Quiz

Next Thursday, Sept 6

- Will focus on terminology and notation (mostly multiple choice)
- Might include something from the reading for that day (PML Ch 2)

Let me know ahead of time if you can't make it

- Excused quizzes will be excluded from your grade

What is Machine Learning?

INFO-4604, Applied Machine Learning
University of Colorado Boulder

August 28-30, 2018

Prof. Michael Paul

Definition

Murphy:

- “a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data”

Definition

Murphy:

- "a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data"

- predict = guess the value(s) of unknown variable(s)
  - (not necessarily prediction of future… c.f. *forecasting*)
- future data = data you haven't seen before

Types of Learning

- Supervised learning
  - Goal: Prediction
- Unsupervised learning
  - Goal: Discovery

Supervised Learning

Learn how to *predict* an output from a given input.

- Given a photo, identify who is in it
- Given an audio clip, identify the song
- Given a patient's medical history, estimate how likely they will need follow-up care within a month

Supervised Learning

Two types of prediction:

- Classification
  - Discrete outputs (typically categorical)
- Regression
  - Continuous outputs (usually)

If you need to brush up on these definitions, read Ch. 1 of *OpenIntro Statistics*.

Classification

- Document classification
  - Is this email spam?
  - Is this tweet positive toward this product?
  - Is this review/article real?
- Image classification
  - Is this a photo of a cat?
  - Which letter or number is written here?
- Object recognition
  - Identify the faces in this image
  - Identify pedestrians in this video

Classification

A classification algorithm is called a **classifier**

Classifiers require examples of inputs paired with outputs

- Called **training data**

Classifiers learn from training examples to map input to output

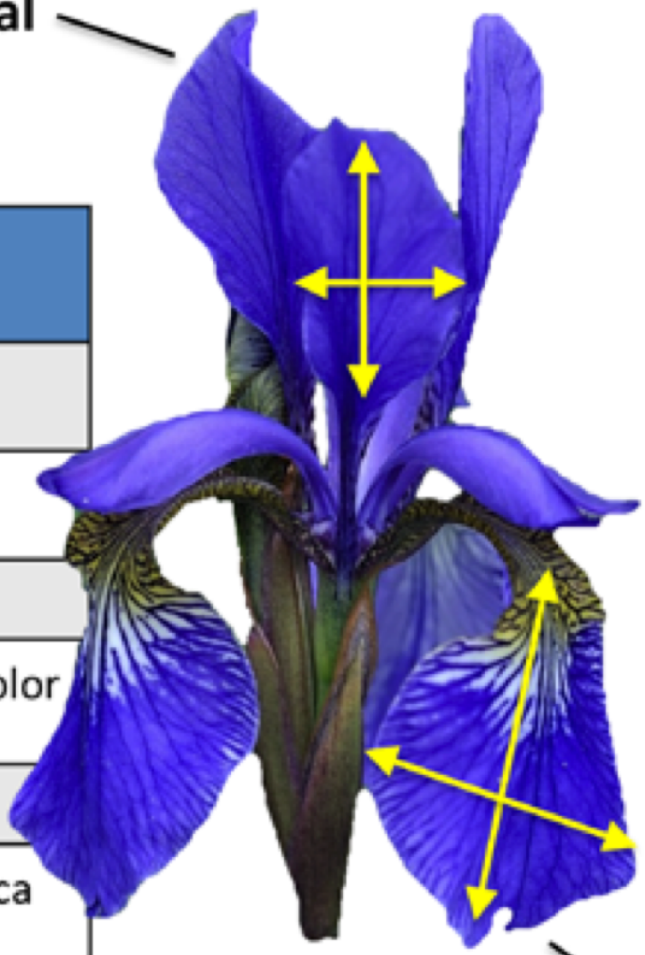
- Then when a classifier encounters new data where the output is unknown, it can make a prediction

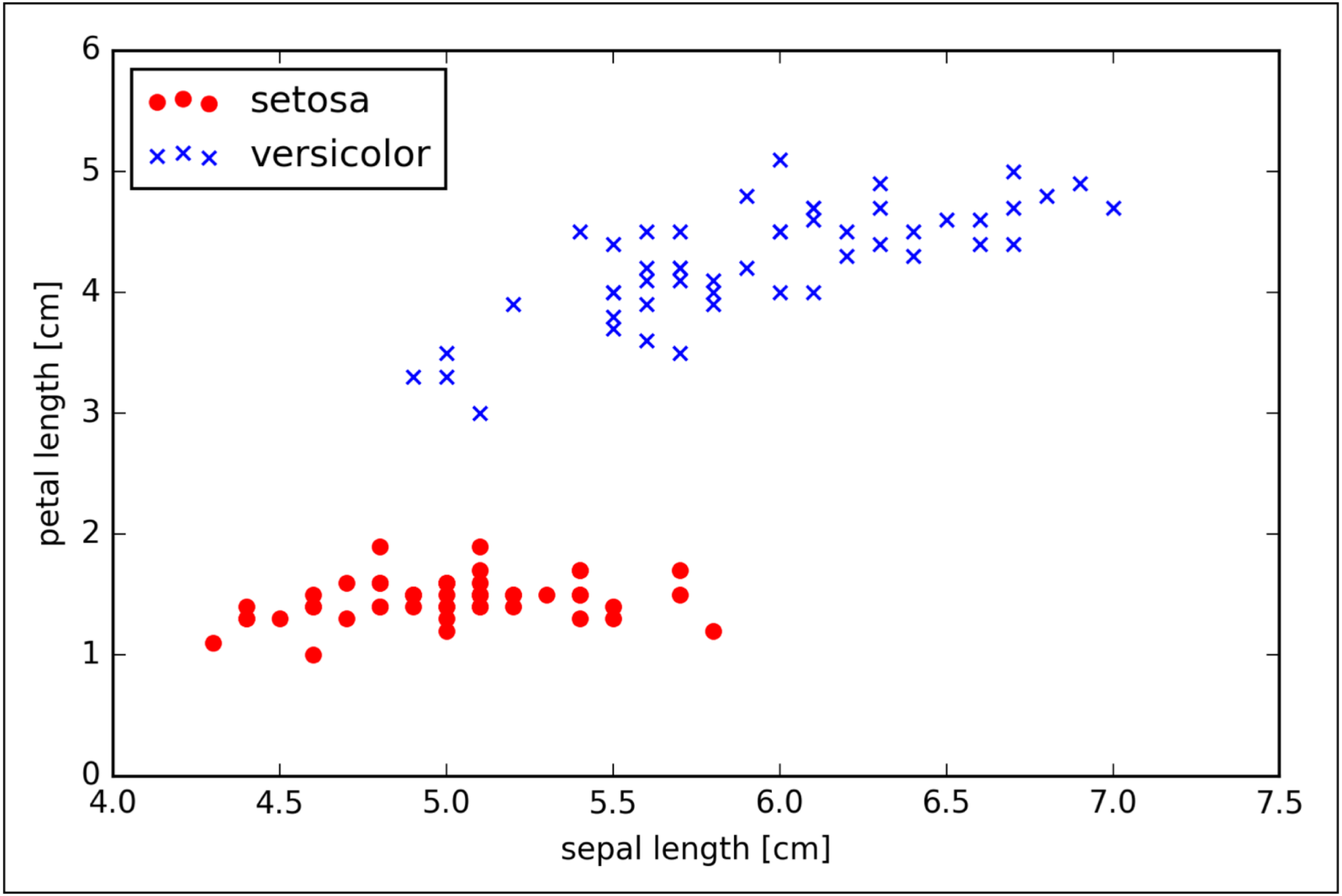
Samples
(instances, observations)

