Hello!

Introduction

David Quigley

CSCI 5622

2021 Fall

About Me

Dr. David Quigley (You can call me David or Dr. Quigley)
Instructor, Dept. of Computer Science
9th year at CU (4th year post-graduate)
Applied Machine Learning Research:
Studying the way students use digital tools in the classroom
- Student Scientific Modeling
- Students Reading to Learn
- Students' epistemological beliefs on the nature of science

About You

Post an introduction on Piazza! Share as much or as little as you want (but I want everyone to post, to make sure everyone can access, use Piazza)

Class Resources

Computing Device
- Laptop recommended, any OS (ish)

Course Canvas Website
canvas.colorado.edu

Python 3 (with external packages)
- Anaconda recommended https://www.anaconda.com/download/

“Dumb” Calculator

Course Logistics

Assigned Readings – Expected *Weekly*

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

Problem Sets – Every 2 - 3 Weeks

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

Midterms – Approximately every 7 – 8 weeks

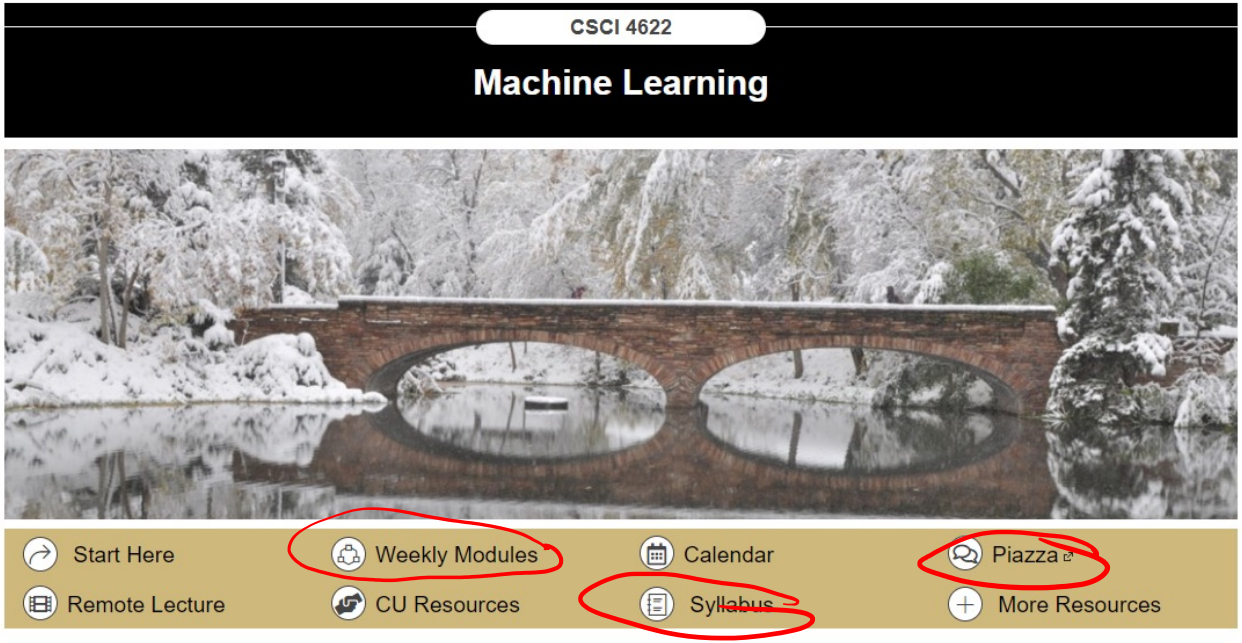
- Announced well in advance
- Midterms = 30% of Grade

Project Updates - Every 5 – 6 weeks (not synced with midterms)

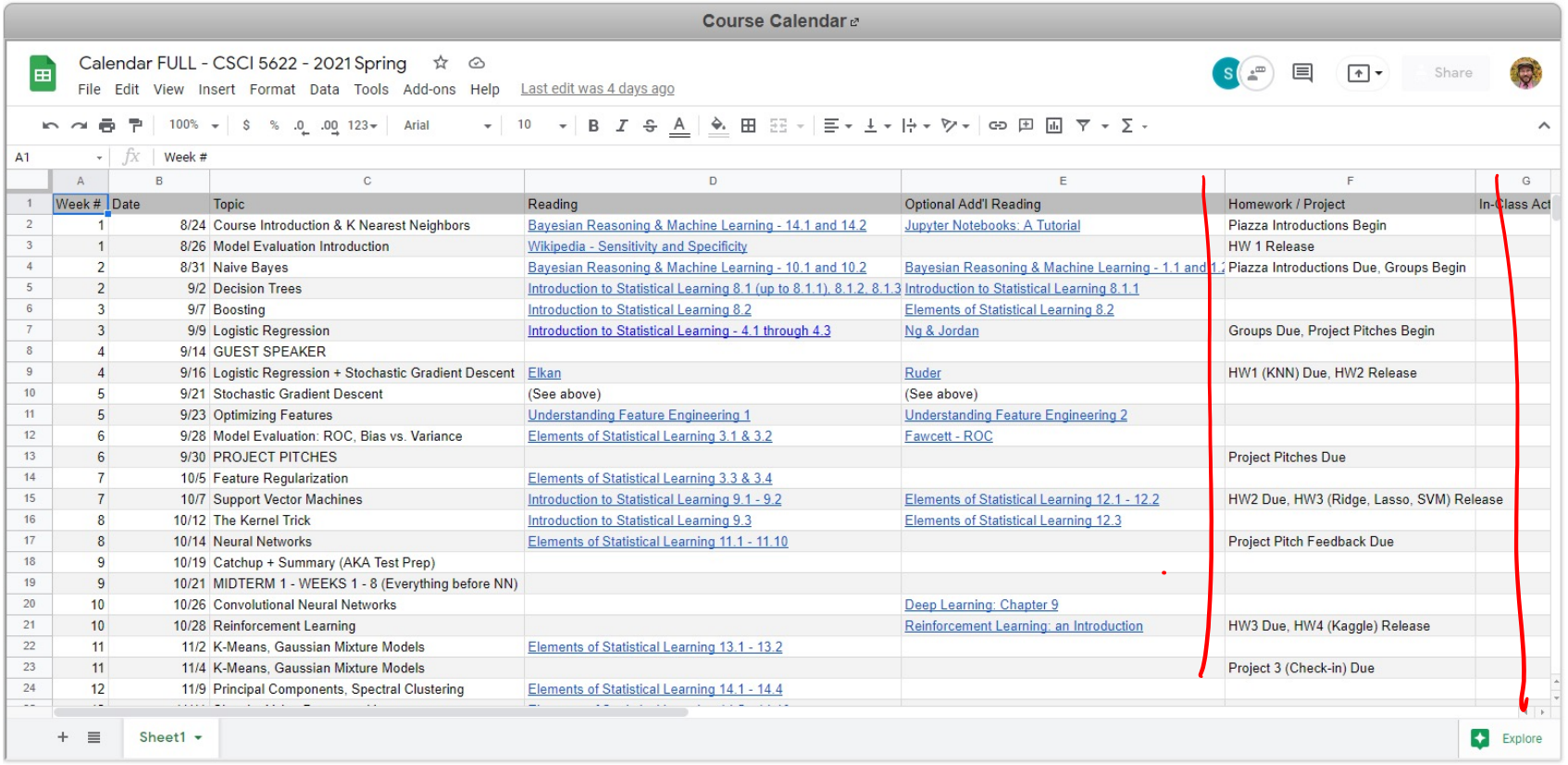
- Be on the lookout for team information
- Projects = 30% of Grade

Course Logistics

Recent Announcements



Calendar



Course Calendar

Calendar FULL - CSCI 5622 - 2021 Spring

File Edit View Insert Format Data Tools Add-ons Help Last edit was 4 days ago

Share

A1 fx Week #

Week #	Date	Topic	Reading	Optional Add'l Reading	Homework / Project	In-Class Act
1	8/24	Course Introduction & K Nearest Neighbors	Bayesian Reasoning & Machine Learning - 14.1 and 14.2	Jupyter Notebooks: A Tutorial	Piazza Introductions Begin	
1	8/26	Model Evaluation Introduction	Wikipedia - Sensitivity and Specificity		HW 1 Release	
2	8/31	Naive Bayes	Bayesian Reasoning & Machine Learning - 10.1 and 10.2	Bayesian Reasoning & Machine Learning - 1.1 and 1.2	Piazza Introductions Due, Groups Begin	
2	9/2	Decision Trees	Introduction to Statistical Learning 8.1 (up to 8.1.1), 8.1.2, 8.1.3	Introduction to Statistical Learning 8.1.1		
3	9/7	Boosting	Introduction to Statistical Learning 8.2	Elements of Statistical Learning 8.2		
3	9/9	Logistic Regression	Introduction to Statistical Learning - 4.1 through 4.3	Ng & Jordan	Groups Due, Project Pitches Begin	
4	9/14	GUEST SPEAKER				
4	9/16	Logistic Regression + Stochastic Gradient Descent	Elkan	Ruder	HW1 (KNN) Due, HW2 Release	
5	9/21	Stochastic Gradient Descent	(See above)	(See above)		
5	9/23	Optimizing Features	Understanding Feature Engineering 1	Understanding Feature Engineering 2		
6	9/28	Model Evaluation: ROC, Bias vs. Variance	Elements of Statistical Learning 3.1 & 3.2	Fawcett - ROC		
6	9/30	PROJECT PITCHES			Project Pitches Due	
7	10/5	Feature Regularization	Elements of Statistical Learning 3.3 & 3.4			
7	10/7	Support Vector Machines	Introduction to Statistical Learning 9.1 - 9.2	Elements of Statistical Learning 12.1 - 12.2	HW2 Due, HW3 (Ridge, Lasso, SVM) Release	
8	10/12	The Kernel Trick	Introduction to Statistical Learning 9.3	Elements of Statistical Learning 12.3		
8	10/14	Neural Networks	Elements of Statistical Learning 11.1 - 11.10		Project Pitch Feedback Due	
9	10/19	Catchup + Summary (AKA Test Prep)				
9	10/21	MIDTERM 1 - WEEKS 1 - 8 (Everything before NN)				
10	10/26	Convolutional Neural Networks		Deep Learning: Chapter 9		
10	10/28	Reinforcement Learning		Reinforcement Learning: an Introduction	HW3 Due, HW4 (Kaggle) Release	
11	11/2	K-Means, Gaussian Mixture Models	Elements of Statistical Learning 13.1 - 13.2			
11	11/4	K-Means, Gaussian Mixture Models			Project 3 (Check-in) Due	
12	11/9	Principal Components, Spectral Clustering	Elements of Statistical Learning 14.1 - 14.4			

Sheet1

Explore

Course Functionality

- Rule #1 – We will be following all campus rules and guidelines for instruction.
- Flexibility – Everyone can access Remote (synchronous), Online (asynchronous)
- Zoom students participate in class, etc.
    - You'll have to use your mic – I can't effectively monitor chat with my setup

Course Functionality

- Rule #1 – We will be following all campus rules and guidelines for instruction.
- Flexibility – Everyone can access Remote (synchronous), Online (asynchronous)
- Zoom students participate in class, etc.
    - You'll have to use your mic – I can't effectively monitor chat with my setup
- I may occasionally have to teach *remotely* or *online* for everyone's safety.

