

> How STS can Improve Data Science

- > Momin M. Malik, PhD <momin_malik@cyber.harvard.edu>
Data Science Postdoctoral Fellow
Berkman Klein Center for Internet & Society at Harvard University

Tufts University STS Lunch Seminar, 23 January 2020
Slides: <https://mominmalik.com/tufts2020.pdf>

› Outline

- › Introduction
- › Setting the stage
- › Overarching STS themes
- › Bias in geotagged tweets
- › Platform effects on Facebook
- › “Prediction” in machine learning
- › Conclusion

› Introduction

› Setting the stage

› Overarching STS themes

› Bias in geotagged tweets

› Platform effects on Facebook

› “Prediction” in machine learning

› Conclusion

› References

» Introduction

» Setting the
stage

» Overarching
STS themes

» Bias in
geotagged
tweets

» Platform
effects on
Facebook

» "Prediction"
in machine
learning

» Conclusion

» References

» Setting the stage

What is “data science”?

Introduction

Setting the stage

Overarching STS themes

Bias in geotagged tweets

Platform effects on Facebook

“Prediction” in machine learning

Conclusion

References

ANATOMY OF A DATA SCIENTIST

SALARY
Average salary of data scientist is \$126,000/year

EDUCATION
68% of all data scientists have at least a Master's degree
46% of data scientists have a PhD

BENEFITS
Harvard Business Review ranked data science the "Sexiest Job of the 21st Century"
One of the fastest growing careers in the United States
95% of data science graduates found jobs since 2011

RESPONSIBILITIES
Consultant
Extract, clean, and analyze data from various sources
Solve problems
Build software tools
Communicate findings to management

CAREER POSSIBILITIES
The majority of data scientists work in the technology industry
Other industries include marketing, consulting, healthcare and other technology, finance, government, gaming, and more.

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21st century requires a mixture of multi-disciplinary skills ranging from an understanding of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand why a data scientist is so highly paid is equally hard. So here is a little cheat sheet on what the modern data scientist really is.

MATH & STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning (decision trees, neural networks, linear & logistic regression)
- Unsupervised learning (clustering, dimensionality reduction)
- Optimization (gradient descent, etc.)

PROGRAMMING & DATABASE

- Computer science fundamentals
- Scripting languages (e.g. Python, Perl)
- Database management packages (e.g. R, Oracle, SQL and NoSQL)
- Relational algebra
- Parallel computers and parallel programming
- MapReduce concepts
- Hadoop and Hive/Hy
- Custom software
- Experience with cloud (AWS, etc.)

COMMUNICATION & VISUALIZATION

- Ability to engage with non-technical management
- Strong public skills
- Transferable data communication skills (lecturing, writing, etc.)
- Visual art design
- Strategic, creative, critical, innovative and critical analysis
- Knowledge of key communication tools (e.g. Power, PPT, Tableau)

DOMAIN KNOWLEDGE & SOFT SKILLS

- Fluency about the business
- Current about data
- Influence without authority
- Problem solver
- Strategic, creative, critical, innovative and critical analysis

FRANKENSTEIN'S DATA SCIENTIST

What exactly is Frankenstein's Data Scientist made up of?

EYES
Asking the right questions is often what sets you apart from the rest. It's important to be able to know what questions to ask and what you're going to find.

BRAIN
Any analyst that deals with a novel or uncharted data set only and uncharted results might not succeed in the long run. History, Spoken, Read and SAS.

MOUTH
To get ahead in business you need to be a great communicator. We demand that Data Scientists are capable of making their findings clear and easy to understand for those who need to make decisions based on their findings.

HEART
All projects and decisions approach to your work to solve business problems from a different angle will set you apart.

HANDS
The best data scientists are those who are able to take a problem and solve it in a way that is both simple and innovative. They are able to take a problem and solve it in a way that is both simple and innovative.

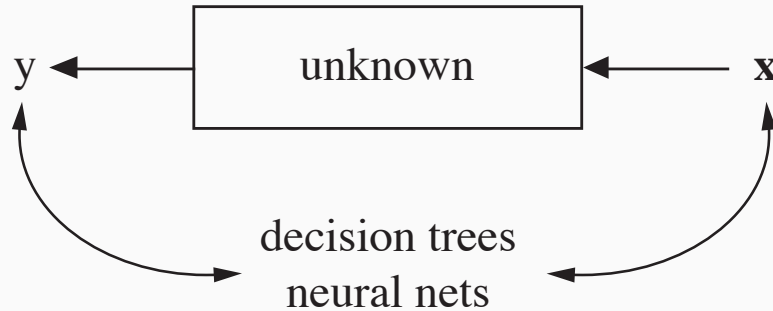
FEET
You need to be able to get up and move. You need to be able to move. You need to be able to move. You need to be able to move.

DATA SCIENTIST MUST-HAVE SKILLS

- Machine Learning
- Statistical Modeling
- Exploratory Analysis
- Clustering
- Regression Analysis
- Computer Science Fundamentals
- Database Management System
- Data Visualization
- Python
- Big Data
- Storytelling skills
- Convert data-based insights into decisions
- Collaborative with Sr. Management
- Knowledge of tools like Tableau
- Visual art design

Applied statistics and applied machine learning, mostly in business

> (What is machine learning?)



> An instrumental use of correlations to *mimic* the output of a target process, rather than understand the *relationship* between inputs and outputs

Breiman, 2001. See also Jones, 2018.

> My background

> Introduction

> Setting the stage

> Overarching STS themes

> Bias in geotagged tweets

> Platform effects on Facebook

> "Prediction" in machine learning

> Conclusion

> References



"We check our **e-mails** regularly, make **mobile phone calls**... We may post **blog entries** accessible to anyone, or maintain friendships through **online social networks**. Each of these transactions leaves **digital traces** that can be compiled into comprehensive pictures of both individual and group behavior, with the **potential to transform our understanding of our lives, organizations, and societies.**"



YOU KEEP ON USING THESE DATA

I DO NOT THINK THEY MEAN WHAT YOU THINK THEY MEAN

› Introduction

› Setting the stage

› Overarching STS themes

› Bias in geotagged tweets

› Platform effects on Facebook

› “Prediction” in machine learning

› Conclusion

› References

› Overarching STS themes

> STS theme: Imagination

> Introduction

> Setting the stage

> Overarching STS themes

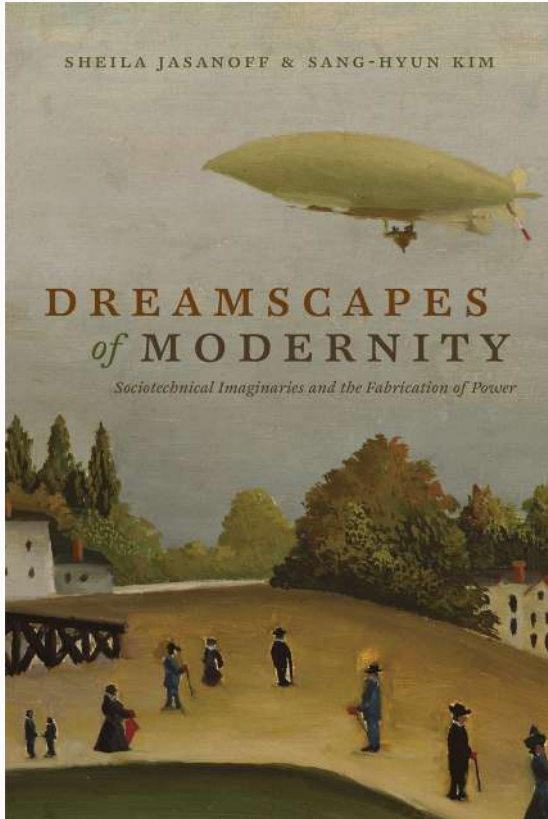
> Bias in geotagged tweets

> Platform effects on Facebook

> "Prediction" in machine learning

> Conclusion

> References



> Imagination “operates at an intersubjective level, uniting members of a social community in shared perceptions of **futures that should or should not be realized.**”

