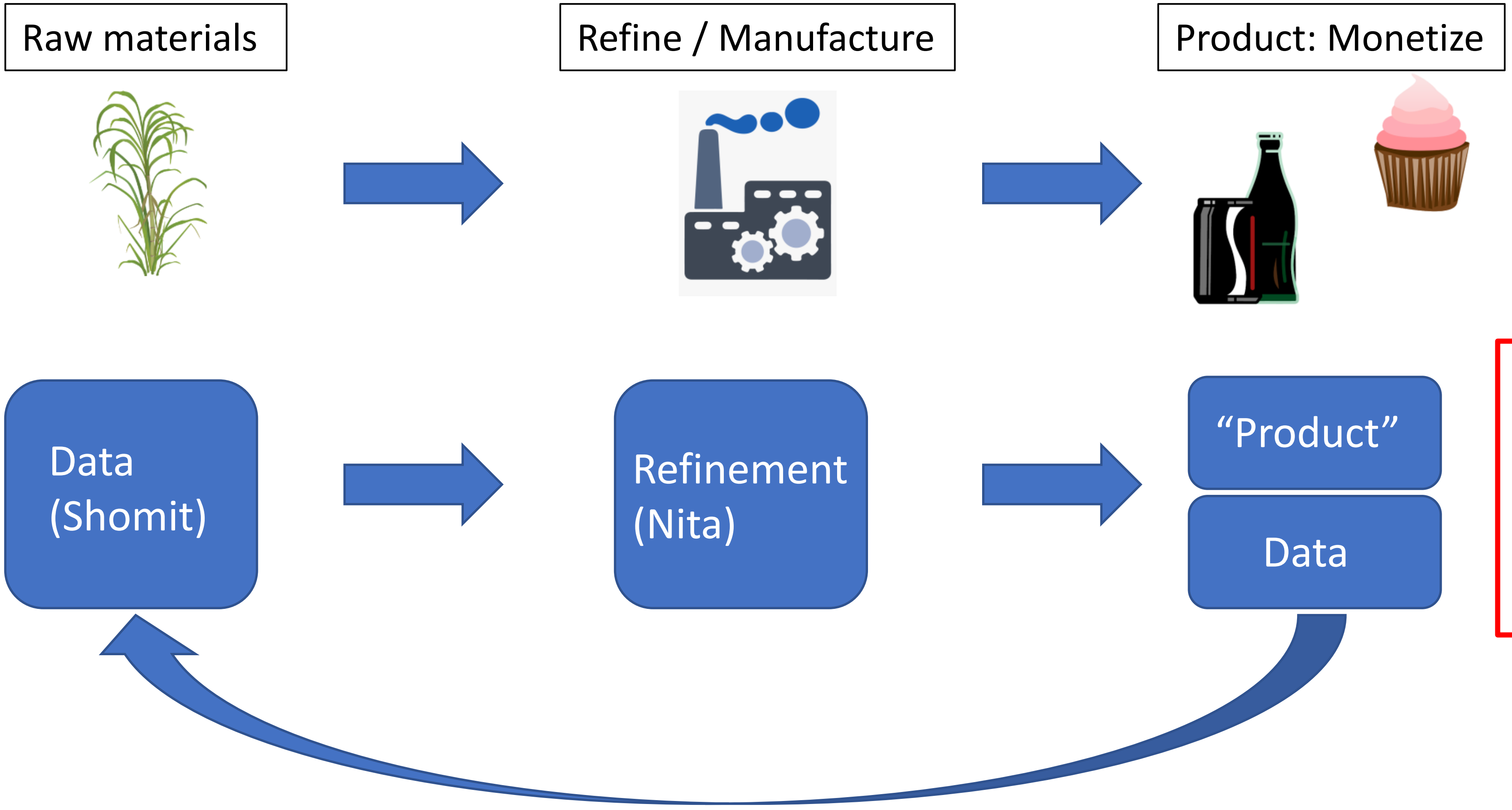
Multi-dimensional Data

What is it, and how do I get it?

Shomit Ghose – shomit@berkeley.edu – Oct 19, 2020

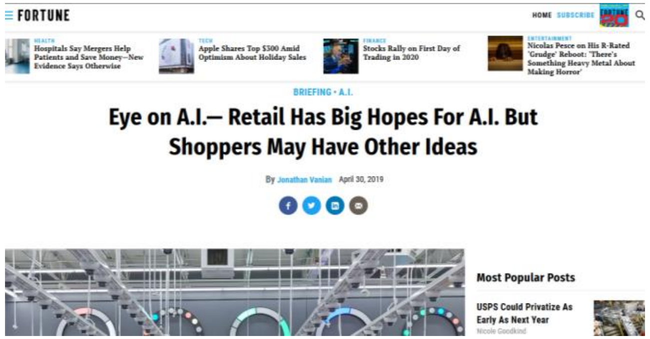
Canonical Business Model: Updated



No matter what business you’re in you need to do this ☺

But Microsoft's Spencer says he doesn't consider Sony and Nintendo his main competition anymore, largely because neither of those Japanese companies owns its own top-end global cloud infrastructure akin to Microsoft's Azure platform. One of Microsoft's main selling points for the new Xbox will be integration with its xCloud technology, which is meant to allow you to play the same game across a console, a desktop PC and a mobile device.

"When you talk about Nintendo and Sony, we have a ton of respect for them, but we see Amazon and Google as the main competitors going forward," Spencer said. "That's not to disrespect Nintendo and Sony, but the traditional gaming companies are somewhat out of position. I guess they could try to re-create Azure, but we've invested tens of billions of


FORTUNE
HOME SUBSCRIBE

HEALTH
Hospitals Say Mergers Help Patients and Save Money—New Evidence Says Otherwise

TECH
Apple Shares Top $300 Amid Optimism About Holiday Sales

FINANCE
Stocks Rally on First Day of Trading in 2020

ENTERTAINMENT
Nicolas Pesce on His R-Rated 'Grudge' Reboot: 'There's Something Heavy Metal About Making Horror'

BRIEFING • A.I.

