McMaster University ENGINEERING

W BOOTH SCHOOL OF ENGINEERING PRACTICE AND TECHNOLOGY

SEP 785: Machine Learning

Lecture 1: Introduction

Instructor: Dr. Dalia Mahmoud, PhD
(Mechanical Engineering, McMaster University)
Email: mahmoudd@mcmaster.ca

Land Acknowledgement

We recognize and acknowledge that students of McMaster University meet and learn on the traditional territories of the Mississauga and Haudenosaunee nations, and within the lands protected by the "Dish With One Spoon" wampum, an agreement to peaceably share and care for the resources around the Great Lakes.

McMaster University

Course Main Objective

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

Provide practical machine learning skills and a solid understanding of the underlying mathematics for application in research and professional work after graduation.

Course Intended Learning Outcomes

- Explain the principles of **supervised learning** and evaluate its suitability for tasks such as **classification** and **regression**.
- Design and implement **machine learning pipelines** using **Python**, integrating appropriate **preprocessing techniques**.
- Select, apply, and interpret **evaluation metrics** to **assess model performance** in classification and regression contexts.
- Utilize advanced techniques like **ensemble** methods, **clustering algorithms**, and recommendation systems to solve practical problems.
- **Debug learning algorithms** and understand what goes on beneath the hood.

Course Assessment

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

In class participation	5
Assignments	35
Final Project	60
Total	100

Course Suggested Topics

- Introduction, Terminology and Baseline
- Data Pre-processing
- Linear Discriminant Analysis
- Classifier Performance and Model Selection
- K Nearest Neighbors Classification
- Bayesian Classifiers
- Decision Trees
- Boosting and AdaBoot
- Logistic Regression
- Support Vector Machines

Lecture Contents

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

- Introduction
- Terminology
- Setting up Jupitar notebook
- Data !!

Lecture Intended Learning Outcomes

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

- Explain the **motivation** to study **machine learning**.
- Identify whether a given problem could be solved using **supervised machine learning** or not.
- Differentiate between **supervised** and **unsupervised** machine learning.
- Explain machine learning terminology such as **features**, **targets**, **predictions.**
- Differentiate between **classification** and **regression** problems.
- Compare different **types of Data**.

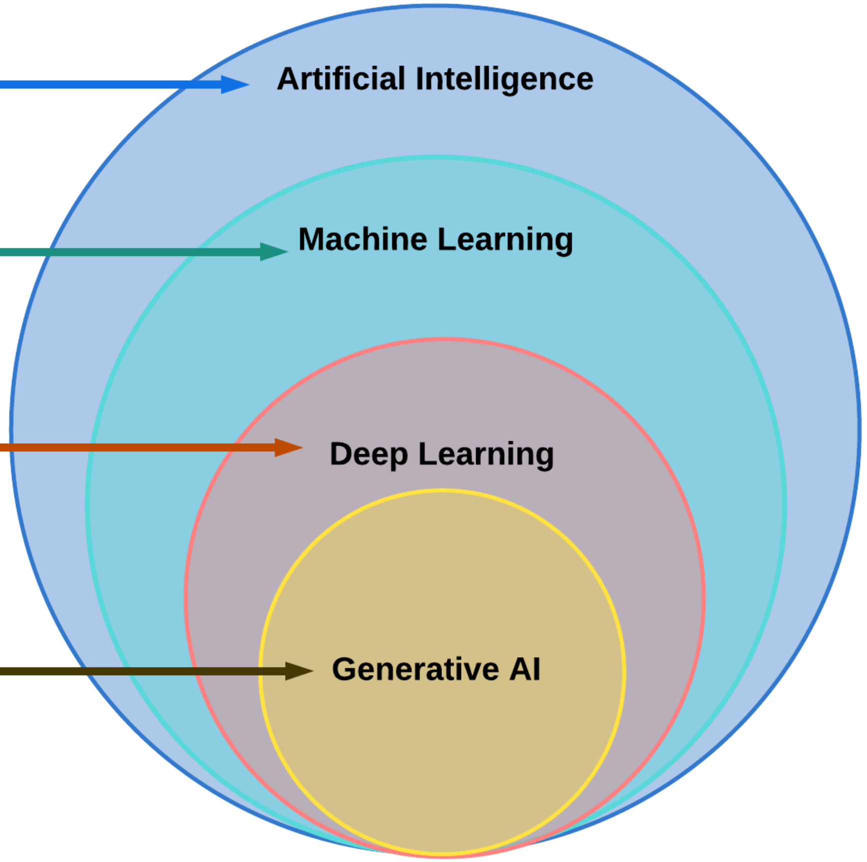
The bigger picture

Artificial Intelligence
AI involves techniques that equip computers to emulate human behavior, enabling them to learn, make decisions, recognize patterns, and solve complex problems in a manner akin to human intelligence.

Machine Learning
ML is a subset of AI, uses advanced algorithms to detect patterns in large data sets, allowing machines to learn and adapt. ML algorithms use supervised or unsupervised learning methods.

Deep Learning
DL is a subset of ML which uses neural networks for in-depth data processing and analytical tasks. DL leverages multiple layers of artificial neural networks to extract high-level features from raw input data, simulating the way human brains perceive and understand the world.

Generative AI
Generative AI is a subset of DL models that generates content like text, images, or code based on provided input. Trained on vast data sets, these models detect patterns and create outputs without explicit instruction, using a mix of supervised and unsupervised learning.



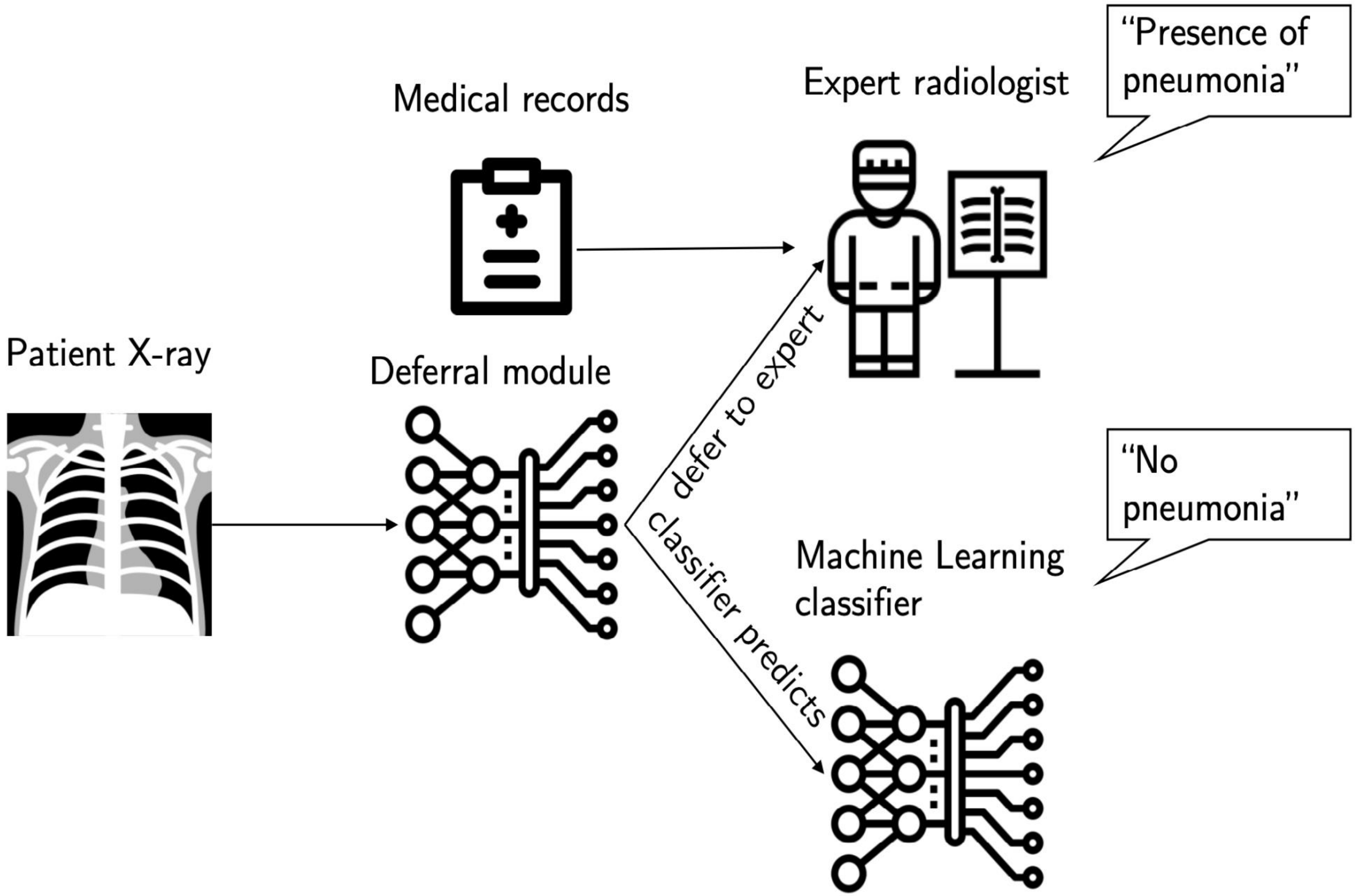
Unraveling AI Complexity - A Comparative View of AI, Machine Learning, Deep Learning, and Generative AI.

