

# Data Visualization

Lisa Ann Yu  
UC Berkeley, Psych 7  
8/2/2018

# About Me

- ❖ Worked with Yang as a Research Assistant
- ❖ MS in Statistics

# About Me

- ❖ Worked with Yang as a Research Assistant
- ❖ MS in Statistics
- ❖ Really like **Questions and Answers, candy**

# Overview

"Most of us need to listen to the music to understand how beautiful it is. But often that's how we present statistics: we just show the notes, we don't play the music."

- Hans Rosling

<https://www.youtube.com/watch?v=jbkSRLYSojo>

# Overview

- ❖ **What** is data visualization?
- ❖ **Why** visualize data?
- ❖ **How** to visualize data?

# What is data visualization?

Is this a data visualization?

# Napoleon's Moscow Invasion

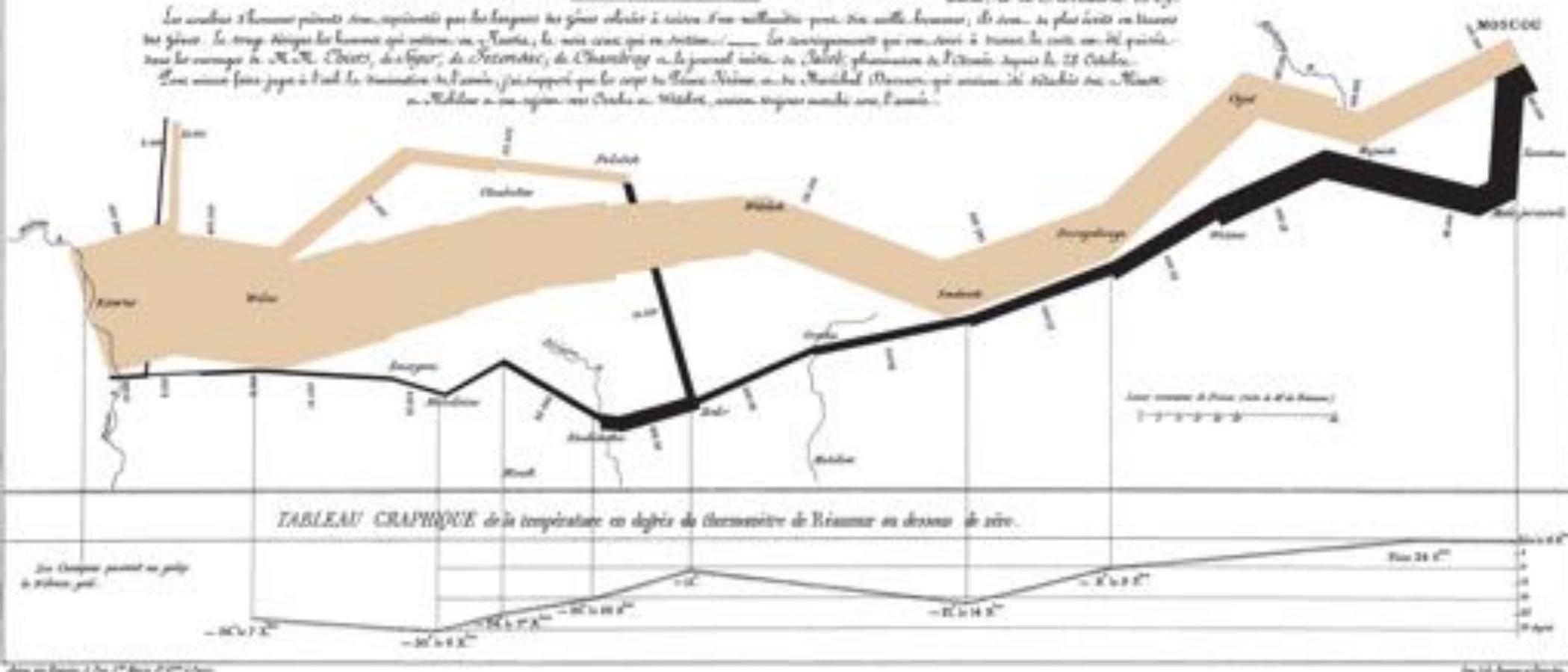
*Carte figurative des pertes subies en hommes de l'Armée Française dans la campagne de Russie, 1812-1813.*

*Dessin par N. RÉMond, Ingénieur Général sortant en Chemin de fer.*

*Paris, le 20 Novembre 1869.*

*Les années d'hiver présentent sans équivoque que les dangers des guerres coloniales à cause d'un millionne pour les soldats français; ils sont au plus élevé en Russie et plus faible. Le temps d'hiver le froid qui atteint en Russie, le vent avec qui se batte... Les combats qui sont assez à faire la mort au fil de l'épée, mais le courage de N.D. Cossat, du général de Tocqueville, de Chabrolle et le journaliste de Saint-Flour placent de l'Armée française depuis le 28 Octobre.*

*Cette carte fait juge à l'ail la domination de l'armée française sur le corps de l'armée russe, mais rappelle que le corps de l'armée russe a été dévasté par l'assaut de l'armée française qui a détruit les villes de Moscou et de Smolensk et qui a vaincu les forces russes dans l'armée.*



A word cloud visualization comparing Donald Trump's and Hillary Clinton's speeches. The left side shows words from Trump's speech, while the right side shows words from Clinton's speech. The size of each word indicates its frequency or importance in the respective speech.

**Trump's Speech (Left):**

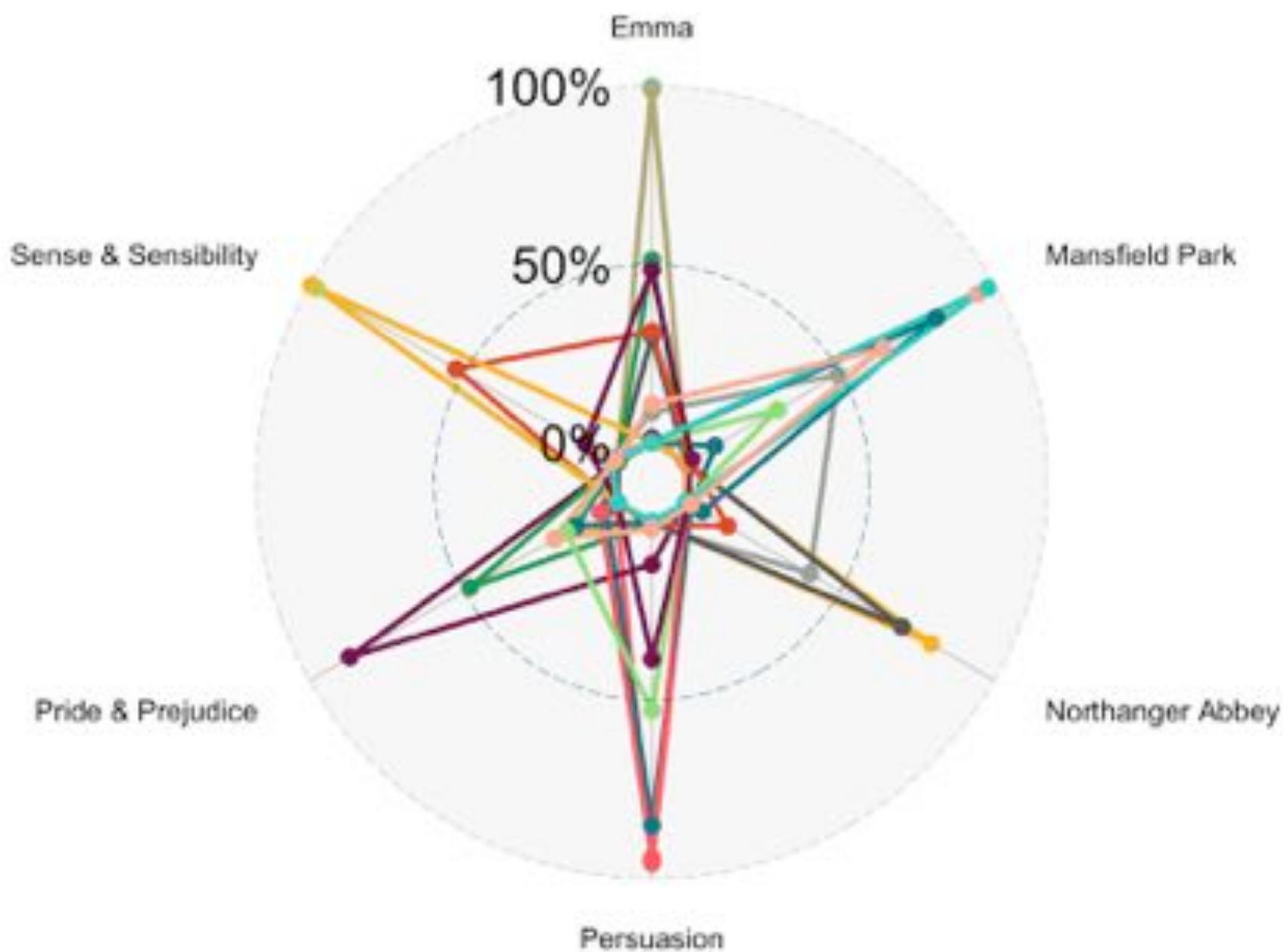
- re
- Hillary
- bad
- think
- person
- say
- back
- Okay
- talk
- two
- good
- upresident
- America
- call
- great
- ev
- want
- win
- make
- lot
- big
- really
- deal
- world
- everybody
- number
- something
- remember
- job
- right
- believe
- mean
- happen

**Clinton's Speech (Right):**

- re
- Hillary
- bad
- think
- person
- say
- back
- Okay
- talk
- two
- good
- upresident
- America
- call
- great
- ev
- want
- win
- make
- lot
- big
- really
- deal
- world
- everybody
- number
- something
- remember
- job
- right
- believe
- mean
- happen

9

- anne
- catherine
- charles
- edward
- elizabeth
- emma
- fanny
- harriet
- henry
- isabella
- jane
- john
- louisa
- lucy
- maria
- mary
- smith
- thomas
- william



# Michael Jordan Career Scoring

<https://public.tableau.com/en-us/s/gallery/michael-jordan-career-scoring?gallery=votd>

# What is Data Visualization?

Discuss with a classmate

# What is Data Visualization?

"the graphical representation of information and data"

The image displays two word clouds side-by-side, representing different political speeches or statements.

**Left Word Cloud:**

- Top words: year, ever, going, time, much, China, million, amazing, well, guy, billion, way, nice, one, now, nobody, Trump, thing, even, country, know, many, some, will, tell, love, said, up, problem, better, happened, care, little, came, known, trade, Mexico, look, got, see, people, let, day, actually, got, State, wall, pay, done.

**Right Word Cloud:**

- Top words: Hillary, bad, think, always, person, say, back, Okay, talk, two, good, up, president, run, America, great, ev, want, need, make, lot, big, really, give, deal, world, everybody, coming, Clinton, week, bring, job, United, right, believe, mean, happen.

14

# Why visualize data?

# Why visualize data?

“a picture is worth a thousand words”

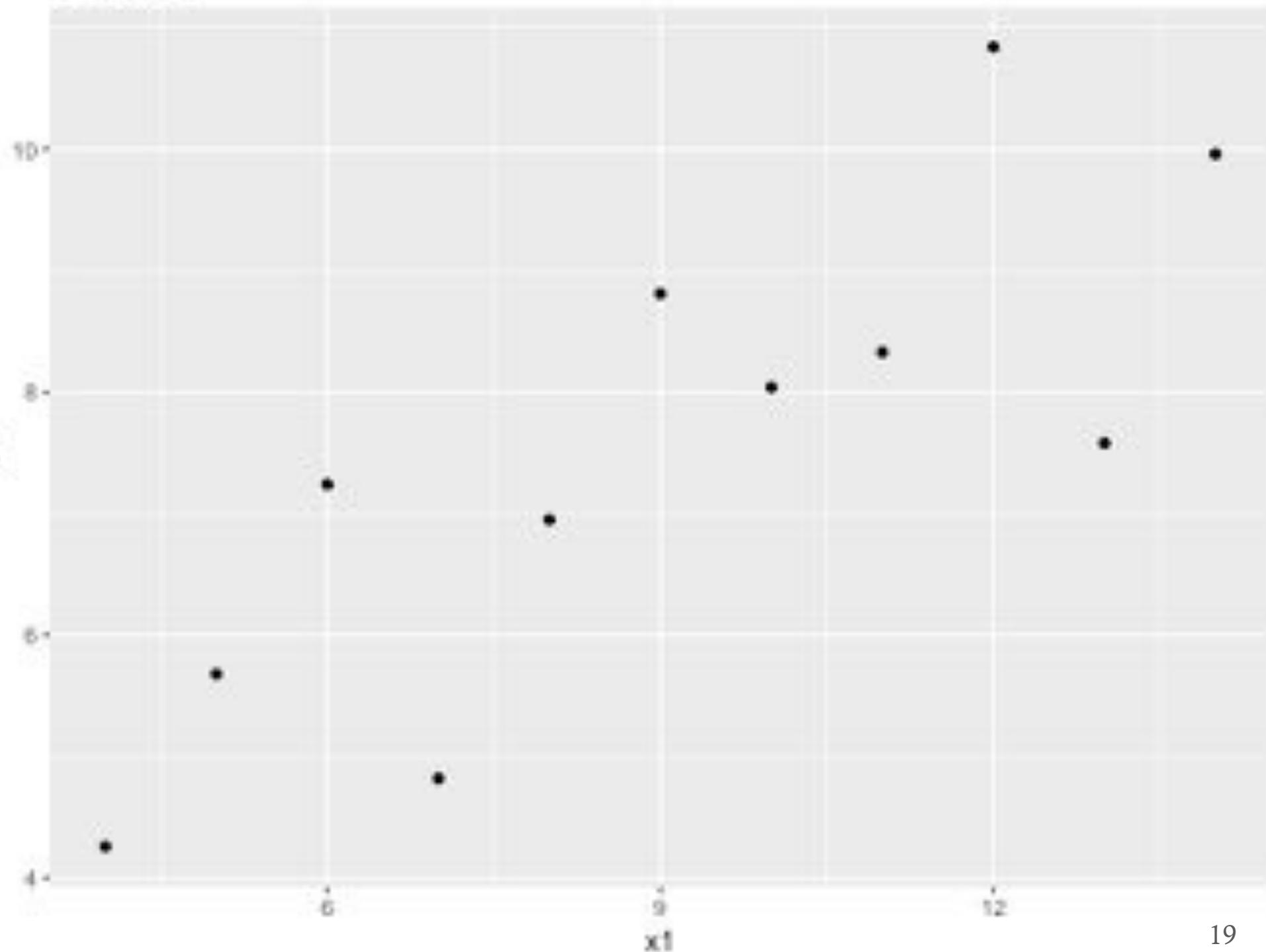
# Why visualize data?

