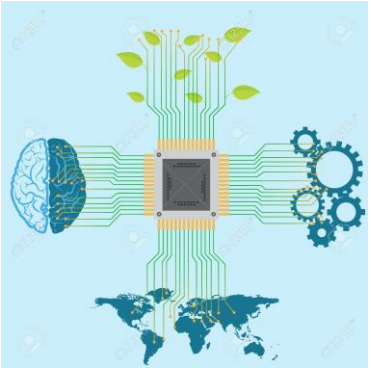
Introduction to Data Science

Data science: Lots of hype recently!

Interest over time. Web Search for « **Data science** » in the **past 10 years**.

Google Trends

Why the hype arround Data Science?



Data proliferation

- Unprecedented Data Generation Rates
- Need for Advanced Analytical Tools

Technological Advancements

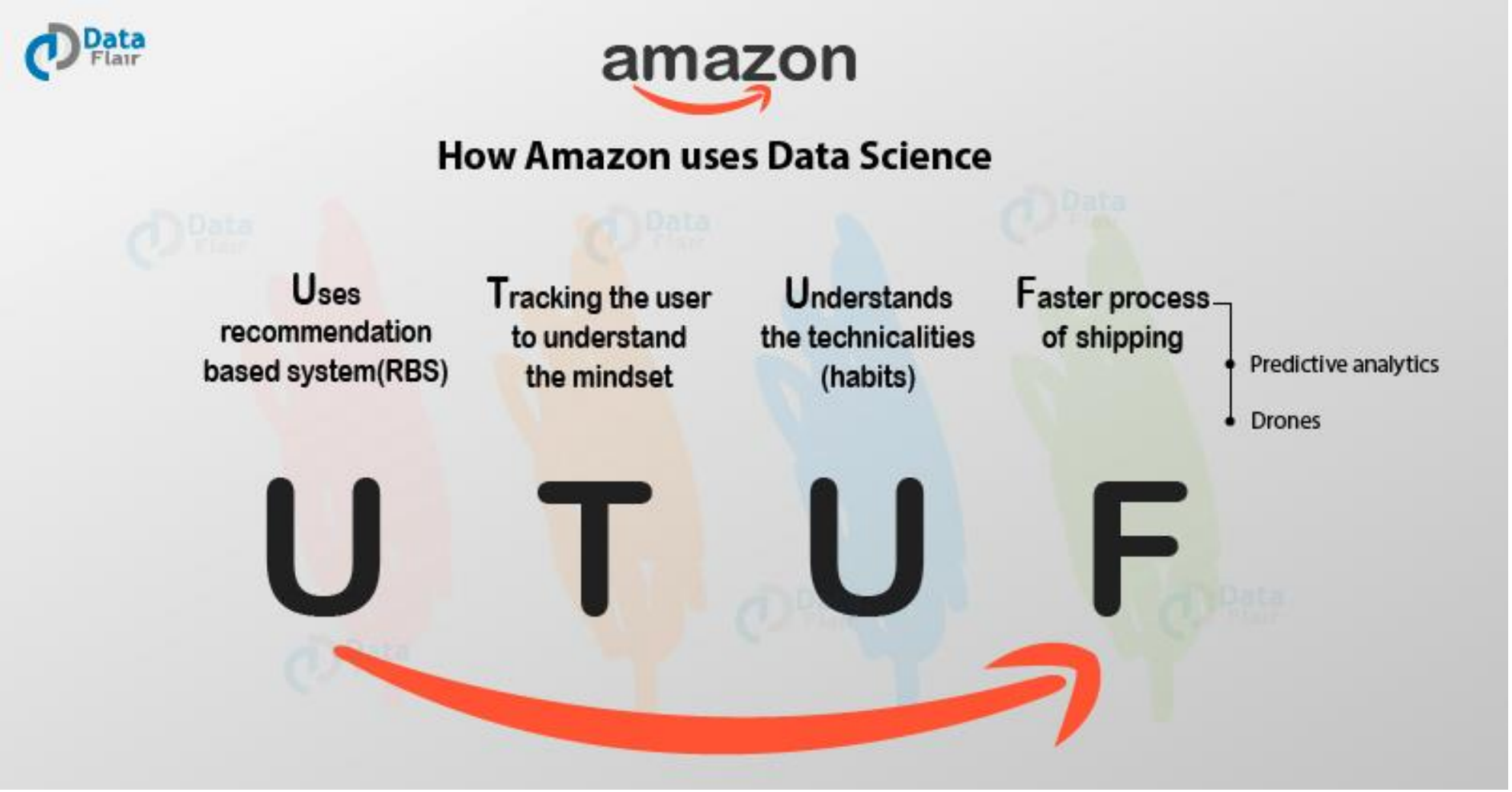
- High-Performance Computing
- Advanced Machine Learning Algorithms

Real-World Impact

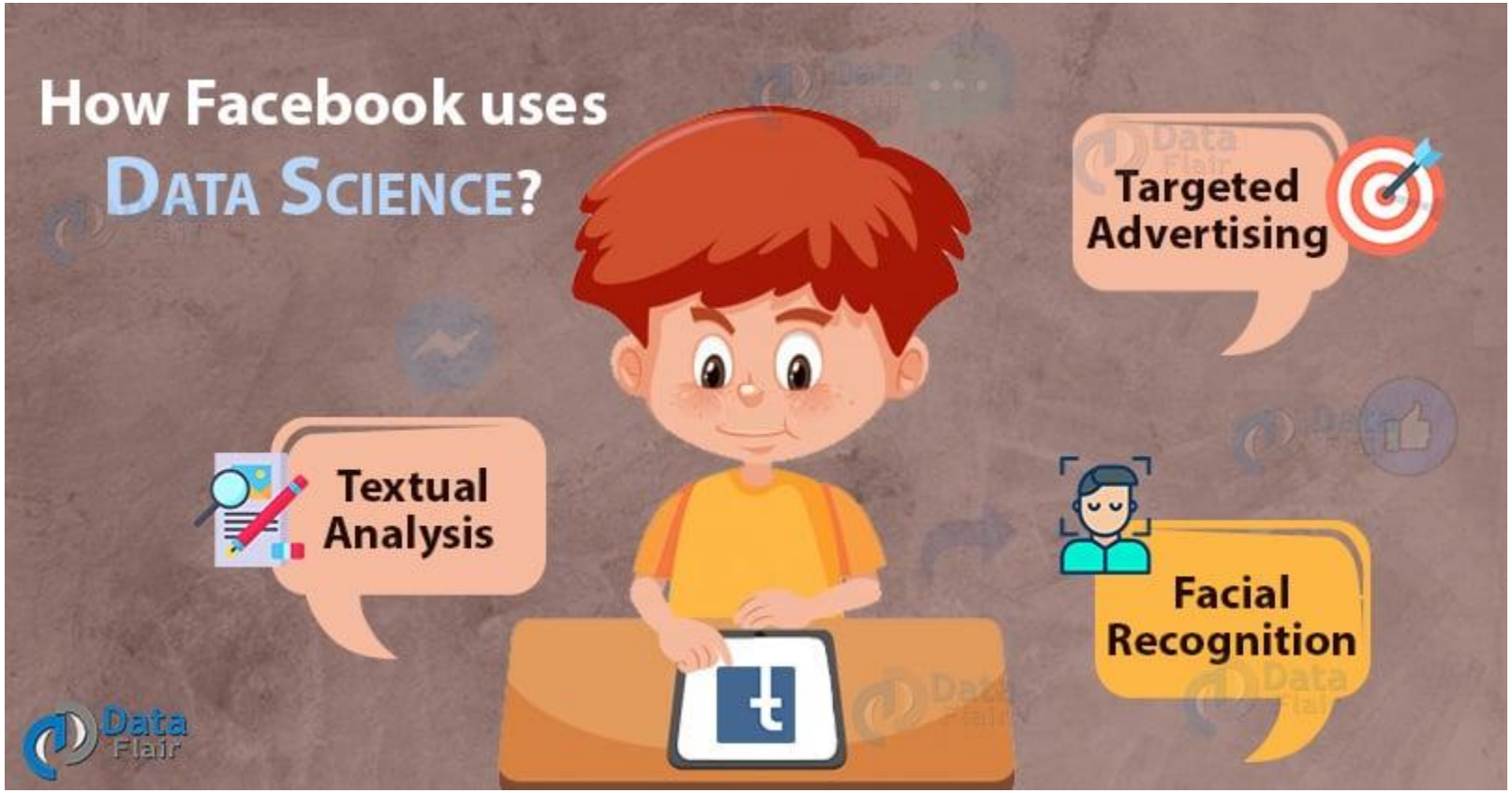
- Tangible Benefits in Various Industries
- From Healthcare to Finance to E-Commerce