(Created by Dr. Lily Popova Zhuhadar, 07, 29, 2023)

Machine Learning Applications: Healthcare and medical diagnosis

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

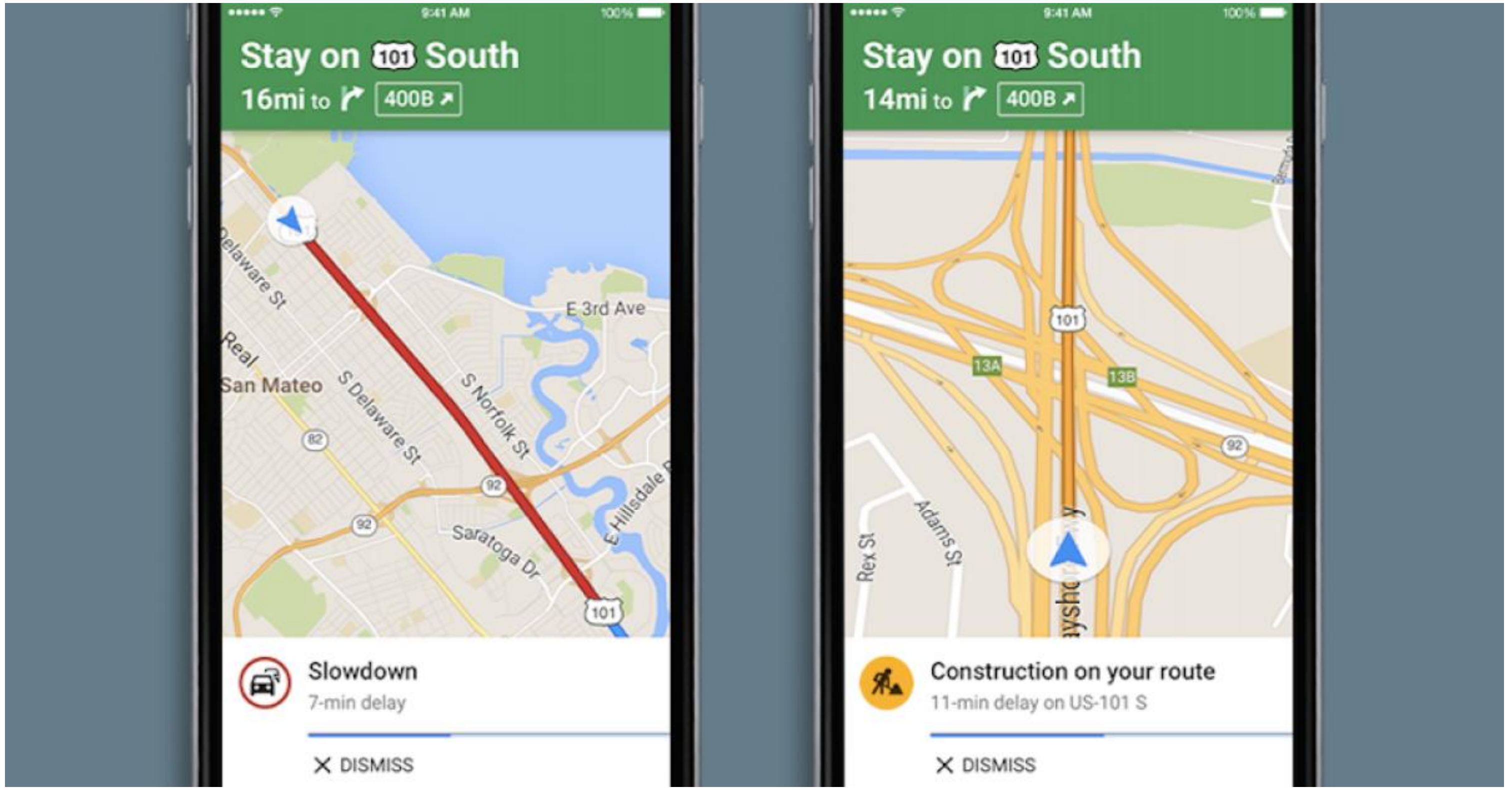
Improving medical and diagnostics paved the way for thorough analysis and improved treatment diagnosis.



Machine Learning Applications : Traffic Alerts

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

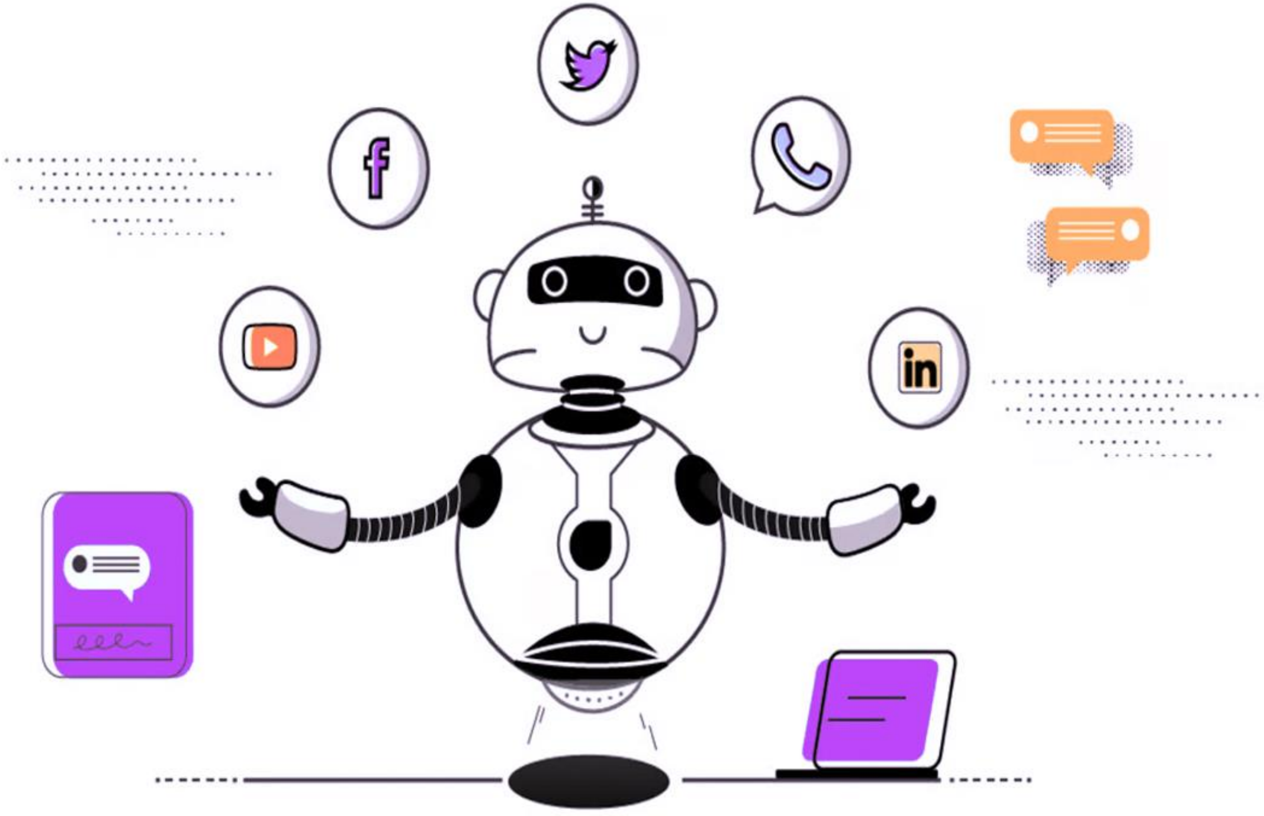
Google Maps utilizes cutting-edge Machine Learning methods and historical knowledge to predict traffic.