- ❖ Identify
  - ❖ Trends
  - ❖ Relationships
  - ❖ Outliers

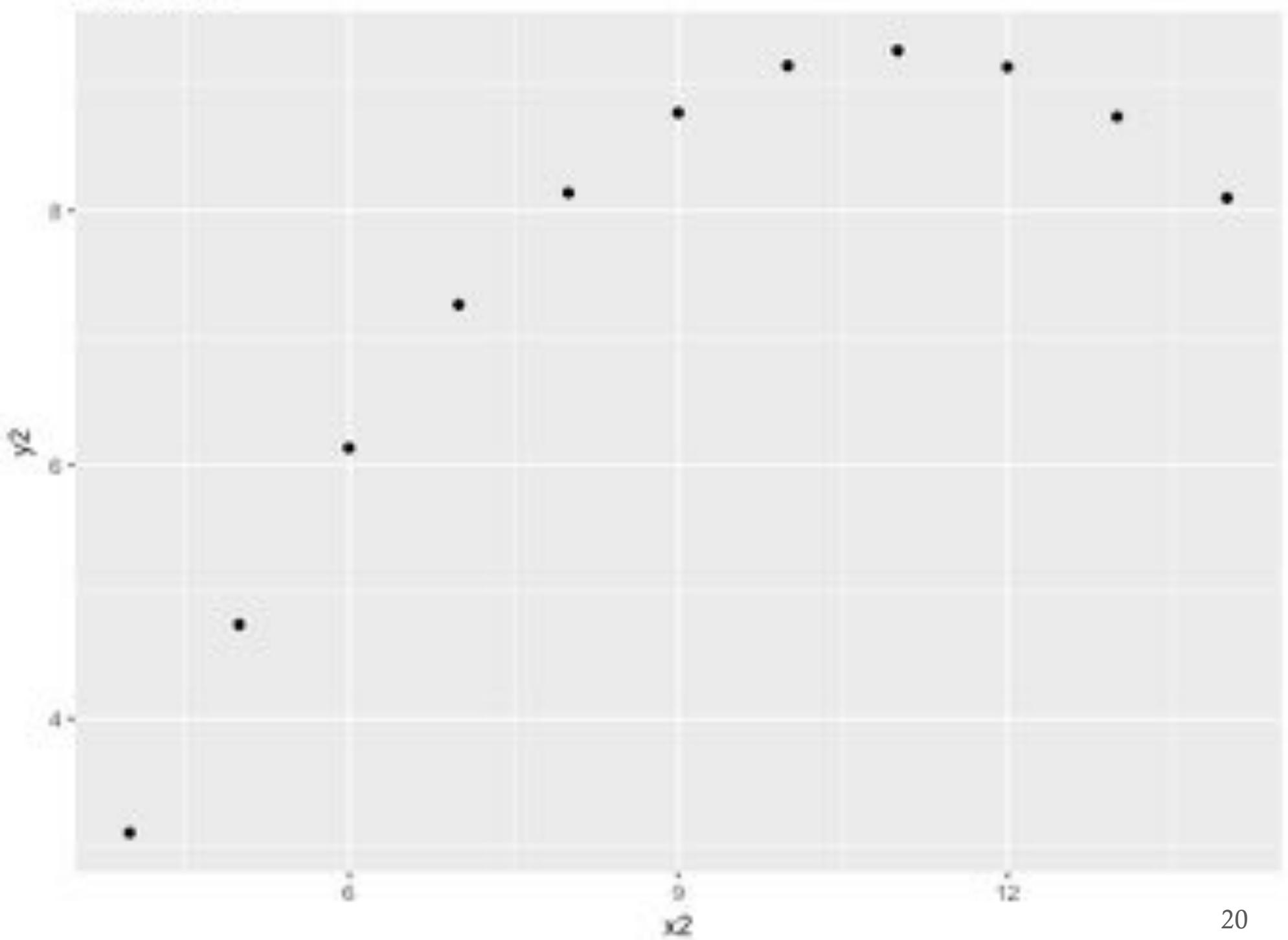
# Anscombe Dataset

x1	y1		x2	y2		x3	y3		x4	y4
10	8.04		10	9.14		10	7.46		8	6.58
8	6.95		8	8.14		8	6.77		8	5.76
13	7.58		13	8.74		13	12.74		8	7.71
9	8.81		9	8.77		9	7.11		8	8.84
11	8.33		11	9.26		11	7.81		8	8.47
14	9.96		14	8.1		14	8.84		8	7.04
6	7.24		6	6.13		6	6.08		8	5.25
4	4.26		4	3.1		4	5.39		19	12.5
12	10.84		12	9.13		12	8.15		8	5.56
7	4.82		7	7.26		7	6.42		8	7.91
5	5.68		5	4.74		5	5.73		8	6.89

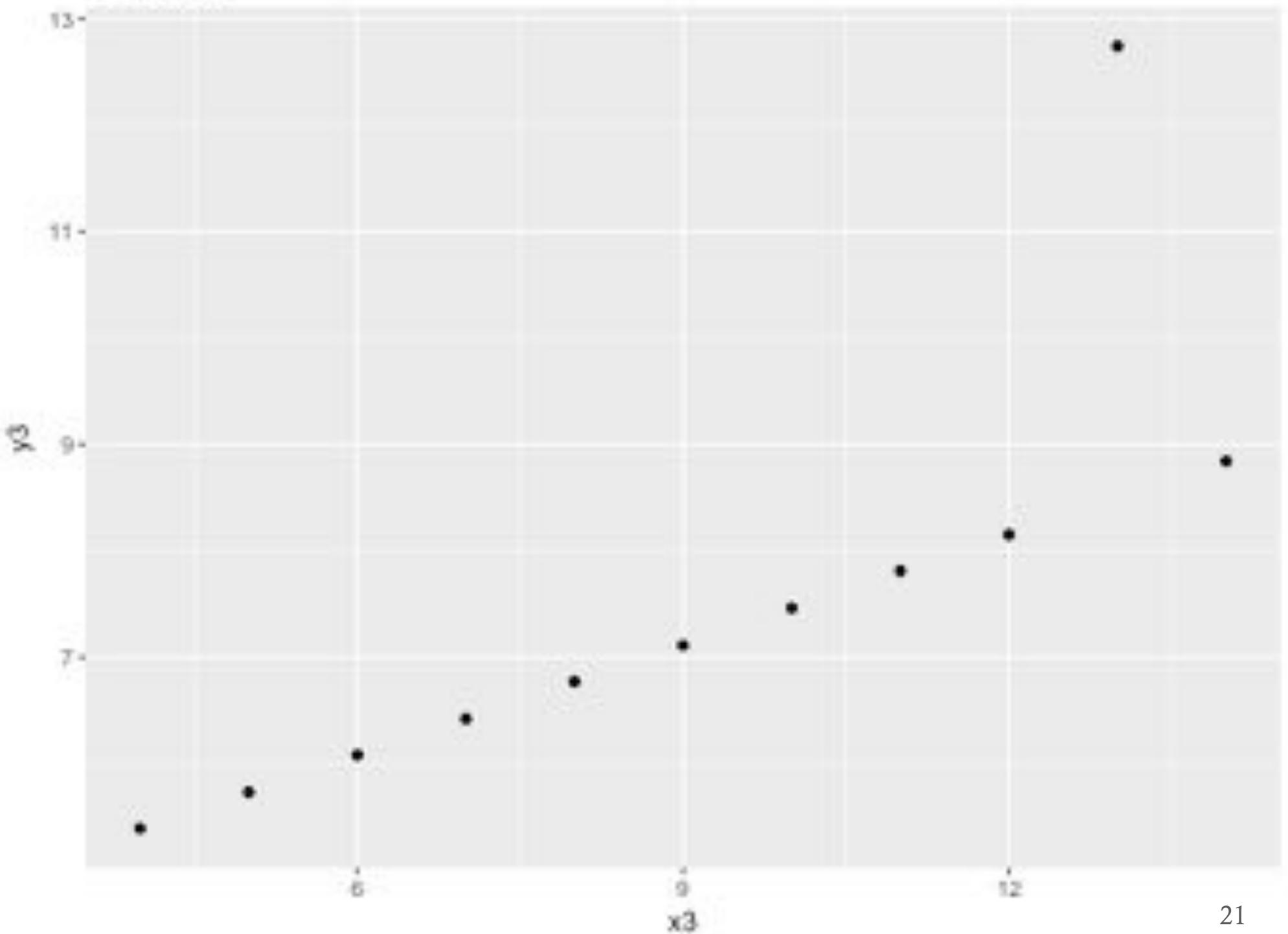
Dataset 1



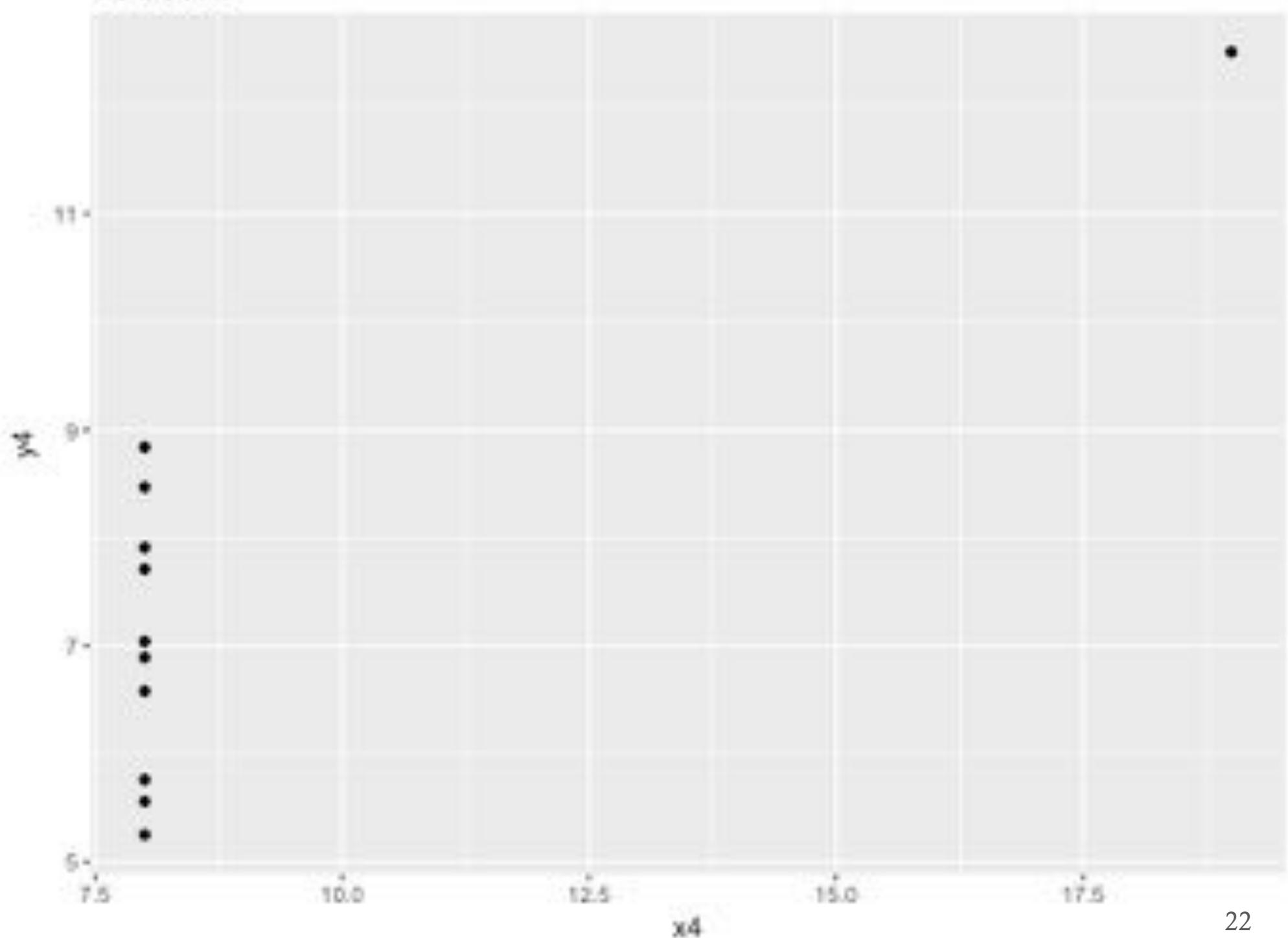
Dataset 2



Dataset 3

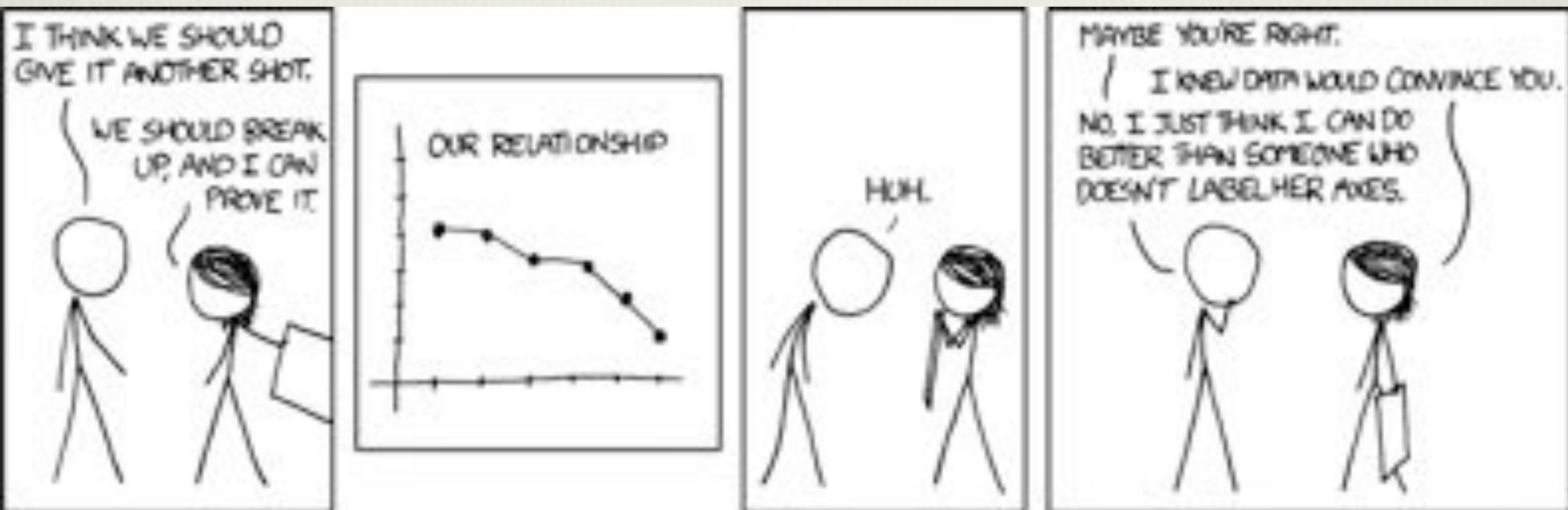


Dataset 4



# How to visualize data?

# How not to visualize data



# How to visualize data?

- ❖ What choices must be made when creating a data visualization?

