

Understanding Machine Learning

* Machine Learning is a subset of Artificial Intelligence that enables programs to learn patterns from data without being explicitly programmed.
* The performance of Machine Learning algorithms improves with more data, but there is a point of diminishing returns.

Key Concepts: Features and Targets

* Features are the input variables used for predictions, while the target is the outcome we want to predict.
* An example is the iris dataset, which uses features like sepal length and width to predict flower species.

Types of Machine Learning:

* Supervised Learning: Involves labeled datasets to predict outcomes (e.g., spam detection).
* Unsupervised Learning: Works with unlabeled data to find patterns (e.g., customer segmentation).

Machine Learning Workflow

* The first step is defining the problem statement, such as classifying different breeds of dogs.
* Next, data collection is crucial, requiring a large set of labeled data for effective model training.

Basic Machine Learning Vocabulary

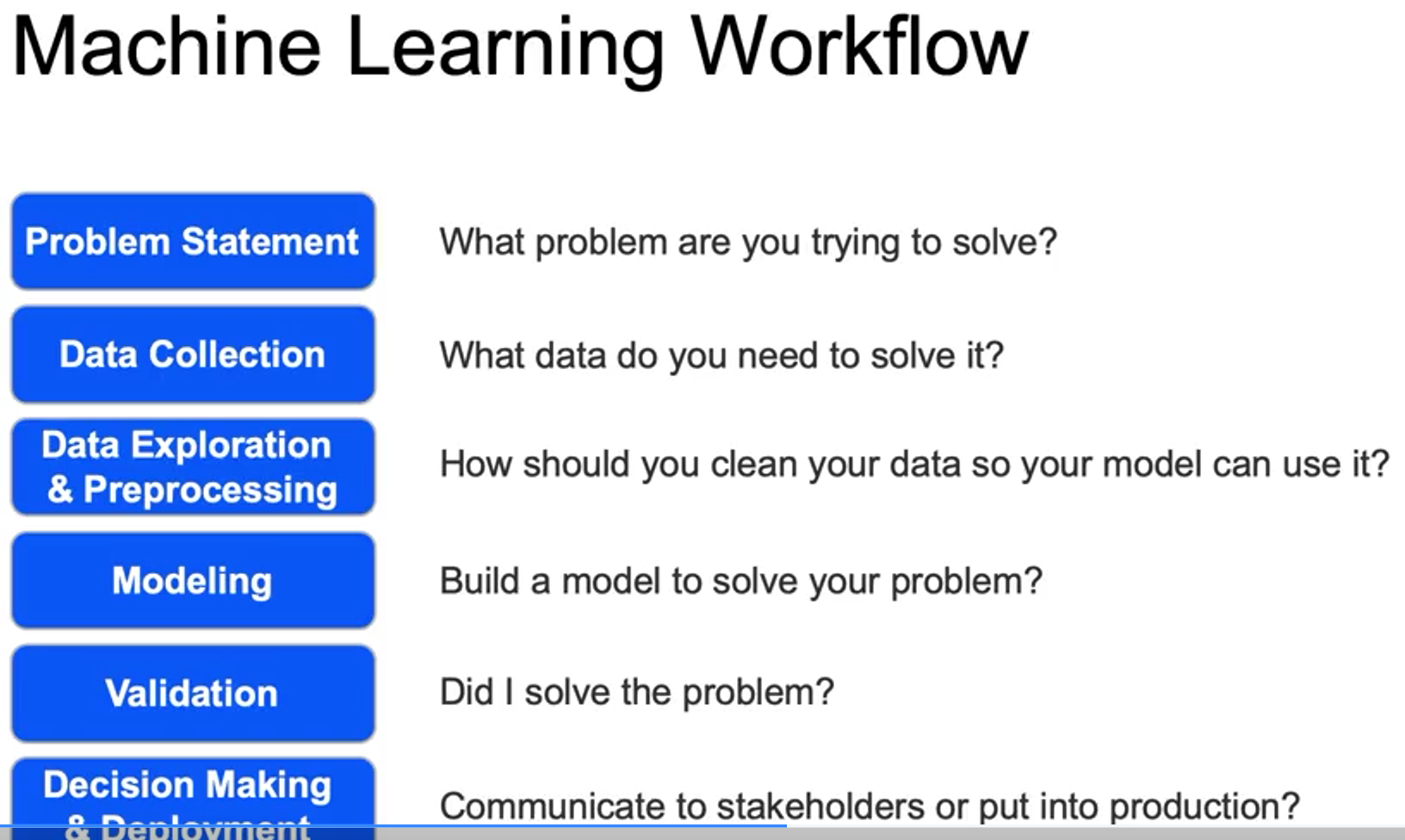
* The target variable is the value being predicted, like the species of an iris flower in a dataset.
* Features are the explanatory variables used to predict the target variable, such as sepal and petal dimensions.

