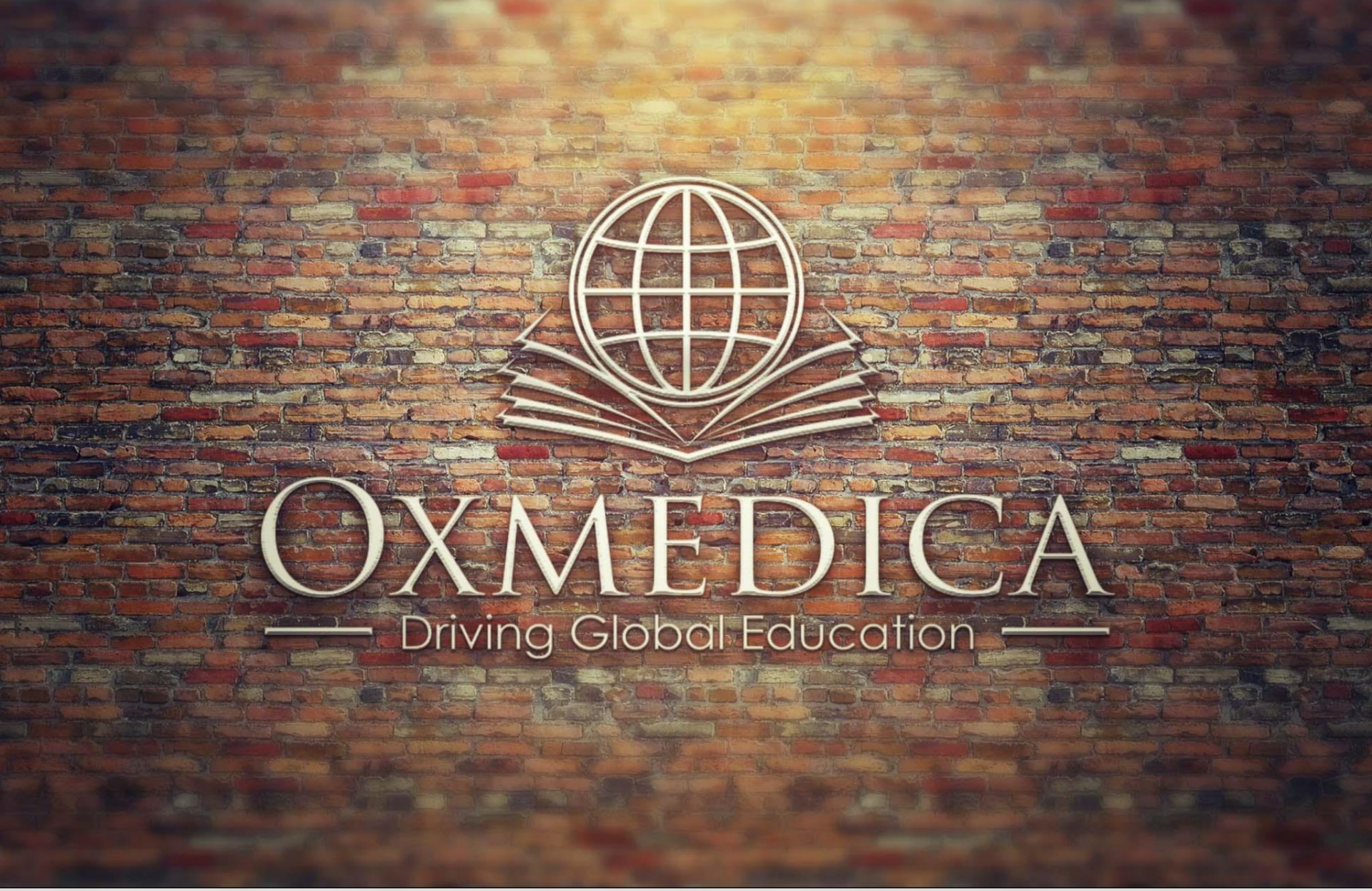


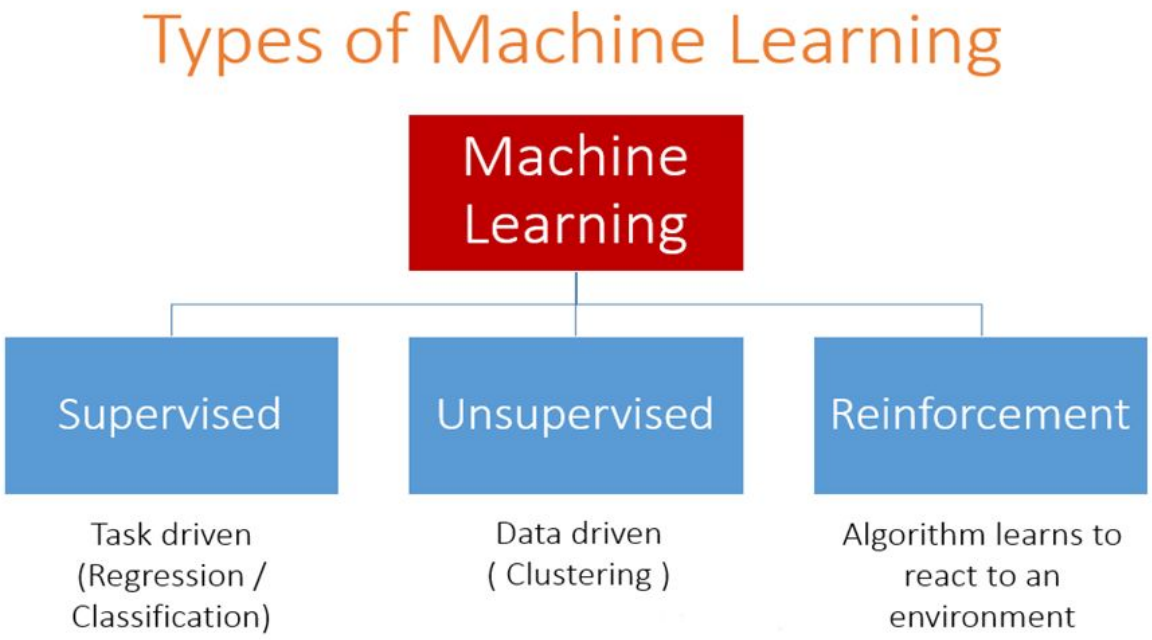
Unsupervised Learning

Overview of today

1. Understanding unsupervised learning: theory
2. The *k*-means clustering algorithm
3. Practical applications of clustering

Part A: Theory

Recap of different types of machine learning



Source

Supervised machine learning

- Dataset includes a target variable/feature.
  Q. What is a variable?

