



جامعة باجي مختار - عنابة
BADJI MOKHTAR - ANNABA UNIVERSITY

Course :

Foundations of Data Science

Prepared by : Dendani Bilal

bilal.dendani@univ-annaba.dz

Speciality :

Data Science

Objectifs :

- Provide a solid foundation in Data Science, with a focus on essential mathematical tools, particularly Linear Algebra.
- Prepare students to tackle more advanced courses in Data Analysis and Machine Learning by equipping them with the necessary theoretical knowledge and practical skills.

Prerequisites:

-  Basic knowledge of Mathematics
-  Statistics, and  Programming.

Course information :

Resources for Data science course:

 Moodle (Main Resource):

<https://elearning-facsci.univ-annaba.dz/course/view.php?id=2157>

 Google Classroom:

<https://classroom.google.com/c/ODE4MTMyNTUxNjg4>

 Discussion and Question : during course & email bilal.dendani@univ-annab.dz

 Sharing the course once completed

 Grading:

 Exam (60%) +  Practical work (40%)

Course Content

Chapter 1. Introduction to Data Science

- What is Data Science?
- Origins and Challenges of Data Science
- Facets and types of data
- How Data Science works
- Use cases and application domains
- The Big Data and Data science ecosystem

Chapter 2. The Data Science Process

- Roles and responsibilities in a Data Science project
- Overview of the Data Science project life cycle
- **Step 1:** Define research objectives and create a project charter
- **Step 2:** Data collection
- **Step 3:** Data cleaning, integration, and transformation
- **Step 4:** Exploratory Data Analysis (EDA)
- **Step 5:** Model building
- **Step 6:** Presenting results and developing applications on top of them

Course Content

Chapter 3: Tools and Technologies Used in Data Science

- Data storage tools
- Data preparation tools
- Data visualization tools
- Notebook IDE tools
- Comprehensive Data Science platforms

Chapter 4: Fundamentals of Linear Algebra

- **Vectors and Vector Spaces**
 - Definition and operations on vectors
 - Vector spaces and subspaces
- **Matrices**
 - Definition, types of matrices, and operations (addition, multiplication, inversion)
 - Special matrices (diagonal, orthogonal, identity matrices)
- **Systems of Linear Equations**
 - Solving linear systems (Gauss-Jordan method, LU decomposition)

Course Content

Chapter 5: Linear Models

- **Simple Linear Regression**
 - Simple linear regression model and parameter estimation
 - Interpretation of results, errors, and diagnostics
- **Multiple Linear Regression**
 - Extension to models with several explanatory variables
 - Model selection and regularization (Ridge, Lasso)

Chapter 6: Advanced Linear Algebra

- **Eigenvalues and Eigenvectors**
 - Calculation and interpretation of eigenvalues and eigenvectors
 - Applications in dimensionality reduction (Principal Component Analysis – PCA)
- **Singular Value Decomposition (SVD)**
 - Theory and computation of singular value decomposition
 - Practical applications (image compression, collaborative filtering)

Chapter 7: Numerical Methods in Data Science

- **Optimization Techniques**
 - Introduction to optimization methods (Gradient Descent, Newton-Raphson)
 - Applications in Machine Learning
- **Numerical Algorithms**
 - Numerical solution of systems of equations
 - Root-finding techniques and numerical integration

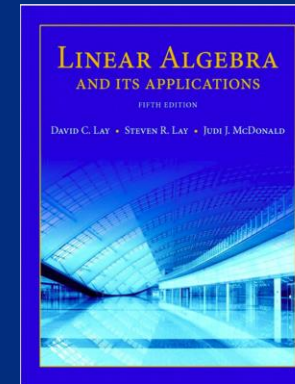
References



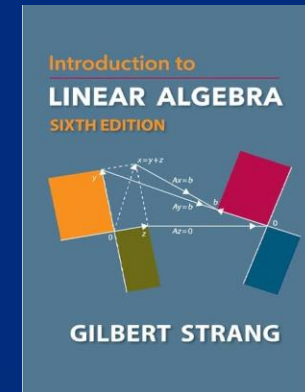
Dietrich, D., "Data science & big data analytics: discovering, analyzing, visualizing and presenting data", Wiley, 2015.



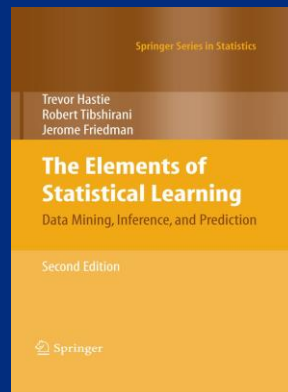
Lutz, M., Biernat, E., "Data Science: fondamentaux et études de cas: Machine Learning avec Python et R", Editions Eyrolles, 2015.



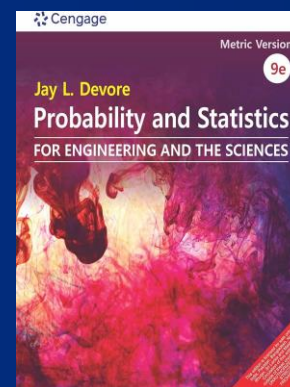
«Linear Algebra and Its Applications» by David & Steven Lay, and Judi McDonald



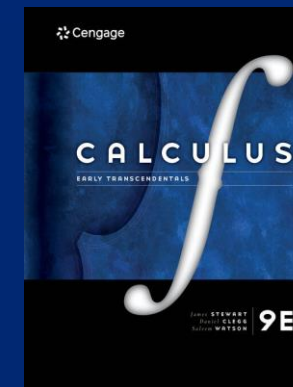
"Introduction to Linear Algebra" de Gilbert Strang.



"The Elements of Statistical Learning: Data Mining, Inference, and Prediction" de Trevor Hastie, Robert Tibshirani, et Jerome Friedman



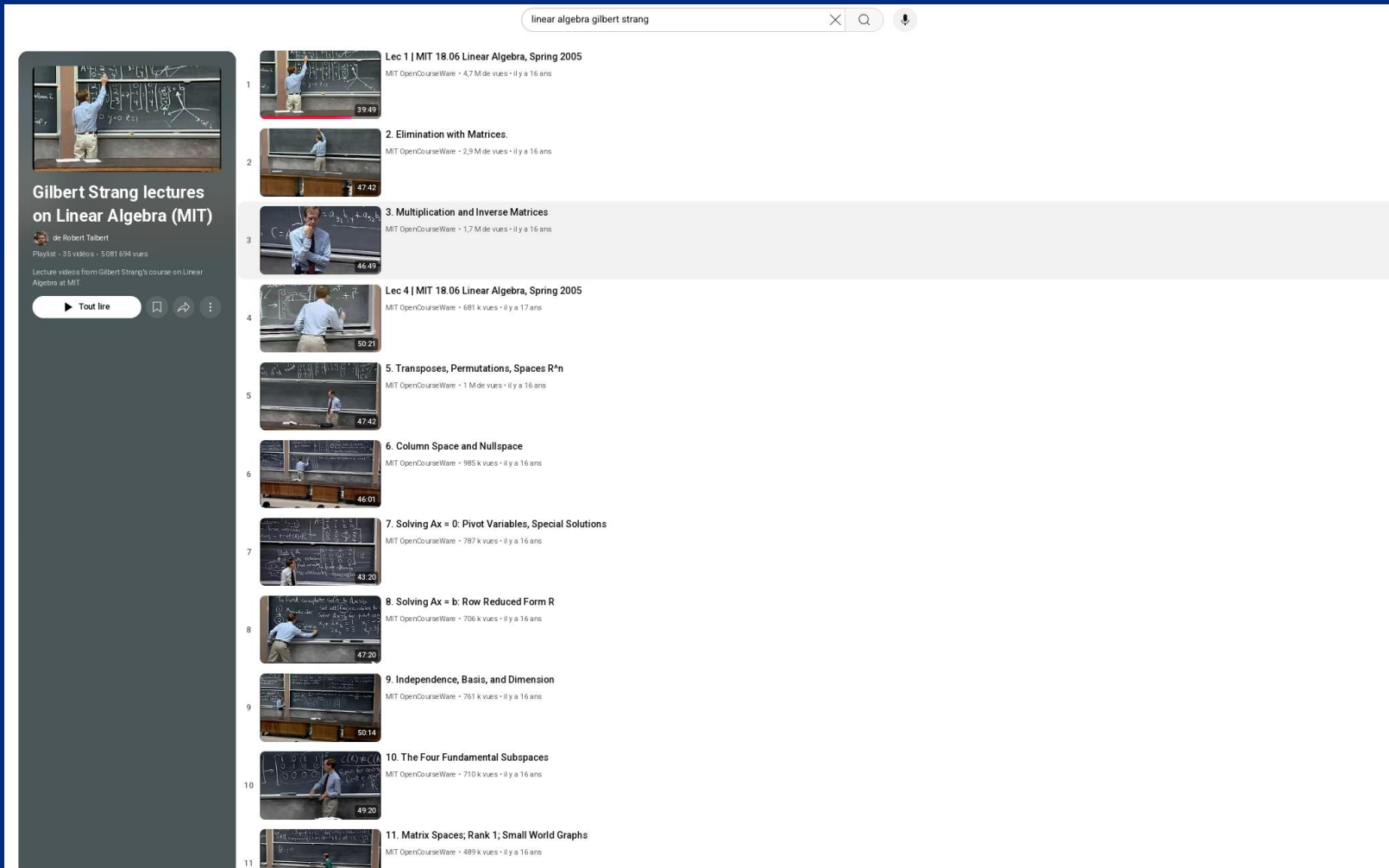
"Probability and Statistics for Engineering and the Sciences" de Jay L. Devore.



"Calculus: Early Transcendentals" de James Stewart.

Linear Algebra youtube free resource

 YouTube Channel: [MIT OpenCourseWare – Gilbert Strang Linear Algebra](#)



The screenshot shows the YouTube channel page for "Gilbert Strang lectures on Linear Algebra (MIT)". The channel is owned by Robert Talbert and has 35 videos with 5,081,694 views. The page displays a list of 11 lecture videos, each with a thumbnail, title, and view count. The search bar at the top shows "linear algebra gilbert strang".

| Lecture Number | Title | Views | Age |
|----------------|--|-------|--------|
| 1 | Lec 1 MIT 18.06 Linear Algebra, Spring 2005 | 4.7 M | 16 ans |
| 2 | 2. Elimination with Matrices. | 2.9 M | 16 ans |
| 3 | 3. Multiplication and Inverse Matrices | 1.7 M | 16 ans |
| 4 | Lec 4 MIT 18.06 Linear Algebra, Spring 2005 | 681 k | 17 ans |
| 5 | 5. Transposes, Permutations, Spaces \mathbb{R}^n | 1 M | 16 ans |
| 6 | 6. Column Space and Nullspace | 985 k | 16 ans |
| 7 | 7. Solving $Ax = 0$: Pivot Variables, Special Solutions | 787 k | 16 ans |
| 8 | 8. Solving $Ax = b$: Row Reduced Form R | 706 k | 16 ans |
| 9 | 9. Independence, Basis, and Dimension | 761 k | 16 ans |
| 10 | 10. The Four Fundamental Subspaces | 710 k | 16 ans |
| 11 | 11. Matrix Spaces; Rank 1; Small World Graphs | 489 k | 16 ans |

Chapter 1

Introduction to Data Science

Prepared by :
Dr. Bilal Dendani



جامعة باجي مختار - عنابة
BADJI MOKHTAR - ANNABA UNIVERSITY

Dr. DENDANI Bilal



Chapter 1 : Introduction to Data Science

- What is data?
- What is Data Science?
- Key definitions of domains related to Data Science
 - What is Big Data?
 - What are Artificial Intelligence, Machine Learning, and Deep Learning?
- The scientific origins of Data Science
- A brief history of the emergence of Data Science
- The challenges of Data Science
- Facets and types of data (structured, unstructured, and semi-structured)
- How Data Science works
- Use cases and application domains
- The Big Data and Data Science ecosystem

0. What is data ?

- Data refers to **raw elements** or **unprocessed facts** such as numbers, symbols, text, or images.
- When observed without interpretation, these remain simple and unorganized data.
- Once **analyzed** and **placed in context**, however, data becomes **information**, carrying meaning and relevance.



