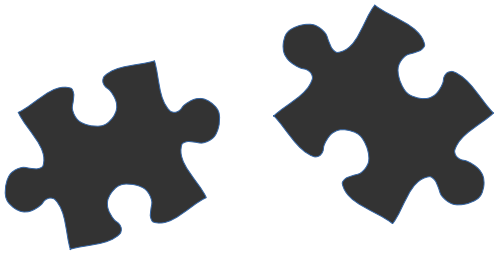
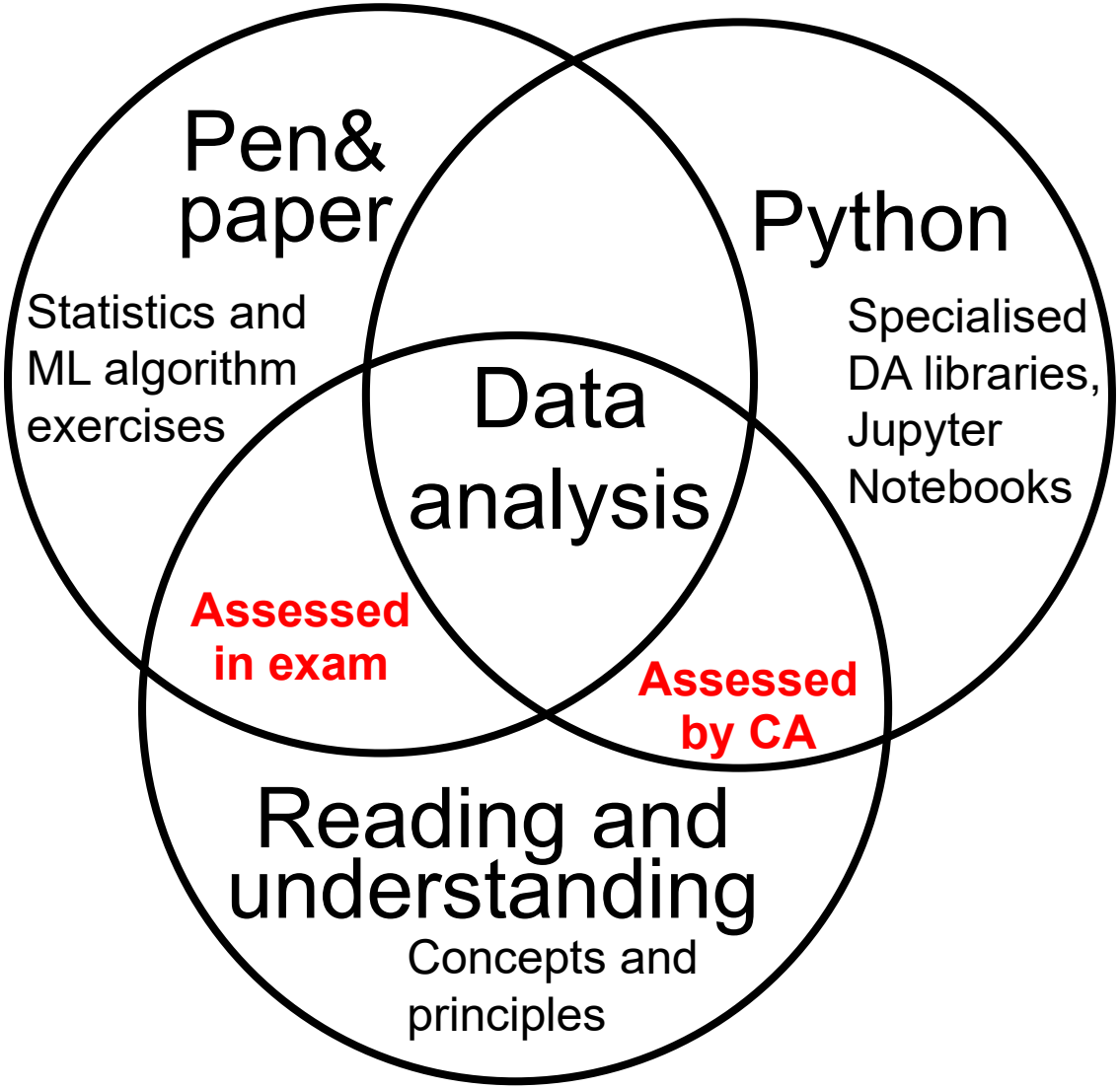
Data Analysis: Introduction

TU Dublin Tallaght Campus,
Department of Computing

In this module

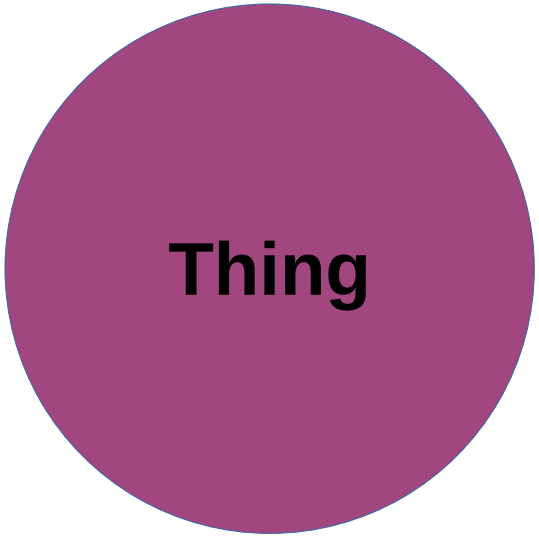


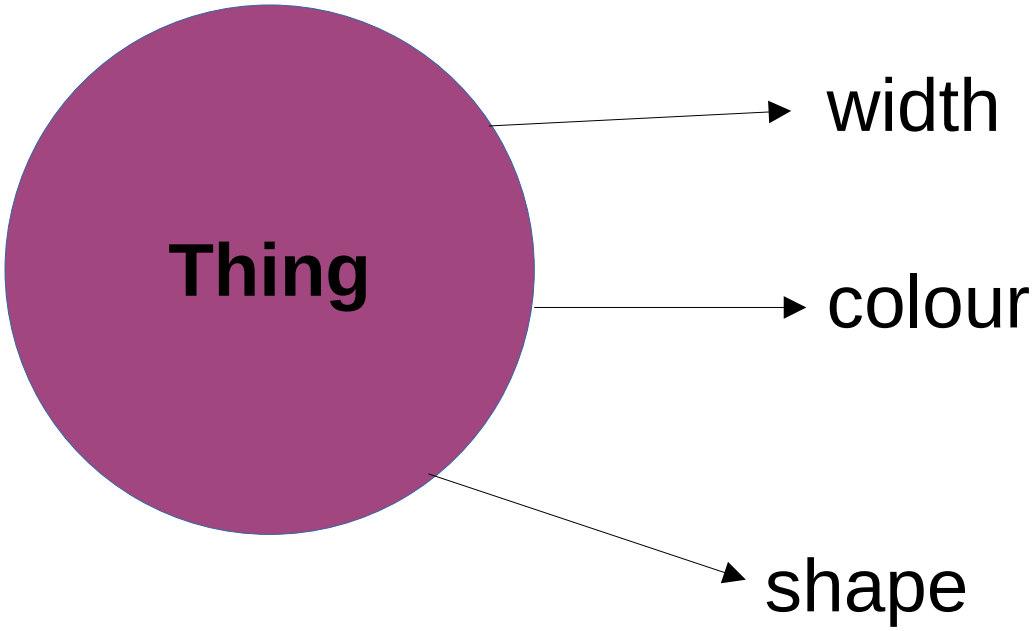
you will find

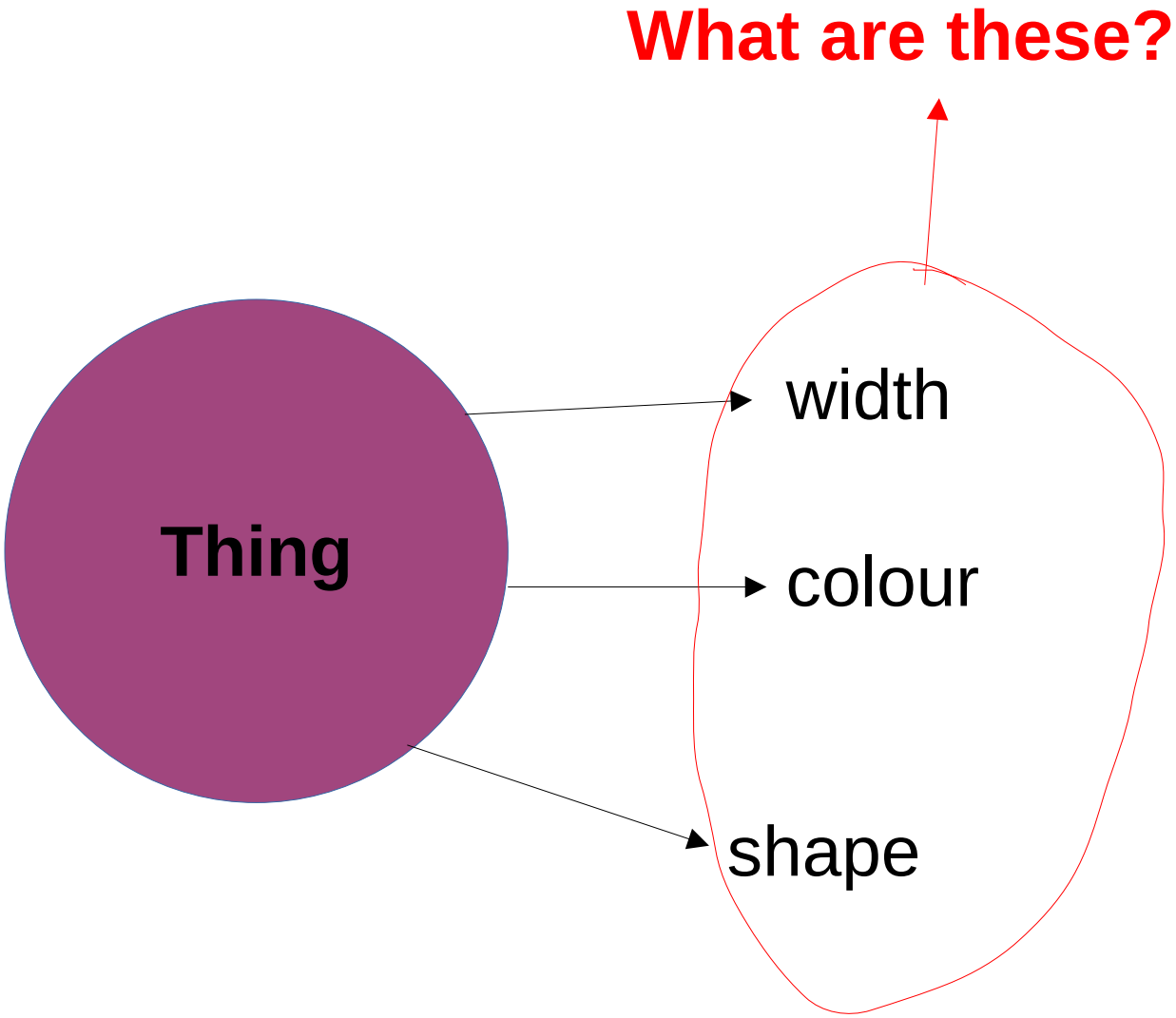

Pen&
paper
Statistics and
ML algorithm
exercises
Python
Specialised
DA libraries,
Jupyter
Notebooks
Data
analysis
Assessed
in exam
Assessed
by CA
Reading and
understanding
Concepts and
principles