Eye on A.I.— Retail Has Big Hopes For A.I. But Shoppers May Have Other Ideas

By Jonathan Vanian April 30, 2019

Most Popular Posts

USPS Could Privatize As Early As Next Year
Nicole Goodkind

Big Data is a Well-Understood Topic. What Do Amazon and Google Understand That Others Do Not?

Multi-dimensional Data – Rule 1

If data is good, more data is better

Data is an **anti-commodity**: the more you have the more it's worth

Multi-dimensional Data – Rule 2

Everything is a data signal for everything

Predicting from Multi-dimensional Data

Coming apart? Cultural distances in the United States over time

Marianne Bertrand and Emir Kamenica*

December 2018

Abstract

We analyze temporal trends in cultural distance between groups in the US defined by income, education, gender, race, and political ideology. We measure cultural distance between two groups as the ability to infer an individual's group based on his or her (i) media consumption, (ii) consumer behavior, (iii) time use, or (iv) social attitudes. Gender difference in time use

Multi-dimensionality: What Do the Following Sets of Attributes Tell Us?

High income

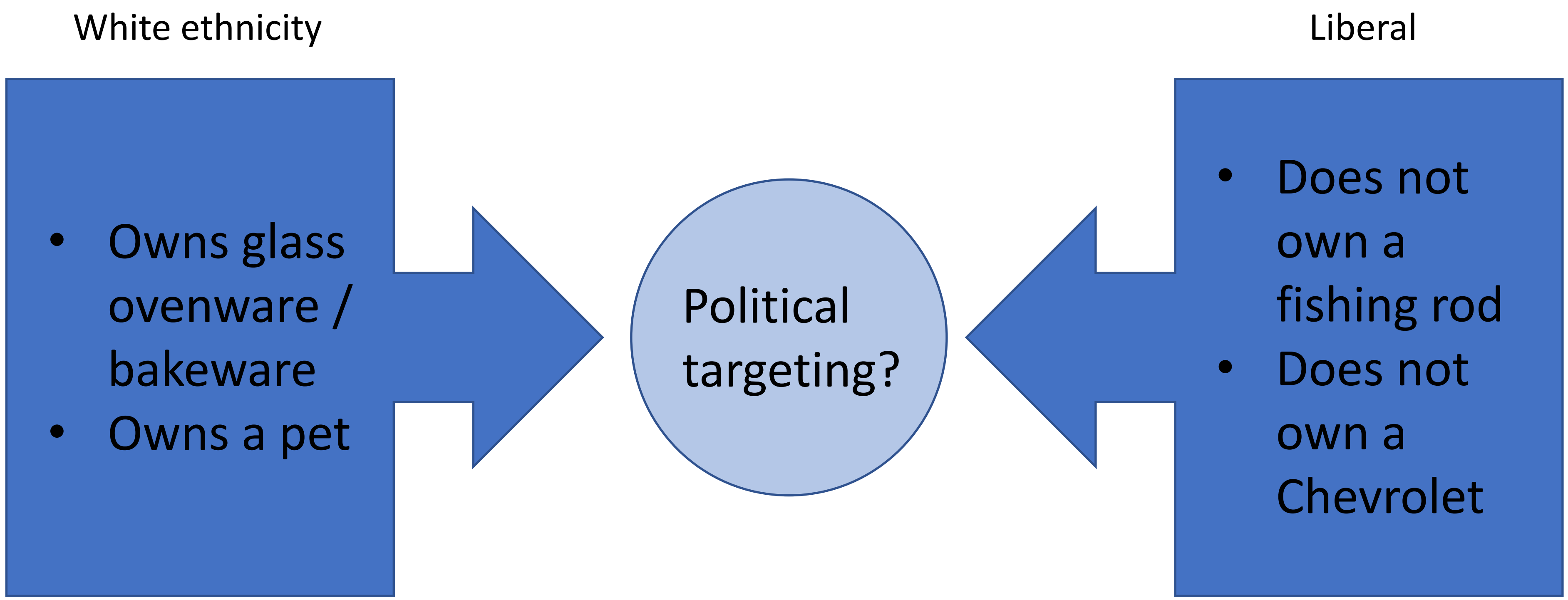
- Uses dishwasher detergent
- Voted in the election
- Uses Netflix
- Verizon Wireless customer
- Has a passport

Good loan risk?

High education

- Owns an iPhone
- Traveled in continental US
- Watched Super Bowl
- Trusts people
- Not afraid to walk at night

Multi-dimensionality: What Do the Following Sets of Attributes Tell Us?



