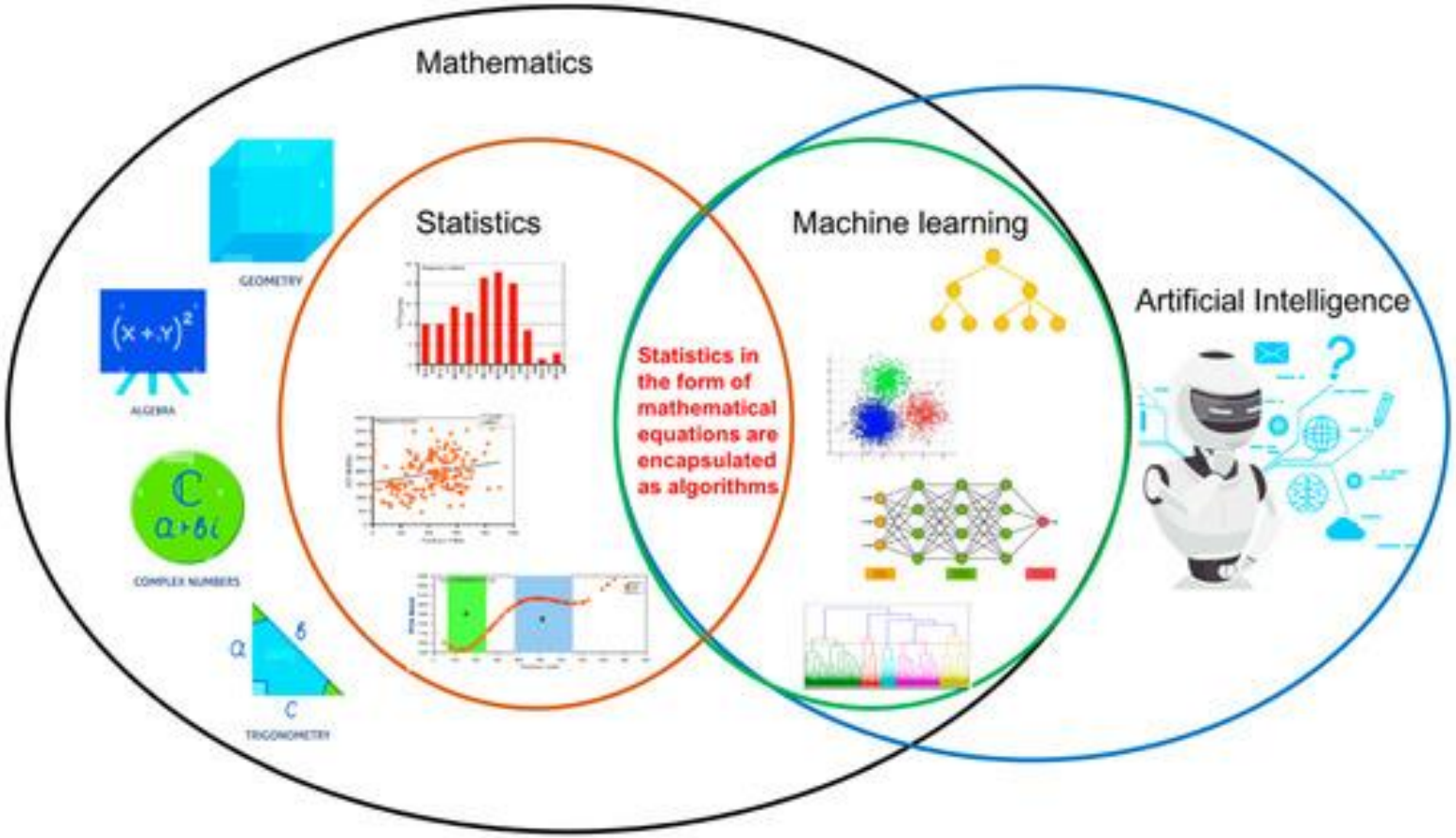
Data Science

1

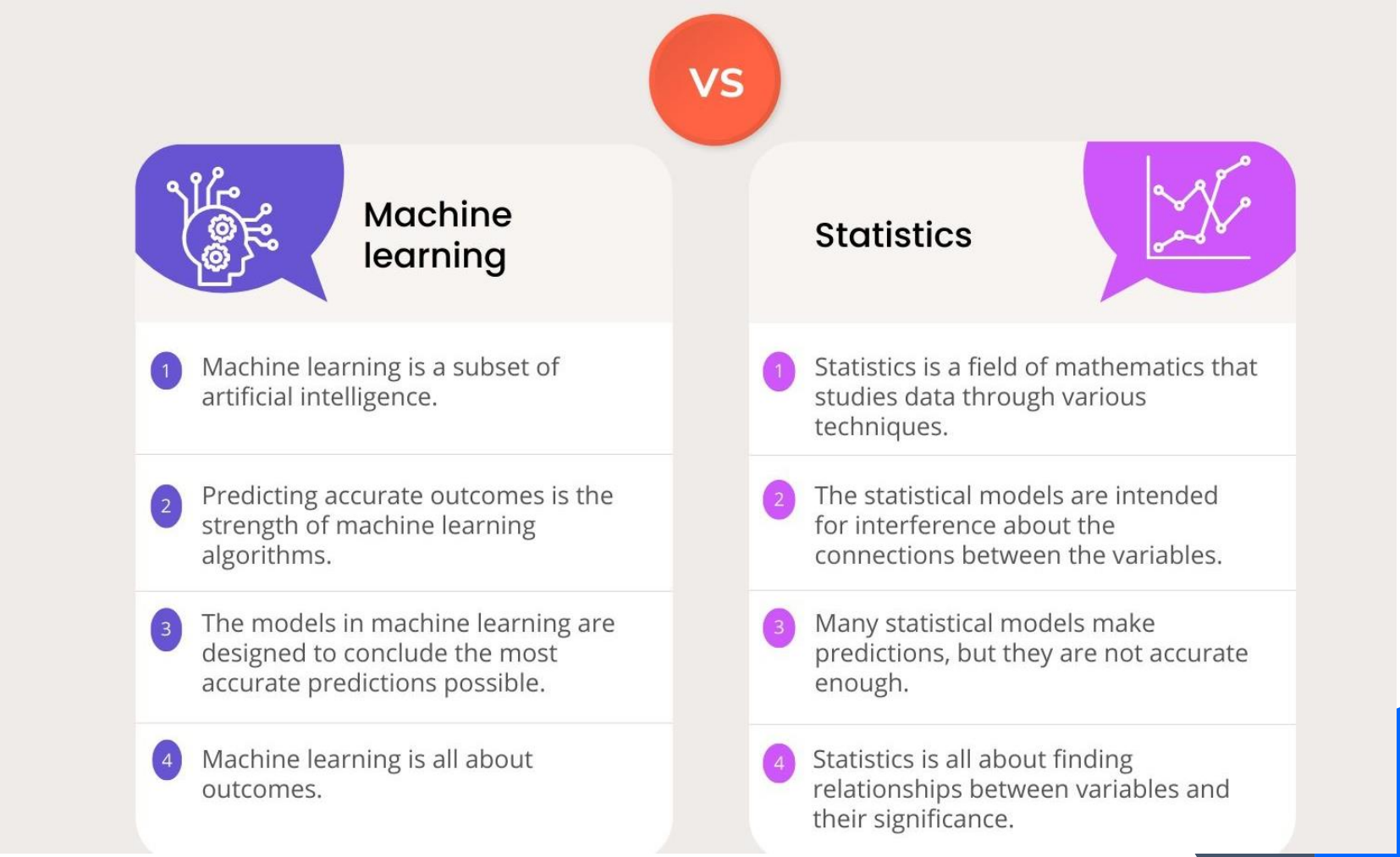
Evolution of Big Data, Data Science and Data Analytics

Transformation from conventional statistics to big data



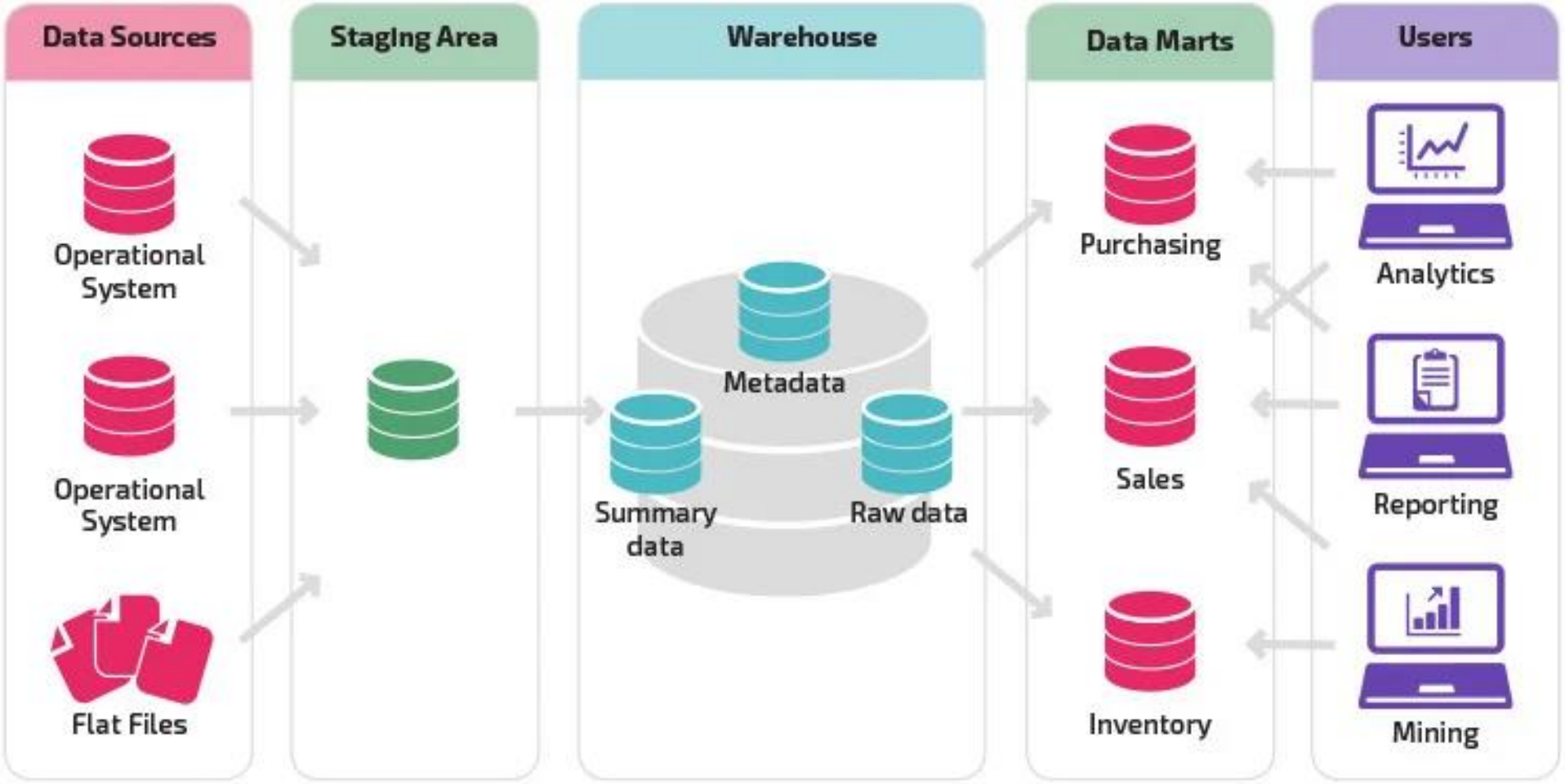
Source: https://www.mdpi.com/2075-4418/12/10/2526

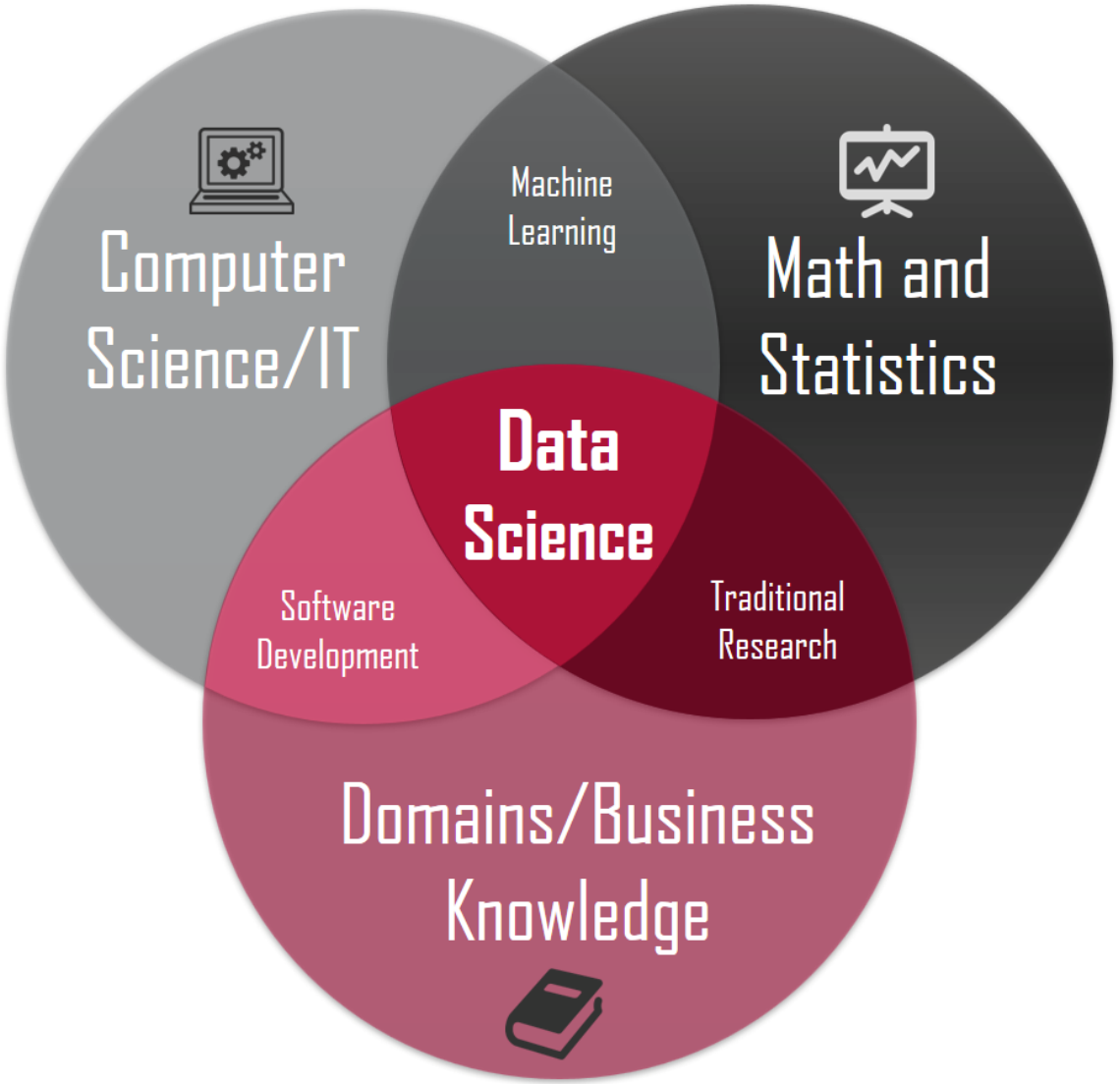
Transformation from conventional statistics to big data



THE 10 Vs OF BIG DATA

Big Data Repositories

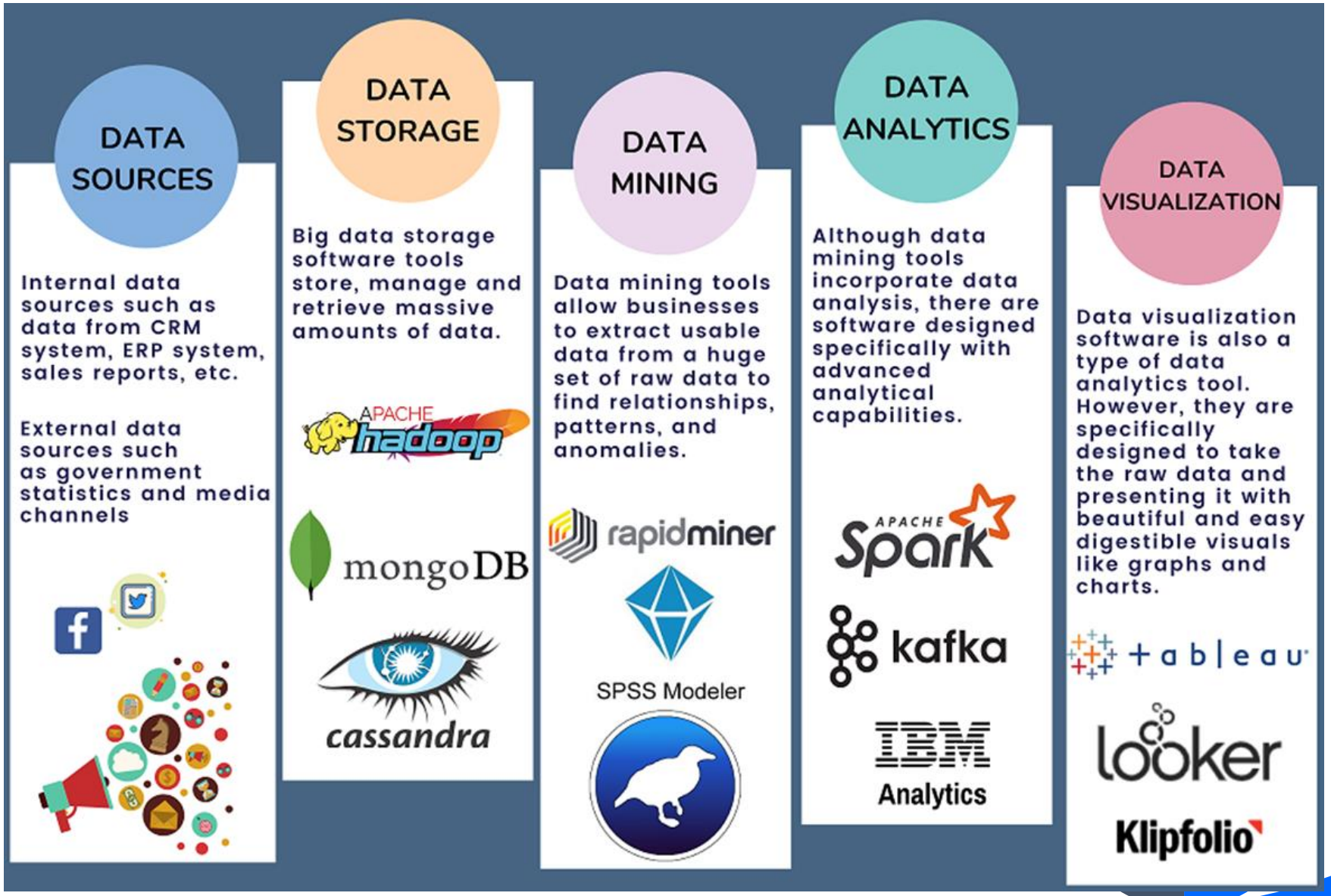




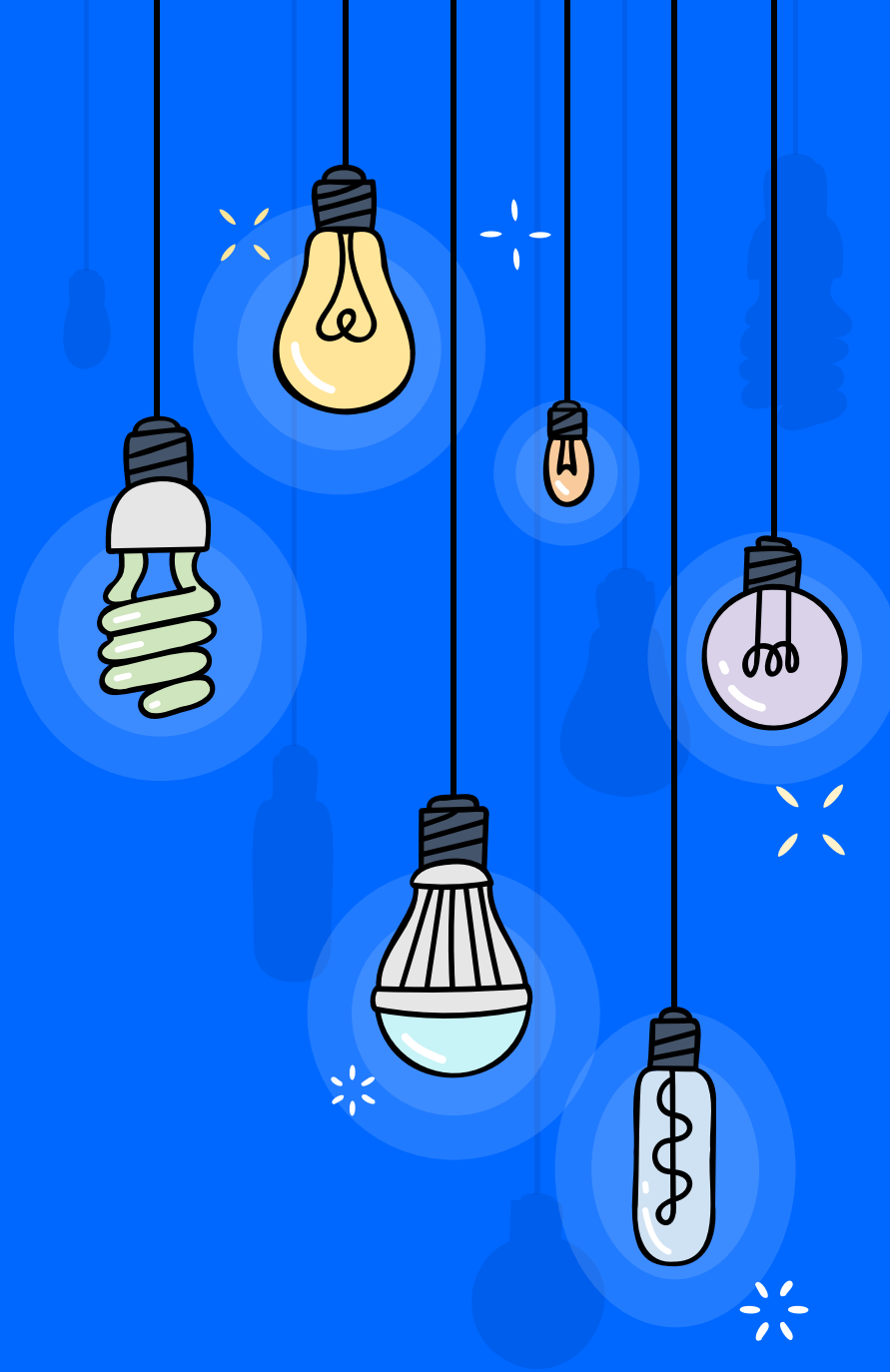
DATA SCIENCE

Domain of study that deals with **vast volumes of data** using **modern tools** and techniques to find unseen patterns, **derive meaningful information**, and make business decisions.

