
PLURALSIGHT

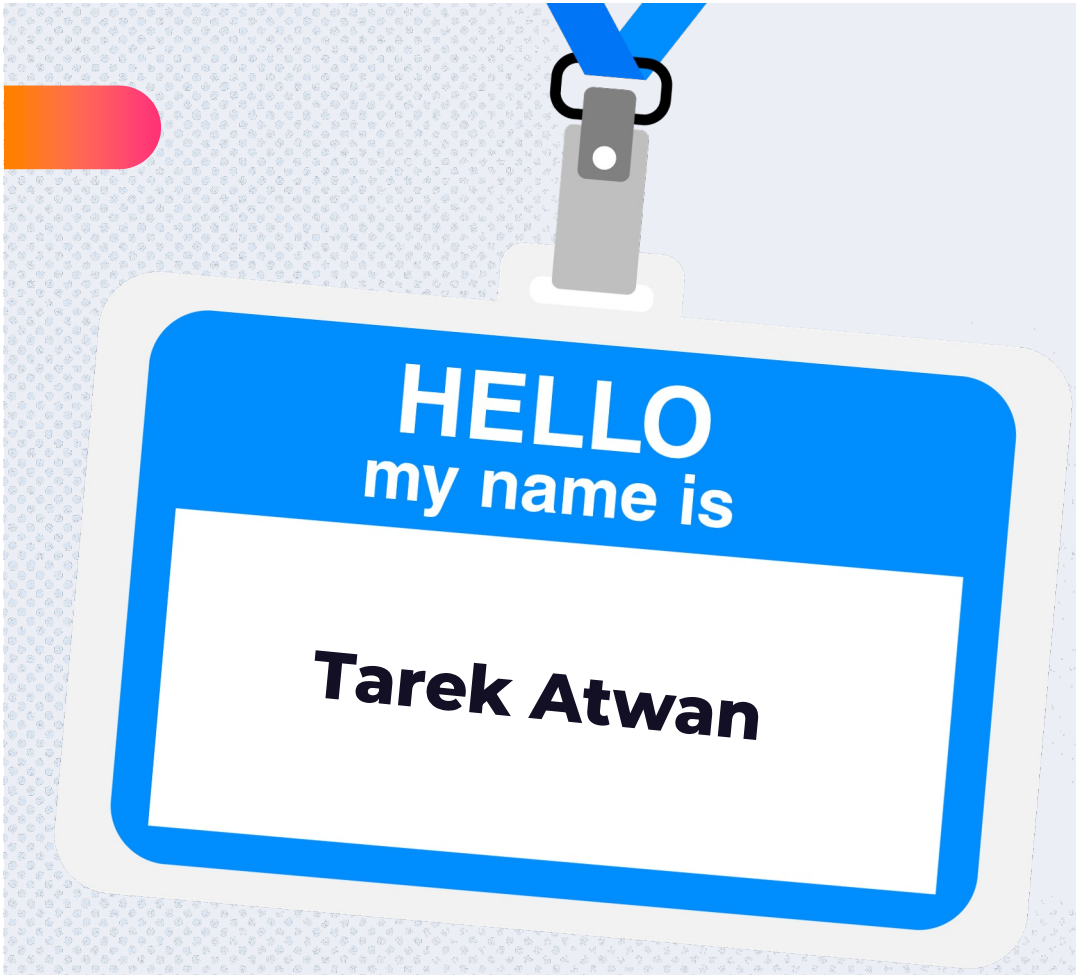
M&T Bank

Data Academy

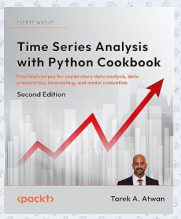
Week 1

Tarek Atwan
Instructor, Pluralsight

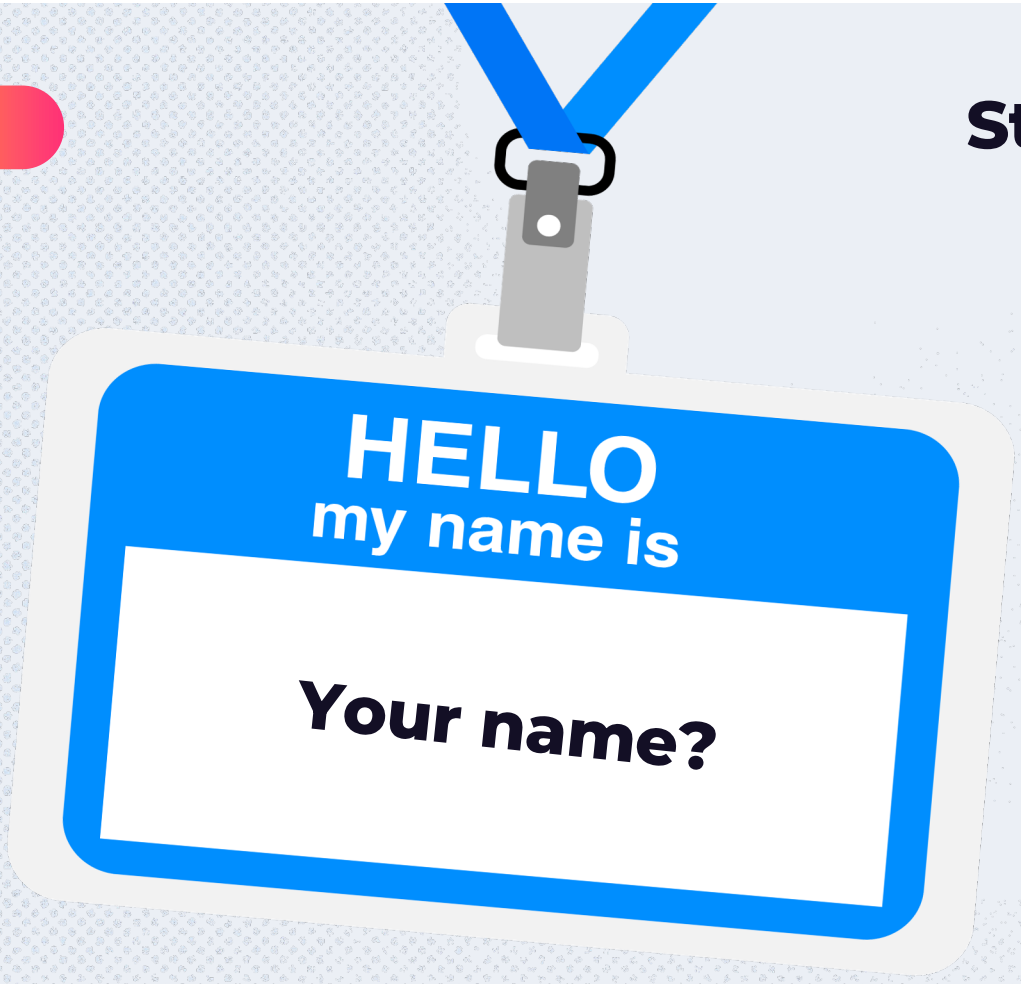
PLURALSIGHT



About Me:



- Book Author
- 17+ Years Consulting Services
- 4+ Years Instructor
- 2 Startups
- World Traveler
- Gym Rat



Student Instructions

- Job title?
- What are your expectations from the Data Academy?
- What is your related experience, if any?
- Any Fun fact?

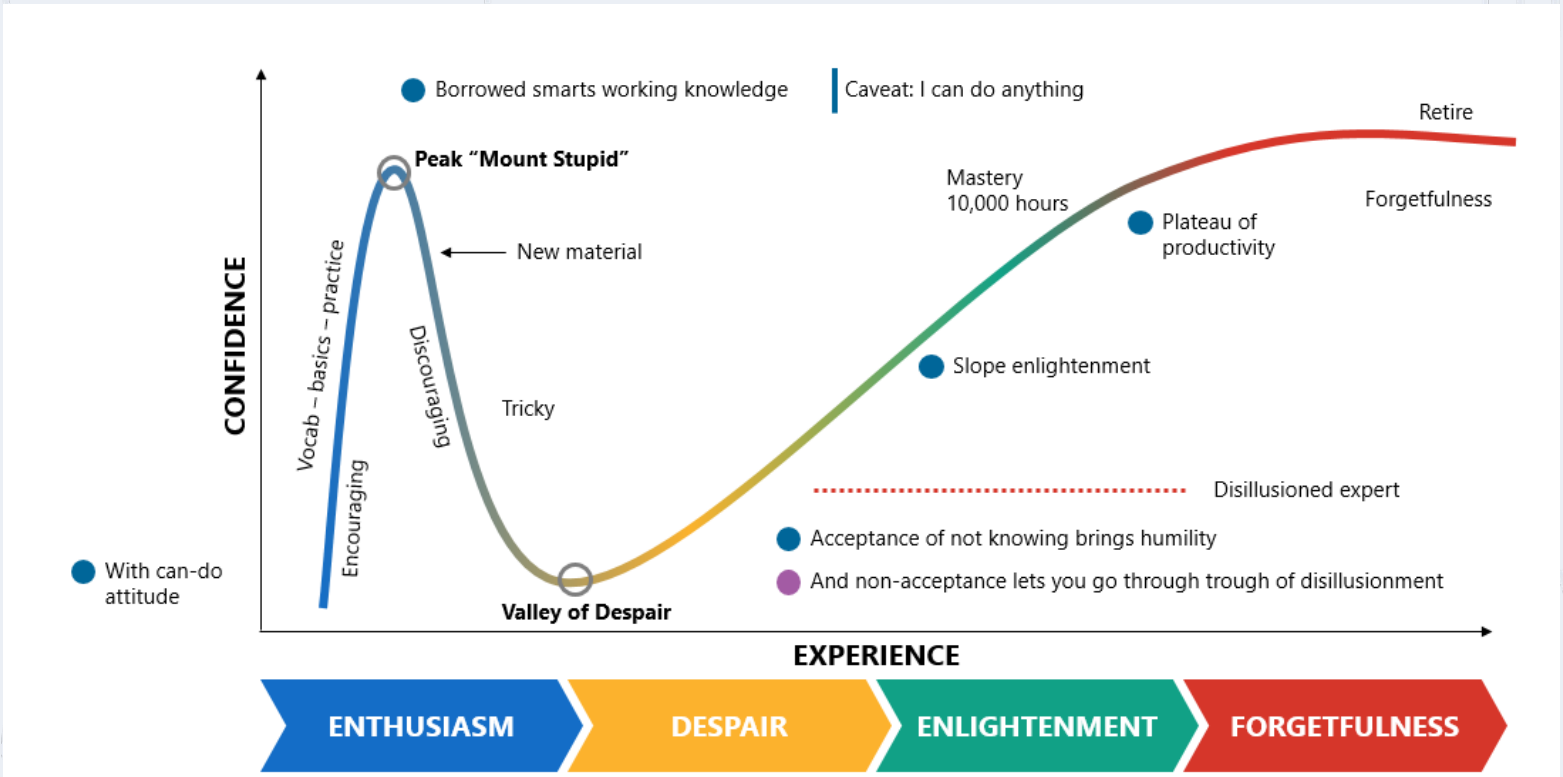
Objectives

At the end of this bootcamp, you will:

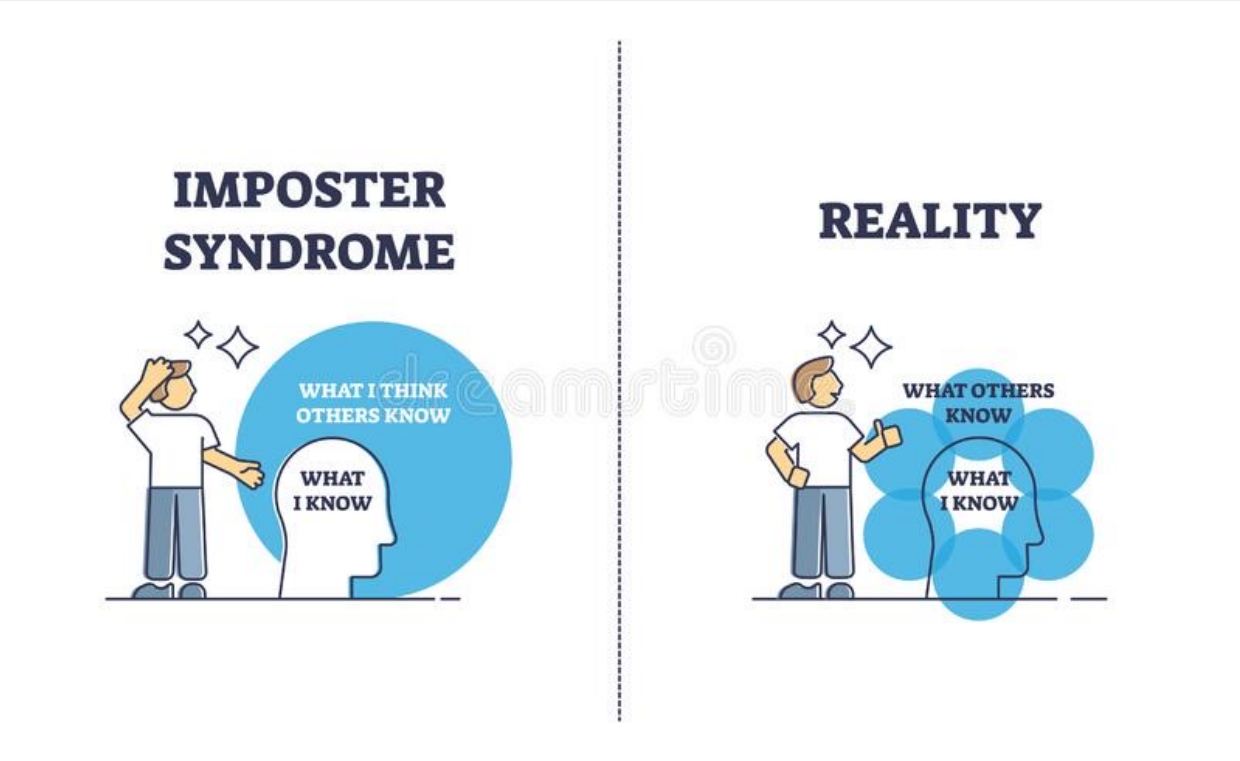
Become a **data ninja** – you will sense bad data, master new weapons, torture the data until it confesses the truth!

DIVING INTO ANALYTICS

Dunning Kruger Effect



Imposter Syndrome



TYPES OF ANALYTICS

What is Data Analytics

- Data analytics is the science of analyzing **raw data** to discover useful **information**, which can be used to optimize processes, improve decision-making, and foster business growth.

- It involves using data, techniques, and tools to identify patterns and trends, generating actionable insights that support informed decision-making

Type of Analytics

- **Descriptive** What happened?
- **Diagnostic** Why did it happen?
- **Predictive** What is likely to happen in the future?
- **Prescriptive** What action should be taken?

Data-Driven Decision Making

What is it?

- Using Facts, Metrics, and Data to Guide Strategic Business Decisions

Why is it important?

- Enhance Accuracy
- Reduce Risks
- Fosters Proactive Strategies

Data-driven decision making refers to the process of using data, facts, metrics, and insights to guide strategic business decisions that align with goals, objectives, and initiatives

Being a Data-Driven Organization

What is it?

- Leverage data at every level for better decision making

Why is it important?

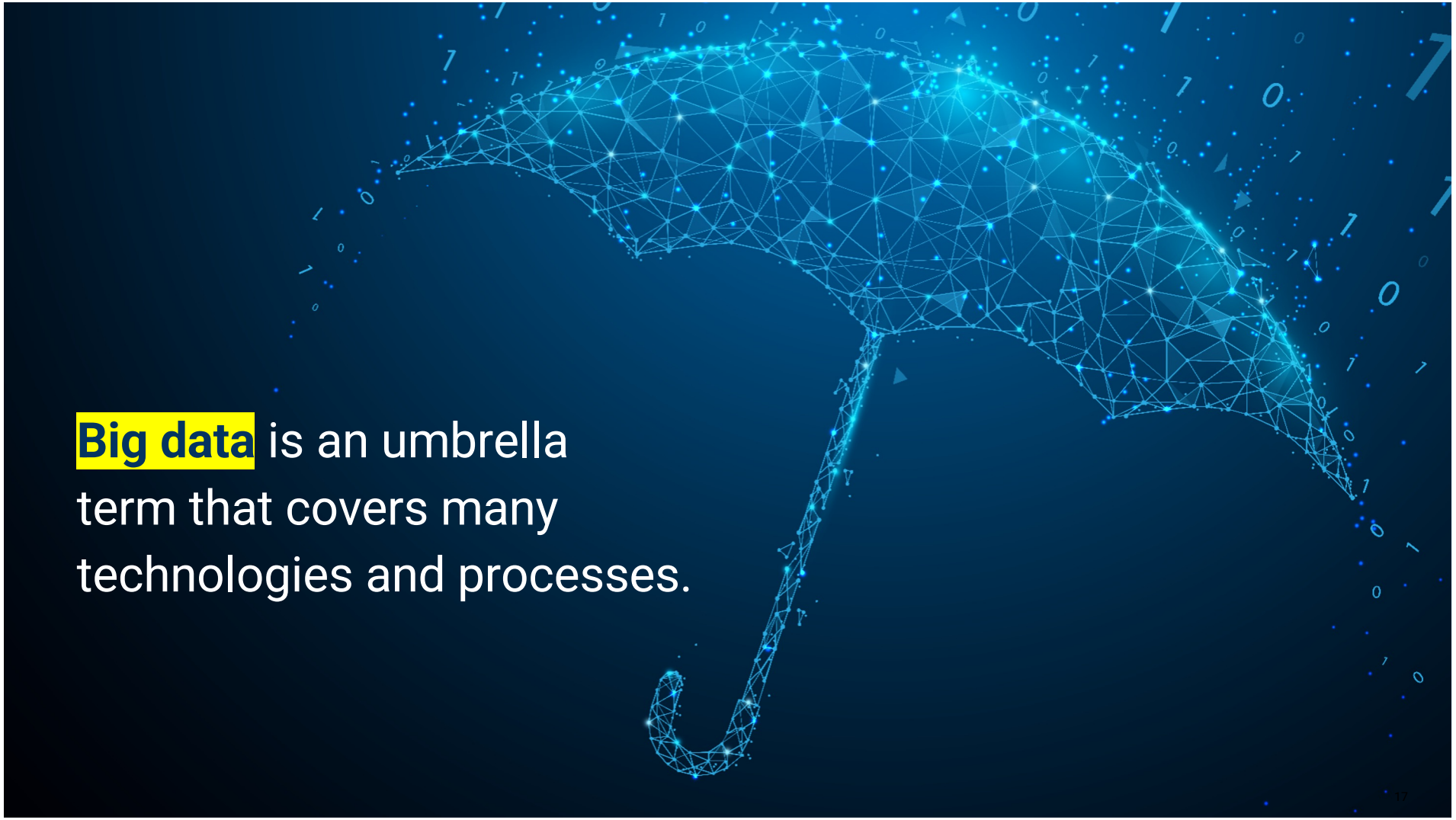
- Ensures higher quality
- Efficient operations
- Increased innovation
- Data democratization

Data democratization

Empowering individuals to use data for informed decisions

A data-driven organization is one that uses data insights to guide its decision-making processes. It is a long-term continuous process that requires addressing business needs, people mindset, and culture change

Big Data

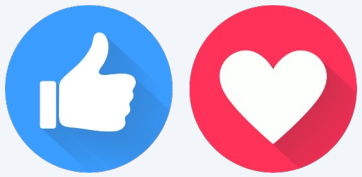

Big data is an umbrella
term that covers many
technologies and processes.