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

SKLearn Scratch

What is Machine Learning (ML)?

Fed the Machine to do a Task for Us

Fancy Statistics / LA

Dependable Fortune Telling

Dependable(?)

Predict Future From Past & Current Data

Automatic optimization

Pattern Recognition

Abstract insights from data

Signal Processing

Improving Performance from Experience

data (?) functions

Pairing Q's and A's to get Rules

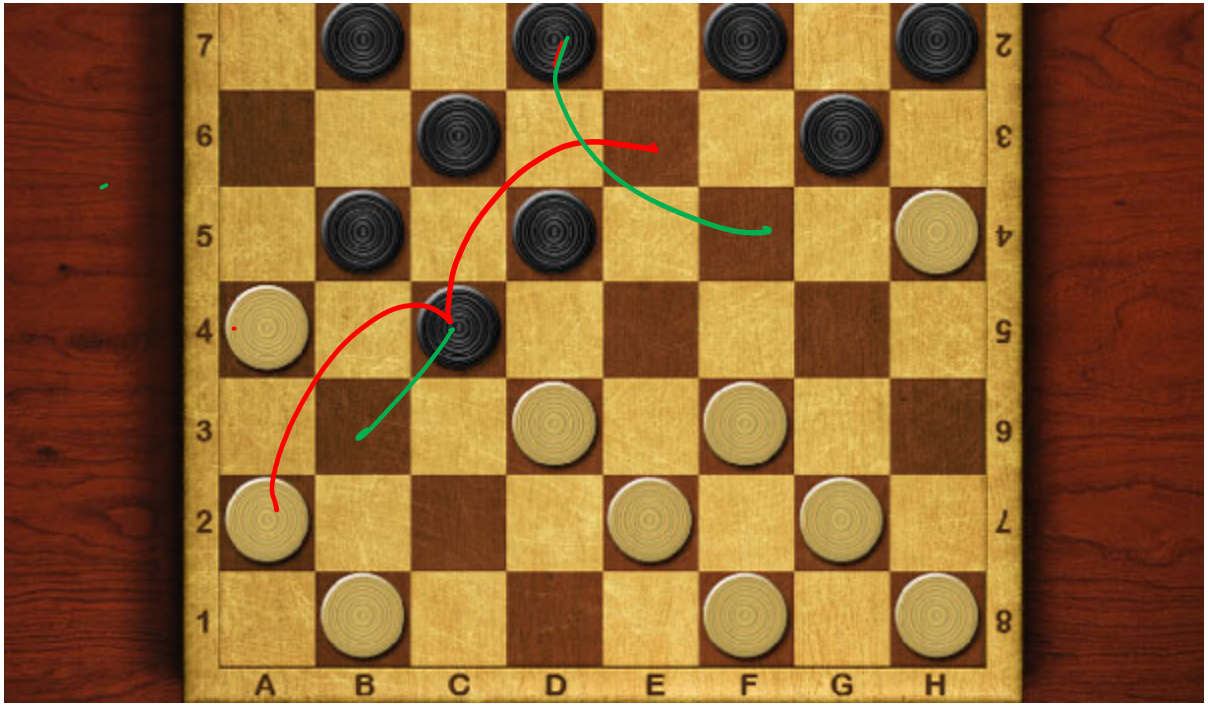
Instrumental Tool for the "4th Paradigm of Knowledge"

Extrapolate from Answered to unanswered

Look at the world from a machine's Perspective

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
2
3
4
5
6
7
8

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

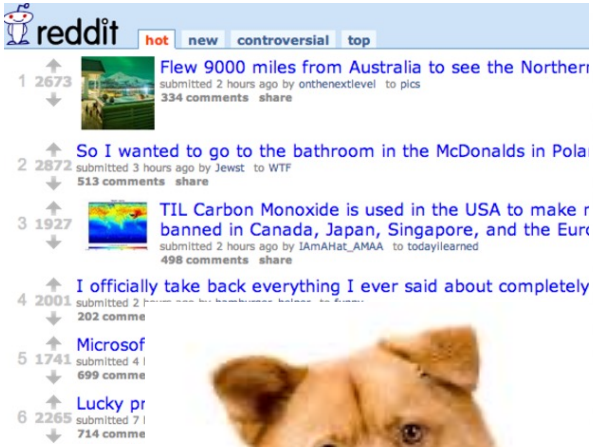
5 1741 Microsoft Security Essentials fails anti-virus certification test, Redmond challenges results (theverge.com)
submitted 4 hours ago by ken27238 to technology
699 comments share

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

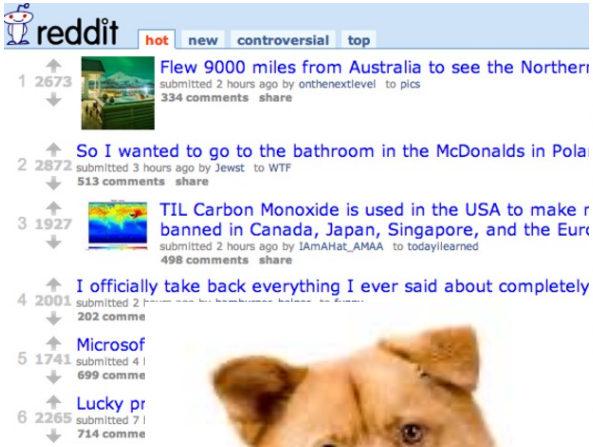
What is “Task”?


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)
202 comme
5 1741
Microsof
699 comme
6 2265
Lucky pr
714 comme

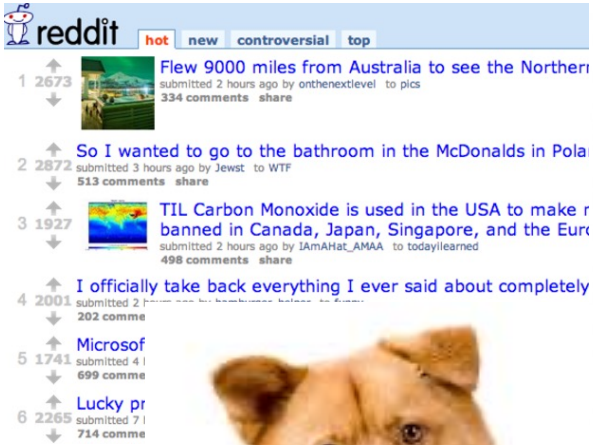
What is “Task”?

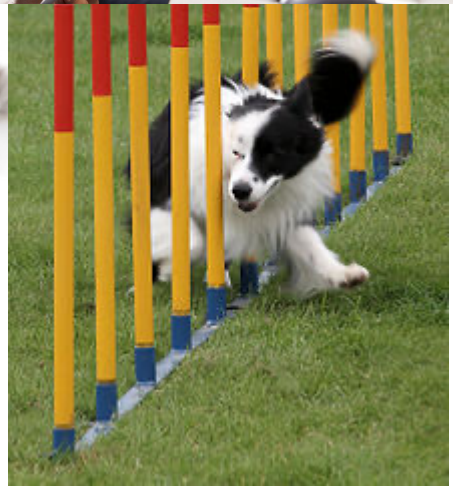

reddit
hot
new
controversial
top
1 2673
Flew 9000 miles from Australia to see the Northern
submitted 2 hours ago by onthenextlevel to pics
334 comments share
2 2872
So I wanted to go to the bathroom in the McDonalds in Pola
submitted 3 hours ago by Jewst to WTF
513 comments share
3 1927
TIL Carbon Monoxide 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 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”?

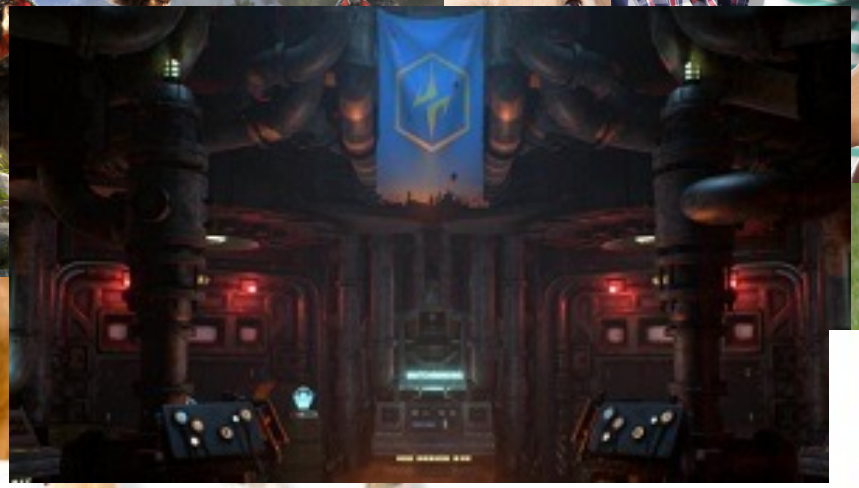

reddit
hot
new
controversial
top
1 2673
Flew 9000 miles from Australia to see the Northern
submitted 2 hours ago by onthenextlevel to pics
334 comments share
2 2872
So I wanted to go to the bathroom in the McDonalds in Pola
submitted 3 hours ago by Jewst to WTF
513 comments share
3 1927
TIL Carbon Monoxide 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

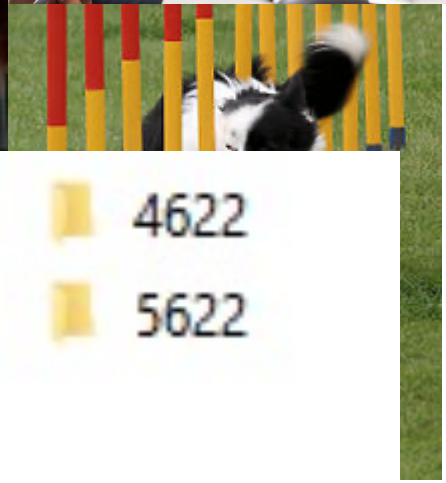


What is “Task”?


reddit
1 2673
So I wa
2 2872
submitted 3
513 comme
3 1927
I officia
4 2001
submitted 2
202 comme
Microsof
5 1741
submitted 4
699 comme
Lucky p
6 2265
submitted 7
714 comme

FUNCOM




4622
5622

What is “Experience”?

