Introduction to machine learning

- **Outline:**

  1. What is machine learning?

  2. Types of machine learning

  3. Regression, classification and clustering

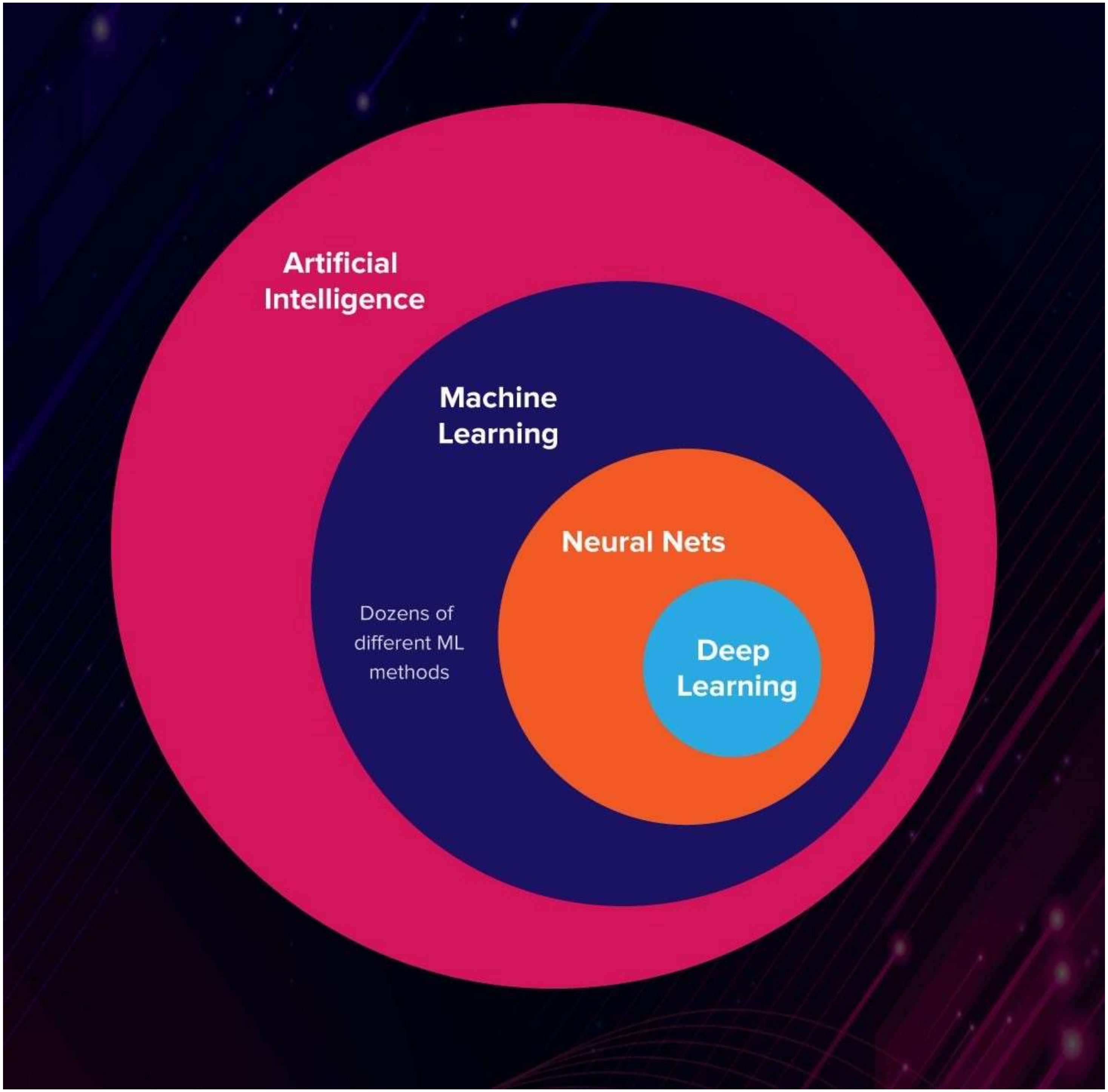
Introduction to machine learning

- **Outline:**

  1. What is machine learning?

  2. Types of machine learning

  3. Regression, classification and clustering

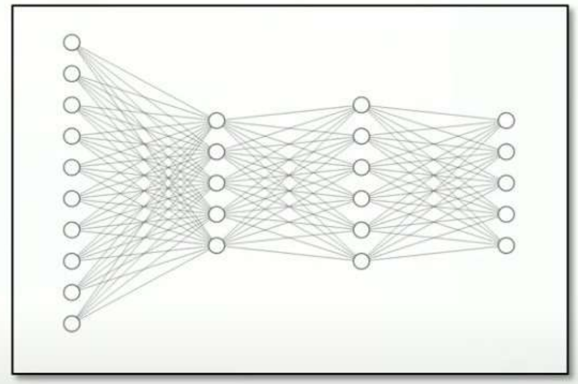

Artificial Intelligence
Machine Learning
Dozens of different ML methods
Neural Nets
Deep Learning

Three forces brought AI to life

Big Data

Compute Power

Machine Learning Algorithms



