The need for data analytics

→ can be used to gain competitive advantage

→ using insights to adjust performance

→ e.g. moneyball, formula 1

→ AI, deep learning, machine learning etc.

→ recommend, predict, determine, understand.

data analytics can help if:

correct data is available

and the problem can be defined.

The age of big data

Data is the (i) quantities, characters or symbols on which operations are performed by a computer.

(ii) which may be stored and transmitted in the form of electrical signals,

(iii) and recorded on magnetic, optical, or mechanical rerording material

Raw/unprocessed data = a collection of numbers and characters, which needs to be cleaned in order to be provided as input to data analysis tools

→ contain imperfections that first need to be corrected

The **rational data model** → organizes data into tables of columns & rows with a unique key for each row

→ data related by common field

→ extract data using **queries** = request for data

→ structured query language (SQL)

↳ int. standard for defining database queries

↳ relationships help enforce accuracy by reducing duplication

the **database** = a collection of information that is organized
so that it can be easily **accessed**, = data can be read
managed and **updated**. = data can be added or updated

↳ data are organized into rows, columns & tables,
using some formal **data scheme**
↳ structure of the database described
in formal language supported by the
database management system.

DATA WAREHOUSES

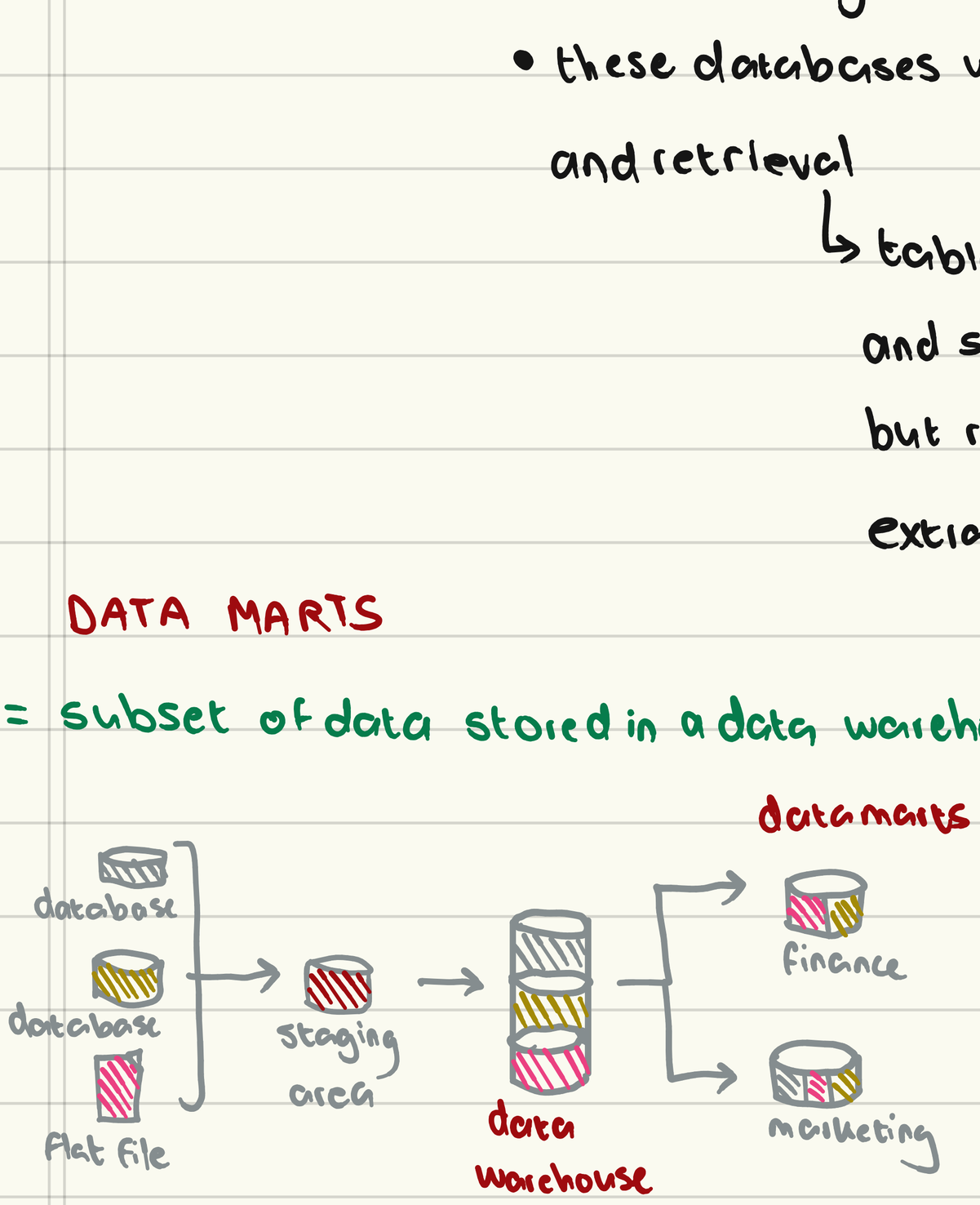
= central repositories of integrated data that stores current
and historical data in one place

