# Software Engineering For Data Science (SEDS)

**Class: 2nd Year 2nd Cycle**
**Branch: IASD**

**Dr. Belkacem KHALDI| ESI-SBA**

## Lecture 01:

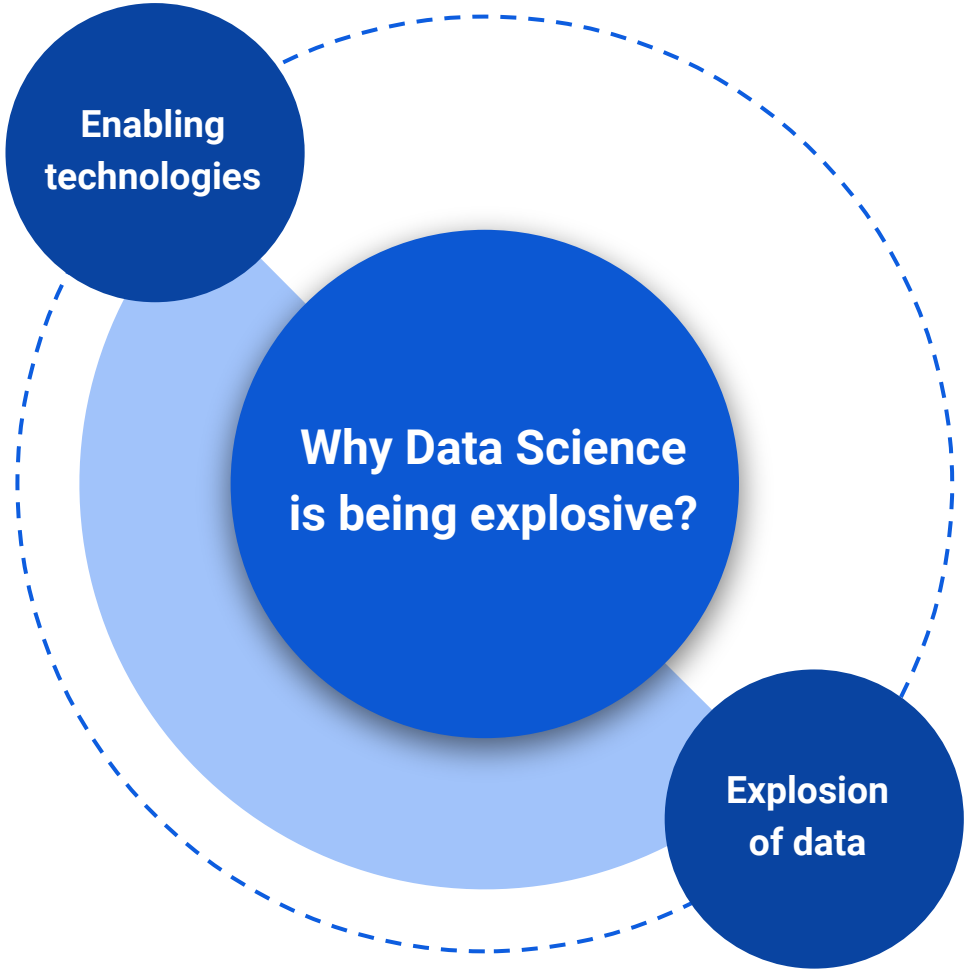
# Introduction To Data Science

# Introduction to Data Science

- Explosion of Data Science
- Brief History of Data Science
- Why Data Science?
- What is Data Science?
- Data Science Landscape
- Data Science Project Methodologies
- Specializations in and around data science
- Top Data Science Tools

# Explosion of Data Science

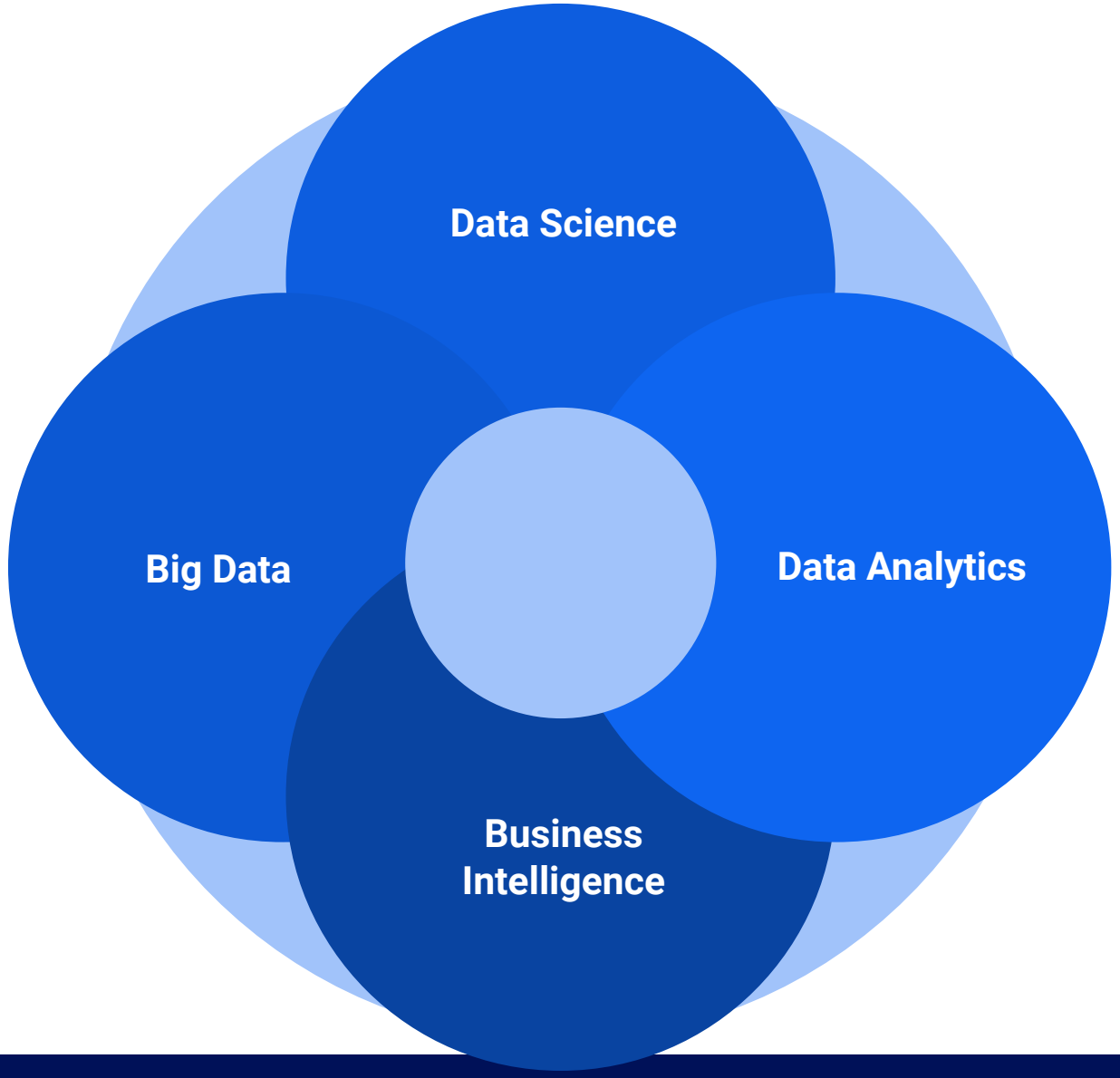
- Storage capacity
- Computing hardware;
- Algorithms;
- 350 years of statistics;
- 100 years of numerical analysis.