➤ STS theme: Instruments

➤ Introduction

➤ Setting the stage

➤ Overarching STS themes

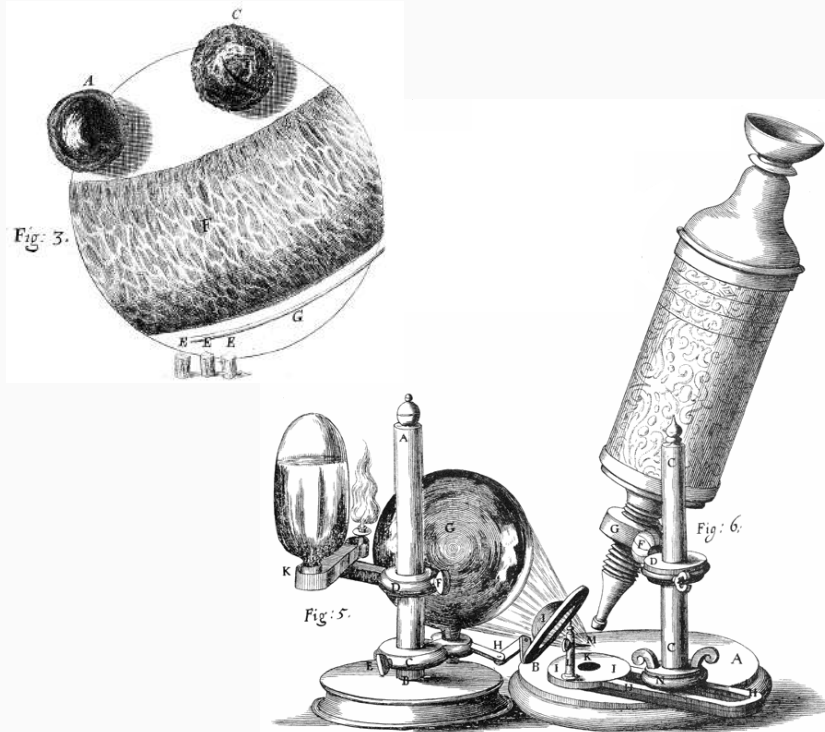
➤ Bias in geotagged tweets

➤ Platform effects on Facebook

➤ "Prediction" in machine learning

➤ Conclusion

➤ References



“The incongruity of [the natural object a specimen was supposed to represent, and the specimen] generated... a peculiar need to perpetually bring the object back to an initial stage of examination, whereby the experiment was constantly stating its own discursive authority in **an attempt to do away with the shortcomings of a yet-imperfect instrument.**” (Szekely, 2011)

Robert Hooke (1665). *Micrographia: or some phyſiological deſcriptions of minute bodies made by magnifying glaſſes. With obſervations and inquiries thereupon.*

> STS theme: Social construction

> Introduction

> Setting the stage

> Overarching STS themes

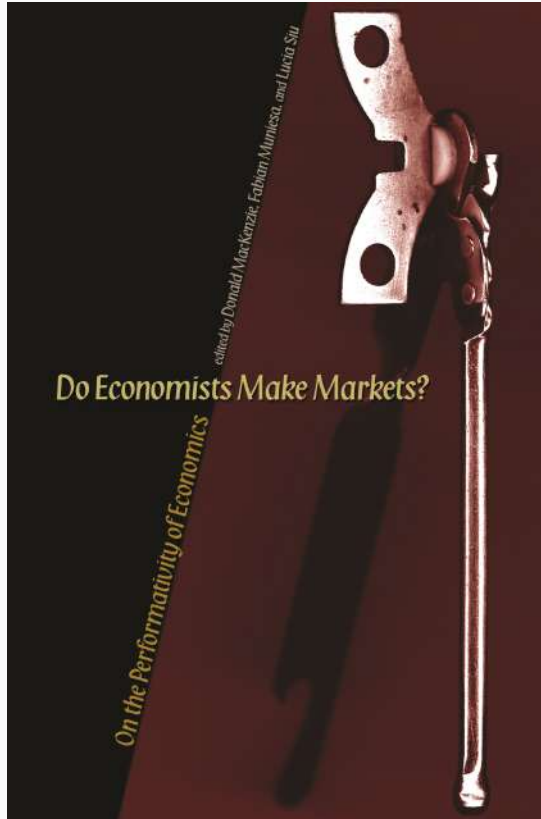
> Bias in geotagged tweets

> Platform effects on Facebook

> "Prediction" in machine learning

> Conclusion

> References



“the *performativity thesis* is that economics produces a body of formal models and transportable techniques that, when carried out into the world by its professionals and popularizers, **reformats and reorganizes the phenomena the models purport to describe...**” (Healy, 2015)

