Hi! PARIS

PARIS ARTIFICIAL INTELLIGENCE FOR SOCIETY

HEC PARIS

INSTITUT POLYTECHNIQUE DE PARIS

Inria

Welcome to the Course 🚀

Data Science Awareness

Lecturer - Soobash Daiboo - Senior ML Engineer

Work Experience

Observatoire de Paris | PSL

ENGIE

Capgemini

Hi! PARIS

Contact Info

www.linkedin.com/in/soobash-daiboo/

www.github.com/soobash

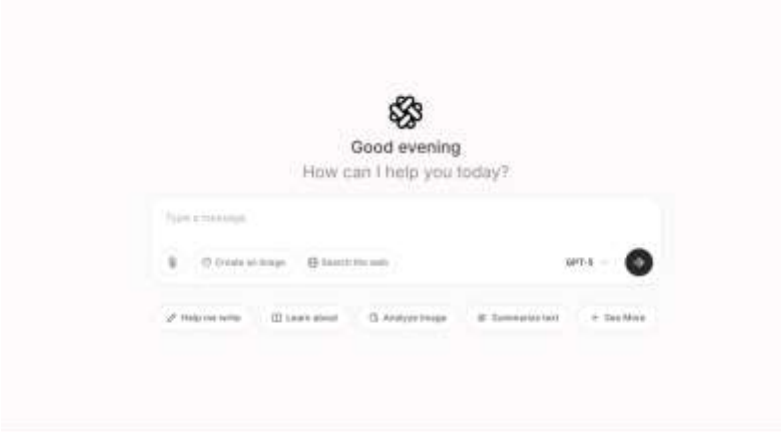
soobash@gmail.com

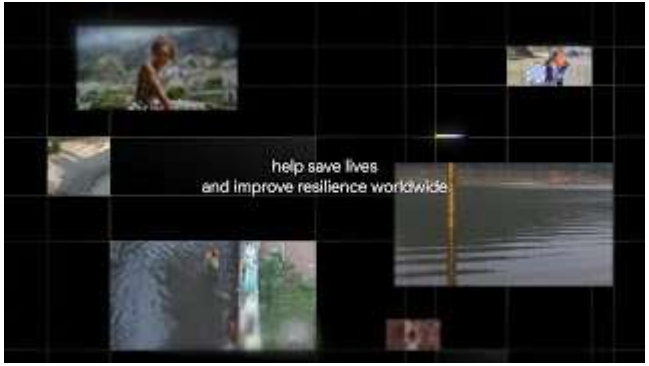
Agenda

I. Introduction to Data Science

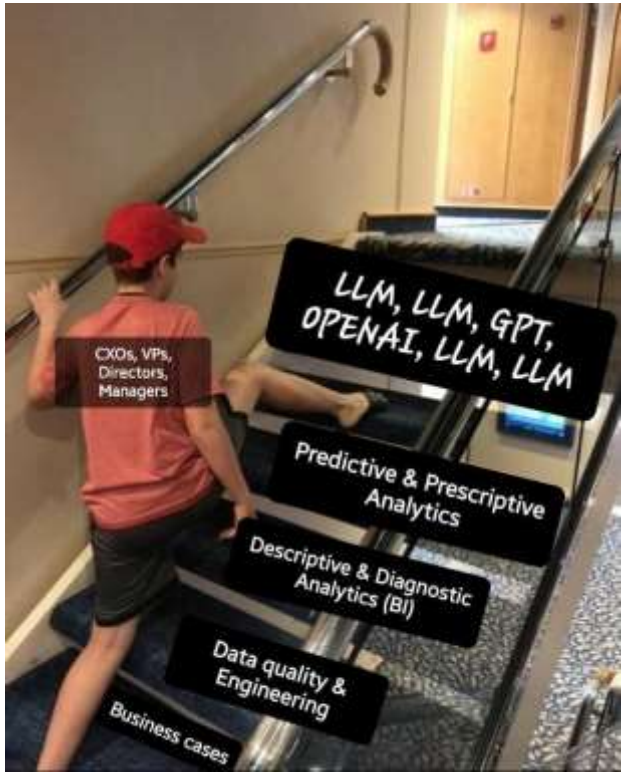
Why having Data Science skills is important ? (1/2)

1 Because DS is everywhere





Why having Data Science skills is important ? (2/2)



2 You need to be aware of the models that are adapted to specific business problems

What is Data Science ?

Data science is the study of **data** to extract meaningful **insights** for **business**.

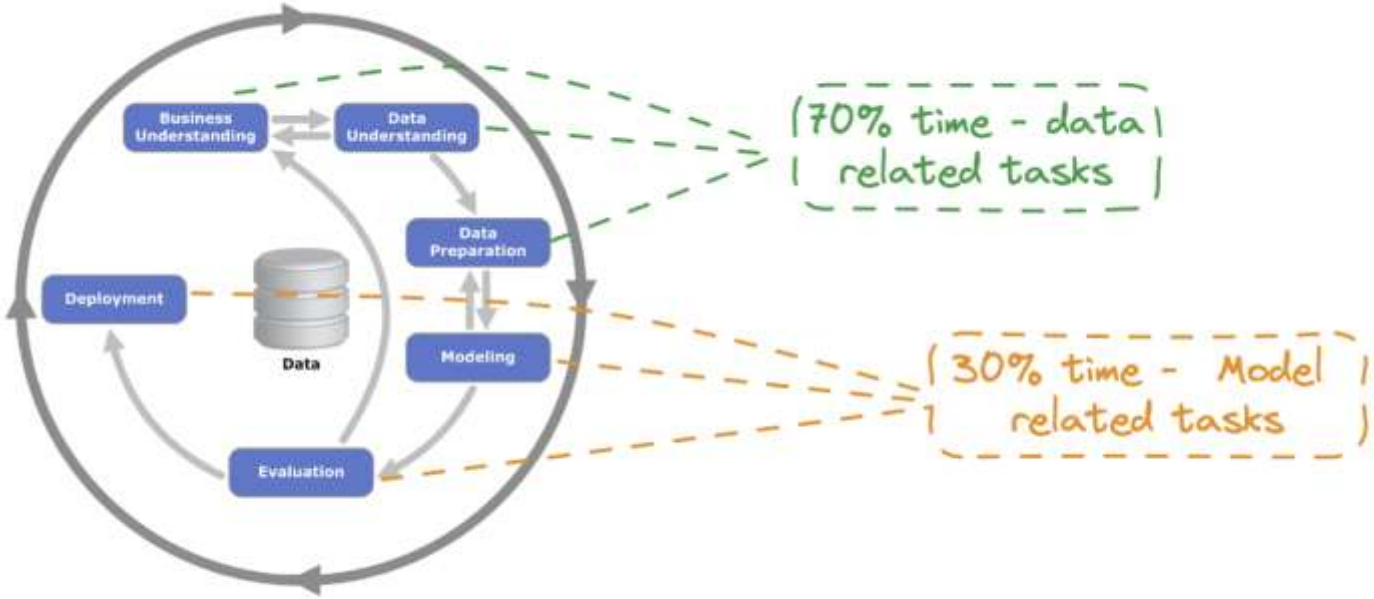
What is being a Data Scientist ?

A **data scientist** is a professional who creates programming code and combines it with statistical knowledge to create insights from data.

A mindset:

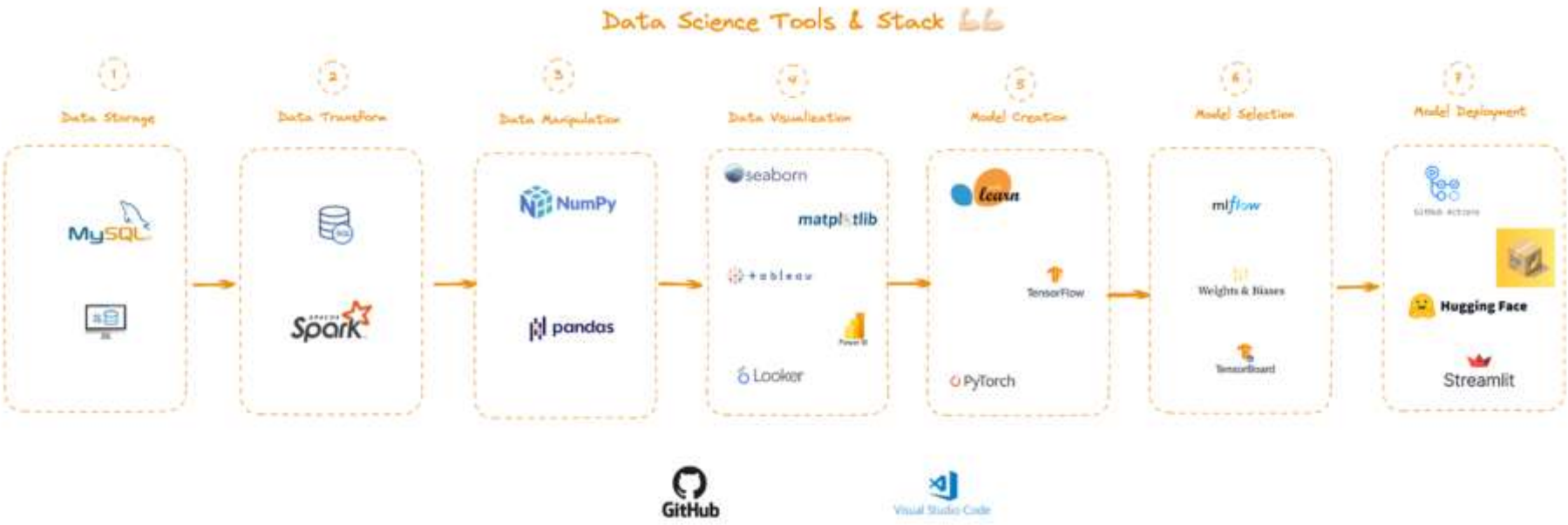
- Learn to fail - iterative process - test & learn (next slide)
- “All I Know Is That I Know Nothing” - Socrates
- Look for answers in a proactive manner:
  - Open-source
  - Stackoverflow
  - ChatGpt
  - Mentors and colleagues
  - Documentation
  - Articles

