

Translating Questions



Data Analysis

What statistics should do about big data: problem forward not solution backward

 Jeff Leek  2013/05/29

There has been a lot of discussion among statisticians about big data and what statistics should do to get involved. Recently [Steve M. and Larry W.](#) took up the same issue on their blog. I have been thinking about this for a while, since I work in genomics, which almost always comes with “big data”. It is also one area of big data where statistics and statisticians have played a huge role.

A question that naturally arises is, “why have statisticians been so successful in genomics?” I think a major reason is the phrase I borrowed from [Brian C.](#) (who may have borrowed it from [Ron B.](#))

problem first, not solution backward

One of the reasons that “big data” is even a term is that there is that data are less expensive than they were a few years ago. One example is the dramatic drop in the price of [DNA-sequencing](#). But there are many many more examples. The quantified self movement and Fitbits, Google Books, social network data from Twitter, etc. are all areas where data that cost us a huge amount to collect 10 years ago can now be collected and stored very cheaply.

As statisticians we look for generalizable principles; I would say that you have to zoom pretty far out to generalize from social networks to genomics but here are two:





A year as told by fitbit

 Nick Strayer

 Dec 27, 2017  11 min read

 [visualization](#) [wearables](#) [time series](#)

I managed to wear a fitbit the entirety of 2017, this is exciting for a few reasons: one I have commitment problems, and two: it's a lot of data that I have to play with. While fitbit's app has some nice pretty graphs, they make it rather hard to actually dump all of your data into something nice like a csv.





Shijing Yao

Follow

Senior Machine Learning Scientist @ Airbnb

May 2 · 11 min read

Categorizing Listing Photos at Airbnb

Large-scale deep learning models are changing the way we think about images of homes on our platform.

Authors: *Shijing Yao, Qiang Zhu, Phillippe Siclait*



- When I run more do I lose weight?
- Are customers more likely to click on ads with puppies?
- Do I need to take an umbrella with me when I leave the house today?



- What or who am I trying to understand with data?
- What measurements do I have on those people or objects that help me answer the question?
- How do the data I have limit the type of question I can answer?
- What is the type of data science question we are trying to answer?



- What or who am I trying to understand with data?
- What measurements do I have on those people or objects that help me answer the question?
- How do the data I have limit the type of question I can answer?
- What is the type of data science question we are trying to answer?



- What or who am I trying to understand with data?
- What measurements do I have on those people or objects that help me answer the question?
- How do the data I have limit the type of question I can answer?
- What is the type of data science question we are trying to answer?