Arthur Lee Samuel

Born	December 5, 1901 Emporia, Kansas
Died	July 29, 1990 (aged 88) Stanford, California

Arthur Lee Samuel *(1959)*

Machine Learning *the "field of study that gives computers the ability to learn without being explicitly programmed".*

The truth

- Machines don't learn

(inputs are slightly distorted → output can be wrong)

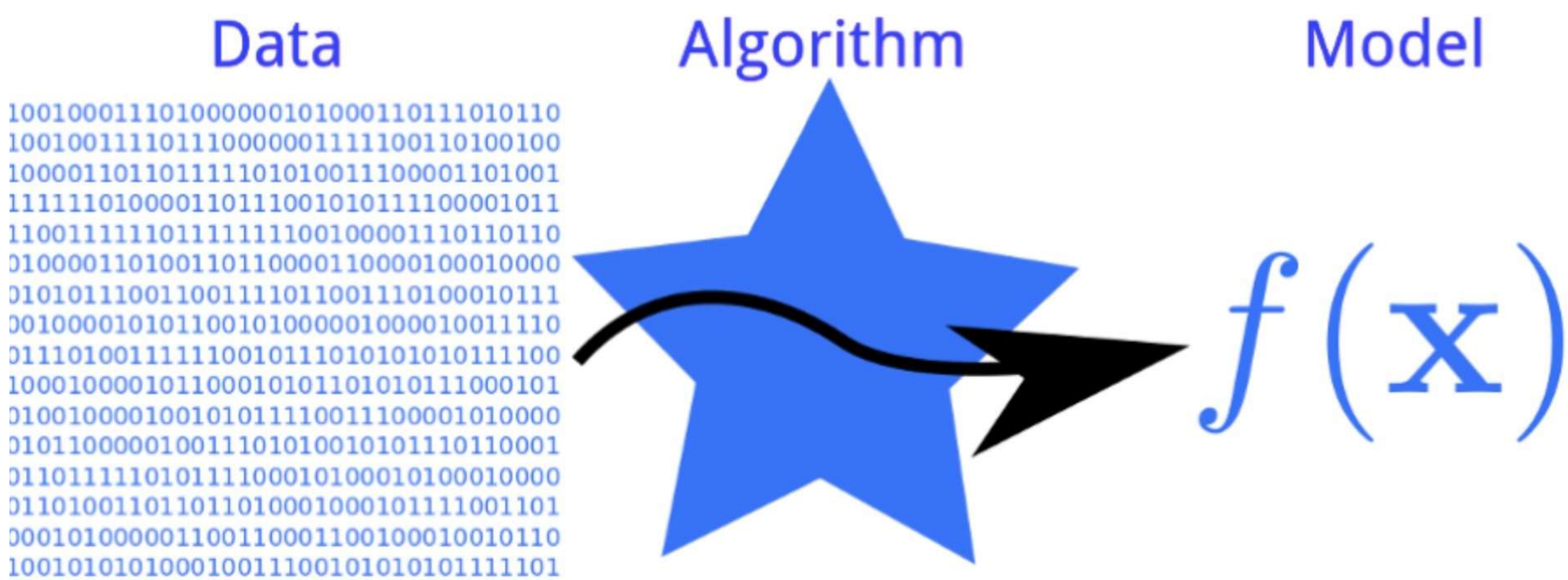
- In 1959, Arthur Samuel (American pioneer in computer gaming and AI) coined the term "machine learning" while at IBM
- In 1960, IBM used new cool term "machine learning" to attract clients and talented employees

