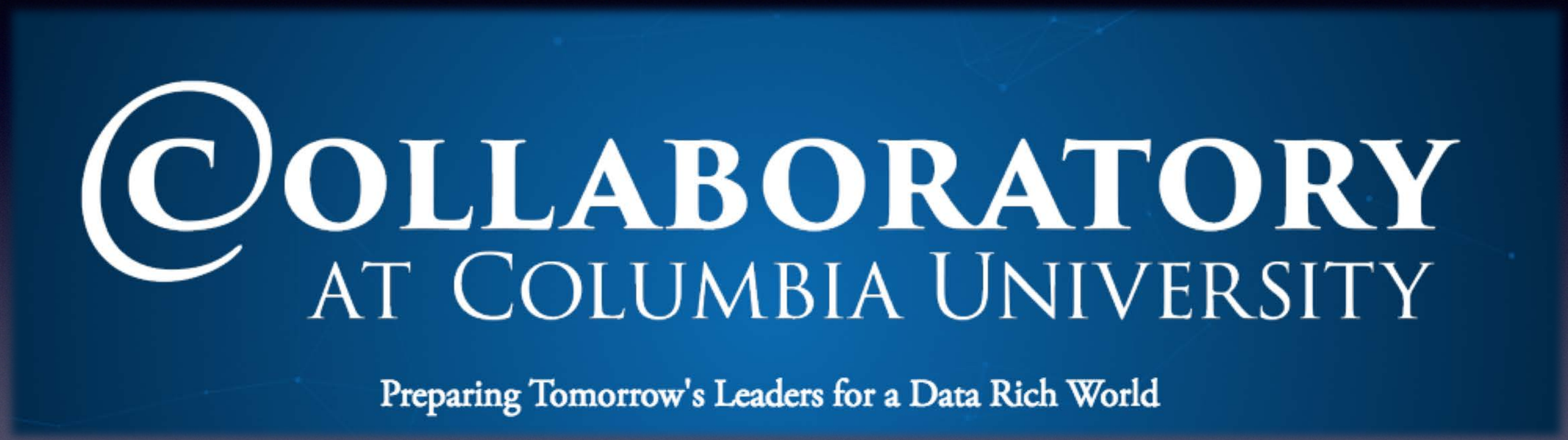
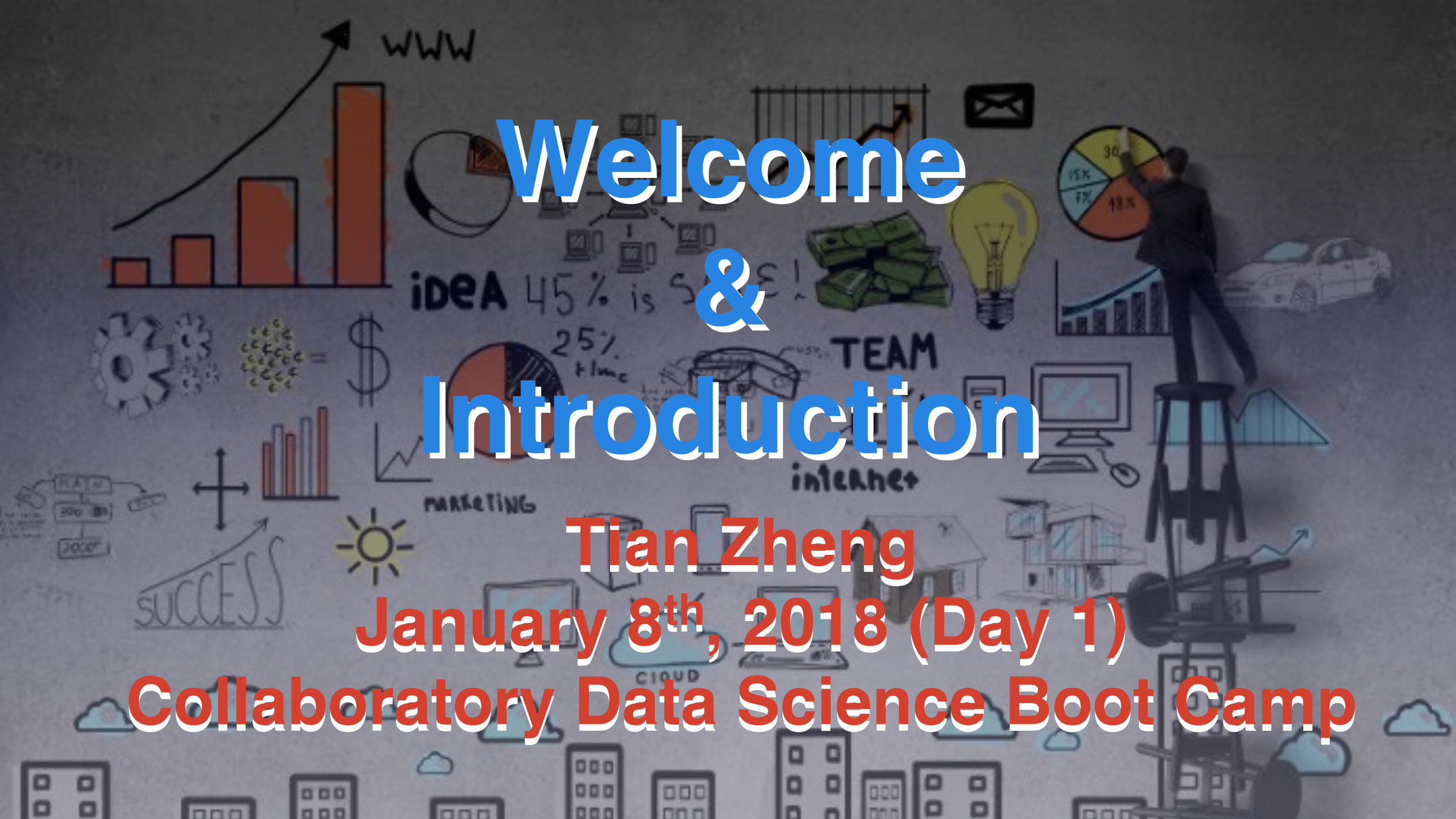

@COLLABORATORY
AT COLUMBIA UNIVERSITY
Preparing Tomorrow's Leaders for a Data Rich World



Welcome & Introduction

Tian Zheng

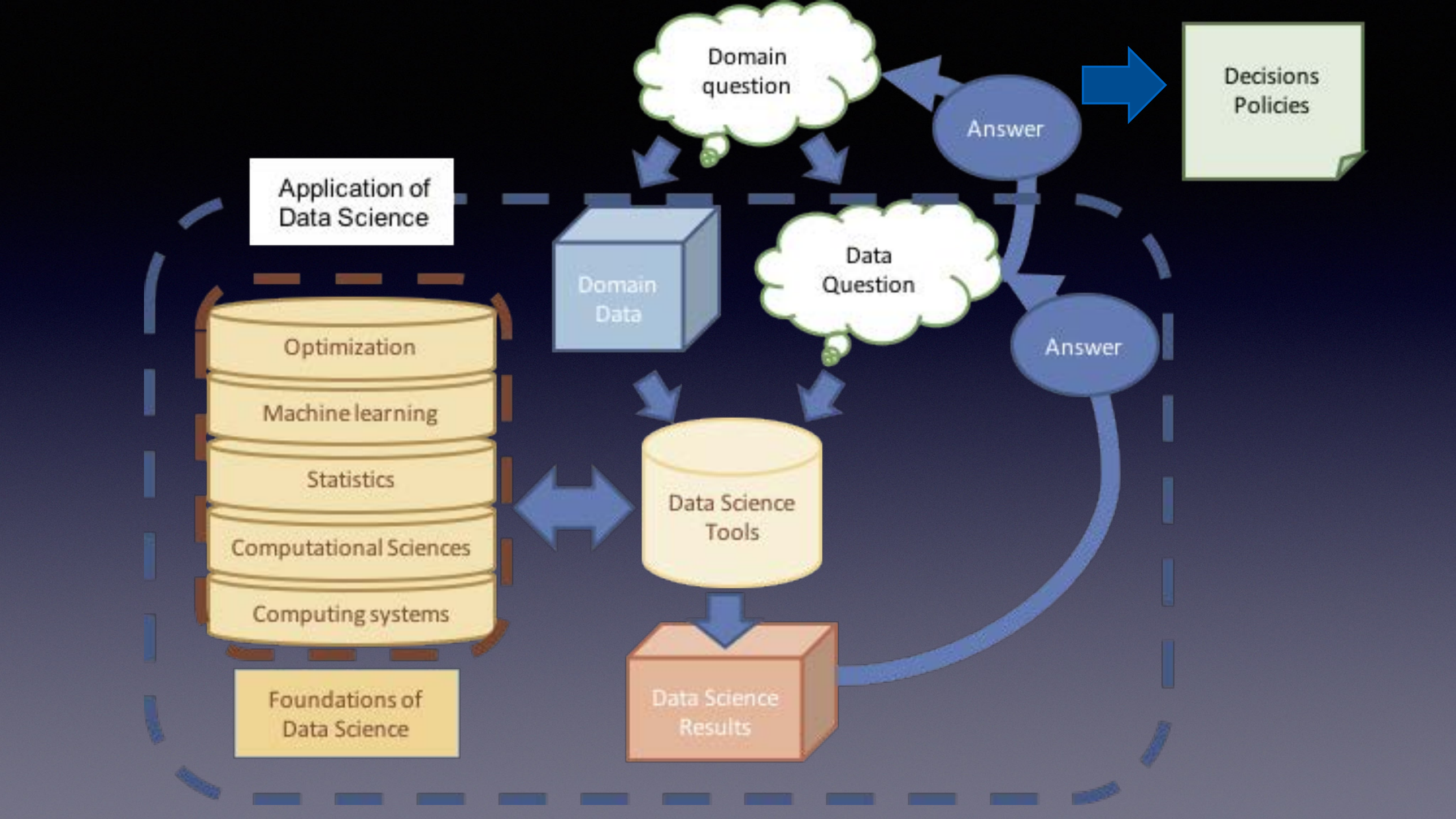
January 8th, 2018 (Day 1)

Collaboratory Data Science Boot Camp

What is data science?

- Data Science represents a new approach to
  - Acquire knowledge,
  - Collect evidence,
  - Form decisions,
  - Make predictions.
- The end points are:

  knowledge, evidence,
  decisions and predictions.
- Driven by breakthroughs in technologies.
- Enabling faster solutions to traditional evidence-based practices.
- Creating solutions that would not be otherwise possible.


Domain question
Answer
Decisions Policies
Application of Data Science
Domain Data
Data Question
Answer
Optimization
Machine learning
Statistics
Computational Sciences
Computing systems
Data Science Tools
Data Science Results
Foundations of Data Science

Data Science Skill Set

- How to **think** about data versus problem:
  - Mathematics/Statistics/Machine Learning
- How to **handle** data
  - Technologies: Python, Java, Hadoop, Spark, etc
- Teamwork and collaboration skills - how to **work** with others.
- How to turn data into business intelligence:

  find **value** in your data
  - Innovation, intellectual curiosity
  - Problem-solving skills
- How to convince others about your data science results
  - Visualization, story telling
  - **Communication** skills

Where do you start?

- **Basics**
  - **Statistics and Probability**
  - **Become comfortable with computing**
  - **Understand modeling and inference**
  - **Machine learning and algorithms**
  - **Visualization**

Skills listed in job postings on indeed.com

What would be the most useful skill(s)?

The “skills” that make a data scientist

- **Problem identification** via data exploration, discussion and team work.
- **Problem solving** by using existing skills or new skills
  - **learn new things “on the job”**
  - **learn from your peers.**
- **Present** your codes, your results *and* your story.

Project-based learning

- **Course projects**
- **Kaggle competition**
- **Coding meet-ups**
- **Hackathons**
- **Project-based bootcamps**

Learning by doing

- Data analysis is a decision making process.
- Investigate: from data to answers
- Learn from others: collaborate and discuss
- Data competitions
  - **Common Task Framework**
  - Active discussion board
  - Winners are required to share full solutions
- Class projects or meet ups
  - Common Task Framework
  - Team discussion
  - Feedback on all areas of a project, not just the implementation.

Learning by making mistakes

- “Nothing in the world is worth having or worth doing unless it means effort, pain, difficulty...”

  — Theodore Roosevelt

- “If you get stuck, start with Google.”

  — Hadley Wickham

- “If you have some ideas, just try it. It won’t break the computer (probably).”

  — Tian Zheng

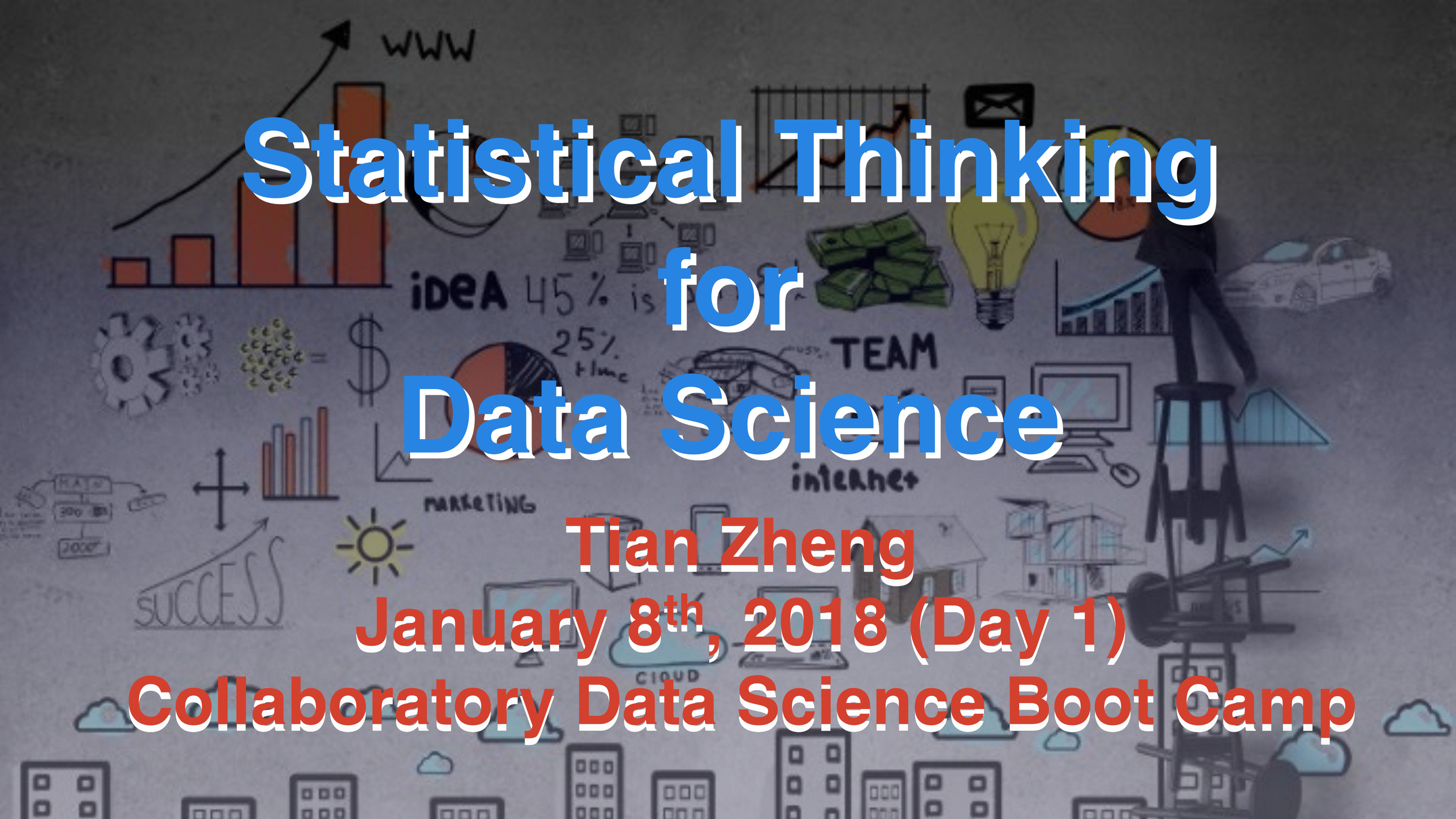
the instructors' answer, *where instructors collectively construct a single answer*

This seems to be an encoding issue. Are you using a Windows machine?

```
ff.all<-Corpus(DirSource(folder.path, encoding="UTF-8"))
```

edit · good answer | 0

Bootcamp logistics (let’s go to GitHub)



Statistical Thinking for Data Science

Tian Zheng

January 8th, 2018 (Day 1)

Collaboratory Data Science Boot Camp

Credit: http://chi2innovations.com/blog/discover-stats-blog-series/statistics-last-dark-art/

Science Isn't Broken

It's just a hell of a lot harder than we give it credit for.

By Christie Aschwanden

Filed under Scientific Method

Published Aug 19, 2015

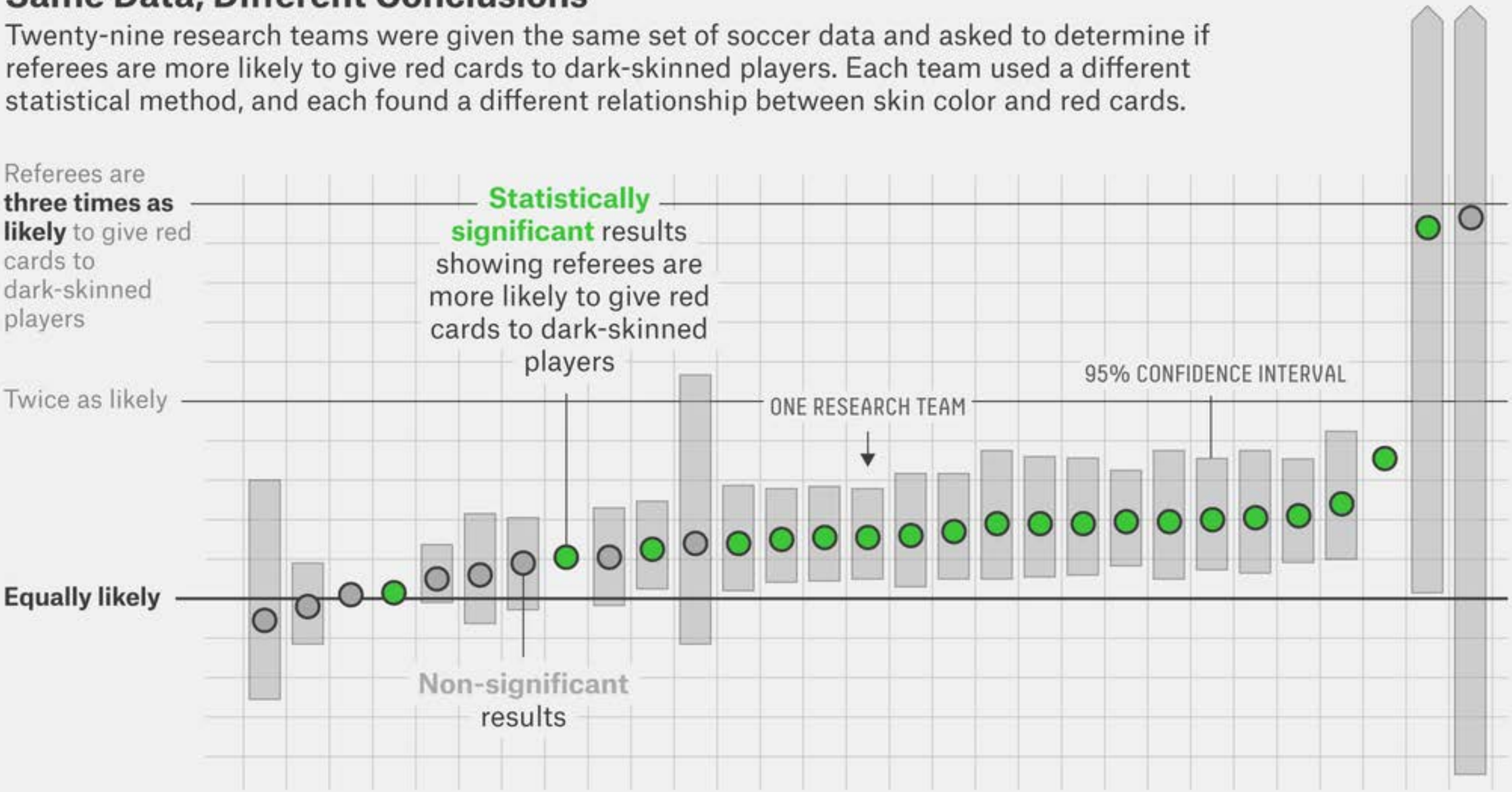
http://fivethirtyeight.com/features/science-isnt-broken/#part1

If you ask different questions you get different answers - one more way science isn't broken it is just really hard

http://simplystatistics.org/2015/08/20/if-you-ask-different-quetions-you-get-different-asnwers-one-more-way-science-isnt-broken-it-is-just-really-hard/

Same Data, Different Conclusions

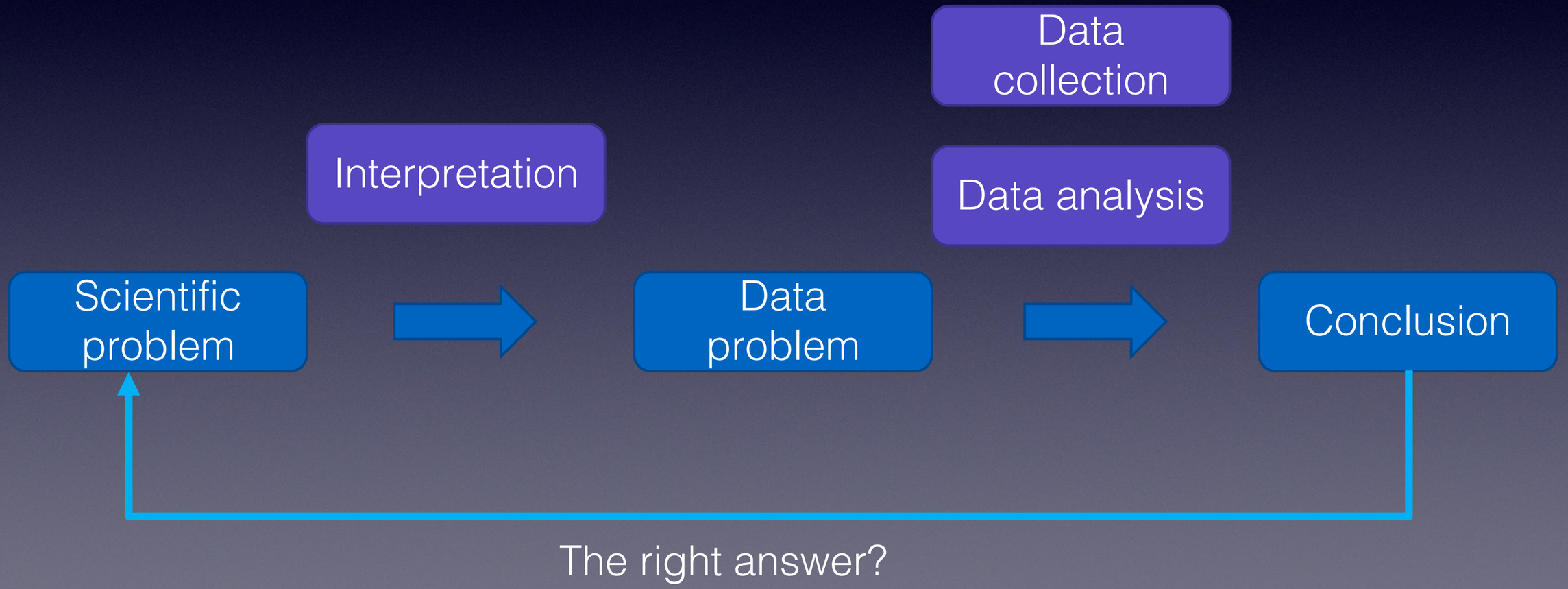
Twenty-nine research teams were given the same set of soccer data and asked to determine if referees are more likely to give red cards to dark-skinned players. Each team used a different statistical method, and each found a different relationship between skin color and red cards.



FIVETHIRTYEIGHT

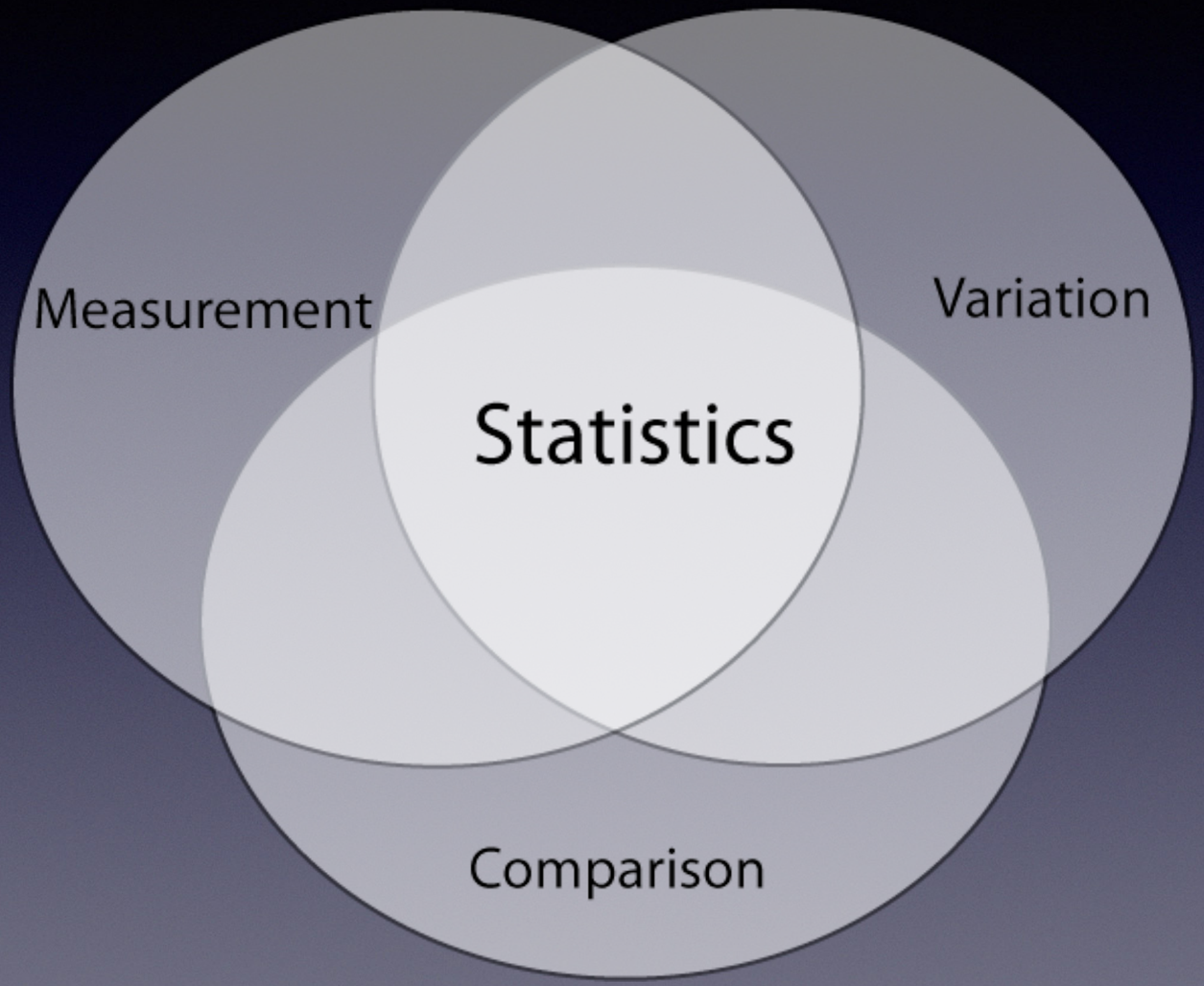
SOURCE: BRIAN NOSEK ET AL.

