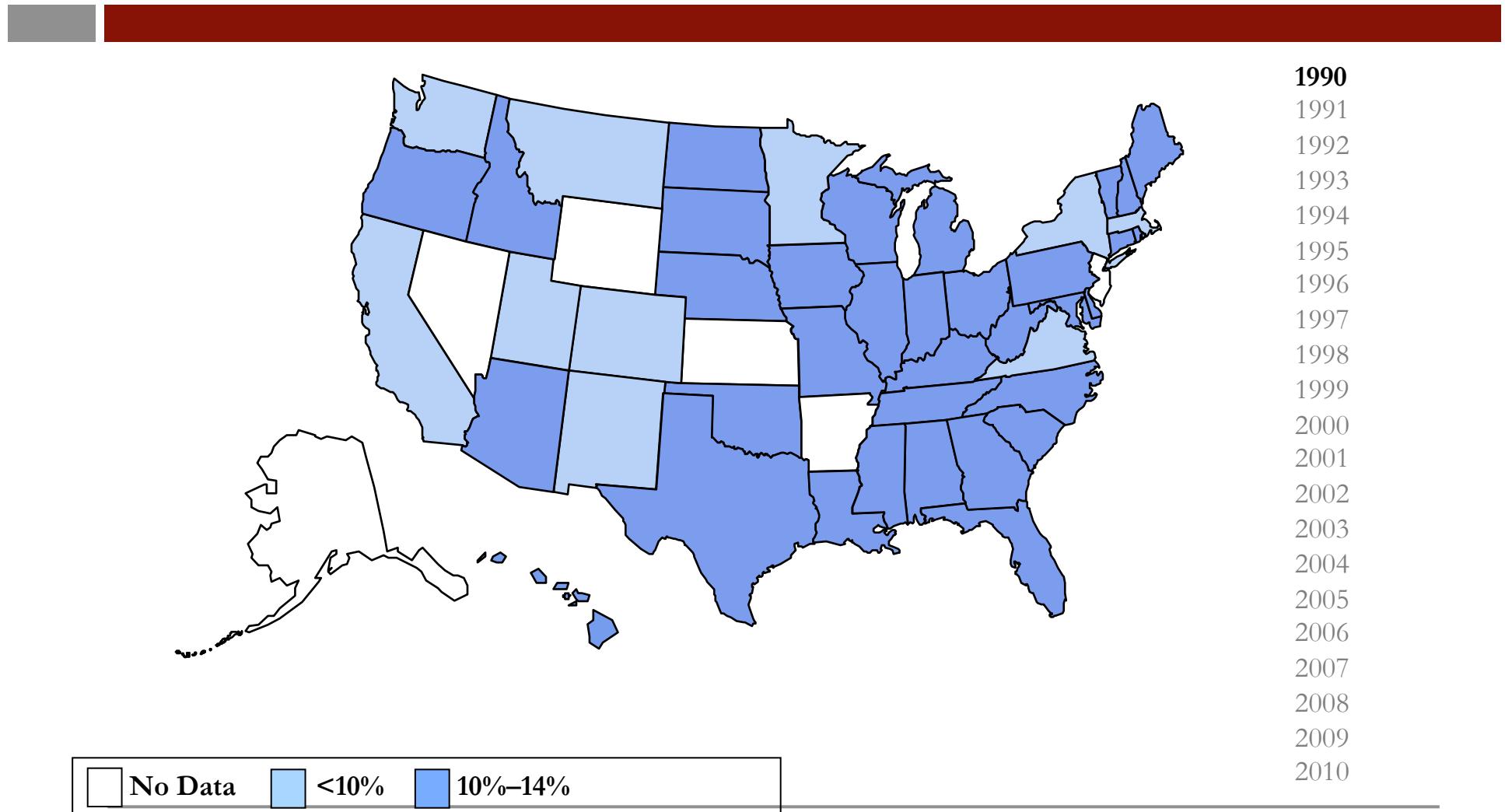


UNDERSTANDING FOOD

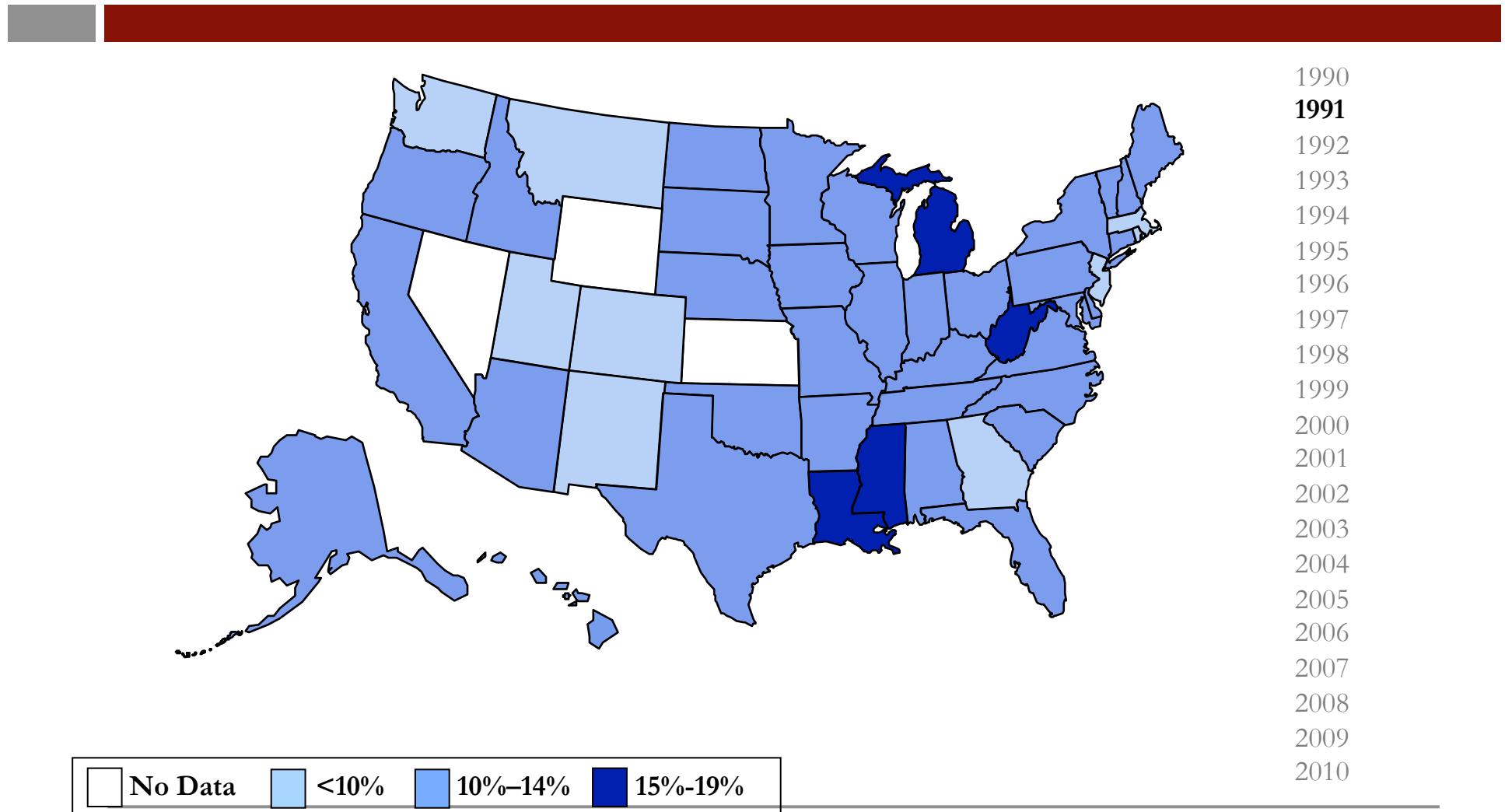
Nutritional Education with Data

15.071x – The Analytics Edge

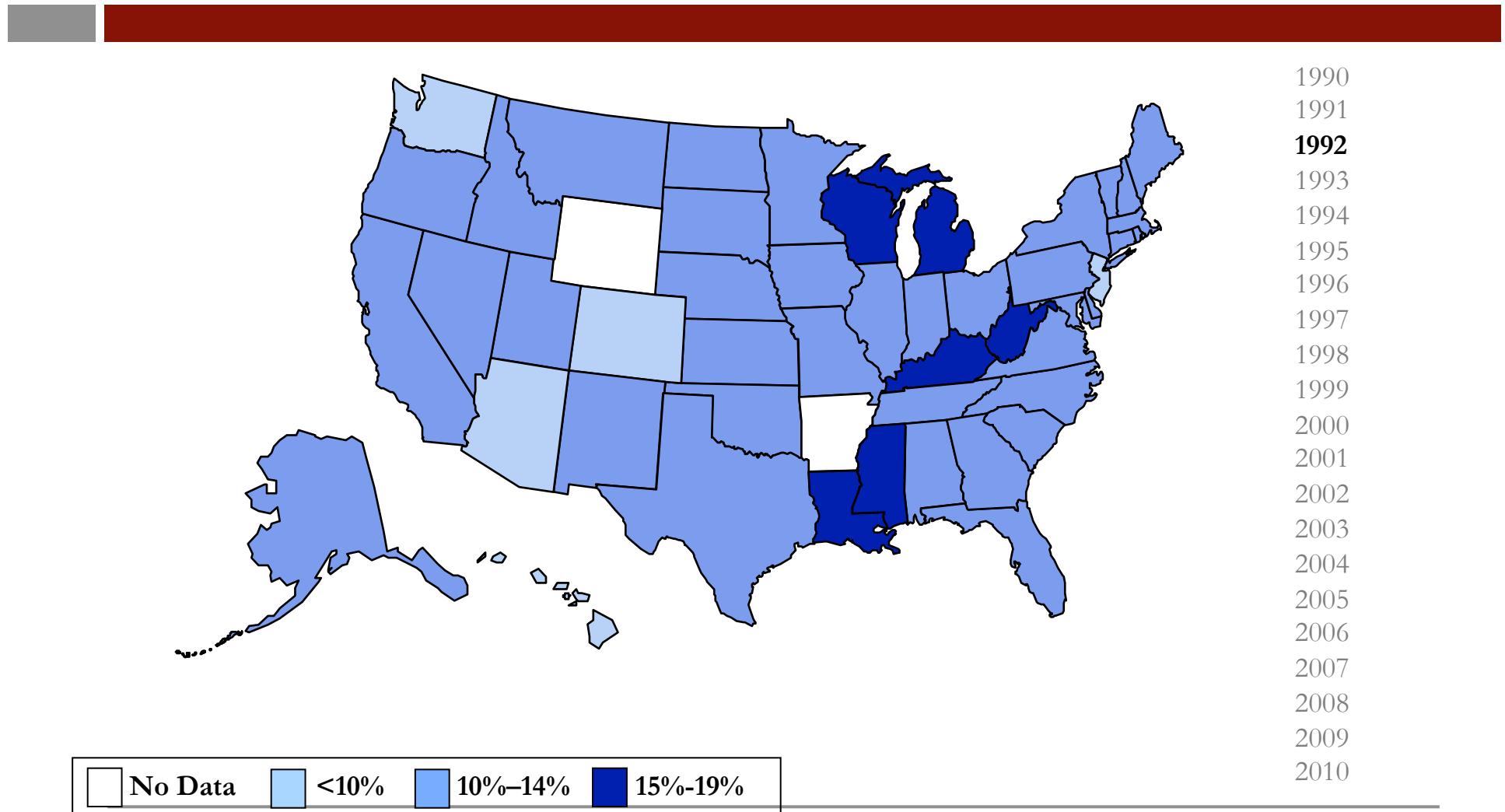
Obesity Trends Among U.S. Adults



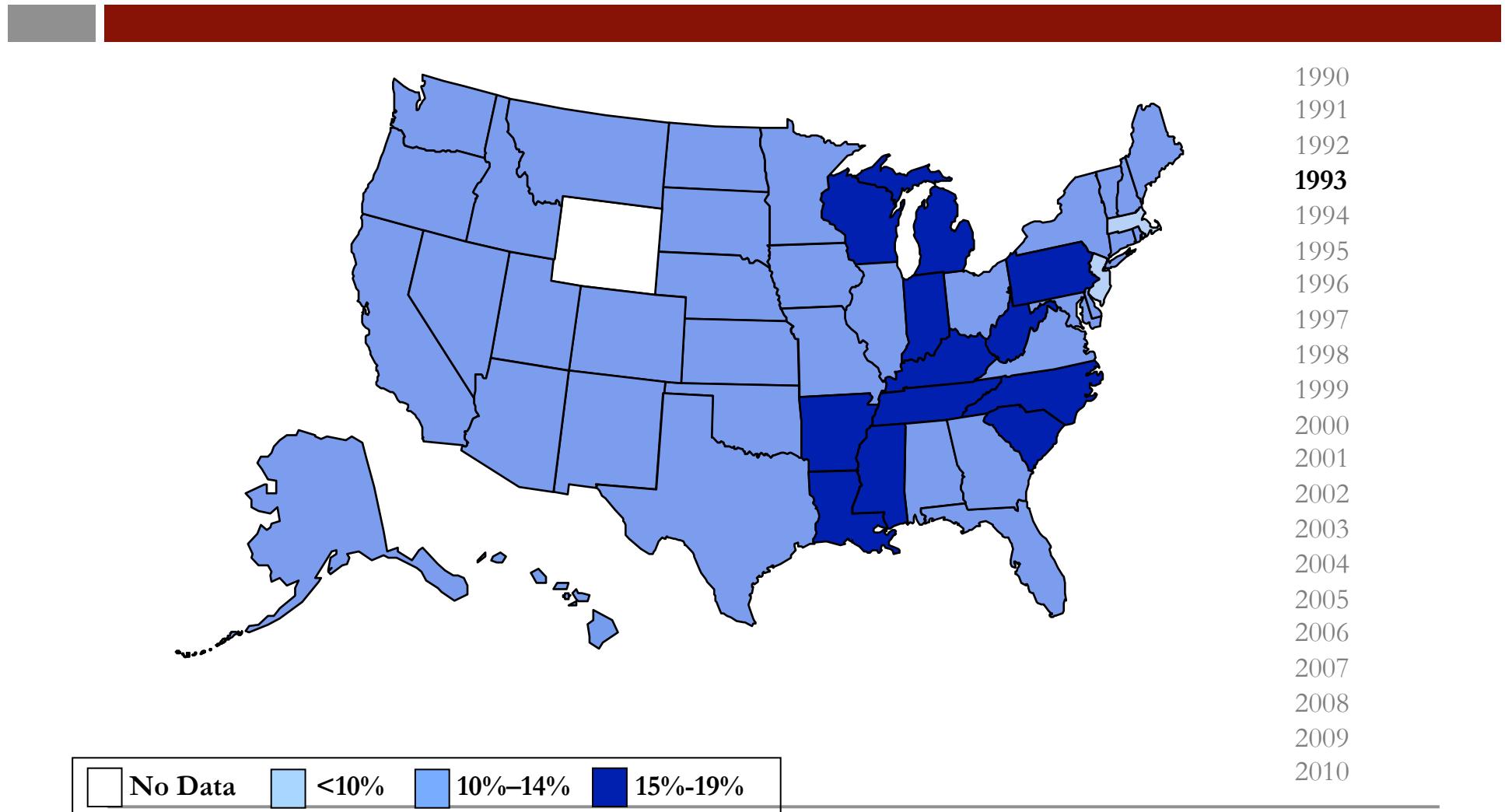
Obesity Trends Among U.S. Adults



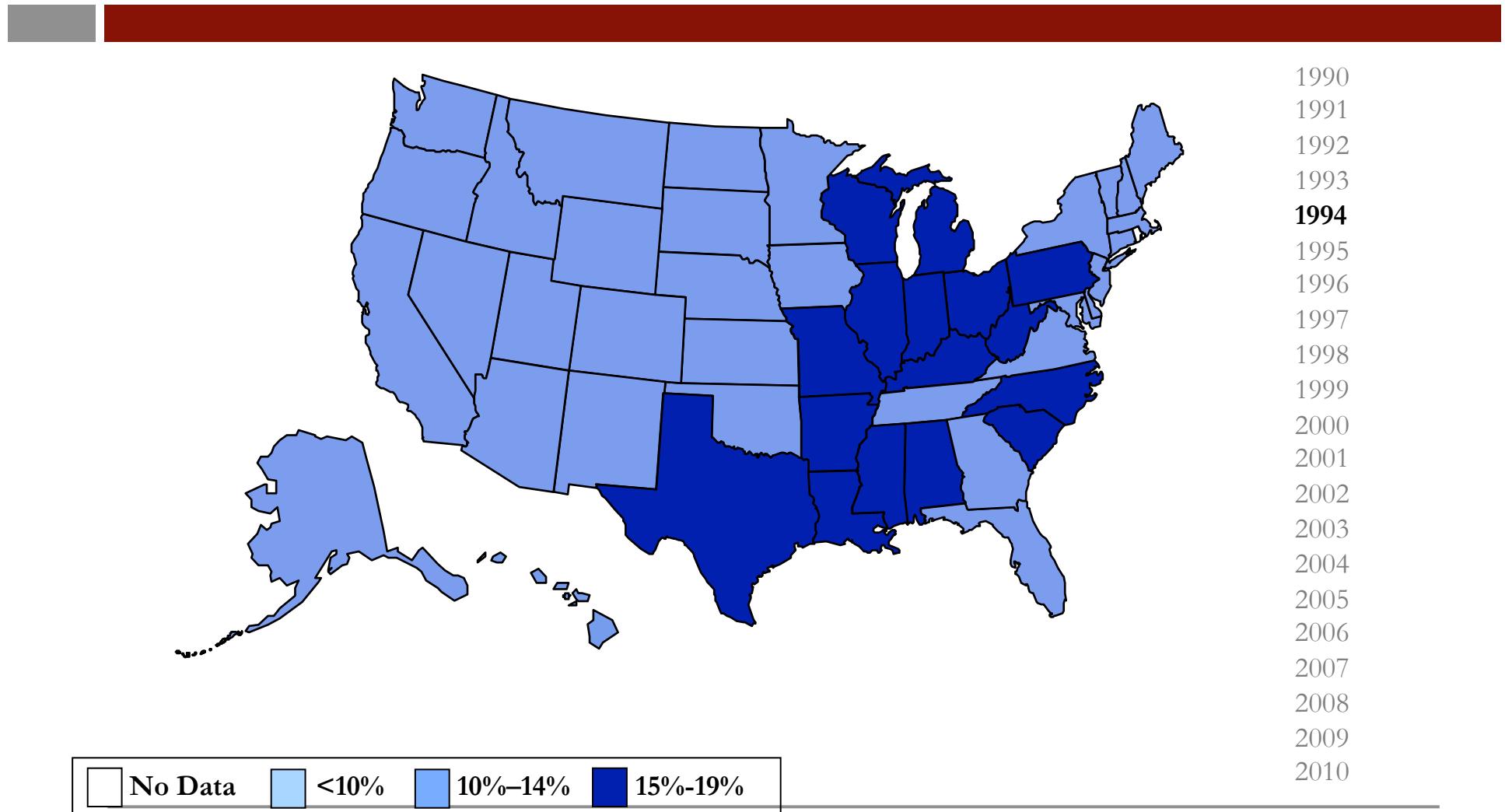
Obesity Trends Among U.S. Adults



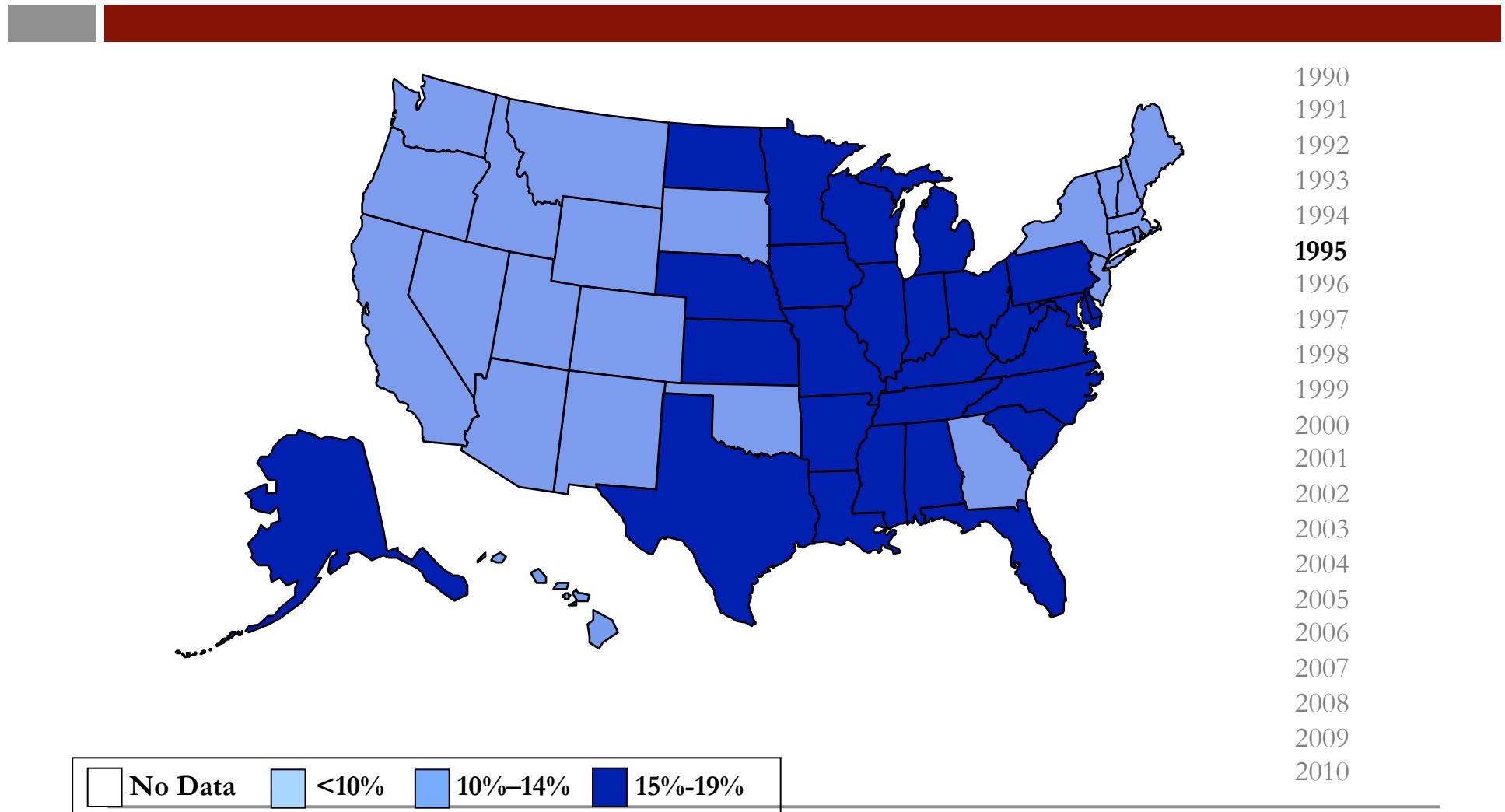
Obesity Trends Among U.S. Adults



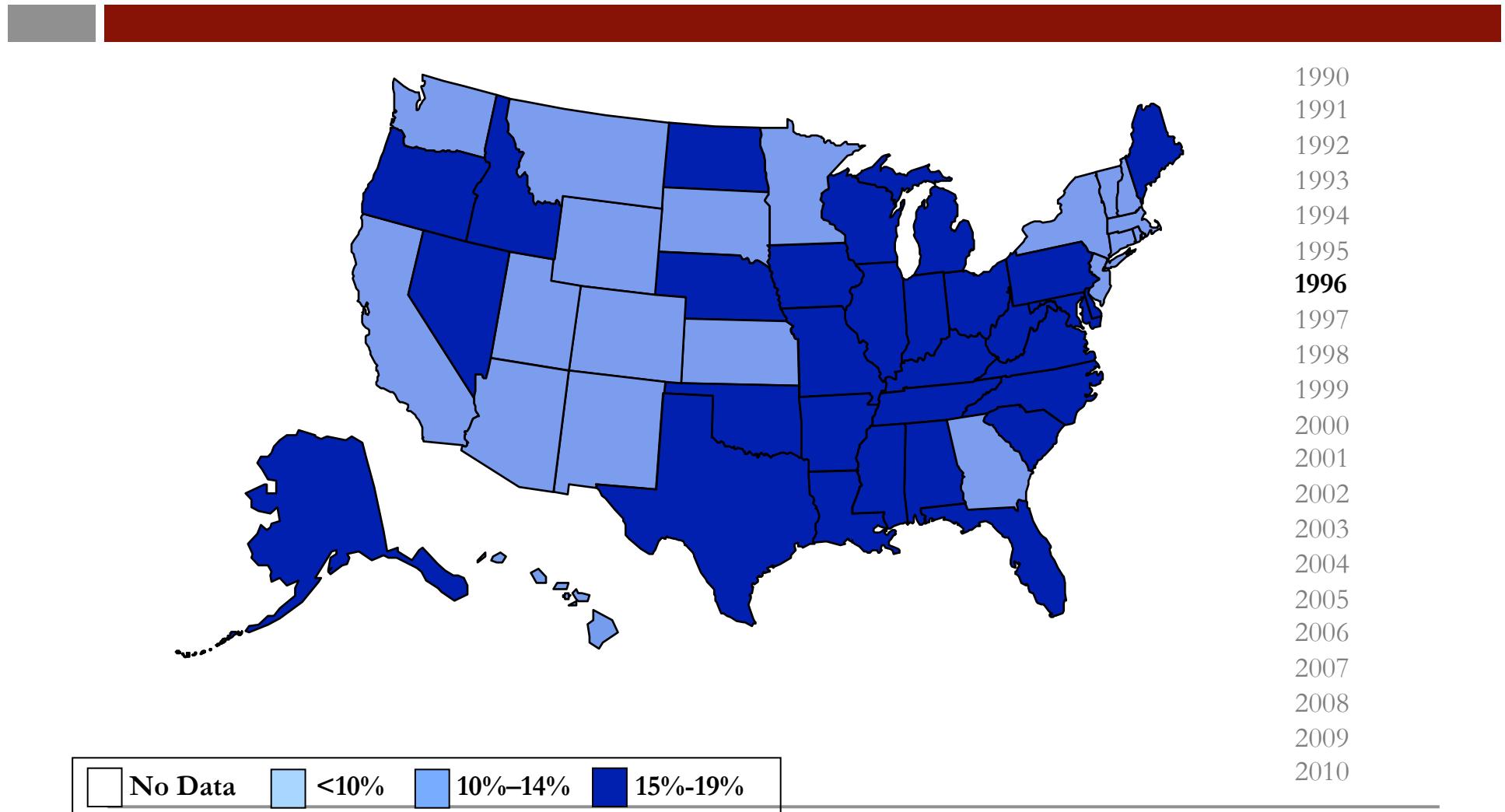
Obesity Trends Among U.S. Adults



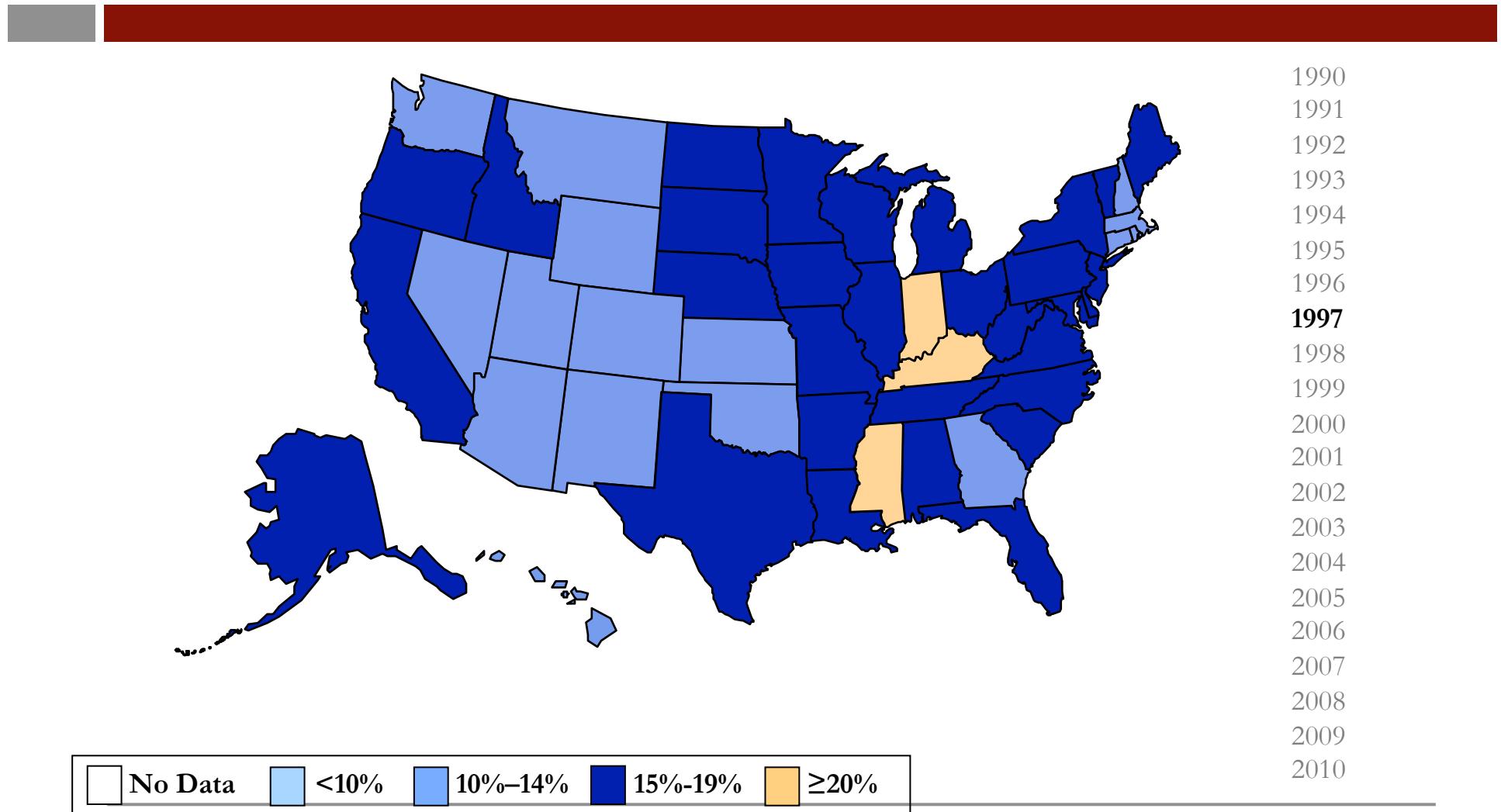
Obesity Trends Among U.S. Adults



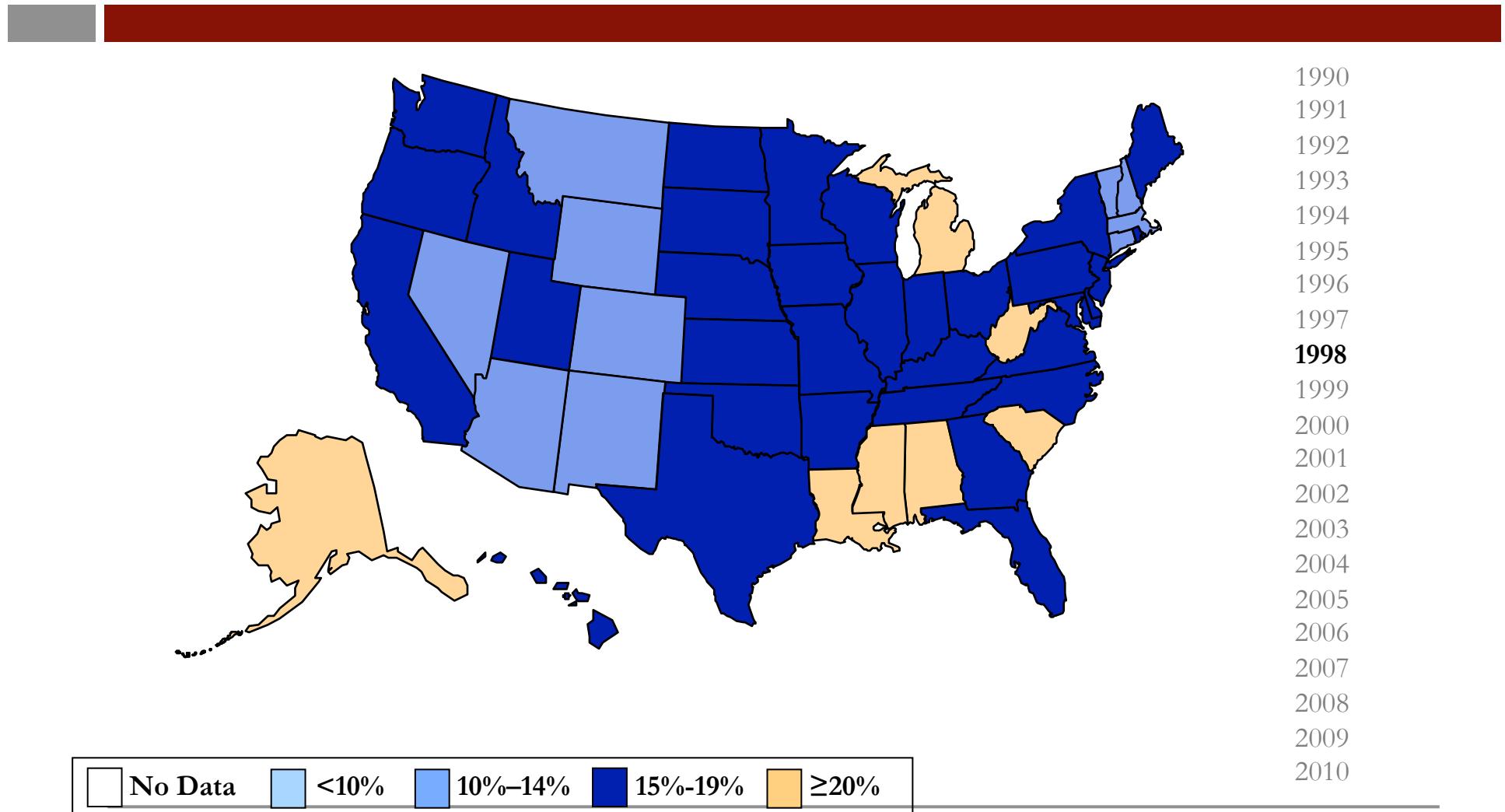
Obesity Trends Among U.S. Adults



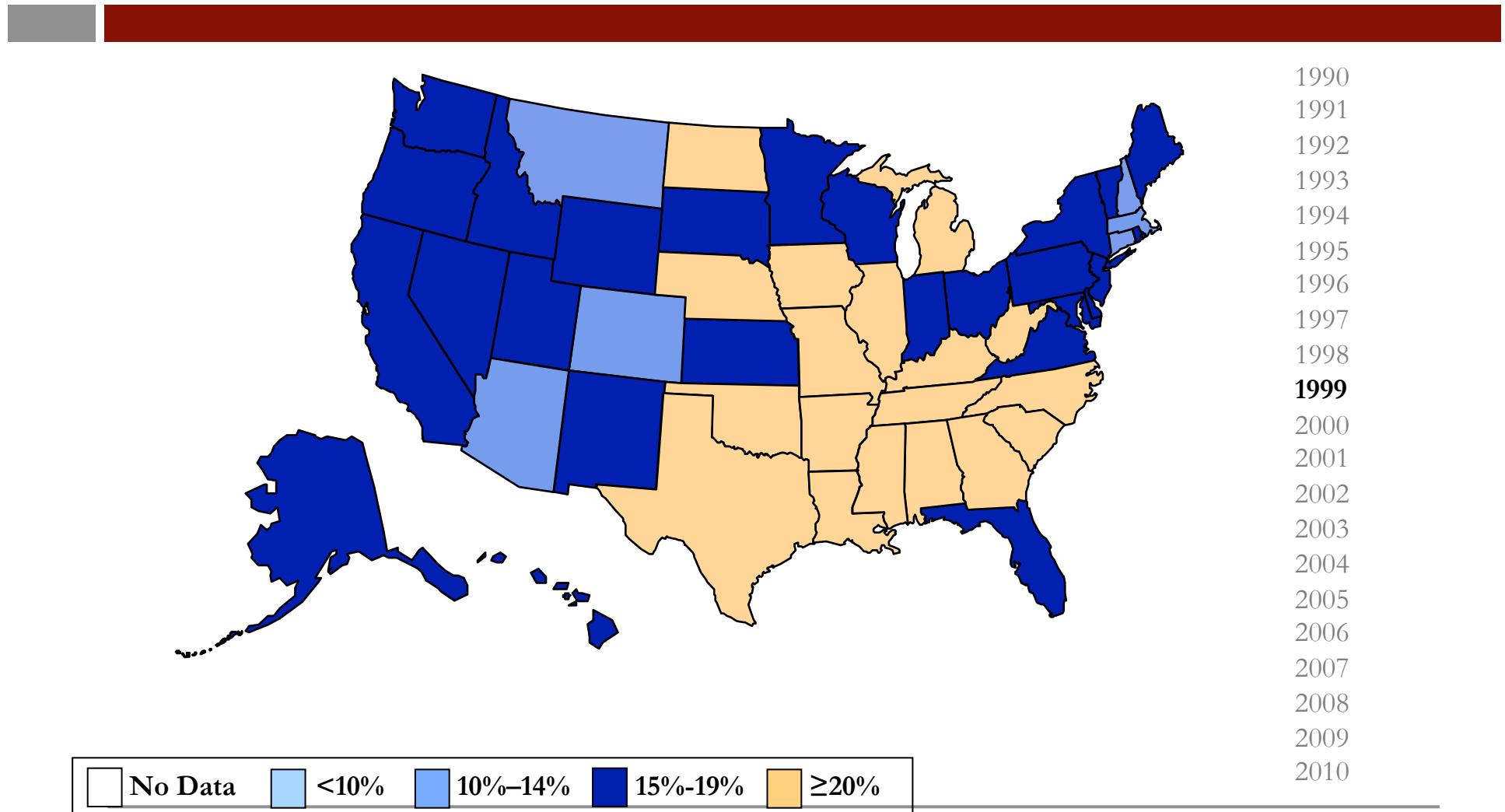
Obesity Trends Among U.S. Adults



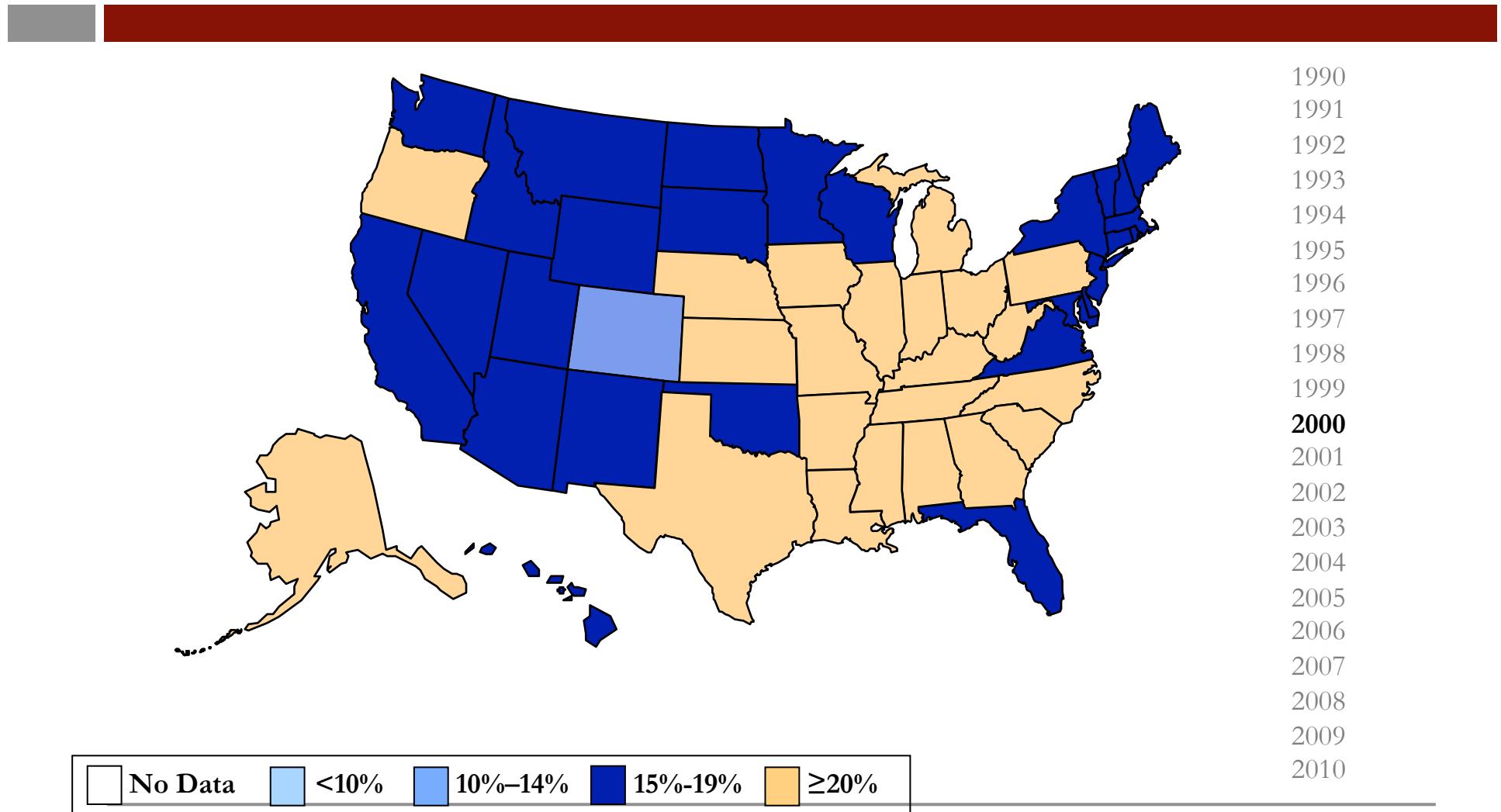
Obesity Trends Among U.S. Adults



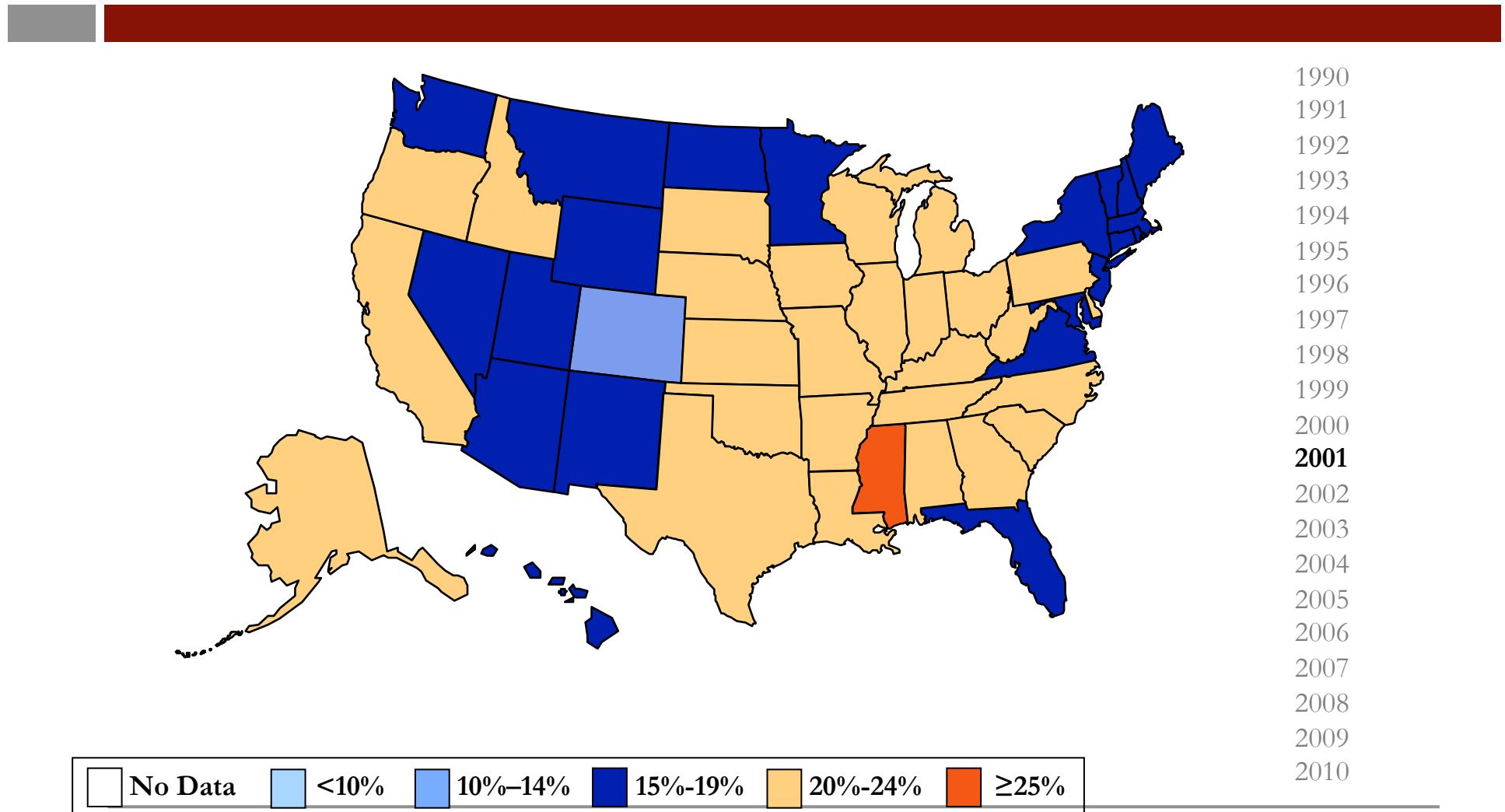
Obesity Trends Among U.S. Adults



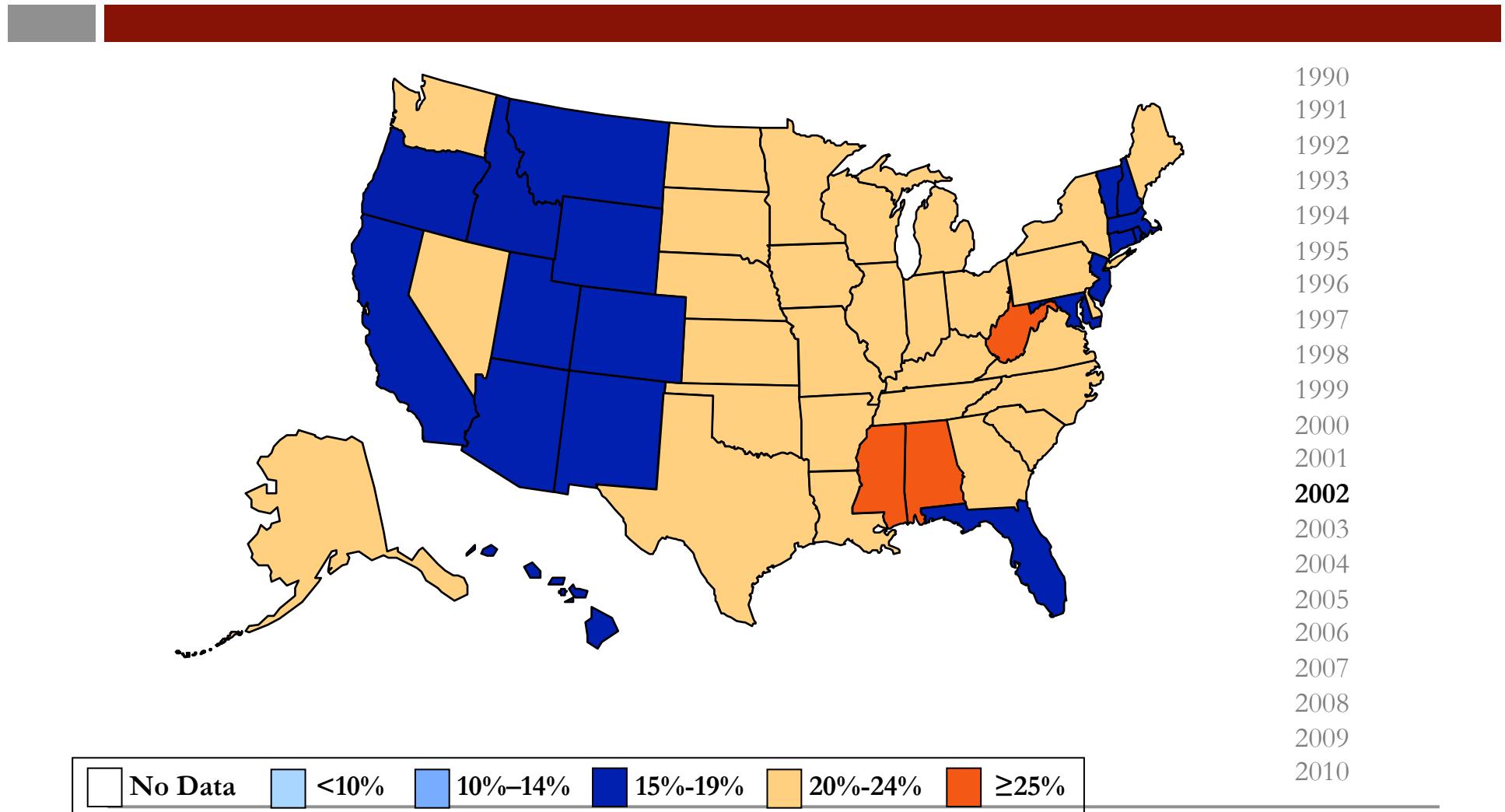
Obesity Trends Among U.S. Adults



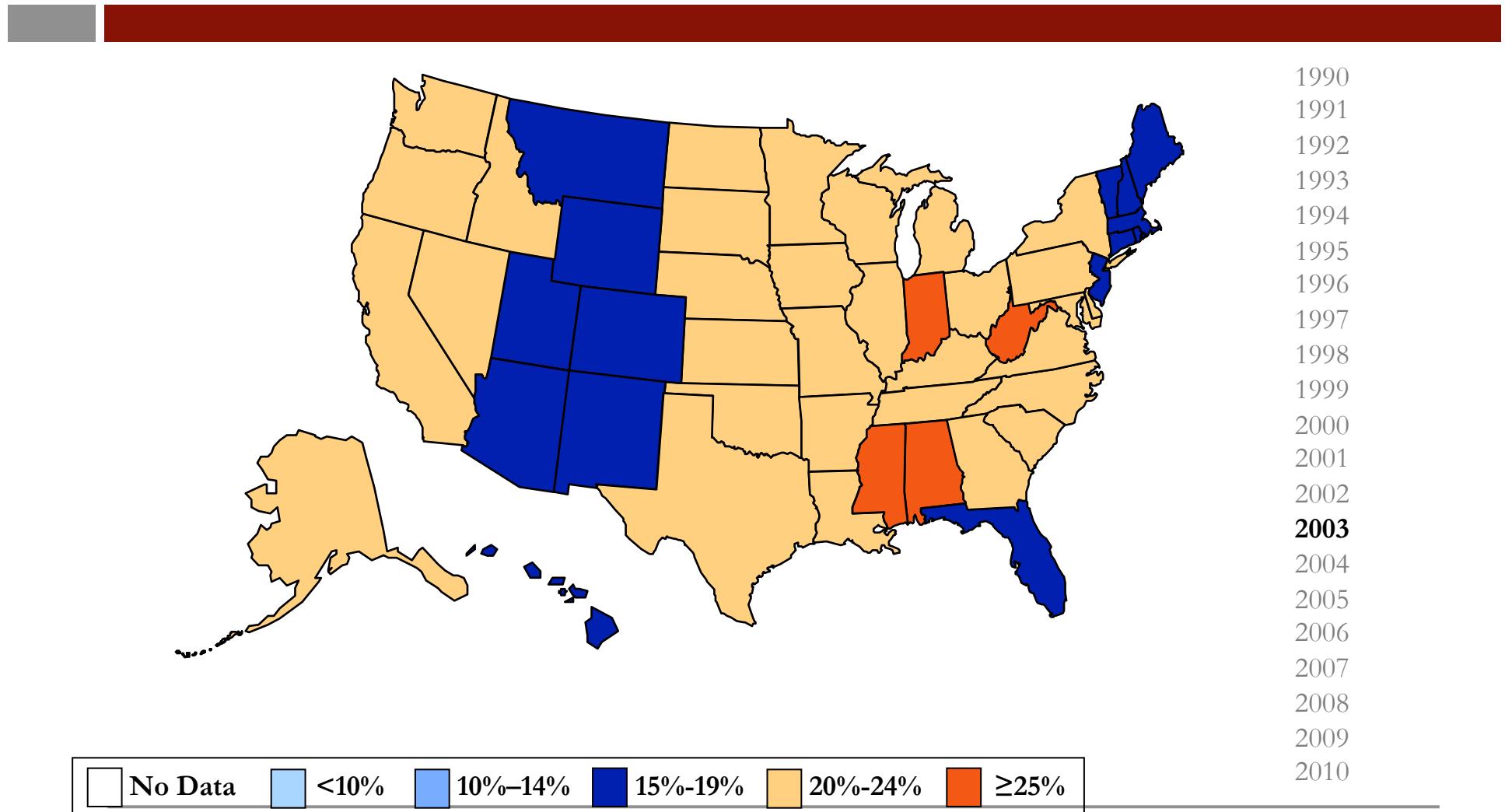
Obesity Trends Among U.S. Adults



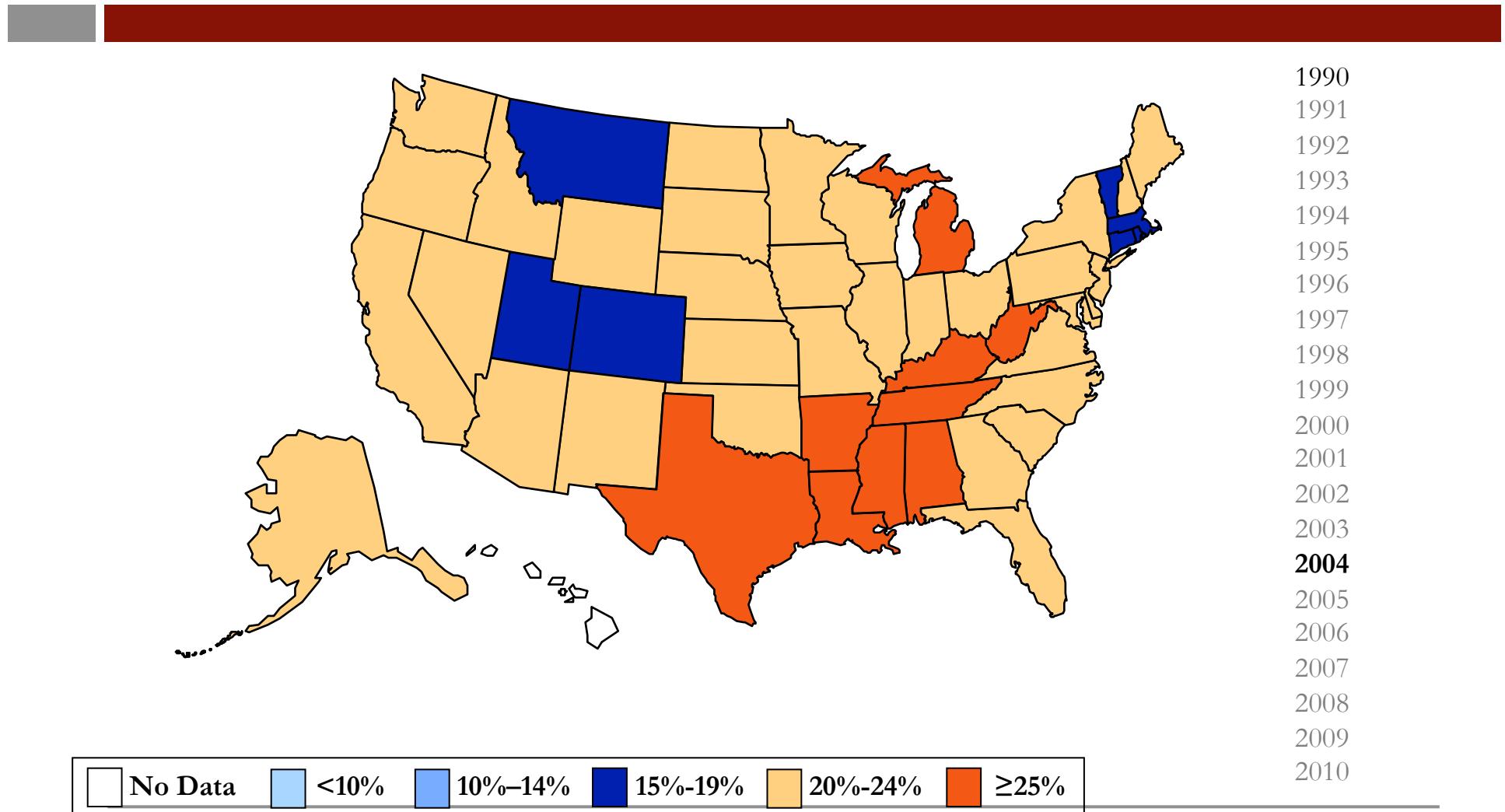
Obesity Trends Among U.S. Adults



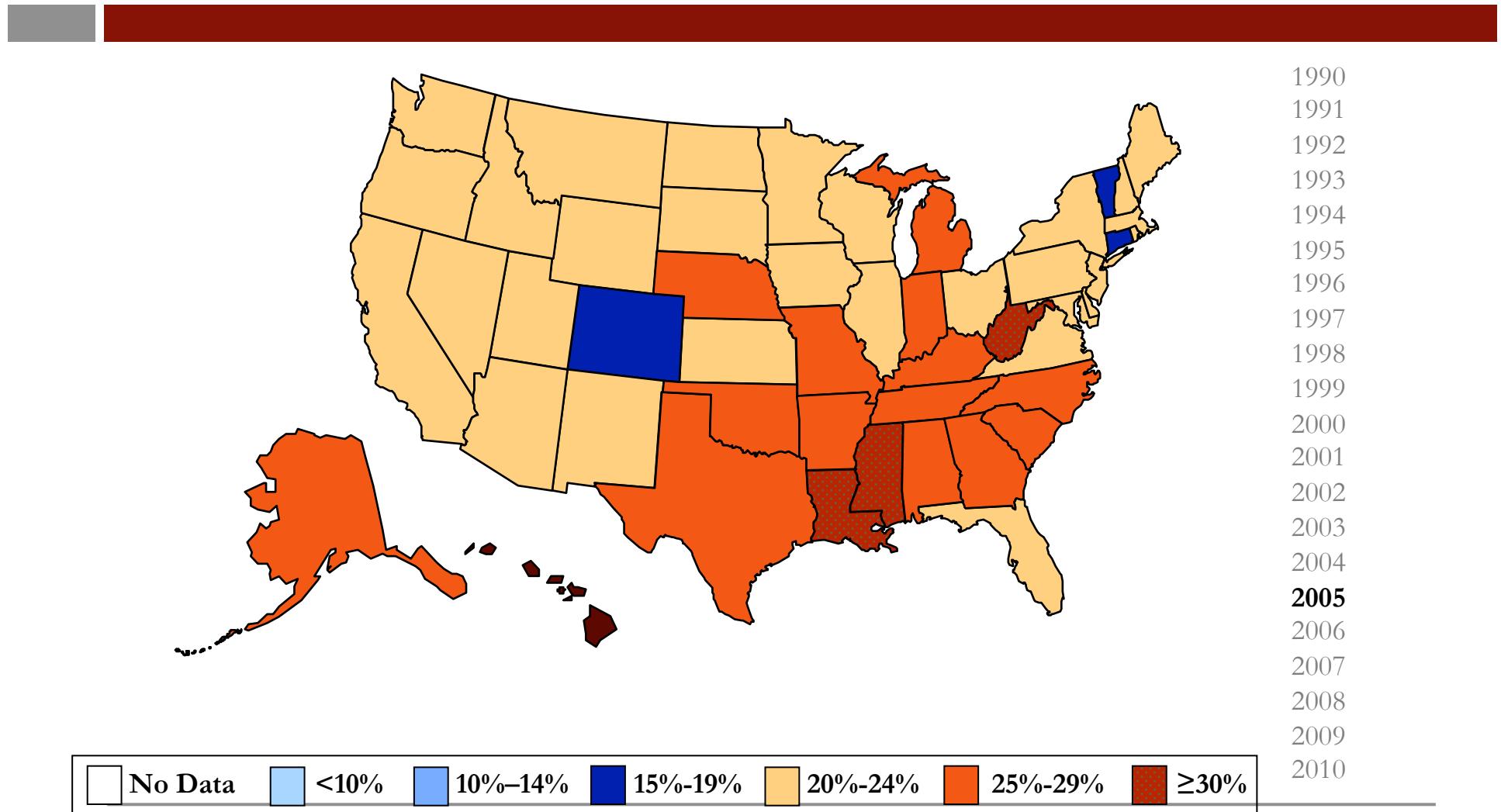
Obesity Trends Among U.S. Adults



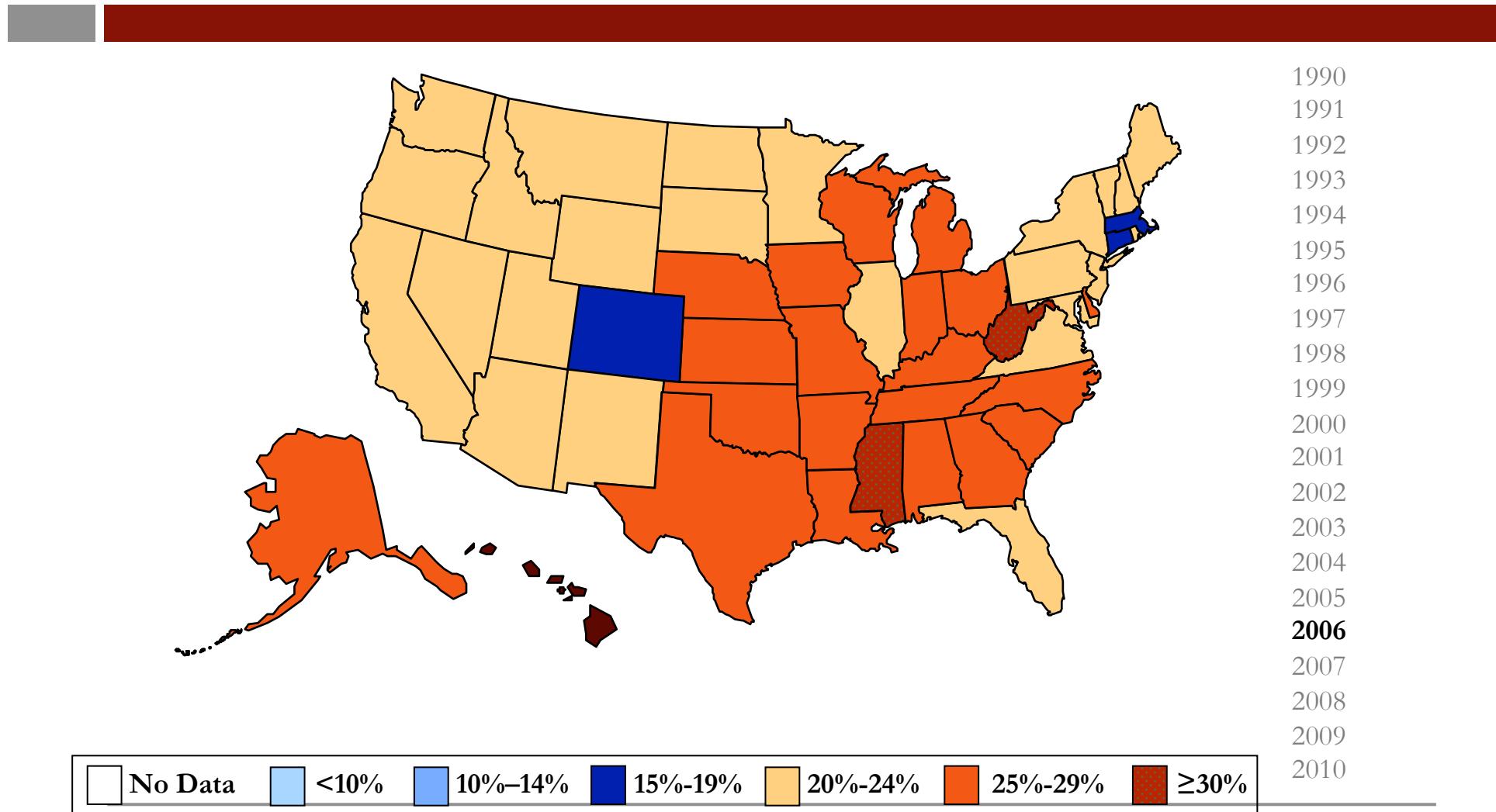
Obesity Trends Among U.S. Adults



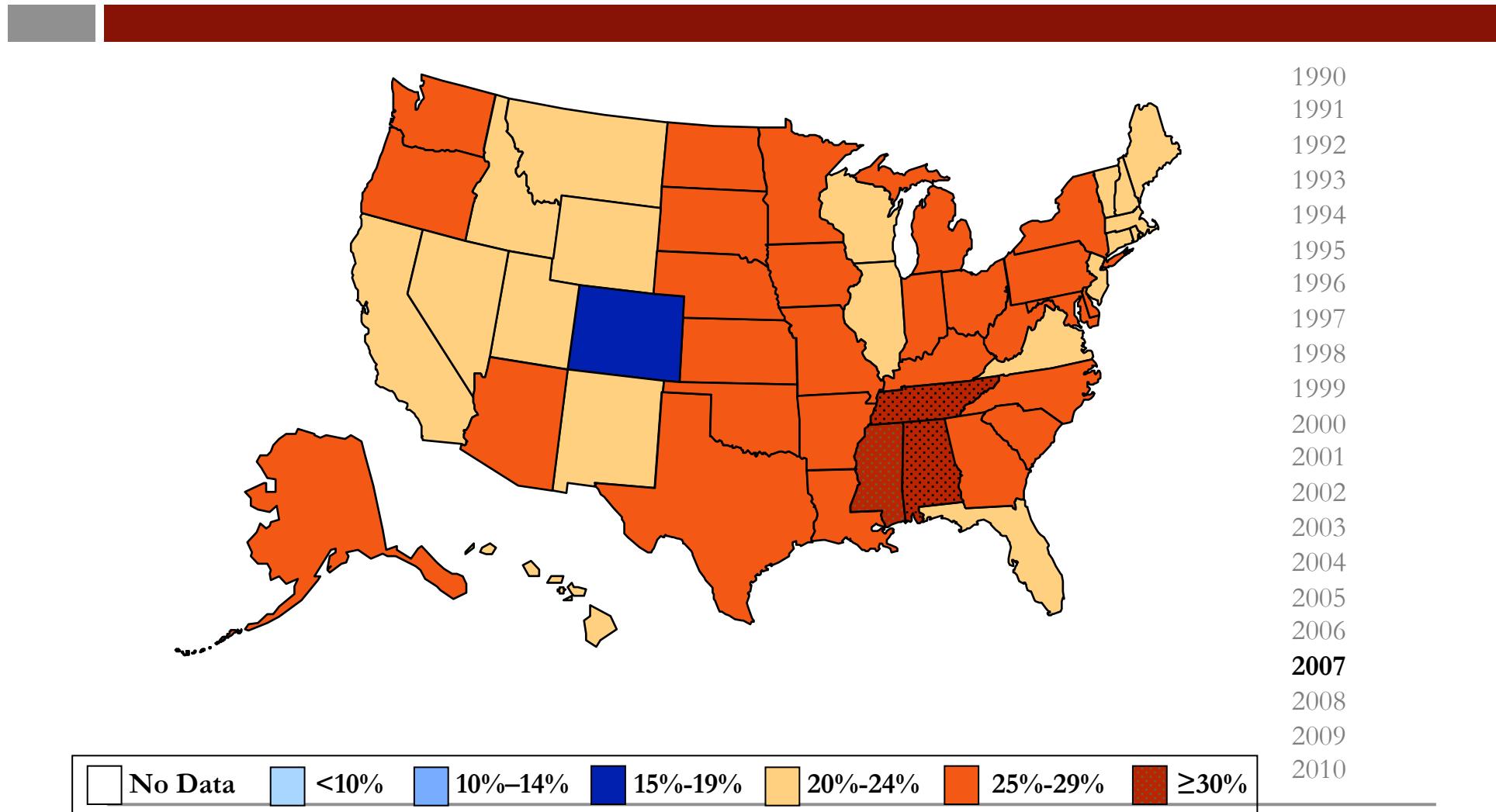
Obesity Trends Among U.S. Adults



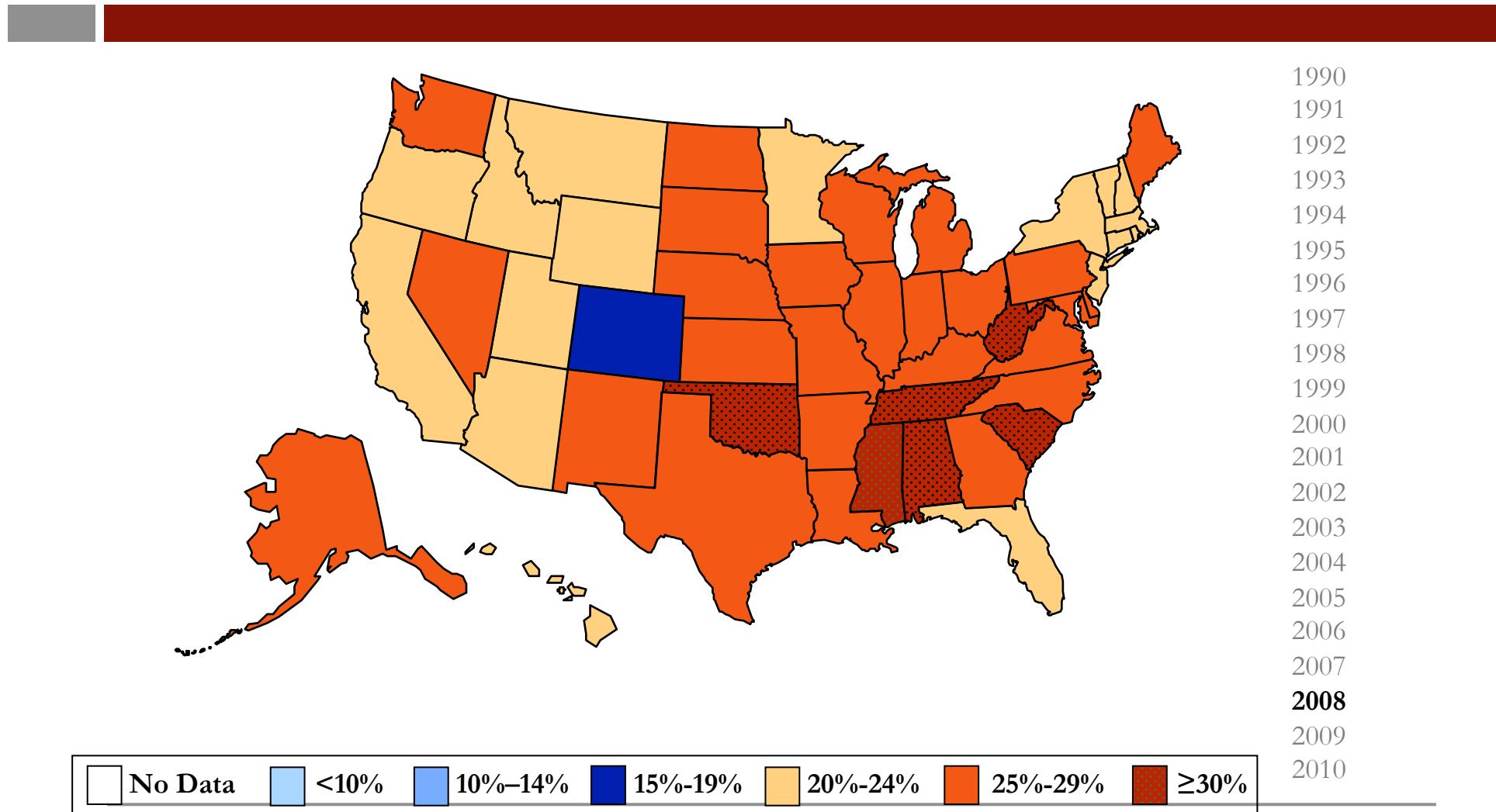
Obesity Trends Among U.S. Adults



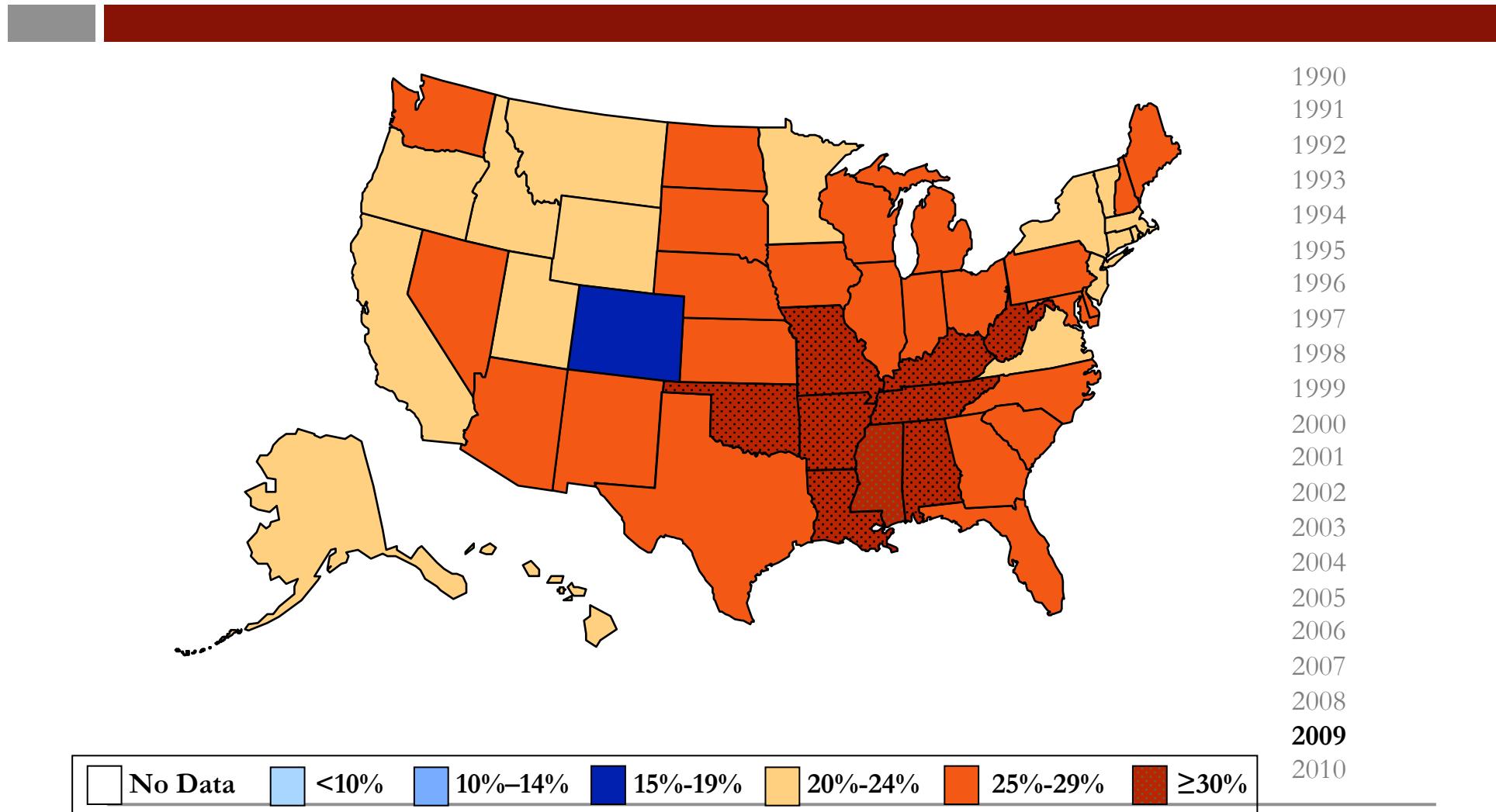
Obesity Trends Among U.S. Adults



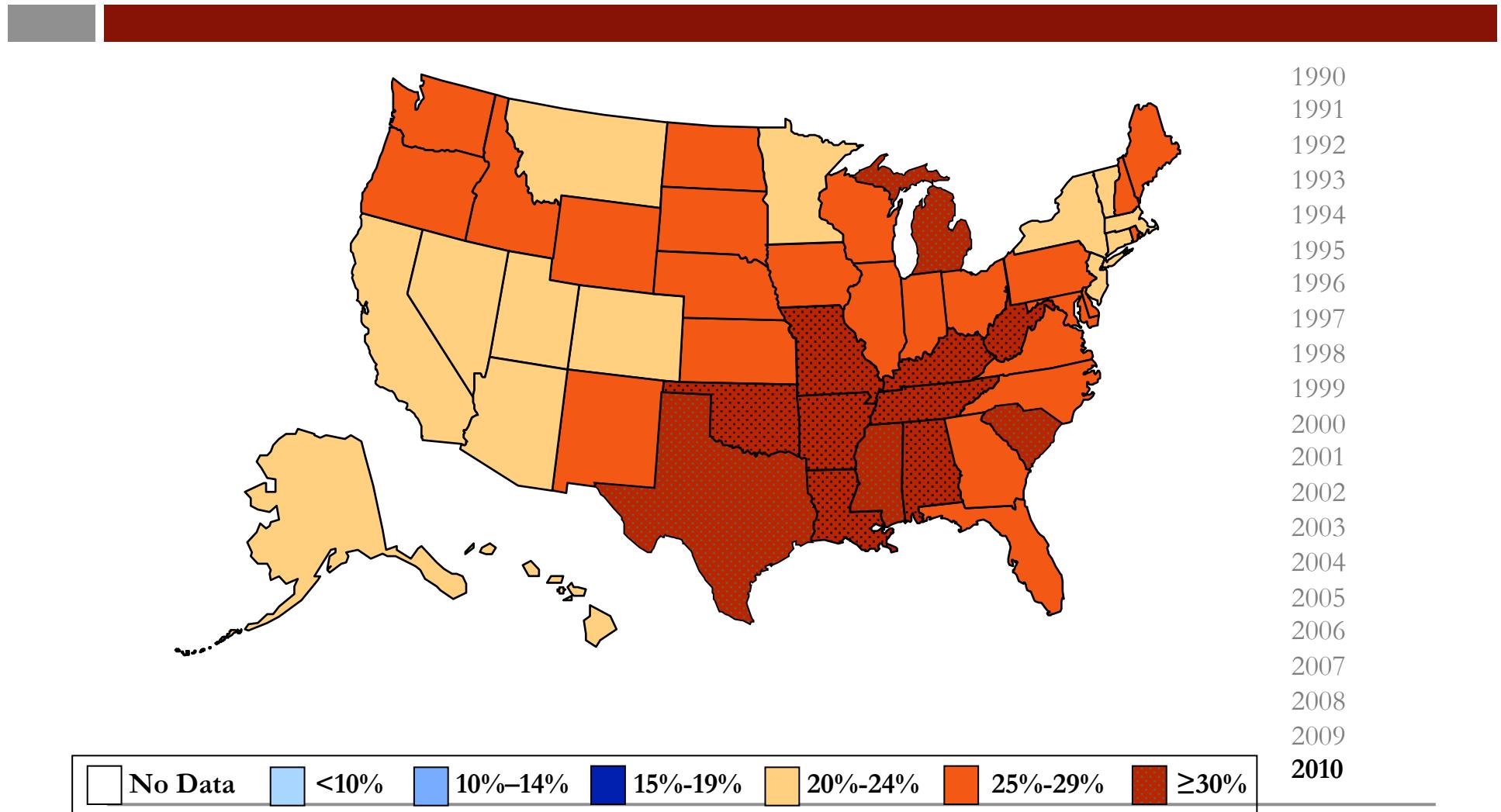
Obesity Trends Among U.S. Adults



Obesity Trends Among U.S. Adults



Obesity Trends Among U.S. Adults



Obesity

- More than 35% of US adults are obese
 - Obesity-related conditions are some of the leading causes of preventable death (heart disease, stroke, type II diabetes)
- Worldwide, obesity has nearly doubled since 1980
- 65% of the world's population lives in countries where overweight and obesity kills more people than underweight

Nutrition

- Good nutrition is essential for a person's overall health and well-being, and is now more important than ever
- Hundreds of nutrition and weight-loss applications
 - 15% of adults with cell phones use health applications on their devices
- These apps are powered by the USDA Food Database

USDA Food Database

- The United States Department of Agriculture distributes a database of nutritional information for over 7,000 different food items
- Used as the foundation for most food and nutrient databases in the US
- Includes information about all nutrients
 - Calories, carbs, protein, fat, sodium, . . .

Summarizing by Group: tapply

- The tapply function takes three arguments

```
tapply(argument1, argument2, argument3)
```

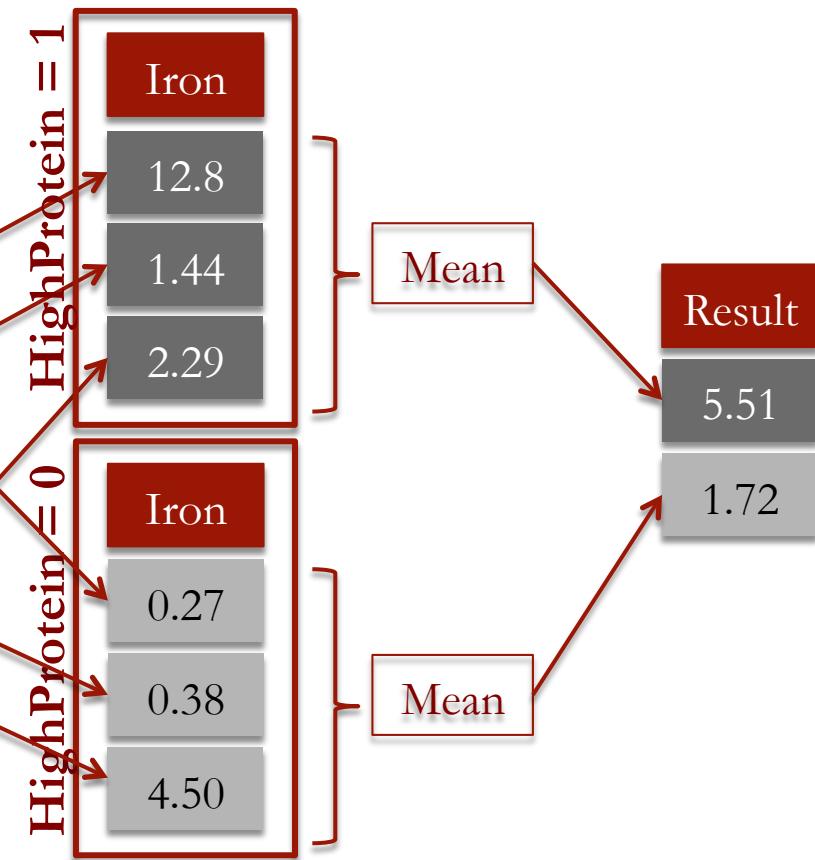
- Group argument 1 by argument 2 and apply argument 3
- To compute the average amount of iron, sorted by high and low protein

```
tapply(USDA$Iron, USDA$HighProtein, mean, na.rm=TRUE)
```

What exactly does tapply do?

```
tapply(USDA$Iron, USDA$HighProtein, mean, na.rm=TRUE)
```

Food	Iron	HighProtein
1	0.27	0
2	12.8	1
3	1.44	1
4	0.38	0
5	4.50	0
6	2.29	1



Summarizing by Group: tapply

- The tapply function takes three arguments

```
tapply(argument1, argument2, argument3)
```

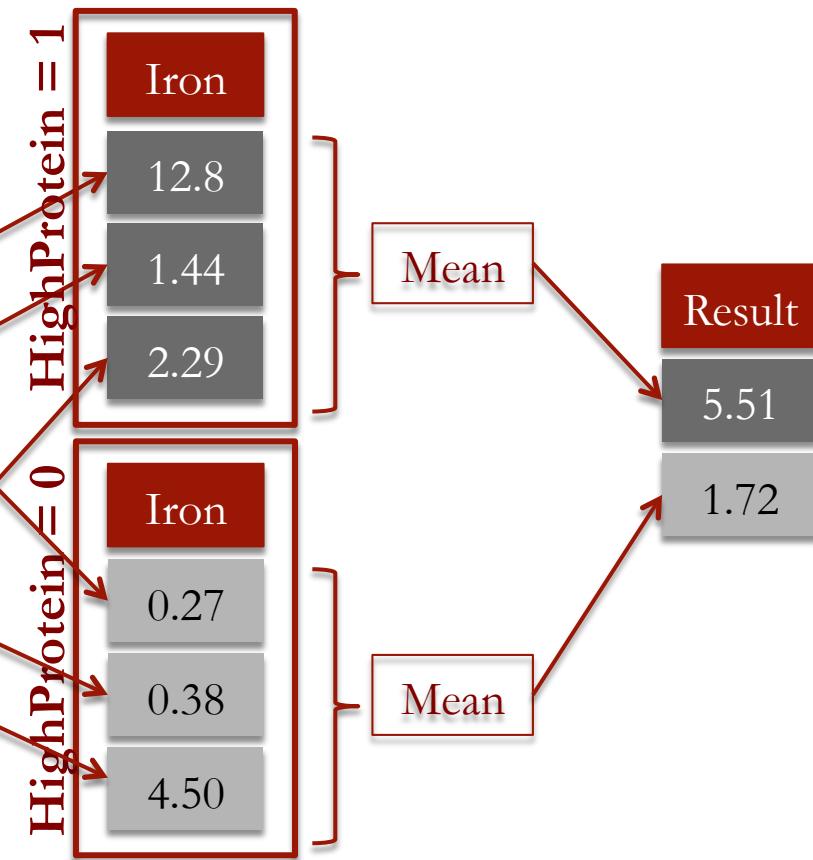
- Group argument 1 by argument 2 and apply argument 3
- To compute the average amount of iron, sorted by high and low protein

```
tapply(USDA$Iron, USDA$HighProtein, mean, na.rm=TRUE)
```

What exactly does tapply do?

```
tapply(USDA$Iron, USDA$HighProtein, mean, na.rm=TRUE)
```

Food	Iron	HighProtein
1	0.27	0
2	12.8	1
3	1.44	1
4	0.38	0
5	4.50	0
6	2.29	1





MONEYBALL

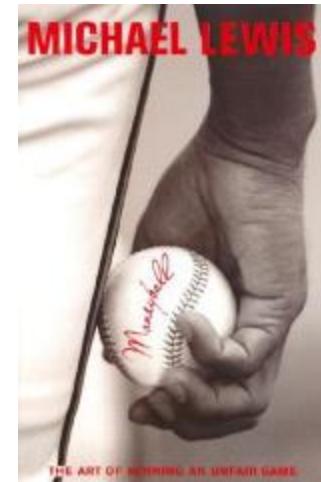
The Power of Sports Analytics

15.071 – The Analytics Edge

The Story

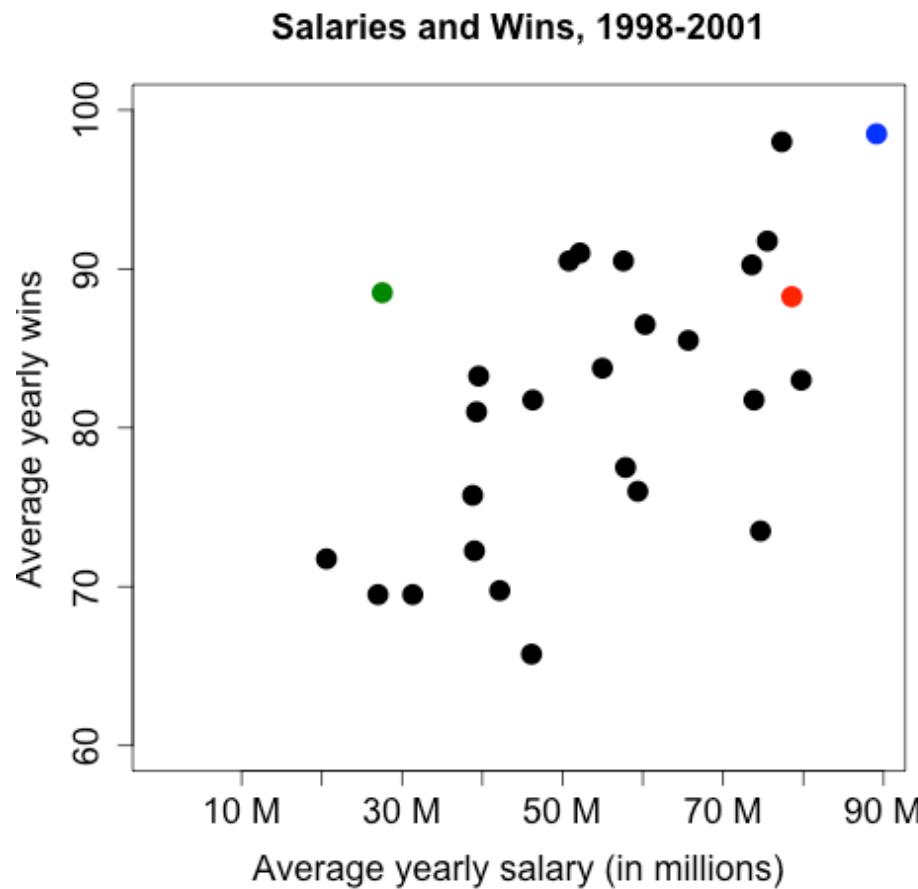
- *Moneyball* tells the story of the Oakland A's in 2002
 - One of the poorest teams in baseball
 - New ownership and budget cuts in 1995
 - But they were improving

Year	Win %
1997	40%
1998	46%
1999	54%
2000	57%
2001	63%



- How were they doing it?
 - Was it just luck?
- In 2002, the A's lost three key players
- Could they continue winning?

The Problem



- Rich teams can afford the all-star players
- How do the poor teams compete?

Competing as a Poor Team



- Competitive imbalances in the game
 - Rich teams have four times the salary of poor teams
- The Oakland A's can't afford the all-stars, but they are still making it to the playoffs. How?
- They take a quantitative approach and find undervalued players

A Different Approach



- The A's started using a different method to select players
- The traditional way was through scouting
 - Scouts would go watch high school and college players
 - Report back about their skills
 - A lot of talk about speed and athletic build
- The A's selected players based on their statistics, not on their looks
 - “The statistics enabled you to find your way past all sorts of sight-based scouting prejudices.”
 - “We’re not selling jeans here”

The Perfect Batter

The A's



A catcher who couldn't throw
Gets on base a lot

The Yankees



A consistent shortstop
Leader in hits and stolen bases

The Perfect Pitcher

The A's



Unconventional delivery
Slow speed

The Yankees



Conventional delivery
Fast speed

Billy Beane



- The general manager since 1997
- Played major league baseball, but never made it big
 - Sees himself as a typical scouting error
- Billy Beane succeeded in using analytics
 - Had a management position
 - Understood the importance of statistics – hired Paul DePodesta (a Harvard graduate) as his assistant
 - Didn't care about being ostracized

Taking a Quantitative View

- Paul DePodesta spent a lot of time looking at the data
- His analysis suggested that some skills were undervalued and some skills were overvalued
- If they could detect the undervalued skills, they could find players at a bargain



The Goal of a Baseball Team

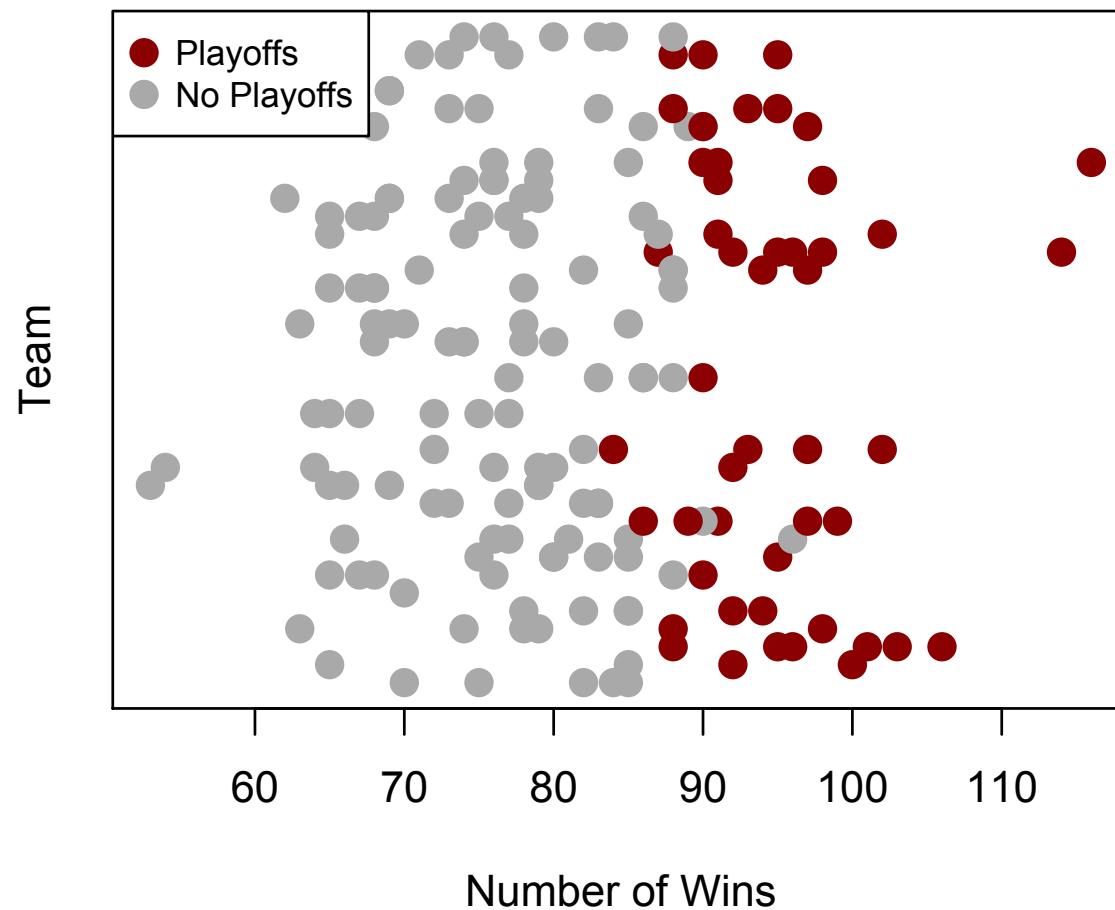


Making it to the Playoffs



- How many games does a team need to win in the regular season to make it to the playoffs?
- “Paul DePodesta reduced the regular season to a math problem. He judged how many wins it would take to make it to the playoffs: 95.”

Making it to the Playoffs



Data from
all teams
1996-2001

Winning 95 Games



- How does a team win games?
- They score more runs than their opponent
- But how many more?
- The A's calculated that they needed to score 135 more runs than they allowed during the regular season to expect to win 95 games
- Let's see if we can verify this using linear regression

The Goal of a Baseball Team



Scoring Runs



- How does a team score more runs?
- The A's discovered that two baseball statistics were significantly more important than anything else
 - On-Base Percentage (OBP)
 - Percentage of time a player gets on base (including walks)
 - Slugging Percentage (SLG)
 - How far a player gets around the bases on his turn (measures power)

Scoring Runs



- Most teams focused on Batting Average (BA)
 - Getting on base by hitting the ball
- The A's claimed that:
 - On-Base Percentage was the most important
 - Slugging Percentage was important
 - Batting Average was overvalued
- Can we use linear regression to verify which baseball stats are more important to predict runs?