Data Science Technologies



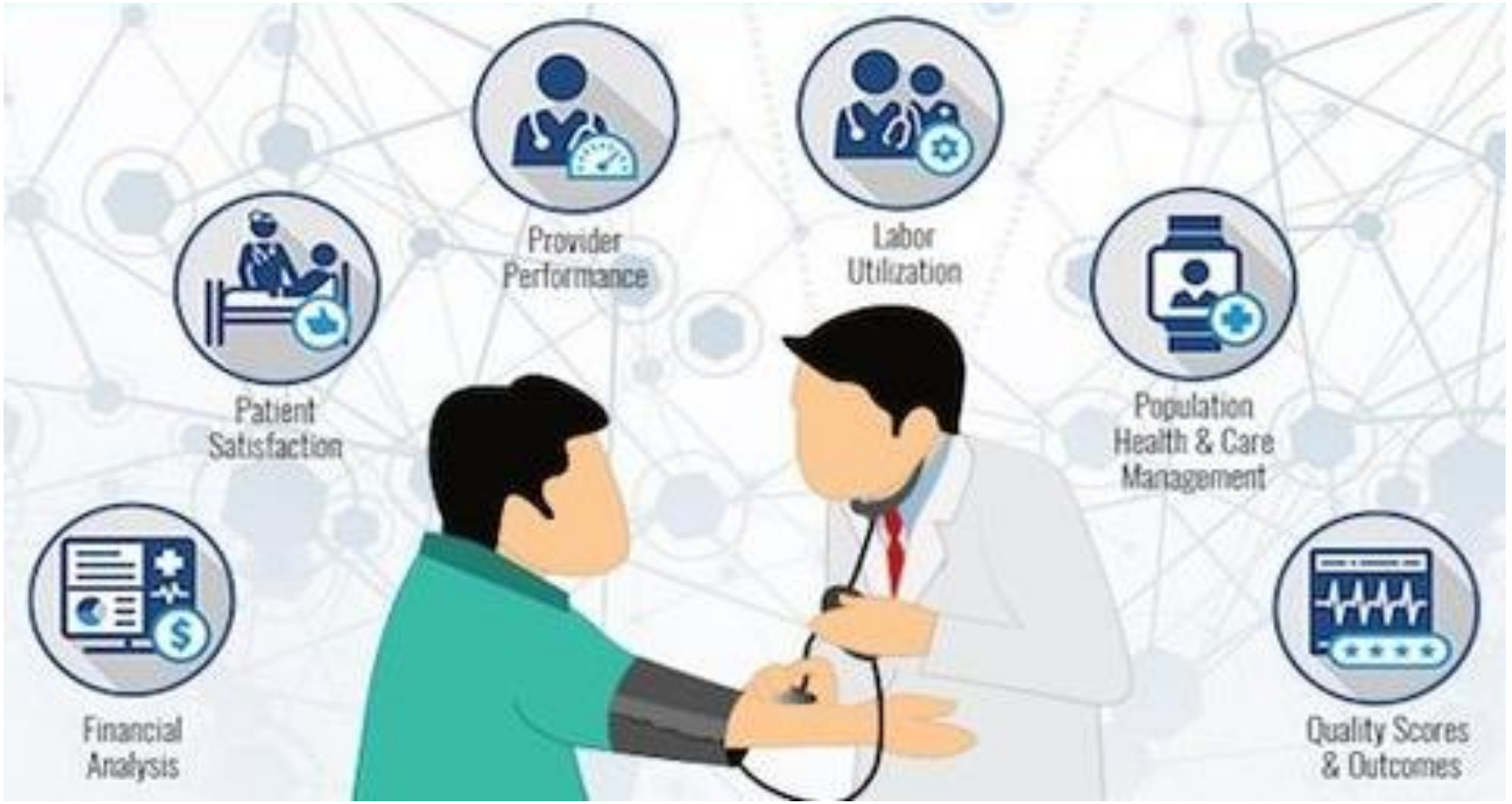
2 Implementation of Big Data in the Public Sector



Data Science Applications

Healthcare

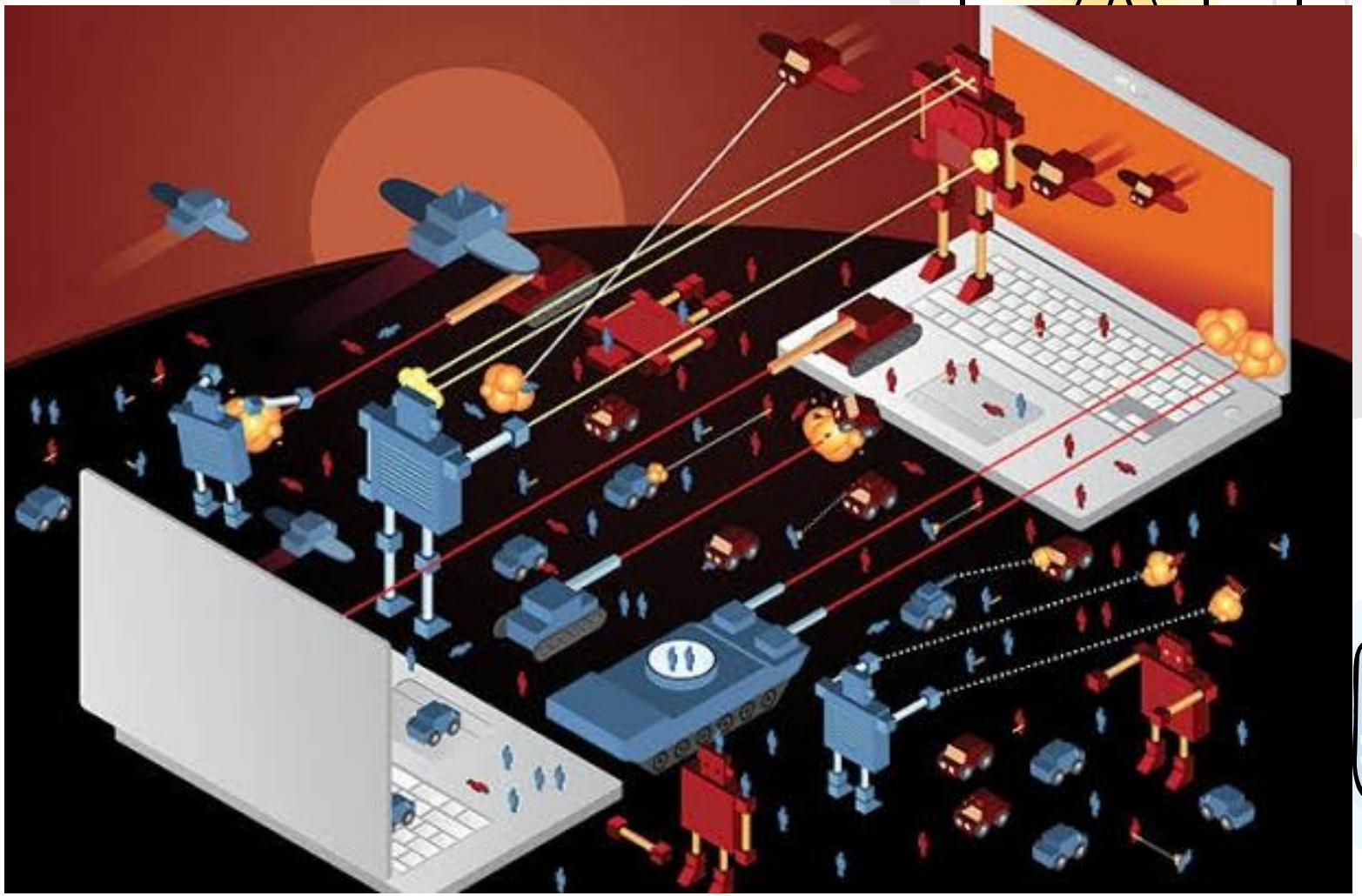
Healthcare companies are using data science to build sophisticated medical instruments to detect and cure diseases.



Data Science Applications

Gaming

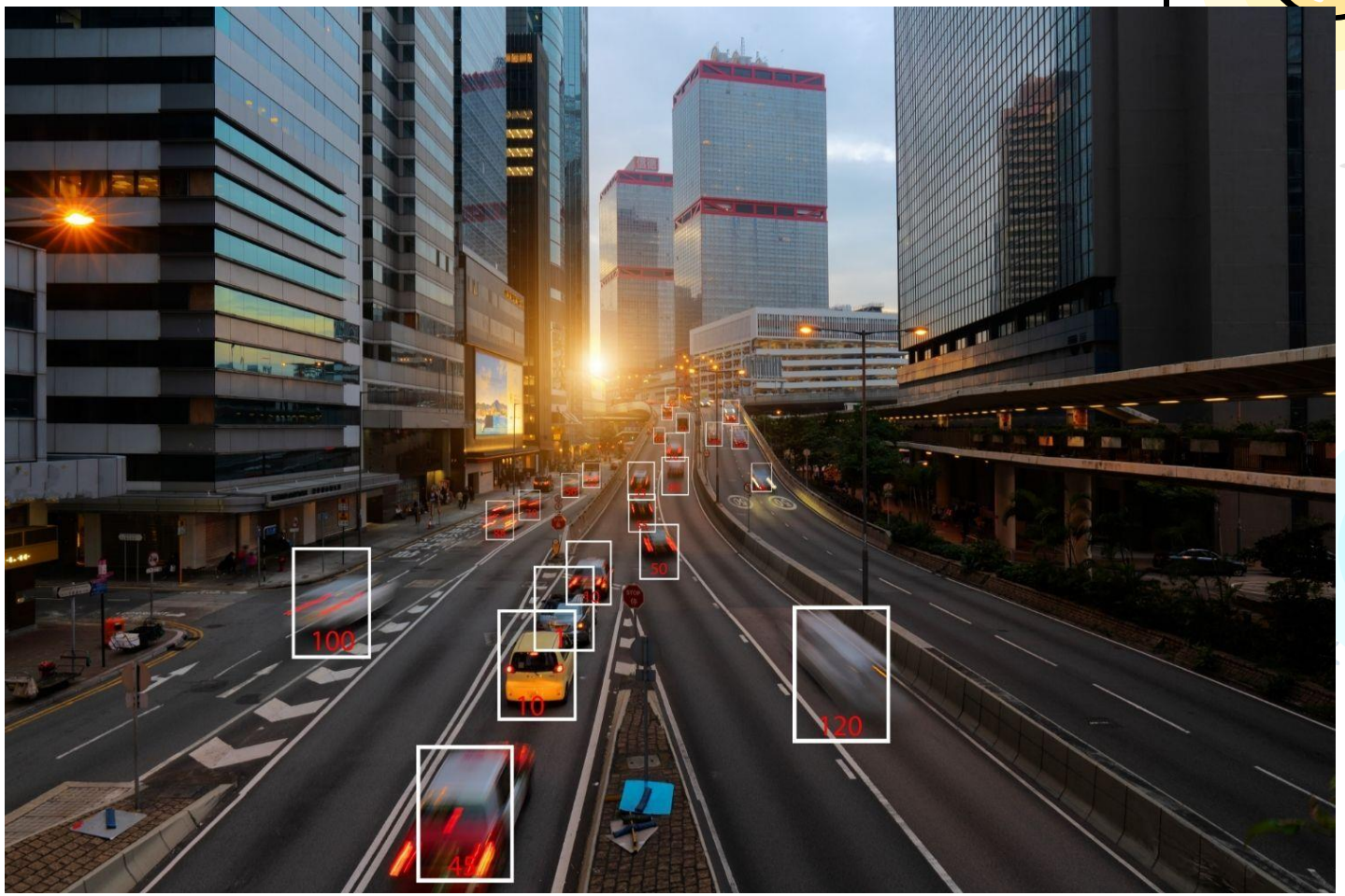
Video and computer games are now being created with the help of data science and that has taken the gaming experience to the next level.



Data Science Applications

Image recognition

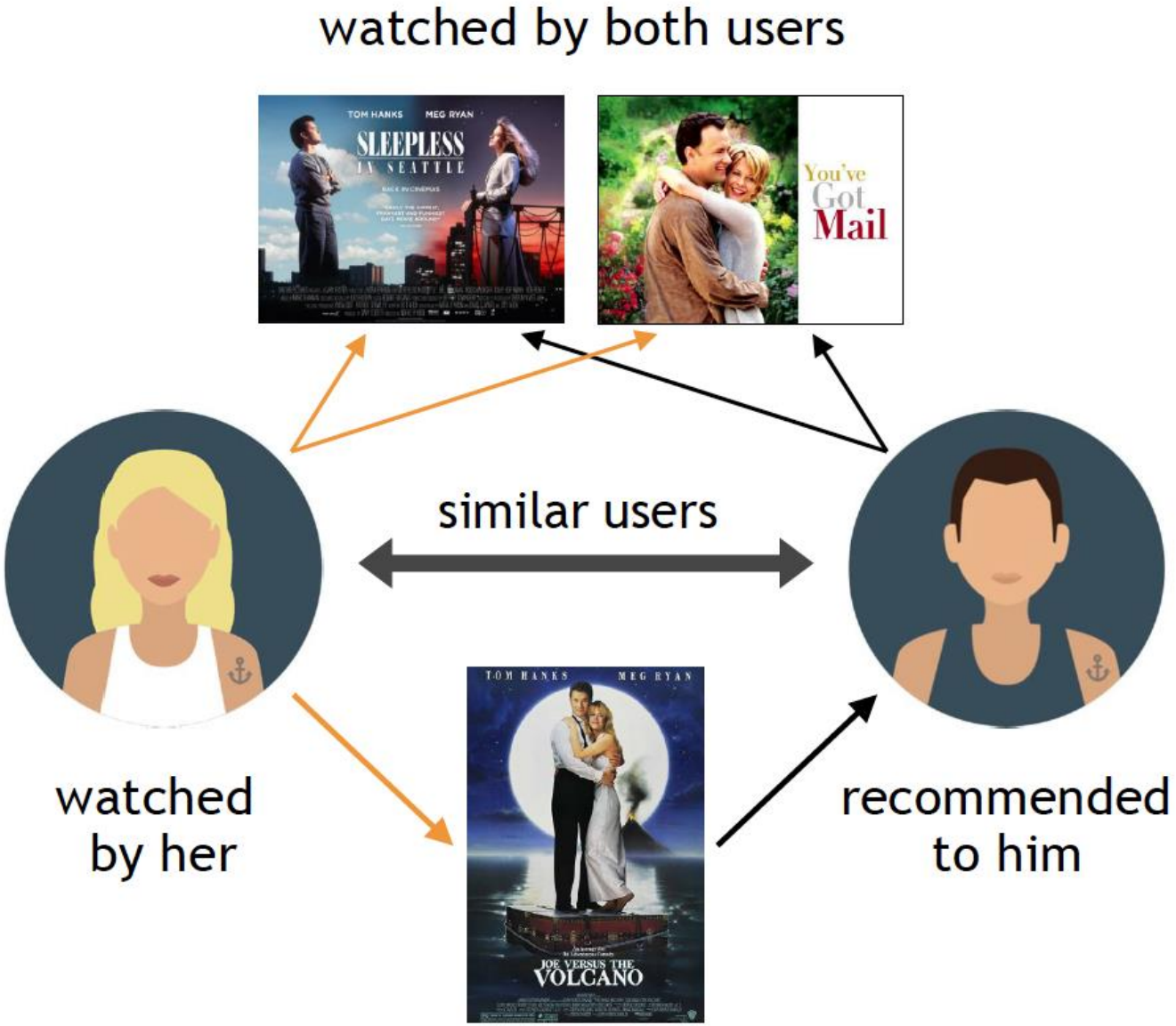
Identifying patterns in images and detecting objects in an image is one of the most popular data science applications.



Data Science Applications

Recommendation systems

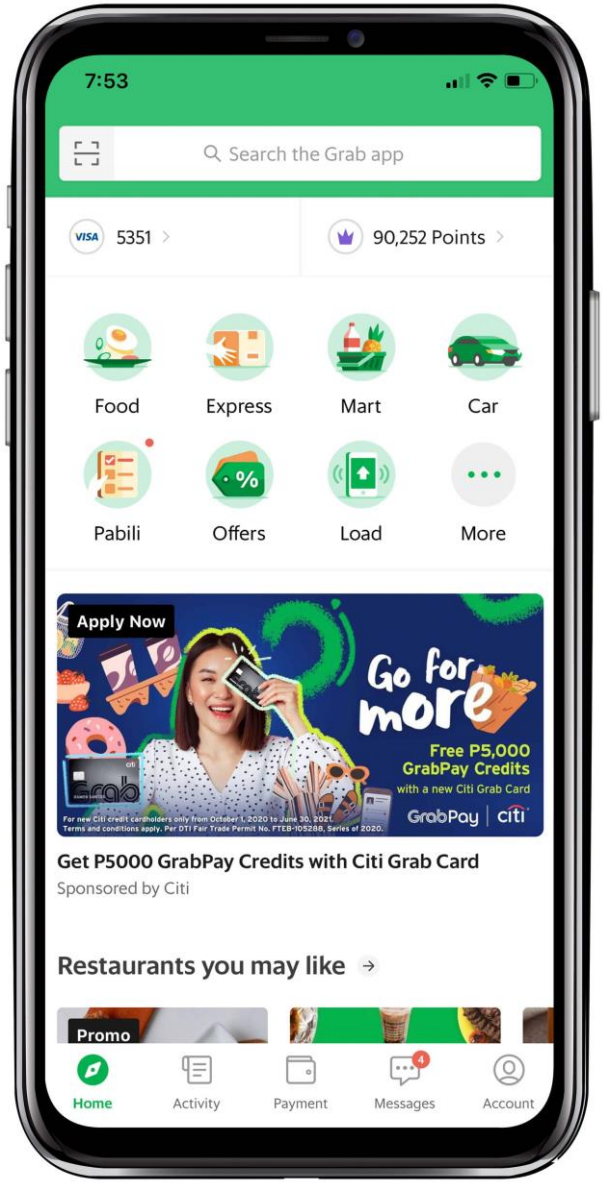
Netflix and Amazon give movie and product recommendations based on what you like to watch, purchase, or browse on their platforms.



Data Science Applications

Logistics

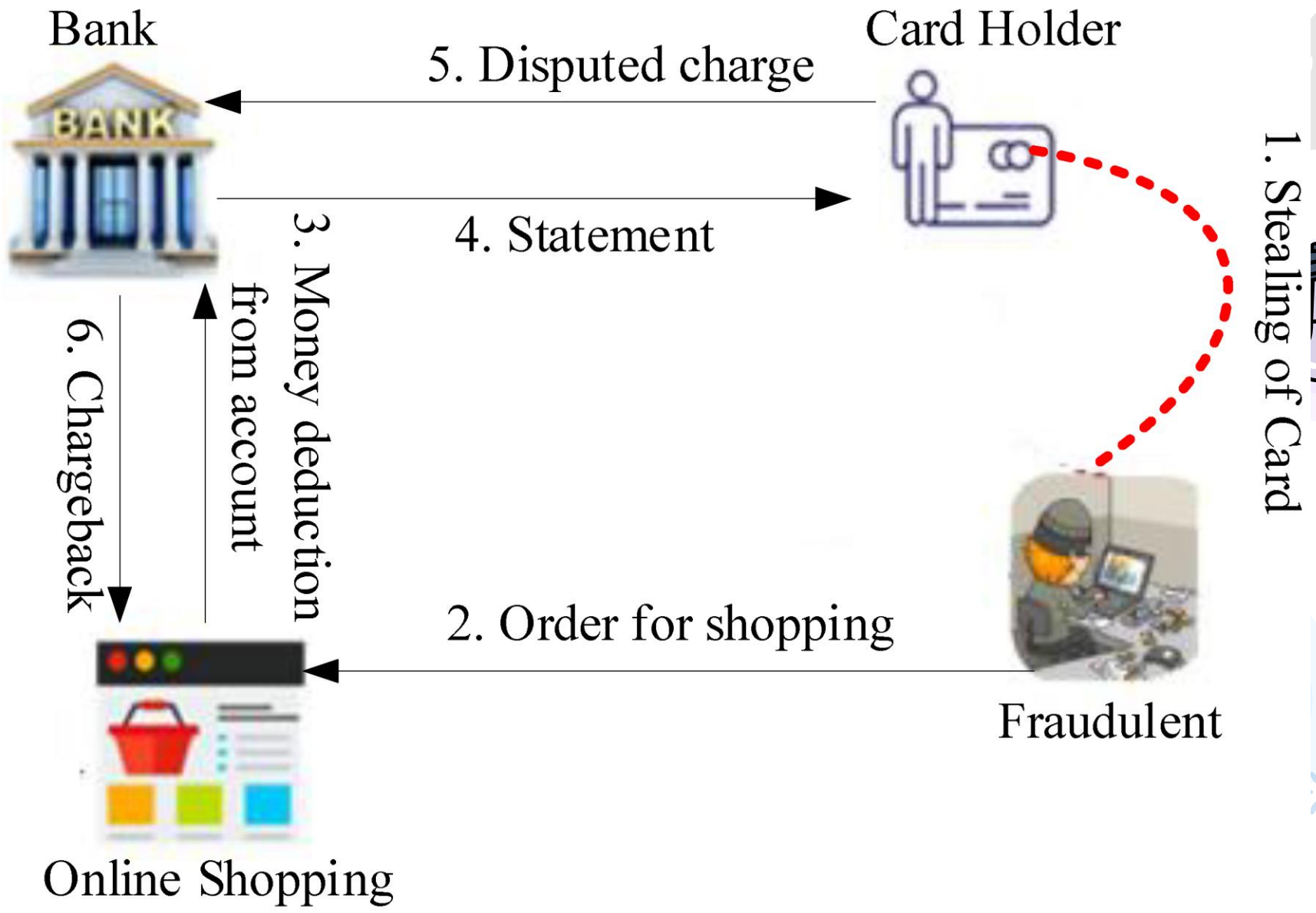
Data Science is used by logistics companies to optimize routes to ensure faster delivery of products and increase operational efficiency.



