Advanced Data Analytics

Agenda

1. Data (Raw vs Traditional vs Big)
2. Data Science
3. Data vs Business vs BI
4. AI/ML/DL

Data

Raw Data

- may called **primary data**
- cannot be analysed straight away
- untouched, accumulated, and stored

- can be collected in a number of ways:
  - **manuale** (as surveys: how much you like a product on a scale of 1-10)
  - **automatic** (as cookies)

- **data preprocessing** needs to be performed on **raw data** to obtain meaningful info
- There are operations that can convert raw data into a more understandable format

Traditional Data & Big Data

- stored in a **digital** format

- can be used as a **base** for performing **analysis** and decision making

- can be divided into:
  - **Numerical**: easily manipulated (e.g. added) that gives us useful info
  - **Categorical**: hold no numerical value

Raw Data and Relativity

DIKW Pyramid: Data, Information, Knowledge, Wisdom

- illustrates the process of transforming **raw data** into **actionable insights**
- Flow starts from bottom to up with increasing values such as:
  - **hindsight** (looking back)
  - **foresight** (looking forward)
  - **insight** (looking broad and deep)

- Data is collected, processed, and analyzed to generate **information**,
  - **information** is then used to develop **knowledge** and **wisdom**
- Using **Analytics techniques**, organizations can extract valuable **insights** from their data and use them to improve their operations, products, and services.

Data

- Dealing with data is the **first step**
  - when **solving** a business **problem** or researching,
  - so it is important to know what you are looking at

Data (Traditional & Big)

- Either Data or Big Data
  - it is your first port of call for business problem-solving,
  - so it is important to know what you are dealing with

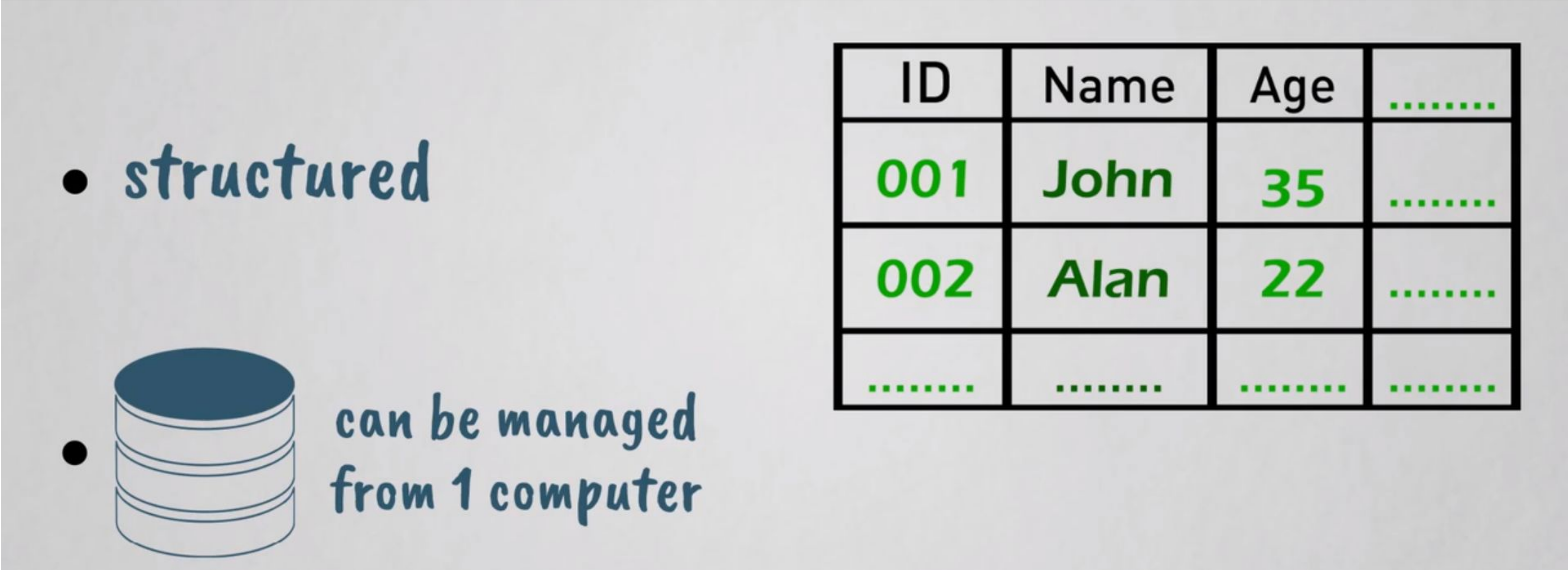
365 DataScience Infographic Columns

Each describes a stage of solving business task process

1. Working with Traditional Data
2. Working with Big Data
3. Doing Business Intelligence (BI)
4. Applying Traditional Data Science Techniques
5. Using Machine Learning (ML) Techniques

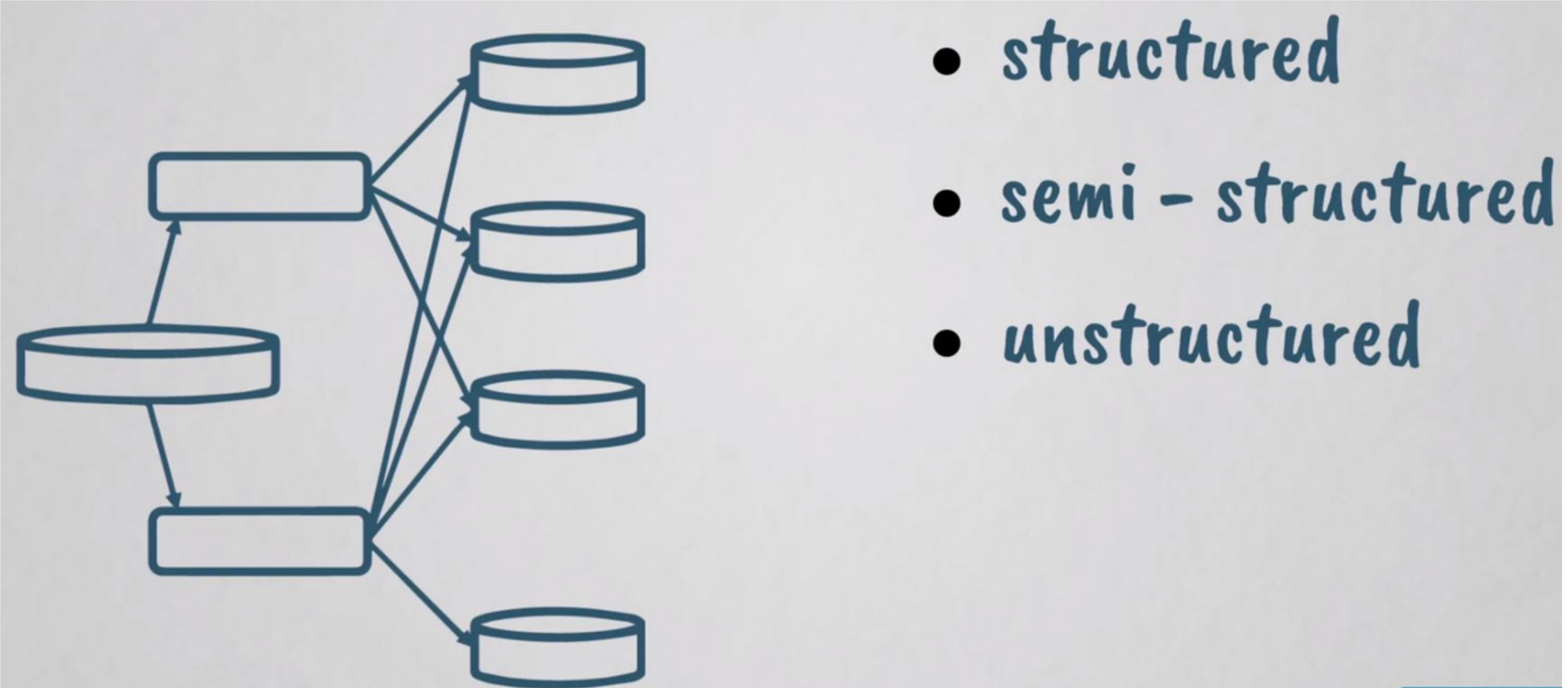
Traditional Data

- is **structured** and stored in **databases**
- in the form of tables containing **numeric** or **text** values
- can be managed from **one computer**



Big Data - 1

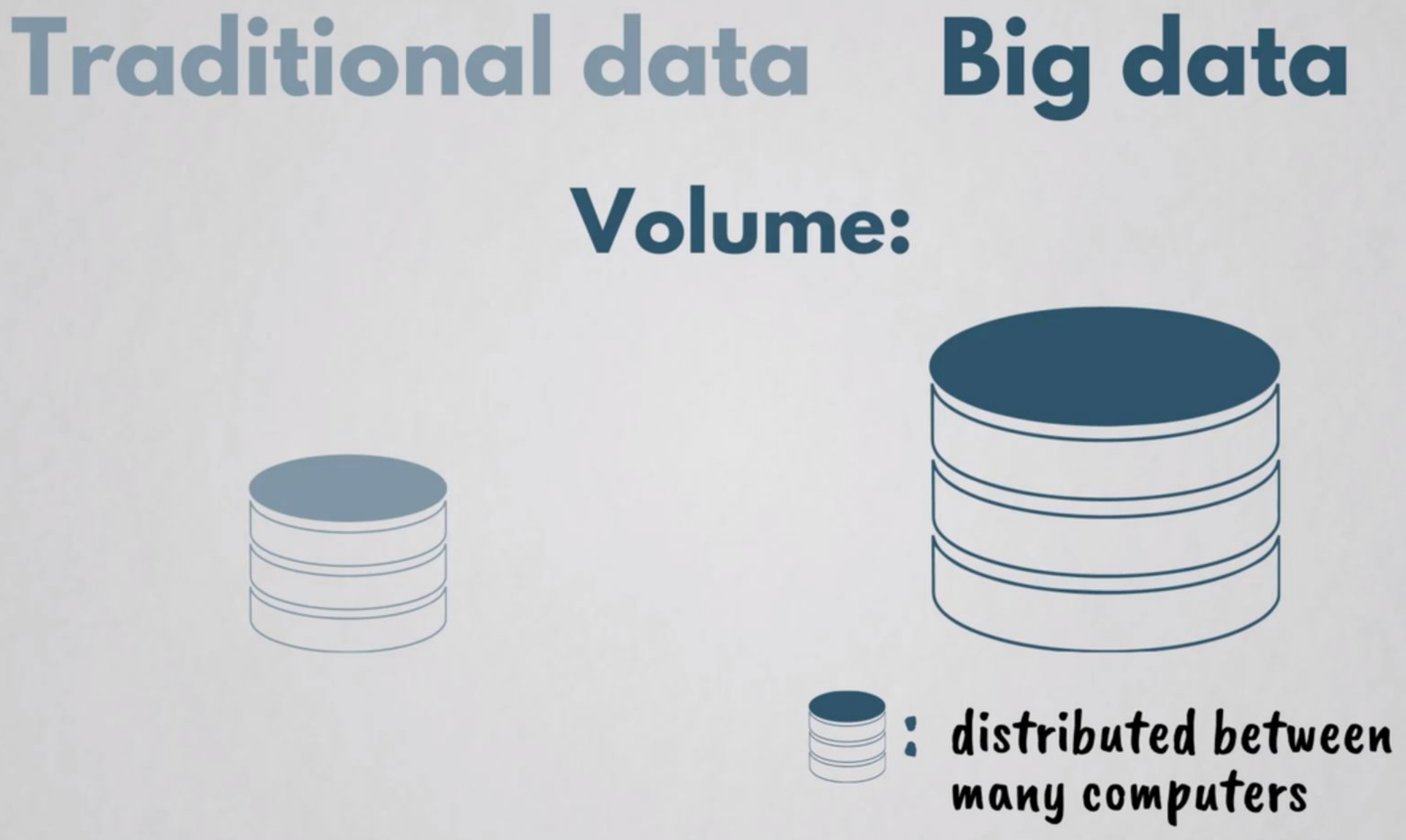
- a term reserved for extremely **large** data
  - not just humongous in terms of **volume**
- could be in various format (its **variety**):
  - **structured**: SQL databases
  - **semi-structured**: NoSQL databases, HTML, XML, JSON, . . .
  - **unstructured**: images, audio, video, . . .



