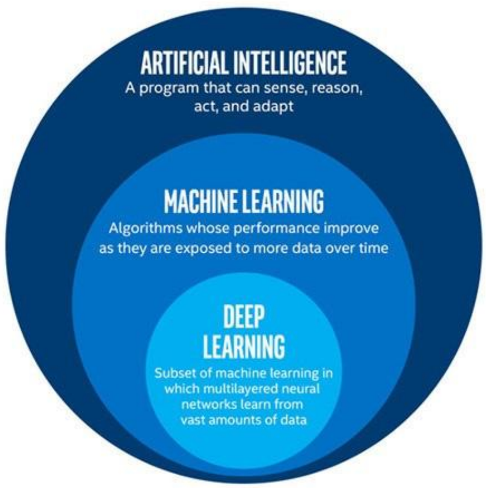
EXPLORATORY DATA ANALYSIS FOR MACHINE LEARNING

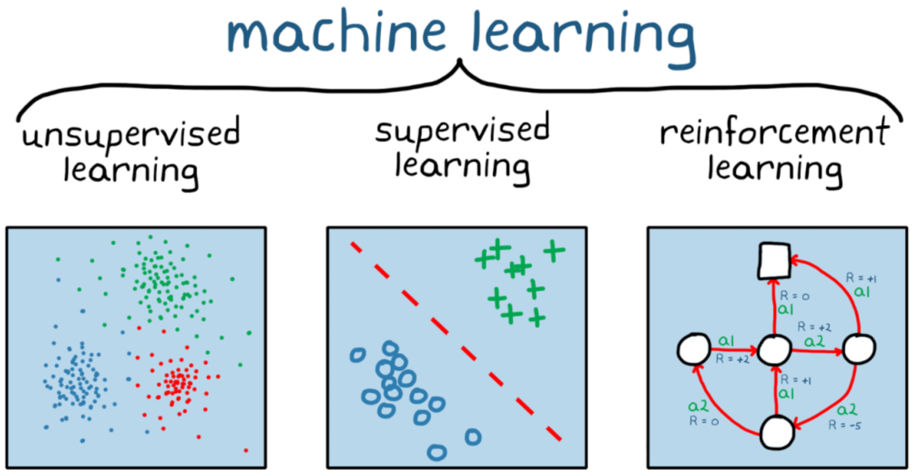
A Brief History of Modern AI and its Applications

Definitions and Relationships



- **AI (Artificial Intelligence):** Systems simulating human intelligence (sense, reason, act, adapt).
- **ML (Machine Learning):** Subfield of AI, enables machines to learn from data instead of explicit programming.
- **DL (Deep Learning):** Subfield of Machine Learning, uses multi-layer neural networks, automatically extracts features, improves with large datasets.

Machine Learning



- Learns patterns from data, improves over time. May reach diminishing returns with excessive data.
- **Types:**
    - **Supervised Learning:** Labeled data → prediction (spam, fraud detection).
    - **Unsupervised Learning:** Unlabeled data → discover hidden structures (customer segmentation).

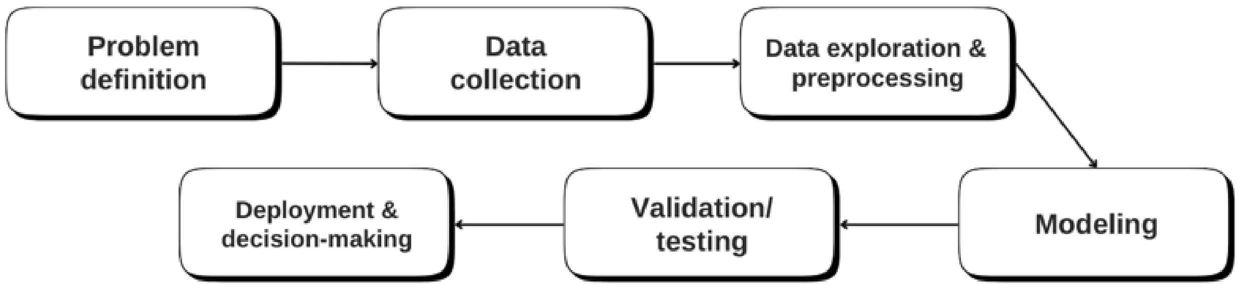
Deep Learning vs. Traditional Machine Learning

- **Traditional Machine Learning:** Requires manual feature engineering; struggles with complex data (e.g., images with 65k features).
- **Deep Learning:** Automatically extracts features, excels with images/language.
- **Comparison:**
    - DL → stronger with large datasets, less feature engineering.
    - Traditional Machine Learning → better for small or dynamic datasets.

