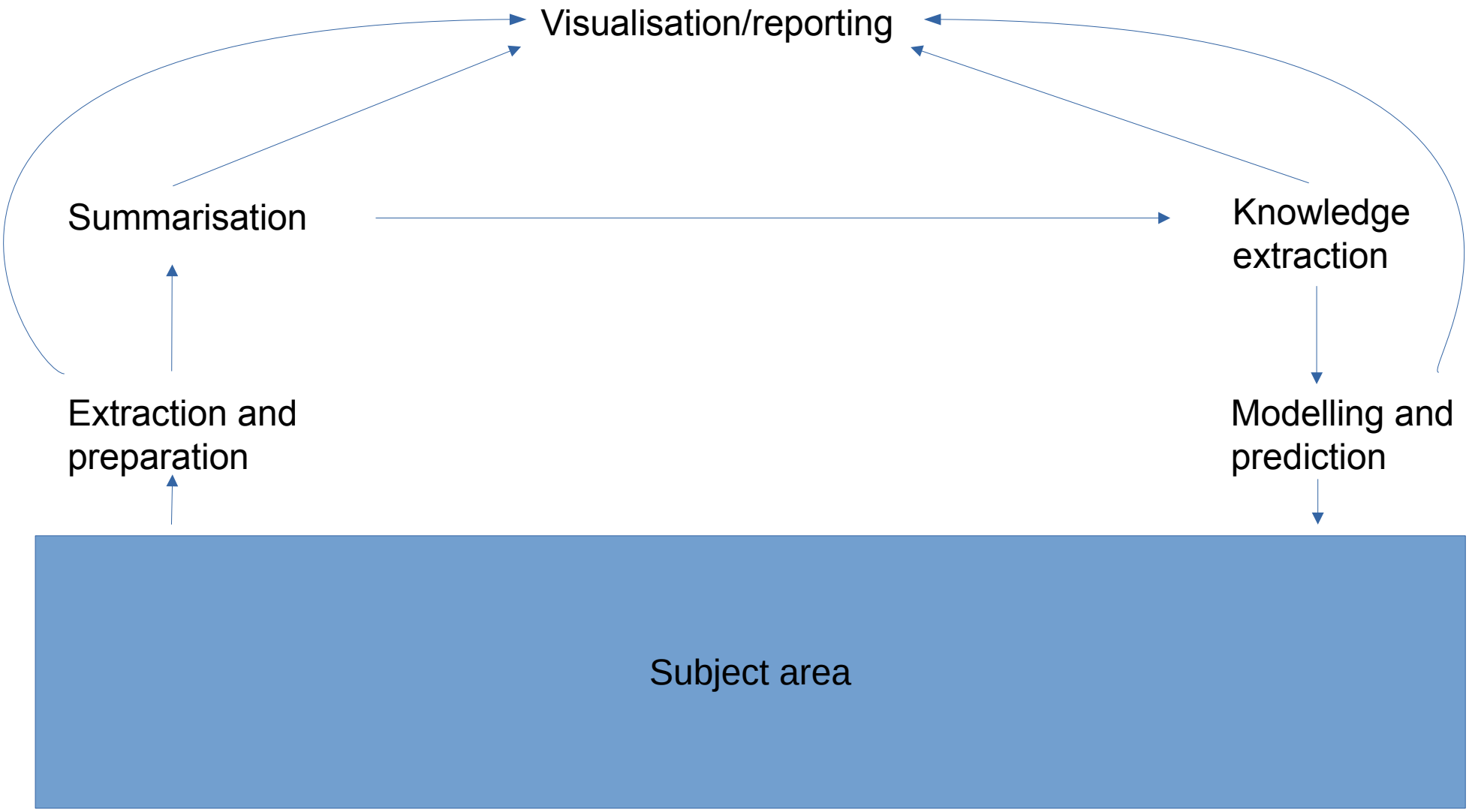
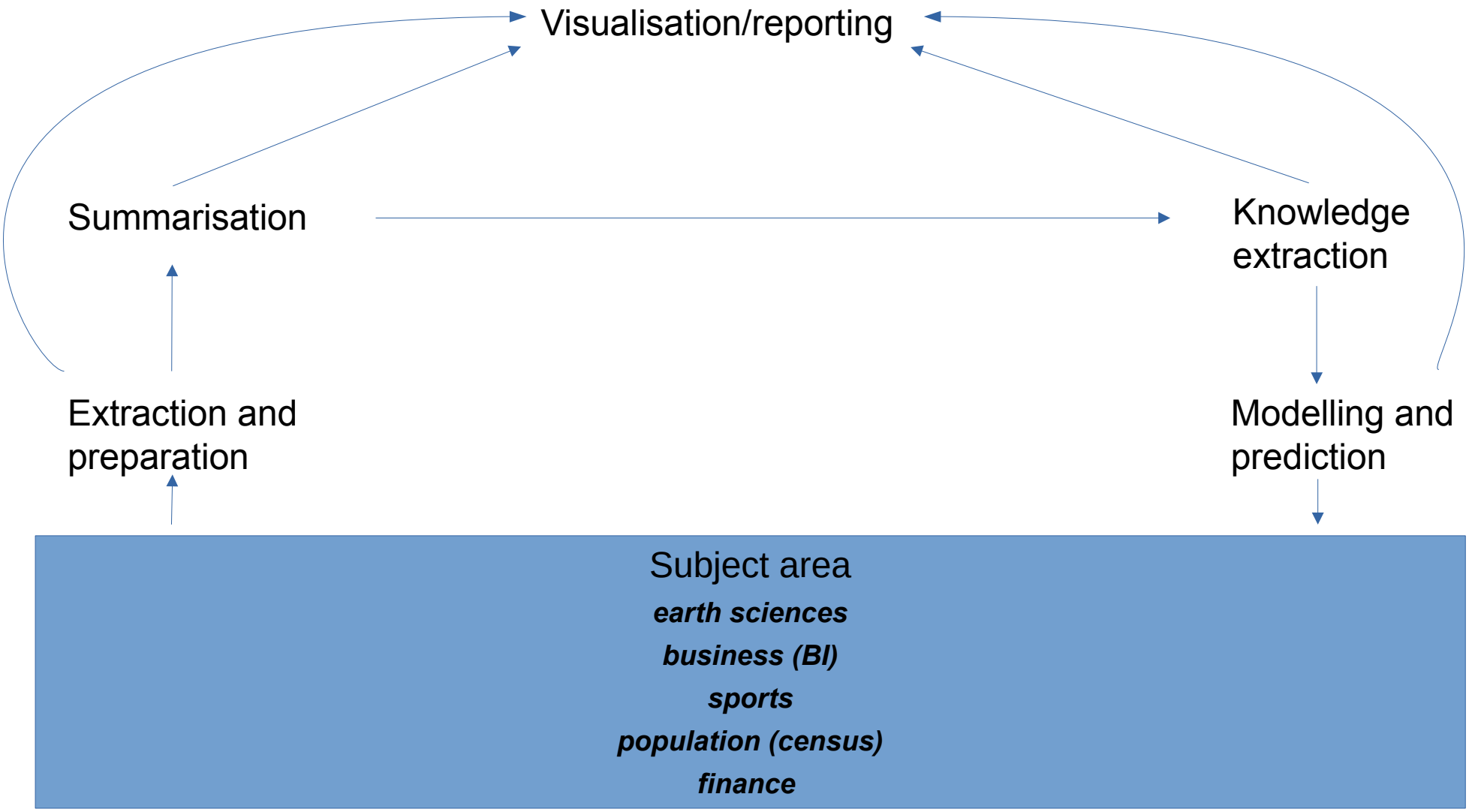
Data Analysis: Introduction

