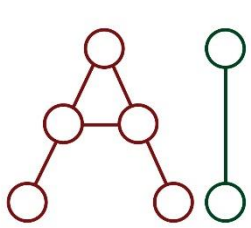


Introduction to Machine Learning

and Introduction to the ChE 197/297 Course

Assoc. Prof. Karl Ezra Pilario, Ph.D.

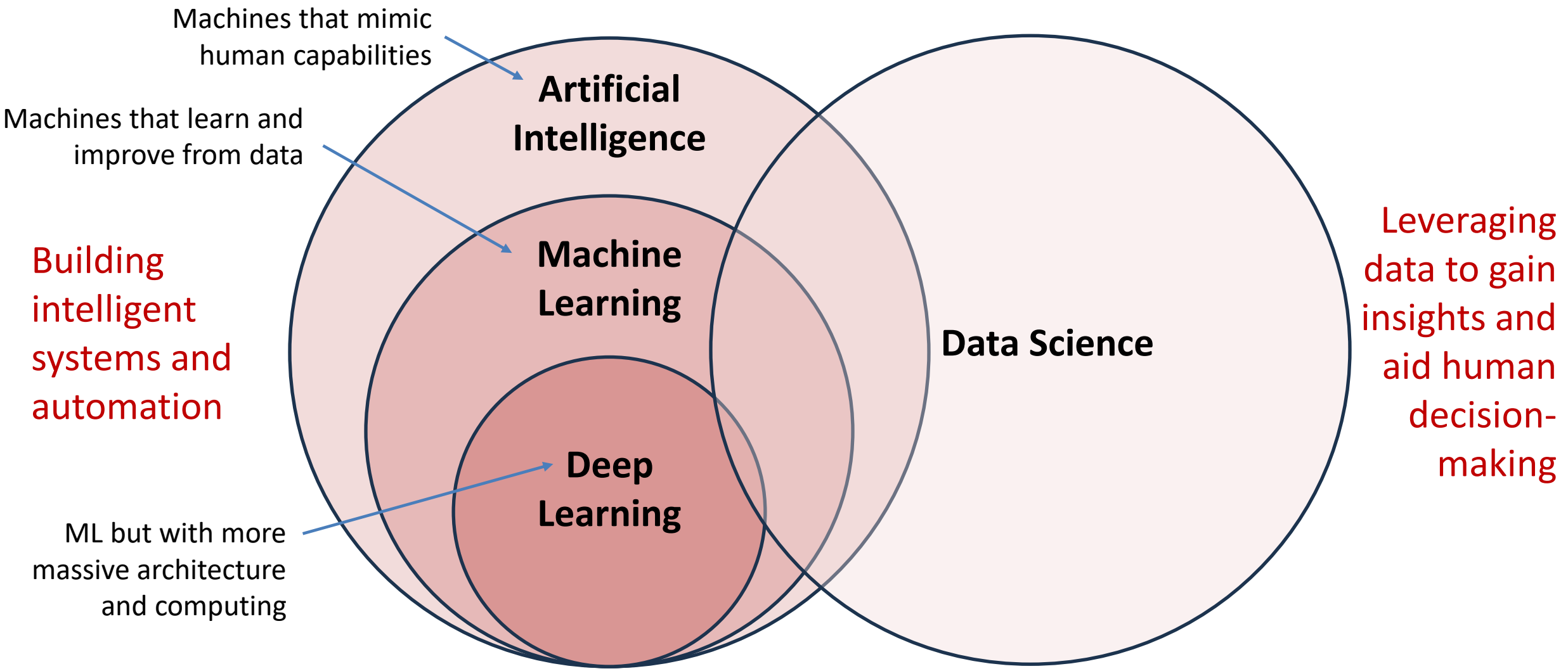
Process Systems Engineering Laboratory
Department of Chemical Engineering
University of the Philippines Diliman

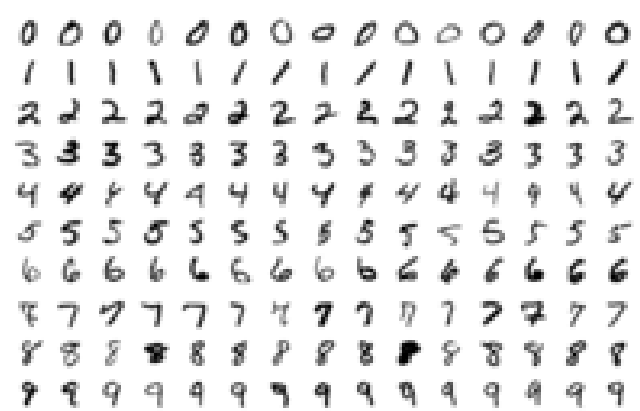
Outline

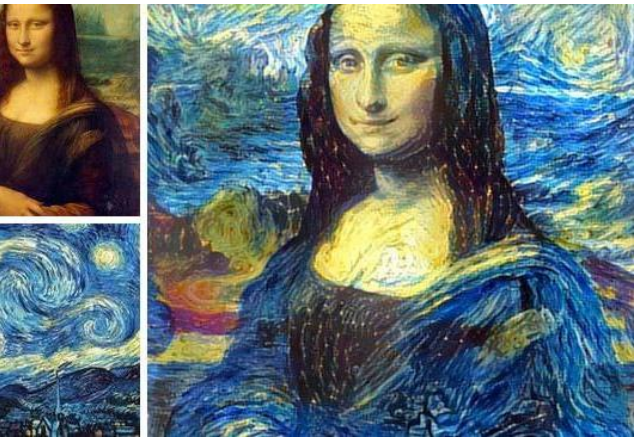
- What is AI / ML / DS?
  - Why only now?
  - Why use them in your industry or field?
  - How to turn data into decisions?
  - Types of Learning Problems

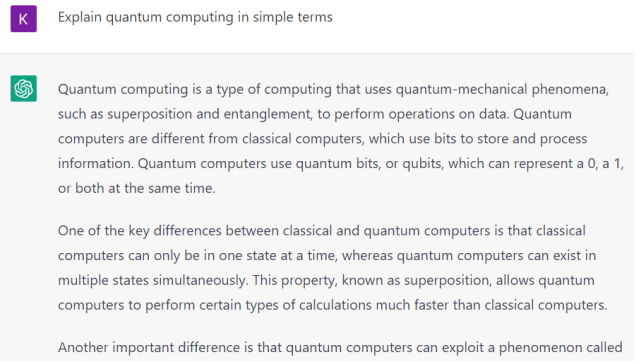
- Intro to the Course
  - Course Delivery
  - Course Content
  - Course Requirements
  - Software

What is AI / ML / DS?





