Introduction – Machine Learning

David Quigley

CSCI 4622

About Me

Dr. David Quigley (You can call me David or Dr. Quigley)
Assistant Teaching Professor, Dept. of Computer Science
12th year at CU (7th year post-graduate)
Applied Machine Learning Research:
Studying the way students use digital tools for learning

- Student use of LMS
- Student Scientific Modeling
- Students Reading to Learn
- Students' epistemological beliefs on the nature of science

About the Course Staff

- TBD

About You

Post an introduction on Piazza in the “Personal Introductions” discussion! Share as much or as little as you want (but I want everyone to post, to make sure everyone can access, use Piazza).

(This is a participation grade, due by Wednesday 1/22)

Class Resources

Computing Device
 - Laptop recommended, any OS (ish)

Course Canvas Website
canvas.colorado.edu

Python 3 (with external packages)
 - Google Colab as a default https://colab.research.google.com/
 - Anaconda recommended https://www.anaconda.com/download/

"Dumb" Calculator

Course Logistics

Weekly Participation– Expected *Weekly*

- Weekly activities have already begun to appear (Piazza Introduction)
- Weekly Participation = 10% of Grade

Assignments (Problem Sets / Homeworks) – Every 2 - 3 Weeks

- See collaboration policy, submission deadlines
- Problem Sets = 30% of Grade

Midterms – Approximately every 7 – 8 weeks

- Already on the Calendar
- Specifics outlined in the week or two before the exam
- Midterms = 30% of Grade

Project Updates - Every 3 – 5 weeks (not synced with midterms)

- Be on the lookout for team information
- Projects = 30% of Grade

Course Functionality – Campus Protocols

- Rule #1 – We will be following all campus rules and guidelines for instruction.
    - Campus Closures (Weather, etc.)

- Flexibility – Life happens, and you can have an interruption while keeping up with class
    - Participation will be over time
    - OH are a good time to review missed materials – you can make them “by appointment”

Course Functionality

- I may occasionally have to adjust class for everyone’s safety.
    - Try to set up your notifications to see Canvas and / or Piazza notifications.

Course Functionality

- I may occasionally have to adjust class for everyone’s safety.
    - Try to set up your notifications to see Canvas and / or Piazza notifications.

Course Logistics

- Weekly Activity – Piazza Introduction
  - Due Friday, August 29
- Problem Set 1 – releases Next Week (Friday)
  - Due Friday, September 12 @ 11:59 PM (i.e. as late as possible)
- Orienting yourself to the syllabus, calendar, weekly modules, etc.
  - Maybe you've started doing course readings? That'd be great!
    - Don't get *too* far ahead on course readings - you'll get yourself lost!

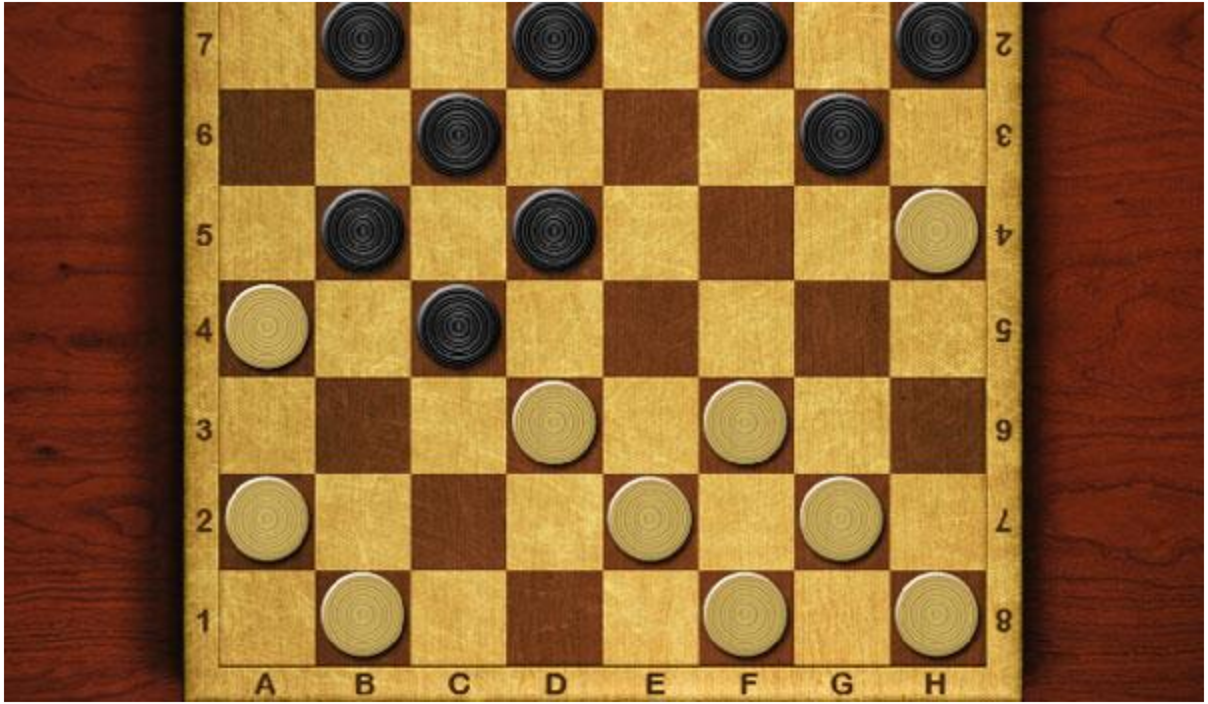
Goals for this Course (see syllabus)

- Explain a problem from an ML perspective
- Select which ML techniques and approaches are best suited to your problem
- Prepare your data to implement the chosen approach
- Apply your ML approach to generate a solution
- Evaluate the results of your solution and share them with others
- Implement common solutions in Python

What is Machine Learning (ML)?

What is ML? Through History…

Arthur Samuel (1959) - Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed.


7
6
5
4
3
2
1
A B C D E F G H

What is ML? Through History...

Well-Posed Learning Problem (Tom Mitchell, 1998) - A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

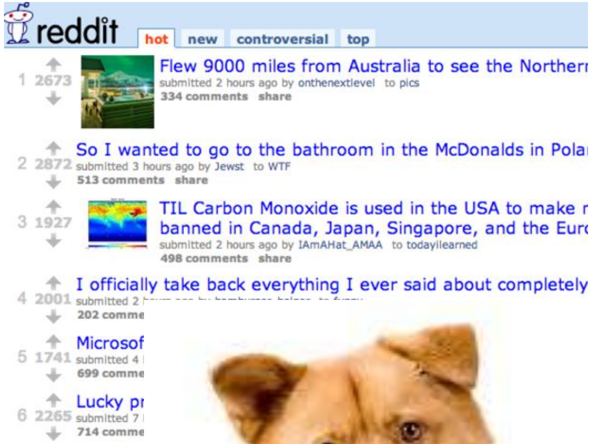
What is "Task"?

reddit
hot
new
controversial
top
1
2673
Flew 9000 miles from Australia to see the Northern Lights. From a hot tub. (i.imgur.com)
submitted 2 hours ago by onthenextlevel to pics
334 comments share
2
2872
So I wanted to go to the bathroom in the McDonalds in Poland. This is what I found. (imgur.com)
submitted 3 hours ago by Jewst to WTF
513 comments share
3
1927
TIL Carbon Monoxide is used in the USA to make meat appear fresher. This practise is banned in Canada, Japan, Singapore, and the European Union. (en.wikipedia.org)
submitted 2 hours ago by IAmAHat_AMAA to todayilearned
498 comments share
4
2001
I officially take back everything I ever said about completely hating all sports (i.imgur.com)
submitted 2 hours ago by hamburger_helper to funny
202 comments share
5
1741
Microsoft Security Essentials fails anti-virus certification test, Redmond challenges results (theverge.com)
submitted 4 hours ago by ken27238 to technology
699 comments share
6
2265
Lucky prospector finds huge gold nugget 177 Oz/5.5kg (Australia) (abc.net.au)
submitted 7 hours ago by LuckyBdx4 to worldnews
714 comments share

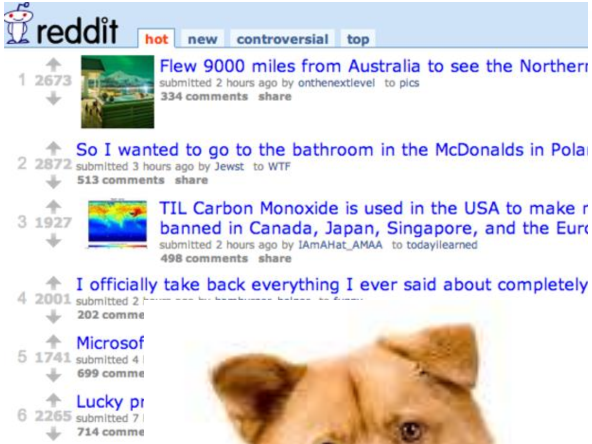
What is “Task”?


reddit
hot
new
controversial
top
1 2673
Flew 9000 miles from Australia to see the Northern Lights. From a hot tub. (i.imgur.com)
submitted 2 hours ago by onthenextlevel to pics
334 comments share
2 2872
So I wanted to go to the bathroom in the McDonalds in Poland. This is what I found. (imgur.com)
submitted 3 hours ago by Jewst to WTF
513 comments share
3 1927
TIL Carbon Monoxide is used in the USA to make meat appear fresher. This practise is banned in Canada, Japan, Singapore, and the European Union. (en.wikipedia.org)
submitted 2 hours ago by IAmAHat_AMAA to todayilearned
498 comments share
4 2001
I officially take back everything I ever said about completely hating all sports (i.imgur.com)
4 2001
202 comme
5 1741
Microsof
submitted 4
699 comme
6 2265
Lucky pr
submitted 7
714 comme

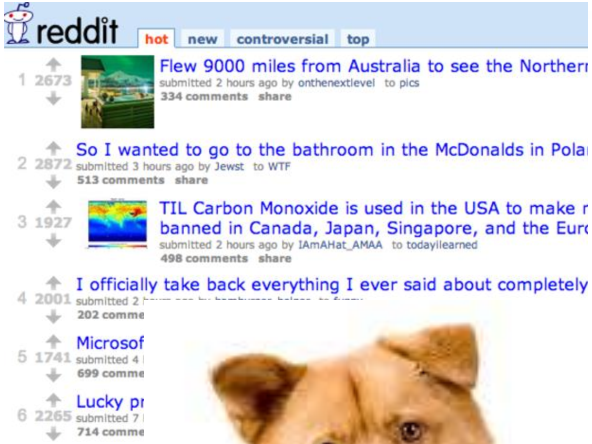
What is “Task”?


reddit
hot
new
controversial
top
1 2673
Flew 9000 miles from Australia to see the Northerr
submitted 2 hours ago by onthenextlevel to pics
334 comments share
2 2872
So I wanted to go to the bathroom in the McDonalds in Pola
submitted 3 hours ago by Jewst to WTF
513 comments share
3 1927
TIL Carbon Monoxide is used in the USA to make r
banned in Canada, Japan, Singapore, and the Eur
submitted 2 hours ago by IAmAHat_AMAA to todayilearned
498 comments share
4 2001
I officially take back everything I ever said about completely
202 comme
5 1741
Microsof
699 comme
6 2265
Lucky pr
714 comme

What is “Task”?


reddit
hot
new
controversial
top
1 2673
Flew 9000 miles from Australia to see the Northern
submitted 2 hours ago by onthenextlevel to pics
334 comments share
2 2872
So I wanted to go to the bathroom in the McDonalds in Pola
submitted 3 hours ago by Jewst to WTF
513 comments share
3 1927
TIL Carbon Monoxide is used in the USA to make r
banned in Canada, Japan, Singapore, and the Eur
submitted 2 hours ago by IAmAHat_AMAA to todayilearned
498 comments share
4 2001
I officially take back everything I ever said about completely
202 comme
5 1741
Microsof
699 comme
6 2265
Lucky pr
714 comme

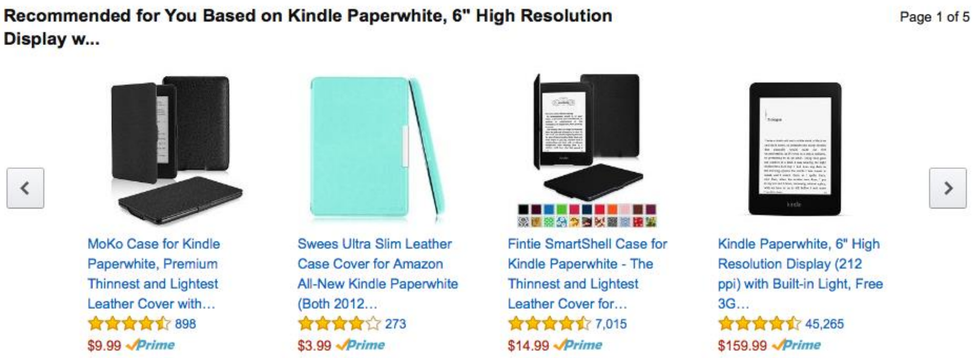
What is "Task"?


reddit
hot
new
controversial
top
1 2673
Flew 9000 miles from Australia to see the Northern
submitted 2 hours ago by onthenextlevel to pics
334 comments share
2 2872
So I wanted to go to the bathroom in the McDonalds in Pola
submitted 3 hours ago by Jewst to WTF
513 comments share
3 1927
TIL Carbon Monoxide is used in the USA to make
banned in Canada, Japan, Singapore, and the Eur
submitted 2 hours ago by IAmAHat_AMAA to todayilearned
498 comments share
4 2001
I officially take back everything I ever said about completely
202 comme
5 1741
Microsof
699 comme
6 2265
Lucky pr
714 comme

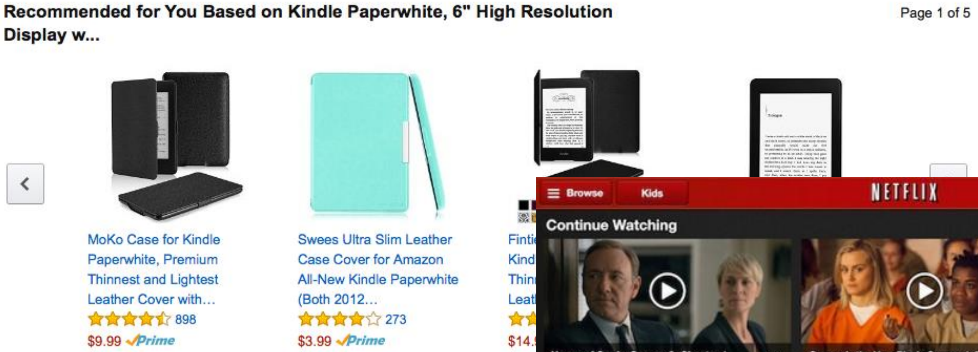
What is ML? Through History...

Well-Posed Learning Problem (Tom Mitchell, 1998) - A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

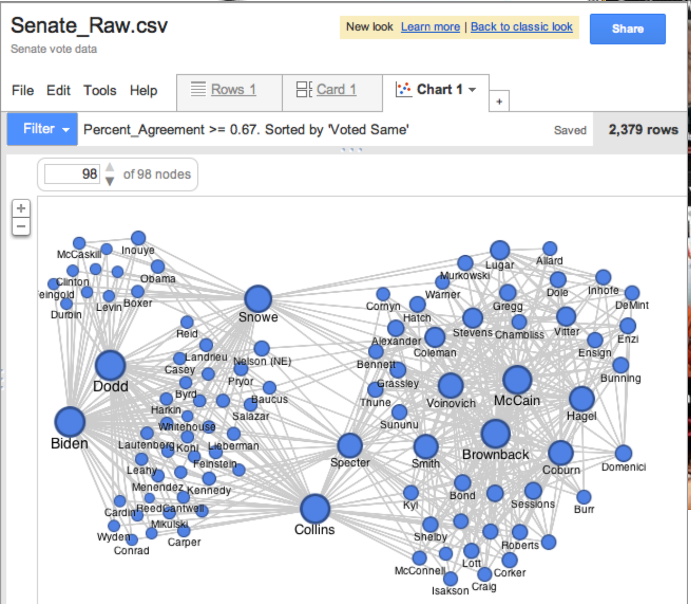
What is “Experience”?



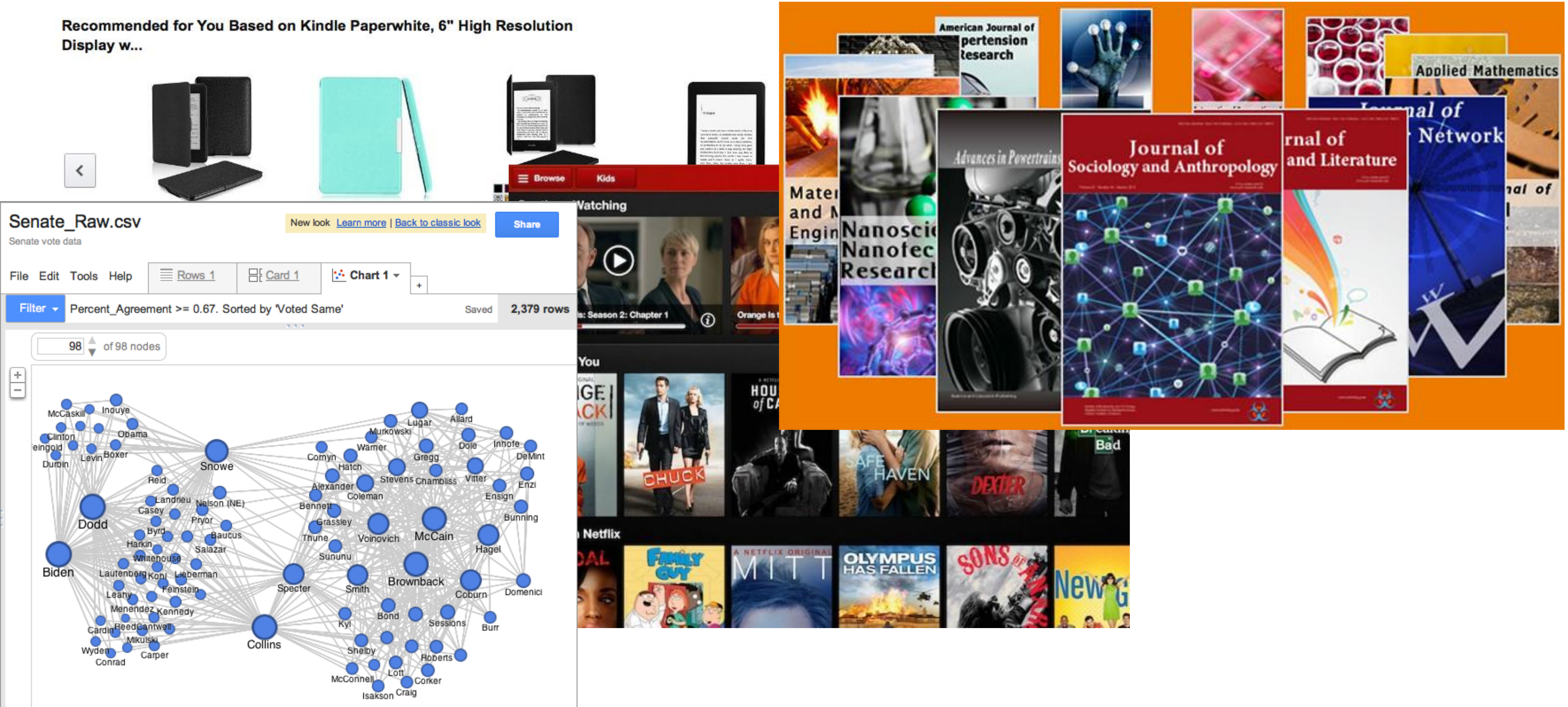
What is "Experience"?



