What is Data Mining?

What is Data Mining?

“Data mining is the process of extracting hidden patterns from data.”

What is Data Mining?

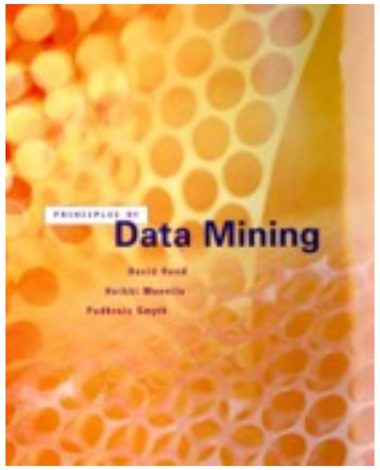
“Data mining is the process of extracting hidden patterns from data.”

“An Unethical Econometric practice of massaging and manipulating the data to obtain the desired results.”

What is Data Mining?

“Data mining is the process of extracting hidden patterns from data.”

“An Unethical Econometric practice of massaging and manipulating the data to obtain the desired results.”

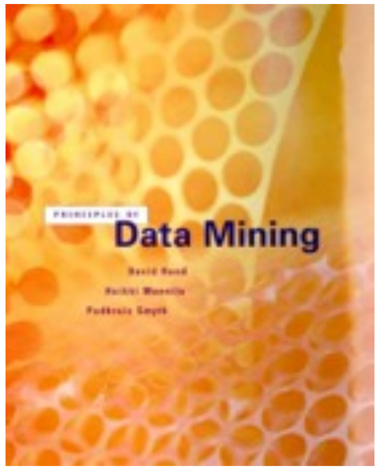


“Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.”

What is Data Mining?

“Data mining is the process of extracting hidden patterns from data.”

“An Unethical Econometric practice of massaging and manipulating the data to obtain the desired results.”



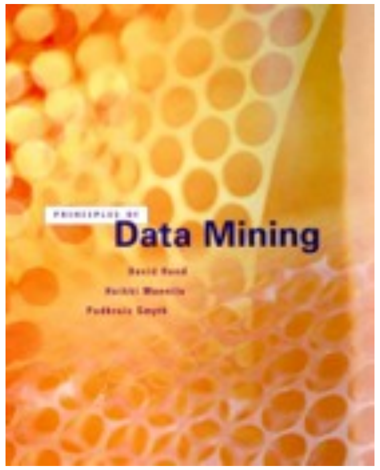
“Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.”

“Data mining, in a broad sense, is the set of techniques for analyzing and understanding data.”

What is Data Mining?

“Data mining is the process of **extracting hidden patterns** from data.”

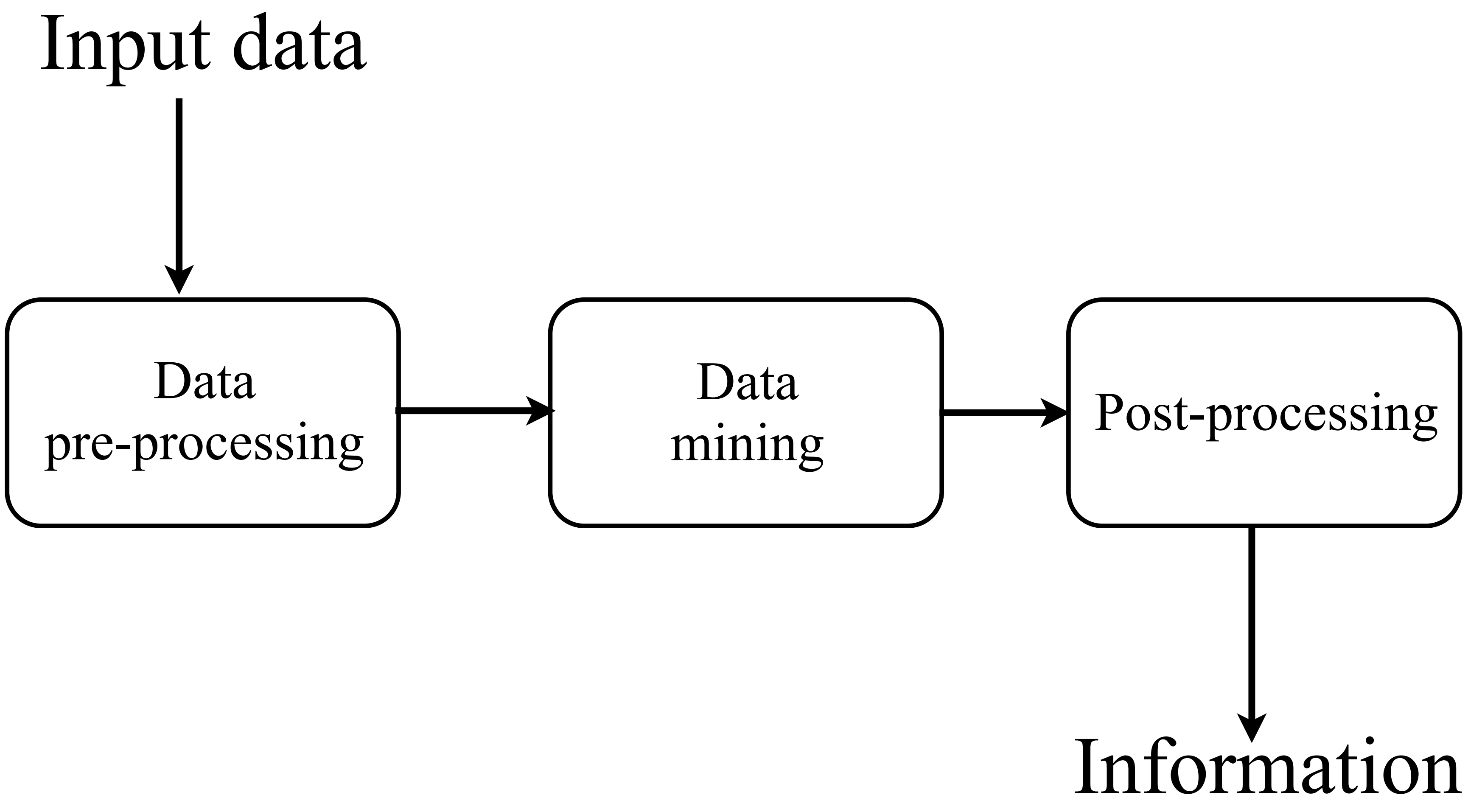
~~“An Unethical Econometric practice of massaging and manipulating the data to obtain the desired results.”~~



