Data Marketing Introduction

Session 1 - Understand Big Data

Introduction

Steven VINCENT - Burgundy Scool of Business-
Jan. 2025 -

Professional Experience

MSc Data Science &
Organizational Behavior
2016 - 2020

Data Engineer
PwC Luxembourg
06.2020 - 03.2022

Data Engineer
Société Générale
03.2022 - Today

The Different Data Roles

Data Analyst
In charge of building dashboards and analysis

ML Engineer
In charge of building machine learning pipelines

Data Engineer
In charge of developing Big data ETL data pipelines and data acquisition

BI Engineer
In charge of Data Warehouse, views, reports (Low Code ETL development)

Data Scientist
In charge of building predictions models and getting data insights

Data Architect
In charge of choosing and improving the data infrastructure

DataOps Engineer
In charge of data change management (CI/CD, Devops, release)

Session 1 - Syllabus

Understand Big Data

Add value to data

Introduction to Google Analytics

Google Analytics certification

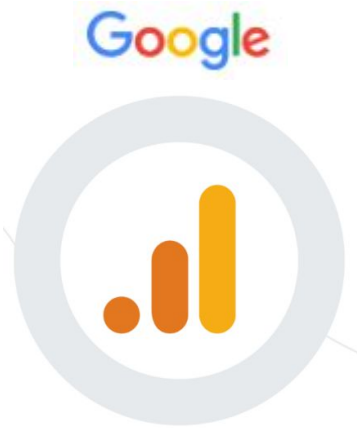
Steven VINCENT - Burgundy Scool of Business-
Jan. 2025 -

Certification Google Analytics

4 training session

Google Analytics online courses

GA4 training video



[Google Analytics Certification]

Exam info

How to get the certification ?

- Obtain a score with a minimum of 80% correct answers
- Do not exit the evaluation window before the end

Certification content ?

- 50 questions about Google Analytics
- Unable to go back to previous question
- Answer questions within the allotted time
- If you fail, you must wait 24 hours before retaking the assessment.

Assessment duration ?

- 75 minutes for 50 questions

Question type ?

- MCQ only

Answer all the questions!

Understand BIG DATA

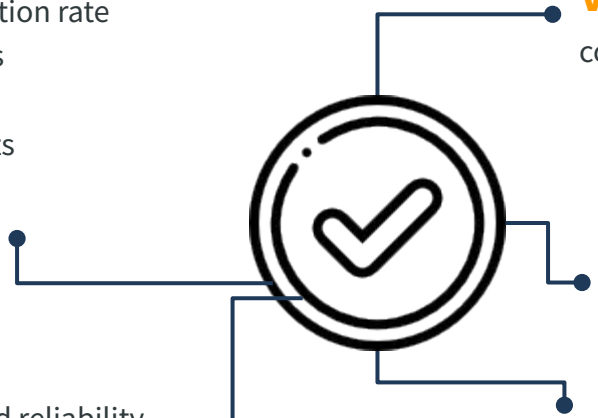
Steven VINCENT - Burgundy Scool of Business-
Jan. 2025 -

Definition and challenges of Big Data

Big Data : process of capturing, merging, and analyzing large and diverse data sets to understand current business practices and seek new opportunities to improve future performance.

Value : how businesses generate value :

- Improve customer retention rate
- Managing negative news
- Improving healthcare
- Reduce carbon footprints
- Create personalized ads



Volume : the considerable amount of data collected in “big data” systems

Velocity : the pace of data flow, both in and out of a business

Variety : the combination of structured and unstructured data collected in “big data” systems

Veracity : the accuracy and reliability of data collected in “big data” systems

Data types and format

Structured data : numbers or text stored and organized in a structure of columns and rows.

→ Excel - ID in a database - rating/opinion

Semi-structured data : data not conventionally formatted and not referenced

→ Email - Tweet - File

Unstructured data : data storing information that is unorganized, difficult to extract, and in any other form that is difficult to organize in a structured format.