Recommended for You Based on Kindle Paperwhite, 6" High Resolution Display w...

Page 1 of 5

MoKo Case for Kindle Paperwhite, Premium Thinnest and Lightest Leather Cover with…

898

$9.99 Prime

Swees Ultra Slim Leather Case Cover for Amazon All-New Kindle Paperwhite (Both 2012…

273

$3.99 Prime

Fintie SmartShell Case for Kindle Paperwhite - The Thinnest and Lightest Leather Cover for…

7,015

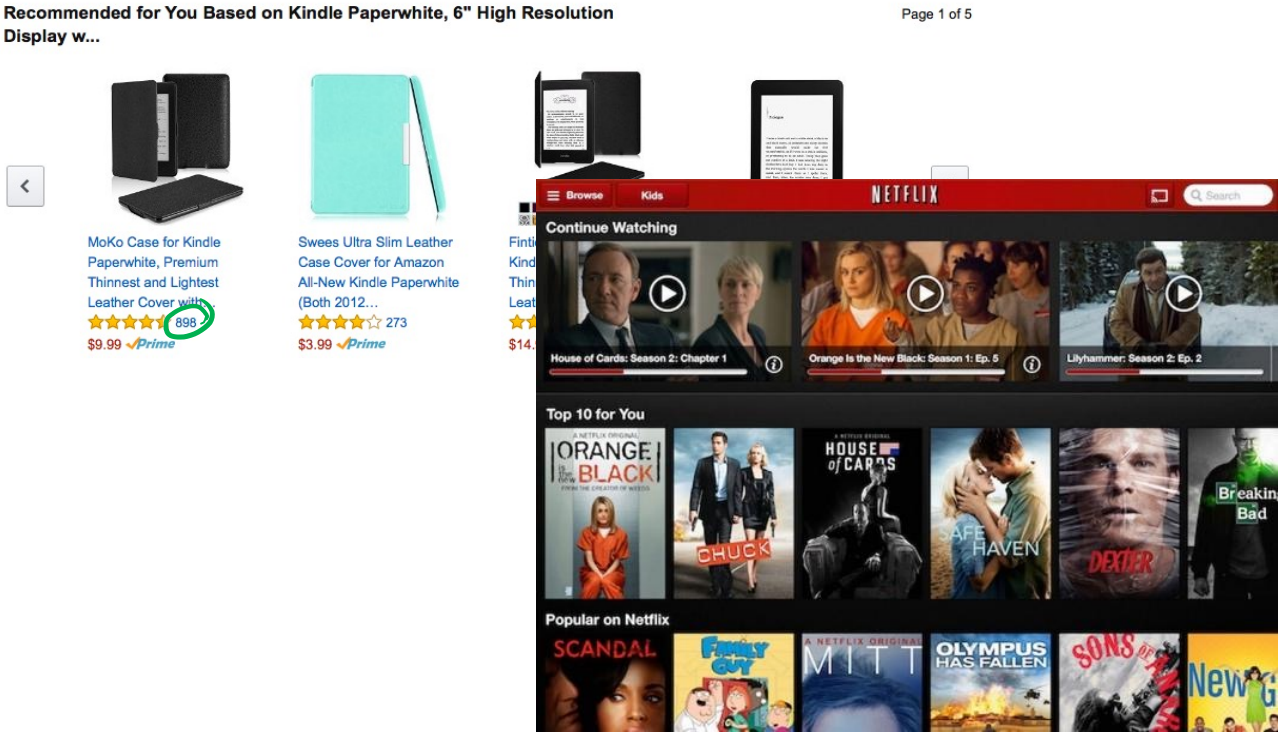
$14.99 Prime

Kindle Paperwhite, 6" High Resolution Display (212 ppi) with Built-in Light, Free 3G…

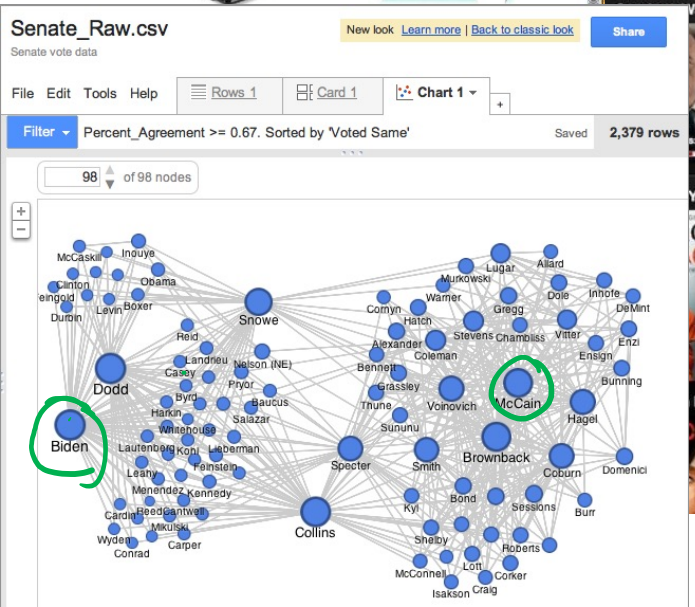
45,265

$159.99 Prime

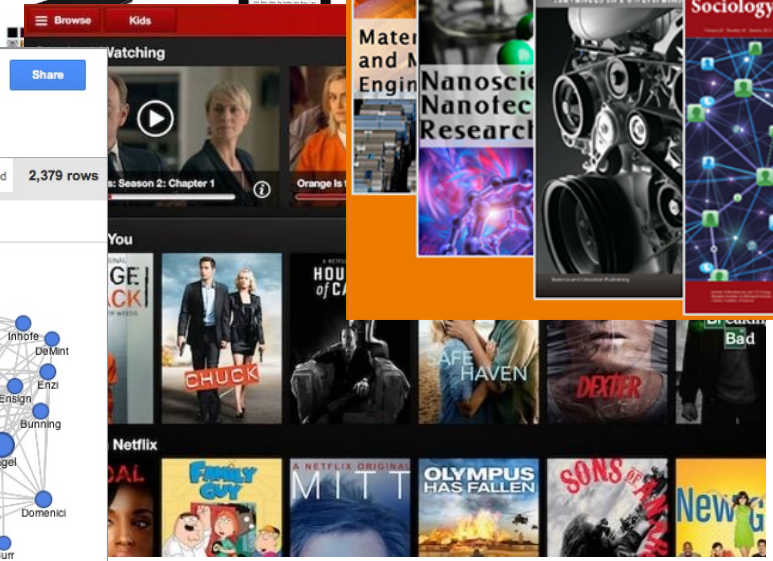
What is “Experience”?

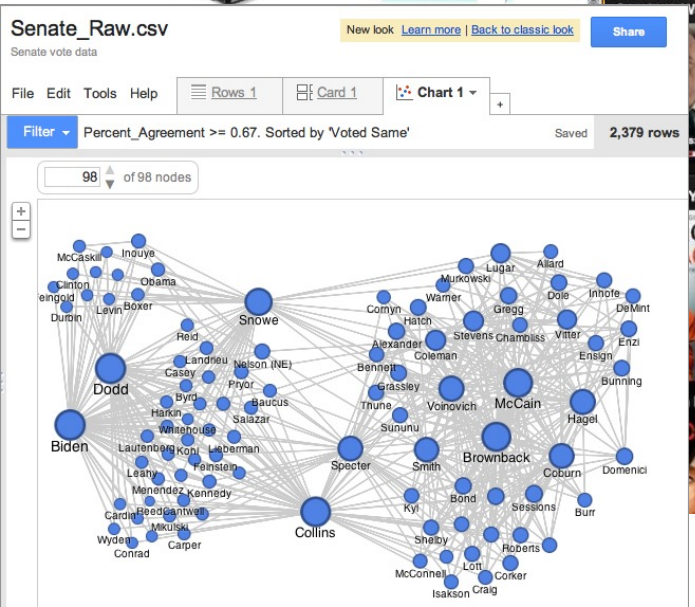


What is “Experience”?

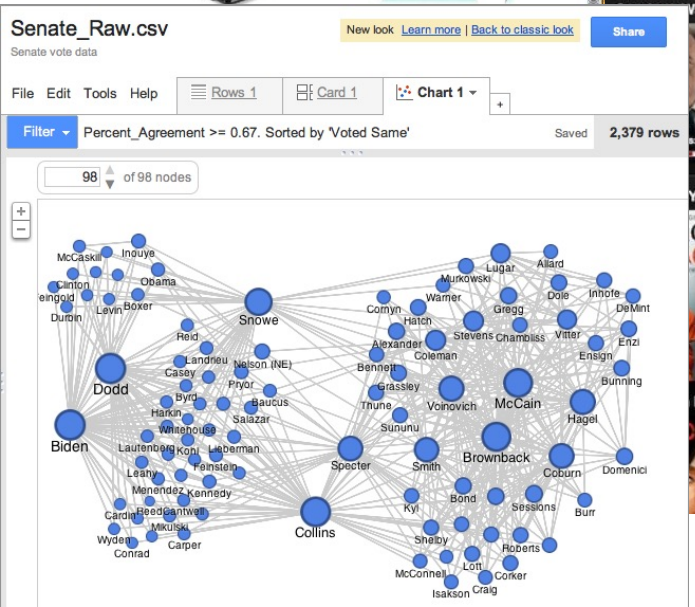


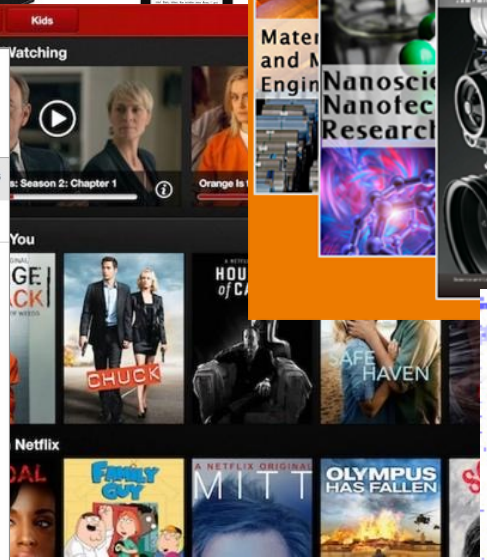
What is “Experience”?