➤ STS themes: Power, Co-construction

➤ Introduction

➤ Setting the stage

➤ Overarching STS themes

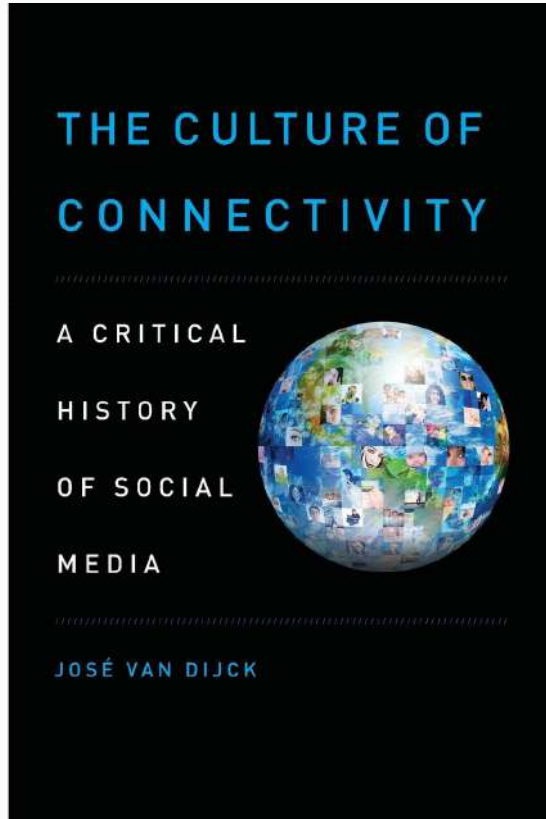
➤ Bias in geotagged tweets

➤ Platform effects on Facebook

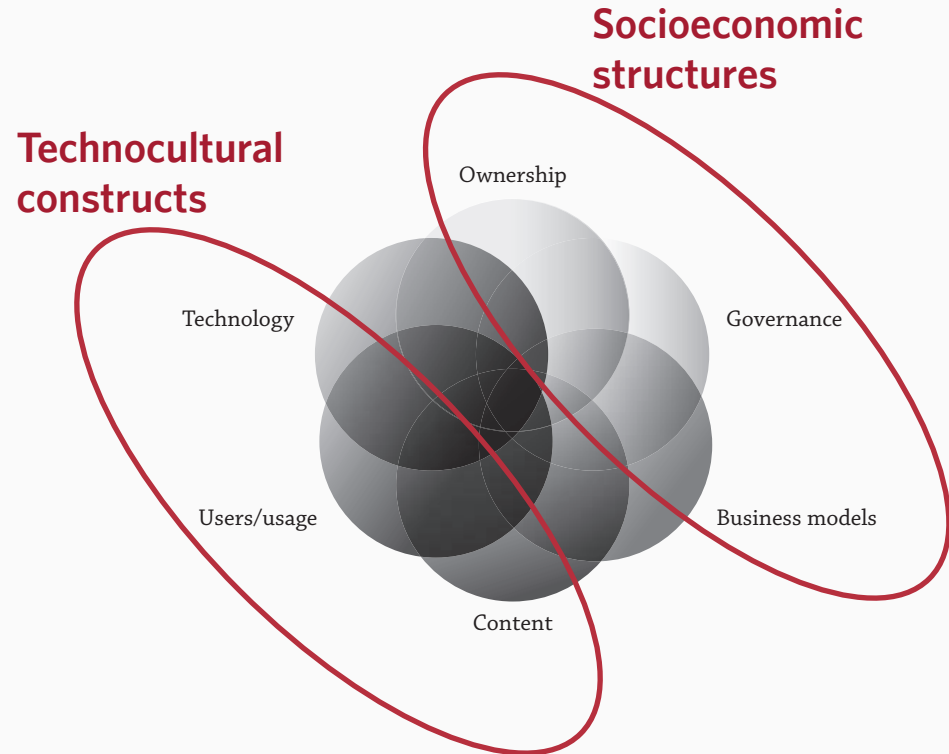
➤ "Prediction" in machine learning

➤ Conclusion

➤ References



How STS can Improve Data Science





› Introduction

› Setting the stage

› Overarching STS themes

› Bias in geotagged tweets

› Platform effects on Facebook

› “Prediction” in machine learning

› Conclusion

› References

› Bias in geotagged tweets

> Instruments, power

> Introduction

> Setting the stage

> Overarching STS themes

> Bias in geotagged tweets

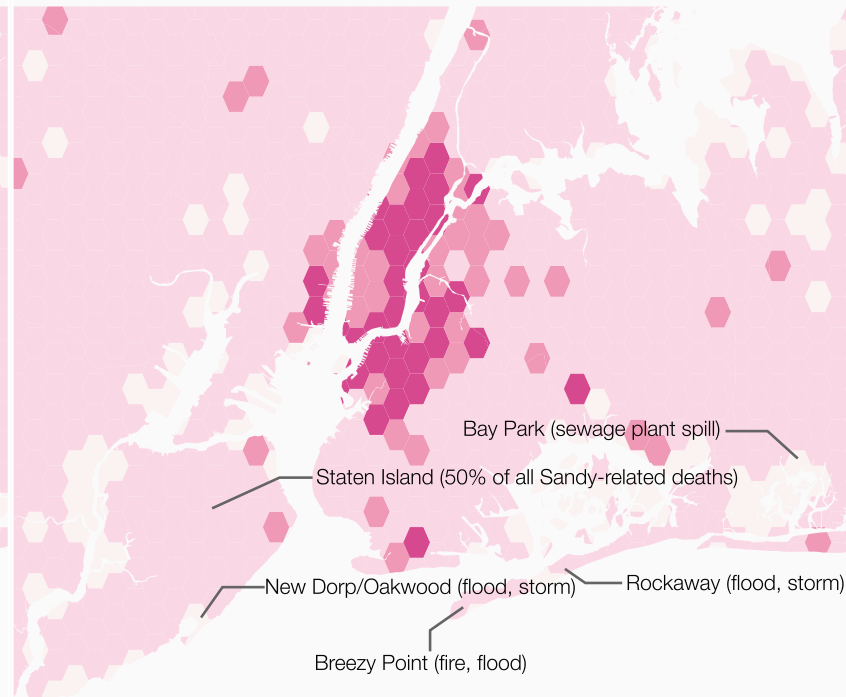
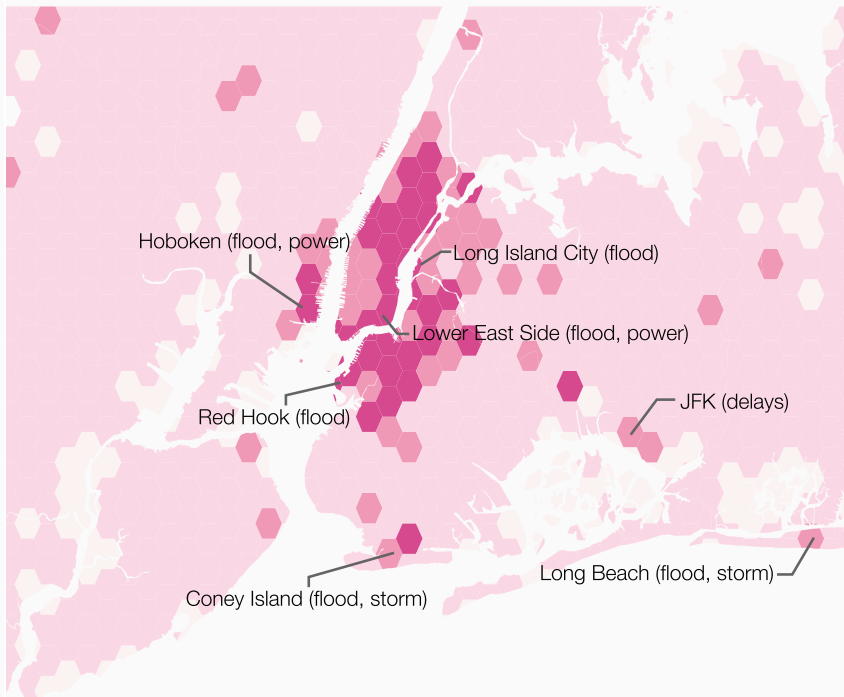
> Platform effects on Facebook

> "Prediction" in machine learning

> Conclusion

> References

Hurricane Sandy, tweets vs. damage/deaths



Shelton et al., 2014.

> What do instruments capture?

> Introduction

> Setting the stage

> Overarching STS themes

> Bias in geotagged tweets

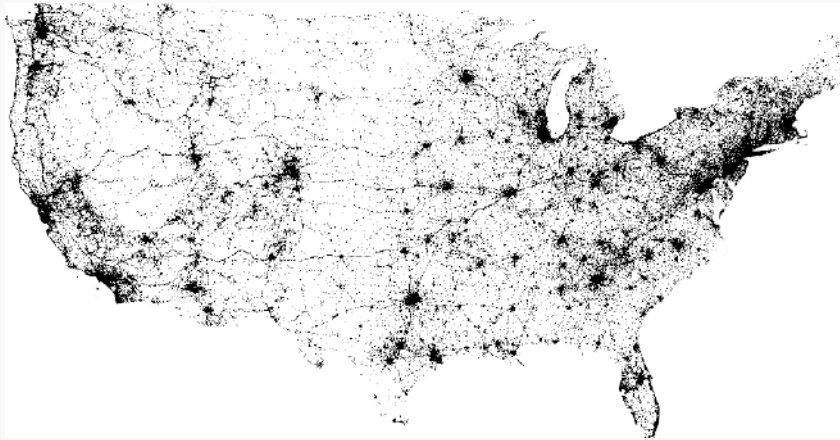
> Platform effects on Facebook

> "Prediction" in machine learning

> Conclusion

> References

Geotagged tweets



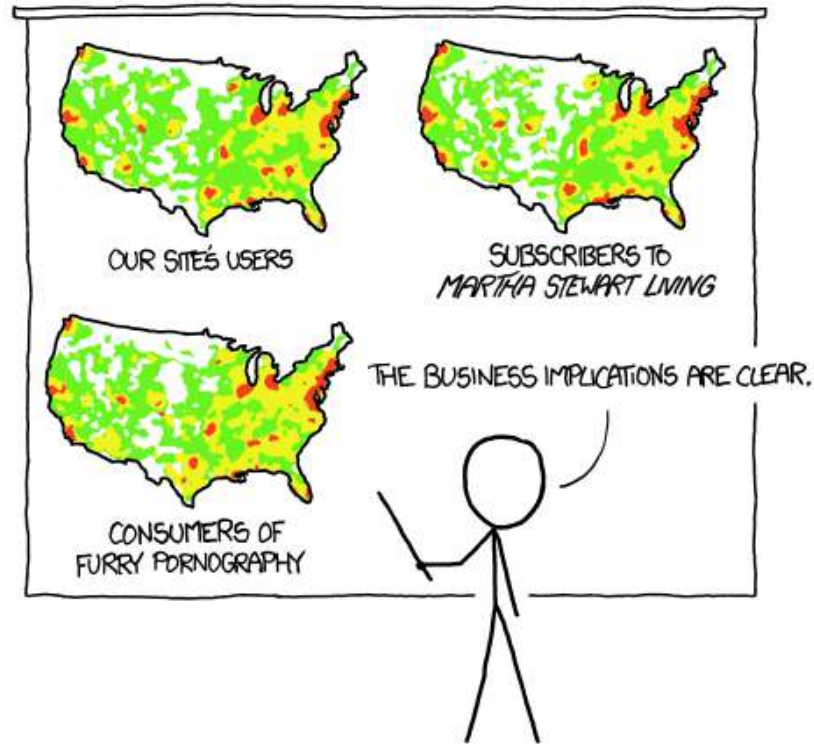
Adapted from Eric Fischer (2009), Contiguous United States geotag map, <https://flic.kr/p/a7WMWS>.

Population



Population density in 2010 US Census. Adapted from 'Nighttime Population Distribution Wall Map' by Geography Division, U.S. Department of Commerce / Economics and Statistics Administration / U.S. Census Bureau. Each square represents 1,000 people.

➤ What do instruments capture?