Data Science Applications

Fraud detection

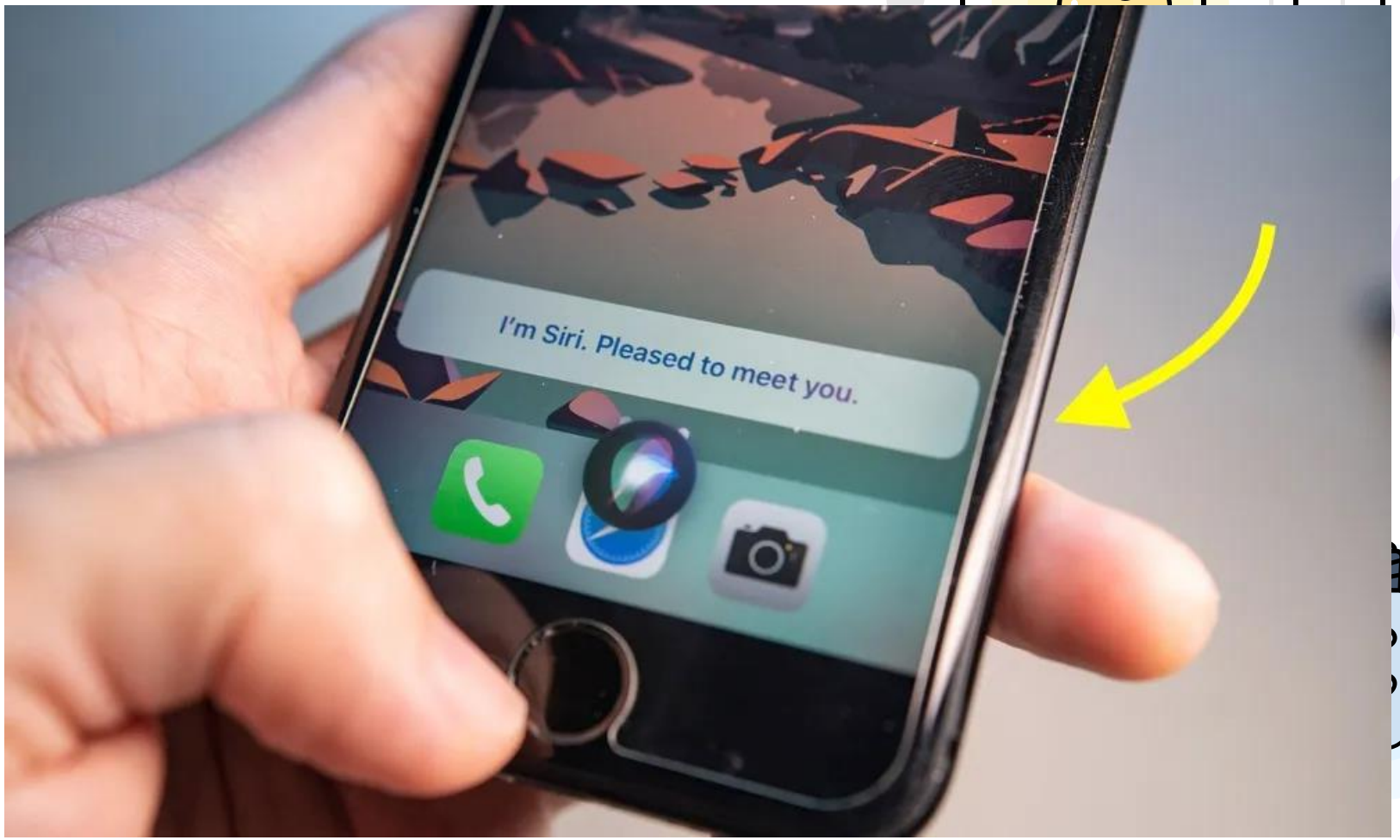
Banking and financial institutions use data science and related algorithms to detect fraudulent transactions.



Data Science Applications

Speech recognition

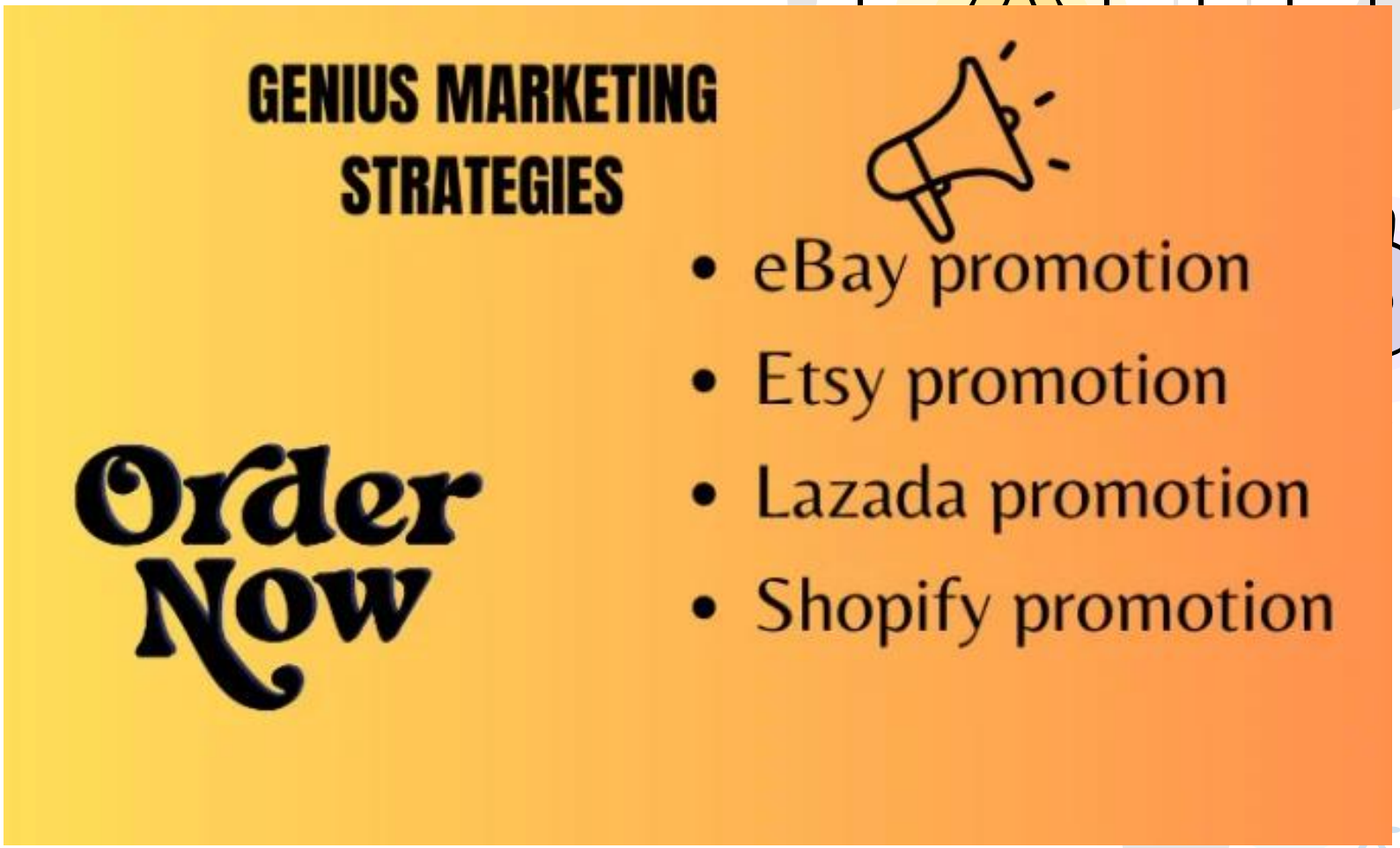
Speech recognition is dominated by data science techniques. We may see the excellent work of these algorithms in our daily lives. Example, help of a virtual speech assistant like Google Assistant or Siri



Data Science Applications

Targeted advertising

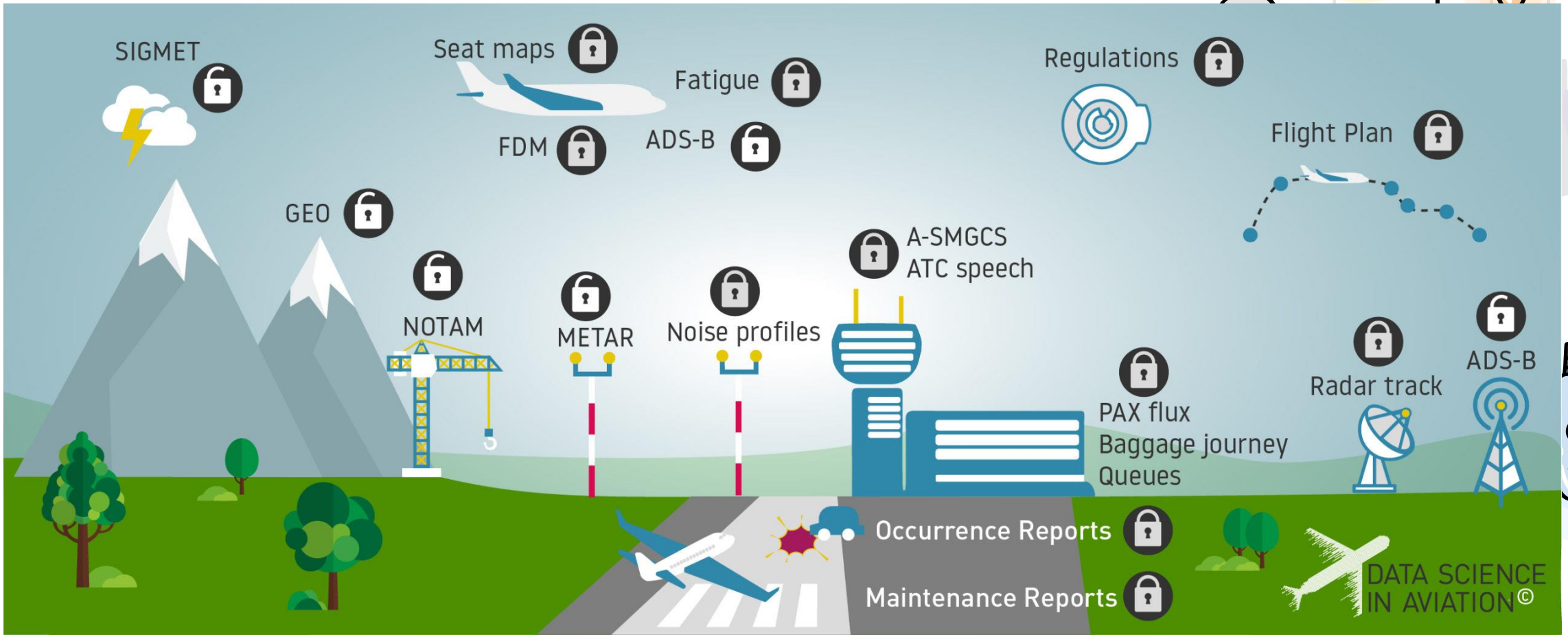
From display banners on various websites to digital billboards at airports, data science algorithms are utilised to identify almost anything. This is why digital advertisements have a far higher CTR (Call-Through Rate) than traditional marketing.



Data Science Applications

Airline Route Planning

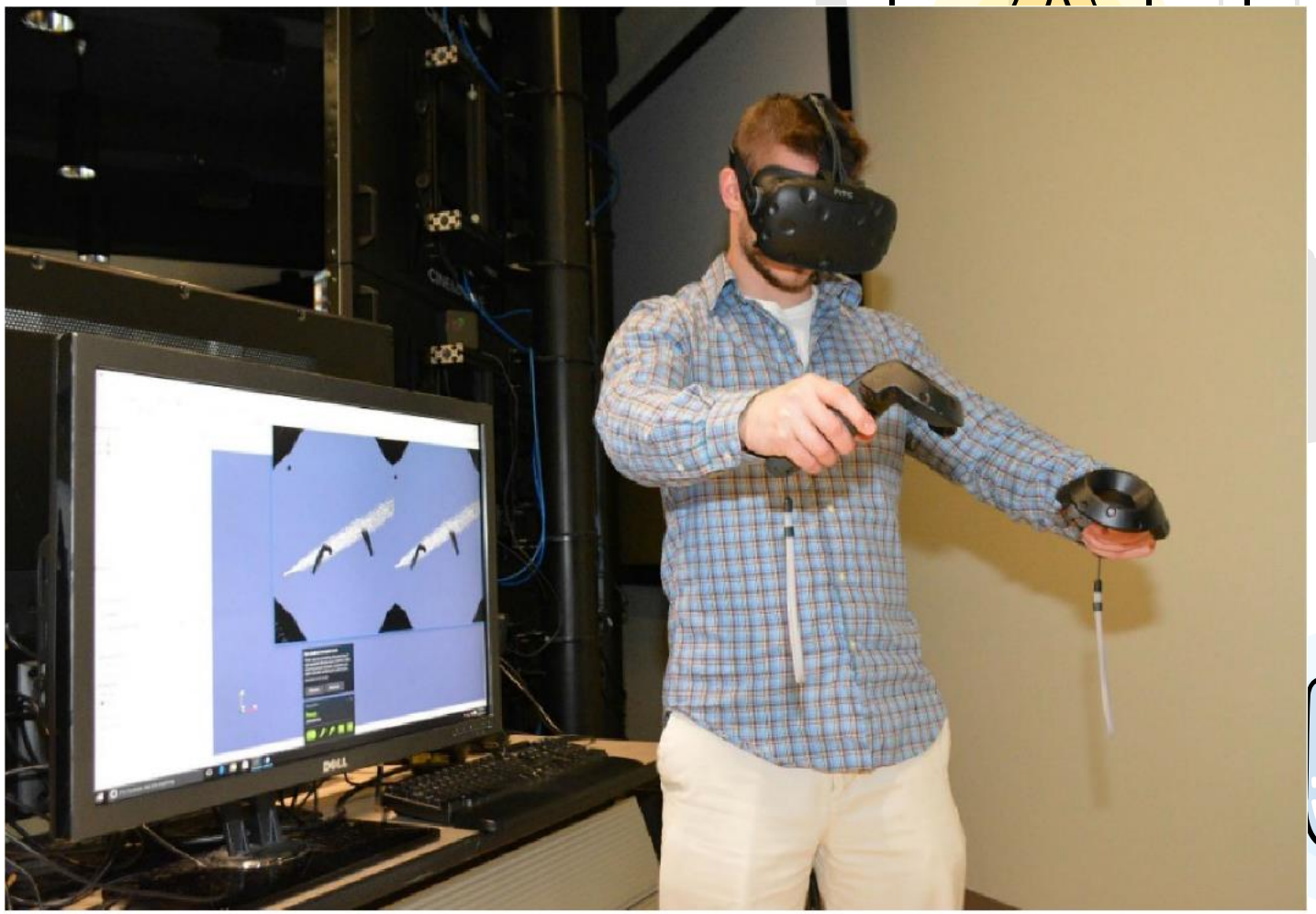
it is easier to predict flight delays for the airline industry, which is helping it grow. It also helps to determine whether to land immediately at the destination or to make a stop in between.



Data Science Applications

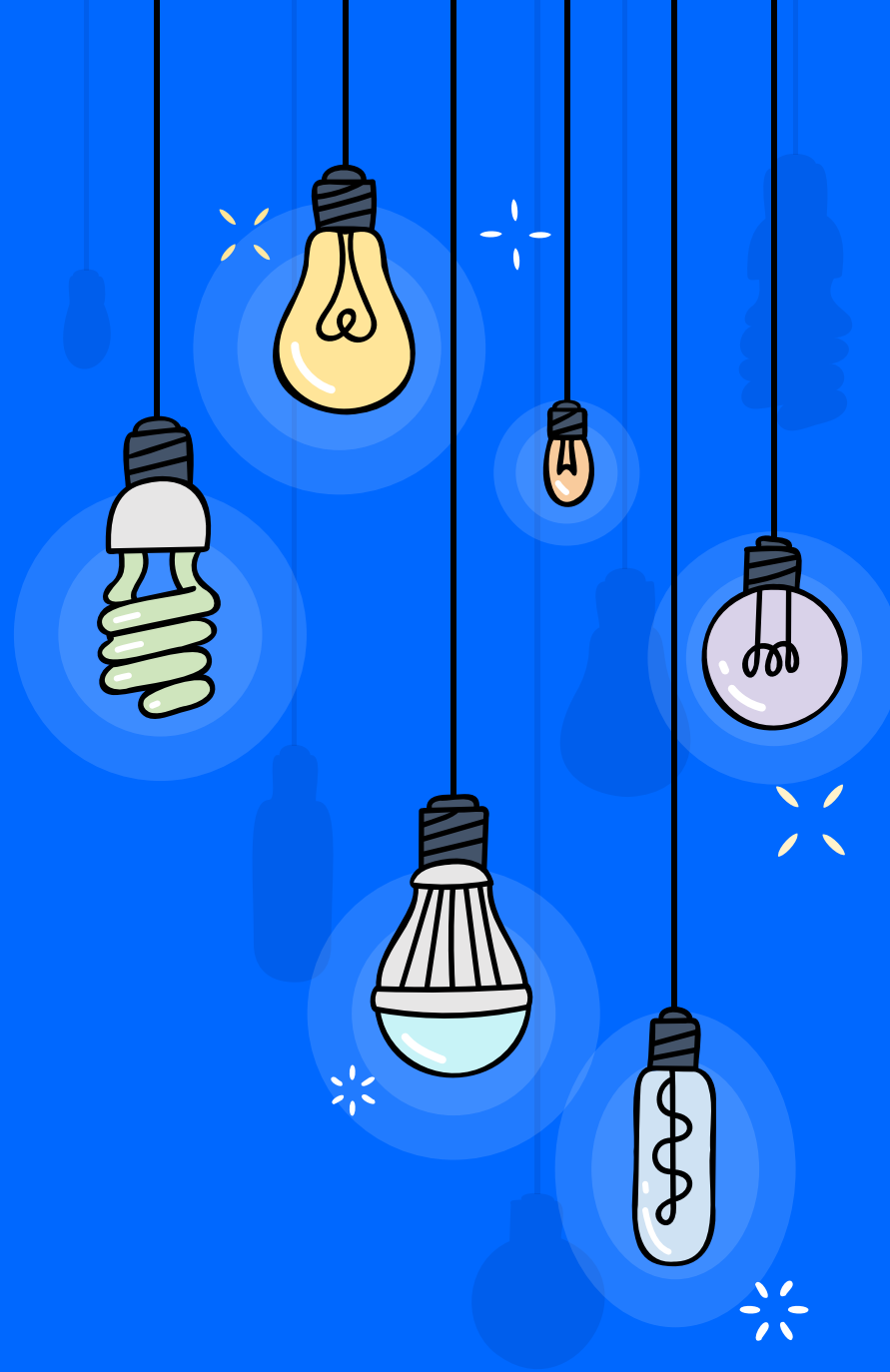
Augmented Reality

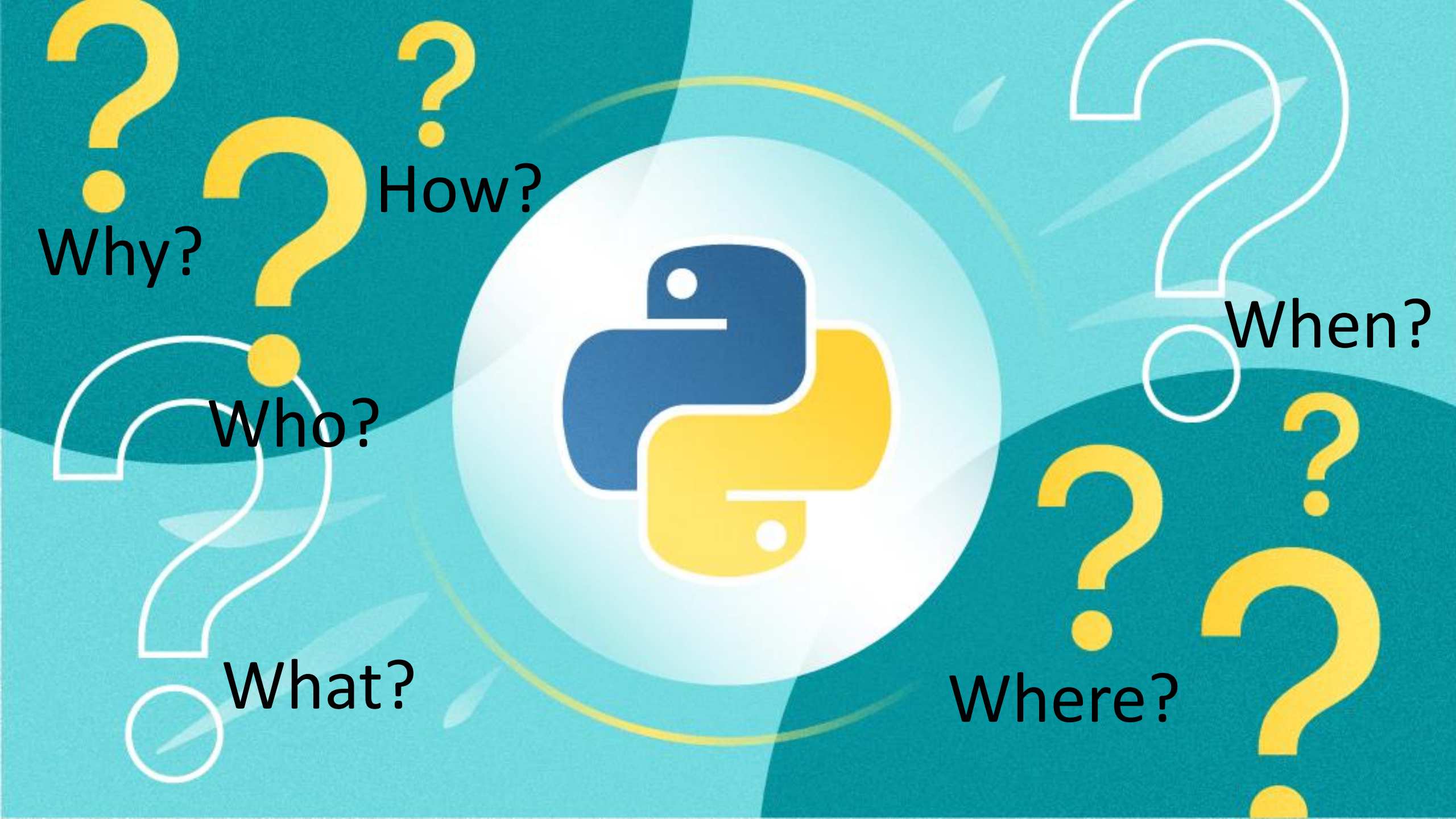
A virtual reality headset incorporates computer expertise, algorithms, and data to create the greatest viewing experience possible. The popular game Pokemon GO is a minor step in that direction.



3

Introduction to Programming Language




How?
Why?
When?
Who?
What?
Where?

INTRODUCTION

WHAT? → A general purpose language, high level, interpreted language
With easy syntax and dynamic semantics

WHO? → Created by Guido Van Rossum

WHEN? → In 1989

WHY? → Easy, free, applications, huge libraries and support, simplicity,
Open source, portability, embeddable, object orientation

WHERE → Google, Dropbox, NETFLIX, BitTorrent, NASA

HOW? → Pydev, Pycharm, Sublime Text, Visual Studio Code, Vim, etc…

TIOBE INDEX JULY 2023

Jul 2023	Jul 2022	Change	Programming Language	Ratings	Change
1	1		Python	13.42%	-0.01%
2	2		C	11.56%	-1.57%
3	4	↑	C++	10.80%	+0.79%
4	3	↓	Java	10.50%	-1.09%
5	5		C#	6.87%	+1.21%
6	7	↑	JavaScript	3.11%	+1.34%
7	6	↓	Visual Basic	2.90%	-2.07%
8	9	↑	SQL	1.48%	-0.16%
9	11	↑	PHP	1.41%	+0.21%
10	20	⇈	MATLAB	1.26%	+0.53%

Introduction to Google Colaboratory

Welcome To Colaboratory

File Edit View Insert Runtime Tools Help

Share

Table of contents

Getting started

Data science

Machine learning

More Resources

Featured examples

Section

+ Code + Text Copy to Drive

Connect

Welcome to Colab!

