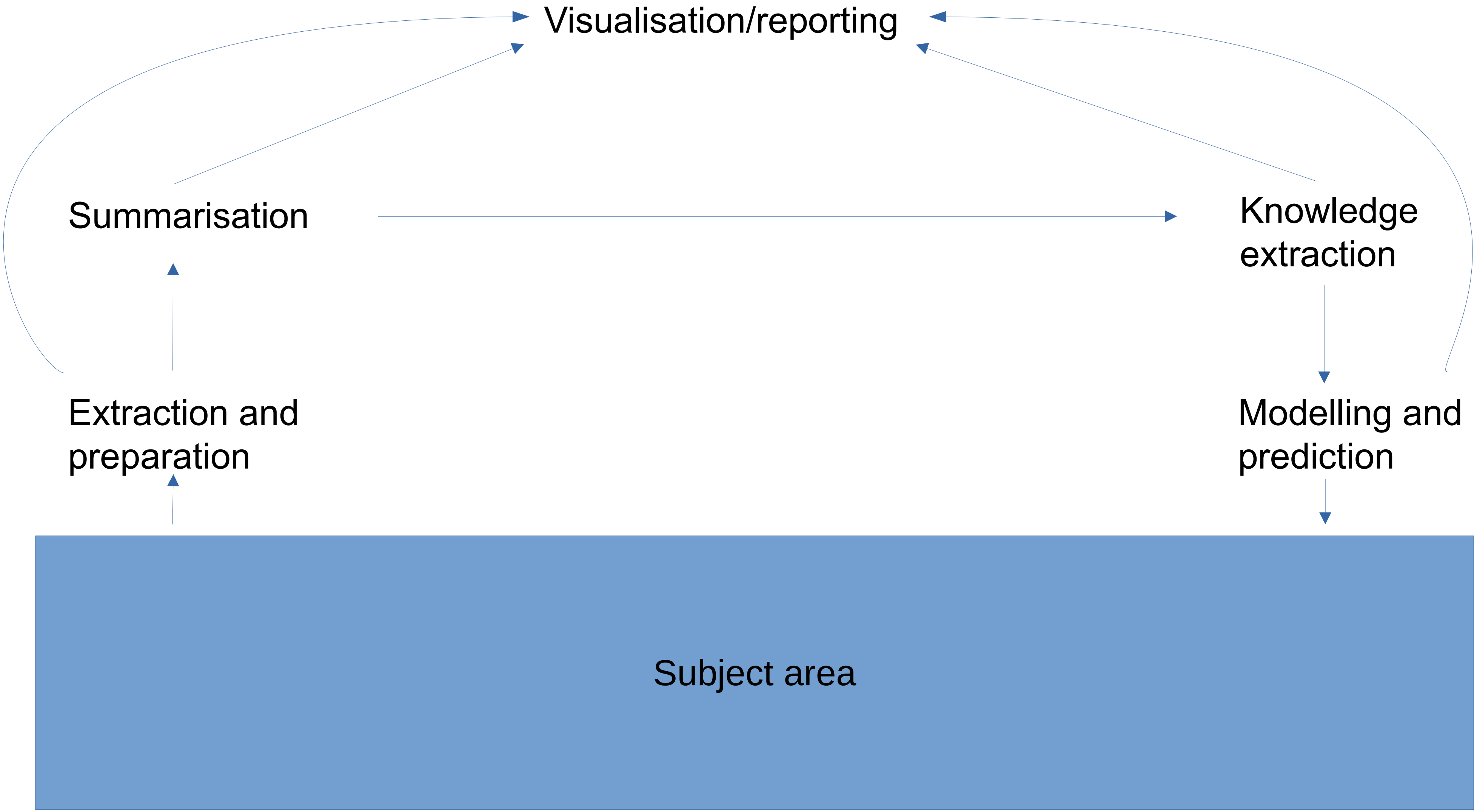
Data Analysis: Introduction

TU Dublin Tallaght,
Department of Computing

The data cycle



Visualisation/reporting

Summarisation

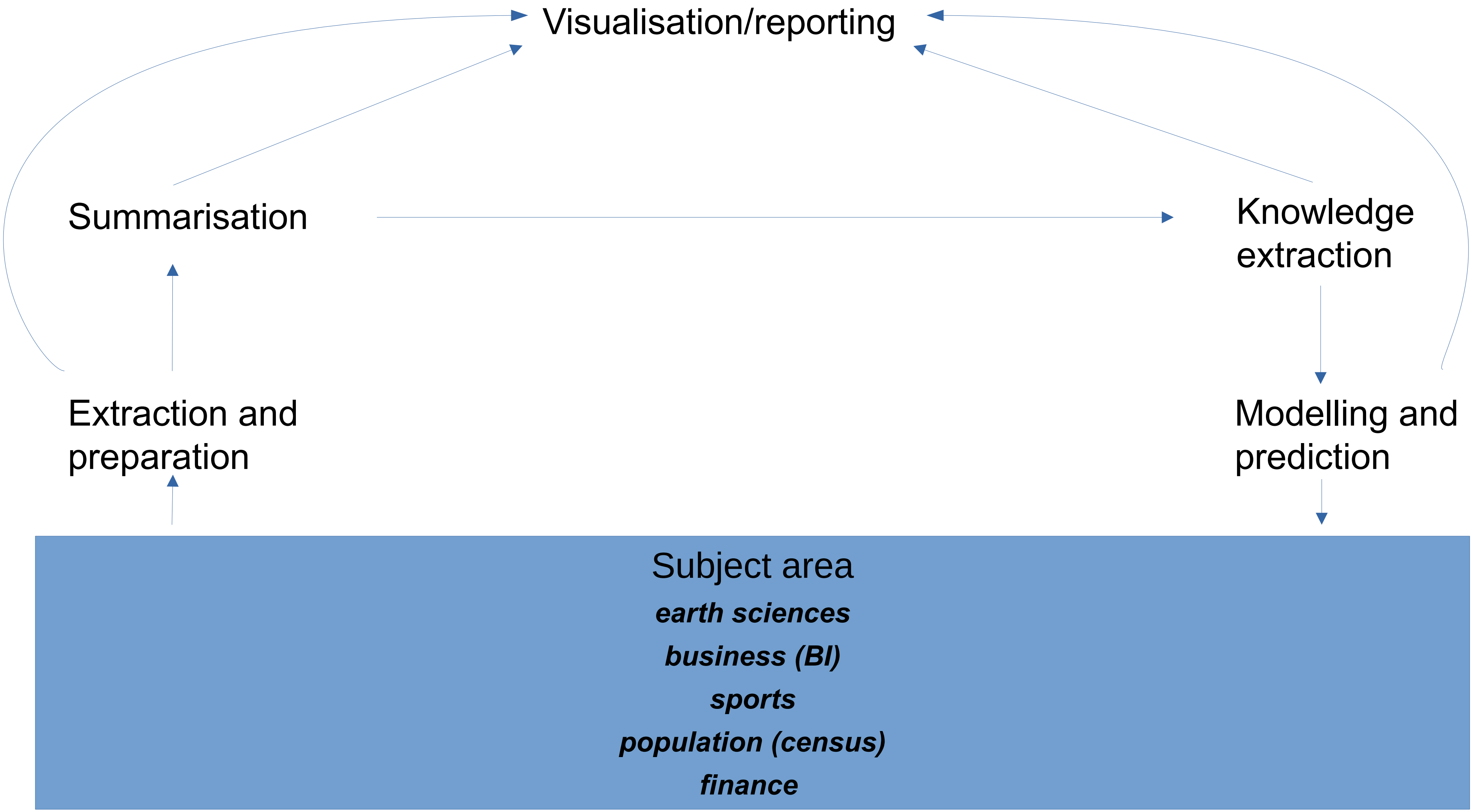
Knowledge extraction

Extraction and preparation

Modelling and prediction

Subject area

The data cycle



Visualisation/reporting

Summarisation

Knowledge extraction

Extraction and preparation

Modelling and prediction

Subject area

earth sciences

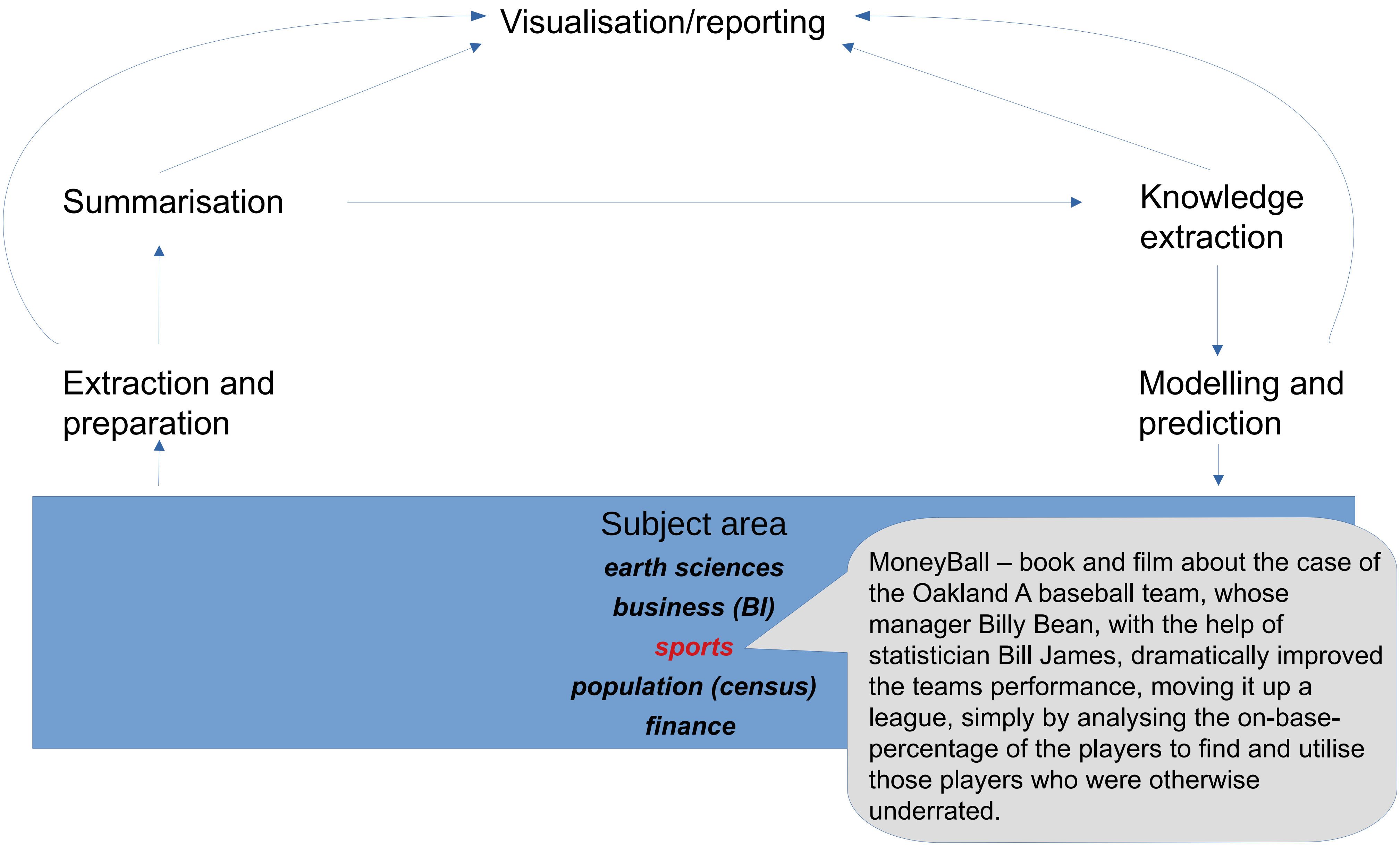
business (BI)