Retrieving Data from CSV Files

* CSV (Comma Separated Values) files consist of rows of data separated by commas, easily read in Pandas using pd.read\_csv().

Working with JSON Files

* JSON (JavaScript Object Notation) files are commonly used for data storage and resemble Python dictionaries.
* To read JSON files, use pd.read\_json()



**Working with SQL Databases**

* SQL (Structured Query Language) is used for relational databases with a fixed schema, such as Microsoft SQL Server, MySQL, and Oracle DB.
* Python libraries like sqlite3 and SQLAlchemy allow connections to SQL databases, enabling data retrieval using SQL queries.

**Working with NoSQL Databases**

* NoSQL databases are non-relational and can store data in various formats, often using JSON.
* Examples include document databases (like MongoDB) and graph databases, which maintain relationships between data points.

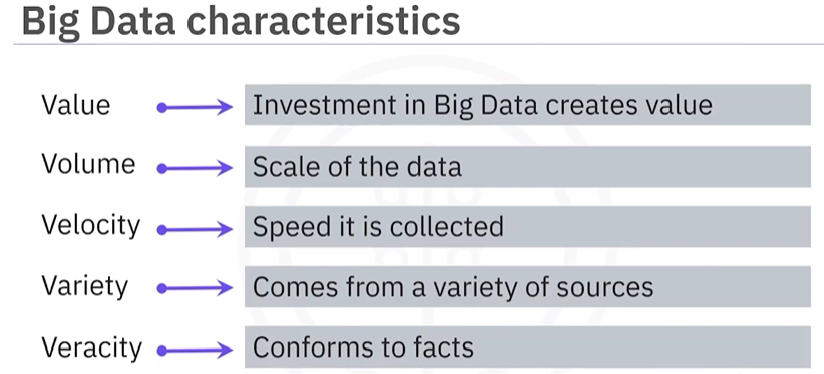
**Handling Missing Values**

* Removing Data: You can remove entire rows to clean the dataset quickly, but this may lead to loss of important information if many rows are missing values.
* Imputation: Replacing null values with the mean or median helps retain data but introduces uncertainty as these are estimates of the true values.

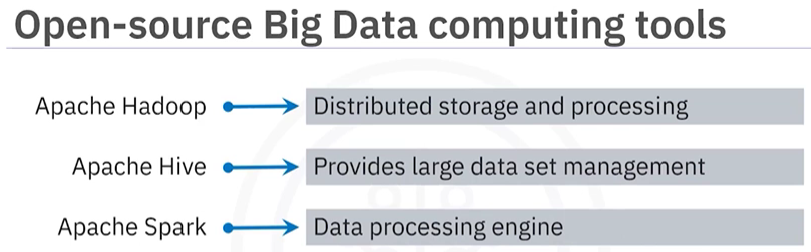
**Outliers**

* Definition: Outliers are distinct observations that can skew model predictions, such as an unusually high sales figure compared to typical values.
* Identification: Use visualizations like histograms, density plots, and box plots to detect outliers, and apply statistical methods to define their boundaries based on percentiles and interquartile ranges.

**BIG DATA:** think about how many videos are uploaded to YouTube every minute. This rapid flow of information is called "velocity." The sheer amount of data we create daily, like the 2.5 quintillion bytes from all digital devices, represents "volume." And the different types of data, such as text, images, and videos, showcase "variety." By understanding these aspects, businesses can better connect with customers and improve their services.