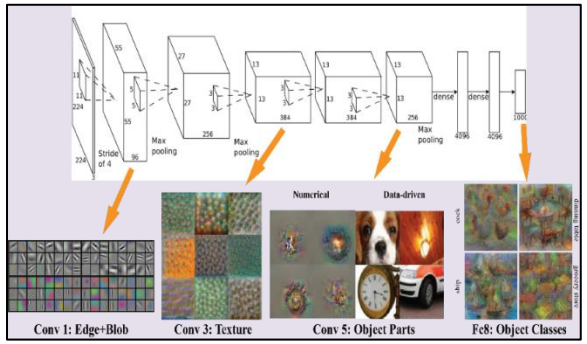




Machine Learning is everywhere…

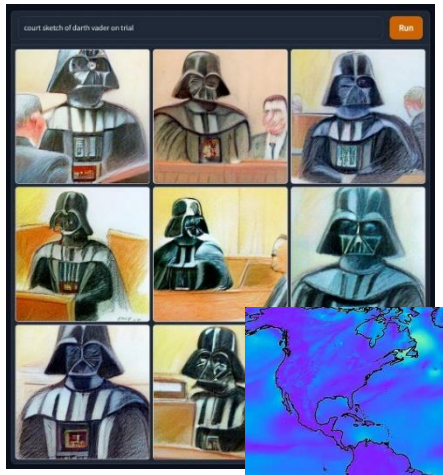
- Filtering emails as spam / not spam
- Handwritten digits recognition
- Speech recognition, Natural Language Processing (NLP)
- Social media (Face recognition, News Feed Ranking, etc.)
- Image / Object Recognition, Image Segmentation
- Recommender systems (movies, products, videos, webpages, bookings)
- Finance (Stock market prediction, customer behaviour, etc.)
- Transportation (Self-driving cars, travel demand modelling)
- Healthcare (Early diagnostics, hospital demand forecasting)
- Bioinformatics (Protein folding and structure prediction, Gene function prediction, Biomedical image analysis)
- Chemometrics (GC-MS data analysis, drug discovery, compound classification, chemical property prediction)

IBM Watson
Jeopardy, **2011**



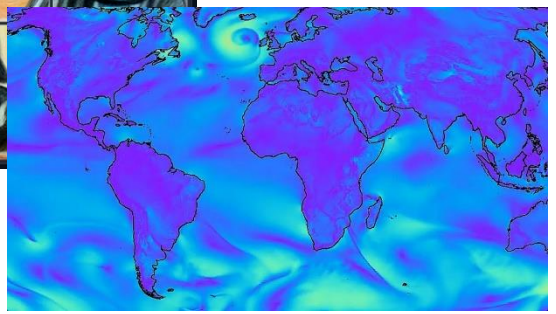
AlexNet
ImageNet Visual Recognition Challenge, **2012**