Figure 1. Data vs information

1. What is data science?

- Data Science is a discipline that leverages data to address complex problems and support informed decision-making.
- It is the **process of transforming** raw data into **meaningful** and actionable **insights**.
- This involves the **collection**, **analysis**, and **interpretation** of data to **identify patterns**, **generate predictions**, and guide strategic decisions.
- Data Science represents a blend of programming skills, statistical knowledge, and domain expertise, applied together to extract value and solve real-world challenges from data.



1. What is data science?

Data science is an interdisciplinary academic field that uses **statistics**, **scientific computing**, **scientific methods**, processing, scientific visualization, algorithms, and systems to extract or extrapolate knowledge from potentially noisy, structured, or unstructured data.

https://fr.wikipedia.org/wiki/Science_des_donn%C3%A9es

https://fr.wikipedia.org/wiki/Science_des_donn%C3%A9es

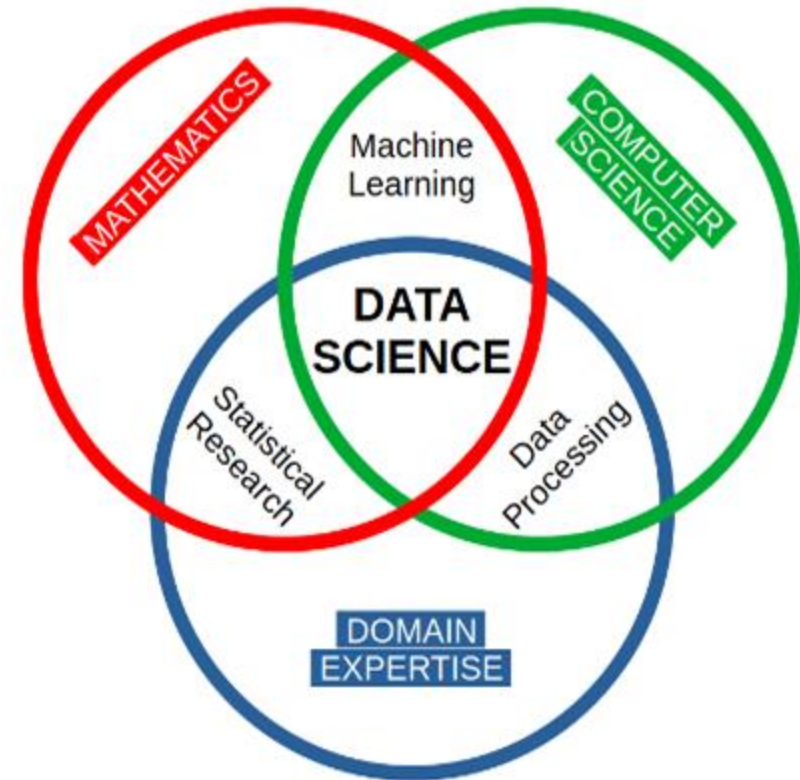


Figure 2. Data science inter-disciplinary field

2. Examples of data science applications

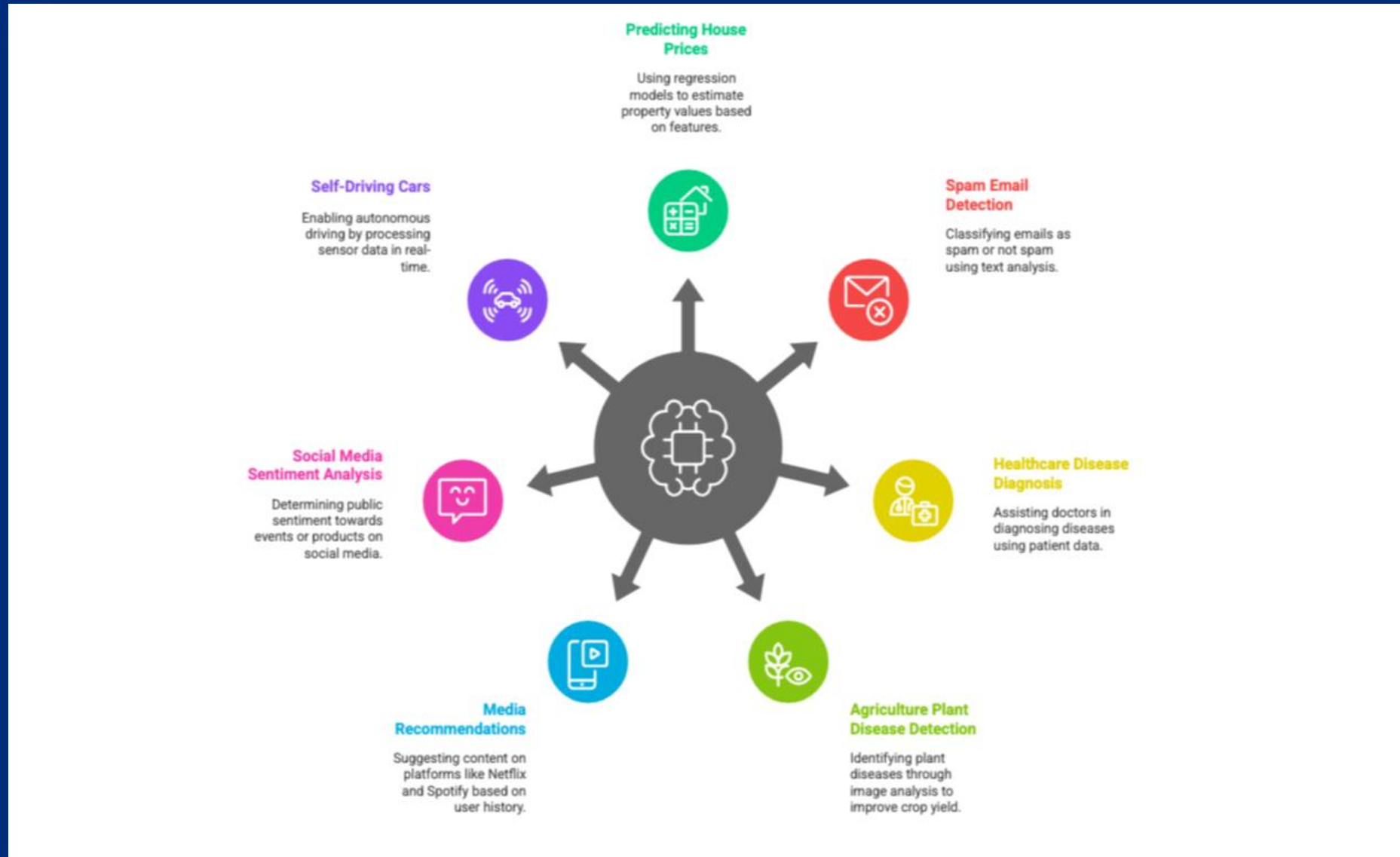


Figure 3. Data science examples of applications

3. What is Big Data?

- For the past few years, we have been hearing about the phenomenon of **Big Data**, often translated as “**massive data**.”
- Big Data is a field that has emerged to handle the **vast quantities of data generated** every day.
- The term Big Data refers to an **accumulation** of data so vast and **complex** that it cannot be processed using traditional database management tools.

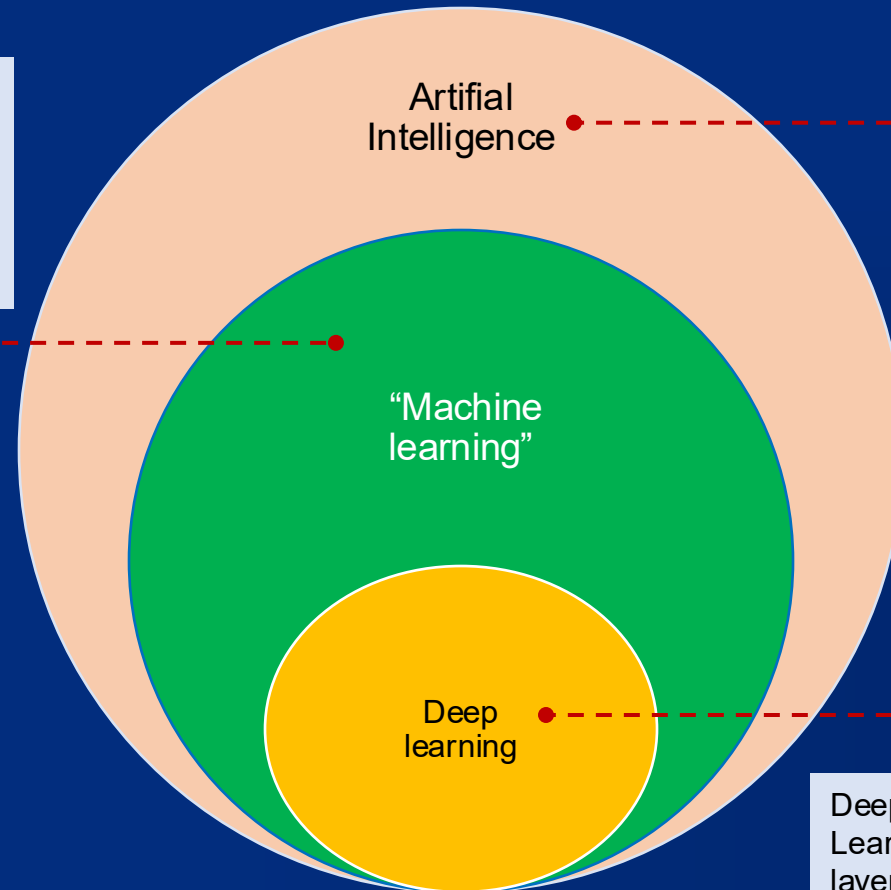


Figure 4. Big Data are everywhere

4. What are Artificial Intelligence, Machine Learning, and Deep Learning?

Machine Learning (ML) is a subfield of Artificial Intelligence that focuses on developing algorithms and mathematical models that enable machines to automatically learn patterns from data and improve their performance on tasks without being explicitly programmed.

Artificial Intelligence (AI) refers to the capability of machines or systems to perform tasks that typically require human intelligence, such as reasoning, planning, problem-solving, learning, and creativity.



