Machine Learning and Data Mining

Lecture 0.1: Introduction, Class Logistics, and Sources of Error

1

Who is this guy?

2

CS434 – ML + DM
Today's Learning Objectives

Be able to answer:

- What is machine learning (ML)?
- When is it appropriate to use?

- How is this class structured?
- How will I be assessed / graded?

- How can I categorize different ML techniques?
- What is the standard ML formalism?
- Why must we make assumptions to learn?
- What are sources of error in ML?

- What is the idea behind k-nearest neighbors?

3

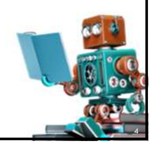
What is Machine Learning?

"Field of study that gives computers the ability to learn without being explicitly programmed"
- Arthur Samuel

"Any change in a system that allows it to perform better the second time on repetition of the same task or on another task drawn from the same distribution"
- Herbert Simon

"Algorithms that automatically detect patterns in data and use the uncovered patterns to predict future data or other outcomes of interest"
- Kevin Murphy

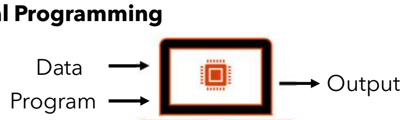
"Algorithms that improve their performance (P) at some task (T) with experience (E)"
- Tom Mitchell



4

What is Machine Learning? (Arthur Samuel's Answer)

Traditional Programming



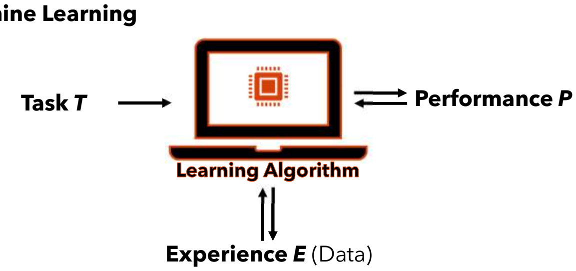
Machine Learning

"Field of study that gives computers the ability to learn without being explicitly programmed"
- Arthur Samuel

5

What is Machine Learning? (Tom Mitchell's Answer)

Machine Learning

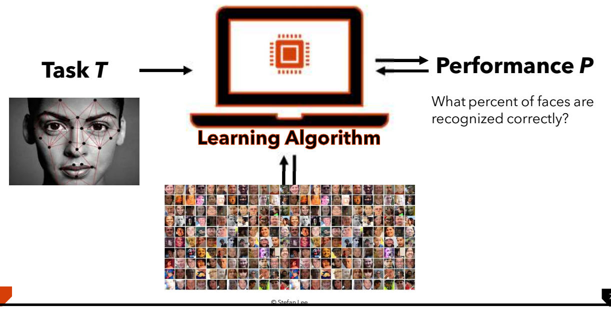


"Algorithms that improve their performance (P) at some task (T) with experience (E)"
- Tom Mitchell

6

What is Machine Learning? (Tom Mitchell's Answer)

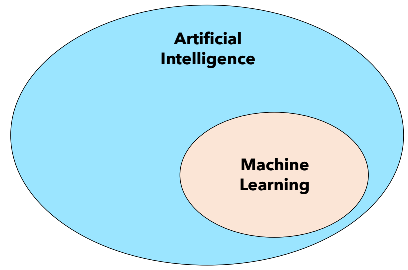
Machine Learning



CS 434 | © Stefan Lee | 7

7

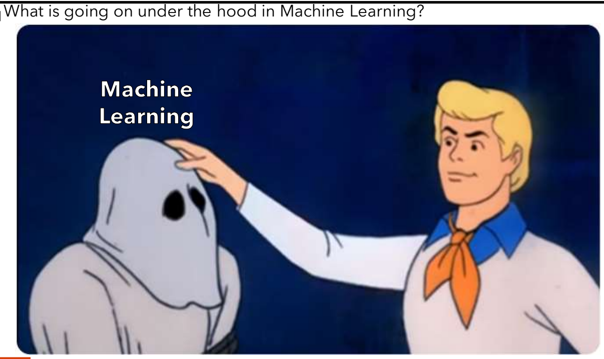
What is Machine Learning? (Relative to AI)



CS 434 | © Stefan Lee | 8

8

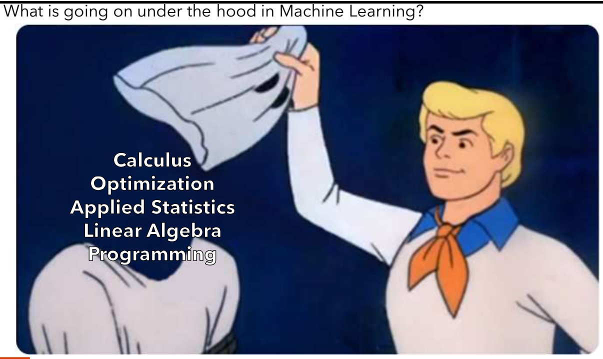
What is going on under the hood in Machine Learning?



CS 434 | © Stefan Lee | 9

9

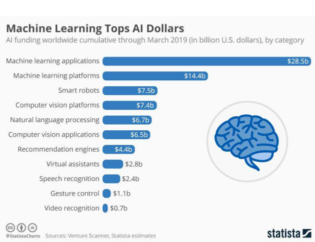
What is going on under the hood in Machine Learning?

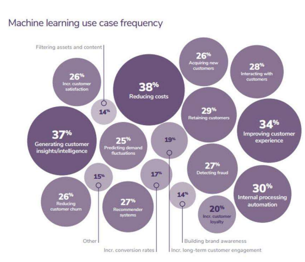


CS 434 | © Stefan Lee | 10

10

What is Machine Learning? (Relative to Business)





CS 434 | © Stefan Lee | 11

11

Where is Machine Learning used?

- **Machine learning is already the preferred approach to:**
  - Speech recognition
  - Natural Language Processing
  - Computer Vision
  - Robotic Control
  - Recommender Systems
  - …
- **This trend is growing**
  - Improved machine learning algorithms
  - Increased data capture and availability
  - Increased computing power
  - Increased *demand* for personalization to users and environments

CS 434 | © Stefan Lee | 12

12

When do we need learning?

13

When do we need learning?

1) Situations where humans can perform a task but can't describe how programmatically.

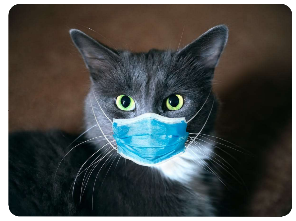


Image Classification

14

When do we need learning?

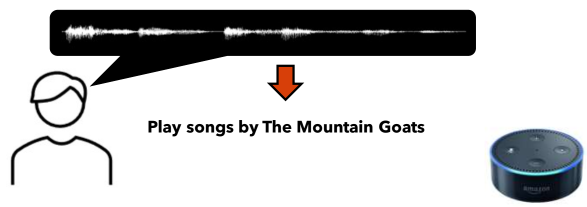
1) Situations where humans can perform a task but can't describe how programmatically.

Video Segmentation

15

When do we need learning?

1) Situations where humans can perform a task but can't describe how programmatically.