→ Social media posts - video - image - sound file

Data storage format : how a data type is represented

→ Example:

- Numeric - Integer: 45 - Decimal: 12.5
- Character string: data
- Unicode character string: data
- Boolean: True
- Date and Time: 01/01/2022 12:01:34

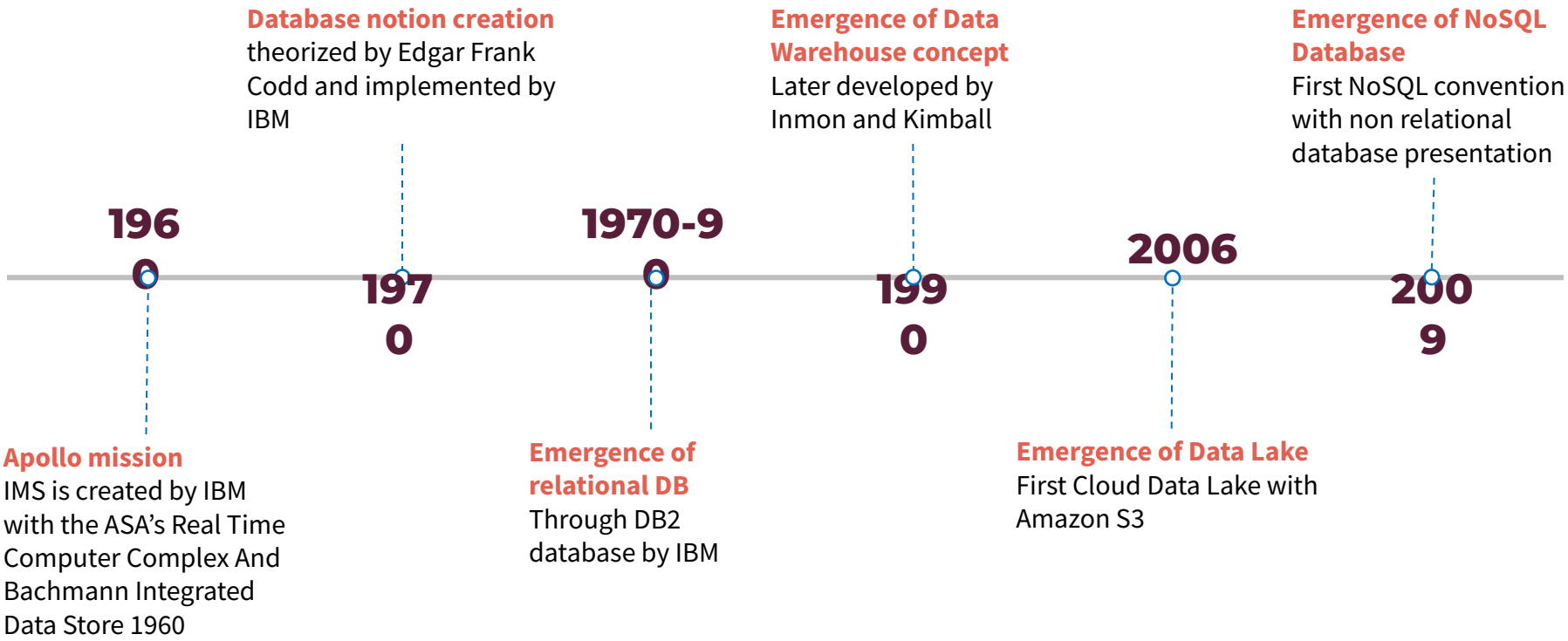
The data format is important for creating databases. The specific format of a column allows specific operations to be performed

```
SELECT
    symbol, name, exchange, assetType, ipoDate, delistingDate
FROM
    default.ticker_listing
WHERE
    symbol IN ('AAPL', 'GOOGL', 'AMZN', 'MSFT')
```