Some success stories…

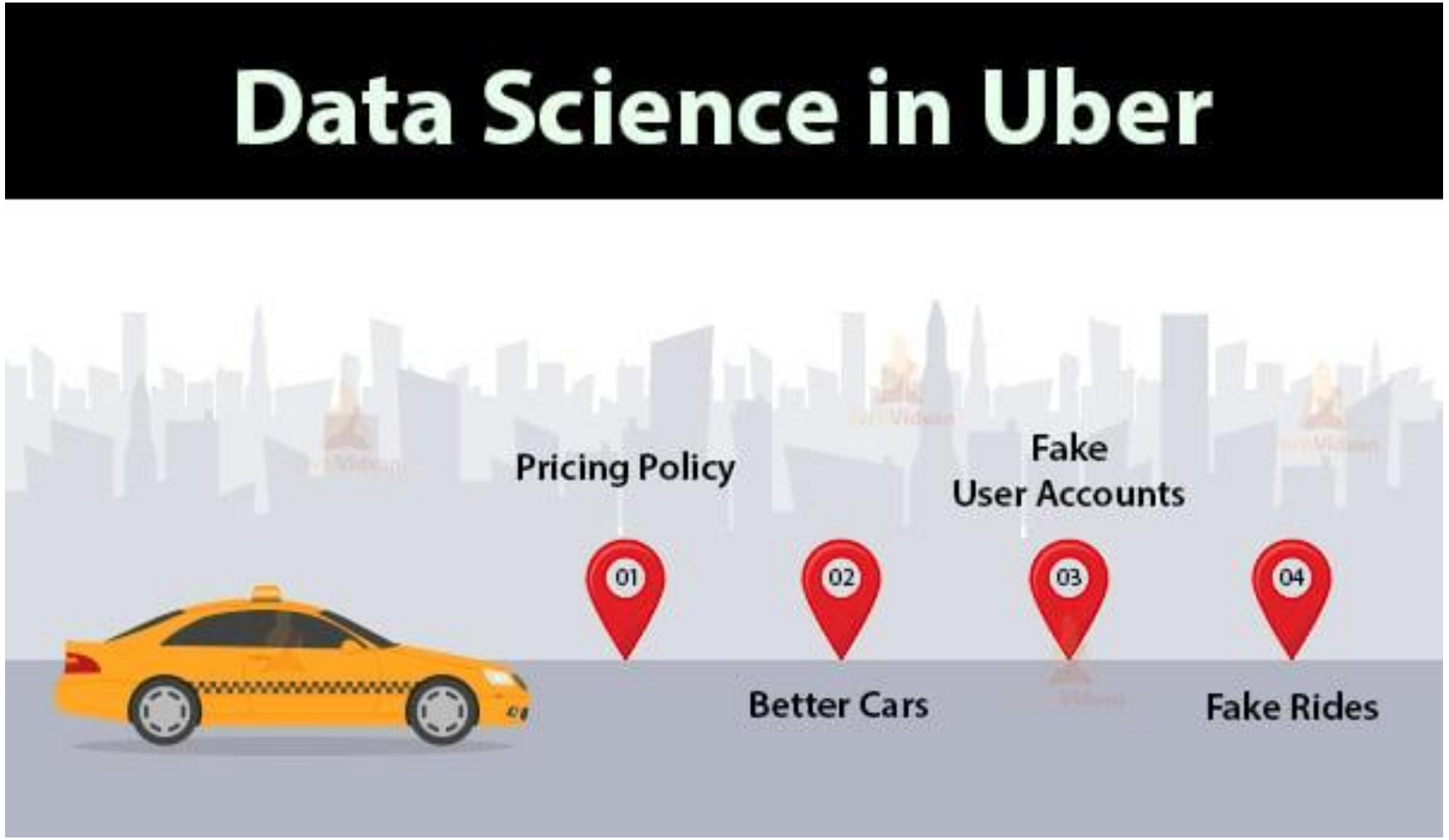
Amazon – Transforming E-commerce with Data Science