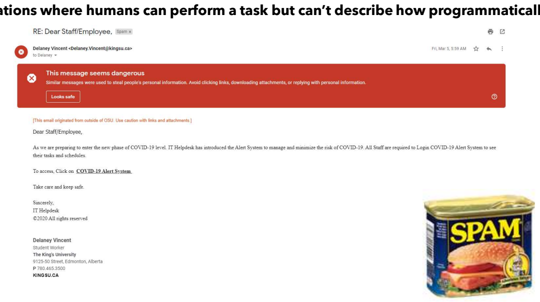


Speech Recognition

16

When do we need learning?

1) Situations where humans can perform a task but can't describe how programmatically.



Spam Detection

17

When do we need learning?

2) Situations where the desired function is different for each individual user

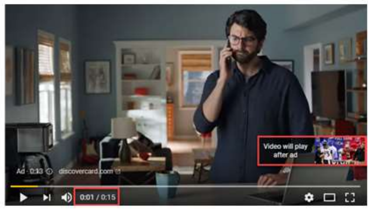
Recommendation Systems

18

When do we need learning?

2) Situations where the desired function is different for each individual user

Ad Targeting



Ad Targeting

19

When do we need learning?

2) Situations where the desired function is different for each individual user

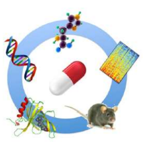
Can I ask who is speaking so I can better understand your voice in the future?

Personalization

20

When do we need learning?

3) Situations where human experts do not have sufficient knowledge (or time)



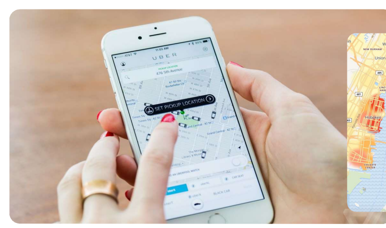
Based on molecular structure, predict effect.

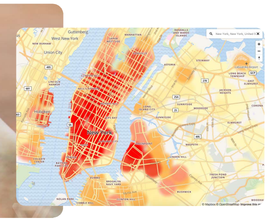
Drug / Material Discovery

21

When do we need learning?

4) Situations where the function of interest is changing rapidly.





Anticipating Demand

22

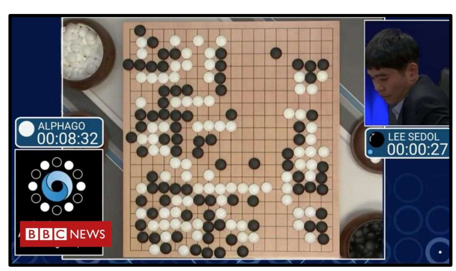
When do we need learning?

4) Situations where the function of interest is changing rapidly.

Predicting Stock Markets

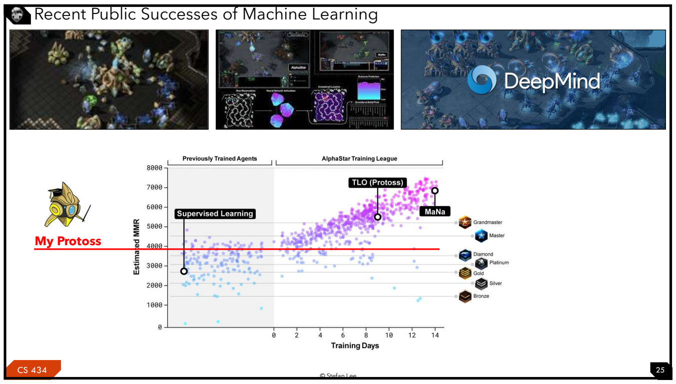
23

Recent Public Successes of Machine Learning

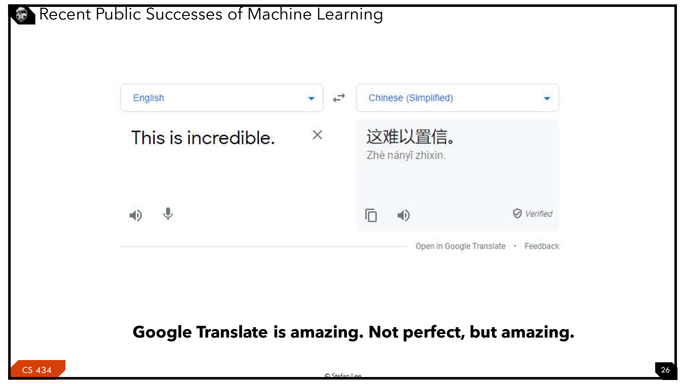


AlphaGo

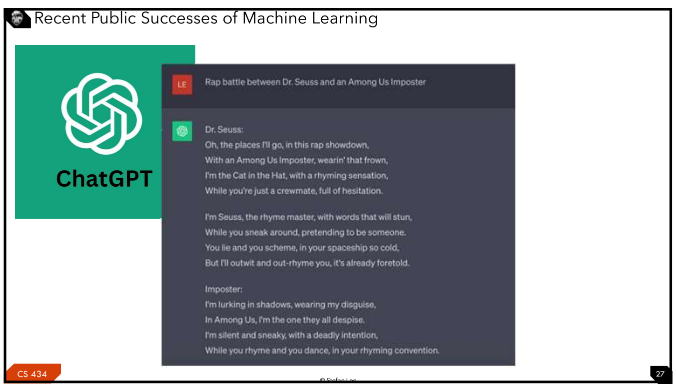
24



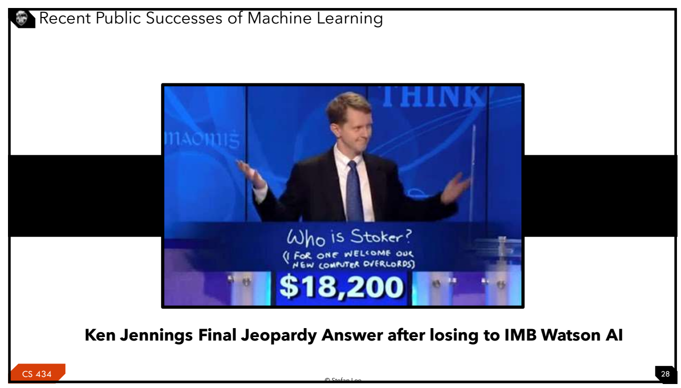
25



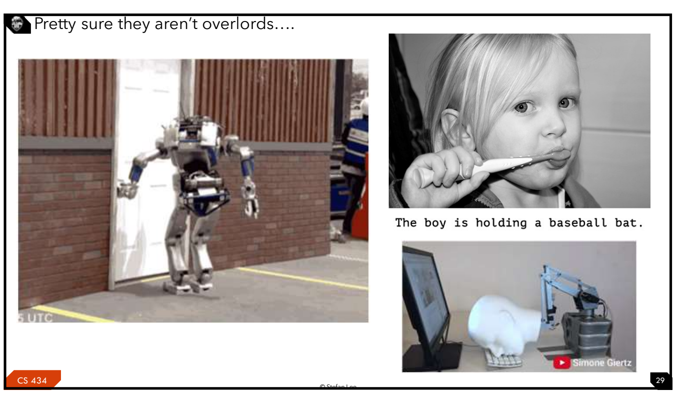
26



27



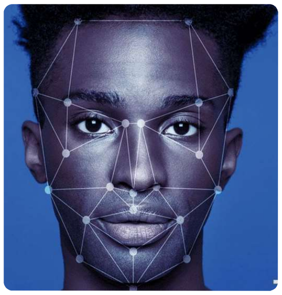
28



29

30

Recent Public Failures of Machine Learning



" we found empirical evidence for the existence of demographic differentials in the majority of the face recognition algorithms we studied"

