CSDS 452 Causality and Machine Learning

Lecture 1: Introduction

Instructor: Jing Ma

Fall 2024, CDS@CWRU

Outline

- **Course Logistics**
- Course Description and Schedule
- Grading Policies
- Introduction of Causality

Course Format

- Tue/Thu 4:00-5:15pm (Eastern Time), Nord 211
- Attendance is required
  - If you are under special conditions and cannot attend the class, please contact the instructor in advance
- Class information and course materials will be updated in Canvas

Instructor

- Jing Ma
- Assistant Professor in CDS@CWRU
- Office hour: Tue 3-4pm
- Office: Olin 511

Webpage: https://jma712.github.io/
Email: jing.ma5@case.edu

Teaching Assistant

- Yiran Qiao
- Email: yxq350@case.edu
- TA Office Hours: See Canvas

Course Communication

- CWRU Canvas (automatically enrolled)
  - Course announcement
  - Lecture materials (slides / reading materials / homework / project)
  - Assignment submission
- If email me: please put CSDS 452 in the subject

Outline

- Course Logistics
- **Course Description and Schedule**
- Grading Policies
- Introduction of Causality

Goal of this Course

- Causal inference is important!

"Causality is very important for the next steps of progress of machine learning."

-- By Yoshua Bengio, Turing Award-wining scientist

Goal of this Course

- Understand fundamental concepts and methods of causal inference
  - Wide coverage of causal inference techniques
- Understand the methodologies and applications of the combination of these two areas
  - Learn how to use machine learning to help casual inference
  - Learn how to use causal inference to help machine learning
- Prepare students for doing cutting-edge research in machine learning, causal inference, and related fields
  - Open the door to the amazing job opportunities in academia or industry in computer and data sciences

What You Will Do and Get

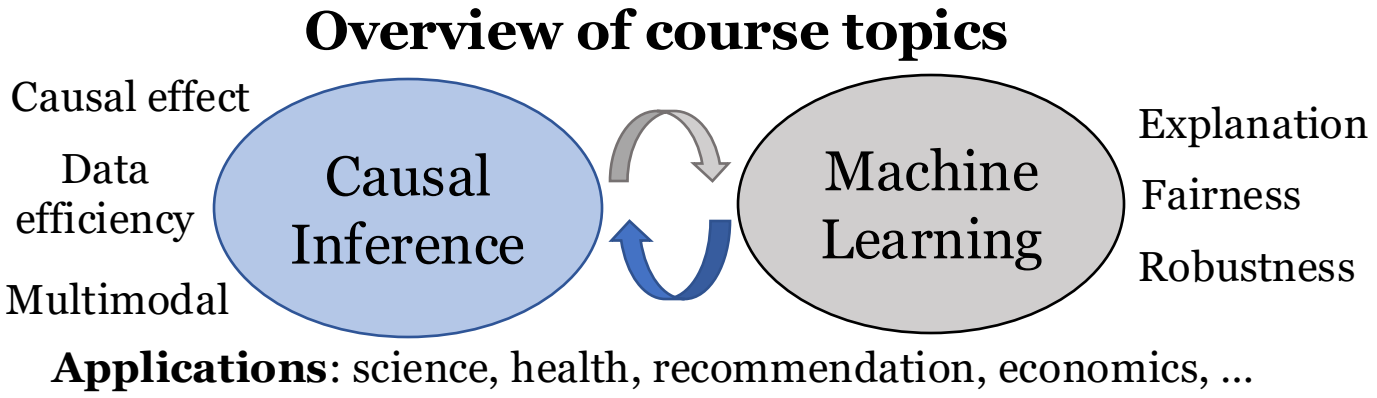
- Read and share related papers in top venues
- Get hands-on project experience by developing practical methods
- Be encouraged to express your thoughts, confusions, and suggestions in our discussion

Course Prerequisites

- Machine Learning
- Basic mathematical skills in Probability and Statistics
- Coding skills in Pythons for machine learning
- (Recommended but not required) Graphical Models
- (Recommended but not required) Causal Inference

Structure of this course

- Three major topics will be covered:
  - Introduction to causal inference
  - Using machine learning to help causal inference
  - Using causal inference to help machine learning



Outline

- Course Logistics
- Course Description and Schedule
- **Grading Policies**
- Introduction of Causality

Grading Policy

- Homework (35%)
- Paper presentation (20%)
- Course project (40%)
- Class participation (5%)

Final Grade: A (>= 90), B (>= 80 and < 90), C (>= 70 and <80), D (>=60 and <70), F(<60).

Homework

- Homework (35%)
  - Students will perform homework individually
  - 3 homework, may include reading report, problem-solving, or small coding tasks
  - Discussion allowed, but must independently write up solutions

- **Late Submission Policy:** Assignments that are turned in late will be receiving penalties based on the formula:
  - final score = raw score * (1 – 0.2 * number of days late).

Paper Presentation

- Paper presentation (20%)
  - (1~2 people) Present one/several papers in related areas in class
  - 18min=15min presentation+3min QA
  - The selected papers need to be confirmed with the instructor and TA at least 2 week before the presentation day
  - The presentation slides need to be sent to the instructor and TA the night before your presentation
  - The presentation will be graded by the instructor, the TA, and all other students, based on the clarity, content, critiques and insightful comments, timing, and handling of questions

Paper Presentation: What to Include

- Motivation of the paper
- Key ideas
- Proposed method (high-level)
- Experimental results
- Strengths and weakness of the paper
- Optional: The connection between your selected paper and your course project

Where to Find Good Papers

- Data Mining Venues
  - KDD, ICDM, SDM, PKDD, TKDE, TKDD, DAMI
- Web and Information Retrieval Venues
  - WWW, SIGIR, WSDM, CIKM, TOIS
- Database Venues
  - VLDB, SIGMOD, ICDE
- Machine Learning/AI Venues
  - ICML, NeurIPS, AISTATS, ICLR, AAAI, IJCAI, JMLR, TPAMI

Course Project

- Course project (40%)
  - 1~3 people (collaboration encouraged!)
  - Presentation form:
    - a proposal in pdf (10%)
      - within 2 pages, latex ACM Format
    - a report in PDF (15%)
      - within 6 pages, latex ACM Format
    - a 18min (15min presentation + 3min QA) presentation in class (15%)

Course Project: Potential Topics

- 1. Survey and empirical comparison of different algorithms in related fields
- 2. Improve existing algorithms by brainstorming the shortcomings
- 3. Investigate a new problem that is seldom studied before
- 4. Contribute a new dataset that may benefit the related research
- 5. Theoretical analysis of existing algorithms and provide rigorous results
- Ideally, the project should be helpful for your own research
  Think about your teammates and potential topics now!

Class Participation

- Class participation (5%)
  - 5 pop quizzes at randomly selected lectures, students can miss at most one quiz to get full participation points
    - 4&5 -> 5 points
    - 3 -> 3 points
    - 2 -> 2 points
    - 1 -> 1 points
    - 0 -> 0 points
- If you are in a special condition (e.g., illness, visa issues, …) and have difficulty in attending class, please contact me before the class.

Reading Material

- Books & papers
- Causal inference
  - Imbens, Guido W, and Donald B Rubin. 2015. Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge University Press.
  - Pearl J. Causality[M]. Cambridge university press, 2009.
  - Miguel Hernan, Jamie Robins. 2023. “Causal Inference: What If”
- Papers from top venues
  - ICML, NeurIPS, KDD, AISTATS, …

Reading materials will be updated after each class!

AI Usage Policy

- **Restrictions on AI in Problem Solving**: Students are not permitted to use AI tools to solve problems or complete assignments that are intended to assess individual understanding and application of the course material.
  - calculations, coding, analyzing case studies, generating responses to theoretical questions, entire report writing

- **Permissible Uses of AI**: AI can be utilized as a tool for language editing and enhancing the clarity of student-written work.
  - assist with grammar, punctuation, and style in writing
- However, the original thoughts and content must be generated by the student.

Week 1	**Aug 27**	**Introduction to course**
	Aug 29	Potential outcome
Week 2	Sep 3	Structural causal model
	Sep 5	Causal effect estimation (1)
Week 3	Sep 10	Causal effect estimation (2)
	Sep 12	Unobserved confounders (1)
Week 4	Sep 17	Unobserved confounders (2)
	Sep 19	Causal inference: a broader view
Week 5	Sep 24	Bayesian Additive Regression Trees
	Sep 26	Neural network based causal inference (1)
Week 6	Oct 1	Neural network based causal inference (2)
	Oct 3	Causal effect learning on graph
Week 7	Oct 8	Causal inference in time-series data
	Oct 10	Applications of Causal inference in different domains
Week 8	Oct 15	Other ML for Causal Inference
	Oct 17	Causal generalization (1)
Week 9	Oct 22 (Fall break)	
	Oct 24	Causal generalization (2)
Week 10	Oct 29	Causal explanation
	Oct 31	Causal fairness
Week 11	Nov 5	Trustworthy AI and causal inference
	Nov 7	Domain-specific machine learning
Week 12	Nov 12	Social responsibility of AI
	Nov 14	Other causal inference for machine learning
Week 13	Nov 19	Paper presentation
	Nov 21	Paper presentation
Week 14	Nov 26	Paper presentation
	Nov 28 (Thanksgiving)	
Week 15	Dec 3	Final project presentation
	Dec 5 (Last Day of Class)	Final project presentation

If you have any suggestions/comments/questions/concerns, feel free to let me know!