sports

population (census)

finance

The data cycle



Visualisation/reporting

Summarisation

Knowledge extraction

Extraction and preparation

Modelling and prediction

Subject area

earth sciences

business (BI)

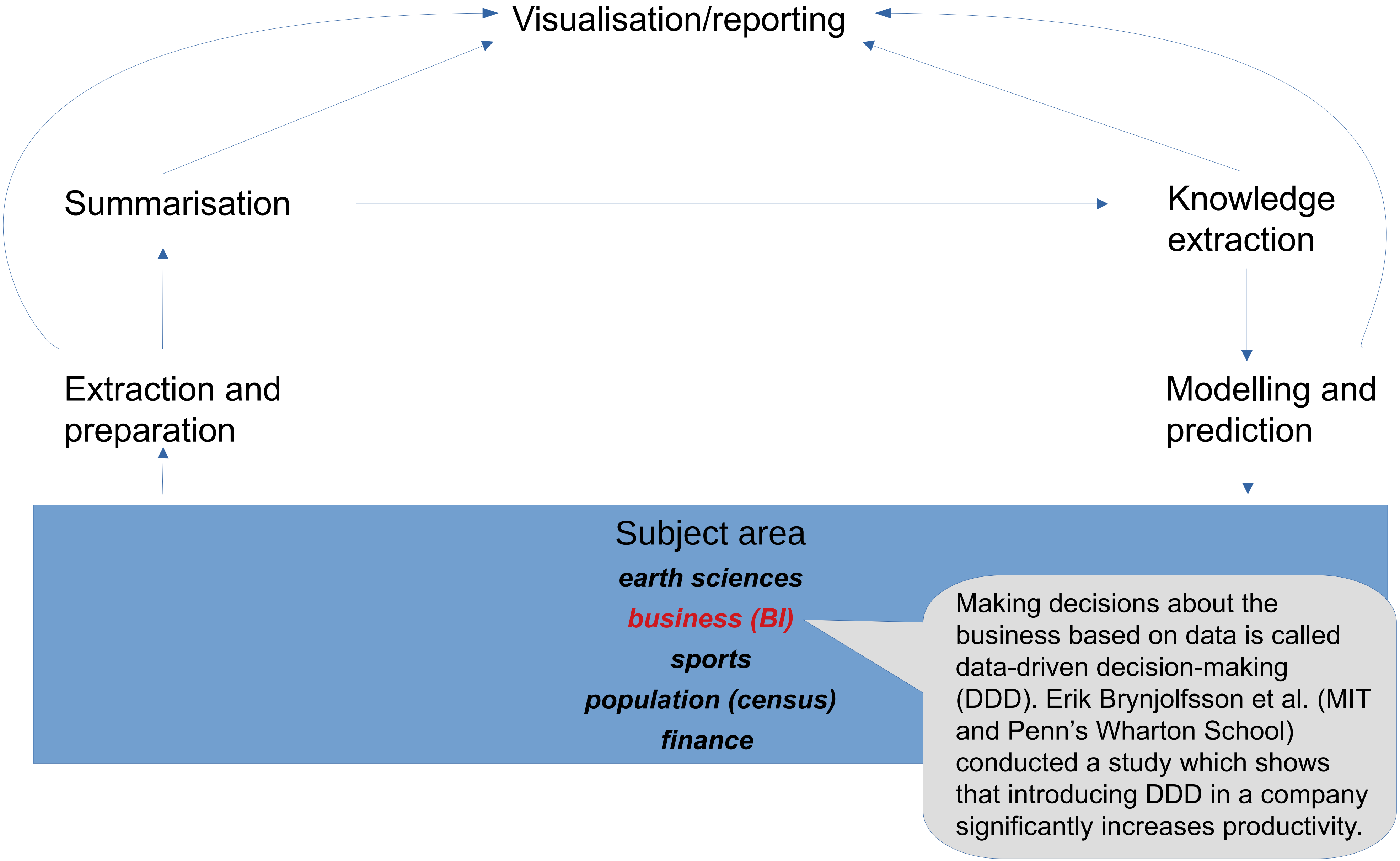
sports

population (census)

finance

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



Visualisation/reporting

Summarisation

Knowledge extraction

Extraction and preparation

Modelling and prediction

Subject area

earth sciences

business (BI)