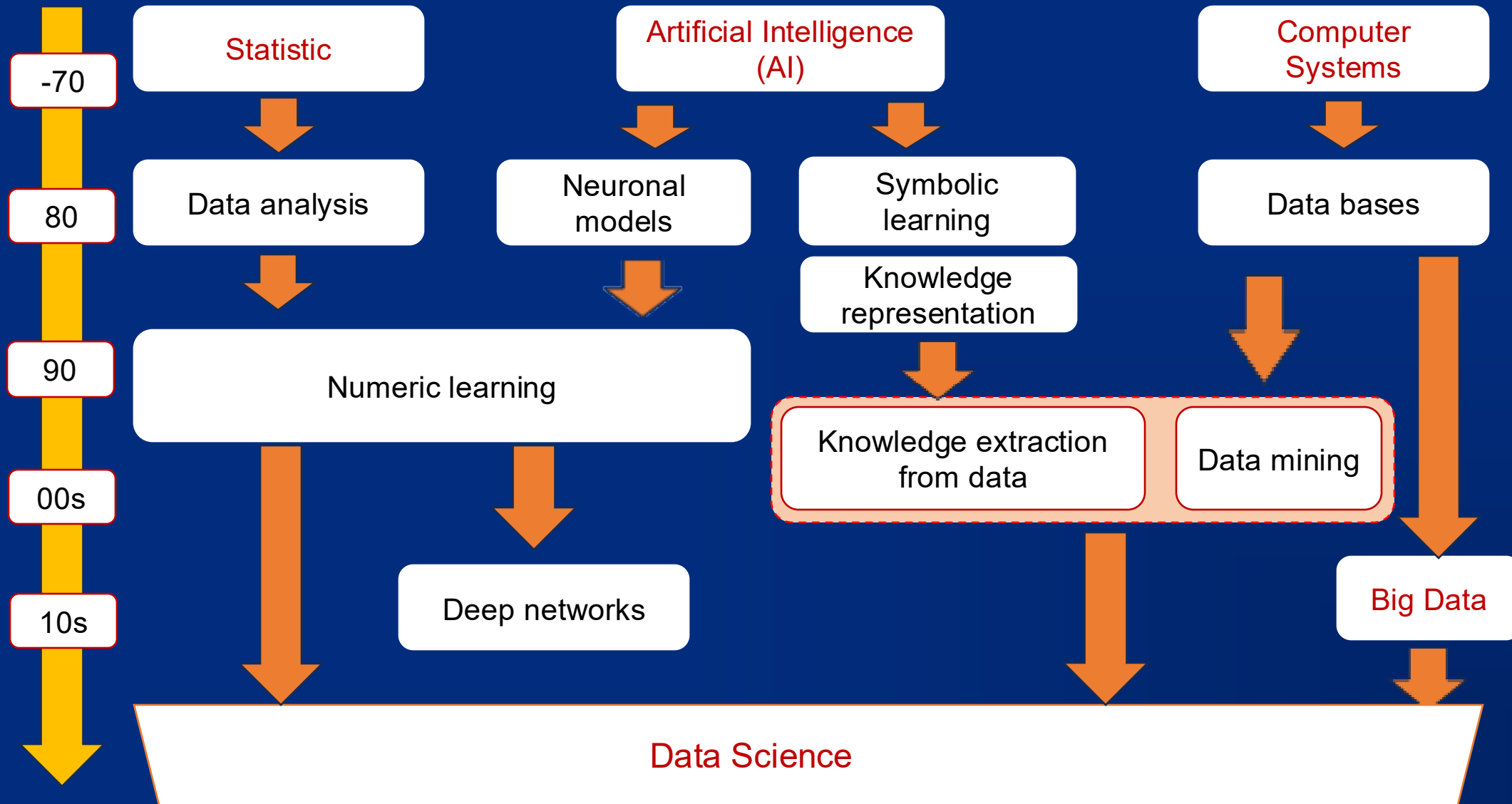
Deep Learning (DL) is a process and a subset of Machine Learning that uses neural networks with multiple hidden layers to model complex patterns in data.

Figure 5. Artificial Intelligence (AI), Machine Learning (ML) and Deep Learning (DL)





5. The Scientific Origins of Data Science



6. History related to Data Science

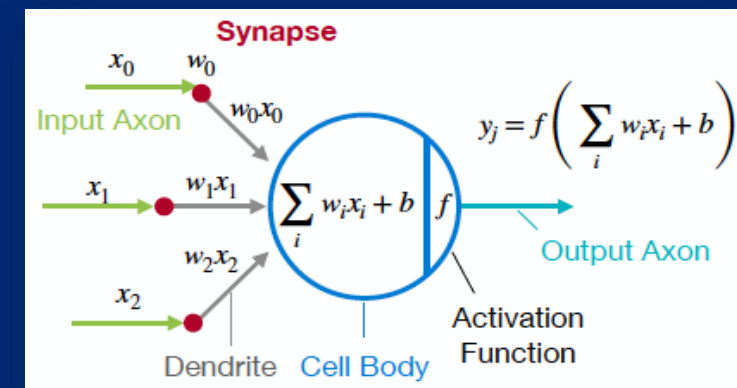
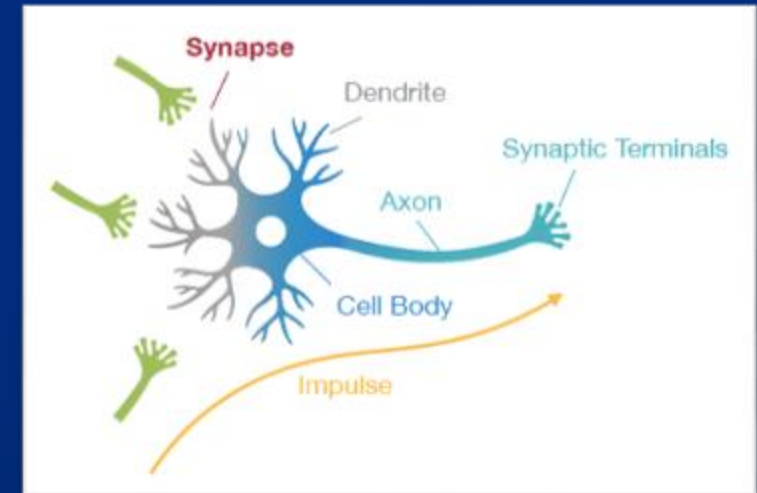
1943 – The Formal Neuron by McCulloch and Pitts



Warren Sturgis McCulloch
(1898 – 1969)



Walter Harry Pitts, Jr.
(1923 – 1969)



https://images.slideplayer.com/22/6379712/slides/slide_8.jpg

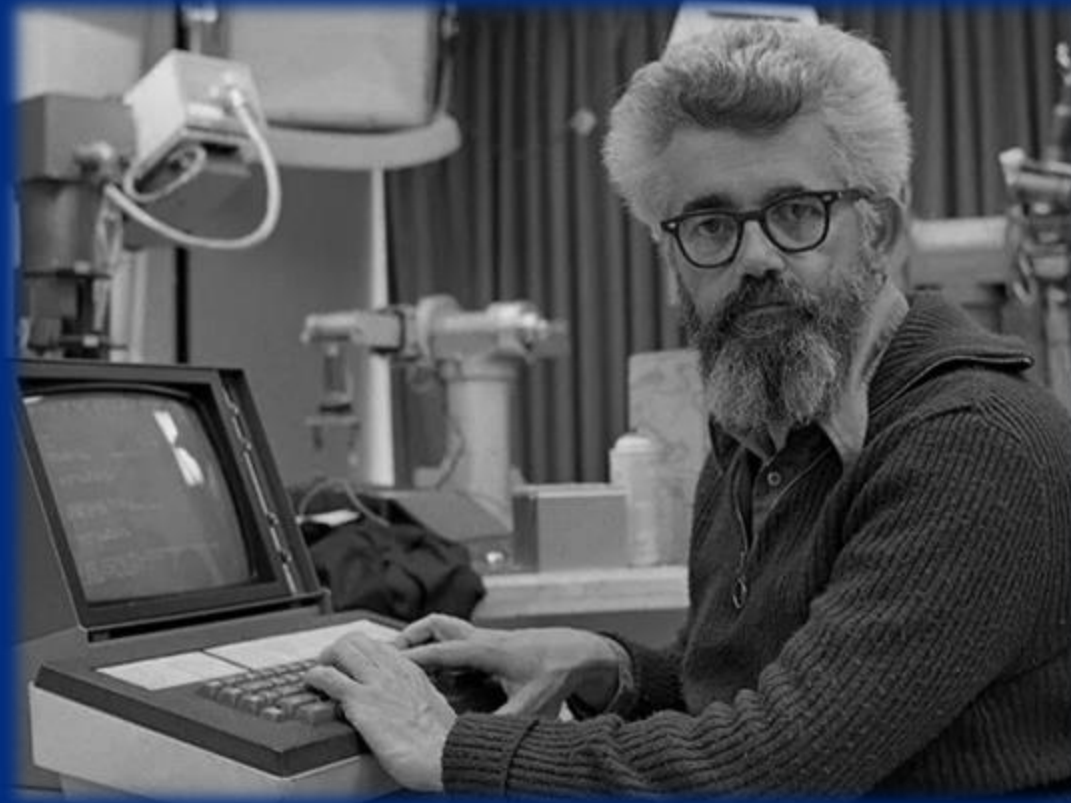
<https://efficientml.ai>

1956

John McCarthy

Artificial Intelligence

“The science and engineering of making intelligent machines, brilliant computer programs”. -John McCarthy-



<https://www.independent.co.uk/news/obituaries/john-mccarthy-computer-scientist-known-as-the-father-of-ai-6255307.html>

1959

Arthur Samuel

What is machine learning?

“Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed.”



<https://www.ibm.com/history/early-games>

1963

Alan Turing

- Alan Turing proposed an innovative approach to Artificial Intelligence.
- He suggested designing a program that simulates the brain of a child, which could then be trained and educated to gradually reach the intellectual maturity of an adult.



1997: Deep blue

The day Deep Blue defeated Garry Kasparov in chess

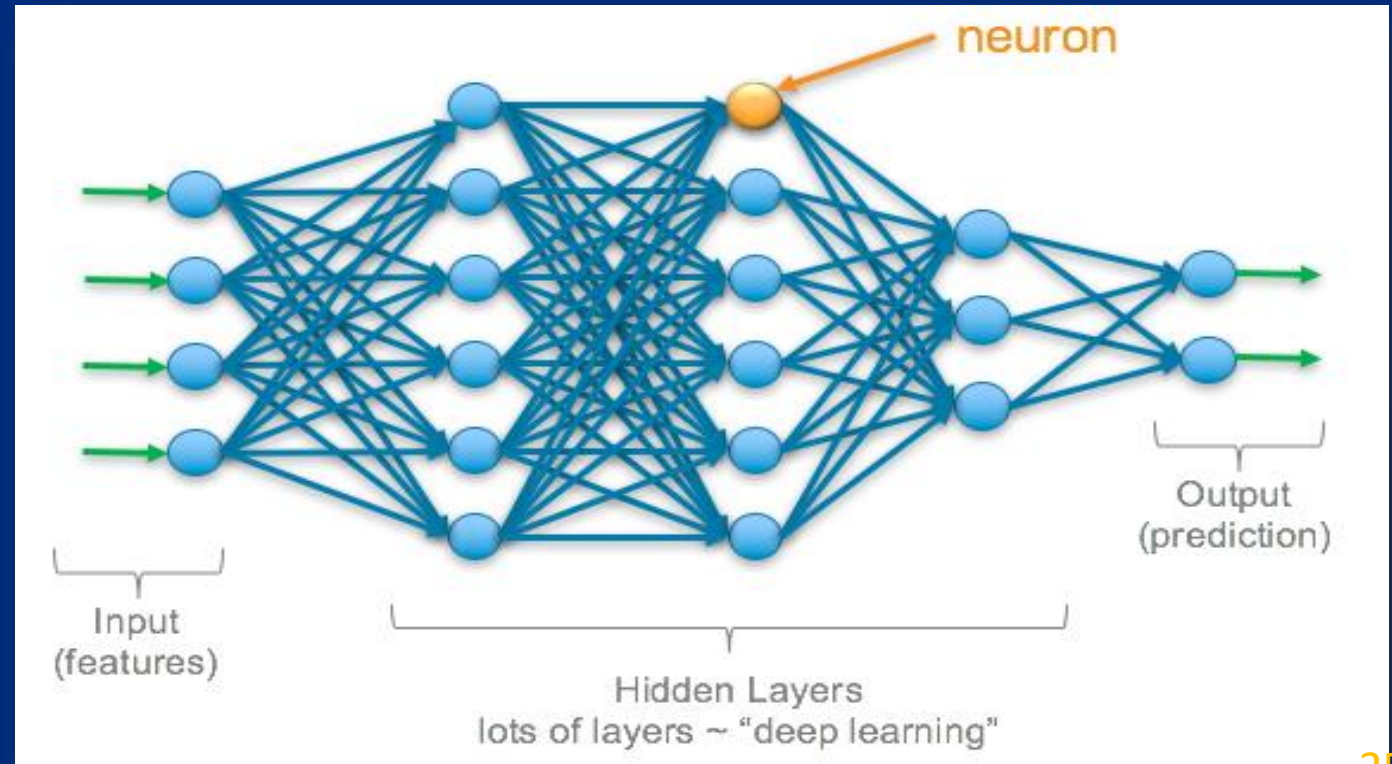


2005 Deep Learning

Geoffrey Hinton (born December 6, 1947) is a Canadian researcher specializing in artificial intelligence, particularly in artificial neural networks. He is a member of the Google Brain team (2013-2023) and a professor in the Department of Computer Science at the University of Toronto. He was one of the first to apply the backpropagation algorithm for training multilayer neural networks. He is one of the leading figures in the deep learning community.