Brief Self Introduction

- Your Name
- Your Program & year
- Optional:
  - research interest
  - why are you interested in this course
  - or anything else!

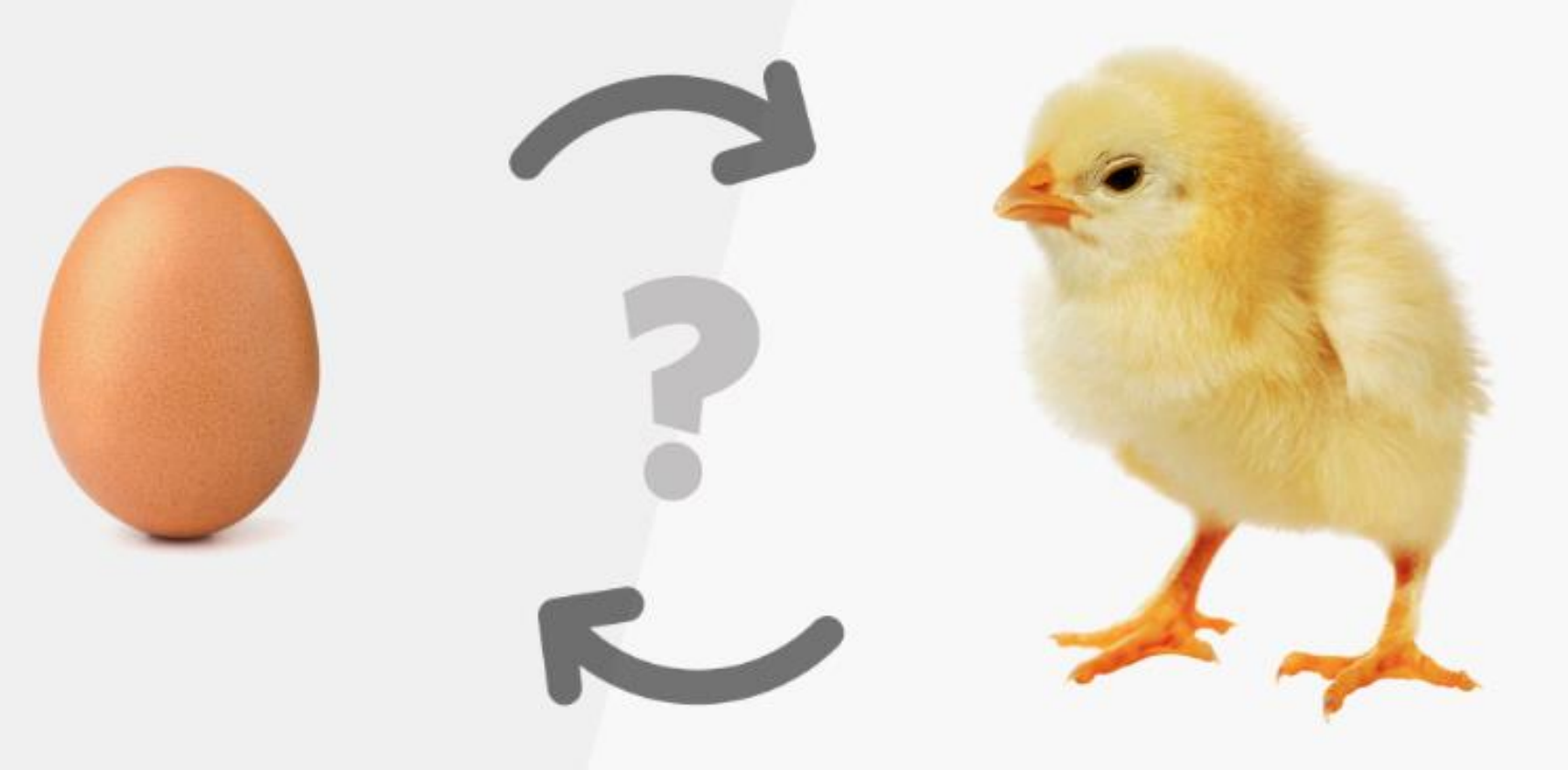
Outline

- Course Logistics
- Course Description and Schedule
- Grading Policies
- **Introduction of Causality**

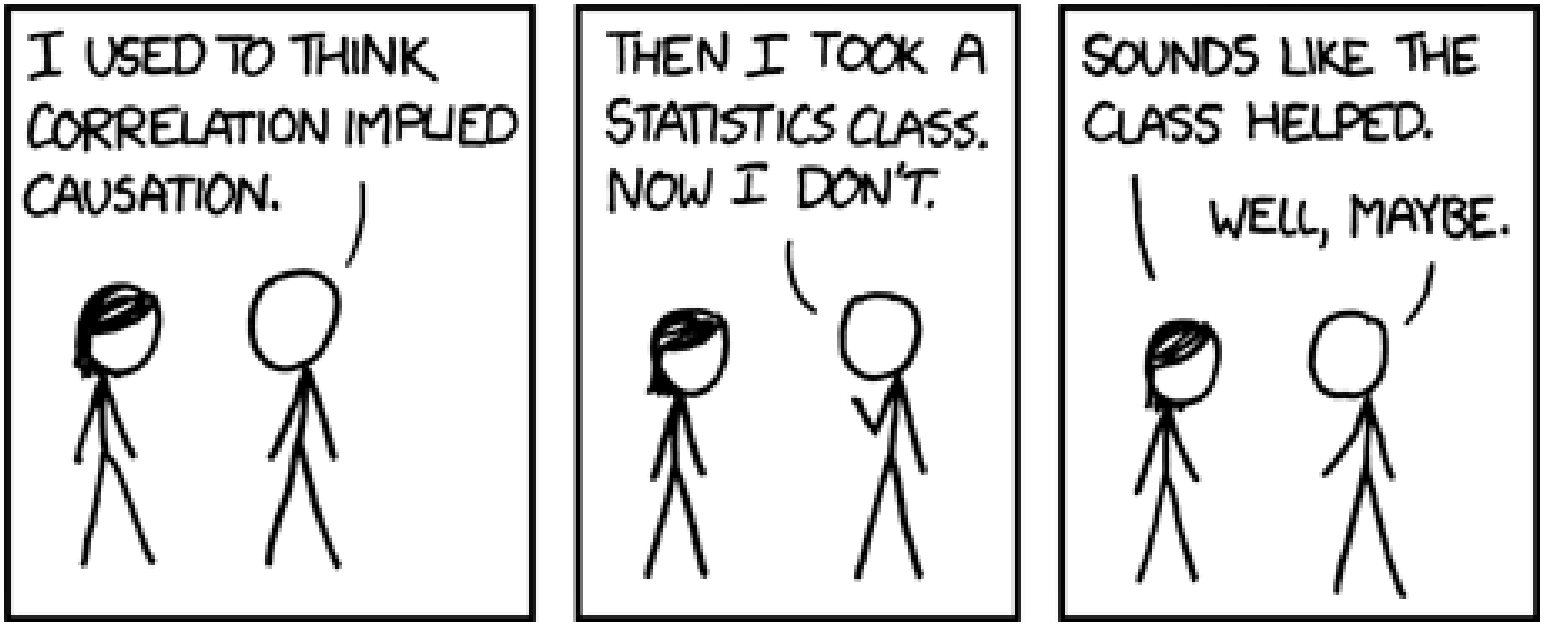
Today's Outline

- What is causality? Why is causality important?
- Introduction to ML
- The connections between ML and causal inference

What is causality?



Dependence vs. causation



http://imgs.xkcd.com/comics/correlation.png

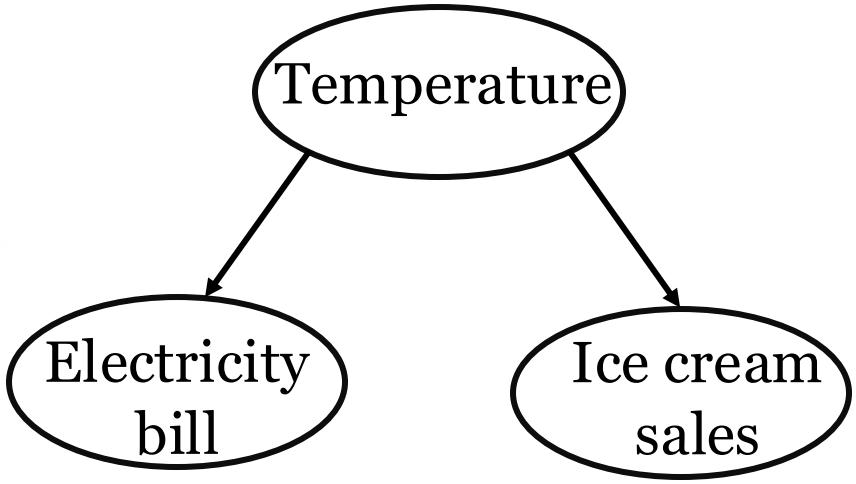
Causality ➜ dependence !

X and Y are associated iff $\exists x_1 \neq x_2$
$\mathrm{P}(\mathrm{Y}|X = x_1) \neq \mathrm{P}(\mathrm{Y}|X = x_2)$

Dependence ➜ causality ?