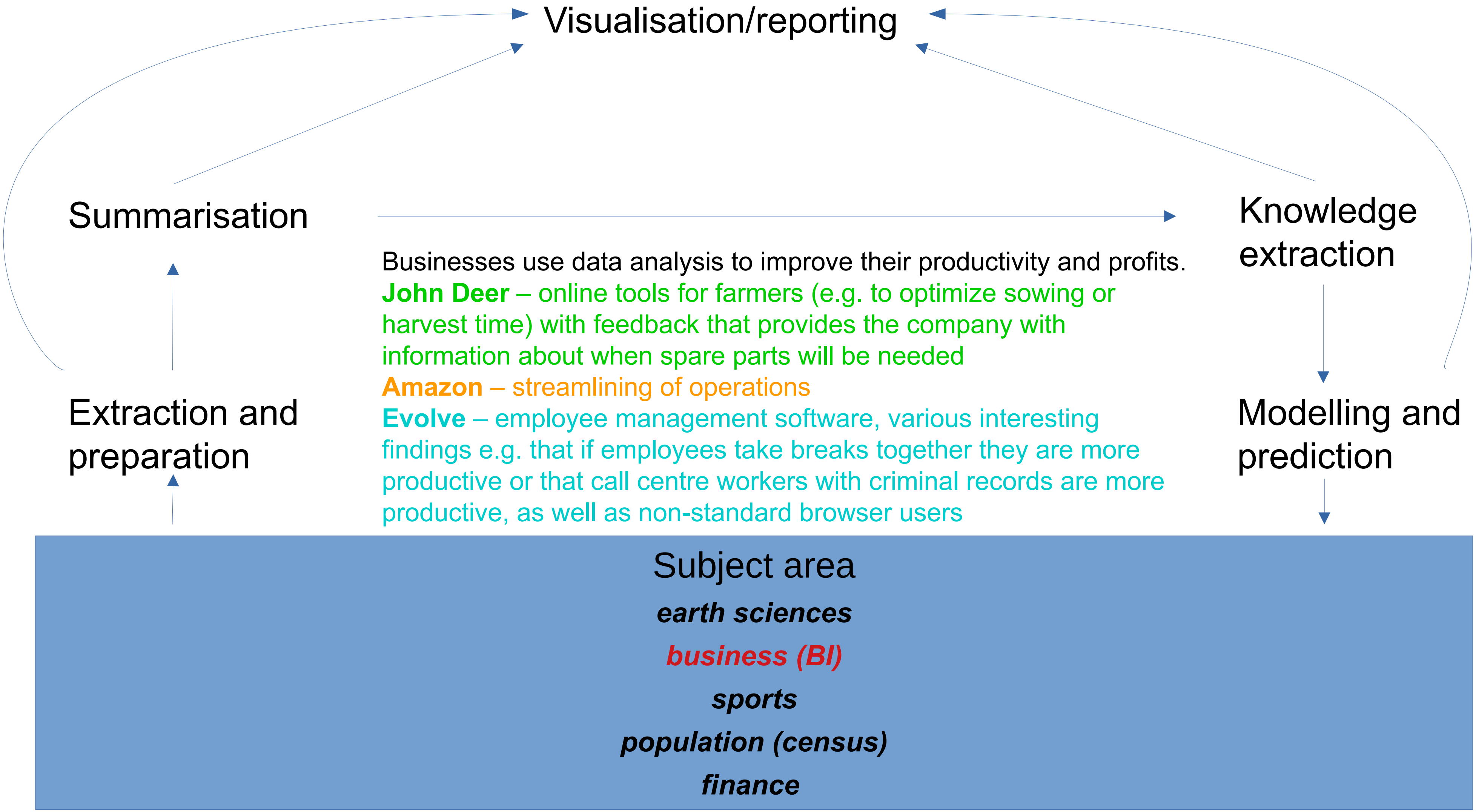
sports

population (census)

finance

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.

The data cycle



Businesses use data analysis to improve their productivity and profits.

John Deer – online tools for farmers (e.g. to optimize sowing or harvest time) with feedback that provides the company with information about when spare parts will be needed

Amazon – streamlining of operations

Evolve – employee management software, various interesting findings e.g. that if employees take breaks together they are more productive or that call centre workers with criminal records are more productive, as well as non-standard browser users

Subject area

earth sciences

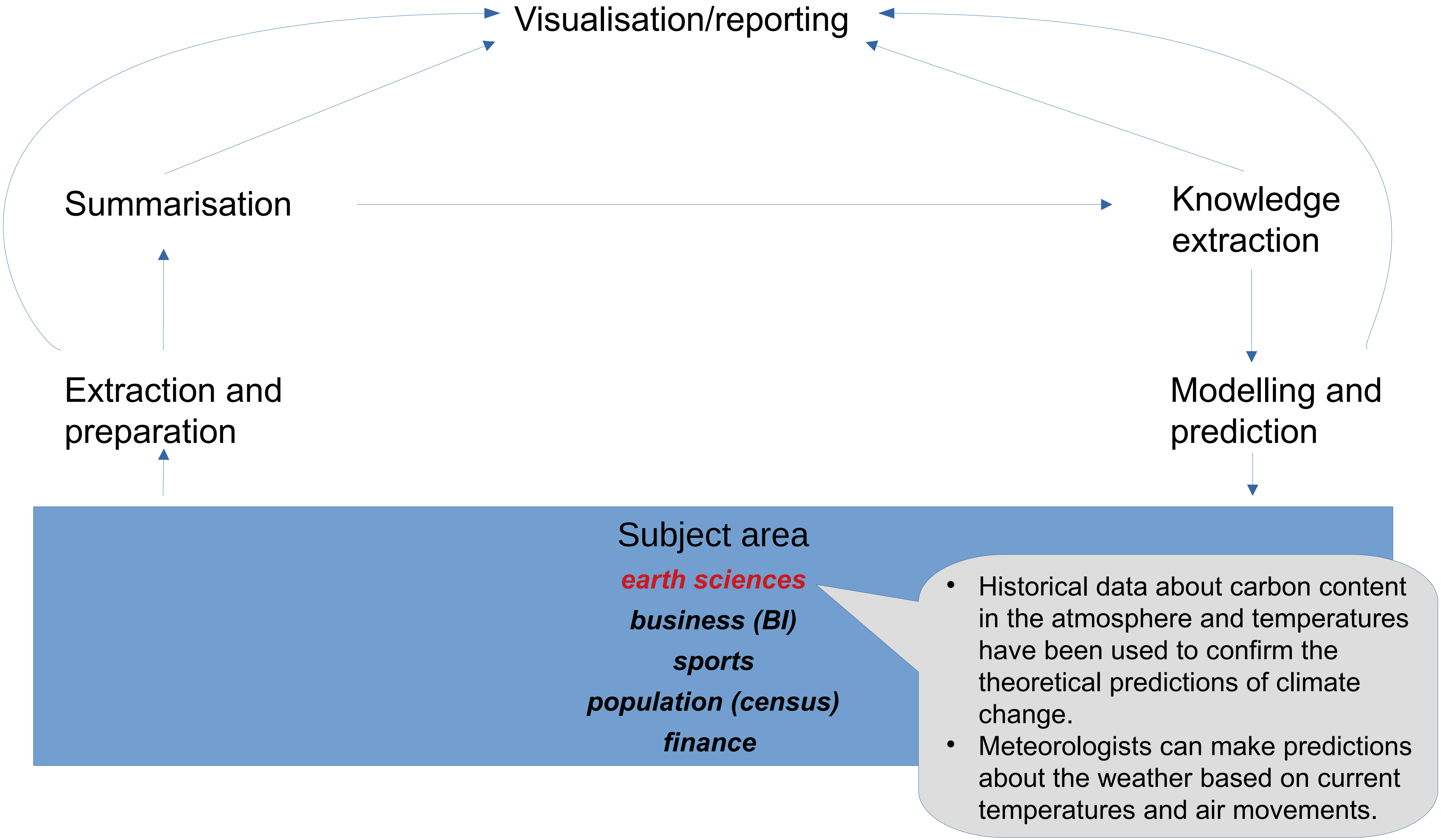
business (BI)

sports

population (census)

finance

The data cycle



Visualisation/reporting

Summarisation

Knowledge extraction

Extraction and preparation

Modelling and prediction

Subject area

earth sciences

business (BI)