SQL query example

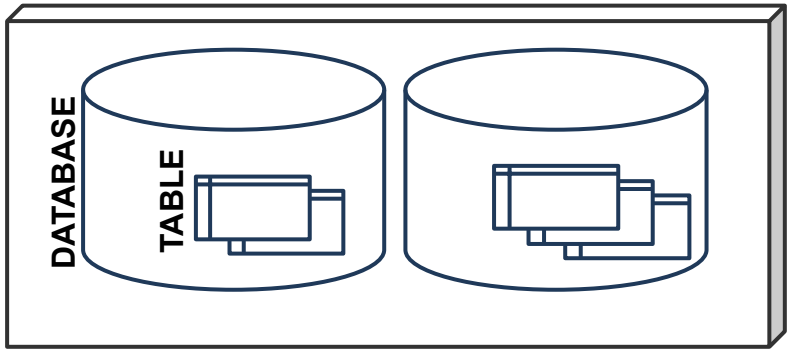
Data storage

History of database



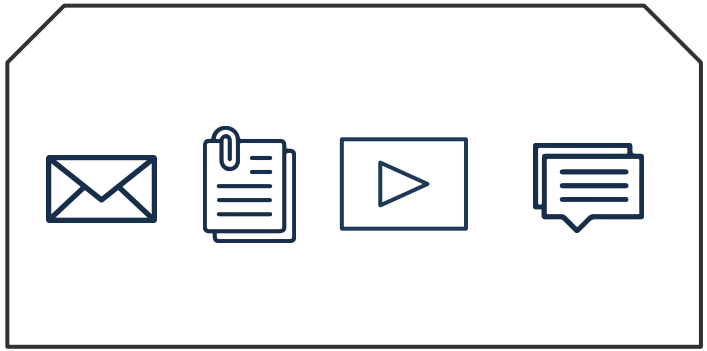
Analytical data storage

DATA WAREHOUSE



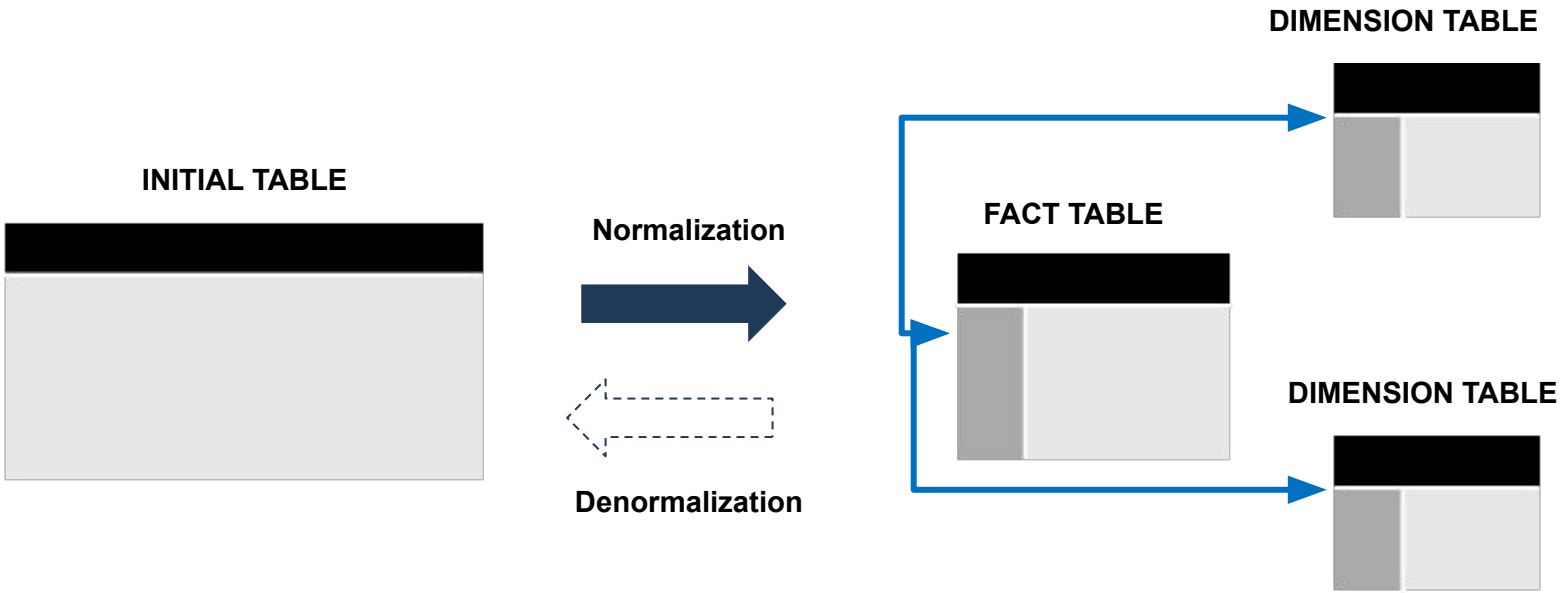
Structured Data

DATA LAKE

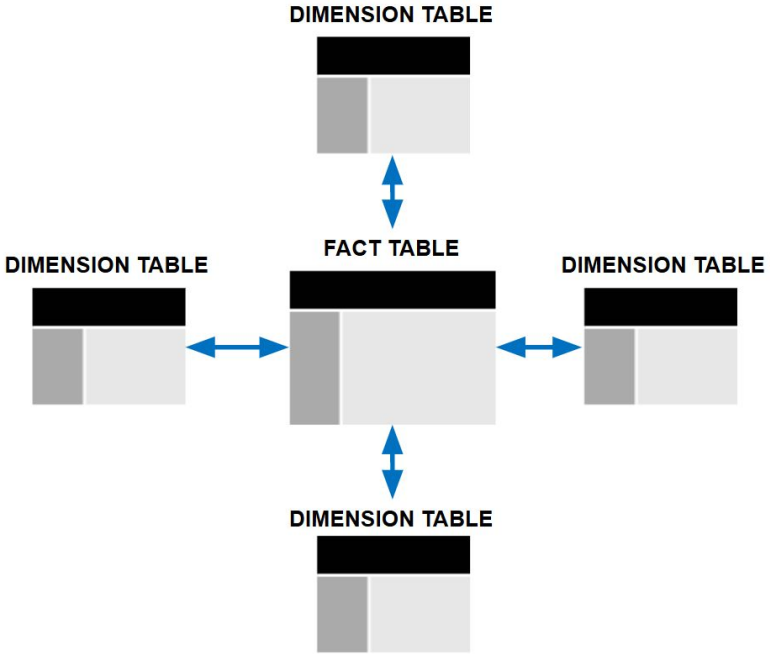


Unstructured Data

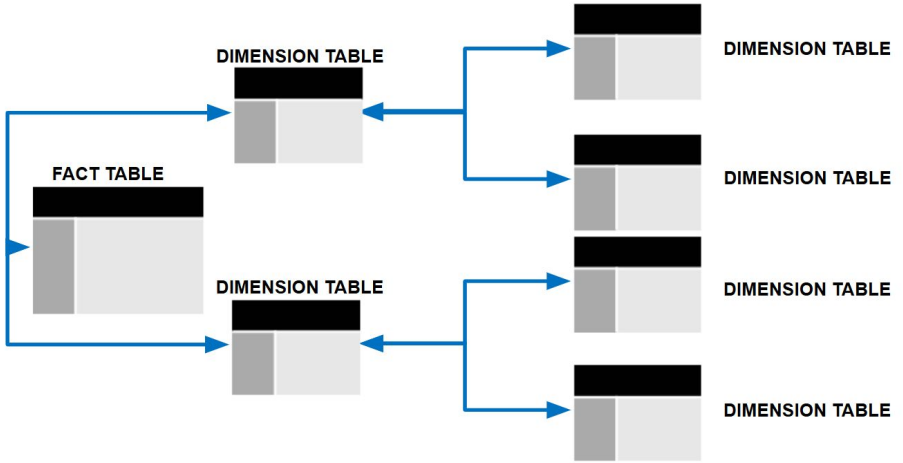
RELATIONAL DATABASE



Data storage optimization



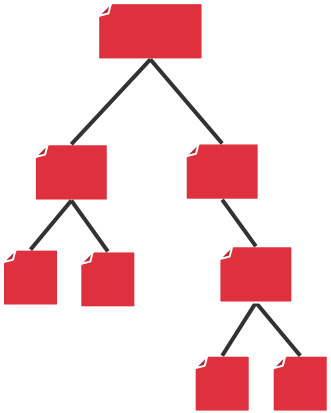
Star Schema



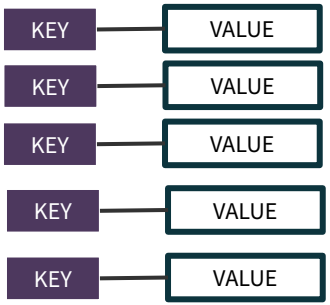
Snowflake Schema

NOSQL DATABASE

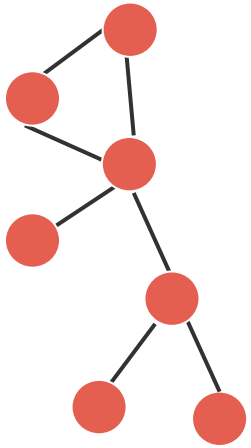
Document



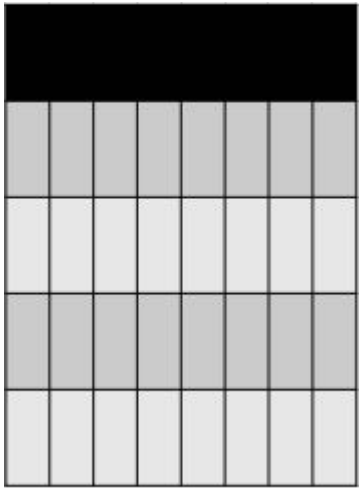
Key-Value



Graph



Wide-column



Add value to data

Steven VINCENT - Burgundy Scool of Business-
Jan. 2025 -

Data sources

Primary data : information directly collected for a specific purpose.
→ Ex: collect Name, first name, email address of the customer

Secondary data : the data has already been collected, often for other purposes or by another organization.
→ Ex: file purchase, public data.