Some **Real-life** Multi-dimensional Signals: Loan Risk

- Email address (eponymous?)
- Cell battery: strength, charging patterns
- Loan time-stamp
- Phone: contacts' responsiveness
- Errors in filling Web forms: spelling, (letter) case
- Typing speed

Frankly, We Do Give a Damn: The Relationship Between Profanity and Honesty

Social Psychological and
Personality Science
2017, Vol. 8(7) 816-826

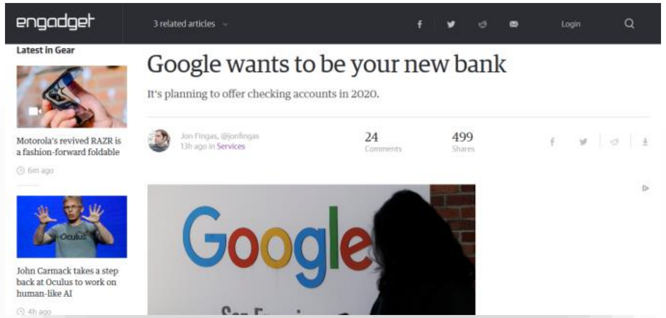
DOI: 10.1177/1948550616681055
journals.sagepub.com/home/spp
SAGE

Gilad Feldman[1], Huiwen Lian[2], Michal Kosinski[3], and David Stillwell[4]

Abstract

There are two conflicting perspectives regarding the relationship between profanity and dishonesty. These two forms of norm-violating behavior share common causes and are often considered to be positively related. On the other hand, however, profanity is often used to express one's genuine feelings and could therefore be negatively related to dishonesty. In three studies, we explored the relationship between profanity and honesty. We examined profanity and honesty first with profanity behavior and lying on a scale in the lab (Study 1; $N = 276$), then with a linguistic analysis of real-life social interactions on Facebook (Study 2; $N = 73,789$), and finally with profanity and integrity indexes for the aggregate level of U.S. states (Study 3; $N = 50$ states). We found a consistent positive relationship between profanity and honesty; profanity was associated with less lying and deception at the individual level and with higher integrity at the society level.

Multi-dimensional Data in Action?


engadget
3 related articles
Login
Latest in Gear
Google wants to be your new bank
It's planning to offer checking accounts in 2020.
Jon Fingas, @jonfingas
13h ago in Services
24 Comments
499 Shares
Motorola's revived RAZR is a fashion-forward foldable
6m ago
John Carmack takes a step back at Oculus to work on human-like AI
4h ago

Multi-dimensional Data in Action?

Multi-dimensional Data

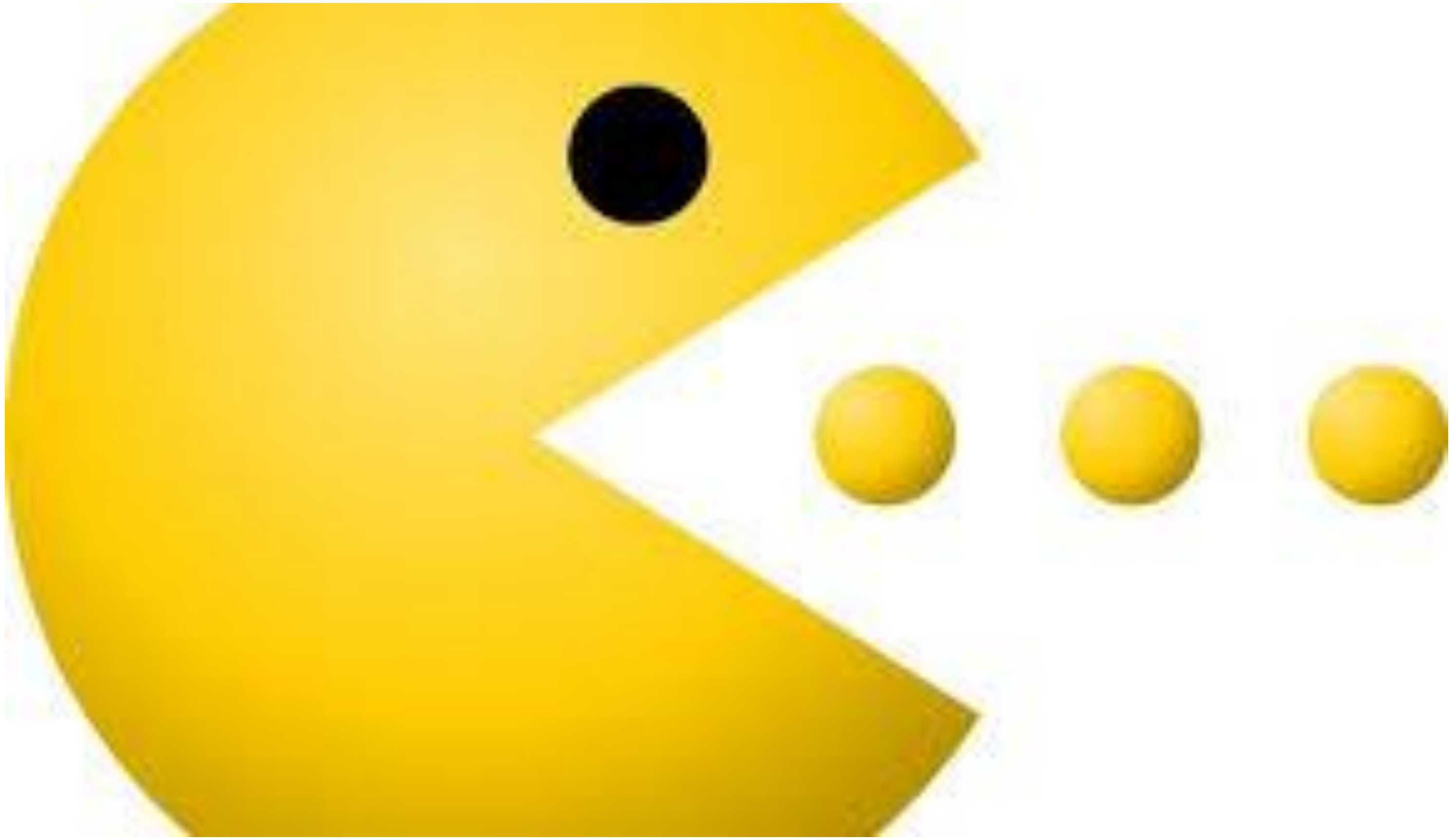
Some (Shomit) definitions

- “Data richness”: your own sources (n → ∞)
- “Data hungriness”: from outside sources (n → ∞)