X is a cause of Y iff $\exists x_1 \neq x_2$
$\mathrm{P}(\mathrm{Y}|do(X = x_1)) \neq \mathrm{P}(\mathrm{Y}|do(X = x_2))$

Example

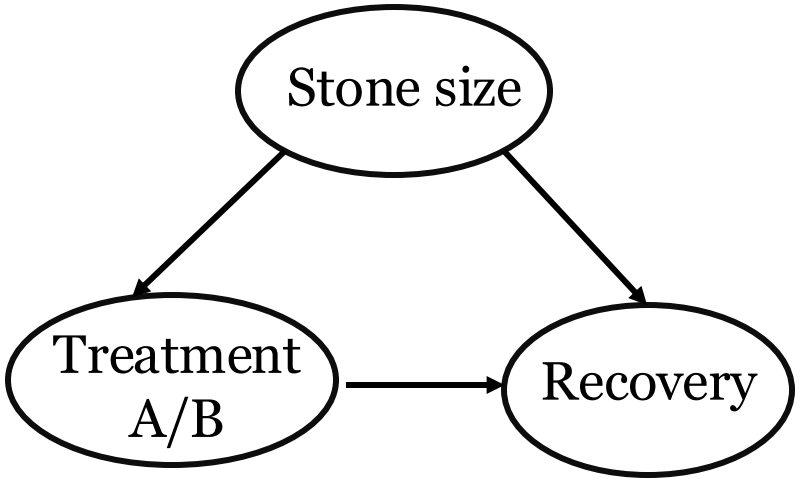


Higher electricity bill leads to higher ice cream sales?

Simpson's paradox

A trend appears in several groups of data but **disappears or reverses** when the groups are combined.

	Treatment A	Treatment B
Small stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**

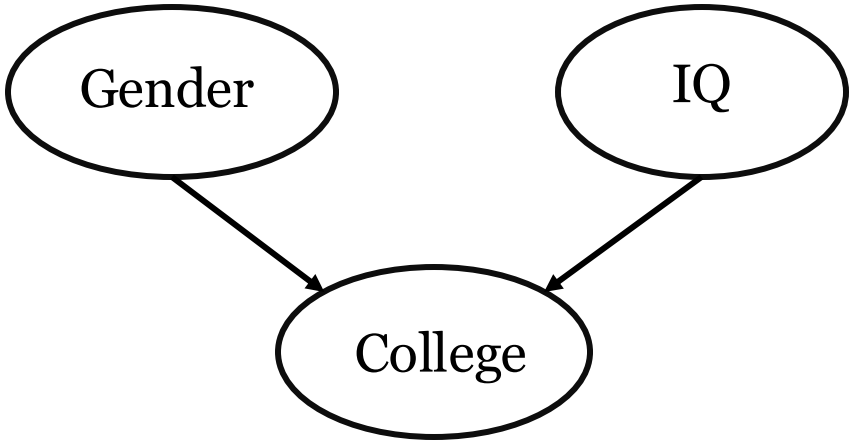


- Doctors tend to give patients with large stones treatment A, and the patients with small stones treatment B
- Larger stones has less recovery rate

"Confounding effect"

Strange Dependence

- Let's go back 50 years; maybe you'll find female college students are smarter than male ones on average. Why?

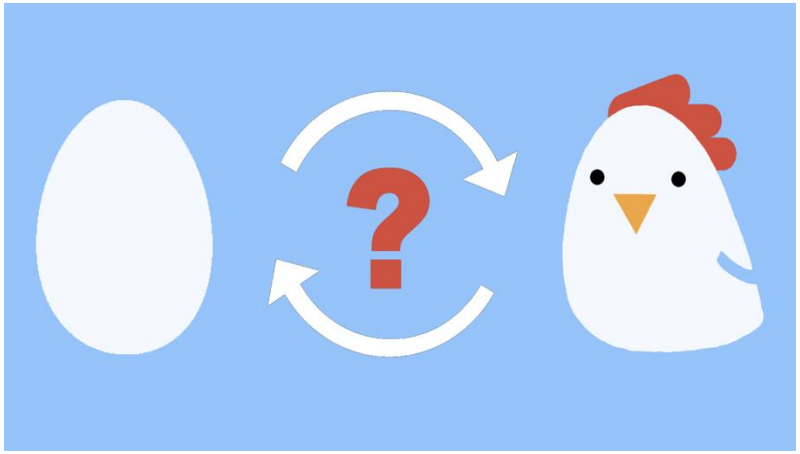


Why is causal inference important?

Thinking from a Causal View

- What is the cause of a certain variable of interest? How significant is the effect of a cause on an outcome?
- What changes can I make? What is the consequence after I make the changes?
- "What if?" e.g., "What if I had only studied harder for that exam?"

Main Problems in Causal Inference



Causal discovery

What causally affects what?

Causal effect estimation How significant is the causal effect?

Introduction to Machine Learning

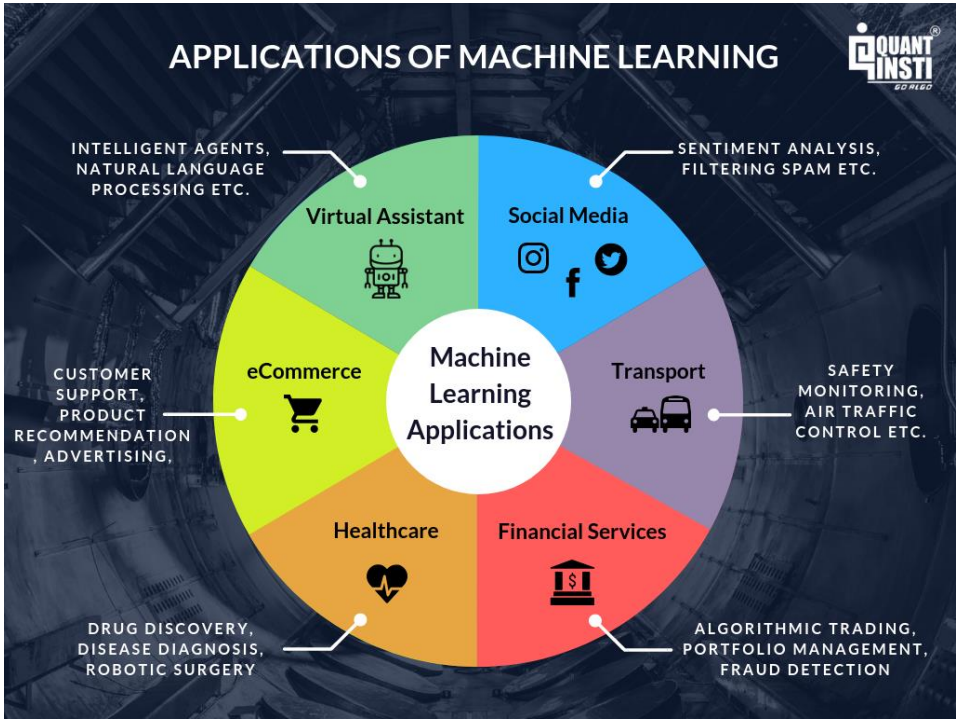
Machine Learning

- Artificial intelligence (AI)
  - The intelligence of machines or software, as opposed to the intelligence of human beings or animals.

- Machine learning
  - Arthur Samuel (1959): the field of study that gives computers the ability to learn without being explicitly programmed
  - Tom Mitchell (1998): The goal is identifying the underlying mechanisms and algorithms that allow improving our knowledge with more data

Why is Machine Learning Important?

- Solve problems automatically and efficiently almost everywhere
- Application domains:
  - healthcare,
  - finance,
  - law,
  - politics,
  - logistics,
  - entertainment,
  - autonomous driving...



https://www.linkedin.com/pulse/machine-learning-its-applications-vidhi-kapoor/

Introduction to ML

- Basic categories of ML
  - Supervised learning
  - Unsupervised learning
  - Reinforcement learning

