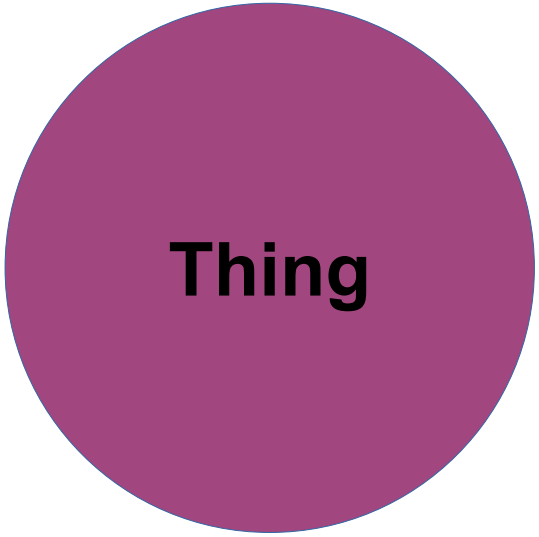
Data Analysis: Introduction

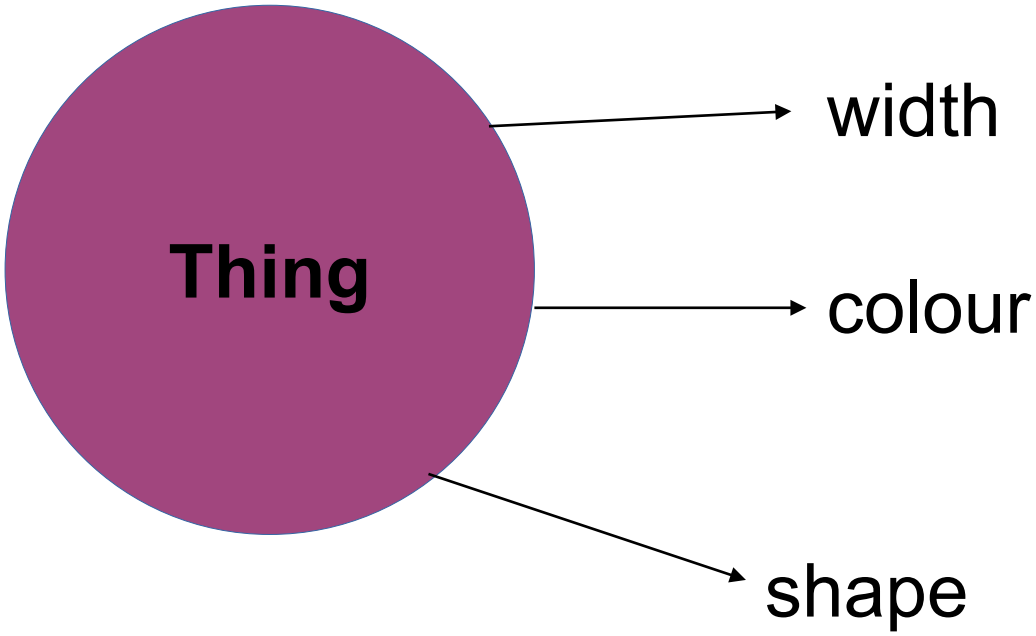
TU Dublin Tallaght Campus,

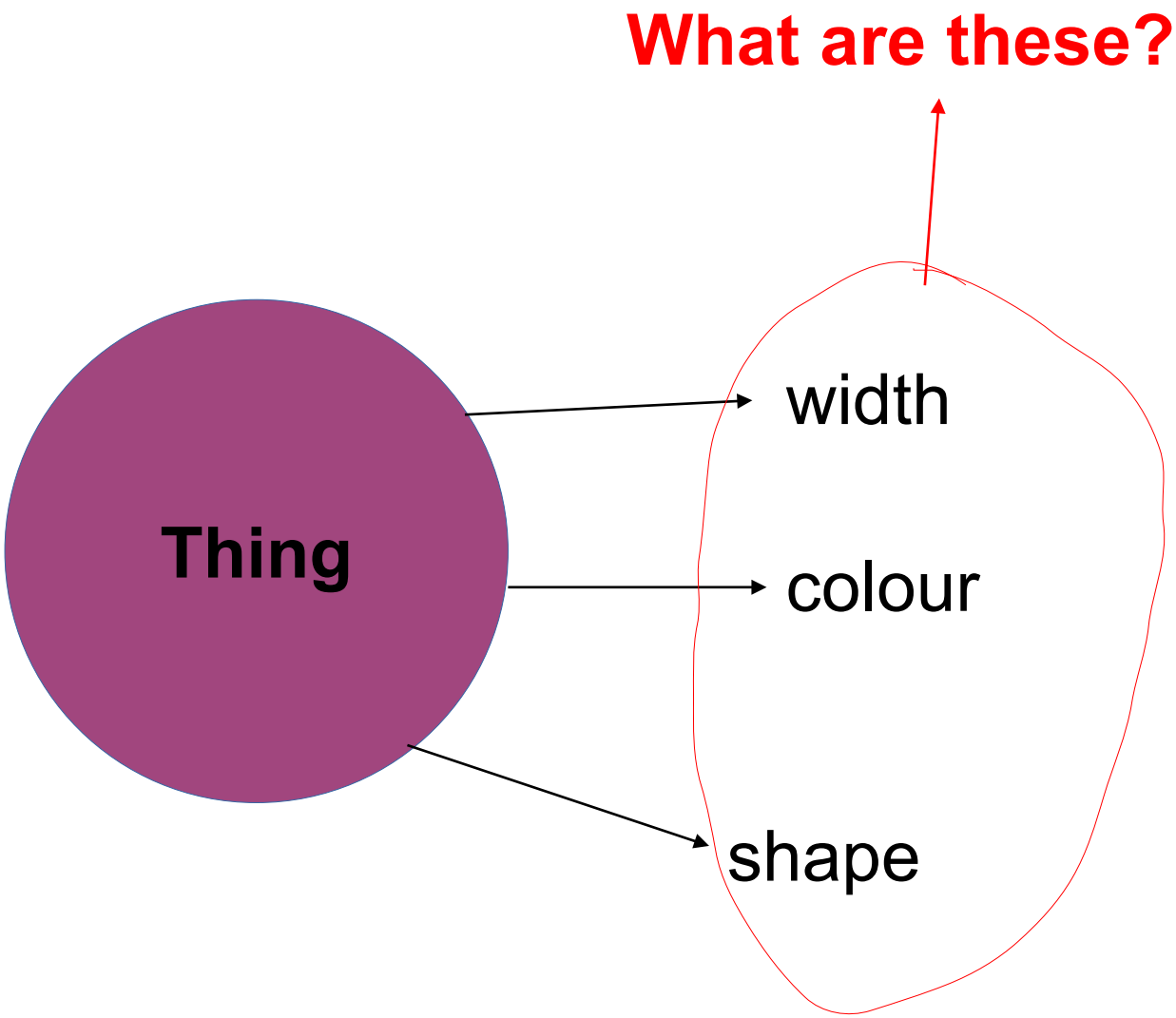
Department of Computing

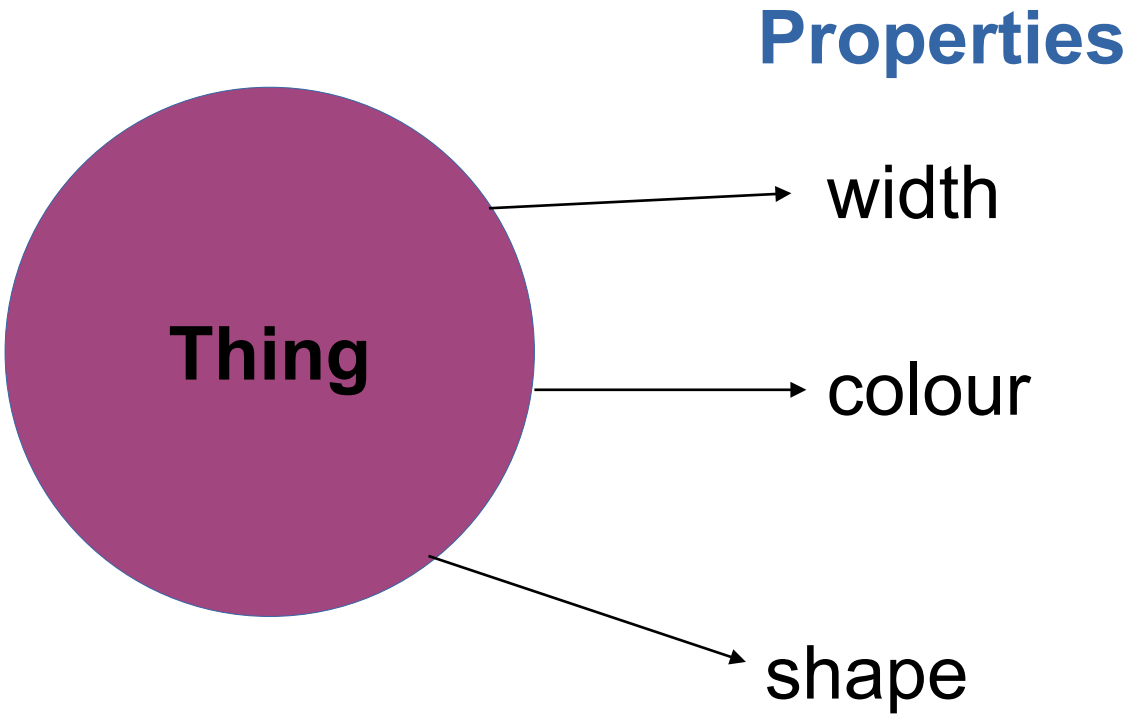
What are we dealing with

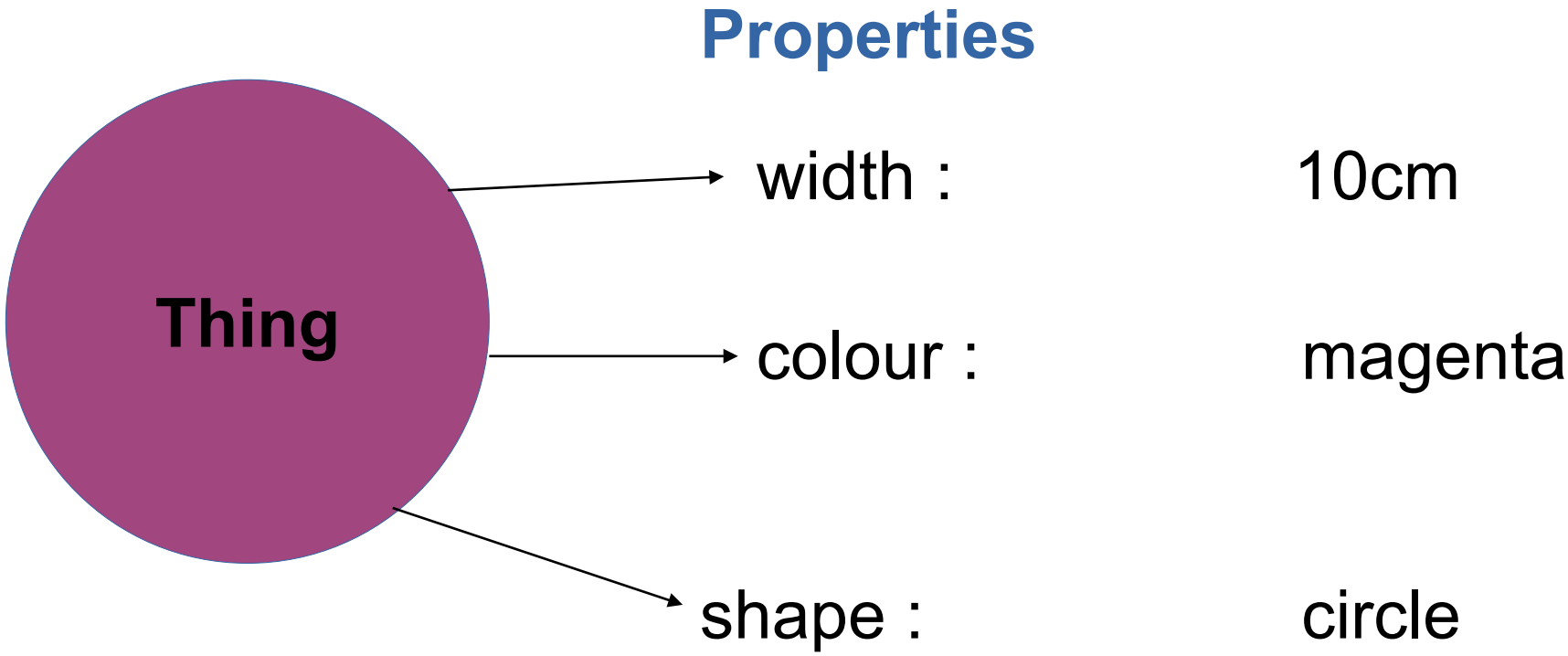
?

