Introduction to

Data Science

JProf. Dr. Claudia Wagner

6TH ESWC Summer School
September 2016

About Me

Since 2009:

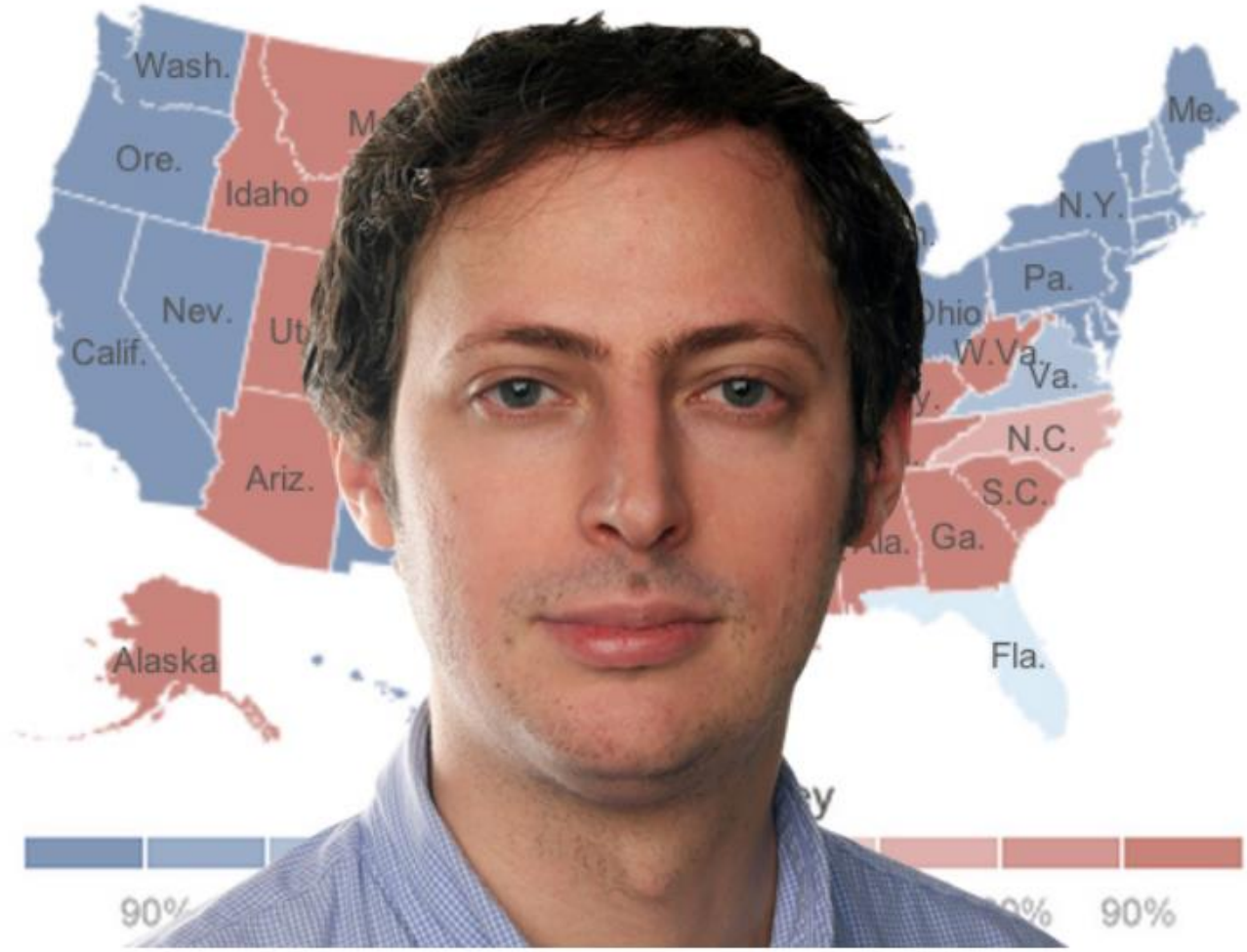
PhD Computer Science at TU Graz
Interning at KMI, HP, Xerox PARC

Since 2013:

Post Doc at GESIS – CSS department

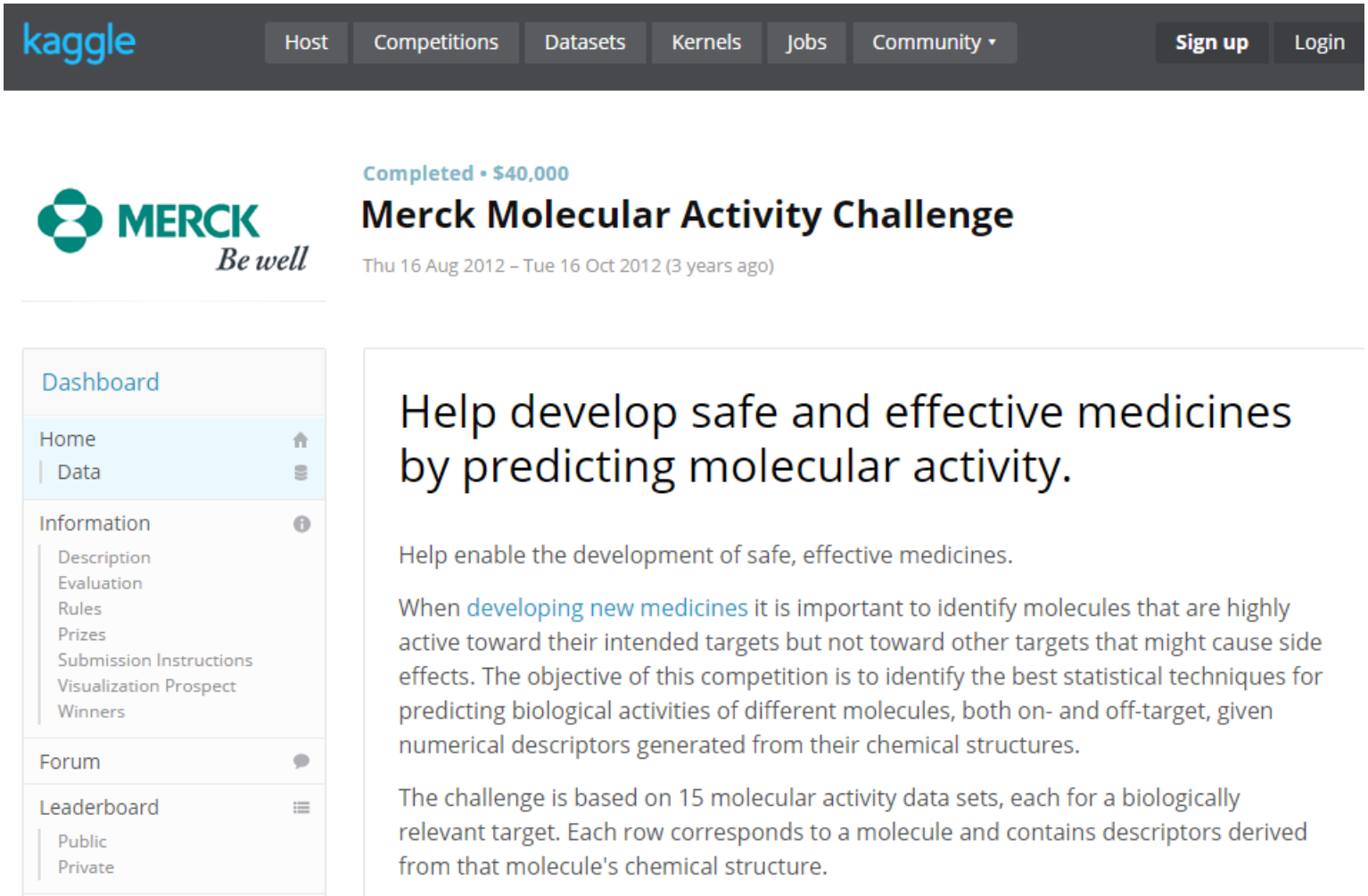
Since 2016:

Assistant Prof at University Koblenz Landau – WEST Institute
Head of Data Science Team at GESIS



Hanspeter Pfister, Data Science course, Introduction https://drive.google.com/drive/folders/0BxYkKyLxfsNVd0xicUVDS1dIS0k

Success Stories



Personalized Advertisement

https://flic.kr/p/ks7QzP

Google Flu Trends (GFT)

- Idea: CDC data about flu cases in US → find search terms that predict CDC data (~1000 observations and 50 million search terms)

- But GFT failed: underestimated of the 2009 H1N1 flu & overestimated the 2012-2013 flu season's cases by 140%

- Problems:
  - Spurious correlations
  - Search behavior changes over time
  - Transparency → what is measured?

Measurement Bias

- **Measurement** is the assignment of a number to a characteristic of an object.
  - E.g. flu cases per state, political leaning of a state

- **Problem:** often we cannot directly observe what we want to measure in organic data.

- How to ensure the quality of measurements?

Quality of Measurements

- Reliability: how consistent and stable is our measurement?
  - Intra-rater reliability
  - Inter-rater reliability
  - Test-retest reliability

- Validity: do we measure what we want to measure?
  - Face validity
  - Construct validity
  - Criterion-based validity

> Current Issue > vol. 112 no. 47 > Shihao Yang, 14473–14478, doi: 10.1073/pnas.1515373112

CrossMark click for updates

Accurate estimation of influenza epidemics using Google search data via ARGO

Shihao Yang[a], Mauricio Santillana[b,c,1], and S. C. Kou[a,1]

Author Affiliations

Edited by Wing Hung Wong, Stanford University, Stanford, CA, and approved September 30, 2015 (received for review August 6, 2015)

Significance

Big data generated from the Internet have great potential in tracking and predicting massive social activities. In this article, we focus on tracking influenza epidemics. We propose a model that utilizes publicly available Google search data to estimate current influenza-like illness activity level. Our model outperforms all available Google-search–based real-time tracking models for influenza epidemics at the national level of the United States, including Google Flu Trends. Our model is flexible, self-correcting, robust, and scalable, making it a potentially powerful tool that can be used for estimation and prediction at multiple temporal and spatial resolutions for other social events.

CHRIS ANDERSON MAGAZINE 06.23.08 12:00 PM

THE END OF THEORY: THE DATA DELUGE MAKES THE SCIENTIFIC METHOD OBSOLETE

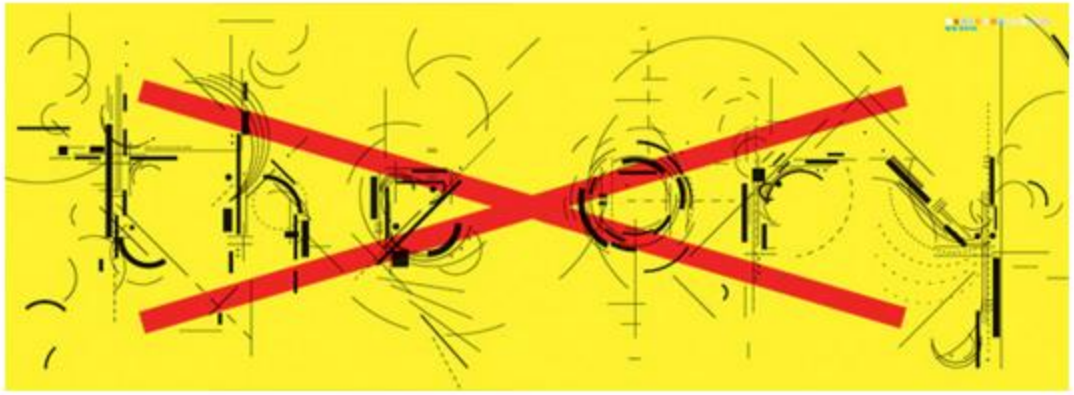


Illustration: Marian Bantjes

BASIC CONCEPTS

Who is a Data Scientist?