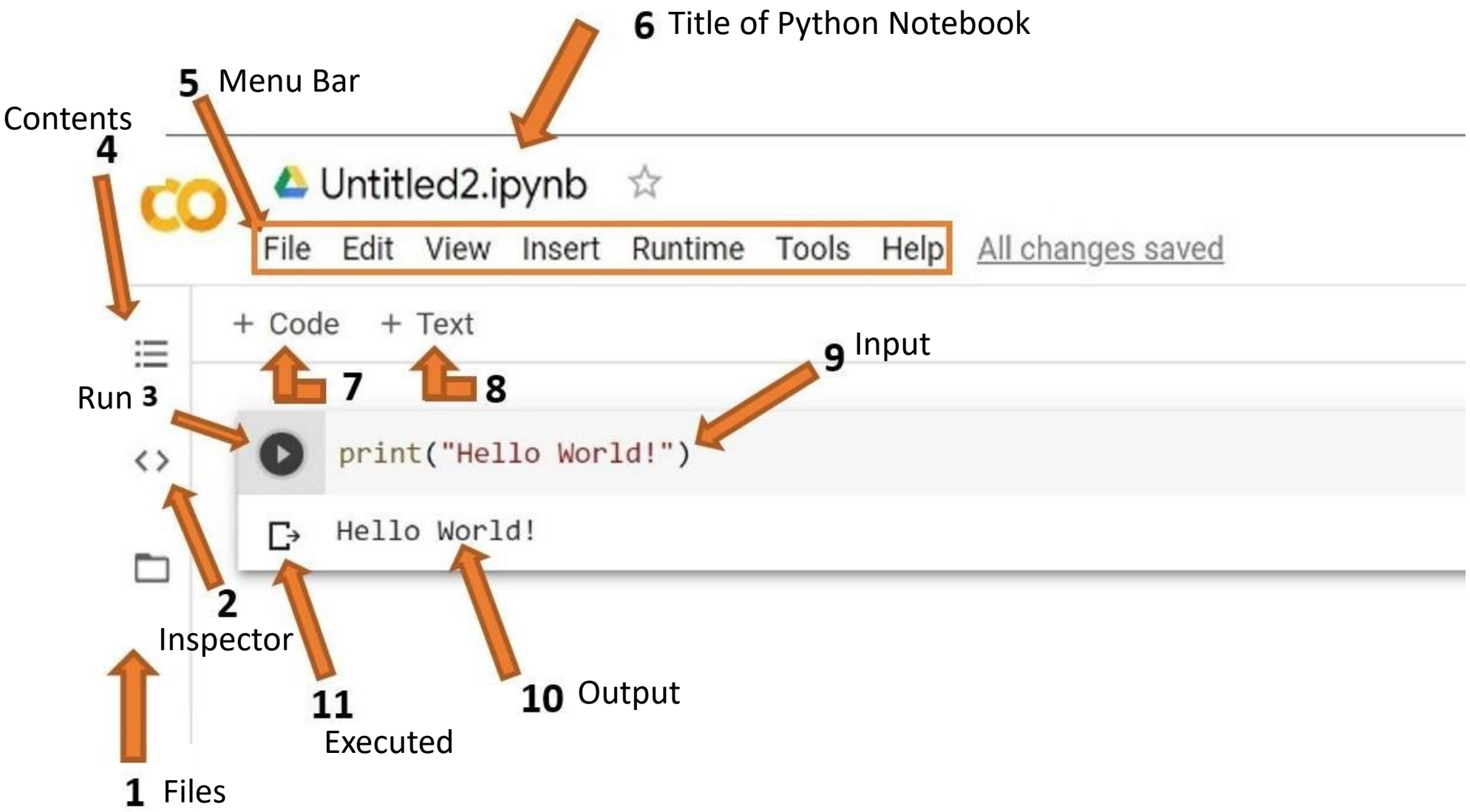
If you're already familiar with Colab, check out this video to learn about interactive tables, the executed code history view, the command palette.

What is Colab?

Colab, or "Colaboratory", allows you to write and execute Python in your browser, with

- Zero configuration required

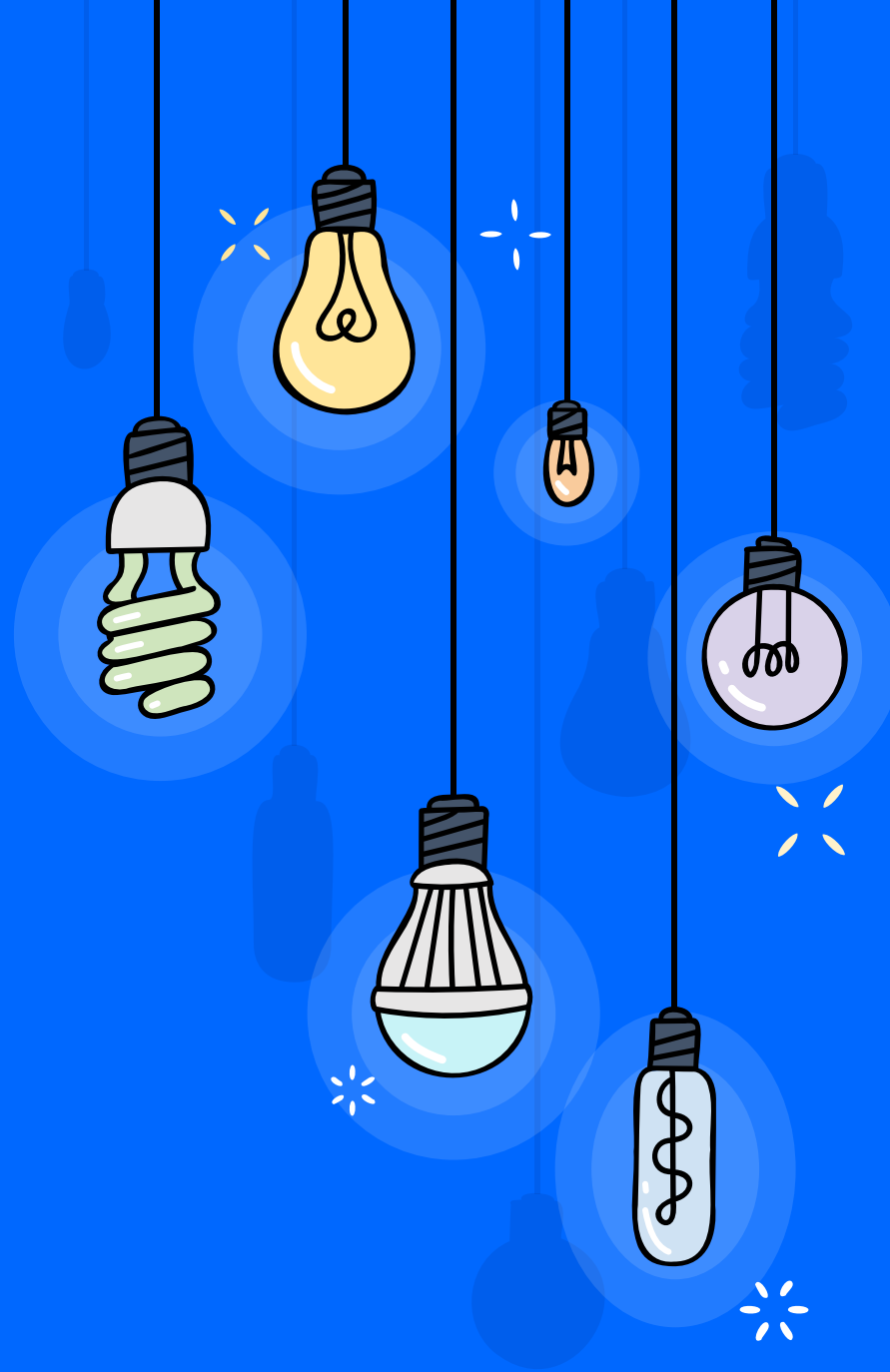
Introduction to Google Colaboratory



Hello!

LET'S LEARN PYTHON

4 LifeCycle of Data Science Field

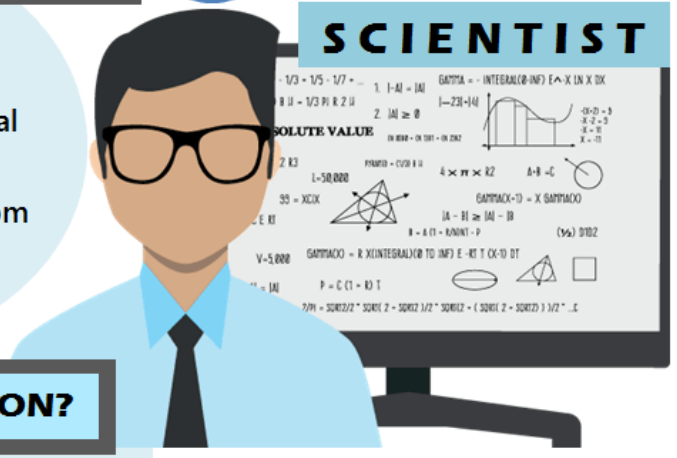


Who is a Data Scientist?

WHO AM I?

I am a part analyst & part artist. I use my analytical and technical abilities to extract meaning / insights from massive data sets.

D A T A S C I E N T I S T



WHAT DO I DO?

1. I cleanse existing raw data & build models to predict future data.
2. I go beyond merely collecting and reporting data, to look at data from multiple angles & give meaning to it.
3. I identify the correct business problem(s) & offer solutions (via visualizations, reports or blogs) by best applying the data.

WHAT DO I RELY ON?

1. Analytics
2. Predictive Models
3. Statistical Analysis & Modeling
4. Data Mining
5. Sentiment Analysis
6. What-if Analysis

THE PROCESS I FOLLOW

- Define Problem
- Structure Data
- Use Programming Language

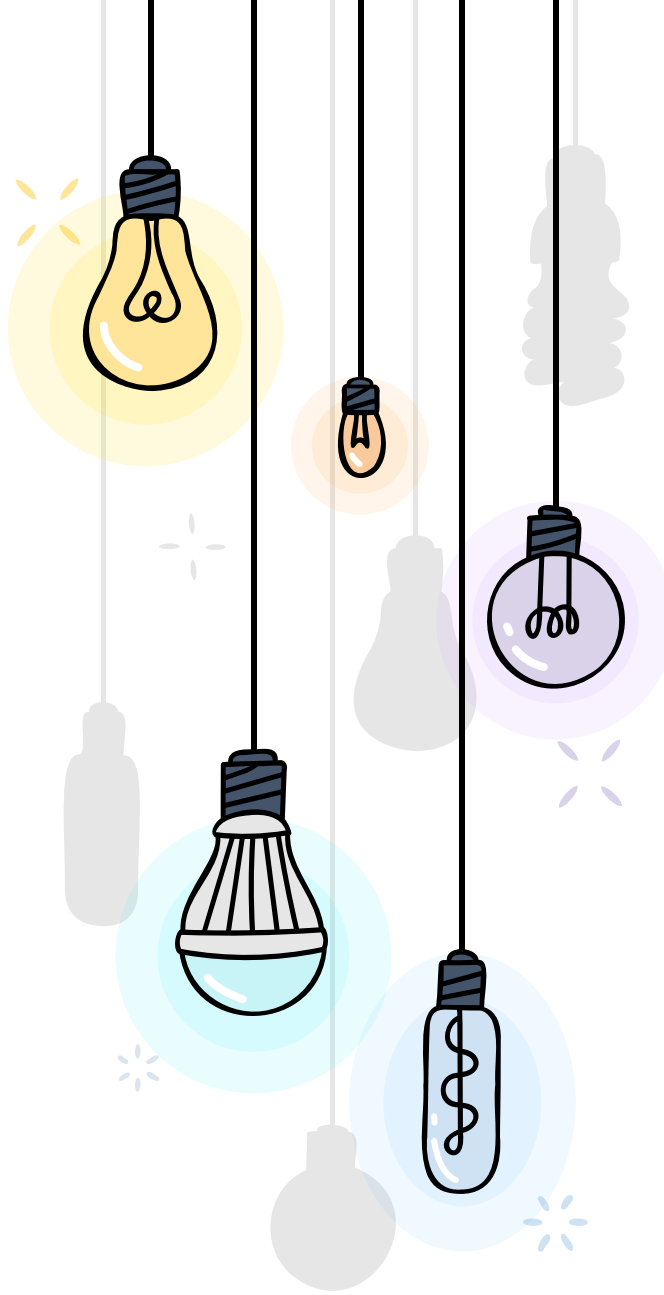
WHAT DO I EARN?

After oil & gas geologists, mine is the 2nd highest paid job in the world!

100,000 to 150,000

HOW DO I HELP ORGANIZATIONS TODAY?

- Increase data accuracy
- Develop strategies
- Improve operational efficiency
- Reduce costs
- Mitigate risks
- Offer personalized products/services



Why Become a Data Scientist?

According to Glassdoor and Forbes, **demand for data scientists will increase by 28 percent by 2026**, which speaks of the profession's durability and longevity, so if you want a secure career, data science offers you that chance.

Comparison

DATA SCIENTIST

Job role: Determine what the problem is, what questions need answers, and where to find the data. Also, they mine, clean, and present the relevant data.

Skills needed: Programming skills (SAS, R, Python), storytelling and data visualization, statistical and mathematical skills, knowledge of Hadoop, SQL, and Machine Learning.

DATA ANALYST

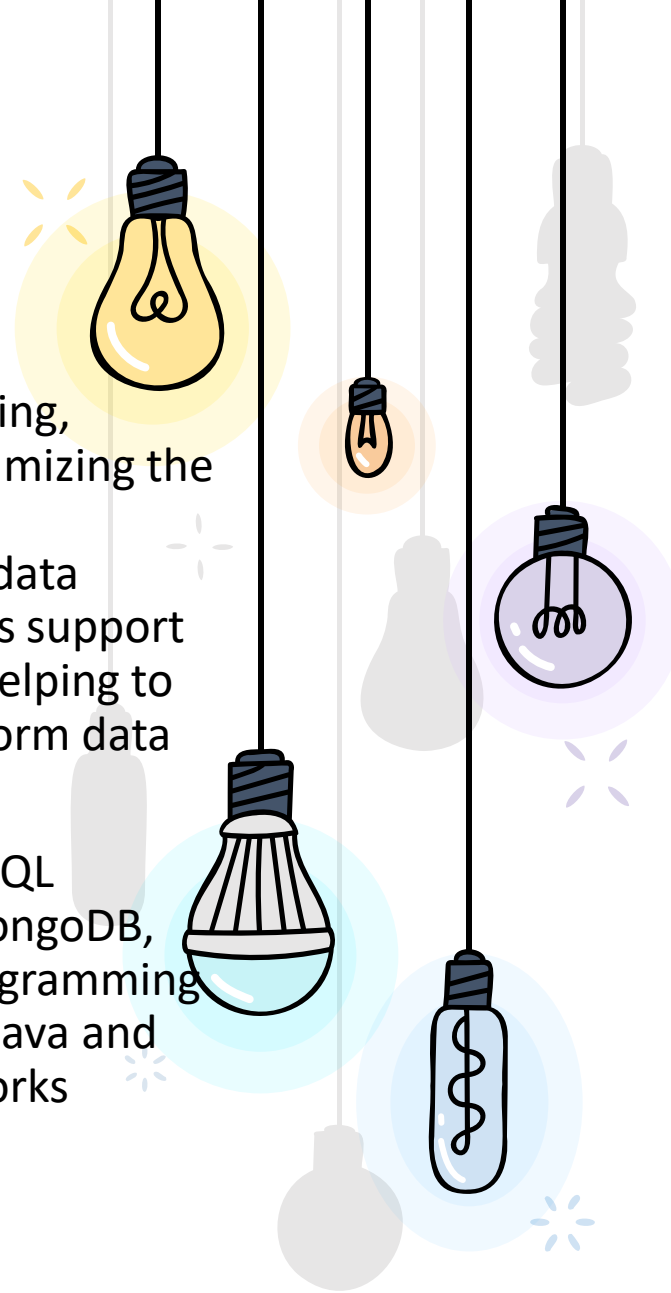
Job role: Analysts bridge the gap between the data scientists and the business analysts, organizing and analyzing data to answer the questions the organization poses. They take the technical analyses and turn them into qualitative action items.

Skills needed: Statistical and mathematical skills, programming skills (SAS, R, Python), experience in data wrangling and data visualization.

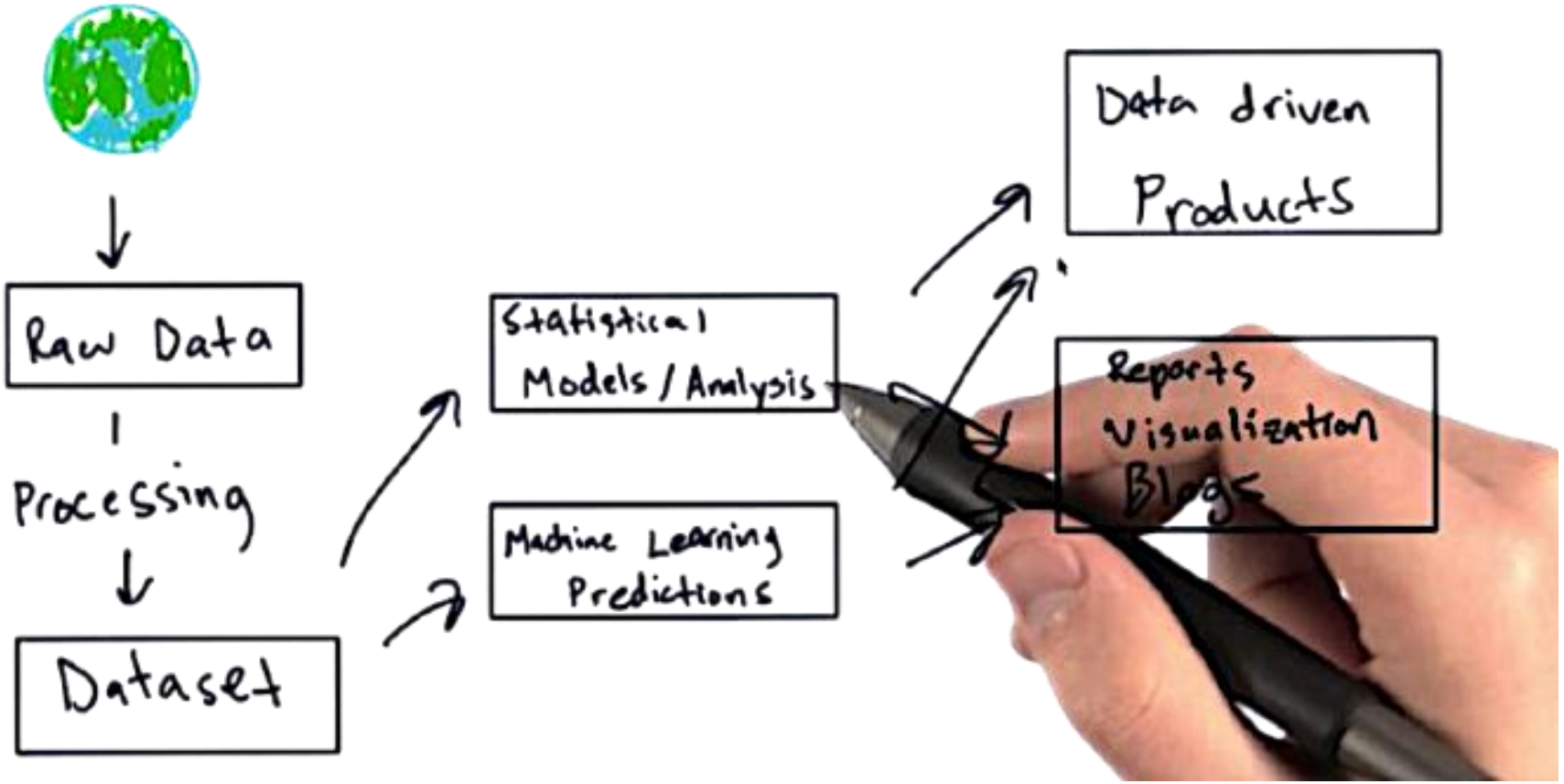
DATA ENGINEER

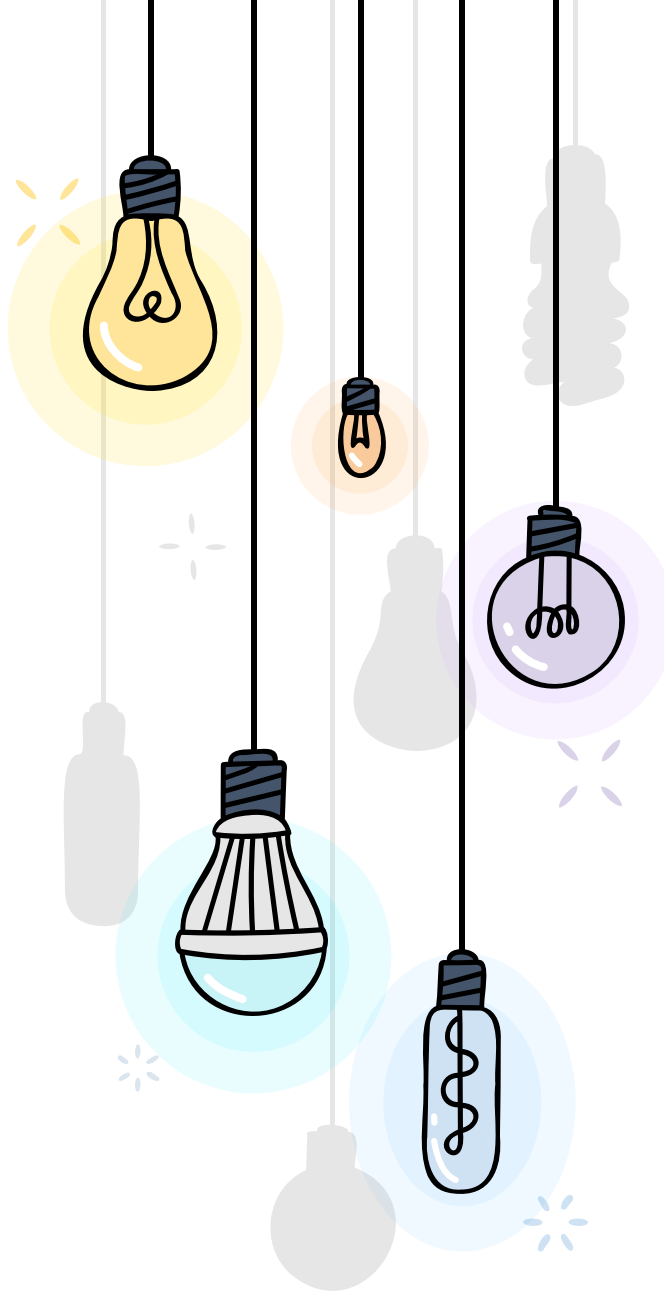
Job role: Focus on developing, deploying, managing, and optimizing the organization's data infrastructure and data pipelines. Engineers support data scientists by helping to transfer and transform data for queries.