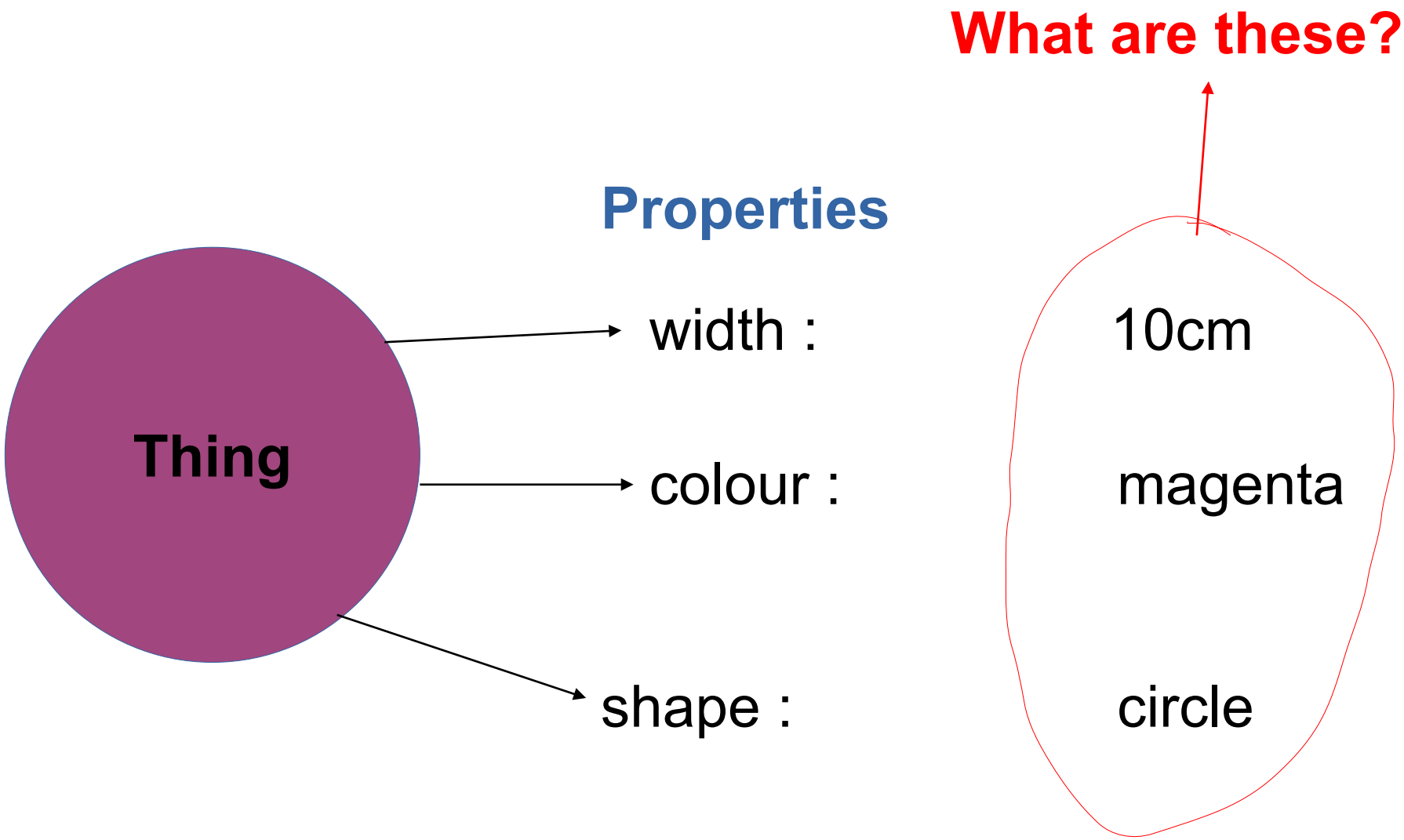

Thing

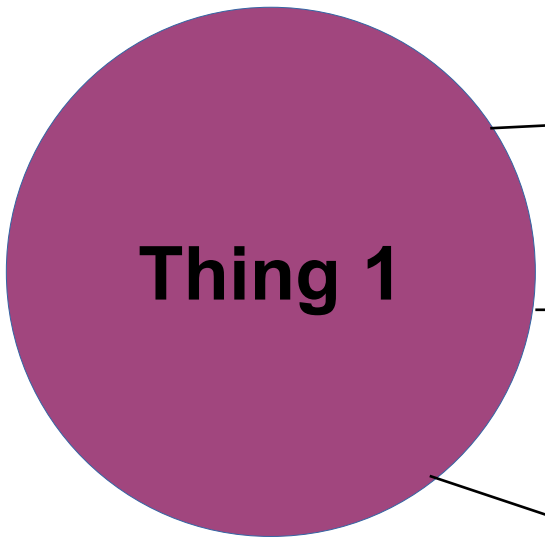

Thing
width
colour
shape


What are these?
Thing
width
colour
shape


Properties
Thing
width
colour
shape

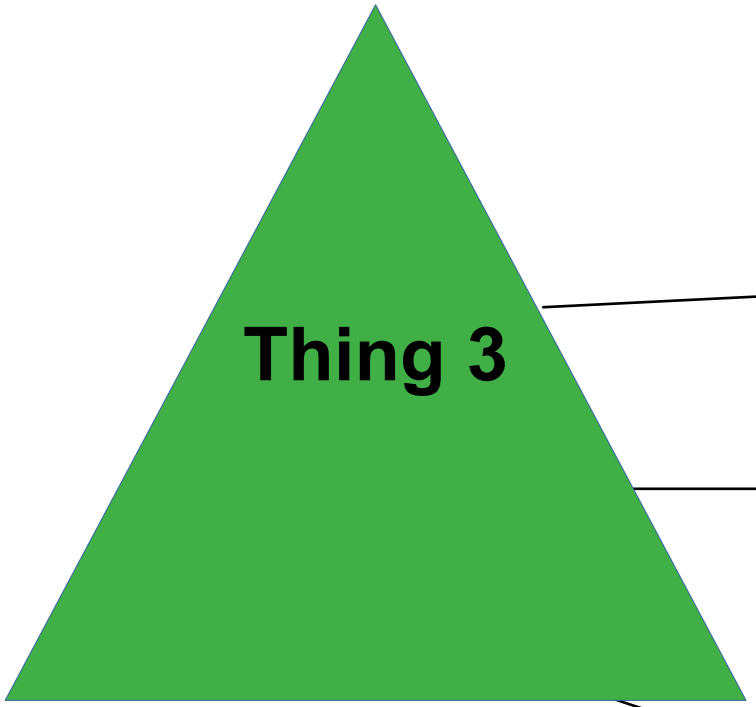

Properties
Thing
width :
10cm
colour :
magenta
shape :
circle


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

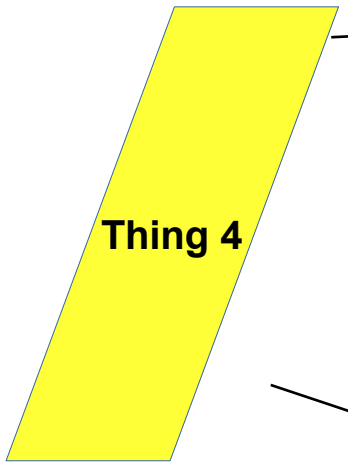


Properties	Values
width :	10cm
colour :	magenta
shape :	circle

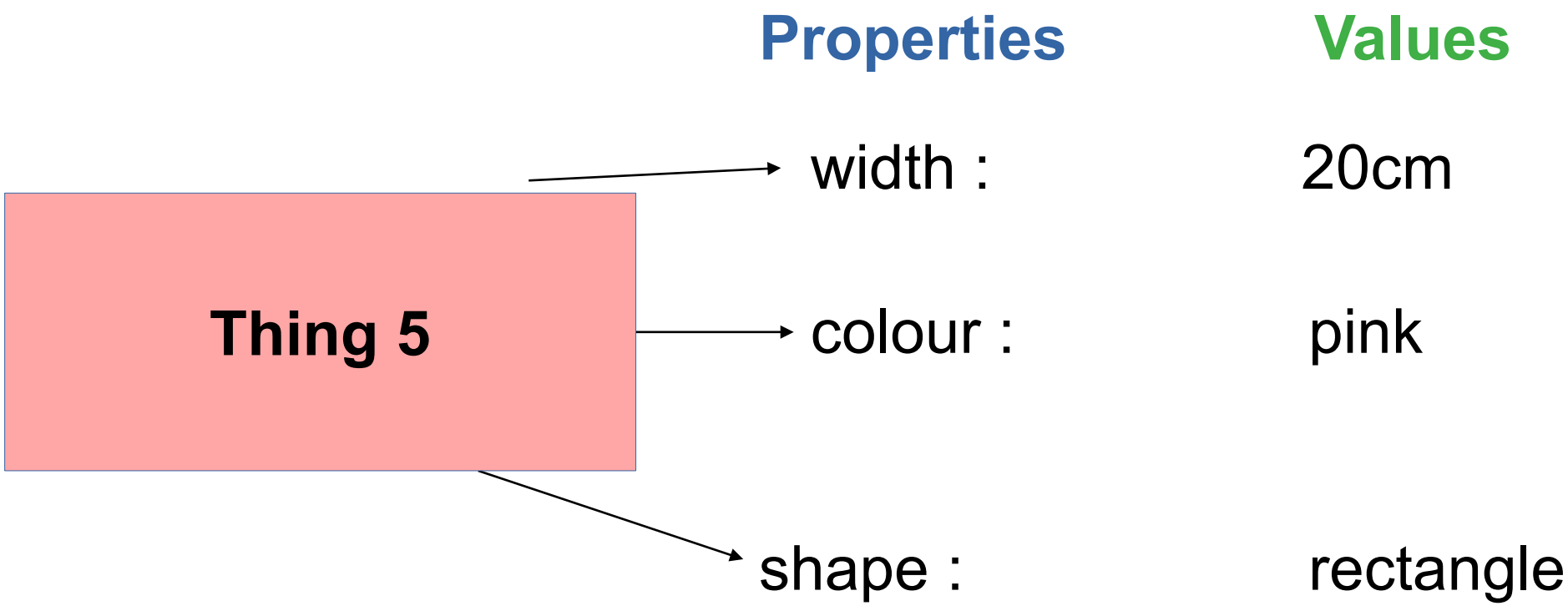