Facebook – Using Data to Revolutionize Social Networking & Advertising

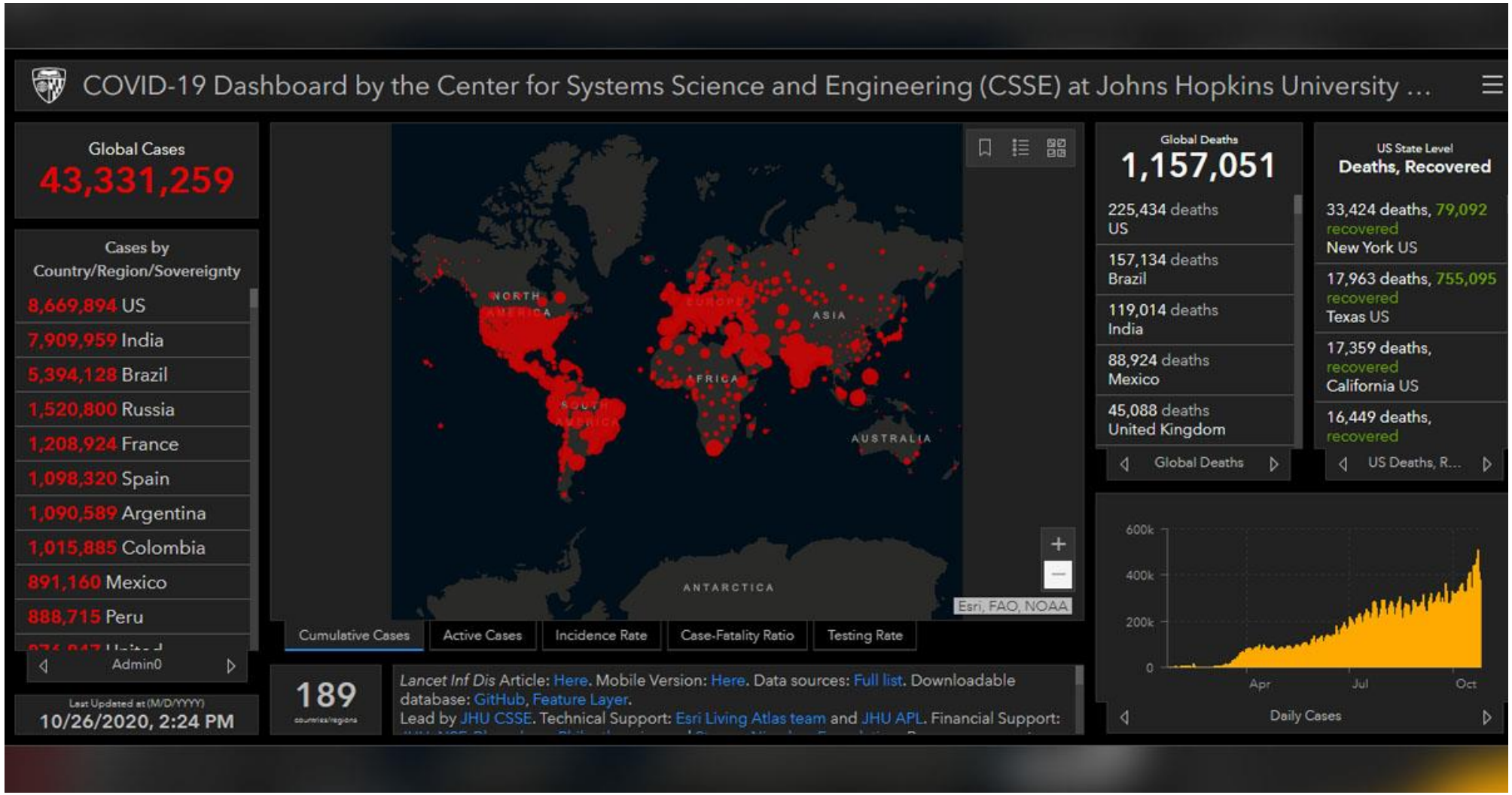


Uber – Using Data to Make Rides Better



Spotify – Revolutionizing Music Streaming

Covid-19 – Using Data to Track the Pandemic and Forecast Hotspots



Rolls-Royce – Data science in manufacturing

How Rolls-Royce 100x'ed the speed of their engineering design processes

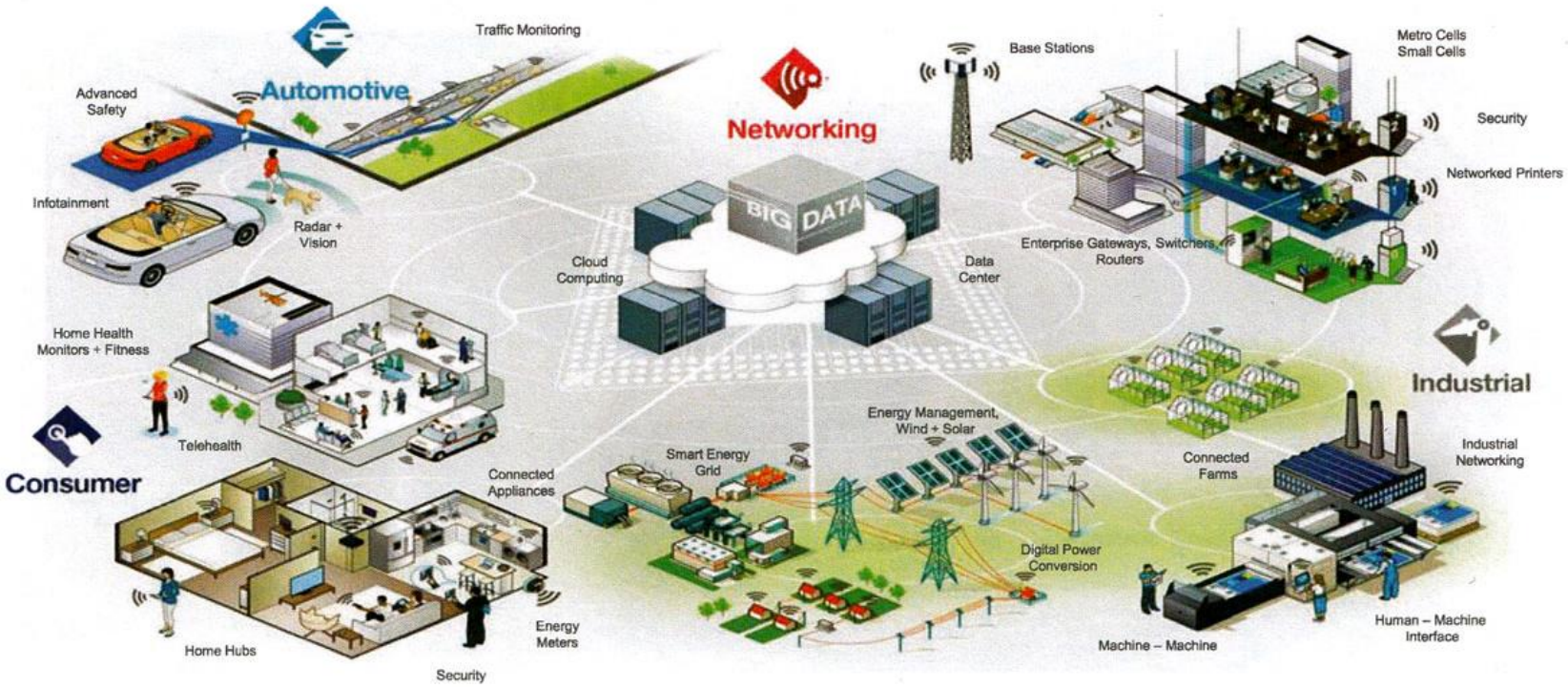
By leveling up their data skills, Rolls-Royce could identify and automate manual processes and save its engineers time to work on more valuable initiatives. By increasing the skills of both technical and non-technical roles, Rolls-Royce was able to help everyone better improve their Python, Power BI, and general data literacy.

The power of data

- “In God we trust. All others must bring data.”
- “Without data you're just another person with an opinion.”

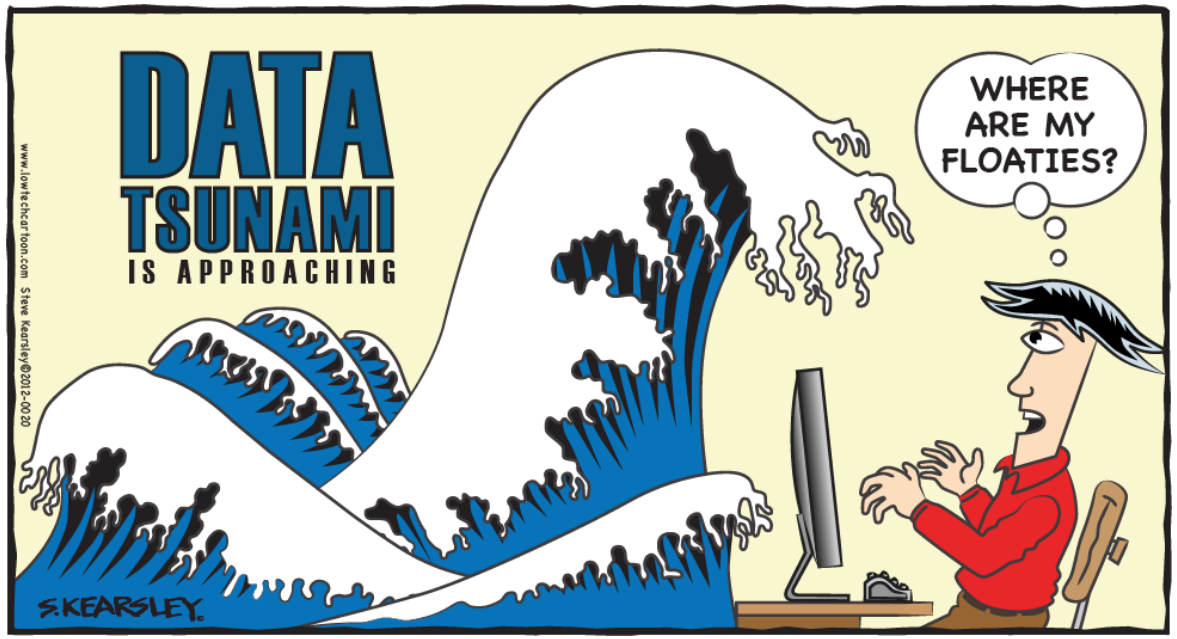
Dr. W. Edwards Deming

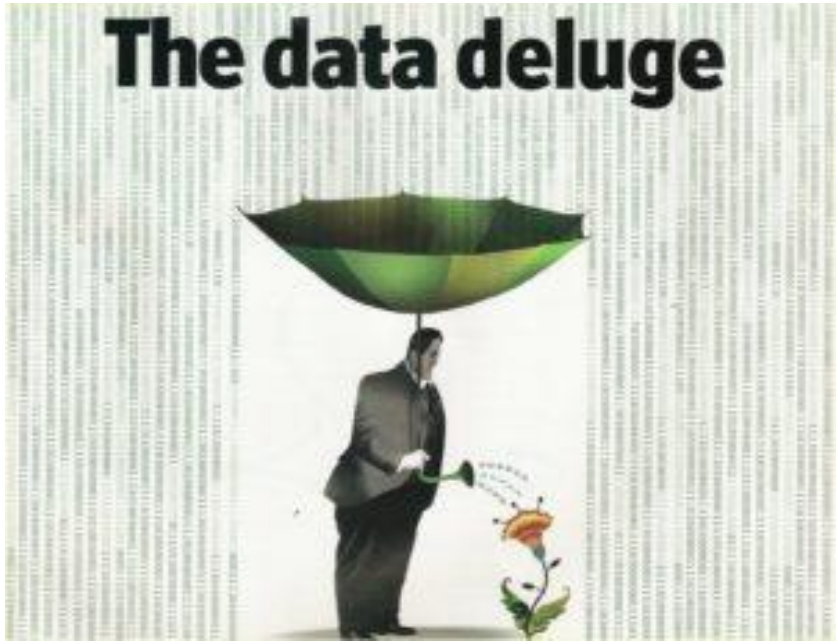
Where does data come from?