- *Patrick Grother, Lead author of NIST Study on Facial Recognition*

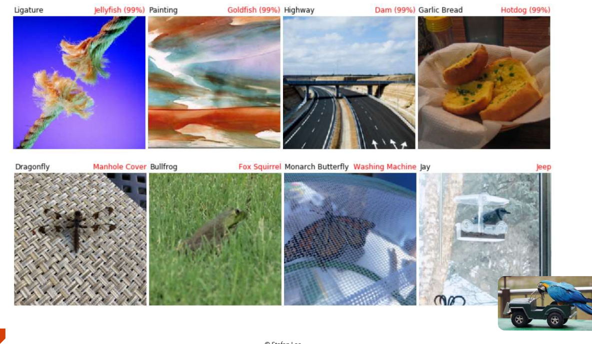
Tested facial recognition products from >90 companies. For many, found false-identification rate **100x** higher for individuals of African or Asian descent than for Caucasians.

Many groups lobby for banning facial recognition technology in public places.

CS 434 © Stefan Lee 31

31

Recent Public Failures of Machine Learning



CS 434 © Stefan Lee 32

32

Question Break!

Someone approaches you with a business proposal. They want you to build a machine learning system to predict whether a student will pass a class or not based on their transcripts. But due to FERPA, they have no data from previous classes and lack the budget to procure any data. Also, to protect their innovative idea, they won't let you ask for donated data publicly.

They are offering you 40% of the business equity – which they **insist** will be worth a fortune right after you finish. Obviously, no pay for now.

A Accept the job and start planning how big of a yacht you will buy with your share.

B Decline the job and avoid talking to this person again in the future.

C

D

CS 434 © Stefan Lee 33

33

CS434 – ML + DM

Today's Learning Objectives

Be able to answer:

- ~~What is machine learning (ML)?~~
- ~~When is it appropriate to use?~~

- How is this class structured?
- How will I be assessed / graded?

- How can I categorize different ML techniques?
- What is the standard ML formalism?
- Why must we make assumptions to learn?
- What are sources of error in ML?

- What is k-nearest neighbors?

© Stefan Lee 34

34

CS434 – Instructional Team

Instructor: Stefan Lee
Class Period: M\W 4-5:50pm in COVL 216 (You're here. Well done. Kudos)
Office Hours: M 11:30-1:00pm on Zoom or KEC3013

Teaching Assistant Team (Office hours TBD)

Eric Slyman

Xiangxi Shi

Jing Wang

Ian Tassin

CS 434 © Stefan Lee 35

35

CS434 – What do you need to do for points?

Grading / Assessment:

- **Quiz Participation** (10%) – Regular conceptual quizzes / surveys. Graded on completion -- where completion means making a reasonable effort at responding to the questions, judged at the TA's discretion. Always due on Sundays.

- **Assignments** (50%) – Five assignments over the term each due on Canvas. First one out today and due next Wednesday.

- **Exams** (40%) – One midterm and a final – 20% each. In-class. Open book / note.

CS 434 © Stefan Lee 36

36

CS434 – Don't be late or cheat.

Late Work Policy:
Late assignments will be accepted within 48 hours of the deadline with a penalty. You will receive 95% of your regular grade for late submission submitted within 24 hours of the deadline. Late submission submitted between 24 and 48 hours past the deadline will receive 90% of your regular grade. No submission will be accepted after 48 hours. No late quizzes or exams will be accepted.

Collaboration Policy:
Students are expected to work independently. High-level discussion of topics is allowed, but any low-level details, sharing of code, or collaborative programming is prohibited. Likewise, the use of code from online resources is prohibited. You do not want to get caught doing this.

Only Exception: Work however closely as you want for homework 0.

37

CS434 – Rough Schedule

Week 0	Introduction and Course Logistics	HW0 Out.
Week 1	K-Nearest Neighbor and MLE/MAP	HW0 Due. HW1 Out.
Week 2	MAP (cont.) and Linear Regression	
Week 3	Polynomial Regression and Logistic Regression	HW1 Due. HW2 Out.
Week 4	Perceptron and SVM	
Week 5	SVM (cont.) and Kernel Methods	HW2 Due.
Week 6	Midterm (Review + Exam)	**Midterm Exam.** HW3 Out.
Week 7	Multiclass Classifiers and Neural Networks	
Week 8	Decision Trees and Ensemble Methods	HW3 Due. HW4 Out.
Week 9	Ensemble Methods (cont.) and Clustering	
Week 10	Dimensionality Reduction and Final Review	HW4 Due.
Finals Week		**Final Exam.**
Forever after	Bask in the glory of your newfound machine learning knowledge	

Schedule subject to change

38

CS434 – Background Material

Am I well prepared for this course?
Maybe? Are you comfortable with most of these (or willing to learn):

- Manipulating probabilities and distributions [**Probability**]
- Taking derivatives and reasoning about integrals [**Calculus**]
- Performing common matrix operations [**Linear Algebra**]
- Standard computer science concepts like… [**Computer Science**]
  - Programming (skeleton code will be in Python)
  - Computational complexity (big O notation)
- Handling abstract mathematical concepts **[Discrete Math]**
  - Proving things by manipulating equations

Homework 0 is out today. Not worth many points and graded on effort but should give you a sense for if you'll need to review some things for this class.

39

CS434 – Reaching for Success

How to do well in this class?

Come to class on time. Pay attention. Take notes. Ask questions during lecture or afterwards via Discord or in office hours. Rewatch recordings for parts you missed.

Start your homework early. You will *always have a homework to do but we don't plan for it to actually take the full two weeks you have to do it.* That said, don't wait till the last day – you will very probably not do well on it.

If you don't know background material but are still keen to learn ML, ask for help and study hard.

Just like saying "Eat healthy, sleep eight hours, avoid stress, and exercise regularly", I know this is obvious advice. Following it is the hard part.

40

CS434 – Machine Learning and Data Mining

What this class is:

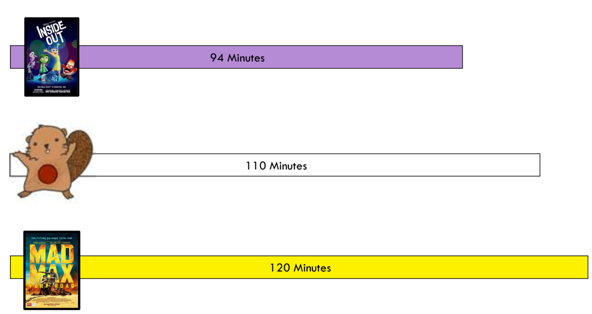
An introduction to the fundamentals of machine learning. You will do math to derive and understand learning algorithms. You will implement learning algorithms from scratch and apply them to provided datasets.

What this class isn't:

A tutorial on how to do machine learning in framework X, Y, or Z. Also, this is not a deep learning class.

41

CS434 – Day-to-Day



42

CS434 – Day-to-Day

110 Minutes

50 Minutes | 50 Minutes

1-5 min admin | 5-10 min break

43

Question Break!

