CL205: Artificial Intelligence and Data Science Overview

Mani Bhushan

Chemical Engineering
IIT Bombay

29 July 2024

Welcome

The most important message for this course:

- "Tou wa ittoki no haji, shiranu wa matsudai no haji": Japanese Saying
  - "To ask a question is a moment's shame, to remain ignorant is a shame forever."
- "The only stupid question is the one not asked."
- So, keep the questions flowing.

Revolutions

- French Revolution: Late 1700s
- American Revolution: Late 1700s
- Indian Freedom Revolution: Mid 1800, Mid 1900
- ⋮
- Green Revolution: Foodgrains, 1960s ...
- White Revolution: Milk, 1970s ...
- Industrial Revolution: Mid 1700s to early 1800s

An Ongoing Industrial Revolution

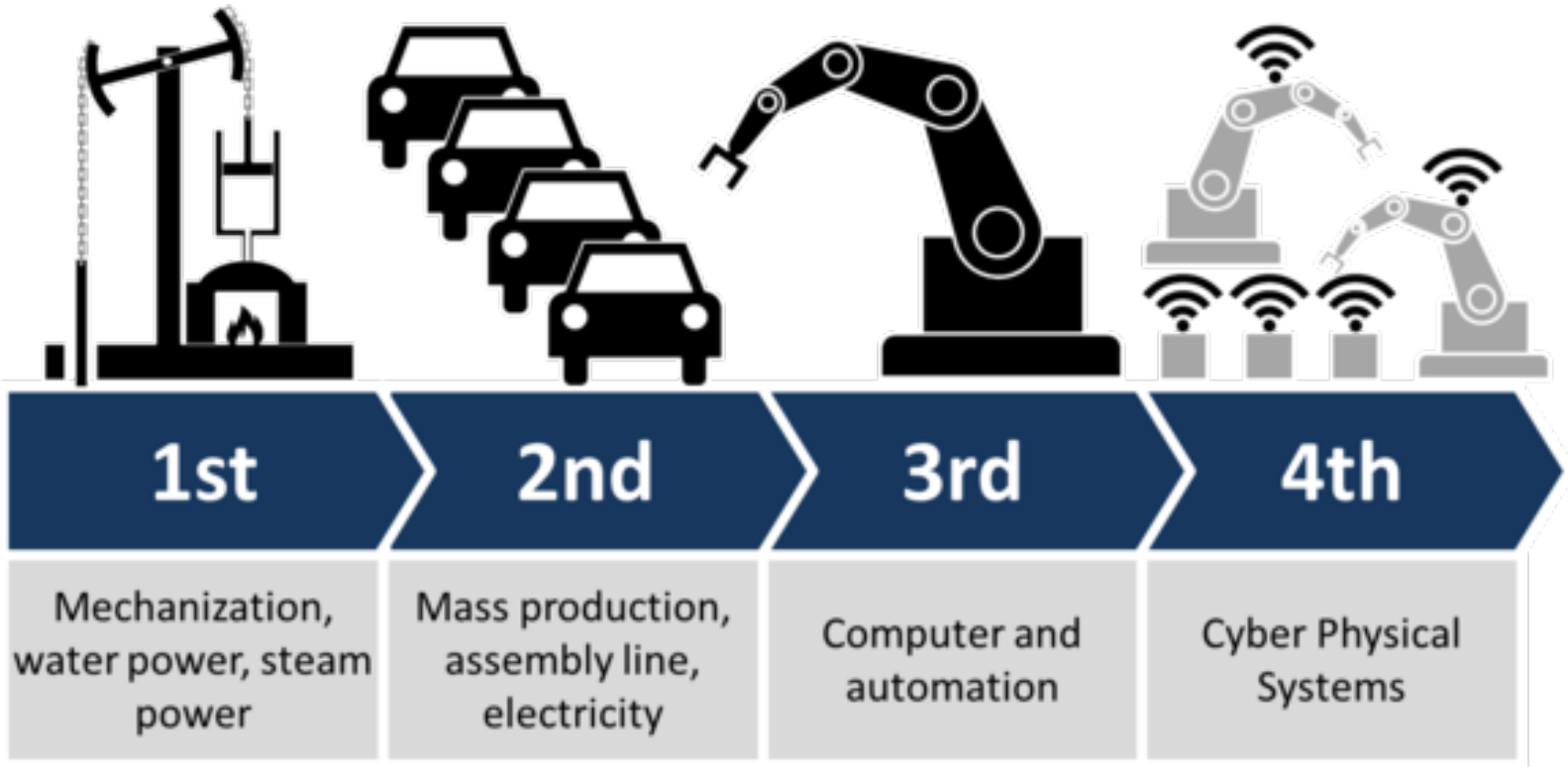


Figure: https://fr.wikipedia.org/wiki/Industrie_4.0

- We are in the middle of an Industrial Revolution: Industry 4.0

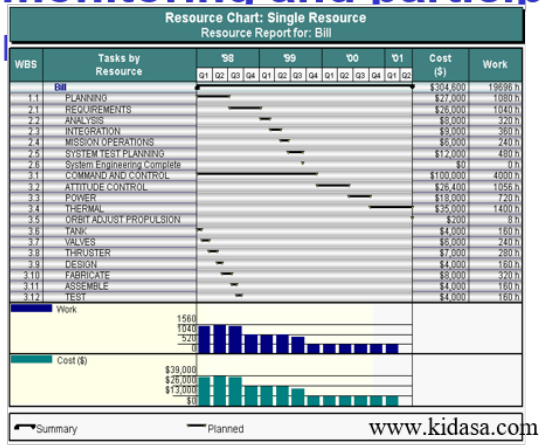
Industry 4.0

- Characterized by smart, efficient production
- Fueled by (industrial) internet of things, Digitization, and Artificial Intelligence