The world is digital : Every interaction, transaction, and communication now leaves a digital footprint.

The tsunami of data keeps growing!





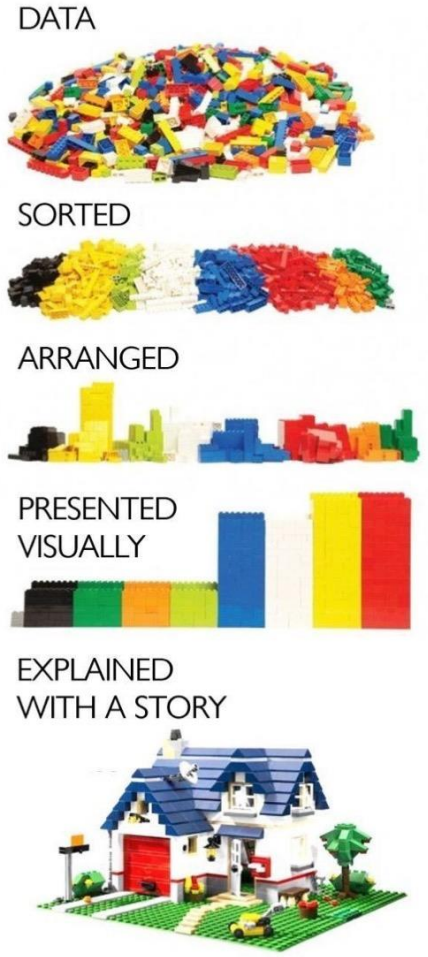
According to the latest estimate, approximately 328.77 million terabytes of data are created each day in 2023; 181 zettabytes of data will be generated in 2025

Drowning in data but starving for insights

"Data are just summaries of thousands of stories—tell a few of those stories to help make the data meaningful."

~ Dan Heath,

- Data itself is useless. It's only when you analyze it, and make it actionable that it becomes valuable.



The real value add is in the storytelling to make the data comprehensible and actionable by humans.

Terence Kawaja

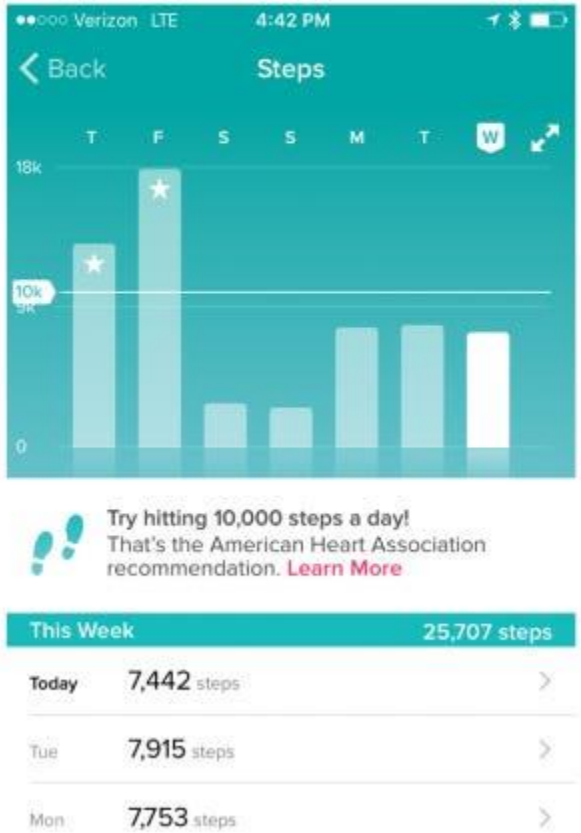
From a post by Geoffrey Colon

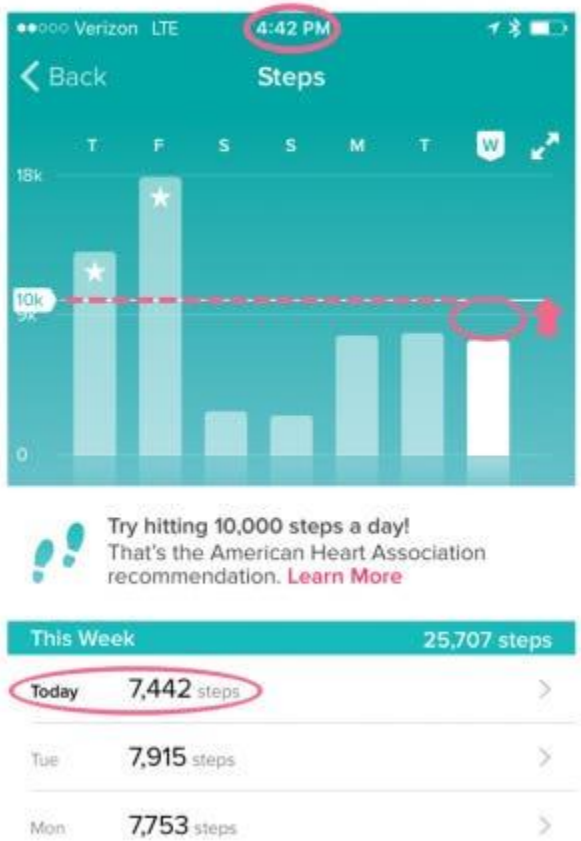
Drowning in data but starving for insights

DATA → INFORMATION → INSIGHT → ACTION

STEPS

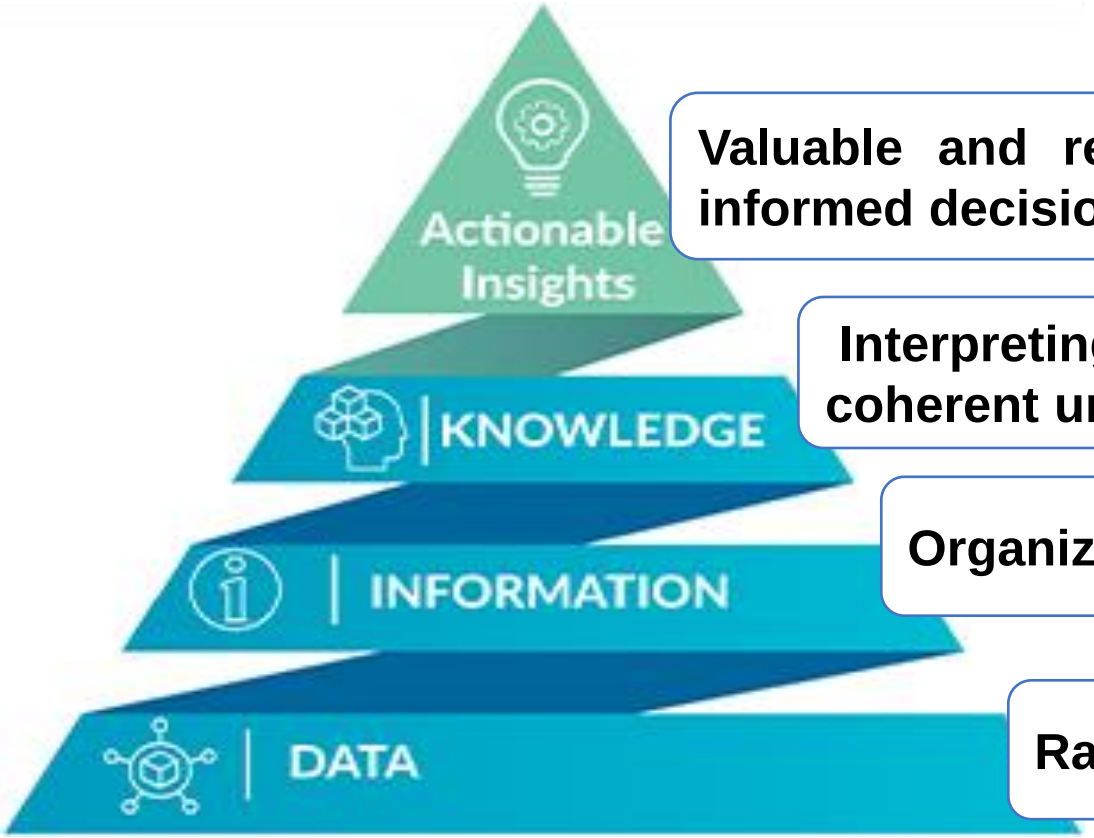
7,442 steps





Towards actionable insights

The insight pyramid



Valuable and relevant information used to make informed decisions

Interpreting and linking information to form a coherent understanding.