"A data scientist is someone who knows more statistics than a computer scientist and more computer science than a statistician."

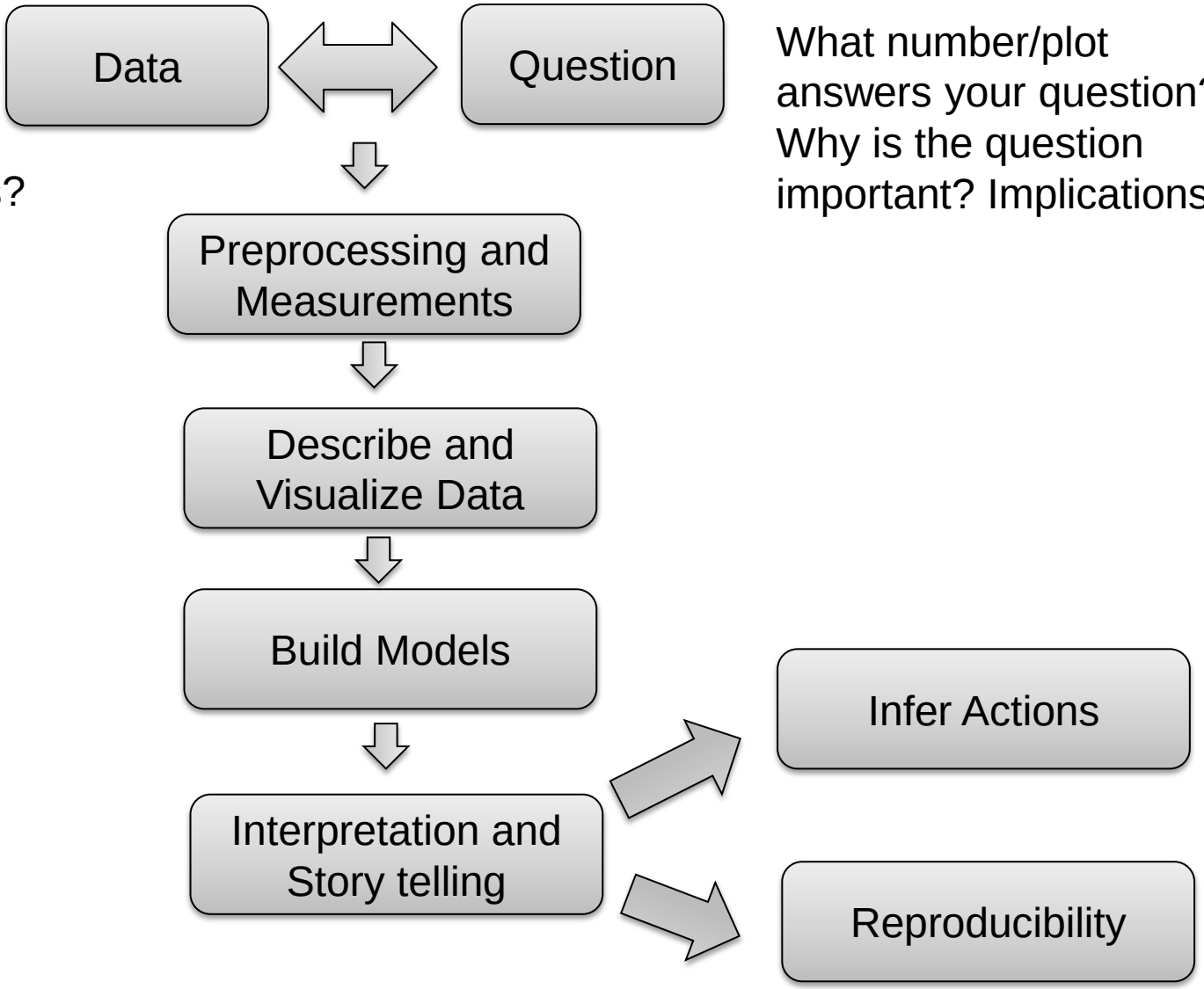
- Josh Blumenstock

- Skills needed
  - Statistics, machine learning, ability to handle big data
  - Scientific curiosity & methodology, story telling, creativity, visualization skills and so on

Is data science a new field?

Activities of Data Scientists

How was the data collected? Sampling Bias? Measurement Bias?



What number/plot answers your question? Why is the question important? Implications?

Collect Organic Data

- Data Dumps
- APIs and Sparql endpoints
- Web Scraping
  - Parse HTML
  - Dynamically loaded content
    - E.g. PhantomJS or Selenium

Sample Bias

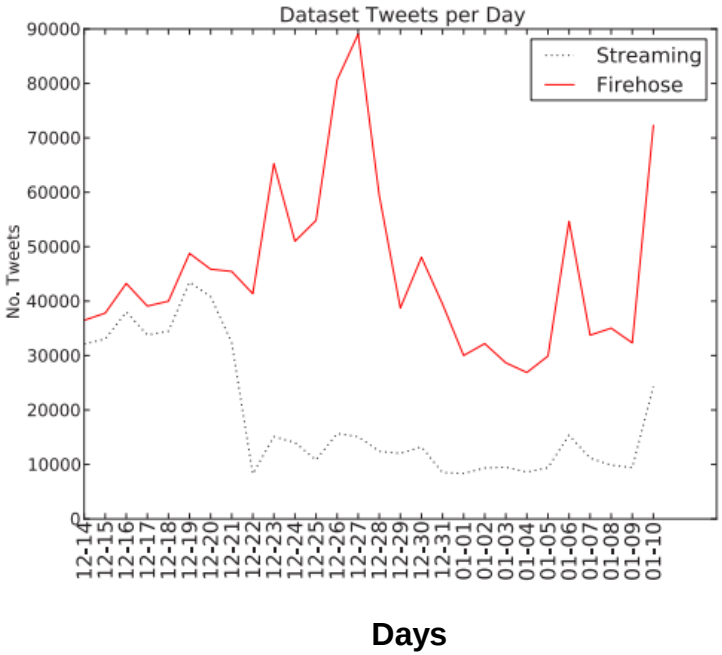
- How to get a random sample of Twitter users?
  - Random sample of tweets from Streaming API
  - Streaming API gives us a random sample of tweets (max 1% of total traffic)

  - Problems?
  - More active users are more likely to be included

  - Better Approach?