sports

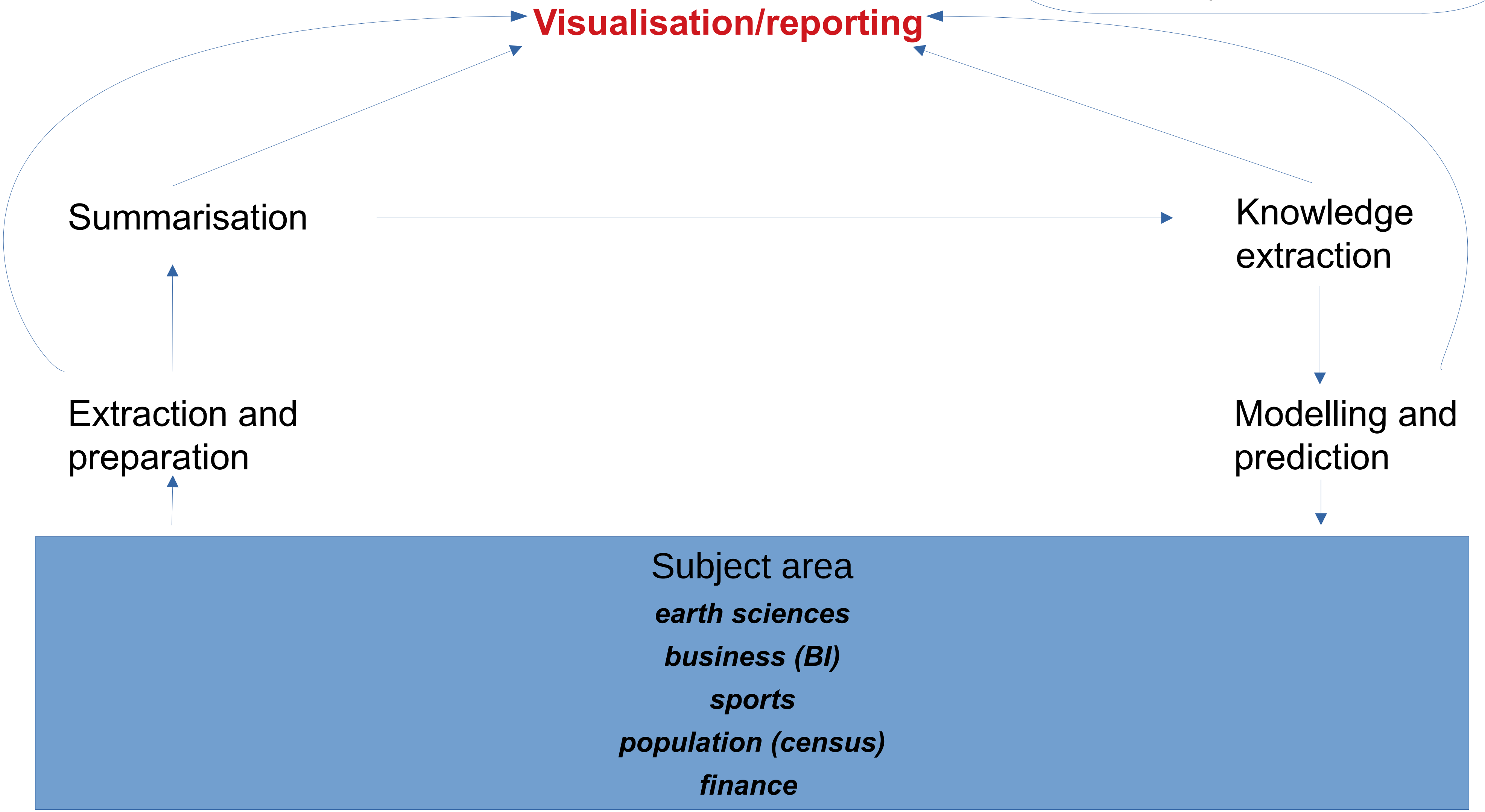
population (census)

finance

- 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

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



Visualisation/reporting

Summarisation

Knowledge extraction

Extraction and preparation

Modelling and prediction

Subject area

earth sciences

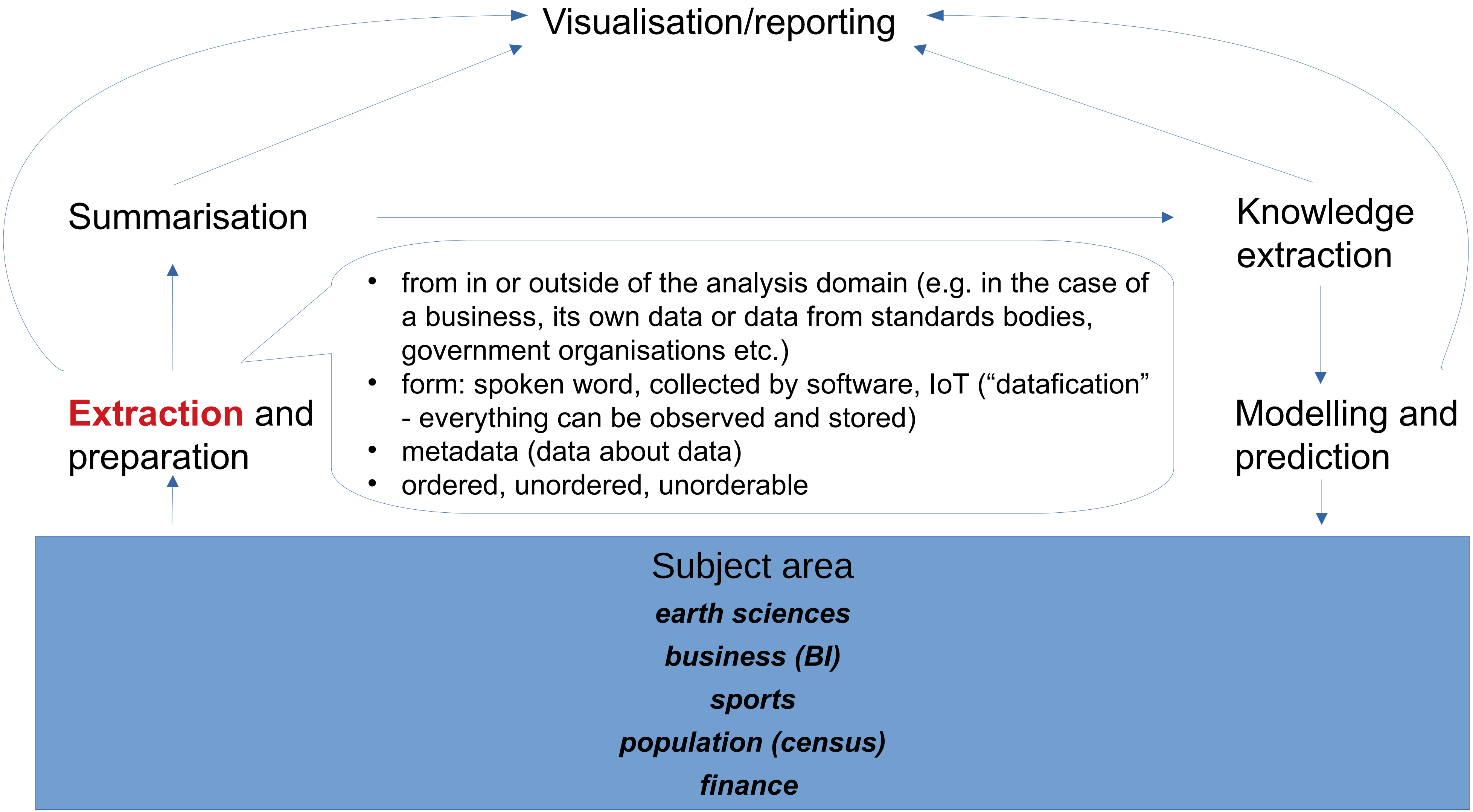
business (BI)

sports

population (census)