X1 (feature)	X2 (feature)	X3 (feature)	X4 (feature)	Y (target)

- Aim is to use the feature to predict the target

- Have lot of '**training**' examples and have to learn the pattern. Then given '**test**' examples and you have to predict the target

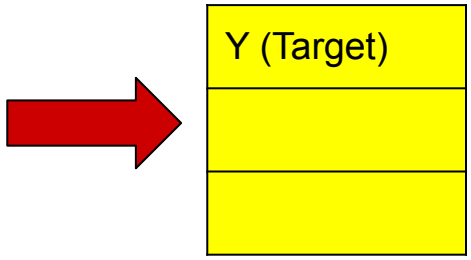
Supervised Learning

Training Examples

X1 (feature)	X2 (feature)	X3 (feature)	X4 (feature)	Y (target)

Test Examples: Use the features to predict the target

X1 (feature)	X2 (feature)	X3 (feature)	X4 (feature)



Unsupervised machine learning

- Unlabeled data - no target variable

X1 (feature)	X2 (feature)	X3 (feature)	X4 (feature)

- Aim is to find patterns in the data
- Main areas are **clustering** and **dimensionality reduction**
- Unsupervised learning is said to be ‘harder’ because we have to learn without labels. Also cannot validate our results.

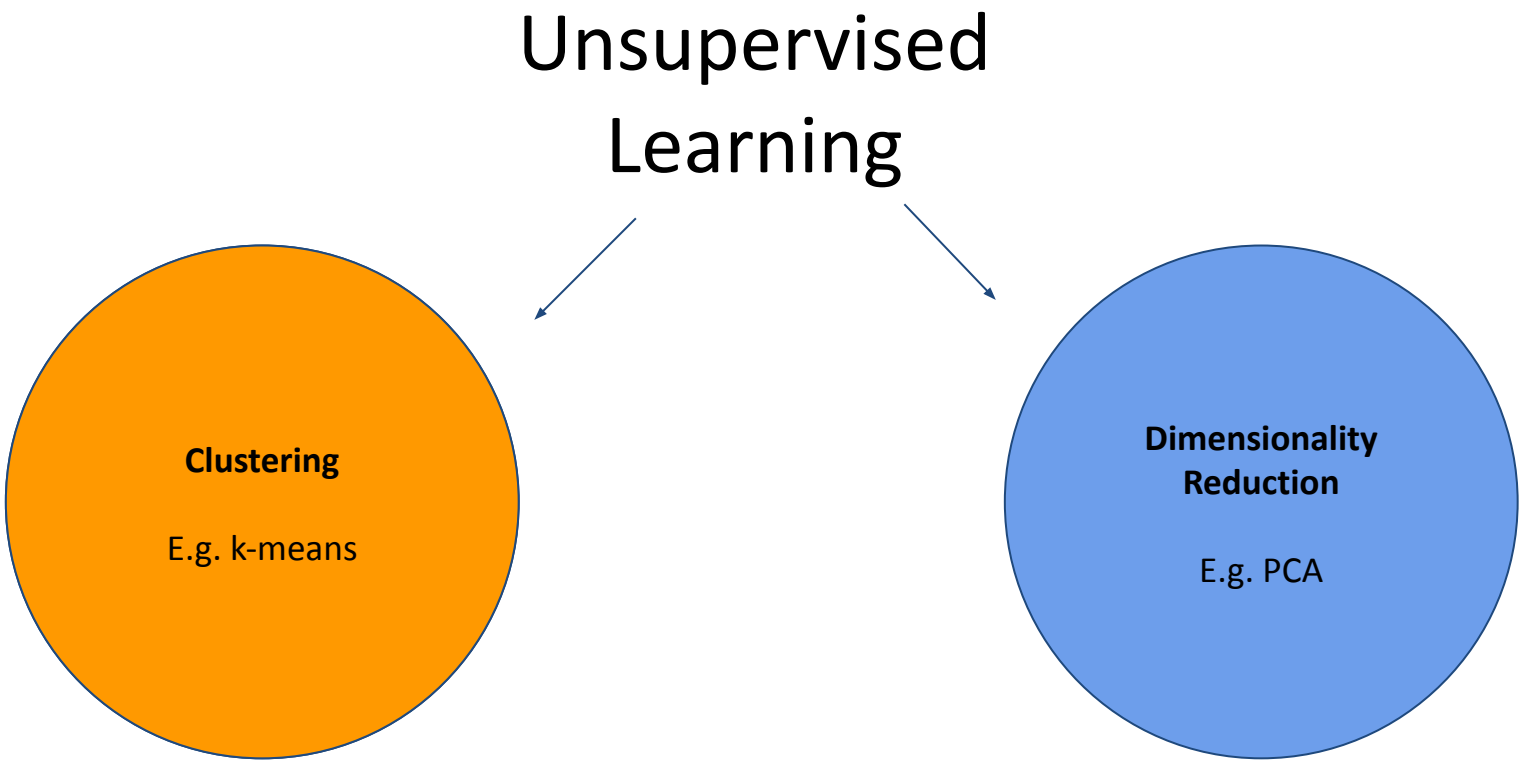
Formal definition of unsupervised learning

- “Supervision” is the act of helping the machine learn by providing some information to it.
- Normally this is providing the labels of the target so that it can learn to differentiate things easily
- Unsupervised learning is learning in the absence of human help.
- Learning patterns from unlabeled data

Mathematical definition of datasets

See the board.