[https://public.tableau.com/profile/lisa.ann.yu4798#!/?vzheme/EOPCollegeMobility/IncomebyTier?  
publish=yes](https://public.tableau.com/profile/lisa.ann.yu4798#!/?vzheme/EOPCollegeMobility/IncomebyTier?publish=yes)

# How to visualize data?

- ❖ Data
- ❖ Chart type
- ❖ Details
  - ❖ Order of variables
  - ❖ Color
  - ❖ Labels: Title, Axes, Legend
  - ❖ Annotations
  - ❖ Interactivity

# How to visualize data?

- ❖ Data
- ❖ Chart type
- ❖ Details
  - ❖ Order of variables
  - ❖ Color
  - ❖ Labels: Title, Axes, Legend
  - ❖ Annotations
  - ❖ Interactivity

# Example Dataset

- ❖ Mobility Rate: Which colleges do the best job of bringing students from lower income quintiles up to higher income quintiles?
- ❖ Aggregate Data: at the college-level
- ❖ Raw Data: kids' (students) and parents' incomes and college student attended



# Chart Type

School	MR20%	MR1%	Tier	Public/ Private
UC Berkeley	4.89%	0.764%	Highly Selective	Public
Stanford	2.25%	0.663%	Ivy Plus	Private
UCLA	5.60%	0.451%	Other Elite	Public

# Example Dataset

- ❖ **MR20%:** MR K Top 20% | P Bottom 20%
  - ❖ Kids whose parents were in the lowest 20% of the income distribution but who personally had income levels in the **top 20%**
- ❖ **MR1%:** MR K Top 1% | P Bottom 20%
  - ❖ Kids whose parents were in the lowest 20% of the income distribution but who personally had income levels in the **top 1%**
- ❖ **Tier:** Ivy Plus, Other Elite, Highly Selective, Selective, Nonselective, Two-Year, < 2-Year
- ❖ **Public/Private**

# Data Decisions

- ❖ Which variables to use?
- ❖ Format of data? (i.e. counts, percentages)

# How to visualize data?

- ❖ Data
- ❖ Chart type
- ❖ Details
  - ❖ Order of variables
  - ❖ Color
  - ❖ Labels: Title, Axes, Legend
  - ❖ Annotations
  - ❖ Interactivity

# Chart Type

- ❖ Depends on **number of variables** and **type of data**
- ❖ Types of data
  - ❖ **Continuous**
  - ❖ **Categorical**

# Chart Type

- ❖ Depends on **number of variables** and **variable type**
- ❖ Types of variables
  - ❖ Continuous
    - ❖ Perform arithmetic operations (i.e. +, -, \*, /)
  - ❖ Categorical
    - ❖ Categories/Can't perform arithmetic operations

# Chart Type

- ❖ Depends on **number of variables** and **type of data**
- ❖ Types of data
  - ❖ Numeric
    - ❖ Height (e.g. in inches)
    - ❖ # of students in each course at Berkeley
  - ❖ Categorical
    - ❖ Race
    - ❖ Gender (at birth)

# Chart Type

School	MR20%	MR1%	Tier	Public/ Private
UC Berkeley	4.89%	0.764%	Highly Selective	Public
Stanford	2.25%	0.663%	Ivy Plus	Private
UCLA	5.60%	0.451%	Other elite	Public

# Chart Type: Example Dataset

- ❖ MR20%
- ❖ MR1%
- ❖ Tier
- ❖ Public/Private

# Chart Type: Example Dataset

- ❖ MR20%: **continuous**
- ❖ MR1%: **continuous**
- ❖ Tier: **categorical**
- ❖ Public/Private: **categorical**

**2 continuous, 2 categorical variables**

# Step Back: Chart Types

Discuss with a classmate

- a. Histogram
- b. Choropleth
- c. Scatterplot
- d. Boxplot
- e. Pie chart
- f. Radar plot
- g. Bar plot

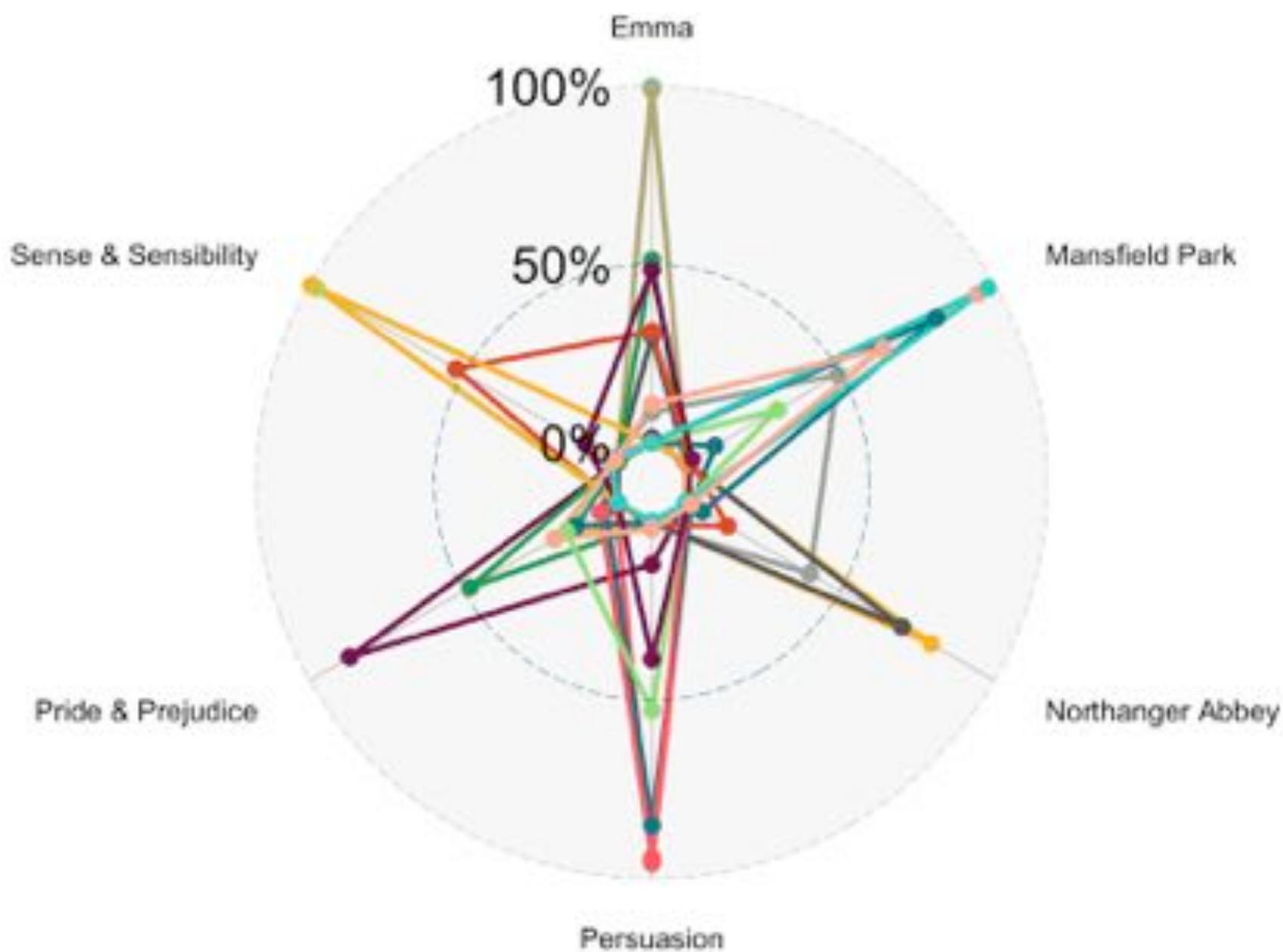
# Step Back: Chart Types

<https://public.tableau.com/profile/lisa.ann.yu4798#!/vizhome/DataVizLecture/ChartTypes?publish=yes>

# Step Back: Chart Types

- a. Histogram: 5
- b. Choropleth: 1
- c. Scatterplot: 4
- d. Boxplot: 2
- e. Pie chart: 3
- f. Radar plot: **Not pictured**
- g. Bar plot: 6

- anne
- catherine
- charles
- edward
- elizabeth
- emma
- fanny
- harriet
- henry
- isabella
- jane
- john
- louisa
- lucy
- maria
- mary
- smith
- thomas
- william



# Chart Type

- ❖ One **continuous** variable
  - ❖ MR20%

# Chart Type

- ❖ One **continuous** variable
  - ❖ MR20%

Histogram

Boxplot

# Chart Type

- ❖ One **categorical** variable
  - ❖ Tier

# Chart Type

- ❖ One **categorical** variable
  - ❖ Tier

**Bar Graph (y-axis: counts)**

# Chart Type

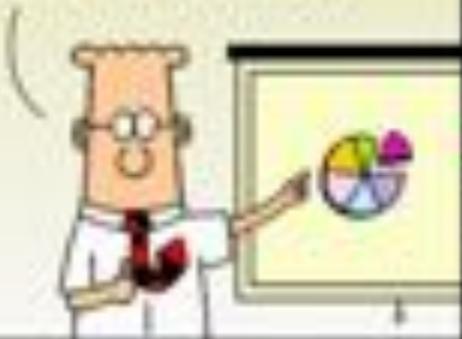
- ❖ One **categorical** variable
  - ❖ Tier

**Pie Chart**

HOW DOES THE  
DATA LOOK  
THIS WEEK?

VISUALIZING THE DATA IN PIE-CHART FORMAT MADE  
SUBPAR NUMBERS SEEM OKAY.

I DIDN'T HAVE  
ANYTHING USEFUL  
TO SAY SO I MADE  
THIS PIE CHART.



OOOH!  
OOOH!

IT MUST  
BE TRUE  
BECAUSE  
IT'S PIE.

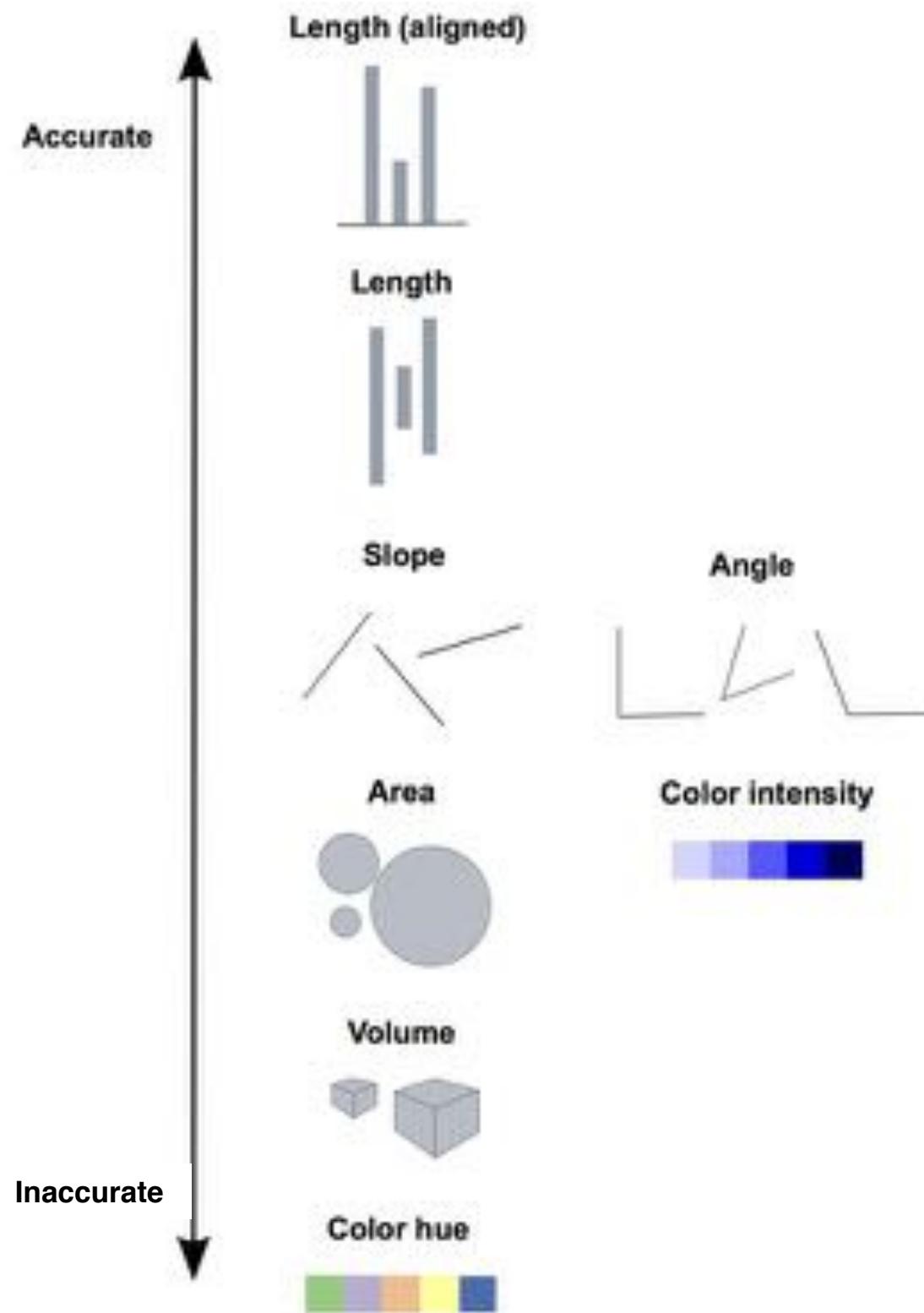
THAT  
WORKED  
TOO  
WELL.

I PLEDGE  
MY LIFE  
AND MY  
FORTUNE  
TO THE PIE!