- Increasing automation:
  - Autonomous decision making as well as opposed to simply automating routine (repeatable) tasks

Industrial Operation Earlier

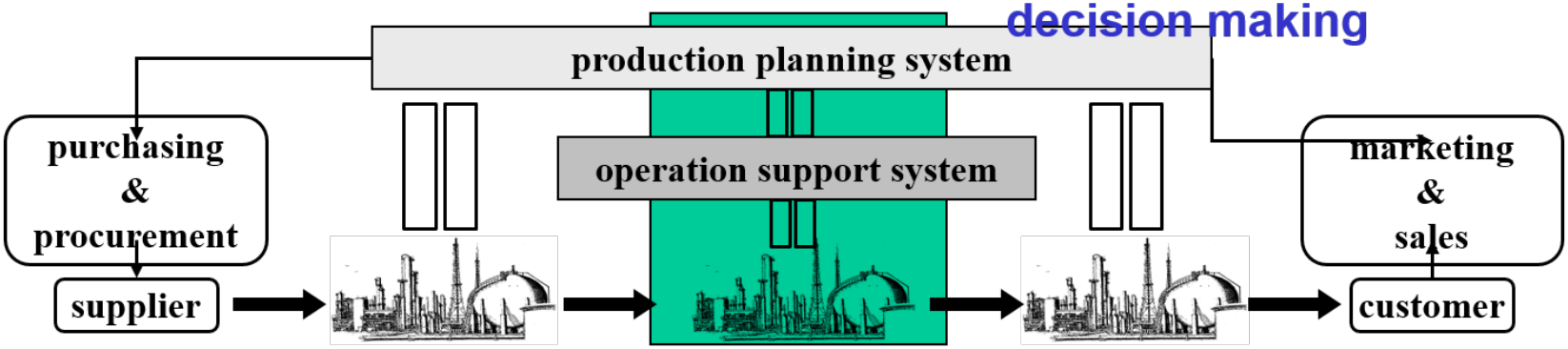
Current View of Control Room in a Manufacturing Facility

digital control systems and multivariable predictive control: enterprise wide monitoring and participative decision

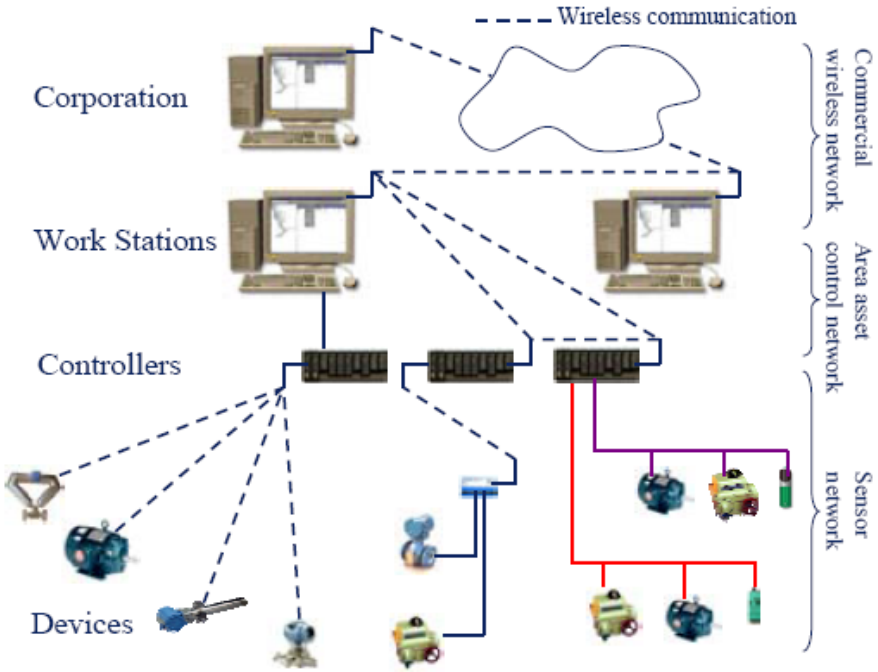


production planning and scheduling

supply chain conscious decision making



The Future



- Industrial internet of things (IIOT): Cloud, sensors, transmitters, fieldbus
- Digital twin
- Challenges:
  - Security: cyber threats
  - Robustness: field interference, high power signals in inflammable regions
  - Reliability and power

Automation Examples: Modern Day

- What's common to[1]:
  - Alphabet- Waymo
  - General Motors- Cruise
  - Microsoft- Volkswagon
  - Uber Technologies- Motional
  - Tesla
- *Self driving cars*

[1] https://yoshalawfirm.com/blog/5-top-self-driving-car-manufacturers/

It is happening in our discipline [2] as well ...

