JOHNS HOPKINS
CAREY BUSINESS SCHOOL

Data Science and Business Intelligence

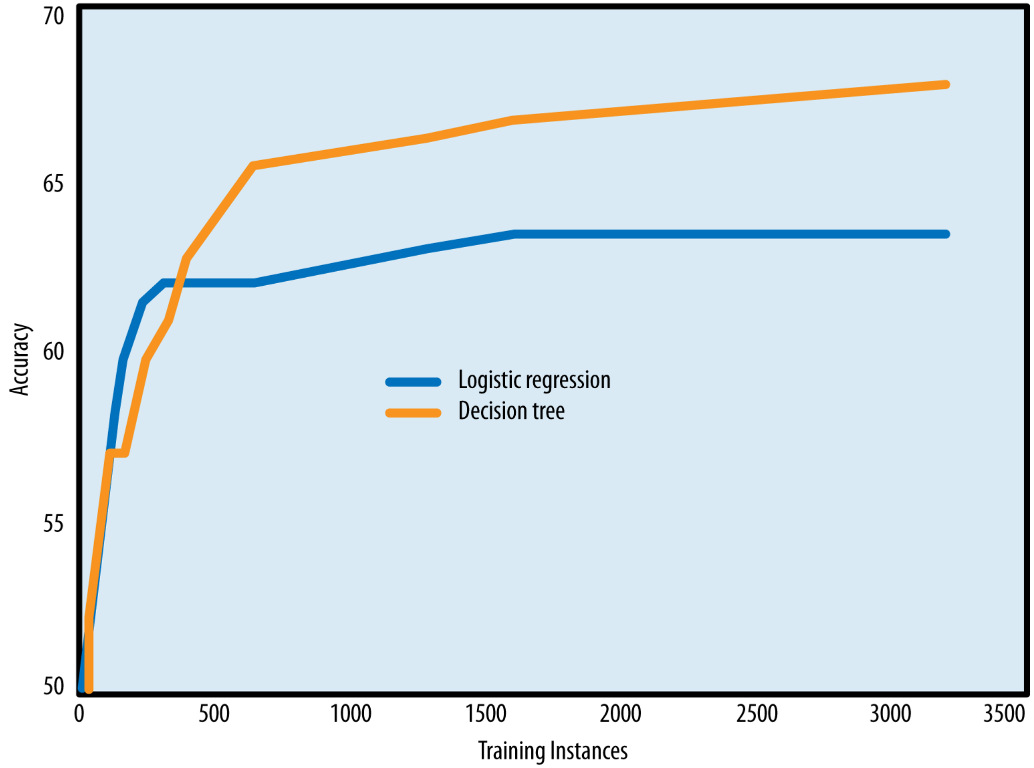
BU.330.780

Session 7

Instructor: Changmi Jung, Ph.D.

Data Driven Decision Making

How big is big? – Do you really need 'big' data?



- Learning curve is a plot of the **generalization performance** (test data) against **the amount of training data**.
- Learning curve may give recommendations on **how much to invest in** training data.

Wider is better, and Diversity matters!
But the value of additional data diminishes.
How can big data help?

Data Stocks vs. Flows

nature International weekly journal of science

Access

To read this story in full you will need to login or make a payment (see right).

Letter

Nature **457**, 1012-1014 (19 February 2009) | doi:10.1038/nature07634; Received 14 August 2008; Accepted 13 November 2008; Published online 19 November 2008; Corrected 19 February 2009

Detecting influenza epidemics using search engine query data

Jeremy Ginsberg[1], Matthew H. Mohebbi[1], Rajan S. Patel[1], Lynnette Brammer[2], Mark S. Smolinski[1] & Larry Brilliant[1]

1. Google Inc., 1600 Amphitheatre Parkway, Mountain View, California 94043, USA
2. Centers for Disease Control and Prevention, 1600 Clifton Road, NE, Atlanta, Georgia 30333, USA

Correspondence to: Matthew H. Mohebbi[1] Correspondence and requests for materials should be addressed to J.G. or M.H.M. (Email: flutrends-