➤ Introduction

➤ Setting the stage

➤ Overarching STS themes

➤ Bias in geotagged tweets

➤ Platform effects on Facebook

➤ "Prediction" in machine learning

➤ Conclusion

➤ References

➤ Modeling population vs. users

➤ Users, and noise proportional to population:

$$U_i = \alpha P_i + \varepsilon_i P_i$$

➤ Take a log transformation:

$$\log U_i = \log \alpha + \log P_i + \varepsilon'_i$$

➤ Compare to a linear model:

$$\log U_i = \beta_0 + \beta_1 \log P_i + \varepsilon'_i$$

➤ Introduction

➤ Setting the stage

➤ Overarching STS themes

➤ Bias in geotagged tweets

➤ Platform effects on Facebook

➤ "Prediction" in machine learning

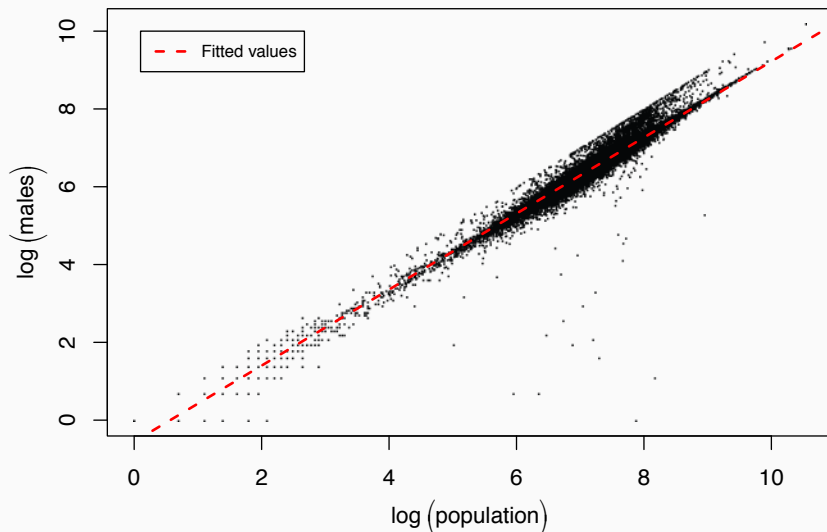
➤ Conclusion

➤ References

Result: Not proportional

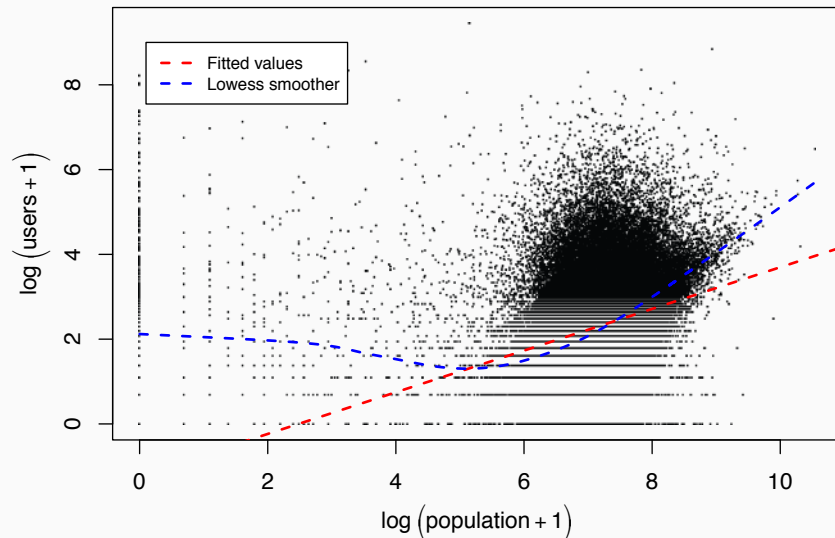
(Each dot is a *Census block group*)

Relationship between male population and total population
(null case)



This shows the model is good for capturing things that are proportional to population.

Relationship between population and geotag users



Geotagged tweet users are clearly *not* proportional to population.

Introduction

Setting the stage

Overarching STS themes

Bias in geotagged tweets

Platform effects on Facebook

"Prediction" in machine learning

Conclusion

References

› Identifying other differences

› Spatial multivariate modeling of biases

Geotagged tweet users associated with:

- ↓ Rural, poor, elderly, non-coastal
- ↑ Asian, Hispanic, black

› ...but these are only the demographics we can access. E.g., harassment of women on Twitter likely discourages geotag use

› Effects of this research?

- › Almost 100 citations in 4 years, all being used to say, “hey, we can’t just use tweets to study population”
- › Exactly my goal!
- › Many problems with the model, but specifics don’t matter as much, and basic point will be robust

› Introduction

› Setting the stage

› Overarching STS themes

› Bias in geotagged tweets

› Platform effects on Facebook

› “Prediction” in machine learning

› Conclusion

› References

» Introduction

» Setting the
stage

» Overarching
STS themes

» Bias in
geotagged
tweets

» Platform
effects on
Facebook

» “Prediction”
in machine
learning

» Conclusion

» References

» Platform effects on Facebook

> Power, co-construction

> Introduction

> Setting the stage

> Overarching STS themes

> Bias in geotagged tweets

> Platform effects on Facebook

> "Prediction" in machine learning

> Conclusion

> References



Markets Insider, Business Insider (2018)

- > Platforms: not neutral utilities or research environments
- > Platform engineers try to shape user behavior towards desirable ends

> “People you may know”

The screenshot shows the Facebook 'People you may know' interface. At the top is a search bar with the Facebook logo and the text 'Search Facebook'. Below this is a section titled 'People you may know' containing six user profiles. Each profile includes a profile picture, the user's name, location, mutual friends, and two buttons: 'Add Friend' and 'Remove'.

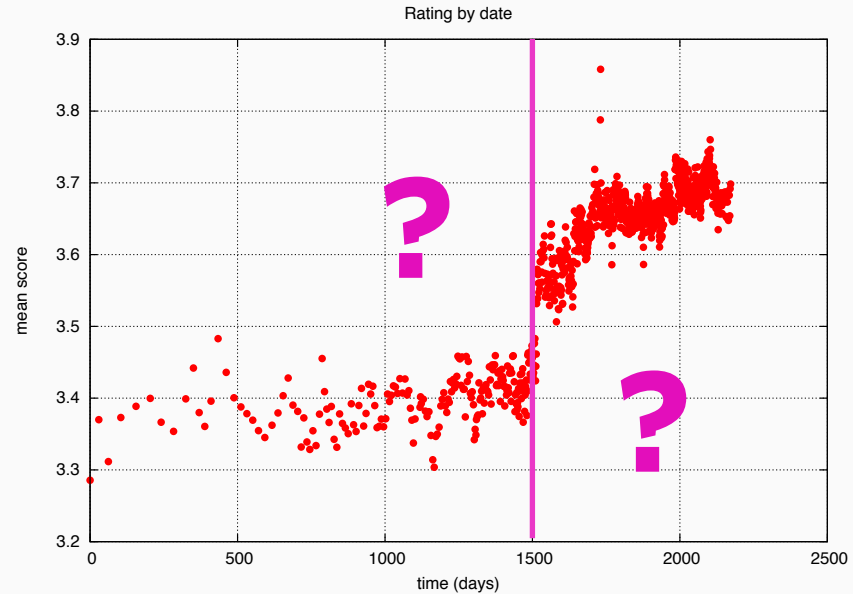
Name	Location	Mutual Friends	Action Buttons
Sara Anderson Severance	Denver, Colorado	Rachelle Albright and 10 other mutual friends	Add Friend, Remove
Anne Walker (Anne Anderson)		Sarah Frederick and 6 other mutual friends	Add Friend, Remove
Paul Dube		Ryan Dube is a mutual friend.	Add Friend, Remove
Mark Rieder	Lord Beaverbrook High School	Justin Pot is a mutual friend.	Add Friend, Remove
Nancy Mescher		Maggie Flynn is a mutual friend.	Add Friend, Remove
Becky Williams Swenson	Denver, Colorado	Rachelle Albright and 3 other mutual friends	Add Friend, Remove

Dann Abright, makeuseof.com

“Facebook uses its data on the structure of social relations to routinely suggest lists of **‘people you may know’** to users, with **the goal of encouraging users to add those people to their network...**” (Healy, 2015)

➤ DS research: Platform effects

- When we measure behavior, what are we really measuring? Social structure/behavior, or the effects of platform design and governance?
- Use discontinuities from data artifacts to make causal estimates



Average Netflix movie ratings over time. Each point averages 100,000 rating instances.

› Data artifacts and causal inference

- › Regression Discontinuity (RD) Design or Interrupted Time Series (ITS) estimate causality

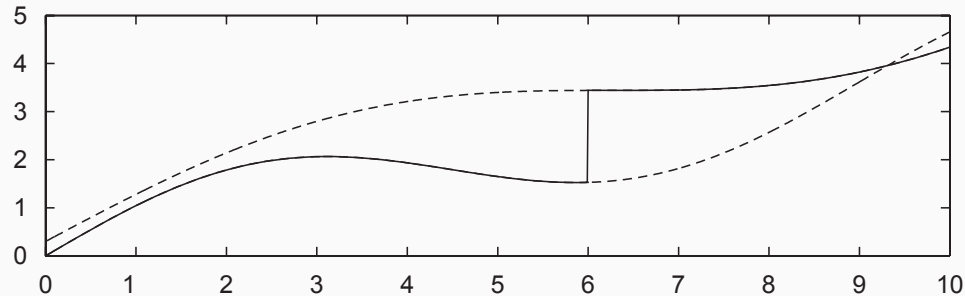


Fig. 2 from Imbens and Lemieux (2008): Potential and observed outcome regression functions.