Multi-dimensional Data Business Model

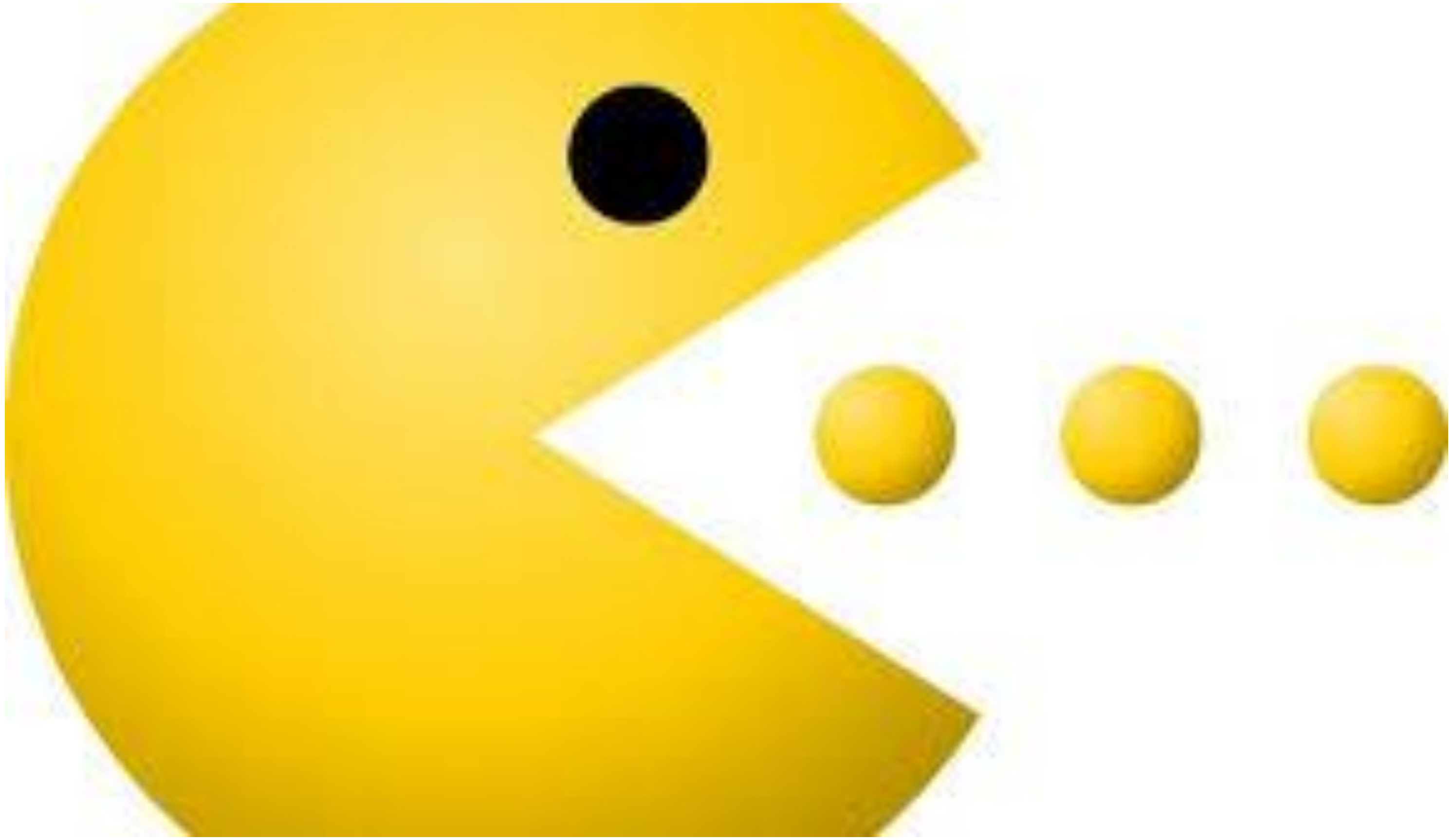
- Must first understand market opportunity and business model
- This will define what data is needed
  - Proprietary sources are essential
- Don’t start with the data *and then* look for the market opportunity and business model
  - A common misstep!
- No merit in just being a (commodity) data broker

Data-Richness: Proprietary Data Sources



- The “n+1” model
- Consider Facebook adding the “Like” button
- Or Amazon’s purchase of Whole Foods
  - An offline shopping signal that complements their existing online signal
- Or Google’s expansion from search, to email, to maps, to self-driving cars…
  - … to their purchase of Nest and Fitbit

Data-Richness: Proprietary Data Sources



- An infinite appetite for ever more dimensions
  - “n+1”
- The higher the dimensionality, the more defensible your business
- Data richness IS part of your product roadmap

