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

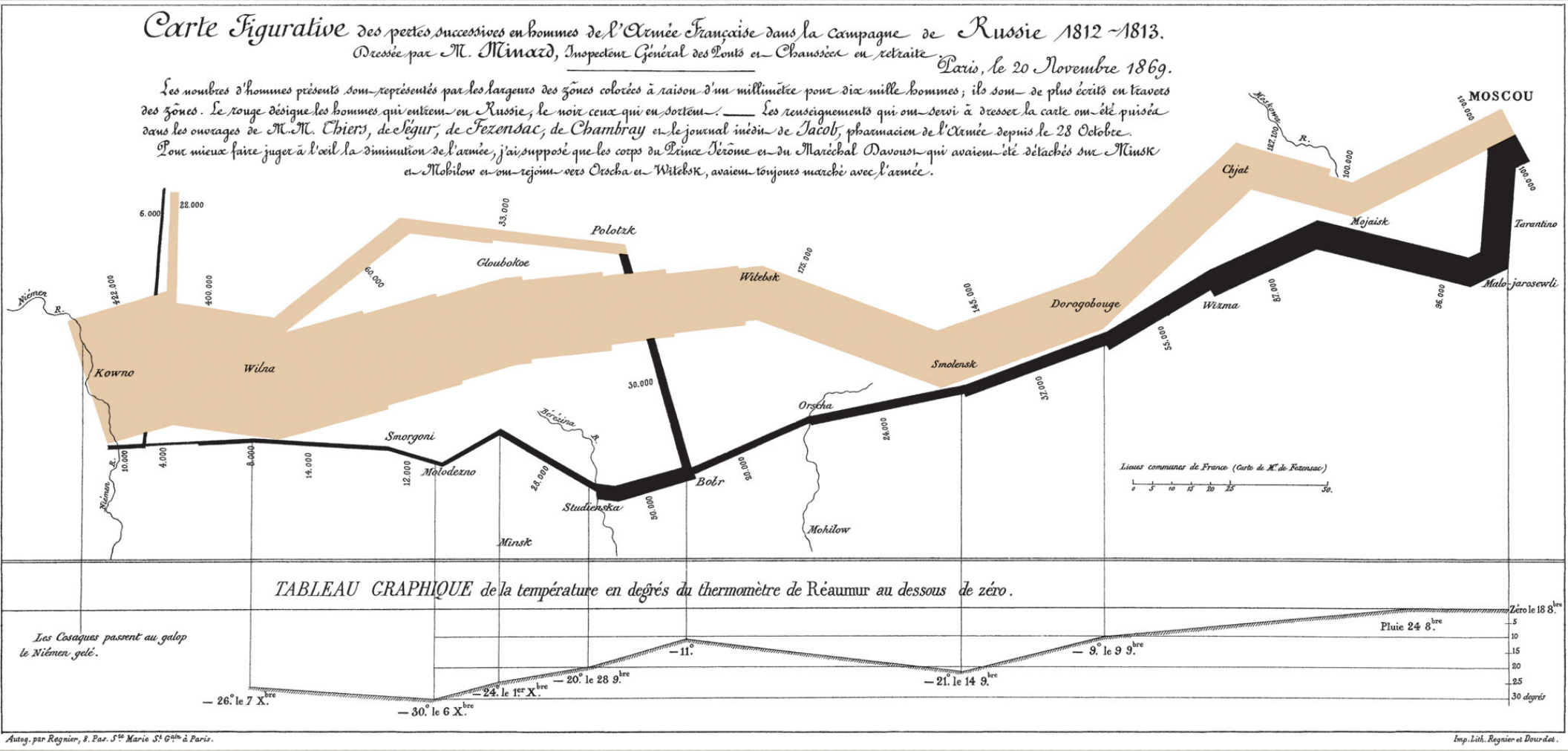
Overview

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

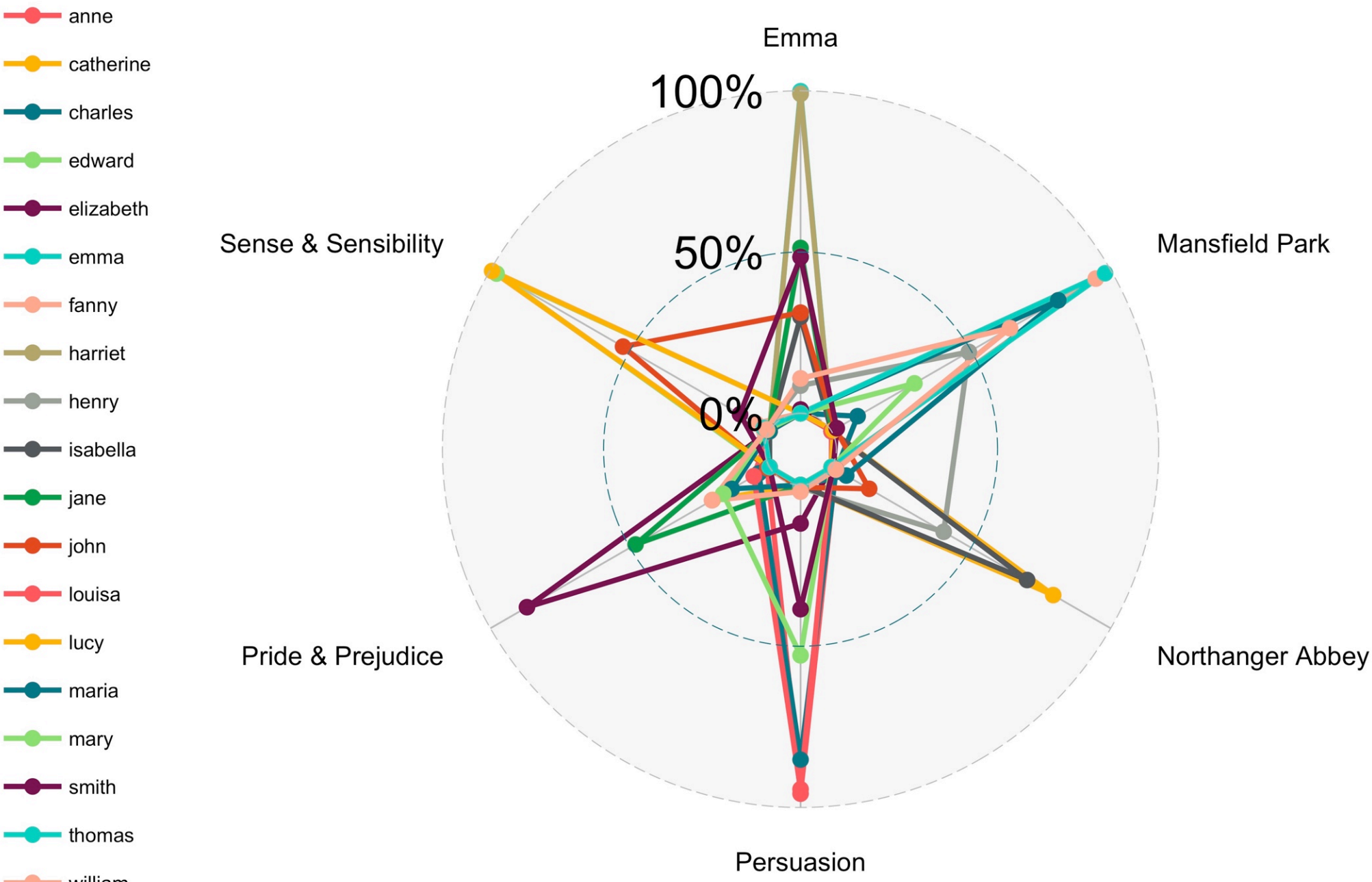
What is data visualization?

Is this a data visualization?

Napoleon's Moscow Invasion




re
Hillary
bad
Obama
think
work
always
person
Okay
say
back
talk
two
go
good
take
president
America
never
want
great
win
need
ago
every
probably
Iowa
make
last
lot
big
really
put
give
made
deal
world
week
coming
everybody
Clinton
number
something
remember
bring
job
United
right
Oh
believe
getting
maybe
mean
happen
year
tremendous
ever
poll
going
New
time
much
China
million
us
Thank
talking
money
guy
well
billion
way
now
one
nice
nobody
incredible
Trump
thing
even
many
come
little
care
build
know
country
came
will
tell
love
said
problem
saying
place
look
see
people
let
got
State
wall
done
actually
day
Mexico
trade


anne
catherine
charles
edward
elizabeth
emma
fanny
harriet
henry
isabella
jane
john
louisa
lucy
maria
mary
smith
thomas
william
Emma
100%
50%
0%
Mansfield Park
Northanger Abbey
Persuasion
Pride & Prejudice
Sense & Sensibility

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"


re
Hillary
bad
Obama
think
work
always
person
Okay
say
back
first
go
talk
two
take
good
never
president
America
want
great
win
need
make
lotbig
really
deal
world
everybody
number
something
remember
job
right
believe
mean
happen
year
ever
poll
going
time
much
China
million
usThank
money
guy
well
way
now
one
Trump
thing
even
many
come
know
country
love
will
tell
said
look
see
people
let
got
State
wall
done
day
actually
Mexico
trade
problem
little
care
build
talking
billion
nobody
incredible
saying
place
better
somebody

Why visualize data?

Why visualize data?

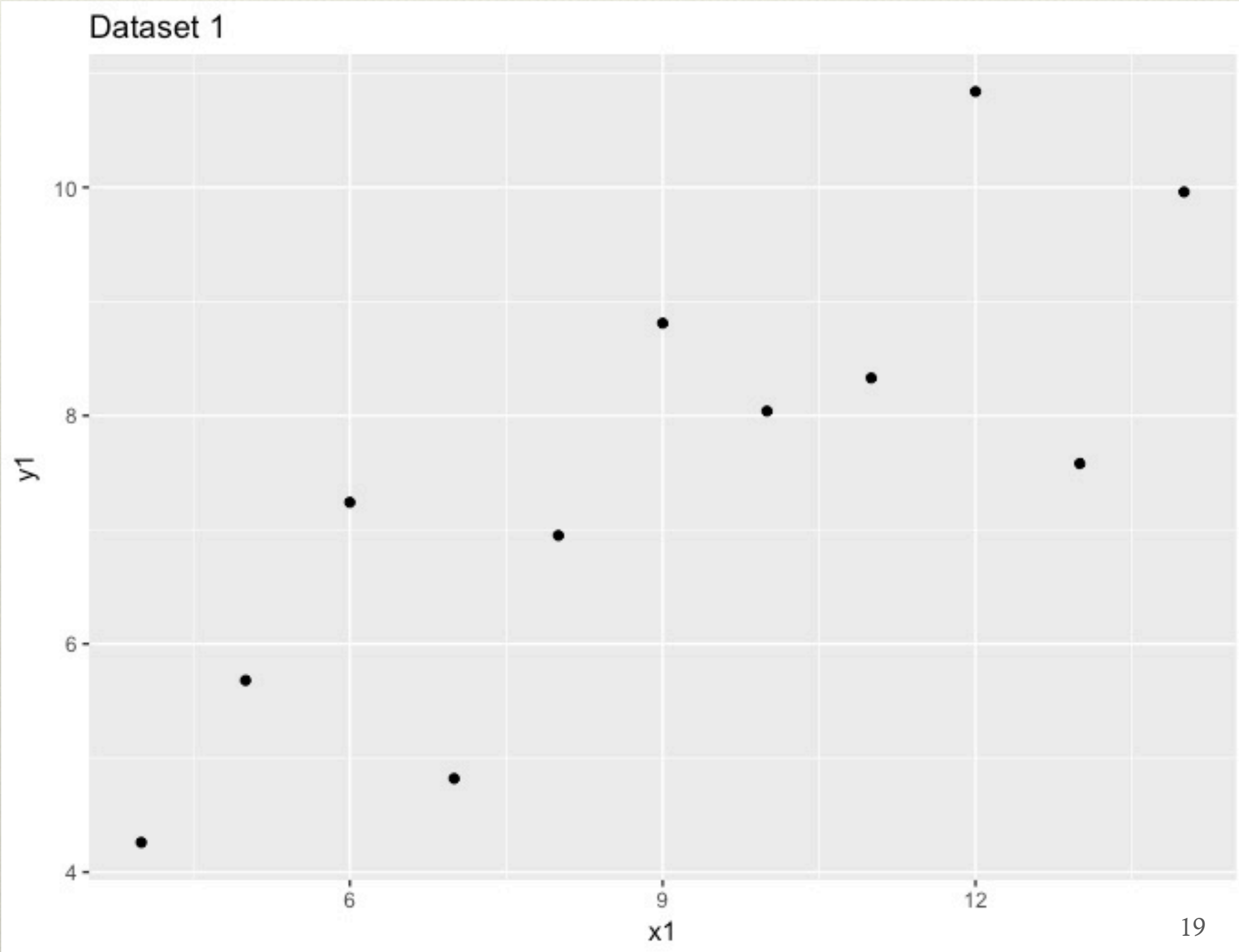
"a picture is worth a thousand words"