- › The difference between "before" and "after" estimates the *local average treatment effect*

Facebook's "People you may know"

Introduction

Setting the stage

Overarching STS themes

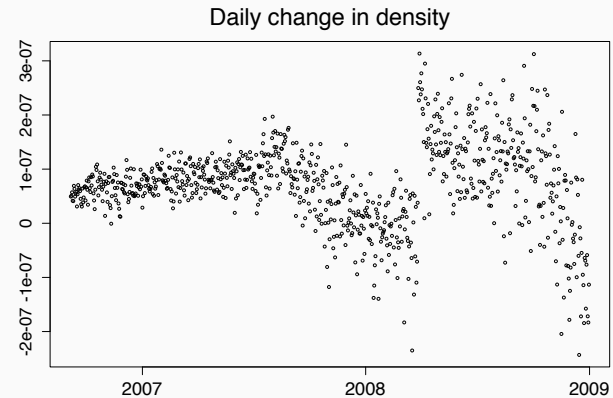
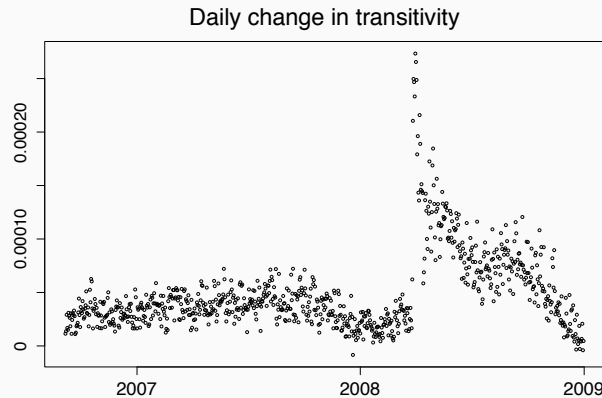
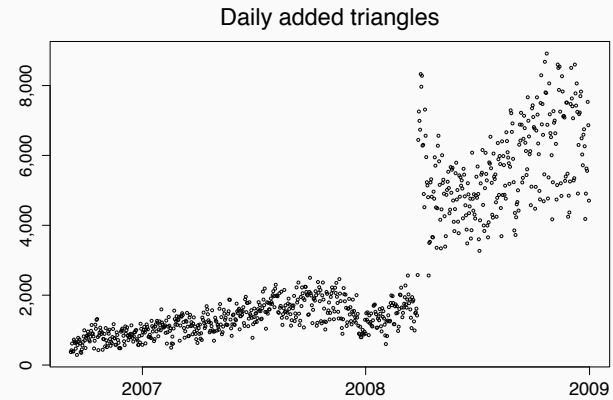
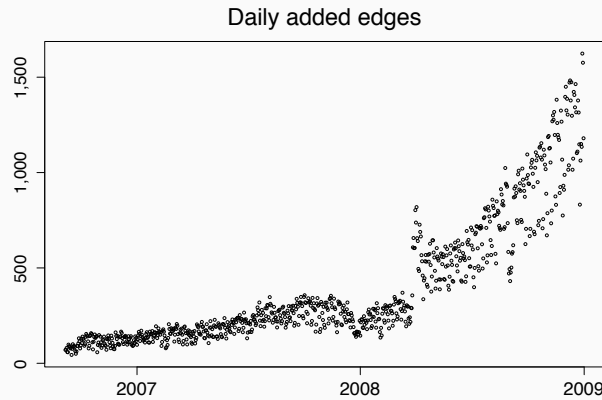
Bias in geotagged tweets

Platform effects on Facebook

"Prediction" in machine learning

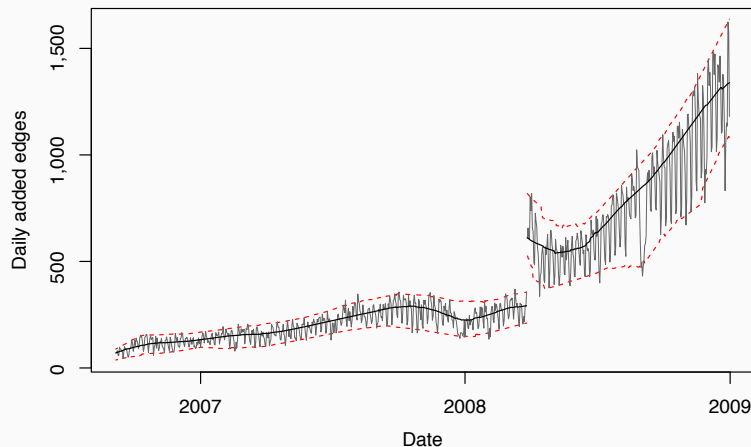
Conclusion

References

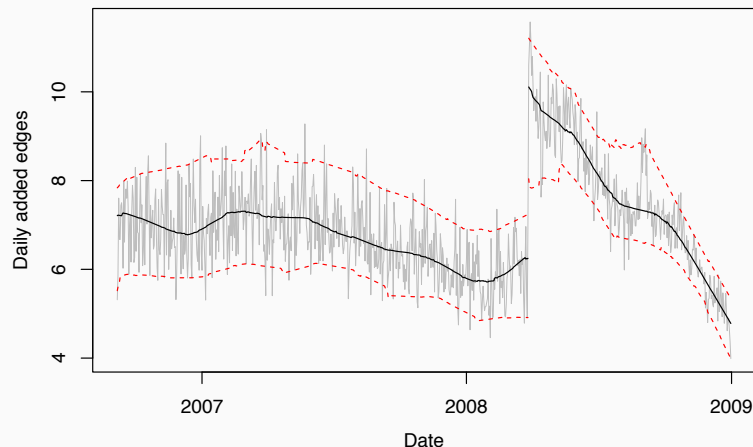


➤ Model the effects of PYMK

➤ Facebook links: +300 new edges per day (~200%)



➤ Triangles: +3.8 triangles per edge (~64%)



› Effects of this research?

- › My goal was to *demonstrate social construction in modeling terms*
- › Not sure if that was successful...
- › Inspired (at least) two independent quantitative research projects, following up with the idea of platform effects

› Introduction

› Setting the stage

› Overarching STS themes

› Bias in geotagged tweets

› Platform effects on Facebook

› "Prediction" in machine learning

› Conclusion

› References

› Introduction

› Setting the stage

› Overarching STS themes

› Bias in geotagged tweets

› Platform effects on Facebook

› “Prediction” in machine learning

› Conclusion

› References

› “Prediction” in machine learning

> Imagination

> Introduction

> Setting the stage

> Overarching STS themes

> Bias in geotagged tweets

> Platform effects on Facebook

> "Prediction" in machine learning

> Conclusion

> References



> Prediction seems scary powerful

The image is a screenshot of a web page from MIT Technology Review. At the top left is the MIT Technology Review logo. At the top right are navigation links: 'Topics+', 'The Download', 'Magazine', and 'Events'. Below the navigation is a horizontal line. Underneath the line is the sub-header 'Intelligent Machines' and the main title 'Software Predicts Tomorrow's News by Analyzing Today's and Yesterday's'. Below the title is a short paragraph: 'Prototype software can give early warnings of disease or violence outbreaks by spotting clues in news reports.' Below that is the author and date: 'by Tom Simonite February 1, 2013'. At the bottom of the article preview is a sentence: 'A method of using online information to accurately predict the future could transform many industries.' The phrase 'accurately predict the future' is highlighted with a red box.

➤ Predict... the future?

➤ Introduction

➤ Setting the stage

➤ Overarching STS themes

Bias in geotagged tweets

➤ Platform effects on Facebook

➤ "Prediction" in machine learning

➤ Conclusion

➤ References

Predicting the Future With Social Media

Sitaram Asur
Social Computing Lab
HP Labs
Palo Alto, California
Email: sitaram.asur@hp.com

Bernardo A. Huberman
Social Computing Lab
HP Labs
Palo Alto, California
Email: bernardo.huberman@hp.com

Mar 2010

Abstract—In recent years, social media has become ubiquitous and important for social networking and content sharing. And yet, the content that is generated from these websites remains largely untapped. In this paper, we demonstrate how social media content can be used to predict real-world outcomes. In particular, we use the chatter from Twitter.com to forecast box-office revenues for movies. We show that a simple model built from

This paper reports on such a study. Specifically we consider the task of predicting box-office revenues for movies using the chatter from Twitter, one of the fastest growing social networks in the Internet. Twitter¹, a micro-blogging network, has experienced a burst of popularity in recent months leading to a huge user-base, consisting of several tens of millions of

Predicting the Future — Big Data, Machine Learning, and Clinical Medicine

Ziad Obermeyer, M.D., and Ezekiel J. Emanuel, M.D., Ph.D.