# Chart Type

## ~~Pie Chart?~~

- ❖ Humans not good at comparing areas
- ❖ Best: lengths, position
- ❖ Worst: volume, area



# Chart Type

- ❖ Two **continuous** variables
  - ❖ MR20%
  - ❖ MR1%

# Chart Type

- ❖ Two **continuous** variables
  - ❖ MR20%
  - ❖ MR1%

## Scatterplot

# Chart Type

- ❖ Two **continuous** variables, one **categorical** variable
  - ❖ MR20%
  - ❖ MR120%
  - ❖ Tier

# Chart Type

- ❖ Two **continuous** variables, one **categorical** variable
  - ❖ MR20%
  - ❖ MR1%
  - ❖ Tier

## Color Scatterplot

# Chart Type

- ❖ One **continuous** variable, one **categorical** variable
  - ❖ MR20%
  - ❖ Tier

# Chart Type

- ❖ One **continuous** variable, one **categorical** variable
  - ❖ MR20%
  - ❖ Tier

**Bar Graph**  
**(y-axis: median percentage)**

# Chart Type

- ❖ Two **continuous** variables, two **categorical** variables
  - ❖ MR20%
  - ❖ MR120%
  - ❖ Tier
  - ❖ Public/Private

# Chart Type

- ❖ Two **continuous** variables, one **categorical** variable
  - ❖ MR20%
  - ❖ MR1%
  - ❖ Tier
  - ❖ Public/Private

**Side-by-side Bar Chart**  
**(y-axis: median percentage)**

# How to visualize data?

- ❖ Data
- ❖ Chart type
- ❖ **Details**
  - ❖ Order of variables
  - ❖ Color
  - ❖ Labels: Title, Axes, Legend
  - ❖ Annotations
  - ❖ Interactivity

# Interpret

Discuss with a neighbor

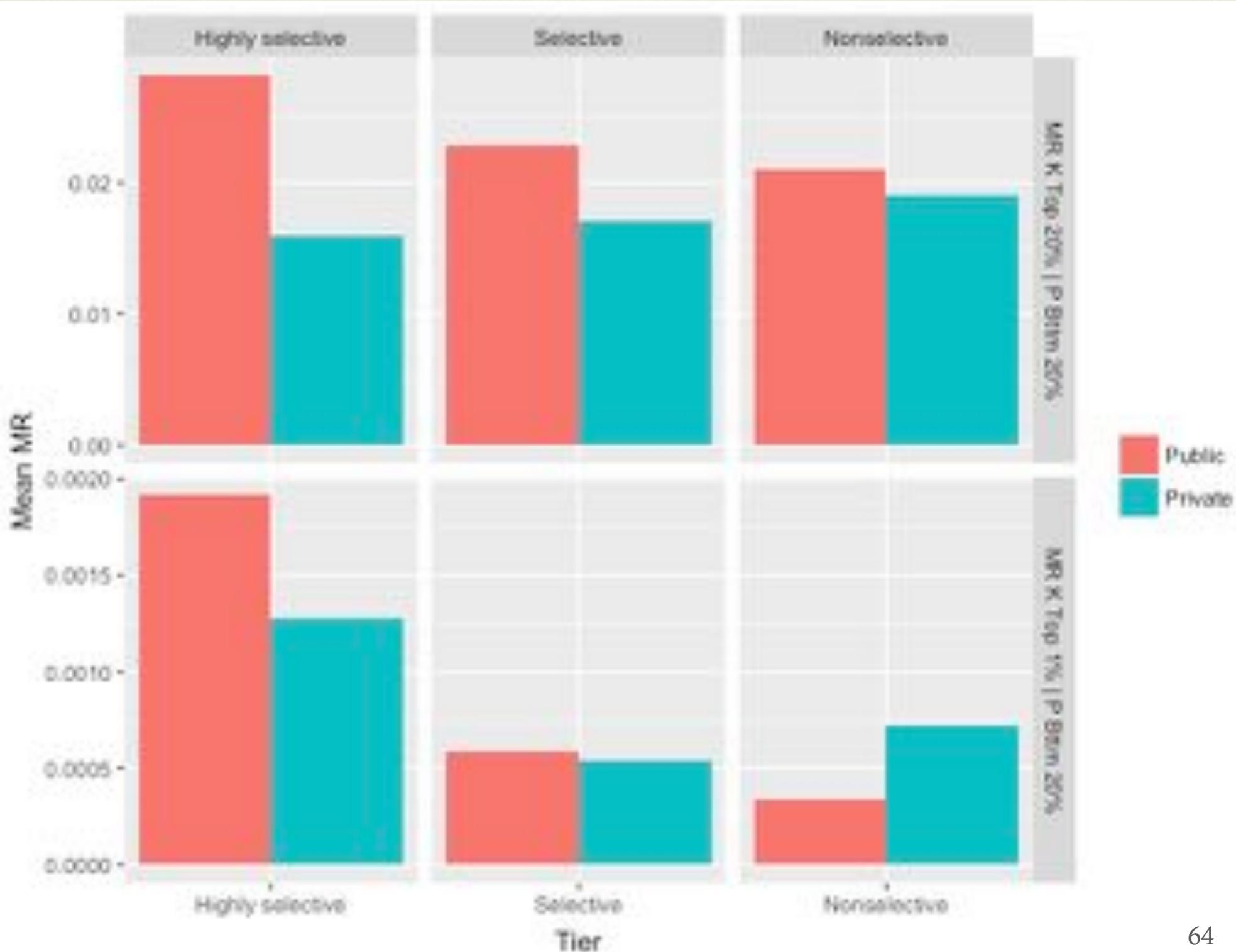
Note: I simplified the tiers

# Interpret

- ❖ In general, public schools are better than private schools at bringing students from the bottom of the income distribution to the top.
- ❖ Especially true for Highly Selective schools
- ❖ Difference between public and private nonexistent for Nonselective schools at the MR20%

# How to visualize data?

- ❖ Data
- ❖ Chart type
- ❖ **Details**
  - ❖ Order of variables
  - ❖ Color
  - ❖ Labels: Title, Axes, Legend
  - ❖ Annotations
  - ❖ Interactivity



# Resources

- ❖ *The Visual Display of Quantitative Information* by Edward Tufte
- ❖ *The Big Book of Dashboards* by Andy Cotgreave, Jeffrey Shaffer, Steve Wexler

# Questions?

# References

- ❖ "Equality of Opportunity" Project (Chetty, Friedman, Hendren): "Mobility Report Cards: The Role of Colleges in Intergenerational Mobility":  
<http://www.equality-of-opportunity.org/data/>
- ❖ janeaustenr CRAN package:  
<https://cran.r-project.org/web/packages/janeaustenr/index.html>
- ❖ <http://paldhous.github.io/ucb/2016/dataviz/week2.html>

# References (images)

- ❖ [https://www.reddit.com/r/dataisbeautiful/  
comments/4phood/trump\\_speeches\\_wordcloud\\_oc](https://www.reddit.com/r/dataisbeautiful/comments/4phood/trump_speeches_wordcloud_oc)
- ❖ <https://xkcd.com/833/>
- ❖ <https://goo.gl/images/MKYKvg>
- ❖ [https://ils.unc.edu/courses/2013\\_summerI/  
inls261\\_001/sessions/04.Spreadsheets/  
14.data\\_display/04.14b.spreadsheets.display.html](https://ils.unc.edu/courses/2013_summerI/inls261_001/sessions/04.Spreadsheets/14.data_display/04.14b.spreadsheets.display.html)

# *Text Analytics*

Lisa Ann Yu  
UC Berkeley, Psych 7  
8/2/2018

# Overview

0 What is text analytics?

0 How to analyze text?

    0 Pipeline

    0 Methods

        0 Pitfalls

# What is text analytics?

# What is text analytics?

“drawing meaning out of written communication”

# What is text analytics?

“drawing meaning out of written communication”

Written communication: any language

# What is text analytics?

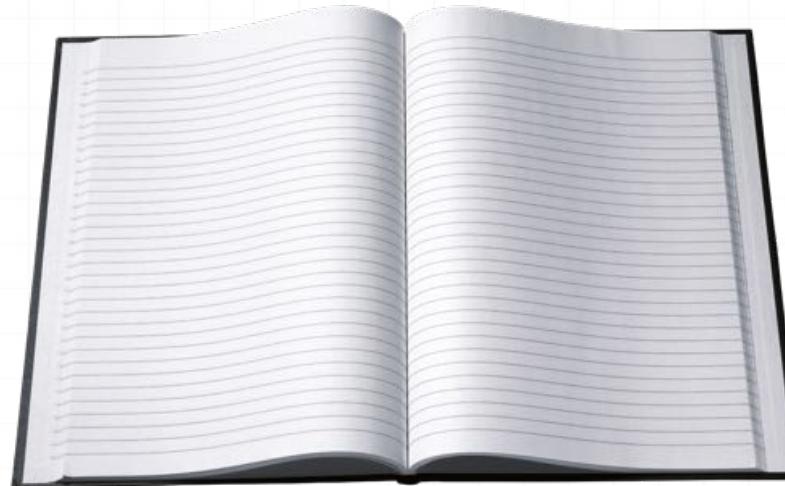
“drawing meaning out of written communication”

Oldest form:

# What is text analytics?

“drawing meaning out of written communication”

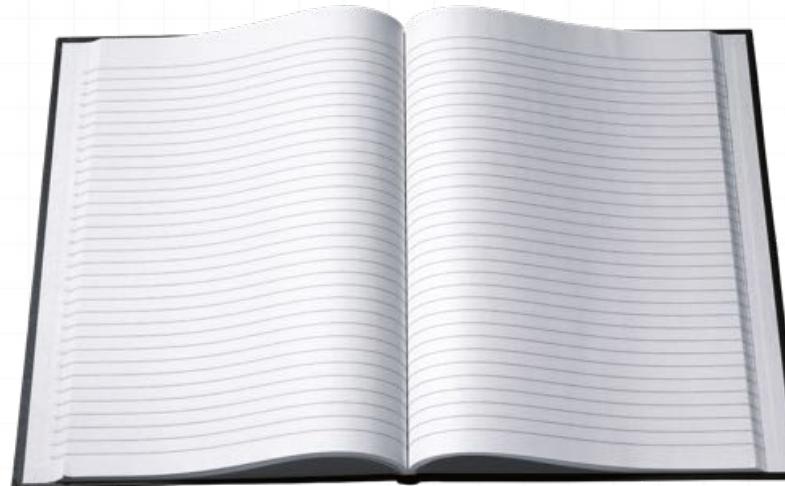
Oldest form: **reading**



# What is text analytics?

“drawing meaning out of written communication”

Oldest form: **reading – takes too long**



# What is text analytics?

Example forms of written communication: ?

# What is text analytics?

Example forms of written communication:

- 0 Books
- 0 Surveys
- 0 Emails
- 0 Twitter
- 0 State of Union speeches
- 0 Professor/class reviews
- 0 Yelp reviews
- 0 Song lyrics

# How to analyze text?

# Example Datasets

1. Trump's Tweets between 2009 and 2018
2. Trump's Speeches during campaign



# Donald Trump

Roles prior to presidency?



# Donald Trump

Roles prior to presidency:

- 0 Businessman: Trump Organization, Trump Tower
- 0 Miss Universe and Miss USA pageants
- 0 Reality TV Producer of *Apprentice, Celebrity Apprentice*
- 0 Coauthor of *The Art of the Deal*

# Questions of Interest

1. What are the most common words?
  - Specifically: over all tweets in this dataset
2. How does sentiment/emotion change over time?
  - Specifically: as Trump changes industries
3. What is the underlying content of a document?
  - Specifically: examining Trump's speeches
4. Which documents are similar to each other?
  - Specifically: examining Trump's speeches

# Example Dataset 1

0 Trump's Tweets from May 2009 – Jan 2018

0 Variables:

0 Created\_at (date-time)

0 Tweet Text

# Example Dataset 1

3	2016-11-08 00:17:57	Big news to share in New Hampshire tonight! Polls loo...
4	2016-11-08 03:43:54	Unbelievable evening in New Hampshire - THANK YO...
5	2016-11-08 04:27:18	@detroitnews: .@hankaTrump in Michigan: "This is yo...
6	2016-11-08 04:29:10	@DonaldJTrumpjr: Thanks New Hampshire!!! #NH #M...
7	2016-11-08 06:42:36	Today we are going to win the great state of MICHIGA...
8	2016-11-08 11:43:34	TODAY WE MAKE AMERICA GREAT AGAIN!
9	2016-11-08 16:39:36	VOTE TODAY! Go to to find your polling location. We ...
10	2016-11-08 18:03:49	We need your vote. Go to the POLLS! Let's continue th...
11	2016-11-08 18:23:39	#ElectionDay
12	2016-11-08 21:18:04	I will be watching the election results from Trump To...
13	2016-11-08 21:28:24	Just out according to @CNN: "Utah officials report voti...
14	2016-11-08 21:31:20	Don't let up, keep getting out to vote - this election i...
15	2016-11-08 23:03:42	Still time to #VoteTrump! #Voted #ElectionNight

Date Time Created	Tweet Text
11/8/16 0:08	Thank you Pennsylvania! Going to New Hampshire now and on to Michigan. Watch PA rally here: The big vote tomorrow!
11/8/16 0:16	Today in Florida, I pledged to stand with the people of Cuba and Venezuela in their fight against oppression- cont:
11/8/16 0:17	Big news to share in New Hampshire tonight! Polls looking great! See you soon. Unbelievable evening in New Hampshire - THANK YOU! Flying to Grand Rapids, Michigan now.
11/8/16 3:43	Watch NH rally here:
11/8/16 4:27	@detroitnews: .@IvankaTrump in Michigan: "This is your movement"™ @realDonaldTrump @DonaldJTrumpJr: Thanks New Hampshire!!!
11/8/16 4:29	#NH #NewHampshire #MAGA
11/8/16 6:42	Today we are going to win the great state of MICHIGAN and we are going to WIN back the White House! Thank you MI!
11/8/16 11:43	TODAY WE MAKE AMERICA GREAT AGAIN!
11/8/16 16:39	VOTE TODAY! Go to to find your polling location. We are going to Make America Great Again! #VoteTrump #ElectionDay
11/8/16 18:03	We need your vote. Go to the POLLS! Let's continue this MOVEMENT! Find your poll location: #ElectionDay #VoteTrump
11/8/16 18:23	#ElectionDay
11/8/16 21:18	I will be watching the election results from Trump Tower in Manhattan with my family and friends. Very exciting!
11/8/16 21:28	Just out according to @CNN: "Utah officials report voting machine problems across entire country"
11/8/16 21:31	Don't let up, keep getting out to vote - this election is FAR FROM OVER! We are doing well but there is much time left. GO FLORIDA! Still time to #VoteTrump!
11/8/16 23:03	#iVoted #ElectionNight @DonaldJTrumpJr: FINAL PUSH! Eric and I doing dozens of radio interviews. We can win this thing! GET OUT AND VOTE! #MAGA
11/8/16 23:20	#ElectionDay hté;
11/8/16 23:20	@EricTrump: Join my family in this incredible movement to #MakeAmericaGreatAgain!! Now it is up to you! Please #VOTE for America! :é;