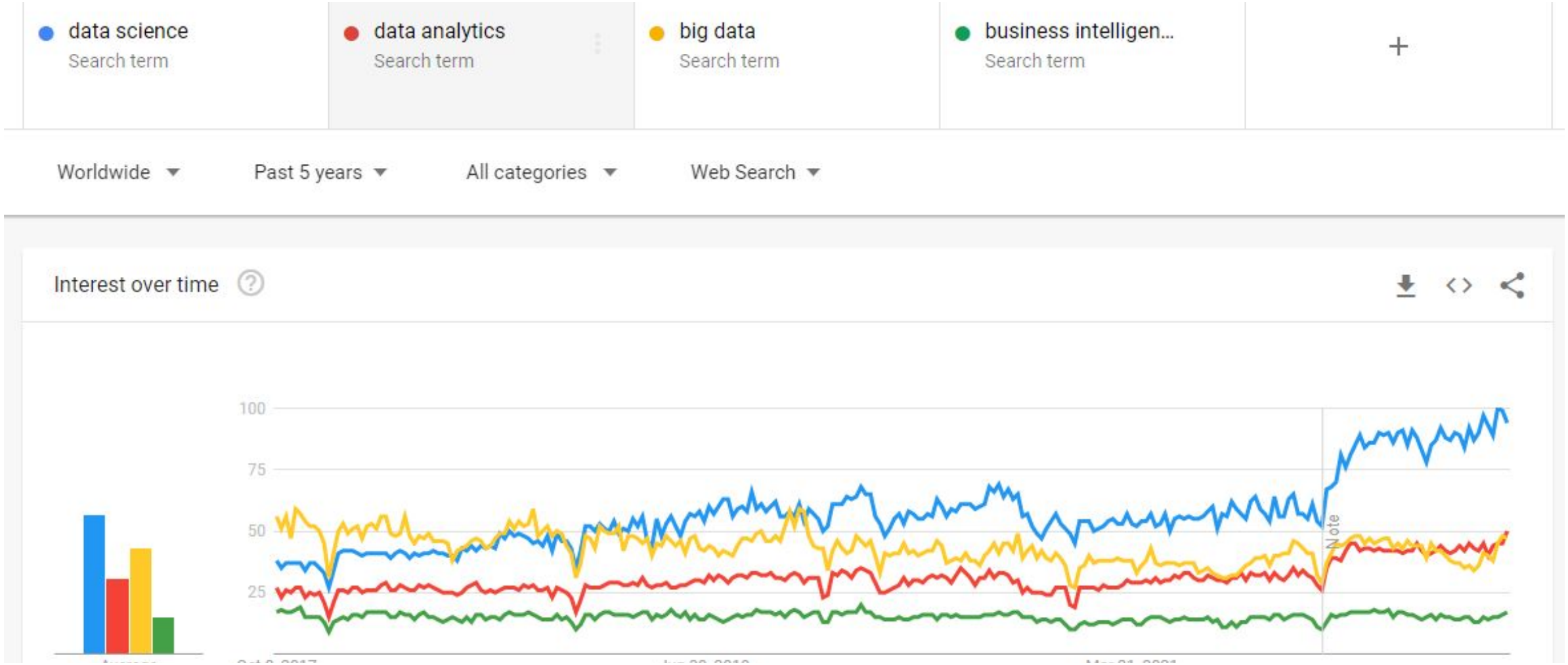

- Sensor technology (e.g. weather, smart vehicules, smart homes, …);
- Purchase histories (customer loyalty programs, fraud detection, …);
- Data from Smartphones (≈8 billion) generating GPS traces, trillions of photos each year, ..
- DNA sequencing . . .

# Explosion of Data Science: Buzz Words

**Buzz Words due to constant evolution of data science**

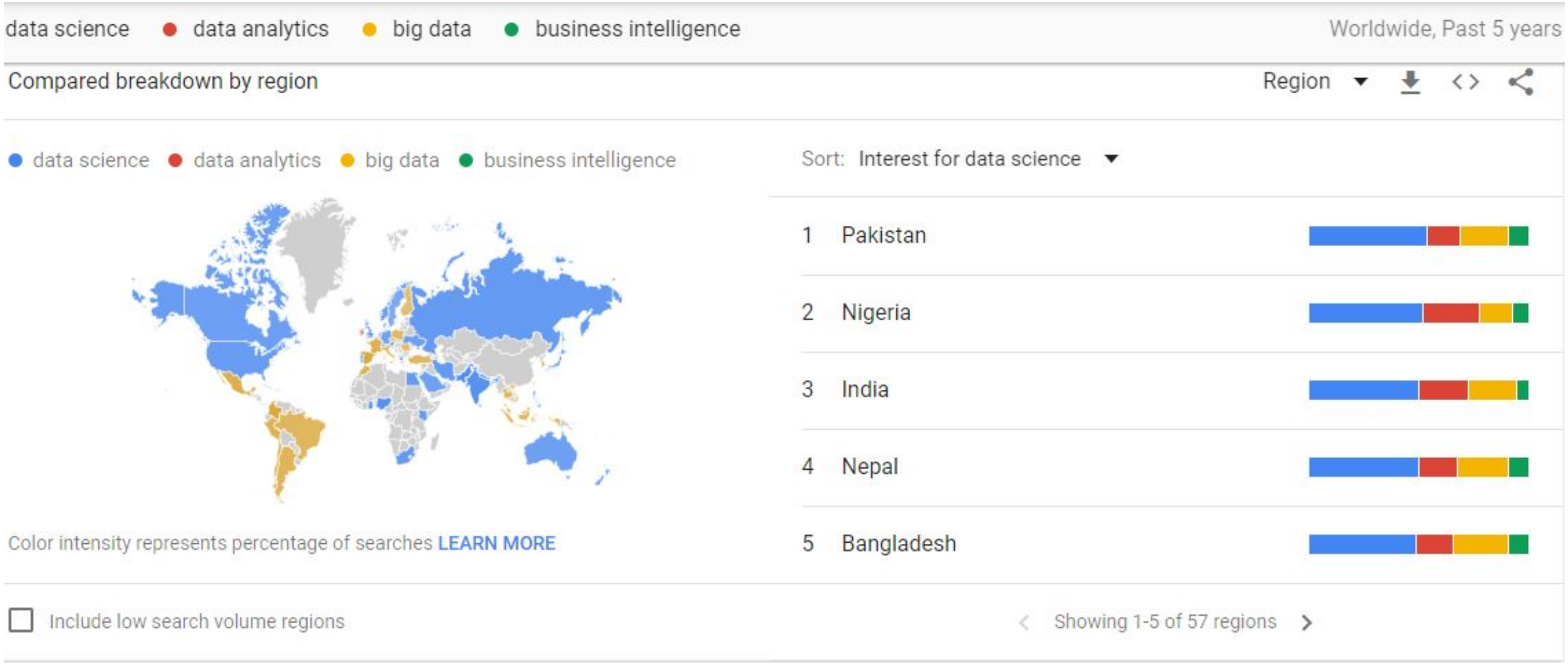


# Explosion of Data Science: Buzz Words




https://trends.google.com/

# Explosion of Data Science: Buzz Words




https://trends.google.com/

# Explosion of Data Science: Buzz Words

## Big Data

A term that describes large, hard-to-manage volumes of data – both structured and unstructured – that inundate businesses on a day-to-day basis.

- Structured or unstructured.
- Stored in Databases, Cloud, or Warehouses systems.

## Business Intelligence (BI)

A technology-driven process for analyzing data and delivering actionable information that helps make informed business decisions.

- Data Analysis
- Visual Dashboard
- Reporting with Visuals

## Data Analytics

The use of tools and processes to combine and examine datasets to identify patterns and develop actionable insights.