Streaming API versus Firehose

How to get a random sample of tweets for certain keywords? Just use the Streaming API?

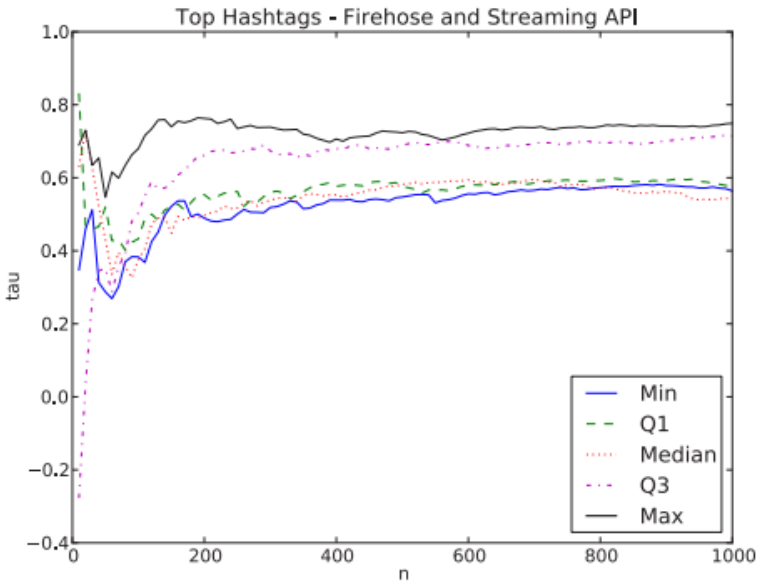


When Firehose spikes the Streaming API coverage was reduced

Probably the threshold was changed here

F Morstatter, J Pfeffer, H Liu, KM Carley, **Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose**, ICWSM 2013

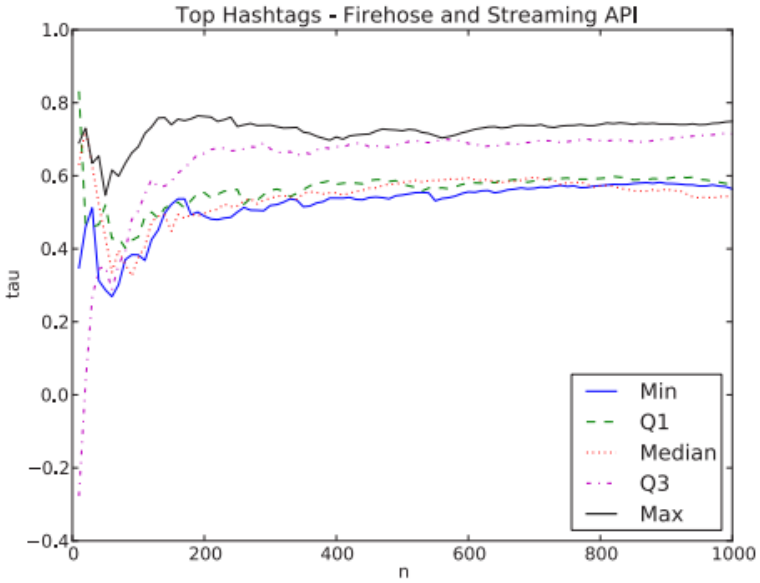
Streaming API versus Firehose



Number of top hashtags

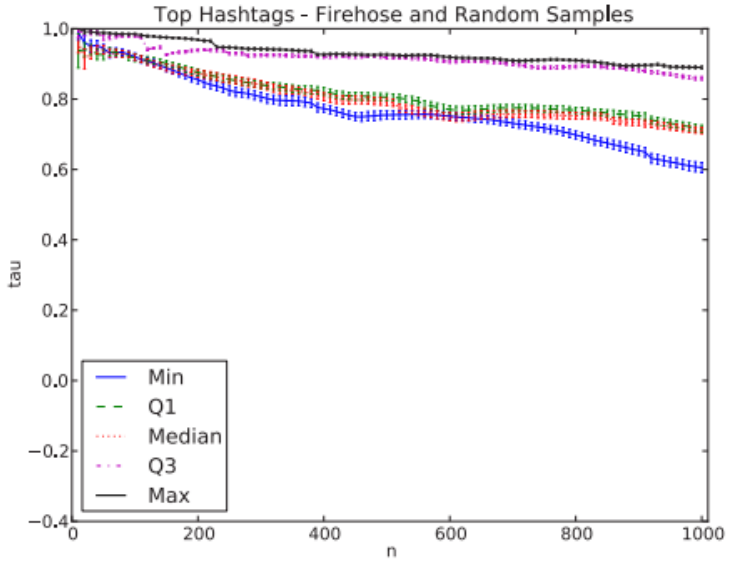
F Morstatter, J Pfeffer, H Liu, KM Carley, **Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose**, ICWSM 2013