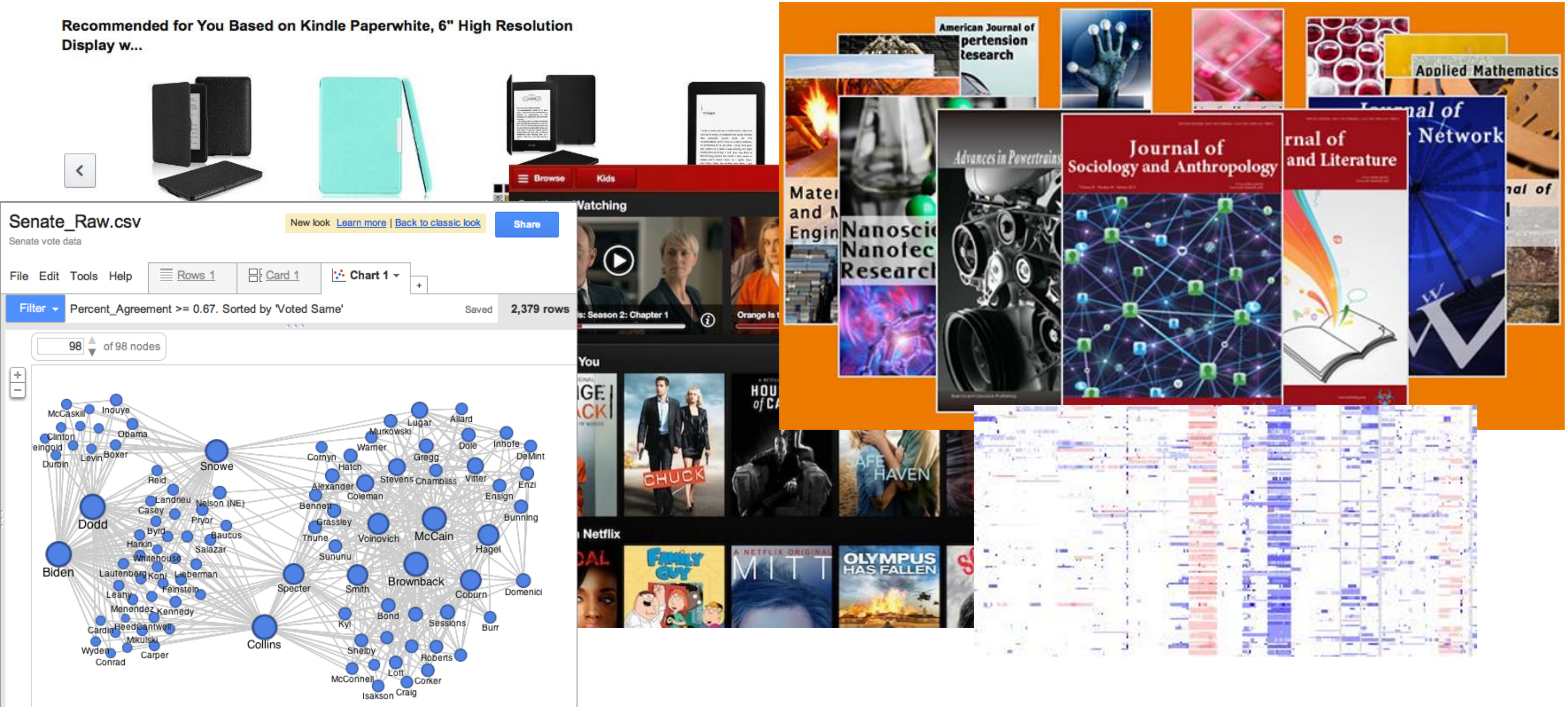
What is "Experience"?



What is “Experience”?



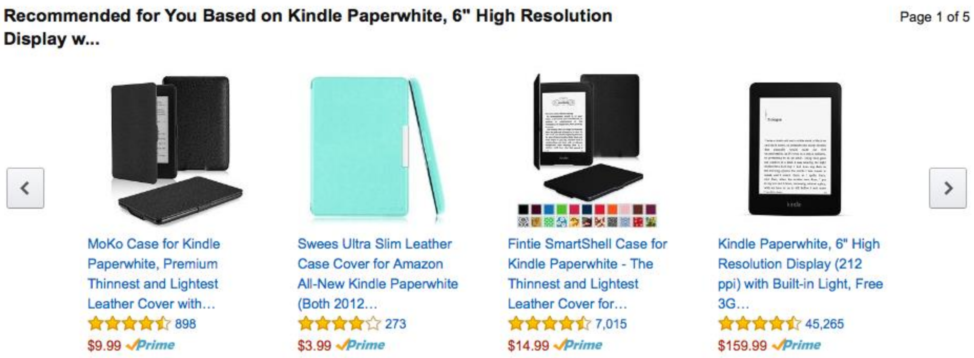
What is "Experience"?



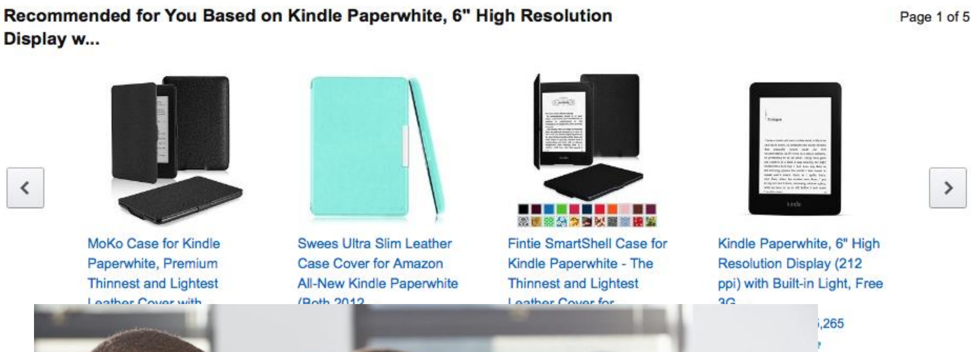
What is ML? Through History…

Well-Posed Learning Problem (Tom Mitchell, 1998) - A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

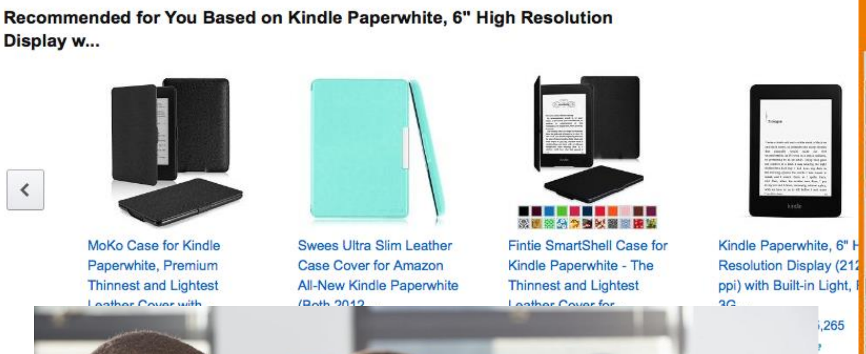
What is “[Improved] Performance”?



What is “[Improved] Performance”?



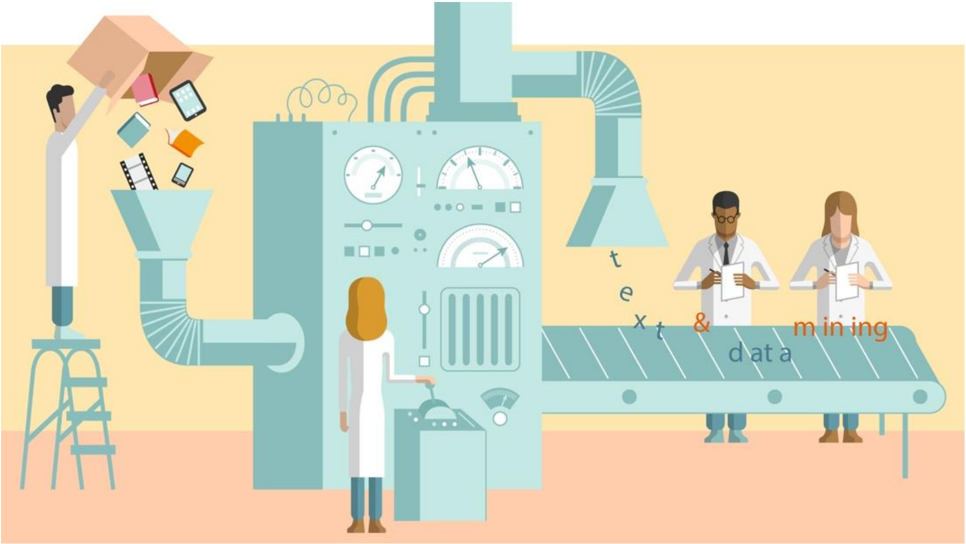
What is “[Improved] Performance”?



What this course isn't – Deep Learning (5922)

We'll cover these areas, but not in depth

What this course isn't – Data Mining (4502)

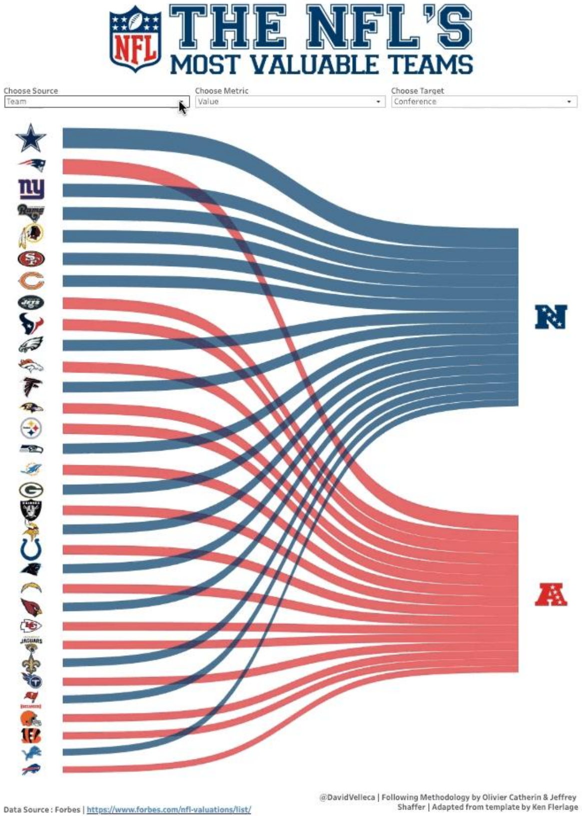


We'll cover these areas, but not in depth

What this course isn't – NLP (3832)

We'll cover these areas, but not in depth

What this course isn't – Info Viz (INFO 4602)



Machine Learning is Math

Data (X) → Hidden Relationship (Z) → Answer (Y)

$X = (X_1, X_2, \dots X_n)$, each entry in X is a *feature*

Y is a *response*

Z is the *mapping* from *features* to *response*

Machine Learning is Math

Data (X) → Hidden Relationship (Z) → Answer (Y)

$X = (X_1, X_2, \ldots X_n)$, each entry in X is a *feature*
Y is a *response*
Z is the *mapping* from *features* to *response*

According to theory, there is a Z to map every X (of infinite size) to the real Y

Machine Learning is Math

Data (X) → Hidden Relationship (Z) → Answer (Y)

$X = (X_1, X_2, \ldots X_n)$, each entry in X is a *feature*

Y is a *response*

Z is the *mapping* from *features* to *response*

According to theory, there is a Z to map every X (of infinite size) to the real Y

Machine Learning is an approximation of Z

(Sometimes we care about trying to discover / measure the true Z, sometimes not)

Problem Space – Housing Market

Problem Space – Housing Market

X_i =

Y =

Machine Learning is Math

Data (X) → Hidden Relationship (Z) → Answer (Y)

$X = (X_1, X_2, \dots X_n)$, each entry in X is a *feature*
Y is a *response*
Z is the *mapping* from *features* to *response*

According to theory, there is a Z to map every X (of infinite size) to the real Y

Machine Learning is an approximation of Z
(Sometimes we care about trying to discover / measure the true Z, sometimes not)

Problem Space – Housing Market

X_i =

Y =

Problem Space – Housing Market

X_1	X_2	X_3	X_4	Y
Size (Sq. Ft.)	**# Bed**	**# Bath**	**Year Built**	**Price ($)**
1200	1	1.5	1998	200,000
1800	2	2	1985	450,000
800	1	1	2017	250,000
2500	3	2	1975	500,000
2800	4	2.5	1983	400,000
…	…	…	…	…

Problem Space: Washing Machines

X

Y

Problem Space – Washing Machines?

Supervised Learning

Find Patterns in *fully observed* data, then try to predict *partially observed* data.

Supervised Learning

Find Patterns in *fully observed* data, then try to predict *partially observed* data.

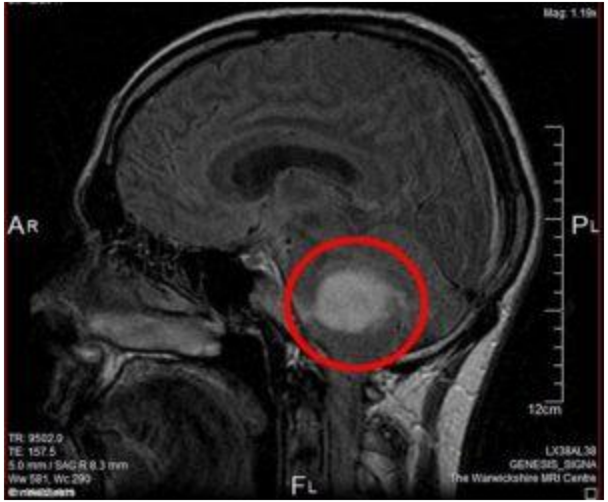
Approximate Z, as f = X → Y

Supervised Learning

Find Patterns in *fully observed* data, then try to predict *partially observed* data.

Approximate Z, as $f = X \rightarrow Y$

$D = \{(X_i, Y_i)\}_{i=1 \rightarrow n}$



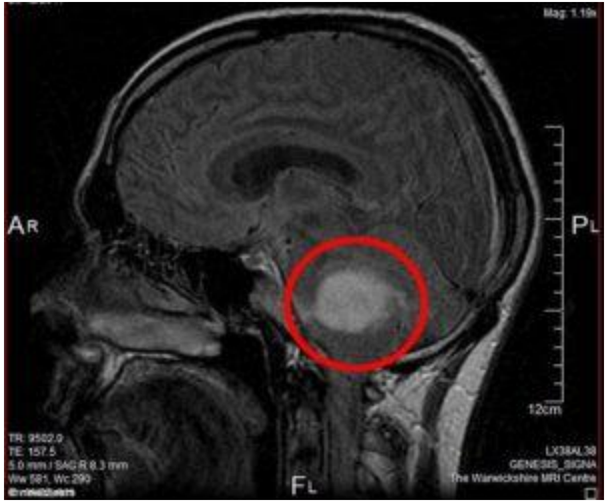
Supervised Learning

Find Patterns in *fully observed* data, then try to predict *partially observed* data.

Approximate Z, as $f = X \rightarrow Y$

$D = \{(X_i, Y_i)\}_{i=1 \rightarrow n}$

We are never able able to think about our *predictions* as *fact.*



Unsupervised Learning

Find *hidden structure* in data when Y is not formally observed.

Discover Z

$D = \{(X_i)\}_{i=1 \rightarrow n}$

Discrete Answer Space

$y \in \{1,2,\ldots C\}$ (i.e. Y is a "class")

$y = f(x) = \text{argmax } p(y = c \mid x, D)$

Continuous Answer Space

$y \in \mathbb{R}$ (i.e. Y is a Real number)

Course Project Interlude

- An “Applied ML” project
  - Finding a problem space that can be defined in ML terms
  - Finding a dataset that can be used to try and answer questions in that problem space
  - … Trying to answer the questions in that problem space

Project Milestones

- Milestone 1: Group Formation and Problem Scoping
- Milestone 2: Formal Pitch
- Milestone 2.1: Pitch Feedback
- Milestone 3: Midway Update
- Milestone 4: Final Report + Presentation

Project Milestone 1

- Group Member Names
    - most groups should be 2 or 3 people
    - if you want to work alone, you will have to demonstrate why it is important / necessary (e.g. access to confidential research data).
- A team name (so we have something to call your group in Canvas - this can be changed later)
- Problem Space
- Data / Data Plan
    - if you don't already have access to your dataset you must discuss how you plan to have the data by the time you give your official pitch (Milestone 2).

Problem Space – College Admissions

The following scenario isn't really true, but it's close to what we do in college admissions…

I am trying to decide if a student should be admitted to my university. I have their SAT and ACT scores and their HS GPA. I also have the history of students who have attended in the past, their SAT / ACT / HS GPA as well as whether or not they graduated from my university. I only want to admit new students if they will graduate.

Problem Space – College Admissions

$X_i =$

$Y =$

My First ML Algorithm – K-Nearest Neighbors

Classifying a new / unknown student x:

Given my training set D, find the K students that are "nearest" to x and assign x to the label y held by the majority of those students.

What does "nearest" mean?