Main types of unsupervised learning



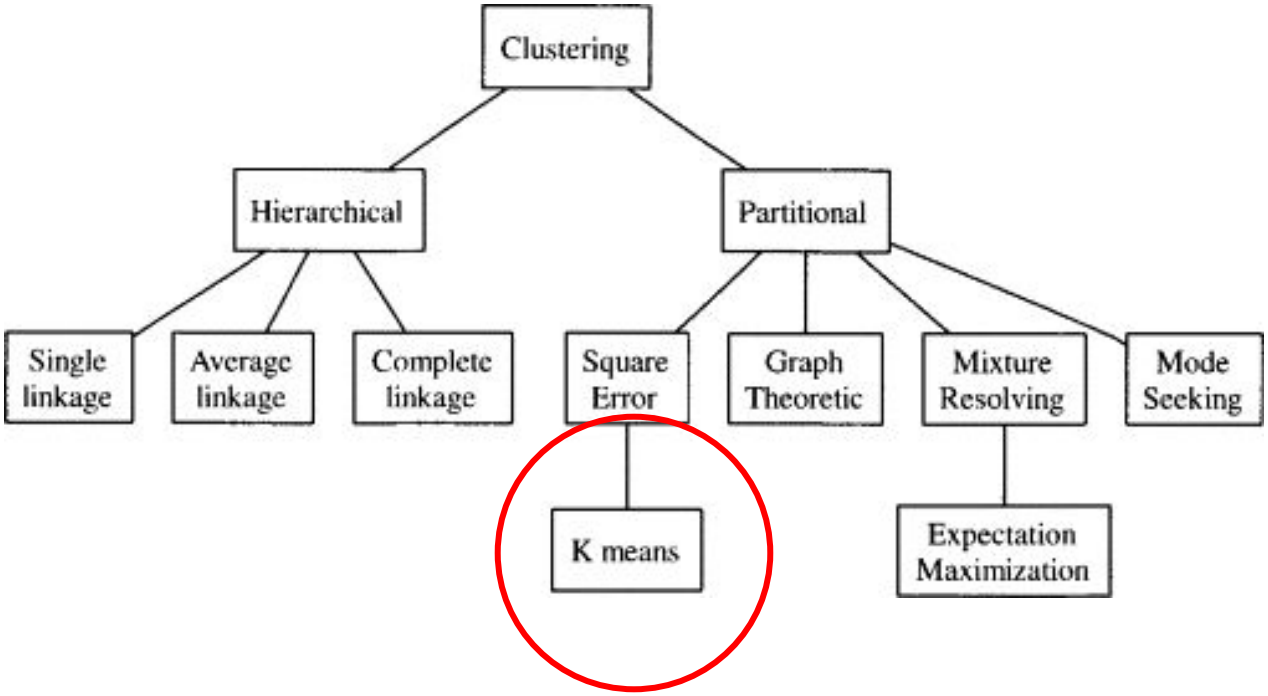
Clustering

What is clustering and why would we want to do it?

- Separating an unlabeled dataset of examples into ‘k’ clusters
- Items within each cluster should be similar to each other. E.g. you might separate cancer patients by their type of cancer or separate text documents into similar groups
- Identifies structure in the data
- Useful for many downstream tasks. Often a useful first step in machine learning pipelines.
- Can you think of any other use cases off the top of your heads?

Types of clustering algorithm

There are **many** different clustering algorithms out there!



Source

Types of clustering algorithm

There are **many** different clustering algorithms out there!

Review of clustering methods with applications

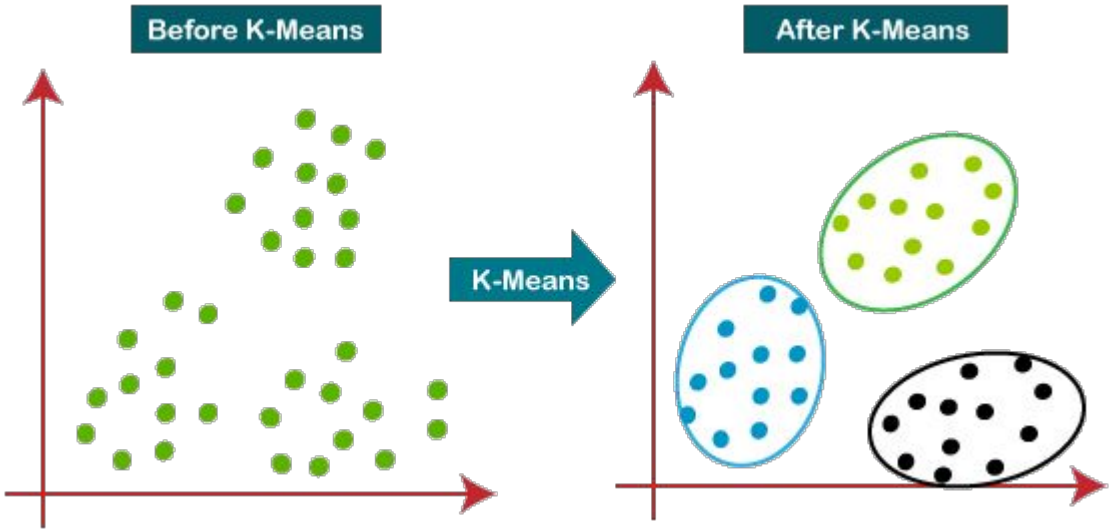
Authors

Oxford Internet Institute, University of Oxford

Contents

K-means: Overview

Aim: Partition your data points into a set S of k clusters to minimise the within-cluster sum of squares (WCSS) (the withinin-cluster variance)



Source

K-means: Mathematical derivation

⚠️This looks deceptively hard!⚠️

- The mathematical optimisation k-means is trying to solve is actually quite complex when written down but the actual application is quite easy.
- See the notes on my website for more details

K-means: In practice

Actually not too hard!

1. Label all the examples with a random label from $j = 1, \ldots, k$
2. Iterate the following until no further changes in class label
   a. Calculate the centroid of each cluster, μ_j
   b. Reassign all points to the closest cluster based on Euclidean distance. Note that you could use other distance metrics too.

Simulation example

Questions about the algorithm

Tasks in pairs: Strengths and limitations

- 5 minutes
- Discuss and write down the strengths and limitations of clustering in general and the specific k-means algorithm
- **Hint:** Think about it from a technical and application point of view

Discussion

Strengths and limitations

✅ Simple, efficient and usually gets decent results.

❌ You have to define K before starting the clustering!
It is not obvious how to do this… often require domain specialists

❌ Very depending on the initialisation!

Algorithms have been developed to improve this including the well-known *k-means*++ algorithm (Stanford algorithm and now the default in sklearn).

Questions

Part B: Practical applications

Individual task: Research

- Find separate applications of clustering in your assigned field.
- Report back to the group (90 seconds) explaining the applications you find.

Marketing	**Social Media**	**Transportation**	**Urban Planning**	**Education**
Agriculture	**Energy**	**Environmental Science**	**Retail**	**Manufacturing**
Entertainment	**Cybersecurity**	**Sports Analytics**	**Real Estate**	**Restaurants**