More than just “lost in translation”



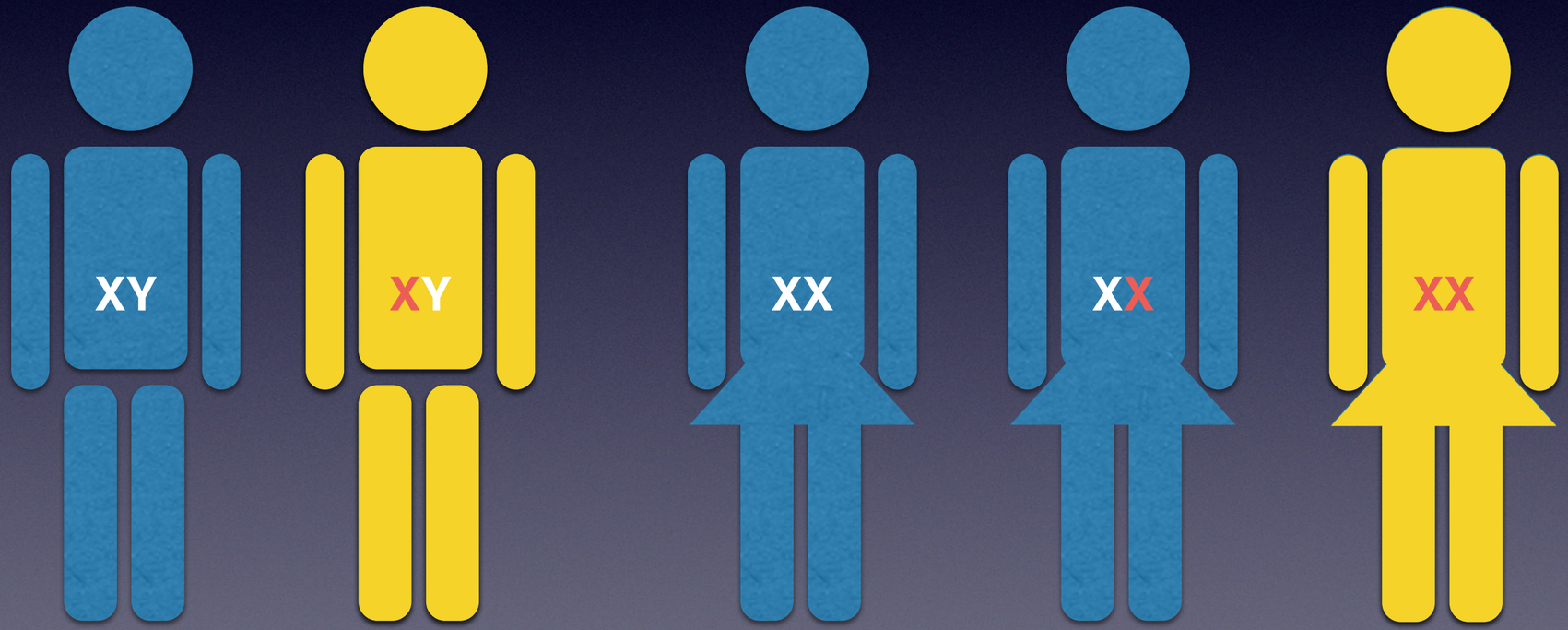
In a statistics curriculum

- Describe and compare variation
- Describe and model association

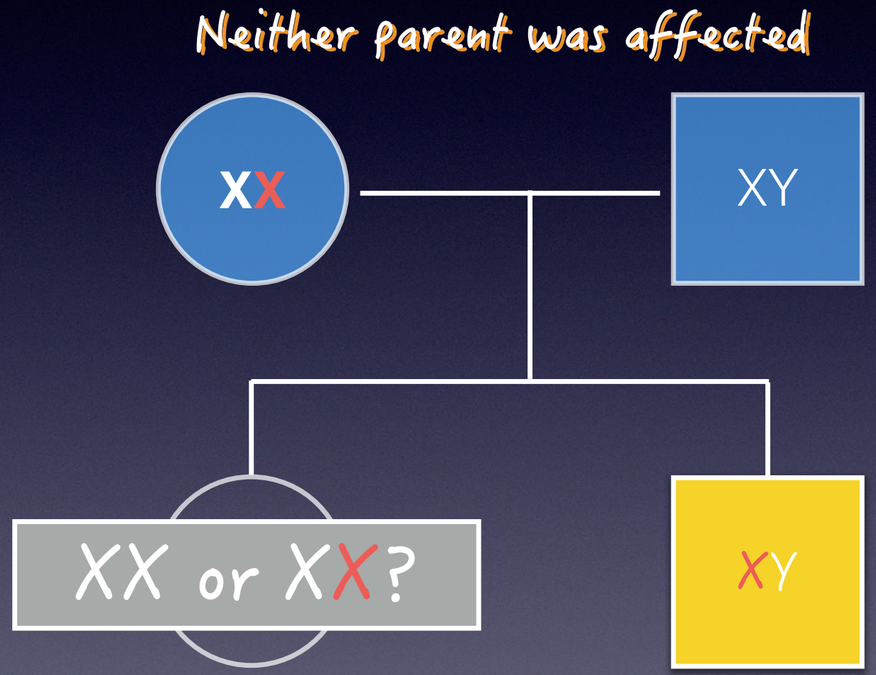


Conditional Probability and Bayes' formula

Hemophilia example



A woman with an affected brother



$P(XX)=P(X\mathbf{X})=0.5$?

Now we observe some data

A: XX

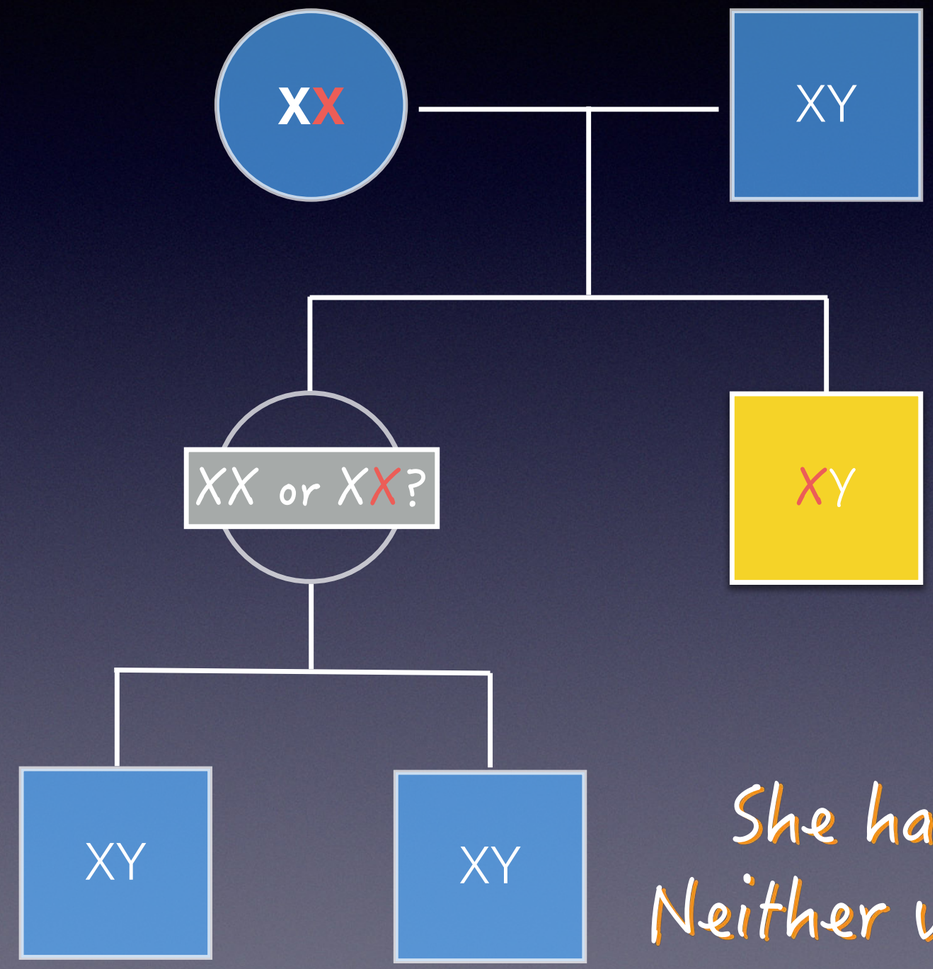
"not A": XX

B: 2 sons with XY

$P(B|A) = 1$

$P(B|\text{not } A) = (1/2)(1/2)$

$= 1/4$



She has two sons.
Neither was affected

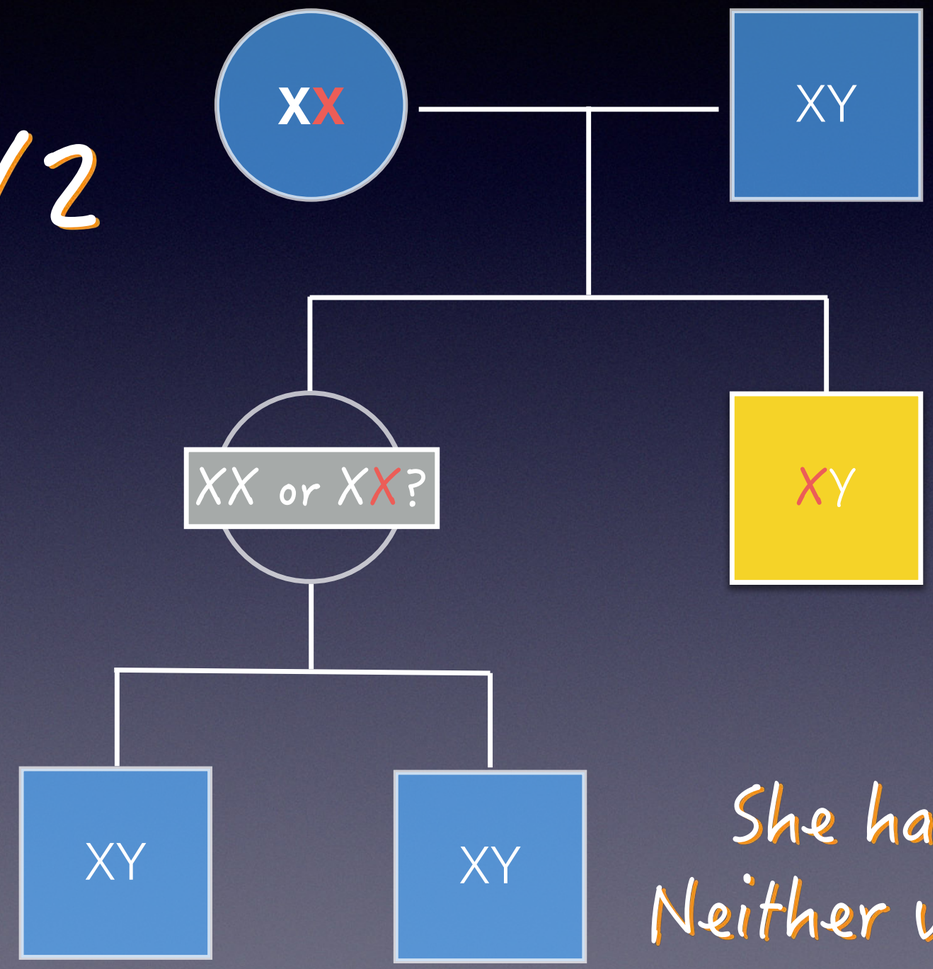
Bayes' formula

$P(A) = P(\text{not } A) = 1/2$

$P(B|A) = 1$

$P(B|\text{not } A) = 1/4$

$P(A|B) = ?$



She has two sons.
Neither was affected

Bayes' formula

P(A) = P(not A) =1/2
P(B|A) = 1
P(B|not A) = 1/4

P(A|B) = ?

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

More data

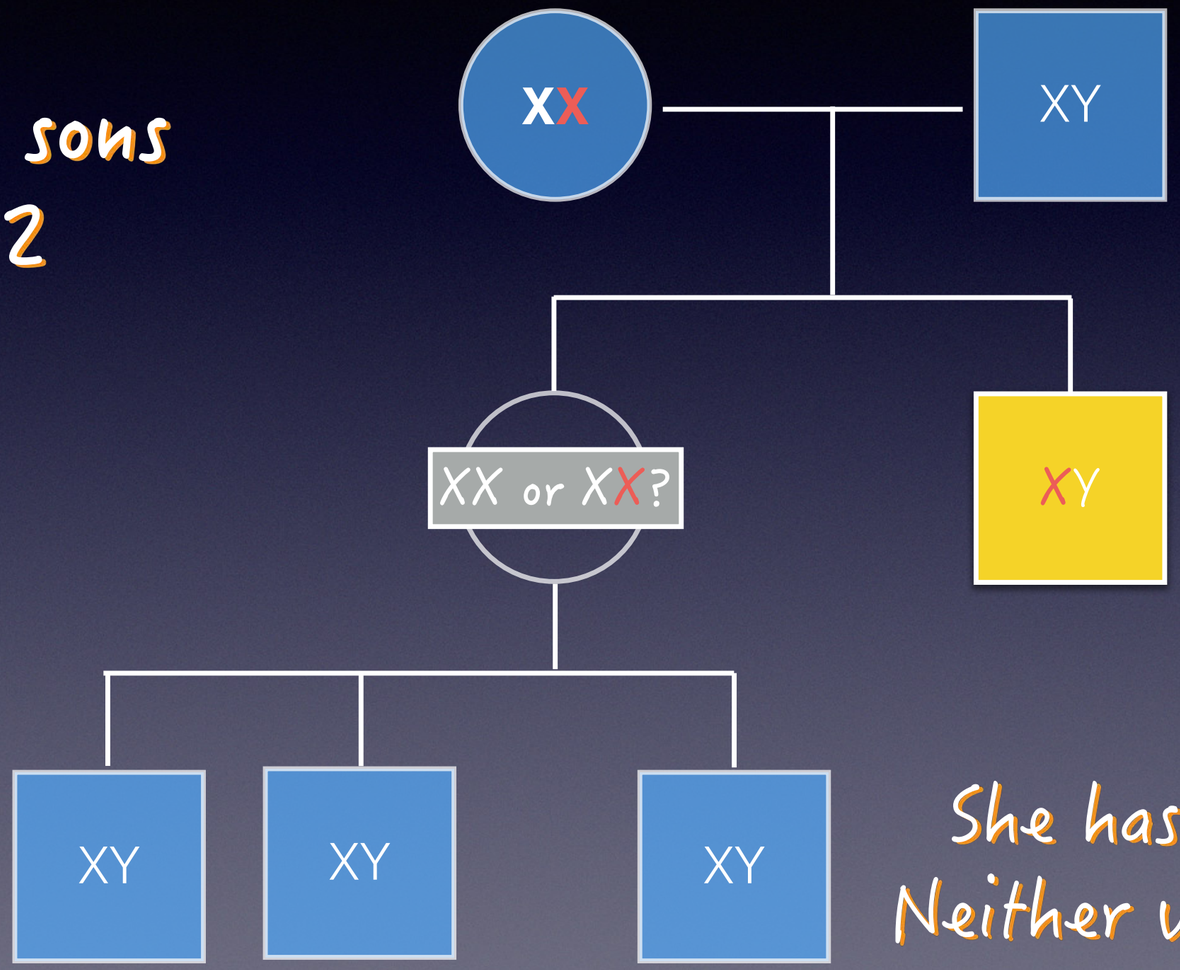
Now B: 3 unaffected sons

P(A) = P(not A) =1/2

P(B|A) = 1

P(B|not A) = 1/8

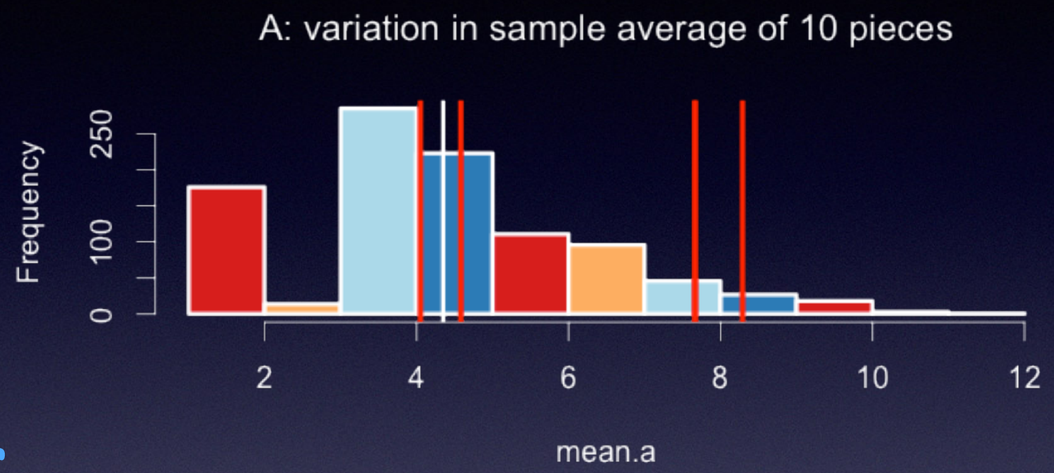
P(A|B) = ?



She has three sons.
Neither was affected

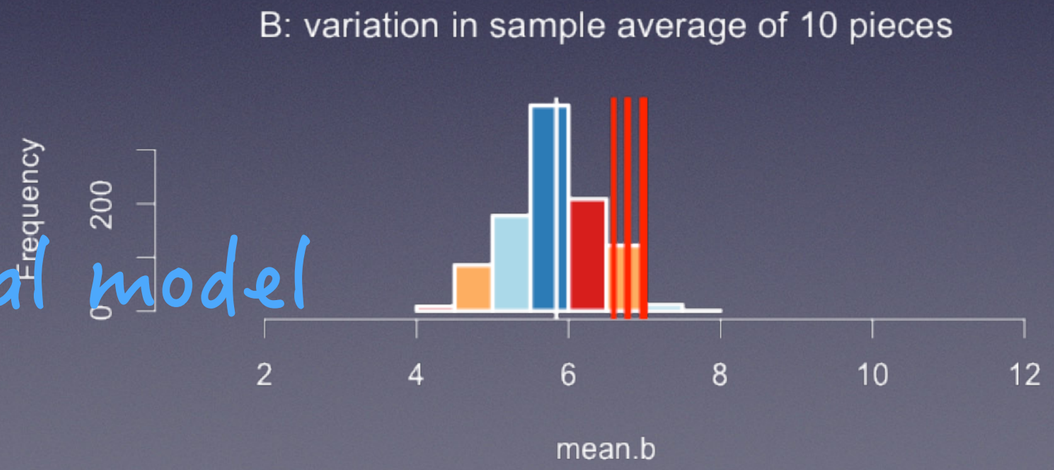
Sampling and Statistical Estimation

Recall the lego example



Sampling distributions

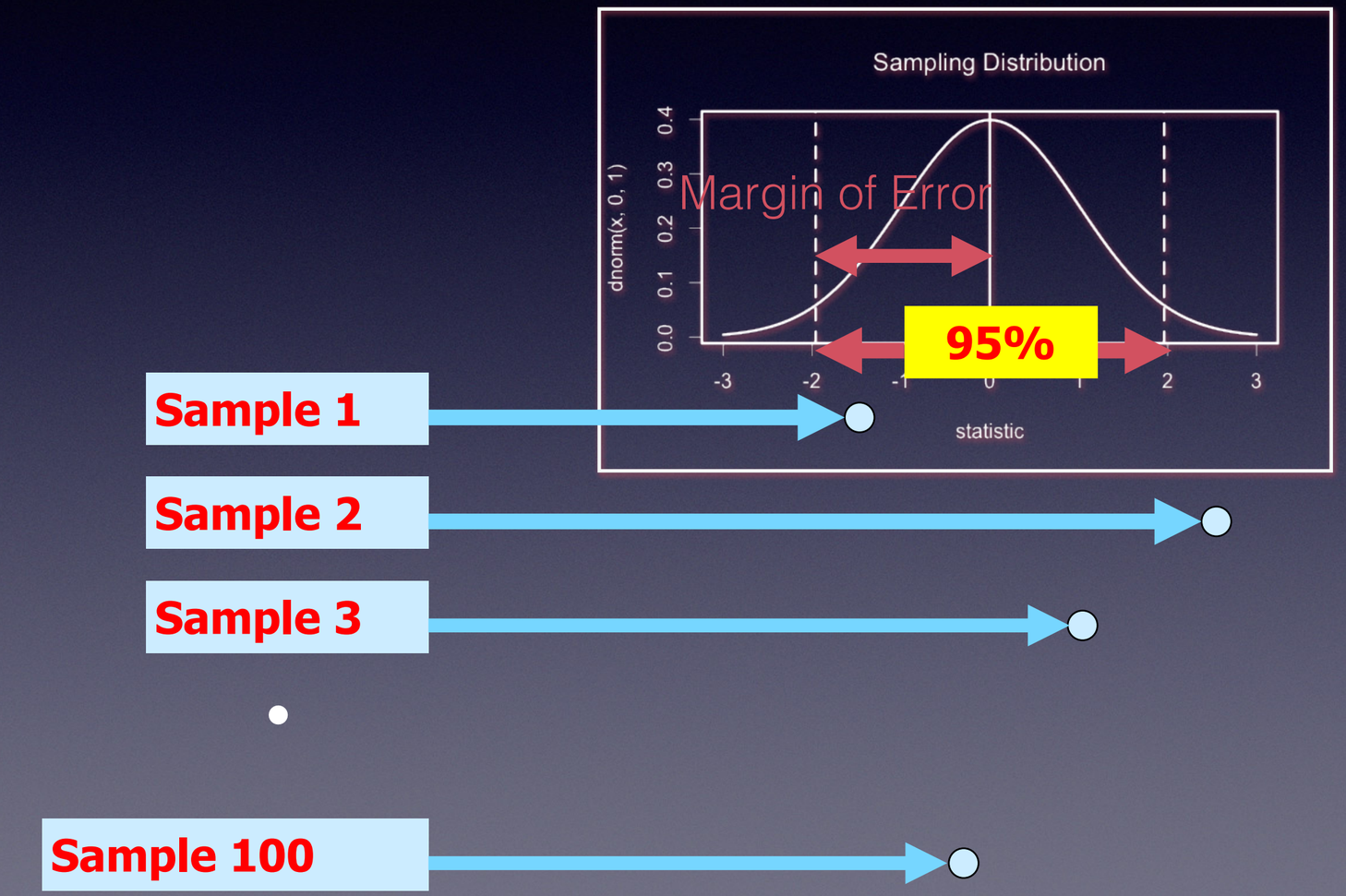
- by simulations
- by experiments
- by mathematical model



