

University of Koblenz • Landau, Germany

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

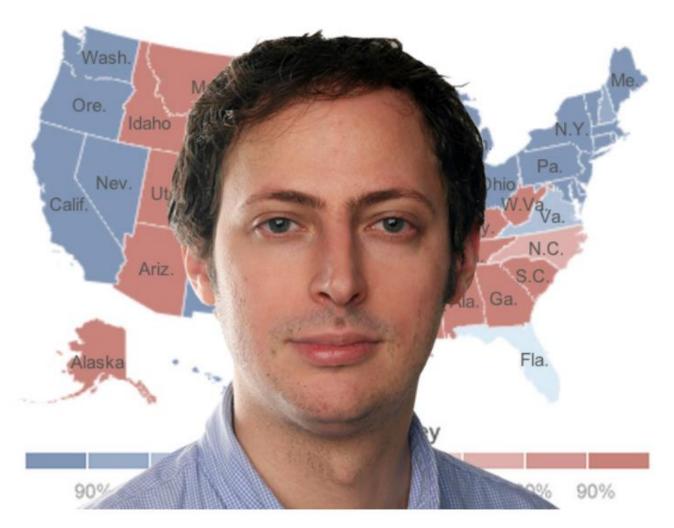
Since 2013:

Post Doc at GESIS – CSS department

Since 2016:

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

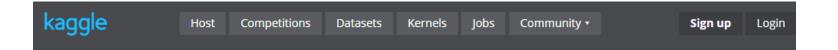




Hanspeter Pfister, Data Science course, Introduction https://drive.google.com/drive/folders/0BxYkKyLxfsNVd0xicUVDS1dIS0k

Success Stories



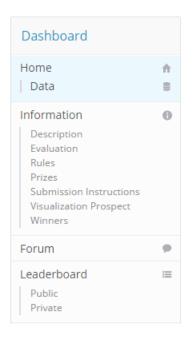




Completed • \$40,000

Merck Molecular Activity Challenge

Thu 16 Aug 2012 - Tue 16 Oct 2012 (3 years ago)



Help develop safe and effective medicines by predicting molecular activity.

Help enable the development of safe, effective medicines.

When developing new medicines it is important to identify molecules that are highly active toward their intended targets but not toward other targets that might cause side effects. The objective of this competition is to identify the best statistical techniques for predicting biological activities of different molecules, both on- and off-target, given numerical descriptors generated from their chemical structures.

The challenge is based on 15 molecular activity data sets, each for a biologically relevant target. Each row corresponds to a molecule and contains descriptors derived from that molecule's chemical structure.



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 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



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 a

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

Abstract Full Text Authors & Info Figures SI Metrics Related Content PDF PDF + SI

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 D6.23.08 12:00 PM

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

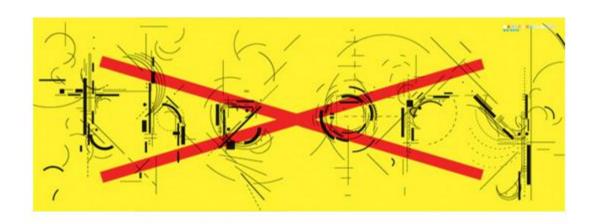


Illustration: Marian Bantjes



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



Activities of Data Scientists



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

Data Question

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

Preprocessing and Measurements



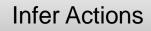
Describe and Visualize Data



Build Models



Interpretation and Story telling



Reproducibility

Collect Organic Data



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

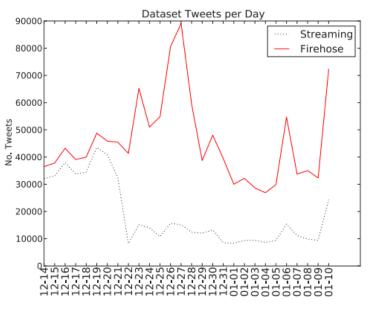
Sample Bias



- 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?