AlphaGo
Game of Go, **2016**



DALL-E
2021, 2022

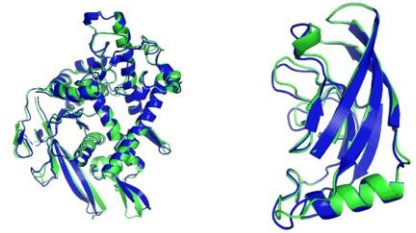
GraphCast
2023



IBM Deep Blue
Chess, **1997**

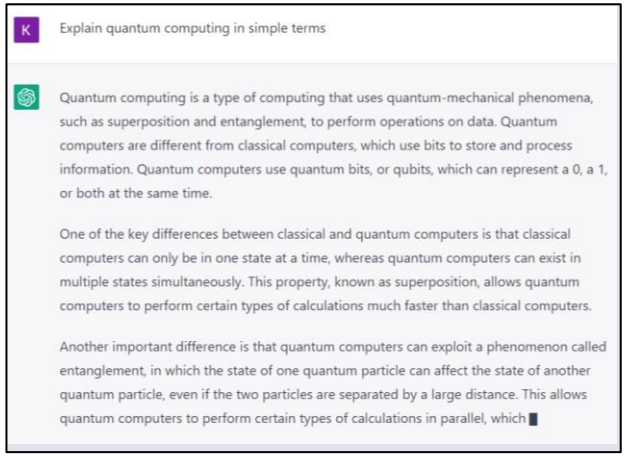
AlphaStar
StarCraft II, **2019**

OpenAI Five
Dota 2, **2019**



AlphaFold
Protein Structure Prediction, **2016**, **2018**

ChatGPT
2022

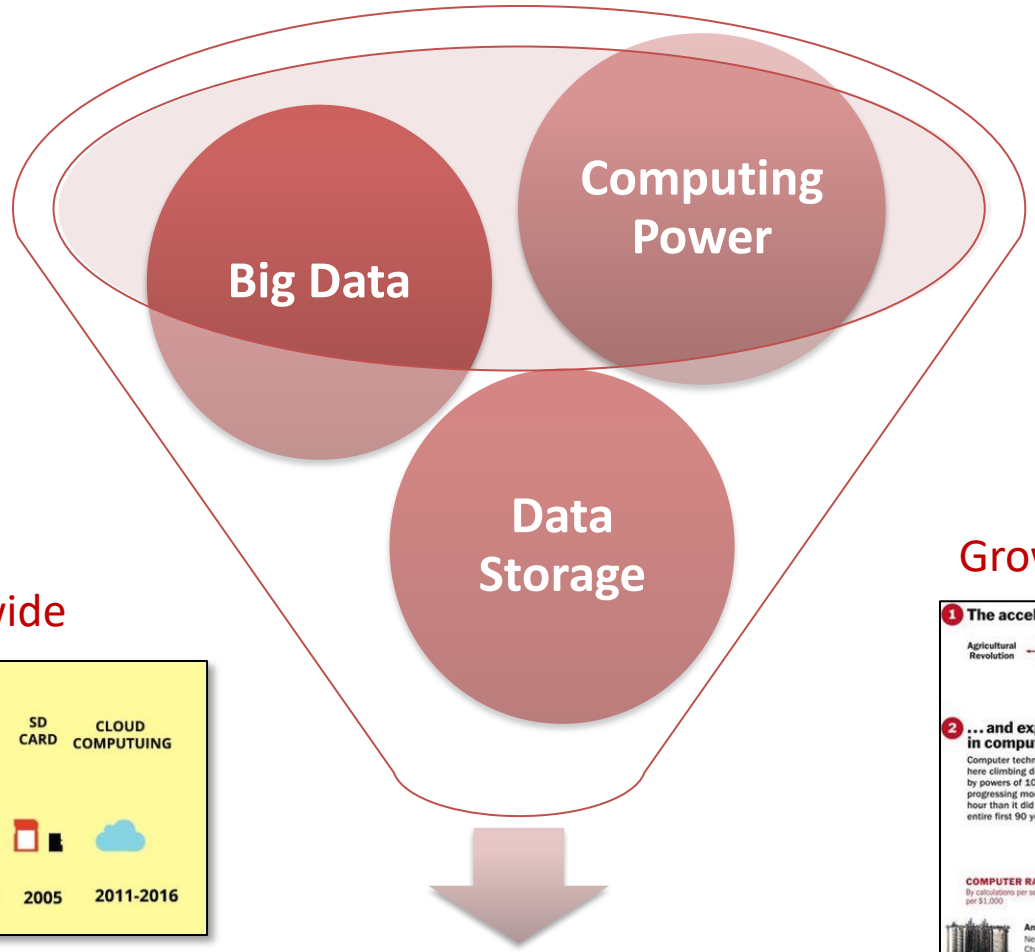


Machine Learning,

Data Science,

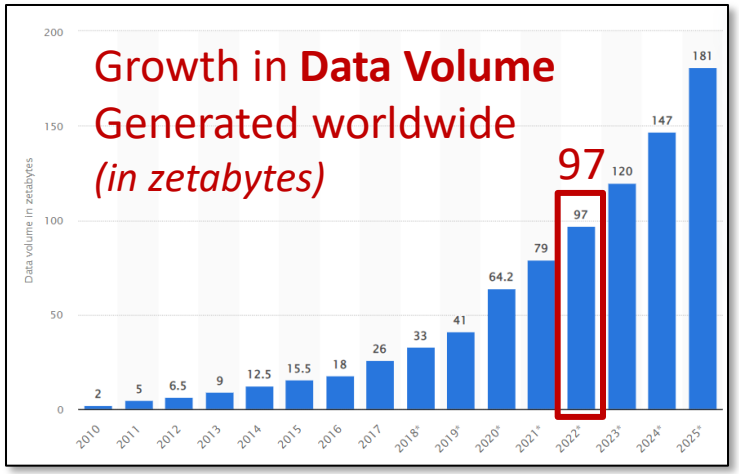
Data Analytics,

…why only now?



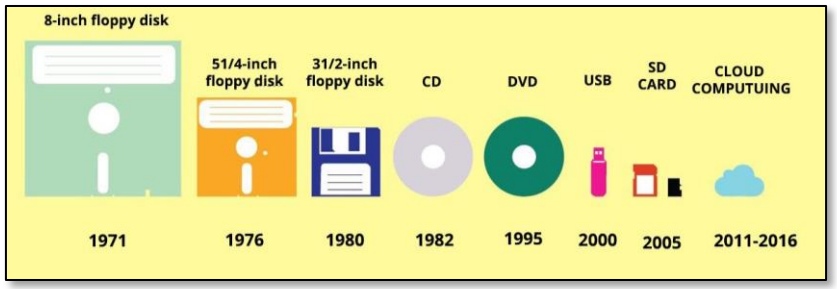
Machine Learning +

Practical Applications

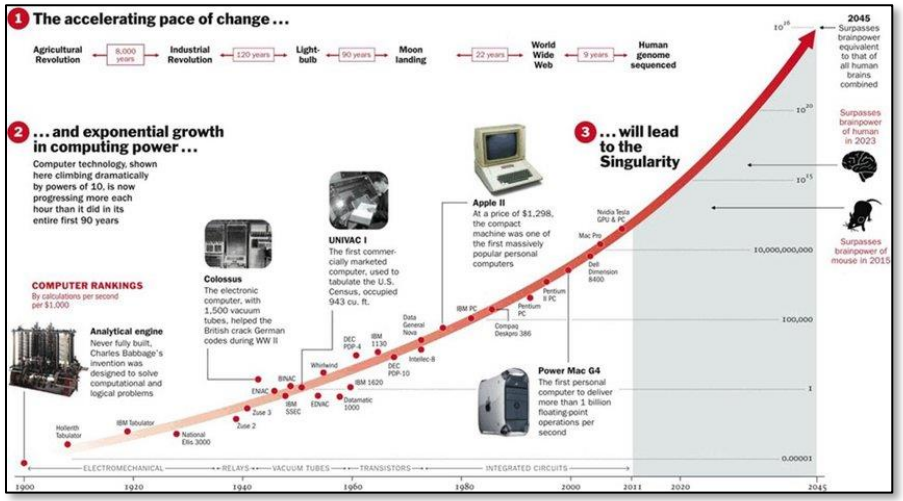


2022

Growth in **Data Storage** worldwide



Growth in **Computing Power** worldwide



Machine Learning,

Data Science,

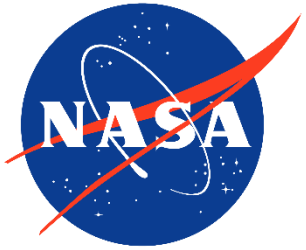
Data Analytics,

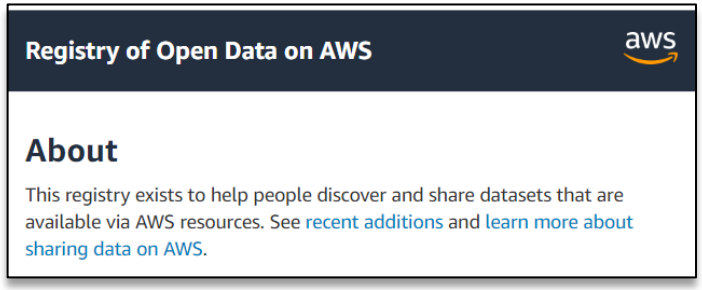
…why only now?

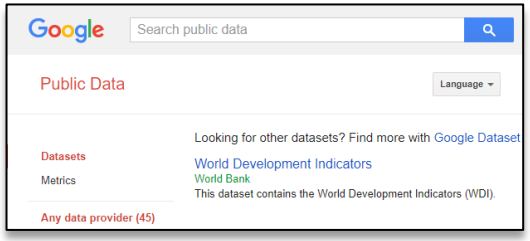
We are currently DROWNING[1] in data!

- There are about 1 trillion web pages.
- 1 hr of video is uploaded to Youtube every second.
- Human genomes have a length of 3.8×10^9 base pairs.
- Walmart handles more than 1 million transactions per hour.
- Etc…

Popular websites where we can get publicly available data:



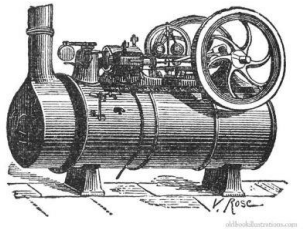




[1] Venkatasubramanian (2009). DROWNING IN DATA: Informatics and Modeling Challenges in a Data-Rich Networked World. *AIChE Journal.*
[2] Murphy (2012). Machine Learning: A Probabilistic Perspective. *MIT Press.*

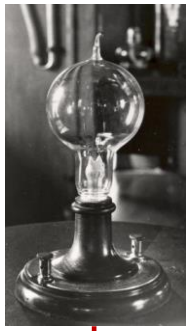
“AI is the new electricity.”

- Andrew Ng, *2017*



Steam Engine
(1700s)

Light Bulb
(Edison)
1879



Only 3% of US households had electricity
1900

50% of US households now had electricity
1920

Electricity

1820 1850 1880 1910

The Between Times
(Agrawal et al, 2022)

AlexNet (2012)

ChatGPT (2022)

A.I.