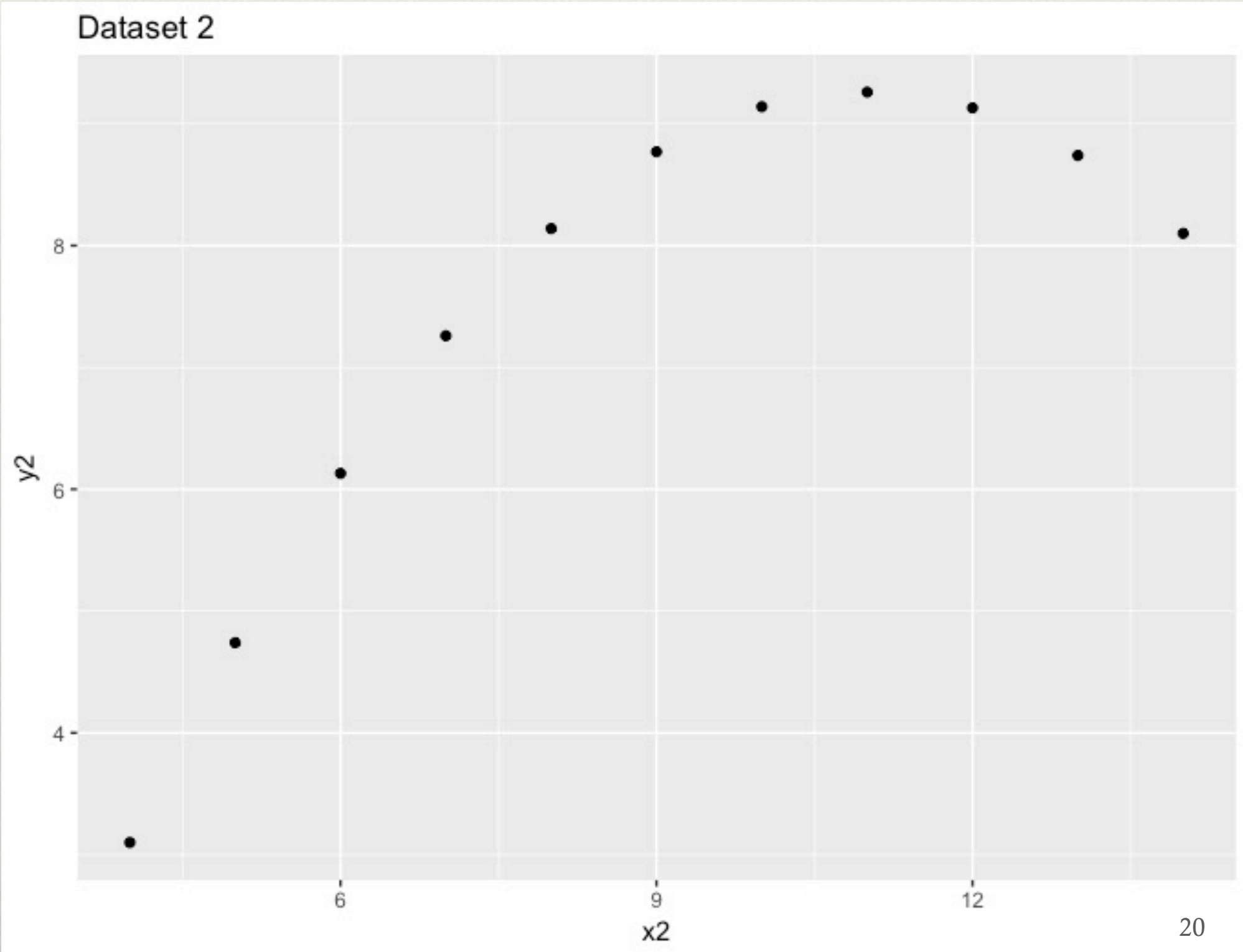
Why visualize data?

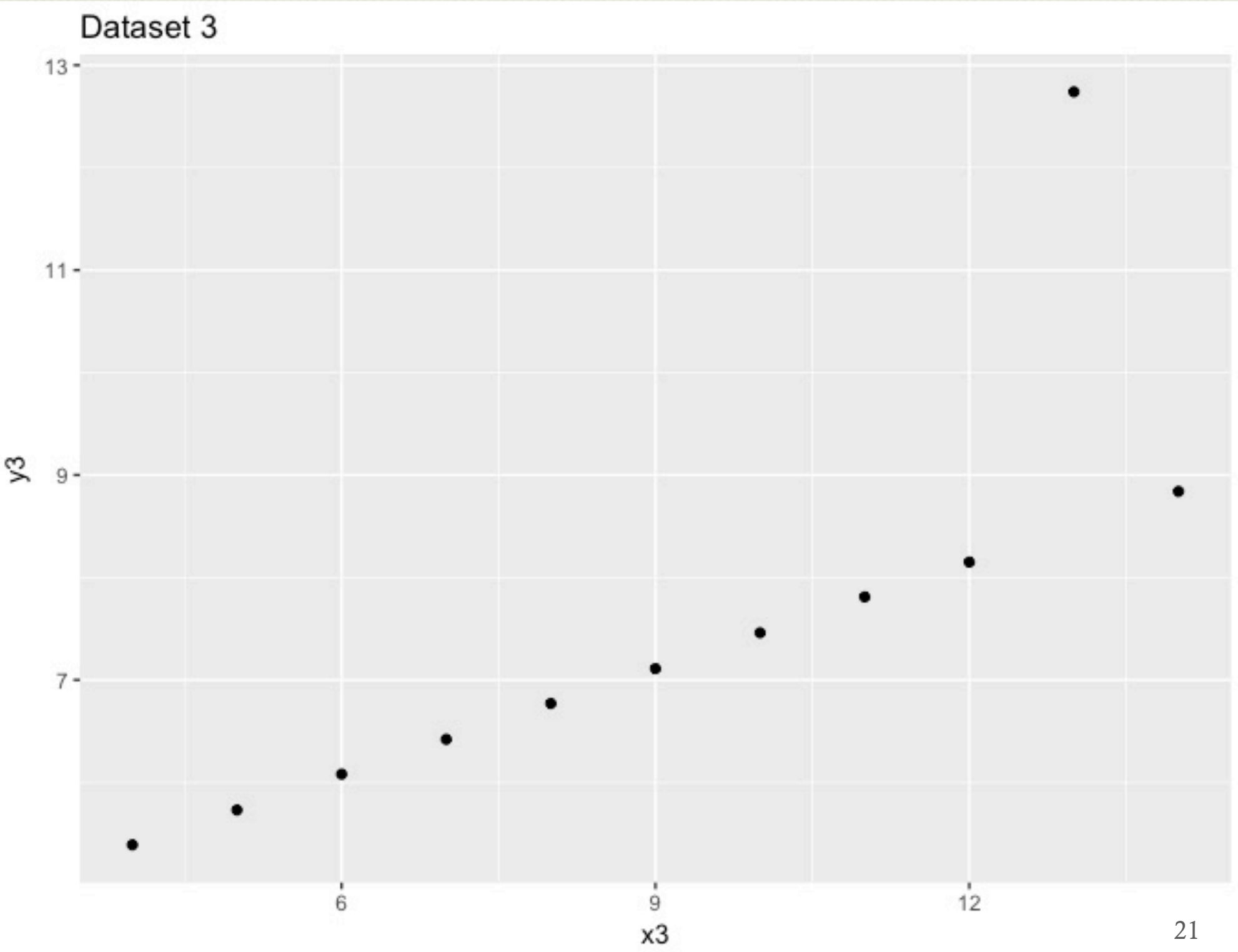
- Identify
  - Trends
  - Relationships
  - Outliers