Allowing Runs

- We can use pitching statistics to predict runs allowed
 - Opponents On-Base Percentage (OOBP)
 - Opponents Slugging Percentage (OSLG)
- We get the linear regression model
 $\text{Runs Allowed} = -837.38 + 2913.60(\text{OOBP}) + 1514.29(\text{OSLG})$
- $R^2 = 0.91$
- Both variables significant

Predicting Runs and Wins



- Can we predict how many games the 2002 Oakland A's will win using our models?
- The models for runs use team statistics
- Each year, a baseball team is different
- We need to estimate the new team statistics using past player performance
 - Assumes past performance correlates with future performance
 - Assumes few injuries
- We can estimate the team statistics for 2002 by using the 2001 player statistics

Predicting Runs Scored



- At the beginning of the 2002 season, the Oakland A's had 24 batters on their roster
- Using the 2001 regular season statistics for these players
 - Team OBP is 0.339
 - Team SLG is 0.430
- Our regression equation was

$$RS = -804.63 + 2737.77(\text{OBP}) + 1584.91(\text{SLG})$$

- Our 2002 prediction for the A's is

$$RS = -804.63 + 2737.77(0.339) + 1584.91(0.430) = 805$$

Predicting Runs Allowed

- At the beginning of the 2002 season, the Oakland A's had 17 pitchers on their roster
- Using the 2001 regular season statistics for these players
 - Team OOBP is 0.307
 - Team OSLG is 0.373
- Our regression equation was

$$RA = -837.38 + 2913.60(\text{OOBP}) + 1514.29(\text{OSLG})$$

- Our 2002 prediction for the A's is

$$RA = -837.38 + 2913.60(0.307) + 1514.29 (0.373) = 622$$

Predicting Wins



- Our regression equation to predict wins was

$$\text{Wins} = 80.8814 + 0.1058(\text{RS} - \text{RA})$$

- We predicted

- $\text{RS} = 805$

- $\text{RA} = 622$

- So our prediction for wins is

$$\text{Wins} = 80.8814 + 0.1058(805 - 622) = 100$$

The Oakland A's

- Paul DePodesta used a similar approach to make predictions
- Predictions closely match actual performance

	Our Prediction	Paul's Prediction	Actual
Runs Scored	805	800 – 820	800

The Oakland A's

- Paul DePodesta used a similar approach to make predictions
- Predictions closely match actual performance

	Our Prediction	Paul's Prediction	Actual
Runs Scored	805	800 – 820	800
Runs Allowed	622	650 – 670	653

The Oakland A's

- Paul DePodesta used a similar approach to make predictions
- Predictions closely match actual performance

	Our Prediction	Paul's Prediction	Actual
Runs Scored	805	800 – 820	800
Runs Allowed	622	650 – 670	653
Wins	100	93 – 97	103

- The A's set a League record by winning 20 games in a row
- Won one more game than the previous year, and made it to the playoffs

The Goal of a Baseball Team



Why isn't the goal
to win the World
Series?

Luck in the Playoffs



- Billy and Paul see their job as making sure the team makes it to the playoffs – after that all bets are off
 - The A's made it to the playoffs in 2000, 2001, 2002, 2003
 - But they didn't win the World Series
- Why?
- “Over a long season the luck evens out, and the skill shines through. But in a series of three out of five, or even four out of seven, anything can happen.”

Is Playoff Performance Predictable?



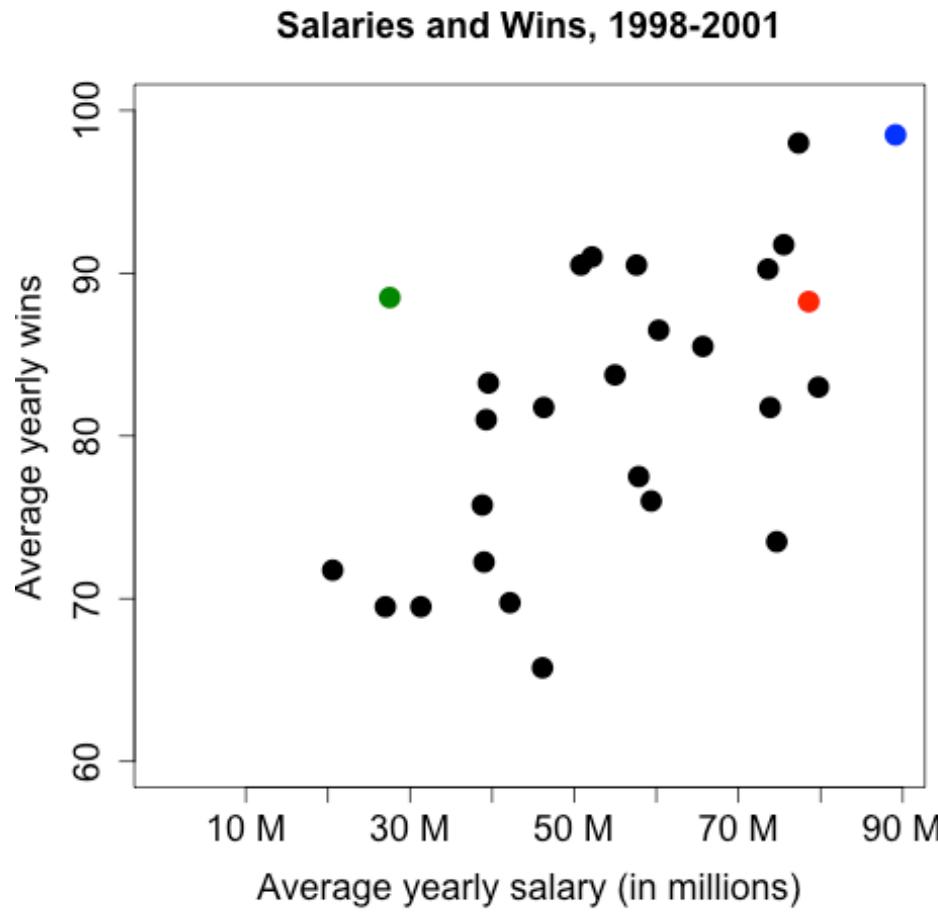
- Using data 1994-2011 (8 teams in the playoffs)
- Correlation between winning the World Series and regular season wins is 0.03
- Winning regular season games gets you to the playoffs
- But in the playoffs, there are too few games for luck to even out
- *Logistic regression* can be used to predict whether or not a team will win the World Series

Other Moneyball Strategies



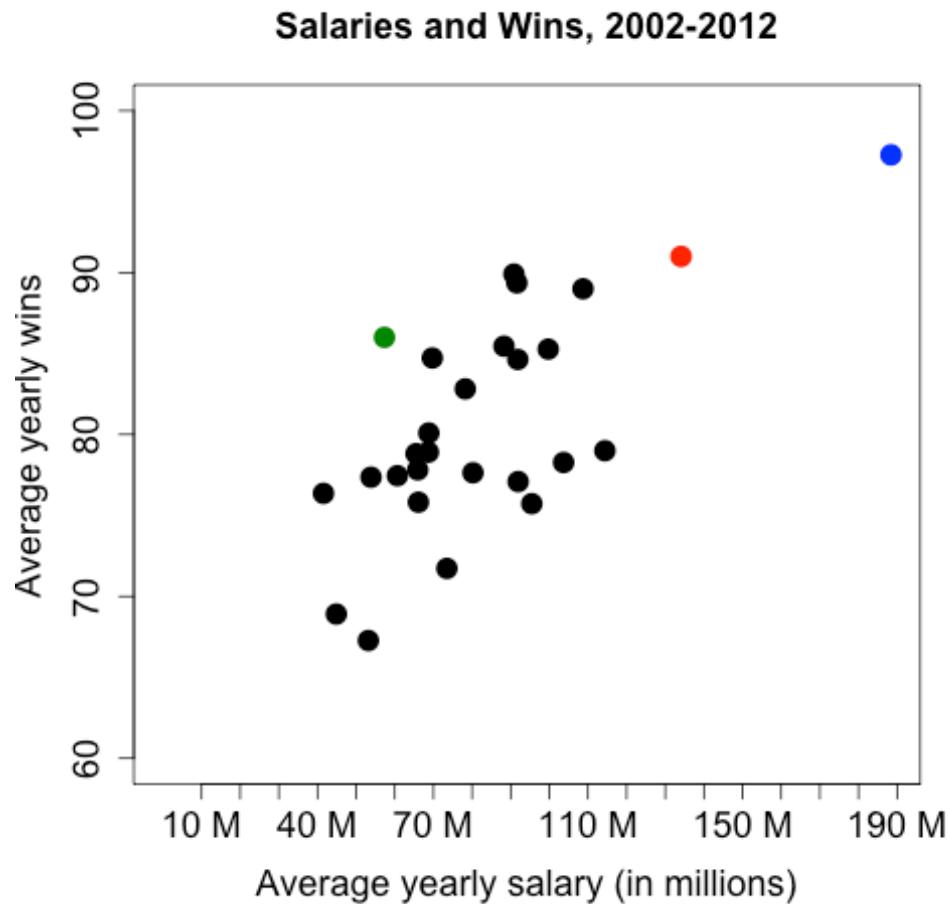
- *Moneyball* also discusses:
 - How it is easier to predict professional success of college players than high school players
 - Stealing bases, sacrifice bunting, and sacrifice flies are overrated
 - Pitching statistics do not accurately measure pitcher ability – pitchers only control strikeouts, home runs, and walks

Where was Baseball in 2002?



- Before Moneyball techniques became more well-known, the A's were an outlier
- 20 more wins than teams with equivalent payrolls
- As many wins as teams with more than double the payroll

Where is Baseball Now?



- Now, the A's are still an efficient team, but they only have 10 more wins than teams with equivalent payrolls
- Fewer inefficiencies

Sabermetrics

- Sabermetrics is a more general term for Moneyball techniques
- There has been a lot of work done in this field
 - Baseball Prospectus (www.baseballprospectus.com)
 - Value Over Replacement Player (VORP)
 - Defense Independent Pitching Statistics (DIPS)
 - *The Extra 2%: How Wall Street Strategies Took a Major League Baseball Team from Worst to First*
 - A story of the Tampa Bay Rays
 - Game-time decisions: batting order, changing pitchers, etc.

Other Baseball Teams and Sports



- Every major league baseball team now has a statistics group
- The Red Sox implemented quantitative ideas and won the World Series for the first time in 86 years
- Analytics are also used in other sports, although it is believed that more teams use statistical analysis than is publically known

The Analytics Edge



- Models allow managers to more accurately value players and minimize risk
 - “In human behavior there was always uncertainty and risk. The goal of the Oakland front office was simply to minimize the risk. Their solution wasn’t perfect, it was just better than ... rendering decisions by gut feeling.”
- Relatively simple models can be useful



THE STATISTICAL SOMMELIER

An Introduction to Linear Regression

15.071 – The Analytics Edge

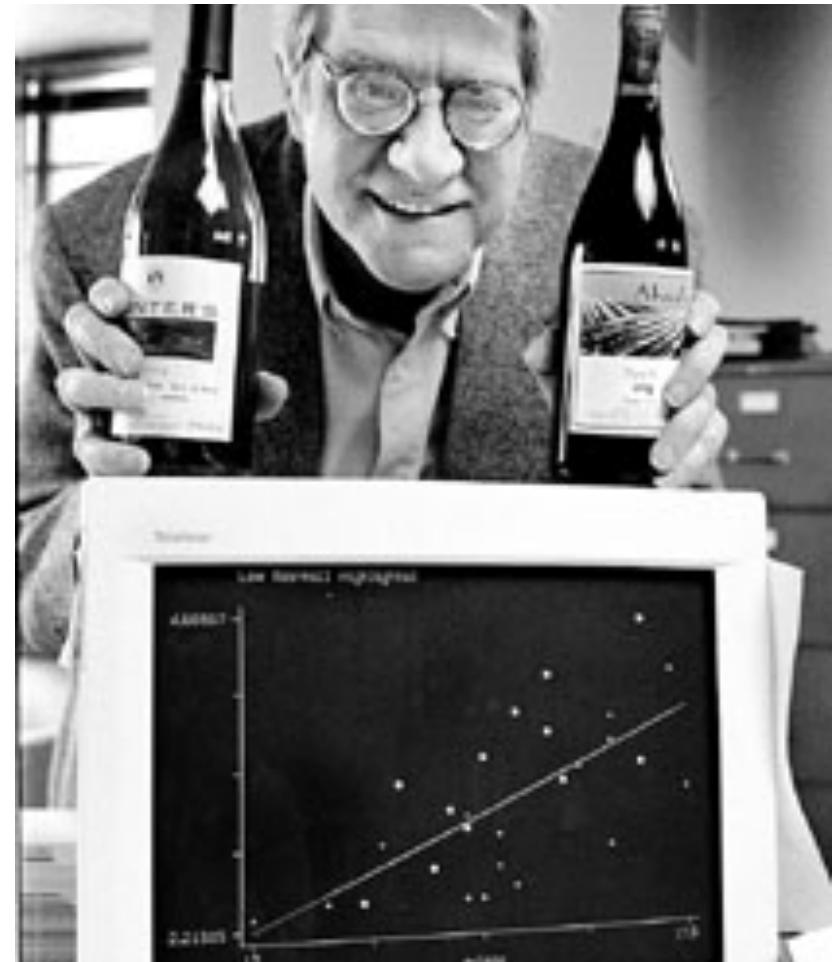
Bordeaux Wine



- Large differences in price and quality between years, although wine is produced in a similar way
- Meant to be aged, so hard to tell if wine will be good when it is on the market
- Expert tasters predict which ones will be good
- Can analytics be used to come up with a different system for judging wine?

Predicting the Quality of Wine

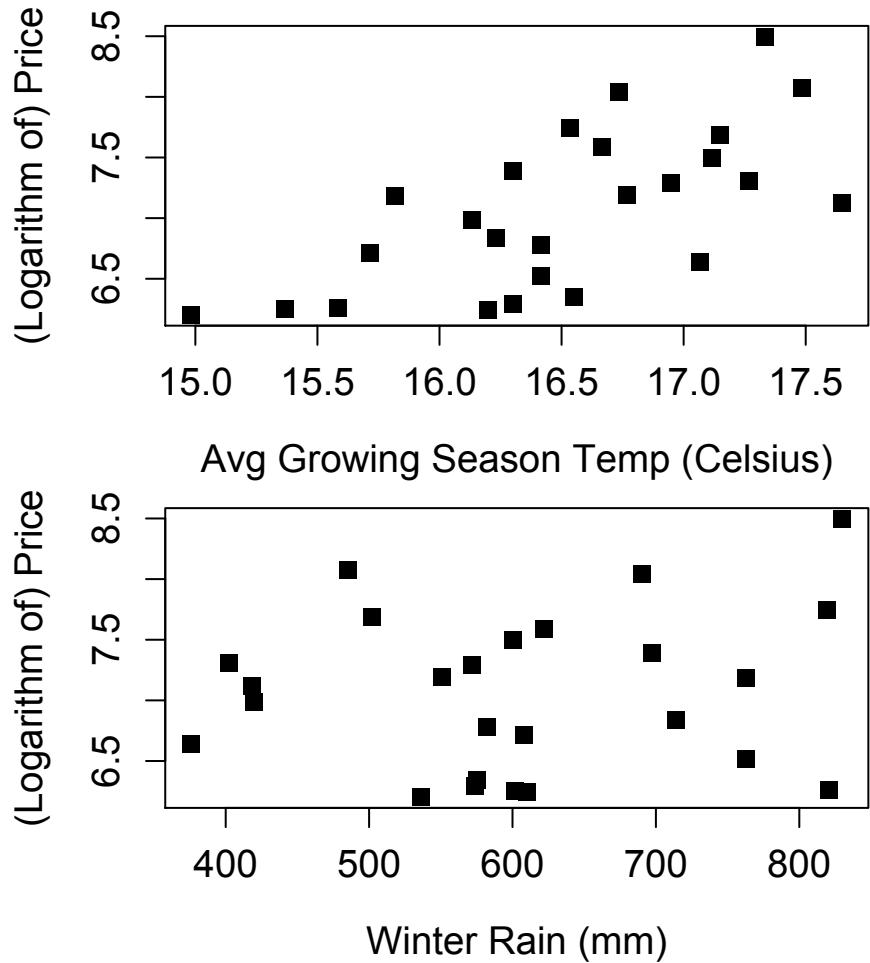
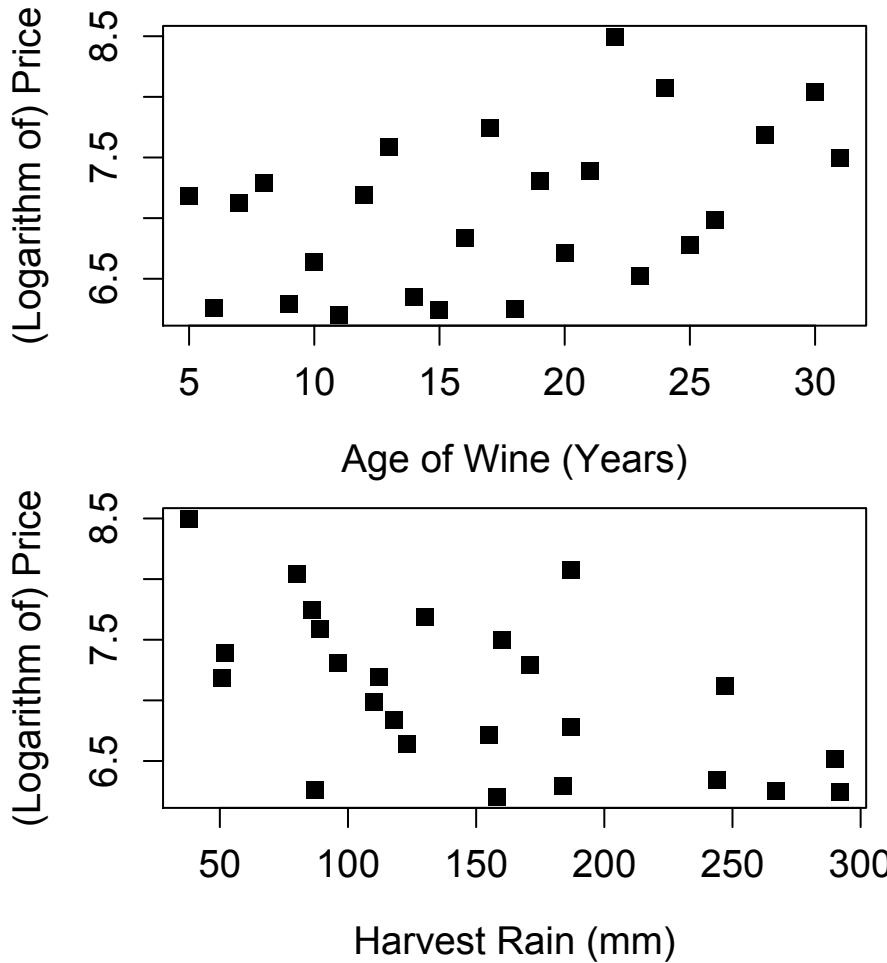
- March 1990 - Orley Ashenfelter, a Princeton economics professor, claims he can predict wine quality without tasting the wine



Building a Model

- Ashenfelter used a method called **linear regression**
 - Predicts an outcome variable, or *dependent variable*
 - Predicts using a set of *independent variables*
- Dependent variable: typical price in 1990-1991 wine auctions (approximates quality)
- Independent variables:
 - Age – older wines are more expensive
 - Weather
 - Average Growing Season Temperature
 - Harvest Rain
 - Winter Rain

The Data (1952 – 1978)

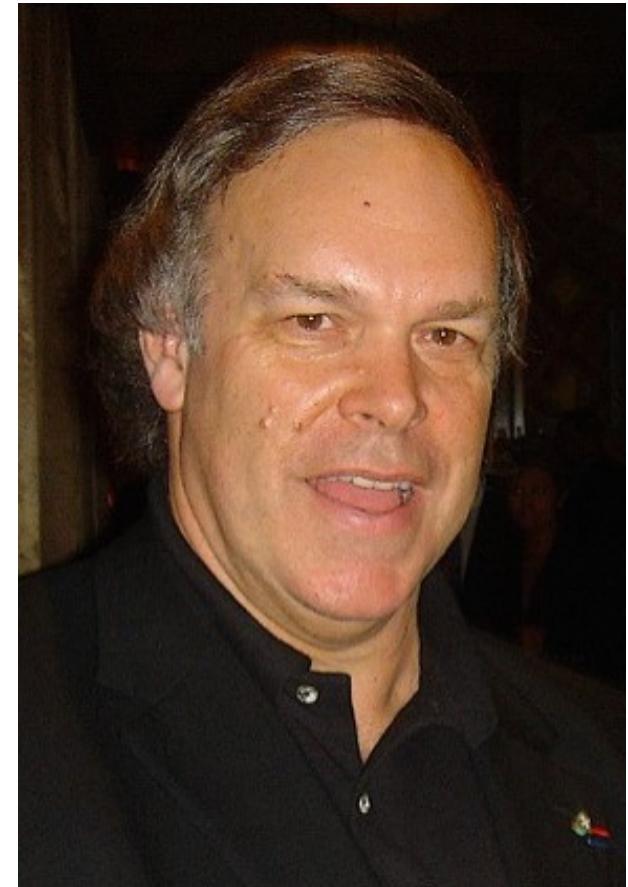


The Expert's Reaction

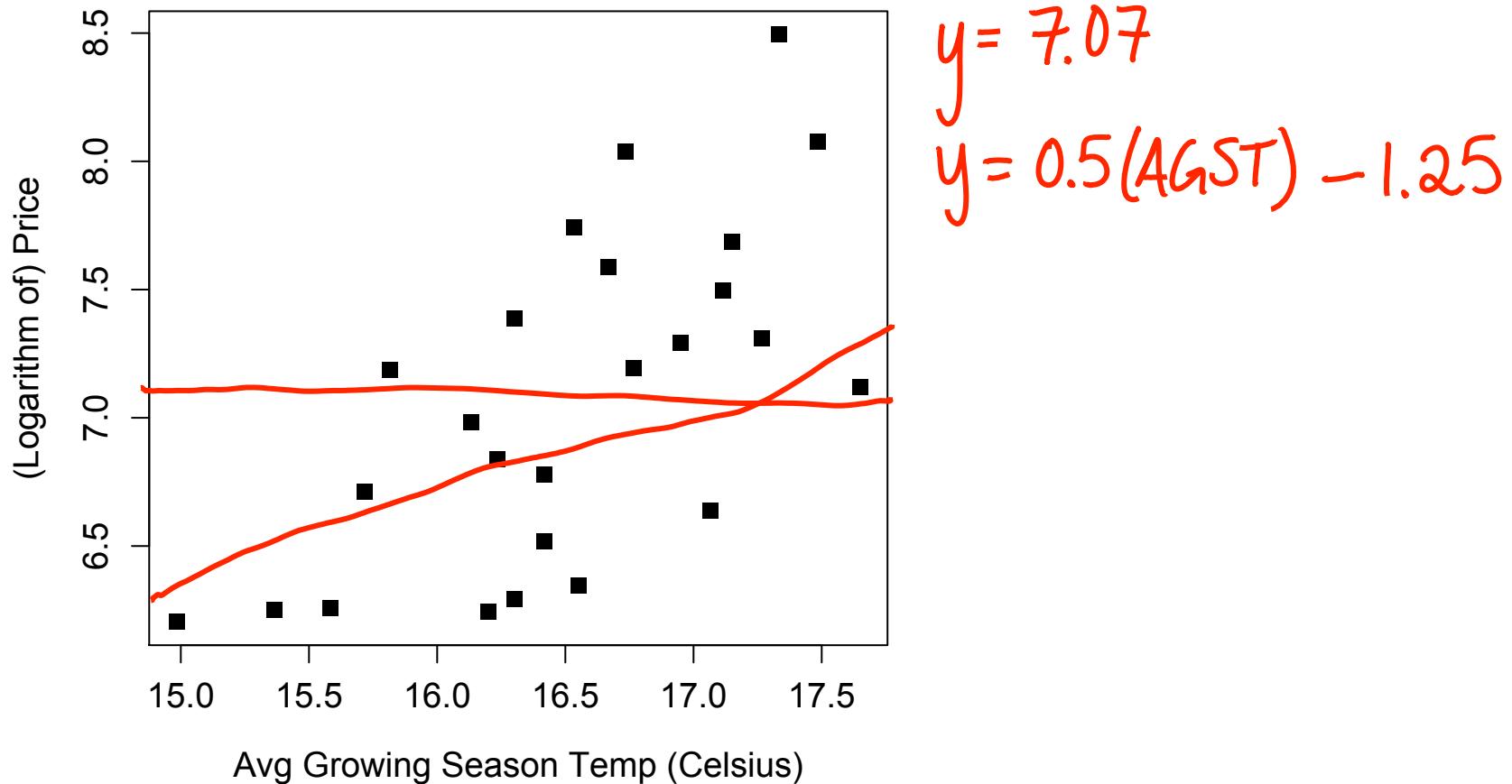
Robert Parker, the world's most influential wine expert:

“Ashenfelter is an absolute total sham”

“rather like a movie critic who never goes to see the movie but tells you how good it is based on the actors and the director”



One-Variable Linear Regression



The Regression Model

- One-variable regression model

$$y^i = \beta_0 + \beta_1 x^i + \epsilon^i$$

y^i = dependent variable (wine price) for the i^{th} observation

x^i = independent variable (temperature) for the i^{th} observation

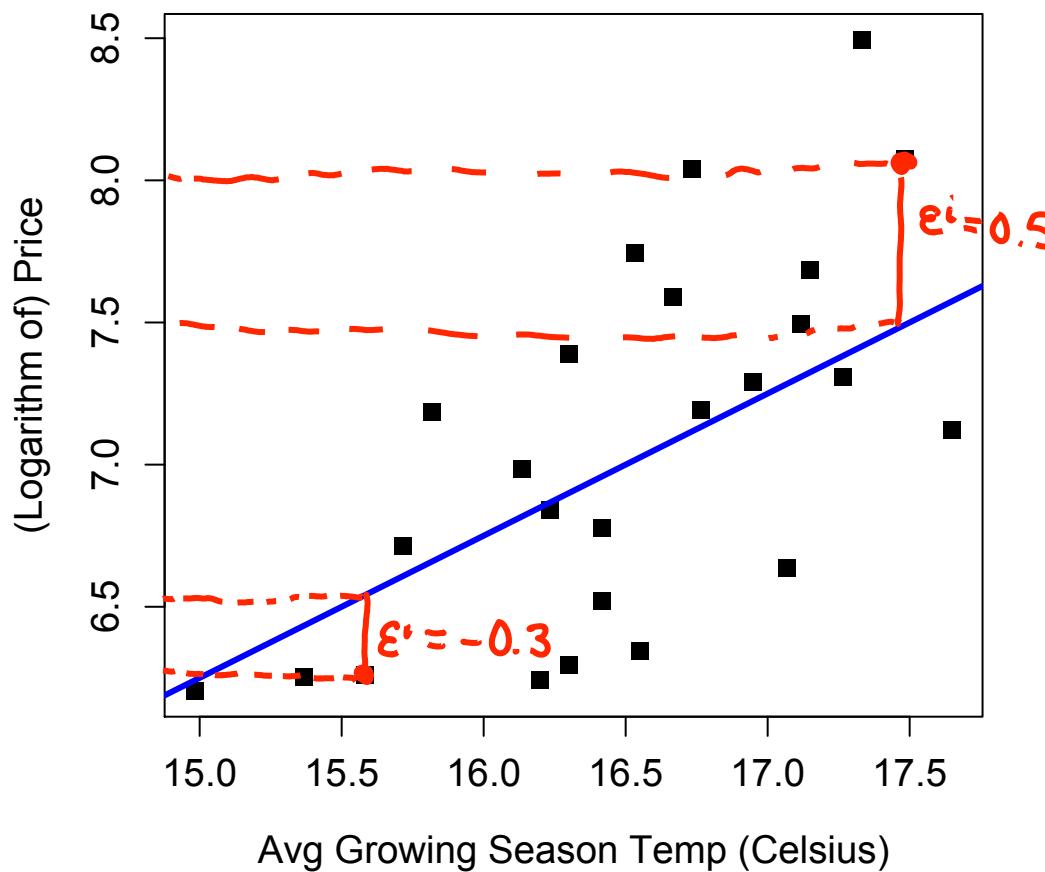
ϵ^i = error term for the i^{th} observation

β_0 = intercept coefficient

β_1 = regression coefficient for the independent variable

- The best model (choice of coefficients) has the smallest error terms

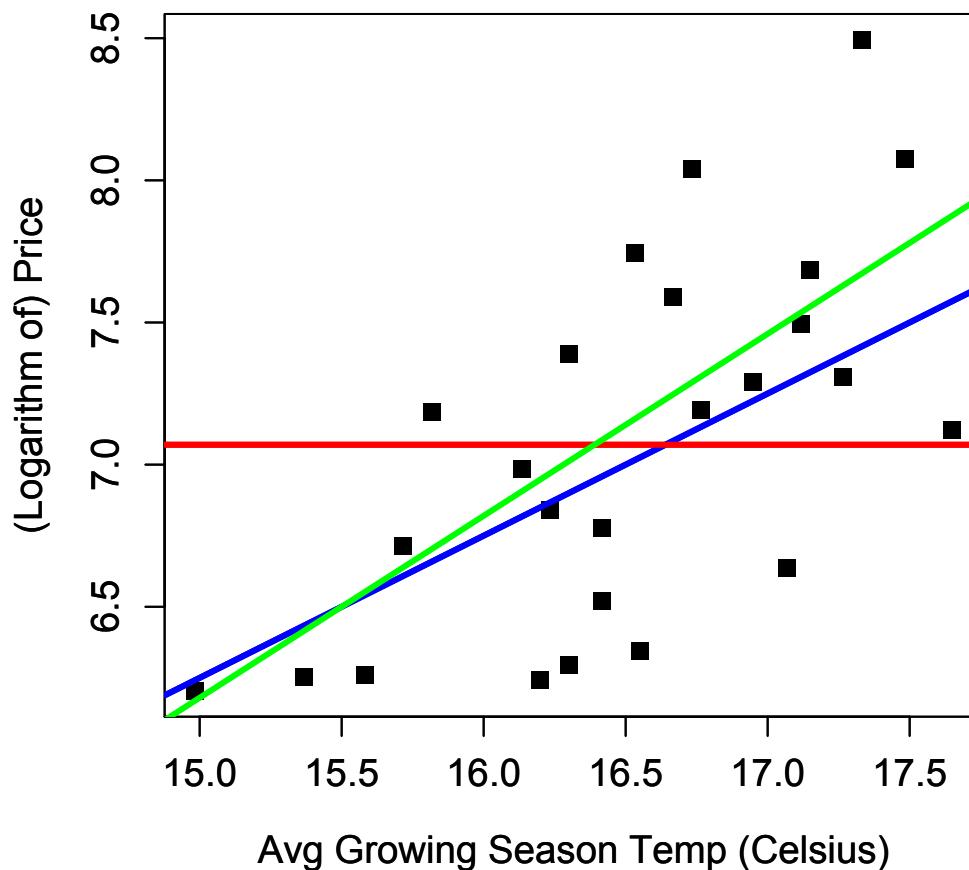
Selecting the Best Model



$$SSE = (e^1)^2 + (e^2)^2 + \dots + (e^N)^2$$

$N = \# \text{data points}$

Selecting the Best Model



SSE = 10.15

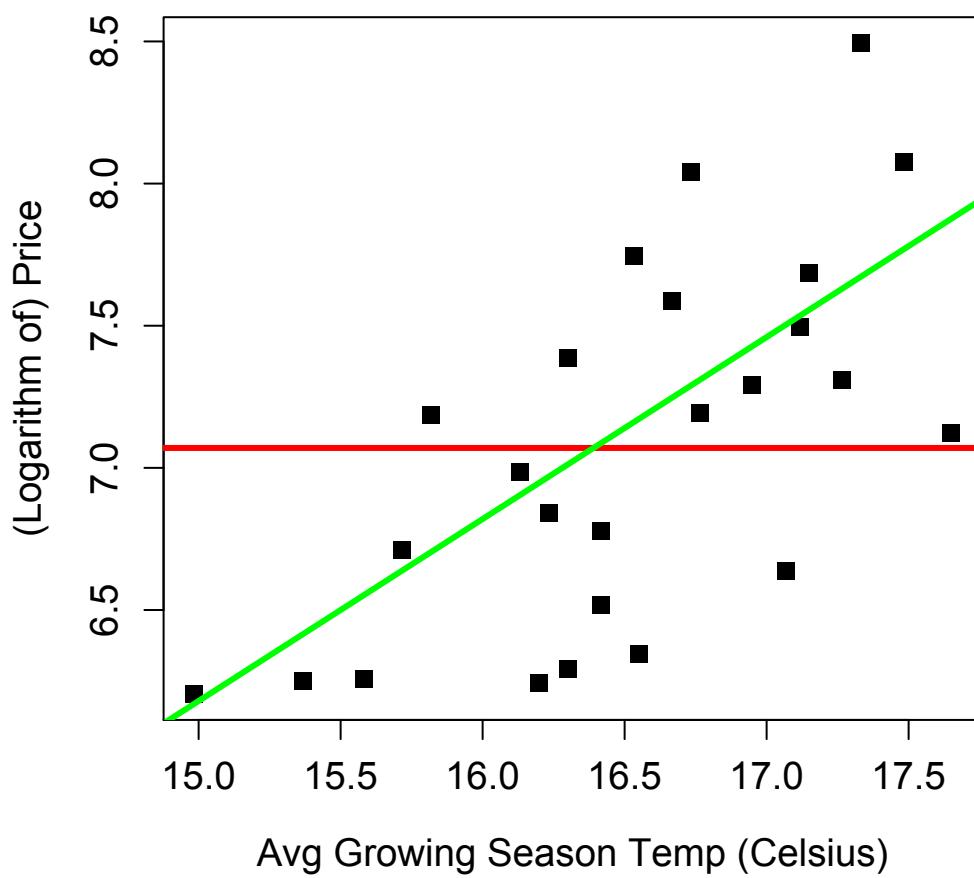
SSE = 6.03

SSE = 5.73

Other Error Measures

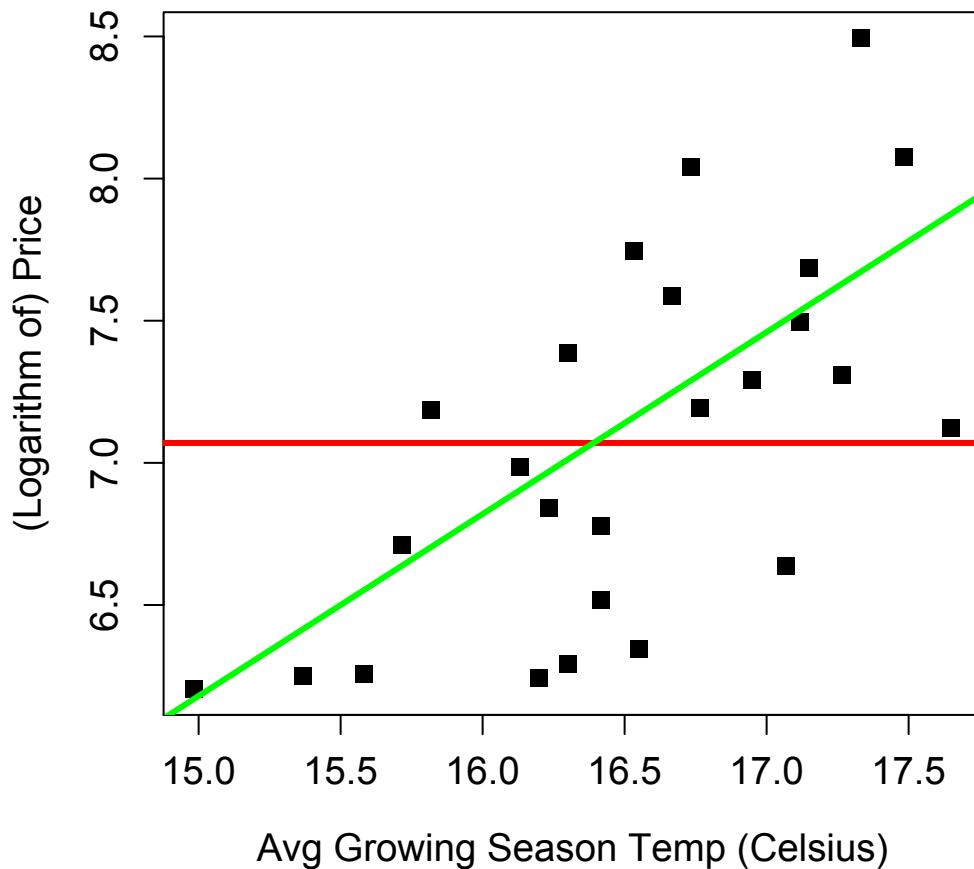
- SSE can be hard to interpret
 - Depends on N
 - Units are hard to understand
- Root-Mean-Square Error (RMSE)
$$RMSE = \sqrt{\frac{SSE}{N}}$$
- Normalized by N, units of dependent variable

R^2



- Compares the best model to a “baseline” model
- The **baseline model** does not use any variables
 - Predicts same outcome (price) regardless of the independent variable (temperature)

R²



$$SSE = 5.73$$

$$SST = 10.15$$

$$R^2 = 1 - \frac{SSE}{SST}$$

$$= 1 - \frac{5.73}{10.15}$$

$$= 0.44$$

Interpreting R²

$$R^2 = 1 - \frac{SSE}{SST}$$

$$0 \leq SSE \leq SST$$
$$0 \leq SST$$

- R² captures value added from using a model
 - R² = 0 means no improvement over baseline
 - R² = 1 means a perfect predictive model
- Unitless and universally interpretable
 - Can still be hard to compare between problems
 - Good models for easy problems will have R² ≈ 1
 - Good models for hard problems can still have R² ≈ 0

Available Independent Variables



- So far, we have only used the Average Growing Season Temperature to predict wine prices
- Many different independent variables could be used
 - Average Growing Season Temperature
 - Harvest Rain
 - Winter Rain
 - Age of Wine (in 1990)
 - Population of France

Multiple Linear Regression

- Using each variable on its own:
 - $R^2 = 0.44$ using Average Growing Season Temperature
 - $R^2 = 0.32$ using Harvest Rain
 - $R^2 = 0.22$ using France Population
 - $R^2 = 0.20$ using Age
 - $R^2 = 0.02$ using Winter Rain
- Multiple linear regression allows us to use all of these variables to improve our predictive ability

The Regression Model

- Multiple linear regression model with k variables

$$y^i = \beta_0 + \beta_1 x_1^i + \beta_2 x_2^i + \dots + \beta_k x_k^i + \epsilon^i$$

y^i = dependent variable (wine price) for the i^{th} observation

x_j^i = j^{th} independent variable for the i^{th} observation

ϵ^i = error term for the i^{th} observation

β_0 = intercept coefficient

β_j = regression coefficient for the j^{th} independent variable

- Best model coefficients selected to minimize SSE

Adding Variables

Variables	R ²
Average Growing Season Temperature (AGST)	0.44
AGST, Harvest Rain	0.71
AGST, Harvest Rain, Age	0.79
AGST, Harvest Rain, Age, Winter Rain	0.83
AGST, Harvest Rain, Age, Winter Rain, Population	0.83

- Adding more variables can improve the model
- Diminishing returns as more variables are added

Selecting Variables

- Not all available variables should be used
 - Each new variable requires more data
 - Causes *overfitting*: high R^2 on data used to create model, but bad performance on unseen data
- We will see later how to appropriately choose variables to remove

Understanding the Model and Coefficients

Coefficients:

		Estimate	Std. Error	t value	Pr(> t)	Estimate Std. Error
(Intercept)		-4.504e-01	1.019e+01	-0.044	0.965202	
AvgGrowingSeasonTemp		6.012e-01	1.030e-01	5.836	1.27e-05	***
HarvestRain		-3.958e-03	8.751e-04	-4.523	0.000233	***
Age		5.847e-04	7.900e-02	0.007	0.994172	
WinterRain		1.043e-03	5.310e-04	1.963	0.064416	.
FrancePopulation		-4.953e-05	1.667e-04	-0.297	0.769578	

→ Signif. codes:	0	***	0.001	**	0.01	*
		0.05	.	0.1	'	1

Correlation

A measure of the linear relationship between variables

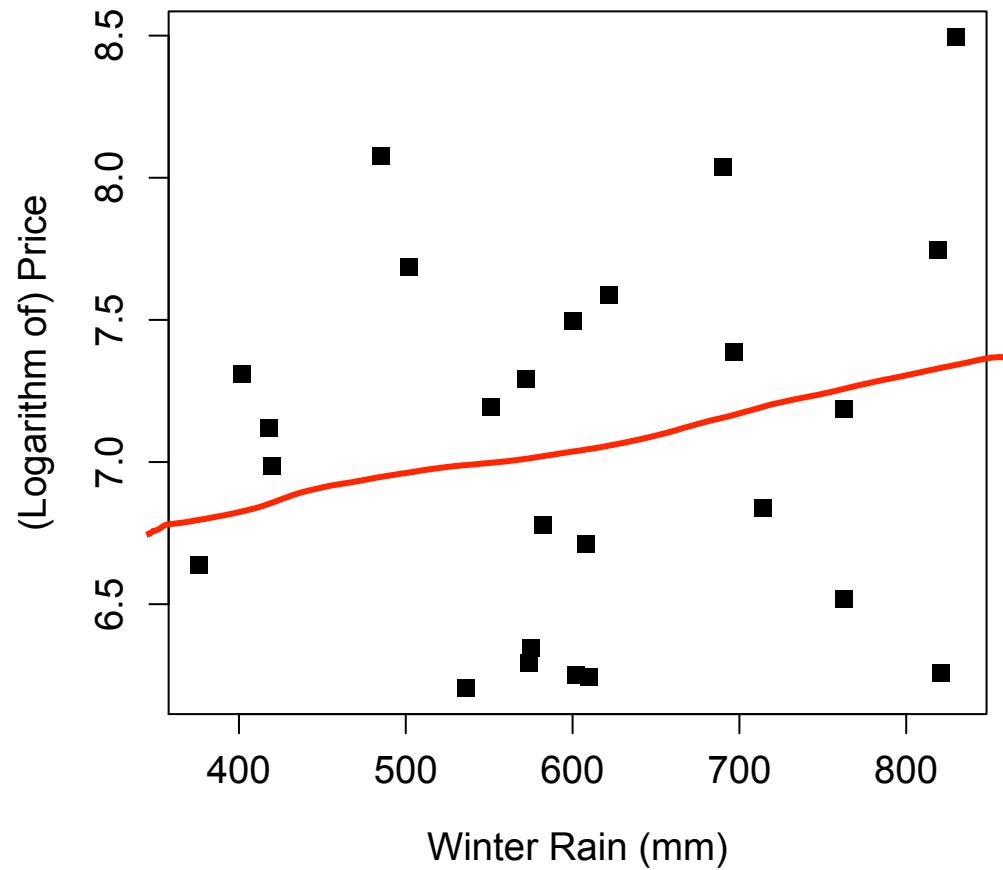
+1 = perfect positive linear relationship

0 = no linear relationship

-1 = perfect negative linear relationship

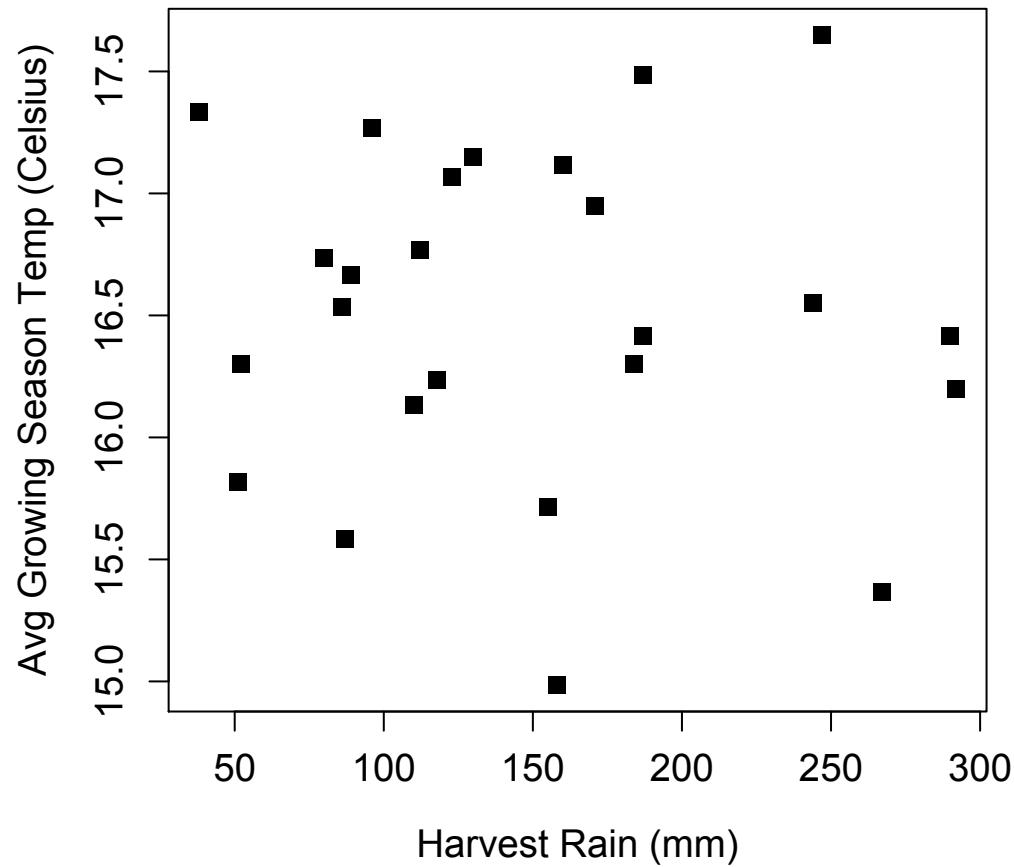
Examples of Correlation

cor
= 0.14



Examples of Correlation

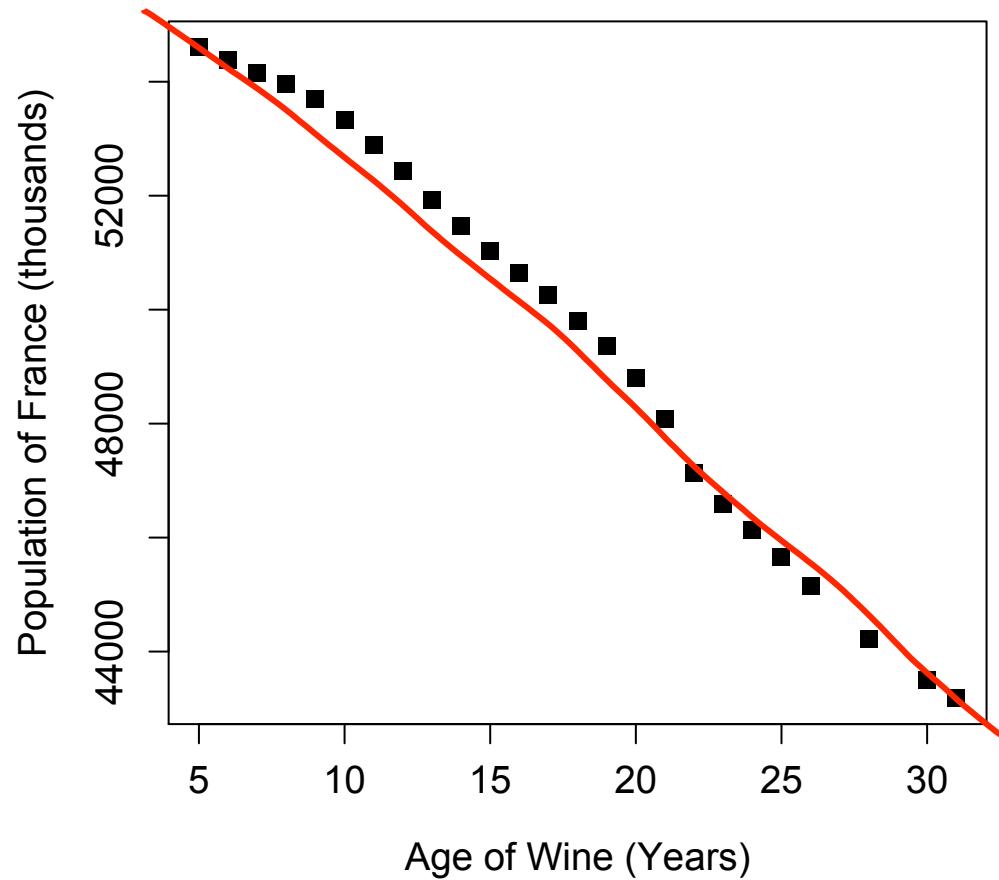
cor
 ≈ -0.06



Examples of Correlation



cor
= -0.99



Predictive Ability



- Our wine model had a value of $R^2 = \underline{0.83}$
- Tells us our accuracy on the data that we used to build the model
out-of-sample
- But how well does the model perform on new data?
 - Bordeaux wine buyers profit from being able to predict the quality of a wine years before it matures

Out-of-Sample R^2

Variables	Model R^2	Test R^2
Avg Growing Season Temp (AGST)	0.44	0.79
AGST, Harvest Rain	0.71	-0.08
AGST, Harvest Rain, Age	0.79	0.53
→ AGST, Harvest Rain, Age, Winter Rain	<u>0.83</u>	<u>0.79</u>
AGST, Harvest Rain, Age, Winter Rain, Population	0.83	0.76

- Better model R^2 does not necessarily mean better test set R^2
- Need more data to be conclusive
- Out-of-sample R^2 can be negative!

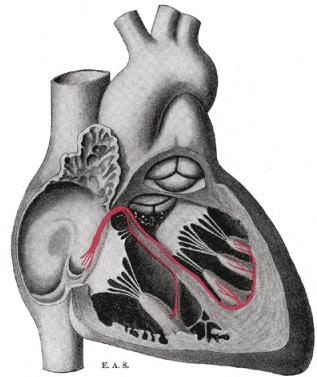
The Results

- **Parker:**
 - 1986 is “very good to sometimes exceptional”
- **Ashenfelter:**
 - 1986 is mediocre
 - 1989 will be “the wine of the century” and 1990 will be even better!
- In wine auctions,
 - 1989 sold for more than twice the price of 1986
 - 1990 sold for even higher prices!
- Later, Ashenfelter predicted 2000 and 2003 would be great
- Parker has stated that “2000 is the greatest vintage Bordeaux has ever produced”

The Analytics Edge



- A linear regression model with only a few variables can predict wine prices well
- In many cases, outperforms wine experts' opinions
- A quantitative approach to a traditionally qualitative problem



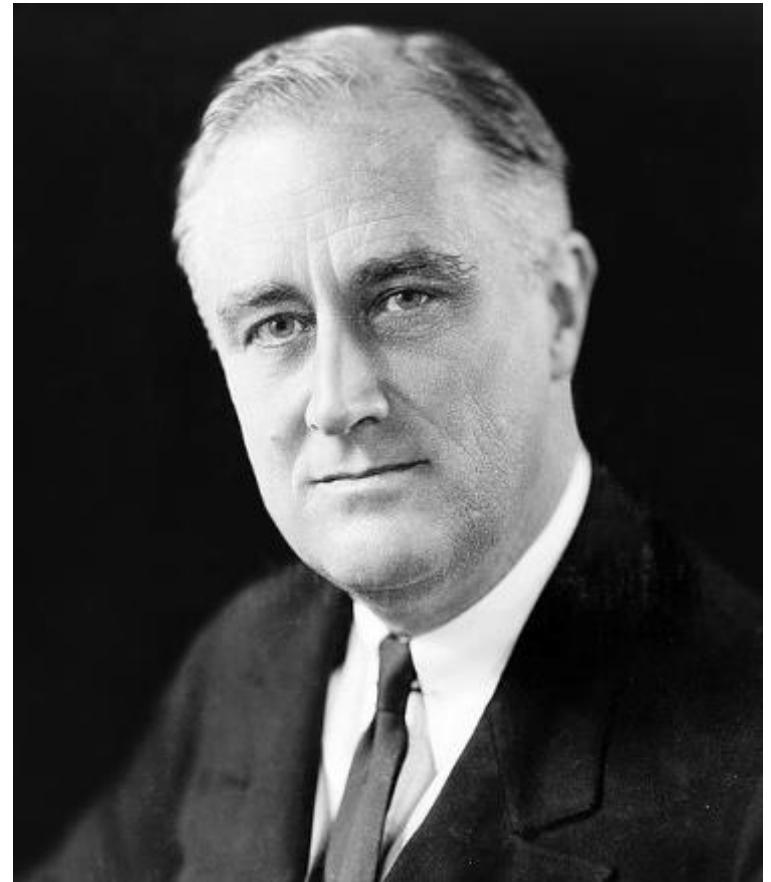
THE FRAMINGHAM HEART STUDY

Evaluating Risk Factors to Save Lives

15.071x – The Analytics Edge

Franklin Delano Roosevelt (FDR)

- President of the United States, 1933-1945
 - Longest-serving president
 - Led country through Great Depression
 - Commander in Chief of U.S. military in World War II
- Died while president, April 12, 1945



FDR's Blood Pressure

- Before presidency, blood pressure of 140/100
 - Healthy blood pressure is less than 120/80
 - Today, this is already considered high blood pressure
- One year before death, 210/120
 - Today, this is called Hypertensive Crisis, and emergency care is needed
 - FDR's personal physician:
“A moderate degree of arteriosclerosis, although no more than normal for a man of his age”
- Two months before death: 260/150
- Day of death: 300/190

Early Misconceptions

- High blood pressure dubbed *essential hypertension*
 - Considered important to force blood through arteries
 - Considered harmful to lower blood pressure
- Today, we know better

“Today, presidential blood pressure numbers like FDR’s would send the country’s leading doctors racing down hallways ... whisking the nation’s leader into the cardiac care unit of Bethesda Naval Hospital.”

-- Daniel Levy, Framingham Heart Study Director

How Did we Learn?



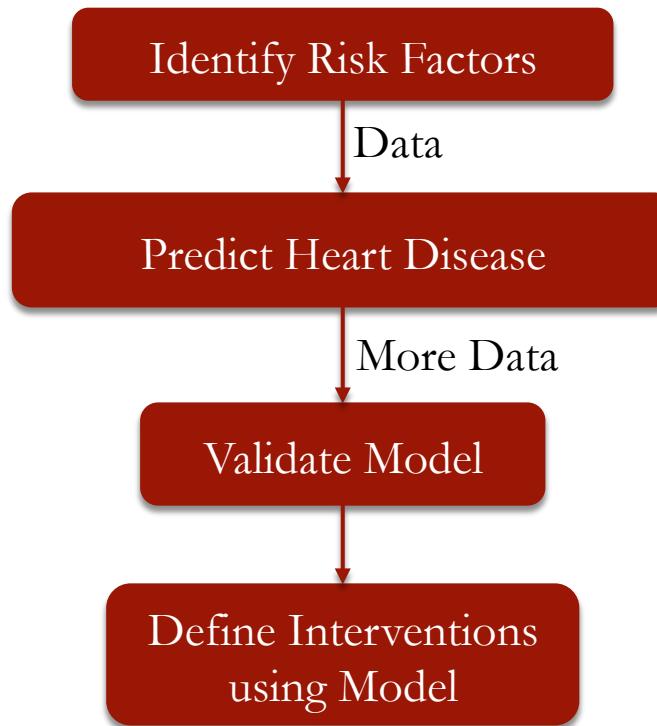
- In late 1940s, U.S. Government set out to better understand cardiovascular disease (CVD)
- Plan: track large cohort of initially healthy patients over time
- City of Framingham, MA selected as site for study
 - Appropriate size
 - Stable population
 - Cooperative doctors and residents
- 1948: beginning of Framingham Heart Study

The Framingham Heart Study



- 5,209 patients aged 30-59 enrolled
- Patients given questionnaire and exam every 2 years
 - Physical characteristics
 - Behavioral characteristics
 - Test results
- Exams and questions expanded over time
- We will build models using the Framingham data to predict and prevent heart disease

Analytics to Prevent Heart Disease



Coronary Heart Disease (CHD)



- We will predict 10-year risk of CHD
 - Subject of important 1998 paper, introducing the Framingham Risk Score
- CHD is a disease of the blood vessels supplying the heart
- Heart disease has been the leading cause of death worldwide since 1921
 - 7.3 million people died from CHD in 2008
 - Since 1950, age-adjusted death rates have declined 60%

Risk Factors

- *Risk factors* are variables that increase the chances of a disease
- Term coined by William Kannell and Roy Dawber from the Framingham Heart Study
- Key to successful prediction of CHD: identifying important risk factors

Hypothesized CHD Risk Factors



- We will investigate risk factors collected in the first data collection for the study
 - Anonymized version of original data
- Demographic risk factors
 - *male*: sex of patient
 - *age*: age in years at first examination
 - *education*: Some high school (1), high school/GED (2), some college/vocational school (3), college (4)

Hypothesized CHD Risk Factors



- Behavioral risk factors
 - *currentSmoker, cigsPerDay*: Smoking behavior
- Medical history risk factors
 - *BPmeds*: On blood pressure medication at time of first examination
 - *prevStroke*: Previously had a stroke
 - *prevHyp*: Currently hypertensive
 - *diabetes*: Currently has diabetes

Hypothesized CHD Risk Factors

- Risk factors from first examination
 - *totChol*: Total cholesterol (mg/dL)
 - *sysBP*: Systolic blood pressure
 - *diaBP*: Diastolic blood pressure
 - *BMI*: Body Mass Index, weight (kg)/height (m)²
 - *heartRate*: Heart rate (beats/minute)
 - *glucose*: Blood glucose level (mg/dL)

An Analytical Approach



- Randomly split patients into training and testing sets
- Use logistic regression on training set to predict whether or not a patient experienced CHD within 10 years of first examination
- Evaluate predictive power on test set

Model Strength



- Model rarely predicts 10-year CHD risk above 50%
 - Accuracy very near a baseline of always predicting no CHD
- Model can differentiate low-risk from high-risk patients ($AUC = 0.74$)
- Some significant variables suggest interventions
 - Smoking
 - Cholesterol
 - Systolic blood pressure
 - Glucose

Risk Model Validation



- So far, we have used *internal validation*
 - Train with some patients, test with others
- Weakness: unclear if model generalizes to other populations
- Framingham cohort white, middle class
- Important to test on other populations

Framingham Risk Model Validation

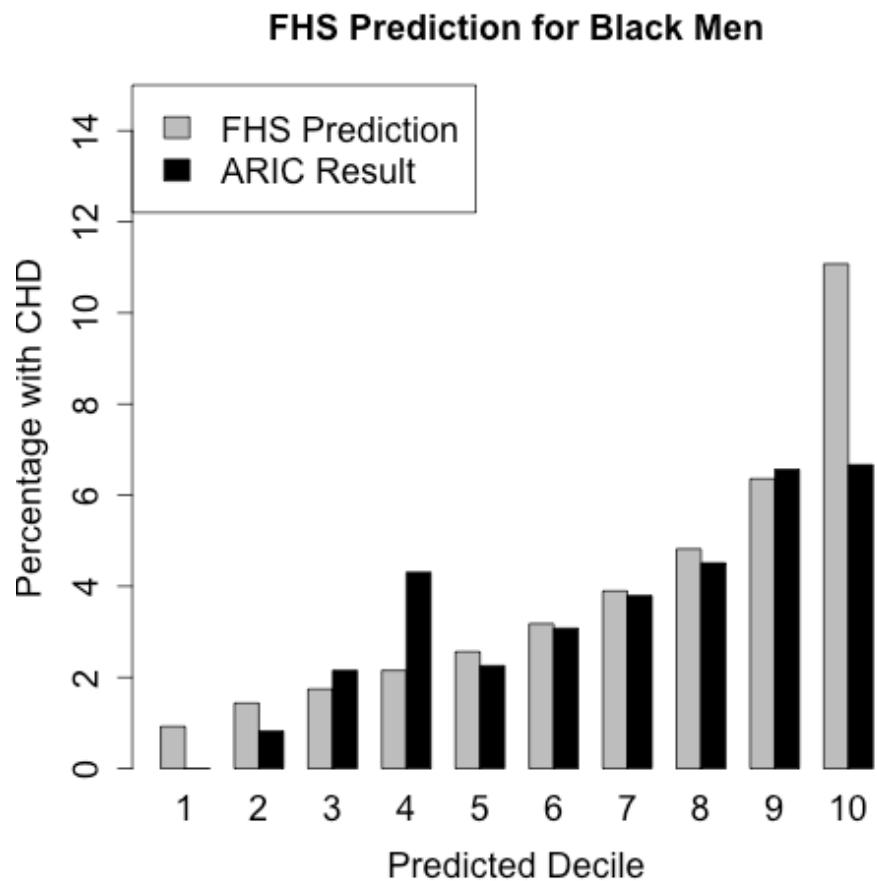
- Framingham Risk Model tested on diverse cohorts

Study	Population
Atherosclerosis Risk in Communities (ARIC) Study	White and Black
Honolulu Heart Program (HHP)	Japanese American
Puerto Rico Heart Health Program (PR)	Hispanic
Strong Heart Study (SHS)	Native American

- Cohort studies collecting same risk factors
- Validation Plan
 - Predict CHD risk for each patient using FHS model
 - Compare to actual outcomes for each risk decile

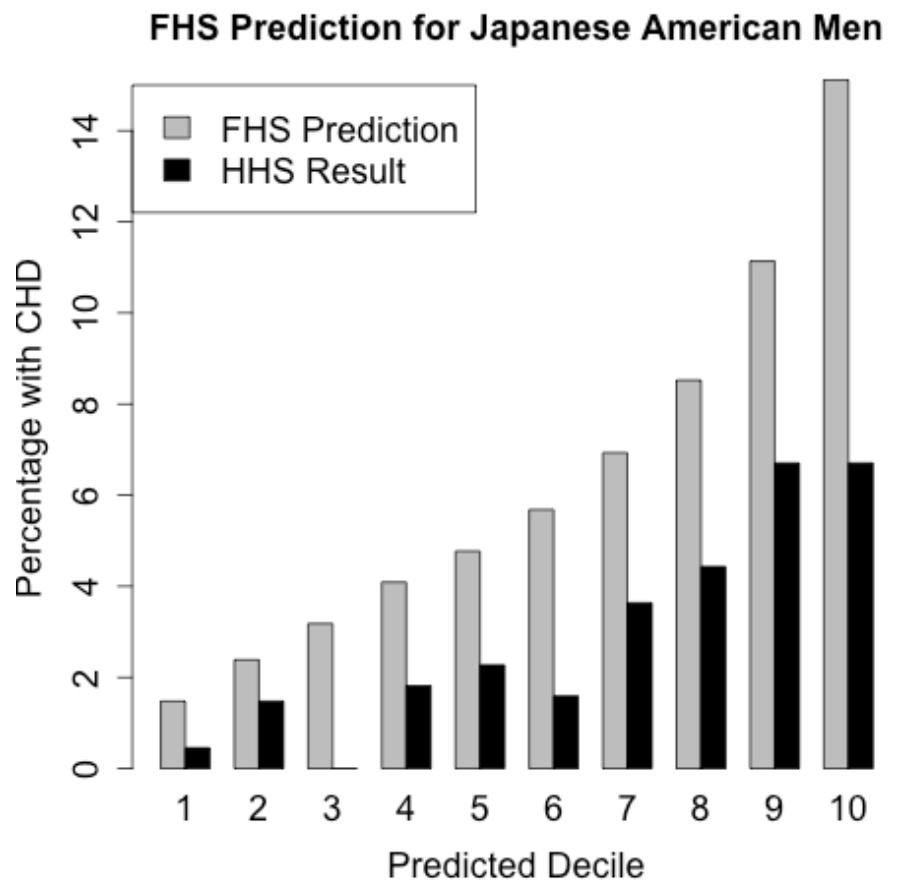
Validation for Black Men

- 1,428 black men in ARIC study
- Similar clinical characteristics, except higher diabetes rate
- Similar CHD rate
- Framingham risk model predictions accurate



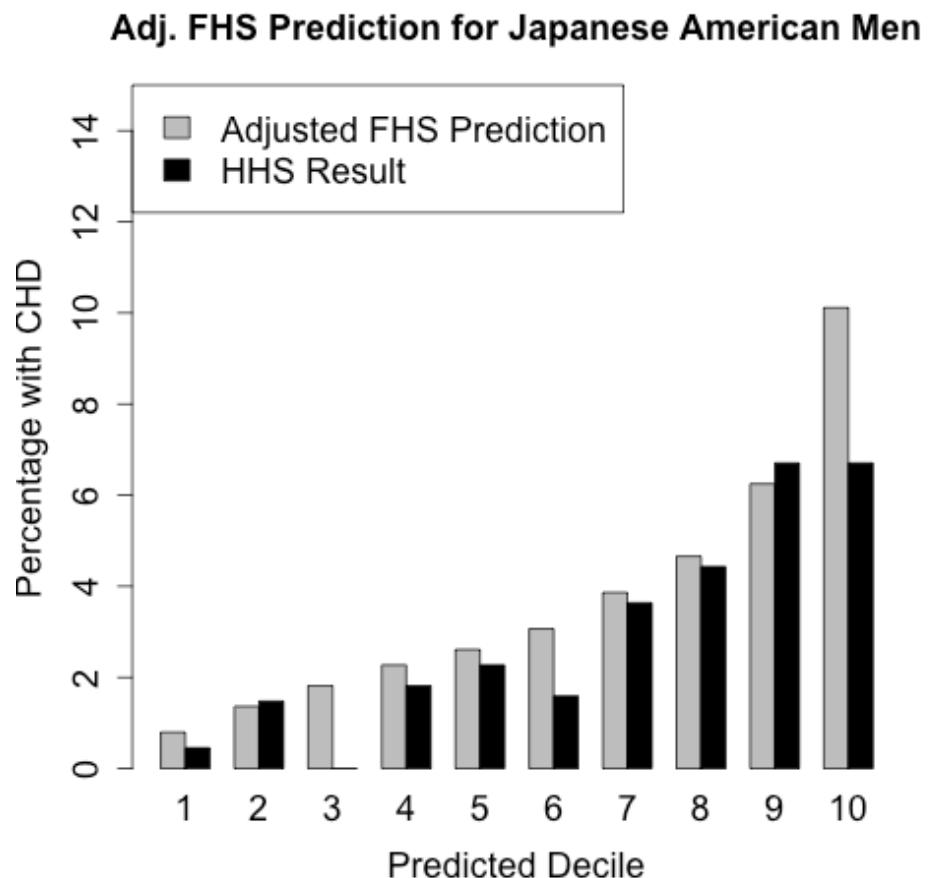
Validation for Japanese American Men

- 2,755 Japanese American men in HHS
- Lower CHD rate
- Framingham risk model systematically overpredicts CHD risk

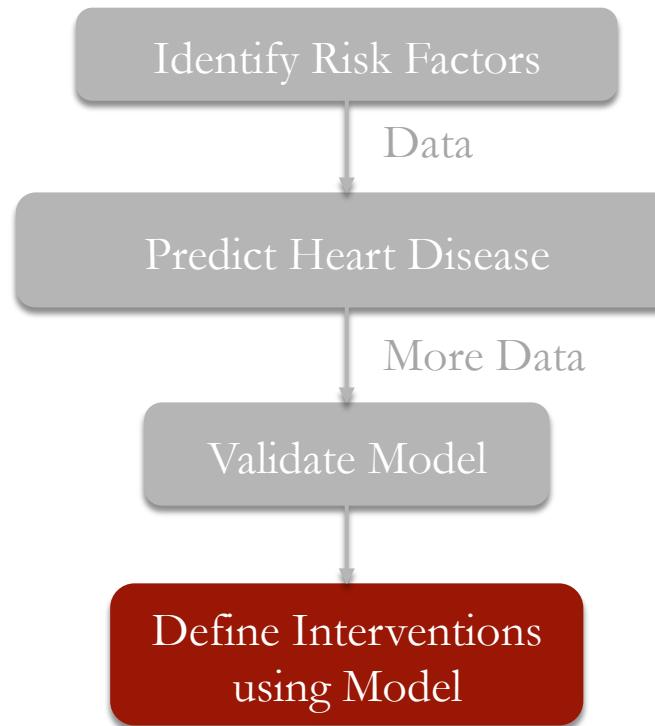


Recalibrated Model

- Recalibration adjusts model to new population
- Changes predicted risk, but does not reorder predictions
- More accurate risk estimates



Interventions



Drugs to Lower Blood Pressure

- In FDR's time, hypertension drugs too toxic for practical use
- In 1950s, the diuretic chlorothiazide was developed
- Framingham Heart Study gave Ed Freis the evidence needed to argue for testing effects of BP drugs
- Veterans Administration (VA) Trial: randomized, double blind clinical trial
- Found decreased risk of CHD
- Now, >\$1B market for diuretics worldwide

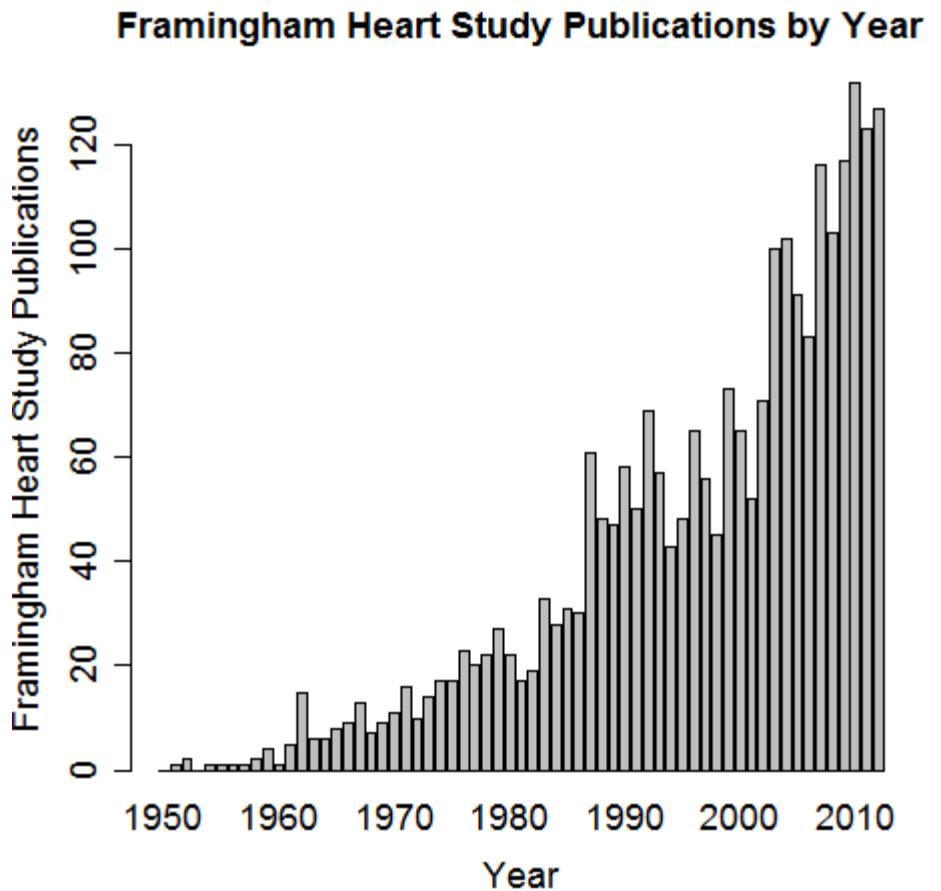
Drugs to Lower Cholesterol



- Despite Framingham results, early cholesterol drugs too toxic for practical use
- In 1970s, first statins were developed
- Study of 4,444 patients with CHD: statins cause 37% risk reduction of second heart attack
- Study of 6,595 men with high cholesterol: statins cause 32% risk reduction of CVD deaths
- Now, > \$20B market for statins worldwide

The Heart Study Through the Years

- More than 2,400 studies use Framingham data
- Many other risk factors evaluated
 - Obesity
 - Exercise
 - Psychosocial issues
 - ...
- *Texas Heart Institute Journal*: top 10 cardiology advances of 1900s



Available Online

Risk Assessment Tool for Estimating Your 10-year Risk of Having a Heart Attack

The risk assessment tool below uses information from the Framingham Heart Study to predict a person's chance of having a heart attack in the next 10 years. This tool is designed for adults aged 20 and older who do not have heart disease or diabetes. To find your risk score, enter your information in the calculator below.

Age:

years

Gender:

Female Male

Total Cholesterol:

mg/dL

HDL Cholesterol:

mg/dL

Smoker:

No Yes

Systolic Blood Pressure:

mm/Hg

Are you currently on any medication to treat high blood pressure?

No Yes

Calculate Your 10-Year Risk

 TOP

Total cholesterol - Total cholesterol is the sum of all the cholesterol in your blood. The higher your total cholesterol, the greater your risk for heart disease. Here are the total values that matter to you:

Less than 200 mg/dL 'Desirable' level that puts you at lower risk for heart disease. A cholesterol level of 200 mg/dL or greater increases your risk.

200 to 239 mg/dL 'Borderline-high.'

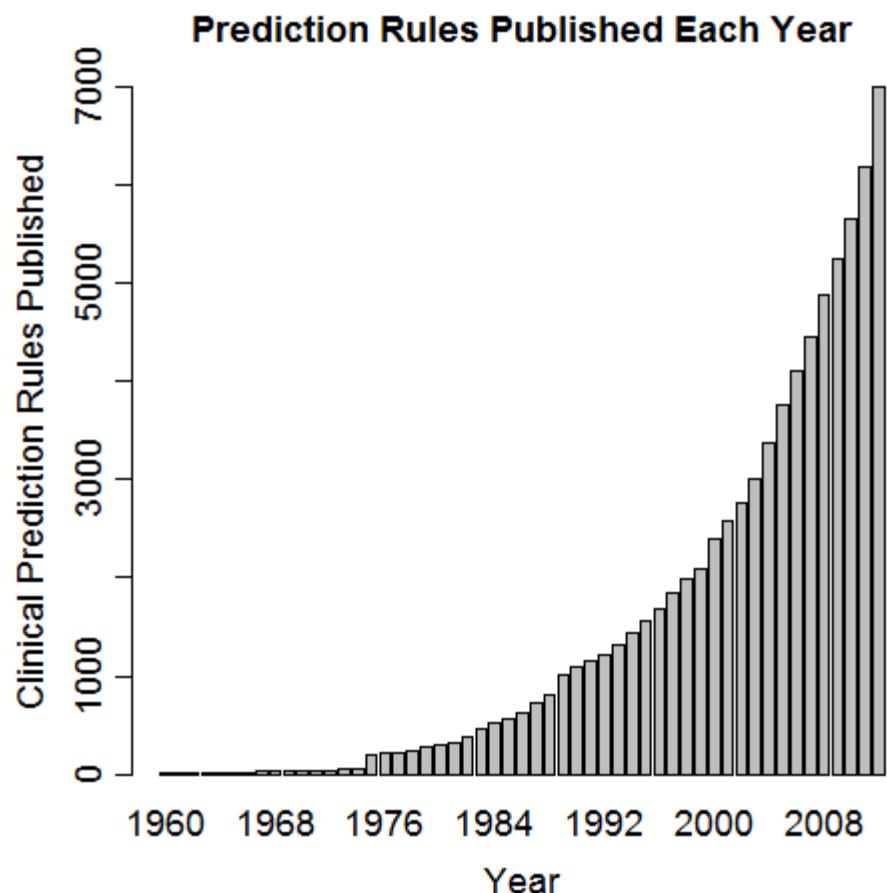
Research Directions and Challenges



- Second generation enrolled in 1971, third in 2002
 - Enables study of family history as a risk factor
- More diverse cohorts begun in 1994 and 2003
- Social network analysis of participants
- Genome-wide association study linking studying genetics as risk factors
- Many challenges related to funding
 - Funding cuts in 1969 nearly closed study
 - 2013 sequester threatening to close study

Clinical Decision Rules

- Paved the way for *clinical decision rules*
- Predict clinical outcomes with data
 - Patient and disease characteristics
 - Test results
- More than 75,000 published across medicine
- Rate increasing





MODELING THE EXPERT

An Introduction to Logistic Regression

15.071 – The Analytics Edge

Ask the Experts!



- Critical decisions are often made by people with expert knowledge
- Healthcare Quality Assessment
 - Good quality care educates patients and controls costs
 - Need to assess quality for proper medical interventions
 - No single set of guidelines for defining quality of healthcare
 - Health professionals are experts in quality of care assessment

Experts are Human



- Experts are limited by memory and time
- Healthcare Quality Assessment
 - Expert physicians can evaluate quality by examining a patient's records
 - This process is time consuming and inefficient
 - Physicians cannot assess quality for millions of patients

Replicating Expert Assessment



- Can we develop analytical tools that replicate expert assessment on a large scale?
- Learn from expert human judgment
 - Develop a model, interpret results, and adjust the model
- Make predictions/evaluations on a large scale
- Healthcare Quality Assessment
 - Let's identify poor healthcare quality using analytics

Claims Data

Medical Claims

Diagnosis, Procedures,
Doctor/Hospital, Cost

Pharmacy Claims

Drug, Quantity, Doctor,
Medication Cost

- Electronically available
- Standardized
- Not 100% accurate
- Under-reporting is common
- Claims for hospital visits can be vague

Creating the Dataset – Claims Samples

Claims Sample

- Large health insurance claims database
- Randomly selected 131 diabetes patients
- Ages range from 35 to 55
- Costs \$10,000 – \$20,000
- September 1, 2003 – August 31, 2005

Creating the Dataset – Expert Review

Claims Sample

Expert Review

- Expert physician reviewed claims and wrote descriptive notes:

“Ongoing use of narcotics”

“Only on Avandia, not a good first choice drug”

“Had regular visits, mammogram, and immunizations”

“Was given home testing supplies”

Creating the Dataset – Expert Assessment



Claims Sample

Expert Review

Expert Assessment

- Rated quality on a two-point scale (poor/good)

“I’d say **care was poor** – poorly treated diabetes”

“No eye care, but overall I’d say **high quality**”

Creating the Dataset – Variable Extraction



Claims Sample

Expert Review

Expert Assessment

Variable Extraction

- Dependent Variable
 - **Quality of care**
- Independent Variables
 - ongoing use of **narcotics**
 - **only on Avandia**, not a good first choice drug
 - Had **regular visits, mammogram, and immunizations**
 - Was given **home testing supplies**

Creating the Dataset – Variable Extraction



Claims Sample

Expert Review

Expert Assessment

Variable Extraction

- Dependent Variable
 - Quality of care
- Independent Variables
 - Diabetes treatment
 - Patient demographics
 - Healthcare utilization
 - Providers
 - Claims
 - Prescriptions

Predicting Quality of Care

- The dependent variable is modeled as a binary variable
 - 1 if low-quality care, 0 if high-quality care
- This is a *categorical variable*
 - A small number of possible outcomes
- Linear regression would predict a continuous outcome
- How can we extend the idea of linear regression to situations where the outcome variable is categorical?
 - Only want to predict 1 or 0
 - Could round outcome to 0 or 1
 - But we can do better with logistic regression

Logistic Regression

- Predicts the probability of poor care
 - Denote dependent variable “PoorCare” by y
 - $P(y = 1)$
- Then $P(y = 0) = 1 - P(y = 1)$
- Independent variables x_1, x_2, \dots, x_k
- Uses the Logistic Response Function

$$P(y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k)}}$$

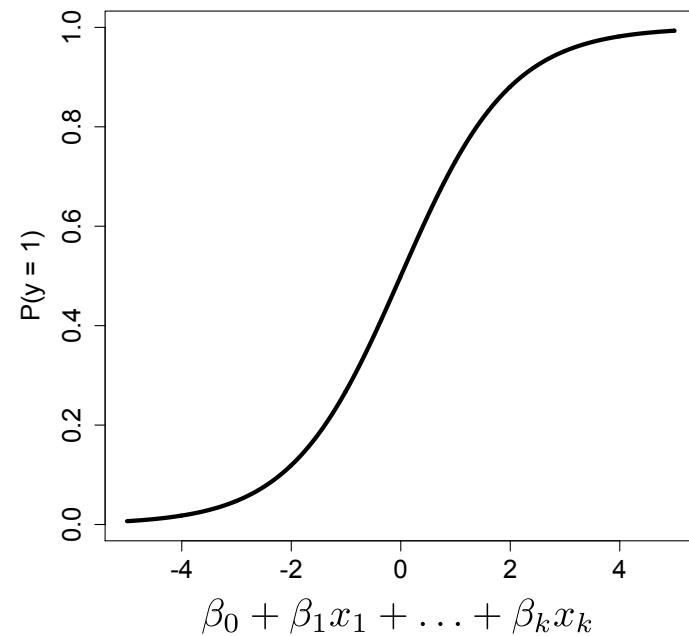
- Nonlinear transformation of linear regression equation to produce number between 0 and 1

Poor Care = 1
Good Care = 0

Understanding the Logistic Function

$$P(y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k)}}$$

- Positive values are predictive of class 1
- Negative values are predictive of class 0



Understanding the Logistic Function

$$P(y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k)}}$$

- The coefficients are selected to
 - Predict a high probability for the poor care cases
 - Predict a low probability for the good care cases

Understanding the Logistic Function

$$P(y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k)}}$$

- We can instead talk about Odds (like in gambling)

$$\text{Odds} = \frac{P(y = 1)}{P(y = 0)}$$

- Odds > 1 if $y = 1$ is more likely
- Odds < 1 if $y = 0$ is more likely

The Logit

- It turns out that

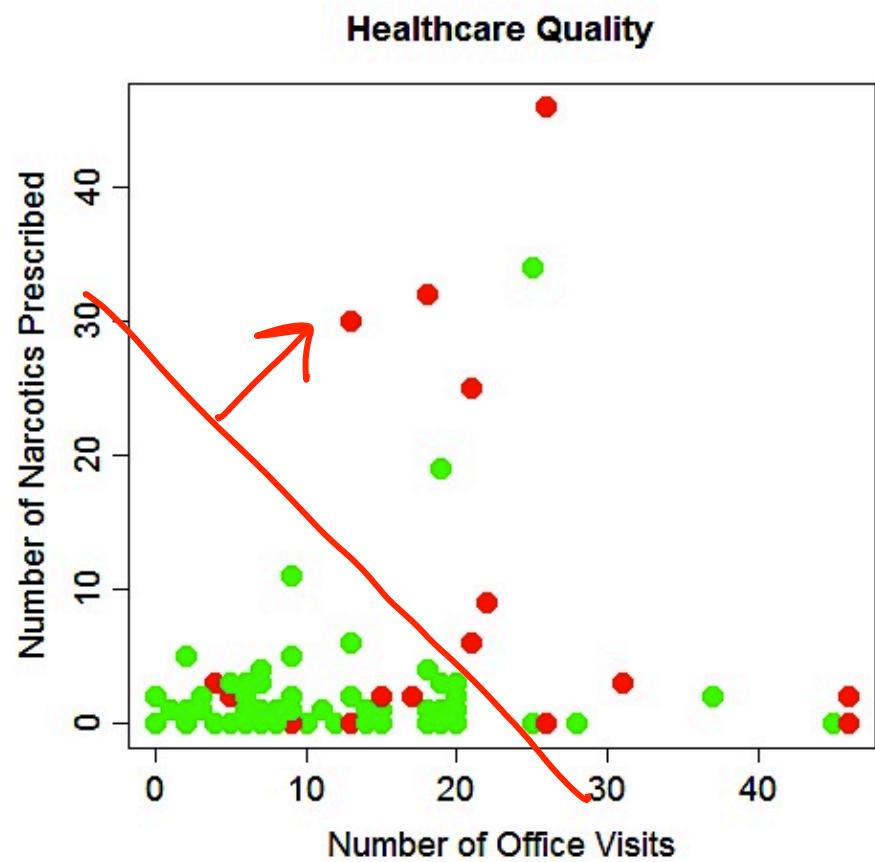
$$\text{Odds} = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k}$$

$$\log(\text{Odds}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

- This is called the “Logit” and looks like linear regression
- The bigger the Logit is, the bigger $P(y = 1)$

Model for Healthcare Quality

- Plot of the independent variables
 - Number of Office Visits
 - Number of Narcotics Prescribed
- Red are poor care
- Green are good care



Threshold Value

- The outcome of a logistic regression model is a probability
- Often, we want to make a binary prediction
 - Did this patient receive poor care or good care?
- We can do this using a *threshold value* t
- If $P(\text{PoorCare} = 1) \geq t$, predict poor quality
- If $P(\text{PoorCare} = 1) < t$, predict good quality
- What value should we pick for t ?

