1495 UNIVERSITY OF ABERDEEN | CELEBRATING 525 YEARS 1495 – 2020

ABERDEEN 2040

Data Mining

Data Mining & Visualisation

Lecture 2

2025

Today…

- What is data mining?
- Data types
- Levels of measurement

What is Data Mining?

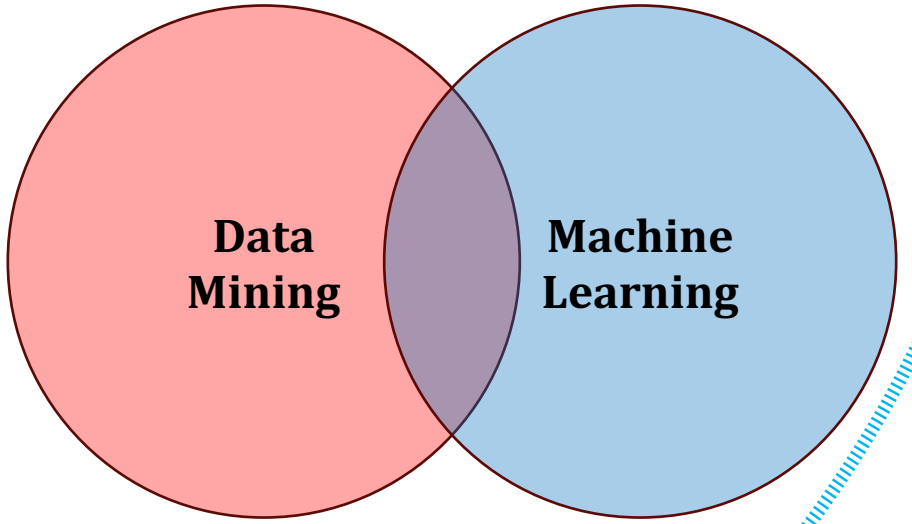
What is Data Mining?

- Process of discovering patterns, understanding trends, and extracting useful information from large datasets

- Examples: Detecting fraud, generating personalised recommendations, grouping similar products or people

Data Mining ≠ Machine Learning

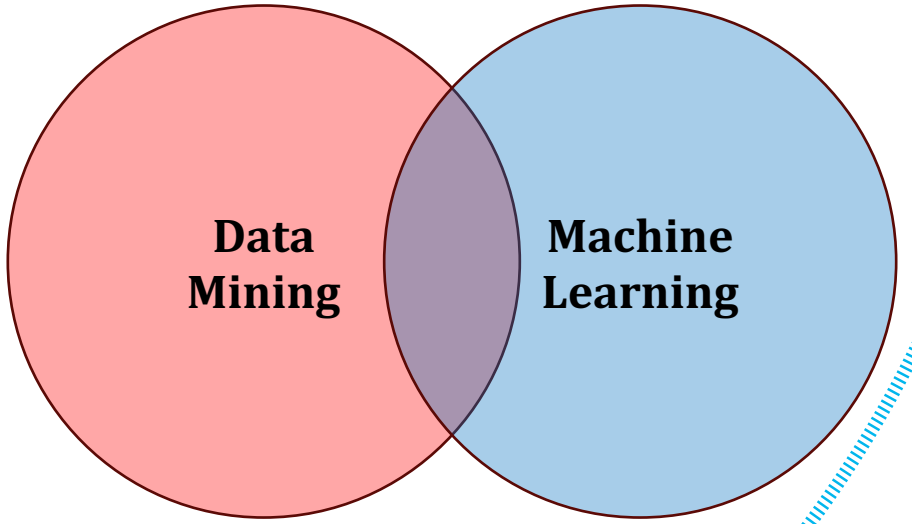
- Considerable overlaps between Data Mining and Machine Learning
- Many of the same approaches and techniques used
- But underlying aims and end-goals are different!



Data Mining ≠ Machine Learning

- ML might be simplified as solving a particular problem
  - E.g. given a dataset, can we predict outcome Y?