- Deep learning

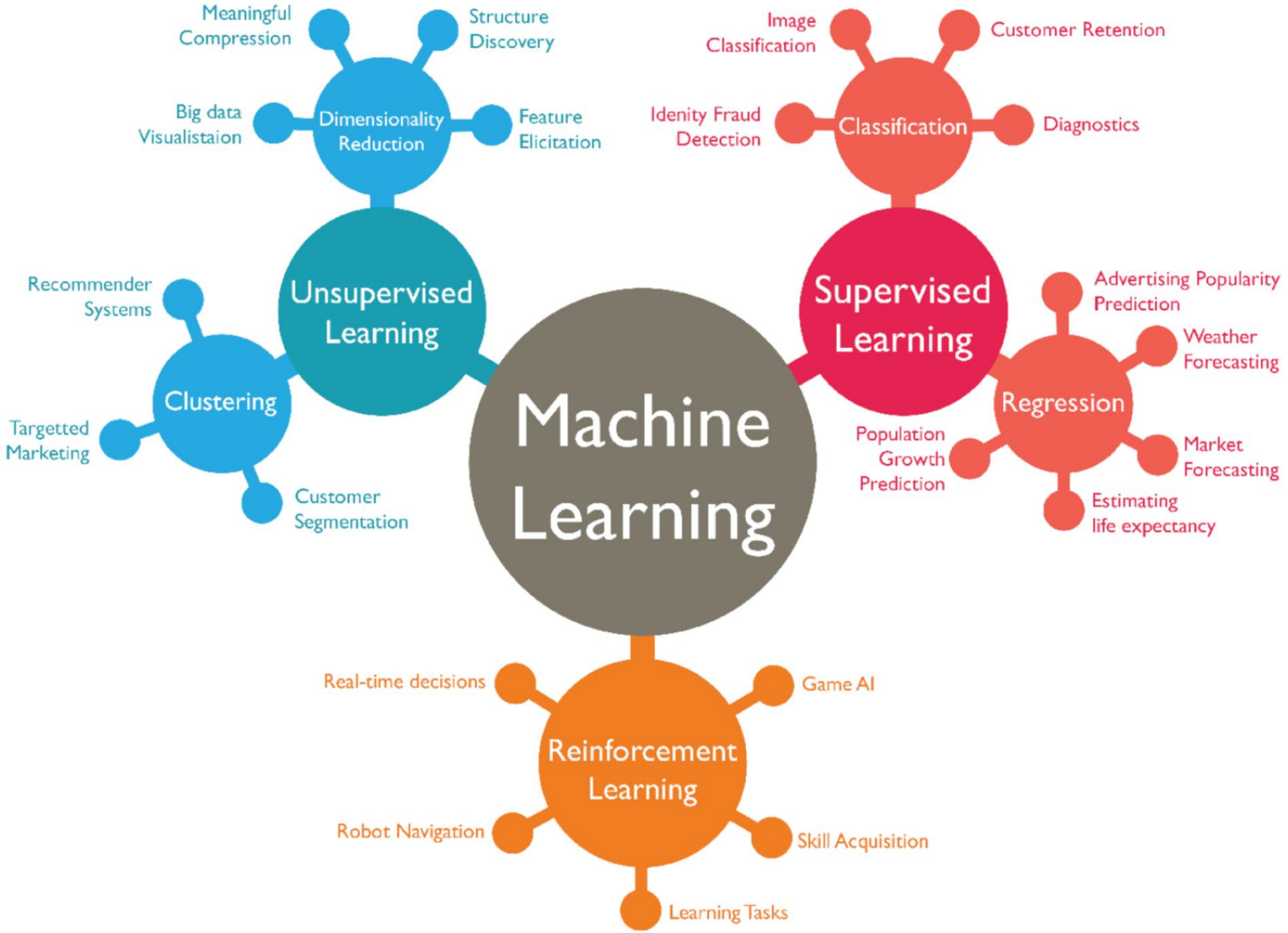
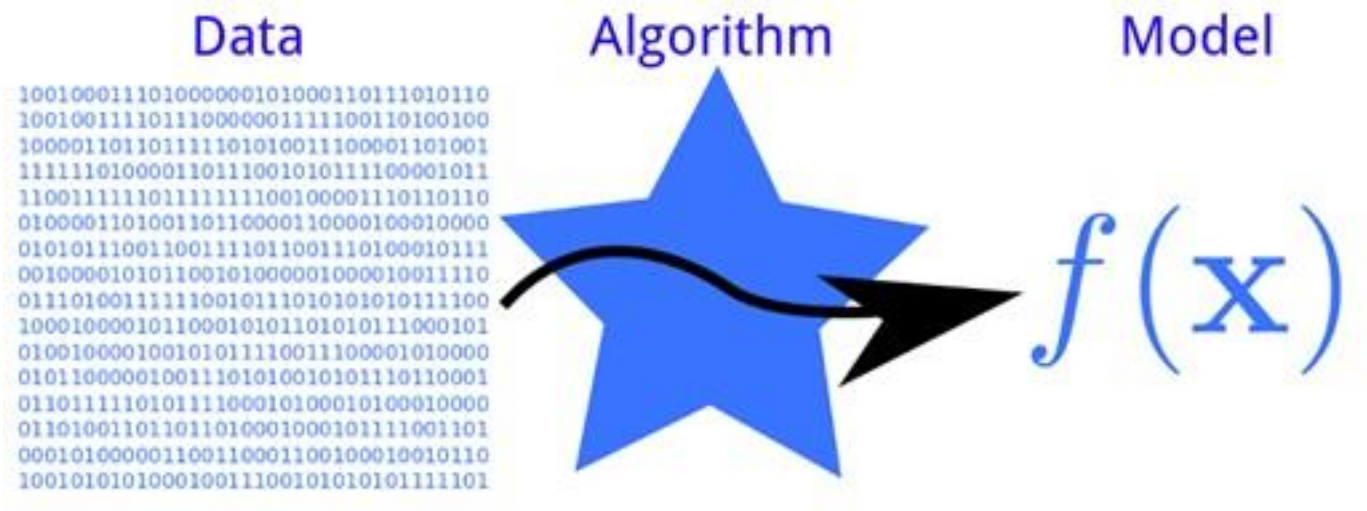


Image via Abdul Rahid

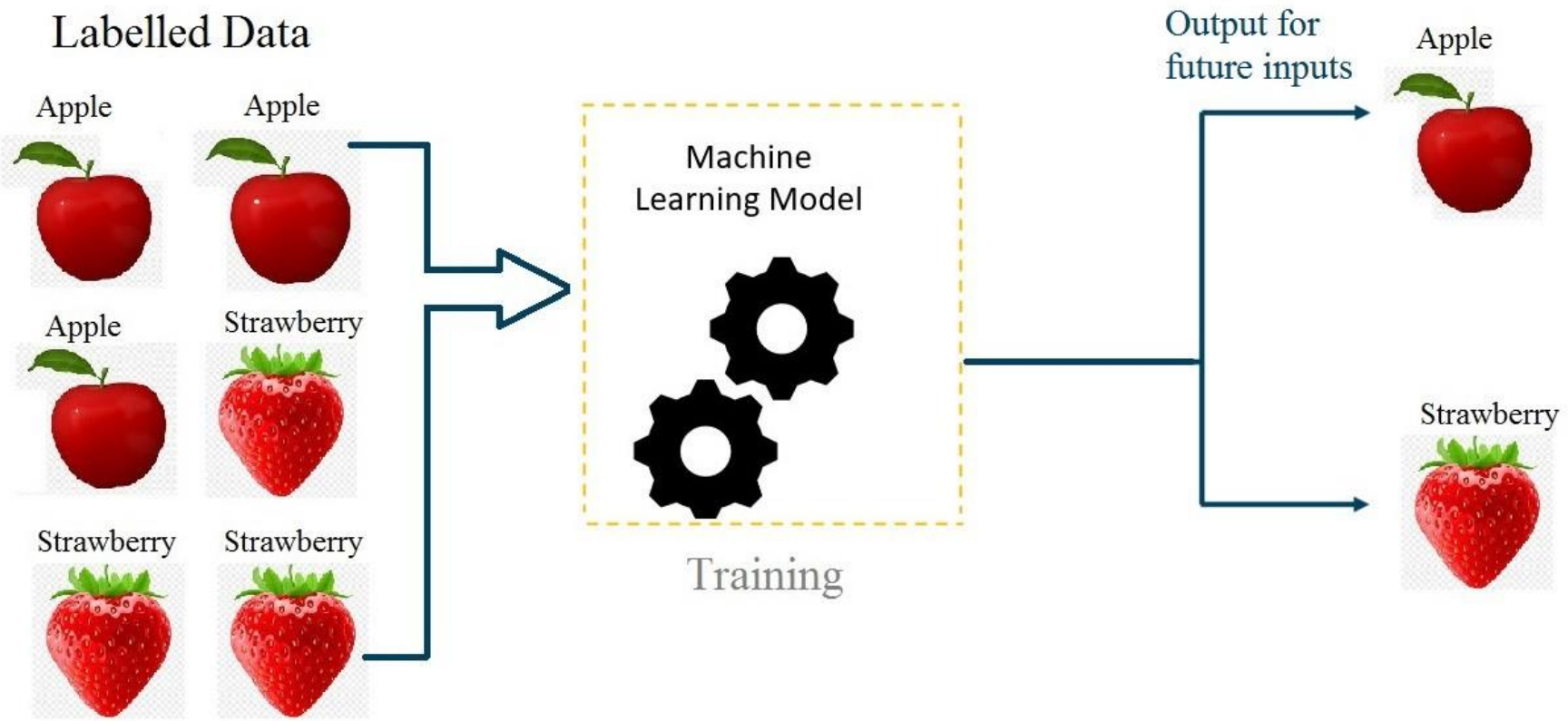
Three components in ML

- Data: what input data do you have?
- Task: what knowledge type do you want to seek from data? (prediction? Description?)
- Algorithm



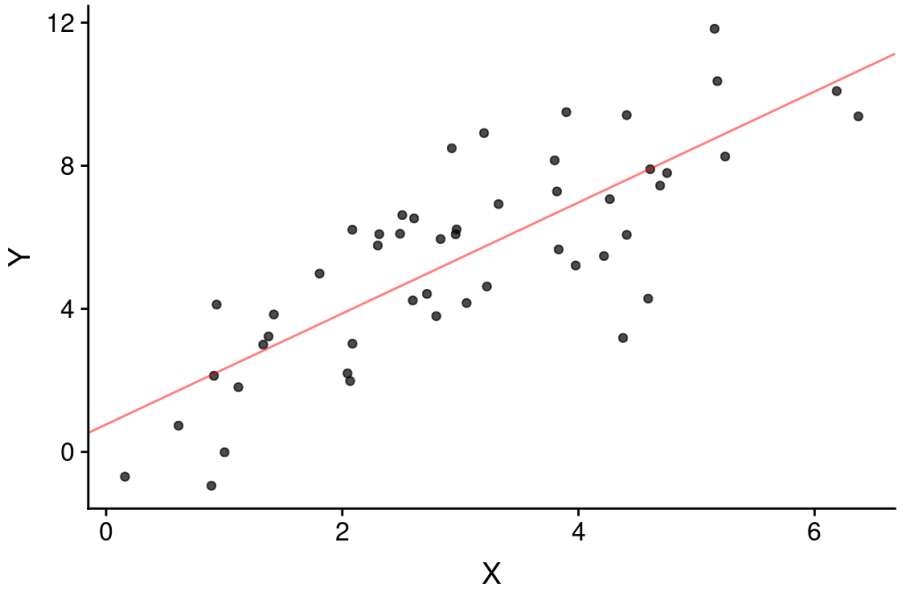
https://www.datasciencecentral.com/how-to-choose-a-machine-learning-model-some-guidelines/