Factors Driving AI Growth

- Availability of **big data**.
- **Increased computing power** (GPUs, cloud).
- **Accessible tools** (TensorFlow, PyTorch).

Basic Machine Learning Workflow Steps:



Historical Context

- **1956:** AI introduced at Dartmouth Conference.
- **1950s–70s:** Perceptron, Arthur Samuel's Machine Learning → failed machine translation → first AI Winter.
    - Main reasons: lack of powerful computing systems and algorithms, high expectations that could not be met → lost faith → major powers like America cut funding.
- **1980s:** Expert systems boom → limited adaptability → second AI Winter.
    - Main reasons: expert system revealed many limitations (high development and maintenance cost, lack of learnability and extensibility), collapse of the specialized Lisp machine market - base of AI, again high expectations that could not be met.
- **1990s–2000s:** Machine Learning success in speech recognition, search, robotics; 1996: Deep Blue beat chess champion.
- **2006:** Deep learning breakthrough → deeper neural networks feasible.
- **2009:** ImageNet database with millions of labeled images.
- **2012:** AlexNet → major breakthrough in computer vision.
- **Today:** Strong progress in NLP, computer vision, translation, and deep learning.

Real-World Applications

- **Advertising:** Personalized marketing.
- **Retail:** Supply chain optimization.
- **Transportation:** Self-driving cars, logistics.

- **Smart Homes:** Voice-enabled entertainment, security.
- **Healthcare:** Diagnostics, drug discovery.
- **Finance:** Algorithmic trading, fraud detection.
- **Government:** Smart cities, citizen services, threat detection.
- **Society:** Maps & navigation (Google Maps, Waze), dynamic pricing (Uber/Lyft), social media recommendations and ads.

Retrieving and Cleaning Data

Retrieving Data

Data sources	Definition	Read command	Write command
CSV files	Comma - separated values	`pd.read_csv("file.csv")`	`df.to_csv("file.csv", index = False)`
JSON files	Key–value / nested format	`pd.read_json("file.json")`	`df.to_json("file.json", orient="records")`
SQL databases	Relational tables	`pd.read_sql(query, conn)`	`df.to_sql("table", conn, if_exists="replace")`
NoSQL databases	Non-relational (JSON-like)	MongoDB: `collection.find()` (via PyMongo)	`collection.insert_many(df.to_dict("records"))`
APIs/ Cloud	Remote web data (JSON/CSV)	`pd.read_json(url)` or `pd.read_csv(url)`	Upload via API client

Data Cleaning Importance

- **Purpose:** Essential for reliable ML; prevents garbage-in, garbage-out.

- **Common Issues:** Duplicates, inconsistent text, missing values, outliers, poor data management.

Handling Duplicates

- Decide if duplicates are valid; filter carefully while retaining original data for analysis.

Handling Missing Values

- **Remove:** Drop rows (may lose information).
- **Impute:** Replace with mean/median (introduces uncertainty).
- **Mask:** Treat as a separate category (assumes similarity).

Handling Outliers

- **Definition:** Extreme values that skew predictions.
- **Identification:** Visualizations (histogram, boxplot), interquartile range.
- **Analysis:** Investigate before removing; some provide insights.

Residuals & Outlier Detection

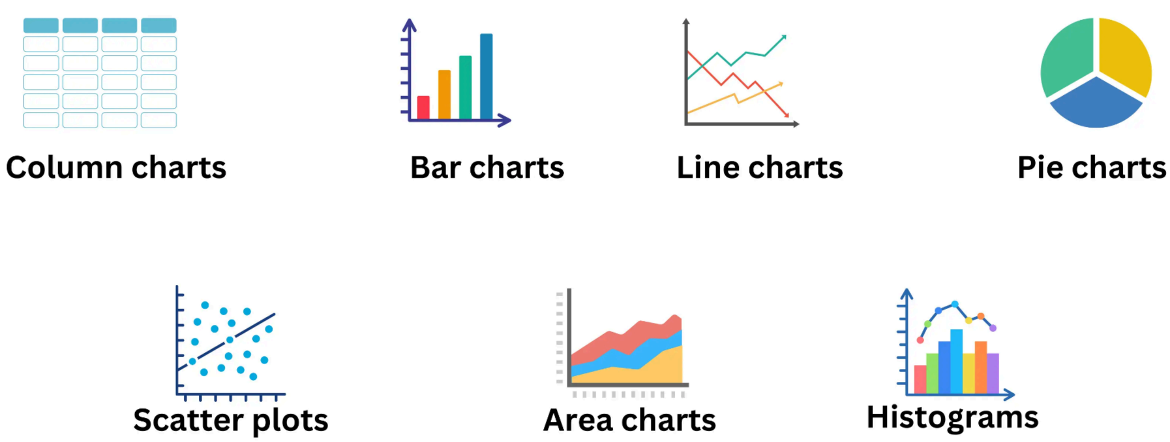
- **Residuals:** Difference between actual and predicted values; indicate model errors.
- **Standardized/Studentized residuals:** Assess impact on predictions.
- **Strategies:** Remove, transform, reassign, predict outlier values, or use robust models.