finance

The data cycle



Visualisation/reporting

Summarisation

Knowledge extraction

Extraction and preparation

- 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, unorderable

Modelling and prediction

Subject area

earth sciences

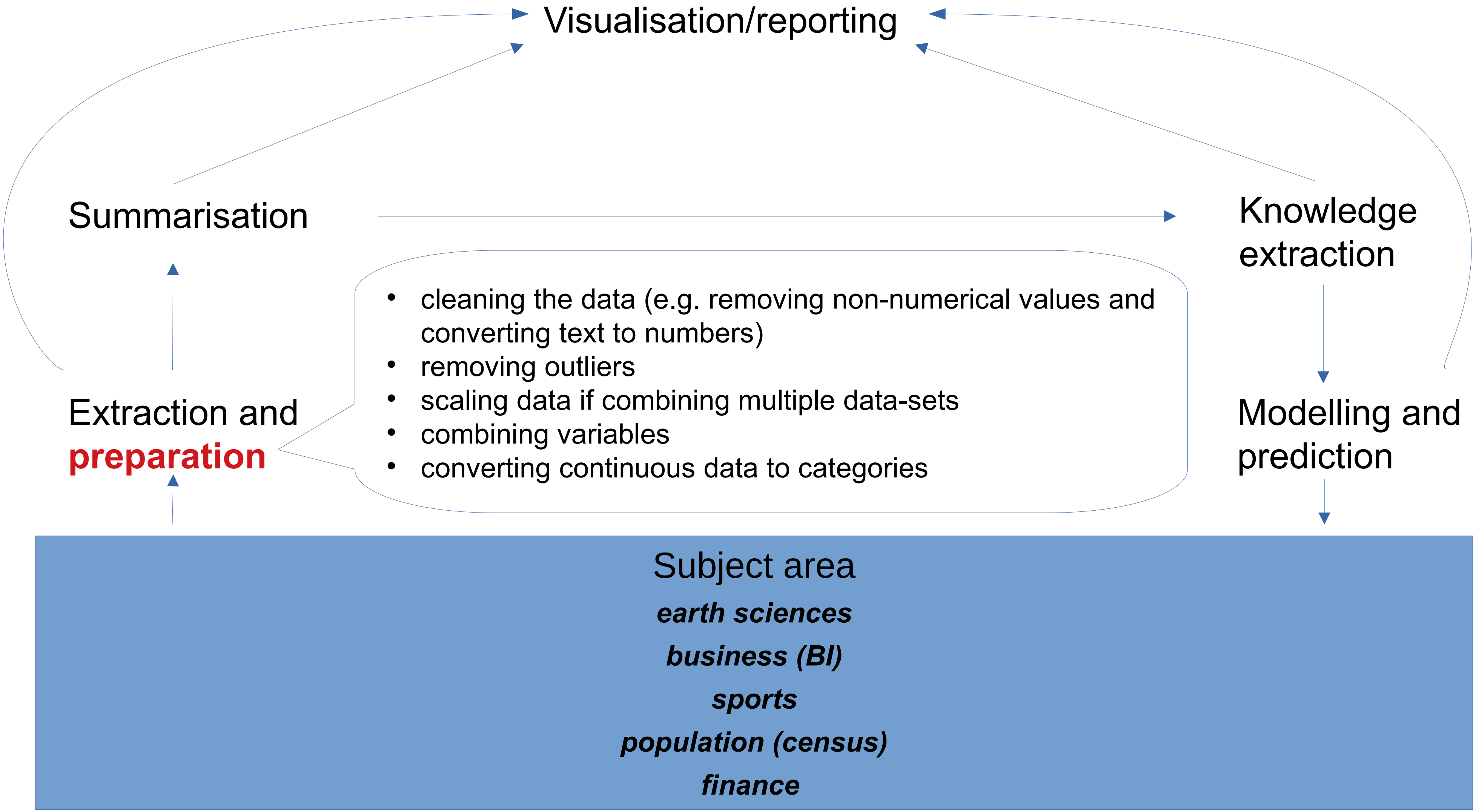
business (BI)

sports

population (census)

finance

The data cycle



Visualisation/reporting

Summarisation

Knowledge extraction

Extraction and **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

Modelling and prediction

Subject area

earth sciences

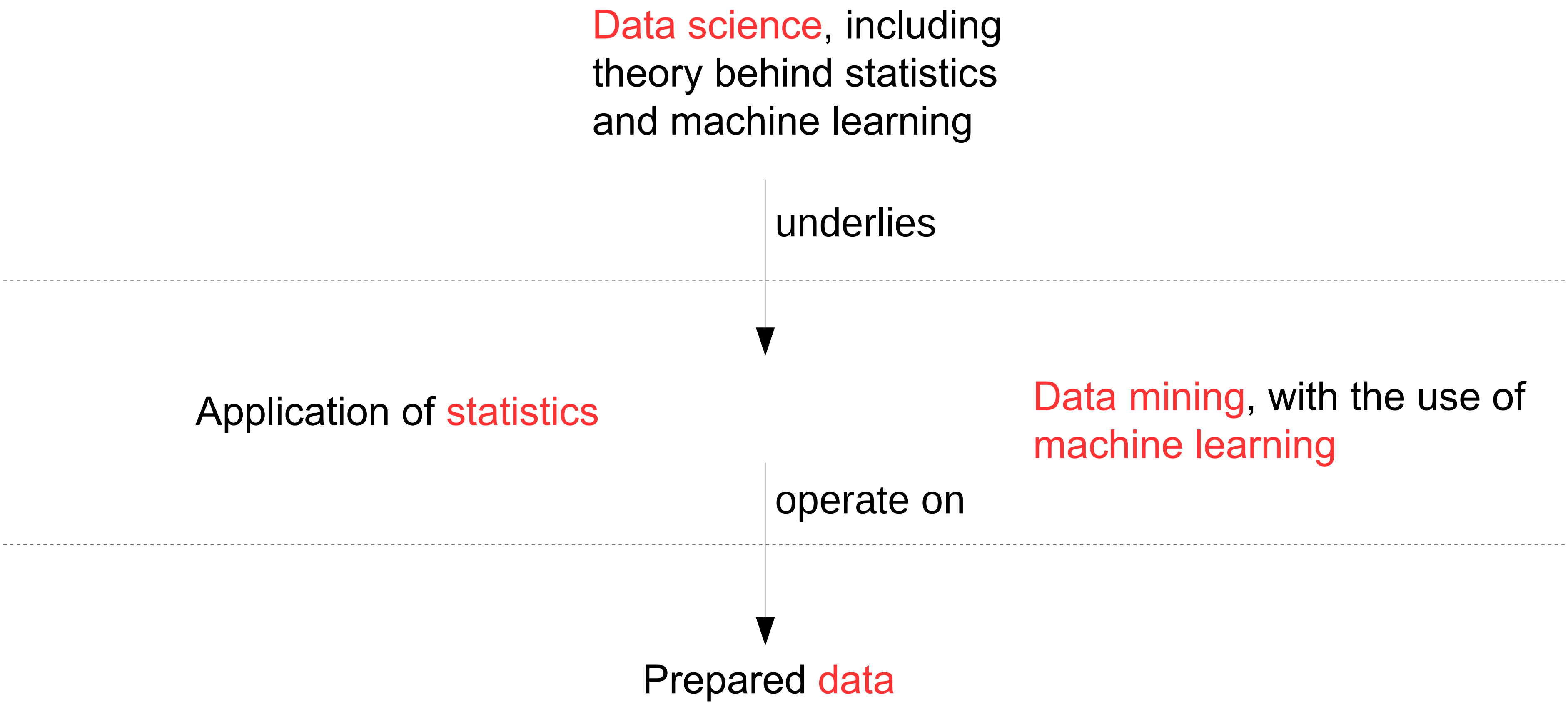
business (BI)

sports

population (census)

