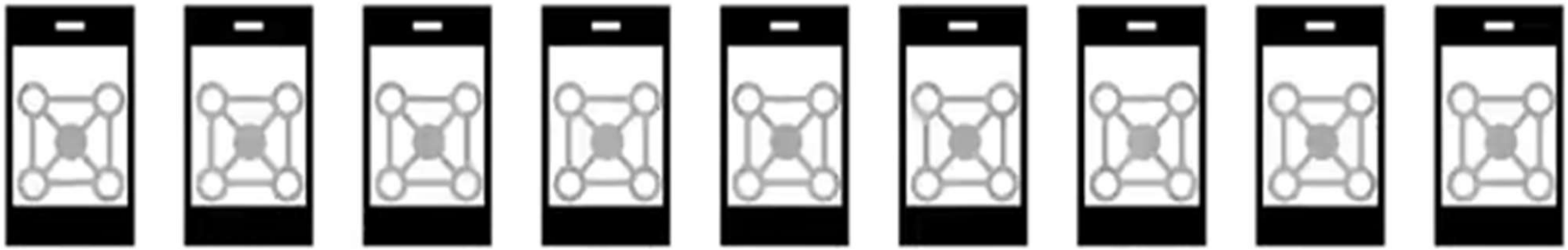
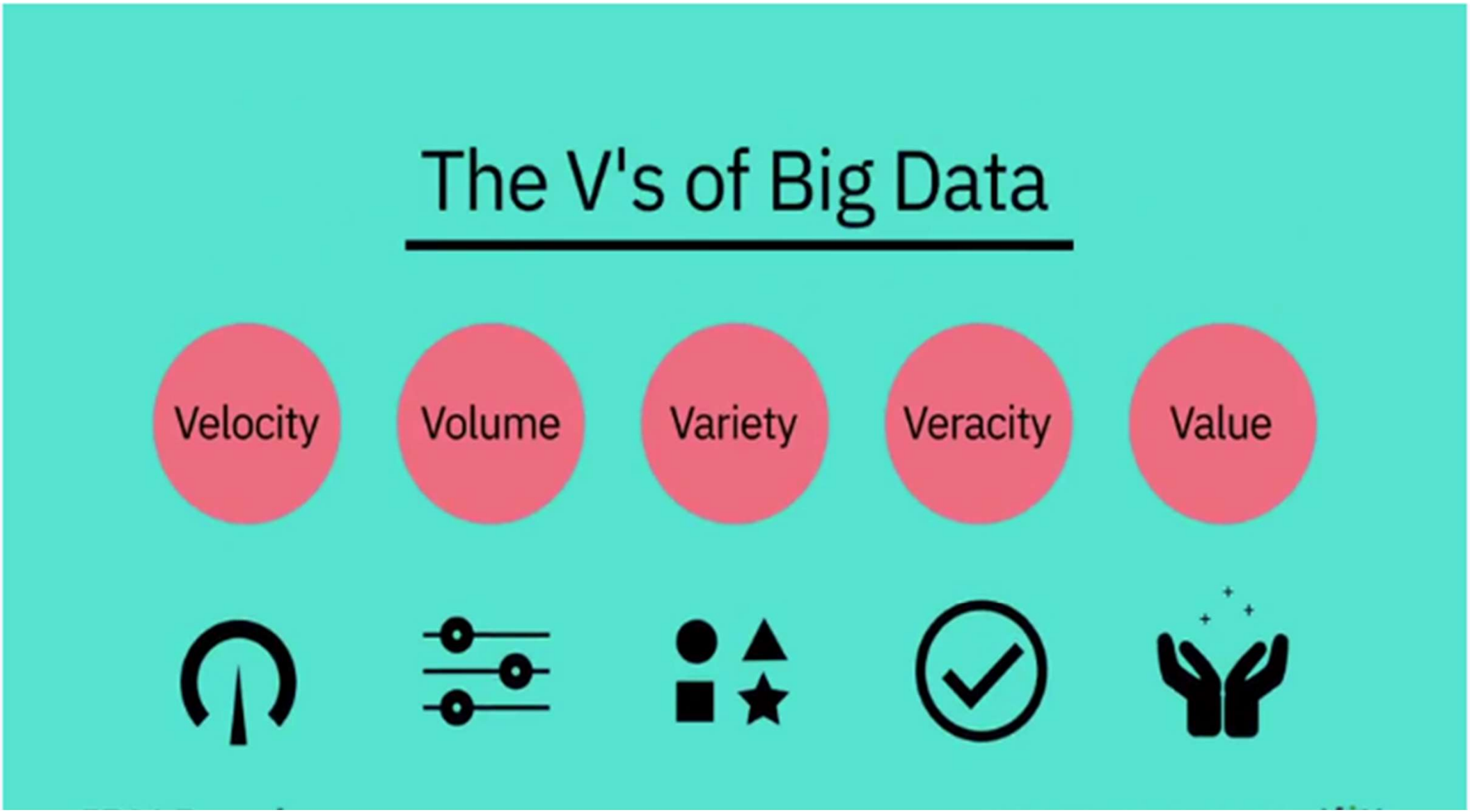
# Big Data

"Big Data refers to the dynamic, large and disparate volumes of data being created by people, tools, and machines. It requires new, innovative, and scalable technology to collect, host, and analytically process the vast amount of data gathered in order to derive real-time business insights that relate to consumers, risk, profit, performance, productivity management, and enhanced shareholder value."

