INTRODUCTION

2110573 Pattern Recognition

Sections

- 21 : Undergrad
- 1 : Masters

- Switch sections if wrong, if full please tell me to increase.

Mycourseville and discord

Discord

https://discord.gg/2usmexXj

TA office hours: 10-11.30 pm Mondays, Tuesdays, Fridays – These are official working hours

Private DMs (unrelated to personal issues will be ignored)

MyCourseVille

https://www.mycourseville.com?q=courseville/course/46136

Password: nya

For homework submission

Github

https://github.com/ekapolc/Pattern_2024

For slides and homework instructions

Playlist

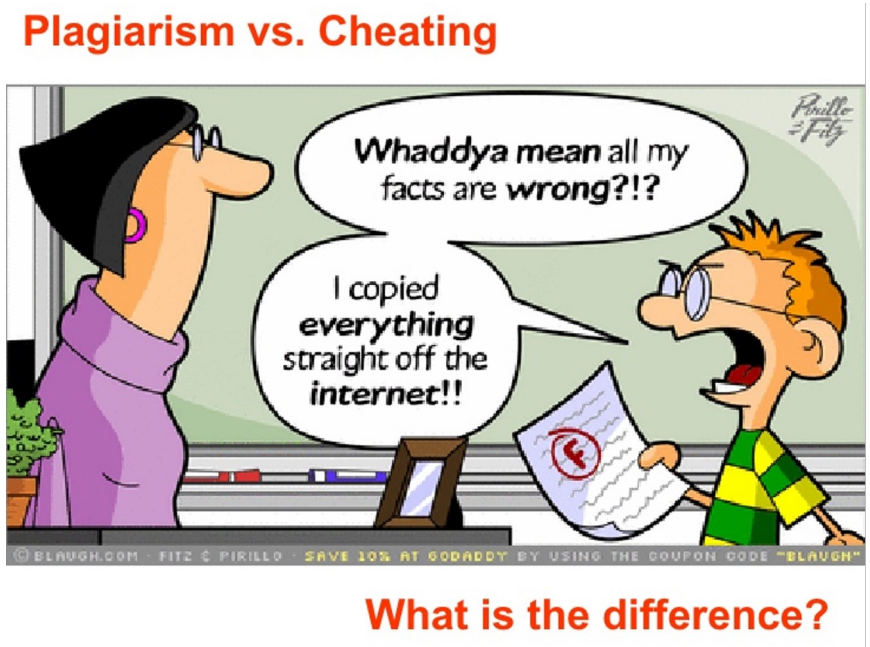
https://www.youtube.com/playlist?list=PLcBOyD1N1T-OpGooU_P9nFL9l3l6liDgu

Syllabus

คาบเรียนที่	เนื้อหา	การบ้านและควิซ
1 - 11/1	Introduction, K-mean	เริ่มHW1
2 - 28/1	Regression, MLE & MAP	
3 - 25/1	Naive Bayes & GMM	ส่งHW1, Quiz 1, เริ่มHW2
4 - 1/2	GMM, EM, ELBO, Dimensionality reduction I (PCA)	
5 - 8/2	Dimensionality reduction II (LDA, RP) and visualization techniques (t-sne, UMAP, PHATE)	ส่งHW2, Quiz 2, เริ่มHW3
6 - 15/2	SVM, NN I	
7 - 22/2	NN II (CNN & Recurrent)	ส่งHW3, Quiz 3, เริ่มHW4
8 - 29/2	NN III (Architectures) & Pytorch demo	เริ่มHW5
9 - 7/3	Midterm week - No midterm for this class	
10 - 14/3	Transformers & Self-supervised I	ส่งHW4, Quiz 4
11 - 21/3	Self-supervised learning II	ส่งHW5, Quiz 5, ส่ง course project proposal, เริ่มHW6
11 - 28/3	Generative models I (GAN, VAE)	เริ่มHW7
12 - 4/4	Generative models II (Diffusion)	ส่งHW6, Quiz 6, เริ่ม HW8
13 - 11/4	Reinforcement Learning	ส่งHW7, Quiz 7
14 - 18/4	No regular class - meeting/progress presentation with project mentors	Course project progress
15 - 25/4	Tricks of the trade: machine learning in the real world + Guest	ส่งHW8, Quiz 8
Some time during final exam	Project presentation No final exam for this class	ส่งcourse project

Plagiarism Policy

- You shall not show other people your code or solution
- Copying will result in a score of zero for both parties on the assignment
- Many of these algorithms have code available on the internet, do not copy paste the codes



Grades

การส่งการบ้านสาย

สายไม่เกิน 6 ชม. -0.5 คะแนน

สายไม่เกิน 24 ชม. -2 คะแนน

ถ้าส่งสายเกิน 24 ชม.จะไม่ได้รับการตรวจ

เกณฑ์การวัดผล

Attendance and in-class activities 10%

Quizzes 20%

Homework 40%

Project 30%

การตัดเกรด

> 85% A

> 80% B+

> 75% B

> 70% C+

> 65% C

> 60% D+

> 55% D

< 55% F

Notes regarding homework submission

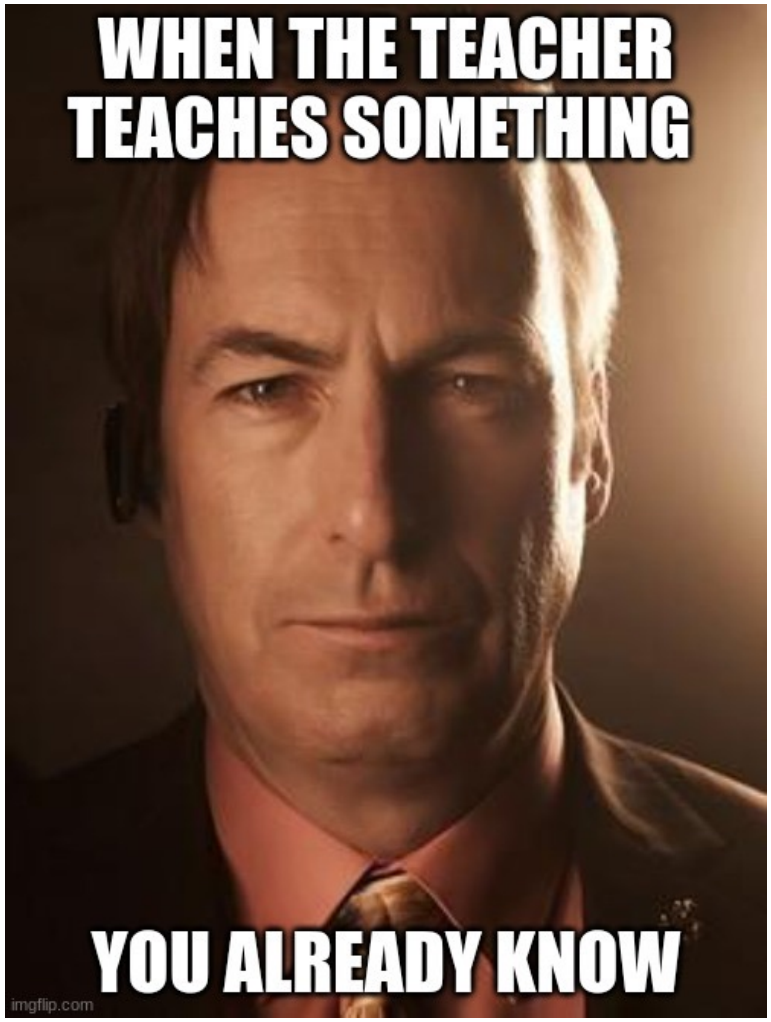
- Please submit everything as pdf
  - The TAs will mostly grade this
  - If you have a collab/python code – then export as pdf
  - Combine the materials into a single pdf file

- Additional materials
  - Can submit additional code or results in a .zip file

Course project

- <= 5 people
- Topic of your choice
  - Can be implementing a paper
  - Extension of a homework
  - Project for other courses with an additional machine learning component
  - Your current research (with additional scope)
  - Or work on a new application
  - Must already have existing data! No data collection!
- Topics need to be pre-approved
  - Details about the procedure TBA

Why study machine learning?



The machine learning trend 2015

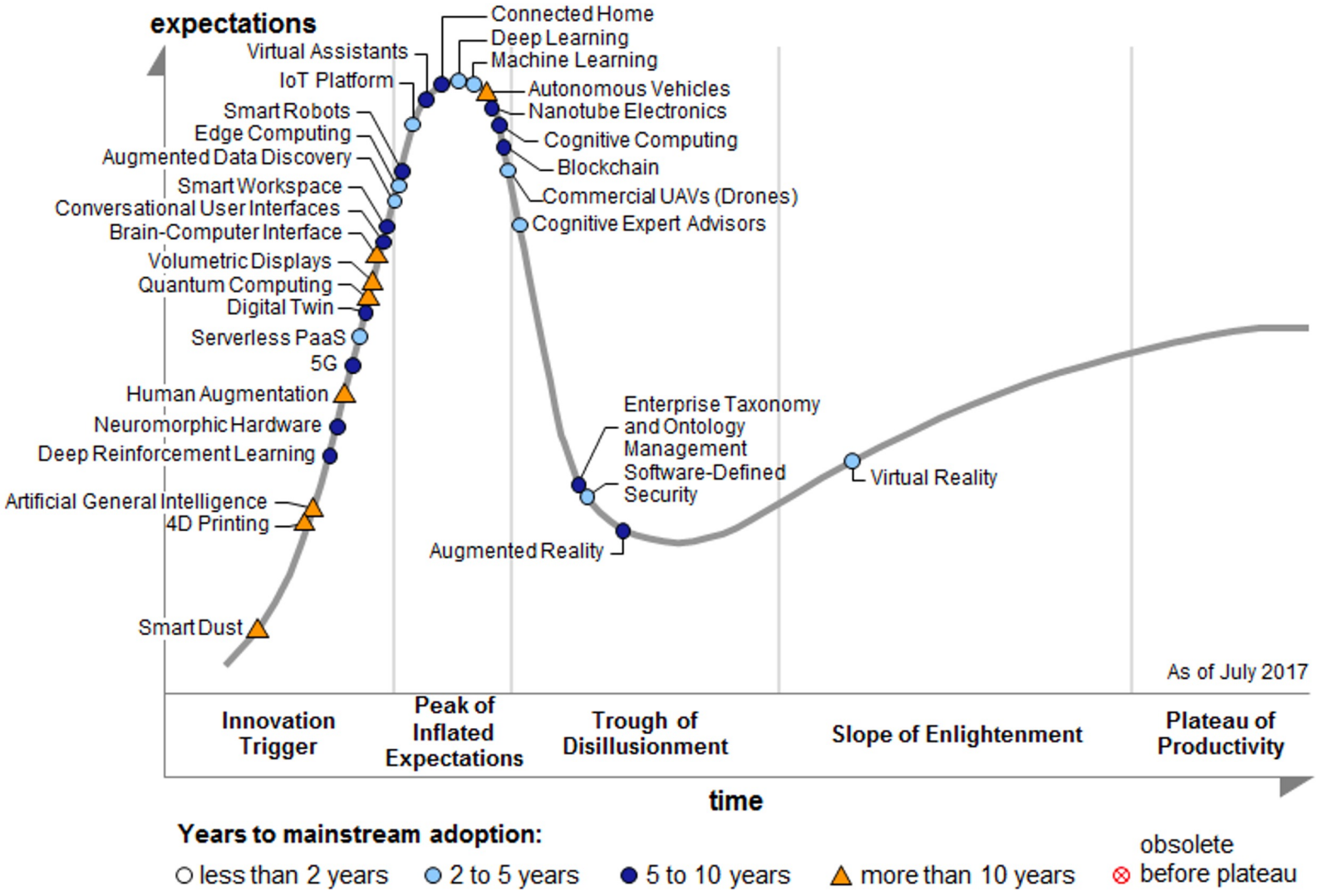
http://www.gartner.com/newsroom/id/3114217

The machine learning trend 2016

Source: Gartner (July 2016)

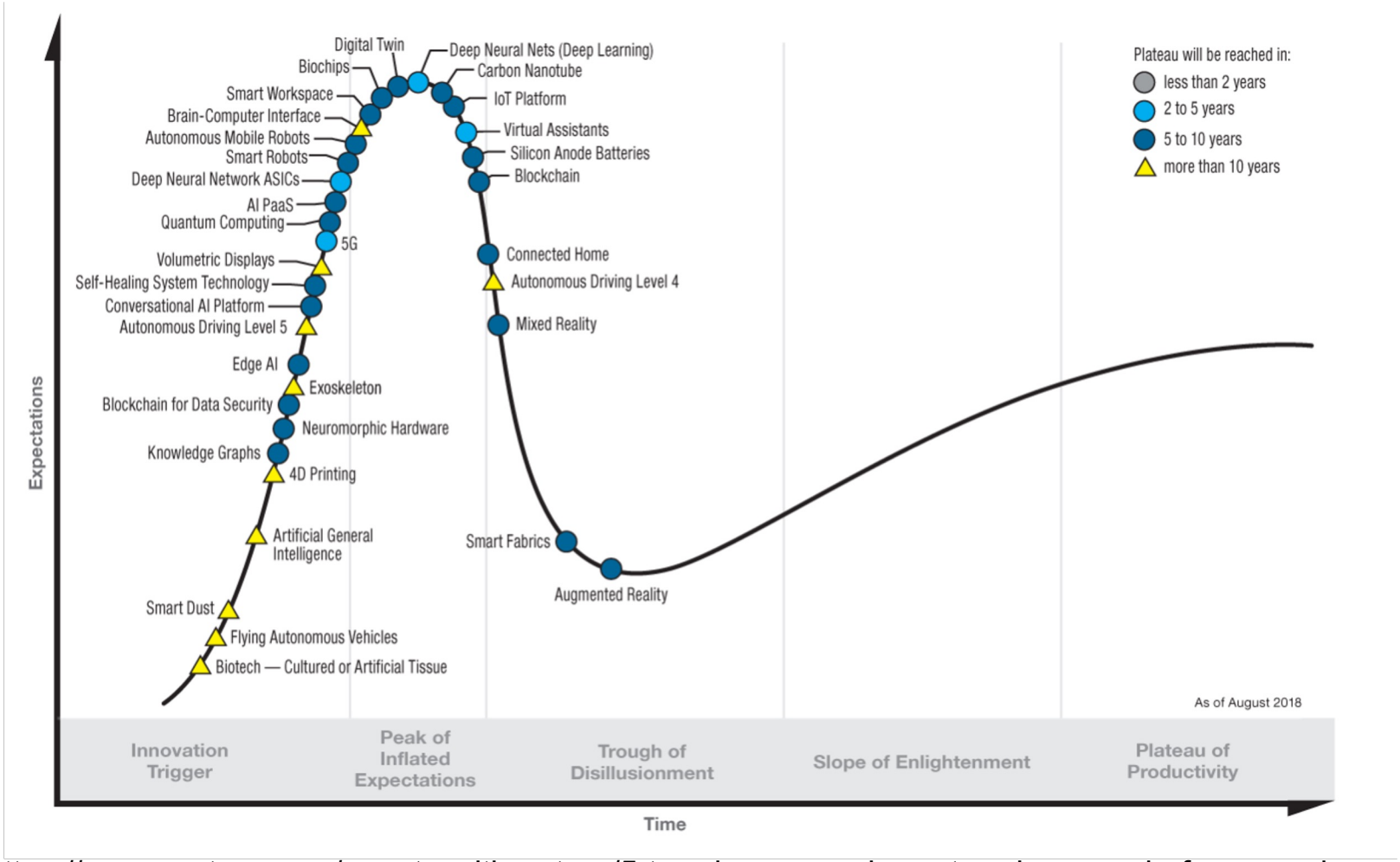
http://www.gartner.com/newsroom/id/3412017

The machine learning trend 2017



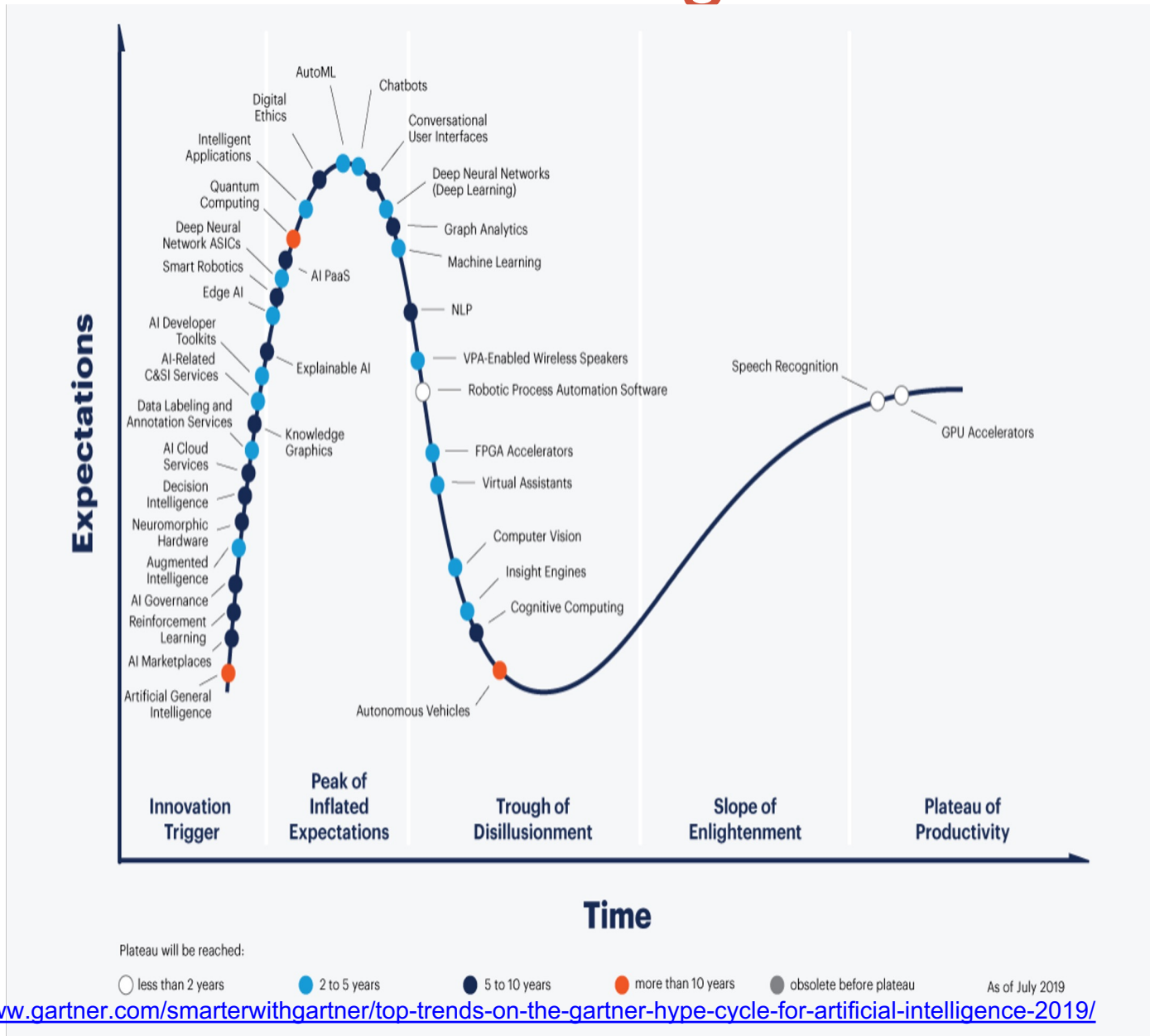
http://www.gartner.com/smarterwithgartner/top-trends-in-the-gartner-hype-cycle-for-emerging-technologies-2017/

