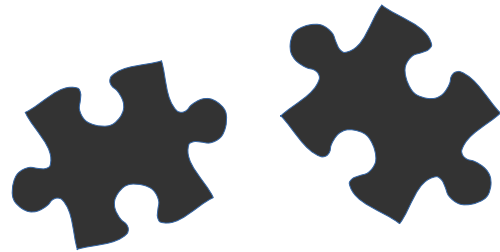


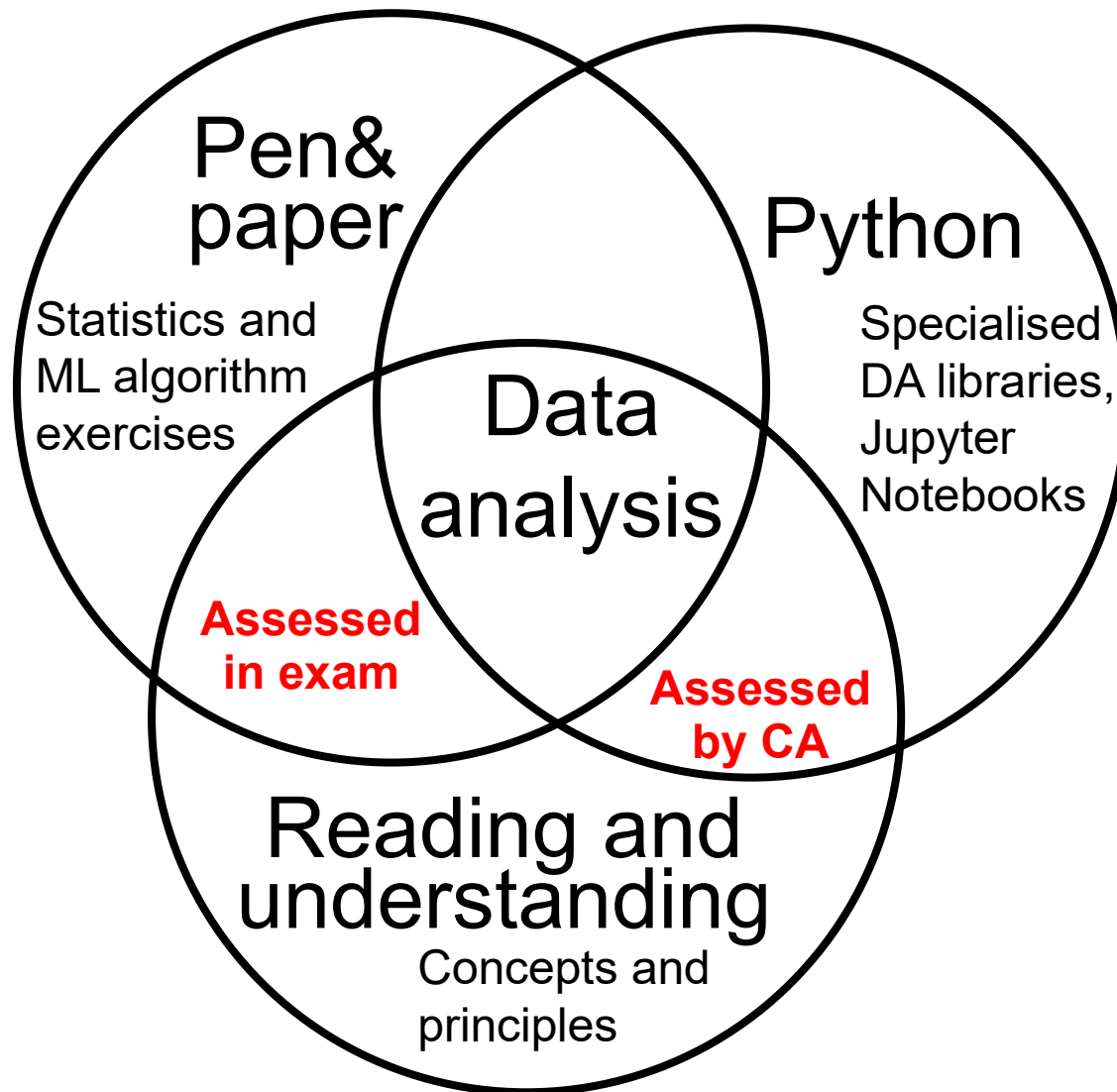
Data Analysis: Introduction

TU Dublin Tallaght Campus,
Department of Computing

In this module



you will find

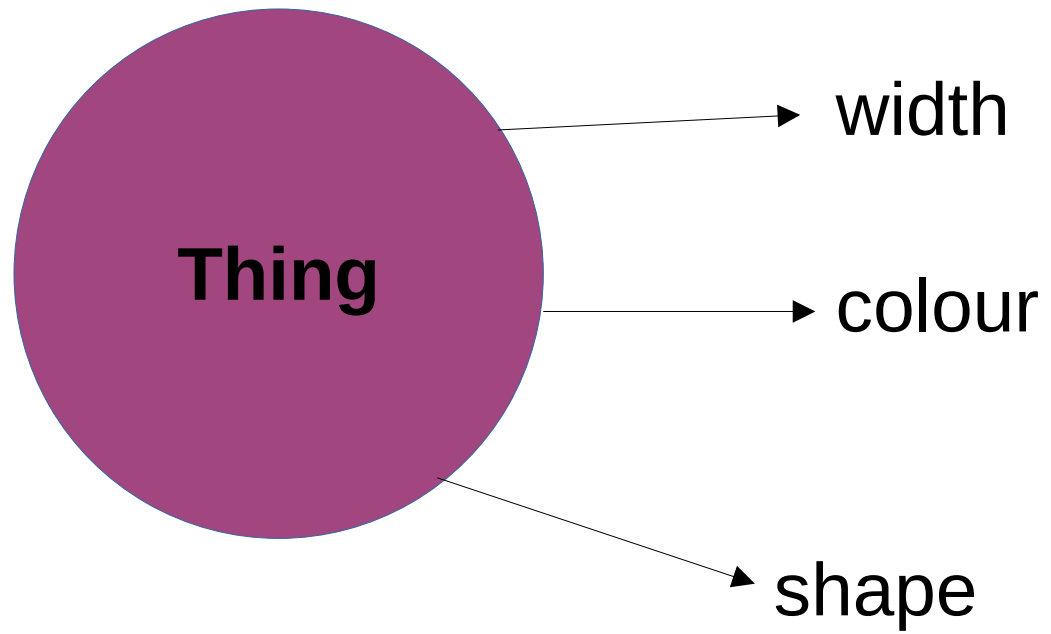


What are we
dealing with

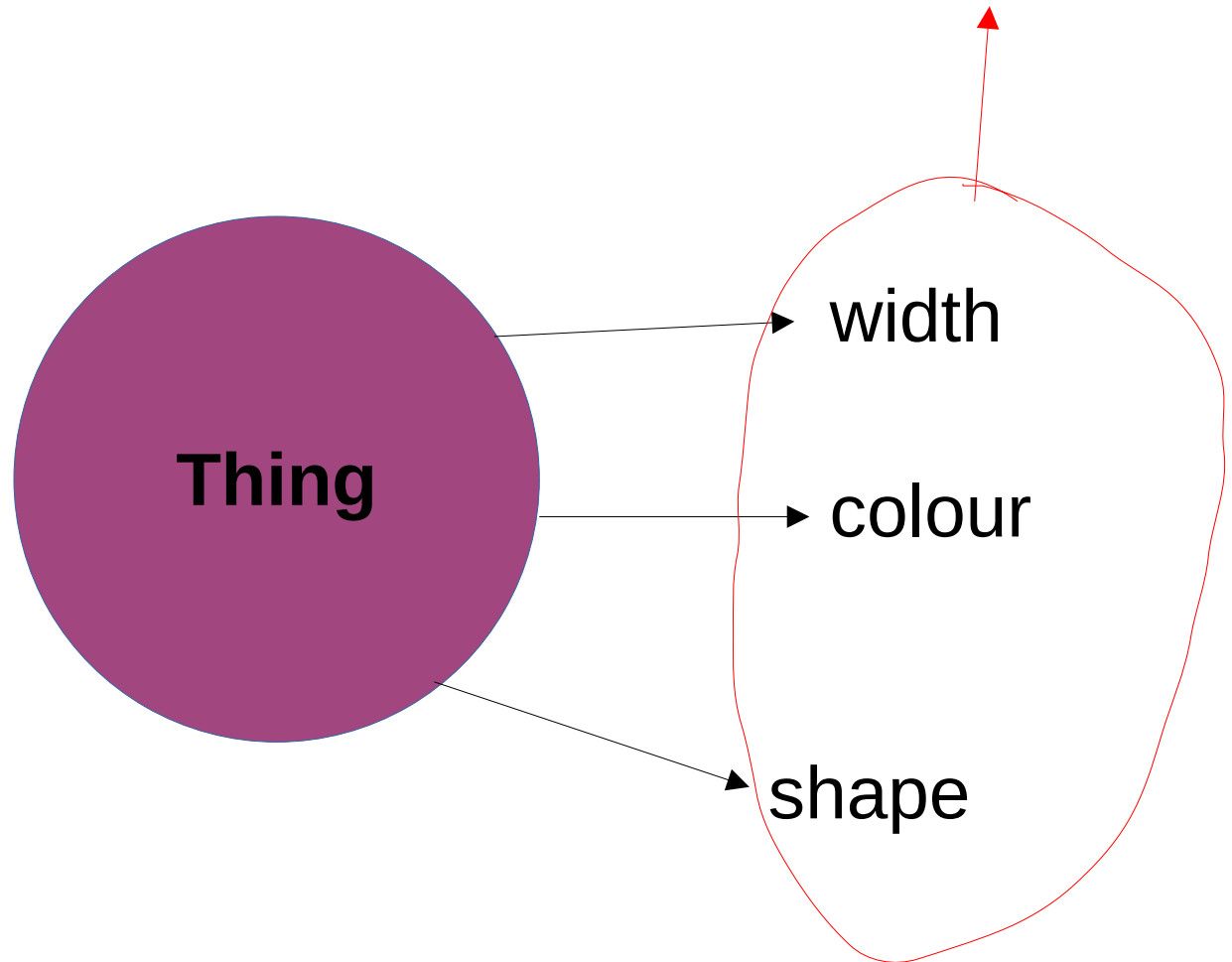




