The Fundamentals of DATA

- Data quality is crucial as models learn from the provided data.

- Models can't differentiate between true and false unless explicitly trained to do so.

- **Types of Data:**

    - **Structured Data:** Organized in databases or spreadsheets with clear features (e.g., height, weight, age). Models learn from these explicit features to make predictions.
    - **Unstructured Data:** Features are not predefined; the model must learn them. Examples include text, images, and videos. For instance, a health score model might learn features from a block of text.
- **Machine Learning Models:**

    - **Supervised Learning:** Requires labeled data (e.g., health scores).
    - **Unsupervised Learning:** Doesn't need labeled data, often used for clustering.
    - **Reinforcement Learning:** Based on rewards and feedback.
- Structured data often requires fewer examples compared to unstructured data, which needs extensive examples to identify features.

- **Neural Networks:**

    - These allow models to work with unstructured data by learning features similarly to the human brain.
- **Data Collection Methods:**

    - **Public Datasets:** Available online, e.g., Kaggle.
    - **Proprietary Datasets:** Owned by companies, e.g., Meta's user data.
    - **User Data:** Collected from services or devices, e.g., Netflix recommendation
- Processing large datasets requires significant computational power, often needing GPUs.

- Ensuring data quality and diversity to avoid biases is critical. For example, Google's issue with stereotypical image search results for "doctor" and "nurse."

- Don’t use all data for training. Split between train, test and validate. Or use cross validation. Split data into folds (5 or 10). Give different folds at different types and then test on last fold.