- Yokogawa Electric Corporation- JSR Corporation
- AI was used to autonomously run a chemical plant for 35 days
- Using reinforcement learning based control

[2] https://www.yokogawa.com/in/news/press-releases/2022/2022-03-22/

But Comes with Perils[3]

- 2001, Waste Management System, Maroochy Shire, Queensland, Australia
- An insider released 265,000 gallons of sewage on the Maroochy Shire
  - Raw sewage spilled into local parks, rivers, residential areas, hotel grounds
- Insider an engineer for firm that installed supervisory control and data acquisition (SCADA) radio-controlled sewage equipment
- Issued radio commands to sewage equipment he helped install
- Caught due to a traffic violation
- Not the usual IT (virus, server attack) type intrusion

[3] https://www.industrialcybersecuritypulse.com/facilities/throwback-attack-an-insider-releases-265000-gallons-of-sewage-on-the-maroochy-shire/

Don't Believe Blindly ...

- Lawyer cites fake cases generated by xxxx in legal brief[4]
- xxxx: ChatGPT
- Requirement of knowing the subject hasn't gone away
  - Ability to distinguish meaningful answers from garbage solutions

[4] www.legaldive.com/news/chatgpt-fake-legal-cases-generative-ai-hallucinations/651557/

So What is AI and Data Science?

Statutory Warning

Confusion Ahead

Areas of Interest

- Machine Learning
- Artificial Intelligence
- Data Science
- Data Analytics

Machine Learning

Figure: https://www.istockphoto.com/photos/machine-learning

- Focus on getting machines to learn
- Use of data to enable machines to learn tasks (classification, regression, etc.) across various domains
- Can be considered subset of artificial intelligence

Artificial Intelligence

- Create machines which think and act like humans: as intelligent as humans
- Sentient machine: "having the ability to use your senses to see and feel"

Figure: `https://en.wikipedia.org/wiki/Terminator_2:_Judgment_Day`

What is Artificial Intelligence

- Not just data, but also use of expert systems, domain knowledge, rule bases, to create artificial (machine) intelligence
- Used more in context of getting machines to behave like humans: Speech, emotions, actions,...
- The holy grail for a long time: many movies
  - Terminator series
  - Enthiran (Robot): Rajnikant
  - AI: Boy in the movie (by Steven Spielberg)

Turing Test for AI

- Alan Turing's (1950) definition of intelligence:
  A computer (machine) is intelligent if after carrying out a conversation (without seeing the other person/computer), a human interrogator cannot say whether the person he/she is conversing with is human or machine.
- Turing test: Can't distinguish between a human and a machine
- Some interesting AI examples:
  `https://dataconomy.com/2021/03/which-ai-closest-passing-turing-test/`

Passing Turing Test not Necessary

- Alternative View of AI:
  - Systems that act or think rationally and not necessarily mimic a human
  - Involves mathematics and engineering
  - Focus on studying underlying principles of intelligence
  - Modern focus
- Not necessary to pass Turing test
- Similar to quest for "artificial flight"
  - Understand aerodynamics, not imitate birds

Data Science

- The scientific basis of approaches
- Field of study that aims to use scientific approach to extract meaning and insights from data
- Catalyst Example

Catalyst A Yield	Catalyst B Yield
78.1	77.1
82.0	83.0
93.4	78.4
82.4	81.5
81.4	86.3
88.3	87.4
91.0	90.5
95.2	92.5

- Being able to say that yield with Catalyst A is better than with B? and maybe identify reason: *Obtain insights from data*

Probability and Statistics

- Probability and Statistics provide a foundation for data science and AI
- Involves dealing with incomplete, noisy, and erroneous datasets
- Establishing models in presence of uncertainty
- Making predictions/deductions in presence of uncertainty
- Data Science: Linear Algebra, Optimization, also relevant

Data Analysis

Figure: https://www.heavy.ai/technical-glossary/graphical-representation

- Data Analytics:
  - Sometimes used in a limited sense of analysis- summary, visualization.
  - Not necessarily training computers/extracting models.

