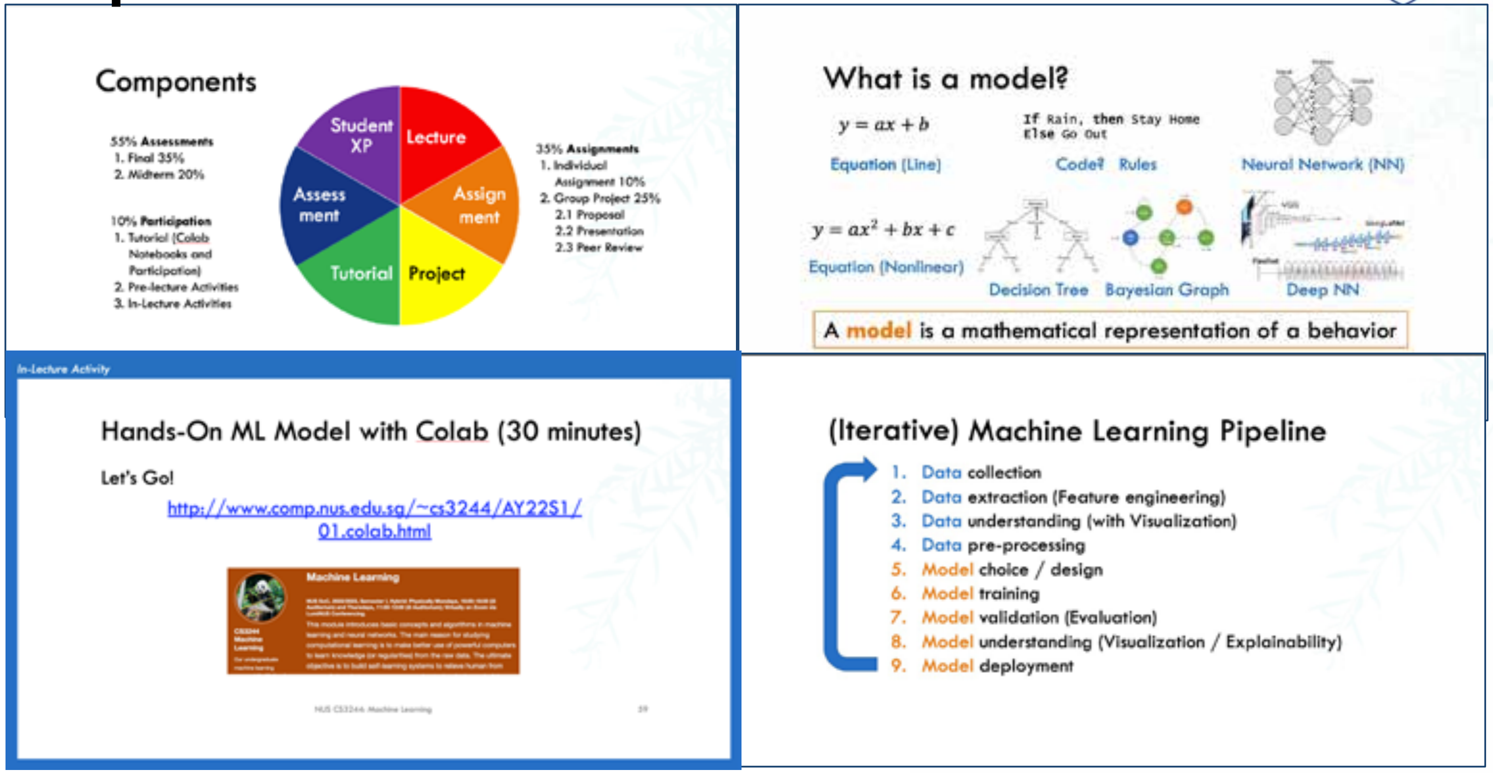
Paradigms of ML and kNN

2

CS 3244
Machine Learning

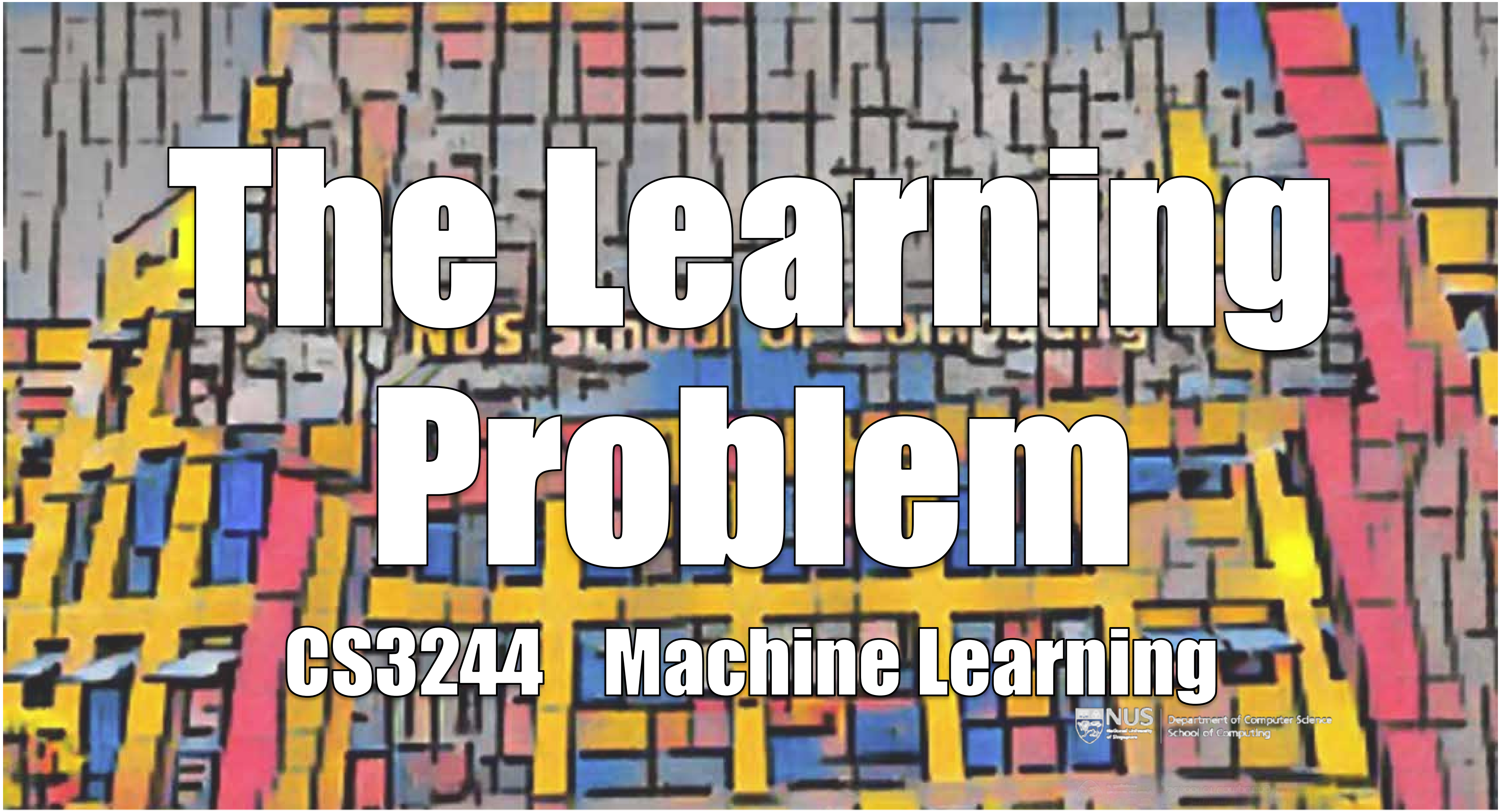
NUS National University of Singapore | Computing

Recap from Week 01



Forecast for Week 02

- Understand how we can define the learning problem and when it is appropriate to apply
- Dissect the components of a machine learning model
- Relate tribes of ML as means of picking a learning algorithm and hypothesis space.
- Understand an instance of the analogizer tribe: k nearest neighbors, and how its hyperparameters affect its decision making.


The Learning
Problem
CS3244 Machine Learning
NUS
Department of Computer Science
School of Computing

So, what it is anyways?

How can we define Machine Learning?

Maybe through examples?

Let’s give three.

#1. What is a tree?

#1. A tree is…

Hard to give a complete mathematical definition

But even a 3-year-old can distinguish trees from non-trees

We'll say that the 3-year-old has learned from data

Photo credits: [Wikipedia]

1. A pattern exists
2. Difficult to pin down formally (mathematically)
3. We have data for it

By Erwin Soo from Singapore. (the trees..)

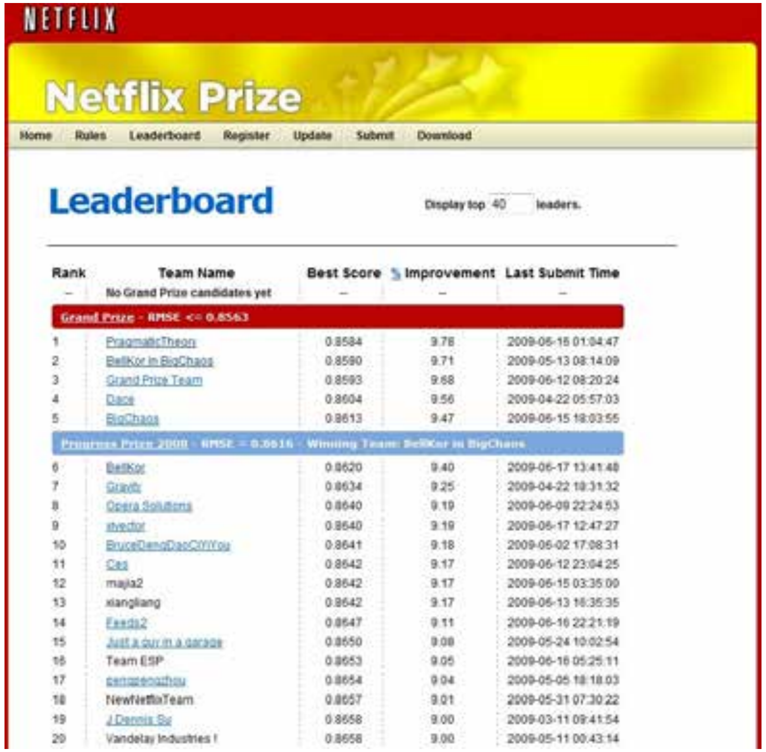
In the #lectures threads, answer or upvote

Q1: Which one of the three criteria is absolutely essential to have for ML to work? Pick one.

{ 1. *A pattern exists*, 2. *Difficult to pin down formally*, 3. *We have data* }

Q2: What happens when the essential criterion is missing?

#2. The Netflix prize*


NETFLIX

Netflix Prize

Home Rules Leaderboard Register Update Submit Download

Leaderboard

Display top 40 leaders.

Rank	Team Name	Best Score	Improvement	Last Submit Time
–	No Grand Prize candidates yet	–	–	–
Grand Prize - RMSE <= 0.8563				
1	PragmaticTheory	0.8584	9.76	2009-06-16 01:04:47
2	BellKor in BigChaos	0.8590	9.71	2009-05-13 08:14:09
3	Grand Prize Team	0.8593	9.68	2009-06-12 08:20:24
4	Dace	0.8604	9.56	2009-04-22 05:57:03
5	BigChaos	0.8613	9.47	2009-06-15 18:03:55
Progress Prize 2008 - RMSE = 0.8616 - Winning Team: BellKor in BigChaos				
6	BellKor	0.8620	9.40	2009-06-17 13:41:48
7	Gravity	0.8634	9.25	2009-04-22 18:31:32
8	Opera Solutions	0.8640	9.19	2009-06-09 22:24:53
9	xlvector	0.8640	9.19	2009-06-17 12:47:27
10	BruceDengDaoCiYiYou	0.8641	9.18	2009-06-02 17:08:31
11	Cas	0.8642	9.17	2009-06-12 23:04:25
12	majia2	0.8642	9.17	2009-06-15 03:35:00
13	xiangliang	0.8642	9.17	2009-06-13 16:35:35
14	Feeds2	0.8647	9.11	2009-06-16 22:21:19
15	Just a guy in a garage	0.8650	9.08	2009-05-24 10:02:54
16	Team ESP	0.8653	9.05	2009-06-16 05:25:11
17	pengpengzhou	0.8654	9.04	2009-05-05 18:18:03
18	NewNetflixTeam	0.8657	9.01	2009-05-31 07:30:22
19	J Dennis Su	0.8658	9.00	2009-03-11 09:41:54
20	Vandelay Industries !	0.8658	9.00	2009-05-11 00:43:14

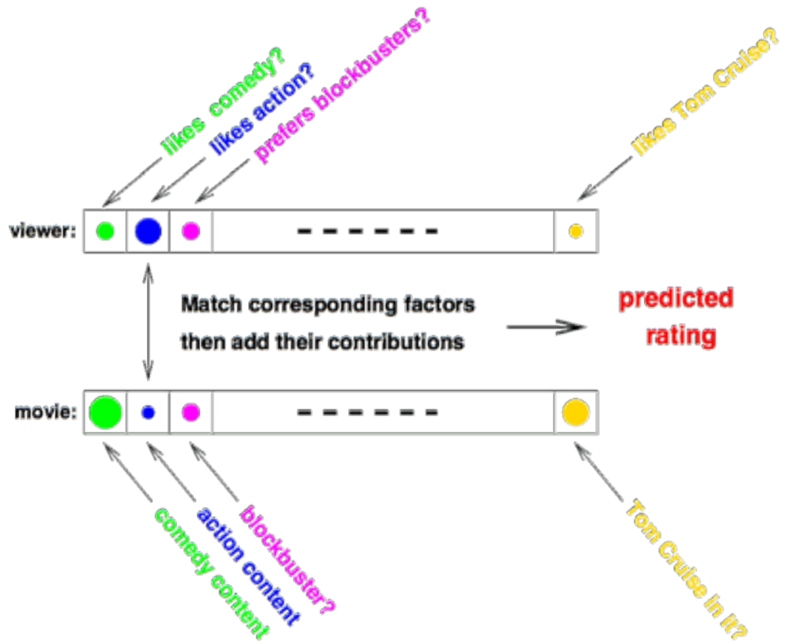
Predicting how a viewer will rate a movie

10% improvement = win a USD $ 1 M prize

* (BTW, Netflix never used the algorithm due to engineering costs.)

Photo credits: Wired magazine blog

Previous ratings predict future ratings



Viewer taste and movie content imply viewer ratings
⇒ A pattern exists

No magic formula to predict ratings
⇒ Difficult to pin down formally (mathematically)

Netflix has data. We can learn to identify movie genres and viewer preferences
⇒ We have data for it

#3. Credit Approval

Would you approve a credit line for this person?