Machine Learning Applications : Chatbot

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

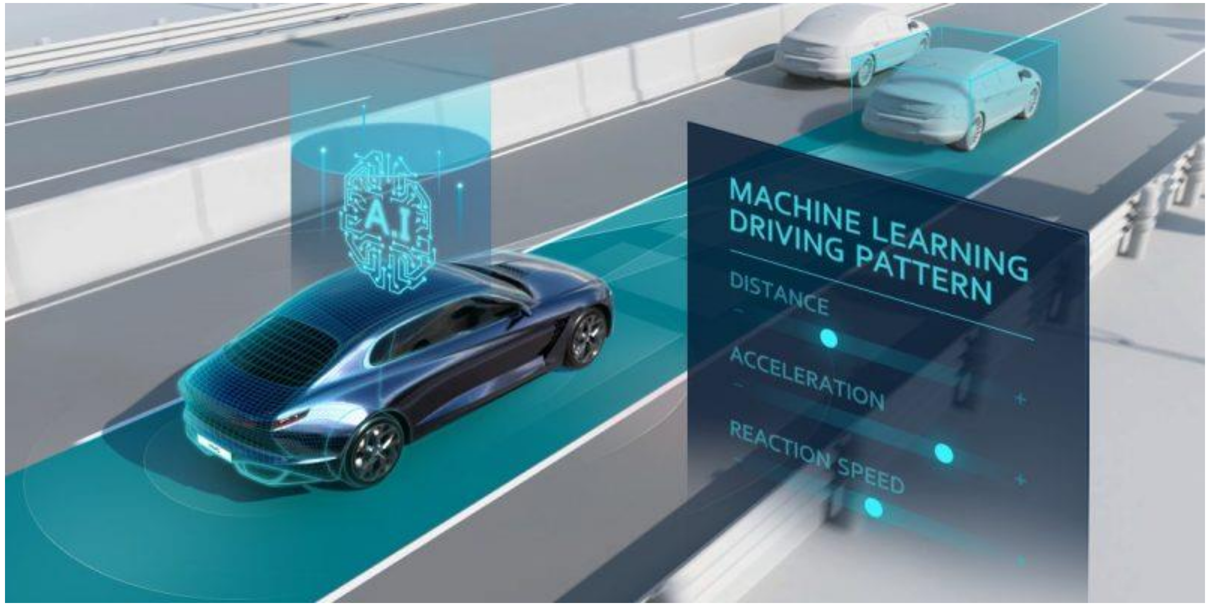
Chatbots can interpret the context of a conversation using Machine Learning and then react appropriately.



Machine Learning Applications : Self Driving Cars

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

- The role of Machine Learning in autonomous vehicles enables the vehicle to learn from data and make predictions about the world.
- Machine Learning algorithms can predict the behavior of objects, pedestrians, people, and other vehicles on the road.



Machine Learning Applications : Ads Recommendations

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

- Machine Learning predicts which ads are most relevant and effective for users.
- Machine Learning can segment users into different groups, allowing advertisers to tailor their ads and improve their relevance.

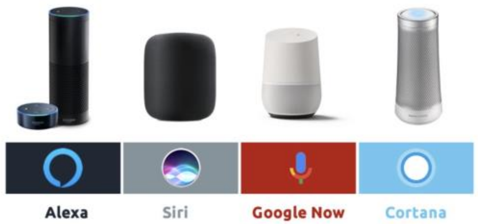
Machine Learning is All around us

McMaster University

ENGINEERING

W Booth School of Engineering Practice and Technology

Voice assistants

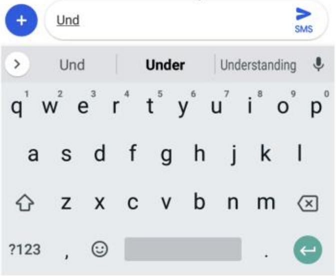


Google news

Recommendation systems

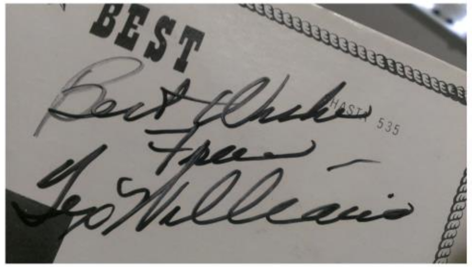
Face recognition

Auto-completion



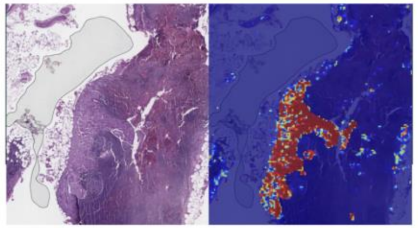
Stock market prediction

Character recognition



Self-driving car

Cancer diagnosis



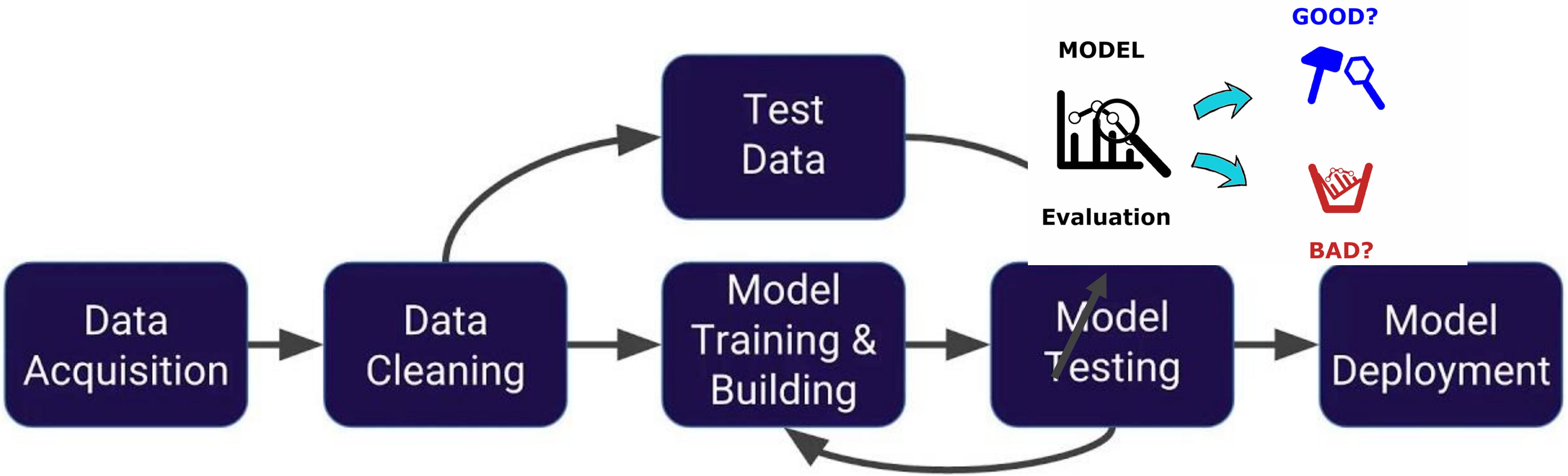
Drug discovery

AlphaGo

Why Use Machine Learning

- Imagine writing a program for spam identification, i.e., whether an email is spam or non-spam.
- Traditional programming
  - Come up with rules using human understanding of spam messages.
  - Time consuming and hard to come up with robust set of rules.
- Machine learning
  - Collect large amount of data of spam and non-spam emails and let the machine learning algorithm figure out rules.
- With machine learning, you’re likely to
  - Save time
  - Customize and scale products