Supervised Learning



Essence of Supervised Learning

- Given: data pairs (x_i, y_i), $i = 1, \ldots, n$
- Learning: fitting curve $f(x)$ on training data
- Goal: learn a good $f(x)$ to generalize well (i.e., predict unseen examples well)

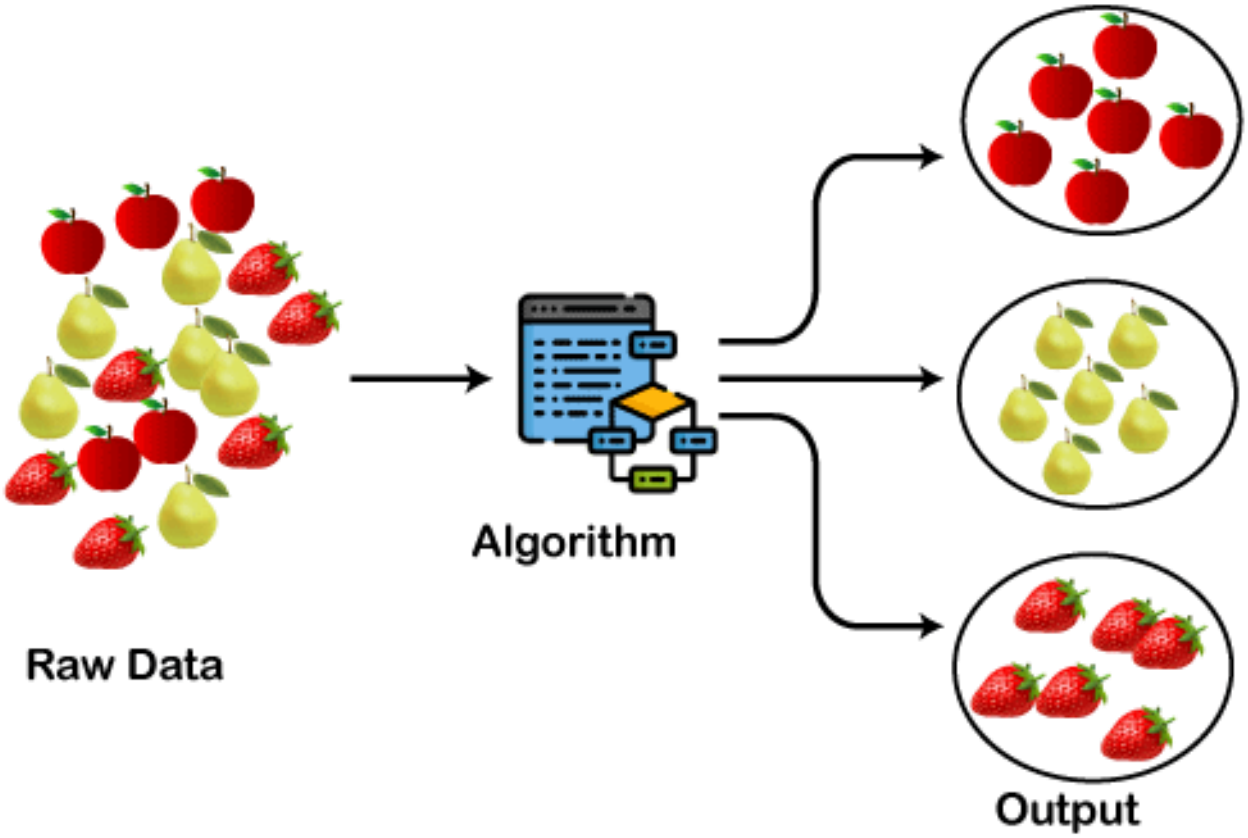


Supervised Learning Algorithms

- Nearest-neighbor
- Decision trees
- Linear/nonlinear regression
- Neural networks/deep learning

Unsupervised Learning

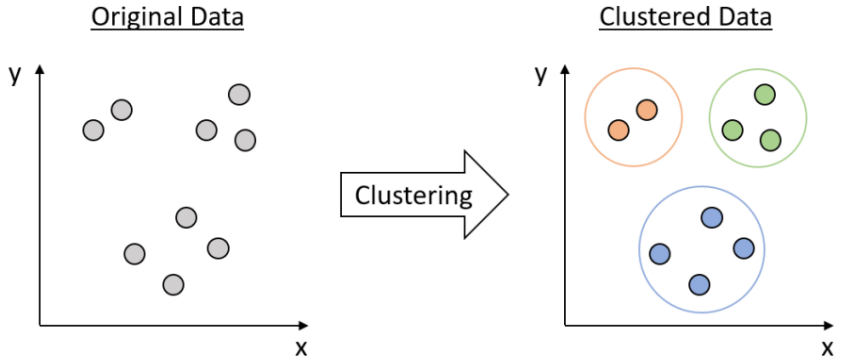
- No supervision (label) is provided

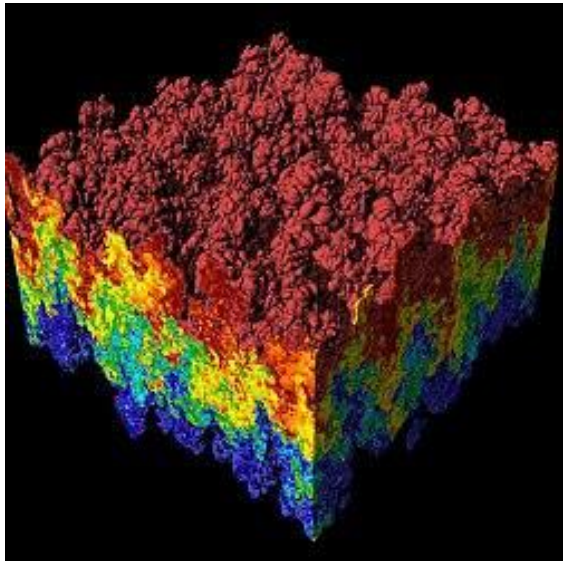


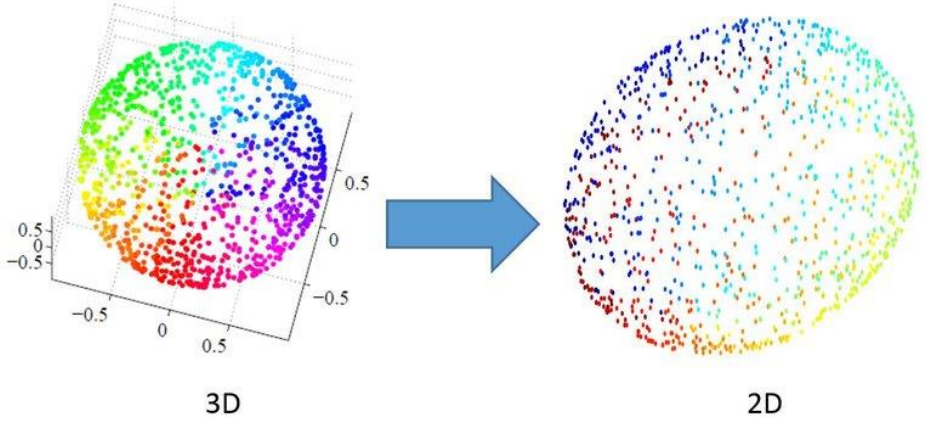
https://www.javatpoint.com/clustering-in-machine-learning