What is machine learning?

- The process of solving a practical problem by:
  - Gathering a dataset
  - Based on that dataset building a statistical model which is assumed to be used somehow to solve the practical problem.

Machine learning algorithm

- Finding a mathematical formula based on a collection of inputs (i.e, “training data”)
- Applying formula to training inputs → produces the desired outputs.
- Applying formula to novel inputs → generates the correct outputs.
- New inputs come from the same or a similar statistical distribution.



Introduction

Scope & Motivation

- **Machine Learning: What?**
  - **Introductory example**: *When to play golf?*
    - **Collect data**
      - Consulting experts (e.g., golf players)
      - Watching players
      - Collecting weather data, etc.

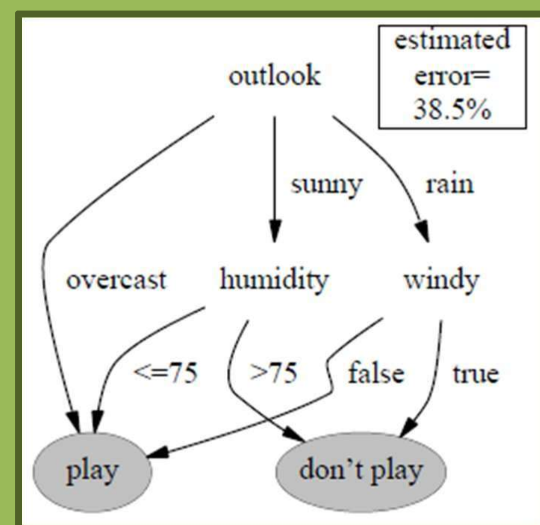
From [Menzies, 2002]

```
golf.names
Play, Don't Play.

outlook: sunny, overcast, rain.
temperature: continuous.
humidity: continuous.
windy: true, false.

golf.data
sunny, 85, 85, false, Don't Play
sunny, 80, 90, true, Don't Play
overcast, 83, 88, false, Play
rain, 70, 96, false, Play
rain, 68, 80, false, Play
rain, 65, 70, true, Don't Play
overcast, 64, 65, true, Play
sunny, 72, 95, false, Don't Play
sunny, 69, 70, false, Play
rain, 75, 80, false, Play
sunny, 75, 70, true, Play
overcast, 72, 90, true, Play
overcast, 81, 75, false, Play
rain, 71, 96, true, Don't Play
```

Introduction

Scope & Motivation

- **Machine Learning: What?**
  - **Introductory example**:
    When to play golf?
    - **Create a model** using one/ several classifiers
      - E.g., decision trees
    - **Evaluate model**
      - E.g., classification error



There's a **lot more to the machine learning process**... We're just getting started ☺

Introduction

Scope & Motivation

- **Machine Learning: What?**
  - A **branch of Artificial Intelligence** (AI)

> "***Machine learning (ML)*** *is concerned with the design and development of algorithms and techniques that* ***allow computers to "learn"****. The* ***major focus*** *of ML research is to* ***extract information from data automatically****, by computational and statistical methods. It is thus closely related to data mining and statistics".*
> [Svensson and Söderberg, 2008]