If homework zero seems hard to me, Professor Stefan is telling me that I should:

A Drop the class and give up on my dreams of understanding machine learning ☹

B View the recorded review sessions in areas you find challenging and ask questions in Discord

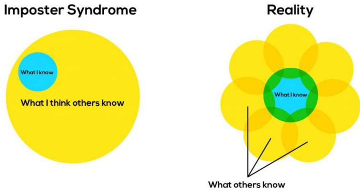
C Consider whether the extra work of refreshing or learning these skills fits into my life

D Invest all my money in Dogecoin
How could you think this is what I was saying? So wrong. Much not listening. (This is not financial advice).

44

CS434 – Machine Learning and Data Mining

It is okay not to know something



If you don't recognize a symbol or notation I use, let me know. I'm happy to explain it and then you will – easy fix.

45

CS434 – ML + DM

Today's Learning Objectives

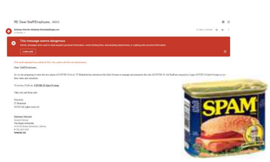
Be able to answer:

- ~~What is machine learning (ML)?~~
- ~~When is it appropriate to use?~~
- ~~How is this class structured?~~
- ~~How will I be assessed / graded?~~
- How can I categorize different ML techniques?
- What is the standard ML formalism?
- Why must we make assumptions to learn?
- What are sources of error in ML?
- What is the idea behind k-nearest neighbors?

46

Machine Learning In a Nutshell

- **Thousands of machine learning algorithms**
  - Hundreds new every year
- **Decades of ML research oversimplified:**
  - All of Machine Learning:
    - Learn a mapping from input to output **g: X → Y**
      X: emails, Y: {spam, notspam}



47

Machine Learning In a Nutshell

Input: x (images, text, emails…)

Output: y (spam or non-spam…)

(Unknown) Target Function
f: X → Y (the "true" mapping / reality)

Data
$(x_1,y_1), (x_2,y_2), \ldots, (x_N,y_N)$ (our observations of the world)

Model / Hypothesis Class
$\mathcal{H}$ = {g: X → Y} (the space of possible models)

48

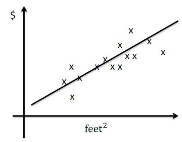
Machine Learning Topics

Supervised Learning
Learn to predict output from input given example (x,y) pairs

Regression: Predicting a (or multiple) continuous values as output

$x \rightarrow g(x) \rightarrow y$

Continuous



Example: Predicting the price of a house based on square footage

CS 434 © Stefan Lee 49

49

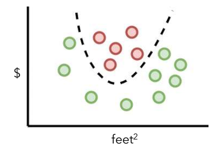
Machine Learning Topics

Supervised Learning
Learn to predict output from input given example (x,y) pairs

Classification: Predicting a (or multiple) discrete variable as output

$x \rightarrow g(x) \rightarrow y$

Discrete



Example: Predicting if a house will sell based on price and square footage

CS 434 © Stefan Lee 50

50

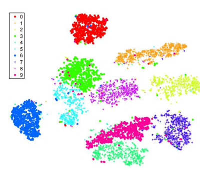
Machine Learning Topics

Unsupervised Learning
Only given a set of input instances (x)

Clustering: Discover self-similar groups within the data

$x \rightarrow g(x) \rightarrow y$

Discrete
"Cluster"
Label



Example: Find groups of houses with similar properties

CS 434 © Stefan Lee 51

51

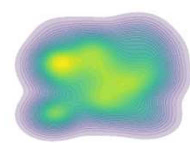
Machine Learning Topics

Unsupervised Learning
Only given a set of input instances (x)

Density Estimation: Estimate the underlying distribution generating our data

$x \rightarrow g(x) \rightarrow y$

Continuous
Density
Estimate
y=P(x)



Example: Estimate how likely is a given house with these properties

CS 434 © Stefan Lee 52

52

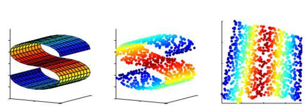
Machine Learning Topics

Unsupervised Learning
Only given a set of input instances (x)

Dimensionality Reduction: Represent high-dim data as low-dim data

$x \rightarrow g(x) \rightarrow y$

Continuous
representation
of x



Example: Finding a 2D subspace in a 3D feature space

CS 434 © Stefan Lee 53

53

Machine Learning Topics

Reinforcement Learning
Multiple interactions with an environment yield a reward signal. No optimal outputs are provided. Goal: learn a policy that optimizes rewards

$x \rightarrow g(x) \rightarrow y$

Observation

Action
(continuous or
discrete)

Example: Navigating in an environment to a goal.

Barely going to touch this one.

CS 434 © Stefan Lee 54

54

No Assumptions → No Learning

It's clear we are interested in mappings from input to output. Let's get concrete.

Millow (legally distinct from Zillow)

Goal: Predict whether a house will sell within a month

Your records contain d attributes for each house…

- Lets assume binary: HasPool? >300sqft?, IsHaunted?...
- Implies $x \in \{0,1\}^d$

… and whether it sold within a month.

- Implies y $\in \{0,1\}$

How many possible mappings from x to y are there?

55

No Assumptions → No Learning

How many possible mappings $g: \{0,1\}^d \rightarrow \{0,1\}$ from x to y are there?

What is a mapping? For each unique x, provide an output y.

x	y	x	y	x	y
0 0 … 0	0	0 0 … 0	1	0 0 … 0	1
0 0 … 1	1	0 0 … 1	1	0 0 … 1	0
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
1 1 … 1	0	1 1 … 1	0	1 1 … 1	0

…

Making no assumptions about the function -- 2^{2^d} possible!

56

No Assumptions → No Learning

Will data save us?

Each unique data point fills in one of these rows for us. Say we have n.

x	y
0 0 … 0	1
0 0 … 1	
.	
.	
.	
1 1 … 1	0

Still 2^{2^d-n} possible mappings. Far, far too many.

We need to make some sort of assumption about the relationship between input and output.

57

Question Break!

You want to predict shoe size from someone's height. You are given a dataset of height/shoe-size pairs.

This is an example of a:

- **A** supervised classification problem
- **B** unsupervised clustering problem
- **C** supervised regression problem
- **D** unsupervised density estimation problem

58

Question Break!

You want to estimate how likely it is for someone to have a certain shoe size / height combo. You are given a dataset of height/shoe-size pairs.

This is an example of a:

- **A** supervised classification problem
- **B** unsupervised clustering problem
- **C** supervised regression problem
- **D** unsupervised density estimation problem

59

Question Break!

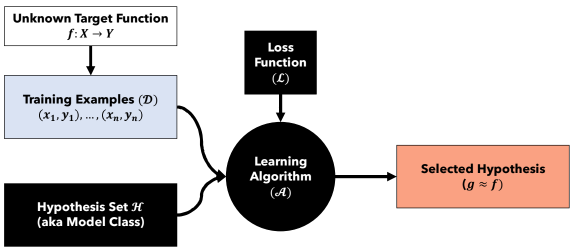
You want to group people by their shoe-size and height. You are given a dataset of height/shoe-size pairs.