2010 2040 2070 2100

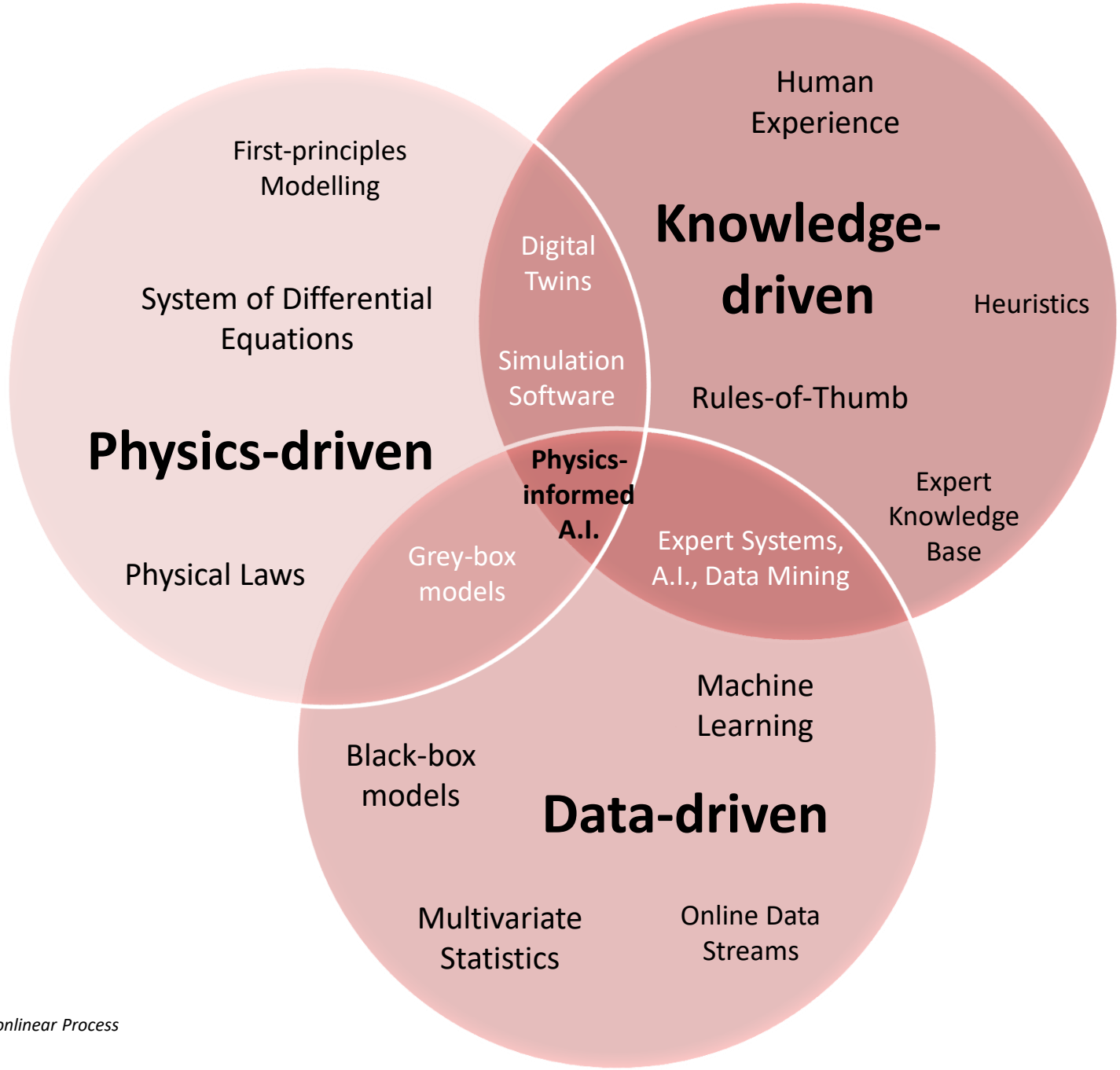
The Between Times?
(Agrawal et al, 2022)

Source:
https://www.wsj.com/video/andrew-ng-ai-is-the-new-electricity/56CF4056-4324-4AD2-AD2C-93CD5D32610A
Agrawal, Gans, Goldfarb (2022). Power and Prediction: The Disruptive Economics of AI.

Why adopt AI/ML/DS in your industry or field?

Three approaches to engineering problems:

1. Physics-driven Methods
2. Knowledge-driven Methods
3. Data-driven Methods

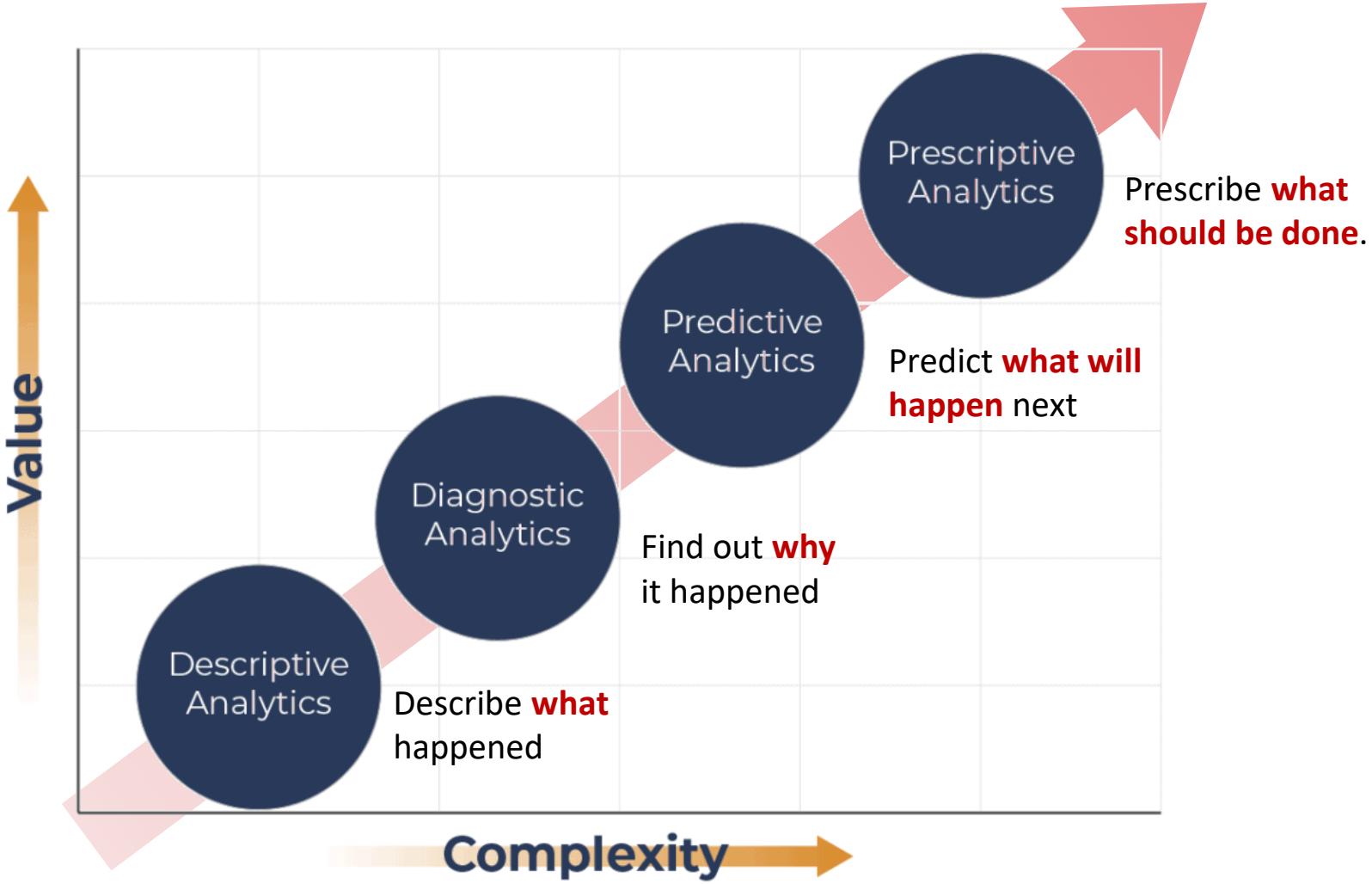


Reference: Pilario et al. (2020), *A Review of Kernel Methods for Feature Extraction in Nonlinear Process Monitoring.* MDPI: Processes, https://doi.org/10.3390/pr8010024

How to turn data into decisions?