Data Science Lifecycle



Cross-industry standard process for data mining

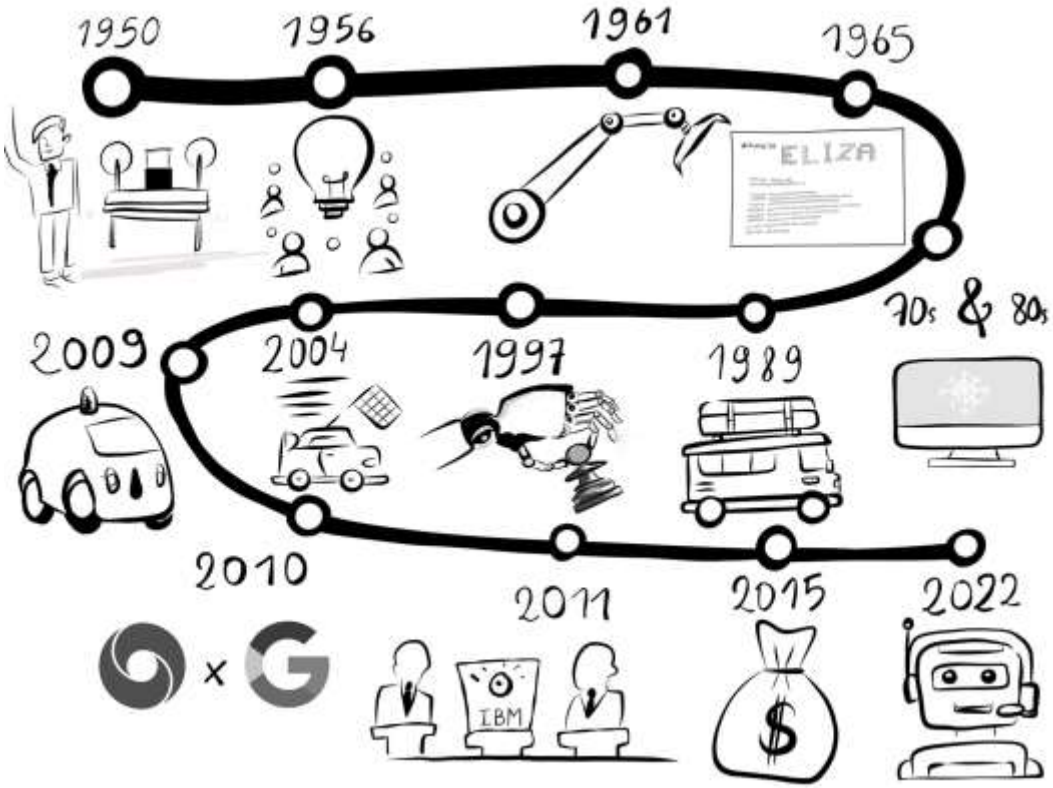
Data Scientists Tools & Stack :



II. History of Data Science & AI

> *"The history of artificial intelligence dates back to the **1950s**, when computer scientist Alan Turing proposed that machines could think and perform tasks like humans. Since then, AI has evolved rapidly, with advancements in areas such as natural language processing, machine learning, and robotics. AI has also enabled the development of self-driving cars and intelligent personal assistants. AI is now being used in a variety of industries, from healthcare to finance and beyond."* **GPT3** (an Artificial Intelligence from the Company OpenAI) - https://chat.openai.com/chat

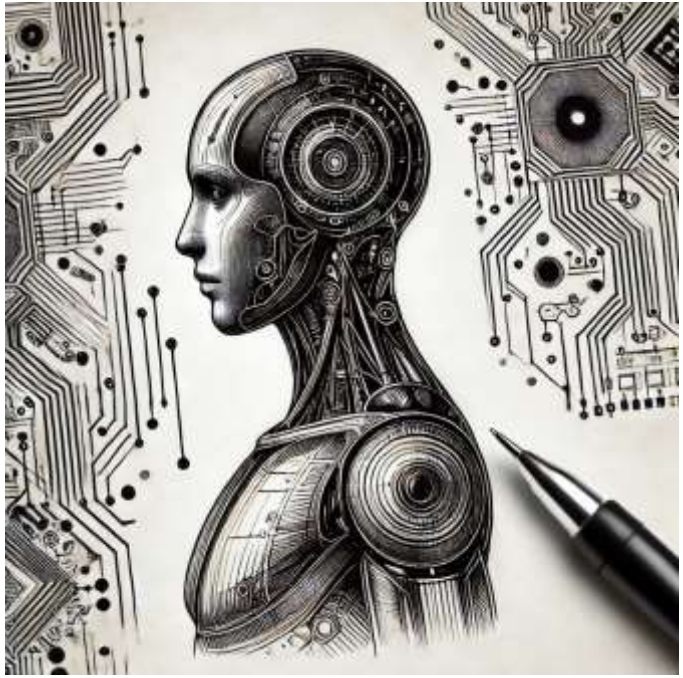
The AI Timeline



What is next ? AGI (Artificial General Intelligence) - 1/2

"AGI will be reached in 5 years, give or take, maybe slightly longer - but no one knows exactly when or what it will mean for society." - Sam Jan, 2024

"AGI is inevitable, but not coming next year nor with LLMs" - Yann June, 2024



LLMs are Impressive… Sometimes

Human: If Tom is 3 inches taller than Dick, and Dick is 2 inches taller than Harry, how much taller is Tom than Harry?
AI: Tom is 5 inches taller than Harry.

Human: Can Tom be taller than himself?
AI: No, Tom cannot be taller than himself.

Human: Can a sister be taller than her brother?
AI: Yes, a sister can be taller than her brother.

Human: Can two siblings each be taller than the other?
AI: Yes, two siblings can each be taller than the other.

Human: On a map, which compass direction is usually left?
AI: On a map, north is usually to the left.

Human: Can fish run?
AI: No, fish cannot run.

Human: If the door is locked, what must you do first before opening it?
AI: If the door is locked, you must first unlock it before opening it.

Human: Which was invented first, cars, ships, or planes?
AI: Cars were invented first, followed by planes and ships.

What is next ? AGI (Artificial General Intelligence) - 2/2

Yann LeCun in • Abonné
VP & Chief AI Scientist at Meta
2 mois •

Nice article in Financial Time where I explain that Auto-Regressive LLM are insufficient to reach human-level intelligence (or even cat-level intelligence).

But alternative architectures that I call "objective driven" may reach human-level intelligence one day.
They use world models based on JEPA (Joint Embedding Predictive Architectures, which are not generative).
With this, we may have systems that
1. understand the physical world
2. have persistent memory
3. can reason
4. can plan, perhaps hierarchically.
Four essential characteristics necessary for intelligent behavior, which humans and many animals exhibit.

FT article

Why Data Science is trending Now ? 🧑‍🔬🖥️🔍

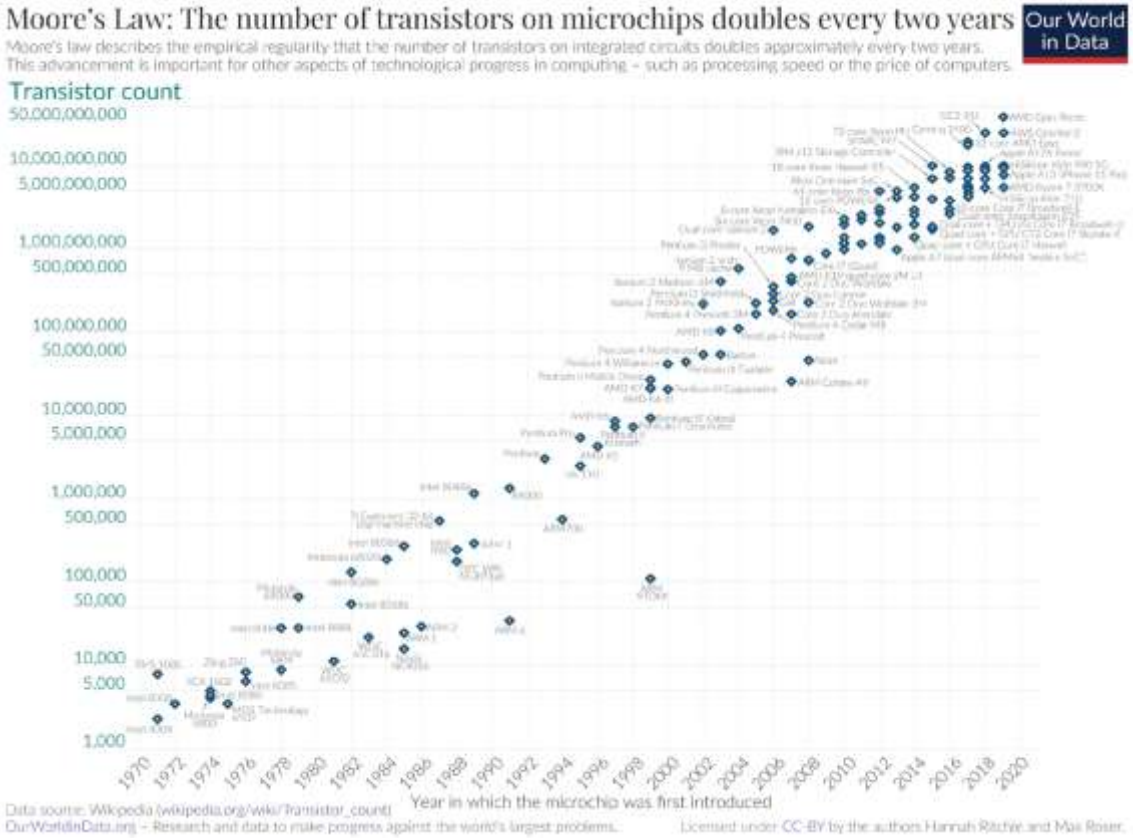
The 3 forces that brought Data Science to Life:

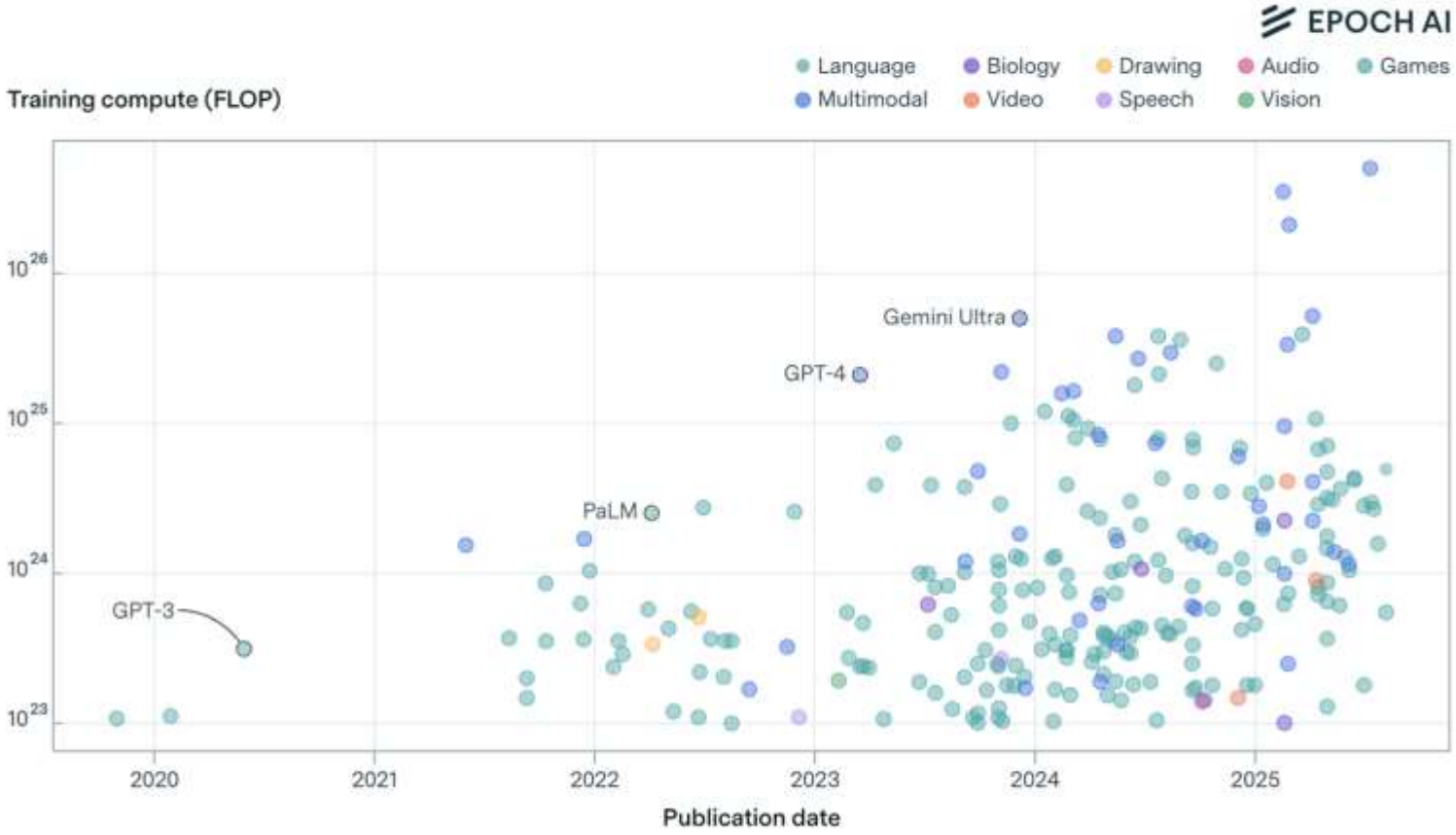
+

+

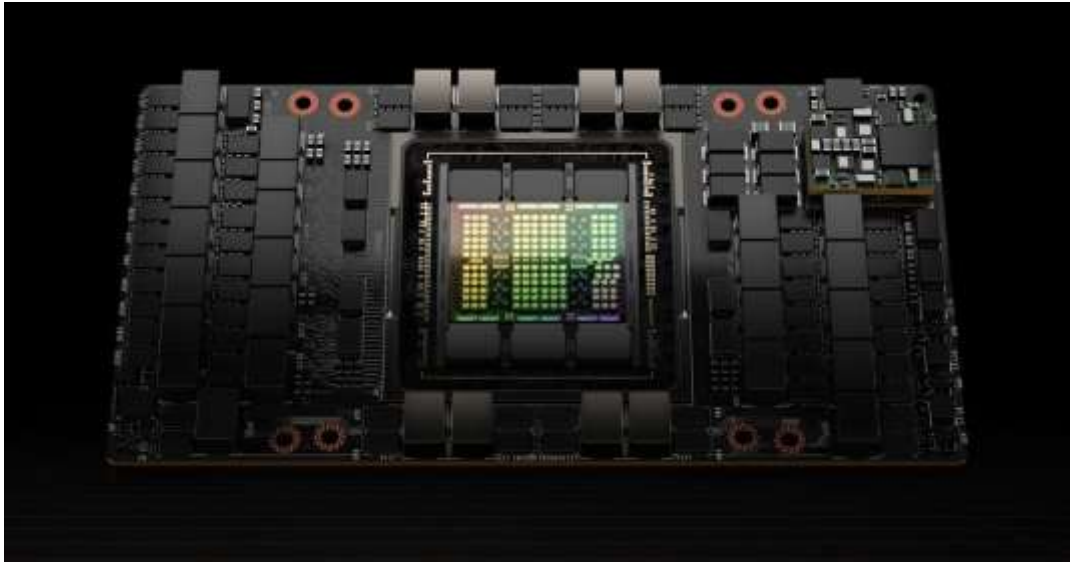
I – Computing Power

- One of the best way to understand the increase in computational powers in the recent years is to look at Moore's Law phenomenon.
- Moore's Law is an empirical observation made by Intel co-founder Gordon Moore in 1965 that the number of transistors on an integrated circuit board would double approximately every two years. This has held true for the past several decades and has resulted in exponential growth in computing power. Moore's Law has been an important driving force in the development of the computer industry, with companies constantly striving to create smaller and faster processors to stay ahead of the competition.




EPOCH AI
Language
Biology
Drawing
Audio
Games
Multimodal
Video
Speech
Vision
Training compute (FLOP)
10^{26}
10^{25}
10^{24}
10^{23}
Gemini Ultra
GPT-4
PaLM
GPT-3
2020
2021
2022
2023
2024
2025
Publication date


WHEN EVERYONE DIGS FOR GOLD
SHOVELS $99
NVIDIA
Microsoft
Meta
SELL
SHOVELS



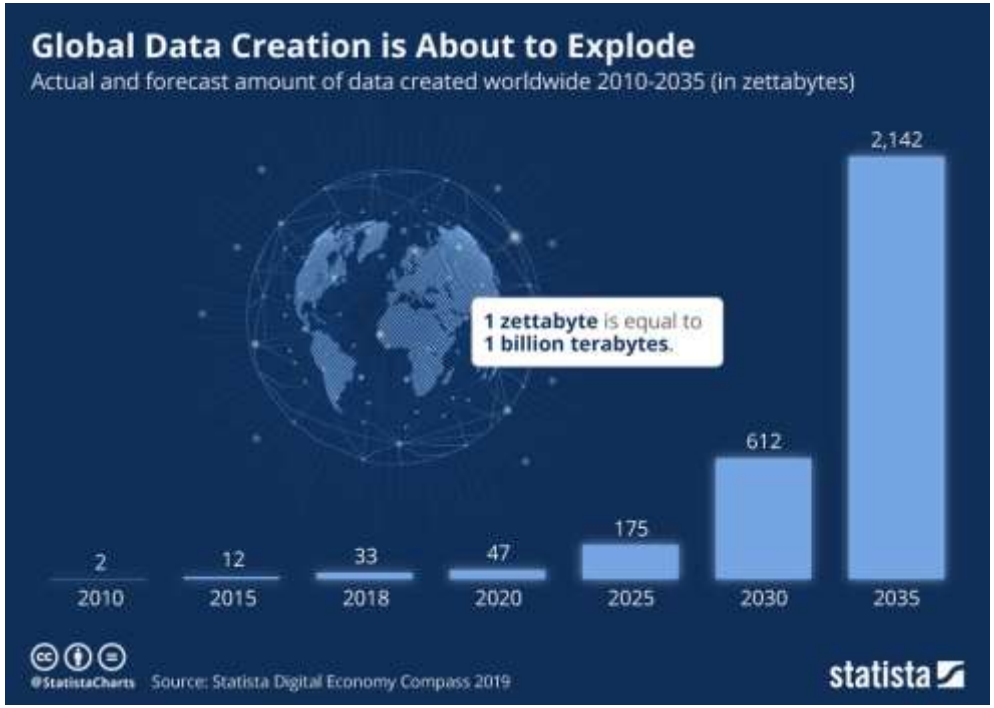
a Hugging Face Space by open-llm-leaderboard

Jensen Huang-Favoring Moore's Law Over Customer Feedback

https://www.youtube.com/watch?v=6Uc-EiQ2xnU

II – Data Boom

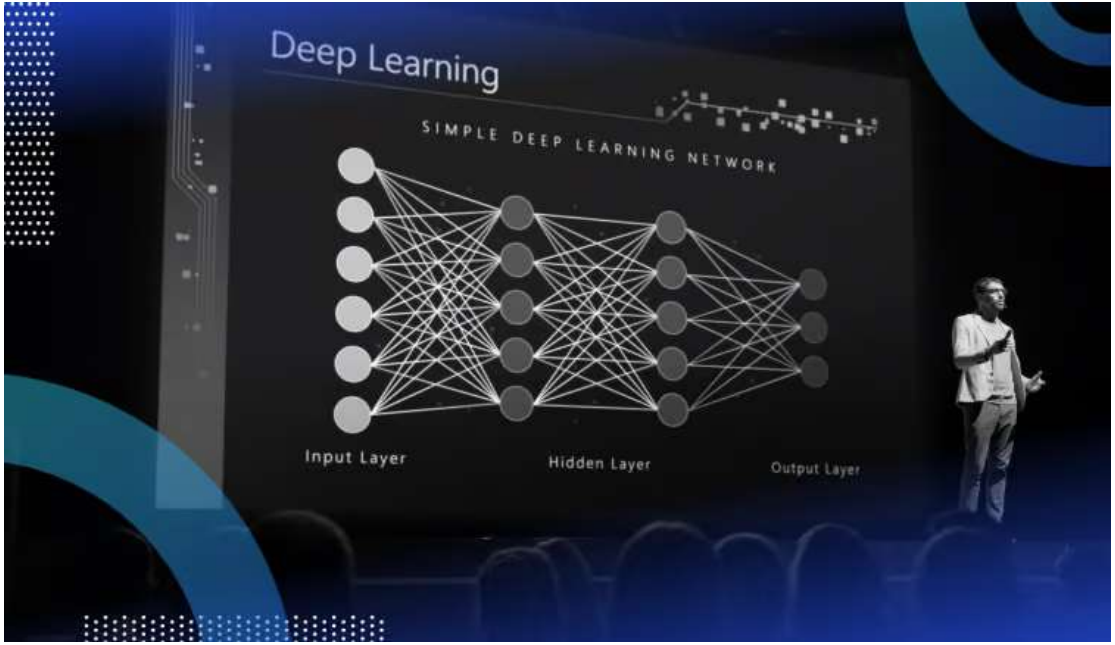
- To put the rapid growth of data into sharper focus, years ago (2013) IBM reported 90% of the data in the world today had been created in the last two years alone. International Data Corporation (IDC) forecasts that by 2025, global data will grow to 163 zettabytes (or a trillion gigabytes). That's 10 times the 16.1 zettabytes of data generated in 2016.