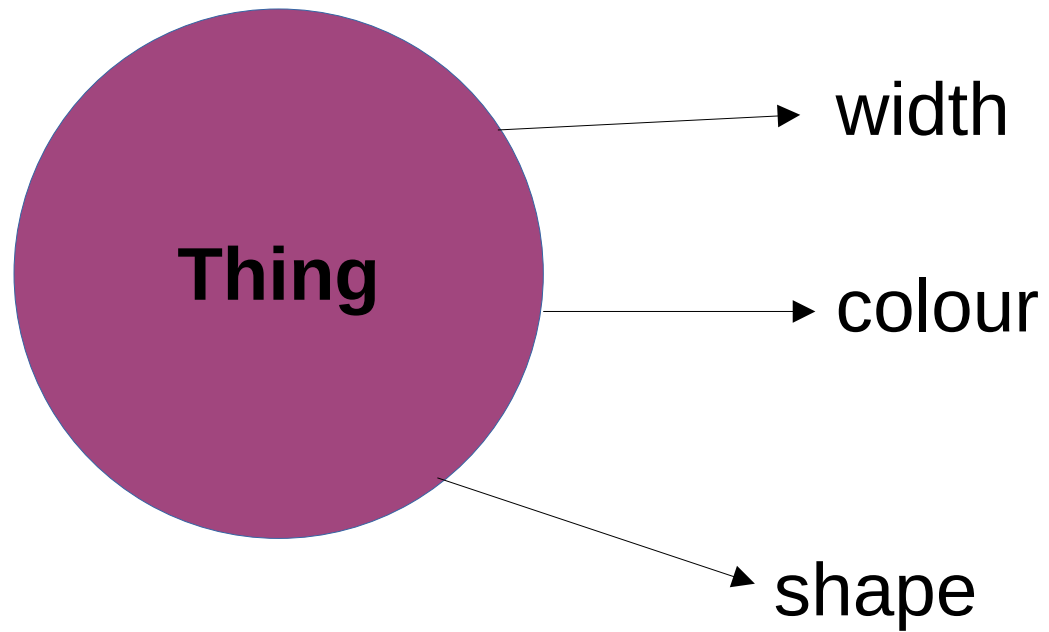
Thing



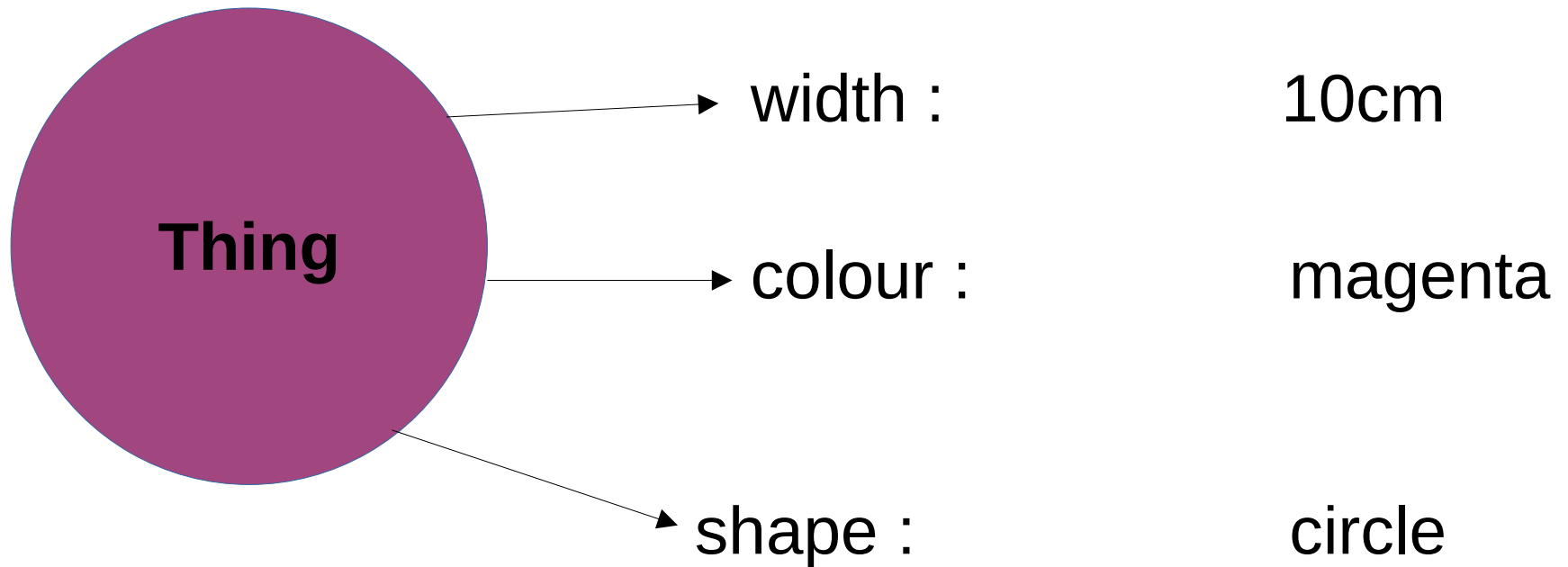
What are these?



Properties

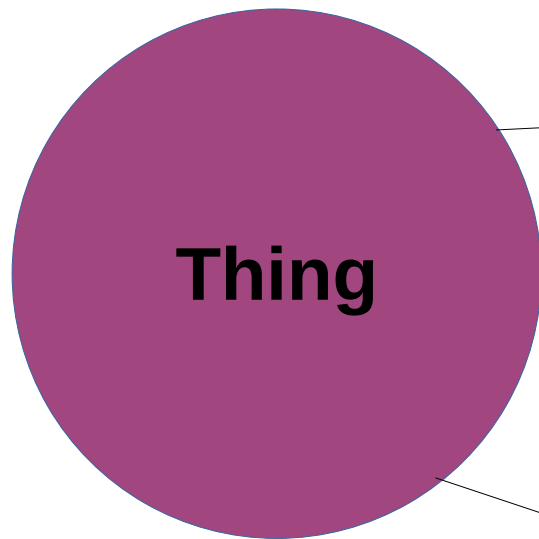


Properties



What are these?

Properties



→ width :

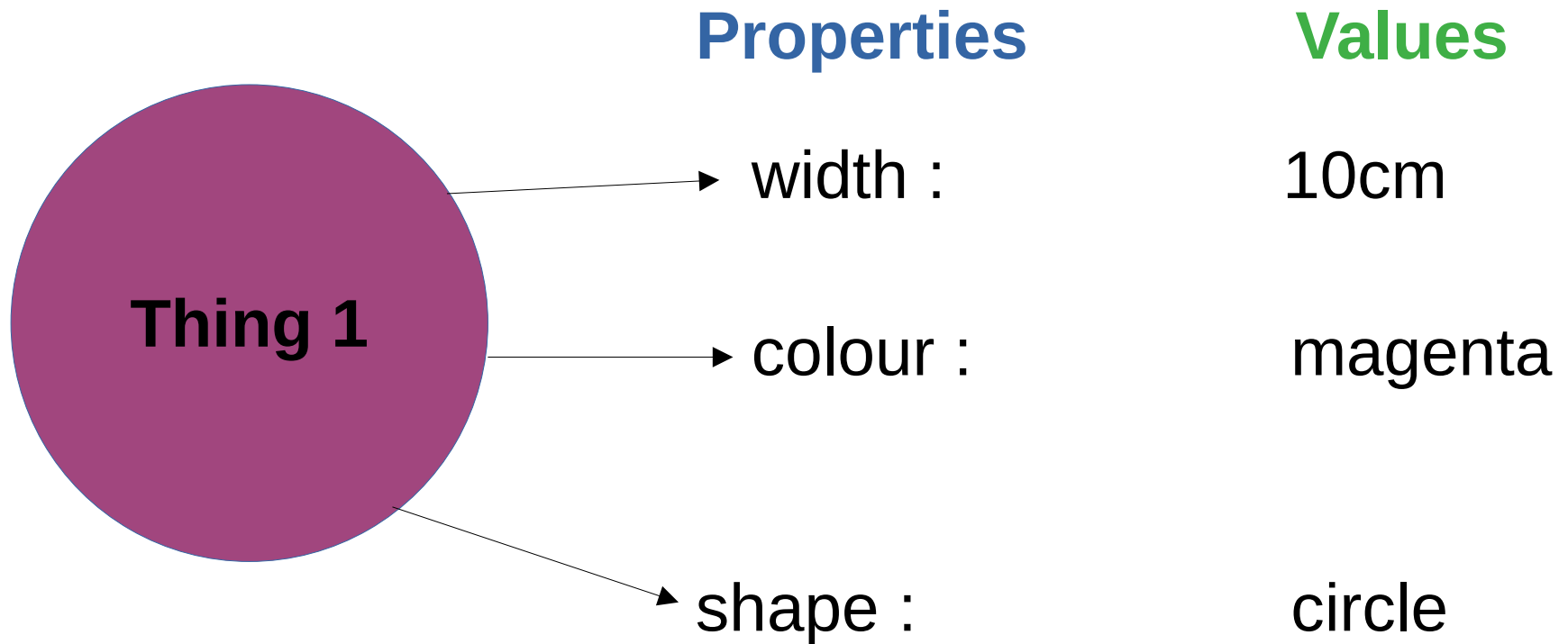
→ colour :

→ shape :

10cm

magenta

circle





Properties

Values

→ width :