Discussion

Recap questions

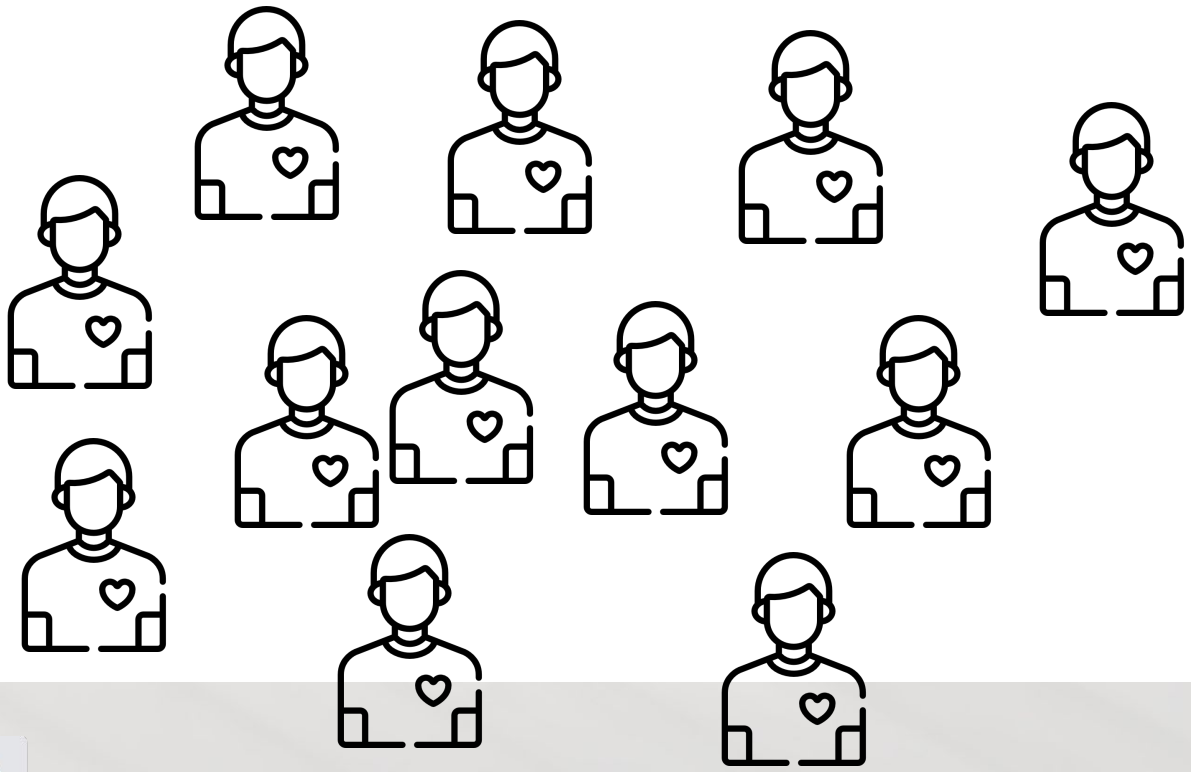
1. What are the types of machine learning?
2. What defines unsupervised learning?
3. What are the different types of clustering algorithm?
4. Why might we want to do clustering?
5. How does k-means work?
6. What are the pros and cons of the k-means algorithm?

Example applications of clustering in healthcare and finance

Healthcare example

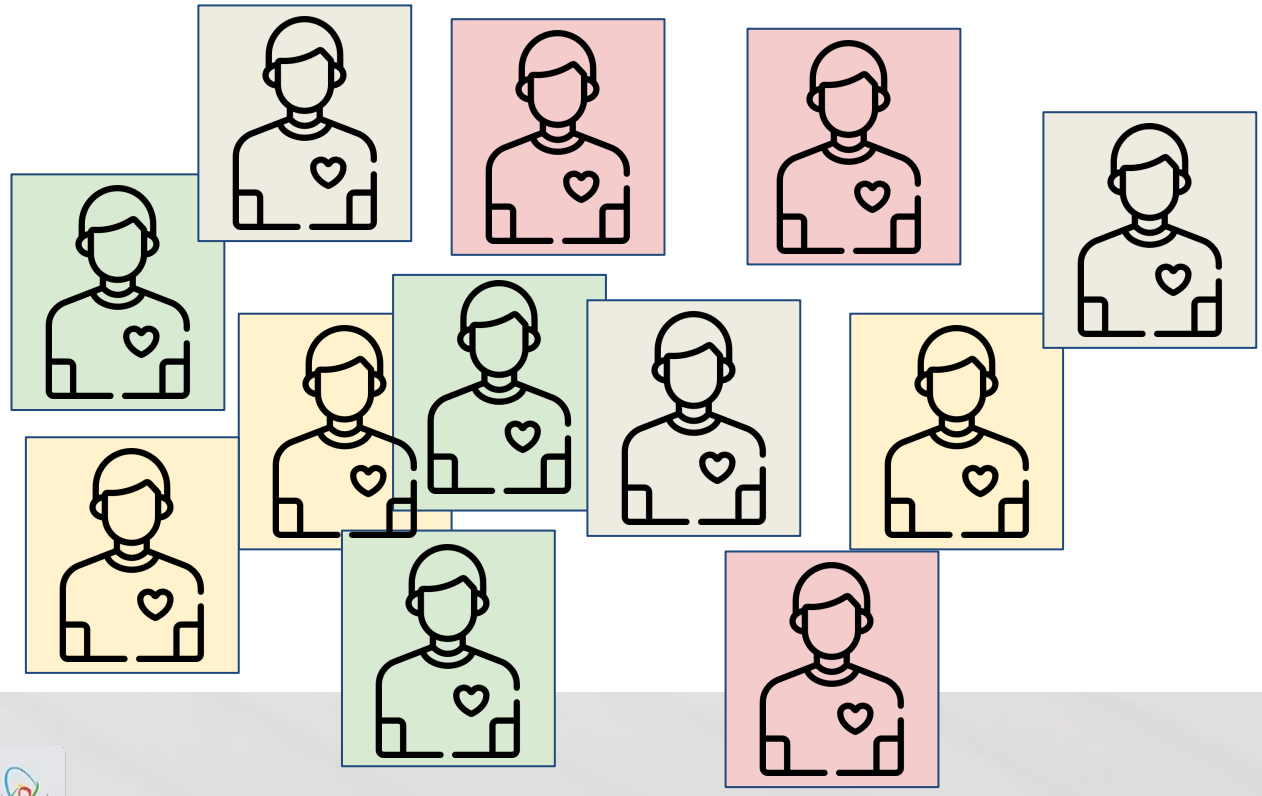
Patient clustering

- Patients may have many different symptoms
- Can use these symptoms to classify them into different risk groups
- This can also be used to automatically detect outliers