finance

The data analysis landscape



Data science, including theory behind statistics and machine learning

underlies

Application of statistics

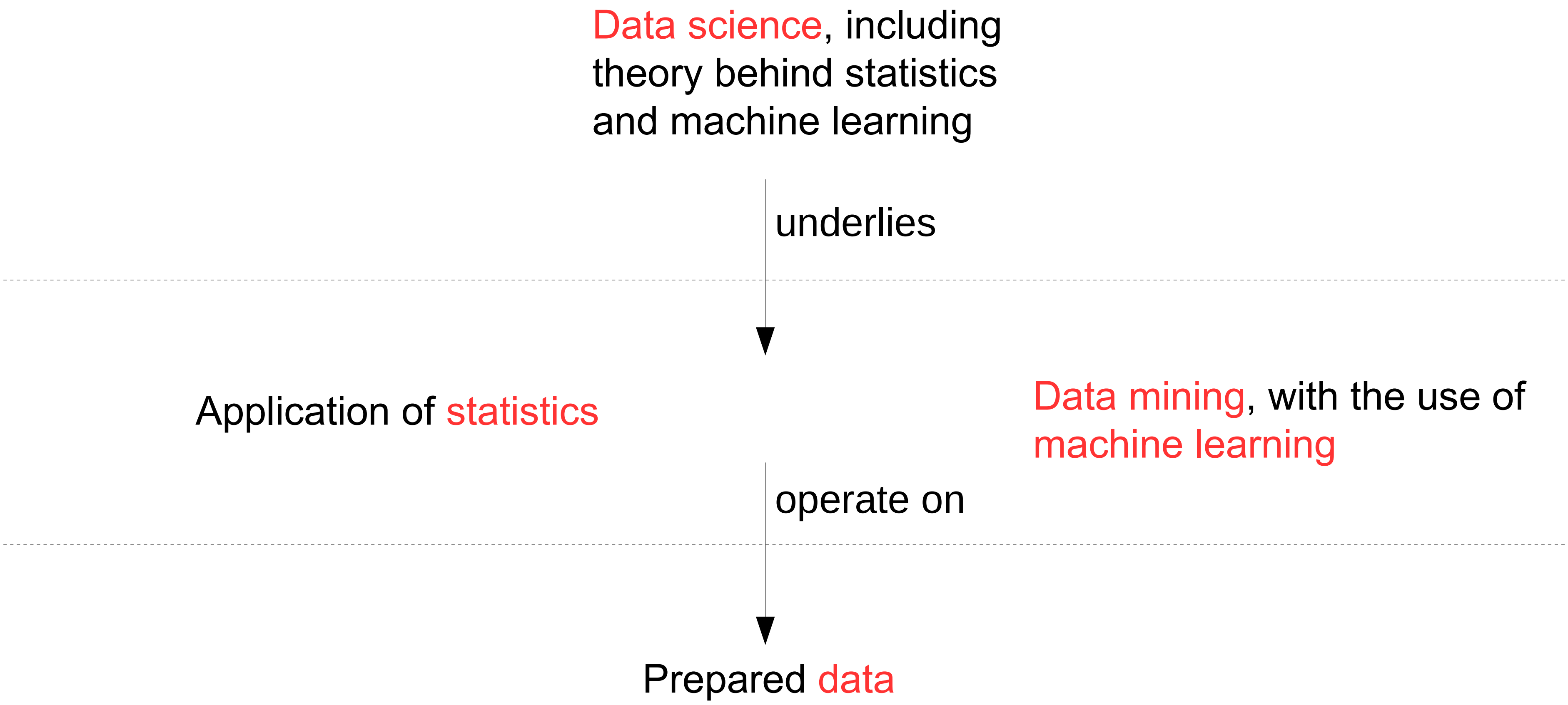
Data mining, with the use of machine learning

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 science, including theory behind statistics and machine learning

underlies

Application of statistics

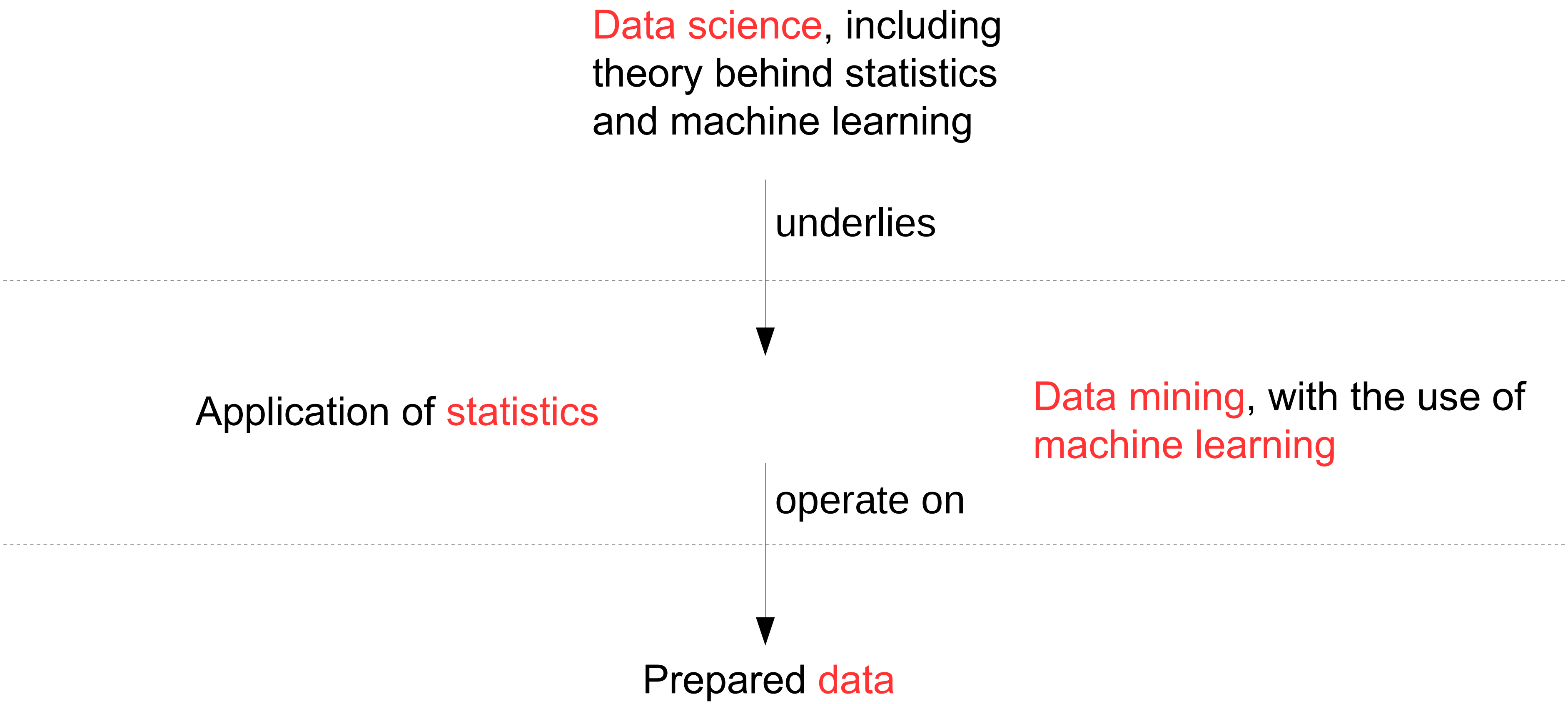
Data mining, with the use of machine learning

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

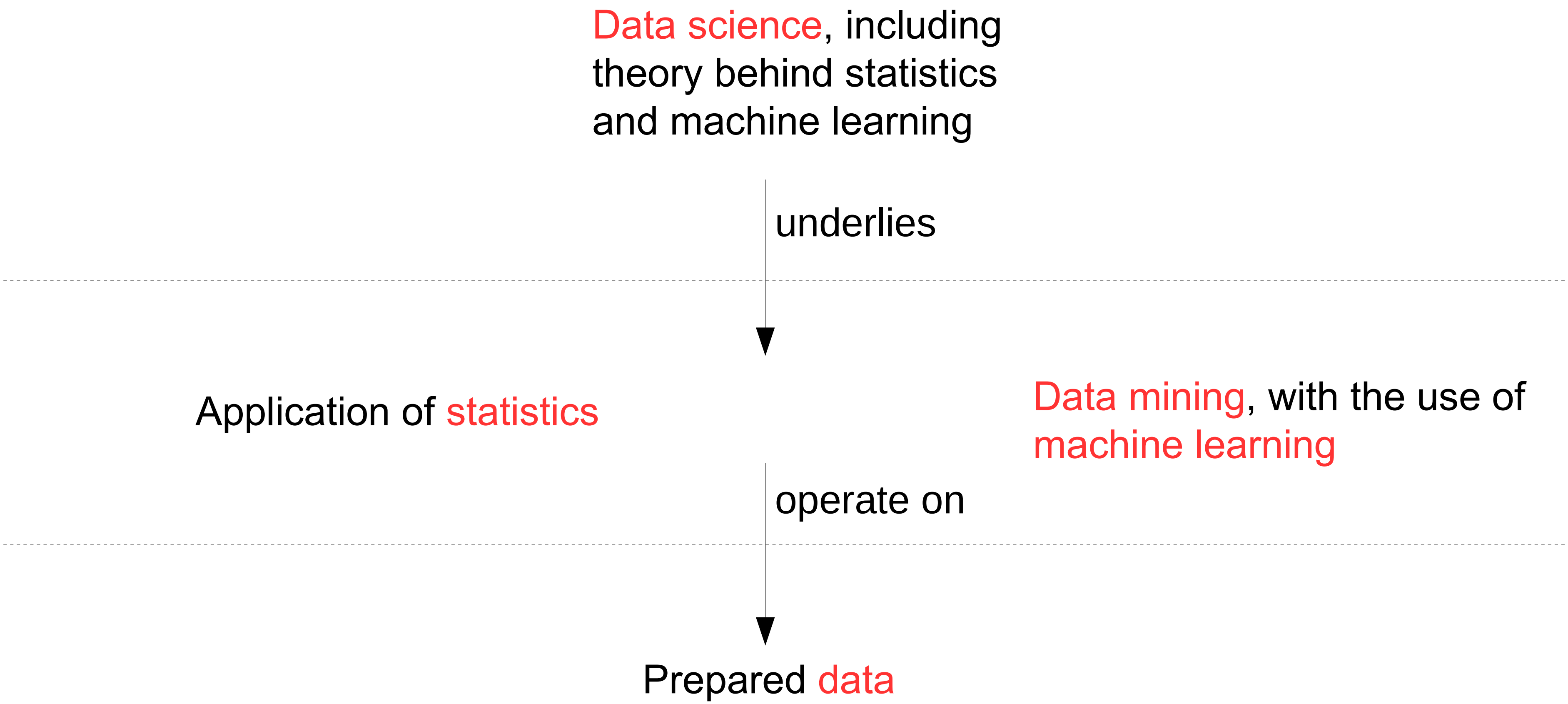
Data mining, with the use of machine learning