Big Data Can Be Anything

Email

Stocks

Tweets

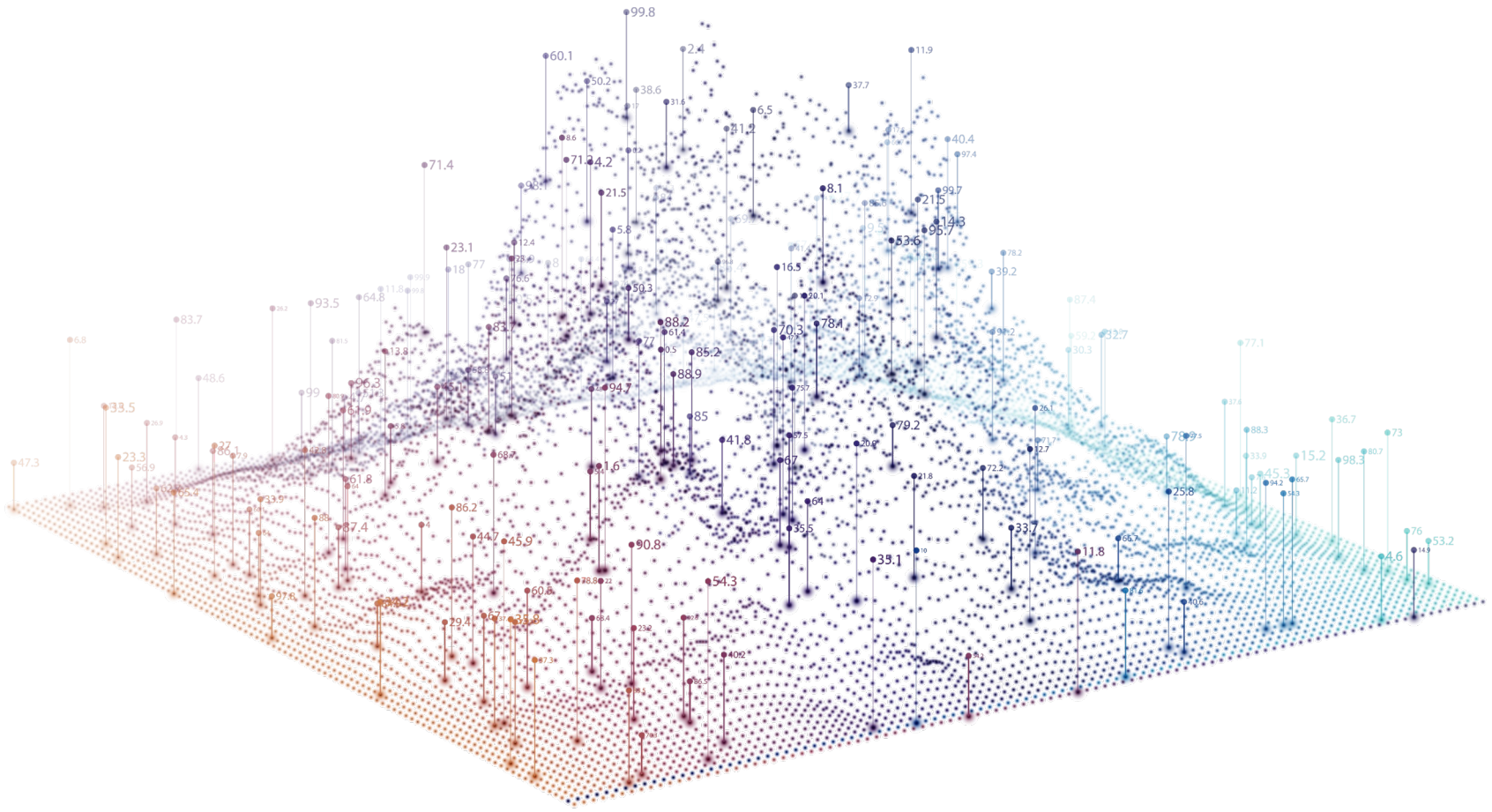


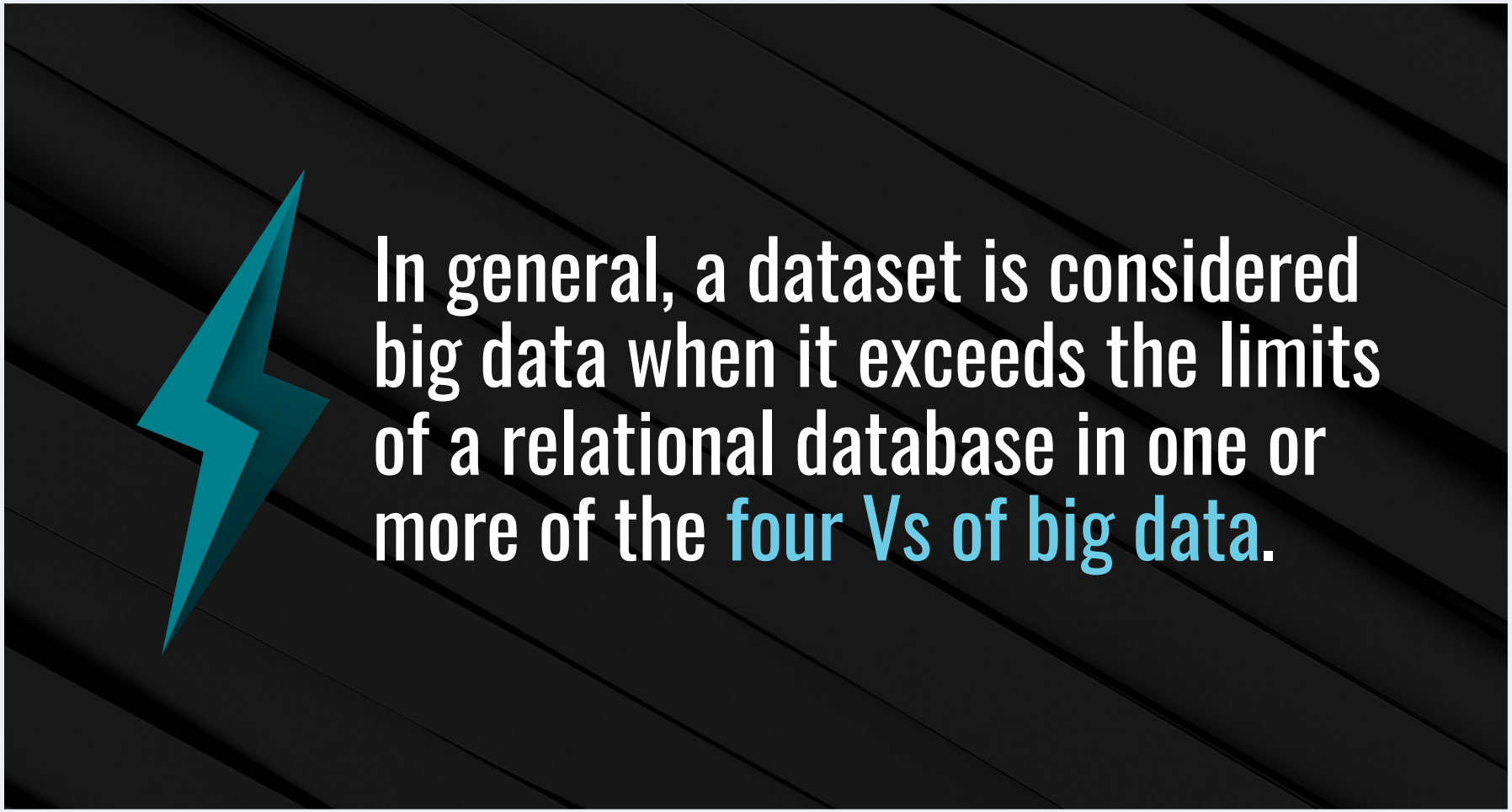
Social media posts

Supply chain alerts

Cell towers

At what point is a dataset large enough to be considered big data?




In general, a dataset is considered big data when it exceeds the limits of a relational database in one or more of the four Vs of big data.

Intro to Big Data

The four **V**s of big data:

01 **Volume:** The size of the data.

02 **Velocity:** How quickly the data comes in.

03 **Variety:** The diversity of data.

04 **Veracity:** The uncertainty of data.

What is Big Data

What is it?

- Massive volume of data that can't be processed using traditional databases or techniques
- Volume, Variety, Velocity, Veracity (Four V's)

Why is it important?

- Uncover hidden patterns and correlations
- Uncover deeper insights
- **The importance of big data lies in how it is used.**

What are some challenges?

- Storage
- Processing
- Analysis
- Visualization

Diversity in Data Sources

Location

- On-Premises
- Cloud

Storage Type (Examples)

- Database (Relational, Non-Relational)
  - MySQL, SQL Server, MongoDB
- File-based
  - Excel, CSV, JSON

Data Forms and Speeds

Forms:

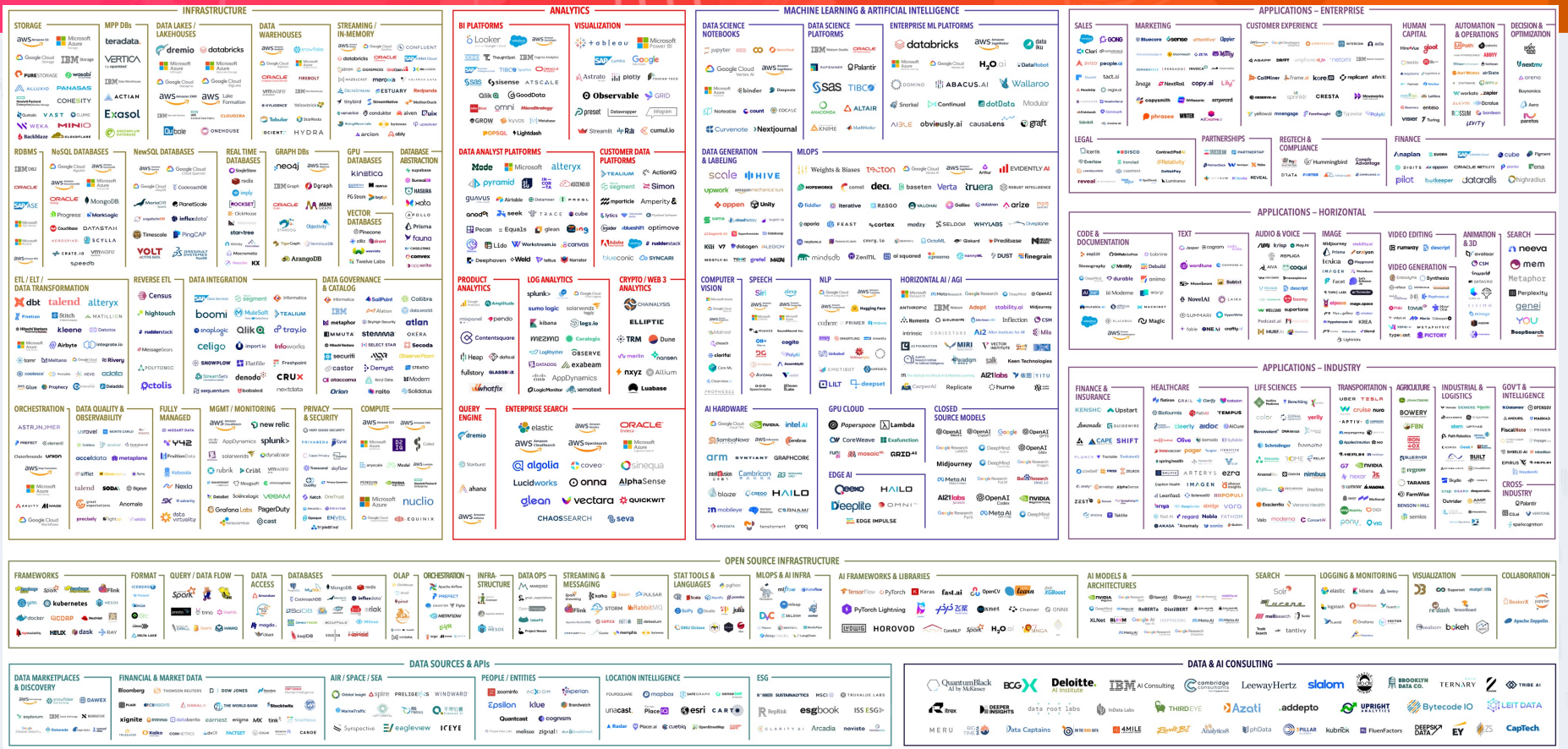
- **Structured** – Highly organized like (example: Excel Spreadsheet)
- **Unstructured** – No pre-defined structure (example: Emails or a Social Media Post)
- **Semi-Structured** – A hybrid (example: XML or JSON)

Speeds:

- **Interval-Based**
  - Daily, Weekly, Monthly
- **Real-Time**

Different forms and speeds serve varied analytics needs

2023 Data Analytics, ML, and AI Landscape



Distinctiveness of Data Projects

- **Complexity**: Data projects often involve multiple stakeholders, varied data sources, and complex transformations, making them inherently different and often more complex than other types of projects.

Components in Data Analytics Projects

When estimating the **effort level of a data project**, consider the following common components:

- **Project Scope**: Clearly define the project's objectives, deliverables, and boundaries to understand the overall effort required.
- **Data Collection and Preparation**: Assess the complexity of data sources, data quality, and data integration tasks. This includes data cleaning, transformation, and consolidation.
- **Data Modeling and Analysis**: Determine the level of complexity for data modeling, statistical analysis, and machine learning algorithms. This includes selecting appropriate techniques, testing and validation, and interpreting results.
- **Visualization and Reporting**: Evaluate the effort needed to design and develop interactive dashboards, reports, and visualizations that effectively communicate insights to stakeholders.

Components in Data Analytics Projects

When estimating the **effort level of a data project**, consider the following common components:

- **Infrastructure and Tools**: Consider the time and resources required to set up and configure the necessary hardware, software, and cloud services for data storage, processing, and analysis.
- **Project Management and Collaboration**: Estimate the effort for project planning, task management, team coordination, and stakeholder communication. This includes identifying dependencies, managing risks, and tracking progress.
- **Testing and Deployment**: Account for the time and resources needed to test the solution, address any issues or bugs, and deploy the final product to production.
- **Maintenance and Support**: Plan for ongoing maintenance, monitoring, and support of the data project, including updates, bug fixes, and user assistance.

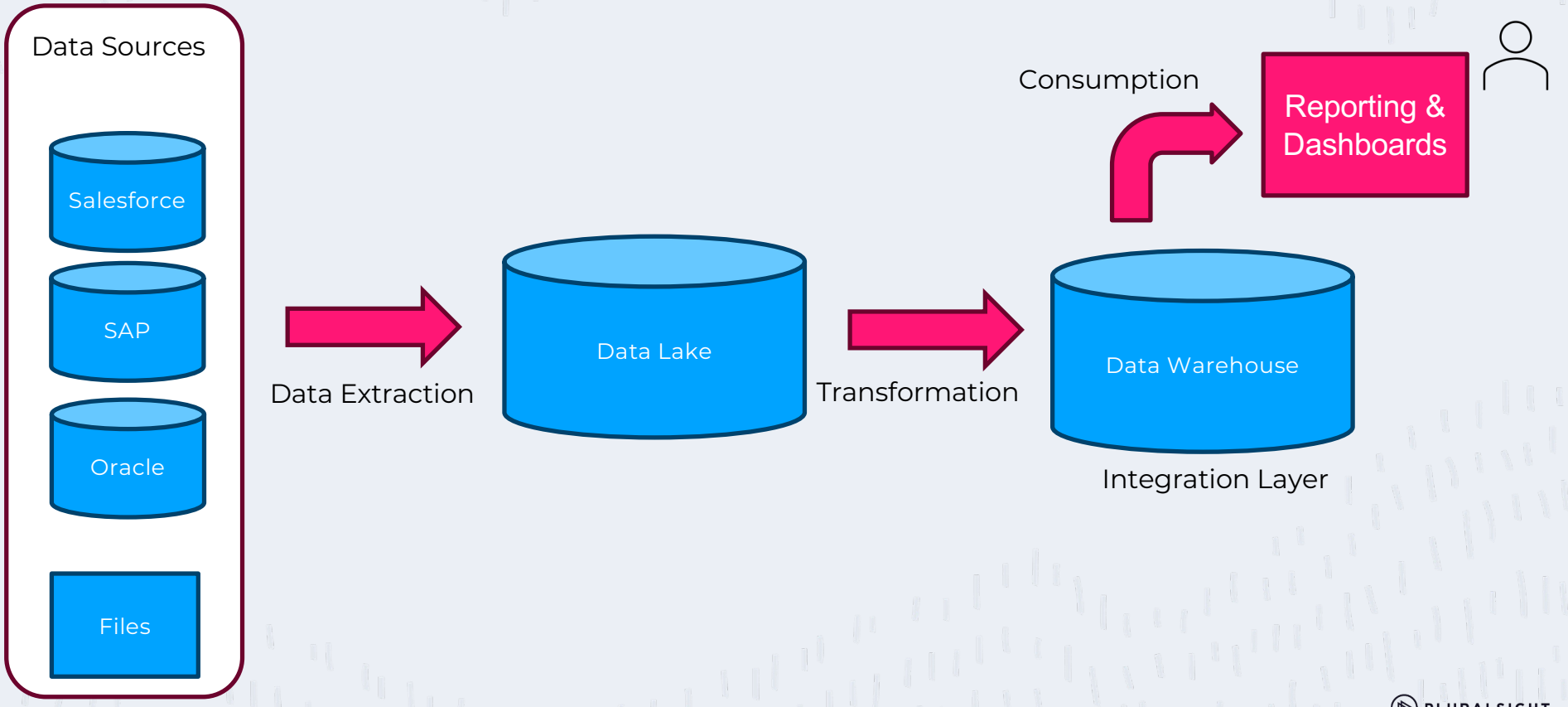
Components in Data Analytics Projects

When estimating the **effort level of a data project**, consider the following common components:

- **Training and Documentation**: Consider the effort required to train users on the new solution, create user guides, and document the project's processes and workflows.
- **Data Governance and Security**: Assess the effort needed to ensure data privacy, compliance with regulations, and the implementation of appropriate security measures.

By considering these components, you can more accurately estimate the effort level of your data project and plan for its successful execution.