Unsupervised Learning

- Clustering
- Visualization
- Dimensional reduction

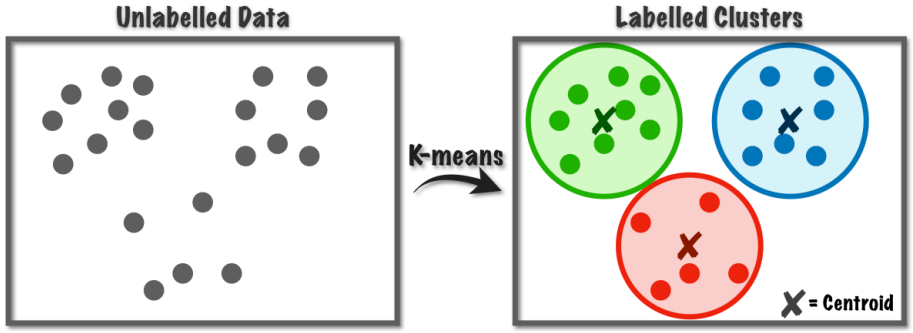


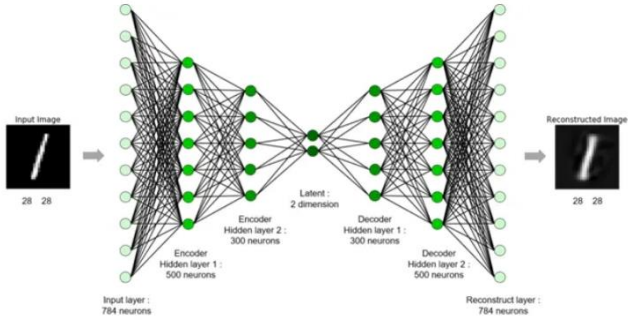


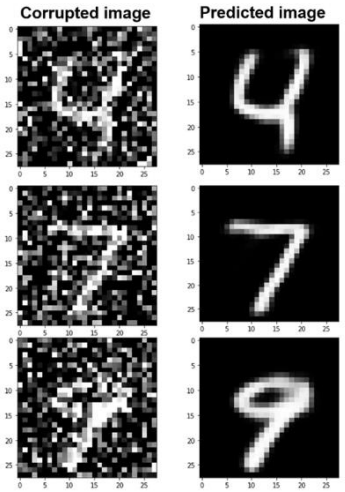


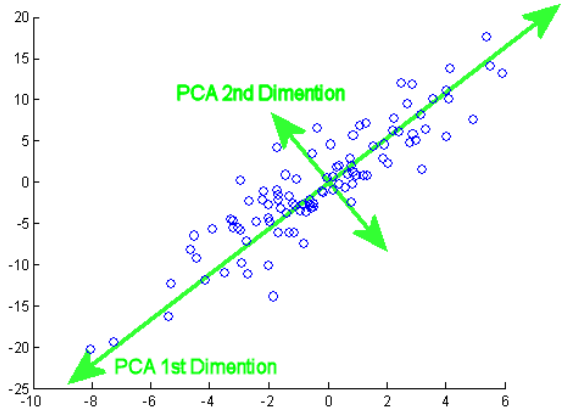
Unsupervised Learning Algorithms

- K-means clustering
- Principle component analysis (PCA)
- Autoencoder



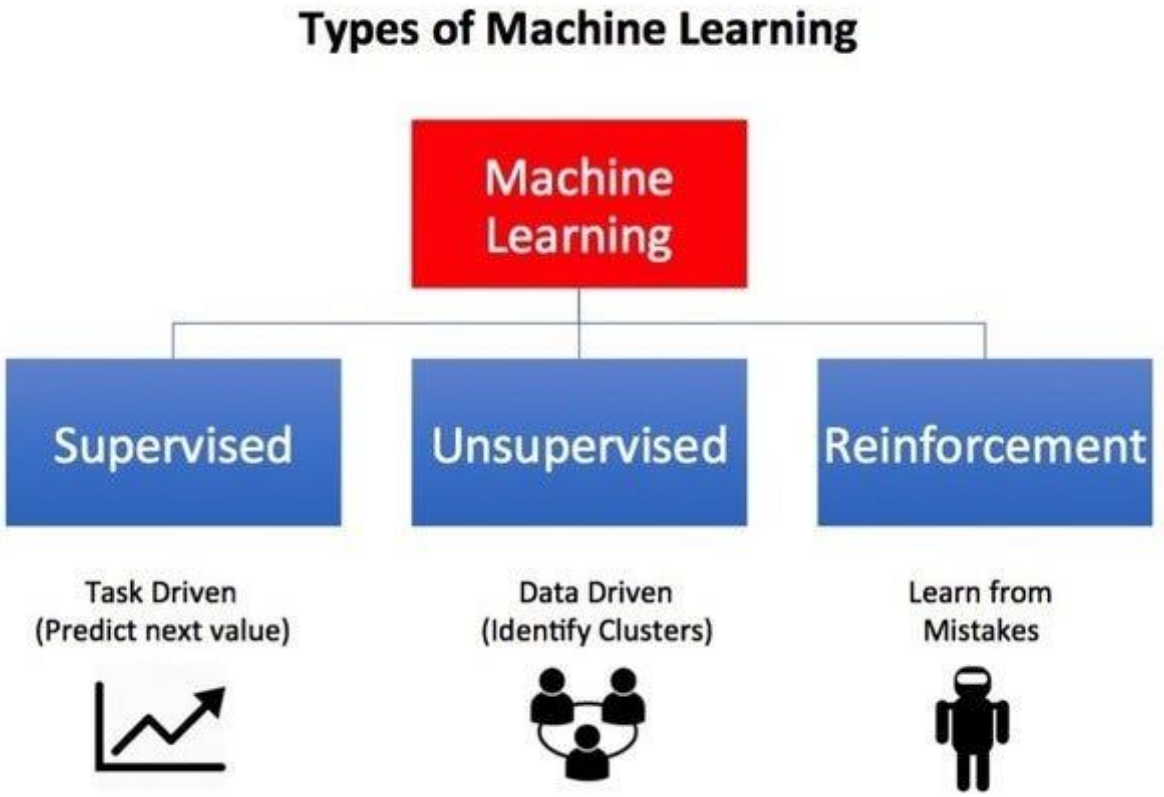






Reinforcement Learning

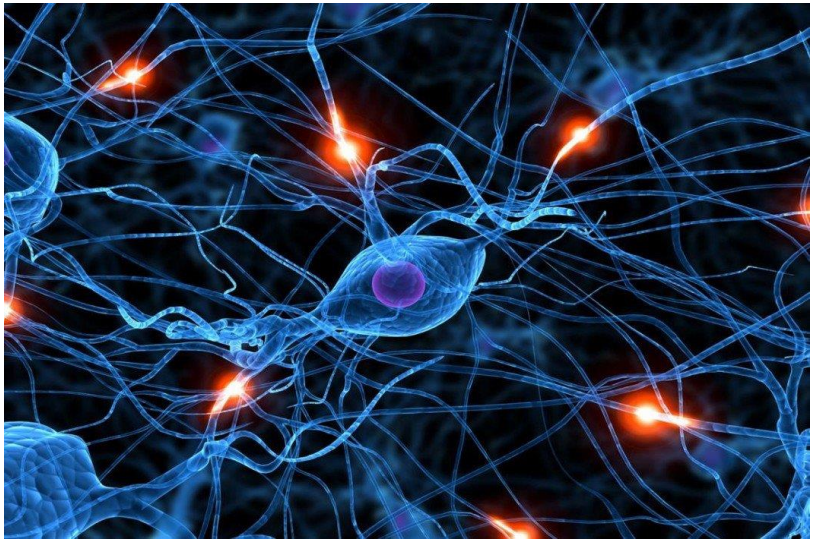
- Model receives feedback during learning

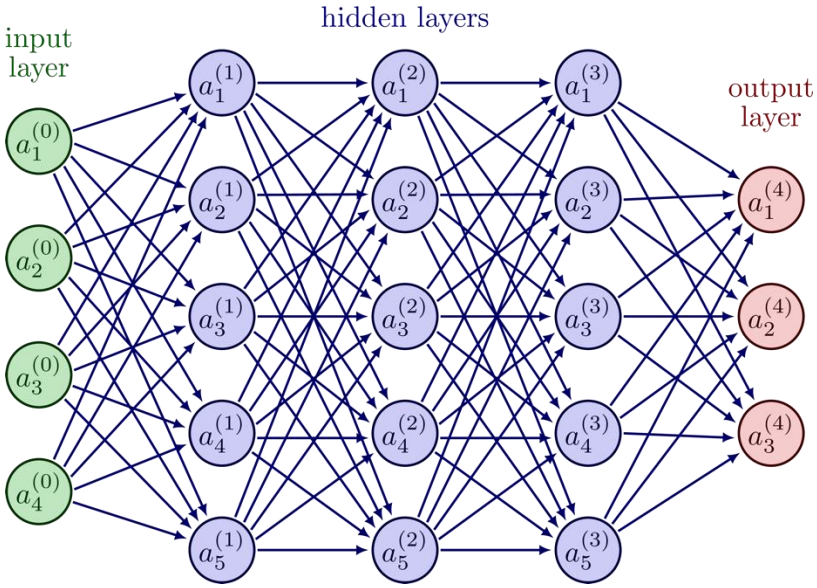


https://medium.datadriveninvestor.com/reinforcement-learning-cea620b89676

Deep Learning

- Consist of interconnected nodes, or "neurons”
- Organized into layers. Each neuron processes and transmits information to other neurons
- Functional approximation & feature extractor
- Supervised/unsupervised /reinforcement learning
- Very powerful!