→ optimised for analytics

↳ previous model: • data stored in numerous separate databases
within one organization
• these databases were optimised for storage
and retrieval
↳ tables with minimum redundancy
and smallest no. of fields make updates fast,
but requires multiple operations to
extract useful analytics.

DATA MARTS

= subset of data stored in a data warehouse orientated for a specific business need.



DATA WAREHOUSES VS DATA MARTS

same: designed for structured data = clearly defined data types.: easily searchable

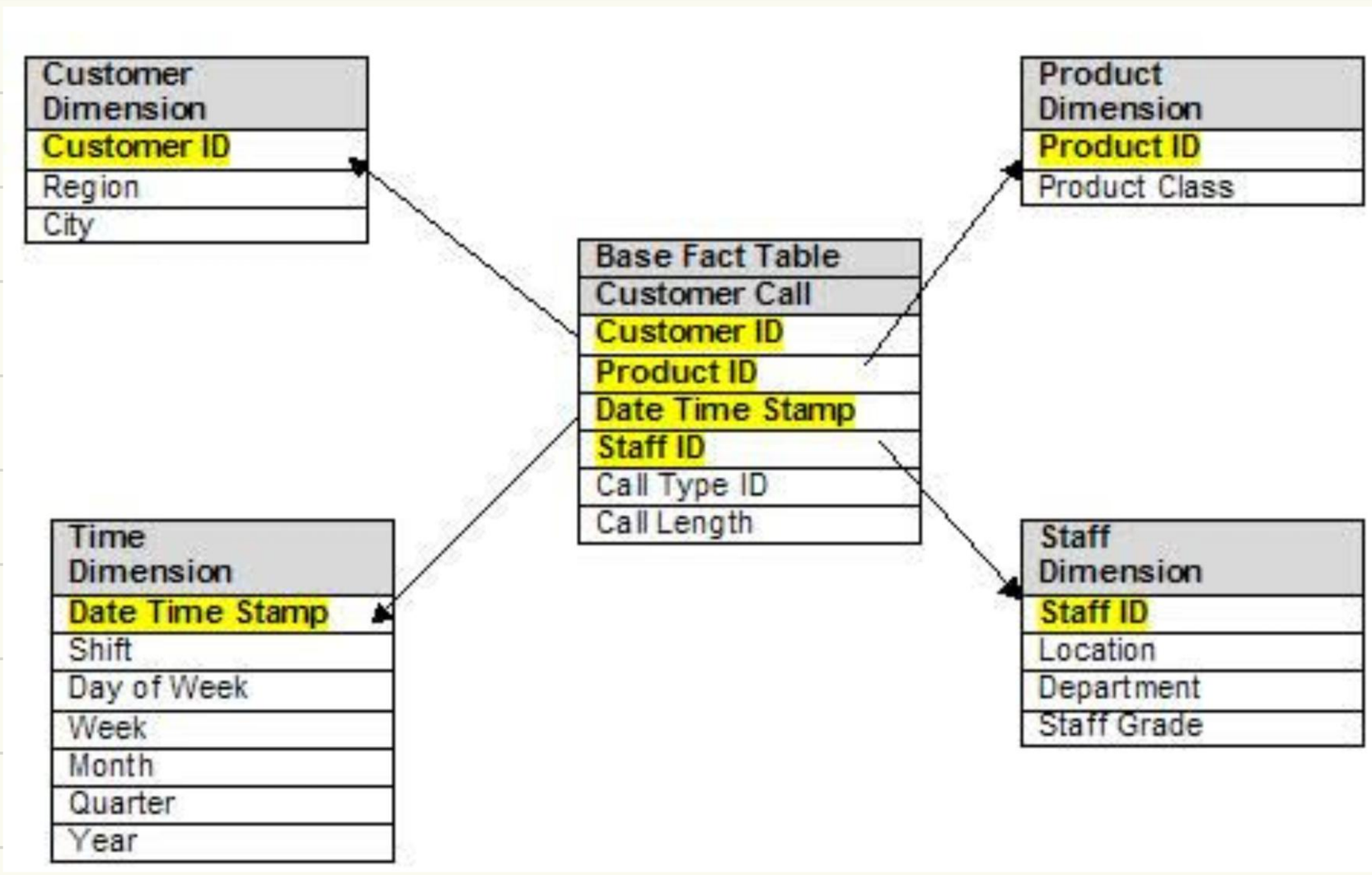
and speed of access = read-only

differences	DATA WAREHOUSE	DATA MART
Stores data from	multiple subject areas	one subject area
contains	detailed info	detailed info + summaries
integrates	all data sources	only info for a given subject
uses a dimensional model	not necessarily	yes, e.g. star schema, snowflake

☆ THE STAR SCHEMA

consists of ① a large central table (fact table) → store observations/events.
containing the bulk of the data,
with no redundancy,

and ② a set of smaller attendant tables (dimension tables),
one for each dimension



BIG DATA characterised by...

1. **Volume** → size of generated and stored data (e.g. GB, TB)
2. **Variety** → type and nature of data in in the original unstructured (= to be defined) raw forms.
   → types: seq + time data, data streams, engineering design data, multimedia data, graph and network data.

→ new challenges: how to store + analyse data.

characterise data:

Structured data ● **Semi-structured data** ● **Unstructured data** →

ID	Name	Age	Degree
1000	Bruce	19	Communication
1002	Kamala	18	Accounting
1003	T'Challa	23	Psychology

organized, easily searchable

```
<University>
 <ID = "1000">
  <Name = "Bruce">
  <Age = "19">
  <Program = "Communication">
 <ID = "1000">
  <Name = "Bruce">
  <Age = "19">
  ....
```

The university has 5600 students. Bruce (ID: 1000), 19 years old Communication study. Kamala with ID 1002, majoring in Accounting and is 18 years old. T'Challa from Psychology study program, 23 years old, ID 1003

↳ lacks predefined model - not easily searchable

↳ more common, but older tech not designed for it.

NoSQL database:

→ stores data as objects with attributes
→ using object notation language (JSON)
→ vs rational table-based model

advantage: set of attributes of each object is encapsulated within the object
→ individual objects can have diff attributes

Relational database example

ID	Name	Home	Cell phone
1000	Eleven	051 123 4567	082 111 2222
1002	Dustin	NULL	123 456 7891

