Reclaiming the humanity of data science:

Adventures in data with

Ron Paul, Facebook, emojis, and journalism

Hamdan Azhar

Cornell University

ORIE 4741 Guest Lecture

December 15, 2020

My goals

- Humanize data science by telling you about my journey as well as actual problems I've worked on

- Share some lessons I've learned along the way

- Give you practical tips for how to become a data scientist

The state of data science today

TECHNICAL KNOWLEDGE

Programming Languages:	Python, R, SQL, PySpark, Scala
Specialties:	Regression Analysis, Clustering Analysis, Segmentation Analysis, Decision Trees, Deep Learning, ETL Pipelines, Business Intelligence, AWS, Big Data, Spark
IDEs and Tools:	Tableau, Alteryx, R Studio, Power BI, MS PowerPoint, MS Excel, Git, Jupyter
Data Mining Libraries:	pandas, numpy, scikit-learn, nltk, scipy, matplotlib, seaborn, plotly, ggplot2
Certification:	IBM Data Science Professional, IBM Applied Data Science, Udemy Pyspark, AWS Certified Solutions Architect Associate (P), Tableau Desktop Specialist (P)

TECHNICAL SKILLS

Programming: R, Python, SQL
Data management: SQL Server, Oracle SQL, ETL - Oracle Data Integrator
Data Visualization: Tableau, OBIEE 11g and 12c, Qlik Sense, Cognos, Business Objects
Scheduling: Qlik Nprinting, Cognos Publisher, Oracle BI Publisher
Packages: NumPy, Pandas, matplotlib, SciPy, Scikit-Learn
Tools: Microsoft Office, JIRA, HP Quality Center, Confluence, R Studio, Jupyter, Spyder
Predictive Models: Regressions, Classification Models, Time-Series Analysis and Forecasting Techniques
Machine Learning: Univariate Analysis, Multivariate Analysis, Statistical Modelling, Hypothesis Testing
Expertise: Business Data Modelling, Data Extraction and Cleaning, Excel Reporting, Agile, Waterfall, SDLC

TECHNOLOGY AND SOFTWARE SKILLS

Languages: SQL, PLSQL, PostgreSQL, Python (Seaborn, Bokeh, Plotly, Matplotlib, NumPy, PySpark, PyTorch, TensorFlow, Scikit Learn, BeautifulSoup), JSON, Javascript, R (R- Shinny, Keras, Caret, e1071, RWeka)
BI Tools: Tableau, R, Weka, Minitab, Matlab, PowerBI, Datastage (ETL), Excel (StatTool, Power Queries, Pivot, vlookup), Snowflake, Databricks, Jupyter
Statistics: Time series forecasting Bayes theorem, Hypothesis testing, A/B testing, causation, collinearity, probability, regression
Big Data: NoSQL, MongoDB, OLAP, OLTP, Amazon Web Services (AWS) S3, Hadoop, HDFS, Hive, Spark, Map Reduce, Cassandra, AWS

The original promise of data science

Harvard Business Review

Data Scientist: The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data. by Thomas H. Davenport and D.J. Patil

From the Magazine (October 2012)

The original promise of data science

"I keep saying the sexy job in the next ten years will be statisticians. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s?

The ability to take data – to be able to **understand it**, to **process it**, to **extract value from it**, to **visualize it**, to **communicate it** is going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is the ability to understand that data and extract value from it.

I think statisticians are part of it, but it's just a part. You also want to be able to visualize the data, communicate the data, and utilize it effectively. But I do think those skills – of being able to access, understand, and communicate the insights you get from data analysis – are going to be extremely important. Managers need to be able to access and understand the data themselves."

- Hal Varian, *McKinsey Quarterly*, January 2009

Most of data science is deeply human

1. Find a problem to solve

2. Get the data necessary to answer this question

3. Clean the data

4. Analyze the data / <u>build models</u>

5. Tell a compelling story

6. Use the story to drive impact

Human Skills

Technical Skills

Human Skills

* Here, we focus on data science for understanding. There's also data science for prediction, which is gradually becoming more engineering than science and can afford to be more purely technical.
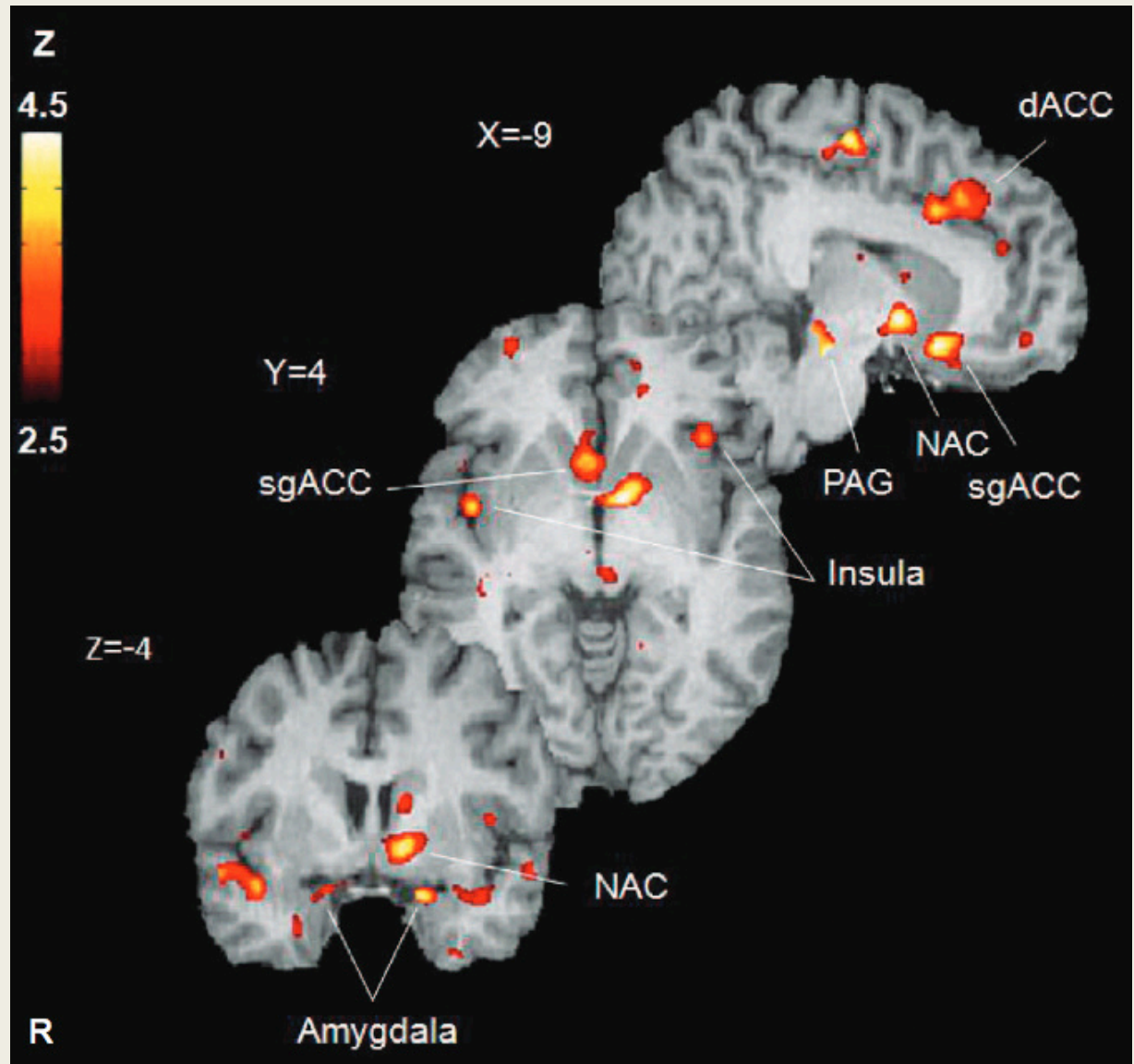
My journey
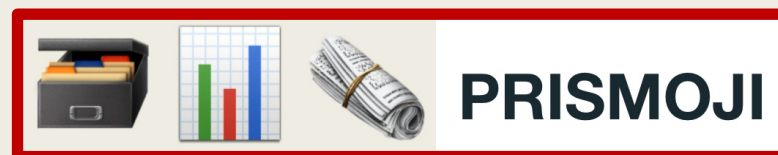
"Soon, data sets—large, unseemly, raw collections of numbers—began to tell me their stories. Whether these stories concerned demographics, human physiology, or the economy, they shed light on our understanding of the world. I saw firsthand how the language of statistics worked, and the immense power it held."