every minute of the day:
•Snapchat users share **527,760** photos
•More than **120** professionals join LinkedIn
•Users watch **4,146,600** YouTube videos
•**456,000** tweets are sent on Twitter
•Instagram users post **46,740** photos

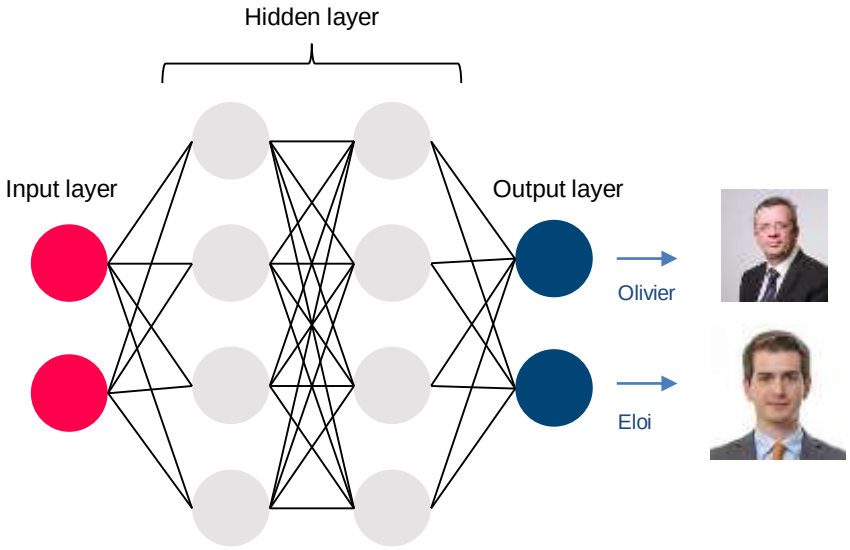
TikTok has **45.26 million** daily active users in 2022

III – Better Algorithms & Tools



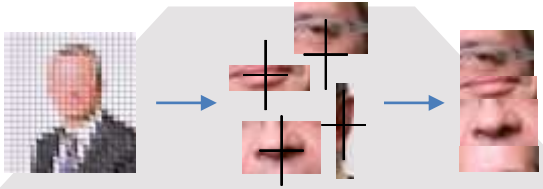
Deep Learning is inspired by the structure and function of the brain

Learning
A neuronal network that "learns" faces needs to train on million examples to be able to identify a face in a crowd or saturated landscape



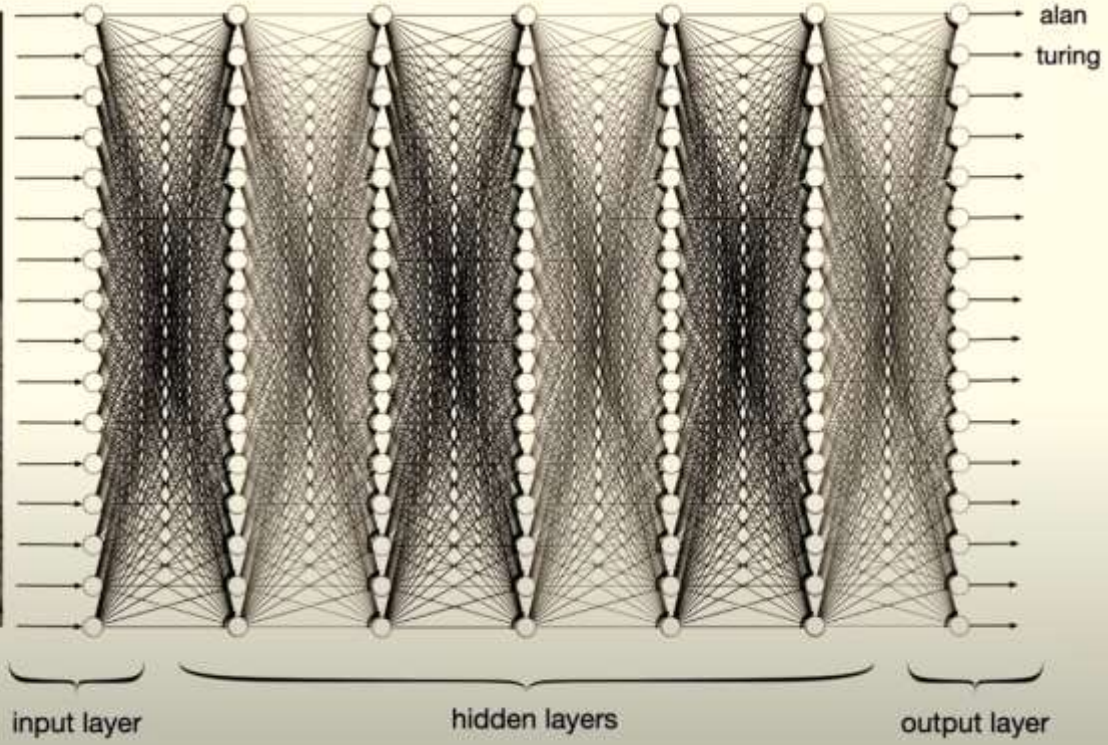
Recognition

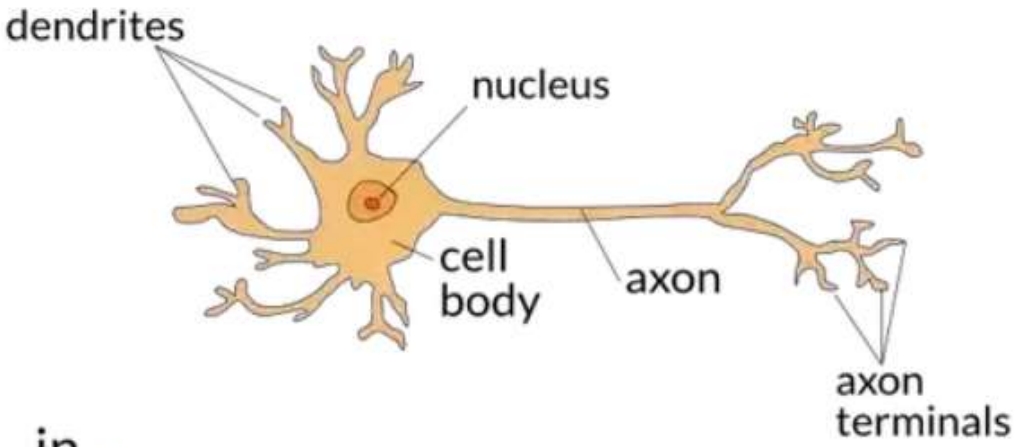
A picture of a face injected in the network is sequentially transformed by the different neuronal layers to be able to identify the face as an output

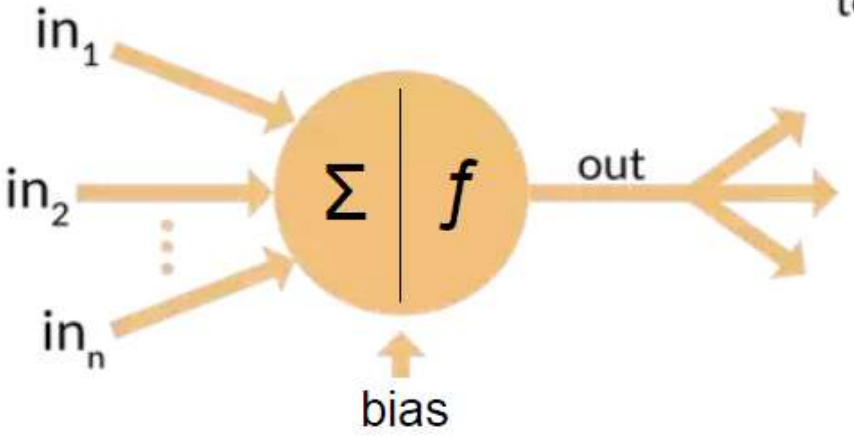


Each layer progressively identifies more complex face characteristics

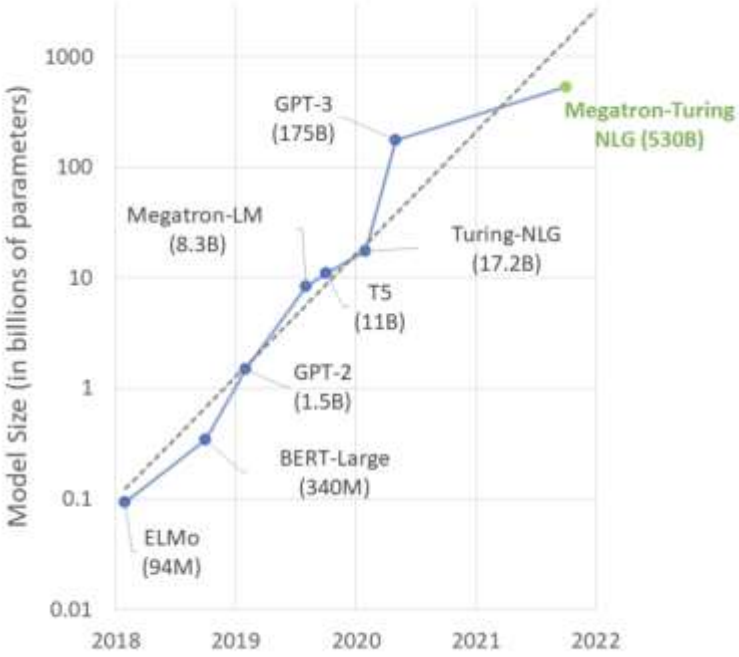
Neural Networks




dendrites
nucleus
cell
body
axon
axon
terminals


in1
in2
inn
Σ
f
out
bias

Large Language Models: A New Moore's Law?



