**Module 1: A Brief History of Modern AI and Its Applications**

**1.1. Artificial Intelligence (AI)**

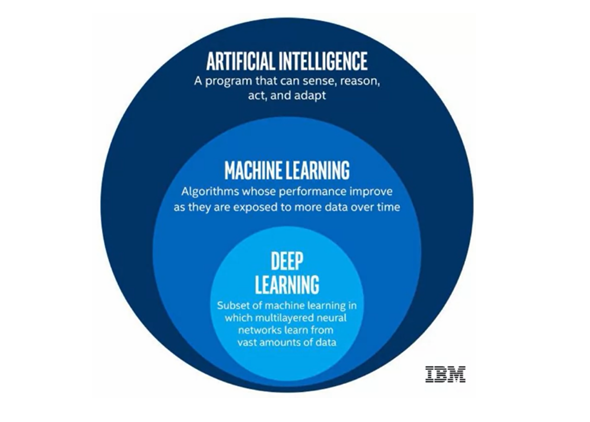
* Định nghĩa: AI là ngành khoa học và kỹ thuật tạo ra hệ thống có khả năng thực hiện những công việc vốn cần trí tuệ con người (như suy luận, học tập, nhận dạng).
* Giải thích: Máy tính thường chỉ làm theo lệnh, nhưng AI có thể “thích ứng” với dữ liệu và môi trường.
* Ví dụ: Google Translate, hệ thống gợi ý Netflix.
* **Key takeaway**: AI = mục tiêu tổng quát về trí thông minh nhân tạo.

**1.2. Machine Learning (ML)**

* Định nghĩa: ML là phương pháp của AI cho phép máy tính học từ dữ liệu thay vì lập trình quy tắc cứng.
* Giải thích: Thay vì viết “nếu email chứa chữ *lottery* thì spam”, ta đưa máy hàng ngàn email đã dán nhãn để nó học.
* Ví dụ: Bộ lọc spam Gmail.
* **Key takeaway**: ML = học từ dữ liệu.

**1.3. Deep Learning (DL)**

* Định nghĩa: DL là nhánh con của ML dựa trên mạng nơ-ron nhiều tầng (deep neural networks).
* Giải thích: DL tự động học đặc trưng (features) từ dữ liệu, thay vì con người phải thiết kế thủ công.
* Ví dụ: Nhận diện khuôn mặt Facebook.
* **Key takeaway**: DL = cách mạnh nhất hiện nay để đạt ML.

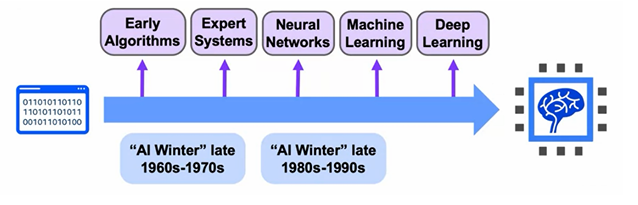


**1.4. Lịch sử AI**

* 1950: Turing Test.
* 1956: Dartmouth – khai sinh AI.
* 1957: Perceptron.
* 1966, 1973: AI Winters (thất vọng do kỳ vọng cao).
* 1980s: Expert Systems, backpropagation (1986).
* 2012+: Deep Learning bùng nổ nhờ GPU + Big Data.

**Key takeaway chung:** Hiểu lịch sử = hiểu vì sao ML/DL bùng nổ ngày nay.

Timeline bên dưới:



**Module 2: Retrieving and Cleaning Data**

**2.1. Retrieving Data**

* Định nghĩa: Retrieving data = quá trình lấy dữ liệu từ nguồn (database, file, API).
* Giải thích: ML không quan tâm nguồn, nhưng ta phải đưa về dạng bảng (DataFrame).

**Lý thuyết:** Dữ liệu có 3 loại:

* **Có cấu trúc:** bảng SQL, CSV.
* **Bán cấu trúc:** JSON, XML.
* **Không cấu trúc:** văn bản, ảnh, video.
* Ví dụ:

SQL:



CSV:



* **Key takeaway:** Nguồn dữ liệu khác nhau → cuối cùng vẫn phải quy về bảng sạch.

**2.2. Missing Data**

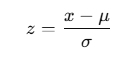
* **Định nghĩa:** Missing data = giá trị bị thiếu trong bảng. Trường hợp dữ liệu không có giá trị (NaN, None).
* **Giải thích:** Nếu không xử lý, mô hình sẽ “crash” hoặc học sai.
* **Công thức:** không có công thức cố định, nhưng dùng imputation: mean, median, most frequent.
* **Ví dụ:** Điền median bằng SimpleImputer(strategy="median") .
* **Key takeaway:** Thiếu ít → drop, thiếu nhiều → impute.

**Cách xử lý:**

* Xoá (drop).
* Điền (impute): mean, median, mode, hoặc mô hình dự đoán.
* Tạo cờ “missing” (biến nhị phân).