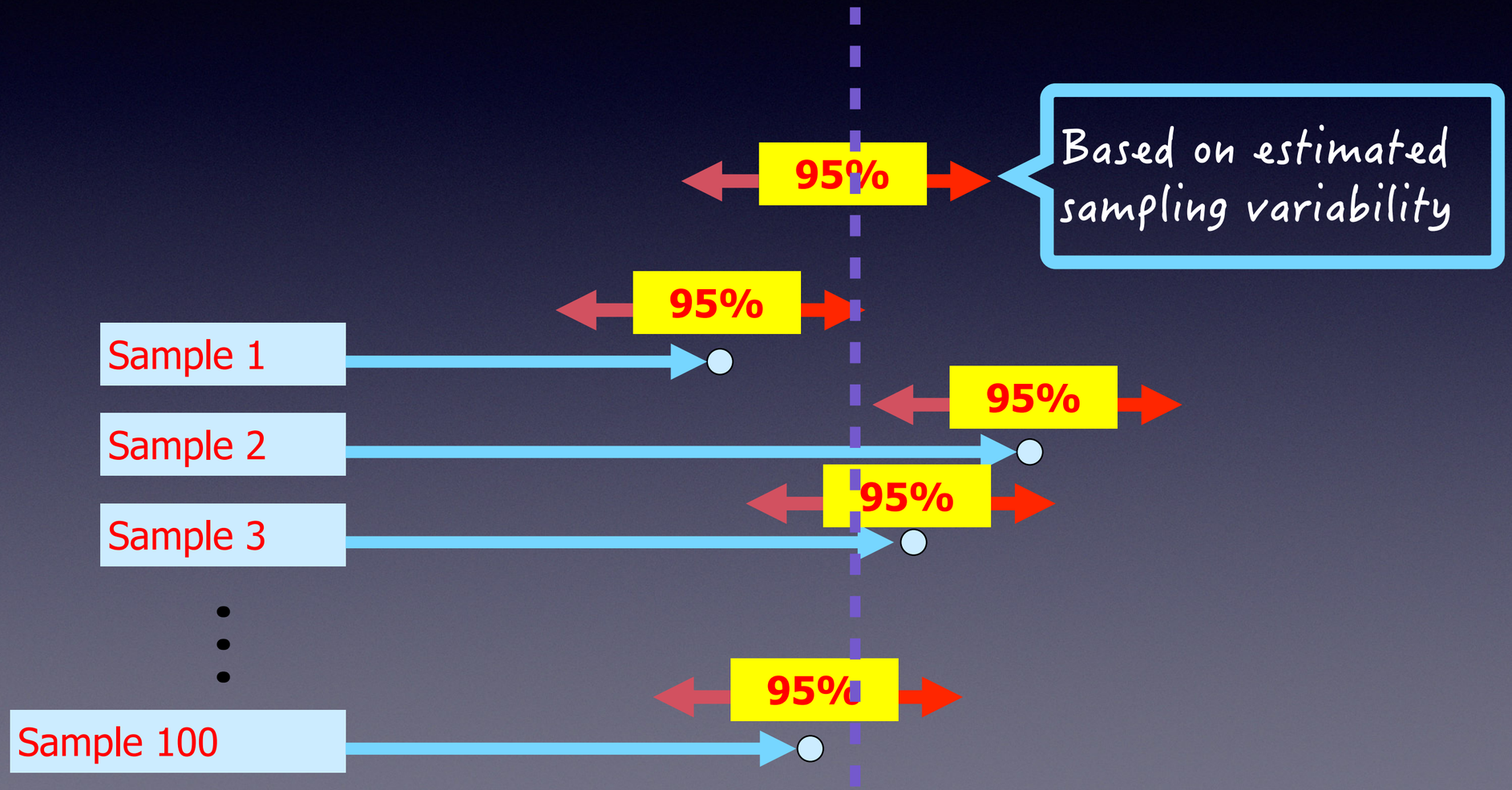
Sampling distribution model

- Probability model
- Describe the **variation** of a statistic (e.g., an estimate of a population quantity) due to sampling
  - Center
  - Variability
- Factors affect sampling distributions
  - Population variation for the quantity of interest
  - Data generation process
  - Sample size

Sampling distribution and confidence interval

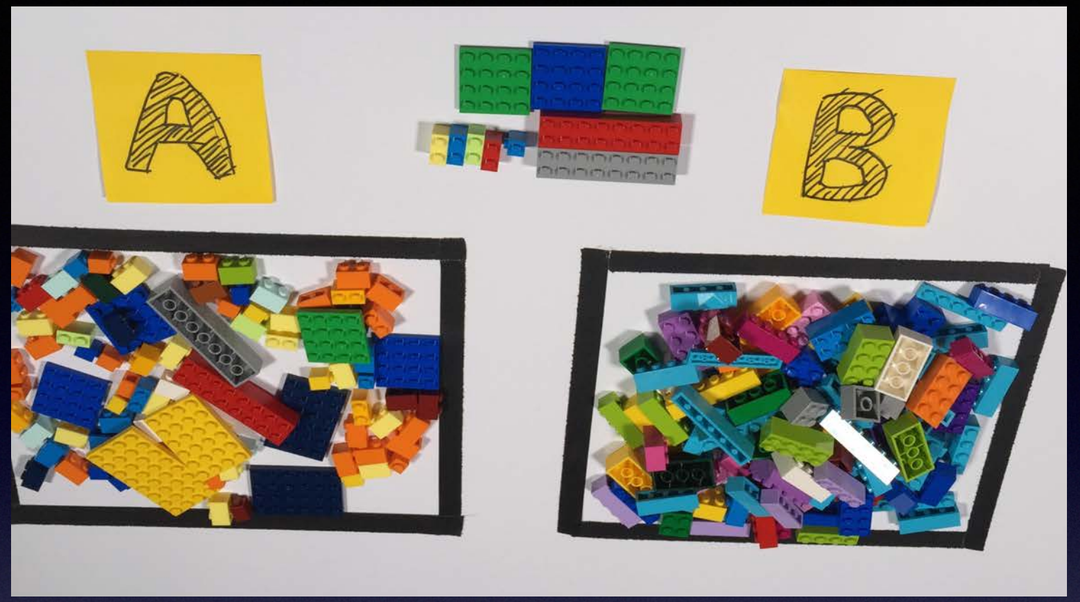


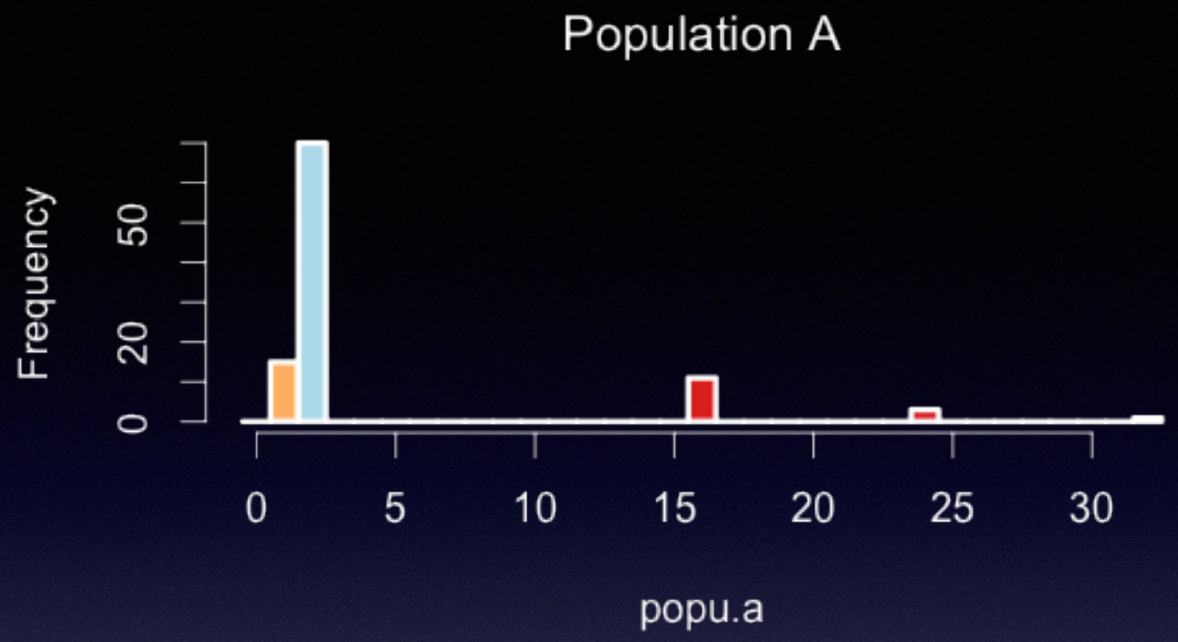
Sampling distribution and confidence interval

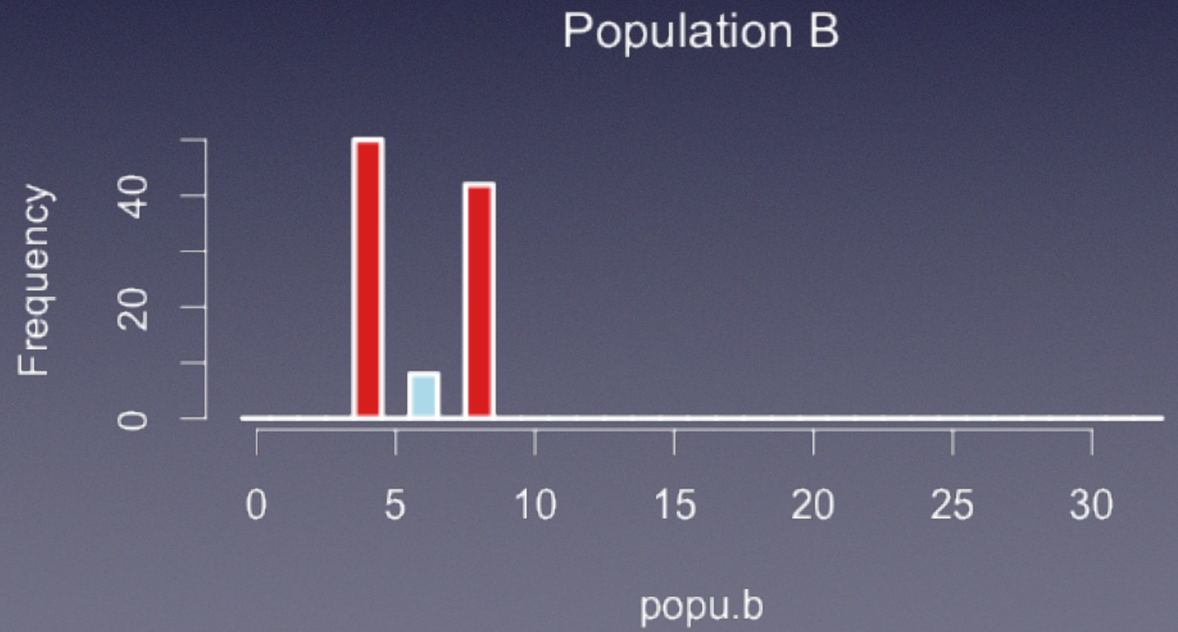


Lego demo

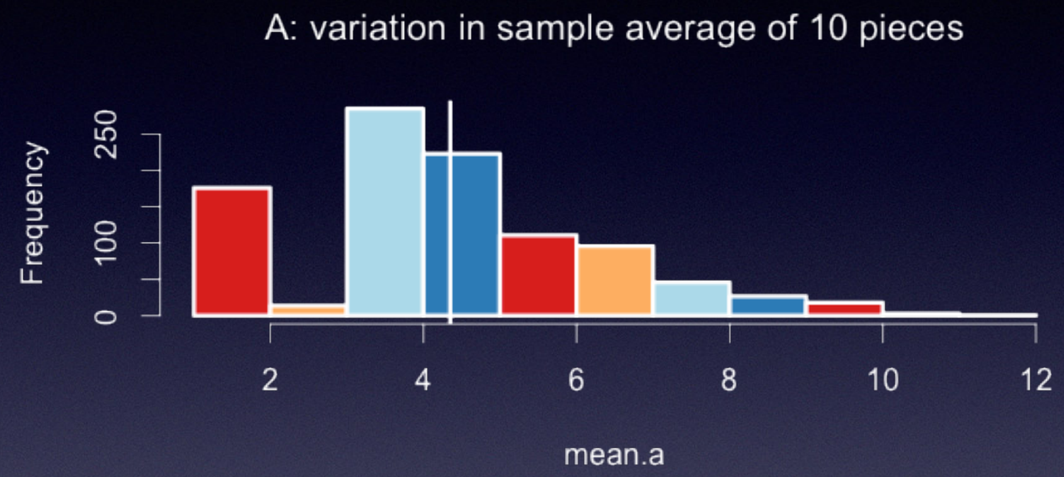
- From a population of 200 lego pieces, we create two populations.
- Each population has 100 pieces.
- X, Variable of interest is the number of points on a sampled lego piece.
  - **Population A**: a mixed population of very small pieces and very large pieces
  - **Population B**: a relatively more homogeneous population
- Population A: the average for X is 4.35
- Population B: the average for X is 5.84
- For illustration, we randomly sample pieces from these two populations by hand.

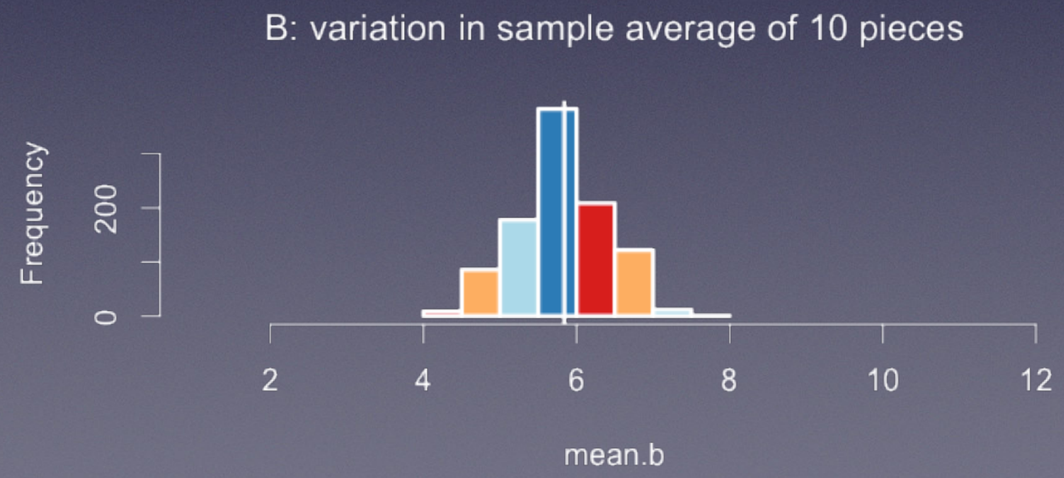



Population A
Frequency
0
20
50
0
5
10
15
20
25
30
popu.a

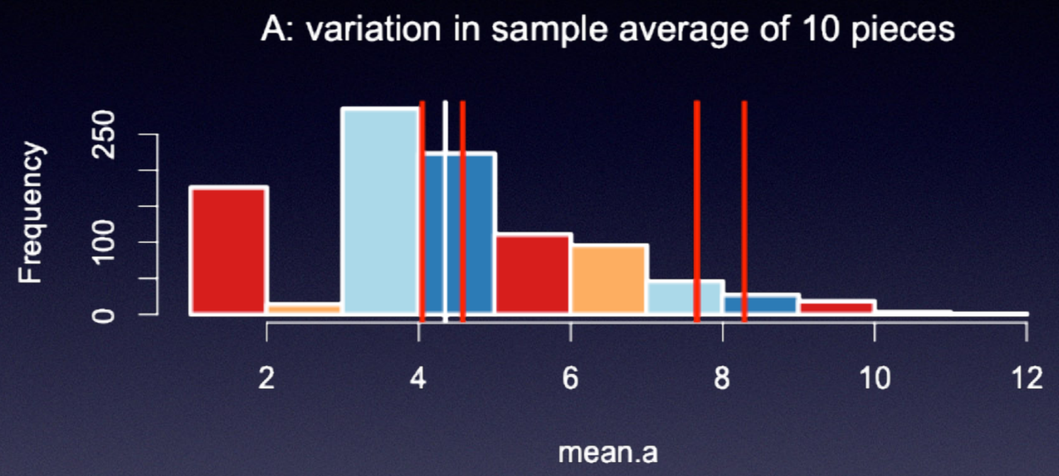

Population B
Frequency
0
20
40
0
5
10
15
20
25
30
popu.b

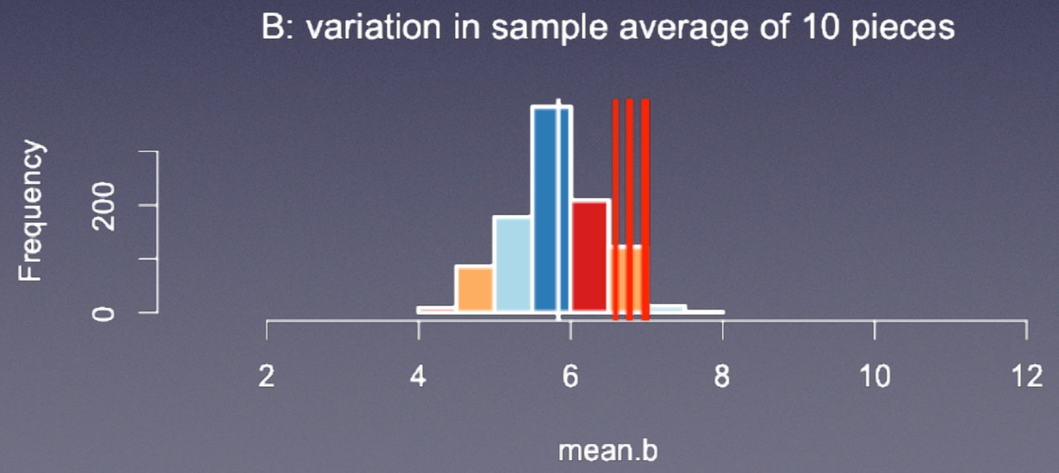
Sampling distribution of sample means (computer simulation)





Sample means from real samples

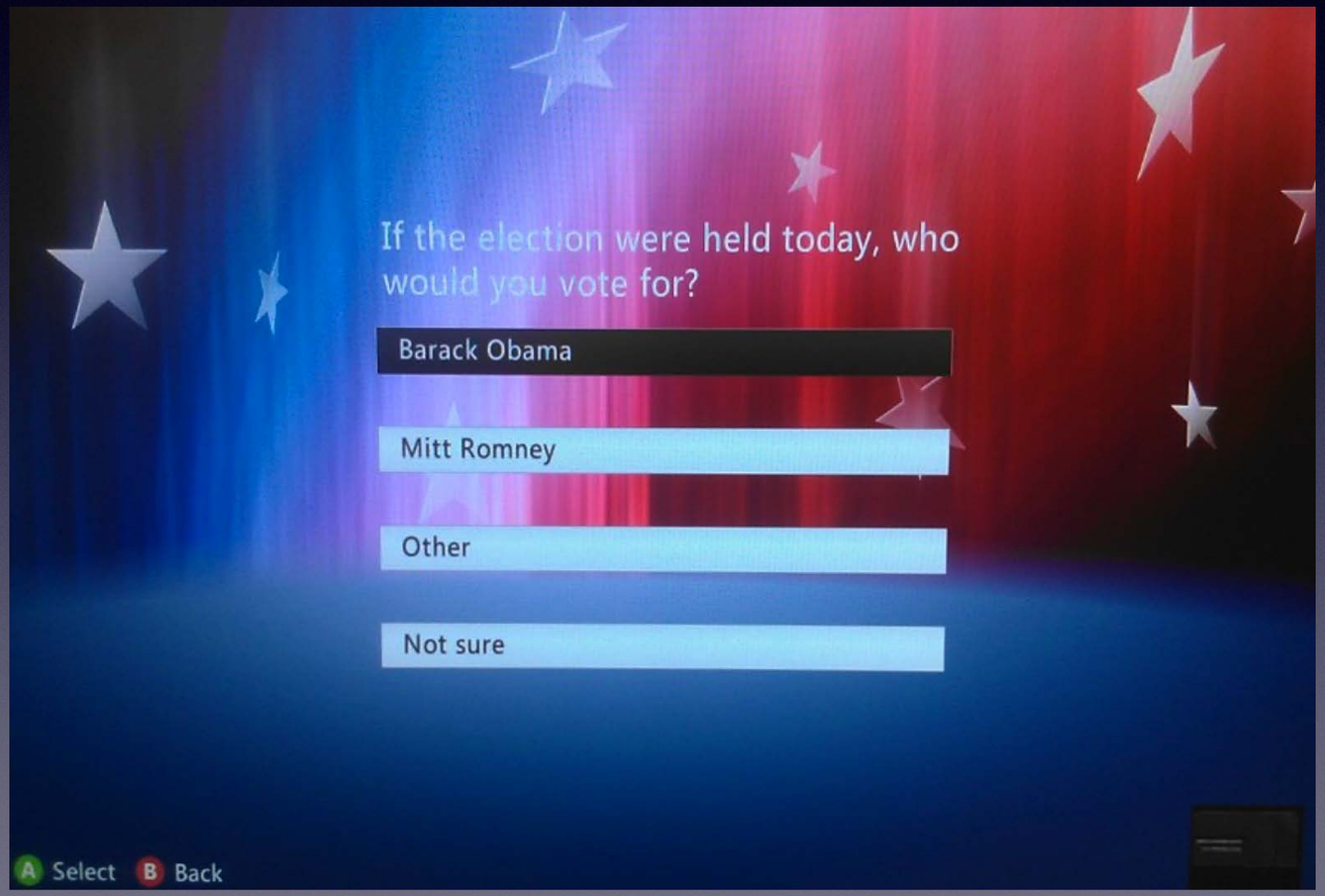


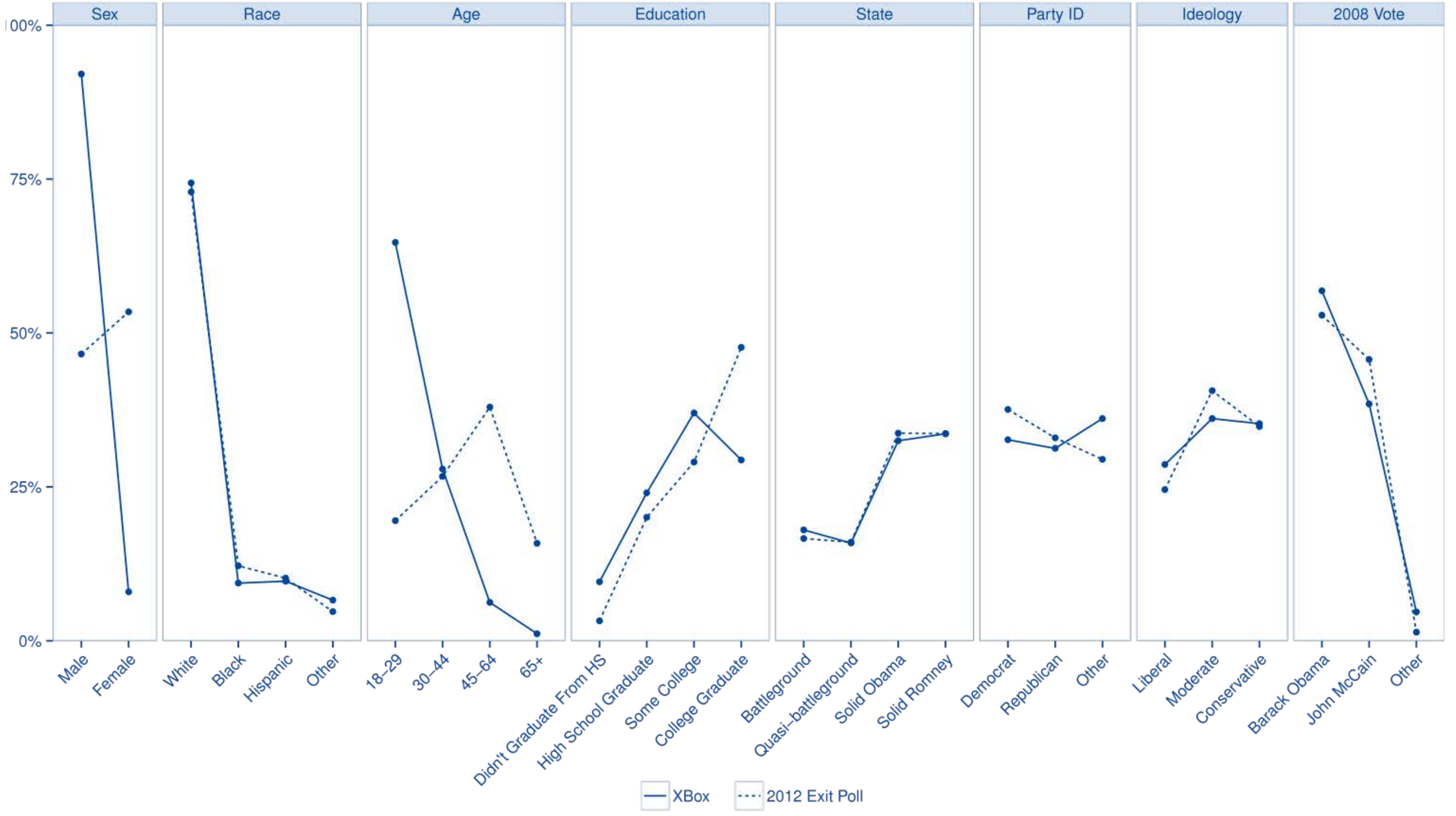


Bias in Sample estimates

- Estimates based on a sample differ from the truth.
  - Sampling bias or Selection bias in sampled observations on a variable
  - occurs when the probability that an individual is selected is associated with the variable's value.
- We need to model the selection process for understanding sampling bias, in addition to sampling variability.