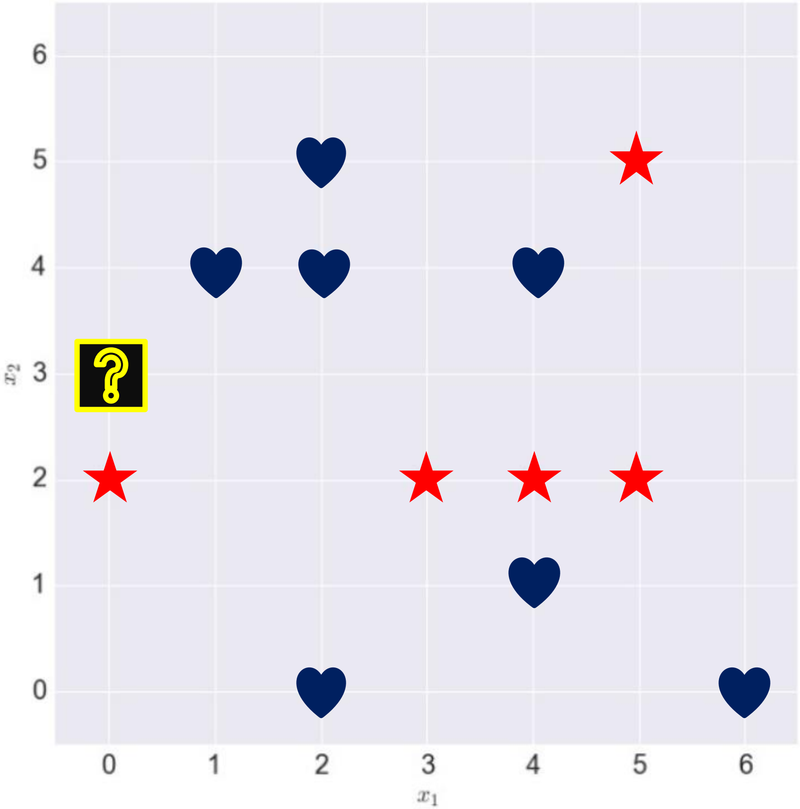
Prediction – College Admission

Student	SAT	ACT	GPA	Graduated?
A	1200	26	3.2	Yes
B	1450	28	3.5	Yes
C	1000	20	3.0	Yes
D	730	15	2.0	No
NEW	720	16	2.2	???

Prediction – College Admission

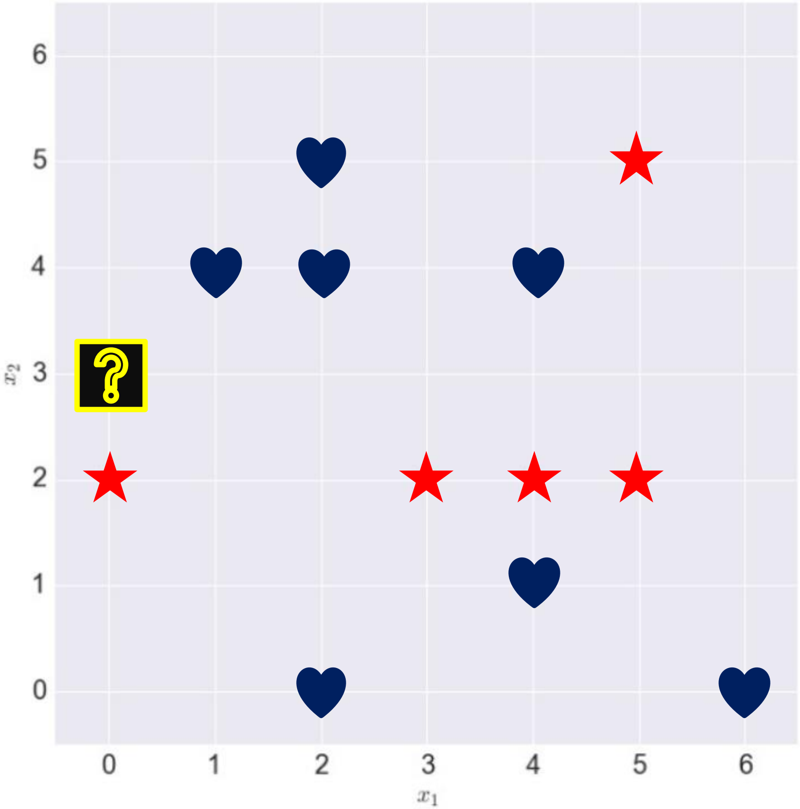
Student	X_1	X_2	Y
A	1200	26	1
B	1450	28	1
C	1000	20	1
D	730	15	-1
NEW	720	16	???

KNN – Which Point is Closest?



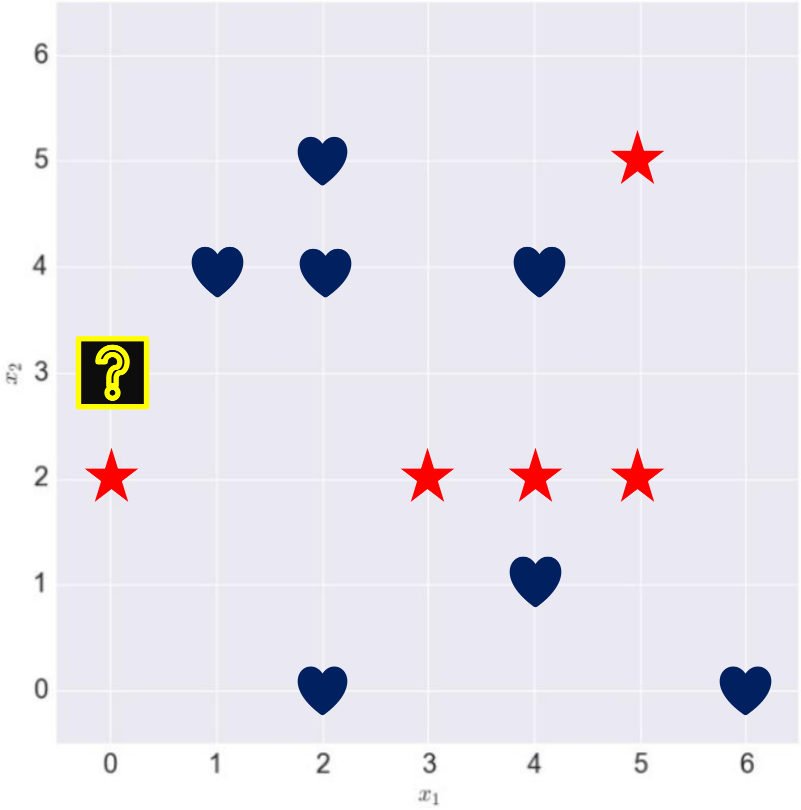
KNN – Manhattan/Taxicab Distance (L_1 Norm)

Manhattan Distance: $|x_i - x|$



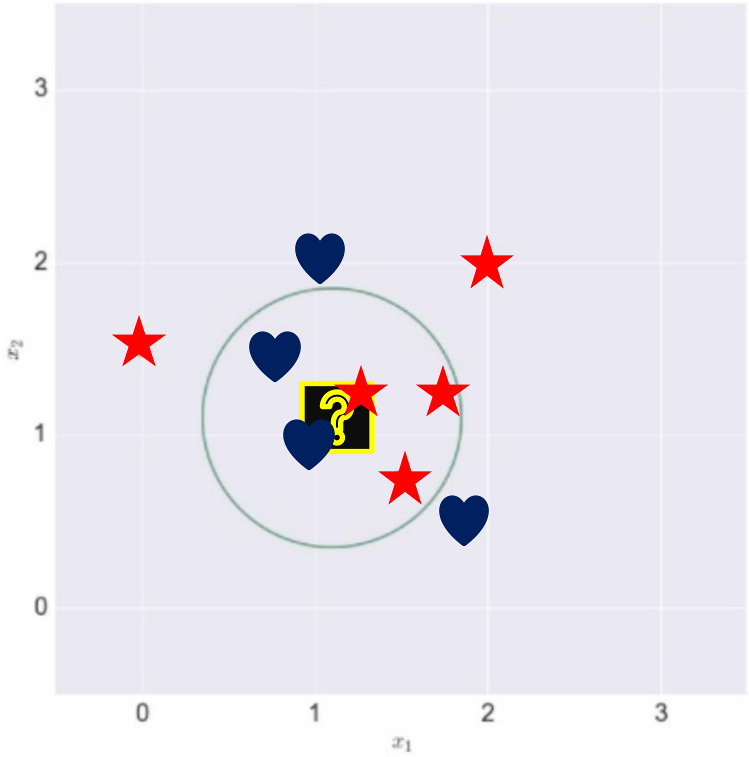
KNN – Euclidian Distance (L_2 Norm)

Euclidian Distance: $\sqrt{(\|x_i - x\|^2)}$



KNN – Euclidian Distance (L_2 Norm)

Euclidian Distance: $\sqrt{(\|x_i - x\|^2)}$

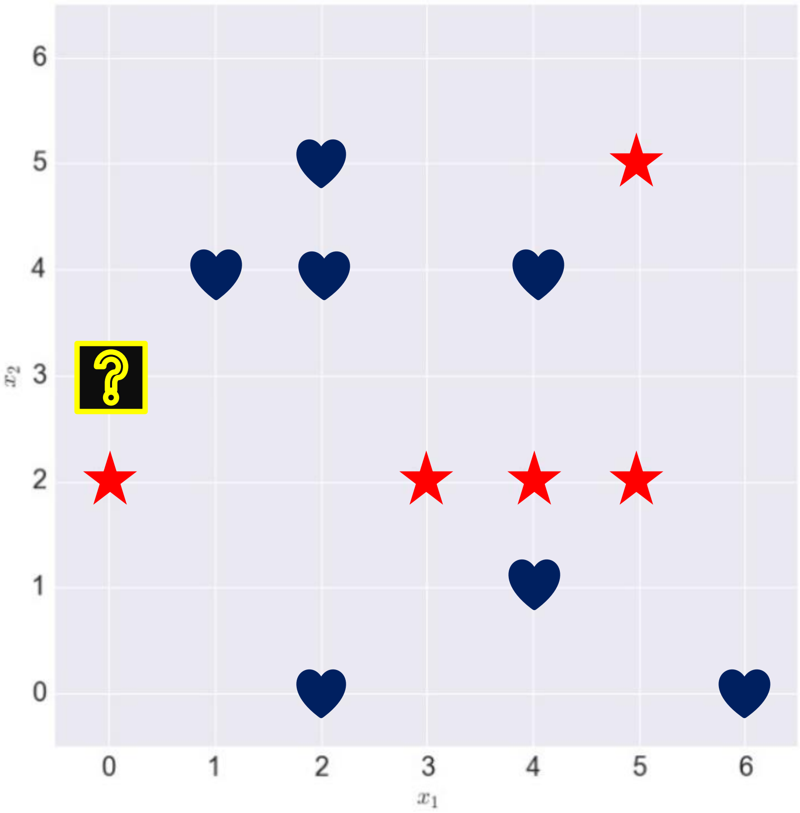


Find the Nearest Neighbors

K = 1, Point to classify at (0,3)

Nearest Neighbor

Prediction

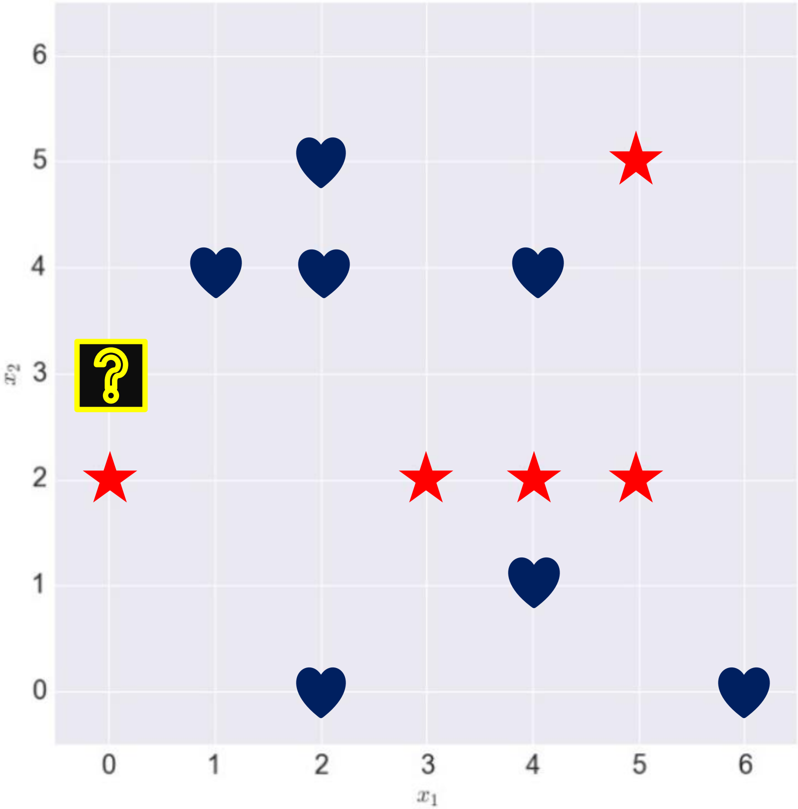


Find the Nearest Neighbors

K = 2

Nearest Neighbors

Prediction



Edge Case – K is even

This *often* messes up a K-Nearest Neighbors classification technique in a binary setting

What are you going to do?

Edge Case – K is even

This *often* messes up a K-Nearest Neighbors classification technique in a binary setting.

- Most common solution: K is an odd number

Edge Case – K is even

This *often* messes up a K-Nearest Neighbors classification technique in a binary setting.

- Most common solution: K is an odd number

(We'll explore a case in a few minutes where this may not generalize)

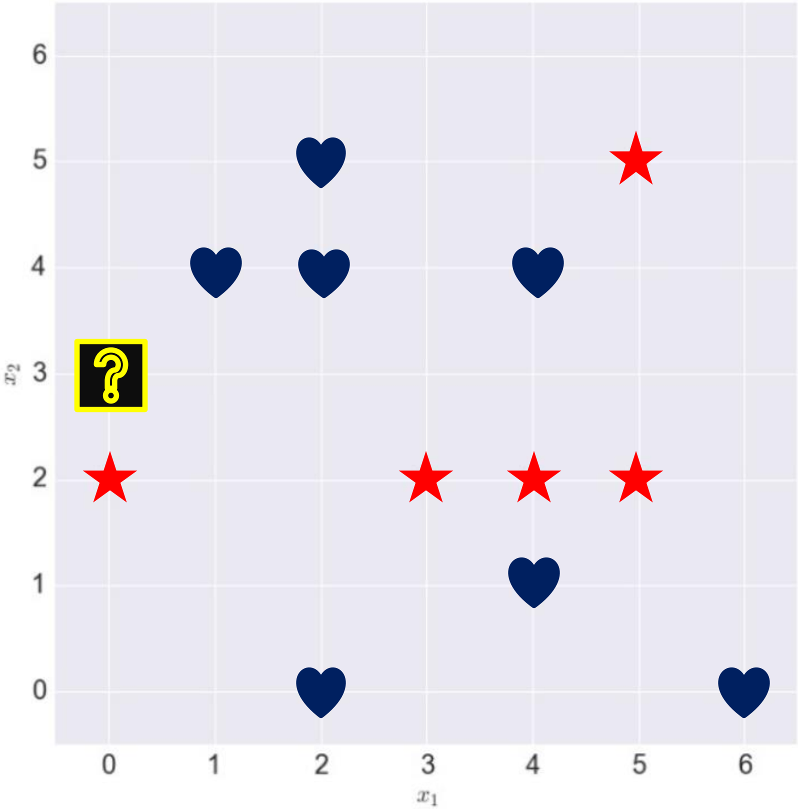
- A safety net: Set up defaults if there's a tie
    - A common default: Whichever case is more common
    - An ethical consideration: Whichever case is safer / more ethical

Find the Nearest Neighbors

K = 3

Nearest Neighbors

Prediction

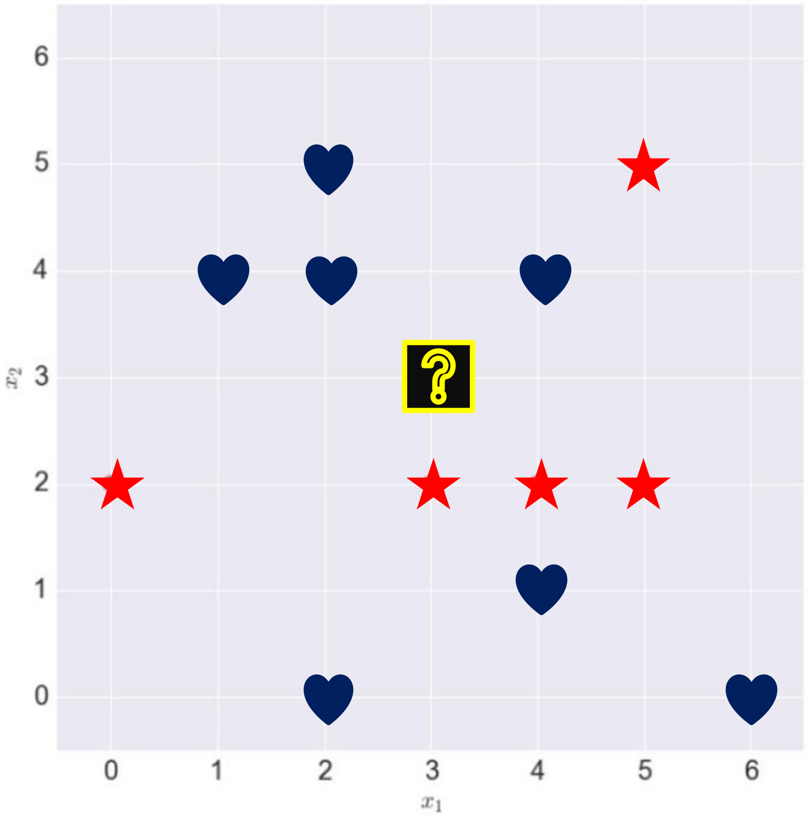


Find the Nearest Neighbors

K = 3

Nearest Neighbors

Prediction



Edge Case – Multiple cases equidistant

Sometimes you're looking for, say, the 3 nearest neighbors, but you end up finding there are 2 or 3 equally distant neighbors at that 3rd nearest distance.