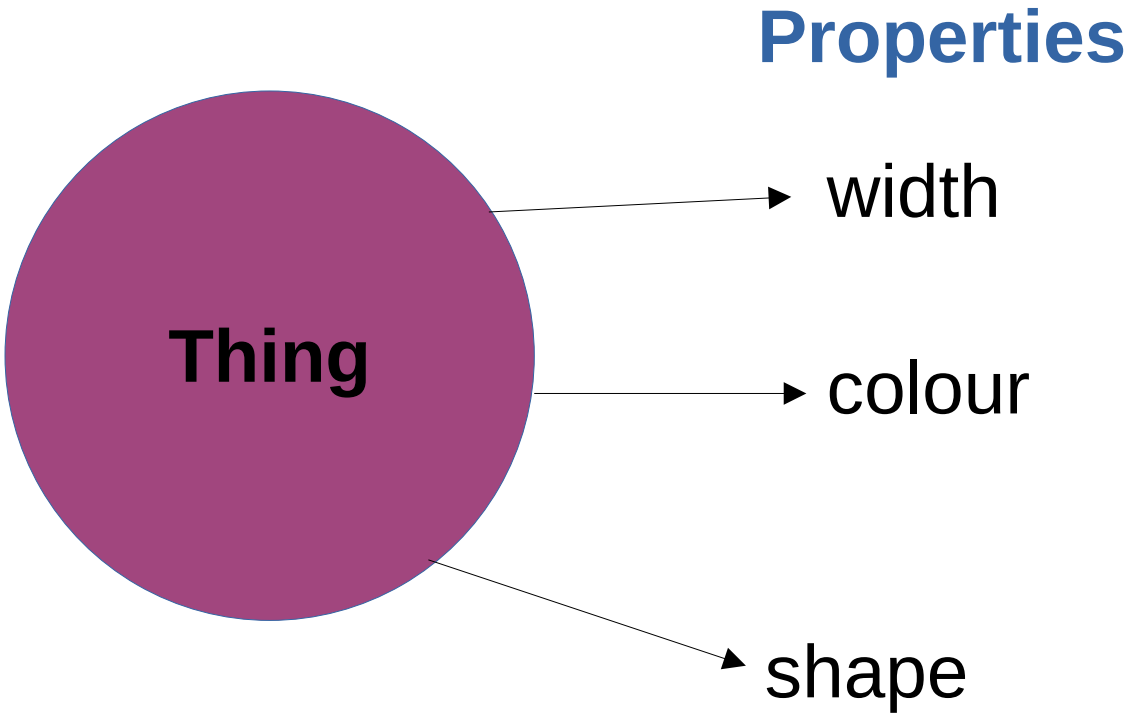
What are we dealing with

?