Pictorially

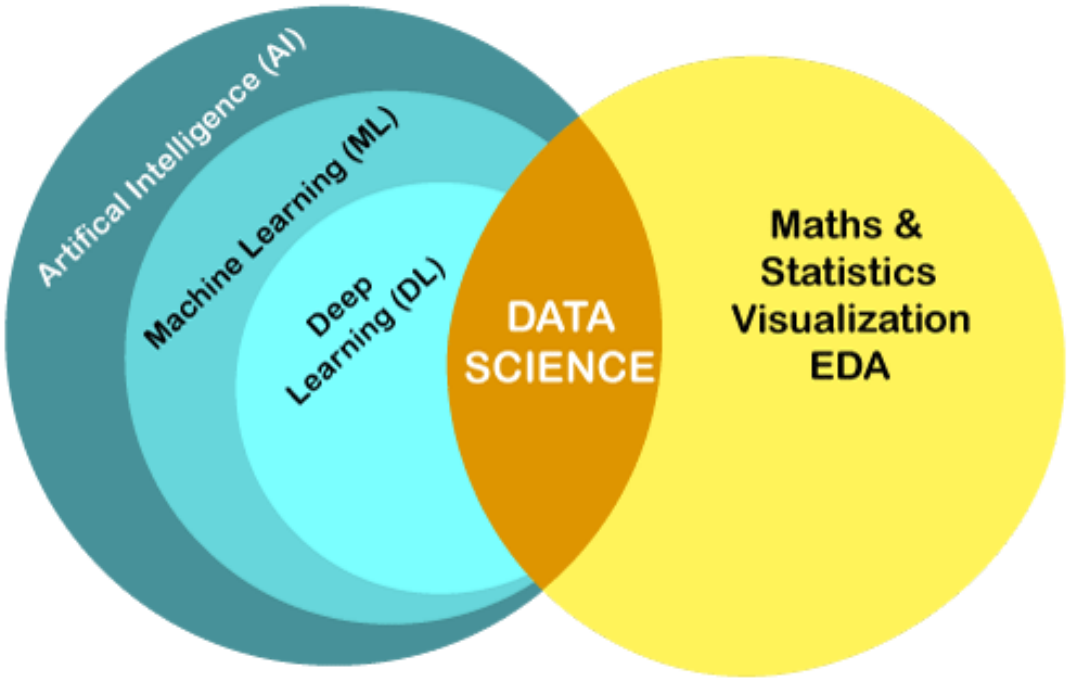


Figure: https://www.javatpoint.com/data-science-vs-machine-learning

Various terms used interchangeably

It takes two to tango

- Probability/statistics idea as tools for understanding and framing questions
  - Have been around for a while
  - Several innovative extensions
- Use of programming tools/libraries
  - Wide availability of the same
  - Python, R, ...
- Availability of computing hardware
  - Laptops, cloud
  - Moore's law: #transistors on an IC double in two years.
- Increasing amount of data: big data
  - Internet of things (IOT)
  - Industrial internet of things (IIOT): Industry 4.0
- Result: Tremendous interest in this area

Recent Advances

- Big data
- Cloud Computing
- Deep Neural Networks
- Convolution and Recurrent Neural Networks
- Applications in
  - Image processing
  - Natural language processing
  - Large language models (LLM)
- Deepblue, Watson, Srini, AlphaGo, Google assistance, Google lens, ...

Types of Problems: Example I

- Impact of temperature on viscosity of toluene-tetralin blend[5]. Table gives data for 0.4 molar fraction of toluene.

Temp. (^oC)	Viscosity (mPa.S)
24.9	1.1330
35.0	0.9772
44.9	0.8532
55.1	0.7550
65.2	0.6723
75.2	0.6021
85.2	0.5420
95.2	0.5074

- Problem: Predict viscosity given the temperature.

[5] Byers and Williams, 1987

Types of Problems: Example II

- Given income, family, education data of hundreds of past customers and their loan repayment record,
- For a new loan applicant, predict if the customer will default on loan or not

Types of Problems

- Fundamentally two types of problems
- Regression
  - Predicting quantitative value of a response (dependent) variable of interest: Output
  - Use of independent/regressor variables or features: Inputs
  - Example: Predict car mileage given engine capacity, car weight, car dimensions, number of gears, etc ..
  - Use of its own past values: Inputs (Time series)
- Classification
  - Predicting qualitative value (label) to a sample: Output
  - Use of independent/regressor variables or features: Inputs
  - Example: Given height, weight, BMI, bone density, etc. of a person, predict whether person is male or female.
- In both types: Given vector x, predict y.

Maybe Another Type as Well

- Feature Engineering/Selection
  - Selecting the right set of variables or combinations/transformations thereof
  - Ultimately useful for some purpose
  - Implicit in any modeling activity
- Monk Problem
  - A monk climbs up from base of hill (@9AM) to the top (@9PM) on day 1: moving at arbitrary speeds.
  - The monk climbs down from top of hill (@9AM) to the base (@9PM) on day 2: moving at arbitrary speeds.
  - Would there be a place on the way such that he is at that place at the same time on both days?