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

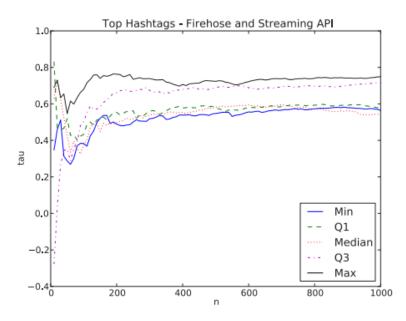
Probably the threshold was changed here

Days

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





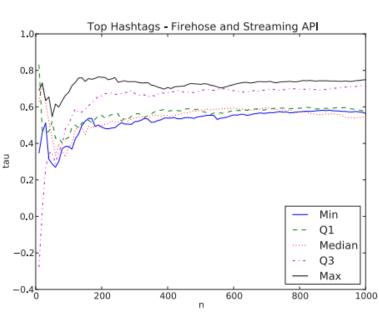


Number of top hashtags

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

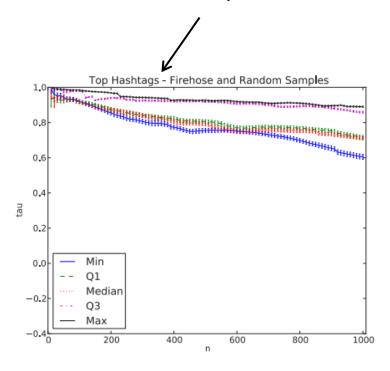
Streaming API versus Firehose





Number of top hashtags

Uniform random samples of tweets



Number of top hashtags

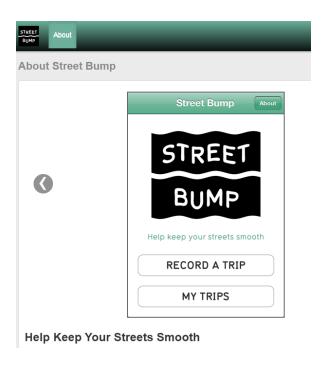
F Morstatter, J Pfeffer, H Liu, KM Carley, Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose, ICWSM 2013





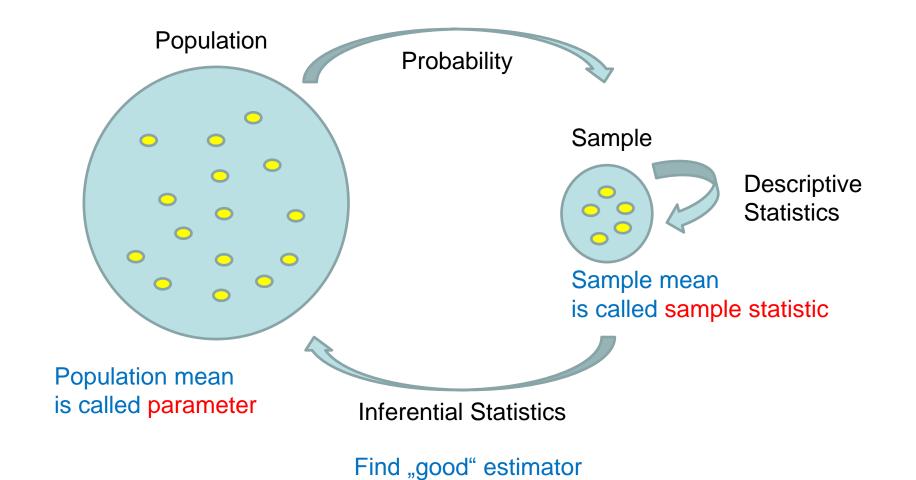
What if we access to all data of one system?

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



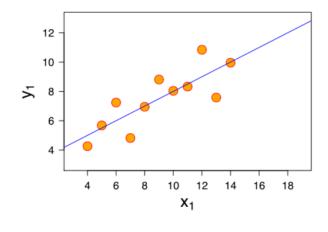


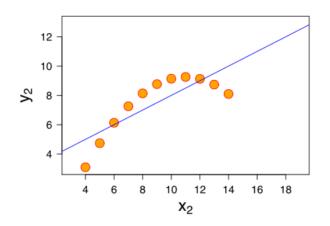


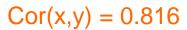
Anscombe's quartet

I		II		III		IV	
x	у	X	у	X	у	X	у
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