Properties	Values
width :	10cm
colour :	blue
shape :	rectangle



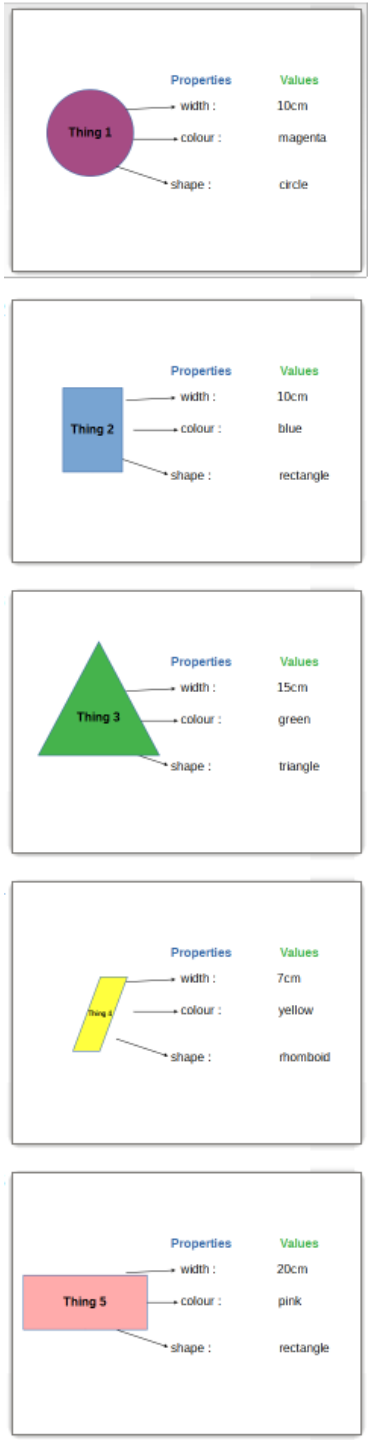
Properties	Values
width :	15cm
colour :	green
shape :	triangle



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)

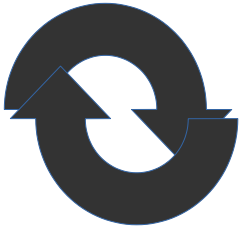


WE HAVE
4 properties (variables,
attributes, features)

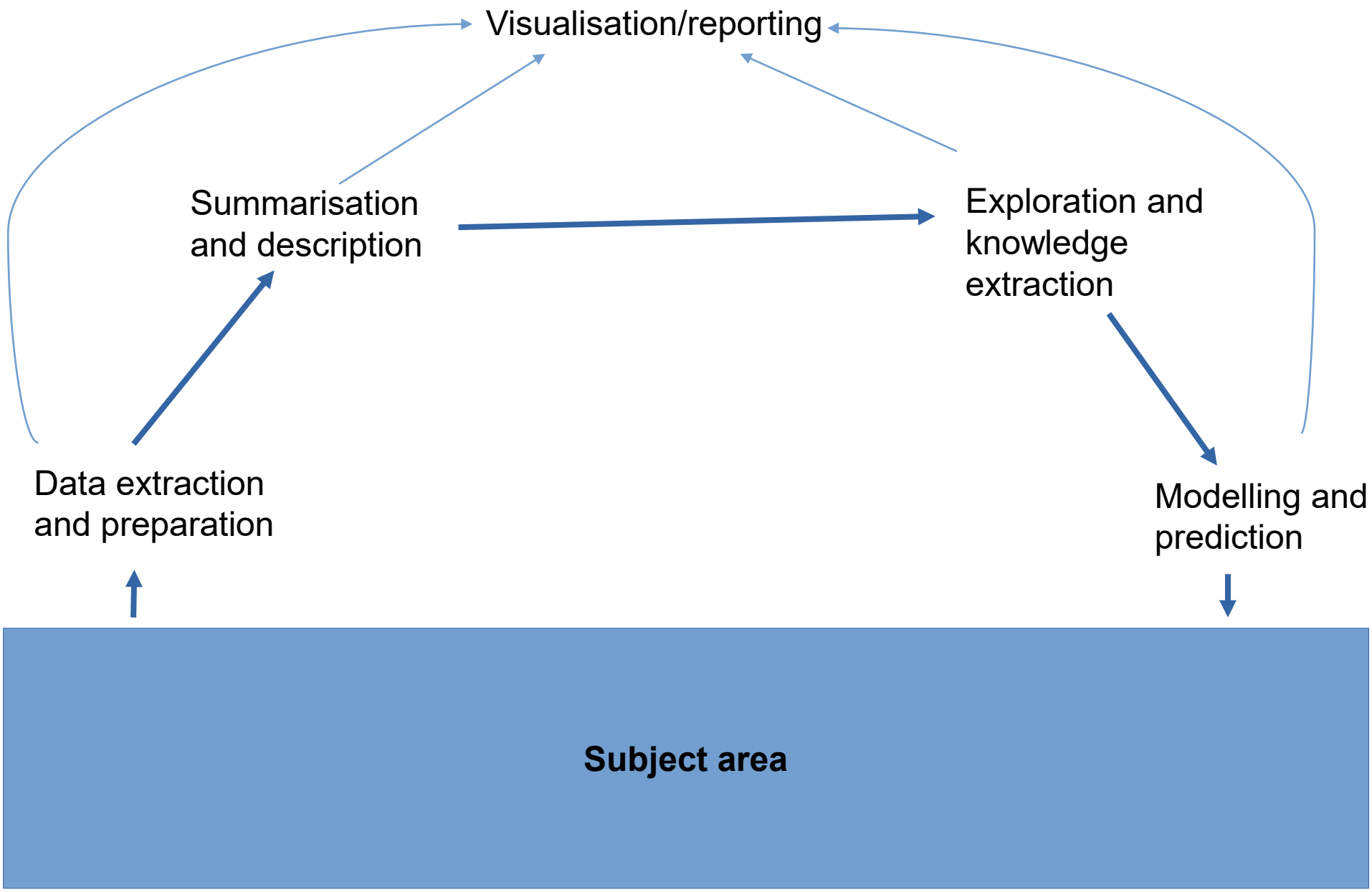
ID	width (cm)	colour	shape
1	10	magenta	circle
2	10	blue	rectangle
3	15	green	triangle
4	7	yellow	rhomboid
5	20	pink	rectangle