The machine learning trend 2018



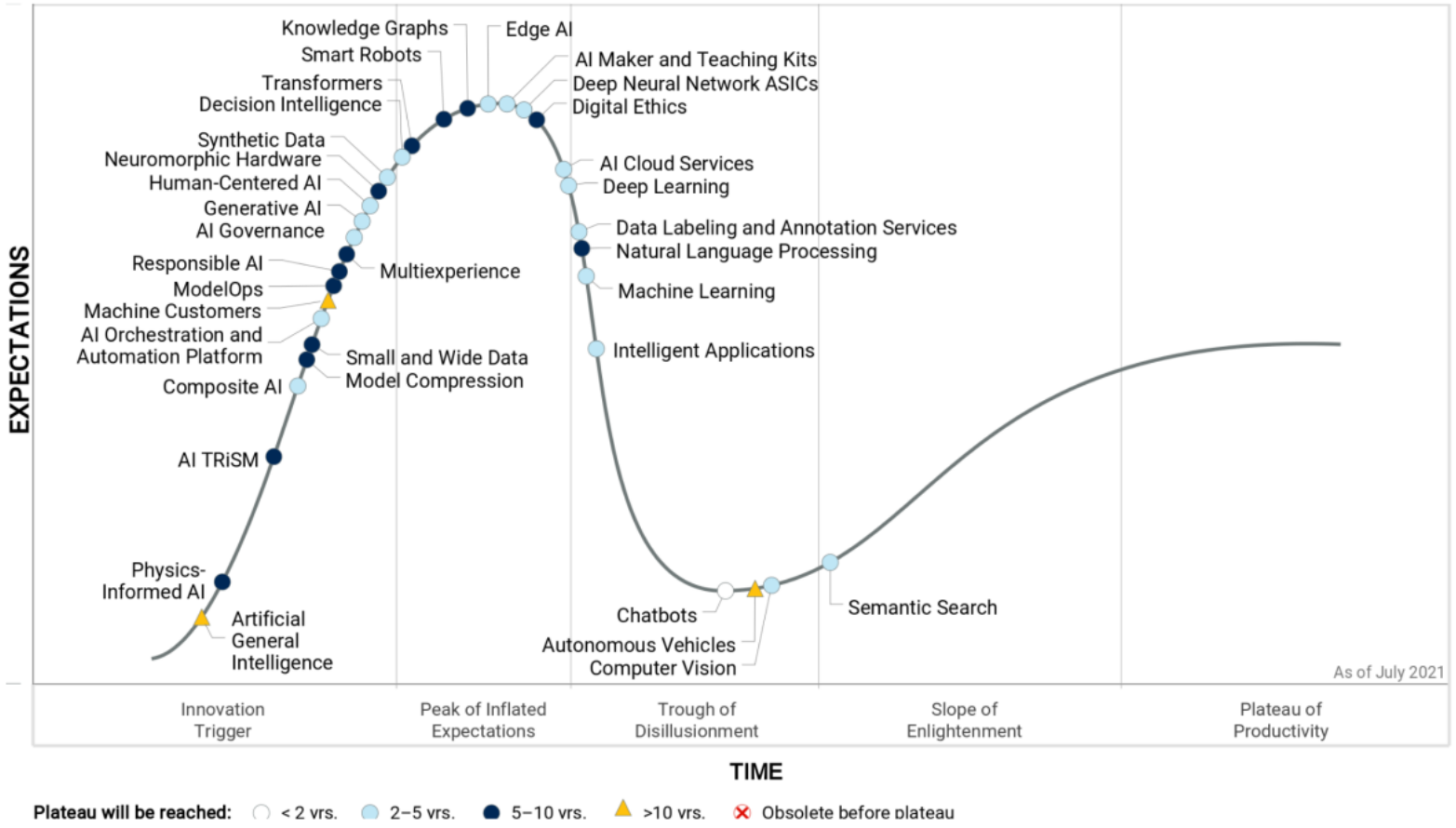
https://www.gartner.com/smarterwithgartner/5-trends-emerge-in-gartner-hype-cycle-for-emerging-technologies-2018/

The machine learning trend 2019



https://www.gartner.com/smarterwithgartner/top-trends-on-the-gartner-hype-cycle-for-artificial-intelligence-2019/

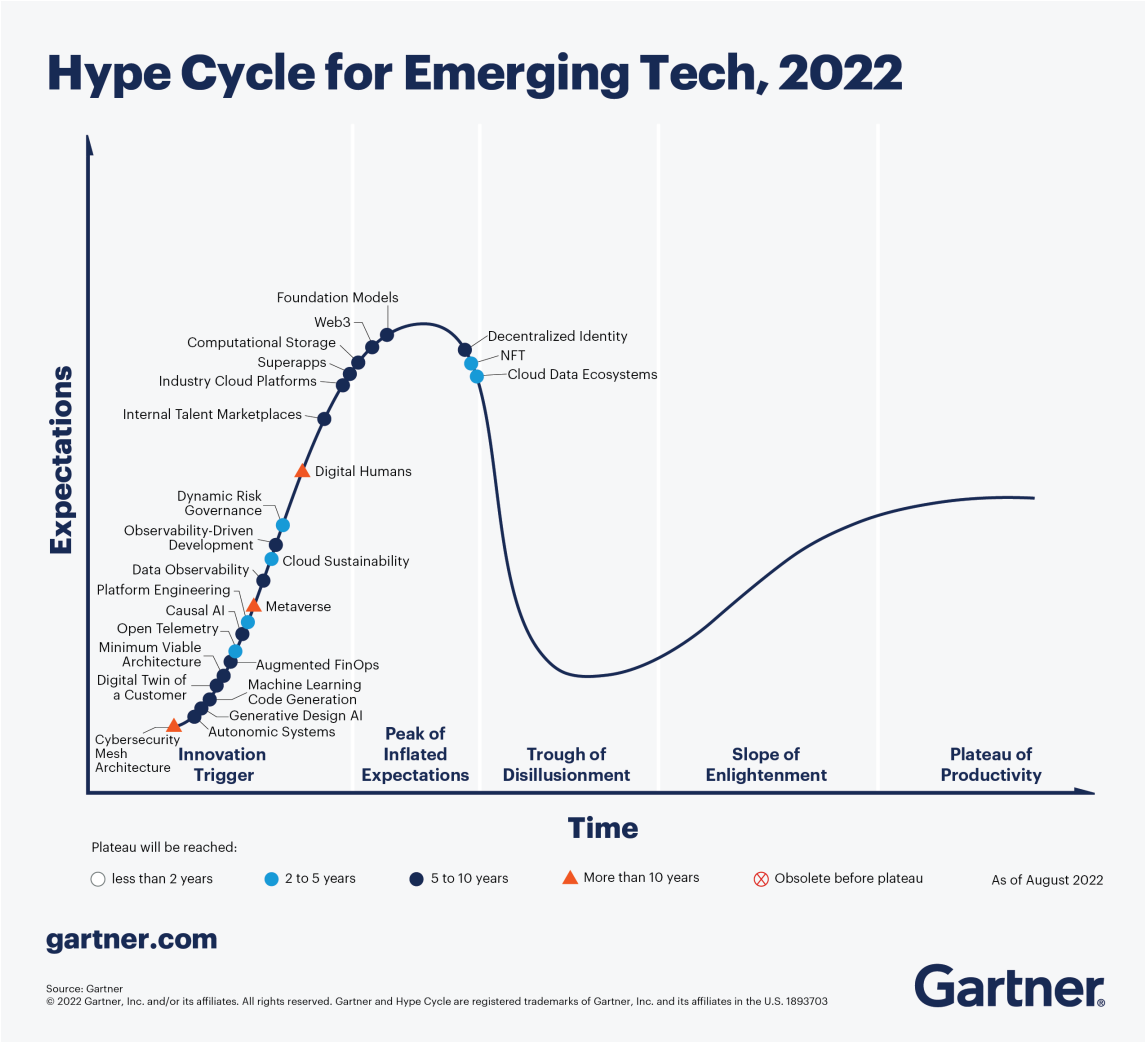
The machine learning trend 2021



Gartner

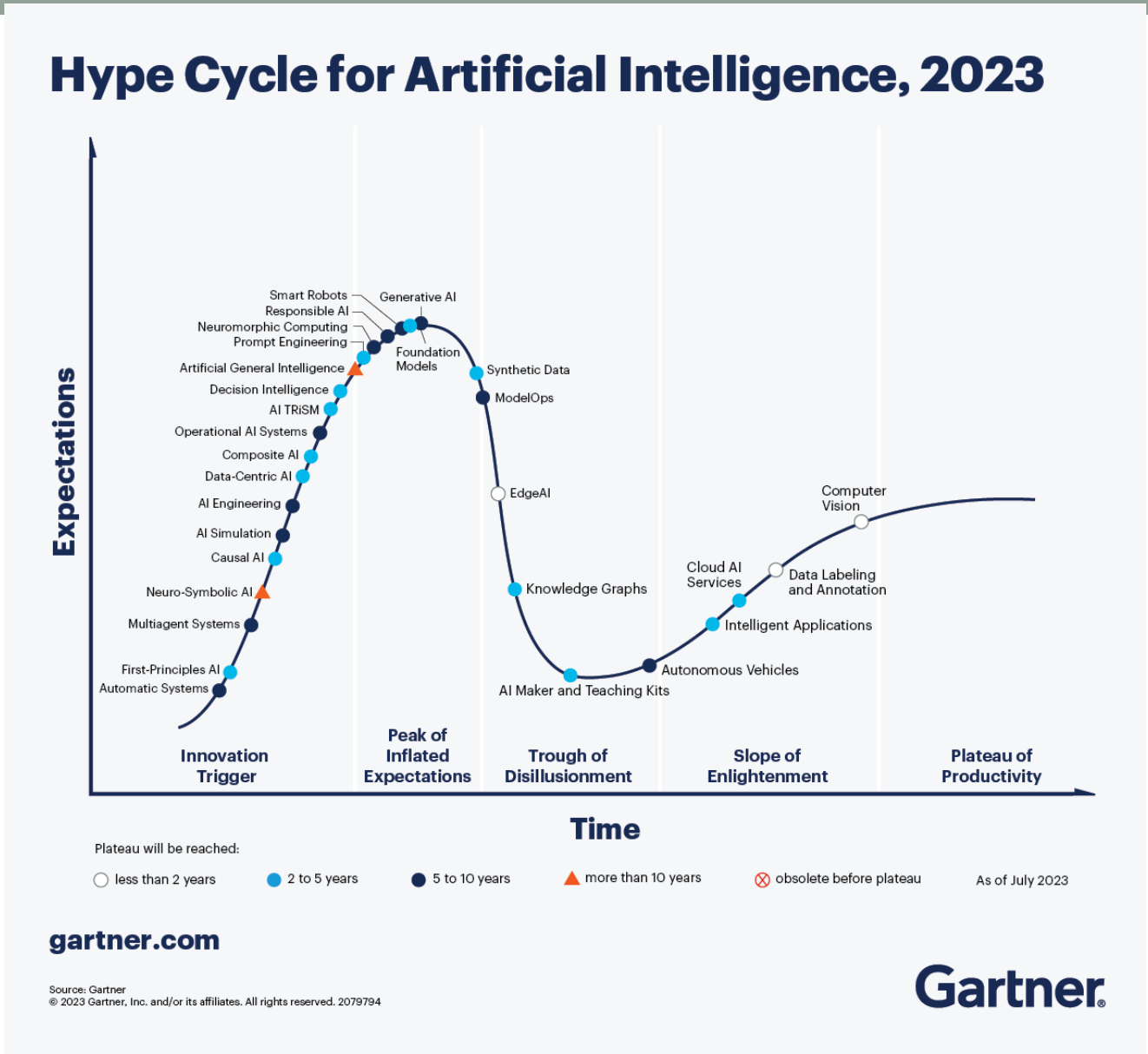
https://www.gartner.com/en/articles/the-4-trends-that-prevail-on-the-gartner-hype-cycle-for-ai-2021

The machine learning trend 2022



https://www.gartner.com/en/articles/what-s-new-in-the-2022-gartner-hype-cycle-for-emerging-technologies

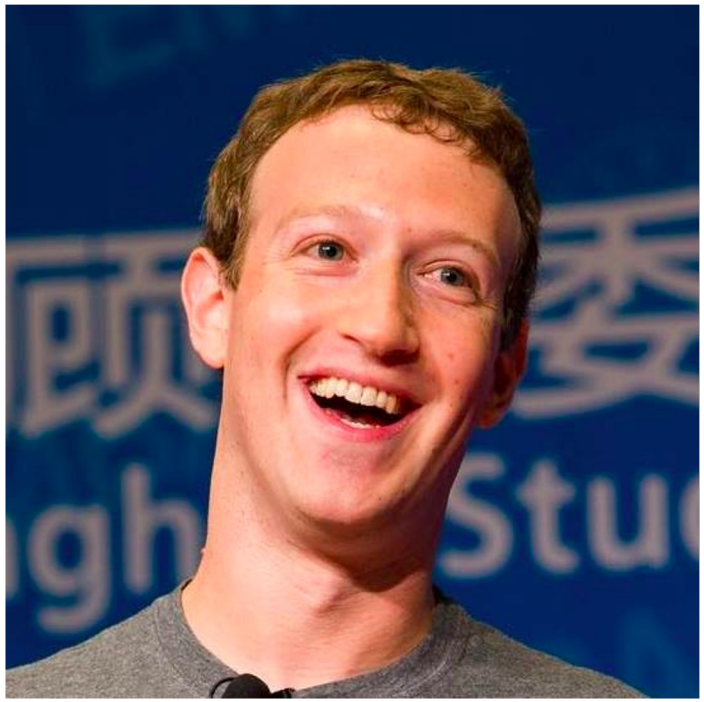
2023



https://www.gartner.com/en/articles/what-s-new-in-artificial-intelligence-from-the-2023-gartner-hype-cycle

- “If I were to guess like what our biggest existential threat is, it’s probably that. So we need to be very careful with the artificial intelligence. There should be some regulatory oversight maybe at the national and international level, just to make sure that we don’t do something very foolish.”

- “I think people who are naysayers and try to drum up these doomsday scenarios — I just, I don’t understand it. It’s really negative and in some ways I actually think it is pretty irresponsible”



Darren Cunningham @dcunni · 6h

Zuckerberg blasts @elonmusk warnings against artificial intelligence as 'pretty irresponsible' bizjournals.com/sanjose/news/2… @svbizjournal #ai

Facebook CEO Mark Zuckerberg blasts Tesla CEO Elon Musk's warn...

"People who are naysayers and try to drum up these doomsday scenarios — I just, I don't understand it," the Facebook CEO said. "It's really negative

bizjournals.com

30 296 566

Elon Musk ✓

@elonmusk

Following

Replying to @dcunni @SVbizjournal

I've talked to Mark about this. His understanding of the subject is limited.

8:07 AM - 25 Jul 2017

Poll

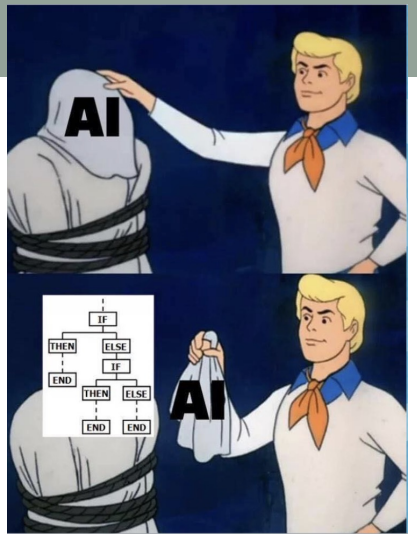
What is Pattern Recognition?

- "Pattern recognition is a branch of machine learning that focuses on the recognition of patterns and regularities in data, although it is in some cases considered to be nearly synonymous with machine learning."

wikipedia

- What about
  - AI
  - Data mining
  - Knowledge Discovery in Databases (KDD)
  - Statistics
  - Data science

What is AI?



- Classical definition
  - A system that appears intelligent
- Populace definition

- Probably what the field means right now
  - ML
    - narrow AI
      - Specialized

https://techsauce.co/pr-news/tcas-use-cloud-computing-and-ai-for-admission

Artificial General Intelligence (AGI)

- "hypothetical ability of an intelligent agent to understand or learn any intellectual task that a human being can."

Wikipedia

Can continue to learn new skills on its own.

Probably not of much interest besides philosophical debates

Works done in baby steps

ML vs PR vs DM vs KDD

- “The short answer is: None. They are … concerned with the same question: how do we learn from data?”