The new kid (giant) on the block: Generative AI

Generative AI, also known as Generative Artificial Intelligence, refers to a class of AI models and algorithms designed to generate new data that is similar to the data they were trained on. These models are capable of creating original content, such as images, text, audio, and videos, rather than just recognizing patterns or making decisions based on existing data.

ChatGPT

What is Learnt-1?

- Aim: To predict y given x.
- Translates to: Learn a model relating y to x.
- Without loss of generality: y is a scalar, and x is a vector.
- Regression problems:

$$y = f(x_1, x_2, ..., x_n, p_1, p_2, ..., p_m)$$

- Classification problems:

$h(x_1, x_2, ..., x_n, p_1, p_2, ..., p_m) \geq 0$, if data belongs to class 1 (y=0)
$h(x_1, x_2, ..., x_n, p_1, p_2, ..., p_m) < 0$, if data belongs to class 2 (y=1)

What is Learnt-2?

- Need to identify an appropriate f, h
  - Choice of model: Neural Network, support vector machine, ...
  - Trial and error: Experience
  - Computers better at trying out various choices
- Given an f or h
  - Choose the parameters $p_1, p_2, ..., p_m$
  - An optimization problem

Popular Models

- Simple regression models, Neural networks, decision trees, support vector machines, Bayesian belief networks, etc.

Types of Learning Paradigms

Based on types of learning (training) datasets available

- Supervised learning: Both inputs, outputs i.e. labels given
- Unsupervised learning: Only inputs given; labels not given
- Semi-supervised learning
  - For some samples: Both inputs, outputs given
  - For other samples: Only inputs given
- Reinforcement learning
  - Both inputs, outputs not given.
  - Algorithm has to generate inputs, apply to system and see the output (reward)
  - Figure out which are good inputs for a given situation