This is an example of a:

- **A** supervised classification problem
- **B** unsupervised clustering problem
- **C** supervised regression problem
- **D** unsupervised density estimation problem

60

Machine Learning In a Nutshell



Observations of some phenomenon | Designed by us | Result of our learning

CS 434 · 61

61

Machine Learning In a Nutshell

Key Elements of A Machine Learning Application:

- **Representation / Model Class / Hypothesis Space ($\mathcal{H}$)**
  - How do we represent the function g we are trying to learn?
- **Evaluation / Objective / Loss Function ($\mathcal{L}$)**
  - How can we measure if a specific model is good or bad?
- **Optimization / Learning Algorithm ($\mathcal{A}$)**
  - How do we find models in our hypothesis space that perform well under our loss function?
- **Dataset / Experience / Training Examples ($\mathcal{D}$)**
  - What real-world phenomenon are we modelling?

CS 434 · 62

62

Representation / Model Class / Hypothesis Space ($\mathcal{H}$)

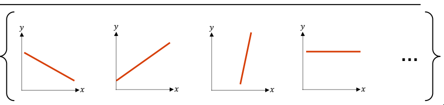
How do we represent the function g we are trying to learn?

- Linear Models
- Polynomial Models
- Decision trees / forests
- Sets of rules / Logic programs
- Instances (non-parametric models)
- Graphical models (Bayes/Markov nets)
- Neural networks
- Support vector machines
- Model ensembles
- Kernel Regression
- Etc.

Picking a model class defines a space of possible models we might learn:

Example

Linear Hypothesis Space
$\{g(x) | g(x) = b + wx\}$



CS 434 · 63

63

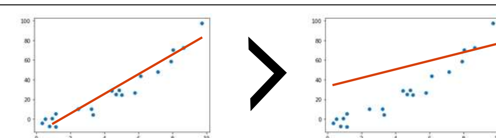
Evaluation / Objective / Loss Function ($\mathcal{L}$)

How can we measure if a specific model is good or bad?

- Accuracy
- Precision and recall
- Squared error
- Likelihood
- Posterior probability
- Cost / Utility
- Margin
- Entropy
- K-L divergence
- Etc.

Defining an objective or loss function induces an ordering of models in the hypothesis space

Example



CS 434 · 64

64

Optimization / Learning Algorithm ($\mathcal{A}$)

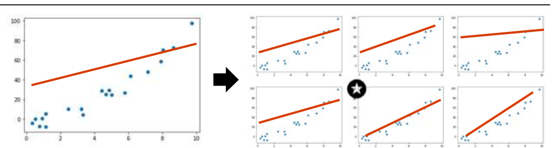
How do we find models in our hypothesis space that perform well under our loss function?

Discrete/Combinatorial optimization
Search Techniques
Graph algorithms (cuts, flows, etc)
…

Continuous optimization
Convex/Non-convex optimization
Linear programming
…

Learning algorithms determine how to search through the hypothesis space

Example

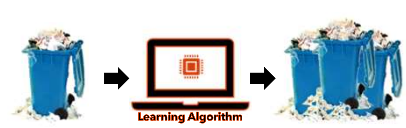


CS 434 · 65

65

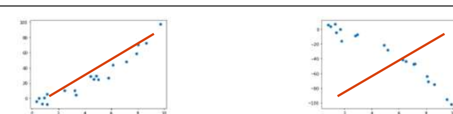
Dataset / Experience / Training Examples ($\mathcal{D}$)

What real-world phenomenon are we modelling?



Along with the loss function, determines what a "good" model looks like

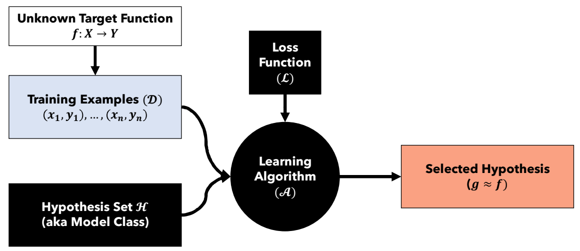
Example



CS 434 · 66

66

Machine Learning In a Nutshell



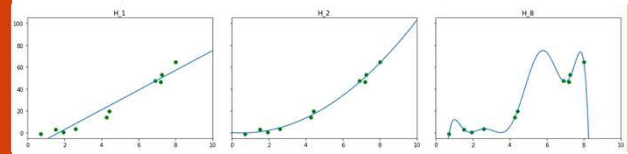
Observations of some phenomenon | Designed by us | Result of our learning

CS 434 | 67

67

Question Break!

Green points are our observed data, which fit do you think is best?



A Left One (First-degree polynomial fit)

B Middle One (2nd-degree polynomial fit)

C Right One (8th-degree polynomial fit)

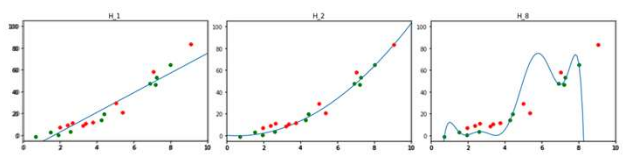
D I wasn't paying attention and the quiet woke me up. What are we talking about again?

CS 434 | 68

68

Overfitting, Underfitting, and Model Selection

We don't really care about loss on training data… we want our model to generalize.

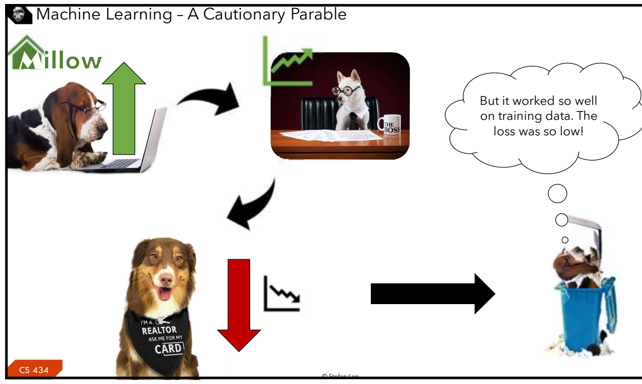


● Training data ● New data

CS 434 | 69

69

Machine Learning – A Cautionary Parable

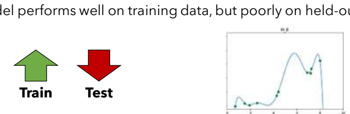


CS 434 | 70

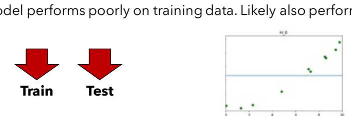
70

Overfitting, Underfitting, and Model Selection

Overfitting: Model performs well on training data, but poorly on held-out testing data.



Underfitting: Model performs poorly on training data. Likely also performs poorly on test data.



Model Selection: The exercise of finding a hypothesis space / model class that neither underfits or overfits. ← One of the reasons doing ML well is hard.