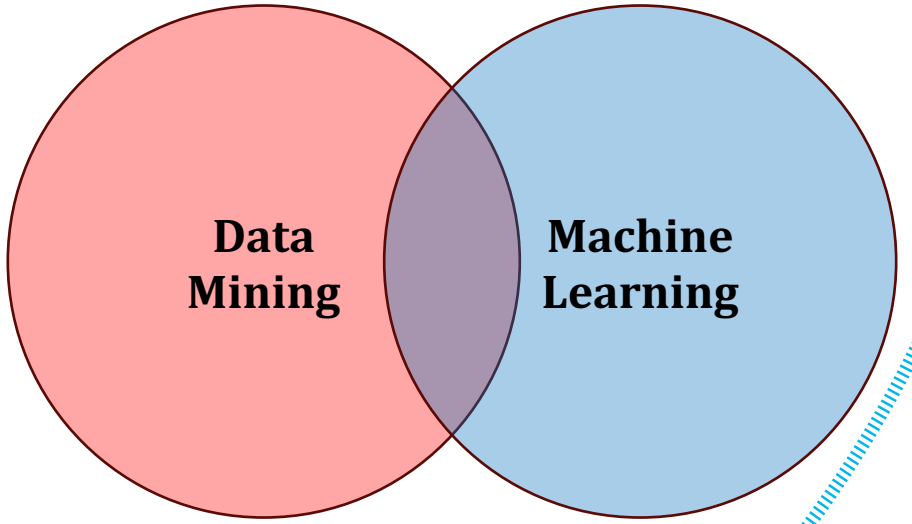
  - Identify task
  - Collect relevant data
  - Feature engineering & data pre-processing
  - Train model(s)
  - Evaluate model(s)
  - Deploy model
  - Ongoing refinement & optimisation



Data Mining ≠ Machine Learning

- DM concerns the broader processes of understanding the data itself
  - E.g. given a dataset, why does outcome Y occur?

  - Identify relevant questions
  - Collect relevant data
  - Exploratory data analysis (EDA) & Data Visualisation
  - Data cleaning & pre-processing
  - Train interpretable model(s) or conduct statistical tests
  - Analyse & report outcomes to stakeholders
  - Ongoing monitoring and reporting



Data Mining ≠ Machine Learning

While overlaps exist, we will focus on:

- **Inferring** from past data
- **Interpreting** the models we build
- The **practicalities** of using these techniques

Again, **understanding your data** and **asking the right questions!**

Data Mining: Example Scenario

Let's say we work for a mobile developer as a Data Scientist

The CEO might want to know:

- What makes people upgrade to premium membership?
  - If we discount premium membership, do we make more money?
- What makes people keep using the app (retention)?
  - Where are users most likely to stop using the app?

Data Types

Data Types

- There are lots of different **types** of data

Try to think of a few different types!

Data Types

- There are lots of different **types** of data
- E.g.:
  - Numbers and measurements
  - Dates & times
  - Spatial information (e.g. latitudes & longitudes)
  - Ordered groups
  - Unordered groups
  - Multimedia (images, video, audio)

Categorical vs Quantitative Data

- Categorical data refers to data that can be divided into groups.

  E.g.: Gender, race, age group, and educational level

- Quantitative data is data that can be represented numerically, including anything that can be counted, measured, or given a numerical value.

  E.g.: Age, height, number of students in a lecture

Discrete vs Continuous Data

- Discrete data refers to data that can only assume specific values that cannot be subdivided.

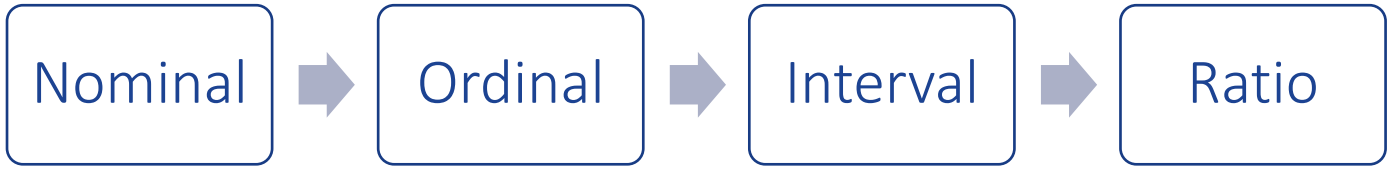
- Continuous data can be any numeric value and can be meaningfully split into smaller parts.

- Categorical data is discrete, but quantitative data can be discrete or continuous!