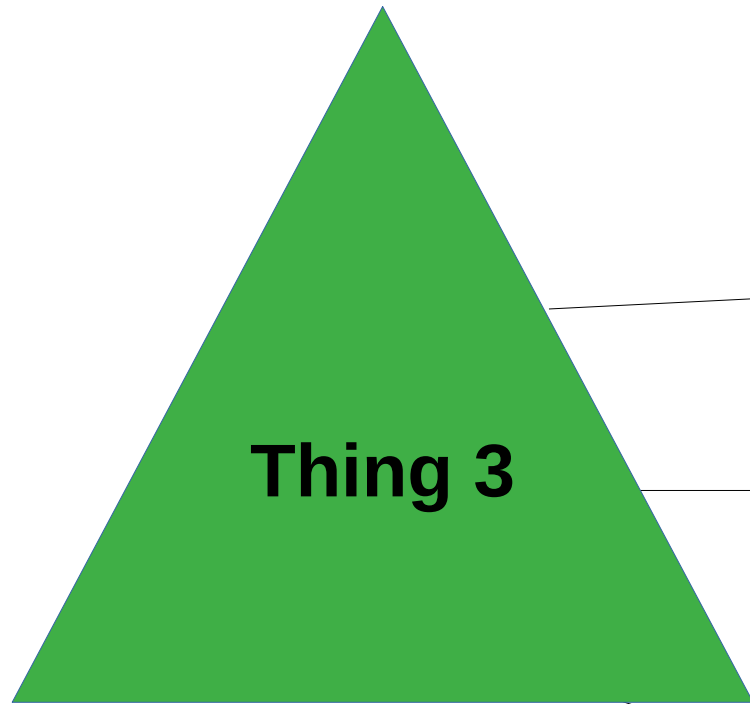
10cm

→ colour :

blue

→ shape :

rectangle



Properties

Values

→ width :

15cm

→ colour :

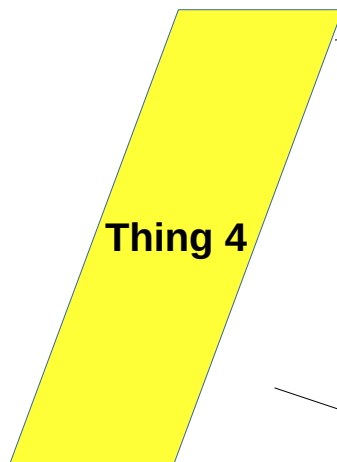
green

→ shape :

triangle

Properties

Values



width :

7cm

colour :

yellow

shape :

rhomboid

Properties

Values



width :

20cm

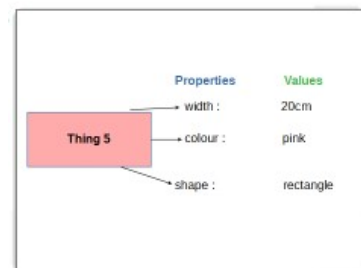
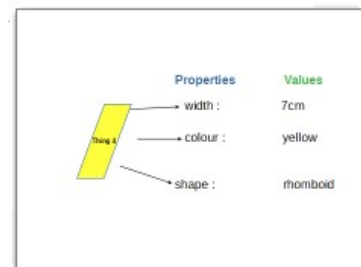
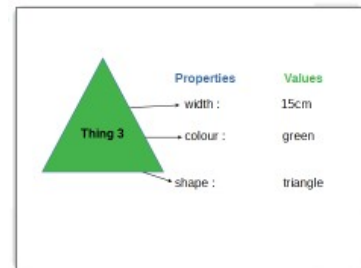
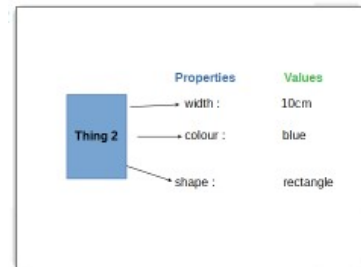
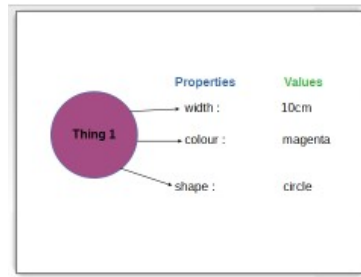
colour :

pink

shape :

rectangle

WE HAVE
5 things
(instances,
observations,
examples)

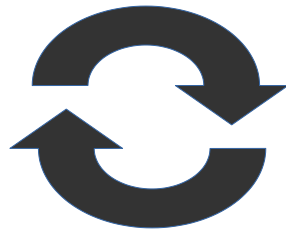


WE HAVE
4 properties (variables,
attributes, features)

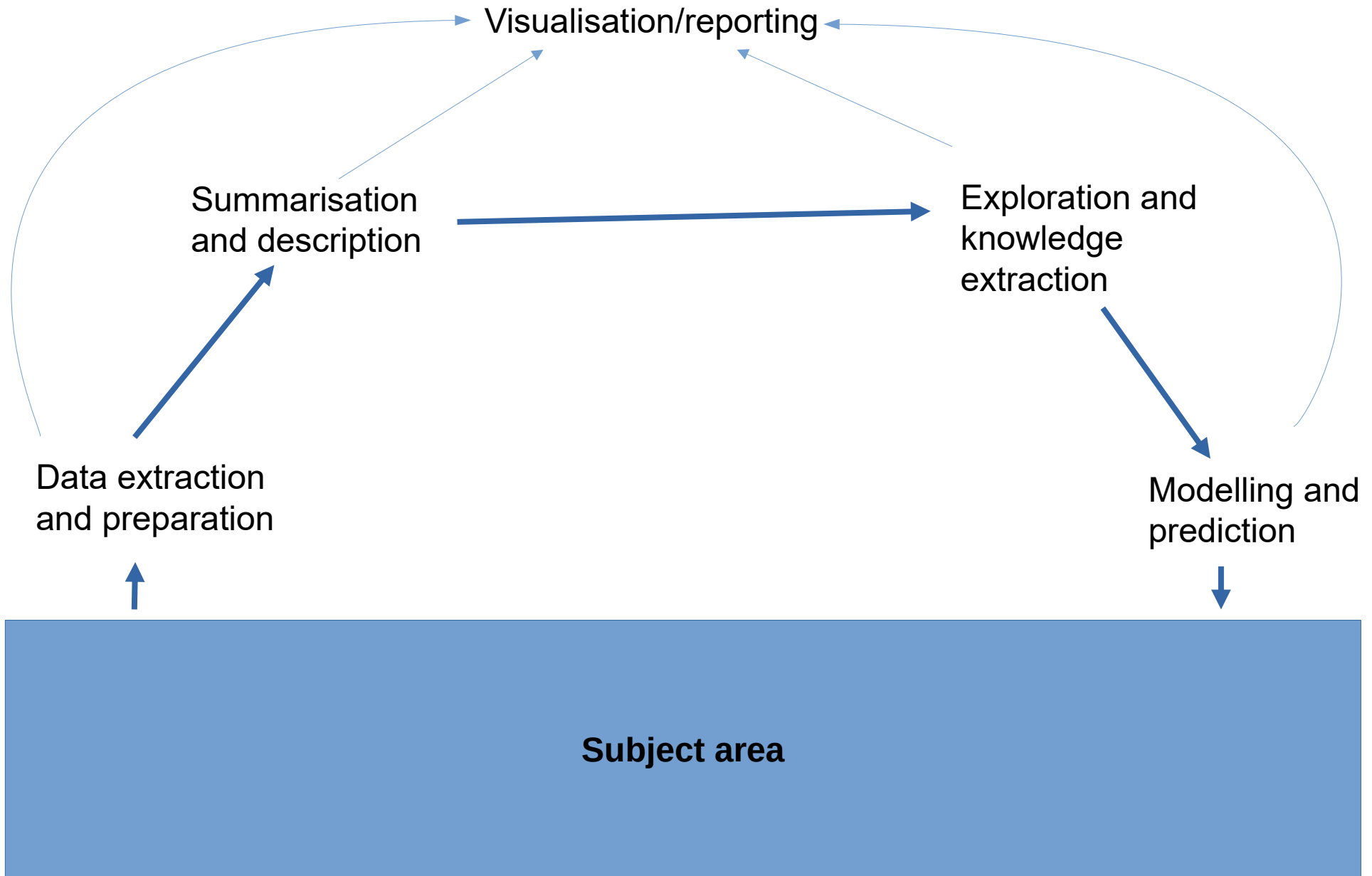
ID	width (cm)	colour	shape
1	10	magenta	circle
2	10	blue	rectangle
3	15	green	triangle
4	7	yellow	rhomboid
5	20	pink	rectangle