NoSQL database example

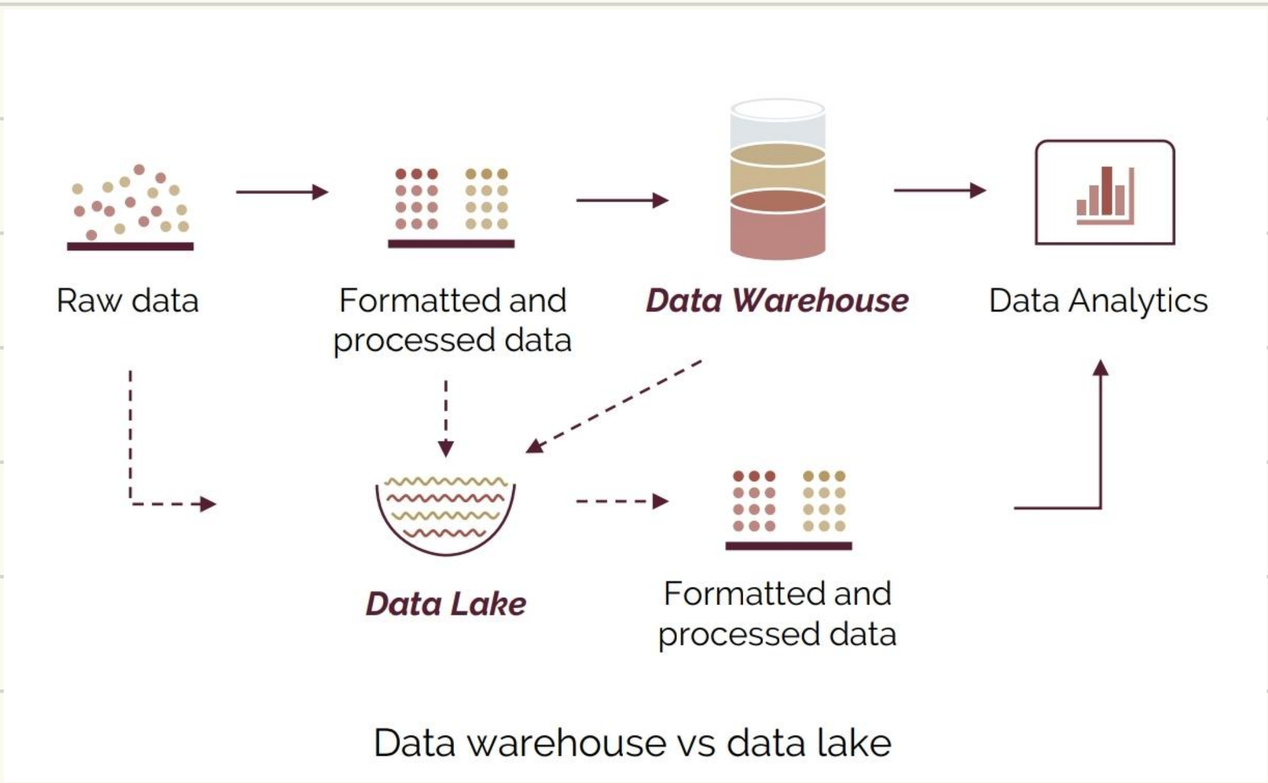
Key	Value
<ID = "1000">	<Name = "Eleven"> <Age = "19"> <Home = "051 123 4567"> <Cell = "082 111 2222">
<ID = "1002">	<Name = "Dustin"> <Age = "19"> <Cell = "123 456 7891">

Data lakes

→ a system / repository that can store raw data, as well as processed / transformed data - usually as object blobs or files

(object blobs → collection of binary data stored as a single entity)

→ can include:
- structured data from rational databases
- semi-structured data e.g. CSV, logs, XML, JSON
- unstructured data e.g. emails, documents
- binary data e.g. images, video, audio



Data warehouse vs data lake

③ **Velocity** → speed at which data is generated + processed (e.g. real-time)

stationary	non-stationary
Fixed data set	Data set not static
All data available for analysis	Data instances arrive continuously
Over time new data	Requires real-time model analysis
∴ analysis redone / model redeveloped	/ model adaption

Data analytics toolbox

inputs → process → outputs

Need: → data needs to be rather seen as an asset → reveal meaningful info
→ need to gain value from data