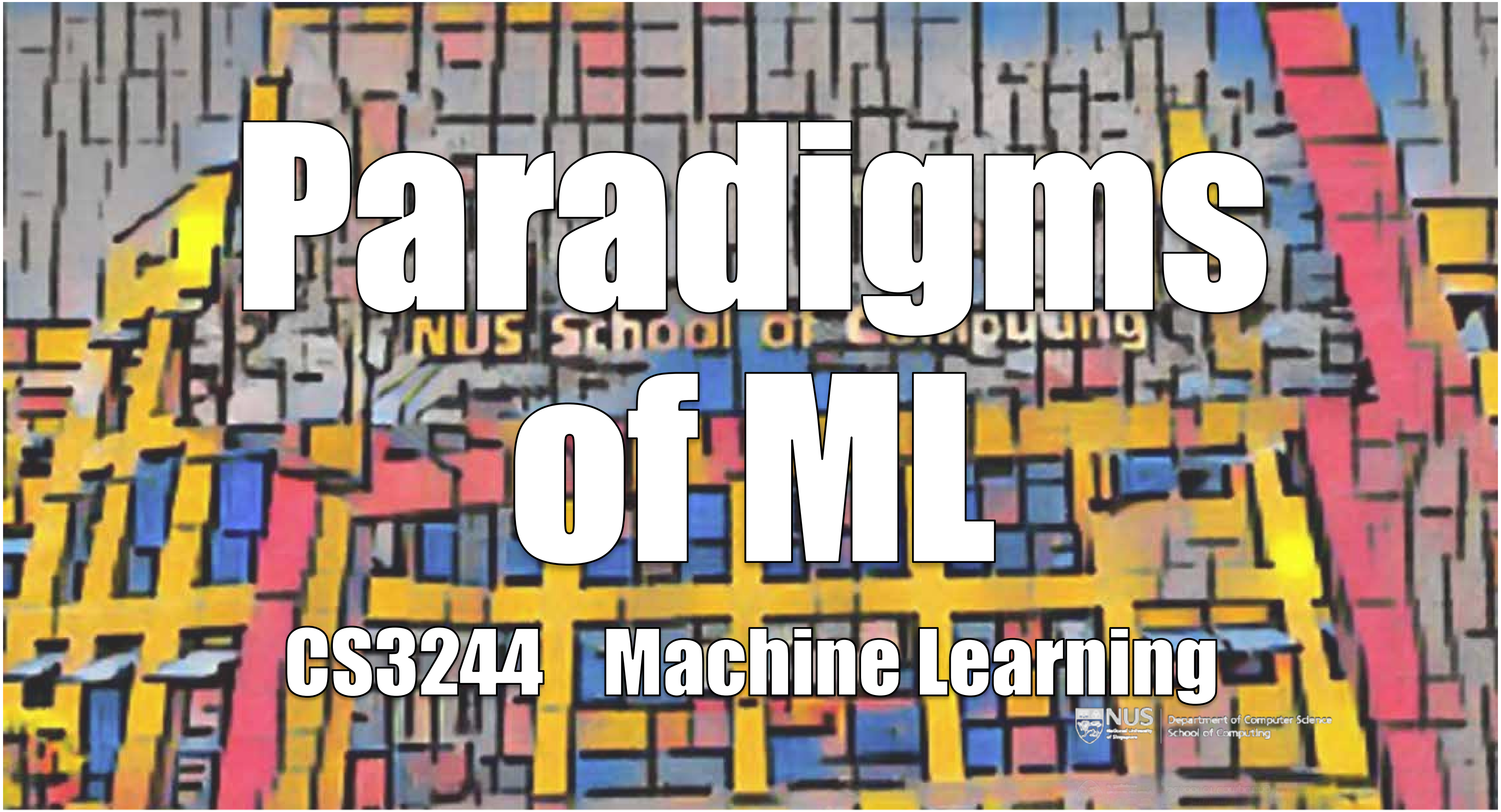
Criterion	Value
Age	32 years
Gender	Male
Salary	40 K
Debt	26 K
…	…
Years in Job	1 year
Years at Current Residence	3 years

Using criteria to approve or not
⇒ A pattern exists

No magic credit approval formula
⇒ Difficult to pin down formally (mathematically)

Banks have lots of data, on customers and whether they defaulted or not
⇒ We have data for it

Artwork credits: Project DebtRelief


Paradigms
of ML
CS3244 Machine Learning
NUS
Department of Computer Science
School of Computing

Basic premise of learning

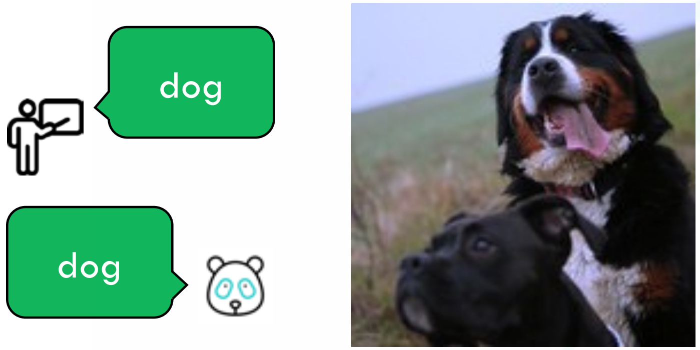
"using a set of observations to uncover an underlying pattern"

Three general variations of the learning problem:

- Supervised Learning [W01–04, 09, 10]
- Unsupervised Learning [W12]
- Reinforcement Learning [not included]

1. Supervised Learning

Each input instance comes with a corresponding *answer* (a.k.a. *ground truth* or *label*).



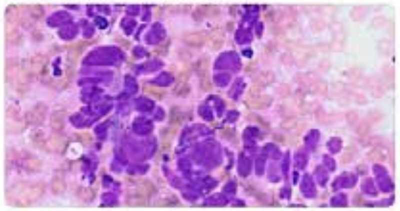
Pre-Lecture Activity from last week

ML application: predict shipments' time of delivery in Amazon
Link: https://aws.amazon.com/blogs/industries/how-to-predict-shipments-time-of-delivery-with-cloud-based-machine-learning-models/
Paradigm: Supervised learning. Training was conducted with 6.3 million rows of data.
Tribe: Symbolists. The transit time model uses the XGBoost algorithm, which is an implementation of gradient boosted trees algorithm, to attempt to accurately predict the transit time.

> **Amazon Web Services**
> How to Predict Shipments' Time of Delivery with Cloud-based Machine Learning Models | Amazon Web Services
> Transportation and logistics businesses have seen an upsurge in trade over the last few years, with increasing ecommerce and online sales driving demand. This welcome growth has, however, been accompanied with an ever-more demanding consumer looking for certainty and reliability in their service delivery alongside a more competitive landscape, forcing prices down with tightening margins. [...]
> Mar 24th, 2021

ML Application: Use of ML for medical diagnosis (https://www.neuraldesigner.com/learning/examples/breast-cancer-diagnosis#TrainingStrategy)
Paradigm: Supervised Learning. Data set provided had both input and output.
Tribe: Connectionists. A Neural Network was used, containing a Scaling, Perceptron & Probabilistic layer.

> **neuraldesigner.com**
> Breast cancer diagnosis machine learning example
> This example aims to assess whether a lump in a breast could be malignant (cancerous) or benign (non-cancerous) from digitized images of a fine-needle aspiration biopsy. (213 kB)



Gmail Smart Compose
https://ai.googleblog.com/2018/05/smart-compose-using-neural-networks-to.html
It offers sentence completion suggestions as we type an email.

- **Paradigm:** Unsupervised Learning as there are no pre-labelled completion suggestions but they are determined from patterns in training data.
- **Tribe:** Connectionists as it employs the RNN-LM model for suggestions, which uses neural networks to learn to predict the next word.

> **Google AI Blog**
> Smart Compose: Using Neural Networks to Help Write Emails
> Posted by Yonghui Wu, Principal Engineer, Google Brain Team Last week at Google I/O , we introduced Smart Compose , a new feature in Gmail...
> Google AI

Components of Learning

Formalization

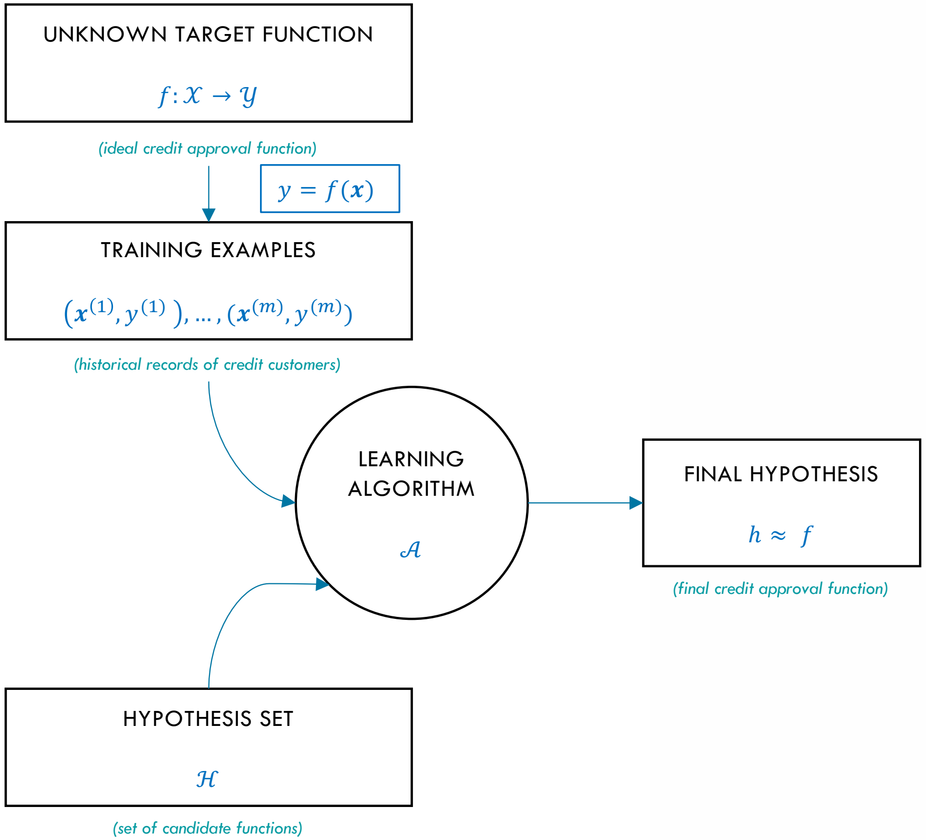
- Input: $\boldsymbol{x} = \{x_1, x_2, \dots, x_n\}$ *(customer application)*
- Output: y *(profitable customer?)*
- Target Function: $f : \mathcal{X} \to \mathcal{Y}$ *(ideal credit approval formula)*
- Data: $D = (\boldsymbol{x}^{(1)}, y^{(1)}), (\boldsymbol{x}^{(2)}, y^{(2)}), \dots, (\boldsymbol{x}^{(m)}, y^{(m)})$ *(historical records)*

- Hypothesis: $h: \mathcal{X} \to \mathcal{Y}$ *(formula to be used)*

$\mathcal{X}$, $\mathcal{Y}$ and D are given by the learning problem;
The target f is fixed but unknown.

We learn the function h from the data D.

🤞 Data Matrix Indices
- Index n: # of features (cols)
- Index m: # of instances (rows)

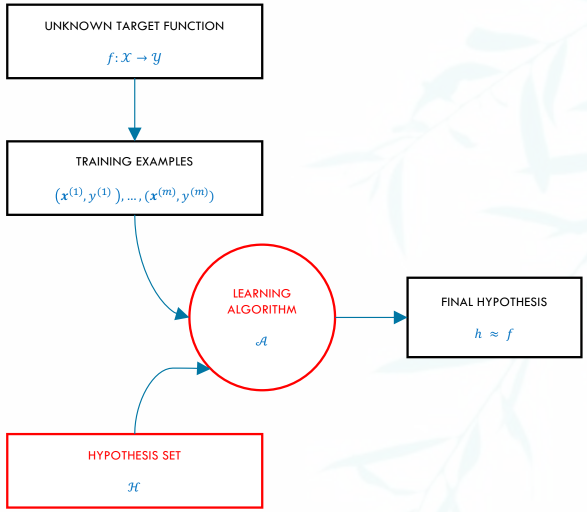

UNKNOWN TARGET FUNCTION
$f: \mathcal{X} \to \mathcal{Y}$
(ideal credit approval function)
$y = f(\boldsymbol{x})$
TRAINING EXAMPLES
$(\boldsymbol{x}^{(1)}, y^{(1)}), \ldots, (\boldsymbol{x}^{(m)}, y^{(m)})$
(historical records of credit customers)
LEARNING ALGORITHM
$\mathcal{A}$
FINAL HYPOTHESIS
$h \approx f$
(final credit approval function)
HYPOTHESIS SET
$\mathcal{H}$
(set of candidate functions)

Solution Components

There are 2 components of the learning algorithm:

1. The *Hypothesis set* $\mathcal{H} = \{h_1, h_2, \dots, h_{|\mathcal{H}|}\}$

2. The *Learning algorithm* $\mathcal{A}$ selects $h \in \mathcal{H}$



Together they are referred to as the *learning model*.
We choose $\mathcal{H}$ and the learning algorithm $\mathcal{A}$ chooses h.

Visualizing the input

UNKNOWN TARGET FUNCTION
$f: \mathcal{X} \rightarrow \mathcal{Y}$

TRAINING EXAMPLES
$(\boldsymbol{x}^{(1)}, y^{(1)}), \ldots, (\boldsymbol{x}^{(m)}, y^{(m)})$

Let's zoom in on the inputs to learning:

X and Y could be single, continuous values, which is then easy to plot:

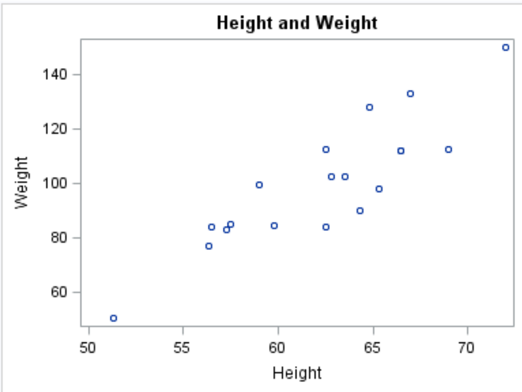


Image Courtesy support.sas.com

But X could also be a vector, having many dimensions. Y could as well. But often this generalizes, so we can show simple plots of X being 2D and Y being 1D.

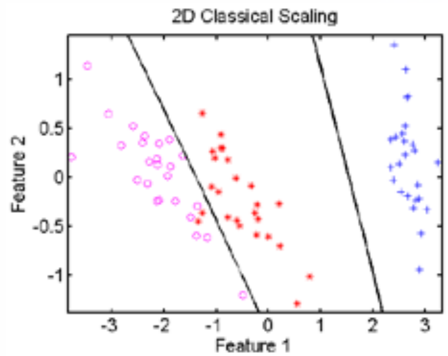


Image Courtesy: 37steps.com

The Data Matrix

We can visualize the training examples and labels thus as a data matrix, where each example vector (transposed) is stacked on top of each other.

Fisher's *Iris* Data

Sepal length	Sepal width	Petal length	Petal width	Species
4.9	3.0 $x^{(0)}$	1.4	0.2	I. setosa $y^{(0)}$
4.7	3.2 $x^{(1)}$	1.3	0.2	I. setosa $y^{(1)}$
4.6	3.1	1.5	0.2	*I. setosa*
5.0	3.6	1.4	0.3	*I. setosa*
5.4	3.9	1.7	0.4	*I. setosa*
4.6	3.4	1.4	0.3	*I. setosa*
5.0	3.4	1.5	0.2	*I. setosa*
4.4	2.9	1.4	0.2	*I. setosa*
4.9	3.1	1.5	0.1	*I. setosa*
6.2	3.4	5.4	2.3	*I. virginica*
5.9	3.0	5.1	1.8	*I. virginica*

Iris versicolor

Iris virginica

Images and tables courtesy: Wikipedia

Differentiating estimates from ground truth

Unknown Target Function $f: \mathcal{X} \rightarrow \mathcal{Y}$

Hypothesized Target Function $h \approx f$

Your Turn: How do we denote what $h(x)$ outputs?

$$h(x) = ?$$

Differentiating estimates from ground truth

Unknown Target Function $f: \mathcal{X} \rightarrow \mathcal{Y}$

Hypothesized Target Function $h \approx f$

Your Turn: How do we denote what $h(x)$ outputs?

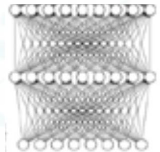
Stats uses the hat operator to denote estimates: i.e., $h(x) = \hat{y}$.
You'll also encounter notation where $\hat{f}$ stands for h.

Also, we'll use the $\mathscr{script}$ font to denote "the space of all possible": e.g., $\mathcal{X}$ means all possible input values.