Machine Learning Process



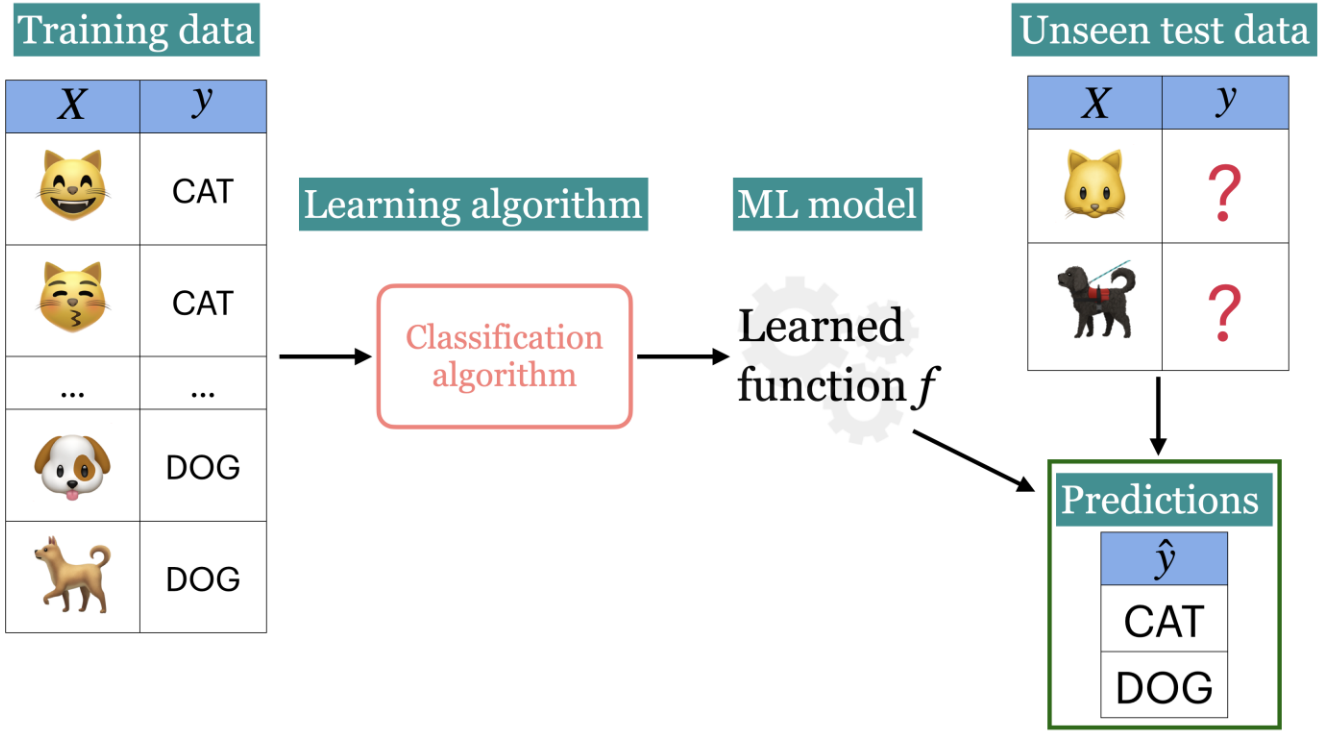
Types of machine learning

- Here are some typical learning problems.
- Supervised learning (Gmail spam filtering)
  - Training a model from input data and its corresponding targets to predict targets for new examples.
- Unsupervised learning (Google News)
  - Training a model to find patterns in a dataset, typically an unlabeled dataset.
- Reinforcement learning (AlphaGo)
  - A family of algorithms for finding suitable actions to take in a given situation in order to maximize a reward.
- Recommendation systems (Amazon item recommendation system)
  - Predict the “rating” or “preference” a user would give to an item.

What is supervised machine learning (ML)?

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

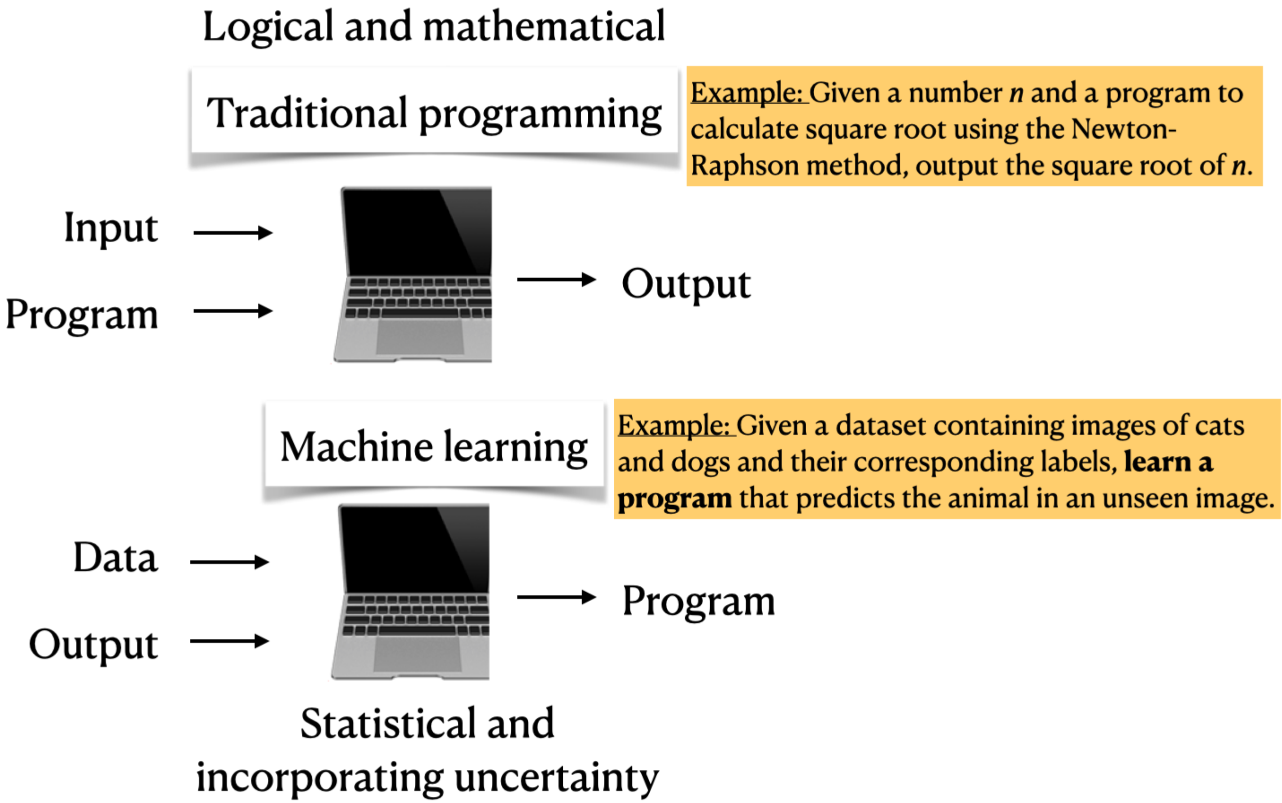
- Training data comprises a set of observations (X) and their corresponding targets (y).



- We wish to find a model function f that relates X to y.
- We use the model function to predict targets of new examples.