Source: https://iterationinsights.com/article/where-to-start-with-the-4-types-of-analytics/

- Applying machine learning to your data is not enough.
- Don’t just let your data speak, let it change the way you do things. **The goal is prescriptive analytics!**
- Getting through each stage of analytics requires more and more effort, but also **more returns**.



The disruptive power of AI?

Source: Agrawal, Gans, Goldfarb (2022). Power and Prediction: The Disruptive Economics of AI.

- Currently, most AI solutions are just "point-solutions".
- For AI to be truly disruptive in any organization, *entire systems* currently in place *must radically change*.
- Barriers exist in each stage. AI adoption is not easy.

Prescribe **what should be done**.

Point-solutions

- Adopt AI for an existing procedure, all else remains the same.
- e.g. At airports, AI can be used to predict aircraft arrivals and congestions.

Application-solutions

- Adopt AI to create a new procedure, all else remains the same.
- e.g. Combine the *air traffic* congestion prediction to *land traffic* congestion prediction (e.g. Waze), then create an app that tells people the best time to leave home for their flight.

System-solutions

- Adopt AI to create a new procedure, but only if all other procedures change.
- e.g. Big airports must be willing to *sacrifice terminal amenities* because, with good AI predictions, travelers now have no reason to stay long in airport terminals.

Outline

- What is AI / ML / DS?
  - …
  - **Types of Learning Problems**

- Intro to the Course (AI 221)
  - Course Delivery
  - Course Content
  - Course Requirements
  - Software

Types of Learning Problems

A simple example…

Supervised Learning

These are images of dogs.

These are images of cars.

Now, what is this an image of?

Unsupervised Learning

Here are some images…

Is there an image that does not belong?

Are there images with similar patterns?

Types of Learning Problems

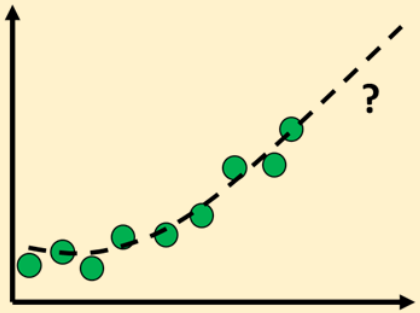
Supervised Learning

Learn a mapping or a function:

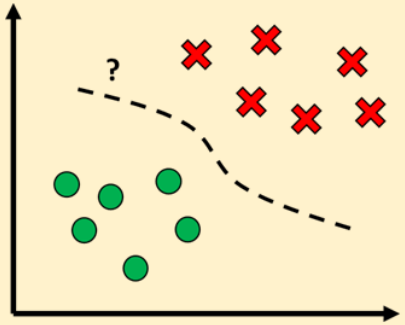
$$y = f(x)$$

from inputs (x) to outputs (y), *given a labelled set of input-output examples* (or).

Regression



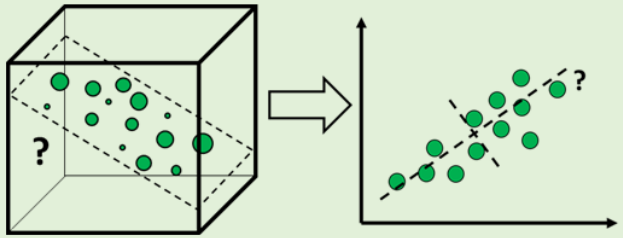
Classification



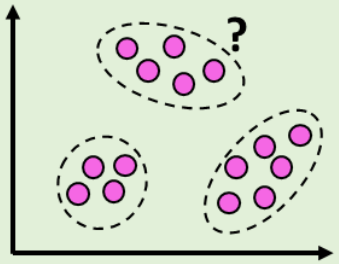
Unsupervised Learning

Discover *patterns or structure* from a data set ( ) without any label information.

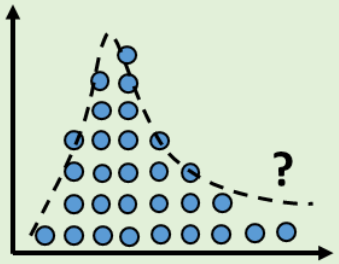
Dimensionality Reduction



Clustering



Density Estimation

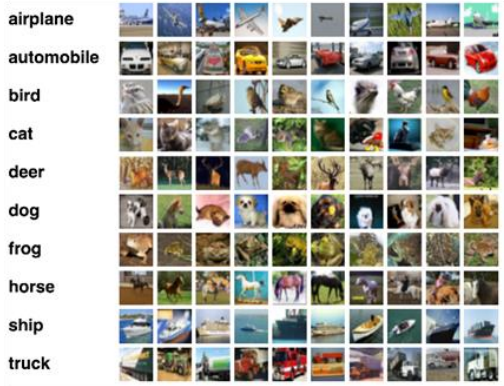


Types of Learning Problems

Semi-Supervised Learning

Goal: Make a computer learn from both labelled and unlabelled data.

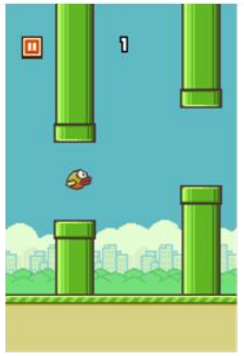
Labelled Data

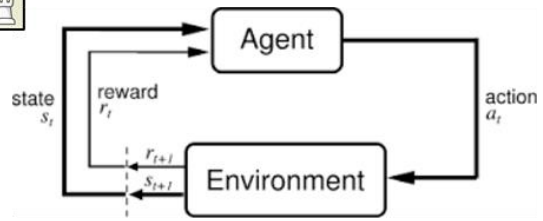


Unlabelled Data

Reinforcement Learning

Goal: Make a computer learn by letting it interact with the environment.