By now, it's almost old news: big data will transform medicine. It's essential to remember, however, that data by themselves are useless. To be useful, data must be analyzed, interpreted, and acted on. Thus, it is algorithms —

not data sets — that will prove transformative. We believe, therefore, that attention has to shift to new statistical tools from the field of machine learning that will be critical for anyone practicing medicine in the 21st century.

First, it's important to understand what machine learning is not. Most computer-based algorithms in medicine are "expert systems" — rule sets encoding knowledge on a given topic, which are applied to draw conclusions

1216

N ENGL J MED 375:13 NEJM.ORG SEPTEMBER 29, 2016

The New England Journal of Medicine
Downloaded from nejm.org at Harvard Library on November 8, 2018. For personal use only. No other uses without permission.
Copyright © 2016 Massachusetts Medical Society. All rights reserved.



predict verb

pre-dict | \pri-'dikt \

predicted; predicting; predicts

Definition of *predict*

transitive verb

: to declare or indicate **in advance**

especially : **foretell** on the basis of observation, experience, or scientific reason

intransitive verb

: to make a **prediction**

↓ Other Words from *predict*

↓ Synonyms

↓ Choose the Right Synonym

> “Prediction” is not prediction!

- > “*It’s not prediction at all!* I have not found a single paper predicting a future result. All of them claim that a prediction could have been made; i.e. they are *post-hoc* analysis and, needless to say, negative results are rare to find.” -Gayo-Avello, “I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper”, 2012

> Introduction

> Setting the stage

> Overarching STS themes

> Bias in geotagged tweets

> Platform effects on Facebook

> “Prediction” in machine learning

> Conclusion

> References

➤ “Prediction” is correlation

➤ Introduction

➤ Setting the stage

➤ Overarching STS themes

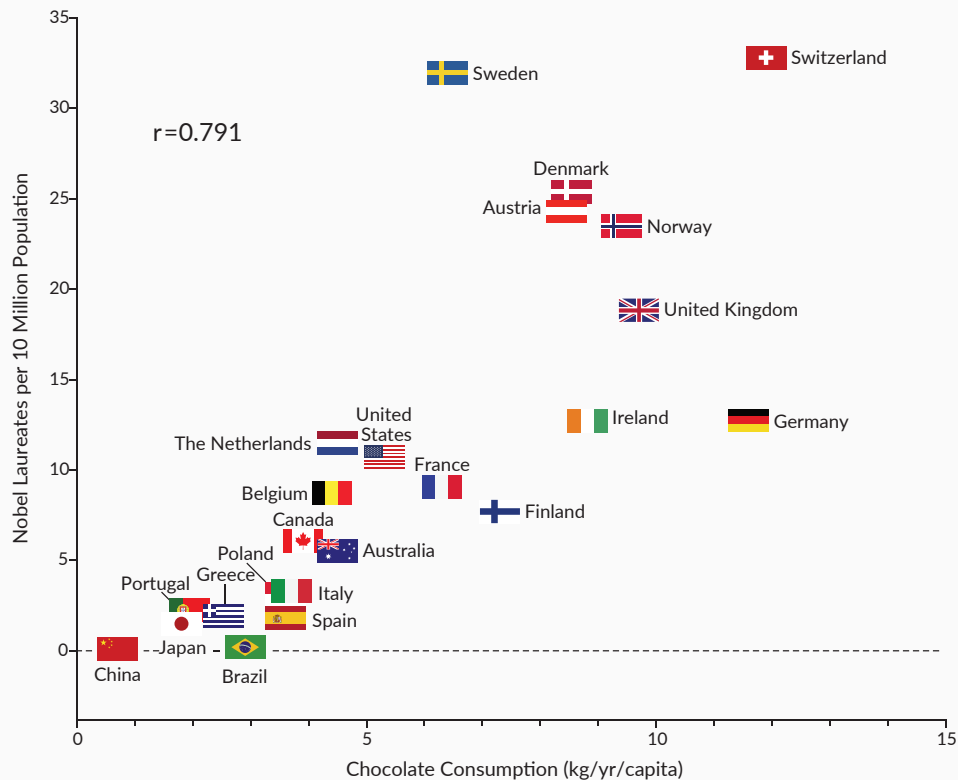
➤ Bias in geotagged tweets

➤ Platform effects on Facebook

➤ “Prediction” in machine learning

➤ Conclusion

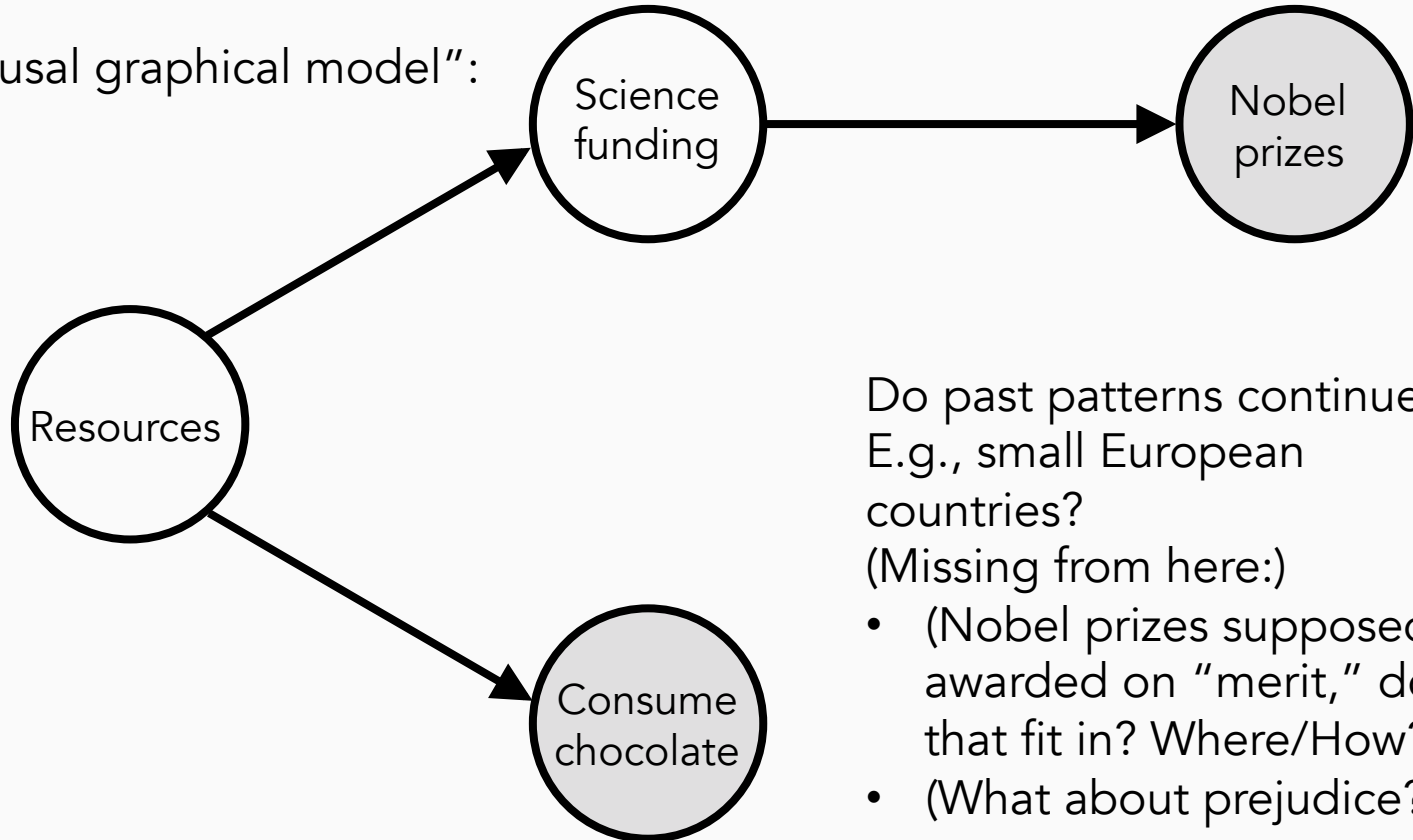
➤ References



Messerli, 2012

> Correlation is not causation

A "causal graphical model":



Do past patterns continue?
E.g., small European countries?

(Missing from here:)

- (Nobel prizes supposedly awarded on "merit," does that fit in? Where/How?)
- (What about prejudice?)