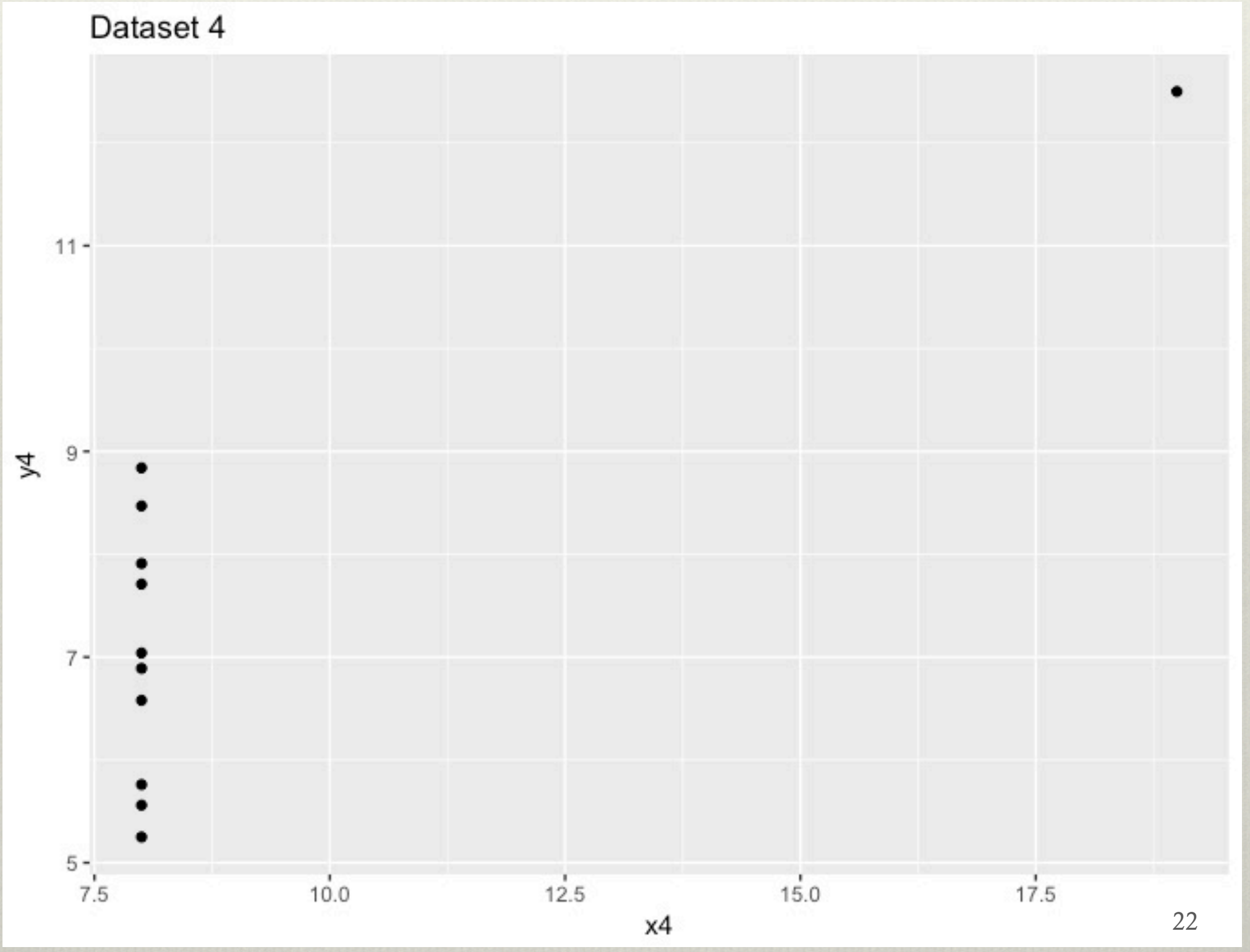
Anscombe Dataset

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
10
8
6
4
y1
6
9
12
x1

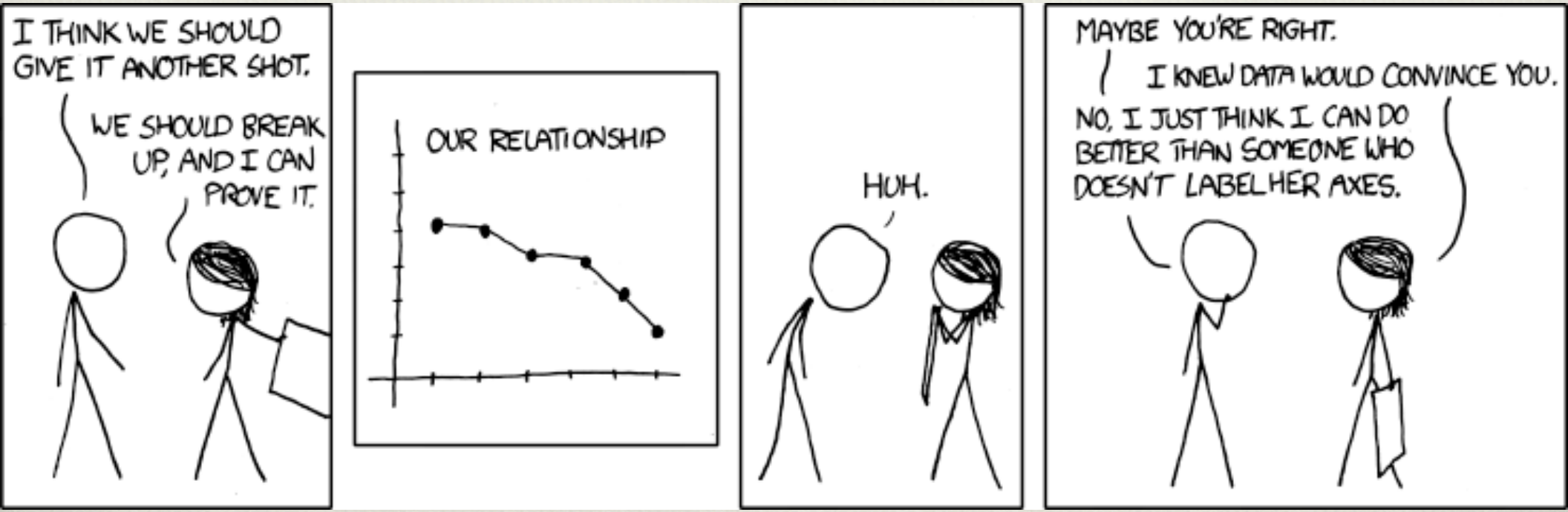

Dataset 2
y2
8
6
4
6
9
12
x2


Dataset 3
13
11
9
7
y3
6
9
12
x3


Dataset 4
11
9
7
5
y4
7.5
10.0
12.5
15.0
17.5
x4

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#!/vizhome/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**

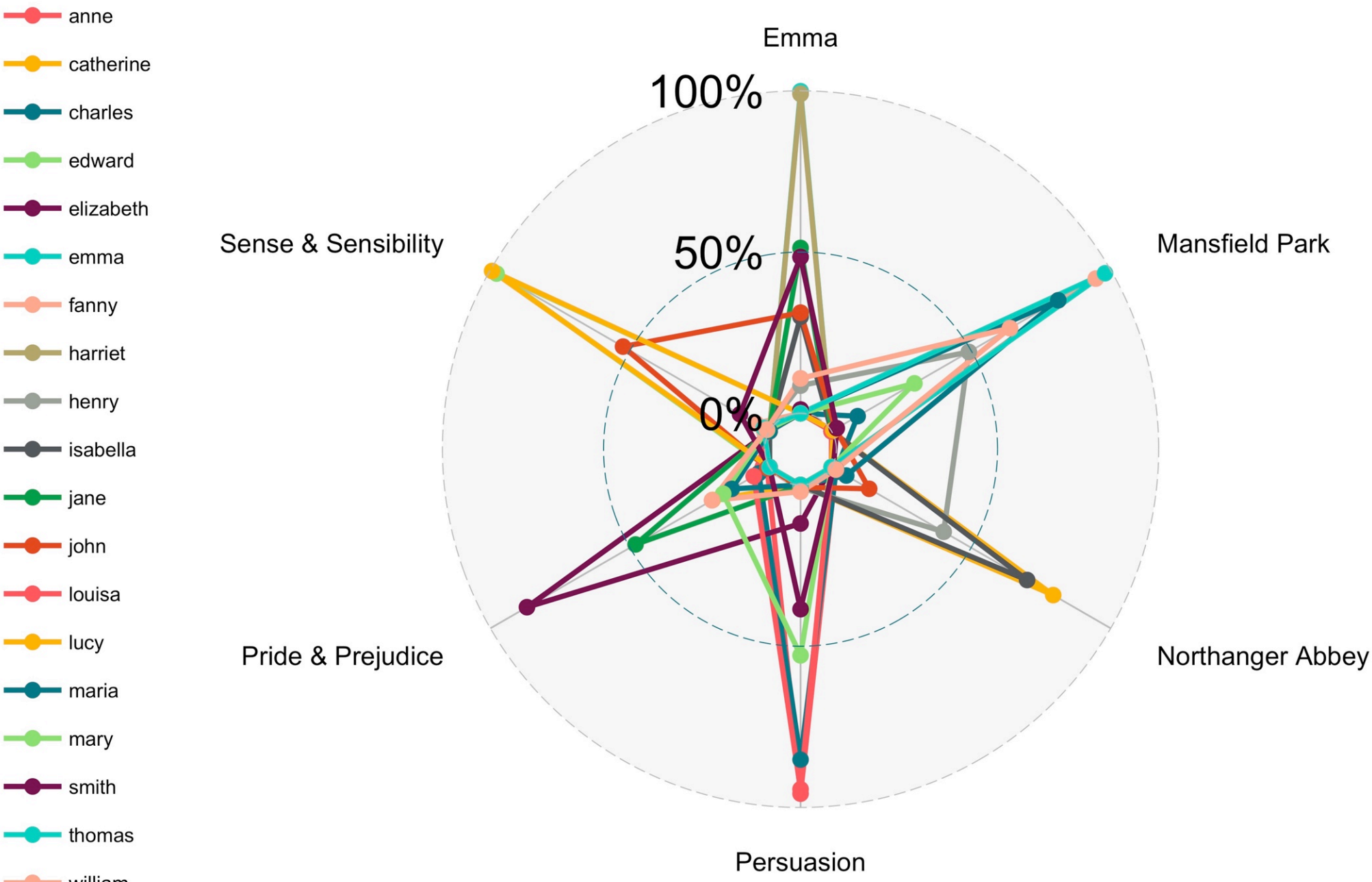

anne
catherine
charles
edward
elizabeth
emma
fanny
harriet
henry
isabella
jane
john
louisa
lucy
maria
mary
smith
thomas
william
Emma
100%
50%
0%
Mansfield Park
Northanger Abbey
Persuasion
Pride & Prejudice
Sense & Sensibility

Chart Type

- One **continuous** variable
  - MR20%

Chart Type

- One **continuous** variable
  - MR20%

Histogram

Boxplot

Chart Type

- One **categorical** variable
  - Tier

Chart Type

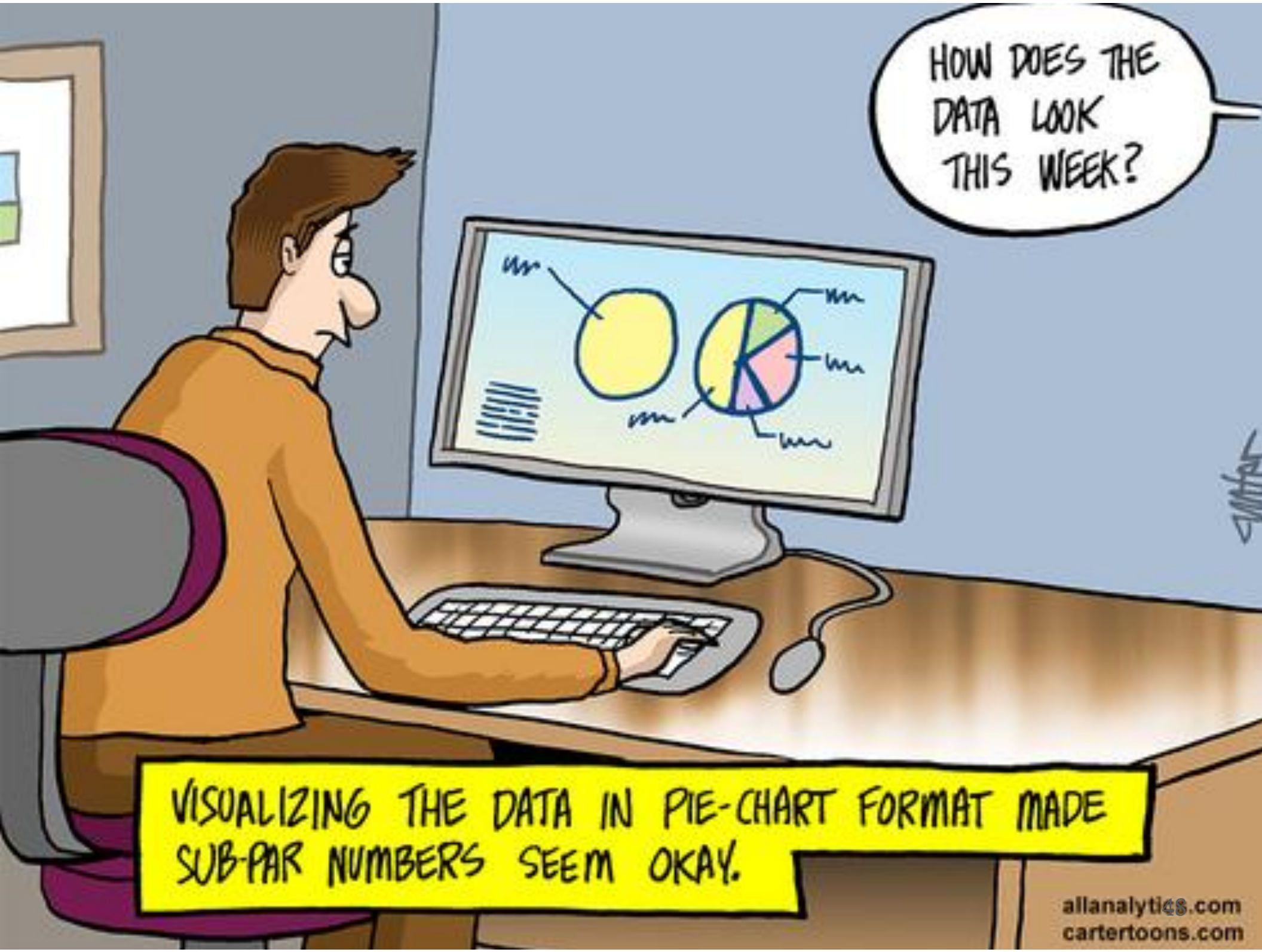
- One **categorical** variable
  - Tier

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 SUB-PAR NUMBERS SEEM OKAY.
allanalytics.com
cartertoons.com