Organized data that conveys meaning

Raw and unprocessed with no meaning

Data, Information, Knowledge and Insights

Data

150, "John Smith", 2.5,...

Information

John Smith's height is 150 cm.

Knowledge

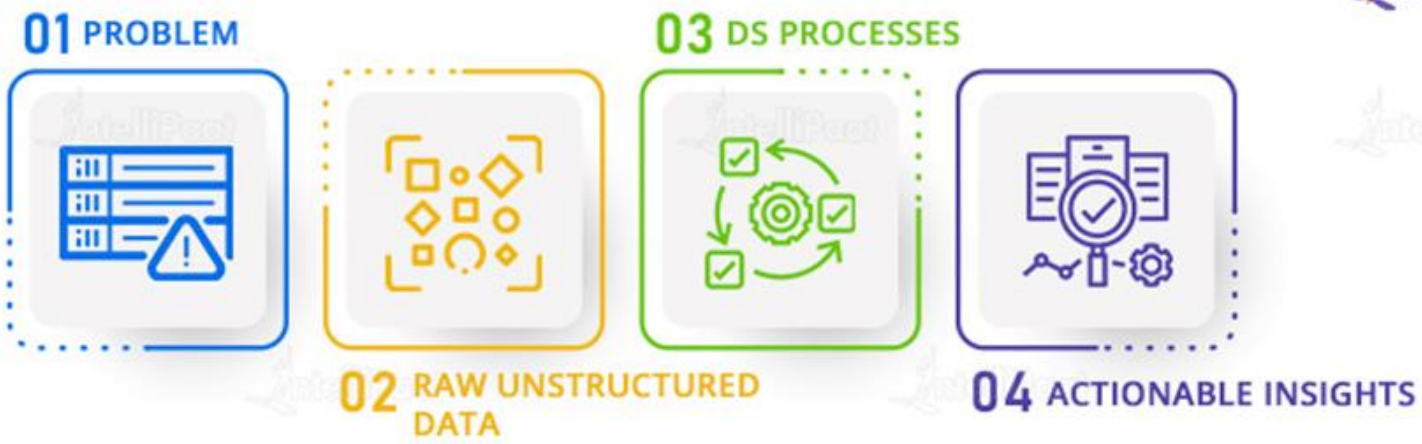
John Smith's height is below average for his age group, indicating a potential growth issue.

Insights

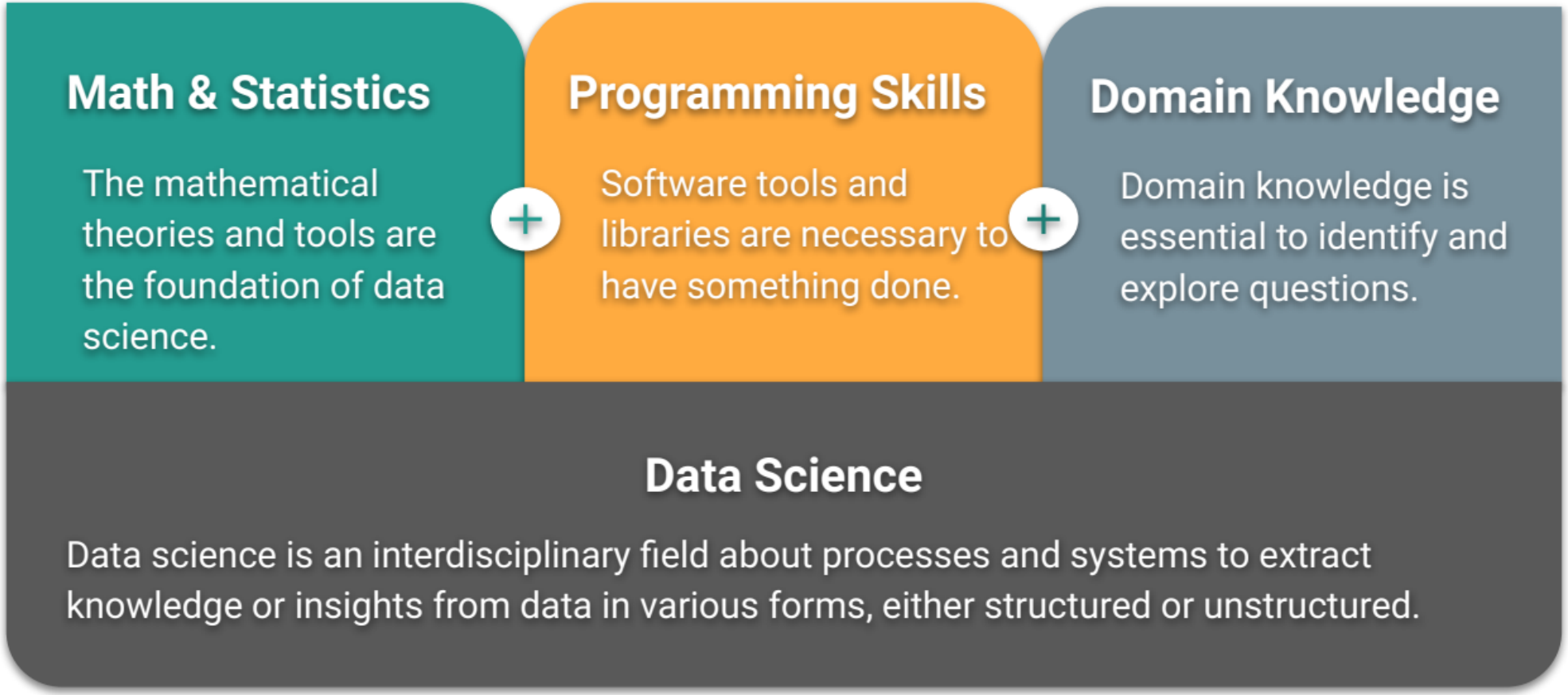
John's growth issue might be related to a medical condition that requires specialized treatment, and early intervention can significantly improve his prognosis.