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

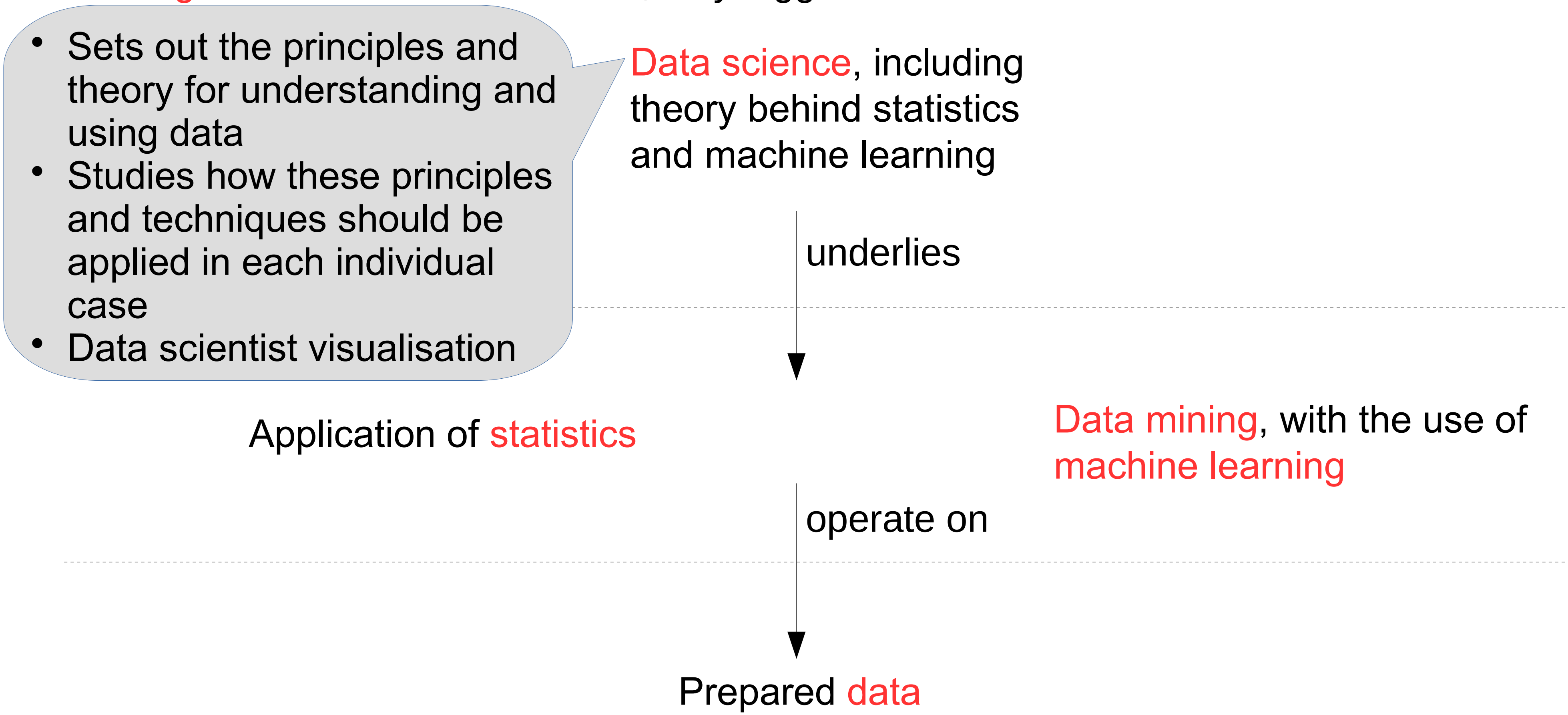
- In the last 20 years the data cycle is ‘intensifying’
- Growing processing power
- Almost limitless storage capacity
- Connectivity with large bandwidths
- Techniques have developed on this new wave of possibilities

- 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

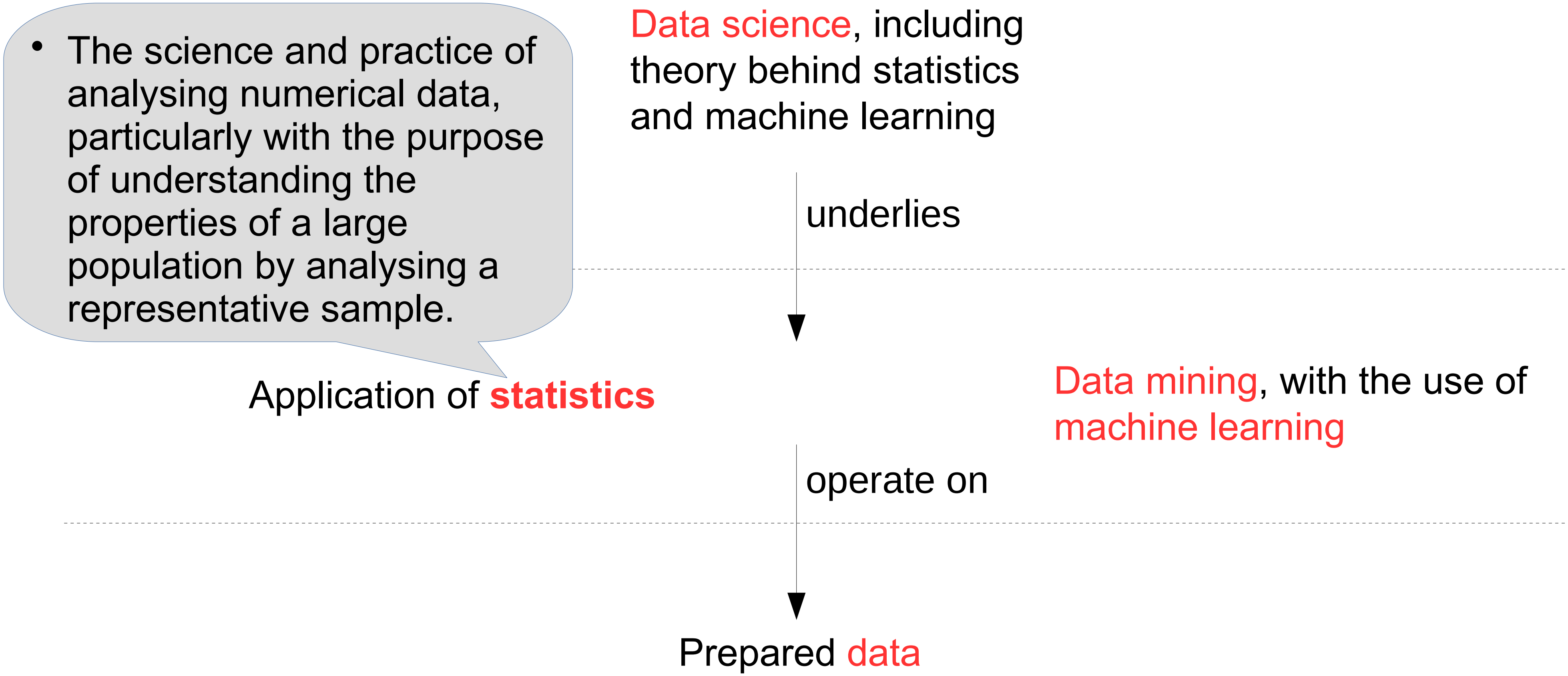
Data mining, with the use of machine learning