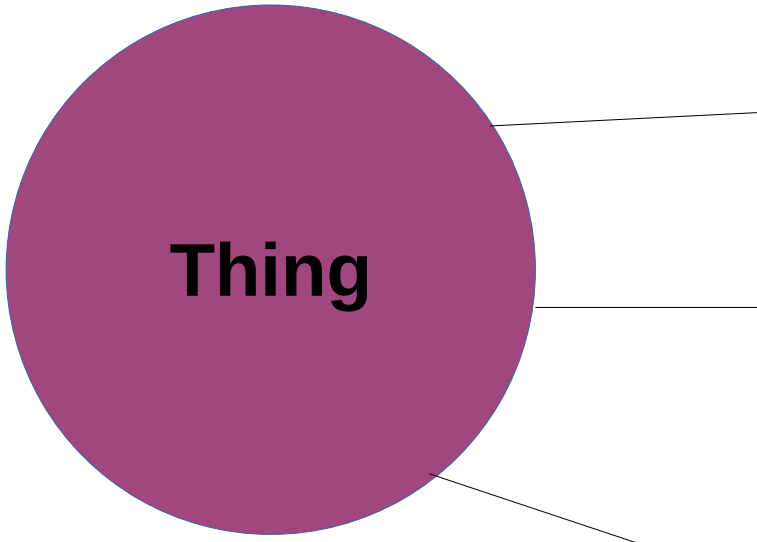

Thing


Thing
width
colour
shape


What are these?
Thing
width
colour
shape


Properties
Thing
width
colour
shape

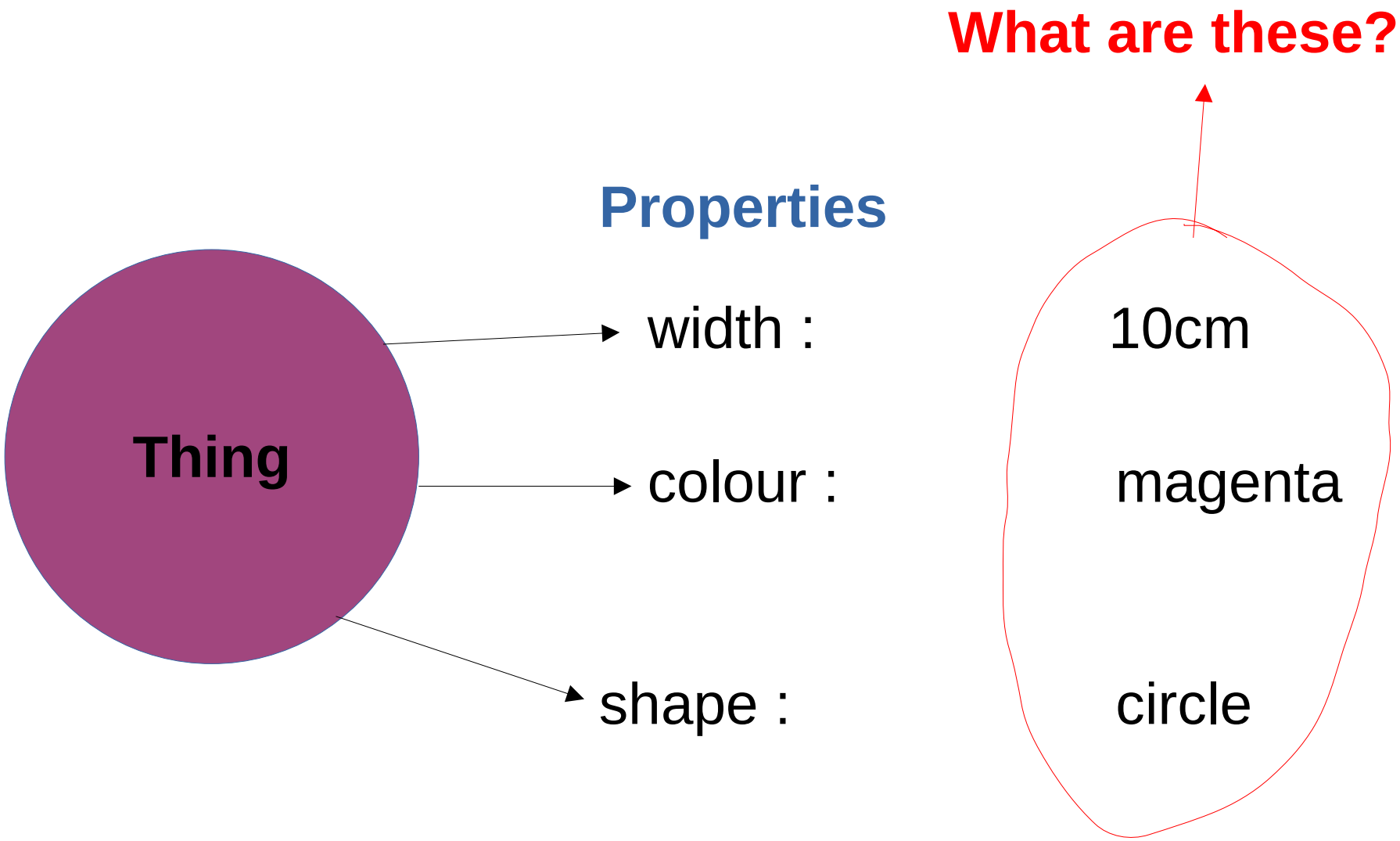
Properties

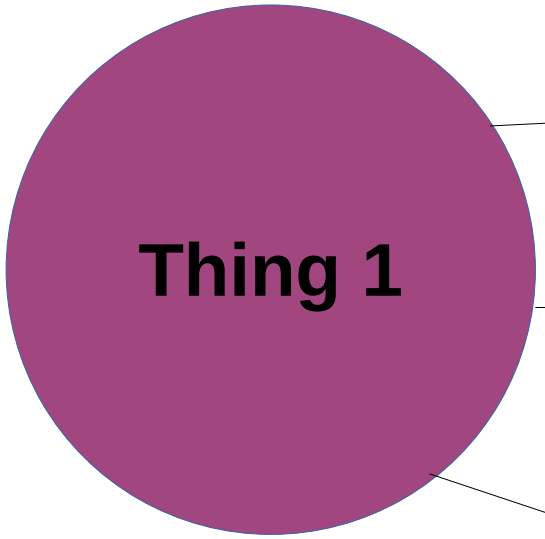


width : 10cm

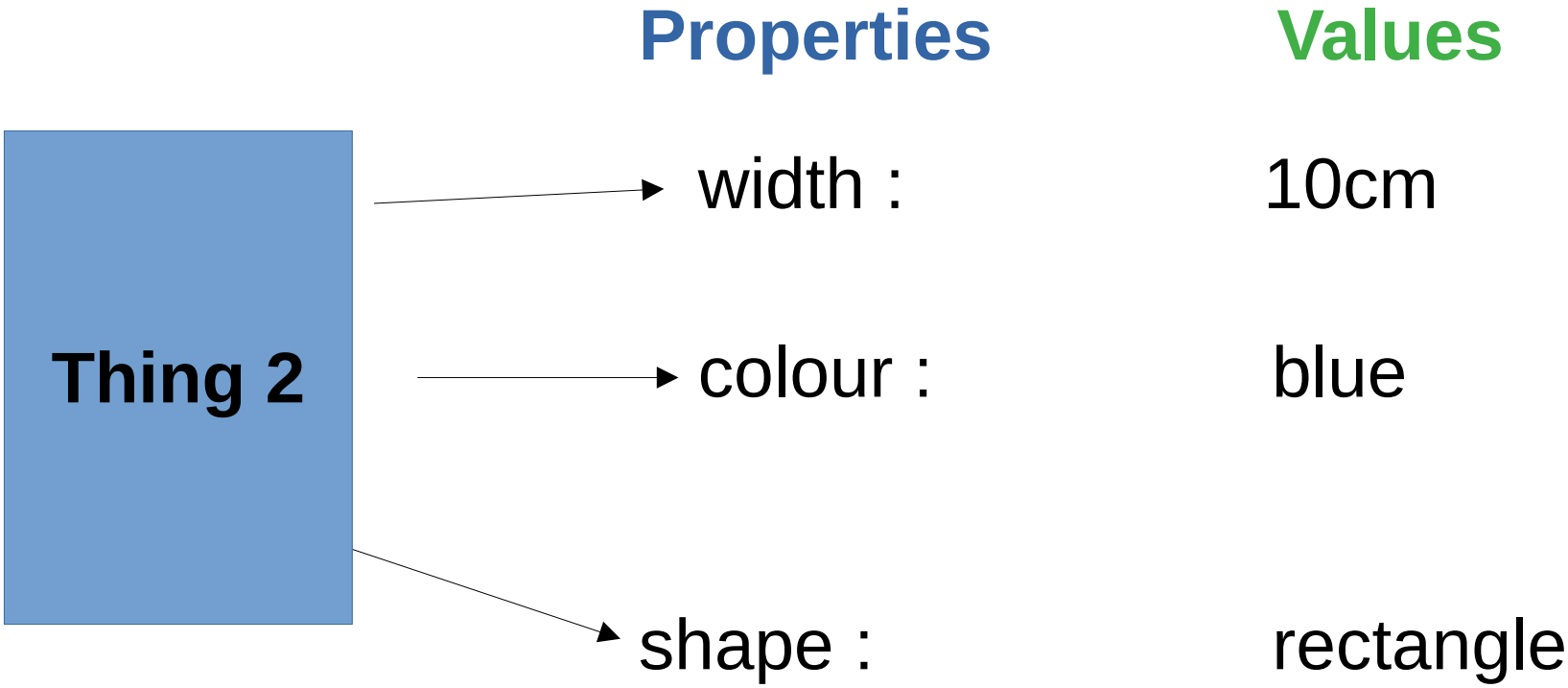
colour : magenta

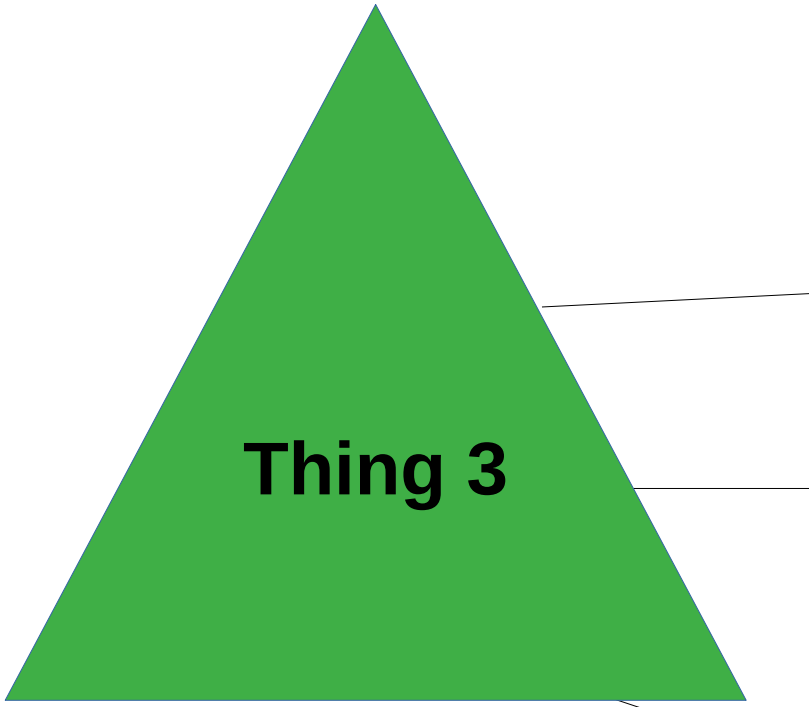
shape : circle


What are these?
Properties
Thing
width :
10cm
colour :
magenta
shape :
circle

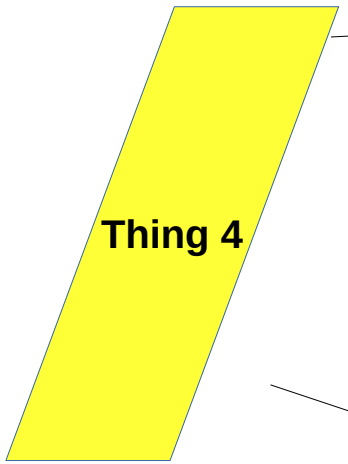


Properties	Values
width :	10cm
colour :	magenta
shape :	circle

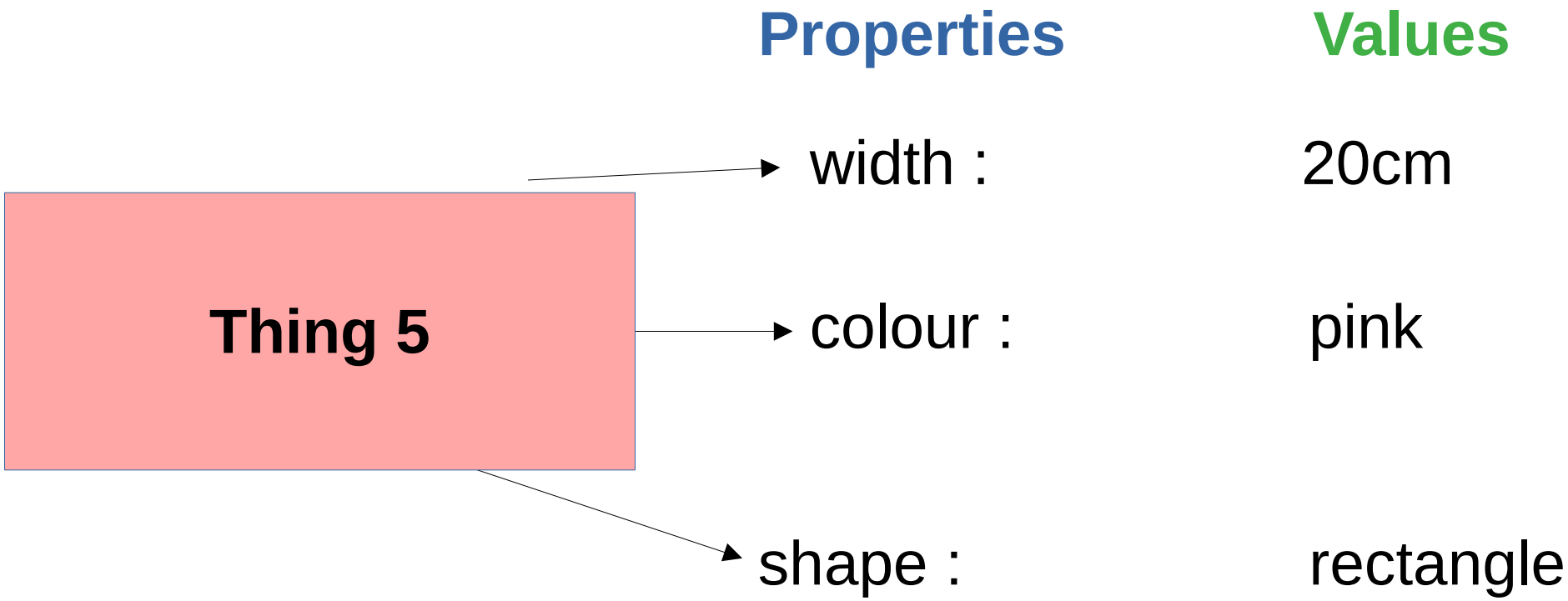

Properties
Values
Thing 2
width :
10cm
colour :
blue
shape :
rectangle



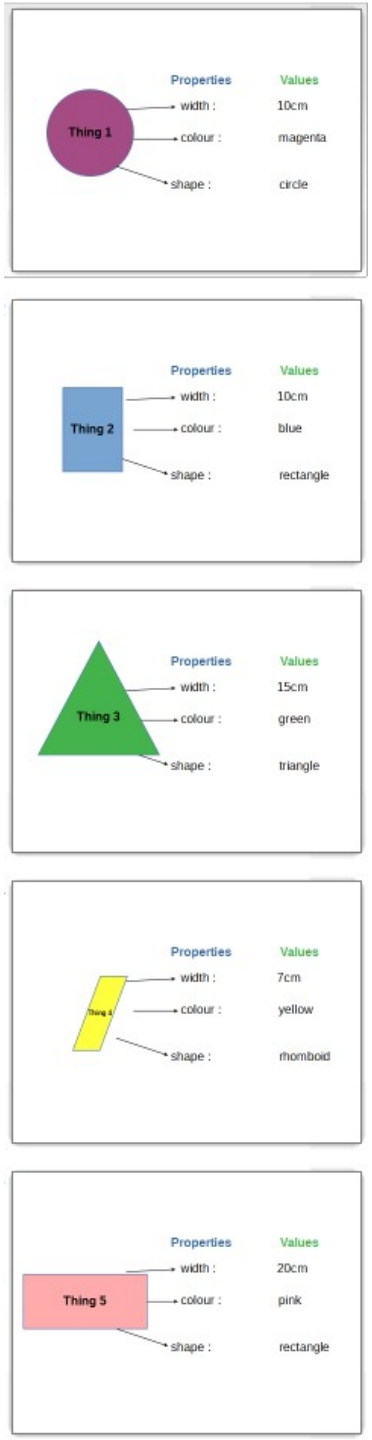
Properties	Values
width :	15cm
colour :	green
shape :	triangle



Properties	Values
width :	7cm
colour :	yellow
shape :	rhomboid


Properties
Values
Thing 5
width :
20cm
colour :
pink
shape :
rectangle

WE HAVE
5 things
(instances,
observations,
examples)

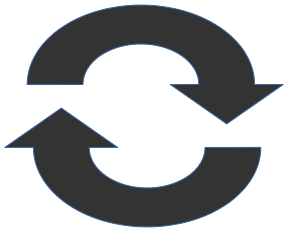


WE HAVE
4 properties (variables,
attributes, features)