Here is an example of top deep learning models across times and the number of parameters used for their predictions ➞

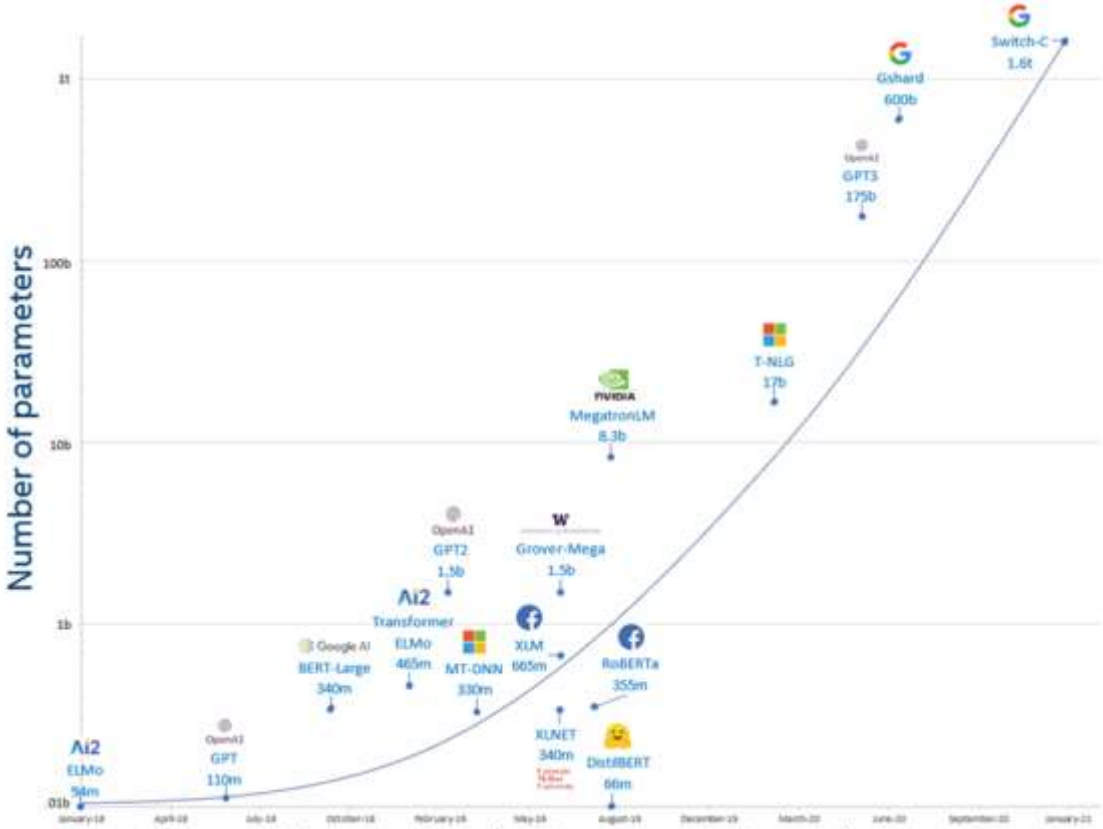
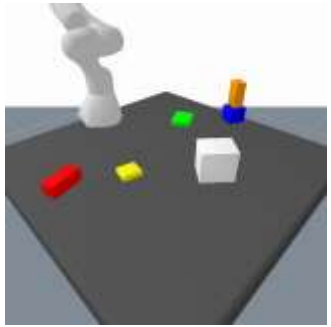
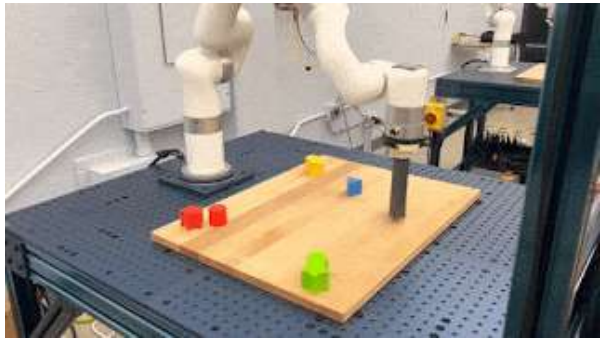


Figure 1: Exponential growth of number of parameters in DL models

Advanced models, example in 2022:

PaLM-E: An embodied multimodal language model



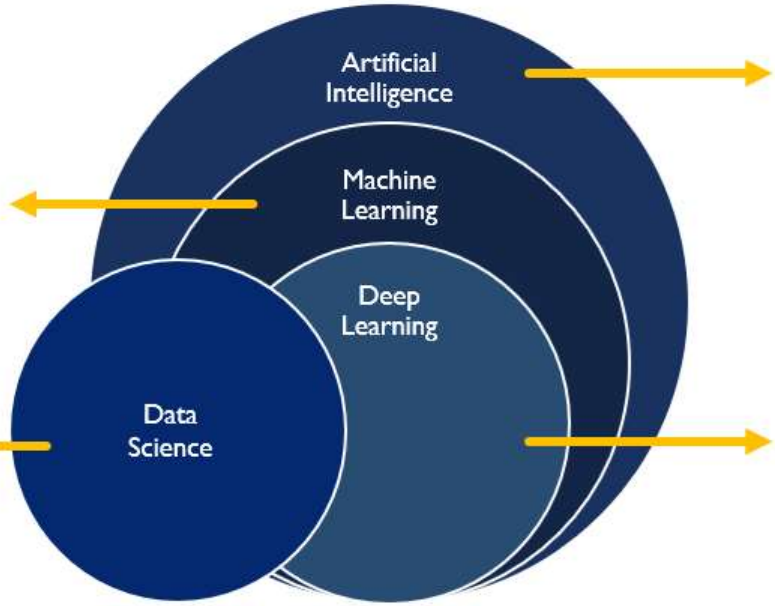


III. Key Concepts Behind the term « Data Science »

Data Science vs Machine Learning vs Deep Learning vs AI

Machine Learning is a subset of AI, and builds a model based on training data to make predictions

Data Science is a subset of AI. It is an area of statistics, scientific methods, etc. to extract meaning and insights from data..

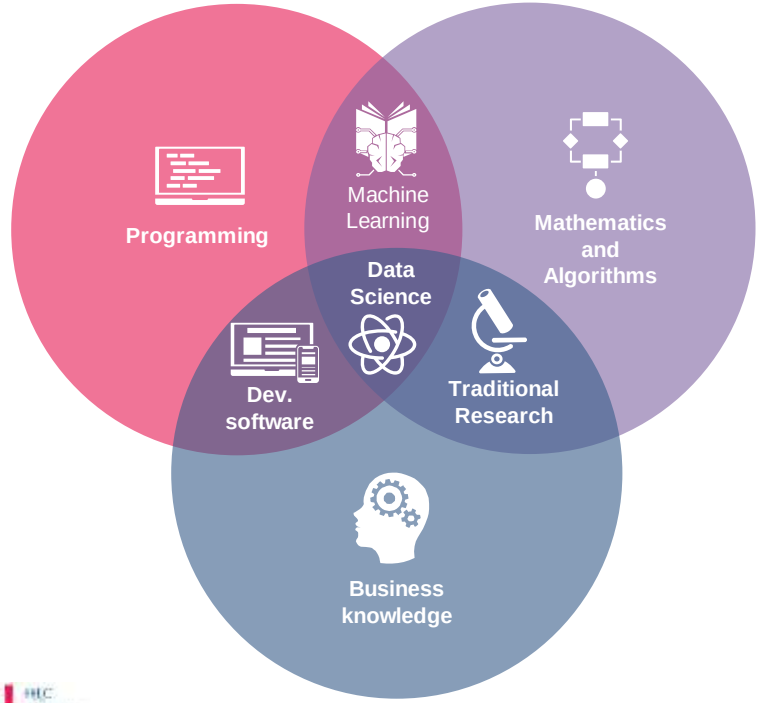


Artificial Intelligence means creating smart machines to mimic human behavior

Deep learning is a subset of ML, a class of ML algorithms to solve complex problems.

Data Science is at the intersection of several disciplines, it allows us to solve problems and make predictions from data

KEY SKILLS



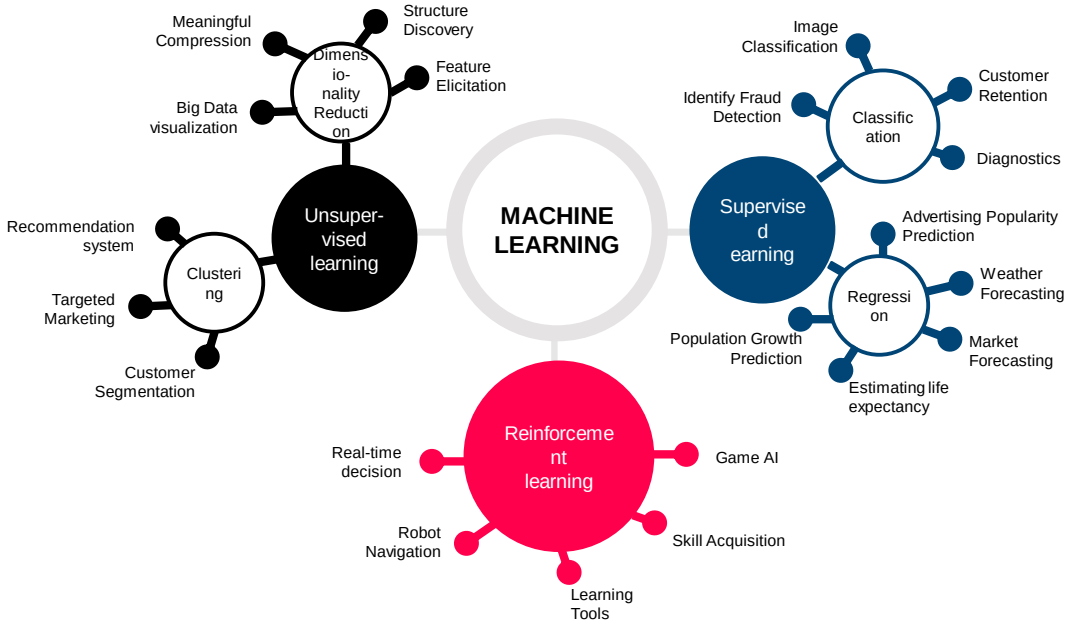
COMMUNICATION

Know how to communicate analyses and model outputs with the right visualizations

KNOW-HOW

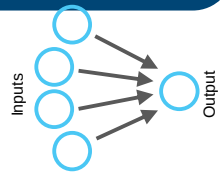
Knowing how to translate data into added value, curiosity, critical thinking, entrepreneurial spirit