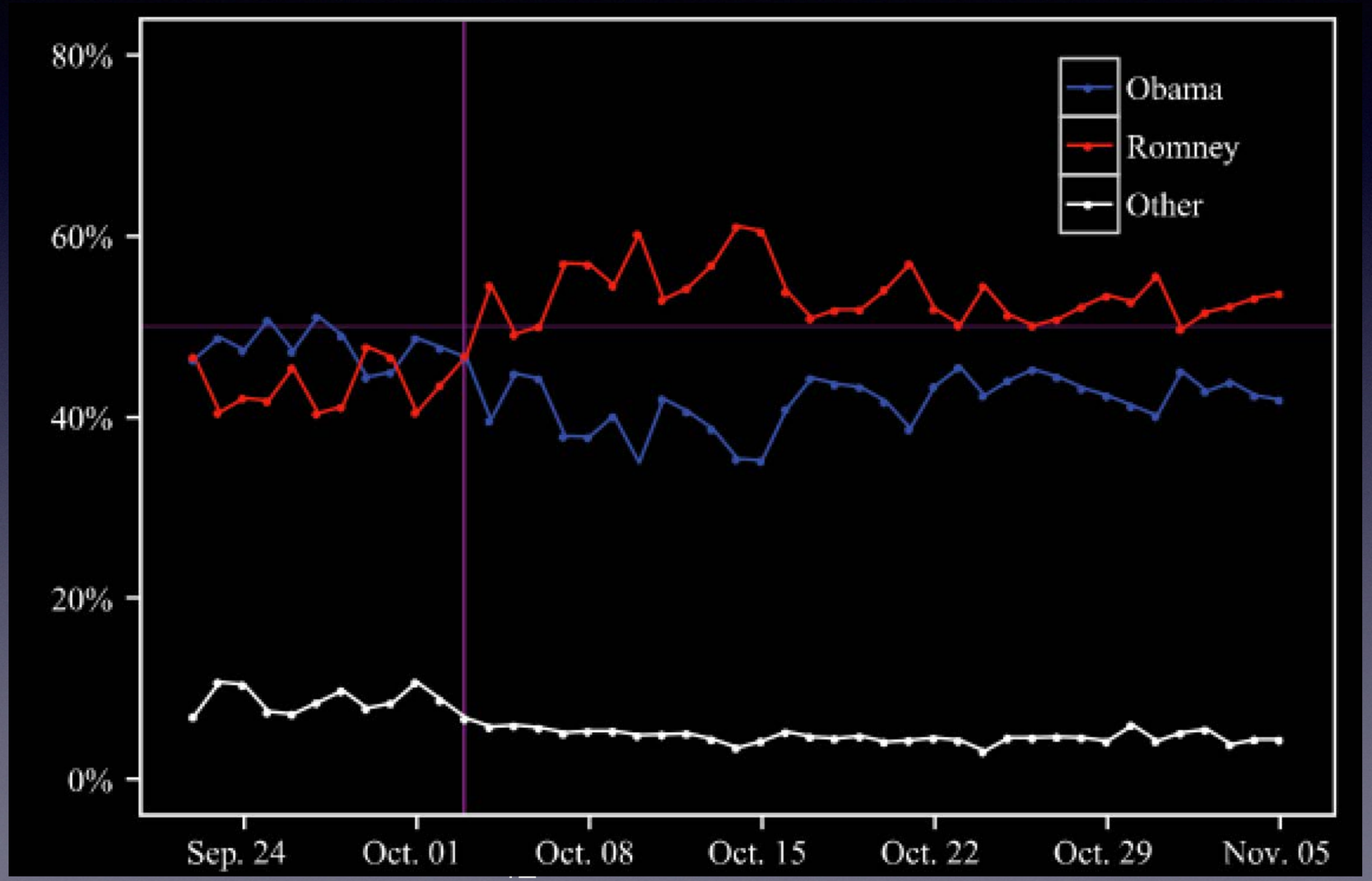
Surveying
a non-representative sample?



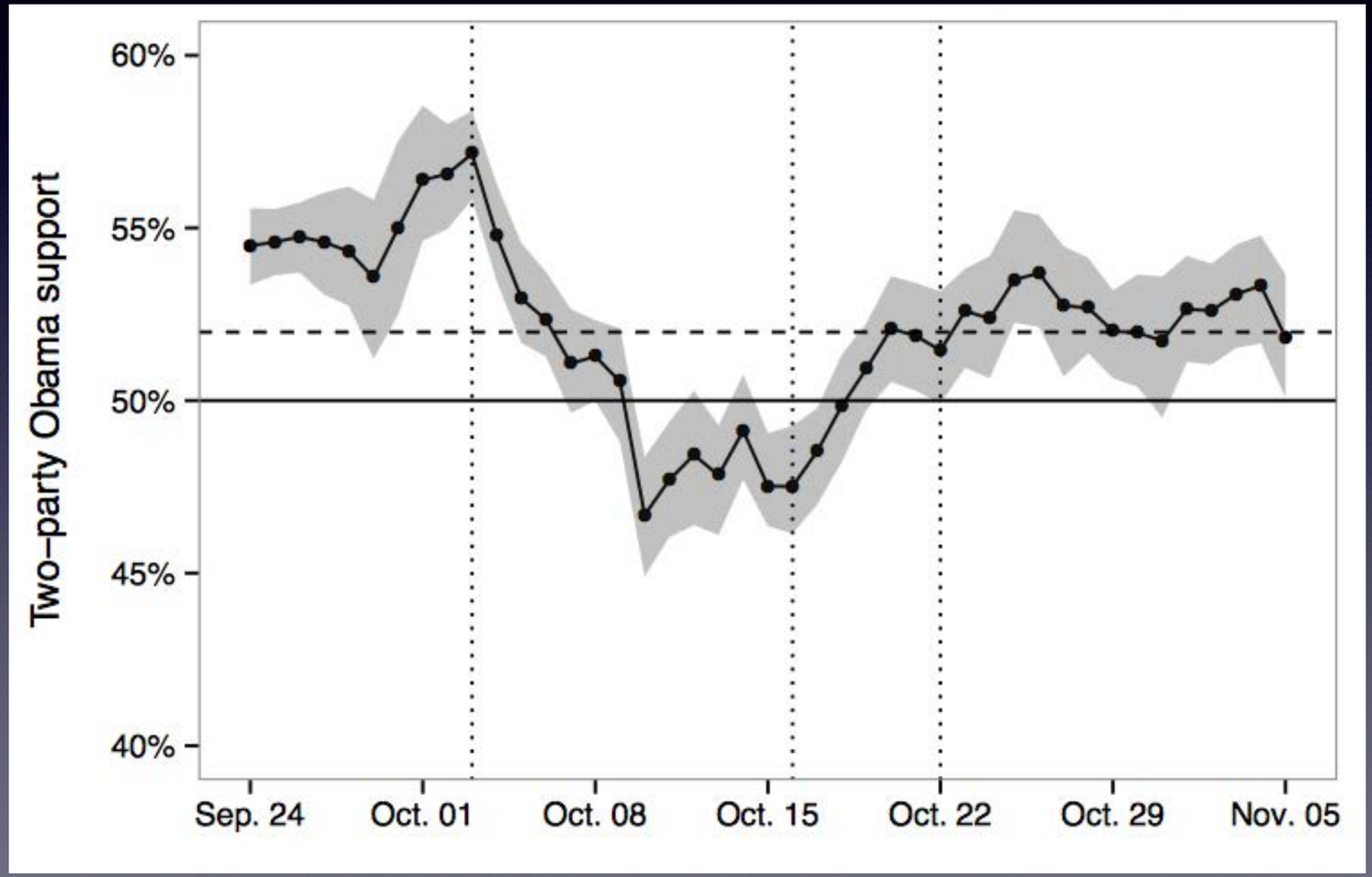

Sex
Race
Age
Education
State
Party ID
Ideology
2008 Vote
100%
75%
50%
25%
0%
Male
Female
White
Black
Hispanic
Other
18–29
30–44
45–64
65+
Didn't Graduate From HS
High School Graduate
Some College
College Graduate
Battleground
Quasi–battleground
Solid Obama
Solid Romney
Democrat
Republican
Other
Liberal
Moderate
Conservative
Barack Obama
John McCain
Other
XBox
2012 Exit Poll

A successful replication of the 1936 Literary Digest error

345,858
respondents



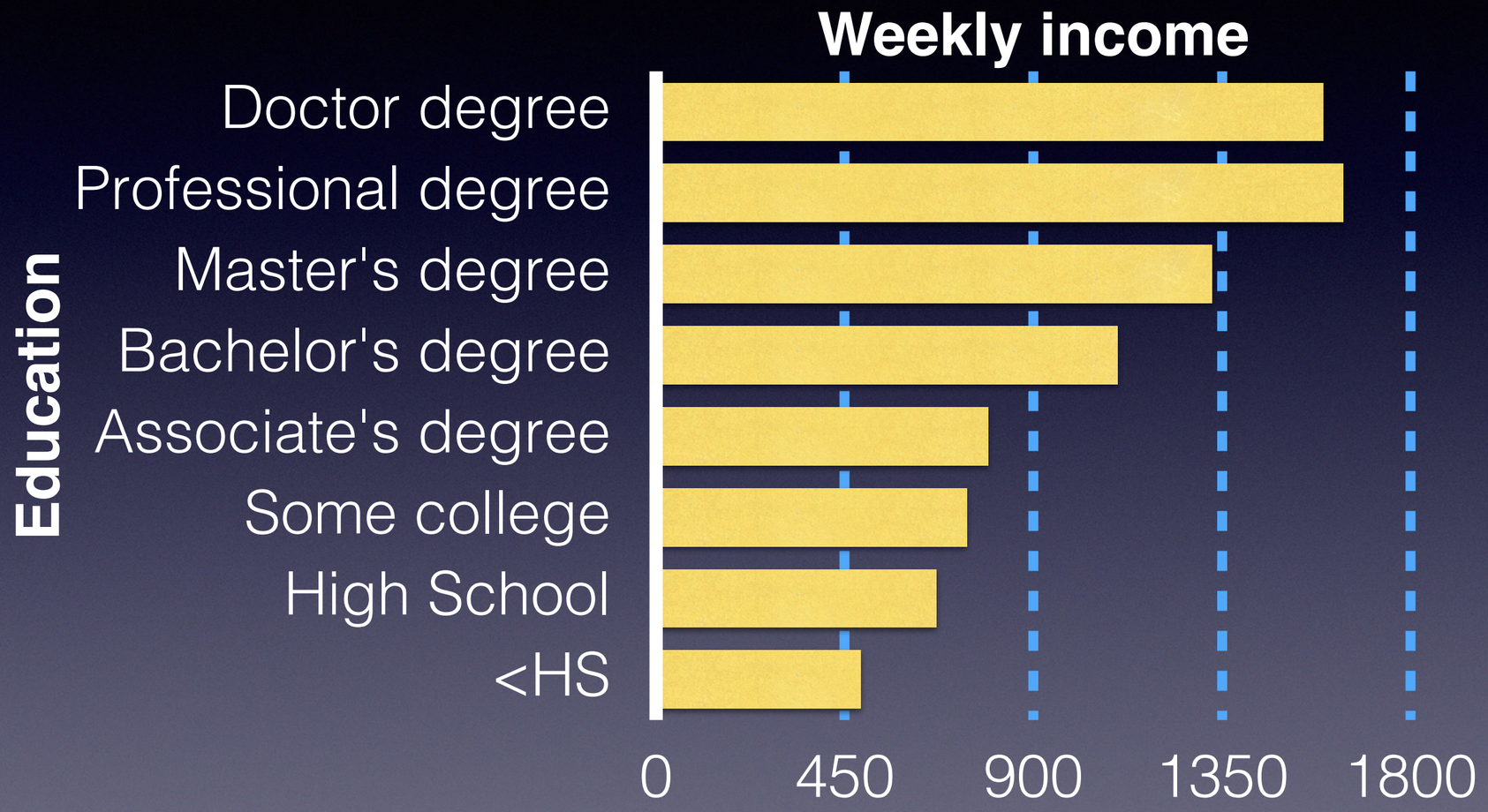
Adjusted using multilevel regression and poststratification*



Adjusted using 2008 exit poll statistics

Association: analysis of variance

Education versus Income



Data source: Bureau of Labor Statistics (2014)

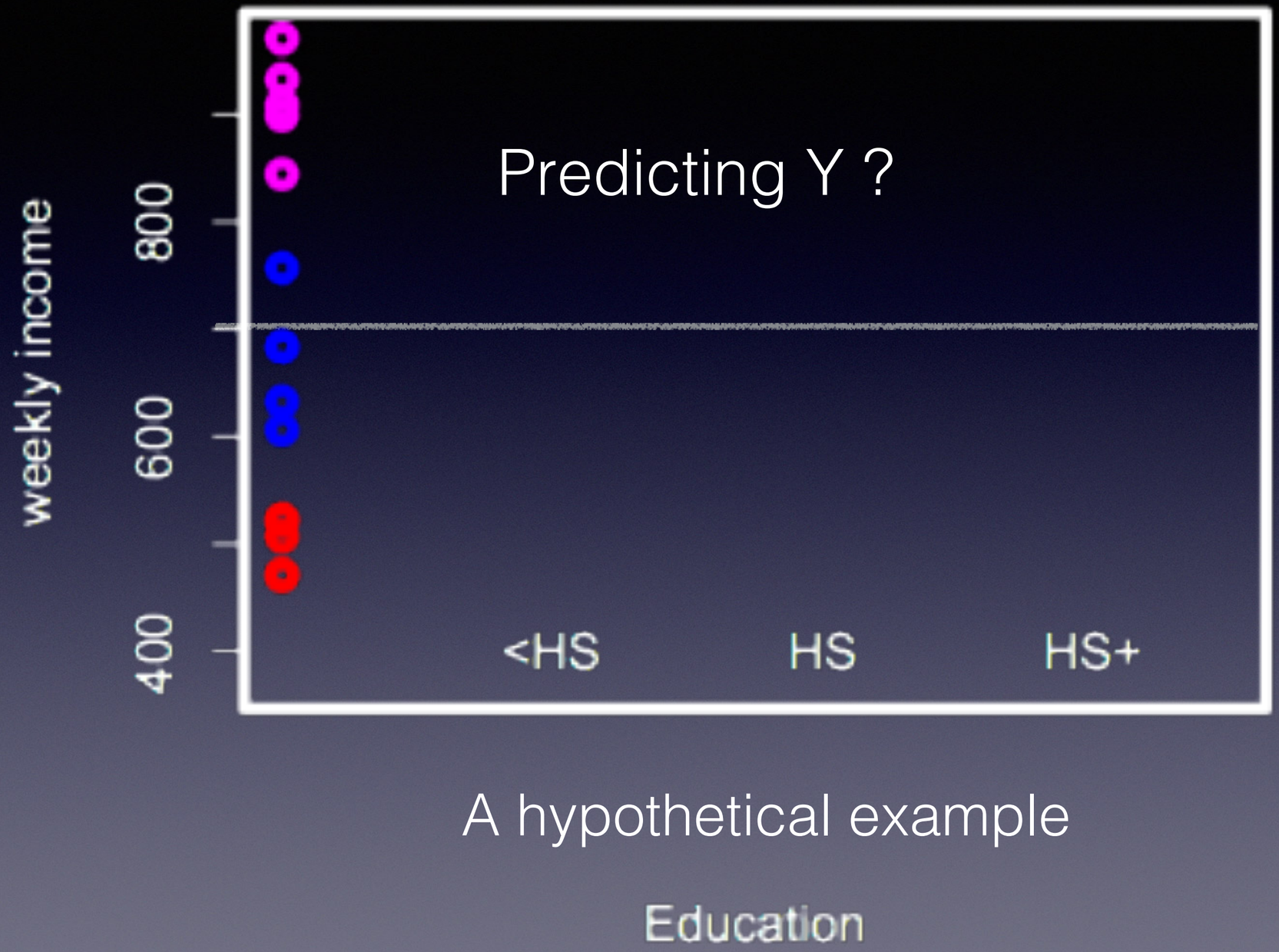
A hypothetical example

Education

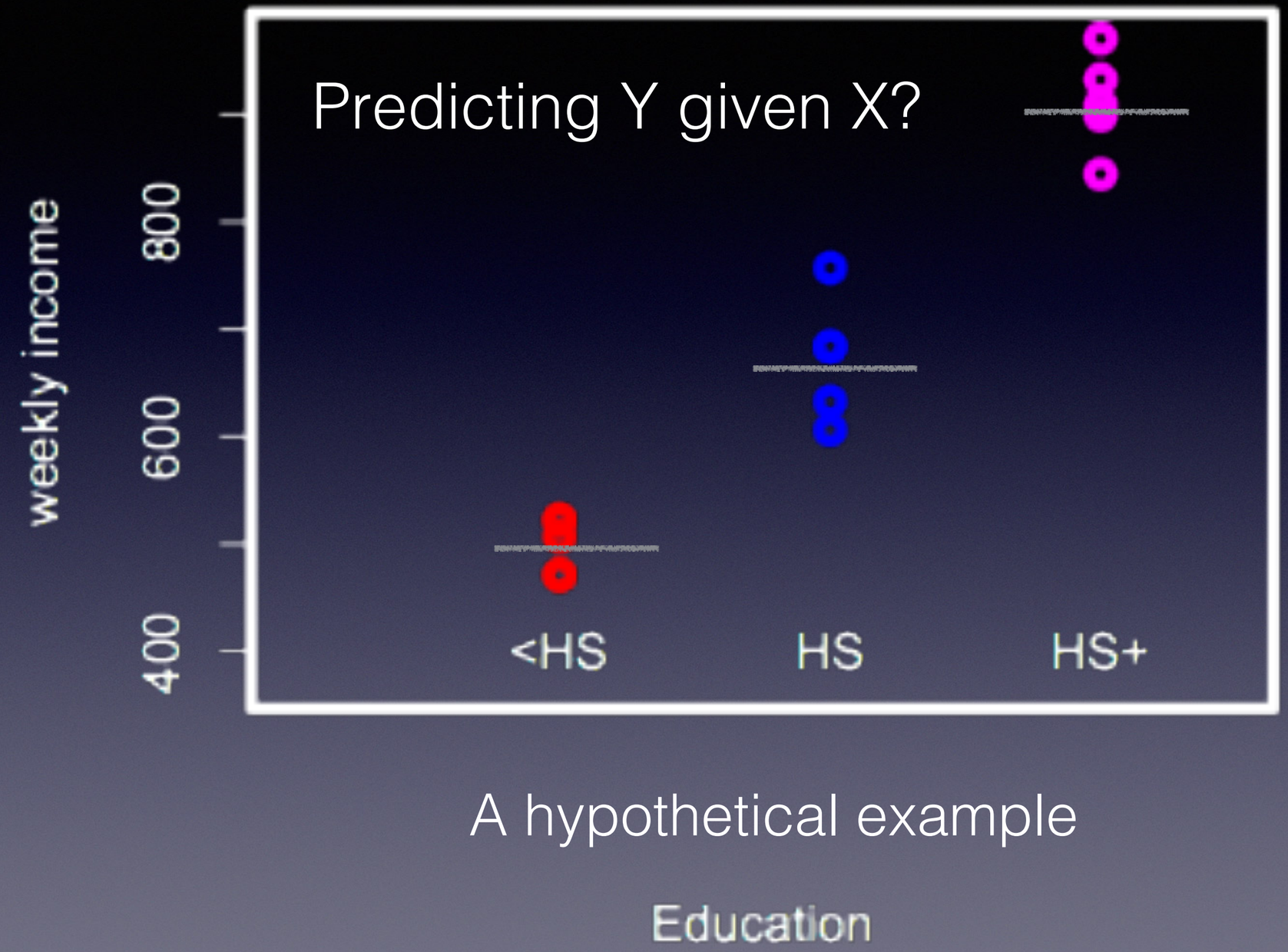
A hypothetical example

Analysis of variance (ANOVA)

- It concerns the total variation in Y (the quantitative variable) — Income.
- Considering a variable X, we would like to know how much variation in Y can be *explained* by X.
- Not necessarily a causal relation.