Streaming API versus Firehose



Number of top hashtags

Uniform random samples of tweets



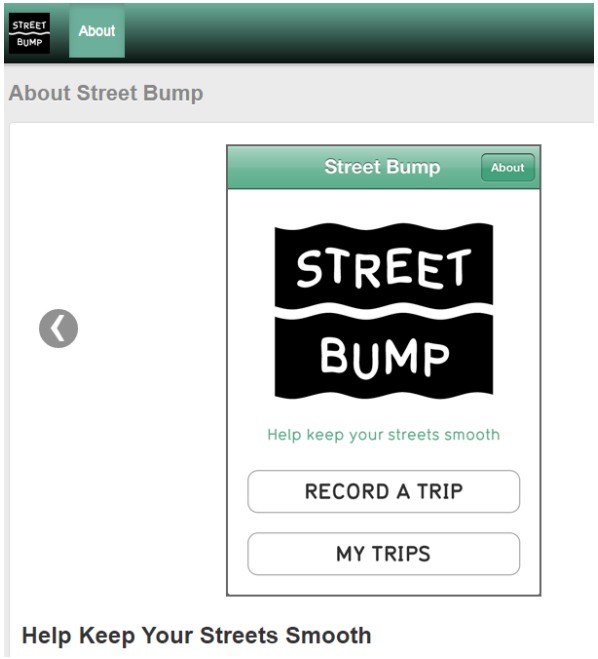
Number of top hashtags

F Morstatter, J Pfeffer, H Liu, KM Carley, **Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose**, ICWSM 2013

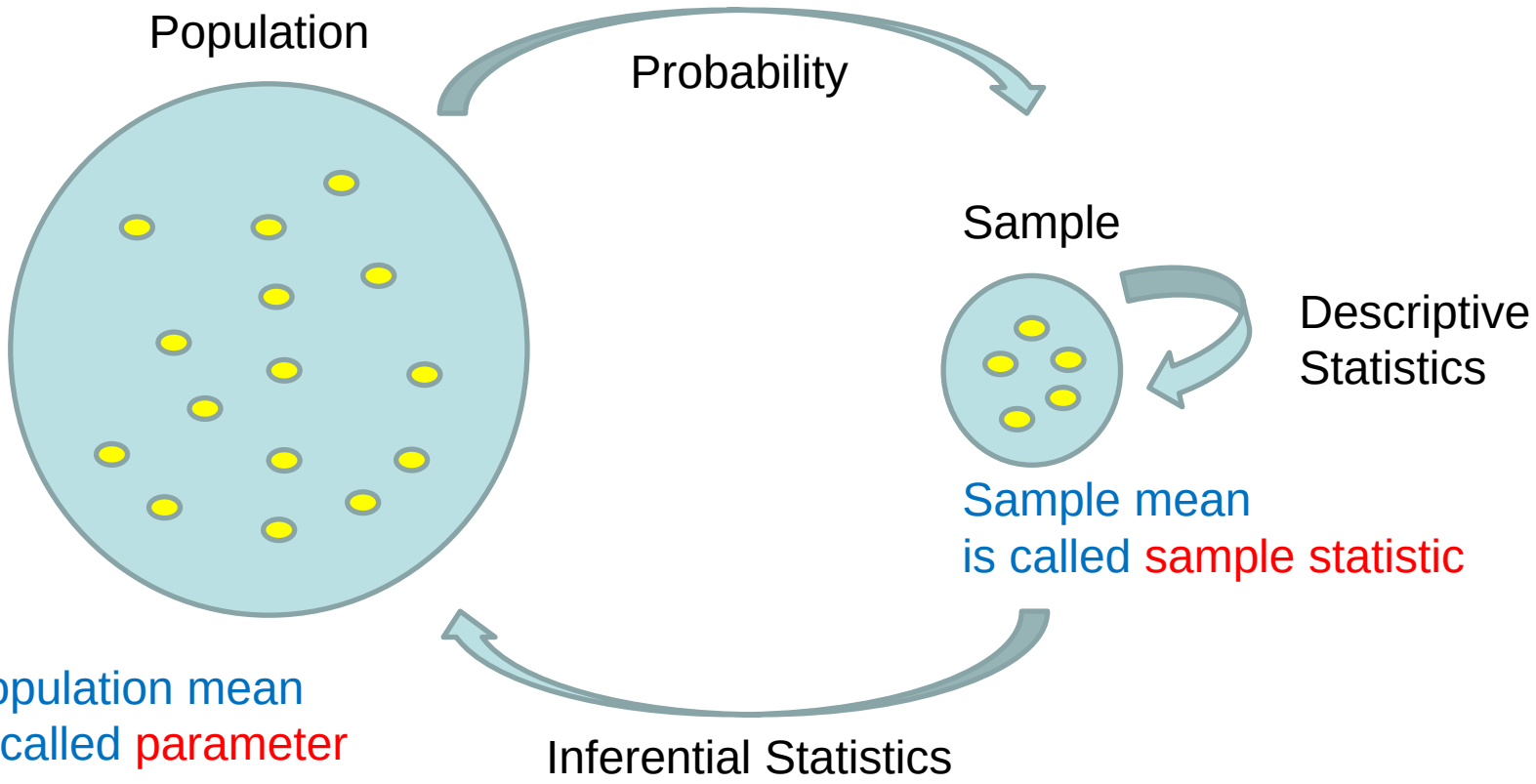
Data Bias

What if we access to all data of one system?

Can we use this data to inform the city about where to fix streets?