Data Warehouse Project Architecture

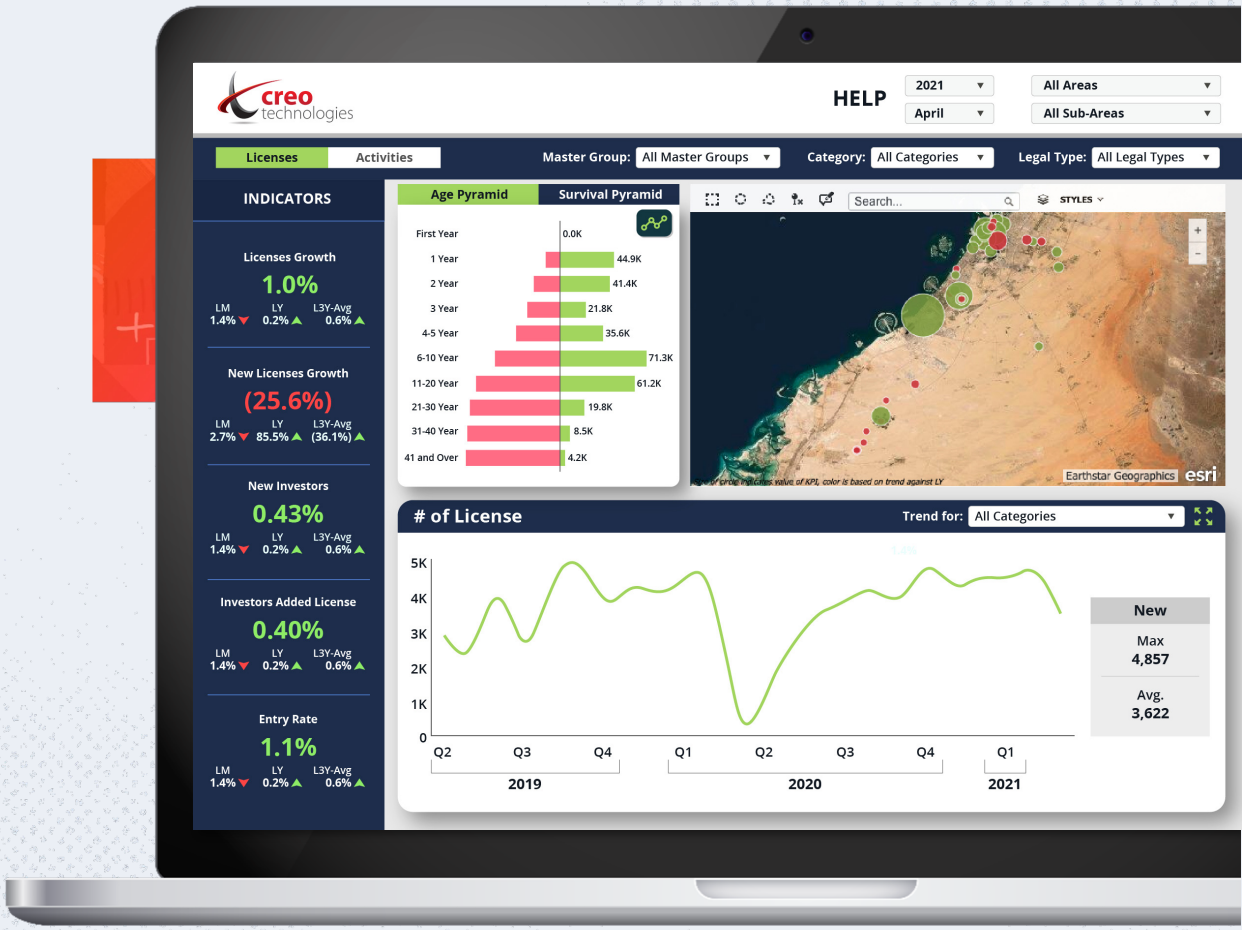


DATA VISUALIZATION AND STORYTELLING

PROJECT PORTFOLIO MANAGEMENT

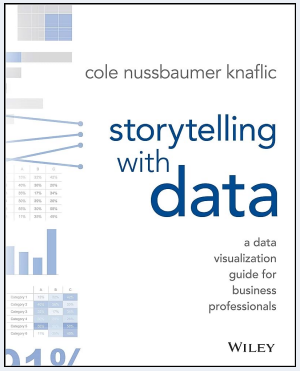
COVID-19 DASHBOARD

SALES



Storytelling with Data

1.Understand the context: Before creating a data story, it is essential to understand the context of the data, including the audience, the purpose of the story, and the message that needs to be conveyed
2.Choose an effective visual display: Selecting the right visual display is crucial to ensure that the data is presented in a clear and understandable way. The choice of visual display should be based on the type of data and the message that needs to be conveyed
3.Eliminate clutter: Clutter can distract the audience from the message and make the data story difficult to understand. It is essential to eliminate unnecessary elements and simplify the visual display to ensure that the message is clear
4.Direct the audience's attention: Highlighting the most important parts of the data and directing the audience's attention to them is crucial to ensure that the message is understood. This can be achieved through the use of color, size, and other visual cues
5.Think like a designer: Design thinking involves considering the audience's needs and preferences when creating a data story. It is essential to create a story that is visually appealing, easy to understand, and engaging
6.Leverage the power of storytelling: Storytelling is a powerful tool that can help make data more relatable and memorable. By incorporating a narrative into the data story, it is possible to create a more engaging and impactful message



Data Puke

"**Data puke**" is a term used in the data visualization field to describe charts, graphs, or dashboards that present an overwhelming amount of data without clear organization, focus, or meaningful interpretation.

Essentially, it's when a visualization offers a lot of data but little to no insight or clarity.

USER EXPERIENCE

User Interface (UI) and User Experience (UX) in Data Visualization

User Interface (UI)

- UI design is concerned with the visual elements and interactive features of a data visualization tool or application.
- It focuses on creating a visually appealing and intuitive interface for users to interact with the data.
- UI designers work on aspects such as layout, color schemes, typography, and iconography to ensure a cohesive and engaging user experience

User Experience (UX)

- UX design is responsible for the overall experience and satisfaction of users when interacting with a data visualization tool or application.
- It focuses on understanding user needs, goals, and behaviors to create a seamless and meaningful experience.
- UX designers work on aspects such as information architecture, interaction design, and usability testing to ensure that the data visualization tool is effective and easy to use

Role of UI and UX in Storytelling

User Interface (UI)

- Facilitate Interaction
- Present Information
- Guide Navigation

User Experience (UX)

- Enhance Navigation
- Ensure Satisfaction
- Drive Engagement

EXCEL TO DATABASES

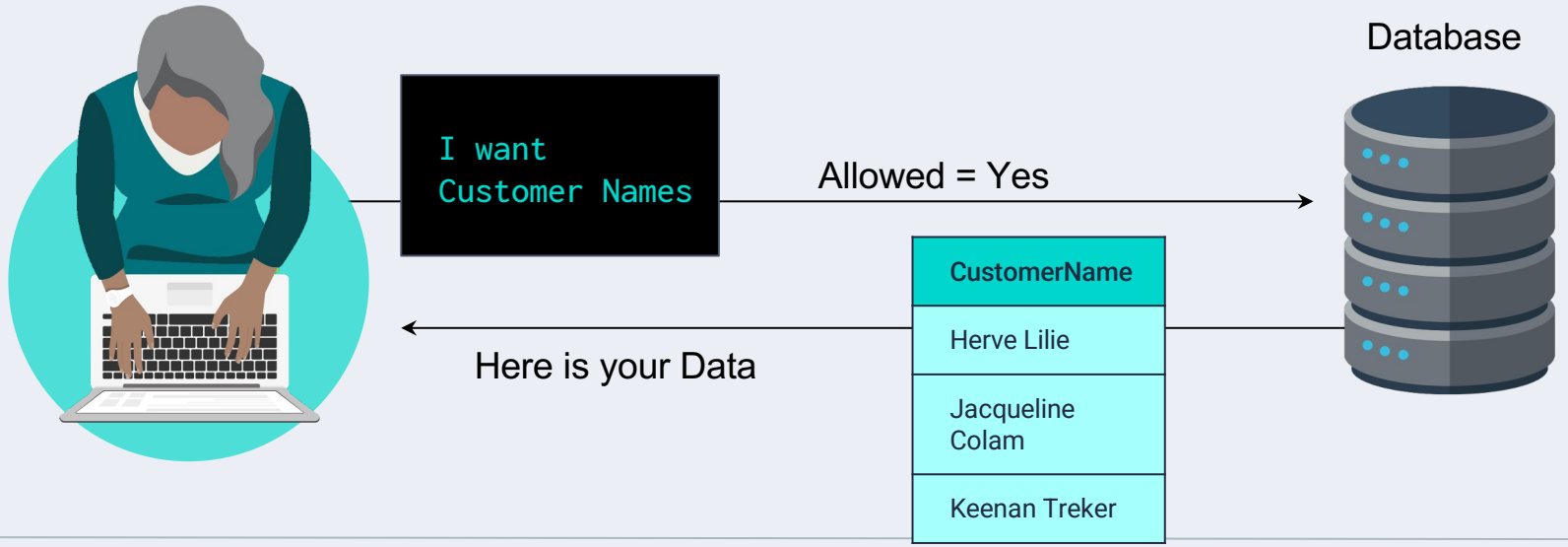
Excel Shortcomings & Limitations

- **Scalability**: Not suitable for handling very large datasets
- **Data Integrity**: Limited support for data validations
- **Concurrency**: Excel does not handle multiple users simultaneously well
- **Data Security**: Excel has limited security features.
- **Data Recovery**: Excel lacks sophisticated transaction management and recovery systems
- **Complex Relations & Queries**: Excel is not designed to handle complex relationships between different sets of data

A **database** is an organized collection of information or data, stored electronically in a computer system. It is designed to **efficiently** store, retrieve, update, and manage data.

Databases

The primary purpose of a database is to provide a way to store and retrieve data in a structured and controlled manner, ensuring **data integrity**, **security**, and **accessibility**.



What is SQL

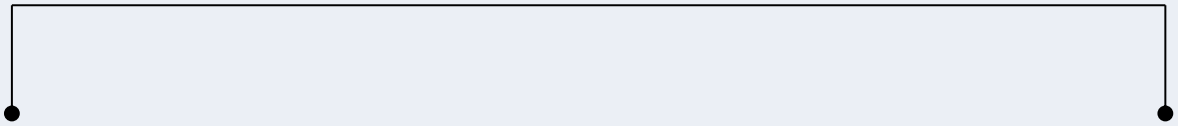
SQL (often pronounced "sequel") stands for Structured Query Language.

It is a powerful tool that enables programmers to create, populate, manipulate, and access databases. It also provides an easy method for dealing with server-side storage.

What is SQL

Data using SQL is stored in tables on the server, much like spreadsheets you would create in Microsoft Excel.

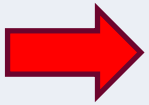
This makes the data easy to visualize and search.



Customer_ID	Date_ID
d005458dtsf	6/26/2019
d007sfgs847	8/3/2018
d004fgsfh445	12/3/2018

Order_ID	Customer_ID	Date_ID
10001	d005458dtsf	6/26/2019
10002	d007sfgs847	8/3/2018
10003	d004fgsfh445	12/3/2018