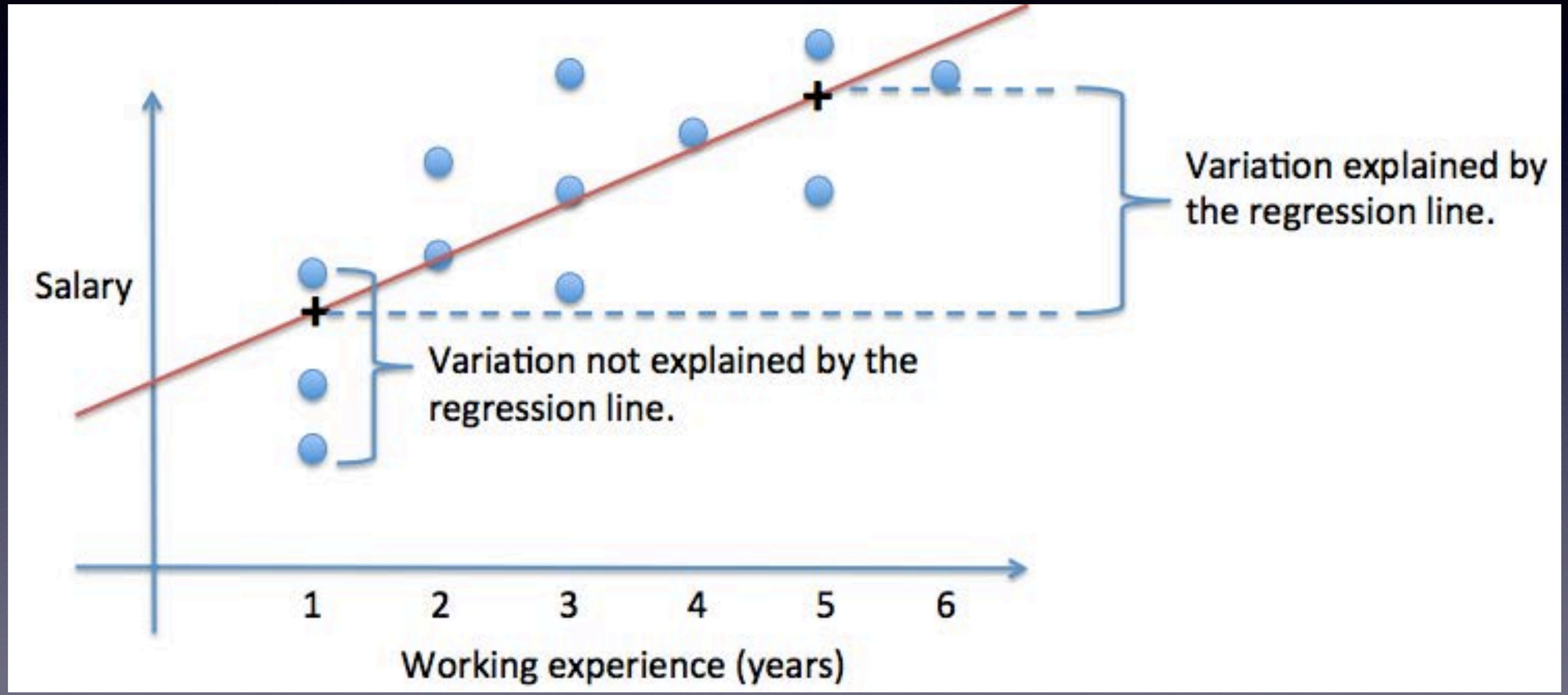


A hypothetical example

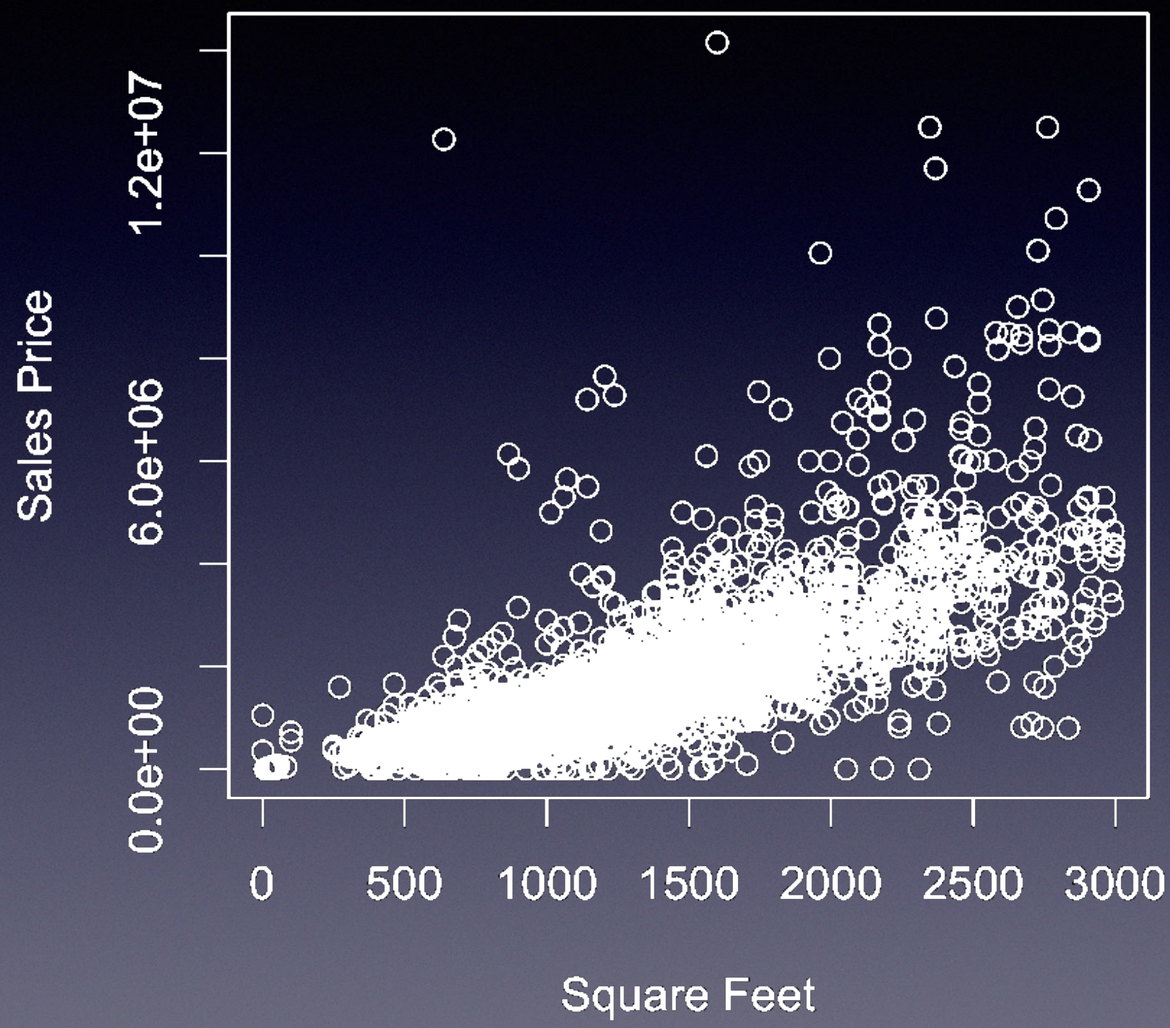


A hypothetical example

ANOVA and regression



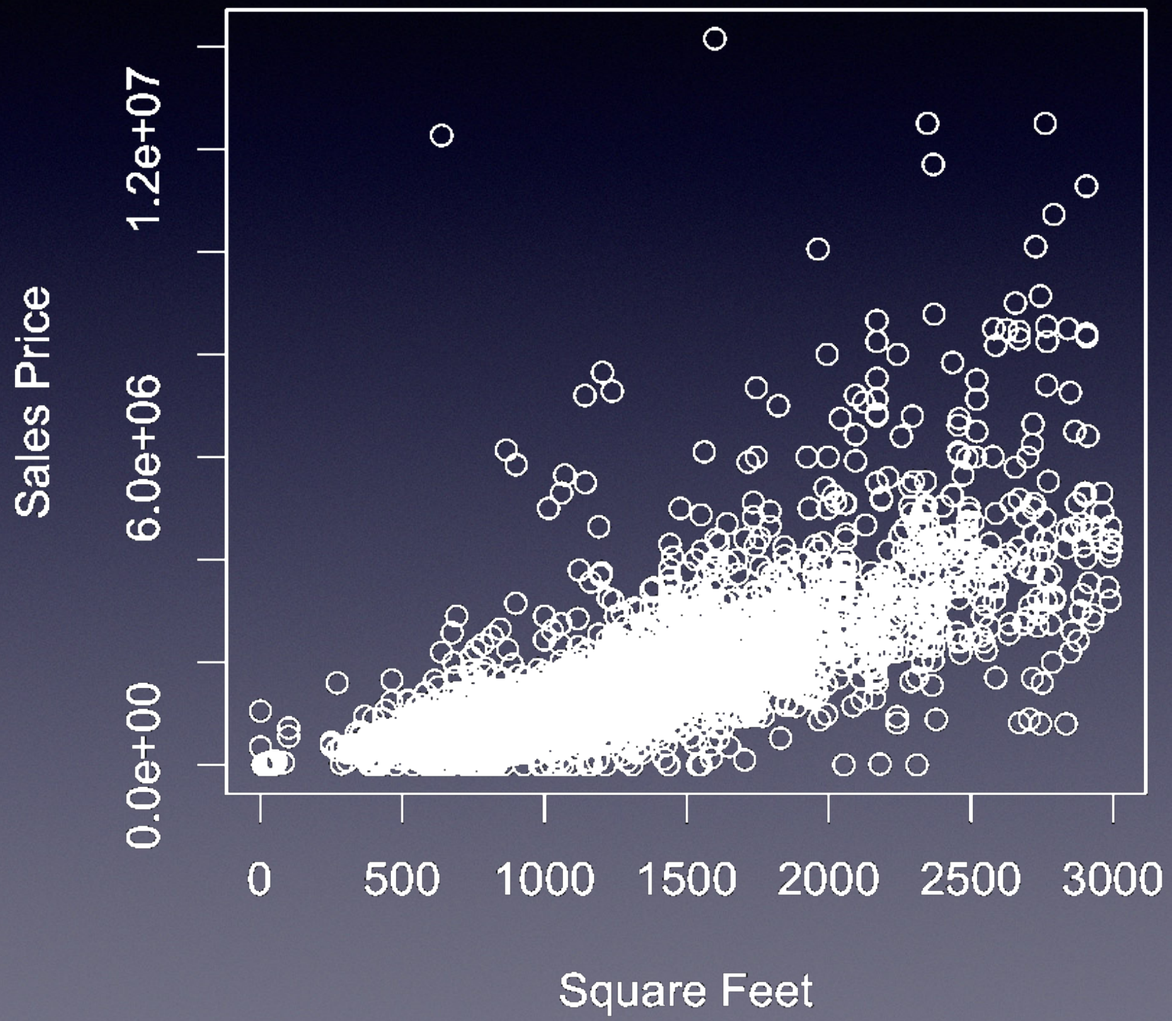
Manhattan Condo Prices



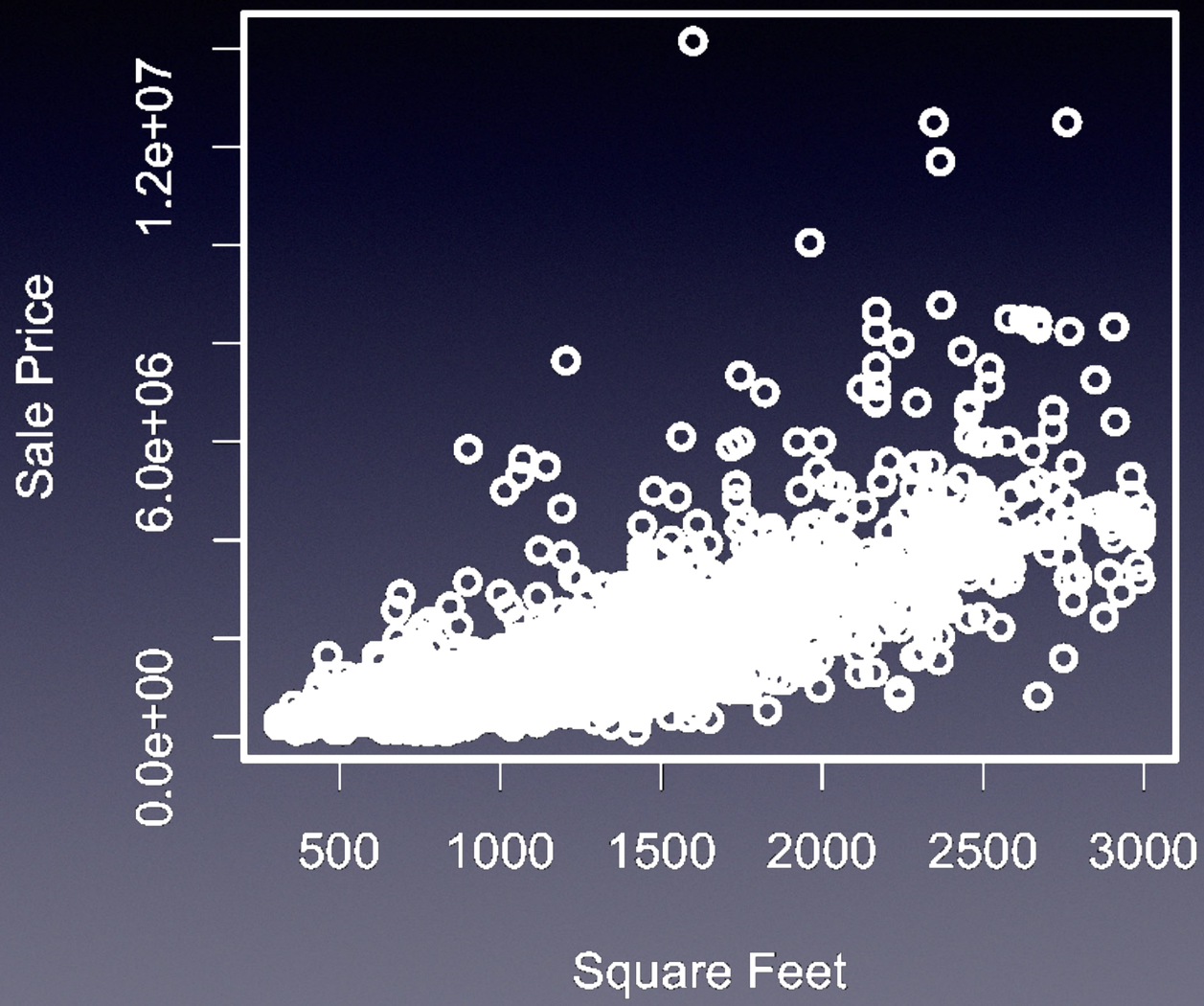
Manhattan Condo Prices

- From NYC Open Data
- Year 2009
- Condo building with elevator
- 3553 apartments

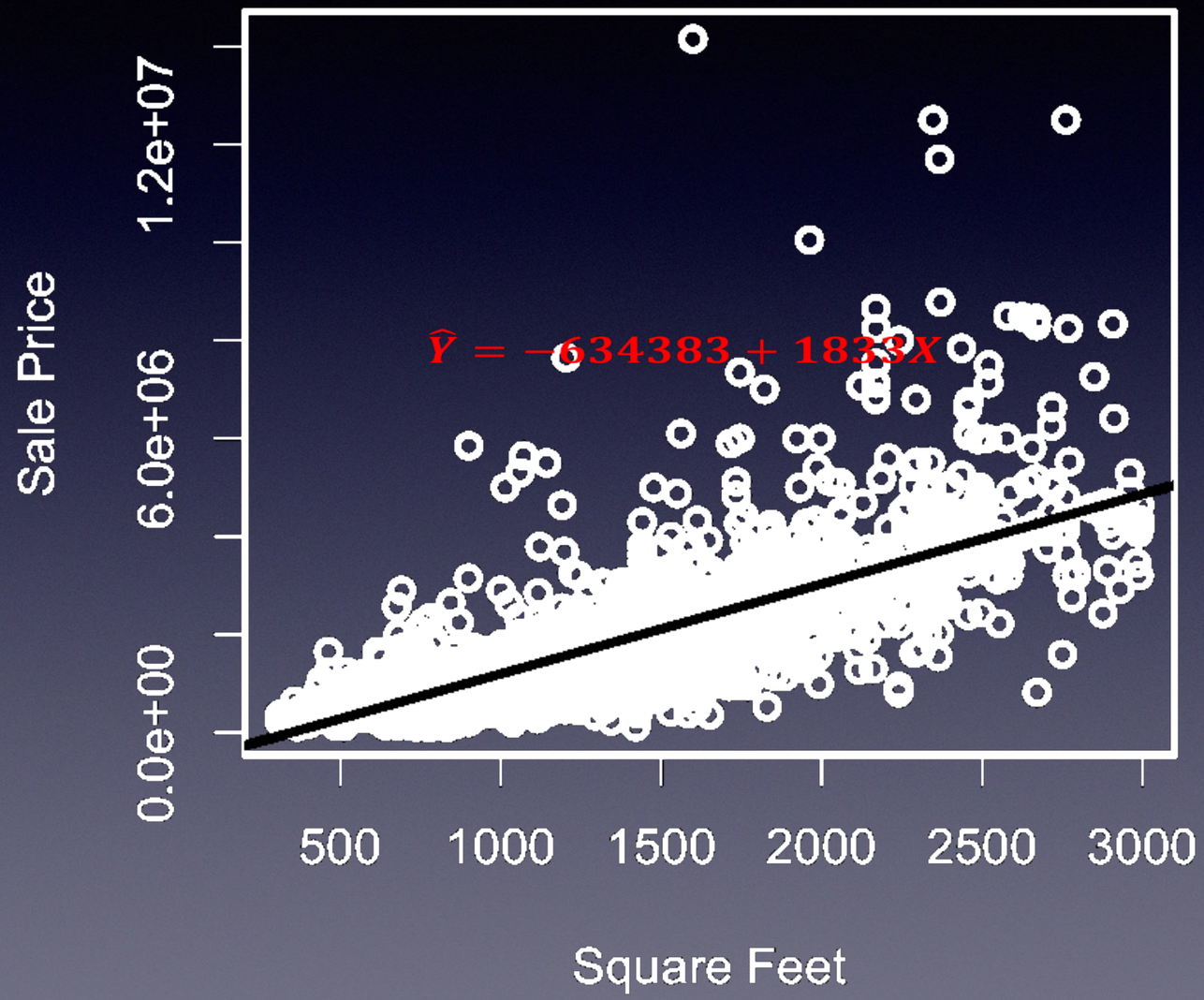
Manhattan Condo Prices



Manhattan Condo Prices

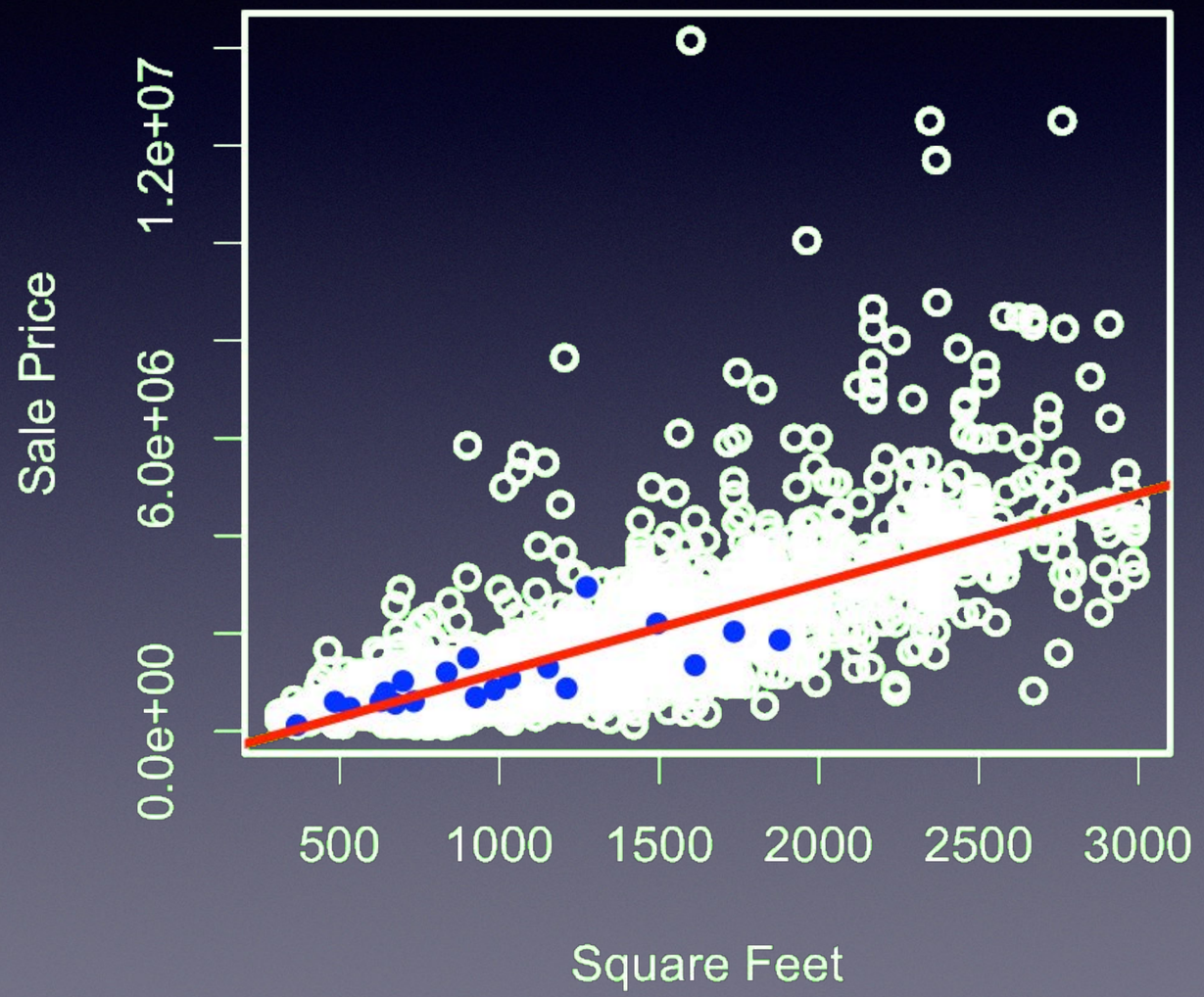


Manhattan Condo Prices

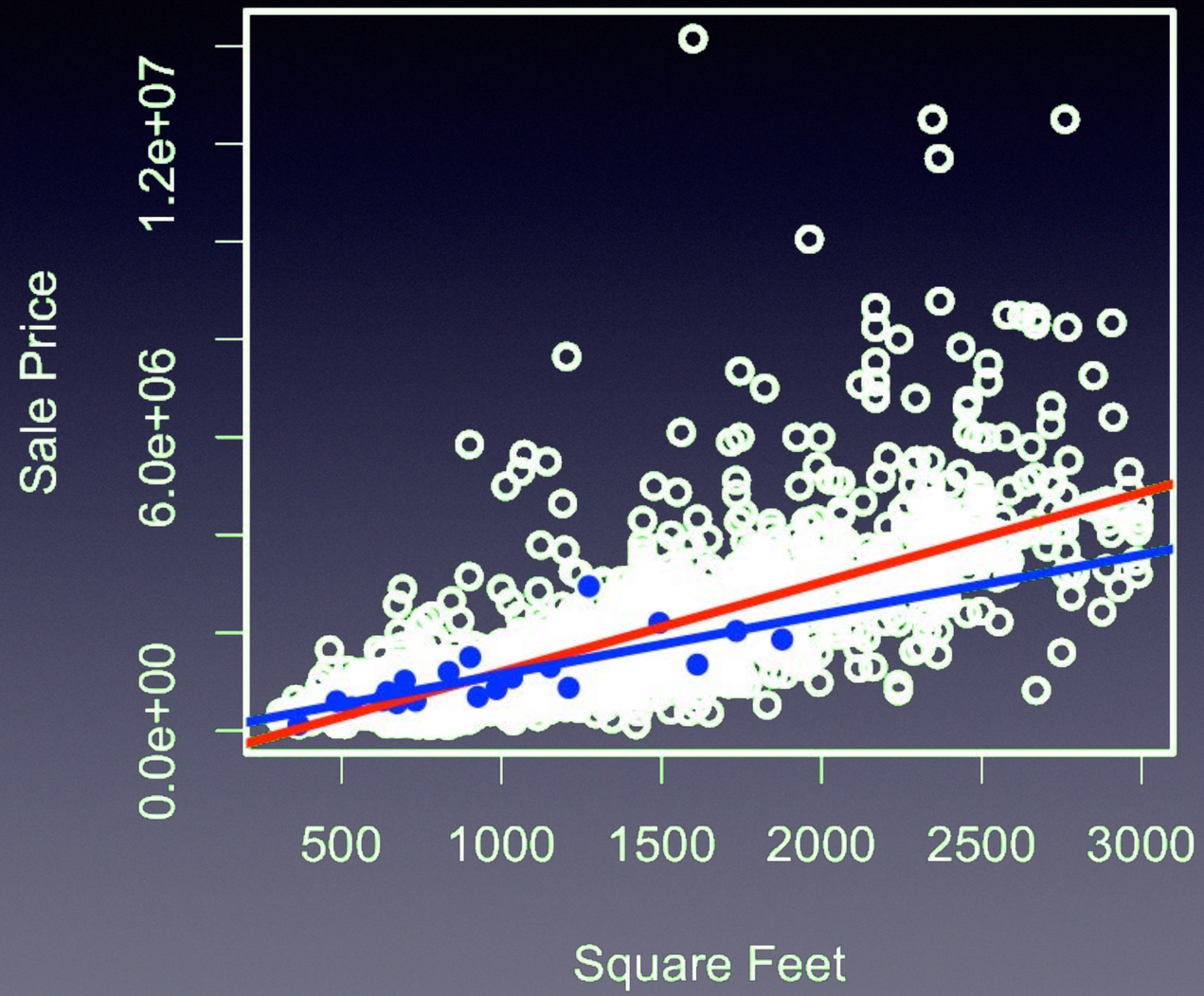


$R^2 = 0.63$

Sampling variability in regression estimates

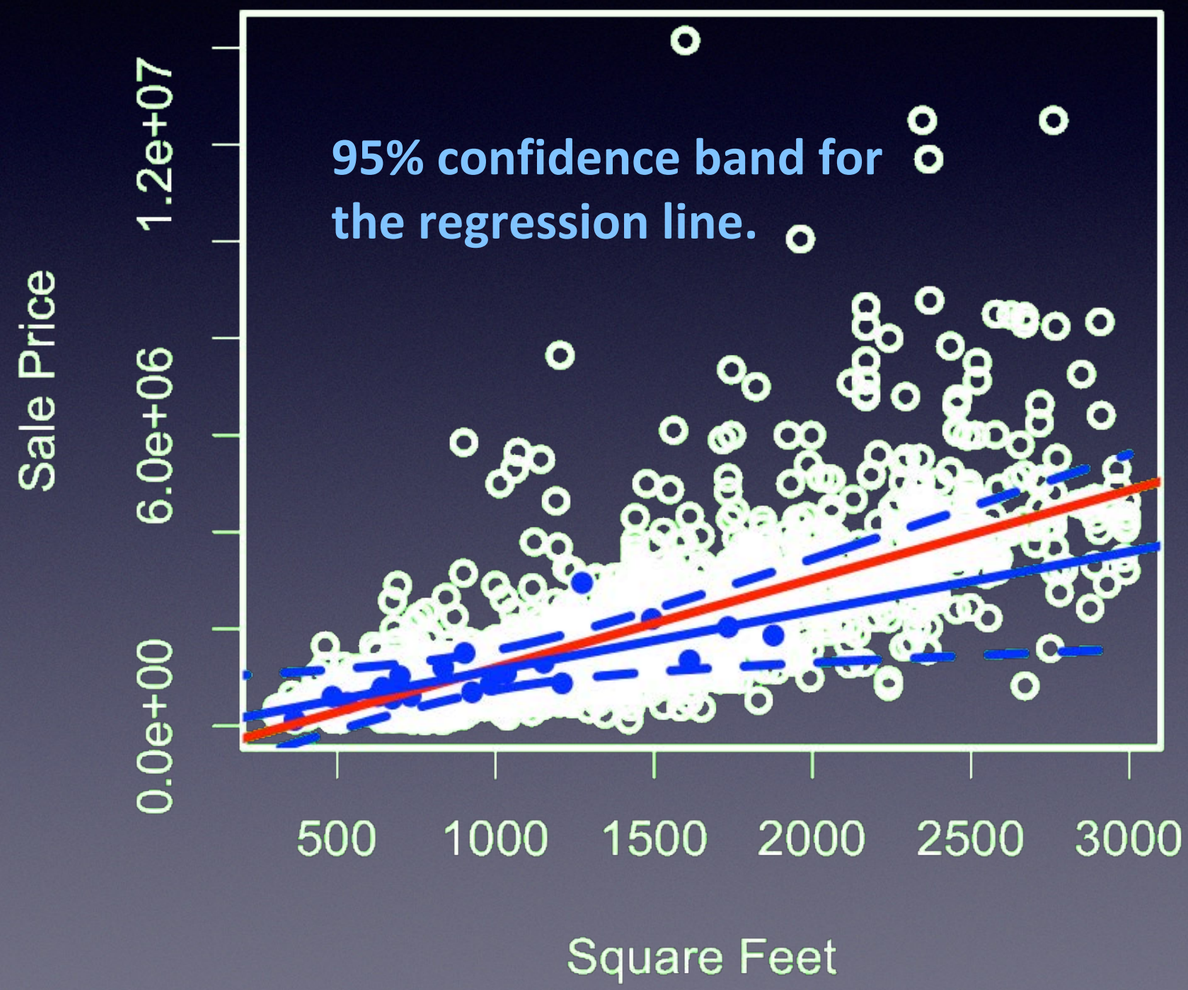


Sampling variability in regression estimates



Sampling variability in regression estimates

- **The confidence band centers at the sample estimate.**
- **It represents interval estimate for the regression line.**
- **Other inference on regression estimates can also be carried out.**



Ten Simple Rules for Effective Statistical Practice

Rule 1: Statistical Methods Should Enable Data to Answer Scientific Questions.

Rule 2: Signals Always Come with Noise.

Rule 3: Plan Ahead, Really Ahead.

Rule 4: Worry about Data Quality.

Rule 5: Statistical Analysis Is More Than a Set of Computations.

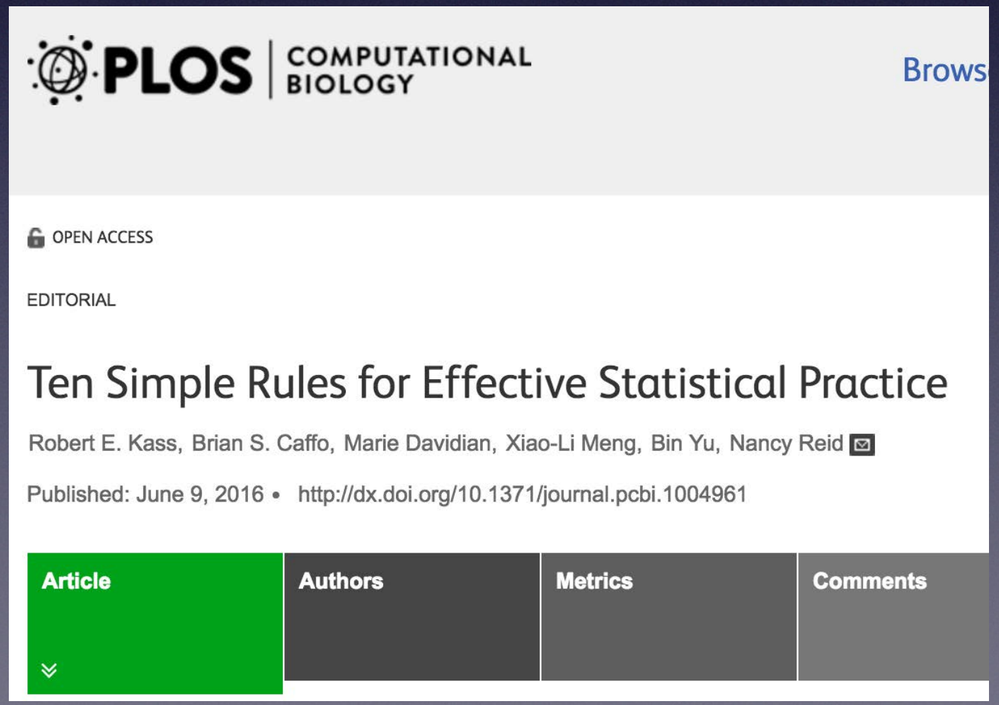
Rule 6: Keep it Simple.

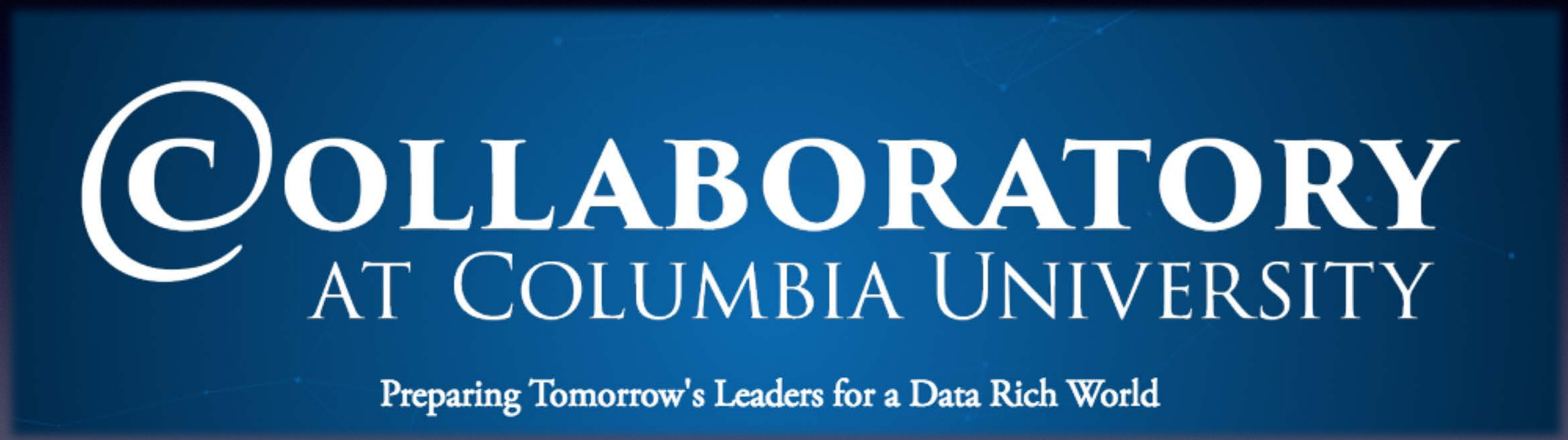
Rule 7: Provide Assessments of Variability.