CS 434 | 71

71

A Practitioner's ML Loop (for supervised learning)

Set up a supervised learning problem

Data collection of input-output pairs where we know the output is right

Representation: Choose how to represent the data

Modelling: Choose a hypothesis space / model class

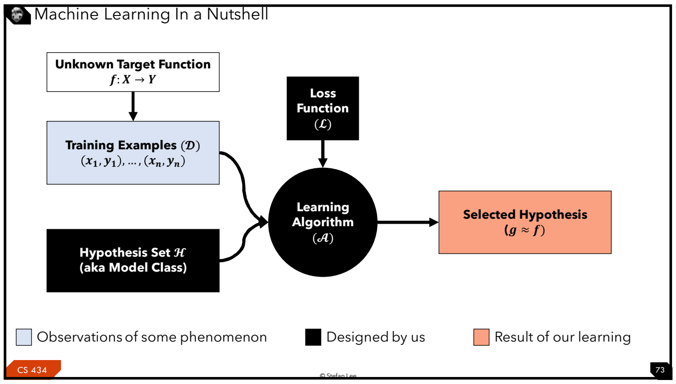
Learning/Training/Estimation: Find the best member of the chosen hypothesis class

Model Selection: Try different models and evaluate them on held out data (more on this later)

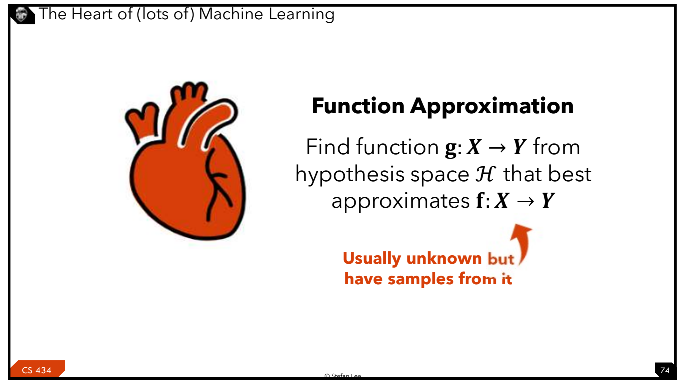
If happy, stop. Else repeat/refine one or more of the above.