Ernst and Young

# Foundations of Big Data



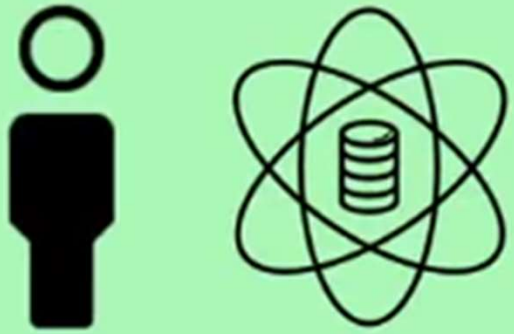

The V's of Big Data
Velocity
Volume
Variety
Veracity
Value

# Big Data characteristics

| | |
|---|---|
| Value | Investment in Big Data creates value |
| Volume | Scale of the data |
| Velocity | Speed it is collected |
| Variety | Comes from a variety of sources |
| Veracity | Conforms to facts |

# Big data

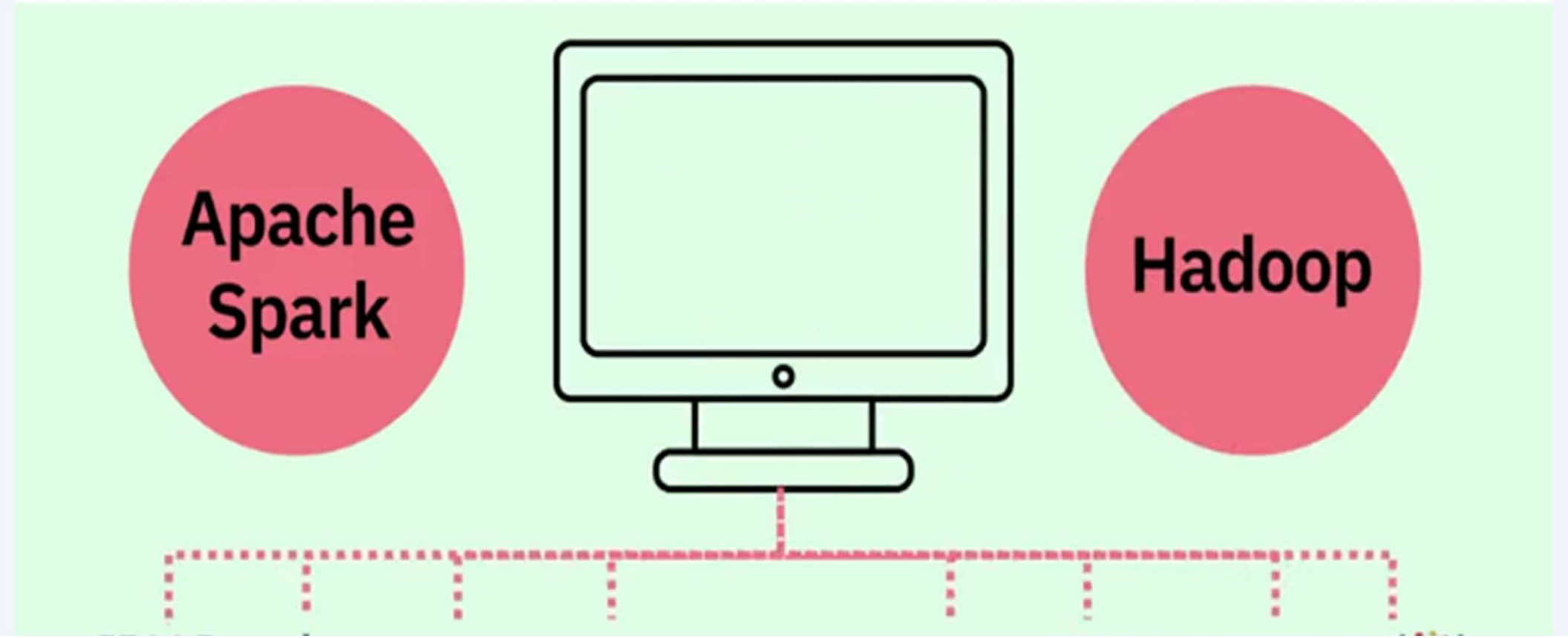
# Data scientists



IBM Developer

SKILLS NETWORK

# Alternative tools



# Big Data Processing Tools

The Big Data processing technologies provide ways to work with large sets of structured, semi-structured, and unstructured data so that value can be derived from big data.



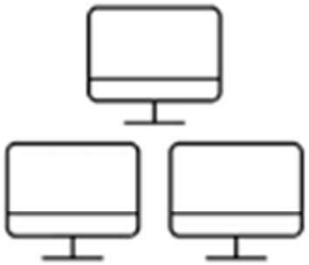
# Big Data Processing Tools


**Apache Hadoop**
**a collection of tools that provides distributed storage and processing of big data**

# Hadoop


Distributed storage and processing of large datasets across clusters of computers.

Node



Cluster

# Hadoop


Hadoop provides a **reliable, scalable,** and **cost-effective** solution for storing data with no format requirements.

**Benefits include:**

**Better real-time data-driven decisions:**
Incorporates emerging data formats not traditionally used in data warehouses

**Improved data access and analysis:**
Provides real-time, self-service access to stakeholders

**Data offload and consolidation:**
Optimizes and streamlines costs by consolidating data, including cold data, across the organization

# Hadoop

## Distributed File System


Hadoop Distributed File System, or HDFS, is a storage system for big data that runs on multiple commodity hardware connected through a network.