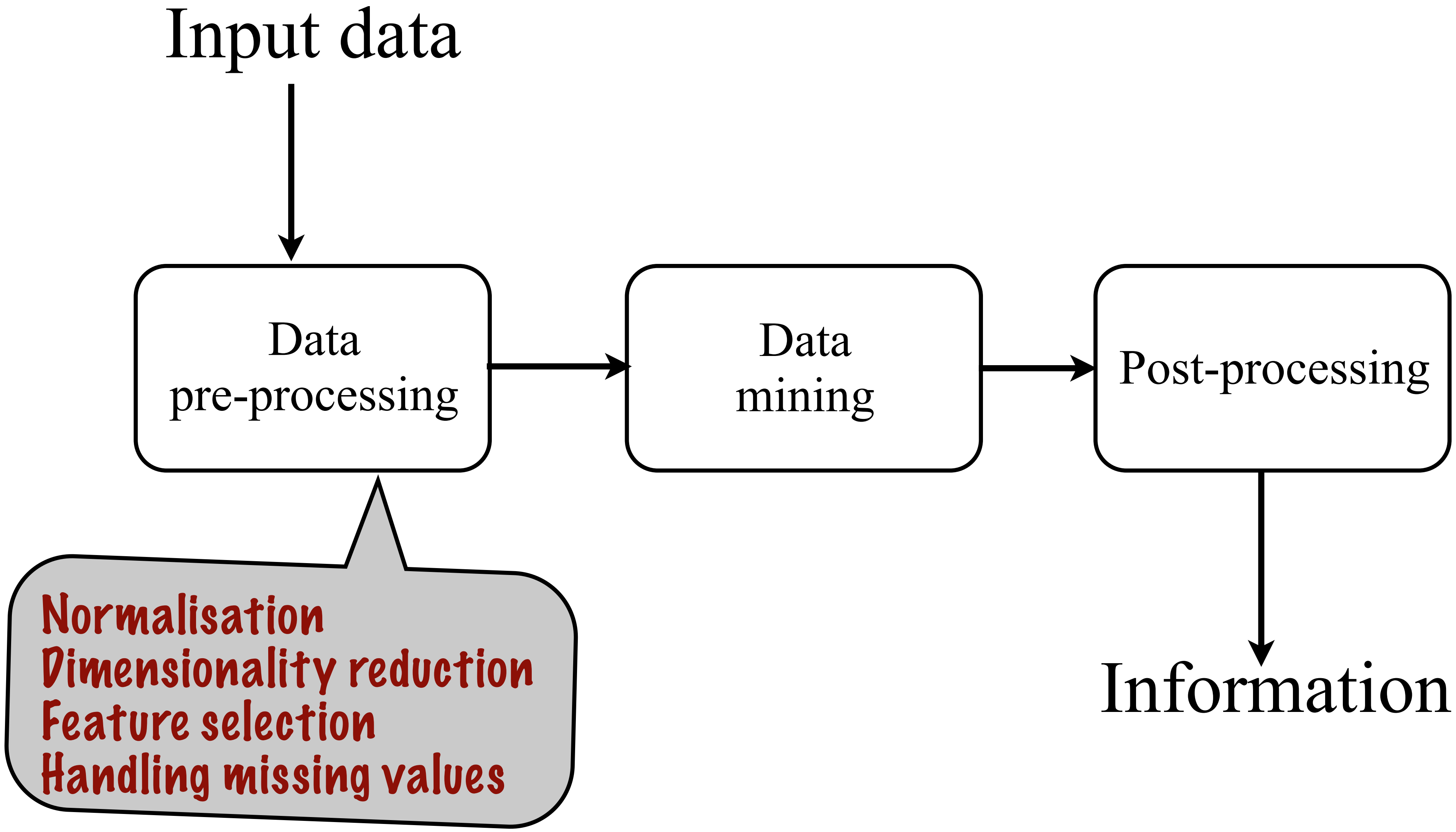
“Data mining is the **analysis** of (often large) observational data sets to find **unsuspected relationships** and to **summarize the data** in novel ways that are both **understandable** and **useful** to the data owner.”

“Data mining, in a broad sense, is the set of techniques for **analyzing** and **understanding** data.”