Traditional programming vs. ML

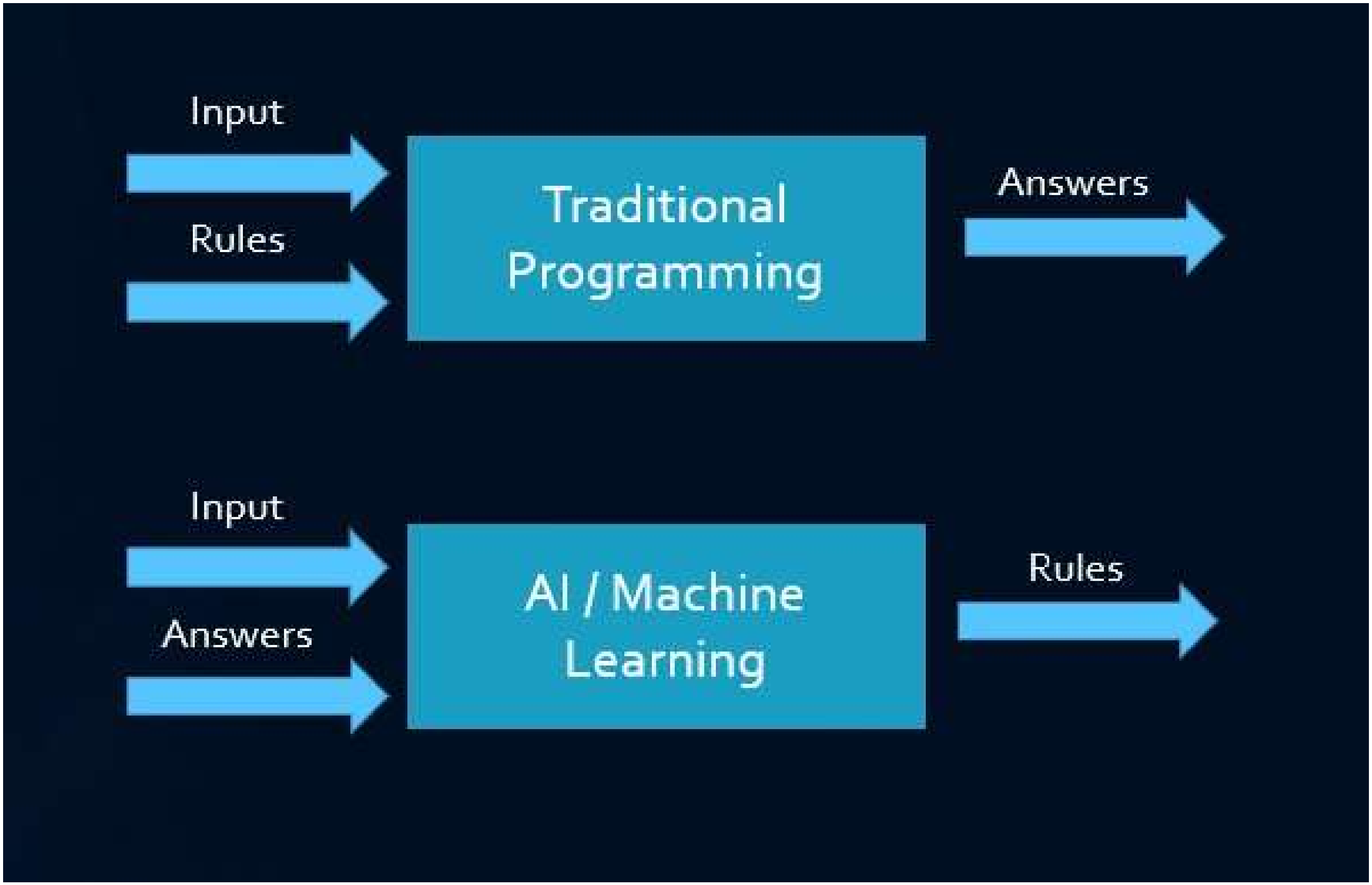
Traditional programming

- Feeding computer with rules
- Computer utilizes computing
- Coming up with answers

Machine learning

- Feeding computer with huge amount of data
- Computer processes the data
- Coming up with trained model that can solve the unseen problems of the real world

Traditional programming vs. ML



Example of quadratic equation solving

- Given $a \neq 0, b, c$
- $ax^2 + bx + c = 0$
- Find x?