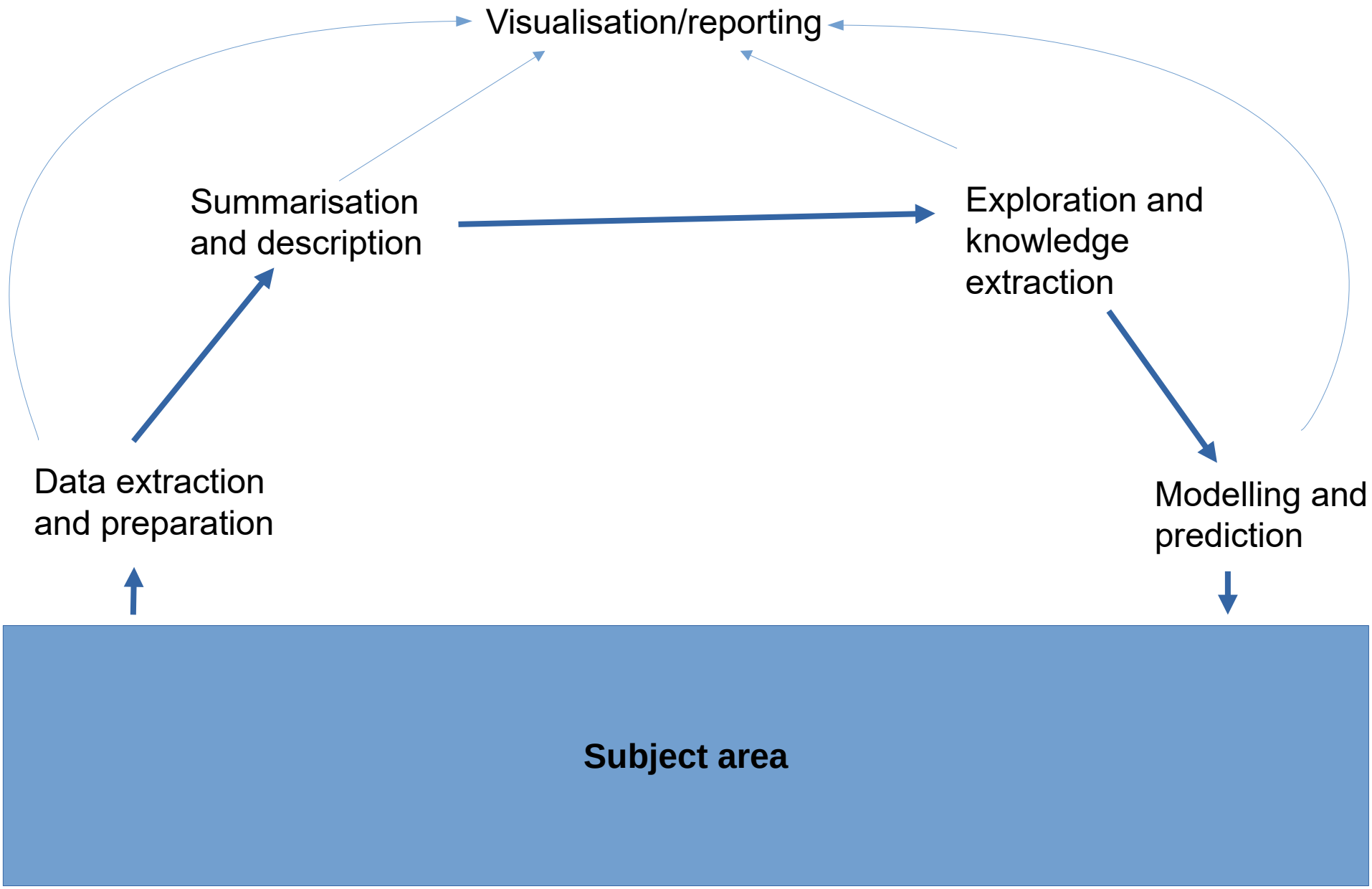
ID	width (cm)	colour	shape
1	10	magenta	circle
2	10	blue	rectangle
3	15	green	triangle
4	7	yellow	rhomboid
5	20	pink	rectangle