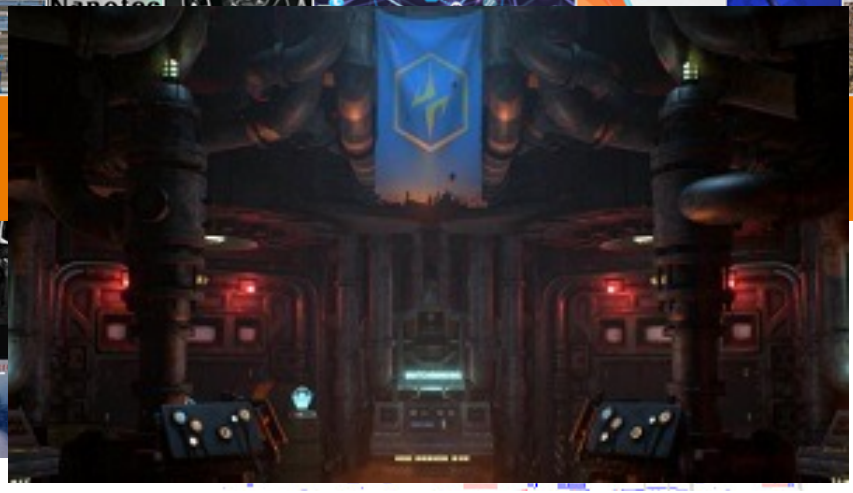


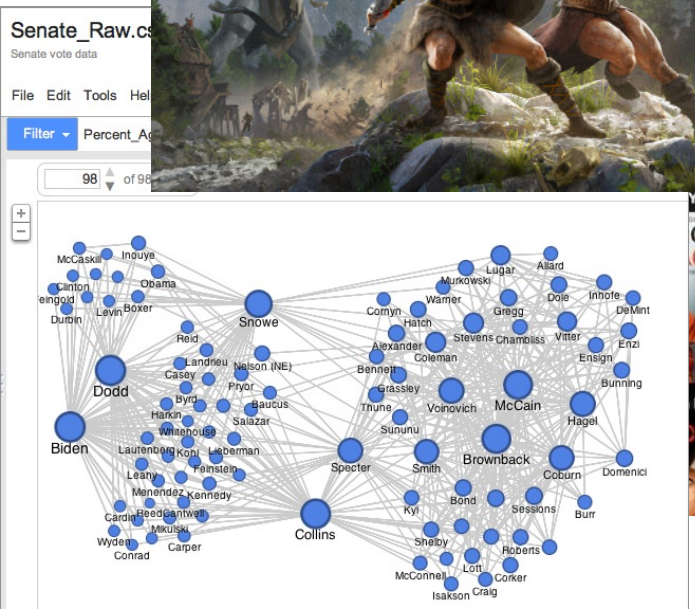
What is “Experience”?

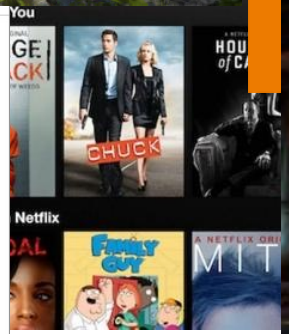




What is “Experience”?







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 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”?

Recommended for You Based on Kindle Paperwhite, 6" High Resolution Display w...

Page 1 of 5

MoKo Case for Kindle Paperwhite, Premium Thinnest and Lightest Leather Cover with…

4.5 out of 5 stars 898

$9.99 Prime

Swees Ultra Slim Leather Case Cover for Amazon All-New Kindle Paperwhite (Both 2012…

4.0 out of 5 stars 273

$3.99 Prime

Fintie SmartShell Case for Kindle Paperwhite - The Thinnest and Lightest Leather Cover for…

4.5 out of 5 stars 7,015

$14.99 Prime

Kindle Paperwhite, 6" High Resolution Display (212 ppi) with Built-in Light, Free 3G…

4.5 out of 5 stars 45,265

$159.99 Prime

What is “[Improved] Performance”?

Recommended for You Based on Kindle Paperwhite, 6" High Resolution Display w...

Page 1 of 5

MoKo Case for Kindle Paperwhite, Premium Thinnest and Lightest

Swees Ultra Slim Leather Case Cover for Amazon All-New Kindle Paperwhite

Fintie SmartShell Case for Kindle Paperwhite - The Thinnest and Lightest

Kindle Paperwhite, 6" High Resolution Display (212 ppi) with Built-in Light, Free

,265

What is “[Improved] Performance”?

What is “[Improved] Performance”?

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 *algorithm* that maps *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 *algorithm* that maps *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 *algorithm* that maps *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

END

X_i = # Bedrooms, # Bath, AC, Location, "Safety", Sqft, Pets, Pool, HoA, Parking, Schools

Y = Price

Thursday

CSCI 5622

- David Quigley
- Assistant Teaching Professor!

Course Logistics

- Weekly Activity – Piazza Introduction
    - Due Tuesday, August 31, 3:55 PM (i.e. before class)
- Problem Set 1 – releases Now!
    - Due Thursday, September 16, 3:55PM (i.e. before class)
- 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! Plus I may be swapping for a new book as we move along.

Problem Space – Housing Market

END

X_i = # Bedrooms, # Bath, AC, Location, "Safety", Sqft, Pets, Pool, HoA, Parking, Schools

Y = Price

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 – Student Learning

- How might you gather information from this population?

Survey Registrar Extracurricular activities
Quizzes Observation Interviews
Directory Ask Parents

Problem Space – Student Learning

X_i = grades

age

Name

location

"Demographic" — Nationality, gender, race, personality, SES

Y = Acumen for subject

IQ

Graduation

Problem Space – Student Learni

Chapter 1

Ecology and the fate of the black-footed ferret

Species and habitats

1 Planet Earth is the only known planet in the universe that is home to living things. Living things, like plants and animals, are called organisms. You are an organism, and plants, fungi, and tiny single-celled bacteria are also organisms! Think about what you, as an organism, need to survive. You might think about food and water, but you have other needs too. Earth is unique because it supports organisms by providing them with their five basic needs. Yes, organisms need water and food. Also, most organisms need air, energy and a proper space to live called a habitat. When they have what they need, organisms are better able to survive in their environments.

2 The organisms in a particular area and all the non-living parts of that area make up an ecosystem. The five factors needed for survival are found there: food for energy, water, air, and living spaces, or habitats. Desert, rainforests, prairies, and coral reefs are all examples of ecosystems. What is the ecosystem where you live?

3 Earth has millions of different types of organisms, which are called species. Species are groups of similar organisms that typically live in similar environments and interact with each other. They also interact with the non-living parts of the environment - like air, water, and soil. All the interactions among species and their habitats can get very complex. Ecology is the study of these interactions. Ecologists study prairie dog towns, insect colonies, rain forests, migrating animal herds, bacteria colonies growing on the bottom of the ocean, and so on. These scientists hope to understand more about interactions among species and habitats.

Ecosystem Balance: The Black-Footed Ferret

4 The fate of the black-footed ferret is an example of the complex interactions between species. The furry long-necked creatures look like a cross between a raccoon and a weasel. By the late 1970s, scientists thought black-footed ferrets were extinct in the wild. Then, over the years, they discovered that ferret numbers mirrored the number of prairie dogs in the area. Why? It's because prairie dogs make up 90 percent of what ferrets eat. When the prairie dogs disappeared, so did the ferrets.

A black-footed ferret
Credit: J. Michael Lockhart / USFWS

5 Habitat loss is one reason for the decline of prairie dog colonies and thus, the decline of the ferret. New houses and agricultural croplands have taken over their villages. Ranchers, farmers, and gardeners have actively killed them because they consider them pests. At the same time, diseases swept through the prairie dog colonies and wiped out many of them.

6 Without a reliable food supply the ferret population almost died out. Scientists tried to breed ferrets in captivity, but were not successful. So people pretty much gave up on the black-footed ferret. Then, in 1981, a rancher's dog discovered ferrets in Wyoming. Ecologists trapped as many of the animals as they could. They tried to get these survivors to breed in captivity. Over many generations, this group and its offspring successfully produced more than 4,800 animals. The ecologists have succeeded in reintroducing some of the captive ferrets into 12 different places. Ferrets are starting to make a comeback in the American West because there are more prairie dogs. In addition to a reliable food supply, another key to ferret survival is habitat protection. Prohibiting human activities in certain areas, including building houses and roads, logging, and grazing livestock would allow both prairie dog colonies and ferret populations to rebound. The idea is to preserve the environment in its natural state so that existing ecosystems can continue to thrive as nature intended.

< Previous Next >

Chapter 1 assessment

Check Your Understanding of Chapter 1

A balanced ecosystem is important because

- a) It gives scientists something to study.
- b) It allows living and non-living things to thrive.
- c) It tells scientists when they need to make changes for a more dynamic ecosystem.
- d) The Department of Fish and Wildlife will release more hunting licenses.

Answer key: b | Processing level: SM | Category: Inference | Paragraphs: 1,2,3,4,5,6

What is the BEST way we can help the ferrets?

- a) Teach them to eat other prey.
- b) Move ferrets to a place where there are still prairie dogs.
- c) Protect the prairie dogs in areas where ferrets live.
- d) Teach them to hibernate so they need less food

Answer key: c | Processing level: GC | Category: Main point | Paragraphs: 6

Please select the most appropriate answer to complete the following sentence:
As prairie dog numbers increased _____

- a) ferrets started to make a comeback in the American West.
- b) so did the numbers of raccoons and weasels.
- c) more new houses were built and more agricultural croplands were cultivated.
- d) ecologists learned more about species and habitats.