- **Solving:**

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Example of quadratic equation solving

- Given $y = ax^2 + bx + c$

when x=	0	y=	1.1
when x=	1	y=	5.9
when x=	2	y=	16.8
when x=	3	y=	33.9

- **Solve for** x = **8**?

- Using ML → $y = 3.078x^2 + 1.701x + 1.106$

  Expected 1.1,5.9,16.8,33.9
  Got 1.106,5.884,16.817,33.906

- x = 8 -→ y =?

Bài tập

- Lập trình tính căn bậc 2 của một số S cho trước theo công thức Newton.

$$x_{n+1} = \frac{1}{2} * \left(x_n + \frac{S}{x_n}\right)$$

- n là số vòng lặp
- Cách ước lượng x_0:

➢ Biểu diễn S dưới dạng: $S = a * 10^{2n} (a > 0)$

➢ Tính xấp xỉ căn bậc 2 của S

$$2 * 10^n \quad (a < 10)$$
$$6 * 10^n \quad (a >= 10)$$

Kiểm tra

- Tính căn bậc 2 của số S = 785286
- Cho biết $x'' = 600$

When ML is used?

A real-world problem is a candidate for the application of machine learning if -

1. Historical data exists in a huge amount
2. A pattern exists in the data
3. Extremely hard to pin down a solution mathematically

Introduction to machine learning

- **Duration:** 2 hrs
- **Outline:**

1. What is machine learning?
2. **Types of machine learning**
3. Regression, classification and clustering

Types of ML

- **Supervised learning:** training data includes desired outputs
  - Classification
  - Regression
- **Unsupervised learning:** training data does not include desired outputs
  - Clustering
  - Association
- **Reinforcement learning:** rewards from a sequence of actions

Supervised learning

- **Dataset:** set of labeled examples (labeled data)

$$\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$$

Feature vector — Label

[height weight gender age] — {normal, thin, fat}

- **Goal:** to produce a model that takes a **feature vector** as input and outputs the label for this feature vector.

Unsupervised learning

- **Dataset:** set of unlabeled examples

$$\{\mathbf{x}_i\}_{i=1}^{N}$$

Feature vector

[height weight gender age]

- **Goal:** to produce a model that takes a feature vector as input and transforms it to another vector or to a value used to solve a practical problem.

- Ex: clustering, dimensionality reduction, outlier detection

Reinforcement

- The machine is capable of perceiving the state of the environment around as a feature vector.
- The machine can execute actions in every state.
- Different action brings different rewards → move the machine to the other state
- **Goal:** to learn a policy (function f ~ model in supervised learning) that takes the feature vector of a state as input and outputs an optimal action (=action maximizes the expected average reward)
- **Applications:** sequential decision making, long-term goal. Ex: game playing, robotics, resource management, logistics.

Introduction to machine learning

- **Duration:** 2 hrs
- **Outline:**

1. What is machine learning?
2. Types of machine learning
3. **Regression, classification and clustering**

General introduction

- Based on the type of output variable and the type of problem to be addressed, ML algorithms are categorized into:

  - Classification – supervised learning
  - Regression – supervised learning
  - Clustering – unsupervised learning

Classification

- Automatically assigning a **label** to an **unlabeled example**
- Taking a collection of **labeled examples** as inputs and producing a **model** that can take an unlabeled example as input and output a label (or a number label)
- Label = member of a finite set of **classes**
- **Ex:** Covid-19 detection

Clustering

- Automatically assigning a **label** to examples given an unlabeled dataset
- Dividing the examples into a number of **groups** such that examples in the same groups are more similar to other examples in the same group than those in other groups.
- **Goal:** to segregate groups with similar features and assign them into **clusters**.