https://fr.wikipedia.org/wiki/Geoffrey_Hinton

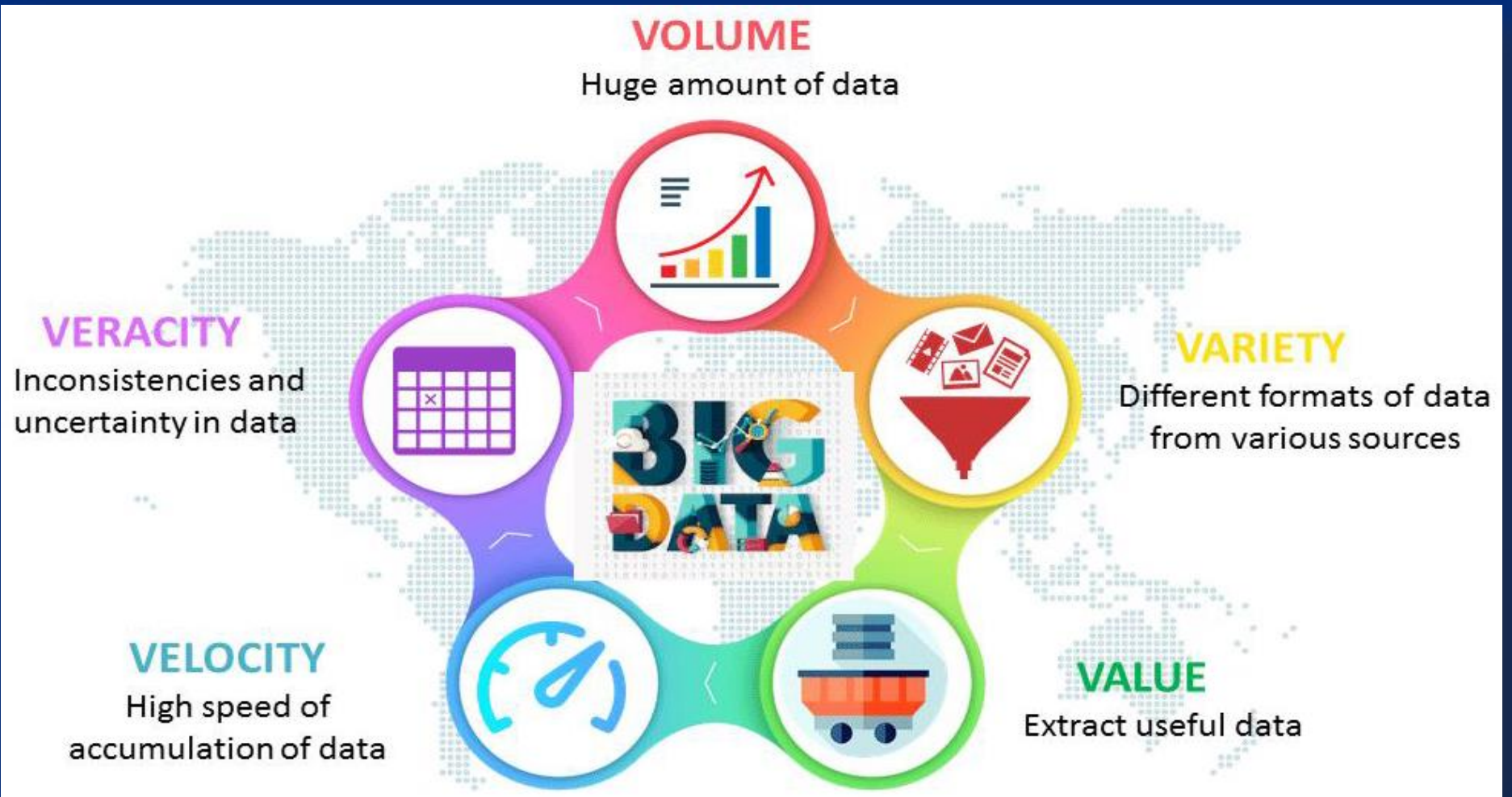


<https://srrghn.medium.com/deep-learning-common-architectures-6071d47cb383>

Figure 10. Deep neural network overview

2010 Big Data

The term *Big Data* refers to the massive and highly complex accumulation of data that exceeds the capacity of traditional database management tools to store, process, and analyze effectively



<https://iwconnect.com/algorithms-for-solving-big-data-use-cases/>

Figure 11. Big data 5 Vs

2015 Alpha Go

- Developed by Google DeepMind to play the complex board game Go.
- Combined deep neural networks and reinforcement learning.
- Learned strategies from human games and self-play.
- Defeated world champion Lee Sedol in 2016.
- Marked a major breakthrough in artificial intelligence and data-driven learning.



In October 2015, AlphaGo won a match against Fan Hui with a score of 5 to 0.

Dr. DENDANY Bilal
https://fr.wikipedia.org/wiki/Match_AlphaGo_-_Lee_Sedol#/media/Fichier:FanHui.jpg



AlphaGo



AlphaGo defeated Lee Sedol in all games except the fourth one, between March 9 and 15, 2016.

https://fr.wikipedia.org/wiki/Lee_Sedol

2022

Open AI

OpenAI (“AI” standing for *artificial intelligence*) is an American artificial intelligence company founded in 2015 in San Francisco, California.

Its mission is to develop and promote artificial general intelligence (AGI) that is *safe and beneficial for all humanity*.

OpenAI’s launch of ChatGPT in November 2022 sparked a global interest in conversational agents and generative AI, reaching 100 million users in just two months.



<https://fr.wikipedia.org/wiki/OpenAI>

2023

MidJourney and the Rise of Generative AI Art

- MidJourney is an AI tool that creates images from text prompts using deep learning.
- It became popular in 2023 for producing realistic and artistic visuals.
- Uses large datasets and machine learning models to understand and generate creative content.
- Demonstrates how data science powers creativity and innovation in the arts.
- Represents the rise of generative AI, where machines can mimic human imagination.






<https://www.midjourney.com/home>

e

7. Challenges of Data Science

1. Volume and Complexity of Data

-  **Big Data volume** represents the exponential growth of data that makes storage, processing, and analysis increasingly complex.
-  **Variety** Data comes in many forms — structured, semi-structured, and unstructured, requiring diverse tools and methods.
-  **Velocity** Data must often be processed in real time or near real time to enable quick decision-making.

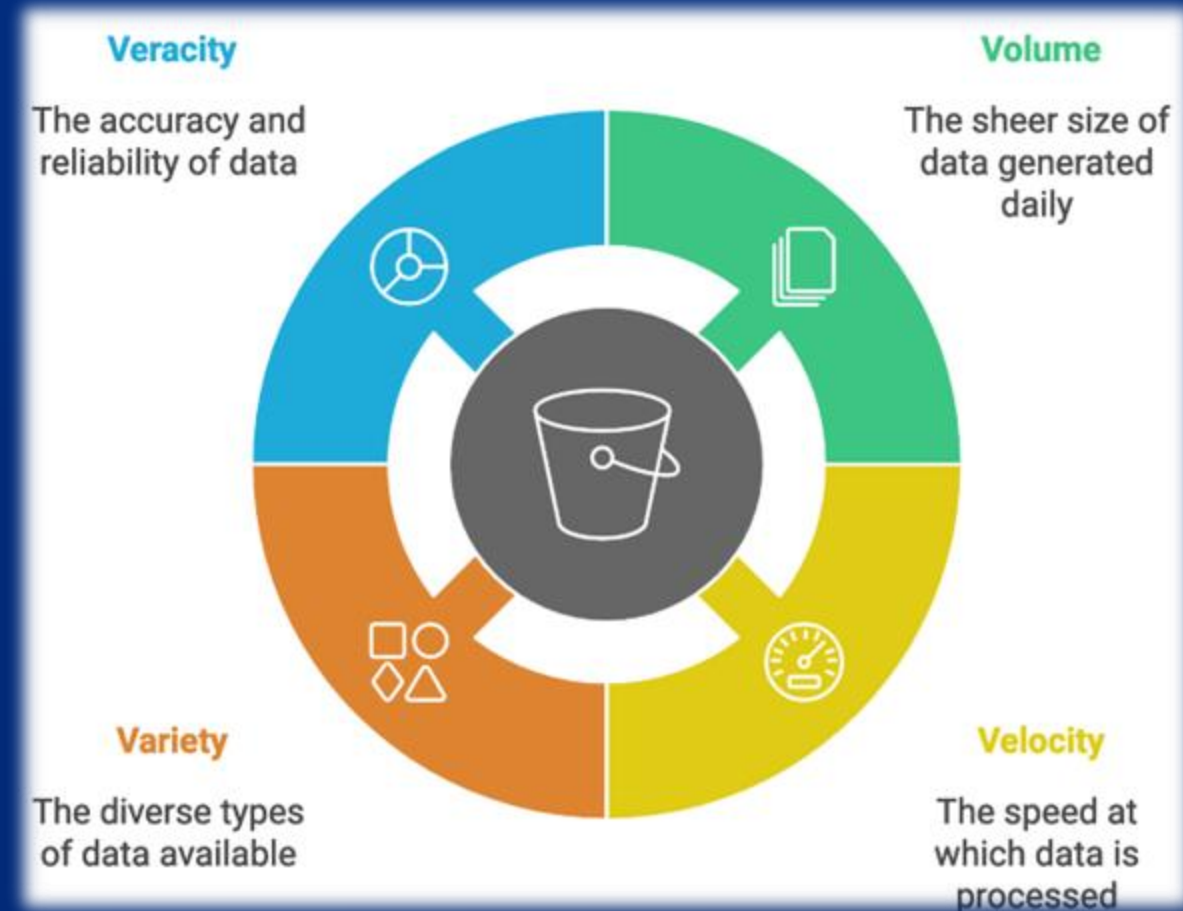





Figure 12. Data science challenges: volume and complexity of data

7. Challenges of Data Science

2. Data Quality:

-  **Accuracy:** Incorrect or incomplete data can lead to biased or unreliable results.
-  **Consistency:** Data must be consistent across sources to ensure the reliability of analyses.
-  **Relevance:** It is essential to select data that is relevant to the specific question or problem being addressed.

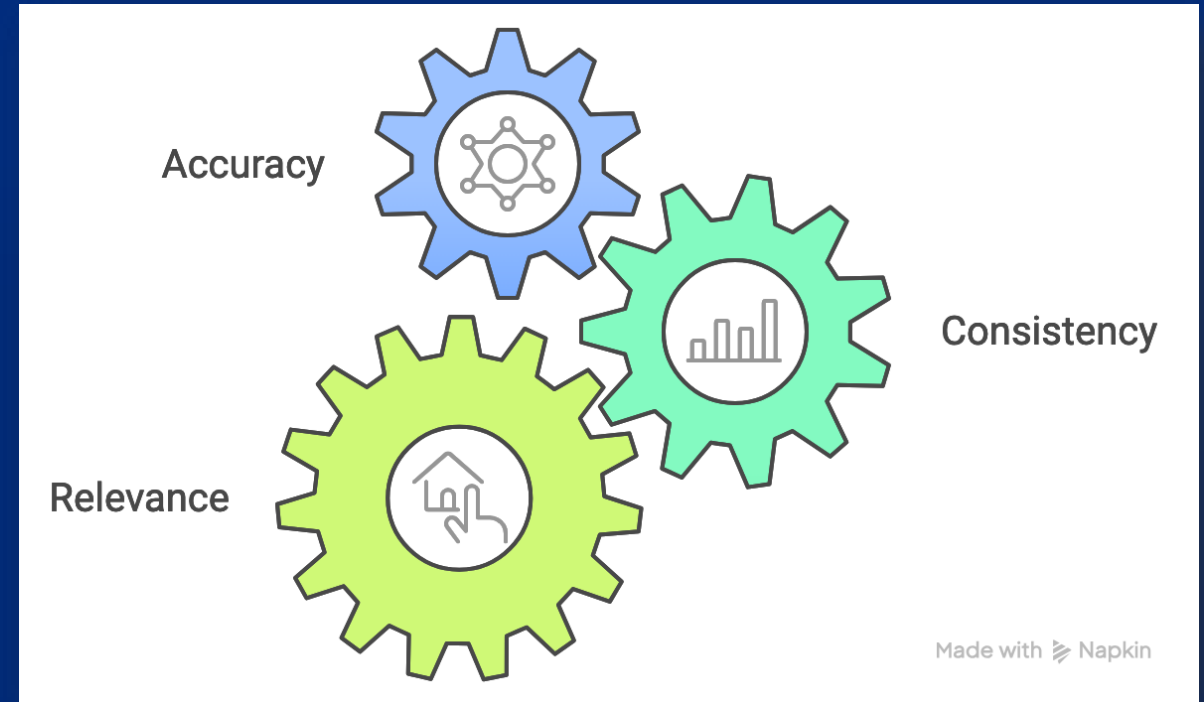


Figure 13. Data quality challenge

7. Challenges of Data Science

3. Bias and Ethics in Data Analysis

Algorithmic Bias: Machine learning and artificial intelligence models are often biased if the training data itself contains biases.