> Correlations can fail

- > Non-causal correlations can fit the data really well!
- > Google Flu Trends: half flu detector, half winter detector



➤ Not obvious usage of “predict”

➤ Introduction

➤ Setting the stage

➤ Overarching STS themes

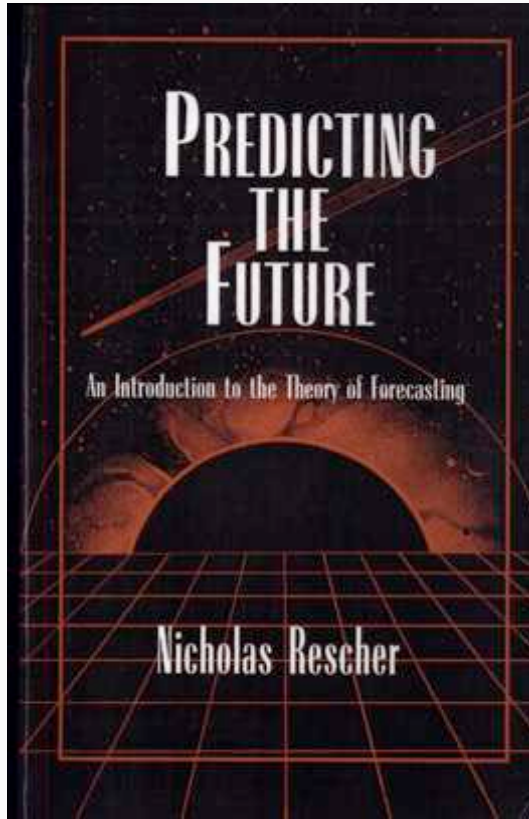
➤ Bias in geotagged tweets

➤ Platform effects on Facebook

➤ “Prediction” in machine learning

➤ Conclusion

➤ References



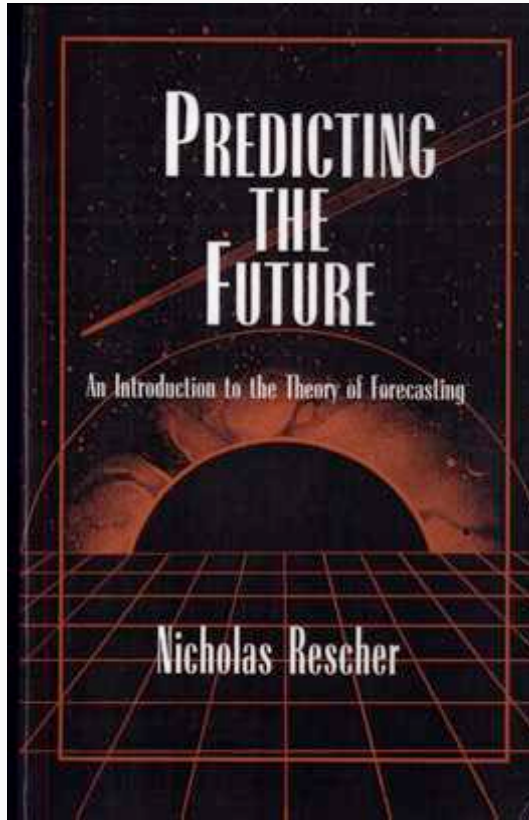
How STS can Improve Data Science

88 ■ PREDICTING THE FUTURE

TABLE 6.1: A SURVEY OF PREDICTIVE APPROACHES

Predictive Approaches	Linking Mechanism	Methodology Of Linkage
UNFORMALIZED/JUDGMENTAL		
judgmental estimation	expert informants	informed judgment
FORMALIZED/INFERENTIAL		
RUDIMENTARY (ELEMENTARY)		
trend projection	prevailing trends	projection of prevailing trends
curve fitting	geometric patterns	subsumption under an established pattern
circumstantial analogy	comparability groupings	assimilation to an analogous situation
SCIENTIFIC (SOPHISTICATED)		
indicator coordination	causal correlations	statistical subsumption into a correlation
law derivation (nomic)	accepted laws (deterministic or statistical)	inference from accepted laws
phenomenological modeling (analogical)	formal models (physical or mathematical)	analogizing of actual (“real-world”) processes with presumably isomorphic model process

› But has rhetorical power



“In all times and places, decision makers have looked to predictive counselors of some sort—putative experts, be they religious or secular, to guide them regarding the auguries of the gods, the stars, or the inexorable decrees of fate or of nature.”

› Introduction

› Setting the stage

› Overarching STS themes

› Bias in geotagged tweets

› Platform effects on Facebook

› “Prediction” in machine learning

› Conclusion

› References

› Leveraging inconsistencies

- › The expectation of Mean Squared Error (MSE) can be *decomposed* into three terms: the irreducible error, the square of the amount by which a model misses the “truth”, and the noisiness of the model
- › Decreasing the noisiness of the model, if greater than the amount by which it departs from the “truth”, can improve prediction
- › We can simulate a “toy” example of this
 - The “truth” is a model we use to generate data. But when making predictions, leaving out noisy causal inputs (“false” models) can make better predictions than does using the “true” model! (Shmueli, 2010)

› Introduction

› Setting the stage

› Overarching STS themes

› Bias in geotagged tweets

› Platform effects on Facebook

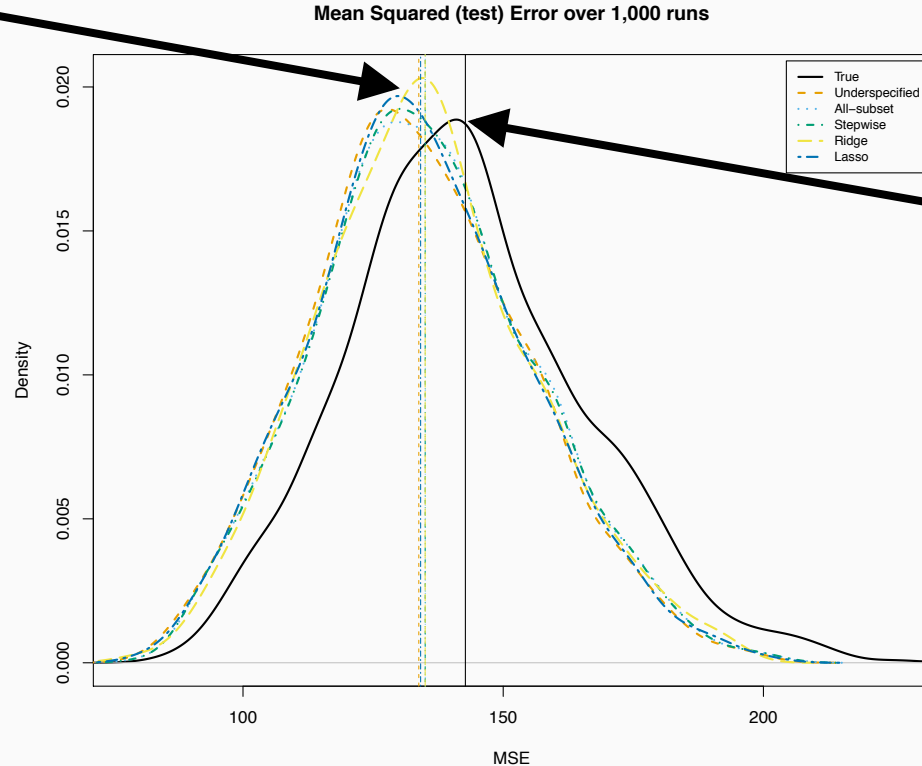
› “Prediction” in machine learning

› Conclusion

› References

➤ A 'true' model predicting worse

Error, over thousands of simulations, of correlation-optimizing models that end up leaving out noisy (but still causal) variables



Error, over thousands of simulations, of the model that generated the data, fit back to the data

➤ Introduction

➤ Setting the stage

➤ Overarching STS themes

➤ Bias in geotagged tweets

➤ Platform effects on Facebook

➤ "Prediction" in machine learning

➤ Conclusion

➤ References

> Dependencies also matter

- > Machine learning uses cross-validation (splitting data, fitting a model on the “training” set and reporting performance on recovering the signal in held-out “test” set) to judge performance
- > If data points are not independent (e.g., in a time series, observations will not depart too far from previous values; or in a social network, people’s outcomes are related to that of their network neighbors), then splitting data into training and test may not work
- > Test error will be a better reflection of general performance than training error, but can still *vastly* underestimate generalizability
- > I demonstrate over 10,000 simulations from a multivariate normal distribution, where dimensions have a correlation of 0.5