Answer key: a | Processing level: LC | Category: Fact | Paragraphs: 6

Answer the questions above to check your understanding

Check your answers

Chapter 1 lesson

Linking Ideas

First, listen to the lesson below:

Linking Ideas lesson. See also: Common transition words (from student toolbox)

In this lesson, you will learn about linking related ideas by identifying pronouns and synonyms that writers sometimes use in place of familiar words. Linking these pronouns and synonyms to the words they replace (referents) will help you keep track of what you are read in the sentence.

Chapter 2

What's in an ecosystem

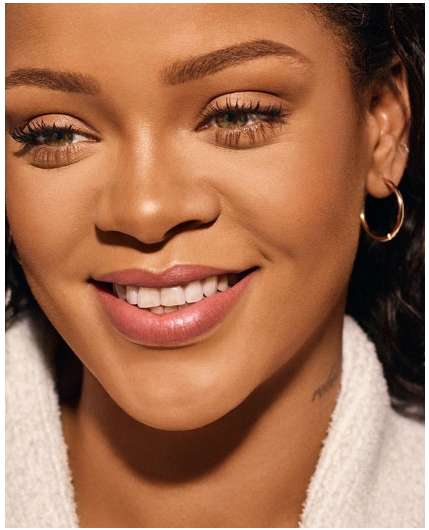
Biotic and Abiotic Factors

1 An ecosystem is made up of living and non-living parts. As we learned in Chapter 1, an ecosystem is a complex group of organisms that interact with each other in a particular environment. The organisms, which may include plants, animals, insects, fish & reptiles, interact with each other, but they also interact with the non-living environment – the air, soil and water. Ecology includes the study of these two main categories of interactions.

2 Factors are what ecologists call the parts of the ecosystem that interact or effect each other. When a mountain lion eats a deer, ecologists refer to the lion and the deer as biotic (living) factors because they have an effect on each other. When water quality in Sandy Creek affects the health of a fish called the white sucker, ecologists refer to the water as an abiotic (non-living) factor affecting the survival of the white sucker. Interactions are not always beneficial to everyone involved. In the example of the lion and the deer, the lion benefited but the deer did not.

Problem Space – Sentiment

- Manager for Roc Nation / Rihanna
- *What is public sentiment about her newest album?*
    - *How might you gather this information?*



Social Media

Hits on Videos
Comments

Streams Likes

Purchases

Solicit feedback

News Articles

Radio Stations

Concert Tickets

Problem Space – Sentiment

X_i =

Y =

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.

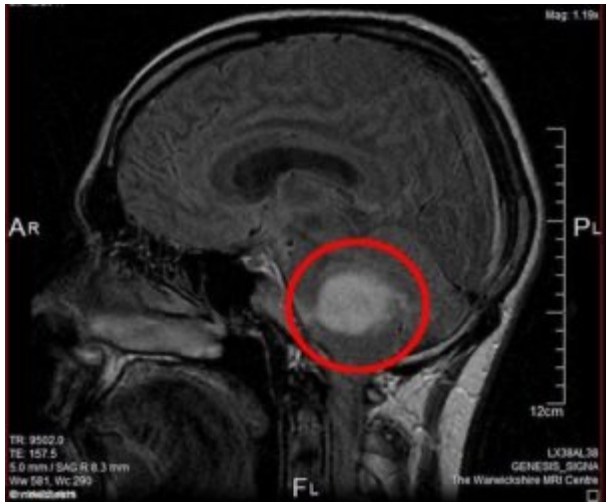
Approximate Z, as f = X → Y

Supervised Learning

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

Approximate Z, as f = X → Y

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



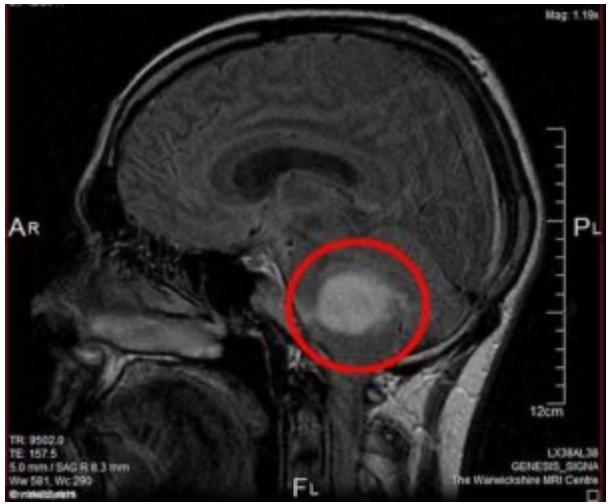
Supervised Learning

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*.



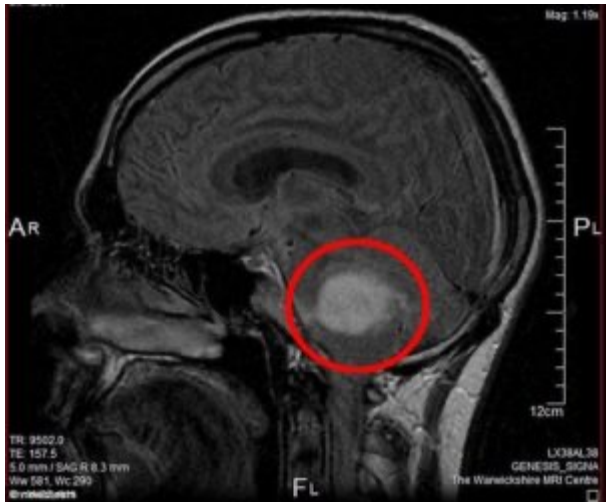
Supervised Learning

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*. So...



Unsupervised Learning

Find *hidden structure* in data that cannot be formally observed.

Discover Z

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

Discrete Answer Space

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

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

Classification

Continuous Answer Space

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

$y = w^t x + w^0$

w = weight (w^t is weight for each element t)

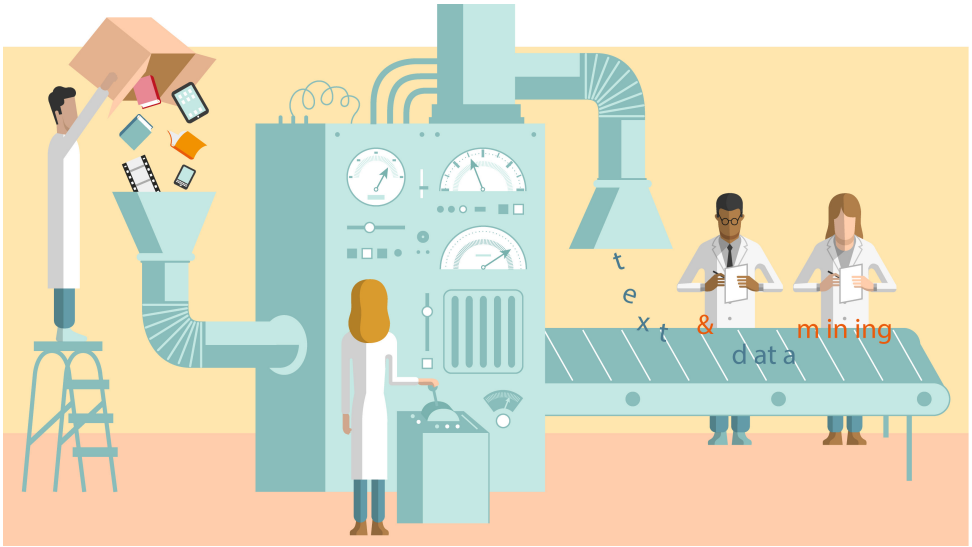
w^0 = bias

Regression

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

We'll cover these areas, but not in depth

What this course isn't – Data Mining (5502)

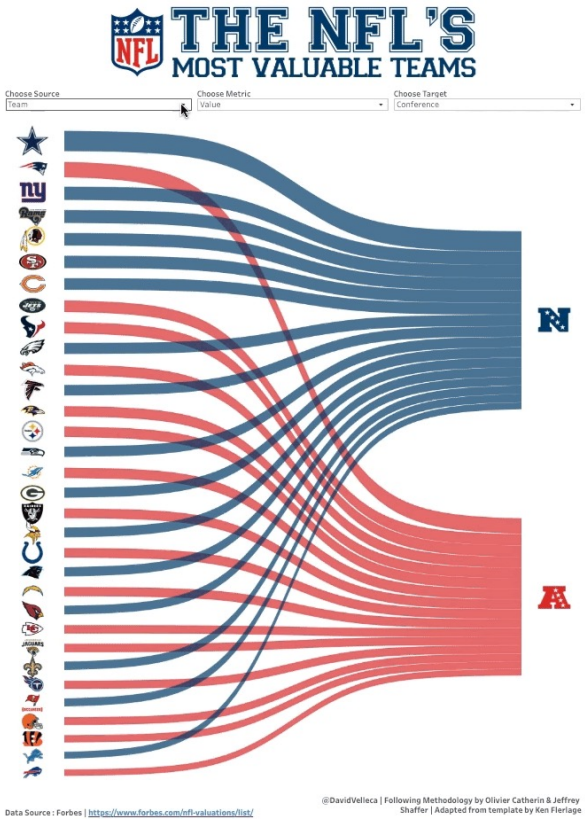


We'll cover these areas, but not in depth

What this course isn't – NLP (5832)

We'll cover these areas, but not in depth

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



For Today...

Install Python 3, numpy, matplotlib, pandas, and jupyter, sklearn

- EASIEST WAY: Install Anaconda
    - https://www.anaconda.com/distribution/
    - This is the platform we will be using to check Problem Sets

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.