Threshold Value

- Often selected based on which errors are “better”
- If t is **large**, predict poor care rarely (when $P(y=1)$ is large)
 - More errors where we say good care, but it is actually poor care
 - Detects patients who are receiving the worst care
- If t is **small**, predict good care rarely (when $P(y=1)$ is small)
 - More errors where we say poor care, but it is actually good care
 - Detects all patients who might be receiving poor care
- With no preference between the errors, select $t = 0.5$
 - Predicts the more likely outcome

Selecting a Threshold Value

Compare actual outcomes to predicted outcomes using a *confusion matrix (classification matrix)*

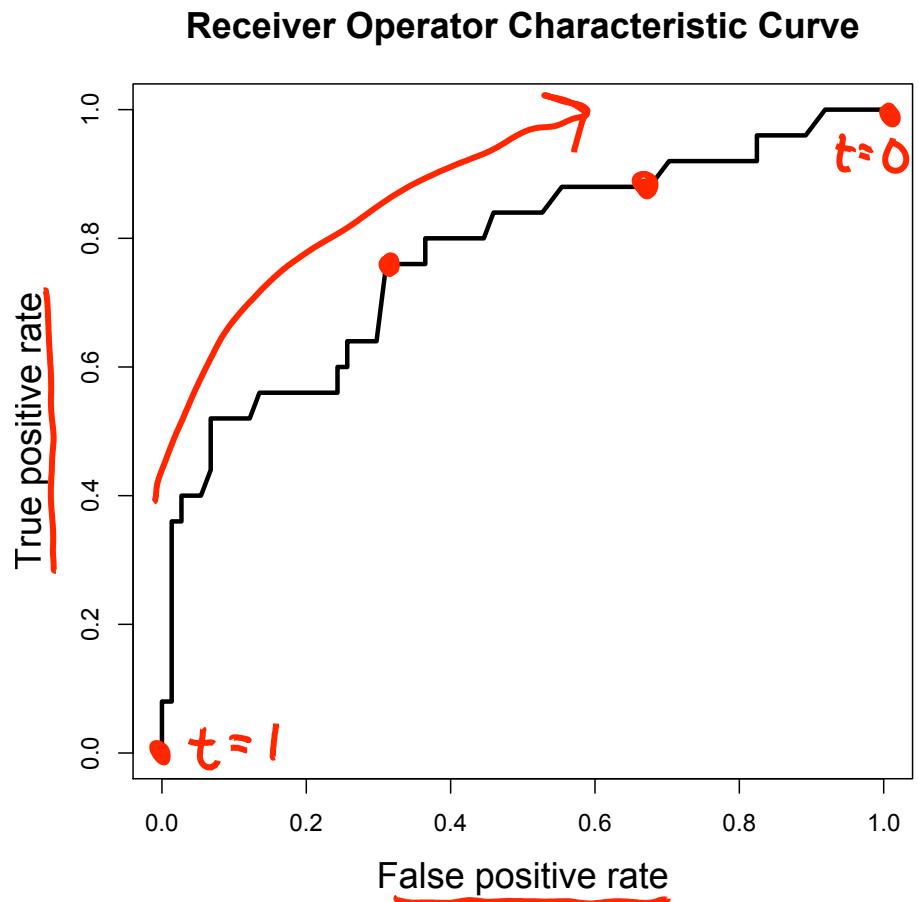
	Predicted = 0	Predicted = 1
Actual = 0	True Negatives (TN)	False Positives (FP)
Actual = 1	False Negatives (FN)	True Positives (TP)

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

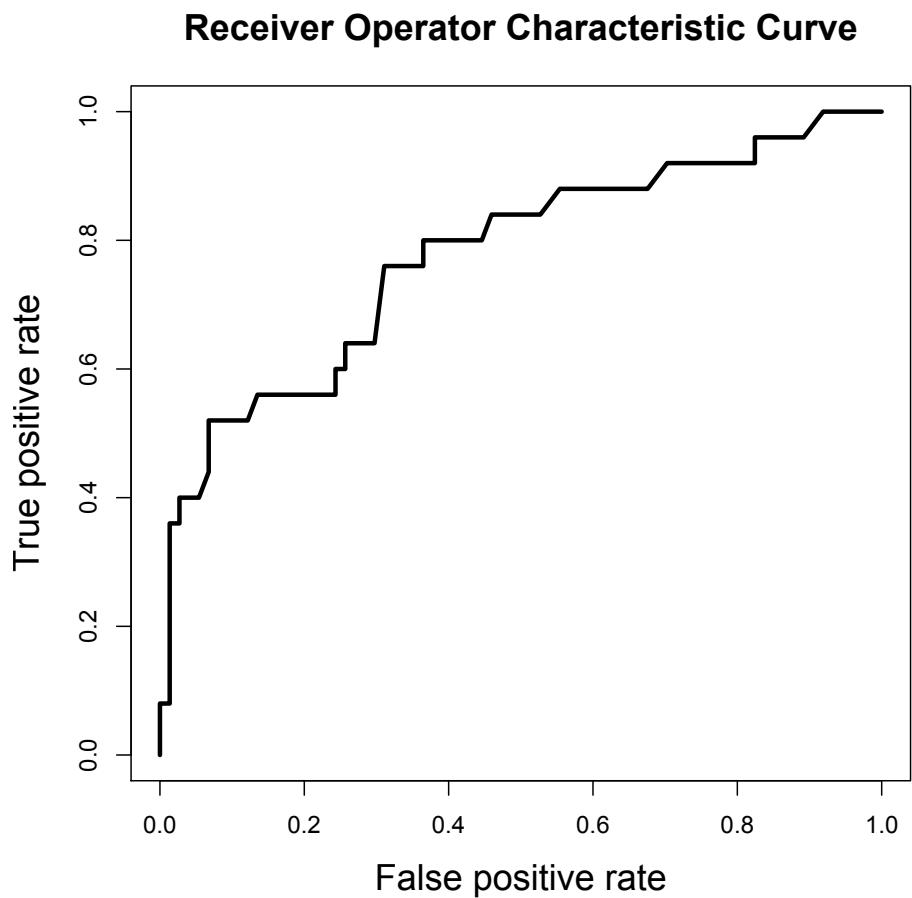
Receiver Operator Characteristic (ROC) Curve

- True positive rate (sensitivity) on y-axis
 - Proportion of poor care caught
- False positive rate ($1 - \text{specificity}$) on x-axis
 - Proportion of good care labeled as poor care



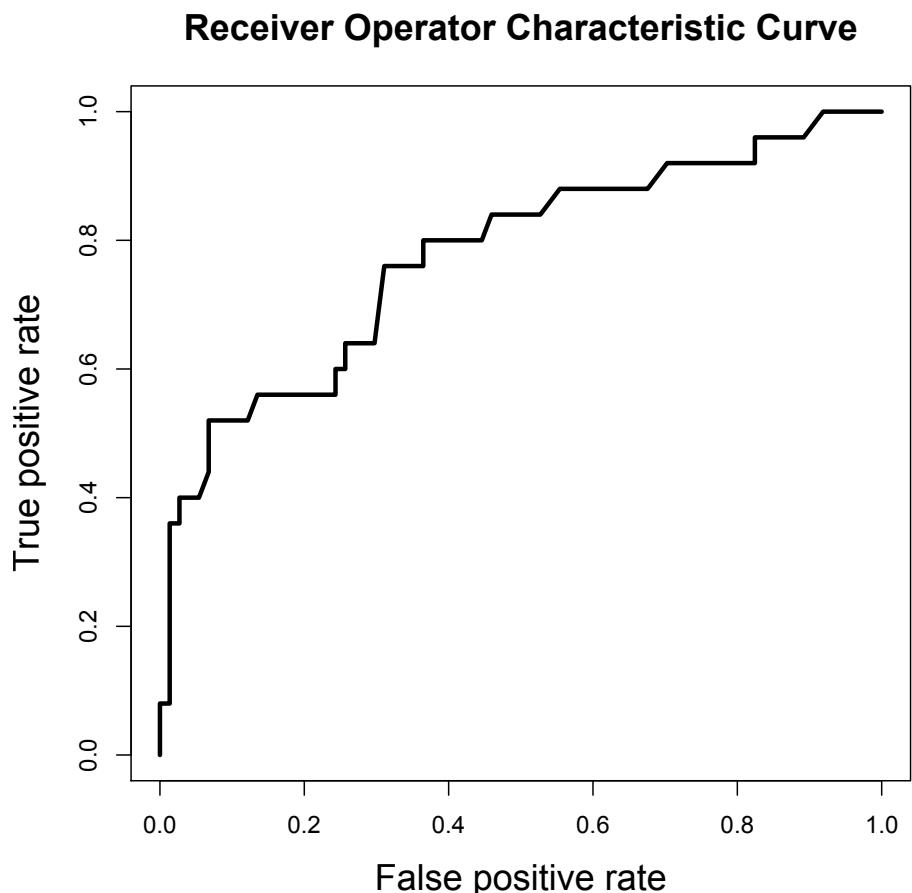
Selecting a Threshold using ROC

- Captures all thresholds simultaneously
- **High threshold**
 - High specificity
 - Low sensitivity
- **Low Threshold**
 - Low specificity
 - High sensitivity



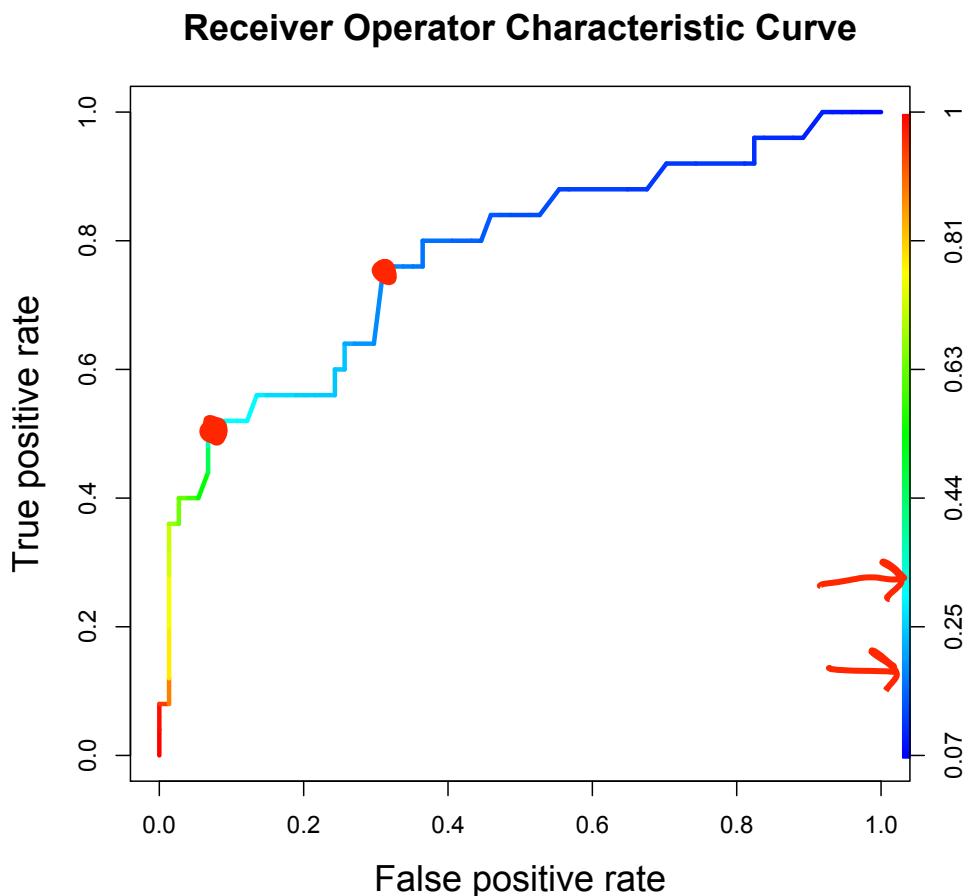
Selecting a Threshold using ROC

- Choose **best threshold** for **best trade off**
 - cost of failing to detect positives
 - costs of raising false alarms



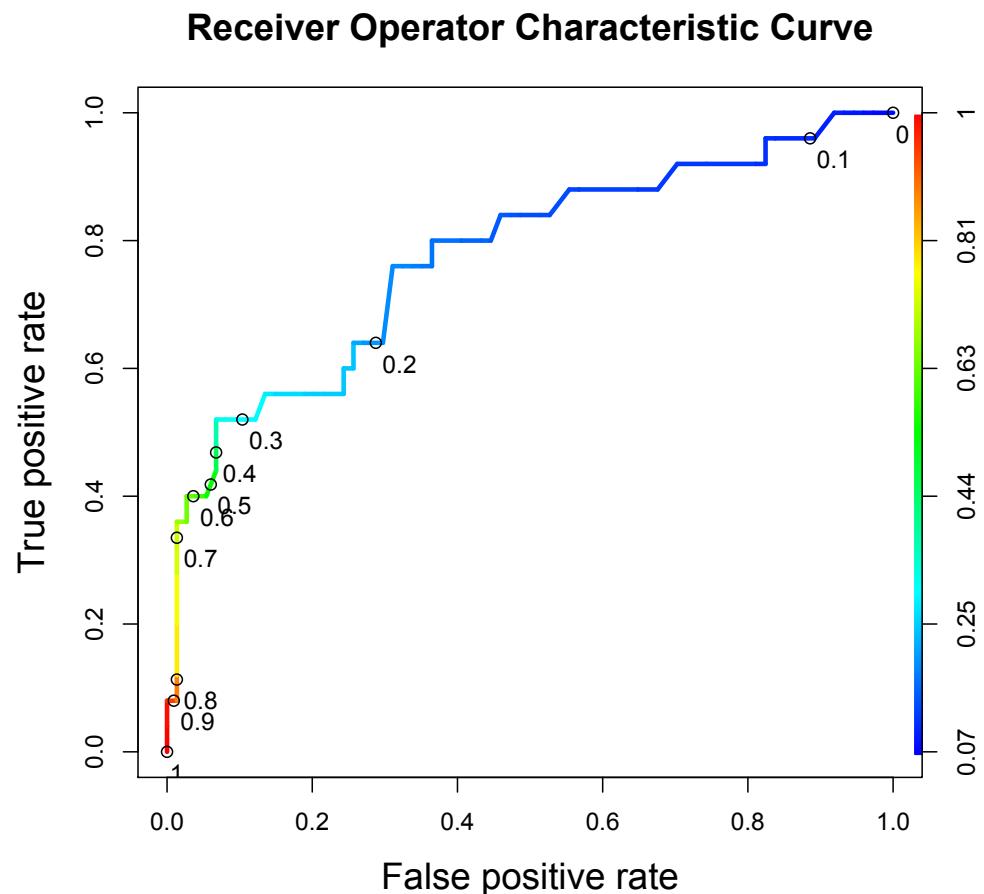
Selecting a Threshold using ROC

- Choose **best threshold** for **best trade off**
 - cost of failing to detect positives
 - costs of raising false alarms



Selecting a Threshold using ROC

- Choose **best threshold** for **best trade off**
 - cost of failing to detect positives
 - costs of raising false alarms



Interpreting the Model



- Multicollinearity could be a problem
 - Do the coefficients make sense?
 - Check correlations
- Measures of accuracy

Compute Outcome Measures

Confusion Matrix:

	Predicted Class = 0	Predicted Class = 1
Actual Class = 0	True Negatives (TN)	False Positives (FP)
Actual Class = 1	False Negatives (FN)	True Positives (TP)

N = number of observations

$$\text{Overall accuracy} = (\text{TN} + \text{TP})/\text{N}$$

$$\text{Overall error rate} = (\text{FP} + \text{FN})/\text{N}$$

$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN})$$

$$\text{False Negative Error Rate} = \text{FN}/(\text{TP} + \text{FN})$$

$$\text{Specificity} = \text{TN}/(\text{TN} + \text{FP})$$

$$\text{False Positive Error Rate} = \text{FP}/(\text{TN} + \text{FP})$$

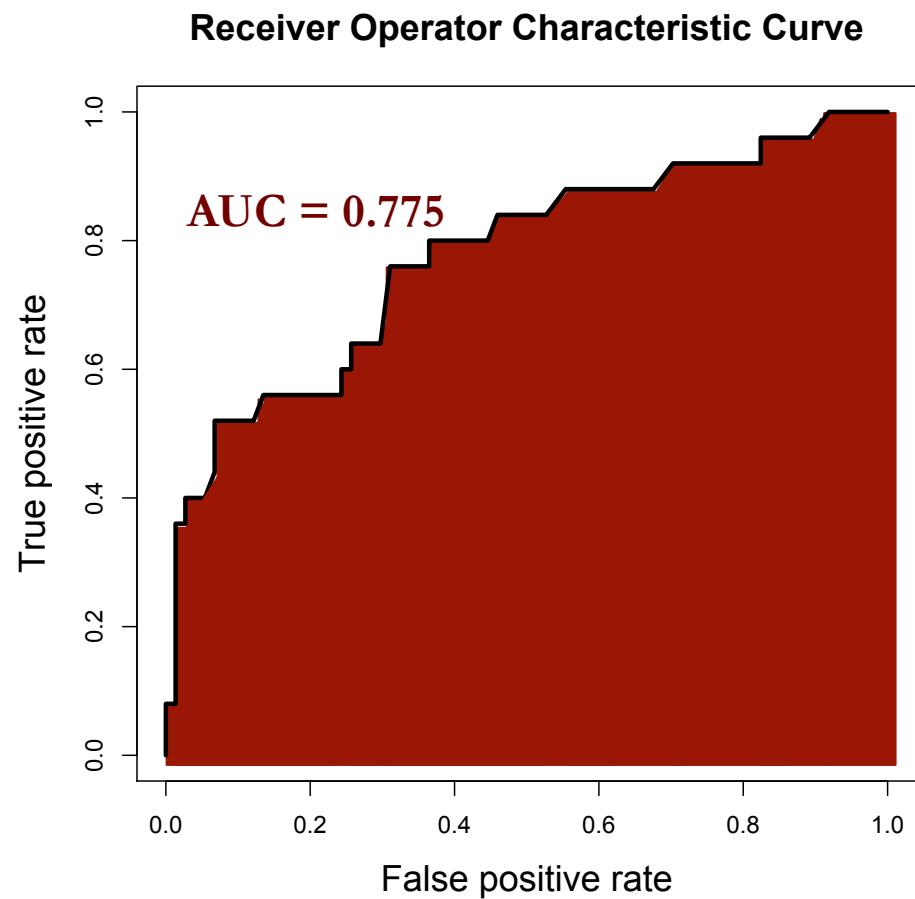
Making Predictions

- Just like in linear regression, we want to make predictions on a test set to compute out-of-sample metrics
> `predictTest = predict(QualityLog, type="response", newdata=qualityTest)`
- This makes predictions for probabilities
- If we use a threshold value of 0.3, we get the following confusion matrix

	Predicted Good Care	Predicted Poor Care
Actually Good Care	19	5
Actually Poor Care	2	6

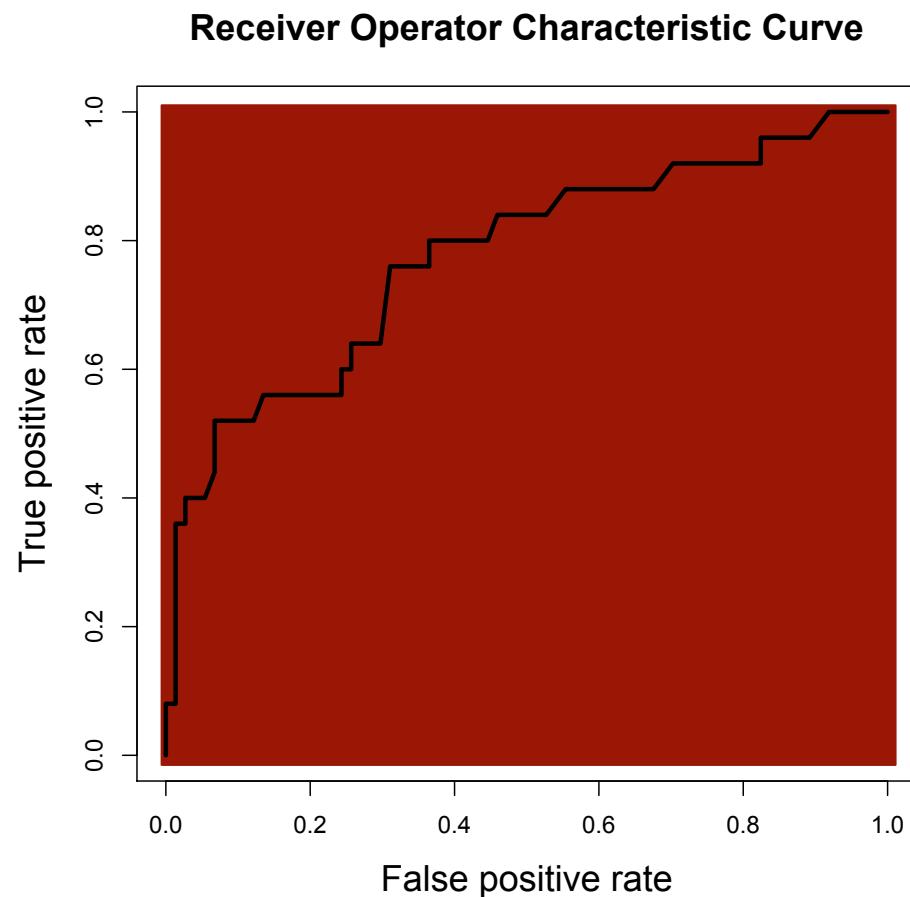
Area Under the ROC Curve (AUC)

- Just take the area under the curve
- Interpretation
 - Given a random positive and negative, proportion of the time you guess which is which correctly
- Less affected by sample balance than accuracy



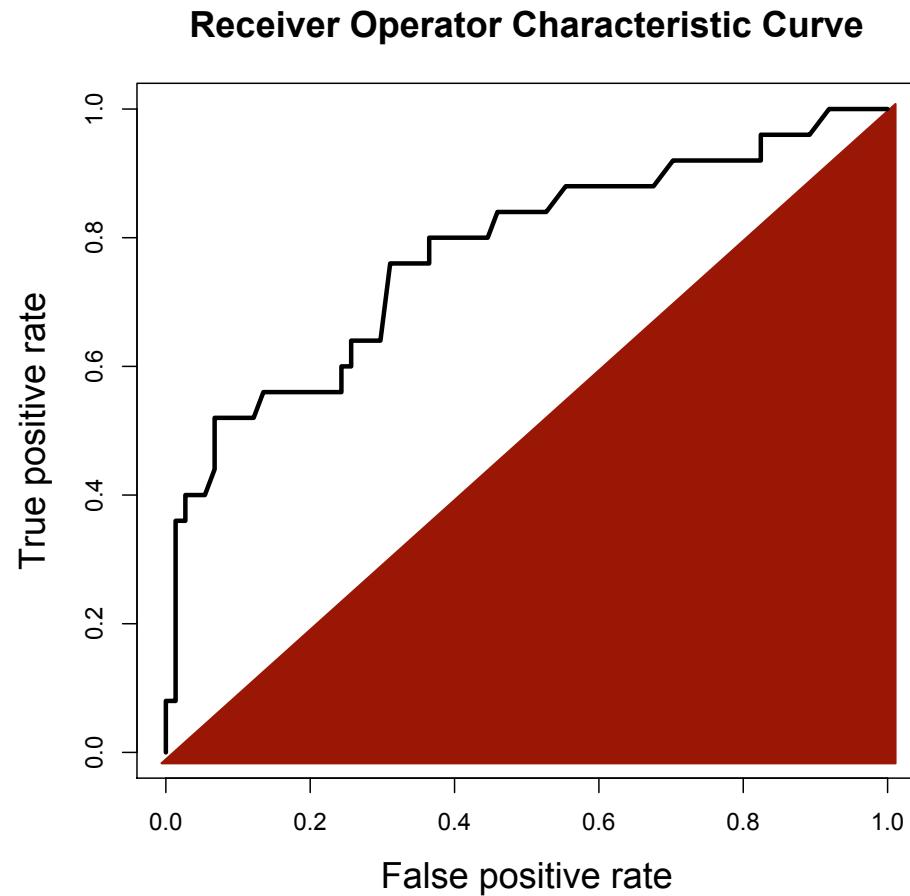
Area Under the ROC Curve (AUC)

- What is a good AUC?
 - Maximum of 1
(perfect prediction)



Area Under the ROC Curve (AUC)

- What is a good AUC?
 - Maximum of 1
(perfect prediction)
 - Minimum of 0.5
(just guessing)



Conclusions



- An expert-trained model can accurately identify diabetics receiving low-quality care
 - Out-of-sample accuracy of 78%
 - Identifies most patients receiving poor care
- In practice, the probabilities returned by the logistic regression model can be used to prioritize patients for intervention
- Electronic medical records could be used in the future

The Competitive Edge of Models



- While humans can accurately analyze small amounts of information, models allow larger scalability
- Models do not replace expert judgment
 - Experts can improve and refine the model
- Models can integrate assessments of many experts into one final unbiased and unemotional prediction

Claims Data

Medical Claims

Diagnosis, Procedures,
Doctor/Hospital, Cost

Pharmacy Claims

Drug, Quantity, Doctor,
Medication Cost

- Electronically available
- Standardized
- Not 100% accurate
- Under-reporting is common
- Claims for hospital visits can be vague

Creating the Dataset – Claims Samples

Claims Sample

- Large health insurance claims database
- Randomly selected 131 diabetes patients
- Ages range from 35 to 55
- Costs \$10,000 – \$20,000
- September 1, 2003 – August 31, 2005

Creating the Dataset – Expert Review

Claims Sample

Expert Review

- Expert physician reviewed claims and wrote descriptive notes:

“Ongoing use of narcotics”

“Only on Avandia, not a good first choice drug”

“Had regular visits, mammogram, and immunizations”

“Was given home testing supplies”

Creating the Dataset – Expert Assessment



Claims Sample

Expert Review

Expert Assessment

- Rated quality on a two-point scale (poor/good)

“I’d say **care was poor** – poorly treated diabetes”

“No eye care, but overall I’d say **high quality**”

Creating the Dataset – Variable Extraction



Claims Sample

Expert Review

Expert Assessment

Variable Extraction

- Dependent Variable
 - **Quality of care**
- Independent Variables
 - ongoing use of **narcotics**
 - **only on Avandia**, not a good first choice drug
 - Had **regular visits, mammogram, and immunizations**
 - Was given **home testing supplies**

Creating the Dataset – Variable Extraction



Claims Sample

Expert Review

Expert Assessment

Variable Extraction

- Dependent Variable
 - Quality of care
- Independent Variables
 - Diabetes treatment
 - Patient demographics
 - Healthcare utilization
 - Providers
 - Claims
 - Prescriptions

Predicting Quality of Care

- The dependent variable is modeled as a binary variable
 - 1 if low-quality care, 0 if high-quality care
- This is a *categorical variable*
 - A small number of possible outcomes
- Linear regression would predict a continuous outcome
- How can we extend the idea of linear regression to situations where the outcome variable is categorical?
 - Only want to predict 1 or 0
 - Could round outcome to 0 or 1
 - But we can do better with logistic regression

Logistic Regression

- Predicts the probability of poor care
 - Denote dependent variable “PoorCare” by y
 - $P(y = 1)$
- Then $P(y = 0) = 1 - P(y = 1)$
- Independent variables x_1, x_2, \dots, x_k
- Uses the Logistic Response Function

$$P(y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k)}}$$

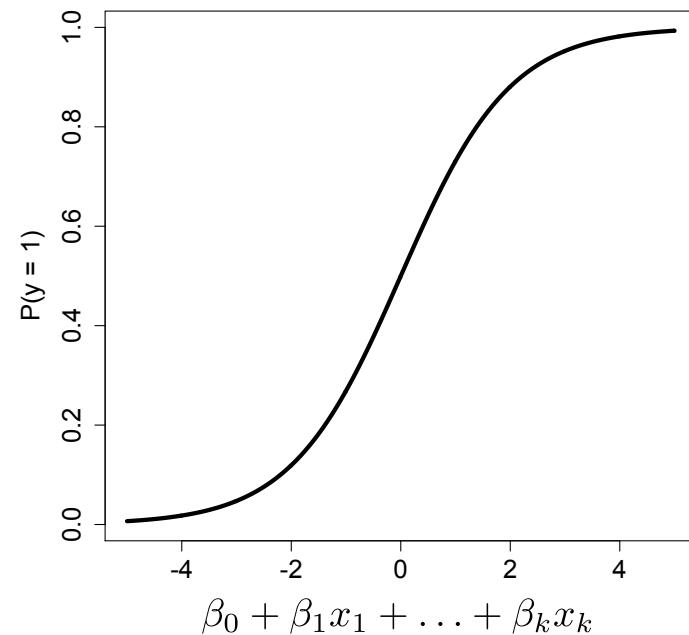
- Nonlinear transformation of linear regression equation to produce number between 0 and 1

Poor Care = 1
Good Care = 0

Understanding the Logistic Function

$$P(y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k)}}$$

- Positive values are predictive of class 1
- Negative values are predictive of class 0



Understanding the Logistic Function

$$P(y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k)}}$$

- The coefficients are selected to
 - Predict a high probability for the poor care cases
 - Predict a low probability for the good care cases

Understanding the Logistic Function

$$P(y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k)}}$$

- We can instead talk about Odds (like in gambling)

$$\text{Odds} = \frac{P(y = 1)}{P(y = 0)}$$

- Odds > 1 if $y = 1$ is more likely
- Odds < 1 if $y = 0$ is more likely

The Logit

- It turns out that

$$\text{Odds} = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k}$$

$$\log(\text{Odds}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

- This is called the “Logit” and looks like linear regression
- The bigger the Logit is, the bigger $P(y = 1)$



ELECTION FORECASTING

Predicting the Winner Before any Votes are Cast

15.071 – The Analytics Edge

United States Presidential Elections

- A president is elected every four years
- Generally, only two competitive candidates
 - Republican
 - Democratic

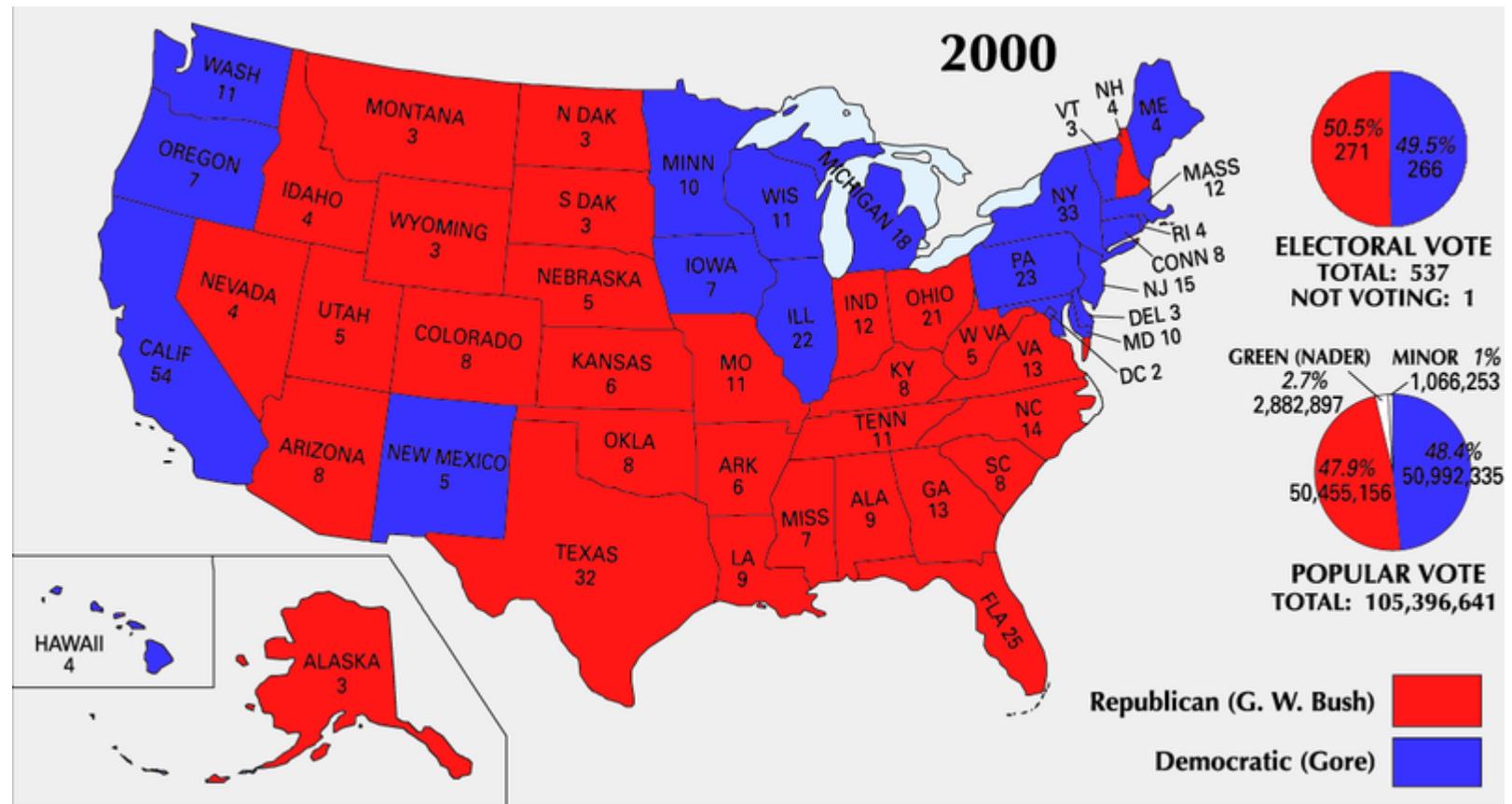


The Electoral College



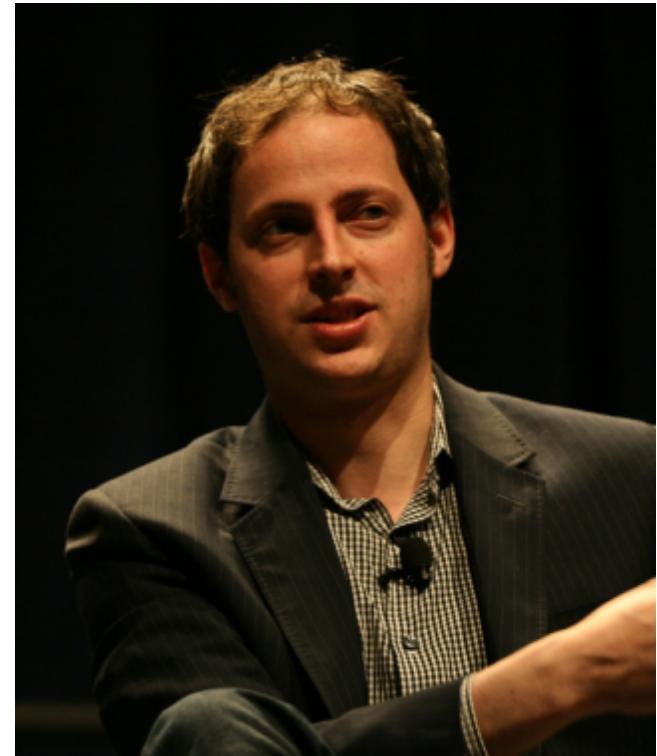
- The United States have 50 states
- Each assigned a number of *electoral votes* based on population
 - Most votes: 55 (California)
 - Least votes: 3 (multiple states)
 - Reassigned periodically based on population change
- Winner takes all: candidate with the most votes in a state gets all its electoral votes
- Candidate with most electoral votes wins election

2000 Election: Bush vs. Gore



Election Prediction

- Goal: Use polling data to predict state winners
- Then-*New York Times* columnist Nate Silver famously took on this task for the 2012 election



The Dataset

- Data from RealClearPolitics.com
- Instances represent a state in a given election
 - *State*: Name of state
 - *Year*: Election year (2004, 2008, 2012)
- Dependent variable
 - *Republican*: 1 if Republican won state, 0 if Democrat won
- Independent variables
 - *Rasmussen, SurveyUSA*: Polled R% - Polled D%
 - *DiffCount*: Polls with R winner – Polls with D winner
 - *PropR*: Polls with R winner / # polls



ELECTION FORECASTING

Predicting the Winner Before any Votes are Cast

15.071 – The Analytics Edge

Simple Approaches to Missing Data

- Delete the missing observations
 - We would be throwing away more than 50% of the data
 - We want to predict for all states
- Delete variables with missing values
 - We want to retain data from Rasmussen/SurveyUSA
- Fill missing data points with average values
 - The average value for a poll will be close to 0 (tie between Democrat and Republican)
 - If other polls in a state favor one candidate, the missing one probably would have, too

Multiple Imputation

- Fill in missing values based on non-missing values
 - If Rasmussen is very negative, then a missing SurveyUSA value will likely be negative
 - Just like *sample.split*, results will differ between runs unless you fix the random seed
- Although the method is complicated, we can use it easily through R's libraries
- We will use Multiple Imputation by Chained Equations (mice) package



KEEPING AN EYE ON HEALTHCARE COSTS

The D2Hawkeye Story

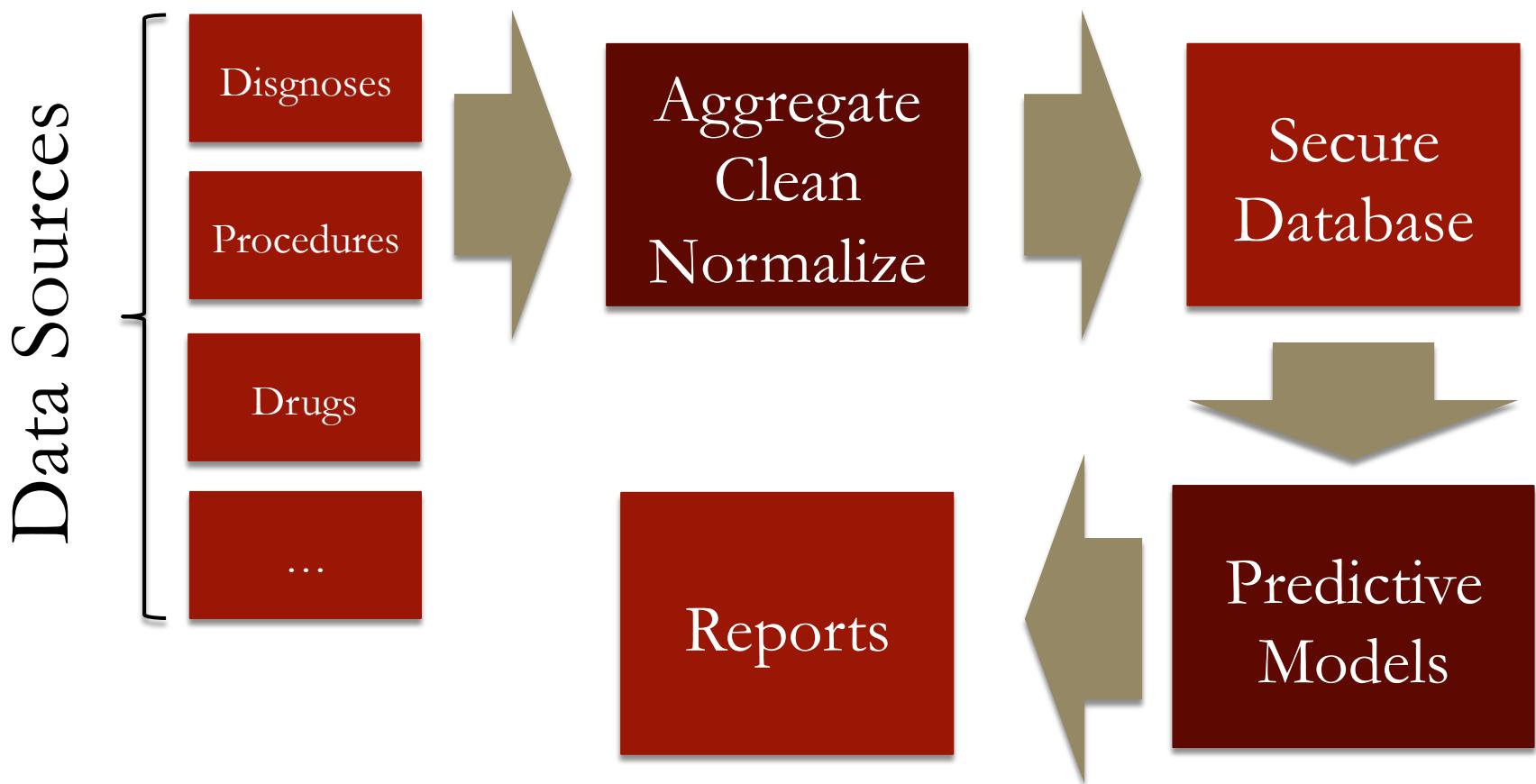


D2Hawkeye



- Founded by Chris Kryder, MD, MBA in 2001
- Combine expert knowledge and databases with analytics to improve quality and cost management in healthcare
- Located in Massachusetts USA, grew very fast and was sold to Verisk Analytics in 2009

D2Hawkeye



Healthcare Case Management

- D2Hawkeye tries to improve healthcare case management
 - Identify high-risk patients
 - Work with patients to manage treatment and associated costs
 - Arrange specialist care
- Medical costs often relate to severity of health problems, and are an issue for both patient and provider
- Goal: improve the quality of cost predictions

Impact



- Many different types of clients
 - Third party administrators of medical claims
 - Case management companies
 - Benefit consultants
 - Health plans
- **Millions of people** analyzed monthly through analytic platform in 2009
- **Thousands of employers** processed monthly

Pre-Analytics Approach

- Human judgment – MDs manually analyzed patient histories and developed
- Limited data sets
- Costly and inefficient
- Can we use analytics instead?



Data Sources

- Healthcare industry is data-rich, but data may be hard to access
 - Unstructured – doctor's notes
 - Unavailable – hard to get due to differences in technology
 - Inaccessible – strong privacy laws around healthcare data sharing
- What is available?

Data Sources



- Claims data
 - Requests for reimbursement submitted to insurance companies or state-provided insurance from doctors, hospitals and pharmacies.
- Eligibility information
- Demographic information

Claims Data

ClaimType	ProviderName	DiagCode	DiagDesc	SourceDiagCode	SourceDiagDesc	ProcNDCCode	ProcNDCDesc	ServiceDate	PaidAmount
DEN	SOUTHEASTERN MINNESOTA ORAL & MAX	DD0238	Dental Diseases	5206	Unspecified Anomaly of Tooth Position	DD007	Anesthesia - General	4/22/2005	\$ -
DEN	ASSOCIATED ORAL & MAXILLOFACIAL SURGEONS PA	DD0238	Dental Diseases	5206	Disturbances in ToOther Eruption	DD025	Dental	7/8/2005	\$ 272.68
DEN	CENTRAL FLORIDA ORAL SURGERY	DD0238	Dental Diseases	5206	Disturbances in ToOther Eruption	DD025	Dental	11/11/2005	\$ 568.13
Med	ALPHARETTA INTERNA	DD0004	ENT and Upper Resp Disorders	4610	Acute Maxillary Sinusitis	DD147	Office Visit - Established Patient	5/26/2005	\$ 125.85
Med	CUMMING FAMILY MEDICINE	DD0170	Neurotic and Personality Disorders	30000	Neurotic Disorders- 30000	DD149	Office Visit - New Patient	6/20/2005	\$ -
Med	ATLANTA WOMENS HEALTH GROUP- 582483738.20	DD0102	Screening	V776	Special Screening for Cystic Fibrosis	DD077	Lab - Blood Tests	7/29/2005	\$ 1.52

Claims Data



- Rich, structured data source
- Very high dimension
- Doesn't capture all aspects of a persons treatment or health – many things must be inferred
- Unlike electronic medical records, we do not know the results of a test, only that a test was administered

D2Hawkeye's Claims Data



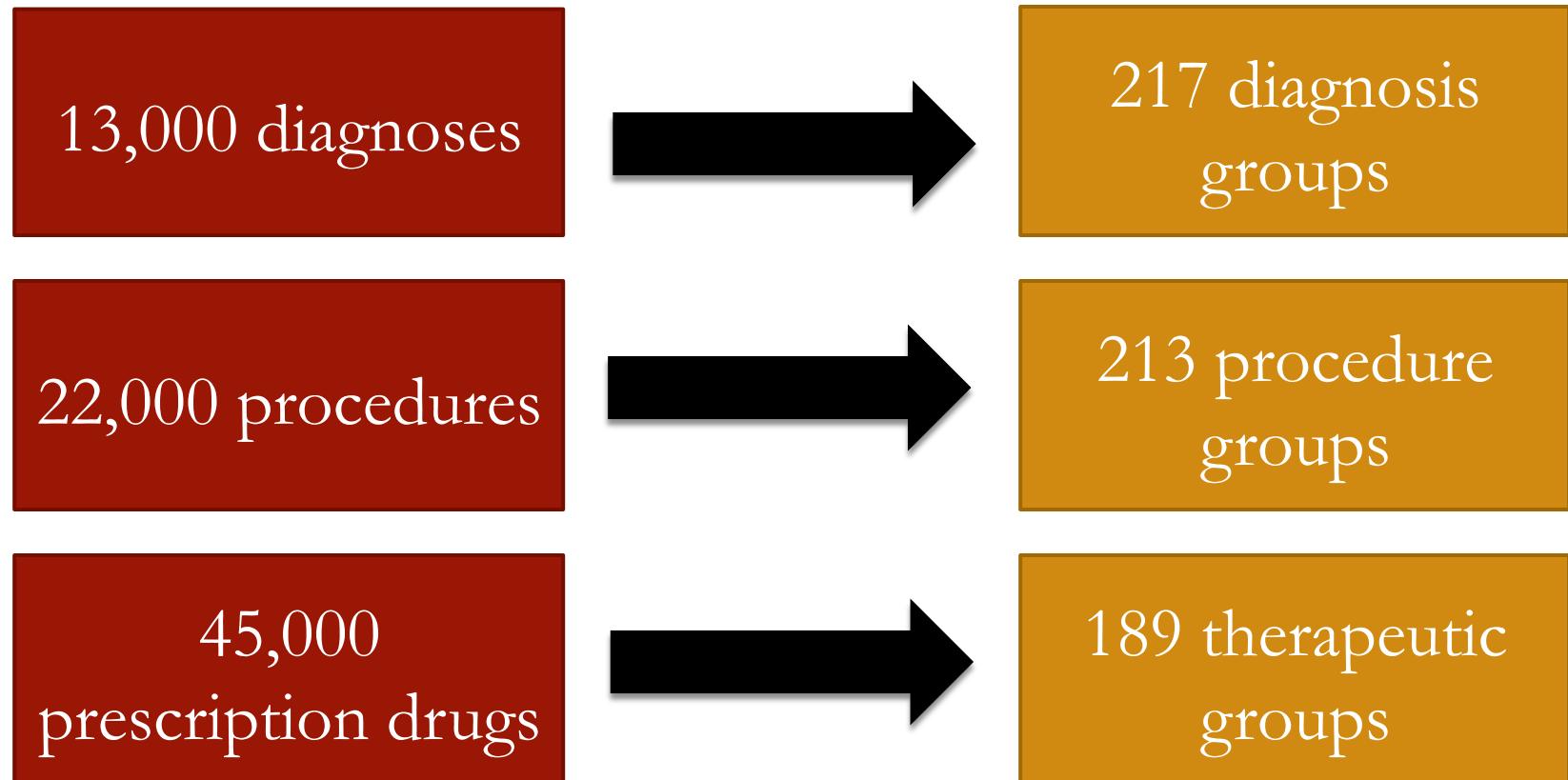
- Available: claims data for **2.4 million people** over a span of **3 years**

“Observation”
Period
2001-2003

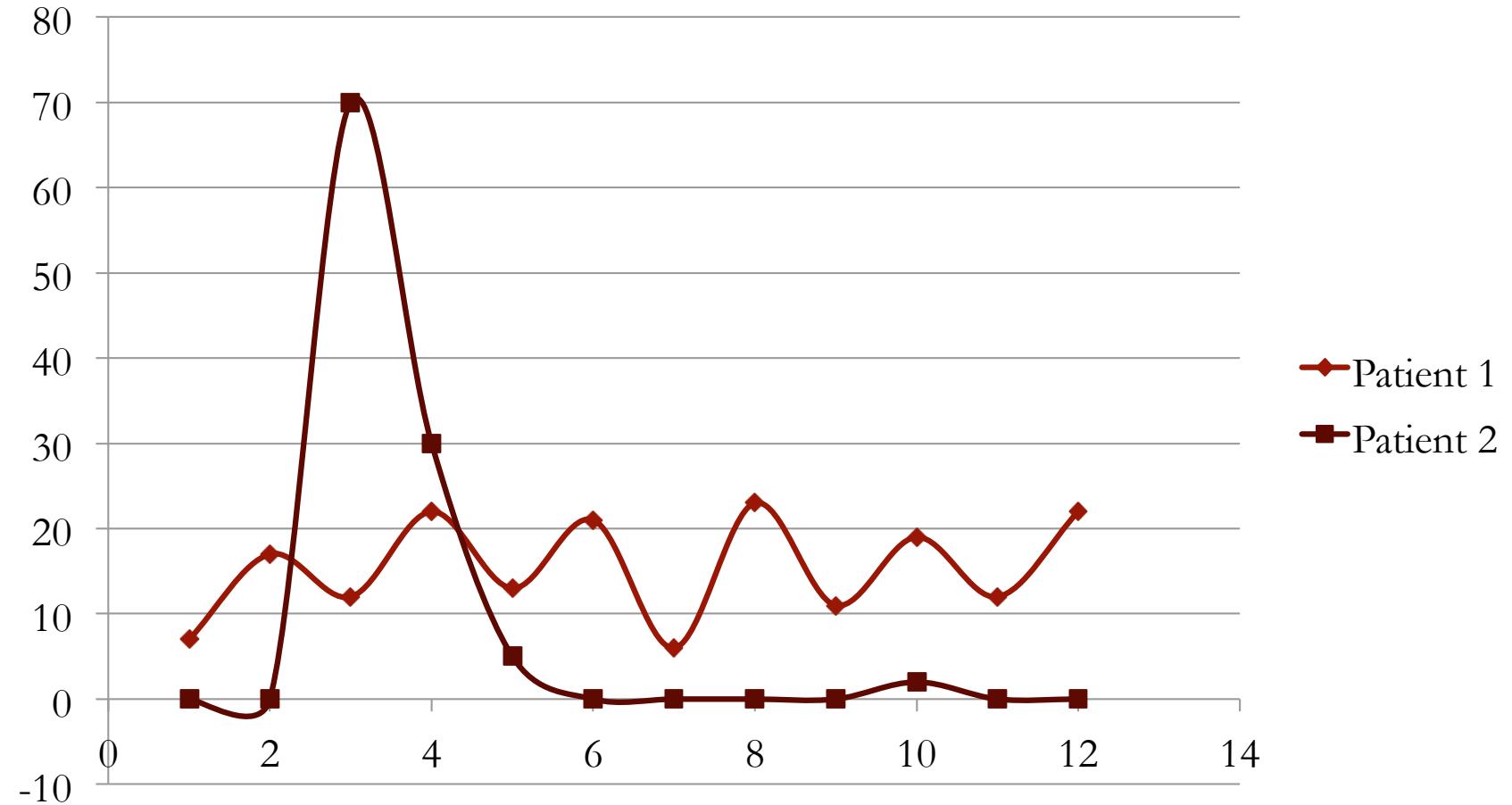
“Results”
Period
2003-2004

- Include only people with data for at least 10 months in both periods – **400,000 people**

Variables



Variables – Cost Profiles



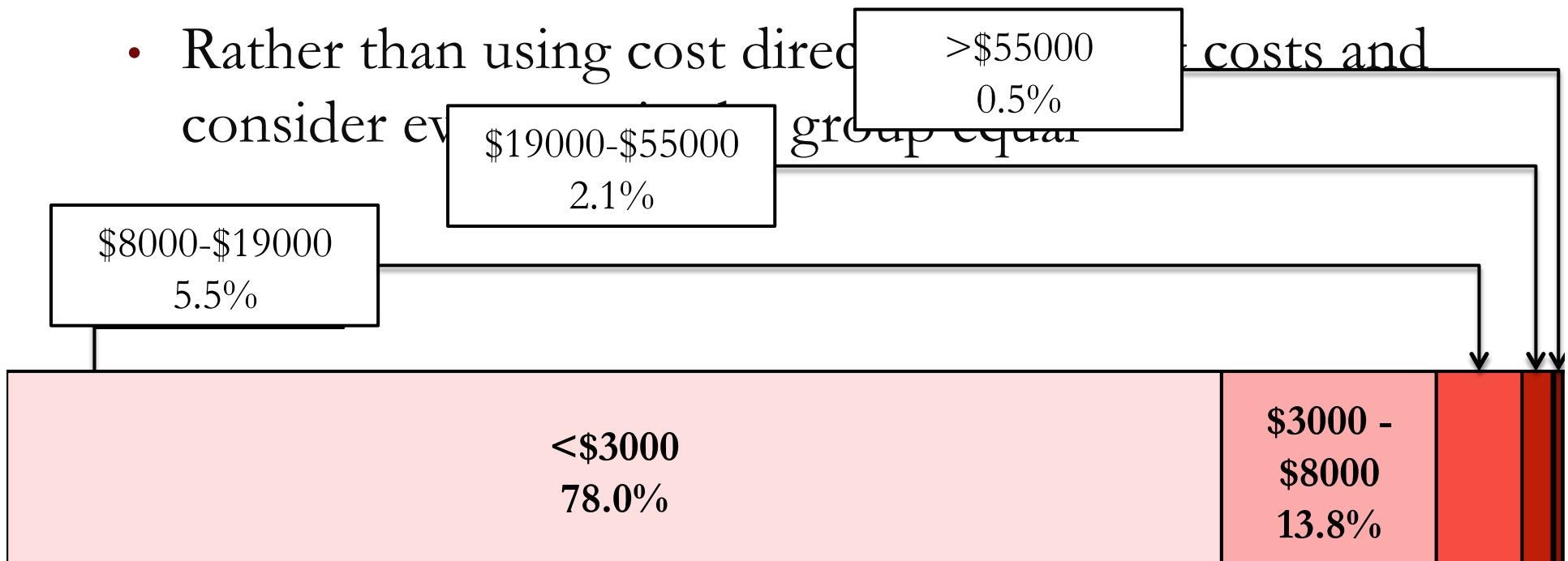
Additional Variables



- Chronic condition cost indicators
- 269 medically defined risk rules
 - Interactions between illnesses *obesity – depression*
 - Interactions between diagnosis and age
 - Noncompliance to treatment
 - Illness severity
- Gender and age

Cost Variables

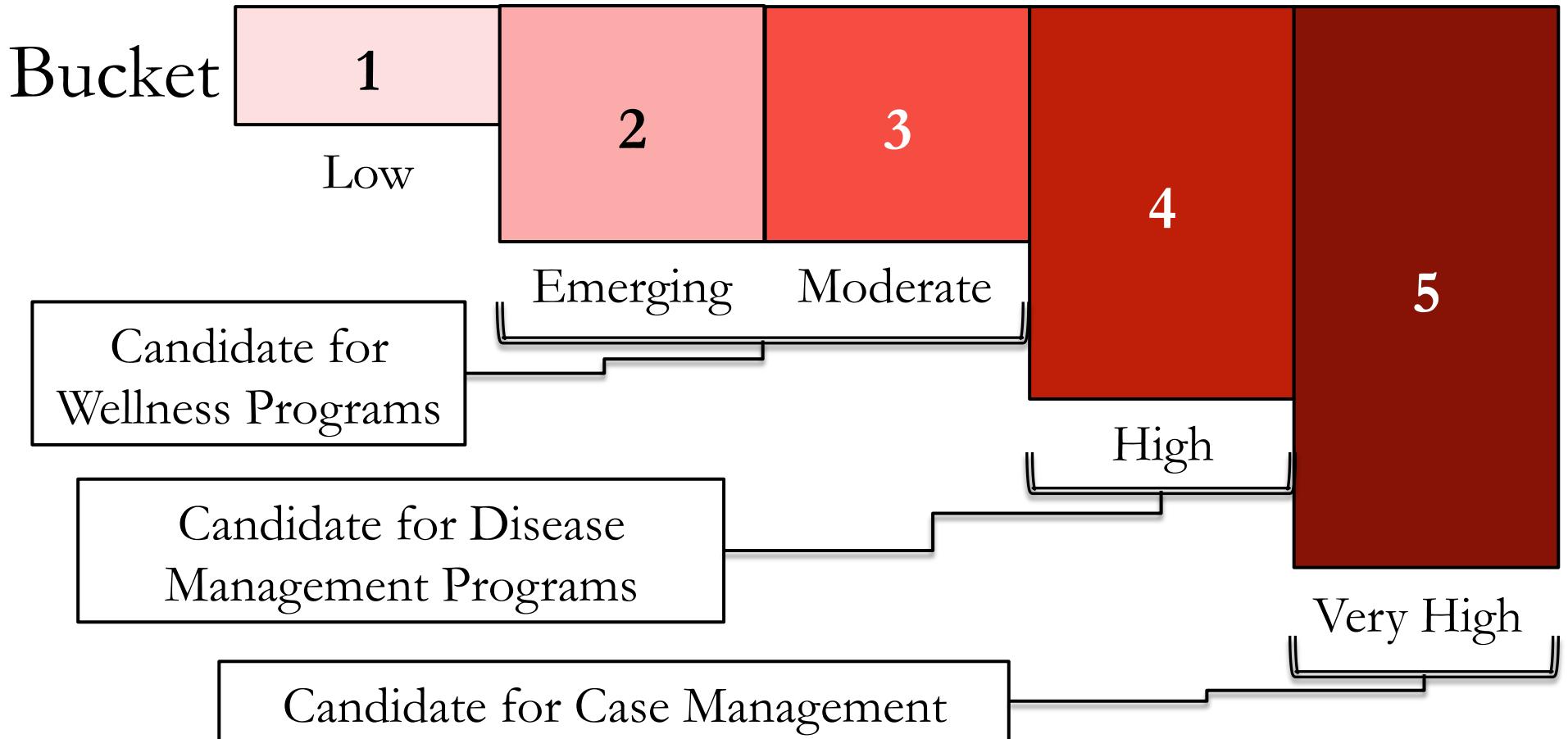
- Rather than using cost directed at costs and consider even group equal



Bucket



Medical Interpretation of Buckets



Error Measures

- Typically we use R^2 or accuracy, but others can be used
- In case of D2Hawkeye, failing to classify a **high-cost patient** correctly is **worse** than failing to classify a **low-cost patient** correctly
- Use a “penalty error” to capture this asymmetry

Penalty Error

- Key idea: use asymmetric penalties
- Define a “penalty matrix” as the cost of being wrong

		Outcome				
		1	2	3	4	5
Forecast	1	0	2	4	6	8
	2	1	0	2	4	6
	3	2	1	0	2	4
	4	3	2	1	0	2
	5	4	3	2	1	0

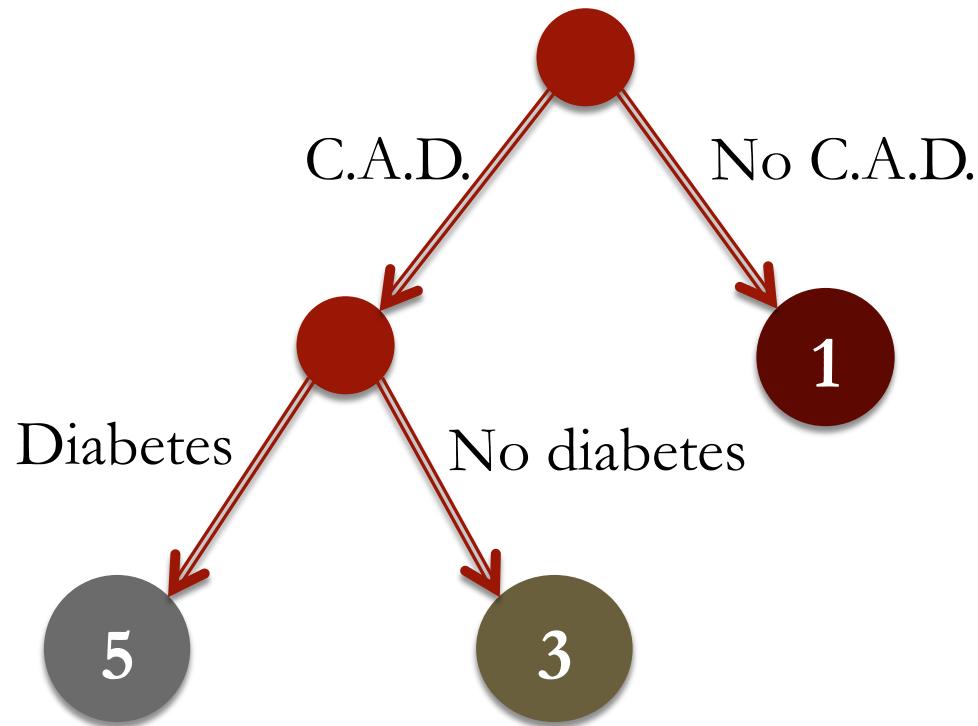
Baseline



- Baseline is to simply predict that the cost in the next “period” will be the cost in the current period
- Accuracy of 75%
- Penalty Error of 0.56

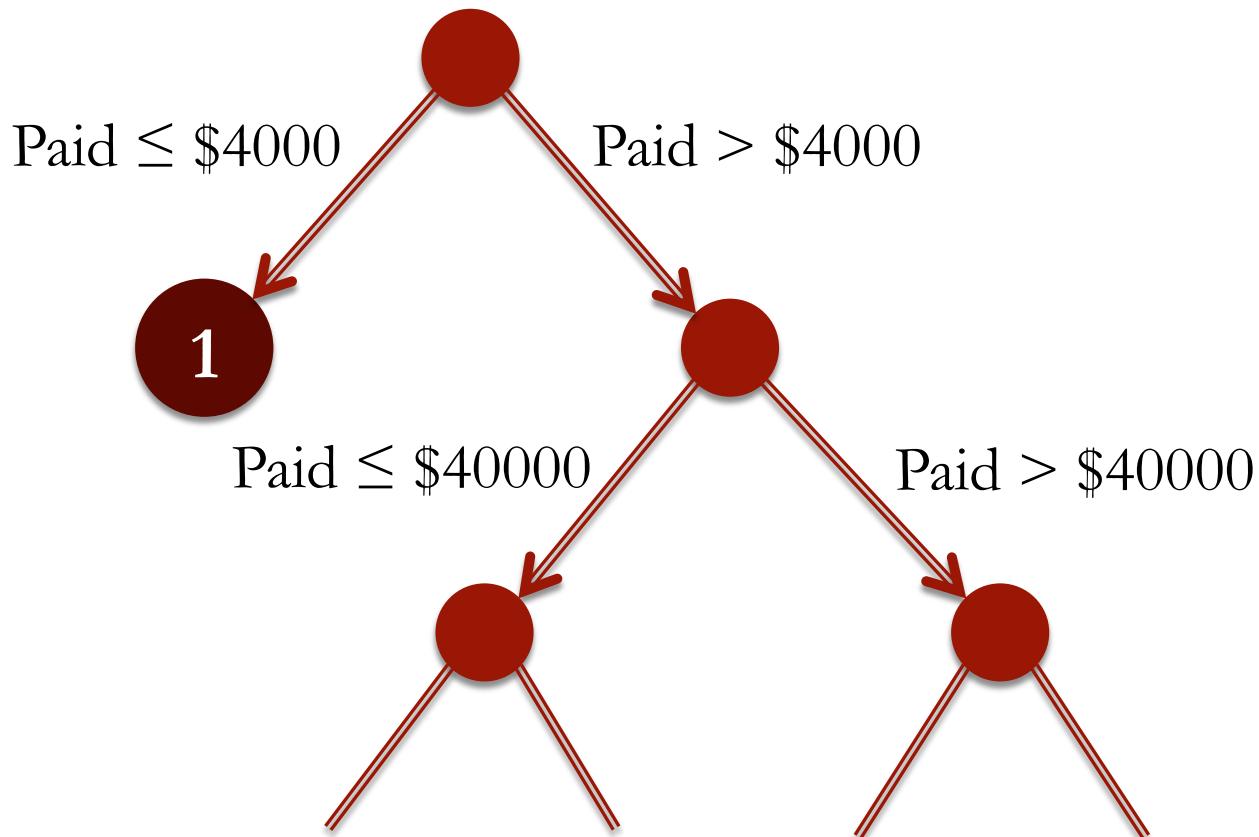
Multi-class Classification

- We are predicting a bucket number
- Example



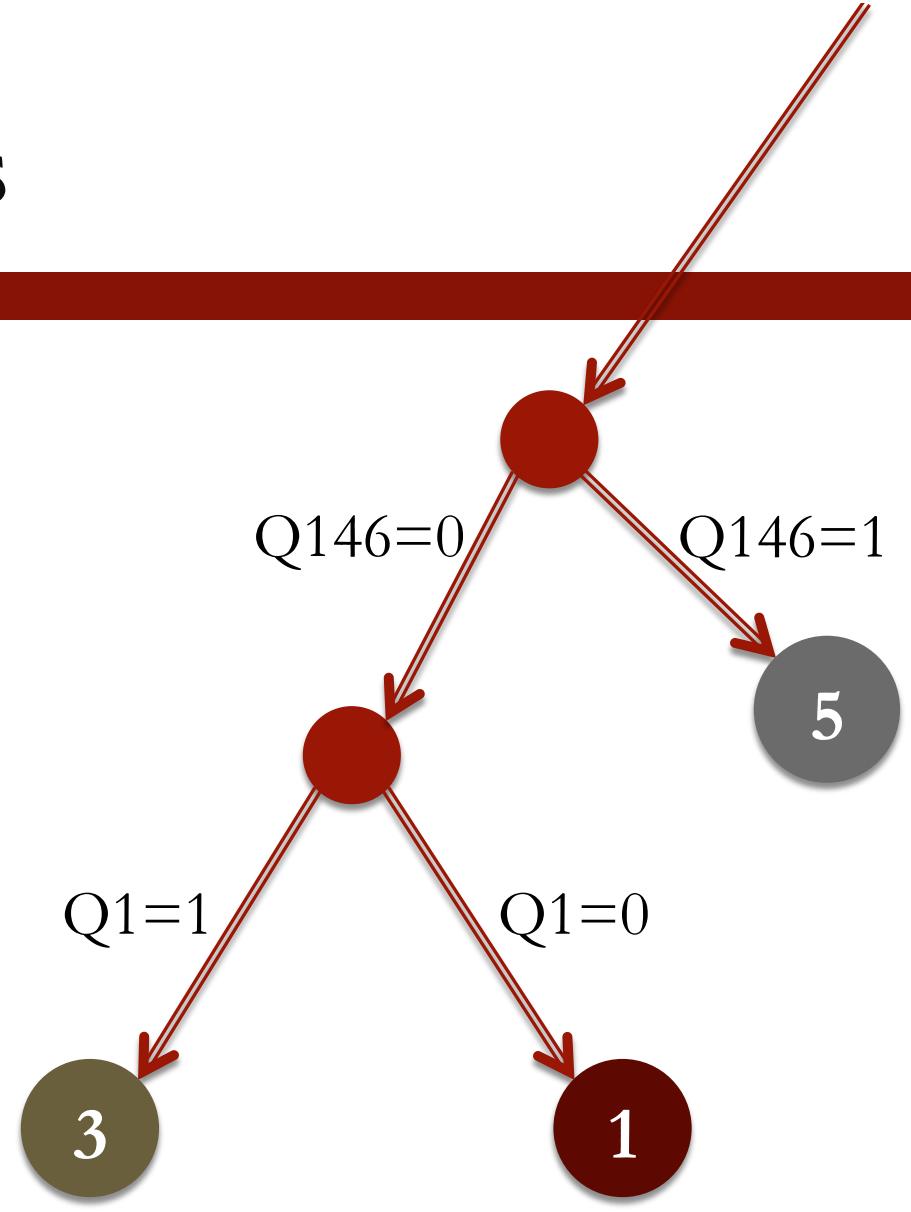
Most Important Factors

- First splits are related to cost



Secondary Factors

- Risk factors
- Chronic Illness
- “Q146”
 - Asthma + depression
- “Q1”
 - Risk factor indicating hylan injection
 - Possible knee replacement or arthroscopy



Example Groups for Bucket 5



- Under 35 years old, between \$3300 and \$3900 in claims, C.A.D., but no office visits in last year
- Claims between \$3900 and \$43000 with at least \$8000 paid in last 12 months, \$4300 in pharmacy claims, acute cost profile and cancer diagnosis
- More than \$58000 in claims, at least \$55000 paid in last 12 months, and not an acute profile

Results

Bucket	Accuracy		Penalty Error	
	Trees	Baseline	Trees	Baseline
All	80%	75%	0.52	0.56
1	85%	85%	0.42	0.44
2	60%	31%	0.89	0.96
3	53%	21%	1.01	1.37
4	39%	19%	1.01	1.72
5	30%	23%	1.01	1.88

Insights



- **Substantial improvement** over the baseline
- **Doubled accuracy** over baseline in some cases
- Smaller accuracy improvement on **bucket 5**, but
much lower penalty

Analytics Provide an Edge



- Substantial improvement in D2Hawkeye's ability to identify patients who need more attention
- Because the model was interpretable, physicians were able to improve the model by identifying new variables and refining existing variables
- Analytics gave D2Hawkeye an edge over competition using “last-century” methods



“Location, location, location!”

Regression Trees for Housing Data

Boston

- Capital of the state of Massachusetts, USA
- First settled in 1630
- 5 million people in greater Boston area, some of the highest population densities in America.



Boston



Housing Data



- A paper was written on the relationship between **house prices** and **clean air** in the late 1970s by David Harrison of Harvard and Daniel Rubinfeld of U. of Michigan.
- “Hedonic Housing Prices and the Demand for Clean Air” has been cited ~1000 times
- Data set widely used to evaluate algorithms.

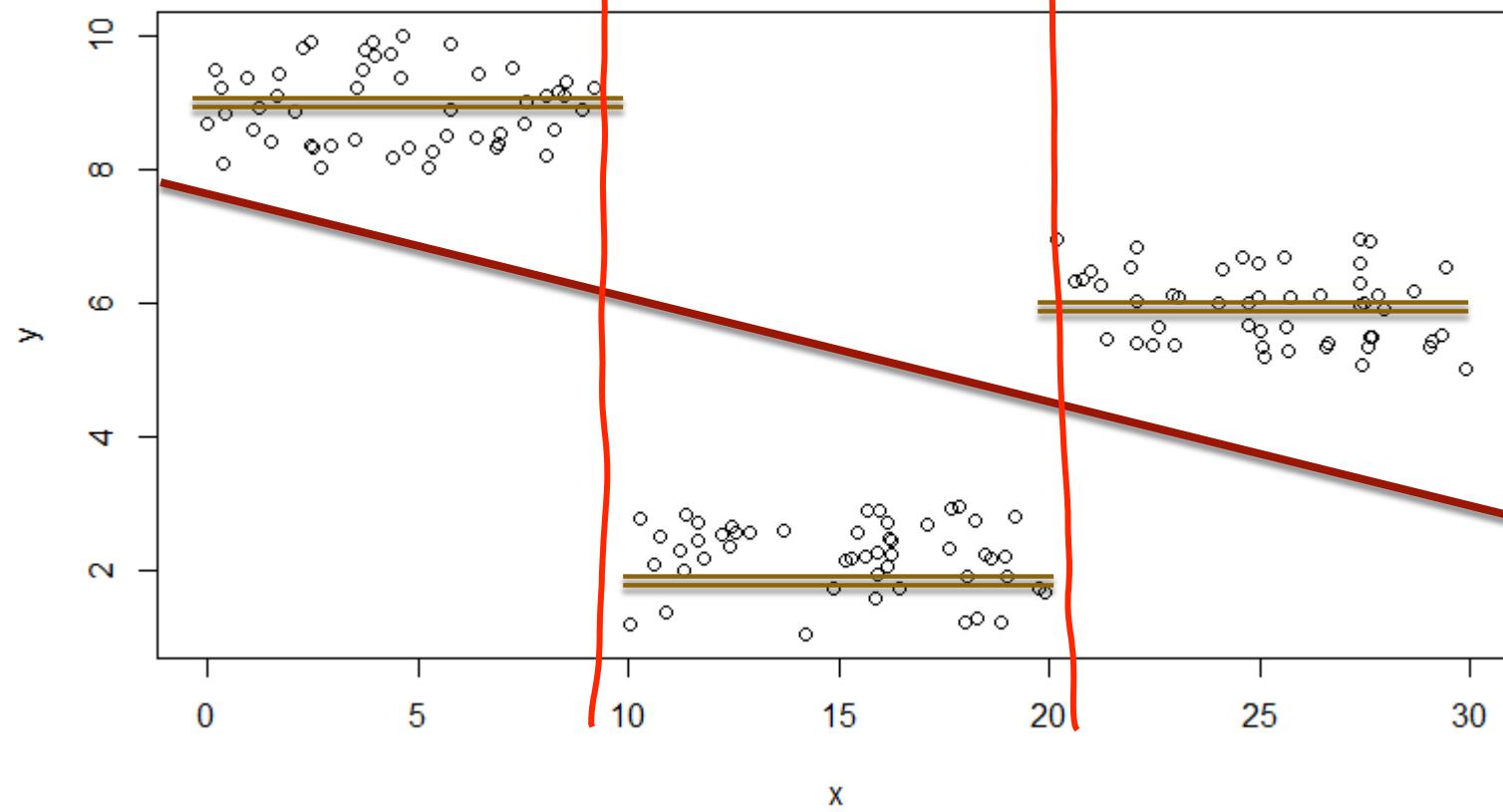
The R in CART

- In the lecture we mostly discussed **classification trees** – the output is a factor/category
- Trees can also be used for **regression** – the output at each leaf of the tree is no longer a category, but a number
- Just like classification trees, **regression trees** can capture **nonlinearities** that linear regression can't.

Regression Trees

- With Classification Trees we report the average outcome at each leaf of our tree, e.g. if the outcome is “true” 15 times, and “false” 5 times, the value at that leaf is:
$$\frac{15}{15+5} = 0.75 \geq 0.5 \rightarrow \text{true}$$
- With Regression Trees, we have continuous variables, so we simply report the average of the values at that leaf:
$$3, 4, 5 = 4$$

Example



Housing Data



- We will explore the dataset with the aid of trees.
- Compare linear regression with regression trees.
- Discussing what the “cp” parameter means.
- Apply cross-validation to regression trees.

Understanding the data



- Each entry corresponds to a census **tract**, a statistical division of the area that is used by researchers to break down towns and cities.
- There will usually be multiple census tracts per **town**.
- **LON** and **LAT** are the longitude and latitude of the center of the census tract.
- **MEDV** is the median value of owner-occupied homes, in thousands of dollars.

Understanding the data



- **CRIM** is the per capita crime rate
- **ZN** is related to how much of the land is zoned for large residential properties
- **INDUS** is proportion of area used for industry
- **CHAS** is 1 if the census tract is next to the Charles River
- **NOX** is the concentration of nitrous oxides in the air
- **RM** is the average number of rooms per dwelling

Understanding the data



- **AGE** is the proportion of owner-occupied units built before 1940
- **DIS** is a measure of how far the tract is from centers of employment in Boston
- **RAD** is a measure of closeness to important highways
- **TAX** is the property tax rate per \$10,000 of value
- **PTRATIO** is the pupil-teacher ratio by town

The “cp” parameter

- “cp” stands for “**complexity parameter**”
- Recall the first tree we made using LAT/LON had many splits, but we were able to trim it without losing much accuracy.
- Intuition: having too many splits is bad for generalization, so we should penalize the **complexity**

The “cp” parameter

- Define **RSS**, the **residual sum of squares**, the sum of the square differences

$$RSS = \sum_{i=1}^n (y_i - f(x_i))^2$$

- Our goal when building the tree is to minimize the RSS by making splits, but we want to penalize too many splits. Define **S** to be the number of splits, and λ (lambda) to be our penalty. Our goal is to find the tree that minimizes

$$\sum_{Leaves} (\text{RSS at each leaf}) + \lambda S$$

The “cp” parameter

- λ (lambda) = 0.5

Splits	RSS	Total Penalty
0	5	5
1	$2 + 2 = 4$	$4 + 0.5*1 = 4.5$
2	$1+0.8+2 = 3.8$	$3.8 + 0.5*2 = 4.8$

The “cp” parameter

$$\sum_{Leaves} (\text{RSS at each leaf}) + \lambda S$$

- If pick a large value of λ , we won't make many splits because we pay a big price for every additional split that outweighs the decrease in “error”
- If we pick a small (or zero) value of λ , we'll make splits until it no longer decreases error.

The “cp” parameter

- The definition of “cp” is closely related to λ
- Consider a tree with no splits – we simply take the average of the data. Calculate RSS for that tree, let us call it **RSS(no splits)**

$$c_p = \frac{\lambda}{\text{RSS}(\text{no splits})}$$



PREDICTIVE CODING

Bringing Text Analytics to the Courtroom

15.071 – The Analytics Edge

Enron Corporation



- U.S. energy company from Houston, Texas
- Produced and distributed power
- Market capitalization exceeded \$60 billion
- *Forbes*: Most Innovative U.S. Company, 1996-2001
- Widespread accounting fraud exposed in 2001
 - Led to bankruptcy, the largest ever at that time
 - Led major accounting firm Arthur Andersen to dissolve
- Symbol of corporate corruption

California Energy Crisis

- California is most populous state in United States
- In 2000-2001, plagued by blackouts despite having plenty of power plants
- Enron played a key role in causing crisis
 - Reduced supply to state to cause price spikes
 - Made trades to profit from the market instability
- Federal Energy Regulatory Commission (FERC) investigated Enron's involvement
 - Eventually led to \$1.52 billion settlement
 - Topic of today's recitation

The eDiscovery Problem

- Enron had millions of electronic files
- Leads to the *eDiscovery* problem: how we find files relevant to a lawsuit?
 - In legal parlance, searching for *responsive* documents
- Traditionally, keyword search followed by manual review
 - Tedious process
 - Expensive, time consuming
- More recently: *predictive coding (technology-assisted review)*
 - Manually label some of the documents to train models
 - Apply models to much larger set of documents

The Enron Corpus



- FERC publicly released emails from Enron
- > 600,000 emails, 158 users (mostly senior management)
- Largest publicly available set of emails
- Dataset we will use for predictive coding
- We will use labeled emails from the 2010 Text Retrieval Conference Legal Track
 - *email* – text of the message
 - *responsive* – does email relate to energy schedules or bids?

Predictive Coding Today



- In legal system, difficult to change existing practices
 - System based on *past precedent*
 - eDiscovery historically performed by keyword search coupled with manual review
- 2012 U.S. District Court ruling: predictive coding is legitimate eDiscovery tool
- Use likely to expand in coming years



TURNING TWEETS INTO KNOWLEDGE

An Introduction to Text Analytics



Twitter

- Twitter is a social networking and communication website founded in 2006
- Users share and send messages that can be no longer than 140 characters long
- One of the Top 10 most-visited sites on the internet
- Initial Public Offering in 2013
- Valuation ~\$31 billion



Impact of Twitter

- Use by protestors across the world
- Natural disaster notification, tracking of diseases
- Celebrities, politicians, and companies connect with fans and customers
- Everyone is watching!



2013 AP Twitter Hack



The Associated Press @AP

7m

Breaking: Two Explosions in the White House and Barack Obama is injured

[Expand](#) [Reply](#) [Retweet](#) [Favorite](#) [More](#)

- The Associated Press is a major news agency that distributes news stories to other news agencies
- In April 2013 someone tweeted the above message from the main AP verified Twitter account
- S&P500 stock index fell 1% in seconds, but the White House rapidly clarified

Understanding People

- Many companies maintain online presences
- Managing public perception in age of instant communication essential
- Reacting to changing sentiment, identifying offensive posts, determining topics of interest...
- How can we use analytics to address this?



Using Text as Data

- Until now, our data has typically been
 - Structured
 - Numerical
 - Categorical
- Tweets are
 - Loosely structured
 - Textual
 - Poor spelling, non-traditional grammar
 - Multilingual



hannah @lawlöff

MY ELECTRIC HAS WENT OUT AND A GIANT SPIDER IS COMING 4
ME AND mY ONLY SOURCE OF LIGHT IS THE FLASHLIGHT ON MY
PHONE GOD BLESS @Apple

[Expand](#)



matt @clairvoyant

WHYCANT I GO BACK TO IOS6 ITS NOT THAT BIG A DEAL @Apple
I LIKE YOUR OLD OPERATING SYSTEM BETTER

[Expand](#)

Text Analytics



- We have discussed why people care about textual data, but how do we handle it?
- Humans can't keep up with Internet-scale volumes of data
 - ~500 million tweets per day!
- Even at a small scale, the cost and time required may be prohibitive

How Can Computers Help?

- Computers need to understand text
- This field is called **Natural Language Processing**
- The goal is to understand and derive meaning from human language
- In 1950, Alan Turing proposes a test of machine intelligence: passes if it can take part in a real-time conversation and cannot be distinguished from a human



History of Natural Language Processing

- Some progress: “chatterbots” like ELIZA
- Initial focus on understanding grammar
- Focus shifting now towards statistical, machine learning techniques that learn from large bodies of text
- Modern “artificial intelligences”: Apple’s Siri and Google Now



Why is it Hard?



- Computers need to understand text
- Ambiguity:
 - “I put my **bag** in the **car**. **It** is large and blue”
 - “**It**” = **bag**? “**It**” = **car**?
- Context:
 - Homonyms, metaphors
 - Sarcasm
- In this lecture, we’ll see how we can build analytics models using text as our data

Sentiment Mining - Apple

- **Apple** is a computer company known for its laptops, phones, tablets, and personal media players
- Large numbers of fans, large number of “haters”
- Apple wants to monitor how people feel about them over time, and how people receive new announcements.
- **Challenge:** Can we correctly classify tweets as being negative, positive, or neither about Apple?



Creating the Dataset



- Twitter data is publically available
 - Scrape website, or
 - Use special interface for programmers (API)
 - Sender of tweet may be useful, but we will ignore
- Need to construct the outcome variable for tweets
 - Thousands of tweets
 - Two people may disagree over the correct classification
 - One option is to use **Amazon Mechanical Turk**

Amazon Mechanical Turk



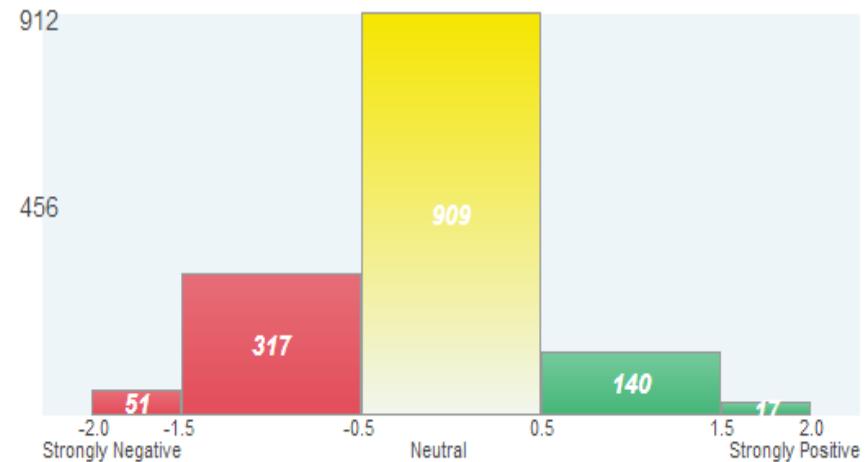
- Break tasks down into small components and distribute online
- People can sign up to perform the tasks for a fee
 - Pay workers, e.g. \$0.02 per classified tweet
 - Amazon MTurk serves as a broker, takes small cut
- Many tasks require human intelligence, but may be time consuming or require building otherwise unneeded capacity

Our Human Intelligence Task

- Actual question we used:

Judge the sentiment expressed by the following item toward the software company "Apple"

- Workers could pick from
 - Strongly Negative (-2)
 - Negative (-1)
 - Neutral (0)
 - Positive (+1)
 - Strongly Positive (+2)
- Five workers labeled each tweet



Our Human Intelligence Task



- For each tweet, we take the average of the five scores.
 - “LOVE U @APPLE” (1.8)
 - “@apple @twitter Happy Programmers' Day folks!” (0.4)
 - “So disappointed in @Apple Sold me a Macbook Air that WONT run my apps. So I have to drive hours to return it. They wont let me ship it.” (-1.4)
- We have labels, but how do we build independent variables from the text of a tweet?

A Bag of Words

- Fully understanding text is difficult
- Simpler approach:

Count the number of times each words appears

- “This course is great. I would recommend this course to my friends.”

THIS	COURSE	GREAT	...	WOULD	FRIENDS
2	2	1	...	1	1

A Simple but Effective Approach



- One feature for each word - a simple approach, but effective
- Used as a baseline in text analytics projects and natural language processing
- Not the whole story though - preprocessing can dramatically improve performance!

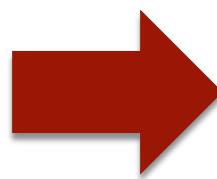
Cleaning Up Irregularities

- Text data often has many inconsistencies that will cause algorithms trouble
- Computers are very literal by default – Apple, APPLE, and ApPLe will all be counted separately.
- Change all words to either lower-case or upper-case

Apple	APPLE	ApPLe	→	apple
apple	apple	apple		3

Cleaning Up Irregularities

- Punctuation also causes problems – basic approach is to remove everything that isn't a,b,...,z
- Sometimes punctuation **is** meaningful
 - Twitter: **@apple** is a message to Apple, **#apple** is about Apple
 - Web addresses: **www.website.com/somepage.html**
- Should tailor approach to the specific problem

@Apple	APPLE!	--apple--		apple
apple	apple	apple		3

Removing Unhelpful Terms

- Many words are frequently used but are only meaningful in a sentence - “**stop words**”
 - Examples: *the, is, at, which...*
 - Unlikely to improve machine learning prediction quality
 - Remove to reduce size of data
- Two words at a time?
 - “**The Who**” → “ ”
 - “**Take That**” → “**Take**”



Stemming

- Do we need to draw a distinction between the following words?
argue argued argues arguing
- Could all be represented by a common **stem**, argu
- Algorithmic process of performing this reduction is called **stemming**
- Many ways to approach the problem

Stemming

- Could build a **database of words** and their stems
 - **Pro:** handles exceptions
 - **Con:** won't handle new words, bad for the Internet!
- Can write a **rule-based** algorithm
 - e.g. if word ends in “ed”, “ing”, or “ly”, remove it
 - **Pro:** handles new/unknown words well
 - **Con:** many exceptions, misses words like **child** and **children** (but would get other plurals: **dog** and **dogs**)

Stemming

- The second option is widely popular
 - “**Porter Stemmer**” by Martin Porter in 1980, still used!
 - Stemmers have been written for many languages
- Other options include machine learning (train algorithms to recognize the roots of words) and combinations of the above

Real example from data:

“*by far the best customer care service I have ever received*”

➡ “*by far the best custom care servic I have ever receiv*”

Sentiment Analysis Today



- Over 7,000 research articles have been written on this topic
- Hundreds of start-ups are developing sentiment analysis solutions
- Many websites perform real-time analysis of tweets
 - “tweetfeel” shows trends given any term
 - “The Stock Sonar” shows sentiment and stock prices

Text Analytics in General



- Selecting the specific features that are relevant in the application
- Applying problem specific knowledge can get better results
 - Meaning of symbols
 - Features like number of words

The Analytics Edge



- Analytical sentiment analysis can replace more labor-intensive methods like polling
- Text analytics can deal with the massive amounts of unstructured data being generated on the internet
- Computers are becoming more and more capable of interacting with humans and performing human tasks
- In the next lecture, we'll discuss IBM Watson, an impressive feat in the area of Text Analytics



MAN VS. MACHINE

How IBM Built a *Jeopardy!* Champion

15.071x – The Analytics Edge

A Grand Challenge



- In 2004, IBM Vice President Charles Lickel and co-workers were having dinner at a restaurant
- All of a sudden, the restaurant fell silent
- Everyone was watching the game show *Jeopardy!* on the television in the bar
- A contestant, Ken Jennings, was setting the record for the longest winning streak of all time (75 days)

A Grand Challenge



- Why was everyone so interested?
 - *Jeopardy!* is a quiz show that asks complex and clever questions (puns, obscure facts, uncommon words)
 - Originally aired in 1964
 - A huge variety of topics
 - Generally viewed as an impressive feat to do well
- No computer system had ever been developed that could even come close to competing with humans on *Jeopardy!*

A Tradition of Challenges



- IBM Research strives to push the limits of science
 - Have a tradition of inspiring and difficult challenges
- Deep Blue – a computer to compete against the best human chess players
 - A task that people thought was restricted to human intelligence
- Blue Gene – a computer to map the human genome
 - A challenge for computer speed and performance

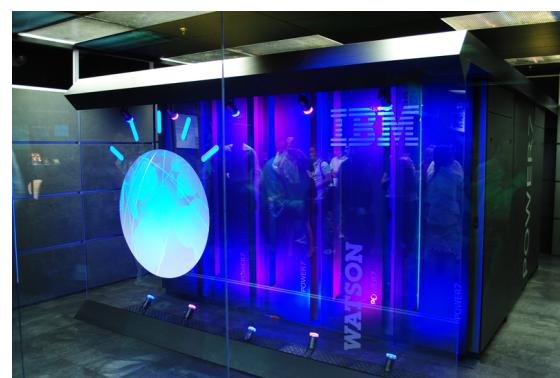
The Challenge Begins



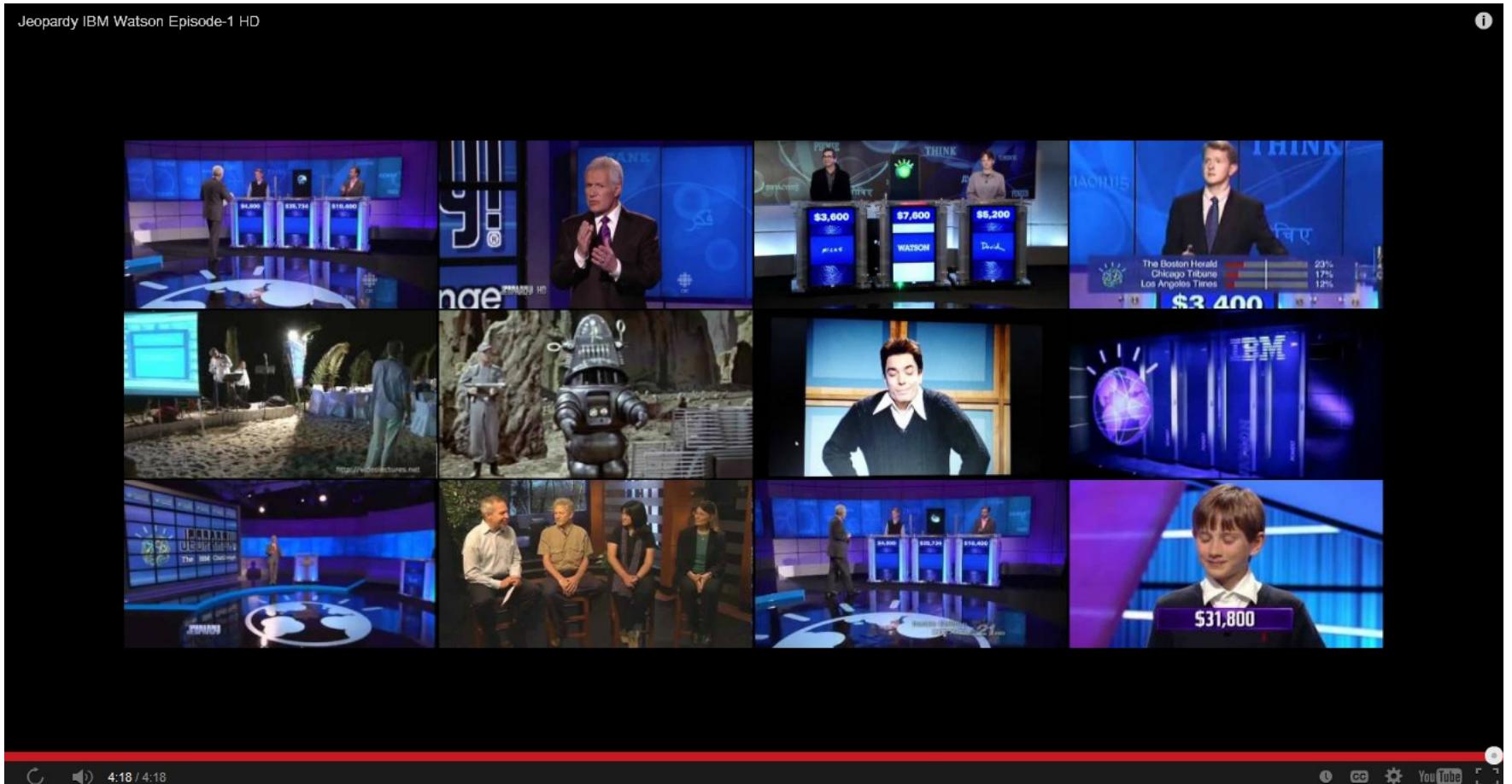
- In 2005, a team at IBM Research started creating a computer that could compete at *Jeopardy!*
 - No one knew how to beat humans, or if it was even possible
- Six years later, a two-game exhibition match aired on television
 - The winner would receive \$1,000,000

The Contestants

- Ken Jennings
 - Longest winning streak of 75 days
- Brad Rutter
 - Biggest money winner of over \$3.5 million
- Watson
 - A supercomputer with 3,000 processors and a database of 200 million pages of information



The Match Begins



The Game of *Jeopardy!*

- Three rounds per game
 - Jeopardy
 - Double Jeopardy (dollar values doubled)
 - Final Jeopardy (wager on response to one question)
- Each round has five questions in six categories
 - Wide variety of topics (over 2,500 different categories)
- Each question has a dollar value – the first to buzz in and answer correctly wins the money
 - If they answer incorrectly they lose the money



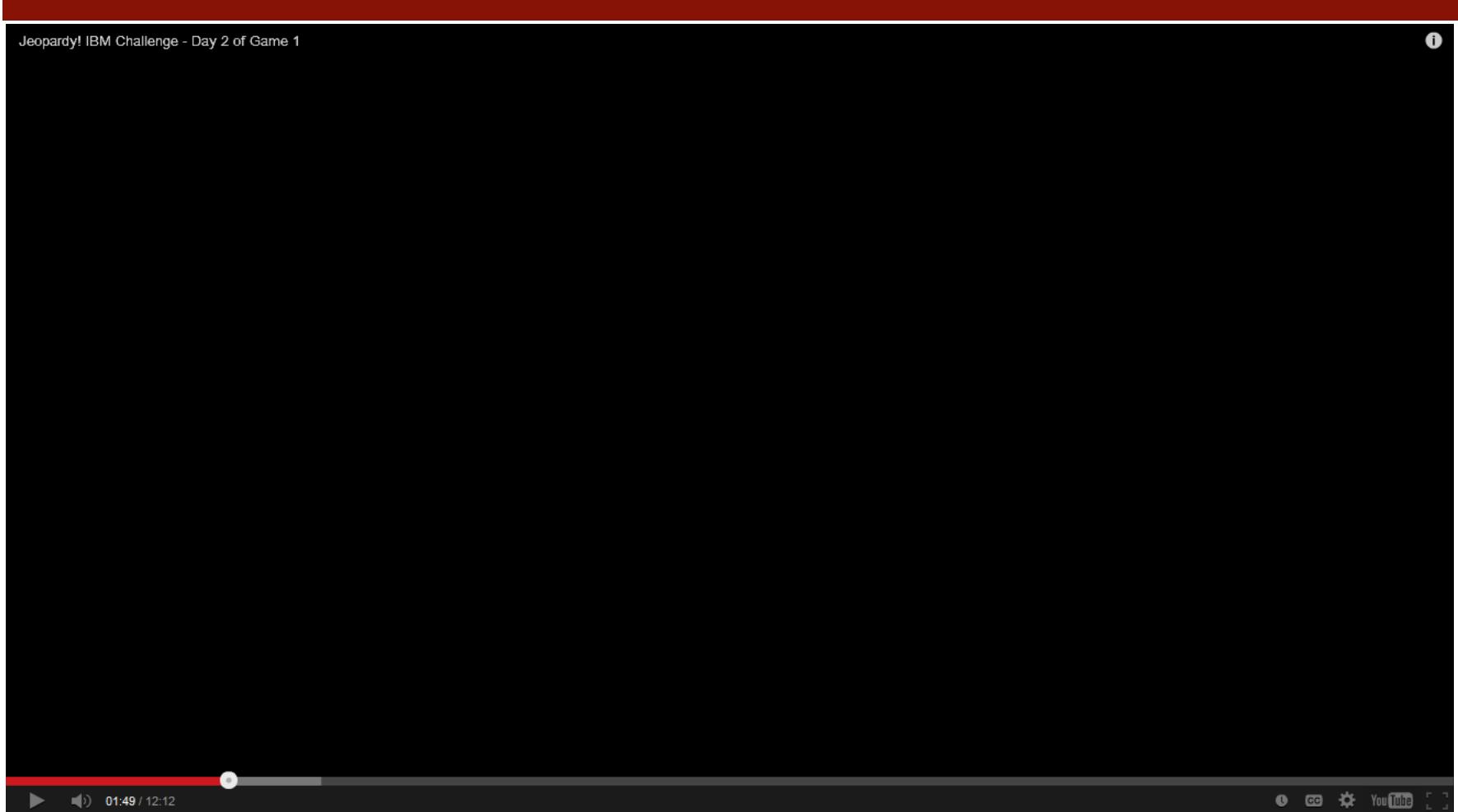
Example Round

THE DINOSAURS	NOTABLE WOMEN	OXFORD ENGLISH DICTIONARY	NAME THAT INSTRUMENT	BELGIUM	COMPOSERS BY COUNTRY
\$200	\$200	\$200	\$200	\$200	\$200
\$400	\$400	\$400	\$400	\$400	\$400
\$600	\$600	\$600	\$600	\$600	\$600
\$800	\$800	\$800	\$800	\$800	\$800
\$1000	\$1000	\$1000	\$1000	\$1000	\$1000

Jeopardy! Questions

- Cryptic definitions of categories and clues
- Answer in the form of a question
 - Q: Mozart's last and perhaps most powerful symphony shares its name with this planet.
 - A: What is Jupiter?
 - Q: Smaller than only Greenland, it's the world's second-largest island.
 - A: What is New Guinea?

Watson Playing Jeopardy



Why is Jeopardy Hard?



- Wide variety of categories, purposely made cryptic
- Computers can easily answer precise questions
 - What is the square root of $(35672-183)/33$?
- Understanding natural language is hard
 - Where was Albert Einstein born?
 - Suppose you have the following information:
“One day, from his city views of Ulm, Otto chose a water color to send to Albert Einstein as a remembrance of his birthplace.”
 - Ulm? Otto?

A Straightforward Approach



- Let's just store answers to all possible questions
- This would be impossible
 - An analysis of 200,000 previous questions yielded over 2,500 different categories
- Let's just search Google
 - No links to the outside world permitted
 - It can take considerable skill to find the right webpage with the right information

Using Analytics



- Watson received each question in text form
 - Normally, players see and hear the questions
- IBM used analytics to make Watson a competitive player
- Used over 100 different techniques for analyzing natural language, finding hypotheses, and ranking hypotheses

Watson's Database and Tools

- A massive number of data sources
 - Encyclopedias, texts, manuals, magazines, Wikipedia, etc.
- Lexicon
 - Describes the relationship between different words
 - Ex: “Water” is a “clear liquid” but not all “clear liquids” are “water”
- Part of speech tagger and parser
 - Identifies functions of words in text
 - Ex: “Race” can be a verb or a noun
 - He won the race by 10 seconds.
 - Please indicate your race.

How Watson Works

- Step 1: Question Analysis
 - Figure out what the question is looking for
- Step 2: Hypothesis Generation
 - Search information sources for possible answers
- Step 3: Scoring Hypotheses
 - Compute confidence levels for each answer
- Step 4: Final Ranking
 - Look for a highly supported answer

Step 1: Question Analysis



- What is the question looking for?
- Trying to find the Lexical Answer Type (LAT) of the question
 - Word or noun in the question that specifies the type of answer
- Ex: “Mozart’s last and perhaps most powerful symphony shares its name with **this planet**.”
- Ex: “Smaller than only Greenland, **it’s** the world’s second-largest island.”

Step 1: Question Analysis



- If we know the LAT, we know what to look for
- In an analysis of 20,000 questions
 - 2,500 distinct LATs were found
 - 12% of the questions do not have an explicit LAT
 - The most frequent 200 explicit LATs cover less than 50% of the questions
- Also performs **relation detection** to find relationships among words, and **decomposition** to split the question into different clues

Step 2: Hypothesis Generation



- Uses the question analysis from Step 1 to produce candidate answers by searching the databases
- Several hundred candidate answers are generated
- Ex: “Mozart’s last and perhaps most powerful symphony shares its name with **this planet**.”
 - Candidate answers: Mercury, Earth, Jupiter, etc.

Step 2: Hypothesis Generation



- Then each candidate answer plugged back into the question in place of the LAT is considered a hypothesis
 - Hypothesis 1: “Mozart’s last and perhaps most powerful symphony shares its name with **Mercury**.”
 - Hypothesis 2: “Mozart’s last and perhaps most powerful symphony shares its name with **Jupiter**.”
 - Hypothesis 3: “Mozart’s last and perhaps most powerful symphony shares its name with **Earth**.”

Step 2: Hypothesis Generation



- If the correct answer is not generated at this stage, Watson has no hope of getting the question right
- This step errors on the side of generating a lot of hypotheses, and leaves it up to the next step to find the correct answer

Step 3: Scoring Hypotheses



- Compute *confidence levels* for each possible answer
 - Need to accurately estimate the probability of a proposed answer being correct
 - Watson will only buzz in if a confidence level is above a threshold
- Combines a large number of different methods

Lightweight Scoring Algorithms



- Starts with “lightweight scoring algorithms” to prune down large set of hypotheses
- Ex: What is the likelihood that a candidate answer is an instance of the LAT?
 - If this likelihood is not very high, throw away the hypothesis
- Candidate answers that pass this step proceed the next stage
 - Watson lets about 100 candidates pass into the next stage

Scoring Analytics

- Need to gather supporting evidence for each candidate answer
- Passage Search
 - Retrieve passages that contain the hypothesis text
 - Let's see what happens when we search for our hypotheses on Google
 - Hypothesis 1: "Mozart's last and perhaps most powerful symphony shares its name with **Mercury**."
 - Hypothesis 2: "Mozart's last and perhaps most powerful symphony shares its name with **Jupiter**."

Passage Search



Web Images Maps Shopping News More ▾ Search tools

About 938,000 results (0.55 seconds)

[Mercury: Mozart's Jupiter Symphony - The Front Row](#)

www.thefrontrow.org/.../1349112026-Mercury-Mozarts-Jupiter-Sympho... ▾

Oct 1, 2012 - Antoine Plante, artistic director of the period-instruments group **Mercury** - The Orchestra Redefined, talks about the program of **symphonies** and ...

[Mozarts Jupiter Symphony | Mercury \(formerly Mercury Baro...](#)

www.artshound.com › MUSIC ▾

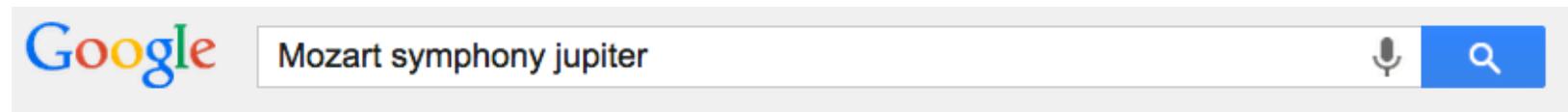
Opening the **Mercury** season at the Wortham Center's Cullen Theatre on Saturday, October 6, 2012 will be a program featuring **Mozart's "Jupiter" Symphony**.

[Event - Mozart's "Jupiter" Symphony Mercury Houston - The ...](#)

mercuryhouston.org/events/7/ ▾

Mercury combines the forces of Haydn and **Mozart** for a memorable concert event, highlighted by **Mozarts** iconic Jupiter **Symphony**. A wonderful way to kick off ...

Passage Search



Web Images Maps Shopping Videos More ▾ Search tools

About 1,440,000 results (0.31 seconds)

[Symphony No. 41 \(Mozart\) - Wikipedia, the free encyclopedia](#)

[en.wikipedia.org/wiki/Symphony_No._41_\(Mozart\)](http://en.wikipedia.org/wiki/Symphony_No._41_(Mozart)) ▾

It was the last **symphony** that he composed, and also the longest. The work is nicknamed the **Jupiter Symphony**. This name stems not from **Mozart** but rather was ...

[Instrumentation](#) - [Composition and premiere](#) - [Movements](#) - [Notes](#)

[W. A. Mozart - Symphony No. 41 "Jupiter" in C major \(Harmon...](#)

www.youtube.com/watch?v=zK5295yEQMQ ▾

Feb 11, 2012 - W. A. Mozart - Symphony No. 41 "Jupiter" in C major, K. 551 (1788): 1. Allegro vivace, 4/4 2. Andante cantabile, 3/4 in F major 3. Menuetto: ...

[Mozart - Symphony No. 41 in C, K. 551 \[complete\] \(Jupiter\) - ...](#)

www.youtube.com/watch?v=bnK3kh8ZEgA ▾

Feb 21, 2012 - Wolfgang Amadeus Mozart completed his **Symphony** No. 41 in C major, K. 551, on 10 August 1788. It was the last **symphony** that he composed ...