The 3 MAIN types of Machine Learning



Supervised learning

learn the relationship between input and output variables using annotated examples (e.g., predicting the value of a house)

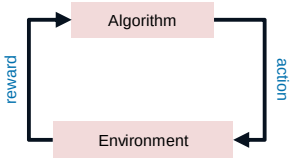


Unsupervised learning

Explore the data and find a structure, without annotated examples (e.g. customer segmentation)

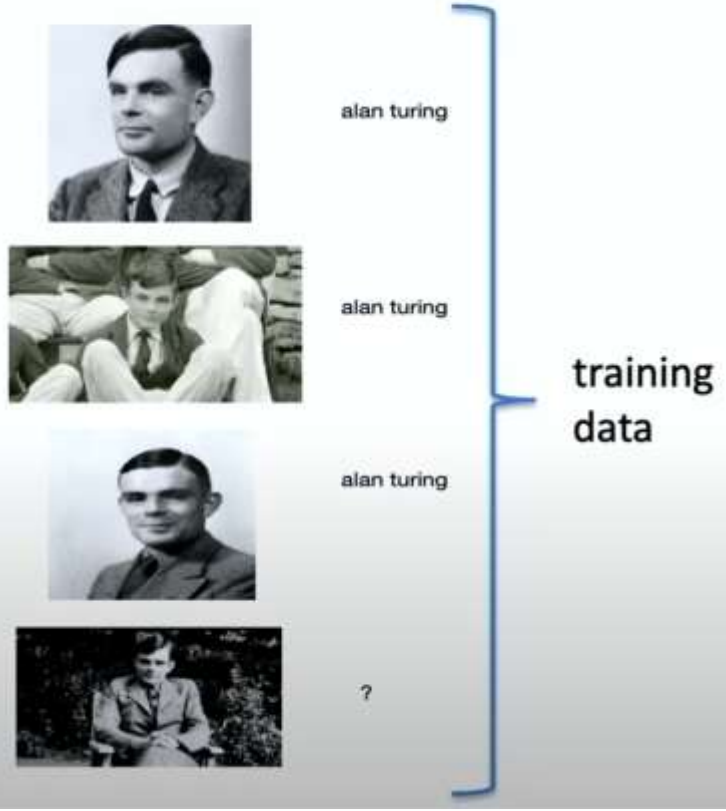
Reinforcement learning

Learn by interacting with the environment and try to maximize a reward (e.g. automatic robot vacuum cleaner)



Machine Learning

- Goal is to learn a mapping from **inputs** to **outputs**
- Simplest technique is **supervised learning**
- Uses **training data**
- How to do the training?
  - **neural nets** and **deep learning** a popular current approach



Technical steps specific to the development of a data science project

DESIGN AND QUALIFICATION → **DEVELOPMENT OF A PROTOTYPE AND INDUSTRIALIZATION** → **RUN**

1

Collecting data

Internet of Things
Programming interface (API)
Open data
Web content extraction
Databases

2

Refine the framework

Refine the business purpose
Type of model
Evaluation measures
Nature of variables

3

Preparing the data

Data mining and cleaning
Data visualization
Variable normalization
Feature engineering

4

Training and evaluating a model

Training
Model evaluation (metrics)

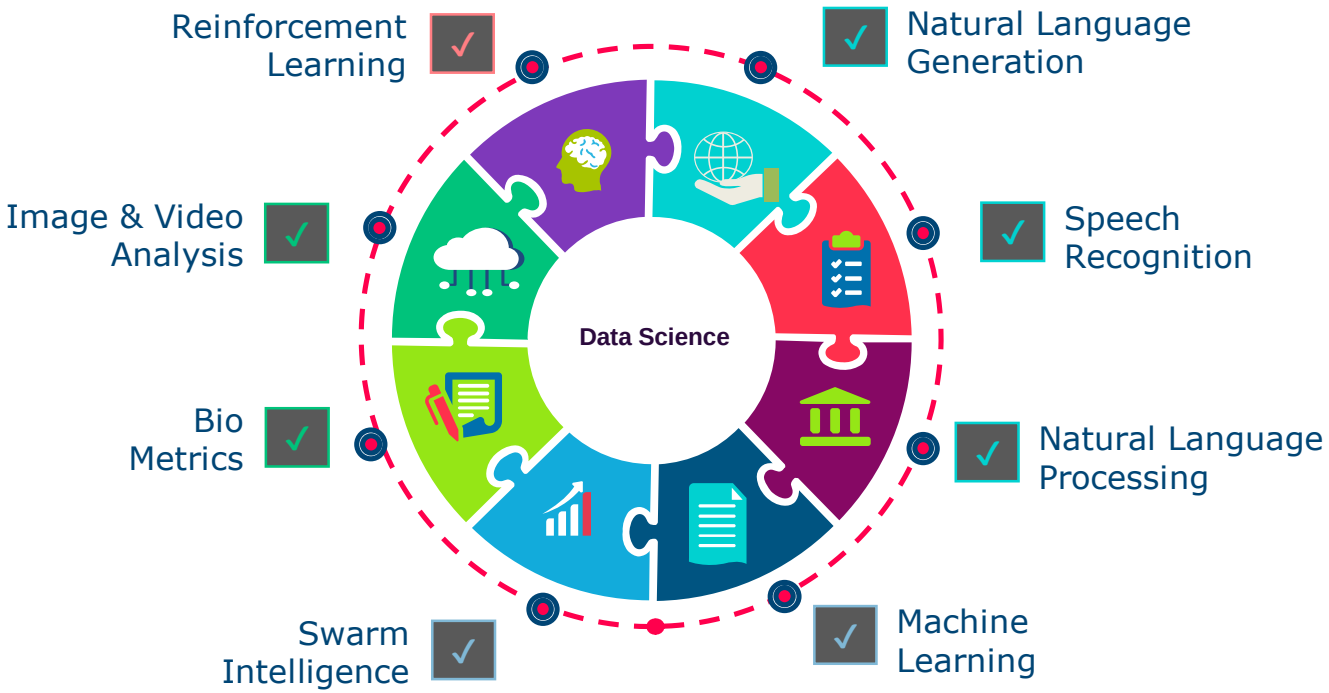
Iterate

5

Industrialize/Deploy

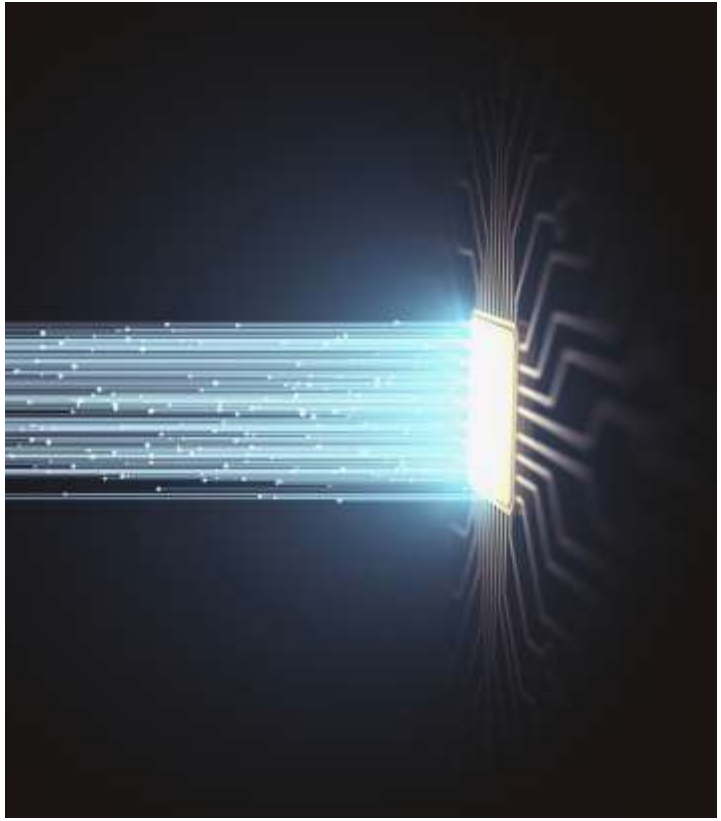
Code optimization
Automation of data ingestion
Integration into business processes

Overview of Data Science concept application



Key success factors for a data project

- ACCESSIBLE DATA SOURCES
- QUALITY DATA
- ANSWER TO A BUSINESS NEED
- INVOLVEMENT OF THE BUSINESS
- AGILE FRAME - FAIL FAST