- Hamdan Azhar (2008)

Ron Paul 2012

Using data to inform voter contact in a presidential campaign
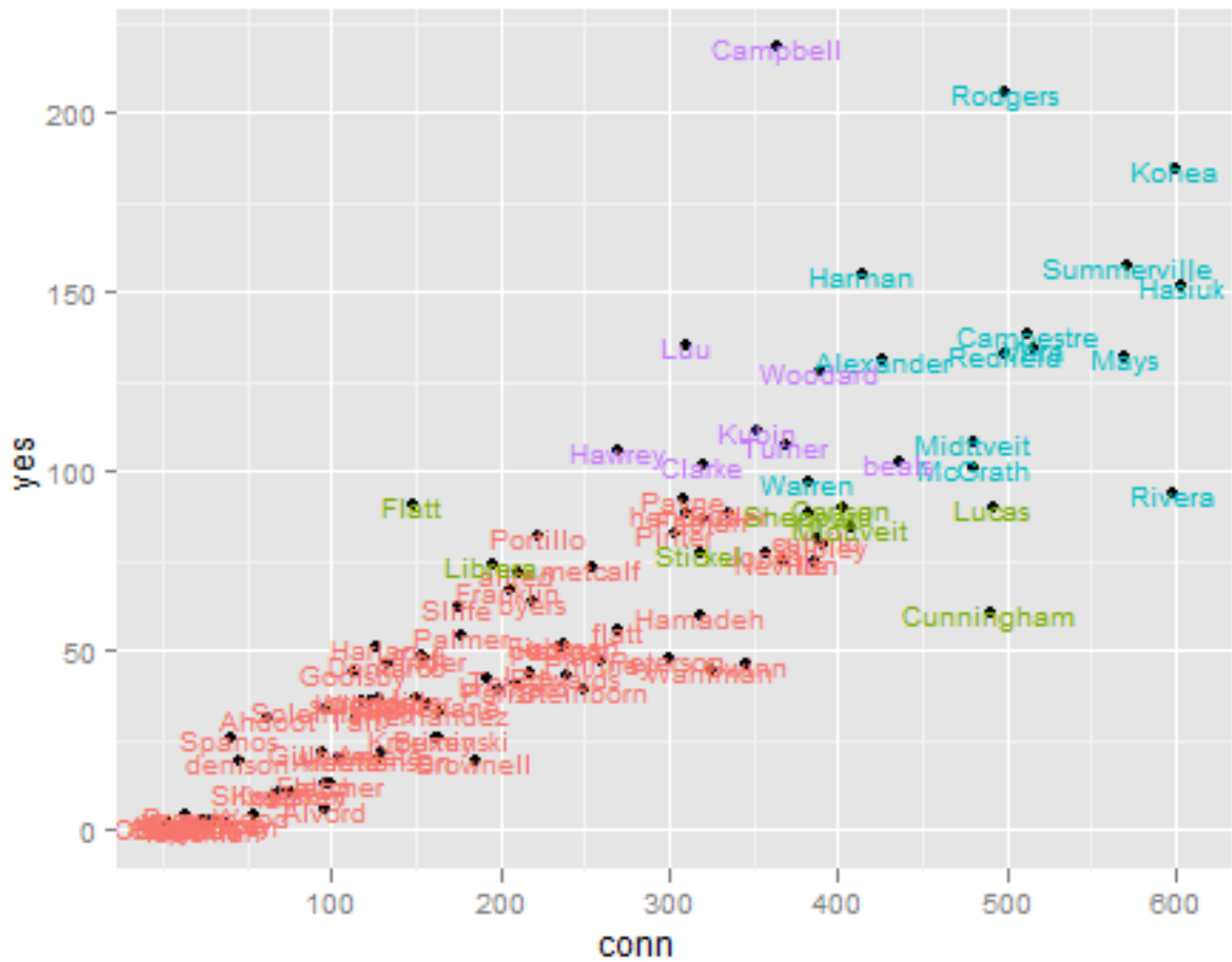
2012: RP Support rate, ID data
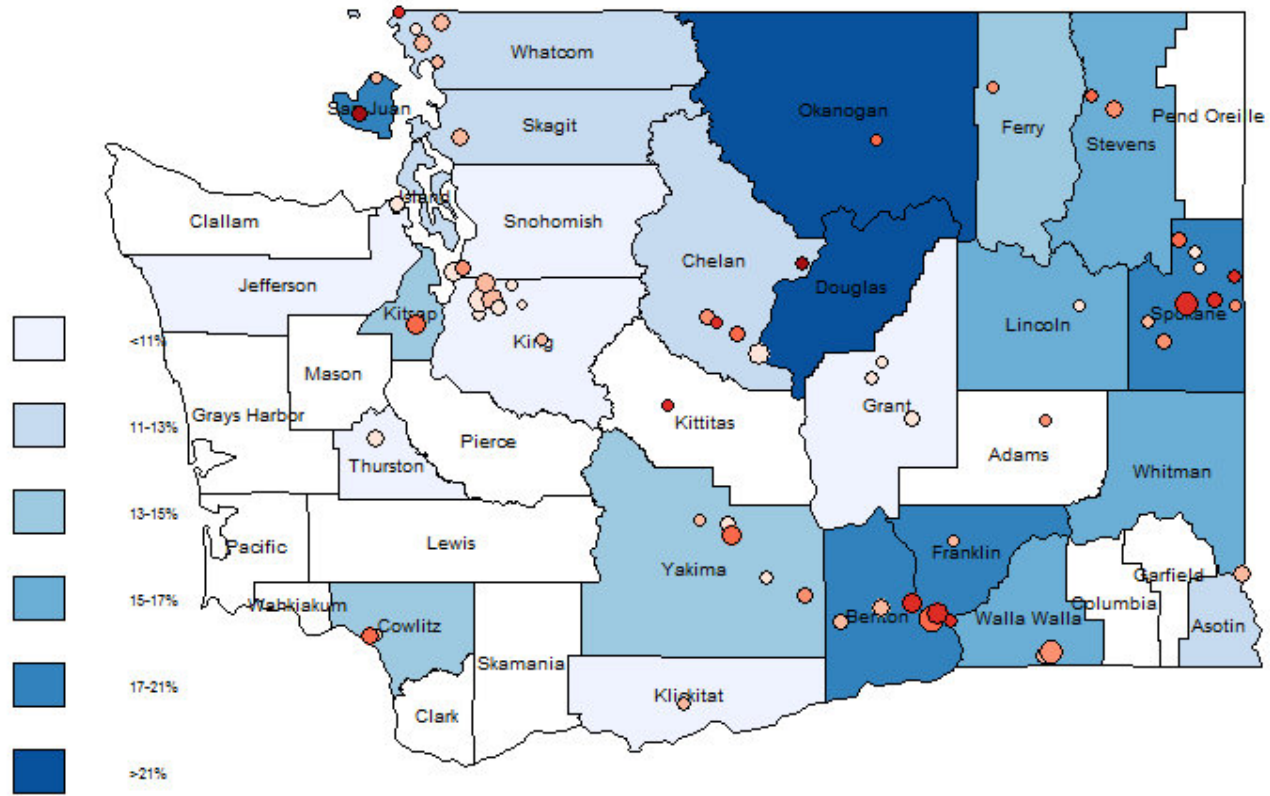
12,547 answered surveys, Feb. 28, 2012, F