A classifier can have parameters Θ

$h = ax_1 + b$

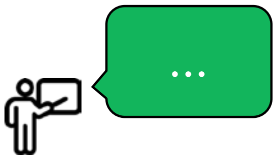
$h = \Theta(\Theta\left(\ldots\left(\Theta(\boldsymbol{x})\right)\right)$



And *hyperparameters*: parameters that influence the learning process.

Photo by CHUTTERSNAP on Unsplash

2. Unsupervised Learning



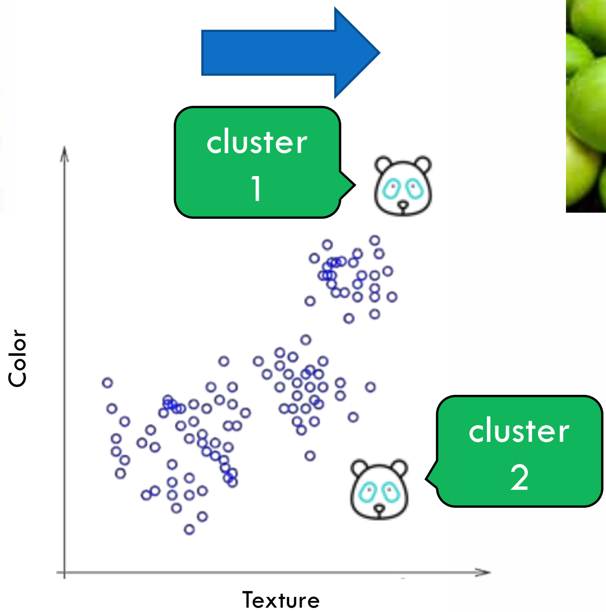


Photo credits: The Globe and Mail, annaterrenal @ DeviantArt,

Link: https://www.geeksforgeeks.org/how-does-netflix-use-machine-learning/
Paradigm: Netflix uses unsupervised learning. It uses data such as your history in movies/TV shows watched and recommends movies and TV shows that are watched by the cluster which consists of people with similar interests in shows watch.
Tribe: Analogizers. The problem of content recommendation based on people that watched and enjoyed identical movies and TV shows is similar to that of the K nearest neighbours algorithm that is associated with Analogizers tribe of machine learning.

GeeksforGeeks

How Does Netflix Use Machine Learning? - GeeksforGeeks

A Computer Science portal for geeks. It contains well written, well thought and well explained computer science and programming articles, quizzes and practice/competitive programming/company interview Questions.

Dec 20th, 2020

Other Applications of Unsupervised Learning

Density Estimation:

- dataset assumed to be representative of a larger population
- goal is to estimate the probability of data

Outlier detection:

- uncharacteristic data
- noise filtering

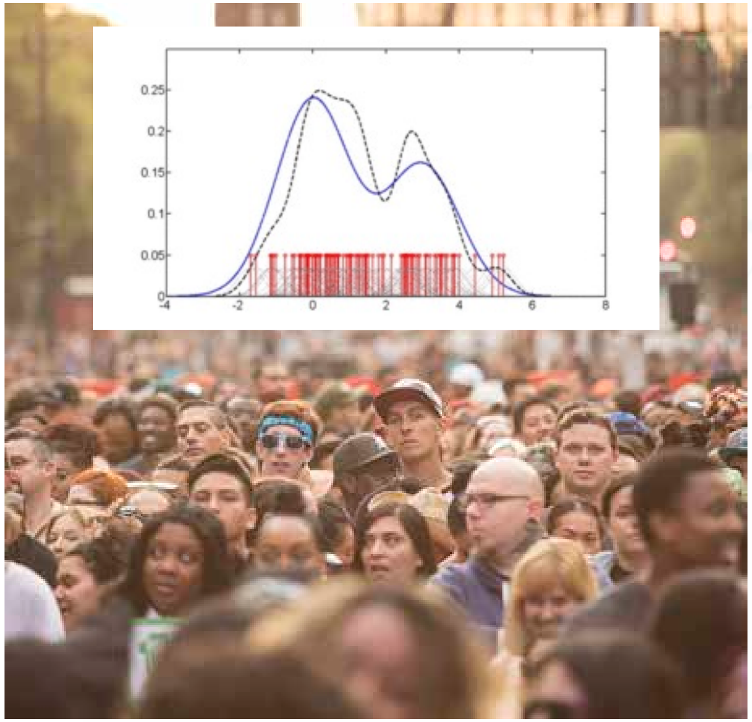


Photo by Stephen Leonardi on Unsplash , GIF from Wikipedia

3. Reinforcement Learning

Instead of (*input, correct output*),
we get guidance on our behaviors.

Photo credits: ibtimes.com

Characteristics of Reinforcement Learning

What makes reinforcement learning different from other ML paradigms?

- There is no supervisor, only a reward signal
- Feedback is delayed, not instantaneous
- Time matters (sequential and non *i.i.d.* data)
- Agent's actions affect the subsequent data it receives

Screen capture from [OpenAI.com YouTube video]

Examples:

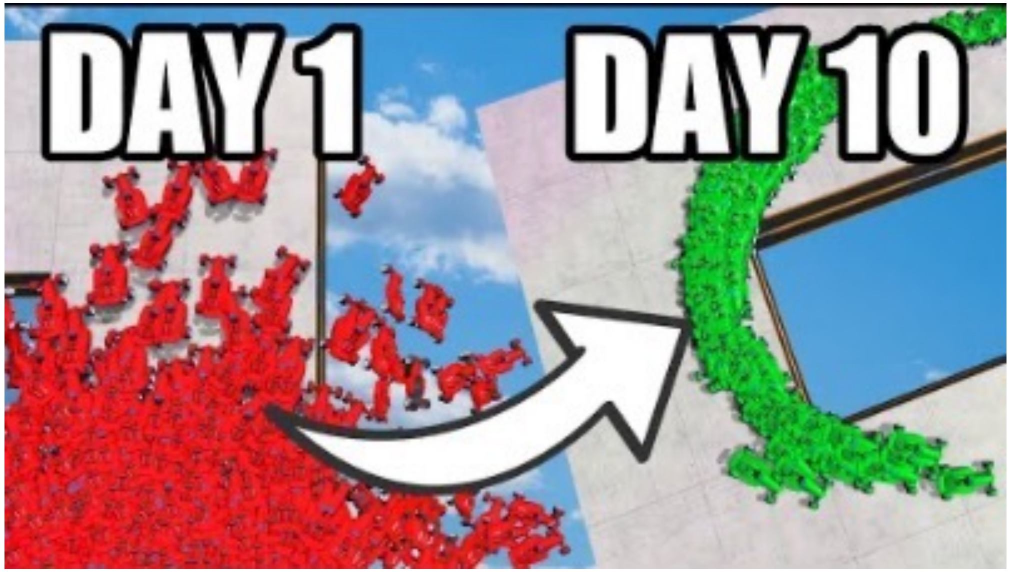
- Guide a helicopter to fly stunts
- Defeat humans at MMORPG
- Manage an investment portfolios
- Make robots walk or do stunts
- Stylising a photograph

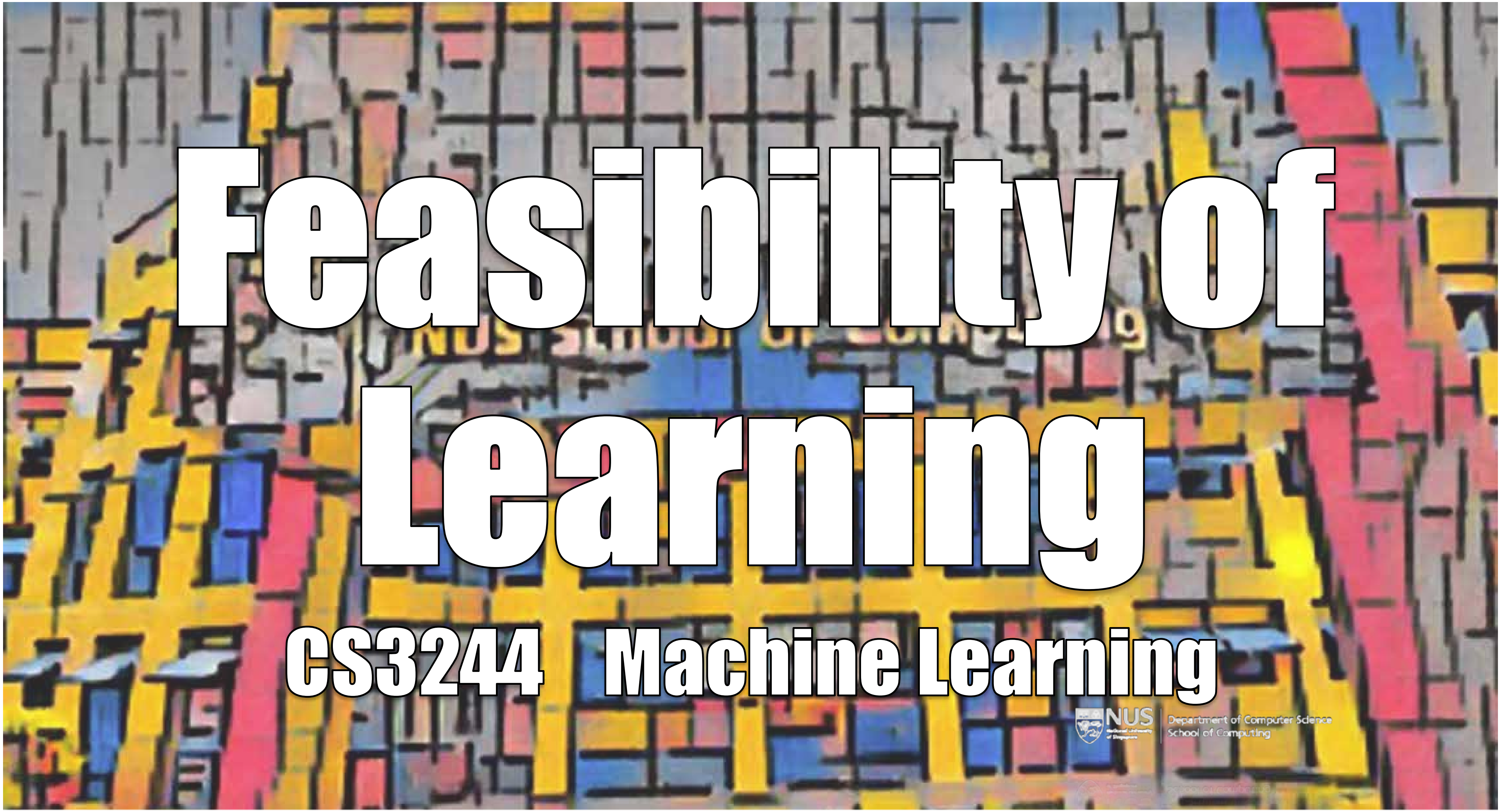
Pre-Lecture Activity from last week

In the article by Parikh, he mentioned that one of the main problems solved by reinforcement learning is game playing. An example of this can be seen in this video - AI Learns to Drive from Scratch in Trackmania.

I think it belongs to the reinforcement learning paradigm. In the video, it is mentioned that the AI learns through the use of a reward system with no supervision. In the case of the video, the AI is rewarded based on the distance/time travelled on the track and other factors like punishment for going off the track.

I think that the tribe that this algorithm belongs to is the symbolists. In the video, it is described that the neural networks' role is to predict the amount of reward associated with the various actions when given a set of inputs and thereafter decide on an action to take.





Feasibility of Learning

CS3244 Machine Learning

NUS
Department of Computer Science
School of Computing

Lim Kopi?

lɪm kopi: [leem-koh-pee] (*adjective*) Chill-out, and to relax. Traditionally means to ‘drink coffee’, but has evolved to become synonymous with asking someone out for a drink, or a chat.

Example: “It’s stifling here in the office on this Monday morning, let’s lim kopi at the café and discuss further?” From Hokkien/Malay.

形容词： 放松、悠闲，休息。 原译为”喝咖啡“，但之后演变为约人出去聊天。

口语例句： （源自福建话、马来语） “星期一的办公室实在是令人窒息，不如，我们一块儿去外面lim kopi, 继续讨论我们的议题？ ”

adjektif. Lepak, berehat. Maksud asalnya ialah ‘minum kopi’, tetapi telah berubah untuk bermaksud mengajak sesorang untuk keluar dan minum bersama-sama.

Contoh bahasa percakapan: “Saya berasa rimas duduk di pejabat pada pagi Isnin, mari lim kopi dan teruskan perbincangan?” Daripada bahasa Hokkien atau Melayu.

லிம் கோப்பி • வினைச்சொல்: சிலுசிலுப்பு, ஓய்வெடுக்க, பொதுவாக கோபி அருந்துதல் என பொருள், ஆனால் நாளடைவில் அது ஒருவரை கோபி அருந்த அழைப்பு விடுவதுபோல் அமைந்து விட்டது.

பேச்சு வழக்கு உதாரணம்: இன்றைய திங்கள் காலை ரொம்ப கடினமாக இருக்கிறது, ஒரு “லிம் கோப்பி” அடித்துகொண்டு உரையாடுவோமா? ஹோக்கியேன் மொழியிலிருந்து வந்த சொல்.

Credits: Zinkie Aw Singapore Lang Rocks . Logo and text used with explicit permission http://singaporelang.rocks/glossary/lim-kopi

Kopi the local way

Sweet
Local
Coffee

Sweet
Local
Tea

Coffee,
No milk,
No sugar

Coffee,
No milk,
Less strong,
No sugar

Coffee
With milk,
No sugar

Coffee,
Sweeter

Iced
Coffee

Iced Coffee,
To go

EnjoyDrink Supervised Learning

Positive (*aka* +ve, +, 1, 1, True)	Negative (*aka* –ve, –, –1, 0, False)

EnjoyDrink Supervised Learning

Positive (*aka* +ve, +, 1, 1, True)

Sweet
Local
Coffee

Negative (*aka* –ve, –, –1, 0, False)

EnjoyDrink Supervised Learning

Positive (*aka* +ve, +, 1, 1, True)

Sweet
Local
Coffee

Iced
Coffee

Negative (*aka* –ve, –, –1, 0, False)

EnjoyDrink Supervised Learning

Positive (*aka* +ve, +, 1, 1, True)

Coffee,
No milk,
No sugar

Iced
Coffee

Negative (*aka* –ve, –, –1, 0, False)

EnjoyDrink Supervised Learning

Positive (*aka* +ve, +, 1, 1, True)