The quest for Meaning: from data to insights

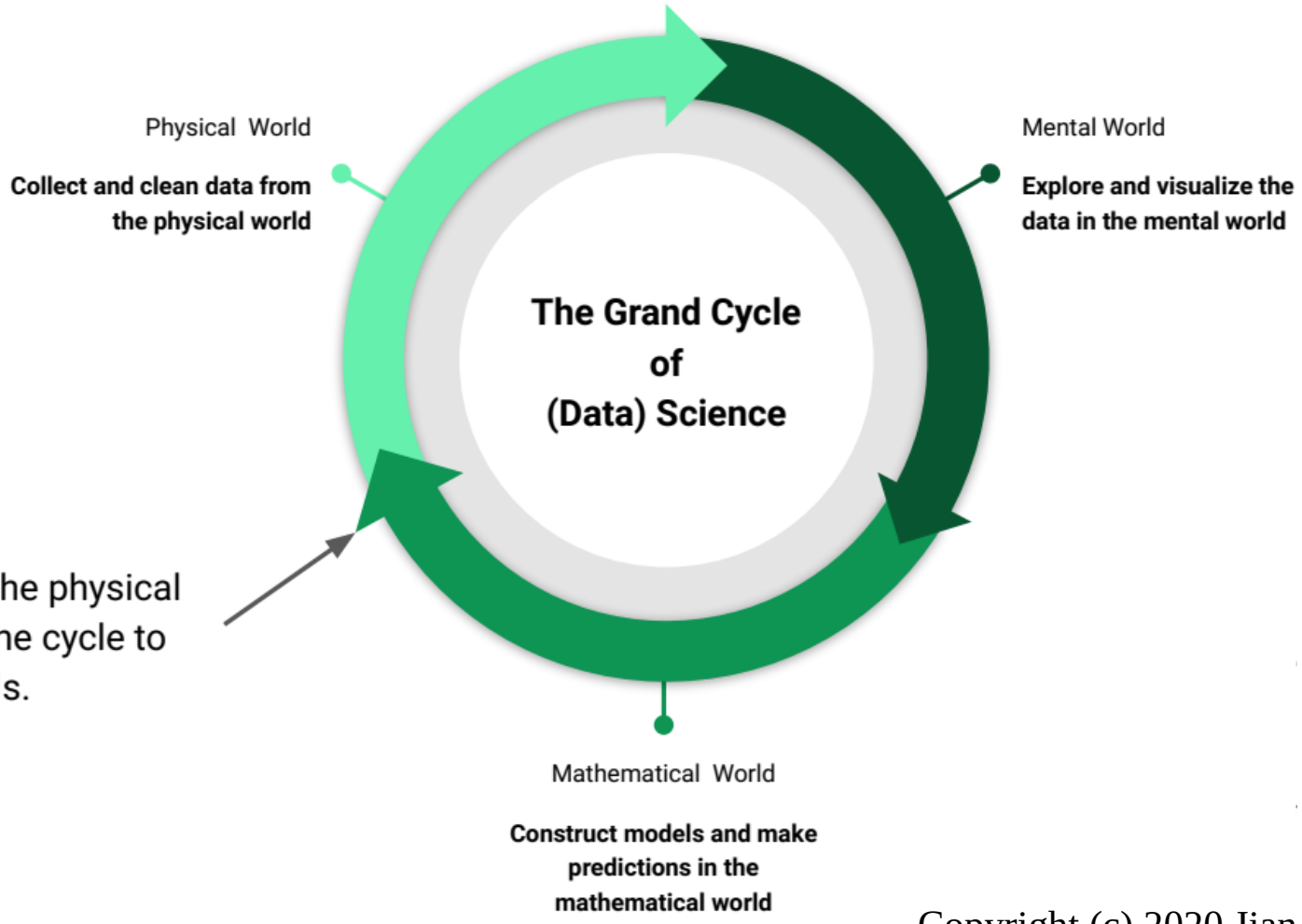
- Data science is the key to making sense of this digital deluge.
- It is the art of transforming data into actionable insights and valuable knowledge.



Data science is multidisciplinary



The data science cycle



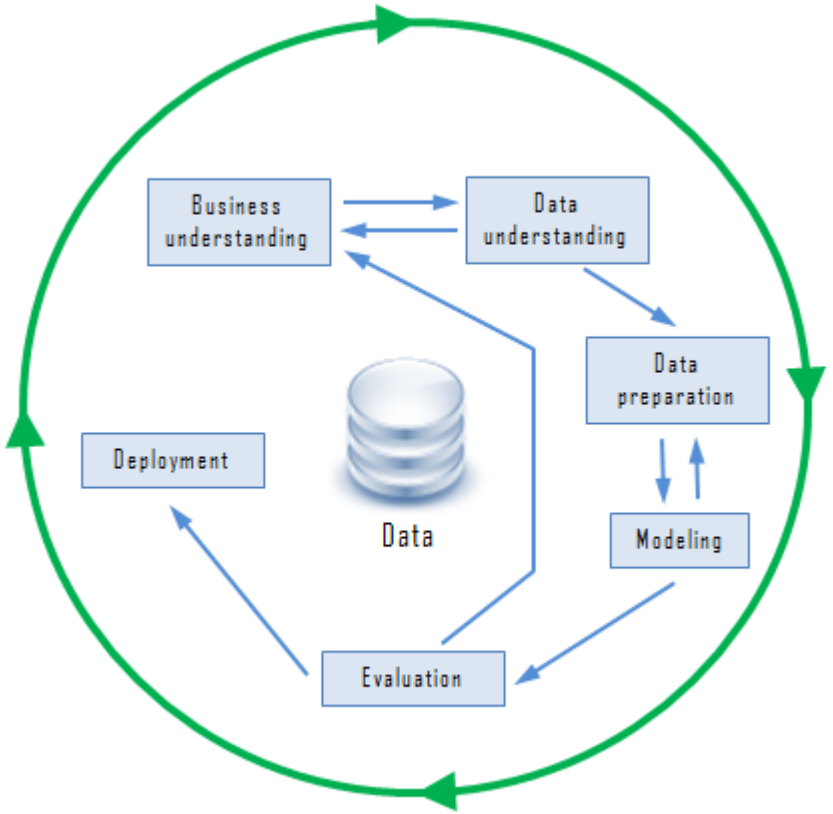
Deploy models in the physical world and repeat the cycle to improve the models.

Data science is a process

CRISP-DM

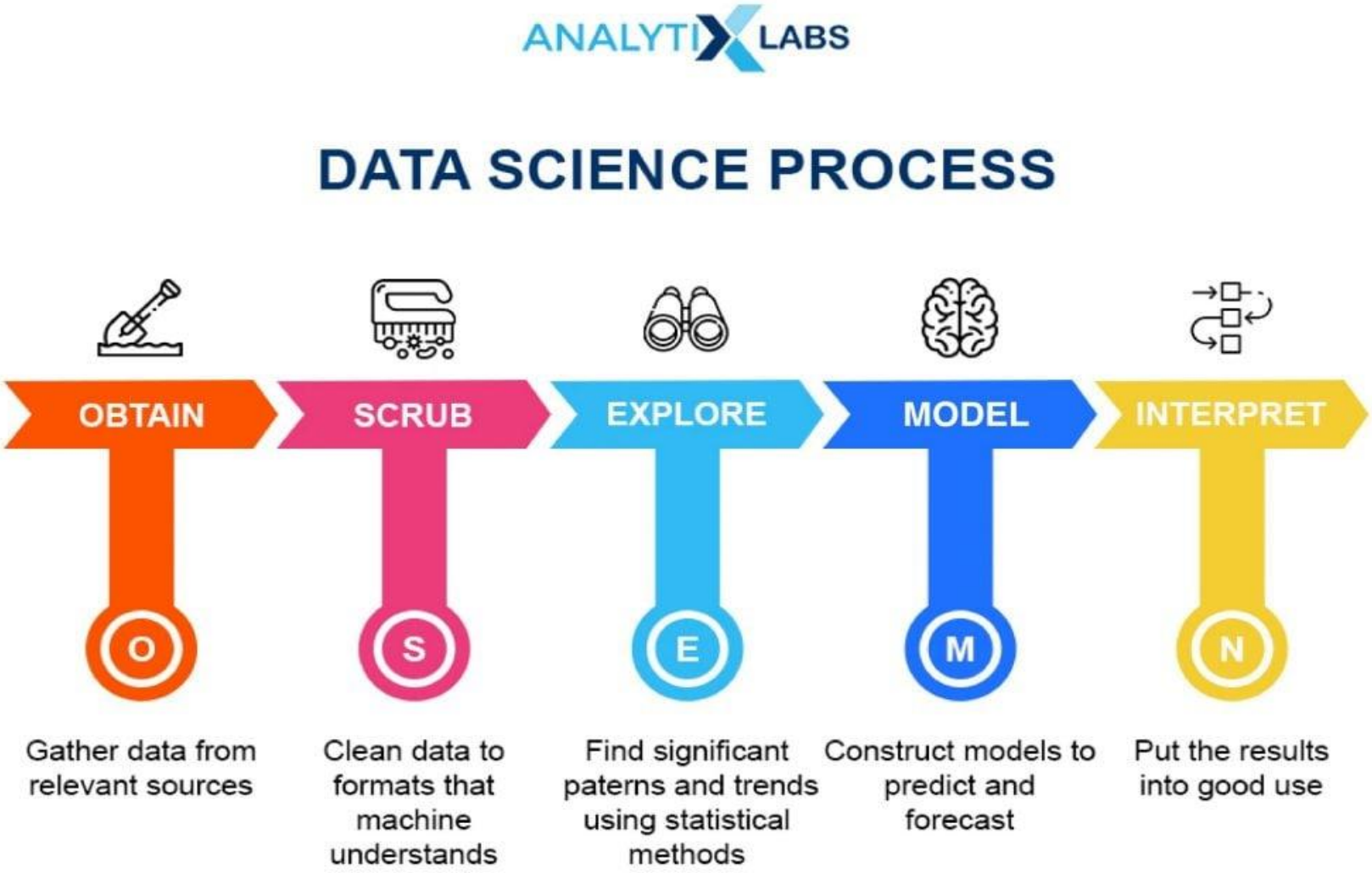
- Cross-industry standard process for data mining
- An open standard process model that describes common steps followed for data mining and analytics projects

- It is the most widely-used analytics model



Cross-industry Standard Process for Data Mining (CRISP-DM)

The OSMEN ("Awesome") DS process!