CS 434 | 72

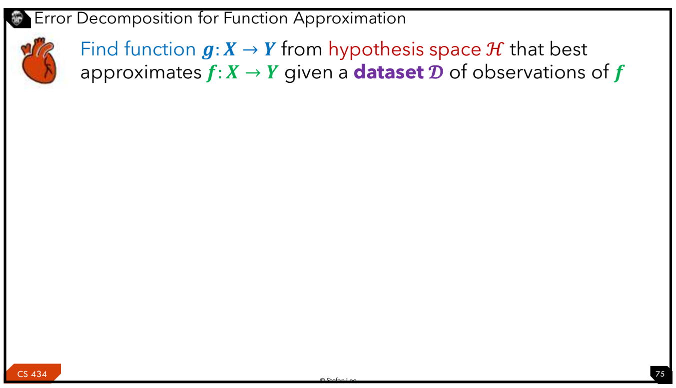
72



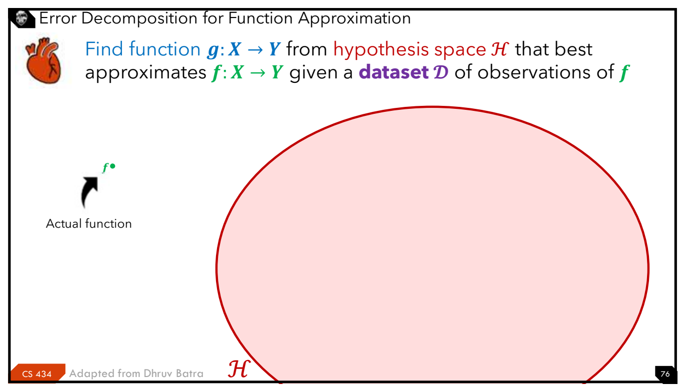
73



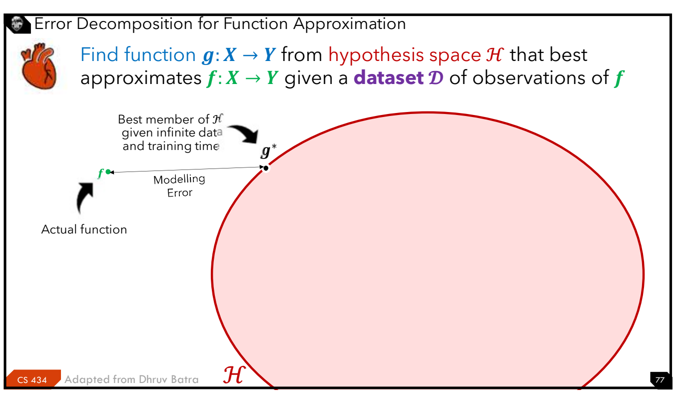
74



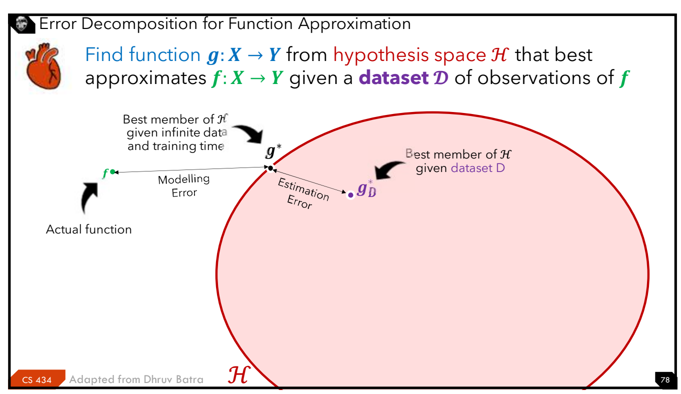
75



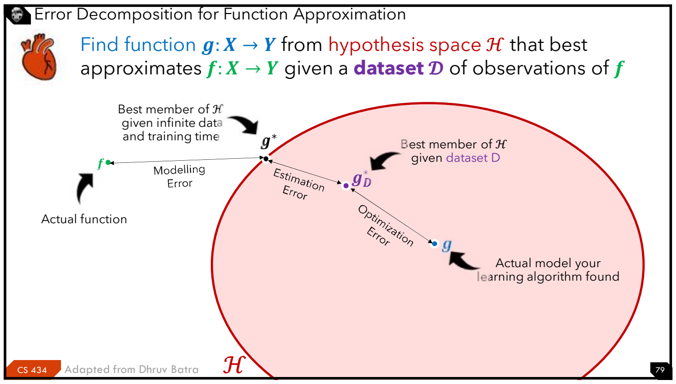
76



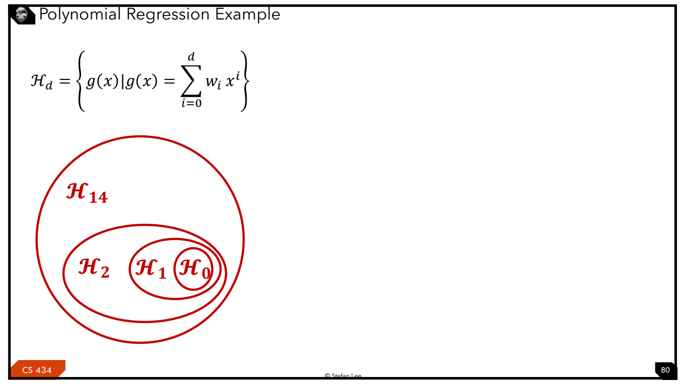
77



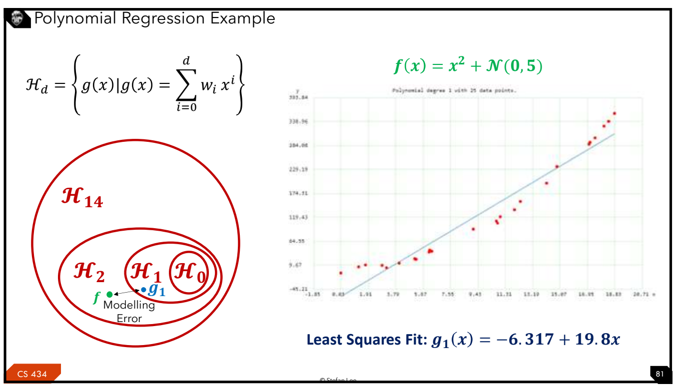
78



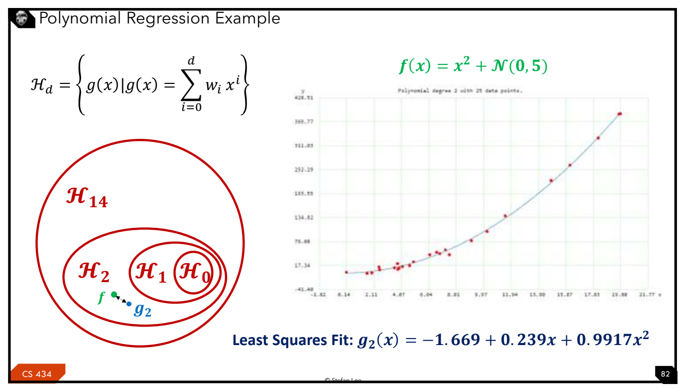
79



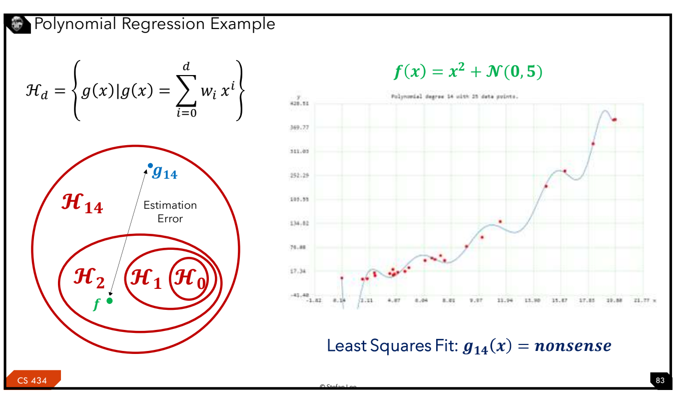
80



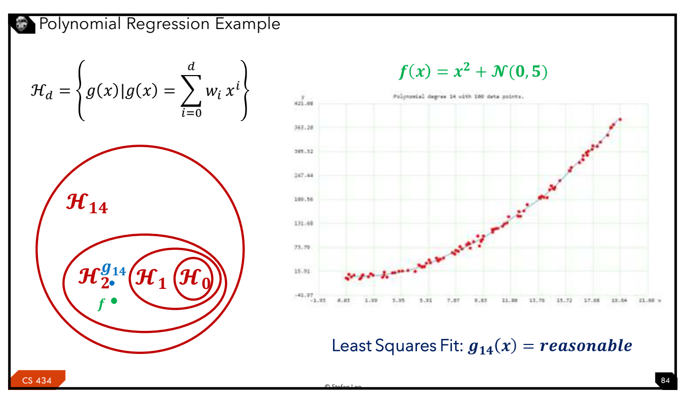
81



82



83



84

Question Break!

Which source(s) of error tend(s) towards zero with infinite data?

- **A** Optimization Error
- **B** Estimation Error
- **C** Modelling Error
- **D** Bayes Error

CS 434 · 85

85

Question Break!

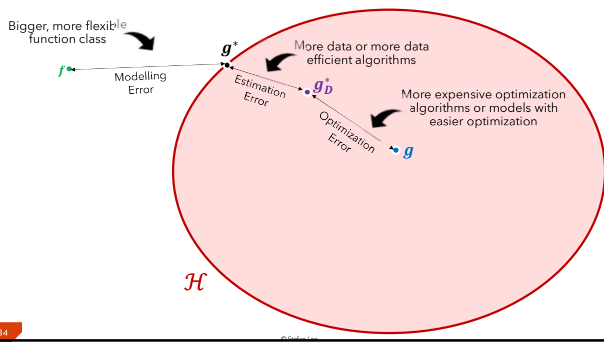
Which source(s) of error tend(s) towards zero as computation goes to infinity?

- **A** Optimization Error
- **B** Estimation Error
- **C** Modelling Error
- **D** Bayes Error

CS 434 · 86

86

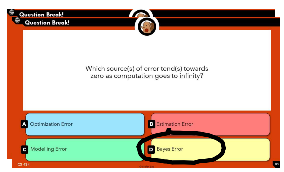
Reducing Sources of Error



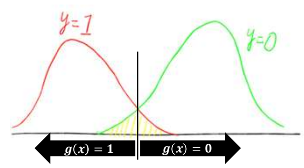
CS 434 · 87

87

What is Bayes Error?



Bayes Error
Irreducible error inherit in the function being approximated – nothing we can fix.



CS 434 · 88

88

What is Bayes Error?

Bayes Error
Irreducible error inherit in the function being approximated – nothing we can fix.

Bayes Optimal Classifier:

Suppose we know the true distribution $P(Y|X)$ and for each x we encounter we predict:

$$\hat{y} = argmax_y\ P(Y = y|X = x)$$

How often would we be wrong?

- It depends on how random $P(y|x)$ is. Consider a binary problem $y \in \{0,1\}$
  - If $P(y = 0 \mid X = x) = 0.6$ → Wrong 40% of the time for $\mathbf{x}$.
- Could we do better? Nope. Proof can be found here: https://en.wikipedia.org/wiki/Bayes_classifier

CS 434 · 89

89

CS434 – ML + DM

Today's Learning Objectives

Be able to answer:

- ~~What is machine learning (ML)?~~
- ~~When is it appropriate to use?~~

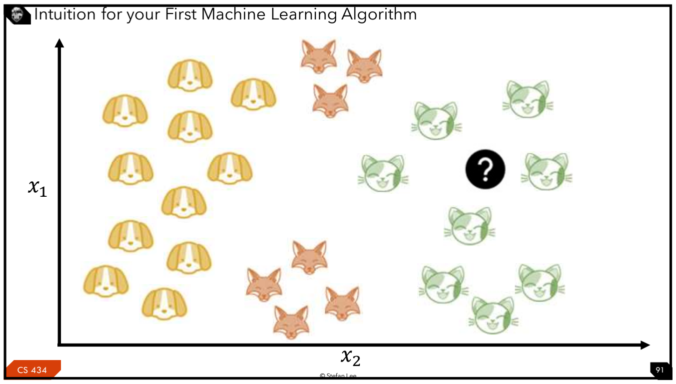
- ~~How is this class structured?~~
- ~~How will I be assessed / graded?~~

- ~~How can I categorize different ML techniques?~~
- ~~What is the standard ML formalism?~~
- ~~Why must we make assumptions to learn?~~
- ~~What are sources of error in ML?~~

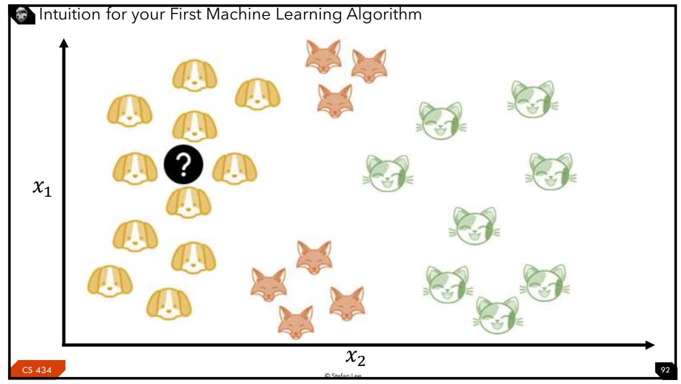
- What is the idea behind k-nearest neighbors?