ID	county	paul.ids	paul	newt	mitt	santorum	novote	numvoters	diffpaul	ra
Clark - 6490	Clark	24	0	0	0	0	0	0	24	
Clark - 2457	Clark	23	0	0	0	0	0	0	23	
Lander - 5	Lander	23	0	0	0	0	0	0	23	
Clark - 6645	Clark	18	0	0	0	0	0	0	18	
Lander - 4	Lander	18	0	0	0	0	0	0	18	
Clark - 2605	Clark	17	0	0	0	0	0	0	17	
Washoe - 6425	Washoe	17	0	0	0	0	0	0	17	
Lander - 2	Lander	12	0	0	0	0	0	0	12	
Washoe - 7541	Washoe	12	0	0	0	0	0	0	12	
Washoe - 7557	Washoe	12	0	0	0	0	0	0	12	
Clark - 3844	Clark	11	0	0	0	0	0	0	11	
Clark - 5577	Clark	10	0	0	0	0	0	0	10	
Clark - 2126	Clark	8	0	0	0	0	0	0	8	
Clark - 6703	Clark	8	0	0	0	0	0	0	8	
Clark - 7503	Clark	8	0	0	0	0	0	0	8	
Washoe - 2080	Washoe	8	0	0	0	0	0	0	8	
Washoe - 6215	Washoe	8	0	0	0	0	0	0	8	
Clark - 3380	Clark	7	0	0	0	0	0	0	7	
Clark - 6705	Clark	7	0	0	0	0	0	0	7	
Clark - 4566	Clark	6	0	0	0	0	0	0	6	
Clark - 4652	Clark	6	0	0	0	0	0	0	6	
Clark - 6608	Clark	6	0	0	0	0	0	0	6	
Lander - 3	Lander	6	0	0	0	0	0	0	6	
Clark - 1643	Clark	5	0	0	0	0	0	0	5	

Role of data in a campaign

Voter Contact

   Phone Bank

   Door Knocking

   Direct Mail $$$$

   ID, Persuasion, GOTV,

   Fundraising, Volunteering

Advertising $$$$

   TV, Digital, Radio, Print

Events $$$$

   Candidate Appearances

   Surrogate Events

Takeaways

- Work on something no one else is working on and very few people care about --> all upside when you're able to turn lemons into lemonade + lots of autonomy

- Ron Paul taught me that data science in the real world was something I enjoyed and wanted to do

Facebook

Using data to prove Facebook ads work and to help advertisers run better Facebook ads
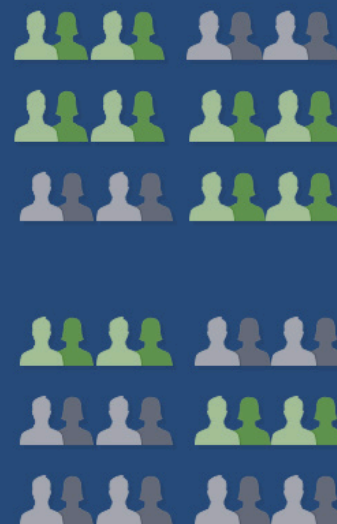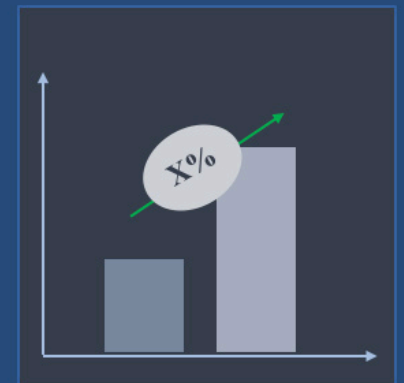
Three questions: the key ingredients for effective data storytelling

- What should I focus on?
- What is the meaning I'm going to give this?
- What am I going to do now?

* With credit to Tony Robbins

THE AD AGENCY BLOODLINE

A closer look at the agencies and internet milestones that have shaped our digital landscape of today.

The Three-Second Audition: How to Ensure Video Ads Make the Cut

PRISMOJI

The world's reaction to Brexit, in emoji
What people really mean, decoded.
qz.com

👍 Like   💬 Comment   ➤ Share

There were other reactions. They were sometimes articulate, but more often stilted as the world grappled with the fallout from Britain's self-inflicted referendum.

Here are a few:

François Hollande, president of France: 🥴

Angela Merkel, chancellor of Germany: 😞

Nicola Sturgeon, first minister of Scotland: 💪

Donald Trump, presumptive Republican presidential nominee: A Trump presidency welcomes you... 👋

Mark Carney, governor of the Bank of England: I've got a £250bn bazooka... 💰

Young people: 😠

Old people: 😏

Boris Johnson, ex-mayor of London, leader of "leave" campaign: 😱

David Cameron, British prime minister: 😟

George Osborne, British chancellor: 👀

Nigel Farage, leader of UKIP, Britain's far-right party: 😵

🐦   📘   💼   ✉️   📞

NEWS IN A WHOLE NEW WAY

Hey, we have an iPhone app now.