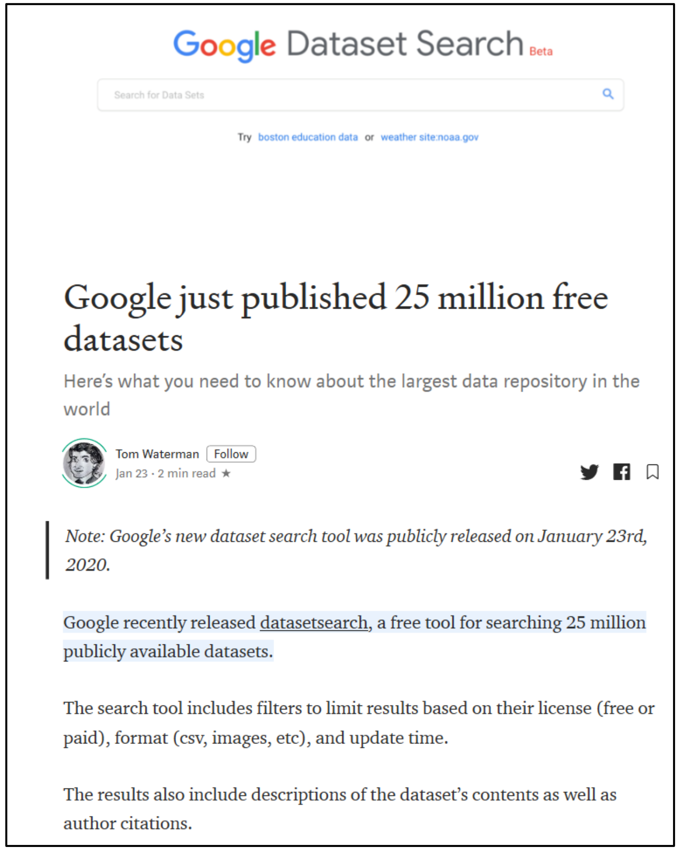
Data-Hungriness

- Human societies have always organized with others of non-competing interests for protection against competing interests
- Every company needs to do the same
- Who are your non-competing-interest parties?
- Implies an extensible data model

Data-Hungriness: Federating Data

- You're in the data business (not the banking, beer or logistics business…)
- So you want to have as much data as possible
  - And increase this every day: "n+1"
- As part of your business strategy, identify who your data partners might be
  - If they're not a direct competitor, partner!
  - By default, today's large data players *are* your direct competitors
- Data-hungriness is part of your product roadmap

https://datasetsearch.research.google.com/


Google Dataset Search Beta

Search for Data Sets

Try boston education data or weather site:noaa.gov

Google just published 25 million free datasets

Here's what you need to know about the largest data repository in the world

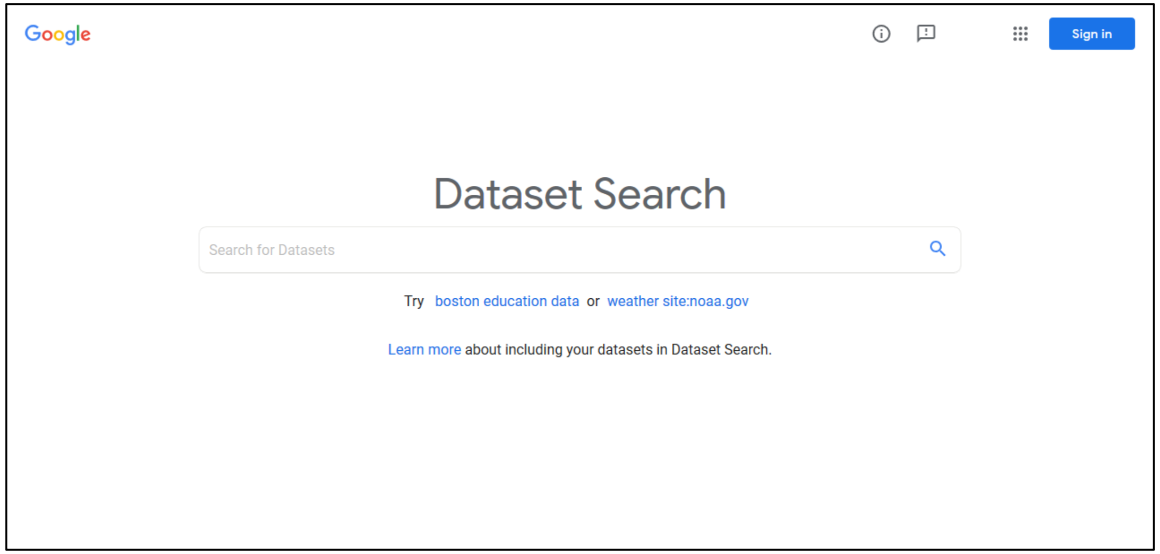
Tom Waterman Follow
Jan 23 · 2 min read ★

> *Note: Google's new dataset search tool was publicly released on January 23rd, 2020.*

Google recently released datasetsearch, a free tool for searching 25 million publicly available datasets.

The search tool includes filters to limit results based on their license (free or paid), format (csv, images, etc), and update time.

The results also include descriptions of the dataset's contents as well as author citations.