Institute of Technology Tallaght,
Department of Computing

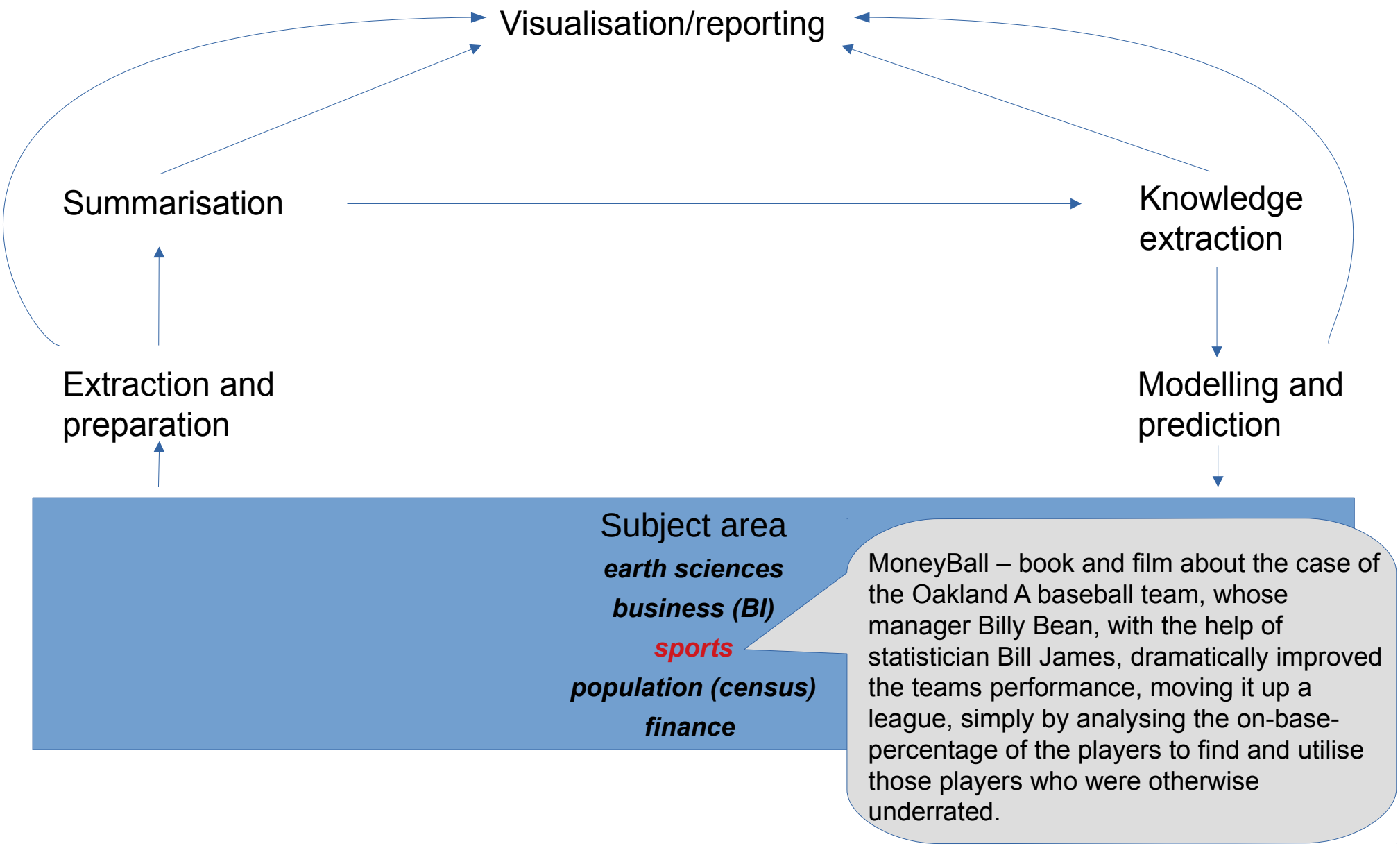
The data cycle



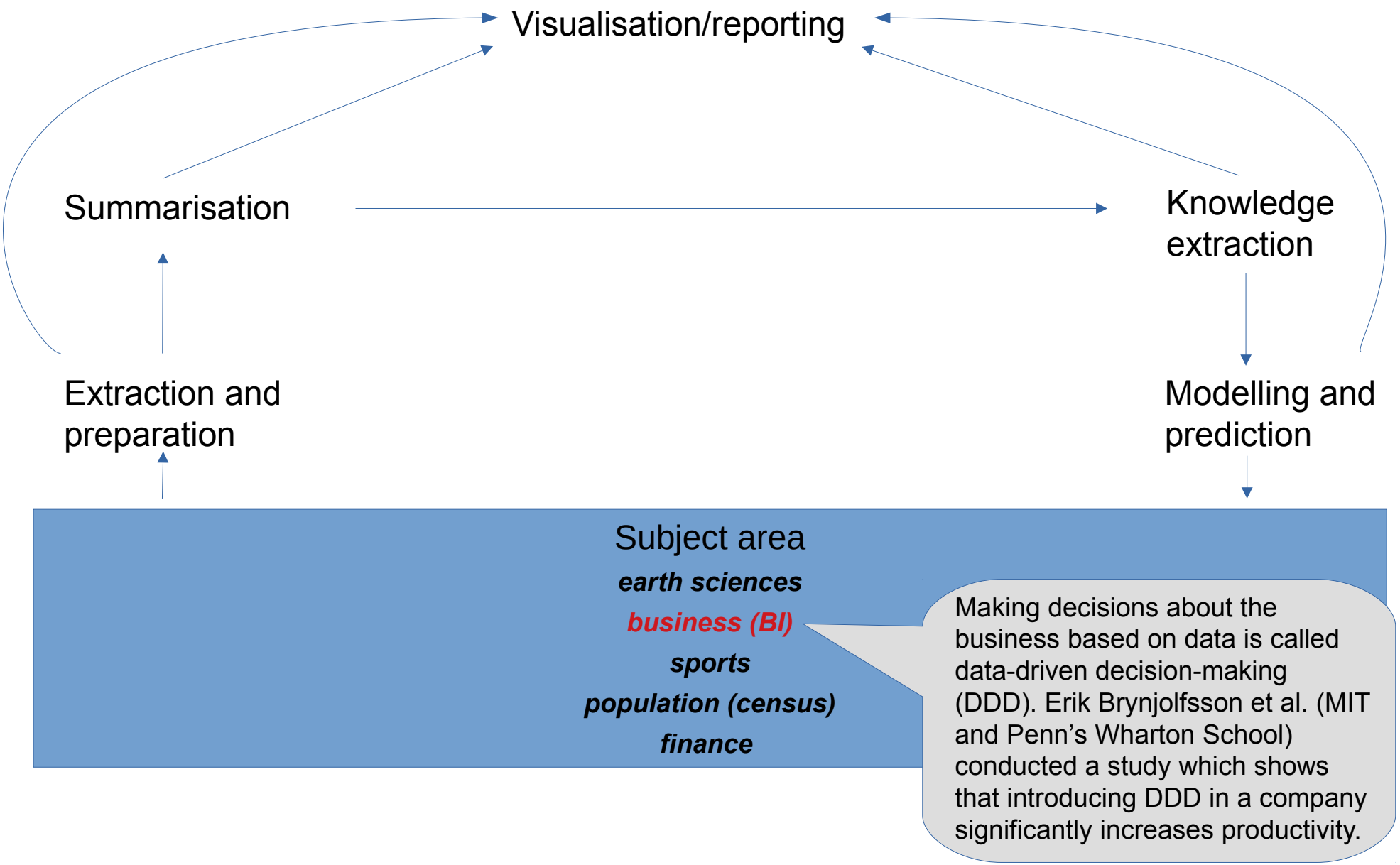
The data cycle



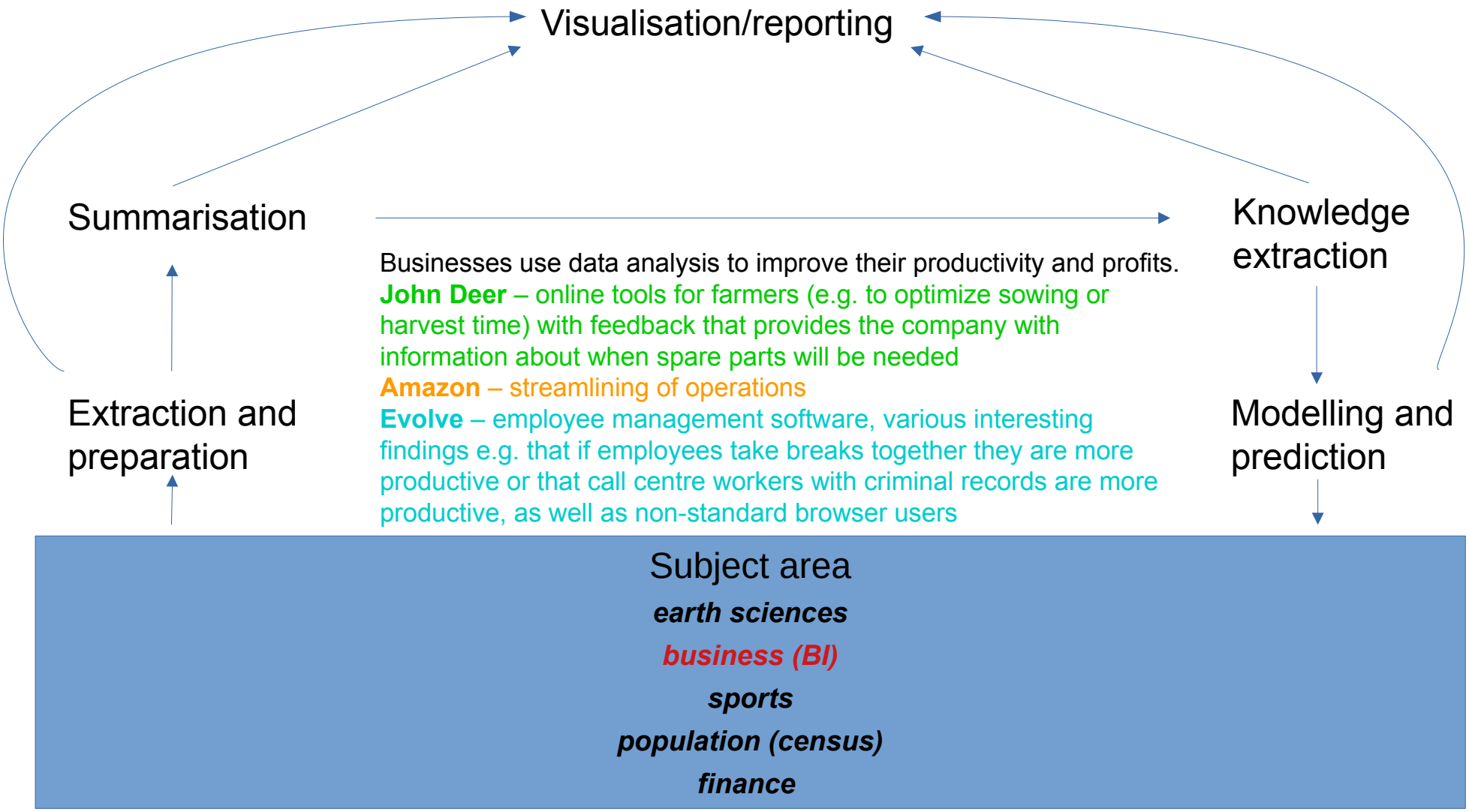
The data cycle



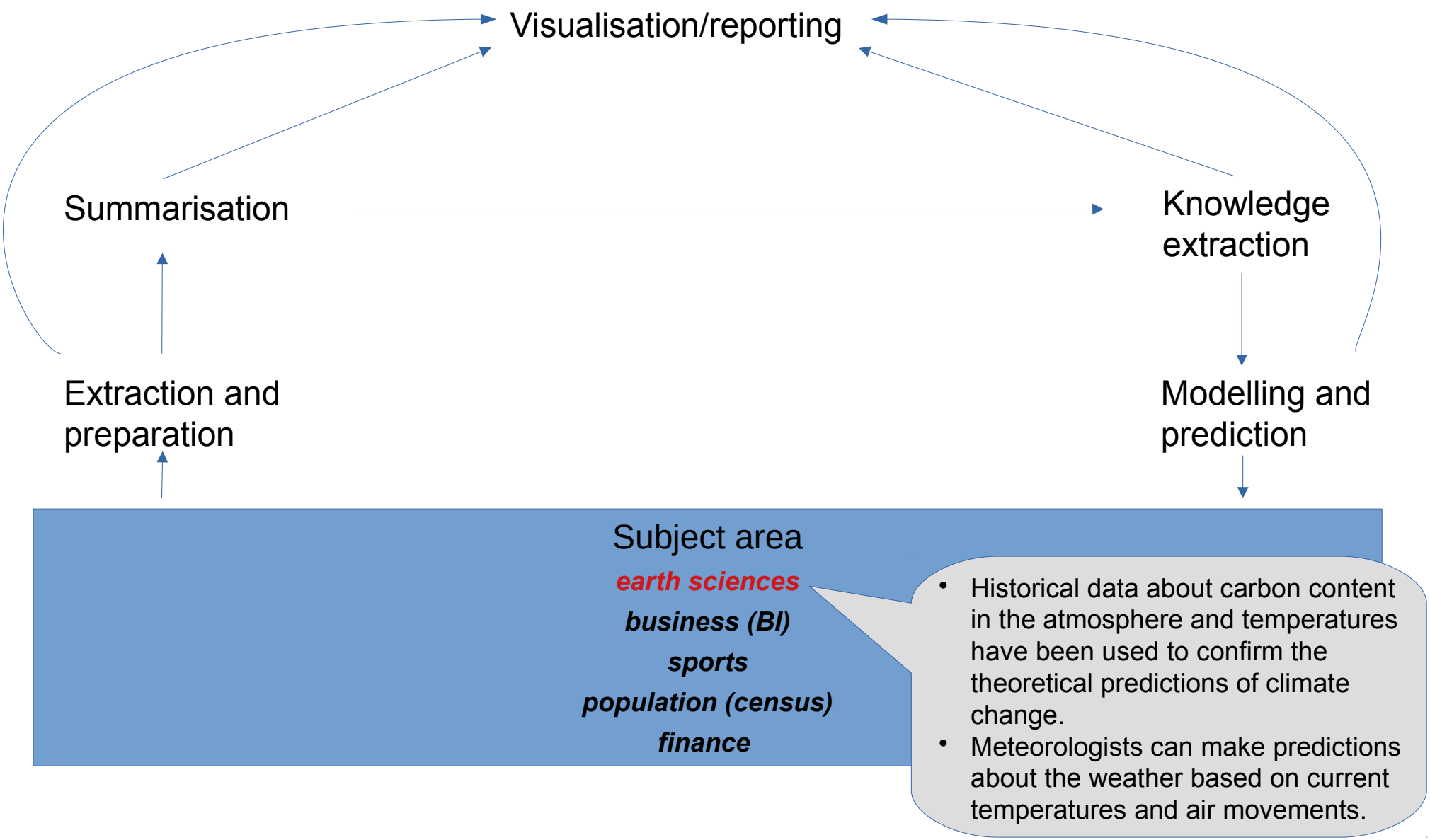
The data cycle



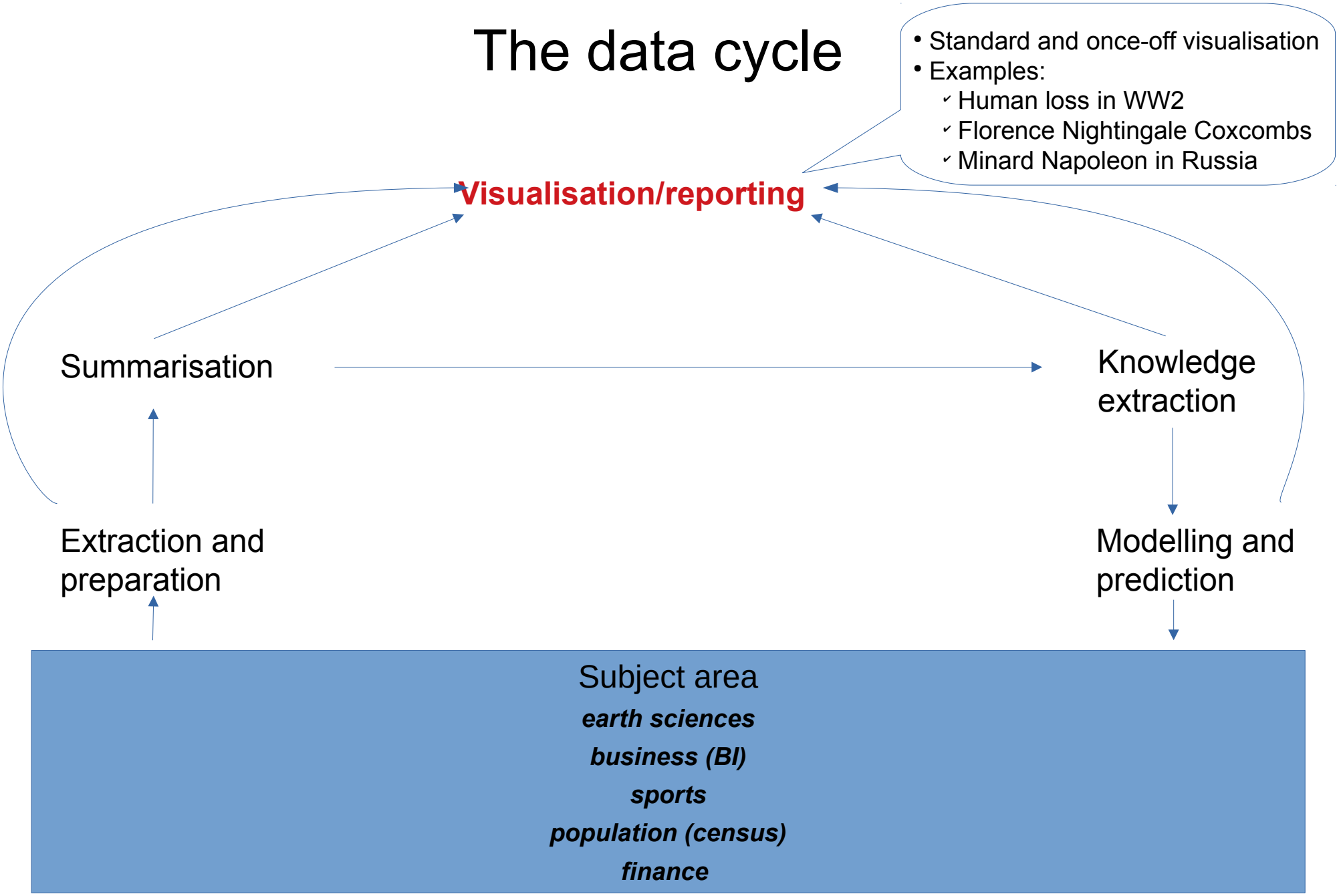
The data cycle



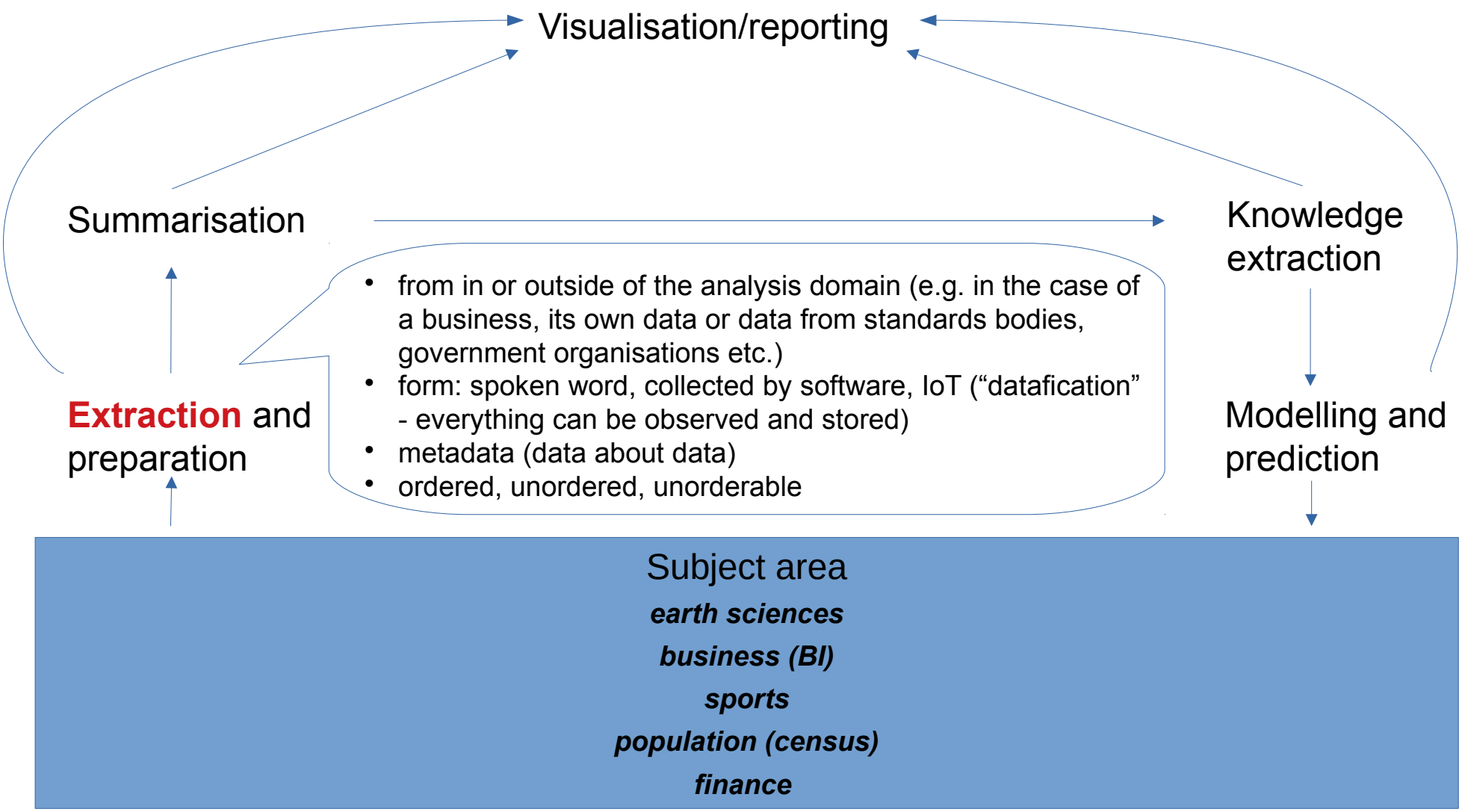
The data cycle



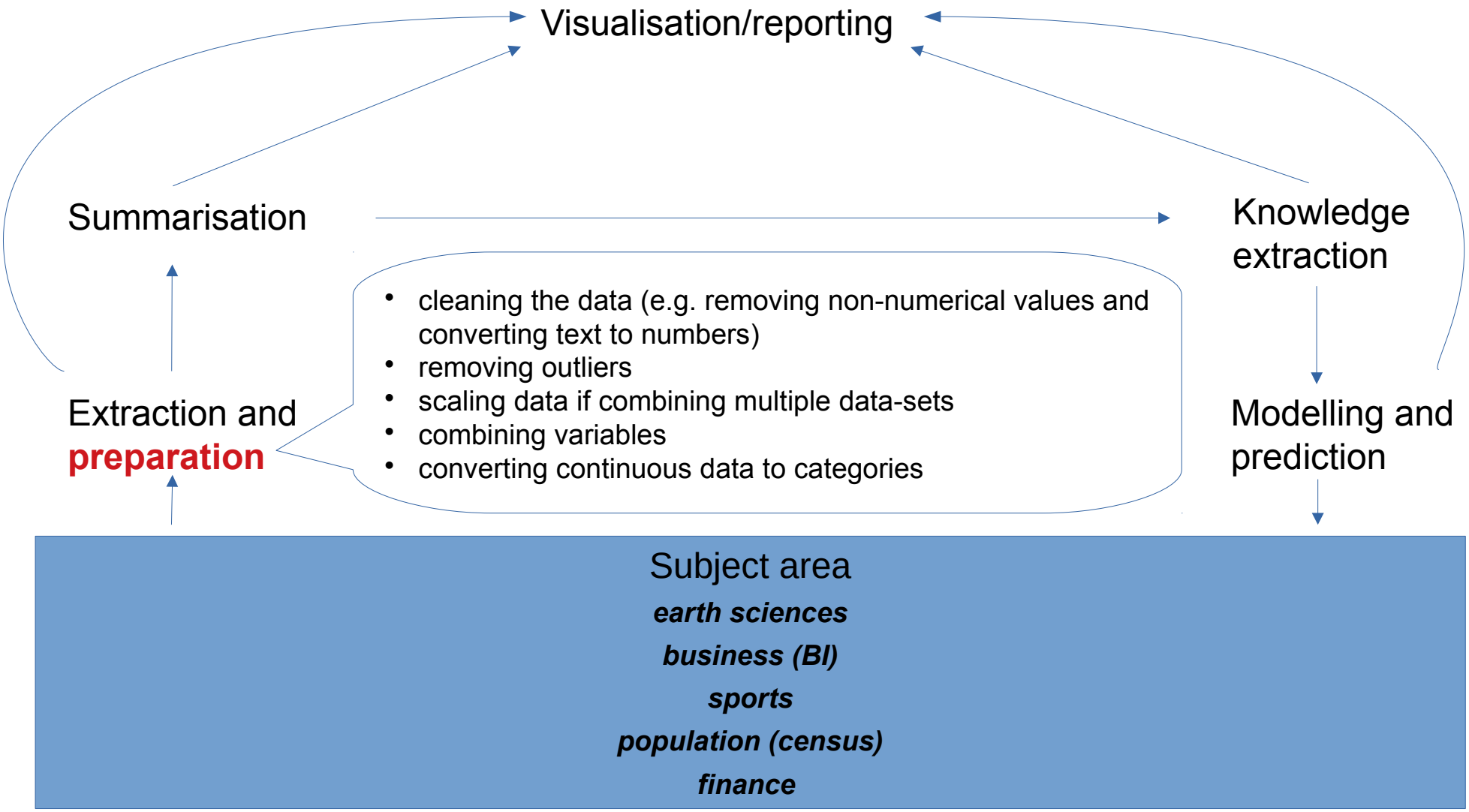
The data cycle



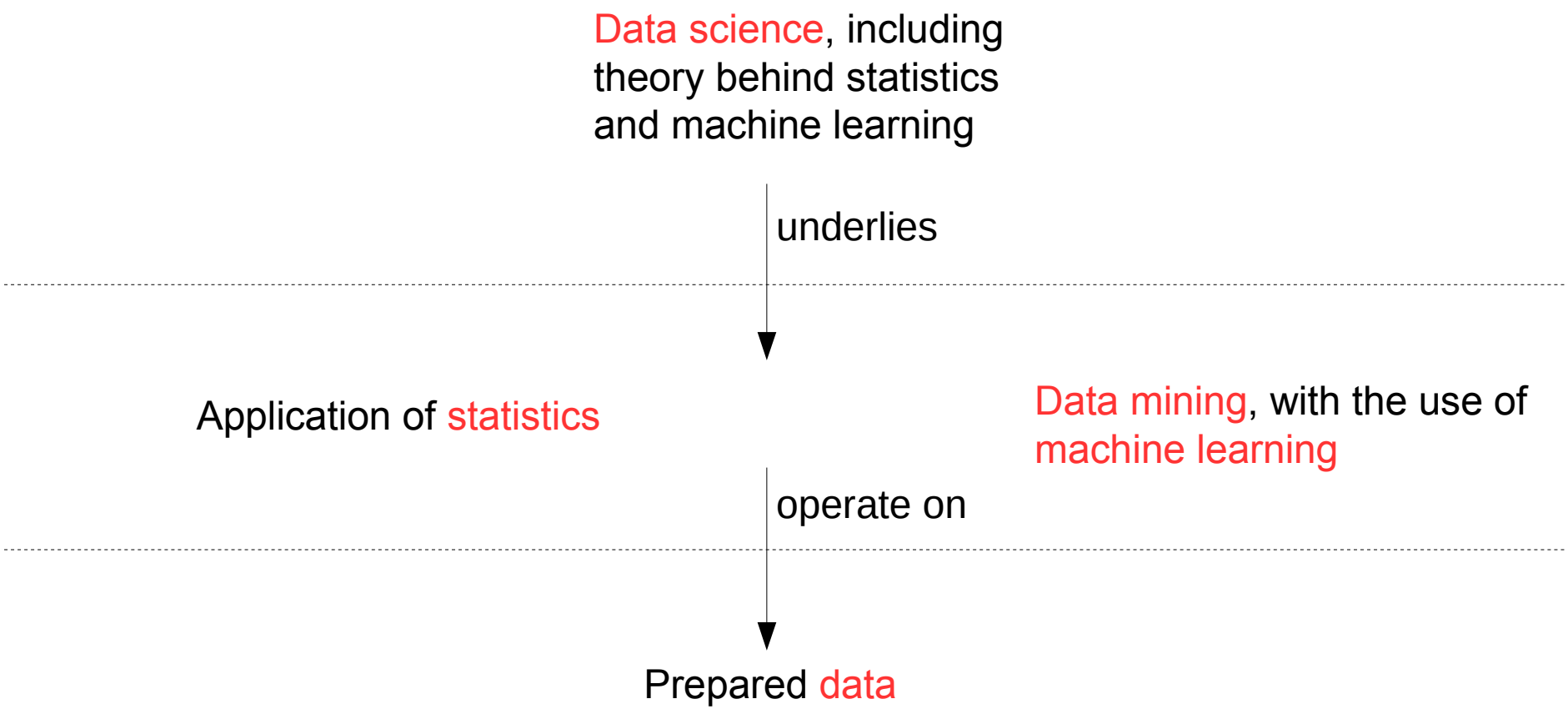
The data cycle



The data cycle



The data analysis landscape



The data analysis landscape