QUARTZ ▶️   Good morning •••

My reaction to this article, in emoji

😐

UK

Here Are the Most Popular Emojis From the #Brexit Reaction

HA

HAMDAN AZHAR
Jun 30 2016, 9:34am

emoji	emoji name	brexit rank	general rank	brexit index*	general index*	overindex**
😂	face with tears of joy	1	1	100	100	
🇬🇧	flag of united kingdom	2	363	87	0.2	400x
👍	thumbs up sign	3	18	26	11	2.3x
👏	clapping hands sign	4	45	24	6	3.9x
❤️	heavy black heart	5	3	21	45	
😭	loudly crying face	6	7	17	29	
😔	pensive face	7	13	14	18	
😩	weary face	8	11	13	22	
😢	crying face	9	27	12	9	1.3x
🙈	see-no-evil monkey	10	24	12	9	1.3x

Hashtag Signature: flag for united kingdom

Indexed Density of Hashtag for given emoji

notmyvote voteremain euref brexit voteleave
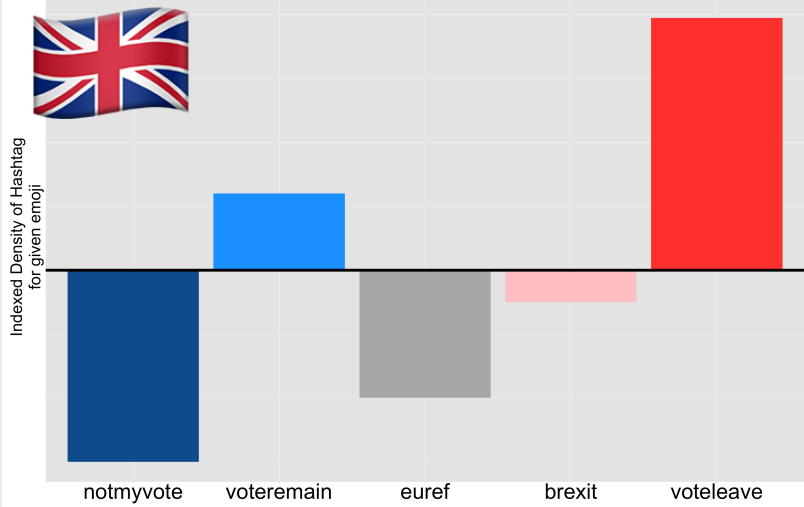
Hashtag Signature: party popper

Indexed Density of Hashtag for given emoji

notmyvote voteremain euref brexit voteleave
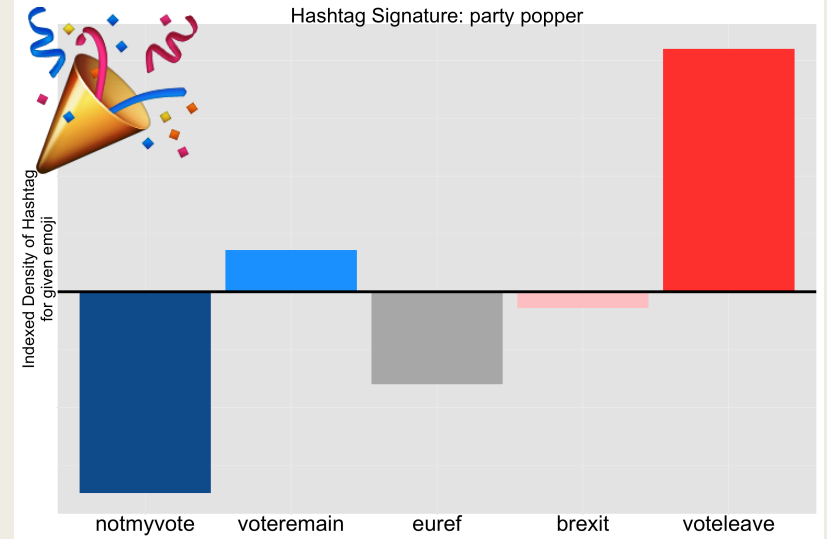
Hashtag Signature: loudly crying face

Indexed Density of Hashtag for given emoji

notmyvote voteremain euref brexit voteleave
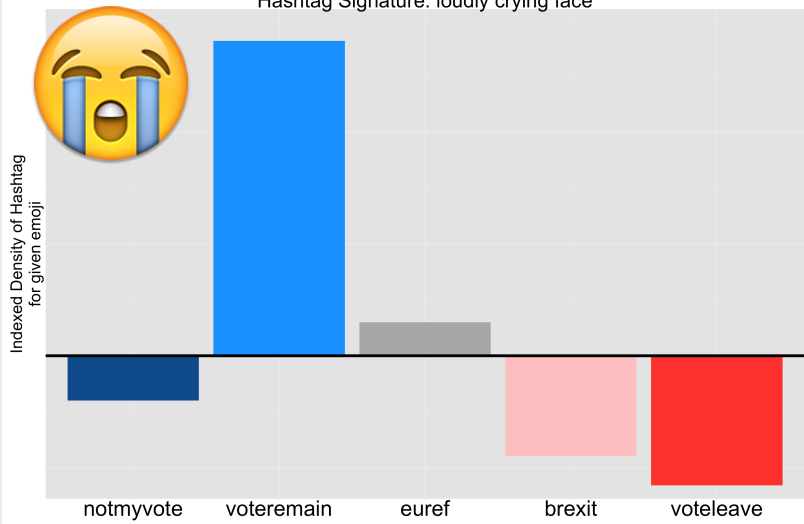
Hashtag Signature: person with folded hands

Indexed Density of Hashtag for given emoji

notmyvote voteremain euref brexit voteleave
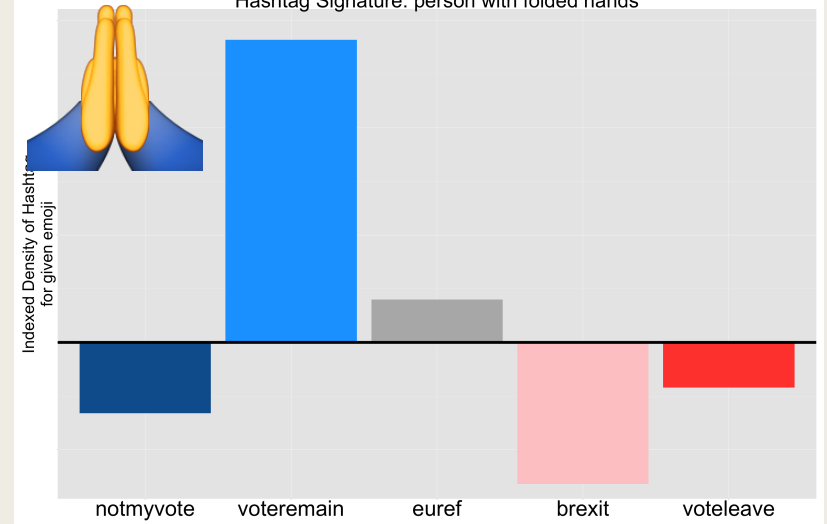
MOTHERBOARD