Exploratory Data Analysis and Feature Engineering

Exploratory Data Analysis (EDA)

- **Purpose:** Summarize dataset characteristics, identify patterns, trends, outliers, and need for cleaning or extra data.
- **Techniques:**
    - **Statistics:** Mean, median, min/max, correlations.
    - **Visualizations:** Histograms, scatter plots, box plots, pair plots, hexbin plots, facet grids.
- **Sampling:**
    - Random sampling for large datasets.
    - Stratified sampling to maintain proportion across categories.

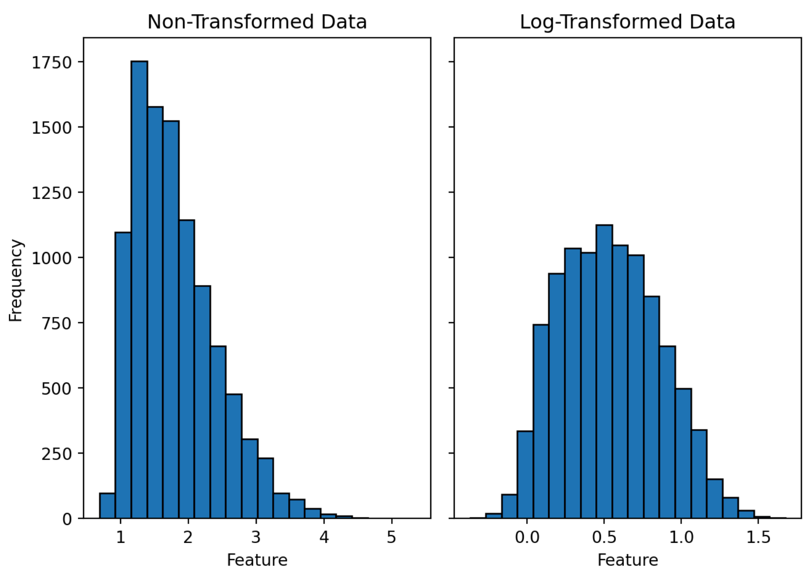
Python Visualization Libraries



- **Matplotlib:** Core plotting library; %matplotlib inline for notebooks.
- **Pandas:** Simplifies plotting on DataFrames.
- **Seaborn:** Built on Matplotlib; easier for aesthetically pleasing, statistical plots.
- **Techniques:** Scatter plots, histograms, boxplots, pair plots, hexbin, facet grids.

Feature Engineering & Variable Transformation

- **Purpose:** Optimize model performance, handle skewed distributions, outliers.
- **Transformations:**



  - **Log transformation:** Normalizes skewed data, handles diminishing returns (e.g., budget vs. box office revenue).
  - **Polynomial features:** Add flexibility (x^2, x^3, …) while keeping the model linear in parameters.
- **Encoding Categorical Features:**
  - **Nominal:** One-hot encoding.
  - **Binary:** 0/1 encoding,
  - **Ordinal:** Integer encoding while respecting order.
- **Feature Scaling:**
  - Standard Scaling (mean=0, std=1)

$$z = \frac{x - \mu}{\sigma}$$

  - Min-Max Scaling (0–1)

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- Robust Scaling (IQR-based).

$$X_{new} = \frac{X - X_{median}}{IQR}$$

- Important for distance-based algorithms like KNN; ensures meaningful comparisons.