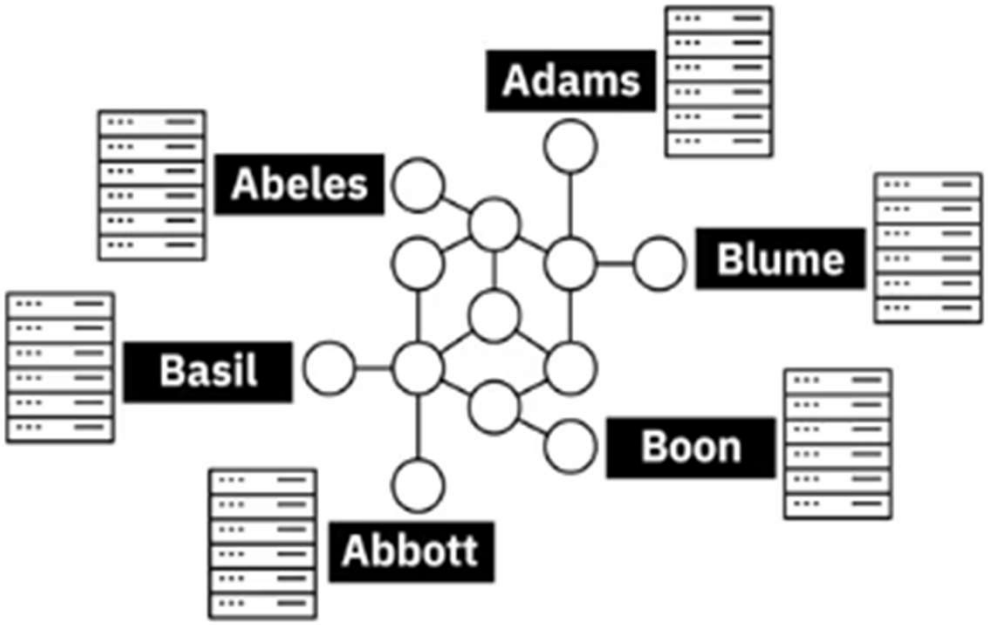
Provides scalable and reliable big data storage by partitioning files over multiple nodes

Splits large files across multiple computers, allowing parallel access to them

Replicates file blocks on different nodes to prevent data loss

# Hadoop

## Distributed File System


- Higher availability
- Better scalability
- Data locality

# Hadoop

## Distributed File System


**Benefits that come from using HDFS include:**

- Fast recovery from hardware failures, because HDFS is built to detect faults and automatically recover.
- Access to streaming data, because HDFS supports high data throughput rates.
- Accommodation of large data sets, because HDFS can scale to hundreds of nodes, or computers, in a single cluster.
- Portability, because HDFS is portable across multiple hardware platforms and compatible with a variety of underlying operating systems.

# Big Data Processing Tools


**Apache Hive**
**a data warehouse for data query and analysis**

# Hive


Hive is an open-source data warehouse software for reading, writing, and managing large data set files that are stored directly in either HDFS or other data storage systems such as Apache HBase.

Queries have high latency → Not suitable for applications that need fast response times

# Hive


Read-based → Not suitable for transaction processing that involves a high percentage of write operations.

Hive is better suited for →

- Data warehousing tasks such as ETL, reporting, and data analysis
- Easy access to data via SQL

# Big Data Processing Tools


**Apache Spark**
**a distributed analytics framework for complex, real-time data analytics**

# Spark

Spark is a general-purpose data processing engine designed to extract and process large volumes of data for a wide range of applications.

- Interactive Analytics
- Streams Processing
- Machine Learning
- Data Integration
- ETL

# Spark


**Key attributes:**

- Has in-memory processing which significantly increases speed of computations
- Provides interfaces for major programming languages such as Java, Scala, Python, R, and SQL
- Can run using its standalone clustering technology
- Can also run on top of other infrastructures, such as Hadoop
- Can access data in a large variety of data sources, including HDFS and Hive
- Processes streaming data fast
- Performs complex analytics in real-time

# What is cloud computing?

Delivery of on-demand computing resources

- Networks
- Servers
- Storage
- Applications
- Services
- Data centers

Over the Internet on a pay-for-use basis

# What is cloud computing?

Applications and data that users access over the Internet rather than locally

- Online web apps
- Secure online business applications
- Storing personal files
  - Google Drive
  - OneDrive
  - Dropbox

# Cloud

Enables us to work with Big Data

On-demand computing resources

Pay-for-use basis

# Cloud characteristics

| | |
|---|---|
| On-demand | Access to processing, storage, and network |
| Network access | Resources access via the Internet |
| Resource pooling | Shared resources dynamically assigned |
| Elasticity | Automatically scales resource access |
| Measured service | Only pay for what you use or reserve |

# Cloud benefits to data science

Addresses computing challenges related to:

- Scalability
- Collaboration
- Accessibility
- Maintenance

Gives instant access to latest technology and tools

# Cloud computing user benefits

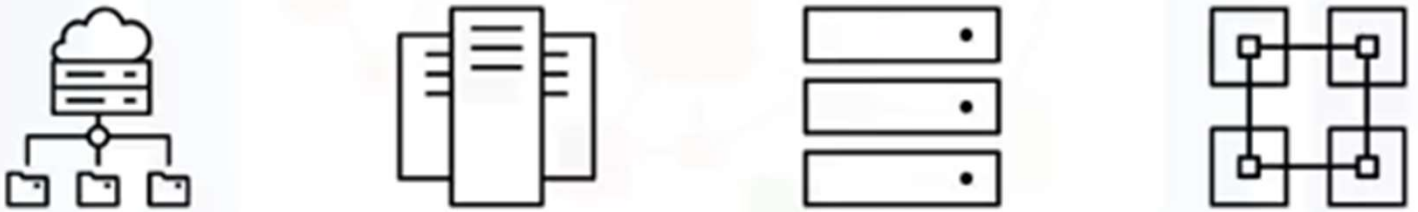
- No need to purchase applications and install them on local computer
- Use online versions of applications and pay a monthly subscription
- More cost-effective
- Access most current software versions
- Save local storage space
- Work collaboratively in real time

# Cloud computing

- Five characteristics
- Three deployment models
- Three service models

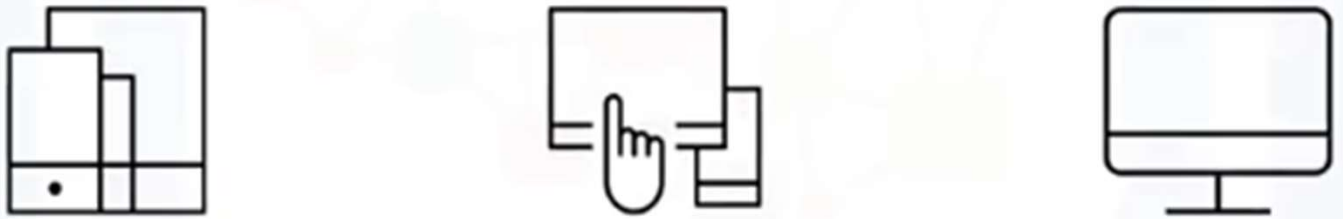
# Cloud computing characteristics

## On-demand self-service



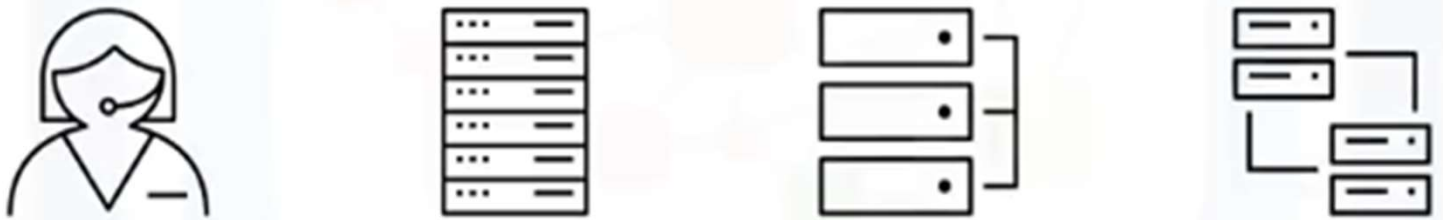
# Cloud computing characteristics

## Broad network access



# Cloud computing characteristics

## Resource pooling

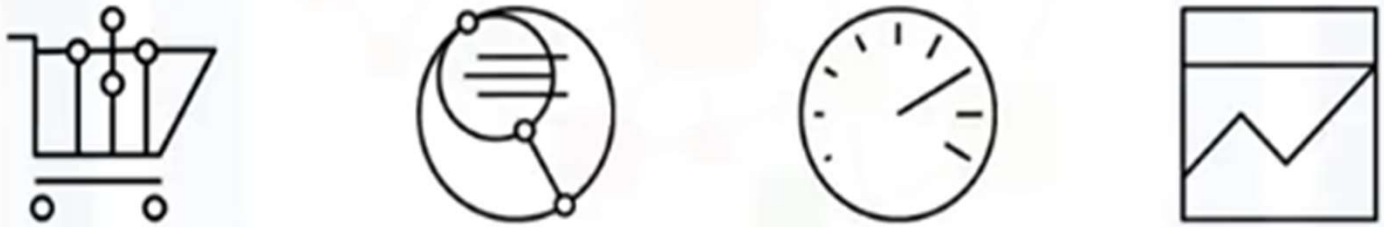


# Cloud computing characteristics

## Rapid elasticity

# Cloud computing characteristics

## Measured service



# Cloud computing characteristics

- Cloud computing is about using technology as a service by leveraging remote systems on-demand over the Internet
- It has changed the way the world consumes compute services
  - Making them more cost-efficient while making organizations more agile in response to change