A Data Scientist's Emoji Guide to Kanye West and Taylor Swift

MADISON MARGOLIN
Aug 10 2016, 2:32pm

Overall Frequency (Per 1,000 Tweets)

Emoji Valence (@taylorswift13 <--> @kanyewest)

higher association with
@taylorswift13

higher association with
@kanyewest

+2000%   +1000%   equal   +1000%

30

20

10

0
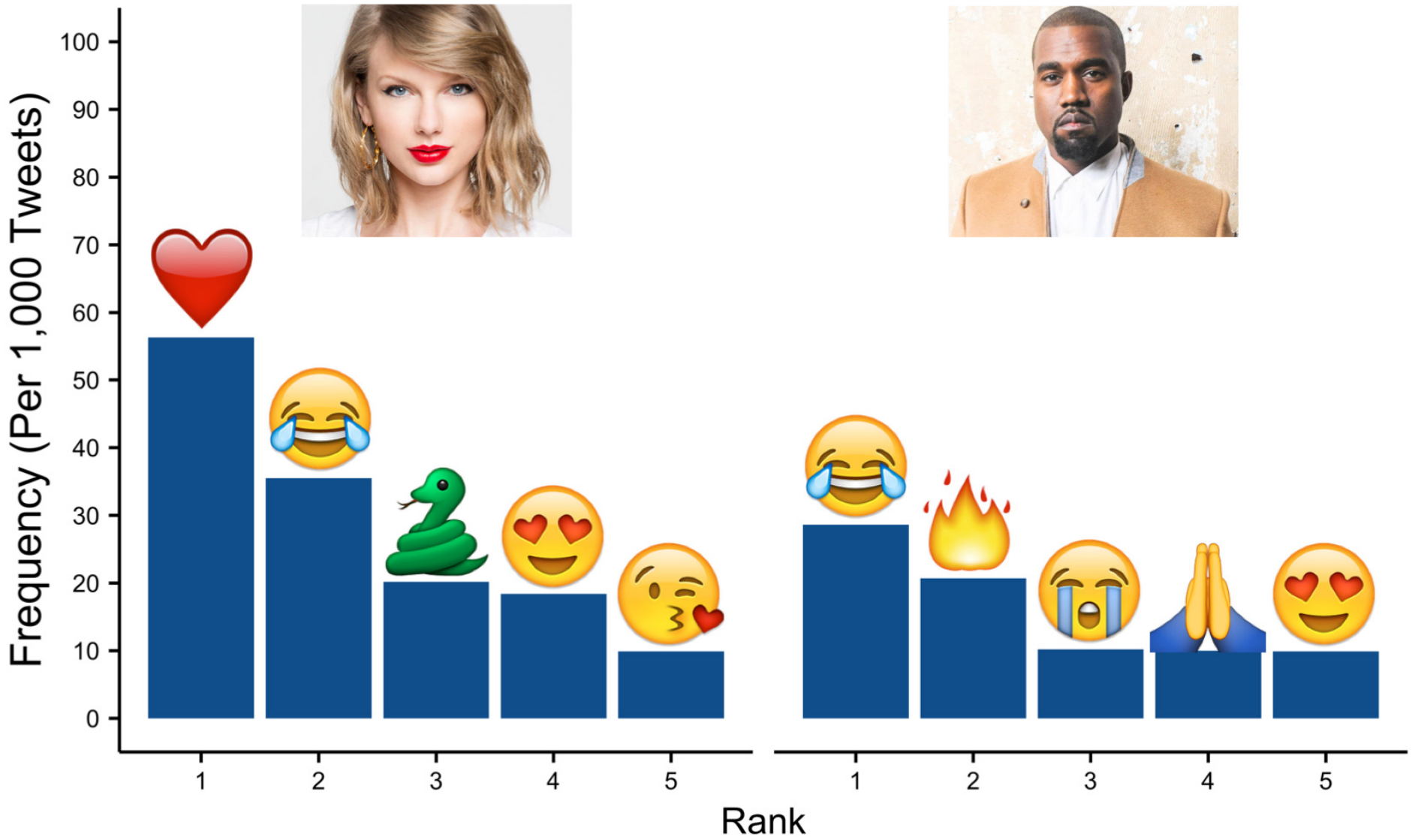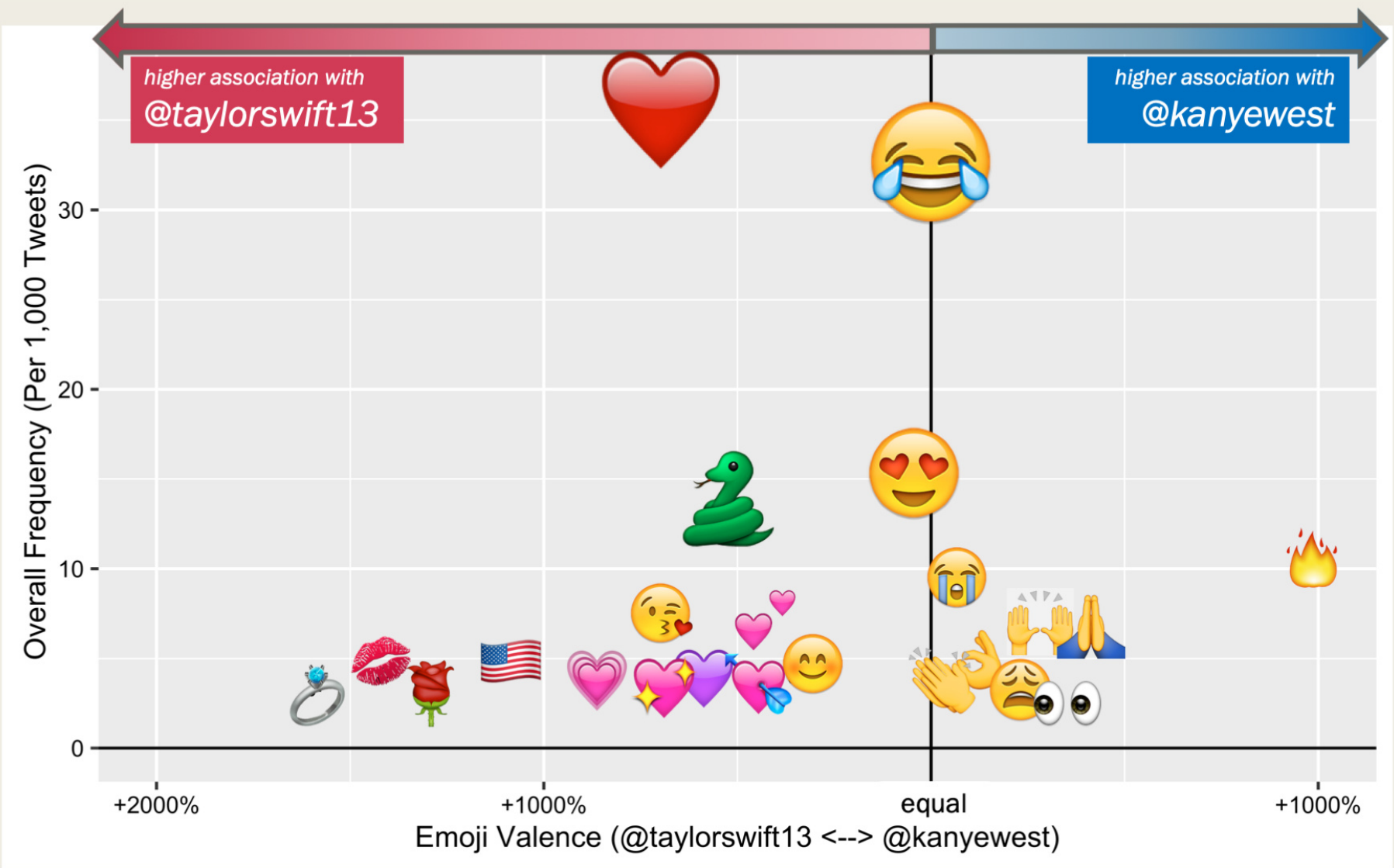
Our common emoji language of #fanlove