(Supervised) machine learning: popular definition

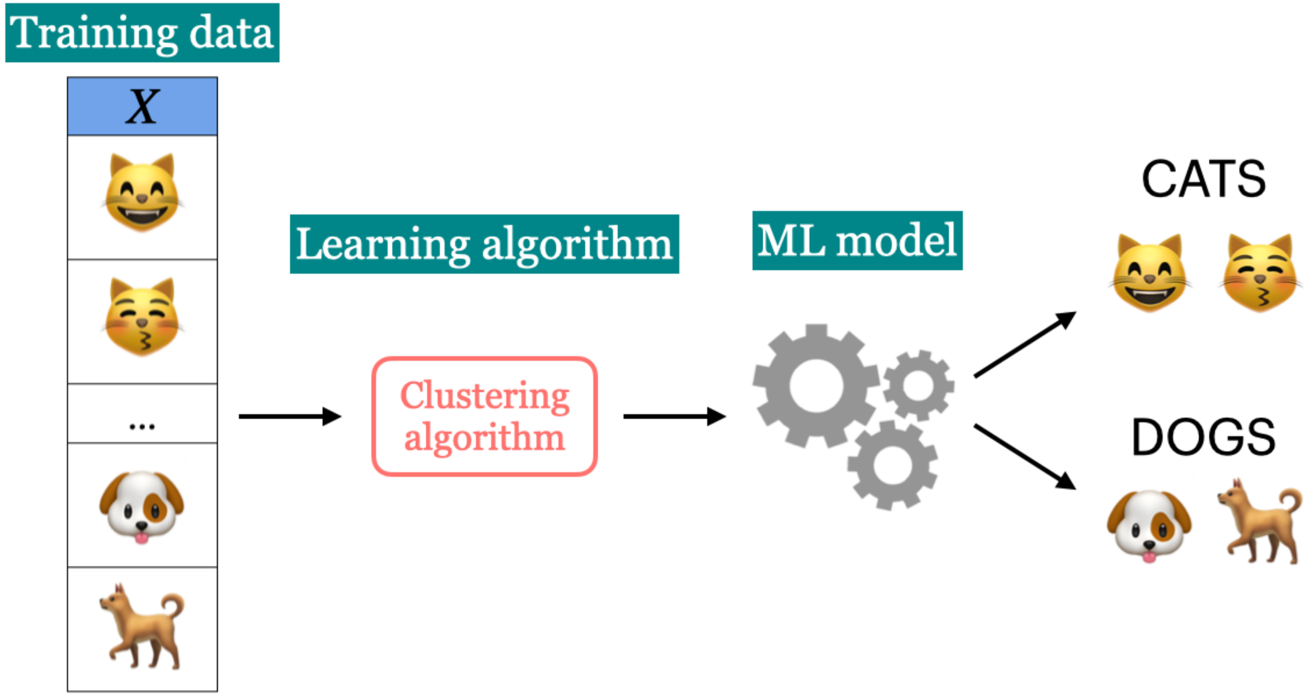
- A field of study that gives computers the ability to learn without being explicitly programmed.-- Arthur Samuel (1959)
- ML is a different way to think about problem solving.



Unsupervised learning

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

- In unsupervised learning training data consists of observations (X) without any corresponding targets. Unsupervised learning could be used to group similar things together in X or to provide concise summary of the data. We'll learn more about this topic in later videos



Lecture Contents

- Introduction
- Terminology
- Setting up Jupitar notebook
- Data !!

Tabular data

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

- In supervised machine learning, the input data is typically organized in a tabular format, where rows are examples and columns are features. One of the columns is typically the target.

X Features (d) — y Target

Examples (n)

ml_experience	class_attendance	lab1	lab2	lab3	lab4	quiz1	quiz2
1	1	91	93	88	92	94	A+
1	0	78	87	88	85	80	not A+
…	…	…	…	…	…	…	…
0	1	69	75	65	80	65	not A+

Features and Targets

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

- Features are relevant characteristics of the problem, usually suggested by experts. Features are typically denoted by X and the number of features is usually denoted by (d).
- Target is the feature we want to predict, typically denoted by (y).

X — Features (d)

y — Target

Examples (n)

ml_ experience	class_ attendance	lab1	lab2	lab3	lab4	quiz1	quiz2
1	1	91	93	88	92	94	A+
1	0	78	87	88	85	80	not A+
...	...	...	...	...	...	...	...
0	1	69	75	65	80	65	not A+

Examples

- A row of feature values. When people refer to an example, it may or may not include the target corresponding to the feature values, depending upon the context. The number of examples is usually denoted by (n) .

X Features (d) — y Target

Examples (n)

ml_ experience	class_ attendance	lab1	lab2	lab3	lab4	quiz1	quiz2
1	1	91	93	88	92	94	A+
1	0	78	87	88	85	80	not A+
...	...	...	...	...	...	...	...
0	1	69	75	65	80	65	not A+

Classification vs. Regression

- In supervised machine learning, there are two main kinds of learning problems based on what they are trying to predict.
- Classification problem: predicting among two or more discrete classes
  - Example1: Predict whether a patient has a liver disease or not
  - Example2: Predict whether a student would get an A+ or not in quiz2.
- Regression problem: predicting a continuous value
  - Example1: Predict housing prices
  - Example2: Predict a student's score in quiz2.

Classification vs. Regression

Predict whether a student would get an A+ or not in quiz2

Classification model

ml_ experience	...	quiz1	quiz2
0	...	90	A+
1	...	91	A+
...	...	...	...
0	...	83	not A+

→ Classification algorithm →

ml_ experience	...	quiz1	quiz2
1	...	86	?

Predict a student's *score* in quiz2

Regression model

ml_ experience	...	quiz1	quiz2
0	...	90	91
1	...	91	92
...	...	...	...
0	...	83	80

→ Regression algorithm →

ml_ experience	...	quiz1	quiz2
1	...	86	?

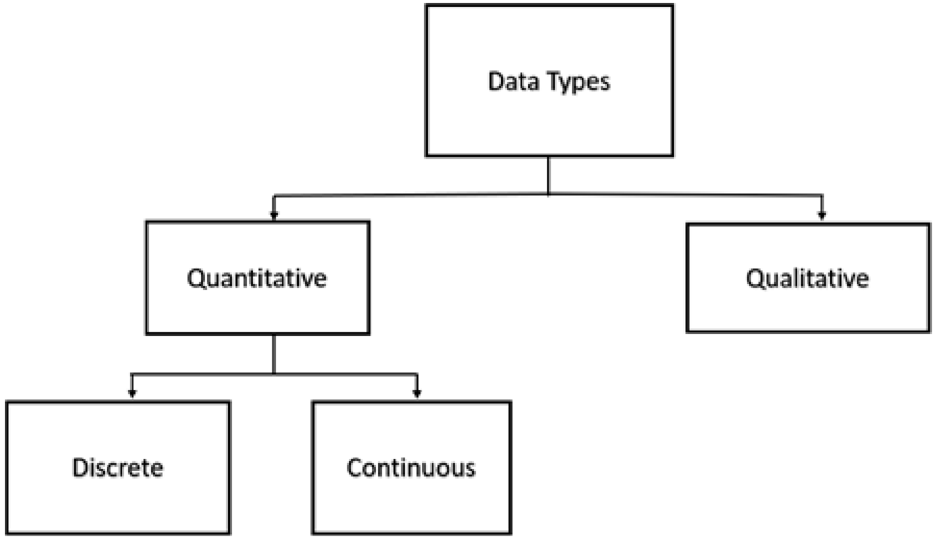
Lecture Contents

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

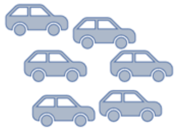
- Introduction
- Terminology
- Setting up Jupitar notebook
- Data !!

Different Types of data types

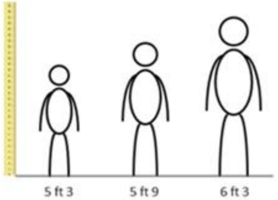
- Quantitative data type:
  - This type of data type consists of numerical values. Anything which is measured by numbers.
  - E.g., Profit, quantity sold, height, weight, temperature, etc.