This table is what we analyse

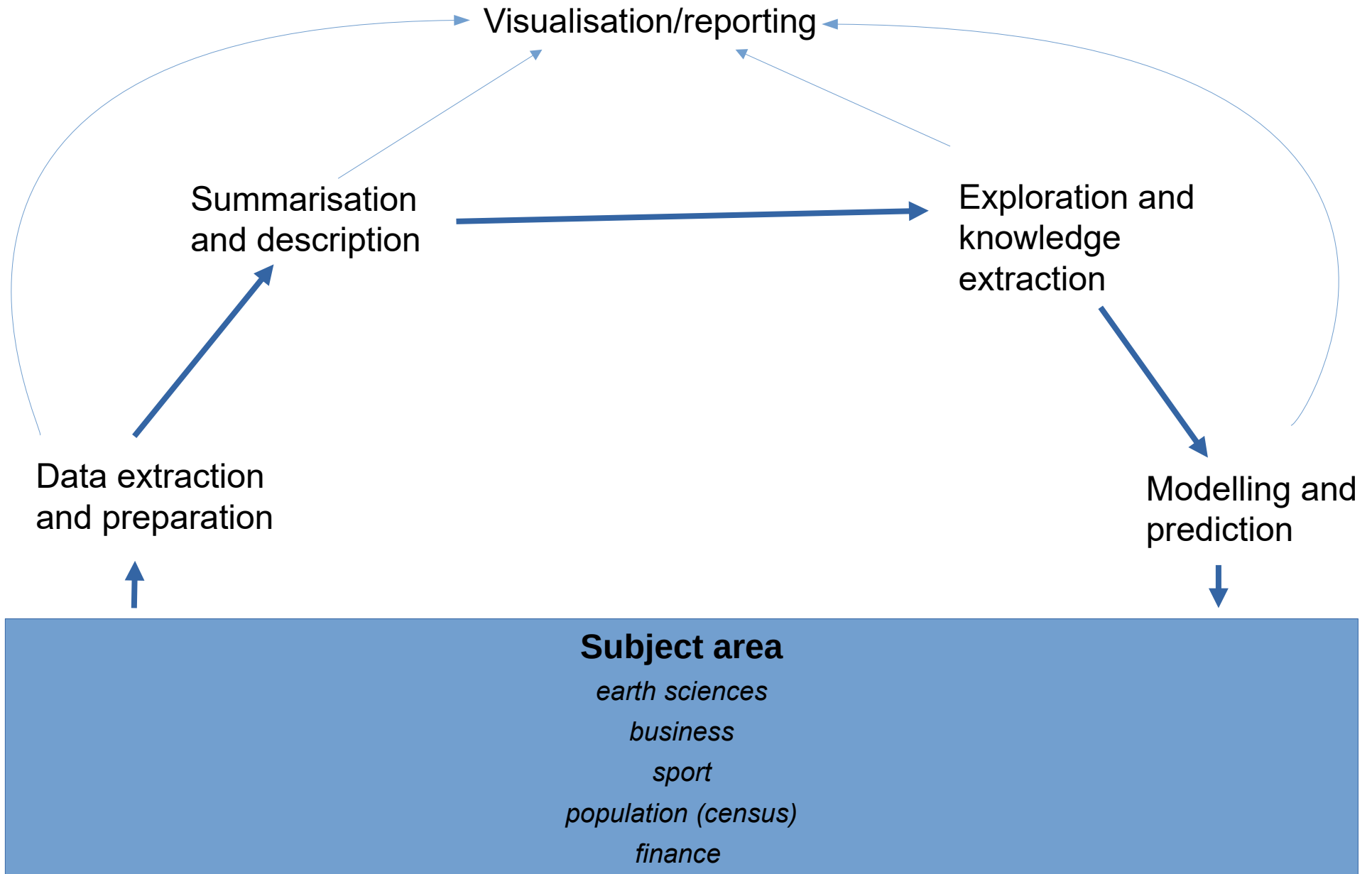
The data cycle



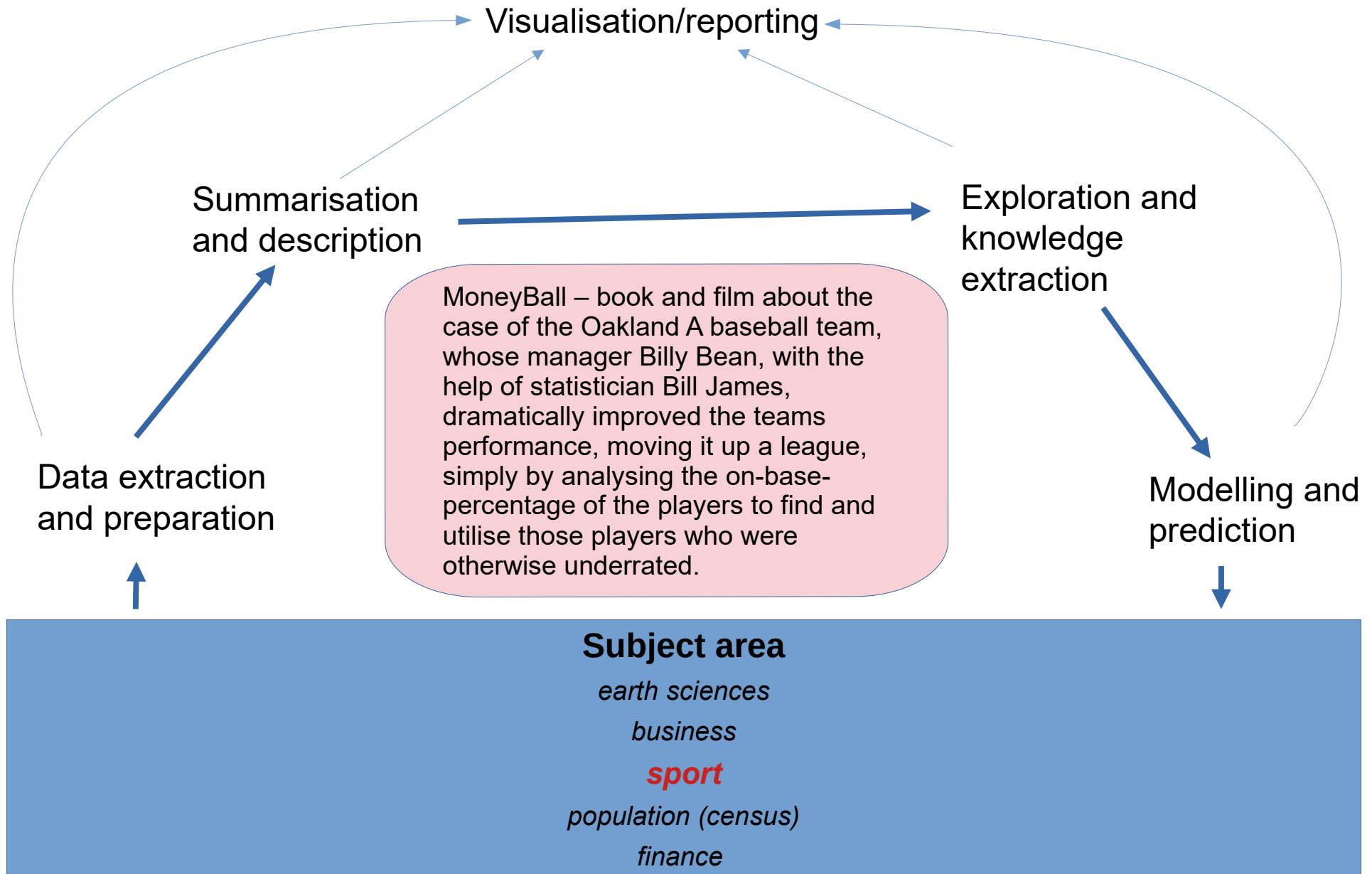
The data cycle



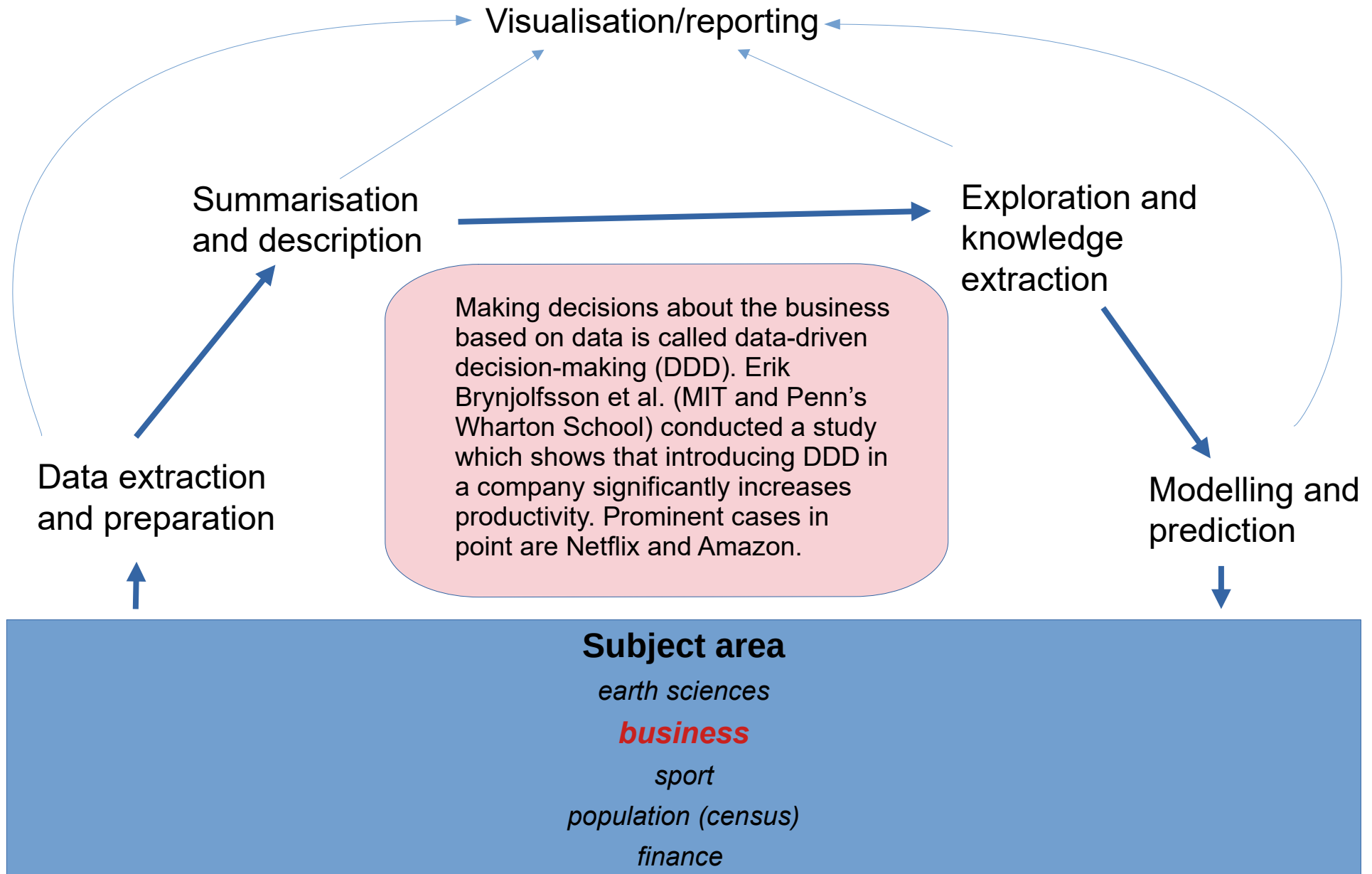
The data cycle



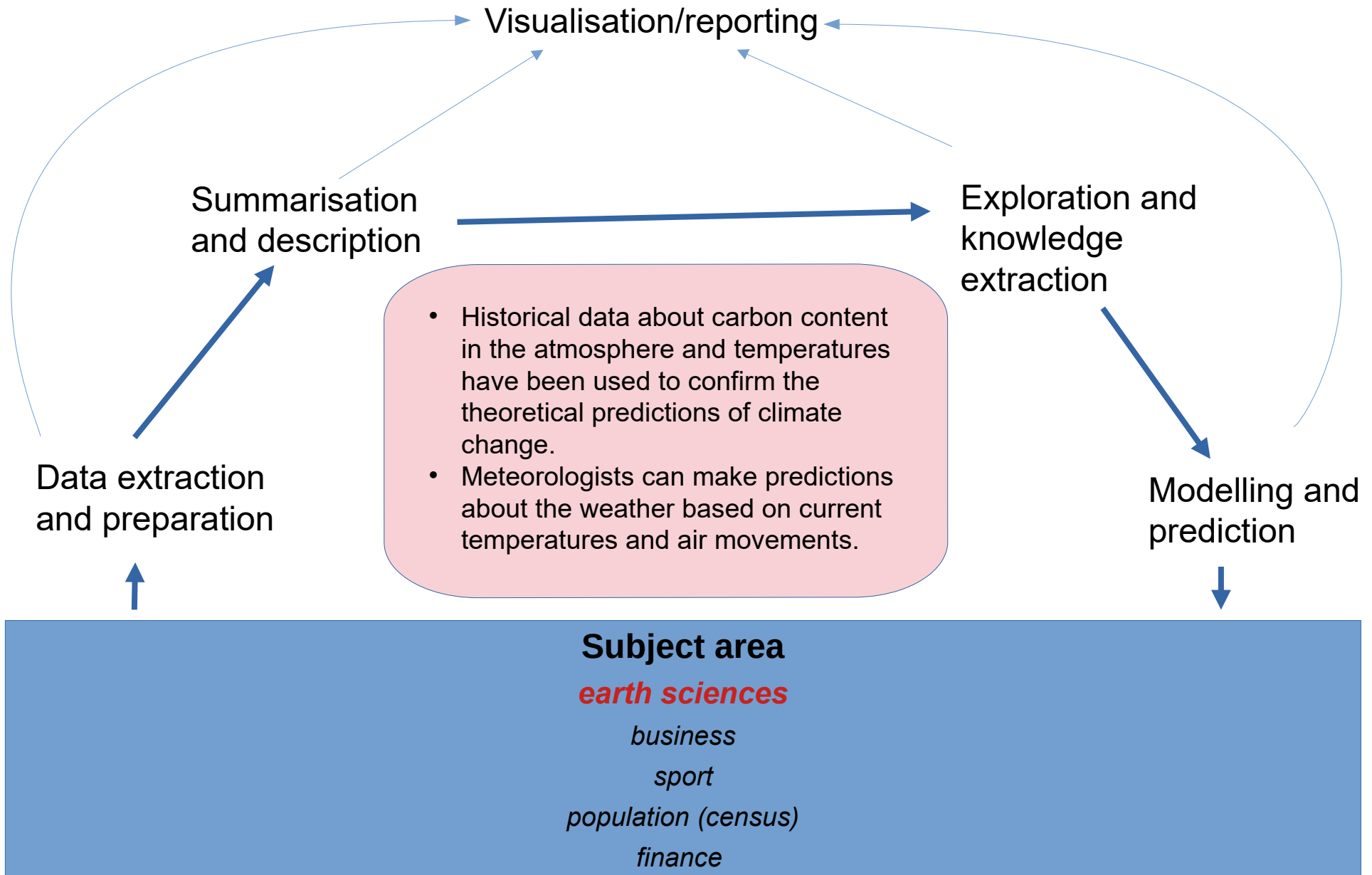
The data cycle



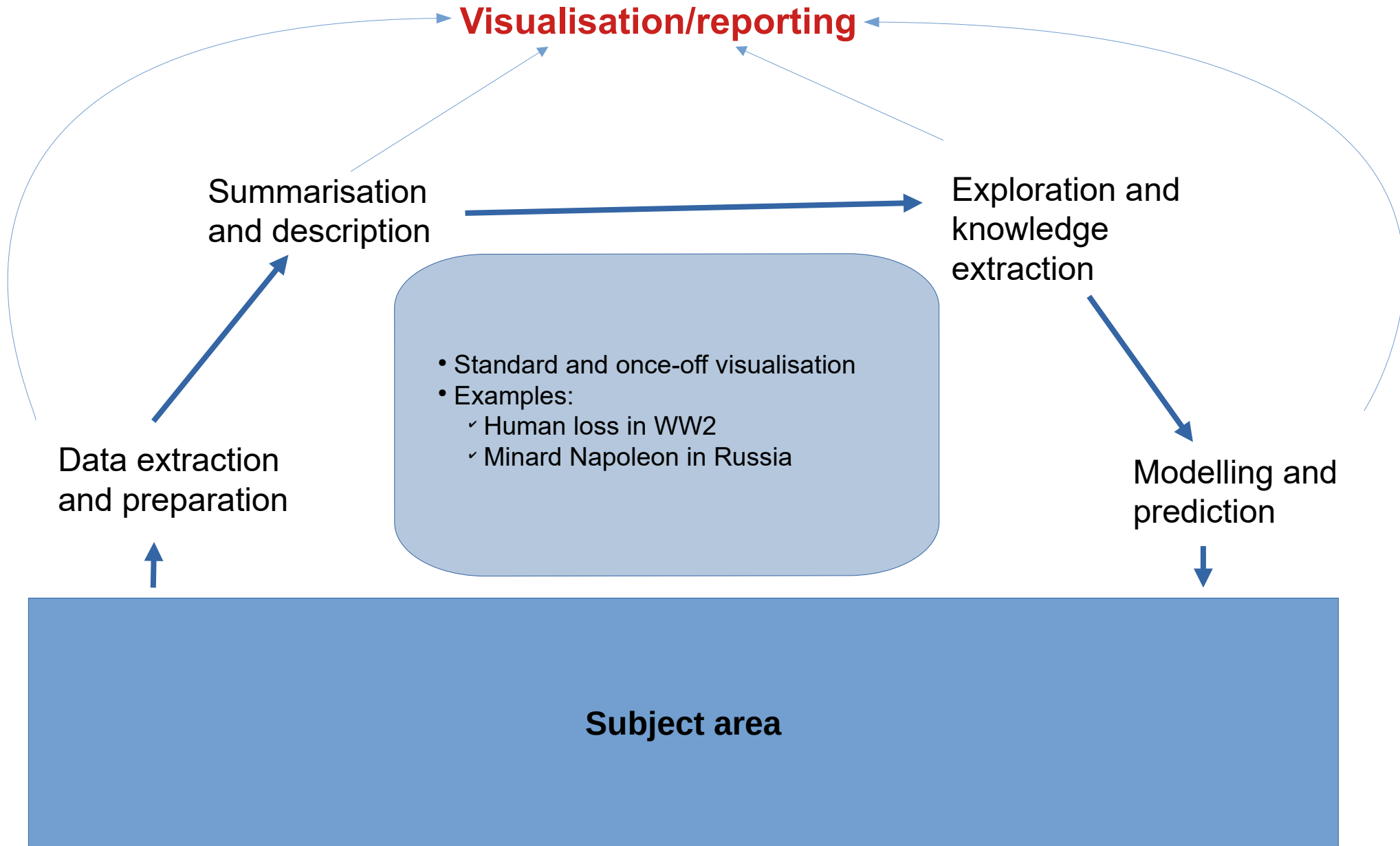
The data cycle



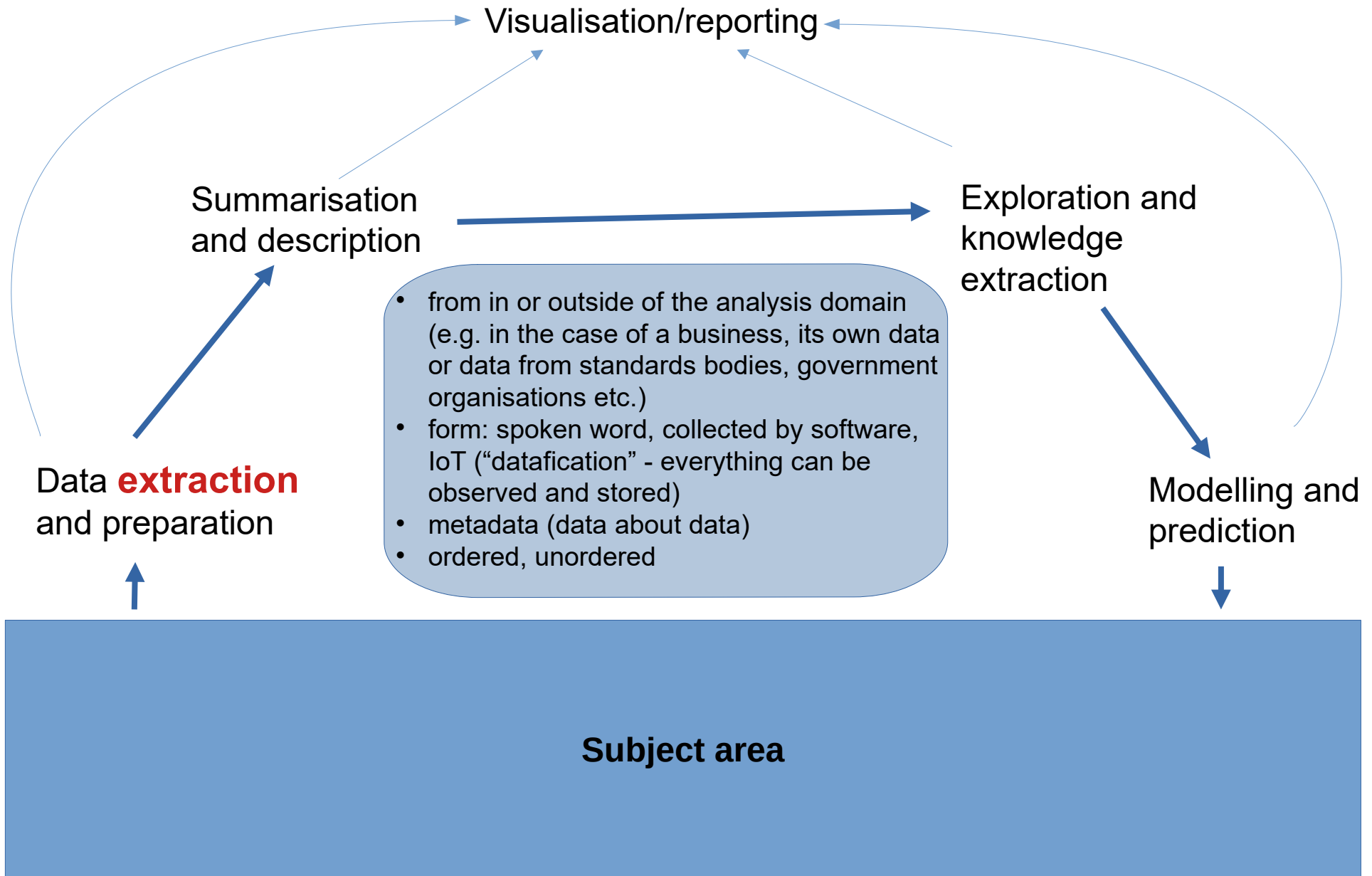
The data cycle



The data cycle



The data cycle




```
graph TD; SA[Subject area] --> DEP[Data extraction and preparation]; DEP --> SD[Summarisation and description]; SD --> VR[Visualisation/reporting]; VR --> EKE[Exploration and knowledge extraction]; EKE --> MP[Modelling and prediction]; MP --> SA; EKE --> VR; SD --> EKE; DEP --> VR
```

The diagram illustrates the data science process, starting from the **Subject area** (represented by a large blue rectangle at the bottom). The process flows through several stages:

- Data extraction and preparation** (indicated by an upward arrow from the Subject area).
- Summarisation and description** (indicated by an upward arrow from Data extraction and preparation).
- Visualisation/reporting** (indicated by an upward arrow from Summarisation and description).
- Exploration and knowledge extraction** (indicated by a horizontal arrow from Summarisation and description).
- Modelling and prediction** (indicated by a downward arrow from Exploration and knowledge extraction, which then points back to the Subject area).