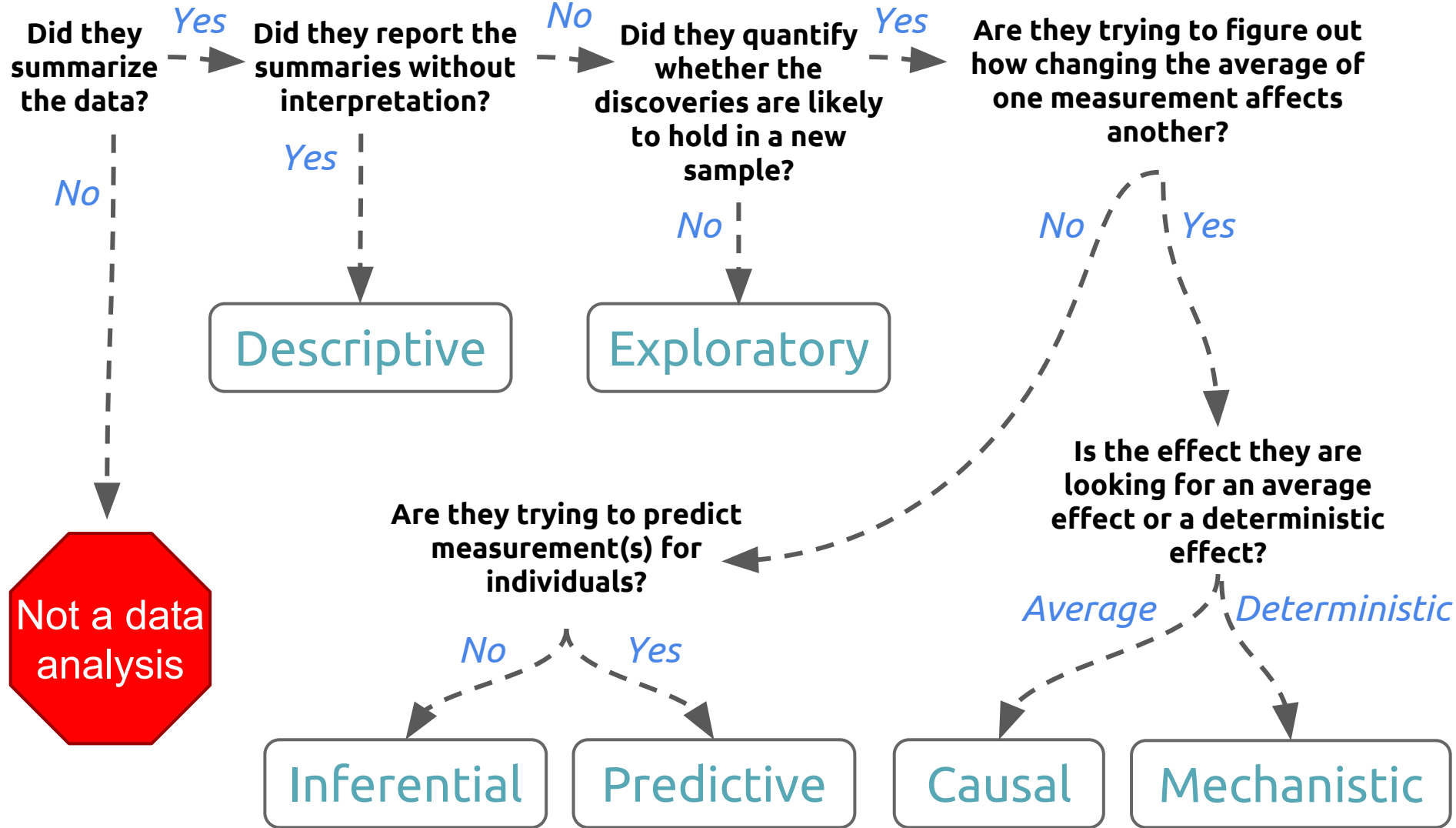


- What or who am I trying to understand with data?
- What measurements do I have on those people or objects that help me answer the question?
- How do the data I have limit the type of question I can answer?
- What is the type of data science question we are trying to answer?



- What or who am I trying to understand with data?
- What measurements do I have on those people or objects that help me answer the question?
- How do the data I have limit the type of question I can answer?
- What is the type of data science question we are trying to answer?





When I run more do I lose weight?

Who?



What?



Limitations?

- Only applies to Jeff
- Missing diet info
- Run data limited

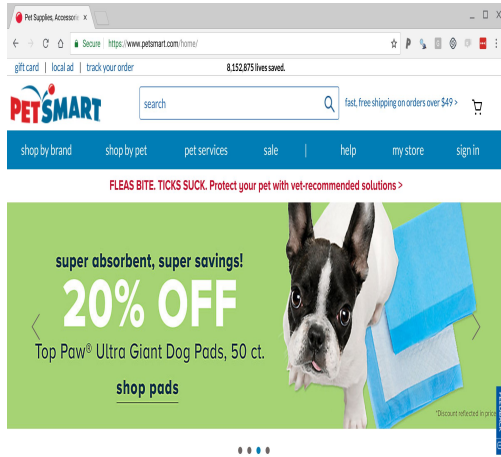
Question Type

Exploratory



Are customers more likely to click on ads with puppies?

Who?



What?



Limitations?

- Applies to customers of Pets.com
- Only one days worth of data

Question Type

Inferential

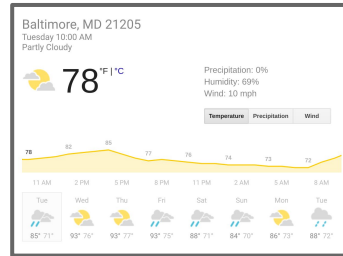


Do I need to take an umbrella with me when I leave the house today?

Who?



What?



Limitations?

- Only applies to one person
- Limited training data
- The future may be different from the past

Question Type

Predictive



David Robinson

Chief Data Scientist at
DataCamp, works in R and
Python.

- Email
- Twitter
- Github
- Stack Overflow

Subscribe

Subscribe to this blog

Recommended Blogs

- DataCamp
- R Bloggers
- RStudio Blog
- R4Stats

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. Trump	Donald J. Trump
Good luck # #OpeningCe pic.twitter.c	Heading to talking abo SHORT CIP
27,391 Likes	4,451 Likes
Aug 5, 2016 at 8:59 PM	Aug 6, 2016 at 11:11 AM

Todd Vaziri @tvaziri

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

Every hyperbolic tweet is from Android (from him).