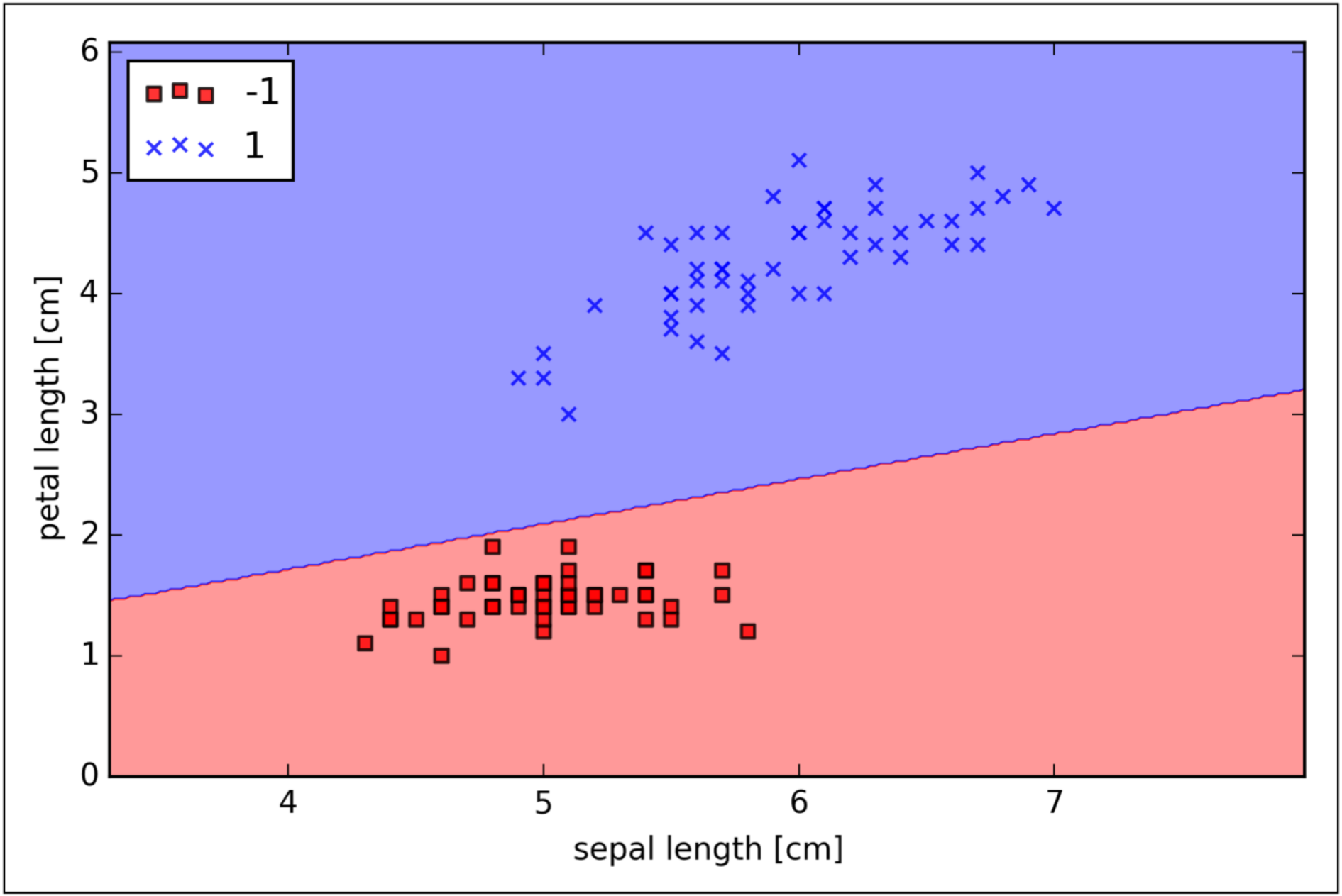
	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

Features
(attributes, measurements, dimensions)

Class labels
(targets)




setosa
versicolor
petal length [cm]
sepal length [cm]
0
1
2
3
4
5
6
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5


-1
1
petal length [cm]
sepal length [cm]
0
1
2
3
4
5
6
4
5
6
7

Let’s build a classifier

Music recommendation:

Will this person like the new Taylor Swift single?

Let's build a classifier

Training data: Does this person like the new Taylor Swift single?

A	B	C	Likes New TSwift
Y	Y	N	Y
Y	N	Y	N
Y	Y	N	Y
Y	N	Y	N
Y	Y	N	Y
N	N	N	N

Let's build a classifier

What are we predicting?

"Will this consumer like the new Taylor Swift single?"

What are the features?

A = does this person have any siblings?

B = did they like Taylor Swift's previous album?

C = do they like Kanye West?

Let's build a classifier

Has Siblings	Previous Purchase	Likes Kanye	Likes New TSwift
Y	Y	N	Y
Y	N	Y	N
Y	Y	N	Y
Y	N	Y	N
Y	Y	N	Y
N	N	N	N

Let's build a classifier: takeaway

Lots of rules match the original data

- Most rules won't work on new data
- Need to be able to **generalize**

This is hard to do without knowing what the variables mean

- A machine learning algorithm won't know what they mean, either (unless you tell it)
- Some heuristics: use rules with lots of evidence; use rules that are simple

Supervised Learning

Recipe for supervised machine learning:

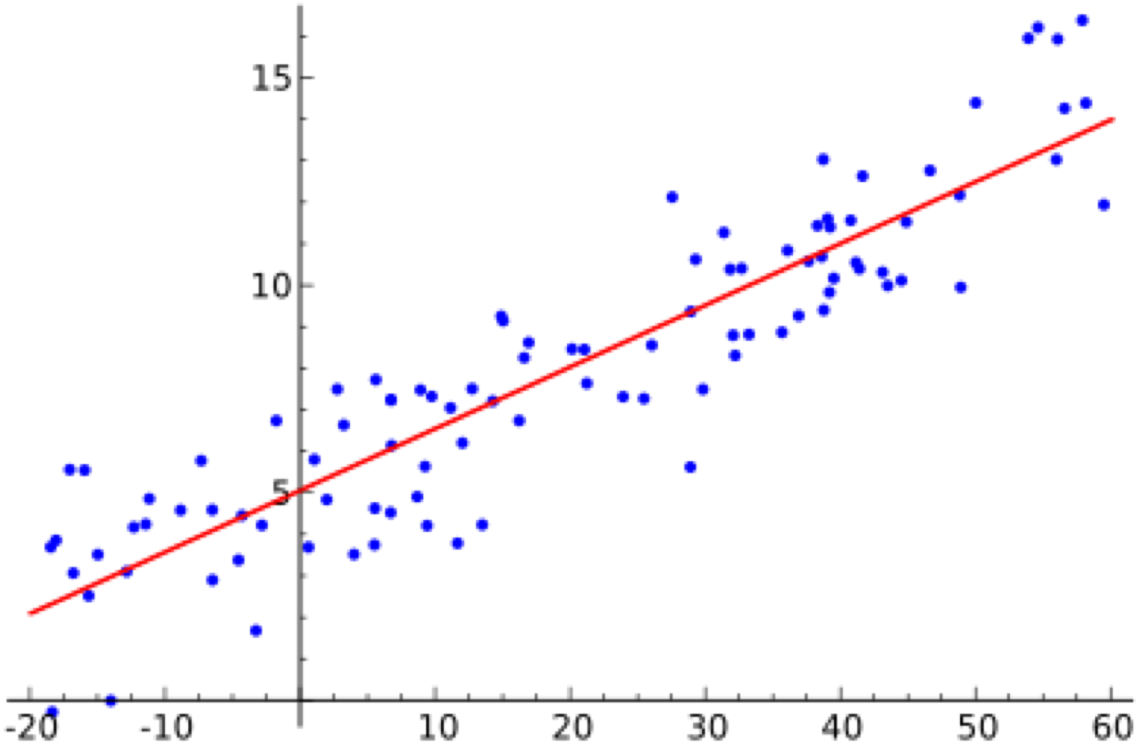
Pattern matching + generalization

Supervised Learning

Two types of prediction:

- Classification
  - Discrete outputs (typically categorical)
- **Regression**
  - **Continuous outputs (usually)**

Regression



Linear regression with one input variable

Regression

Examples:

- Predicting how much money a movie will make
- Forecasting tomorrow's high temperature
- Estimate someone's age based on their face
- Rate how strongly someone likes a product (e.g., in a tweet)

Types of Learning

- Supervised learning
  - Goal: Prediction
- **Unsupervised learning**
  - **Goal: Discovery**

Unsupervised Learning

Finding “interesting” patterns in data

- Not trying to predict any particular variable
- No training data
- Maybe you don’t even know what you’re looking for

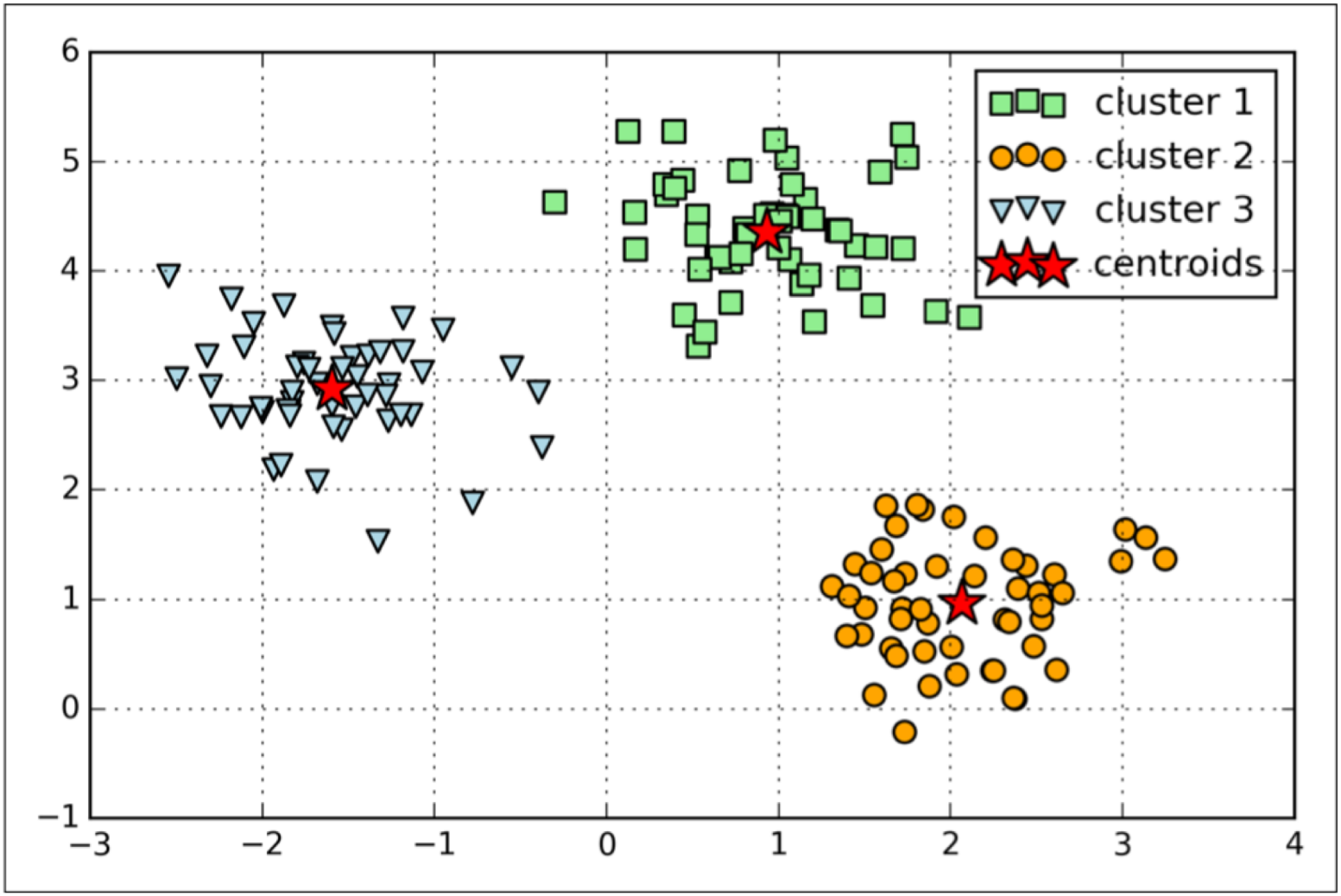
Example: **anomaly detection**

- Trying to identify something unusual (e.g., fraud) but you don’t know what it looks like

Unsupervised Learning

Clustering is an unsupervised learning task that involves grouping data instances into categories

- Similar to classification, but you don’t know what the classes are ahead of time


cluster 1
cluster 2
cluster 3
centroids
6
5
4
3
2
1
0
−1
−3
−2
−1
0
1
2
3
4

Unsupervised Learning

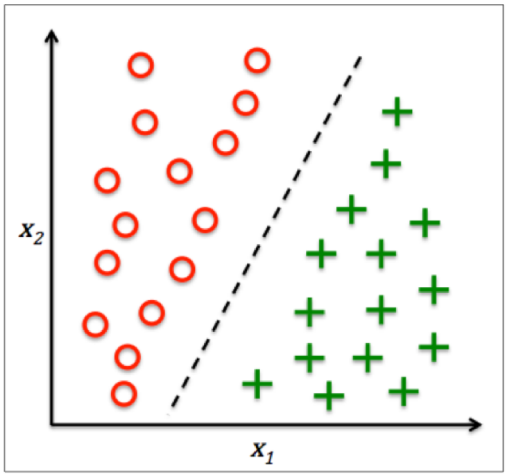
Example: movie recommendation

- Clustering can be used to put people into different groups based on the kinds of movies they like.