Data Quality

Example on US Census Dataset: link

IV. Data Science Use Cases

With a practical use in the daily life

TRANSLATION

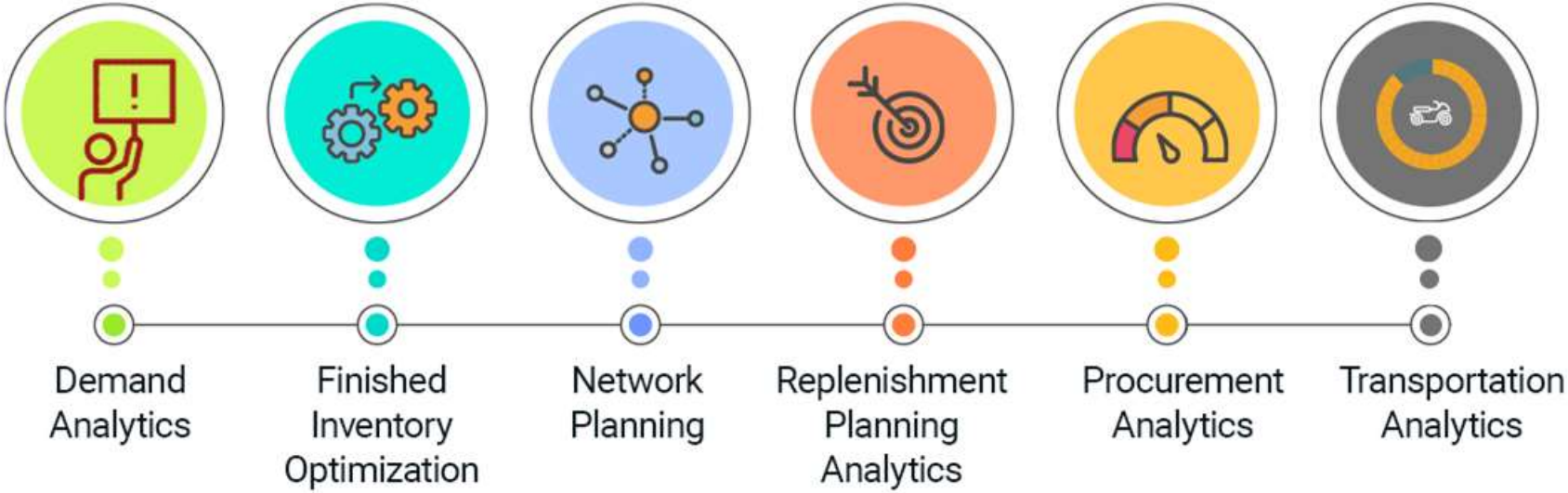
CHATBOTS

VOICE RECOGNITION

FACIAL RECOGNITION

OBJECT DETECTION

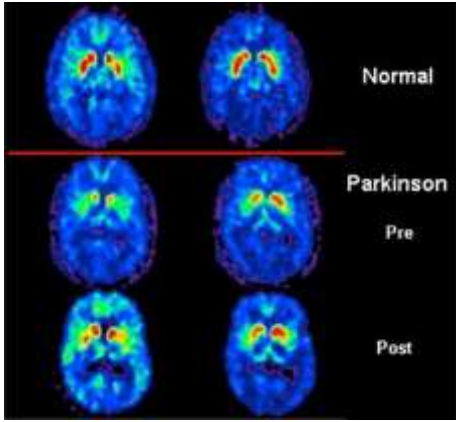
Example of applications of Data Science in Supply Chain



Two Applications use cases in the Hi! PARIS Center:

1) Fire Detection using computer vision:

2) Parkinson Detection on medical images:



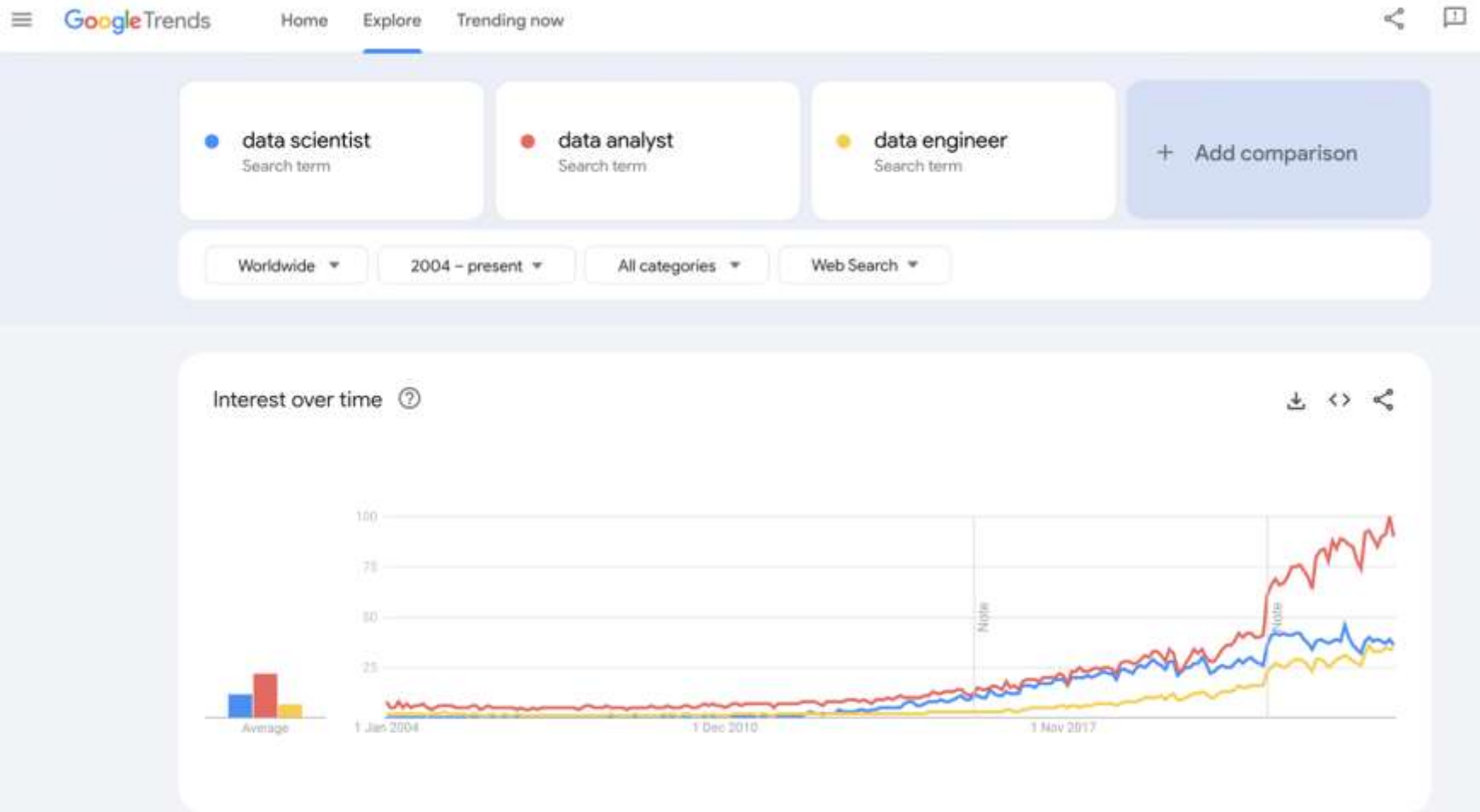
V. The Data Science Market

Evolution of the field

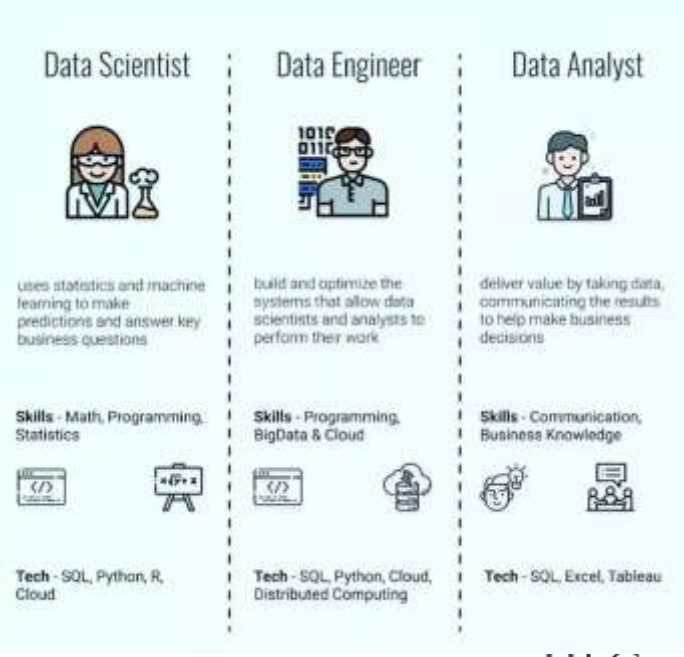
- Employment □ :
  - AI and machine learning jobs have jumped by almost 75% over the past 4 years and are poised to keep growing.
  - The field of artificial intelligence has a tremendous career outlook, with the Bureau of Labor Statistics predicting a 31.4% increase in jobs for data scientists and mathematical science professionals.
- Investments 📈:

🎯 French Government Investment Strategy 2025 in AI:

- 2022: €1.5 billion
- 2025: €2.2 billion
- increase of 46.6%
- 109 billion Euros in AI infrastructure projects part of FRANCE 2030 investment plans

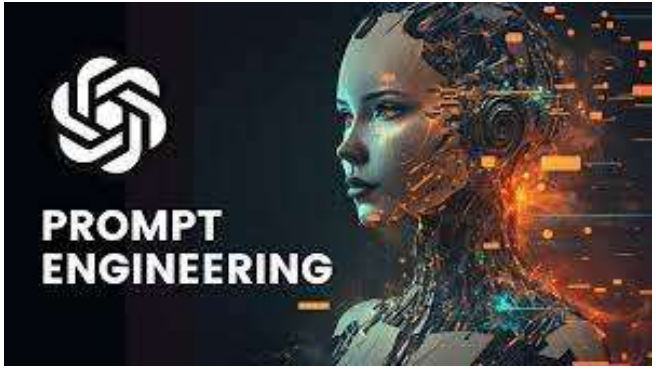

Google Trends
Home
Explore
Trending now
data scientist
Search term
data analyst
Search term
data engineer
Search term
Add comparison
Worldwide
2004 – present
All categories
Web Search
Interest over time
100
75
50
25
Note
Note
Average
1 Jan 2004
1 Dec 2010
1 Nov 2017

Increasing number of Data Science related jobs



New Jobs

New Jobs appearing



G

code a python function to clean the missing value in a dataset

Sure! Here's a function that takes a pandas DataFrame and a list of column names as input and replaces missing values in those columns with the mean of the non-missing values:

Copy code

```
def clean_missing(df, columns):
    for col in columns:
        mean = df[col].mean()
        df[col].fillna(mean, inplace=True)
    return df
```

You can call this function like this:

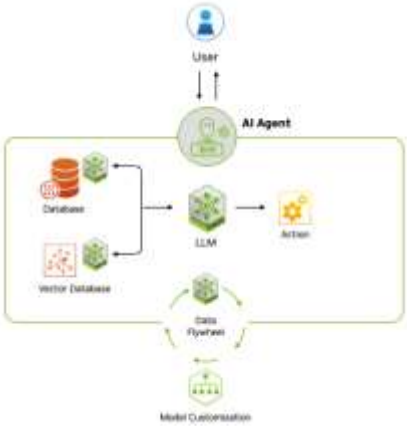
Copy code

```
df = clean_missing(df, ['column1', 'column2', 'column3'])
```

This will replace missing values in the columns 'column1', 'column2', and 'column3' with the mean of the non-missing values in each column.