ARTICLE LINKS

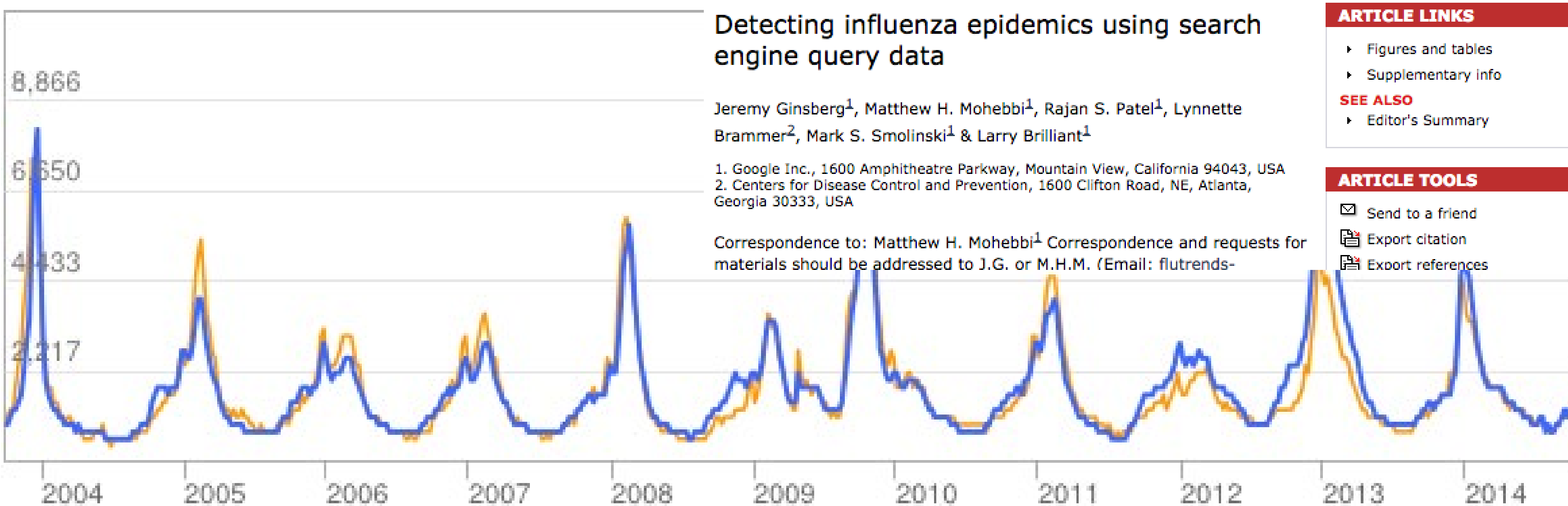
- Figures and tables
- Supplementary info

SEE ALSO

- Editor's Summary

ARTICLE TOOLS

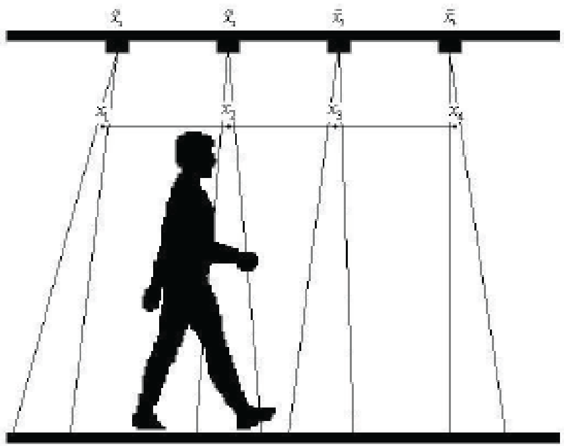
- Send to a friend
- Export citation
- Export references



Source: http://www.google.org/flutrends/about/how.html

Example: Gait Study

Gait impairment is an important indicator/predictor of cognitive and physical function in the elderly. However, the clinical assessment is inconsistent and episodic.



- Unobtrusive sensors or computer vision collect gait data
- Gait velocity and other gait metrics are captured from the gait videos (auto-convert the data into the metrics), which will then feed into the deep learning algorithm that predicts which patients would respond to the shunt surgery.

Cleveland Clinic & IBM are working on predicting NPH (Normal Pressure Hydrocephalus) by using Gait impairment as an indicator in a Deep Learning model [(March 2025)]

Is Data Science A Magic Bullet?

- Any inherent limitations?
- Any risks?
- Any complements?

Let's talk about the dark side!