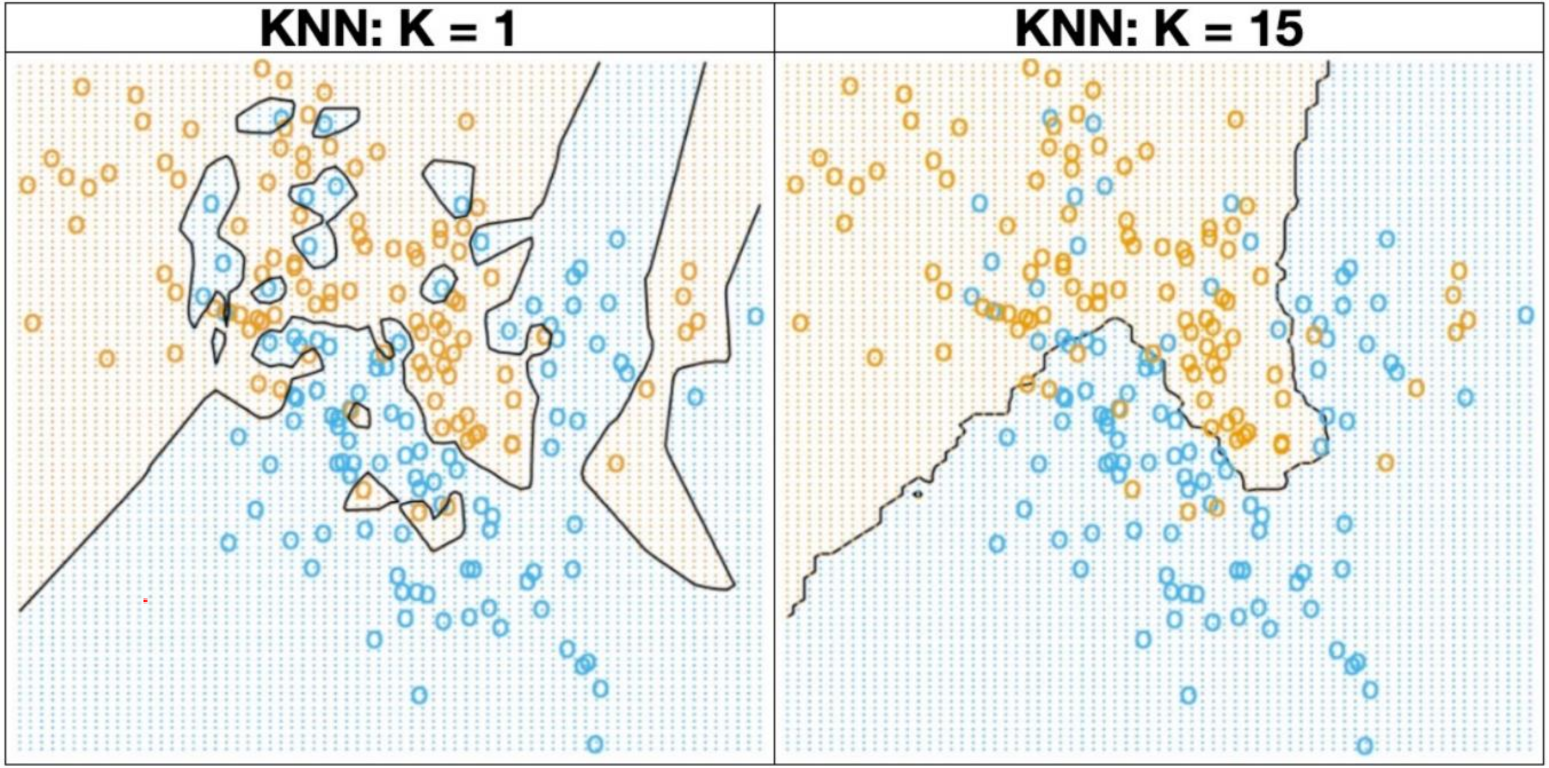
What are you going to do?

Edge Case – Multiple cases equidistant

Sometimes you're looking for, say, the 3 nearest neighbors, but you end up finding there are 2 or 3 equally distant neighbors at that 3rd nearest distance.

- Easiest answer: Whichever one you encounter first in memory!
- Other answers
    - Allow your K to be flexible
        - Will your answer change at K = 4 or K = 5?
    - Fall back on a default rule

My First ML Algorithm – KNN



KNN – Beyond Binary Decisions

Iris Classification

Iris Setosa

Iris Versicolor

Iris Virginica

https://en.wikipedia.org/wiki/Iris_flower_data_set

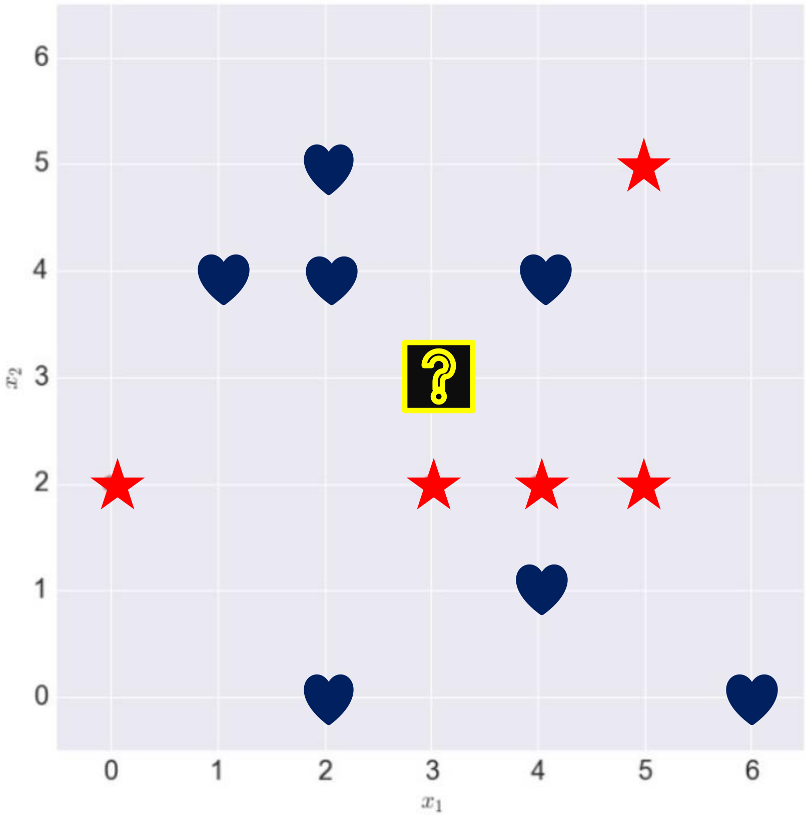
KNN – Beyond Binary Decisions

What are some questions / issues / considerations you see arising in this problem space?

https://en.wikipedia.org/wiki/Iris_flower_data_set

KNN – Beyond Binary Decisions

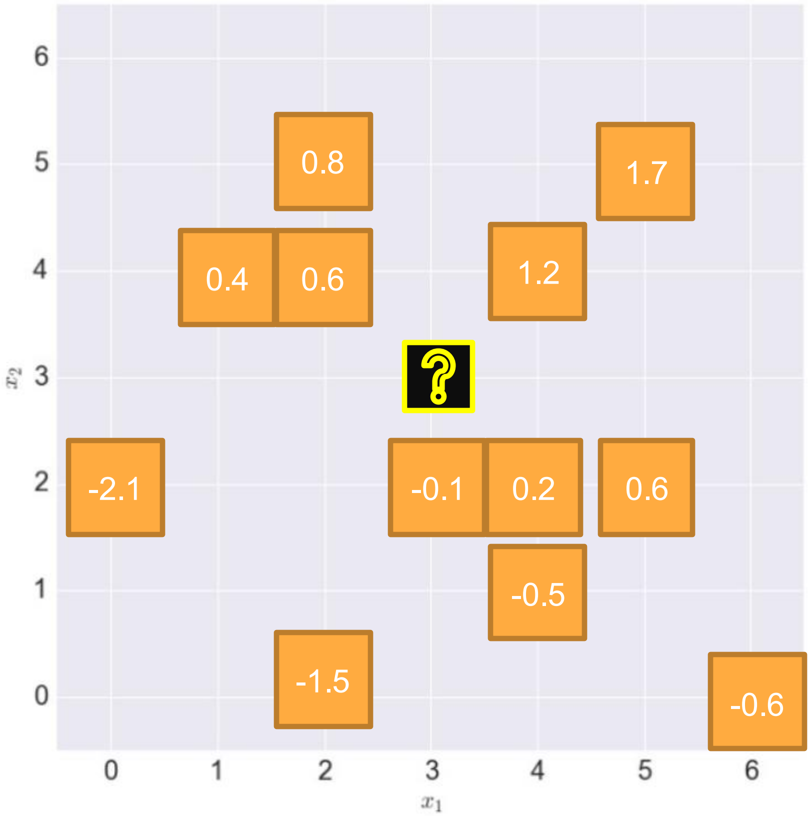
What if the "label" Y we are assigning is not a "class"...



KNN – Beyond Binary Decisions

What if the "label" Y we are assigning is not a "class"...

But is a numeric value?

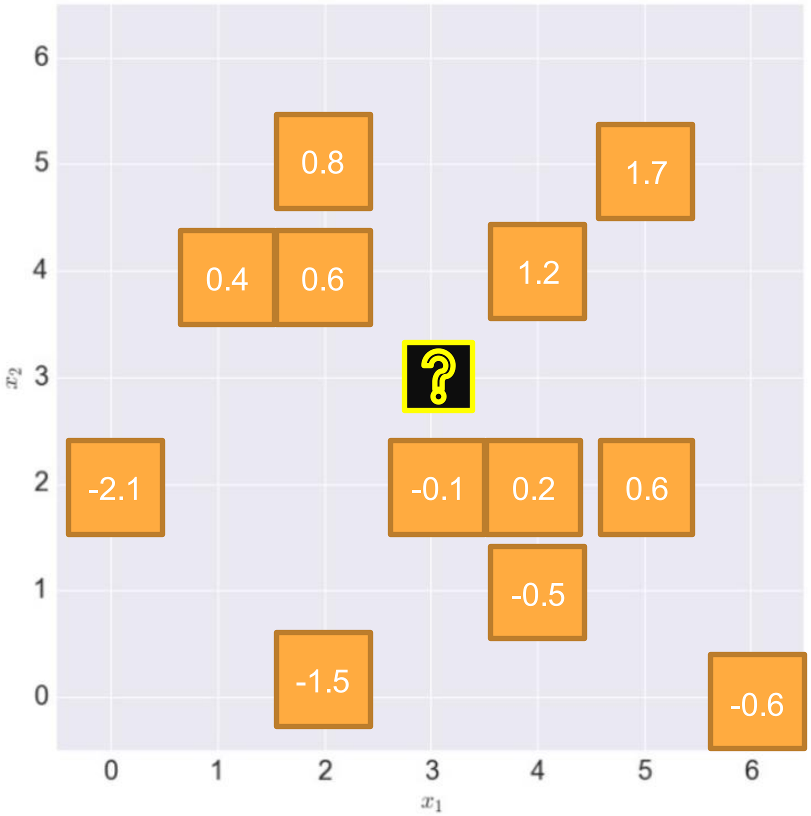


KNN – Beyond Binary Decisions

What if the "label" Y we are assigning is not a "class"...

But is a numeric value?

Find K neighboring points and find the average of their Y values!

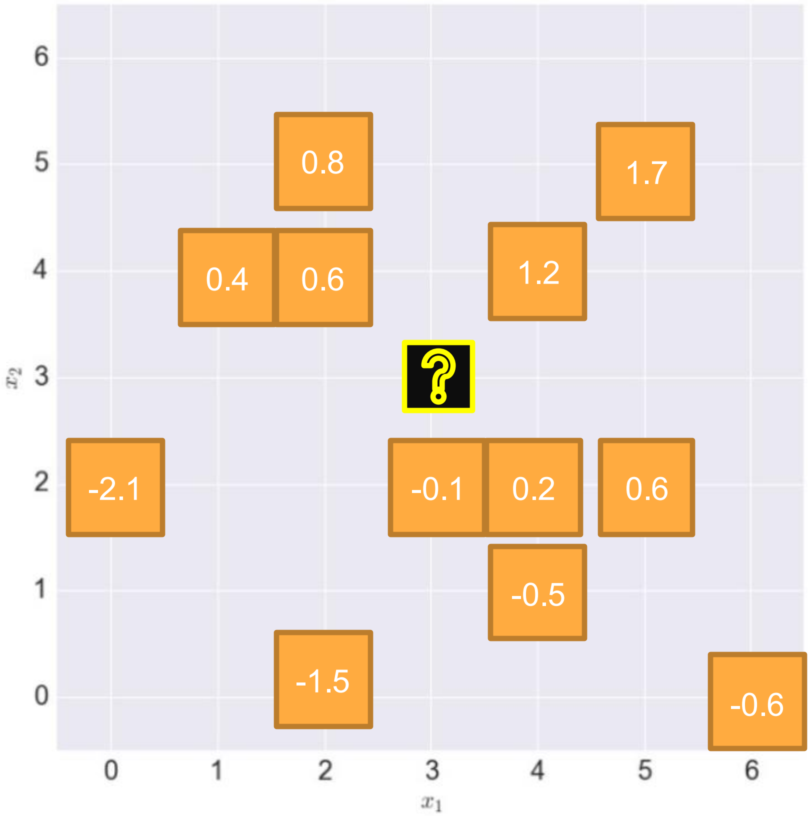


KNN – Beyond Binary Decisions

K = 1, Point to regress at (3,3)

Nearest Neighbor

Prediction

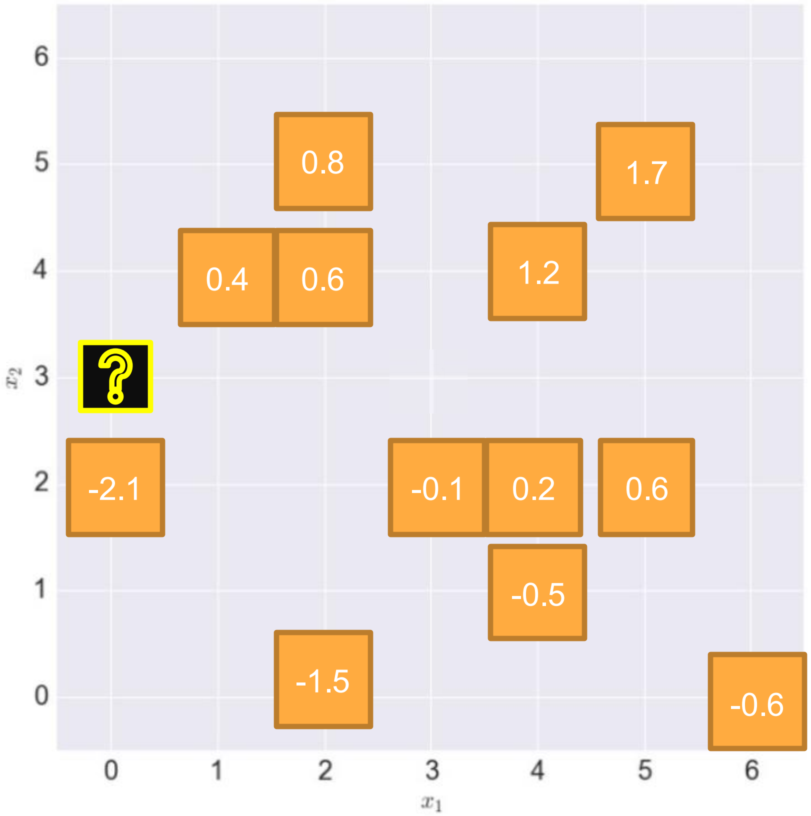


KNN – Beyond Binary Decisions

K = 2, Point to regress at (0,3)

Nearest Neighbor

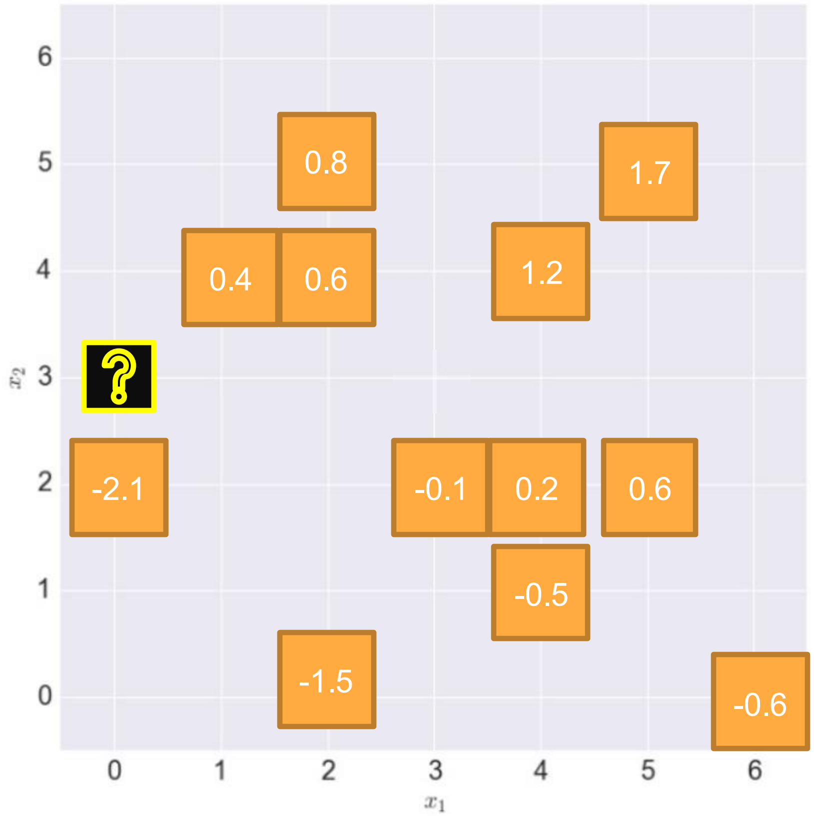
Prediction



KNN – Beyond Binary Decisions

K = 2, Point to regress at (0,3)

Since it is no longer a voting scheme, we no longer worry about ties!