# Pipeline/Method 0

Pipeline: Standard steps to take when analyzing text

Method 0: Most common words (Question of Interest 1)

# Pipeline

1. Standardize case
2. Tokenization
3. Stop word removal
4. \*Punctuation Removal/Stemming

# Pipeline: Standardize case

0 TODAY = Today = today

DateTime Created	Tweet Text
11/8/16 0:08	Thank you Pennsylvania! Going to New Hampshire now and on to Michigan. Watch PA rally here: The big vote tomorrow!
11/8/16 0:16	<b>Today</b> in Florida, I pledged to stand with the people of Cuba and Venezuela in their fight against oppression- cont:
11/8/16 0:17	Big news to share in New Hampshire tonight! Polls looking great! See you soon. Unbelievable evening in New Hampshire - THANK YOU! Flying to Grand Rapids, Michigan now.
11/8/16 3:43	Watch NH rally here:
11/8/16 4:27	@detroitnews: .@IvankaTrump in Michigan: "This is your movement"™ @realDonaldTrump @DonaldJTrumpJr: Thanks New Hampshire!!!
11/8/16 4:29	#NH #NewHampshire #MAGA
11/8/16 6:42	<b>Today</b> we are going to win the great state of MICHIGAN and we are going to WIN back the White House! Thank you MI!
11/8/16 11:43	<b>TODAY WE MAKE AMERICA GREAT AGAIN!</b>
11/8/16 16:39	VOTE <b>TODAY</b> ! Go to to find your polling location. We are going to Make America Great Again! #VoteTrump #ElectionDay
11/8/16 18:03	We need your vote. Go to the POLLS! Let's continue this MOVEMENT! Find your poll location: #ElectionDay #VoteTrump
11/8/16 18:23	#ElectionDay
11/8/16 21:18	I will be watching the election results from Trump Tower in Manhattan with my family and friends. Very exciting!
11/8/16 21:28	Just out according to @CNN: "Utah officials report voting machine problems across entire country"
11/8/16 21:31	Don't let up, keep getting out to vote - this election is FAR FROM OVER! We are doing well but there is much time left. GO FLORIDA! Still time to #VoteTrump!
11/8/16 23:03	#iVoted #ElectionNight
11/8/16 23:20	@DonaldJTrumpJr: FINAL PUSH! Eric and I doing dozens of radio interviews. We can win this thing! GET OUT AND VOTE! #MAGA
11/8/16 23:20	#ElectionDay ht;
11/8/16 23:20	@EricTrump: Join my family in this incredible movement to #MakeAmericaGreatAgain!! Now it is up to you! Please #VOTE for
11/8/16 23:20	America! :)

# Pipeline

1. Standardize case (i.e. all upper or lower case)
2. Tokenization

# Pipeline: Tokenization

- 0 Token: “a meaningful unit of text, such as a word”
- 0 Tokenization: “splitting text into tokens”
- 0 Separator: space, comma, hyphen, period, hashtag

Example: “Still time to #VoteTrump! #IVoted  
#ElectionNight

# Pipeline: Tokenization

“Still time to #VoteTrump! #IVoted #ElectionNight

Separator: space	Separator: #
Still	Still time to
time	VoteTrump!
to	IVoted
#VoteTrump!	ElectionNight
#IVoted	
#ElectionNight	

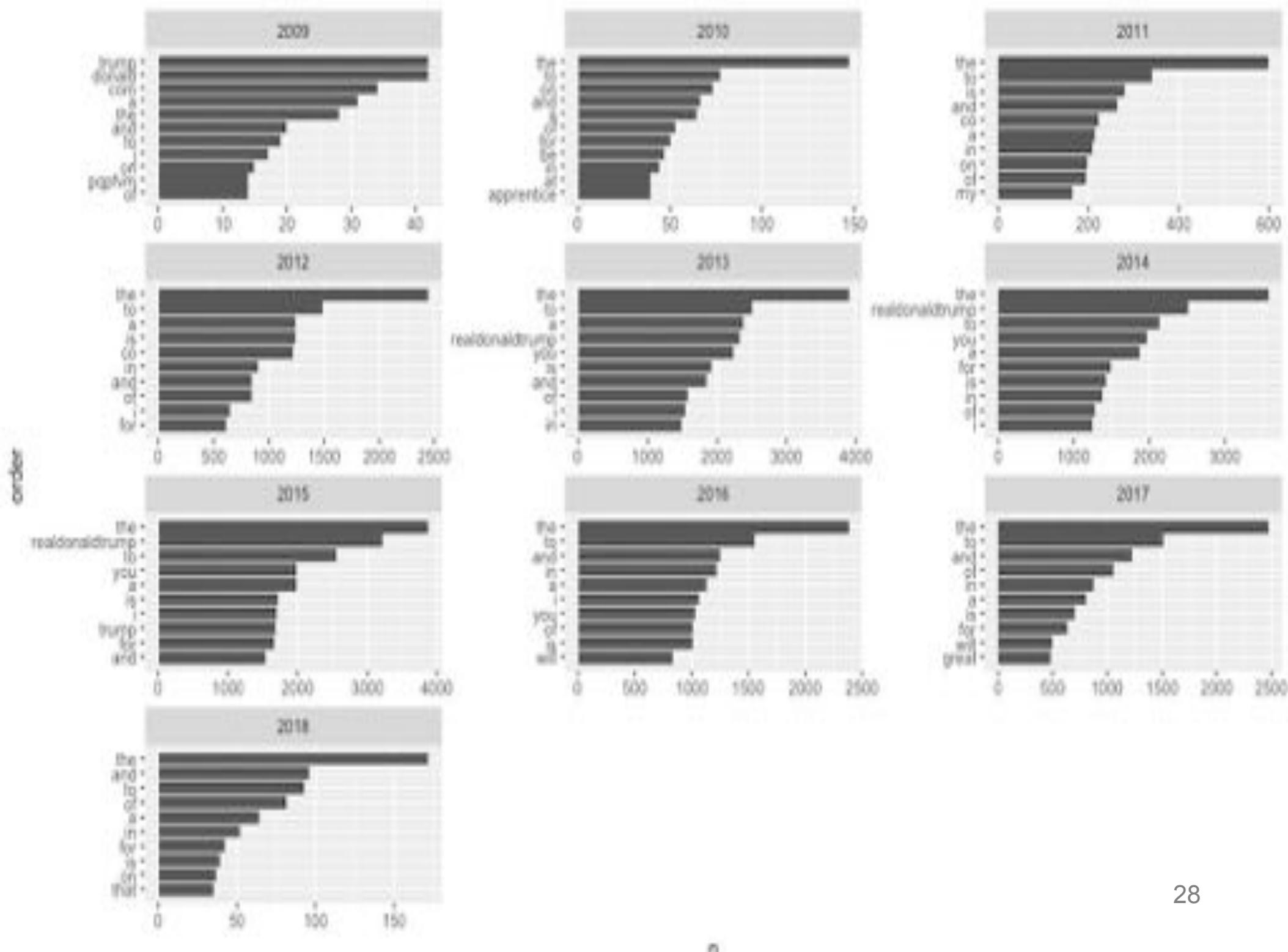
# Pipeline: Tokenization

## 0 Pitfalls

- 0 Depends heavily on token
- 0 Sometimes need two words to completely convey the meaning, i.e. New York

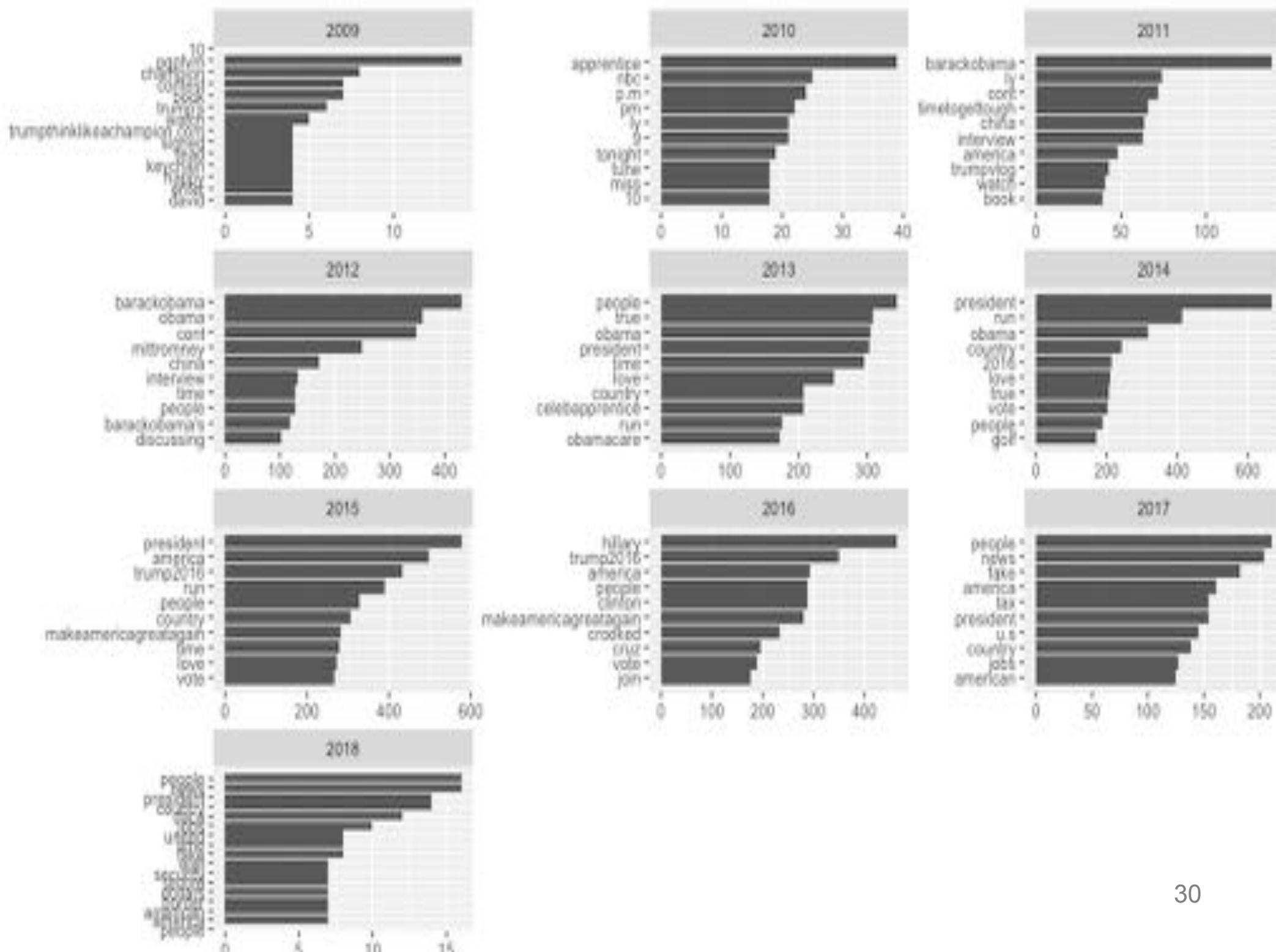
# Pipeline: ?

Created_at	Separator: space
2016-11-08	Still
2016-11-08	time
2016-11-08	to
2016-11-08	#VoteTrump!
2016-11-08	#IVoted
2016-11-08	#ElectionNight
2016-11-09	@IvankaTrump
2016-11-09	Such



# Pipeline: Stop Words

- 0 Def: words “not useful for analysis, typically extremely common words”
- 0 Stop words list comes from lexicon



# Pipeline: Stop Words

## 0 Pitfalls:

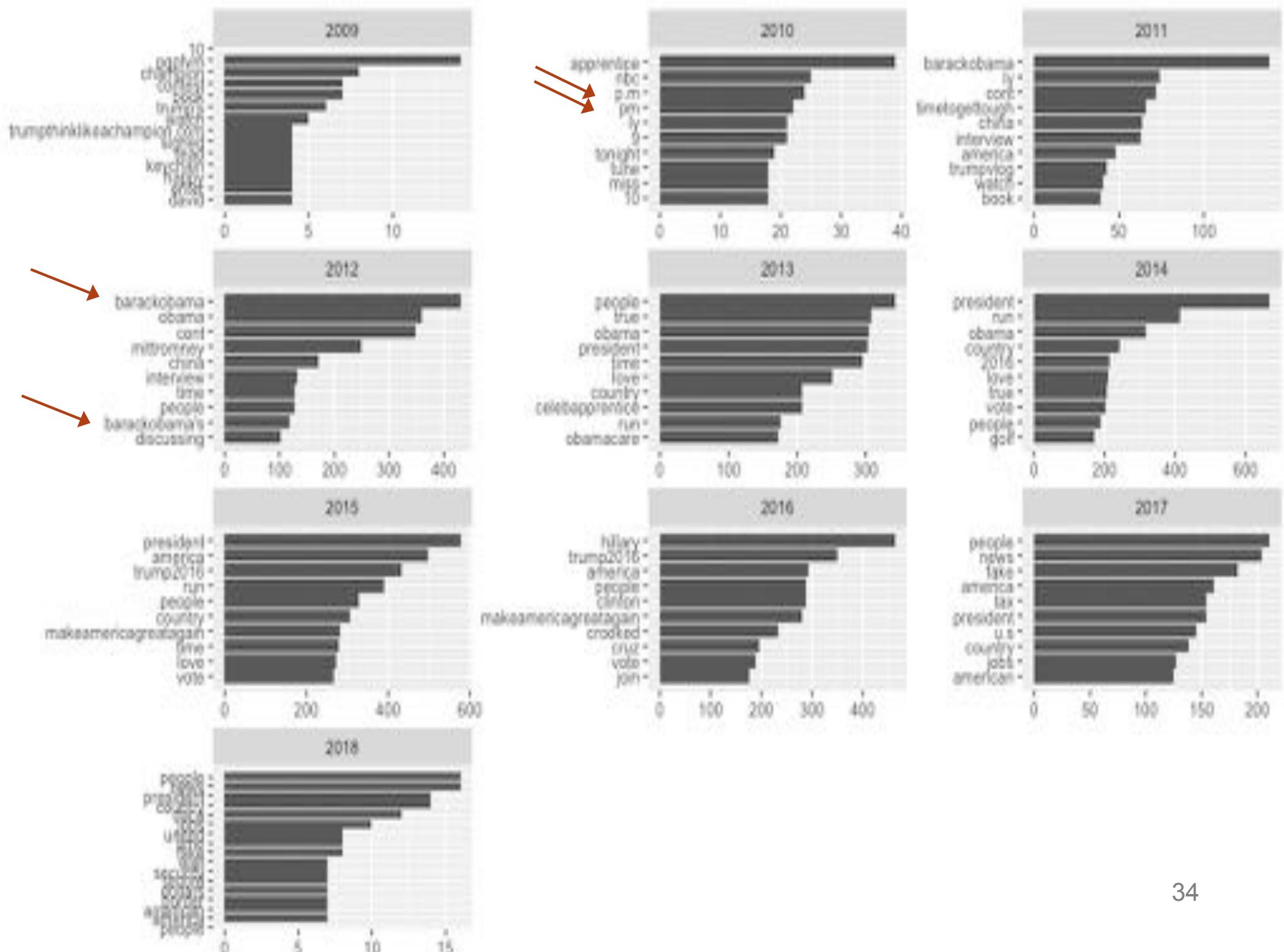
### 0 Document/Topic-specific

0 i.e. realdonaldtrump, donald, trump, t.co (twitter link)