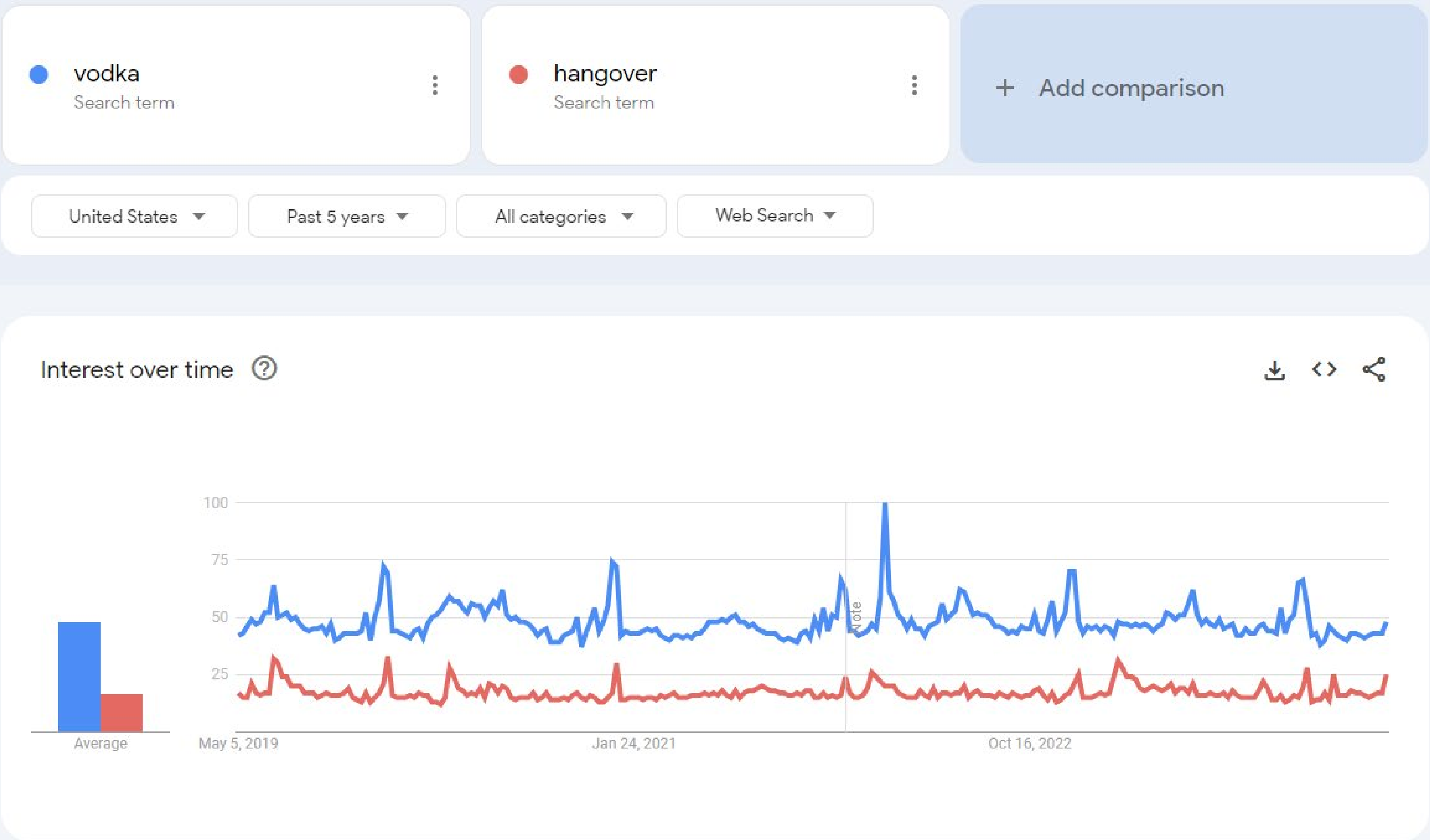
Well, Hal Varian's video first....

Google Trends/Correlates

https://www.youtube.com/watch?v=-I8acYHQ0v0&t=11s

Google Trends

- Google Trends: https://trends.google.com/trends/



Understand limitations (1): Correlation vs. Causation

- Correlation: Variations in one quantity tell us something about variations in the other

- Mathematically, there are several measures for correlation.