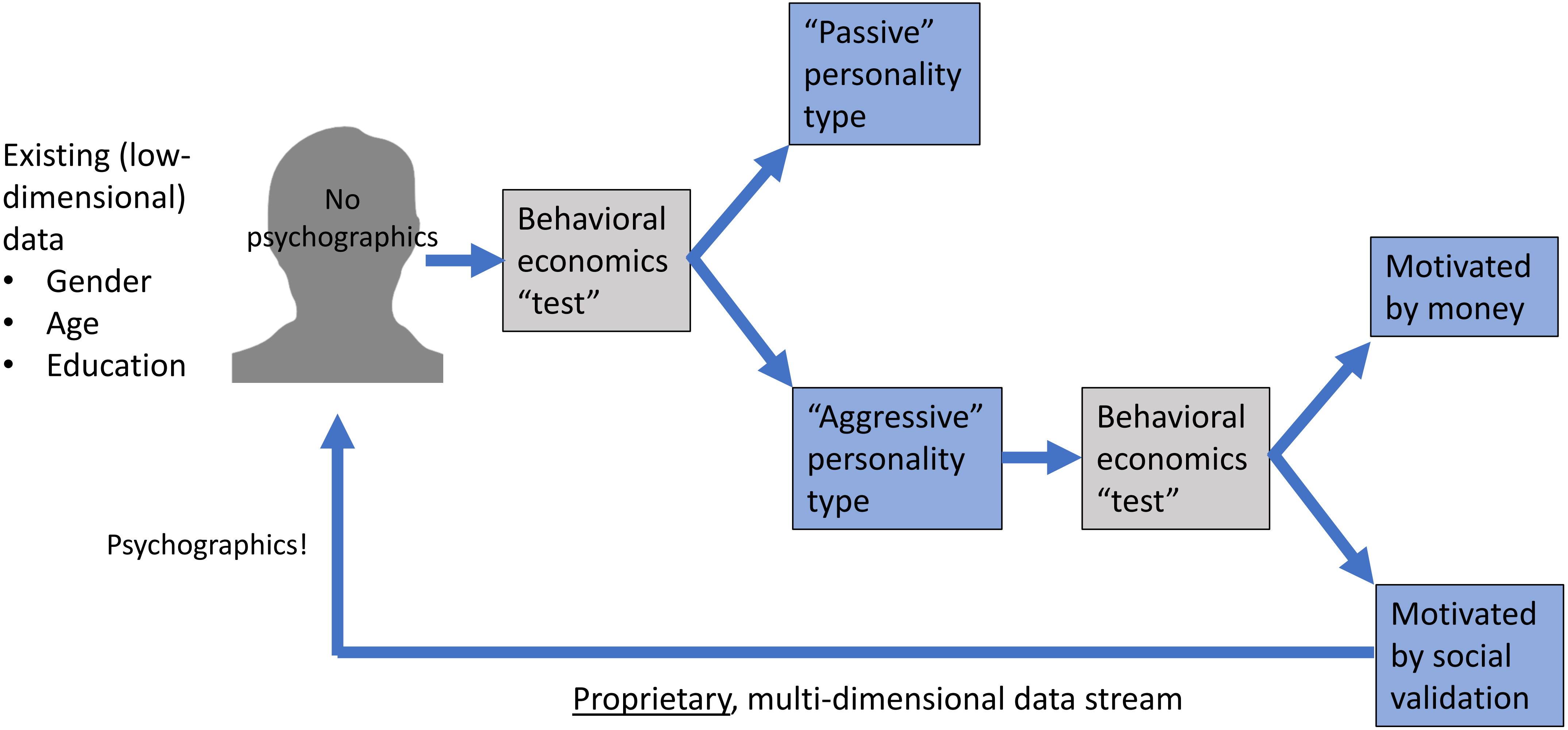
Why Did Google Publish 25 Million Free Data Sets?

A. Because they're devoted to improving the human condition

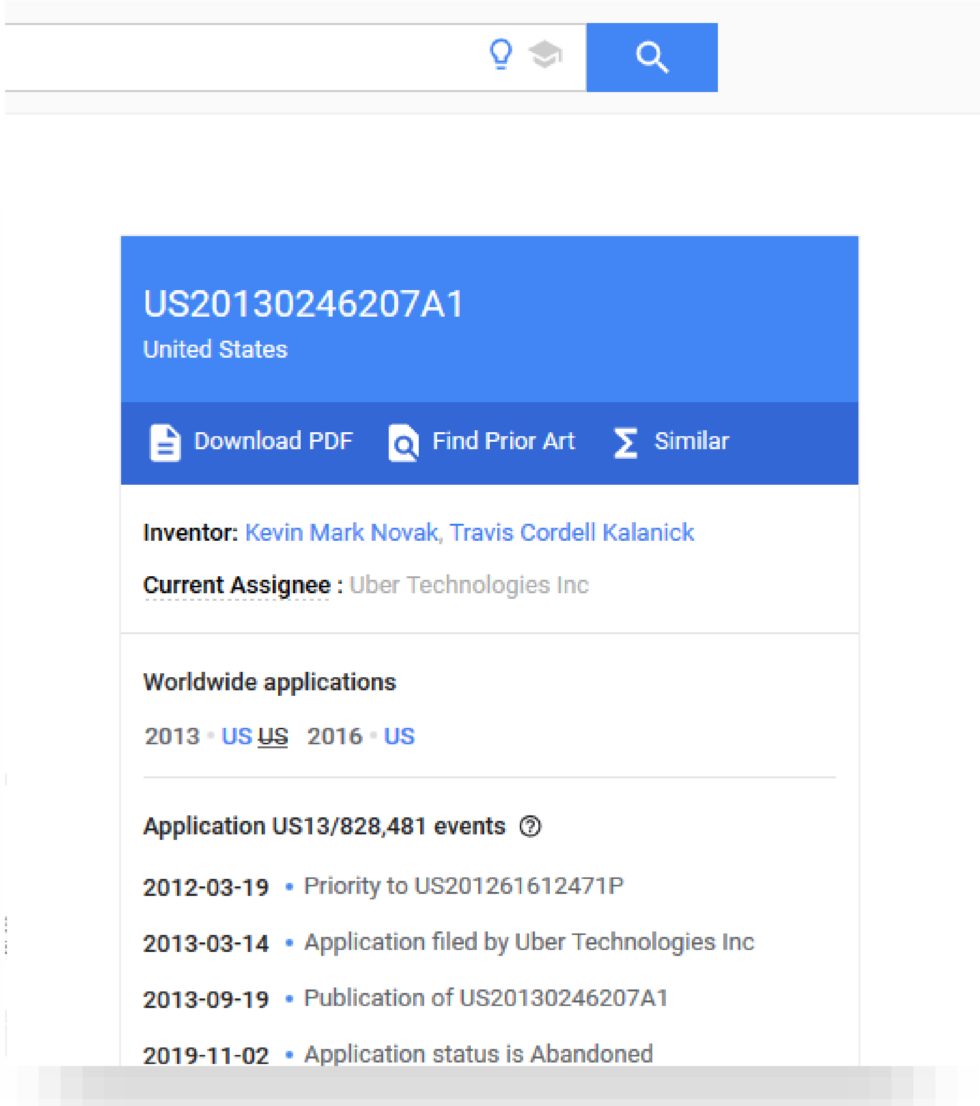
B. Because they want to see what data sets people seek out and use. (<u>Everything</u> is a data signal, and we users will freely provide a signal on what data we find most interesting, which in turn will be of acute interest to Google.)