The KDD process

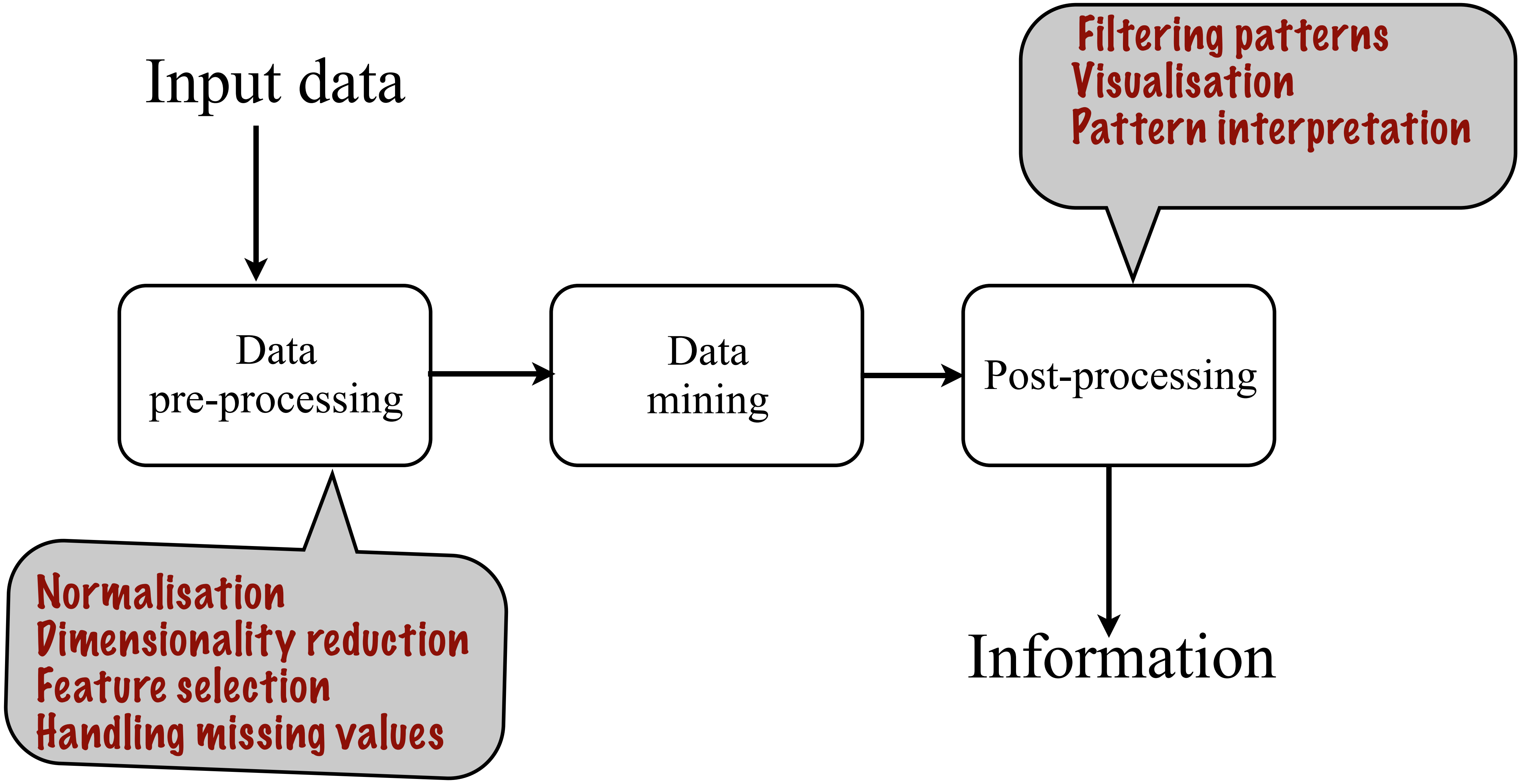


The KDD process



Normalisation
Dimensionality reduction
Feature selection
Handling missing values

The KDD process



Input data

Data pre-processing

Data mining

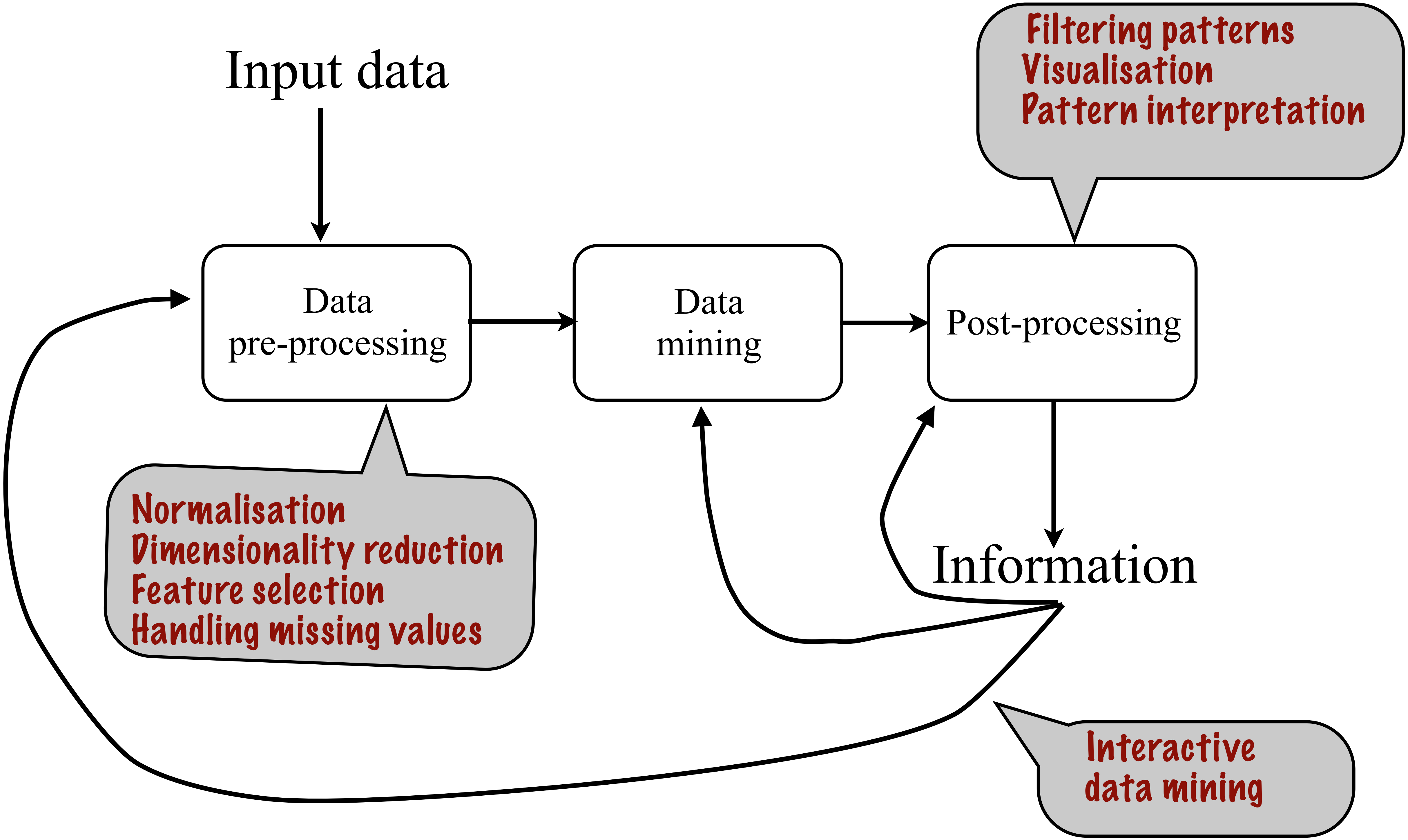
Post-processing

Information

Filtering patterns
Visualisation
Pattern interpretation

Normalisation
Dimensionality reduction
Feature selection
Handling missing values

The KDD process



Input data

Data pre-processing

Data mining

Post-processing

Information

Normalisation
Dimensionality reduction
Feature selection
Handling missing values

Filtering patterns
Visualisation
Pattern interpretation

Interactive data mining

Data Mining vs. Information Retrieval

- IR is answering questions user asked
- DM is answering questions user **didn’t** ask
  - “Show me web pages relevant to this query”
  - “Show me interesting patterns in these web pages’ contents”
    - Vague problem, how to evaluate the results?
    - Typical approach: pre-define some measurable quality of “interestingness”

Data mining's position in sciences

- Data mining uses statistical tools and methods to infer from data
  - Is data mining just fancy name for statistics?
- Data mining uses methods to learn unseen
  - Is data mining just boring name for machine learning?