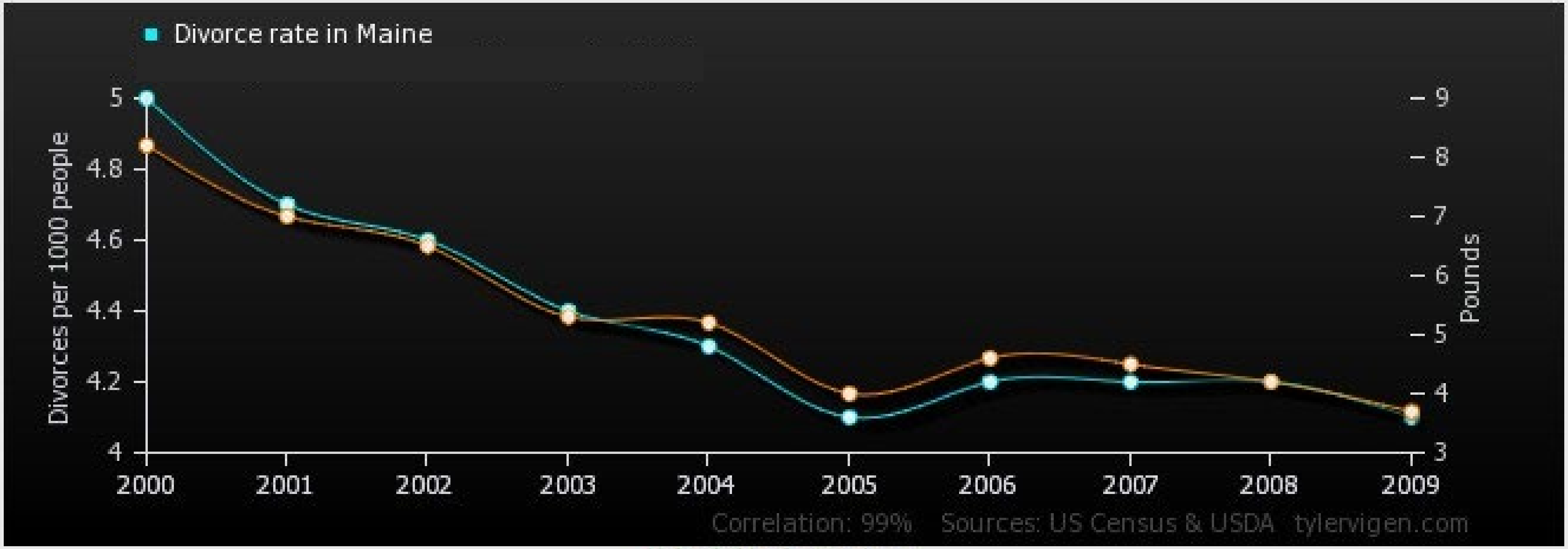
- One of the most widely used one is Pearson's correlation coefficient

$$\rho_{X,Y} = corr(X, Y)$$

$$= \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

Divorce rate in Maine

correlates with



Upload this chart to imgur

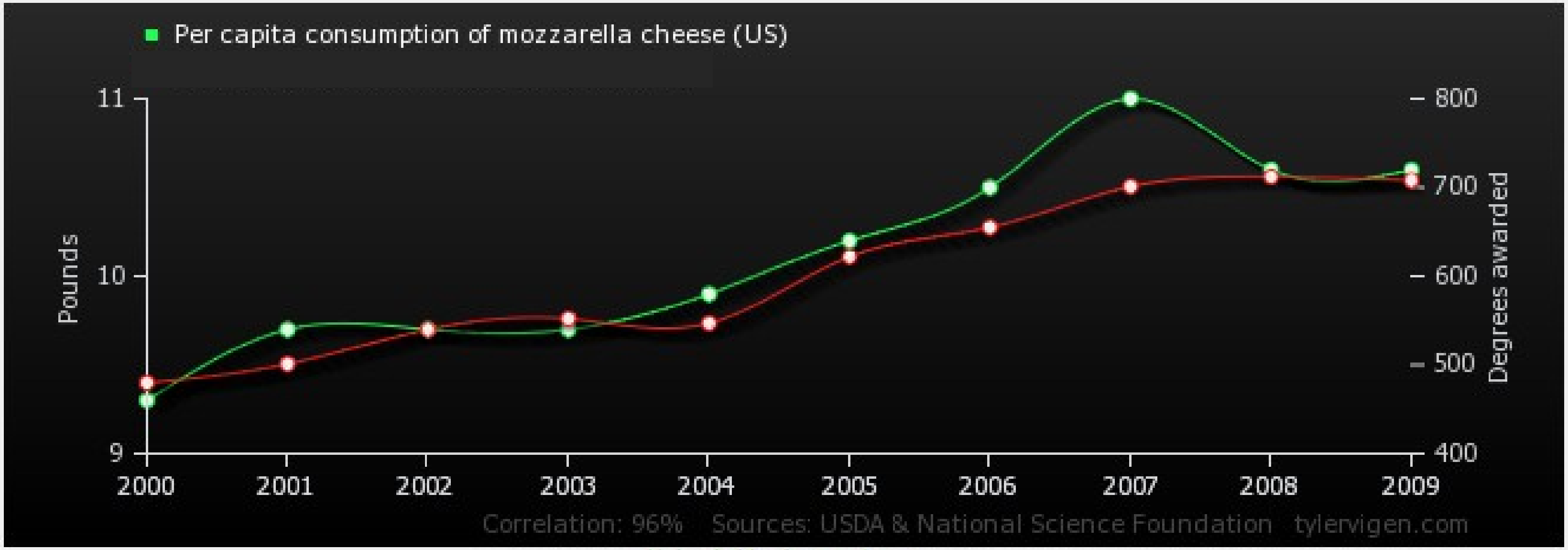
	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Divorce rate in Maine *Divorces per 1000 people (US Census)*	5	4.7	4.6	4.4	4.3	4.1	4.2	4.2	4.2	4.1
	8.2	7	6.5	5.3	5.2	4	4.6	4.5	4.2	3.7