# Cloud deployment models

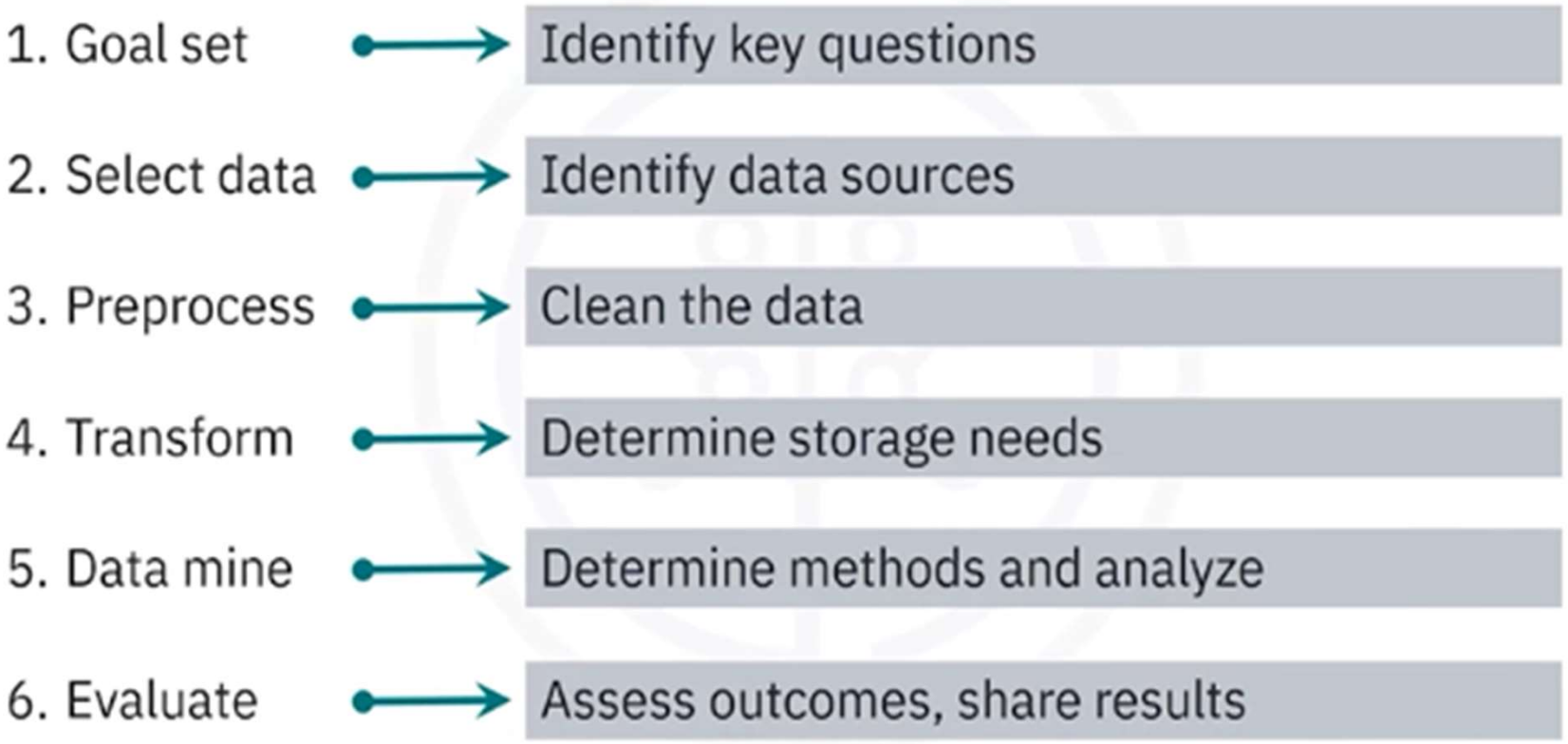
Public cloud

Private cloud

Hybrid cloud

# Cloud service models

# Data mining process



# Recap

In this video, you learned that:

- Cloud computing is the delivery of on-demand computing resources over the Internet on a pay-for-use basis
- Cloud computing is composed of five essential characteristics, three deployment models, and three service models
- Five essential characteristics of cloud computing: on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service
- Three cloud deployment models: public, private, and hybrid
- Three cloud service models are based on three layers in a computing stack: IaaS, PaaS, and SaaS

# Machine Learning

A subset of AI that uses computer algorithms to analyze data and make intelligent decisions based on what it has learned, without being explicitly programmed

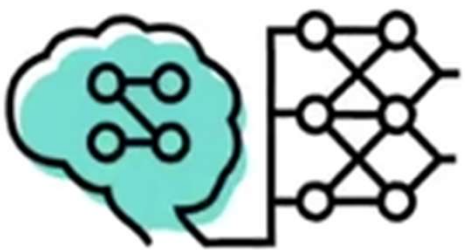
# Machine Learning

A subset of AI that uses computer algorithms to analyze data and make intelligent decisions based on what it has learned, without being explicitly programmed

Trained with large sets of data

They do not follow rules-based algorithms

They learn from examples



# Deep Learning

A specialized subset of Machine Learning that uses layered neural networks to simulate human decision-making

# Deep Learning

A specialized subset of Machine Learning that uses layered neural networks to simulate human decision-making