This table is what we analyse

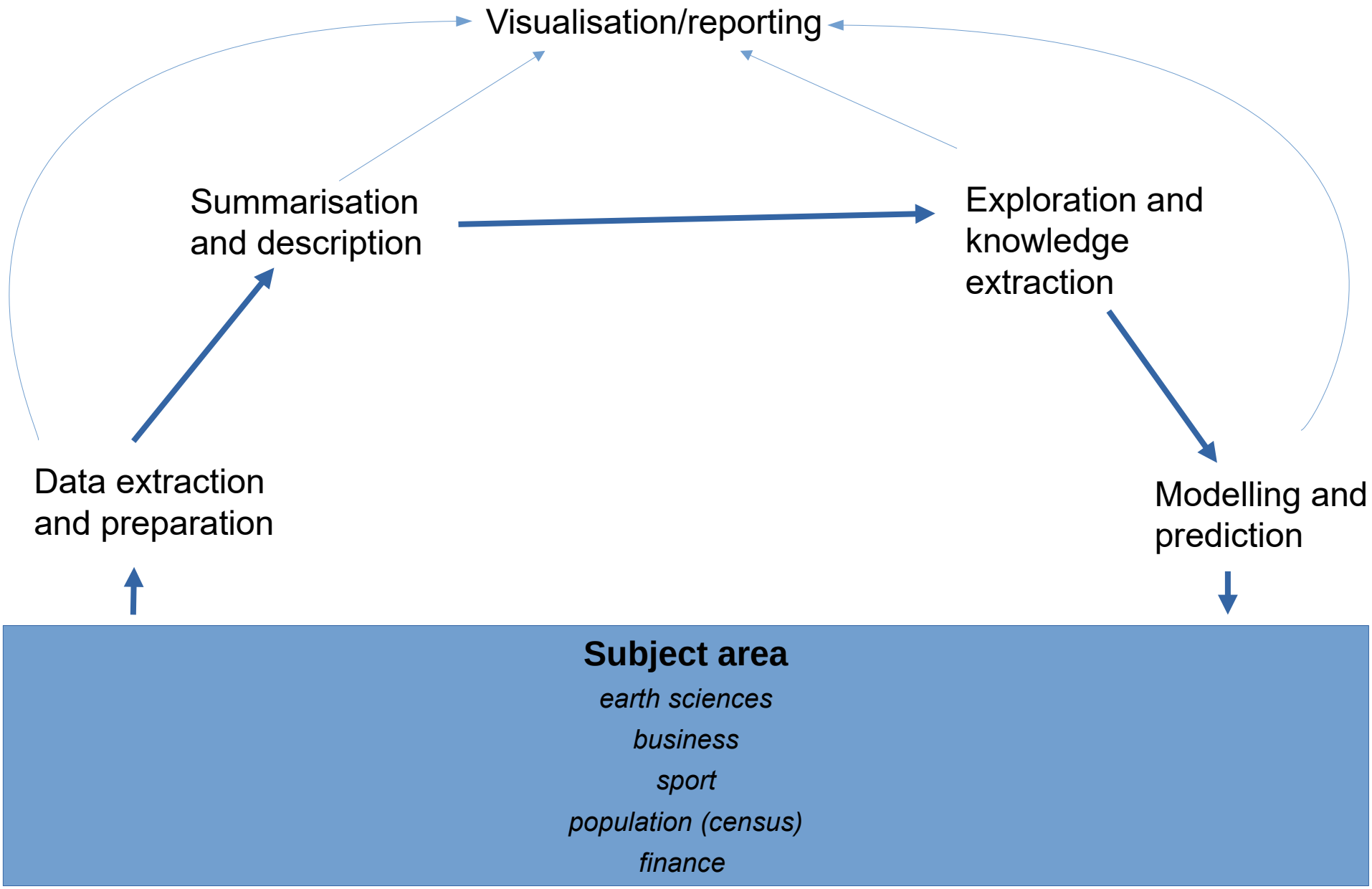
The data cycle



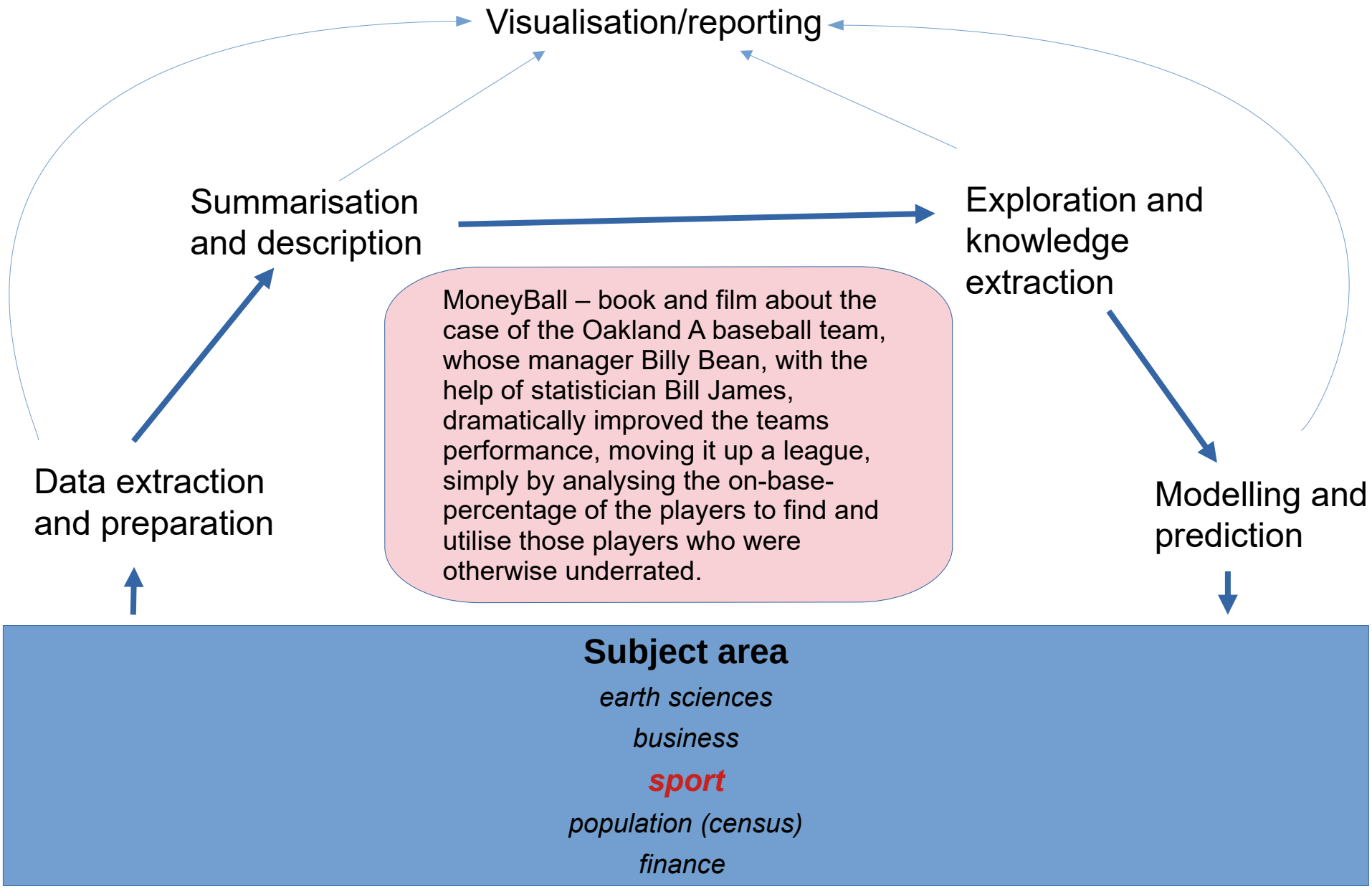
The data cycle



The data cycle

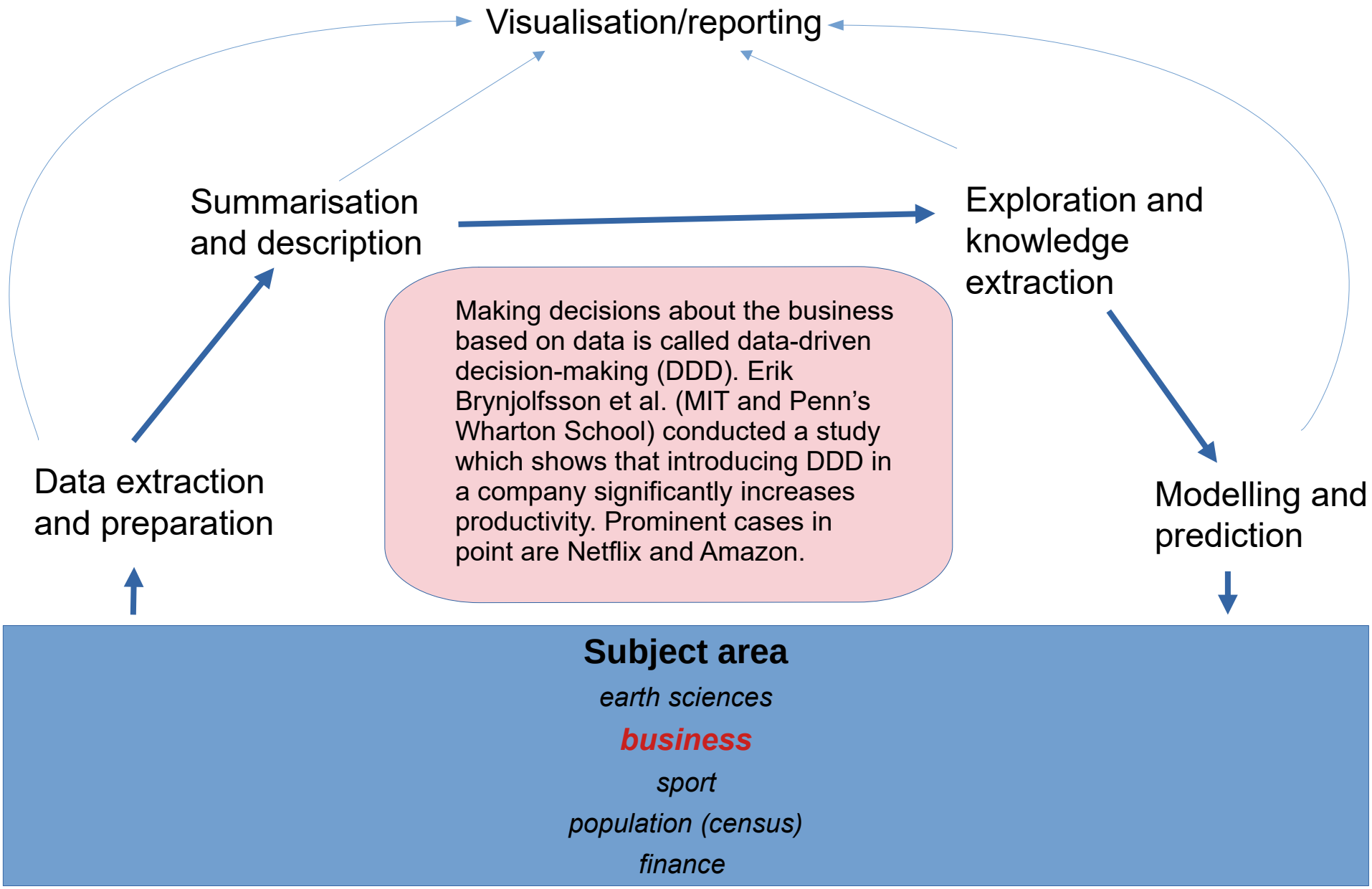


The data cycle



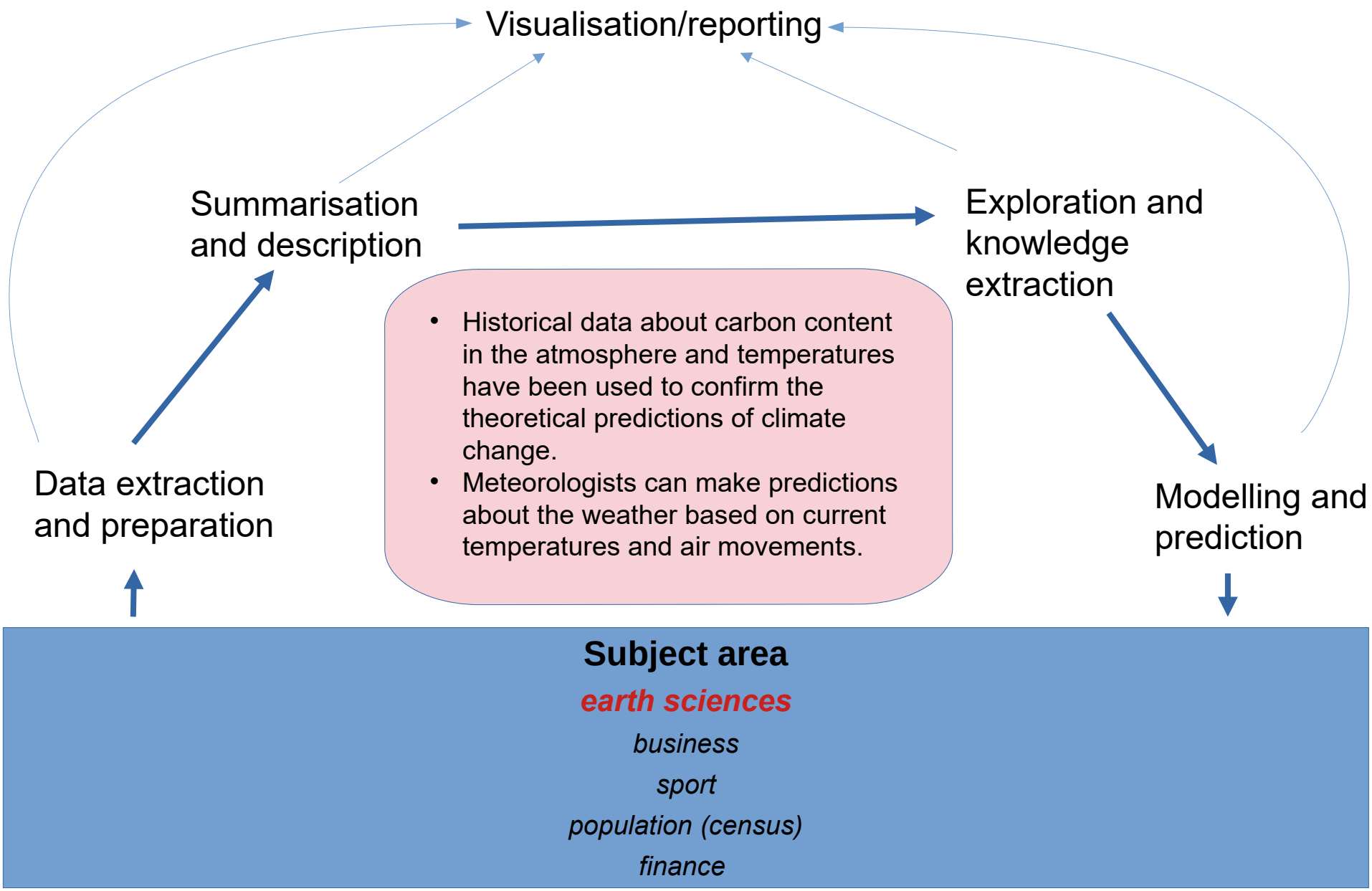
MoneyBall – book and film about the case of the Oakland A baseball team, whose manager Billy Bean, with the help of statistician Bill James, dramatically improved the teams performance, moving it up a league, simply by analysing the on-base-percentage of the players to find and utilise those players who were otherwise underrated.

The data cycle



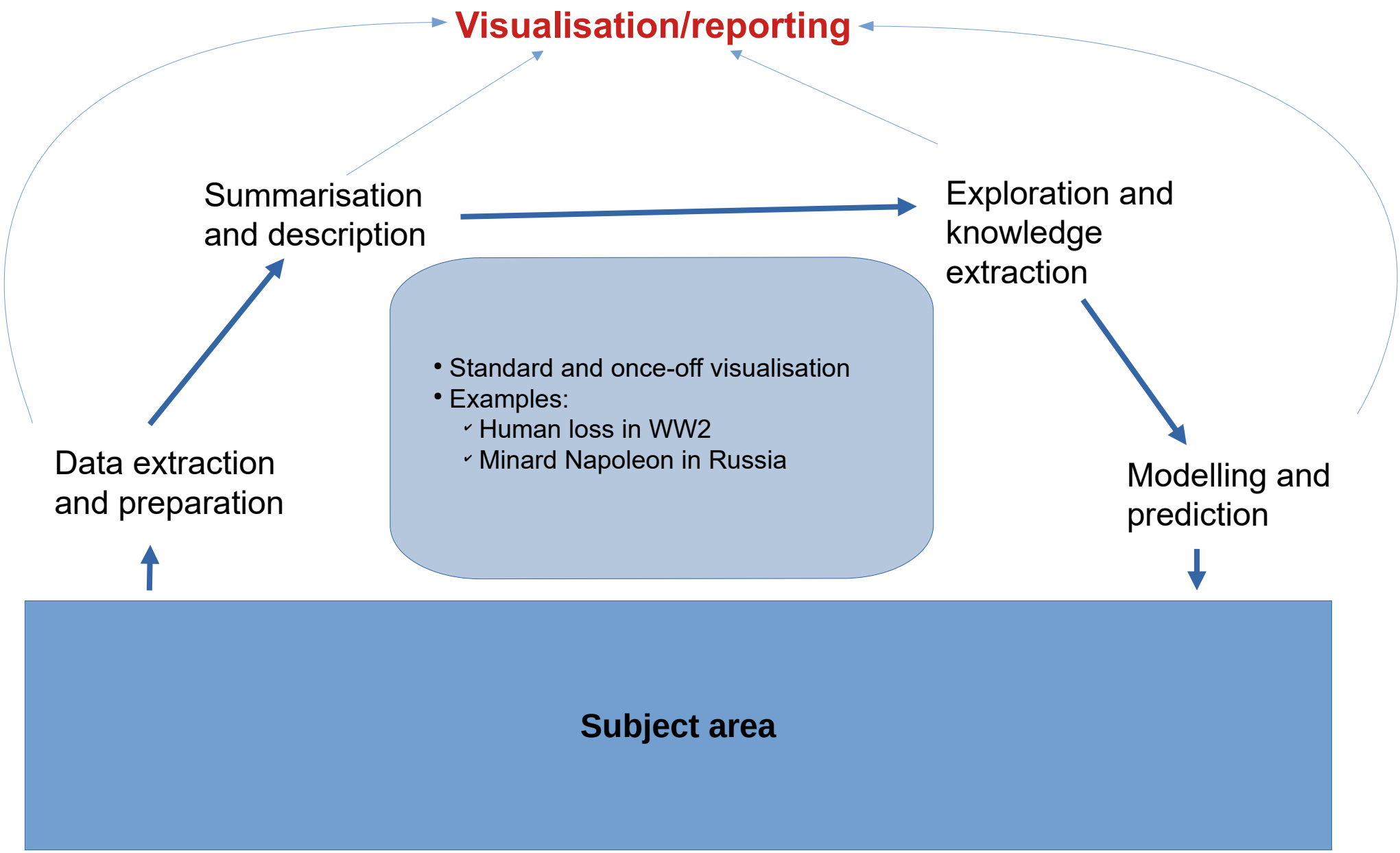
Making decisions about the business based on data is called data-driven decision-making (DDD). Erik Brynjolfsson et al. (MIT and Penn's Wharton School) conducted a study which shows that introducing DDD in a company significantly increases productivity. Prominent cases in point are Netflix and Amazon.