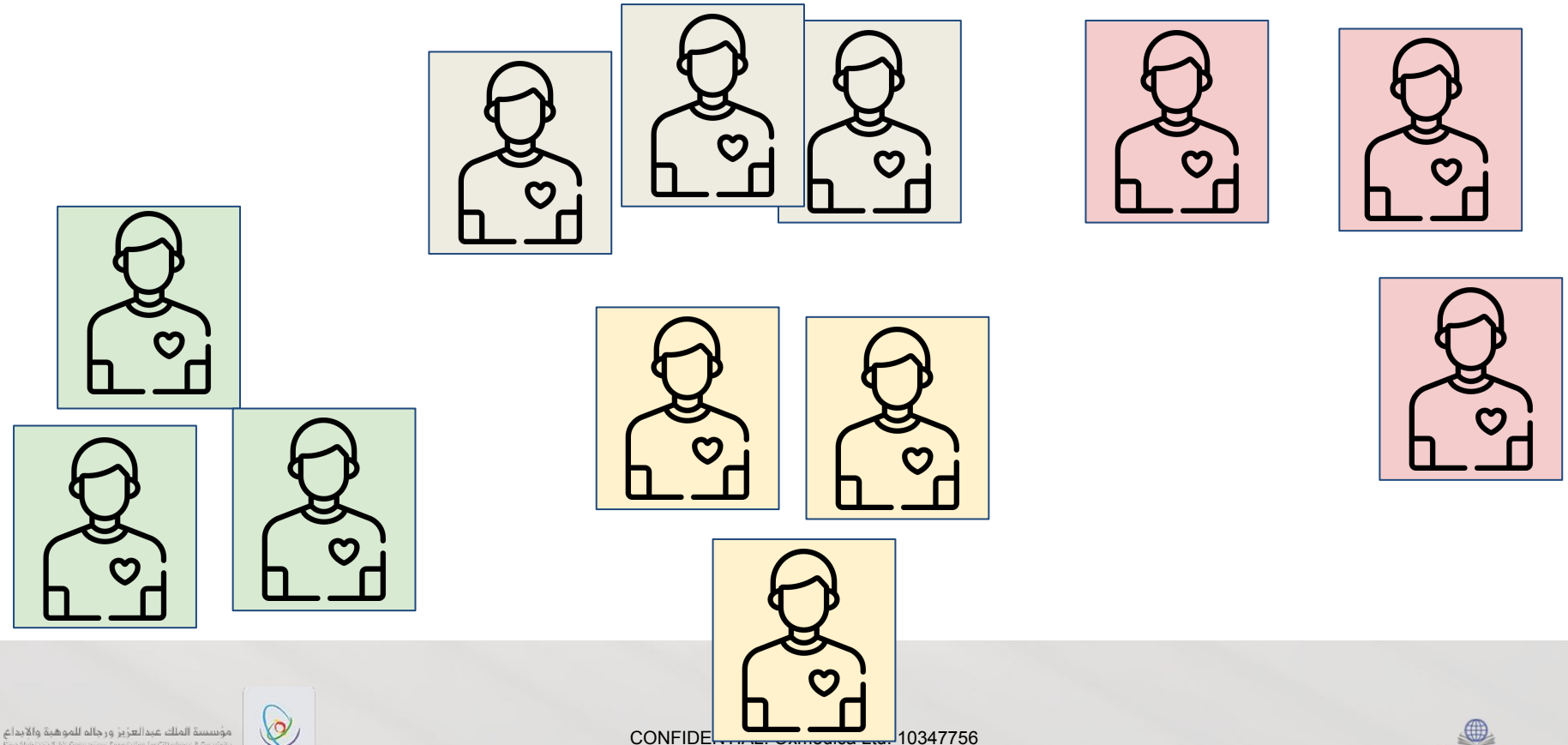
Patient clustering

- Patients may have many different symptoms
- Can use these symptoms to classify them into different risk groups
- This can also be used to automatically detect outliers

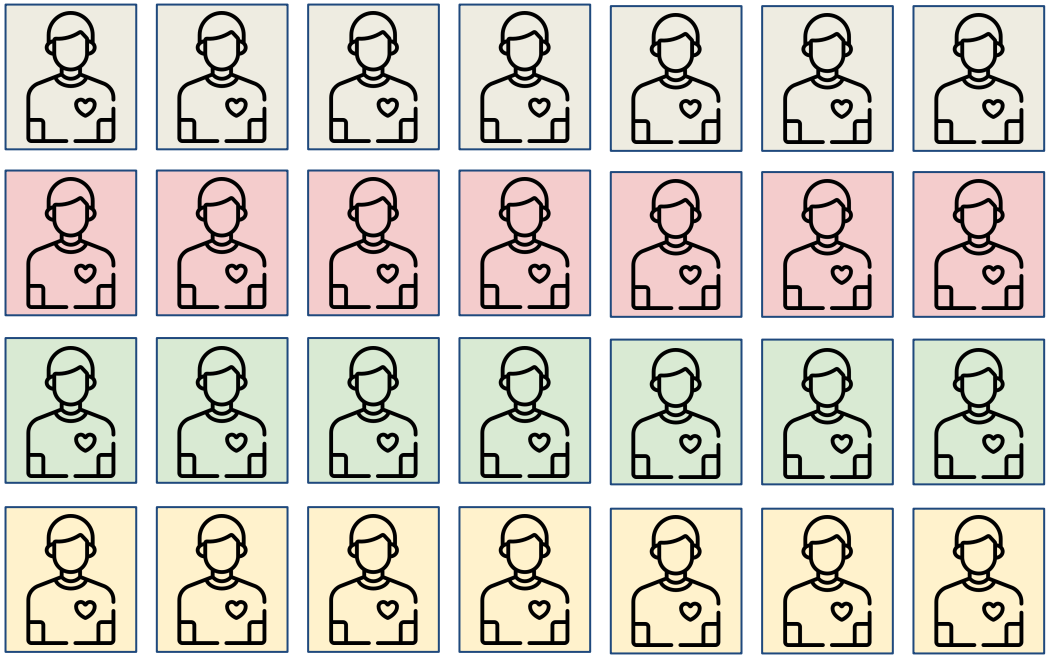


Patient clustering

- Use the clusters to help you work out what disease people have
- Useful approach for sub-diseases e.g. helping identify cancer groups