Formal def: ① process of extracting useful insights from raw data

② using an exploratory/data-driven approach

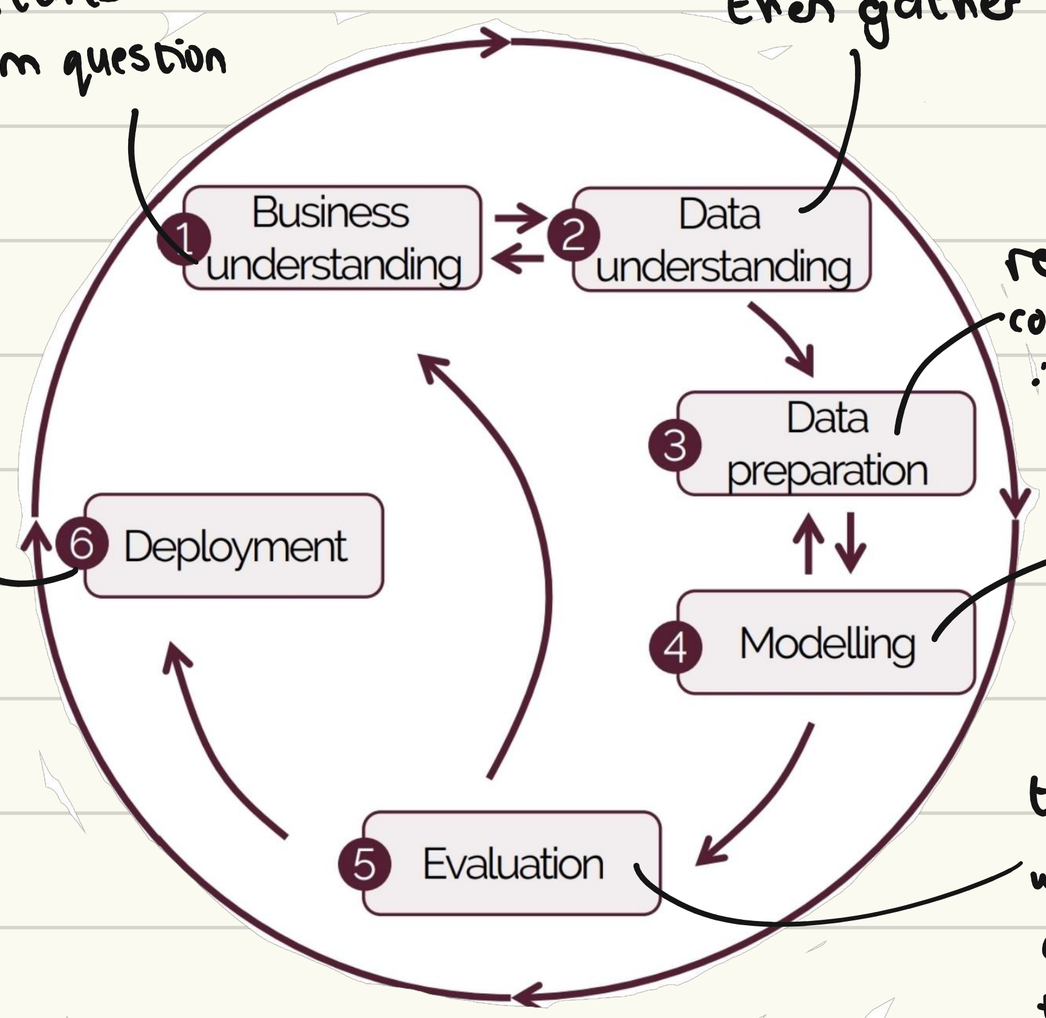
③ that shares similarities w/ data mining

↳ process, uses algorithms to discover hidden patterns + valuable info from small / very large databases.

CRISP-DM Cross Industry Process for Data Mining

data problem starts w/ business problem question

relevant data identified + collected, then gather insights



relevant data sources combined in table-like structure ∴ easier to do data analytics

one-off: present results

else: integrate model within a process in the organization

several models constructed, best model selected

tests performed to determine whether model performance conforms to quality requirements + business success criteria

Analytics types

PAST:
- descriptive analytics — what happened - understand past
- diagnostic analytics — why did it happen - issue tree - multiple analyses

FUTURE:
- predictive analytics — what is likely to happen - forecast
- prescriptive analytics — what can be done