7. Challenges of Data Science

4. Data Security

- **Cyberattacks and Data Breaches:**

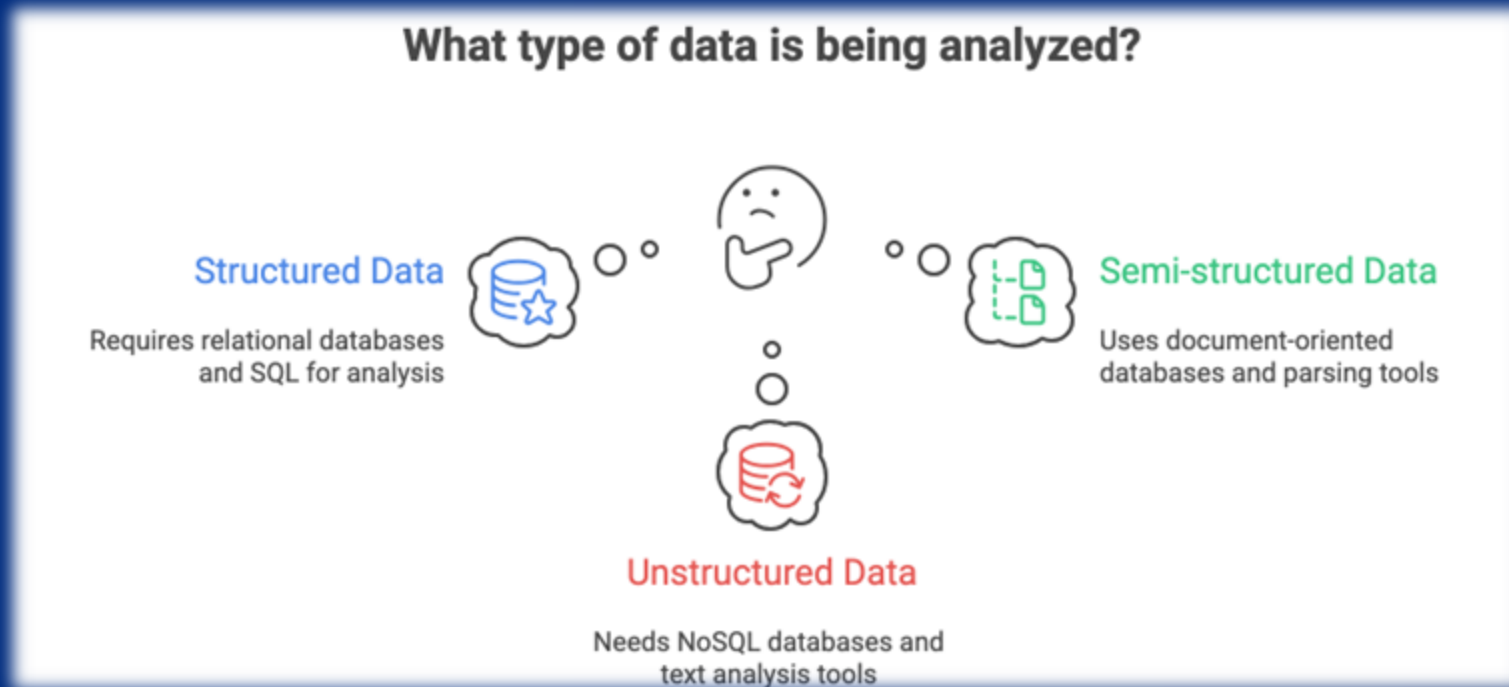
Data science often relies on massive volumes of information stored in databases or cloud infrastructures. This makes it an attractive target for cyberattacks. Ensuring the security of infrastructures, data integrity, and system resilience is therefore essential.

- **Protection Against Manipulation:**

Data can also be manipulated to produce biased or dishonest results, such as fake news or misinformation.

8. From Challenges to Understanding Data

- Before solving the challenges of data science, such as data security, quality, and bias, we must first *understand what kind of data we are dealing with*.
- Data can be **structured**, **unstructured**, or **semi-structured**, and each type requires different tools, storage systems, and analysis methods.



8. Facets and Types of Data (structured data)

1. Structured Data

Data can be identified according to their structure.

Definition of Structured Data:

Structured data describes a property (e.g., name, address, credit card number) of an entity (e.g., customer, product) according to a fixed model or template.

Examples:

- Data stored in spreadsheets (e.g., Excel files).
- Records stored in the tables of a relational database.
- Each property is easily distinguishable from the others.
- It corresponds to a unit within the structure (e.g., a column in a table).

8. Facets and Types of Data (unstructured data)

Definition (Unstructured Data):

Unstructured data describes an entity that does not have a defined structure because its properties cannot be easily distinguished from one another.

Examples:

- A text document is unstructured.
- Description of an entity's properties embedded in a rich context.
- No direct access to these properties.

8. Facets and Types of Data (semi-structured data)

Definition

Semi-structured data has a structure where entities and their properties can be easily distinguished, BUT the organization of the structure is not as strict as that of a database table.

- **Examples:**
 - XML document
 - JSON file
 - HTML page

Xml

```
<book id="bk101">
  <author>Gambardella, Matthew</author>
  <title>XML Developers Guide</title>
  <genre>Computer</genre>
  <price>44.95</price>
  <publish_date>2000-10-01</publish_date>
</book>
```

Json

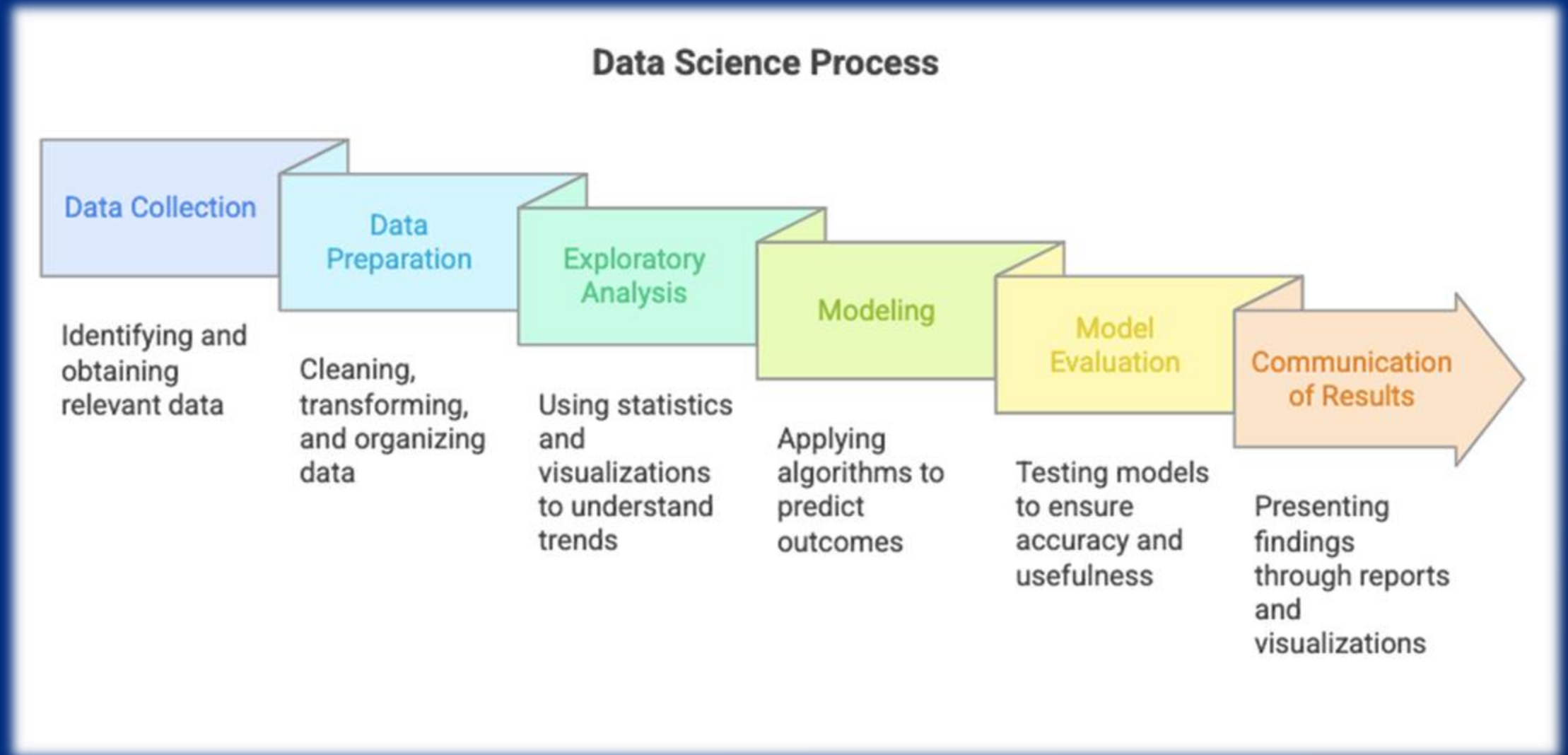
```
{
  "nom": "Alice",
  "âge": 30,
  "adresse": {
    "rue": "123 Rue de Paris",
    "ville": "Paris"
  },
  "téléphones": ["123-456-7890", "098-765-4321"]
}
```

9. How Does Data Science Work?

- Data science is a discipline that operates by following a structured process of collecting, processing, analyzing, and interpreting data.
- It relies on a series of key steps to extract actionable insights from data.
- It's a bit like a **detective solving a mystery** by using clues.



9. How Does Data Science Work?



Data Science Process

The data science process refers to the general steps used to extract information and insights from data. It is a methodological approach applied across all types of data science projects.

- The typical steps include:
 - **Data Collection:** Identifying and acquiring relevant data.
 - **Data Preparation:** Cleaning, transforming, and organizing data to make it usable.
 - **Exploratory Data Analysis (EDA):** Using statistics and visualizations to understand trends and relationships within the data.
 - **Modeling:** Applying algorithms to predict outcomes or classify data.
 - **Model Evaluation:** Testing models to ensure their accuracy, performance, and usefulness.
 - **Communication of Results:** Presenting insights through reports, visualizations, or dashboards.

10. Use cases and domain applications

Healthcare:

- Disease detection from medical images.
- Analysis of medical records to predict hospitalizations.
- Creation of personalized treatments through genomic analysis (e.g., *IBM Watson in oncology*).

Marketing and Recommendations:

- Using machine learning algorithms to personalize offers and recommendations.
- Examples: Netflix (movie/series recommendations), Amazon (suggested products).