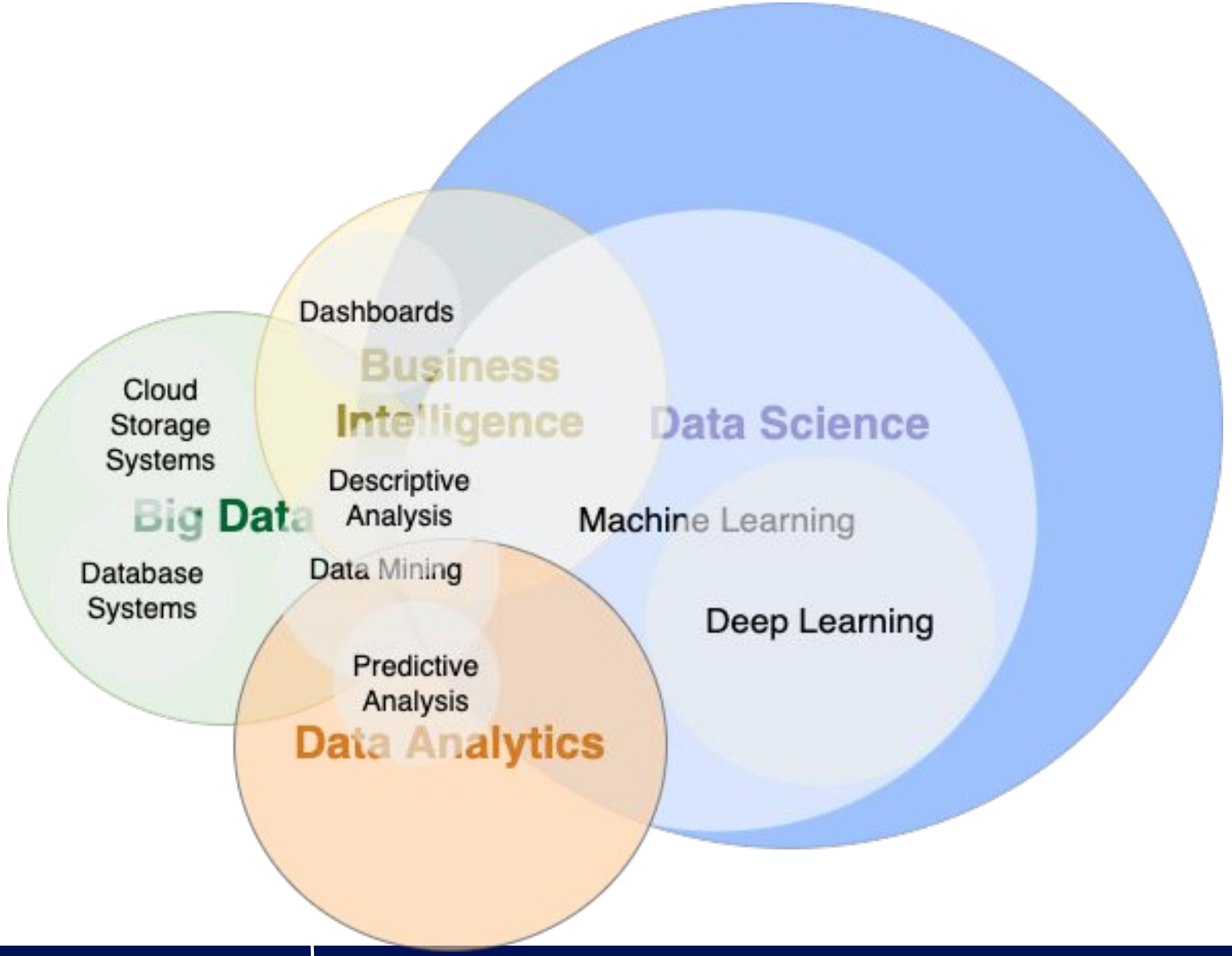
- Data Analysis
- Classical Statistical Models Creation.
- Results Communication.

## Data Science

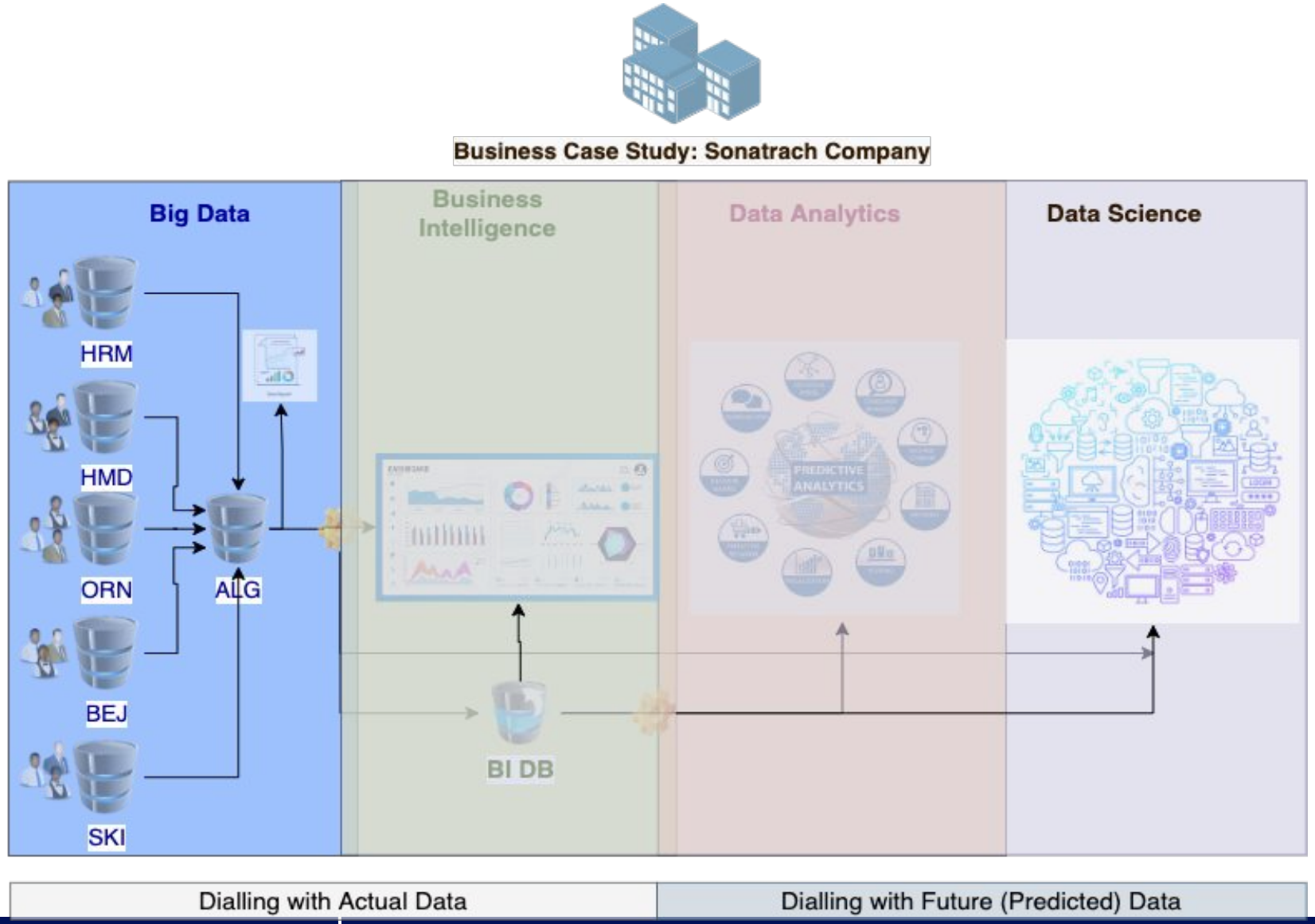
The field of applying advanced analytics techniques to extract valuable information from data for business decision-making, strategic planning and other uses.

- Data Analysis
- ML and DL Models Creation..
- Models Deployment
- Results Communication

# Data Science: Buzz Words



# Data Science: Buzz Words



# Brief History of Data Science (1\3)

**John Tukey**

*The Future of Data Analysis*, a published paper where he envisions a new field for learning insights from data.

**1962**

**1977**

**John Tukey**

*Exploratory Data Analysis*, Book Publishing, which is a key part of data science today.

**1991**

**Guido Van Rossum**

Publishing **Python** programming language online for the first time, which becomes later the 1st most-used data science language.

**1993**

**Ross Ihaka and Robert Gentleman**

The **R** programming language is publicly released, which goes on to become the 2nd most-used data science general-purpose language.

**1996**

**IFCS-96 Conference, Kobe, Japan**

The 5th conference on "Data Science, Classification and Related Methods" – possibly the first time "**data science**" was used to refer to something similar to modern data science.

**IFCS:** International Federation of Classification Societies.

# Brief History of Data Science (2\3)

**Jeff Wu**

Proposition to rename “**statistics**” to "**data science**" in an inauguration lecture at the University of Michigan.

**Jeff Hammerbacher and DJ Patil**

use the term "**data scientist**" in **job** postings after trying to come up with a good job title for their work.


**Data Science Masters Degrees**

Universities begin offering **masters degrees** in **data science**; **data science job** postings explode to new heights; big breakthroughs are made in **deep learning** with releasing new packages and publications.

**1997**

**2001**

**2008**

**2010**

**2010s**

**William Cleveland**

Publishing of an interesting paper describing a new field, "**data science**," which expands on data analysis.

**Kaggle.com**

The launch of **kaggle.com** as an online **data science community** and **data science competition** website.

kaggle.com

# Brief History of Data Science (3\3)


**Harvard Business Review (HBR)**

Publishing the well-known article entitled "*Data Scientist: The Sexiest Job of the 21st Century*", which adds fuel to the data science fire.

**TensorFlow and PyTorch**

The release of **TensorFlow** (2015) and **PyTorch** (2016) as deep learning and machine learning libraries by **Google** and **Facebook**.

**SageMaker Studio**

The release of **SageMaker studio** by **Amazon**, an entirely complet cloud tool for building, training, deploying, and analysing machine learning models.