High correlation between number of street bumps and wealth of an area

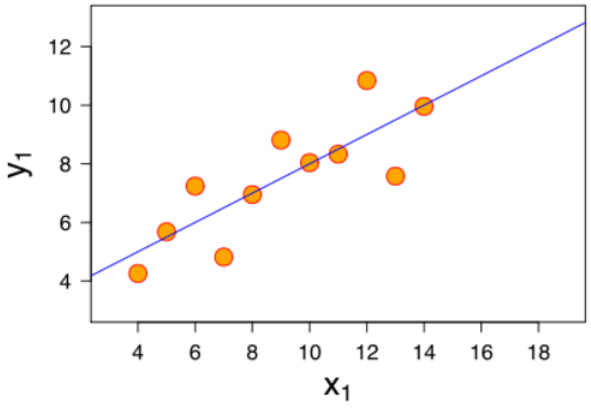


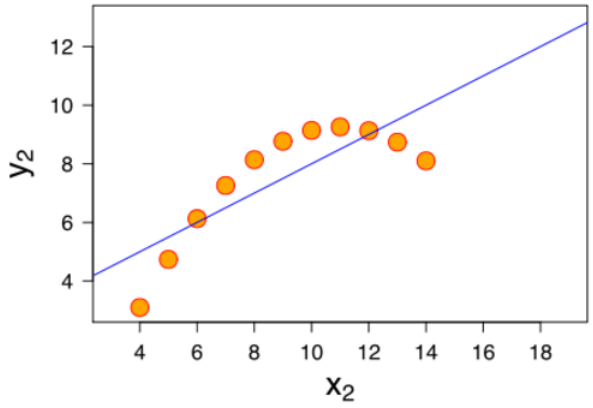
Find „good“ estimator

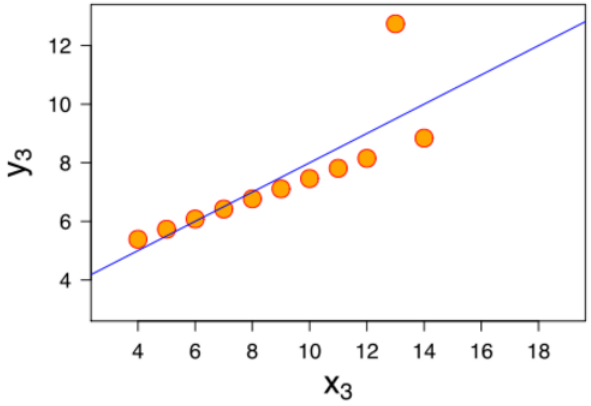
Describe Data

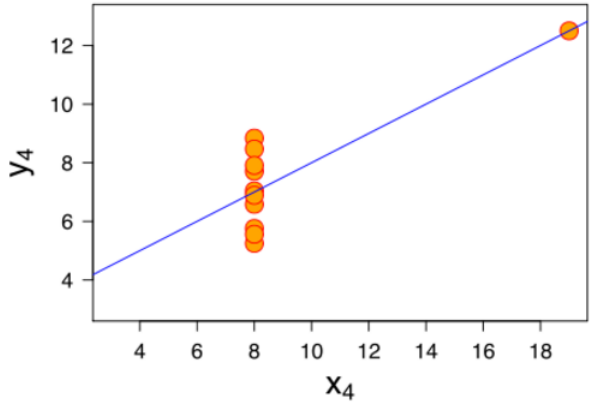
Anscombe's quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89









Cor(x,y) = 0.816

Mean(y) = 7.50

Var(y) = 4.122 or 4.127

Y=4+0.500*X

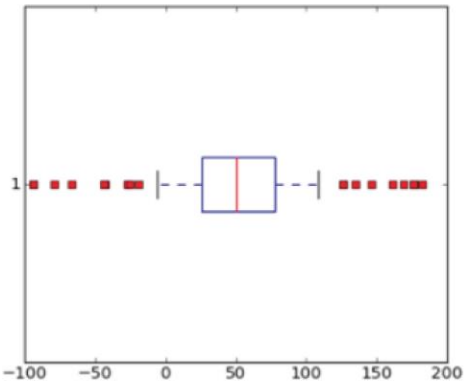
VISUALIZE DATA

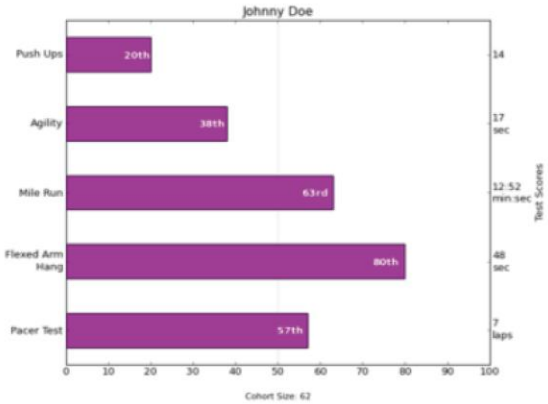
Univariate Data

1 Variable only!

Describe it: e,g, central tendency, dispersion,

Plot it: frequency distributions, bar graph, histogram, pie chart, line, graph, box-and-whisker plot