Traditional Data vs Big Data - Volume

Big Data

- needs a whopping amount of memory space,
- typically distributed between many computers
- Its size is measured in TB, PB, EB



Traditional Data vs Big Data - Variety

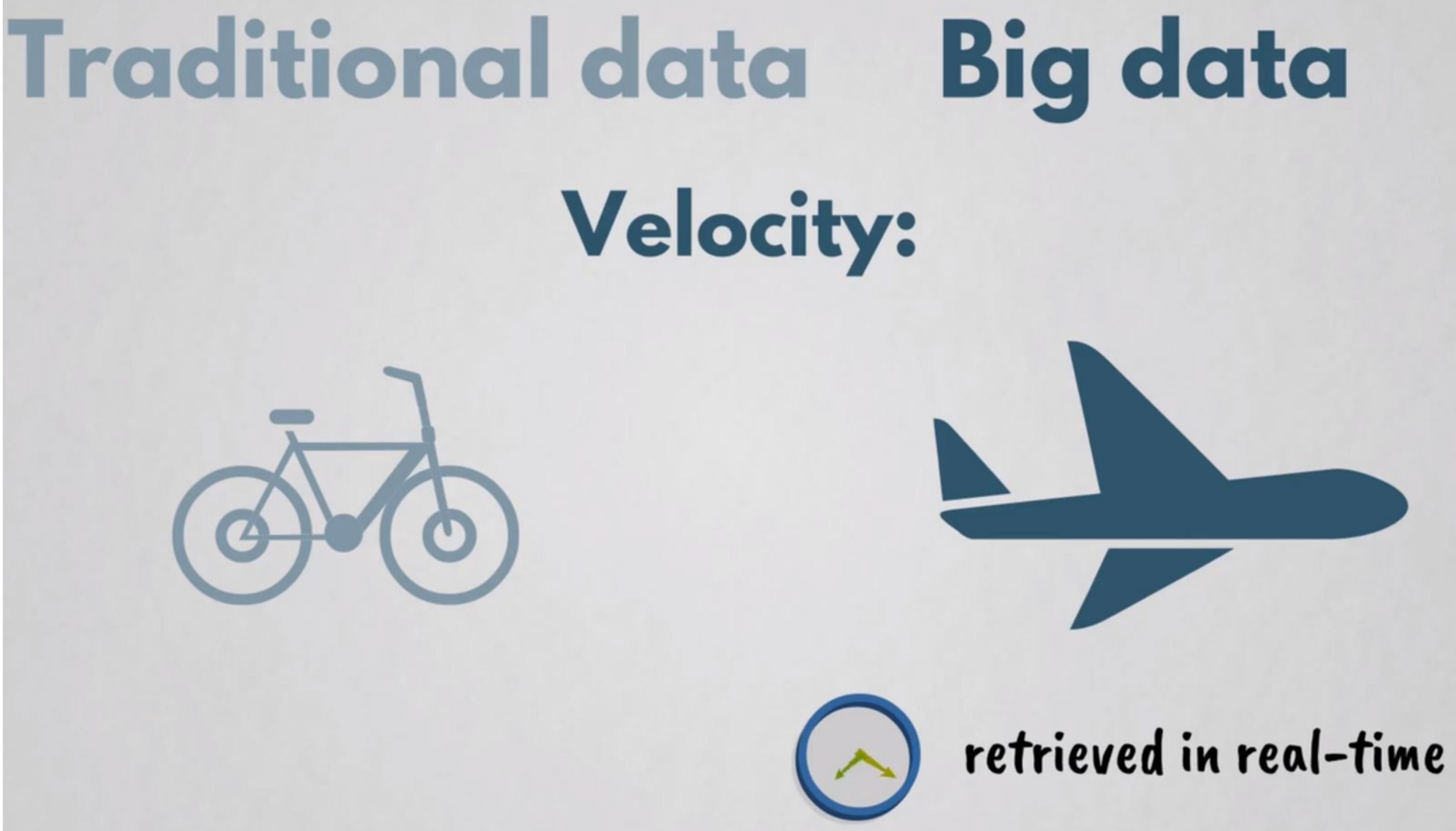
Big Data

- not just numbers and text
- implies dealing with images, audio, video, files, and others

Traditional Data vs Big Data - Velocity

Big Data

- One goal is to make extracting **patterns** from **Big Data** as quickly as possible
  - the progress that has been made in this area is remarkable
- Outputs from huge datasets can be retrieved in real-time
  - this means they can be extracted so quickly,
  - so results could be computed immediately after source data has been obtained



Big Data - 2

- is often characterized with the letter ‘**V**’
- Under different frameworks we may have **3,5,7,11** and even more Vs of **Big Data**

- The main **Vs**:
  - **volume**: amount of data
  - **variety**: number of data types
  - **velocity**: speed of data
- Some other **Vs**:
  - **visualization** process of representing abstract
  - **value** represents the business value to be derived from big data
  - **veracity** (quality) analysis of data is virtually worthless if it’s not accurate
  - **variability** how efficiently it differentiates between noisy or important data

Big Data Note

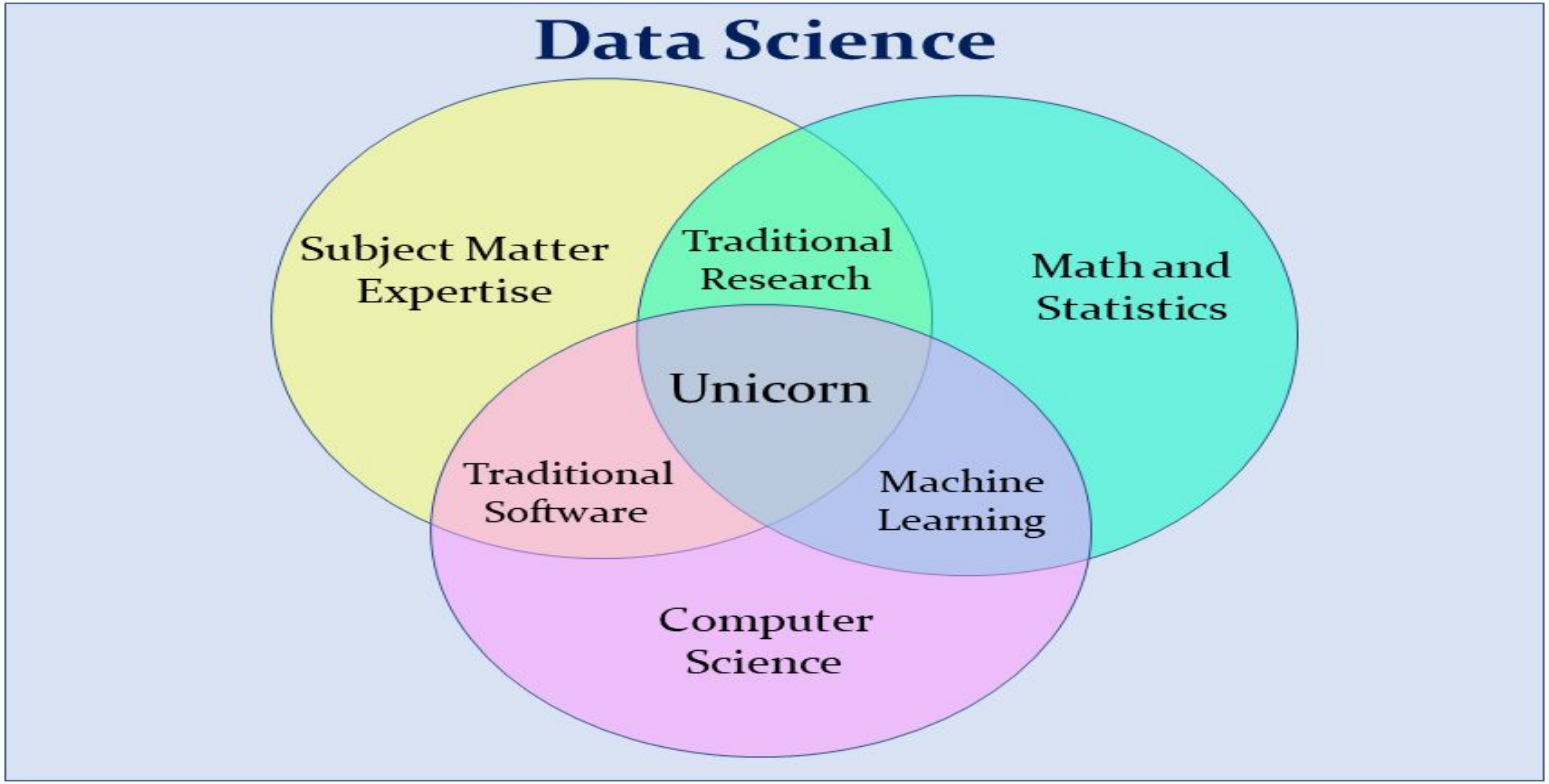
- not just **volume** that defines a data set as ‘big’
- other **Vs** (characteristics) play an important role as well

Data Science

Data Science - 1

- After gathering and organizing all data,
  - it is time to get your hand dirty with **analytics**

Data Science Venn Diagram 2.0



Data Management Evolution (1)

1) Mainframe-based **Hierarchical Databases** became available in the **1960s**
   - bringing more formality to the process of managing data

2) **Relational Databases** emerged in the **1970s**
   - cemented its place at the center of the data management ecosystem during 1980s

3) **Data Warehouses** conceived late in **1980s**
   - early adopters of the concept began deploying data warehouses in the mid-1990s

Data Management Evolution (2)

4) **Hadoop** became available in **2006**
   - followed by **Spark** processing engine and various other big data technologies

5) **NoSQL Databases** started to become available in the same time frame

6) **Data Lakes**
   - have given organizations a broader set of data management choices

7) **Data Lakehouse** concept in **2017** further expanded the options
   - enforce a predefined schema and have data transformation capabilities,
   - allowing semi-structured and unstructured data to be standardized before storage

Why do we stream data?