Most Popular Emojis	Beyoncé	Justin Bieber	DJ Khaled	Drake	Rihanna
1st	😍	😍	🔑	😂	😍
2nd	❤️	❤️	🔥	😢	❤️
3rd	😂	😭	💯	🔥	😂
4th	🐝	🔥	🙏	😍	😭
5th	😭	😂	🙌	❤️	😘
(% of tweets with emojis)	23%	29%	34%	18%	25%

More common in tweets mentioning Clinton

More common in tweets mentioning Trump

Overall Frequency (Per 1,000 Tweets)

Emoji Valence: Log Odds Ratio
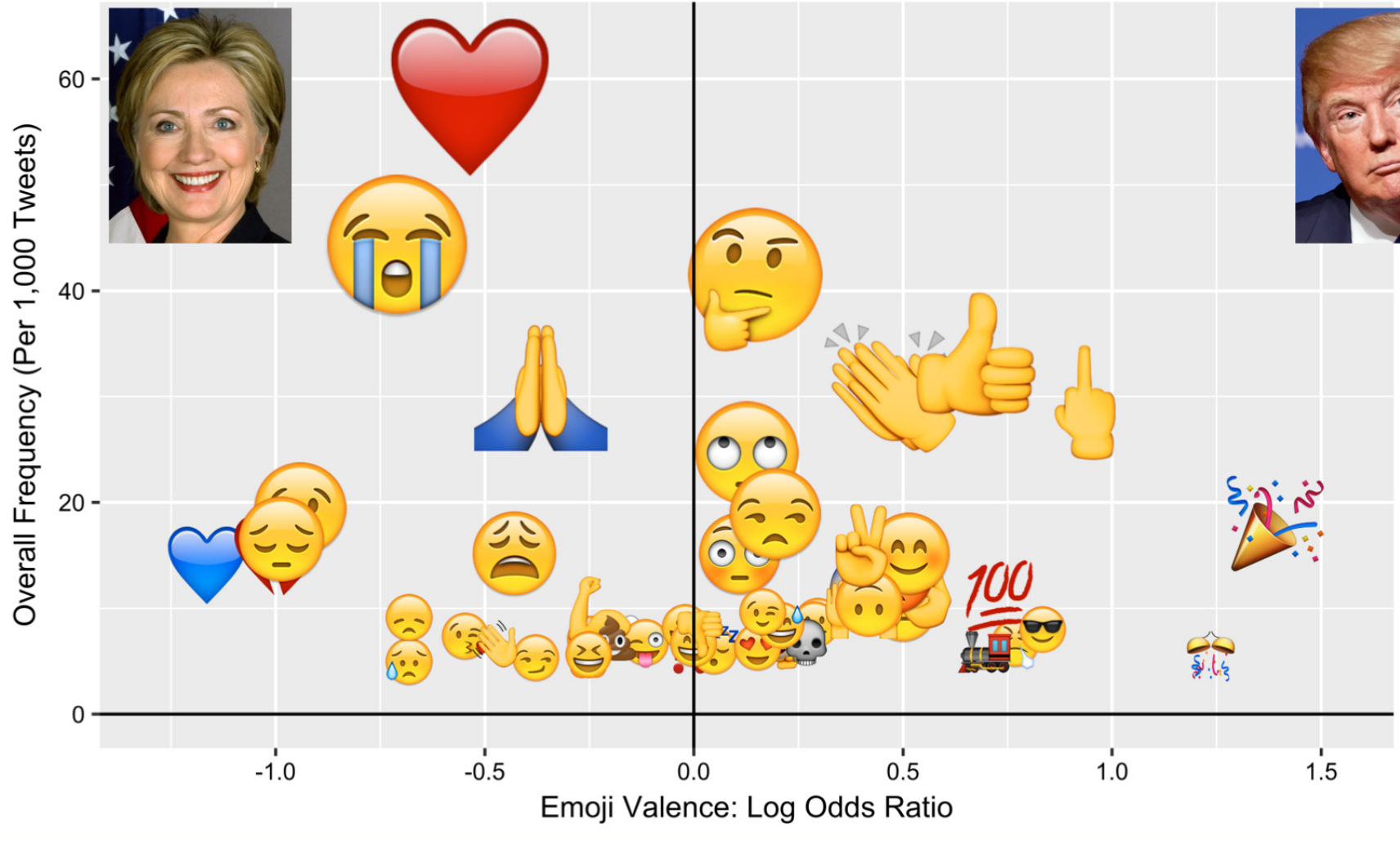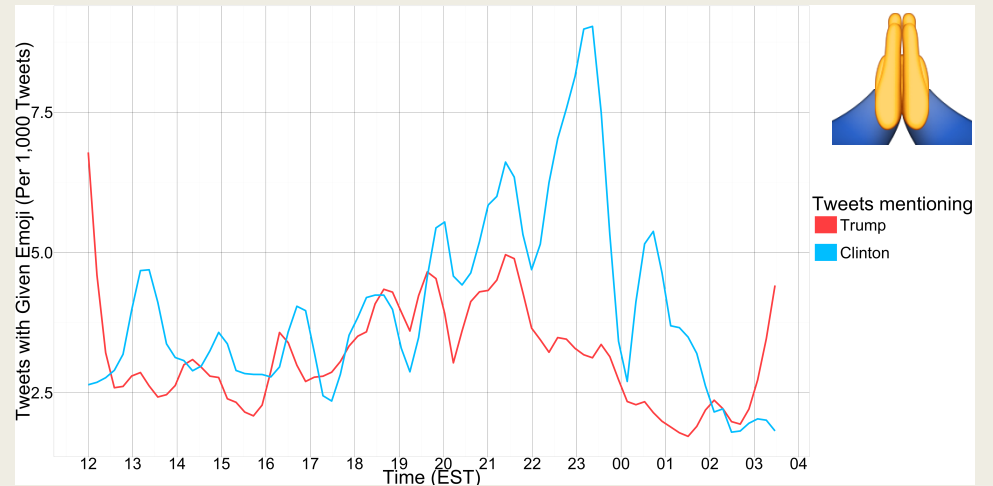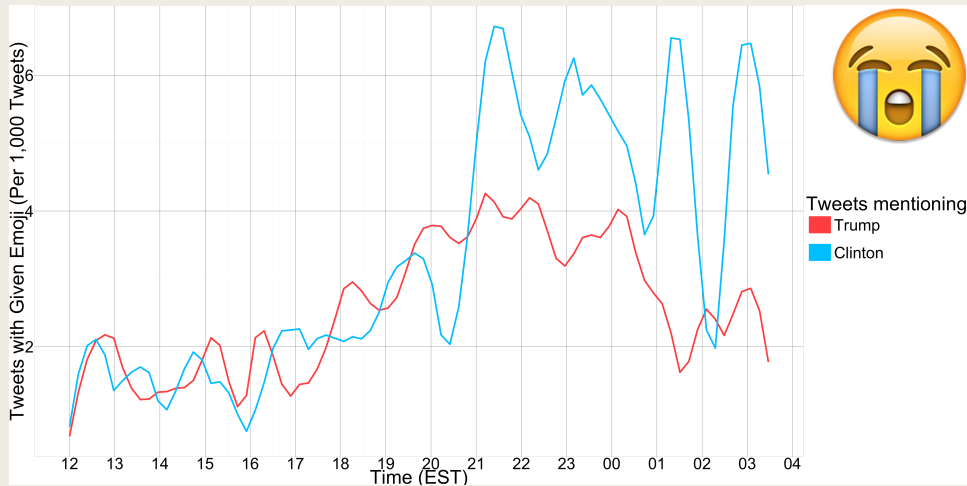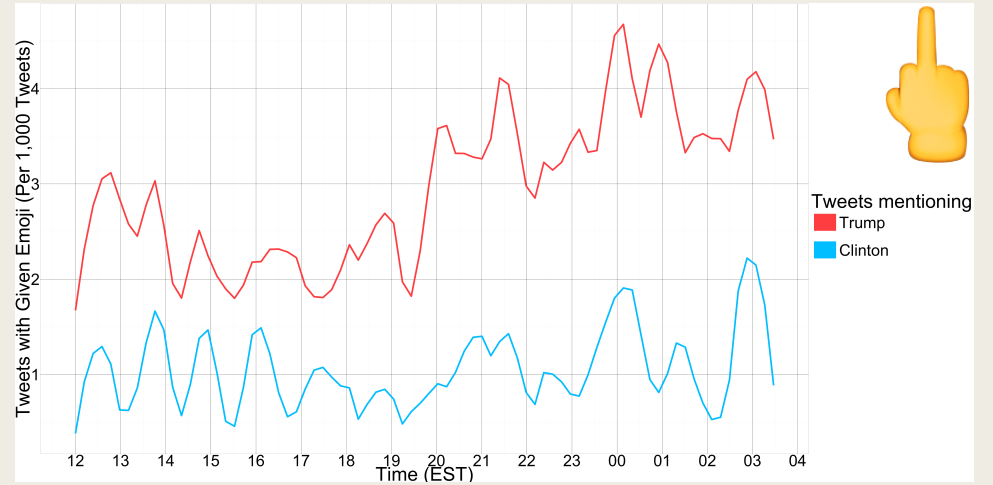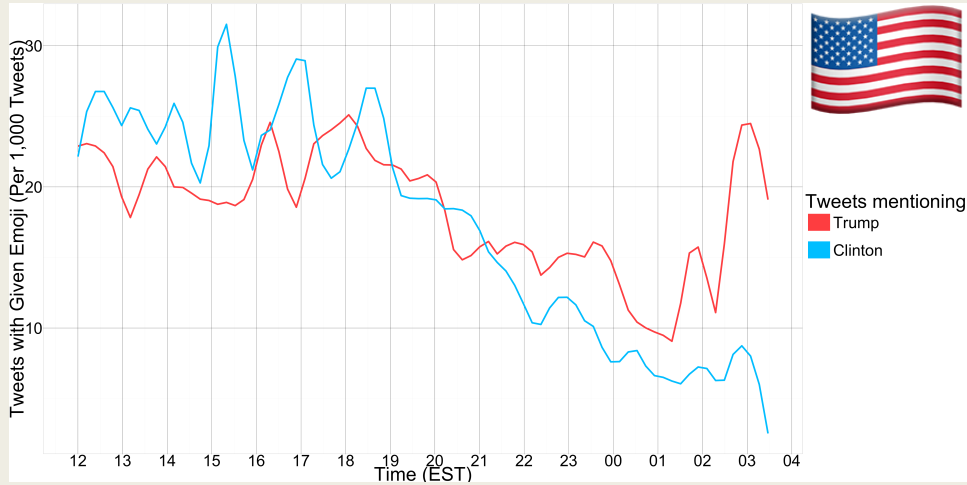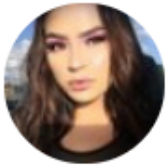
Val
@itsvaleriee___