# Pipeline: ?

# Pipeline: ?

1. Standardize case
2. Tokenization
3. Stop Words Removal
4. ?



# Pipeline: ?

1. Standardize case
2. Tokenization
3. Stop Words Removal
4. Remove punctuation/stemming

# Pipeline: ?

## 0 Punctuation Removal:

0 , . ! ; ? ( ) [ ] { }

0 's

0 Danger: meaning of a word may change if remove ‘

0 i.e. it's vs. its

## 0 Stemming

0 america = american = americans

# Pipeline: ?

# Pipeline

1. Standardize case
2. Tokenization
3. Stop Words
4. Remove punctuation/stemming

# Methods 1-4

# Methods

1. Sentiment Analysis
2. Emotion Analysis
3. TF-IDF
4. Topic Modeling

# Methods

- 1. Sentiment Analysis**
2. Emotion Analysis
3. TF-IDF
4. Topic Modeling

# Sentiment Analysis

- 0 Question of Interest 2: How does sentiment/emotion change over time?
- 0 Time frame here: month

# Sentiment Analysis

- 0 Lexicon/Dictionary:
  - 0 AFINN
  - 0 Bing
  - 0 NRC
  - 0 Many others: SentiWords

# Sentiment Analysis

## 0 Lexicon: AFINN

- 0 Highly positive (5): “hurrah”
- 0 Positive (4): “amazing”
- 0 Neutral (0): “some kind”
- 0 Negative (-4): “catastrophic”
- 0 Highly negative (-5): Cuss words
- 0 N = 2476

# Sentiment Analysis

0 Lexicon: Bing et al.

0 Positive (30%)

0 Negative (70%)

0 N = 6788

# Sentiment Analysis

0 Lexicon: NRC

0 Positive (41%)

0 Negative (59%)

0 N = 5636

# Sentiment Analysis

- 0 Get score/sentiment for every word
- 0 Calculate average score per month or percent of positive words

# Sentiment Analysis

## Election Day Tweets: AFINN

status_id	word	score
<id>	<id>	<int>
x795782371895349250	tonight	NA
x795782371895349250	polls	NA
x795834203430645760	unbelievable	-1
x795834203430645760	evening	NA
x795834203430645760	hampshire	NA
x795834203430645760	flying	NA
x795834203430645760	grand	3
x795834203430645760	rapids	NA
x795834203430645760	michigan	NA
x795834203430645760	watch	NA

# Sentiment Analysis

## Election Day Tweets: AFINN

status_id	word	score
<id>	<id>	<id>
x795781945607278592	fight	-1
x795782371895349250	share	1
x795834203430645760	unbelievable	-1
x795834203430645760	grand	3
x795879172795203584	win	4
x795879172795203584	win	4
x796099494442057728	exciting	3
x796130213826621440	win	4
x796130340180029441	join	1

Date Time Created	Tweet Text
11/8/16 0:08	Thank you Pennsylvania! Going to New Hampshire now and on to Michigan. Watch PA rally here: The big vote tomorrow!
11/8/16 0:16	Today in Florida, I pledged to stand with the people of Cuba and Venezuela in their <b>fight</b> against oppression- cont:
11/8/16 0:17	Big news to <b>share</b> in New Hampshire tonight! Polls looking great! See you soon. <b>Unbelievable</b> evening in New Hampshire - THANK YOU! Flying to Grand Rapids, Michigan now.
11/8/16 3:43	Watch NH rally here:
11/8/16 4:27	@detroitnews: .@IvankaTrump in Michigan: "This is your movement"™ @realDonaldTrump @DonaldJTrumpJr: Thanks New Hampshire!!!
11/8/16 4:29	#NH #NewHampshire #MAGA
11/8/16 6:42	Today we are going to <b>win</b> the great state of MICHIGAN and we are going to WIN back the White House! Thank you MI!
11/8/16 11:43	TODAY WE MAKE AMERICA GREAT AGAIN!
11/8/16 16:39	VOTE TODAY! Go to to find your polling location. We are going to Make America Great Again! #VoteTrump #ElectionDay
11/8/16 18:03	We need your vote. Go to the POLLS! Let's continue this MOVEMENT! Find your poll location: #ElectionDay #VoteTrump
11/8/16 18:23	#ElectionDay
11/8/16 21:18	I will be watching the election results from Trump Tower in Manhattan with my family and friends. Very exciting!
11/8/16 21:28	Just out according to @CNN: "Utah officials report voting machine problems across entire country"
11/8/16 21:31	Don't let up, keep getting out to vote - this election is FAR FROM OVER! We are doing well but there is much time left. GO FLORIDA! Still time to #VoteTrump!
11/8/16 23:03	#iVoted #ElectionNight @DonaldJTrumpJr: FINAL PUSH! Eric and I doing dozens of radio interviews. We can win this thing! GET OUT AND VOTE! #MAGA
11/8/16 23:20	#ElectionDay htâ€¢; @EricTrump: Join my family in this incredible movement to #MakeAmericaGreatAgain!! Now it is up to you! Please #VOTE for
11/8/16 23:20	America! :â€¢

# Sentiment Analysis

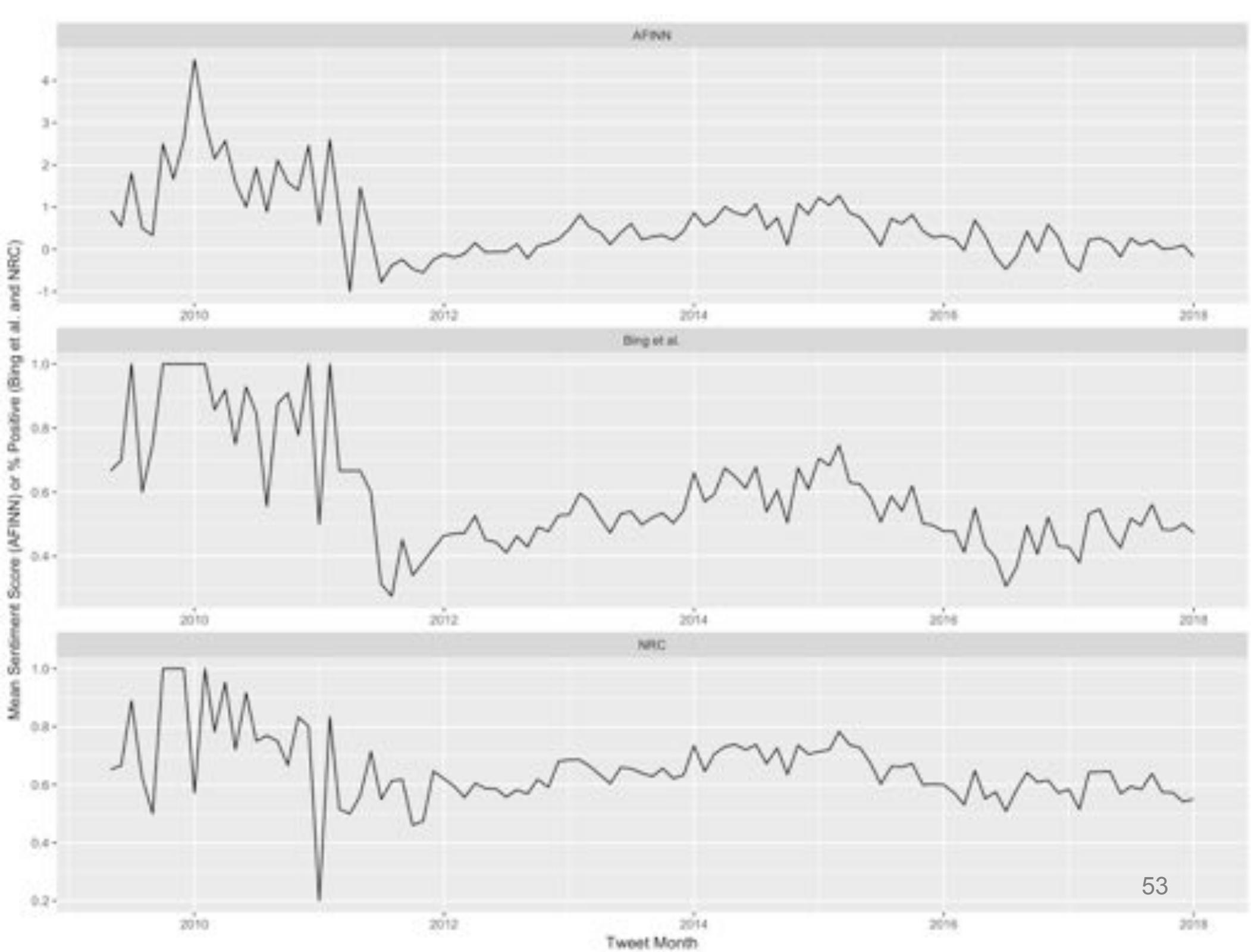
Election Day Tweets: Bing et al.

created_at	word	sentiment
<53: POSIXct>	<chr>	<chr>
2016-11-08 00:16:15	oppression	negative
2016-11-08 03:43:54	unbelievable	negative
2016-11-08 03:43:54	grand	positive
2016-11-08 06:42:36	win	positive
2016-11-08 06:42:36	win	positive
2016-11-08 21:18:04	exciting	positive
2016-11-08 23:20:09	win	positive
2016-11-08 23:20:39	incredible	positive

# Sentiment Analysis

## Election Day Tweets: NRC

created_at	word	sentiment
<53: Positive>	<echo>	<echo>
2016-11-08 00:08:28	vote	negative
2016-11-08 00:08:28	vote	positive
2016-11-08 00:16:15	fight	negative
2016-11-08 00:16:15	oppression	negative
2016-11-08 00:17:57	share	positive
2016-11-08 03:43:54	unbelievable	negative
2016-11-08 03:43:54	flying	positive
2016-11-08 06:42:36	white	positive
2016-11-08 16:39:36	vote	negative
2016-11-08 16:39:36	vote	positive



# Sentiment Analysis

Most Negative Month: NRC

<b>created_at</b>	<b>word</b>	<b>sentiment</b>
<S3: POSIXct>	<chr>	<chr>
2011-01-21 18:25:39	celebrity	negative
2011-01-21 18:25:39	celebrity	positive
2011-01-13 16:14:17	forget	negative
2011-01-13 16:14:17	late	negative
2011-01-12 15:28:31	late	negative

## Most Negative Month by NRC (Jan 2011)

Watch Melania on QVC this morning from 10 a.m. to 11 a.m. with her third line of her "Melania Timepieces & Jewelry" collection...

Busy doing phoners this week with Neil Cavuto, Wolf Blitzer, Fox & Friends, and Larry Kudlow....check out .com/

Don't miss my Fabulous World of Golf now in its second season on Golf Channel beginning tonight at 9 pm ET .ly/gcQjPR

Don't miss my fabulous World of Golf now in its second season on Golf Channel beginning January 31 at 9 pm ET. **Celebrity** matches and more...

Don't **forget** to watch me tonight on **Late** Night with Jimmy Fallon, 12:35 a.m. on NBC. I'll be making a big announcement!

Watch me on **Late** Night with Jimmy Fallon tomorrow night at 12:35 a.m. on NBC--I'll be making a big announcement!

THe people at shouldtrumprun.com have got it right! How are our factories supposed to compete with China and other countries...

...when they have no environmental restrictions! America' s workers need us." .com/

This afternoon I'll be speaking with Neil Cavuto on Your World with Neil Cavuto, 55 p.m. on FOX News.

# Sentiment Analysis

Most Negative Month by NRC using AFINN lexicon

created_at	word	score
<S3: POSIXct>	<chr>	<int>
2011-01-31 14:25:25	miss	-2
2011-01-31 14:25:25	fabulous	4
2011-01-21 18:25:39	miss	-2
2011-01-21 18:25:39	fabulous	4
2011-01-13 16:14:17	forget	-1

## Most Negative Month by NRC (Jan 2011)

Watch Melania on QVC this morning from 10 a.m. to 11 a.m. with her third line of her "Melania Timepieces & Jewelry" collection...

Busy doing phoners this week with Neil Cavuto, Wolf Blitzer, Fox & Friends, and Larry Kudlow....check out .com/

Don't **miss** my **Fabulous** World of Golf now in its second season on Golf Channel beginning tonight at 9 pm ET .ly/gcQjPR

Don't **miss** my **fabulous** World of Golf now in its second season on Golf Channel beginning January 31 at 9 pm ET. Celebrity matches and more...

Don't **forget** to watch me tonight on Late Night with Jimmy Fallon, 12:35 a.m. on NBC. I'll be making a big announcement!

Watch me on Late Night with Jimmy Fallon tomorrow night at 12:35 a.m. on NBC--I'll be making a big announcement!

THe people at shouldtrumprun.com have got it right! How are our factories supposed to compete with China and other countries...

...when they have no environmental restrictions! America' s workers need us." .com/

This afternoon I'll be speaking with Neil Cavuto on Your World with Neil Cavuto, <sup>57</sup> 4 p.m. on FOX News.