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

K = 1

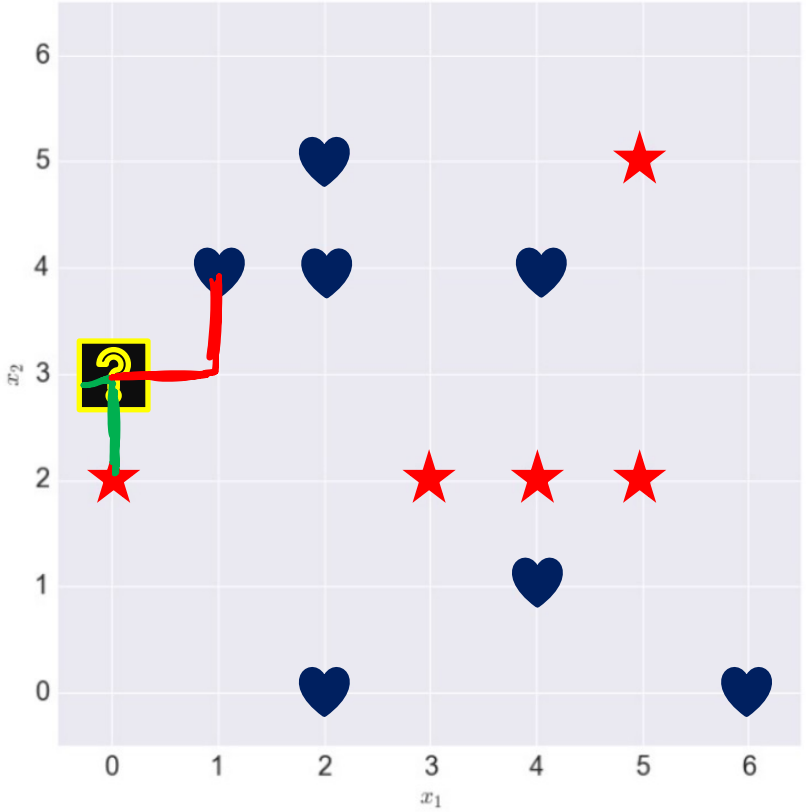
Student	SAT	ACT	GPA	Graduated?
A	1200	26	3.2	Yes
B	1450	28	3.5	Yes
C	1000	20	3.0	Yes
D	730	15	2.0	No
NEW	720	16	2.2	??? NO

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	???

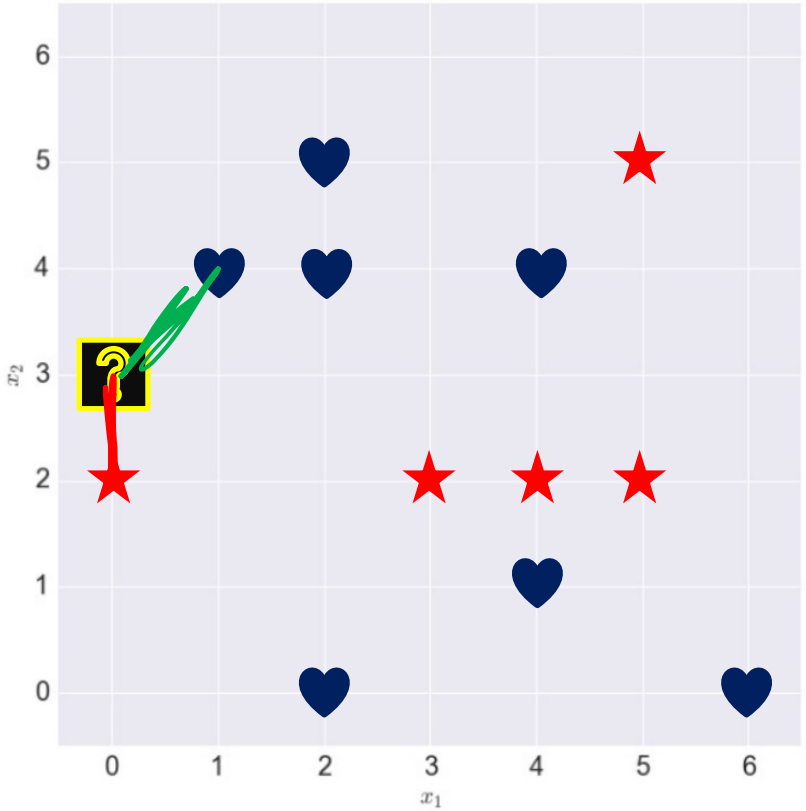
KNN – Manhattan/Taxicab Distance (L_1 Norm)

- Manhattan Distance: $\Sigma_{a=1...n}|x_{ia} - x_a|$



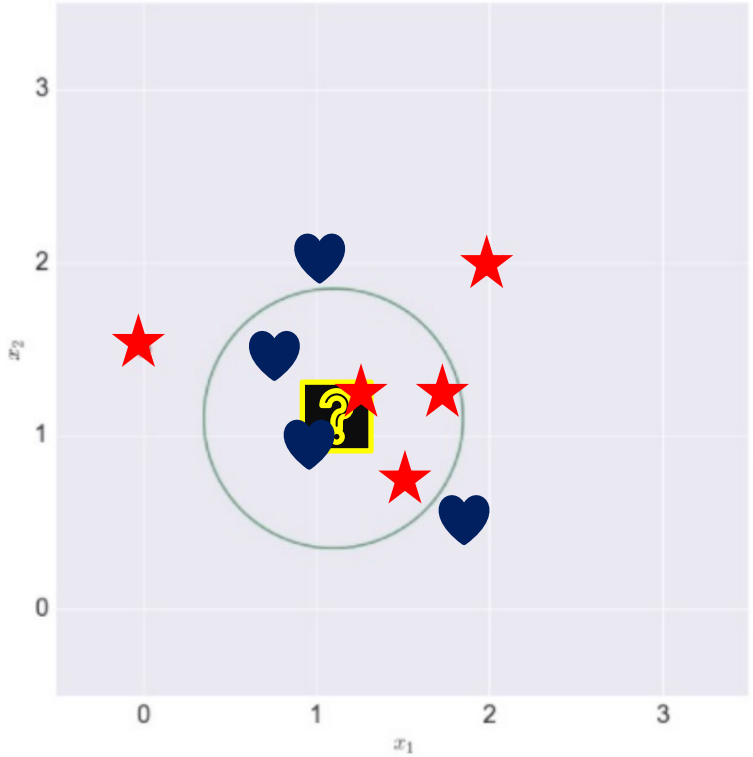
KNN – Euclidian Distance (L_2 Norm)

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



KNN – Euclidian Distance (L_2 Norm)

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



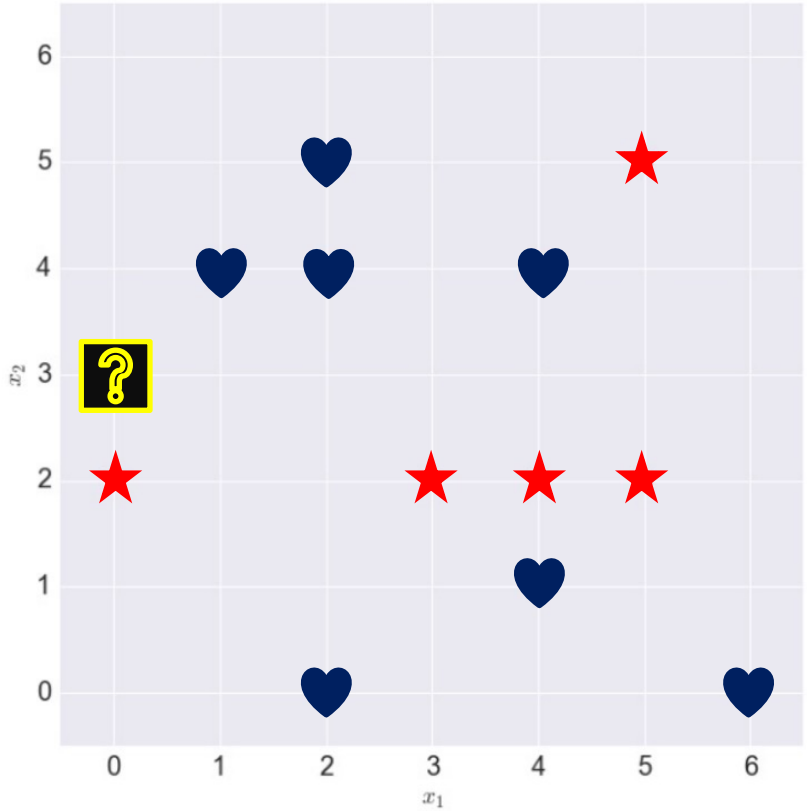
Find the Nearest Neighbors

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

Nearest Neighbor

(0,2)

Prediction



Find the Nearest Neighbors

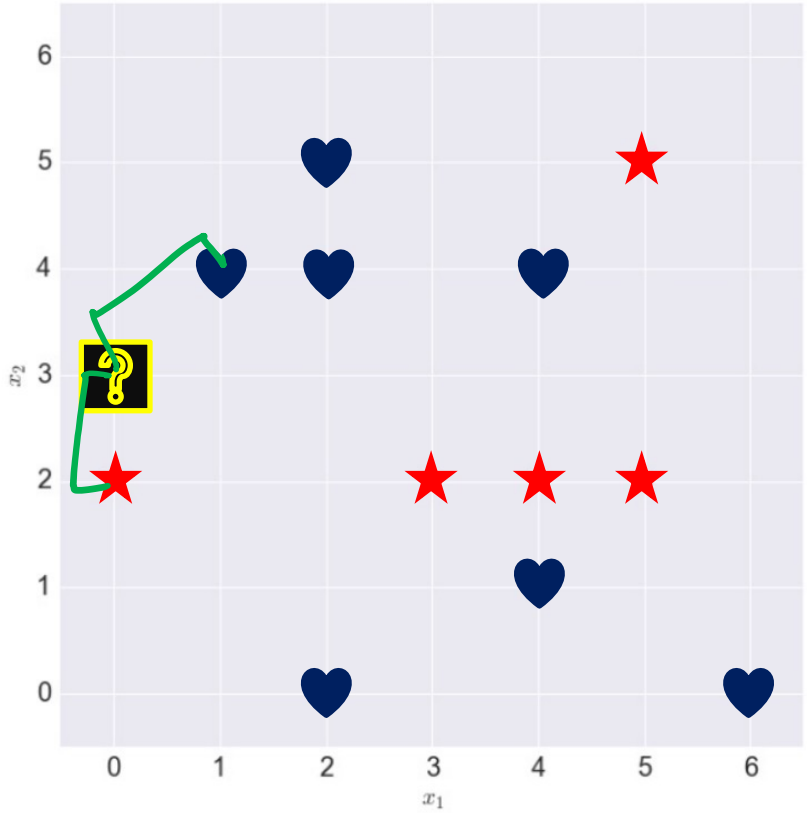
K = 2

Nearest Neighbors

(0,2) (1,4)

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)