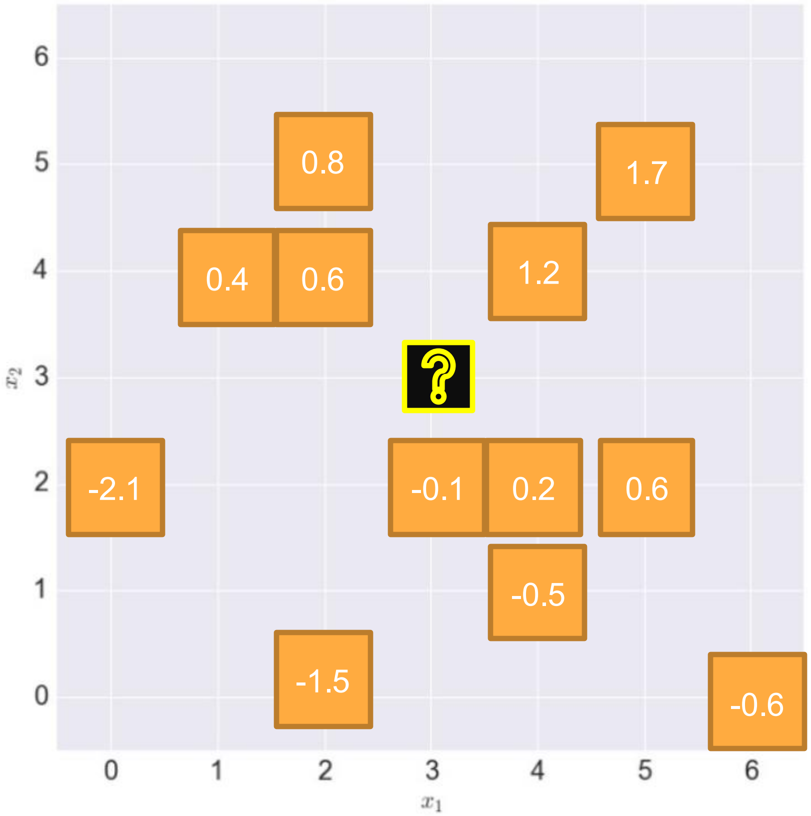


KNN – Beyond Binary Decisions

K = 2, Point to regress at (3,3)

Nearest Neighbor

Prediction

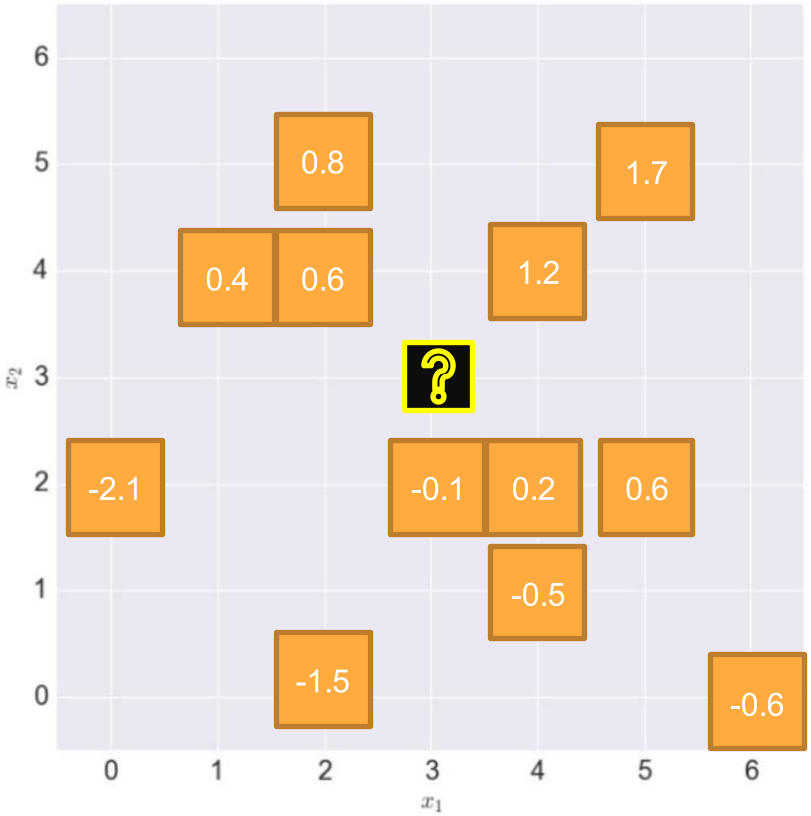


KNN – Beyond Binary Decisions

K = 2, Point to regress at (0,3)

We still have an equidistance problem!

But is it a problem?



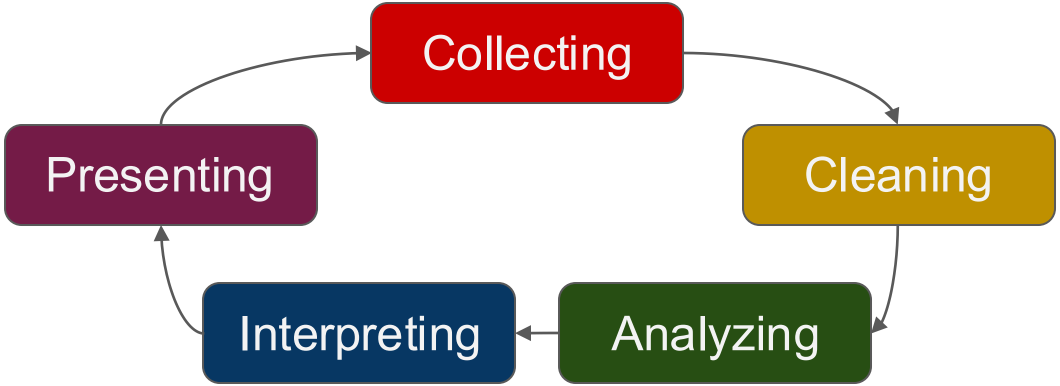
Edge Case – Multiple cases equidistant

Sometimes you're looking for, say, the 3 nearest neighbors, but you end up finding there are 2 or 3 equally distant neighbors at that 3rd nearest distance.

- Easiest answer: Whichever one you encounter first in memory!
- Other answers
    - Allow your K to be flexible
        - Will your answer change at K = 4 or K = 5?
    - Fall back on a default rule

Start-To-Finish ML

Data Analytics Cycle



Adapted from Clow 2012: https://dl.acm.org/doi/pdf/10.1145/2330601.2330636
& Labrinidis & Jagadish, 2012 (as interpreted by Gandomi & Haider, 2015): https://www.sciencedirect.com/science/article/pii/S0268401214001066

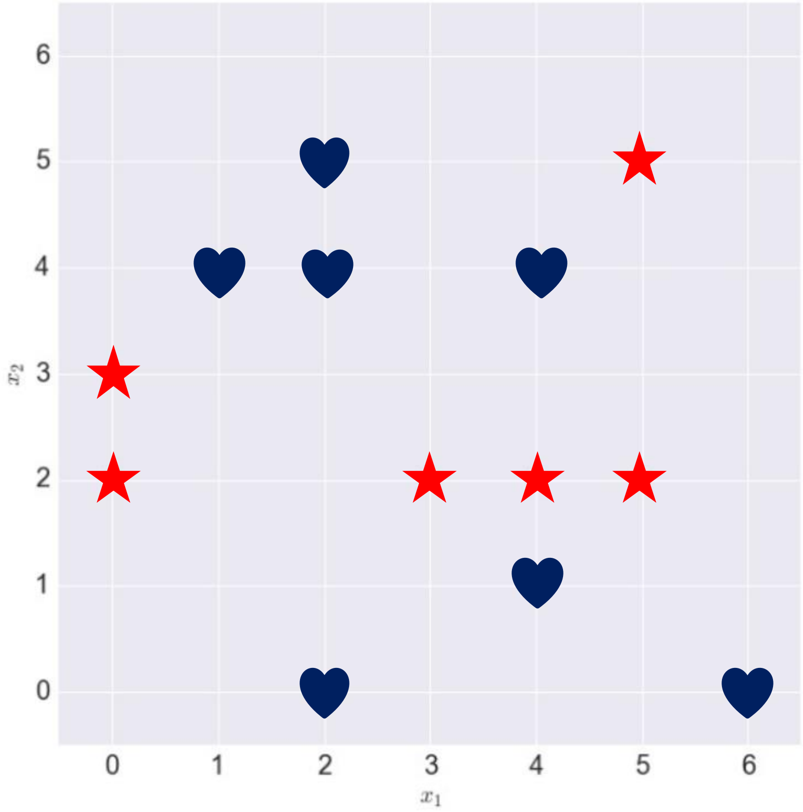
Collecting Data

- Where do “Data” come from?

Our Toy Dataset

X_1 and X_2 are our features X

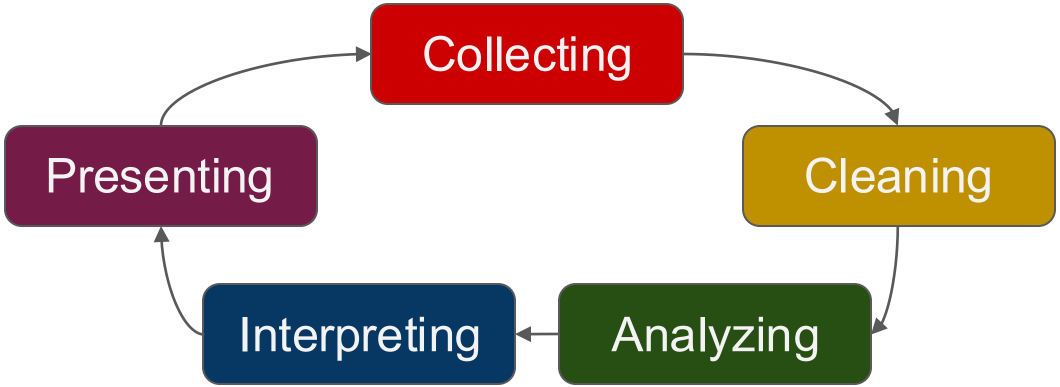
The Shape (Star or Heart) is our outcome Y



Prediction – College Admission

Student	X_1	X_2	Y
A	1200	26	1
B	1450	28	1
C	1000	20	1
D	730	15	-1

Data Analytics Cycle



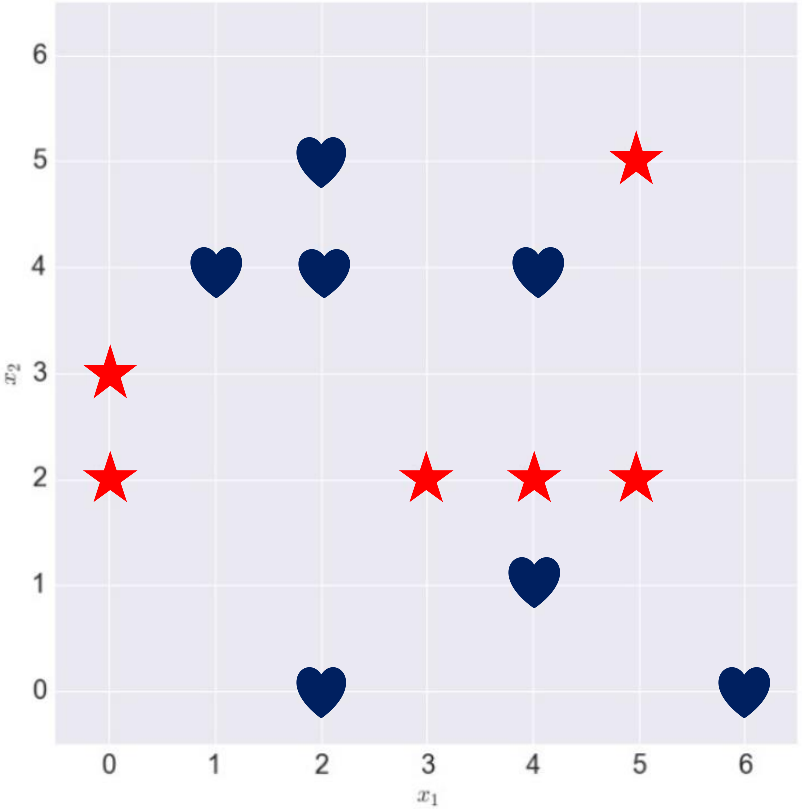
Adapted from Clow 2012: https://dl.acm.org/doi/pdf/10.1145/2330601.2330636
& Labrinidis & Jagadish, 2012 (as interpreted by Gandomi & Haider, 2015): https://www.sciencedirect.com/science/article/pii/S0268401214001066

Our Toy Dataset

X_1 and X_2 are our features X

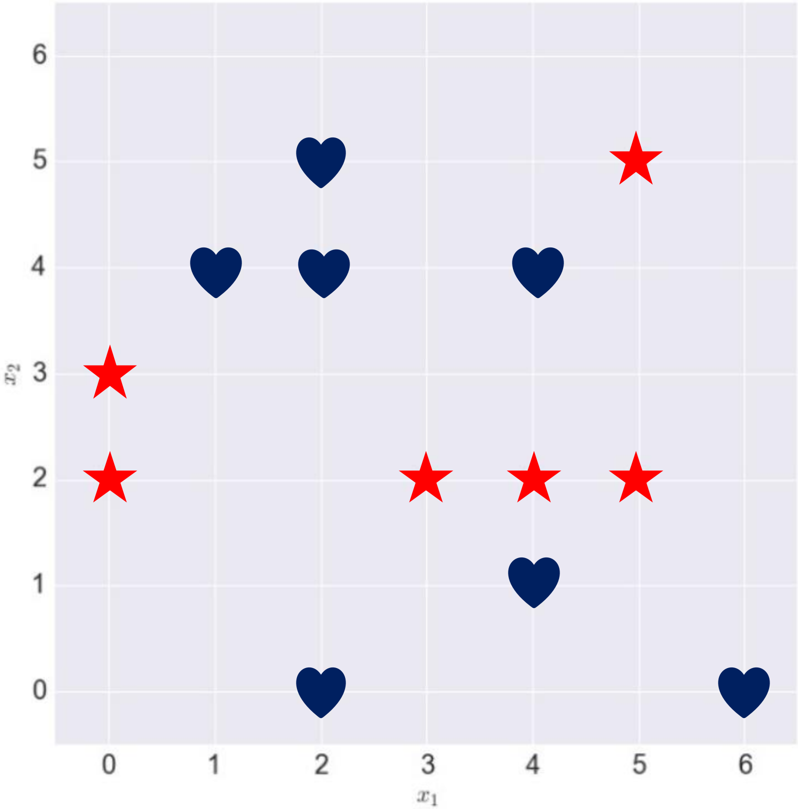
The Shape (Star or Heart) is our outcome Y

Is this going to be easy to feed into Python?



Our Toy Dataset

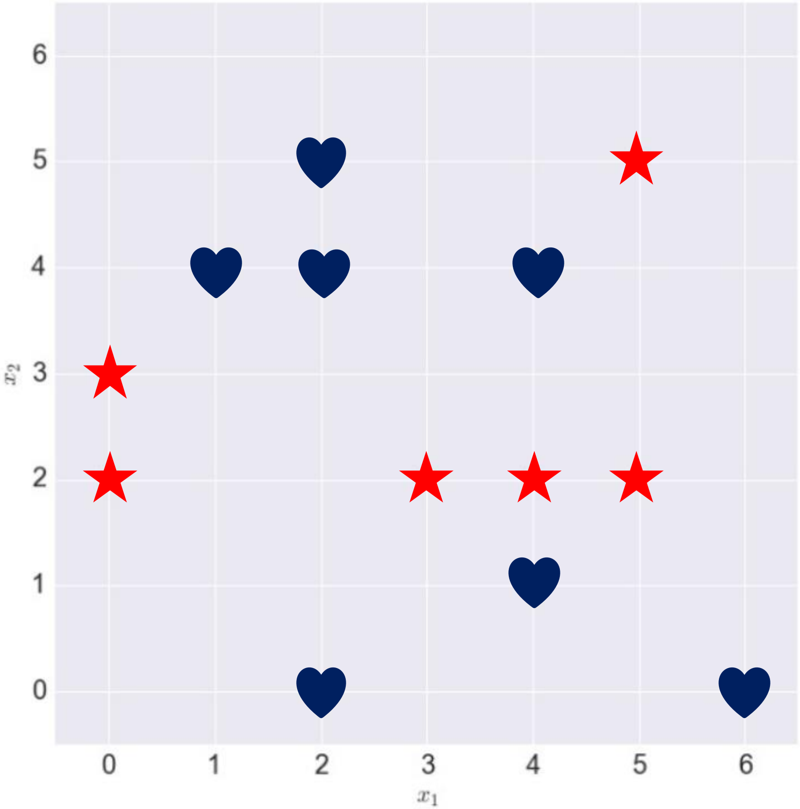
Represent your data in a Table (/ Data Frame / etc.)



Our Toy Dataset

features = [
[0,2],
[0,3],
[1,4],
[2,0],
...
[6,0]
]

labels = [
star,
star,
heart,
heart,
...
heart
]



Prediction – College Admission

What are you going to predict for this case with K = 1, using Manhattan distance?

Student	X_1	X_2	Y
A	1200	26	1
B	1450	28	1
C	1000	20	1
D	730	15	-1
NEW	720	16	???

Prediction – College Admission

What are you going to predict for this NEW case with K = 1, using Manhattan distance?

Student	X_1	X_2	Y
A	1200	26	1
B	1450	28	1
C	1180	20	1
D	730	15	-1
NEW	950	30	???

Prediction – College Admission

The X_1 feature is overpowering the X_2 feature!

Student	**X_1**	**X_2**	**Y**
A	1200	26	1
B	1450	28	1
C	1180	20	1
D	730	15	-1
NEW	950	30	???

Euclidian Distance – N-Dimensional Vector

Euclidian Distance: $||x_i - x||^2$

X_1	X_2	X_3	X_4	Y	Distance
2	5	9832	.005	Positive	
4	82	9421	.008	Positive	
3	17	9321	.04	Negative	
4	90	9128	.001	Negative	

Euclidian Distance – N-Dimensional Vector

Euclidian Distance: $||x_i - x||^2$

X_1	X_2	X_3	X_4	Y	Distance
2	5	9832	.005	Positive	
4	82	9421	.008	Positive	
3	17	9321	.04	Negative	
4	90	9128	.001	Negative	
3	16	9830	.04	???	

Euclidian Distance – N-Dimensional Vector

Euclidian Distance: $||x_i - x||^2$

X_1	X_2	X_3	X_4	Y	Distance
2	5	9832	.005	Positive	126.001
4	82	9421	.008	Positive	171638.001
3	17	9321	.04	Negative	259082
4	90	9128	.001	Negative	498281.0015
3	16	9830	.04	???	

Euclidian Distance – N-Dimensional Vector

Euclidian Distance: $||x_i - x||^2$

X_1	X_2	X_3	X_4	Y	Distance
2	5	9832	.005	Positive	126.001
4	82	9421	.008	Positive	171638.001
3	17	9321	.04	Negative	259082
4	90	9128	.001	Negative	498281.0015
3	16	9830	.04	Positive?	

Min-Max Scaling

Transform X (Data) to X' (Scaled Data)

For (x_i) in X

scale = $\max(x_i) - \min(x_i)$

low = $\min(x_i)$

For $(x_{i,j})$ in (x_i)

$x'_{i,j} = (x_{i,j} - \text{low}) / \text{scale}$

https://en.wikipedia.org/wiki/Feature_scaling

Scaling – N-Dimensional Vector

Euclidian Distance: $||x_i - x||^2$

X_1	X_2	X_3	X_4	Y	Distance
0	0	1	0.103	Positive	1.07
1	0.906	0.416	0.179	Positive	1.87
.5	0.141	0.274	1	Negative	0.52
1	1	0	0	Negative	3.00
.5	0.129	0.997	1	Negative	

Scaling – Housing Market?