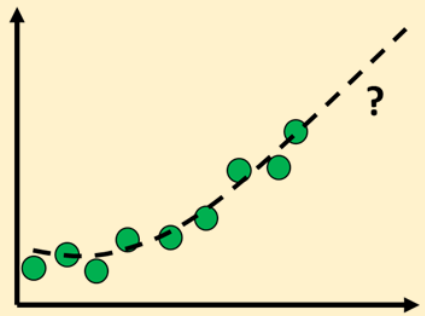
Supervised Learning

Learn a mapping or a function:

$$y = f(x)$$

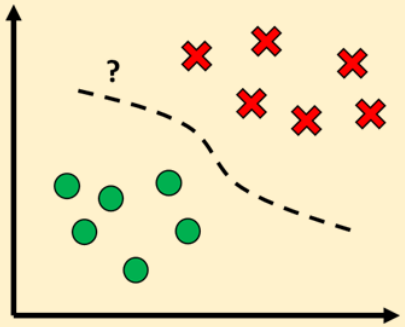
from inputs (x) to outputs (y), *given a labelled set of input-output examples* (● or ✖).

Regression



- **Given:** Training Data $\{x_i, y_i\}_{i=1,2\ldots,N}$
- Target y_i is a **continuous** variable.
- Examples:
  - Forecasting future stock price
  - Forecasting energy resources
  - Prediction of key performance indicators
  - Predicting the properties of molecules based on their structure
  - Predicting the environmental impact of pollutants

Classification

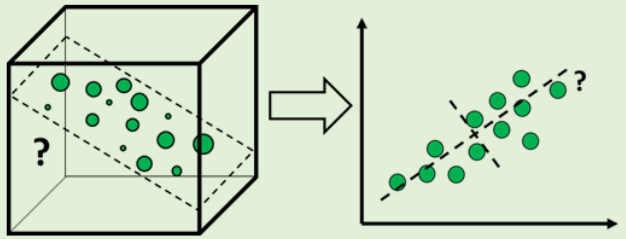


- **Given:** Training Data $\{x_i, y_i\}_{i=1,2\ldots,N}$
- Target y_i is a **categorical** variable.
- Examples:
  - Classifying objects in images
  - Classifying chest X-ray images into COVID positive/negative
  - Handwritten digits recognition
  - Filter e-mails into spam/not spam
  - Classify critical equipment as to healthy or faulty
  - Activity recognition from wearable devices

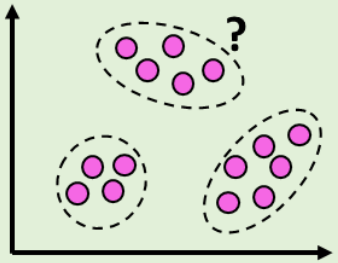
Unsupervised Learning

Discover *patterns or structure* from a data set (●) without any label information.

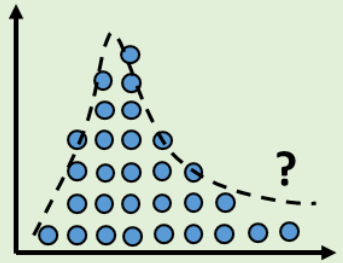
Dimensionality Reduction



Clustering



Density Estimation



Dimensionality Reduction

- **Given:** Data $\{x_i\}_{i=1,2\ldots,N}$
- **Reduce features** but retain the most important information from the original data.
- Examples:
  - Feature Engineering
  - Image compression
  - Filtering noise from signals
  - Source separation in audio
  - Data visualization

Clustering

- **Given:** Data $\{x_i\}_{i=1,2\ldots,N}$
- **Group** similar data points together.
- Examples:
  - Customer segmentation
  - Recommendation systems
  - Identifying fake news
  - Clustering documents, tweets, posts

Density Estimation

- **Given:** Data $\{x_i\}_{i=1,2\ldots,N}$
- **Estimate** the distribution of the data.
- Examples:
  - Anomaly Detection
  - Novelty Detection
  - Generative Models
  - Finding distribution modes
  - Spatio-temporal analytics

Can you identify the type of learning problem?

Regression, Classification, Dimensionality Reduction, Clustering, Density Estimation

Example 1

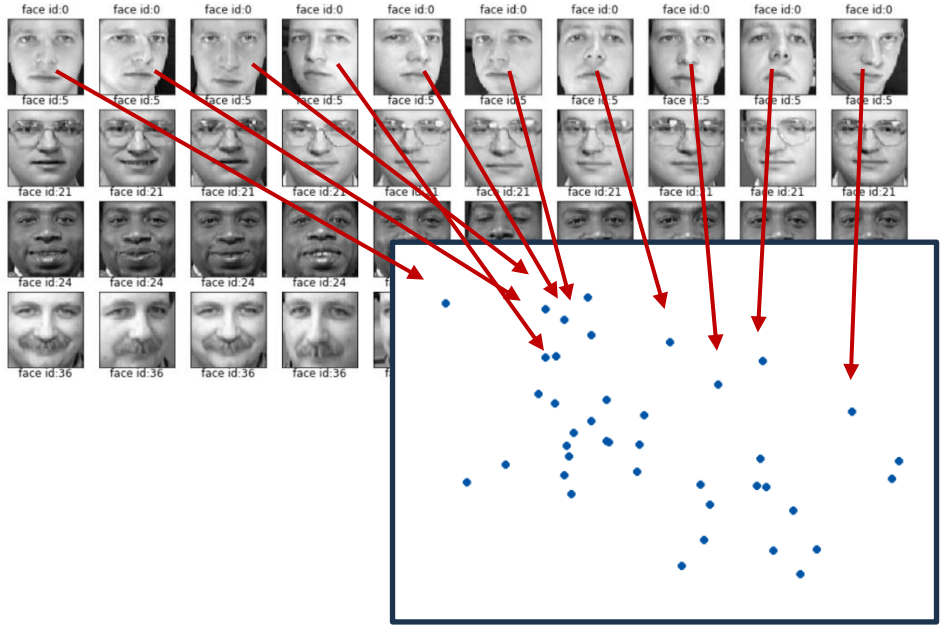
Given the weight of the car, its model year, and horsepower, predict its mileage in miles per gallon (mpg).

car_weight	model_year	horsepower	mileage
1522 kg	2020	150	18 mpg
1930 kg	2017	185	16 mpg
1321 kg	2018	200	21 mpg
2128 kg	2019	168	?
2498 kg	2018	170	15 mpg
1882 kg	2021	155	17 mpg
1956 kg	2019	190	?
1672 kg	2017	182	18 mpg

Answer: **Regression**

Example 2

Given images of faces with varying poses and expressions, *map* each image onto a 2D point so that similar-looking images are closer together on the map.



Answer: **Dimensionality Reduction**

Can you identify the type of learning problem?

Regression, Classification, Dimensionality Reduction, Clustering, Density Estimation

Example 3

Given a tweet, predict whether the sentiment is positive, negative, or neutral.

Tweet	Sentiment
I'm in pain…	Negative
Manifesting a promotion this year!	Positive
It's 2AM. Who's awake?	Neutral
Heavy traffic at EDSA	Negative
Family dinner… So full!	Positive
Spoiler alert: RIP Tony Stark	?
Tesla sucks!	?
It's a boy!	Positive

Answer: **Classification**

Example 4

Given student grades in 5 subjects: Math, Chemistry, Physics, English, and Reading, group the students with similar competencies.