90

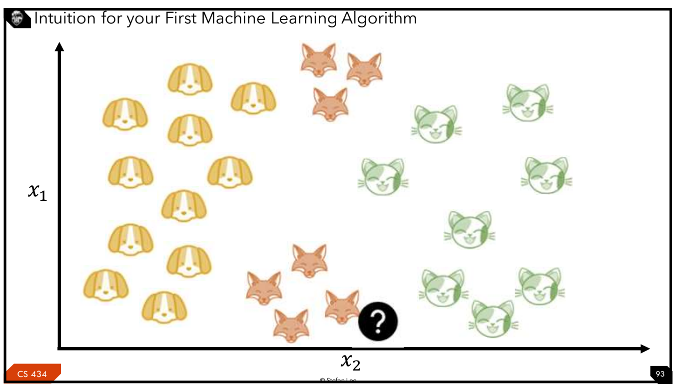
90



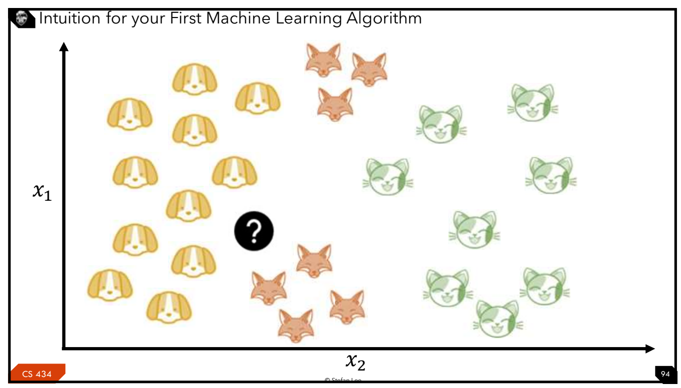
91



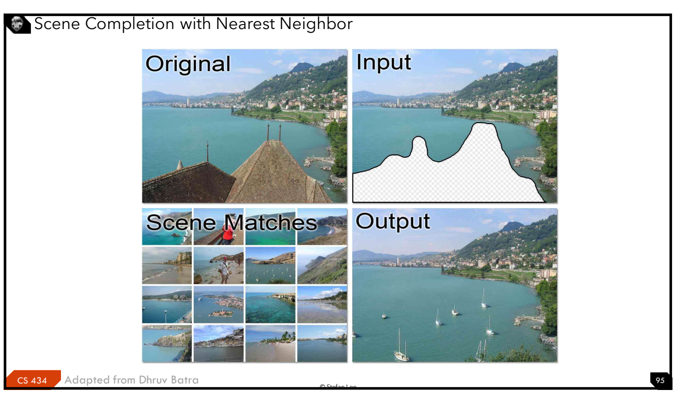
92



93

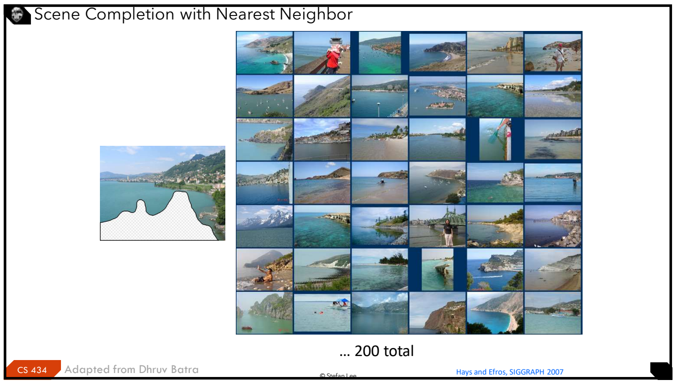


94

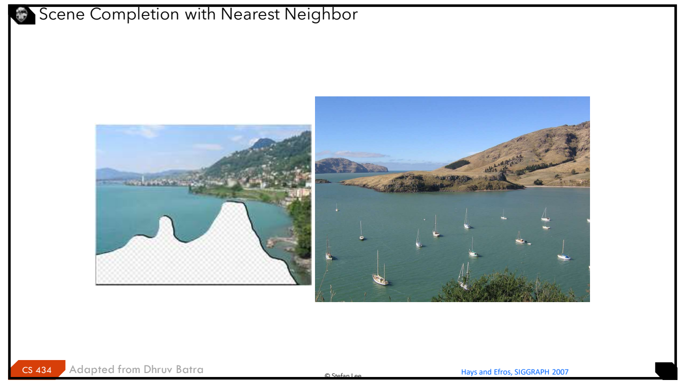


95

96



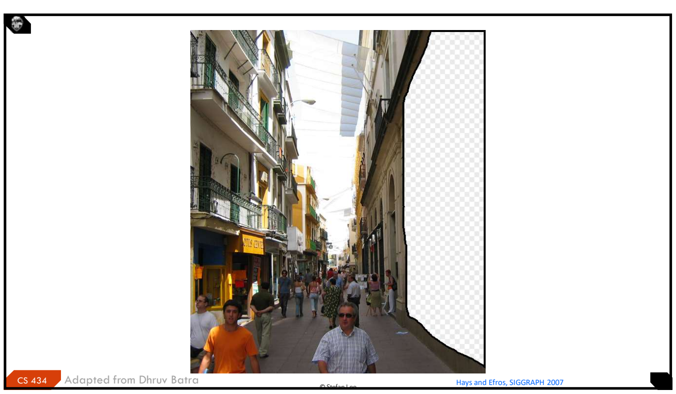
97



98

99

100



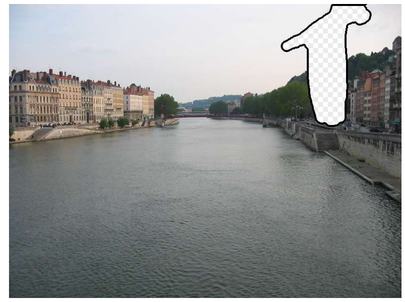
101

102

CS 434 Adapted from Dhruv Batra

Hays and Efros, SIGGRAPH 2007

103



CS 434 Adapted from Dhruv Batra

Hays and Efros, SIGGRAPH 2007

104

CS 434 Adapted from Dhruv Batra

Hays and Efros, SIGGRAPH 2007

105

Next time

Next Time: We'll go more in-depth with kNN and discuss weighting functions, classification/regression, and the curse of dimensionality.

CS 434

106