Prerequisite step: Business understanding

"Far better an approximate answer to the right question, which is often vague, than the exact answer to the wrong question, which can always be made precise."

— John Tukey

What is the Business Question we are trying to answer?

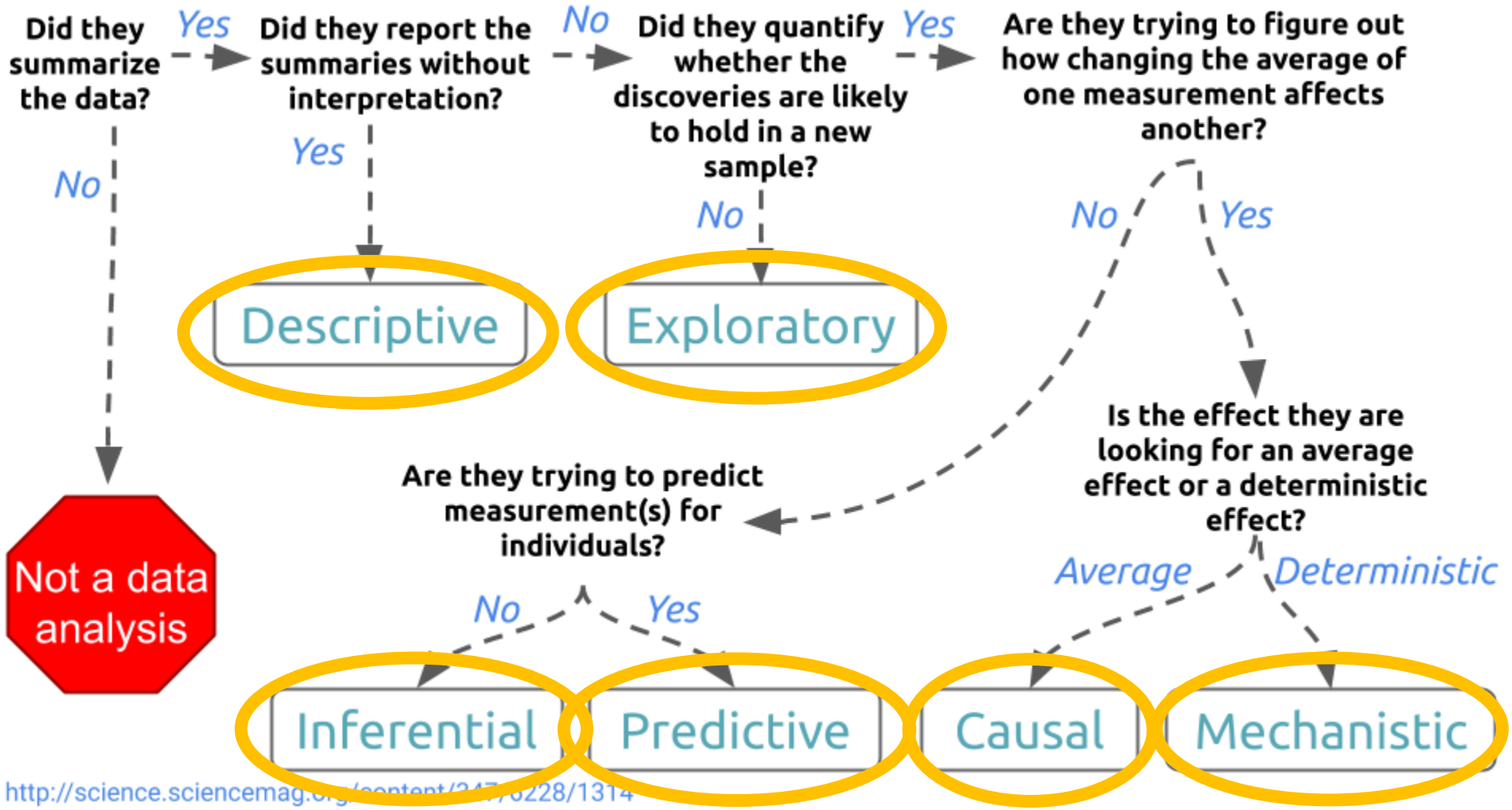
Asking the right question sets up the rest of the path.

Questions are important!

"Good data science is more about the questions you pose of the data rather than data munging and analysis." — Riley Newman

- **First**, formulate the questions that you will use the data to solve.
- The more questions you ask of the data, the more insight you will get.

Types of Data Science questions



http://science.sciencemag.org/content/347/6228/1314

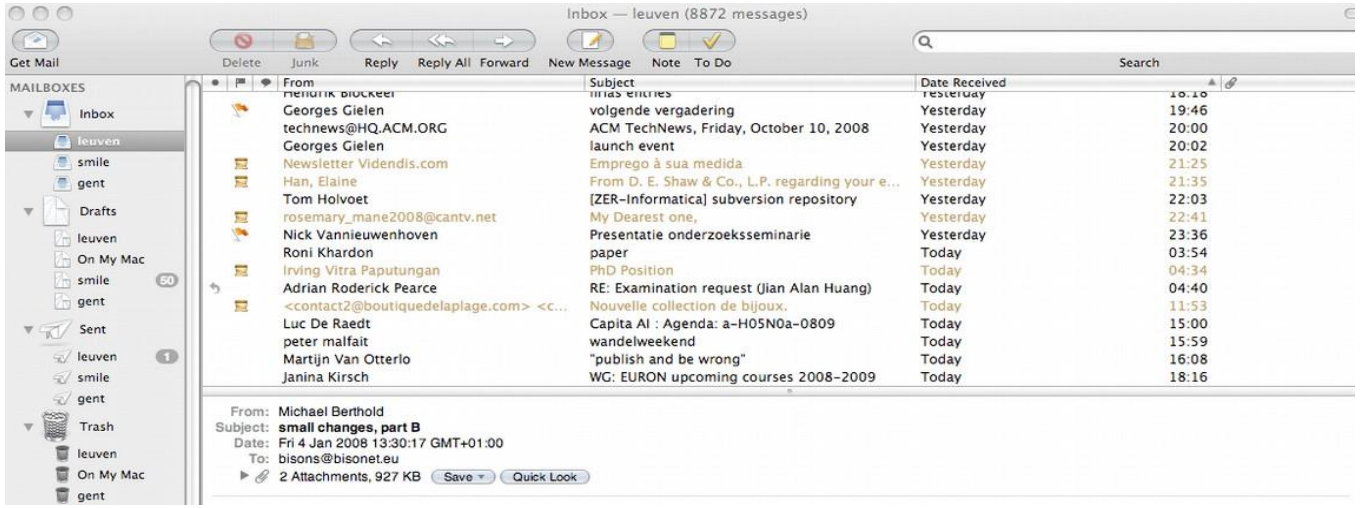
Types of Data Science questions

- **Descriptive:** Describe and summarize data (e.g. Covid-19 statistics)
- **Exploratory:** Search for unknown relationships, new discoveries, **without identifying the cause**
- **Inferential:** What data say about larger population..
- **Predictive:** use Historical data to make **predictions** for the future
- **Causal:** Explore **causation**; what happens to variable X when variable Y change.
- **Mechanistic:** Determine governing principles; what **exact changes** in variables lead to changes in other variables.