Deep Learning algorithms

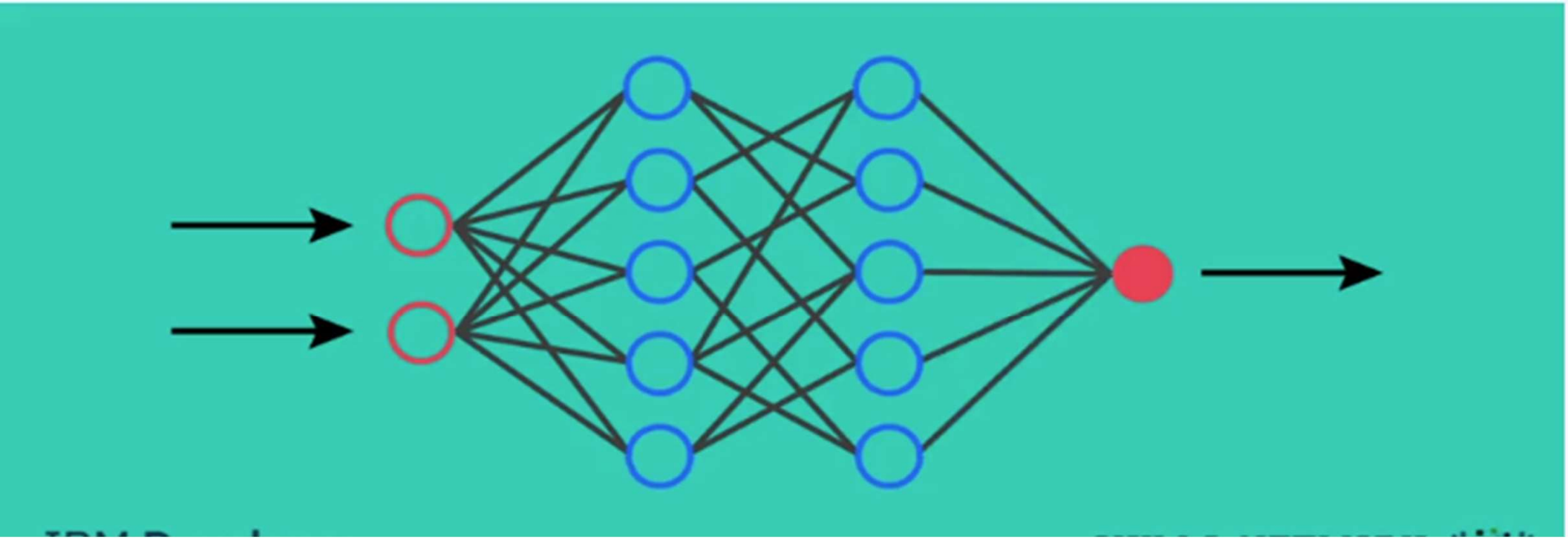
Label and Categorize

# Neural Networks

Take inspiration from biological neural networks, although they work quite a bit differently

# Neural Networks

Take inspiration from biological neural networks, although they work quite a bit differently

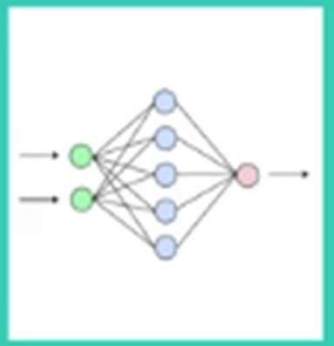


# Understanding the differences in key AI concepts

**Artificial Intelligence**

**Machine Learning**

**Deep Learning**



**Neural Networks**

# Artificial Intelligence and Data Science

**Data Science**
is the process and method for extracting knowledge and insights from large volumes of disparate data.

**It involves**
mathematics
statistical analysis
data visualization
machine learning
and more...

**It could use**
machine learning algorithms

deep learning models

**It's a broad term**
encompasses the entire data processing methodology

AI includes everything that allows computers to learn how to solve problems and make intelligent decisions

Both AI and Data Science can involve the use of **Big Data**

# Generative AI and Data Science

# What you will learn

Describe generative AI and

Explain how data scientists use generative AI in data science

# What is Generative AI?

Produces new data

Creates content:

- Images
- Music
- Language
- Code

# How does Generative AI work?

Generative adversarial networks (GANs )

Variational auto-encoders (VAEs)