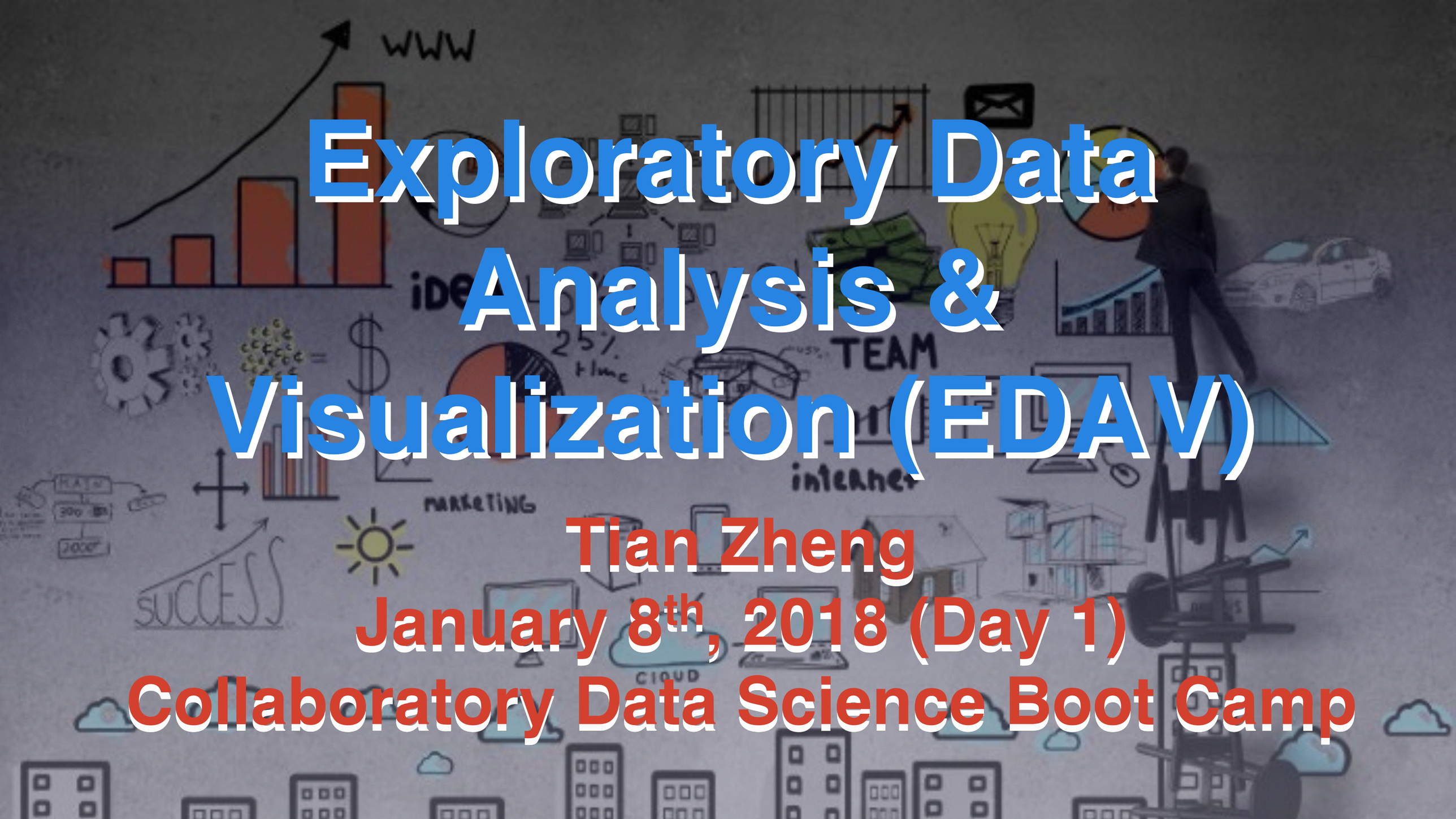
Rule 8: Check Your Assumptions.

Rule 9: When Possible, Replicate!

Rule 10: Make Your Analysis Reproducible.




@COLLABORATORY
AT COLUMBIA UNIVERSITY
Preparing Tomorrow's Leaders for a Data Rich World



Exploratory Data Analysis & Visualization (EDAV)

Tian Zheng

January 8th, 2018 (Day 1)

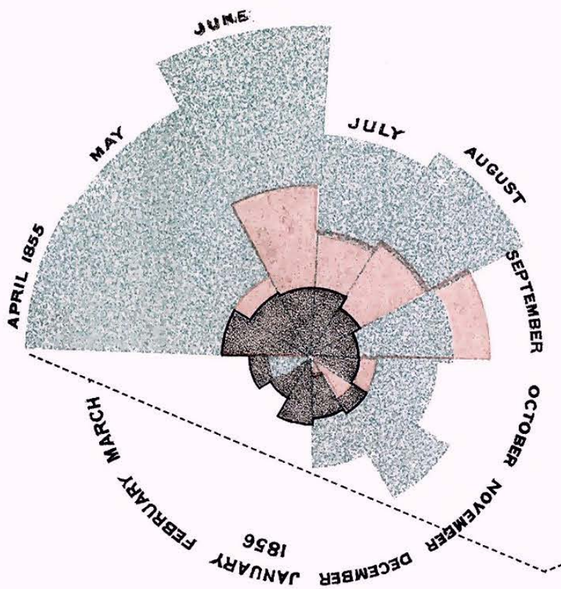
Collaboratory Data Science Boot Camp

Displaying categorial variable

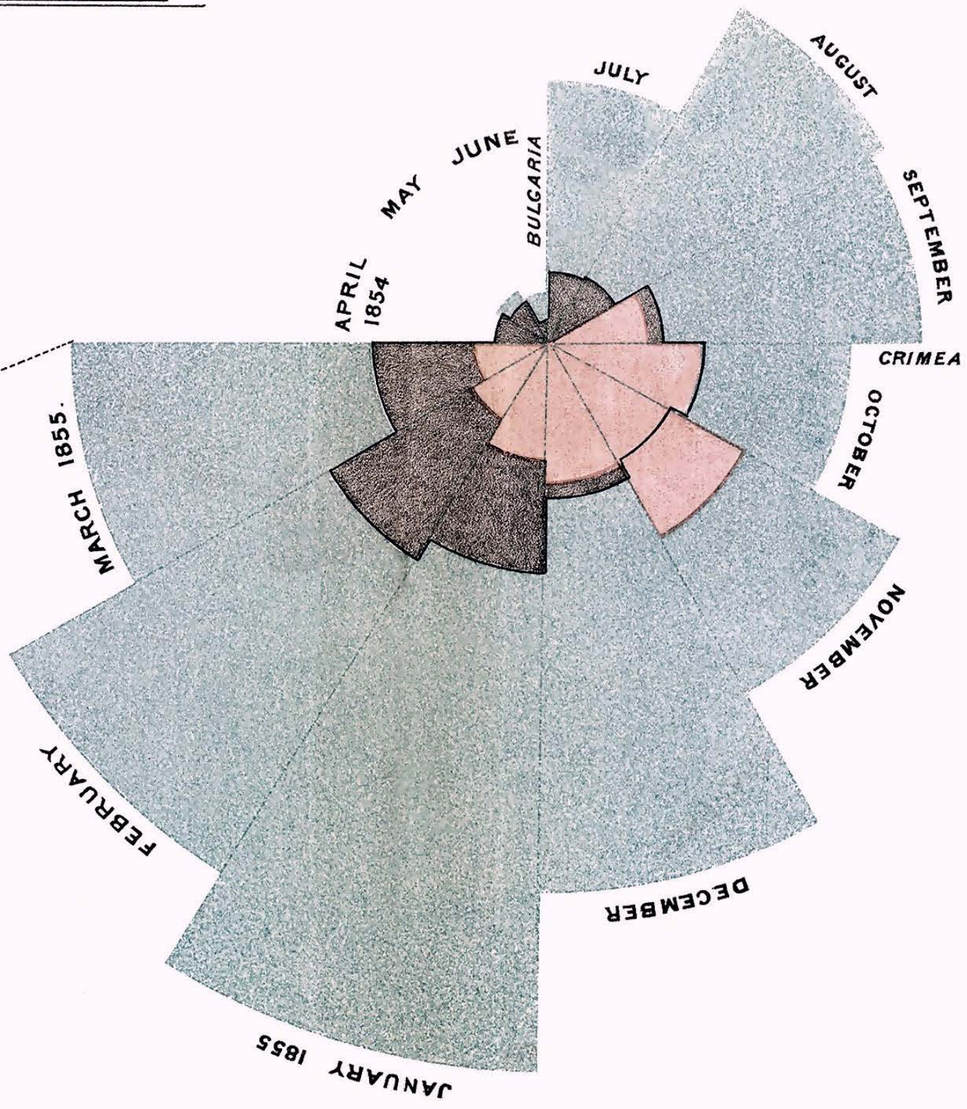
- For categorical variable, we summarize the data using the **counts** of observed occurrences of each value.
- Alternatively, we can use *percentage* or *proportion*.

DIAGRAM OF THE CAUSES OF MORTALITY IN THE ARMY IN THE EAST.

2.
APRIL 1855 TO MARCH 1856.



1.
APRIL 1854 TO MARCH 1855.



The Areas of the blue, red, & black wedges are each measured from the centre as the common vertex.

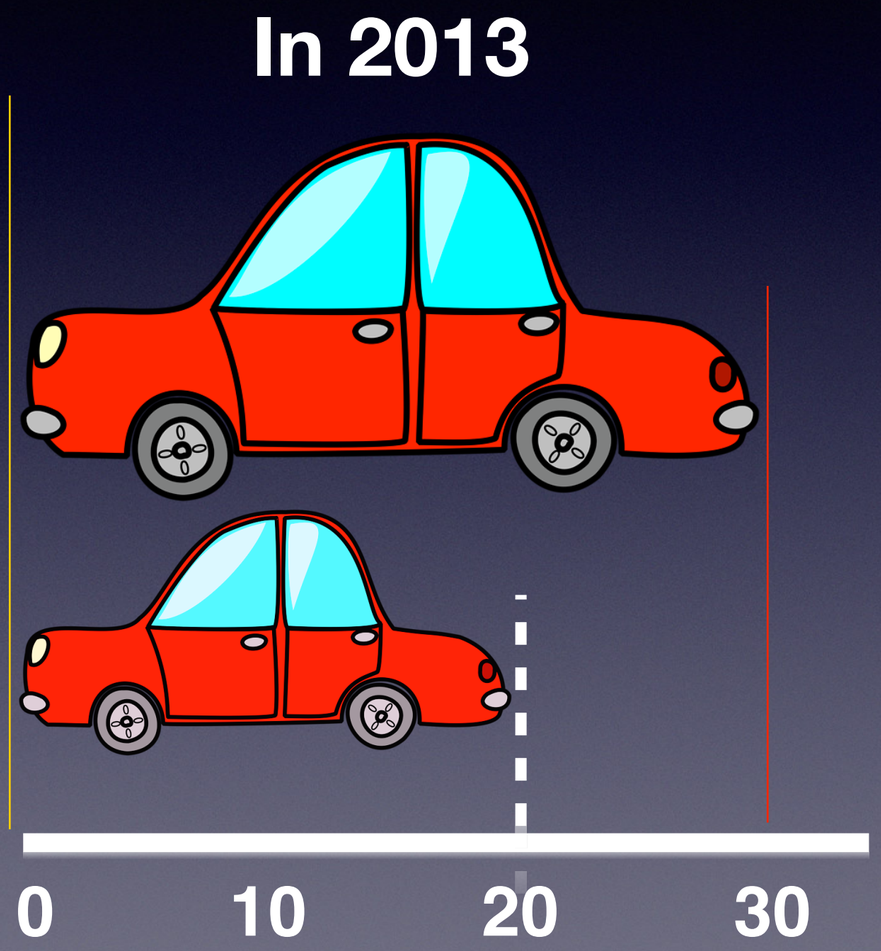
The blue wedges measured from the centre of the circle represent area for area the deaths from Preventible or Mitigable Zymotic diseases; the red wedges measured from the centre the deaths from wounds, & the black wedges measured from the centre the deaths from all other causes.

The black line across the red triangle in Nov.r 1854 marks the boundary of the deaths from all other causes during the month.

In October 1854, & April 1855, the black area coincides with the red; in January & February 1856, the blue coincides with the black.

The entire areas may be compared by following the blue, the red & the black lines enclosing them.

Area principle

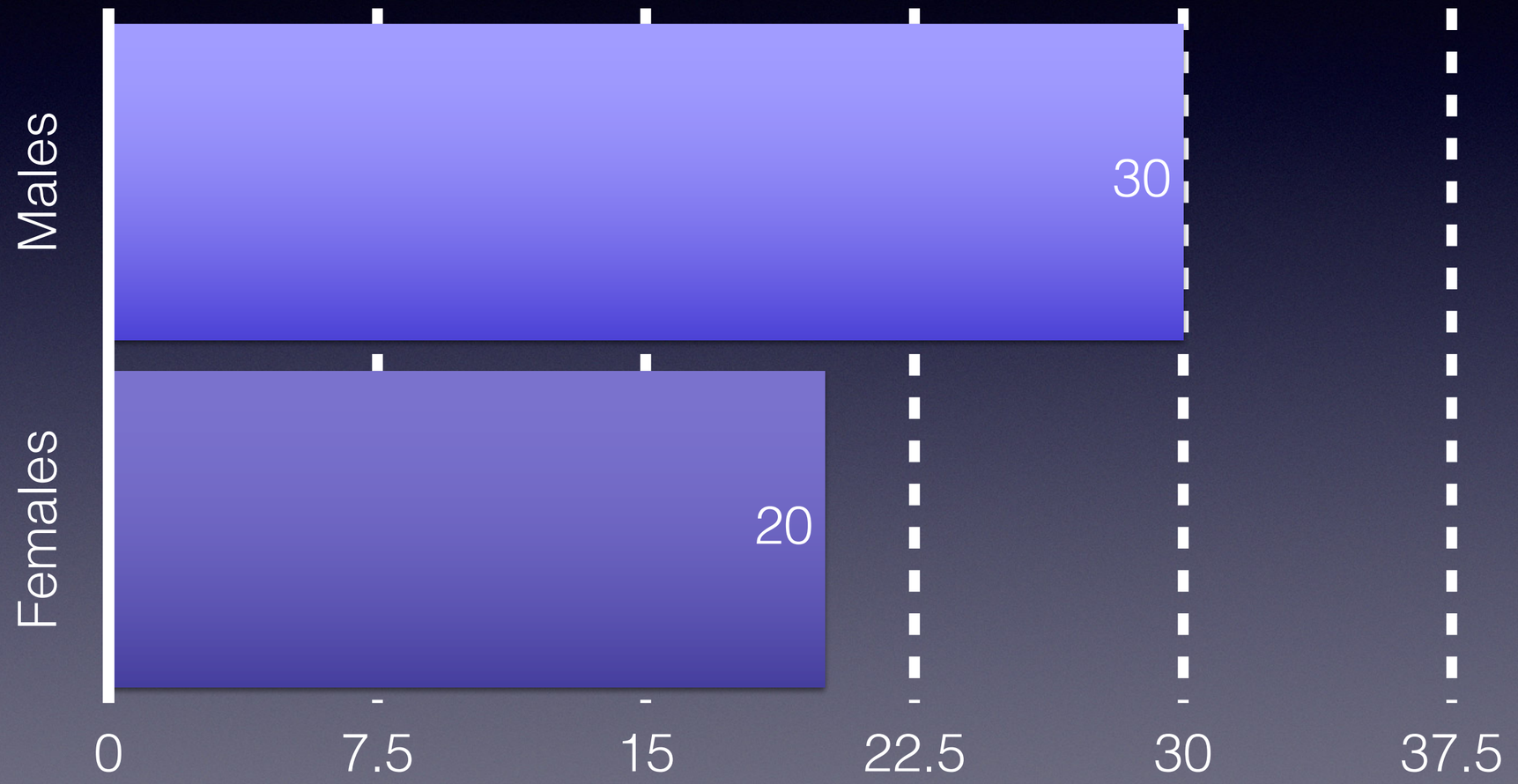


The area principle — the size of the area correlates with the data summaries.

~ 30% of accidental deaths of males were due to automobile accidents.

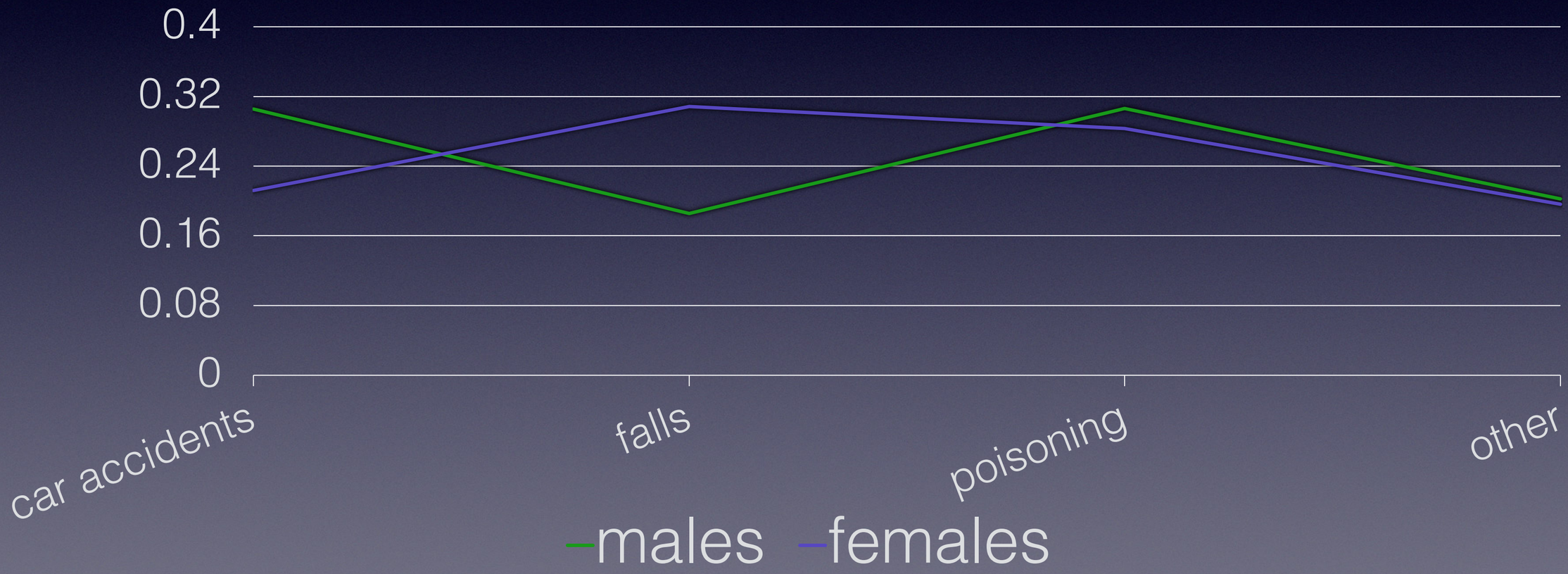
~ 20% of accidental deaths of females were due to automobile accidents.

Area principle



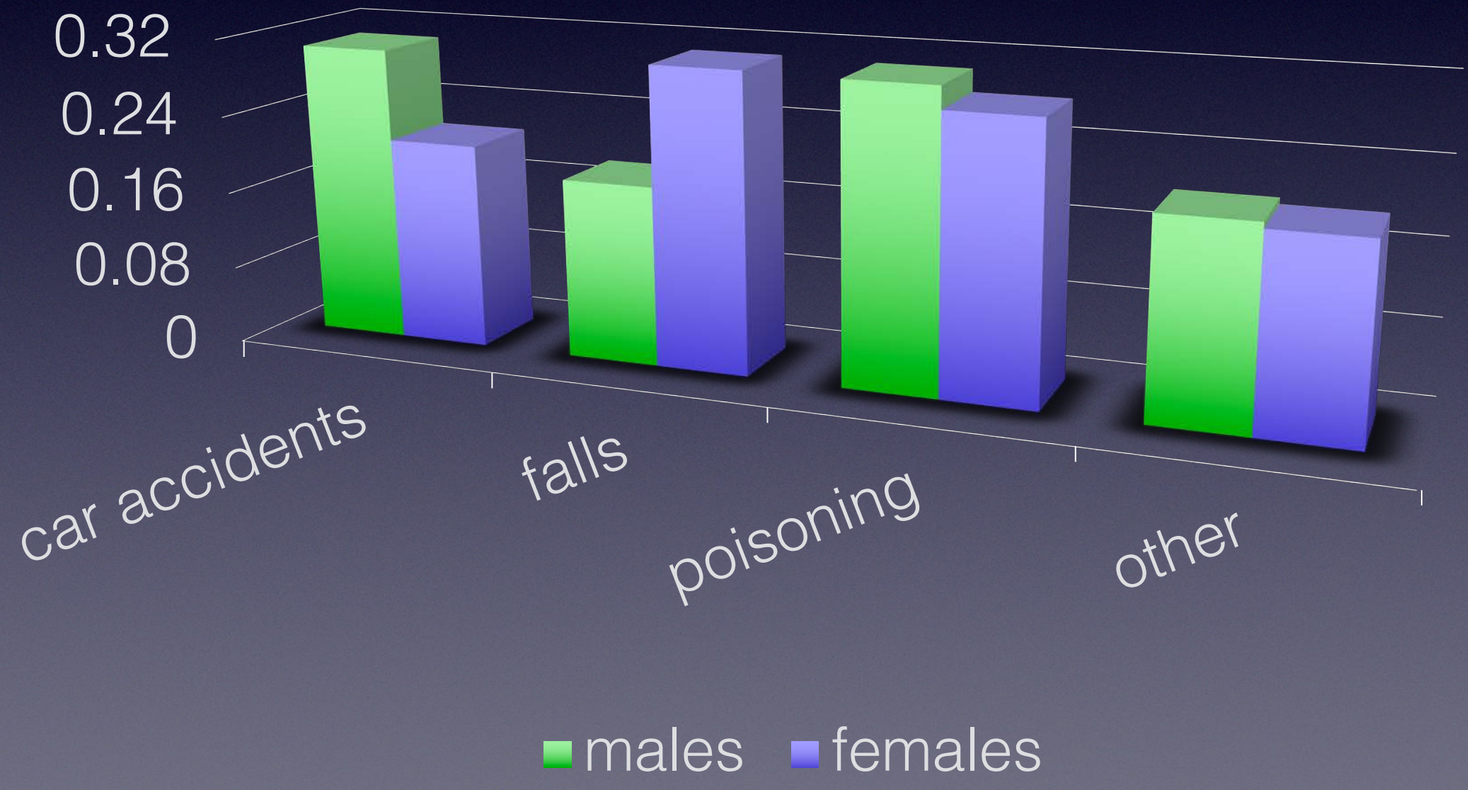
Side by side comparison

Accidental Deaths in 2013 by gender and by different causes (proportions)



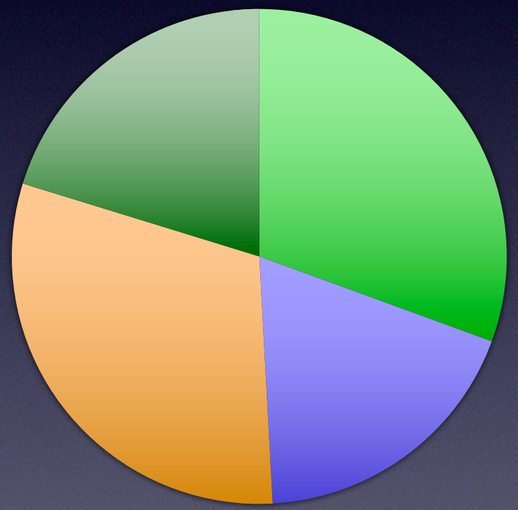
Side by side comparison

Accidental Deaths in 2013 by gender and by different causes (proportions)