**2012** | **2015** | **2015-2016** | **2018** | **2020**

**DJ Patil**

The **1st chief data scientist** of the **US Office of Science and Technology Policy**. His mission was to "*Unleashing the Power of Data to Serve the American People*".

**AutoML**

The release of **cloud AutoML** by **Google**, democratizing a new automatic technique for machine learning and data science.

Google's AutoML

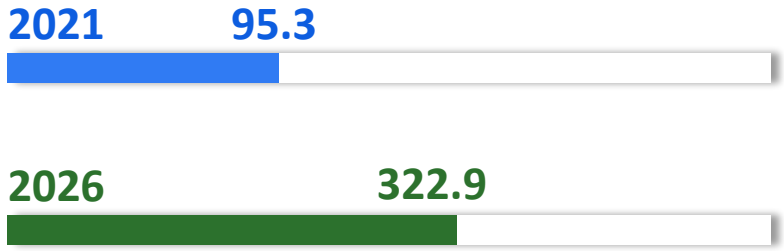
# Brief History of Data Science: Summary

- **Data science** was around for several decades before it became wildly popular.
- **Data science** is actually being used productively thanks to the amount of digital data availability and accessibility.
- **Python** and **R**, the two most extensively used programming languages in **data science**, existed for **15 years** before the topic of **data science** became serious.
- **TensorFlow** from Google and **Pytorch** from Facebook are being extensively enhanced since their first releases as they are actually the most-used python libraries in **Data Science**.
- The rise of **Data Science Competitions** for the 1st time in **2010** by **Kaggle.com**
- In the late **2010s** and early **2020s**, some aspects of **data science** started to become automated (**AutoML**).

# Why Data Science? Markets & Statistics

- According to the **2021 Data Science Platform Market Report by marketsandmarkets.com**:
  - The global **Data Science Platform Market** size was valued **$95.3 billions** in **2021**,
  - The market for data science platform is estimated to reach **$322.9 billions** in **2026**,
  - At a Compound Annual Growth Rate (**CAGR**) of **27.7%** for the period **2021-2026**.

**The global Data Science Platform Market ($ billion)**

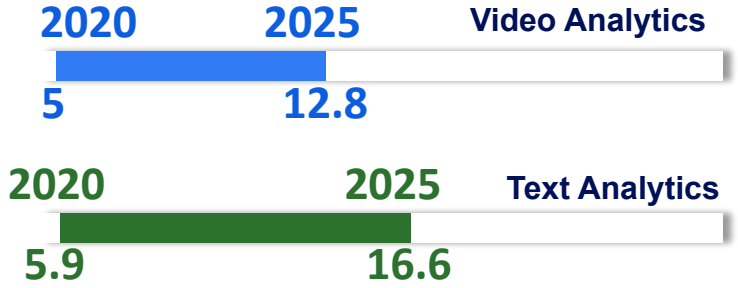



**Compound Annual Growth Rate**


https://www.marketsandmarkets.com/

- According to the **Global Analytics Industry Report 2022** by **researchandmarkets.com**:
  - The **global market** for data science **video analytics** should grow from **$5.0 billions** in **2020** to **$12.8 billions** by **2025**, at compound annual growth rate (CAGR) of **20.6%** for the period of **2020-2025**.
  - The **global market** for data science **text analytics** should grow from **$5.9 billions** in **2020** to **$16.6 billions** by **2025**, at compound annual growth rate (**CAGR**) of **23.0%** for the period of **2020-2025**.

**The global Analytics Industry Report 2022 ($ billion)**




**Compound Annual Growth Rate**

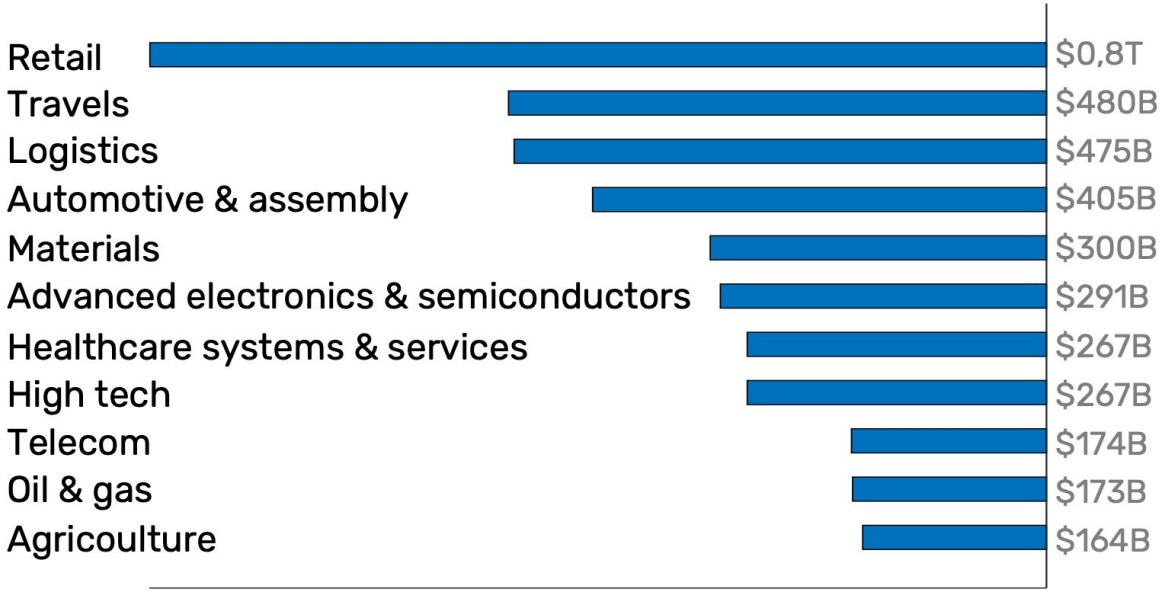


https://www.researchandmarkets.com/

**CAGR:** the mean annual growth rate of an investment over a specified period of time longer than one year

# Why Data Science? Markets & Statistics

- **Data science** applications offer innumerable business benefits:
  - According to the **McKinsey Global Institute** Study, "**Notes from the AI frontier: Modeling the impact of AI on the world economy",** reported on **Sep. 04, 2018**:
    - The Business value that will be created by the AI and Data Science up to 2030 is worth **$13 Trillions**.
    - It is difficult to find an industrial sector that will not benefit from **Data Science** in the near future.


https://www.mckinsey.com/

# Why Data Science? Markets & Statistics

- According to the **Simplilearn** Study , "***Introduction to Data Science: A Beginner's Guide***", reported on **Mar 19, 2022**:
  - Companies that are implementing the ground-breaking technology of Data Science are already taking advantage of it.
    - **Southwest Airlines Co.** saved **One hundred million dollars** by minimizing the idle hours of its planes.
    - **United Parcel Service, Inc.** saved **36 million dollars** by optimizing its fleet,
    - **The U.S., Internal Revenue Service**, saved **2 billion dollars** by enhancing its ability to identify improper payments (fraud payments).


https://www.simplilearn.com/

# Why Data Science? Motivations

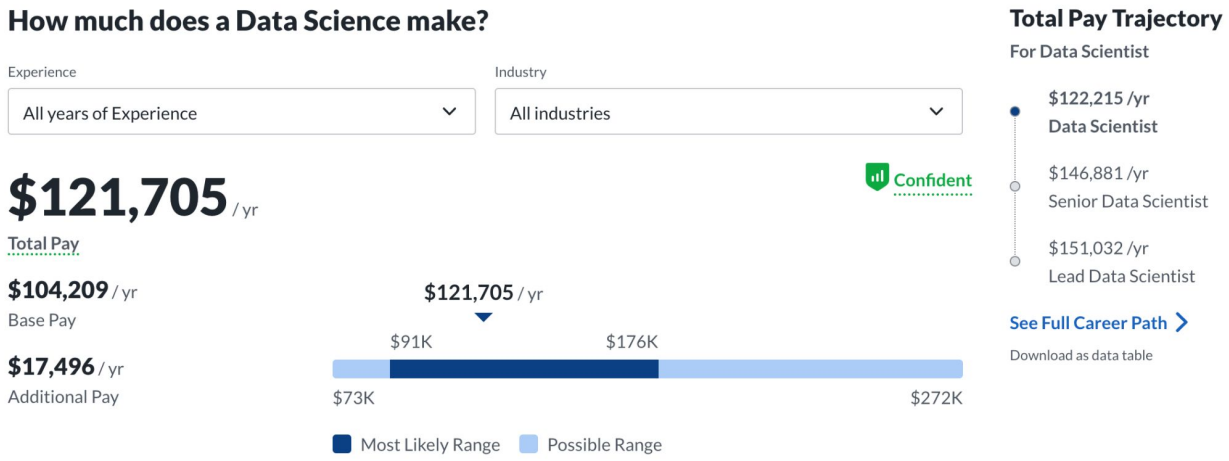
**Data science** has been deemed as the sexiest job of the 21st century.