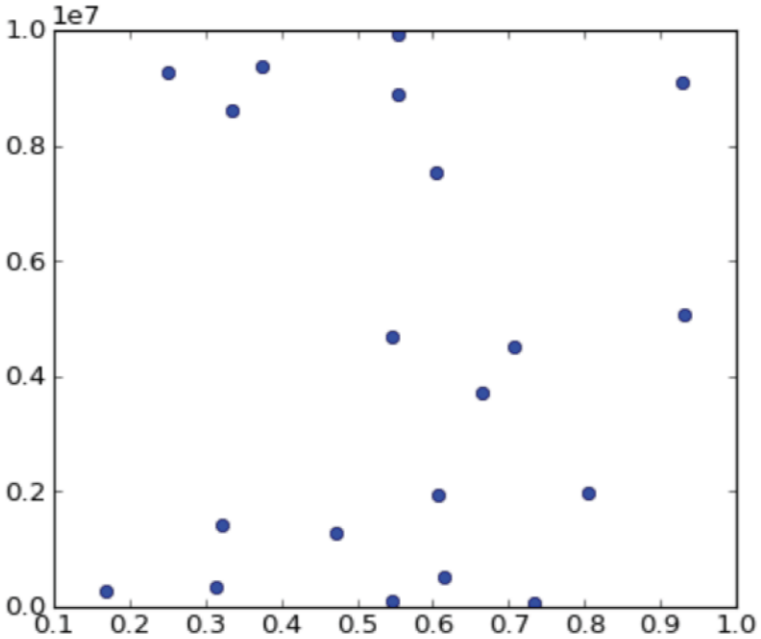




Based on slide from M.Agrawala

Bivariate Data

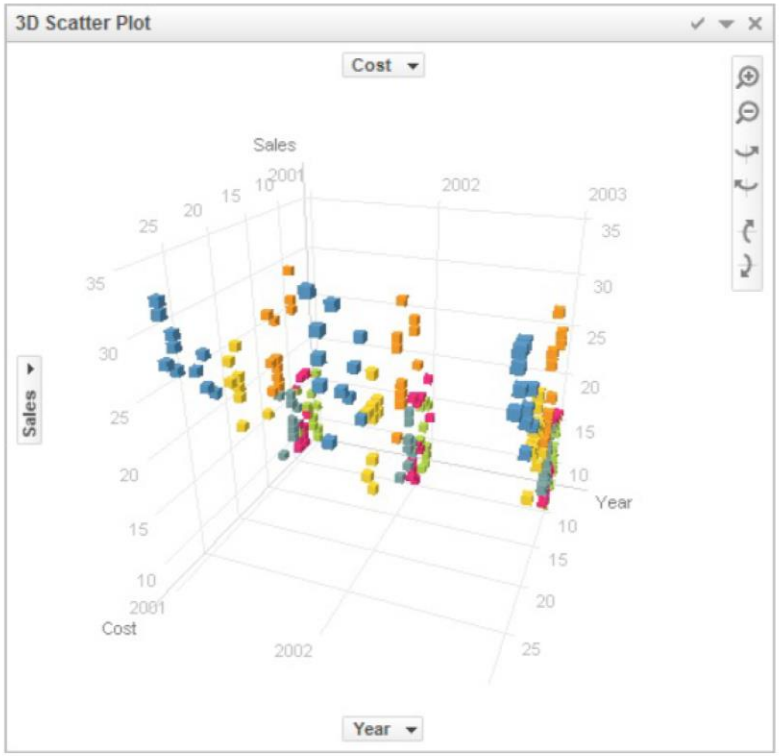
Scatterplot is common



Based on slide from M. Agrawala

Trivariate Data

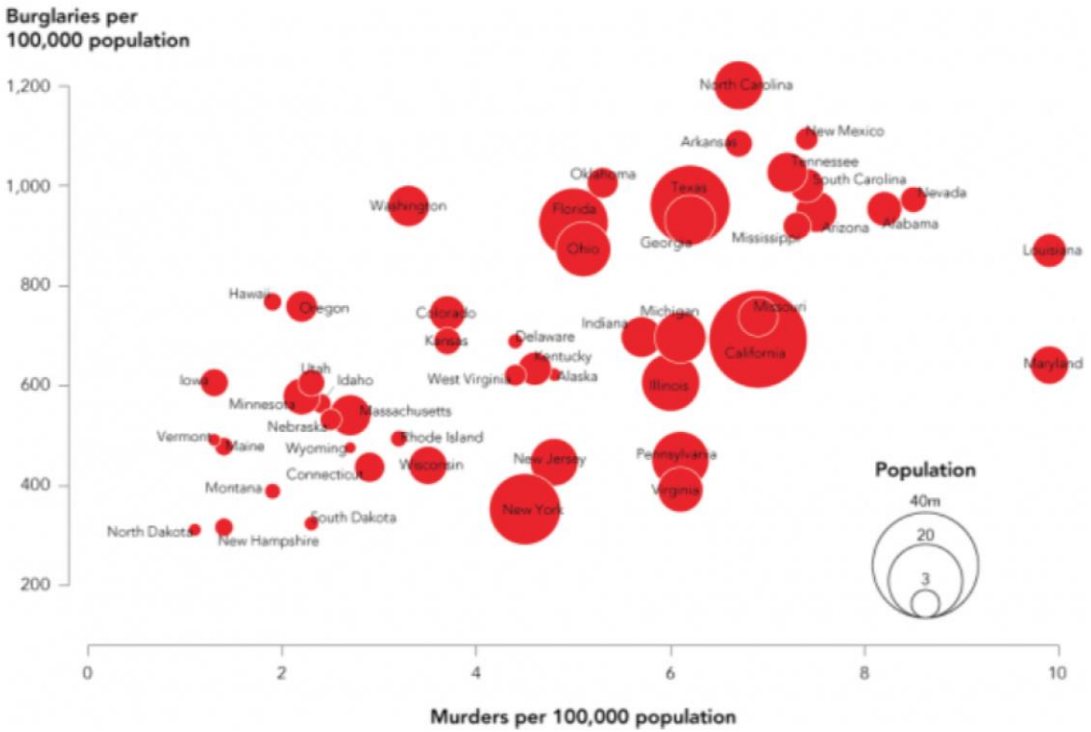
Do NOT use 3D scatterplots!