Finance and Insurance:

- Prediction of financial risks.
- Detection of banking fraud.
- Real-time transaction analysis.

Industry and Predictive Maintenance:

- Using IoT sensors and algorithms to predict equipment failures.
- System monitoring.
- Anomaly detection.

Transportation and Logistics:

- Route optimization.
- Autonomous vehicles.
- Delivery time prediction (e.g., *Uber* or *FedEx*).
- Enhanced driving experience.
- Vehicle monitoring systems.

Security and Surveillance:

- Detection of abnormal behaviors (intelligent surveillance cameras).
- Cyberattack analysis and prevention.

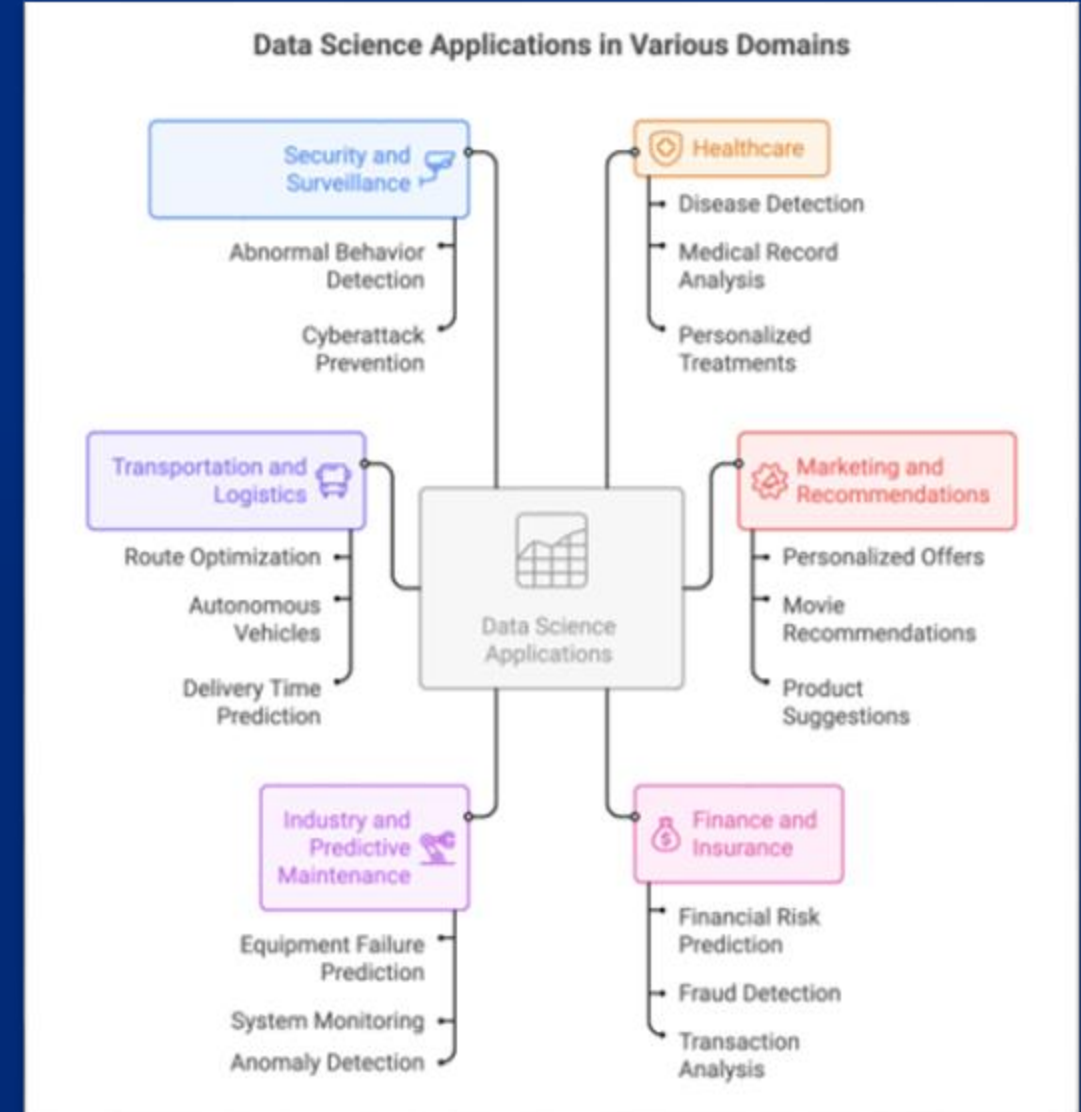


Figure 16. Data science applications in multiple domains

The 4Vs of Big Data

- **Big Data** refers to **massive** and **complex** datasets, often generated in **real time**.
- It requires specific technologies and infrastructures to be processed efficiently.
- **The 4Vs of Big Data:**
 - **Volume:** The massive amount of data generated by users, connected devices, and systems.
 - **Velocity:** The speed at which data is generated, processed, and analyzed.
 - **Variety:** The different types of data — structured, semi-structured, and unstructured.
 - **Veracity:** The quality, accuracy, and reliability of the data.

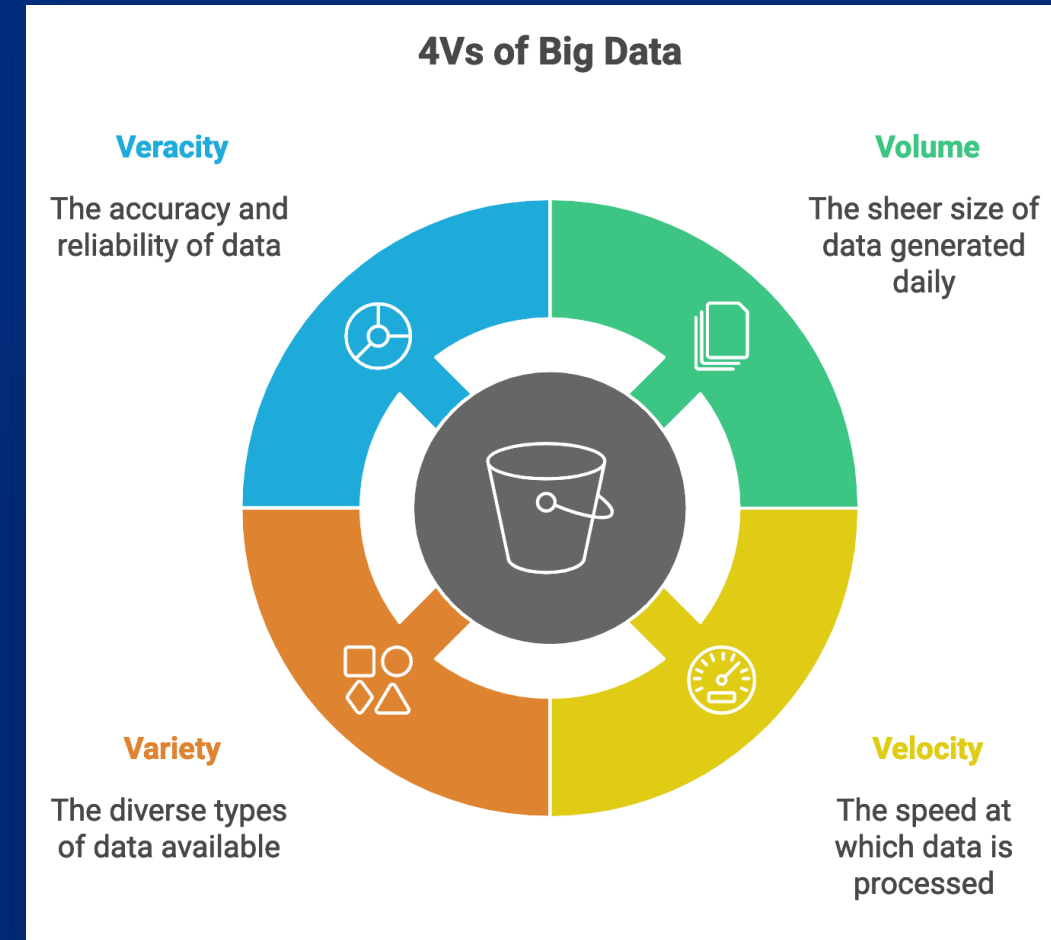
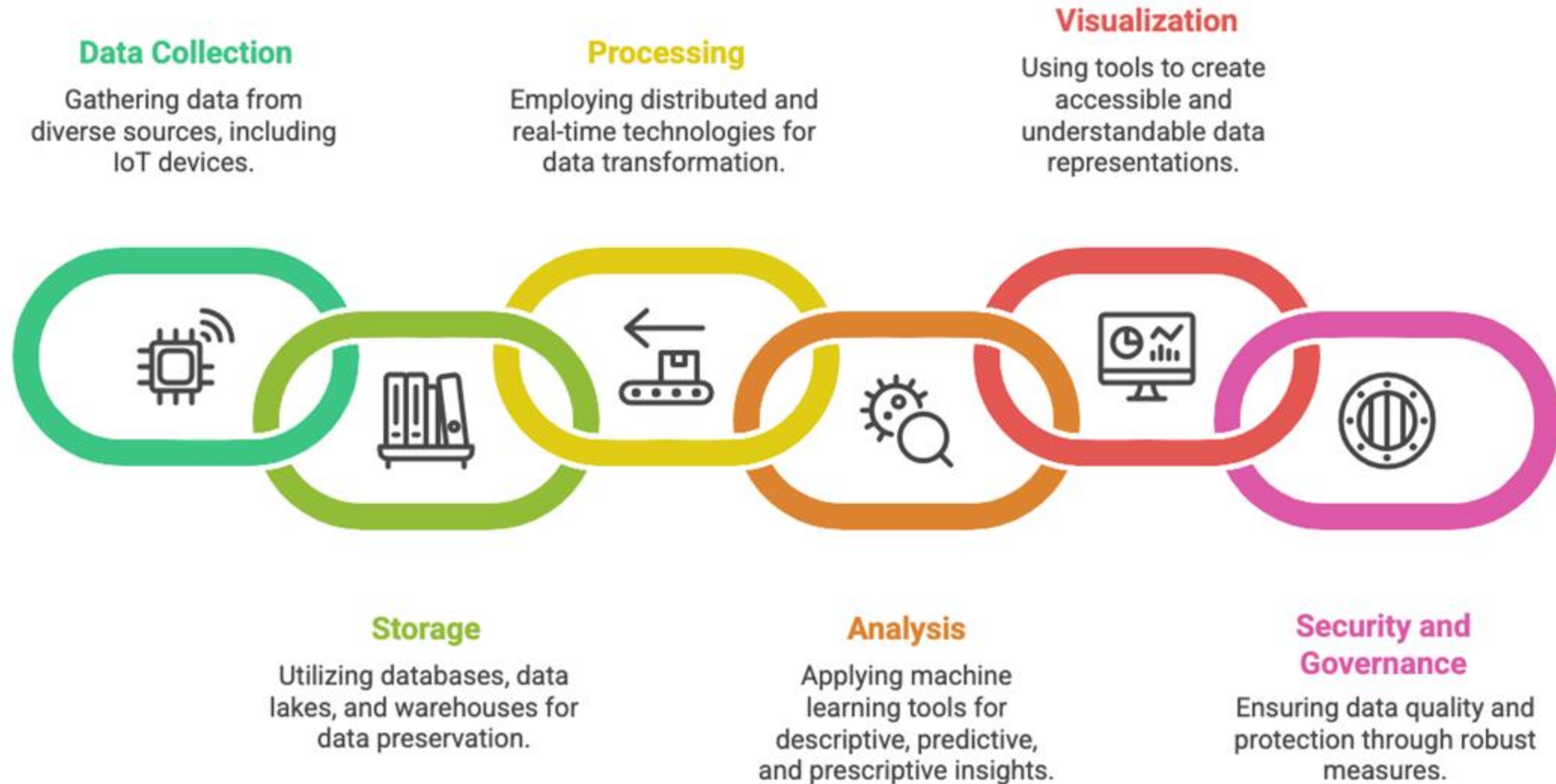