Example of Data science workflow

- **Business question**

SPAM Filter



SPAM email reduces productivity, automatically remove it

How can we effectively distinguish between legitimate and spam emails?

DS workflow: SPAM Filter

- **Obtaining data**
  - Collect messages, in general and from the user, that are spam (negative) and legitimate (positive): acquisition, annotation, …
  - Given a text message, predict whether it is spam or not

→ text categorization, useful in general

→ we want a <u>function</u> from message to {0,1}

→ is called binary classification problem

DS workflow: SPAM Filter

- **Scrubing data**

Given a raw text, convert string data into numerical data one

- Bag of words, TFIDF, Word2Vec

Text Preprocessing

1. Remove Noisy Data: header, footer, HTML, XML, markup data
2. Tokenization: word, character, and subword (n-gram characters)
3. Normalization: converting all words to lowercases

DS workflow: SPAM Filter

- **Exploring the cleaned data (EDA)**
  - Explore to understand
    - the distribution of features,
    - correlations, and patterns related to transactions.

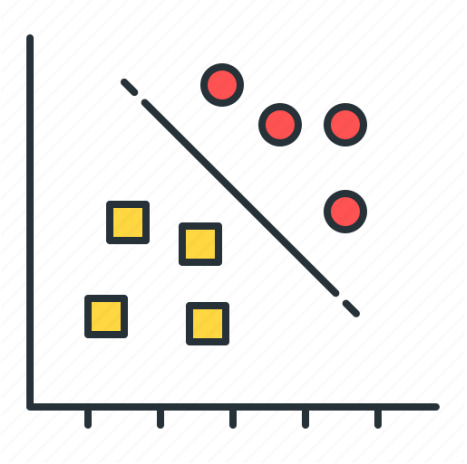
  - Use plots (e.g. histograms, scatter plots, or heatmaps) to visualize transaction amounts, time distributions, and other relevant features to identify potential trends.

DS workflow: SPAM Filter

- **Modeling**
  - We could write a rule-based system, such as

    if Title.contains(“YOU HAVE WON!!!”) then return Spam

  - train a classifier (e.g. naïve bayes)
  - Does it work well? → evaluate



DS workflow: SPAM Filter

- **Modeling**

Evaluate and refine

on unseen emails

		Truth	
		Spam	Legitimate
Predicted as	Spam	150	30
	Legitimate	200	720

False positives

False negatives

DS workflow: SPAM Filter

- **Interpreting (Data Storytelling)**

Data-driven battle against spam

- Data, the weapon that safeguard inboxes

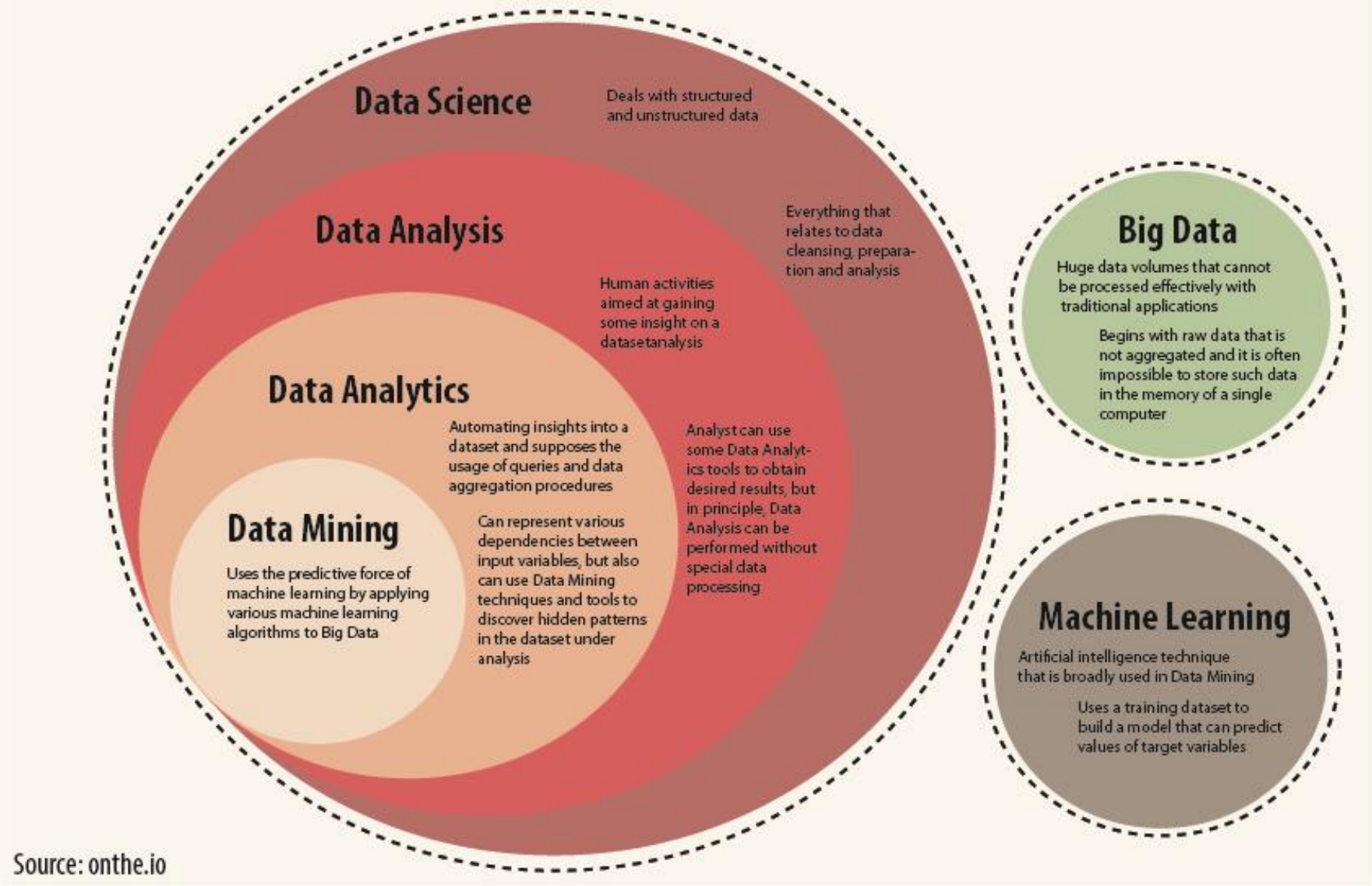
Display :

- a word cloud showcasing prevalent words in spam emails.
- bar charts or histograms depicting the frequency of specific terms or patterns in spam versus non-spam emails.
- Confusion matrix or ROC curve illustrating the model's performance in detecting spams.

DS in summary,

- Form your **Business Problem**
- **Obtain** your data
- **Clean** your data,
- **Explore** your data with visualizations,
- **Model** your data using statistical or machine learning models,
- **iNterpret** your analysis results,

Data Mining & Data Science



Data Mining & Data science

	Data Mining	Data Science
01	More involved with its processes	Broadly focuses on the science of data
02	Primarily used for business purposes	It is essentially implemented for scientific purposes
03	Data mining is a technique that is a part of the KDD process	Data science is a field of study
04	Primarily deals with structured data	It deals with all types of data - structured, unstructured, or semi-structured
05	It is about extracting valuable information from data	It is about collecting, & processing, analyzing & utilizing data in various operations
06	It is a subset of data science as mining activities are in the pipeline of data science	Involves data scraping, cleaning, visualization, stats, etc. Therefore, it is a superset of data mining
07	Its objective is to realize the value of data & make it usable by extracting important Info.	Objective is to build data-dominant products for a venture

Course Content

- **Data Wrangling and Cleaning**
  - Data collection and preprocessing
  - Handling missing data and outliers
  - Data transformation and feature engineering
- **Exploratory Data Analysis (EDA)**
  - Descriptive statistics and summary metrics
  - Correlation analysis and dimensionality reduction
- **Classification and prediction**
  - Supervised learning : SVM…
  - Ensemble learning
  - Evaluation Metrics for Classification Models
  - Model Overfitting and regularization