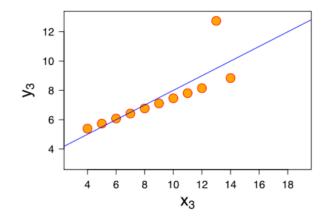


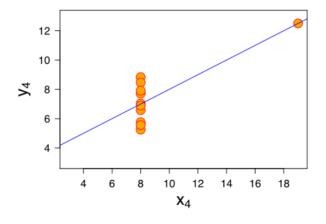




$$Mean(y) = 7.50$$

$$Var(y) = 4.122 \text{ or } 4.127$$







VISUALIZE DATA



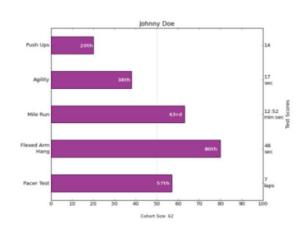


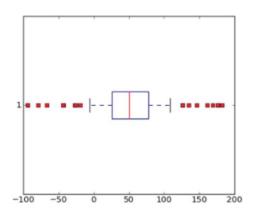
Univariate Data

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

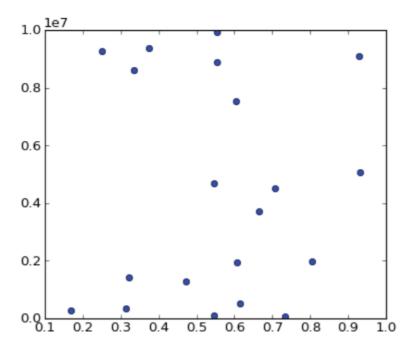


WeST



Bivariate Data

Scatterplot is common



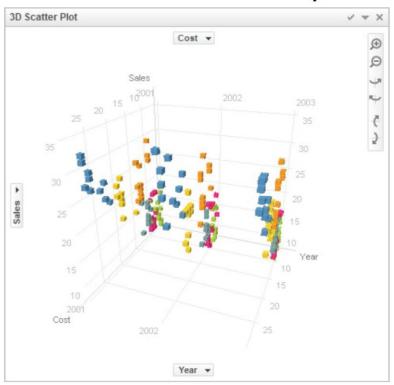
Based on slide from M.Agrawala





Trivariate Data

Do NOT use 3D scatterplots!



Based on slide from M.Agrawala

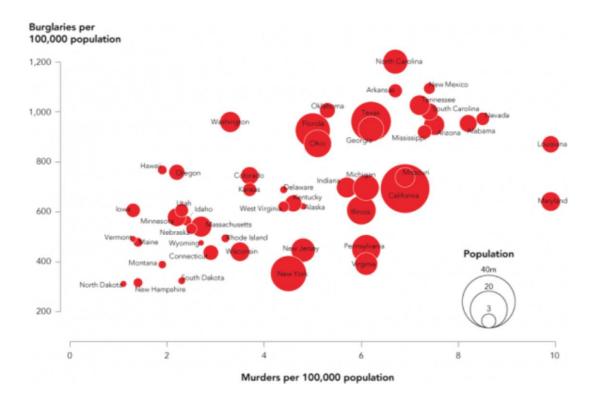


WeST



Trivariate Data

Map the third dimension to some other visual attribute



Based on slide from M.Agrawala



WeST



STORIES AND CONCLUSIONS







Share this idea













6,387,155 Total views

UNIVERSITÄT KOBLENZ:LANDAU



- 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)



Riskiest countries to live in 50 million Population annually 25m Very low risk Very high risk at risk from disasters ← less risk $more risk \rightarrow$ China Has the largest number of people exposed, despite low overall risk East Asia & Pacific Vanuatu 4 The widest range of risk of any A small island nation Japan continent, owing to the varying levels of Good disaster preparedness that is the country development across the continent helps offset high exposure most at risk to disasters, reducing risk Netherlands Exposure levels similar to Bangladesh but better Europe & Central Asia preparedness reduces Seven in ten nations in Europe & central overall risk Asia are low risk, the highest of any region Ecuador Struck by a 7.8 magnitude earthquake on 16 April Latin America & Caribbean Much of the west coast is frequently hit Guatemala by earthquakes and tsunamis The fourth riskiest country to live in. Frequently hit by floods and earthquakes

32

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



WeST

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