Call me a jaded cynic, but I'm going with "B"

Behavioral Economics as Multi-dimensional Data



Uber Case Study…



Uber Patent: Data-hungriness & Data-richness

- Driver profile (type/class of vehicle, mobile device profile, …)
- Passenger profile (historical)
- Time & date: current
- Historical pattern for time & date
- Weather
- Calendar: holiday, first day of school, voting day, …
- Event information
  - What / Where?
  - Number of attendees
  - Historically, is demand higher before or after the event?
- Traffic
- Flight information from airports and airlines
- Social networking information
- News (fire, emergencies)

Data Strategy: Differentiation!

Most important

What are my n+1 proprietary sources?

Second most important ("partially proprietary")

What are my n+1 partner sources?

Easiest to get

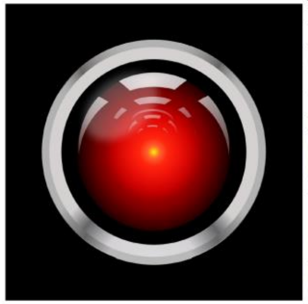
What are my n+1 free sources?

The (Data) Scientific Method: An Infinite Loop of Experimentation

1. Observation: business' strategic goal
2. Hypothesis: identify economic inefficiency
3. Test: data (selection) + method (selection)
4. Analyze: did it work? And return to step 2!

The geographic bias in medical AI tools

23 September 2020, by Shana Lynch



Credit: Pixabay/CC0 Public Domain

Just a few decades ago, scientists didn't think much about diversity when studying new medications. Most clinical trials enrolled mainly white men living near urban research institutes, with the assumption that any findings would apply equally to the rest of the country. Later research

examined clinical applications of machine learning to find that most algorithms are trained on datasets from patients in only three geographic areas, and that the majority of states have no represented patients whatsoever.

"AI algorithms should mirror the community," says Amit Kaushal, an attending physician at VA Palo Alto Hospital and Stanford adjunct professor of bioengineering. "If we're building AI-based tools for patients across the United States, as a field, we can't have the data to train these tools all coming from the same handful of places."

Kaushal, along with Russ Altman, a Stanford professor of bioengineering, genetics, medicine, and biomedical data science, and Curt Langlotz, a professor of radiology and biomedical informatics research, examined five years of peer-reviewed articles that trained a deep-learning algorithm for a diagnostic task intended to assist with patient care. Among U.S. studies where geographic origin could be characterized, they found the majority (71%) used patient data from California, Massachusetts, or New York to train the algorithms. Some 60% solely relied on these three locales. Thirty-four states were not represented at all, while the other 13 states contributed limited data.

The research didn't expose bad outcomes from AI trained on the geographies, but raised questions

STAT Topics Coronavirus Opinion Podcast Newsletters Reports Events TRY STAT PLUS

HEALTH TECH

AI systems are worse at diagnosing disease when training data is skewed by sex

By REBECCA ROBBINS / MAY 25, 2020

Please consider making a contribution to support our coronavirus coverage.

CONTRIBUTE

MOST POPULAR

'It's something I have never seen': How the Covid-19 virus hijacks cells

Racial disparities in automated speech recognition

Allison Koenecke[a], Andrew Nam[b], Emily Lake[c], Joe Nudell[d], Minnie Quartey[e], Zion Mengesha[c], Connor Toups[c], John R. Rickford[c], Dan Jurafsky[c,f], and Sharad Goel[d,1]

[a]Institute for Computational & Mathematical Engineering, Stanford University, Stanford, CA 94305; [b]Department of Psychology, Stanford University, Stanford, CA 94305; [c]Department of Linguistics, Stanford University, Stanford, CA 94305; [d]Department of Management Science & Engineering, Stanford University, Stanford, CA 94305; [e]Department of Linguistics, Georgetown University, Washington, DC 20057; and [f]Department of Computer Science, Stanford University, Stanford, CA 94305

Edited by Judith T. Irvine, University of Michigan, Ann Arbor, MI, and approved February 12, 2020 (received for review October 5, 2019)

Automated speech recognition (ASR) systems, which use sophisticated machine-learning algorithms to convert spoken language to text, have become increasingly widespread, powering popular virtual assistants, facilitating automated closed captioning, and enabling digital dictation platforms for health care. Over the last several years, the quality of these systems has dramatically improved, due both to advances in deep learning and to the collection of large-scale datasets used to train the systems. There is concern, however, that these tools do not work equally well for all subgroups of the population. Here, we examine the ability of five state-of-the-art ASR systems—developed by Amazon, Apple, Google, IBM, and Microsoft—to transcribe structured interviews conducted with 42 white speakers and 73 black speakers. In total, this corpus spans five US cities and consists of 19.8 h of audio matched on the age and gender of the speaker. We found that all five ASR systems exhibited substantial racial disparities, with an average word error rate (WER) of 0.35 for black speakers compared with 0.19 for white speakers. We trace these disparities to the underlying acoustic models used by the ASR systems as the race gap was equally large on a subset of identical phrases spoken by black and white individuals in our corpus. We conclude by proposing strategies—such as using more diverse training datasets that include African American Vernacular English—to reduce these performance differences and ensure speech recognition technology is inclusive.