  Larry Wasserman – CMU Professor

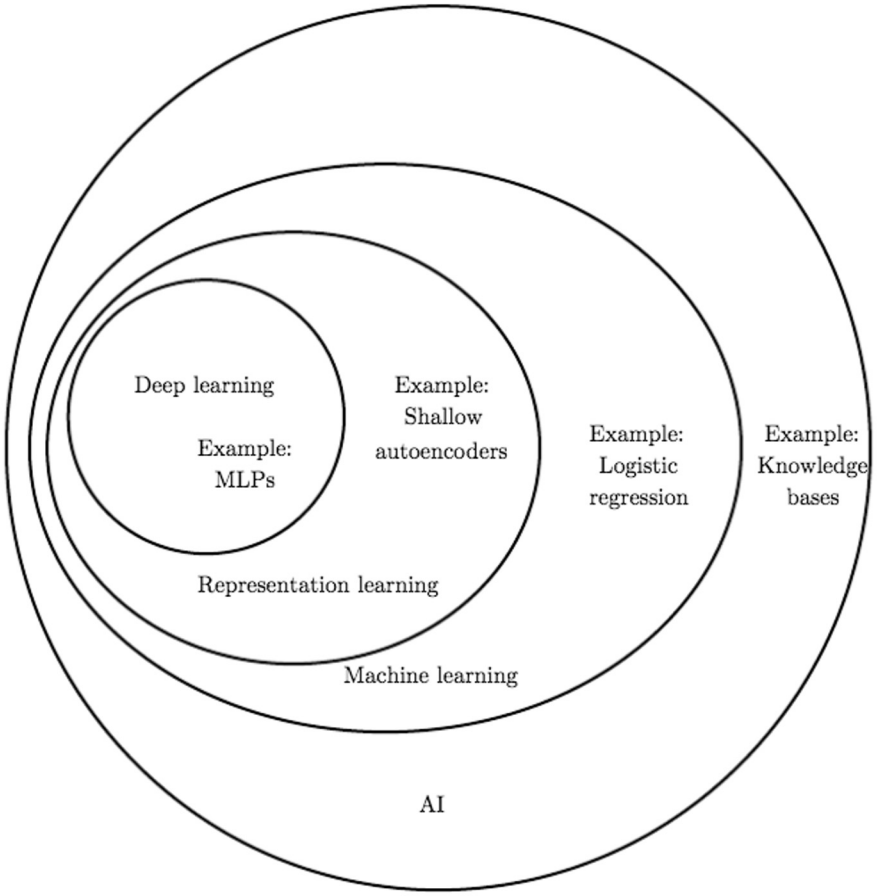
- Nearly identical tools and subject matter

History

- Pattern Recognition started from the engineering community (mainly Electrical Engineering and Computer Vision)
- Machine learning comes out of AI and mostly considered a Computer Science subject
- Data mining starts from the database community

Distinguishing things

- DM – Data warehouse, ETL
- AI – search, swarm intelligence
- PR – Signal processing (feature engineering)



http://www.deeplearningbook.org/

Different terminologies

http://statweb.stanford.edu/~tibs/stat315a/glossary.pdf

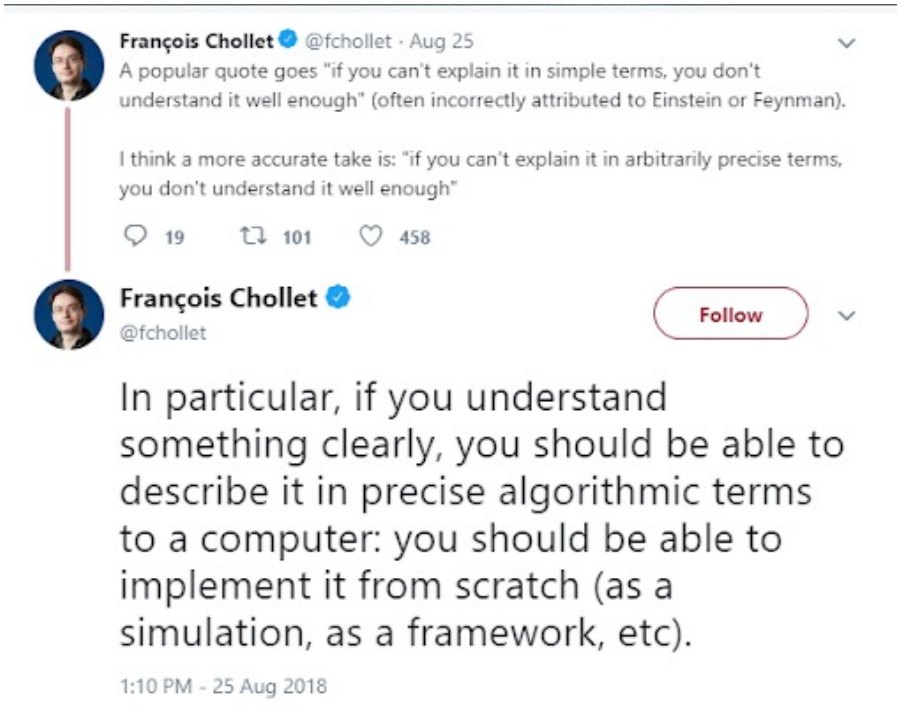
Machine learning	Statistics
network, graphs	model
weights	parameters
learning	fitting
generalization	test set performance
supervised learning	regression/classification
unsupervised learning	density estimation, clustering
large grant = $1,000,000	large grant= $50,000

Merging communities and fields

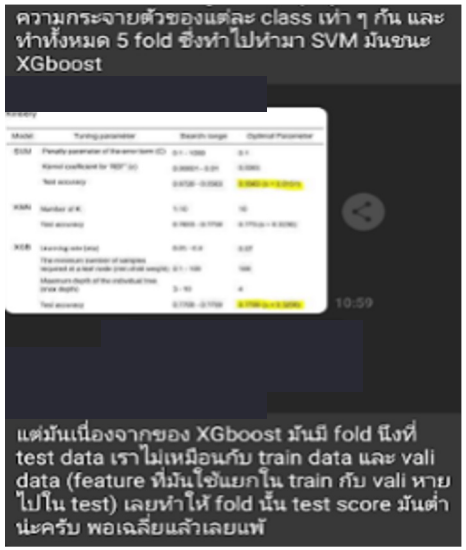
- With the advent of Deep learning the fields are merging and the differences are becoming unclear

Course philosophy

- Going beyond the black box
- In this course you will
  - Understand models on a deeper level
  - Implement stuff from scratch



The danger zone



อาจารย์ครับ ผมมีประโยคตัดเป็นคำ ใส่เป็น word bag ที่มีขนาดเท่ากับ vocab
ใส่เป็นความถี่ของคำในประโยค
แล้วผ่าน TF-IDF แล้วเข้า Multinomial NB มันพังกว่าไม่ผ่าน TF-IDF ครับ ผมสงสัยว่าเป็นเพราะอะไร ผมรู้แค่ว่าค่า prob มันเปลี่ยนจริงๆ มันน่าจะดี เพราะผมทำ TF-IDF บน SVM แล้วผลมันดีครับ

ผมใช้ laplace ด้วยครับ แต่ผลไม่น่าพังขนาดนั้น alpha=1 ครับ

อาจารย์ครับ ควรแบ่ง data ไงดีครับตอนทำ neural net
ผมมี data อยู่ของช่วงวันที่ 9-16 ครับ
คือ ตอนที่ทำ linear กับ pca ผมใช้ train เป็นช่วงวันที่ 9 - 13 ครับ ส่วน test ผมใช้ช่วง 14-15
ตอนทำ neural net ถ้าผมใช้เป็นวันที่ 16 วันเดียวพอไหมครับ หรือควรแบ่ง data ใหม่ครับ
ตอนนี้ training set ผมมีประมาณ 360000 ครับ ส่วน test set มีประมาณ 150000 ครับ

แล้วจึงนำภาพ3D 11440 ไป Train เข้า CNN แล้ว Classify ว่า เป็น 1(Depression Group) หรือ 0(Control Group)

ซึ่งผมก็พยายามปรับพารามิเตอร์ต่าง ๆ ก็ได้ Acc สูงสุดที่ 65%

ผมเลยมามองว่าจะใช้ GRU ด้วย

ปล.การ Train ครั้งก่อน ทำการ shuffle ตาต้าเรียบร้อยแล้วนะครับ

ตรงนี้ GRU จะต้องรับ input ยังไงเหรอครับ คือ รับเป็น 11440 โดยไม่ shuffle แล้วตั้ง batch=143 เพื่อให้มันมองเป็น คน ๆ ไปเหรอครับ ?

หรือผมต้องมี 1 timesteps มาคั่นระหว่าง sample เพื่อให้มันแยกได้

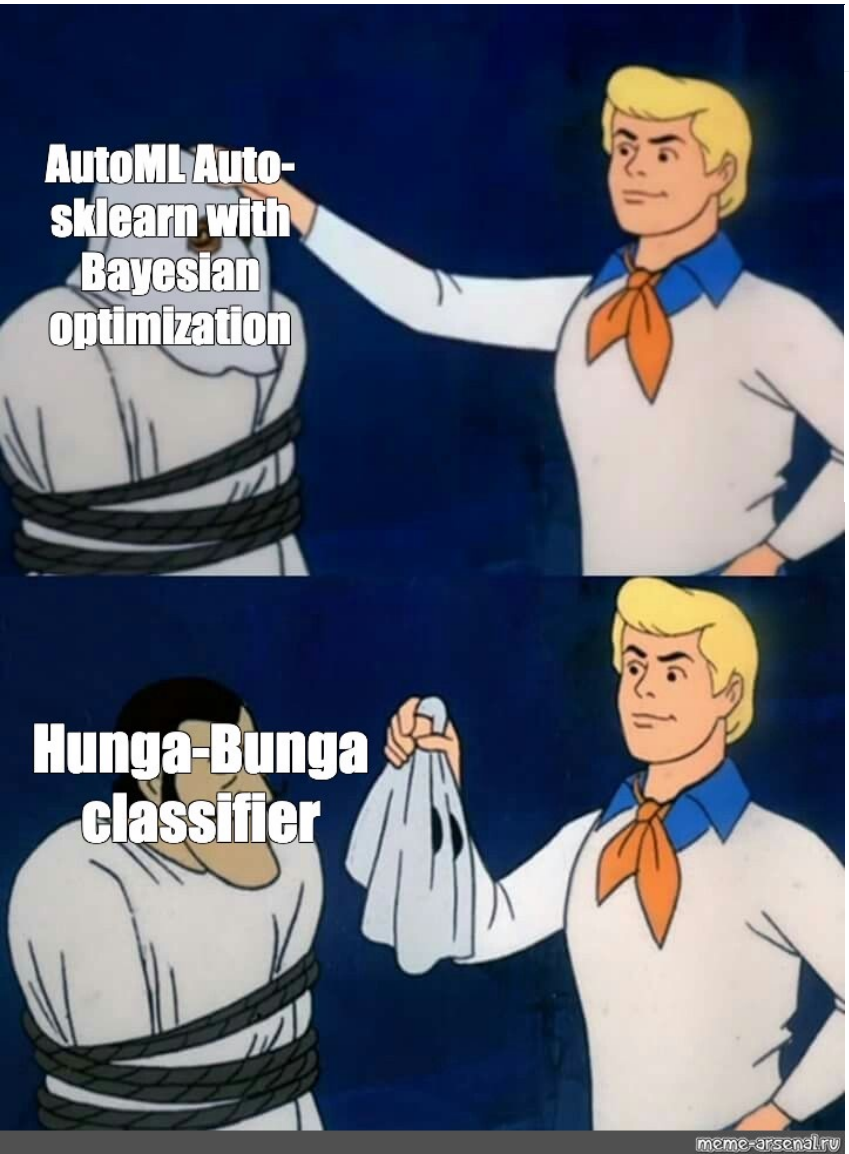
อย่างถ้าเป็นใน HW1 ของ NLP มันจะให้เราทำ model ที่บอกว่าขึ้นประโยคใหม่หรือยัง แต่ของผมจะเป็นแบบนอกจากบอกว่าขึ้นประโยคใหม่แล้ว ยังต้องบอกด้วยว่าเป็น Noun, Verb, หรือ Adj. ประมาณนั้นอะครับ เพราะของผมคือต้องบอกว่าเป็น control หรือ depression โดยใช้ทั้งหมด 143 timesteps

อีกอย่างที่สงสัยคือตอน test น่ะครับ อย่างงี้หมายความว่าถ้าผม test แต่ละ sample ด้วย 143 timestep input ผมก็จะได้ 143 outputs ใช่มั้ยครับ แต่ที่ผมต้องการคือ output ตัวเดียวที่บอกว่าเป็น Depression หรือ Control เพียงค่าเดียว

Driving a car analogy

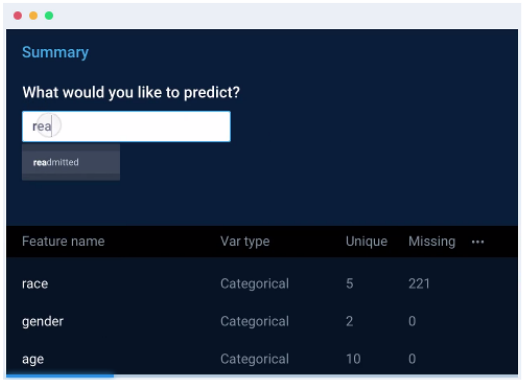
- Just drive without knowing where you are going
- Getting there vs getting there effectively
- Putting the wrong fuel into the car

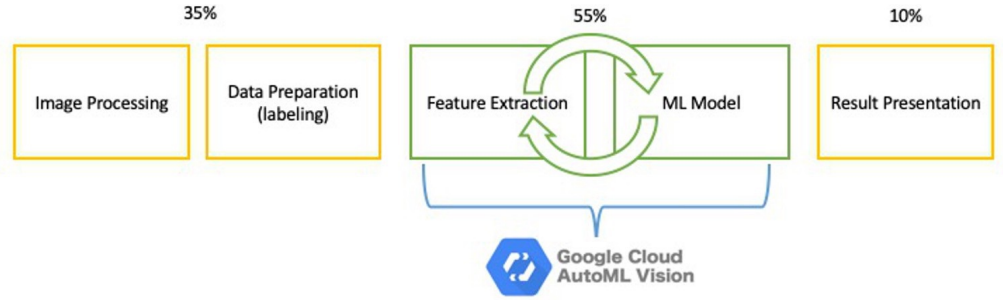
Be better than autoML



DataRobot PRODUCT SOLUTIONS EDUCATION ABOUT WE'RE HIRING! CONTACT US

1. Upload your data
2. Select the target variable
3. Build 100s of models in one click
4. Explore top models and get insights
5. Deploy best model and make predictions





AutoML replaces this stage saving 55% of effort and providing better accuracy

https://towardsdatascience.com/ocr-for-scanned-numbers-using-googles-automl-vision-29d193070c64

Types of machine learning

1. Supervised learning
2. Unsupervised learning
3. Reinforcement learning

0. Pre-machine learning: rule-base

Pre-machine learning: 7-segment display

- **Input:** 7 binary values (0,1) forming a display
- Given **x** = (A, B, C, D, E, F, G)
- **Output:** y, either 0, 1, …, 9 or not a number
- **Task:** write a program (a function F) that maps **x** to y; F(**x**) = y

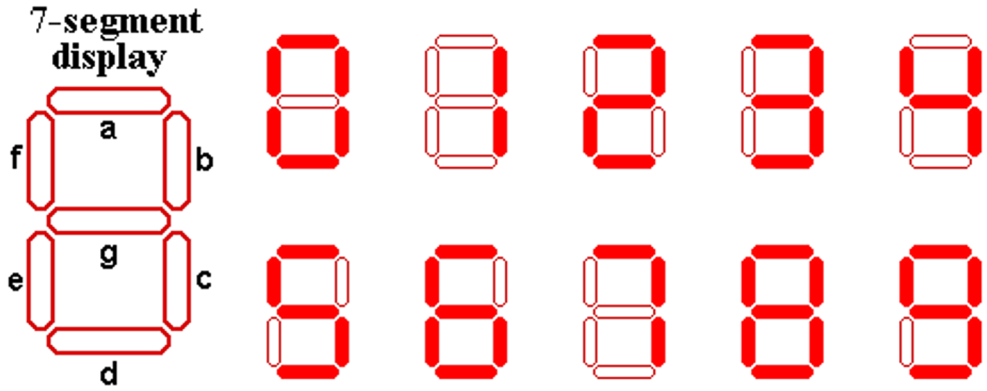


Image from http://www.physics.udel.edu/~watson/scen103/colloq2000/7-seg.html

Mapping function

Y ⟵ X ⟶

Count	a	b	c	d	e	f	g
0	●	●	●	●	●	●	
1		●	●				
2	●	●		●	●		●
3	●	●	●	●			●
4		●	●			●	●
5	●		●	●		●	●
6	●		●	●	●	●	●
7	●	●	●				
8	●	●	●	●	●	●	●
9	●	●	●	●		●	●

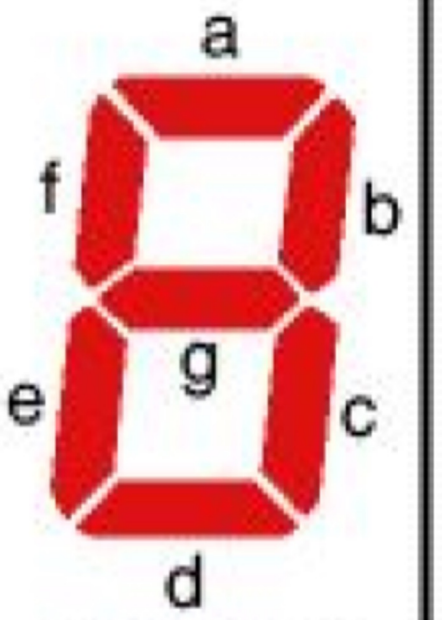
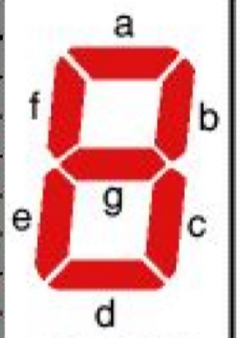


Image from: http://www.instructables.com/id/DIY-7-Segment-Display/

Mapping function

Count	a	b	c	d	e	f	g
0	●	●	●	●	●	●	
1		●	●				
2	●	●		●	●		●
3	●	●	●	●			●
4		●	●			●	●
5	●		●	●		●	●
6	●		●	●	●	●	●
7	●	●	●				
8	●	●	●	●	●	●	●
9	●	●	●	●		●	●



7-segment display

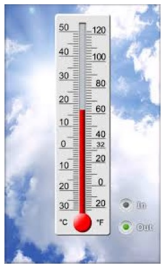
- IF A==1 && B==1 && C==1 && D==1 && E==1 && F==1 && G==0, THEN output(0).
- IF B==1 && C==1, THEN output(1)
- …..
- OTHERWISE, output(“not number”)

F(x)

Learning from data

- Machine learning requires identifying the same ingredients
  - Input, Output, Task

Real world observations



This is the hardest part of data science and the last part to be replaced by machines.

https://cloud.google.com/blog/products/gcp/how-a-japanese-cucumber-farmer-is-using-deep-learning-and-tensorflow

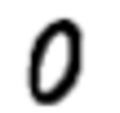
An example

- Handwritten digit recognition
- Input: **x** = 28 x 28 pixel image
- Output: y = digit 0 to 9
- Task: find F(**x**) such that y ≈ F(**x**)

Goal of machine learning is to find the best F(**x**) automatically from data

Supervised learning

- Learn a classifier F from a training set (input-output pairs)
  - $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), (\mathbf{x}_3, y_3), \ldots, (\mathbf{x}_n, y_n)\}$

x	y
	0
	1
	2

Need a training set for training.
Training = finding (optimizing) a good function f

Labeling (i.e., assigning y for each **x** in the training set) is typically done manually.