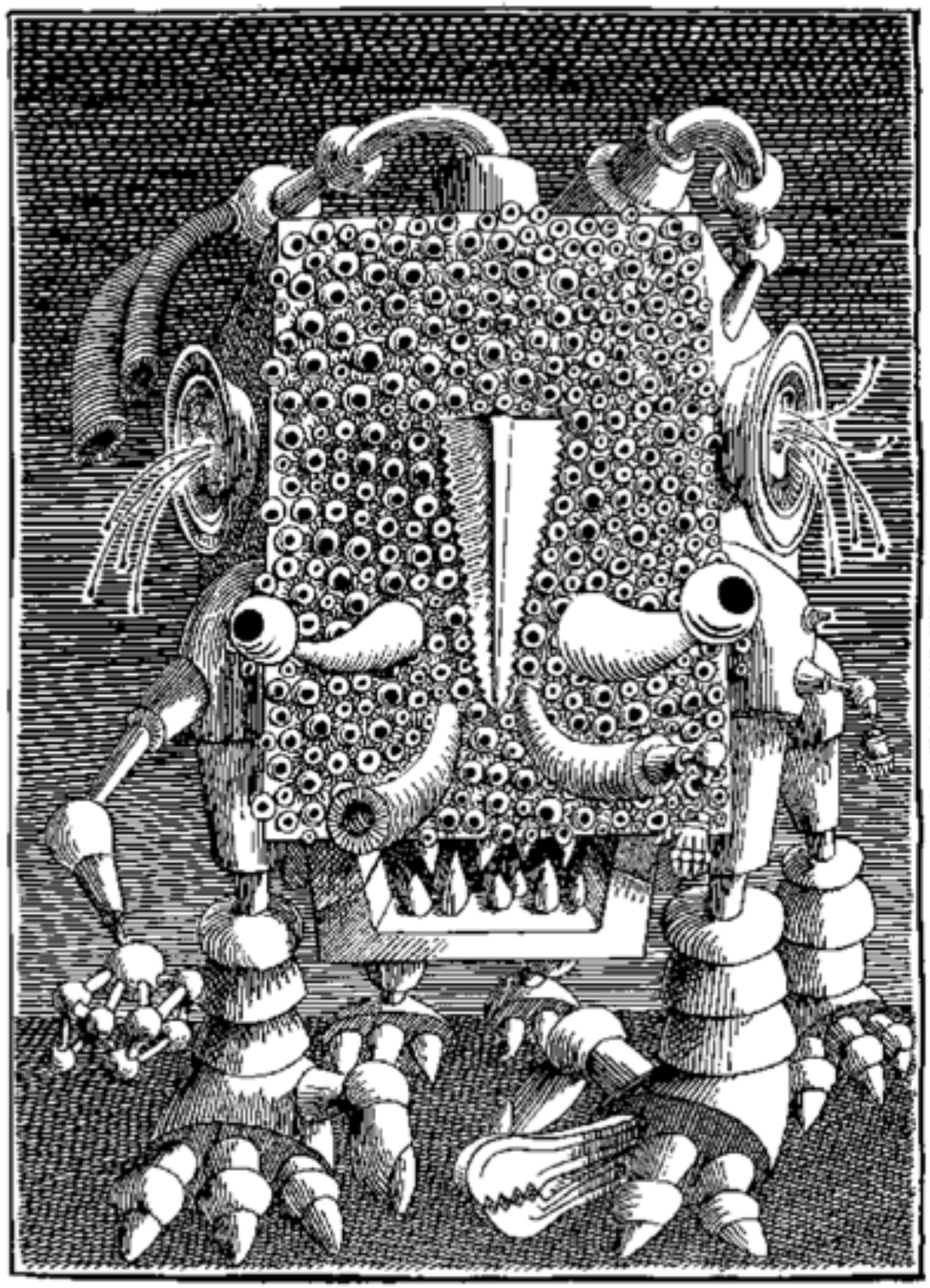
Data mining's position in sciences

- Data mining uses statistical tools and methods to infer from data
  - Is data mining just fancy name for statistics?
- Data mining uses methods to learn unseen
  - Is data mining just boring name for machine learning?

Is data mining a voodoo science?

Why Data Mining?

Why Data Mining?

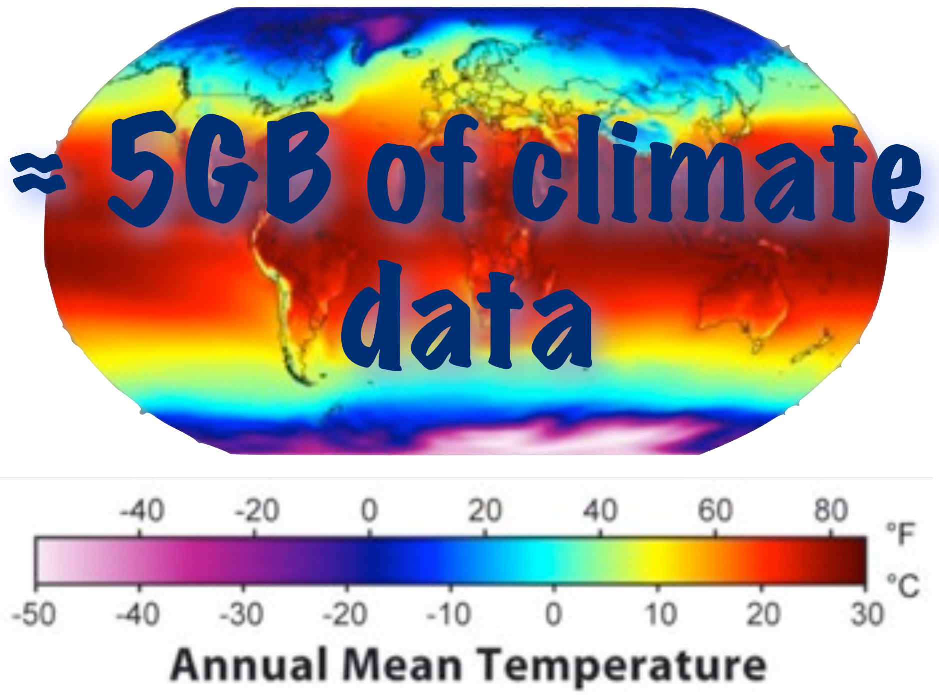


The ”PHT” Pirate wanted all information of the world. But before he realized most of it was useless, he was already buried under it.