- Get real-time info in a dashboard or another means
- Make decisions in real-time

- If data comes in late, it's no longer valuable
  - especially during an emergency

- **Example** in cybersecurity: New York City Cyber Command
  - 5 or 6 TB on weekdays during peak times & 2 or 3 TB on weekends
  - **Security Analysts** can access the data in several ways:
    - run queries in the data warehouse
    - use other tools that will provide **visualizations** of the data
    - Check for example: **Splunk Certified Cybersecurity Defense Analyst**

Data Science - 2

The infographic divides Data Science into three segments:

- BI
- Traditional methods
- ML methods

The infographic divides Data into two segments:

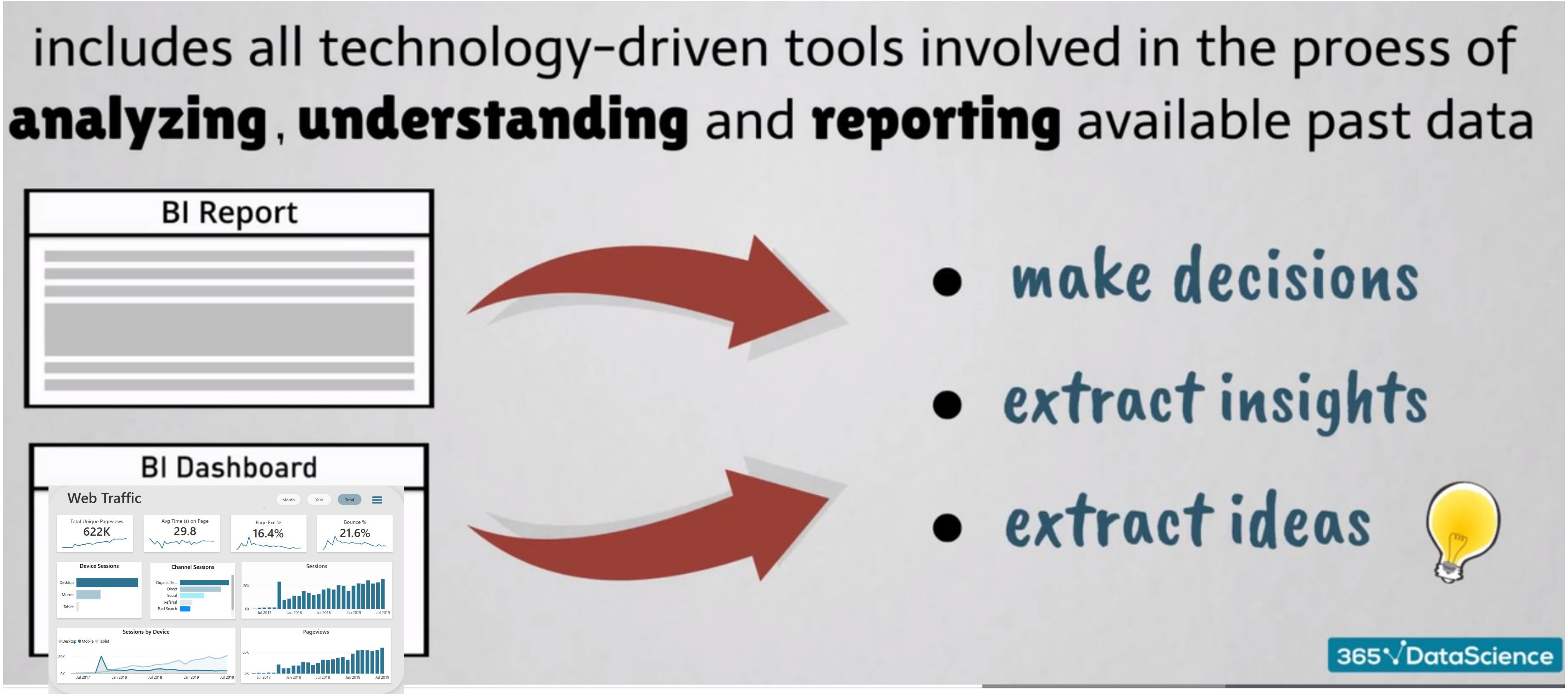
- Traditional Data
- Big Data

Business Intelligence - 1

1st step of applying data science

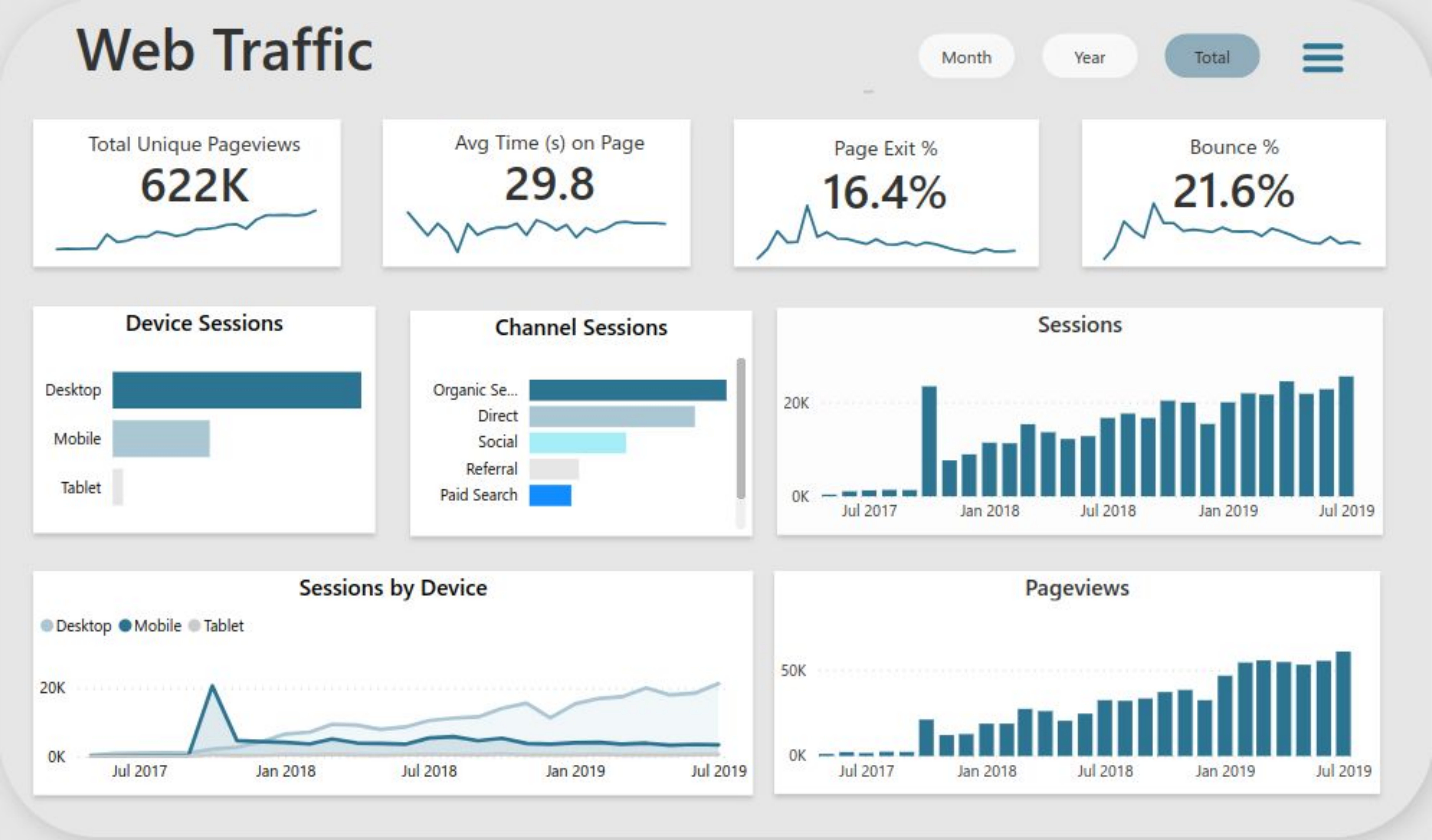
- is to **analyse** the past data that we have acquired
- BI is the discipline we need for this

Business Intelligence - 2



- This would result in having reports or dashboards,
  - which will help in making **strategic** and **tactical business decisions**

Example of BI Dashboard



BI Questions - 1

BI Questions - 2

- **Example**: BI means understanding **how** your sales grew and **why**:
  - Did competitors lose market share?
  - Was there an increase in the price of your products?
  - Did you sell a mix of more expensive products?
  - Were there more profitable client accounts?
  - How did your **profitability margins** behave in the same time frame of the previous year?

- BI is all about:
  - **understanding past** business performance to improve future performance

BI answers questions like

- What happened?
- When did it happen?
- How many units did we sell?
- In which region did we sell the most goods?

BI

- **BI** requires the combination of
    - data skills
    - business knowledge
  - to explain the past performance of the company

BI: Real-life Example

- BI allows you to adjust your strategy to past data as soon as it is available.

- If done right,
  - BI will help to efficiently manage your **shipment logistics** and, in turn,
  - reduce costs and increase profit.

After BI 1

- BI is **worth the time** in the total process
- BI **extracts insights** and ideas about business that will
  - help to **grow**
  - give an **edge over competitors**, giving added stability

- We want to forecast future sales and profitability, as well as expenses
- Once **BI reports and dashboards** are complete and presented, it is time to apply:
  - **Traditional** Methods (Traditional Data Science)
  - or **ML** Methods
- ■ to develop an idea of what will happen

After BI 2

- After the **dashboard** is ready, the **Data Science team** will use:
    - **business analytics** or
    - **data analytics**
  - to develop models that could **predict** future outcomes

- **BI** can be seen as the **preliminary** step of **Predictive Analytics**

Predictive Analytics

There are two branches of **Predictive Analytics**

1. **Traditional Methods:** classical statistical methods for forecasting

2. **ML Methods**

Traditional Methods

Traditional Methods

- relate to **Traditional Data**
- designed **prior to** the existence of **Big Data**,
  - where the technology simply wasn't as advanced as it is today
- involve applying **statistical** approaches to create **Predictive Models**
- perfect for **forecasting** future performance with great accuracy
- there is no denying that these tools are absolutely **applicable today**