# Sentiment Analysis

Most Negative Month by NRC using Bing et al. lexicon

<b>created_at</b>	<b>word</b>	<b>sentiment</b>
<53: POSIXct>	<chr>	<chr>
2011-01-31 14:25:25	miss	negative
2011-01-31 14:25:25	fabulous	positive
2011-01-21 18:25:39	miss	negative
2011-01-21 18:25:39	fabulous	positive

## Most Negative Month by NRC (Jan 2011)

Watch Melania on QVC this morning from 10 a.m. to 11 a.m. with her third line of her "Melania Timepieces & Jewelry" collection...

Busy doing phoners this week with Neil Cavuto, Wolf Blitzer, Fox & Friends, and Larry Kudlow....check out .com/

Don't **miss** my **Fabulous** World of Golf now in its second season on Golf Channel beginning tonight at 9 pm ET .ly/gcQjPR

Don't **miss** my **fabulous** World of Golf now in its second season on Golf Channel beginning January 31 at 9 pm ET. Celebrity matches and more...

Don't forget to watch me tonight on Late Night with Jimmy Fallon, 12:35 a.m. on NBC. I'll be making a big announcement!

Watch me on Late Night with Jimmy Fallon tomorrow night at 12:35 a.m. on NBC--I'll be making a big announcement!

THe people at shouldtrumprun.com have got it right! How are our factories supposed to compete with China and other countries...

...when they have no environmental restrictions! America' s workers need us." .com/

This afternoon I'll be speaking with Neil Cavuto on Your World with Neil Cavuto, 59 p.m. on FOX News.

# Sentiment Analysis

Pitfalls: ?

# Sentiment Analysis

## Pitfalls

- 0 Not every word in the lexicon
- 0 Sentiment depends on context
  - 0 Jane Austen example
  - 0 Solution: Part of Speech tagging

# Methods

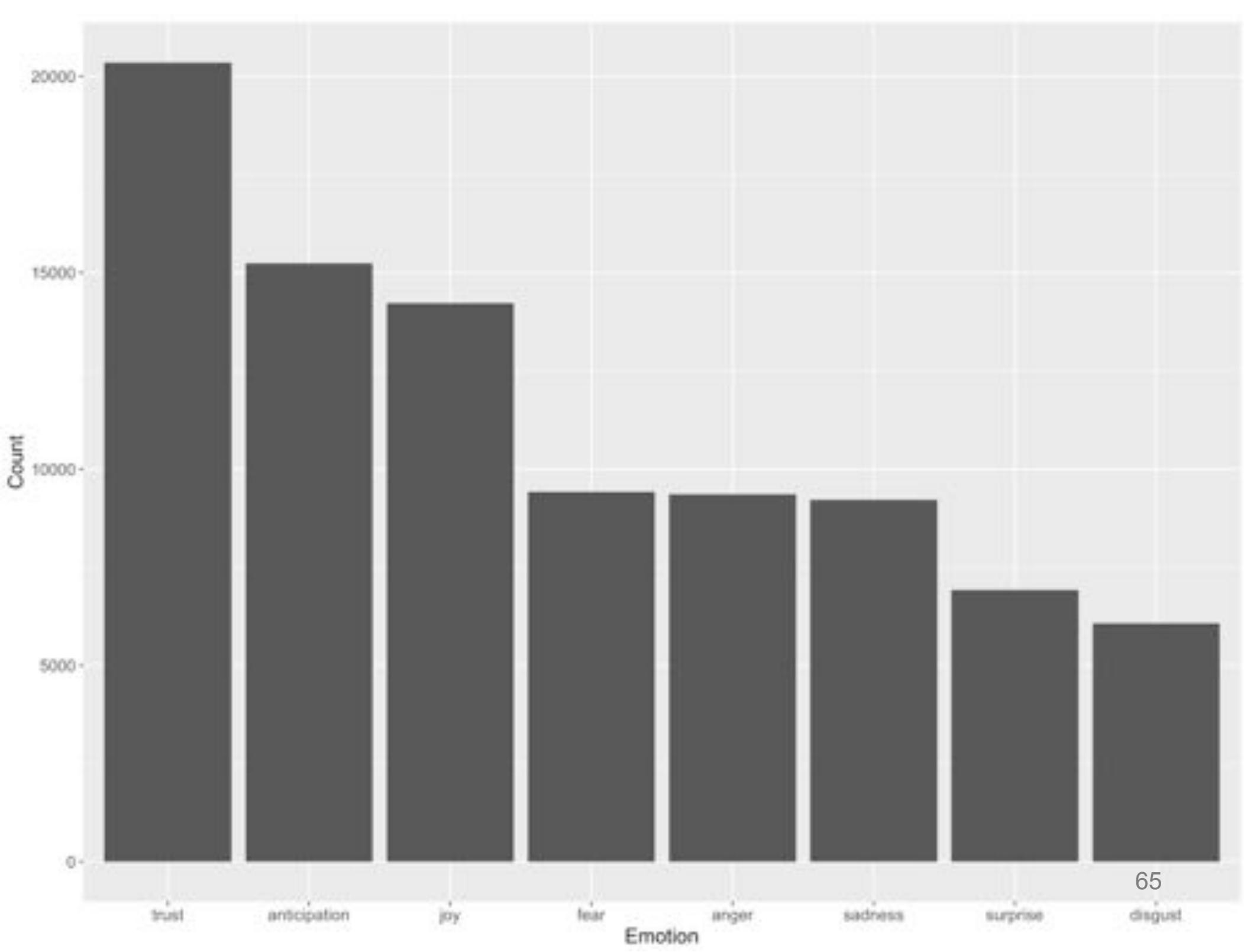
1. Sentiment Analysis
2. Emotion Analysis
3. TF-IDF
4. Topic Modeling

# Emotion Analysis

- 0 Same idea as sentiment analysis but with emotions instead of sentiment
- 0 6 basic emotions?

# Emotion Analysis





# Methods

1. Sentiment Analysis
2. Emotion Analysis
- 3. TF-IDF**
4. Topic Modeling

# TF-IDF

“term frequency-inverse document frequency”

0 Setup: multiple documents

0 Question of Interest 3: What is the underlying content of a document?

# Example Dataset 2

- 0 56 Trump speeches from campaign
  - 0 Dates: June 2015 to November 9, 2016
- 0 Goal: what is the underlying content of each speech?

# Pipeline

## 0 Pipeline Review

# Pipeline

## 0 Pipeline Review

1. Standardize case
2. Tokenization
3. Remove Stop Words
4. Remove Punctuation/Stemming

# Pipeline

## 0 Pipeline Review

1. Standardize case
2. Tokenization
- 3. Remove Stop Words**
4. Remove Punctuation/Stemming
- 5. Calculate word frequency in each Document  
(Method 0)**

## Example Dataset 2 Post-Pipeline

speech_number	word	n
8	i	333
8	the	298
8	to	285
19	and	243
8	a	236
0	the	231
8	and	225
9	the	225
7	the	213
0	and	212

# TF-IDF: General Idea

- 0 Data Level: (Document, Word)
- 0 TF-IDF: How important is this word for this document?
- 0 Term Frequency:
  - 0 How common is this word in this document?
- 0 Inverse Document Frequency:
  - 0 How common is this word in all the other documents?

# TF-IDF: Calculation

0 Data Level: (*document, word*)

0 Term Frequency:

0 # times *word* appears in *document* / total # of words in *document*

0 Inverse Document Frequency:

0  $\ln(\text{total \# of documents} / \text{\# documents containing } \textit{word})$

0  $\text{TF-IDF} = \text{TF} \times \text{IDF}$

# Example

0 Data Level: (Speech 8, “i”)

0 “i” in Speech 8

0 # times “i” appears in Speech 8: 333

0 Total # words in Speech 8: 10,007

0 Total # of documents with the word “i”: 56 (all)

0 Term Frequency: ?

0 Inverse Document Frequency: ?

0 TF-IDF: ?

# Example

0 “i” in Speech 8

0 # times “i” appears in Speech 8: 333

0 Total # words in Speech 8: 10,007

0 Total # of documents with the word “i”: 56 (all)

0 Term Frequency:

0 **# times word appears in document / total number of words in document**

0 Inverse Document Frequency:

0 **ln(# of documents / # documents containing term)**

# Example

0 Term Frequency:

0 # times word appears in document / total number of words in document

$$0 \frac{333}{10,007} = 0.0333$$

0 Inverse Document Frequency:

0  $\ln(\# \text{ of documents} / \# \text{ documents containing term})$

$$0 \ln(56 / 56) = \ln(1) = 0$$

0  $\text{tf-idf} = 0.0333 \times 0 = 0$

# Example

0 “israel” in Speech 1

0 # times “israel” appears in Speech 1: 37

0 Total # words in Speech 1: 2,305

0 Total # of documents with the word “israel”: 6

0 Term Frequency:  $37 / 2305 = 0.0161$

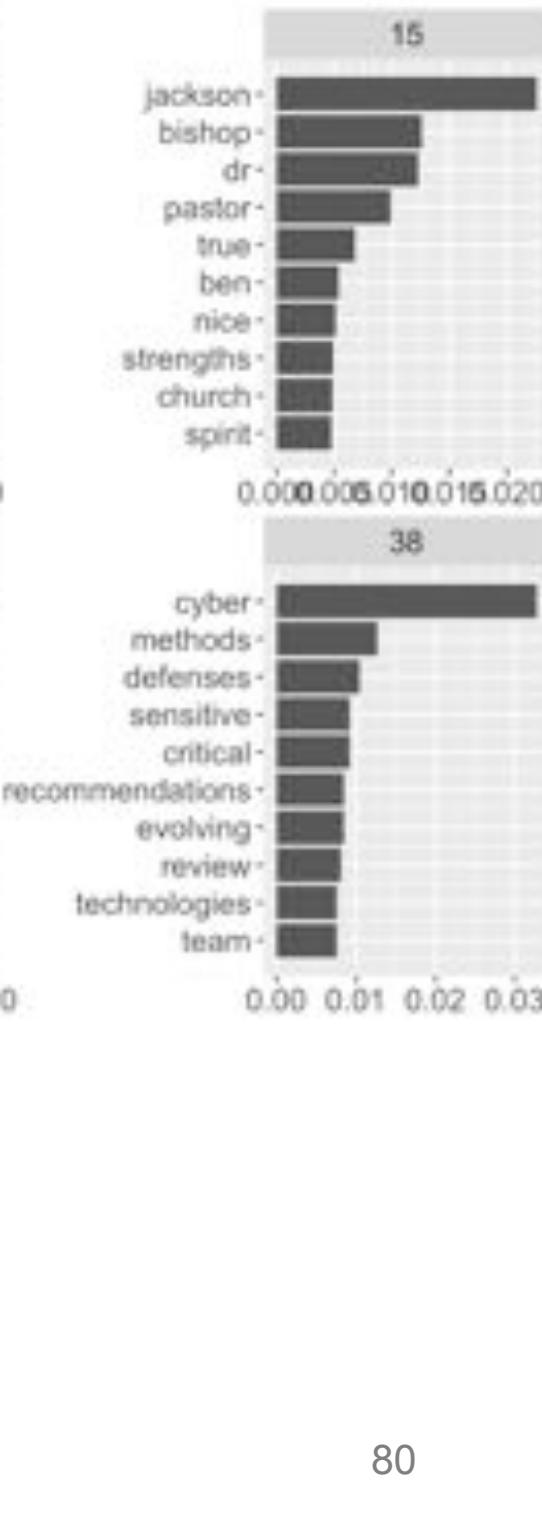
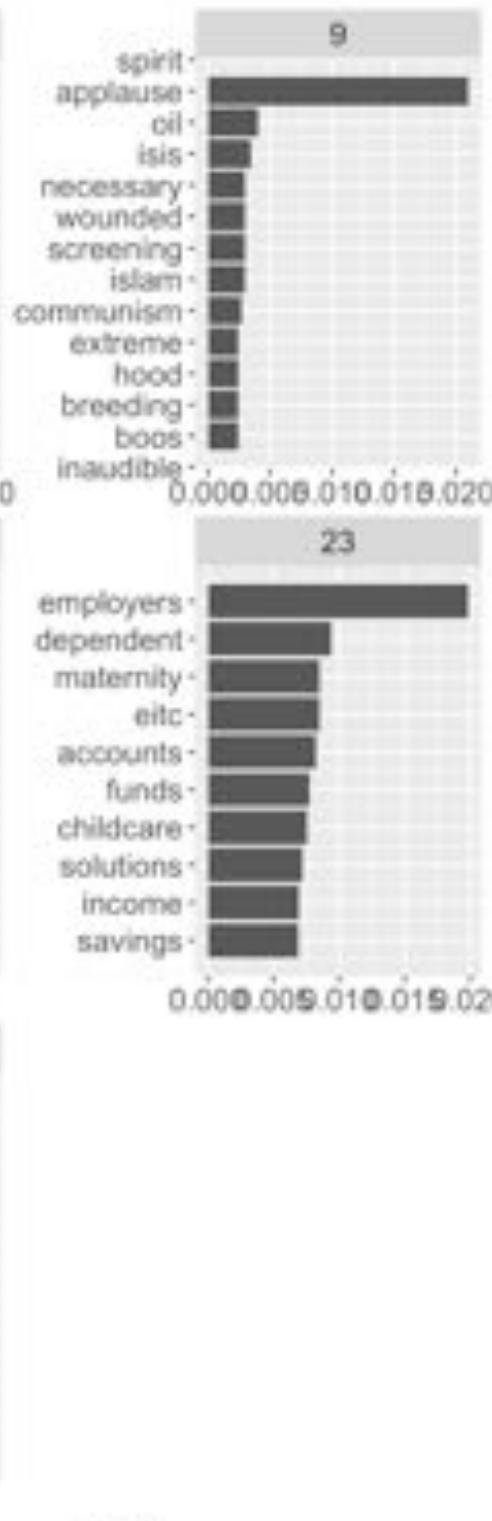
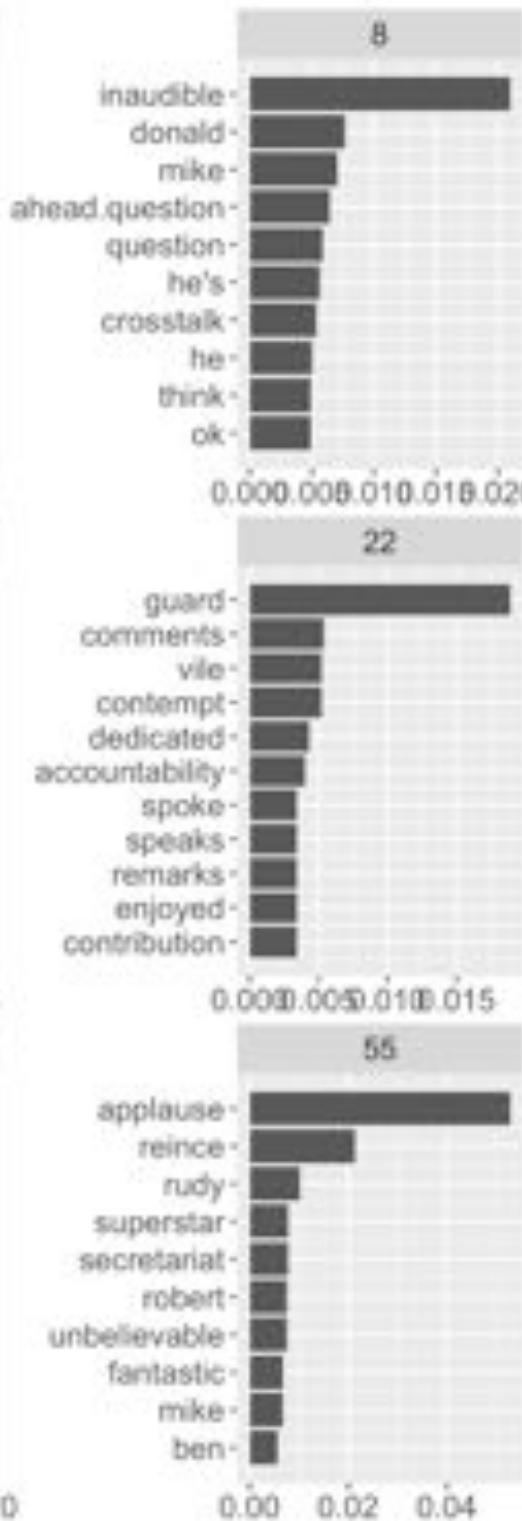
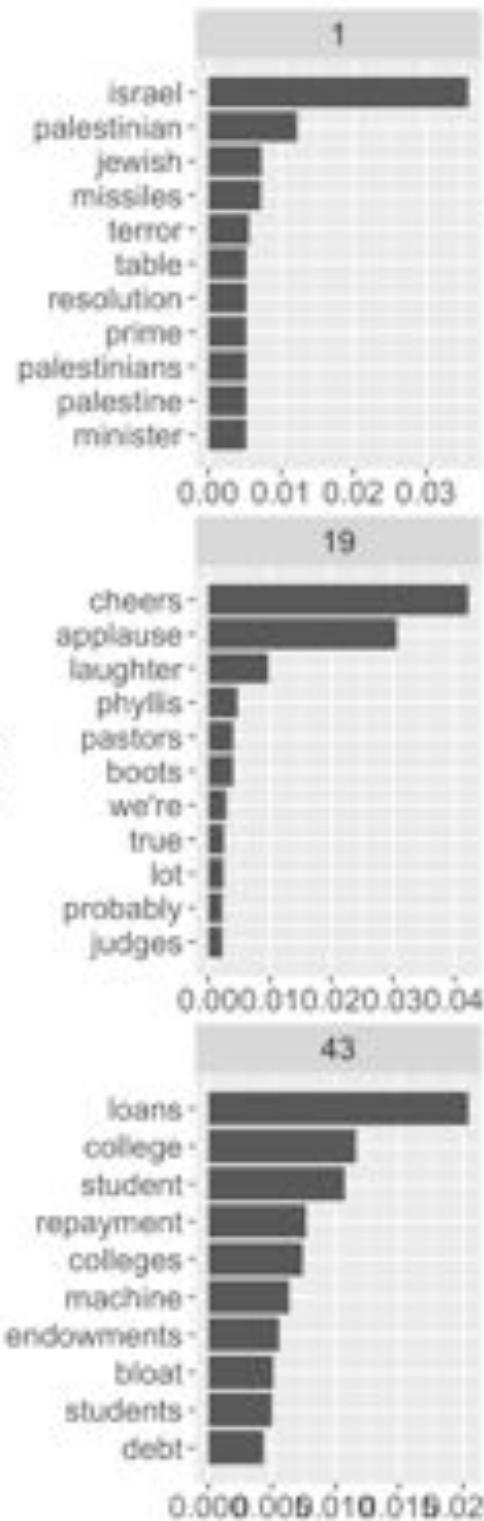
0 Inverse Document Frequency:  $\ln(56 / 6) = 2.234$