Euclidian Distance: $||x_i - x||^2$

# Bedrms	Acres	Sq. Ft.	Radon	New Build?	Distance
2	5	9832	.005	Positive	
4	82	9421	.008	Positive	
3	17	9321	.04	Negative	
4	90	9128	.001	Negative	

Scaling – College Prediction

Student	SAT	ACT	GPA	Graduated?
A	1200 / 1600	26 / 36	3.2 / 4.0	Yes
B	1450 / 1600	28 / 36	3.5 / 4.0	Yes
C	1000 / 1600	20 / 36	3.0 / 4.0	Yes
D	730 / 1600	15 / 36	2.0 / 4.0	No
NEW	720 / 1600	16 / 36	2.2 / 4.0	???

Normalization

Transform X (Data) to X' (Scaled Data)

For (x_i) in X

    mean = avg (x_i)

    dev = stdev (x_i)

    For $(x_{i,j})$ in (x_i)

        $x'_{i,j} = (x_{i,j} - \text{mean}) / \text{dev}$

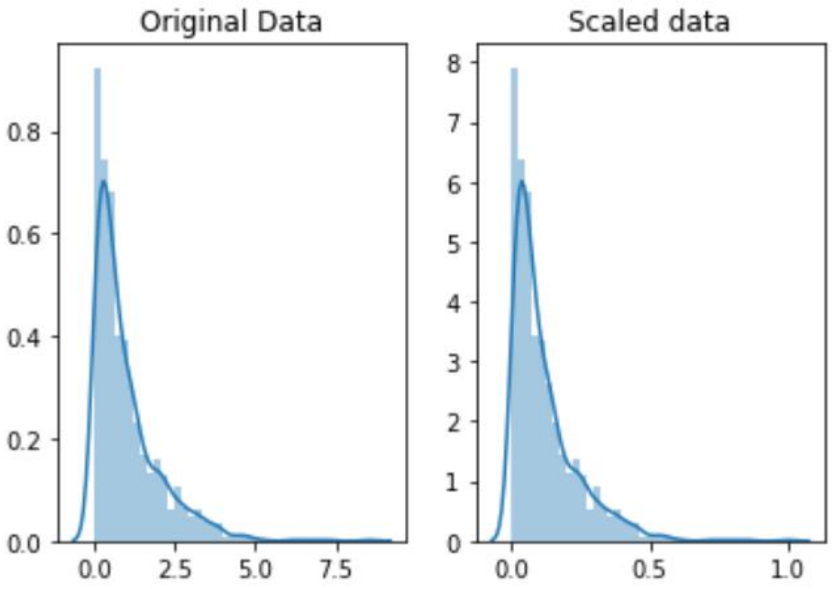
https://en.wikipedia.org/wiki/Feature_scaling

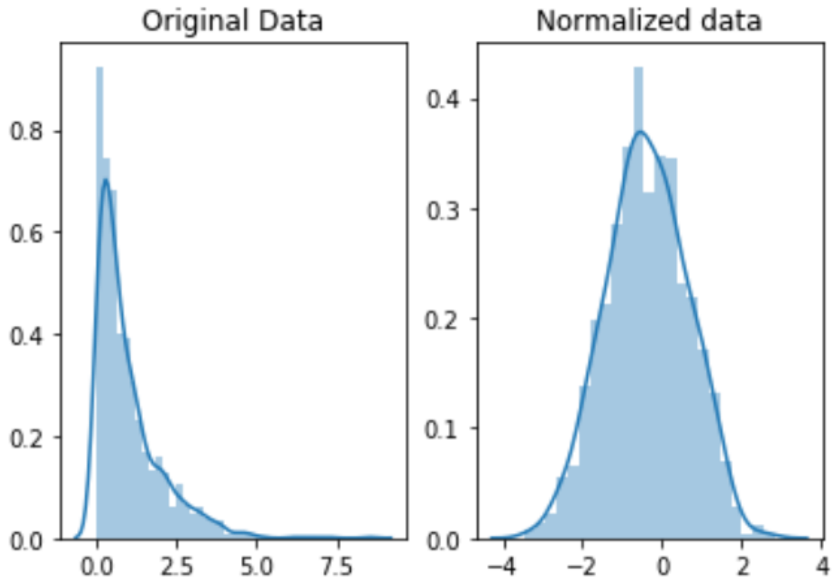
Euclidian Distance – N-Dimensional Vector

Euclidian Distance: $||x_i - x||^2$

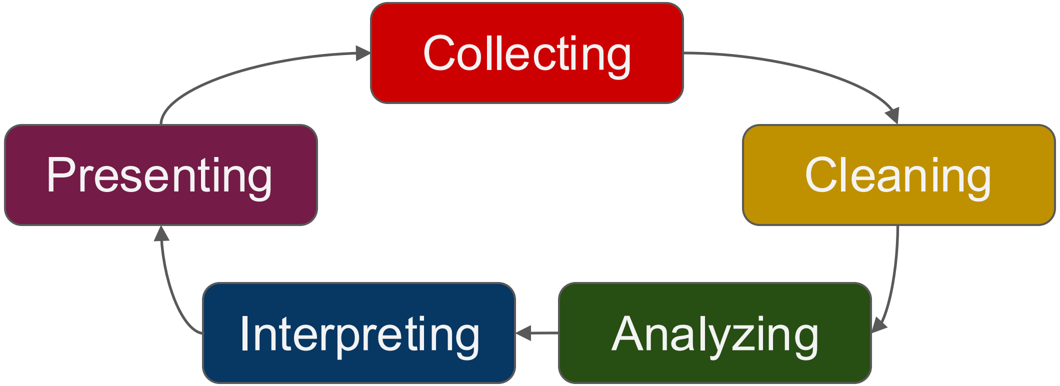
X_1	X_2	X_3	X_4	Y	Distance
-1.30558	-0.9954294	1.368526	-0.4749171	Positive	
0.7833494	0.7665950	-0.0151497	-0.3072993	Positive	
-0.261116	-0.7208282	-0.351810	1.480624	Negative	
0.7833494	0.9496625	-1.001566	-0.6984075	Negative	
-0.2611164	-0.7437116	1.3617933	1.480624	???	

Side Note: Feature Scaling vs. Normalizing





Data Analytics Cycle



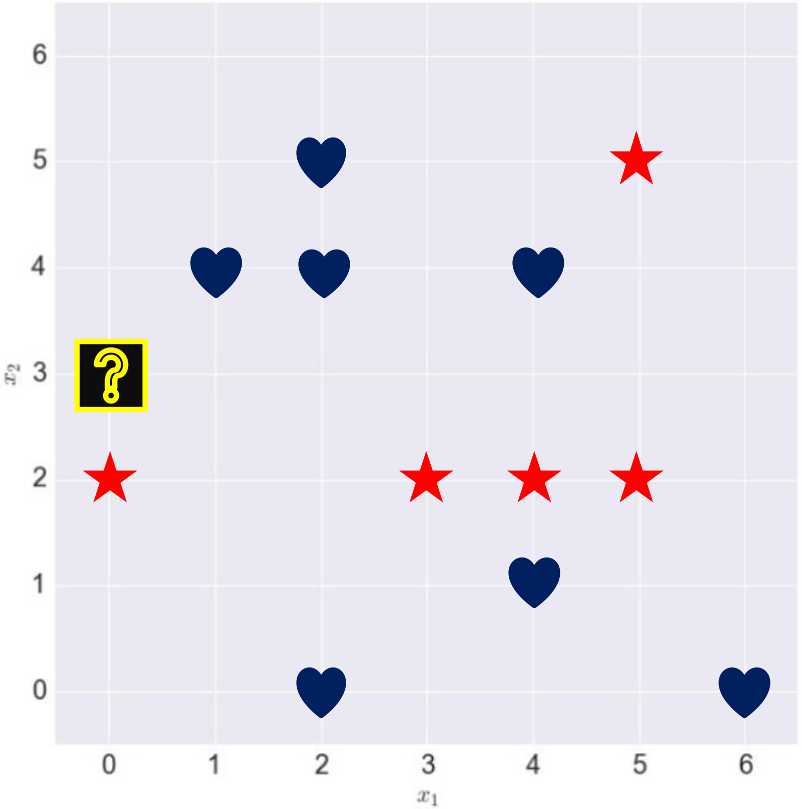
Adapted from Clow 2012: https://dl.acm.org/doi/pdf/10.1145/2330601.2330636
& Labrinidis & Jagadish, 2012 (as interpreted by Gandomi & Haider, 2015): https://www.sciencedirect.com/science/article/pii/S0268401214001066

Find the Nearest Neighbors

Classification

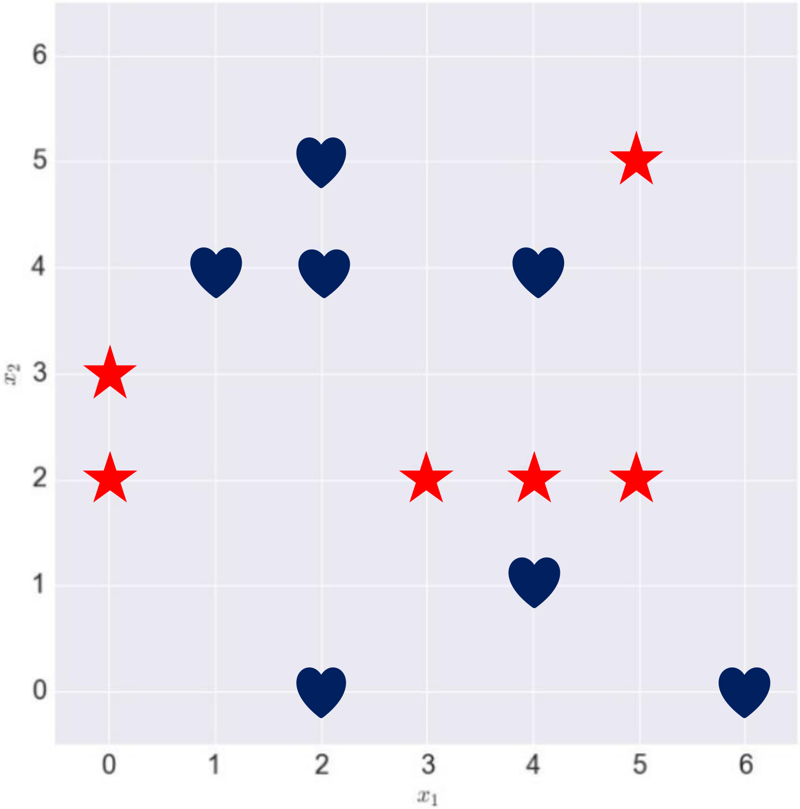
Our objective – Predicting a class

Did we get it right?
How do we know we got it right?

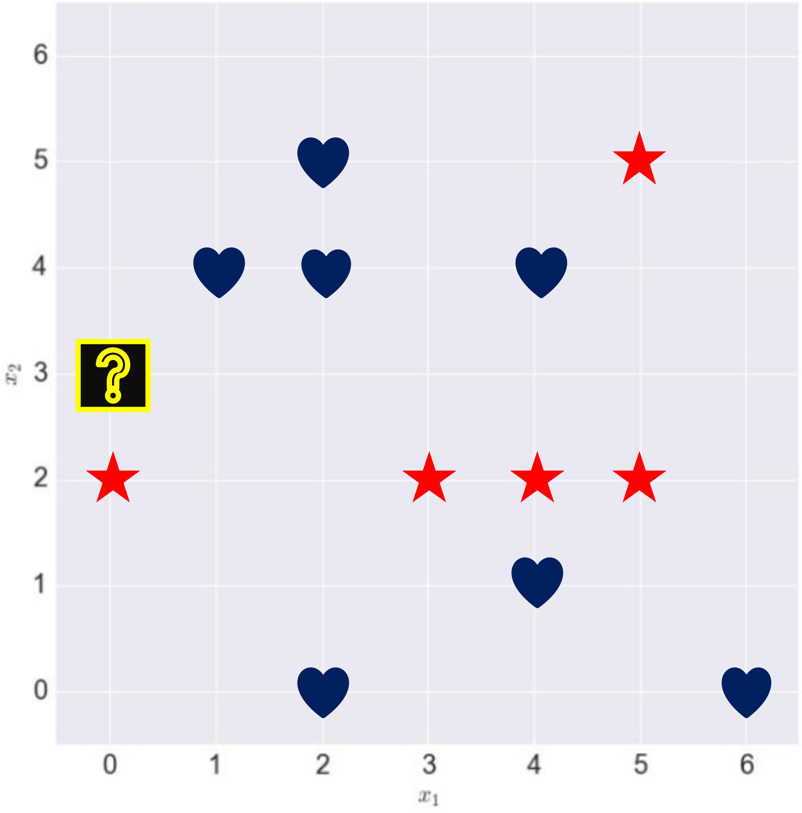


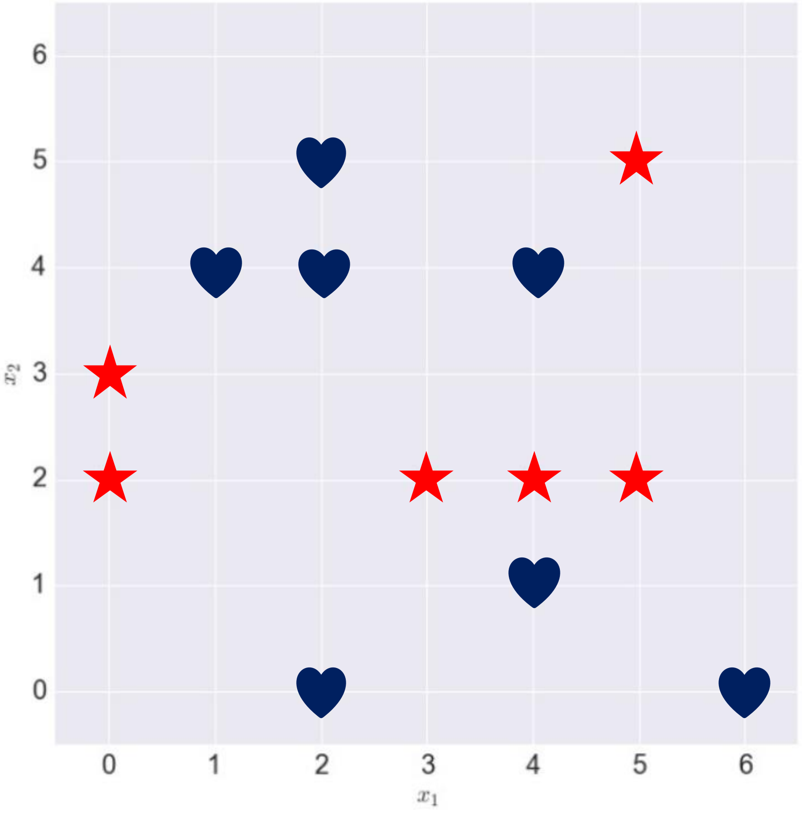
How do we know if it works?

Classify ones we already know the answer to!

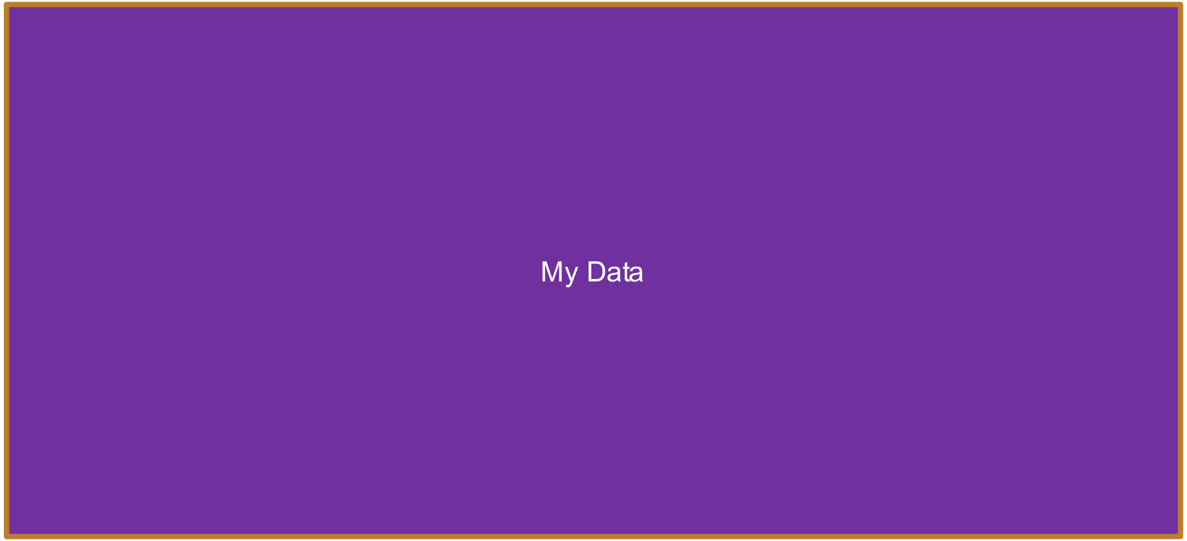


Accuracy





Homework 1 – Training & Test Sets



Train a KNN on everything

K = 1

What does my nearest neighbor to X look like?

Homework 1 – Training & Test Sets

Divide it into training and testing sets!