Princeville, a rural, nearly exclusively African American community in eastern North Carolina; Rochester, a moderate-sized city in Western New York; and the District of Columbia. The second dataset we use is Voices of California (VOC) (26), an ongoing compilation of interviews recorded across the state in both rural and urban areas. We focus our analysis on two California sites: Sacramento, the state capitol; and Humboldt County, a predominately white rural community in Northern California.

In both datasets, the interviews were transcribed by human experts, which we use as the ground truth when evaluating the performance of machine transcriptions. The original recorded interviews contain audio from both the interviewer and the interviewee. Our study is based on a subset of audio snippets that exclusively contain the interviewee and are 5 to 50 s long. We match these snippets across the two datasets based on the age and gender of the speaker and the duration of the snippet. After matching, we are left with 2,141 snippets from each dataset, with an average length of 17 s per snippet, amounting to 19.8 total hours of audio. In the matched dataset, 44% of snippets were of male speakers, and the average age of speakers was 45 y.

We assess the performance of the ASR systems in terms of the word error rate (WER) (27), a standard measure of discrepancy between machine and human transcriptions. Formally, WER is defined as:

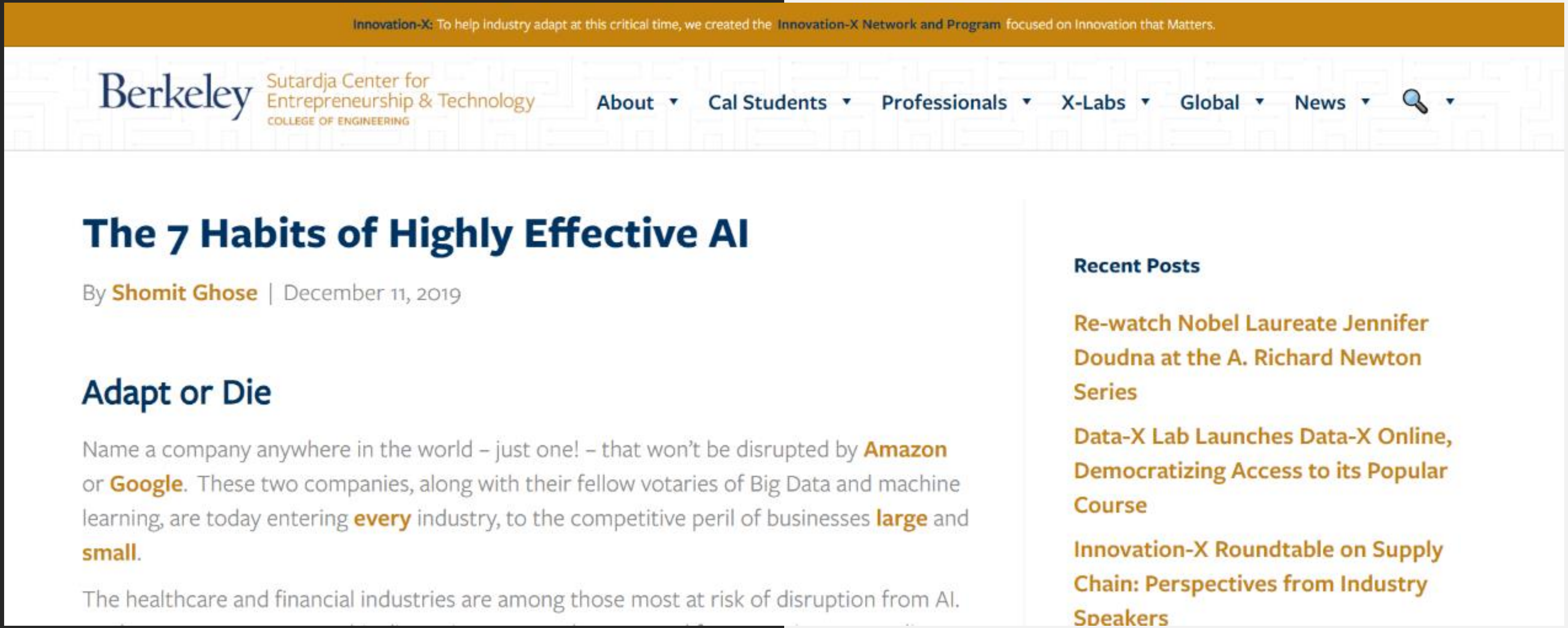
If Training Data is Skewed (Even Unintentionally), It's Not Ethical

Data Ethics : The "Guilty Teenager" Operating Model

Data Ethics: The Guilty Teenager Operating Model

Further Reading

https://scet.berkeley.edu/the-7-habits-of-highly-effective-ai/



Innovation-X: To help industry adapt at this critical time, we created the Innovation-X Network and Program focused on Innovation that Matters.

Berkeley Sutardja Center for Entrepreneurship & Technology
COLLEGE OF ENGINEERING

About ▾ Cal Students ▾ Professionals ▾ X-Labs ▾ Global ▾ News ▾

The 7 Habits of Highly Effective AI

By **Shomit Ghose** | December 11, 2019

Adapt or Die

Name a company anywhere in the world – just one! – that won't be disrupted by **Amazon** or **Google**. These two companies, along with their fellow votaries of Big Data and machine learning, are today entering **every** industry, to the competitive peril of businesses **large** and **small**.

The healthcare and financial industries are among those most at risk of disruption from AI.

Recent Posts

Re-watch Nobel Laureate Jennifer Doudna at the A. Richard Newton Series

Data-X Lab Launches Data-X Online, Democratizing Access to its Popular Course

Innovation-X Roundtable on Supply Chain: Perspectives from Industry Speakers

Mange tak
(Data wins!)

shomit@berkeley.edu