Types of data analytics

→ BIG DATA ANALYTICS → scaling up dat.anely. techs to very large volumes of data

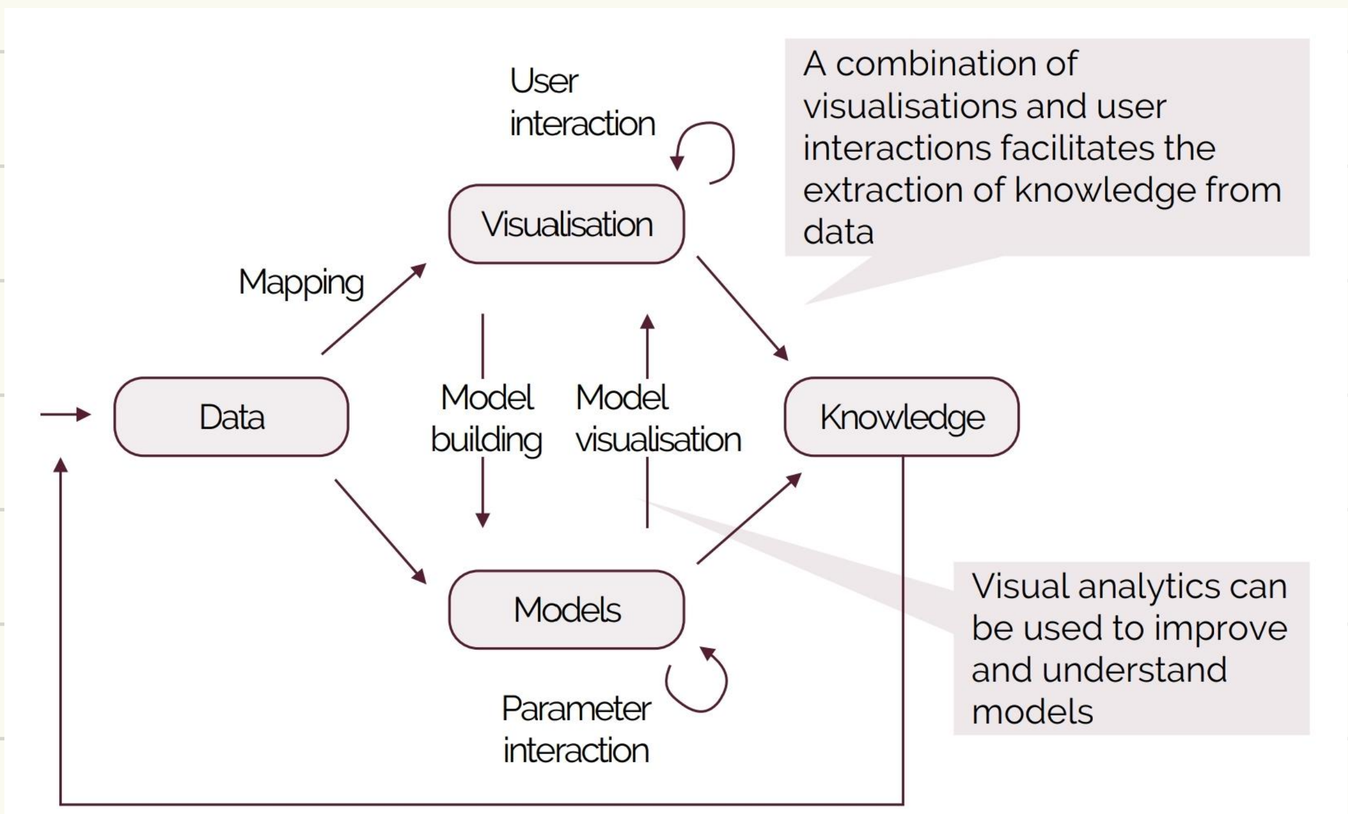
→ main issues:
- CPU time
- computational speed
- memory (prim. + sec)
- data streams
- data fusion of streams.

→ VISUAL ANALYTICS → analytical reasoning + interactive visual interfaces

→ automated analysis techniques + interactive visualization

for: effective understanding, reason, decision-making

→ process of mapping data attributes to visual properties



Abstract view of the different stages and their interactions in the visual analytics process

→ IMAGE ANALYTICS → automated extraction of meaningful information from images

→ predict + anticipate future events

↳ what objects present in image, where in world, what actions.

→ classification ⋯→ prediction ⋯→ post-estimation

→ challenges: images contain variance,

which algorithms need to be invariant to.

– viewpoint, scale, deformation, occlusion,

illumination, background clutter, intra-class.

→ VIDEO ANALYTICS → temporal image analytics, image analytics over time,
that focus on finding temporal patterns btw consecutive images.
→ applications: • dynamic masking
• emotion recognition over time
• object tracking
• behaviour prediction
• tracking movement patterns.

→ TIME SERIES ANALYTICS → identifying patterns from temporal data sequences
time series → observations of a quantity(ies) collected over time
→ finance, economics, engineering etc.

→ AUDIO ANALYTICS → analysing audio signals to extract information.
→ broader than detection ~ understanding
→ applications: • emotion detection
• sound identification
• speech recognition + understanding

→ TEXT ANALYTICS → uses linguistical, statistical, machine learning techniques
to convert unstruc. text into meaningful data for analysis.
→ challenges: • word order affects meaning
• context needed for prediction.
• sarcasm, misspellings, abbr., spelling variants...