Clustering

Regression

- Automatically predicting a **real-valued label** (i.e., *target*) given an **unlabeled example**.
- Taking a collection of **labeled examples** as inputs and producing a **model** that can take an unlabeled example as input and output a target.
- **Ex**: estimating house price based on house features [area, # bedrooms, location, etc]

ML Applications

Software Engineering (SE)

- **Why?**
  - **Simplify software development**
    - "Construction of systems that support classification, prediction, diagnosis, planning, monitoring, requirements engineering, validation, and maintenance"[Menzies, 2002]
      - E.g., Software quality, size and cost prediction, etc.

ML Applications

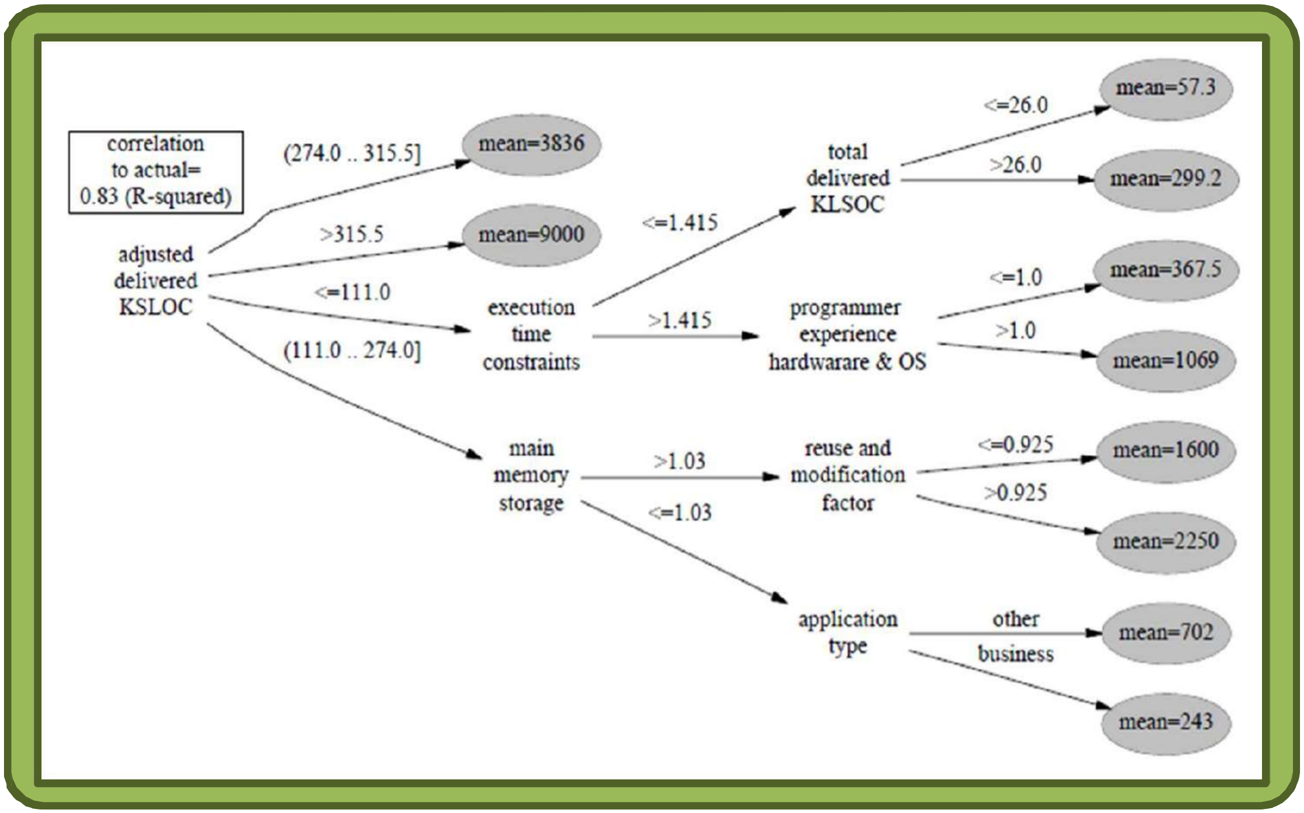
Software Engineering

- **How?**
  - Data collection
    - Company’s past projects, public benchmarks, etc.
  - Methodologies
    - Many of the practical SE applications of machine learning use **decision tree learners** [Menzies, 2002]
      - Knowledge bust be **explicit**

ML Applications

Software Engineering

- **Example: predicting software development time at TRW Aerospace (cited in [Menzies, 2002])**



From [Menzies, 2002]

ML Applications

Software Engineering

- **Example: predicting software development time at TRW Aerospace (cited in [Menzies, 2002])**
  - Developed by Barry W. Boehm, in 1981, when he was TRW's director of Software Research and Technology
  - Data collection
    - **COCOMO-I (Constructive Cost Model) database**: data from **63 software projects** at TRW
      - Projects ranging in size from **2,000 to 100,000 lines of code**, and programming languages ranging from **assembly to PL/I**.