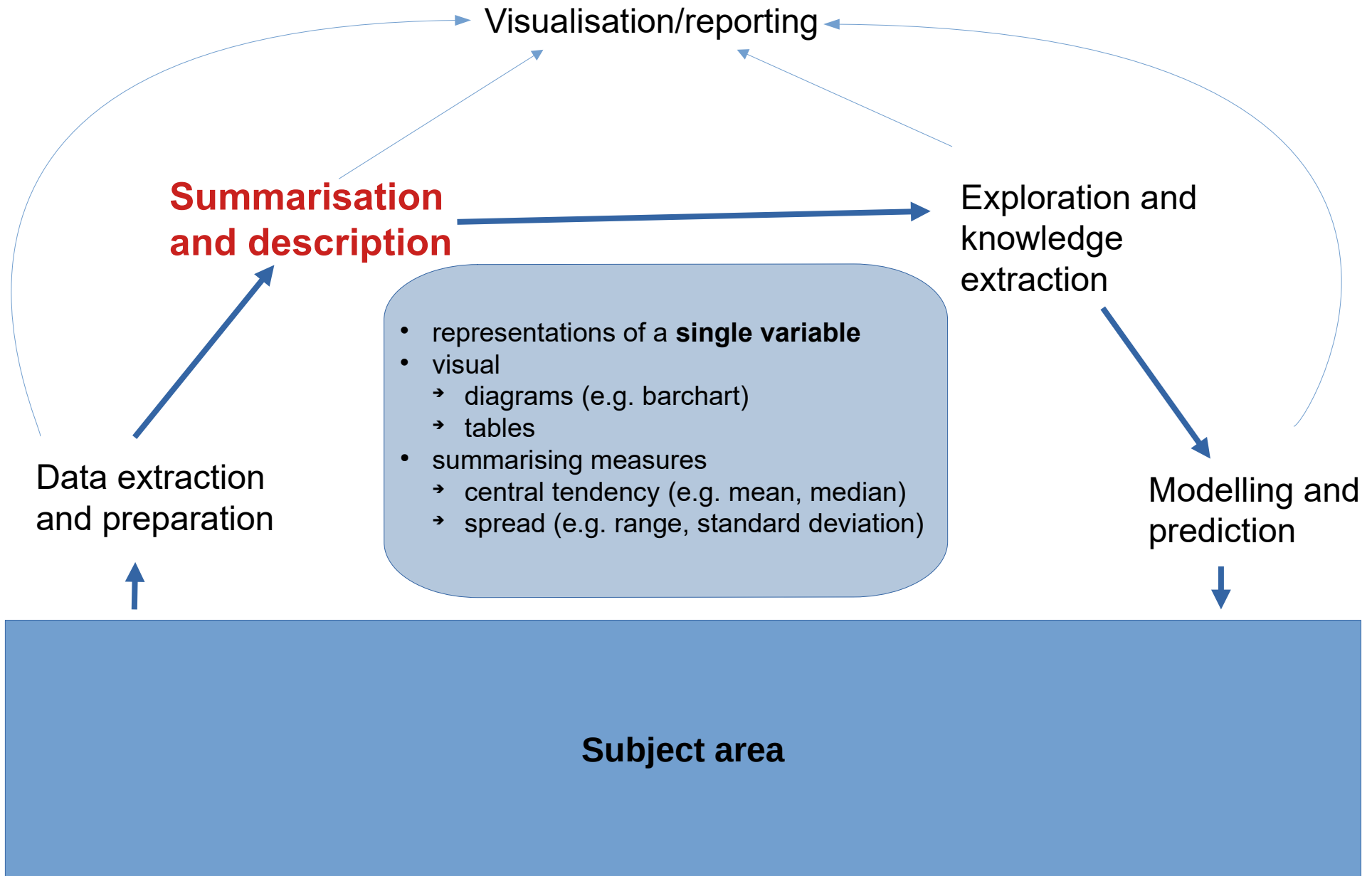
A central box lists the tasks involved in data preparation:

- cleaning the data (e.g. removing non-numerical values and converting text to numbers)
- removing outliers
- scaling data if combining multiple data-sets
- combining variables
- converting continuous data to categories

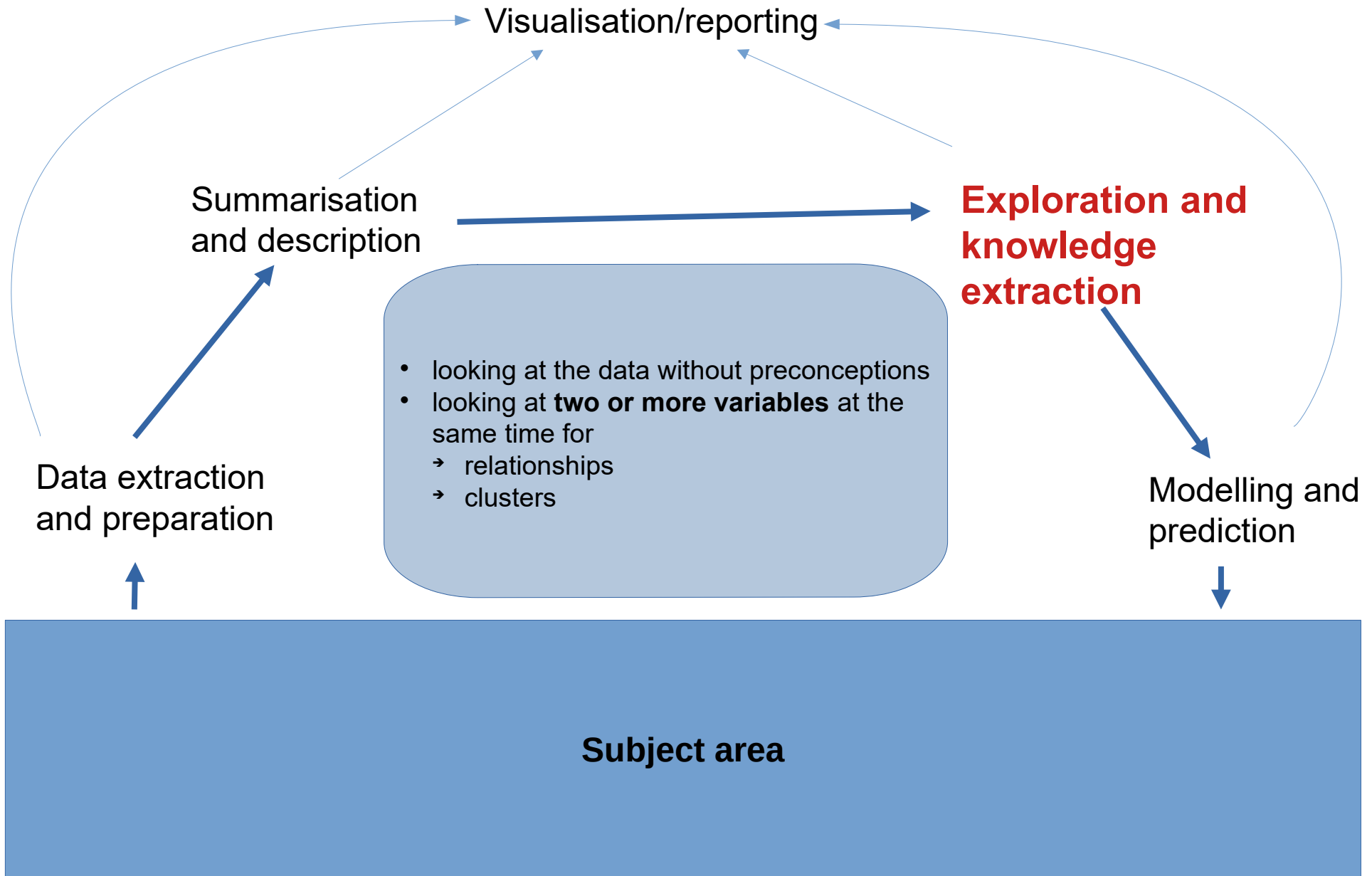
Curved arrows at the top indicate feedback loops from **Visualisation/reporting** back to **Data extraction and preparation** and **Exploration and knowledge extraction**.