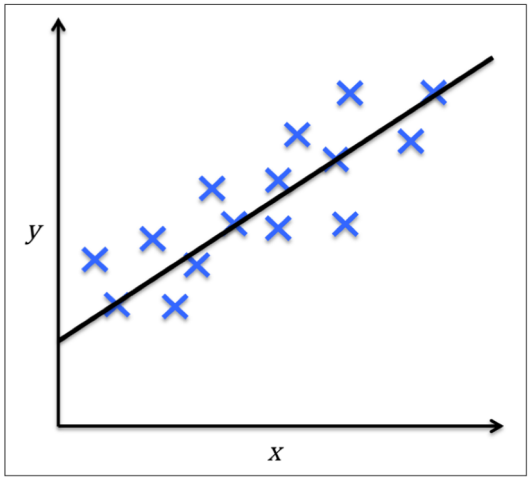
Interest Group 3:	**Interest Group 18:**	**Interest Group 8:**
Trainspotting	Mary Poppins	Pretty Woman
Fargo	Cinderella	Mrs. Doubtfire
Pulp Fiction	The Sound of Music	Ghost
Clerks	Dumbo	Sleepless in Seattle

From Hoffman (2004) “Latent Semantic Models for Collaborative Filtering.”

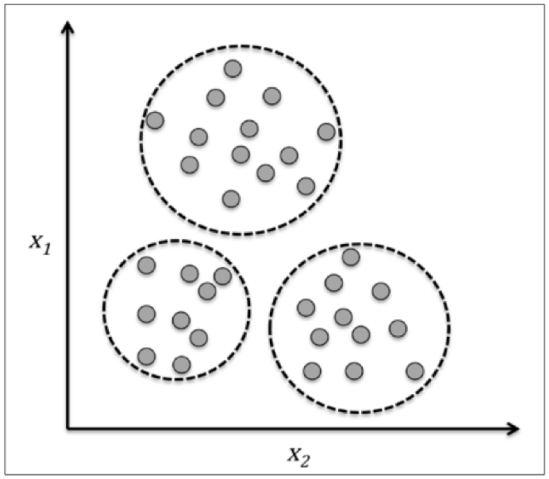
Classification



Regression



Clustering



Semi-supervised Learning

Combines both types of learning

Really just a special case of supervised learning

- You have a specific prediction task, but some of your data has unknown outputs

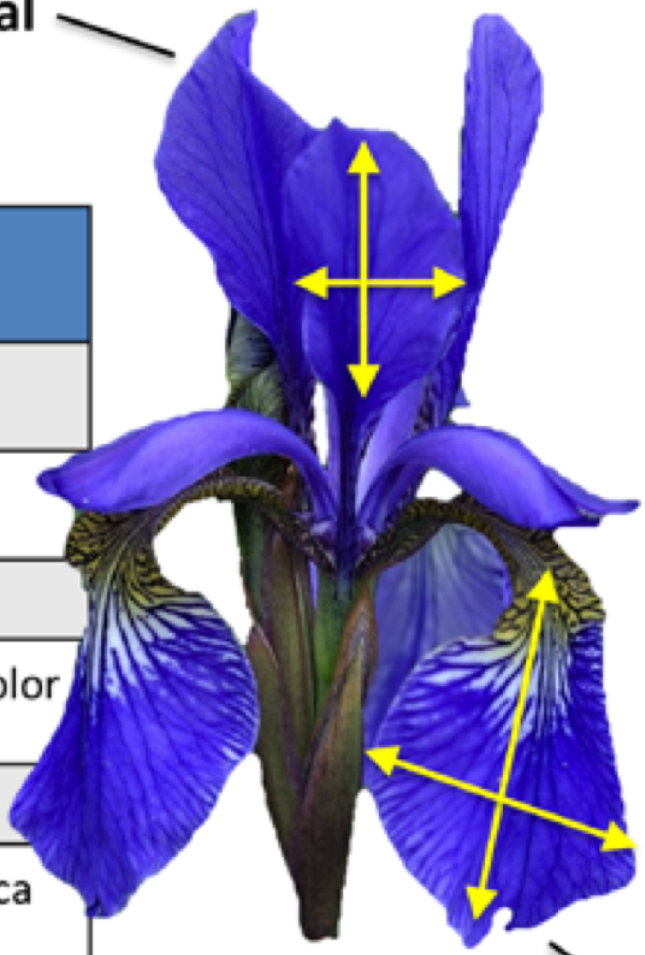
Pause

Samples
(instances, observations)

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

Features
(attributes, measurements, dimensions)

Class labels
(targets)



Terminology

Each data point (i.e., each “thing” you are classifying/regressing/clustering) is called an **instance**

- Alternative name: **observation**
- Also called **examples** or **samples** when used as training data in supervised learning

In a data set, each row corresponds to an instance.

Terminology

The “input” variables are called **features**

- Alternative names: **attributes**, **covariates**
- Also referred to as the **independent** variables

In a data set, each column corresponds to a feature. (Except for the last column, which is the output.)

The list of feature values for an instance is called the instance’s **feature vector**

Terminology

The value of the “output” variable (the “thing” you are trying to predict) is the **label**

- Also called the **dependent** variable

In a data set, this is the final column. (Unless there is more than one label, which is a setting we will consider later in the course.)

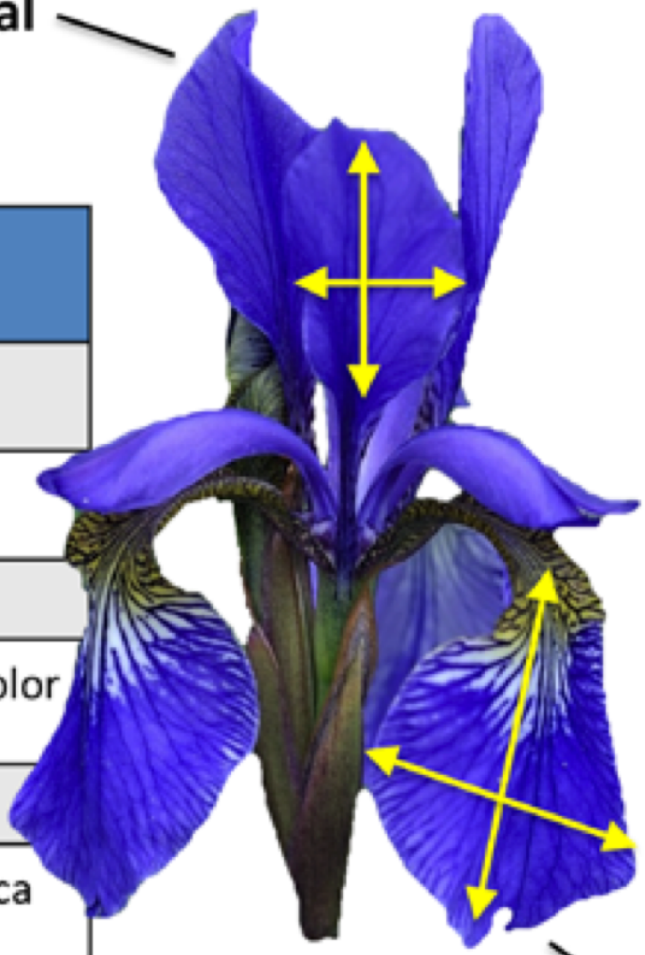
In classification, the possible values the labels can have are called **classes**