Figure 17. The 4Vs of Big Data

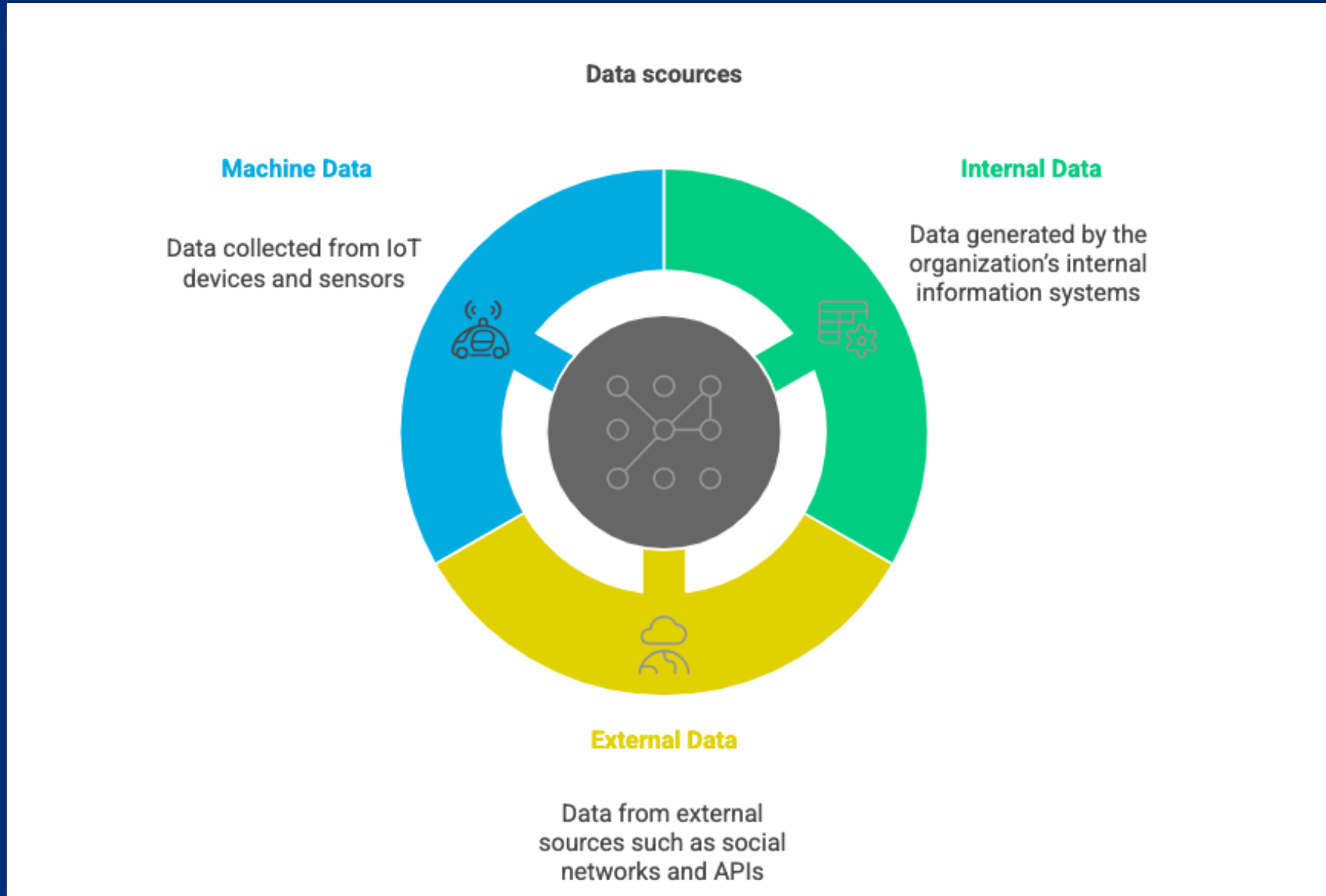
11. The Ecosystem of Big Data and Data Science

- **Data Collection:** Data from internal and external sources, as well as connected devices (IoT).
- **Storage:** Use of databases (SQL, NoSQL), data lakes (centralized storage for structured and unstructured data), and data warehouses (structured data storage such as Amazon Redshift, Google BigQuery).
- **Processing:** Distributed technologies (Hadoop, Spark) and real-time systems (Kafka, Flink).
- **Analysis:** Descriptive, predictive, and prescriptive analysis using machine learning tools (TensorFlow, PyTorch, Scikit-learn).
- **Visualization:** Tools like Tableau, Power BI, and Python libraries such as Matplotlib and Pandas to make insights accessible.
- **Security and Governance:** Ensuring data quality and protection.

Ecosystem of Big Data and Data Science



11.1 Different data sources



11.2 Data Storage types

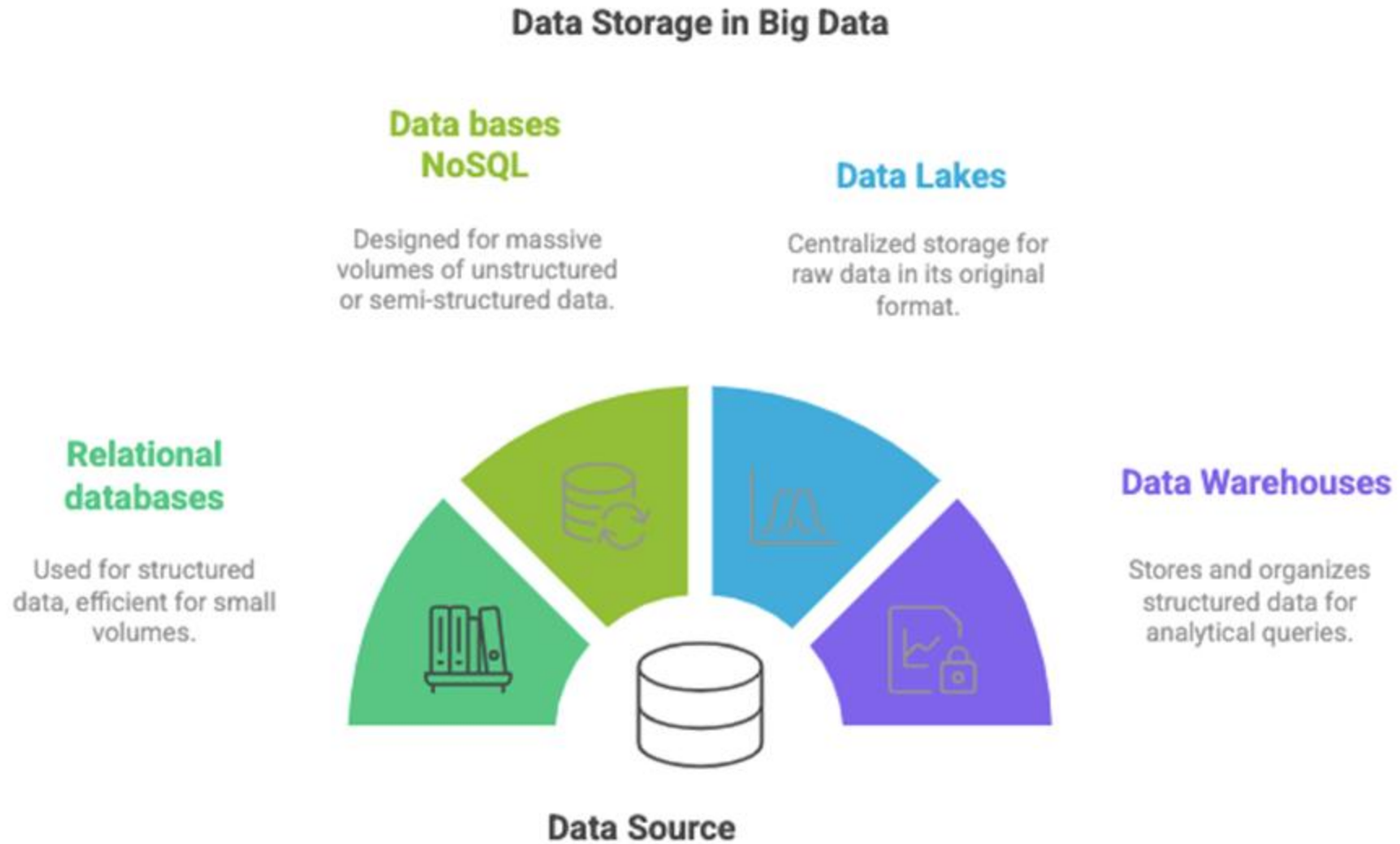


Figure 20. Data storage in Big Data

11.3 Data Processing Tools and Technologies

Data processing in the Big Data ecosystem relies on a combination of tools specialized in handling massive data volumes and performing advanced analytics. These tools include :

ETL (Extract, Transform, Load)

Tools used to extract data from various sources, transform it into a usable format, and load it into data warehouses.

Examples: Talend, Apache NiFi, Informatica.

Distributed Processing Systems

These systems allow large volumes of data to be processed in parallel across multiple machines.

Examples: Hadoop, Spark.

Real-Time Processing Systems

These enable data to be analyzed as it is generated, in real-time.

Examples: Apache Kafka, Apache Flink, Apache Storm.