Coffee,
No milk,
No sugar

Sweet
Local
Tea

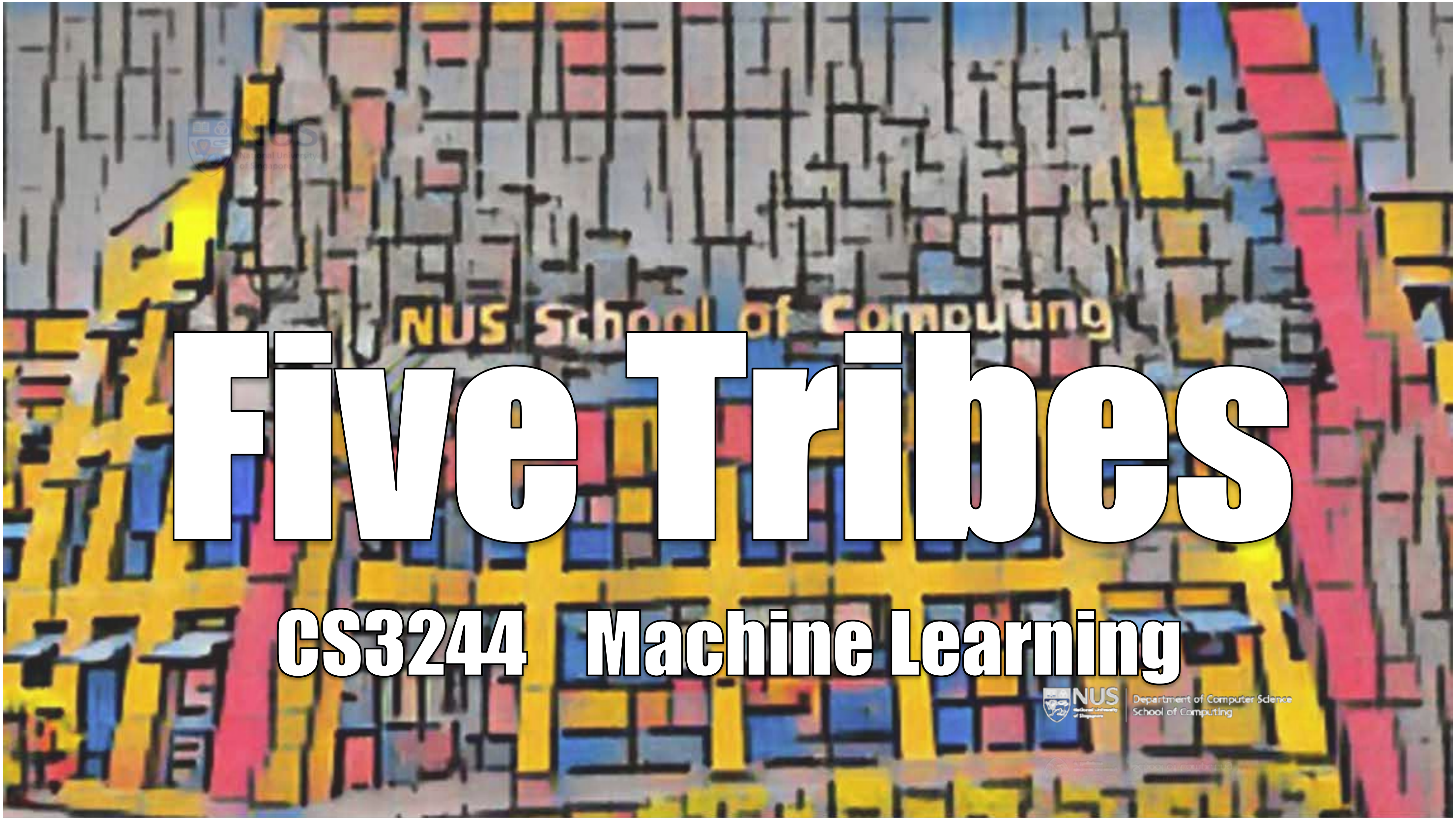
Negative (*aka* –ve, –, –1, 0, False)

Futility of Bias Free Learning

A learner that makes no *a priori* assumptions on f has *no rational basis* for classifying any unseen instances.

Making prior assumptions – Inductive Bias – is the only way to make learning feasible.

Each family ("tribe") of models $\mathcal{A}$ has a different take on its inductive bias, which forms of h it can represent: $\mathcal{H}$.

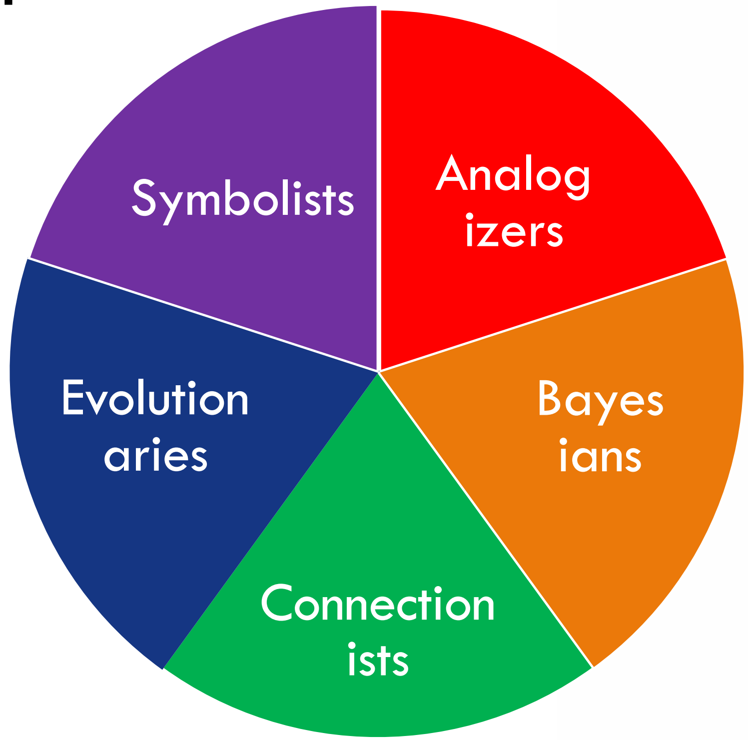


Five Tribes

CS3244 Machine Learning

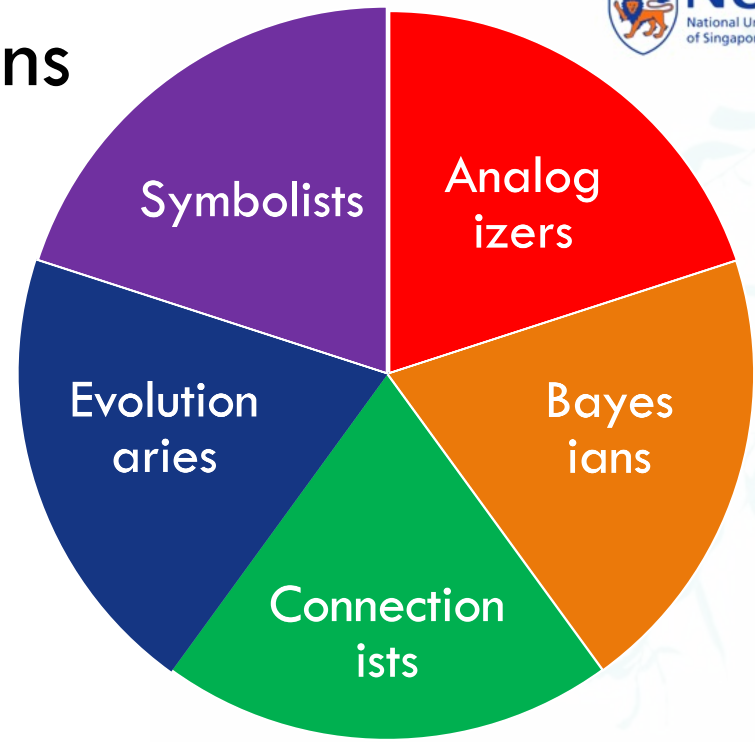
NUS
Department of Computer Science
School of Computing

Different Disciplines, Different Representations



Different Disciplines, Different Representations

- Statistics
- Neuroscience
- Biology
- Philosophy and Linguistics
- Psychology



1. Weighing Evidence: Bayesians

"There is nothing certain, but the uncertain" – Proverb

Bayes Rule:

Posterior: How probable the instance a positive case?

Likelihood: How probable is the data given that our class is positive?

Prior: How probable is an instance to be positive without seeing any data?

$$P(y|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|y)P(y)}{P(\boldsymbol{x})}$$

Marginal: How probable is the evidence under both classes of instances (both positive or negative)

Taught in SoC in CS3243 Foundations of Artificial Intelligence

Solve:
Probabilistic Inference

Evaluate:
Posterior Probability

Representation:
Graphical Models

Pre-Lecture Activity from last week

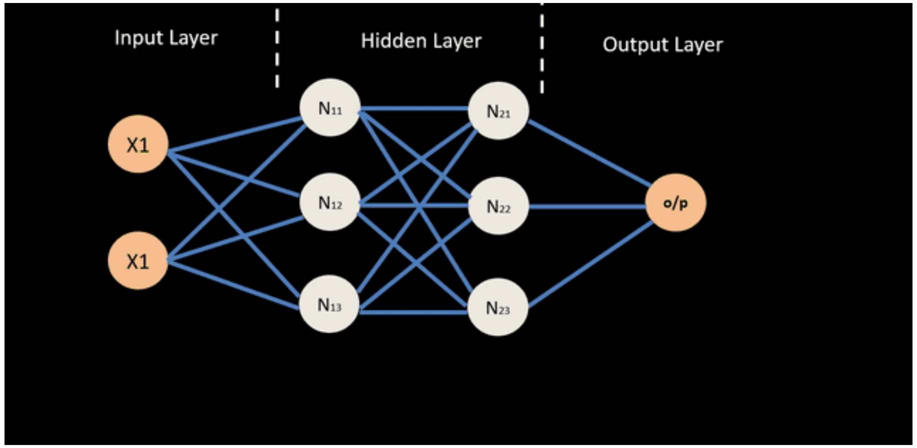
This is one of ML applications on AI customer service. https://emailtree.ai/?gclid=CjwKCAjwsNiIBhBdEiwAJK4khuV00sx6x7VzGpZdZwWWU429YIlv5MD0XRL-AnLjtYX9dMwPrfIluxoCHqUQAvD_BwE

I think this application uses supervised learning paradigm and The Bayesians tribe.

> **EmailTree AI**
>
> Customer Service Automation - EmailTree AI
>
> EmailTree AI helps to reinvent the customer experience with an intelligent end-to-end customer service automation. (62 kB)
>
> EMAILTREE.AI

2. Estimating Parameters: Connectionists

"Small changes add up to big differences."

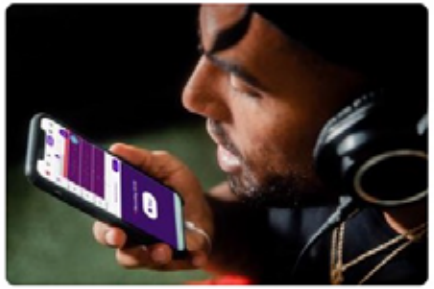


The bulk of our supervised learning. Taught in: [W04, W09–10]

Solve: Gradient Descent
Evaluate: Loss Function
Representation: Neural Network

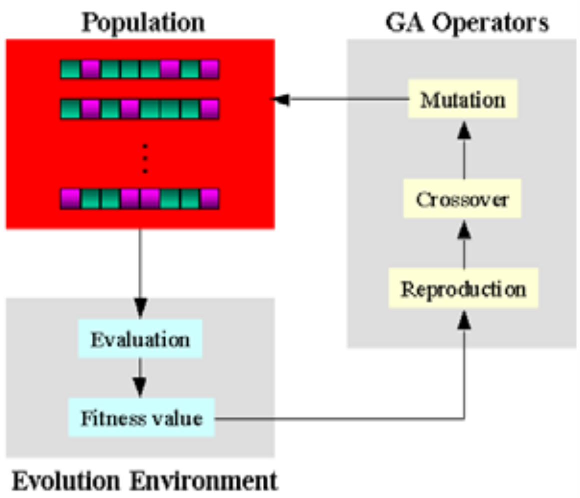
Link: https://towardsdatascience.com/audio-deep-learning-made-simple-automatic-speech-recognition-asr-how-it-works-716cfce4c706
Application: Speech-to-Text
Paradigm: Supervised Learning - training data consists of audio clips of spoken words (input) and text transcript of what is spoken (target label)
Tribe: Connectionists - common approach is to use neural networks to distinguish each character of the words

> **Medium**
>
> **Audio Deep Learning Made Simple: Automatic Speech Recognition (ASR), How it Works**
>
> Speech-to-Text algorithm and architecture, including Mel Spectrograms, MFCCs, CTC Loss and Decoder, in Plain English
>
> **Reading time**
>
> 14 min read
>
> 25 May



3. Structure Learning: Evolutionaries

"It is not the strongest of the species that survives, nor the most intelligent that survives. It is the one that is the most adaptable to change."



Genetic Algorithm Evolution Flow

Solve: Genetic Search
Evaluate: Fitness / Reward
Representation: Genetic Programs

Taught in SoC in CS3243 Foundations of Artificial Intelligence

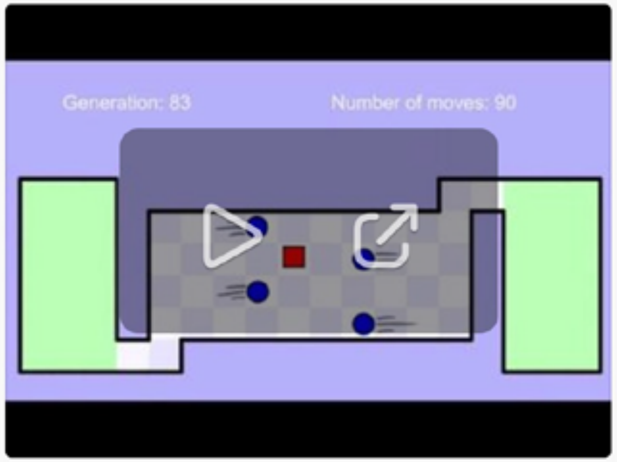
Ying-Hong Liao, and Chuen-Tsai Sun: [An Educational Generic Algorithms Learning Tool]

Links:
https://youtu.be/Yo2SepcNyw4
https://github.com/Code-Bullet/WorldsHardestGameAI

Now that I've seen the other answers, mine is kinda lame, but here we have a youtuber who uses a genetic algorithm to improve on his game playing AI. I believe in this case its paradigm is reinforcement learning, as the agent tries to maximizes its rewards by going the furthest. It belongs to the evolutionaries tribe as it uses the genetic algorithm which mimics how evolution works.

YouTube | Code Bullet

AI Learns to play the Worlds Hardest Game



4. Structured Inference: Symbolists

"Once you have eliminated the impossible, whatever remains, however improbable, must be the truth."

Modus Ponens:

All men are mortal
Socrates is a man

Therefore, Socrates is mortal.

$$\frac{P \rightarrow Q, P}{\therefore Q}$$

Taught in [W03]; also in SoC in CS3243 Foundations of Artificial Intelligence

Solve:	Inverse Deduction
Evaluate:	Accuracy
Representation:	Logic

Pre-Lecture Activity from last week

Additional data on road conditions. There are external data sources that can provide important information that impacts traffic. Think social media posts about sports events in the area, local news about civil protests, or even police scanners about crime scenes, accidents, or road blockages.

Fig. 1. Illustration of probe GPS data and social media data for traffic monitoring.

Traffic Prediction: How Machine Learning Helps Forecast Congestions

Paradigm: supervised learning using random forest. inputs: weather conditions, time period, special conditions of the road, road quality, and holidays; outputs:values that represent the level of congestion

Tribe: The Symbolists. The **random forest** algorithm creates multiple decision trees and merges their data to obtain accurate predictions.

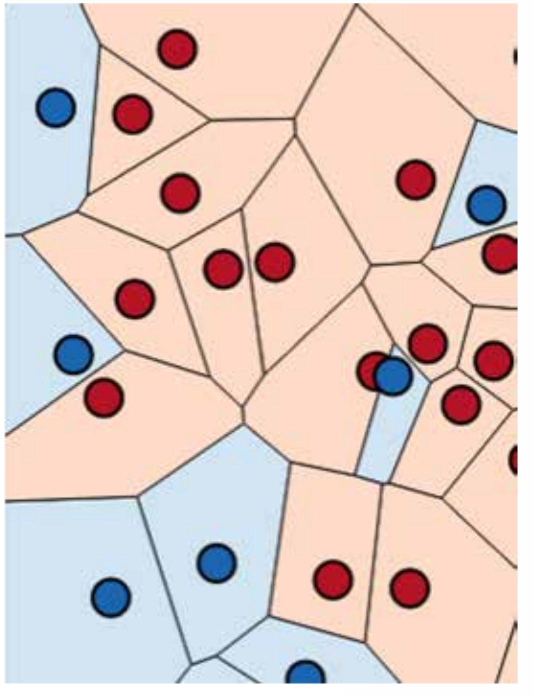
The **random forest** algorithm creates multiple decision trees and merges their data to obtain accurate predictions. It's quite fast and can produce effective results given sufficient training data. When applied to the road congestion problem, this method showed an accuracy of 87.5 percent. In this case, weather conditions, time period, special conditions of the road, road quality, and holidays are used as model input variables.