Follow

How can I go to work in the morning knowing the people I work with voted for trump. 🤔 I just can't.

1:41 AM - 9 Nov 2016

While emojis might seem silly and trivial, people use them to convey real, authentic emotions.

Emojis are a universal language in an era of increasing depersonalization.

Emojis are also data.

And **data science** can help us understand them.

Data journalism

Even the most basic analyses can add tremendous value to the public discourse

David Robinson

Data Scientist at Stack Overflow, works in R and Python.

- ✉ Email
- 🐦 Twitter
- ⚙ Github
- 📑 Stack Overflow

Subscribe

Your email

Subscribe to this blog

Recommended Blogs

- R Bloggers
- RStudio Blog
- R4Stats
- Simply Statistics

Text analysis of Trump's tweets confirms he writes only the (angrier) Android half

I don't normally post about politics (I'm not particularly savvy about polling, which is where data science has had the largest impact on politics). But this weekend I saw a hypothesis about Donald Trump's twitter account that simply begged to be investigated with data:

Donald J. Tru
Good luck #
#OpeningCe
pic.twitter.c

27,391 Likes

Aug 5, 2016 at 8:59 PM

Donald J. Tr
Heading to
talking abo
SHORT CIR

4,451 Likes

Aug 6, 2016 at 11:11 AM

Todd Vaziri ✔
@tvaziri

🐦 Follow

Every non-hyperbolic tweet is from iPhone (his staff).

Every hyperbolic tweet is from Android (from him).

3:20 PM - 6 Aug 2016

↩   ⟲ 10,699   ♥ 15,013

Which are the words most likely to be from Android and most likely from iPhone?