Samples
(instances, observations)

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

Features
(attributes, measurements, dimensions)

Class labels
(targets)



Terminology

In supervised learning:

- a **training instance** (or **training example**) is a feature vector paired with a label
- the **training data** (sometimes **labeled data**) is the table of all training instances

In unsupervised learning, the data set contains feature vectors but no labels (sometimes called **unlabeled data**)

Prediction

A **prediction function** is what you get at the end of learning

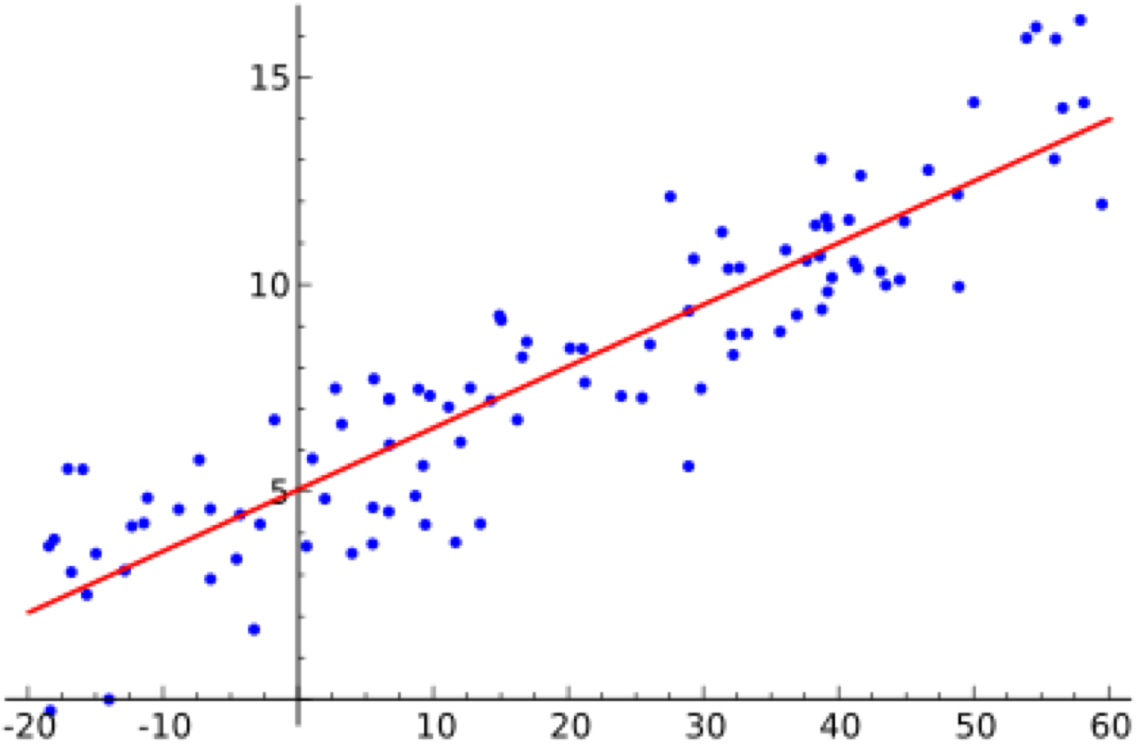
- Sometimes called a **predictor** (but features are also sometimes called *predictor variables*, so this can get confusing)
- Sometimes called a **hypothesis**

A **classifier** is what you call a prediction function if you are doing classification.

Prediction

Example of a simple prediction function:

$y = .17x + 5$



Prediction

Where does this function come from?

Need to learn it so that it is accurate.

What is accurate?

Need to define the **error** or **loss** of a prediction function.

- For classification, this is usually the (negated) probability that the classifier is correct.
- For regression, this is usually measured by how far away the predicted value will be.

Prediction

There is some hypothetical measure of how well a classifier will do on all data it might encounter (the **true error** or **risk**)

But there's probably no way to measure that… usually you can only measure the error or loss on the training data, called the **training error**