This table is what we
analyse

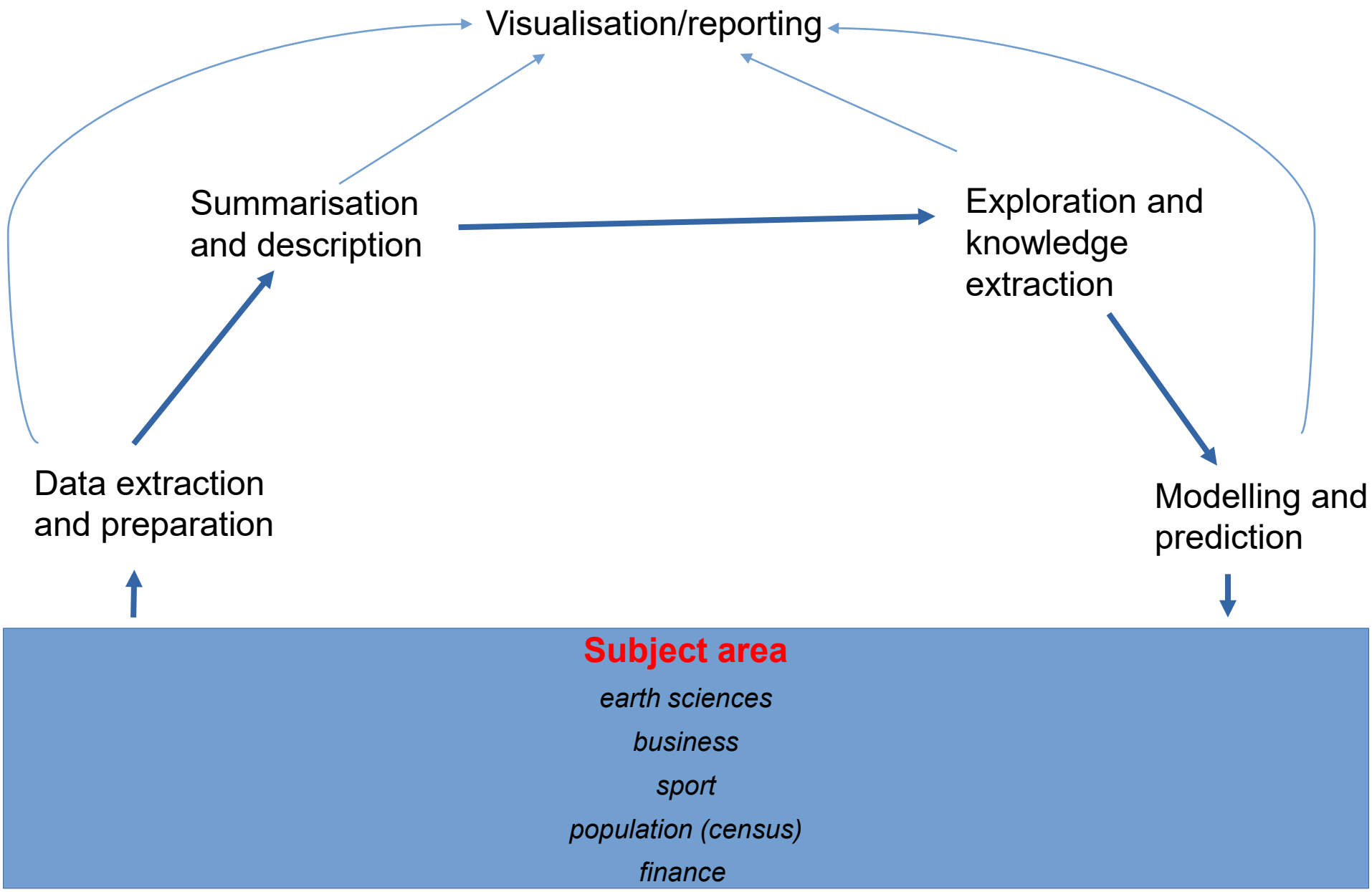
The analysis cycle



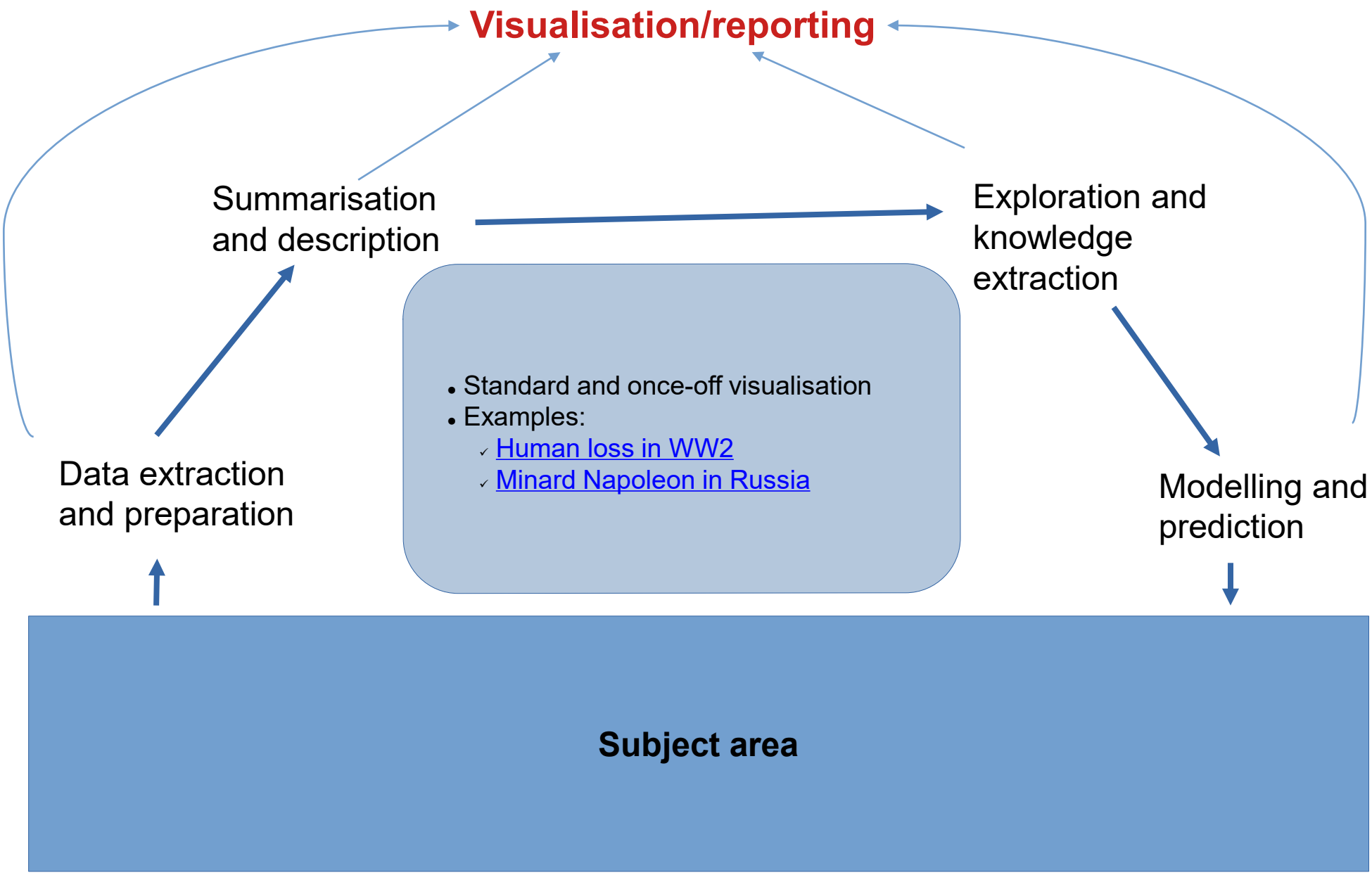
The data analysis cycle



The data analysis cycle

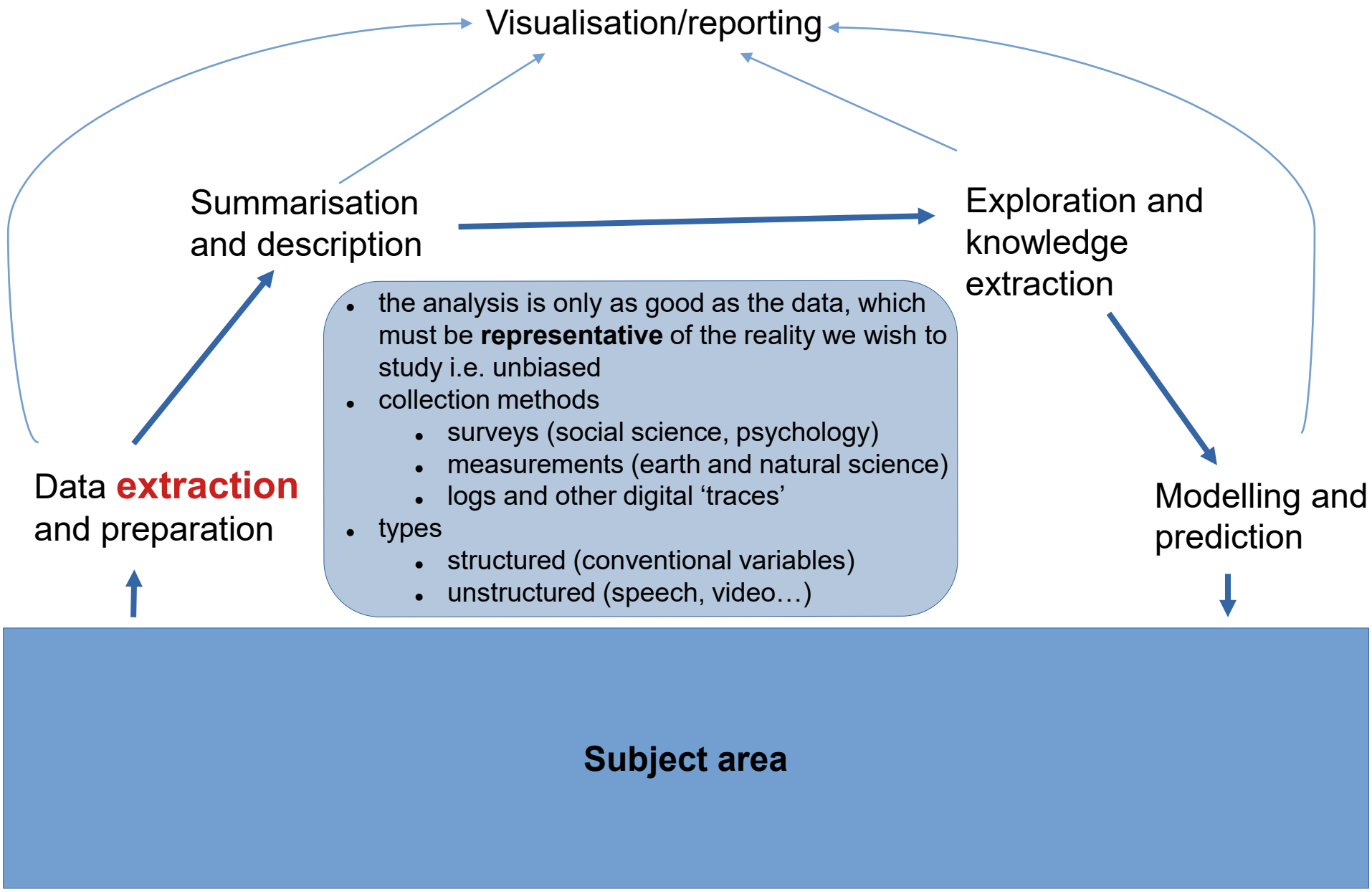


The data analysis cycle



- Standard and once-off visualisation
- Examples:
  - ✓ Human loss in WW2
  - ✓ Minard Napoleon in Russia

The data analysis cycle



- the analysis is only as good as the data, which must be **representative** of the reality we wish to study i.e. unbiased
- collection methods
  - surveys (social science, psychology)
  - measurements (earth and natural science)
  - logs and other digital 'traces'
- types
  - structured (conventional variables)
  - unstructured (speech, video…)

The data analysis cycle

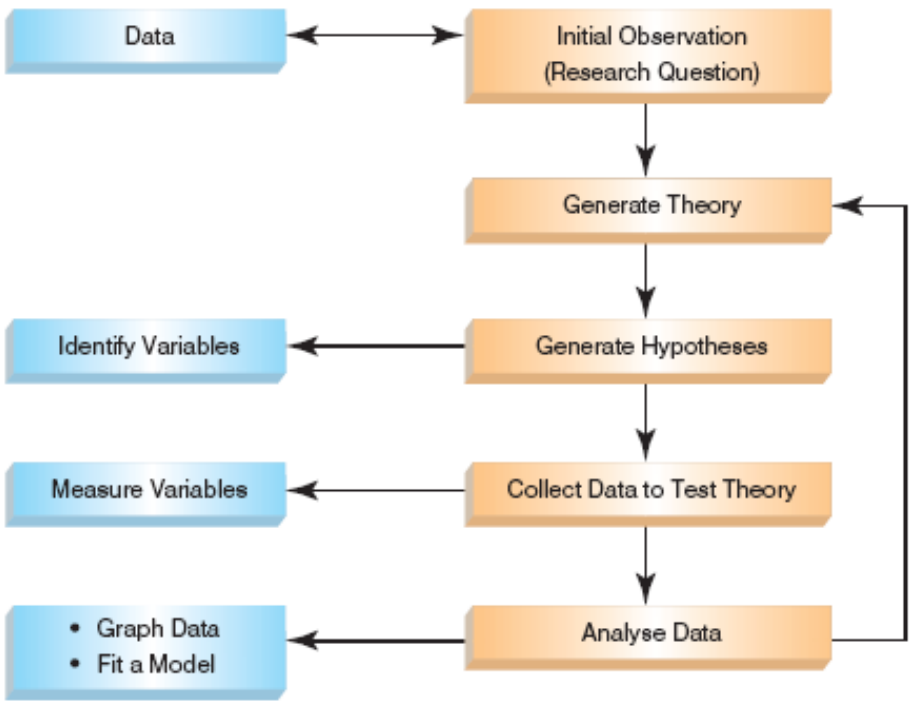
Visualisation/reporting

Summarisation and description

Exploration and knowledge extraction

- the analysis is only as good as the data, which must be **representative** of the reality we wish to study i.e. unbiased

Data **extraction** and preparation



Source: [DSR]

Often, data analysis is performed simply because **data are available**. This is the case with enterprise data warehouses, which are explored for knowledge hidden in data already available as a by-product of a company's commercial activity.