The data cycle

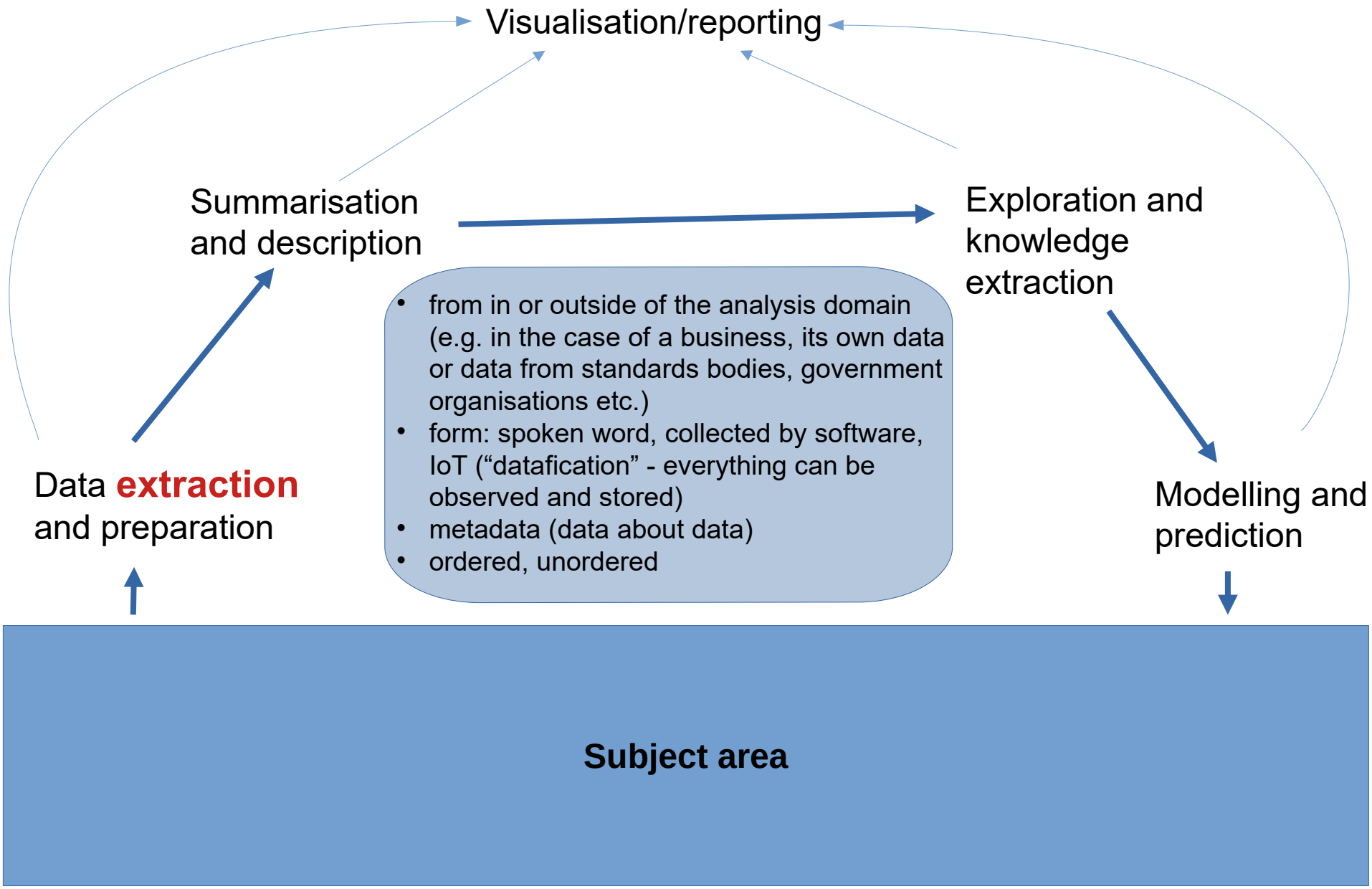


- Historical data about carbon content in the atmosphere and temperatures have been used to confirm the theoretical predictions of climate change.
- Meteorologists can make predictions about the weather based on current temperatures and air movements.

The data cycle

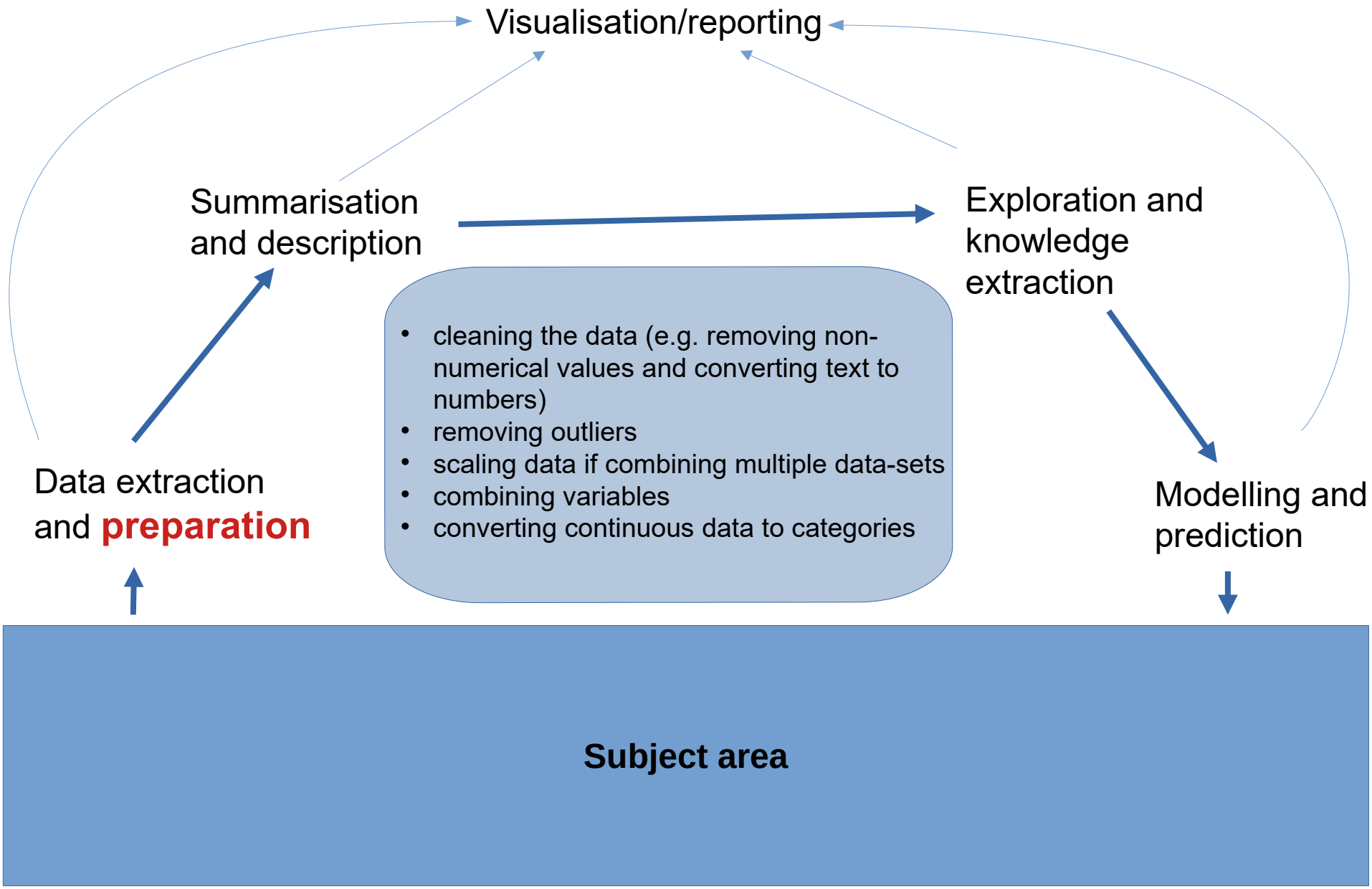


The data cycle



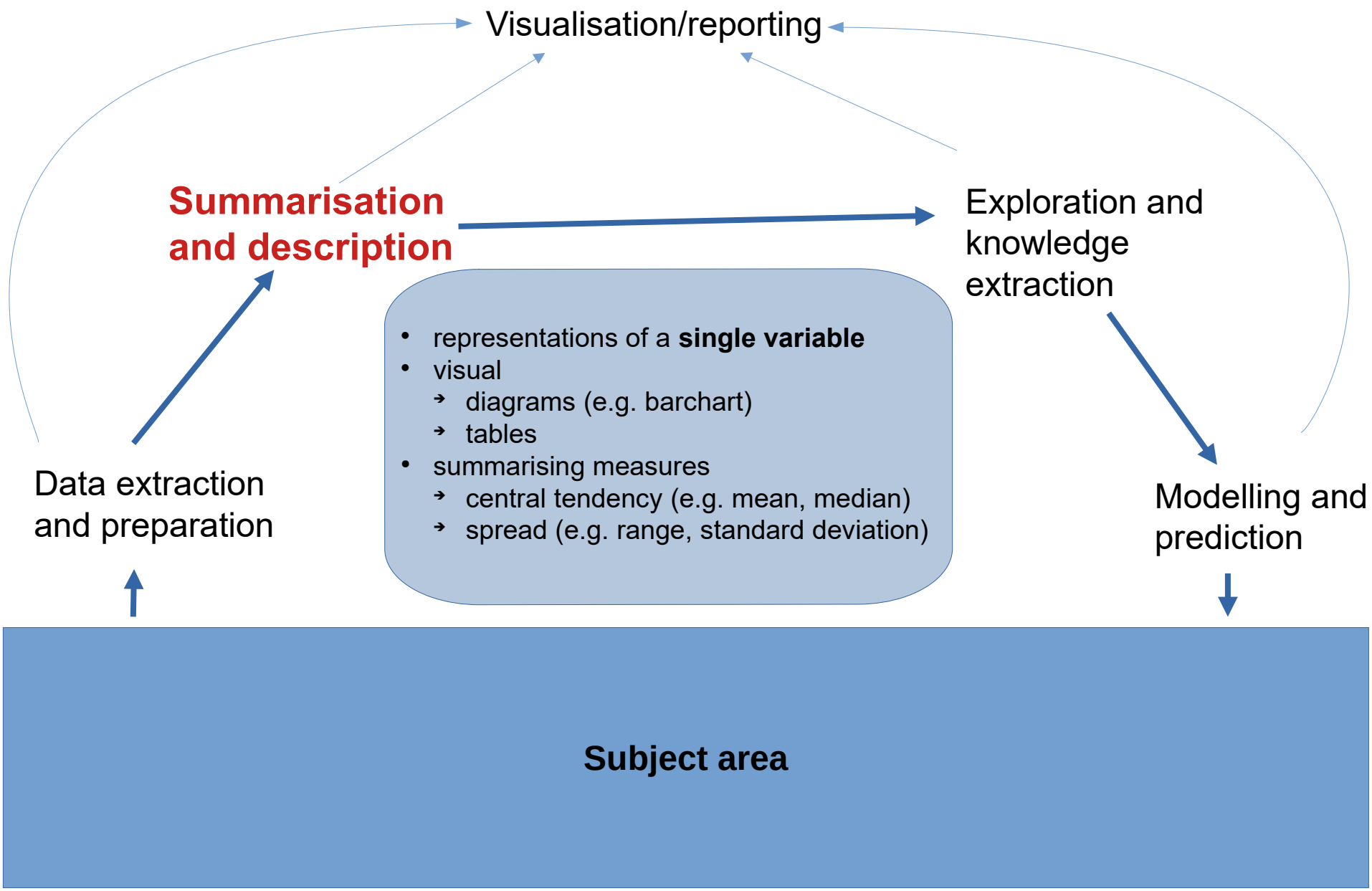
- from in or outside of the analysis domain (e.g. in the case of a business, its own data or data from standards bodies, government organisations etc.)
- form: spoken word, collected by software, IoT (“datafication” - everything can be observed and stored)
- metadata (data about data)
- ordered, unordered