- Alternatively: **empirical error/risk**

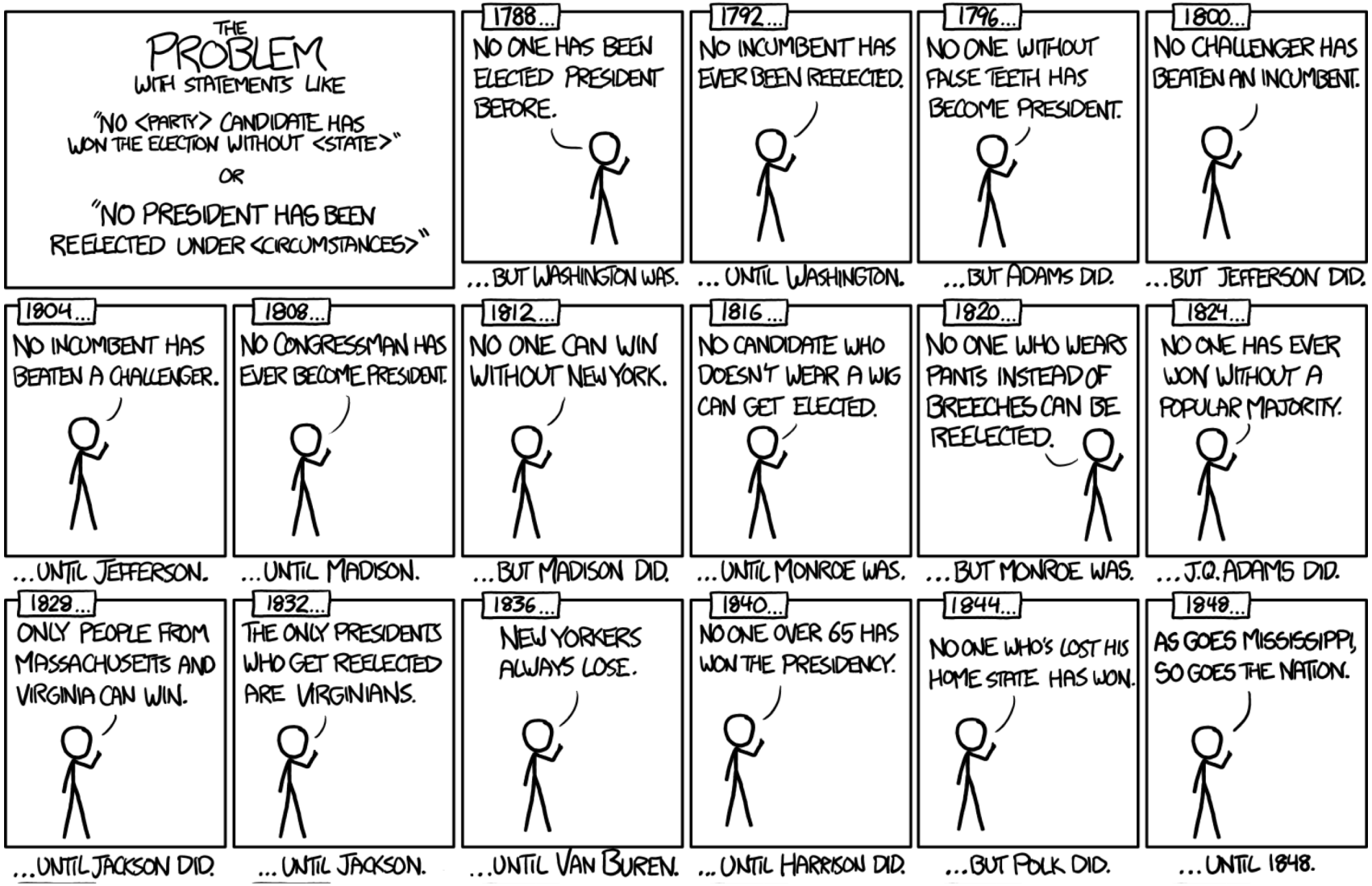
Prediction

Goal of machine learning is to learn a prediction function that minimizes the (true) error.

Since true error is unknown, instead minimize the training error.

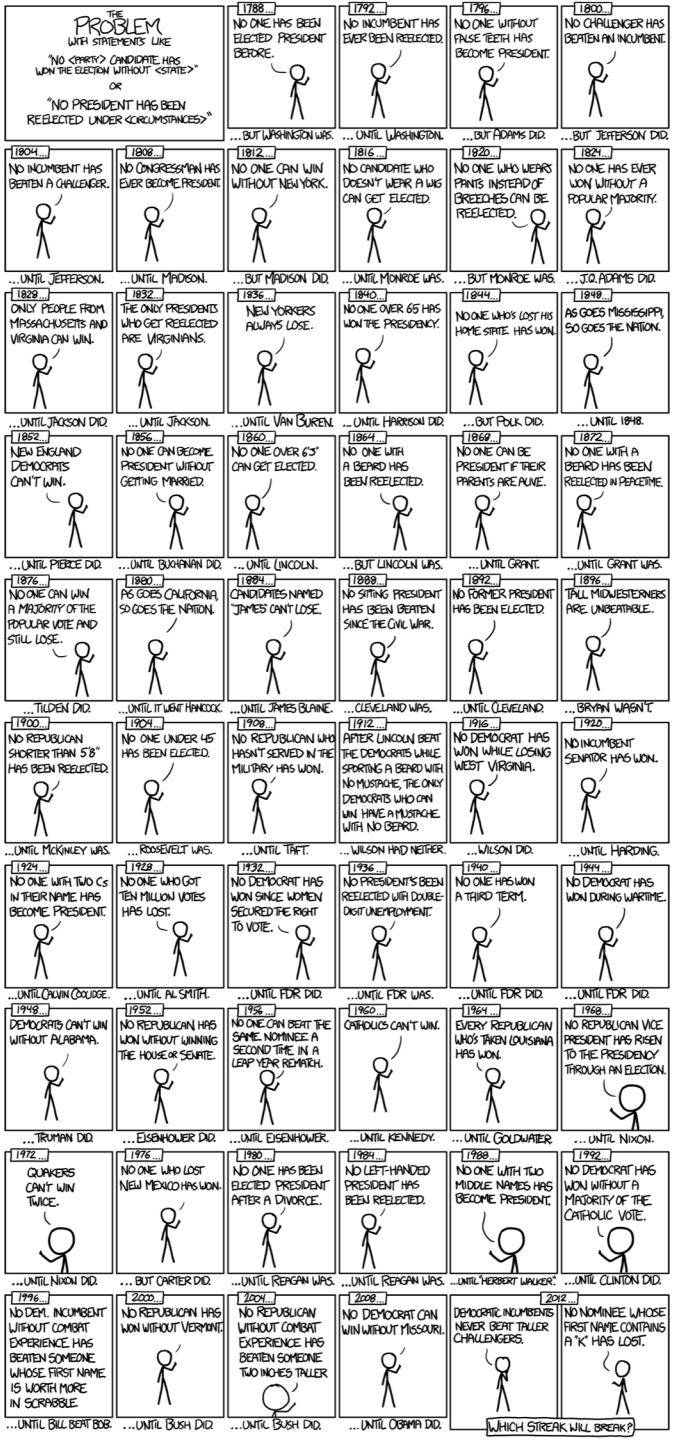
Generalization

Prediction functions that work on the training data might not work on other data



... ...