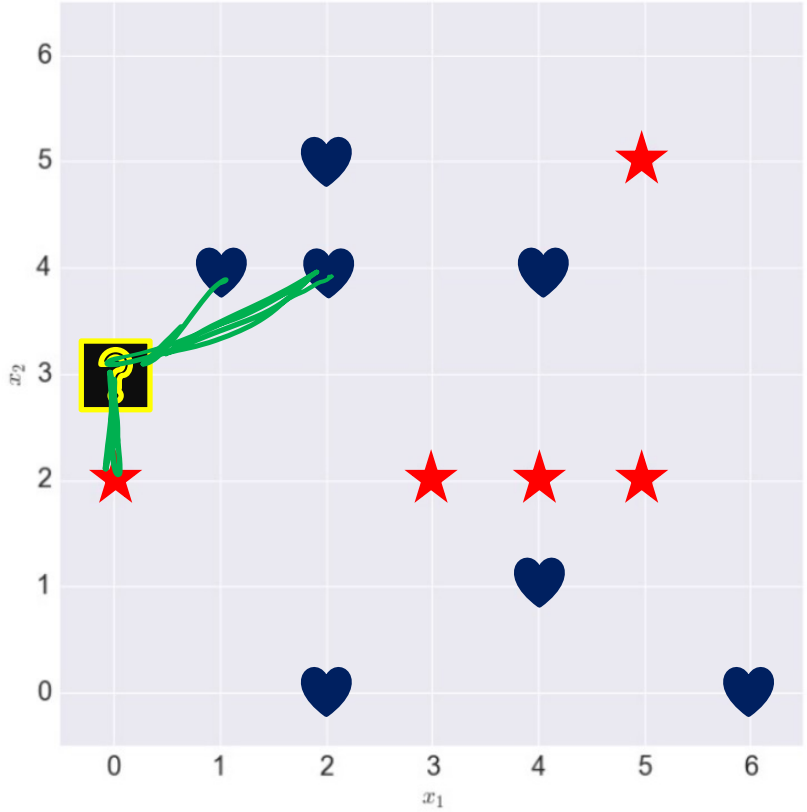
- A safety net: Set up defaults if there's a tie
    - A common default: Whichever case is more common
    - An ethical consideration: Whichever case is safer / more ethical

Find the Nearest Neighbors

K = 3

Nearest Neighbors

Prediction



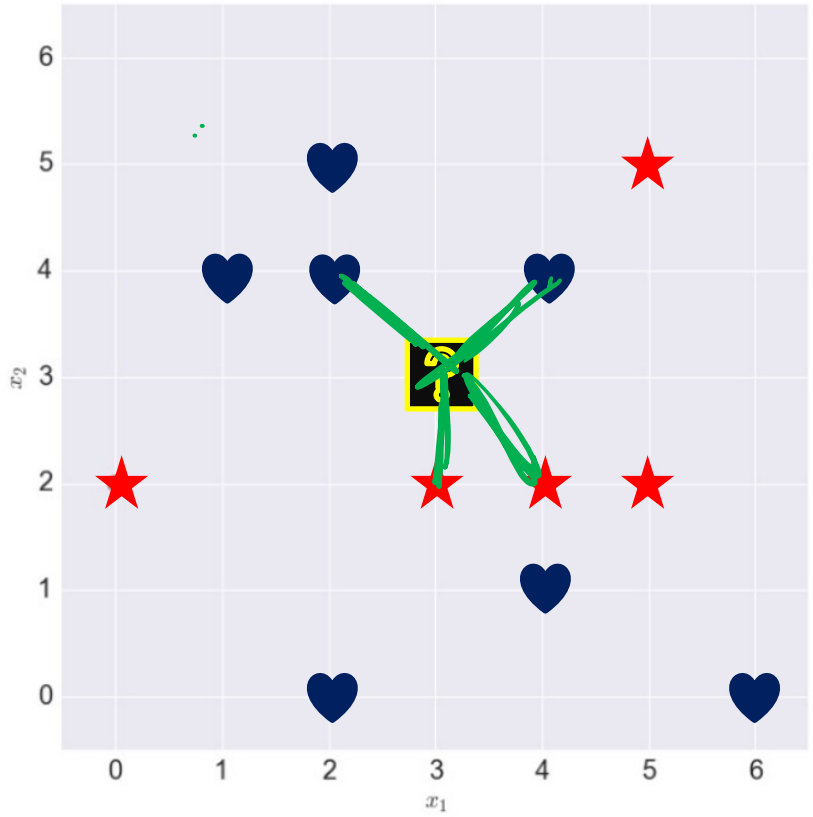
Find the Nearest Neighbors

K = 3

Nearest Neighbors

(3,2)

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.

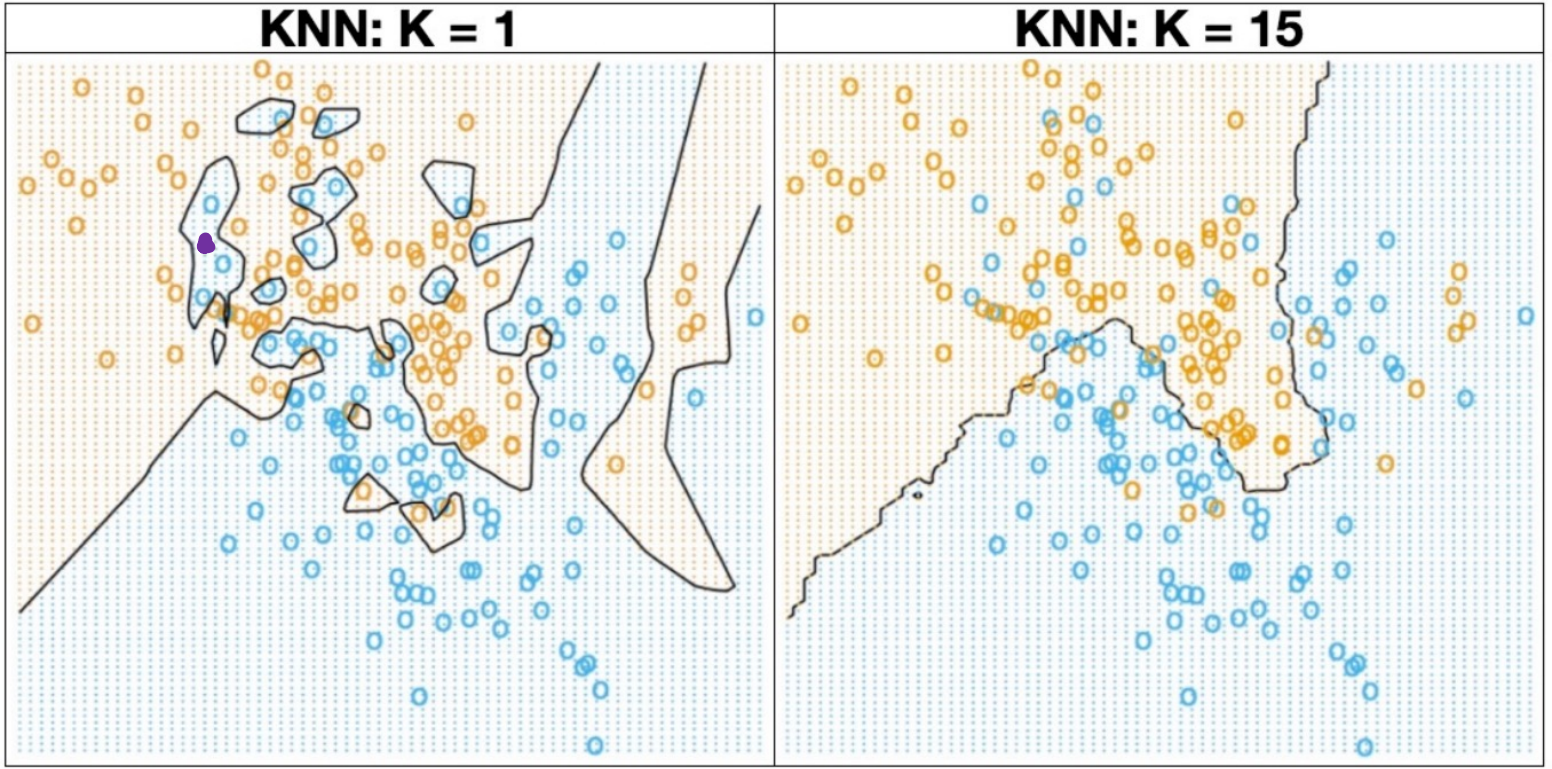
What are you going to do?

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

How do we know if it works?

Evaluating Models

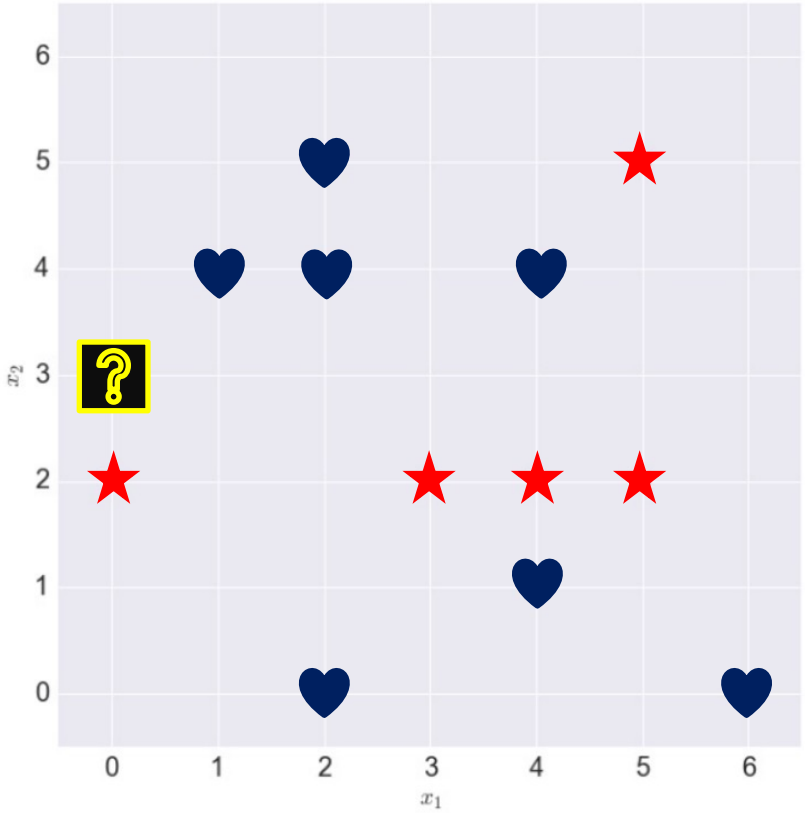
How do we know if it works?