The data cycle



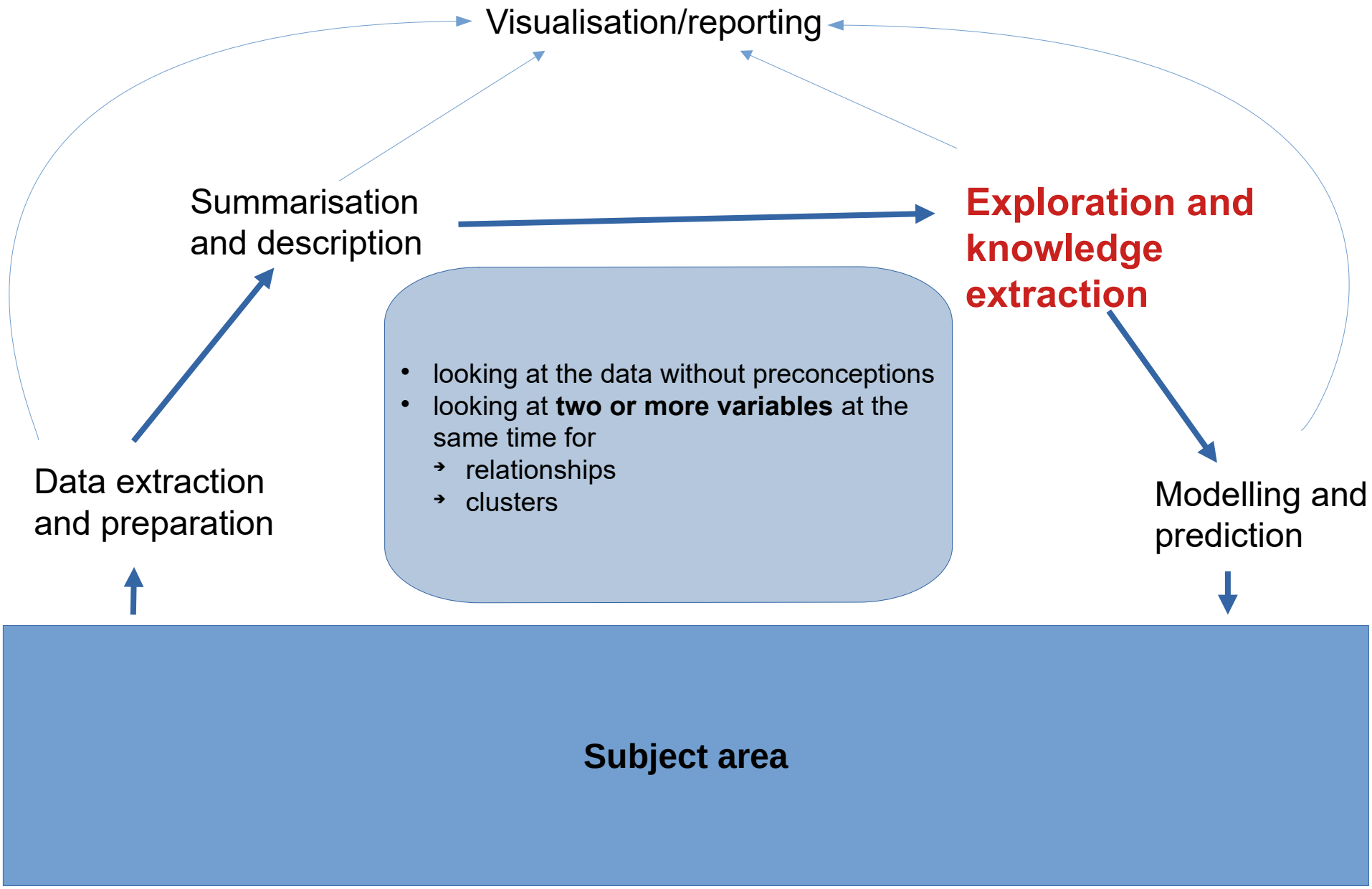
- cleaning the data (e.g. removing non-numerical values and converting text to numbers)
- removing outliers
- scaling data if combining multiple data-sets
- combining variables
- converting continuous data to categories

The data cycle

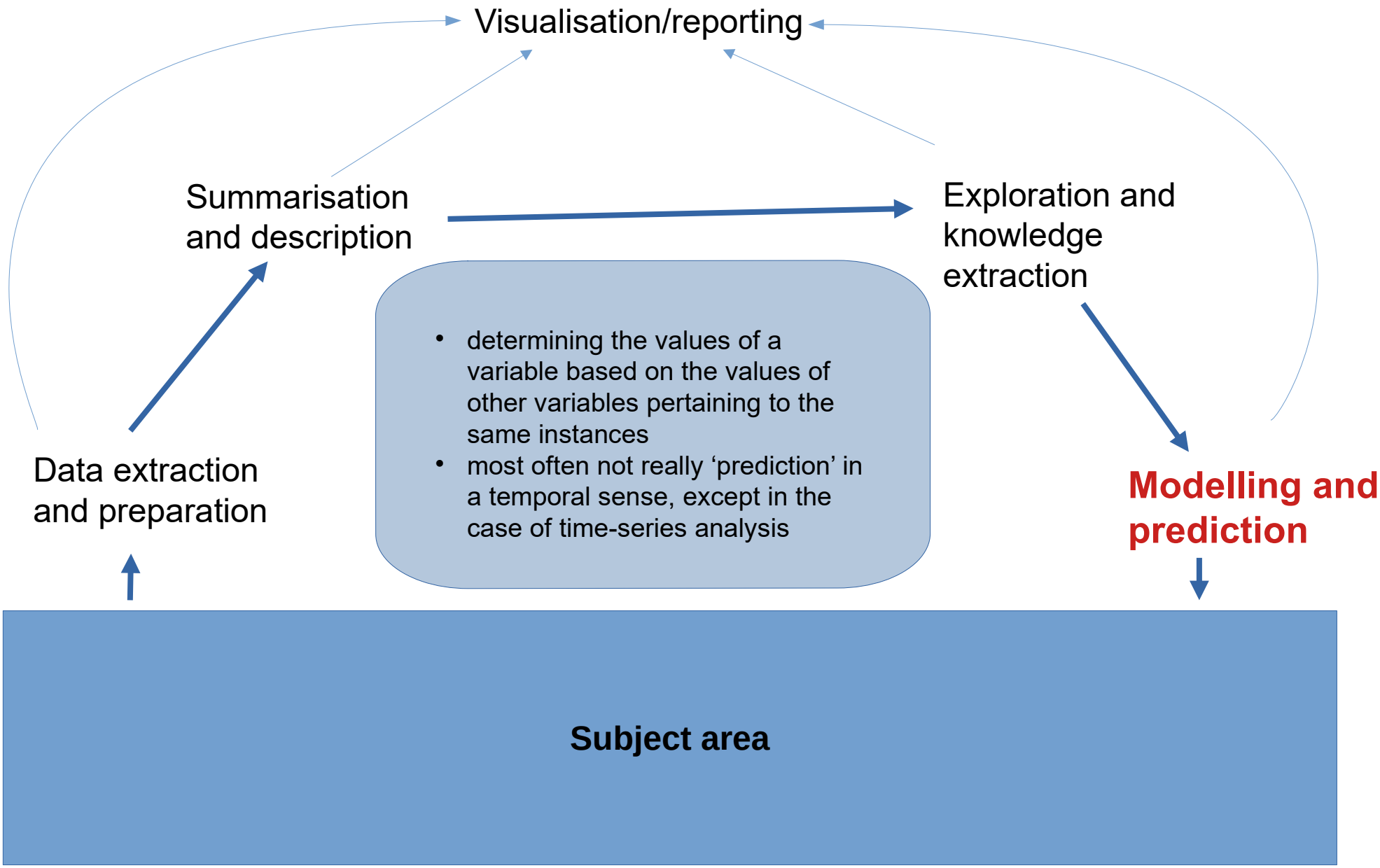


- representations of a **single variable**
- visual
  - diagrams (e.g. barchart)
  - tables
- summarising measures
  - central tendency (e.g. mean, median)
  - spread (e.g. range, standard deviation)