Traditional Methods: Real-life Examples

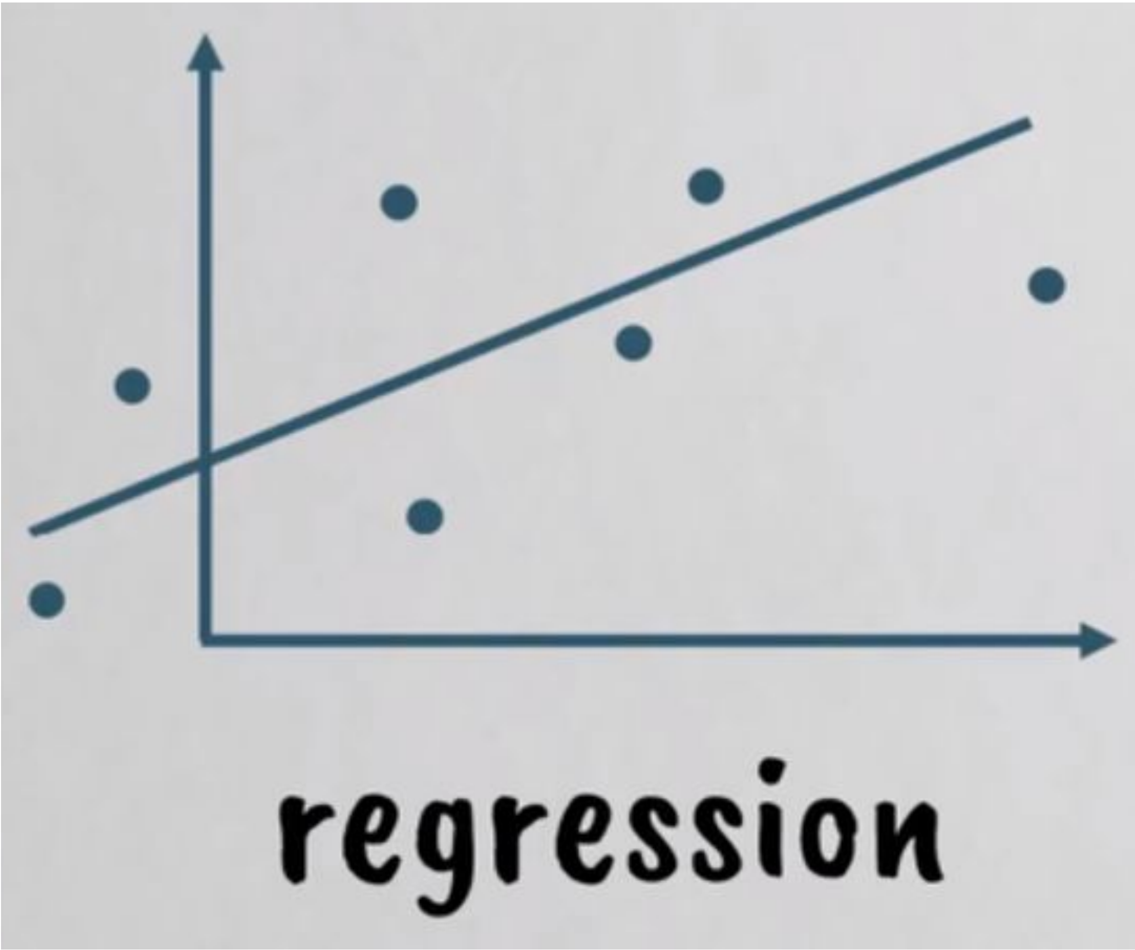
- The application of traditional methods is extremely broad

- Example 1: **Forecasting sales data**
  - using **time series data** to predict a firm’s future expected sales

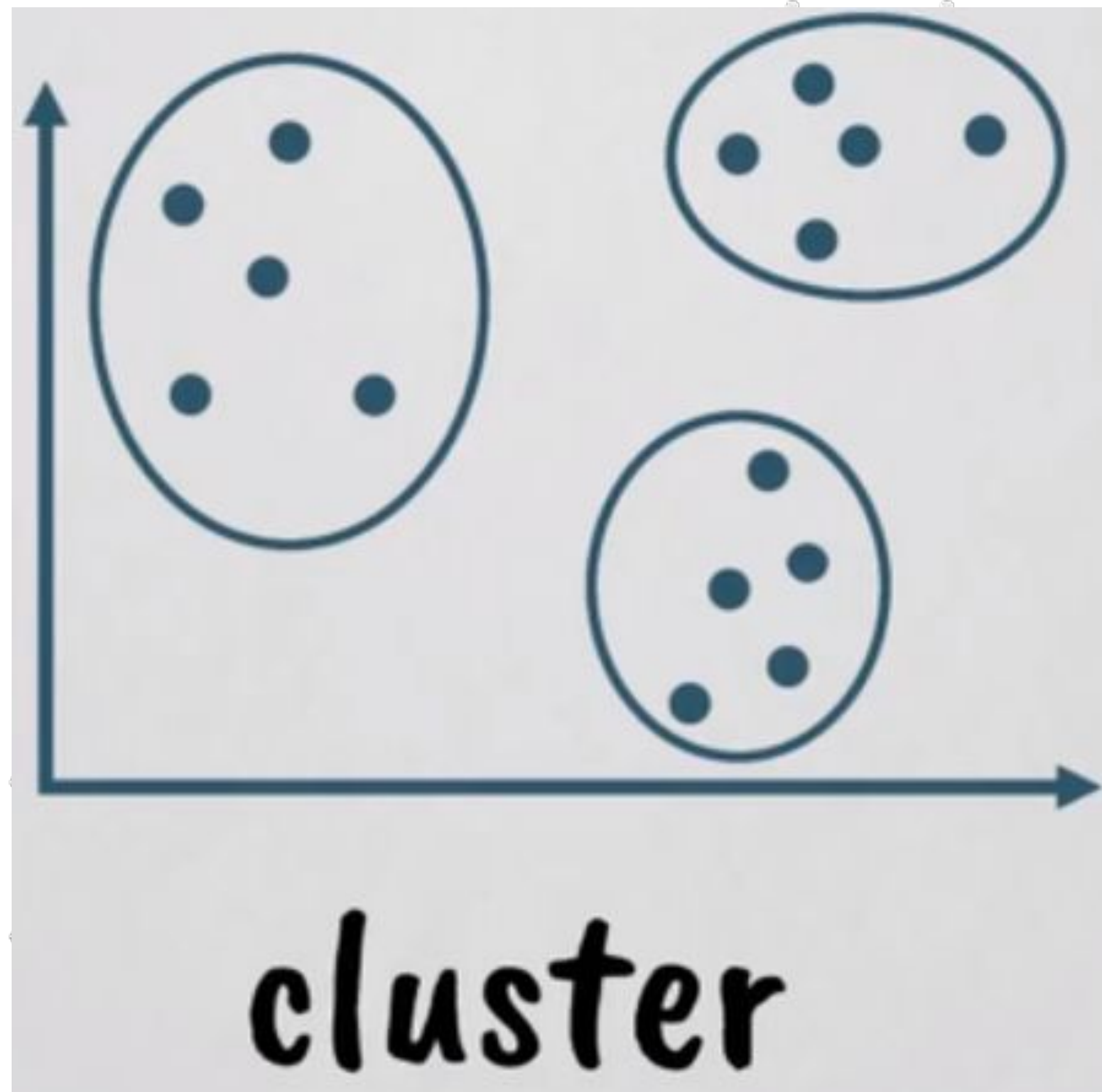
- Example 2: **UX**
  - plot customer satisfaction and customer revenue
  - to find that each **cluster** represents a different geographical location

Traditional Methods: Techniques

- **Regression**: model used for quantifying causal relationships,
  - among different variables included in the analysis



- **Clustering**: grouping the data to analyze meaningful patterns



Do we need AI/ML? (1)

- A lot of **Data Analytics** is backward-looking, nothing wrong with that
  - But instead we're going to use **ML** to generate forward-looking or predictive insights

- Of course, the point of looking at historical data might be to make those decisions
  - Perhaps a **Business Analysts** examine data and they suggest new policies or rules
  - **Ex**: They could suggest that it's possible to raise price of a product in a certain region
  - Now, that **Business Analyst** is making a predictive insight but is that scalable?
  - Can **Business Analysts** make such a decision for every product in every single region?
  - And can they dynamically adjust the price every period?

Do we need AI/ML? (2)

- Here's where the computers get involved:
  - In order to make decisions around predictive insights repeatable, we need ML
  - We need a computer program to derive those insights
  - So **ML** is about making predictive insights from data, many of them at a time

Data
vs Business
vs BI

Business Case Studies

- are real-world experiences of how business people and companies succeed or fail
- examine events that have already happened

- We do not need a dataset to learn from **business cases**
- We could learn from **them** and attempt to prevent making a similar mistake in future

Business Case Studies: Example

Ryanair: cost-friendly budget airline that sells tickets so cheap

- travel to less busy airports
    - far from the city
    - outside business hours
    - tries to keep planes for small times on airfields to save on rent
- charges for almost every small addition
- operates only one type of aircraft to speed out ground crew processes

Data Analysts vs Business Analysts vs BI Analysts

Data Analyst common tasks:

- Identifying and **sourcing** data
- **Cleaning** and preparing data for analysis
- Analyzing data for **patterns** and trends
- **Visualizing data** to make it easier to understand
- Presenting data in such a way that it tells a compelling **story**
- Working with **stakeholders** to define a problem or business need

Data Analysts vs Business Analysts vs BI Analysts

Business Analyst common tasks:

- Training and coaching staff in new systems
- Reviewing processes to identify areas for improvement
- Evaluating a company's current functions and IT structures
- Interviewing team members to identify areas for improvement
- **Creating visuals** and financial models to support business decisions
- Presenting recommendations/findings to management and other key **stakeholders**

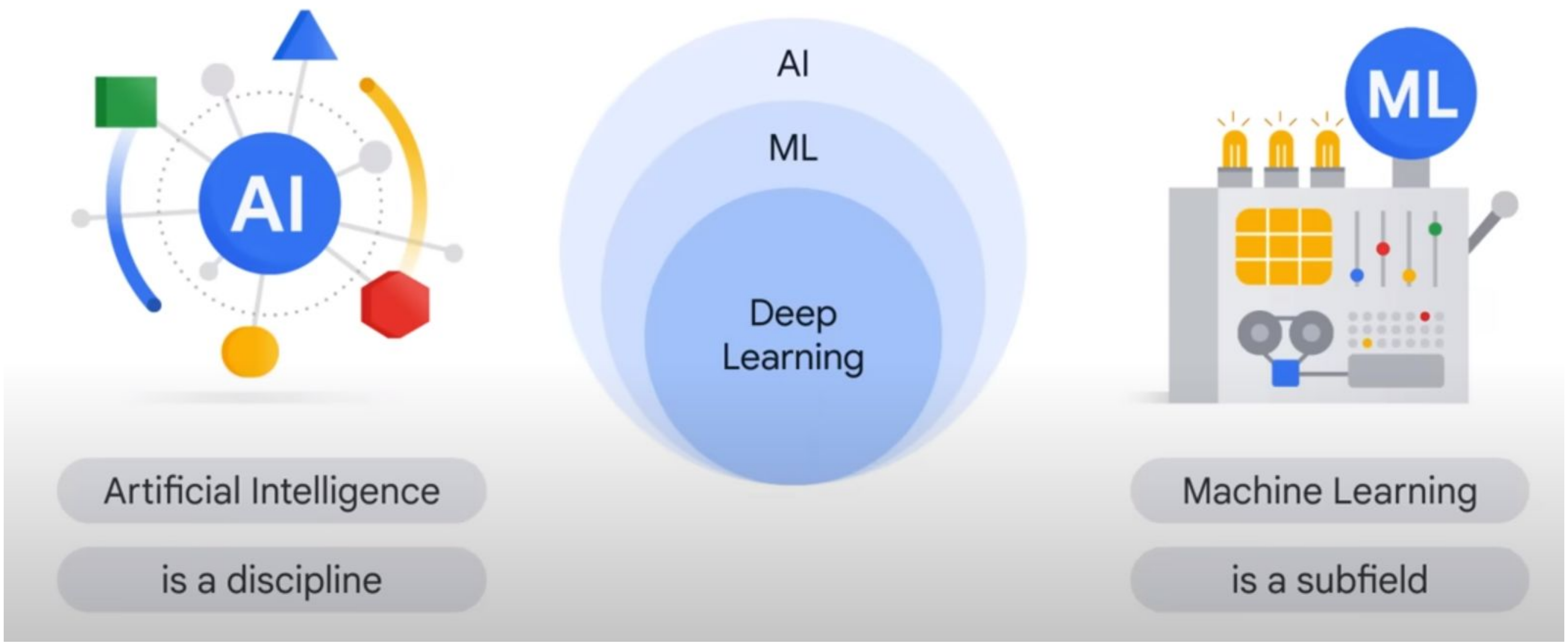
Data Analysts vs Business Analysts vs BI Analysts

BI Analyst:

- Somewhat of a hybrid between **Business Analyst** and **Data Analysts**
- The job of a **BI Analyst** requires to:
  1. understand the essence of a business
  2. strengthen that business through the power of data
- They use **analysis**, modeling, and **visualization** of:
  1. industry trends
  2. competitive landscape
    - to help businesses cut losses and increase profits

AI/ML/DL

AI is the theory and development of computer systems able to perform tasks normally requiring human intelligence.