No. of Laptops



No. of Cars

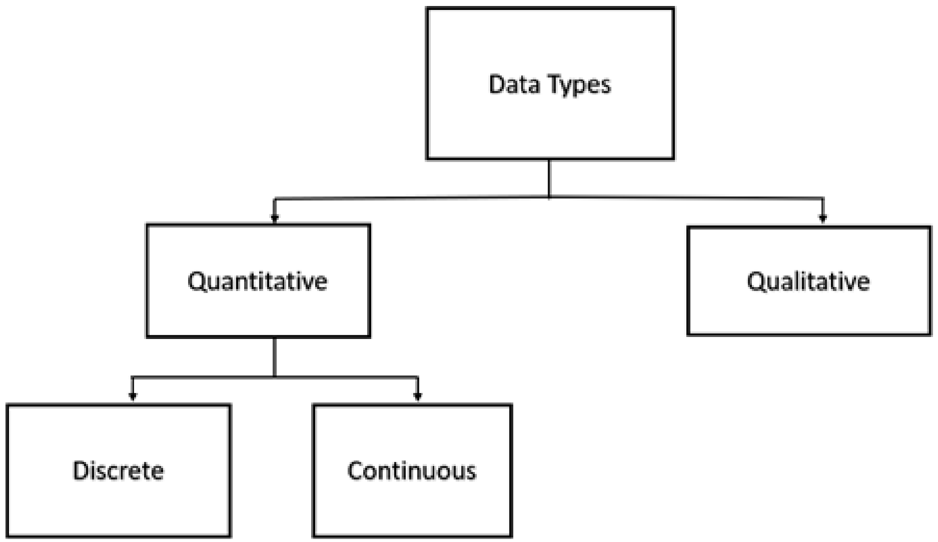


Height

Time

Different Types of data types

- Qualitative data type:
  - These are the data types that cannot be expressed in numbers. This describes categories or groups and is hence known as the categorical data type.



Structured Data

- This type of data is either number or words. This can take numerical values, but mathematical operations cannot be performed on it. This type of data is expressed in tabular format.
- E.g.) Sunny=1, cloudy=2, windy=3 or binary form data like 0 or1, Good or bad, etc.

ID	Name	Age	Degree
1	John	18	B.Sc.
2	David	31	Ph.D.
3	Robert	51	Ph.D.
4	Rick	26	M.Sc.
5	Michael	19	B.Sc.

Unstructured Data

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

- This type of data does not have the proper format and therefore known as unstructured data.
- This comprises textual data, sounds, images, videos, etc.

Other Data Types

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

- Nominal Data Type: This is in use to express names or labels which are not order or measurable.
- E.g., male or female (gender), race, country, etc.

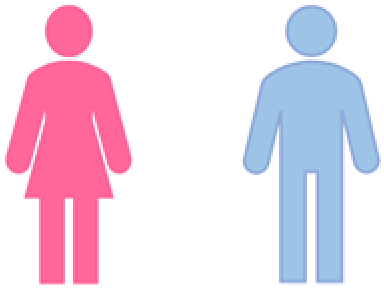
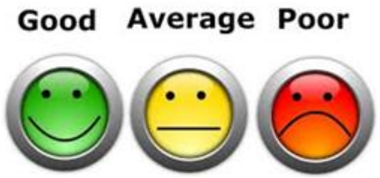


Fig: Gender (Female, Male), An Example Of Nominal Data Type

- Ordinal Data Type: This is also a categorical data type like nominal data but has some natural ordering associated with it.
- E.g., Likert rating scale, Shirt sizes, Ranks, Grades, etc.



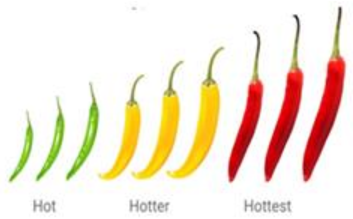


Fig: Rating (Good, Average, Poor), An Example Of Ordinal Data Type

Other Data Types

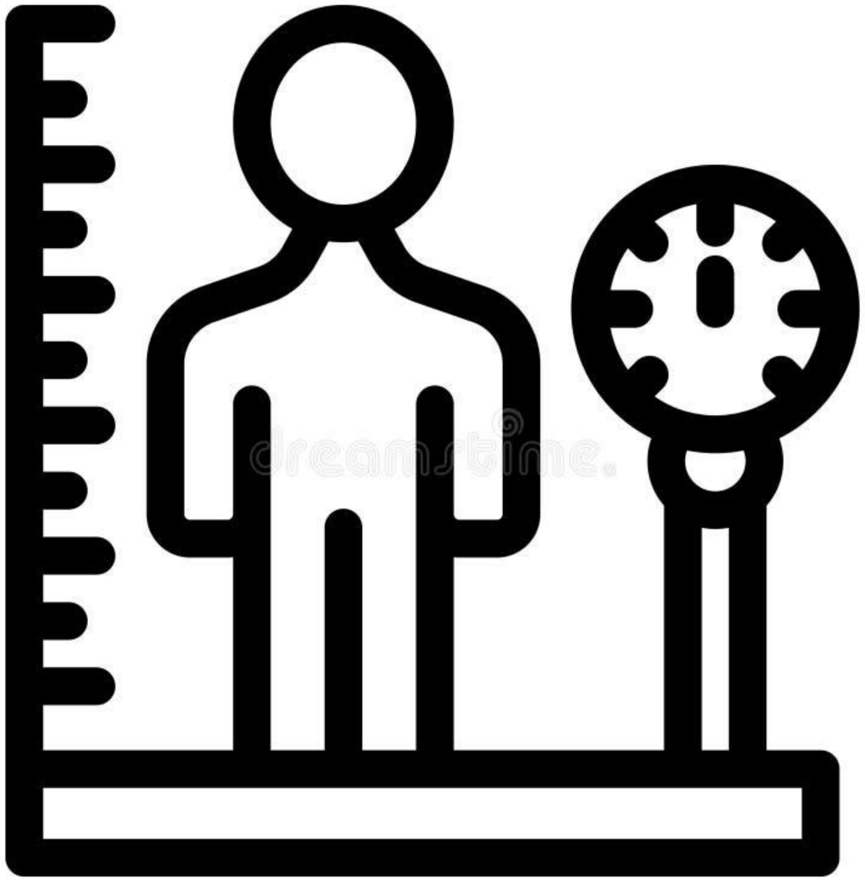
- Interval Data Type: This is numeric data which has proper order and the exact zero means the true absence of a value attached.
- Here zero means not a complete absence but has some value. This is the local scale.
- E.g., Temperature measured in degree Celsius, time, Sat score, credit score, pH, etc.

Fig: Temperature, An Example Of Interval Data Type

Other Data Types

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

- Ratio Data Type: This quantitative data type is the same as the interval data type but has the absolute zero.
- Here zero means complete absence, and the scale starts from zero. This is the global scale.
- E.g., height, weight, etc.