—Stanisław Lem, *The Cyberiad*

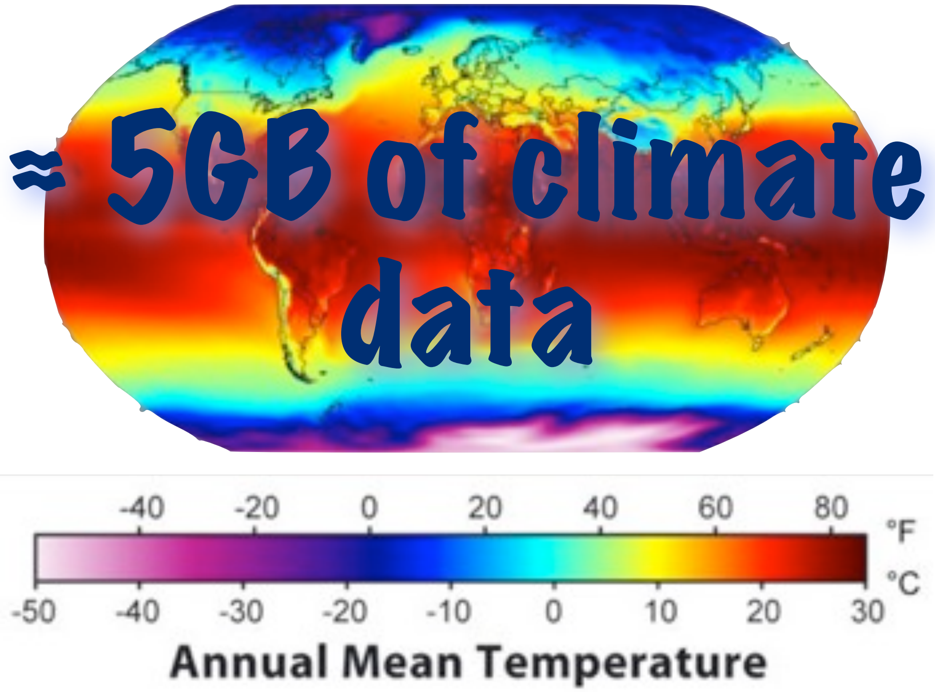
Data, data, data, data, …

Data, data, data, data, …



≈ 5GB of climate data

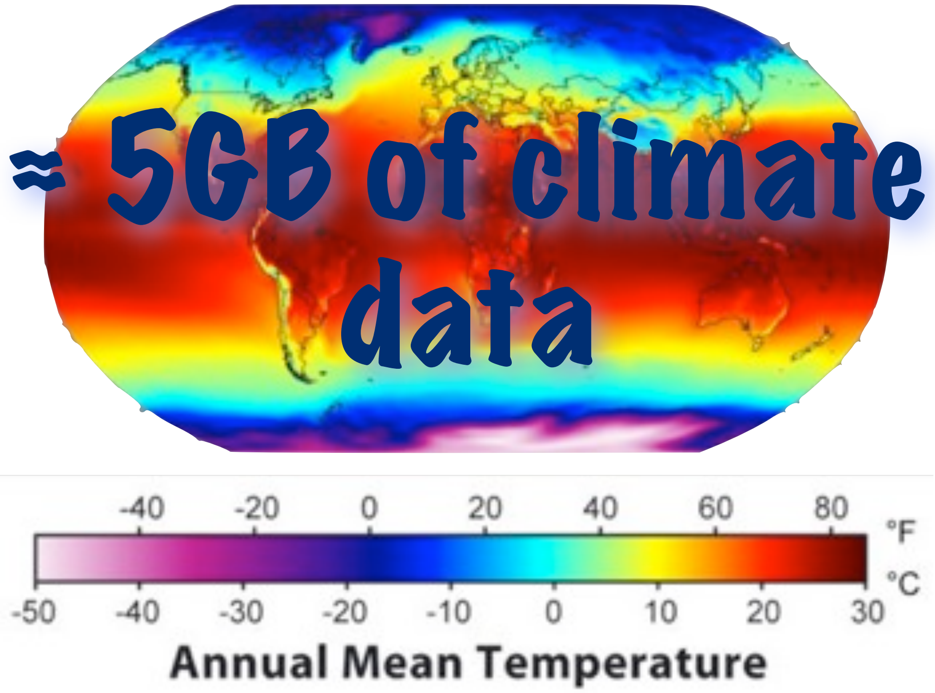
Data, data, data, data, …



≈ 5GB of climate data

40 000 000 000 photos

Data, data, data, data, …



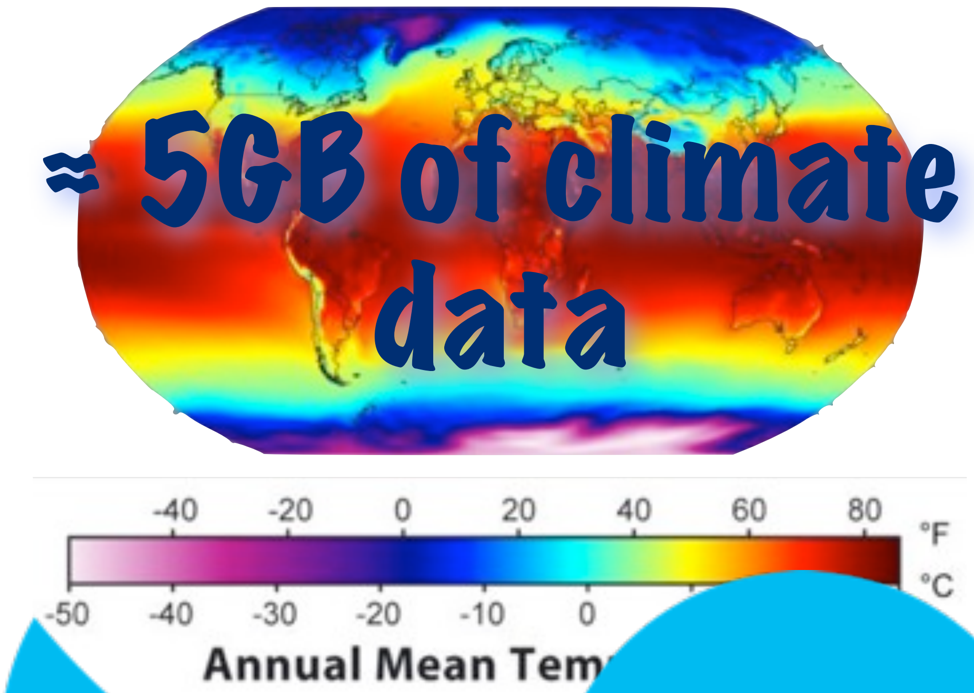
≈ 5GB of climate data

Walmart

1 000 000 customer transactions per hour

40 000 000 000 photos

Data, data, data, data, …



≈ 5GB of climate data

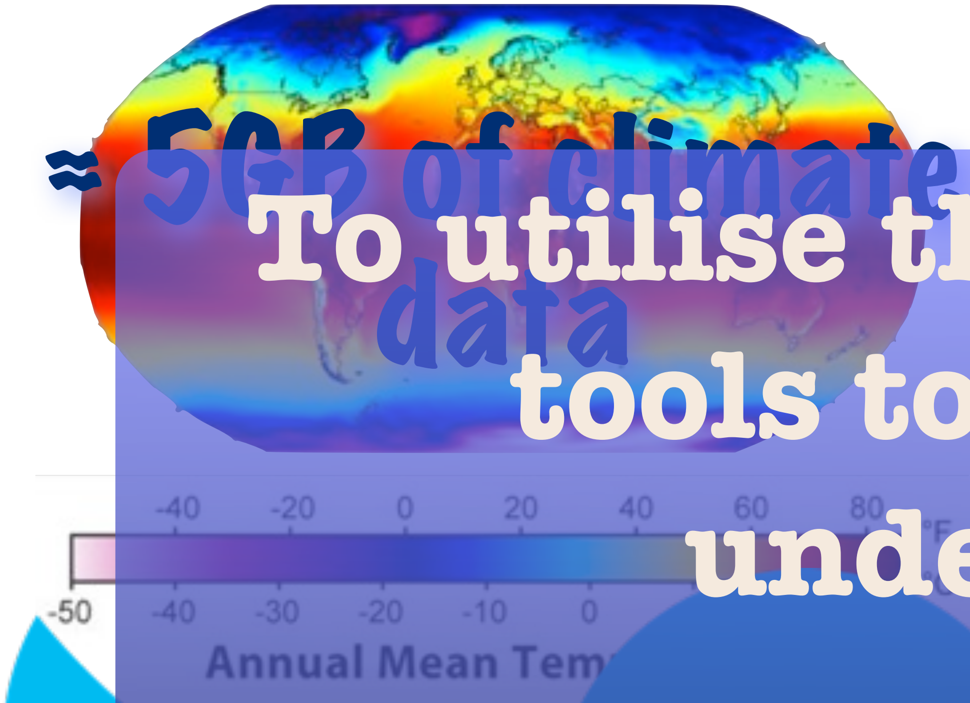
Walmart