Open Data : freely accessible data produced by a private company but above all by a public administration/establishment/community.

Data brokers : collect data themselves or buy it from other companies and aggregate information, legally or not, with data from other sources.

Data quality

Completeness

Existence of the data :
Mandatory presence of a value for the data

Completeness of scope:
Data completeness for a dataset

Exactitude

Veracity:
Compliance of the data with the documentation internal reference

Compliance with standards:
Compliance of repository data values with the list of repository values

Consistency

Consistency of data within data group :
Within the same group of data, consistency of data in relation to other data

Data consistency between data groups :
Between two distinct groups of data, consistency of the data between them

Plausibility:
Detecting an outlier in the data

Uniqueness

Uniqueness:
Compliance with unique data identification rules when it must be unique

Punctuality

Freshness:
Compliance of the last data update

Punctuality:
Compliance between the delivery date and the actual data submission date

Digital Analytics

Data analytics: science which consists of analyzing raw data in order to better interpret information. The analysis process is most often automated and via software to collect and analyze the data

PREDICTIVE

Which will probably happen in the short term.

« What's going to happen?»

What happened to sales the last time we had a hot summer?
How many weather models predict a hot summer this year?

DESCRIPTIVE

Describes what happened over a given period of time.

« What happened?»

Has the number of views increased? Are sales stronger this month than last month?

PRESCRIPTIVE

Suggests an action plan.

« What should I do ? »

We should add an evening shift to the brewery and rent an additional tank to increase production if the probability of a hot summer is measured as the average of these five weather models and the average is greater than 58%.