Skills needed: NoSQL databases (e.g., MongoDB, Cassandra DB), programming languages such as Java and Scala, and frameworks (Apache Hadoop).

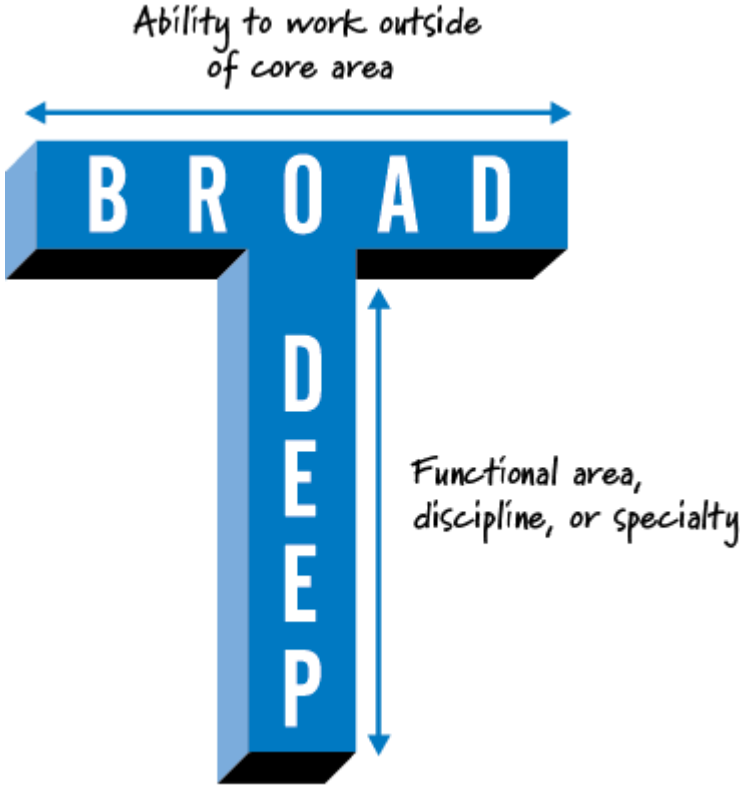


What does a data scientist do?



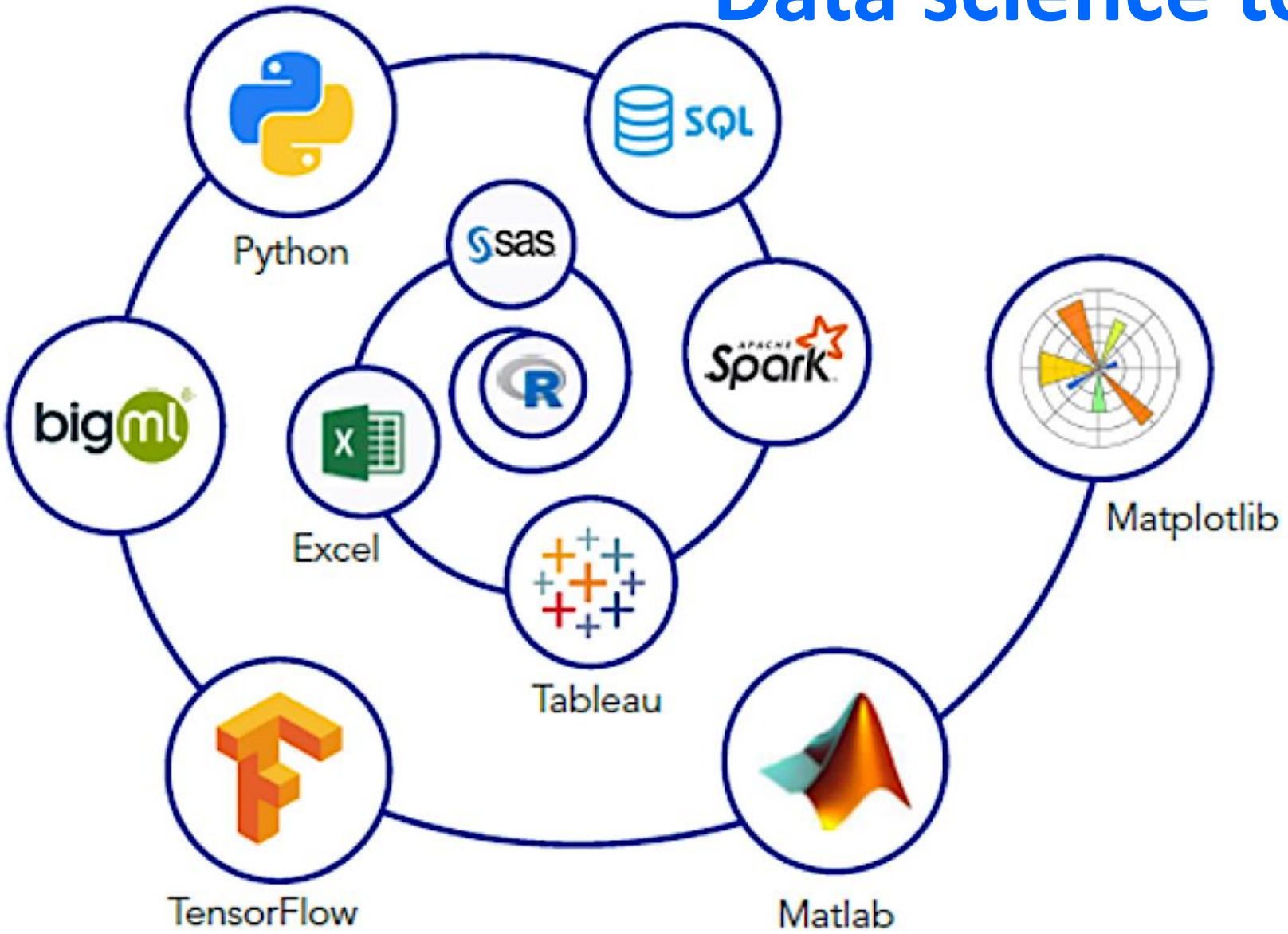


T-Shaped Skillset



https://youtu.be/sJUP4utv3aU

Data science tools

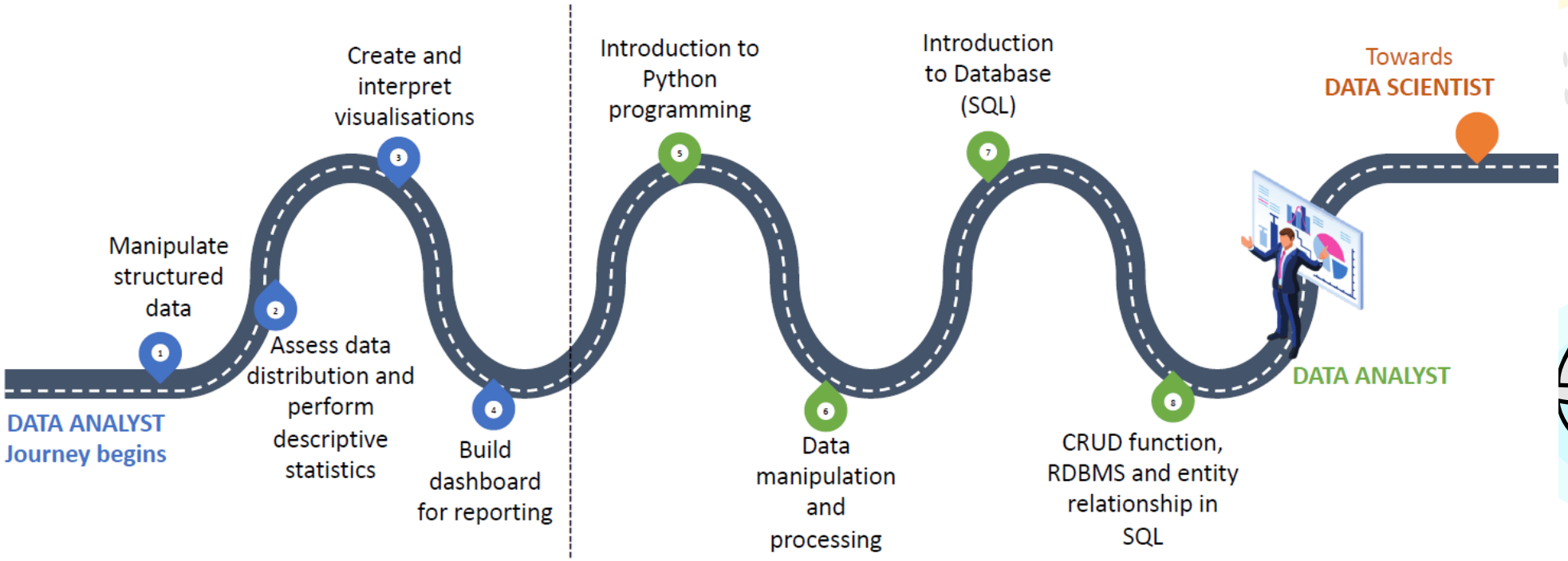


MEDIUM

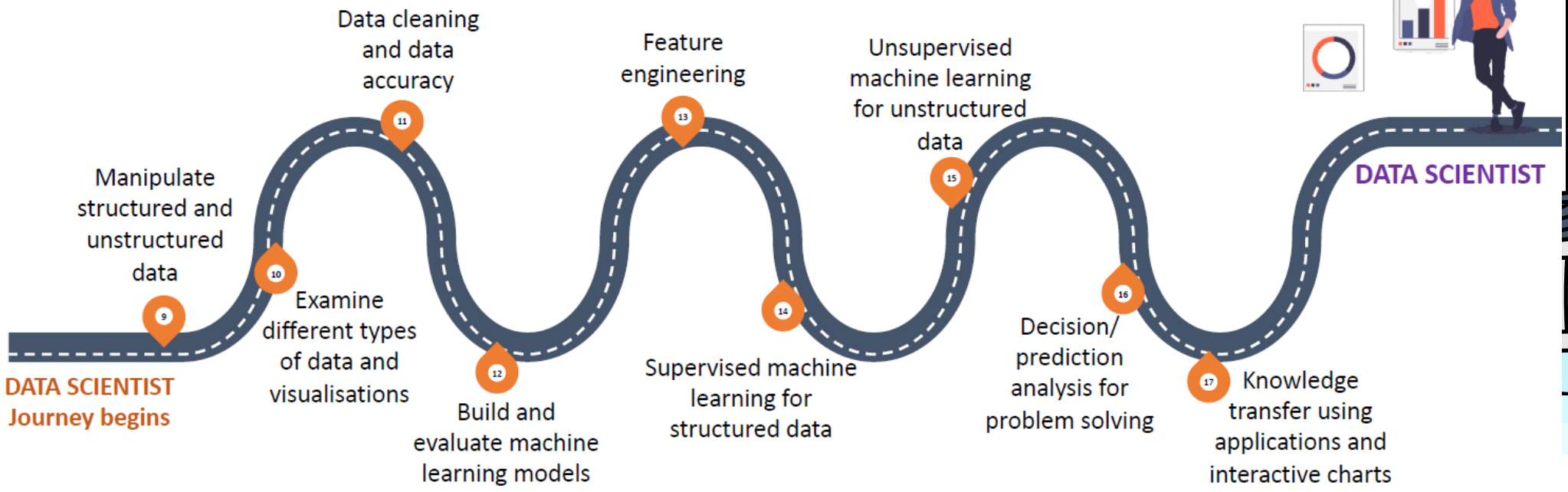
DATA FLAIR

Source: https://data-flair.training/blogs/data-science-toohttps://medium.com/javarevisited/10-essential-tools-data-scientists-should-learn-in-2022-acbae6558643ls/

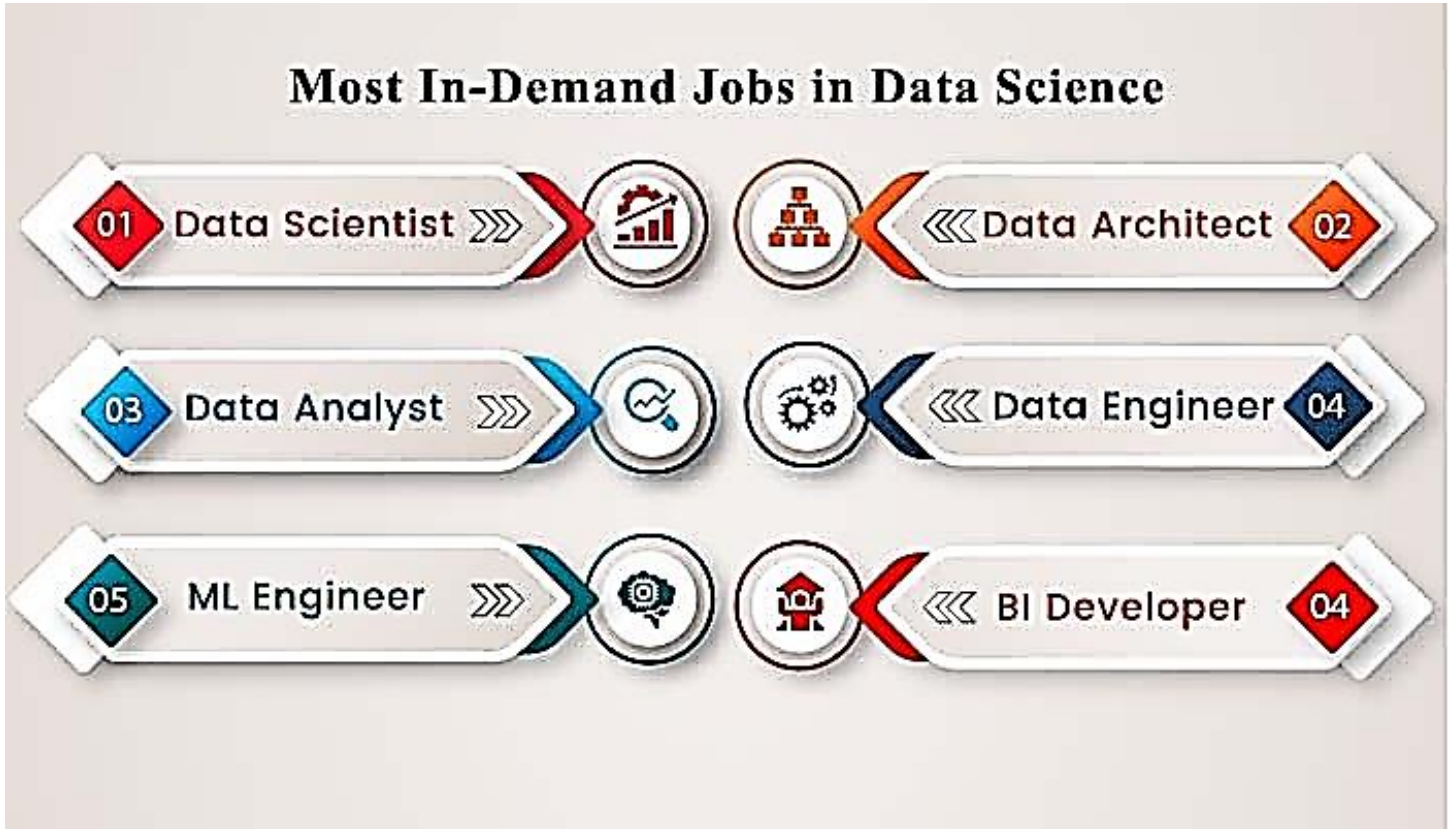
Roadmap & Learning Framework to become a Data Scientist



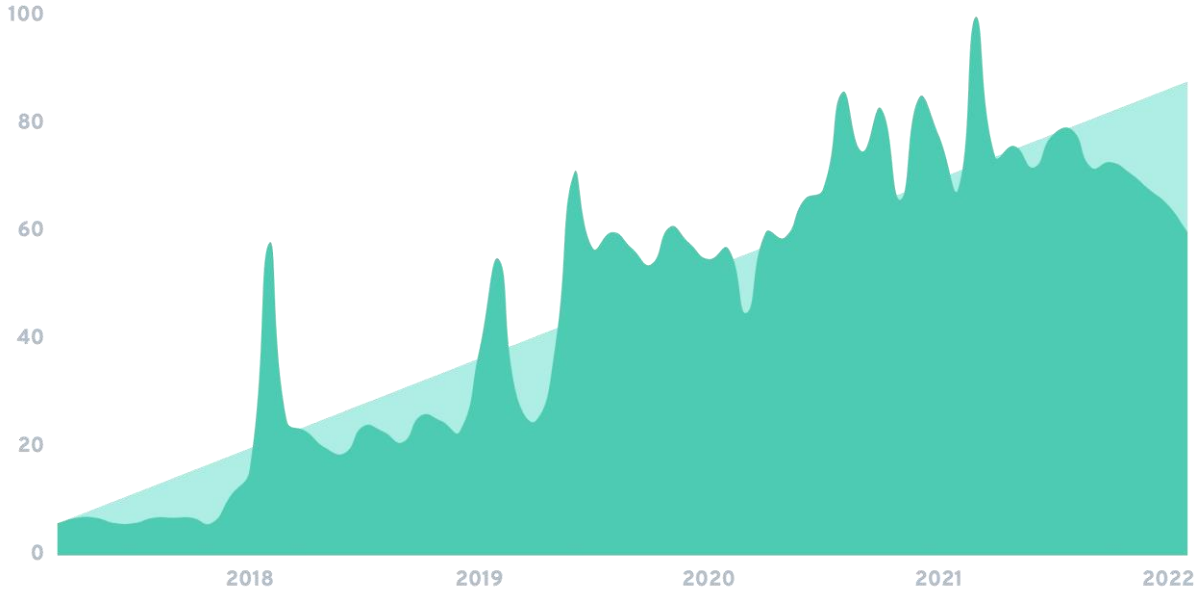
Roadmap & Learning Framework to become a Data Scientist

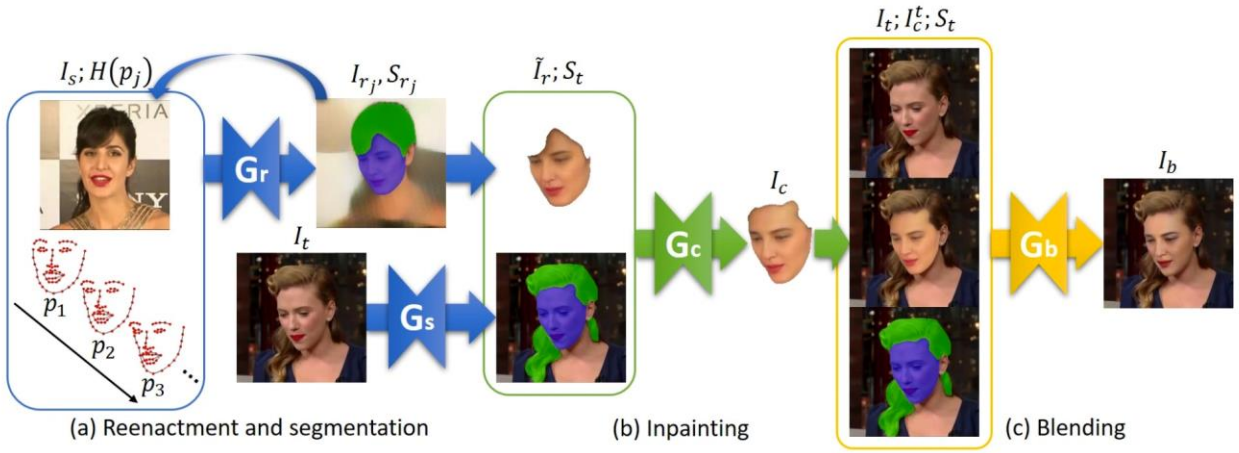


Demand and Opportunities in Data Science



1. Explosion In Deepfake Video And Audio



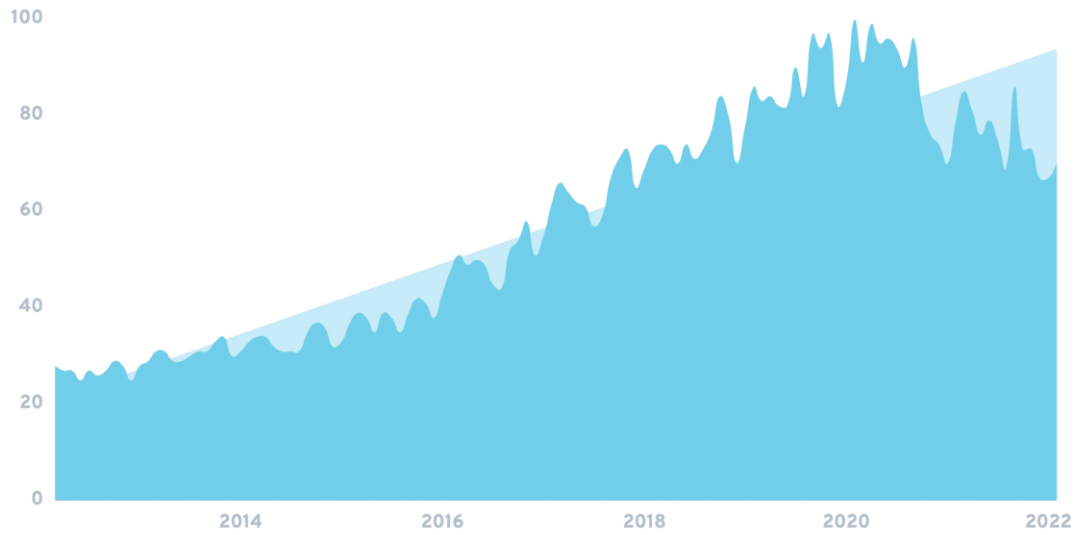


Open source software makes deepfake technology relatively accessible.

There's huge scope for this technology to be used maliciously.

Back in 2019, an AI company **deepfaked popular podcaster Joe Rogan's voice** so effectively it instantly went viral on social media.

2. More Applications Created With Python



“Python” searches have grown by 150% in the last 10 years. Python is on track to become the most popular programming language by 2025.