Based on slide from M.Agrawala

Trivariate Data

Map the third dimension to some other visual attribute

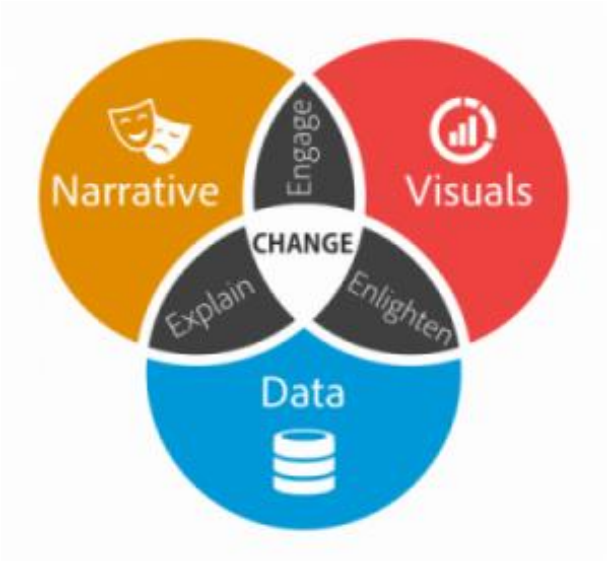


Based on slide from M. Agrawala

STORIES AND CONCLUSIONS


Sheryl Sandberg:
Why we have too few women leaders
TEDWomen 2010 · 14:58 · Filmed Dec 2010
46 subtitle languages
View interactive transcript
Share this idea
Facebook
LinkedIn
Twitter
Link
Email
Embed
6,387,155 Total views

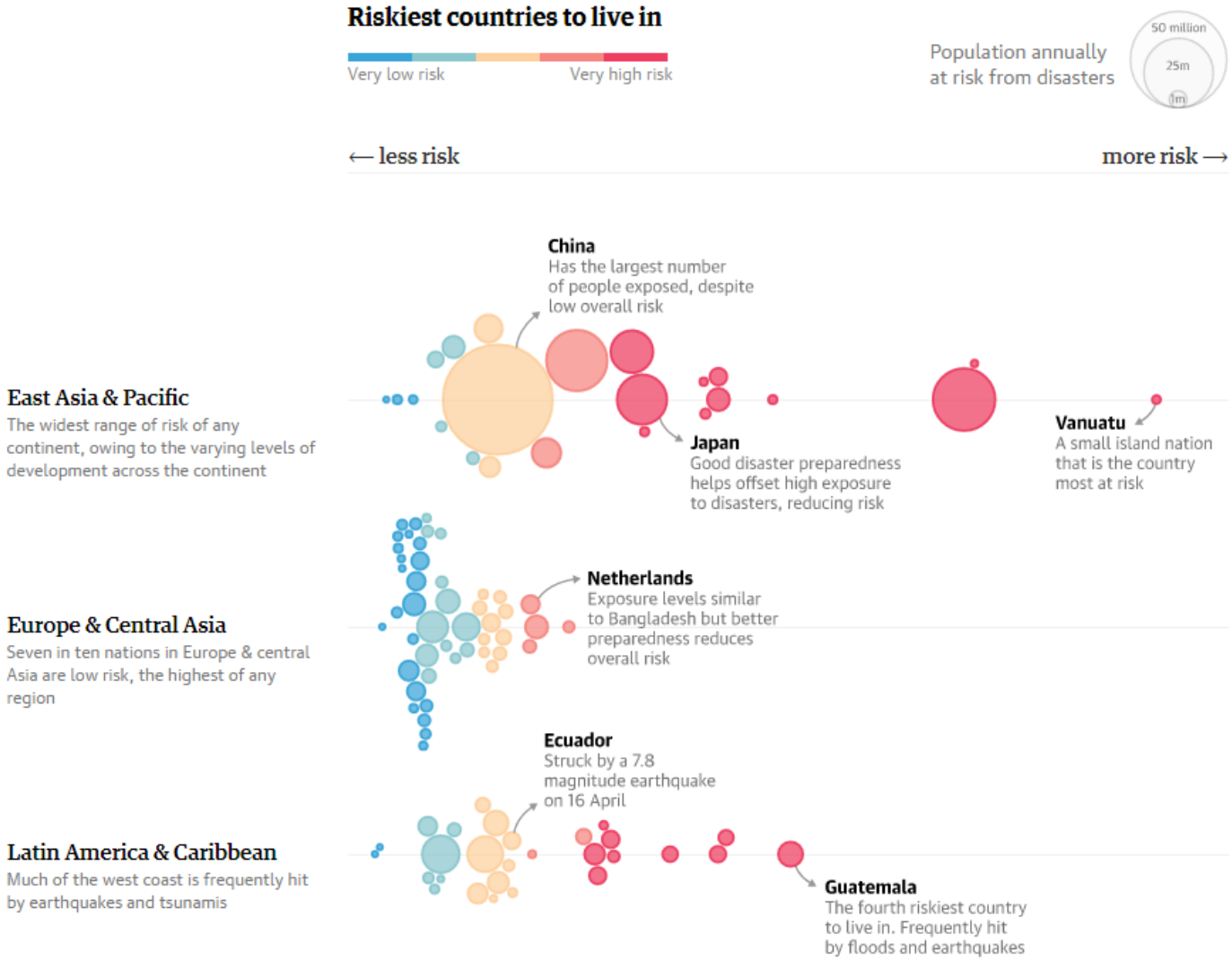
- Stories are hard to forget. Stories connect us.
- Good story drives ***change***
  - Flawed assumption: decisions are based solely on logic and reason



http://www.forbes.com/sites/brentdykes/2016/03/31/data-storytelling-the-essential-data-science-skill-everyone-needs

Storytelling Guidelines

- First decide what's your goal, your main point
- Organizes facts into compelling narrative
  - Narrative is "an account of a series of events, facts, etc., given in order and with the establishing of connections between them."
    - Start with the problem/question
    - Attempt to resolve/answer the problem/question
    - Ends with the resolution
    - Sometimes useful: use a protagonist
- Think about the audience
- Include visualization to supports narrative
- Engage them and make them think (give them 2+2 not 4)



https://www.theguardian.com/global-development/datablog/2016/apr/25/where-is-the-riskiest-place-to-live-floods-storms

Conclusions

- Data Science is exciting
- Lots of potential to improve business and science

- High demand
  - By 2018 the number of data science jobs in US alone will exceed 490k, but there will be fewer than 200k according to a McKinsey study.

- What is needed?
  - Knowledge about data and potential biases
  - Statistic, machine learning, ability to handle big data, scientific methods, story telling, creativity, visualization skills and so on

QUESTIONS