Scoring Analytics

- Determine the degree of certainty that the evidence supports the candidate answers
- More than 50 different scoring components
- Ex: Temporal relationships
 - “In 1594, he took a job as a tax collector in Andalusia”
 - Two candidate answers: Thoreau and Cervantes
 - Thoreau was not born until 1817, so we are more confident about Cervantes

Step 4: Final Merging and Ranking



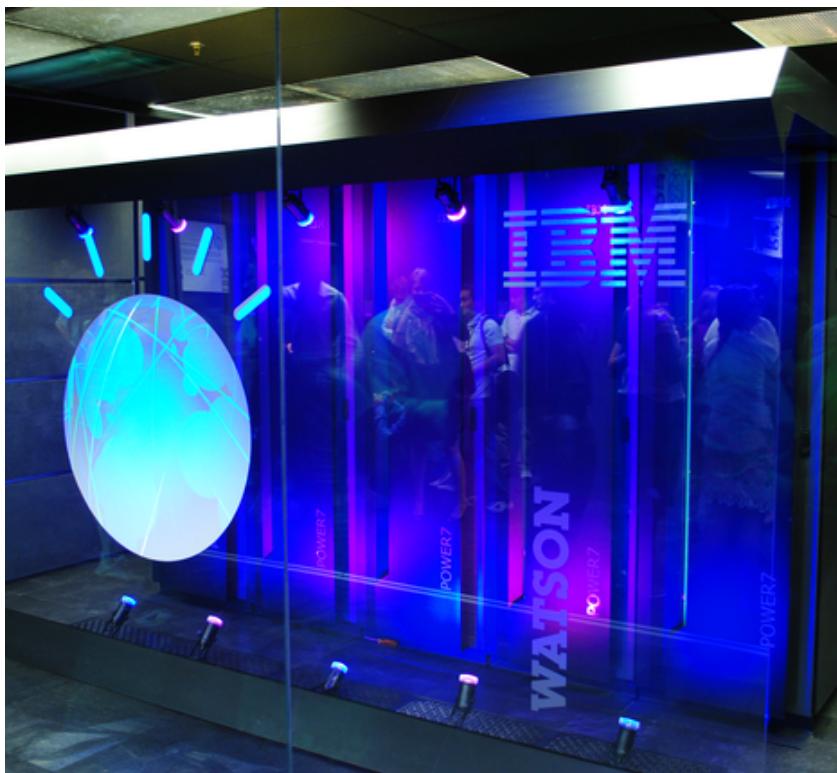
- Selecting the single best supported hypothesis
- First need to merge similar answers
 - Multiple candidate answers may be equivalent
 - Ex: “Abraham Lincoln” and “Honest Abe”
 - Combine scores
- Rank the hypotheses and estimate confidence
 - Use predictive analytics

Ranking and Confidence Estimation



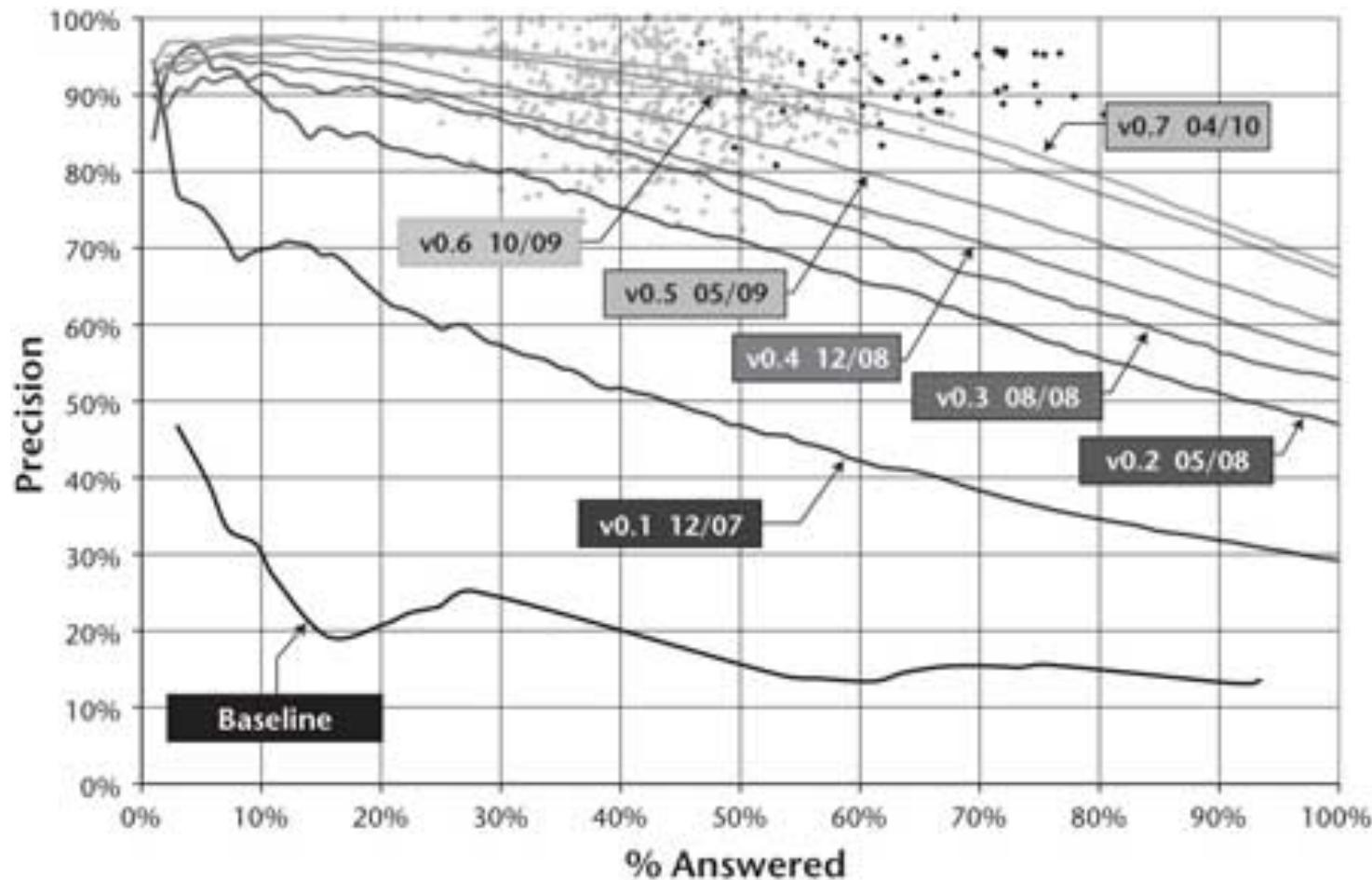
- Training data is a set of historical *Jeopardy!* questions
- Each of the scoring algorithms is an independent variable
- Use logistic regression to predict whether or not a candidate answer is correct, using the scores
- If the confidence for the best answer is high enough, Watson buzzes in to answer the question

The Watson System



- Eight refrigerator-sized cabinets
- High speed local storage for all information
- Originally took over two hours to answer one question
 - This had to be reduced to 2-6 seconds

Progress from 2006 - 2010



Let the games begin!



- The games were scheduled for February 2011
- Two games were played, and the winner would be the contestant with the highest winnings over the two games

The Jeopardy Challenge



The Results

	Ken Jennings	Brad Rutter	Watson
Game 1	\$4,800	\$10,400	\$35,734
Game 2	\$19,200	\$11,200	\$41,413
Total	\$24,000	\$21,600	\$77,147

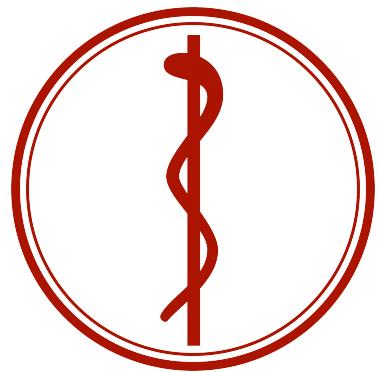
What's Next for Watson

- Apply to other domains
 - Watson is ideally suited to answering questions which cover a wide range of material and often have to deal with inconsistent or incomplete information
- Medicine
 - The amount of medical information available is doubling every 5 years and a lot of the data is unstructured
 - Cancer diagnosis and selecting the best course of treatment
 - MD Anderson and Memorial Sloan-Kettering Cancer Centers

The Analytics Edge



- Combine many algorithms to increase accuracy and confidence
 - Any one algorithm wouldn't have worked
- Approach the problem in a different way than how a human does
 - Hypothesis generation
- Deal with massive amounts of data, often in unstructured form
 - 90% of data is unstructured



Predictive Diagnosis

Clustering to Better Predict Heart Attacks

15.071x – The Analytics Edge

Heart Attacks



- Heart attack is a common complication of coronary heart disease resulting from the interruption of blood supply to part of the heart
- 2012 report from the American Heart Association estimates **about 715,000** Americans have a heart attack every year
 - **Every 20 seconds**, a person has a heart attack in the US
 - Nearly **half** occur without prior warning signs
 - **250,000** Americans die of Sudden Cardiac Death yearly

Heart Attacks

- Well-known symptoms
 - Chest pain, shortness of breath, upper body pain, nausea
- Nature of heart attacks makes it hard to predict, prevent and even diagnose
 - **25%** of heart attacks are silent
 - **47%** of sudden cardiac deaths occur outside hospitals, suggesting many do not act on early warning signs
 - **27%** of respondents to a 2005 survey recognized the symptoms and called 911 for help

Analytics Helps Monitoring



- Understanding the clinical characteristics of patients in whom heart attack was missed is key
- Need for an increased understanding of the patterns in a patient's diagnostic history that link to a heart attack
- Predicting whether a patient is at risk of a heart attack helps monitoring and calls for action
- Analytics helps **understand patterns** of heart attacks and provides **good predictions**

Claims Data

- Claims data offers an expansive view of a patient's health history
 - Demographics, medical history and medications
 - Offers insights regarding a patient's risk
 - May reveal **indicative signals and patterns**
- We will use health insurance claims filed for about 7,000 members from January 2000 – November 2007

Claims Data

- Concentrated on members with the following attributes
 - At least 5 claims with coronary artery disease diagnosis
 - At least 5 claims with hypertension diagnostic codes
 - At least 100 total medical claims
 - At least 5 pharmacy claims
 - Data from at least 5 years
- Yields patients with a high risk of heart attack and a reasonably rich history and continuous coverage

Data Aggregation



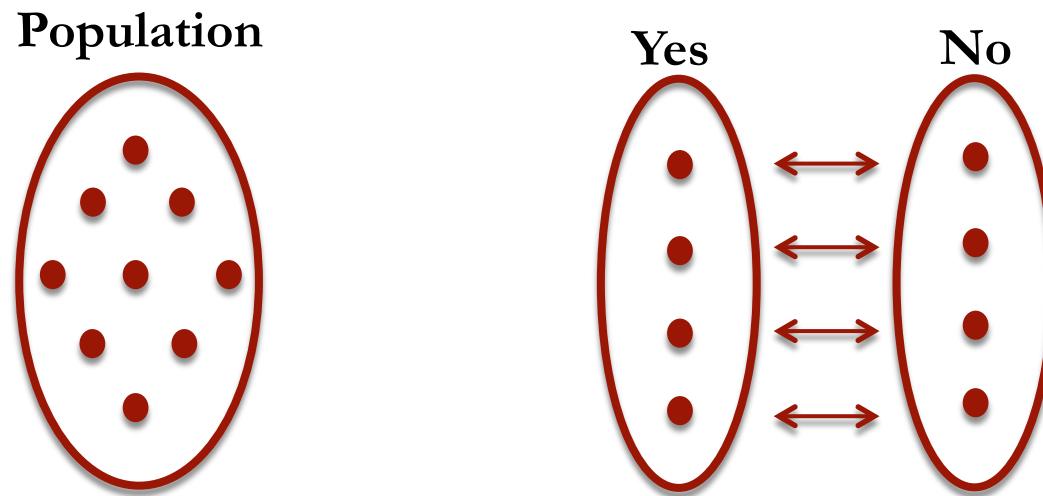
- The resulting dataset includes about **20 million health insurance entries** including individual medical and pharmaceutical records
- Diagnosis, procedure and drug codes in the dataset comprise tens of thousands of attributes
- Codes were aggregated into groups
 - 218 diagnosis groups, 180 procedure groups, 538 drug groups
 - 46 diagnosis groups were considered by clinicians as possible risk factors for heart attacks

Diagnostic History

- We then compress medical records to obtain a chronological representation of a patient's diagnostic profile
 - Cost and number of medical claims and hospital visits by diagnosis
- Observations split into 21 periods, each 90 days in length
 - Examined 9 months of diagnostic history leading up to heart attack/no heart attack event
 - Align data to make observations date-independent while preserving the order of events
 - 3 months ~ 0-3 months before heart attack
 - 6 months ~ 3-6 months before heart attack
 - 9 months ~ 6-9 months before heart attack

Target Variable

- Target prediction is the first occurrence of a heart attack
 - Diagnosis on medical claim
 - Visit to emergency room followed by hospitalization
 - Binary Yes/No



Dataset Compilation

Variables	Description
1	Patient identification number
2	Gender
3-49	Diagnosis group counts 9 months before heart attack
50	Total cost 9 months before heart attack
51-97	Diagnosis group counts 6 months before heart attack
98	Total cost 6 months before heart attack
99-145	Diagnosis group counts 3 months before heart attack
146	Total cost 3 months before heart attack
147	Yes/No heart attack

Cost Bucket Partitioning

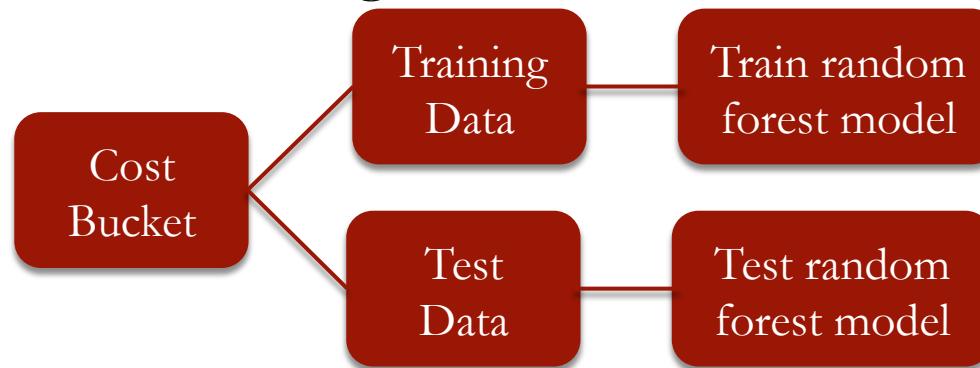
- Cost is a good summary of a person's overall health
- Divide population into similar smaller groups
 - Low risk, average risk, high risk

Bucket	Cost Range	% Data	Members	% with Heart Attack
1	< \$2K	67.56	4,416	36.14
2	\$2K - \$10K	21.56	1,409	43.22
3	> \$10K	10.88	711	38.12

- Build models for each group

Predicting Heart Attacks (Random Forest)

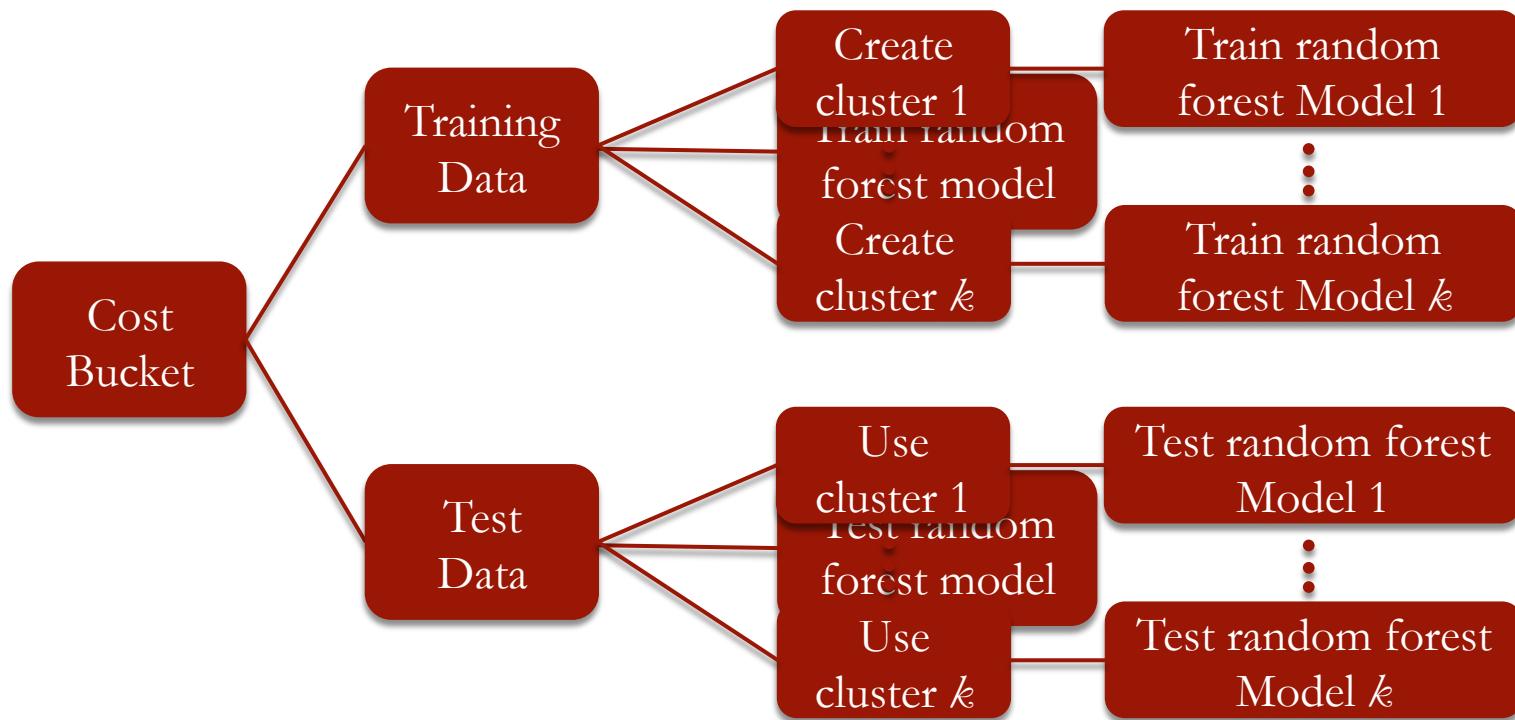
- Predicting whether a patient has a heart attack for each of the cost buckets using the random forest algorithm



Bucket	Random Forest
1	49.63%
2	55.99%
3	58.31%

Incorporating Clustering

- Patients in each bucket may have different characteristics



Clustering Cost Buckets



- Two clustering algorithms were used for the analysis as an alternative to hierachal clustering
 - Spectral Clustering
 - **k -means clustering**

Clustering Cost Buckets

- Two clustering algorithms were used for the analysis as an alternative to hierachal clustering
 - Spectral Clustering
 - ***k-means clustering***

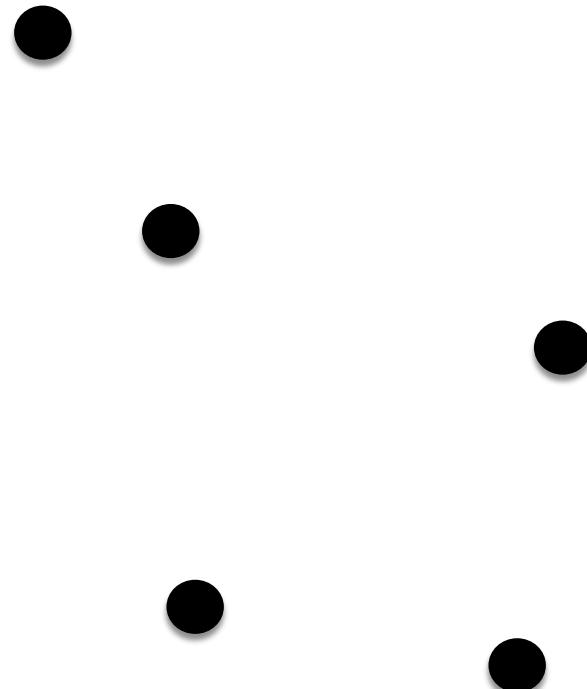
***k*-Means Clustering Algorithm**

1. Specify desired number of clusters k
2. Randomly assign each data point to a cluster
3. Compute cluster centroids
4. Re-assign each point to the closest cluster centroid
5. Re-compute cluster centroids
6. Repeat 4 and 5 until no improvement is made

k -Means Clustering

k -Means Clustering Algorithm

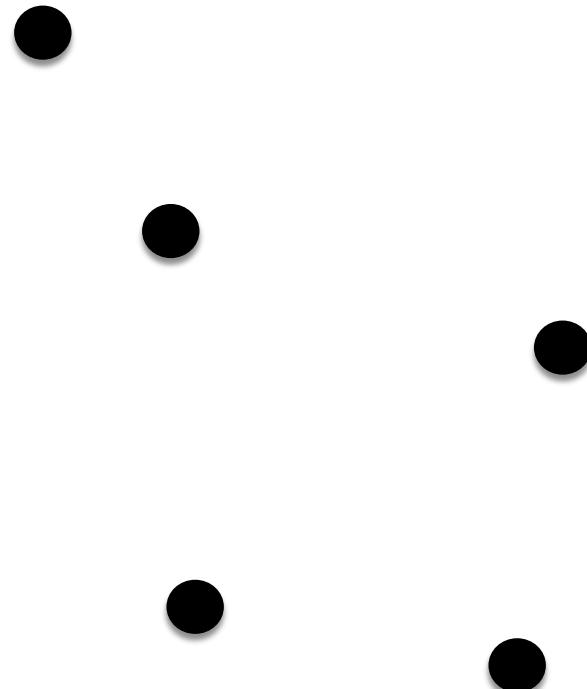
1. Specify desired number of clusters k



k -Means Clustering

k -Means Clustering Algorithm

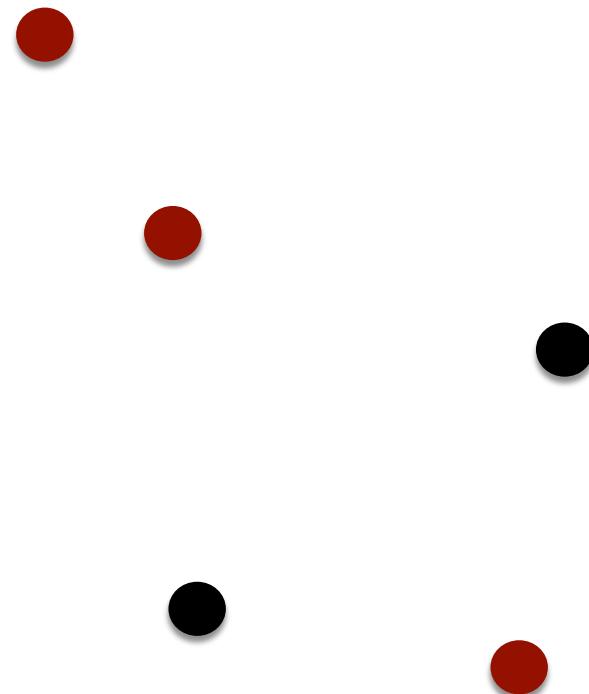
1. Specify desired number of clusters k
2. Randomly assign each data point to a cluster



k -Means Clustering

k -Means Clustering Algorithm

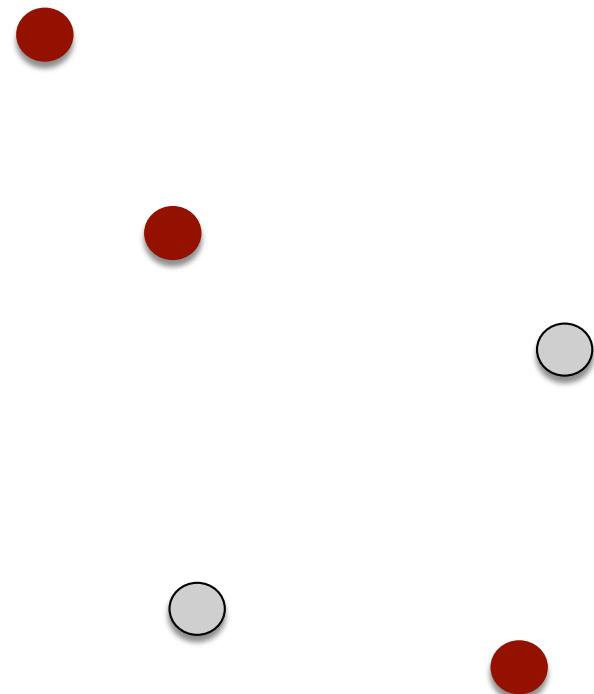
1. Specify desired number of clusters k
2. Randomly assign each data point to a cluster



k -Means Clustering

k -Means Clustering Algorithm

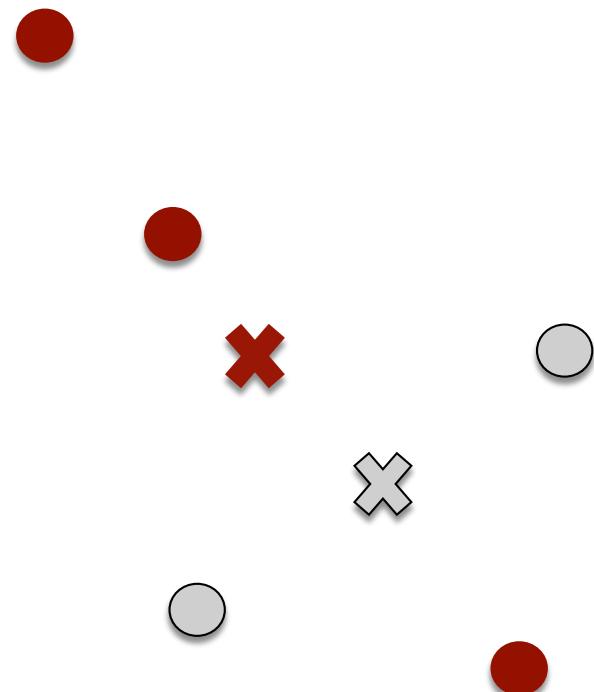
1. Specify desired number of clusters k
2. Randomly assign each data point to a cluster



k -Means Clustering

k -Means Clustering Algorithm

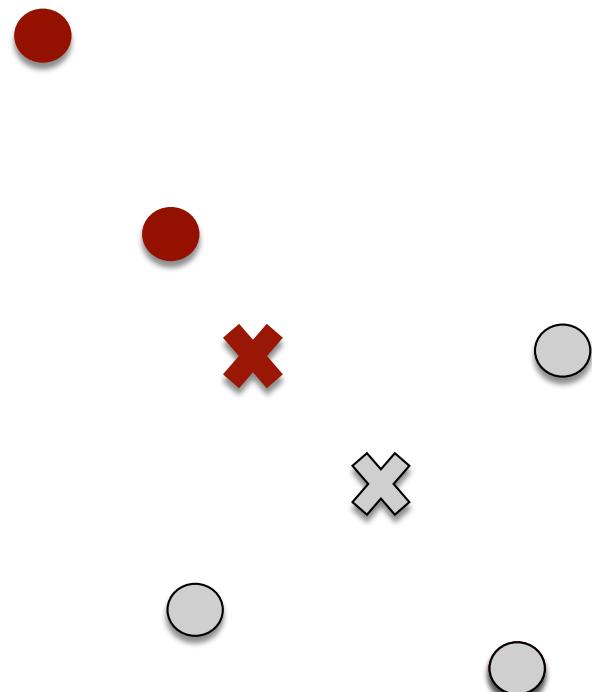
1. Specify desired number of clusters k
2. Randomly assign each data point to a cluster
3. Compute cluster centroids



k -Means Clustering

k -Means Clustering Algorithm

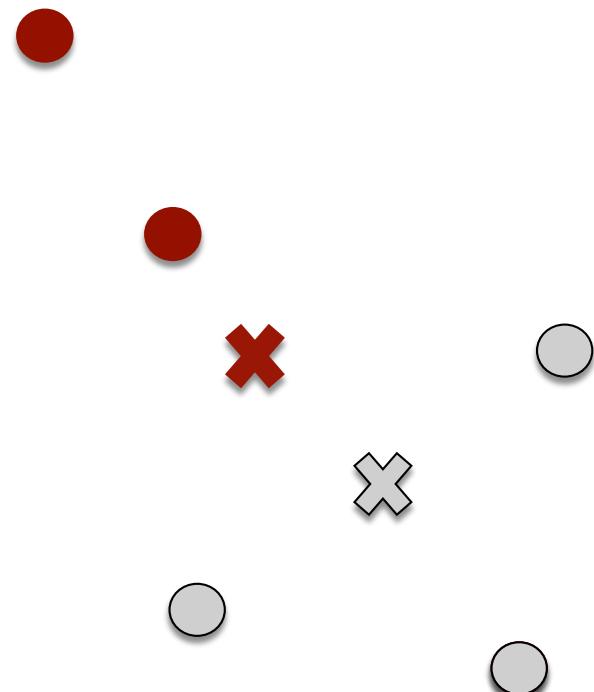
1. Specify desired number of clusters k
2. Randomly assign each data point to a cluster
3. Compute cluster centroids
4. Re-assign each point to the closest cluster centroid



k -Means Clustering

k -Means Clustering Algorithm

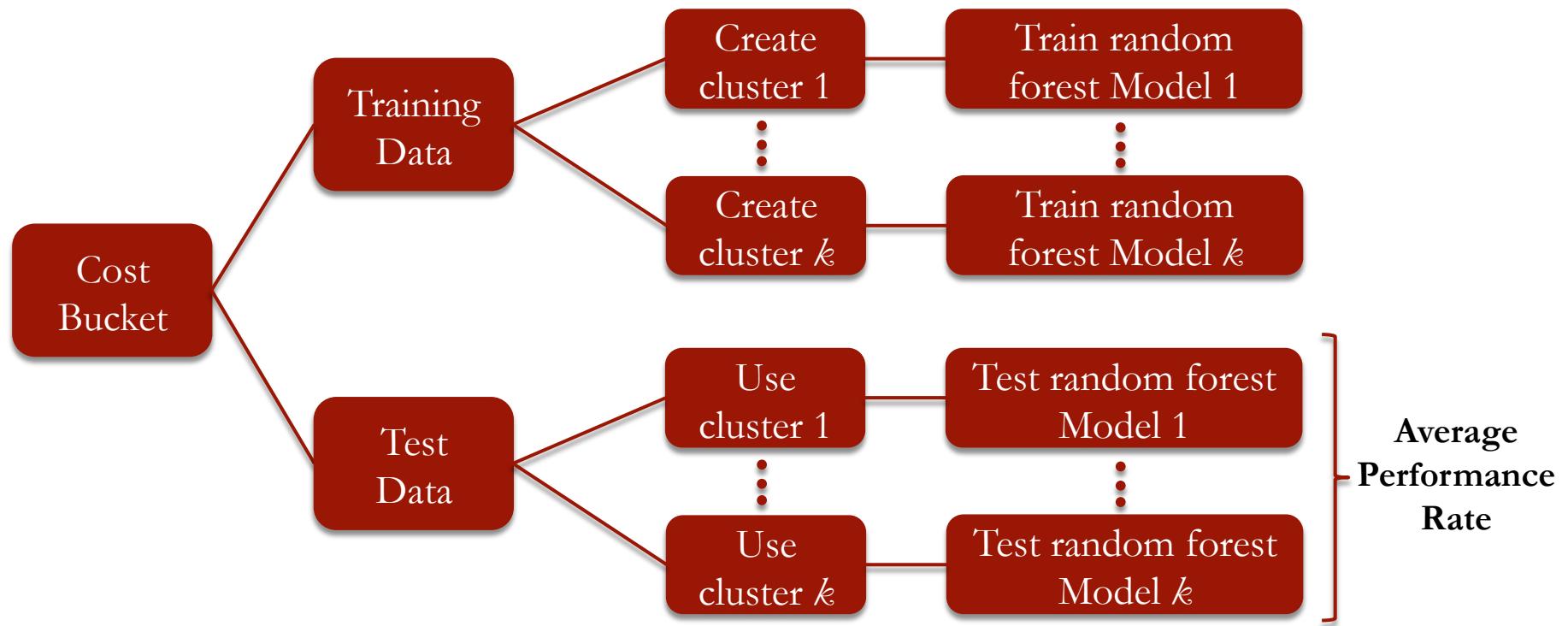
1. Specify desired number of clusters k
2. Randomly assign each data point to a cluster
3. Compute cluster centroids
4. Re-assign each point to the closest cluster centroid
5. Re-compute cluster centroids
6. Repeat 4 and 5 until no improvement is made



Practical Considerations

- The number of clusters k can be selected from previous knowledge or experimenting
- Can strategically select initial partition of points into clusters if you have some knowledge of the data
- Can run algorithm several times with different random starting points
- In recitation, we will learn how to run the k -means clustering algorithm in R

Random Forest with Clustering



Predicting Heart Attacks

- Perform clustering on each bucket using $k=10$ clusters
- Average prediction rate for each cost bucket

Cost Bucket	Random Forest without Clustering	Random Forest with Clustering
1	49.63%	64.75%
2	55.99%	72.93%
3	58.31%	78.25%

Understanding Cluster Patterns

- Clusters are interpretable and reveal unique patterns of diagnostic history among the population

Cost Bucket 2

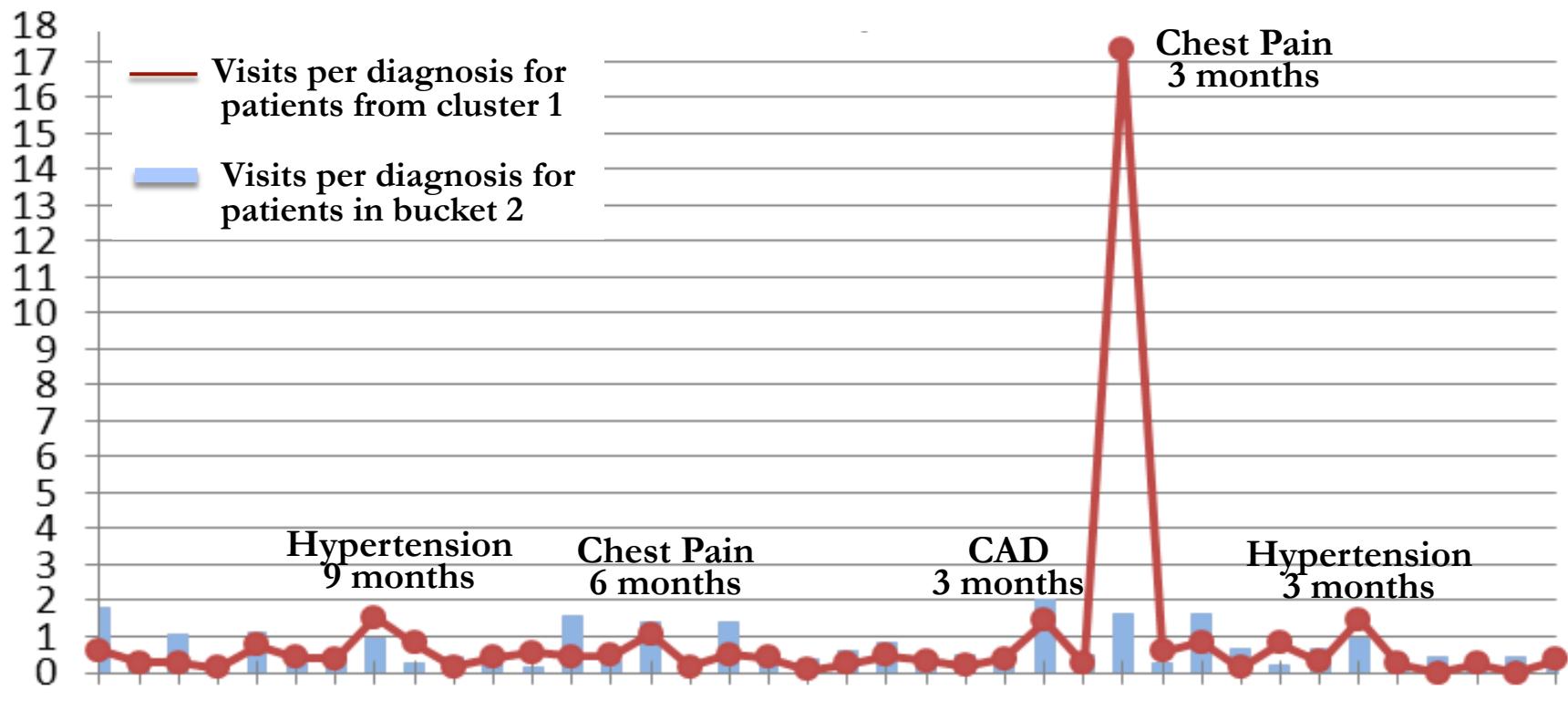
Cluster 1	Chest Pain (3 months)
Cluster 6	Coronary Artery Diseases (3 months)
Cluster 7	Chronic Obstructive Pulmonary Disease

Cost Bucket 3

Cluster 4	Anemia (3, 6, 9 months)
Cluster 5	Hypertension and Cerebrovascular Disease
Cluster 10	Diabetes (3, 6, 9 months)

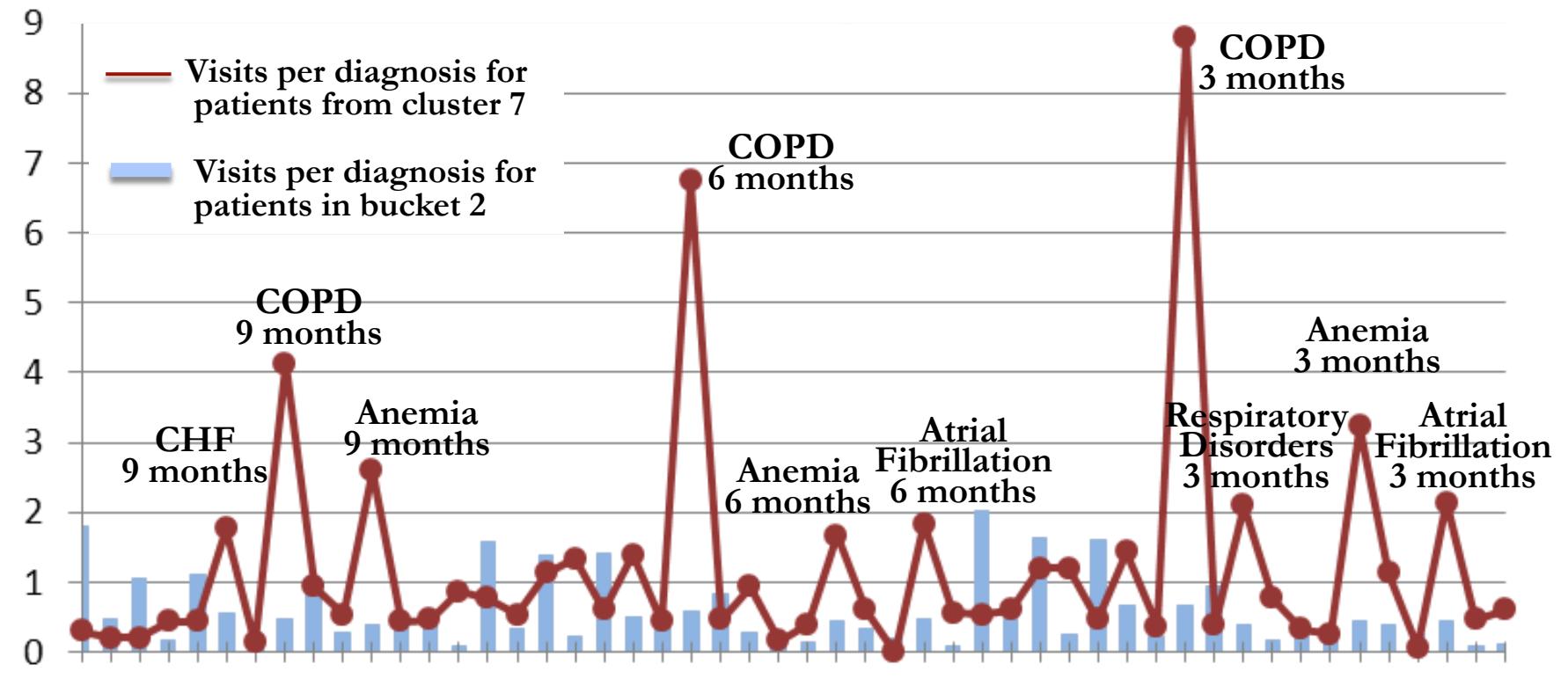
Occurrence of Chest Pain

- Cluster 1 in bucket 2 reflects a temporal pattern of chest pain diagnosed 3 months before a heart attack



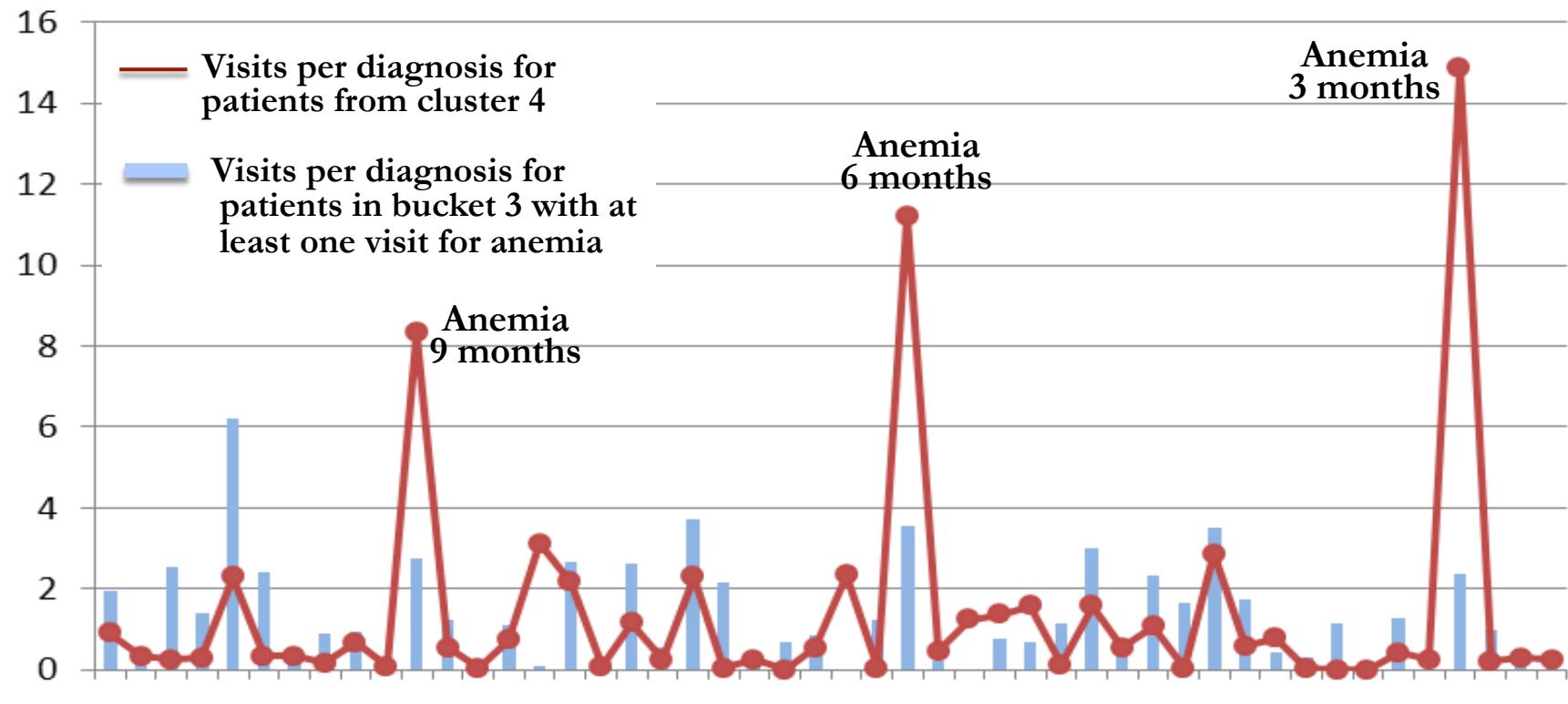
Chronic Obstructive Pulmonary Disease (COPD)

- Patients from Cluster 7 in cost bucket 2 who suffered a heart attack have regular doctor visits for COPD



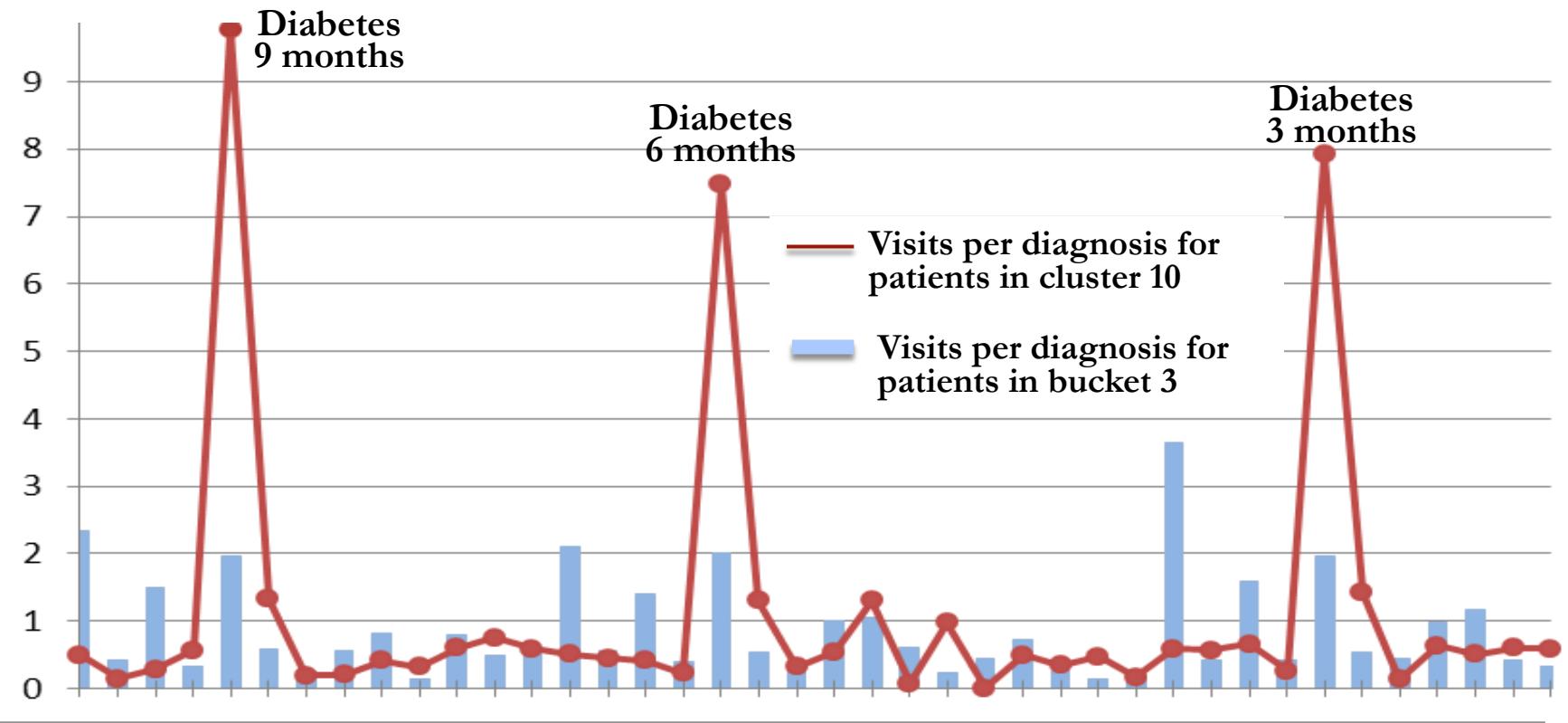
Gradually Increasing Occurrence of Anemia

- Cluster 4 in bucket 3 shows a temporal diagnosis pattern of anemia



Occurrence of Diabetes

- Cluster 10 in bucket 3 shows a temporal diagnosis of diabetes



Impact of Clustering



- Clustering members within each cost bucket yielded better predictions of heart attacks within clusters
- Grouping patients in clusters exhibits temporal diagnostic patterns within 9 months of a heart attack
- These patterns can be incorporated in the diagnostic rules for heart attacks
- Great research interest in using analytics for early heart failure detection through pattern recognition

Analytics for Early Detection



- IBM, Sutter Health and Geisinger Health System partnered in 2009 to research analytics tools in view of early detection

“Our earlier research showed that signs and symptoms of heart failure in patients are often documented **years before** a diagnosis”

“ The **pattern of documentation** can offer clinically useful **signals for early detection** of this deadly disease”

Steve Steinhubl (2013), Cardiologist MD from Geisenger



Recommendations Worth a Million

An Introduction to Clustering

15.071x – The Analytics Edge

Netflix

- Online DVD rental and streaming video service
- More than 40 million subscribers worldwide
- \$3.6 billion in revenue
- Key aspect is being able to offer customers accurate movie recommendations based on a customer's own preferences and viewing history



The Netflix Prize



- From 2006 – 2009 Netflix ran a contest asking the public to submit algorithms to predict user ratings for movies
- Training data set of ~100,000,000 ratings and test data set of ~3,000,000 ratings were provided
- Offered a grand prize of \$1,000,000 USD to the team who could beat Netflix's own algorithm, Cinematch, by more than 10%, measured in RMSE

Contest Rules

- If the grand prize was not yet reached, progress prizes of \$50,000 USD per year would be awarded for the best result so far, as long as it had $>1\%$ improvement over the previous year.
- Teams must submit code and a description of the algorithm to be awarded any prizes
- If any team met the 10% improvement goal, last call would be issued and 30 days would remain for all teams to submit their best algorithm.

Initial Results



- The contest went live on October 2, 2006
- By October 8, a team submitted an algorithm that beat Cinematch
- By October 15, there were three teams with algorithms beating Cinematch
- One of these solutions beat Cinematch by >1%, qualifying for a progress prize

Progress During the Contest



- By June 2007, over 20,000 teams had registered from over 150 countries
- The 2007 progress prize went to team BellKor, with an 8.43% improvement on Cinematch
- In the following year, several teams from across the world joined forces

Competition Intensifies



- The 2008 progress prize went to team BellKor which contained researchers from the original BellKor team as well as the team BigChaos
- This was the last progress prize because another 1% improvement would reach the grand prize goal of 10%

Last Call Announced

- On June 26, 2009, the team BellKor's Pragmatic Chaos submitted a 10.05% improvement over Cinematch

The screenshot shows the Netflix Prize Leaderboard. At the top, it displays "Leaderboard 10.05%" and an option to "Display top 20 leaders". A yellow arrow points to the "% Improvement" column header in the table below. The table has columns for Rank, Team Name, Best Score, % Improvement, and Last Submit Time. The top entry is for "BellKor's Pragmatic Chaos" with a score of 0.8558 and an improvement of 10.05%. Below this, there is a red banner for the "Grand Prize - RMSE <= 0.8563". The remaining entries in the table are: 2. PragmaticTheory (0.8582), 3. BellKor in BigChaos (0.8590), 4. Grand Prize Team (0.8593), 5. Dace (0.8604), and 6. BigChaos (0.8613).

Rank	Team Name	Best Score	% Improvement	Last Submit Time
1	BellKor's Pragmatic Chaos	0.8558	10.05	2009-06-26 18:42:37
Grand Prize - RMSE <= 0.8563				
2	PragmaticTheory	0.8582	9.80	2009-06-25 22:15:51
3	BellKor in BigChaos	0.8590	9.71	2009-05-13 08:14:09
4	Grand Prize Team	0.8593	9.68	2009-06-12 08:20:24
5	Dace	0.8604	9.56	2009-04-22 05:57:03
6	BigChaos	0.8613	9.47	2009-06-23 23:06:52

Predicting the Best User Ratings



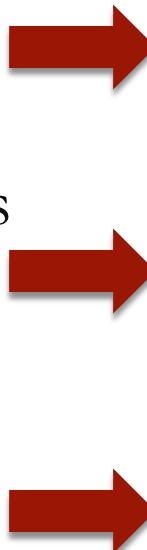
- Netflix was willing to pay over \$1M for the best user rating algorithm, which shows how critical the recommendation system was to their business
- What data could be used to predict user ratings?
- Every movie in Netflix's database has the ranking from all users who have ranked that movie
- We also know facts about the movie itself: actors, director, genre classifications, year released, etc.

Using Other Users' Rankings

	Men in Black	Apollo 13	Top Gun	Terminator
Amy	5	4	5	4
Bob	3		2	5
Carl		5	4	4
Dan	4	2		

- Consider suggesting to Carl that he watch “Men in Black”, since Amy rated it highly and Carl and Amy seem to have similar preferences
- This technique is called **Collaborative Filtering**

Using Movie Information

- We saw that Amy liked “Men In Black”
 - It was directed by Barry Sonnenfeld
 - Classified in the genres of action, adventure, sci-fi and comedy
 - It stars actor Will Smith
 - Consider recommending to Amy:
 - Barry Sonnenfeld’s movie “Get Shorty”
 - “Jurassic Park”, which is in the genres of action, adventure, and sci-fi
 - Will Smith’s movie “Hitch”
- 

This technique is called **Content Filtering**

Strengths and Weaknesses



- Collaborative Filtering Systems
 - Can accurately suggest complex items without understanding the nature of the items
 - Requires a lot of data about the user to make accurate recommendations
 - Millions of items – need lots of computing power
- Content Filtering
 - Requires very little data to get started
 - Can be limited in scope

Hybrid Recommendation Systems

- Netflix uses both collaborative and content filtering
- For example, consider a collaborative filtering approach where we determine that Amy and Carl have similar preferences.
- We could then do content filtering, where we would find that “Terminator”, which both Amy and Carl liked, is classified in almost the same set of genres as “Starship Troopers”
- Recommend “Starship Troopers” to both Amy and Carl, even though neither of them have seen it before

MovieLens Data

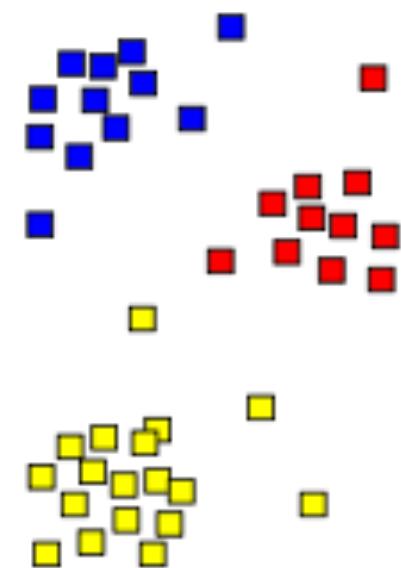
- www.movielens.org is a movie recommendation website run by the GroupLens Research Lab at the University of Minnesota
- They collect user preferences about movies and do collaborative filtering to make recommendations
- We will use their movie database to do content filtering using a technique called clustering

MovieLens Item Dataset

- Movies in the dataset are categorized as belonging to different genres
 - (Unknown)
 - Comedy
 - Film Noir
 - Sci-Fi
 - Action
 - Crime
 - Horror
 - Thriller
 - Adventure
 - Documentary
 - Musical
 - War
 - Animation
 - Drama
 - Mystery
 - Western
 - Children's
 - Fantasy
 - Romance
- Each movie may belong to many genres
- Can we systematically find groups of movies with similar sets of genres?

Why Clustering?

- “Unsupervised” learning
 - Goal is to segment the data into similar groups instead of prediction
- Can also cluster data into “similar” groups and then build a predictive model for each group
 - Be careful not to overfit your model!
This works best with large datasets



Types of Clustering Methods



- There are many different algorithms for clustering
 - Differ in what makes a cluster and how to find them
- We will cover
 - Hierarchical
 - K-means in the next lecture

Distance Between Points

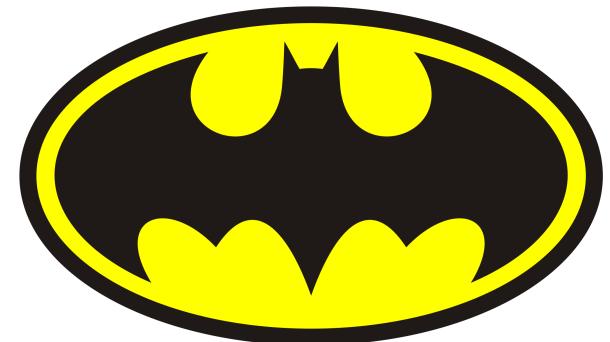
- Need to define distance between two data points
 - Most popular is “Euclidean distance”
 - Distance between points i and j is

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \dots + (x_{ik} - x_{jk})^2}$$

where k is the number of independent variables

Distance Example

- The movie “Toy Story” is categorized as Animation, Comedy, and Children’s
 - Toy Story:
(0,0,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0)
- The movie “Batman Forever” is categorized as Action, Adventure, Comedy, and Crime
 - Batman Forever:
(0,1,1,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0)



Distance Between Points

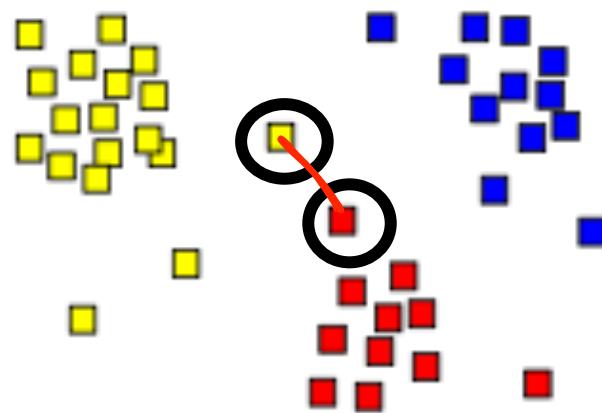
- Toy Story: (0,0,0,1,1,1,0,0,0,0,0,0,0,0,0,0)
- Batman Forever: (0,1,1,0,0,1,1,0,0,0,0,0,0,0,0,0)

$$d = \sqrt{(0-0)^2 + (0-1)^2 + (0-1)^2 + (1-0)^2 + \dots}$$
$$= \sqrt{5}$$

- Other popular distance metrics:
 - Manhattan Distance
 - Sum of absolute values instead of squares
 - Maximum Coordinate Distance
 - Only consider measurement for which data points deviate the most

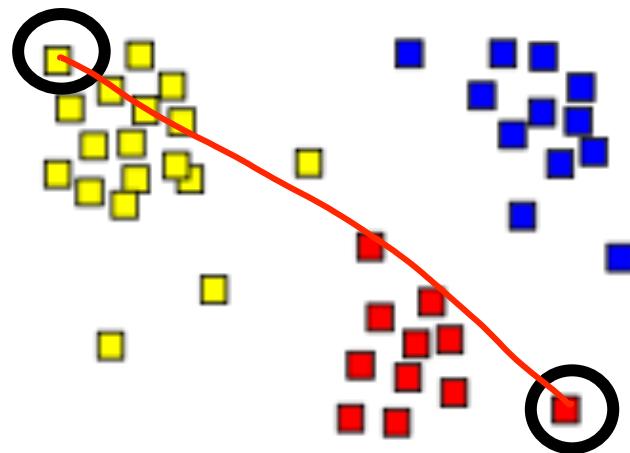
Distance Between Clusters

- Minimum Distance
 - Distance between clusters is the distance between points that are the closest



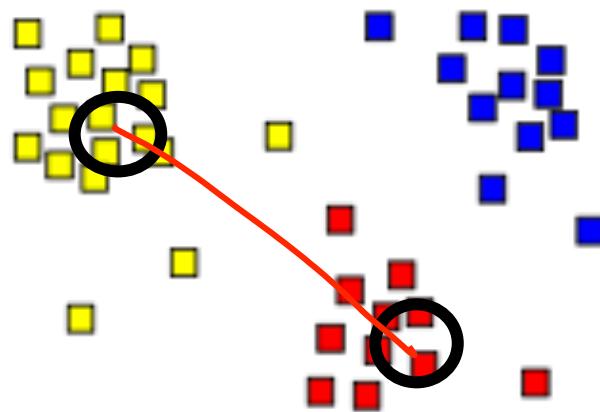
Distance Between Clusters

- Maximum Distance
 - Distance between clusters is the distance between points that are the farthest



Distance Between Clusters

- Centroid Distance
 - Distance between centroids of clusters
 - Centroid is point that has the average of all data points in each component



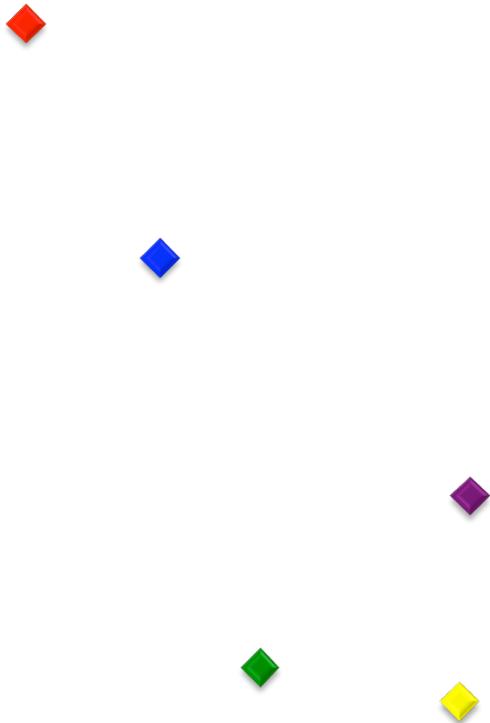
Normalize Data



- Distance is highly influenced by scale of variables, so customary to normalize first
- In our movie dataset, all genre variables are on the same scale and so normalization is not necessary
- However, if we included a variable such as “Box Office Revenue,” we would need to normalize.

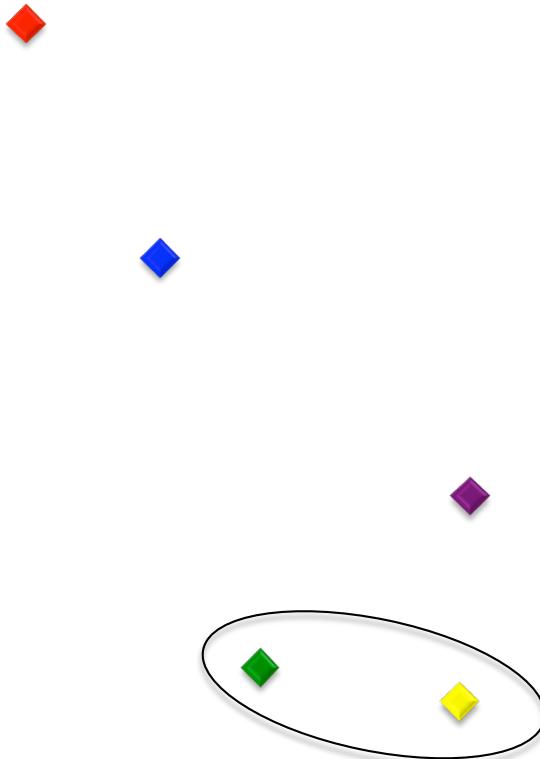
Hierarchical

- Start with each data point in its own cluster



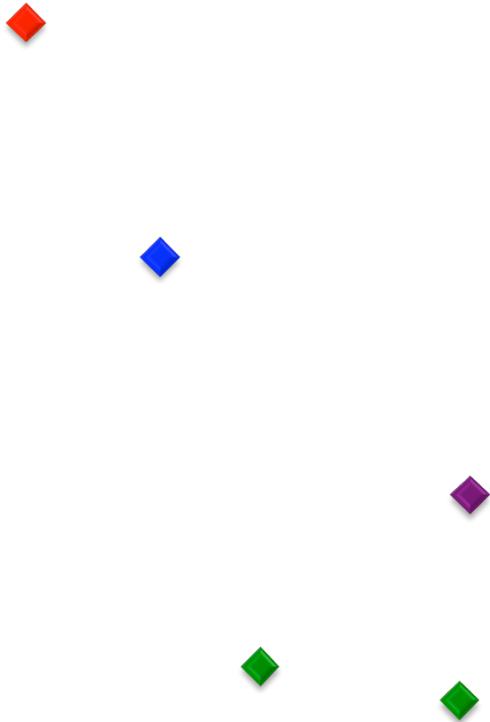
Hierarchical

- Combine two nearest clusters (Euclidean, Centroid)



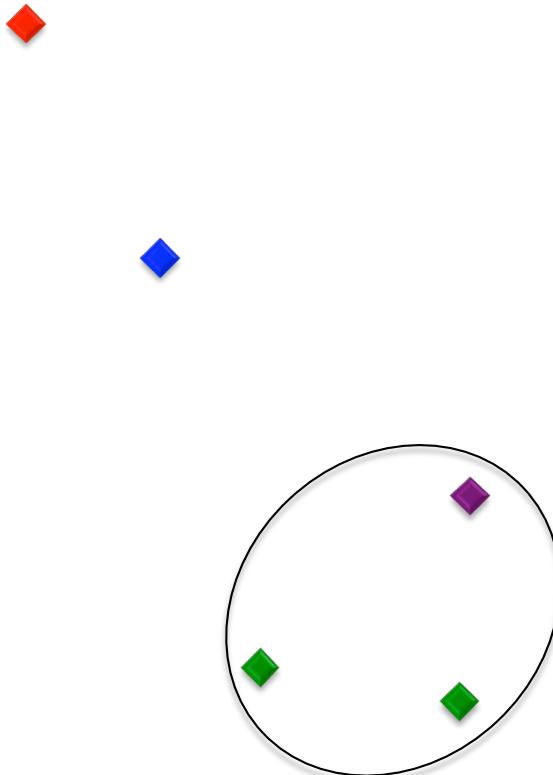
Hierarchical

- Combine two nearest clusters (Euclidean, Centroid)



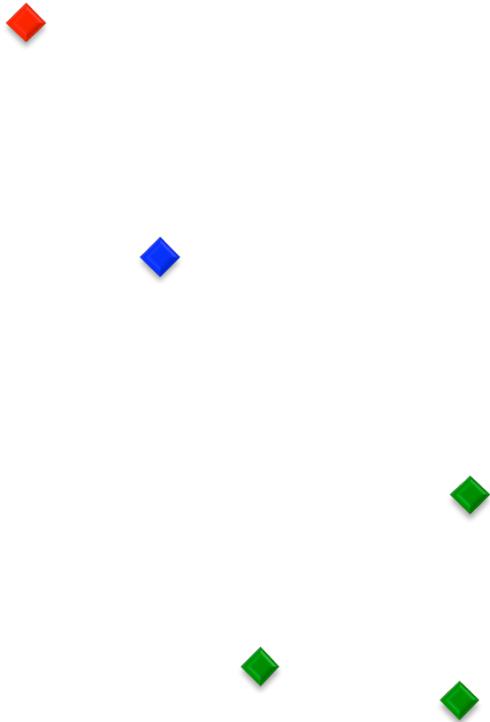
Hierarchical

- Combine two nearest clusters (Euclidean, Centroid)



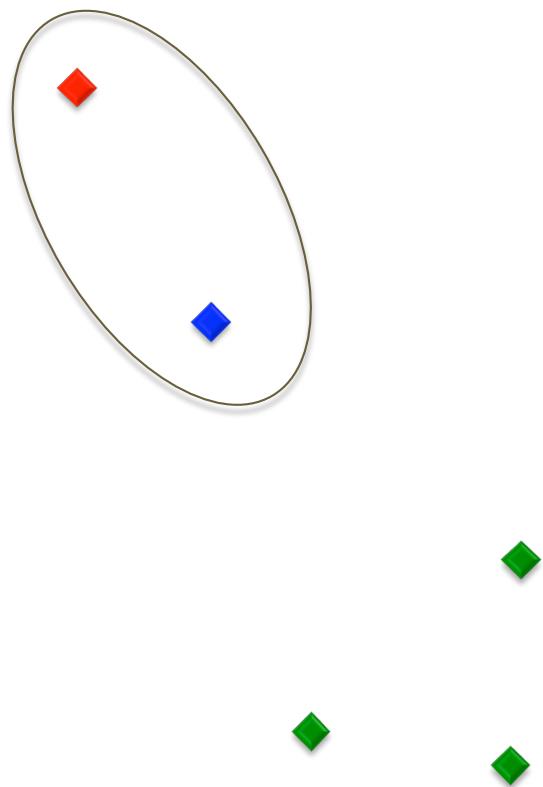
Hierarchical

- Combine two nearest clusters (Euclidean, Centroid)



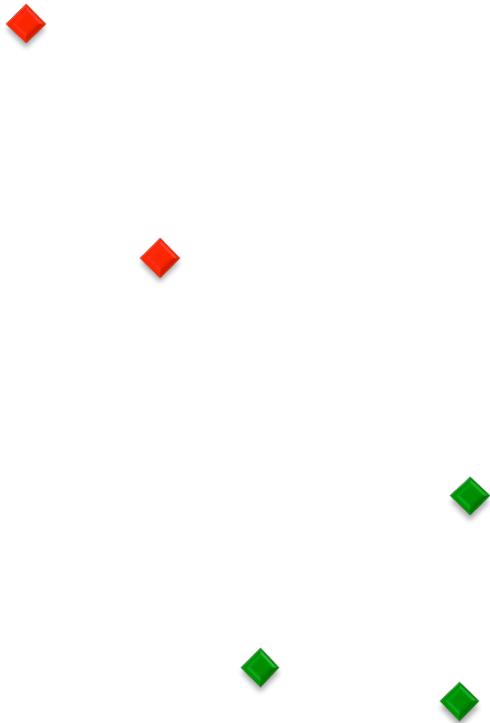
Hierarchical

- Combine two nearest clusters (Euclidean, Centroid)



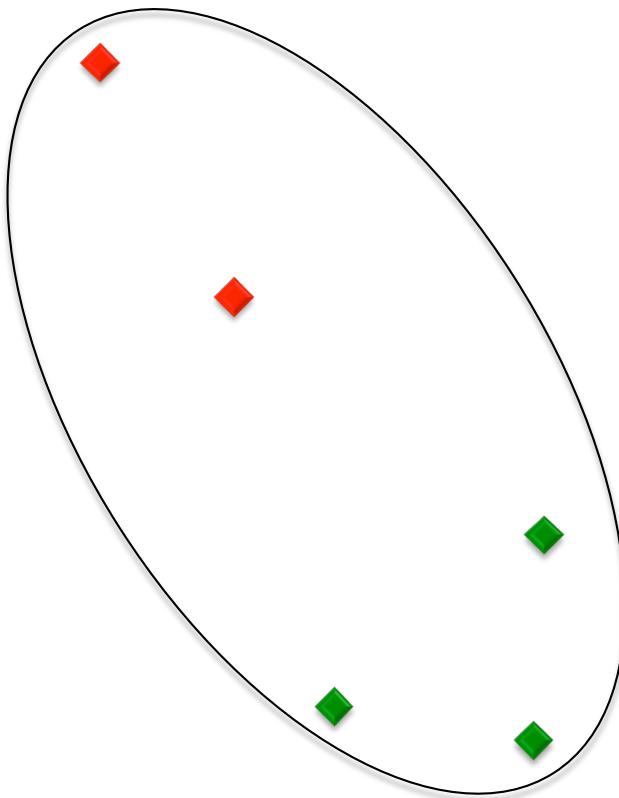
Hierarchical

- Combine two nearest clusters (Euclidean, Centroid)



Hierarchical

- Combine two nearest clusters (Euclidean, Centroid)



Hierarchical

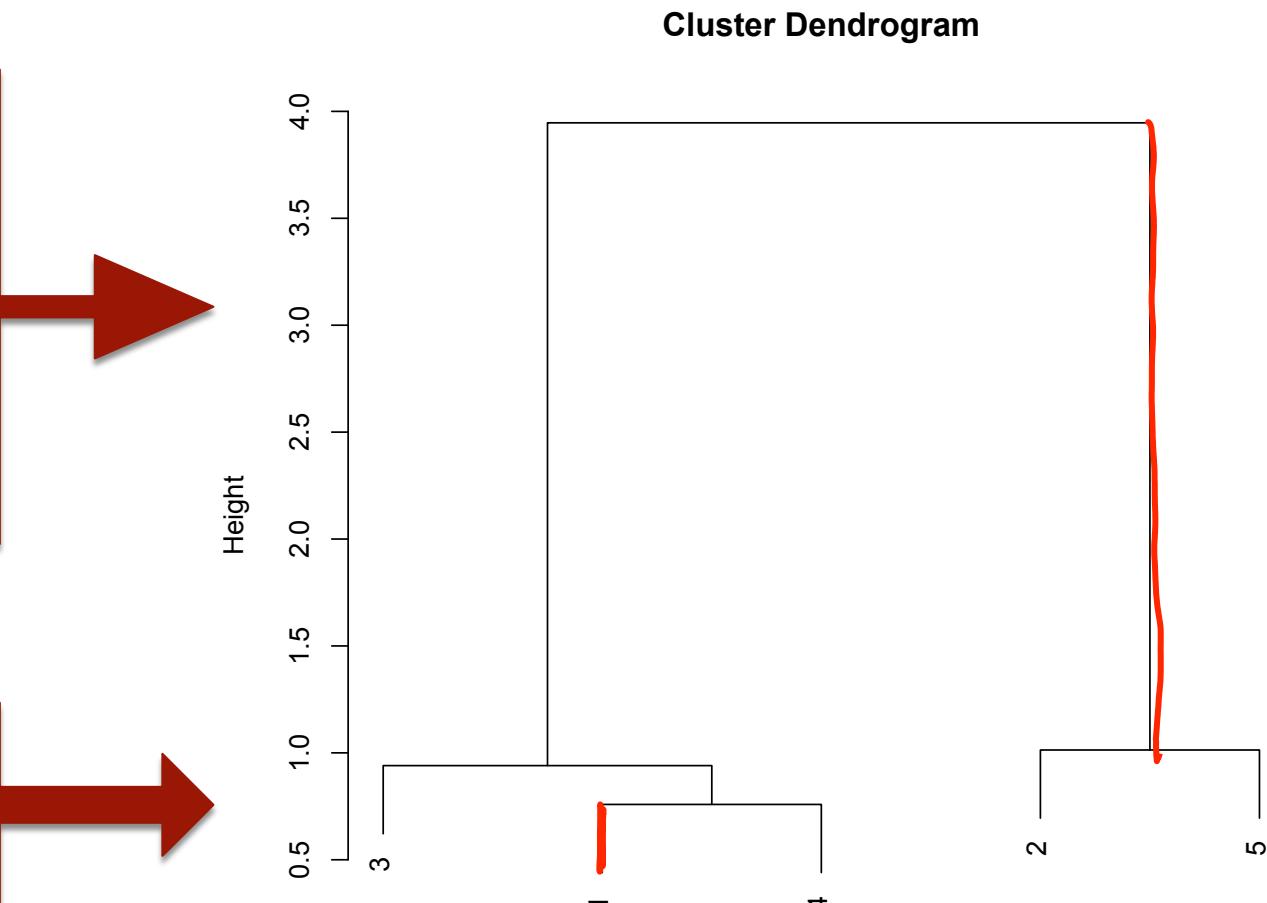
- Combine two nearest clusters (Euclidean, Centroid)



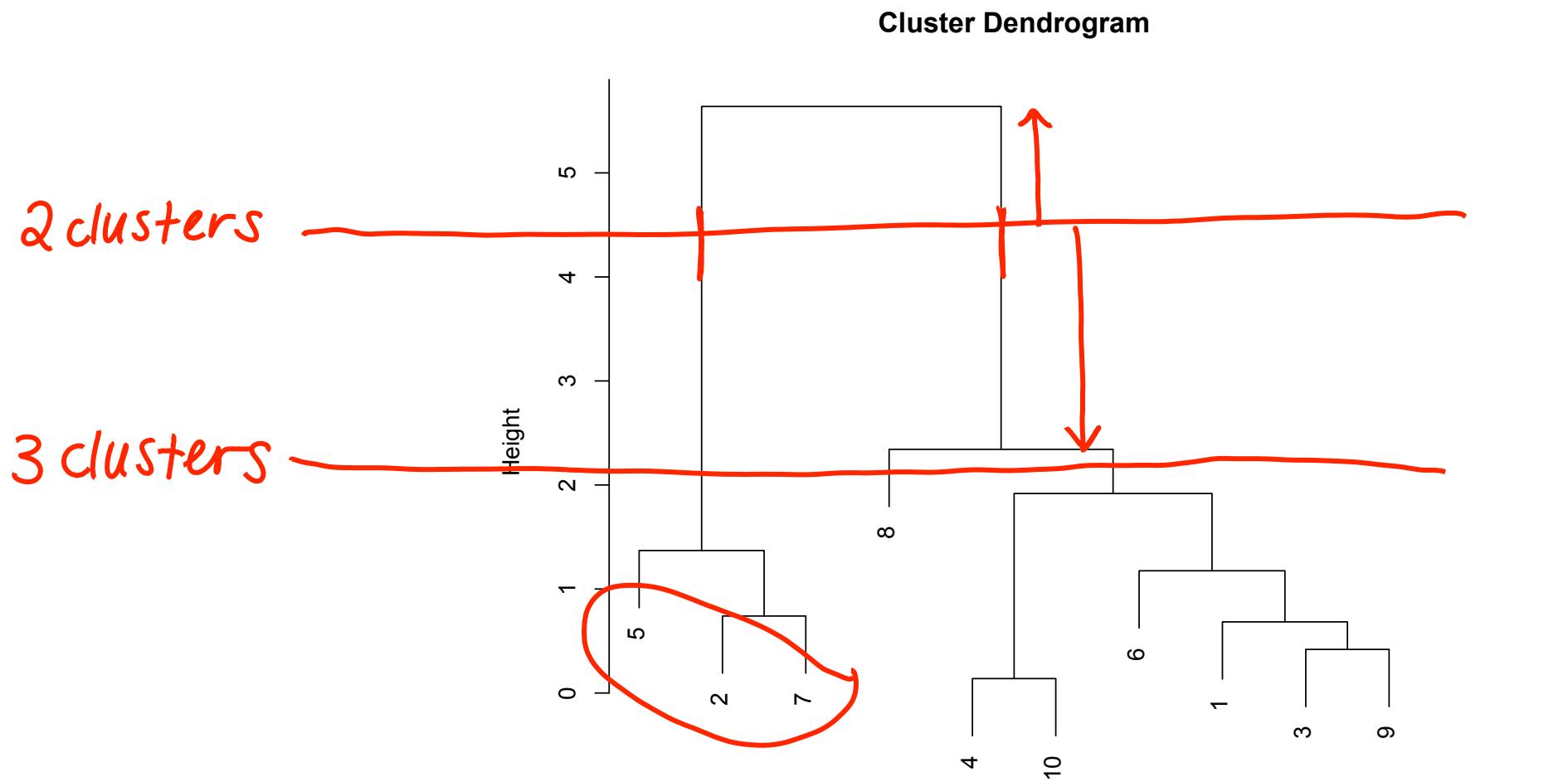
Display Cluster Process

Height of vertical lines represents distance between points or clusters

Data points listed along bottom



Select Clusters



Meaningful Clusters?

- Look at statistics (mean, min, max, . . .) for each cluster and each variable
- See if the clusters have a feature in common that was not used in the clustering (like an outcome)

Beyond Movies: Mass Personalization



- “If I have 3 million customers on the web, I should have 3 million stores on the web”
– Jeff Bezos, CEO of Amazon.com
- Recommendation systems build models about users’ preferences to personalize the user experience
- Help users find items they might not have searched for:
 - A new favorite band
 - An old friend who uses the same social media network
 - A book or song they are likely to enjoy

Cornerstone of these Top Businesses



Recommendation Method Used

- Collaborative Filtering
 - Amazon.com
 - Last.fm
 - Spotify
 - Facebook
 - LinkedIn
 - Google News
 - MySpace
 - **Netflix**
- Content Filtering
 - Pandora
 - IMDB
 - Rotten Tomatoes
 - Jinni
 - Rovi Corporation
 - See This Next
 - MovieLens
 - **Netflix**

The Netflix Prize: The Final 30 Days



- 29 days after last call was announced, on July 25, 2009, the team The Ensemble submitted a 10.09% improvement
- When Netflix stopped accepting submissions the next day, BellKor's Pragmatic Chaos had submitted a 10.09% improvement solution and The Ensemble had submitted a 10.10% improvement solution
- Netflix would now test the algorithms on a private test set and announce the winners

Winners are Declared!

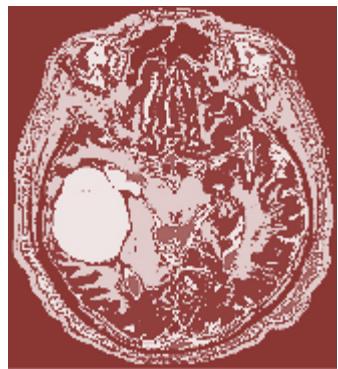
- On September 18, 2009, a winning team was announced
- BellKor's Pragmatic Chaos won the competition and the \$1,000,000 grand prize



The Edge of Recommendation Systems



- In today's digital age, businesses often have hundreds of thousands of items to offer their customers
- Excellent recommendation systems can make or break these businesses
- Clustering algorithms, which are tailored to find similar customers or similar items, form the backbone of many of these recommendation systems



Seeing the Big Picture

Segmenting Images to Create Data

15.071x – The Analytics Edge

Image Segmentation



- Divide up digital images to salient regions/clusters corresponding to individual surfaces, objects, or natural parts of objects
- Clusters should be uniform and homogenous with respect to certain characteristics (color, intensity, texture)
- Goal: Useful and analyzable image representation

Wide Applications



- Medical Imaging
 - Locate tissue classes, organs, pathologies and tumors
 - Measure tissue/tumor volume
- Object Detection
 - Detect facial features in photos
 - Detect pedestrians in footages of surveillance videos
- Recognition tasks
 - Fingerprint/Iris recognition

Various Methods

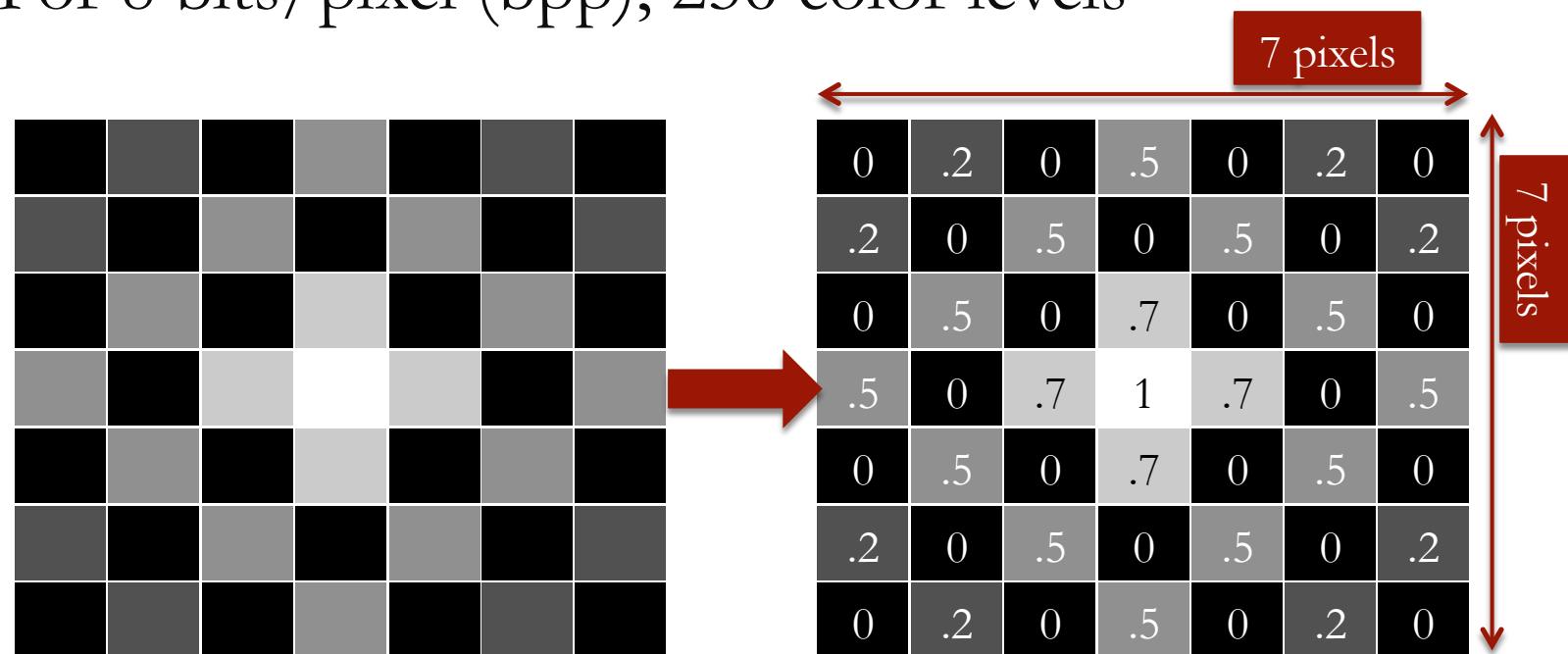
- Clustering methods
 - Partition image to clusters based on differences in pixel colors, intensity or texture
- Edge detection
 - Based on the detection of discontinuity, such as an abrupt change in the gray level in gray-scale images
- Region-growing methods
 - Divides image into regions, then sequentially merges sufficiently similar regions

In this Recitation...