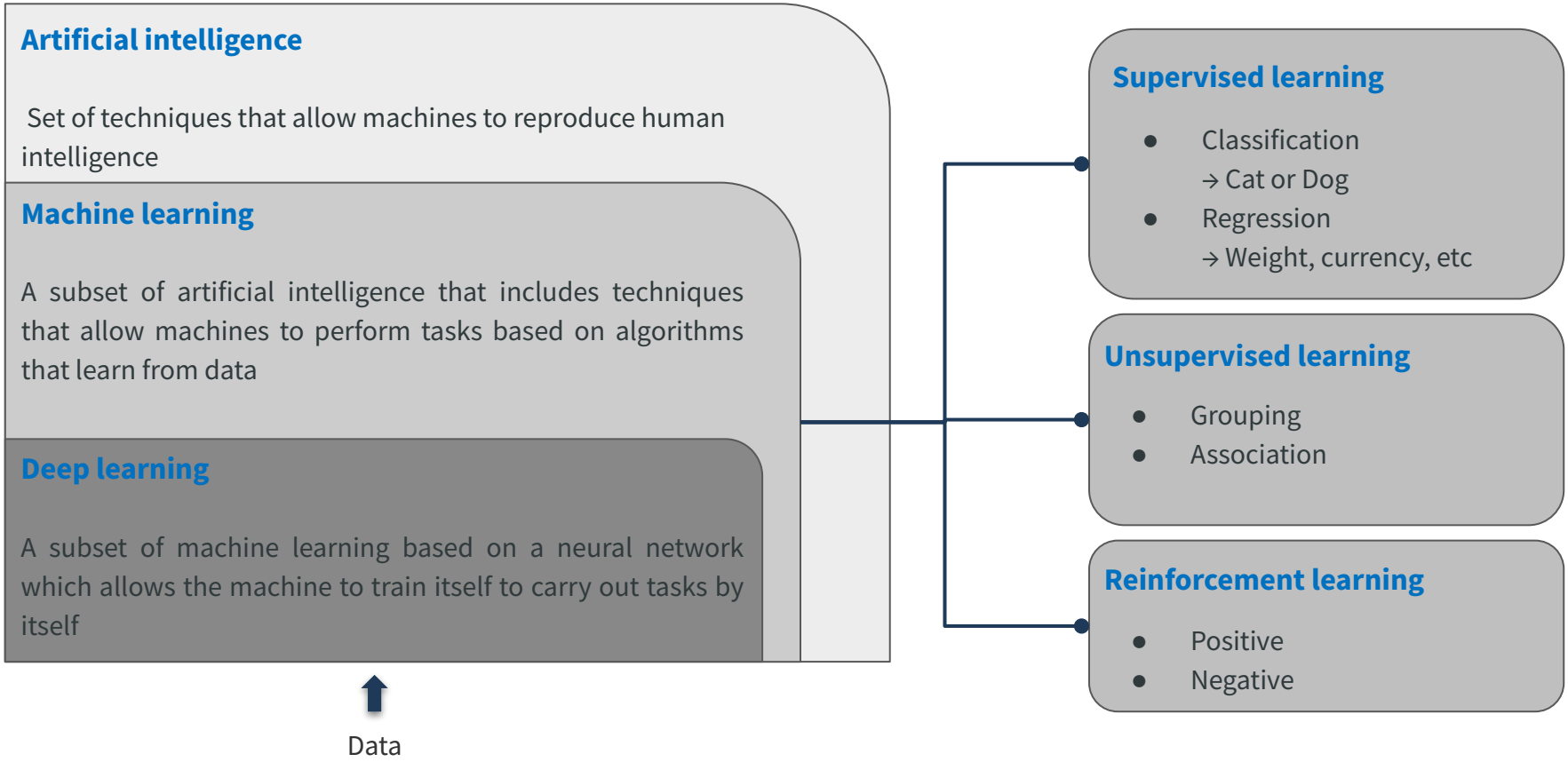
DIAGNOSTIC

Reason why something happened. This involves more diverse data inputs and a bit of guesswork.

« Why did this happen?»

Has the weather affected beer sales? Has this latest marketing campaign had an impact on sales?

Data: basis of artificial intelligence



Large language models

Large Language Models (LLM) are programs trained on massive linguistic data sets to identify characteristics and relationships between similar data elements without human intervention (automatic learning of the structure of the language). human language.)

Microsoft 365 Copilot

GitHub Copilot

ChatGPT

Predictive AI uses machine learning to extrapolate the future.
Generative AI uses machine learning to create content

USE CASE

- **Conversational robots and virtual assistants:** customer chatbots for assistance, monitoring of contacts via a website or personal assistant (answers open questions).
- **Code generation and debugging:** code snippets, identification and correction of code errors, and completion of programs on instructions.
- **Text sentiment analysis:** automation of understanding customer satisfaction.
- **Text classification and clustering:** categorizing and sorting large volumes of data to identify themes and trends
- **Translation** of documents and web pages
- **Synthesis and paraphrase:** summaries of articles, publications, calls, meetings, highlighting important points.
- **Content generation:** writing text, synopsis or new content that can serve as a draft or primer.

General Data Protection Regulation (GDPR)

General Data Protection Regulation (GDPR) :
In French « Règlement Général sur la Protection des Données » or RGPD, regulates the processing of personal data in the territory of the European Union.

Objectives :

- Adapt the legal context to technological developments in society
- Strengthens citizen control over the use of their data
- Harmonize rules in Europe for businesses in order to develop digital technology based on user trust

Who is affected:

Any organization, public or private, regardless of its size, country of establishment and activity, which processes personal data targeting European residents or established within the territory of the European Union.

Data protection

Personal data :
Any information relating to a natural person, identified or identifiable, directly or indirectly, from a single piece of data or the crossing of a set of data.

Processing of personal data :
Operation or set of operations, computerized or physical, relating to personal data, regardless of the process used. It must have a legal and legitimate objective and purpose with respect to the professional activity.

Legal entities are not affected by the GDPR (company contact details, standard email and telephone).

WORKSHOP