According to the Turing test: **AI** is the machine's ability to exhibit intelligence behavior indistinguishable from that of a human

Artificial Intelligence (AI)

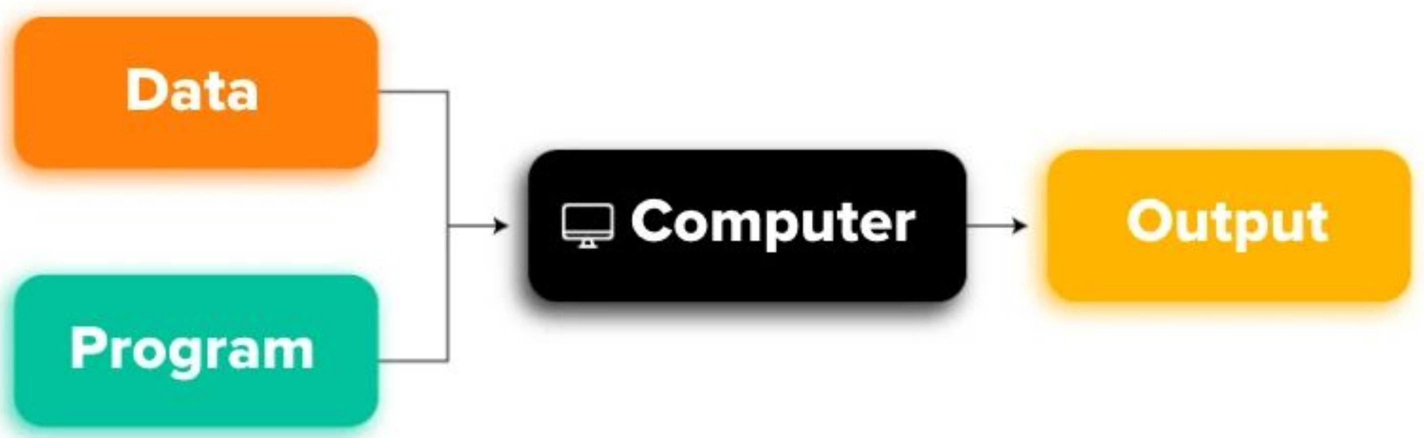
- **AI** is a pretty general term that can have a somewhat philosophical interpretation

- We, as humans, have only managed to reach **AI** through **ML**

- **Data Scientists** are interested in:
  - how tools from **ML** can help improve the accuracy of estimations

ML gives computers the ability to learn without explicit programming.

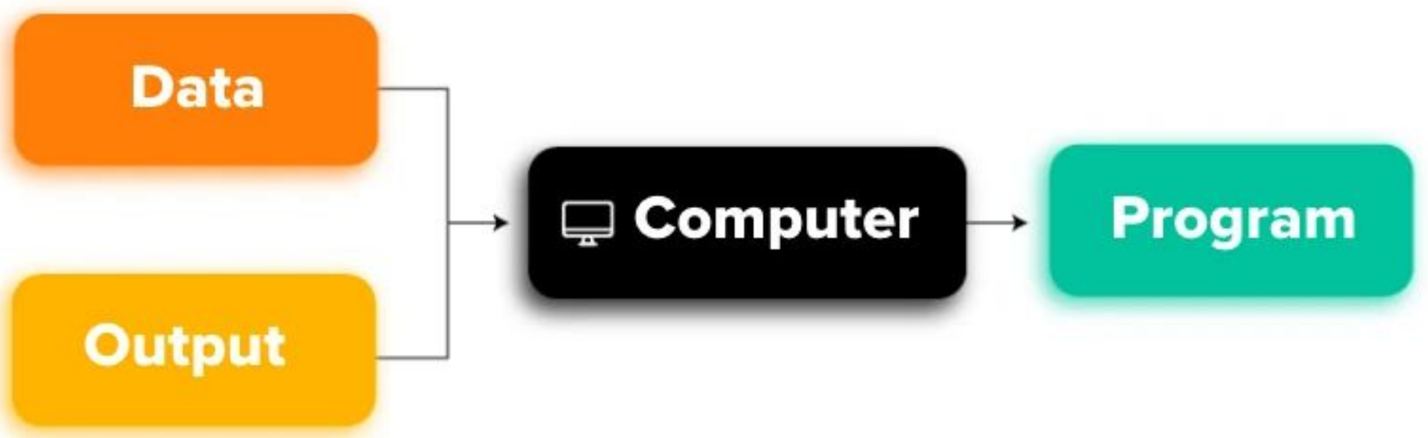
TRADITIONAL PROGRAMMING



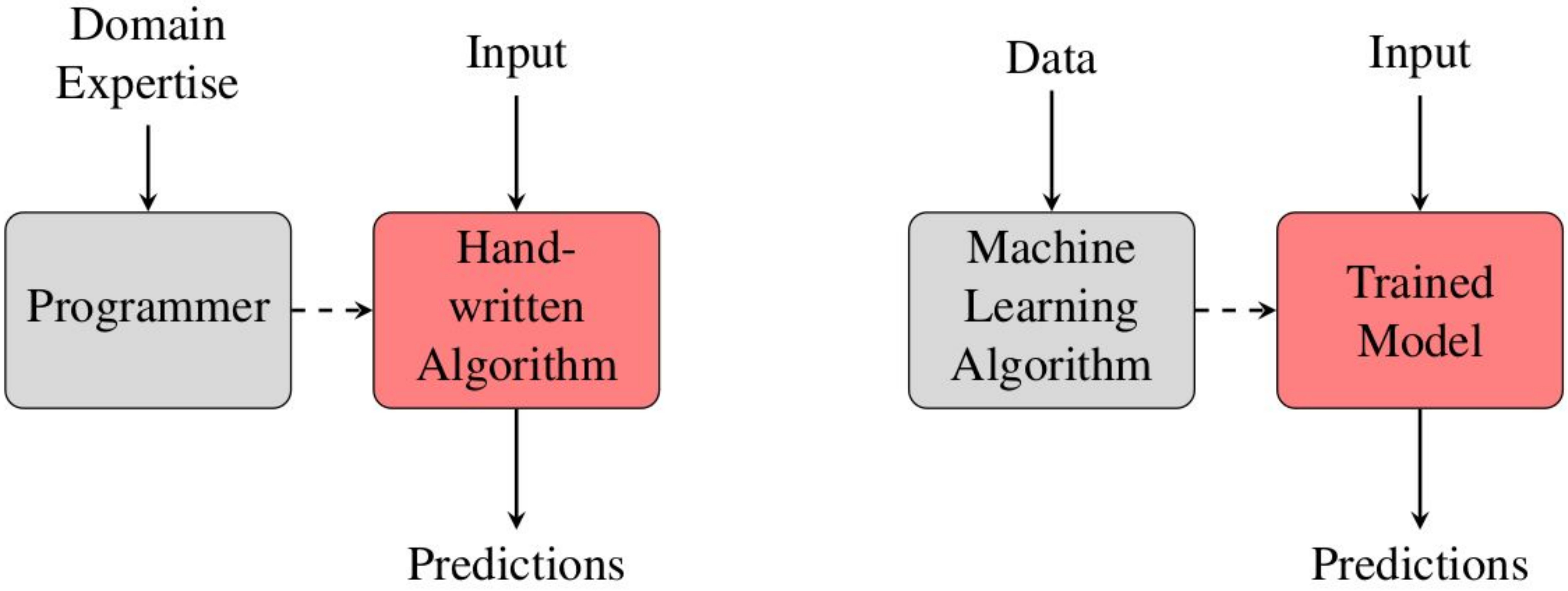
- In **ML**, the responsibility is left to the machine

- **ML** is all about creating/implementing algorithms that let machines
  - receive **data**,
  - perform **calculations**, and
  - apply **statistical** analysis
- to
  - make **predictions**
  - analyze **pattern**
  - give **recommendations**
- with unprecedented accuracy