Different Types of Databases



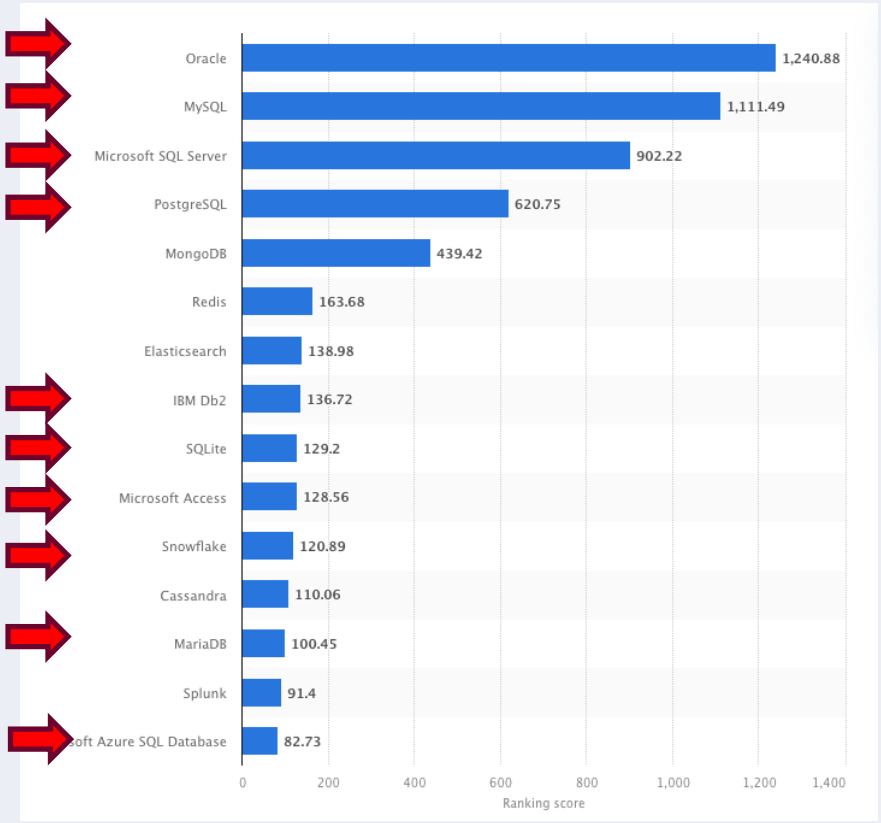
Relational	Key-Value	Document
Vector	Time Series	Graph
Wide Column	Search	Spatial

DB-Engines Ranking as of December 2023

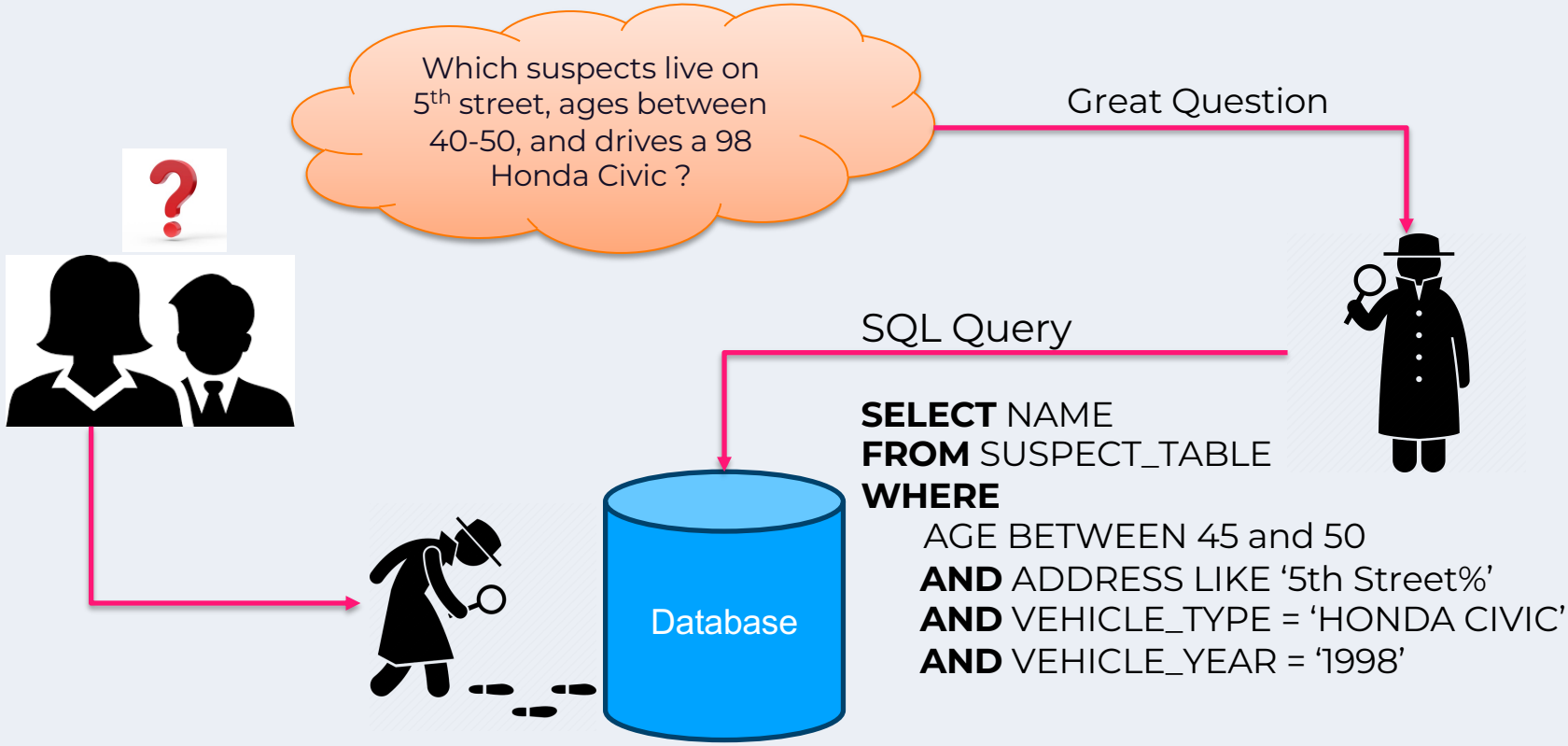
417 systems in ranking, December 2023

Rank			DBMS	Database Model	Score		
Dec 2023	Nov 2023	Dec 2022			Dec 2023	Nov 2023	Dec 2022
1.	1.	1.	Oracle	Relational, Multi-model	1257.41	-19.62	+7.10
2.	2.	2.	MySQL	Relational, Multi-model	1126.64	+11.40	-72.76
3.	3.	3.	Microsoft SQL Server	Relational, Multi-model	903.83	-7.59	-20.52
4.	4.	4.	PostgreSQL	Relational, Multi-model	650.90	+14.05	+32.93
5.	5.	5.	MongoDB	Document, Multi-model	419.15	-9.40	-50.18
6.	6.	6.	Redis	Key-value, Multi-model	158.35	-1.66	-24.22
7.	7.	↑ 8.	Elasticsearch	Search engine, Multi-model	137.75	-1.87	-7.18
8.	8.	↓ 7.	IBM Db2	Relational, Multi-model	134.60	-1.40	-12.02
9.	↑ 10.	9.	Microsoft Access	Relational	121.75	-2.74	-12.08
10.	↑ 11.	↑ 11.	Snowflake	Relational	119.88	-1.12	+5.11

Statisa –Worldwide Ranking of Databases as of September 2023



Using SQL for your Investigation



DATA SCIENCE

CRISP-DM

Cross-Industry Standard Process for Data Mining

- A structured approach to planning, implementing, and deploying data mining activities

What is Generative AI?

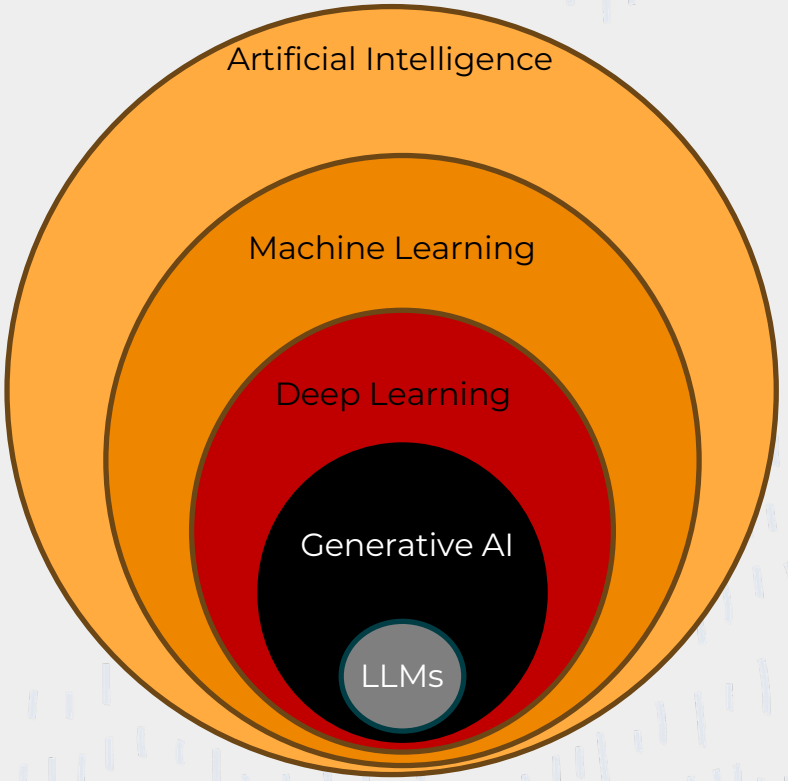
AI refers to the broad concept of machines or computers performing tasks that typically require human intelligence. This includes reasoning, learning, problem-solving, perception, language understanding, etc.

ML is a subset of AI focused on the idea that machines can learn from data, identify patterns, and make decisions with minimal human intervention

DL is a subset of ML that uses neural networks with many layers (deep networks) to model complex patterns in data.

Generative AI refers to a class of AI, often realized through DL, that focuses on generating new content or data that is similar to but distinct from the training data.

LLMs are a type of deep learning model designed to understand, generate, and interact with human language at a large scale. They are trained on vast amounts of text data.