Supervised Learning

- In training data: Both inputs, outputs i.e. labels given

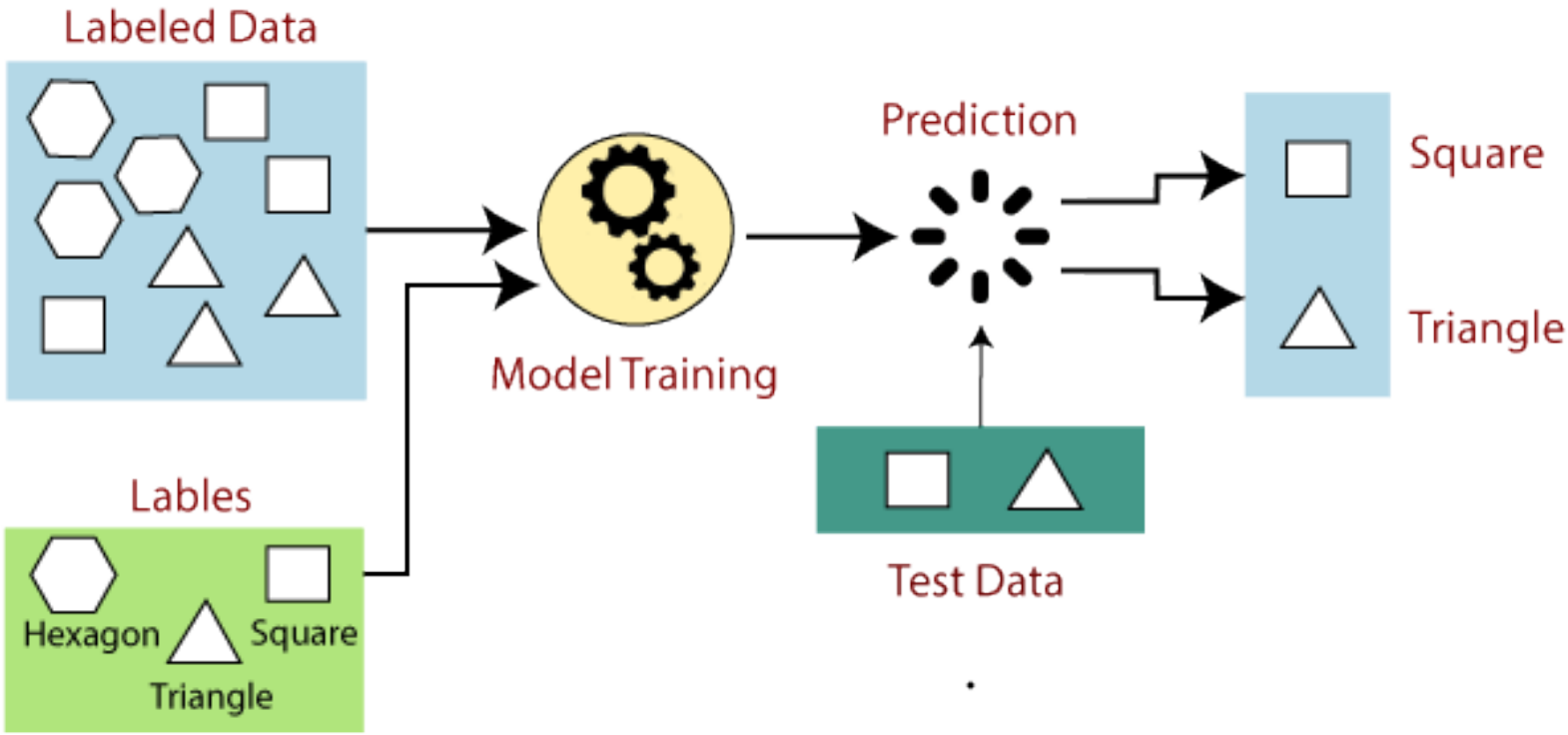


Figure: https://www.javatpoint.com/supervised-machine-learning

Unsupervised Learning

- In training data: Outputs i.e. labels NOT given

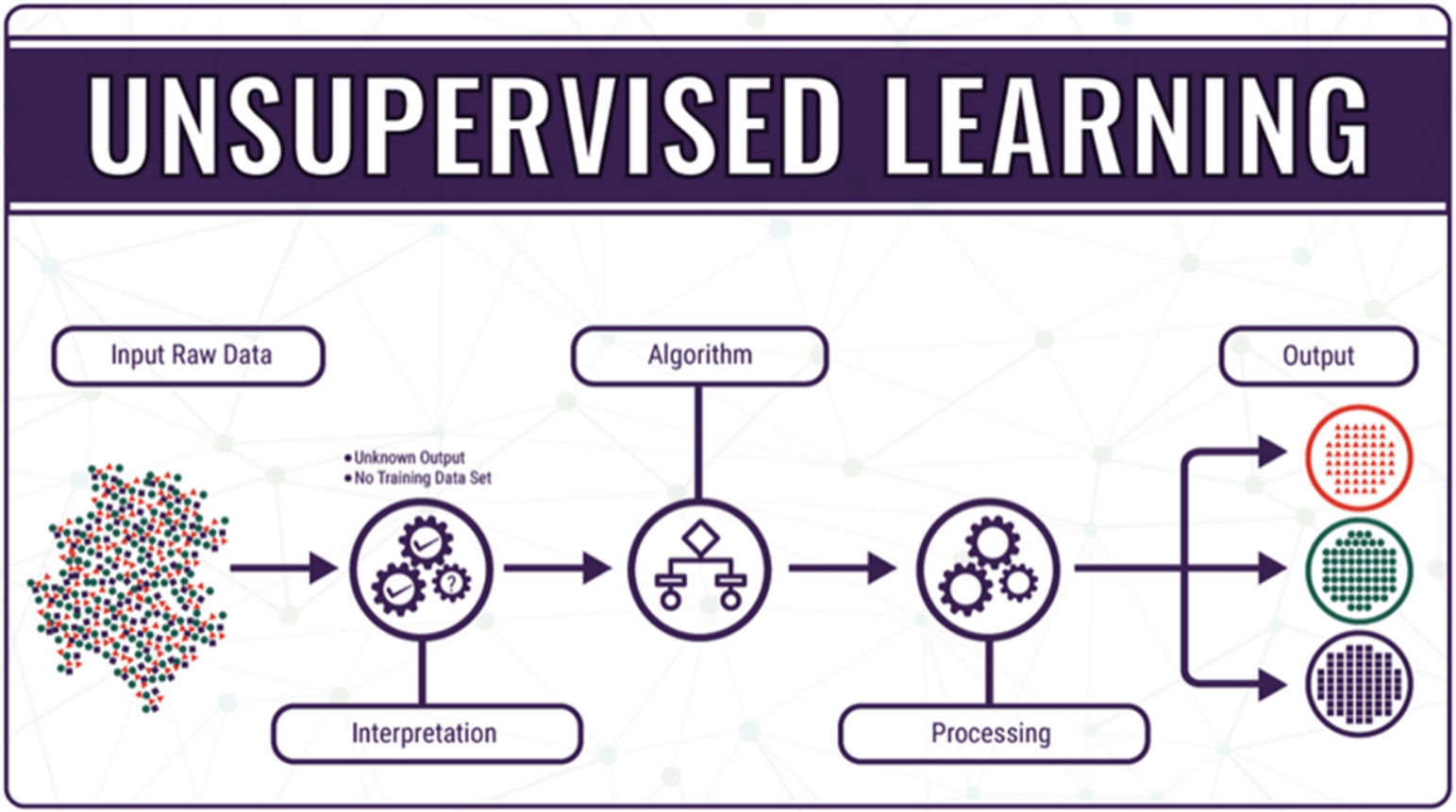


Figure: https://www.researchgate.net/figure/Unsupervised-Learning18_fig3_341703036