MACHINE LEARNING



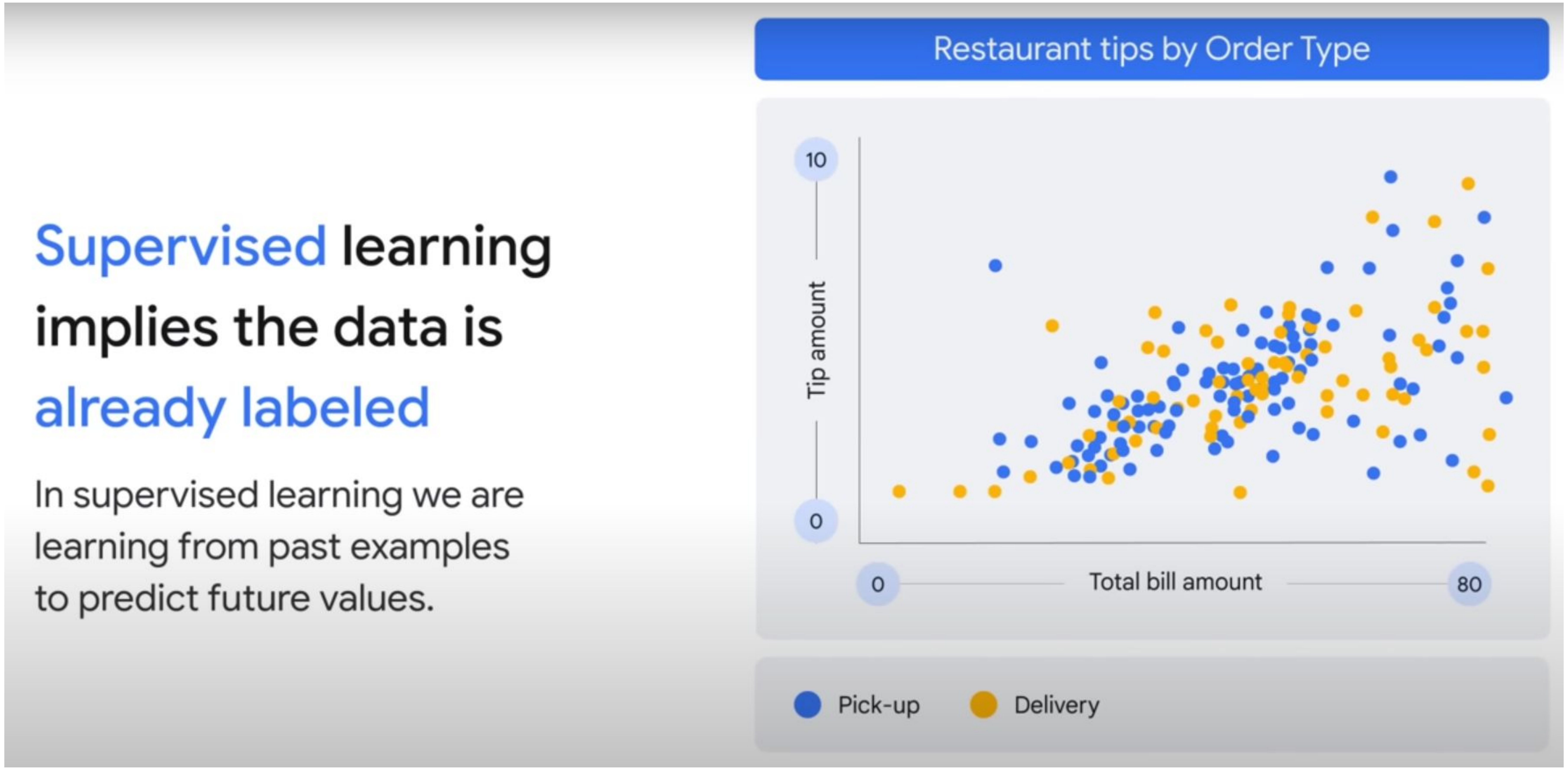
ML



(a) Traditional method

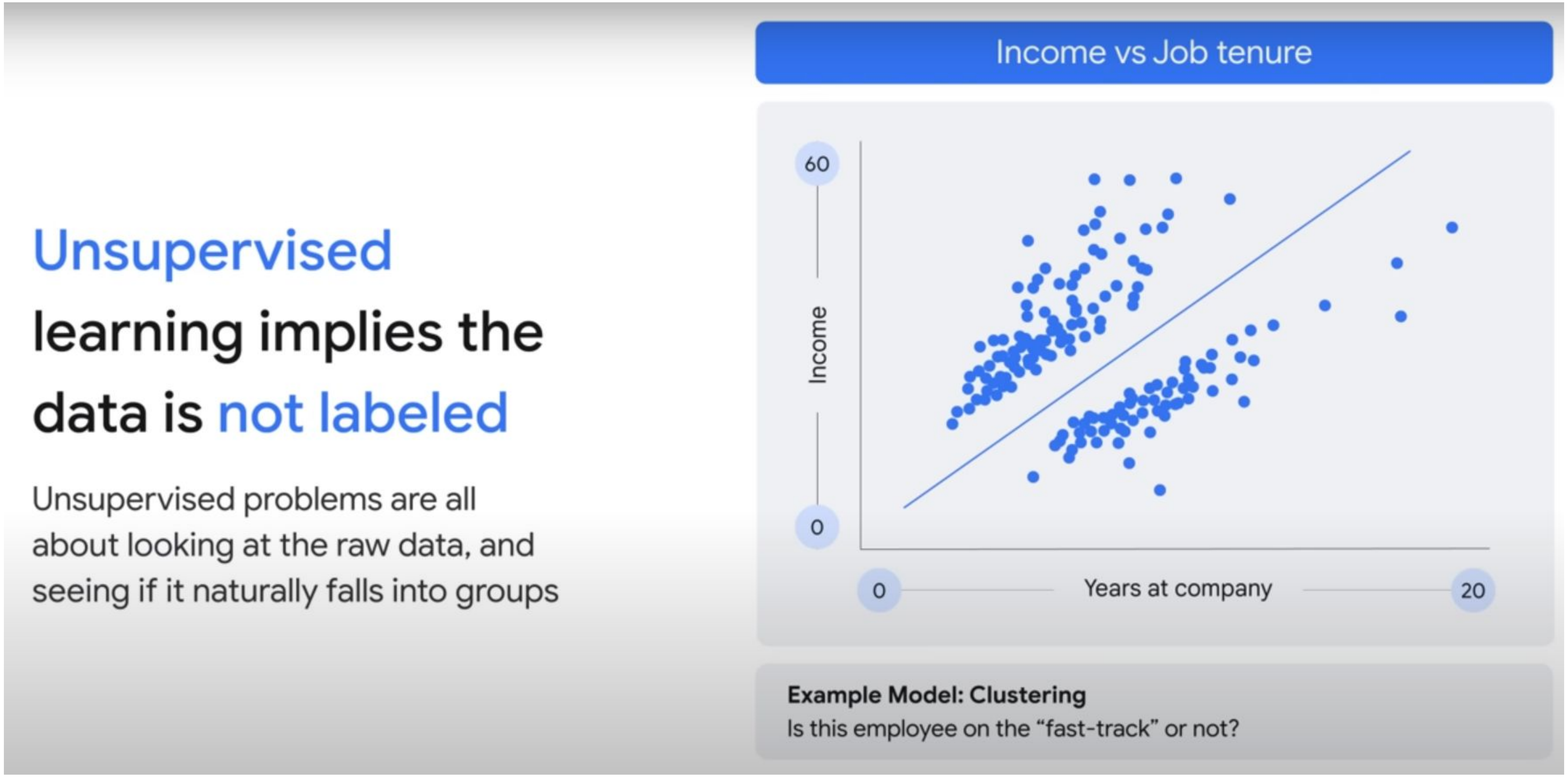
(b) Machine learning pipeline

ML Types: Supervised Learning



- Training an algorithm resembles a teacher supervising her students
- Provides feedback every step of the way
- We use labelled data

ML Types: Unsupervised Learning



- Algorithm trains itself, No teacher to provide feedback
- Algorithm uses unlabelled data

ML Types: Reinforcement learning

- A reward system is introduced

- Every time a student does a task better than it used to in the past
    - they will receive a reward
    - and nothing if the task is not performed better

- Instead of minimizing an error,
    - we maximize a reward

ML: Main Types

Supervised models

Task-driven and identify a **goal**

Classify data

Is an email spam?