- Review **hierarchical** and **k -means** clustering in R
- Restrict ourselves to gray-scale images
 - Simple example of a flower image (flower.csv)
 - Medical imaging application with examples of transverse MRI images of the brain (healthy.csv and tumor.csv)
- Compare the use, pros and cons of all analytics methods we have seen so far

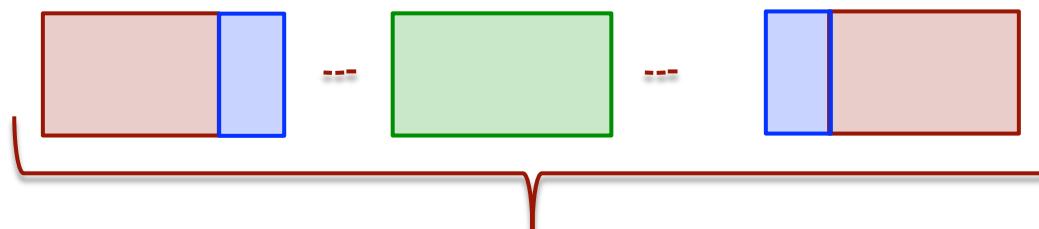
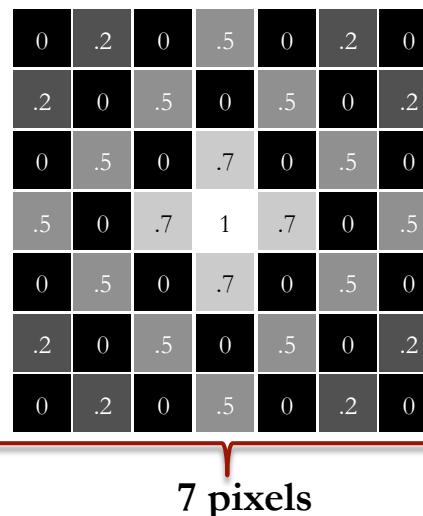
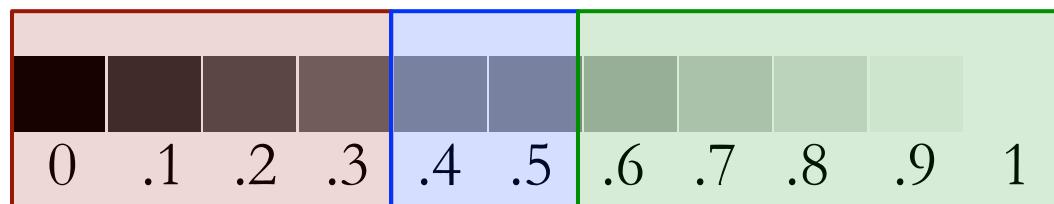
Grayscale Images

- Image is represented as a matrix of pixel intensity values ranging from 0 (black) to 1 (white)
- For 8 bits/pixel (bpp), 256 color levels



Grayscale Image Segmentation

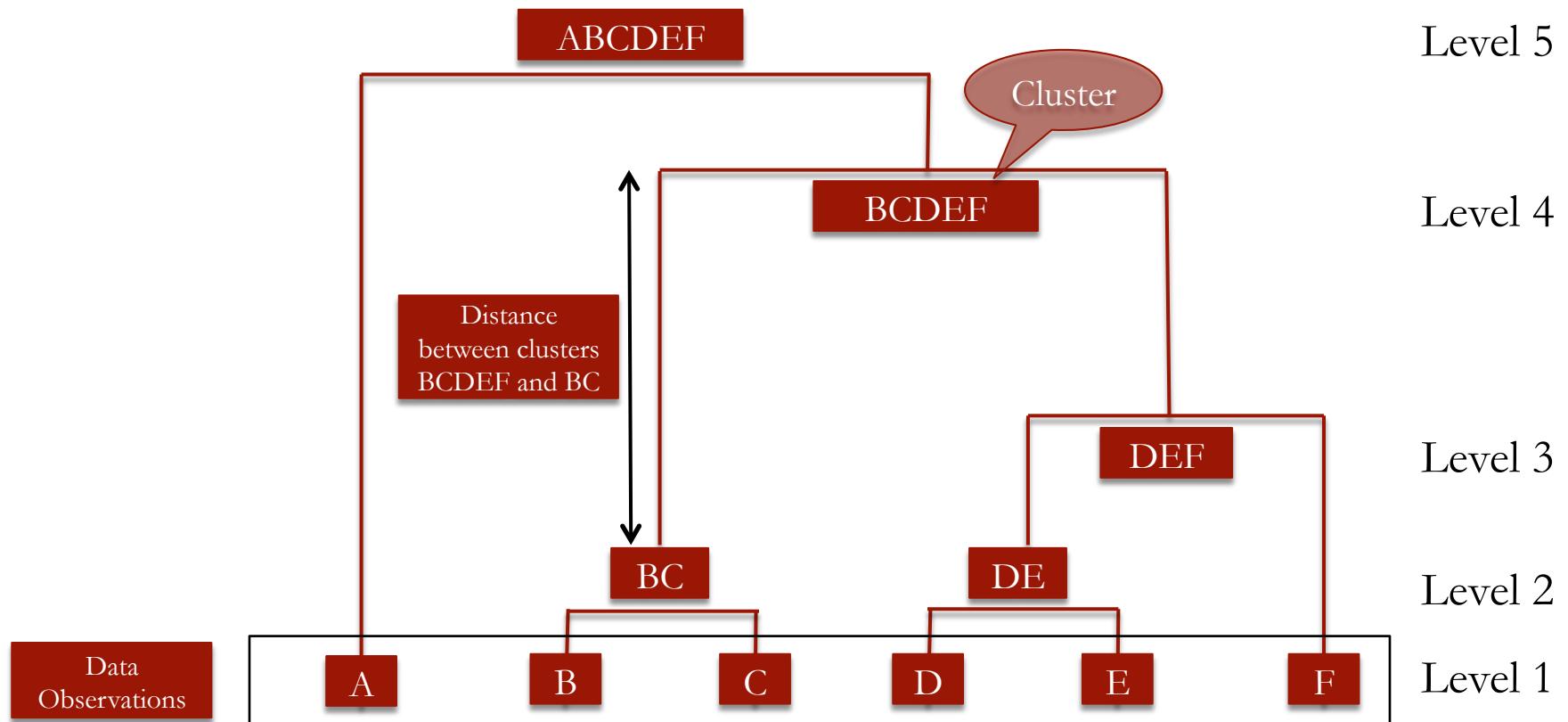
- Cluster pixels according to their intensity values



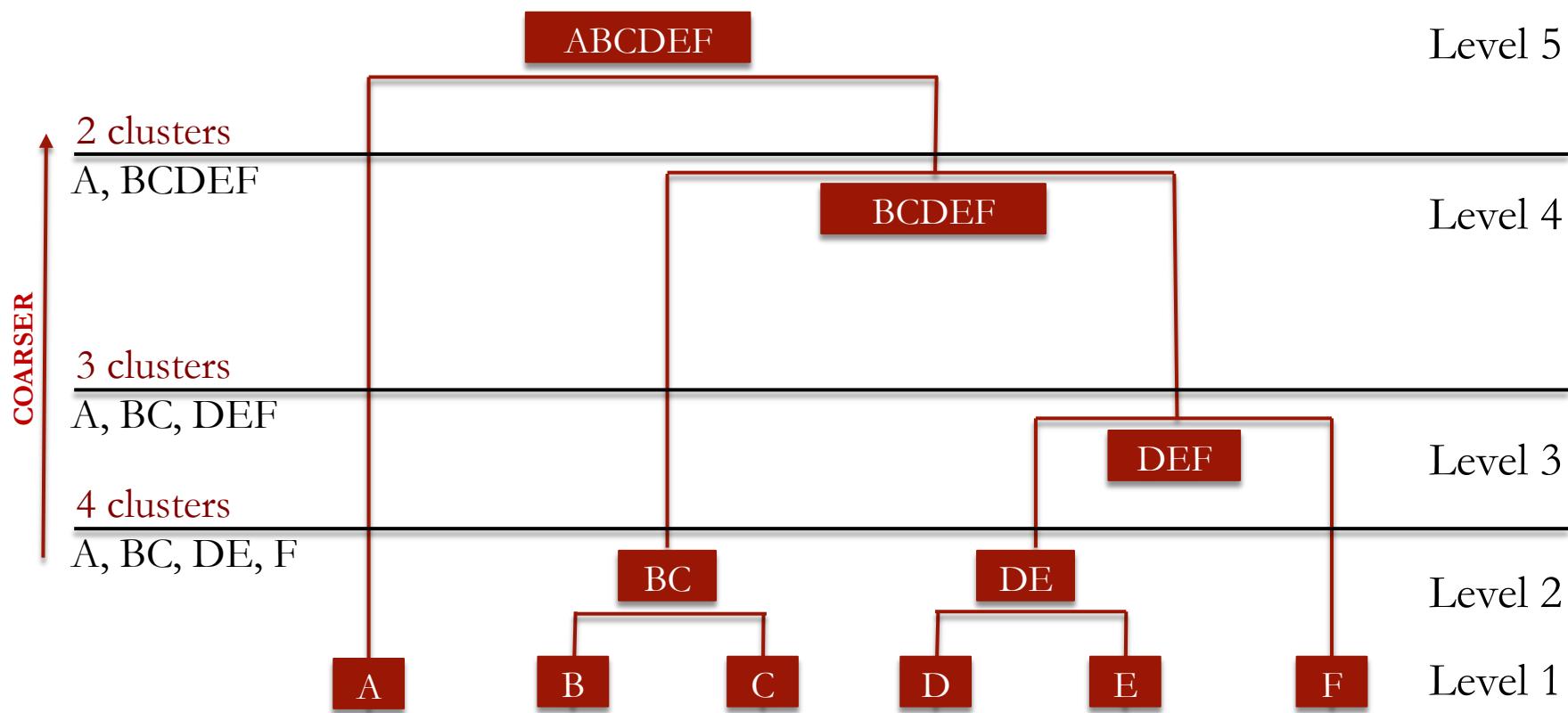
Intensity Vector of size $n = 7 \times 7$

Pairwise distances $n(n-1)/2$

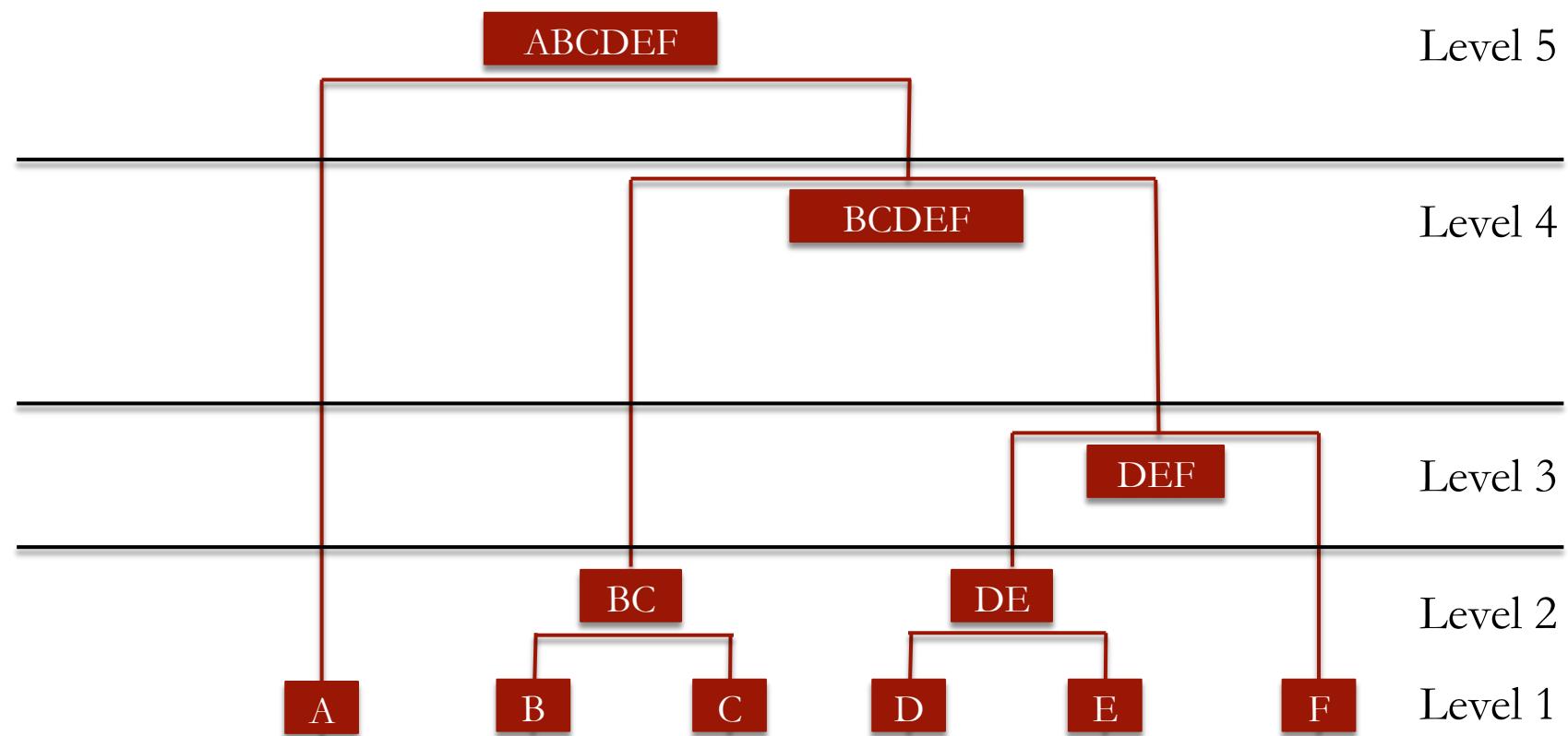
Dendrogram Example



Dendrogram Example

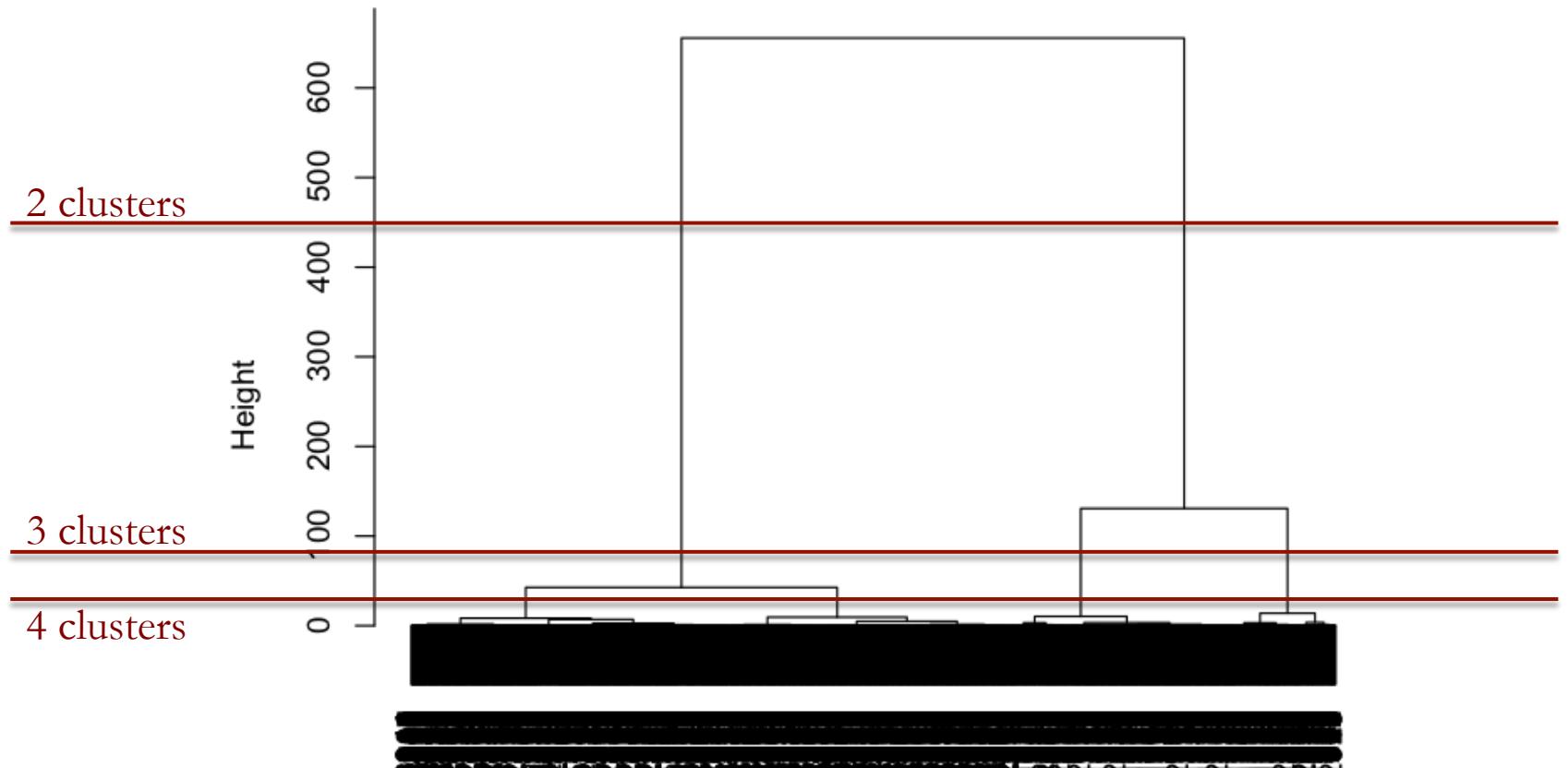


Dendrogram Example



Flower Dendrogram

Cluster Dendrogram



k -Means Clustering

- The k -means clustering aims at partitioning the data into k clusters in which each data point belongs to the cluster whose mean is the nearest

k -Means Clustering Algorithm

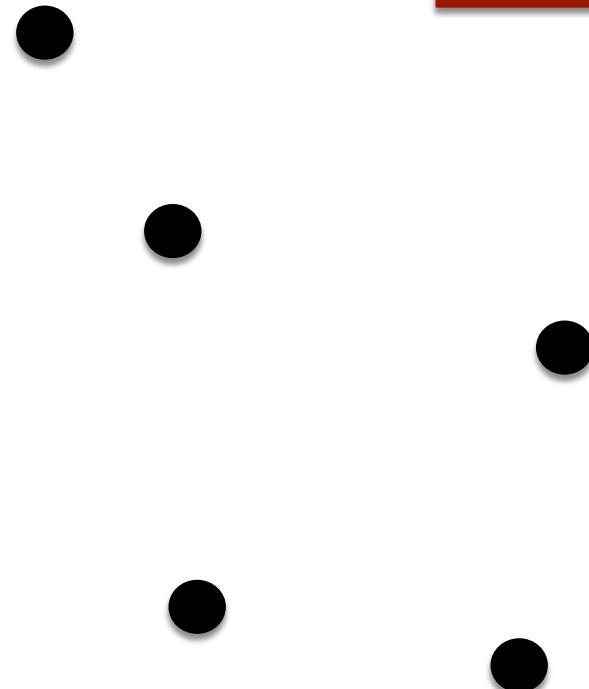
1. Specify desired number of clusters k
2. Randomly assign each data point to a cluster
3. Compute cluster centroids
4. Re-assign each point to the closest cluster centroid
5. Re-compute cluster centroids
6. Repeat 4 and 5 until no improvement is made

k -Means Clustering

k -Means Clustering Algorithm

1. Specify desired number of clusters k

$k = 2$

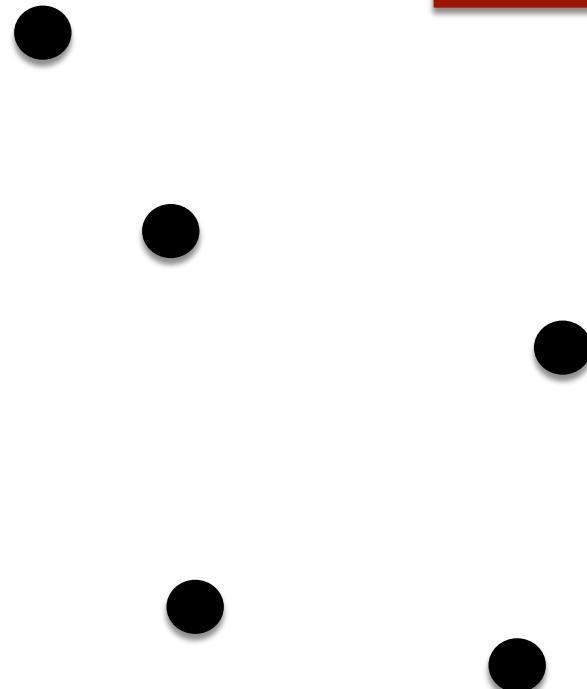


k -Means Clustering

k -Means Clustering Algorithm

1. Specify desired number of clusters k
2. Randomly assign each data point to a cluster

$k = 2$

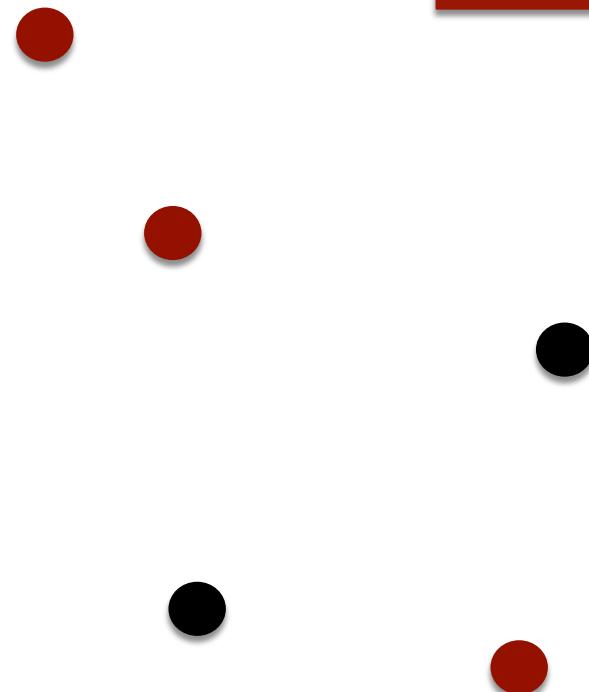


k -Means Clustering

k -Means Clustering Algorithm

1. Specify desired number of clusters k
2. Randomly assign each data point to a cluster

$k = 2$

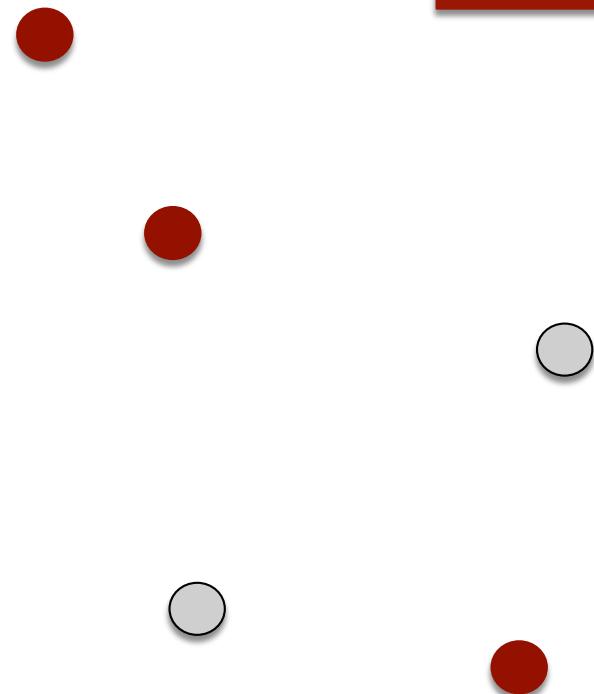


k -Means Clustering

k -Means Clustering Algorithm

1. Specify desired number of clusters k
2. Randomly assign each data point to a cluster

$k = 2$

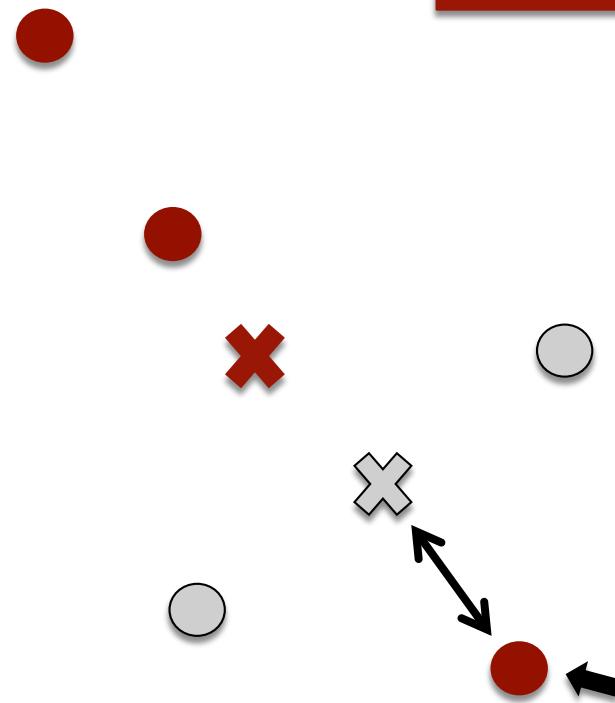


k -Means Clustering

k -Means Clustering Algorithm

1. Specify desired number of clusters k
2. Randomly assign each data point to a cluster
3. Compute cluster centroids
4. Re-assign each point to the closest cluster centroid

$k = 2$

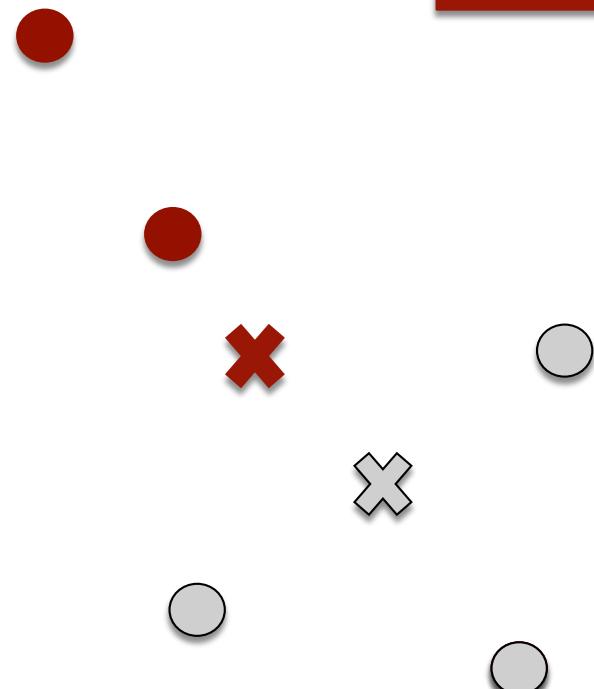


k -Means Clustering

k -Means Clustering Algorithm

1. Specify desired number of clusters k
2. Randomly assign each data point to a cluster
3. Compute cluster centroids
4. Re-assign each point to the closest cluster centroid

$k = 2$

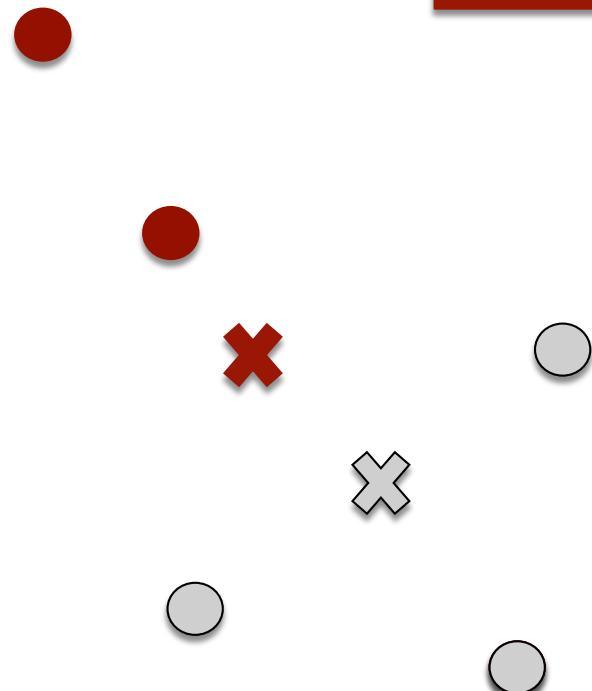


k -Means Clustering

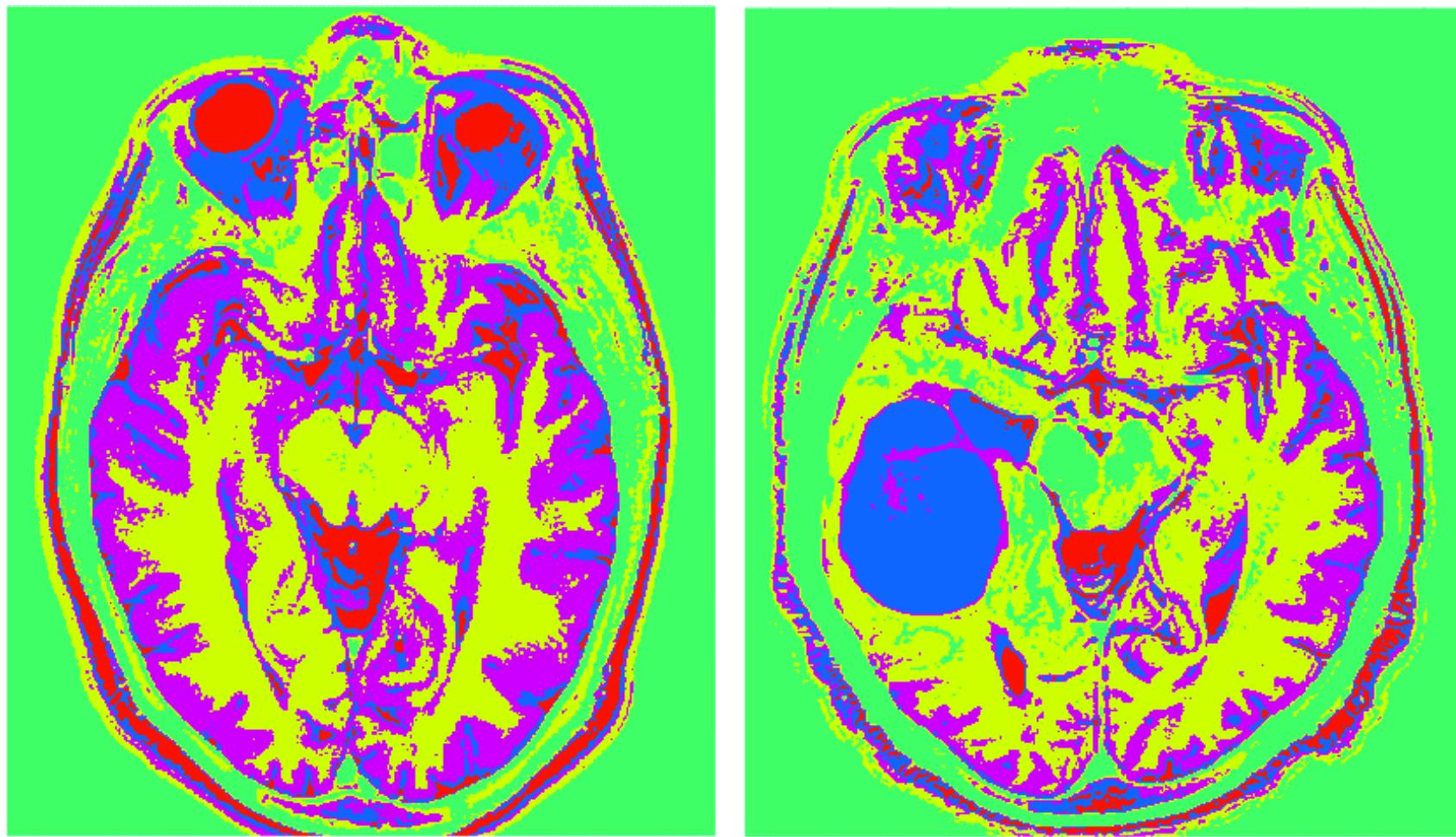
k -Means Clustering Algorithm

1. Specify desired number of clusters k
2. Randomly assign each data point to a cluster
3. Compute cluster centroids
4. Re-assign each point to the closest cluster centroid
5. Re-compute cluster centroids
6. Repeat 4 and 5 until no improvement is made

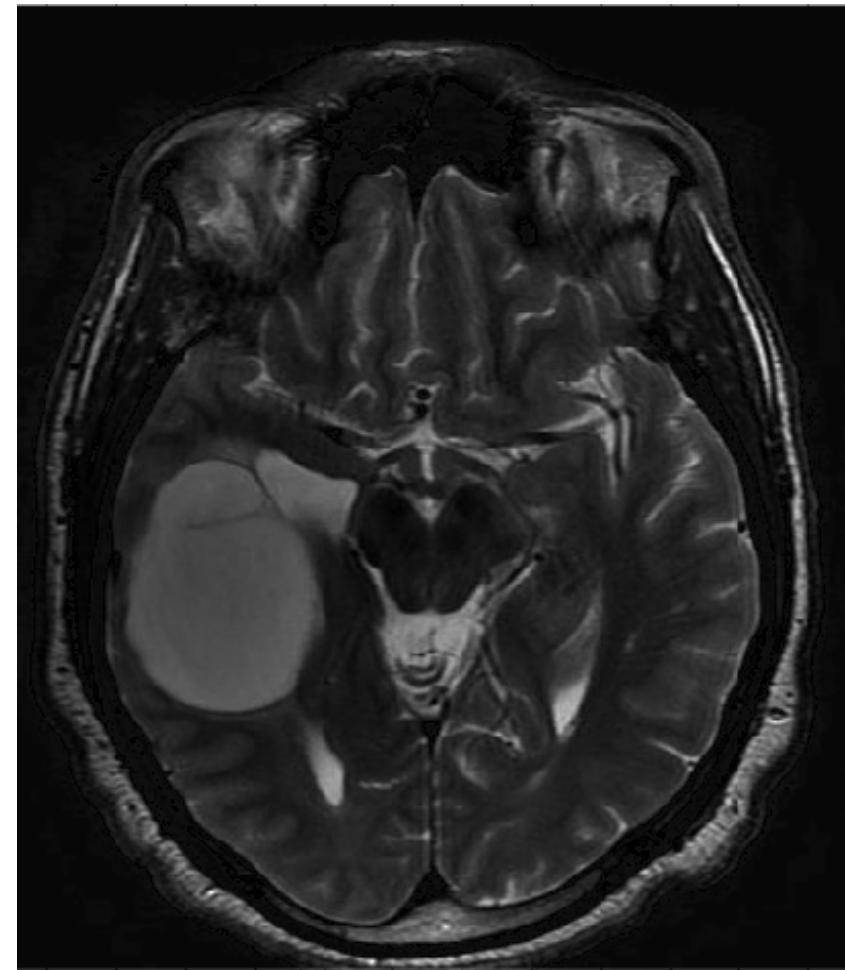
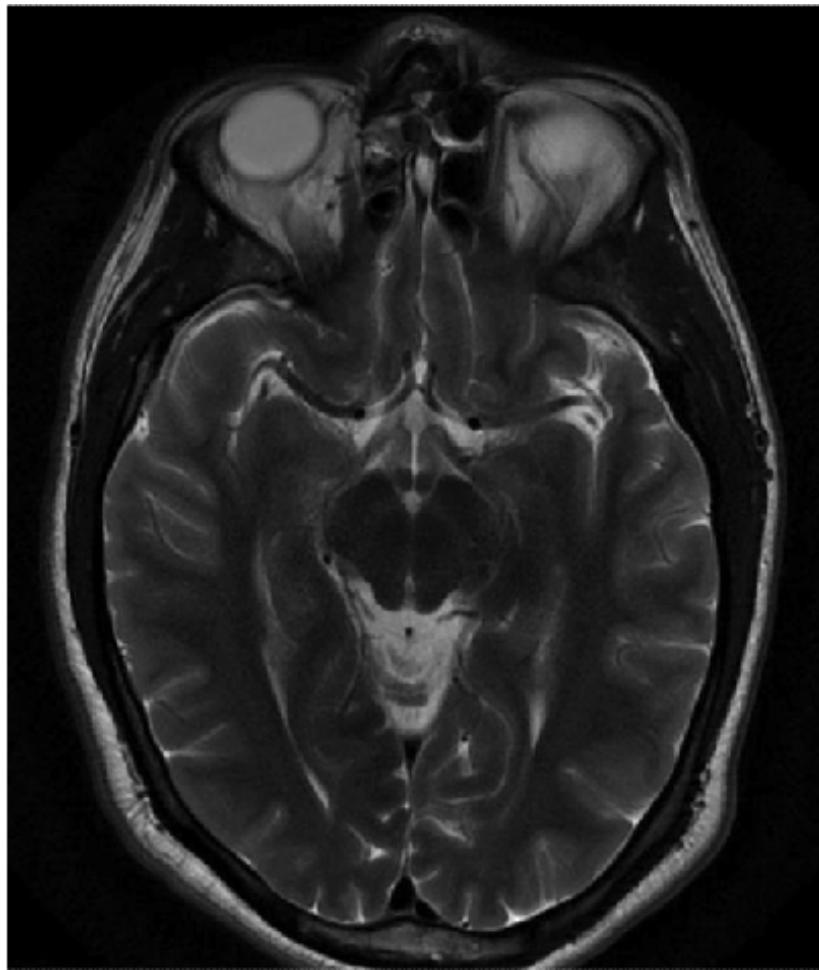
$k = 2$



Segmented MRI Images



T2 Weighted MRI Images

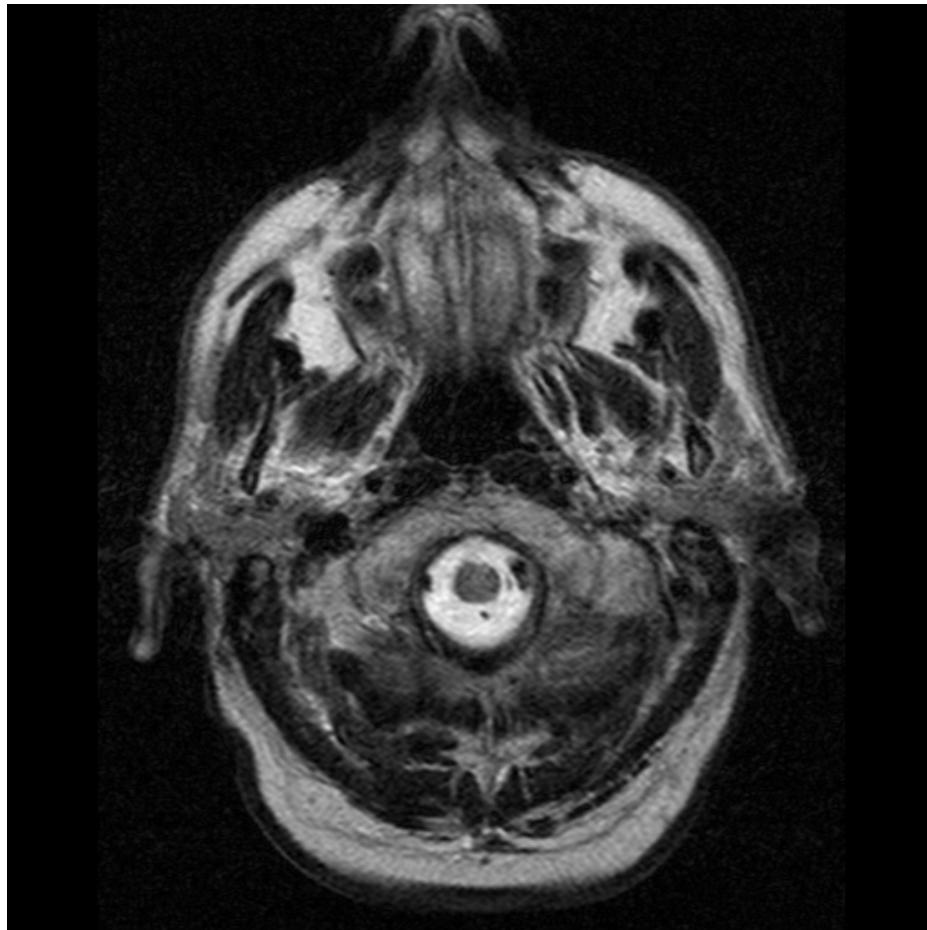


First Taste of a Fascinating Field



- MRI image segmentation is subject of ongoing research
- k -means is a good starting point, but not enough
 - Advanced clustering techniques such as the modified fuzzy k -means (MFCM) clustering technique
 - Packages in R specialized for medical image analysis
<http://cran.r-project.org/web/views/MedicalImaging.html>

3D Reconstruction



- 3D reconstruction from 2D cross sectional MRI images
- Important in the medical field for diagnosis, surgical planning and biological research

Comparison of Methods

	Used For	Pros	Cons
Linear Regression	Predicting a continuous outcome (salary, price, number of votes, etc.)	<ul style="list-style-type: none">• Simple, well recognized• Works on small and large datasets	<ul style="list-style-type: none">• Assumes a linear relationship $Y = a \underbrace{\log(X)}_{+b}$
Logistic Regression	Predicting a categorical outcome (Yes/No, Sell/Buy, Accept/Reject, etc.)	<ul style="list-style-type: none">• Computes probabilities that can be used to assess confidence of the prediction	<ul style="list-style-type: none">• Assumes a linear relationship

Comparison of Methods

	Used For	Pros	Cons
CART	Predicting a categorical outcome (quality rating 1--5, Buy/Sell/Hold) or a continuous outcome (salary, price, etc.)	<ul style="list-style-type: none">• Can handle datasets without a linear relationship• Easy to explain and interpret	<ul style="list-style-type: none">• May not work well with small datasets
Random Forests	Same as CART	<ul style="list-style-type: none">• Can improve accuracy over CART	<ul style="list-style-type: none">• Many parameters to adjust• Not as easy to explain as CART

Comparison of Methods

	Used For	Pros	Cons
Hierarchical Clustering	<ul style="list-style-type: none">Finding similar groupsClustering into smaller groups and applying predictive methods on groups	<ul style="list-style-type: none">No need to select number of clusters a prioriVisualize with a dendrogram	<ul style="list-style-type: none">Hard to use with large datasets
k -means Clustering	Same as Hierarchical Clustering	<ul style="list-style-type: none">Works with any dataset size	<ul style="list-style-type: none">Need to select number of clusters before algorithm



The Analytical Policeman

Visualization for Law and Order

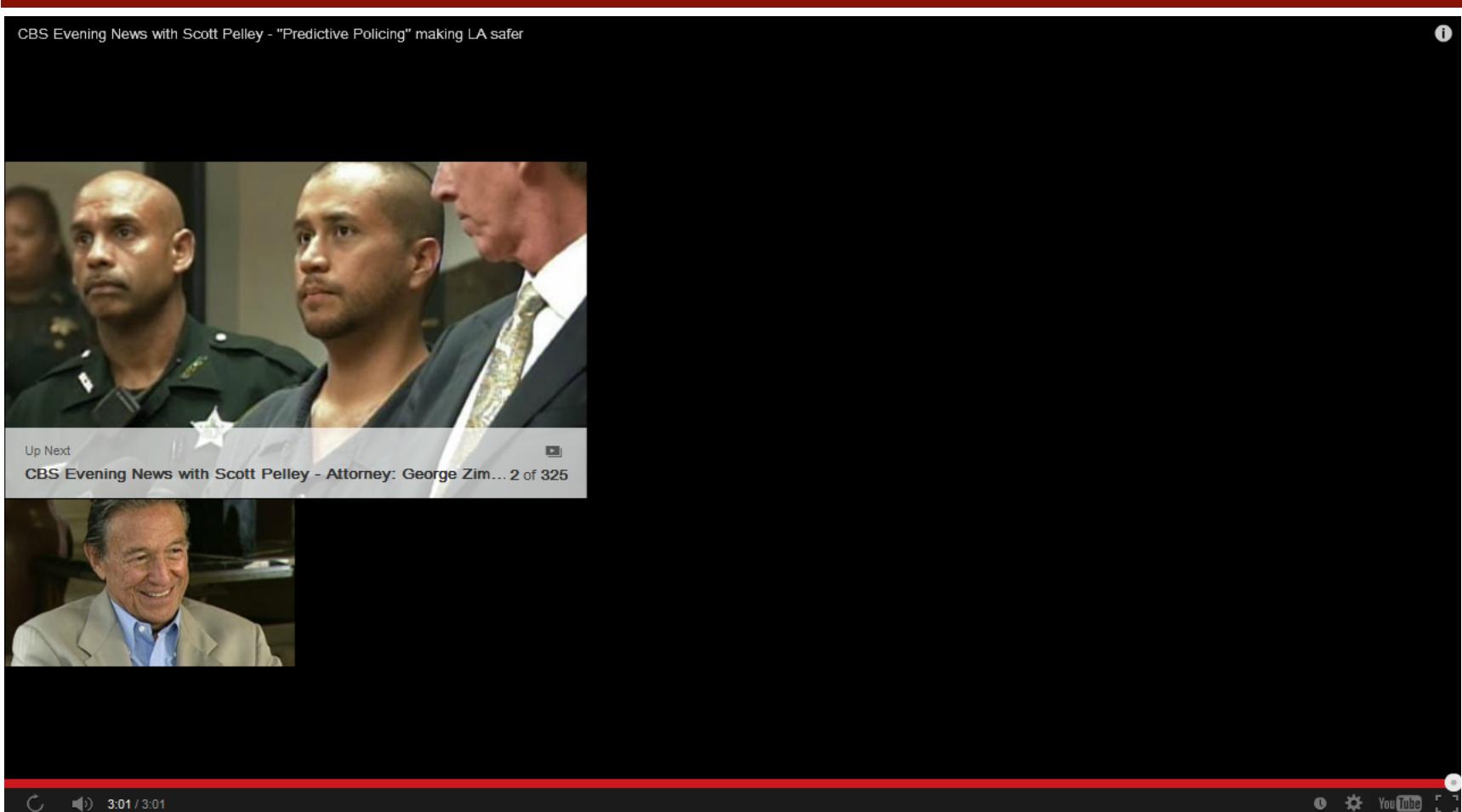
15.071x – The Analytics Edge

The Analytical Policeman



- The explosion of computerized data affects all parts of society, including law and order
- In the past, human judgment and experience was the only tool in identifying patterns in criminal behavior
- Police forces around the US and the world are augmenting human judgment with analytics – sometimes described as “**predictive policing**”

Predictive Policing in the News



Example: Los Angeles Police Dept.



“I’m not going to get more money. I’m not going to get more cops. I have to be better at using what I have, and that’s what **predictive policing** is about... If this old street cop can change the way that he thinks about this stuff, then I know that my [officers] can do the same.”

- Los Angeles Police Chief Charlie Beck

Role of Analytics

- The analytical tools you have learned in this class can be used to make these “predictive policing” models
- However, **communicating** the results of these models is essential – a **linear regression** output table will not be of use to a **policewoman on patrol**
- Visualization bridges the gap between **the data and mathematics** and the **end user**

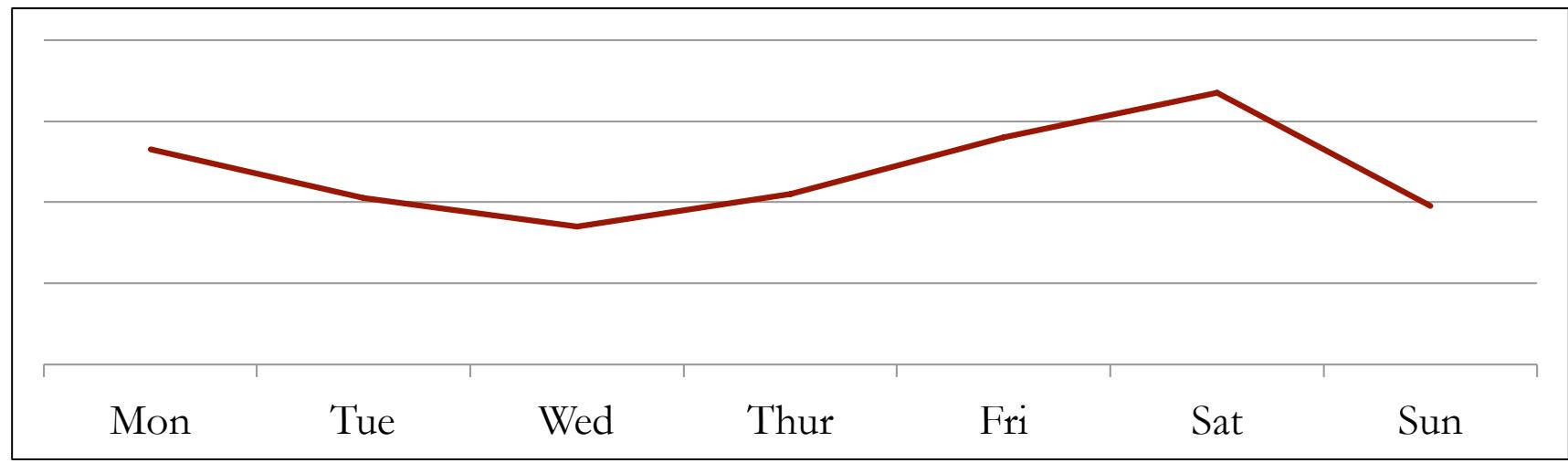
Understanding the Past



- Before we even consider a predictive model, we should try to understand the historical data
- Many cities in the US and around the world provide logs of reported crimes, usually including the time, location, and nature of the event
- We will use data from Chicago about motor vehicle thefts

Crime Over Time

- Suppose we wanted to communicate crime patterns over the course of an average week
- We could display daily averages using a line graph, but this does not seem like it would be too useful



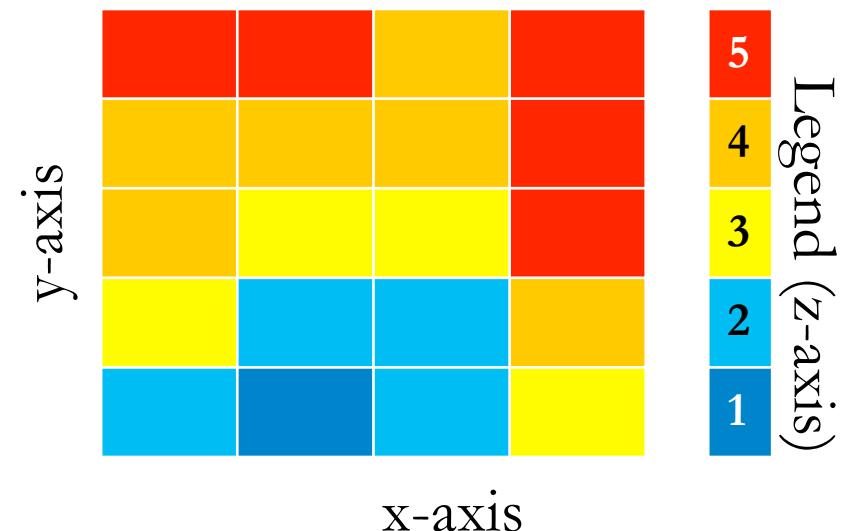
Crime Over Time

- We can replace our **x-axis** with the **hour of the day**, and have a different **line for every day of the week**, but this would be a jumbled mess with **7 lines!**
- We could use no visualization at all, and instead present the information in a table
- This is valid, but how can we make the table more interesting and usable?

	MO	TU	WE	TH
03:00	34	32	31	...
04:00	15	24	22	...
05:00	22	10	33	...
06:00	13	14	19	...
...

Heatmaps

- **Heatmaps** are a way of visualizing data using three attributes. The **x-axis** and **y-axis** are typically displayed horizontally and vertically
- The **third attribute** is represented by shades of color. For example, a **low** number might be **blue**, and a **high** number might be **red**



Heatmaps

- We can pick **different color schemes** based on the **type of data** to convey different messages



- The x-axis and y-axis don't need to be continuous – they can be **categorical**
- We could even combine a heatmap with a **geographical map** – we will discuss this later in the class

A Chicago Crime Heatmap

- We will use Chicago motor vehicle theft data to explore patterns of crime:
 - Over days of the week
 - Over hours of the day
- We're interested in the total number of car thefts that occur in any particular hour of a day of the week over the whole data set

Eye on Crime

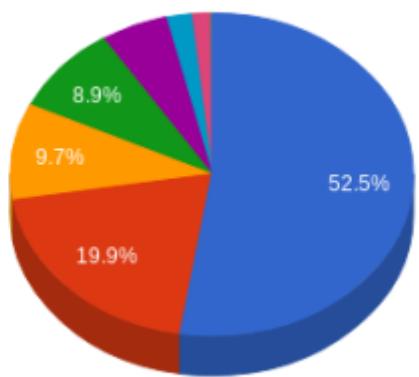


- Criminal activity-related data often has both components of time and location
- Sometimes all that is required is a line chart, but heatmaps can visualize data that would be too big for a table
- Plotting data on maps is much more effective than a table for location based data, and is eye-catching

Predictive Policing



- Many police forces are exploiting their databases to focus finite resources on problem areas
- Not only do analytics help improve policework, the outputs are also good communication tools to decision makers in government and to the wider public
- The application of analytics to data like this is new and growing, with companies like PredPol and Palantir leading the effort



The Good, the Bad, and the Ugly Visualization Recitation

15.071x – The Analytics Edge

Great Power, Great Responsibility



- There are many ways to visualize the same data.
- You have just seen how to make quite attractive visualizations with ggplot2, which has good default settings, but judgement is still required, e.g. do I vary the size, or do I vary the color?
- Excel, etc. can also be used to make perfectly acceptable visualizations – or terrible ones.

What is the difference?



- Good visualizations...

Clearly and accurately convey the key messages in the data

- Bad visualizations...

**Obfuscate the data
(either through ignorance, or malice!)**

What does this mean?



- Visualizations can be used by an analyst for their own consumption, to gain insights.
- Visualizations can also be used to provide information to a decision maker, and/or to convince someone.
- Bad visualizations hide patterns that could give insight, or mislead decision makers.

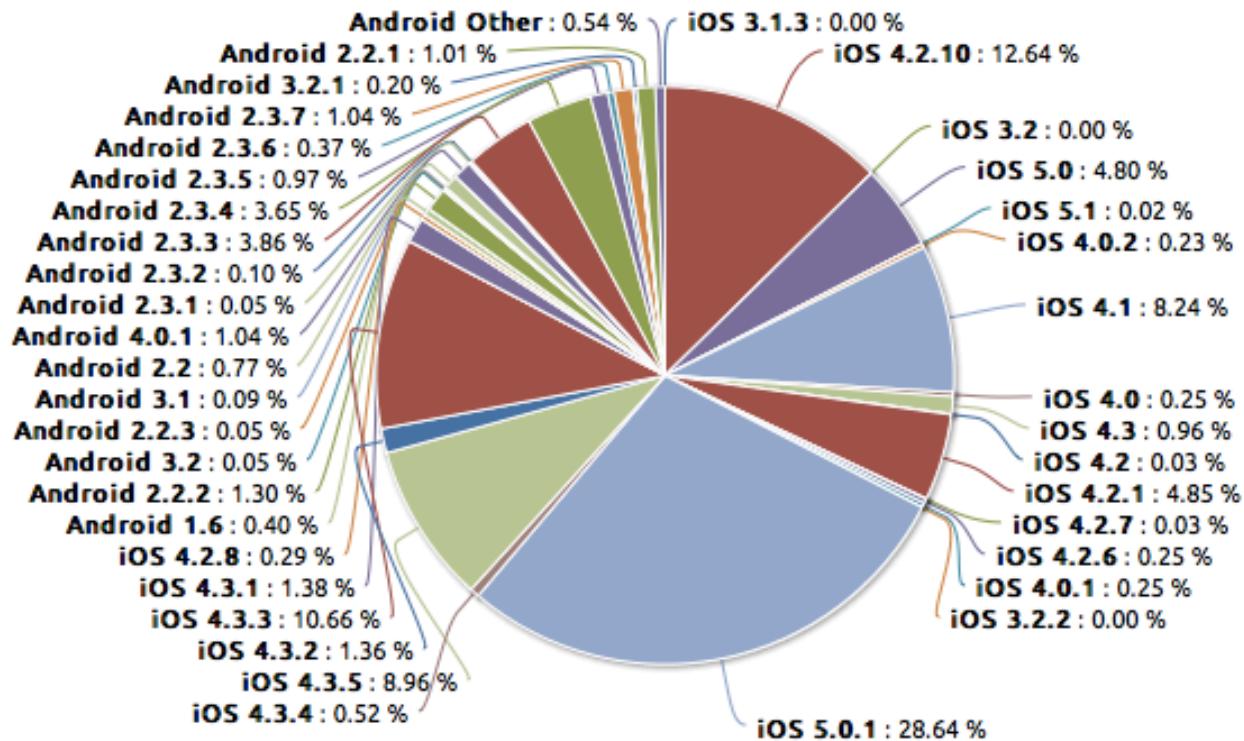
Today



- We will look at some examples of visualizations taken from a variety of sources.
- We'll discuss what is good and bad about them
- We will switch in to R to build better versions ourselves.
- Think for yourself: ultimately subjective!

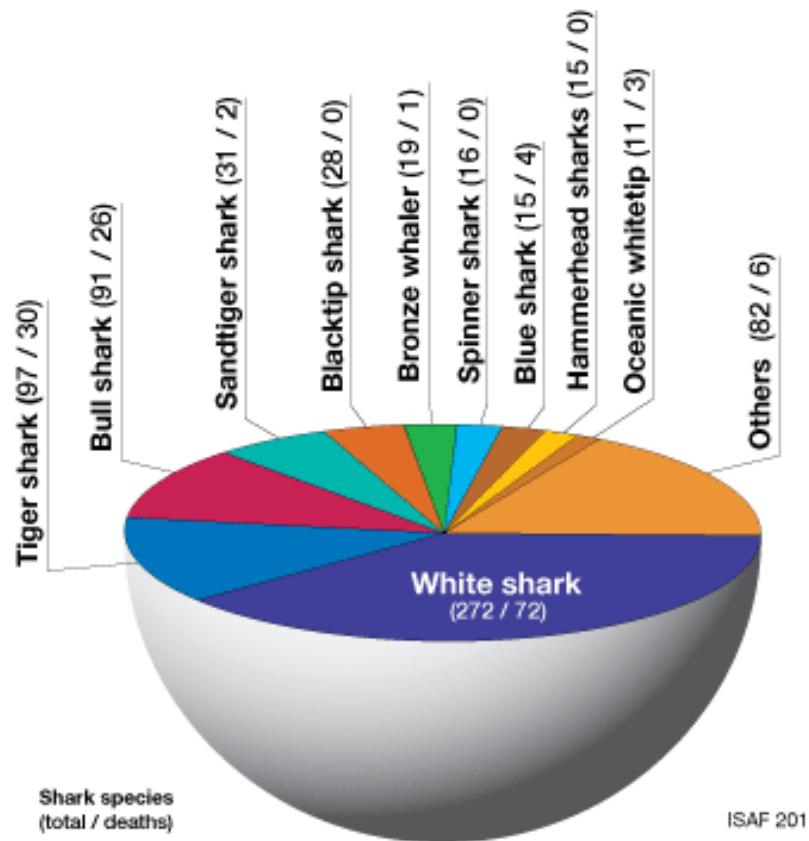
Bad Visualizations?

Crashes by OS Version Normalized (12/1 - 12/15)



Source: <http://www.forbes.com/sites/tomiogeran/2012/02/02/does-ios-crash-more-than-android-a-data-dive/>

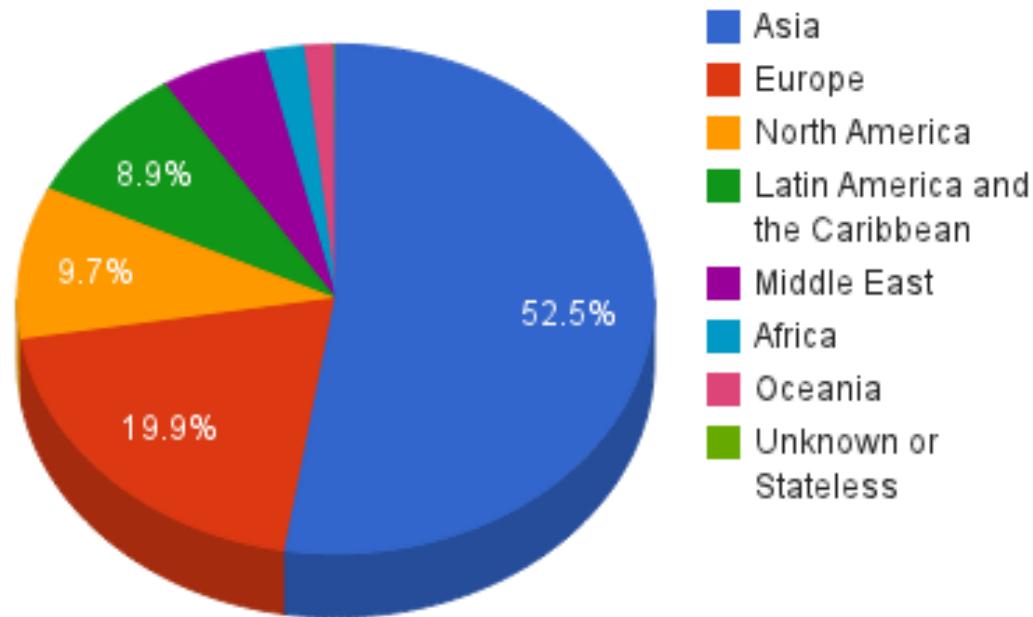
Bad Visualizations?



Source: International Shark Attack File report

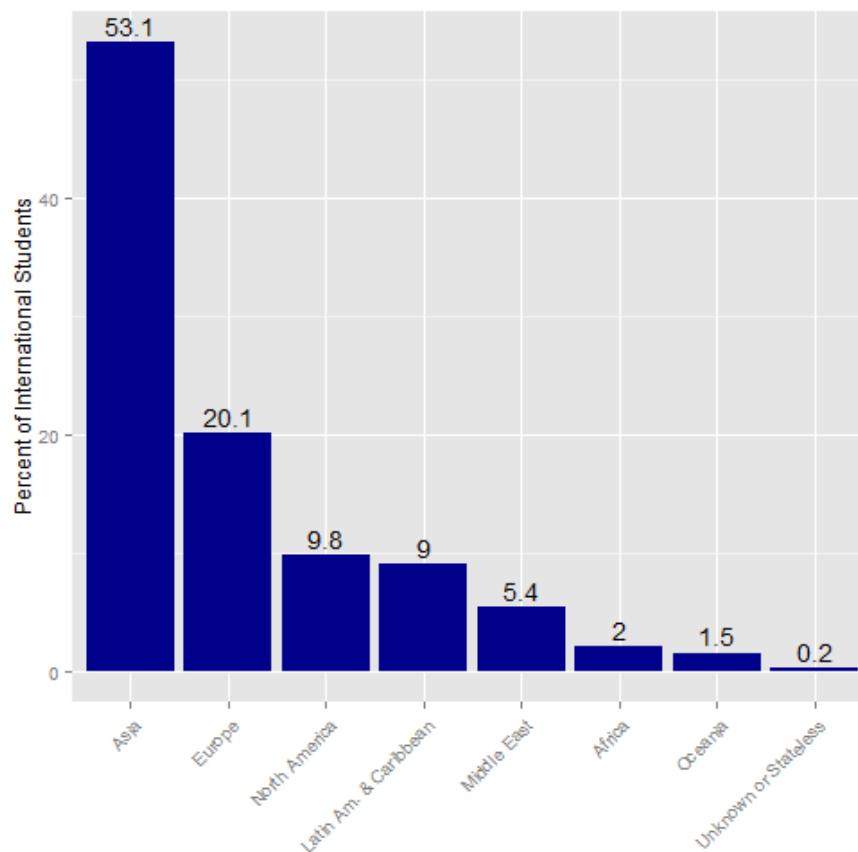
Bad Visualization?

MIT International Graduate Students



- Not all points can be labeled, so data is lost
- Colors are meaningless, are close enough to be confusing, but are still needed to make it at all readable.
- 3D adds nothing, visible volume is larger than true share

Better Visualization?



- All data is visible!
- Don't lose small regions.
- Can easily compare relative sizes
- Something to consider is that, for some people and applications, being not as “visually exciting” is a negative.

On a World Map?

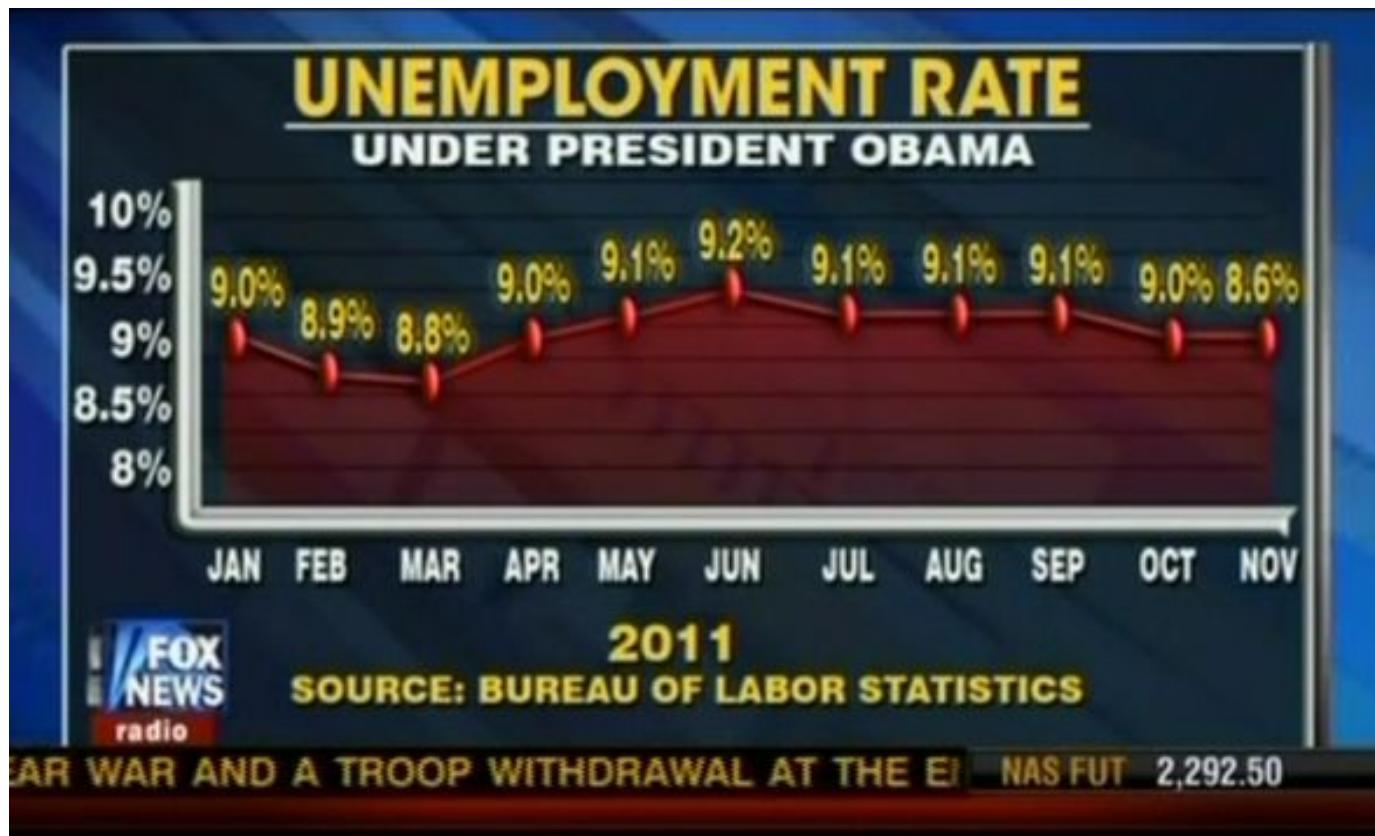
- Possible with this data, but still a bit tedious to create because we need to determine which countries lie in which region.
- Shading all countries in region the same color is misleading – countries in, e.g. Latin America, will send students at different rates.
- We have access to per country data – we will plot it on a world map and see if it is effective.

Bad Scales



Source: BBC

Bad Scales

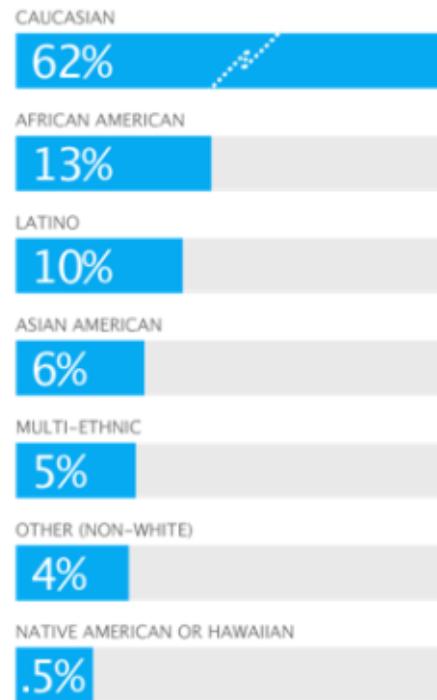


Source: Fox News

Bad Scales

Diversity for 2012 Corps Members

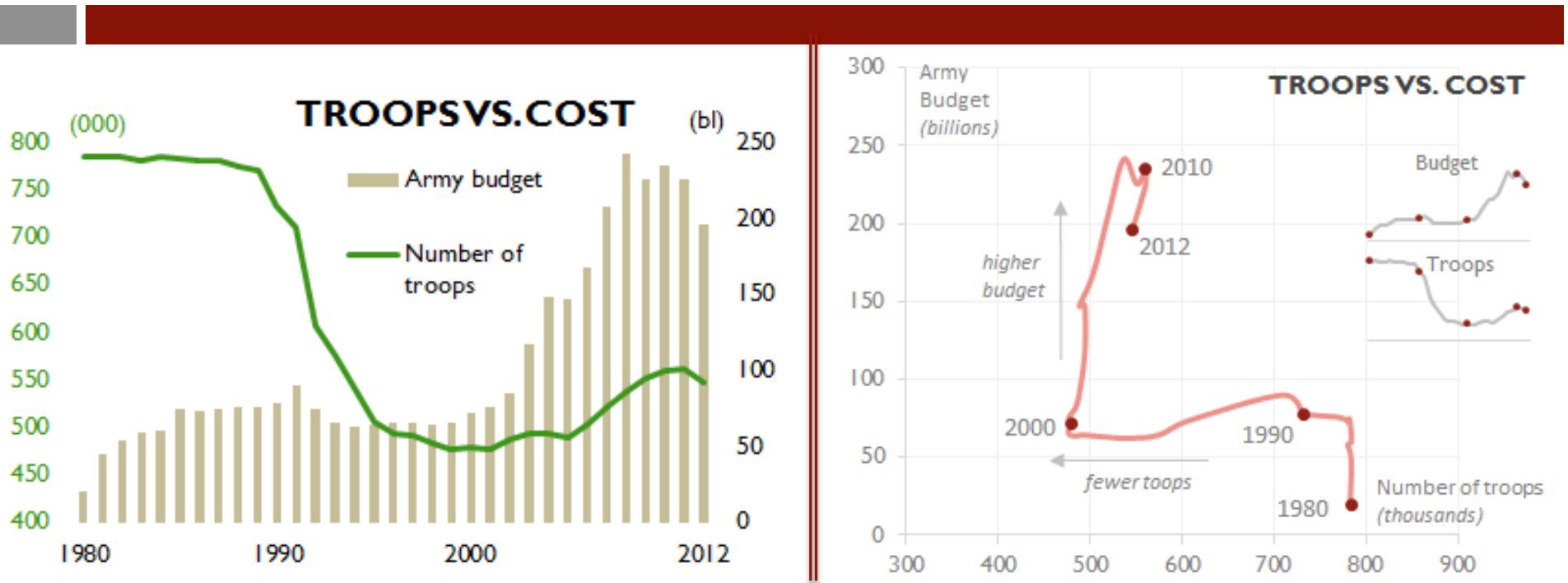
Total people of color: 38%



- “Caucasian” bar is truncated – would be as wide as this slide!
- Every bar has its own scale – compare “Native American” to “African American”.
- Only thing useful is the numbers.
- Minor: mixed precision, unclear rounding applied

<http://www.teachforamerica.org/why-teach-for-america/the-corps/who-we-look-for/the-importance-of-diversity>

Two Rights Make A Wrong



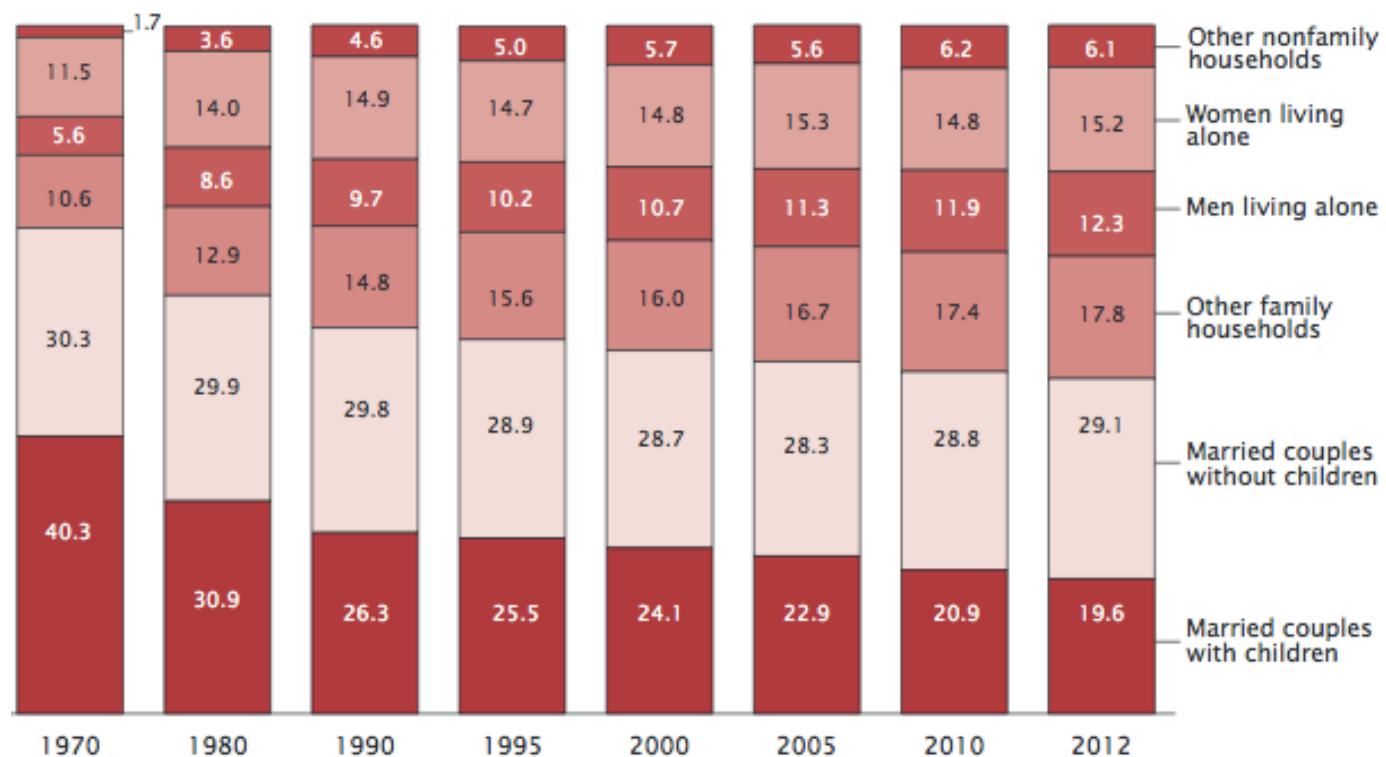
Source: <http://www.excelcharts.com/blog/redraw-troops-vs-cost-time-magazine/>

- Different units suggest (non-existent) crossover in 1995
- Transformation shows true moments of change

Family Matters

Households by Type, 1970 to 2012: CPS

(In percent)

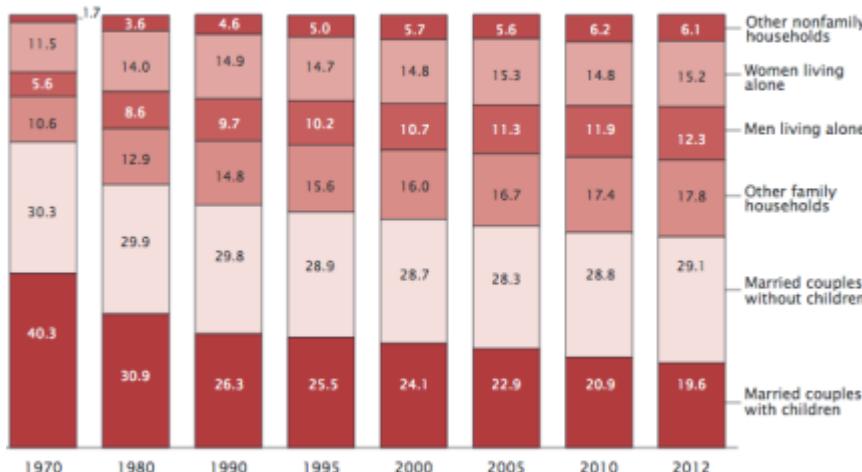


Source: U.S. Census Bureau, Current Population Survey, Annual Social and Economic Supplement, selected years, 1970 to 2012.

Family Matters

Households by Type, 1970 to 2012: CPS

(in percent)



Source: U.S. Census Bureau, Current Population Survey, Annual Social and Economic Supplement, selected years, 1970 to 2012.

- If we are interested in shares within a year, its good.
- If we want to see rates of change, it is pretty much unusable!

- If we want to compare year-to-year, its possible though imperfect.
- Numbers are relative – absolute numbers may reveal, e.g. married couples without children is constant across years.



Visualizing the World

An Introduction to Visualization

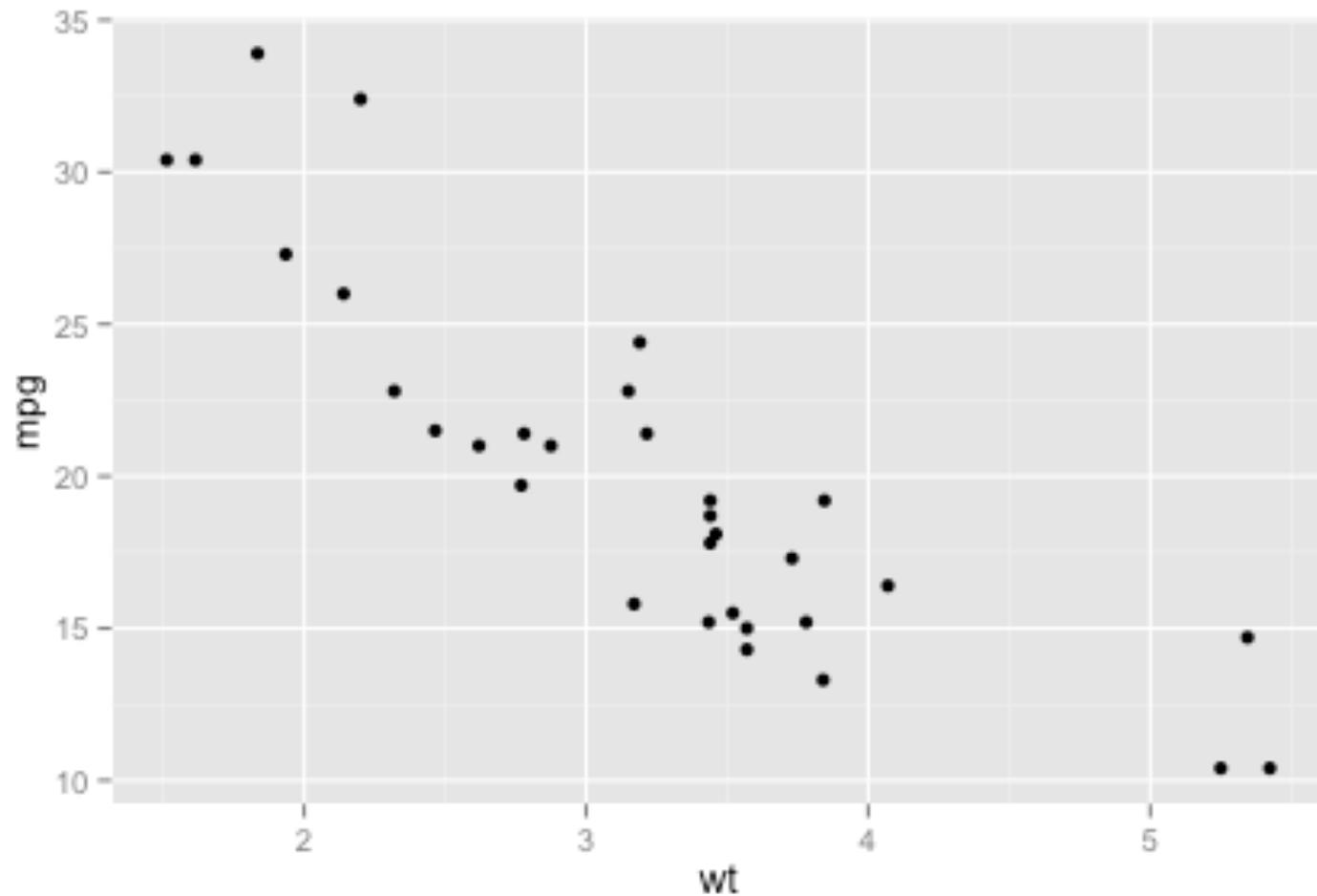
15.071x – The Analytics Edge

Why Visualization?

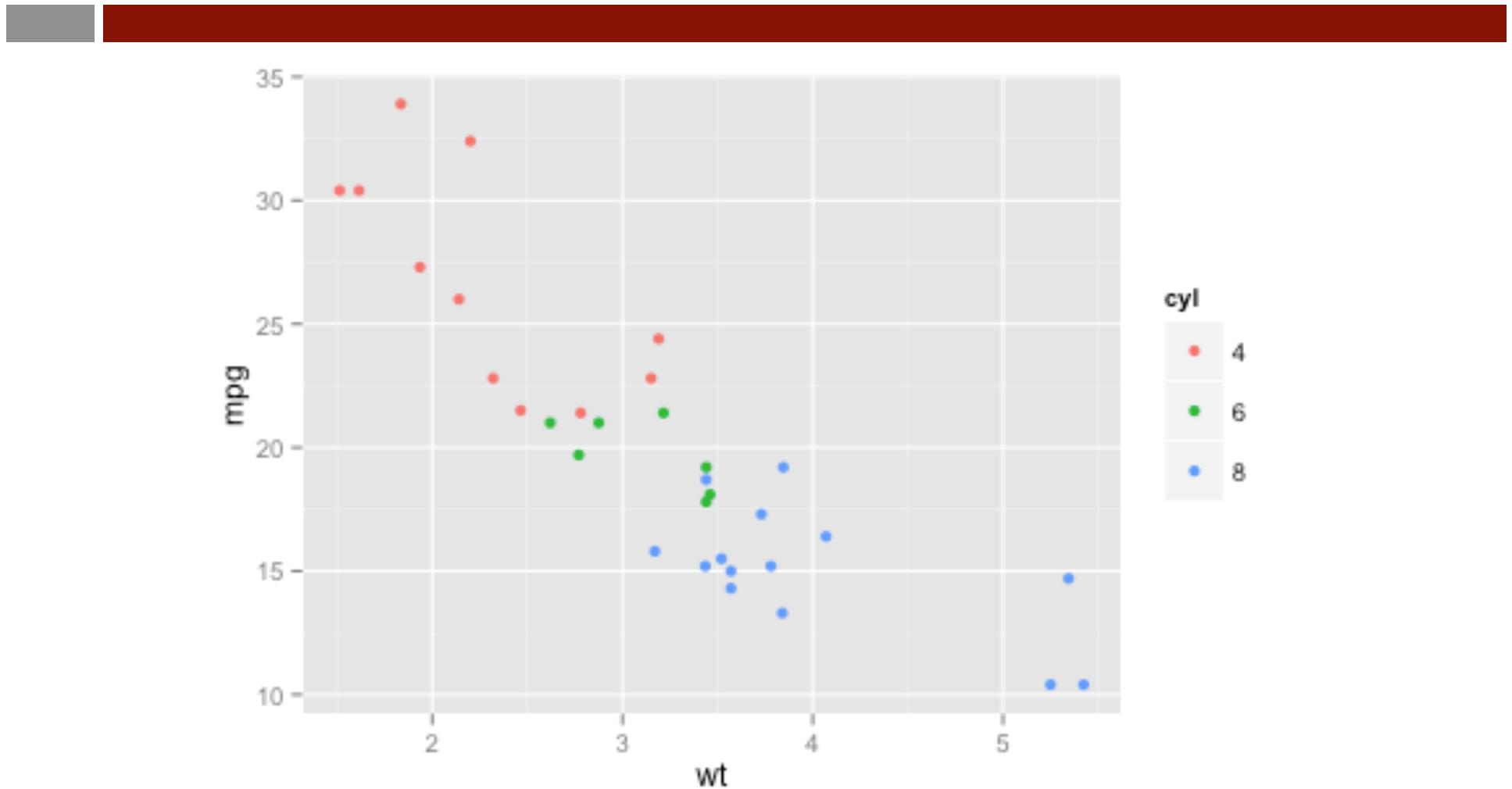


- “The picture-examining eye is the best finder we have of the wholly unanticipated”
-John Tukey
- Visualizing data allows us to discern relationships, structures, distributions, outliers, patterns, behaviors, dependencies, and outcomes
- Useful for initial data exploration, for interpreting your model, and for communicating your results

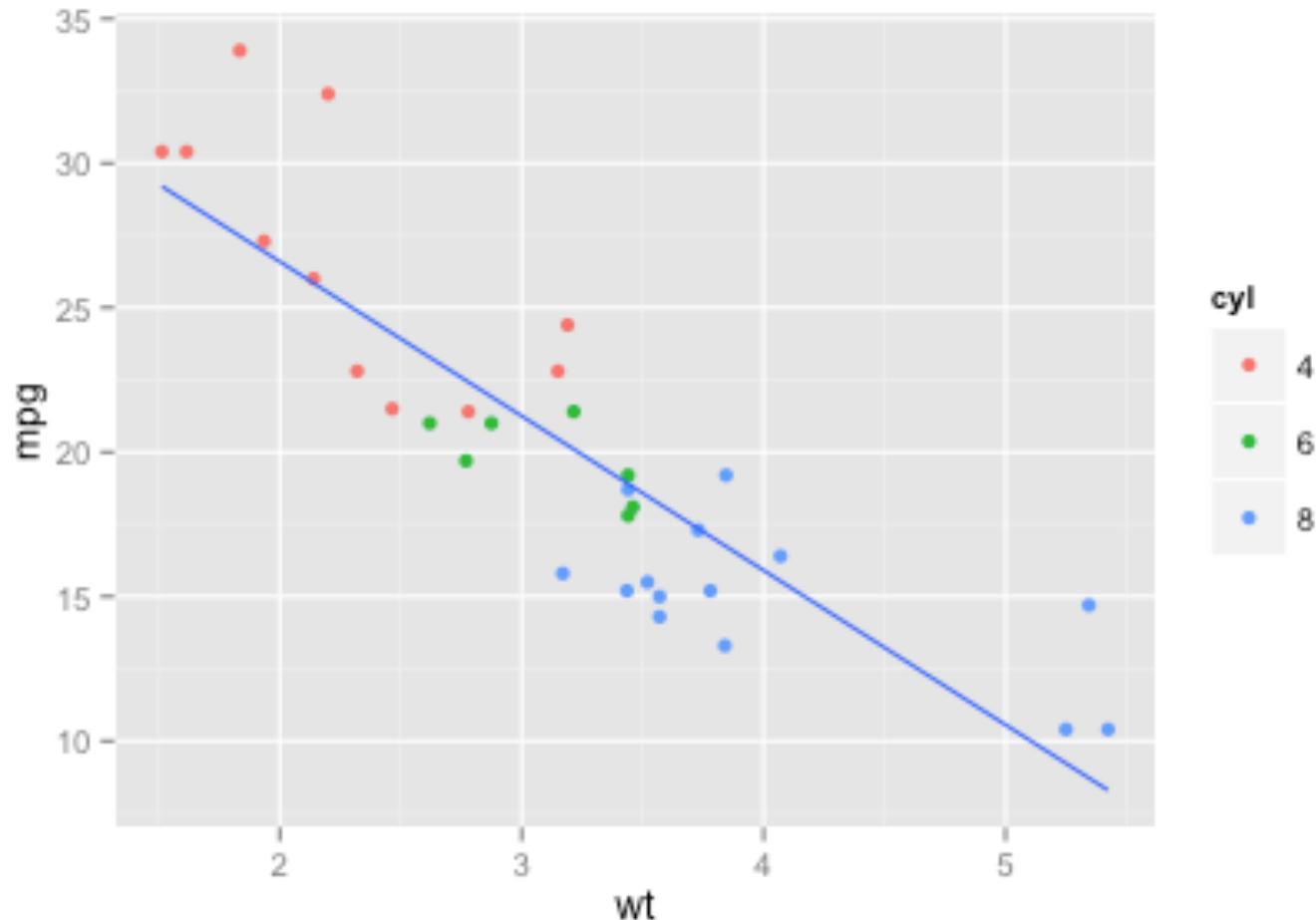
Initial Exploration Shows a Relationship



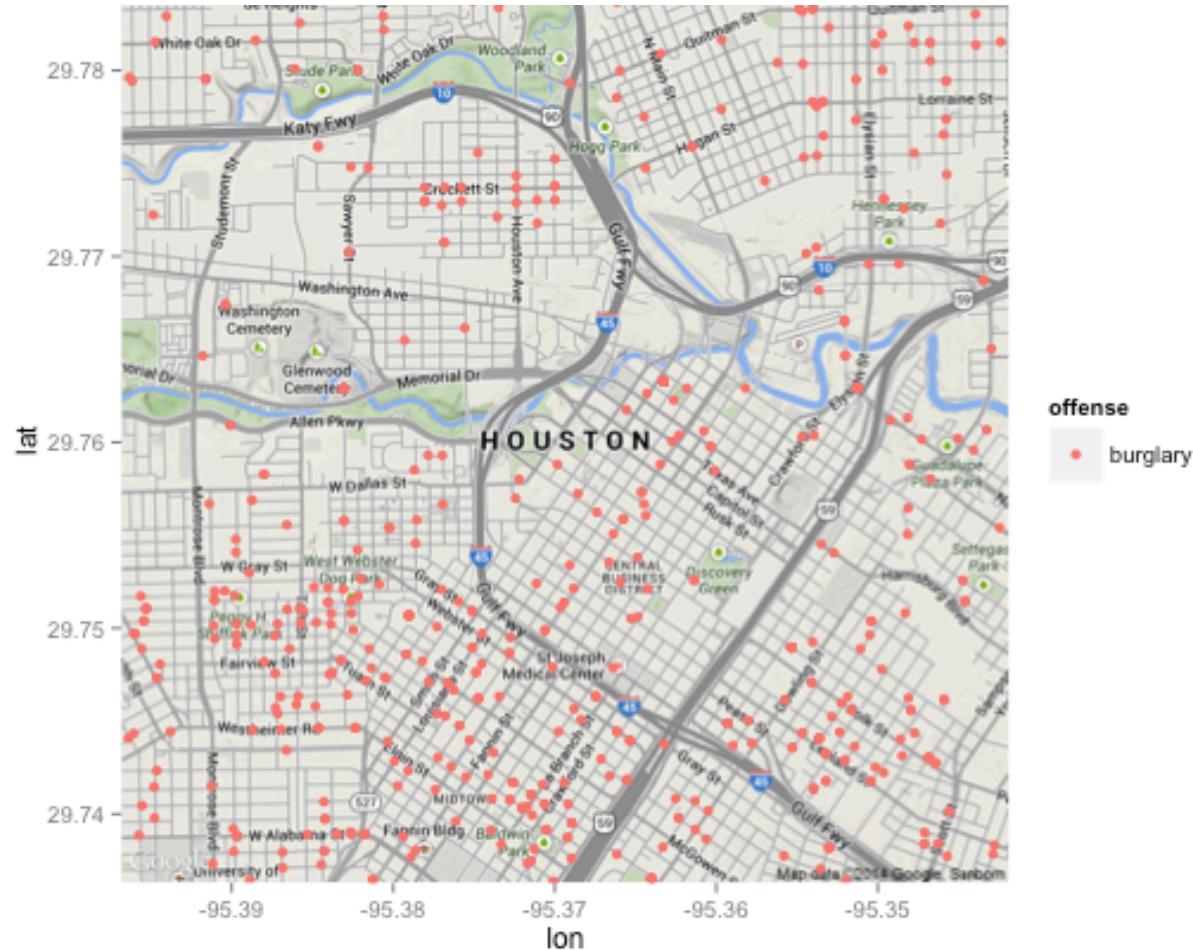
Explore Further: Color by Factor



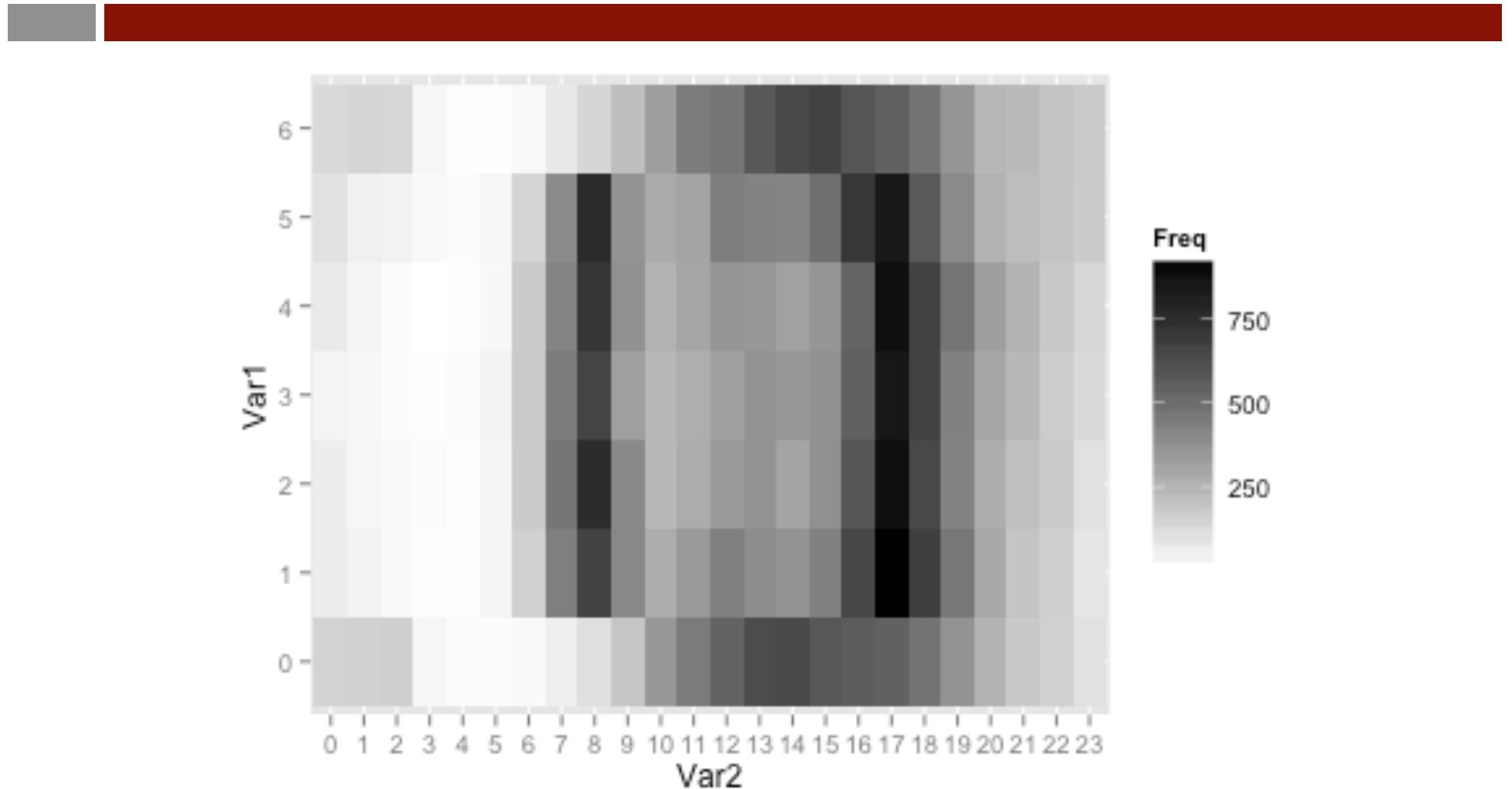
Make a Model. Plot the Regression Line.



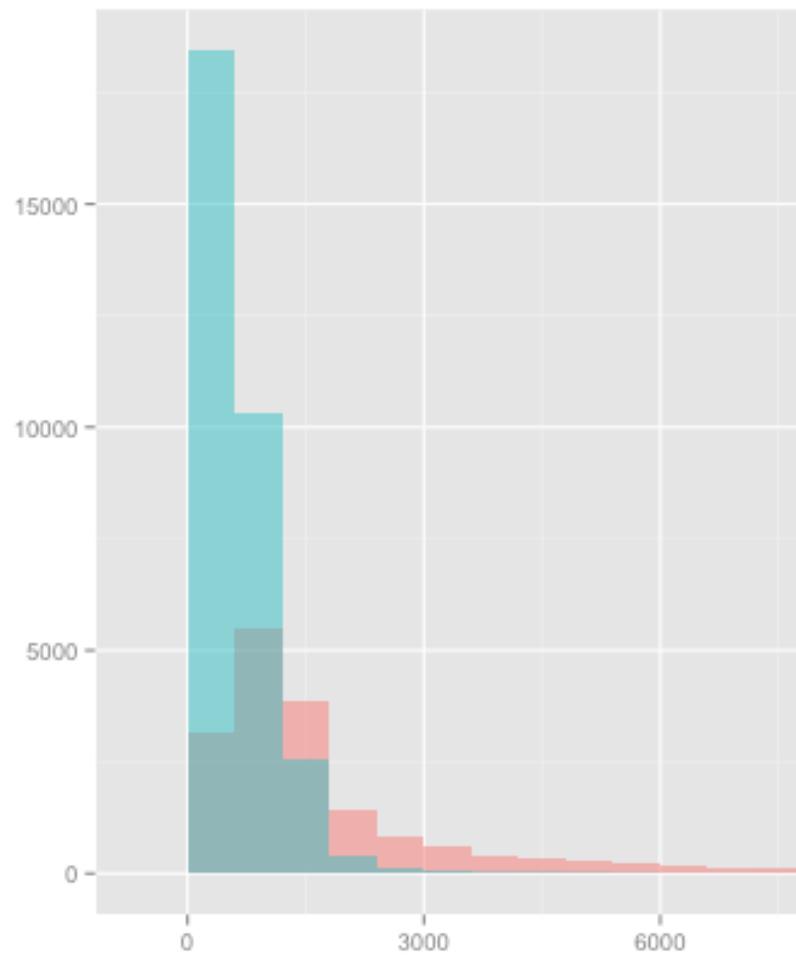
Add Geographical Data to a Map



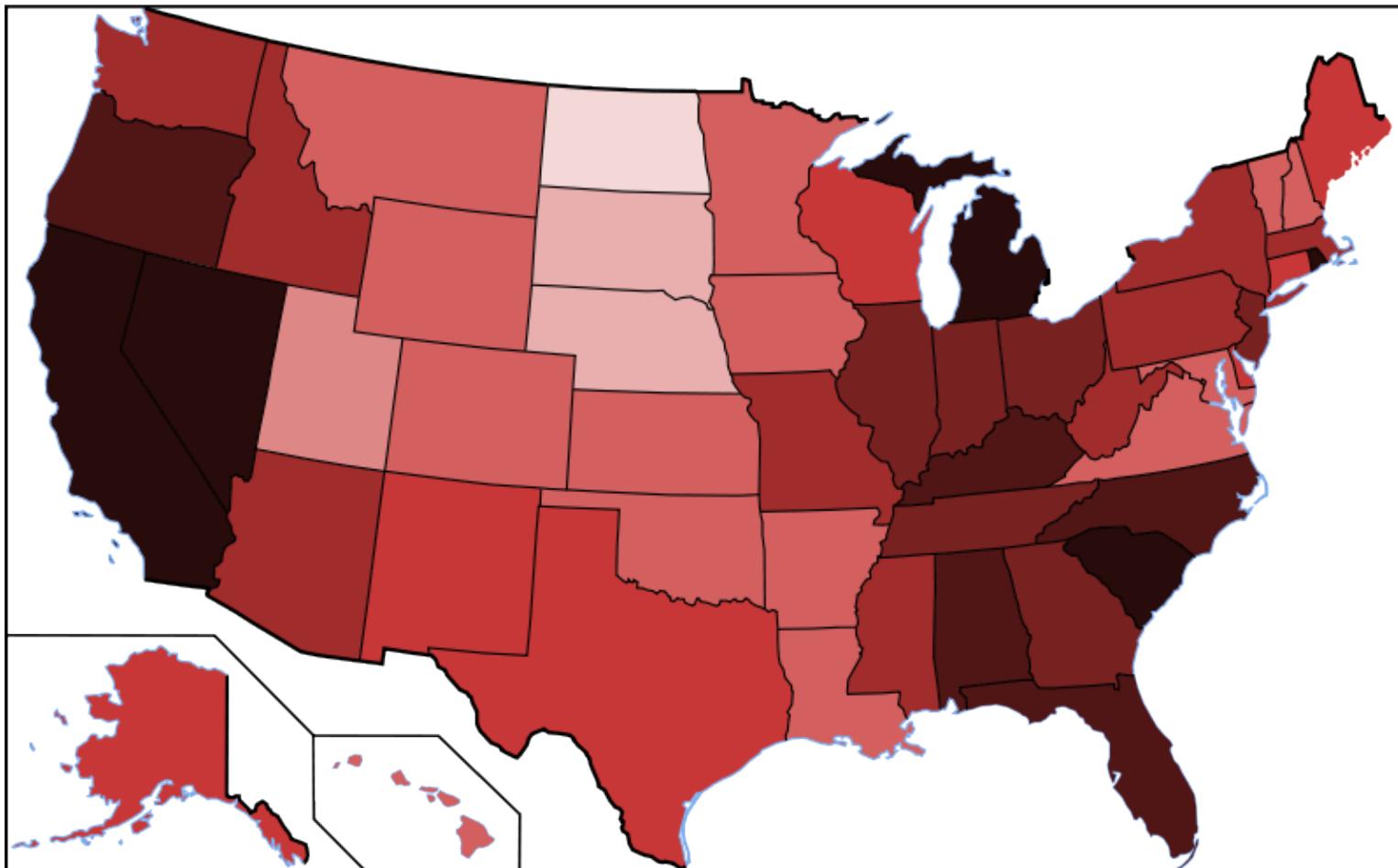
Show Relationships in a Heatmap



Make Histograms. Explore Categories.



Color a Map According to Data



The Power of Visualizations



- This week, we will create all of these visualizations
- We will see how visualizations can be used to
 - Better understand data
 - Communicate information to the public
 - Show the results of analytical models
- In the next video, we will discuss the World Health Organization (WHO), and how they use visualizations

The World Health Organization



“WHO is the authority for health within the United Nations system. It is responsible for providing leadership on global health matters, shaping the health research agenda, setting norms and standards, articulating evidence-based policy options, providing technical support to countries and monitoring and assessing health trends.”

The World Health Report

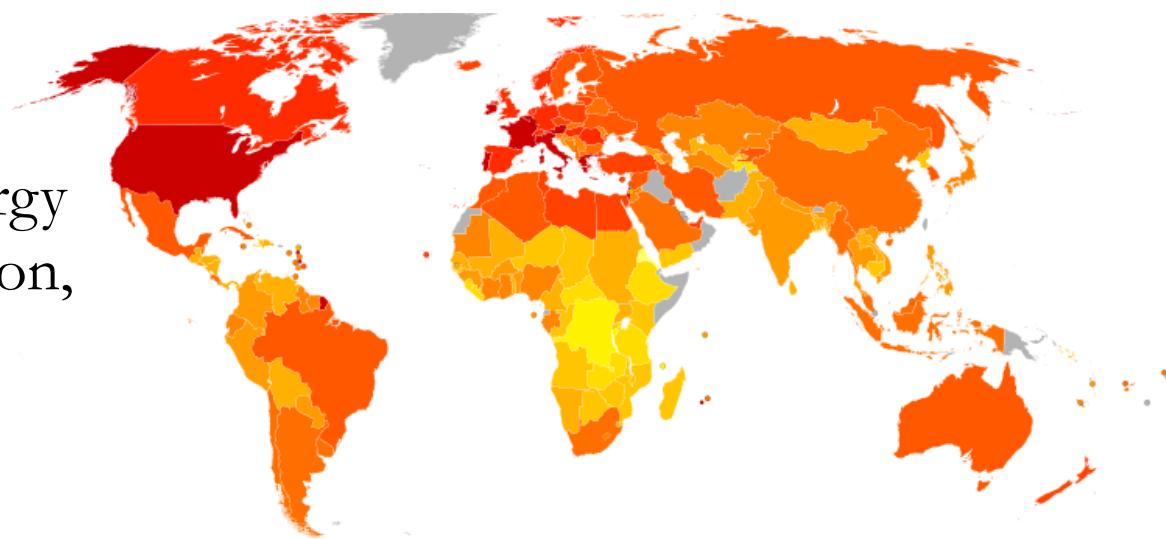
- WHO communicates information about global health in order to inform citizens, donors, policymakers and organizations across the world
- Their primary publication is “World Health Report”
- Each issue focuses on a specific aspect of global health, and includes statistics and experts’ assessments



Online Data Repository

- WHO also maintains an open, online repository of global health data
- WHO provides some data visualizations, which helps them communicate more effectively with the public

World Energy Consumption,
2001-2003



What is a Data Visualization?



- A mapping of data properties to visual properties
- Data properties are usually numerical or categorical
- Visual properties can be (x,y) coordinates, colors, sizes, shapes, heights, . . .