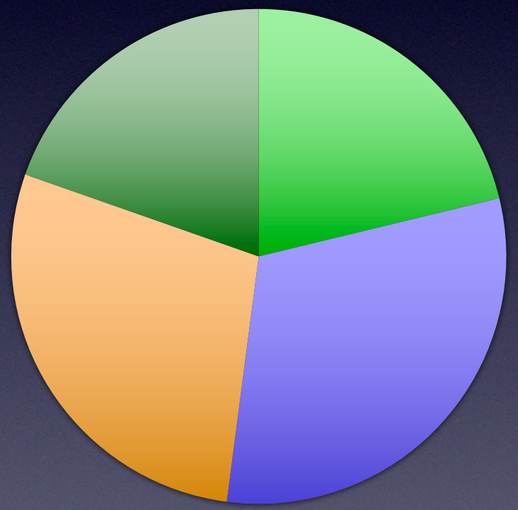


Side by side comparison

Men



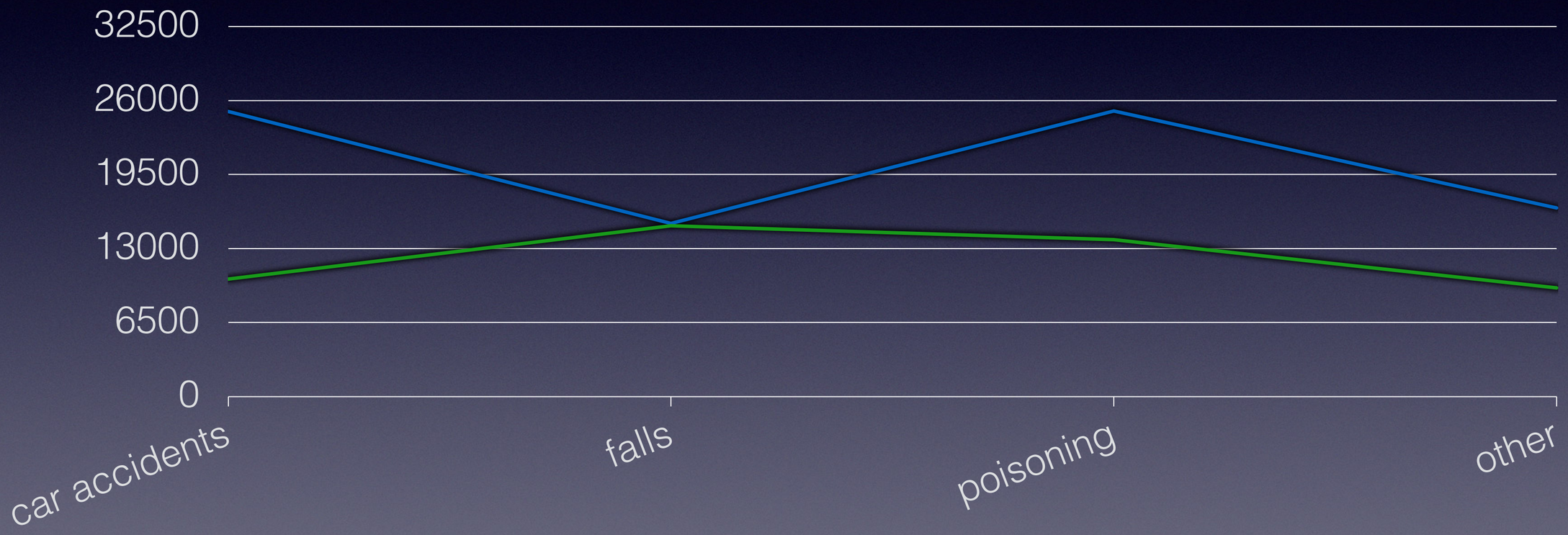
Women



car accidents
falls
poisoning
other

Side by side comparison

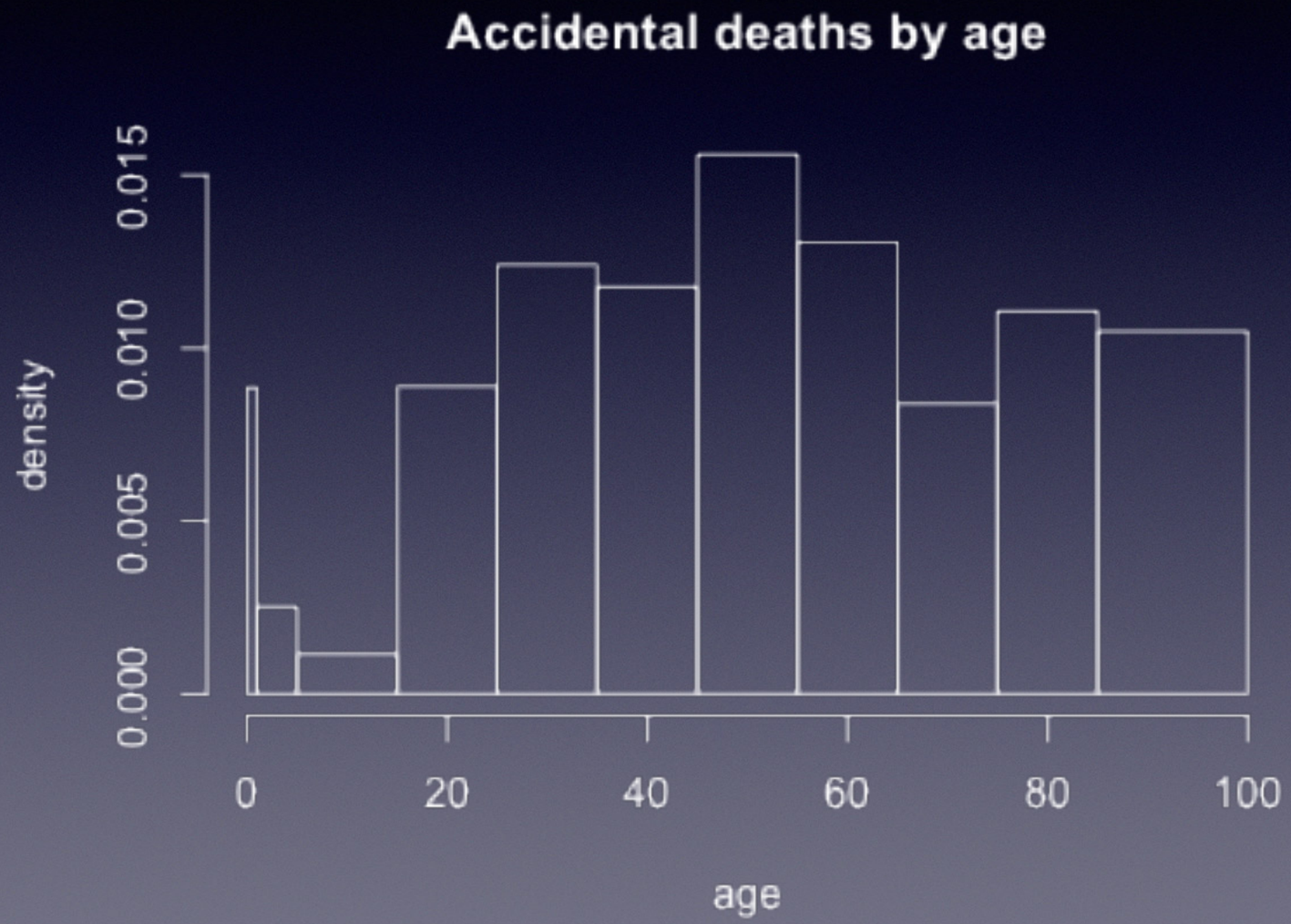
Accidental Deaths in 2013 by gender and by different causes (raw counts)



Displaying quantitative variable

- For quantitative variables, we also summarize the data using **the counts of observed occurrences of values.**
- Different from categorical variables, we may count occurrences **within intervals** rather than individual values.
- We also use percentage or proportion.

Displaying quantitative variable

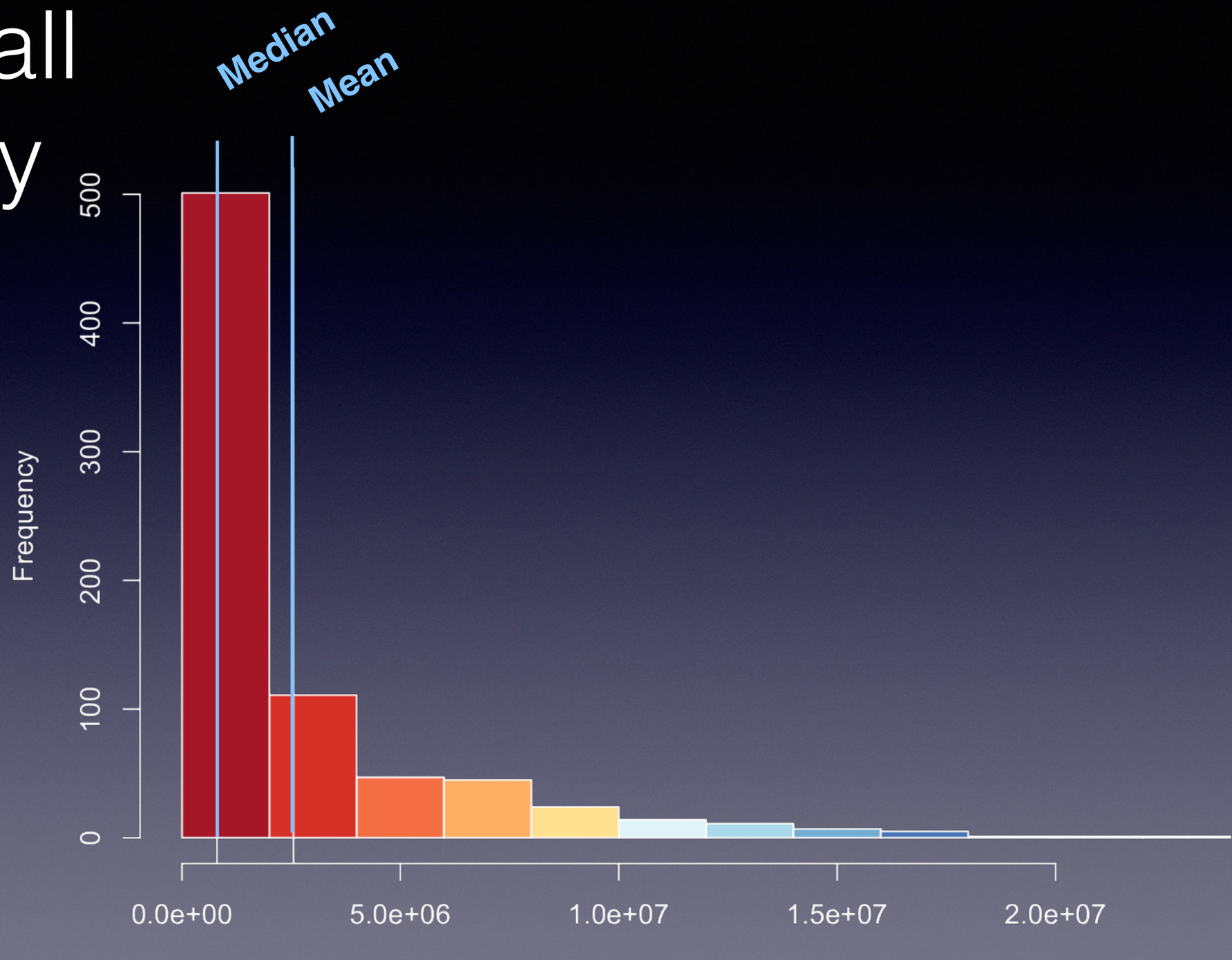


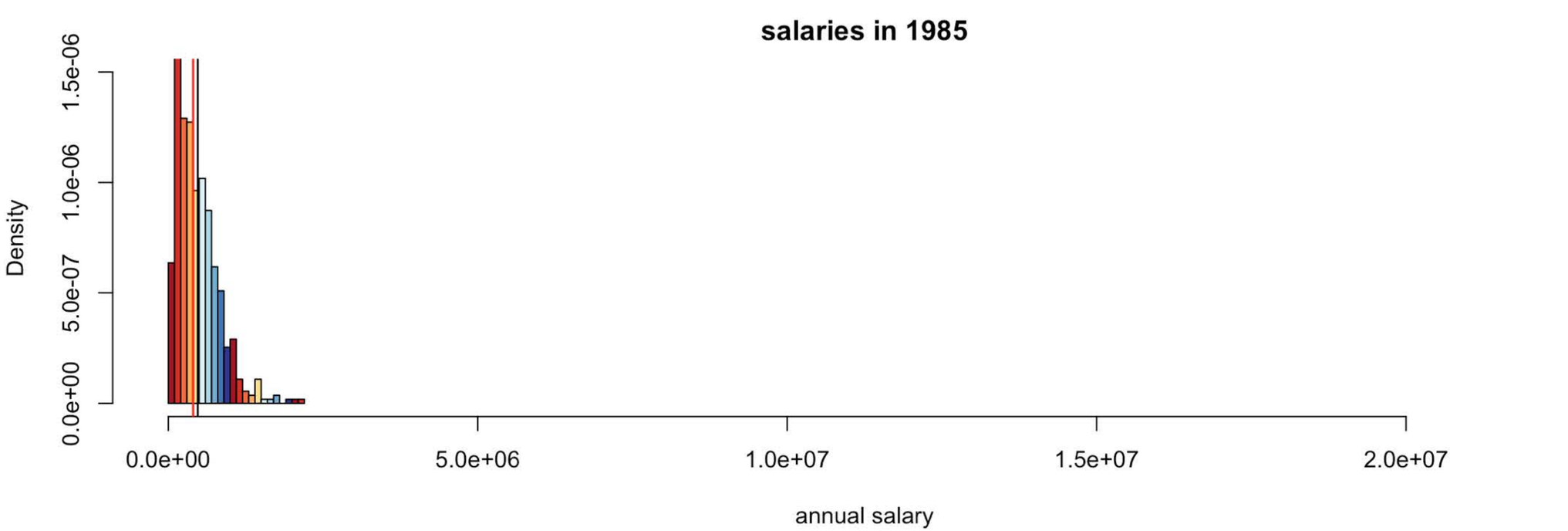
Center of Variation

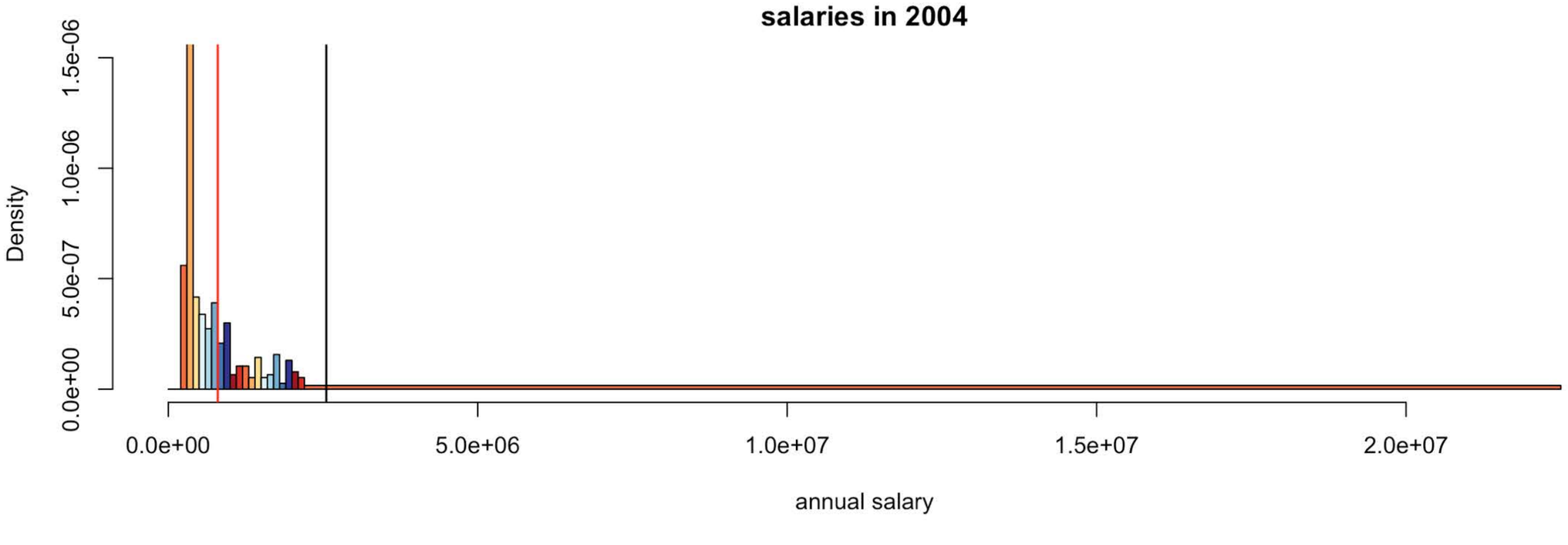
- Summarizing center of variation:
  - **mean** (numerical average)
  - **median** (mid-point)

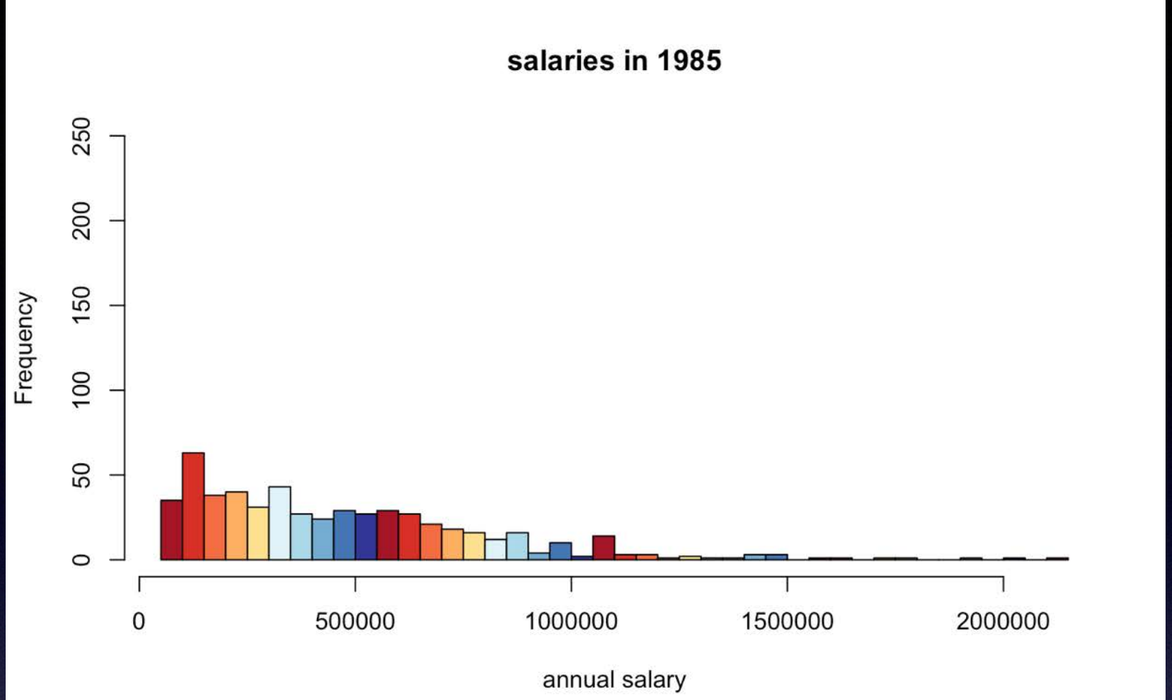
- When the data come with **a few extremely large values**
  - mean is more affected by them than median.
  - Sensitive to outliers.

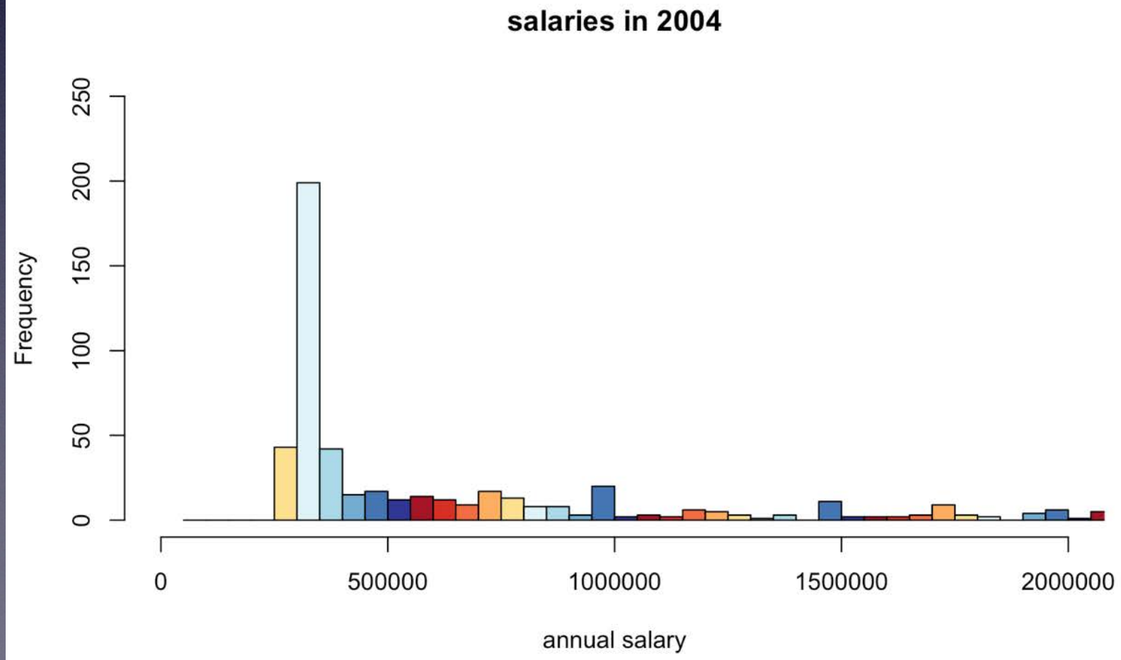
MLB Baseball Player Salary 2004




salaries in 1985
Density
0.0e+00
5.0e-07
1.0e-06
1.5e-06
0.0e+00
5.0e+06
1.0e+07
1.5e+07
2.0e+07
annual salary


salaries in 2004
Density
0.0e+00
5.0e-07
1.0e-06
1.5e-06
0.0e+00
5.0e+06
1.0e+07
1.5e+07
2.0e+07
annual salary

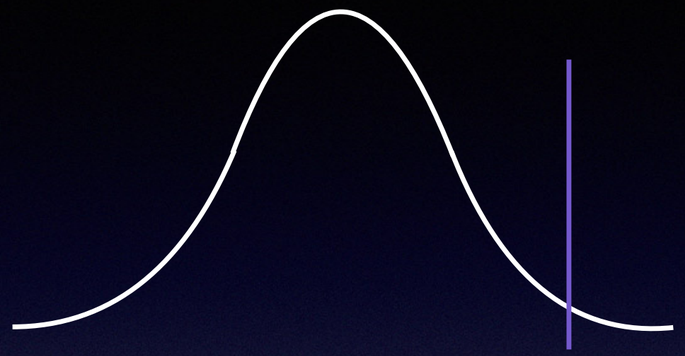

salaries in 1985
Frequency
0
50
100
150
200
250
0
500000
1000000
1500000
2000000
annual salary


salaries in 2004
Frequency
0
50
100
150
200
250
0
500000
1000000
1500000
2000000
annual salary

Summarizing variation
Standard Deviation

- For multiple observed values, **variation** means their **deviation from their center**.
- **Standard deviation**
  - deviation from the **sample mean**
  - "the average squared deviation".
- Standard deviation is a parameter for **normal distributions (the bell curves).**
- It is used as a “**yard stick**” for variation.
- It ***standardizes*** variation.

Standard deviation as a yard stick



- Luis is making 25K a year.
  The income in his city has a mean 20K and standard deviation of 4K.

Luis is 1.25 standard deviation ABOVE the mean in his city

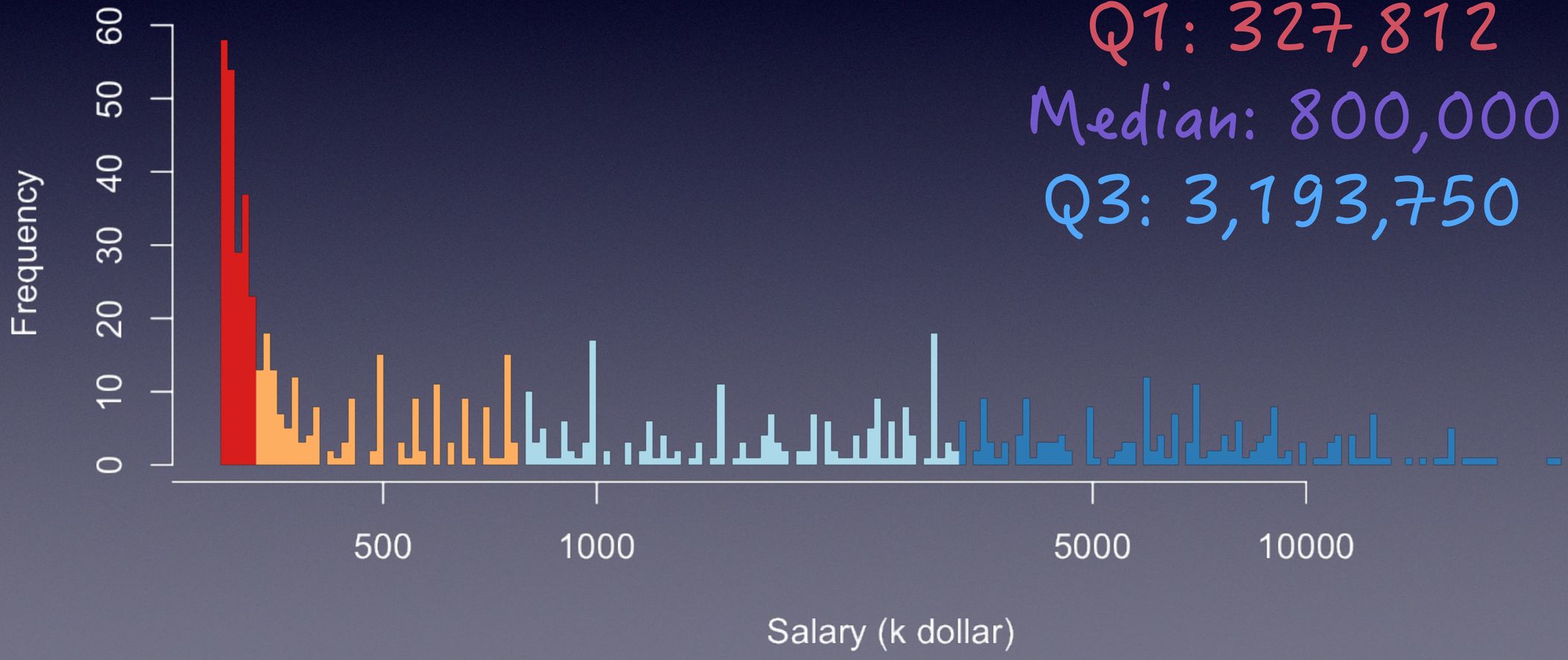
- Miles is also making 25K a year.
  The income in his city has a mean 30K and a standard deviation of 5K.