- cleaning the data (e.g. removing non-numerical values and converting text to numbers)
- removing outliers
- scaling data if combining multiple data-sets
- combining variables
- converting continuous data to categories

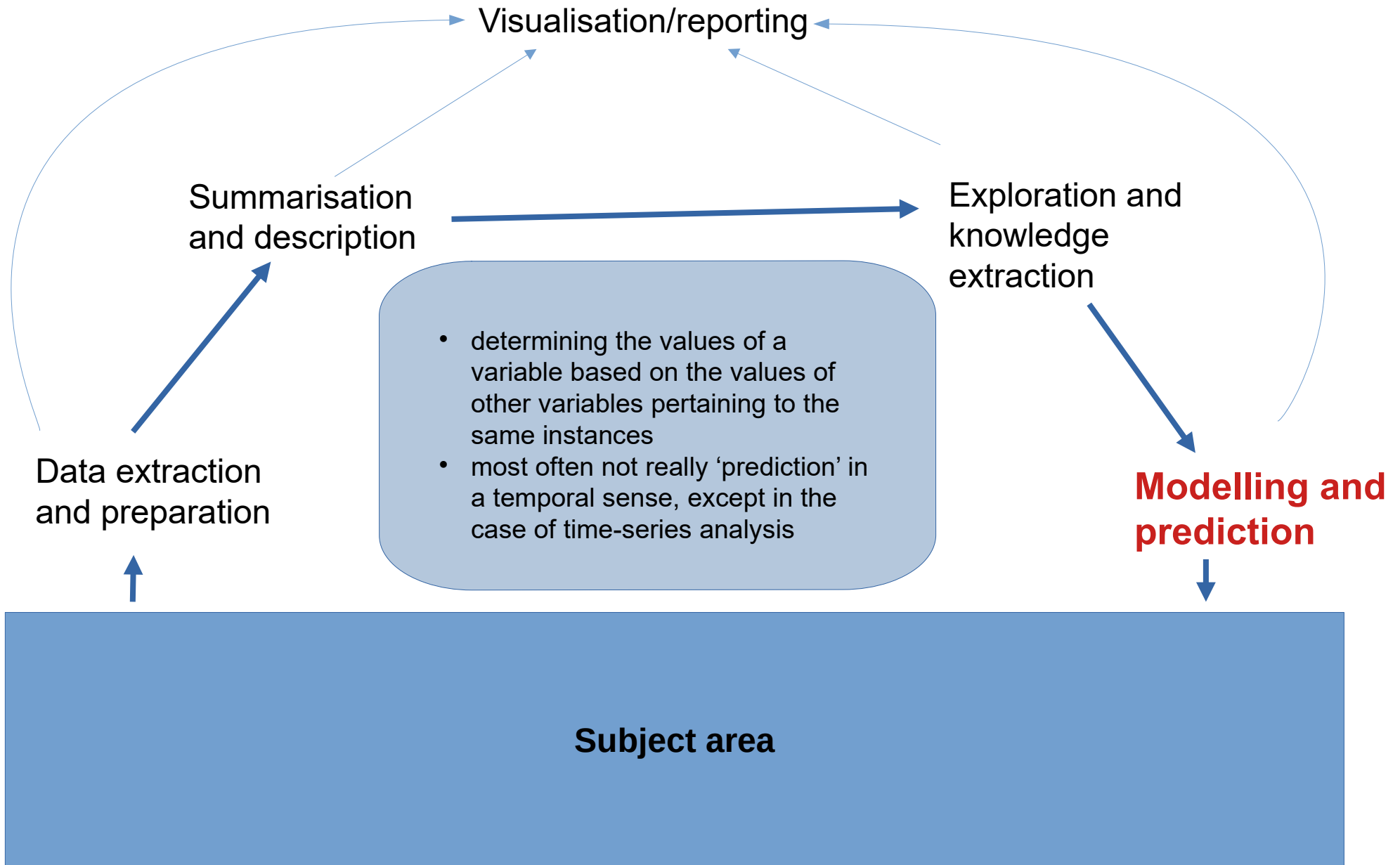
The data cycle



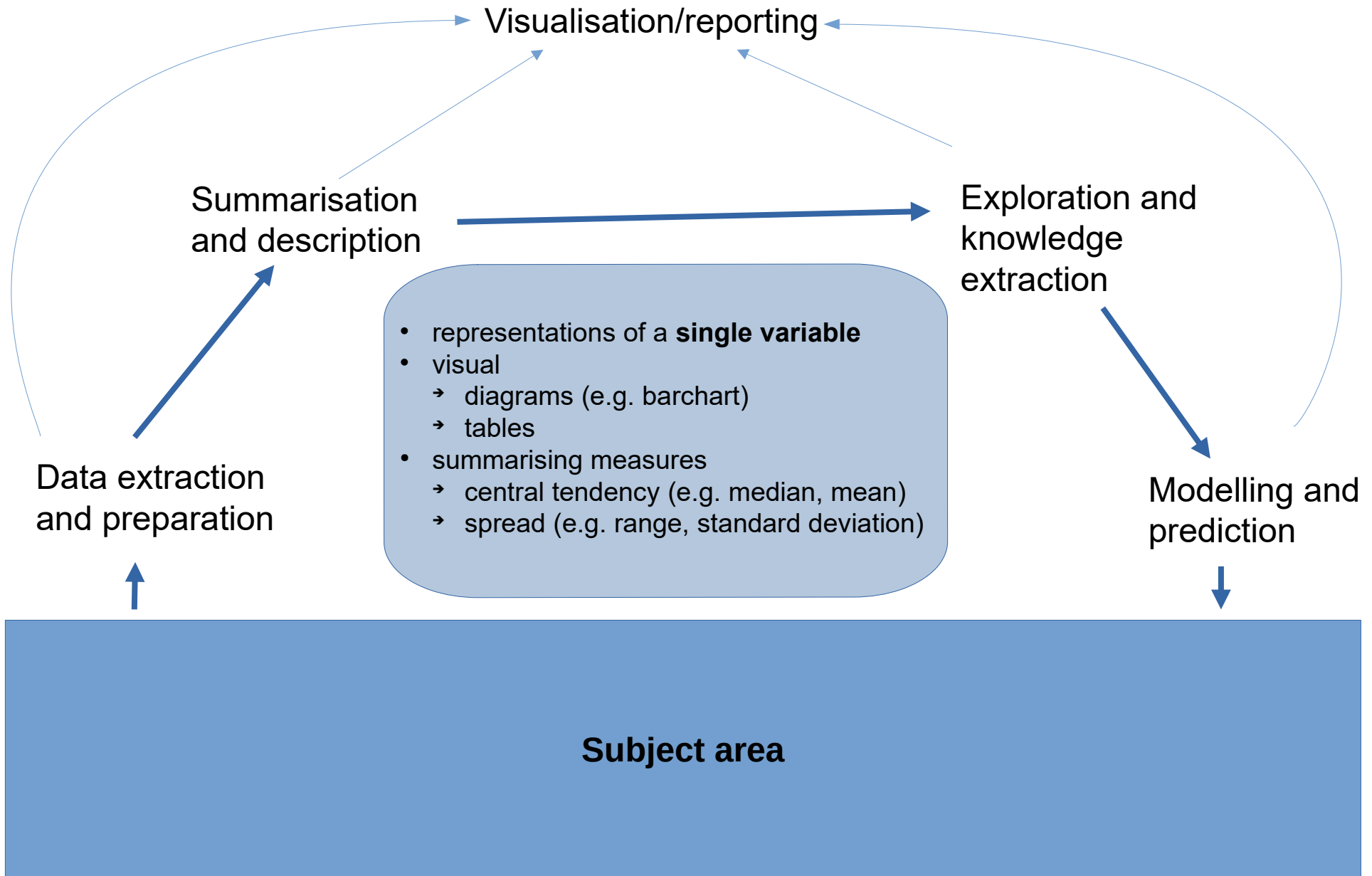
The data cycle



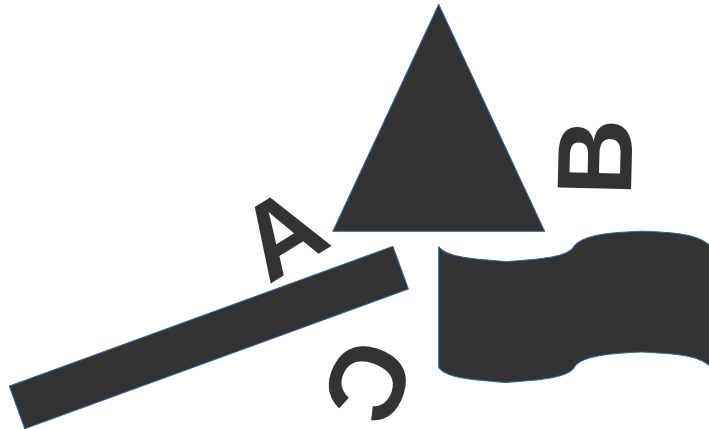
The data cycle



The data cycle



Landscape and terminology relating to data analysis



Landscape and terminology

Data science, including
theory behind statistics
and machine learning

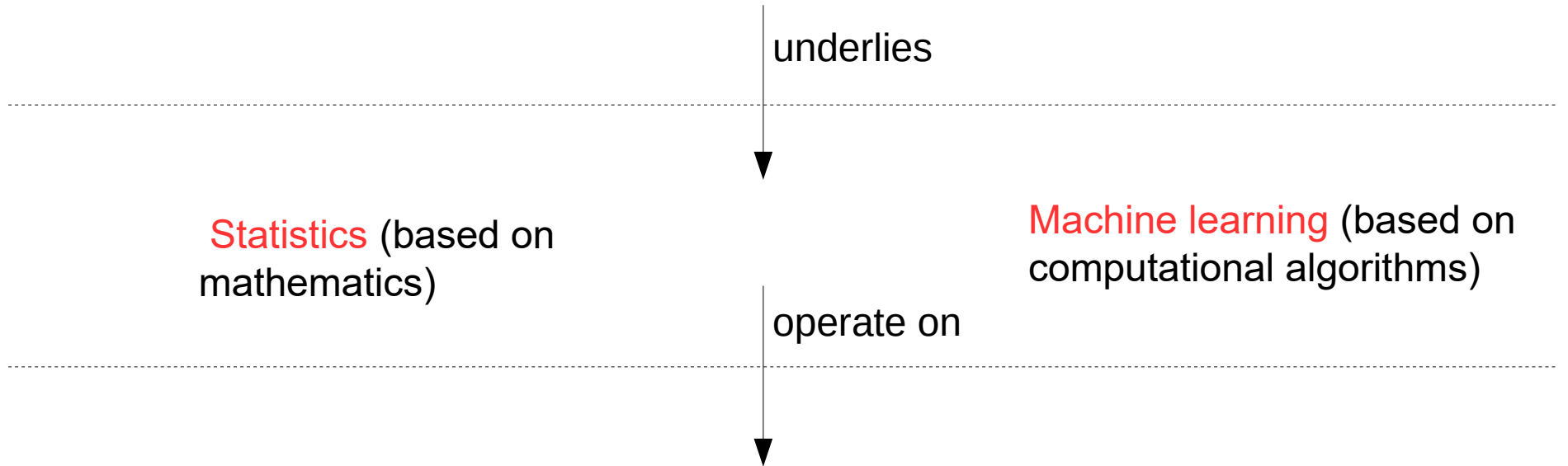
underlies

Statistics (based on
mathematics)

Machine learning (based on
computational algorithms)

operate on

Prepared **data**



Landscape and terminology

- Sets out the principles and theory for understanding and using data
- Studies how these principles and techniques should be applied in each individual case

Data science, including theory behind statistics and machine learning

underlies

Statistics (based on mathematics)

Machine learning (based on computational algorithms)

operate on

Prepared **data**

Landscape and terminology

- The science and practice of analysing numerical data, particularly with the purpose of understanding the properties of a large population by analysing a representative sample.