1 000 000 customer transactions per hour

340 000 000 tweets per day

40 000 000 000 photos

Data, data, data, data, …



≈ 5GB of climate data

1 000 000 customer transactions per hour

340 000 000 tweets per day

40 000 000 000 photos

To utilise this data, we need tools to analyse and understand it.

We need data mining.

Data Mining Applications

Data Mining Applications

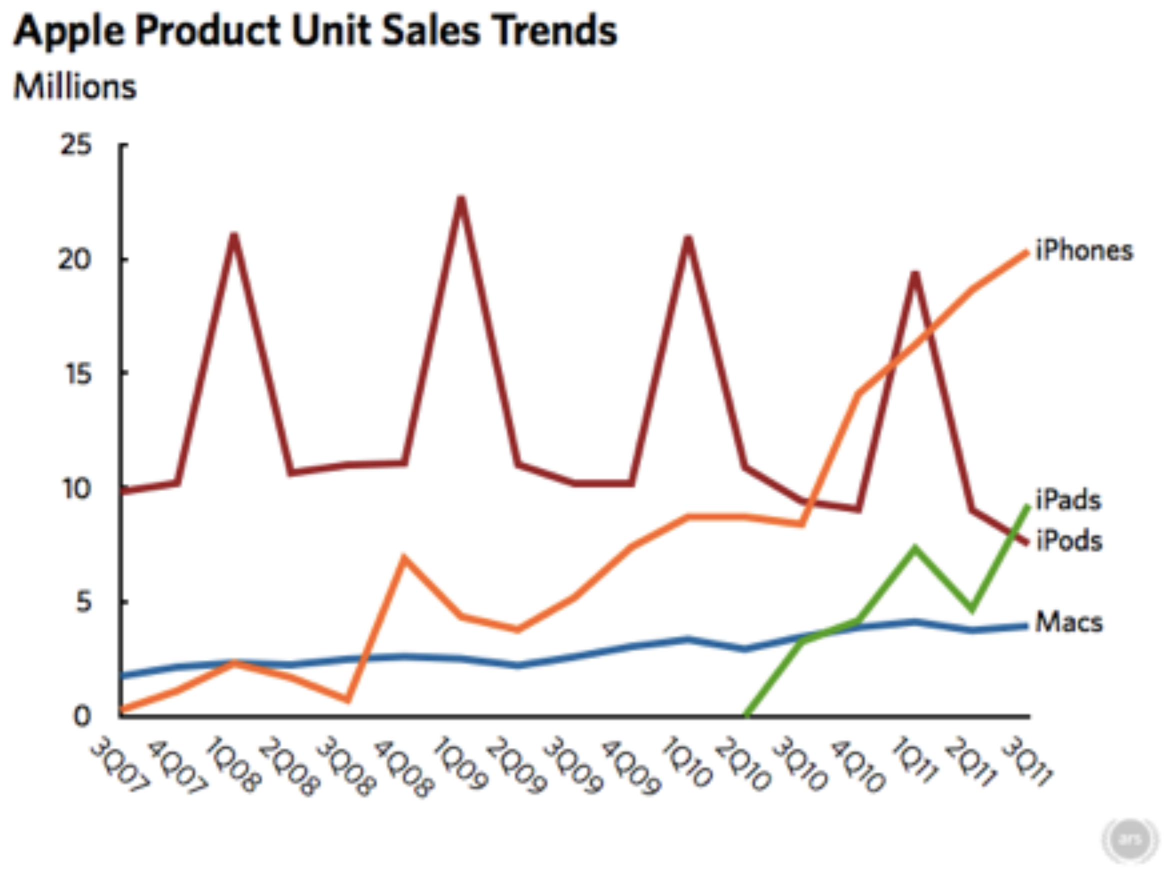
- Business intelligence

Data Mining Applications

- Business intelligence
  - What customers buy together?