****

**Big Data Clusters:**   
In the world of data, a big data cluster is like a team of many computers working together to handle huge amounts of information. Instead of bringing all the data to one computer, which can be slow and overwhelming, the data is split into smaller pieces. Each piece is sent to different computers in the cluster, and they all work on their piece at the same time. This way, they can process a lot of data quickly and efficiently.



focuses on three key open-source technologies used in big data analytics: Apache Hadoop, Apache Hive, and Apache Spark.

**Apache Hadoop**

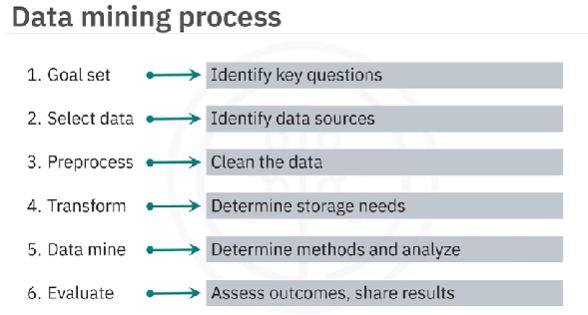
* A framework for distributed storage and processing of large datasets across clusters of computers, allowing for scalability from a single node to many.

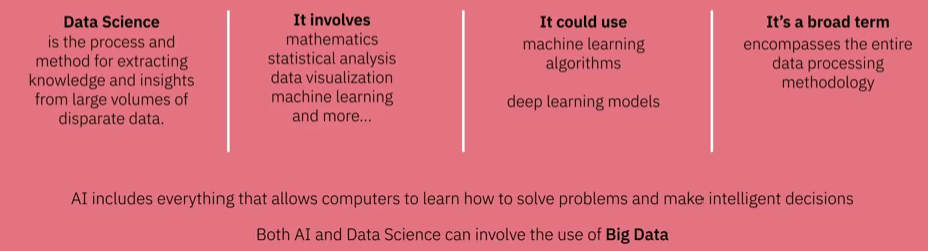
**Apache Hive**

* An open-source data warehouse software built on Hadoop, designed for reading, writing, and managing large datasets.

**Apache Spark**

* A general-purpose data processing engine that excels in real-time analytics and complex data processing.





**Generative AI** **in simple terms:** Generative AI is like a creative artist that can produce new things instead of just looking at what already exists. Imagine a painter who can create beautiful paintings from scratch, or a musician who can compose original songs. In the world of technology, generative AI uses deep learning models to learn from a lot of data and then create new content, such as images, music, or even text that sounds like it was written by a human.

**Neural Networks:**

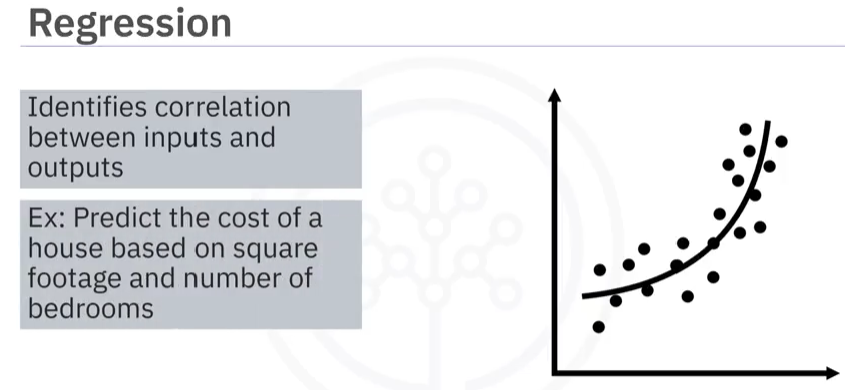
What They Are: Neural networks are computer systems designed to recognize patterns, similar to how our brains work. They consist of layers of interconnected nodes

* **Simple Example**: Imagine you want to teach a computer to recognize pictures of cats and dogs.
  + **Input Layer**: You feed the network images of cats and dogs.
  + **Hidden Layers**: The network has hidden layers that analyze the images, looking for features like fur patterns, shapes, and colors.
  + **Output Layer**: Finally, the network decides if the image is a cat or a dog based on what it learned.