Types of machine learning

1. Supervised learning

   Learn a model F from pairs of (**x**,y)

2. Unsupervised learning

   Discover the hidden structure in unlabeled data **x** (no y)

3. Reinforcement learning

   Train an agent to take appropriate actions in an environment by maximizing rewards

Typical workflow of machine learning

1. Feature extraction (getting the **x**)
2. Modeling
   - Training (getting the function F)
3. Evaluation
   - Metrics (defining what's the best function F)
   - Testing (getting the y for unseen inputs)

Typical workflow of machine learning

- The typical workflow

Real world observations

Squire Trelawney, Dr. Livesey, and the rest of those gentlemen having asked me to write down the whole particulars about Treasure Island, from the beginning to the end, keeping nothing back but the bearings of the island, and that only because there is still treasure not yet lifted, I take up my pen in the year of grace 17— and go back to the time when my father kept the Admiral Benbow inn and the brown old seaman with the sabre cut first took up his lodging under our roof.

I remember him as if it were yesterday, as he came plodding to the inn door, his sea chest following behind him in a hand-barrow; a tall, strong, heavy, nut-brown man, his tarry pigtail falling over the shoulder of his soiled blue coat, his hands ragged and scarred, with black, broken nails, and the sabre cut across one cheek, a dirty, livid white. I remember him looking round the cove and whistling to himself as he did so, and then breaking out in that old sea-song that he sang so often afterwards:

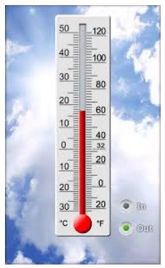
"Fifteen men on the dead man's chest—Yo-ho-ho, and a bottle of rum!" In the high, old tottering voice that seemed to have been tuned and broken at the capstan bars. Then he rapped on the door with a bit of stick like a handspike that he carried, and when my father appeared, called roughly for a glass of rum. This, when it was brought to him, he drank slowly, like a connoisseur, lingering on the taste and still looking about him at the cliffs and up at our signboard.

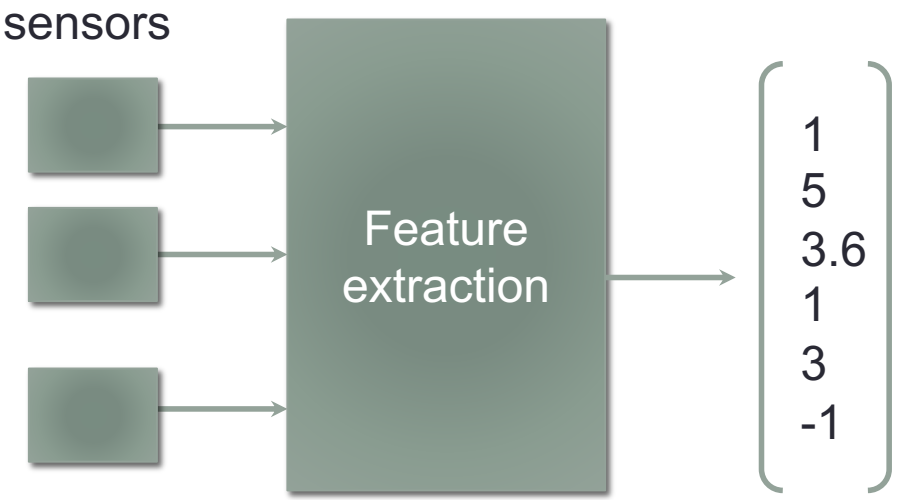
"This is a handy cove," says he at length; "and a pleasant sittyated grog-shop. Much company, mate?"

My father told him no, very little company, the more was the pity.

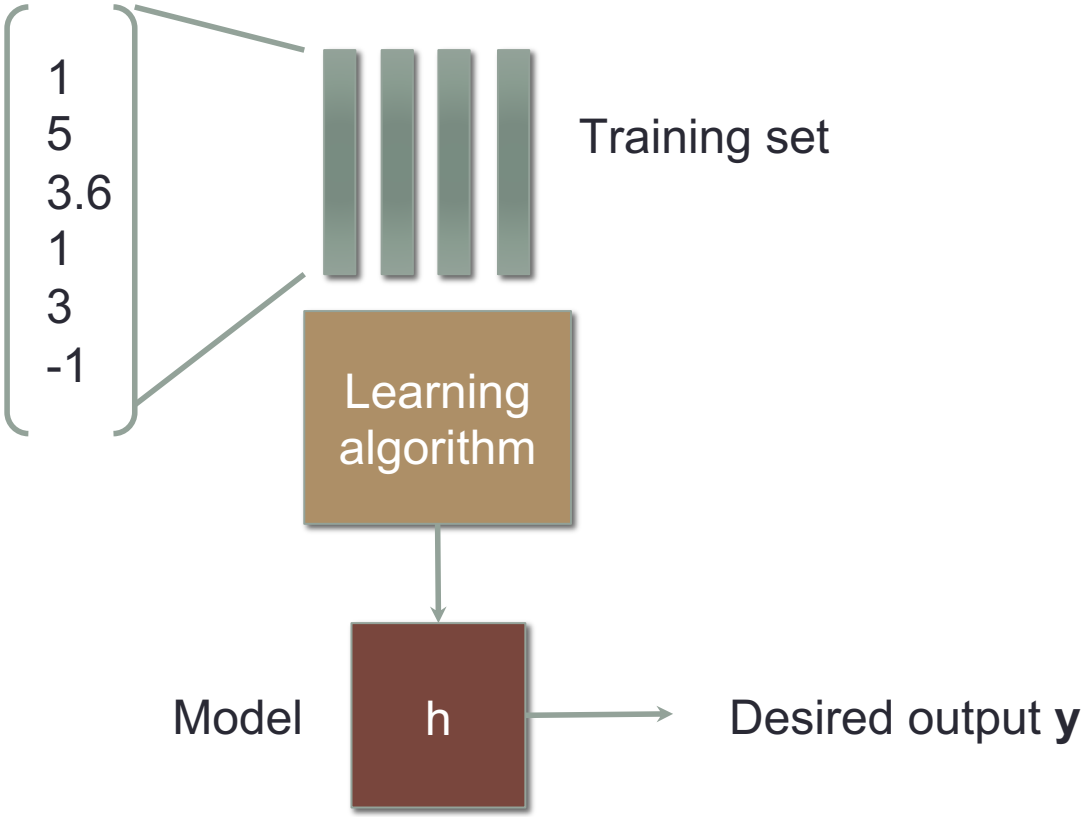
"Well, then," said he, "this is the berth for me. Here you, matey," he cried to the man who trundled the barrow; "bring up alongside and help up my chest. I'll stay here a bit," he continued. "I'm a plain man; rum and bacon and eggs is what I want, and that head up there for to watch ships off. What you mought call me? You mought call me captain. Oh, I see what you're at—there"; and he threw down three or four gold pieces on the threshold. "You can tell me when I've worked through that," says he, looking as fierce as a commander.

And indeed had as his clothes were and coarsely as he spoke, he had none of the appearance of a

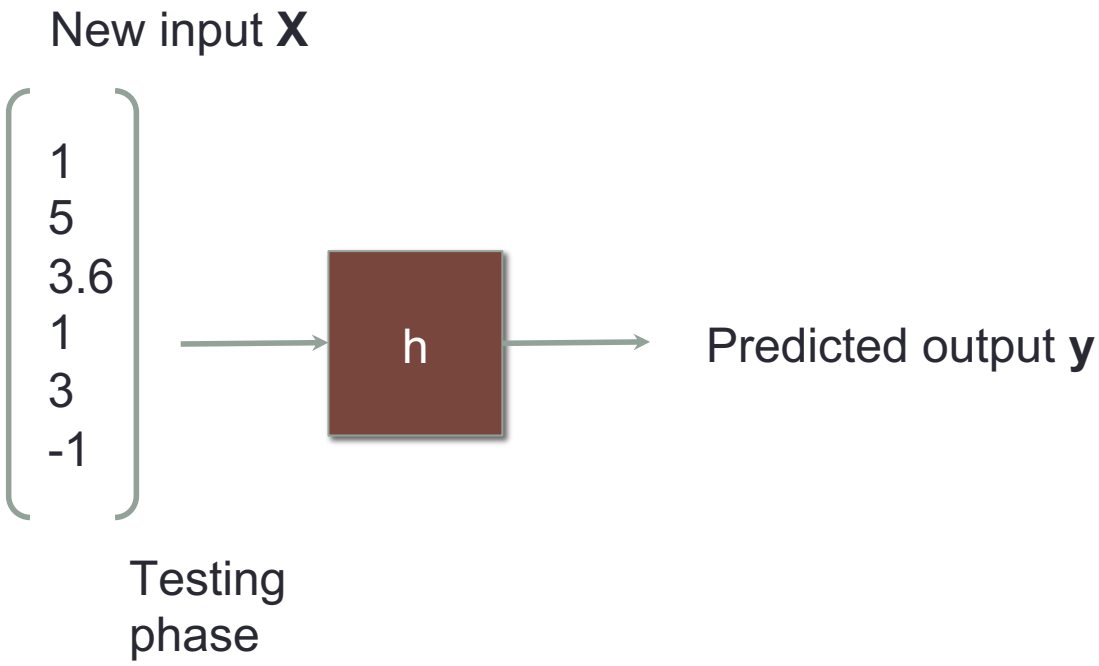




How do we learn from data?



How do we learn from data?

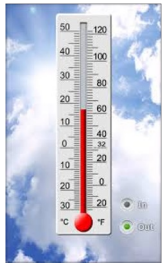


Feature extraction

- The process of extracting meaningful information related to the goal
- A distinctive characteristic or quality
- Example features

data1

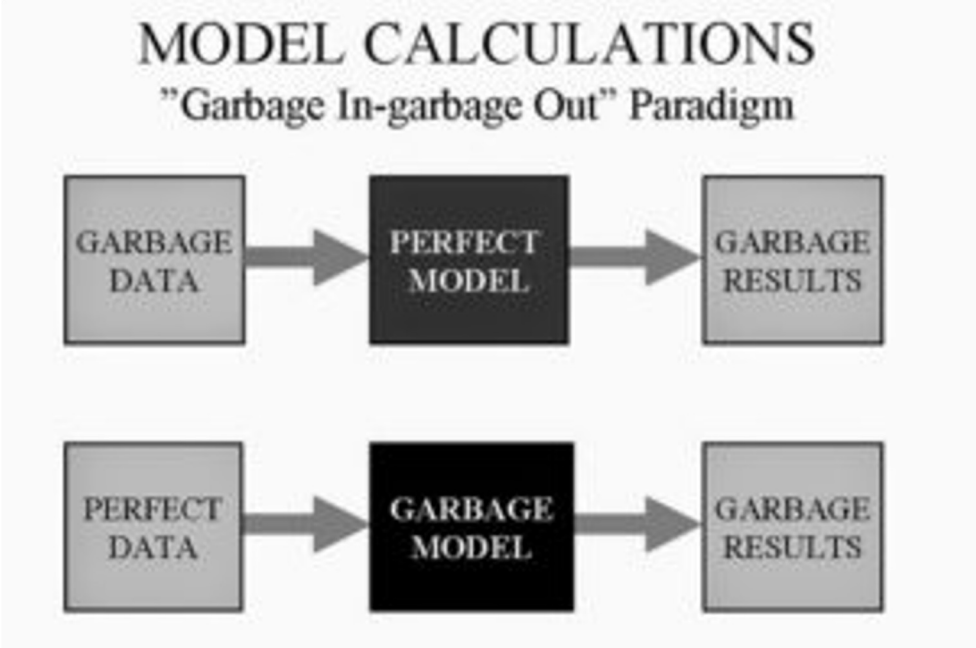
data2



data3

Garbage in Garbage out

- The machine is as intelligent as the data/features we put in
- “Garbage in, Garbage out”
- Data cleaning is often done to reduce unwanted things



https://precisionchiroco.com/garbage-in-garbage-out/

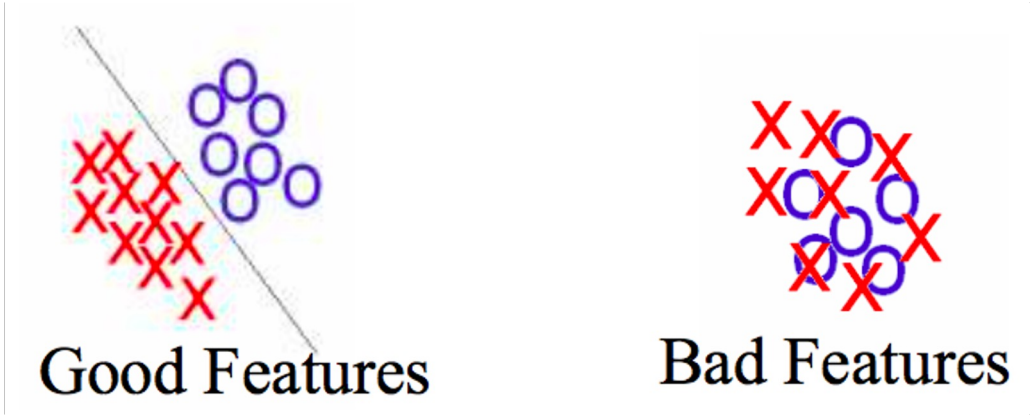
The need for data cleaning

However, good models should be able to handle some dirtiness!

https://www.linkedin.com/pulse/big-data-conundrum-garbage-out-other-challenges-business-platform

Feature properties

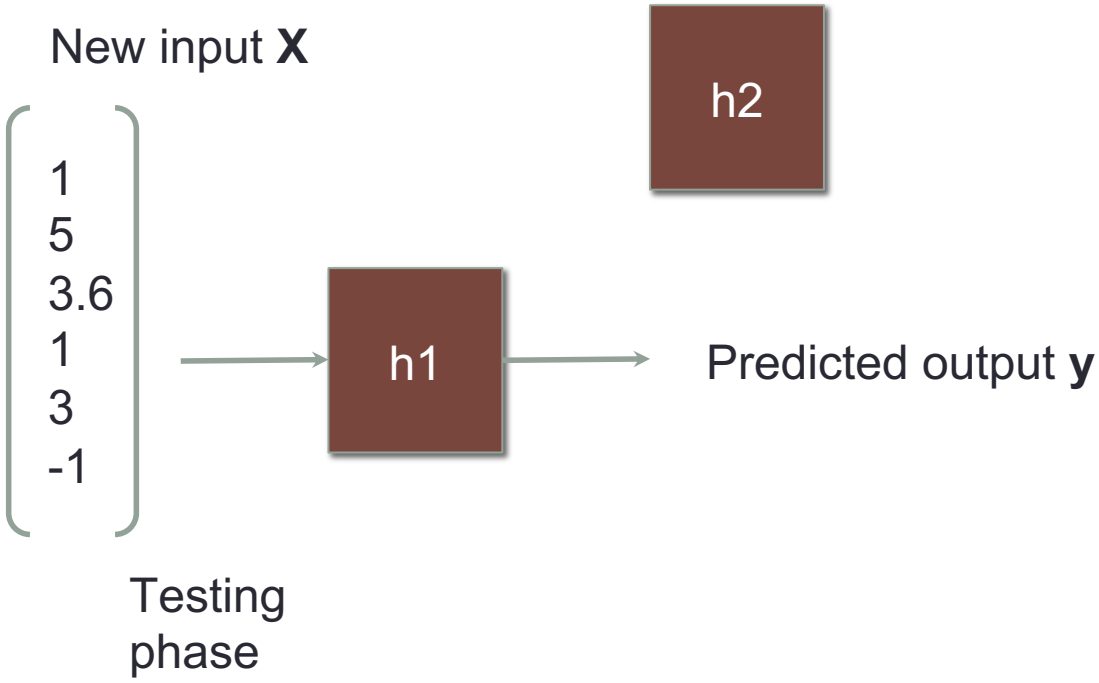
- The quality of the feature vector is related to its ability to discriminate samples from different classes



- In this course, we won't talk much about data/feature issues, since these are domain specific. However, they can be important than modeling.

Model evaluation

How to compare h1 and h2?



Metrics

- Compare the output of the models
  - Errors/failures, accuracy/success
- We want to quantify the error/accuracy of the models
- How would you measure the error/accuracy of the following

Ground truths

- We usually compare the model predicted answer with the correct answer.
- What if there is no real answer?
  - How would you rate machine translation?

ไปไหน

Model A: Where are you going?
Model B: Where to?