Types of Errors

Classified As *Ground Truth*	**C**	**~C**
C		
~C		

Types of Errors

Classified As *Ground Truth*	**C**	**~C**
C	True Positive (Hit)	False Negative (Miss)
~C	False Positive (False Alarm)	True Negative (Correct Rejection)

Types of Errors

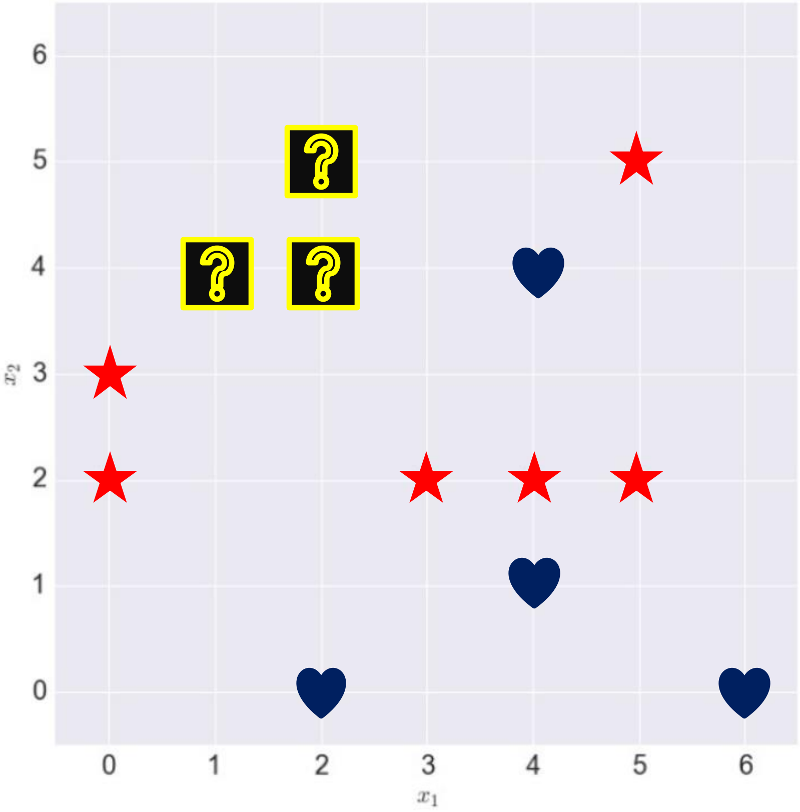
Types of Errors

Classified As *Ground Truth*	**A**	**B**	**C**
A			
B			
C			

Types of Errors

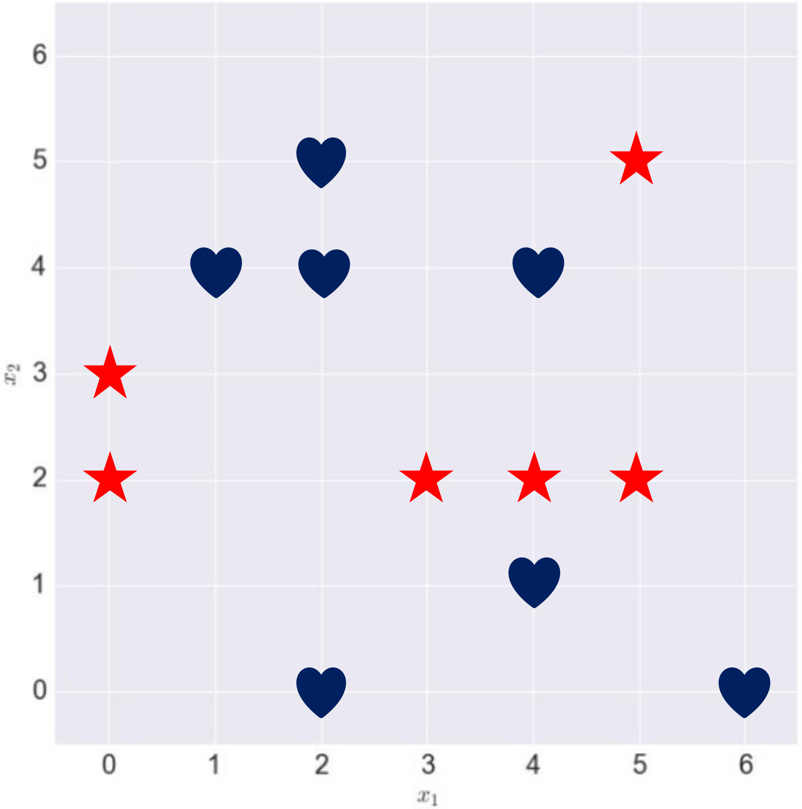
Classified As / *Ground Truth*	**A**	**B**	**C**
A	hit	miss	miss
B	miss	hit	miss
C	miss	miss	hit

How do we know if it works?



How do we know if it works?

Uh oh...

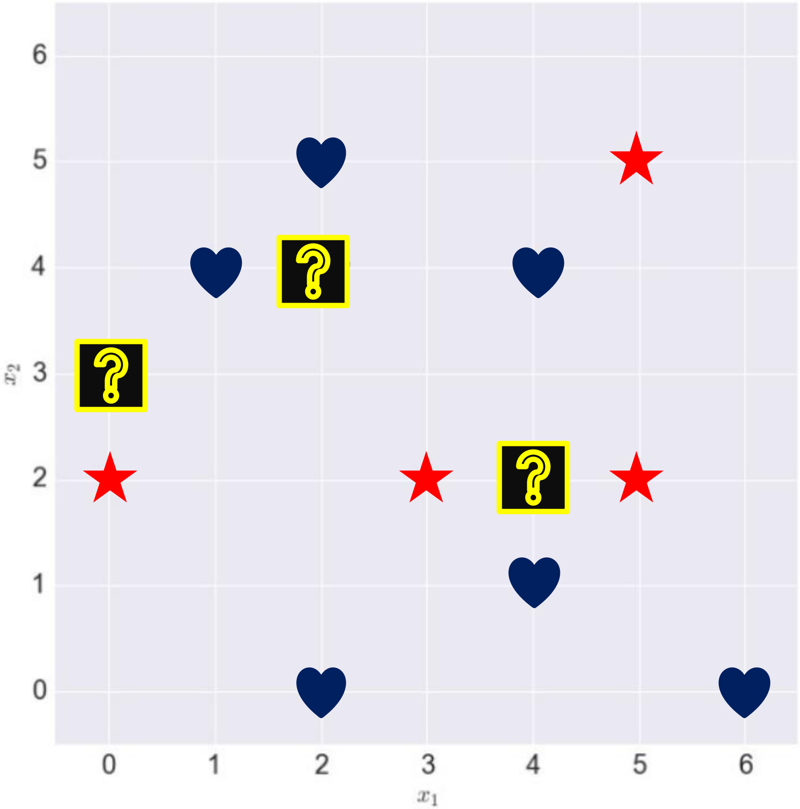


Homework 1 – Training & Test Sets

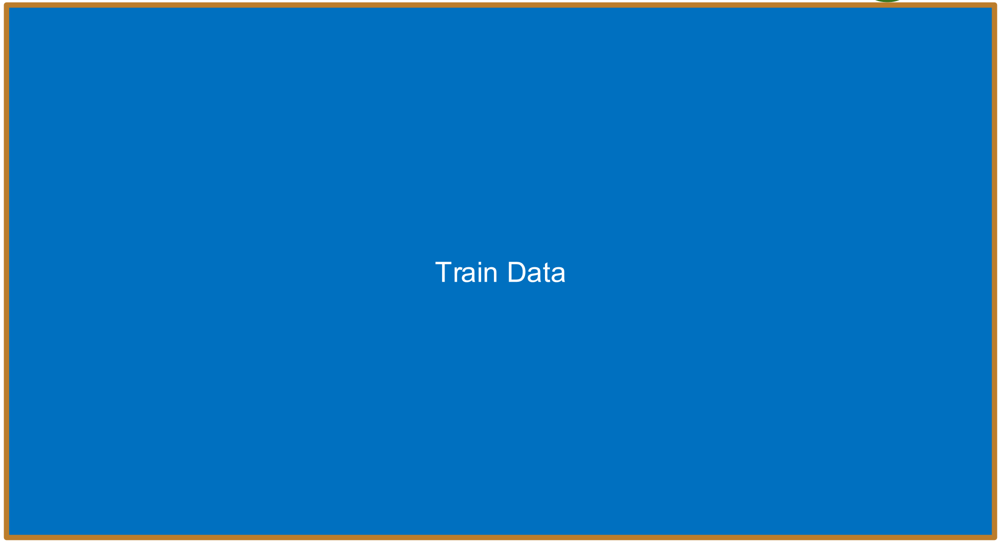
I pull out the last 20% of samples

But what if they were put in order?

How do we know if it works?



Homework 1 – Training & Test Sets



I pull out a random 20% of my data
Now I have something (probably) representative, AND
I'm not just testing inherent bias of my model or data

Euclidian Distance – N-Dimensional Vector

Euclidian Distance: $||x_i - x||^2$

X_1	X_2	X_3	X_4	Y	Distance
-1.30558	-0.9954294	1.368526	-0.4749171	Positive	
0.7833494	0.7665950	-0.0151497	-0.3072993	Positive	
-0.261116	-0.7208282	-0.351810	1.480624	Negative	
0.7833494	0.9496625	-1.001566	-0.6984075	Negative	
-0.2611164	-0.7437116	1.3617933	1.480624	???	

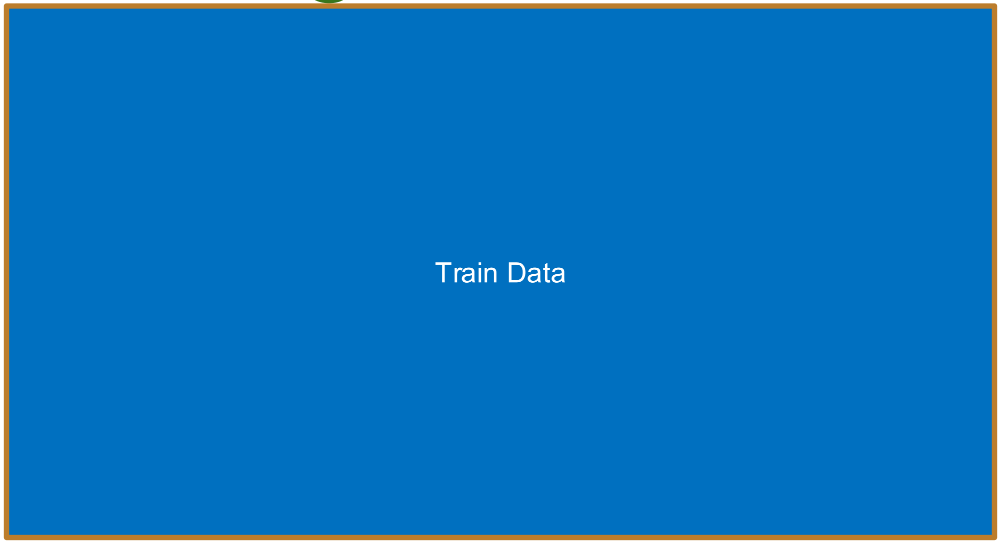
Where do we get our Mean and St. Dev.?

Euclidian Distance – N-Dimensional Vector

Euclidian Distance: $||x_i - x||^2$

	X_1	X_2	X_3	X_4	Y	Distance
	2	5	9832	.005	Positive	
Train	4	82	9421	.008	Positive	
	3	17	9321	.04	Negative	
	4	90	9128	.001	Negative	
Test	3	16	9830	.04	???	

Training & Test Sets

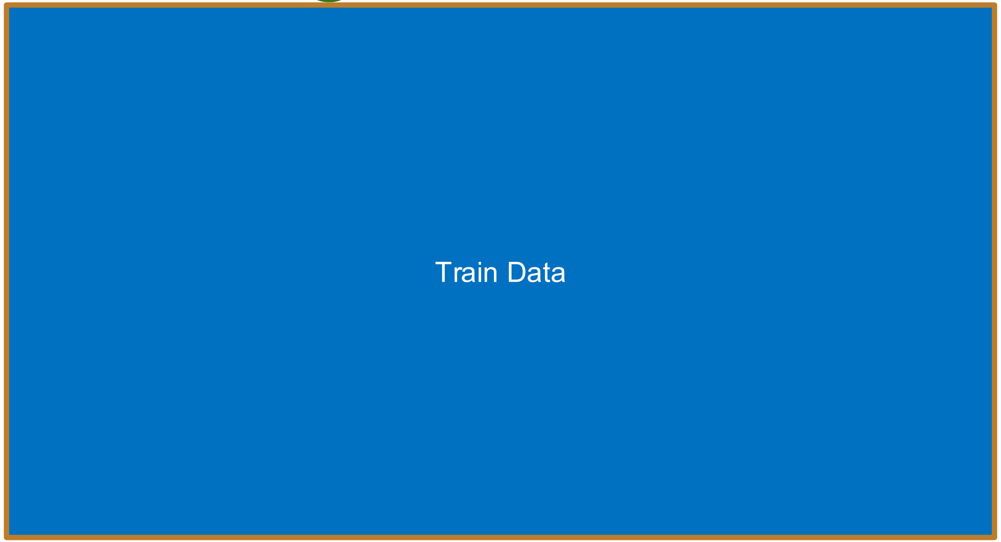


I pull out a random 20% of my data

Now I have something (probably) representative, AND

I'm not just testing inherent bias of my model or data

Training & Test Sets



I have sold a variety of houses (train data)

I develop data transformations based on these data

I apply the same transformations to future samples (test data)

Min-Max Scaling

Transform X_train (Train Data) to X_train' (Scaled Train Data)

For x_i (each column) in X_train

    scale = $\max(x_i) - \min(x_i)$

    low = $\min(x_i)$

    For $x_{i,j}$ (sample) in x_i

        $x'_{i,j} = (x_{i,j} - \text{low}) / \text{scale}$

Transform X_test (Test Data) to X_test' (Scaled Test Data)

For (x_i) in X_test

    For $(x_{i,j})$ in (x_i)

        $x'_{i,j} = (x_{i,j} - \text{low}) / \text{scale}$

https://en.wikipedia.org/wiki/Feature_scaling

Normalization

Transform X_train (Data) to X_train' (Scaled Data)

For (x_i) in X

    mean = avg (x_i)

    dev = stdev (x_i)

    For $(x_{i,j})$ in (x_i)

        $x'_{i,j} = (x_{i,j} - \text{mean}) / \text{dev}$

Transform X_test (Test Data) to X_test' (Scaled Test Data)

For (x_i) in X_test

    For $(x_{i,j})$ in (x_i)

        $x'_{i,j} = (x_{i,j} - \text{mean}) / \text{dev}$

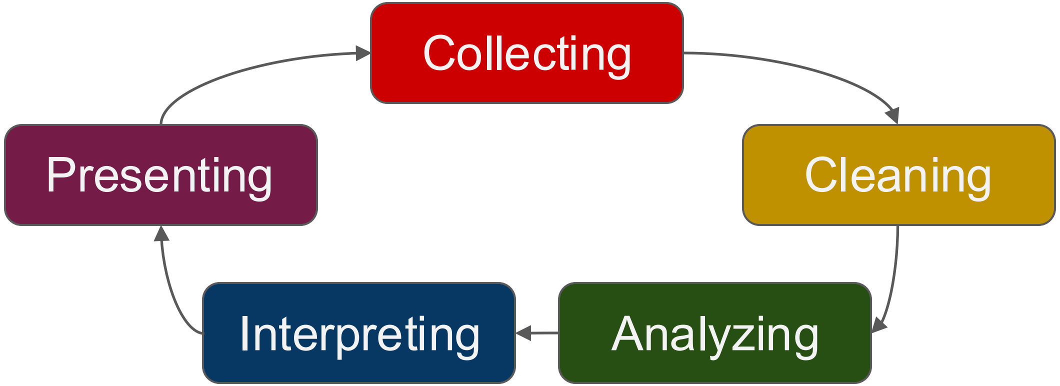
https://en.wikipedia.org/wiki/Feature_scaling

Normalization / Scaling

Normalization and Scaling are important for us to consider
- It will allow us to consider variables on equal footing

Normalization and Scaling are important for us to consider *on a case by case basis*
- Sometimes a “default” transformation won’t make sense

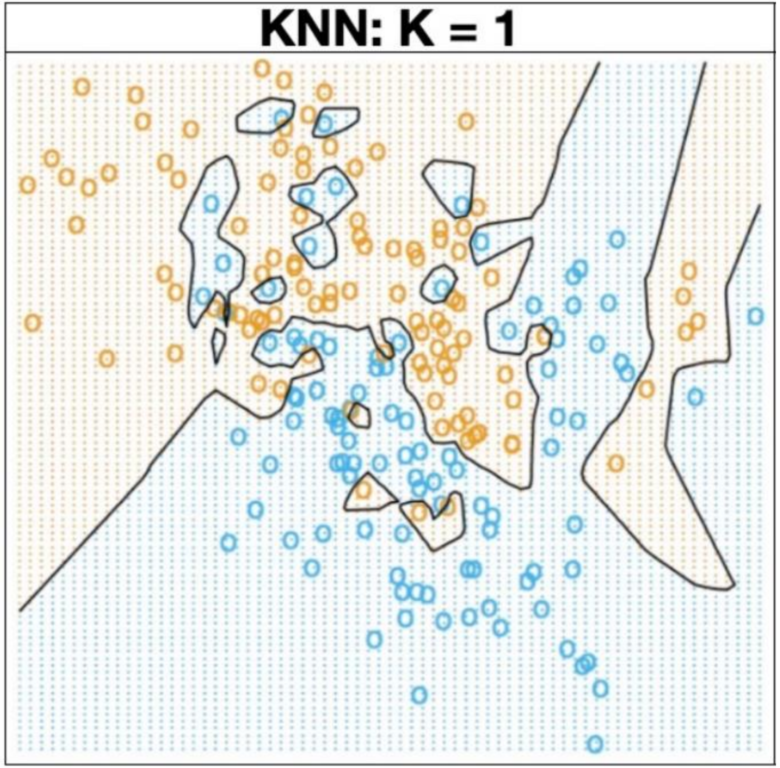
Data Analytics Cycle



Adapted from Clow 2012: https://dl.acm.org/doi/pdf/10.1145/2330601.2330636
& Labrinidis & Jagadish, 2012 (as interpreted by Gandomi & Haider, 2015): https://www.sciencedirect.com/science/article/pii/S0268401214001066

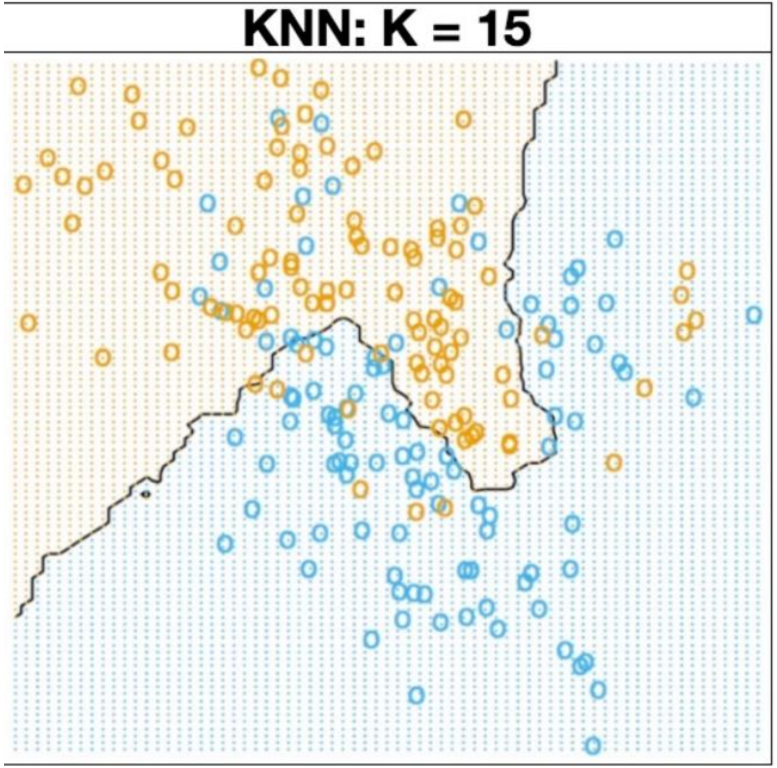
KNN – Small K

- What do you notice about this space?
  - Any odd cases (outliers) still have a significant impact on the prediction



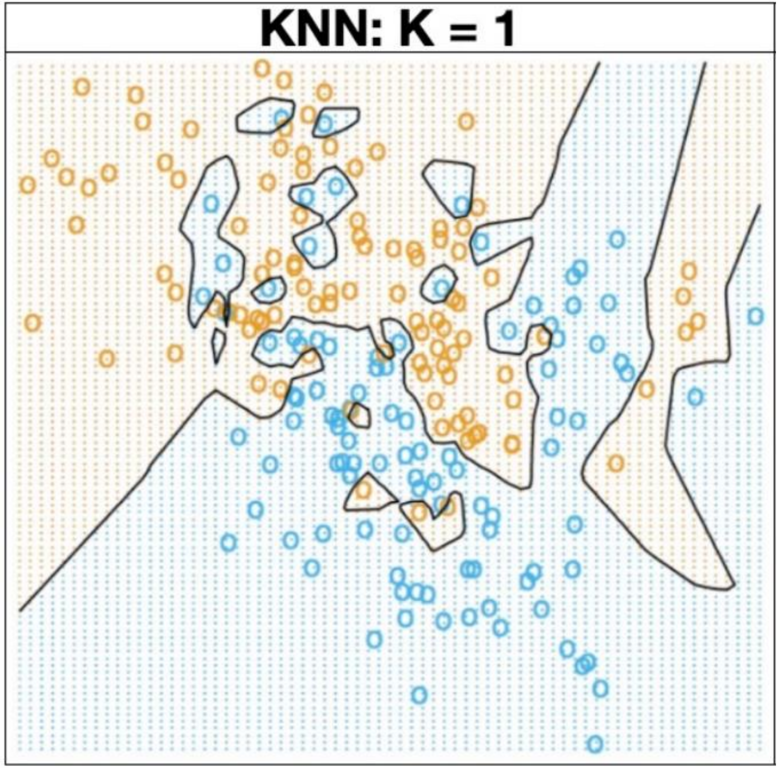
KNN – Large K

- What do you notice about this space?
  - We are very imprecise with our border, we lose a lot of nuance in edge cases.