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**

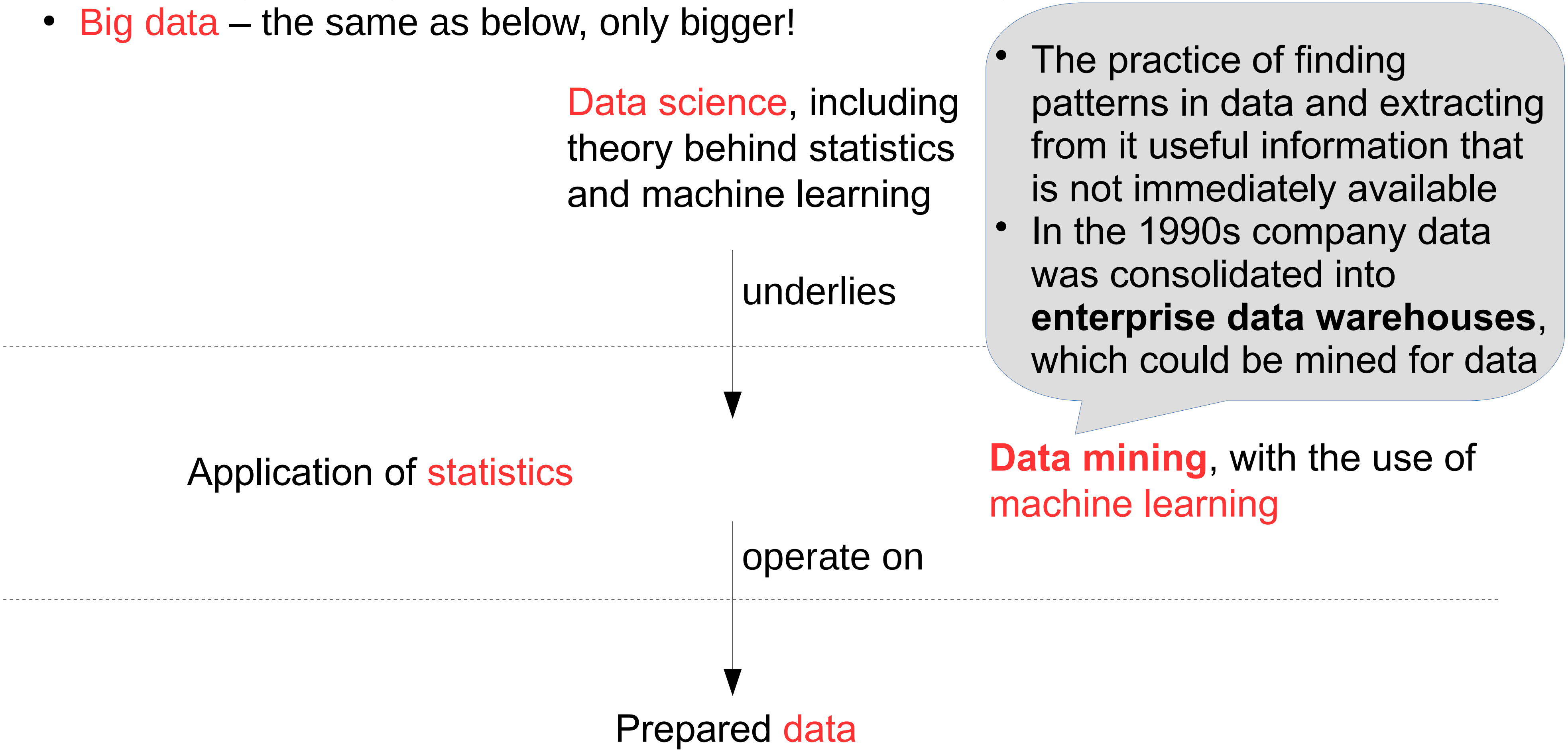
Data mining, with the use of machine learning

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

- 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

Application of statistics

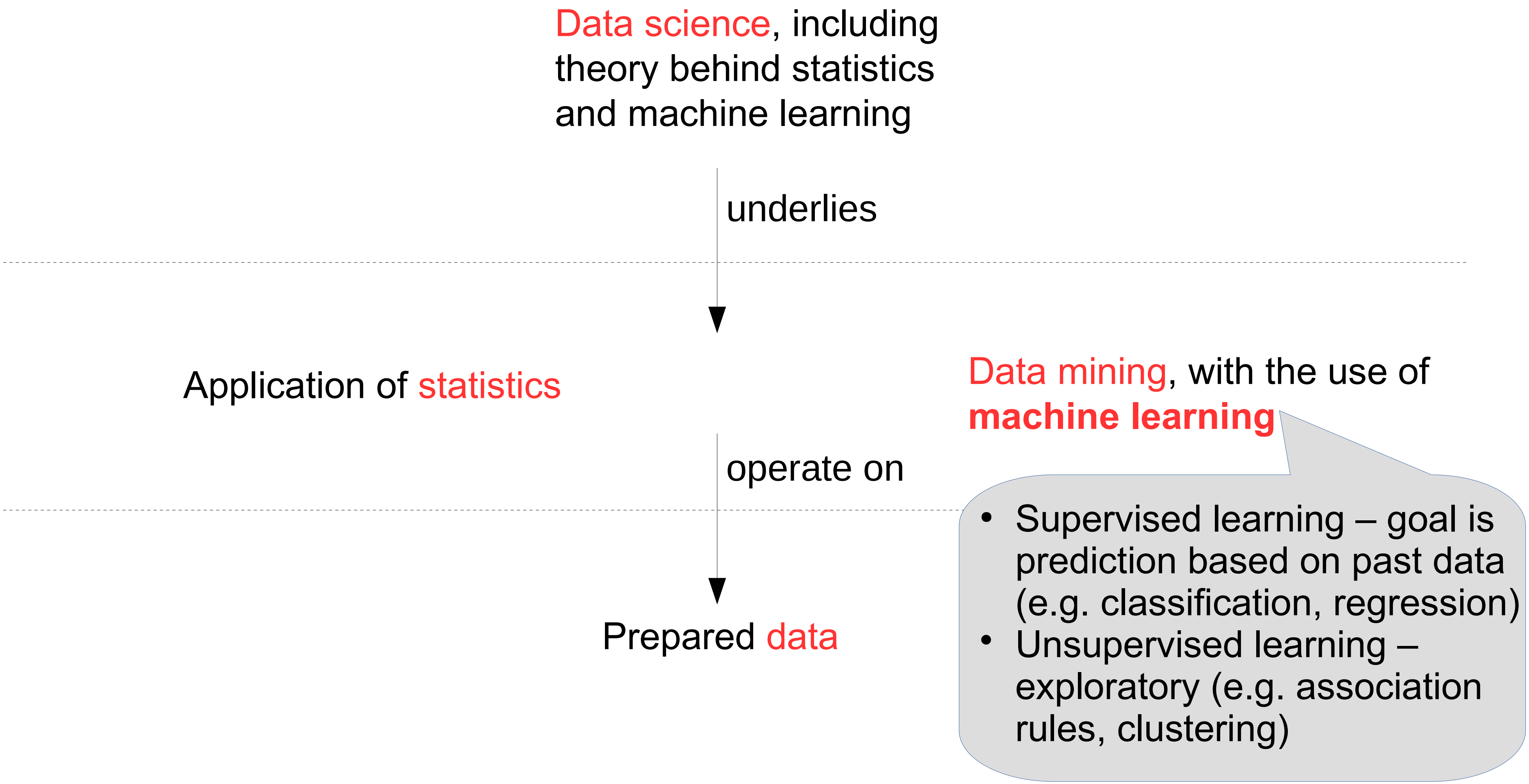
Data mining, with the use of machine learning

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**

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

operate on

Prepared data

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