Replicates underlying features of original data

# Applications of Generative AI

Natural Language Processing

Healthcare

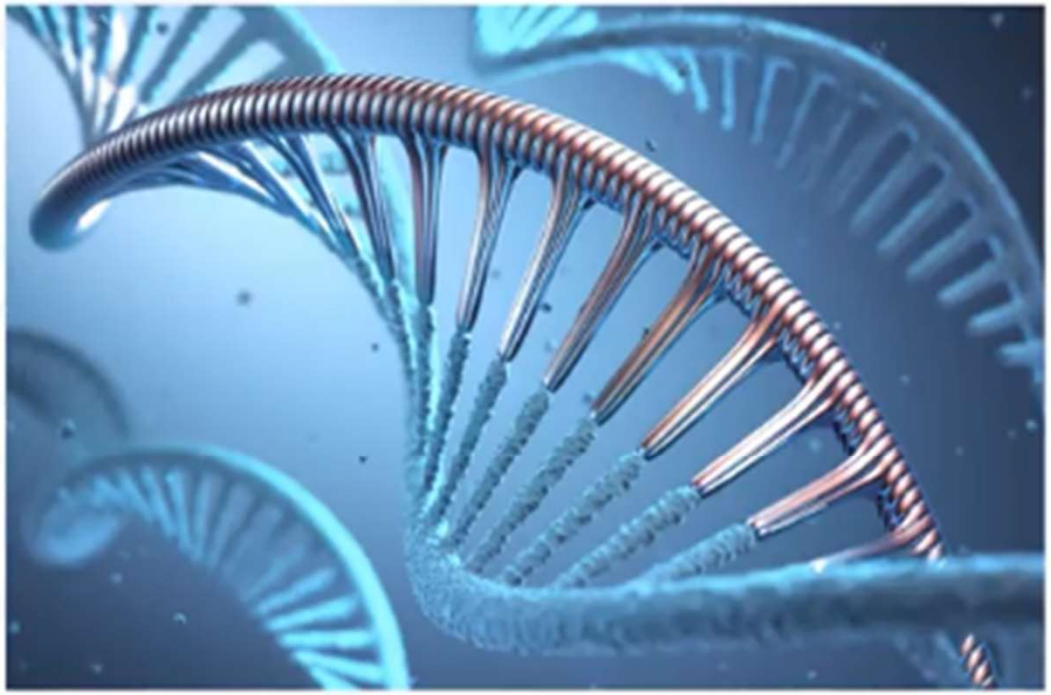
Art and Design

Gaming

Fashion and Retail

# Synthetic data

- Building data models takes a lot of data
- Data sets may not have enough data to build a model
- Generative AI makes data augmentation possible
- Creates data with similar properties
- Use this synthetic data for model training and testing



# Coding automation

Confined by a time when examining data

Generate software code to construct models

Focus on higher-level tasks

# Uncover insights


- Generate insights and reports
- Automate updates
- Enhance decision-making
- IBM Cognos Analytics
- Music
- Language
- Code

# Deep Learning and Machine Learning

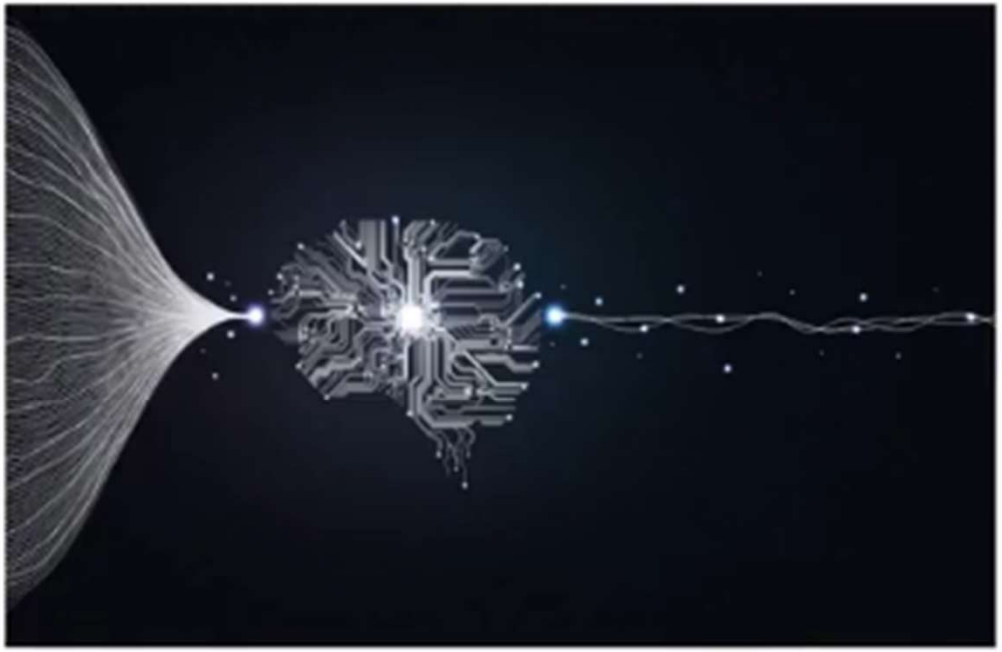
Lesson Summary

# From this lesson

- Terms used in AI
- How data scientists use AI
- Relationship between machine learning and data science
- Regression

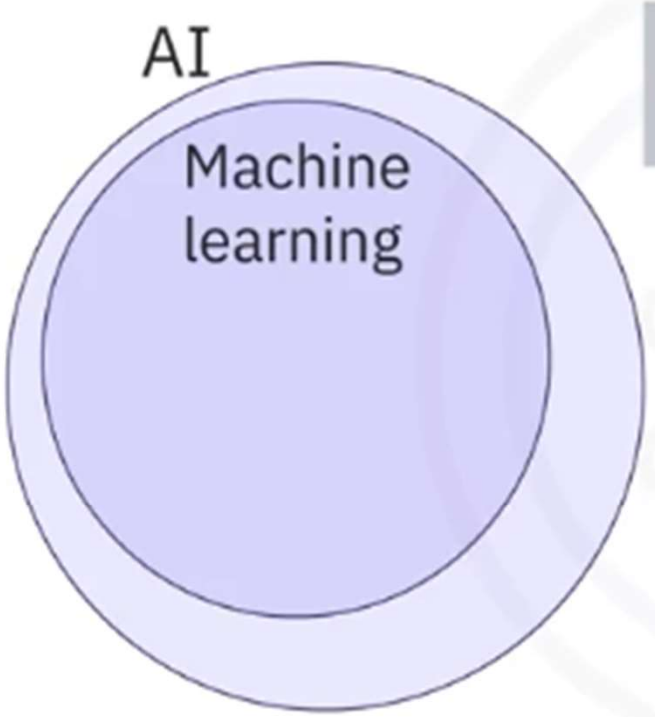
# AI terminology

AI terms you should know:



- Machine learning
- Deep learning
- Neural networks
- Generati

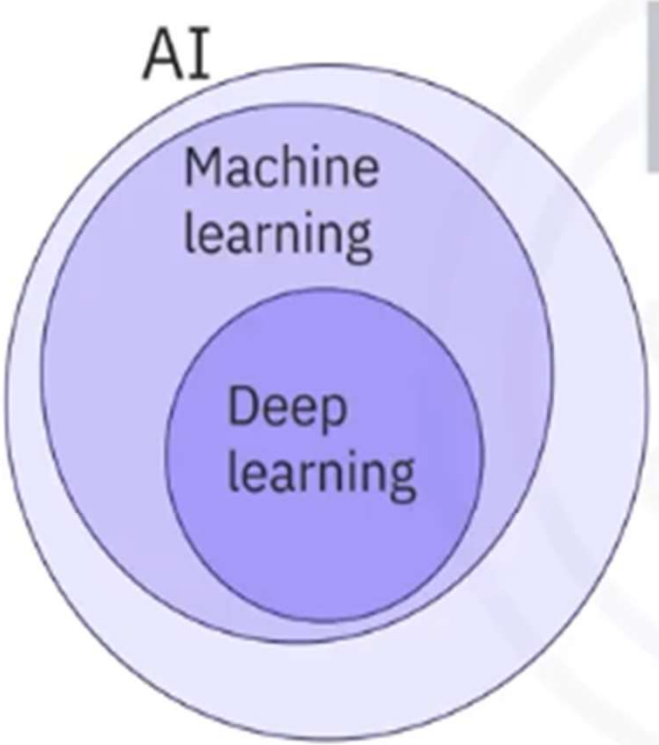
# Machine learning




Using computer algorithms to analyze and make predictions

Analysis happens without the need for explicit programming

# Deep learning




Simulates human decision-making using neural networks

Neural networks:

- Collection of computing units called neurons
- Design inspired by how neurons in the brain behave

# Generative AI

Subset of AI

Focus on new data production

Mimics content created by humans

Can generate new data to use when training and testing a model

# Machine learning algorithms

Applications:

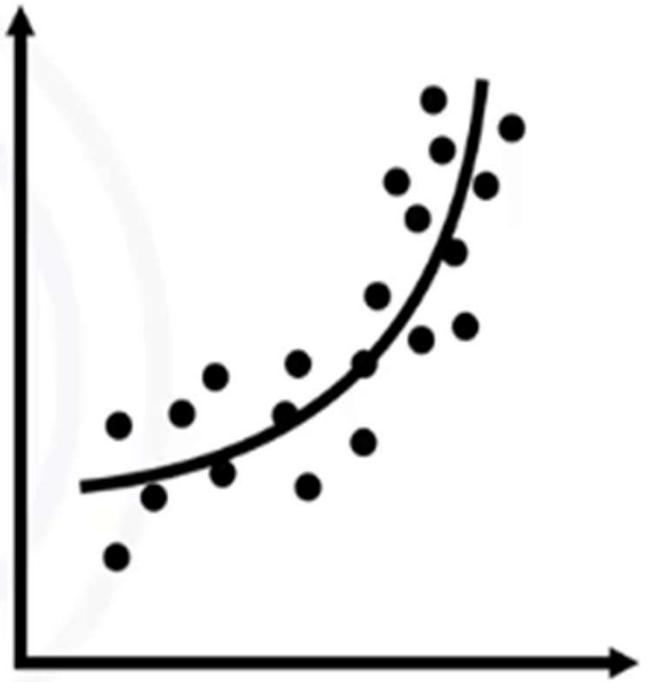
- Make predictions
- Make recommendations

Ex: Predict whether a credit card purchase is fraudulent or not

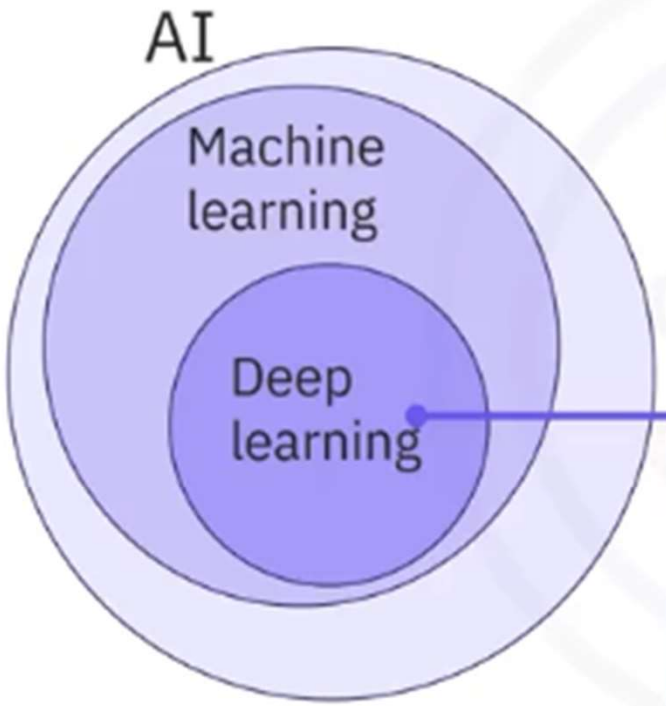
# Regression

Identifies correlation between inputs and outputs

Ex: Predict the cost of a house based on square footage and number of bedrooms



# Recap




Generative AI produces new data from raw data with similar traits

Uses neural networks to teach itself patterns from inputs and outputs

Data scientists use AI along with “Big Data” to make predictions