Levels of Measurement

Levels of Measurement

- Data is often grouped into one of four levels, indicating its precision



- Importantly, the level can impact what analysis you can perform with that data

Levels of Measurement: Nominal

- Unordered classes (categorical). Data can only be categorised.
- May be coded into numeric 'dummy variables'
- Examples: Gender, race, degree program

Levels of Measurement: Ordinal

- Ordered classes (categorical). Data can be categorised and ranked.
- May also be coded into numeric variables
- Examples: Age group, educational level

Levels of Measurement: Interval

- Numerical (quantitative) data, with equal intervals but with no absolute zero

- Examples: Temperature (Celsius), year

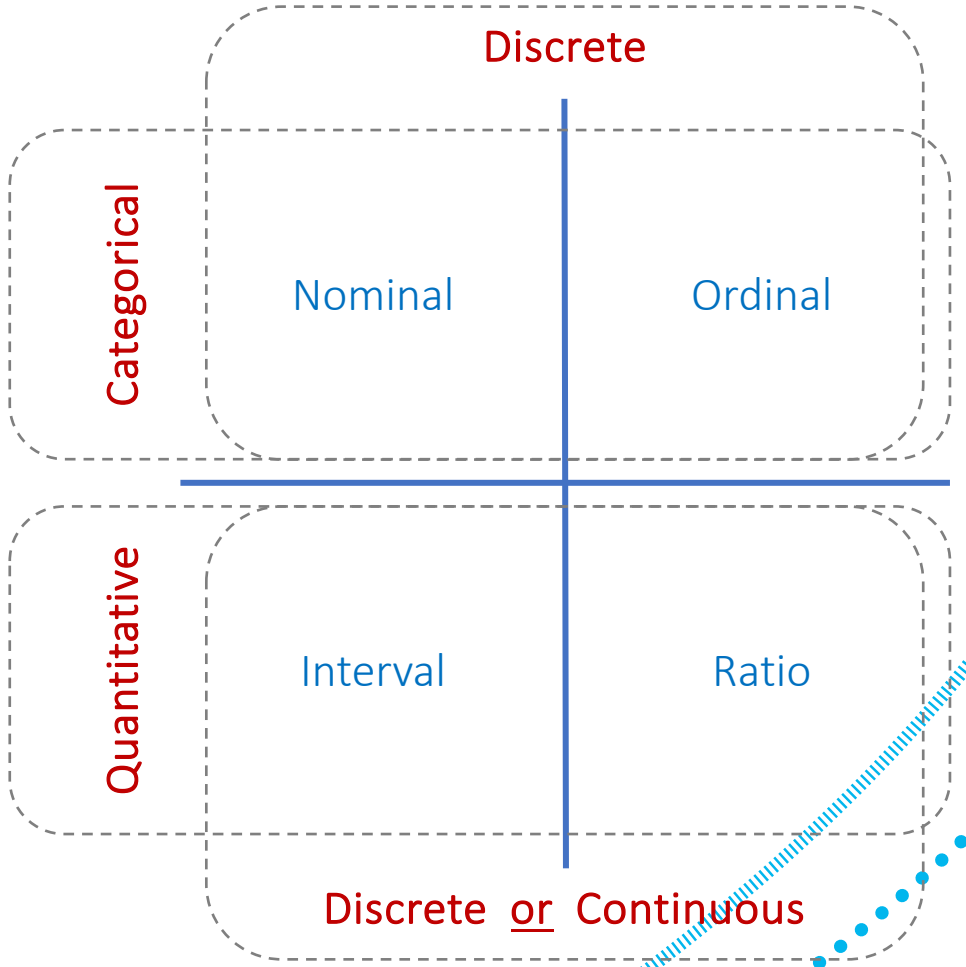
Levels of Measurement: Ratio

- Numerical (quantitative) data, with equal intervals and with an absolute zero

- Examples: Age, weight, height, tempeature (Kelvin)

Levels of Measurement

- Again, level will determine what analysis you can perform with that data
- However: Python won't tell you, and won't correct you!
- Understand your data!



Example: Titanic Dataset

```
# Load the 'titanic' example dataset
titanic = sns.load_dataset("titanic")

# Show the first few rows
titanic.head()
```

[2] ✓ 1.1s

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone
0	0	3	male	22.0	1	0	7.2500	S	Third	man	True	NaN	Southampton	no	False
1	1	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes	False
2	1	3	female	26.0	0	0	7.9250	S	Third	woman	False	NaN	Southampton	yes	True
3	1	1	female	35.0	1	0	53.1000	S	First	woman	False	C	Southampton	yes	False
4	0	3	male	35.0	0	0	8.0500	S	Third	man	True	NaN	Southampton	no	True

Have a think -- What levels are each of the variables at?

Today's Lab – Let's Explore Some Datasets!

- Some Getting Started guides to get up and running with popular Python packages for data analysis.
- Once completed, explore Kaggle to find some datasets that you might be interested in using throughout the course.
- Throughout the labs, you will explore, visualise and analyse datasets of your choosing – so find some interesting ones!