From: https://xkcd.com/1122/


THE PROBLEM WITH STATEMENTS LIKE
"NO <PARTY> CANDIDATE HAS WON THE ELECTION WITHOUT <STATE>"
OR
"NO PRESIDENT HAS BEEN REELECTED UNDER <CIRCUMSTANCES>"
1788... NO ONE HAS BEEN ELECTED PRESIDENT BEFORE. ...BUT WASHINGTON WAS.
1792... NO INCUMBENT HAS EVER BEEN REELECTED. ...UNTIL WASHINGTON.
1796... NO ONE WITHOUT FALSE TEETH HAS BECOME PRESIDENT. ...BUT ADAMS DID.
1800... NO CHALLENGER HAS BEATEN AN INCUMBENT. ...BUT JEFFERSON DID.
1804... NO INCUMBENT HAS BEATEN A CHALLENGER. ...UNTIL JEFFERSON.
1808... NO CONGRESSMAN HAS EVER BECOME PRESIDENT. ...UNTIL MADISON.
1812... NO ONE CAN WIN WITHOUT NEW YORK. ...BUT MADISON DID.
1816... NO CANDIDATE WHO DOESN'T WEAR A WIG CAN GET ELECTED. ...UNTIL MONROE WAS.
1820... NO ONE WHO WEARS PANTS INSTEAD OF BREECHES CAN BE REELECTED. ...BUT MONROE WAS.
1824... NO ONE HAS EVER WON WITHOUT A POPULAR MAJORITY. ...J.Q. ADAMS DID.
1828... ONLY PEOPLE FROM MASSACHUSETTS AND VIRGINIA CAN WIN. ...UNTIL JACKSON DID.
1832... THE ONLY PRESIDENTS WHO GET REELECTED ARE VIRGINIANS. ...UNTIL JACKSON.
1836... NEW YORKERS ALWAYS LOSE. ...UNTIL VAN BUREN.
1840... NO ONE OVER 65 HAS WON THE PRESIDENCY. ...UNTIL HARRISON DID.
1844... NO ONE WHO'S LOST HIS HOME STATE HAS WON. ...BUT POLK DID.
1848... AS GOES MISSISSIPPI, SO GOES THE NATION. ...UNTIL 1848.
1852... NEW ENGLAND DEMOCRATS CAN'T WIN. ...UNTIL PIERCE DID.
1856... NO ONE CAN BECOME PRESIDENT WITHOUT GETTING MARRIED. ...UNTIL BUCHANAN DID.
1860... NO ONE OVER 6'3" CAN GET ELECTED. ...UNTIL LINCOLN.
1864... NO ONE WITH A BEARD HAS BEEN REELECTED. ...BUT LINCOLN WAS.
1868... NO ONE CAN BE PRESIDENT IF THEIR PARENTS ARE ALIVE. ...UNTIL GRANT.
1872... NO ONE WITH A BEARD HAS BEEN REELECTED IN PEACETIME. ...UNTIL GRANT WAS.
1876... NO ONE CAN WIN A MAJORITY OF THE POPULAR VOTE AND STILL LOSE... ...TILDEN DID.
1880... AS GOES CALIFORNIA, SO GOES THE NATION. ...UNTIL IT WENT HANCOCK.
1884... CANDIDATES NAMED "JAMES" CAN'T LOSE. ...UNTIL JAMES BLAINE.
1888... NO SITTING PRESIDENT HAS BEEN BEATEN SINCE THE CIVIL WAR. ...CLEVELAND WAS.
1892... NO FORMER PRESIDENT HAS BEEN ELECTED. ...UNTIL CLEVELAND.
1896... TALL MIDWESTERNERS ARE UNBEATABLE... ...BRYAN WASN'T.
1900... NO REPUBLICAN SHORTER THAN 5'8" HAS BEEN REELECTED. ...UNTIL MCKINLEY WAS.
1904... NO ONE UNDER 45 HAS BEEN ELECTED. ...ROOSEVELT WAS.
1908... NO REPUBLICAN WHO HASN'T SERVED IN THE MILITARY HAS WON. ...UNTIL TAFT.
1912... AFTER LINCOLN, BEAT THE DEMOCRATS WHILE SPORTING A BEARD WITH NO MUSTACHE. THE ONLY DEMOCRATS WHO CAN WIN HAVE A MUSTACHE WITH NO BEARD. ...WILSON HAD NEITHER.
1916... NO DEMOCRAT HAS WON WHILE LOSING WEST VIRGINIA. ...WILSON DID.
1920... NO INCUMBENT SENATOR HAS WON. ...UNTIL HARDING.
1924... NO ONE WITH TWO CS IN THEIR NAME HAS BECOME PRESIDENT. ...UNTIL CALVIN COOLIDGE.
1928... NO ONE WHO GOT TEN MILLION VOTES HAS LOST. ...UNTIL AL SMITH.
1932... NO DEMOCRAT HAS WON SINCE WOMEN SECURED THE RIGHT TO VOTE. ...UNTIL FDR DID.
1936... NO PRESIDENT'S BEEN REELECTED WITH DOUBLE-DIGIT UNEMPLOYMENT. ...UNTIL FDR WAS.
1940... NO ONE HAS WON A THIRD TERM. ...UNTIL FDR DID.
1944... NO DEMOCRAT HAS WON DURING WARTIME. ...UNTIL FDR DID.
1948... DEMOCRATS CAN'T WIN WITHOUT ALABAMA. ...TRUMAN DID.
1952... NO REPUBLICAN HAS WON WITHOUT WINNING THE HOUSE OR SENATE. ...EISENHOWER DID.
1956... NO ONE CAN BEAT THE SAME NOMINEE A SECOND TIME IN A LEAP YEAR REMATCH. ...UNTIL EISENHOWER.
1960... CATHOLICS CAN'T WIN. ...UNTIL KENNEDY.
1964... EVERY REPUBLICAN WHO'S TAKEN LOUISIANA HAS WON. ...UNTIL GOLDWATER.
1968... NO REPUBLICAN VICE PRESIDENT HAS RISEN TO THE PRESIDENCY THROUGH AN ELECTION. ...UNTIL NIXON.
1972... QUAKERS CAN'T WIN TWICE. ...UNTIL NIXON DID.
1976... NO ONE WHO LOST NEW MEXICO HAS WON. ...BUT CARTER DID.
1980... NO ONE HAS BEEN ELECTED PRESIDENT AFTER A DIVORCE... ...UNTIL REAGAN WAS.
1984... NO LEFT-HANDED PRESIDENT HAS BEEN REELECTED. ...UNTIL REAGAN WAS.
1988... NO ONE WITH TWO MIDDLE NAMES HAS BECOME PRESIDENT. ...UNTIL "HERBERT WALKER".
1992... NO DEMOCRAT HAS WON WITHOUT A MAJORITY OF THE CATHOLIC VOTE. ...UNTIL CLINTON DID.
1996... NO DEM. INCUMBENT WITHOUT COMBAT EXPERIENCE HAS BEATEN SOMEONE WHOSE FIRST NAME IS WORTH MORE IN SCRABBLE. ...UNTIL BILL BEAT BOB.
2000... NO REPUBLICAN HAS WON WITHOUT VERMONT. ...UNTIL BUSH DID.
2004... NO REPUBLICAN WITHOUT COMBAT EXPERIENCE HAS BEATEN SOMEONE TWO INCHES TALLER ...UNTIL BUSH DID.
2008... NO DEMOCRAT CAN WIN WITHOUT MISSOURI. ...UNTIL OBAMA DID.
2012...
DEMOCRATIC INCUMBENTS NEVER BEAT TALLER CHALLENGERS.
NO NOMINEE WHOSE FIRST NAME CONTAINS A "K" HAS LOST.
WHICH STREAK WILL BREAK?