Inferential Statistics and Hypothesis Testing

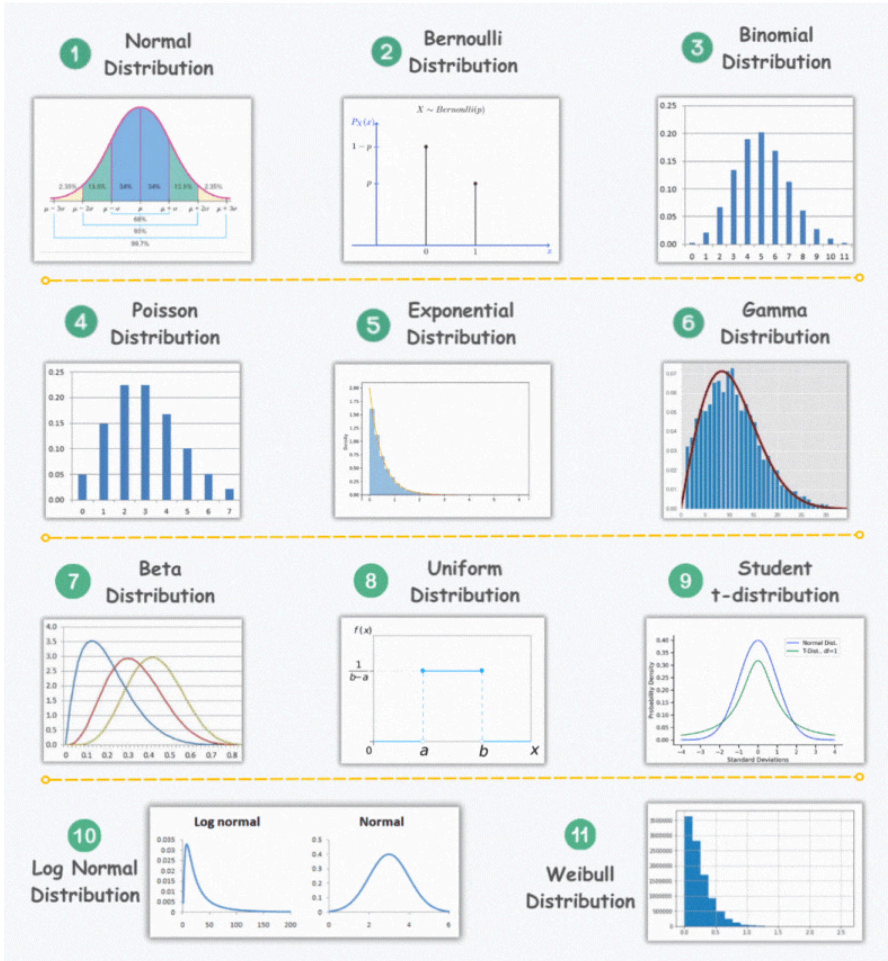
Estimation vs. Inference

- **Estimation:** Provides a point estimate of a parameter (e.g., sample mean = 20).
- **Inference:** Goes further by estimating the population distribution and attaching measures of uncertainty (e.g.,confidence intervals CI = 19%–21%).

Parametric vs. Non-Parametric Models

- **Parametric Models:** Assume a specific distribution, defined by finite parameters (e.g., linear regression, normal distribution).
- **Non-Parametric Models:** Make fewer assumptions, rely more heavily on observed data (e.g., histograms, kernel density).

Common Distributions



Distribution	Definition	Parameters	Example
Uniform	All outcomes equally likely	a,b (min, max)	Dice rolls, lottery
Normal (Gaussian)	Bell-shaped, around the mean	μ (mean), σ (std)	Heights, test scores
Log-Normal	Log values follow Normal	μ,σ (of log)	Income, stock prices
Exponential	Time between random events	λ (rate)	Waiting time for arrivals
Poisson	Event counts in fixed interval	λ (rate)	Number of emails per hour

Frequentist vs. Bayesian Statistics

- **Frequentist:** Relies on repeated sampling. Estimates probabilities without prior assumptions.
- **Bayesian:** Treats parameters as random variables. Combines prior beliefs with observed data → updates to posterior distribution.

Hypothesis Testing

- **Null Hypothesis (H_0):** No effect.
- **Alternative Hypothesis (H_1):** Effect exists.
- **Bayesian Approach:** Produces posterior probabilities instead of strict reject/accept decisions.

Type I and Type II Errors

	Actual - - True/ False	
Predicted - - Positive/ Negative	True Positive	False Positive (Type I)
	False Negative (Type II)	True Negative

- Note: The power of a test = 1 – P(Type II error).

Significance Levels & P-Values

- **Significance Level (α):** Threshold for rejecting H_0 (commonly 0.05).
- **P-Value:** Probability of observing data as extreme as current sample under H_0.
- **Bonferroni Correction:** Adjusts α when running multiple tests to reduce false positives.

Correlation vs. Causation

- **Correlation** :A statistical relationship when two variables change together (increase or decrease). It shows association.

- **Caussation:** Occurs when one variable directly causes a change in the other. It is a cause - effect relationship.
- **Confounding Variables:** A third factor may drive both variables.
- **Spurious Correlations:** Random coincidences.
- **Business Caution:** Use correlation for prediction, but never assume direct cause without deeper analysis.