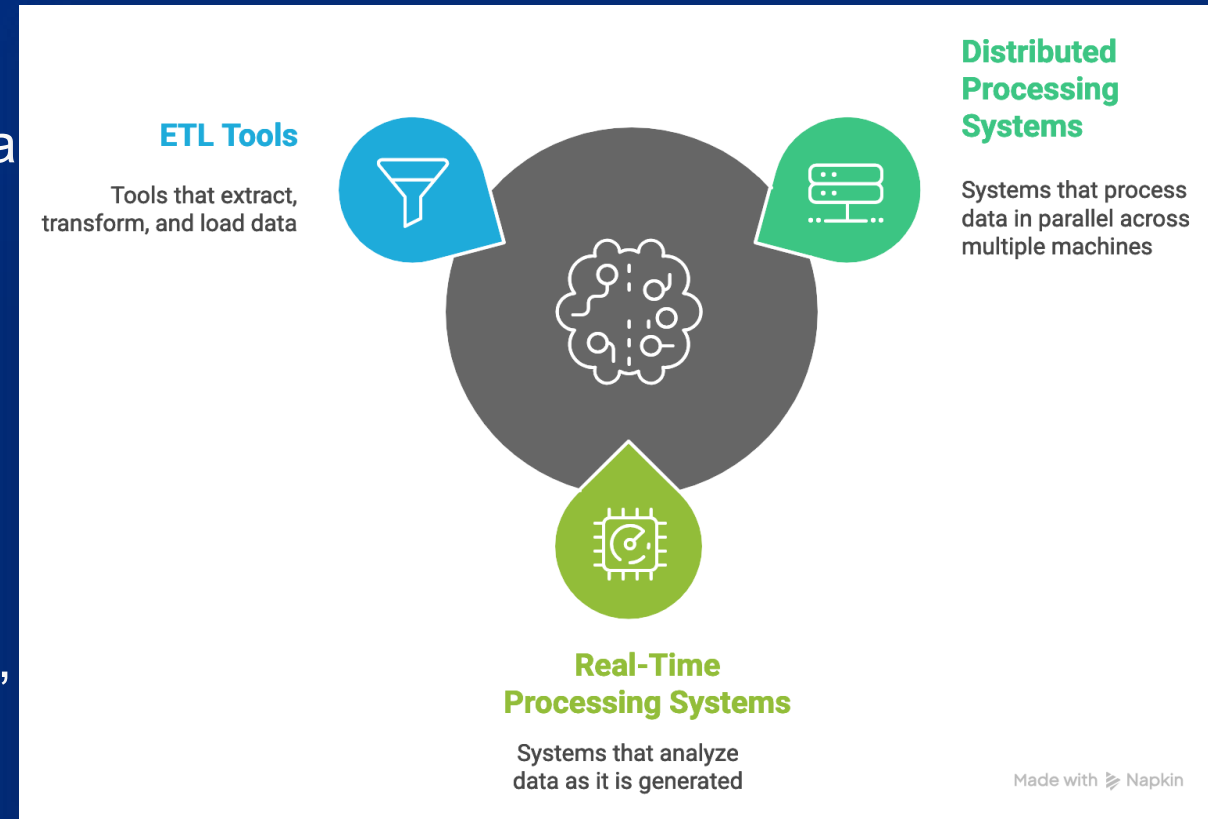


Figure 21. Data processing tools and technologies

11.4 Tools of statistical analysis and Data modeling

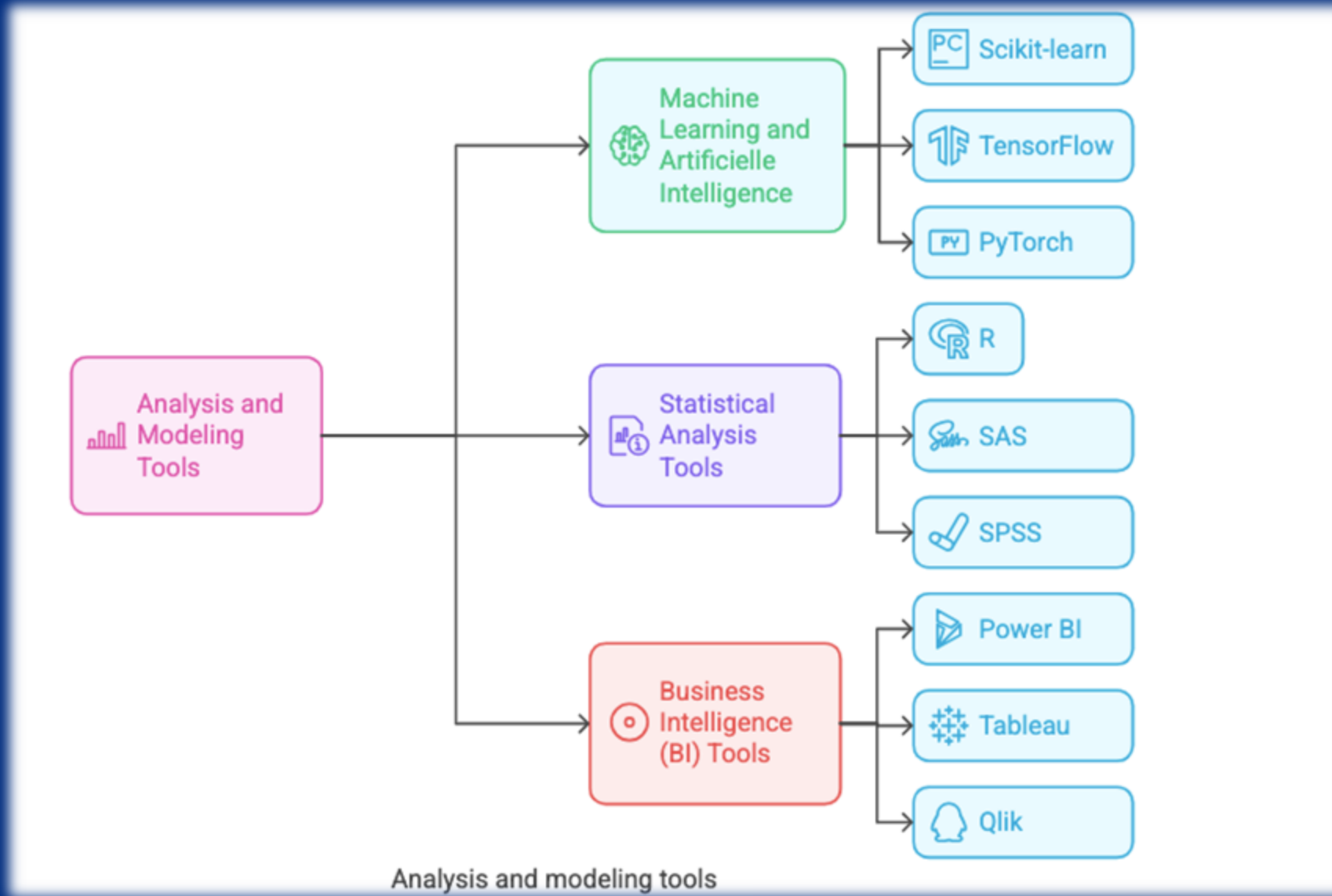


Figure 22. Tools of statistical analysis and Data modeling

11.5 Data Visualization

- **Data visualization** is an essential component for making analytical results understandable and actionable. Visualization tools help transform complex analytical outcomes into charts, dashboards, or interactive maps.
- **Visualization Tools:**
 - **Interactive Dashboards:** Power BI, Tableau
 - **Visualization Libraries:** Matplotlib, D3.js, Plotly

11.6 Data Governance and Security

In the Big Data ecosystem, data governance and security are critical aspects to ensure that data is reliable, high-quality, and well-protected.

- **Data Governance:**
It defines the policies and procedures for managing data properly, including aspects such as **data quality**, **metadata management**, and **regulatory compliance**.
 - *Example:* Implementing **GDPR policies** to protect personal data.
- **Data Security:**
Protecting data from unauthorized access is essential, using techniques such as **encryption**, **access control**, and **system monitoring**.
 - *Example:* Implementing **encryption systems** and **multi-factor authentication** mechanisms.
-  *In data science, strong governance and security ensure that data-driven insights are trustworthy and that sensitive information remains protected.*

11.6 Data governance and security

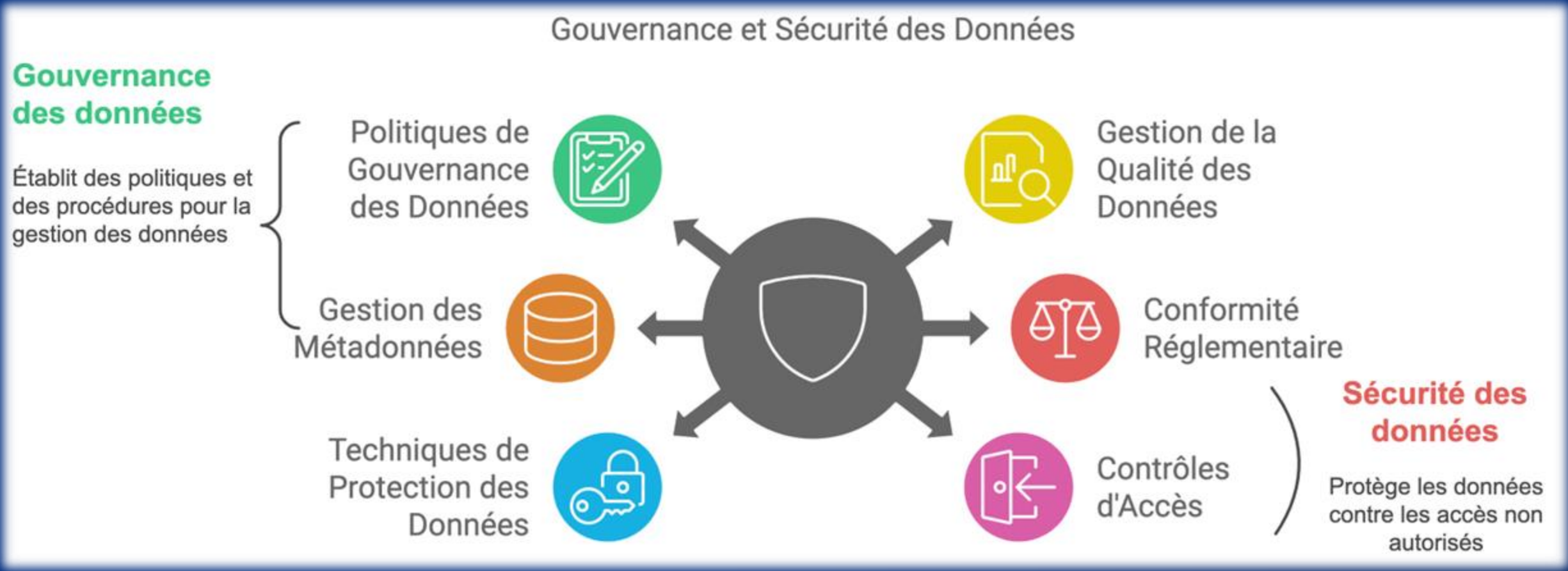


Figure 23. Data governance and security

11.7 Cloud computing in Big data and Data science ecosystem

Cloud technologies play a major role in the Big Data ecosystem. They enable the hosting of storage, processing, and analytical infrastructures in the cloud, providing **virtually unlimited scalability** and **flexible cost management**.

Cloud Service Providers:

Amazon Web Services (AWS): Offers Big Data processing services such as AWS Lambda, Redshift, and S3.

Microsoft Azure: Provides a suite of tools for Big Data, machine learning, and analytics, including Azure Data Lake and Azure Machine Learning.

Google Cloud: Includes services like BigQuery for real-time, large-scale data analysis.

💡 In data science, cloud computing allows organizations to store and process massive datasets efficiently without the need for local infrastructure.

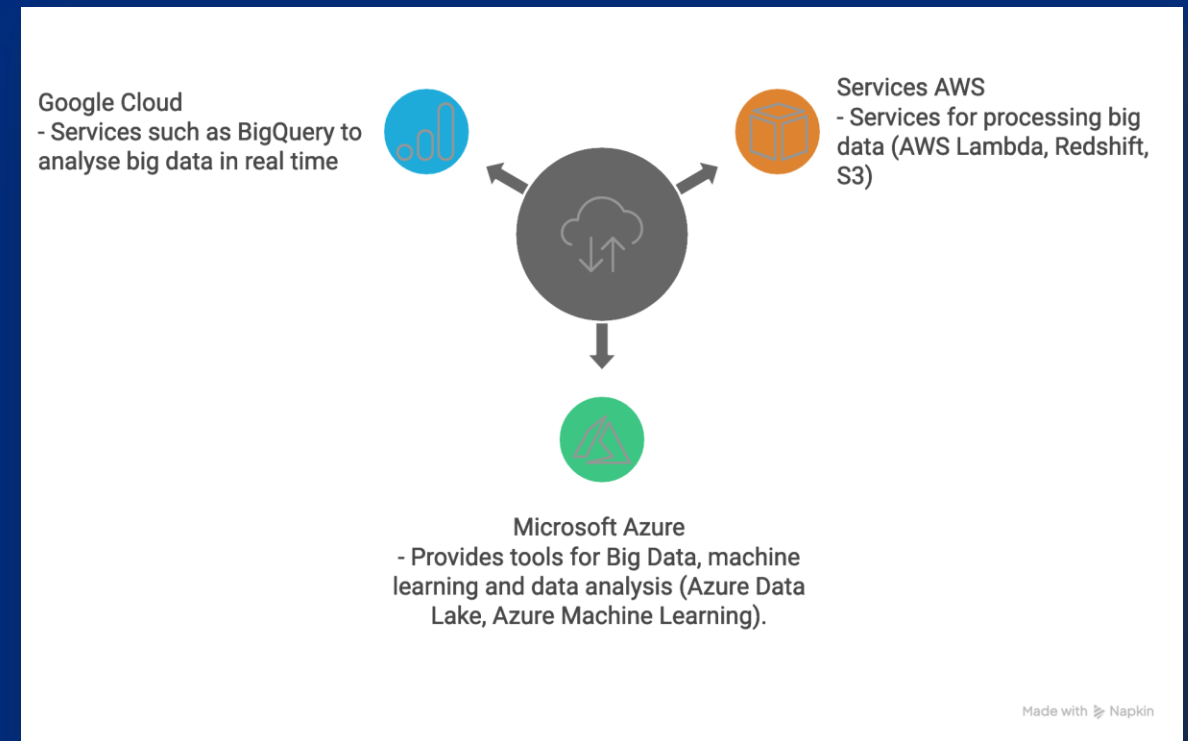


Figure 24. Cloud computing in the big data and data science ecosystem