Designing a metric can be tricky, especially when it's subjective

Ground truths can be hard

Speed limits in the United ...
en.wikipedia.org

Speed limits in Japan - Wikip...
en.wikipedia.org

France lowers speed limit on roads ...
hurriyetdailynews.com

Speed limits in Mexico - Wiki...
en.wikipedia.org

Speed Limit 40 Sign | KirbyBuilt Products
kirbybuilt.com

10km Speed Limit Safety Sig...
officemax.co.nz

Miami Reducing Speed Li...
miamigov.com

Speed limits in Germany - ...
en.wikipedia.org

Speed limits in Thailand - ...
en.wikipedia.org

speed limit on Germany's Autobahn ...
thelocal.de

Speed limit could drop on stretch of ...
beaumontenterprise.com

Motorway speed limits to be reduced to ...
express.co.uk

Slides from https://github.com/goldmermaid/mlrs

Credit to Andrej Karpathy

Labelling

"Label lane lines"

Credit to Andrej Karpathy

Labelling issues

abel lane lines”

How do you annotate lane lines when they do this?

Credit to Andrej Karpathy

Credit to Andrej Karpathy

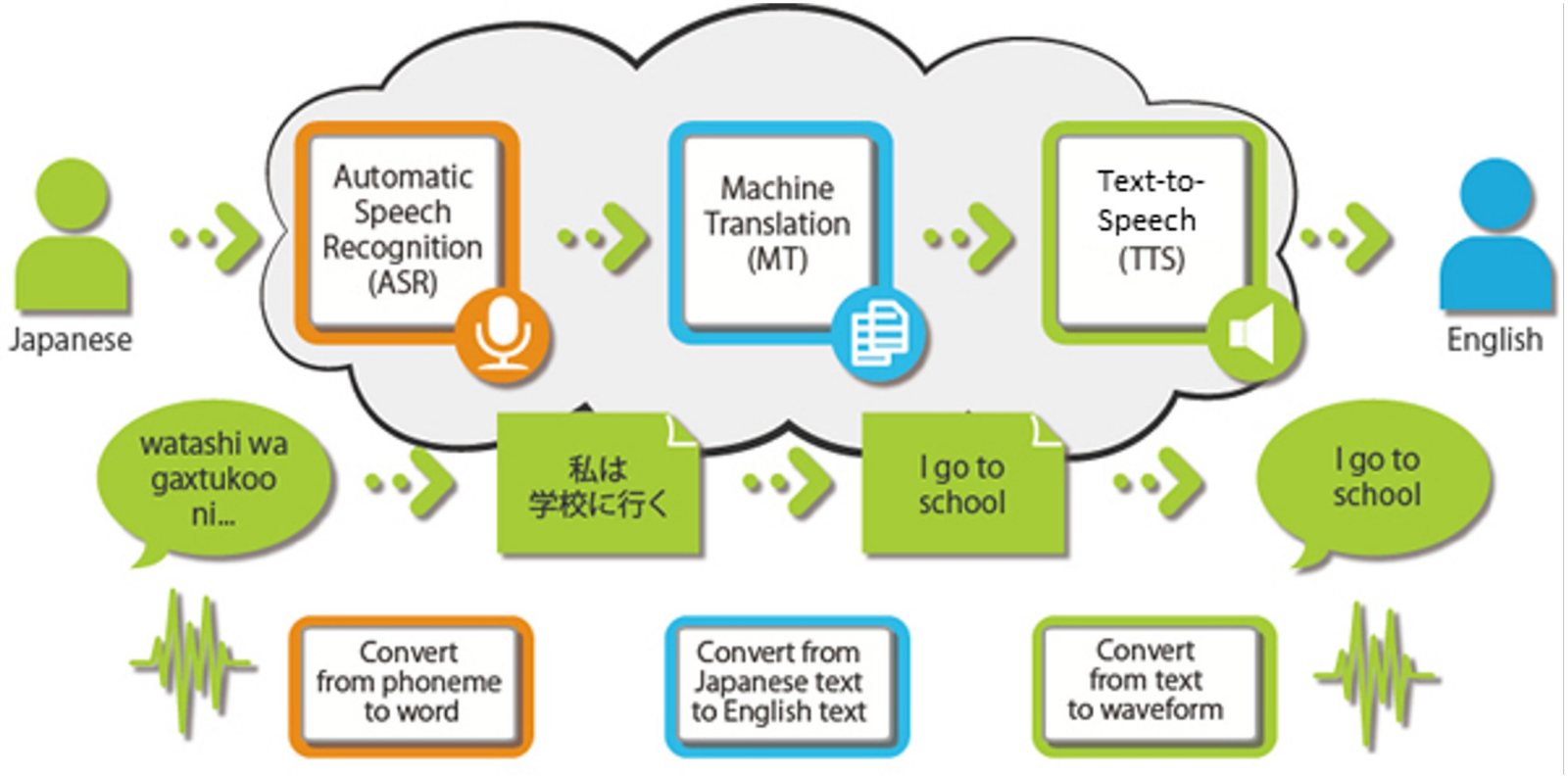
Metrics consideration 1

- Are there several metrics?

- Use the metric closest to your goal but never disregard other metrics.
  - May help identify possible improvements

Metrics consideration 2

- Are there sub-metrics?



http://www.ustar-consortium.com/qws/slot/u50227/research.html

Commonly used metrics

- Error rate
- Accuracy rate

- Precision
- True positive
- Recall
- False alarm
- F score

A detection problem

- Identify whether an event occur
- A yes/no question
- A binary classifier

Smoke detector

Hotdog detector

Evaluating a detection problem

- 4 possible scenarios

		Detector	
		Yes	No
Actual	Yes	True positive	False negative (Type II error)
	No	False Alarm (Type I error)	True negative

True positive + False negative = # of actual yes
False alarm + True negative = # of actual no

- False alarm and True positive carries all the information of the performance.

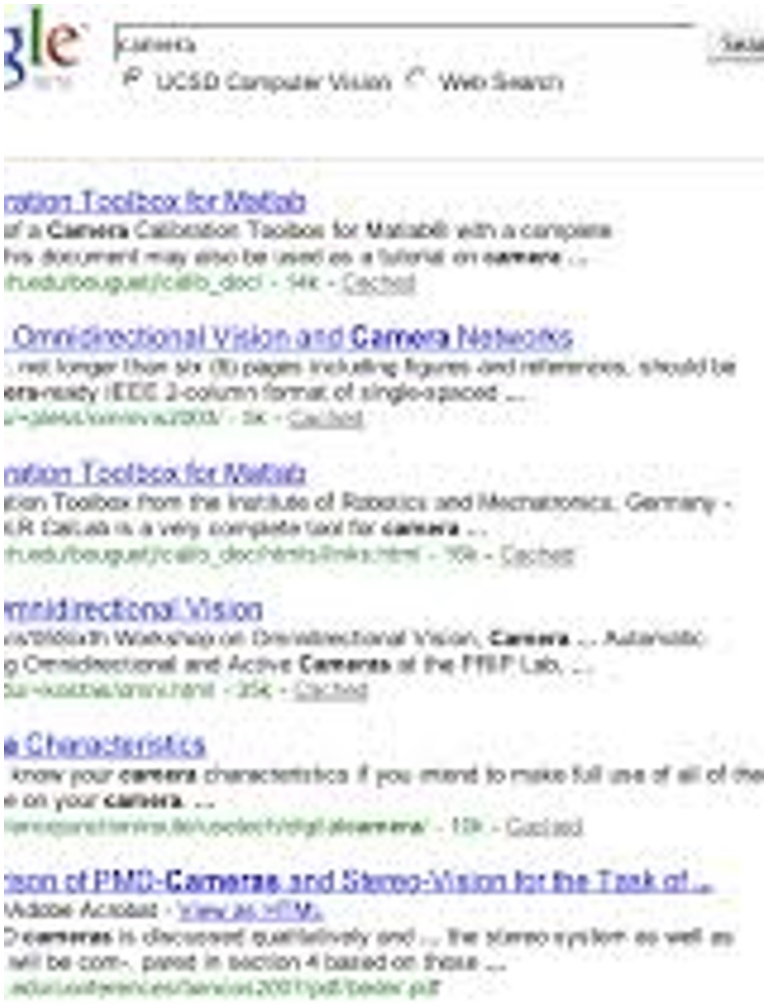
Definitions

Actual	Detector: Yes	Detector: No
Yes	True positive	False negative (Type II error)
No	False Alarm (Type I error)	True negative

- True positive rate (Recall, sensitivity)
  = # true positive / # of actual yes
- False positive rate (False alarm rate)
  = # false positive / # of actual no
- False negative rate (Miss rate)
  = # false negative / # of actual yes
- True negative rate (Specificity)
  = # true negative / # of actual no

- Precision = # true positive / # of predicted positive

Search engine example



A recall of 50% means?

A precision of 50% means?

Recall/precision

- When do you want high recall?
- When do you want high precision?

- Initial screening for cancer
- Face recognition system for authentication
- Detecting possible suicidal postings on social media
- COVID screening: ATK vs PCR

Usually there's a trade off between precision and recall. We will revisit this later

Let's consider a case

- A: no rain predictor has 97% accuracy
  - Always say no rain.

		Detector	
		Rain	No rain
Actual	Rain	0	1
	No rain	0	30

- Accuracy might not be a good metric for biased data
- A good model should be better than stupid baselines

Definitions 2

- F score (F1 score, f-measure)

$$F_1 = 2 \cdot \frac{1}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}.$$

  - A single measure that combines both aspects
  - A harmonic mean between precision and recall (an average of rates)

Note that precision and recall says nothing about the true negative

Evaluating models

- We talked about the training set used to learn the model
- We use a different data set to test the accuracy/error of models – “test set”
- We can still compute the error and accuracy on the training set
- Training error vs Testing error
- We will discuss how we can use these to help guide us later

Other considerations when evaluating models

- Training time
- Testing time
- Memory requirement
- Parallelizability
- Latency

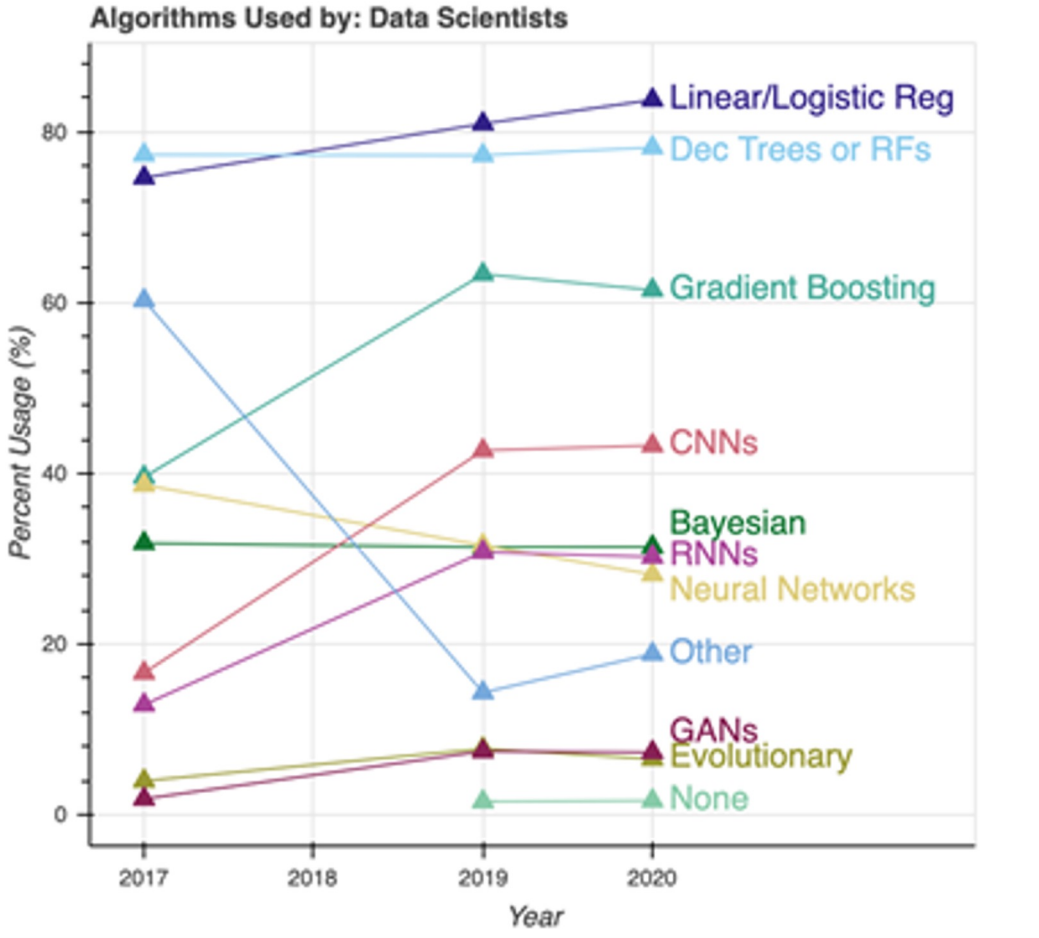
Course walkthrough

Traditional Machine learning

Deep learning

คาบเรียนที่	เนื้อหา	การบ้านและควิซ
1 - 11/1	Introduction, K-mean	เริ่มHW1
2 - 28/1	Regression, MLE & MAP	
3 - 25/1	Naive Bayes & GMM	ส่งHW1, Quiz 1, เริ่มHW2
4 - 1/2	GMM, EM, ELBO, Dimensionality reduction I (PCA)	
5 - 8/2	Dimensionality reduction II (LDA, RP) and visualization techniques (t-sne, UMAP, PHATE)	ส่งHW2, Quiz 2, เริ่มHW3
6 - 15/2	SVM, NN I	
7 - 22/2	NN II (CNN & Recurrent)	ส่งHW3, Quiz 3, เริ่มHW4
8 - 29/2	NN III (Architectures) & Pytorch demo	เริ่มHW5
9 - 7/3	Midterm week - No midterm for this class	
10 - 14/3	Transformers & Self-supervised I	ส่งHW4, Quiz 4
11 - 21/3	Self-supervised learning II	ส่งHW5, Quiz 5, ส่ง course project proposal, เริ่มHW6
11 - 28/3	Generative models I (GAN, VAE)	เริ่มHW7
12 - 4/4	Generative models II (Diffusion)	ส่งHW6, Quiz 6, เริ่ม HW8
13 - 11/4	Reinforcement Learning	ส่งHW7, Quiz 7
14 - 18/4	No regular class - meeting/progress presentation with project mentors	Course project progress
15 - 25/4	Tricks of the trade: machine learning in the real world + Guest	ส่งHW8, Quiz 8
Some time during final exam	Project presentation No final exam for this class	ส่งcourse project

Why anything else besides deep learning



https://medium.com/analytics-vidhya/ongoing-kaggle-survey-picks-the-topmost-data-science-trends-7c19ec7606a1

KNN and K-means clustering

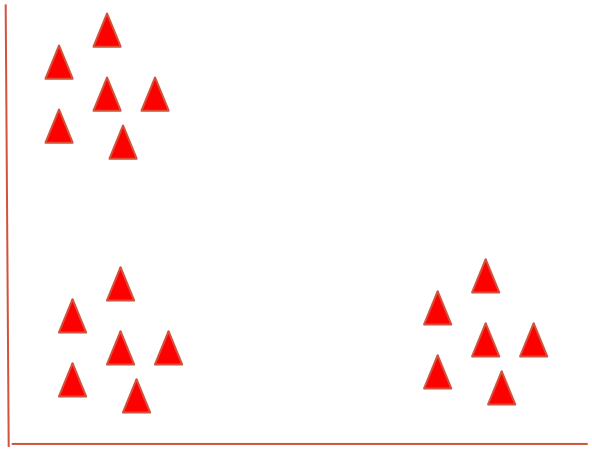
Our first model - Unsupervised learning

Discover the hidden structure in unlabeled data X (no y)

- Customer/product segmentation
- Data analysis for ...
- Identify number of speakers in a meeting recording
- Helps supervised learning in some task