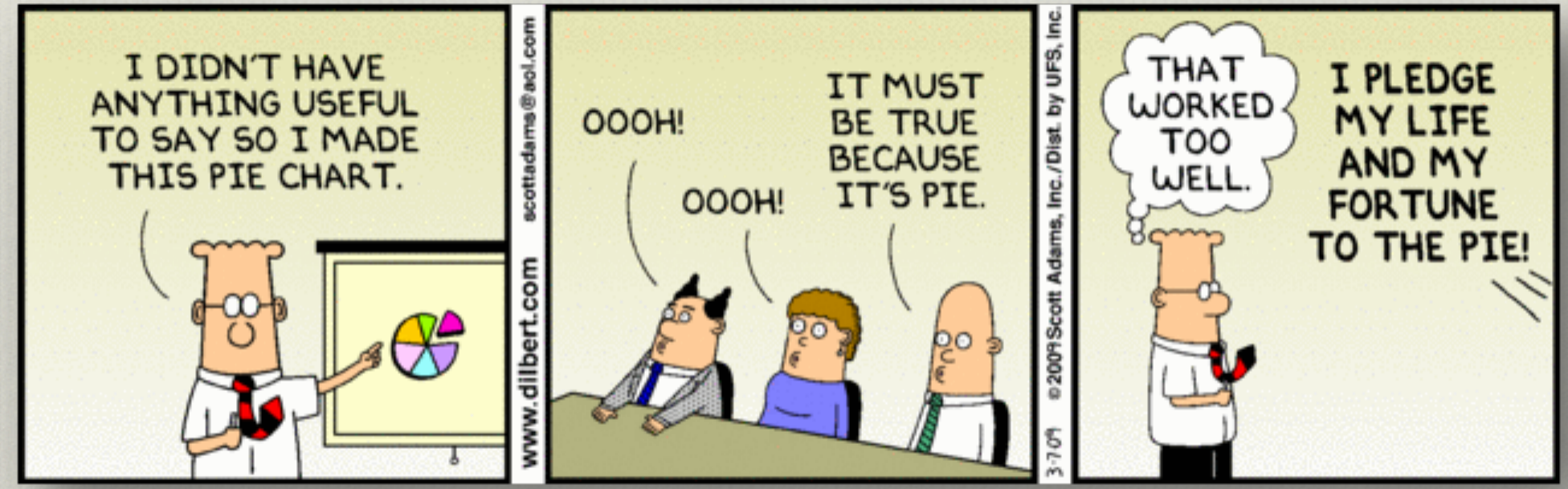

I DIDN'T HAVE ANYTHING USEFUL TO SAY SO I MADE THIS PIE CHART.
www.dilbert.com scottadams@aol.com
OOOH!
OOOH!
IT MUST BE TRUE BECAUSE IT'S PIE.
3-10
©2009 Scott Adams, Inc./Dist. by UFS, Inc.
THAT WORKED TOO WELL.
I PLEDGE MY LIFE AND MY FORTUNE TO THE PIE!

Chart Type

~~Pie Chart?~~

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

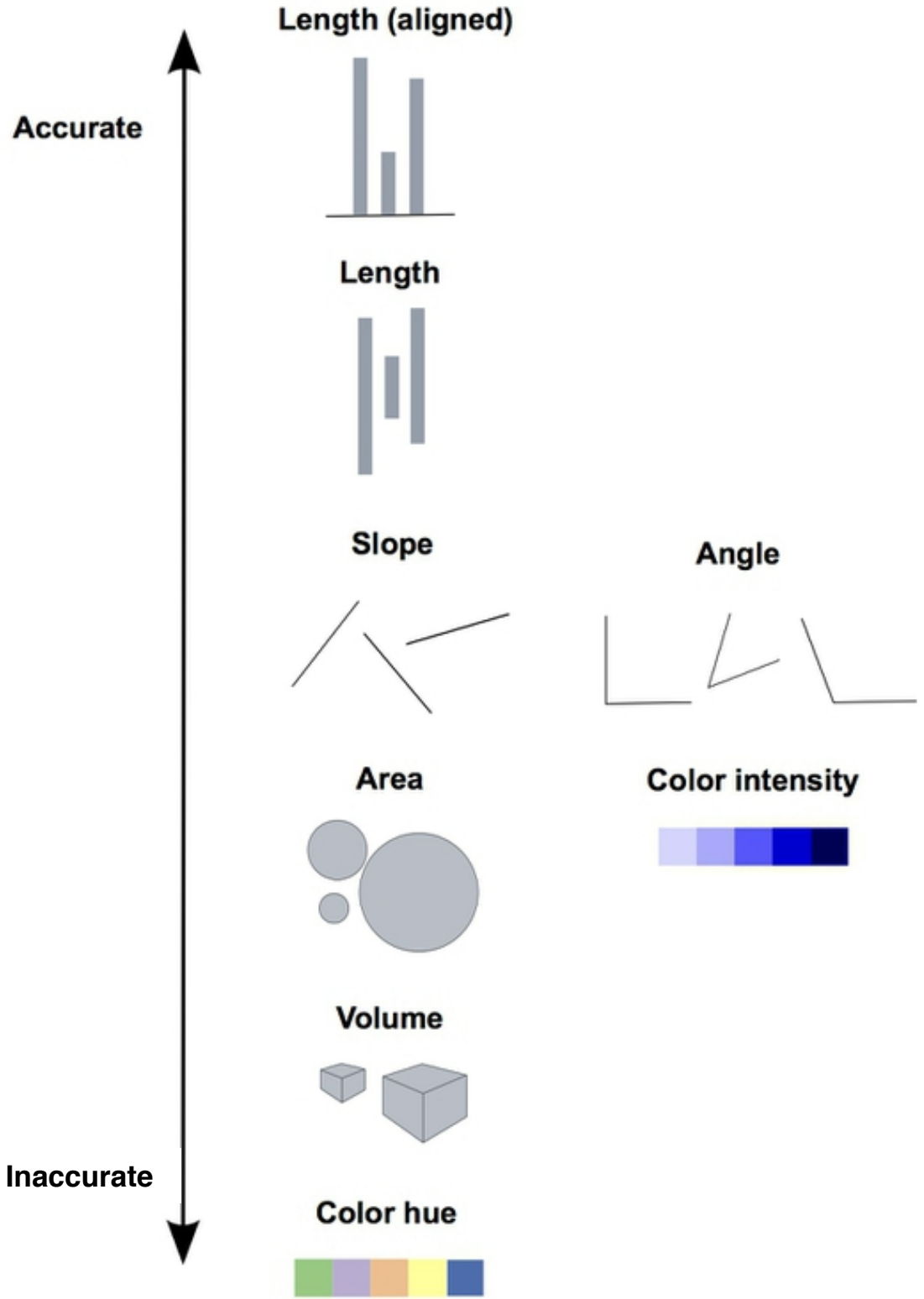

Accurate
Inaccurate
Length (aligned)
Length
Slope
Angle
Area
Color intensity
Volume
Color hue

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

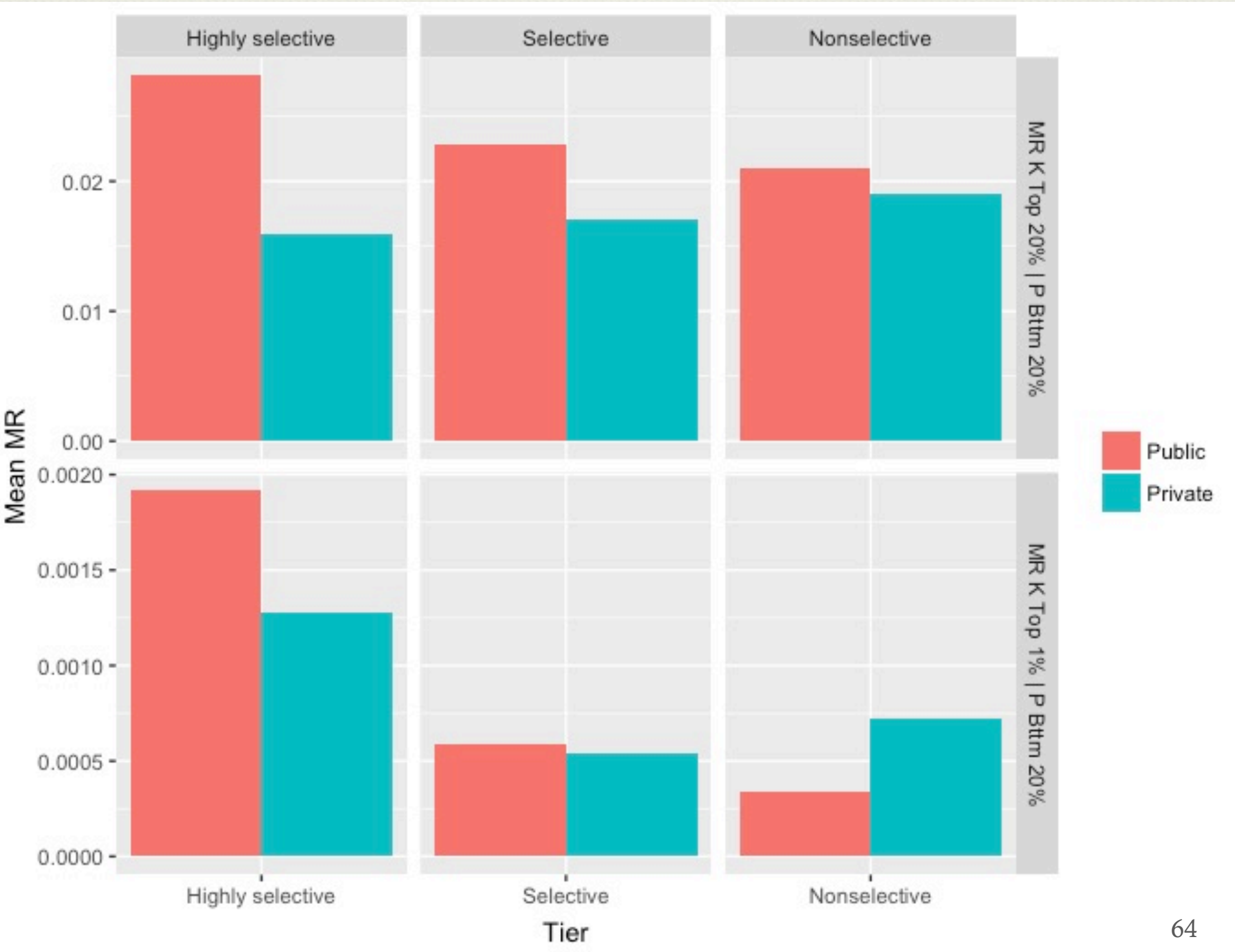
Note: I simplified the tiers

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


Highly selective
Selective
Nonselective
MR K Top 20% | P Bttm 20%
MR K Top 1% | P Bttm 20%
Mean MR
0.02
0.01
0.00
0.0020
0.0015
0.0010
0.0005
0.0000
Highly selective
Selective
Nonselective
Tier
Public
Private

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://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

Text Analytics

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

Overview

- **What** is text analytics?
- **How** to analyze text?
  - Pipeline
  - Methods
    - 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?