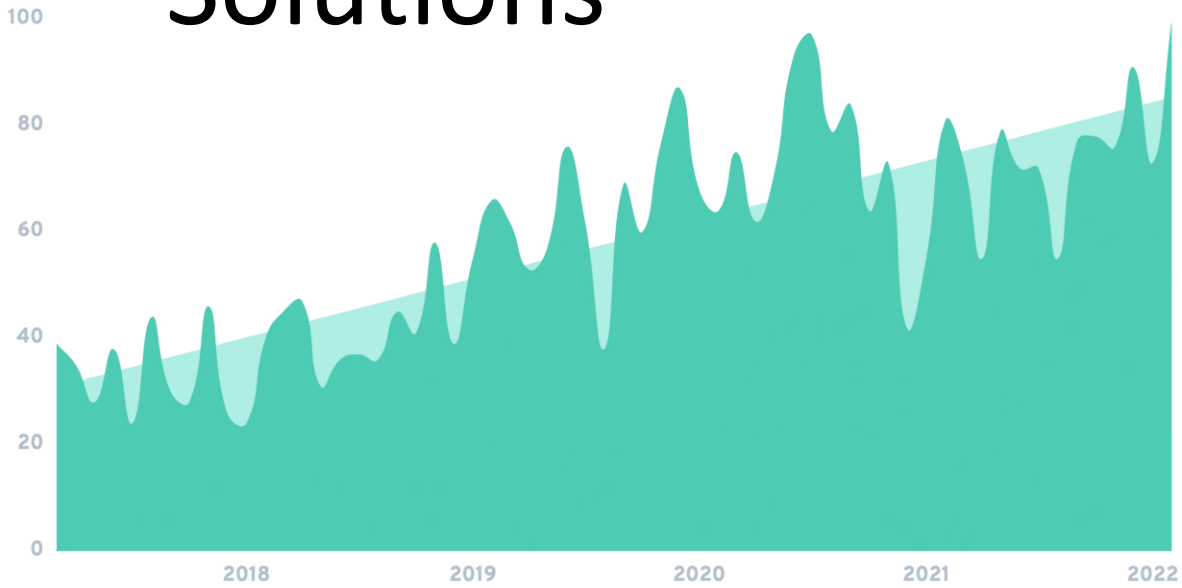
TIOBE
the software quality company

About us Join TI

Products Qu

Aug 2022	Aug 2021	Change	Programming Language
1	2	↑	Python
2	1	↓	C
3	3		Java
4	4		C++
5	5		C#
6	6		Visual Basic
7	7		JavaScript
8	9	↑	Assembly language
9	10	↑	SQL
10	8	↓	PHP

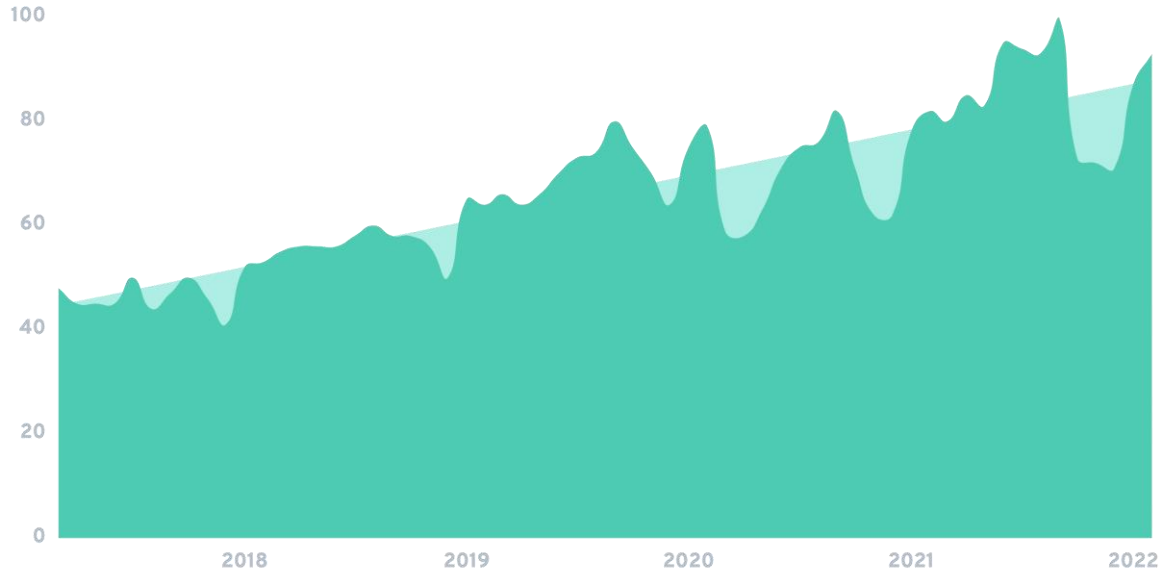
3. Increased Demand For End-To-End AI Solutions



Quicky growing end-to-end AI solutions

- The AI end-to-end solution helps enterprise customers to clean their large data sets and build machine learning models.
- This way, companies can gain valuable, deep learning insights from their massive amounts of data.
- And automate important data management tasks.
- Previously, businesses would have to seek expertise in all the different parts of the process and piece it together themselves

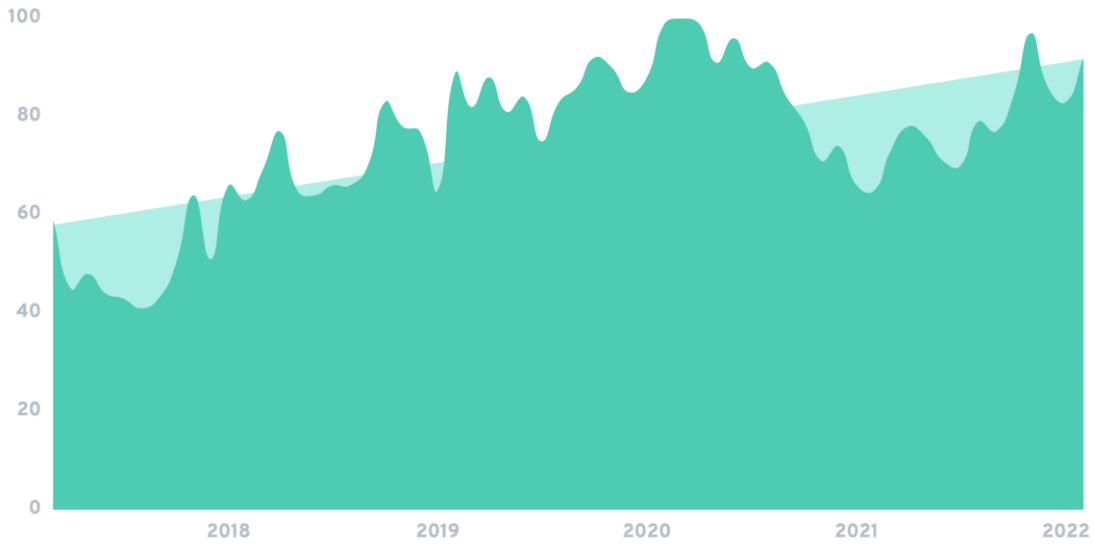
4. Companies Hire More Data Analysts



“Data analyst” searches are up by 93% since 2017. Interest in this data science role displays hockey stick growth.

- Big data is often extremely messy and lacking in proper structure.
- Which is why humans are needed to manually tidy training data before it is ingested by machine learning algorithms.
- AI-produced results are not always reliable or accurate, so machine learning companies often use humans to clean up the final data.
- And write up an analysis of what they find in a way that non-tech stakeholders can understand it.

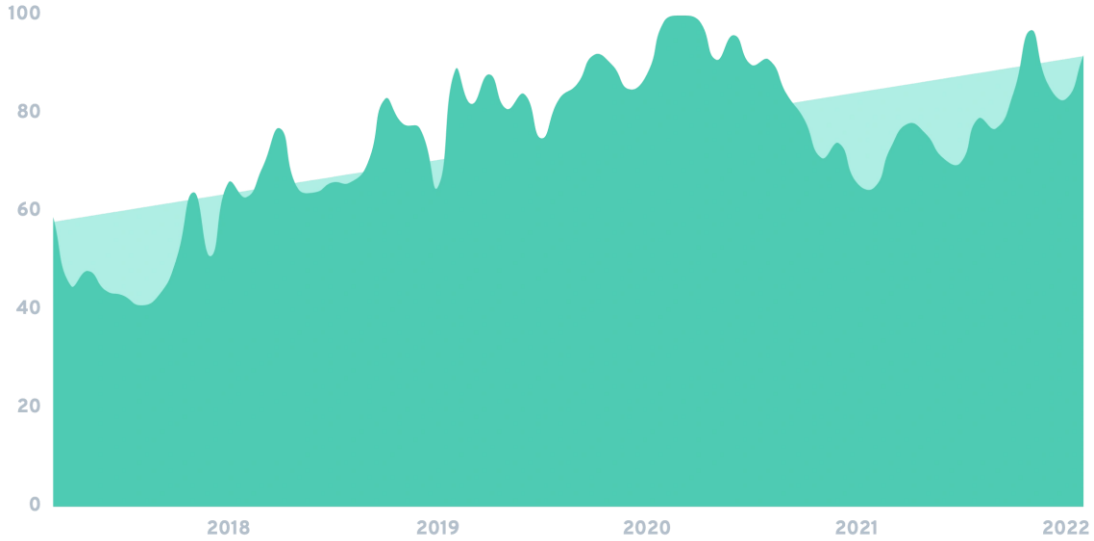
5. Data Scientists Joining Kaggle



Search growth for “Kaggle” has increased by 55% over five years. The data science platform has over 5 million users across 194 countries.

- Kaggle has grown quickly to become the world's largest data science community.
- And with over 8 million users across 194 countries, it’s not slowing down.
- Many budding data scientists now start with Kaggle to begin their machine learning journey.
- Users can even share data sets and enter competitions to solve data science challenges with neural networks.
- Or work with other data scientists to build models in Kaggle’s web-based data science workbench.

6. Increased Interest In Consumer Data Protection



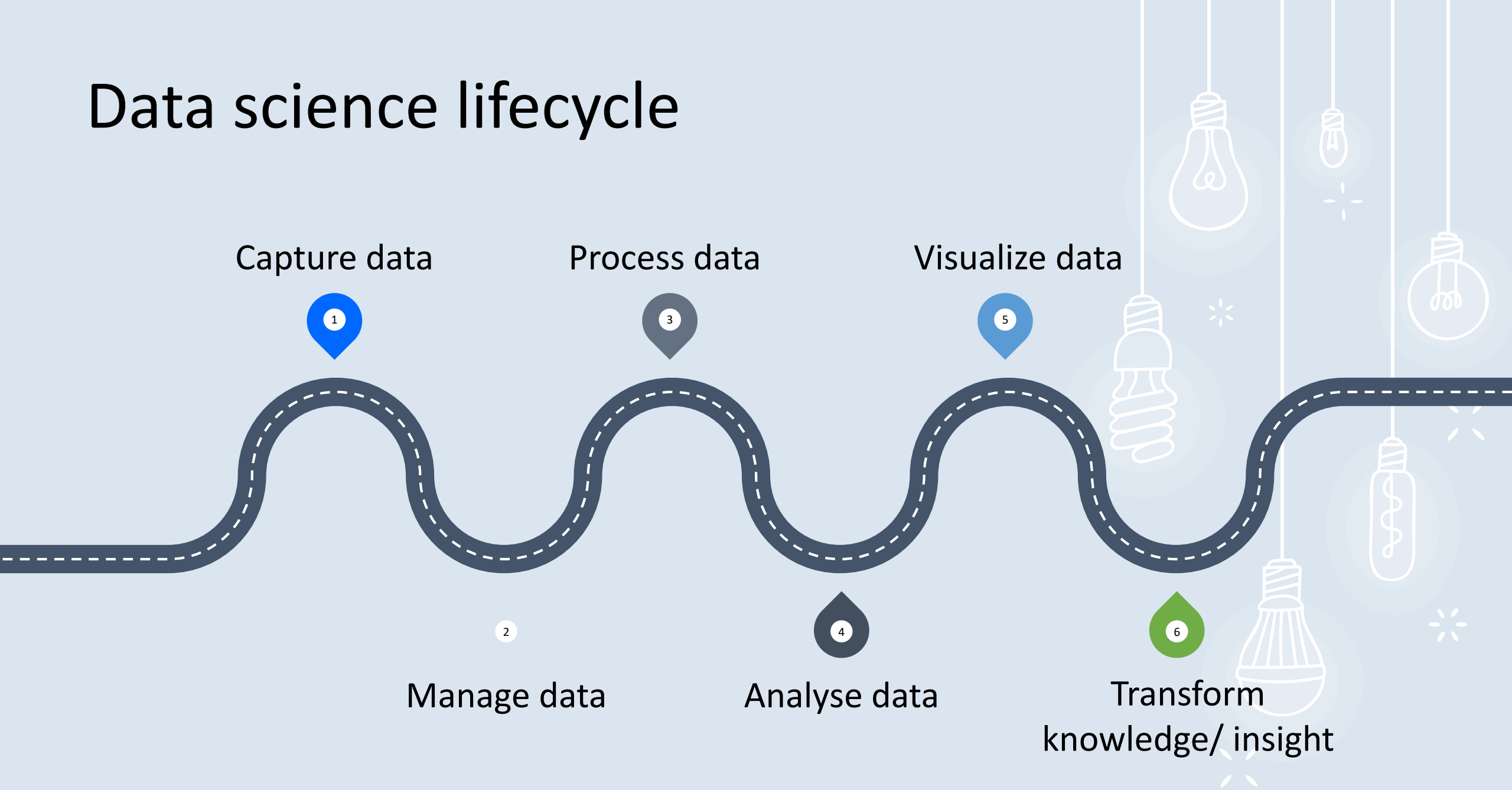
“Data privacy” has seen a search growth of 125% over the last 10 years. People are now searching about their data privacy in greater numbers by the month.

- This broader data privacy trend means that large data sets will soon be walled off and harder to come by.
- And this could become a bane for data science when it comes to the future acquisition and use of consumer data.

5

LifeCycle of Data Science Field (part 2)

Data science lifecycle



Data Science Principles

Statistics

Statistics are at the **core of data science**. A sturdy handle on statistics can help you extract more intelligence and obtain more meaningful results.

Programming

The most common programming languages are **Python, and R**. They support multiple libraries for data science and ML.

Databases

A capable data scientist needs to **understand how databases work**, how to manage them, and how to extract data from them.

Machine learning

Machine learning is the backbone of data science. Data Scientists need to have a **solid grasp of ML** in addition to basic knowledge of statistics.

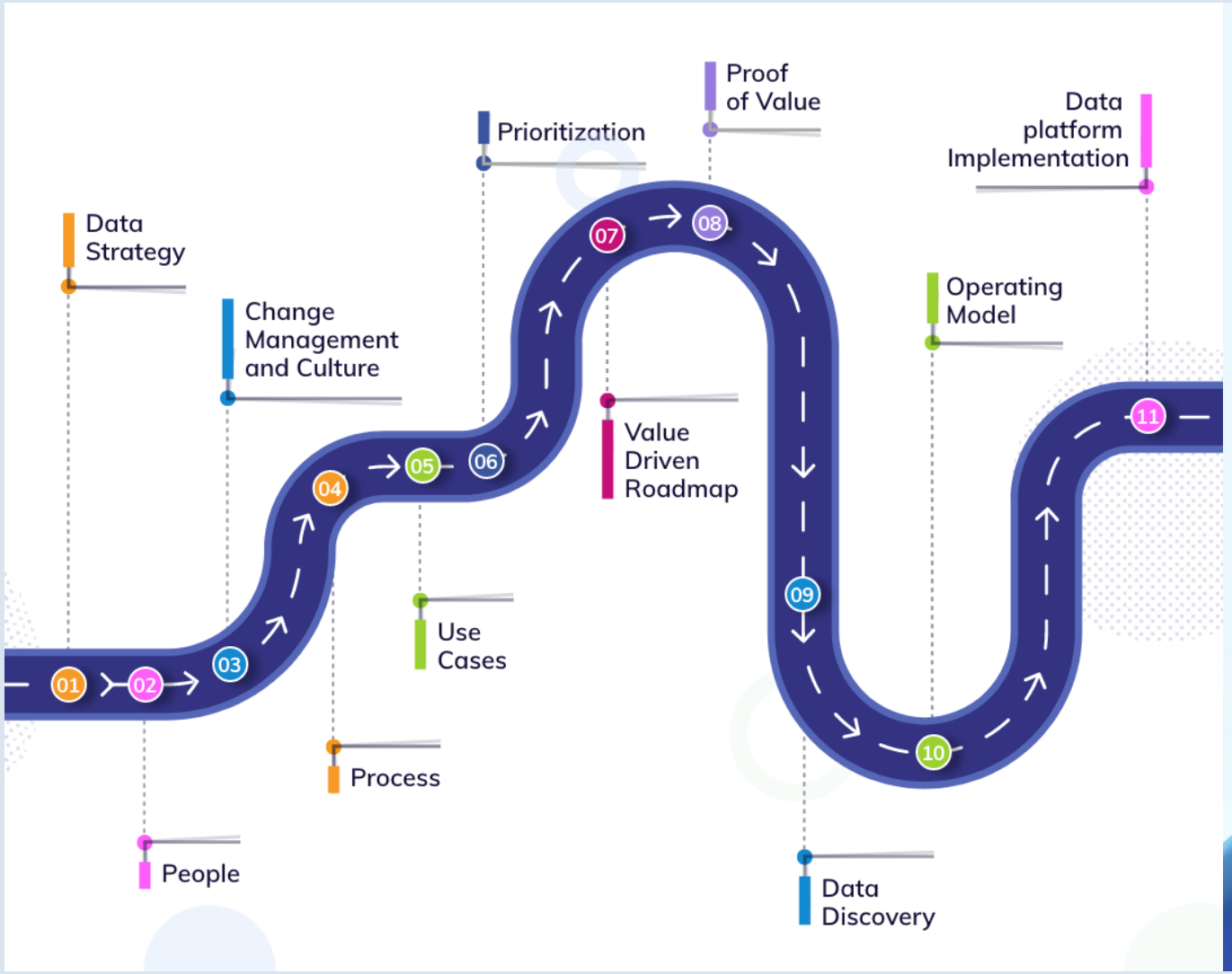
Data story telling

The ability to **effectively communicate insights from a dataset** using narratives and visualizations. It can be used to put data insights into context for and inspire action from your audience.

Communication skill

Being able to communicate effectively is perhaps the most important of all life skills. It is what enables us to **pass information to other people**, and to understand what is said to us.

Strategies to becoming a Data-Driven Organization

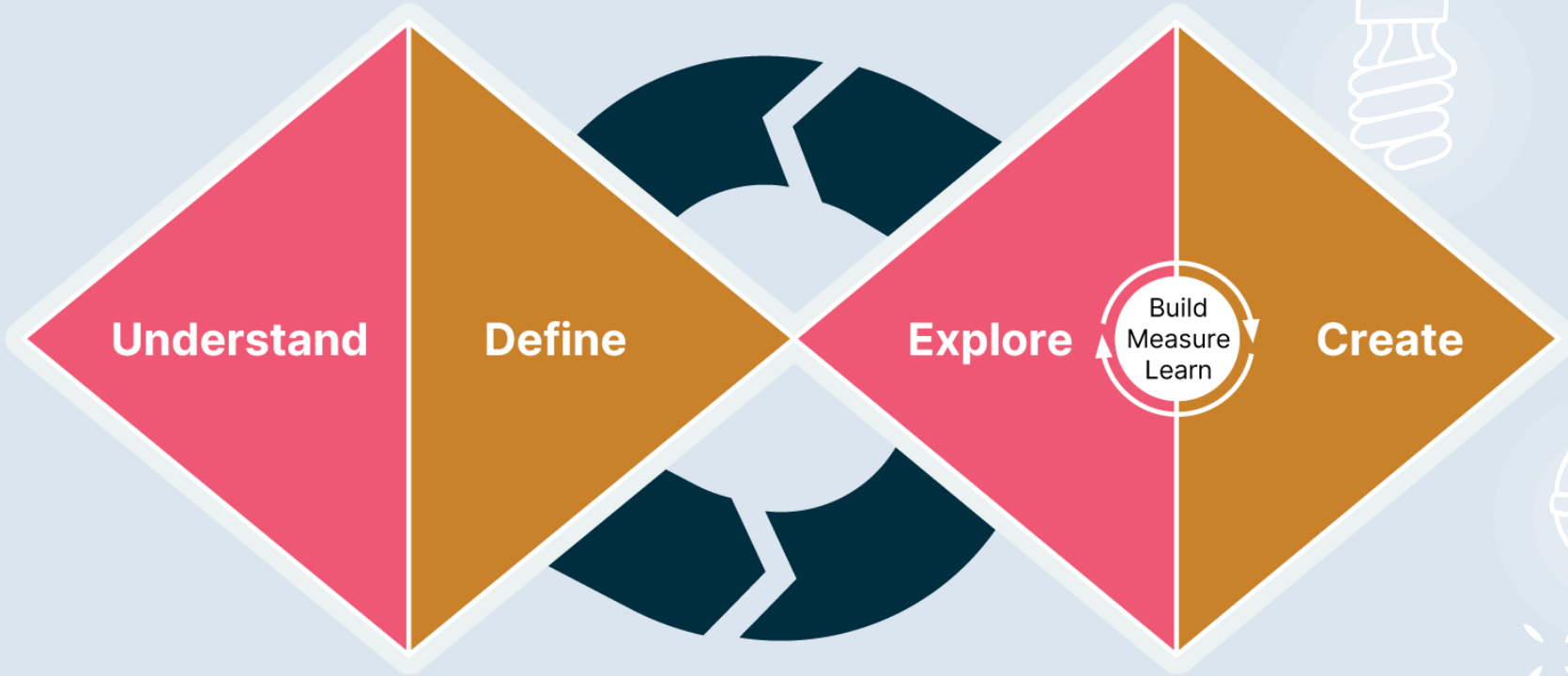


Developing Data Products

Strategy
Understand whys and hows

Solution
Execute the outcome

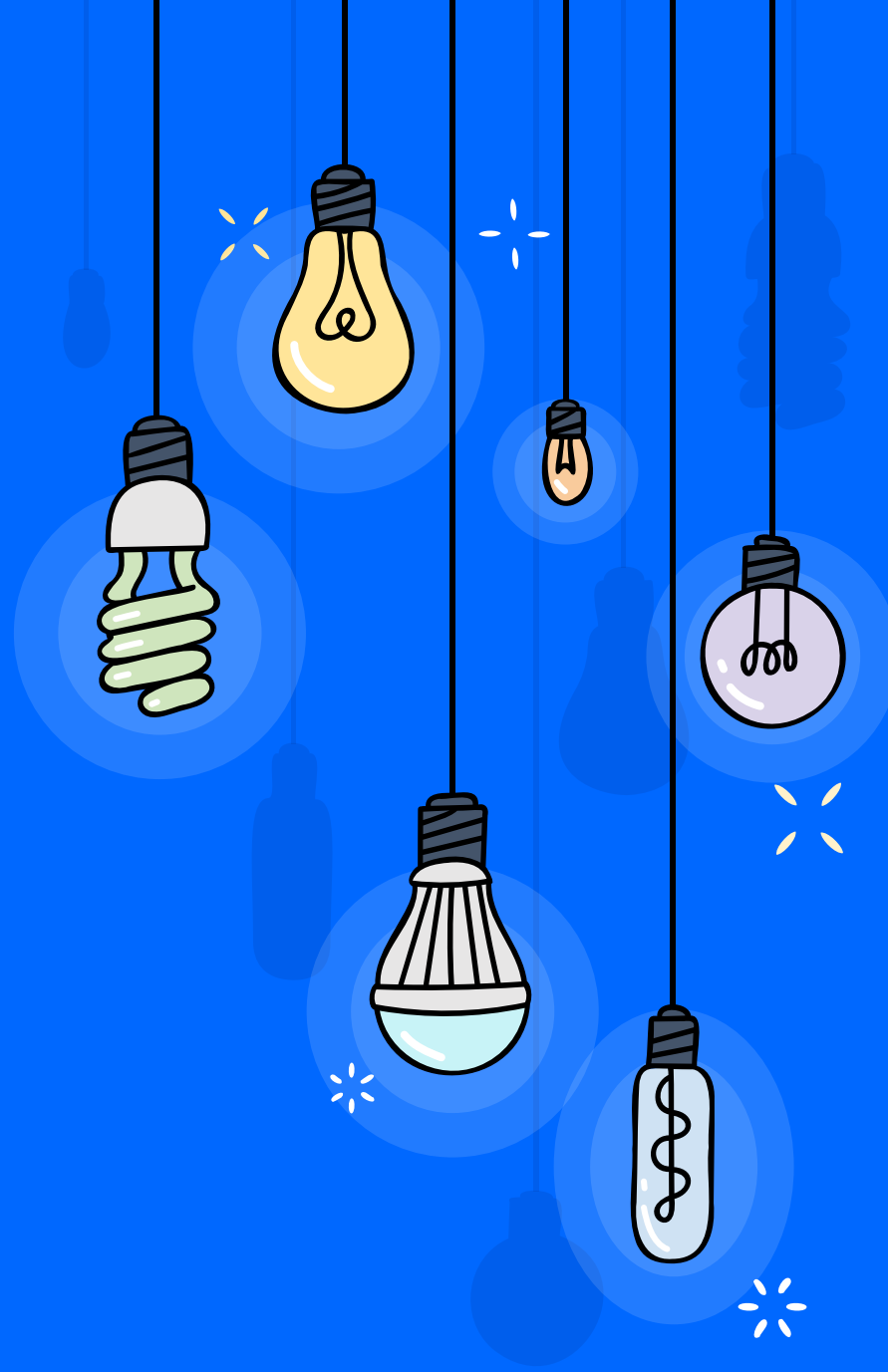
build the **right thing** with lean product development