Find the Nearest Neighbors

Classification

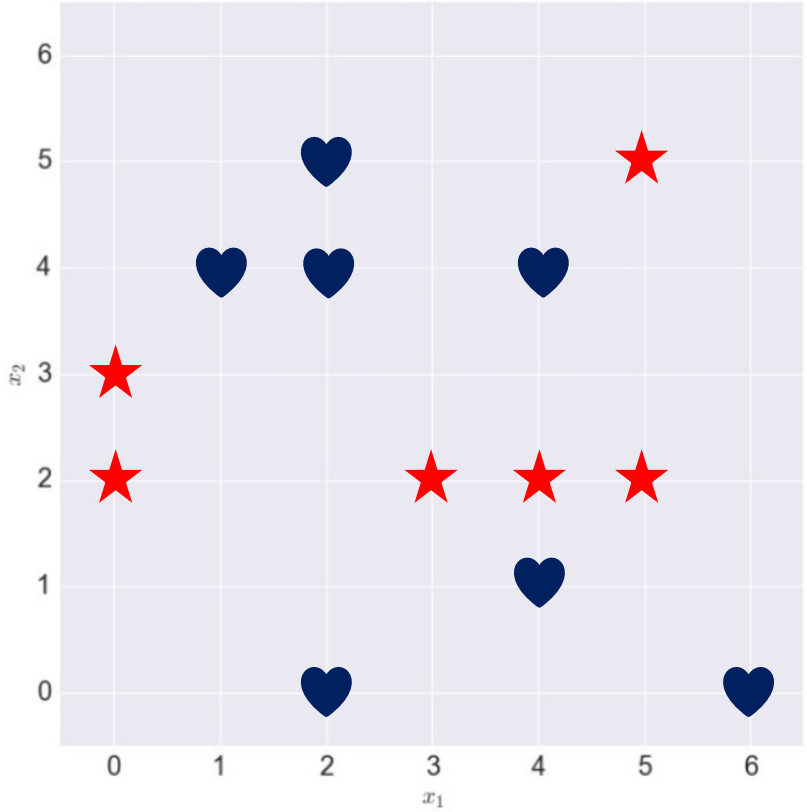
Our objective – Predicting a class

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

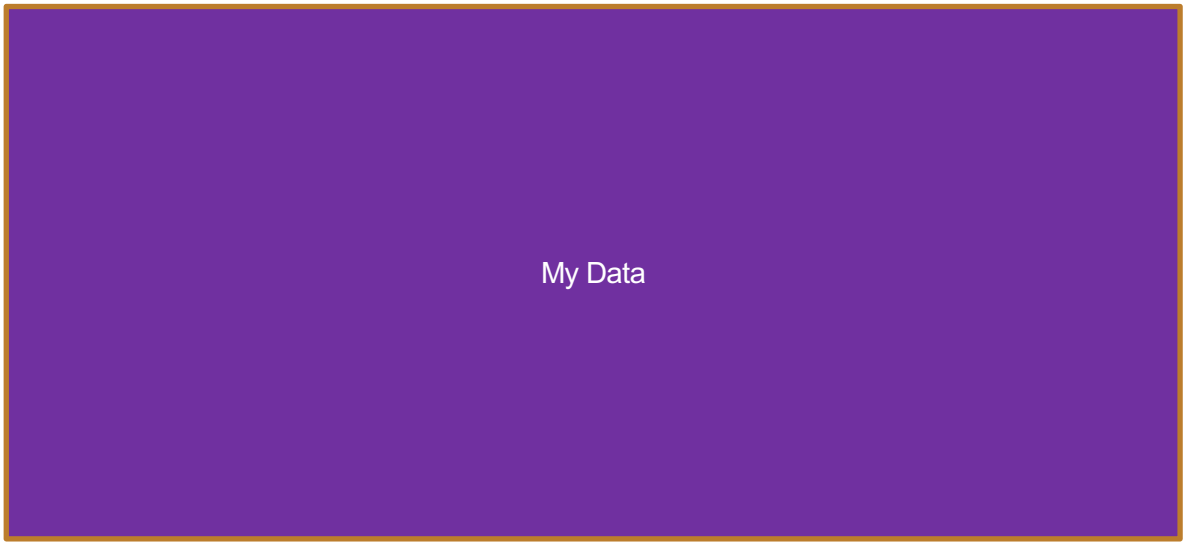


How do we know if it works?

Classify ones we already know the answer to!



Homework 1 – Training & Test Sets



Train a KNN on everything

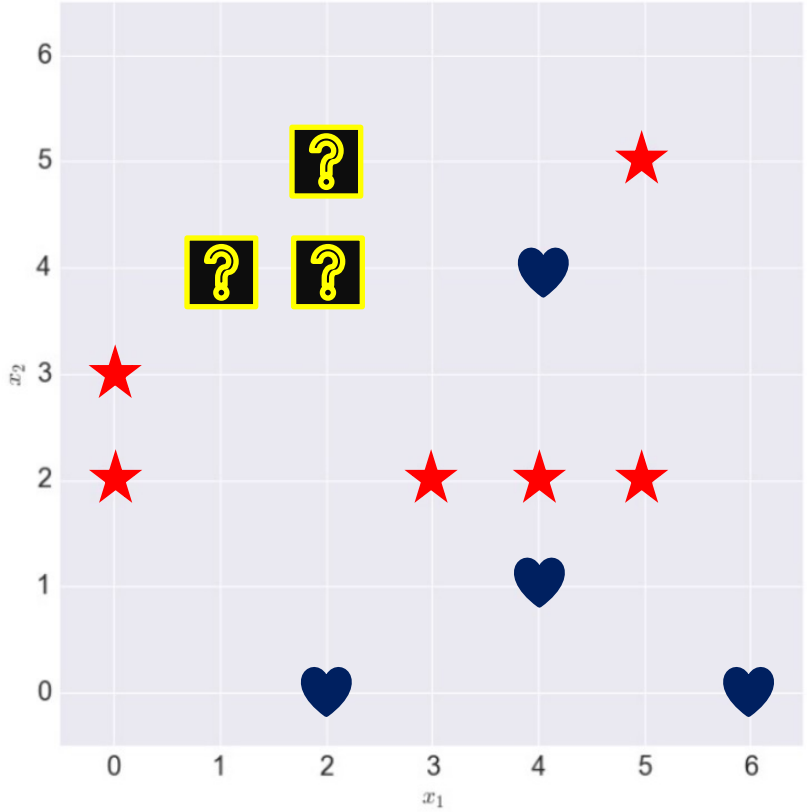
K = 1

What does my nearest neighbor to X look like?

Homework 1 – Training & Test Sets

Divide it into training and testing sets!

How do we know if it works?

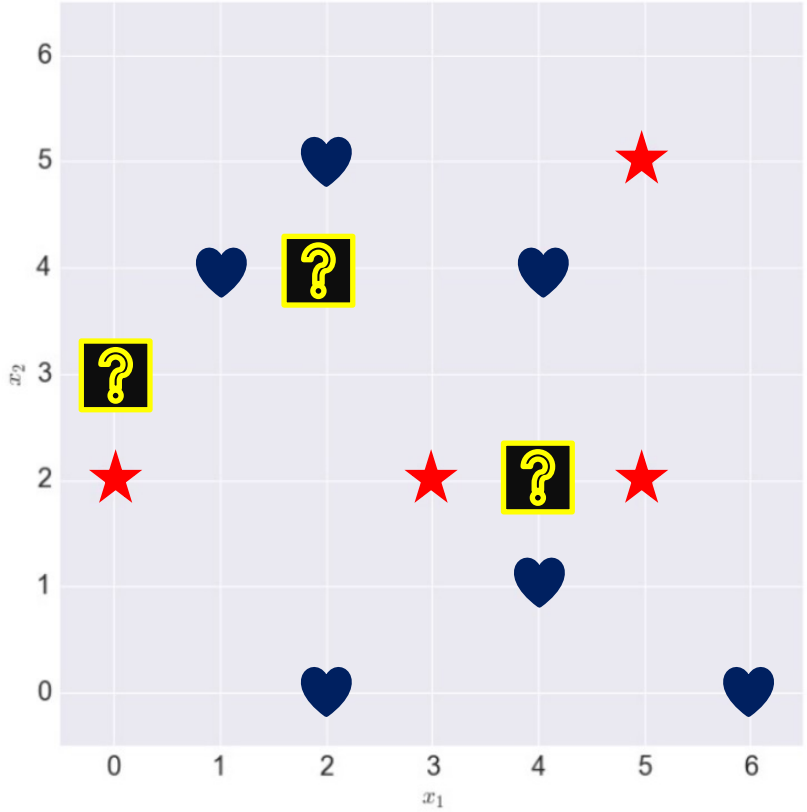


Homework 1 – Training & Test Sets

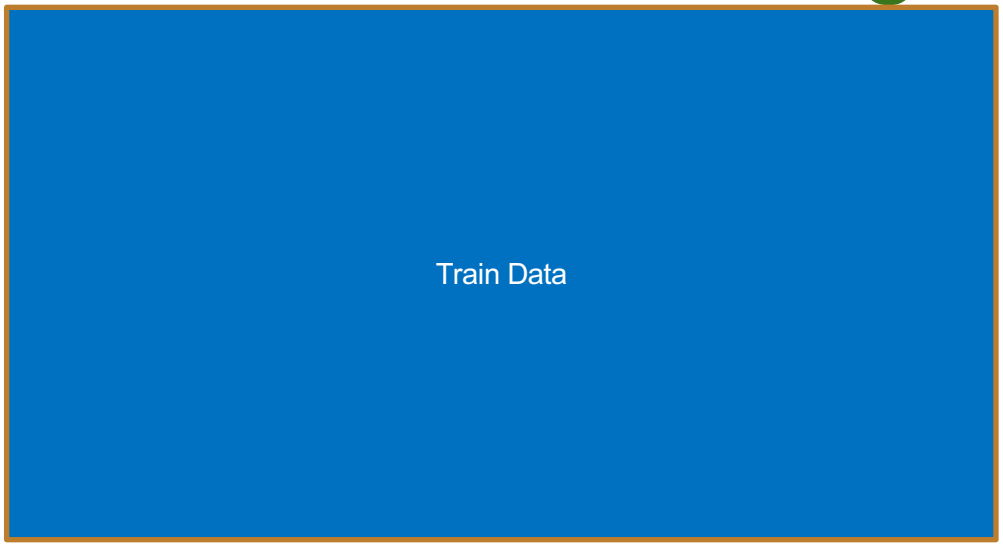
I pull out the last 20% of samples

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 dataset