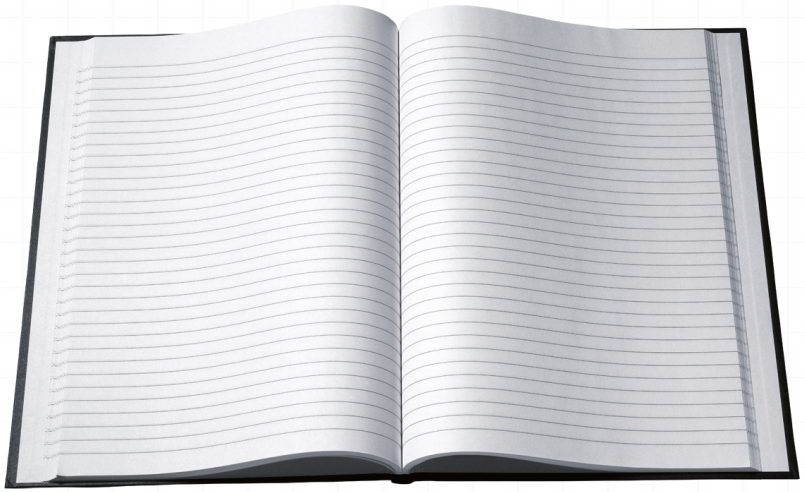
"drawing meaning out of written communication"

Oldest form:

What is text analytics?

"drawing meaning out of written communication"

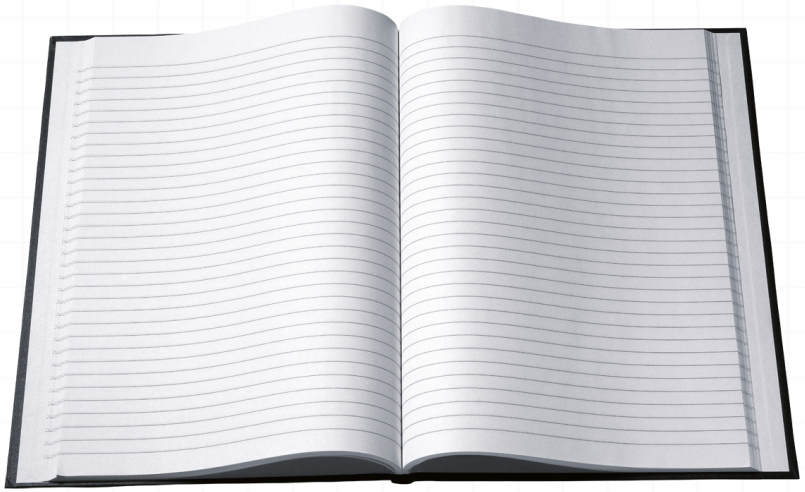
Oldest form: **reading**



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:

- Books
- Surveys
- Emails
- Twitter
- State of Union speeches
- Professor/class reviews
- Yelp reviews
- 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:

- Businessman: Trump Organization, Trump Tower
- Miss Universe and Miss USA pageants
- Reality TV Producer of *Apprentice*, *Celebrity Apprentice*
- 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

- Trump's Tweets from May 2009 – Jan 2018
- Variables:
  - Created_at (date-time)
  - 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: .@IvankaTrump in Michigan: 'This is yo…
6	2016-11-08 04:29:10	@DonaldJTrumpJr: Thanks New Hampshire!!! #NH #N…
7	2016-11-08 06:42:36	Today we are going to win the great state of MICHIGA…
8	2016-11-08 11:43:14	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! #iVoted #ElectionNight

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	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.
11/8/16 3:43	Unbelievable evening in New Hampshire - THANK YOU! Flying to Grand Rapids, Michigan now. Watch NH rally here:
11/8/16 4:27	@detroitnews: .@IvankaTrump in Michigan: â€˜This is your movementâ€™ @realDonaldTrump
11/8/16 4:29	@DonaldJTrumpJr: Thanks New Hampshire!!! #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!
11/8/16 23:03	Still time to #VoteTrump! #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 #ElectionDay htâ€¦
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! :â€¦

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	**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.
11/8/16 3:43	Unbelievable evening in New Hampshire - THANK YOU! Flying to Grand Rapids, Michigan now. Watch NH rally here:
11/8/16 4:27	@detroitnews: .@IvankaTrump in Michigan: â€˜This is your movementâ€™ @realDonaldTrump
11/8/16 4:29	@DonaldJTrumpJr: Thanks New Hampshire!!! #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!
11/8/16 23:03	Still time to #VoteTrump! #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 #ElectionDay htâ€¦
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! :â€¦

Pipeline

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

Pipeline: Tokenization

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

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

Pipeline: Tokenization

"Still time to #VoteTrump! #IVoted #ElectionNight

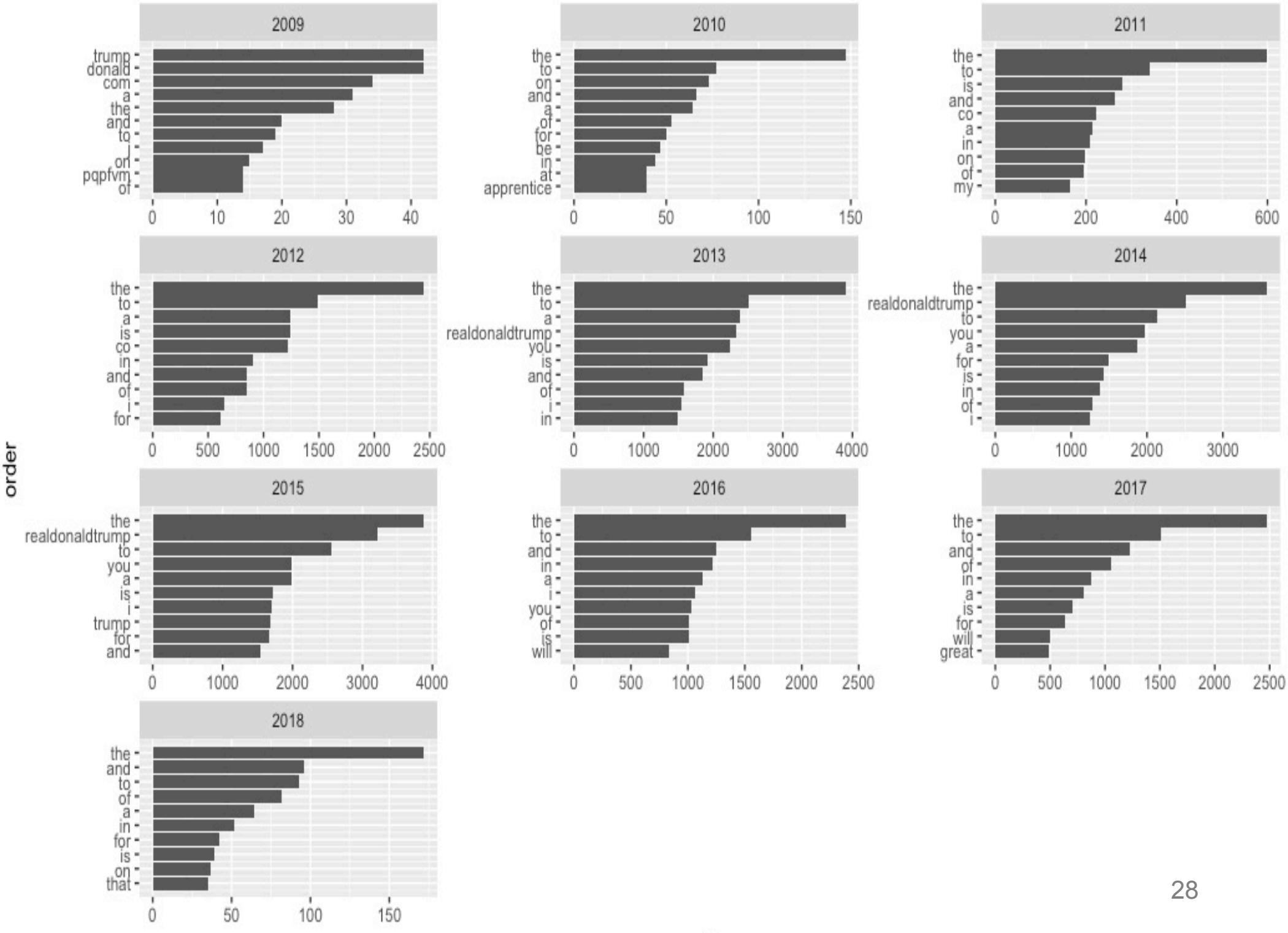
Separator: space	Separator: #
Still	Still time to
time	VoteTrump!
to	IVoted
#VoteTrump!	ElectionNight
#IVoted	
#ElectionNight	

Pipeline: Tokenization

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

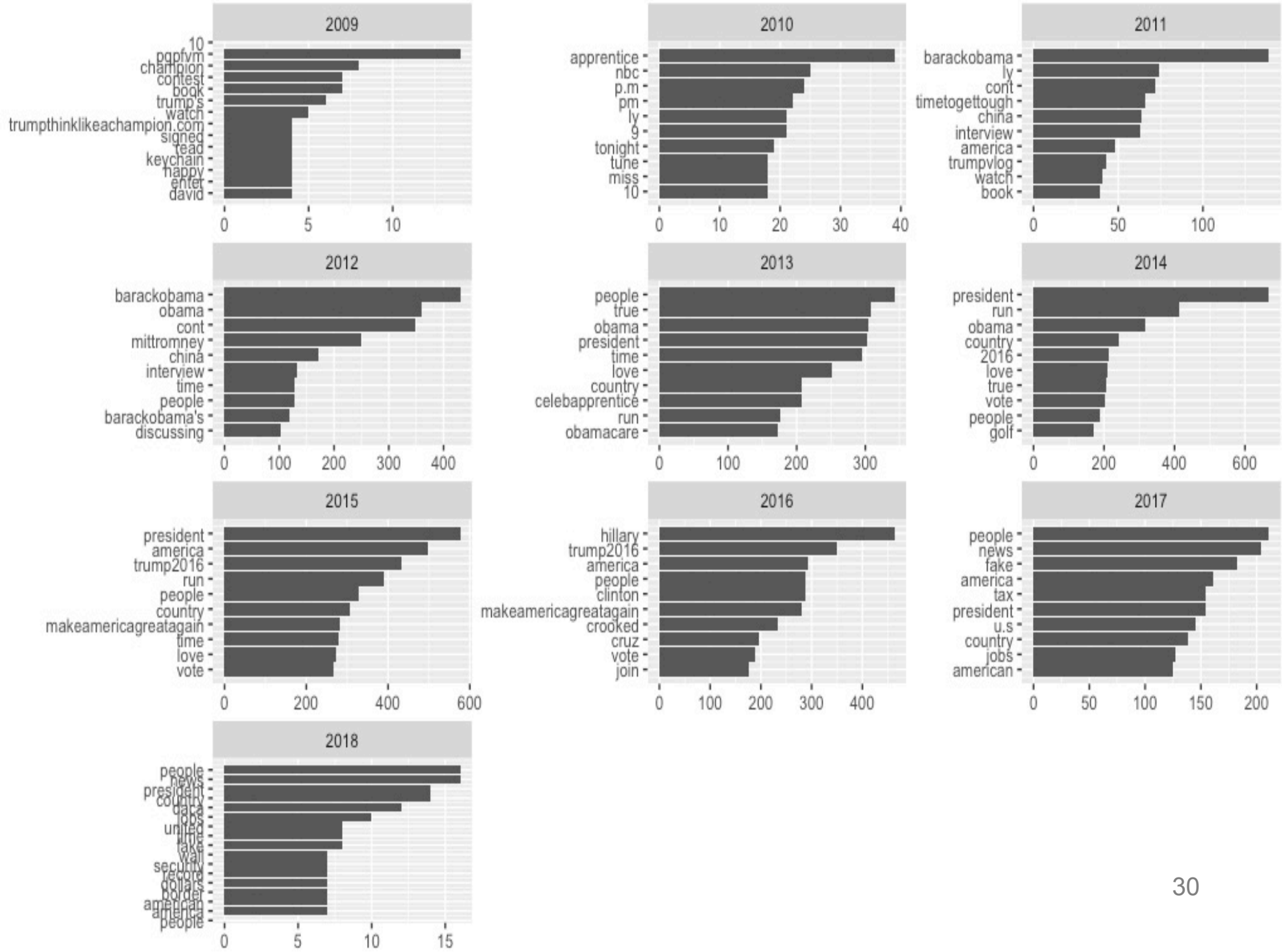
Pipeline: ?