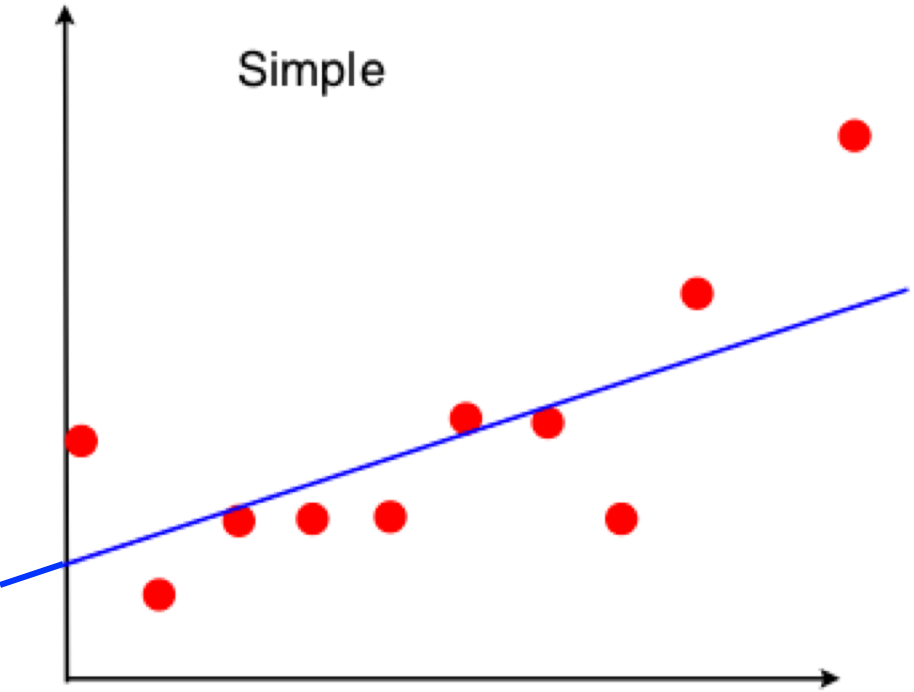
Generalization

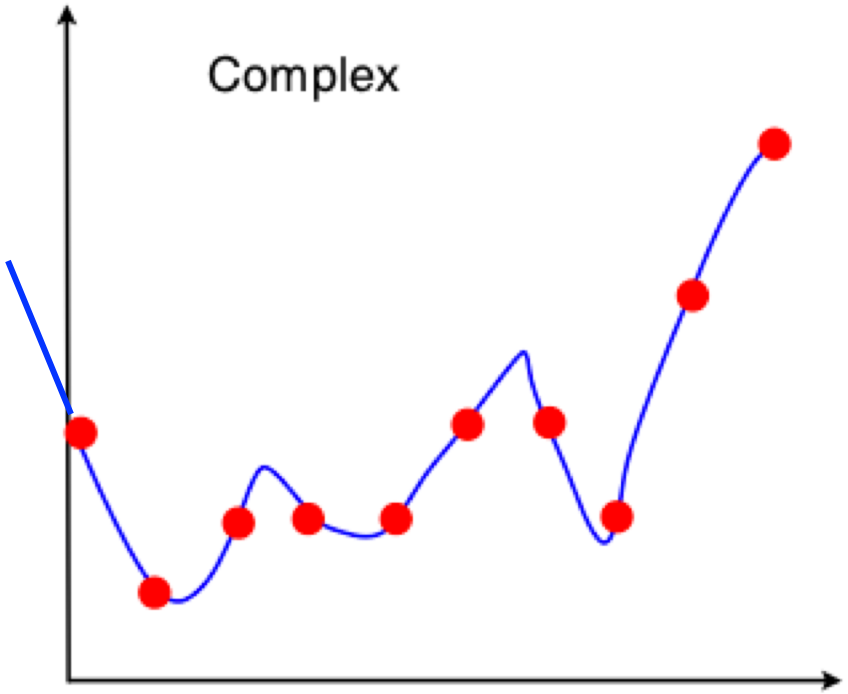
Prediction functions that work on the training data might not work on other data

Minimizing the training error is a reasonable thing to do, but it's possible to minimize it "too well"

- If your function matches the training data well but is not learning general rules that will work for new data, this is called **overfitting**

Generalization





From: https://www.quora.com/Whats-the-difference-between-overfitting-and-underfitting

Generalization

Restrictions on what a classifier can learn is called an **inductive bias**

Inductive biases are an important and necessary ingredient to learning classifiers that will generalize to new data

Generalization

One type of bias: don't use certain features

Has Siblings	Previous Purchase	Likes Kanye	Likes New TSwift
Y	Y	N	Y
Y	N	Y	N
Y	Y	N	Y
Y	N	Y	N
N	Y	N	Y

Generalization

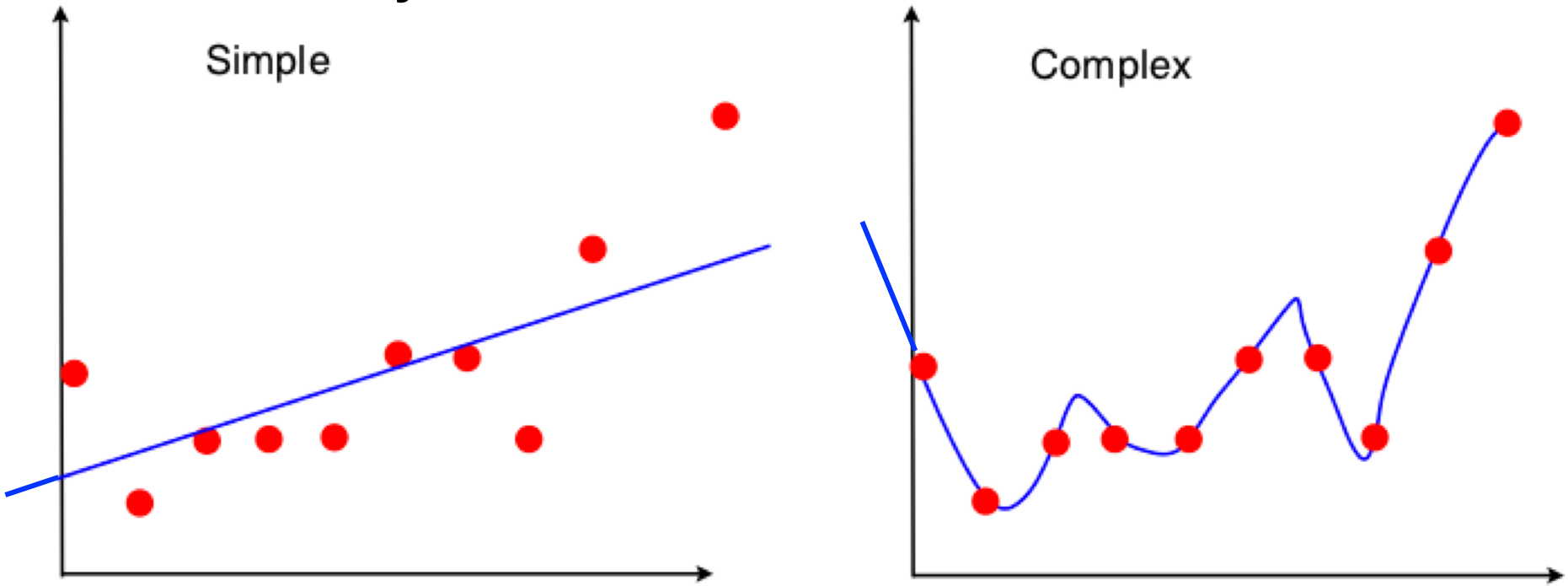
One type of bias: don't use certain features

Has Siblings	Previous Purchase	Likes Kanye	Likes New TSwift
Y	Y	N	Y
Y	N	Y	N
Y	Y	N	Y
Y	N	Y	N
N	Y	N	Y

We suspect that this is probably irrelevant, so don't include it

Generalization

Another type of bias: restrict what kind of function you can learn



Linear functions (lines or planes) are so simple that they won't overfit, even if they aren't perfect on training data

Generalization

We'll discuss other types of inductive bias (some automatic) that can help with generalization throughout the semester

Almost done

Uncertainty

When making a prediction, there is some uncertainty (by definition)

Many machine learning models can estimate the **probability** that an instance has a particular label

Machine Learning in Practice