Student	Math	Chemistry	Physics	English	Reading
1	81	85	88	94	92
2	95	80	94	93	85
3	92	94	89	81	80
4	94	83	90	91	84
5	88	84	90	97	95
6	90	93	88	85	82
7	92	94	91	87	81
8	87	82	85	93	94

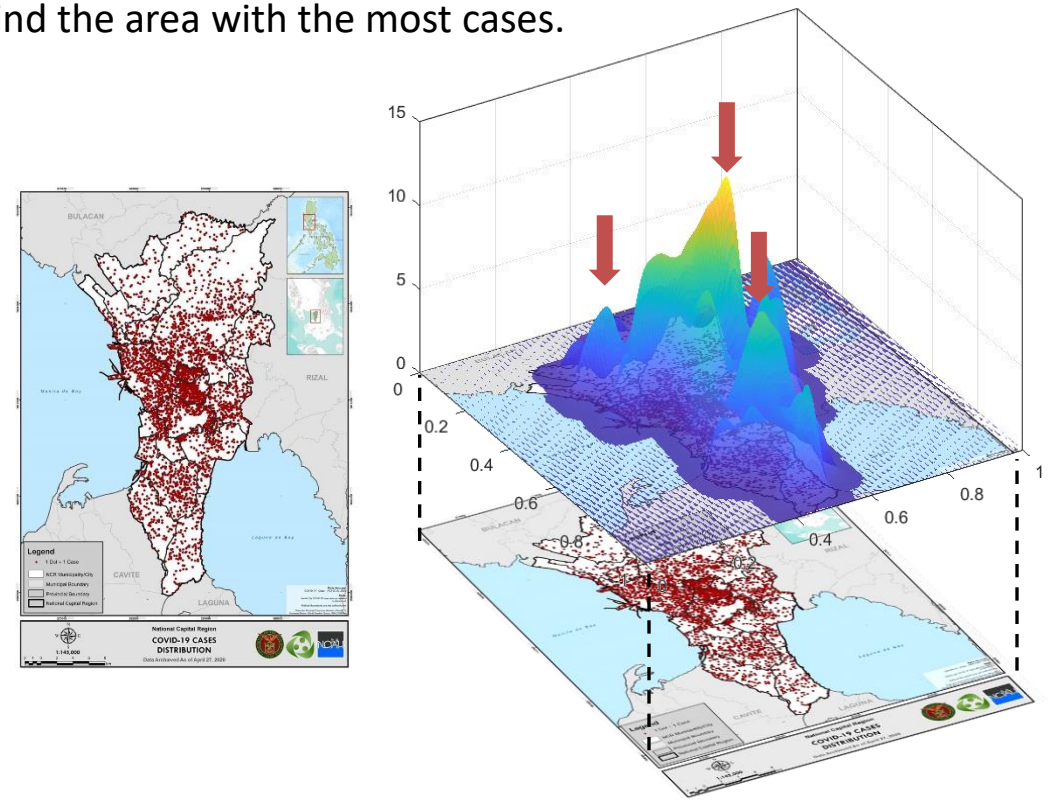
Answer: **Clustering**

Can you identify the type of learning problem?

Regression, Classification, Dimensionality Reduction, Clustering, Density Estimation

Example 5

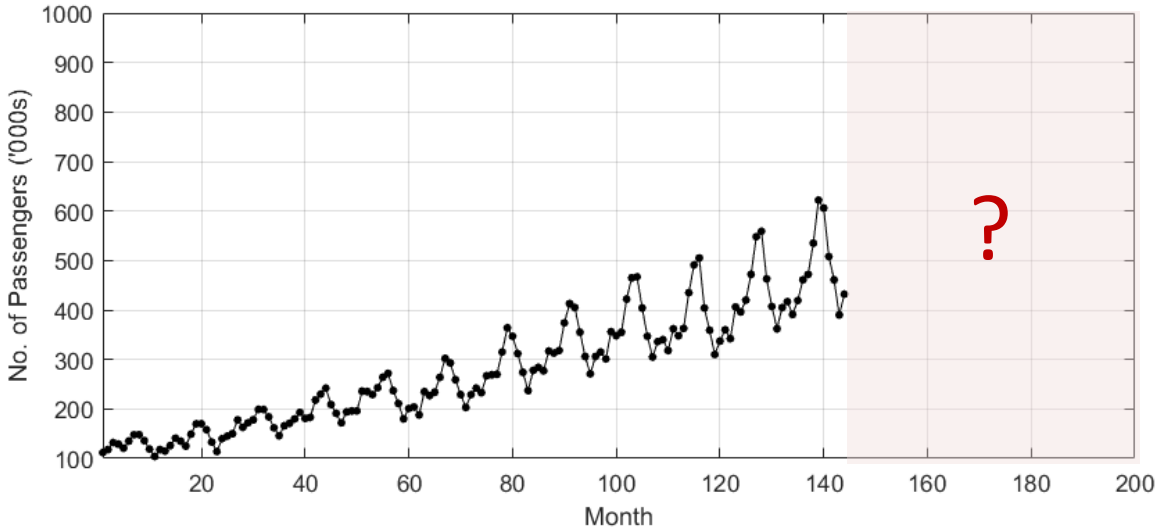
Given the spatial occurrence of Covid cases in Metro Manila, find the area with the most cases.



Answer: Density Estimation

Example 6

Given the number of airline passengers in the previous months, predict the number of passengers for the next few months.



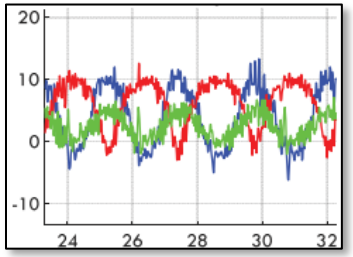
Answer: Regression

Can you identify the type of learning problem?

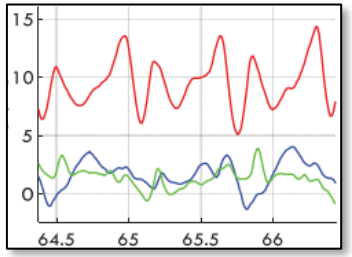
Regression, Classification, Dimensionality Reduction, Clustering, Density Estimation

Example 7

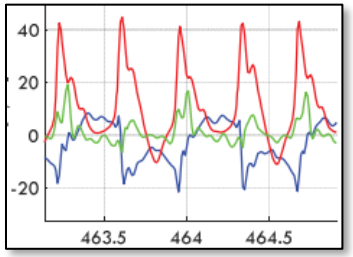
Given smartphone *accelerometer data* from a human doing exercise, predict the kind of exercise being done.