Patient clustering: Can detect outliers



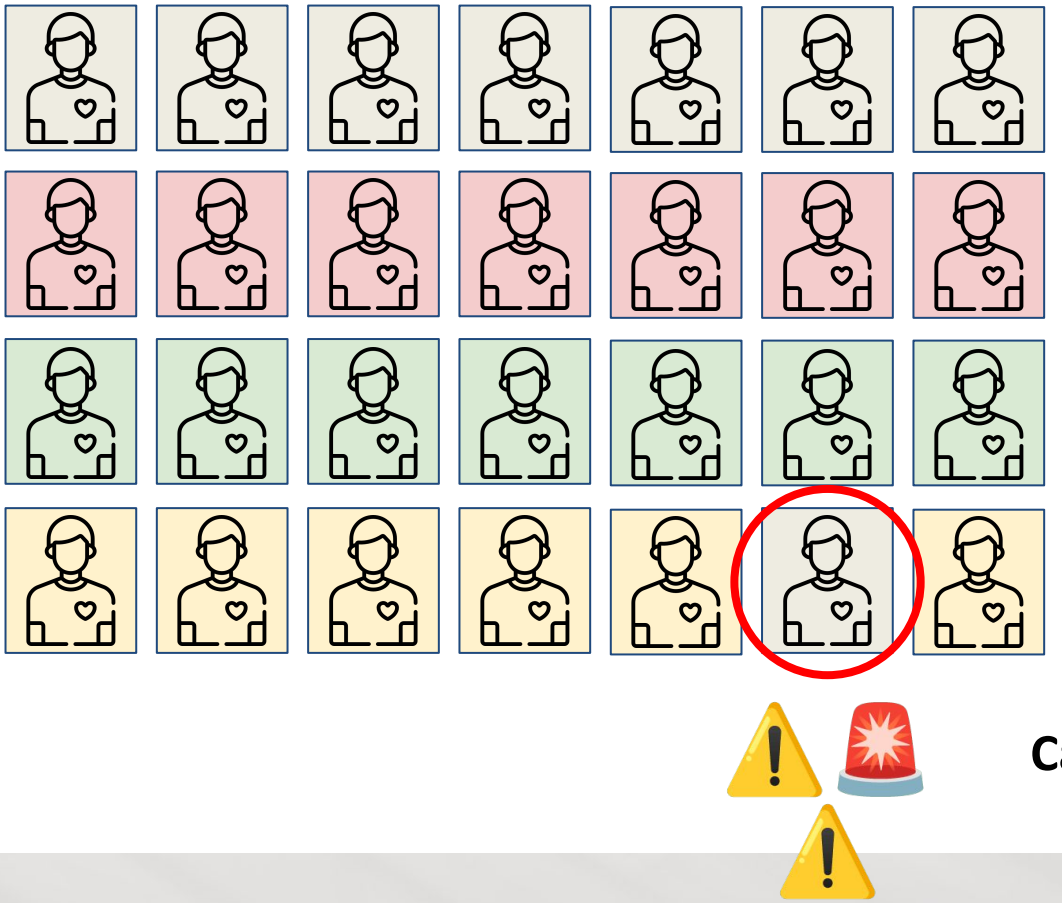
Cancer

Kidney

Mental health

Heart attack

Patient clustering: Can detect outliers



Cancer

Kidney

Mental health

Heart attack

Can help you detect outliers!

Any others?

Finance example

Building a diversified portfolio

- Ideal to build a portfolio of stocks which are have different characteristics to prevent correlated losses.
- Cluster stocks by various metrics e.g. variance, price, risk, industry…etc
- Each group of stocks should be highly correlated within the cluster but have low between-cluster correlation
- Select stocks from different clusters to get a diversified portfolio → lower risk

Building a diversified portfolio

- What kind of risks might this protect the portfolio from?

- What are the limitations of this approach?

Any others?

Clustering in Intensive Care Units

UNSUPERVISED LEARNING APPROACHES FOR IDENTIFYING ICU PATIENT SUBGROUPS: DO RESULTS GENERALISE?

Harry Mayne[1], Guy Parsons[1, 2], and Adam Mahdi[1]

[1]Oxford Internet Institute, University of Oxford
[2]NIHR Academic Clinical Fellow at University of Oxford and Thames Valley Deanery

ABSTRACT

The use of unsupervised learning to identify patient subgroups has emerged as a potentially promising direction to improve the efficiency of Intensive Care Units (ICUs). By identifying subgroups of patients with similar levels of medical resource need, ICUs could be restructured into a collection of smaller subunits, each catering to a specific group. However, it is unclear whether common patient subgroups exist across different ICUs, which would determine whether ICU restructuring could be operationalised in a standardised manner. In this paper, we tested the hypothesis that common ICU patient subgroups exist by examining whether the results from one existing study generalise to a different dataset. We extracted 16 features representing medical resource need and used consensus clustering to derive patient subgroups, replicating the previous study. We found limited similarities between our results and those of the previous study, providing evidence against the hypothesis. Our findings imply that there is significant variation between ICUs; thus, a standardised restructuring approach is unlikely to be appropriate. Instead, potential efficiency gains might be greater when the number and nature of the subunits are tailored to each ICU individually.

Recent work from my laboratory at Oxford

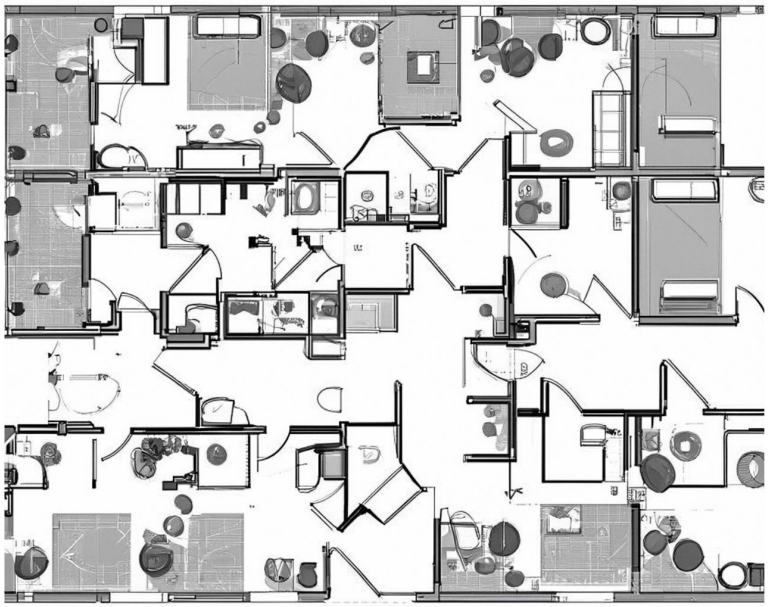
ICU Risks

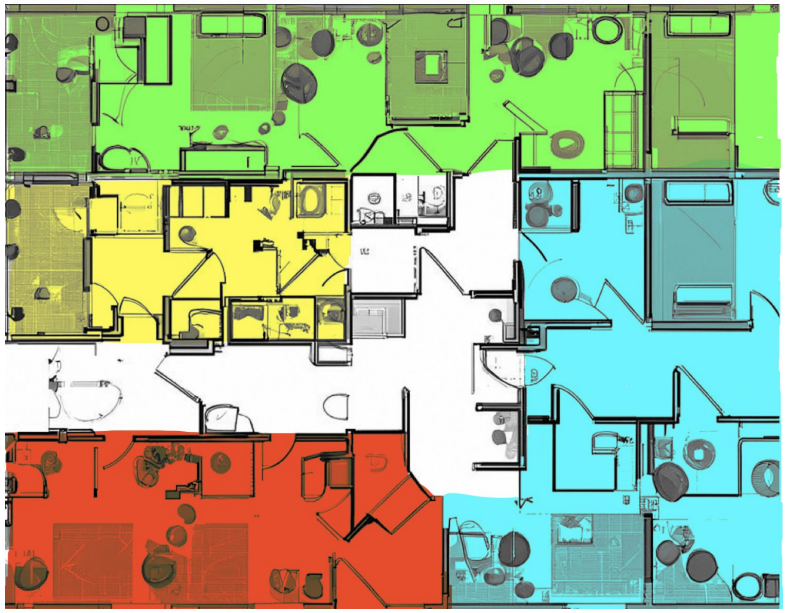
1. Ageing population
2. Advances in medicine
3. Under-investment