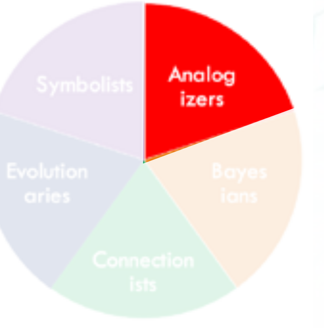
5. Mapping to Novelty: Analogizers

"Birds of a feather flock together"

Model Free:
The Model **is** the Data

Taught later today, also [W04]





Solve:
Constrained Optimization

Evaluate:
Margin

Representation:
Support Vectors

Pre-Lecture Activity from last week

Example: Weather Prediction System Using KNN Classification Algorithm
URL: https://www.ej-compute.org/index.php/compute/article/download/44/18
Tribe: The Analogizers
Paradigm: Unsupervised Learning

Summary – 5 Tribes

Each tribe has a different set of assumptions about data

You can use any tribe's methods to solve an ML problem

But a particular tribe's representations or methods may be more suited to certain problem settings

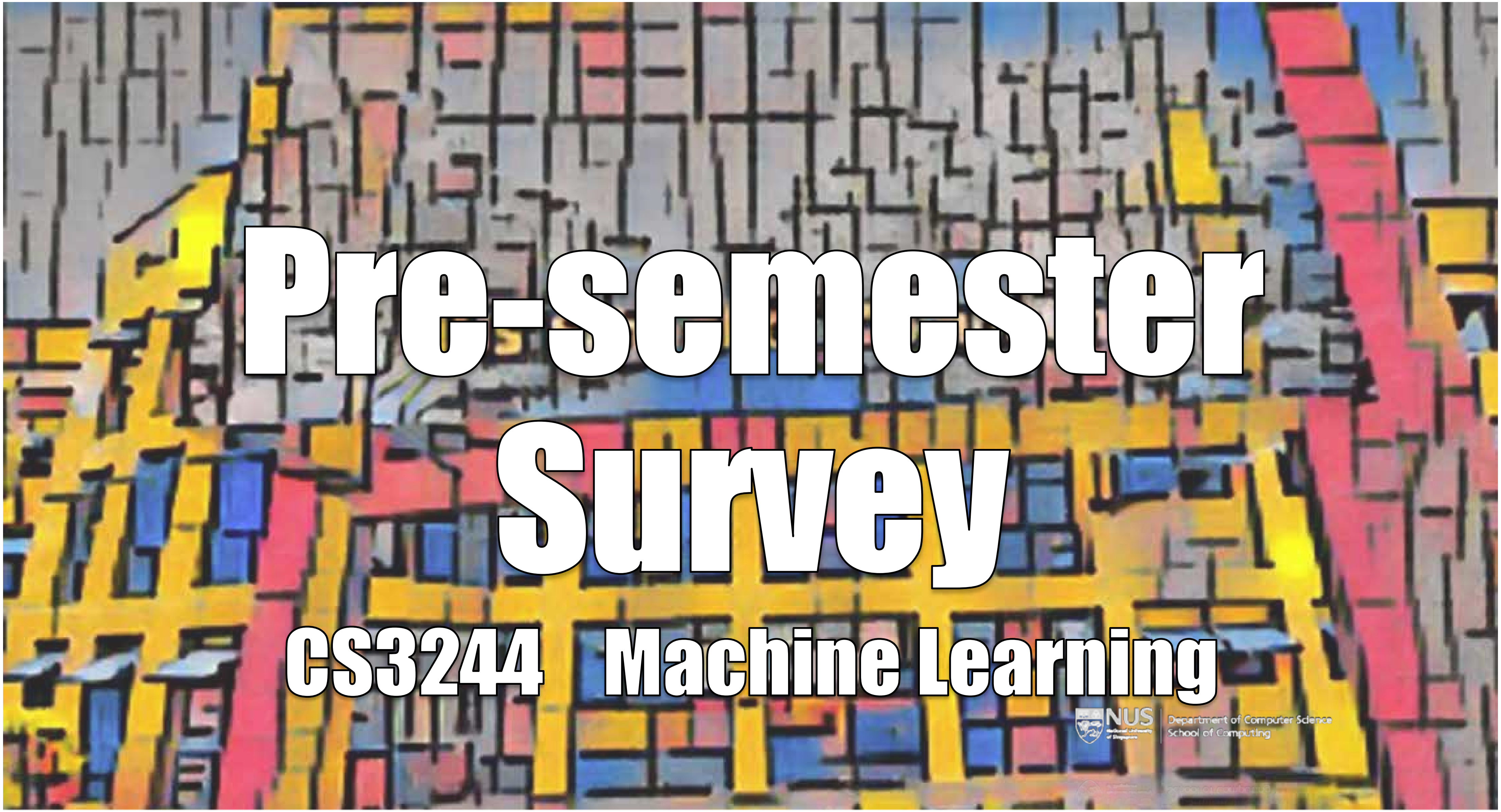
To think about 🕵: How is this last statement related to inductive bias?

The Singularity

The **technological singularity** is a hypothetical point in time at which technological growth will become radically faster and uncontrollable, resulting in unforeseeable changes to human civilization.

In the most popular version an upgradable intelligent agent enters a "runaway reaction" of self-improvement cycles, resulting in a powerful superintelligence that qualitatively far surpasses all human intelligence.

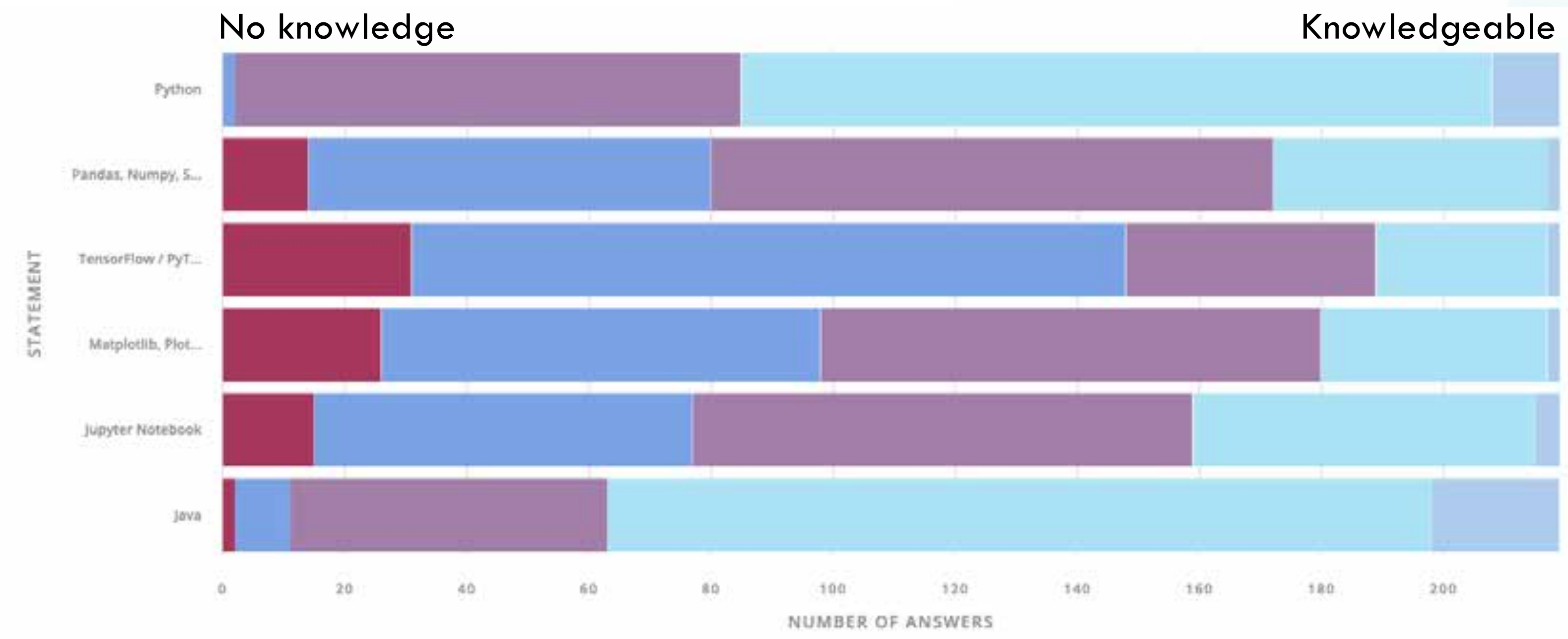
Photo by Clarisse Croset on Unsplash


Pre-semester
Survey
CS3244 Machine Learning
NUS
Department of Computer Science
School of Computing

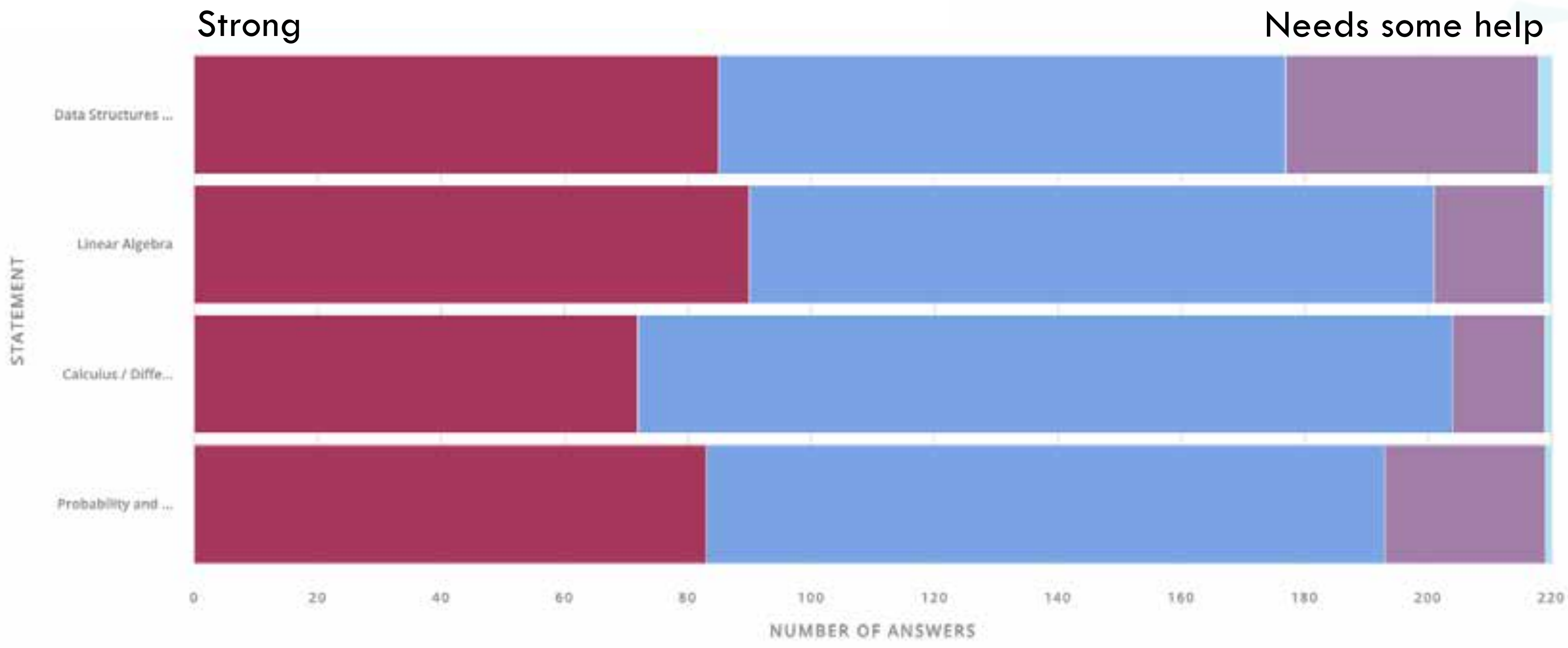
What is your major?

Options	*Count*	*Pct*
Computer Science	118	47.97 %
Data Science	57	23.17 %
Statistics	6	2.44 %
Math	8	3.25 %
Business	3	1.22 %
Science	5	2.03 %
Engineering	39	15.85 %
Other	10	4.07 %

Programming Experience



Prior Requisite Knowledge

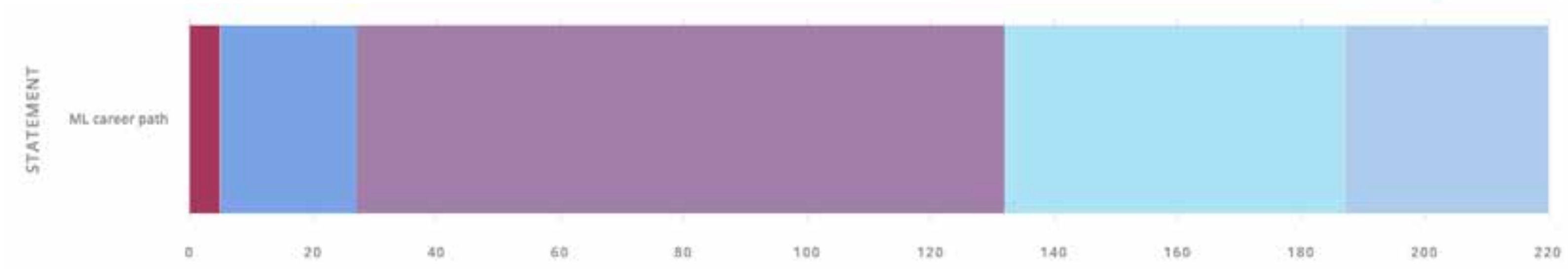


Overseas?