Correlation: 0.992558

Source: http://news.distractify.com/dark/trivial-facts/these-hilarious-graphs-show-unexpected-correlations-between-seemingly-unrelated-statistics/

Per capita consumption of mozzarella cheese (US)

correlates with



Upload this chart to imgur

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Per capita consumption of mozzarella cheese (US) Pounds (USDA)	9.3	9.7	9.7	9.7	9.9	10.2	10.5	11	10.6	10.6
	480	501	540	552	547	622	655	701	712	708

Correlation: 0.958648

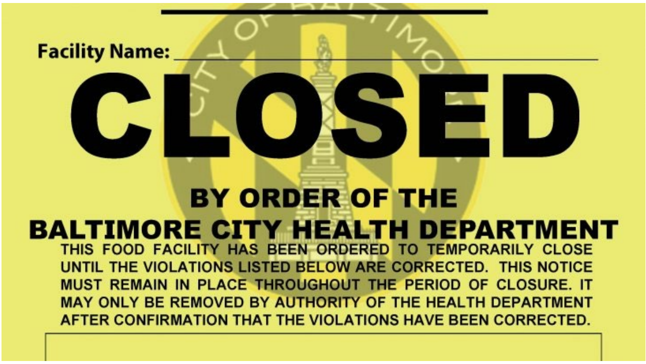
Source: http://news.distractify.com/dark/trivial-facts/these-hilarious-graphs-show-unexpected-correlations-between-seemingly-unrelated-statistics/

Understand limitations (2): Use Context

- Netflix competition example
  - DVD vs. Online streaming

The last DVDs were sent out on Sep. 29th, 2023

- Health-code violation algorithm developed using data from Boston
  - Will it be effective in Orlando?