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


2009
trump
donald
com
a
the
and
to
i
on
pqpfvm
of
0
10
20
30
40
2010
the
to
on
and
a
of
for
be
in
at
apprentice
0
50
100
150
2011
the
to
is
and
co
a
in
on
of
my
0
200
400
600
2012
the
to
a
is
co
in
and
of
i
for
0
500
1000
1500
2000
2500
2013
the
to
a
realdonaldtrump
you
is
and
of
i
in
0
1000
2000
3000
4000
2014
the
realdonaldtrump
to
you
a
for
is
in
of
i
0
1000
2000
3000
2015
the
realdonaldtrump
to
you
a
is
i
trump
for
and
0
1000
2000
3000
4000
2016
the
to
and
in
a
i
you
of
is
will
0
500
1000
1500
2000
2500
2017
the
to
and
of
in
a
is
for
will
great
0
500
1000
1500
2000
2500
2018
the
and
to
of
a
in
for
is
on
that
0
50
100
150
order
n

Pipeline: Stop Words

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


2009
10
pqpfvm
champion
contest
book
trump's
watch
trumpthinklikeachampion.com
signed
read
keychain
happy
enter
david
0
5
10
2010
apprentice
nbc
p.m
pm
ly
9
tonight
tune
miss
10
0
10
20
30
40
2011
barackobama
ly
cont
timetogettough
china
interview
america
trumpvlog
watch
book
0
50
100
2012
barackobama
obama
cont
mittromney
china
interview
time
people
barackobama's
discussing
0
100
200
300
400
2013
people
true
obama
president
time
love
country
celebapprentice
run
obamacare
0
100
200
300
2014
president
run
obama
country
2016
love
true
vote
people
golf
0
200
400
600
2015
president
america
trump2016
run
people
country
makeamericagreatagain
time
love
vote
0
200
400
600
2016
hillary
trump2016
america
people
clinton
makeamericagreatagain
crooked
cruz
vote
join
0
100
200
300
400
2017
people
news
fake
america
tax
president
u.s
country
jobs
american
0
50
100
150
200
2018
people
news
president
country
daca
jobs
united
time
fake
wall
security
record
dollars
border
american
america
people
0
5
10
15

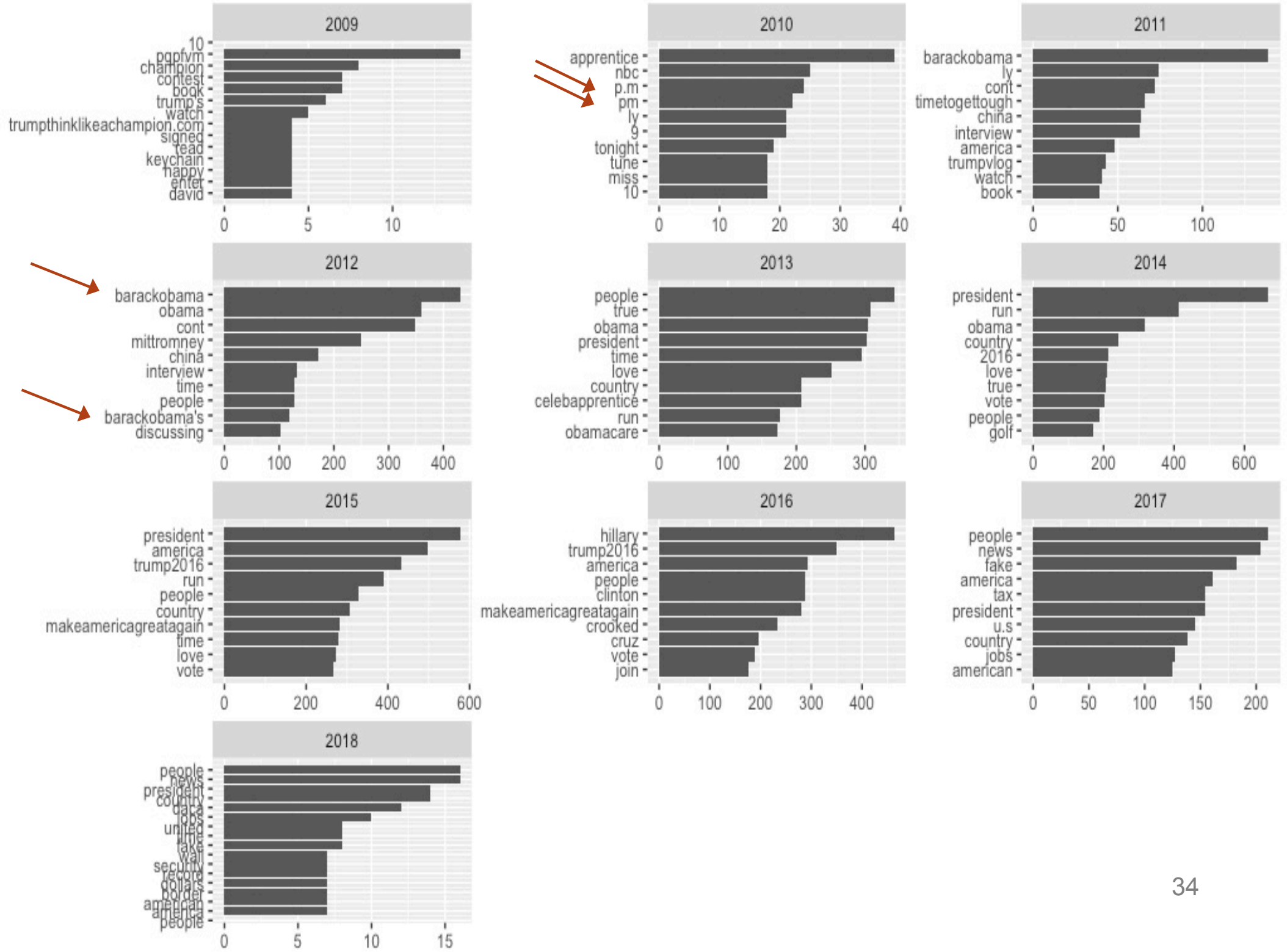
Pipeline: Stop Words

- Pitfalls:
  - Document/Topic-specific
    - i.e. realdonaldtrump, donald, trump, t.co (twitter link)

Pipeline: ?

Pipeline: ?

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


2009
10
pqpfvm
champion
contest
book
trump's
watch
trumpthinklikeachampion.com
signed
read
keychain
happy
enter
david
0
5
10
2010
apprentice
nbc
p.m
pm
ly
9
tonight
tune
miss
10
0
10
20
30
40
2011
barackobama
ly
cont
timetogettough
china
interview
america
trumpvlog
watch
book
0
50
100
2012
barackobama
obama
cont
mittromney
china
interview
time
people
barackobama's
discussing
0
100
200
300
400
2013
people
true
obama
president
time
love
country
celebapprentice
run
obamacare
0
100
200
300
2014
president
run
obama
country
2016
love
true
vote
people
golf
0
200
400
600
2015
president
america
trump2016
run
people
country
makeamericagreatagain
time
love
vote
0
200
400
600
2016
hillary
trump2016
america
people
clinton
makeamericagreatagain
crooked
cruz
vote
join
0
100
200
300
400
2017
people
news
fake
america
tax
president
u.s
country
jobs
american
0
50
100
150
200
2018
people
news
president
country
daca
jobs
united
time
fake
war
security
record
dollars
border
american
america
people
0
5
10
15

Pipeline: ?

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

Pipeline: ?

- Punctuation Removal:
  - , . ! ; ? ( ) [ ] { }
  - ‘s
    - Danger: meaning of a word may change if remove ‘
      - i.e. it’s vs. its
- Stemming
  - 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

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

Sentiment Analysis

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

Sentiment Analysis

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

Sentiment Analysis

- Lexicon: Bing et al.
  - Positive (30%)
  - Negative (70%)
  - N = 6788

Sentiment Analysis

- Lexicon: NRC
  - Positive (41%)
  - Negative (59%)
  - N = 5636

Sentiment Analysis

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

Sentiment Analysis

Election Day Tweets: AFINN

status_id <chr>	word <chr>	score <int>
x795782371895349250	tonight	*NA*
x795782371895349250	polls	*NA*
x79583420343064576O	unbelievable	-1
x79583420343064576O	evening	*NA*
x79583420343064576O	hampshire	*NA*
x79583420343064576O	flying	*NA*
x79583420343064576O	grand	3
x79583420343064576O	rapids	*NA*
x79583420343064576O	michigan	*NA*
x79583420343064576O	watch	*NA*

Sentiment Analysis

Election Day Tweets: AFINN

status_id <chr>	word <chr>	score <int>
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

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	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.
11/8/16 3:43	**Unbelievable** evening in New Hampshire - THANK YOU! Flying to Grand Rapids, Michigan now. Watch NH rally here:
11/8/16 4:27	@detroitnews: .@IvankaTrump in Michigan: â€˜This is your movementâ€™ @realDonaldTrump
11/8/16 4:29	@DonaldJTrumpJr: Thanks New Hampshire!!! #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!
11/8/16 23:03	Still time to #VoteTrump! #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 #ElectionDay htâ€¦
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! :â€¦