Options	Count	Pct
No. I'll be in Singapore.	213	96.82 %
Not sure, I might have to.	4	1.82 %
Yes, I'm not planning to travel to Singapore this semester.	2	0.91 %
Other Responses	1	0.45 %

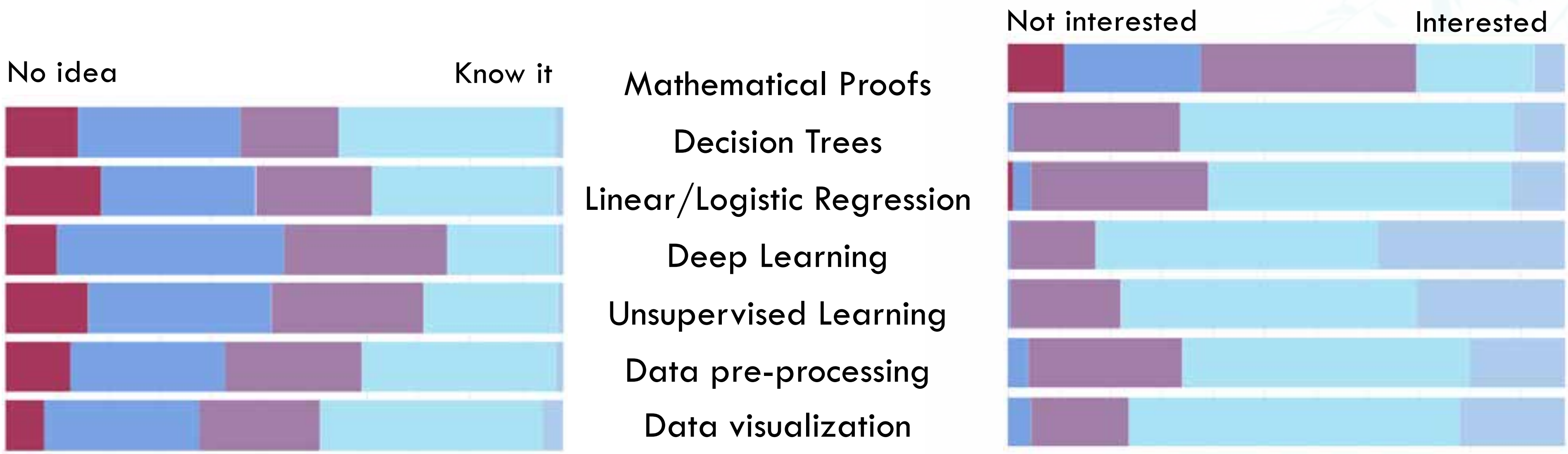
Career Path

Unlikely Likely



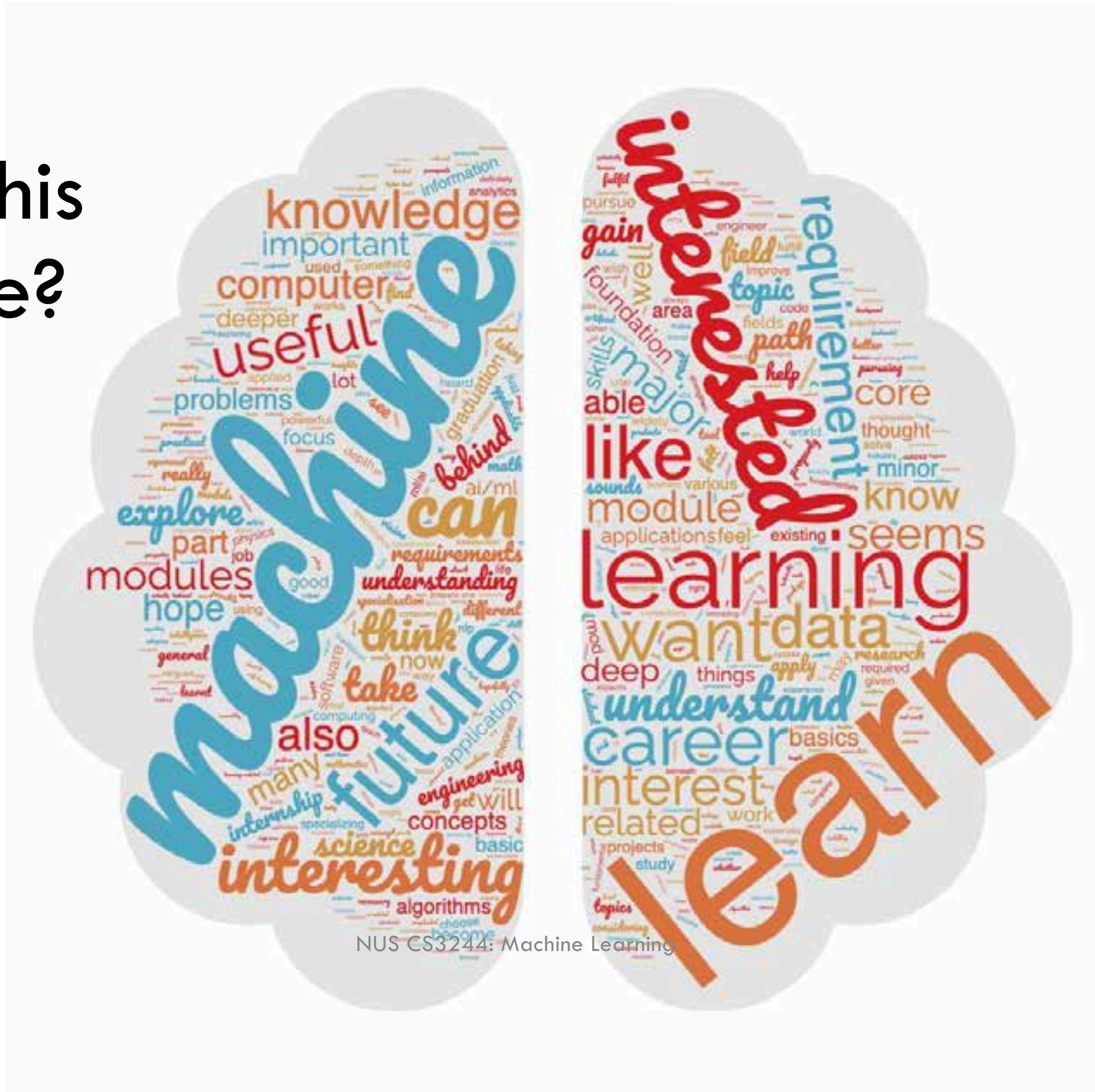
Topics in ML

Prior Knowledge **Your Interest**



Why this module?

NUS
National University of Singapore



Why this module?

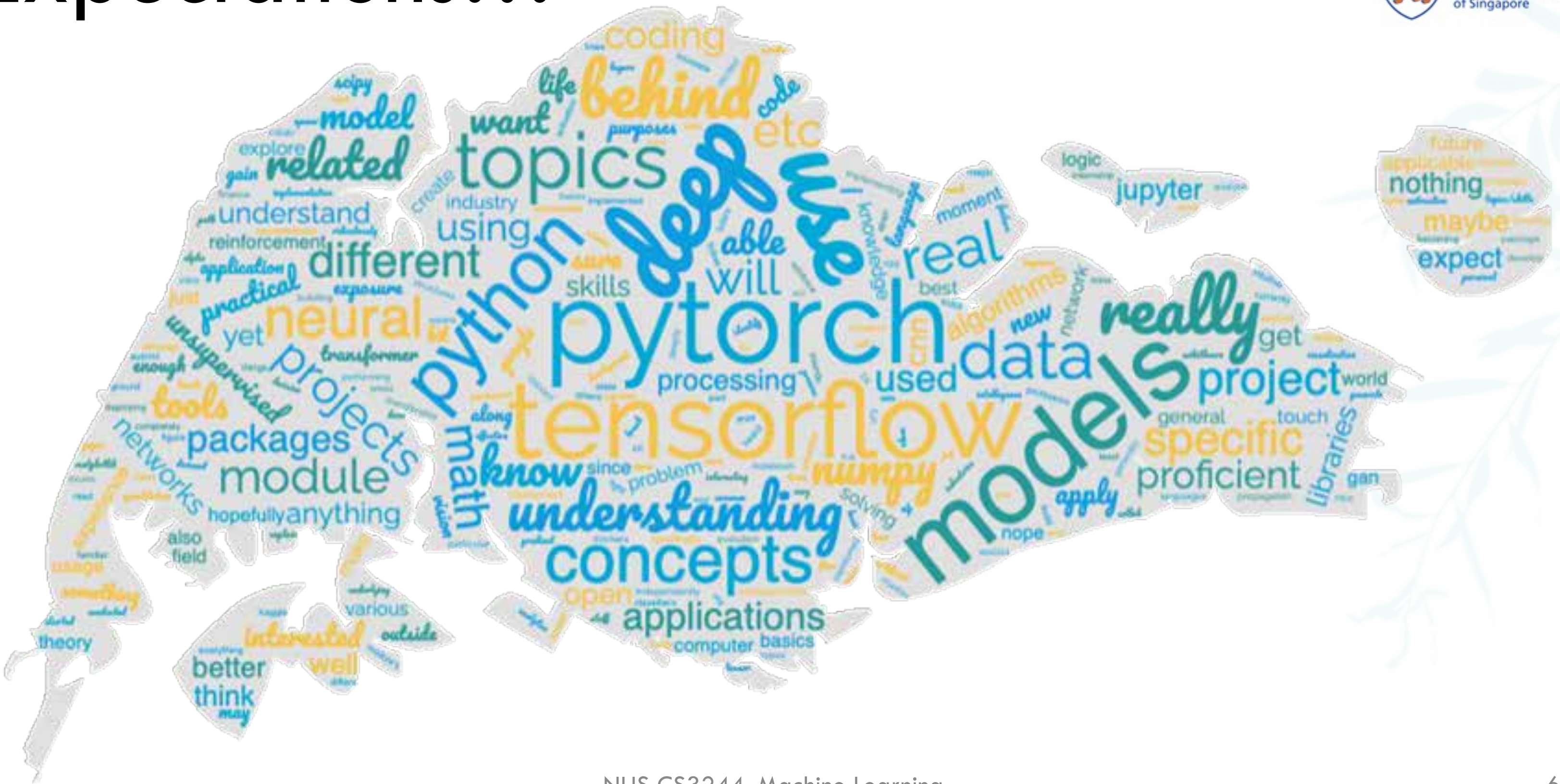
1. It is a core module, 2. I want to gain some understanding of machine learning and why it is at the top of the hierarchy of data science needs, 3. Maybe the module can help me to understand whether I want an ML-related career path.

Saw an operation that was automated with computer vision back in NS, thought it was cool. Also, I heard that AI is math heavy, so that definitely piques my interest

I feel that data is very powerful. Information is widely available but selectively choosing certain pieces of information as data, reshaping the data, analysing the data and then evaluating the data … gain insights … useful to improve existing processes or provide solutions to existing problems

Machine Learning is cool :)

Expectations…



Expectations…

Google colab and how it differs from jupyter notebook

To reinforce my intuitive understanding of ML with more math.

How to know when to use machine learning? and how do I get started with a machine learning project?

Would like to learn the math behind neural network concepts like activation functions, gradient descent, backward propagation

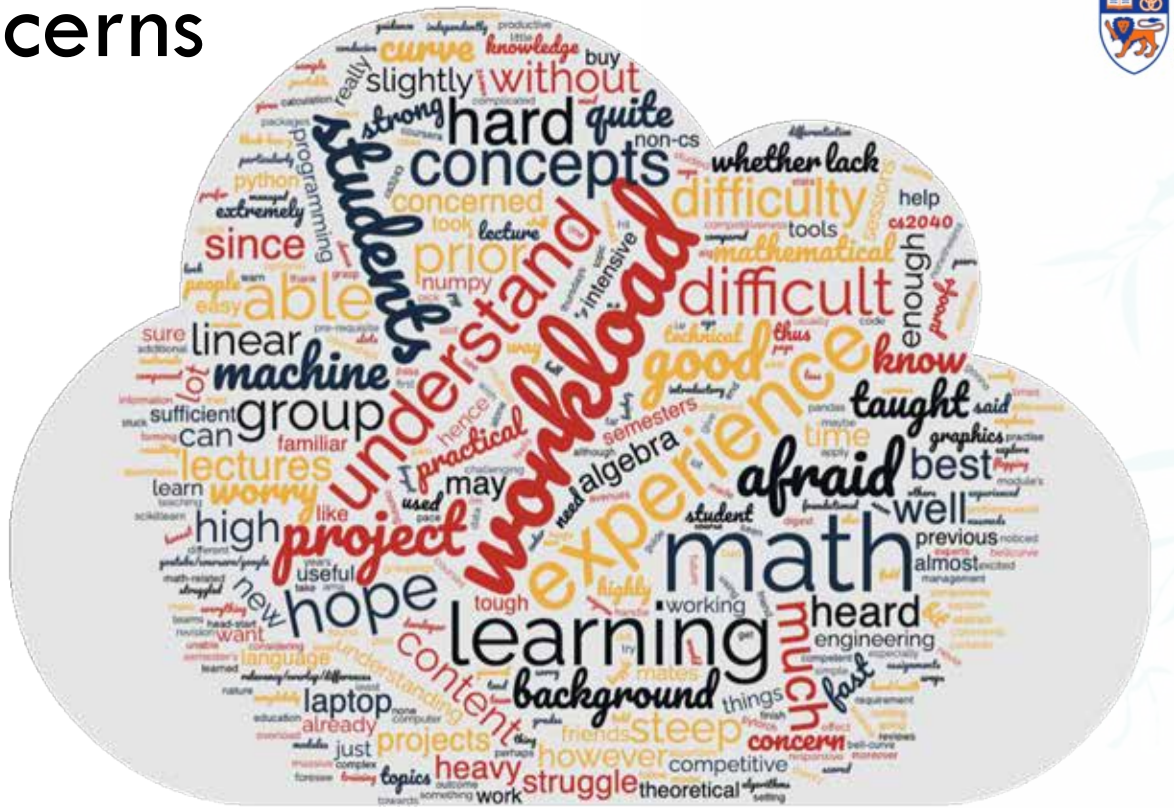
PyTorch and all the basics of ML. I hope to be able to learn how to do real ML projects from this module.

Path way to be ML engineer

Natural Language Processing and Computer Vision

reinforcement learning … Training models to act as an artificial intellegence such as Alpha GO

Your Concerns



Your Concerns

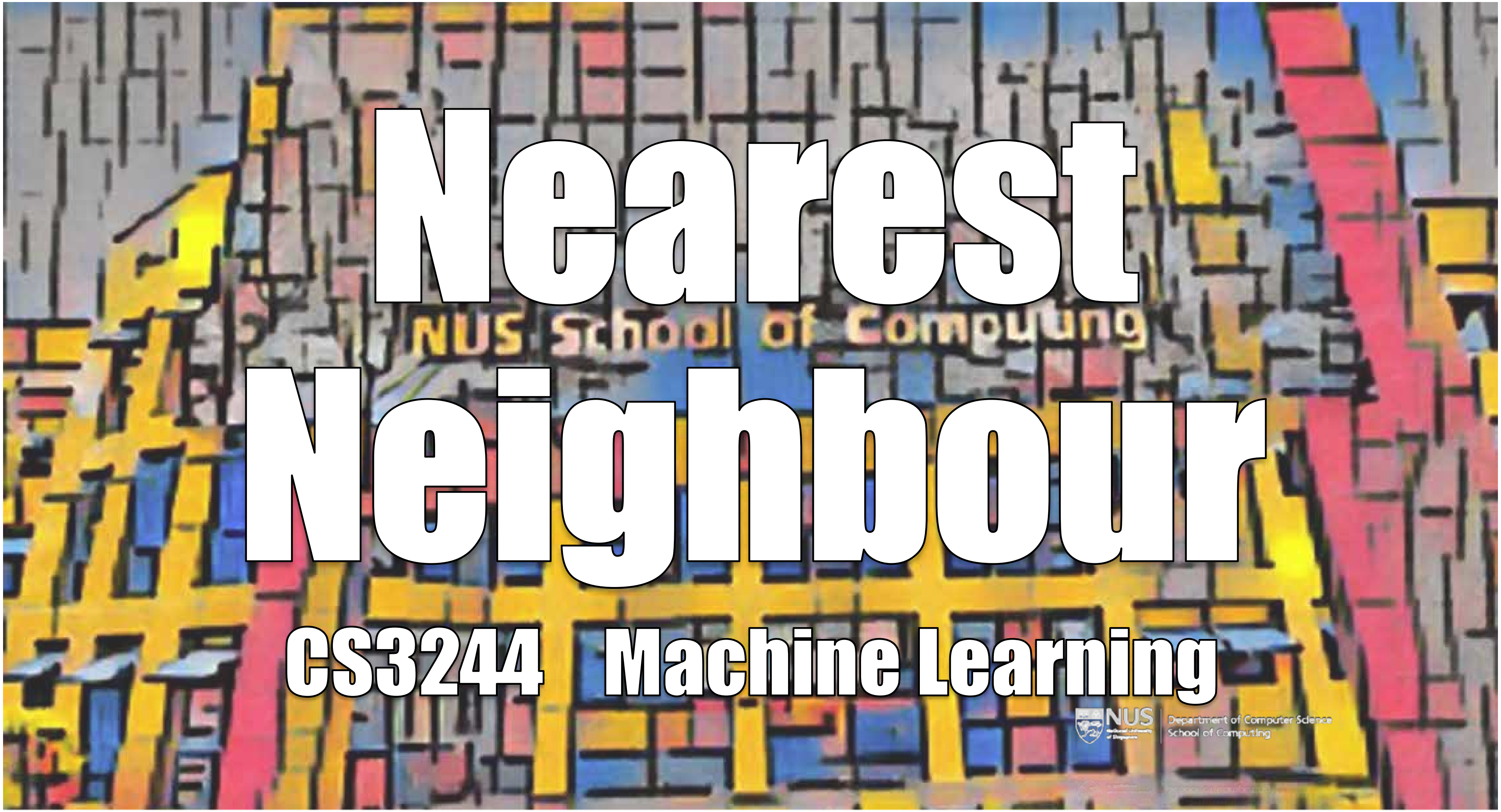
I hope my project mates are strong.