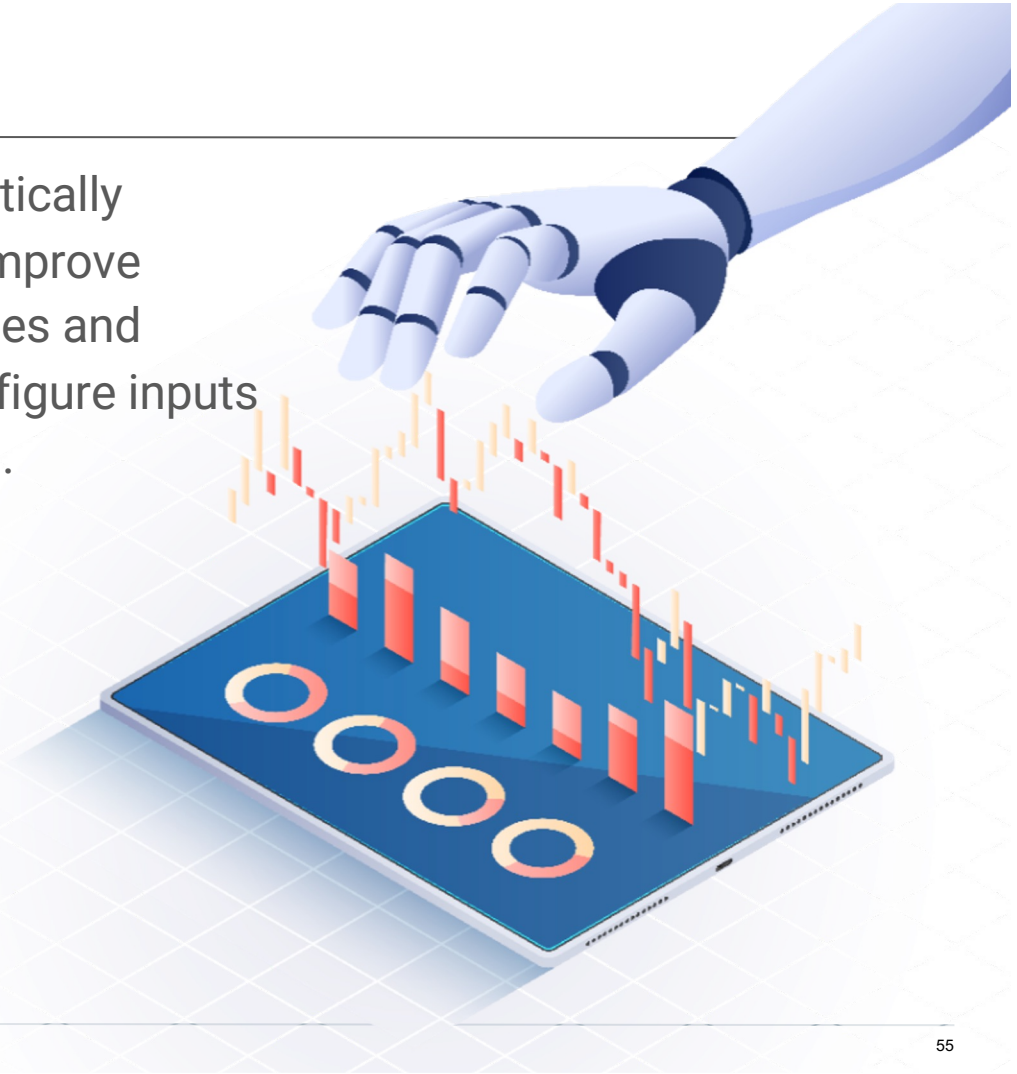
What do you know about machine learning?

Machine learning is a programming approach for designing applications that learn from their inputs and make adjustments based on their outputs.

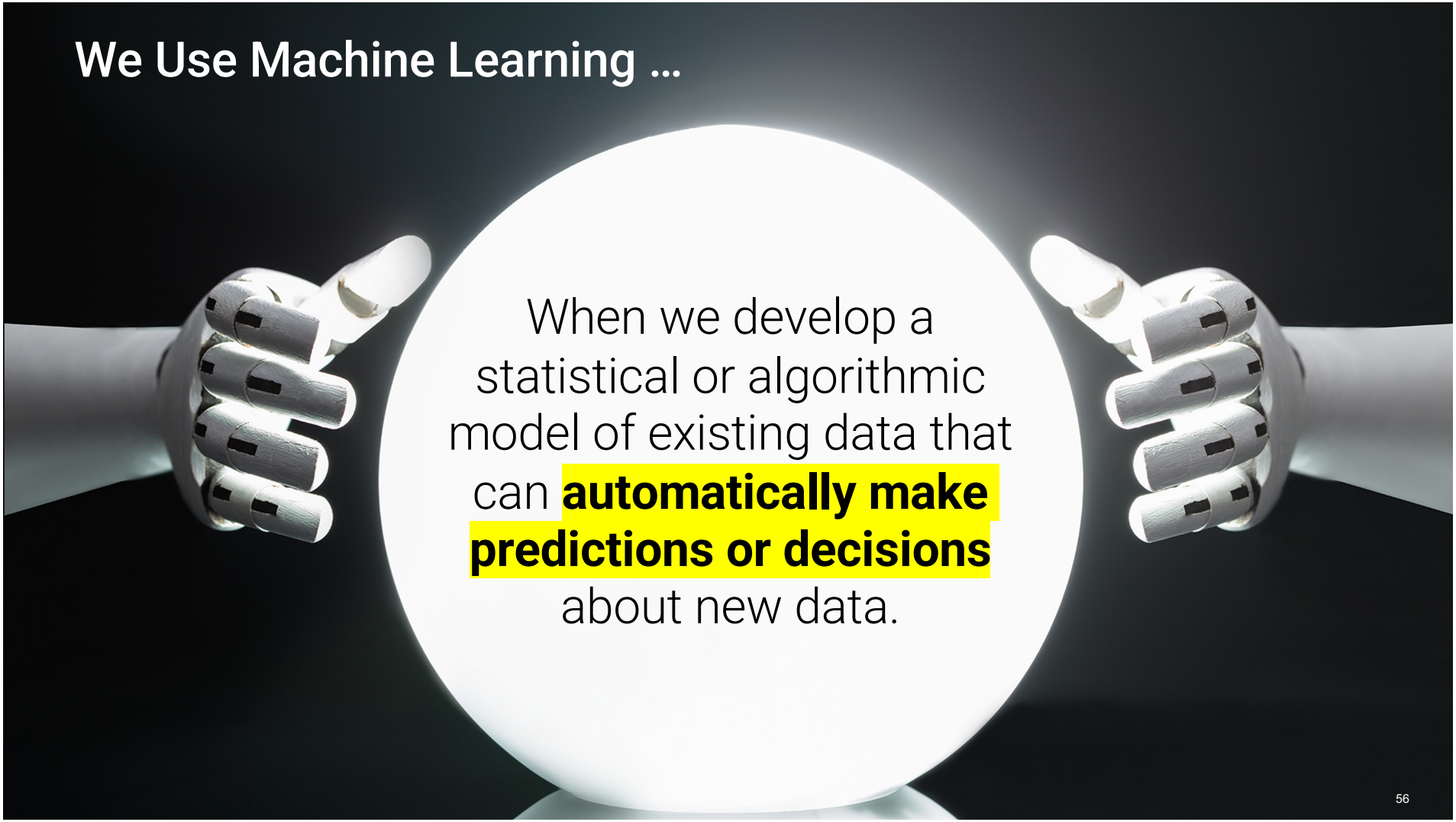
Machine Learning

A machine learning algorithm automatically adapts (automatic configuration) to improve the accuracy and precision of outcomes and predictions, so we do not need to configure inputs and manually change to the algorithm.

Because machine learning algorithms can learn on their own, developers do not need to worry about coding for every scenario.



We Use Machine Learning …



When we develop a statistical or algorithmic model of existing data that can **automatically make predictions or decisions** about new data.

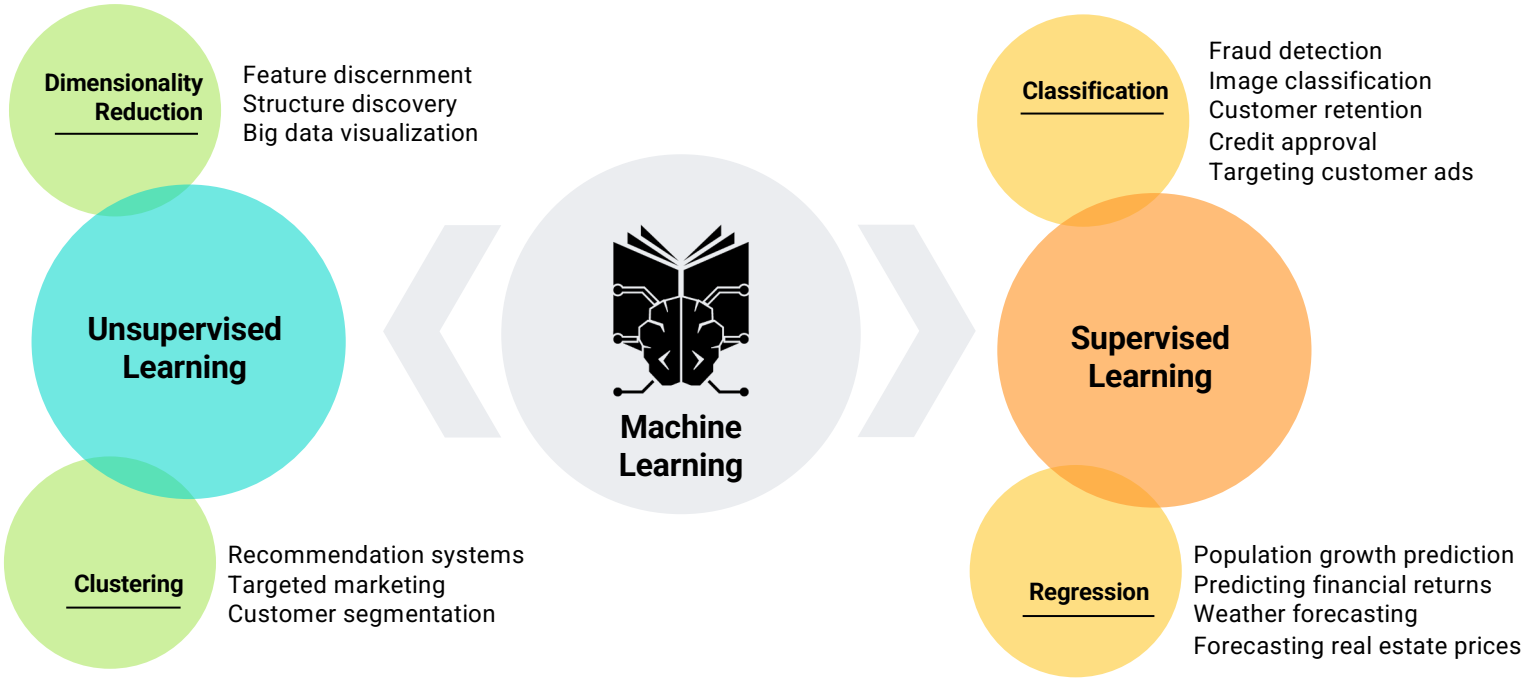
Machine Learning

All machine learning pipelines follow a **Model-Fit-Predict** paradigm where we use a dataset or data model to fit, or train, the algorithm.