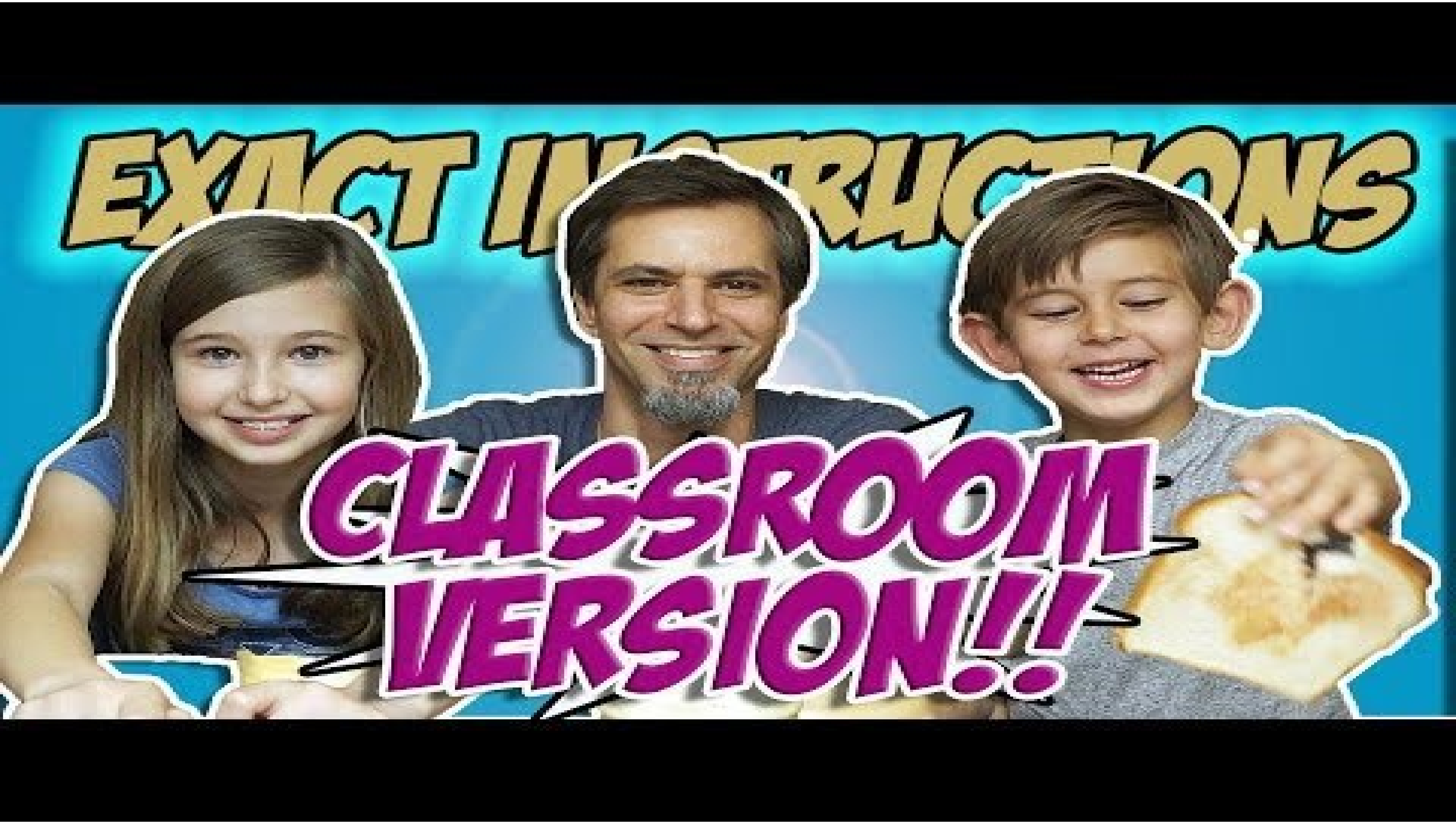


Meet the New Boss: Intuition-based vs. data-driven (evidence-based) decision making

- Traditionally, hiring decisions have been made mostly based on intuition (or HIPPO: Highest Paid Person's Opinion).

- Assume that you propose a plan to implement a “data-driven” hiring process.
  - Discuss a data-driven approach you may suggest. Assume that you have unlimited resources which can be used for collecting any sort of data, hiring data scientists, etc.

Yes, We will definitely have more objective results!!??
Really?



https://www.youtube.com/watch?v=FN2RM-CHkuI

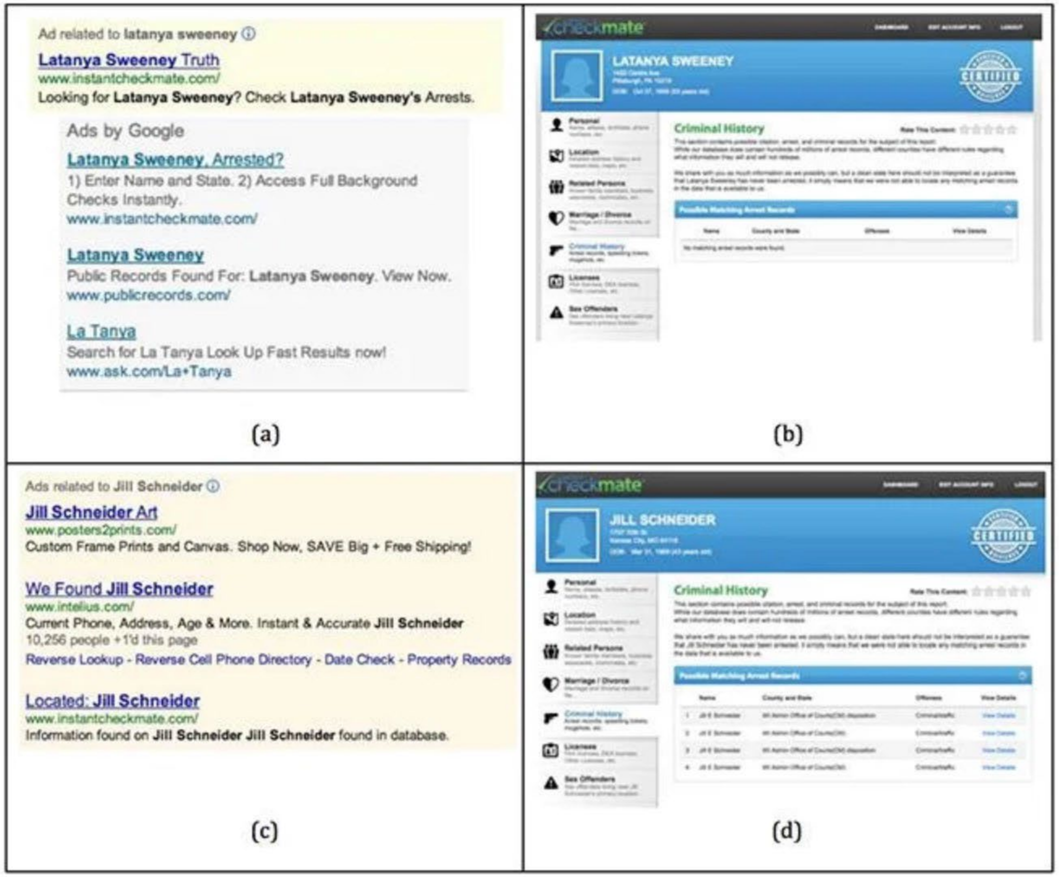
Does Algorithm Really Eliminate Discrimination?