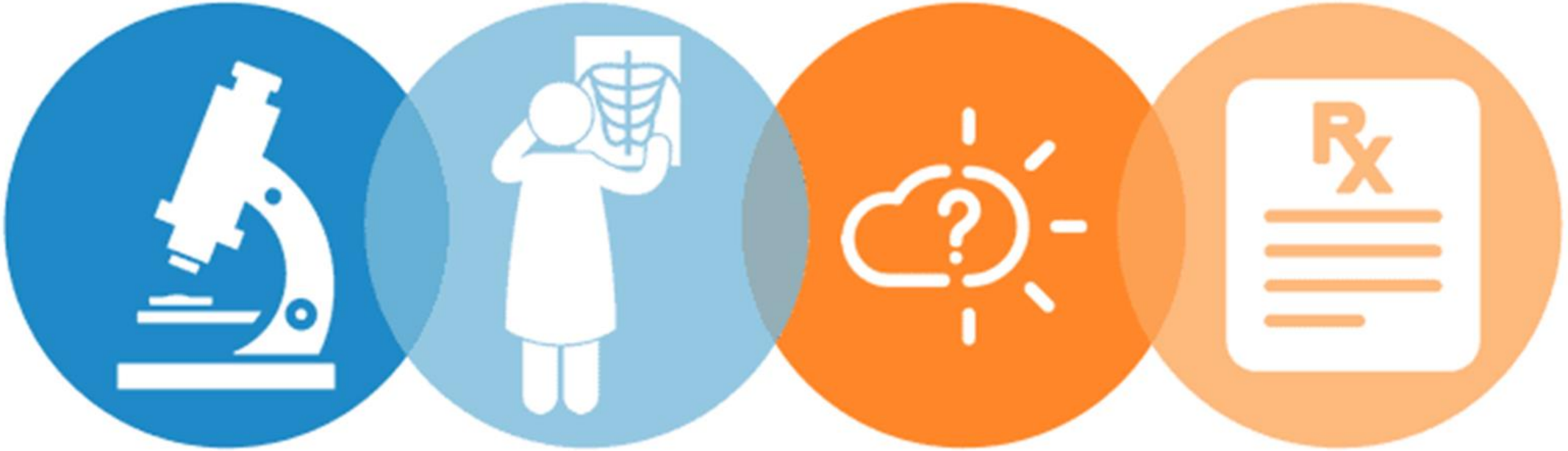
build the **thing right** via continuous integration & delivery

6

Workflow of Data Science



Types of Analytics



DESCRIPTIVE ANALYSIS
Explains what happened

DIAGNOSTIC ANALYSIS
Explains why it happened

PREDICTIVE ANALYSIS
Forecast what might happen

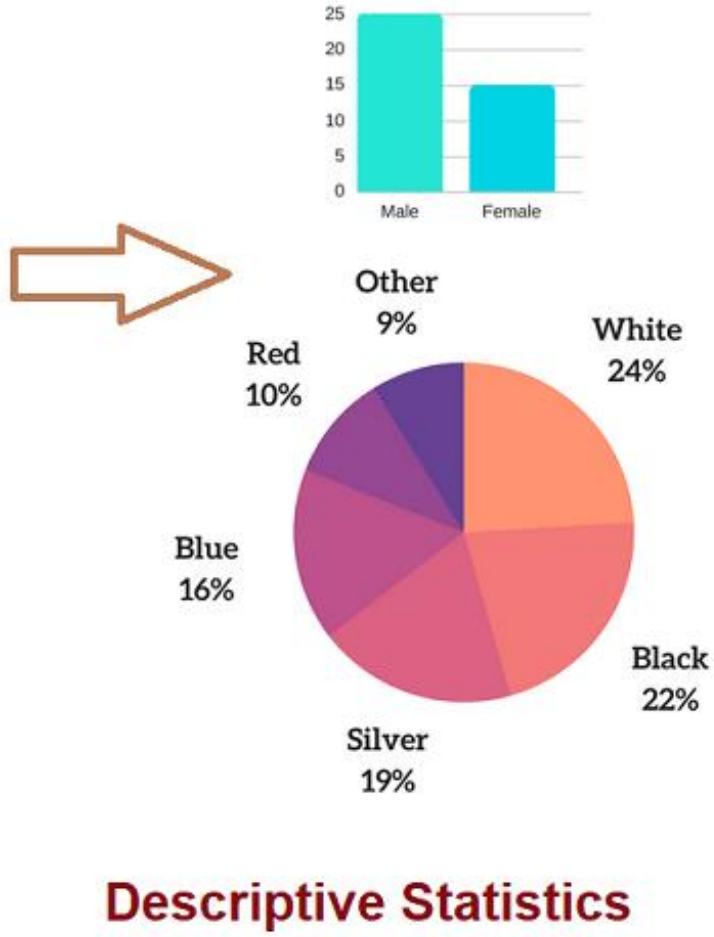
PRESCRIPTIVE ANALYSIS
Recommends an action based on the forecast

Descriptive analysis

- Describe the data
- Common statistics (counts, average, etc..)
- Typical reporting methods:

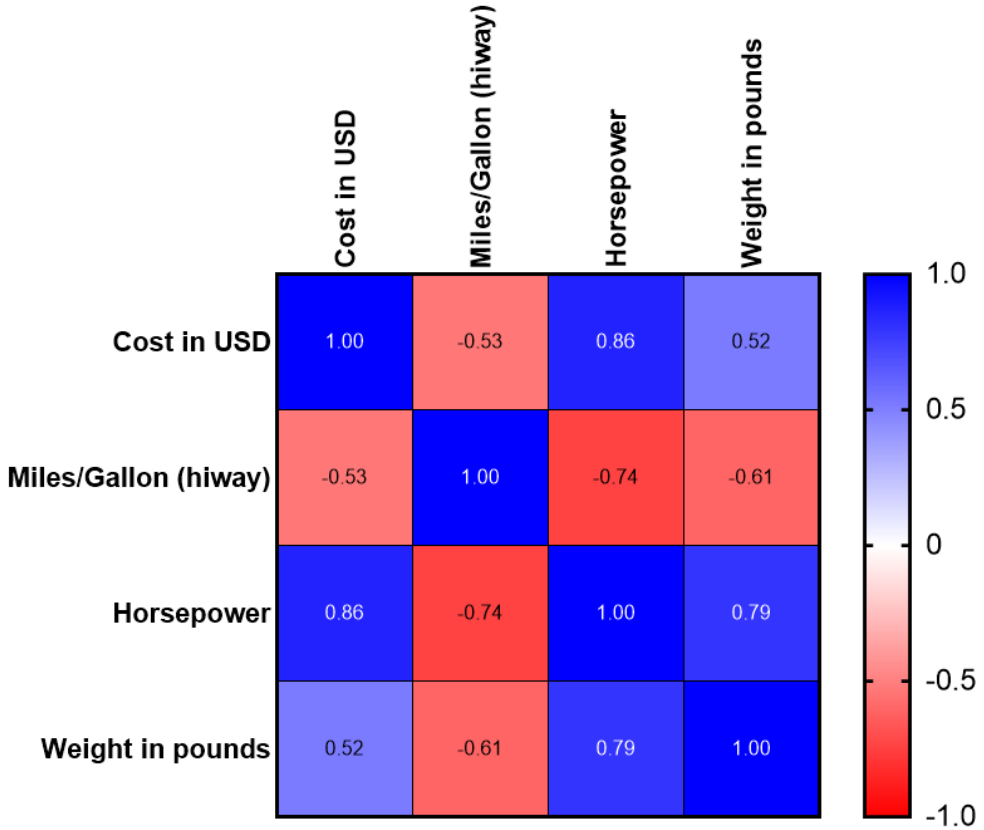
Tables, charts, written narratives

A	B	C
Respondent Number	**Age**	**Gender**
1	22	M
2	37	F
3	45	F
4	62	F
5	28	M
6	45	M
7	88	F
8	61	M
9	95	M
10	27	M
11	39	F
12	43	M
13	55	F
14	59	F



Diagnostics analysis

- Why did it happen?
- Drill-down techniques
- Data/ knowledge discovery
- correlations

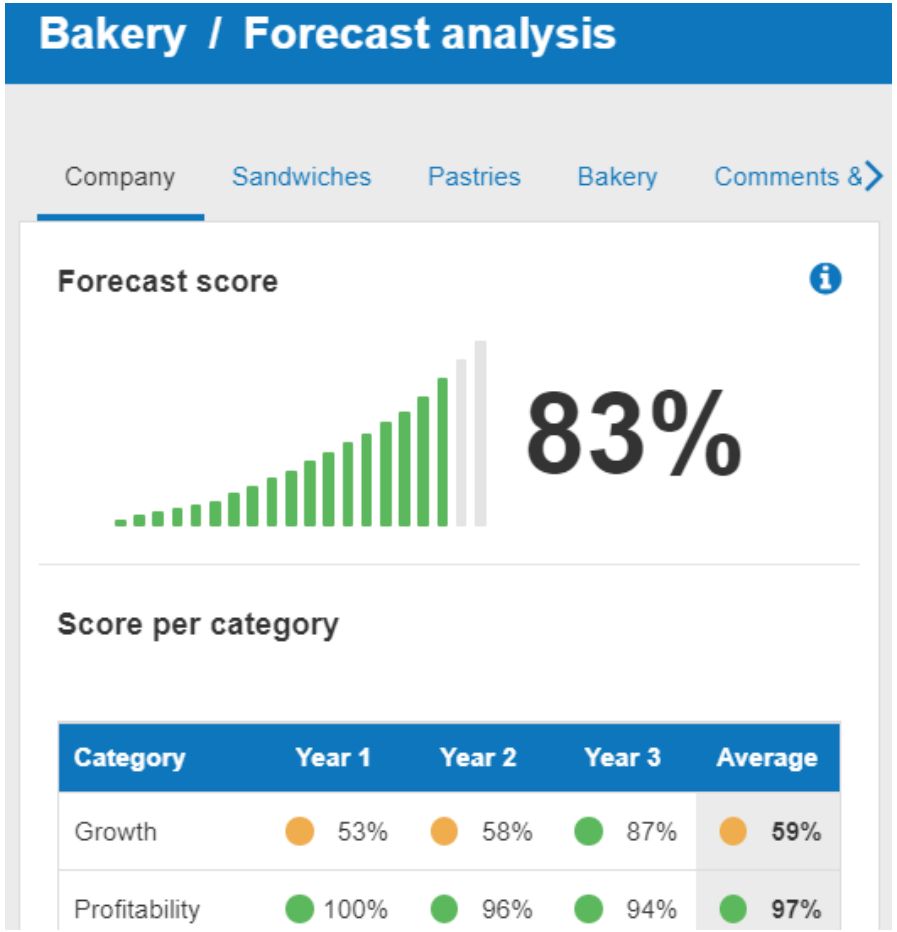


Correlation analysis

Source: https://www.graphpad.com/support/faq/what-is-the-difference-between-correlation-and-linear-regression/

Predictive analysis

- Predict instead of describing or classifying
- Rapid analysis
- Relevant insights
- Ease of use
- All predictive analytics are probabilistic in nature



Forecast analysis

Source: https://www.thebusinessplanshop.com/en/help/dashboard/forecast/

Prescriptive analysis

- What should be done?
- Is characterized by techniques such as

➢Graph analysis

➢Simulation

➢Neural networks

➢Machine learning

Credit card fraud detection

Steps towards Problem Solving using Data Science techniques

1. Get oriented
2. Discover business objectives
3. Define scope
4. Formulate a BA plan
5. Define requirements
6. Technical implementation
7. Perform analysis
8. Knowledge transfer/ solution
9. Implement the knowledge

1. Get oriented

- Clarify the role of a business analyst
- Identify stakeholders
- Understand the project history
- Understand existing systems and business process

2. Discover business objectives

- Discover stakeholder’s expectations
- Understand ‘why’ of the projects
- Reconcile conflicting expectations
- Clarify the business objectives

3. Define scope

- Develop strategies to determine technology and business process change
- Draft and review the scope statement
- Confirm the business objectives

4. Formulate a BA plan

- Choose the most appropriate types of BA deliverables
- Define specific set of deliverables
- Identify timelines for completing deliverables

5. Define requirements

- Collect the data needed
- Define clear, concise, concrete, complete and consistent requirements (5C)

6. Technical implementation

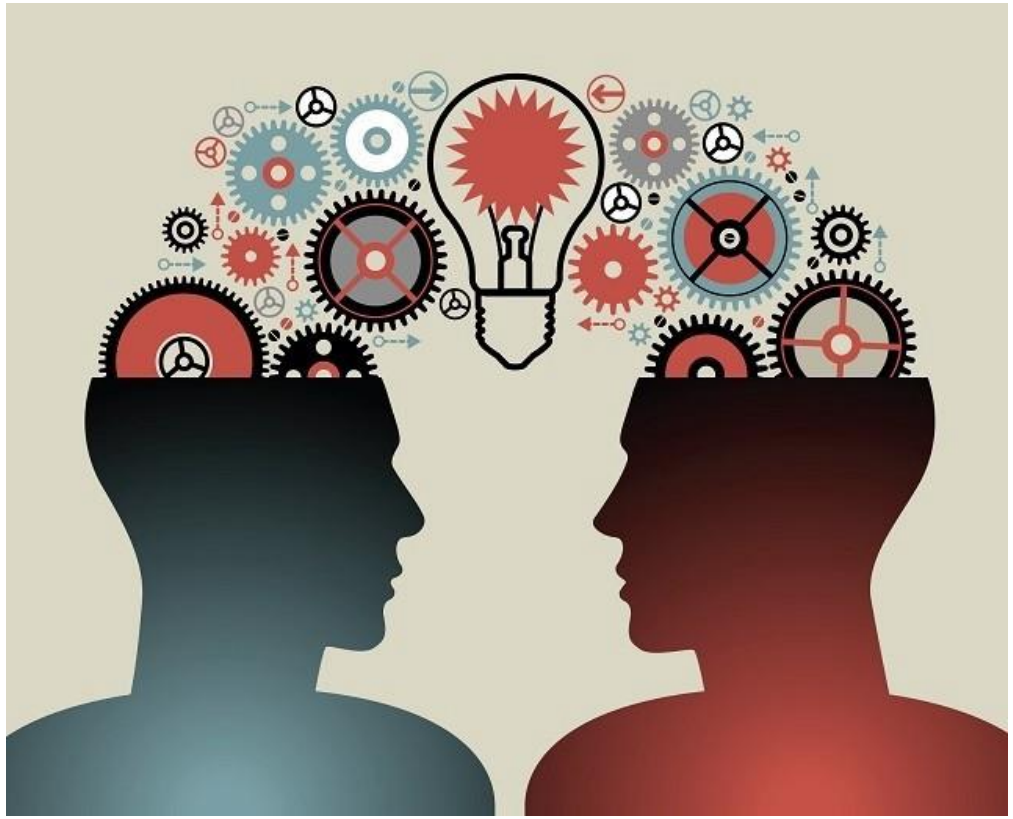
- Decide the final solution design
- Update requirement documentation
- Specify software and other technical implementations

7. Perform analysis

- Perform analysis using the correct data and technical requirement
- Visualize data and interpret the result

8. Knowledge transfer/ solution

- Analyse and develop interim and future business process documentation
- Train the business users
- Lead the change implementation



9. Implement the knowledge

- Business users make changes to existing process to solve problems
- Evaluate the progress using data, metrics and trend to assess measurable business values

Data sources

BIG DATA

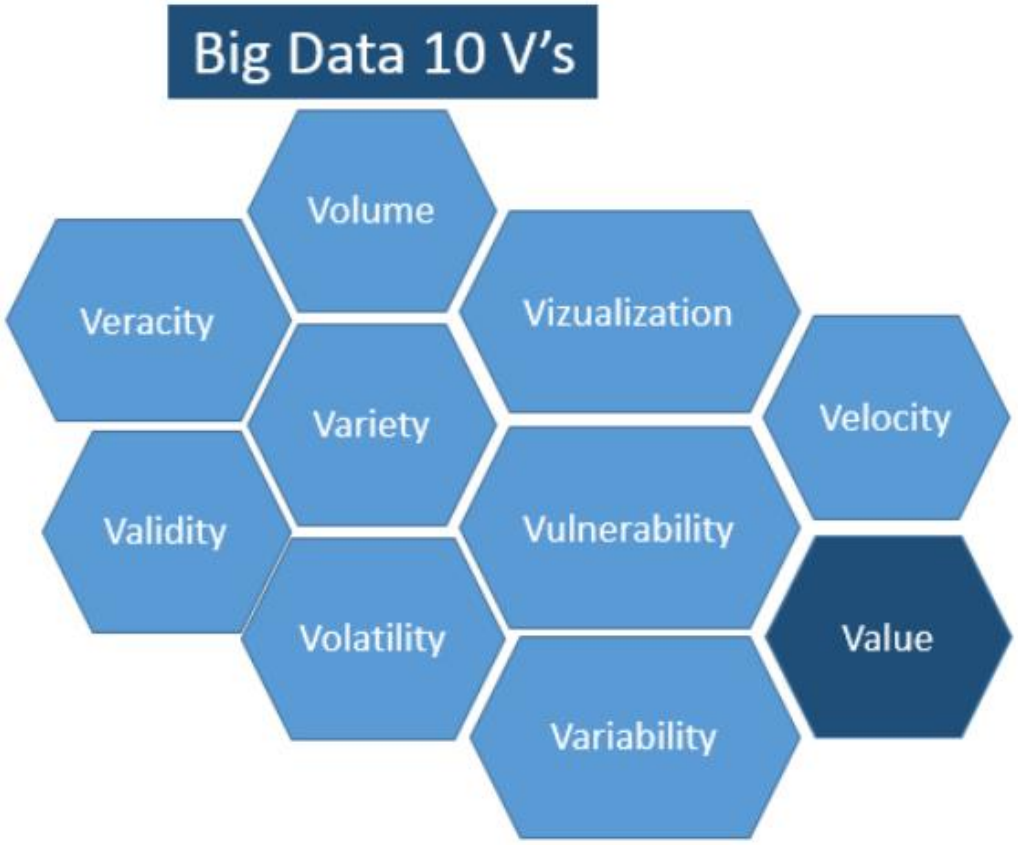
Massive amounts of business data from a wide variety of sources, much of which is available in real time, and much of which is uncertain or unpredictable.

DATA

Numerical or textual facts and figures that are collected through some type of measurement process.

INFORMATION

Result of analyzing data; that is, extracting meaning/ knowledge from data to support evaluation and decision making.