Miles is 1 standard deviation BELOW the mean in his city

Summarizing variation - Quantiles

- Quantiles (or percentile): a value threshold of a quantity that is defined to have a percent of data below it.
  - SAT critical reading, a score 600 is the 79th percentile.
- A set of special percentiles are called **quartiles**, which corresponds to 25%, 50% and 75% percentiles.
  - Quartiles divide data into quarters.

MLB Baseball Player Salary 2004



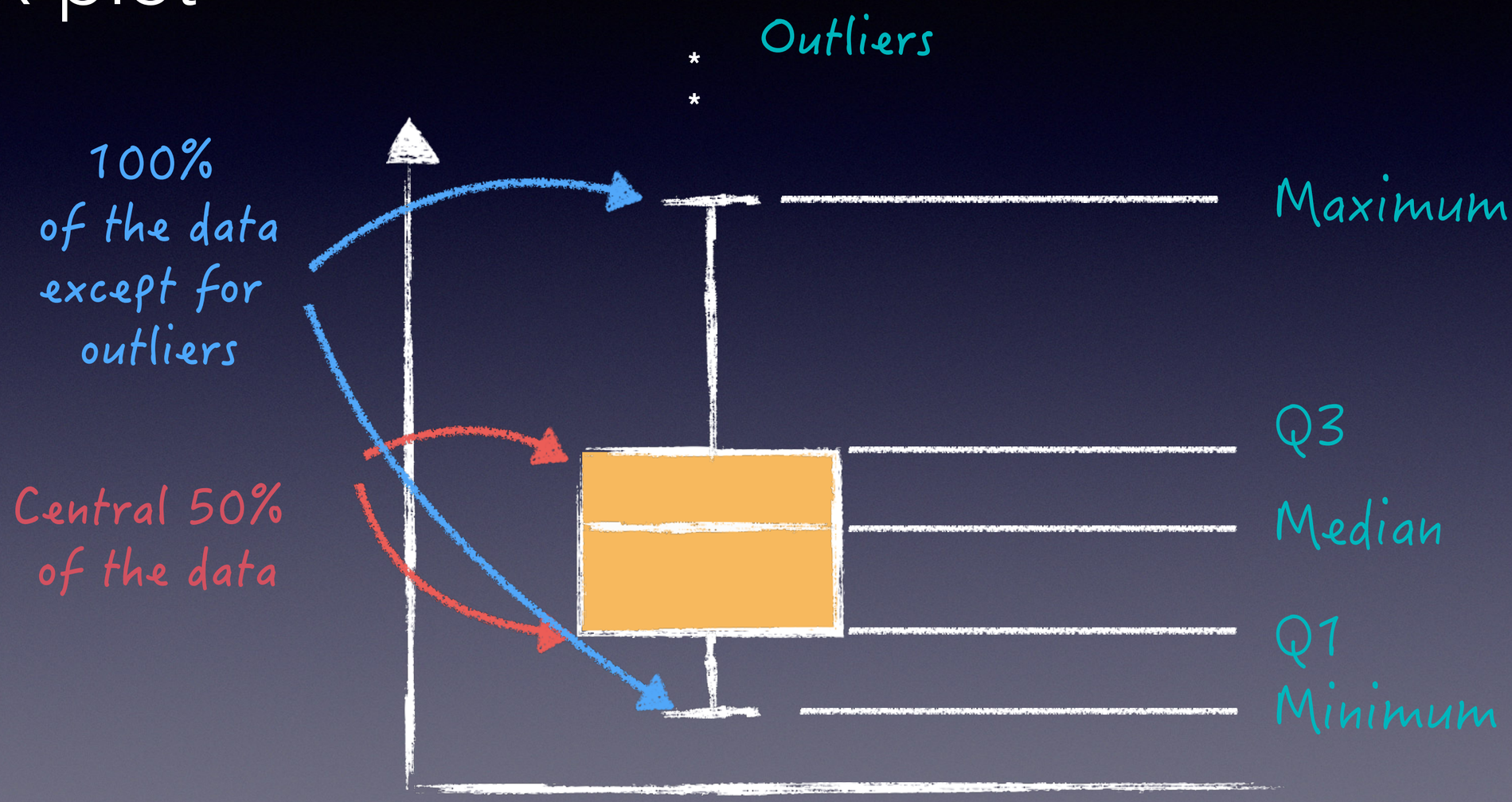
Quartiles:

Q1: 327,812

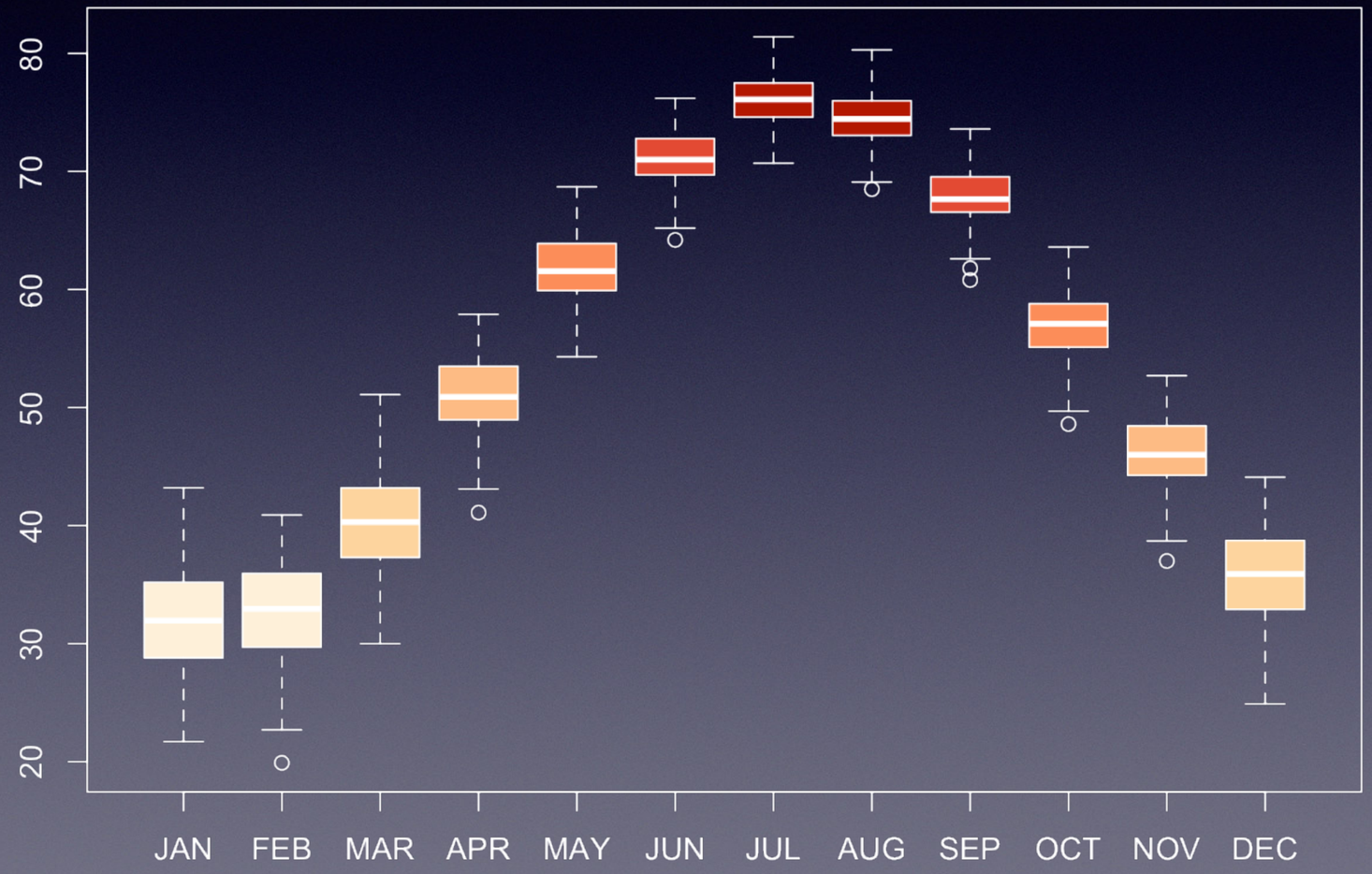
Median: 800,000

Q3: 3,193,750

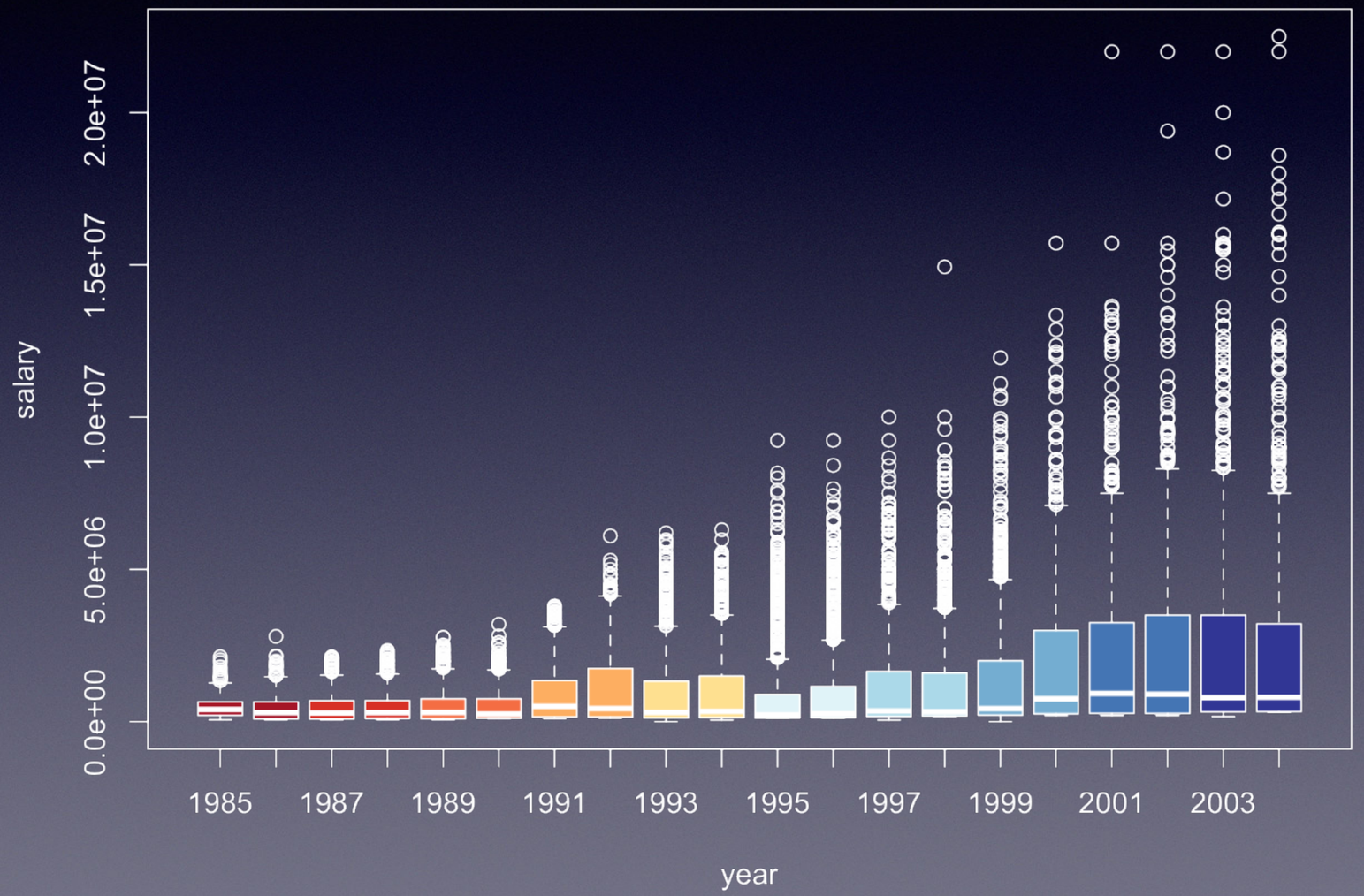
Box plot


Outliers
100%
of the data
except for
outliers
Central 50%
of the data
Maximum
Q3
Median
Q1
Minimum

NYC Monthly Average Temperature 1869-2012



MLB Salary 1985-2004



Describing Association

Association

Certain values of one trait
are **observed more frequently** with
Certain values of another trait.

Association: categorical variables

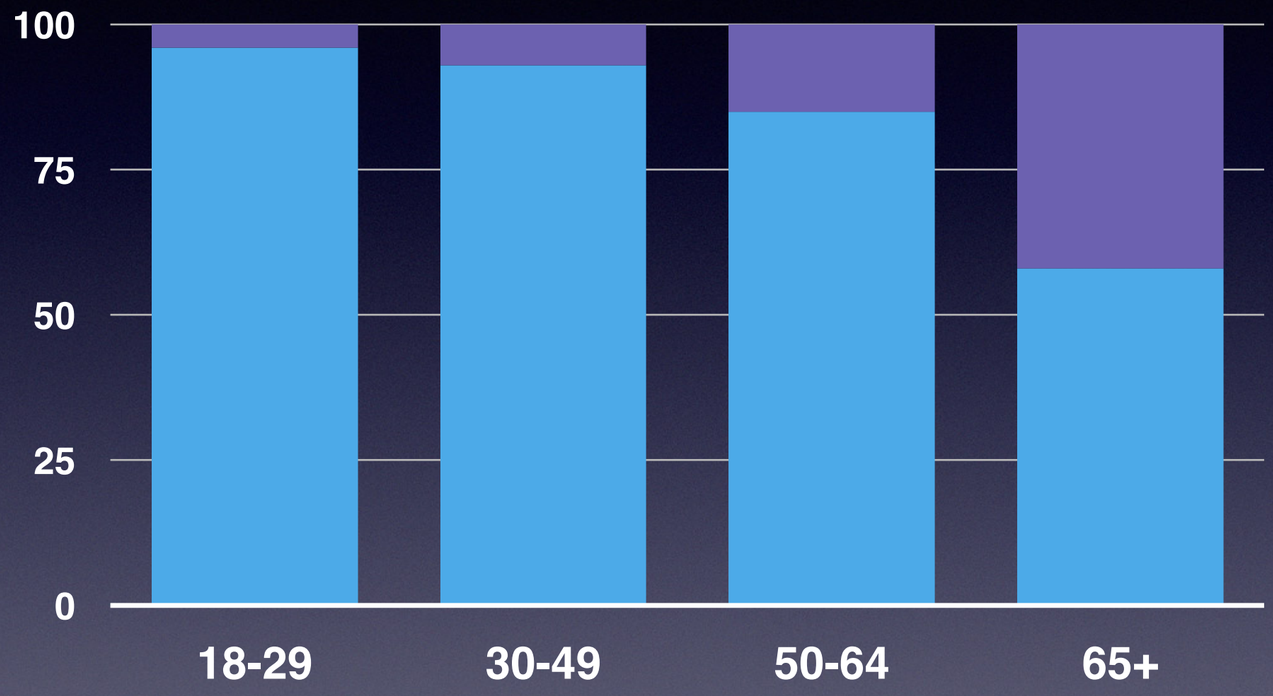
The association of two categorical variables is summarized using counts of **joint occurrences**.

Age	Internet Use: Use Internet	Internet Use: Do not use
18-29	48	2
30-49	93	7
50-64	85	5
65+	29	21

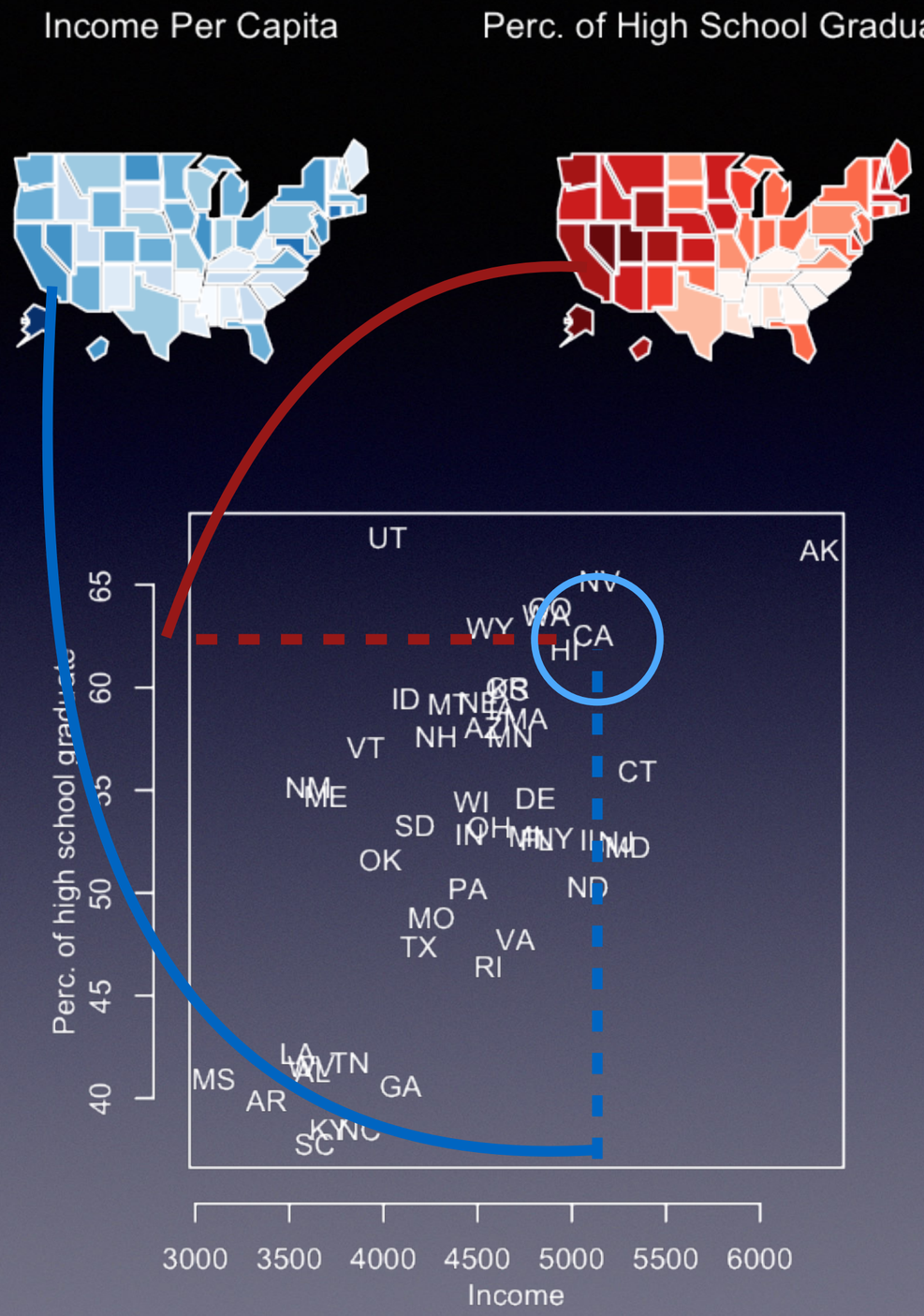
Hypothetical example based on findings from Pew Research Center on Internet Use (August 2012)

Internet Use versus Age

Display conditional frequencies (proportions)



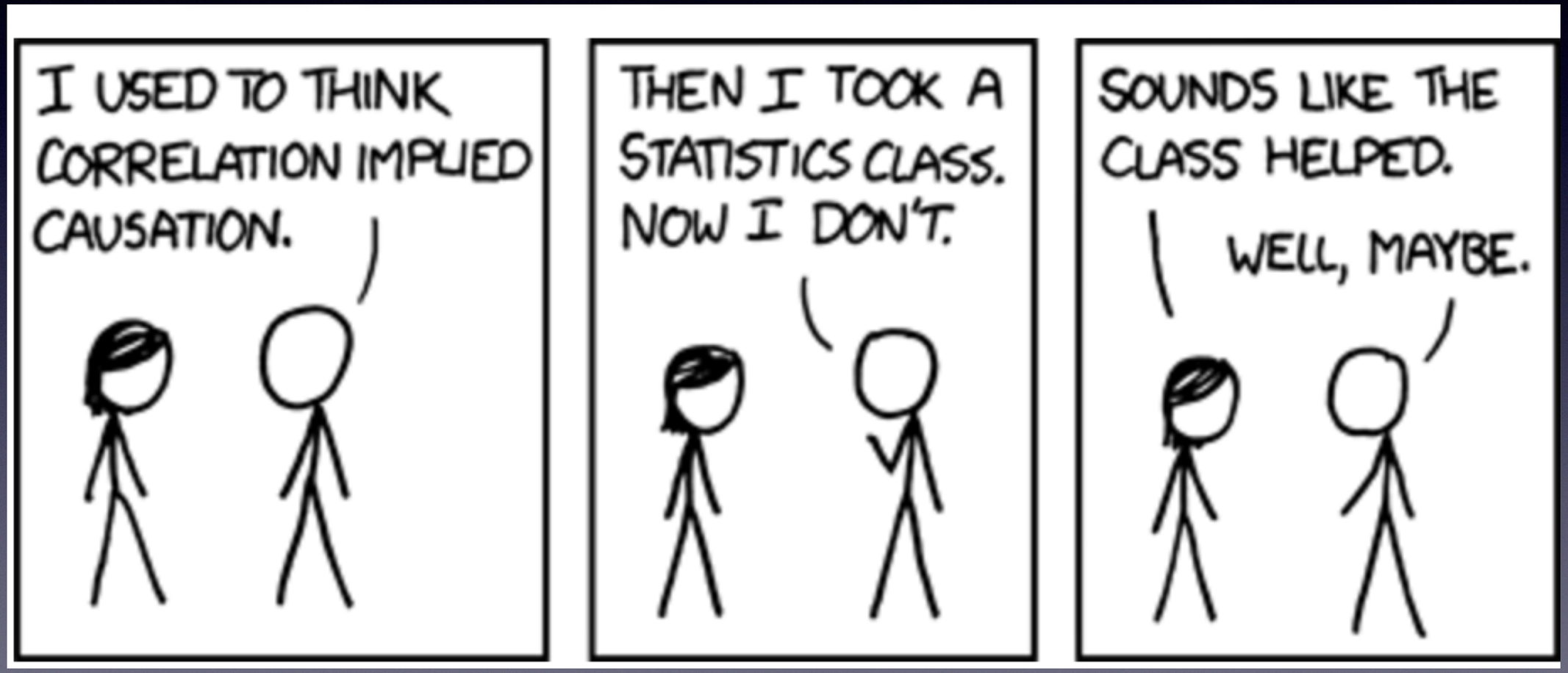
This is a stacked bar chart.



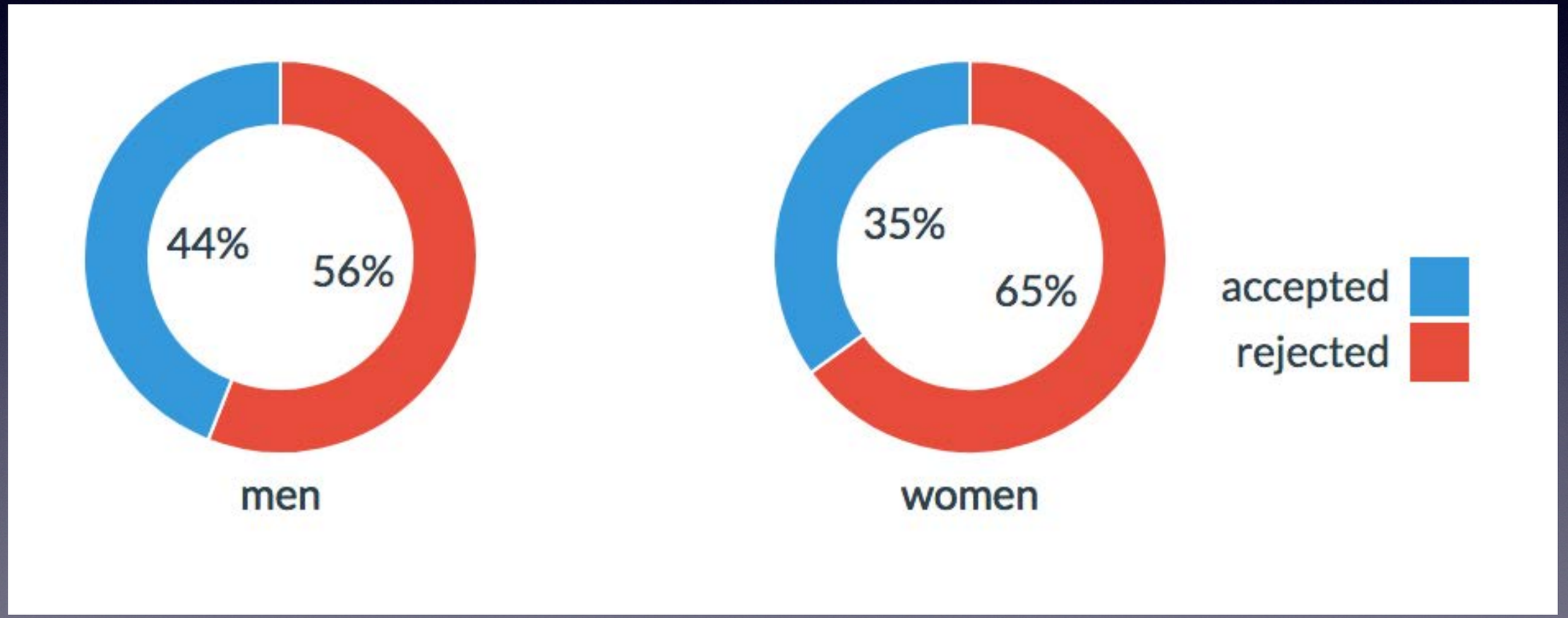
SCATTERPLOT

displays the relation of two quantitative variables

Association = Causation?



In 1973, the University of California-Berkeley was sued for sex discrimination. The numbers looked pretty incriminating: the graduate schools had just accepted 44% of male applicants but only 35% of female applicants.




Departments
A
B
C
D
E
F
combined
Women
Men
82%
67%
34%
34%
23%
6%
62%
63%
36%
33%
26%
7%
combine
accepted
rejected
10 applicants =