Data Mining Applications

- Business intelligence
  - What customers buy together?
  - What are the seasonal trends?



Data Mining Applications

- Business intelligence
  - What customers buy together?
  - What are the seasonal trends?
  - How to make more money?

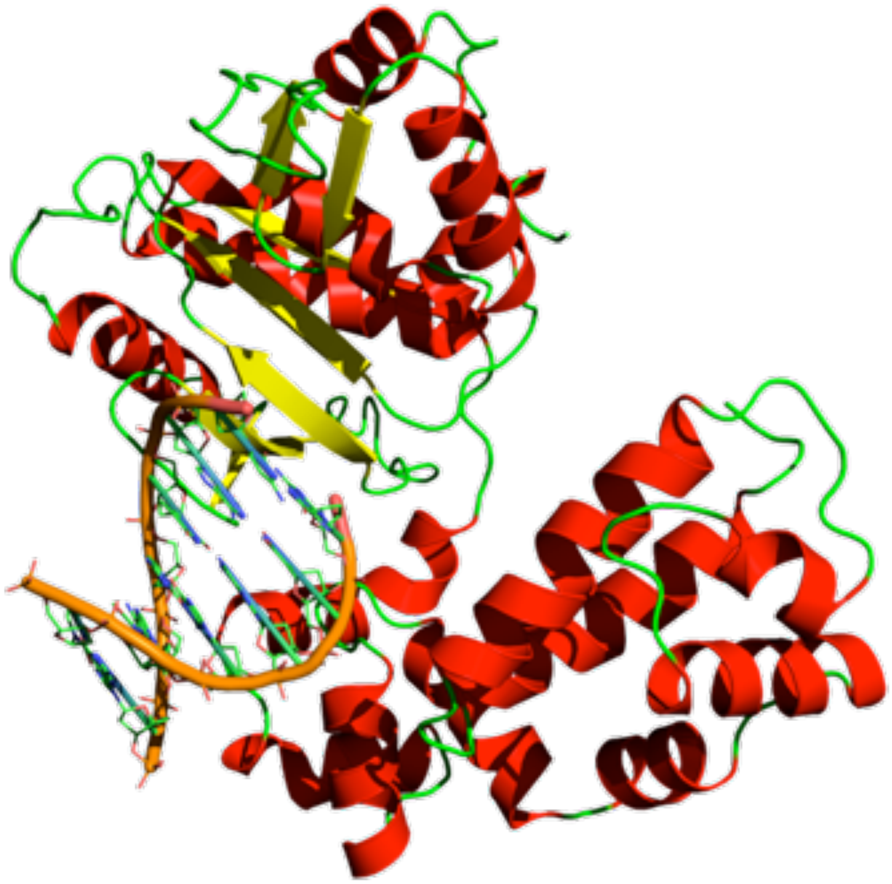
$?

Data Mining Applications

- Business intelligence
  - What customers buy together?
  - What are the seasonal trends?
  - How to make more money?
- Scientific data analysis

Data Mining Applications

- Business intelligence
  - What customers buy together?
  - What are the seasonal trends?
  - How to make more money?
- Scientific data analysis
  - What genes cause diseases?

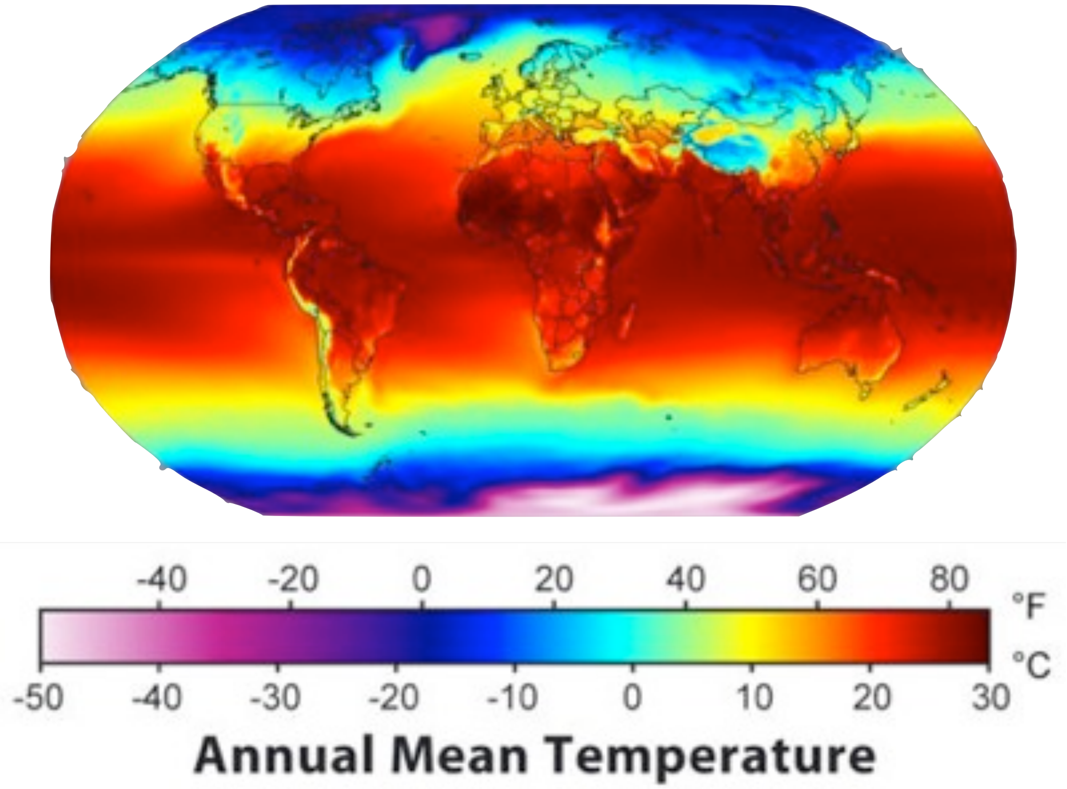


Data Mining Applications

- Business intelligence
  - What customers buy together?
  - What are the seasonal trends?
  - How to make more money?
- Scientific data analysis
  - What genes cause diseases?
  - What species co-inhabit areas?