Converting to Numerical Features

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

- Often want a real-valued example representation:

Age	City	Income
23	Van	22,000.00
23	Bur	21,000.00
22	Van	0.00
25	Sur	57,000.00
19	Bur	13,500.00
22	Van	20,000.00

$\rightarrow$

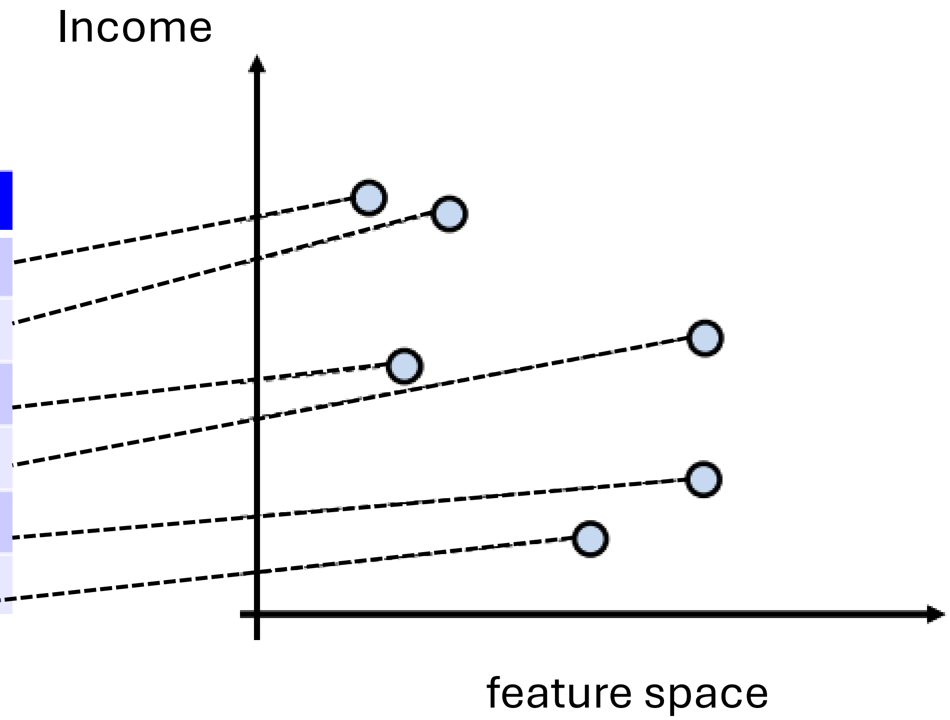
Age	Van	Bur	Sur	Income
23	1	0	0	22,000.00
23	0	1	0	21,000.00
22	1	0	0	0.00
25	0	0	1	57,000.00
19	0	1	0	13,500.00
22	1	0	0	20,000.00

- This is called a “1 of k” encoding (or “one hot” encoding).
- We can now interpret examples as points in space:
  - E.g., first example is at (23,1,0,0,22000).

Data Space

- You can compute a "distance" between examples in feature space. – "Are these examples close to each other?"

Age	Van	Bur	Sur	Income
23	1	0	0	22,000.00
23	0	1	0	21,000.00
22	1	0	0	0.00
25	0	0	1	57,000.00
19	0	1	0	13,500.00
22	1	0	0	20,000.00



Approximating Text with Numerical Features

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

- Bag of words replaces document by word counts:

The **International Conference on Machine Learning** (ICML) is the leading international academic conference in machine learning

ICML	International	Conference	Machine	Learning	Leading	Academic
1	2	2	2	2	1	1

- You can compute a “distance” between documents.
  - To find similar documents or decide if two documents are similar.

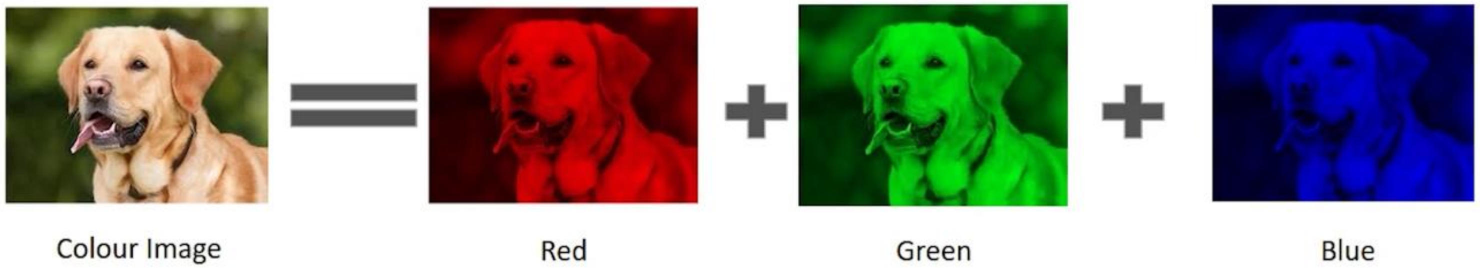
Image Data

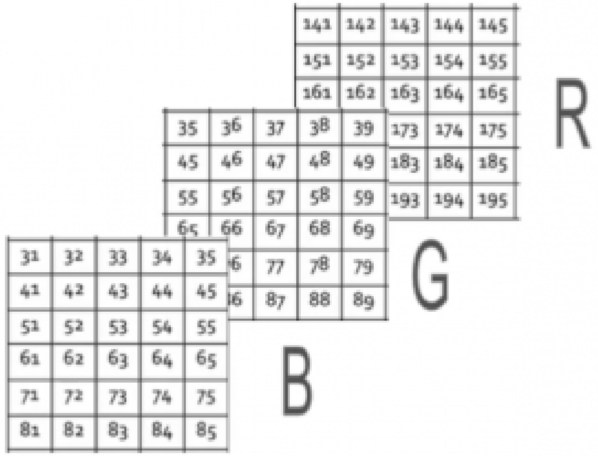
- Images are stored in a computer as a matrix of numbers known as pixel values.
- These pixel values represent the intensity of each pixel.
- In grayscale images, a pixel value of 0 represents black, and 255 represents white.

Image Data

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

- This image comprises many different colors. Almost all colors can be generated from the three primary colors – Red, Green, and Blue. Therefore, we can say that each colored image is a unique composition of these three colors or 3 channels – Red, Green, and Blue.





Data Cleaning

- Ways that data might not be ‘clean’:
  - Noise (e.g., distortion on phone).
  - Outliers (e.g., data entry or instrument error).
  - Missing values (no value available or not applicable)
- Duplicated data (repetitions, or different storage formats).
  - Any of these can lead to problems in analyses.
  - Some ML methods are robust to these.
  - Often, ML is the best way to detect/fix these.

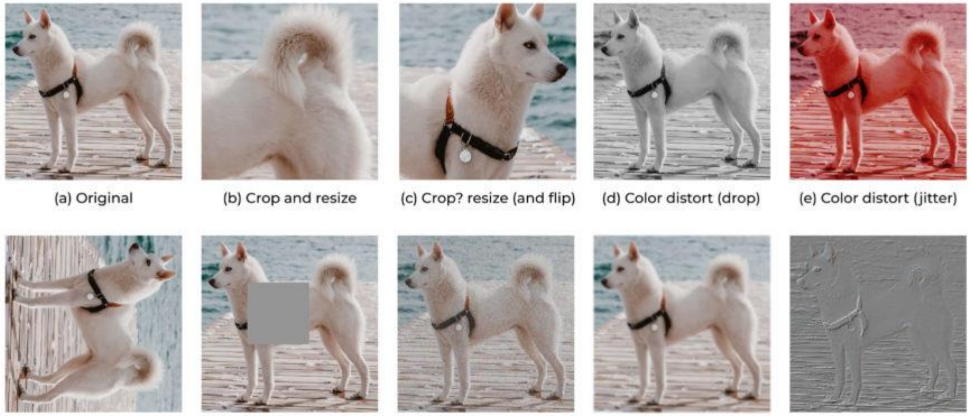
How much data do we need?

- It Depends !!
- The complexity of a model
- The complexity of the learning algorithm
- Labeling needs
- Acceptable error margin
- Input diversity

How to deal with lack of data

McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

Data augmentation is a process of expanding an input dataset by slightly changing the existing (original) examples. It's widely used for image segmentation and classification. Typical image alteration techniques include cropping, rotation, zooming, flipping, and color modifications.