Anscombe's Quartet

X1	Y1
10.0	8.04
8.0	6.95
13.0	7.58
9.0	8.81
11.0	8.33
14.0	9.96
6.0	7.24
4.0	4.26
12.0	10.84
7.0	4.82
5.0	5.68

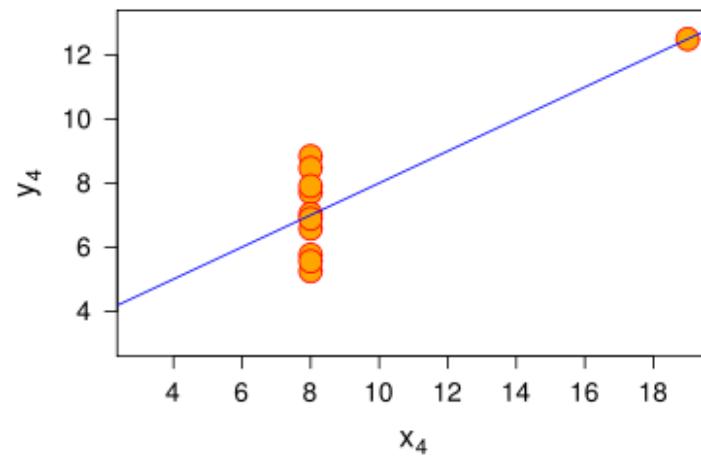
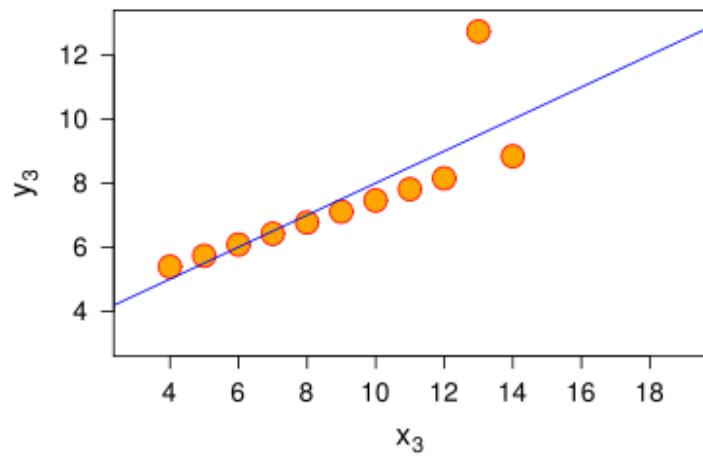
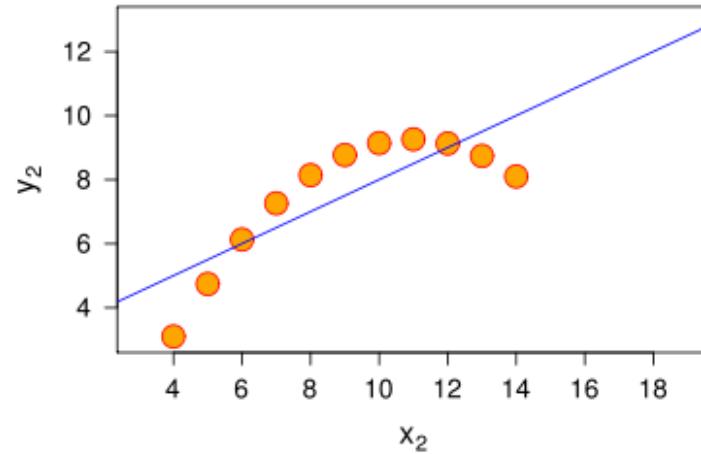
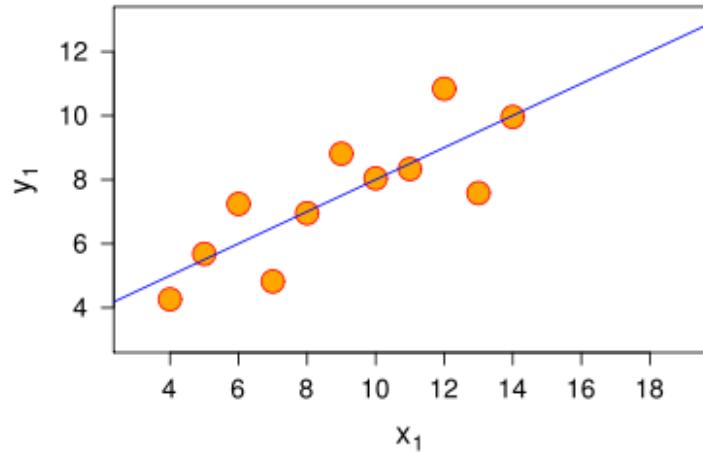
X2	Y2
10.0	9.14
8.0	8.14
13.0	8.74
9.0	8.77
11.0	9.26
14.0	8.10
6.0	6.13
4.0	3.10
12.0	9.13
7.0	7.26
5.0	4.74

X3	Y3
10.0	7.46
8.0	6.77
13.0	12.74
9.0	7.11
11.0	7.81
14.0	8.84
6.0	6.08
4.0	5.39
12.0	8.15
7.0	6.42
5.0	5.73

X4	Y4
8.0	6.58
8.0	5.76
8.0	7.71
8.0	8.84
8.0	8.47
8.0	7.04
8.0	5.25
19.0	12.50
8.0	5.56
8.0	7.91
8.0	6.89

- Mean of X: 9.0
- Variance of X: 11.0
- Mean of Y: 7.50
- Variance of Y: 4.12
- Correlation between X and Y: 0.816
- Regression Equation:
- $Y = 3.00 + 0.500X$

Anscombe's Quartet



ggplot

- “ggplot2 is a plotting system for R, based on the grammar of graphics, which tries to take the good parts of base and lattice graphics and none of the bad parts. It takes care of many of the fiddly details that make plotting a hassle (like drawing legends) as well as providing a powerful model of graphics that makes it easy to produce complex multi-layered graphics.”

-Hadley Wickham, creator, www.ggplot2.org

Graphics in Base R vs ggplot



- In base R, each mapping of data properties to visual properties is its own special case
 - Graphics composed of simple elements like points, lines
 - Difficult to add elements to existing plots
- In ggplot, the mapping of data properties to visual properties is done by adding layers to the plot

Grammar of Graphics

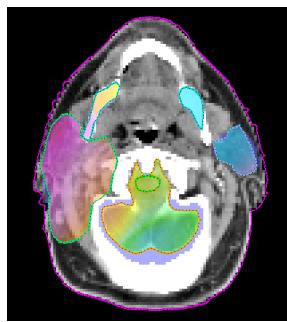


- ggplot graphics consist of at least 3 elements:
 1. **Data**, in a data frame
 2. **Aesthetic mapping** describing how variables in the data frame are mapped to graphical attributes
 - Color, shape, scale, x-y axes, subsets,...
 3. **Geometric objects** determine how values are rendered graphically
 - Points, lines, boxplots, bars, polygons,...

The Analytics Edge



- WHO's online data repository of global health information is used by citizens, policymakers, and organizations across the world
- Visualizing the data facilitates the understanding and communication of global health trends at a glance
- ggplot in R lets you visualize for exploration, modeling, and sharing results



RADIATION THERAPY

An Application of Linear Optimization

15.071x – The Analytics Edge

Cancer



- Cancer is the second leading cause of death in the United States, with an estimated **570,000 deaths** in 2013
- Over **1.6 million new cases** of cancer will be diagnosed in the United States in 2013
- In the world, cancer is also a leading cause of death – **8.2 million deaths** in 2012

Radiation Therapy



- Cancer can be treated using radiation therapy (RT)
- In RT, beams of high energy photons are fired into the patient that are able to kill cancerous cells
- In the United States, about **half of all cancer patients** undergo some form of radiation therapy

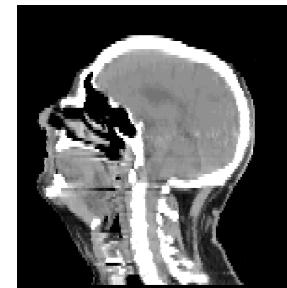
History of Radiation Therapy

- X-rays were discovered by Wilhelm Röntgen in 1895 (awarded the first Nobel Prize in Physics in 1901)
 - Shortly after, x-rays started being used to treat skin cancers
- Radium discovered by Marie and Pierre Curie in 1898 (Nobel Prize in Chemistry in 1911)
 - Began to be used to treat cancer, as well as other diseases



History of Radiation Therapy

- First radiation delivery machines (linear accelerators) developed in 1940
- Computed tomography (CT) invented in 1971
- **Invention of intensity-modulated radiation therapy (IMRT) in early 1980s**



IMRT

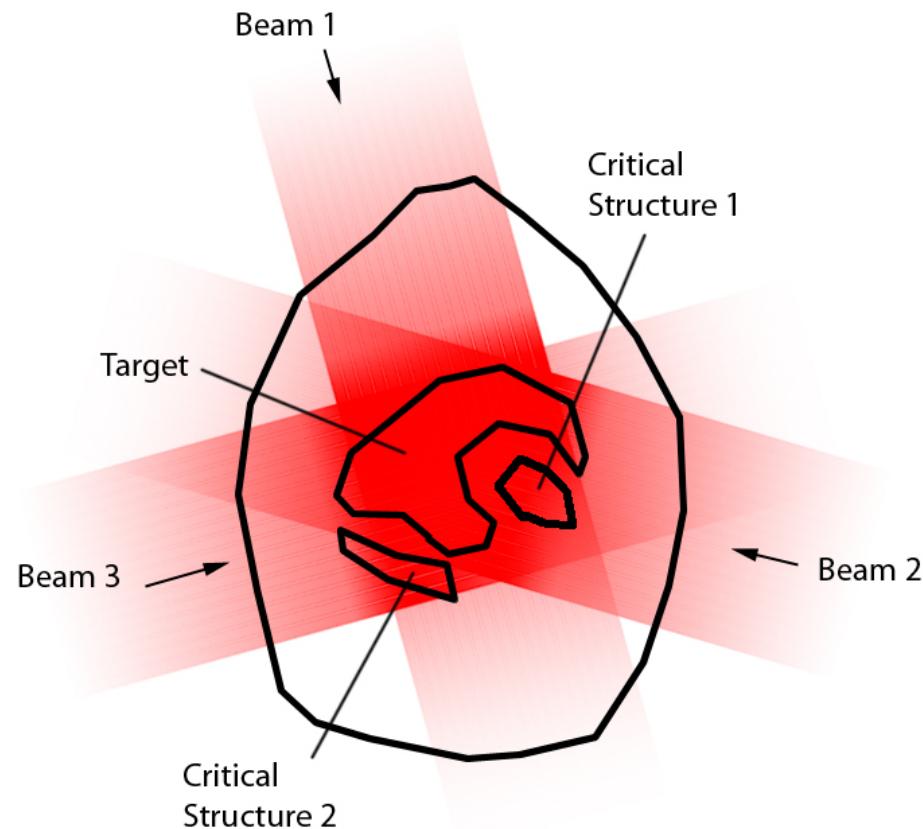


- To reach the tumor, radiation passes through healthy tissue, and damages both healthy and cancerous tissue
- Damage to healthy tissue can lead to undesirable side effects that reduce post-treatment quality of life
- We want the dose to “fit” the tumor as closely as possible, to reduce the dose to healthy tissues

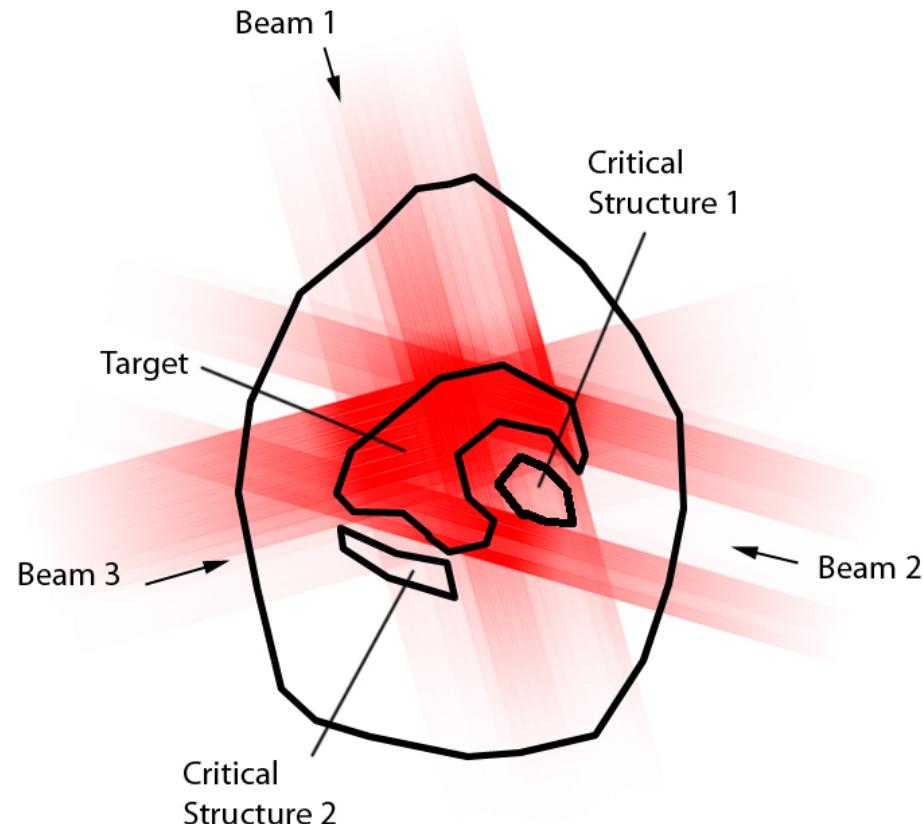
IMRT

- In IMRT, the intensity profile of each beam is non-uniform
- By using non-uniform intensity profiles, the three-dimensional shape of the dose can better fit the tumor
- Let's see what this looks like

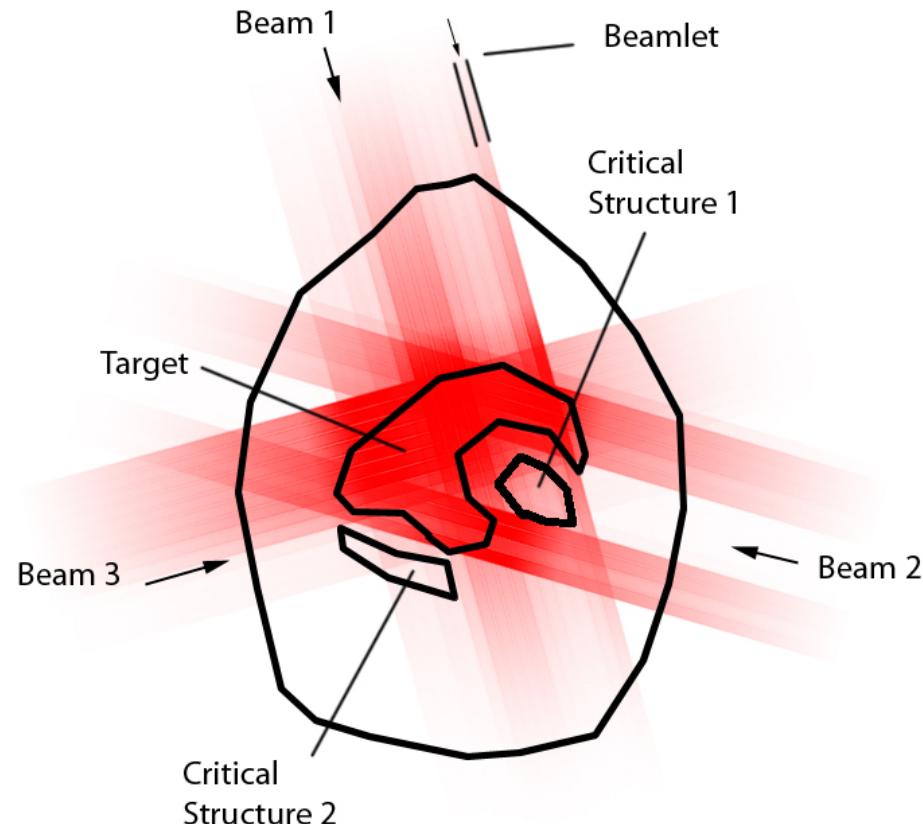
Using Traditional Radiation Therapy



Using IMRT



Using IMRT



Designing an IMRT Treatment



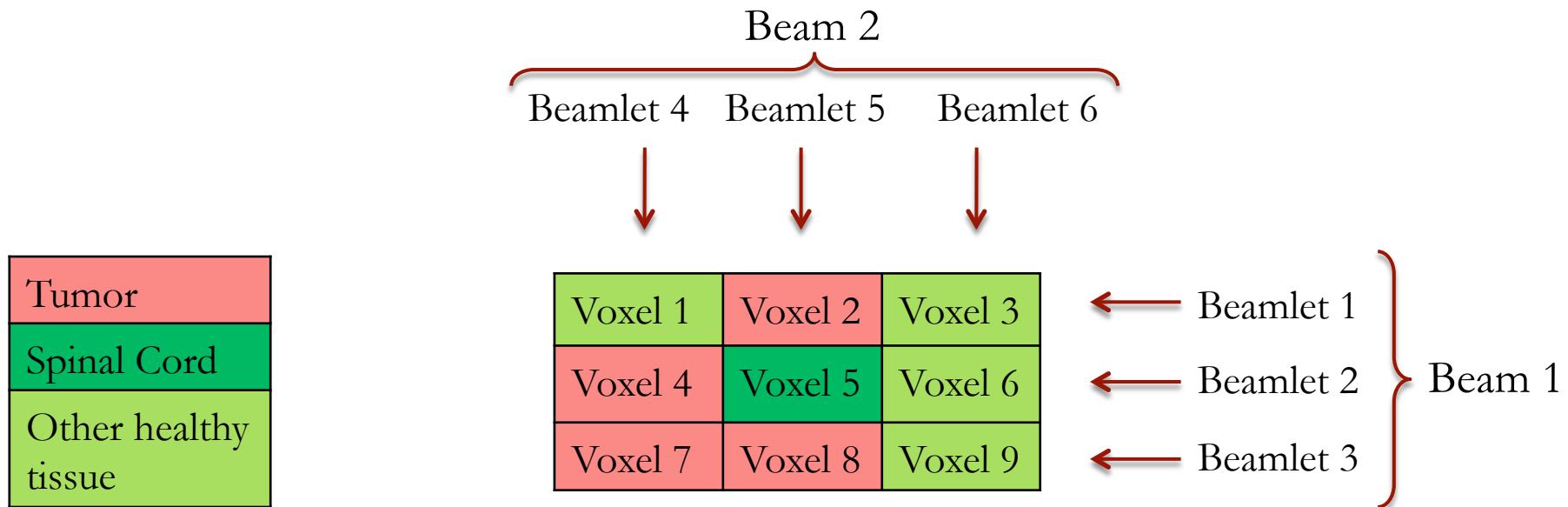
- Fundamental problem:
 - How should the beamlet intensities be selected to deliver a therapeutic dose to the tumor *and* to minimize damage to healthy tissue?

The Data

- Treatment planning starts from a CT scan
 - A radiation oncologist contours (draws outlines) around the tumor and various critical structures
 - Each structure is discretized into voxels (volume elements) – typically 4 mm x 4 mm x 4 mm
- From CT scan, can compute how much dose each beamlet delivers to every voxel



Small Example – 9 Voxels, 6 Beamlets



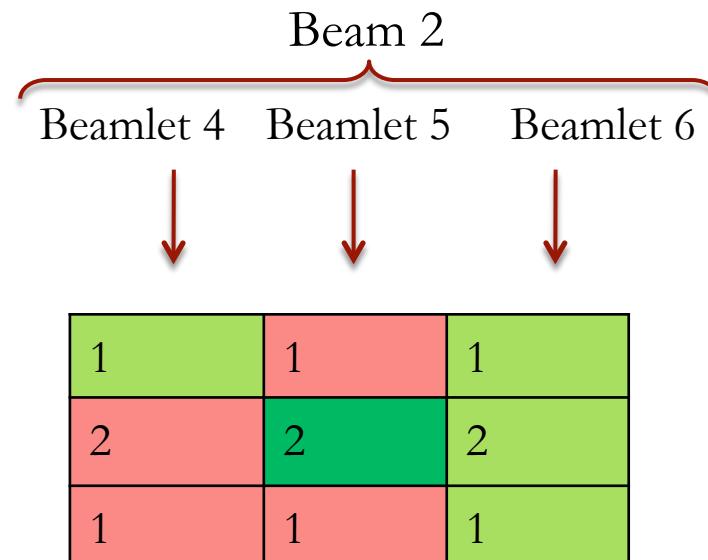
- Minimize total dose to healthy tissue (spinal + other)
- Constraints: tumor voxels at least 7Gy (Gray), spinal cord voxel at most 5Gy

Dose to Each Voxel – Beamlets 1, 2, 3

1	2	2
1	2	2.5
1.5	1.5	2.5

← Beamlet 1
← Beamlet 2
← Beamlet 3 } Beam 1

Dose to Each Voxel – Beamlets 4, 5, 6



Small Example – The Model

1	2	2
1	2	2.5
1.5	1.5	2.5

← Beamlet 1
 ← Beamlet 2
 ← Beamlet 3

Beamlet 4 Beamlet 5 Beamlet 6

1	1	1
2	2	2
1	1	1

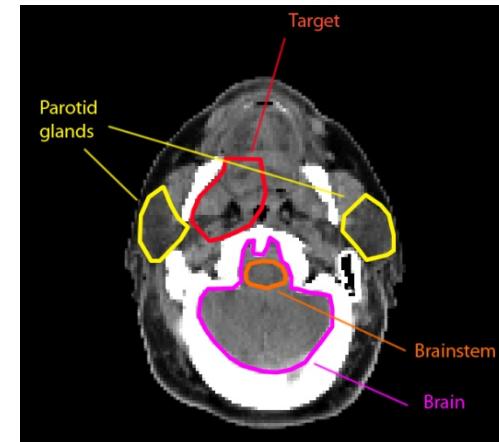
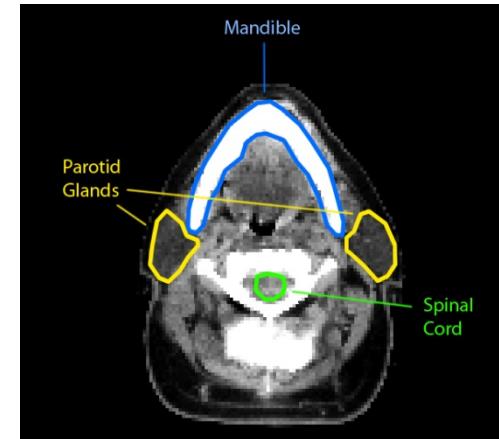
Decisions: $x_1, x_2, x_3, x_4, x_5, x_6$

Minimize $(1+2)x_1 + (2+2.5)x_2 + 2.5x_3 + x_4 + 2x_5 + (1+2+1)x_6$

$$\begin{aligned}
 2x_1 + x_5 &\geq 7 \\
 x_2 + 2x_4 &\geq 7 \\
 1.5x_3 + x_4 &\geq 7 \\
 1.5x_3 + x_5 &\geq 7 \\
 2x_2 + 2x_5 &\leq 5 \\
 x_1, x_2, x_3, x_4, x_5, x_6 &\geq 0
 \end{aligned}$$

A Head and Neck Example

- We will test out this approach on a head-and-neck case
 - Total of 132,878 voxels
 - One target volume (9,777 voxels)
 - Five critical structures: spinal cord, brain, brain stem, parotid glands, mandible (jaw)
 - 5 beams; each beam ~60 beamlets ($1\text{cm} \times 1\text{cm}$) for a total of 328 beamlets



Treatment Plan Criteria

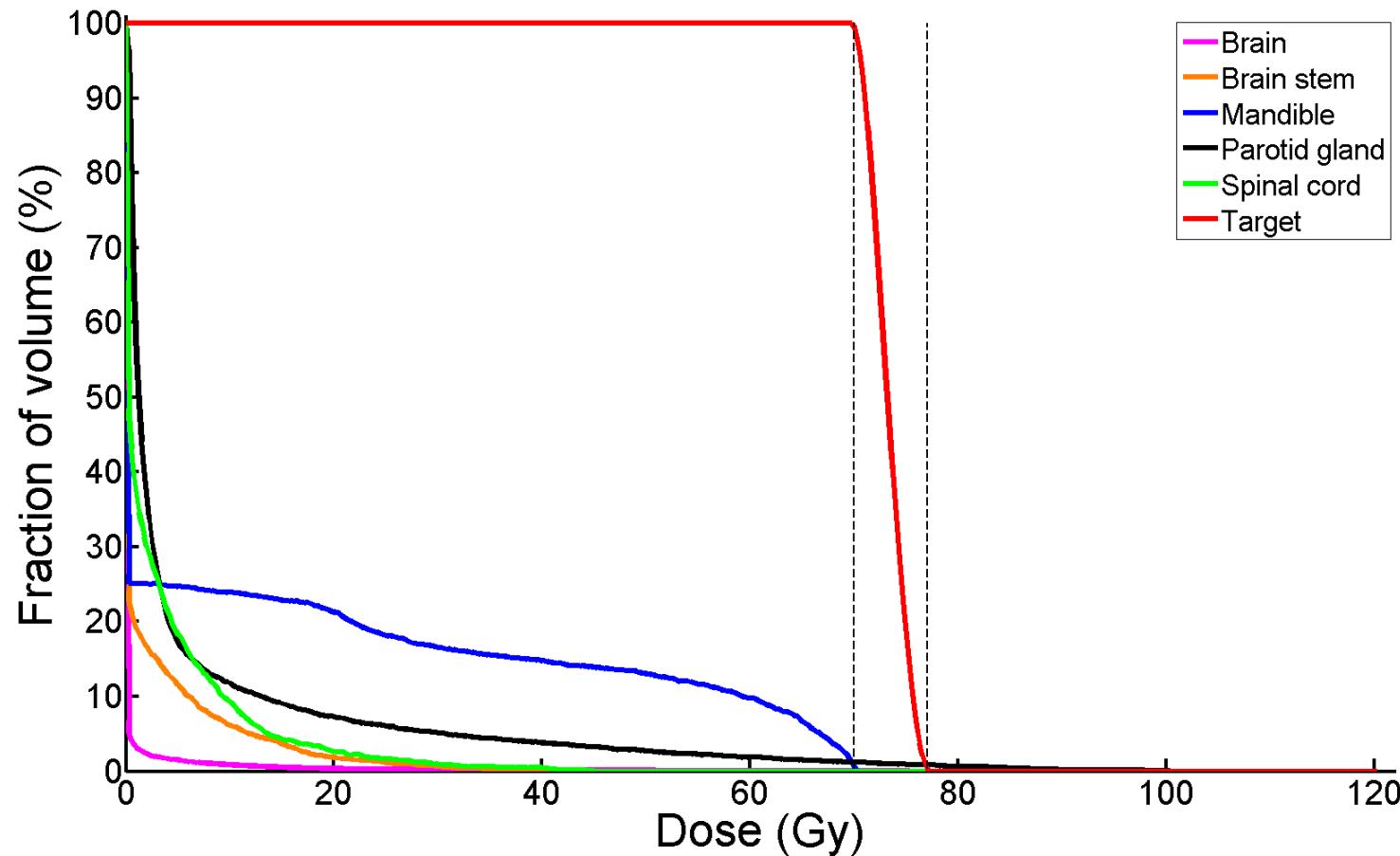
- Dose to whole tumor between 70Gy and 77Gy
- Maximum spinal cord dose at most 45Gy
 - Significant damage to any voxel will result in loss of function
- Maximum brain stem dose at most 54Gy
- Maximum mandible dose at most 70Gy
- Mean parotid gland dose at most 26Gy
 - Parotid gland is a parallel structure: significant damage to any voxel does not jeopardize function of entire organ

The Optimization Problem

minimize Total healthy tissue dose

subject to $70\text{Gy} \leq \text{Dose to voxel } v \leq 77\text{Gy}$, for all tumor voxels v ,
 $\text{Dose to voxel } v \leq 45\text{Gy}$, for all spinal cord voxels v ,
 $\text{Dose to voxel } v \leq 54\text{Gy}$, for all brain stem voxels v ,
 $\text{Dose to voxel } v \leq 70\text{Gy}$, for all mandible voxels v ,
$$\frac{\text{Total parotid dose}}{\text{Num. parotid voxels}} \leq 26\text{Gy},$$
 $w_b \geq 0, \text{ for all beamlets } b.$

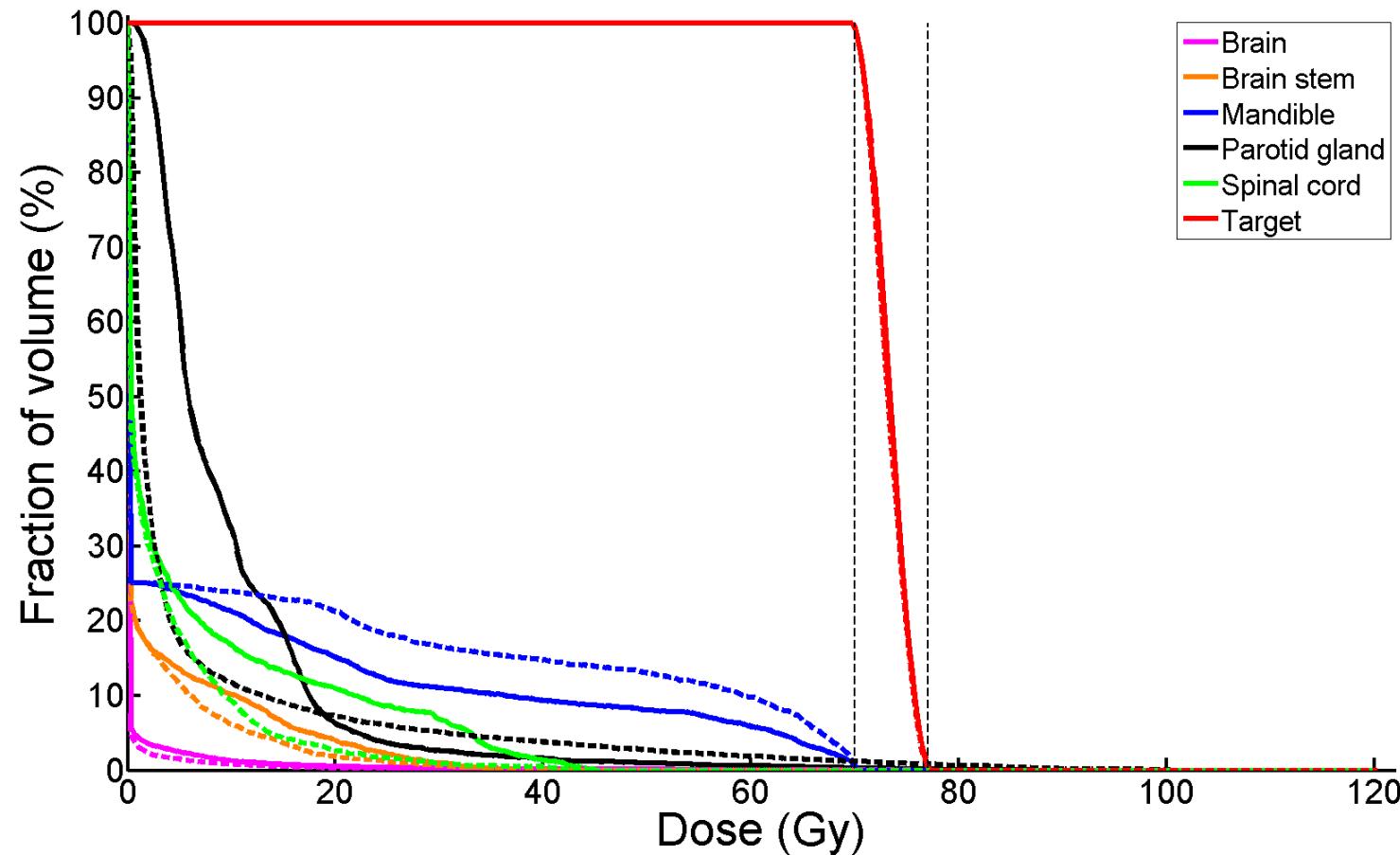
Solution



Exploring Different Solutions

- Mean mandible dose was 11.3Gy – how can we reduce this?
- One approach: modify objective function
 - Current objective is the sum of the total dose
$$T_B + T_{BS} + T_{SC} + T_{PG} + T_M$$
 - Change objective to
$$T_B + T_{BS} + T_{SC} + T_{PG} + 10 \times T_M$$
 - Set mandible weight from 1 (current solution) to 10

New Solution



Sensitivity

- Another way to explore tradeoffs is to modify constraints
 - For example: by relaxing the mandible maximum dose constraint, we may improve our total healthy tissue dose
 - How much does the objective change for different constraints?

Shadow Prices

Organ	Highest shadow price
Parotid gland	0
Spinal cord	96.911
Brain stem	0
Mandible	7399.72

- Parotid gland and brain stem have shadow prices of zero
 - Modifying these constraints is not beneficial
- Mandible has highest shadow price
 - If slight increase in mandible dose is acceptable, total healthy tissue dose can be significantly reduced

IMRT Optimization in Practice



- Radiation machines are connected to treatment planning software that implements and solves optimization models (linear and other types)
 - Pinnacle by Phillips
 - RayStation by RaySearch Labs
 - Eclipse by Varian

Extensions

- Selection of beam angles
 - Beam angles can be selected jointly with intensity profiles using **integer optimization** (topic of next week)
- Uncertainty
 - Often quality of IMRT treatments is degraded due to uncertain organ motion (e.g., in lung cancer, patient breathing)
 - Can manage uncertainty using a method known as **robust optimization**

Efficiency



- Manually designing an IMRT treatment is inefficient and impractical
- Linear optimization provides an *efficient* and *systematic* way of designing an IMRT treatment
 - Clinical criteria can often be modeled using constraints
 - By changing the model, treatment planner can explore tradeoffs

Clinical Benefits

- Ultimately, IMRT benefits the patient
 - In head and neck cancers, saliva glands were rarely spared prior to IMRT; optimized IMRT treatments spare saliva glands
 - In prostate cancer, optimized IMRT treatments reduce toxicities and allow for higher tumor doses to be delivered safely
 - In lung cancer, optimized IMRT reduces risk of radiation-induced pneumonitis



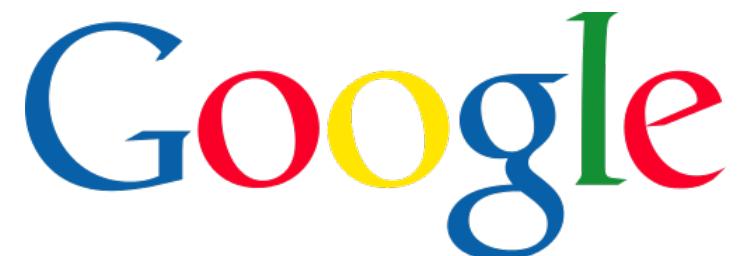
GOOGLE ADWORDS

Optimizing Online Advertising

15.071x – The Analytics Edge

Google Inc.

- Provides products and services related to the Internet
 - Mission: “... to organize the world’s information and make it universally accessible and useful.”
- Most widely known for its search engine
(www.google.com)
 - User enters a query; Google returns links to websites that best fit query



History of Google



- 1996 – Sergei Brin and Larry Page, graduate students at Stanford, working on a research project
 - How to measure importance of any webpage using links on the internet
- 1998 – Incorporated Google as a company and received first funding; database of 60 million webpages
- 2004 – Initial Public Offering
- 2007 – Google acquires YouTube and other companies
- 2013 – **More than 1 billion unique monthly visitors**

Google's Business Model



- Google search engine is free to use, so how does Google make money?
- Answer: **online advertising**

Example of Sponsored Ads

The screenshot shows a Google search results page for the query "nine inch nails tickets". The results are filtered by "Web" and show approximately 4,630,000 results found in 0.27 seconds. The page includes a navigation bar with links to Apple, iCloud, Facebook, Twitter, Wikipedia, Yahoo!, News, and Popular. A "Sign in" button is also present.

Ads related to nine inch nails tickets

- Tickets at StubHub - Sports, Concert, & Theater Tickets**
www.stubhub.com/ ▾
The Only FanProtect Guarantee.
The first choice for ticket bargains – SportsBusiness Daily
StubHub has 1,901 followers on Google+
Concert Tickets - Deals In Your Area - Buy Tickets, Earn Rewards - Sell Tickets
- 40% Nine Inch Nails Tix - Prices Slashed for a Limited-Time**
www.goodseatstickets.com/Nine-Inch-Nails ▾ ★★★★★ 62 seller reviews
Don't Miss out. Buy Tickets today.
Good Seat Tickets has 441 followers on Google+
Compare Us - Get \$15 Coupon - 40% Discount
- 30% Off 9 Inch Nails Tix - Event Happening Soon?**
www.bargainseatsonline.com/ ▾ ★★★★★ 46 seller reviews
Save Up To 30% On Select Tickets. Buy Now!
- nine inch nails Tickets - Ticketmaster**
www.ticketmaster.com ▾ All Tickets ▾ Music Tickets ▾ Alternative Rock ▾
Buy nine inch nails tickets from the official Ticketmaster.com site. Find nine inch nails tour schedule, concert details, reviews and photos.
- NIN Tour**
tour.nin.com/ ▾
Nine Inch Nails. ... DATE, CITY, VENUE, PUBLIC TICKETS. 03.06.2014, Sydney, Australia, Entertainment Centre w/ Queens Of The Stone Age & Brody Dalle ...
Tickets - Sign in - Nin.com presale ticketing fan - Sign up for a nin.com account

Ads

- 50% Off Nine Inch Nails**
nineinchnails.pricesavertickets.com/ ▾
Up To \$75 Off Use Code SAVENOW
Nobody Beats Nine Inch Nails Ticket
- Event Tickets Cheap**
www.ticketliquidator.com/ ▾
★★★★★ 131 seller reviews
Concerts, Sports & Theatre Tickets.
Compare Prices. Save 15% or More.
- Nine Inch Nails Tickets**
www.tickets2get.com/ ▾
Get Great Ticket Selection & Prices on Nine Inch Nails Concerts
- Nine Inch Nails Tickets**
nin.eventicketsexpress.com/ ▾
1 (866) 702 9901
Low Low Prices On All Event Tickets
30-40% Less Than The Competition!
- Nine Inch Nails Concert**
www.amazon.com/dvd ▾
★★★★★ 367 reviews for amazon.com
Save up to 35% on top sellers.
Free Shipping on Qualified Orders.

Google Advertising - AdWords



- Why do companies advertise on Google?
 - Google receives heavy traffic
 - Search pages are formatted in a very clean way
 - Companies can choose which types of queries their ads will be displayed for; better targeting
- 97% of Google's revenues come from AdWords

How does Advertising on Google work?



1. Advertisers place bids for different queries in an auction
2. Based on bids and *quality score* (fit of advertiser and ad to the queries), Google decides price-per-click of each advertiser and each query
3. Google then decides how often to display each ad for each query

Price-per-click (PPC)



- For each query, Google decides each advertiser's **price-per-click (PPC)**
 - How much advertiser pays Google when user clicks ad for that query
- Each advertiser also specifies a **budget**
 - Each time user clicks on advertiser's ad, budget is depleted by PPC amount

Example of price-per-click



Advertiser	Query 1 ("4G LTE")	Query 2 ("largest LTE")	Query 3 ("best LTE network")
AT&T	\$5	\$5	\$20
T-Mobile	\$10	\$5	\$20
Verizon	\$5	\$20	\$25

Advertiser	Budget
AT&T	\$170
T-Mobile	\$100
Verizon	\$160



-T--Mobile-- **verizon**

\$100

$$5 \times \$10 = \frac{\$50}{\$50}$$

Click-through Rate (CTR)



- Advertiser only pays Google if the user *clicks* on the ad
- The probability that a user clicks on an advertiser's ad is the **click-through rate (CTR)**
 - Can also think of as “clicks per user”

Example of click-through rate



Advertiser	Query 1 ("4G LTE")	Query 2 ("largest LTE")	Query 3 ("best LTE network")
AT&T	0.10	0.10	0.08
T-Mobile	0.10	0.15	0.10
Verizon	0.10	0.20	0.20

$$50 \text{ users} \times 0.08 = 4 \text{ users}$$

$$100 \text{ users} \times 0.08 = 8 \text{ users}$$

Average Price Per Display



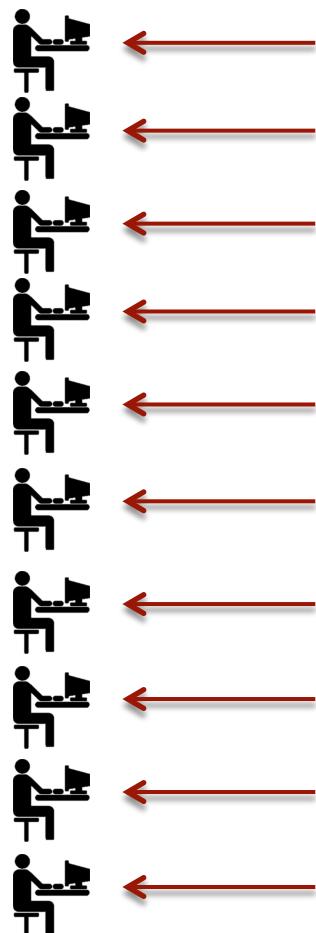
- Average amount that an advertiser pays each time its ad is shown is **PPC x CTR**

How average price per display works

- 
- “best LTE network”
- 
- “best LTE network”
- 
- “best LTE network”
- 
- “best LTE network”
- 
- “best LTE network”
- 
- “best LTE network”
- 
- “best LTE network”
- 
- “best LTE network”
- 
- “best LTE network”
- 
- “best LTE network”
-

Suppose 10 users
search for
“best LTE network”

How average price per display works



verizon
verizon
verizon
verizon
verizon
verizon
verizon
verizon
verizon
verizon

Google decides
to display
Verizon's ad

verizon

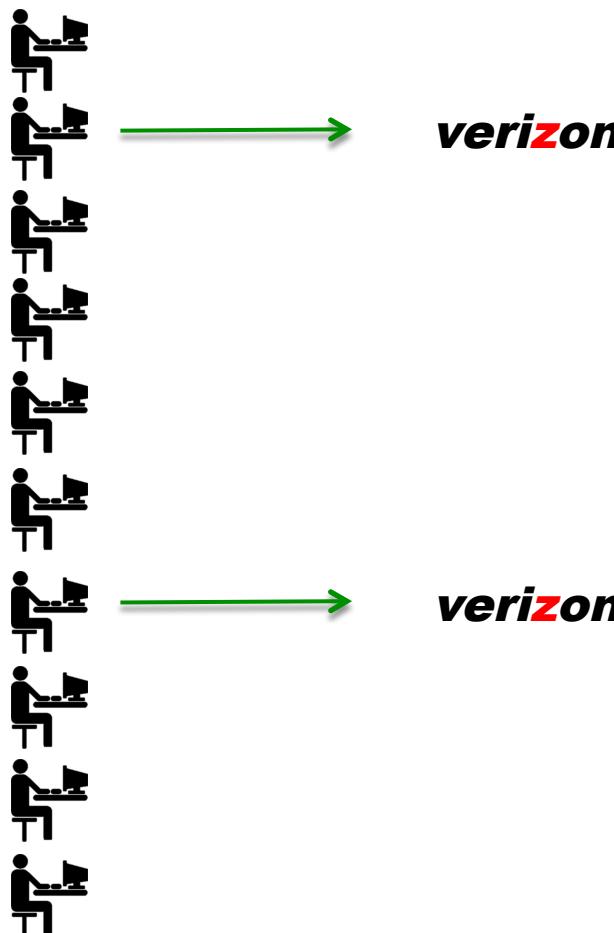
How average price per display works



verizon
verizon
verizon
verizon
verizon
verizon
verizon
verizon
verizon

CTR of Verizon
and “best LTE
network” is 0.2,
so only 2 users click
on the ad

How average price per display works



Verizon pays PPC
for each user:

2 clicks

× \$25 per click

= \$50

How average price per display works



verizon

Verizon pays on
average per user/
display:



verizon

$$\begin{aligned} & \$50 \\ & \div 10 \text{ displays} \\ & = \$5 \text{ per display} \end{aligned}$$

How average price per display works



verizon



verizon

This is exactly the
PPC multiplied
by the CTR:

\$25 per click

× 0.20 clicks per user

= \$5 per user/display

Average price per display for example

Advertiser	Query 1 PPC ("4G LTE")	Query 2 PPC ("largest LTE")	Query 3 PPC ("best LTE network")
AT&T	\$5	\$5	\$20
T-Mobile	\$10	\$5	\$20
Verizon	\$5	\$20	\$25

Advertiser	Query 1 CTR ("4G LTE")	Query 2 CTR ("largest LTE")	Query 3 CTR ("best LTE network")
AT&T	0.10	0.10	0.08
T-Mobile	0.10	0.15	0.10
Verizon	0.10	0.20	0.20

Average price per display for example

Advertiser	Query 1 APPD ("4G LTE")	Query 2 APPD ("largest LTE")	Query 3 APPD ("best LTE network")
AT&T	\$0.50	\$0.50	\$1.60
T-Mobile	\$1.00	\$0.75	\$2.00
Verizon	\$0.50	\$4.00	\$5.00

Query estimates



- Google does not control how many times a query will be requested – driven by users!
- For each query, Google has estimate of number of times query will be requested over a given day

Example of query estimates

Query	Est. # of Requests
“4G LTE”	140
“largest LTE”	80
“best LTE network”	80

Google's problem



- **How many times to display each ad for each query to maximize revenue**

Google's problem

- Objective:
 - Maximize revenue
- Decision:
 - For each advertiser and query, number of times ad will be displayed for that query
- Constraints:
 - Average amount paid by each advertiser cannot exceed budget
 - Total ads for given query cannot exceed estimated number of requests for that query

Problem data

Advertiser	Avg. \$ / Query 1 Ad Display	Avg. \$ / Query 2 Ad Display	Avg. \$ / Query 3 Ad Display
AT&T	\$0.50	\$0.50	\$1.60
T-Mobile	\$1.00	\$0.75	\$2.00
Verizon	\$0.50	\$4.00	\$5.00

Advertiser	Budget
AT&T	\$170
T-Mobile	\$100
Verizon	\$160

Query	Est. # of Requests
Q1 (“4G LTE”)	140
Q2 (“largest LTE”)	80
Q3 (“best LTE network”)	80

Modeling the problem

- Decision variables:

$$x_{A1} \quad x_{A2} \quad x_{A3} \quad x_{T1} \quad x_{T2} \quad x_{T3} \quad x_{v1} \quad x_{v2} \quad x_{v3}$$

- Revenue to Google under ad strategy:

$$0.50x_{A1} + 0.50x_{A2} + 1.60x_{A3} + 1.00x_{T1} + \dots + 5.00x_{v3}$$

- Amount advertiser AT&T pays in ad strategy:

$$0.50x_{A1} + 0.50x_{A2} + 1.60x_{A3} \leq 170$$

- Number of times ad strategy uses query 2:

$$x_{A2} + x_{T2} + x_{v2} \leq 80$$

Let's do it in LibreOffice



Extensions to the problem



- Slates/positions
- Personalization
- Other issues
 - Estimating CTRs
 - How should advertisers bid?

Slates/positions

- Search result page has space for more than one ad
- **Slate:** combination of ads
- Many possible slates: which ones to display?

$x_{\text{advertiser, query}}$

$x_{A1} \quad x_{T1} \quad x_{V1}$

$x_{\text{slate, query}}$

$x_{AT1} \quad x_{AV1} \quad x_{TV1}$

$x_{TA1} \quad x_{VA1} \quad x_{VT1}$

↑↑

Personalization

- In addition to the query, Google can use other information to decide which ad to display:
 - IP address/geographic location
 - Previous Google searches/browser activity on Google
- How do we account for this?

$X_{\text{advertiser}, \text{query}}$

$X_{A|p_i}$

$X_{\text{advertiser}, \text{query}, \text{user profile}}$

$P_1 \ P_2 \ P_3$

AdWords at Google's scale



- We studied a small instance of the ad allocation problem
 - 3 bidders, 3 queries
- We saw how an optimization solution increases revenue by 16% over “common-sense” solution
- In reality, problem is *much* larger
 - Hundreds to thousands of bidders, over \$40 billion
 - Gains from optimization at this scale become *enormous*



REVENUE MANAGEMENT

An Introduction to Linear Optimization

15.071x – The Analytics Edge

Airline Regulation (1938-1978)



- The Civil Aeronautics Board (CAB) set fares, routes, and schedules for all interstate air transport
- Most major airlines favored this system due to guaranteed profits
- Led to inefficiency and higher costs
 - Applications for new routes and fares often delayed or dismissed

Airline Deregulation (1978)

- The administration of President Jimmy Carter passed the Airline Deregulation Act in 1978
- The Act encouraged
 - **More competition:** 52 new airlines between 1980 and 2000
 - **New air routes:** saved passengers an estimated \$10.3 billion each year in travel time
 - **Lower fares:** ticket prices are 40% lower today than they were in 1978
- This led to **more passengers**
 - The number of air passengers increased from 207.5 million in 1974 to 721.1 million in 2010

A Competitive Edge



- More competition led to heavy losses by air carriers
 - Need to lower fares while meeting operating costs
- 9 major carriers and more than 100 smaller airlines went bankrupt between 1978 and 2002
- How did airlines compete?

Discount Fares

- On January 17, 1985 American Airlines (AA) launched its Ultimate Super Saver fares to compete with PeopleExpress
- Need to fill at least a minimum number of seats without selling every seat at discount prices
 - Sell enough seats to cover fixed operating costs
 - Sell remaining seats at higher rates to maximize revenues/profits

How Many Seats to Sell on Discount?

- Passengers have different valuations
 - Business people value flexibility (last-minute/refundable)
 - People seeking getaways value good deals (early birds)
- Sell too many discounted seats
 - Not enough seats for high-paying passengers
- Sell too few discounted seats
 - Empty seats at takeoff implying lost revenue
- How should AA allocate its seats among customers to maximize its revenue?

Let's Start Simple



Ticket Prices



**Lowest Fare
from \$238**

Flights	Departure	Arrival	Choice
3 [+]	12:00 pm JFK	03:10 pm LAX	 \$238 2 Seats left

Early Bird

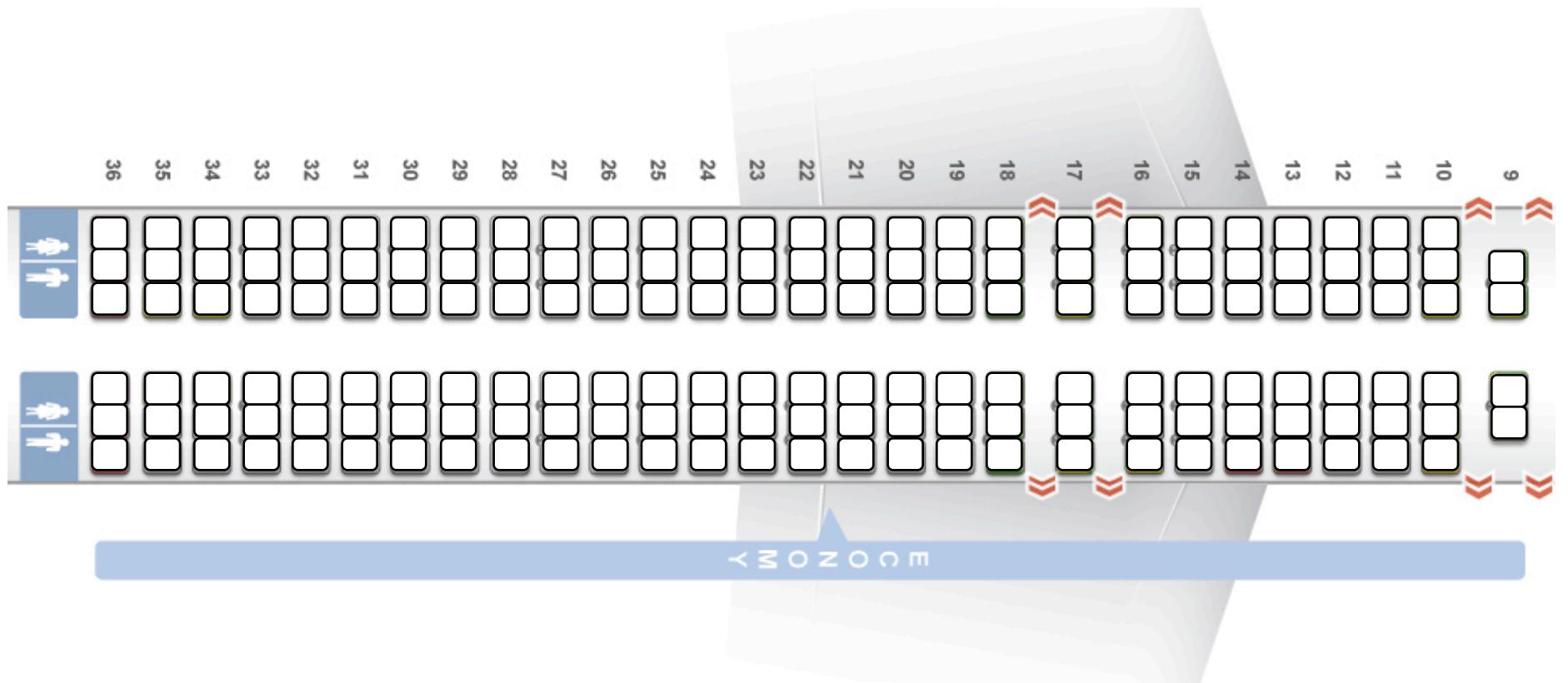
**Lowest Fare
from \$617**

Flights	Departure	Arrival	Choice
3 [+]	12:00 pm JFK	02:55 pm LAX	 \$617

Last minute

Boeing 757-200 Seat Map

- 166 Economy seats



Demand Forecasting

- Demand for different prices can be forecasted using analytics tools, looking at historical data and incorporating models of human behavior
 - Time series methods
 - Linear regression
- Forecasts could be erroneous
 - Need to assess sensitivity to forecast errors
- We'll assume that demand has been forecasted

Myopic Solution

		Price	Demand	Seats to Sell	
JFK	Regular	617	50	50	Capacity 166
LAX	Discount	238	150	116	

- How many discount seats to sell to maximize revenue?

Myopic Solution

		Price	Demand	Seats to Sell	
JFK	Regular	617	100	100	Capacity 166
LAX	Discount	238	150	66	

- How many discount seats to sell to maximize revenue?

Myopic Solution

		Price	Demand	Seats to Sell	
JFK	Regular	617	200	166	Capacity 166
LAX	Discount	238	150	0	

- How many discount seats to sell to maximize revenue?
- This seems simple, but what if we had 100 different flights?
- In the next video, we'll see how to formulate this mathematically

Single Route Example

		Price	Demand	Seats to Sell	
JFK	Regular	617	100		
LAX	Discount	238	150		

Capacity
166

- Problem: Find the optimal number of discounted seats and regular seats to sell to maximize revenue
- Let's formulate the problem mathematically

Step 1. Decisions

		Price	Demand	Seats to Sell	
JFK	Regular	617	100		Capacity 166
LAX	Discount	238	150		

- What are our decisions?
 - Number of regular seats to sell – R
 - Number of discount seats to sell – D

Step 2. Objective

		Price	Demand	Seats to Sell	
JFK	Regular	617	100		
LAX	Discount	238	150		

Capacity
166

- What is our objective?
 - Maximizing total airline revenue
 - Revenue from each type of seat is equal to the number of that type of seat sold times the seat price

$$\max \quad 617 * R + 238 * D$$

Step 3. Constraints

		Price	Demand	Seats to Sell	
JFK	Regular	617	100		
LAX	Discount	238	150		

 Capacity
166

- AA cannot sell more seats than the aircraft capacity
 - Total number of seats sold cannot exceed capacity
$$R + D \leq 166$$
- AA cannot sell more seats than the demand
 - Regular seats sold cannot exceed 100 $R \leq 100$
 - Discount seats sold cannot exceed 150 $D \leq 150$

Step 4. Non-Negativity

		Price	Demand	Seats to Sell	
JFK	Regular	617	100		Capacity 166
LAX	Discount	238	150		

- AA cannot sell a negative number of seats

$$R \geq 0 \quad D \geq 0$$

Problem Formulation

		Price	Demand	Seats to Sell	
JFK	Regular	617	100		
LAX	Discount	238	150		

Capacity
166

Maximize Total airline revenue

Subject to Seats sold cannot exceed capacity

Seats sold cannot exceed demand

Seats sold cannot be negative

Problem Formulation

		Price	Demand	Seats to Sell	
JFK	Regular	617	100		
LAX	Discount	238	150		

Capacity
166

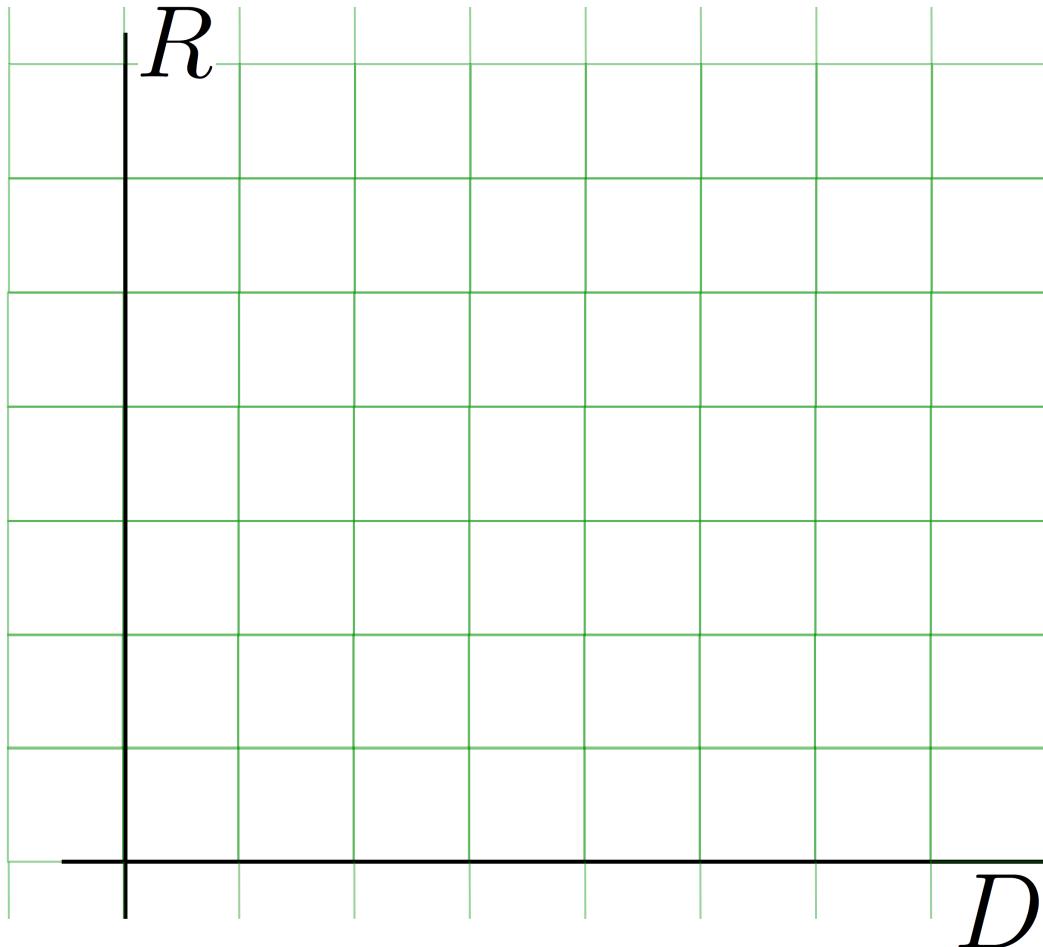
$$\text{Maximize } 617R + 238D$$

$$\text{Subject to } R + D \leq 166$$

$$R \leq 100, D \leq 150$$

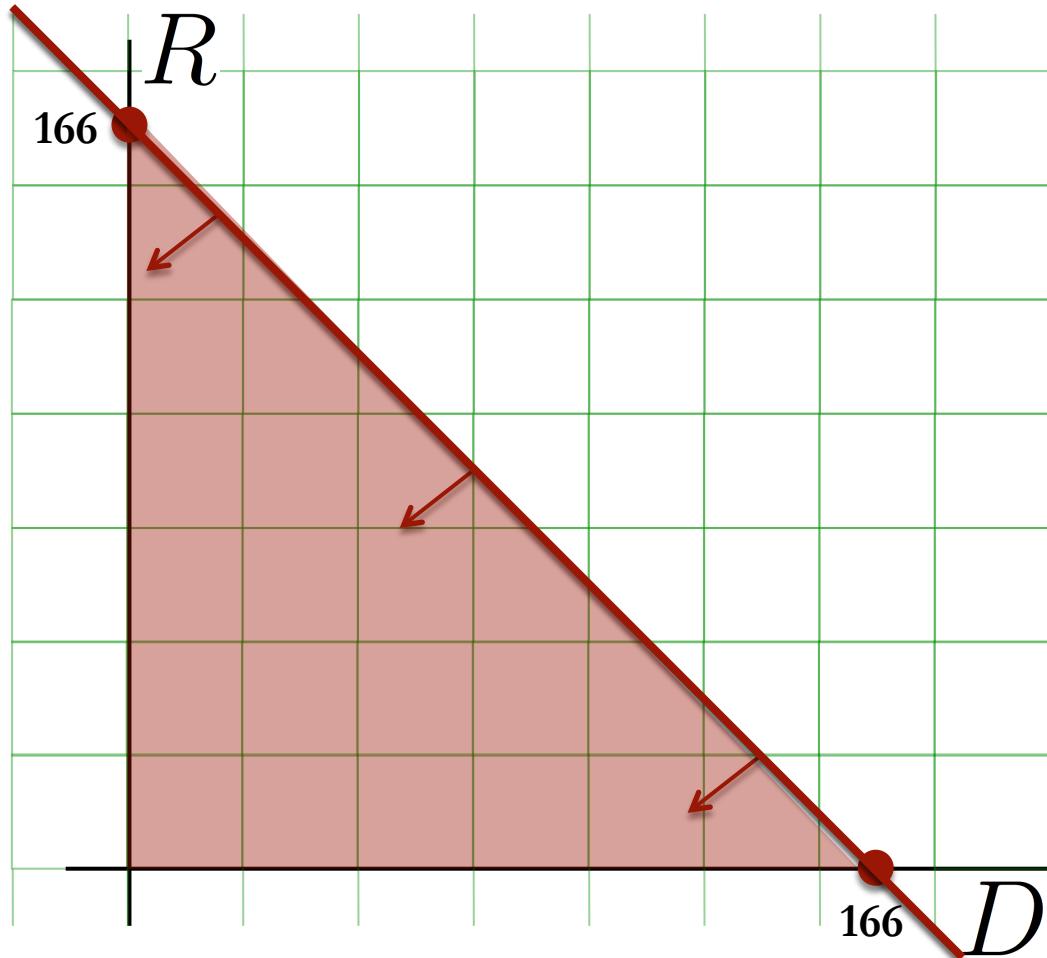
$$R \geq 0, D \geq 0$$

Visualizing the Problem



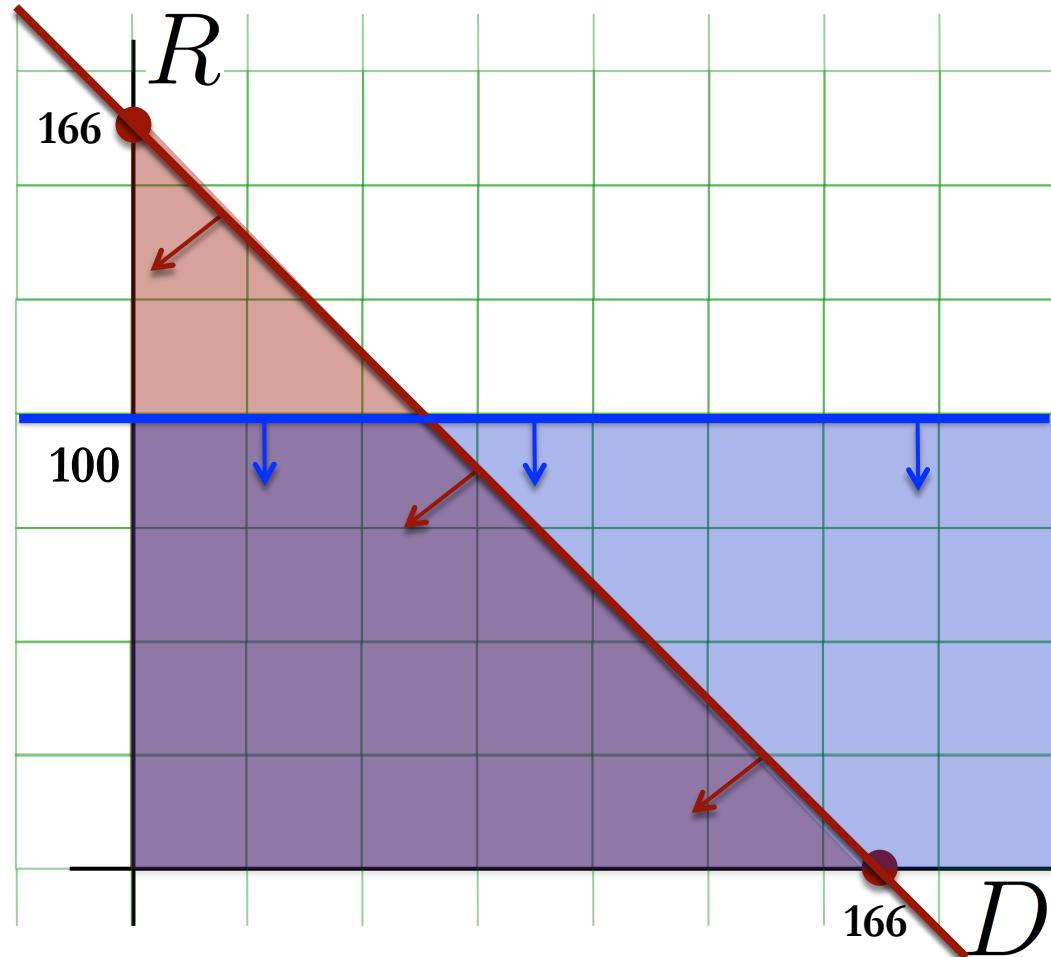
- 2D Representation
- Constraints
 - Non-negativity
 $R \geq 0, D \geq 0$

Visualizing the Problem



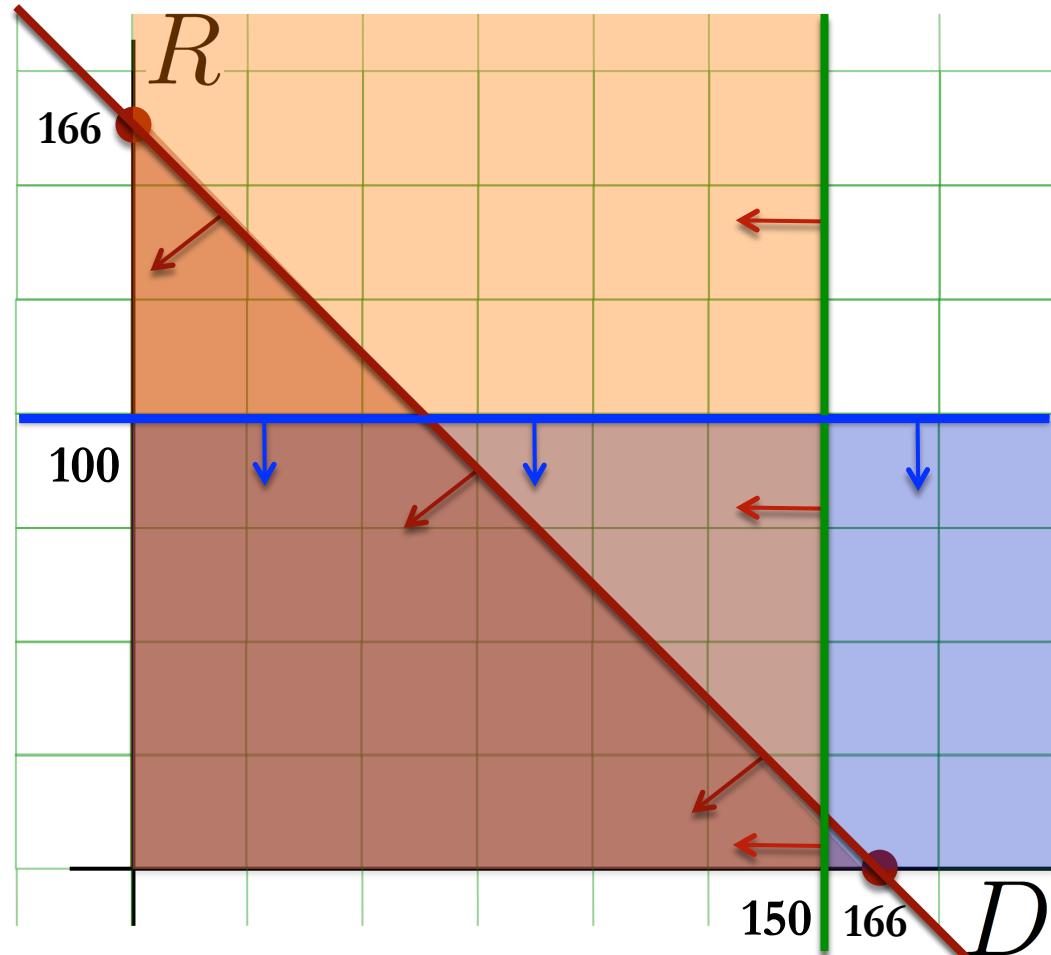
- 2D Representation
- Constraints
 - Non-negativity
 $R \geq 0, D \geq 0$
 - Capacity
 $R + D \leq 166$

Visualizing the Problem



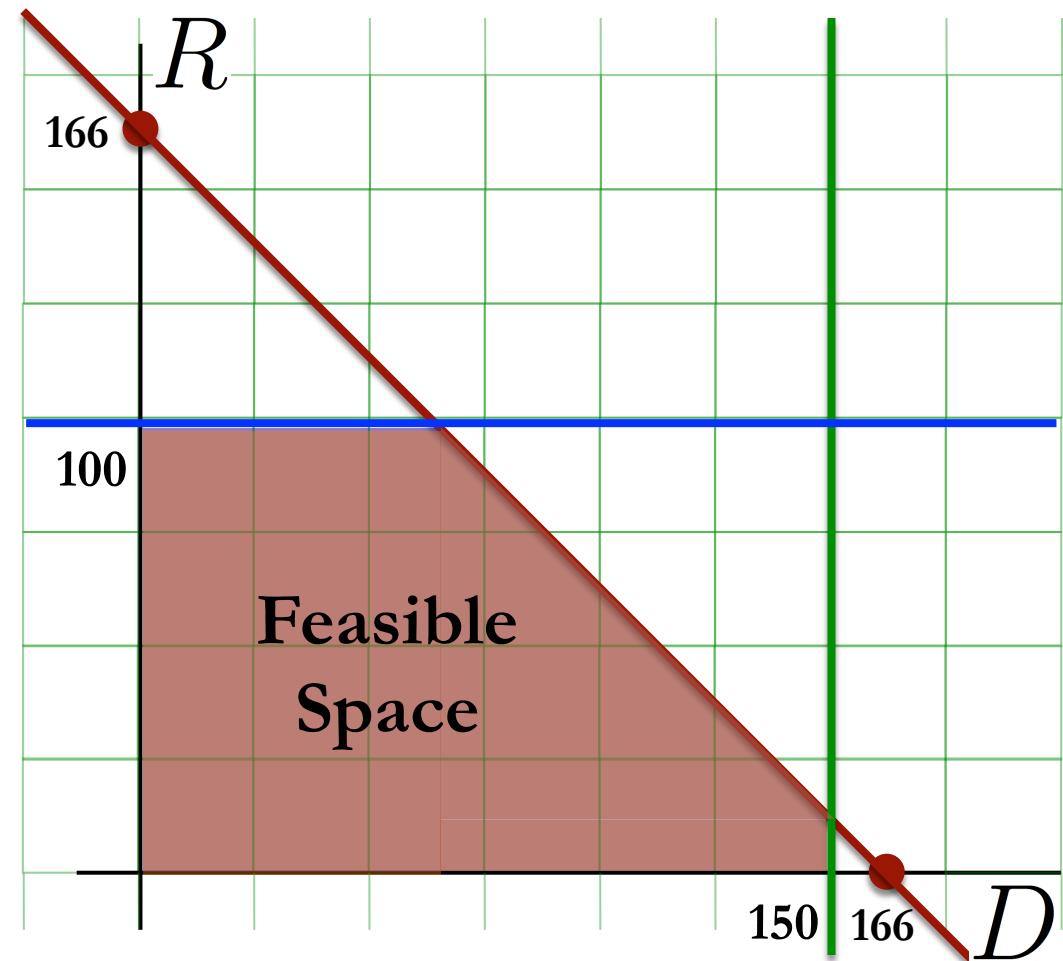
- 2D Representation
- Constraints
 - Non-negativity
 $R \geq 0, D \geq 0$
 - Capacity
 $R + D \leq 166$
 - Demand
 $R \leq 100, D \leq 150$

Visualizing the Problem

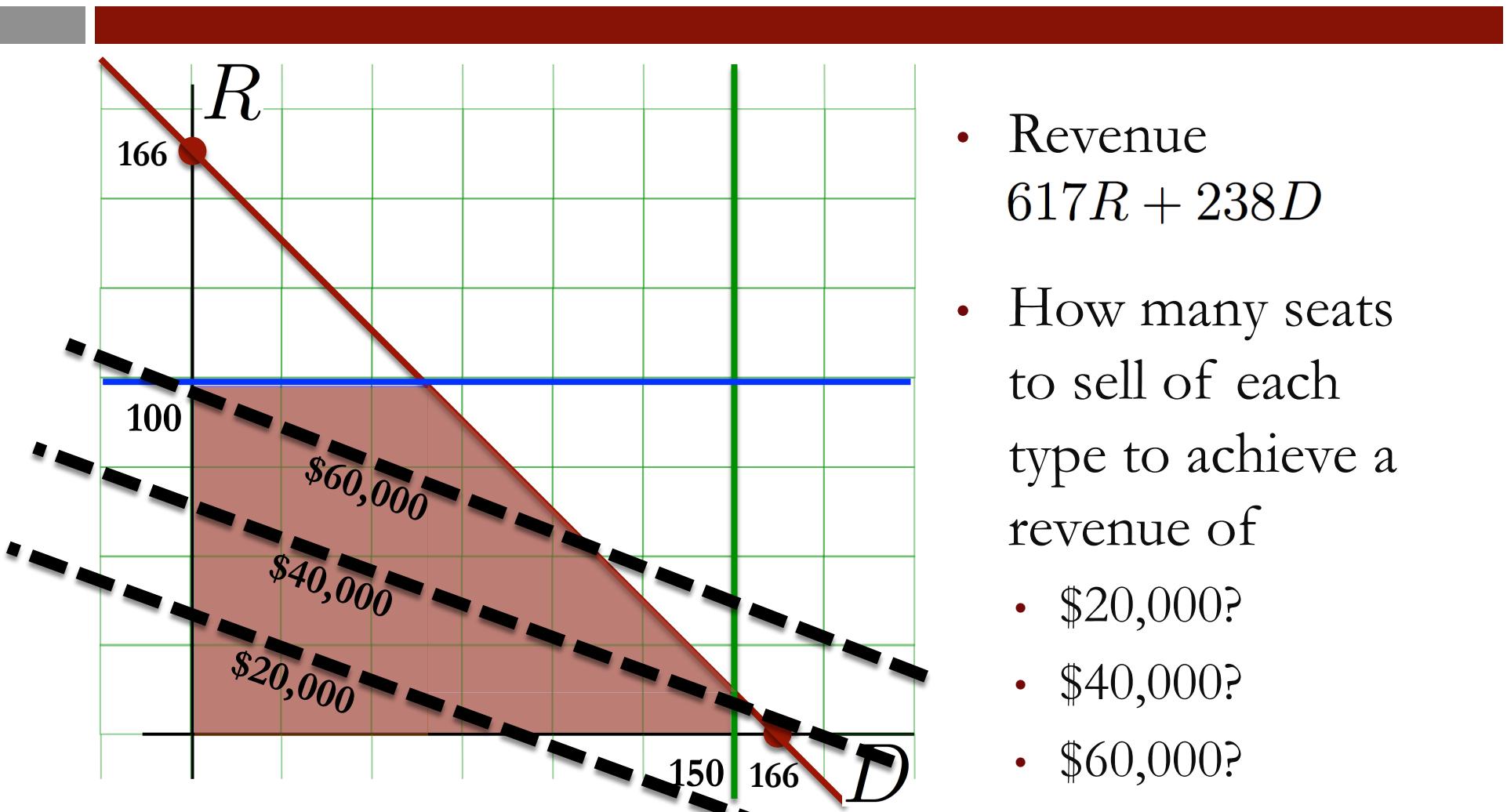


- 2D Representation
- Constraints
 - Non-negativity
 $R \geq 0, D \geq 0$
 - Capacity
 $R + D \leq 166$
 - Demand
 $R \leq 100, D \leq 150$

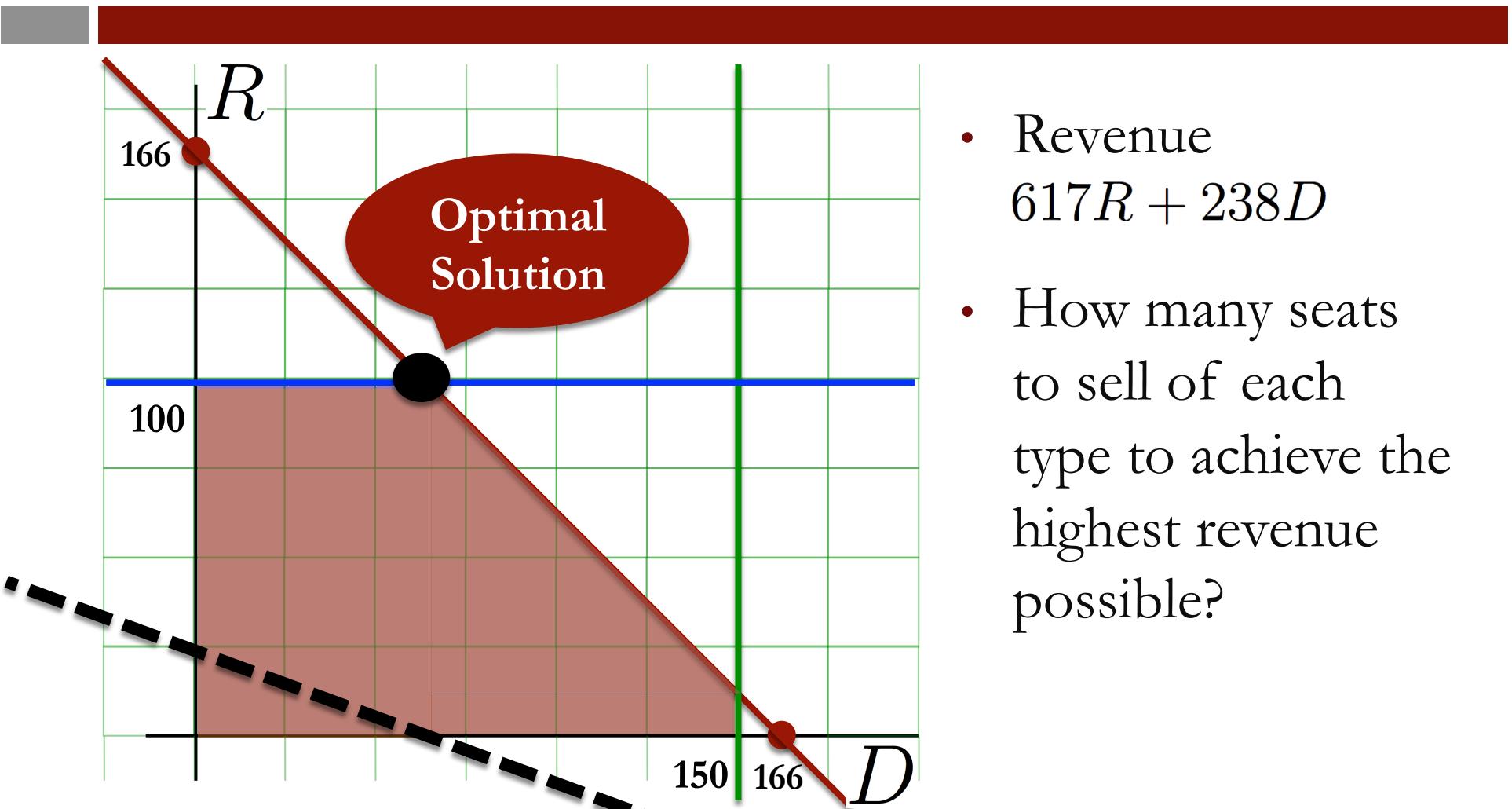
Feasible Space



Possible Solutions



Best Solution

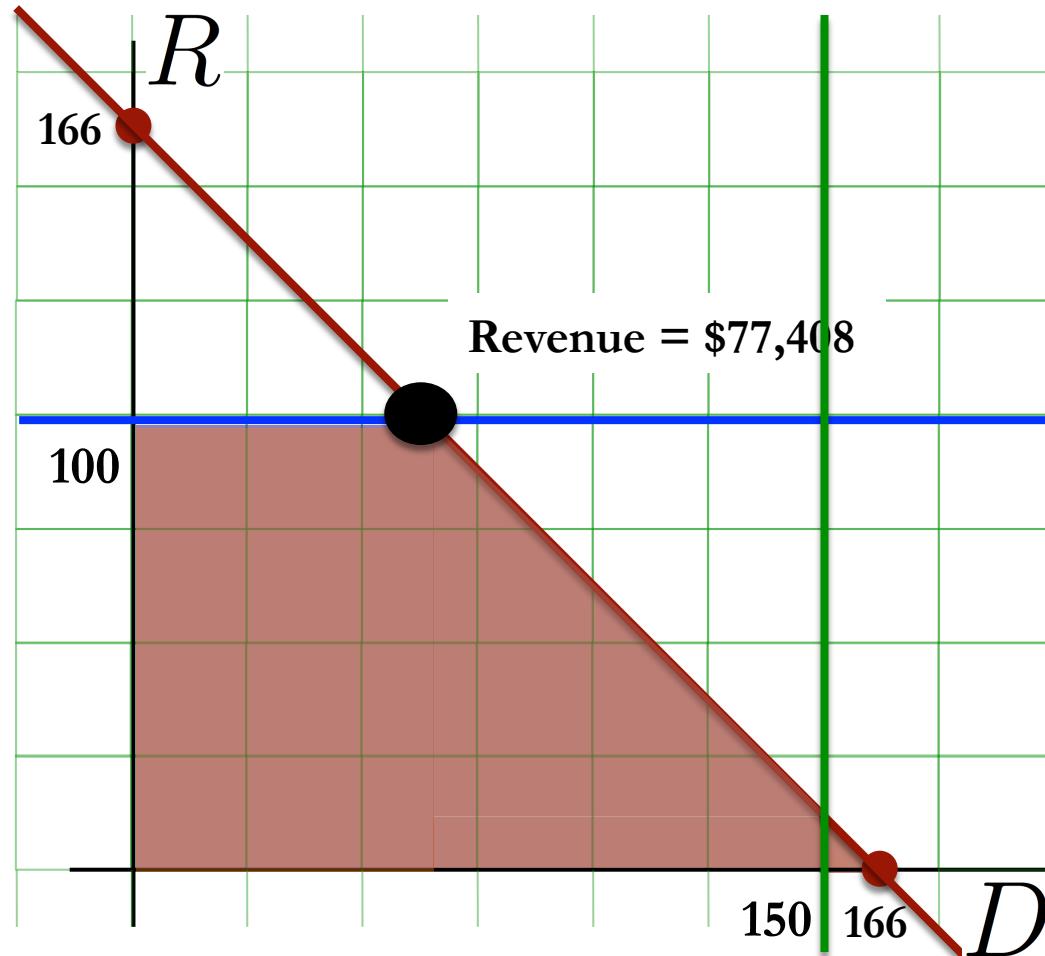


Marketing Decisions

- Management is trying to figure out whether it would be beneficial to invest in marketing its fares
- AA forecasts that its marketing effort is likely to attract one more unit of demand per **\$200 spent**

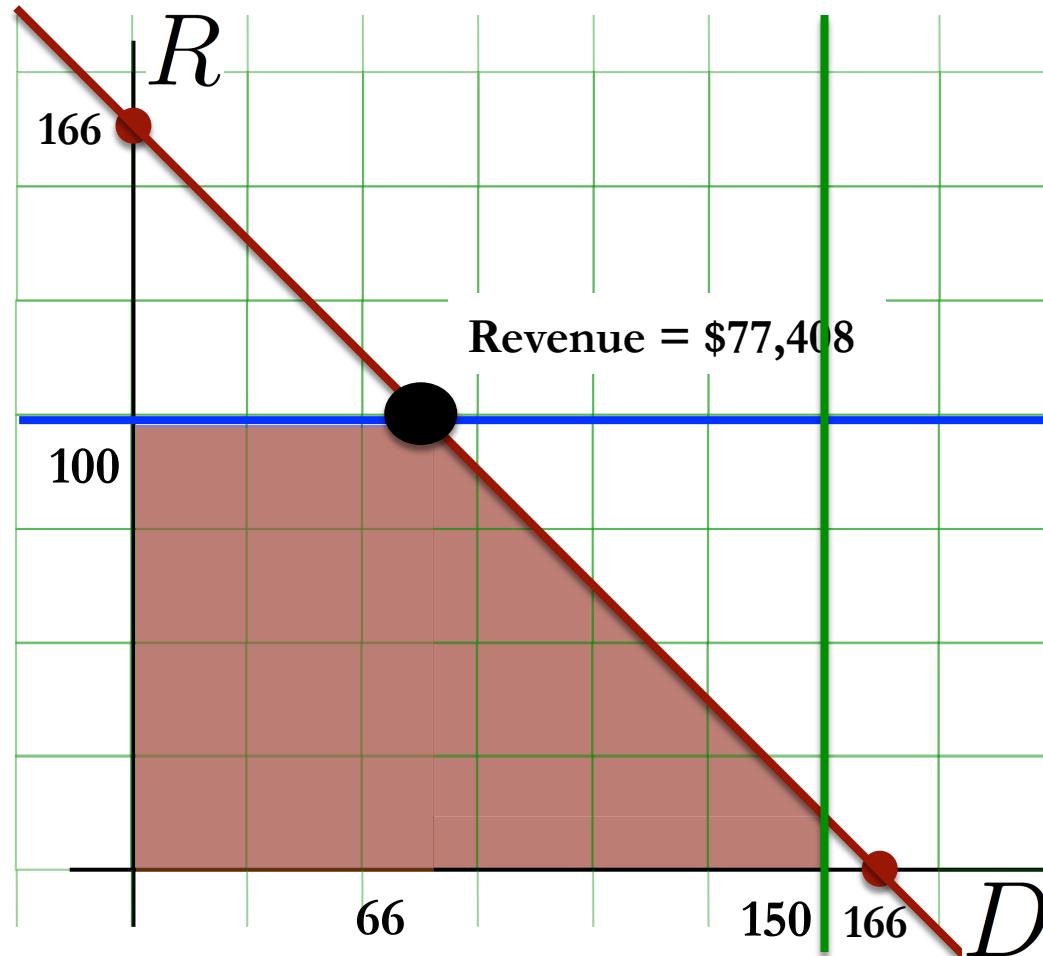
	Marketing Cost/unit	Marginal Revenue
Discount Fare	\$200	
Regular Fare	\$200	

Marketing Discount Fares



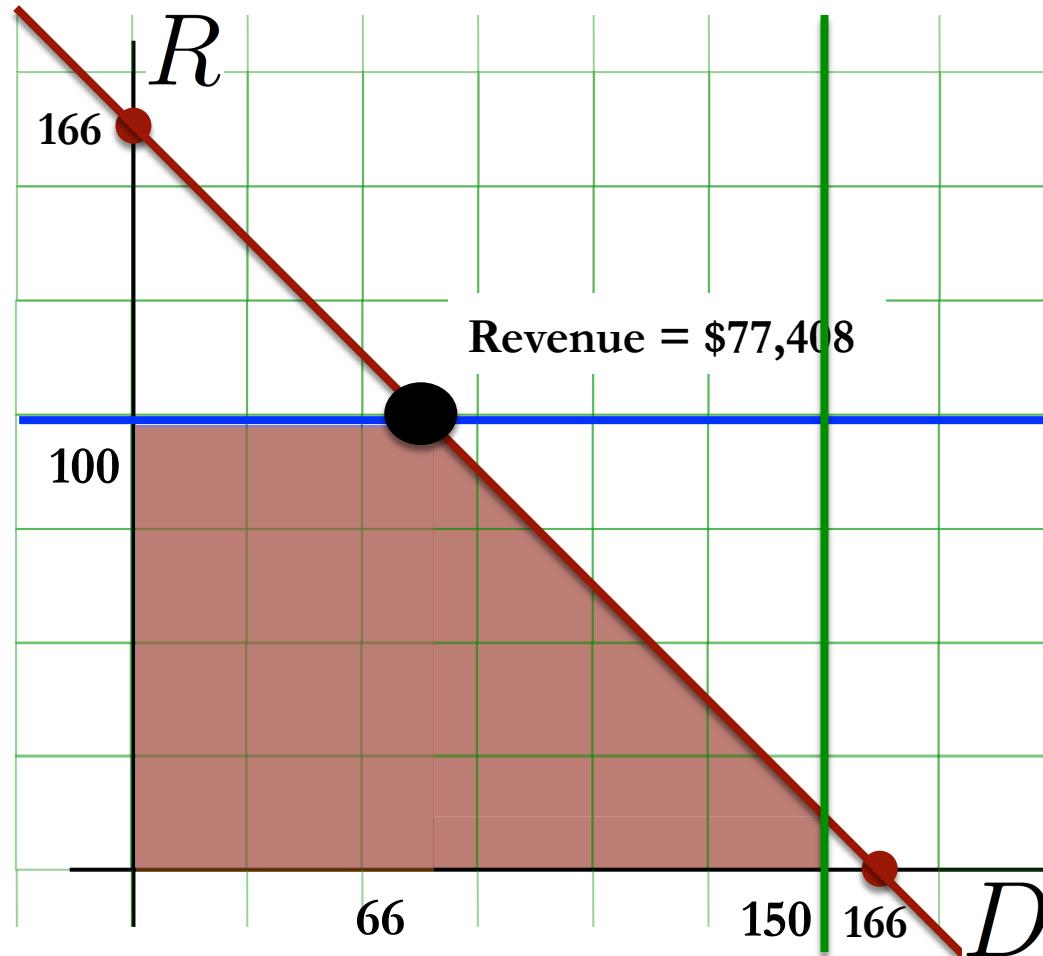
- What if AA increases its marketing budget for discount fares
- Higher demand for discount class
 - 150
 - 175
 - 200

Marketing Discount Fares



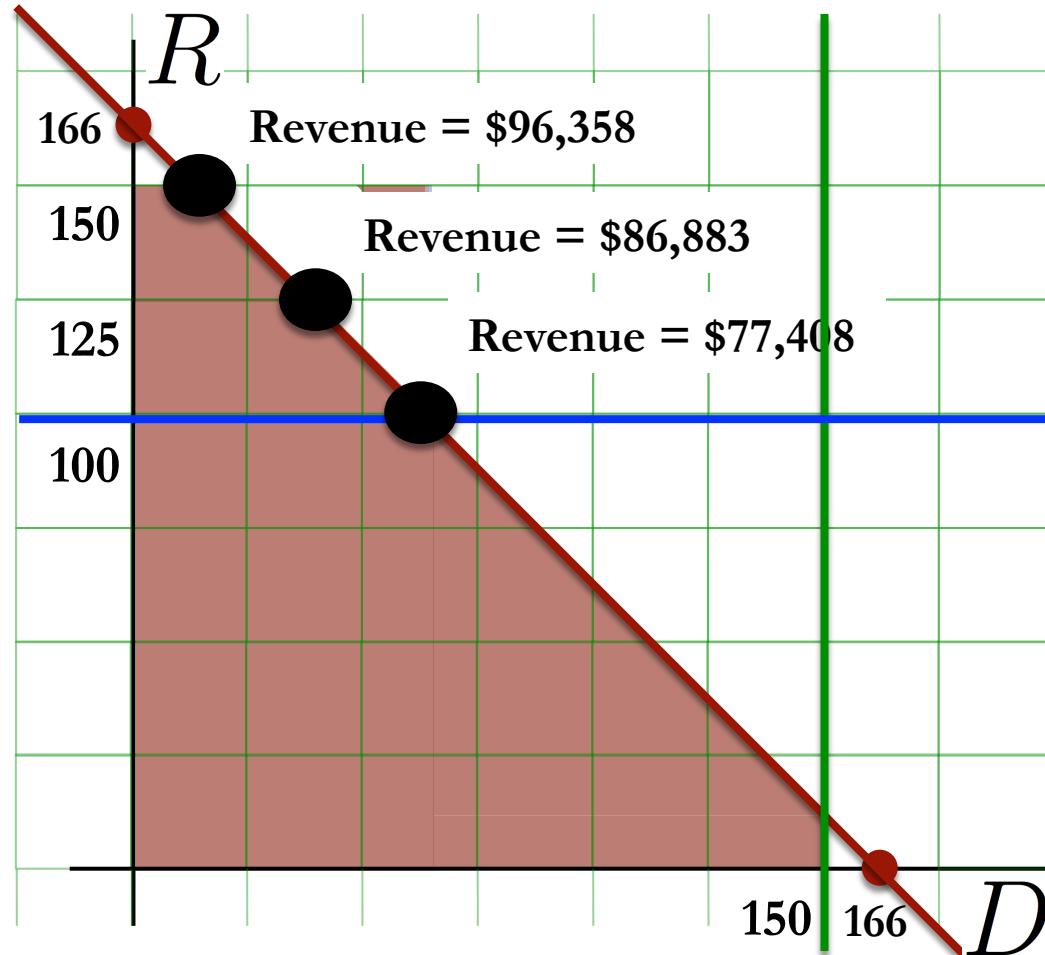
- What if AA decreases its budget to market discount fares?
- Lower demand for discount fare without affecting revenue

Marketing Discount Fares



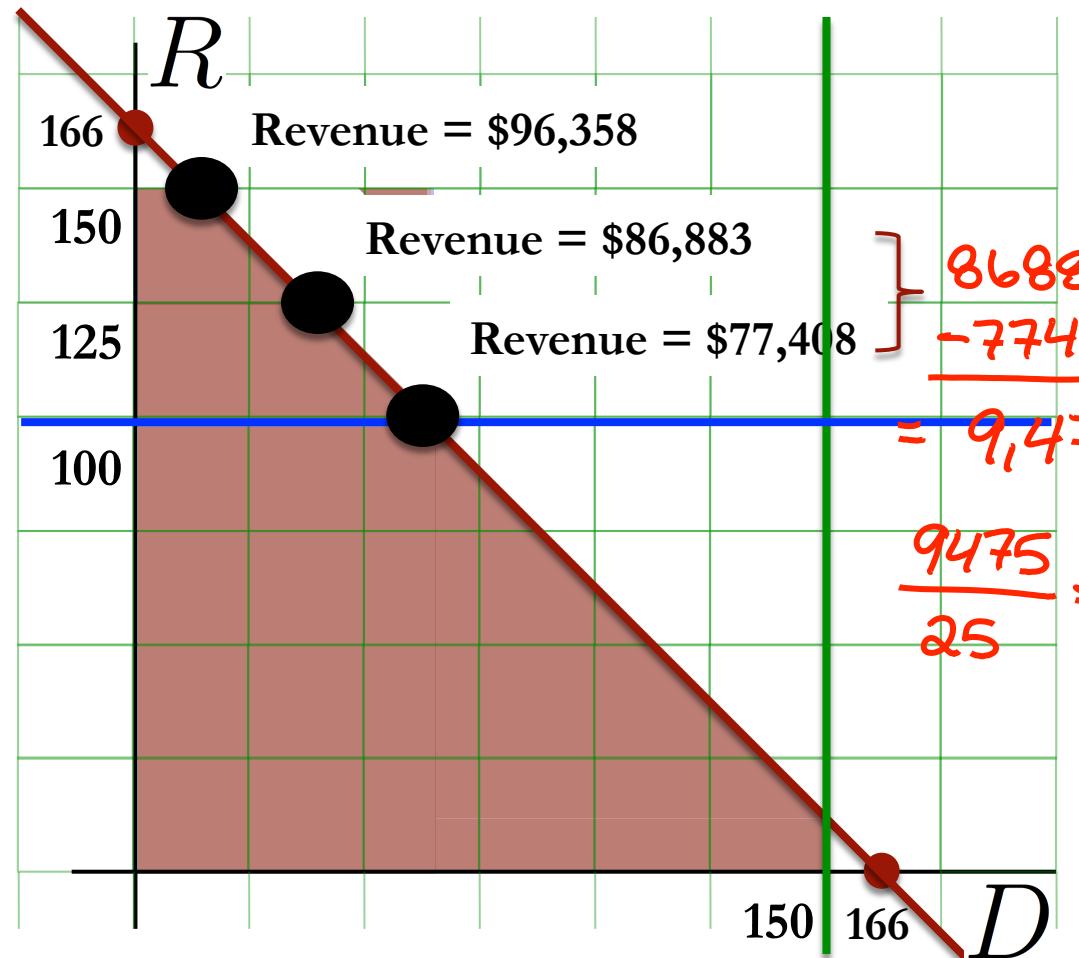
- “Shadow Price”
 - Marginal revenue of increasing discount demand by 1 unit
 - **ZERO** for discount demand greater than 66