- **Collected information** need to be **analyzed** properly in order to get **actionable results**.
- A huge amount of **data** requires s**pecific infrastructures** to be handled.
- A huge amount of **data** requires **computational power** to be analyzed.
- Rising of specific **job titles**: Data Scientist, Senior Data Scientist, Lead Data Scientist, Full Stack Data Scientist, Data Analyst, ….

# Why Data Science? Motivations

**Data science** has been deemed as the sexiest job of the 21st century.

- The estimated total pay for a **Data Scientist** is **$121,705** per year in the **United States** area.
- The rise of **data science competitions** with cash prizes: Kaggle, Analytics Vidhya, HackerRank, DrivenData, AIcrowd, CodaLab, Topcoder, Zindi, Tianchi, ….
- Virtually every aspect of business is now open to **data collection** (operations, manufacturing, supply-chain management, customer behaviour, marketing campaigns, Oil & Gas, Agriculture, …).




https://www.glassdoor.com/

# What is Data Science?

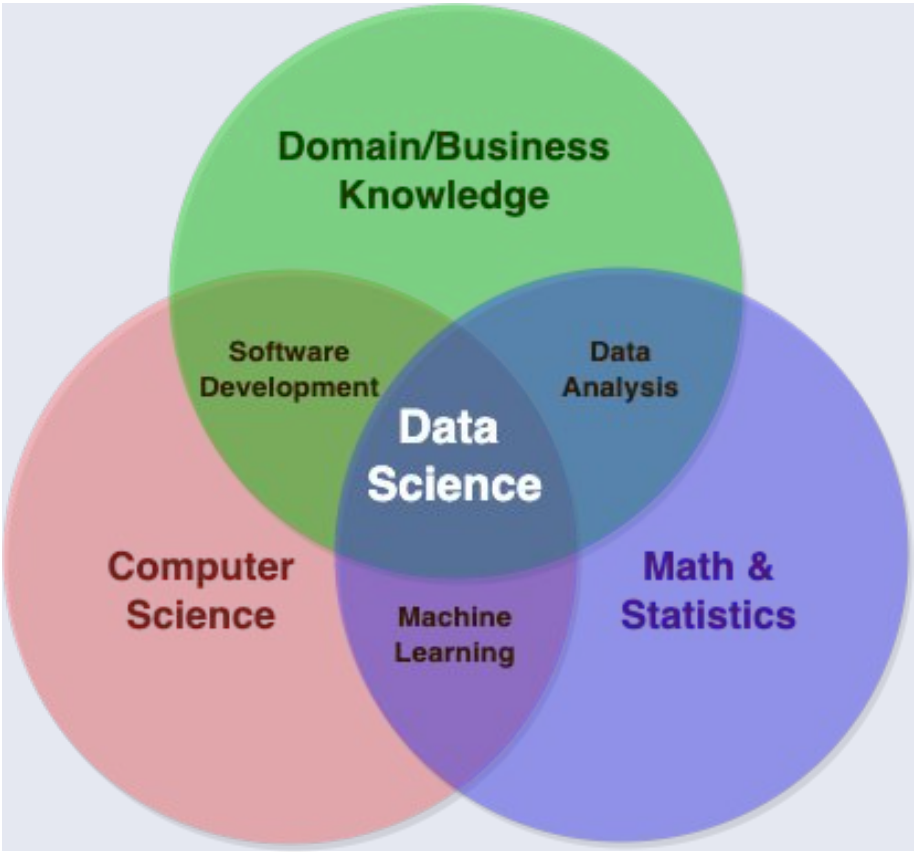
- **A funny definition:**

"*A data scientist is better than any computer scientist at statistics, and better than any statistician at computer programming.*"

This encapsulates the general skills of most data scientists, as well as the history of the field

- **Several Definition:**
  - **Data science** ⇒ the study of data and it is used to develop methods to store, record, and analyse data to effectively get useful information.
  - **Data science** ⇒ a combination of different techniques and theories taken from many fields such as **Math and statistics**, **Computer Science** and **Domains/Business Knowledge**.

- **Data Science** relies on ***software development*** processes + ***data analysis*** techniques ⇒ to build accurate ***artificial intelligence*** and ***machine learning*** models capable of extracting useful data and **predicting the future patterns** and **behaviours**.



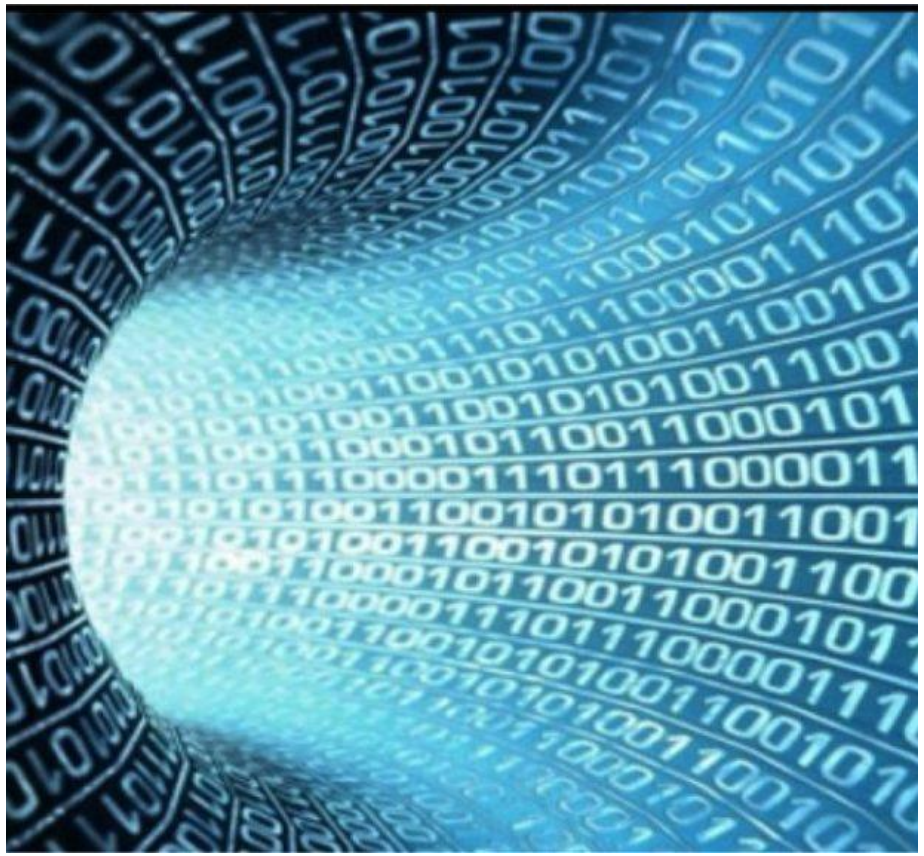
# What is Data Science? All About Data

**Classical Statistical Analytics Methods** are meaningful to be applied if:

1. A **pattern** exists.
   - If a **pattern** does not exist, I do not **learn** anything.
2. We cannot pin it down mathematically ⇒ Absence of **analytical solutions**).
   - If I can describe the **pattern** mathematically, I will not presumably **learn** the best **relation**.

So the Solution is to apply advanced **Data Science Techniques**.

- We have **data** so use it for solving **business** problems.