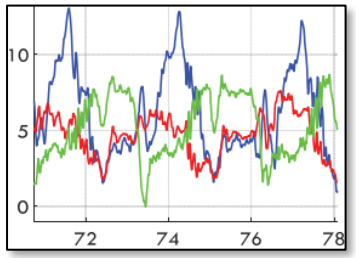
Push-up



Walking



Running



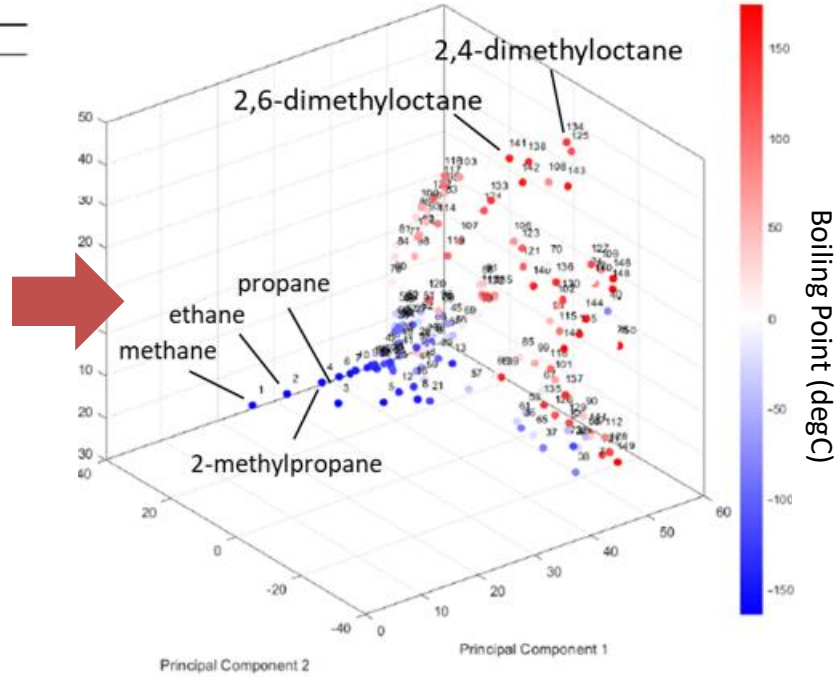
Squats

Answer: **Classification**

Example 8

Given the structural properties of alkane molecules, *map* them onto 3D space based on their similarities, then *predict* their boiling points.

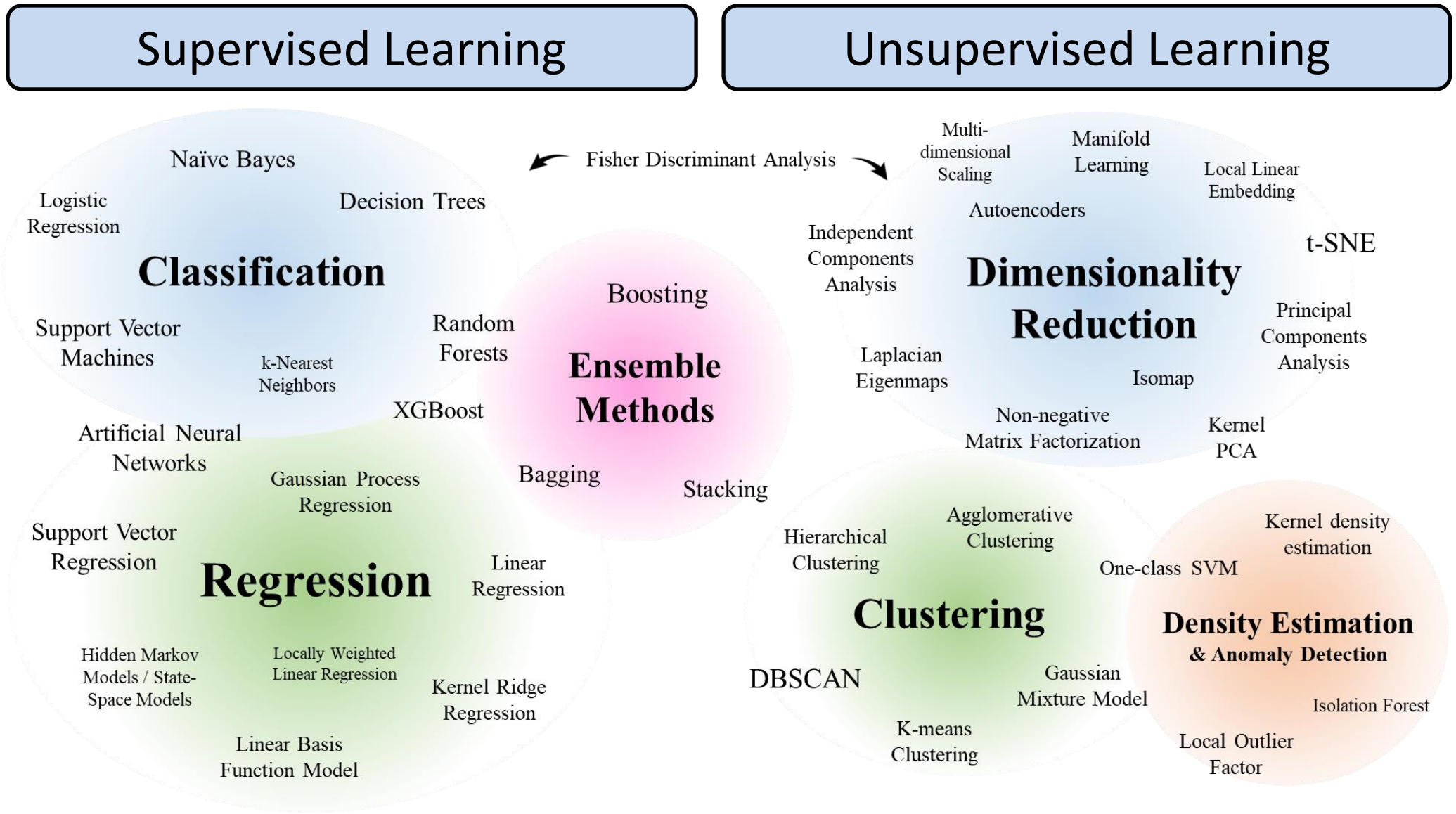
No.	BP	Alkane
1	-164	methane
2	-88.6	ethane
3	-42.1	propane
4	-11.7	2-methylpropane
5	-0.5	butane
6	9.5	2,2-dimethylpropane
7	27.8	2-methylbutane
8	36.1	pentane
9	49.7	2,2-dimethylbutane
10	58	2,3-dimethylbutane
11	60.3	2-methylpentane
12	63.3	3-methylpentane
13	69	hexane
14	80.9	2,2,3-trimethylbutane
15	79.2	2,2-dimethylpentane
16	86.1	3,3-dimethylpentane
17	89.8	2,3-dimethylpentane
18	80.5	2,4-dimethylpentane
19	90	2-methylhexane
20	92	3-methylhexane



Answer: **Dimensionality Reduction + Regression**

Source: https://ieeexplore.ieee.org/document/8472444

Classical Machine Learning Methods



Reference: Pilario et al. (2020), *A Review of Kernel Methods for Feature Extraction in Nonlinear Process Monitoring.* MDPI: Processes, https://doi.org/10.3390/pr8010024

Outline

- What is AI / ML / DS?
  - Why only now?
  - Why use them in your industry or field?
  - How to turn data into decisions?
  - Types of Learning Problems

- Intro to the Course
  - Course Delivery
  - Course Content
  - Course Requirements
  - Software