Logistic regression

Predict a number

Shoe sales for the next three months

Linear regression

Unsupervised models

Data-driven and identify a **pattern**

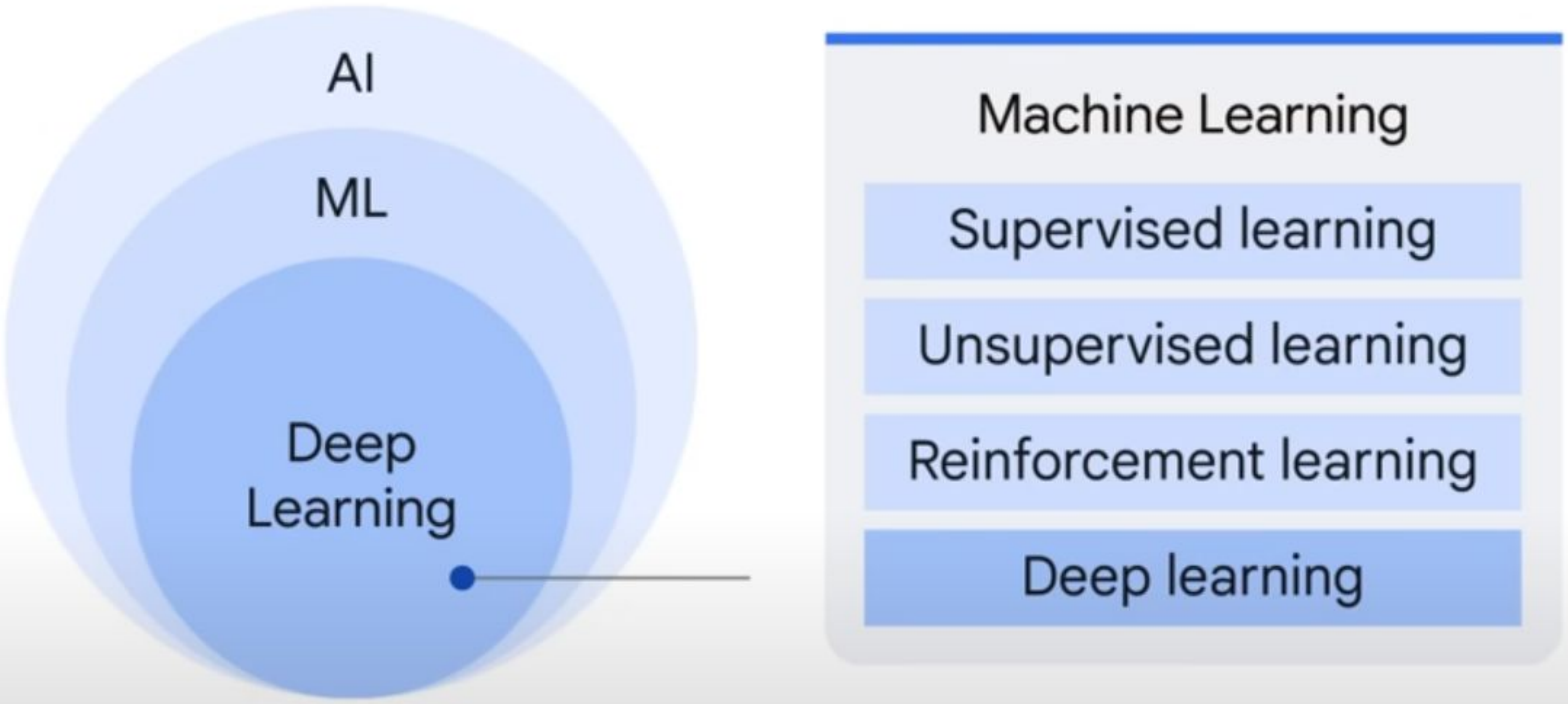
Identify patterns and clusters

Grouping photos

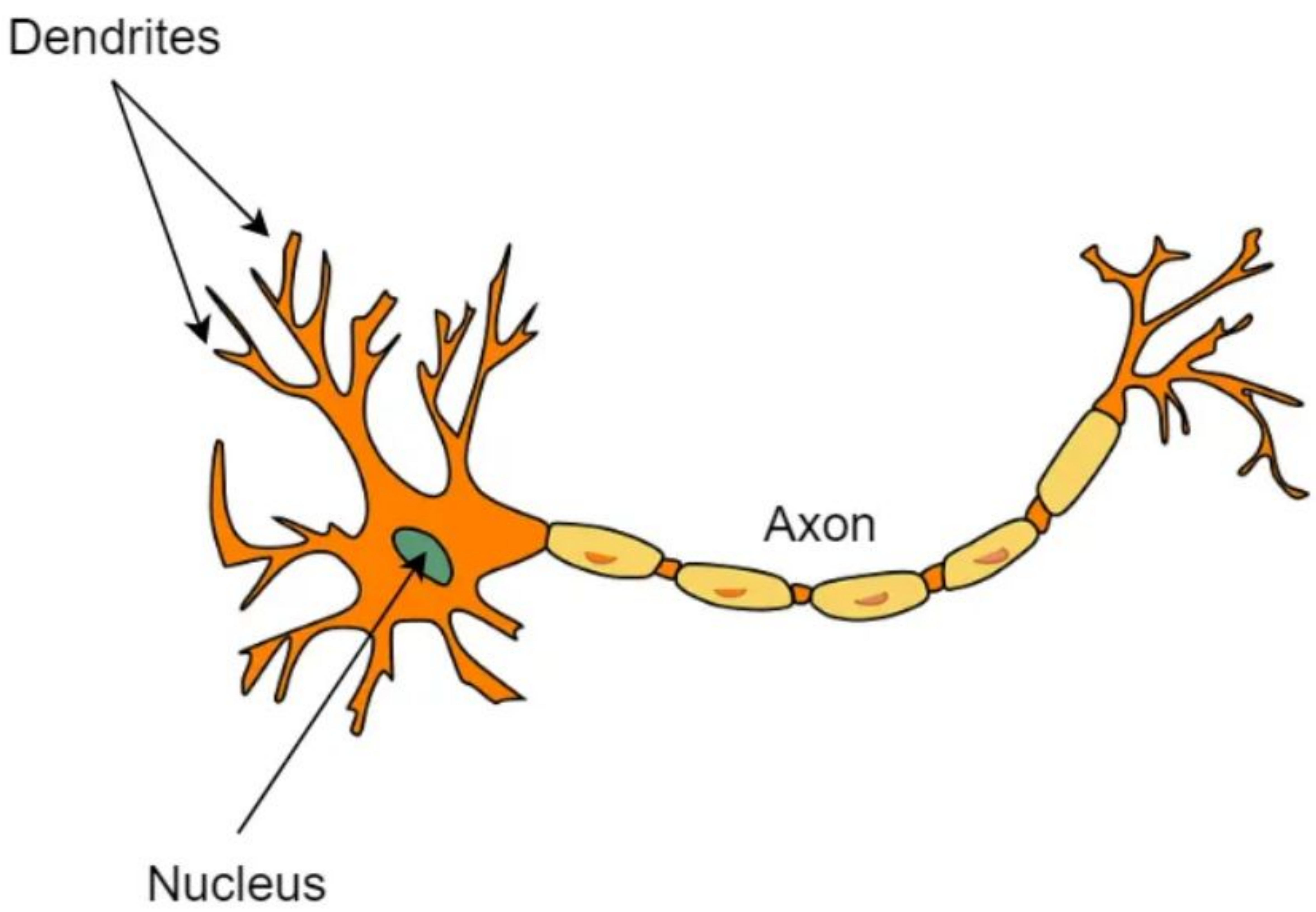
Cluster analysis

Deep Learning (DL)

Deep learning is a subset of ML

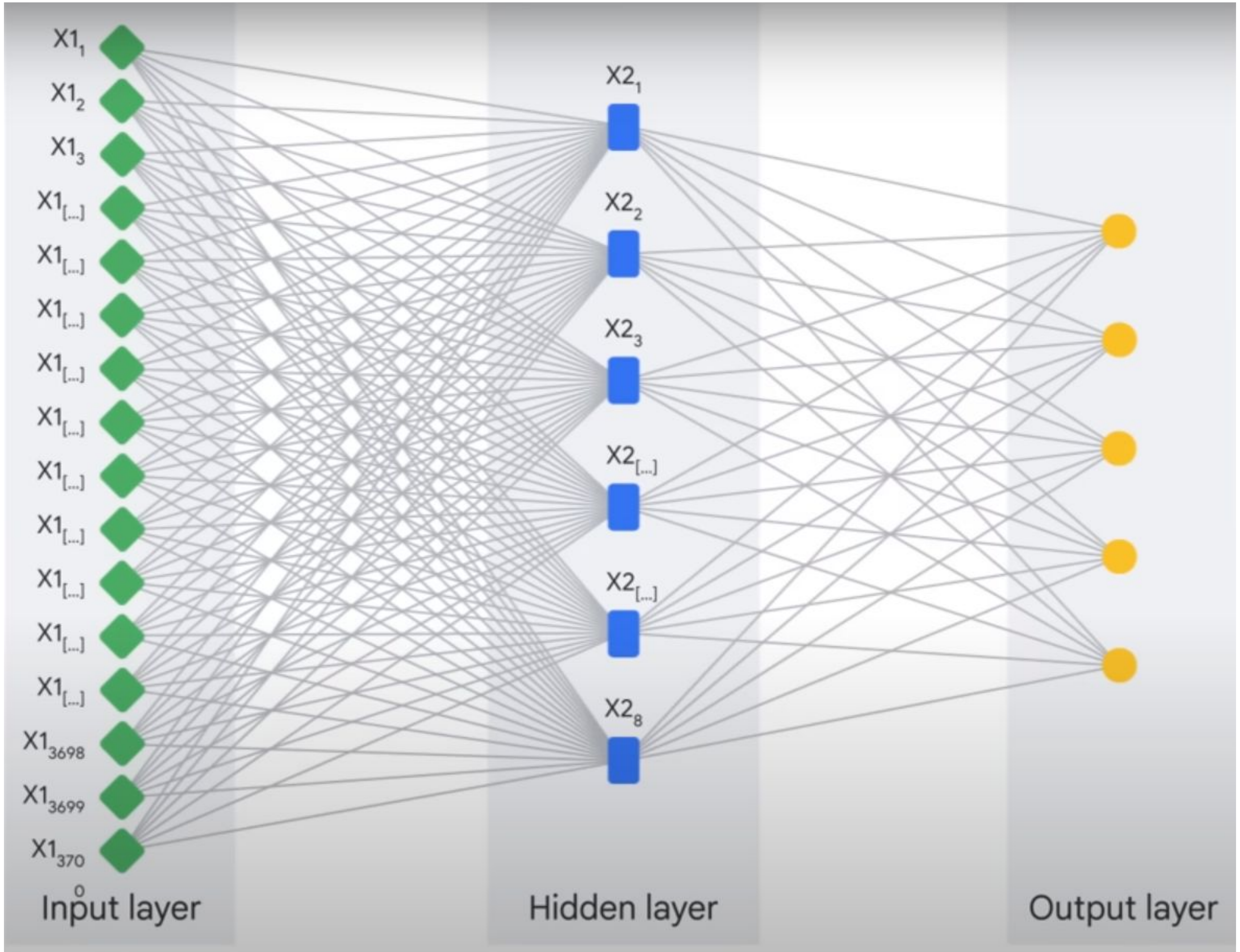


Neural Networks



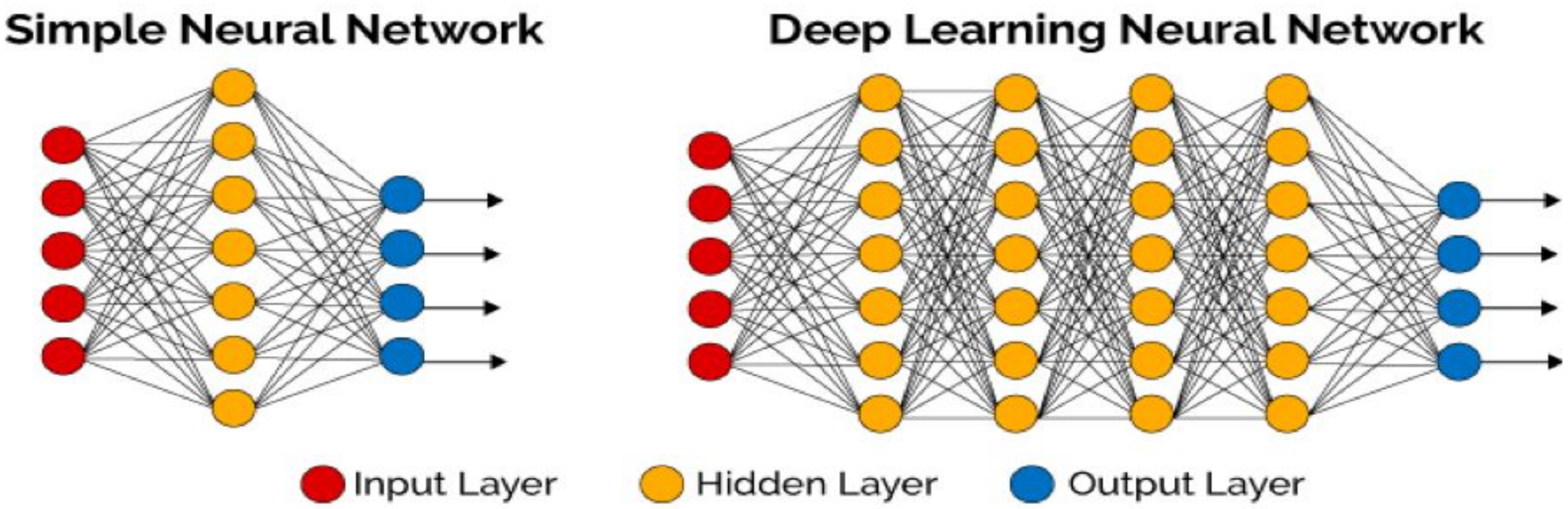
Some studies: Approximately 100 billion neurons in human brain

Artificial Neural Networks (ANNs)