# What is Data Science? All About Data

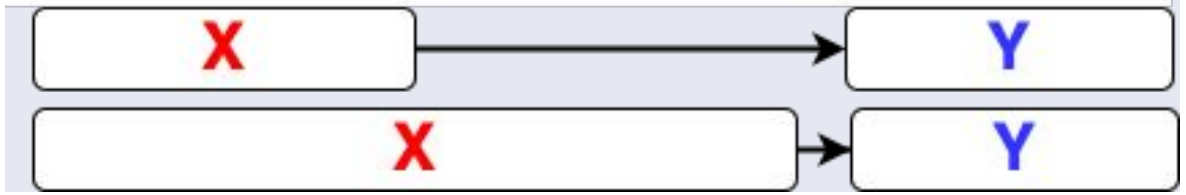
## Data Types

- The data can have different **formats**. The most typical is that of a **table.**

- The **data** can come from a database or from .csv, Excel files…

- **Goal: predict house prices**
  - Learn the relation from **House area** to **Price**.
  - Learn the relation from **House area** AND **#bedrooms** to **Price**

| House area(feet$^2$) | # bedrooms | Price (1000$) |
|---|---|---|
| 523 | 1 | 115 |
| 645 | 1 | 150 |
| 708 | 2 | 210 |
| 1034 | 3 | 280 |
| 2290 | 4 | 355 |
| 2545 | 4 | 440 |

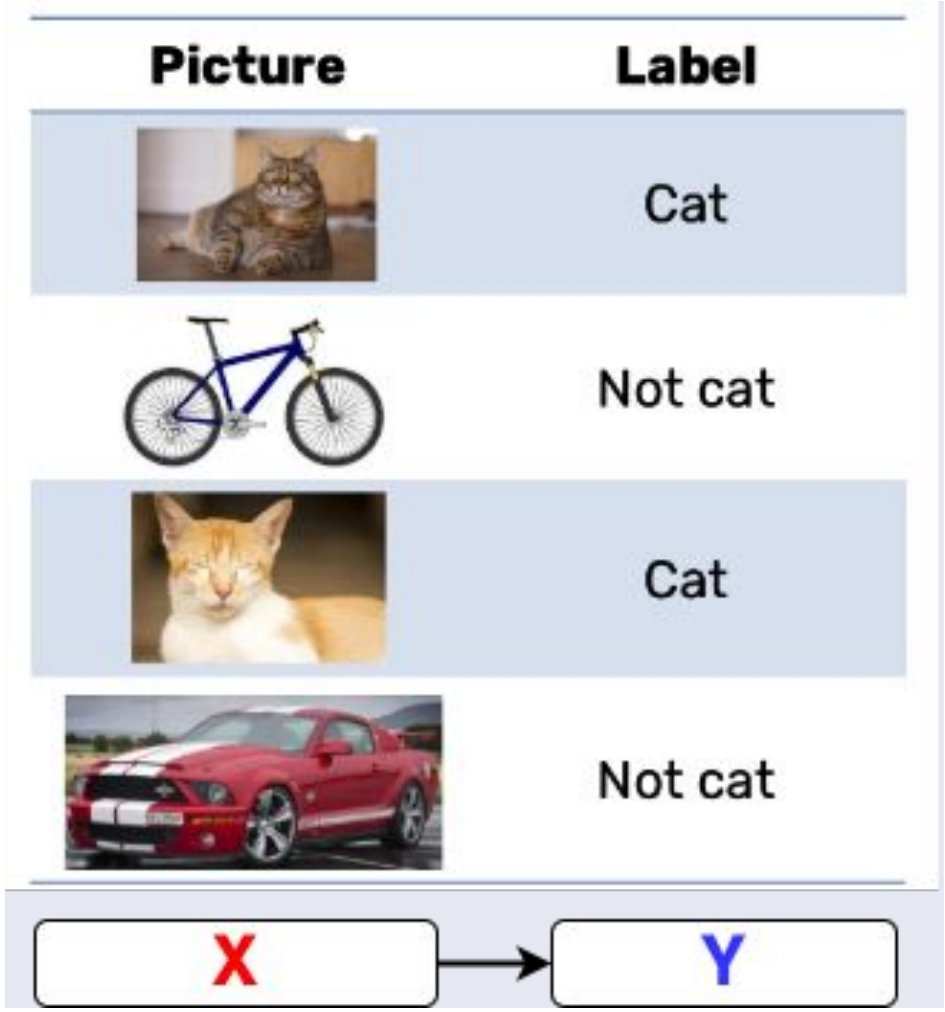


# What is Data Science? All About Data

## Data Types

- Another type of data can be an **image.**

- **Goal: recognize if there is a cat in the image**
  - Learn the relation from a **Picture** to a **Label** (**«class of belonging» (cat vs. not cat))**

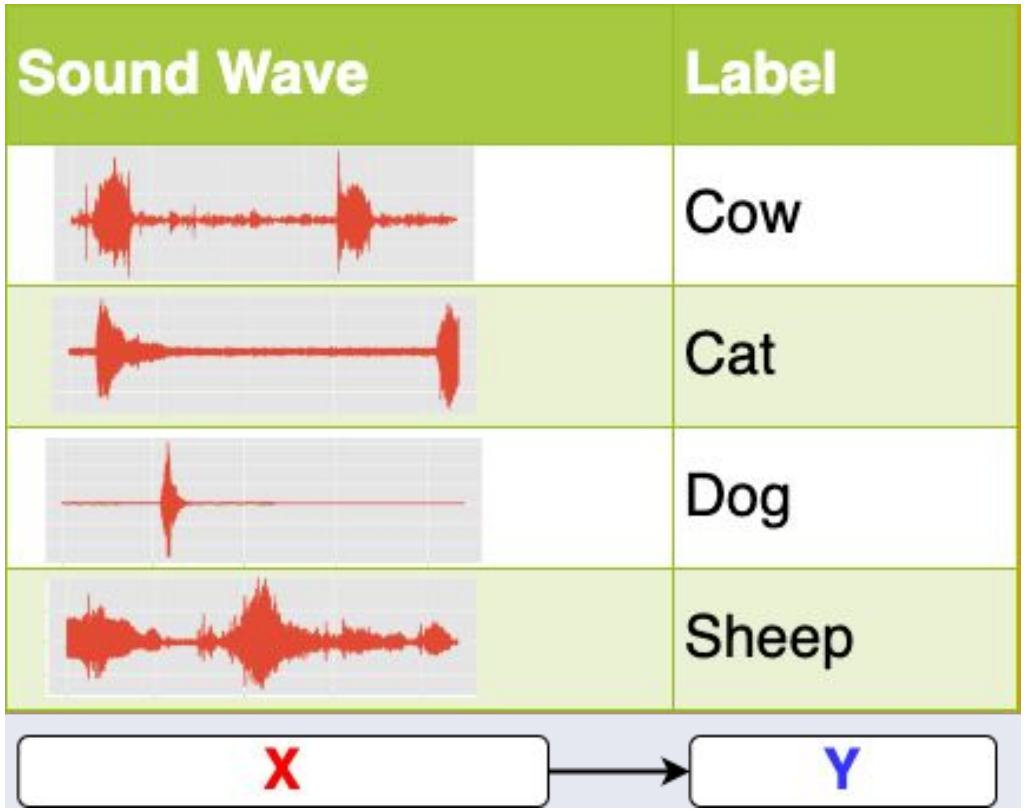


# What is Data Science? All About Data

## Data Types

- Another type of data can be an **Audio.**

- **Goal: recognize animal sound**
  - Learn the relation from a **Sound Wave** to a **Label** (**«class of belonging» (animal))**



# What is Data Science? All About Data

**Data are dirty:**

- Garbage IN, garbage OUT

- **Data problems:**
  - **Missing values**
  - **Not correct values**
  - ....