**2.3. Outliers**

* **Định nghĩa:** Outlier = giá trị khác biệt cực lớn so với phần còn lại.
* **Giải thích:** Có thể do lỗi nhập hoặc bản chất dữ liệu hiếm.
* **Công thức (Z-score):**



* **Ký hiệu:**
  + x: giá trị quan sát.
  + μ: mean.
  + σ: std dev ( độ lệch chuẩn )
* **Ví dụ:** Trong [1,2,2,3,100], số 100 là outlier (z-score cao).
* **Key takeaway:** Không xoá outlier bừa, cần ngữ cảnh.

**2.4. Scaling**

* **Định nghĩa:** Scaling = biến đổi dữ liệu về cùng khoảng/đơn vị.
* **Lý thuyết:** nhiều mô hình (kNN, SVM, NN) nhạy với thang đo → cần scale.
* **Giải thích:** Tránh biến lớn “áp đảo” biến nhỏ.
* **Công thức (Min-Max):**



* **Ví dụ:** Lương (triệu) 20–50 → scale về 0–1.
* **Key takeaway:** Scale để công bằng giữa biến.

**Module 3: Exploratory Data Analysis (EDA) & Feature Engineering**

**3.1. EDA**

* **Định nghĩa:** EDA = phân tích dữ liệu khám phá để hiểu phân phối, mẫu, quan hệ.
* **Thành phần:**

Thống kê mô tả: mean, median, variance, skewness, kurtosis.

Biểu đồ: histogram, boxplot, scatter plot, heatmap.

* **Vai trò:** phát hiện vấn đề (missing, outlier, phân phối lệch), gợi ý hướng xử lý.
* **Giải thích:** Như soi gương dữ liệu trước khi modeling.
* **Ví dụ:** Histogram cho phân phối, boxplot cho ngoại lai, heatmap cho tương quan.
* **Key takeaway:** EDA = bước bắt buộc trước modeling.

**3.2. Feature Engineering**

* **Định nghĩa:** FE = quá trình biến đổi, tạo đặc trưng mới để cải thiện model.
* **Kỹ thuật phổ biến:**
  + Log transform: 
  + Polynomial: thêm .
  + Encoding: one-hot cho nominal, ordinal encoding cho thứ tự.
  + Feature selection: lọc biến ít ý nghĩa.
  + PCA: giảm chiều dữ liệu.
* **Giải thích:** Feature tốt có thể quan trọng hơn model.
* **Key takeaway:** “Garbage in, garbage out” → FE giúp dữ liệu giàu thông tin hơn.

**Module 4: Inferential Statistics & Hypothesis Testing**

**4.1. Estimation**

* **Định nghĩa:** Ước lượng = dùng mẫu để suy ra tham số quần thể.
* **Lý thuyết:** Định lý giới hạn trung tâm → mẫu đủ lớn sẽ có mean gần chuẩn.
* **Công thức mean mẫu:**



* **Ký hiệu:** nnn = số mẫu, xix\_ixi​ = từng giá trị.
* **Ví dụ:** Chiều cao 10 người = 165–175cm → 

**4.2. Confidence Interval**



**Ký hiệu:**

* xˉ: mean mẫu.
* σ : std dev. (Độ lệch chuẩn)
* n: số mẫu.
* zα/2​: giá trị tới hạn (1.96 cho 95%).

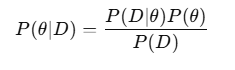
**Ý nghĩa:** Khoảng giá trị mà tham số thật có thể nằm trong.

**Key takeaway:** CI giúp định lượng độ chắc chắn.

**4.3. Hypothesis Testing**

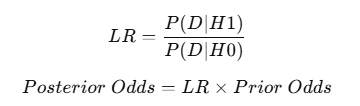
* **Định nghĩa:** kiểm định giả thuyết = cách ra quyết định dựa trên dữ liệu.
* **H0:** giả thuyết gốc (không có khác biệt).
* **H1:** giả thuyết thay thế.
* **p-value:** xác suất quan sát dữ liệu nếu H0 đúng.
* **Key takeaway:** p nhỏ → dữ liệu “không hợp” với H0 → bác bỏ.

**4.4. Bayesian Inference**



* **Ký hiệu:**
  + P(θ): prior (niềm tin ban đầu).
  + P(D∣θ): likelihood (dữ liệu).
  + P(θ∣D): posterior (niềm tin mới).
* **Giải thích:** Posterior = niềm tin cập nhật sau khi có dữ liệu.
* **Key takeaway:** Bayes = framework tự nhiên để học từ dữ liệu.

**4.5. Likelihood Ratio**

****

* **Ý nghĩa:** đo mức độ dữ liệu nghiêng về H1 hay H0.
* **Ví dụ:** nếu prior = 0.2, LR=5 → posterior ~0.56.
* **Key takeaway:** LR = công cụ “bẻ cán cân” giữa H0 & H1.

**Master Takeaways**

1. AI – ML – DL: mục tiêu → phương pháp → công cụ mạnh.
2. 80% thời gian ML = xử lý dữ liệu.
3. EDA/FE quan trọng hơn chọn model phức tạp.
4. Thống kê suy luận giúp tránh kết luận sai.
5. Bayes & LR = công cụ update niềm tin.
6. Project = bằng chứng “tôi hiểu và tôi làm được”.