Large Language Model

- Large language models (LLM)
  - Language model (LM): powerful tools for language processing, machine translation, and question answering
  - Large -- contain tens or hundreds of billions of parameters

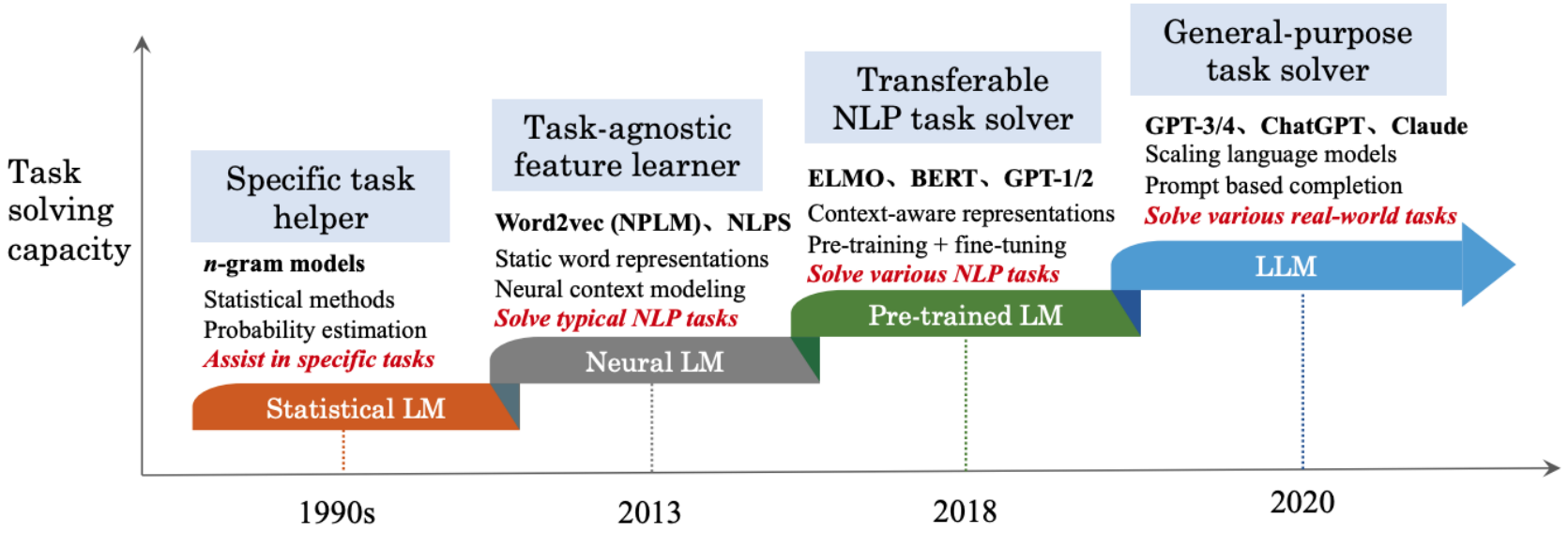
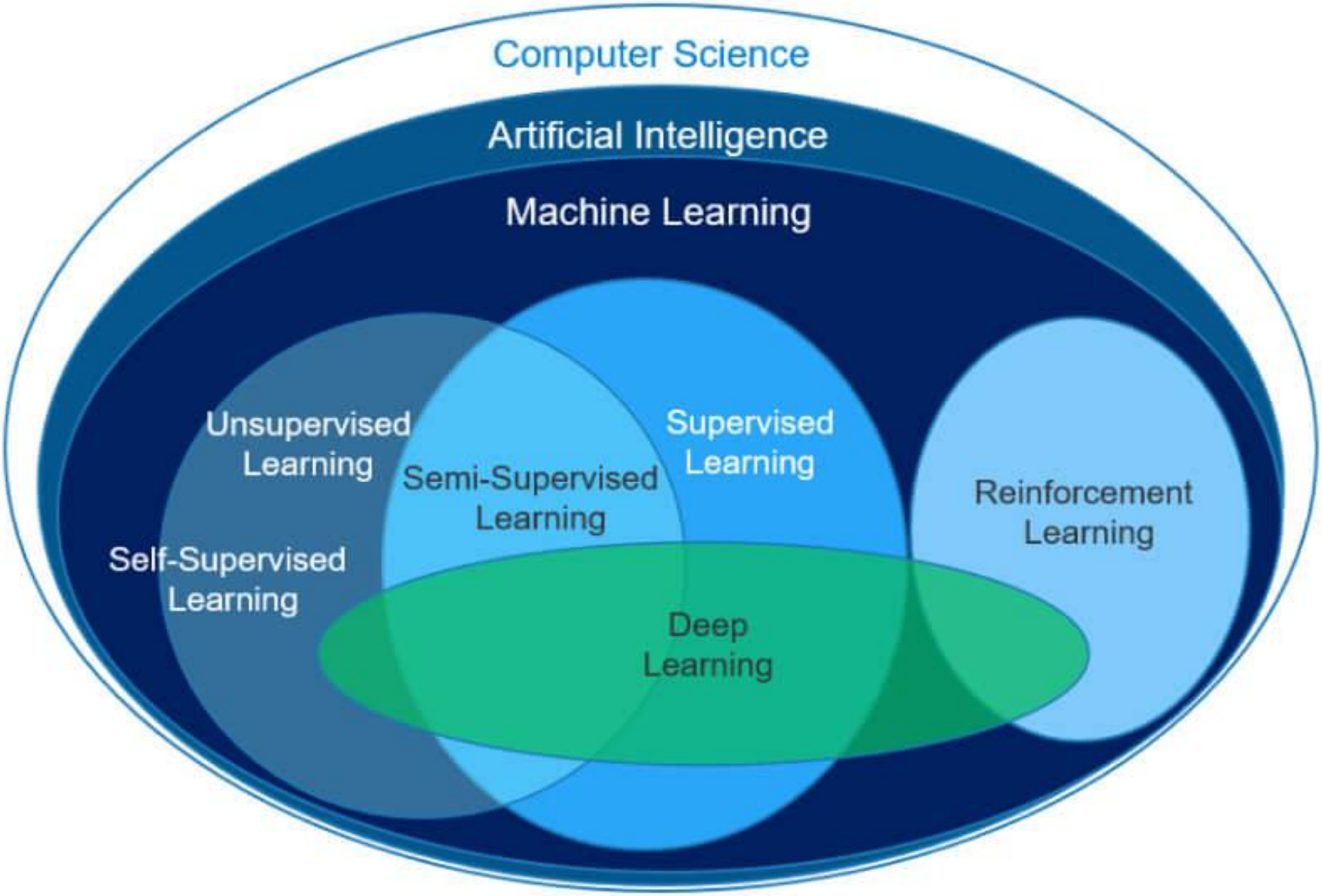


Figure: An evolution process of the four generations of language models (LM) [1]

[1] Zhao, Wayne Xin, et al. "A survey of large language models." *arXiv preprint arXiv:2303.18223* (2023).

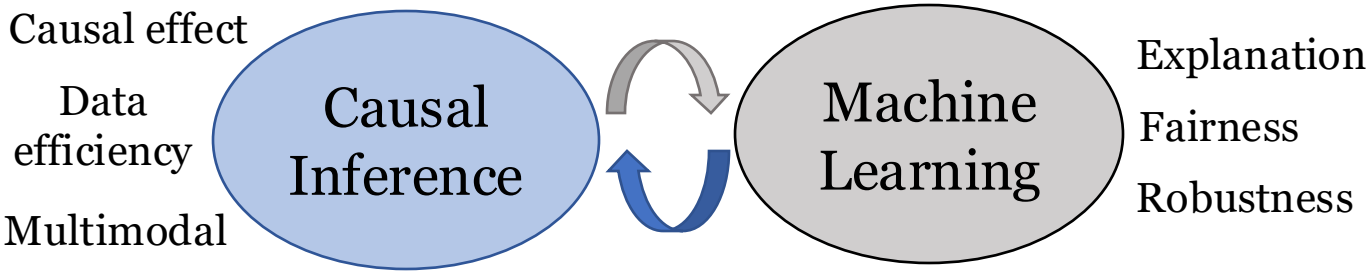


https://www.superannotate.com/blog/supervised-learning-and-other-machine-learning-tasks

Connection between Machine Learning and Causal Inference

ML & Causal Inference

	Causal inference	**Machine Learning**
data	Low dimension	High dimension
	i.i.d.	i.i.d./non-i.i.d.
	Single modality	Single/Multiple modality
prior	Given causal relation function type	Functional approximation
Inference	Causation	Dependency



Applications: science, health, recommendation, economics, …

ML for Causal Inference

- Traditional causal inference is often based on many assumptions and limited scenarios
- ML can help causal inference
  - Non-iid data
  - Unknown causal relation types
  - Multimodality
  - Relaxed assumptions