The older approach is to start with a theory that needs to be tested and collect data with that aim. The picture shows this process, which needs to be paralleled in **sampling assessment** for analysis, even if the data is pre-collected.

The data analysis cycle

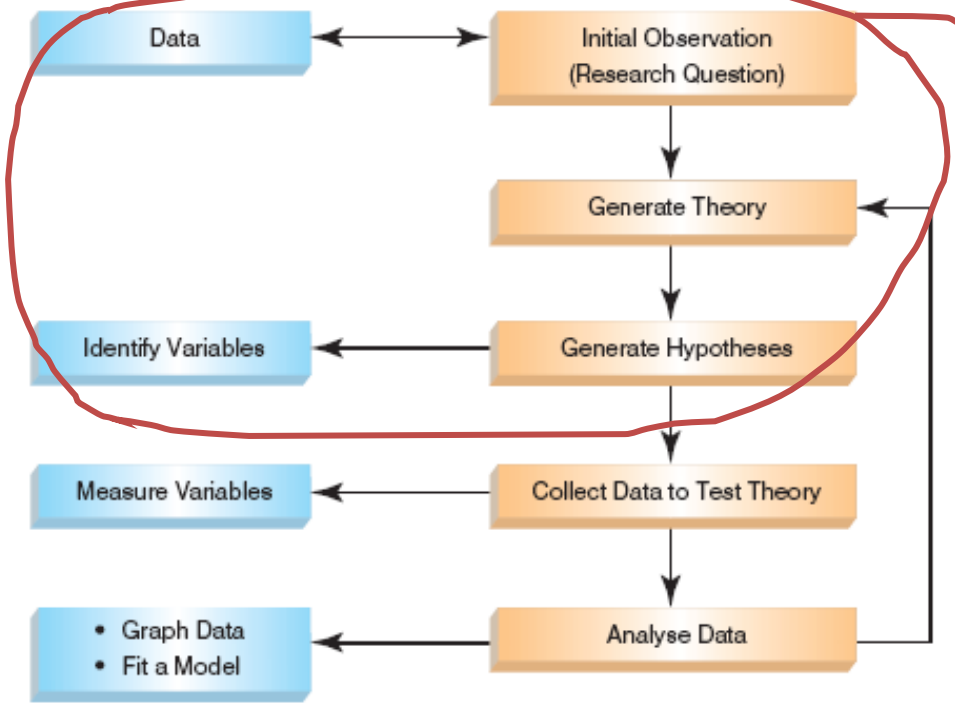
Visualisation/reporting

Summarisation and description

Exploration and knowledge extraction

Data **extraction** and preparation

- the analysis is only as good as the data, which must be **representative** of the reality we wish to study i.e. unbiased



Source: [DSR]

The design of the data collection process (i.e. the **experiment)** must be grounded in **sound conceptual understanding** of the problem domain.

A **theory** is expressed in terms of domain concepts, e.g. 'TU Dublin attracts local students'. A **hypothesis** is expressed as a concrete prediction e.g. 'the majority of students grew up within 10km of Grangegorman'. **Falsification** occurs when a theory or hypothesis is disproved.

The data analysis cycle

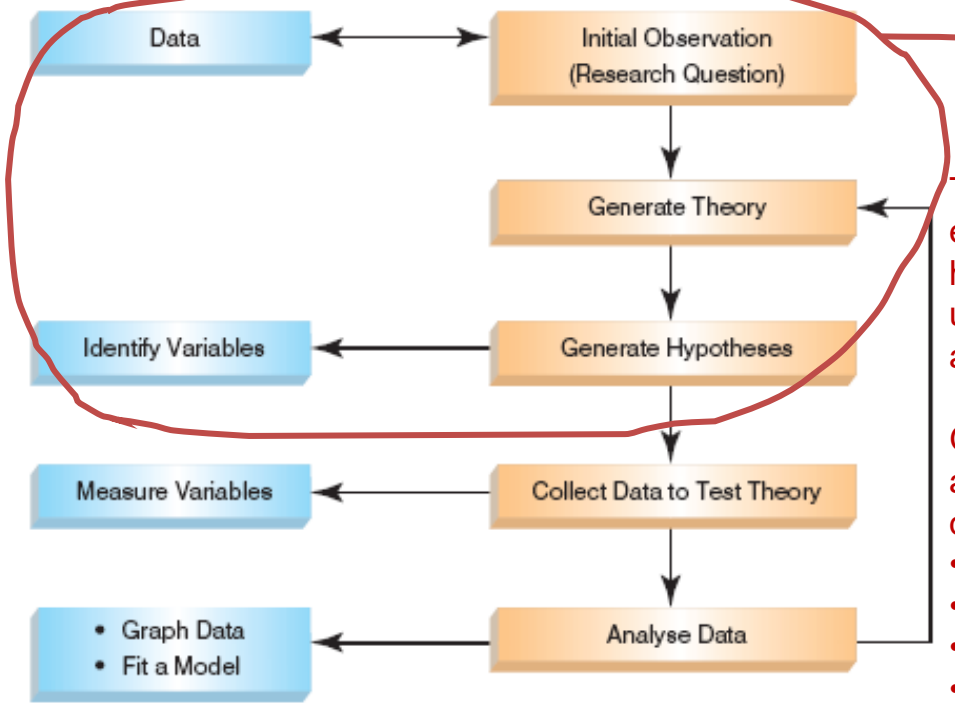
Visualisation/reporting

Summarisation and description

Exploration and knowledge extraction

Data **extraction** and preparation

- the analysis is only as good as the data, which must be **representative** of the reality we wish to study i.e. unbiased



Source: [DSR]

The variables involved in the experiment are identified from the hypothesis (in the example already used these would be 'college name' and 'distance to home').