Semi-Supervised Learning

- Training data: Some part labeled, some unlabeled

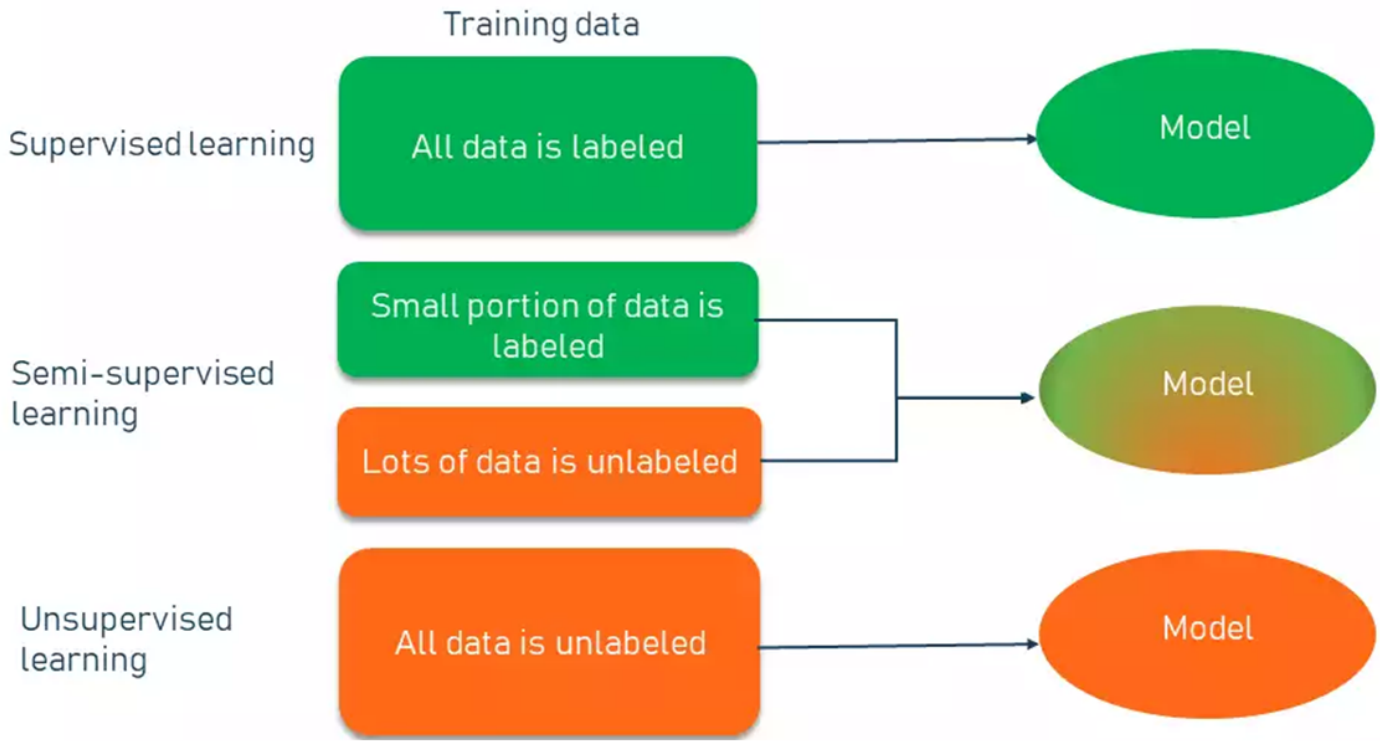


Figure: https://www.altexsoft.com/blog/semi-supervised-learning/

Reinforcement Learning

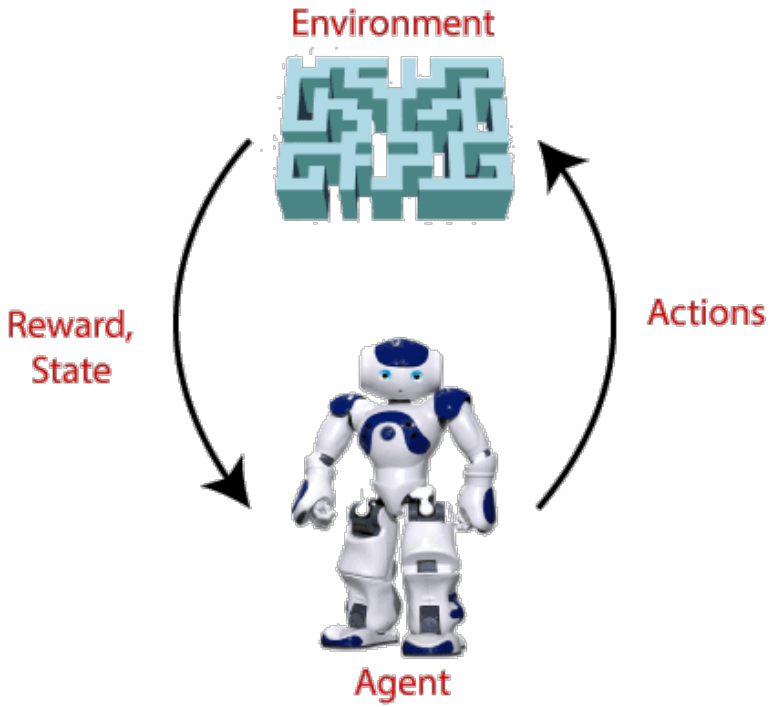


Figure: https://www.javatpoint.com/reinforcement-learning

- Both inputs, outputs not given.
- Algorithm to apply actions on system and see output (reward)
- Figure out which are good inputs for a given situation
- Example: How a baby learns to walk, self driving car