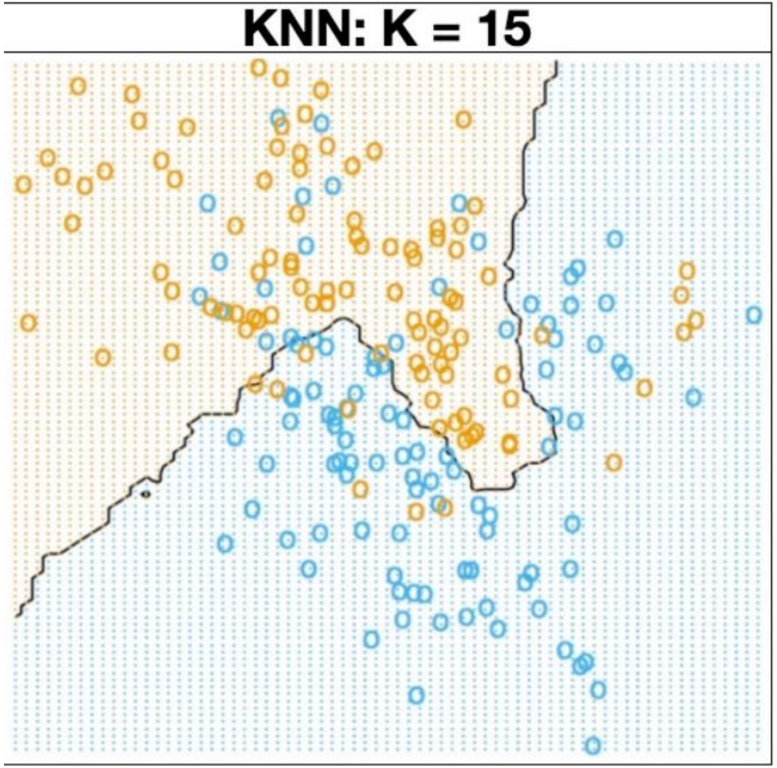
KNN – Small K

- Overfitting – allowing individual samples to have too much power in determining our outcome.



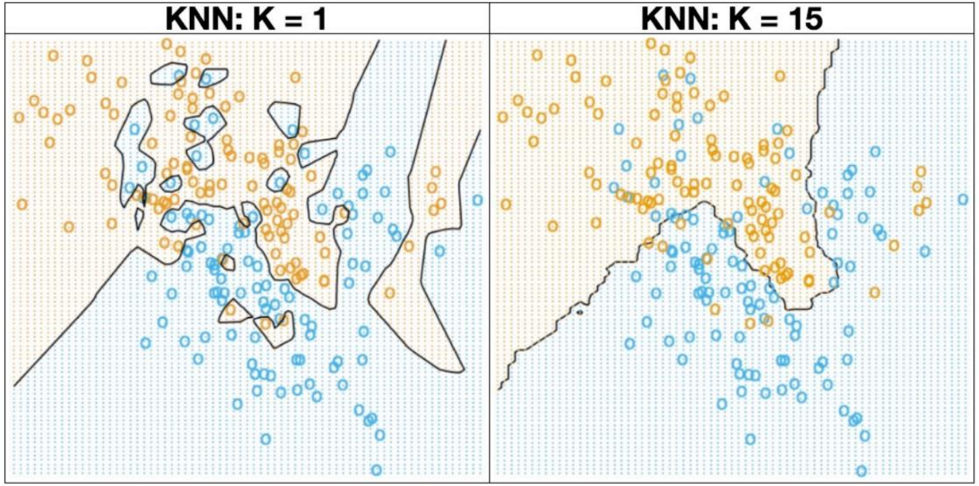
KNN – Large K

- Underfitting – giving individual samples too little power in determining our outcome.



KNN – "Correct" # for K?

- It depends
  - (You're going to get really tired of this answer...)
- Probably more than 1
  - We worry about outliers
- Probably less than 15
  - but it depends on the number of samples in your dataset and the distribution of your data
- 3, 5, 7, etc. are (relatively) common.



Optimizing K using a Validation Set

- Build a model using a first K (K=1)...

Optimizing K using a Validation Set

- Build a model using a first K (K=1)...

- Check how good your K=1 Model does...

Optimizing K using a Validation Set

- Build a model using a first K (K=1)...

- Check how good your K=1 Model does...

- Hang on to that information (accuracy, etc.)

Optimizing K using a Validation Set

- Build a model using a second K (K=2)...

- Check how good your K=2 Model does...

- Compare your K=2 model to your K=1 model. Which one is better?
- Hang on to that information (accuracy, etc.)

Optimizing K using a Validation Set

- For K=k from m to n (usually 1 to n):
    - Build a model using k...

    - Check how good your K=k Model does...

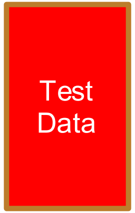
- Compare your m – n models. Which one is best?

Optimizing K using a Validation Set

- You’ve used the Train and Validation data a lot...
    - It’s not very clean to report your “validation” accuracy, because you worked hard finding the K (hyperparameter) that was the very best. That’s “cheating” (from a success perspective, not from a class perspective).

Optimizing K using a Validation Set

- You've used the Train and Validation data a lot...
    - It's not very clean to report your "validation" accuracy, because you worked hard finding the K (hyperparameter) that was the very best. That's "cheating" (from a success perspective, not from a class perspective).
- You need a Test dataset that is untouched to see how good your model *really* does.



Sidebar – Training, Validation, Testing sets

1) Training Set – used in the (usually iterative) training process

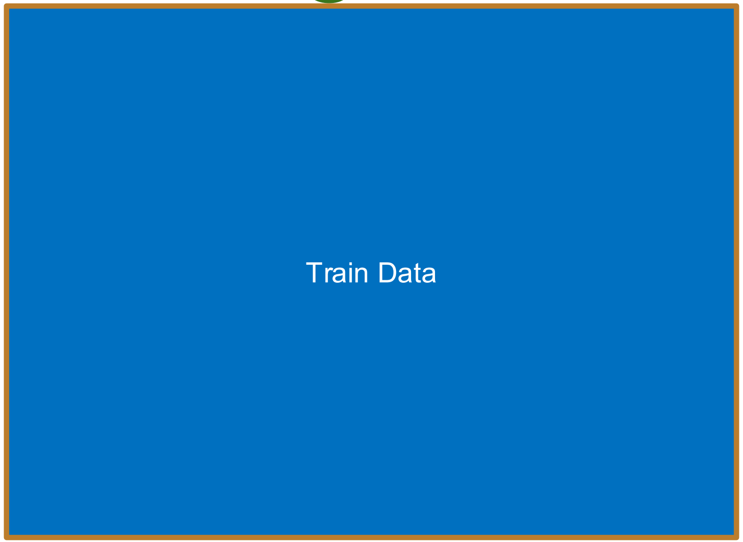
2) Validation Set – kept out of the training process, used to meter your training process

3) Testing Set – Used to evaluate the accuracy, etc. of your learning

Optimizing K using a Validation Set

- You've used the Train and Validation data a lot...
    - It's not very clean to report your "validation" accuracy, because you worked hard finding the K (hyperparameter) that was the very best. That's "cheating" (from a success perspective, not from a class perspective).
- You need a Test dataset that is untouched to see how good your final model *really* does.
    - This is what you would report to others as the expected goodness of your model.

Training & Hold-Out & Test Sets



I pull out another random 20% of my data

If my dataset is smaller, I'm starting to run out of data…

To Be Continued!

K-Nearest Neighbors Algorithmic Complexity

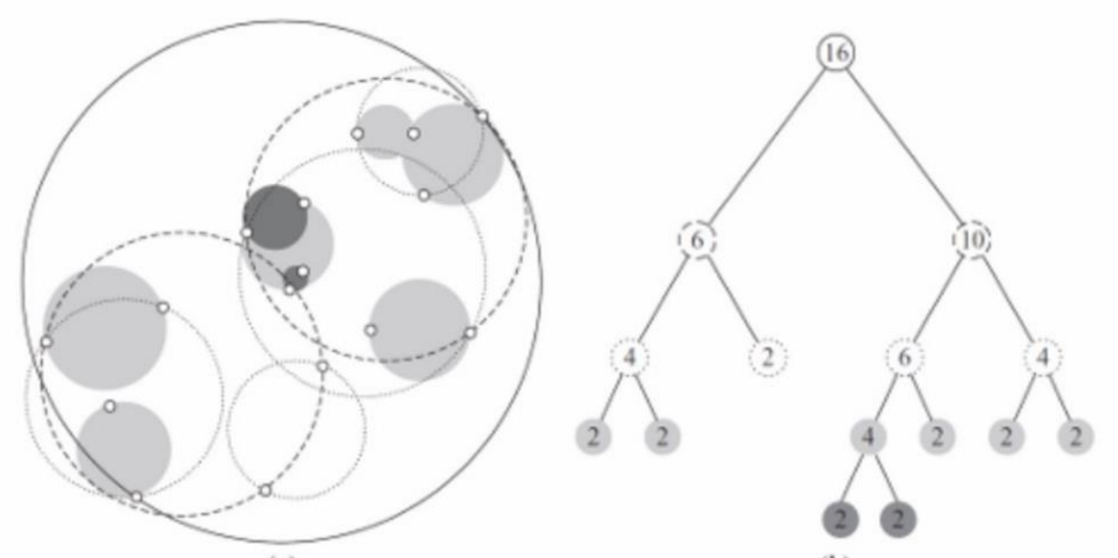
One Query, m training examples, each with D features

$O(m^* D)$

For the Naïve Case

Scikit-learn has additional details in their implementation: https://scikit-learn.org/stable/modules/neighbors.html

K-Nearest Neighbors – Tree Structure



(a) (b)

Problem Space – College Admissions

The following scenario isn't fully true, but it's close to what we do in college admissions…

I am trying to decide if a student should be admitted to my university. I have their SAT and ACT scores and their HS GPA. I also have the history of students who have attended in the past, their SAT / ACT / HS GPA as well as whether or not they graduated. I only want to admit new students if they will graduate.

What is an Error?

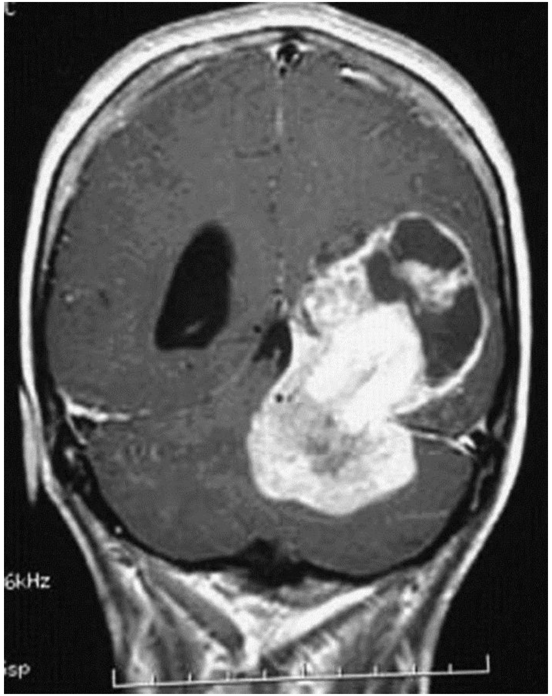
We've looked at trying it out on a test set and getting an "accuracy"

Accuracy = # correct / # total

$$\frac{\sum_n^1 (y_i == \hat{y}_i)}{n}$$

1) What is our "test set"?
2) Are all mistakes created equal?

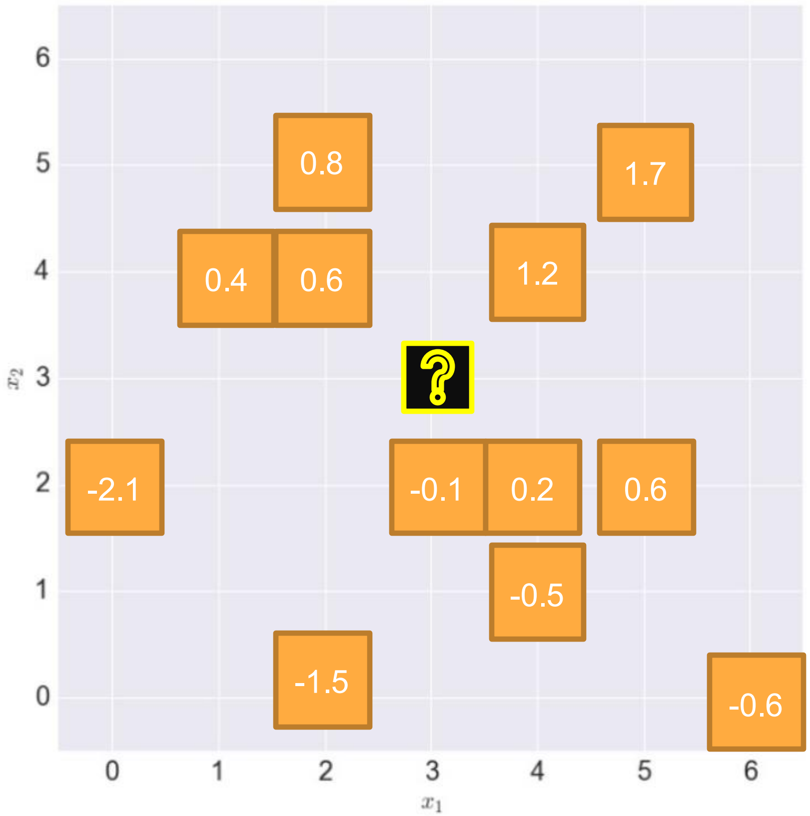
What makes an error?

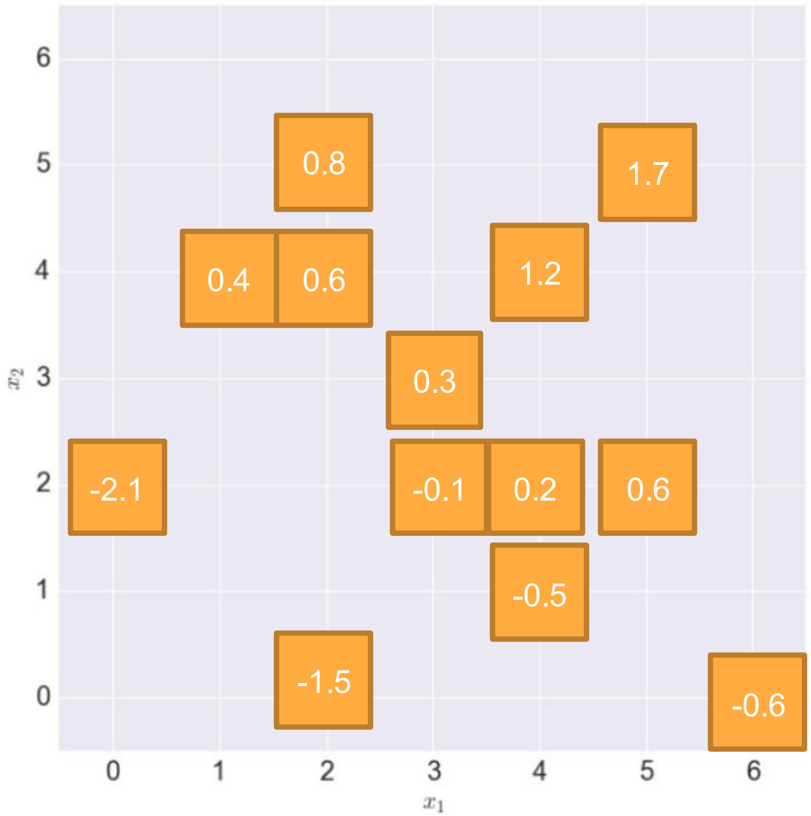


Types of Errors

Classified As *Ground Truth*	**Cancer**	**Not Cancer**
Cancer	True Positive (Hit)	False Negative (Miss)
Not Cancer	False Positive (False Alarm)	True Negative (Correct Rejection)

KNN – Regression





Regression Error

- Accuracy = # correct / # total – Does not capture “close”...

Regression Error

- Accuracy = # correct / # total – Does not capture “close”…
- What’s the difference between our ground truth and our prediction?
    - Absolute Error: $\sum_n^1 |y_i - \hat{y}_i|$
- What’s the *average* distance between our ground truths and predictions?
    - Mean Absolute Error: $\frac{\sum_n^1 |y_i - \hat{y}_i|}{n}$
- What’s the *average* squared distance (we’ll get more into why later)
    - Mean Squared Error: $\frac{\sum_n^1 (y_i - \hat{y}_i)^2}{n}$

Ready for Problem Set 1!