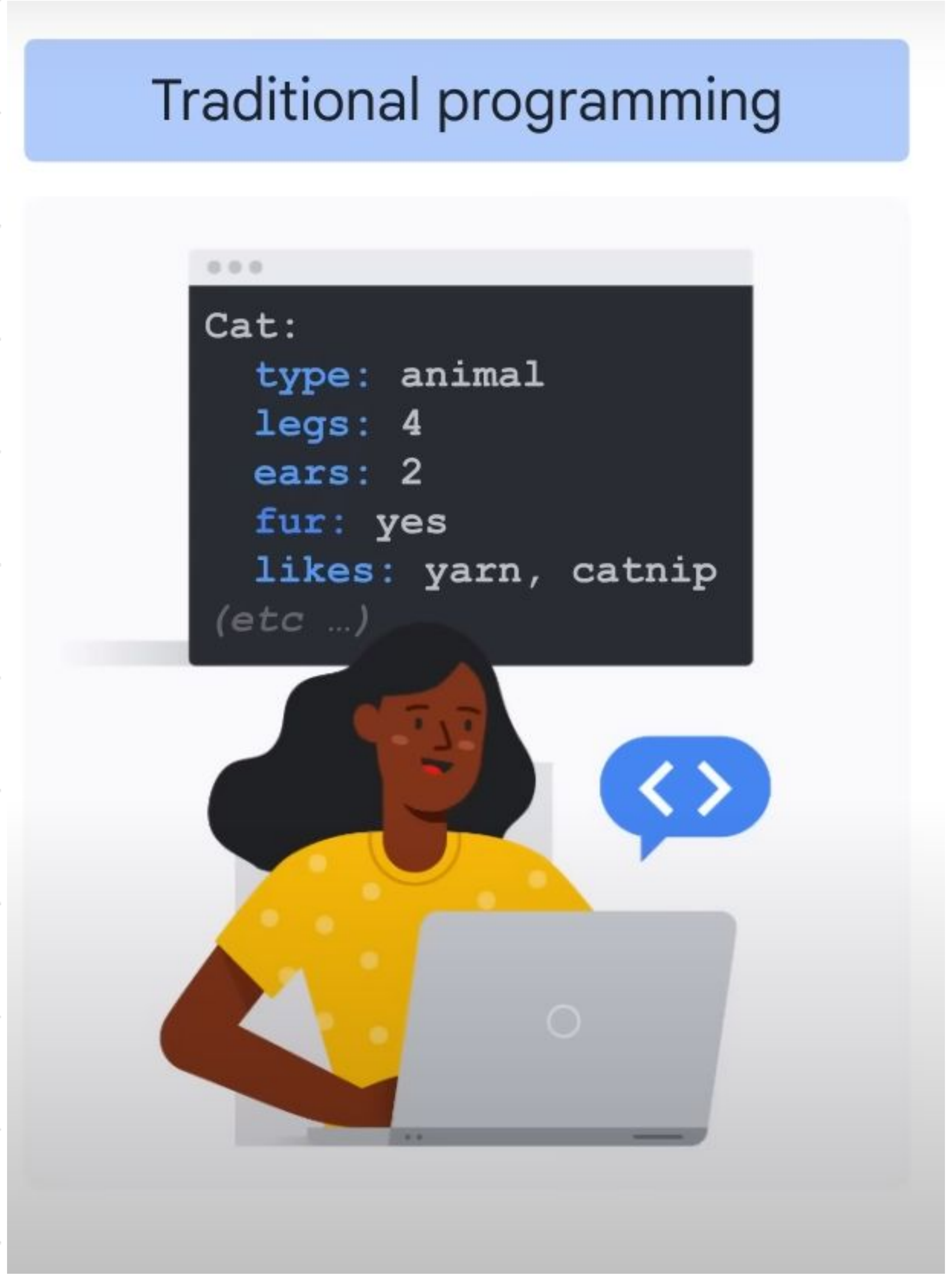


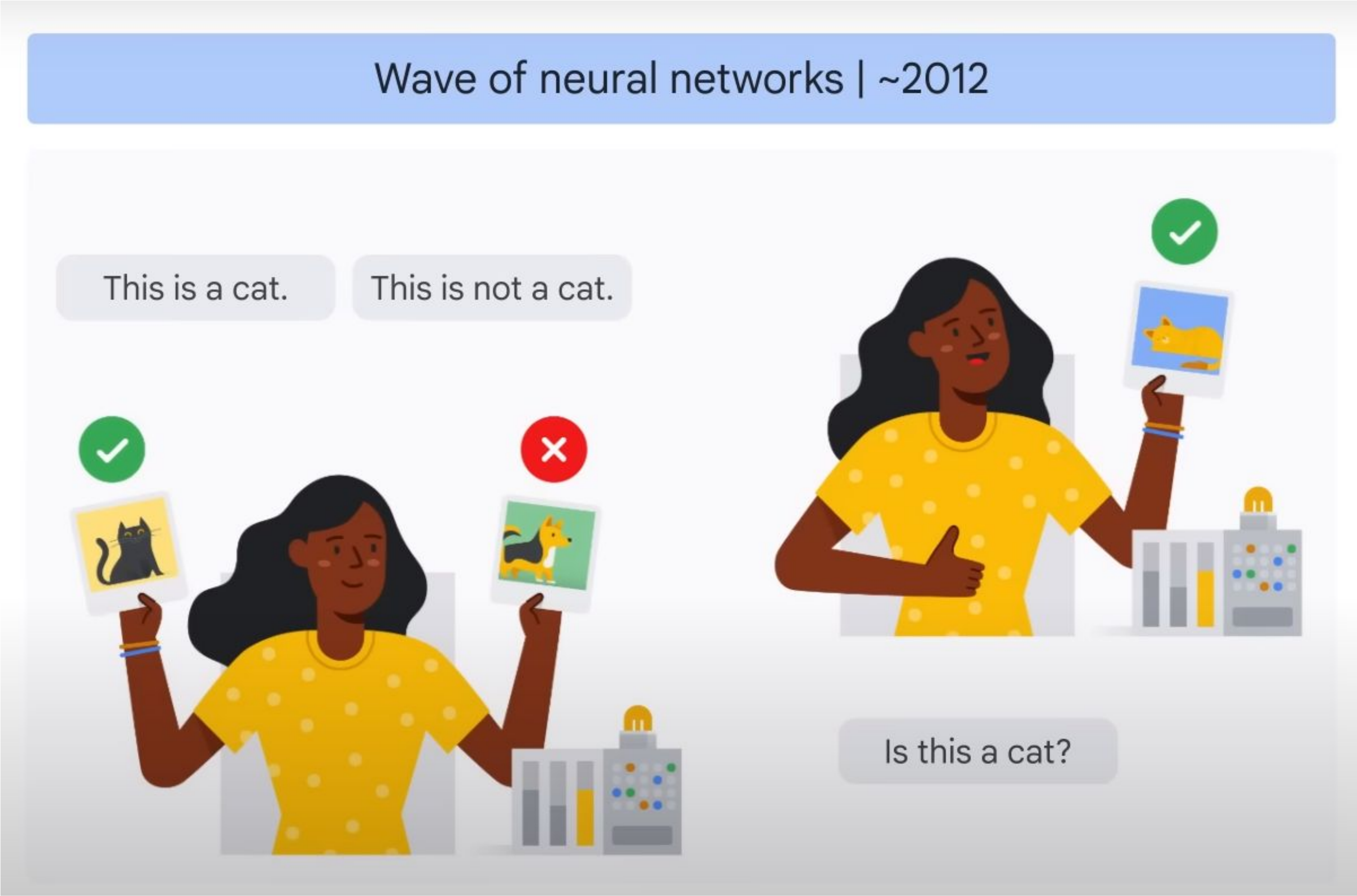
Deep Learning (DL) needs Deep ANN

- **DL**: a modern state-of-the-art approach to ML
- **DL** leverages the power of Artificial Neural Networks
- there is a debate on why the algorithms used outperform all conventional methods

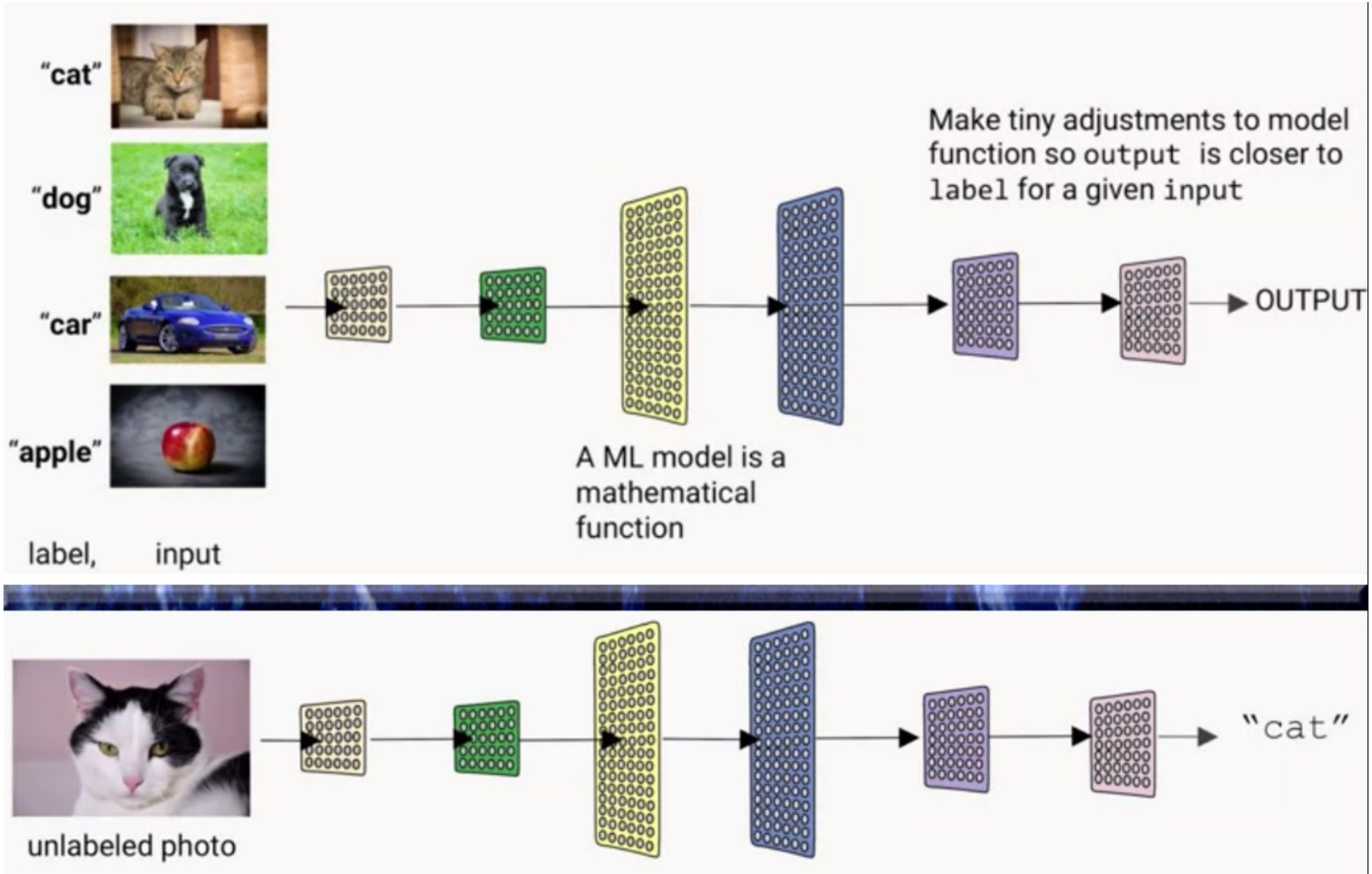


Traditional Program vs. ANNs





ANNs & Supervised Learning



ANNs & Semi-supervised Learning

- ANNs can use both **Labeled** and **Unlabeled data**
  - This is called **Semi-Supervised Learning**

- In **Semi-Supervised Learning**, a ANN is trained on
  - a small amount of **Labeled data** and
  - a large amount of **Unlabeled data**

- **Labeled data** helps the ANN to learn the basic concepts of the task while
- **Unlabeled data** helps the ANN to generalize to new examples

Notes

Note

- The border between Traditional and ML methods can be considered thin (artificial)
- The mathematics behind both is virtually the same
- Nevertheless, we will use this thin boundary to explain better
  - which techniques are considered classical and
  - which are more complex and unconventional

Questions

Links

https://github.com/FCAI-B/da

References

1. https://www.udacity.com/blog/2018/01/4-types-data-science-jobs.html
    - 4 Types of Data Science Jobs
2. https://www.coursera.org/articles/data-analyst-vs-business-analyst
    - Data Analyst vs. Business Analyst
3. https://learn.365datascience.com/courses/intro-to-data-and-data-science
    - 365 Data Science - Introduction to Data and Data Science