I've heard that people taking this module usually already have prior experience in ML, so i'm slightly worried about the bell curve.

I am worried that there may not be enough guidance or sufficient avenues for help if we need it. Also worried about high difficulty and workload.

I have seen past module reviews. Highly theoretical and less practical in the working world. ...Hope this semester's version will be more practical and useful for my future internships.

prefer able to choose teammates too theory, i want more practical teaching team not responsive not enough materials like PYP to practise


Nearest
Neighbour
NUS School of Computing
CS3244 Machine Learning
NUS
Department of Computer Science
School of Computing

5 tribes: The Analogizer

Compare something with something else to assist understanding

To classify a test instance, assign it to its closest neighbor

Train: 🧠 Memorize the training data

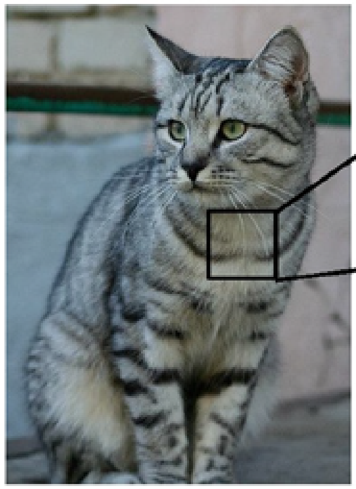
Test: Compute which training instance is closest to the test instance

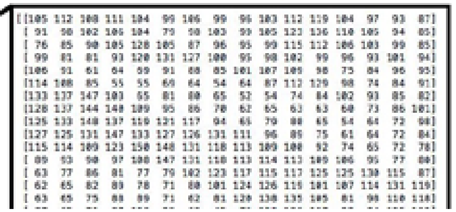
Recap: image representation

The semantic gap: We see a cat…

Recap: image representation

The semantic gap: We see a cat…





…but the computer just sees a grid of numbers.

Here we have a color image with 3 channels (R,G,B) and 256 levels of intensity.

Width × Height × Channels
600 × 800 × 3 = 1.44M integers

CIFAR 10 Dataset

10 classes

10K test instances

50K training instances

Distance Metric

The key to determining the meaning of "nearest"

L1 distance metric: $d\left(x^{(*)}, x^{(1)}\right) = \sum_{n} \left| x_n^{(*)} - x_n^{(1)} \right|$

We can apply this to:

- Over the entire image in grayscale: x_0
- Over each channel separately to produce 3 inputs: x_0, x_1, x_2
- In other ways…

Distance Metric

The key to determining the meaning of "nearest"

L1 distance metric: $d\left(x^{(*)}, x^{(1)}\right) = \sum_{p} \left| x_p^{(*)} - x_p^{(1)} \right|$

$$\left| \text{test image} - \text{training image} \right| = \text{pixel-wise absolute value differences}$$

test image

56	32	10	18
90	23	128	133
24	26	178	200
2	0	255	220

-

training image

10	20	24	17
8	10	89	100
12	16	178	170
4	32	233	112

=

pixel-wise absolute value differences

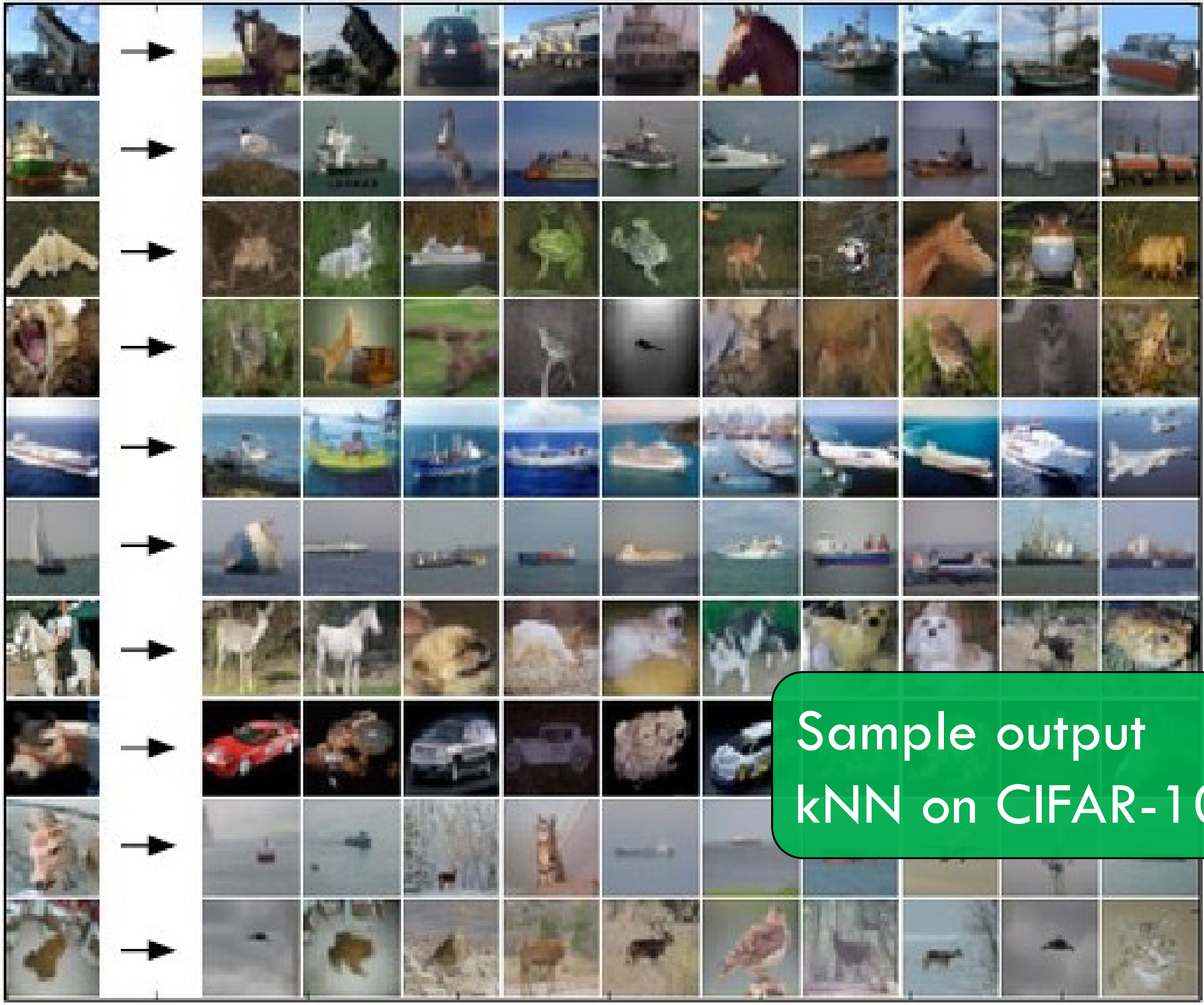
Distance Metric

The key to determining the meaning of "nearest"

$$d(x^{(*)}, x^{(1)}) = \sum_p \left| x_p^{(*)} - x_p^{(1)} \right|$$

L1 distance metric: $$d\left(x^{(*)}, x^{(1)}\right) = \sum_p \left| x_p^{(*)} - x_p^{(1)} \right|$$

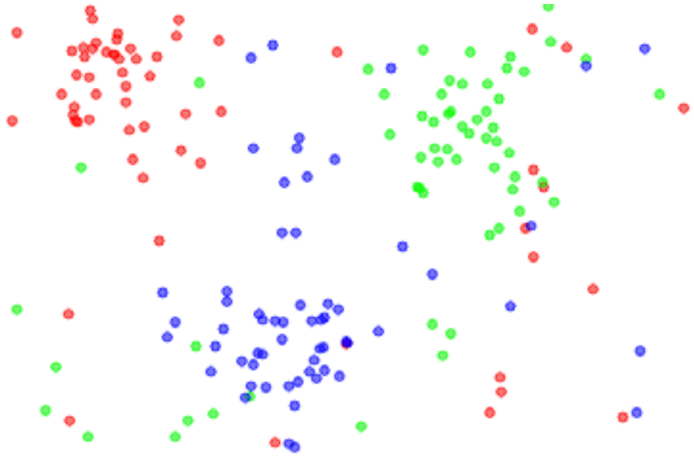
$$\left| \overset{\text{test image}}{\begin{array}{|c|c|c|c|} \hline 56 & 32 & 10 & 18 \\ \hline 90 & 23 & 128 & 133 \\ \hline 24 & 26 & 178 & 200 \\ \hline 2 & 0 & 255 & 220 \\ \hline \end{array}} - \overset{\text{training image}}{\begin{array}{|c|c|c|c|} \hline 10 & 20 & 24 & 17 \\ \hline 8 & 10 & 89 & 100 \\ \hline 12 & 16 & 178 & 170 \\ \hline 4 & 32 & 233 & 112 \\ \hline \end{array}} \right| = \overset{\text{pixel-wise absolute value differences}}{\begin{array}{|c|c|c|c|} \hline 46 & 12 & 14 & 1 \\ \hline 82 & 13 & 39 & 33 \\ \hline 12 & 10 & 0 & 30 \\ \hline 2 & 32 & 22 & 108 \\ \hline \end{array}} \xrightarrow{\text{add}} 456$$

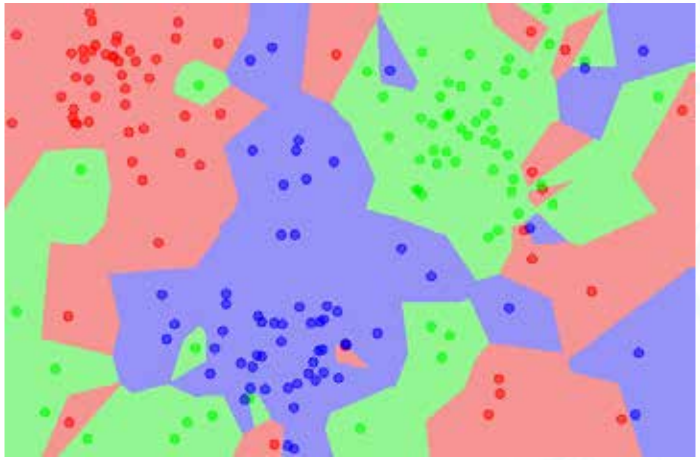


Sample output
kNN on CIFAR-10

So, what does it look like?

Here's a three-class problem, plotted using just two input features:



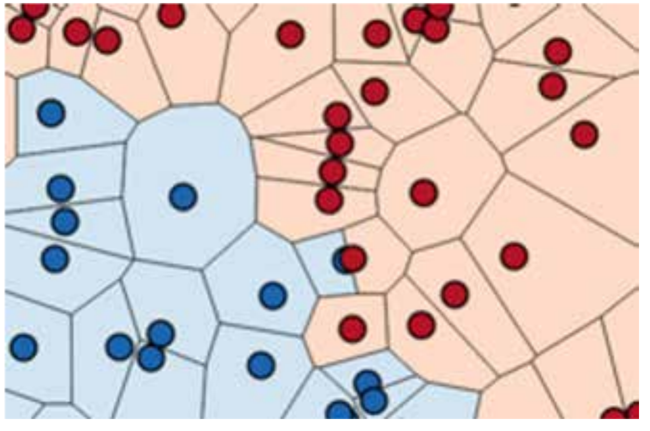


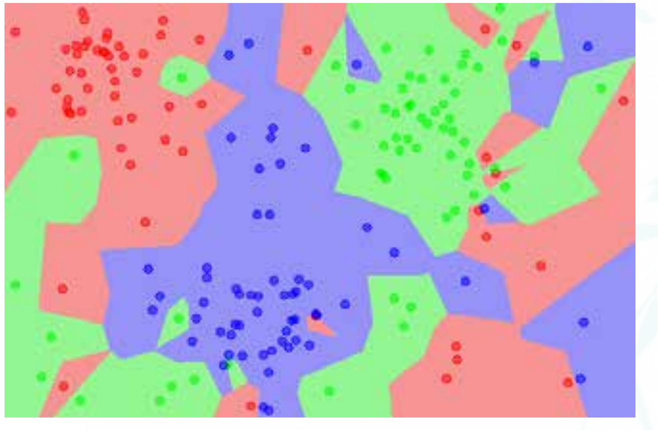
Caution: different meaning of X–Y than in the previous examples!

Photo credits: Agor153 - Own work, CC BY-SA 3.0

So, what does it look like?

Here's a three-class problem, plotted using just two input features:





The decision boundaries form a Voronoi tessellation.

Photo credits: Agor153 - Own work, CC BY-SA 3.0
and http://scott.fortmann-roe.com/docs/BiasVariance.html

Summary – Nearest Neighbor

A model free-algorithm (non-parametric)

Requires an immense amount of time to apply
(needs to cycle through all the training data per test instance)

How can we speed this up?

Doesn't work well in high dimensions (why?)


K nearest neighbours
CS3244 Machine Learning
NUS
Department of Computer Science
School of Computing

Get to know your neighbors

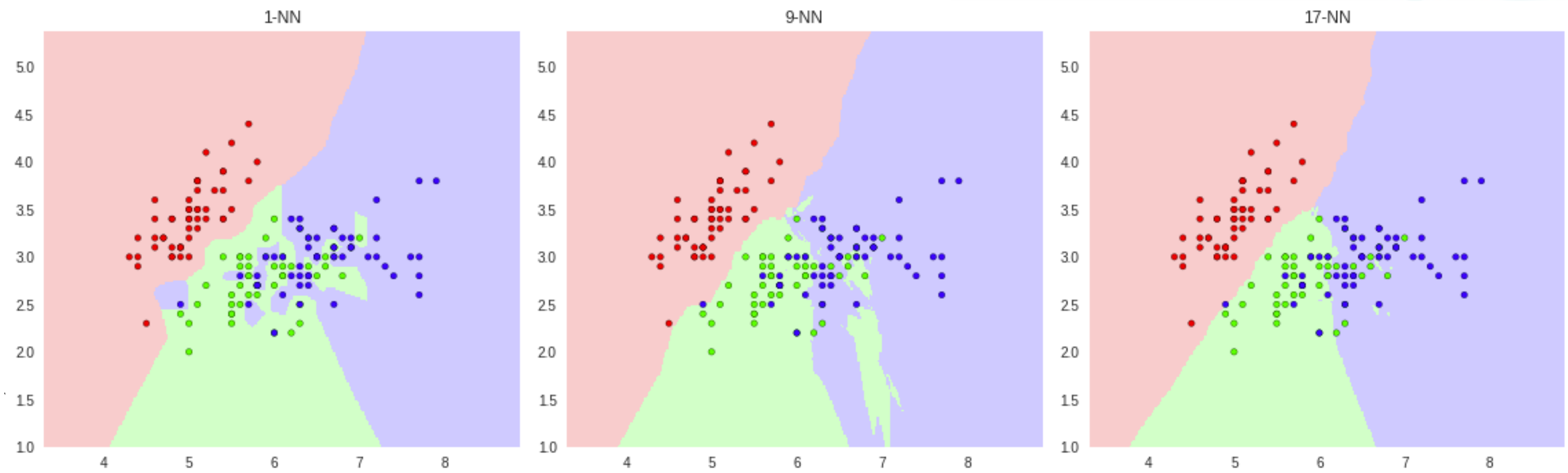
Let the community decide.

Use k closest training data points, rather than the closest.

Usually set k to an odd number to have fewer ties.

Effect of k

Smaller k: complex surface.
Larger k: smoother surface.