Statistics (based on mathematics)

Data science, including theory behind statistics and machine learning

underlies

Machine learning (based on computational algorithms)

operate on

Prepared **data**

Landscape and terminology

Data science, including theory behind statistics and machine learning

underlies

- Supervised learning – goal is prediction based on past data (e.g. classification, regression)
- Unsupervised learning – exploratory (e.g. association rules, clustering)

Statistics (based on mathematics)

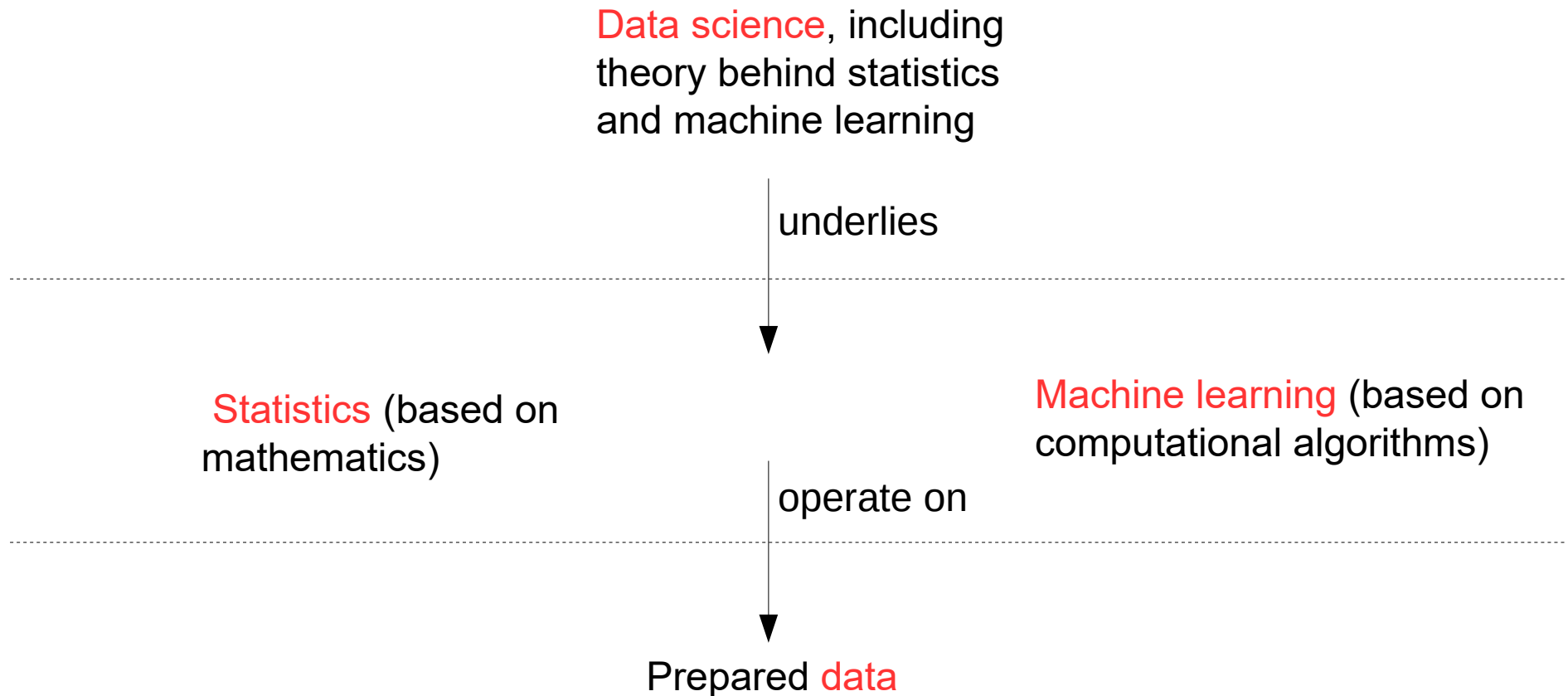
operate on

Machine learning (based on computational algorithms)

Prepared **data**

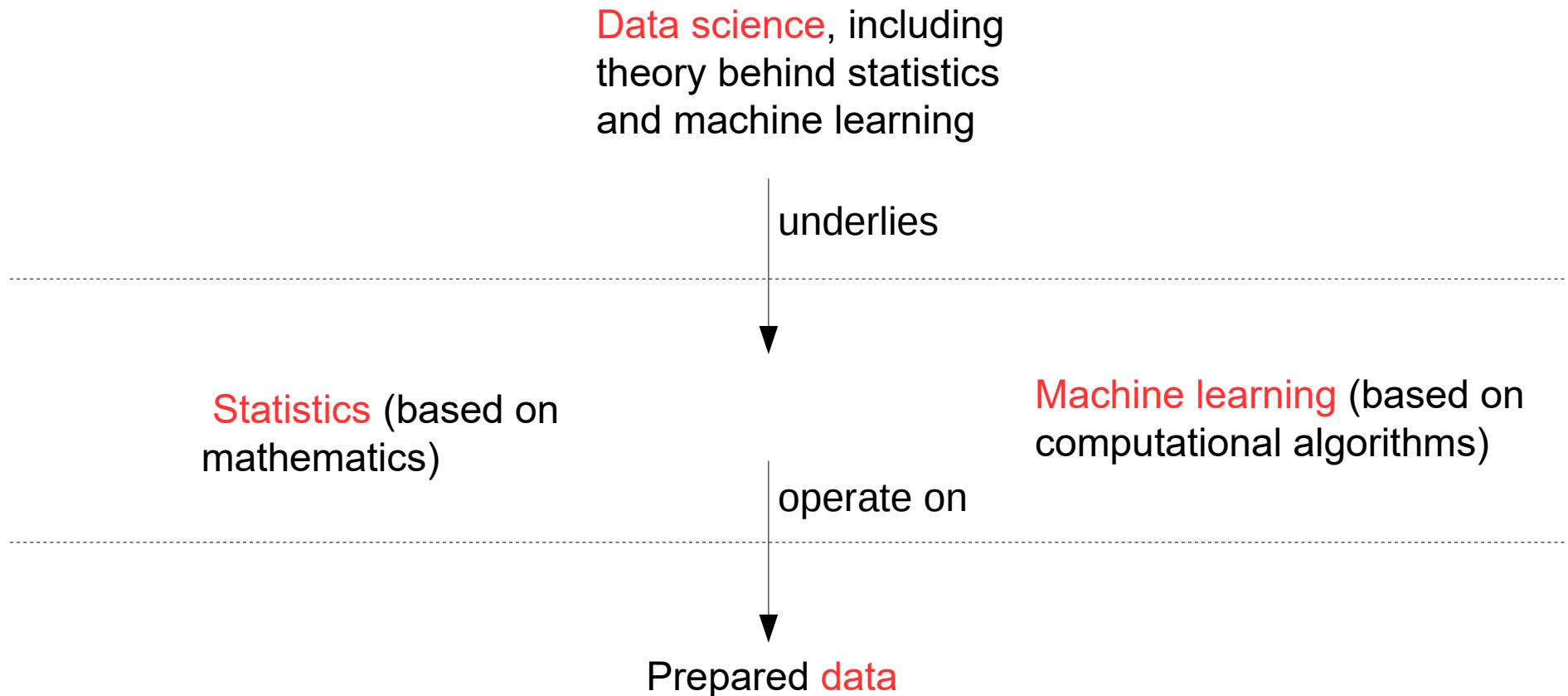
Landscape and terminology

- **Analytics** – techniques and activities based around using data in a problem domain e.g. business analytics, financial analytics (the middle of the ‘sandwich’ below).



Landscape and terminology

- **Analytics** – techniques and activities based around using data in a problem domain e.g. business analytics, financial analytics (the middle of the ‘sandwich’ below).
- **Data analysis** – the activity of analysing data in any way.



Landscape and terminology

- **Analytics** – techniques and activities based around using data in a problem domain e.g. business analytics, financial analytics (the middle of the ‘sandwich’ below).
- **Data analysis** – the activity of analysing data in any way.
- **Data mining** – exploration of (predominantly) enterprise data warehouses (1990s)

Data science, including
theory behind statistics
and machine learning

underlies

Statistics (based on
mathematics)

Machine learning (based on
computational algorithms)

operate on

Prepared **data**

Landscape and terminology

- **Analytics** – techniques and activities based around using data in a problem domain e.g. business analytics, financial analytics (the middle of the ‘sandwich’ below).
- **Data analysis** – the activity of analysing data in any way.
- **Data mining** – exploration of (predominantly) enterprise data warehouses (1990s)
- **Big data** (on next page)

Data science, including
theory behind statistics
and machine learning

underlies

Statistics (based on
mathematics)

Machine learning (based on
computational algorithms)

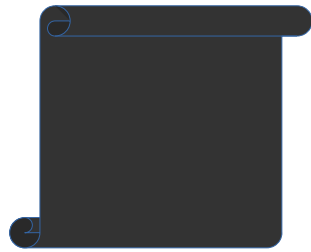
operate on

Prepared **data**

Big Data

- In the last 20 years the data cycle is 'intensifying'
- Growing processing power
- Almost limitless storage capacity
- Connectivity with large bandwidths
- Techniques have developed on this new wave of possibilities
- Big data is at a scale that cannot be processed with conventional technologies.
- 4 Vs
IBM 4Vs of Big Data
- New technologies:
 - Hadoop (Apache)
 - MapReduce (Google)
 - MongoDB etc.
- The data science principles are the same as 'normal sized' data

In this module



you will learn

to...

- ask questions, then investigate if they can be answered by analysing data
- understand and apply methods and techniques for all stages of the data cycle
- decide which methods and techniques to apply depending on scenario
- analyse every individual analysis case on its own merits