0 tf-idf:  $0.0161 \times 2.234 = 0.0359$

# Results

0 Only looked at 10 speeches with highest tf-idf scores

order



tf\_idf

# Pitfalls: ?

# Pitfalls

- 0 Subject to differences in text format (i.e. [applause] included) across documents
- 0 Sensitive to typos
- 0 Sensitive to proper names
- 0 Doesn't take position into account
  - 0 Beginning, middle, end of speech indicates importance
  - 0 N-grams: "don't vote"

# Methods

- 1. Sentiment Analysis**
- 2. Emotion Analysis**
- 3. TF-IDF**
- 4. Topic Modeling**

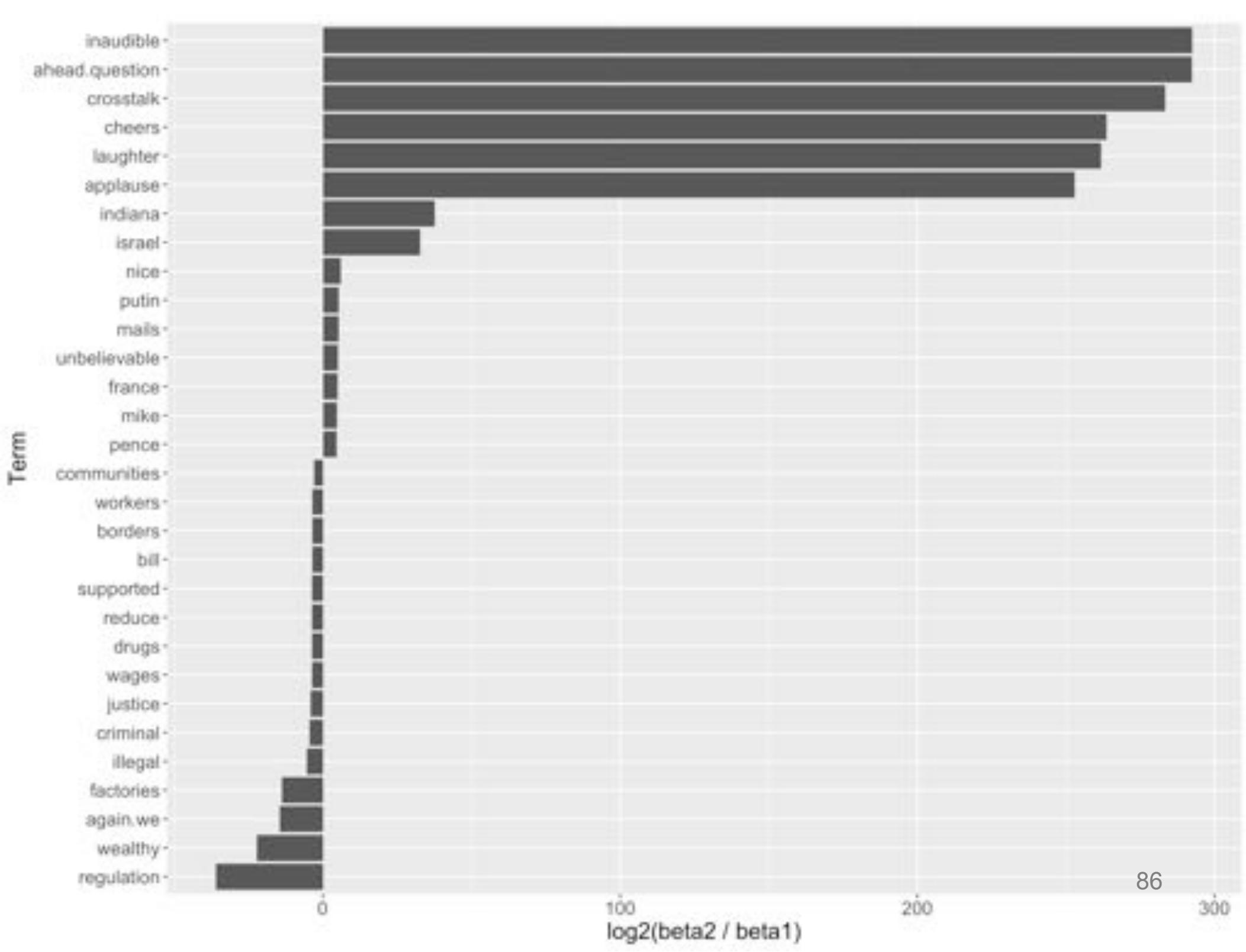
# General Idea

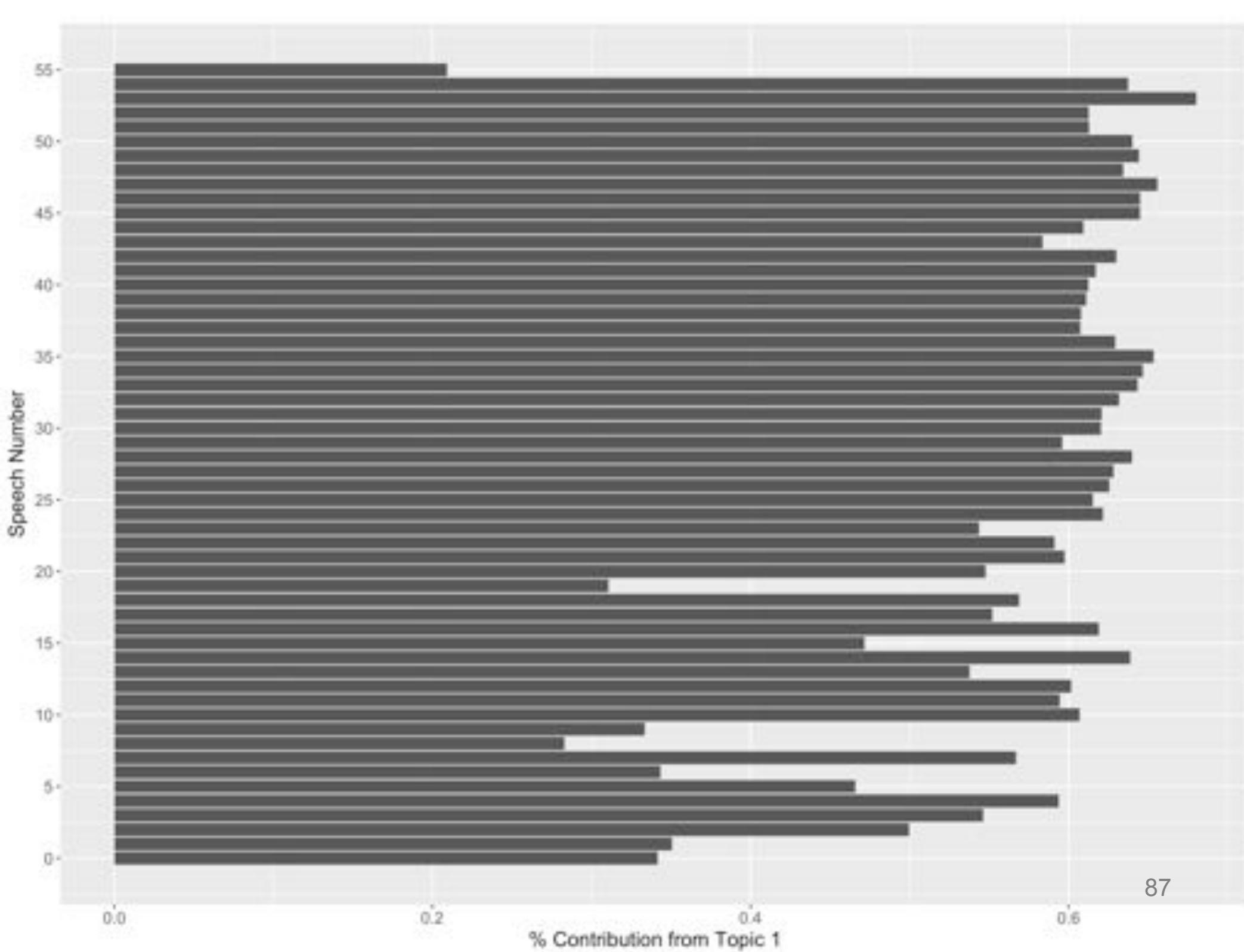
- 0 Setup: several documents
- 0 Question of Interest 4: Which documents are similar to each other?

# Latent Dirichlet Allocation

## 0 Principles

- 0 Every topic is a mixture of words
- 0 Every document is a mixture of topics





# Bonus

# Bonus

0 Natural Language Processing: computer able to understand natural human language

0 Google Assistant

0 Text Generation

0 Shakespeare lines

# Google Assistant

<https://www.cnbc.com/video/2018/05/08/googles-assistant-can-place-phone-calls-and-fool-humans.html>

generated\_output.txt - Editor

Datei Bearbeiten Format Ansicht T

```
HE there, this a generated output poem from the shakespeare machine. have fun!
```

AND CLIFFORD                    He is no more, madam.

LORD POLONIUS    That's the prince of Cassius, and I will be so. I have a  
                      beauty to my lord, and there is a  
                      prince and the command of the bosom of the world.

FALSTAFF                  I am sorry than the poot silent of the world.

LADY MACBETH          I have say'st, a man in my particular, I'll be a montest that I may be a most cause. Therefore, I  
                      would no more be come to this, and I would not have any  
                      tongue to be so found, I can tell to the court of my lord.

# Questions?

# Thank you!

Let's take a selfie

# References

## (Text Analytics, Data)

- 0 <https://www.tidytextmining.com>
- 0 Data source (Trump twitter)
- 0 [https://data.world/data-society/major-speeches-by-donald-trump/workspace/file?filename=speech\\_9.txt](https://data.world/data-society/major-speeches-by-donald-trump/workspace/file?filename=speech_9.txt)
- 0 <https://hlt-nlp.fbk.eu/technologies/sentiwords> (specific instructions for how to cite this)
- 0 <https://searchbusinessanalytics.techtarget.com/definition/natural-language-processing-NLP>
- 0 <https://towardsdatascience.com/deep-learning-with-tensorflow-part-3-music-and-text-generation-8a3fbfdc5e9b>

# References (Images)

- 0 <https://www.pixarpost.com/2014/11/inside-out-character-profiles-anger-joy.html>
- 0 <https://nypost.com/2018/08/01/trump-agrees-to-let-insurers-sell-cheaper-health-plans-than-obamacare/>
- 0 <https://twitter.com/twitter>