- How can it discriminate?

- Algorithm can ”learn” to discriminate
  - Algorithms learn from human behavior, so they reflect the biases we hold.
  - A model is built on “historical” data – “historical” biases in the training data will be learned by the algorithm

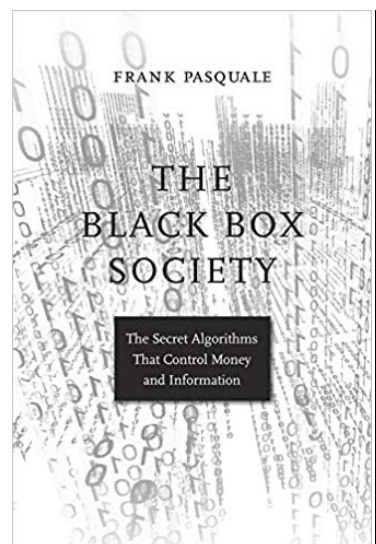
- Example: ad-targeting algorithms
  - High-paying jobs to men but not women
  - Ads for high-interest loans to people in low-income neighborhoods

The Bias



In one study, black-identified names generated different ads than white-identified ones.

Chart courtesy Latanya Sweeney/Harvard University (http://arxiv.org/ftp/arxiv/papers/1301/1301.6822.pdf)



Was Facebook Really Biased?

http://pubsonline.informs.org/journal/mnsc/

MANAGEMENT SCIENCE
Vol. 65, No. 7, July 2019, pp. 2966–2981
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads

Anja Lambrecht,[a] Catherine Tucker[b]

[a] Marketing, London Business School, London NW1 4SA, United Kingdom; [b] Marketing, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142

Contact: alambrecht@london.edu, http://orcid.org/0000-0001-6766-1602 (AL); cetucker@mit.edu, http://orcid.org/0000-0002-1847-4832 (CT)

Received: November 28, 2017
Revised: March 2, 2018
Accepted: March 13, 2018
Published Online in Articles in Advance: April 10, 2019

https://doi.org/10.1287/mnsc.2018.3093

Abstract. We explore data from a field test of how an algorithm delivered ads promoting job opportunities in the science, technology, engineering and math fields. This ad was explicitly intended to be gender neutral in its delivery. Empirically, however, fewer women saw the ad than men. This happened because younger women are a prized demographic and are more expensive to show ads to. An algorithm that simply optimizes cost-effectiveness in ad delivery will deliver ads that were intended to be gender neutral in an apparently discriminatory way, because of crowding out. We show that this empirical regularity extends to other major digital platforms.

History: Accepted by Joshua Gans, business strategy.
Funding: Supported by a National Science Foundation Career Award [Grant 6923256].

Keywords: algorithmic bias • online advertising • algorithms • artificial intelligence

How to Make Algorithm Fairer?

- Assess biases in training data
- More human involvement?
- Who should be treated similarly to whom?
- Who is responsible?

Risks and a mitigation plan should be addressed!

Customer Data and Privacy

Target's predictive model

- Tough to change consumers' shopping habits
- There are some brief period in a person's life when his/her shopping habits become particularly flexible.
- How much does Target know about you?
- Personalized promotion vs. Privacy invasion
- Did target invade customer's privacy? What went wrong?