Marketing Regular Fares



- AA is considering increasing its budget to market regular fares
- Higher demand for regular class
 - 100
 - 125
 - 150

Marketing Regular Fares



- “Shadow Price”
 - Marginal revenue for unit increase in demand of regular seats
 - \$379 for regular demand between 0 and 166

Marketing Decisions

- Management is trying to figure out whether it would be beneficial to invest in marketing its fares
- AA forecasts that its marketing effort is likely to attract one more unit of demand per **\$200 spent**

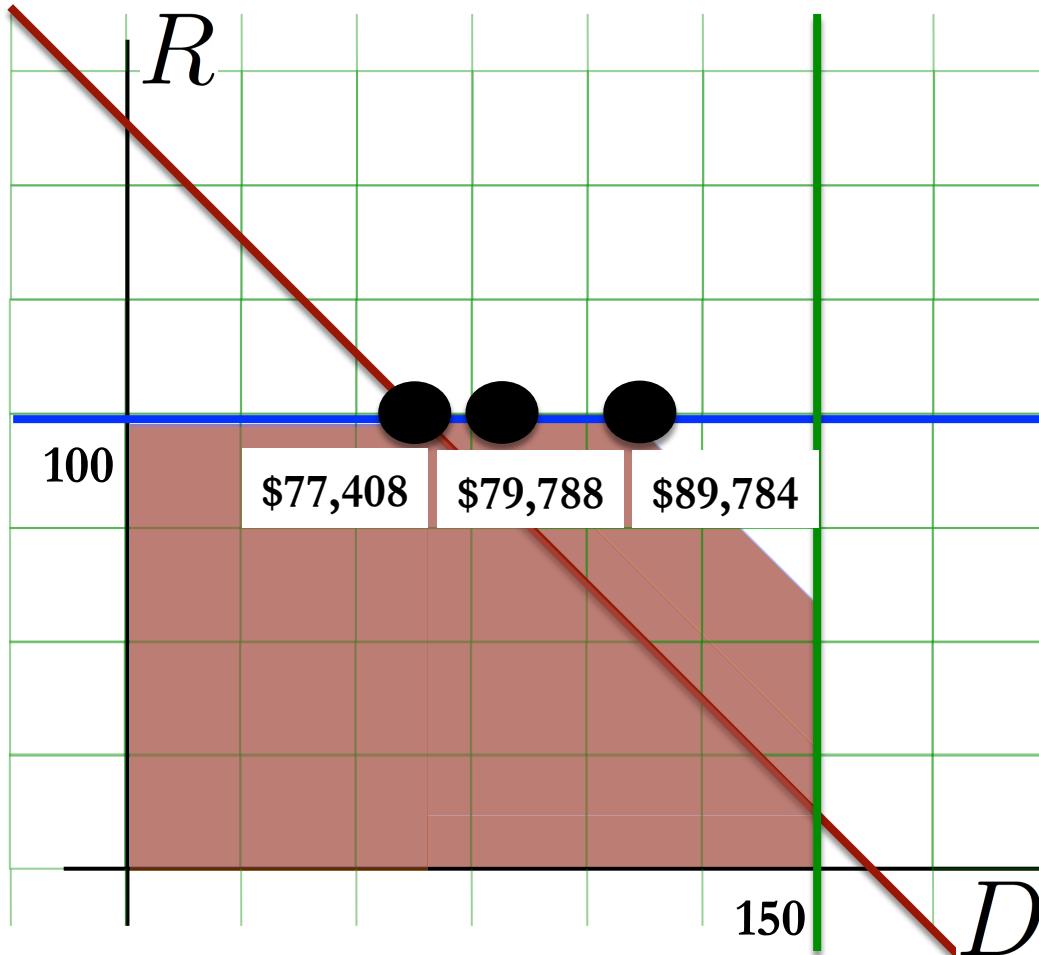
	Marketing Cost/unit	Marginal Revenue
Discount Fare	\$200	0
Regular Fare	\$200	\$379

Capacity Allocation

- Management is trying to figure out whether it would be beneficial to allocate a bigger aircraft for the 6 hour JFK-LAX leg

	Cost/hr	Total Cost	Seats	Revenue
Original Aircraft	\$12,067	\$72,402	166	\$77,408
Boeing 757-200	\$12,765	\$76,590	176	
Boeing 767-300	\$14,557	\$87,342	218	

Aircraft Capacity



- AA is considering increasing its aircraft capacity
 - 166
 - 176
 - 218

Capacity Allocation

- Management is trying to figure out whether it would be beneficial to allocate a bigger aircraft for the 6 hour JFK-LAX leg

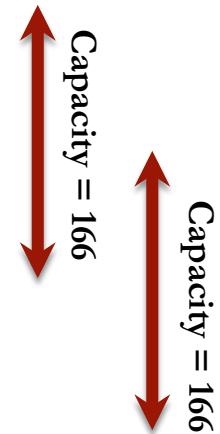
	Total Cost	Revenue	Profit
Original Aircraft	\$72,402	\$77,408	\$5,006
Boeing 757-200	\$76,590	\$79,788	\$3,198
Boeing 767-300	\$87,342	\$89,784	\$2,442

Connecting Flights



Step 1. Decisions

		Price	Demand	Seats to Sell
JFK	Regular	428	80	
	Discount	190	120	
LAX	Regular	642	75	
	Discount	224	100	
DFW	Regular	512	60	
	Discount	190	110	



- Number of regular seats to sell

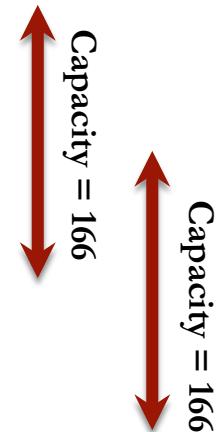
$$R_{\text{JFK-LAX}}, R_{\text{JFK-DFW}}, R_{\text{DFW-LAX}}$$

- Number of discount seats to sell

$$D_{\text{JFK-LAX}}, D_{\text{JFK-DFW}}, D_{\text{DFW-LAX}}$$

Step 2. Objective

		Price	Demand	Seats to Sell
JFK	Regular	428	80	
	Discount	190	120	
LAX	Regular	642	75	
	Discount	224	100	
DFW	Regular	512	60	
	Discount	190	110	

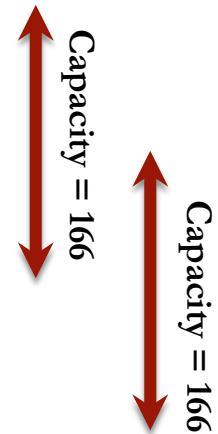


- Maximize total revenue

$$\begin{aligned} & 428R_{\text{JFK-LAX}} + 190D_{\text{JFK-LAX}} \\ & + 642R_{\text{JFK-DFW}} + 224D_{\text{JFK-DFW}} \\ & + 512R_{\text{DFW-LAX}} + 190D_{\text{DFW-LAX}} \end{aligned}$$

Step 3. Constraints

		Price	Demand	Seats to Sell
JFK	Regular	428	80	
LAX	Discount	190	120	
JFK	Regular	642	75	
DFW	Discount	224	100	
DFW	Regular	512	60	
LAX	Discount	190	110	



- AA cannot sell more seats than the aircraft capacity
 - First leg - JFK-DFW

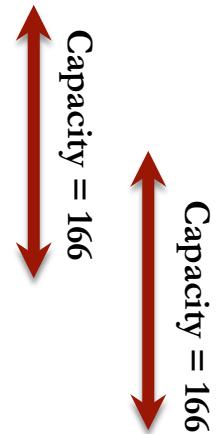
$$R_{\text{JFK-LAX}} + D_{\text{JFK-LAX}} + R_{\text{JFK-DFW}} + D_{\text{JFK-DFW}} \leq 166$$

- Second leg - DFW-LAX

$$R_{\text{JFK-LAX}} + D_{\text{JFK-LAX}} + R_{\text{DFW-LAX}} + D_{\text{DFW-LAX}} \leq 166$$

Step 3. Constraints

		Price	Demand	Seats to Sell
JFK	Regular	428	80	
LAX	Discount	190	120	
JFK	Regular	642	75	
DFW	Discount	224	100	
DFW	Regular	512	60	
LAX	Discount	190	110	



- AA cannot sell more seats than the demand

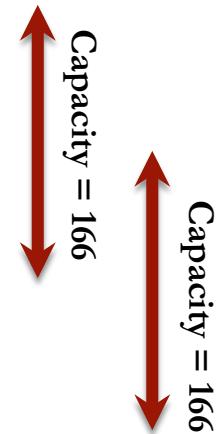
$$R_{JFK-LAX} \leq 80 \quad D_{JFK-LAX} \leq 120$$

$$R_{JFK-DFW} \leq 75 \quad D_{JFK-DFW} \leq 100$$

$$R_{DFW-LAX} \leq 60 \quad D_{DFW-LAX} \leq 110$$

Step 4. Non-Negativity

		Price	Demand	Seats to Sell
JFK	Regular	428	80	
	Discount	190	120	
LAX	Regular	642	75	
	Discount	224	100	
DFW	Regular	512	60	
	Discount	190	110	



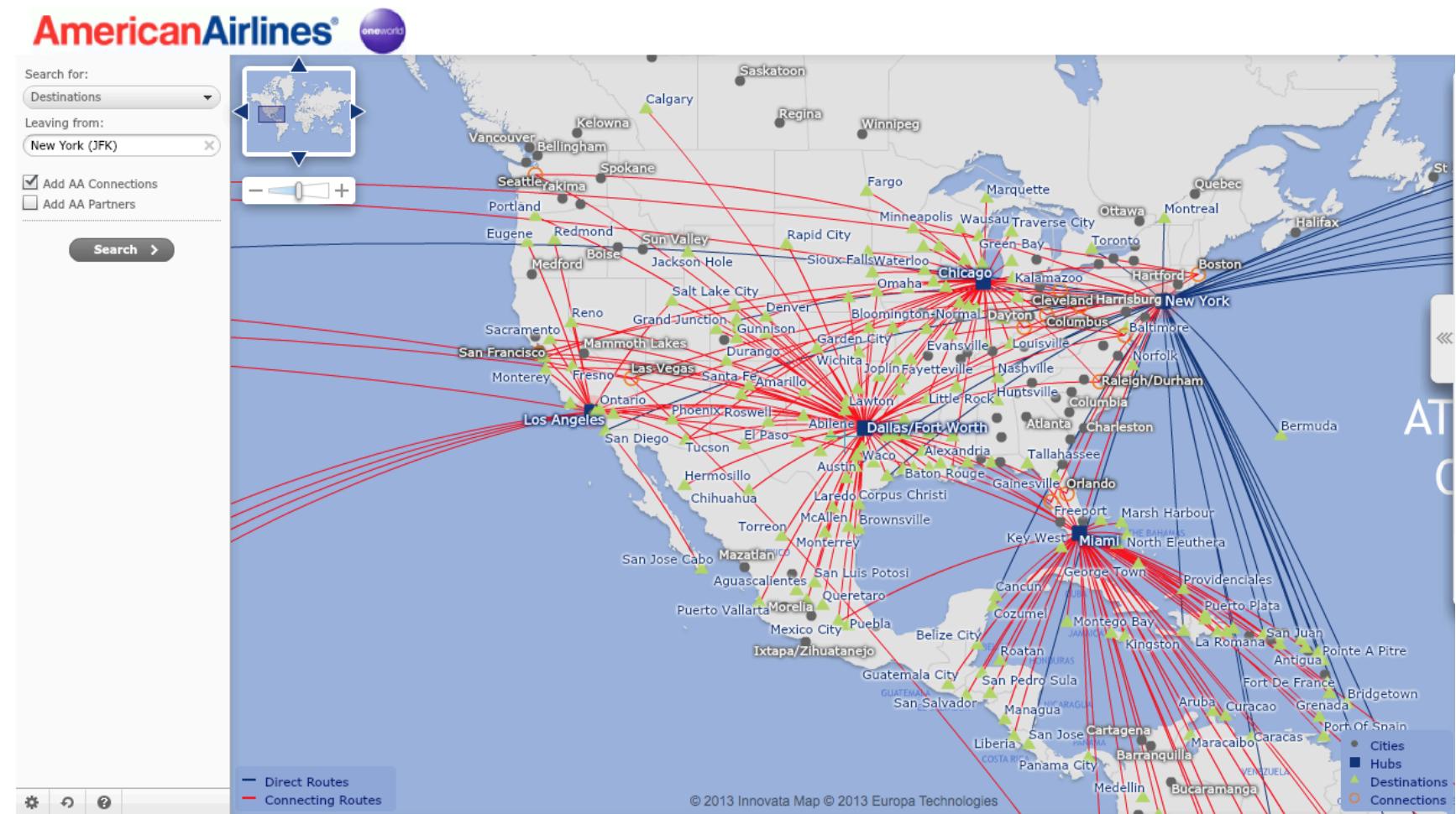
- AA cannot sell a negative number of seats

$$R_{\text{JFK-LAX}} \geq 0 \quad D_{\text{JFK-LAX}} \geq 0$$

$$R_{\text{JFK-DFW}} \geq 0 \quad D_{\text{JFK-DFW}} \geq 0$$

$$R_{\text{DFW-LAX}} \geq 0 \quad D_{\text{DFW-LAX}} \geq 0$$

Complex Network



Multiple Fare Classes

Fare	Domestic Upg.	International Upg.	EQP	EQM	Mileage	Fare	Domestic Upg.	International Upg.	EQP	EQM	Mileage
A	First Class	First Class	1.5	1.0	150%	N	Yes	No	.5	1.0	100%
B	Yes	Yes	1.5	1.0	100%	O	Yes*	No	.5	1.0	100%
C	NA	Business Upgrade	N/A	N/A	N/A	P	First Class Fare	First Class Fare	1.5	1.0	150%
D	NA	Business Fare	1.5	1.0	125%	Q	Yes	No	.5	1.0	100%
E	No	No	N/A	N/A	N/A	R	NA	Business Class Upgrade or waitlist	N/A	N/A	N/A
F	First Class Fare	First Class	1.5	1.0	150%	S	Yes*	No	.5	1.0	100%
G	Government	Government	.5	1.0	100%	T	Coach Award	No	N/A	N/A	N/A
H	Yes*	Waitlist only	1.0	1.0	100%	U	NA	Business Class Award	N/A	N/A	N/A
I	NA	Business Class Fare	1.5	1.0	125%	V	Yes*	No	1.0	1.0	100%
J	NA	Business Class Fare	1.5	1.0	125%	W	Yes*	No	1.0	1.0	100%
K	Yes	No	1.0	1.0	100%	X	First Class Upgrade	Business Class Upgrade	N/A	N/A	N/A
L	Yes	No	1.0	1.0	100%	Y	Yes	Yes	1.5	1.0	100%
M	Yes	No	1.0	1.0	100%	Z	First Class Award	NA	N/A	N/A	N/A

EQP: Elite-Qualifying Points / EQM: Elite-Qualifying Miles

The Competitive Strategy of AA



- PEOPLEExpress could not compete with AA's Ultimate Super Savers fares

“We were a vibrant, profitable company from 1981 to 1985, and then we tipped right over into **losing 50 million a month.**”

“We had been profitable from the day we started until American came at us with Ultimate Super Savers.”

Donald Burr, CEO of PEOPLEExpress (1985)

The Competitive Strategy of AA



- Selling the right seats to the right customers at the right prices

“**Revenue management** is the single most important technical development in transportation management since we entered the era of airline deregulation.”

“We estimate that revenue management has generated **\$1.4 billion in incremental revenue** in the last three years.”

Robert Crandall, former CEO of AA (~1985)

The Edge of Revenue Management



- Sabre Holdings
 - Built revenue management system for AA
 - As of November 2012, ranked 133 among America's largest private companies with \$3.15 billion in sales
 - 400 airlines, 90,000 hotels, 30 car-rental companies
- Today, companies prosper from revenue management
 - Delta airlines increased annual revenue by \$300 million
 - Marriott hotels increased annual revenue by \$100 million



eHarmony

Maximizing the Probability of Love

15.071x – The Analytics Edge

About eHarmony



- Goal: take a scientific approach to love and marriage and offer it to the masses through an online dating website focused on long term relationships
- Successful at matchmaking
 - Nearly 4% of US marriages in 2012 are a result of eHarmony
- Successful business
 - Has generated over \$1 billion in cumulative revenue

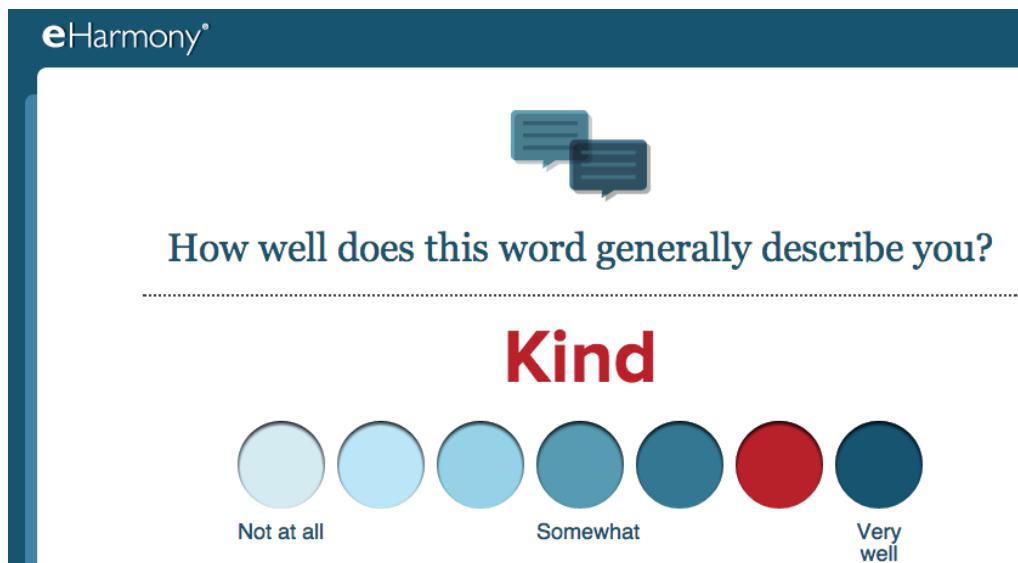
The eHarmony Difference



- Unlike other online dating websites, eHarmony does not have users browse others' profiles
- Instead, eHarmony computes a compatibility score between two people and uses optimization algorithms to determine their users' best matches

eHarmony's Compatibility Score

- Based on 29 different “dimensions of personality” including character, emotions, values, traits, etc.
- Assessed through a 436 question questionnaire
- Matches must meet $>25/29$ compatibility areas



Dr. Neil Clark Warren



- Clinical psychologist who counseled couples and began to see that many marriages ended in divorce because couples were not initially compatible
- Has written many relationship books: “Finding the Love of Your Life”, “The Triumphant Marriage”, “Learning to Live with the Love of Your Life and Loving It”, “Finding Commitment”, and others

Research → Business



- In 1997, Warren began an extensive research project interviewing 5000+ couples across the US, which became the basis of eHarmony's compatibility profile
- www.eHarmony.com went live in 2000
- Interested users may fill out the compatibility quiz, but in order to see matches, members must pay a membership fee to eHarmony

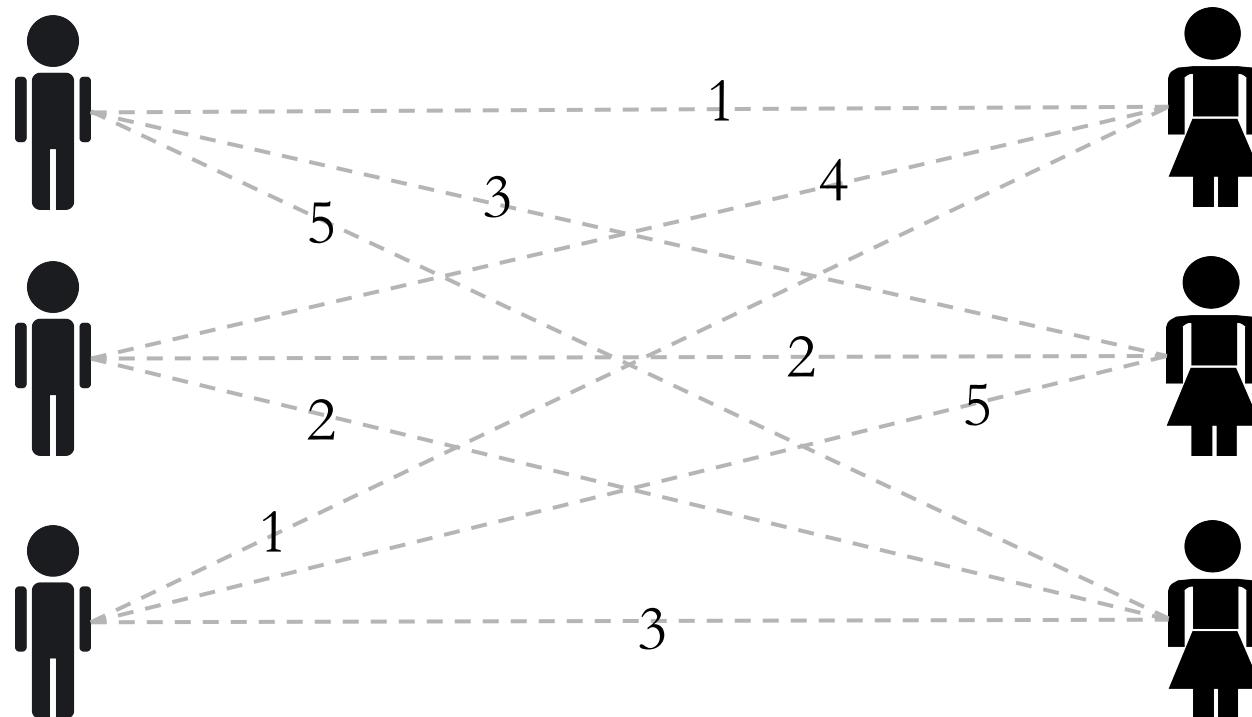
eHarmony Stands Out From the Crowd



- eHarmony was not the first online dating website and faced serious competition
- Key difference from other dating websites: takes a quantitative optimization approach to matchmaking, rather than letting users browse

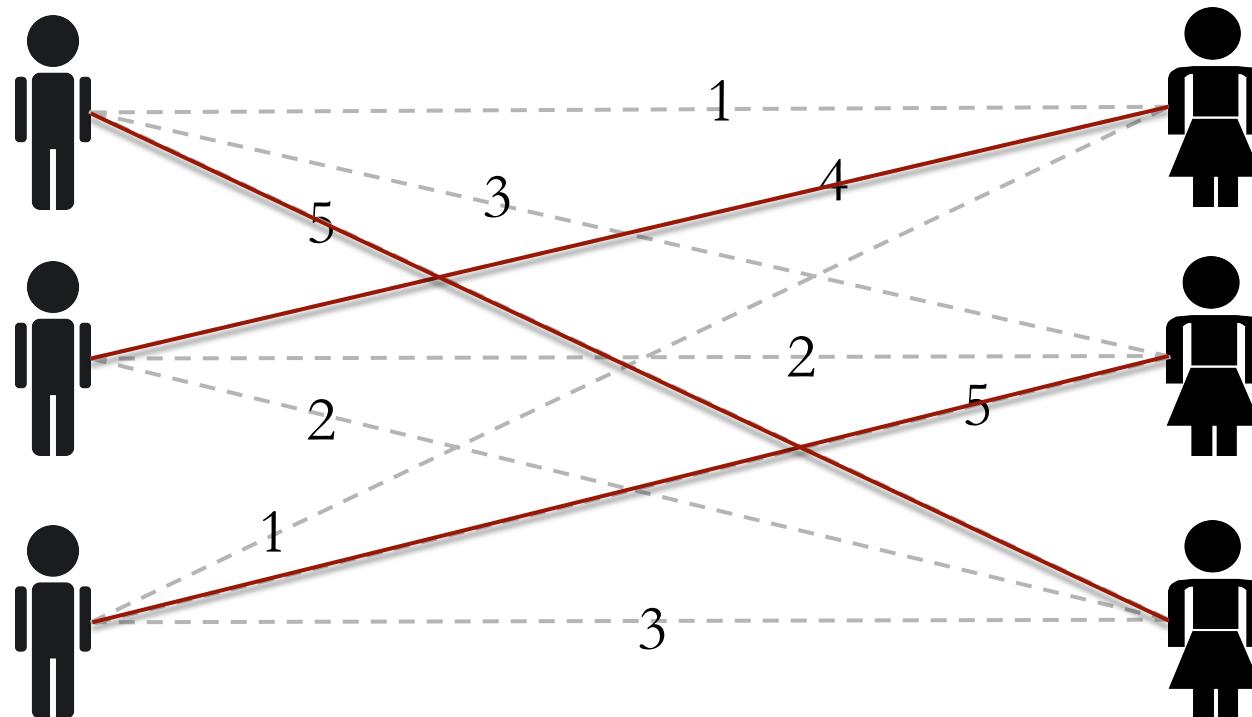
Integer Optimization Example

- Suppose we have three men and three women
- Compatibility scores between 1 and 5 for all pairs



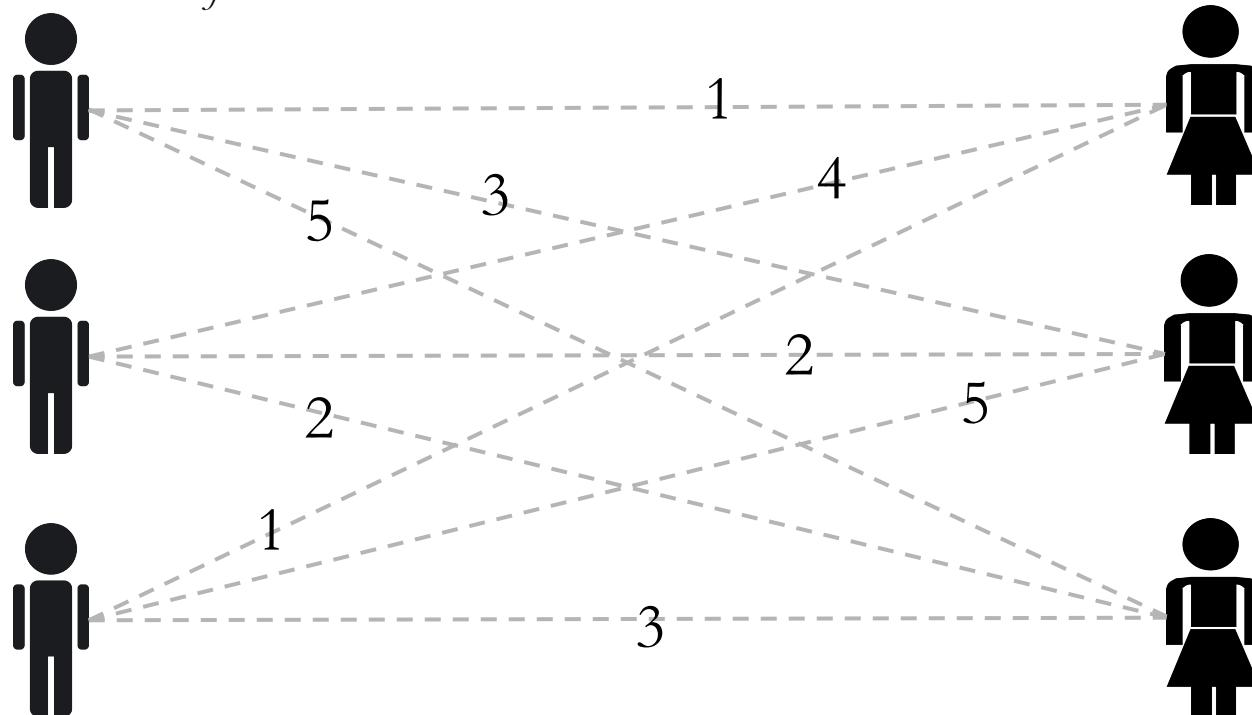
Integer Optimization Example

- How should we match pairs together to maximize compatibility?



Data and Decision Variables

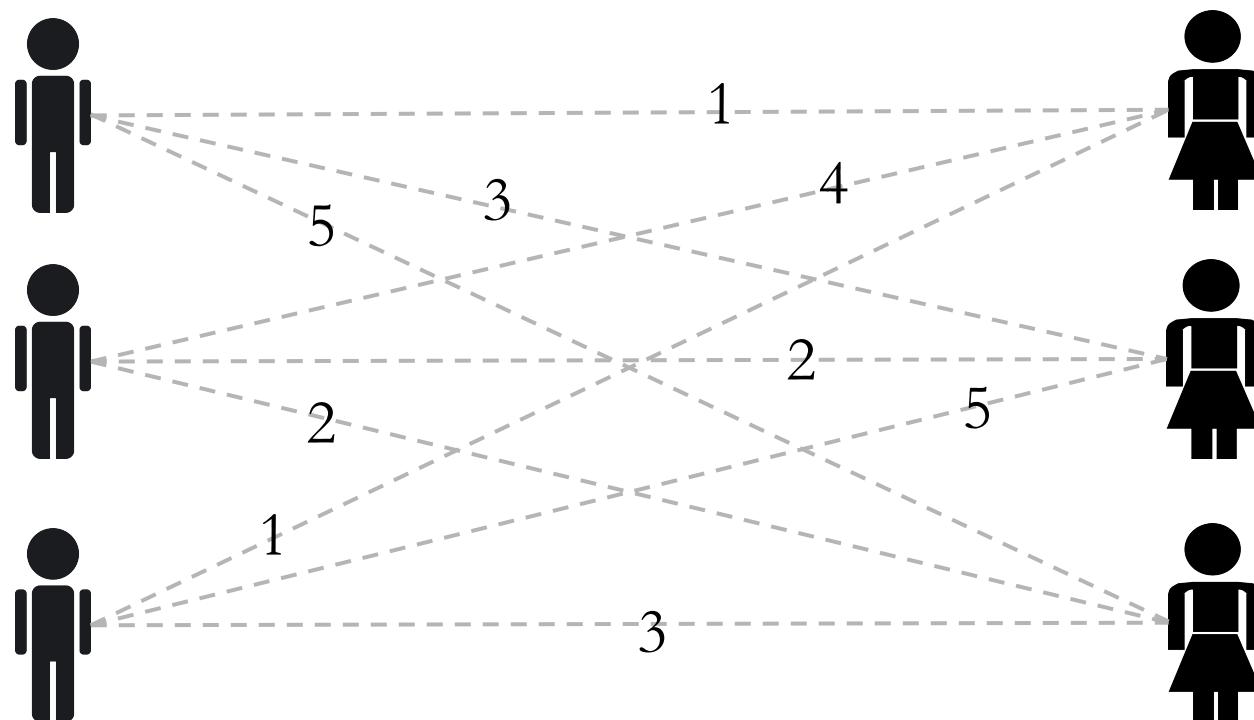
- Decision variables: Let x_{ij} be a binary variable taking value 1 if we match user i and user j together and value 0 otherwise
- Data: Let w_{ij} be the compatibility score between user i and j



Objective Function

- Maximize compatibility between matches:

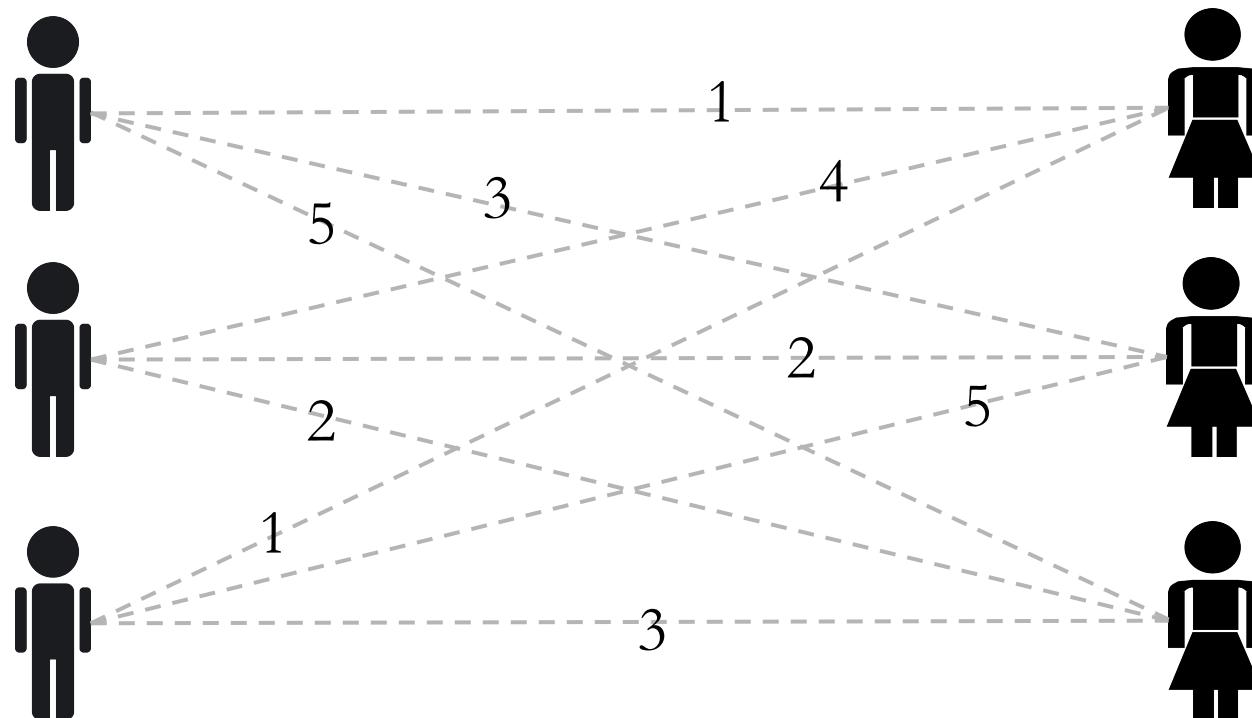
$$\max \quad w_{11}x_{11} + w_{12}x_{12} + w_{13}x_{13} + w_{21}x_{21} + \dots + w_{33}x_{33}$$



Constraints

- Match each man to exactly one woman:

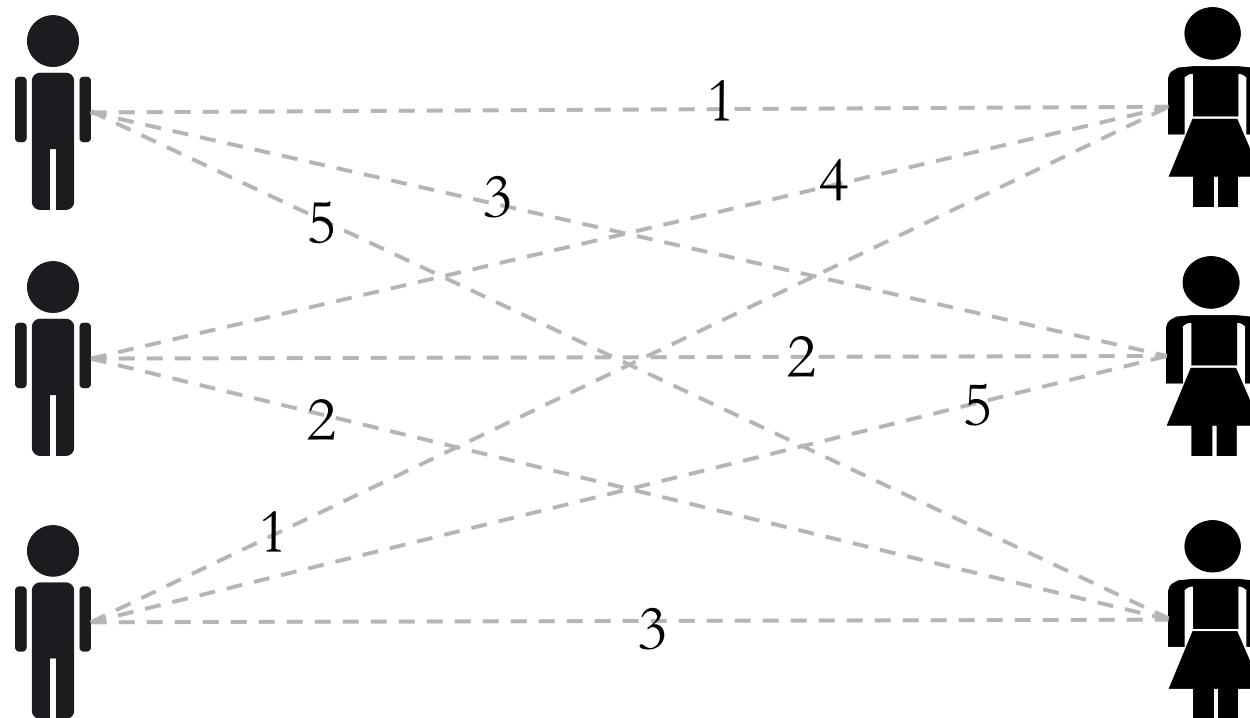
$$x_{11} + x_{12} + x_{13} = 1$$



Constraints

- Similarly, match each woman to exactly one man:

$$x_{11} + x_{21} + x_{31} = 1$$



Full Optimization Problem

$$\max \quad w_{11}x_{11} + w_{12}x_{12} + w_{13}x_{13} + w_{21}x_{21} + \dots + w_{33}x_{33}$$

$$\text{subject to: } x_{11} + x_{12} + x_{13} = 1$$

$$x_{21} + x_{22} + x_{23} = 1$$

$$x_{31} + x_{32} + x_{33} = 1$$

$$x_{11} + x_{21} + x_{31} = 1$$

$$x_{12} + x_{22} + x_{32} = 1$$

$$x_{13} + x_{23} + x_{33} = 1$$

Match every man with
exactly one woman

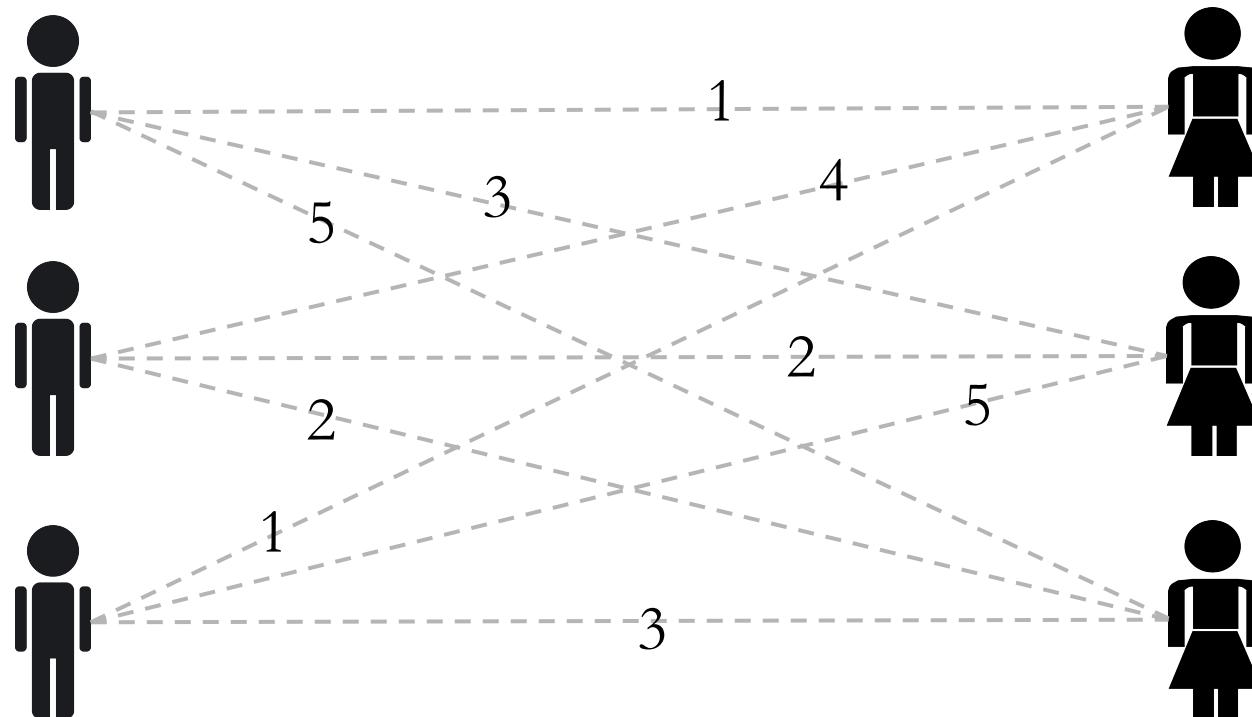
Match every woman
with exactly one man

$x_{11}, x_{21}, x_{31}, x_{12}, x_{22}, x_{32}, x_{13}, x_{23}, x_{33}$ are binary

Extend to Multiple Matches

- Show woman 1 her top two male matches:

$$x_{11} + x_{21} + x_{31} = 2$$



Compatibility Scores



- In the optimization problem, we assumed the compatibility scores were data that we could input directly into the optimization model
- But where do these scores come from?
- “Opposites attract, then they attack”
 - Neil Clark Warren
- eHarmony’s compatibility match score is based on similarity between users’ answers to the questionnaire

Predictive Model

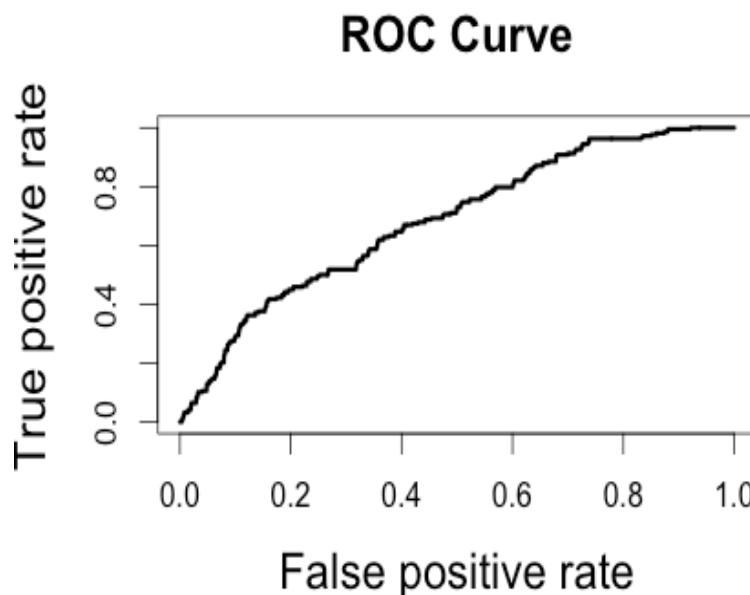
- Public data set from eHarmony containing features for ~275,000 users and binary compatibility results from an interaction suggested by eHarmony
- Feature names and exact values are masked to protect users' privacy
- Try logistic regression on pairs of users' differences to predict compatibility

Reduce the Size of the Problem



- Filtered the data to include only users in the Boston area who had compatibility scores listed in the dataset
- Computed absolute difference in features for these 1475 pairs
- Trained a logistic regression model on these differences

Predicting Compatibility is Hard!



- Model AUC = 0.685

- If we use a low threshold we will predict more false positives but also get more true positives
- Classification matrix for threshold = 0.2:

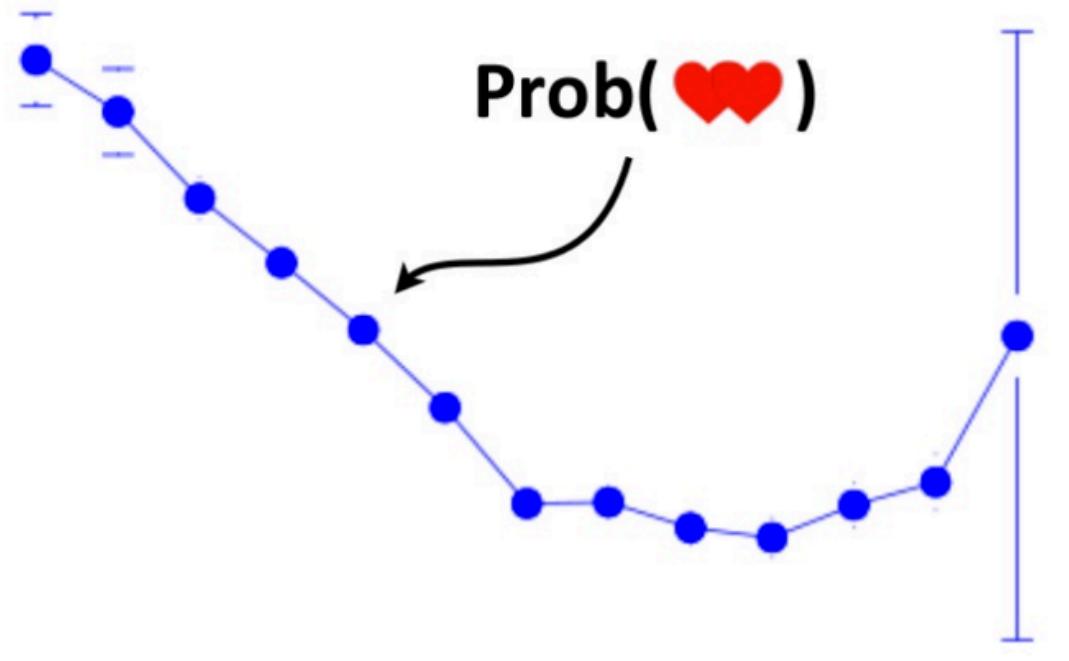
Act \ Pred	0	1
0	1030	227
1	126	92

Other Potential Techniques

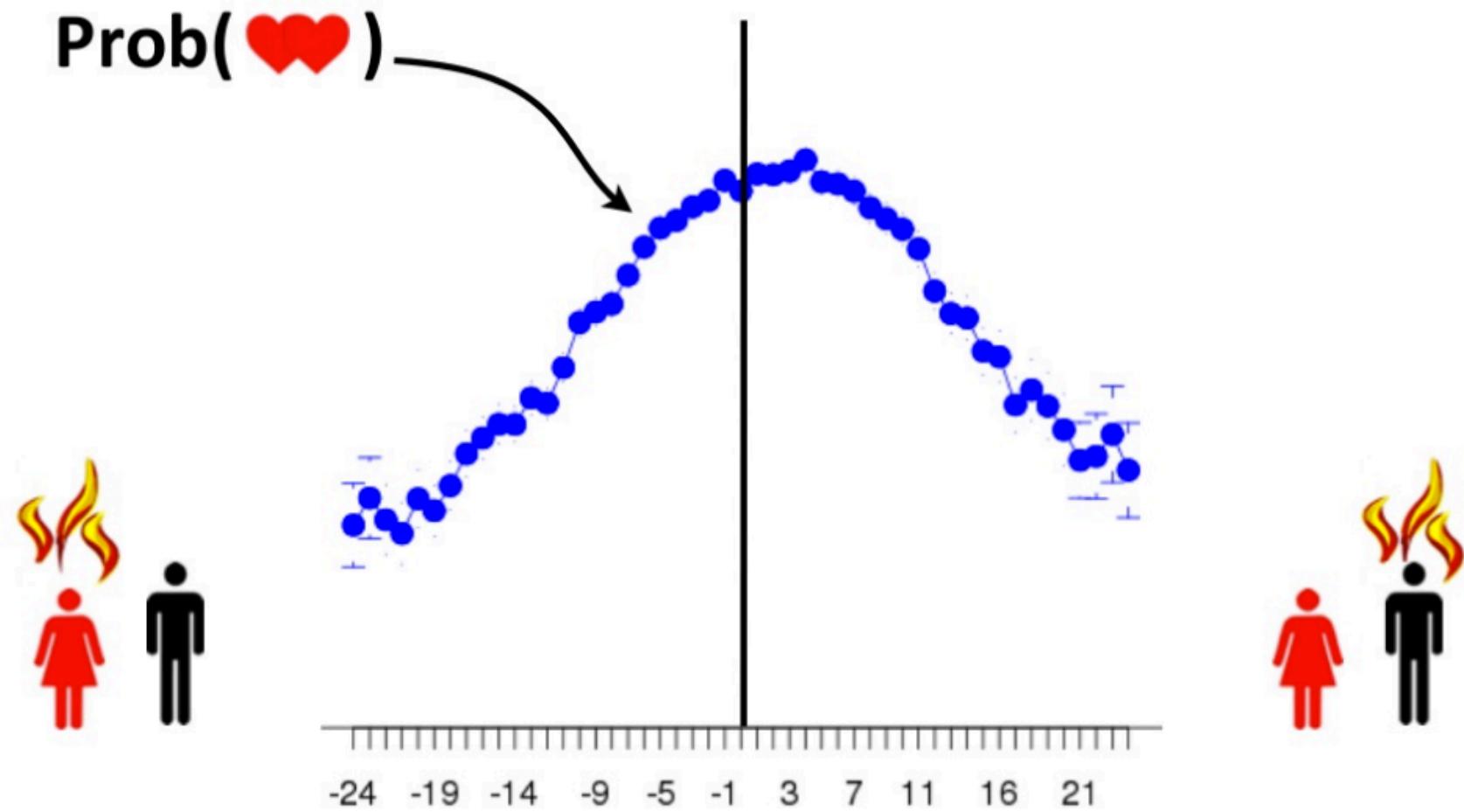


- Trees
 - Especially useful for predicting compatibility if there are nonlinear relationships between variables
- Clustering
 - User segmentation
- Text Analytics
 - Analyze the text of users' profiles
- And much more...

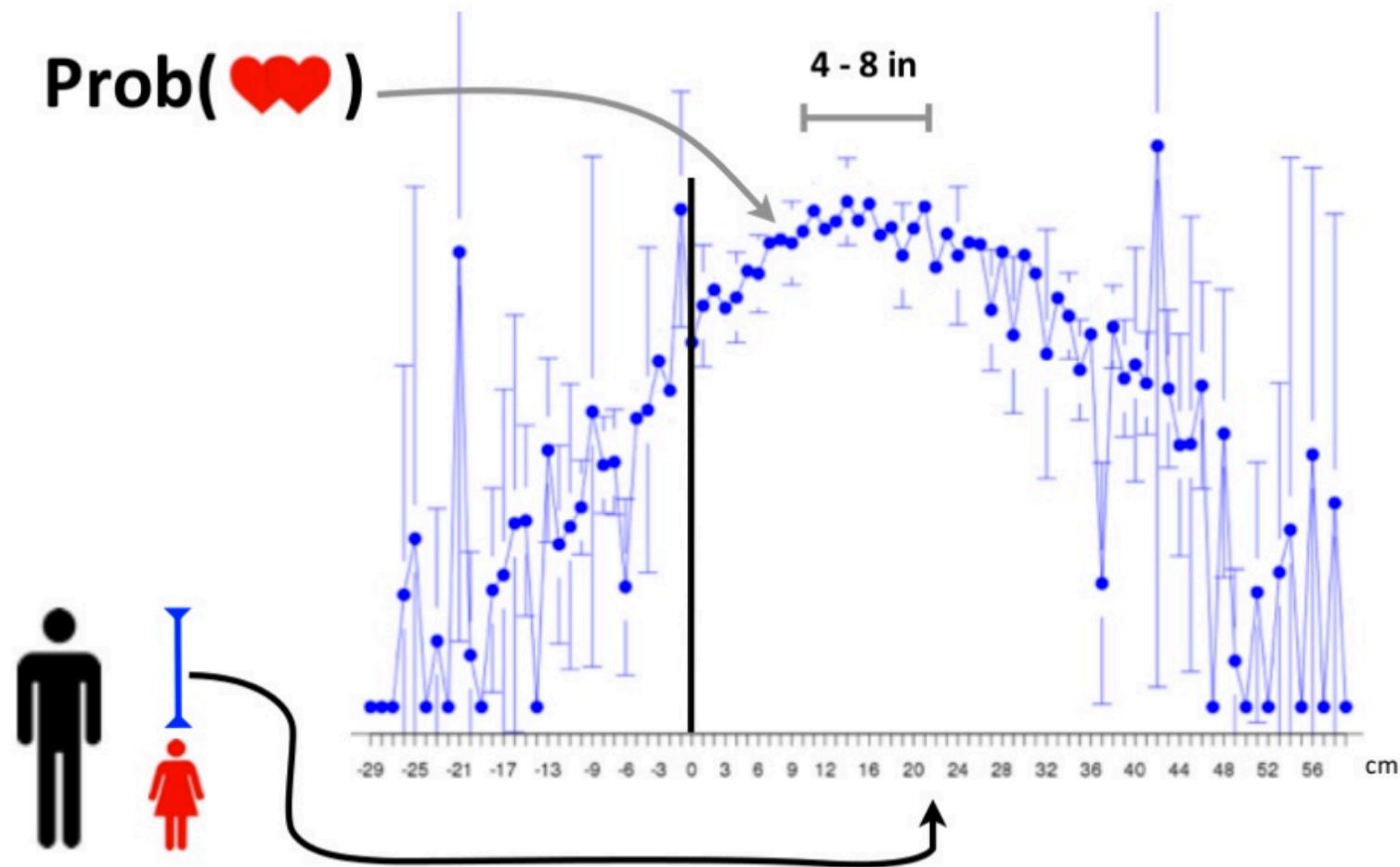
Feature Importance: Distance



Feature Importance: Attractiveness



Feature Importance: Height Difference



How Successful is eHarmony?

- By 2004, eHarmony had made over \$100 million in sales.
- In 2005, 90 eHarmony members married every day
- In 2007, 236 eHarmony members married every day
- In 2009, 542 eHarmony members married every day



eHarmony Maintains its Edge

- 14% of the US online dating market.
- The only competitor with a larger portion is Match.com with 24%.
- Nearly 4% of US marriages in 2012 are a result of eHarmony.
- eHarmony has successfully leveraged the power of analytics to create a successful and thriving business.





OPERATING ROOM SCHEDULING

Making Hospitals Run Smoothly

15.071x – The Analytics Edge

Operating Room Schedules

- Hospitals have a limited number of ORs.
- Operating room managers must determine a weekly schedule assigning ORs to different departments in the hospital.



Difficulties



- Creating an acceptable schedule is a highly political process within the hospital.
- Surgeons are frequently paid on a fee-for-service basis, so changing allocated OR hours directly affects their income.
- The operating room manager's proposed schedule must strike a delicate balance between all the surgical departments in the hospital.

Logistical Issues



- Operating rooms are staffed in 8 hour blocks.
- Each department sets their own target number of allocation hours, which may not be integer.
- Departments may have daily and weekly requirements:
 - Ex) Gynecology needs at least 1 OR per day
 - Ex) Ophthalmology needs at least 2 ORs per week
 - Ex) The oral surgeon is only present on Tuesdays and Thursdays.

Case study: Mount Sinai Hospital

- Has 10 ORs which are staffed Monday – Friday
 - $10 \text{ ORs} \times 5 \text{ days} \times 8 \text{ hours/day} = 400 \text{ hours}$ to assign
 - Must divide these 400 hours between 5 departments:

Department	Weekly Target Allocation Hours
Ophthalmology	39.4
Gynecology	117.4
Oral Surgery	19.9
Otolaryngology	26.3
General Surgery	189.0

Problem Data

- Number of surgery teams from each department available each day:
- Maximum number of ORs required by each department each day:

	M	T	W	R	F
Ophthalmology	2	2	2	2	2
Gynecology	3	3	3	3	3
Oral Surgery	0	1	0	1	0
Otolaryngology	1	1	1	1	1
General Surgery	6	6	6	6	6

	M	T	W	R	F
Ophthalmology	2	2	2	2	2
Gynecology	3	3	3	3	3
Oral Surgery	1	1	1	1	1
Otolaryngology	1	1	1	1	1
General Surgery	6	6	6	6	6

Additional Problem Data

- Weekly requirement on number of ORs each department requires:

	Minimum	Maximum
Ophthalmology	3	6
Gynecology	12	18
Oral Surgery	2	3
Otolaryngology	2	4
General Surgery	18	25

The Traditional Way



- Before the integer optimization method was implemented at Mount Sinai in 1999, the OR manager used graph paper and a large eraser to try to assign the OR blocks.
- Any changes were incorporated by trial and error.
- Draft schedule was circulated to all surgical groups.
- Incorporating feedback from one department usually meant altering another group's schedule, leading to many iterations of this process.

Optimization Problem

- Decisions
 - How many ORs to assign each department on each day.
 - Integer decision variables x_{jk} represent the number of operating rooms department j is allocated on day k .



Objective

- Maximize % of target allocation hours that each department is actually allocated.
- If target allocation hours are t_j for department j, then we want to maximize the sum of $(8 \times x_{jk}) \div t_j$ over all departments and days of the week.

Objective

- Maximize % of target allocation hours that each department is actually allocated.
- If target allocation hours are t_j for department j, then we want to maximize the sum of $(8 \times x_{jk}) \div t_j$ over all departments and days of the week.
 - Ex) If otolaryngology has a target of 37.3 hours per week and we allocate them 4 ORs then their % of target allocation hours = $(8 \times 4) \div 37.3 = 85.8\%$

Constraints

- At most 10 ORs are assigned every day
- The number of ORs allocated to a department on a given day cannot exceed the number of surgery teams that department has available that day
- Meet department daily minimums and maximums
- Meet department weekly minimums and maximums

Ophthalmology	OP
Gynecology	GY
Oral Surgery	OS
Otolaryngology	OT
General Surgery	GS

Constraints

- $x_{OP,M} + x_{GY,M} + x_{OS,M} + x_{OT,M} + x_{GS,M} \leq 10$
- The number of ORs allocated to a department on a given day cannot exceed the number of surgery teams that department has available that day
- Meet department daily minimums and maximums
- Meet department weekly minimums and maximums

Ophthalmology	OP
Gynecology	GY
Oral Surgery	OS
Otolaryngology	OT
General Surgery	GS

Constraints

- $x_{OP, M} + x_{GY, M} + x_{OS, M} + x_{OT, M} + x_{GS, M} \leq 10$
- $0 \leq x_{GY, F} \leq 3$
- $0 \leq x_{OS, W} \leq 0$
- Meet department daily minimums and maximums
- Meet department weekly minimums and maximums

Ophthalmology	OP
Gynecology	GY
Oral Surgery	OS
Otolaryngology	OT
General Surgery	GS

Constraints

- $x_{OP, M} + x_{GY, M} + x_{OS, M} + x_{OT, M} + x_{GS, M} \leq 10$
- $0 \leq x_{OS, W} \leq 3$
- $0 \leq x_{GY, F} \leq 0$
- $0 \leq x_{GS, T} \leq 6$
- Meet department weekly minimums and maximums

Ophthalmology	OP
Gynecology	GY
Oral Surgery	OS
Otolaryngology	OT
General Surgery	GS

Constraints

- $x_{OP,M} + x_{GY,M} + x_{OS,M} + x_{OT,M} + x_{GS,M} \leq 10$
- $0 \leq x_{OS,W} \leq 3$
- $0 \leq x_{GY,F} \leq 0$
- $0 \leq x_{GS,T} \leq 8$
- $3 \leq x_{OP,M} + x_{OP,T} + x_{OP,W} + x_{OP,R} + x_{OP,F} \leq 6$

Ophthalmology	OP
Gynecology	GY
Oral Surgery	OS
Otolaryngology	OT
General Surgery	GS



SPORTS SCHEDULING

An Introduction to Integer Optimization

15.071x – The Analytics Edge

The Impact of Sports Schedules



- Sports is a \$300 billion dollar industry
 - Twice as big as the automobile industry
 - Seven times as big as the movie industry
- TV networks are key to revenue for sports teams
 - \$513 million per year for English Premier League soccer
 - \$766 million per year for NBA
 - \$3 billion per year for NFL
- They pay to have a good schedule of sports games

Sports Schedules



- Good schedules are important for other reasons too
 - Extensive traveling causes player fatigue
 - Ticket sales are better on the weekends
 - Better to play division teams near the end of season
- All competitive sports require schedules
 - **Which pairs** of teams play each other and **when?**

The Traditional Way

- Until recently, schedules mostly constructed by hand
 - Time consuming: with 10 teams, there are over 1 trillion possible schedules (every team plays every other team)
 - Many constraints: television networks, teams, cities, . . .
- For Major League Baseball, a husband and wife team constructed the schedules for 24 years (1981-2005)
 - Used a giant wall of magnets to schedule 2430 games
- Very difficult to add new constraints

Some Interesting Constraints



- In 2008, the owners and TV networks were not the only ones who cared about the schedule
- President Barack Obama and Senator John McCain complained about the schedule
 - National conventions conflicted with game scheduling
- Then, the Pope complained about the schedule!
 - The Pope visited New York on April 20, 2008
 - Mass in Yankee stadium (the traditional location)
- Each of these constraints required a new schedule

An Analytics Approach



- In 1996, “The Sports Scheduling Group” was started
 - Doug Bureman, George Nemhauser, Michael Trick, and Kelly Easton
- They generate schedules using a computer
 - Have been scheduling college sports since 1999
 - Major League Baseball since 2005
- They use optimization
 - Can easily adapt when new constraints are added



SPORTS SCHEDULING

An Introduction to Integer Optimization

15.071x – The Analytics Edge

The Impact of Sports Schedules



- Sports is a \$300 billion dollar industry
 - Twice as big as the automobile industry
 - Seven times as big as the movie industry
- TV networks are key to revenue for sports teams
 - \$513 million per year for English Premier League soccer
 - \$766 million per year for NBA
 - \$3 billion per year for NFL
- They pay to have a good schedule of sports games

Sports Schedules



- Good schedules are important for other reasons too
 - Extensive traveling causes player fatigue
 - Ticket sales are better on the weekends
 - Better to play division teams near the end of season
- All competitive sports require schedules
 - **Which pairs** of teams play each other and **when?**

The Traditional Way

- Until recently, schedules mostly constructed by hand
 - Time consuming: with 10 teams, there are over 1 trillion possible schedules (every team plays every other team)
 - Many constraints: television networks, teams, cities, . . .
- For Major League Baseball, a husband and wife team constructed the schedules for 24 years (1981-2005)
 - Used a giant wall of magnets to schedule 2430 games
- Very difficult to add new constraints

Some Interesting Constraints



- In 2008, the owners and TV networks were not the only ones who cared about the schedule
- President Barack Obama and Senator John McCain complained about the schedule
 - National conventions conflicted with game scheduling
- Then, the Pope complained about the schedule!
 - The Pope visited New York on April 20, 2008
 - Mass in Yankee stadium (the traditional location)
- Each of these constraints required a new schedule

An Analytics Approach



- In 1996, “The Sports Scheduling Group” was started
 - Doug Bureman, George Nemhauser, Michael Trick, and Kelly Easton
- They generate schedules using a computer
 - Have been scheduling college sports since 1999
 - Major League Baseball since 2005
- They use optimization
 - Can easily adapt when new constraints are added

Scheduling a Tournament

- Four teams
 - Atlanta (A) , Boston (B) , Chicago (C) , and Detroit (D)
- Two divisions
 - Atlanta and Boston
 - Chicago and Detroit
- During four weeks
 - Each team plays the other team in its division twice
 - Each team plays teams in other divisions once
- The team with the most wins from each division will play in the championship
- Teams prefer to play divisional games later

An Optimization Approach



- Objective
 - Maximize team preferences (divisional games later)
- Decisions
 - Which teams should play each other each week
- Constraints
 - Play other team in division twice
 - Play teams in other divisions once
 - Play exactly one team each week

Decision Variables

- We need to decide which teams will play each other each week
 - Define variables x_{ijk}
 - If team i plays team j in week k, $x_{ijk} = 1$
 - Otherwise, $x_{ijk} = 0$
- This is called a *binary decision variable*
 - Only takes values 0 or 1

$$x_{AC2} = 1$$

$$x_{AC1} = 0$$

$$x_{AC3} = 0$$

$$x_{AC4} = 0$$

Integer Optimization



- Decision variables can only take integer values
- Binary variables can be either 0 or 1
 - Where to build a new warehouse
 - Whether or not to invest in a stock
 - Assigning nurses to shifts
- Integer variables can be 0, 1, 2, 3, 4, 5, . . .
 - The number of new machines to purchase
 - The number of workers to assign for a shift
 - The number of items to stock

The Formulation

- Objective
 - Maximize team preferences (divisional games later)
- Decisions
 - Which teams should play each other each week
- Constraints
 - Play other team in division twice
 - Play teams in other divisions once
 - Play exactly one team each week

The Formulation

- Objective
 - Maximize team preferences (divisional games later)
- Decisions
 - Binary variables x_{ijk}
- Constraints
 - Play other team in division twice
 - Play teams in other divisions once
 - Play exactly one team each week

The Formulation

- Objective
 - Maximize team preferences (divisional games later)
- Decisions
 - Binary variables x_{ijk}
- Constraints
 - $x_{AB1} + x_{AB2} + x_{AB3} + x_{AB4} = 2$
 - Play teams in other divisions once
 - Play exactly one team each week

Similar constraint for
teams C and D

The Formulation

- Objective
 - Maximize team preferences (divisional games later)
- Decisions
 - Binary variables x_{ijk}
- Constraints
 - $x_{AB1} + x_{AB2} + x_{AB3} + x_{AB4} = 2$
 - $x_{AC1} + x_{AC2} + x_{AC3} + x_{AC4} = 1$
 - Play exactly one team each week

Similar constraint for teams C and D

Similar constraints for teams A and D, B and C, and B and D

The Formulation

- Objective
 - Maximize team preferences (divisional games later)
- Decisions
 - Binary variables x_{ijk}
- Constraints
 - $x_{AB1} + x_{AB2} + x_{AB3} + x_{AB4} = 2$
 - $x_{AC1} + x_{AC2} + x_{AC3} + x_{AC4} = 1$
 - $x_{AB1} + x_{AC1} + x_{AD1} = 1$

Similar constraint for teams C and D

Similar constraints for teams A and D, B and C, and B and D

Similar constraints for every team and week

The Formulation

- Objective
 - Maximize $x_{AB1} + 2x_{AB2} + 4x_{AB3} + 8x_{AB4} + x_{CD1} + 2x_{CD2} + 4x_{CD3} + 8x_{CD4}$
- Decisions
 - Binary variables x_{ijk}
- Constraints
 - $x_{AB1} + x_{AB2} + x_{AB3} + x_{AB4} = 2$
 - $x_{AC1} + x_{AC2} + x_{AC3} + x_{AC4} = 1$
 - $x_{AB1} + x_{AC1} + x_{AD1} = 1$

Similar constraint for teams C and D

Similar constraints for teams A and D, B and C, and B and D

Similar constraints for every team and week

Adding Logical Constraints

- Binary variables allow us to model logical constraints
- A and B can't play in weeks 3 and 4

$$x_{AB3} + x_{AB4} \leq 1$$

- If A and B play in week 4, they must also play in week 2

$$x_{AB2} \geq x_{AB4}$$

- C and D must play in week 1 or week 2 (or both)

$$x_{CD1} + x_{CD2} \geq 1$$

Solving Integer Optimization Problems



- We were able to solve our sports scheduling problem with 4 teams (24 variables, 22 basic constraints)
- The problem size increases rapidly
 - With 10 teams, 585 variables and 175 basic constraints
- For Major League Baseball
 - 100,000 variables
 - 200,000 constraints
 - This would be impossible in LibreOffice
- So how are integer models solved in practice?

Solving Integer Optimization Problems



- Reformulate the problem
 - The sports scheduling problem is solved by changing the formulation
 - Variables are sequences of games
 - Split into three problems that can each be solved separately
- Heuristics
 - Find good, but not necessarily optimal, decisions

Solving Integer Optimization Problems



- General purpose solvers
 - CPLEX, Gurobi, GLPK, COIN-OR
- In the past 20 years, the speed of integer optimization solvers has increased by a factor of 250,000
 - Doesn't include increasing speed of computers
- **A problem that can be solved in 1 second today took 7 years to solve 20 years ago!**

Solving the Sports Scheduling Problem



- When the Sports Scheduling Group started, integer optimization software was not useful
- Now, they can use powerful solvers to generate schedules
- Takes months to make the MLB schedule
 - Enormous list of constraints
 - Need to define priorities on constraints
 - Takes several iterations to get a good schedule

The Analytics Edge



- Optimization allows for the addition of new constraints or structure changes
 - Can easily generate a new schedule based on an updated requirement or request
- Now, all professional sports and most college sports schedules are constructed using optimization

Question ID	Question Text	Possible Answers
96024	Are you good at math?	Yes, No
98059	Do/did you have any siblings?	Yes, Only-child
98078	Do you have a "go-to" creative outlet?	Yes, No
98197	Do you pray or meditate on a regular basis?	Yes, No
98578	Do you exercise 3 or more times per week?	Yes, No
98869	Does life have a purpose?	Yes, No
99480	Did your parents spank you as a form of discipline/punishment?	Yes, No
99581	Are you left-handed?	Yes, No
99716	Do you live alone?	Yes, No
99982	Do you keep check-lists of tasks you need to accomplish?	Check!, Nope
100010	Do you watch some amount of TV most days?	Yes, No
100562	Do you think your life will be better five years from now than it is today?	Yes, No
100680	Have you cried in the past 60 days?	Yes, No
100689	Do you feel like you are currently overweight?	Yes, No
101162	Are you generally more of an optimist or a pessimist?	Optimist, Pessimist
101163	Which parent "wore the pants" in your household?	Mom, Dad
101596	As a kid, did you ever build (or help build) a tree-house?	Yes, No
102089	Do you rent or own your primary residence?	Rent, Own
102289	Does your life feel adventurous?	Yes, No
102674	Do you have any credit card debt that is more than one month old?	Yes, No
102687	Do you eat breakfast every day?	Yes, No
102906	Are you currently carrying a grudge against anyone in your personal life?	Yes, No
103293	Do you have more than one pet?	Yes, No
104996	Do you brush your teeth two or more times every day?	Yes, No
105655	Were you awakened by an alarm clock this morning?	Yes, No
105840	Do you ever treat yourself to "retail therapy"?	Yes, No
106042	Are you taking any prescription medications?	Yes, No
106272	Do you own any power tools? (power saws, drills, etc.)	Yes, No
106388	Do you work 50+ hours per week?	Yes, No
106389	Are you a good/effective liar?	Yes, No
106993	Do you like your given first name?	Yes, No
106997	Do you generally like people, or do most of them tend to get on your nerves pretty easily?	Yay people!, Grrr people
107491	Do you punctuate text messages?	Yes, No
107869	Do you feel like you're "normal"?	Yes, No
108342	Do you spend more time with friends online or in-person?	Online, In-person
108343	Do you feel like you have too much personal financial debt?	Yes, No
108617	Do you live in a single-parent household?	Yes, No
108754	Do both of your parents have college degrees?	Yes, No
108855	Do you enjoy getting together with your extended family?	Yes!, Umm...
108856	Lots of people are around! Are you more likely to be right in the middle of things, or looking for your own quieter space?	Socialize, Space
108950	Are you generally a cautious person, or are you comfortable taking risks?	Cautious, Risk-friendly
109244	Are you a feminist?	Yes, No
109367	Have you ever been poor (however you personally defined it at the time)?	Yes, No
110740	Mac or PC?	Mac, PC
111220	Is your alarm clock intentionally set to be a few minutes fast?	Yes, No
111580	As a teenager, do/did you have parents who were generally more supportive or demanding?	Supportive, Demanding
111848	Did you ever get a straight-A report card in high school or college?	Yes, No
112270	Are you better looking than your best friend?	Yes, No
112478	Do you have any phobias?	Yes, No
112512	Are you naturally skeptical?	Yes, No
113181	Do you meditate or pray on a regular basis?	Yes, No
113583	While driving: music or talk/news radio?	Tunes, Talk
113584	During your average day, do you spend more time interacting with people (face-to-face) or technology?	People, Technology
113992	Do you gamble?	Yes, No
114152	Do you support a particular charitable cause with a lot of your time and/or money?	Yes, No
114386	Are you more likely to over-share or under-share?	TMI, Mysterious
114517	Do you turn a TV on in the morning while getting ready for your day?	Yes, No
114748	Do you drink the unfiltered tap water in your home?	Yes, No
114961	Can money buy happiness?	Yes, No
115195	Do you live within 20 miles of a major metropolitan area?	Yes, No
115390	Has your personality changed much from what you were like as a child?	Yes, No
115602	Were you an obedient child?	Yes, No
115610	Does the "power of positive thinking" actually work?	Yes, No
115611	Do you personally own a gun?	Yes, No
115777	Do you find it easier to start and maintain a new good habit, or to permanently kick a bad habit?	Start, End

115899	Would you say most of the hardship in your life has been the result of circumstances beyond your own control, or has it been mostly the result of your own decisions and actions?	Circumstances,Me
116197	Are you a morning person or a night person?	A.M.,P.M.
116441	Do you have a car payment?	Yes,No
116448	If you had to stop telling *any* lies for 6 months (even the smallest "little-white-lie" would immediately make you violently ill), would it change your life in any noticeable way?	Yes,No
116601	Have you ever traveled out of the U.S.?	Yes,No
116797	Do you take a daily multi-vitamin?	Yes,No
116881	Would you rather be happy or right?	Happy,Right
116953	Do you like rules?	Yes,No
117186	Do you have a quick temper?	Hot headed,Cool headed
117193	Do you work (or attend school) on a pretty standard "9-to-5ish" daytime schedule, or do you have to work unusual hours?	Standard hours,Odd hours
118117	Have you lived in the same state your whole life?	Yes,No
118232	Are you more of an idealist or a pragmatist?	Idealist,Pragmatist
118233	Have you ever had your life genuinely threatened by intentional violence (or the threat of it)?	Yes,No
118237	Do you feel like you are "in over-your-head" in any aspect of your life right now?	Yes,No
118892	Do you wear glasses or contact lenses?	Yes,No
119334	Did you accomplish anything exciting or inspiring in 2013? (comments from the 2012 poll are linked for inspiration)	Yes,No
119650	Which do you really enjoy more: giving or receiving?	Giving,Receiving
119851	Are you in the middle of reading a good book right now?	Yes,No
120012	Does the weather have a large effect on your mood?	Yes,No
120014	Are you more successful than most of your high-school friends?	Yes,No
120194	Your generally preferred approach to starting a new task: read up on everything you can before trying it out, or dive in with almost no knowledge and learn as you go?	Study first,Try first
120379	Do you have (or plan to pursue) a Masters or Doctoral degree?	Yes,No
120472	Science or Art?	Science,Art
120650	Were your parents married when you were born?	Yes,No
120978	As a kid, did you watch Sesame Street on a regular basis?	Yes,No
121011	Changing or losing a job, getting married or divorced, the death of a close relative, moving, a major health issue, bankruptcy...all are life events that can create high stress for people. Have you experienced any of these in 2013?	Yes,No
121699	2013: did you drink alcohol?	Yes,No
121700	2013: did you start a new romantic relationship?	Yes,No
122120	Your significant other takes an extra long look at a very attractive person (of your gender) walking past both of you. Are you upset?	Yes,No
122769	Do you collect anything (as a hobby)?	Yes,No
122770	Do you have more than \$20 cash in your wallet or purse right now?	Yes,No
122771	Do/did you get most of your K-12 education in public school, or private school?	Public,Private
123464	Do you currently have a job that pays minimum wage?	Yes,No
123621	Are you currently employed in a full-time job?	Yes,No
124122	Did your parents fight in front of you?	Yes,No
124742	Do you have to personally interact with anyone that you really dislike on a daily basis?	Yes,No

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[Syllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[Schedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 1 (1/1 point)

eHarmony has a distinguishing feature, that makes it stand out from other online dating websites? What is this?

- eHarmony was the first online dating site.
- eHarmony asks users for personal information.
- eHarmony charges a fee to users.
- eHarmony only suggests matches to users, instead of allowing users to browse. ✓

EXPLANATION

eHarmony was not the first online dating site, and while they ask users for personal information and charge a fee, other online dating sites do that as well. What distinguishes eHarmony is that they only suggest matches to users, instead of allowing users to browse other users' profiles.

[Final Check](#)[Save](#)[Hide Answer](#)

You have used 1 of 2 submissions



[About \(<https://www.edx.org/about-us>\)](#) [Jobs \(<https://www.edx.org/jobs>\)](#)
[Press \(<https://www.edx.org/press>\)](#) [FAQ \(<https://www.edx.org/student-faq>\)](#)
[Contact \(<https://www.edx.org/contact>\)](#)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community/>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)
© 2014 edX, some rights reserved.

[Terms of Service and Honor Code - Privacy Policy](#) (<https://www.edx.org/edx-privacy-policy>)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[yllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[chedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 1 (1/1 point)

How much data do you think Andrew Martin should use to build his model?

- Information from all cases with the same set of justices as those he is trying to predict. Data from cases where the justices were different might just add noise to our problem. ✓
- Only information from the most recent year. Since the justices change every year, only this information would be useful.

EXPLANATION

Andrew Martin should use all data from the cases with the same set of justices. The justices do not change every year, and typically you want to use as much data as you have available.

[Hide Answer](#)*You have used 1 of 1 submissions*

[About \(<https://www.edx.org/about-us>\)](#) [Jobs \(<https://www.edx.org/jobs>\)](#)
[Press \(<https://www.edx.org/press>\)](#) [FAQ \(<https://www.edx.org/student-faq>\)](#)
[Contact \(<https://www.edx.org/contact>\)](#)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community/>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/1082353830440950827>)



(<http://youtube.com/user/edxonline>)

© 2014 edX, some rights reserved.

[Terms of Service and Honor Code - Privacy Policy \(<https://www.edx.org/edx-privacy-policy>\)](#)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[Syllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[Schedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 1 (1 point possible)

In what ways do you think an analytics approach to predicting healthcare cost will improve upon the previous approach of human judgment?

- It will allow D2Hawkeye to analyze millions of patients. ✓
- It will allow D2Hawkeye to make predictions faster than doctors can. ✓
- It will allow D2Hawkeye use all available data (millions of cases) to make decisions. ✓

EXPLANATION

All of the above are correct. There are many advantages to having an analytics approach to predict cost. However, it is important that the models are interpretable, so trees are a great model to use in this situation.

[Hide Answer](#)*You have used 2 of 2 submissions*

[About \(<https://www.edx.org/about-us>\)](#) [Jobs \(<https://www.edx.org/jobs>\)](#)
[Press \(<https://www.edx.org/press>\)](#) [FAQ \(<https://www.edx.org/student-faq>\)](#)
[Contact \(<https://www.edx.org/contact>\)](#)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community/>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)
© 2014 edX, some rights reserved.

[Terms of Service and Honor Code](#) -
[Privacy Policy \(<https://www.edx.org/edx-privacy-policy>\)](#)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[Syllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[Schedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 1 (1/1 point)

What were the goals of IBM when they set out to build Watson?

- To build a computer that could compete with the best human players at *Jeopardy!*. ✓
- To build a computer for *Jeopardy!* that could be used as a contestant on every show.
- To build a computer that could generate questions on *Jeopardy!*.
- To build a computer that could answer questions that are commonly believed to require human intelligence. ✓

EXPLANATION

The main goals of IBM were to build a computer that could answer questions that are commonly believed to require human intelligence, and to therefore compete with the best human players at *Jeopardy!*.

[Check](#)[Save](#)[Hide Answer](#)

You have used 1 of 3 submissions



[About \(https://www.edx.org/about-us\)](#) [Jobs \(https://www.edx.org/jobs\)](#)
[Press \(https://www.edx.org/press\)](#) [FAQ \(https://www.edx.org/student-faq\)](#)
[Contact \(https://www.edx.org/contact\)](#)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



[\(http://www.meetup.com/edX-Global-Community/\)](#)



[\(http://www.facebook.com/EdxOnline\)](#)



[\(https://twitter.com/edXOnline\)](#)



[\(https://plus.google.com/108235383044095082\)](#)



[\(http://youtube.com/user/edxonline\)](#)
© 2014 edX, some rights reserved.

[Terms of Service and Honor Code - Privacy Policy \(https://www.edx.org/edx-privacy-policy\)](#)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[Syllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[Schedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 1

IMPORTANT NOTE: This quick question will NOT count towards your quick question grade, since it is more about the Moneyball story than about analytics. You can still complete the question, but it is out of 0 points.

Which player skills did the A's think were overvalued?

- Pitching Speed ✓
- Fielding ability ✓
- The ability to get on base
- Stealing bases ✓

EXPLANATION

The A's thought pitching speed, fielding ability, and stealing bases were overvalued, while the ability to get on base was undervalued.

[Hide Answer](#)

You have used 2 of 2 submissions



[About \(<https://www.edx.org/about-us>\)](#) [Jobs \(<https://www.edx.org/jobs>\)](#)
[Press \(<https://www.edx.org/press>\)](#) [FAQ \(<https://www.edx.org/stu...-faq>\)](#)
[Contact \(<https://www.edx.org/contact>\)](#)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community/>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)

© 2014 edX, some rights reserved.

Terms of Service and Honor Code -
[Privacy Policy \(<https://www.edx.org/edx-privacy-policy>\)](#)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[yllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[chedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 1 (1/1 point)

In this class, we've learned many different methods for predicting outcomes. Which of the following methods could be used to predict whether or not someone will experience a heart attack?

- Linear Regression
- Logistic Regression ✓
- CART ✓
- Random Forest ✓

EXPLANATION

Logistic Regression, CART, or Random Forest could all be used to predict whether or not someone has a heart attack, since this is a classification problem. Linear regression would be appropriate for a problem with a continuous outcome, such as the amount of time until someone has a heart attack. In this lecture, we'll use random forest, but the other methods could be used too.

[Final Check](#)[Save](#)[Hide Answer](#)*You have used 1 of 2 submissions*

[About \(<https://www.edx.org/about-us>\)](#) [Jobs \(<https://www.edx.org/jobs>\)](#)
[Press \(<https://www.edx.org/press>\)](#) [FAQ \(<https://www.edx.org/student-faq>\)](#)
[Contact \(<https://www.edx.org/contact>\)](#)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



[\(<http://www.meetup.com/edX-Global-Community/>\)](http://www.meetup.com/edX-Global-Community/)



[\(<http://www.facebook.com/EdxOnline>\)](http://www.facebook.com/EdxOnline)



[\(<http://twitter.com/edXOnline>\)](http://twitter.com/edXOnline)



[\(<https://plus.google.com/108235383044095082>\)](https://plus.google.com/108235383044095082)



[\(<http://youtube.com/user/edxonline>\)](http://youtube.com/user/edxonline)
© 2014 edX, some rights reserved.

[Terms of Service and Honor Code -](#)
[Privacy Policy \(<https://www.edx.org/edx-privacy-policy>\)](#)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[Syllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[Schedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 1 (1/1 point)

In the next video, we'll be formulating the IMRT problem as a linear optimization problem. What do you think the decision variables in our problem will be?

- The amount of radiation to deliver to the tumor
- The intensities of the beams
- The intensities of the beamlets ✓
- The shape of the tumor

EXPLANATION

We get to decide the beamlet intensities - these will be the decision variables in our optimization problem. The amount of radiation to the tumor will be computed using the beamlet intensities, but we also want to make sure we know the amount of radiation to healthy tissue. The intensities of the beams would have been the decision variables in traditional radiation therapy, and the shape of the tumor is data.

[Final Check](#)[Save](#)[Hide Answer](#)

You have used 1 of 2 submissions



[About \(<https://www.edx.org/about-us>\)](#) [Jobs \(<https://www.edx.org/jobs>\)](#)
[Press \(<https://www.edx.org/press>\)](#) [FAQ \(<https://www.edx.org/student-faq>\)](#)
[Contact \(<https://www.edx.org/contact>\)](#)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community/>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)

© 2014 edX, some rights reserved.

Terms of Service and Honor Code -
[Privacy Policy](#) (<https://www.edx.org/edx-privacy-policy>)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[yllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[chedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 1 (1/1 point)

About how many years did it take for a team to submit a 10% improvement over Cinematch?

- 0.5
- 1.5
- 2.5 ✓
- 3.5

EXPLANATION

The contest started in October 2006, and ended in July 2009. So it took about 2.5 years for a team to submit a 10% improvement solution.

[Hide Answer](#)*You have used 1 of 1 submissions*

[About \(<https://www.edx.org/about-us>\)](#) [Jobs \(<https://www.edx.org/jobs>\)](#)
[Press \(<https://www.edx.org/press>\)](#) [FAQ \(<https://www.edx.org/stu-faq>\)](#)
[Contact \(<https://www.edx.org/contact>\)](#)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community/>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)
© 2014 edX, some rights reserved.

[Terms of Service and Honor Code - Privacy Policy \(<https://www.edx.org/edx-privacy-policy>\)](#)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[Syllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[Schedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 1 (2/2 points)

Suppose that you are trying to schedule 3 games between 6 teams (A, B, C, D, E, and F) that will occur simultaneously. Which of the following are feasible schedules?

- A plays B, C plays D, and E plays F ✓
- A plays C, B plays D, and C plays F
- A plays F, B plays E, and C plays D ✓
- A plays B, B plays C, and C plays D
- A plays D, B plays E, and C plays F ✓

EXPLANATION

Each of the teams has to play exactly one of the other teams for the games to occur simultaneously. In the second option, C is playing twice, which is impossible. In the fourth option, B and C are both playing twice.

How many different feasible schedules are there?

- 5
- 10
- 15 ✓
- 20
- 25

EXPLANATION

There are 15 different feasible schedules. We can count them by observing that A can play any of the 5 teams. Once this is fixed, we have 4 teams left. There are 3 ways to make two pairs out of 4 teams. So in total, there are $5 \times 3 = 15$ different schedules. Here is a list of all of them:

A plays B, C plays D, E plays F

A plays B, C plays E, D plays F

A plays B, C plays F, D plays E

A plays C, B plays D, E plays F

A plays C, B plays E, D plays F

A plays C, B plays F, D plays E

A plays D, B plays C, E plays F

A plays D, B plays E, C plays F

A plays D, B plays F, C plays E

A plays E, B plays C, D plays F

A plays E, B plays D, C plays F

A plays E, B plays F, C plays D

A plays F, B plays C, D plays E

A plays F, B plays D, C plays E

A plays F, B plays E, C plays D

Final Check

Save

Hide Answer

You have used 1 of 2 submissions



About (<https://www.edx.org/about-us>) Jobs (<https://www.edx.org/jobs>)
Press (<https://www.edx.org/press>) FAQ (<https://www.edx.org/student-faq>)
Contact (<https://www.edx.org/contact>)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community/>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)
© 2014 edX, some rights reserved.

Terms of Service and Honor Code -
Privacy Policy (<https://www.edx.org/edx-privacy-policy>)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[Syllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[Schedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 1 (1 point possible)

The Los Angeles Police Department sees the benefits of predictive policing as:

- Allowing more intelligent officer deployment ✓
- Eliminating the need for police officers
- Preventing crime ✓
- Catching criminals
- Using resources more effectively ✓

EXPLANATION

According to the Los Angeles Police Department, predictive policing does not eliminate the need for police officers or increase the rate at which they catch criminals. It does, however, allow more intelligent officer deployment, prevents crime, and helps them use resources more effectively.

[Hide Answer](#)

You have used 2 of 2 submissions



[About \(<https://www.edx.org/about-us>\)](#) [Jobs \(<https://www.edx.org/jobs>\)](#)
[Press \(<https://www.edx.org/press>\)](#) [FAQ \(<https://www.edx.org/student-faq>\)](#)
[Contact \(<https://www.edx.org/contact>\)](#)



EdX is a non-profit created by founding partners Harvard and MIT. Our mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community/>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)
© 2014 edX, some rights reserved.

[Terms of Service and Honor Code - Privacy Policy \(<https://www.edx.org/edx-privacy-policy>\)](#)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[Syllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[Schedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 1 (1/1 point)

Why was the city of Framingham, Massachusetts selected for this study? Select all that apply.

- It represented all types of people in the United States.
- It had an appropriate size. ✓
- It had a stable population to observe over time. ✓
- It contained an abnormally large number of people with heart disease.
- The doctors and residents in Framingham were willing to participate. ✓

EXPLANATION

The reasons for Framingham being selected for this study are listed on Slide 4 of the previous video: it had an appropriate size, it had a stable population, and the doctors and residents in the town were willing to participate. However, the city did not represent all types of people in the United States (we'll see later in the lecture how to extend the model to different populations) and there were not an abnormally large number of people with heart disease.

[Hide Answer](#)*You have used 2 of 2 submissions*

[About \(<https://www.edx.org/about-us>\)](#) [Jobs \(<https://www.edx.org/jobs>\)](#)
[Press \(<https://www.edx.org/press>\)](#) [FAQ \(<https://www.edx.org/student-faq>\)](#)
[Contact \(<https://www.edx.org/contact>\)](#)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community/>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)
© 2014 edX, some rights reserved.

[Terms of Service and Honor Code - Privacy Policy](#) (<https://www.edx.org/edx-privacy-policy>)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[Syllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[Schedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 1 (1/1 point)

Normally, a scatterplot only allows us to visualize two dimensions - one on the x-axis, and one on the y-axis. In the previous video we were able to show a third dimension on the scatterplot using what attribute?

- Color ✓
- Shape
- Location

EXPLANATION

On slide 3, we show the scatterplot from slide 2, but with the number of cylinders shown by the color of the points. This allows us to visualize a third dimension of our data.

[Hide Answer](#)

You have used 1 of 1 submissions



[About \(https://www.edx.org/about-us\)](#) [Jobs \(https://www.edx.org/jobs\)](#)
[Press \(https://www.edx.org/press\)](#) [FAQ \(https://www.edx.org/stu...-faq\)](#)
[Contact \(https://www.edx.org/contact\)](#)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



[\(http://www.meetup.com/edX-Global-Community/\)](#)



[\(http://www.facebook.com/EdxOnline\)](#)



[\(https://twitter.com/edXOnline\)](#)



[\(https://plus.google.com/108235383044095082\)](#)



[\(http://youtube.com/user/edxonline\)](#)

© 2014 edX, some rights reserved.

[Terms of Service and Honor Code - Privacy Policy \(https://www.edx.org/edx-privacy-policy\)](#)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[Syllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[Schedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 1 (1/1 point)

At which university was the first version of R developed?

- MIT
- University of Auckland ✓
- Stanford University
- Oxford University

EXPLANATION

The first version of R was developed by Robert Gentleman and Ross Ihaka at the University of Auckland in the mid-1990s.

[Hide Answer](#)

You have used 1 of 1 submissions

[About \(<https://www.edx.org/about-us>\)](#) [Jobs \(<https://www.edx.org/jobs>\)](#)
[Press \(<https://www.edx.org/press>\)](#) [FAQ \(<https://www.edx.org/student-faq>\)](#)
[Contact \(<https://www.edx.org/contact>\)](#)

EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever they have Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.

[\(http://www.meetup.com/edX-Global-Community/\)](http://www.meetup.com/edX-Global-Community/)[\(<http://www.facebook.com/EdxOnline>\)](http://www.facebook.com/EdxOnline)[\(<https://twitter.com/edXOnline>\)](https://twitter.com/edXOnline)[\(<https://plus.google.com/108235383044095082>\)](https://plus.google.com/108235383044095082)[\(<http://youtube.com/user/edxonline>\)](http://youtube.com/user/edxonline)

© 2014 edX, some rights reserved.

Terms of Service and Honor Code -
Privacy Policy (<https://www.edx.org/edx-privacy-policy>)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[yllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[chedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 2 (2/2 points)

Suppose that, as in the previous video, regular seats cost \$617 and discount seats cost \$238. We are selling 166 seats. The demand for regular seats is 150 and the demand for discount seats is 150.

How many discount seats should we sell?

 16 16**Answer:** 16**EXPLANATION**

Since regular seats give us more revenue, we should sell enough regular seats to meet the demand. This means that we will sell 150 regular seats. Since our capacity is 166, this leaves 16 seats to sell to discount customers.

What would our total revenue be, for both regular and discount seats, assuming that we have a full plane?

 96358 96358**Answer:** 96358**EXPLANATION**

We would sell 150 seats to regular customers, giving us a revenue of $\$617 * 150$, and 16 seats to discount customers, giving us a revenue of $\$238 * 16$. Our total revenue would be $\$617 * 150 + \$238 * 16 = \$96,358$.

 Check Save Hide Answer

You have used 1 of 4 submissions





(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)

© 2014 edX, some rights reserved.

[Terms of Service and Honor Code](#) -
[Privacy Policy](#) (<https://www.edx.org/edx-privacy-policy>)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[Syllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[Schedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

QUICK QUESTION 2 (2/2 points)

In the previous video, we saw a small example with 3 men and 3 women. We defined a "match" as an assignment of each man to exactly one woman, and each woman to exactly one man. The optimal match in the previous video was to assign man 1 to woman 3, man 2 to woman 1, and man 3 to woman 2.

How many different feasible matches are there in the example with 3 men and 3 women? (HINT: Another feasible match is to assign man 1 to woman 1, man 2 to woman 3, and man 3 to woman 2.)

6

6

Answer: 6

How many different feasible matches are there with 5 men and 5 women? (HINT: First assign man 1 to one of the women. How many choices are there? Then assign man 2 to a woman - how many choices are there now? Repeat this until every man is matched to every woman.)

- 10
- 20
- 50
- 100
- 120 ✓
- 150

EXPLANATION

In the first case, there are 6 possible matches. The first man can be assigned to any of the 3 women (3 choices). Then the second man can be assigned to any of the remaining 2 women (2 choices). The third man is automatically assigned to the remaining woman. So there are a total of $3 \times 2 = 6$ choices.

In the second case, there are 120 possible matches. The first man can be assigned to any of the 5 women (5 choices), then the second man can be assigned to any of the remaining women (4 choices), etc. This gives a total of $5 \times 4 \times 3 \times 2 = 120$ different matches.

You can easily see how the number of possible matches gets very large on online dating sites!

[Check](#)

[Save](#)

[Hide Answer](#)

You have used 1 of 3 submissions





EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community/>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)
© 2014 edX, some rights reserved.

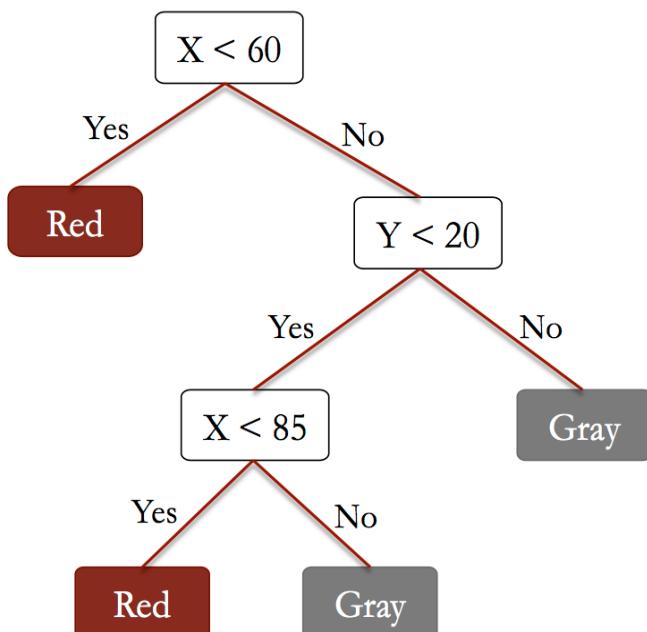
[Terms of Service and Honor Code](#) -
[Privacy Policy](#) (<https://www.edx.org/edx-privacy-policy>)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[yllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[chedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 2 (2/2 points)

Suppose that you have the following CART tree:



How many splits are in this tree?