Example - Customer analysis

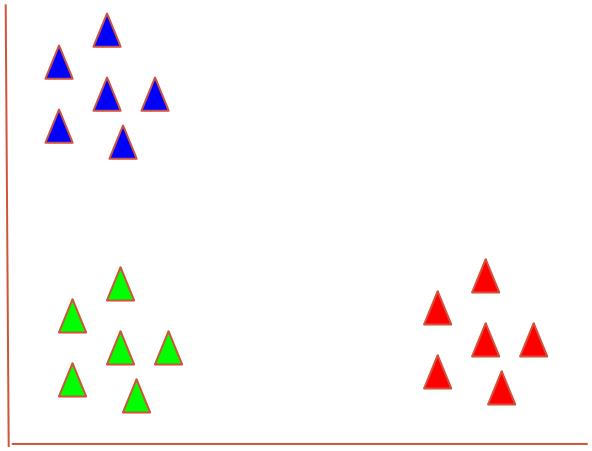
Brand loyalty



Price sensitivity

Example - Customer analysis

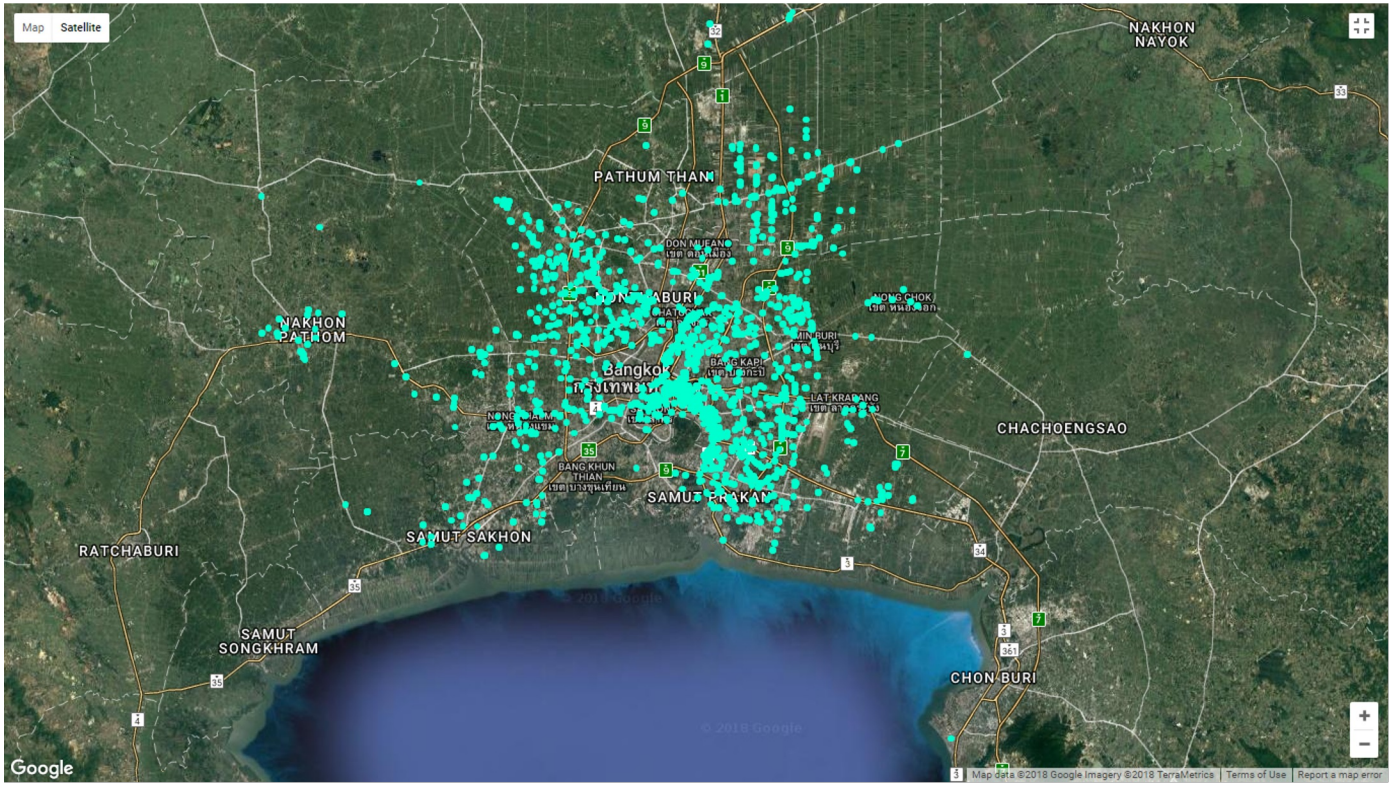
Brand loyalty



Price sensitivity

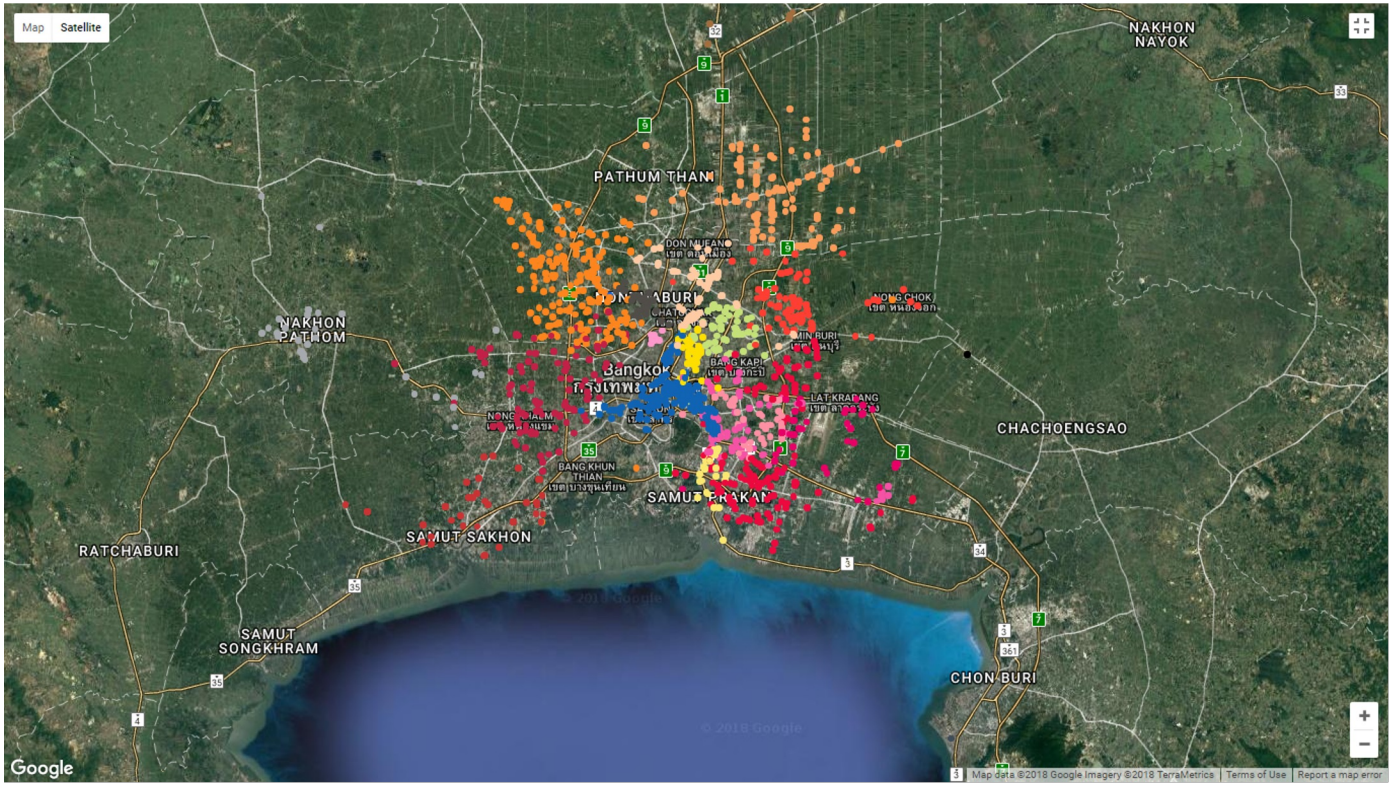
Example - Real Estate segmentation in Thailand

What should be the input feature of this?

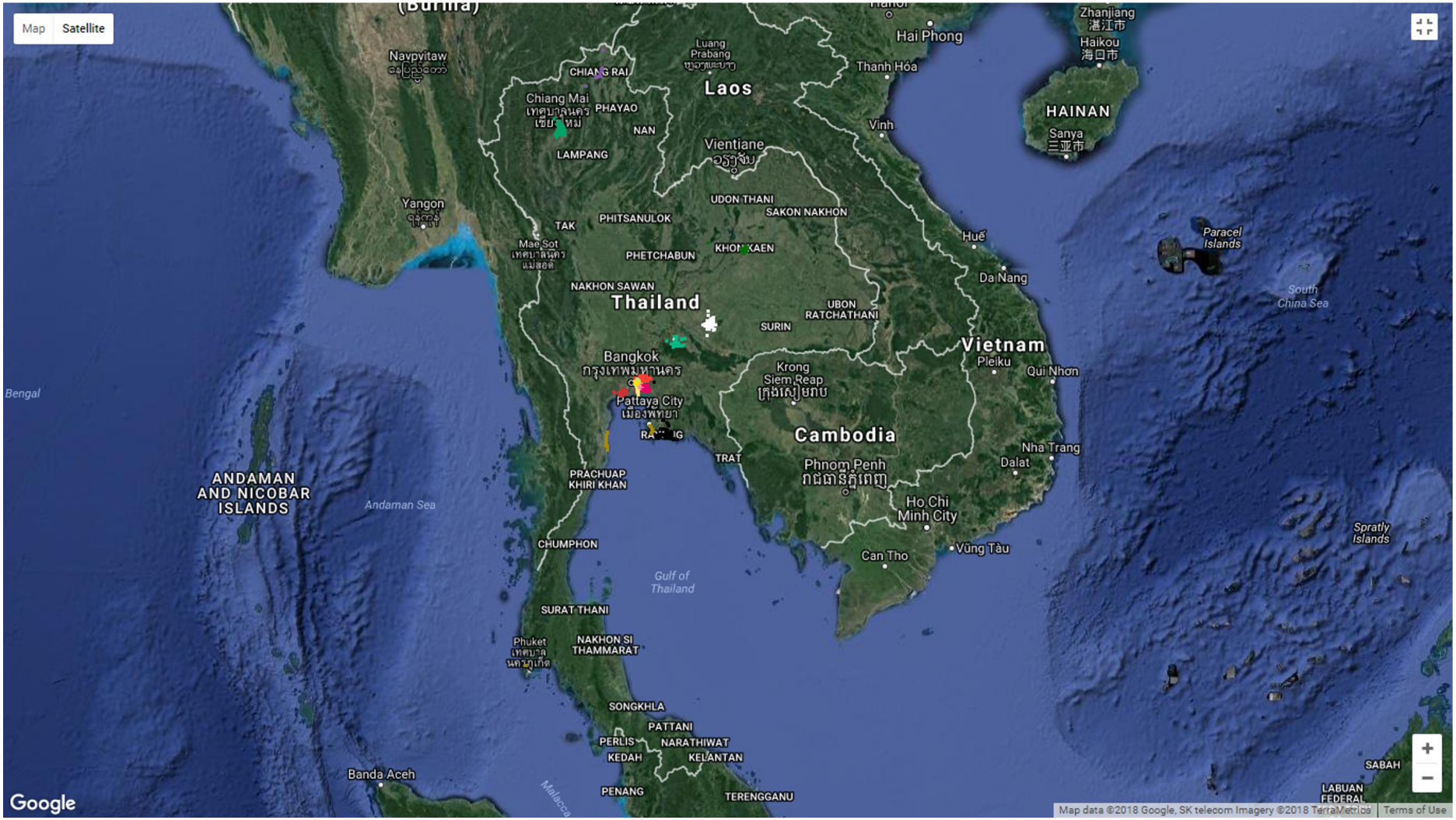


Example - Real Estate segmentation in Thailand

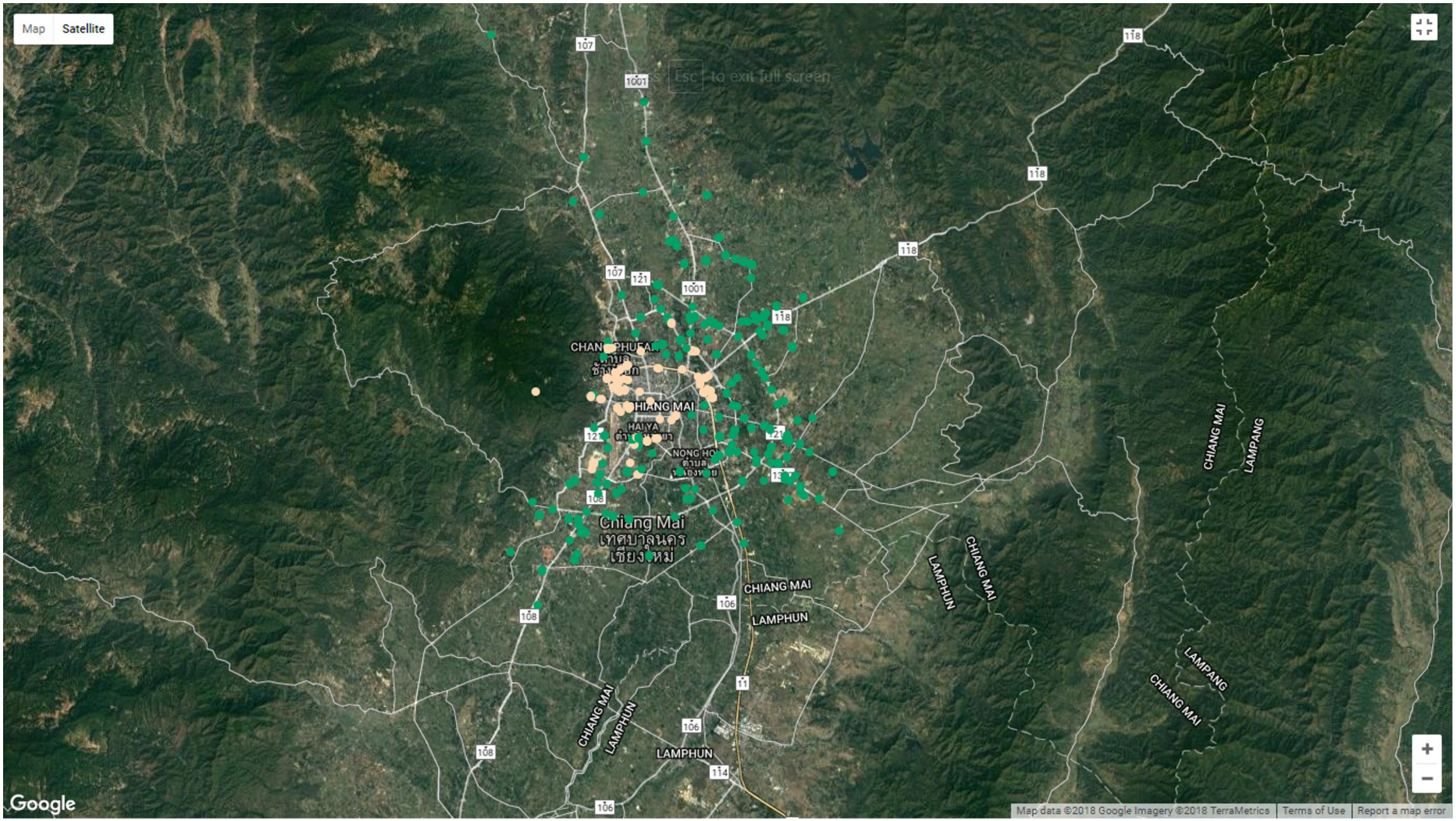
What should be the input feature of this?



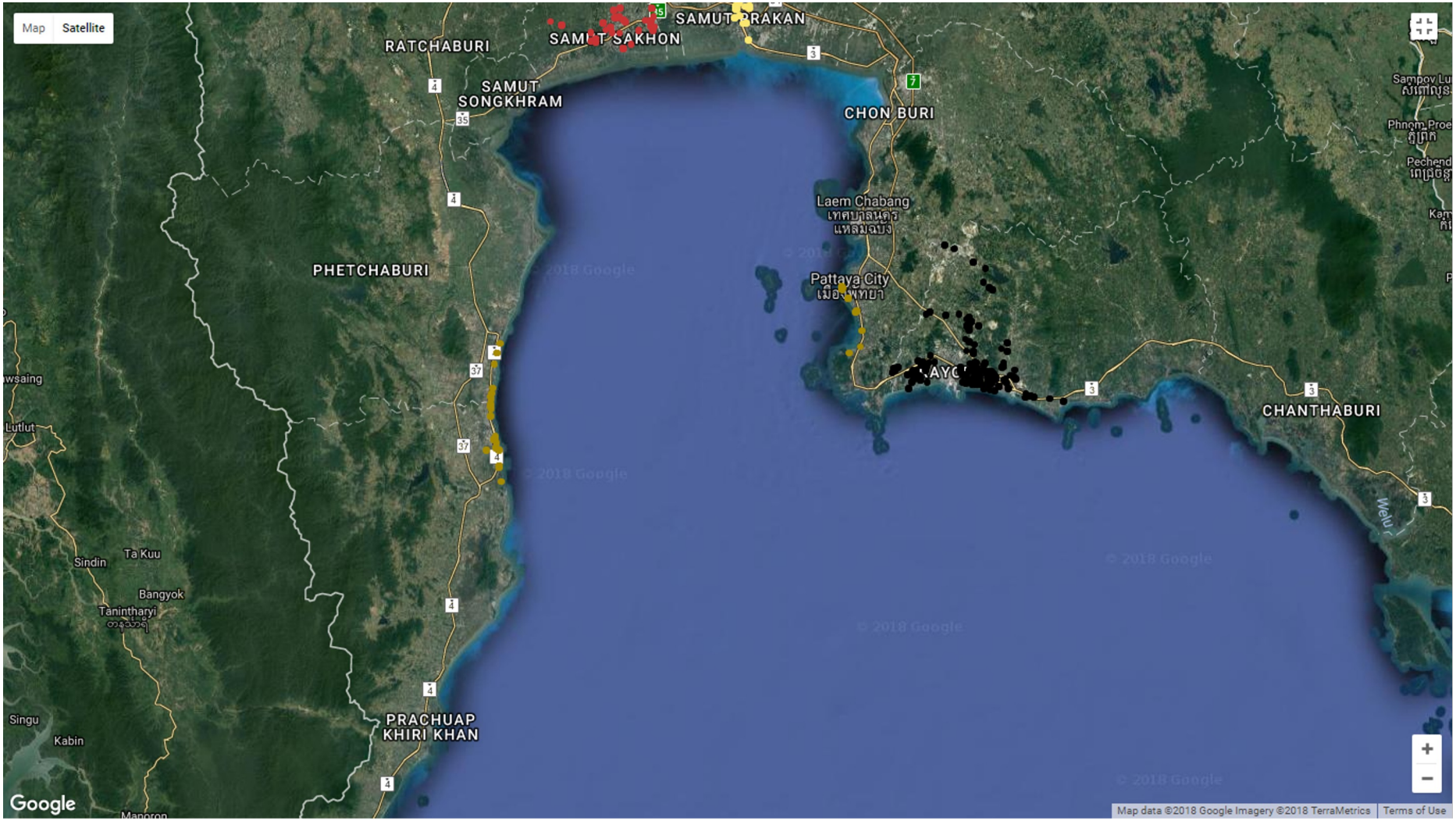
Example - Real Estate segmentation in Thailand