- Analytics – a group of statistical and data mining techniques used in a particular problem domain e.g. business analytics, financial analytics.

Data science, including theory behind statistics and machine learning

underlies

Application of statistics

Data mining, with the use of machine learning

operate on

Prepared data

The data analysis landscape

- Analytics – a group of statistical and data mining techniques used in a particular problem domain e.g. business analytics, financial analytics.
- Data analysis – a generic term for any instance of analysis of data.

Data science, including theory behind statistics and machine learning

underlies

Application of statistics

Data mining, with the use of machine learning

operate on

Prepared data

The data analysis landscape

- Analytics – a group of statistical and data mining techniques used in a particular problem domain e.g. business analytics, financial analytics.
- Data analysis – a generic term for any instance of analysis of data.
- Big data – the same as below, only bigger!

Data science, including theory behind statistics and machine learning

underlies

Application of statistics

Data mining, with the use of machine learning

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 are amounts of data larger than can be processed with conventional technologies.
- New technologies:
  - Hadoop (Apache)
  - MapReduce (Google)
  - MongoDB etc.
- The data science principles are the same as ‘normal sized’ data

- 4 Vs
  IBM 4Vs of Big Data

The data analysis landscape

- Analytics – a group of statistical and data mining techniques used in a particular problem domain e.g. business analytics, financial analytics.
- Data analysis – a generic term for any instance of analysis of data.
- Big data – the same as below, only bigger!

- 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 scientist visualisation

Data science, including theory behind statistics and machine learning

underlies

Application of statistics

Data mining, with the use of machine learning

operate on