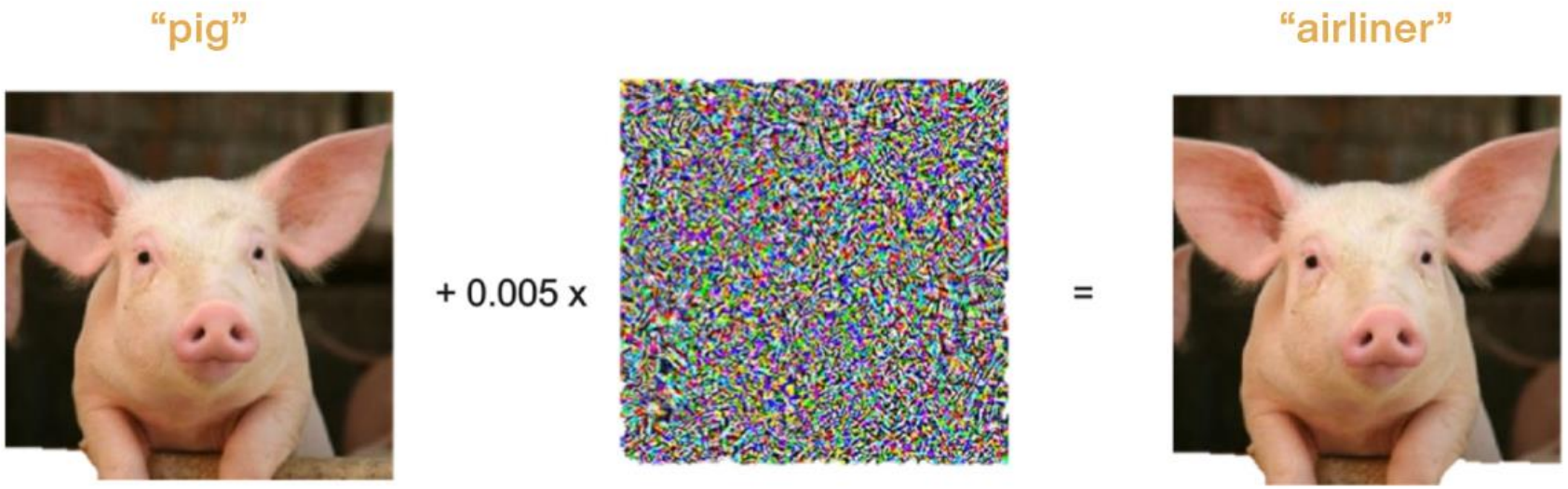
Causal Inference for ML

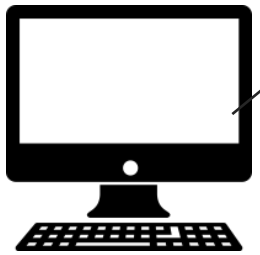
- Traditional ML often only rely on data dependency rather than causation
- Causal Inference can help ML
  - Capture the “cause” of prediction
  - Provide suggestion to change data
  - Infer “what if” cases

Generalization

Train: The animal on with a desert background = camel!
Test: The animal on a grass background = ?

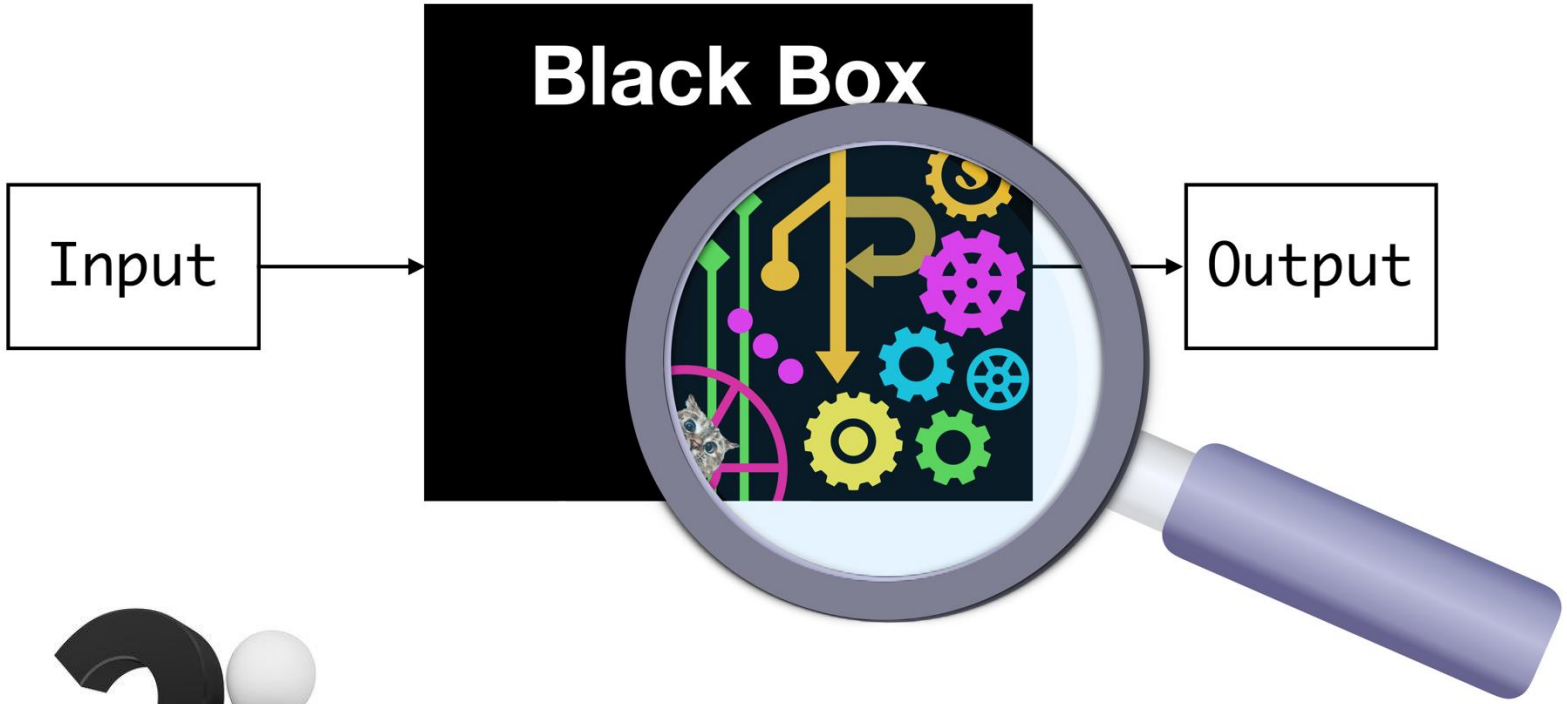
Robustness





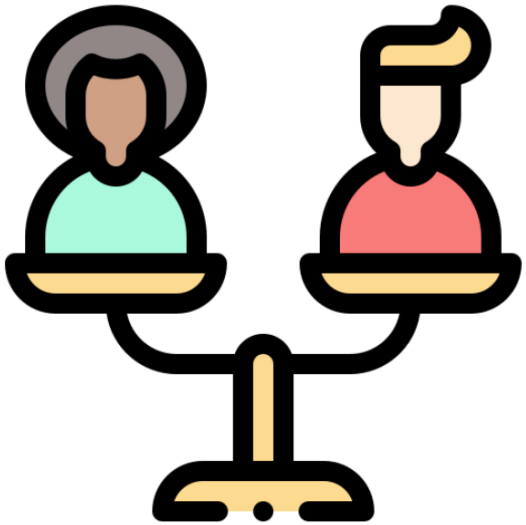
Attack: what perturbation can I make to cause bad prediction?
Defense: Why am I so vulnerable? Can I be more robust by capturing the real "cause"?

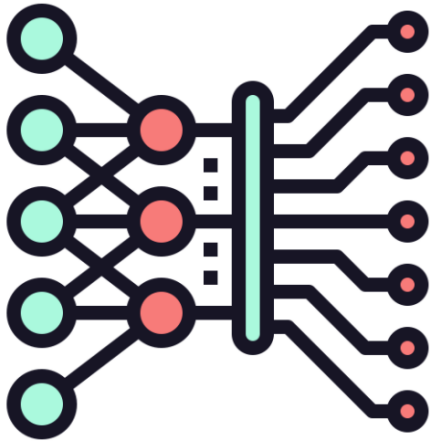
Explanation

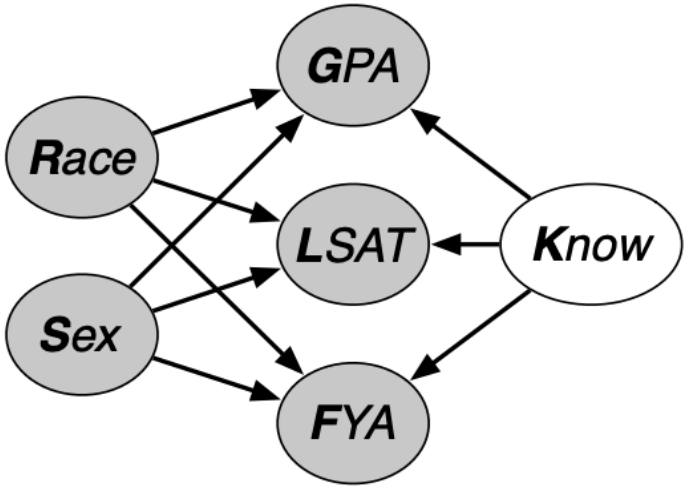


Why does the model make this prediction? Is it reasonable?
What if I change something in the input? Can I get what I want?

Fairness

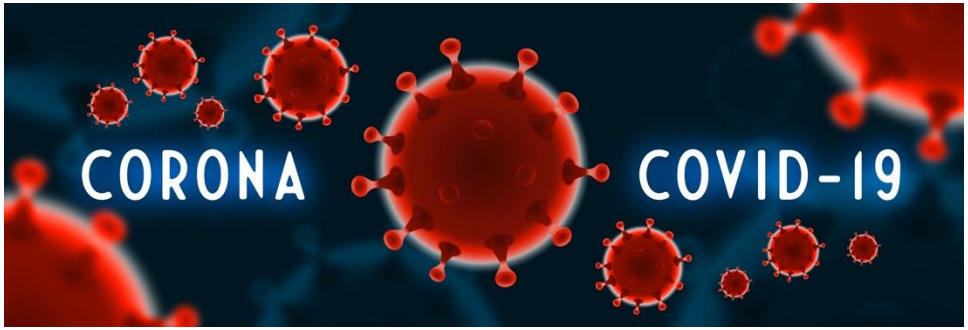






Now we got some predictions! But are they fair for all of us?
What causes unfairness and how can we eliminate it from models?

Applications



Pandemic control

Medicine

Economic analysis

Career plan

AI for Science

Thank you!

jing.ma5@case.edu