Perfect performance

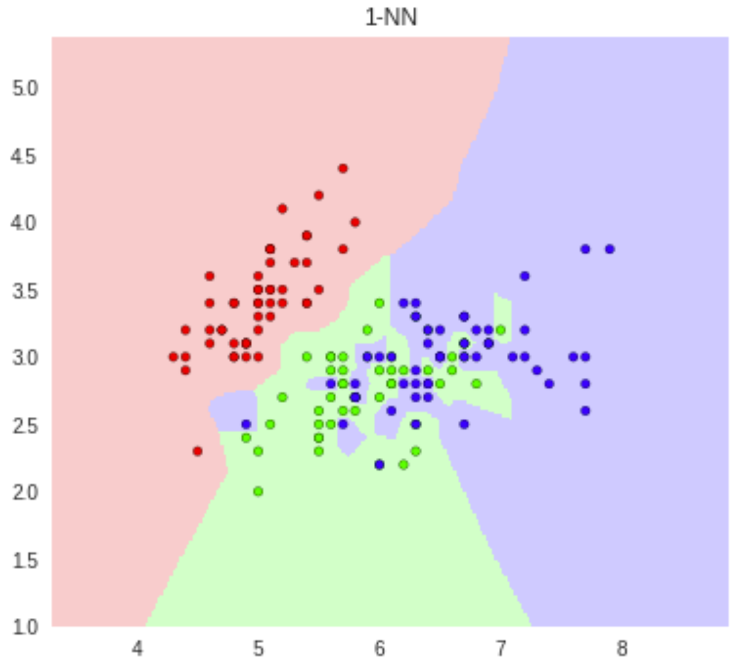
What value of k maximizes performance on the training data?

Perfect performance

What value of k maximizes performance on the training data?

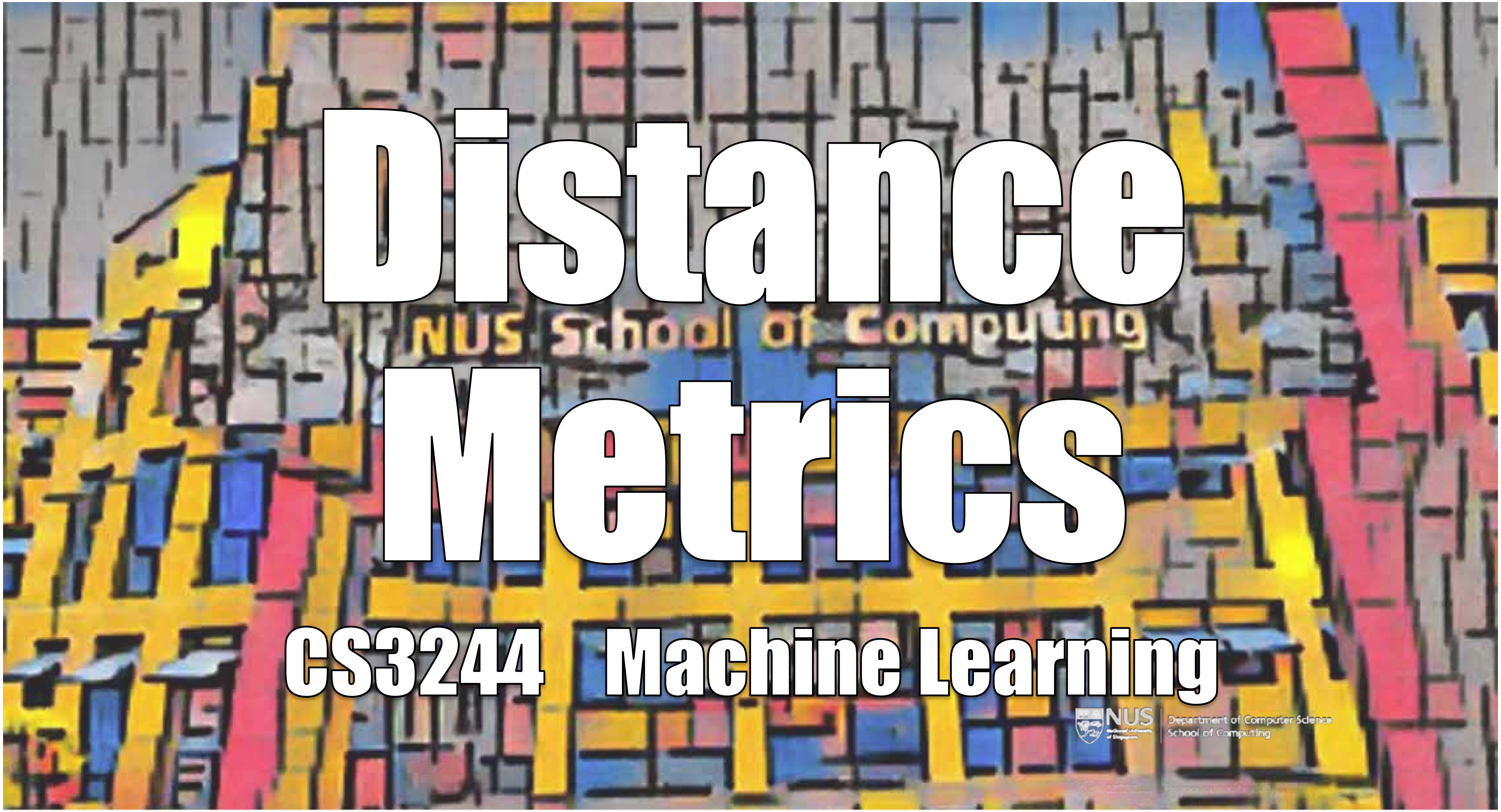
Answer: $k = 1$. Trivially, each point is classified to its own label.

Will it generalize?



Distance Metrics

CS3244 Machine Learning



Distance Metric

Often we'll use Euclidean distance as our first try:

$$L2\left(x^{(*)}, x^{(i)}\right) = \sqrt{\left(x_1^{(*)} - x_1^{(i)}\right)^2 + \left(x_2^{(*)} - x_2^{(i)}\right)^2 + \cdots + \left(x_n^{(*)} - x_n^{(i)}\right)^2}$$

but any *distance metric* (*non-negativity, symmetric, obeys the triangle inequality*) is possible.

Defining your distance metric to fit your data's semantic meaning can be important.

Distance Metrics Behaving Badly

k NN almost never used on images in practice.

Distance metrics on pixels are not informative,
e.g. our L1 distance:

The 3 images on the right have the same distance to the original.

Metric for the Credit Problem

Criterion	Value
Age	32
Gender	~~Male~~
Salary	40,000
Debt	26,000
…	…
Years in Job	1
Years at Current Residence	3

In subgroups here in i3 (or in Zoom breakout rooms), do:

(1 min): Introduce yourself! 🙃

(5 mins): Answer: *Does standard Euclidean Distance (L2) or Manhattan Distance (L1) work well for this problem?* If your team wants, suggest any modifications.

Ask one member to write it to the #lectures thread, and @mention your team members. Be kind and also upvote other teams' responses that you like.

Summary – k NN

k Nearest Neighbor predicts a target input's class by the closest k training examples.

Distance metric and k are *hyperparameters*.

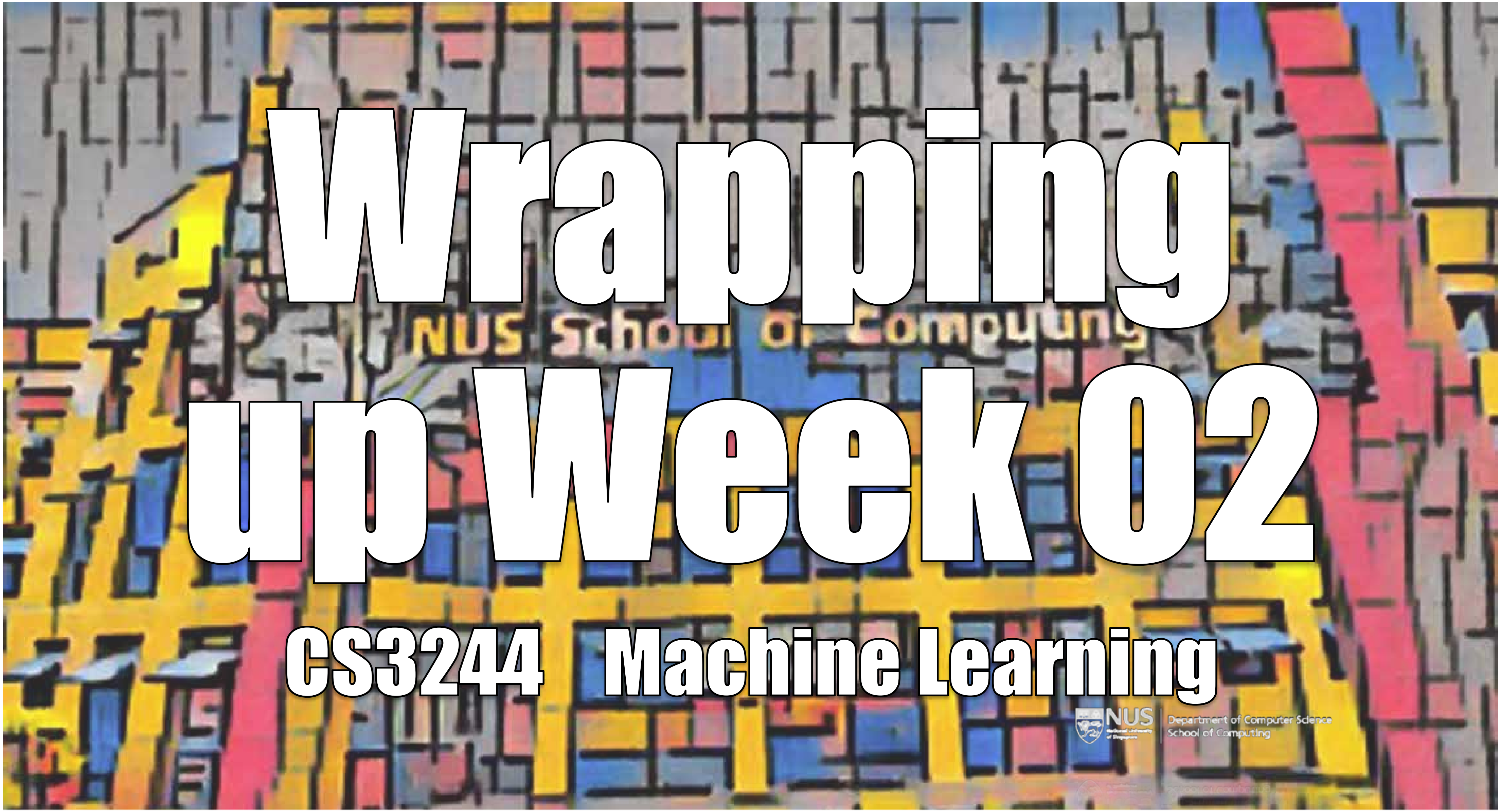
Set hyperparameters by:

- Using our prior knowledge of the data
- Using a reserved part of the training data: the *validation* set
- We'll look into this problem in future lectures

Run on the test set once at the end.

Wrapping up Week 02

CS3244 Machine Learning



What did we learn this week?

- Understand how we can define the learning problem and when it is appropriate to apply
- Dissect the components of a machine learning model
- Relate tribes of ML as means of picking a learning algorithm and hypothesis space.
- Understand an instance of the analogizer tribe: k nearest neighbors, and how its hyperparameters affect its decision making.

Outlook for next week

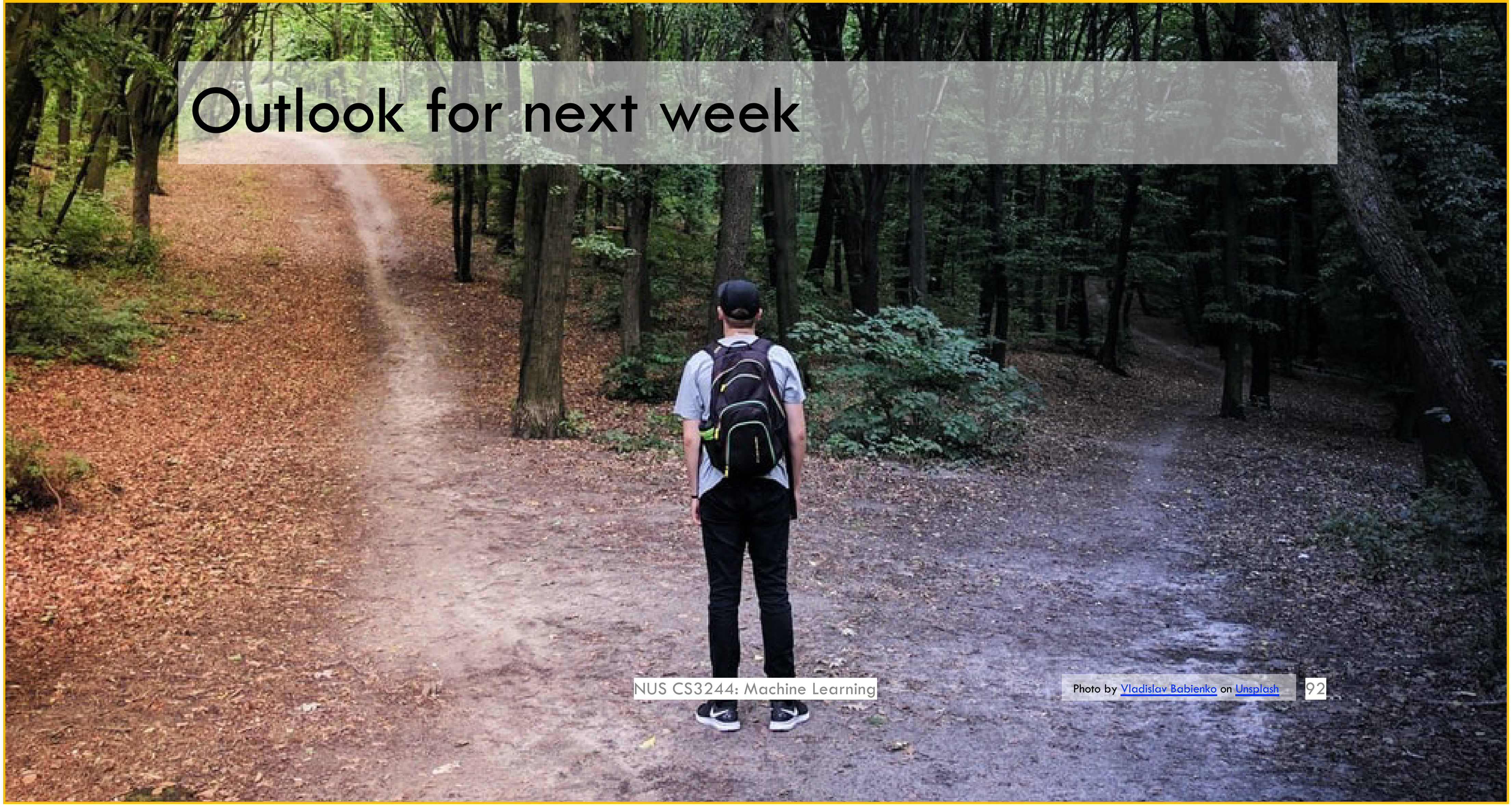


Photo by Vladislav Babienko on Unsplash

Assigned Tasks (due before next Mon)

(10 mins) What context and variables do you think about when making one of the following decisions? What logic do you use to make it?

- Which module you enrolled in?
- What lunch you ate today?
- Which clothes you are wearing now?

Post a 1-2 sentence (not long!) answer to the above in your Slack tutorial group #tg-xx or #tg-na (if you don't have a tutorial group yet).

Assigned Tasks (due before next Mon)

(5 mins) Form your project sub-groups.

- If you're going solo or already have a friend or two, you're done, just fill in the LumiNUS survey.
- If not, attend our 2B session which will also give more details on the project.

(30 mins, optional) For those new to Colab / Jupyter or needing a bit of a math refresher, here's a self-study Colab notebook for you to do.

http://www.comp.nus.edu.sg/~cs3244/AY22S1/02.colab.html