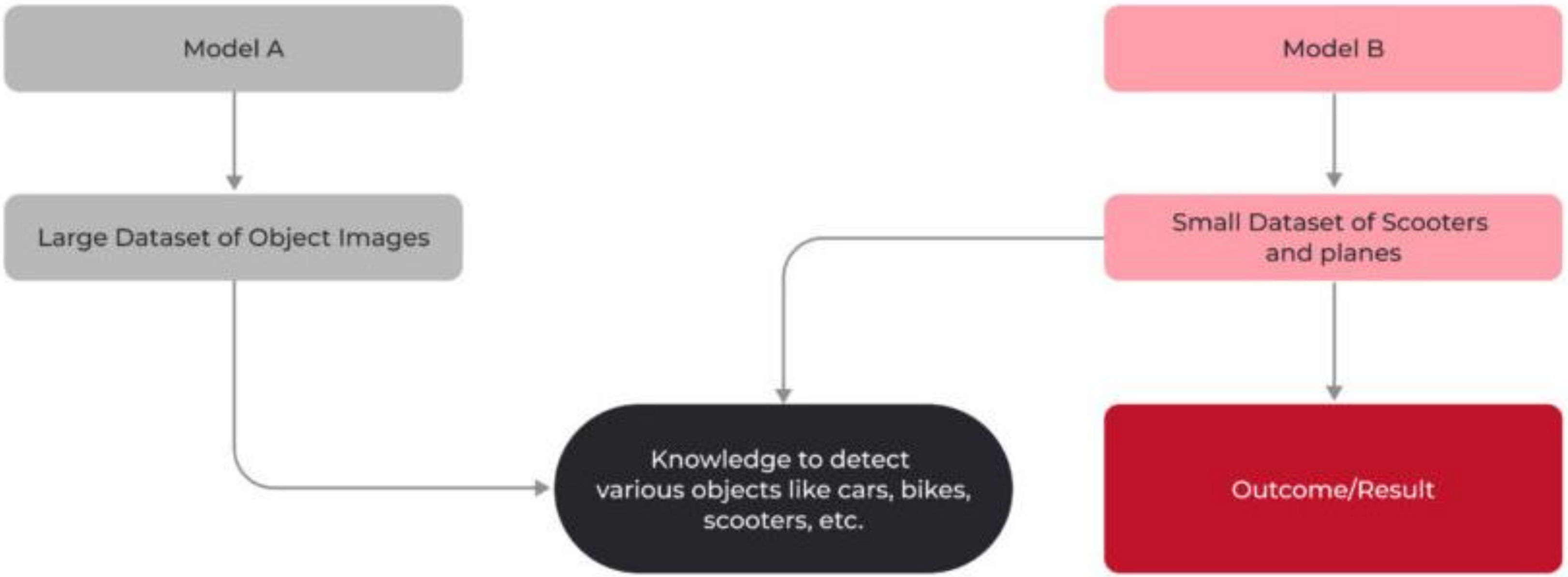
(a) Original (b) Crop and resize (c) Crop? resize (and flip) (d) Color distort (drop) (e) Color distort (jitter)

(f) Rotate {90°, 180°, 270°} (g) Cutout (h) Gaussian noise (i) Gaussian blur (j) Sobel filtering

How to deal with lack of data

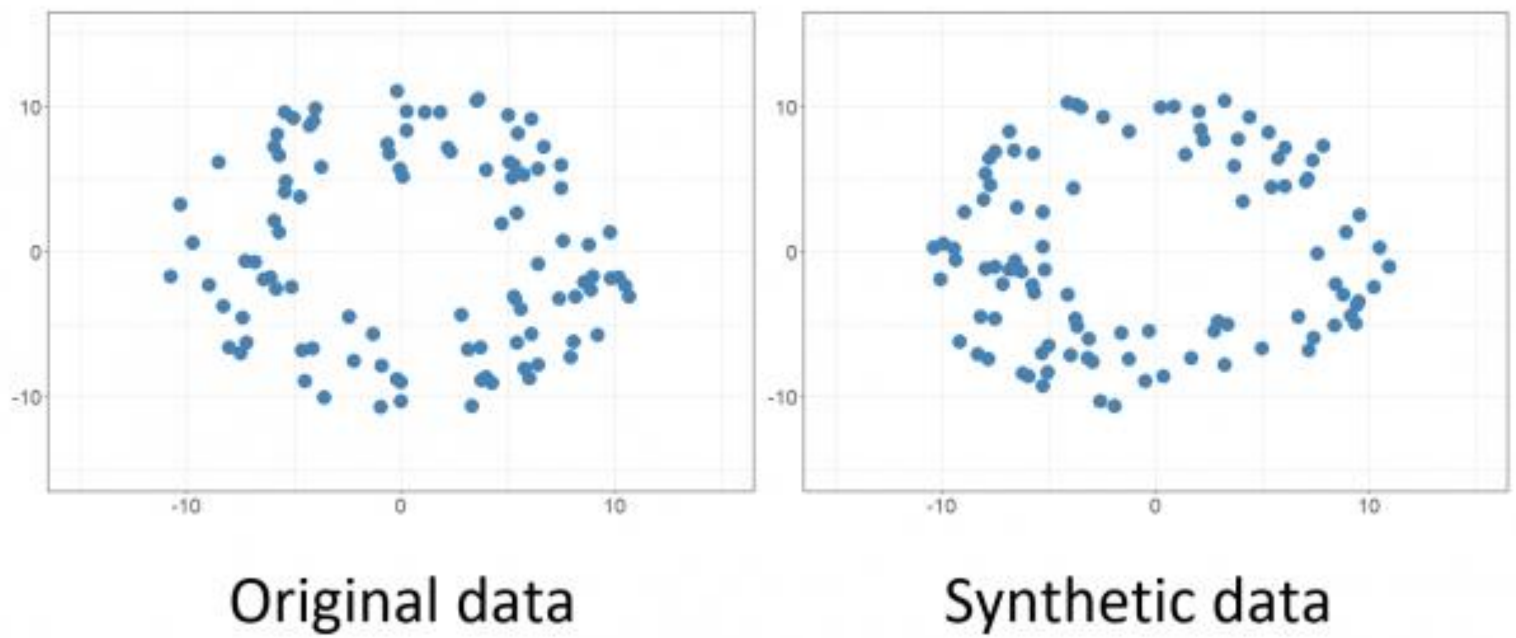
McMaster University
ENGINEERING
W Booth School of Engineering Practice and Technology

Transfer learning is another technique of solving the problem of limited data. This method is based on applying the knowledge gained when working on one task to a new similar task.



How to deal with lack of data

Synthetic data is artificially generated to mimic the characteristics and structure of sensitive real-world data, but without exposing our sensitivities.



The synthetic data retains the structure of the original data but is not the same

Feature Aggregation

- Feature aggregation: – Combine features to form new features:

Van	Bur	Sur	Edm	Cal
1	0	0	0	0
0	1	0	0	0
1	0	0	0	0
0	0	0	1	0
0	0	0	0	1
0	0	1	0	0

BC	AB
1	0
1	0
1	0
0	1
0	1
1	0

- Fewer province “coupons” to collect than city “coupons”.

Feature Transformation

- Mathematical transformations:
  - Discretization (binning): turn numerical data into categorical.

Age
23
23
22
25
19
22

< 20	>= 20, < 25	>= 25
0	1	0
0	1	0
0	1	0
0	0	1
1	0	0
0	1	0

- Only need to collect 3 coupons. – We will see many more transformations (addressing other problems).

Feature Selection

- Remove features that are not relevant to the task.

SID:	Age	Job?	City	Rating	Income
3457	23	Yes	Van	A	22,000.00
1247	23	Yes	Bur	BBB	21,000.00
6421	22	No	Van	CC	0.00
1235	25	Yes	Sur	AAA	57,000.00
8976	19	No	Bur	BB	13,500.00
2345	22	Yes	Van	A	20,000.00

- Student ID is probably not relevant (do not need to collect these coupons).

Course References

- Based on the materials from
  - Lectures prepared from Dr. Jeff Fortuna Lecture notes and slides (SEP 785 Fall2024)
  - UBC CPSC 330 prepared by Dr. Varada Kolhatkar
  - UBC CPS 340 prepared by Dr. Mark Schmidt
  - UofT CSC 2515   prepared by Dr. David Duvenaud

- Abu-Mostafa, Yaser S., Malik Magdon-Ismail, and Hsuan-Tien Lin. Learning from data. Vol. 4. New York: AMLBook, 2012.
- Kuhn, M. "Applied predictive modeling." (2013).

McMaster University ENGINEERING

W BOOTH SCHOOL OF ENGINEERING PRACTICE AND TECHNOLOGY

SEP 785: Machine Learning

Lecture 1: Introduction

Thank you !!

Questions ???