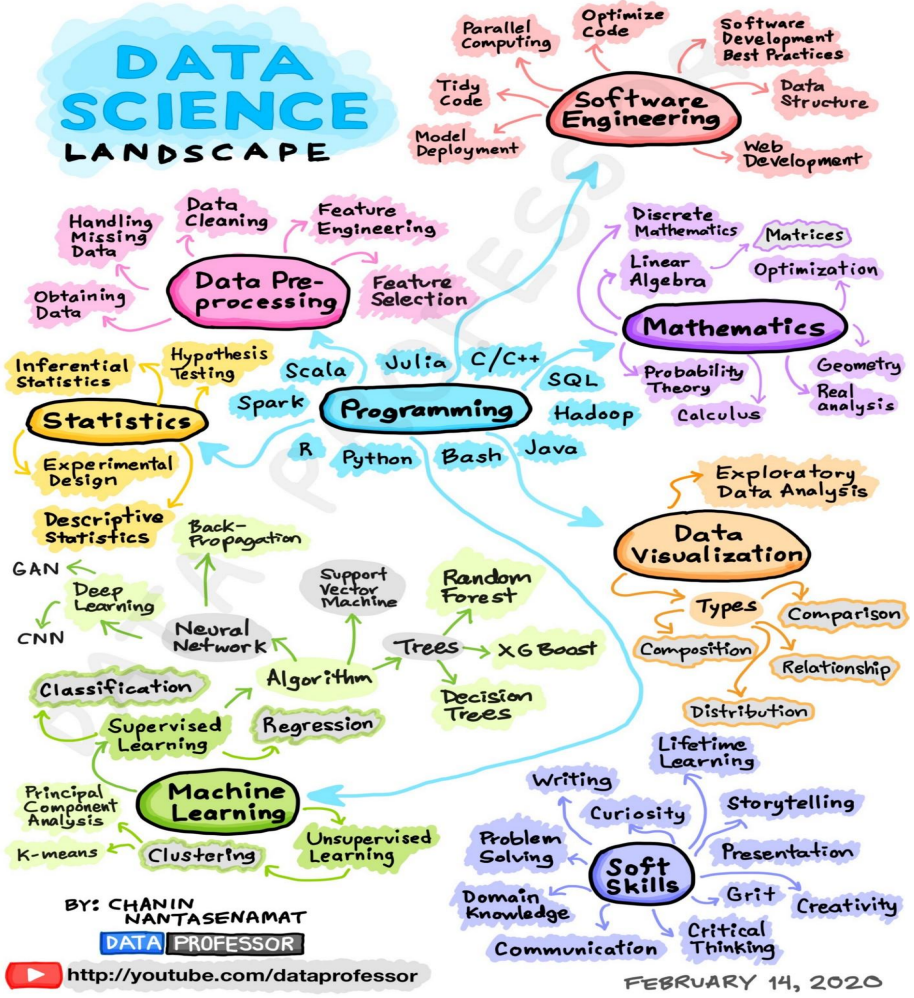
- **Different data types (Structured, Unstructured)**
  - Images, audio, text

| **House area**(feet$^2$) | **# bedrooms** | **Price** (1000$) |
|---|---|---|
| 523 | 1 | 115 |
| 645 | 1 | 0,001 |
| 708 | unknown | 210 |
| 1034 | 3 | unknown |
| unknown | 4 | 355 |
| 2545 | unknown | 440 |

# Data Science Landscape

## A Lot of Skills and Techs to Acquire and Develop

1. **Programming**
   - Python: The most dominant Prog. in Data Science.
2. **Data Pre-processing**
   - Collecting, organizing, and preparing data (25-75% time consuming).
3. **ML and DL algorithms**
   - ML (Clustering, Classification, Regression)
   - DL (NNs, LSTMs, CNNs, GANs, NLP....)
4. **Software Engineering**
   - Code versioning, Reproducible and scalable software Modules, and advanced programming techniques.
5. **Data Visualisation**
   - Visual Data Exploration Analysis: maps, charts, and graphs.
6. **Statistics and Maths.**
   - Basis of data science: Probability theory, Linear Algebra, Statistics Inference, Descriptive Statistics, Calculus, ....
7. **Soft Skills**
   - Other skills such as communication, writing, Problem Solving, ...

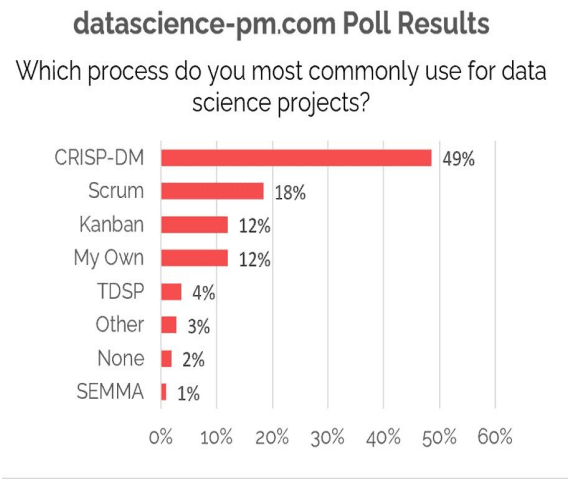


# Data Science Project Methodologies

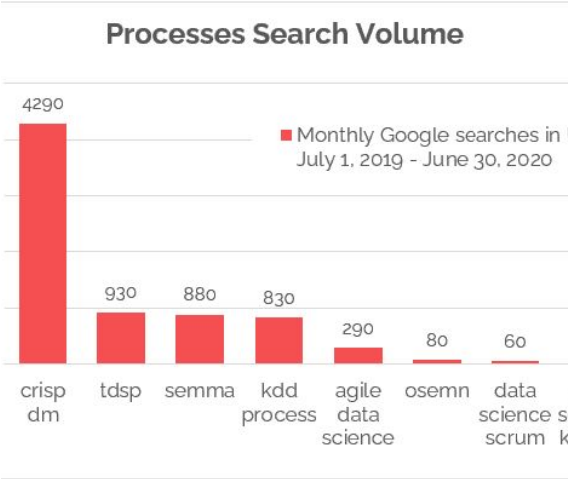
- Working on a large **data science projects** ⇒ Organizing the project into a **process of steps**.

- This especially helps when working as a **team** ⇒ Each team (Group) member(s) focus on a specific Process Step.

- There are a number of **science project management strategies**, the most well-applied one:
  - **CRISP-DM**: **CR**oss-**I**ndustry **S**tandard **P**rocess for **D**ata **M**ining.




**2020 Poll Results**



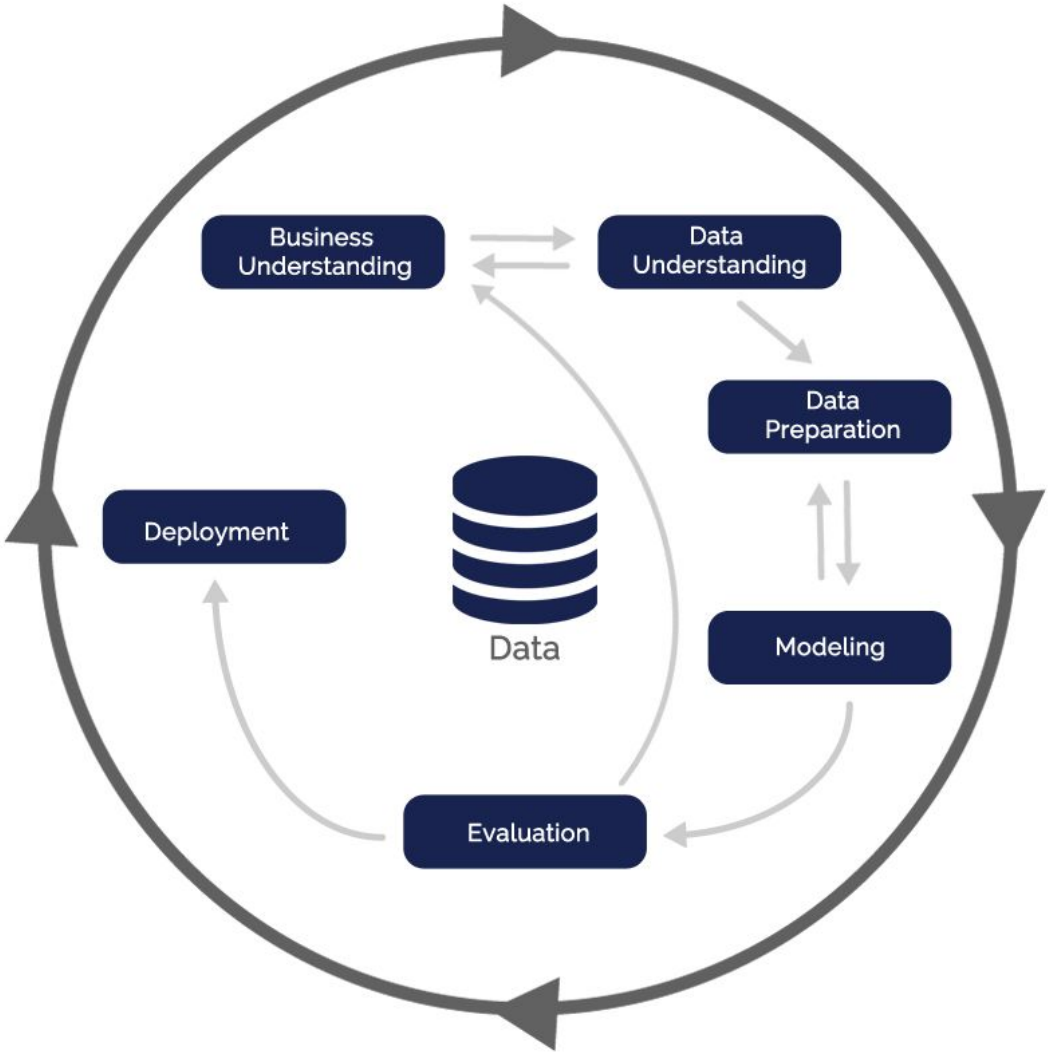

**Google Searches**

https://www.datascience-pm.com/

# Data Science Project Methodologies

## CRISP-DM:

- **CRISP-DM** has been around since the late **1990s**.
- It was created before **data science** existed as its own field. However, it's still widely used for **data science projects management**.
- It's a **six-step process**:
  - Business Understanding,
  - Data Understanding,
  - Data Preparation,
  - Modeling,
  - Evaluation, and
  - Deployment.