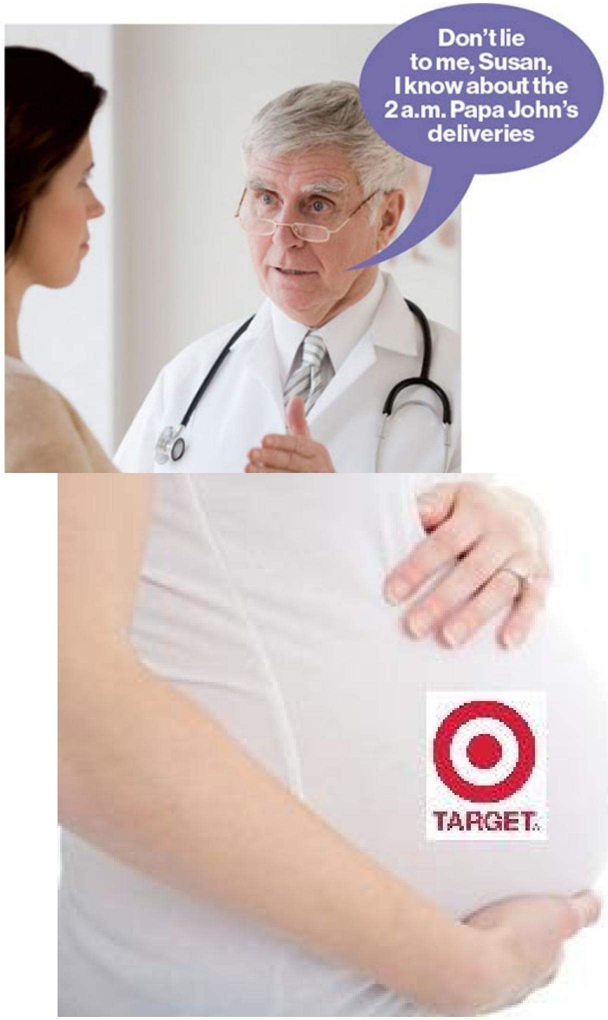


Image from http://www.forbes.com/sites/kashmirhill/2012/02/16/could-target-sell-its-pregnancy-prediction-score/

What Do We Mean by Deployment?

Making the predictive model available for use

- Embedding the model in a web application
- Connecting it to an Enterprise system
- Deliver predictions via API
- Running the job in batch
- What's the business purpose and implication?

Deployment Mode

Deployment Plan

1. Model integration (package it for call, build interface, etc.)
2. Environment setting – cloud vs. on-premise (or hybrid), real time vs. batch?
3. Capture errors and latency, and track model drift – **create a feedback loop**

- Feasible timeline
- Human involvement

*Predictive Models live in **dynamic environments***

Key Issues in Deployment

Category	Potential Issue	Mitigation Strategy
Data Pipeline	Input data format/structure changes	Validate input schema; add input checks
Concept Drift	Model degrades as data patterns change	Schedule regular monitoring & retraining
Scalability	High latency or slow predictions under load	Use caching, load balancing, model optimization
Interpretability	Users distrust predictions without explanation	Provide explainability tools (e.g., SHAP, LIME)
Integration	Model outputs incompatible with downstream systems	Involve IT/dev early; test integration endpoints
Security & Privacy	Sensitive data leakage or unauthorized access	Access controls, encryption, compliance checks
Maintenance	No one owns the model post-deployment	Assign clear owner/team for monitoring & updates

Implications for Your Group Project

- Is your dataset enough? What other information would have made your model work better? Are those achievable if given a longer timeline?
- Is it a causal relationship or a simple correlation? Don't be tempted to interpret your results as "causal".
- Be sure to discuss the limitations of your model or overall project.
- What are the specific business decisions your project suggests? Are there any other soft goals to be considered?
- Plan your feedback loop!
- Address what business implications you can make and potential risks in deployment, as well (integration with the current workflow, how to mitigate potential discrimination, etc.).

Next week: Group project presentation!

- Each team will have up to **18 minutes** for the presentation, including Q&A.
- Your presentations will be recorded for our TAs to review together, but will not be shared with anyone else.
- Every team member must participate in the presentation: each member should present at least some portion of the project.
- Your grade will be based on the final presentation (20%), the report (80%), and peer evaluation. The peer evaluation sheet will become available next week.
- Deliverables: Presentation slide deck (or link), Report, and R codes (in the RMD file), too.
  - A compiled (knitted) RMD file is highly recommended, but you may submit the RMD file itself.

What do I look for in Your Presentation?

- Clarity – clearly explain your problem (why the problem is important) and how you arrived at the choice of the model, etc. Assume that your audience doesn't have a DS background.

- Provide support materials (visualization or other credible resources) to justify your rationale if there are any important assumptions you made.

- Correctness and Research, whenever needed – learn your project domain and cite any resources you referred to.

- Keep track of the time and don't go over the given time.

- Attitude – you must be excited about sharing your work, right? Be confident, and please **do not read from scripts**.