      - Projects were based on the **waterfall model**

ML Applications

Software Engineering

- **Example: predicting software development time at TRW Aerospace (cited in [Menzies, 2002])**
  - Feature Extraction
    - Example of features
      - Estimated thousand source lines of code (KSLOC), complexity, memory constraints, personnel experience (SE capability, applications experience), …
      - Of the 40 attributes in the dataset, only six were deemed significant by the learner
      - Output: software development time (in person months)
  - Methodology
    - CART tree learner

ML Applications

Software Engineering

- **Other examples**
  - Software quality, size and cost prediction, etc.
  - Predicting fault-prone modules
  - ...

ML Applications

Software Engineering

- **Domain specificities**
  - **Data starvation**
    - Particularly acute for newer, smaller software companies
      - Lack the resources to collect and maintain such data
    - → **Knowledge farming**: **farm** knowledge by growing datasets from domain models [Menzies, 2002] (not discussed in this course)
      - Use of **domain models** as a ***seed*** to ***grow*** **data sets** using exhaustive or monte carlo simulations.
      - Then, mine data with machine learning
      - → Out of the scope of this course

ML Applications

Comm. Networks

- **Why?**
  - **Implementation of “intelligent” network protocols**
    - E.g., intelligent routing mechanisms, network anomaly detection, reliability assessment of communication networks, link quality prediction in wireless sensor networks (WSN), etc.