3:20 PM - Aug 6, 2016

14.1K 10.2K people are talking about this



Tweets
38.5K

Following
47

Followers
53.8M

Likes
25

Moments
6

Donald J. Trump ✓

@realDonaldTrump

45th President of the United States of America 🇺🇸

📍 Washington, DC

🔗 [Instagram.com/realDonaldTrump](https://www.instagram.com/realDonaldTrump)

📅 Joined March 2009

[Tweet to Donald J. Trump](#)

👤 24 Followers you know



Tweets

Tweets & replies

Media



Donald J. Trump ✓ @realDonaldTrump · 2h

Lou Dobbs: "This cannot go forward...this Special Counsel with all of his conflicts, with his 17 Angry Democrats, without any evidence of collusion by the Trump Campaign and Russia. The Dems are the ones who should be investigated." Thank you Lou, so true!



8.3K



7.2K



26K



Donald J. Trump ✓ @realDonaldTrump · 2h

Strzok started the illegal Rigged Witch Hunt - why isn't this so-called "probe" ended immediately? Why aren't these angry and conflicted Democrats instead looking at Crooked Hillary?



9.1K



7.6K



28K



<https://twitter.com/realDonaldTrump>



 Search all documentation...

Docs

Basics

Accounts and users

Tweets

Direct Messages

Media

Trends

Geo

Ads

Metrics

Stay Informed

Staying informed about changes to our APIs is important for those developing on the platform and can be critical to maintaining your applications. We have a number of channels to help you stay in-the-loop.

[Learn how >](#)

Search Tweets

Use the Search API to find historical Tweets. Free to enterprise versions available.

Filter realtime Tweets

Get only the Tweets you need by using advanced filtering tools with the realtime streaming API.

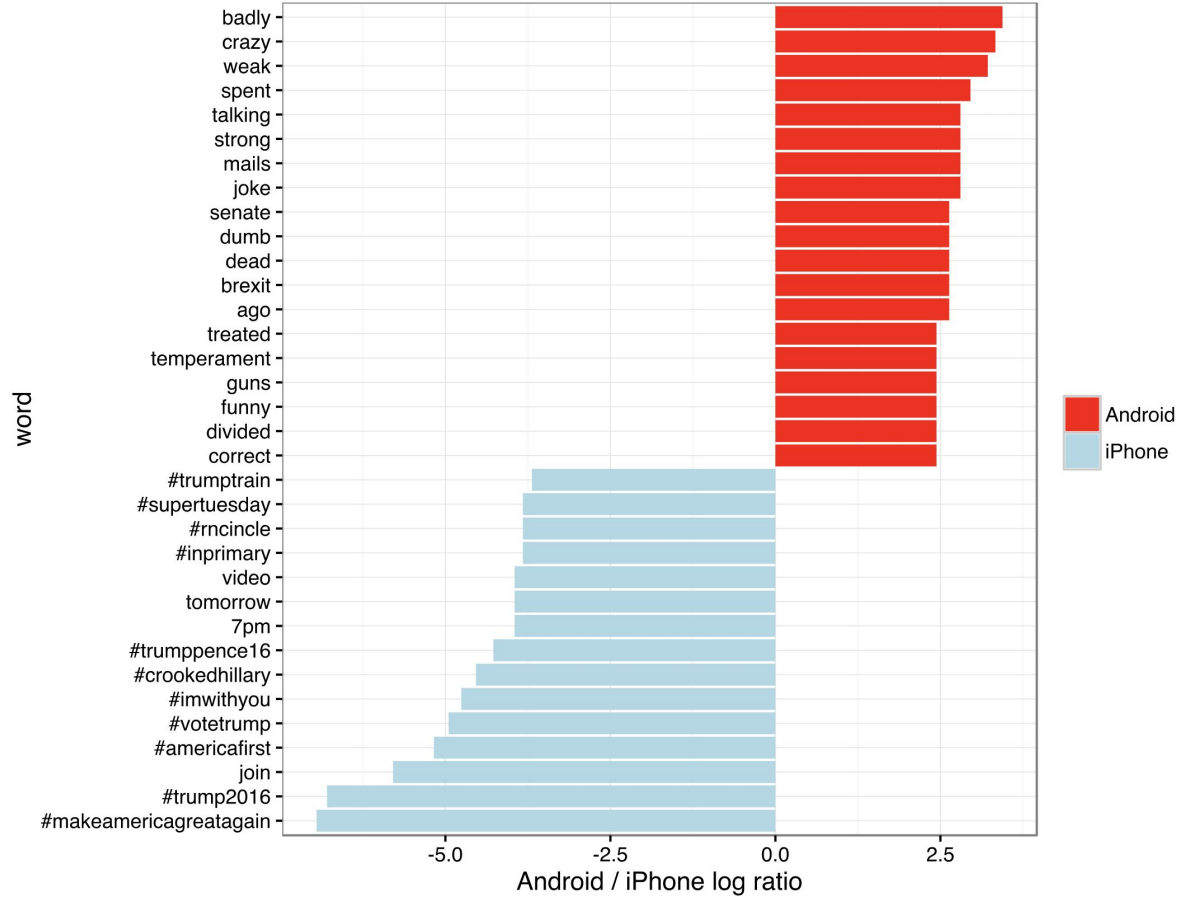
Are the Android and iPhone tweets clearly different?