**Deep Learning:**

What It Is: Deep learning is a more advanced version of neural networks that uses many layers

* **Simple Example**: Continuing with the cat and dog example:
  + In a deep learning model, there might be several hidden layers. The first layer might learn to identify edges, the next layer might learn to recognize shapes, and deeper layers might learn to identify specific features like ears or tails.
  + This multi-layer approach allows the model to become very good at distinguishing between cats and dogs, even in tricky images.

**Regression** in machine learning is a statistical method used to model and analyze the relationship between a dependent variable (the outcome we want to predict) and one or more independent variables (the factors that influence the outcome). **Regression** helps us understand how one variable (house size) affects another (price).

**Exploratory Data Analysis** is like being a detective with data. Imagine you have a big box of puzzle pieces (your data), and before you start putting the puzzle together, you want to see what pieces you have. EDA helps you look at your data to understand its main features, like finding out how many pieces you have, what colours they are, and if any pieces are missing or damaged. You can use graphs and charts to visualize this information, making it easier to spot patterns or trends, just like seeing the picture on the puzzle box.

***Estimation:***

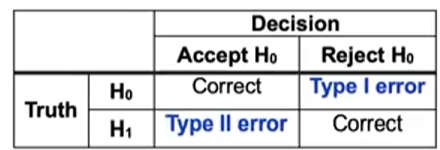
* **Definition**: Estimation involves calculating a value (like a mean or average) based on a sample of data.
* **Purpose**: It provides a single value that represents a characteristic of a larger population.
* **Example**: If you measure the heights of 10 students and find the average height to be 5 feet 6 inches, that average is your estimate of the average height of all students.

***Inference:***

* **Definition**: Inference goes beyond estimation by assessing the reliability of the estimate and making conclusions about the population based on the sample.
* **Purpose**: It provides a range of values (like a confidence interval) and helps understand the uncertainty associated with the estimate.
* **Example**: Continuing from the previous example, if you say, "I estimate the average height is 5 feet 6 inches, and I’m 95% confident that the true average height is between 5 feet 5 inches and 5 feet 7 inches," you are making an inference.

**Hypothesis testing** is a statistical method used to make inferences or draw conclusions about a population based on sample data.

* **Null Hypothesis (H0)**: A statement that there is no effect or no difference, often representing a specific value (e.g., the mean is equal to 5).
* **Alternative Hypothesis (H1 or Ha)**: A statement that indicates the presence of an effect or a difference (e.g., the mean is greater than 5).



* In simple terms, when we conduct a hypothesis test, we start with a null hypothesis (the idea we think is true) and an alternative hypothesis (what we think might be true instead). A Type I error happens when we mistakenly reject the null hypothesis when it is actually true. Imagine you have a coin that you believe is fair (50% chance of heads). If you toss it and decide it’s unfair (reject the null hypothesis) when it really is fair, that’s a Type I error.
* On the other hand, a Type II error occurs when we fail to reject the null hypothesis when it is actually false. Using the same coin example, if you toss the coin and think it’s fair (accept the null hypothesis) when it’s actually biased, that’s a Type II error.

**Correlation:**

* **What it is:** When two things happen at the same time or change together.
* **Example:** If you notice that when it rains, more people use umbrellas, that's correlation. It means rain and umbrella use are related, but it doesn't mean rain causes people to use umbrellas.

**Causation:**

* **What it is:** When one thing directly causes another thing to happen.
* **Example:** If you water a plant (X) and it grows (Y), that's causation. Watering the plant causes it to grow.

**Significance level and p-values**

* A significance level is a probability threshold below which the null hypothesis can be rejected. You must choose the significance level before computing the test statistic. It is usually .01 or .05.
* A p-value is the smallest significance level at which the null hypothesis would be rejected. The confidence interval contains the values of the statistic for which we accept the null hypothesis.