Good understanding of the domain and consequent good experimental design ensure:

- **reliability** (types)
- **validity** (types)
- **accuracy** of measurement
- absence of **confounding** variables

The data analysis cycle

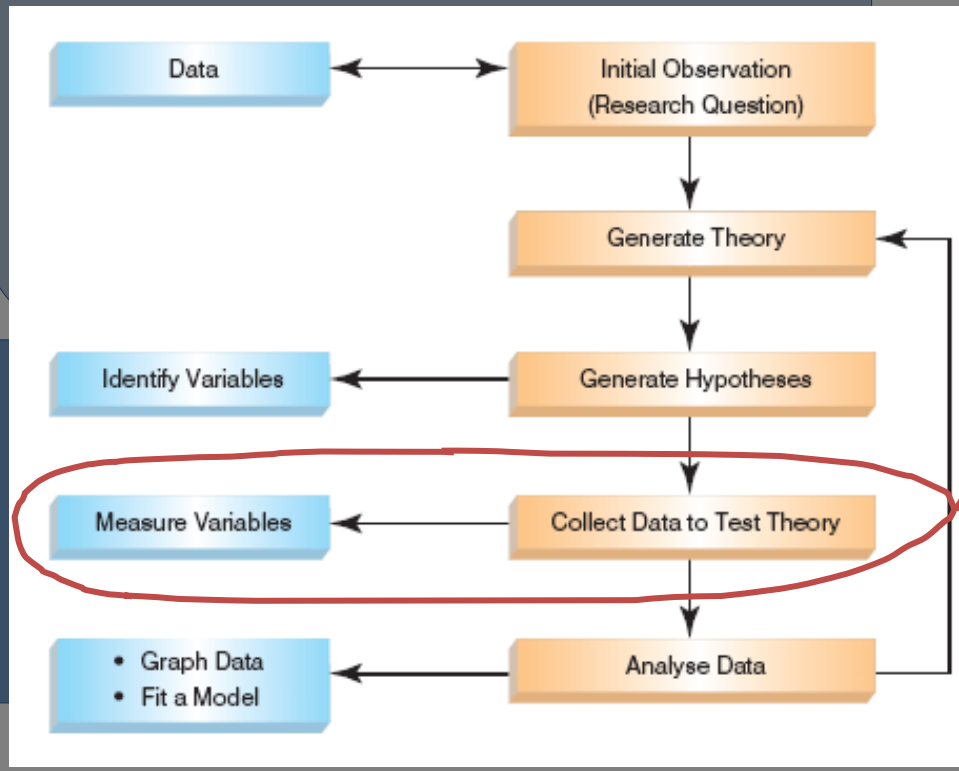
Visualisation/reporting

Summarisation and description

Exploration and knowledge extraction

- the analysis is only as good as the data, which must be **representative** of the reality we wish to study i.e. unbiased

Data **extraction** and preparation



Source: [DSR]

The instances for which data is collected represent a **sample** of the **population** of all instances. Two important (but not only) factors in how unbiased a sample will be are its **size** and **sampling method** (some type of random sampling is desirable).

Small samples may lead to wrong conclusions. Large samples can be wasteful.

The data analysis cycle

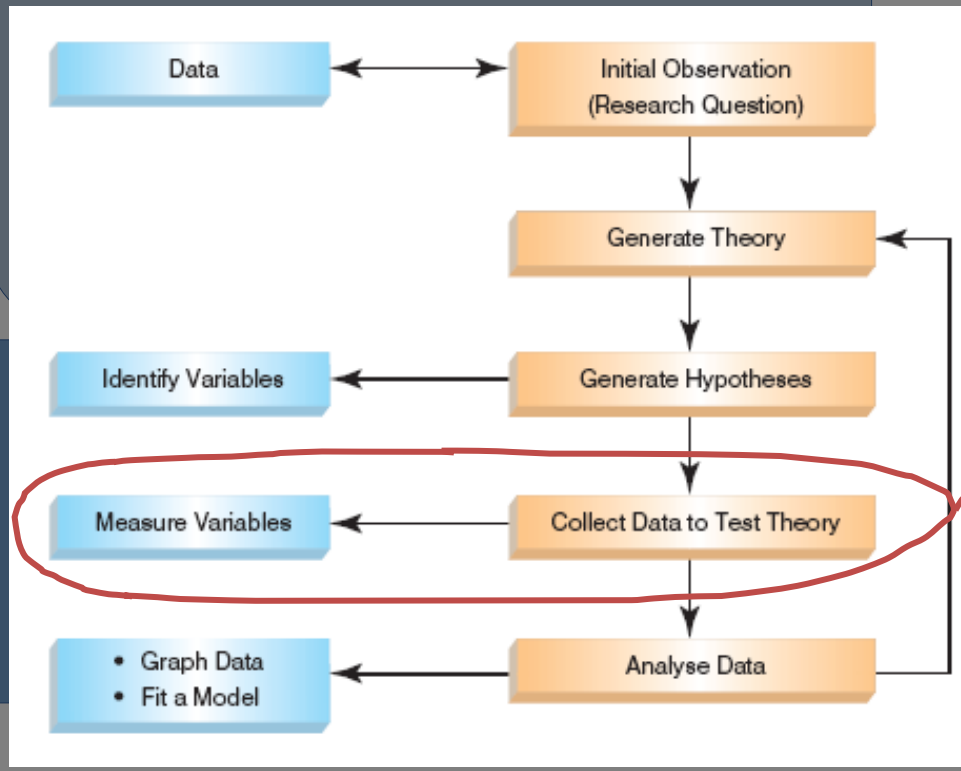
Visualisation/reporting

Summarisation and description

Exploration and knowledge extraction

- the analysis is only as good as the data, which must be **representative** of the reality we wish to study i.e. unbiased

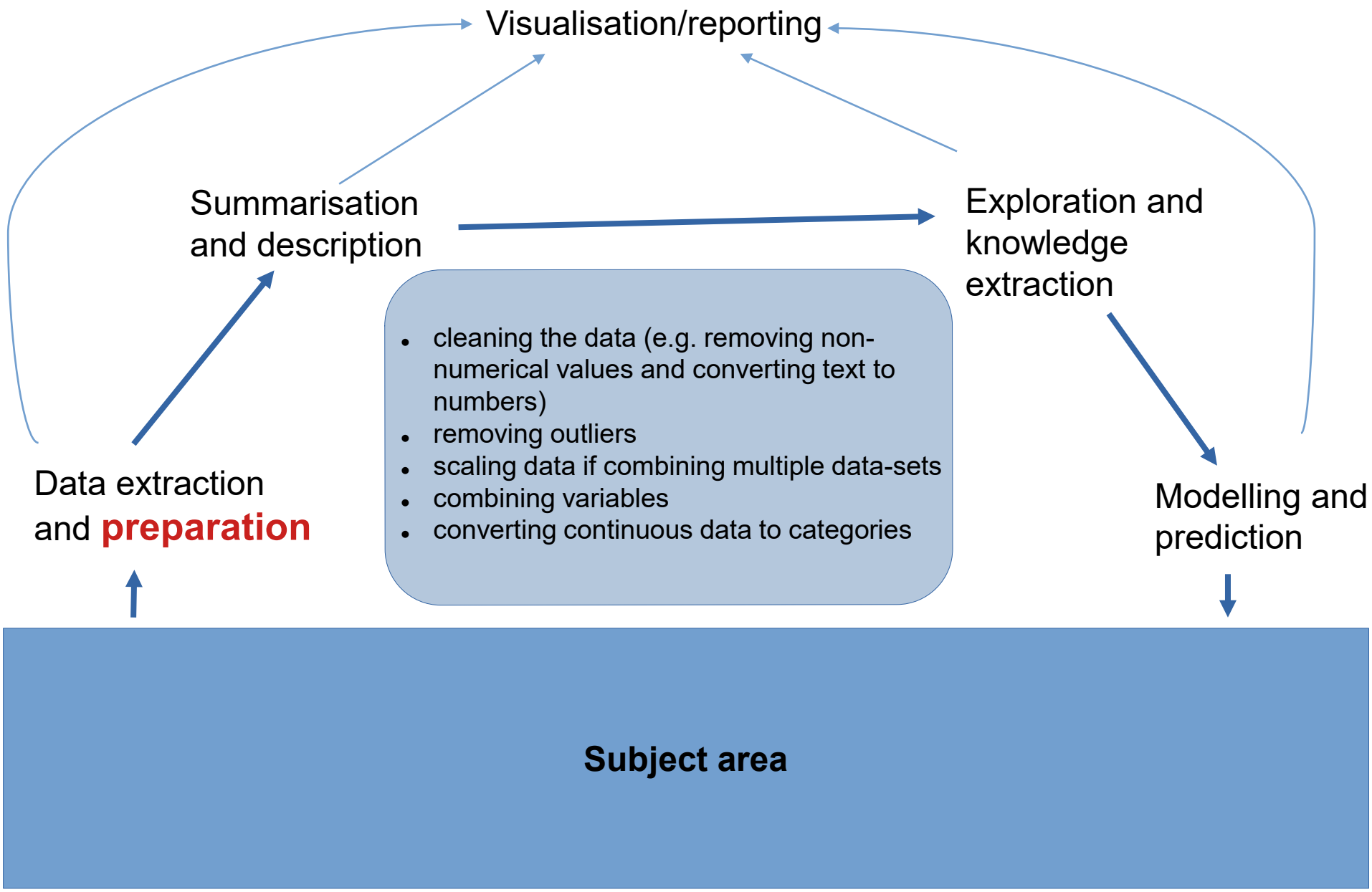
Data **extraction** and preparation



Source: [DSR]