→ Inefficiency

→ Poorer quality of care

→ Excessive pressure on clinicians





Group Identification: Results

Cluster 1

48.18%

Relatively healthy

Near perfect survival

Cluster 2

33.68%

Weaker patients

Survive with long-term health problems

Cluster 3

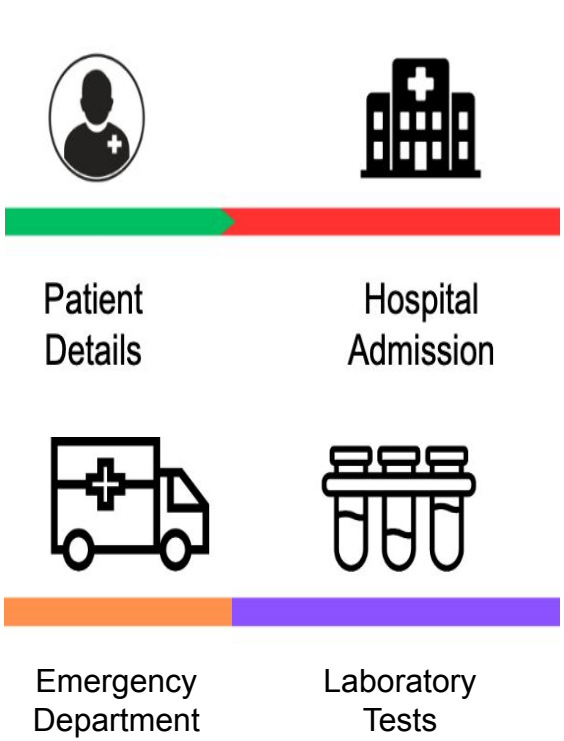
18.14%

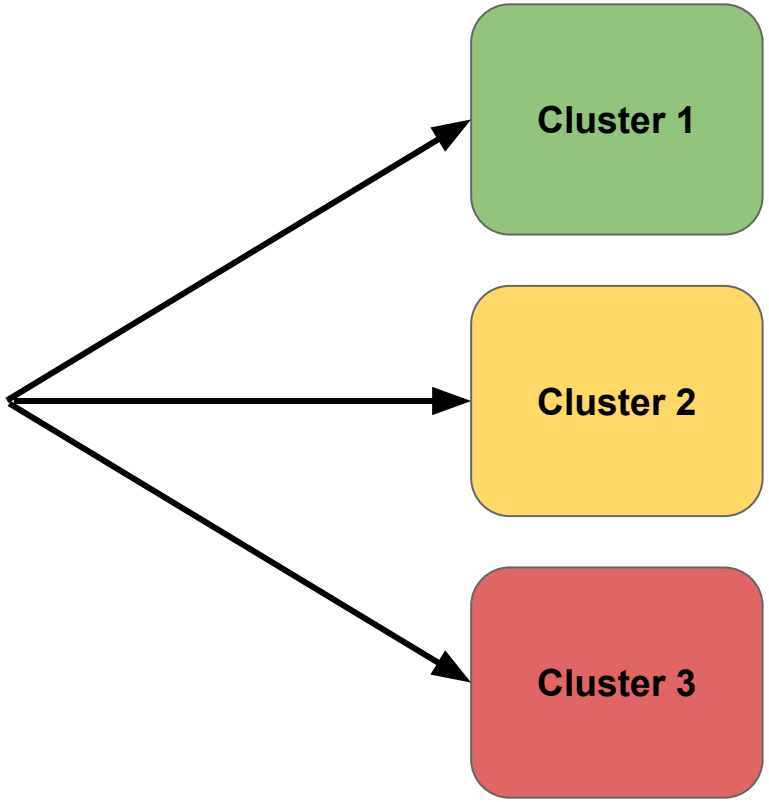
Severe patients

76.19% morality

2

Assigning Patients at ICU Admission





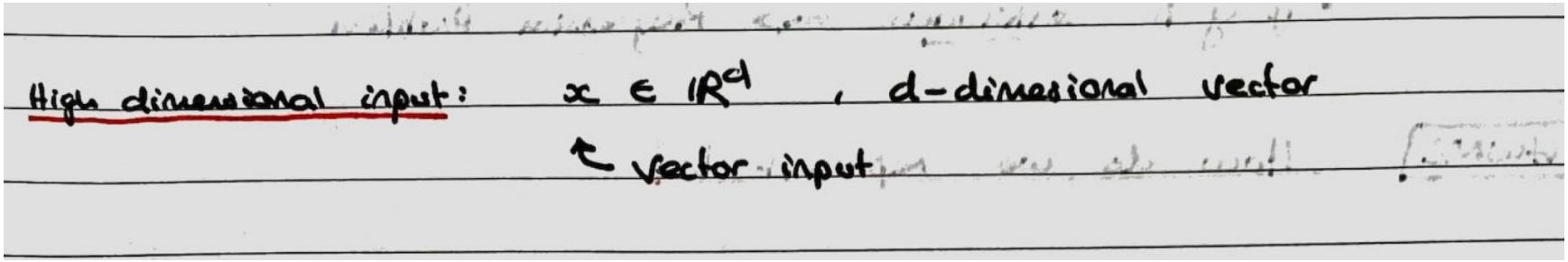
Recap questions

1. What are the types of machine learning?
2. What defines unsupervised learning?
3. How can unsupervised learning be used in healthcare?
4. How can unsupervised learning be used in finance?
5. What are the risks of these approaches?

[EXTRA]

Dimensionality Reduction

What is dimensionality reduction and why would we want to do it?



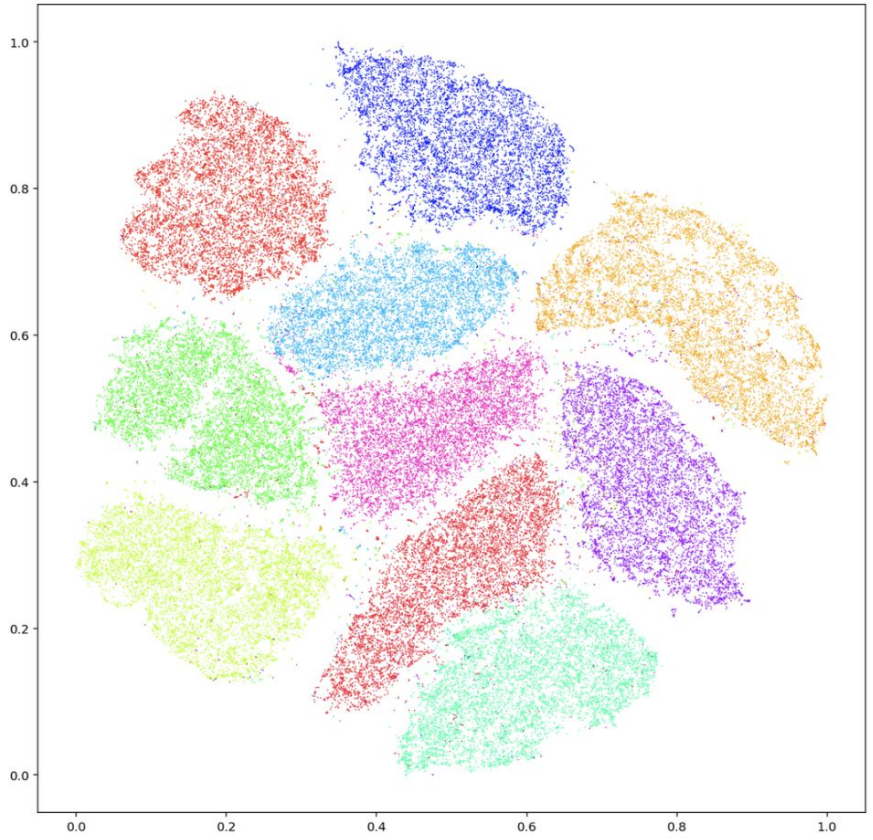
- Datasets can often be ‘high-dimension’ → What do you think this might mean?
- Why might this be problematic?
- The general idea of dimensionality reduction is to keep as much of the information as possible but in fewer dimensions.