Once the algorithm has been trained, we can use the model and the algorithm to make actual predictions.

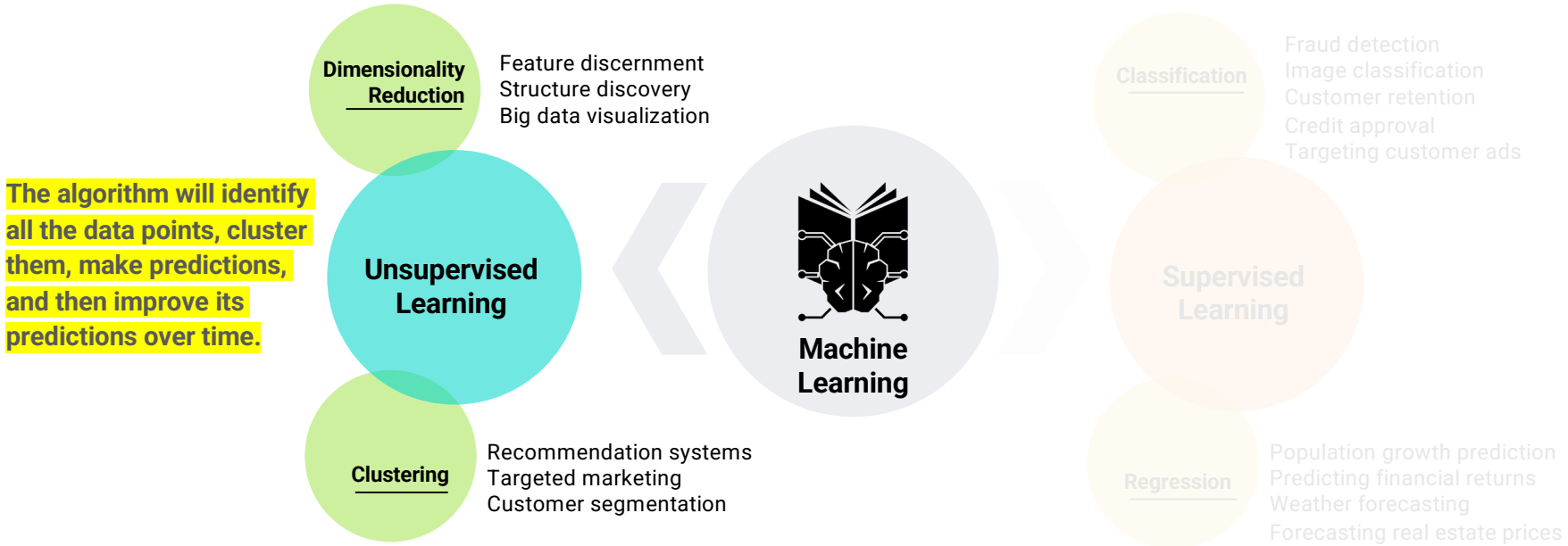
Machine Learning

We've learned that machine learning has **two main approaches**:



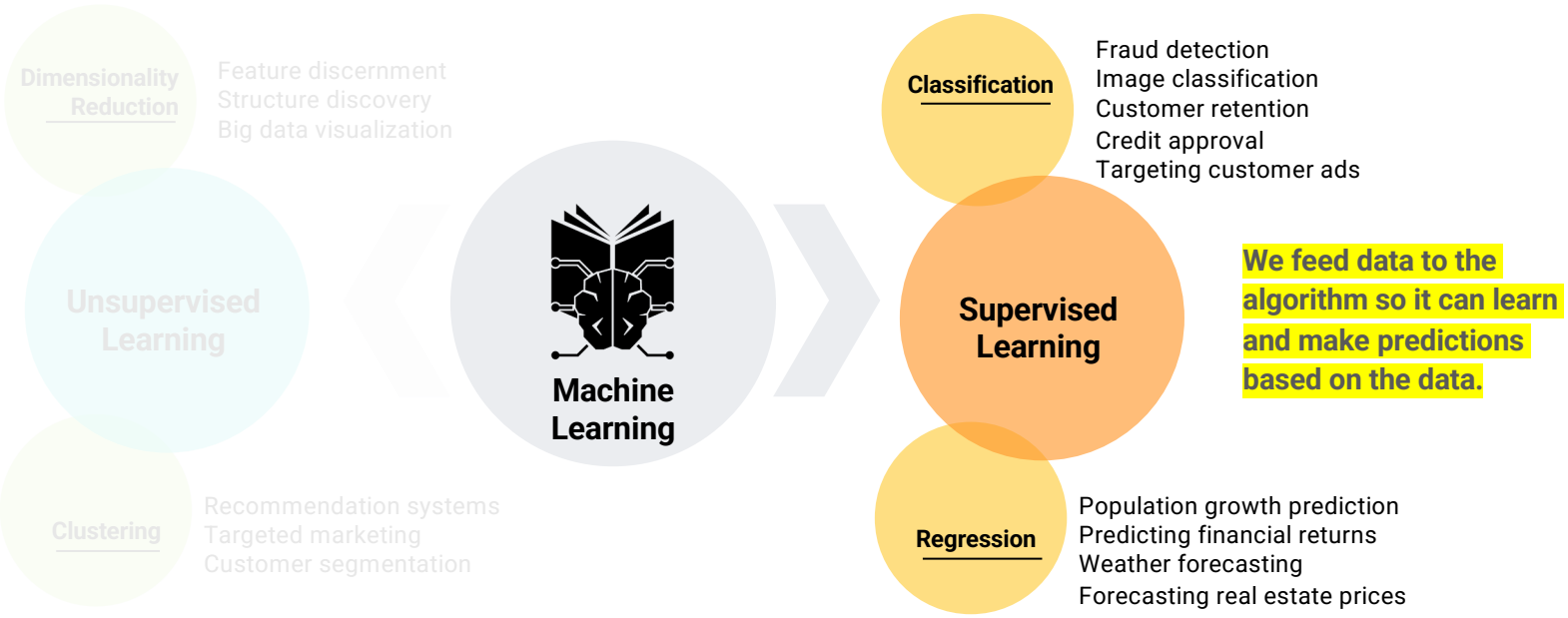
Machine Learning

We've already covered **unsupervised learning**, which is when an intelligent algorithm learns as it goes, without having observed any type of data before.



Machine Learning

This week, we'll cover **supervised learning**.



Types of ML

We can group all of these models into two main buckets:

01 Supervised learning

The algorithm learns on a **labeled dataset**, where each example in the dataset is tagged with the answer.

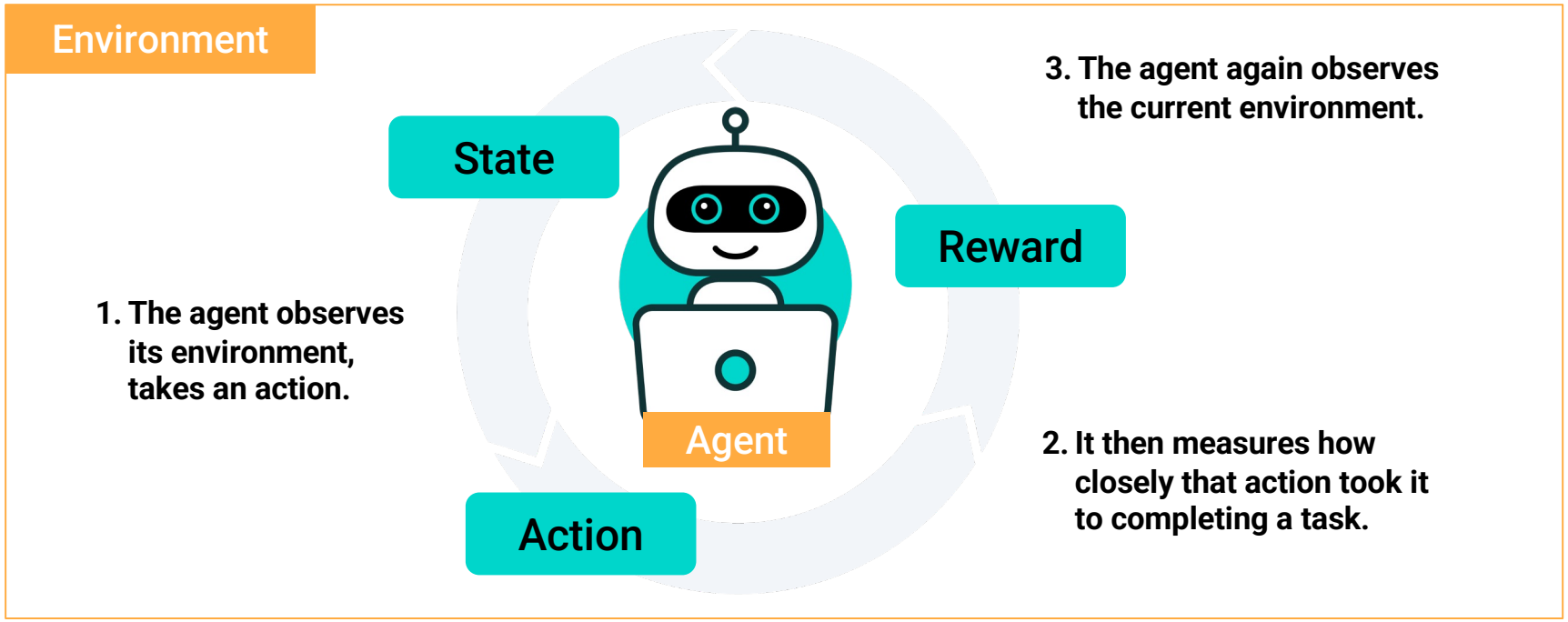
This provides an answer key that can be used to evaluate the accuracy of the training data.

02 Unsupervised learning

The algorithm tries to make sense of an **unlabeled dataset** by extracting features and patterns on its own.

Reinforcement Learning

This third type of machine learning algorithm is used less frequently but still has important applications in finance.



Three Types of ML