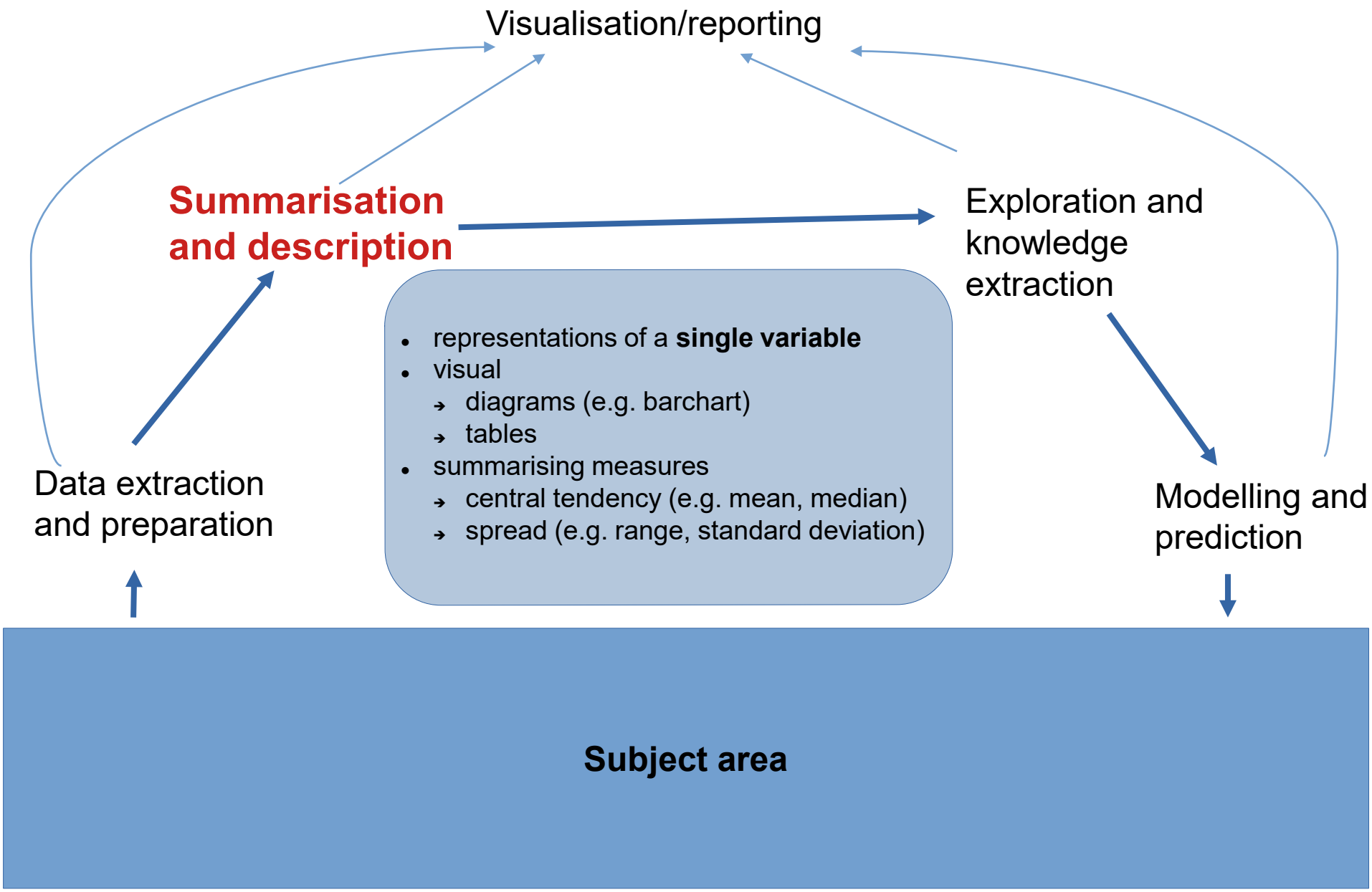
In the case of predictive models, an undesirable state of affairs, reflected in the data, can be **perpetuated** by a model that uses them (e.g. racism or misogyny). This, however, is a problem inherent in the data i.e. one of initial data selection rather than a problem in sampling.

The data analysis cycle

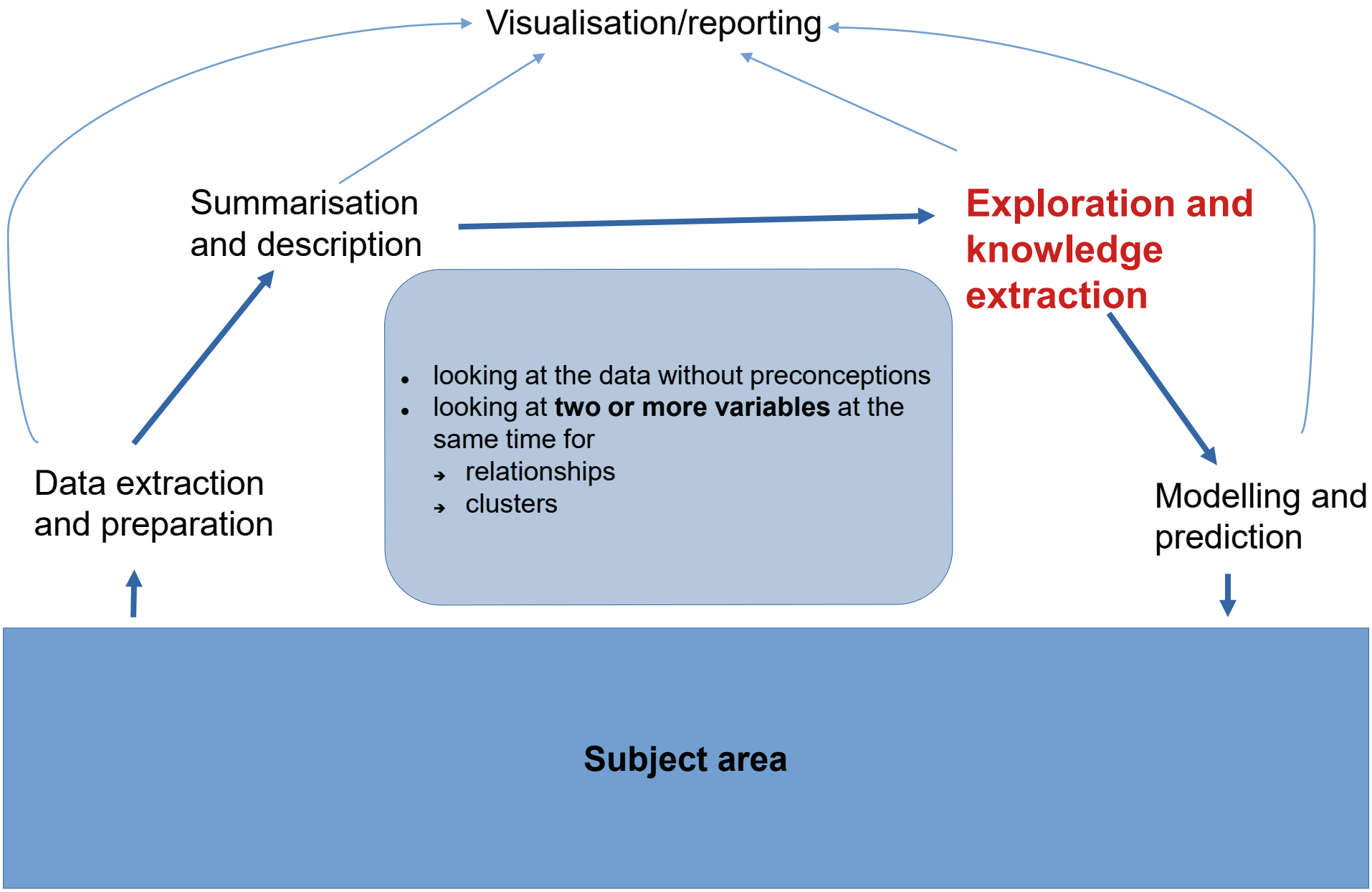


- cleaning the data (e.g. removing non-numerical values and converting text to numbers)
- removing outliers
- scaling data if combining multiple data-sets
- combining variables
- converting continuous data to categories