**CRISP-DM Process Diagram**

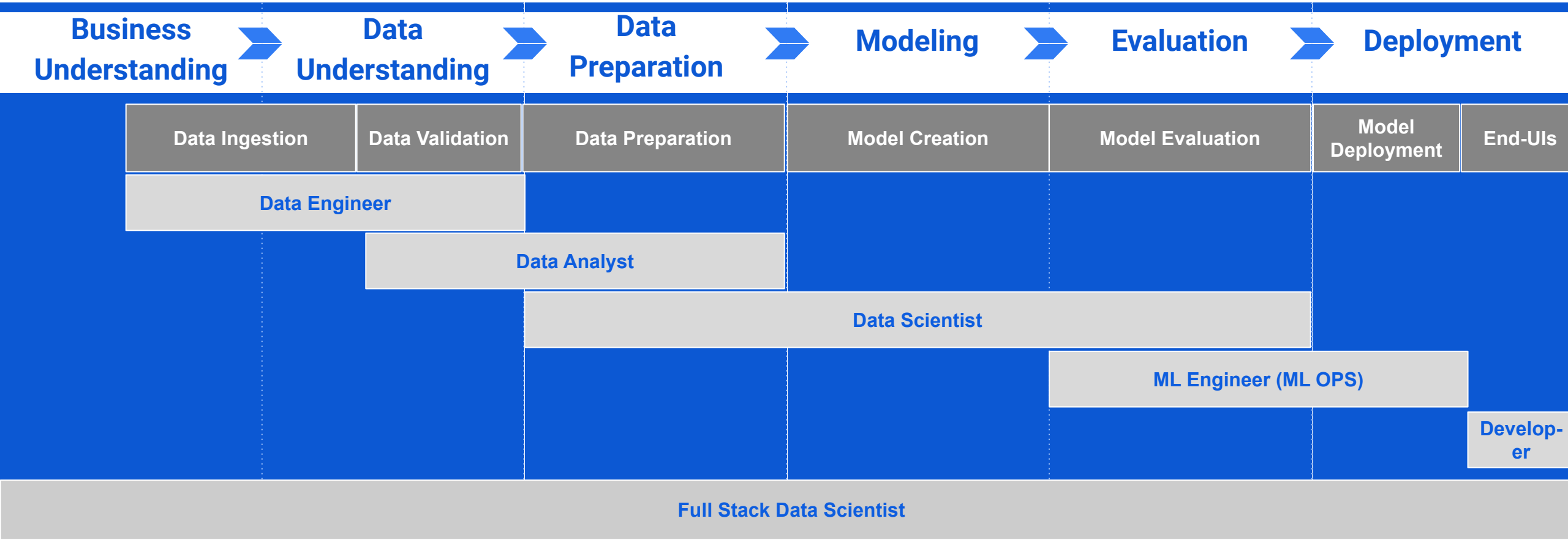
# Data Science Project Methodologies

## CRISP-DM:

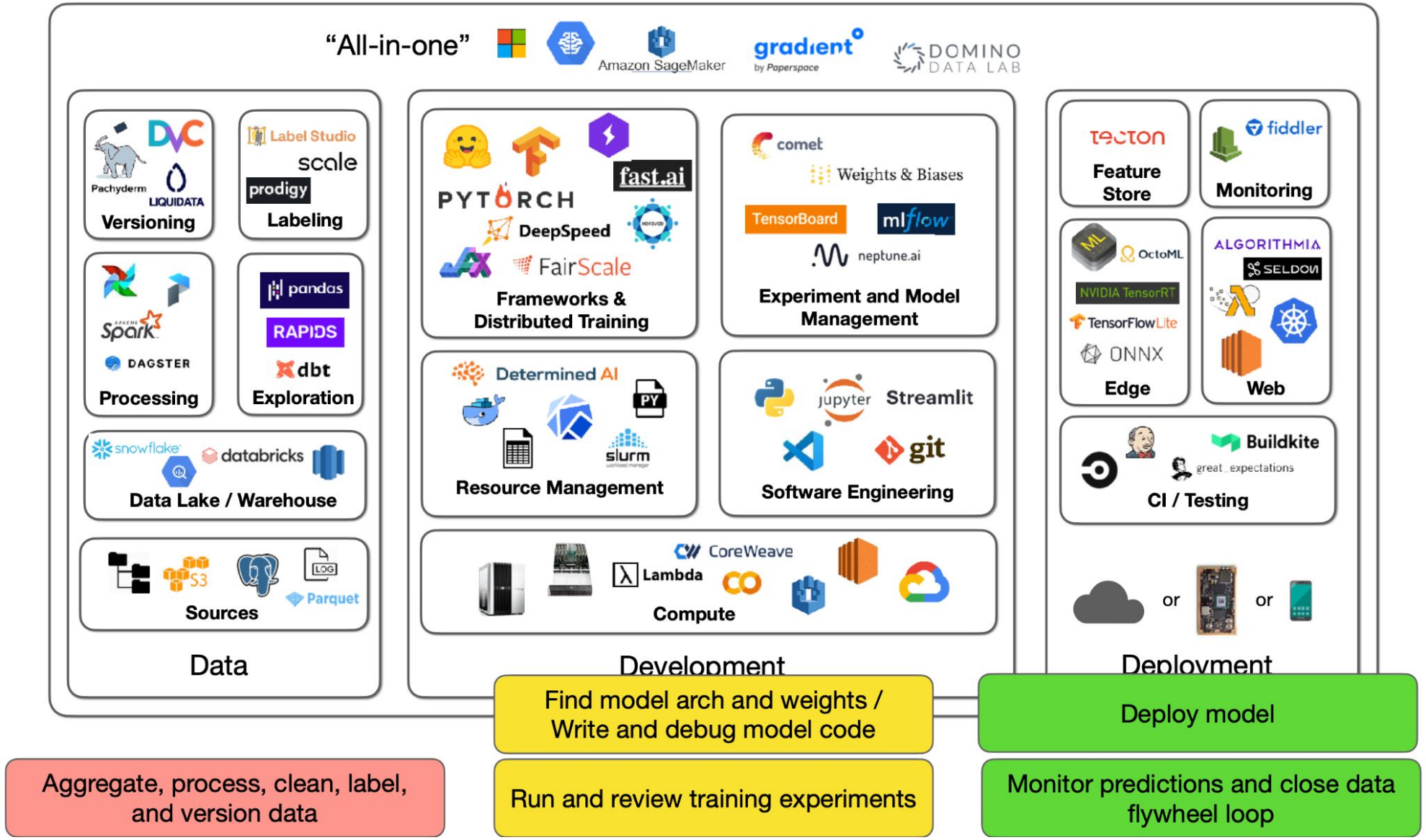
| Business Understanding | Data Understanding | Data Preparation | Modeling | Evaluation | Deployment |
|---|---|---|---|---|---|
| 1. Determine Business Objectives<br>2. Determine Analytics Outcomes<br>3. Access Situation<br>4. Produce Project Plan | 1. Data Ingestion<br>2. Collect Initial Data Set<br>3. Describe Data<br>4. Explore Data<br>5. Data Validation | 1. Select Data<br>2. Clean Data<br>3. Construct Data<br>4. Integrate Data<br>5. Format Data<br>6. Add Labels if Needed | 1. Select Modeling Techniques<br>2. Build Selected Models<br>3. Assess Models<br>4. Select the best Model | 1. Apply Test Set<br>2. Interpret Results<br>3. Review Process<br>4. Final Model Training and Optimization | 1. Analyse Customer ‘ s Environment<br>2. Develop End-UIs<br>3. Model Deployment<br>4. Model Monit. & Maint.<br>5. Produce Final Report<br>6. Review Project |

# Specializations in and around data science

**Several jobs and functions out there are being in and around data science:**



# Top Data Science Tools



# Thanks for your attention


Dr. Belkacem KHALDI
e-mail: b.khaldi@esi-sba.dz