Sentiment Analysis

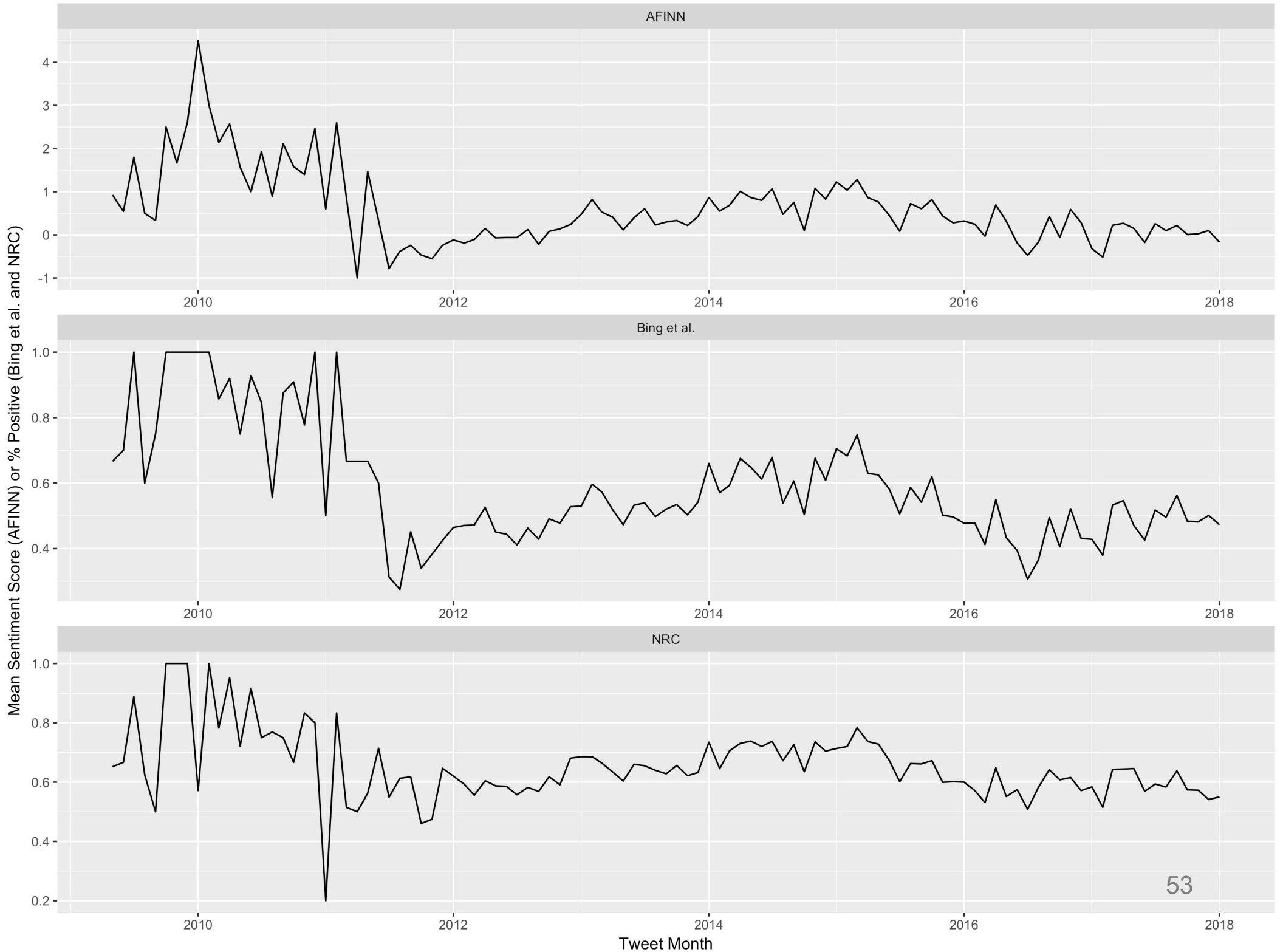
Election Day Tweets: Bing et al.

created_at <S3: POSIXct>	word <chr>	sentiment <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 <S3: POSIXct>	word <chr>	sentiment <chr>
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


AFINN
Bing et al.
NRC
Mean Sentiment Score (AFINN) or % Positive (Bing et al. and NRC)
Tweet Month
2010
2012
2014
2016
2018

Sentiment Analysis

Most Negative Month: NRC

created_at <S3: POSIXct>	word <chr>	sentiment <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, 4 p.m. on FOX News.

Sentiment Analysis

Most Negative Month by NRC using AFINN lexicon

created_at <S3: POSIXct>	word <chr>	score <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, 4 p.m. on FOX News.

Sentiment Analysis

Most Negative Month by NRC using Bing et al. lexicon

created_at <S3: POSIXct>	word <chr>	sentiment <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, 4 p.m. on FOX News.

Sentiment Analysis

Pitfalls: ?

Sentiment Analysis

Pitfalls

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

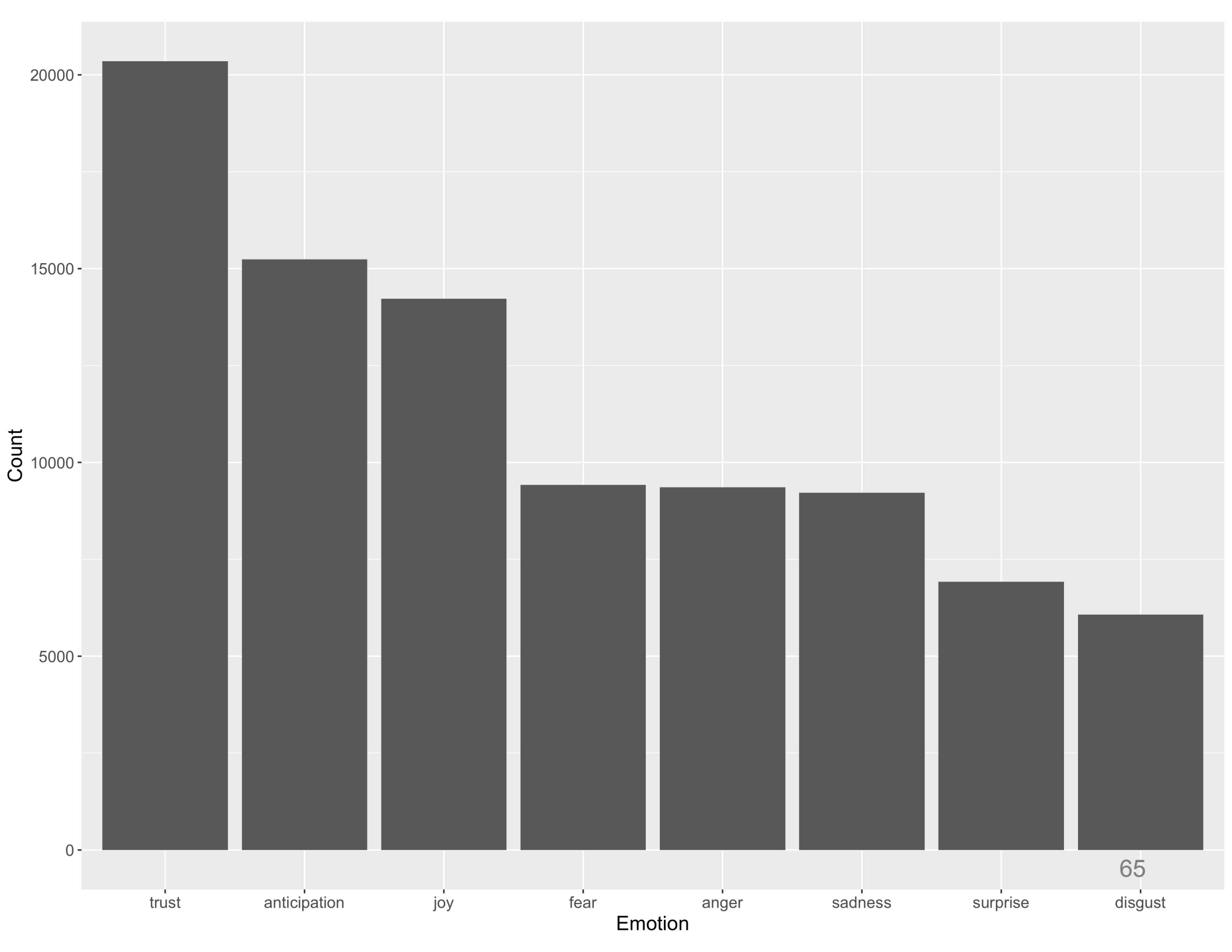
Methods

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

Emotion Analysis

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

Emotion Analysis


20000
15000
10000
5000
0
Count
65
trust
anticipation
joy
fear
anger
sadness
surprise
disgust
Emotion

Methods

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

TF-IDF

"term frequency-inverse document frequency"

- Setup: multiple documents
- Question of Interest 3: What is the underlying content of a document?

Example Dataset 2

- 56 Trump speeches from campaign
  - Dates: June 2015 to November 9, 2016
- 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 <dbl>	word <chr>	n <int>
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

- Data Level: (Document, Word)

- TF-IDF: How important is this word for this document?
- Term Frequency:
  - How common is this word in this document?
- Inverse Document Frequency:
  - How common is this word in all the other documents?

TF-IDF: Calculation

- Data Level: (*document*, *word*)

- Term Frequency:
  - # times *word* appears in *document* / total # of words in *document*
- Inverse Document Frequency:
  - ln(total # of documents / # documents containing *word*)
- TF-IDF = TF x IDF

Example

- Data Level: (Speech 8, “i”)

- “i” in Speech 8
  - # times “i” appears in Speech 8: 333
  - Total # words in Speech 8: 10,007
  - Total # of documents with the word “i”: 56 (all)
- Term Frequency: ?
- Inverse Document Frequency: ?
- TF-IDF: ?

Example

- "i" in Speech 8
  - # times "i" appears in Speech 8: 333
  - Total # words in Speech 8: 10,007
  - Total # of documents with the word "i": 56 (all)
- Term Frequency:
  - **# times word appears in document / total number of words in document**
- Inverse Document Frequency:
  - **ln(# of documents / # documents containing term)**

Example

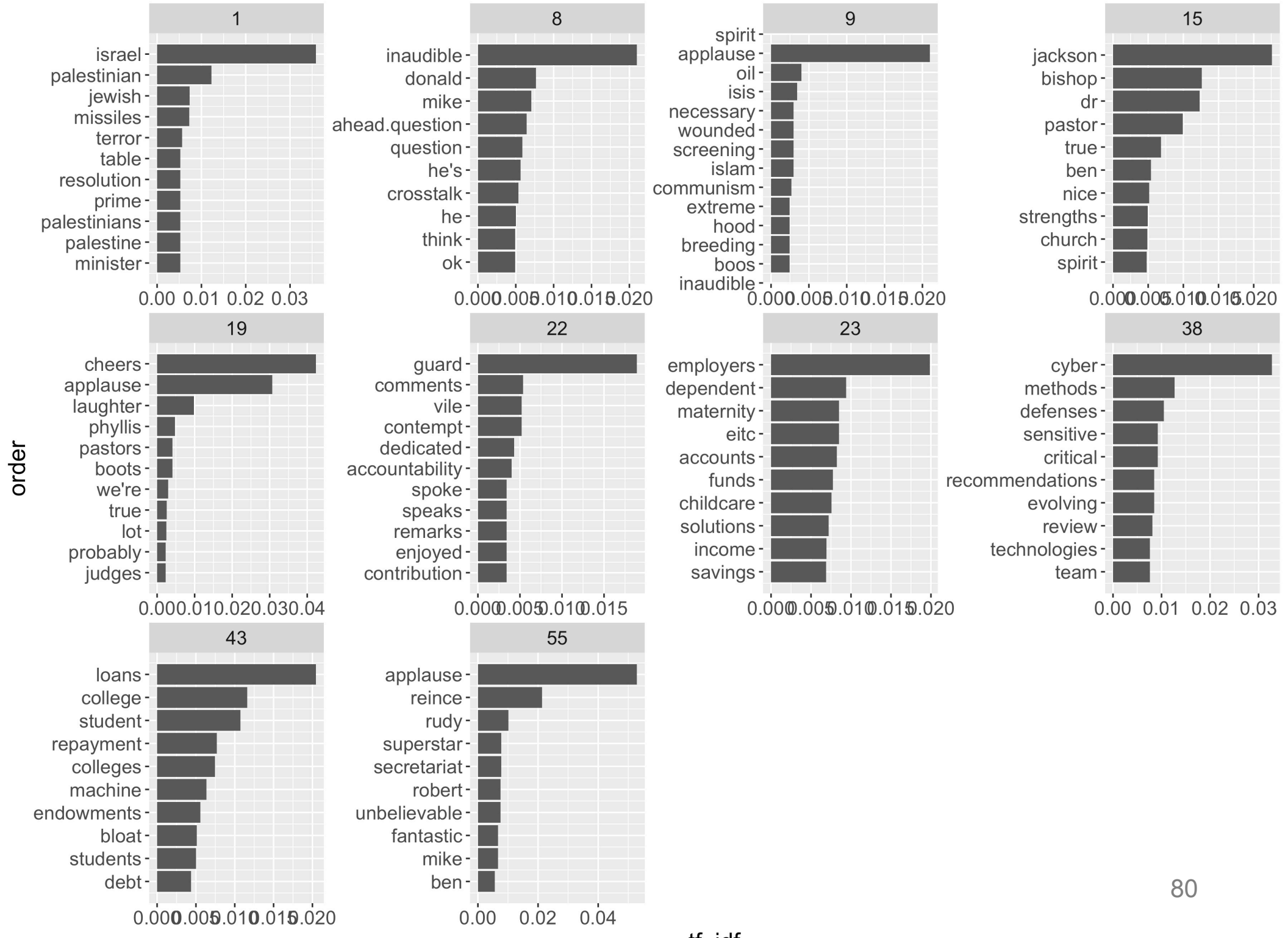
- Term Frequency:
  - **# times word appears in document / total number of words in document**
  - **333 / 10,007 = 0.0333**
- Inverse Document Frequency:
  - **ln(# of documents / # documents containing term)**
  - **ln(56 / 56) = ln(1) = 0**
- **tf-idf = 0.0333 x 0 = 0**