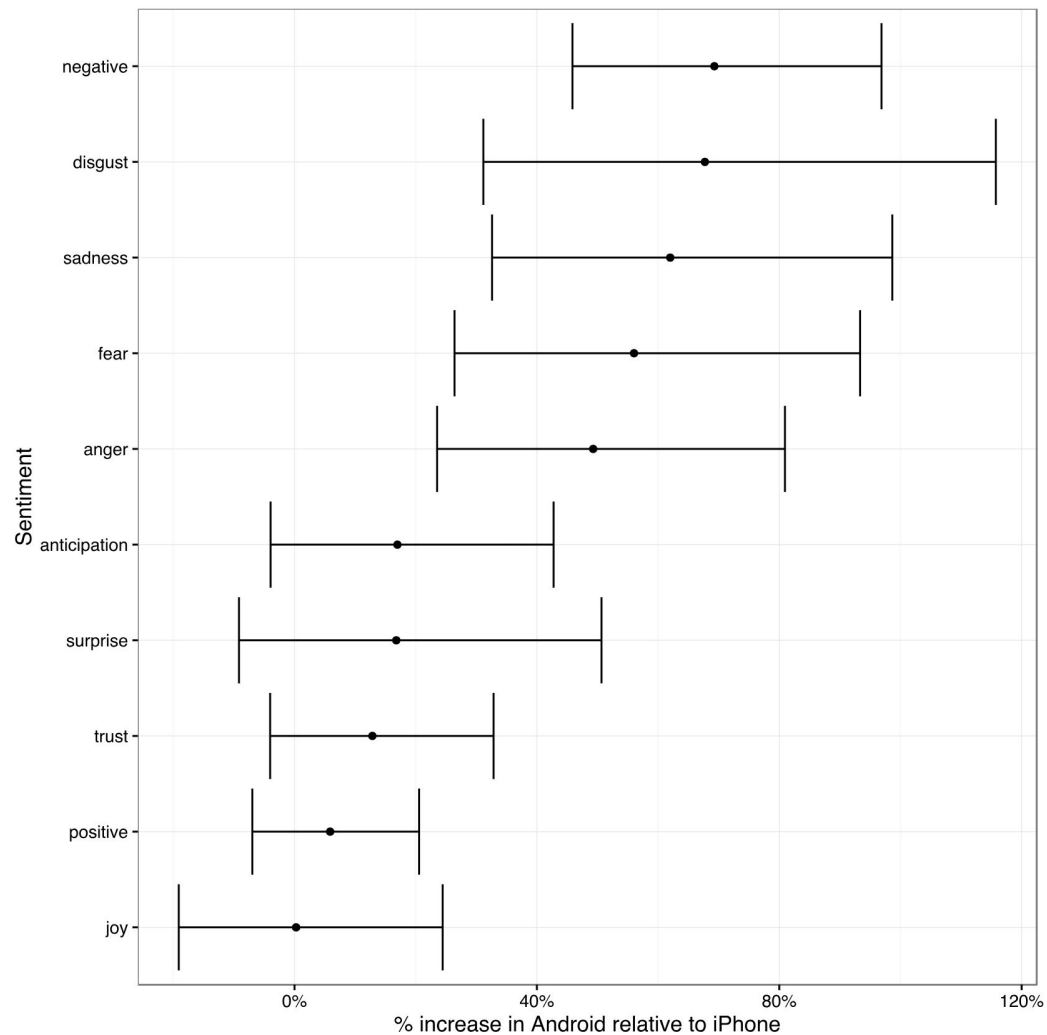


Are the Android tweets angrier and more negative?



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





Source: <http://varianceexplained.org/r/trump-tweets/>