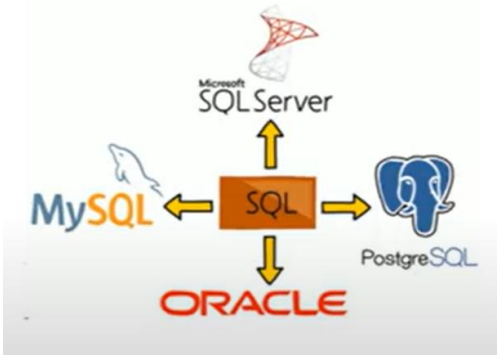
Data sources

Flat files

Google Sheets

CSV

Relational Databases



NoSQL Databases

mongoDB.

redis

APACHE HBASE

CouchDB

Cloud

IBM Bluemix™

Microsoft

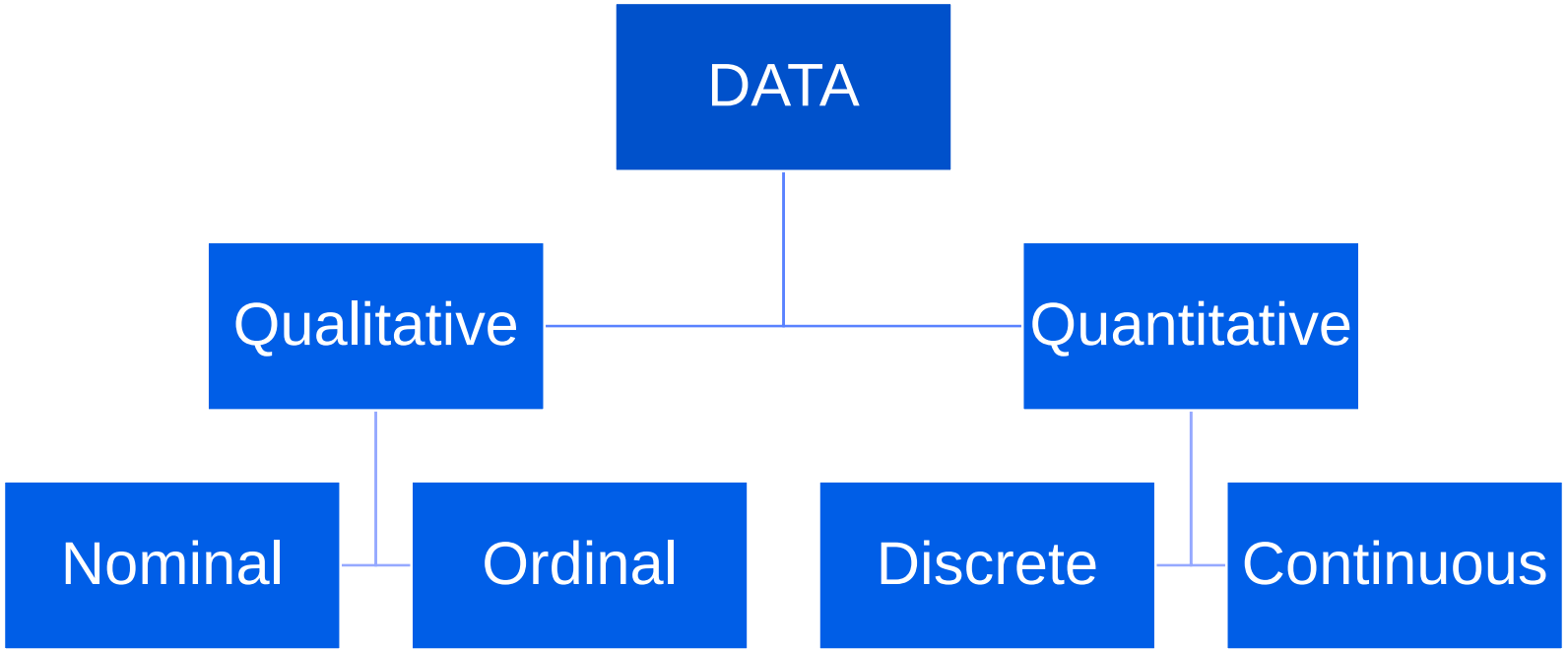
amazon web services™

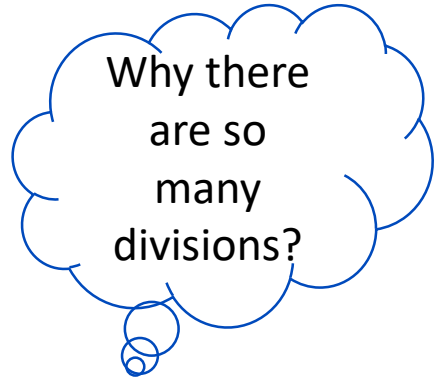
Google

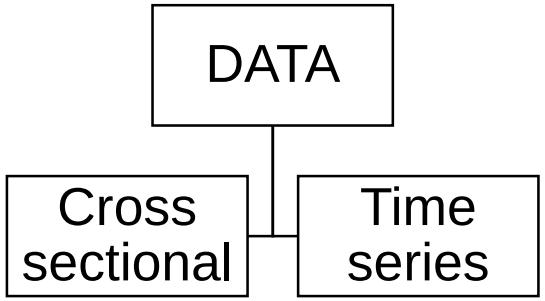
ORACLE

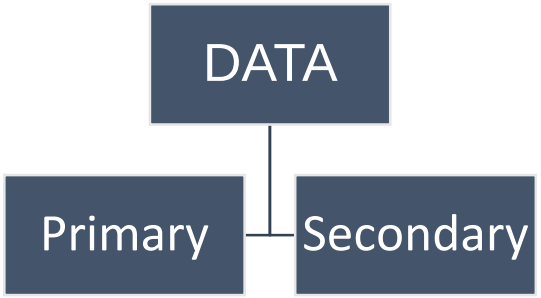
OPENSHIFT

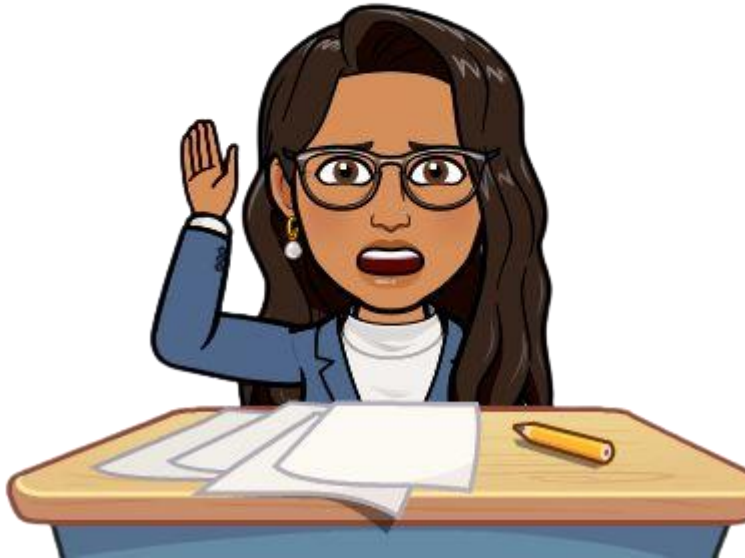
Types of data



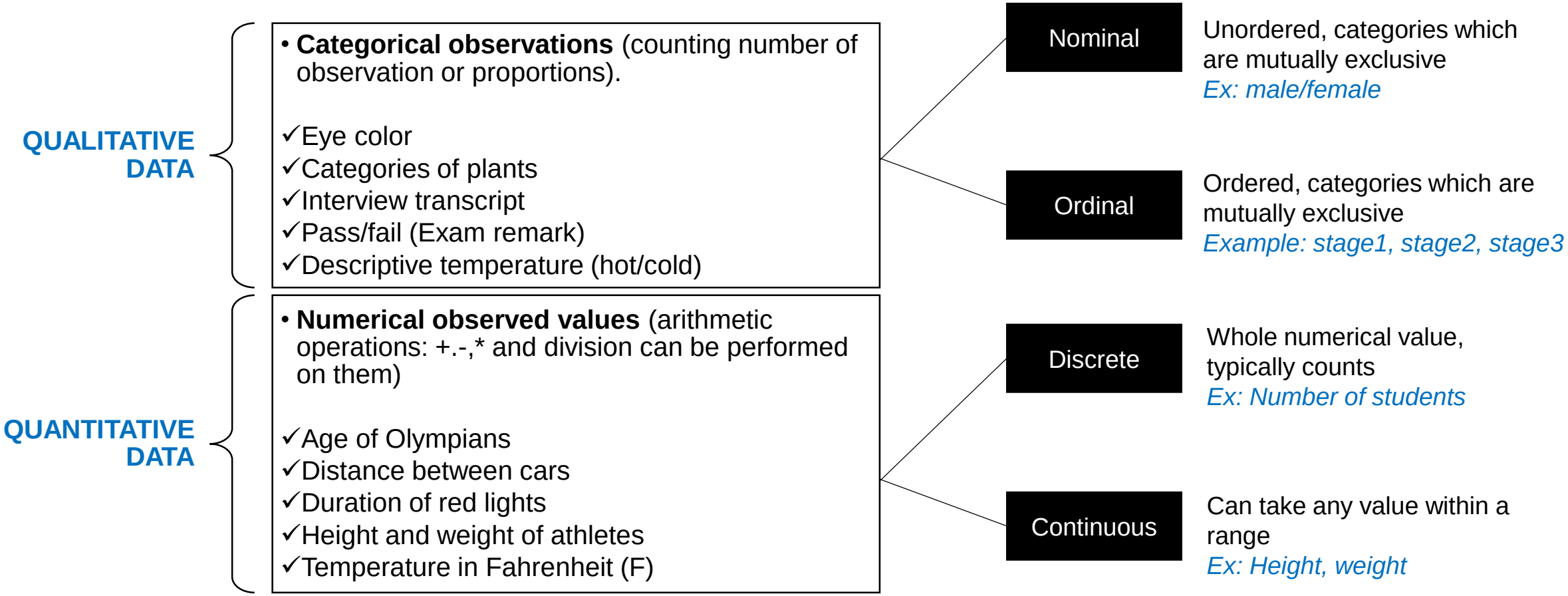








Types of data



Types of data

CROSS SECTIONAL

- Considers the **same variables over a certain period of time**

➢Opening prices of 10 selected stocks on a given date is an example of cross-sectional data. It is to be noted that cross-sectional data should have the same or similar characteristics for comparison

TIME SERIES

- Uses **different data for a given point in time**

➢The daily opening prices of a share over a half-yearly period. It is notable that a too long or a too-short gap of observation of time series data may lead to erratic results

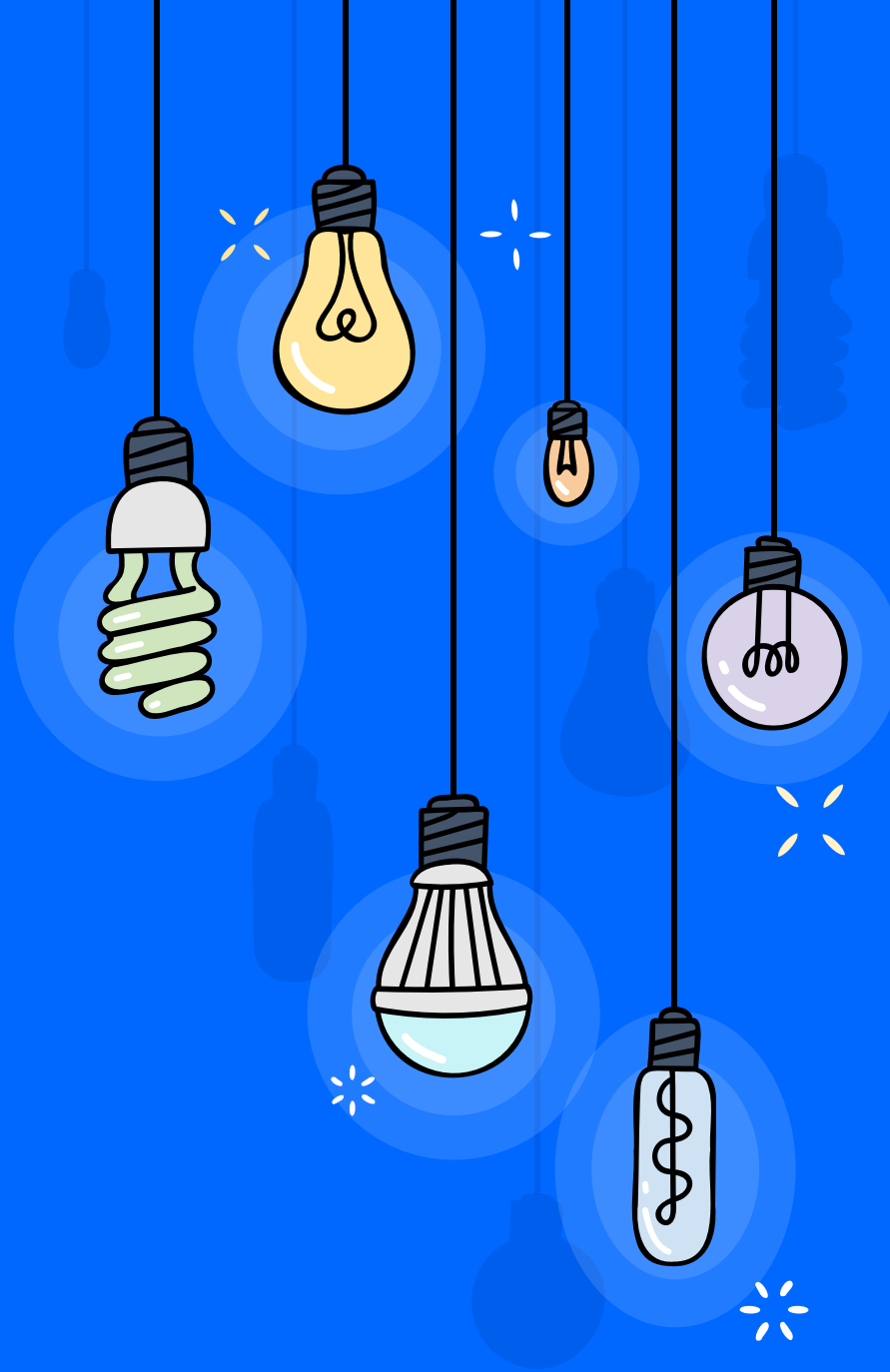
Types of data

	PRIMARY DATA	SECONDARY DATA
Meaning	Primary data refers to the first hand data gathered by the researcher himself.	Secondary data means data collected by someone else earlier.
Data	Real time data	Past data
Process	Very involved	Quick and easy
Source	Surveys, observations, experiments, questionnaire, personal interview	Government publications, websites, books, journal articles, internal records etc.
Cost effectiveness	Expensive	Economical
Collection time	Long	Short
Specific	Always specific to the researcher's needs.	May or may not be specific to the researcher's need.
Available in	Crude form	Refined form
Accuracy and reliability	More	Relatively less

Data Visualisation and Model Development

7

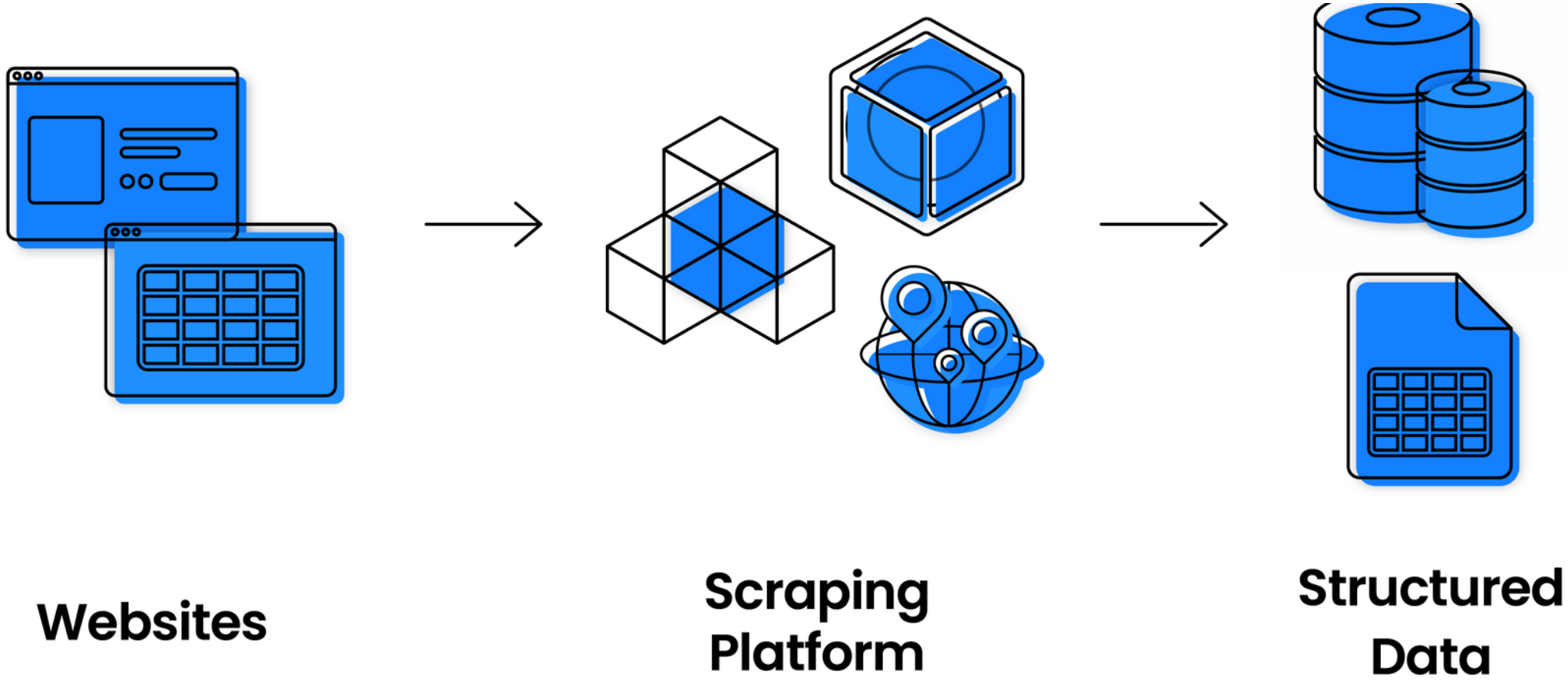
Data Collection Activities



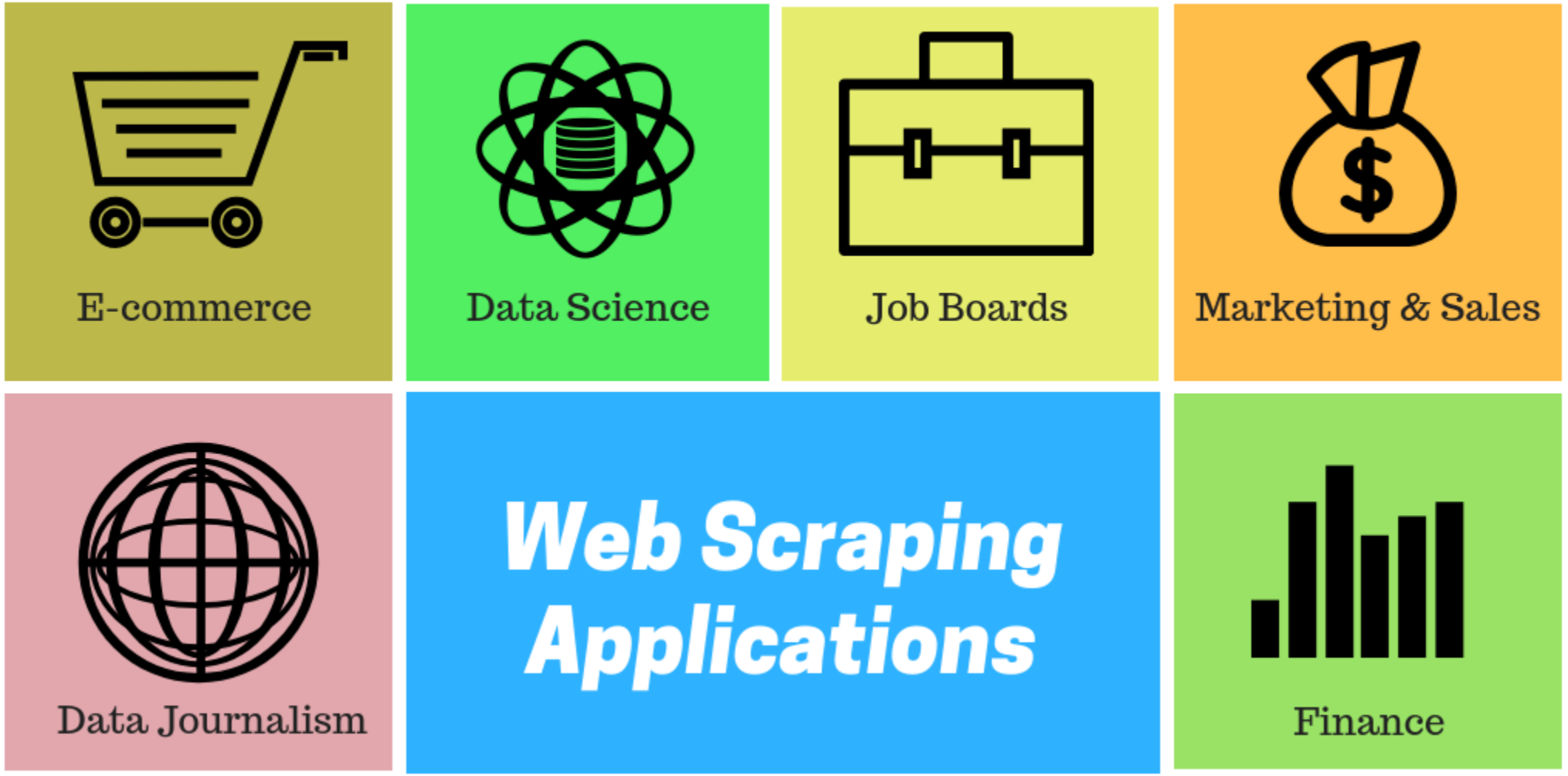
Free Open-Source Databases

1. **GitHub**: Many researchers and organizations share their datasets on GitHub. You can use the search feature to find repositories containing datasets relevant to your needs.
2. **UCI Machine Learning Repository:** The University of California, Irvine hosts a collection of datasets specifically curated for machine learning tasks. You can find a wide range of datasets here.
3. **Google Dataset Search:** Google provides a dedicated search engine to find datasets. It scours the internet for publicly available datasets on various topics.
4. **Data.gov:** This is the official open data portal of the United States government. It provides access to a vast collection of datasets across different domains.
5. **World Bank Open Data:** The World Bank offers free access to global development data. It includes data on economic, social, and environmental topics.
6. **Amazon AWS Public Datasets:** Amazon Web Services provides a collection of public datasets that you can access for free. The datasets cover various domains, including biology, economics, climate, and more.
7. **Open Data Portals:** Many cities and governments have their own open data portals, where they publish datasets related to their jurisdictions. Examples include data.gov.uk for the UK and data.gov.in for India.
8. **Reddit Datasets:** Reddit has a community-driven list of interesting datasets that users have compiled over time.
9. **Quandl:** Quandl is a platform that hosts a large collection of financial and economic datasets.
10. **DataHub:** DataHub is an open data platform that provides access to a wide range of datasets in different formats.
11. **Eurostat:** Eurostat is the statistical office of the European Union, providing access to a wealth of European statistics.
12. **NASA's Open Data Portal:** If you are interested in space-related datasets, NASA offers an open data portal with various datasets related to space exploration and Earth sciences.

Web scraping for data



Web scraping Applications



What is next in DATA SCIENCE?

1.Machine Learning (ML) Advancements
- ML algorithms are becoming more sophisticated, enabling higher accuracy and performance.
- Reinforcement learning, transfer learning, and self-supervised learning are gaining prominence.
- The integration of ML with big data and cloud computing is unlocking new possibilities.

2.Explainable AI (XAI)
- With AI becoming more pervasive, the demand for explainable models is increasing.
- Interpretable AI ensures transparency, accountability, and compliance with regulations.
- XAI techniques like LIME, SHAP, and attention mechanisms are gaining traction.

3.Edge Computing and IoT Convergence
- The proliferation of IoT devices is generating massive data streams.
- Data science is moving closer to the edge, reducing latency and enhancing real-time insights.
- Edge AI empowers autonomous devices, enabling new applications in various domains.

4.Data Ethics and Privacy
- As data collection grows, ensuring ethical use and safeguarding privacy is paramount.
- Data scientists must navigate ethical dilemmas while handling sensitive information.
- Regulatory frameworks like GDPR and CCPA are reshaping data science practices.

5.Automated Machine Learning (AutoML)
- AutoML platforms streamline the model development process.
- Democratizing data science, enabling non-experts to leverage its power.
- Hyperparameter optimization, feature engineering, and model selection made more accessible.

6.Quantum Computing Impact
- Quantum computing promises unprecedented computational power.
- Data scientists are exploring quantum algorithms for optimization and data analysis.
- Quantum machine learning holds the potential to revolutionize various industries.

7.Multi-modal Data Analysis
- Data is increasingly diverse, including text, images, audio, and video.
- Integrating multi-modal data unlocks new insights and enhances predictive capabilities.
- Transfer learning between modalities is an exciting area of research.

8.Data Science Collaboration
- Collaboration between data scientists, domain experts, and stakeholders is crucial.
- Interdisciplinary teams can tackle complex problems with a holistic approach.
- Enhanced communication and teamwork will drive impactful data-driven solutions.

Industrial Revolution IR5.0

THANK YOU!
All The Best ☺