Example

- "israel" in Speech 1
  - # times "israel" appears in Speech 1: 37
  - Total # words in Speech 1: 2,305
  - Total # of documents with the word "israel": 6
- Term Frequency: **37 / 2305 = 0.0161**
- Inverse Document Frequency: ln(**56 / 6) = 2.234**
- tf-idf: **0.0161 x 2.234 = 0.0359**

Results

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


1
israel
palestinian
jewish
missiles
terror
table
resolution
prime
palestinians
palestine
minister
0.00 0.01 0.02 0.03
8
inaudible
donald
mike
ahead.question
question
he's
crosstalk
he
think
ok
0.000 0.005 0.010 0.015 0.020
9
spirit
applause
oil
isis
necessary
wounded
screening
islam
communism
extreme
hood
breeding
boos
inaudible
0.000 0.005 0.010 0.015 0.020
15
jackson
bishop
dr
pastor
true
ben
nice
strengths
church
spirit
0.000 0.005 0.010 0.015 0.020
19
cheers
applause
laughter
phyllis
pastors
boots
we're
true
lot
probably
judges
0.000 0.010 0.020 0.030 0.04
22
guard
comments
vile
contempt
dedicated
accountability
spoke
speaks
remarks
enjoyed
contribution
0.000 0.005 0.010 0.015
23
employers
dependent
maternity
eitc
accounts
funds
childcare
solutions
income
savings
0.000 0.005 0.010 0.015 0.020
38
cyber
methods
defenses
sensitive
critical
recommendations
evolving
review
technologies
team
0.00 0.01 0.02 0.03
43
loans
college
student
repayment
colleges
machine
endowments
bloat
students
debt
0.000 0.005 0.010 0.015 0.020
55
applause
reince
rudy
superstar
secretariat
robert
unbelievable
fantastic
mike
ben
0.00 0.02 0.04
order
tf_idf

Pitfalls: ?

Pitfalls

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

Methods

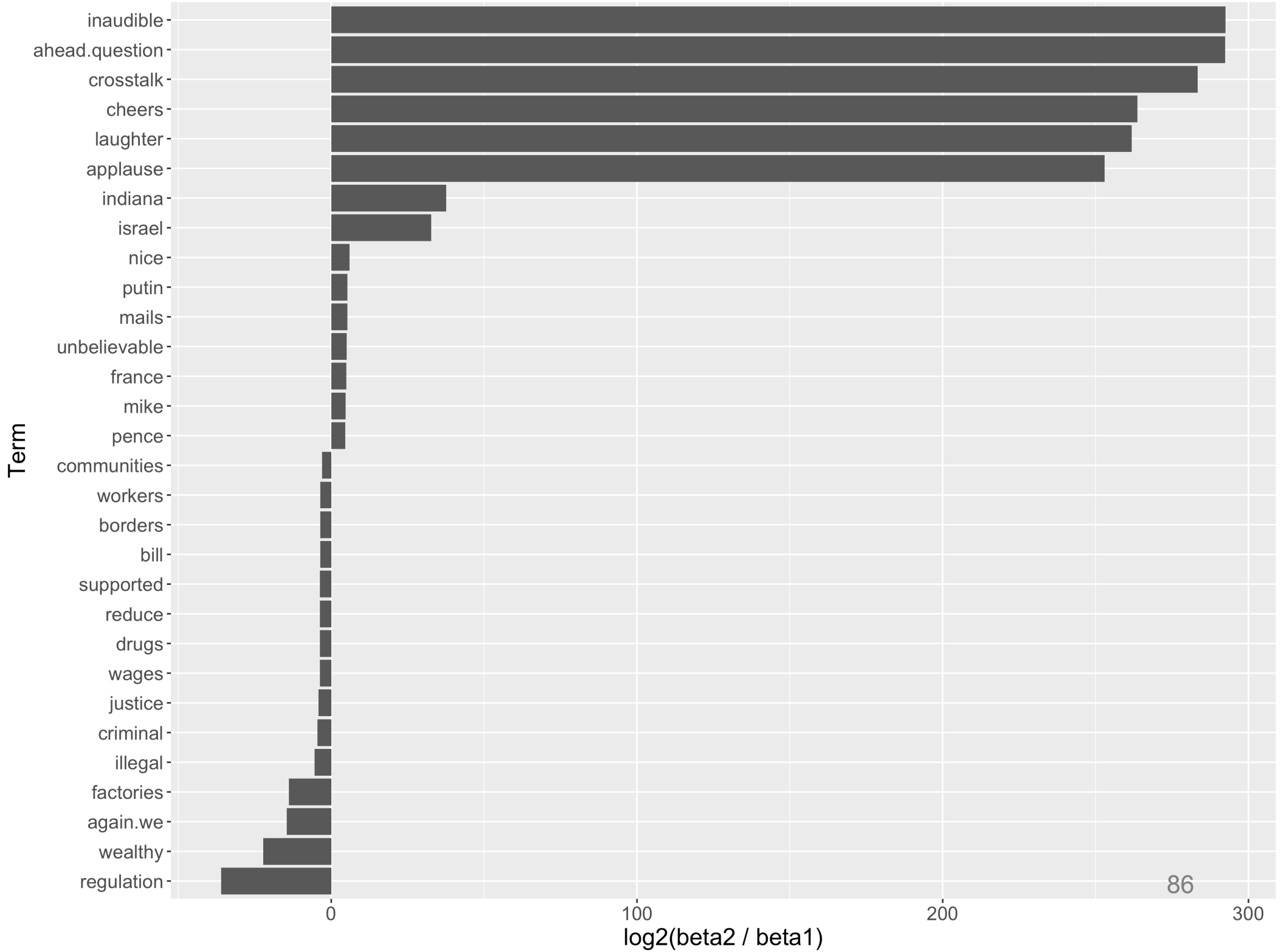
1. Sentiment Analysis
2. Emotion Analysis
3. TF-IDF
4. **Topic Modeling**

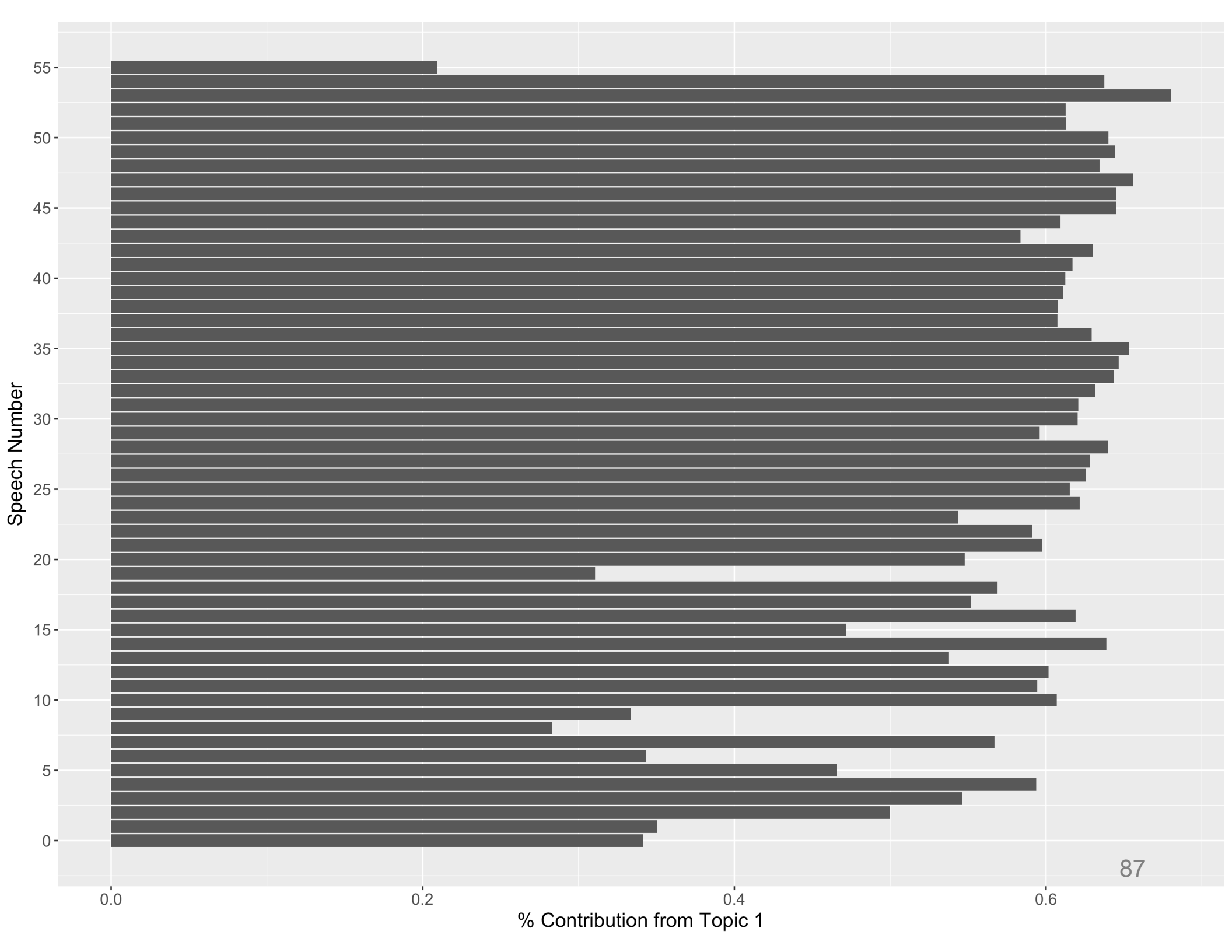
General Idea

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

Latent Dirichlet Allocation

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


inaudible
ahead.question
crosstalk
cheers
laughter
applause
indiana
israel
nice
putin
mails
unbelievable
france
mike
pence
communities
workers
borders
bill
supported
reduce
drugs
wages
justice
criminal
illegal
factories
again.we
wealthy
regulation
Term
0
100
200
300
log2(beta2 / beta1)
86


Speech Number
55
50
45
40
35
30
25
20
15
10
5
0
0.0
0.2
0.4
0.6
% Contribution from Topic 1
87

Bonus

Bonus

- Natural Language Processing: computer able to understand natural human language
  - Google Assistant
- Text Generation
  - 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 ?

```
Hi 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)

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

References (Images)

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