ML Applications

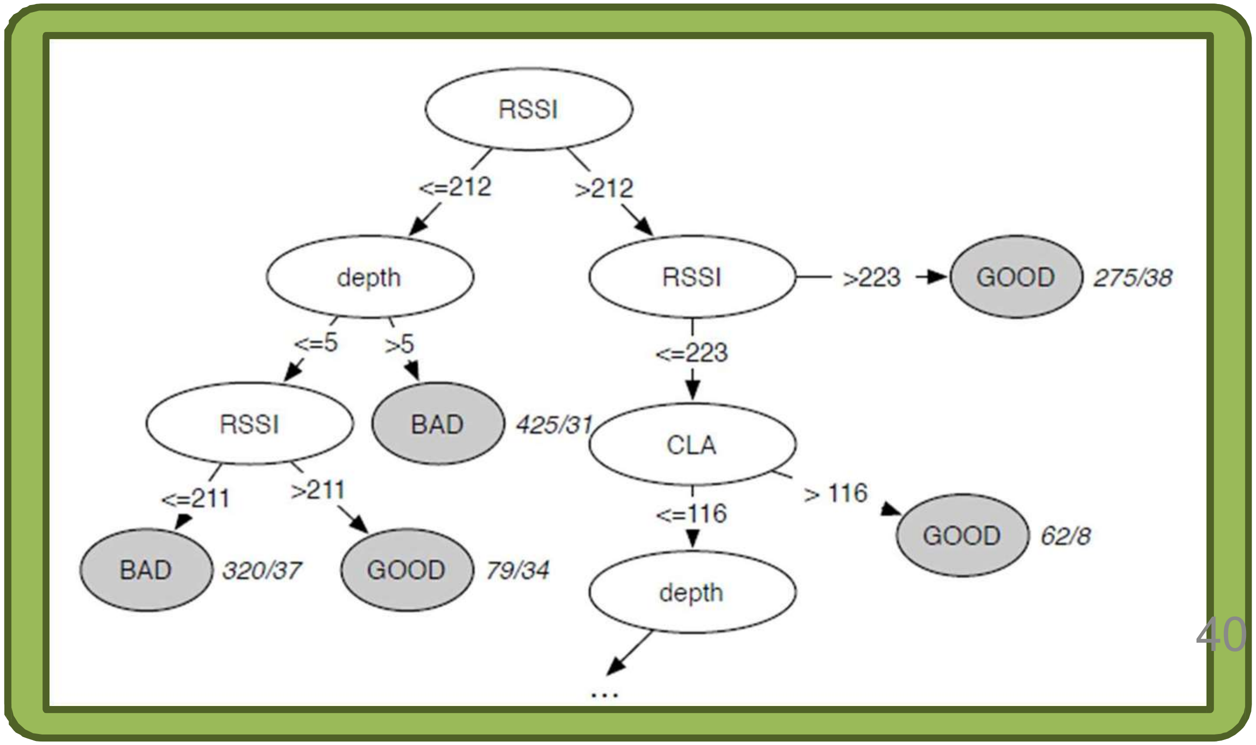
Comm. Networks

- **How?**
  - Data collection
    - Features typically collected at node links
    - Data often manually or semi-automatically annotated (e.g., link quality)
  - Methodologies
    - Both accuracy and interpretability are aimed at → rule induction, decision trees and functional classification algorithms (e.g., SVM) are often useful

ML Applications

Comm. Networks

- **Example: MetricMap: link quality estimation in WSN (cited in [Förster and Murphy, 2010])**



ML Applications

Comm. Networks

- **Example:** **MetricMap: link quality estimation in WSN** **(cited in [Förster and Murphy, 2010])**
  - Developed by Wang et al. at Princeton University in 2006
  - Data collection
    - MistLab sensor network testbed
    - Acquisition of link samples and desired features available at the nodes
    - Link annotation: good or bad, according to its Link Quality Indication (LQI) value (indicator of the strength and quality of a received packet, introduced in the 802.15.4 standard)

ML Applications

Comm. Networks

- **Example: MetricMap: link quality estimation in WSN (cited in [Förster and Murphy, 2010])**
  - Feature Extraction
    - Locally available information, e.g., RSSI (received signal strength indication) levels of incoming packets, CLA (channel load assessment), etc.
  - Methodologies
    - Classification: decision trees (C4.5), using the WEKA workbench
  - Evaluation
    - Algorithm outperformed standard routing protocols in terms of delivery rate and fairness

ML Applications

Comm. Networks

- **Other examples**
  - Intelligent routing mechanisms
  - Network anomaly detection
  - Reliability assessment of communication networks
  - ...

Bài tập về nhà (cá nhân)

- Tìm hiểu về công cụ Teachable Machine
  https://teachablemachine.withgoogle.com
- 1. Áp dụng Teachable Machine xây dựng hệ thống nhận dạng
- 2. Tìm hiểu công cụ khác khác tương tự như Teachable Machine

  Yêu cầu:

  Chọn 2-3 mẫu

- Tự thuyết trình kết hợp mô tả hệ thống, quay video, up youtube, nộp link vào assignment đã tạo sẵn ở MS Teams