The data analysis cycle

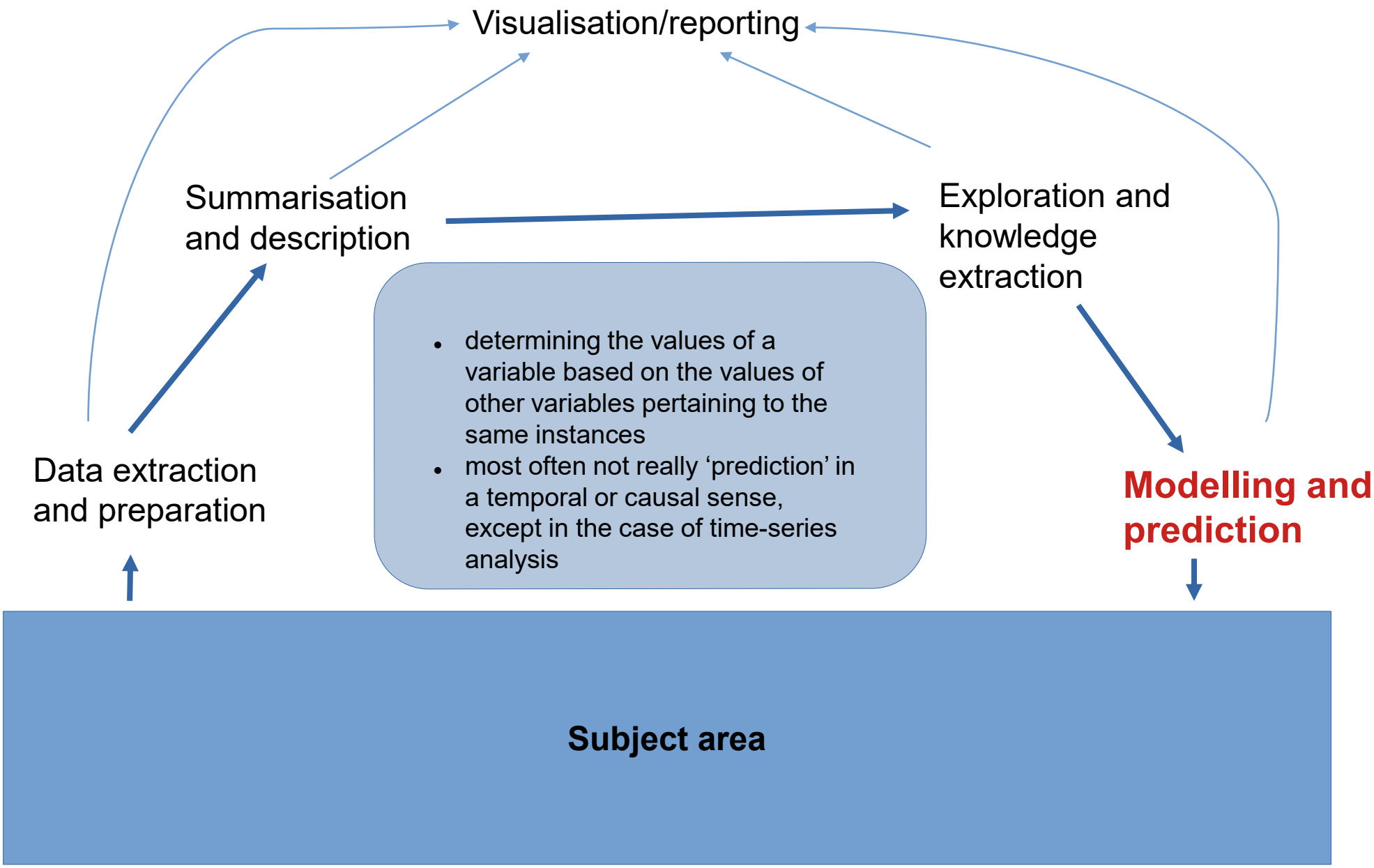


The data analysis cycle



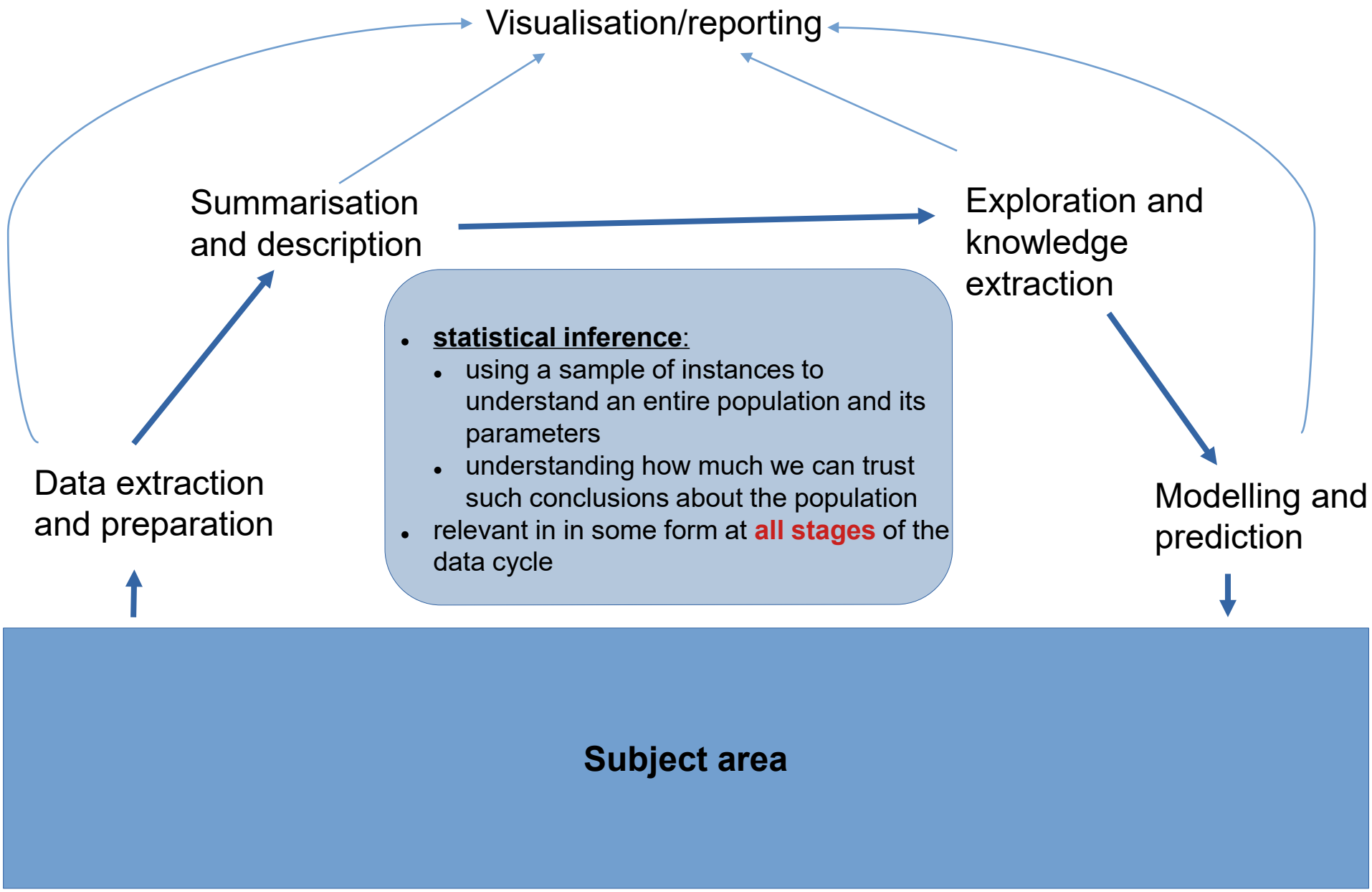
- looking at the data without preconceptions
- looking at **two or more variables** at the same time for
  - relationships
  - clusters

The data analysis cycle



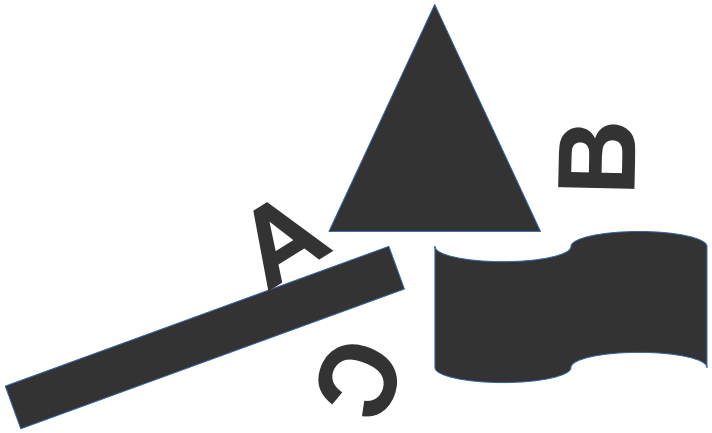
- determining the values of a variable based on the values of other variables pertaining to the same instances
- most often not really ‘prediction’ in a temporal or causal sense, except in the case of time-series analysis

The data analysis cycle



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

Statistics in the data analysis landscape



The data analysis landscape

Data science: concepts and theory behind the analysis of data

underlies

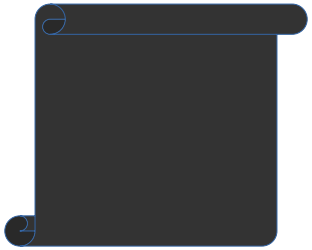
Statistics: mathematically-based description of and inquiry into data

Machine learning: computationally-based algorithms, for predictive modelling and exploration

operate on

Prepared data

In this module



you will learn

to…

- Present the **question** you are interested in
  in a way that makes sense to conduct a statistical analysis
- Inspect and **prepare** the data you have
  to support a statistical analysis
- **Describe** the data you have
  in a way that demonstrates the constraints placed on the analysis by the data
- Conduct a **statistical analysis**
  using appropriate statistical tests
- **Interpret the outcomes** of your statistical analysis
  drawing appropriate conclusions
- **Report** on the findings of your statistical analysis
  in a way that makes sense for your consumer

References

Some pictures in this presentation were taken from the following books. They are cited using the keys shown in square brackets.

[DSR] *Discovering Statistics Using R*, by Andy Field, Jeremy Miles and Zöe Field, Sage, 2012.