Two main use cases

1. Storage

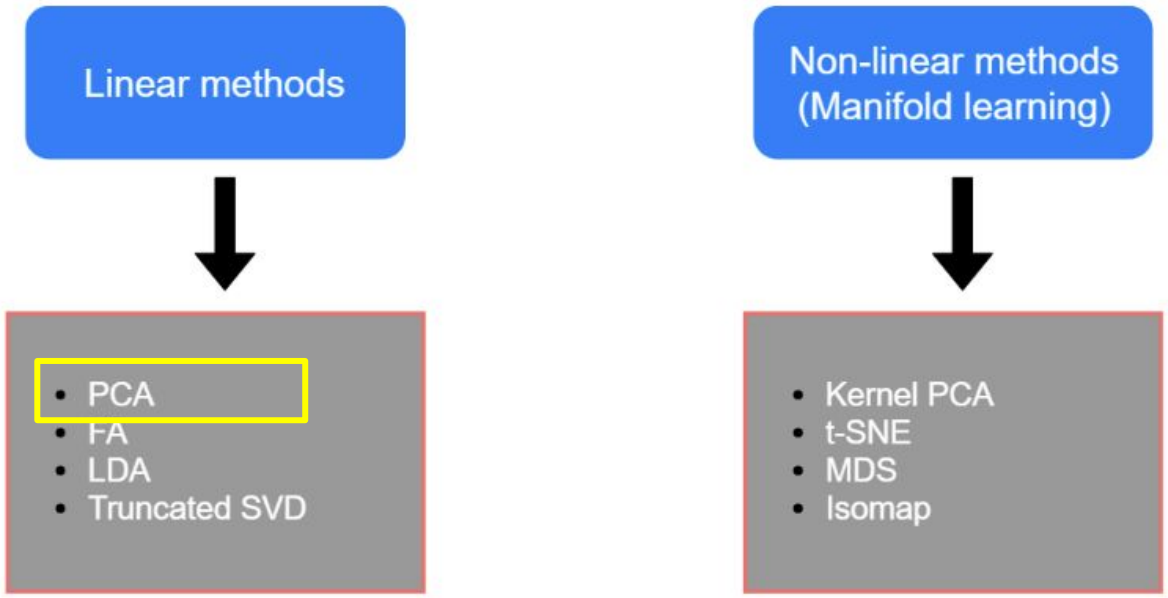
2. Visualisation
(more common for student applications)

Visualisation of 784 dimensional data in 2D (MNIST)



Types of dimensionality reduction

Also lots of types of dimensionality reduction techniques.

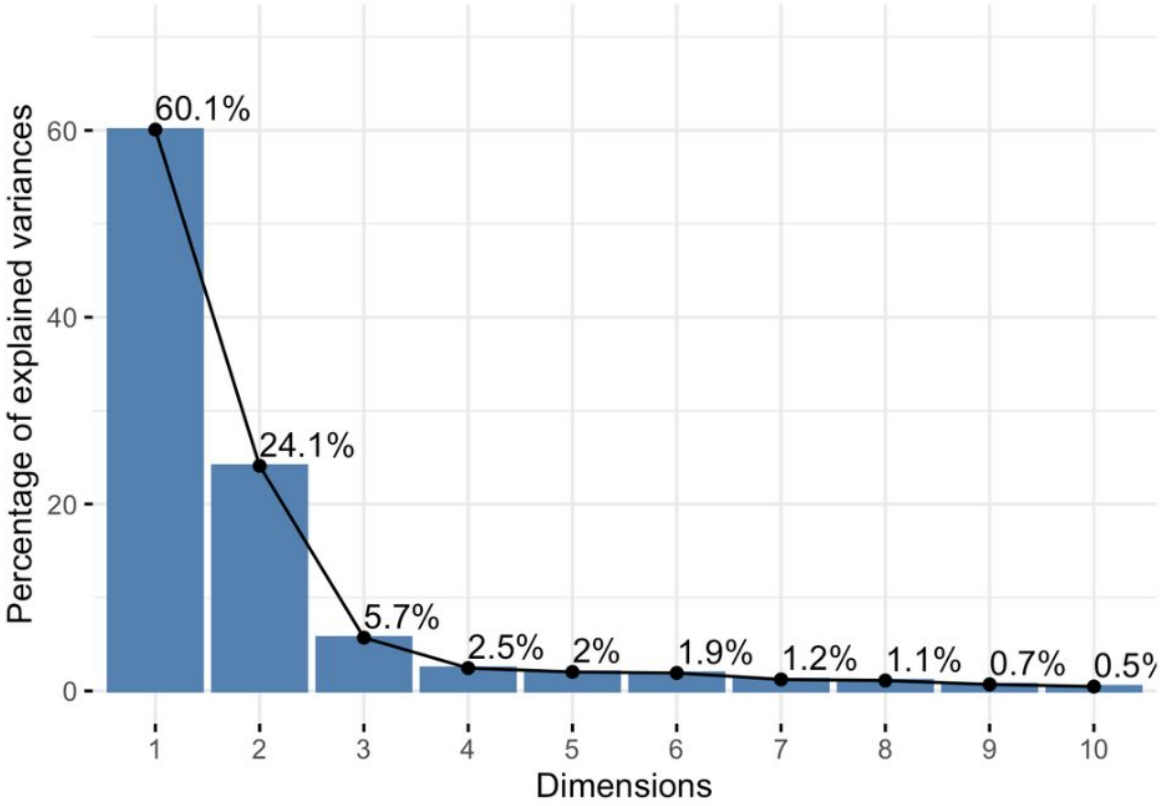


Source

Principal Components Analysis (PCA)

Car example on the board

Proportion of variance explained plot (scree plot)



The maths behind PCA on the board
(only if time…)

EXTRA: Case study

Individual task