You can also use other strategies to impute missing values, such as using the median or the most frequent value. Let me know if you have any questions!

New ways of doing business with Automation (Eg. Agentic AI, Muti-agentic AI)

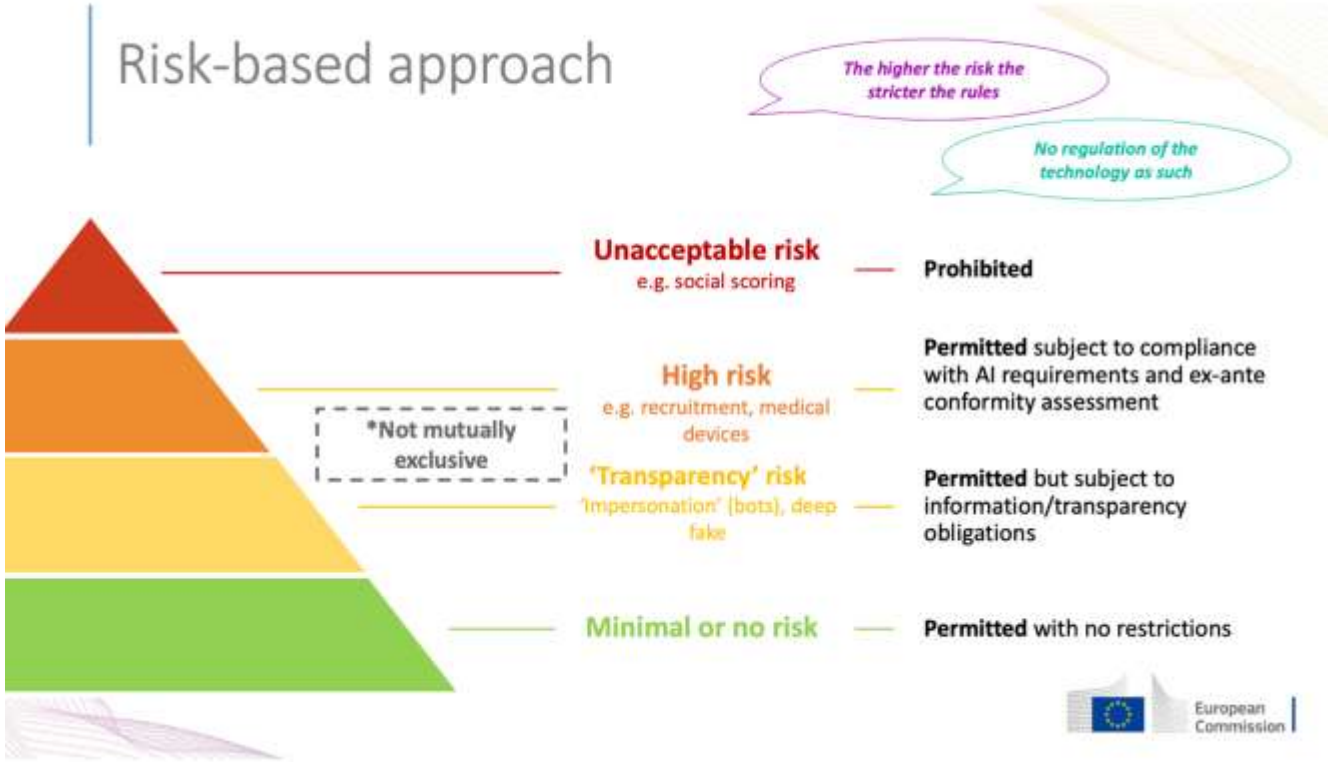


VI. GDPR & EU - AI Act

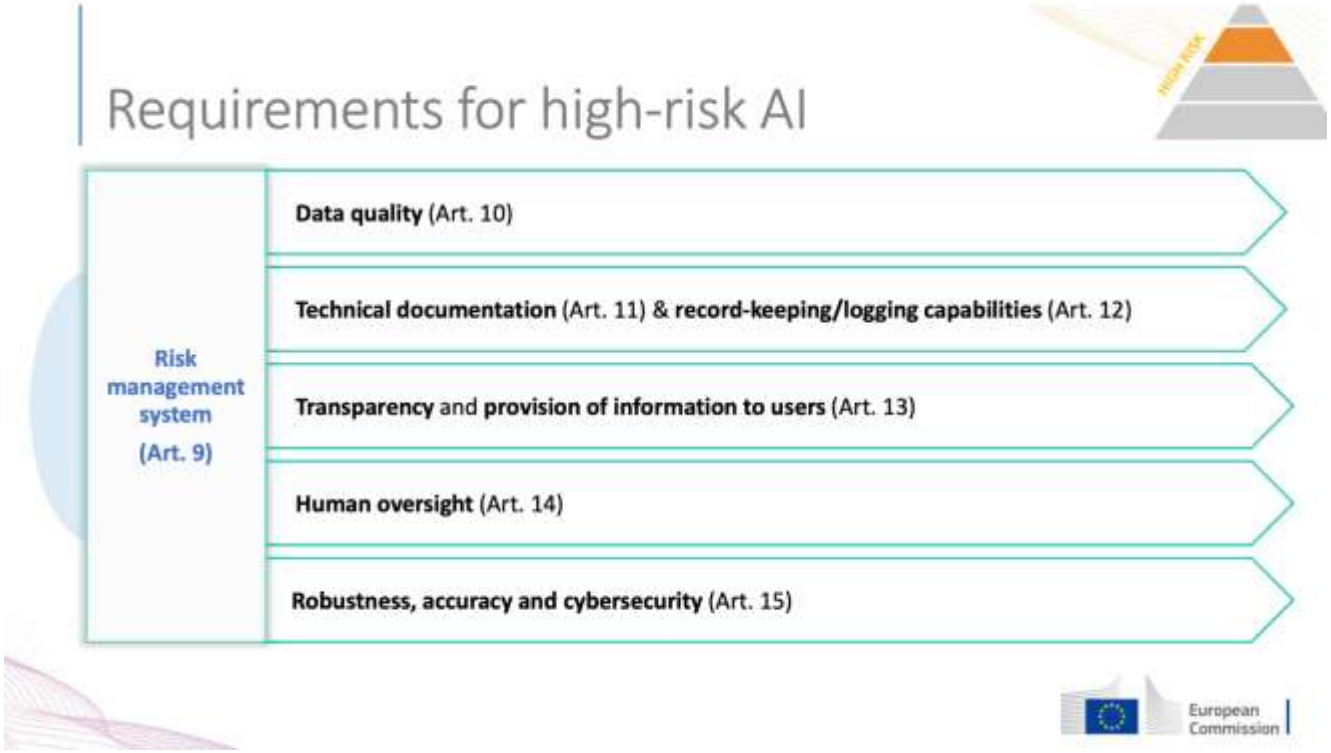
GDPR


Key Take-aways From the Seven Principles of GDPR
1
Be Transparent With Data
Implied consent is a big no-no under the GDPR.
2
Limit Data to What You Need
No snooping up data just because you can.
3
Limiting Kept Data
Do we need all this data? If the answer is no, delete it.
4
Data Must be Accurate
Make sure that data is accurate and up-to-date.
5
Limit Storage of Personal Data
Don't keep it longer than you need it.
6
Integrity and Confidentiality
Use encryption, 2FA, and tamper-evident logging.
7
Accountability
Keep a paper trail to demonstrate compliance.
= GDPR COMPLIANCE!

The AI Act - Risk level

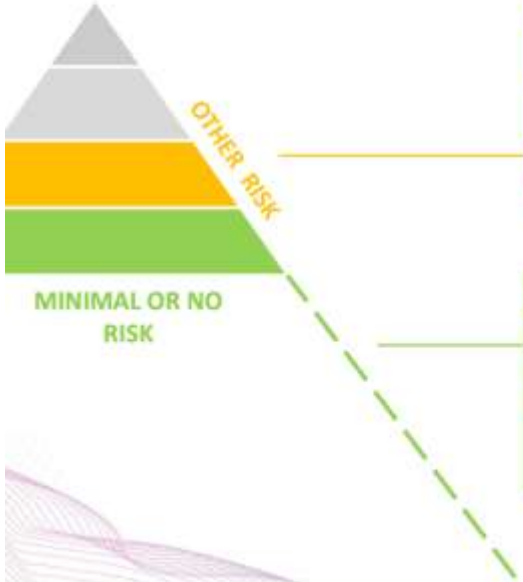


The AI Act - High Risk



The AI Act - Low Risk

Disclosure obligations for non high-risk



Transparency obligations for certain AI systems (Art. 52)

- **Notify humans** that they are **interacting with an AI system** unless this is evident
- **Notify humans** that they are **exposed to emotional recognition or biometric categorisation systems**
- Apply **label to deep fakes**

Possible voluntary codes of conduct (Art. 69)

- No mandatory obligations
- Commission and Board to encourage drawing up of codes of conduct (**voluntary application of requirements for high-risk AI systems or other requirements)**

European Commission

Extra Mile

Readings:

•Michael Haenlein, Andreas Kaplan. (2019). *A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence*. California Management Review - link

•Avery Artsman. (2016). *How Google's self-driving car project rose from a crazy idea to a top contender in the race toward a driverless future*. Business Insider - link

Movies:

https://en.wikipedia.org/wiki/The_Imitation_Game

https://en.wikipedia.org/wiki/AlphaGo_(film)

https://en.wikipedia.org/wiki/Coded_Bias

Thank you!