Data Mining Applications

- Business intelligence
  - What customers buy together?
  - What are the seasonal trends?
  - How to make more money?
- Scientific data analysis
  - What genes cause diseases?
  - What species co-inhabit areas?
  - What happens if average temperature raises?



Data Mining Applications

- Business intelligence
  - What customers buy together?
  - What are the seasonal trends?
  - How to make more money?
- Scientific data analysis
  - What genes cause diseases?
  - What species co-inhabit areas?
  - What happens if average temperature raises?
- And anything else where you have data…

Data Mining Applications

- Business intelligence
  - What customers buy together?
  - What are the seasonal trends?
  - How to make more money?
- Scientific data analysis
  - What genes cause diseases?
  - What species co-inhabit areas?
  - What happens if average temperature raises?
- And anything else where you have data…
  - Who Barack Obama had to persuade to vote him?

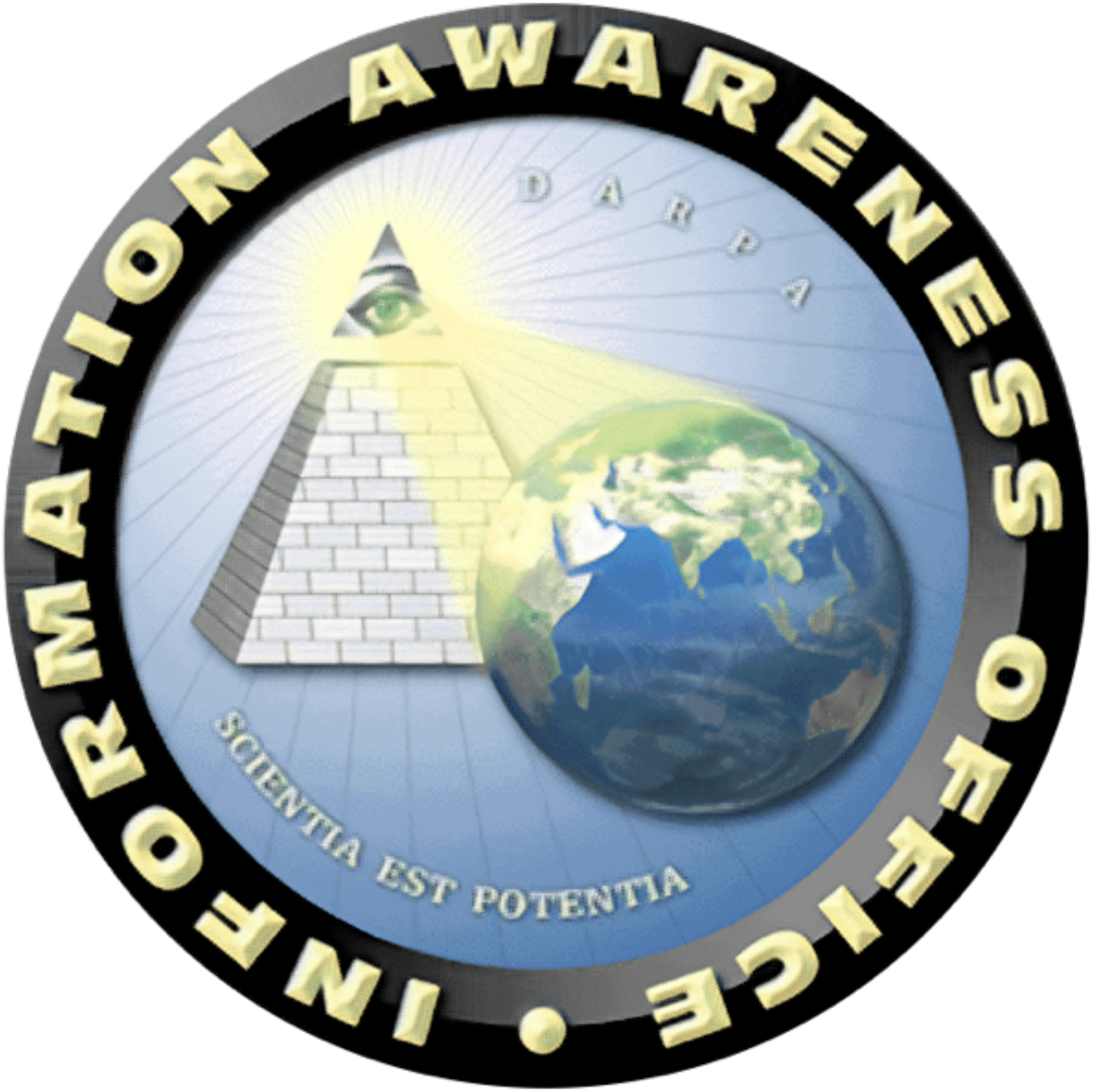
Data Mining Applications

- Business intelligence
  - What customers buy together?
  - What are the seasonal trends?
  - How to make more money?
- Scientific data analysis
  - What genes cause diseases?
  - What species co-inhabit areas?
  - What happens if average temperature raises?
- And anything else where you have data…
  - Who Barack Obama had to persuade to vote him?
  - Is there a problem in the International Space Station?

National Security

National Security

National Security



Data mining and privacy

- Very important!
  - Everybody wants to do data mining, nobody wants to be data mined
  - Google's privacy policies, deleting your information from Facebook, EU and the right to be forgotten, …
- Often imposed by laws
  - Medical data, personal information records, …
- Governments use data mining techniques for national security
  - Profiling

Summary

- We’re collecting more and more data
  - Most of it is uninteresting—but how to find what’s interesting
- Scientific method: form hypothesis, collect data, test hypothesis
- DM approach: collect data and let the computer find what (interesting) hypotheses hold in it