Answer: 3



For which data observations should we predict "Red", according to this tree?

- If X is less than 60, and Y is any value. ✓
- If X is greater than or equal to 60, and Y is greater than or equal to 20.
- If X is greater than or equal to 85, and Y is less than 20.
- If X is greater than or equal to 60 and less than 85, and Y is less than 20. ✓

EXPLANATION

This tree has three splits. The first split says to predict "Red" if X is less than 60, regardless of the value of Y. Otherwise, we move to the second split. The second split says to check the value of Y - if it is greater than or equal to 20, predict "Gray". Otherwise, we move to the third split. This split checks the value of X again. If X is less than 85 (and greater than or equal to 60 by the first split) and Y is less than 20, then we predict "Red". Otherwise, we predict "Gray".

[Final Check](#)[Save](#)[Hide Answer](#)

You have used 2 of 3 submissions



About (<https://www.edx.org/about-us>) Jobs (<https://www.edx.org/jobs>)
Press (<https://www.edx.org/press>) FAQ (<https://www.edx.org/student-faq>)
Contact (<https://www.edx.org/contact>)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community/>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)
© 2014 edX, some rights reserved.

Terms of Service and Honor Code -
Privacy Policy (<https://www.edx.org/edx-privacy-policy>)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[Syllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[Schedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 2 (1 point possible)

A common problem in analytics is that you have some data available, but it's not the ideal dataset. This is the case for this problem, where we only have claims data. Which of the following pieces of information would we ideally *like to have* in our dataset, but are *not included* in claims data? (Select all that apply.)

- Blood test results ✓
- Drugs prescribed to the patient
- Physical exam results (weight, height, blood pressure, etc.) ✓

EXPLANATION

In claims data, we have drugs prescribed to the patient, but we don't have blood test results or physical exam results.

[Hide Answer](#)

You have used 2 of 2 submissions



[About \(<https://www.edx.org/about-us>\)](#) [Jobs \(<https://www.edx.org/jobs>\)](#)
[Press \(<https://www.edx.org/press>\)](#) [FAQ \(<https://www.edx.org/student-faq>\)](#)
[Contact \(<https://www.edx.org/contact>\)](#)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community/>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)
© 2014 edX, some rights reserved.

Terms of Service and Honor Code -
Privacy Policy (<https://www.edx.org/edx-privacy-policy>)



[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[yllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[chedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 2 (1/1 point)

For which of the following reasons is *Jeopardy!* challenging?

- A wide variety of categories. ✓
- Expert knowledge is required in a few specific categories.
- Speed is required - you have to buzz in faster than your competitors. ✓
- The categories and clues are often cryptic. ✓

EXPLANATION

Jeopardy! is challenging because there are a wide variety of categories, speed is required, and the categories and clues are cryptic. Expert knowledge is not generally required.

[Final Check](#)[Save](#)[Hide Answer](#)*You have used 2 of 3 submissions*

[About \(<https://www.edx.org/about-us>\)](#) [Jobs \(<https://www.edx.org/jobs>\)](#)
[Press \(<https://www.edx.org/press>\)](#) [FAQ \(<https://www.edx.org/stu-faq>\)](#)
[Contact \(<https://www.edx.org/contact>\)](#)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community/>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)

© 2014 edX, some rights reserved.

[Terms of Service and Honor Code](#) -
[Privacy Policy](#) (<https://www.edx.org/edx-privacy-policy>)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[yllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[chedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 2 (2 points possible)

Which of the following dependent variables are categorical?

- Deciding whether to buy, sell, or hold a stock ✓
- The weekly revenue of a company
- The winner of an election with two candidates ✓
- The day of the week with the highest revenue ✓
- The number of daily car thefts in New York City
- Whether or not revenue will exceed \$50,000 ✓

EXPLANATION

The weekly revenue of a company is not categorical, since it has a large number of possible values, on a continuous range. The number of daily car thefts in New York City is also not categorical because the number of car thefts could range from 0 to hundreds.

On the other hand, the other options each have a limited number of possible outcomes.

Which of the following dependent variables are binary?

- Deciding whether to buy, sell, or hold a stock
- The weekly revenue of a company
- The winner of an election with two candidates ✓
- The day of the week with the highest revenue
- The number of car thefts in New York City
- Whether or not revenue will exceed \$50,000 ✓

EXPLANATION

The only variables with two possible outcomes are the winner of an election with two candidates, and whether or not revenue will exceed \$50,000.

[Hide Answer](#)*You have used 3 of 3 submissions*



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)

© 2014 edX, some rights reserved.

[Terms of Service and Honor Code](#) -

[Privacy Policy](#) (<https://www.edx.org/edx-privacy-policy>)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[Syllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[Schedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 2 (1/1 point)

If a baseball team scores 713 runs and allows 614 runs, how many games do we expect the team to win?

Using the linear regression model constructed during the lecture, enter the number of games we expect the team to win:

91

Answer: 91

EXPLANATION

Our linear regression model was

$$\text{Wins} = 80.94 + 0.10 * (\text{Run Difference})$$

Here, the run difference is 99, so our prediction is

$$\text{Wins} = 80.94 + 0.10 * 99 = 91 \text{ games.}$$

[Check](#)[Save](#)[Hide Answer](#)

You have used 1 of 3 submissions



[About \(<https://www.edx.org/about-us>\)](#) [Jobs \(<https://www.edx.org/jobs>\)](#)
[Press \(<https://www.edx.org/press>\)](#) [FAQ \(<https://www.edx.org/student-faq>\)](#)
[Contact \(<https://www.edx.org/contact>\)](#)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community/>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edxonline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)

© 2014 edX, some rights reserved.

[Terms of Service and Honor Code - Privacy Policy \(<https://www.edx.org/edx-privacy-policy>\)](#)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[Syllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[Schedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 2 (2/2 points)

In the previous video, we discussed how we split the data into three groups, or buckets, according to cost.

Which bucket has the most data, in terms of number of patients?

- Cost Bucket 1 ✓
- Cost Bucket 2
- Cost Bucket 3

Which bucket probably has the densest data, in terms of number of claims per person?

- Cost Bucket 1
- Cost Bucket 2
- Cost Bucket 3 ✓

EXPLANATION

Cost Bucket 1 contains the most patients (see slide 7 of the previous video), and Cost Bucket 3 probably has the densest data, since these are the patients with the highest cost in terms of claims.

[Final Check](#)[Save](#)[Hide Answer](#)

You have used 1 of 2 submissions



[About \(<https://www.edx.org/about-us>\)](#) [Jobs \(<https://www.edx.org/jobs>\)](#)
[Press \(<https://www.edx.org/press>\)](#) [FAQ \(<https://www.edx.org/student-faq>\)](#)
[Contact \(<https://www.edx.org/contact>\)](#)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, physics, health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community/>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/1082353830440950821>)



(<http://youtube.com/user/edxonline>)

© 2014 edX, some rights reserved.

[Terms of Service and Honor Code - Privacy Policy](#) (<https://www.edx.org/edx-privacy-policy>)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[Syllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[Schedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 2 (2/2 points)

In the previous video, we constructed the optimization problem (see the last slide).

If the beamlet intensity of the first beamlet is set to 3, how much radiation will that beamlet deliver to tumor voxels?

Answer: 6

How much radiation will it deliver to healthy tissue voxels?

Answer: 9**EXPLANATION**

Beamlet 1 hits one tumor voxel, and two healthy tissue voxels. At unit intensity, it delivers a dose of 2 to the tumor voxel, a dose of 2 to the first healthy tissue voxel, and a dose of 1 to the second healthy tissue voxel. At intensity 3, this means that it will deliver a dose of $2 \times 3 = 6$ to the tumor voxel, and $2 \times 3 + 1 \times 3 = 9$ to the healthy tissue voxels.

[Check](#)[Save](#)[Hide Answer](#)

You have used 1 of 4 submissions



[About \(<https://www.edx.org/about-us>\)](#) [Jobs \(<https://www.edx.org/jobs>\)](#)
[Press \(<https://www.edx.org/press>\)](#) [FAQ \(<https://www.edx.org/student-faq>\)](#)
[Contact \(<https://www.edx.org/contact>\)](#)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



[\(http://www.meetup.com/edX-Global-Community/\)](#)



[\(http://www.facebook.com/EdxOnline\)](#)



[\(https://twitter.com/edXOnline\)](#)



[\(https://plus.google.com/1082353830440950827\)](#)



[\(http://youtube.com/user/edxonline\)](#)

© 2014 edX, some rights reserved.

[Terms of Service and Honor Code - Privacy Policy \(<https://www.edx.org/edx-privacy-policy>\)](#)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[Syllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[Schedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 2 (2/2 points)

Let's consider a recommendation system on Amazon.com, an online retail site.

If Amazon.com constructs a recommendation system for books, and would like to use the same exact algorithm for shoes, what type would it have to be?

- Collaborative Filtering ✓
- Content Filtering

If Amazon.com would like to suggest books to users based on the previous books they have purchased, what type of recommendation system would it be?

- Collaborative Filtering
- Content Filtering ✓

EXPLANATION

In the first case, the recommendation system would have to be collaborative filtering, since it can't use information about the items. In the second case, the recommendation system would be content filtering since other users are not involved.

[Hide Answer](#)

You have used 1 of 1 submissions



[About \(<https://www.edx.org/about-us>\)](#) [Jobs \(<https://www.edx.org/jobs>\)](#)
[Press \(<https://www.edx.org/press>\)](#) [FAQ \(<https://www.edx.org/student-faq>\)](#)
[Contact \(<https://www.edx.org/contact>\)](#)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



[\(<http://www.meetup.com/edX-Global-Community/>\)](http://www.meetup.com/edX-Global-Community/)



[\(<http://www.facebook.com/EdxOnline>\)](http://www.facebook.com/EdxOnline)



[\(<https://twitter.com/edXOnline>\)](https://twitter.com/edXOnline)



[\(<https://plus.google.com/108235383044095082>\)](https://plus.google.com/108235383044095082)



[\(<http://youtube.com/user/edxonline>\)](http://youtube.com/user/edxonline)
© 2014 edX, some rights reserved.

[Terms of Service and Honor Code - Privacy Policy \(<https://www.edx.org/edx-privacy-policy>\)](#)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[Syllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[Schedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

QUICK QUESTION 2 (4/4 points)

For each of the decisions below, indicate if the decision variables would be binary, integer, or neither.

1) We have 20 students, and we want to assign them to one of two groups.

- Binary ✓
- Integer
- Neither

2) The owner of 5 clothing stores needs to decide how many shirts, pants, and hats to send to each store, given historical sales data.

- Binary
- Integer ✓
- Neither

3) After try-outs, the coach of a basketball team needs to decide which people should make the team (15 people tried out).

- Binary ✓
- Integer
- Neither

4) A fertilizer company is trying to decide how much (in grams) of three different compounds to add to each bag of fertilizer.

- Binary
- Integer
- Neither ✓

EXPLANATION

The first and third decisions require binary decision variables, since they are both assignment problems. In the first case, we'll have a binary decision variable for each student (20 decision variables). In the third case, we'll have a binary decision variable for each person (15 decision variables).

The second decision requires integer decision variables, since the owner needs to decide how many of each item to send to each store (15 decision variables). Since fractional items would not make sense, the decisions are integer.

The fourth decision does not need binary nor integer decision variables, because the amount in grams can be fractional.

[Final Check](#)[Save](#)[Hide Answer](#)*You have used 1 of 2 submissions*

About (<https://www.edx.org/about-us>) Jobs (<https://www.edx.org/jobs>)
Press (<https://www.edx.org/press>) FAQ (<https://www.edx.org/student-faq>)
Contact (<https://www.edx.org/contact>)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community/>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)

© 2014 edX, some rights reserved.

Terms of Service and Honor Code -
Privacy Policy (<https://www.edx.org/edx-privacy-policy>)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[Syllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[Schedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 2 (1/1 point)

For which of the following situations would a heat map be an appropriate visualization choice?

- Determining if crime is higher or lower on warmer days
- Visualizing the areas on a geographical map with the most crime ✓
- Comparing crime counts by police district and time throughout a city ✓
- Analyzing which months of the year have the most crime on average

EXPLANATION

A heatmap would be useful for the middle two options, because they are trying to visualize crime counts relative to two variables. For the first option, you could use a basic scatterplot with time on the x-axis and amount of crime on the y-axis. For the last option, you could use a bar plot with a bar for each month and the height being the average amount of crime in that month.

[Hide Answer](#)

You have used 2 of 2 submissions



[About \(<https://www.edx.org/about-us>\)](#) [Jobs \(<https://www.edx.org/jobs>\)](#)
[Press \(<https://www.edx.org/press>\)](#) [FAQ \(<https://www.edx.org/student-faq>\)](#)
[Contact \(<https://www.edx.org/contact>\)](#)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community/>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)
© 2014 edX, some rights reserved.

[Terms of Service and Honor Code - Privacy Policy](#) (<https://www.edx.org/edx-privacy-policy>)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[yllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[chedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 2 (2/2 points)

Are "risk factors" the independent variables or the dependent variables in our model?

- Independent Variables ✓
- Dependent Variables
- Neither

EXPLANATION

Risk factors are the independent variables in our model, and are what we will use to predict the dependent variable.

In many situations, a dataset is handed to you and you are tasked with discovering which variables are important. But for the Framingham Heart Study, the researchers had to collect data from patients. In a situation like this one, where data needs to be collected by the researchers, should the risk factors be defined before or after the data is collected?

- Before ✓
- After

EXPLANATION

The researchers should first hypothesize risk factors, and then collect data corresponding to those risk factors. Of course, they could always define more risk factors later and collect more data, but this data would take longer to collect.

[Hide Answer](#)

You have used 1 of 1 submissions



[About \(<https://www.edx.org/about-us>\)](#) [Jobs \(<https://www.edx.org/jobs>\)](#)
[Press \(<https://www.edx.org/press>\)](#) [FAQ \(<https://www.edx.org/student-faq>\)](#)
[Contact \(<https://www.edx.org/contact>\)](#)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



[\(<http://www.meetup.com/edX-Global-Community/>\)](http://www.meetup.com/edX-Global-Community/)



[\(<http://www.facebook.com/EdxOnline>\)](http://www.facebook.com/EdxOnline)



[\(<https://twitter.com/edXOnline>\)](https://twitter.com/edXOnline)



[\(<https://plus.google.com/108235383044095082>\)](https://plus.google.com/108235383044095082)



[\(<http://youtube.com/user/edxonline>\)](http://youtube.com/user/edxonline)

© 2014 edX, some rights reserved.

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)

Help

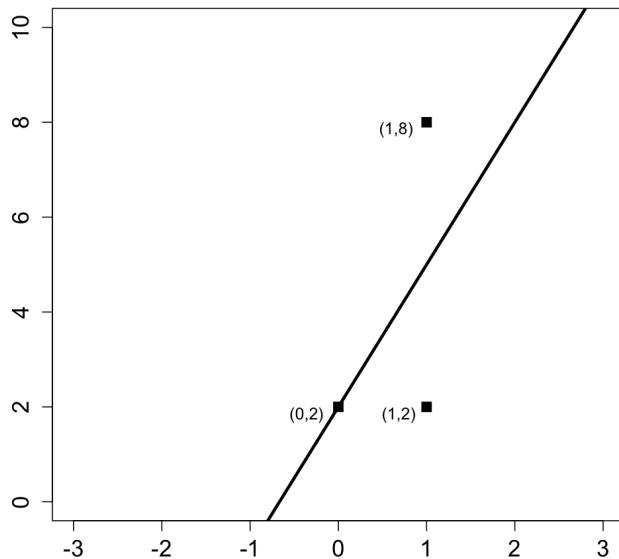
[yllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[chedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

QUICK QUESTION 2 (4/4 points)

The following figure shows three data points and the best fit line

$$y = 3x + 2.$$

The x-coordinate, or "x", is our independent variable and the y-coordinate, or "y", is our dependent variable.



Please answer the following questions using this figure.

What is the baseline prediction?

4

4

Answer: 4

EXPLANATION

The baseline prediction is the average value of the dependent variable. Since our dependent variable takes values 2, 2, and 8 in our data set, the average is $(2+2+8)/3 = 4$.

What is the Sum of Squared Errors (SSE) ?

18

18

Answer: 18

EXPLANATION

The SSE is computed by summing the squared errors between the actual values and our predictions. For each value of the independent variable (x), our best fit line makes the following predictions:

If $x = 0$, $y = 3(0) + 2 = 2$,

If $x = 1$, $y = 3(1) + 2 = 5$.

Thus we make an error of 0 for the data point (0,2), an error of 3 for the data point (1,2), and an error of 3 for the data point (1,8). So we have

$$\text{SSE} = 0^2 + 3^2 + 3^2 = 18.$$

What is the Total Sum of Squares (SST) ?

24

24

Answer: 24

EXPLANATION

The SST is computed by summing the squared errors between the actual values and the baseline prediction. From the first question, we computed the baseline prediction to be 4. Thus the SST is:

$$\text{SST} = (2 - 4)^2 + (2 - 4)^2 + (8 - 4)^2 = 24.$$

What is the R^2 of the model?

0.25

0.25

Answer: 0.25

EXPLANATION

The R^2 formula is:

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}}$$

Thus using our answers to the previous questions, we have that

$$R^2 = 1 - \frac{18}{24} = 0.25.$$

You have used 1 of 5 submissions

Your answers have been saved but not graded. Click 'Check' to grade them.





EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community/>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)
© 2014 edX, some rights reserved.

[Terms of Service and Honor Code](#) -

[Privacy Policy](#) (<https://www.edx.org/edx-privacy-policy>)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[yllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[chedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 2 (1/1 point)

Which of these problems is the LEAST likely to be a good application of natural language processing?

- Processing medical records to extract symptoms
- Judging the winner of a poetry contest ✓
- Flagging customer reviews on Amazon as problematic
- Automatically organizing emails based on their content

EXPLANATION

Judging the winner of a poetry contest requires a deep level of human understanding and emotion. Perhaps someday a computer will be able to accurately judge the winner of a poetry contest, but currently the other three tasks are much better suited for natural language processing.

[Hide Answer](#)*You have used 2 of 2 submissions*

[About \(<https://www.edx.org/about-us>\)](#) [Jobs \(<https://www.edx.org/jobs>\)](#)
[Press \(<https://www.edx.org/press>\)](#) [FAQ \(<https://www.edx.org/students/faq>\)](#)
[Contact \(<https://www.edx.org/contact>\)](#)



 EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



[\(<http://www.meetup.com/edX-Global-Community/>\)](#)



[\(<http://www.facebook.com/EdxOnline>\)](#)



[\(<https://twitter.com/edXOnline>\)](#)



[\(<https://plus.google.com/108235383044095082>\)](#)



[\(<http://youtube.com/user/edxonline>\)](#)
© 2014 edX, some rights reserved.

[Terms of Service and Honor Code - Privacy Policy \(<https://www.edx.org/edx-privacy-policy>\)](#)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[Syllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[Schedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 2 (1/1 point)

Why is it particularly helpful for WHO to provide data visualizations?

- There is no other way to display information shown in visualizations like the Energy Consumption one.
- When communicating information to the general public, a visualization like the Energy Consumption one is much easier to absorb than a table of numbers would be.
- Visualizations can easily be used by policymakers and others who wish to present data from WHO.

EXPLANATION

While there are other ways to display the data given in many visualizations (like tables), visualizations help to better communicate data to the public and can easily be used by others in presentations.

[Final Check](#)[Save](#)[Hide Answer](#)

You have used 1 of 2 submissions



[About \(<https://www.edx.org/about-us>\)](#) [Jobs \(<https://www.edx.org/jobs>\)](#)
[Press \(<https://www.edx.org/press>\)](#) [FAQ \(<https://www.edx.org/student-faq>\)](#)
[Contact \(<https://www.edx.org/contact>\)](#)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



[\(http://www.meetup.com/edX-Global-Community/\)](http://www.meetup.com/edX-Global-Community/)



[\(<http://www.facebook.com/EdxOnline>\)](http://www.facebook.com/EdxOnline)



[\(<https://twitter.com/edXOnline>\)](https://twitter.com/edXOnline)



[\(<https://plus.google.com/108235383044095082>\)](https://plus.google.com/108235383044095082)



[\(<http://youtube.com/user/edxonline>\)](http://youtube.com/user/edxonline)
© 2014 edX, some rights reserved.

[Terms of Service and Honor Code - Privacy Policy \(<https://www.edx.org/edx-privacy-policy>\)](#)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[yllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[chedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 2 (1/1 point)

Which of the following are recommended variable names in R? (Select all correct answers.)

- SquareRoot2 ✓
- Square Root2
- Square_Root2
- Square2.Root ✓
- 2SquareRoot

EXPLANATION

SquareRoot2 and Square2.Root are recommended variable names. The second option is not recommended because it has a space, the third option is not recommended (although, it is acceptable) because it has an underscore, and the fifth option is not recommended (although, it is acceptable) because it starts with a number.

[Hide Answer](#)*You have used 2 of 2 submissions*

About (<https://www.edx.org/about-us>) Jobs (<https://www.edx.org/jobs>)
Press (<https://www.edx.org/press>) FAQ (<https://www.edx.org/student-faq>)
Contact (<https://www.edx.org/contact>)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community/>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edxonline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)
© 2014 edX, some rights reserved.



Terms of Service and Honor Code -
[Privacy Policy](#) (<https://www.edx.org/edx-privacy-policy>)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[Syllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[Schedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 3 (2/2 points)

In the previous video, we set up an optimization problem with 2 different types of tickets.

How many decision variables would we have if there were 4 different types of tickets?

4

4

Answer: 4

How many constraints would we have if there were 4 different types of tickets (with two different types of tickets, our model has 5 constraints: one capacity constraint, two demand constraints, and two non-negativity constraints)?

9

9

Answer: 9**EXPLANATION**

If our model had 4 different types of tickets, we would have four decision variables, one for each type of ticket. We would have 9 constraints, since we would need one capacity constraint, 4 demand constraints, and 4 non-negativity constraints.

[Check](#)[Save](#)[Hide Answer](#)

You have used 1 of 3 submissions



[About \(<https://www.edx.org/about-us>\)](#) [Jobs \(<https://www.edx.org/jobs>\)](#)
[Press \(<https://www.edx.org/press>\)](#) [FAQ \(<https://www.edx.org/student-faq>\)](#)
[Contact \(<https://www.edx.org/contact>\)](#)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



[\(http://www.meetup.com/edX-Global-Community/\)](#)



[\(http://www.facebook.com/EdxOnline\)](#)



[\(https://twitter.com/edXOnline\)](#)



[\(https://plus.google.com/1082353830440950827\)](#)



[\(http://youtube.com/user/edxonline\)](#)

© 2014 edX, some rights reserved.

Terms of Service and Honor Code -
[Privacy Policy \(<https://www.edx.org/edx-privacy-policy>\)](#)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[yllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[chedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 3 (1/1 point)

While we don't have all of the data we would ideally like to have in this problem (like test results), we can define new variables using the data we do have. Which of the following were new variables defined to help predict healthcare cost?

- Variables to capture chronic conditions ✓
- Noncompliance to treatment ✓
- Illness severity ✓
- Interactions between illnesses ✓

EXPLANATION

All of these variables were defined using the claims data to improve cost predictions. This shows how the intuition of experts can be used to define new variables and improve the model.

[Check](#)[Save](#)[Hide Answer](#)*You have used 1 of 3 submissions*

About (<https://www.edx.org/about-us>) Jobs (<https://www.edx.org/jobs>)
Press (<https://www.edx.org/press>) FAQ (<https://www.edx.org/students-faq>)
Contact (<https://www.edx.org/contact>)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community/>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/1082353830440950827>)



(<http://youtube.com/user/edxonline>)
© 2014 edX, some rights reserved.

[Terms of Service and Honor Code](#) -
[Privacy Policy](#) (<https://www.edx.org/edx-privacy-policy>)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[yllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[chedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 3 (1/1 point)

Which of the following two questions do you think would be EASIEST for a computer to answer?

- Was Abraham Lincoln generally considered a happy man?
- What year was Abraham Lincoln born? ✓

EXPLANATION

The second question would be the easiest, because the answer is a fact. The first question is much more subjective.

[Hide Answer](#)*You have used 1 of 1 submissions*

[About \(<https://www.edx.org/about-us>\)](#) [Jobs \(<https://www.edx.org/jobs>\)](#)
[Press \(<https://www.edx.org/press>\)](#) [FAQ \(<https://www.edx.org/student-faq>\)](#)
[Contact \(<https://www.edx.org/contact>\)](#)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all anywhere in the world, wherever there is Internet access. EdX's frne MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community/>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)

© 2014 edX, some rights reserved.

[Terms of Service and Honor Code - Privacy Policy \(<https://www.edx.org/edx-privacy-policy>\)](#)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[yllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[chedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 3 (3/3 points)

Suppose the coefficients of a logistic regression model with two independent variables are as follows:

$$\beta_0 = -1.5, \quad \beta_1 = 3, \quad \beta_2 = -0.5$$

And we have an observation with the following values for the independent variables:

$$x_1 = 1, \quad x_2 = 5$$

What is the value of the Logit for this observation? Recall that the Logit is $\log(\text{Odds})$.

Answer: -1**EXPLANATION**

The Logit is just $\log(\text{Odds})$, and looks like the linear regression equation. So the Logit is $-1.5 + 3*1 - 0.5*5 = -1$.

What is the value of the Odds for this observation? Note that you can compute e^x , for some number x , in your R console by typing `exp(x)`. The function `exp()` computes the exponential of its argument.

Answer: 0.3678794**EXPLANATION**

Using the value of the Logit from the previous question, we have that $\text{Odds} = e^{-1} = 0.3678794$.

What is the value of $P(y = 1)$ for this observation?

Answer: 0.2689414**EXPLANATION**

Using the Logistic Response Function, we can compute that $P(y = 1) = 1/(1 + e^{-\text{Logit}}) = 1/(1 + e^{-1}) = 0.2689414$.

You have used 1 of 5 submissions



About (<https://www.edx.org/about-us>) Jobs (<https://www.edx.org/jobs>)
Press (<https://www.edx.org/press>) FAQ (<https://www.edx.org/student-faq>)
Contact (<https://www.edx.org/contact>)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community/>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)
© 2014 edX, some rights reserved.

[Terms of Service and Honor Code](#) -
[Privacy Policy](#) (<https://www.edx.org/edx-privacy-policy>)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)

Help

[yllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[chedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

QUICK QUESTION 3 (2/2 points)

If a baseball team's OBP is 0.311 and SLG is 0.405, how many runs do we expect the team to score?

Using the linear regression model constructed during the lecture, enter the number of runs we expect the team to score:

Answer: 689

EXPLANATION

Our linear regression model was:

$$\text{Runs Scored} = -804.63 + 2737.77 * (\text{OBP}) + 1584.91 * (\text{SLG})$$

Here, OBP is 0.311 and SLG is 0.405, so our prediction is:

$$\text{Runs Scored} = -804.63 + 2737.77 * 0.311 + 1584.91 * 0.405 = 689 \text{ runs}$$

If a baseball team's opponents OBP (OOBP) is 0.297 and opponents SLG (OSLG) is 0.370, how many runs do we expect the team to allow?

Using the linear regression model constructed during the lecture, enter the number of runs we expect the team to allow:

 =**Answer:** 588

EXPLANATION

Our linear regression model was:

$$\text{Runs Allowed} = -837.38 + 2913.60 * (\text{OOBP}) + 1514.29 * (\text{OSLG})$$

Here, OOBP is 0.297 and OSLG is 0.370, so our prediction is:

$$\text{Runs Scored} = -837.38 + 2913.60 * (.297) + 1514.29 * (.370) = 588 \text{ runs}$$

You have used 1 of 4 submissions

About (<https://www.edx.org/about-us>) Jobs (<https://www.edx.org/jobs>)
Press (<https://www.edx.org/press>) FAQ (<https://www.edx.org/student-faq>)
Contact (<https://www.edx.org/contact>)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community/>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)
© 2014 edX, some rights reserved.

Terms of Service and Honor Code -
Privacy Policy (<https://www.edx.org/edx-privacy-policy>)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[yllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[chedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 3 (1/1 point)

K-means clustering differs from Hierarchical clustering in a couple important ways. Which of the following statements are true?

- In k-means clustering, you have to pick the number of clusters you want before you run the algorithm. ✓
- In k-means clustering, you can pick the number of clusters you want after the algorithm is done, just like in Hierarchical clustering.

EXPLANATION

In k-means clustering, you have to pick the number of clusters before you run the algorithm, but the computational effort needed is much less than that for hierarchical clustering (we'll see this in more detail during the recitation).

[Hide Answer](#)*You have used 1 of 1 submissions*

[About \(<https://www.edx.org/about-us>\)](#) [Jobs \(<https://www.edx.org/jobs>\)](#)
[Press \(<https://www.edx.org/press>\)](#) [FAQ \(<https://www.edx.org/student-faq>\)](#)
[Contact \(<https://www.edx.org/contact>\)](#)



EdX is a non-profit created by founding partners Harvard and MIT, whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



[\(http://www.meetup.com/edX-Global-Community/\)](http://www.meetup.com/edX-Global-Community/)



[\(<http://www.facebook.com/EdxOnline>\)](http://www.facebook.com/EdxOnline)



[\(<https://twitter.com/edXOnline>\)](https://twitter.com/edXOnline)



[\(<https://plus.google.com/108235383044095082>\)](https://plus.google.com/108235383044095082)



[\(<http://youtube.com/user/edxonline>\)](http://youtube.com/user/edxonline)
© 2014 edX, some rights reserved.

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[yllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[chedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 3 (1/1 point)

In our optimal solution, we are giving the maximum allowed dose to the spinal cord (5). If we were to relax this, how much could we decrease the objective? Change the right-hand-side (RHS) of the spinal cord constraint to 6, and re-solve the model. By how much did we decrease the objective? (Hint: the previous objective value was 22.75)

0.58333

0.58333

Answer: 0.583333**EXPLANATION**

If you change the RHS of the spinal cord constraint to 6 and re-solve the model (Tools->Solver, then hit solve) the new objective is 22.666667. So we decreased the objective by 0.5833333.

[Final Check](#)[Save](#)[Hide Answer](#)*You have used 2 of 3 submissions*

[About \(<https://www.edx.org/about-us>\)](#) [Jobs \(<https://www.edx.org/jobs>\)](#)
[Press \(<https://www.edx.org/press>\)](#) [FAQ \(<https://www.edx.org/student-faq>\)](#)
[Contact \(<https://www.edx.org/contact>\)](#)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community/>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)
 © 2014 edX, some rights reserved.



(<http://youtube.com/user/edxonline>)
 © 2014 edX, some rights reserved.



Terms of Service and Honor Code -

Privacy Policy (<https://www.edx.org/edx-privacy-policy>)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[Syllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[Schedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

QUICK QUESTION 3 (1/1 point)

In the previous video, we discussed how clustering is used to split the data into similar groups. Which of the following tasks do you think are appropriate for clustering?

- Dividing search results on Google into categories based on the topic ✓
- Grouping players into different "types" of basketball players that make it to the NBA ✓
- Predicting the winner of the Major League Baseball World Series

Help

EXPLANATION

The first two options are appropriate tasks for clustering. Clustering probably wouldn't help us predict the winner of the World Series.

[Final Check](#)[Save](#)[Hide Answer](#)

You have used 1 of 2 submissions



[About \(<https://www.edx.org/about-us>\)](#) [Jobs \(<https://www.edx.org/jobs>\)](#)
[Press \(<https://www.edx.org/press>\)](#) [FAQ \(<https://www.edx.org/student-faq>\)](#)
[Contact \(<https://www.edx.org/contact>\)](#)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community/>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)
© 2014 edX, some rights reserved.



Terms of Service and Honor Code -
Privacy Policy (<https://www.edx.org/edx-privacy-policy>)



[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[Syllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[Schedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

QUICK QUESTION 3 (2/2 points)

Suppose we had two more teams in our tournament (for a total of 6 teams). Each division would have 3 teams. So each team plays TWO teams twice (the teams in their division), and each team plays three teams once (the teams in the other division). This means that the tournament will last for 7 weeks. How many decision variables would we have?

105

105

Answer: 105

EXPLANATION

We would have 105 decision variables because we have 7 weeks, and 15 different pairs of teams.

Help

How many constraints would we have? Don't include the constraints that force the variables to be binary when counting the constraints here. (HINT: We would have 6 division constraints, since each pair in each division needs to play twice.)

57

57

Answer: 57

EXPLANATION

We would have 6 division constraints, 9 non-division constraints (each of the three teams in one division has to play each of the three teams in the other division), and 42 constraints to make sure each team only plays one team each week (6 teams times 7 weeks).

[Final Check](#)[Save](#)[Hide Answer](#)

You have used 3 of 4 submissions





(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)

© 2014 edX, some rights reserved.

[Terms of Service and Honor Code](#) -
[Privacy Policy](#) (<https://www.edx.org/edx-privacy-policy>)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[yllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[chedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 3 (2/2 points)

Create a new line plot, like the one in Video 3, but add the argument "linetype=2". So the geom_line part of the plotting command should look like:

```
geom_line(aes(group=1), linetype=2)
```

What does this do?

- Makes the line thicker
- Changes the color of the line to blue
- Makes the line dashed ✓
- Makes the line lighter in color

Now, change the alpha parameter to 0.3 by replacing "linetype=2" with "alpha=0.3" in the plot command. What does this do?

- Makes the line thicker
- Changes the color of the line to blue
- Makes the line dashed
- Makes the line lighter in color ✓

EXPLANATION

The linetype parameter makes the line dashed, and the alpha parameter makes the line lighter in color, or more transparent. The two plots can be generated with the following commands:

```
ggplot(WeekdayCounts, aes(x = Var1, y = Freq)) + geom_line(aes(group=1), linetype=2) + xlab("Day of the Week") + ylab("Total Motor Vehicle Thefts")
```



```
ggplot(WeekdayCounts, aes(x = Var1, y = Freq)) + geom_line(aes(group=1), alpha=0.3) + xlab("Day of the Week") + ylab("Total Motor Vehicle Thefts")
```

[Final Check](#)[Save](#)[Hide Answer](#)

You have used 1 of 2 submissions



MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)
© 2014 edX, some rights reserved.

[Terms of Service and Honor Code](#) -
[Privacy Policy](#) (<https://www.edx.org/edx-privacy-policy>)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[Syllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[Schedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 3 (2/2 points)

In the previous video, we computed the following confusion matrix for our logistic regression model on our test set with a threshold of 0.5:

FALSE TRUE

0	1069	6
1	187	11

Using this confusion matrix, answer the following questions.

What is the sensitivity of our logistic regression model on the test set, using a threshold of 0.5?

Answer: 0.05555556

What is the specificity of our logistic regression model on the test set, using a threshold of 0.5?

Answer: 0.9944186

EXPLANATION

Using this confusion matrix, we can compute that the sensitivity is $11/(11+187)$ and the specificity is $1069/(1069+6)$.



[Check](#)

[Save](#)

[Hide Answer](#)

You have used 3 of 5 submissions





(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)

© 2014 edX, some rights reserved.

[Terms of Service and Honor Code](#) -
[Privacy Policy](#) (<https://www.edx.org/edx-privacy-policy>)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[yllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[chedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 3 (1/1 point)

Suppose we add another variable, Average Winter Temperature, to our model to predict wine price. Is it possible for the model's R² value to go down from 0.83 to 0.80?

- No, the model's R² value can only decrease to 0.81 by adding new variables.
- No, the model's R² value can not decrease at all by adding new variables. ✓
- Yes, the R² value could decrease to 0.80.

EXPLANATION

The model's R² value can never decrease from adding new variables to the model. This is due to the fact that it is always possible to set the coefficient for the new variable to zero in the new model. However, this would be the same as the old model. So the only reason to make the coefficient non-zero is if it improves the R² value of the model, since linear regression picks the coefficients to minimize the error terms, which is the same as maximizing the R².

[Hide Answer](#)

You have used 1 of 1 submissions



[About \(<https://www.edx.org/about-us>\)](#) [Jobs \(<https://www.edx.org/jobs>\)](#)
[Press \(<https://www.edx.org/press>\)](#) [FAQ \(<https://www.edx.org/student-faq>\)](#)
[Contact \(<https://www.edx.org/contact>\)](#)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



[\(<http://www.meetup.com/edX-Global-Community/>\)](#)



[\(<http://www.facebook.com/EdxOnline>\)](#)



[\(<https://twitter.com/edXOnline>\)](#)



[\(<https://plus.google.com/108235383044095082>\)](#)



[\(<http://youtube.com/user/edxonline>\)](#)
 © 2014 edX, some rights reserved.

[Terms of Service and Honor Code -](#)
[Privacy Policy \(<https://www.edx.org/edx-privacy-policy>\)](#)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[Syllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[Schedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 3 (1/1 point)

For each tweet, we computed an overall score by averaging all five scores assigned by the Amazon Mechanical Turk workers. However, Amazon Mechanical Turk workers might make significant mistakes when labeling a tweet. The mean could be highly affected by this.

Which of the three alternative metrics below would best capture the typical opinion of the five Amazon Mechanical Turk workers, would be less affected by mistakes, and is well-defined regardless of the five labels?

- An overall score equal to the median (middle) score ✓
- An overall score equal to the majority score
- An overall score equal to the minimum score

EXPLANATION

The correct answer is the first one - the median would capture the typical opinion of the workers and tends to be less affected by significant mistakes. The majority score might not have given a score to all tweets because they might not all have a majority score (consider a tweet with scores 0, 0, 1, 1, and 2). The minimum score does not necessarily capture the typical opinion and could be highly affected by mistakes (consider a tweet with scores -2, 1, 1, 1, 1).

[Hide Answer](#)

You have used 2 of 2 submissions



[About \(<https://www.edx.org/about-us>\)](#) [Jobs \(<https://www.edx.org/jobs>\)](#)
[Press \(<https://www.edx.org/press>\)](#) [FAQ \(<https://www.edx.org/student-faq>\)](#)
[Contact \(<https://www.edx.org/contact>\)](#)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all anywhere in the world, wherever there is Internet access. EdX's free MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community/>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)
© 2014 edX, some rights reserved.

Terms of Service and Honor Code -
[Privacy Policy \(<https://www.edx.org/edx-privacy-policy>\)](#)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[yllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[chedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 3 (2/4 points)

If the values in a vector are displayed in quotes, is the vector a character vector or a numerical vector?

- Character ✓
- Numerical
- This can't be determined from the information given.

EXPLANATION

If the values of a vector are displayed in quotes, then it must be a character vector.

If you wanted to add new observations to a data frame, which function should you use?

- data.frame ✗
- cbind
- rbind ✓

EXPLANATION

To combine two data frames with the same variable values, you should use rbind.

If you want to combine two vectors into a data frame, which function should you use?

- data.frame ✓
 - cbind
 - rbind ✗
- 

EXPLANATION

To combine two vectors into a data frame, you should use data.frame.

If you want to add a variable to your data frame, which function should you use?

- data.frame
- cbind ✓
- rbind

EXPLANATION

To add a variable to your data frame, you should use cbind.

[Hide Answer](#)

You have used 1 of 1 submissions



[About](https://www.edx.org/about-us) (<https://www.edx.org/about-us>) [Jobs](https://www.edx.org/jobs) (<https://www.edx.org/jobs>)
[Press](https://www.edx.org/press) (<https://www.edx.org/press>) [FAQ](https://www.edx.org/student-faq) (<https://www.edx.org/student-faq>)
[Contact](https://www.edx.org/contact) (<https://www.edx.org/contact>)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community/>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)
© 2014 edX, some rights reserved.

[Terms of Service and Honor Code](#) -
[Privacy Policy](#) (<https://www.edx.org/edx-privacy-policy>)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[yllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[chedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 3.1 (1/1 point)

Suppose you have a subset of 20 observations, where 14 have outcome A and 6 have outcome B. What proportion of observations have outcome A?

Answer: 0.7

EXPLANATION

The fraction of observations that have outcome A is $14/20 = 0.7$.

[Final Check](#)[Save](#)[Hide Answer](#)*You have used 2 of 3 submissions*

QUICK QUESTION 3.2 (3/3 points)

The following questions ask about the subset of 20 observations from the previous question.

If we set the threshold to 0.25 when computing predictions of outcome A, will we predict A or B for these observations?

- A ✓
- B

If we set the threshold to 0.5 when computing predictions of outcome A, will we predict A or B for these observations?

- A ✓
- B

If we set the threshold to 0.75 when computing predictions of outcome A, will we predict A or B for these observations?

- A
- B ✓

EXPLANATION

Since 70% of these observations have outcome A, we will predict A if the threshold is below 0.7, and we will predict B if the threshold is above 0.7.

[Hide Answer](#)*You have used 1 of 1 submissions*

About (<https://www.edx.org/about-us>) Jobs (<https://www.edx.org/jobs>)
Press (<https://www.edx.org/press>) FAQ (<https://www.edx.org/student-faq>)
Contact (<https://www.edx.org/contact>)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community/>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)
© 2014 edX, some rights reserved.

Terms of Service and Honor Code -
Privacy Policy (<https://www.edx.org/edx-privacy-policy>)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[Syllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[Schedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 4 (2/2 points)

In the previous video, we solved our optimization problem in LibreOffice. In your spreadsheet, change the demand for regular seats to 50 (cell D5). Then re-solve the model.

What is the new optimal objective value?

Answer: 58458

Now change the demand of regular seats to 200. What is the new optimal objective value?

Answer: 102422

EXPLANATION

For each of these questions, change the value in cell D5 to the new demand. Then select "Solver..." in the "Tools" menu, and hit solve. The problem should re-solve, and the new objective value is in the Objective cell (B8).

[Check](#)[Save](#)[Hide Answer](#)

You have used 1 of 4 submissions



About (<https://www.edx.org/about-us>) Jobs (<https://www.edx.org/jobs>)
 Press (<https://www.edx.org/press>) FAQ (<https://www.edx.org/student-faq>)
 Contact (<https://www.edx.org/contact>)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community/>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)

© 2014 edX, some rights reserved.

[Terms of Service and Honor Code](#) - [Privacy Policy](#) (<https://www.edx.org/edx-privacy-policy>)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[yllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[chedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 4 (3/3 points)

Compute the AUC of the CART model from the previous video, using the following command in your R console:

```
as.numeric(performance(pred, "auc")@y.values)
```

What is the AUC?

0.6927105

0.6927105

Answer: 0.6927105

EXPLANATION

If you run the command given above after going through the commands in Video 4, you get an AUC of 0.6927.

Now, recall that in Video 4, our tree had 7 splits. Let's see how this changes if we change the value of minbucket.

First build a CART model that is similar to the one we built in Video 4, except change the minbucket parameter to 5. Plot the tree.

How many splits does the tree have?

16

16

Answer: 16

**EXPLANATION**

You can build a CART model with minbucket=5 by using the following command:

```
StevensTree = rpart(Reverse ~ Circuit + Issue + Petitioner + Respondent + LowerCourt + Unconst, method="class", data = Train,
control=rpart.control(minbucket=5))
```

If you plot the tree with prp(StevensTree), you can see that the tree has 16 splits! This tree is probably overfit to the training data, and is not as interpretable.

Now build a CART model that is similar to the one we built in Video 4, except change the minbucket parameter to 100. Plot the tree.

How many splits does the tree have?

1

1

Answer: 1

EXPLANATION

You can build a CART model with minbucket=100 by using the following command:

```
StevensTree = rpart(Reverse ~ Circuit + Issue + Petitioner + Respondent + LowerCourt + Unconst, method="class", data = Train, control=rpart.control(minbucket=100))
```

If you plot the tree with prp(StevensTree), you can see that the tree only has one split! This tree is probably not fit well enough to the training data.

[Check](#)

[Save](#)

[Hide Answer](#)

You have used 1 of 5 submissions



[About](#) (<https://www.edx.org/about-us>) [Jobs](#) (<https://www.edx.org/jobs>)
[Press](#) (<https://www.edx.org/press>) [FAQ](#) (<https://www.edx.org/student-faq>)
[Contact](#) (<https://www.edx.org/contact>)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community/>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)
© 2014 edX, some rights reserved.

[Terms of Service and Honor Code](#) -
[Privacy Policy](#) (<https://www.edx.org/edx-privacy-policy>)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[Syllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[Schedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 4 (2/2 points)

The image below shows the penalty error matrix that we discussed in the previous video.

		Outcome				
		1	2	3	4	5
Forecast	1	0	2	4	6	8
	2	1	0	2	4	6
	3	2	1	0	2	4
	4	3	2	1	0	2
	5	4	3	2	1	0

We can interpret this matrix as follows. Suppose the actual outcome for an observation is 3, and we predict 2. We find 3 on the top of the matrix, and go down to the second row (since we forecasted 2). The penalty error for this mistake is 2. If for another observation we predict (forecast) 4, but the actual outcome is 1, that is a penalty error of 3.

What is the worst mistake we can make, according to the penalty error matrix?

- We predict 5 (very high cost), but the actual outcome is 1 (very low cost).
- We predict 1 (very low cost), but the actual outcome is 5 (very high cost). 

EXPLANATION

The highest cost is 8, which occurs when the forecast is 1 (very low cost), but the actual cost is 5 (very high cost). It would be much worst for us to ignore an actual high cost observation than accidentally predict high cost for someone who turns out to be low cost.

What are the "best" types of mistakes we can make, according to the penalty error matrix?

- Mistakes where we predict one cost bucket HIGHER than the actual outcome. 
- Mistakes where we predict one cost bucket LOWER than the actual outcome.

EXPLANATION

We are happier with mistakes where we predict one cost bucket higher than the actual outcome, since this just means we are being a little overly cautious.

[Hide Answer](#)

You have used 1 of 1 submissions



[About](https://www.edx.org/about-us) (<https://www.edx.org/about-us>) [Jobs](https://www.edx.org/jobs) (<https://www.edx.org/jobs>)
[Press](https://www.edx.org/press) (<https://www.edx.org/press>) [FAQ](https://www.edx.org/student-faq) (<https://www.edx.org/student-faq>)
[Contact](https://www.edx.org/contact) (<https://www.edx.org/contact>)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community/>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)
© 2014 edX, some rights reserved.

[Terms of Service and Honor Code](#) -

[Privacy Policy](#) (<https://www.edx.org/edx-privacy-policy>)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[yllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[chedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 4 (2/2 points)

Select the LAT of the following Jeopardy question: NICHOLAS II WAS THE LAST RULING CZAR OF THIS ROYAL FAMILY (Hint: The answer is "The Romanovs")

- NICHOLAS II
- THE LAST RULING CZAR
- THIS ROYAL FAMILY 

Select the LAT of the following Jeopardy question: REGARDING THIS DEVICE, ARCHIMEDES SAID, "GIVE ME A PLACE TO STAND ON, AND I WILL MOVE THE EARTH" (Hint: The answer is "A lever")

- THIS DEVICE 
- ARCHIMEDES
- A PLACE
- THE EARTH

EXPLANATION

The LAT in the first question is "THIS ROYAL FAMILY" and the LAT in the second question is "THIS DEVICE". Remember that if you replace the LAT with the correct answer, the sentence should make sense.

[Final Check](#)[Save](#)[Hide Answer](#)

You have used 1 of 2 submissions



[About \(<https://www.edx.org/about-us>\)](#) [Jobs \(<https://www.edx.org/jobs>\)](#)
[Press \(<https://www.edx.org/press>\)](#) [FAQ \(<https://www.edx.org/student-faq>\)](#)
[Contact \(<https://www.edx.org/contact>\)](#)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



[\(<http://www.meetup.com/edX-Global-Community/>\)](http://www.meetup.com/edX-Global-Community/)



[\(<http://www.facebook.com/EdxOnline>\)](http://www.facebook.com/EdxOnline)



[\(<https://twitter.com/edXOnline>\)](https://twitter.com/edXOnline)



[\(<https://plus.google.com/1082353830440950827>\)](https://plus.google.com/1082353830440950827)



[\(<http://youtube.com/user/edxonline>\)](http://youtube.com/user/edxonline)

© 2014 edX, some rights reserved.

[Terms of Service and Honor Code - Privacy Policy \(<https://www.edx.org/edx-privacy-policy>\)](#)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[Syllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[Schedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 4 (1/1 point)

Suppose you are the General Manager of a baseball team, and you are selecting TWO players for your team. You have a budget of \$1,500,000, and you have the choice between the following players:

Player Name	OBP	SLG	Salary
Eric Chavez	0.338	0.540	\$1,400,000
Jeremy Giambi	0.391	0.450	\$1,065,000
Frank Menechino	0.369	0.374	\$295,000
Greg Myers	0.313	0.447	\$800,000
Carlos Pena	0.361	0.500	\$300,000

Given your budget and the player statistics, which TWO players would you select?

- Eric Chavez
- Jeremy Giambi ✓
- Frank Menechino
- Greg Myers
- Carlos Pena ✓

EXPLANATION

We would select Jeremy Giambi and Carlos Pena, since they give the highest contribution to Runs Scored.

We would not select Eric Chavez, since his salary consumes entire budget, and although he has the highest SLG, there are players with better OBP.

We would not select Frank Menechino since even though he has a high OBP, his SLG is low.

We would not select Greg Myers since he is dominated by Carlos Pena in OBP and SLG, but has a much higher salary.

[Check](#)[Save](#)[Hide Answer](#)

You have used 1 of 3 submissions



MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)
© 2014 edX, some rights reserved.

[Terms of Service and Honor Code](#) -
[Privacy Policy](#) (<https://www.edx.org/edx-privacy-policy>)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[yllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[chedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 4 (1/1 point)

As we saw in the previous video, the clusters can be used to find interesting patterns of health in addition to being used to improve predictive models. By changing the number of clusters, you can find more general or more specific patterns.

If you wanted to find more unusual patterns shared by a small number of people, would you increase or decrease the number of clusters?

- Increase ✓
- Decrease

EXPLANATION

If you wanted to find more unusual patterns, you would increase the number of clusters since the clusters would become smaller and more patterns would probably emerge.

[Hide Answer](#)

You have used 1 of 1 submissions



[About \(<https://www.edx.org/about-us>\)](#) [Jobs \(<https://www.edx.org/jobs>\)](#)
[Press \(<https://www.edx.org/press>\)](#) [FAQ \(<https://www.edx.org/student-faq>\)](#)
[Contact \(<https://www.edx.org/contact>\)](#)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community/>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)
 © 2014 edX, some rights reserved.

[Terms of Service and Honor Code - Privacy Policy \(<https://www.edx.org/edx-privacy-policy>\)](#)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[yllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[chedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 4 (1/1 point)

In the previous video, we discussed a Head and Neck case with 132,878 voxels total, 9,777 voxels in the tumor, and 328 beamlets.

How many decision variables does our optimization model have?

Answer: 328**EXPLANATION**

Our decision variables are for the intensities of the beamlets. So in this case, we would have 328 decision variables, which is the number of beamlets.

[Final Check](#)[Save](#)[Hide Answer](#)*You have used 1 of 2 submissions*

About (<https://www.edx.org/about-us>) Jobs (<https://www.edx.org/jobs>)
Press (<https://www.edx.org/press>) FAQ (<https://www.edx.org/student-faq>)
Contact (<https://www.edx.org/contact>)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



[\(http://www.meetup.com/edX-Global-Community/\)](http://www.meetup.com/edX-Global-Community/)



[http://www.facebook.com/EdxOnline\)](http://www.facebook.com/EdxOnline)



[\(https://twitter.com/edXOnline\)](https://twitter.com/edXOnline)



<https://plus.google.com/108235383044095082>



<http://youtube.com/user/edxonline>
© 2014 edX, some rights reserved.

[Terms of Service and Honor Code - Privacy Policy](#) (<https://www.edx.org/edx-privacy-policy>)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[Syllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[Schedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 4 (1/1 point)

The movie "The Godfather" is in the genres action, crime, and drama, and is defined by the vector: (0,1,0,0,0,0,1,0,1,0,0,0,0,0,0,0,0,0)

The movie "Titanic" is in the genres action, drama, and romance, and is defined by the vector: (0,1,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0)

What is the distance between "The Godfather" and "Titanic", using euclidean distance?

Answer: 1.414214**EXPLANATION**

The distance between these two movies is the square root of 2. They have a difference of 1 in two genres - crime and romance.

You have used 1 of 3 submissions

[About \(<https://www.edx.org/about-us>\)](#) [Jobs \(<https://www.edx.org/jobs>\)](#)
[Press \(<https://www.edx.org/press>\)](#) [FAQ \(<https://www.edx.org/student-faq>\)](#)
[Contact \(<https://www.edx.org/contact>\)](#)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community/>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)
© 2014 edX, some rights reserved.

[Terms of Service and Honor Code](#) -

[Privacy Policy](#) (<https://www.edx.org/edx-privacy-policy>)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[Syllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[Schedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 4 (1/1 point)

Suppose we want to add a constraint that teams A and B must play in week 4 (we want the last game to be a divisional one). Given the current model, which of the following constraints would model this correctly? Select all that apply.

- $x_{AB4} = 0$
- $x_{AB4} = 1$ ✓
- $x_{AB1} + x_{AB2} + x_{AB3} = 1$ ✓
- $x_{AB2} = 1$
- $x_{AB1} + x_{AB2} = 1$

EXPLANATION

The second and third constraints would both model this correctly. We can either force the decision variable for week 4 to 1, or we can make sure that only one game is played in the earlier weeks.

[Check](#)[Save](#)[Hide Answer](#)

You have used 1 of 3 submissions



[About \(<https://www.edx.org/about-us>\)](#) [Jobs \(<https://www.edx.org/jobs>\)](#)
[Press \(<https://www.edx.org/press>\)](#) [FAQ \(<https://www.edx.org/student-faq>\)](#)
[Contact \(<https://www.edx.org/contact>\)](#)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



[\(<http://www.meetup.com/edX-Global-Community/>\)](http://www.meetup.com/edX-Global-Community/)



[\(<http://www.facebook.com/EdxOnline>\)](http://www.facebook.com/EdxOnline)



[\(<https://twitter.com/edXOnline>\)](https://twitter.com/edXOnline)



[\(<https://plus.google.com/108235383044095082>\)](https://plus.google.com/108235383044095082)



[\(<http://youtube.com/user/edxonline>\)](http://youtube.com/user/edxonline)
© 2014 edX, some rights reserved.

[Terms of Service and Honor Code - Privacy Policy \(<https://www.edx.org/edx-privacy-policy>\)](#)



[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[yllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[chedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 4 (1/1 point)

For which of the following models should external validation be used? Consider both the population used to train the model, and the population that the model will be used on.

- A model to predict obesity risk. Data from a random sample of California residents was used to build the model, and we want to use the model to predict the obesity risk of all United States residents. ✓
- A model to predict the stress of MIT students. Data from a random sample of MIT students was used to build the model, and we want to use the model to predict the stress level of all MIT students.
- A model to predict the probability of a runner winning a marathon. Data from all runners in the Boston Marathon was used to build the model, and we want use the model to predict the probability of winning for all people who run marathons. ✓

EXPLANATION

In the first and third models, we are using a special sub-population to build the model. While we can use the model for that sub-population, we should use external validation to test the model on other populations. The second model uses data from a special sub-population, but the model is only intended for that sub-population, so external validation is not necessary.

[Final Check](#)[Save](#)[Hide Answer](#)

You have used 1 of 2 submissions



[About \(<https://www.edx.org/about-us>\)](#) [Jobs \(<https://www.edx.org/jobs>\)](#)
[Press \(<https://www.edx.org/press>\)](#) [FAQ \(<https://www.edx.org/student-faq>\)](#)
[Contact \(<https://www.edx.org/contact>\)](#)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



[\(http://www.meetup.com/edX-Global-Community/\)](http://www.meetup.com/edX-Global-Community/)



[\(http://www.facebook.com/EdxOnline\)](http://www.facebook.com/EdxOnline)



[\(https://twitter.com/edXOnline\)](https://twitter.com/edXOnline)



[\(https://plus.google.com/1082353830440950827\)](https://plus.google.com/1082353830440950827)



[\(http://youtube.com/user/edxonline\)](http://youtube.com/user/edxonline)
© 2014 edX, some rights reserved.

[Terms of Service and Honor Code -](#)

[Privacy Policy \(<https://www.edx.org/edx-privacy-policy>\)](#)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[yllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[chedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 4 (3/3 points)

In R, use the dataset [wine.csv](#) ([/c4x/MITx/15.071x/asset/wine.csv](#)) to create a linear regression model to predict Price using HarvestRain and WinterRain as independent variables. Using the summary output of this model, answer the following questions:

What is the "Multiple R-squared" value of your model?

Answer: 0.3177

What is the coefficient for HarvestRain?

Answer: -4.971e-03

What is the intercept coefficient?

Answer: 7.865

EXPLANATION

In R, create the model by typing the following line into your R console:

```
modelQQ4 = lm(Price ~ HarvestRain + WinterRain, data=wine)
```

Then, look at the output of `summary(modelQQ4)`. The Multiple R-squared is listed at the bottom of the output, and the coefficients can be found in the coefficients table.

[Check](#)[Save](#)[Hide Answer](#)

You have used 1 of 5 submissions





EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)

© 2014 edX, some rights reserved.

[Terms of Service and Honor Code](#) -

[Privacy Policy](#) (<https://www.edx.org/edx-privacy-policy>)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[yllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[chedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 4 (3/3 points)

For each of the following questions, pick the preprocessing task that we discussed in the previous video that would change the sentence "Data is useful AND powerful!" to the new sentence listed in the question.

New sentence: Data useful powerful!

- Cleaning up irregularities (changing to lowercase and removing punctuation)
- Removing stop words ✓
- Stemming

New sentence: data is useful and powerful

- Cleaning up irregularities (changing to lowercase and removing punctuation) ✓
- Removing stop words
- Stemming

New sentence: Data is use AND power!

- Cleaning up irregularities (changing to lowercase and removing punctuation)
- Removing stop words
- Stemming ✓

EXPLANATION

The first new sentence has the stop words "is" and "and" removed. The second new sentence has the irregularities removed (no capital letters or punctuation). The third new sentence has the words stemmed - the "ful" is removed from "useful" and "powerful".

[Final Check](#)[Save](#)[Hide Answer](#)

You have used 1 of 2 submissions





(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)

© 2014 edX, some rights reserved.

[Terms of Service and Honor Code](#) -
[Privacy Policy](#) (<https://www.edx.org/edx-privacy-policy>)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[yllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[chedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 4 (1/1 point)

In R, change the shape of your points to the number 15. What shape are the points now?

- Circles
- Diamonds
- Crosses
- Squares ✓
- Stars

EXPLANATION

If you type:

```
scatterplot + geom_point(shape = 15)
```

where scatterplot is the plot we created in the previous video, you can see that the points are squares.

[Final Check](#)[Save](#)[Hide Answer](#)
You have used 1 of 2 submissions


[About \(<https://www.edx.org/about-us>\)](#) [Jobs \(<https://www.edx.org/jobs>\)](#)
[Press \(<https://www.edx.org/press>\)](#) [FAQ \(<https://www.edx.org/student-faq>\)](#)
[Contact \(<https://www.edx.org/contact>\)](#)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community/>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)
© 2014 edX, some rights reserved.

[Terms of Service and Honor Code - Privacy Policy \(<https://www.edx.org/edx-privacy-policy>\)](#)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[yllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[chedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 4 (4/4 points)

If you want to know how many observations are in your data frame, which function should you use?

- str ✓
- summary
- I can't figure out this information using either of these functions.

EXPLANATION

If you use the `str` function (in the video, we typed `str(WHO)` in our R console) the first line tells you how many observations are in your data frame.

If you want to know the mean value of a numerical variable, which function should you use?

- str
- summary ✓
- I can't figure out this information using either of these functions.

EXPLANATION

If you use the `summary` function (in the video, we typed `summary(WHO)` in our R console) you can see a statistical summary of each variable. For numerical variables, the mean value is listed.

If you want to know the standard deviation of a numerical variable, which function should you use?

- str
- summary
- I can't figure out this information using either of these functions. ✓

EXPLANATION

Neither the `str` function nor the `summary` function provides the standard deviation value of a variable. We'll see how to compute this value in the next video.

In the previous video, we actually used a variable name that is NOT recommended! What was the name of this variable? Type it exactly how we did in R.

Answer: WHO_Europe

EXPLANATION

In the previous video, we called our subset "WHO_Europe". This is actually not a recommended variable name because of the underscore!

[Final Check](#)[Save](#)[Hide Answer](#)

You have used 1 of 2 submissions



About (<https://www.edx.org/about-us>) Jobs (<https://www.edx.org/jobs>)
Press (<https://www.edx.org/press>) FAQ (<https://www.edx.org/student-faq>)
Contact (<https://www.edx.org/contact>)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community/>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)

© 2014 edX, some rights reserved.

[Terms of Service and Honor Code](#) -
[Privacy Policy](#) (<https://www.edx.org/edx-privacy-policy>)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[yllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[chedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 4.1 (1/1 point)

In R, create a logistic regression model to predict "PoorCare" using the independent variables "StartedOnCombination" and "ProviderCount". Use the training set we created in the previous video to build the model.

If you haven't already loaded and split the data in R, please run these commands in your R console to load and split the data set. Remember to first navigate to the directory where you have saved "quality.csv".

```
quality = read.csv("quality.csv")  
  
install.packages("caTools")  
  
library(caTools)  
  
set.seed(88)  
  
split = sample.split(quality$PoorCare, SplitRatio = 0.75)  
  
qualityTrain = subset(quality, split == TRUE)  
  
qualityTest = subset(quality, split == FALSE)
```

Then recall that we built a logistic regression model to predict PoorCare using the R command:

```
QualityLog = glm(PoorCare ~ OfficeVisits + Narcotics, data=qualityTrain, family=binomial)
```

You will need to adjust this command to answer this question, and then look at the summary(QualityLog) output.

What is the coefficient for "StartedOnCombination"?



Answer: 1.95230

EXPLANATION

To construct this model in R, use the command:

```
Model = glm(PoorCare ~ StartedOnCombination + ProviderCount, data=qualityTrain, family=binomial)
```

If you look at the output of summary(Model), the value of the coefficient (Estimate) for StartedOnCombination is 1.95230.

You have used 1 of 5 submissions

QUICK QUESTION 4.2 (1/1 point)

StartedOnCombination is a binary variable, which equals 1 if the patient is started on a combination of drugs to treat their diabetes, and equals 0 if the patient is not started on a combination of drugs. All else being equal, does this model imply that starting a patient on a combination of drugs is indicative of poor care, or good care?

- Poor Care ✓
 Good Care

EXPLANATION

The coefficient value is positive, meaning that positive values of the variable make the outcome of 1 more likely. This corresponds to Poor Care.

[Hide Answer](#)

You have used 1 of 1 submissions



About (<https://www.edx.org/about-us>) Jobs (<https://www.edx.org/jobs>)
Press (<https://www.edx.org/press>) FAQ (<https://www.edx.org/student-faq>)
Contact (<https://www.edx.org/contact>)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community/>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)

© 2014 edX, some rights reserved.

[Terms of Service and Honor Code](#) -
[Privacy Policy](#) (<https://www.edx.org/edx-privacy-policy>)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[yllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[chedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 5 (2/2 points)

Using the visualization we created in the previous video, answer the following questions:

Suppose that our demand for regular seats remains the same (100) but our demand for discount seats goes down to 100. Will our optimal solution change?

- Yes
- No ✓
- I can't answer this question using the visualization.

Now suppose that our demand for regular seats remains the same (100) but our demand for discount seats goes down to 50. Will our optimal solution change?

- Yes ✓
- No
- I can't answer this question using the visualization.

EXPLANATION

In the first case, our optimal solution will not change because we are only offering 66 discount seats. So even if the demand goes down to 100, we are not meeting the demand. But in the second case, we can only offer 50 discount seats. So our airplane will not be full, and our optimal solution will change to 100 regular seats and 50 discount seats.

[Hide Answer](#)

You have used 1 of 1 submissions



[About \(<https://www.edx.org/about-us>\)](#) [Jobs \(<https://www.edx.org/jobs>\)](#)
[Press \(<https://www.edx.org/press>\)](#) [FAQ \(<https://www.edx.org/student-faq>\)](#)
[Contact \(<https://www.edx.org/contact>\)](#)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community/>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)
 © 2014 edX, some rights reserved.

[Terms of Service](#) and [Honor Code](#) -

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[yllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[chedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 5 (2/2 points)

IMPORTANT NOTE: When creating random forest models, you might still get different answers from the ones you see here even if you set the random seed. This has to do with different operating systems and the random forest implementation.

Let's see what happens if we set the seed to two different values and create two different random forest models.

First, set the seed to 100, and re-run the random forest model, exactly like we did in the video. What is the accuracy of the model?

Answer: 0.6882353

Now, set the seed to 200, and re-run the random forest model, exactly like we did in the video. What is the accuracy of this model?

Answer: 0.7058824

EXPLANATION

You can create the models and compute the accuracies with the following commands in R:

```
set.seed(100)
```

```
StevensForest = randomForest(Reverse ~ Circuit + Issue + Petitioner + Respondent + LowerCourt + Unconst, data = Train,  
ntree=200, nodesize=25 )
```



```
PredictForest = predict(StevensForest, newdata = Test)
```

```
table(Test$Reverse, PredictForest)
```

and then repeat it, but with `set.seed(200)` first.

As we see here, the random component of the random forest method can change the accuracy. The accuracy for a more stable dataset will not change very much, but a noisy dataset can be significantly affected by the random samples.

You have used 1 of 4 submissions



[About](https://www.edx.org/about-us) (<https://www.edx.org/about-us>) [Jobs](https://www.edx.org/jobs) (<https://www.edx.org/jobs>)
[Press](https://www.edx.org/press) (<https://www.edx.org/press>) [FAQ](https://www.edx.org/student-faq) (<https://www.edx.org/student-faq>)
[Contact](https://www.edx.org/contact) (<https://www.edx.org/contact>)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)
© 2014 edX, some rights reserved.

[Terms of Service and Honor Code](#) -
[Privacy Policy](#) (<https://www.edx.org/edx-privacy-policy>)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[yllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[chedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 5 (1/1 point)

What were the most important factors in the CART trees to predict cost?

- Cost ranges from the previous year ✓
- Risk factors
- Chronic conditions
- Number of office visits last year

EXPLANATION

The most important variables in a CART tree are at the top of the tree - in this case, they are the cost ranges from the previous year.

[Final Check](#)[Save](#)[Hide Answer](#)*You have used 1 of 2 submissions*

[About \(<https://www.edx.org/about-us>\)](#) [Jobs \(<https://www.edx.org/jobs>\)](#)
[Press \(<https://www.edx.org/press>\)](#) [FAQ \(<https://www.edx.org/stu-faq>\)](#)
[Contact \(<https://www.edx.org/contact>\)](#)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community/>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)
© 2014 edX, some rights reserved.

[Terms of Service and Honor Code](#) -
[Privacy Policy \(<https://www.edx.org/edx-privacy-policy>\)](#)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[Syllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[Schedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 5 (1/1 point)

To predict which candidate answer is correct, we said that Watson uses logistic regression. Which of the other techniques that we have learned could be used instead?

- Linear Regression
- CART ✓
- Random Forests ✓

EXPLANATION

CART and Random Forests are both techniques that are also used for classification, and could provide confidence probabilities.

[Final Check](#)[Save](#)[Hide Answer](#)

You have used 1 of 2 submissions



[About \(<https://www.edx.org/about-us>\)](#) [Jobs \(<https://www.edx.org/jobs>\)](#)
[Press \(<https://www.edx.org/press>\)](#) [FAQ \(<https://www.edx.org/student-faq>\)](#)
[Contact \(<https://www.edx.org/contact>\)](#)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community/>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/1082353830440950827>)



(<http://youtube.com/user/edxonline>)
© 2014 edX, some rights reserved.

[Terms of Service and Honor Code - Privacy Policy](#) (<https://www.edx.org/edx-privacy-policy>)



[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[Syllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[Schedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 5 (1/2 points)

In your spreadsheet from Video 3, make sure that you have solved the original small example problem (change the spinal cord limit back to 5 and re-solve if you have changed it, and make sure the objective value is 22.75).

Now, change the weight for the spinal cord term in the objective to 5.

Without re-solving, what does the objective value of the current solution change to?

Answer: 42.75

EXPLANATION

The term `SUMPRODUCT(B14:B19;F5:F10)` in the objective (corresponding to Voxel 5) should now be `5*SUMPRODUCT(B14:B19;F5:F10)`. This changes the objective value to 42.75.

Now re-solve the model. What does the objective change to?

Answer: 25.6666667

EXPLANATION

You can resolve the model by going to Solver, and just hitting Solve. The new optimal objective function value is 25.666667.

Notice how we are now giving a smaller dose to the spinal cord!



[Hide Answer](#)

You have used 4 of 4 submissions



MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)
© 2014 edX, some rights reserved.

[Terms of Service and Honor Code](#) -
[Privacy Policy](#) (<https://www.edx.org/edx-privacy-policy>)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[yllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[chedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 5 (2/2 points)

Suppose you are running the Hierarchical clustering algorithm with 212 observations.

How many clusters will there be at the start of the algorithm?

Answer: 212

How many clusters will there be at the end of the algorithm?

EXPLANATION

The Hierarchical clustering algorithm always starts with each data point in its own cluster, and ends with all data points in the same cluster. So there will be 212 clusters at the beginning of the algorithm, and 1 cluster at the end of the algorithm.

You have used 1 of 3 submissions

[About \(<https://www.edx.org/about-us>\)](#) [Jobs \(<https://www.edx.org/jobs>\)](#)
[Press \(<https://www.edx.org/press>\)](#) [FAQ \(<https://www.edx.org/student-faq>\)](#)
[Contact \(<https://www.edx.org/contact>\)](#)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



[\(<http://www.meetup.com/edX-Global-Community/>\)](http://www.meetup.com/edX-Global-Community/)



[\(<http://www.facebook.com/EdxOnline>\)](http://www.facebook.com/EdxOnline)



[\(<http://twitter.com/edXOnline>\)](http://twitter.com/edXOnline)



[\(<https://plus.google.com/108235383044095082>\)](https://plus.google.com/108235383044095082)



[\(<http://youtube.com/user/edxonline>\)](http://youtube.com/user/edxonline)

© 2014 edX, some rights reserved.

[Terms of Service and Honor Code - Privacy Policy \(<https://www.edx.org/edx-privacy-policy>\)](#)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[Syllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[Schedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

QUICK QUESTION 5 (1/1 point)

In the previous video, our heatmap was plotting squares out in the water, which seems a little strange. We can fix this by removing the observations from our data frame that have Freq = 0.

Take a subset of LatLonCounts, only keeping the observations for which Freq > 0, and call it LatLonCounts2.

Redo the heatmap from the end of Video 5, using LatLonCounts2 instead of LatLonCounts. You should no longer see any squares out in the water, or in any areas where there were no motor vehicle thefts.

How many observations did we remove?

952

952

Answer: 952

EXPLANATION

You can take a subset of LatLonCounts, only keeping the observations for which Freq > 0 with the following command:

```
LatLonCounts2 = subset(LatLonCounts, Freq > 0)
```

Then, you can generate the new heatmap with the following command:

```
ggmap(chicago) + geom_tile(data=LatLonCounts2, aes(x = Long, y = Lat, alpha=Freq), fill="red")
```

The number of observations in LatLonCounts2 is 686, and the number of observations in LatLonCounts is 1638. These numbers can be found by using nrow or str.

[Check](#)

[Save](#)

[Hide Answer](#)

You have used 1 of 3 submissions





(<http://youtube.com/user/edxonline>)
© 2014 edX, some rights reserved.

[Terms of Service and Honor Code](#) -
[Privacy Policy](#) (<https://www.edx.org/edx-privacy-policy>)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[yllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[chedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 5 (1/1 point)

In Video 3, we built a logistic regression model and found that the following variables were significant (or almost significant) for predicting ten year risk of CHD: male, age, number of cigarettes per day, whether or not the patient previously had a stroke, whether or not the patient is currently hypertensive, total cholesterol level, systolic blood pressure, and blood glucose level. Which **one** of the following variables would be the most dramatically affected by a behavioral intervention? HINT: Think about how much control the patient has over each of the variables.

- Male
- Age
- Number of Cigarettes per day ✓
- Previously had a Stroke
- Hypertensive
- Total Cholesterol Level
- Systolic Blood Pressure
- Blood Glucose Level

EXPLANATION

The number of cigarettes smoked per day would be the most dramatically affected by a behavioral intervention. This is a variable that the patient has the ability to control the most.

[Check](#)[Save](#)[Hide Answer](#)

You have used 1 of 3 submissions



[About \(<https://www.edx.org/about-us>\)](#) [Jobs \(<https://www.edx.org/jobs>\)](#)
[Press \(<https://www.edx.org/press>\)](#) [FAQ \(<https://www.edx.org/student-faq>\)](#)
[Contact \(<https://www.edx.org/contact>\)](#)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



[\(<http://www.meetup.com/edX-Global-Community/>\)](#)



[\(<http://www.facebook.com/EdxOnline>\)](#)



[\(<https://twitter.com/edxonline>\)](#)



[\(<https://plus.google.com/108235383044095082>\)](#)



[\(<http://youtube.com/user/edxonline>\)](#)

© 2014 edX, some rights reserved.

[Terms of Service and Honor Code -](#)

[Privacy Policy \(<https://www.edx.org/edx-privacy-policy>\)](#)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[yllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[chedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 5 (3/3 points)

Using the data set [wine.csv](#) (/c4x/MITx/15.071x/asset/wine.csv), what is the correlation between HarvestRain and WinterRain?

-0.2754409

-0.2754409

Answer: -0.2754409

EXPLANATION

First load the dataset into R and call it "wine":

```
> wine = read.csv("wine.csv")
```

Then you can compute the correlation between HarvestRain and WinterRain by typing the following command into your R console:

```
> cor(wine$HarvestRain, wine$WinterRain)
```

Now, use the dataset wine.csv to create a linear regression model to predict Price using HarvestRain and WinterRain as independent variables, like you did in the previous quick question. Using the summary output of this model, answer the following questions:

Is the coefficient for HarvestRain significant?

- Yes ✓
- No
- I can't answer this question using the summary output.



Is the coefficient for WinterRain significant?

- Yes
- No ✓
- I can't answer this question using the summary output.

EXPLANATION

You can create the model and look at the summary output with the following command:

```
model = lm(Price ~ WinterRain + HarvestRain, data=wine)
```

```
summary(model)
```

From the summary output, you can see that HarvestRain is significant (two stars), but WinterRain is not (no stars).

[Final Check](#)[Save](#)[Hide Answer](#)

You have used 1 of 2 submissions



About (<https://www.edx.org/about-us>) Jobs (<https://www.edx.org/jobs>)
Press (<https://www.edx.org/press>) FAQ (<https://www.edx.org/student-faq>)
Contact (<https://www.edx.org/contact>)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community/>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)
© 2014 edX, some rights reserved.

Terms of Service and Honor Code -
Privacy Policy (<https://www.edx.org/edx-privacy-policy>)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[Syllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[Schedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 5 (2/2 points)

Given a corpus in R, how many commands do you need to run in R to clean up the irregularities (removing capital letters and punctuation)?

2

2

Answer: 2

How many commands do you need to run to stem the document?

1

1

Answer: 1**EXPLANATION**

In R, you can clean up the irregularities with two lines:

```
corpus = tm_map(corpus, tolower)
```

```
corpus = tm_map(corpus, removePunctuation)
```

And you can stem the document with one line:

```
corpus = tm_map(corpus, stemDocument)
```

[Check](#)[Save](#)[Hide Answer](#)

You have used 1 of 3 submissions





(<http://youtube.com/user/edxonline>)
© 2014 edX, some rights reserved.

[Terms of Service and Honor Code](#) -
[Privacy Policy](#) (<https://www.edx.org/edx-privacy-policy>)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[Syllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[Schedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

QUICK QUESTION 5 (1/1 point)

Create the fertility rate versus population under 15 plot again:

```
ggplot(WHO, aes(x = FertilityRate, y = Under15)) + geom_point()
```

Now, color the points by the Region variable. You can add `scale_color_brewer(palette="Dark2")` to your plot if you are having a hard time distinguishing the colors (this color palette is often better if you are colorblind). To use this option, your plot command would be the following:

```
ggplot(WHO, aes(x = FertilityRate, y = Under15)) + geom_point() + scale_color_brewer(palette="Dark2")
```

To find out more about using ggplot in a colorblind-friendly way, please see [this website \(http://bconnelly.net/2013/10/creating-colorblind-friendly-figures/\)](http://bconnelly.net/2013/10/creating-colorblind-friendly-figures/).

One region in particular has a lot of countries with a very low fertility rate and a very low percentage of the population under 15. Which region is it?

- Africa
- Americas
- Eastern Mediterranean
- Europe ✓
- South-East Asia
- Western Pacific

EXPLANATION

You can color the points by region if you adjust the command to the following:

```
ggplot(WHO, aes(x = FertilityRate, y = Under15, color=Region)) + geom_point()
```

Most of the countries in Europe have a very low fertility rate and a very low percentage of the population under 15.

[Final Check](#)[Save](#)[Hide Answer](#)

You have used 1 of 2 submissions



MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)
© 2014 edX, some rights reserved.

[Terms of Service and Honor Code](#) -
[Privacy Policy](#) (<https://www.edx.org/edx-privacy-policy>)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[Syllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[Schedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

QUICK QUESTION 5 (3/3 points)

What is the mean value of the "Over60" variable?

11.16

11.16

Answer: 11.16

EXPLANATION

You can compute this value by either typing `mean(WHO$Over60)` in your R console, or by typing `summary(WHO$Over60)` in your R console. The output is 11.16.

Which country has the smallest percentage of the population over 60?

- Japan
- United Arab Emirates (UAE) ✓
- Sierra Leone
- Cuba
- Luxembourg
- Mali

Help

EXPLANATION

To get this value, you should type `which.min(WHO$Over60)` in your R console. The output is 183. Then, to see the name of the 183rd country in your data frame, type `WHO$Country[183]` in your R console. The output is United Arab Emirates.

Which country has the largest literacy rate?

- Japan
- United Arab Emirates (UAE)
- Sierra Leone
- Cuba ✓
- Luxembourg
- Mali

EXPLANATION

To get this value, you should type `which.max(WHO$LiteracyRate)` in your R console. The output is 44. Then, to see the name of the 44th country in your data frame, type `WHO$Country[44]` in your R console. The output is Cuba.

[Check](#)

[Save](#)

[Hide Answer](#)

You have used 1 of 3 submissions



About (<https://www.edx.org/about-us>) Jobs (<https://www.edx.org/jobs>)
Press (<https://www.edx.org/press>) FAQ (<https://www.edx.org/students/faq>)
Contact (<https://www.edx.org/contact>)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community/>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/1082353830440950821>)



(<http://youtube.com/user/edxonline>)
© 2014 edX, some rights reserved.

Terms of Service and Honor Code -
Privacy Policy (<https://www.edx.org/edx-privacy-policy>)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[Syllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[Schedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 6 (1/1 point)

In your spreadsheet, change the capacity to 250 in the capacity constraint, the regular demand to 150, and the discount demand to 150. Then re-solve the model.

What is the objective value of the optimal solution?

116350

116350

Answer: 116350**EXPLANATION**

You can change the values in the capacity constraint RHS (cell D11), the regular demand (cell D5), and the discount demand (cell D6) and resolve the model by selecting "Solver..." in the "Tools" menu. After it finishes solving, the objective value can be found in the blue cell (B8)

[Check](#)[Save](#)[Hide Answer](#)

You have used 1 of 3 submissions



[About \(<https://www.edx.org/about-us>\)](#) [Jobs \(<https://www.edx.org/jobs>\)](#)
[Press \(<https://www.edx.org/press>\)](#) [FAQ \(<https://www.edx.org/student-faq>\)](#)
[Contact \(<https://www.edx.org/contact>\)](#)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



[\(http://www.meetup.com/edX-Global-Community/\)](http://www.meetup.com/edX-Global-Community/)



[\(<http://www.facebook.com/EdxOnline>\)](http://www.facebook.com/EdxOnline)



[\(<http://twitter.com/edXOnline>\)](http://twitter.com/edXOnline)



[\(<https://plus.google.com/108235383044095082>\)](https://plus.google.com/108235383044095082)



[\(<http://youtube.com/user/edxonline>\)](http://youtube.com/user/edxonline)

© 2014 edX, some rights reserved.

[Terms of Service and Honor Code - Privacy Policy \(<https://www.edx.org/edx-privacy-policy>\)](#)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[Syllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[Schedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 6 (1/1 point)

Plot the tree that we created using cross-validation. How many splits does it have?

Answer: 1**EXPLANATION**

If you follow the R commands from the previous video, you can plot the tree with prp(StevensTreeCV).

The tree with the best accuracy only has one split! When we were picking different minbucket parameters before, it seemed like this tree was probably not doing a good job of fitting the data. However, this tree with one split gives us the best out-of-sample accuracy. This reminds us that sometimes the simplest models are the best!

[Final Check](#)[Save](#)[Hide Answer](#)

You have used 2 of 3 submissions

[About \(<https://www.edx.org/about-us>\)](#) [Jobs \(<https://www.edx.org/jobs>\)](#)
[Press \(<https://www.edx.org/press>\)](#) [FAQ \(<https://www.edx.org/student-faq>\)](#)
[Contact \(<https://www.edx.org/contact>\)](#)

EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.

[\(<http://www.meetup.com/edX-Global-Community/>\)](http://www.meetup.com/edX-Global-Community/)[\(<http://www.facebook.com/EdxOnline>\)](http://www.facebook.com/EdxOnline)[\(<http://twitter.com/edXOnline>\)](http://twitter.com/edXOnline)[\(<http://plus.google.com/108235383044095082>\)](http://plus.google.com/108235383044095082)[\(<http://youtube.com/user/edxonline>\)
© 2014 edX, some rights reserved.](http://youtube.com/user/edxonline)Terms of Service and Honor Code -
Privacy Policy (<https://www.edx.org/edx-privacy-policy>)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[yllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[chedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 6 (1/1 point)

Which of the following is MOST LIKELY to be a topic of Sabermetric research?

- Evaluating how the attitude of managers influences player performance
- Determining the correlation between scouting predictions and player performance
- Predicting how many home runs the Oakland A's will hit next year ✓

EXPLANATION

Sabermetric research tries to take a quantitative approach to baseball. Predicting how many home runs the Oakland A's will hit next year is a very quantitative problem. While the other two topics could be an area of Sabermetric research, they are more qualitative.

While Moneyball made the use of analytics in sports very popular, baseball is not the only sport for which analytics is used. Analytics is currently used in almost every single sport, including basketball, soccer, cricket, and hockey.

Basketball: The study of analytics in basketball, called APBRmetrics, is very popular. There have been many books written in this area, including "Pro Basketball Forecast" by John Hollinger and "Basketball on Paper" by Dean Oliver. There are also several websites dedicated to the study of basketball analytics, including 82games.com. We'll talk more about basketball during recitation.

Soccer: The soccer analytics community is currently growing, and new data is constantly being collected. Many argue that it is much harder to apply analytics to soccer, but there are several books and websites on the topic. Check out "The Numbers Game: Why Everything You Know about Football is Wrong" by Chris Anderson and David Sally, as well as the websites socceranalysts.com and soccermetrics.net.

Cricket: There are several websites dedicated to building models for evaluating player performance in cricket. Check out cricmetric.com and www.impactindexcricket.com.

Hockey: Analytics are used in hockey to track player performance and to better shape the composition of teams. Check out the websites hockeyanalytics.com and lighthousehockey.com.

[Final Check](#)[Save](#)[Hide Answer](#)

You have used 1 of 2 submissions





(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)

© 2014 edX, some rights reserved.

[Terms of Service and Honor Code](#) -
[Privacy Policy](#) (<https://www.edx.org/edx-privacy-policy>)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[yllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[chedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 6 (3/3 points)

Using the table function in R, please answer the following questions about the dataset "movies".

How many movies are classified as comedies?

Answer: 502

How many movies are classified as westerns?

Answer: 27

How many movies are classified as romance AND drama?

Answer: 97

EXPLANATION

You can answer these questions by using the following commands:



```
table(movies$Comedy)
```

```
table(movies$Western)
```

```
table(movies$Romance, movies$Drama)
```

You have used 1 of 3 submissions





EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)

© 2014 edX, some rights reserved.

[Terms of Service and Honor Code](#) -

[Privacy Policy](#) (<https://www.edx.org/edx-privacy-policy>)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[yllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[chedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 6 (1/1 point)

Redo the map from Video 6, but this time fill each state with the variable GunOwnership. This shows the percentage of people in each state who own a gun.

Which of the following states has the highest gun ownership rate? To see the state labels, take a look at the World Atlas map [here](http://www.worldatlas.com/webimage/testmaps/usenames.htm) (<http://www.worldatlas.com/webimage/testmaps/usenames.htm>).

- California
- Montana ✓
- Texas
- Louisiana
- Missouri

EXPLANATION

You can generate the gun ownership plot using the following command:

```
ggplot(murderMap, aes(x = long, y = lat, group=group, fill = GunOwnership)) + geom_polygon(color="black") +
  scale_fill_gradient(low = "black", high = "red", guide="legend")
```

Of these five states, the one that is the most red is Montana.

[Final Check](#)[Save](#)[Hide Answer](#)

You have used 1 of 2 submissions



[About \(<https://www.edx.org/about-us>\)](#) [Jobs \(<https://www.edx.org/jobs>\)](#)
[Press \(<https://www.edx.org/press>\)](#) [FAQ \(<https://www.edx.org/student-faq>\)](#)
[Contact \(<https://www.edx.org/contact>\)](#)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



[\(http://www.meetup.com/edX-Global-Community/\)](#)



[\(http://www.facebook.com/EdxOnline\)](#)



[\(https://twitter.com/edXOnline\)](#)



[\(https://plus.google.com/108235383044095082\)](#)



[\(http://youtube.com/user/edxonline\)](#)
 © 2014 edX, some rights reserved.

[Terms of Service and Honor Code - Privacy Policy \(<https://www.edx.org/edx-privacy-policy>\)](#)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[yllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[chedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 6 (1 point possible)Which of the following are NOT valid values for an out-of-sample (test set) R^2 ?

- 7.0
- 0.3
- 0.0
- 0.6
- 1.0
- 2.4 ✓

EXPLANATIONThe formula for R^2 is

$$R^2 = 1 - \frac{SSE}{SST},$$

where SST is calculated using the average value of the dependent variable on the training set.

Since SSE and SST are the sums of squared terms, we know that both will be positive. Thus SSE/SST must be greater than or equal to zero. This means it is not possible to have an out-of-sample R^2 value of 2.4.However, all other values are valid (even the negative ones!), since SSE can be more or less than SST, due to the fact that this is an out-of-sample R^2 , not a model R^2 .[Hide Answer](#)

You have used 2 of 2 submissions



About (<https://www.edx.org/about-us>) Jobs (<https://www.edx.org/jobs>)
 Press (<https://www.edx.org/press>) FAQ (<https://www.edx.org/student-faq>)
 Contact (<https://www.edx.org/contact>)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community/>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)
 © 2014 edX, some rights reserved.

[Terms of Service and Honor Code](#) -

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[yllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[chedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 6 (1/1 point)

In the previous video, we showed a list of all words that appear at least 20 times in our tweets. Which of the following words appear at least 100 times? (HINT: use the findFreqTerms function)

- app
- buy
- freak
- ipad
- iphon ✓
- itun ✓
- like
- love
- new ✓
- think

EXPLANATION

To answer this question, you need to run the following command in R:

```
findFreqTerms(frequencies, lowfreq=100)
```

This outputs the words that appear at least 100 times in our tweets. They are "iphon", "itun", and "new".

[Check](#)[Save](#)[Hide Answer](#)

You have used 1 of 4 subn

ns



About (<https://www.edx.org/about-us>) Jobs (<https://www.edx.org/jobs>)
 Press (<https://www.edx.org/press>) FAQ (<https://www.edx.org/student-faq>)
 Contact (<https://www.edx.org/contact>)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community/>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)

© 2014 edX, some rights reserved.

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[yllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[chedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 7 (1/2 points)

In this quick question, we'll perform some sensitivity analysis on the connecting flights problem.

Previously, we said that American Airlines could market their fares to increase demand. It costs \$200 in advertising to increase demand by one unit.

Is it worth it to market the discount fares from JFK to DFW?

- Yes. American Airlines should market the discount fares from JFK to DFW to increase demand by 50.
- Yes. American Airlines should market the discount fares from JFK to DFW to increase demand by 10.
- No. American Airlines should not market the discount fares from JFK to DFW because even though the revenue increases, it does not exceed the costs.
- No. American Airlines should not market the discount fares from JFK to DFW because the revenue does not increase at all by increasing the demand for these tickets. ✓

EXPLANATION

You can answer this question without re-solving the model by noticing that we are not meeting the demand for discount fares from JFK to DFW at all. The demand could increase by 100, and we still would not offer more than 11 discount fares.

Alternatively, you could change the demand for discount fares, and re-solve the model. The solution does not change, regardless of how much you increase the demand.

Is it worth it to market the regular fares from JFK to LAX?

- Yes. American Airlines should market the regular fares from JFK to LAX to increase demand by 50.
- Yes. American Airlines should market the regular fares from JFK to LAX to increase demand by 10. ✗
- No. American Airlines should not market the regular fares from JFK to LAX because even though the revenue increases, it does not exceed the costs. ✓
- No. American Airlines should not market the regular fares from JFK to LAX because the revenue does not increase at all by increasing the demand for these tickets.

EXPLANATION

In the current solution, we are meeting the demand for regular tickets from JFK to LAX. If we increase the demand by 10, we offer 10 more regular tickets, but our revenue only increases by \$140, which does not exceed the cost of \$2000. If we increase the demand by 50, to 130, we only offer 91 seats. Therefore, American Airlines should not market the regular fares from JFK to LAX because even though the revenue increases, it does not exceed the costs.

[Hide Answer](#)

You have used 2 of 2 submissions



[About](https://www.edx.org/about-us) (<https://www.edx.org/about-us>) [Jobs](https://www.edx.org/jobs) (<https://www.edx.org/jobs>)
[Press](https://www.edx.org/press) (<https://www.edx.org/press>) [FAQ](https://www.edx.org/student-faq) (<https://www.edx.org/student-faq>)
[Contact](https://www.edx.org/contact) (<https://www.edx.org/contact>)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community/>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)

© 2014 edX, some rights reserved.

[Terms of Service and Honor Code](#) -
[Privacy Policy](#) (<https://www.edx.org/edx-privacy-policy>)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[yllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[chedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 7 (1/1 point)

IMPORTANT NOTE: This question uses the original model with the independent variables "OfficeVisits" and "Narcotics". Be sure to use this model, instead of the model you built in Quick Question 4.

Compute the test set predictions in R by running the command:

```
predictTest = predict(QualityLog, type="response", newdata=qualityTest)
```

You can compute the test set AUC by running the following two commands in R:

```
ROCRpredTest = prediction(predictTest, qualityTest$PoorCare)
```

```
auc = as.numeric(performance(ROCRpredTest, "auc")@y.values)
```

What is the AUC of this model on the test set?

0.7994792

0.7994792

Answer: 0.7994792

EXPLANATION

If you run the two commands given above in your R console, you can see the value of the AUC of this model on the test set by just typing auc in your console. The value should be 0.7994792.

The AUC of a model has the following nice interpretation: given a random patient who actually received poor care, and a random patient who actually received good care, the AUC is the percentage of time that our model will classify which is which correctly.



Final Check

Save

Hide Answer

You have used 2 of 3 submissions





(<http://youtube.com/user/edxonline>)
© 2014 edX, some rights reserved.

[Terms of Service and Honor Code](#) -
[Privacy Policy](#) (<https://www.edx.org/edx-privacy-policy>)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[yllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[chedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 7 (1/1 point)

Run the cutree function again to create the cluster groups, but this time pick k = 2 clusters. It turns out that the algorithm groups all of the movies that only belong to one specific genre in one cluster (cluster 2), and puts all of the other movies in the other cluster (cluster 1). What is the genre that all of the movies in cluster 2 belong to?

- Action
- Adventure
- Animation
- Children's
- Comedy
- Crime
- Documentary
- Drama ✓
- Fantasy
- Film Noir
- Horror
- Musical
- Mystery
- Romance
- Sci-Fi
- Thriller
- War
- Western

EXPLANATION



You can redo the cluster grouping with just two clusters by running the following command:

```
clusterGroups = cutree(clusterMovies, k = 2)
```

Then, by using the tapply function just like we did in the video, you can see the average value in each genre and cluster. It turns out that all of the movies in the second cluster belong to the drama genre.

Alternatively, you can use colMeans or lapply as explained below Video 7.

[Check](#)[Save](#)[Hide Answer](#)*You have used 1 of 3 submissions*

[About](https://www.edx.org/about-us) (<https://www.edx.org/about-us>) [Jobs](https://www.edx.org/jobs) (<https://www.edx.org/jobs>)
[Press](https://www.edx.org/press) (<https://www.edx.org/press>) [FAQ](https://www.edx.org/student-faq) (<https://www.edx.org/student-faq>)
[Contact](https://www.edx.org/contact) (<https://www.edx.org/contact>)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)
© 2014 edX, some rights reserved.

[Terms of Service and Honor Code](#) -
[Privacy Policy](#) (<https://www.edx.org/edx-privacy-policy>)

[Courseware \(/courses/MITx/15.071x/1T2014/courseware\)](#)[Course Info \(/courses/MITx/15.071x/1T2014/info\)](#)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum\)](#)[Progress \(/courses/MITx/15.071x/1T2014/progress\)](#)[yllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](#)[chedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](#)

Help

QUICK QUESTION 7 (1/1 point)

In the previous video, we used CART and Random Forest to predict sentiment. Let's see how well logistic regression does. Build a logistic regression model (using the training set) to predict "Negative" using all of the independent variables. You may get a warning message after building your model - don't worry (we explain what it means in the explanation).

Now, make predictions using the logistic regression model:

```
predictions = predict(tweetLog, newdata=testSparse, type="response")
```

where "tweetLog" should be the name of your logistic regression model. You might also get a warning message after this command, but don't worry - it is due to the same problem as the previous warning message.

Build a confusion matrix (with a threshold of 0.5) and compute the accuracy of the model. What is the accuracy?

0.8056338

0.8056338

Answer: 0.8197183

EXPLANATION

You can build a logistic regression model in R by using the command:

```
tweetLog = glm(Negative ~ ., data=trainSparse, family="binomial")
```

Then you can make predictions and build a confusion matrix with the following commands:

```
predictLog = predict(tweetLog, newdata=testSparse, type="response")
```

=
table(testSparse\$Negative, predictLog > 0.5)

The accuracy is $(254+37)/(254+46+18+37) = 0.8197183$, which is worse than the baseline. If you were to compute the accuracy on the training set instead, you would see that the model does really well on the training set - this is an example of over-fitting. The model fits the training set really well, but does not perform well on the test set. A logistic regression model with a large number of variables is particularly at risk for overfitting.

Note that you might have gotten a different answer than us, because the `glm` function struggles with this many variables. The warning messages that you might have seen in this problem have to do with the number of variables, and the fact that the model is overfitting to the training set. We'll discuss this in more detail in the Homework Assignment.

Is this worse or better than the baseline model accuracy of 84.5%? Think about the properties of logistic regression that might make this the case!

Check

Save

Hide Answer

You have used 1 of 5 submissions



About (<https://www.edx.org/about-us>) Jobs (<https://www.edx.org/jobs>)
Press (<https://www.edx.org/press>) FAQ (<https://www.edx.org/student-faq>)
Contact (<https://www.edx.org/contact>)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community/>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)

© 2014 edX, some rights reserved.

Terms of Service and Honor Code -
Privacy Policy (<https://www.edx.org/edx-privacy-policy>)