The data cycle

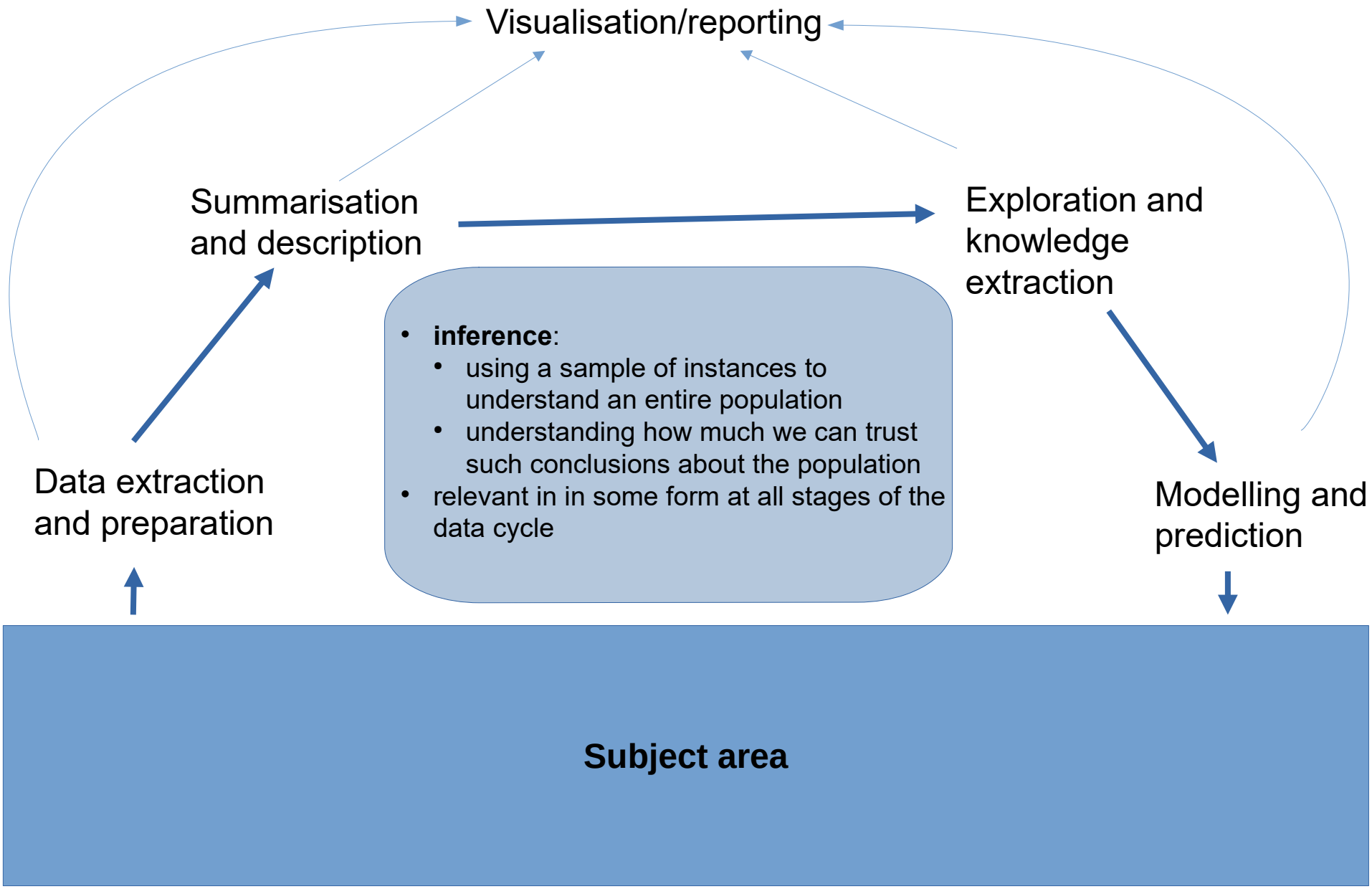


The data cycle



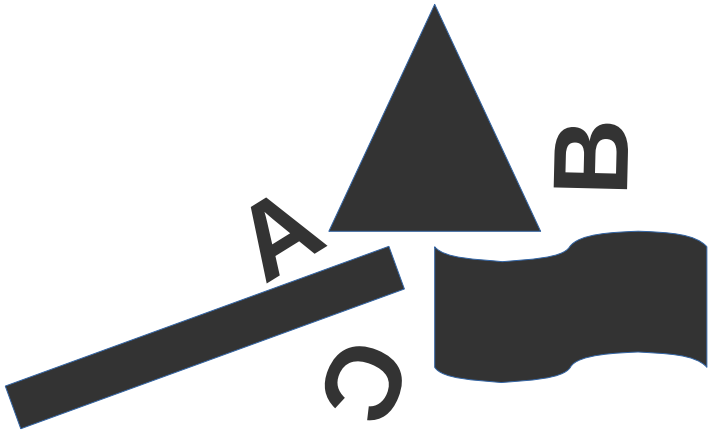
- determining the values of a variable based on the values of other variables pertaining to the same instances
- most often not really ‘prediction’ in a temporal sense, except in the case of time-series analysis

The data cycle



- **inference**:
  - using a sample of instances to understand an entire population
  - understanding how much we can trust such conclusions about the population
- relevant in in some form at all stages of 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.

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

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)

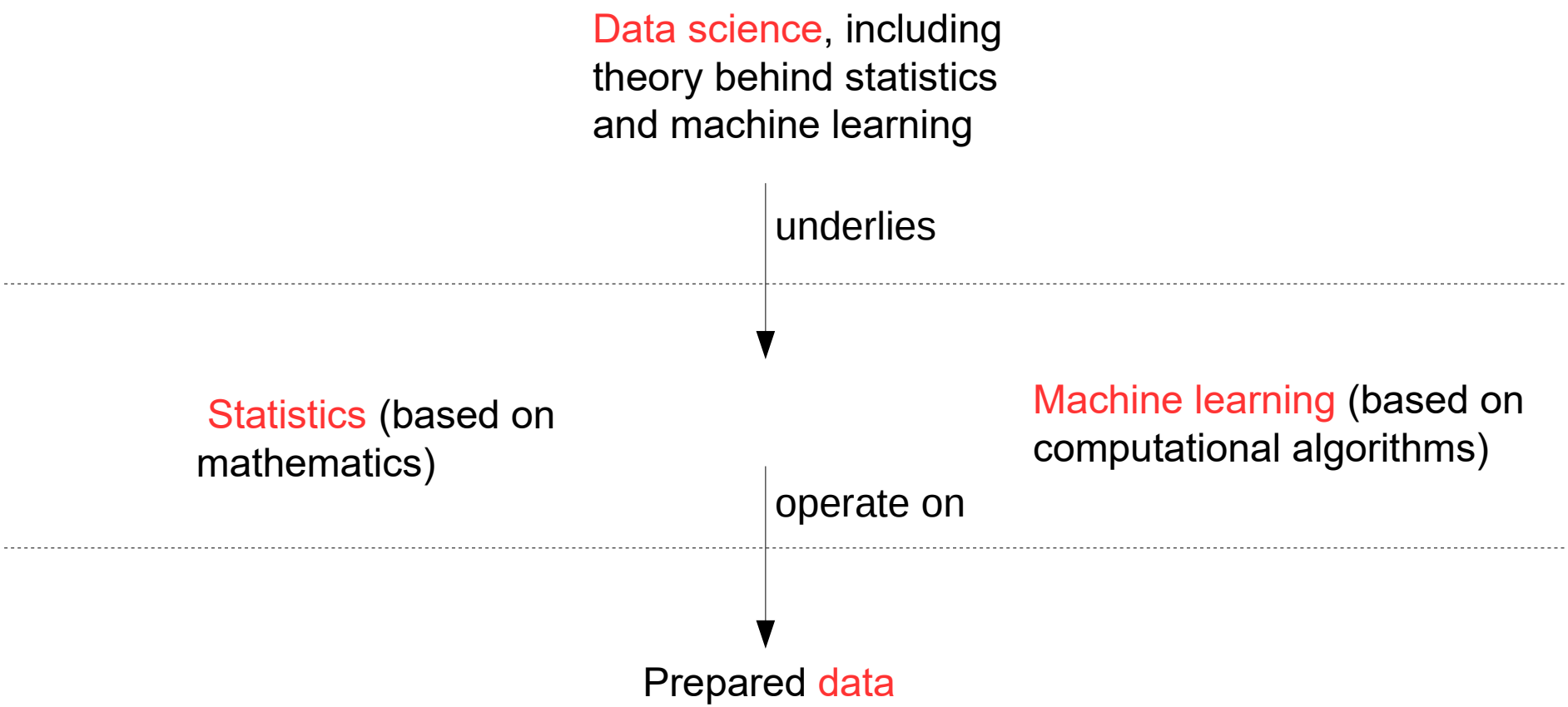
Machine learning (based on computational algorithms)

operate on

Prepared data

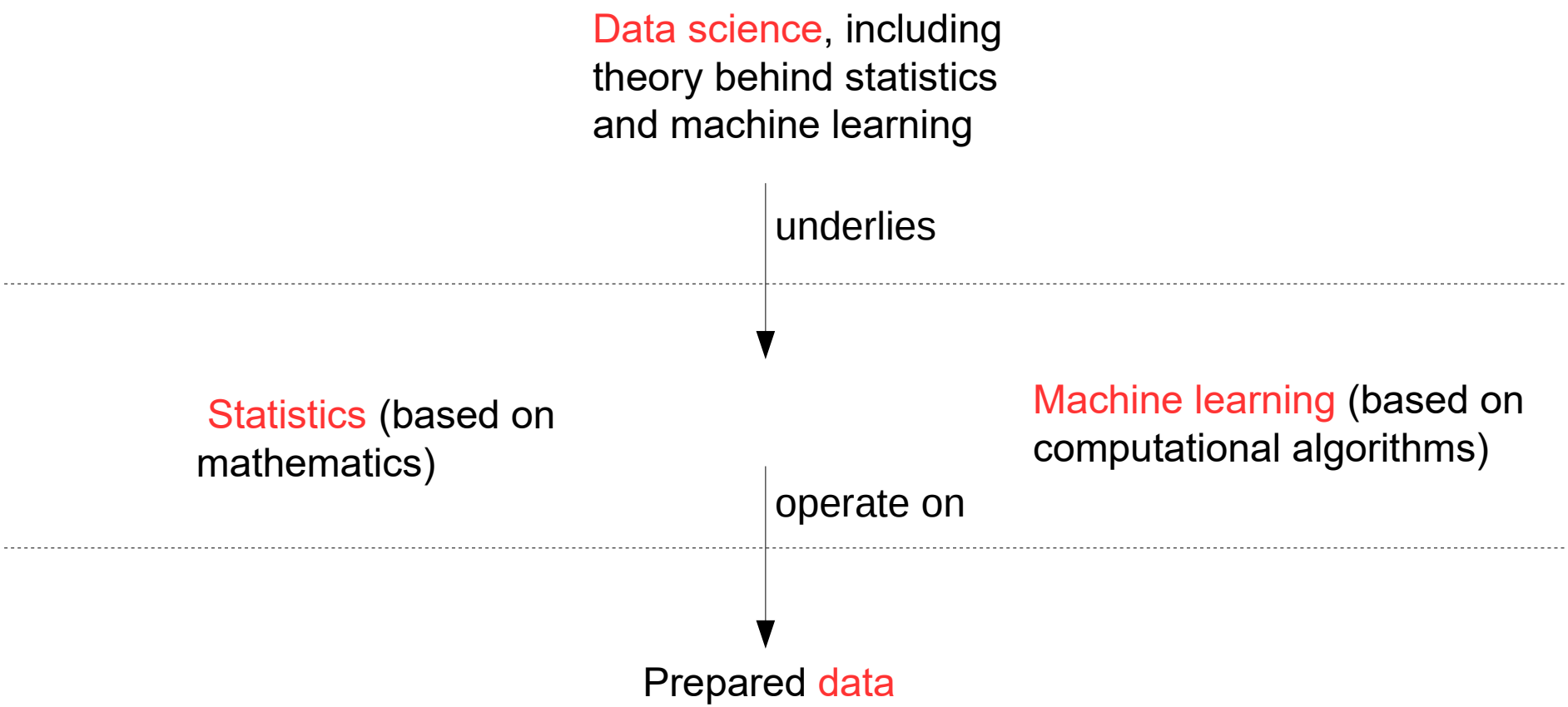
Landscape and terminology

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



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

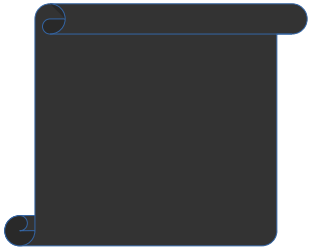
Prepared data

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