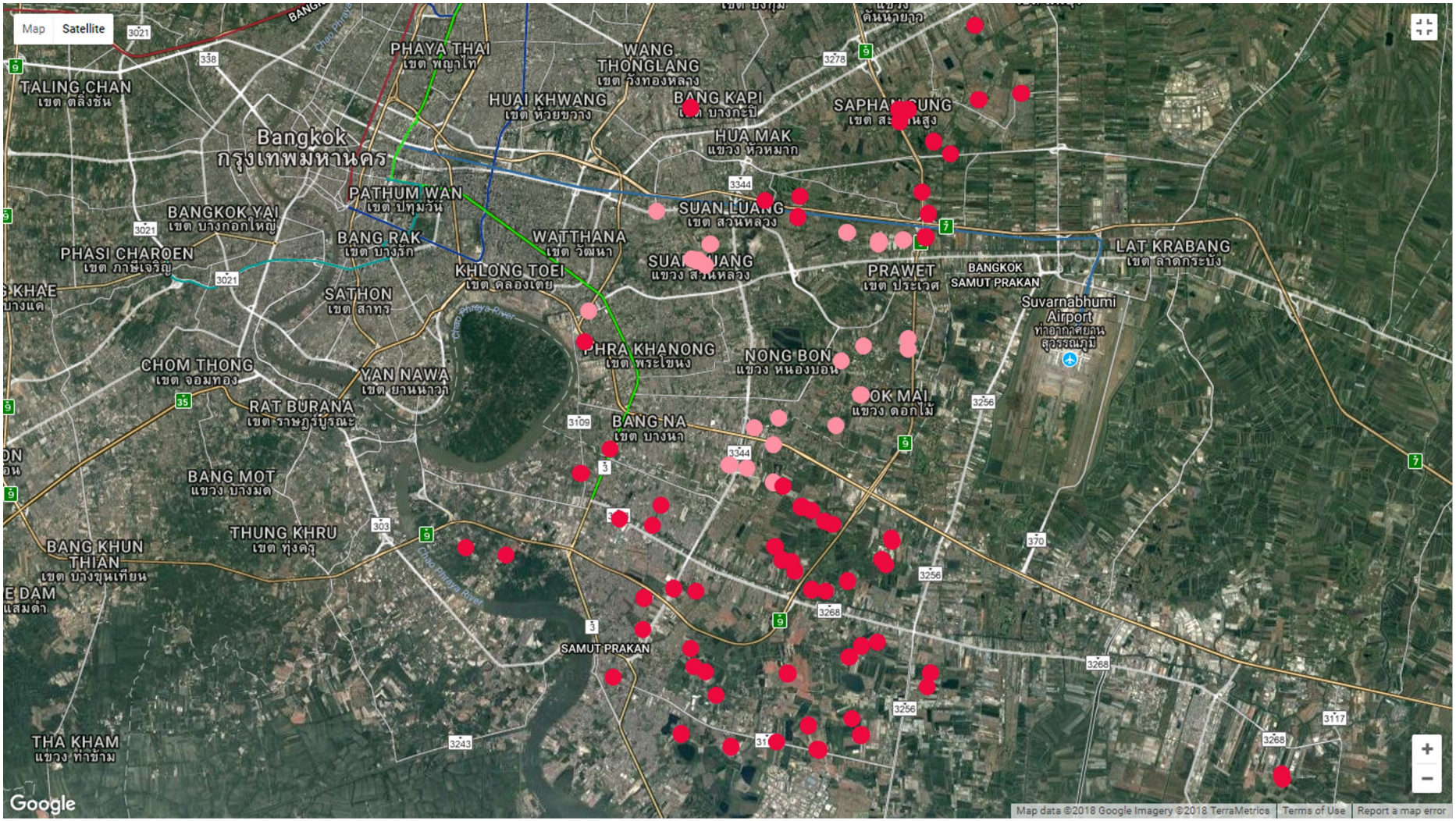
Example - Real Estate segmentation in Thailand



Example - Real Estate segmentation in Thailand



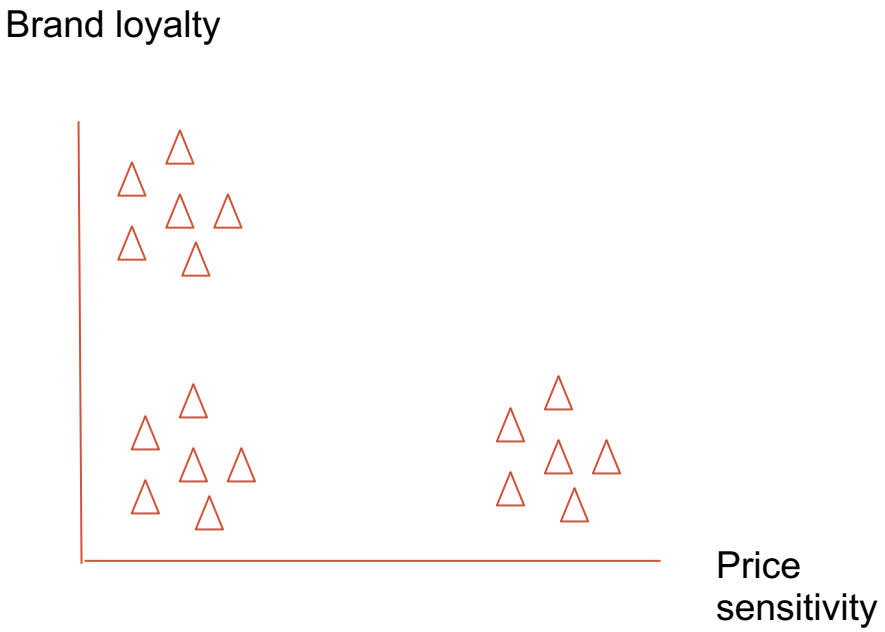
Example - Real Estate segmentation in Thailand



K-mean clustering

Clustering - task that tries to automatically discover groups within the data

Too hard…



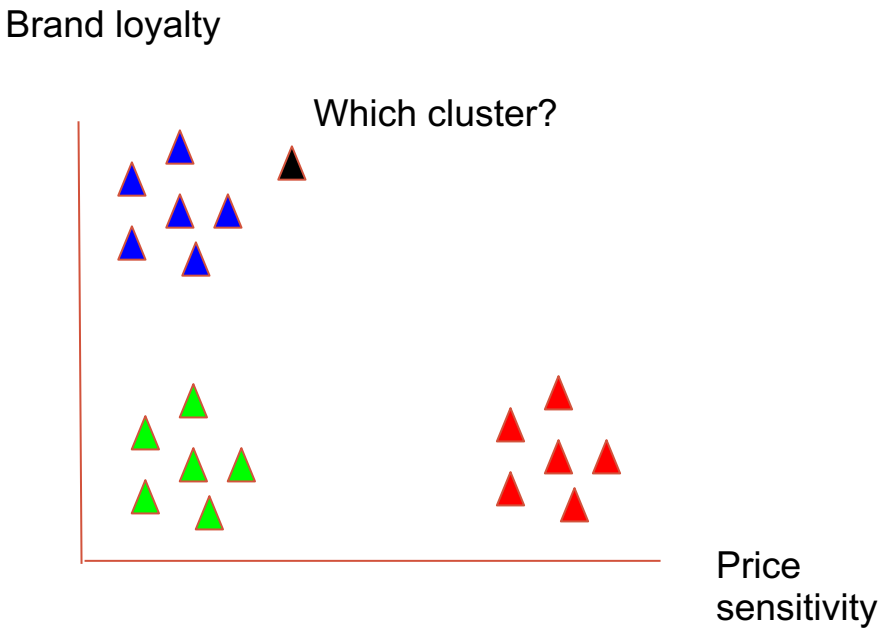
K-mean clustering

Clustering - task that tries to automatically discover groups within the data

Too hard…

Easier if we know the grouping beforehand (supervised)

How?



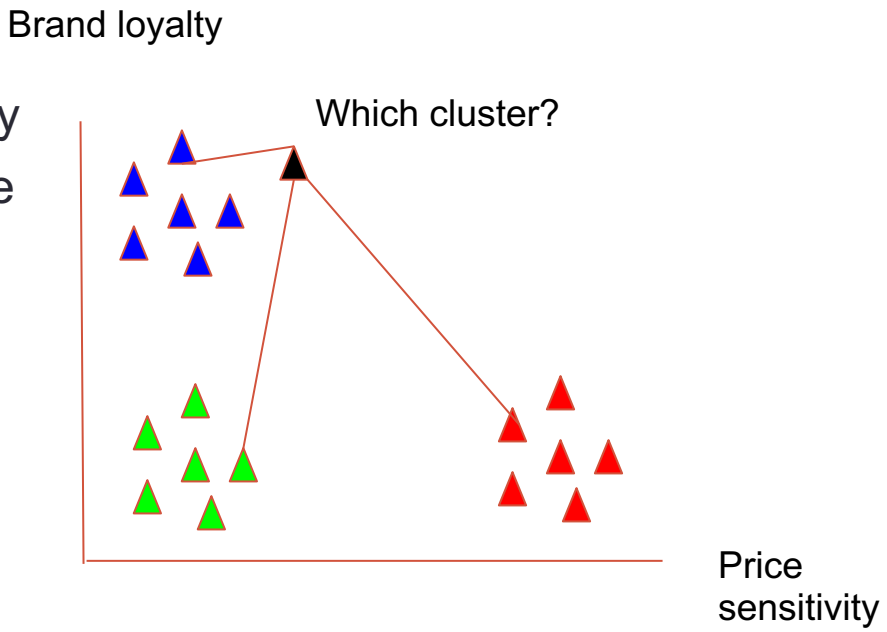
Nearest Neighbour classification

Find the closest training data, assign the same label as the training data

Given query data

For every point in the training data

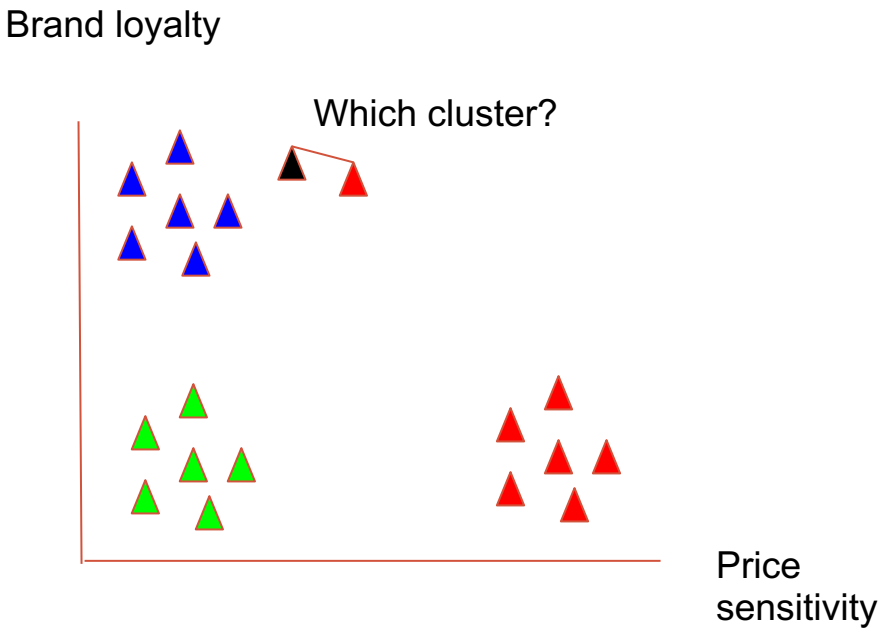
- Compute the distance with the query
- Assign label of the smallest distance



K-Nearest Neighbour (kNN) classification

Nearest Neighbour is susceptible to noise in the training data

Use a voting scheme instead



K-Nearest Neighbour (kNN) classification

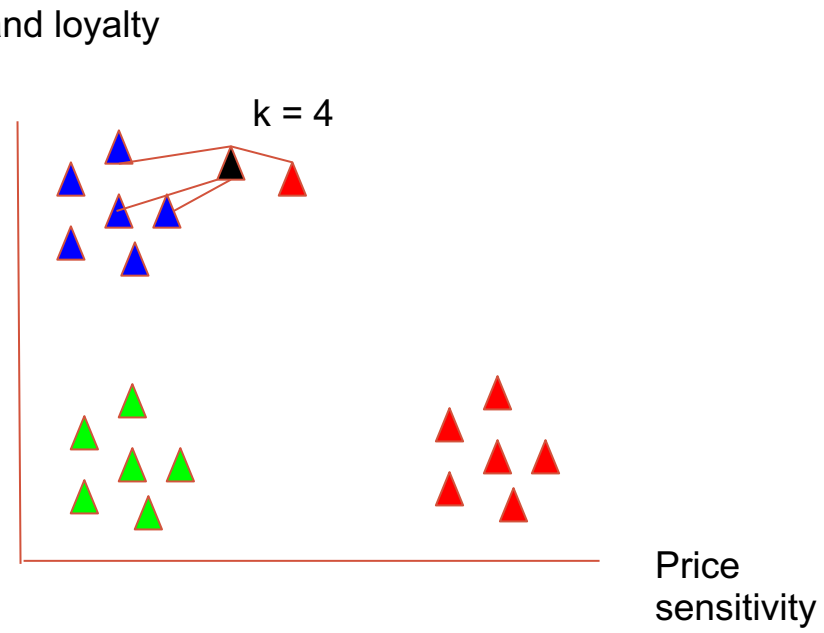
Nearest Neighbour is susceptible to noise in the training data

Use a voting scheme instead

Given query data

For every point in the training data

- Compute the distance with the query
- Assign label of the smallest distance
  - Assign label by voting



The votes can be weighted by the inverse distance (weighted k-NN)

Closest?

We need some kind of distance or similarity measures

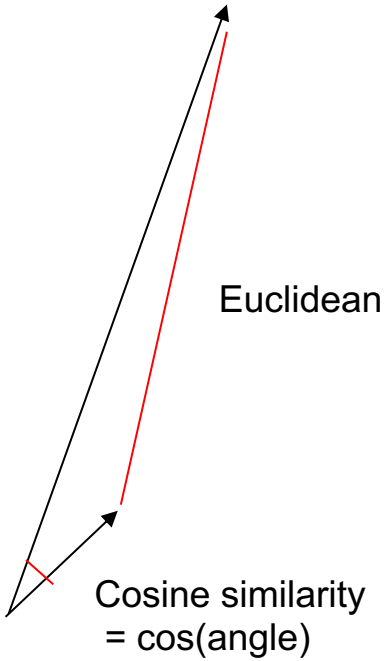
F($\mathbf{X_1}$,$\mathbf{X_2}$) = d

$$X_1 = [x_{1,1}, x_{1,2}, \ldots, x_{1,n}]$$

$$X_2 = [x_{2,1}, x_{2,2}, \ldots, x_{2,n}]$$

Euclidean distance $\sqrt{\Sigma_i (x_{1,i} - x_{2,i})^2}$

Cosine similarity $\frac{X_1 \cdot X_2}{|X_1||X_2|} = \frac{\Sigma_i x_{1,i} x_{2,i}}{\sqrt{\Sigma_i x_{1,i}^2} \sqrt{\Sigma_i x_{2,i}^2}}$



Many more distances, Jaccard distance, Earth mover distance

KNN runtime

For every point in the training data

Compute the distance with the query

Find the K closest data points

Assign label by voting

O(N)

O(JN) - If we have J queries

Expensive

Ways to make it faster

- Kernelized KNN
- Locally Sensitive Hashing (LSH)
- Use centroids

Centroids

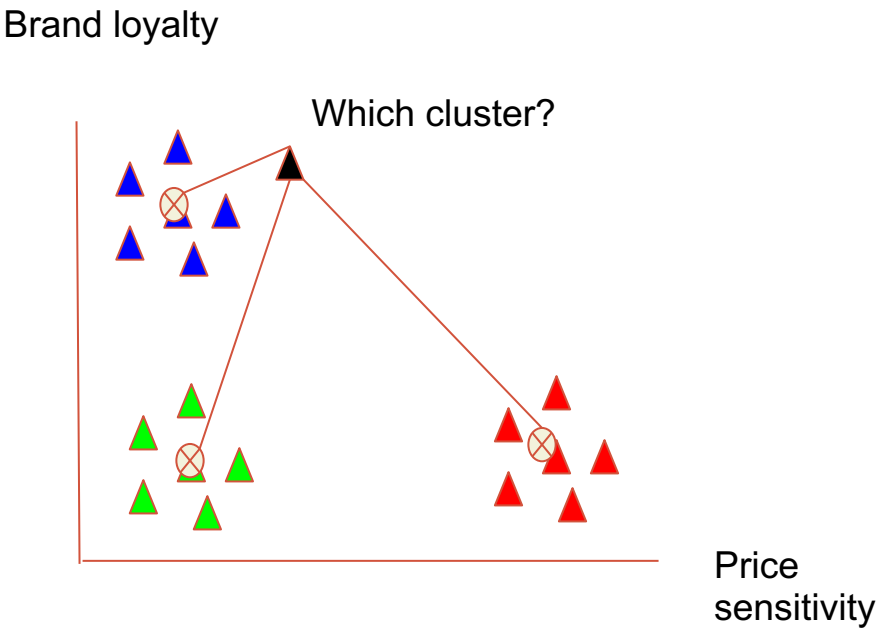
Basically, the representative of the cluster

Find the mean location of the cluster by averaging

Can use mode or median depending on the data

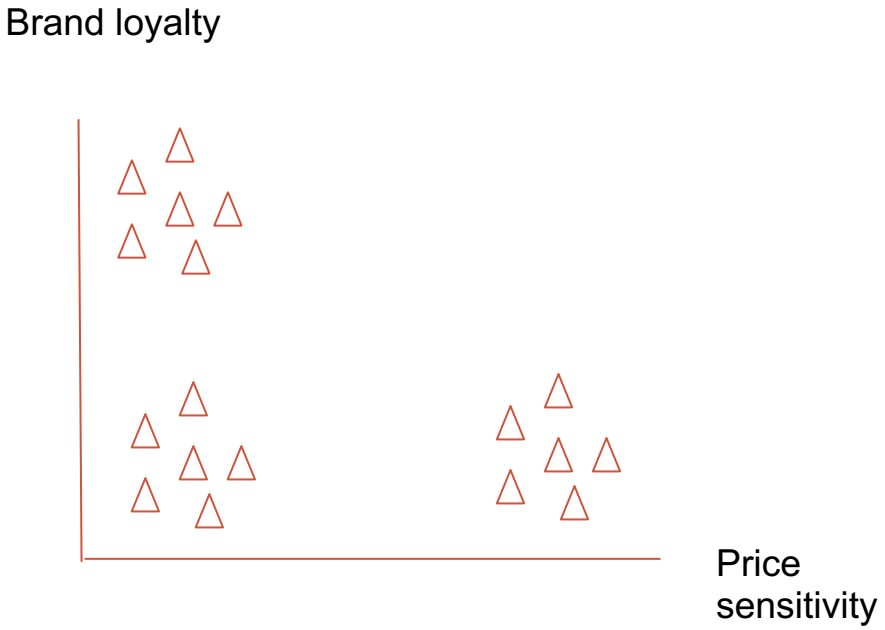
O(JL)