Time for a Quiz

Application	**Problem Type**	**Learning Paradigm**
Predict viscosity of bio-diesel knowing fatty acid composition of vegetable oils used to prepare bio-diesel (lab experimental data available)	Regression	Supervised
Determine whether catalyst is deactivated or not in a reactor	Classification	Supervised/ Unsupervised
Cluster large number of sensors in a nuclear reactor in smaller groups so that sensors within a group have high correlation	Regression/ Classification	Unsupervised

Quiz (II)

Application	Problem Type	Learning Paradigm
Predicting the impending failure (yes/no) of a pump used to pump oil/gas from a well [Historical instances of pump failure not available]	Classification	Unsupervised
Predicting the impending failure (yes/no) of a pump used to extract oil/gas from a well [Several historical instances of pump failure available]	Classification	Supervised

Quiz (III)

Predicting the time-to-failure of a pump used to extract oil/gas from a well [Several historical instances of pump failure available]	Regression	Supervised
Learning to operate the cooling water inlet flow valve so that reactor temperature does not vary	—	Reinforcement

Summary

- Data analysis/machine learning: A vast field
- Evolving fast
  - Availability of newer, larger, standard datasets: In many domains
  - Massive computational power
  - Transfer learning concept: Pre-trained models
  - Easily available codes/libraries
  - ChatGPT

But ...

- Still role for a domain expert
- Not yet possible to: dump all data as input and get output
- Feature engineering, data preprocessing, choice of tool, validation, understanding strengths and limitations,...: Requires domain expert
- Explainable Artificial Intelligence
- AI/ML: Keep fundamentals strong
- This course: Focus on probability and statistics

Thank You

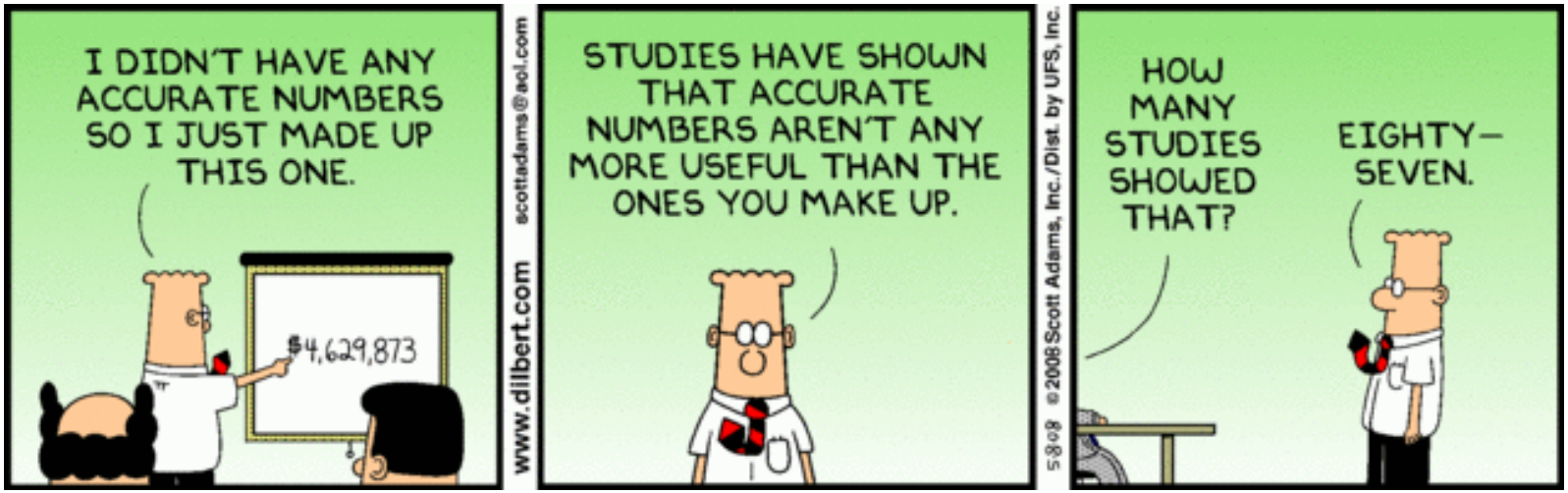


Figure: http://dilbert.com/strip/2008-05-08