> Introduction

> Setting the stage

> Overarching STS themes

> Bias in geotagged tweets

> Platform effects on Facebook

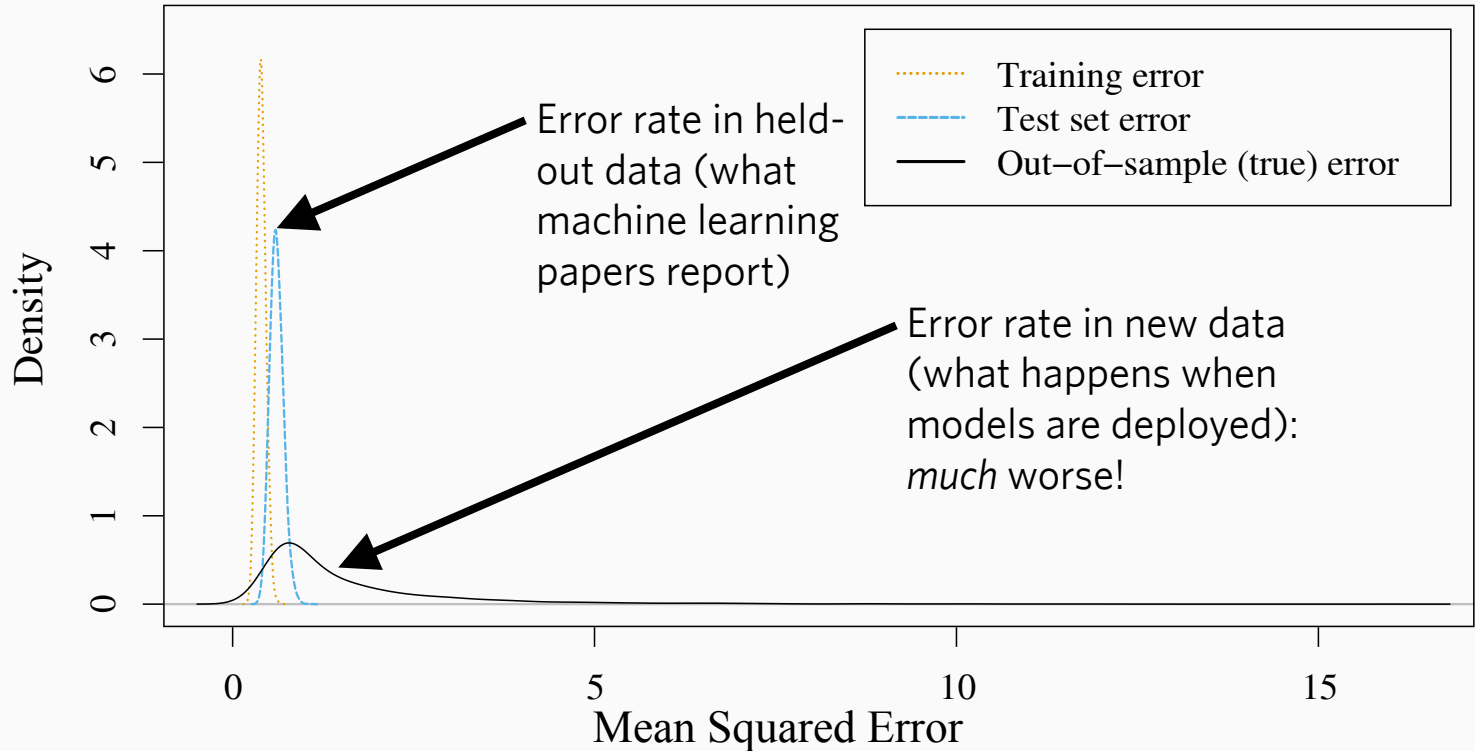
> “Prediction” in machine learning

> Conclusion

> References

> Dependencies also matter

Distribution of error across simulations



> Introduction

> Setting the stage

> Overarching STS themes

> Bias in geotagged tweets

> Platform effects on Facebook

> "Prediction" in machine learning

> Conclusion

> References

» Introduction

» Setting the
stage

» Overarching
STS themes

» Bias in
geotagged
tweets

» Platform
effects on
Facebook

» "Prediction"
in machine
learning

» Conclusion

» References

» Conclusion

› Data science is powerful

- › ...and currently wielded by existing structures of power.
- › Power comes not from correspondence to “reality” or “truth,” but from a complex web of interrelationships
- › Find out what those relationships are, find inconsistencies, articulate those consistencies in quantitative ways

› Introduction

› Setting the stage

› Overarching STS themes

› Bias in geotagged tweets

› Platform effects on Facebook

› “Prediction” in machine learning

› Conclusion

› References

> Thank you!

> Introduction

> Setting the stage

> Overarching STS themes

> Bias in geotagged tweets

> Platform effects on Facebook

> "Prediction" in machine learning

> Conclusion

> References



References

Introduction

Setting the stage

Overarching STS themes

Bias in geotagged tweets

Platform effects on Facebook

"Prediction" in machine learning

Conclusion

References

- Breiman, Leo. "Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author)." *Statistical Science* 16, no. 3 (2001): 199–231. <https://dx.doi.org/10.1214/ss/1009213726>.
- Efron, Bradley, and Carl Morris. "Stein's Paradox in Statistics." *Scientific American* 236, no. 5 (1977): 119–127. <https://dx.doi.org/10.1038/scientificamerican0577-119>.
- Gayo-Avello, Daniel. "'I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper': A Balanced Survey on Election Prediction using Twitter Data." 2012. <https://arxiv.org/abs/1204.6441>.
- Healy, Kieran. "The Performativity of Networks." *European Journal of Sociology* 56, no. 2 (2015): 175–205. <https://dx.doi.org/10.1017/S0003975615000107>.
- Jasanoff, Sheila, and Sang-Hyun Kim, eds. *Dreamscapes of Modernity: Sociotechnical Imaginaries and the Fabrication of Power*. Chicago, IL: The University of Chicago Press, 2015.
- Jones, Matthew L. "How We Became Instrumentalists (Again): Data Positivism since World War II." *Historical Studies in the Natural Sciences* 48, no. 5 (2018): 673–684. <https://dx.doi.org/10.1525/hsns.2018.48.5.673>.
- Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. "The Parable of Google Flu: Traps in Big Data Analysis." *Science* 343, no. 6176 (2014): 1203–1205. <https://dx.doi.org/10.1126/science.1248506>.
- Lazer, David, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. "Computational Social Science". *Science* 323, no. 5915 (2009): 721–723. <https://dx.doi.org/10.1126/science.1167742>.

- Malik, Momin M., and Jürgen Pfeffer. "Identifying Platform Effects in Social Media Data." In *Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM-16)*, 241–249. AAAI Press, 2016. Updated version at http://mominmalik.com/malik_chapter2.pdf.
- Malik, Momin M., Hemank Lamba, Constantine Nakos, and Jürgen Pfeffer. "Population Bias in Geotagged Tweets." In *Papers from the 2015 ICWSM Workshop on Standards and Practices in Large-Scale Social Media Research (ICWSM-15 SPSM)*, 18–27. AAAI Press, 2015. Updated version at http://mominmalik.com/malik_chapter1.pdf.
- Messerli, Franz H. "Chocolate Consumption, Cognitive Function, and Nobel Laureates." *The New England Journal of Medicine*, 367 (2012): 1562–1564. <https://dx.doi.org/10.1056/NEJMon1211064>.
- Mullainathan, Sendhil, and Jann Spiess. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives* 31, no. 2 (2017): 87–106. <https://dx.doi.org/10.1257/jep.31.2.87>.
- Rescher, Nicholas. *Predicting the Future: An Introduction to the Theory of Forecasting*. State University of New York Press, 1998.
- Shelton, Taylor, Ate Poorthuis, Mark Graham, and Matthew Zook. "Mapping the Data Shadows of Hurricane Sandy: Uncovering the Sociospatial Dimensions of 'Big Data'." *Geoforum* 52 (2014): 167–179. <http://dx.doi.org/10.1016/j.geoforum.2014.01.006>.
- Shmueli, Galit. "To Explain or to Predict?" *Statistical Science* 25, no. 3 (2010): 289–310. <https://dx.doi.org/10.1214/10-STS330>.
- Szekely, Francisc. "Unreliable Observers, Flawed Instruments, 'Disciplined Viewings': Handling Specimens in Early Modern Microscopy." *Parergon* 28, no. 1 (2011): 155–176. <https://dx.doi.org/10.1353/pgn.2011.0032>.
- van Dijck, José. *The Culture of Connectivity: A Critical History of Social Media*. New York, NY: Oxford University Press, 2013.