L - number of clusters

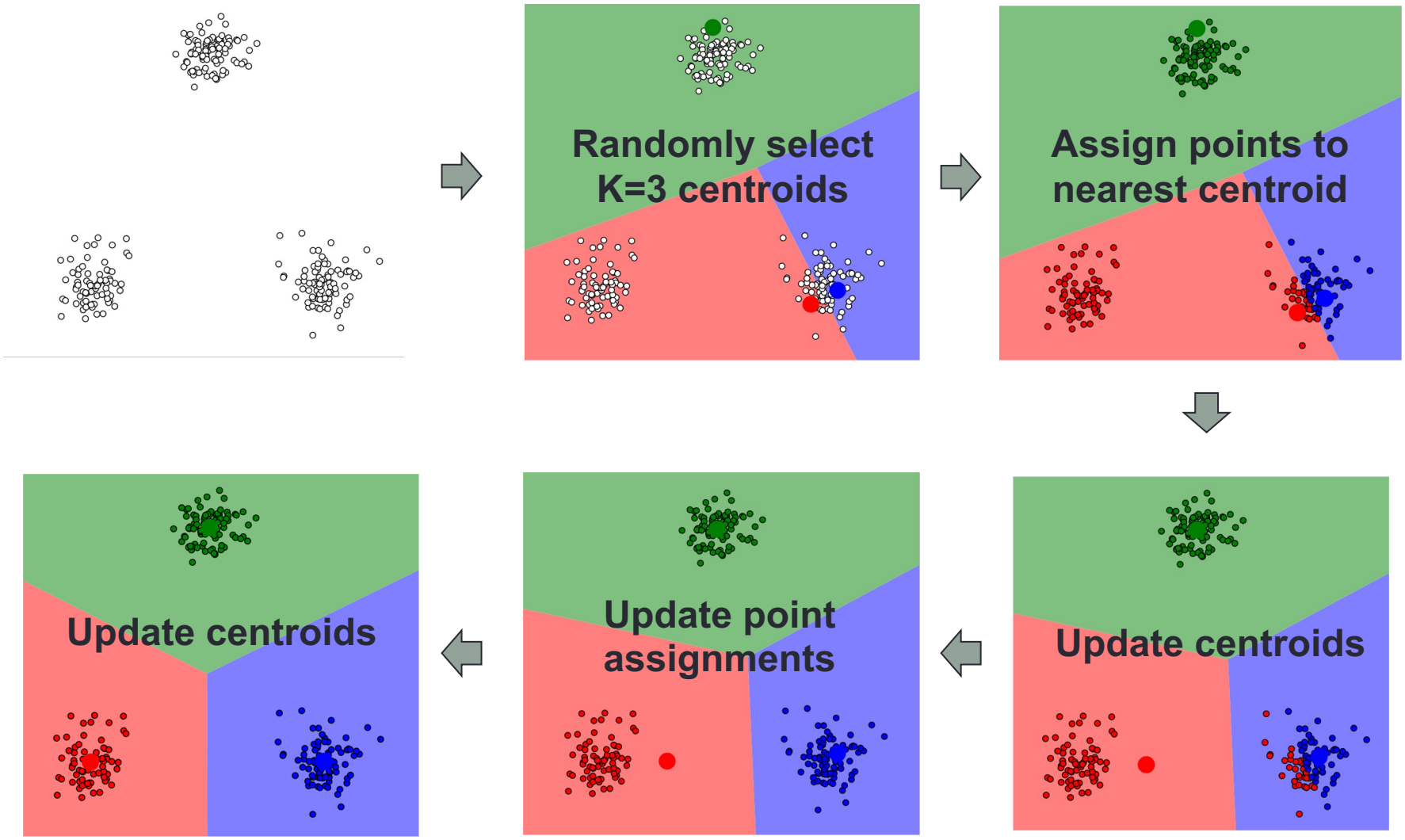


K-mean clustering

1. Randomly init k centroids by picking from data points
2. Assign each data points to centroids
3. Update centroids for each cluster
4. Repeat 2-3 until centroids does not change



An Illustration Of K-Mean Clustering



From https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

Characteristics of K-means

- **The number of clusters, *K*, is specified in advance.**
- **Always converge to a (local) minimum.**
  - **Poor starting centroid locations can lead to incorrect minima.**

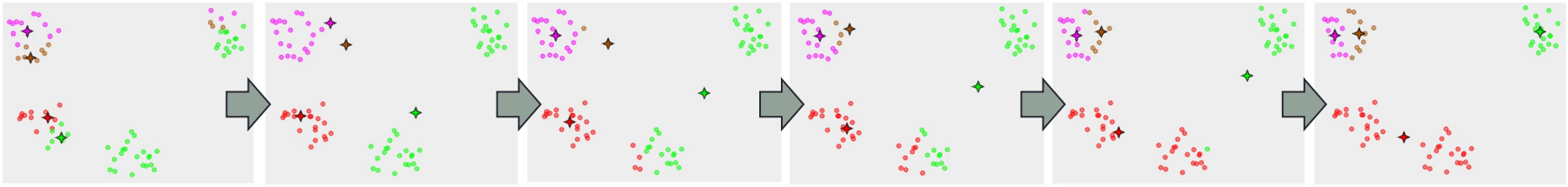
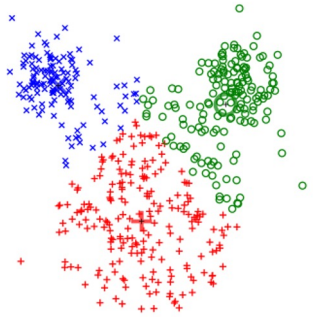
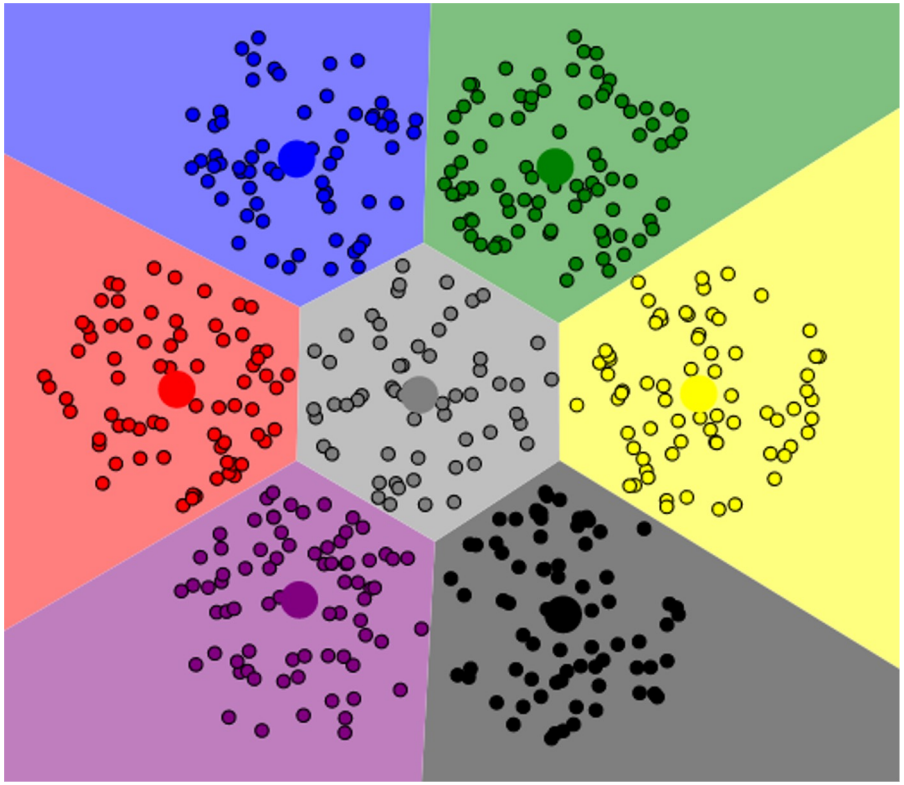


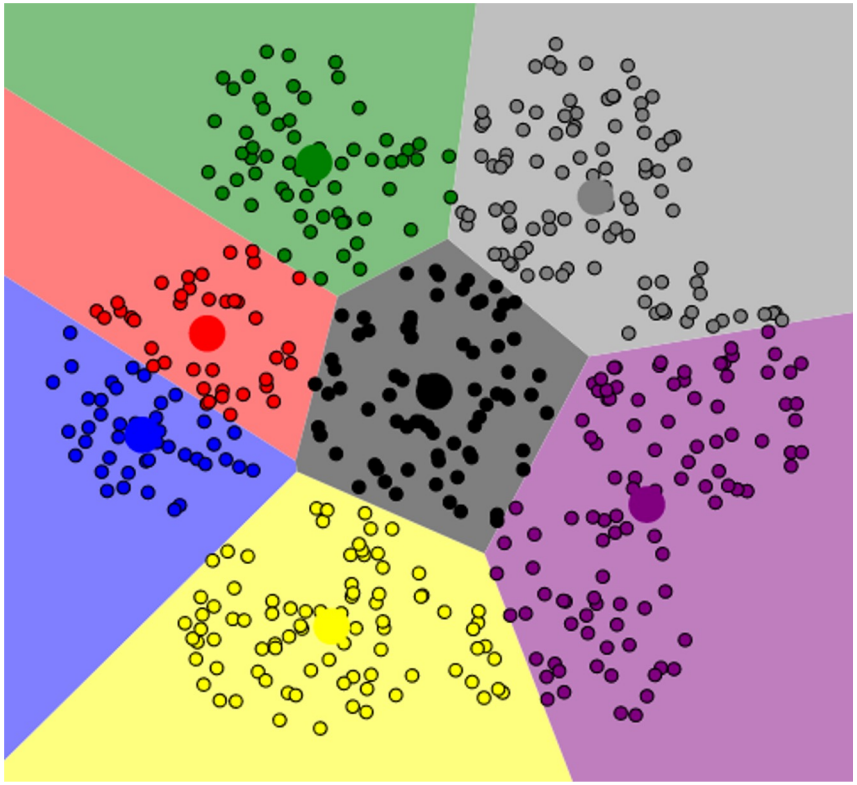
Image from https://en.wikipedia.org/wiki/K-means_clustering

- **The model has several implicit assumptions:**
  - **Data points scatter around cluster's centers.**
  - **Boundary between adjacent clusters is always halfway between the cluster centroids.**



Effect of bad initializations



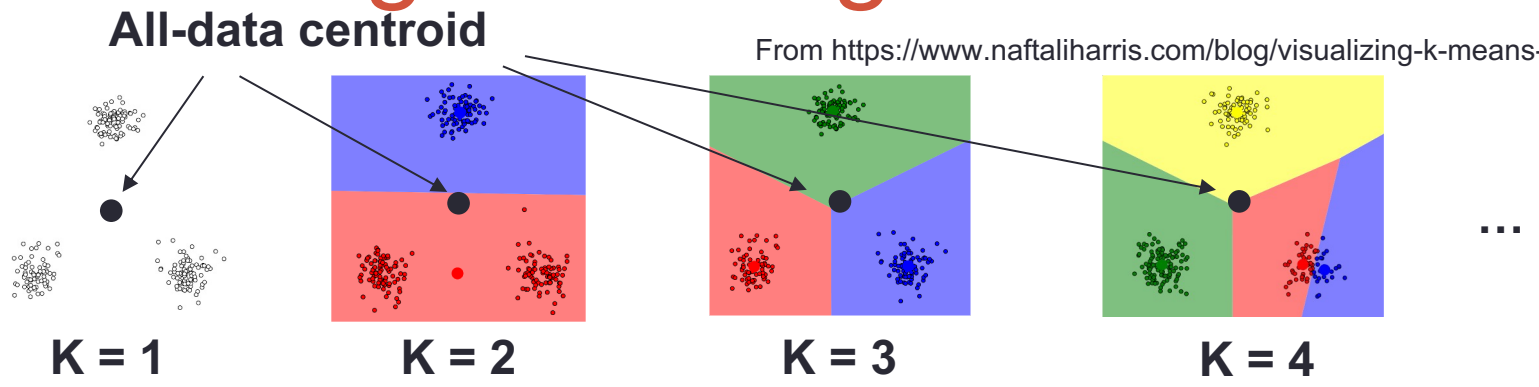


Solution, try different randomization and pick the best

From https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

Clustering Metrics?

Selecting K - Using Elbow method



fraction of explained variance = $\frac{\text{between-cluster variance}}{\text{all-data variance}}$

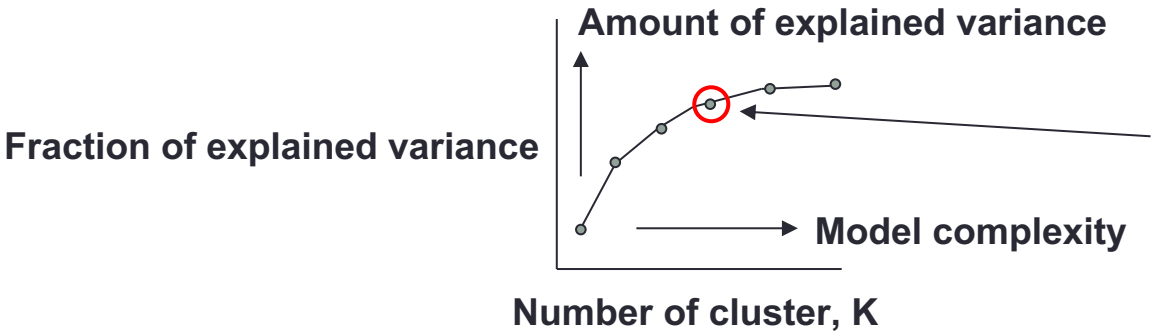
(note: this - denotes Euclidean distance)

between-cluster variance = $\sum_{i=1}^{K} \frac{n_i (M_i - M)^2}{N - 1}$, where n_i = size of i^{th} cluster,

M_i = centroid of i^{th} cluster, and

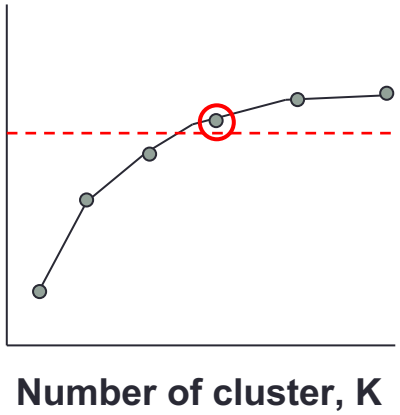
M = all-data centroid.

all-data variance = $\sum_{i=1}^{N} \frac{(x_i - M)^2}{N - 1}$, where x_i = i^{th} data point and N = # of data.



The elbow method chooses K where increasing complexity doesn't yield much in return.

Selecting K - other methods



Choose minimal K that explains at least 95% of the all-data variance.

K = 2
K = 3
K = 4
⋮

→ Training K-mean Clustering Model → Testing / Cross-validation →

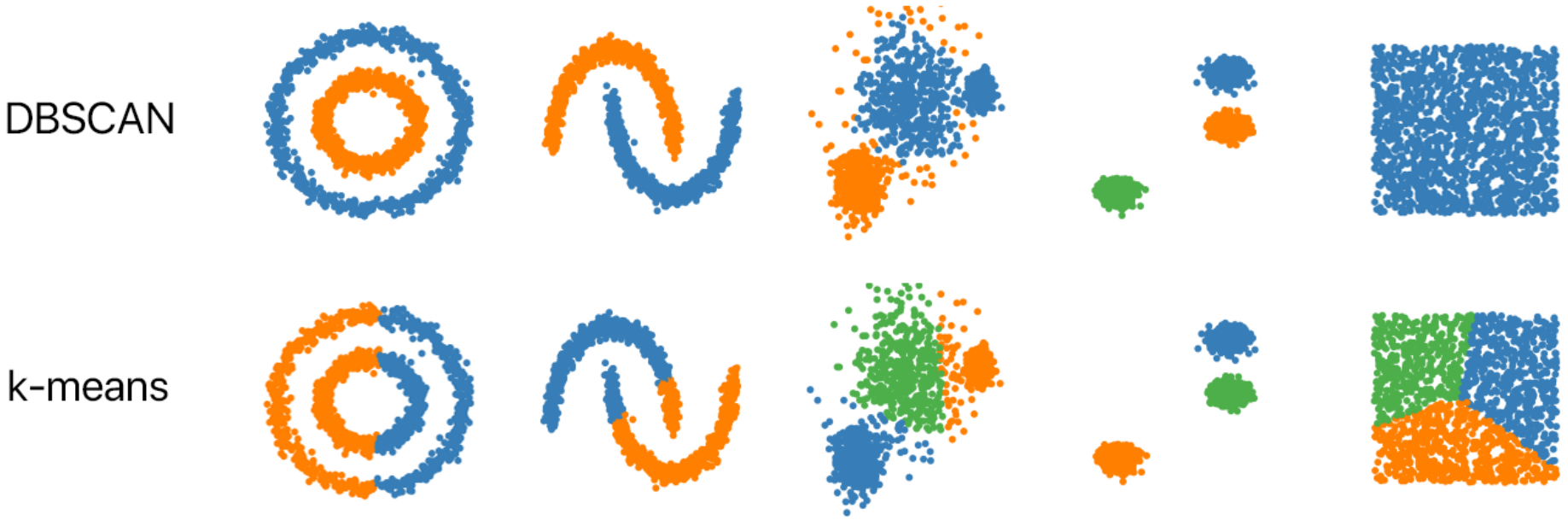
K	Accuracy
2	50%
3	68%
4	83%
⋮	⋮

Choose K that maximizes certain objective (e.g. accuracy on testing data)

Best method

Summary

- Other clustering methods
  - K-mode, K-median
  - Spectral clustering (clustering in embedding space)
  - DBScan (clustering by “density”– very robust, no need for k*)



https://github.com/NSHipster/DBSCAN