Prepared data

The data analysis landscape

- Analytics – a group of statistical and data mining techniques used in a particular problem domain e.g. business analytics, financial analytics.
- Data analysis – a generic term for any instance of analysis of data.
- Big data – the same as below, only bigger!

- 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

Application of **statistics**

Data mining, with the use of machine learning

operate on

Prepared data

The data analysis landscape

- Analytics – a group of statistical and data mining techniques used in a particular problem domain e.g. business analytics, financial analytics.
- Data analysis – a generic term for any instance of analysis of data.
- Big data – the same as below, only bigger!

Data science, including theory behind statistics and machine learning

underlies

Application of statistics

operate on

Prepared data

- The practice of finding patterns in data and extracting from it useful information that is not immediately available
- In the 1990s company data was consolidated into **enterprise data warehouses**, which could be mined for data

Data mining, with the use of machine learning

The data analysis landscape

- Analytics – a group of statistical and data mining techniques used in a particular problem domain e.g. business analytics, financial analytics.
- Data analysis – a generic term for any instance of analysis of data.
- Big data – the same as below, only bigger!

Data science, including theory behind statistics and machine learning

underlies

Application of statistics

operate on

Prepared data

Data mining, with the use of **machine learning**

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

“Learning” Data Analysis

- Asking questions, then investigating if they can be answered by analysing data
- Methods and techniques for all the stages of the data cycle
- Understanding when to apply the various methods and techniques
- Adopting the ‘every case is different’ approach