HAMDAN AZHAR   BACKCHANNEL   06.26.2019 08:03 PM
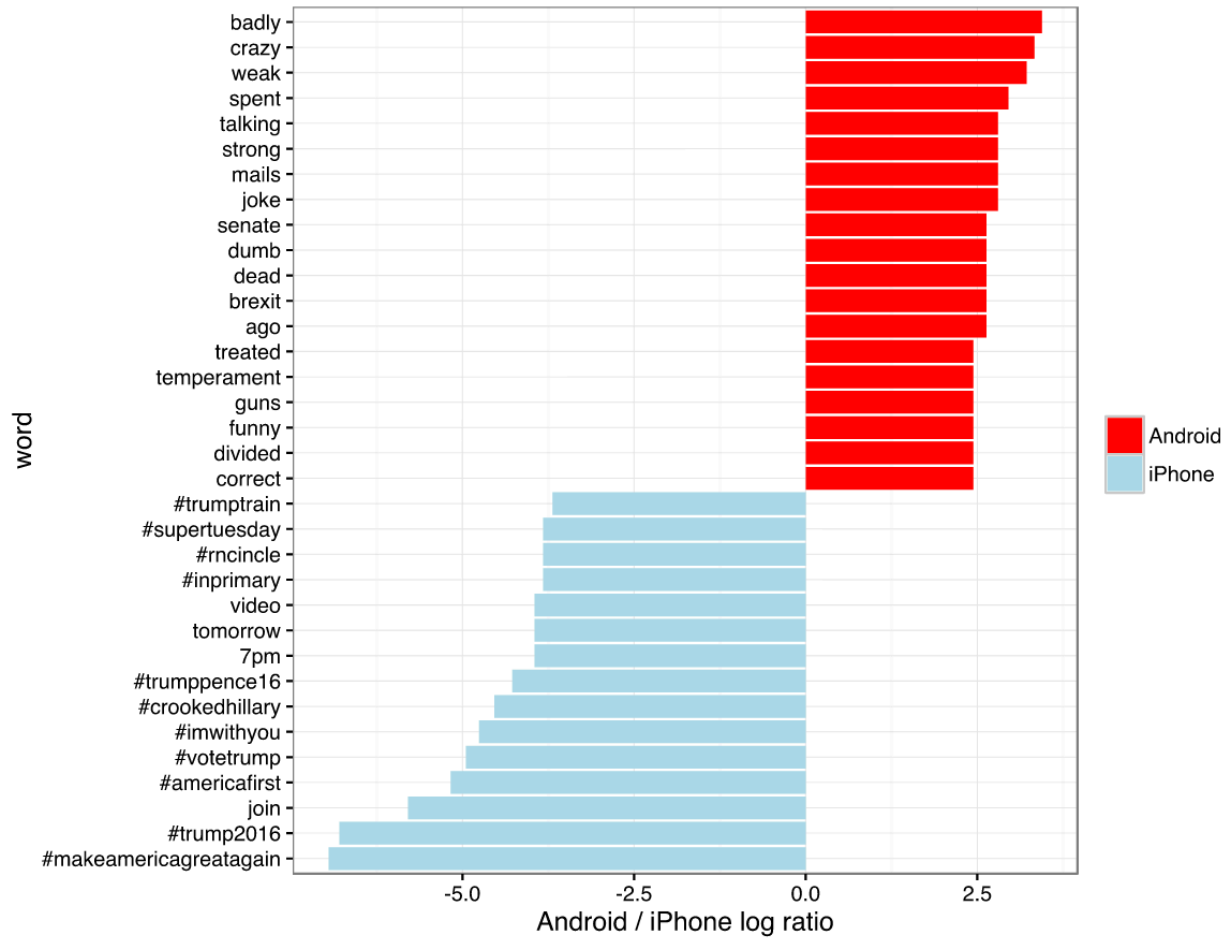
Politicians Don't Trust Facebook—Unless They're Campaigning

An analysis of more than 500 campaign websites reveals routine data sharing that belies politicians' tough talk on data privacy.

81% US senators have Facebook pixels on their campaign websites.

■ Republicans ■ Democrats ■ Independents ■ No Pixel

No Pixel
18

Republicans
47

Democrats
33
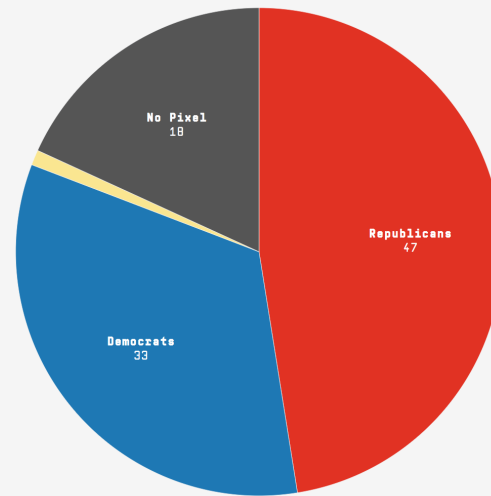
Some thoughts on data science

Most of data science is deeply human

1. Find a problem to solve

2. Get the data necessary to answer this question

3. Clean the data

4. Analyze the data / build models

5. Tell a compelling story

6. Use the story to drive impact

Human Skills

Technical Skills

Human Skills

* Here, we focus on data science for understanding. There's also data science for prediction, which is gradually becoming more engineering than science and can afford to be more purely technical.

Most of data science is deeply human

1. Find a problem to solve

2. Get the data necessary to answer this question

3. Clean the data

4. Analyze the data / build models

5. Tell a compelling story

6. Use the story to drive impact

Human Skills

80% of the time is spent on this

Why do we spend 80% of our time talking about this???

Technical Skills

80% of the impact comes from this

Human Skills

* Here, we focus on data science for understanding. There's also data science for prediction, which is gradually becoming more engineering than science and can afford to be more purely technical.

Data science as an end-to-end process

- **Data Strategist / Consultant:** Understands what the priorities of the business are and how data science can help advance those priorities

- **Data Project Manager:** Designs and drives a project plan and roadmap to advance the above strategy

Data Engineers

- **Data Sourcer:** Acquires the data necessary for the project

- **Data Cleaner:** Processes and cleans the data to make sure it is as accurate as possible and is ready to be analyzed

- **Data Analyst:** Conducts exploratory data analysis to understand the data and identify key trends and patterns

ML Engineers

- **Data Modeler:** Builds statistical models and algorithms to better understand the data or predict something from the data

- **Data Visualizer:** Designs intuitive and aesthetically appealing visualizations to make the data accessible

- **Data Insights Professional:** Identifies the key story, insights, and implications based on the analysis

- **Data Storyteller:** Communicates the story coming out of the analysis (in writing, or in powerpoint, or in a presentation)

- **Data Project Socializer:** Socializes the story to appropriate stakeholders and identifies next steps and recommended action items

The spectrum of data sophistication

- **Level 1:** Data, what's that??

- **Level 2:** We have a lot of data and we have no idea what to do with it.

- **Level 3:** We are doing some basic things with data but we need to do more interesting things and find more insights to get more buy-in from our stakeholders.

- **Level 4:** We have clean data and we know what we want and we need to build a model and implement it in production.

- **Level 5:** We have a model running and we need you to improve it by 1%.

Data science at BlackRock

Thank you!

hamdan.azhar@gmail.com