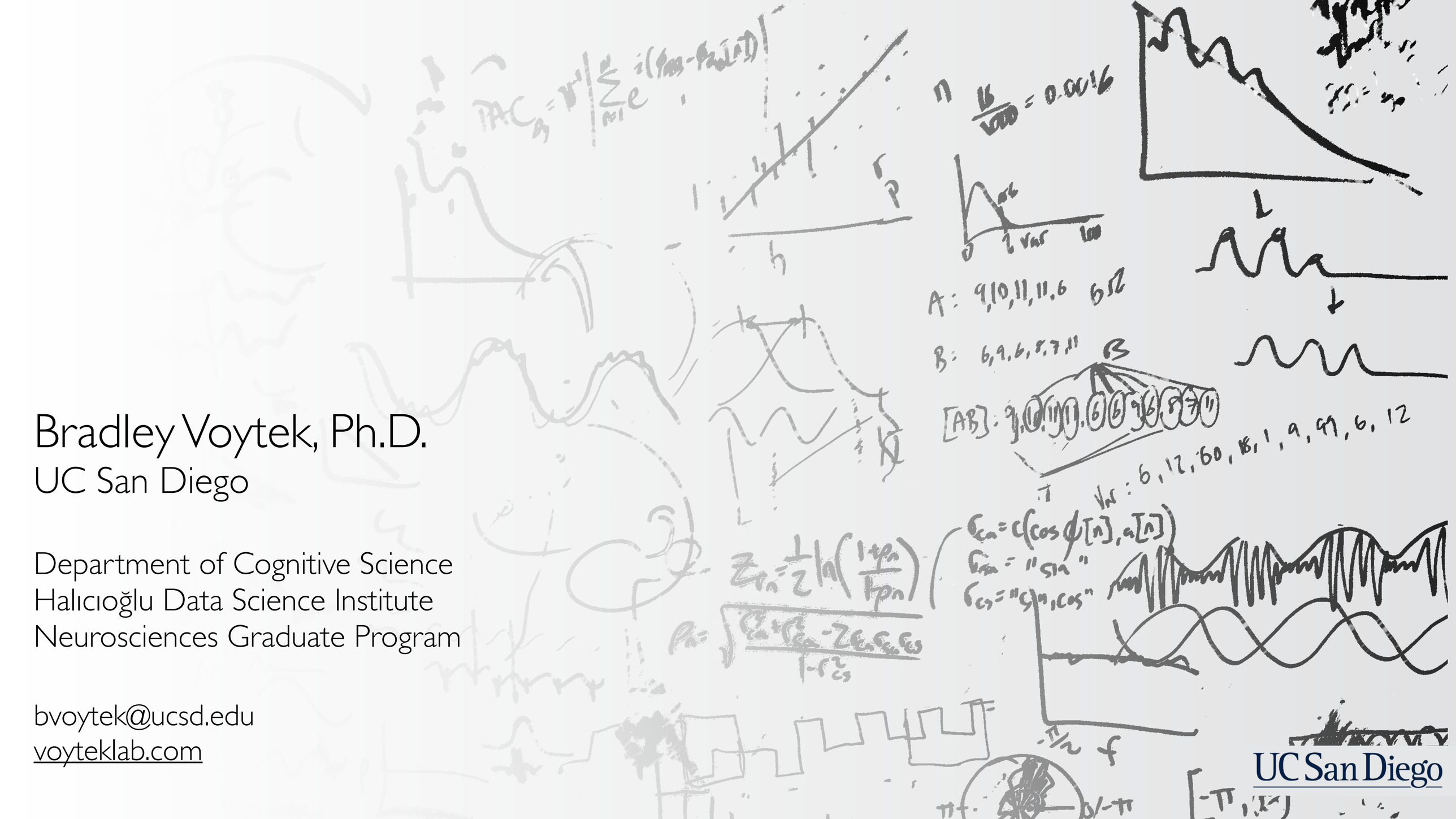
Bradley Voytek, Ph.D.

UC San Diego

Department of Cognitive Science
Halıcıoğlu Data Science Institute
Neurosciences Graduate Program

bvoytek@ucsd.edu
voyteklab.com



UC San Diego

Reading # 1

50 years of Data Science

David Donoho

Sept. 18, 2015
Version 1.00

It's a long read! Give yourself time.

When reading…

Consider:

-What points were made that you had not previously considered?
-What points do you agree with? With which do you disagree?
-What do you not understand? (Ask questions about these in section!)
-What points/topics do you think are missing?
-Does anything differ with what you've learned in other classes? Other papers/books?
-What ideas did this reading give you? What do you want to go learn more about?

Reading quiz rules

Data Science Student Society (DS3)

Source: https://ds3.ucsd.edu/

DATA SCIENCE ALLIANCE

Patricia Lopez

Executive Director

Dr. Adir Mancebo Jr.

Data Project Manager

Kallyn Hobmann

Communications Manager

Phish Jeffrey Moore

Principal Brand Designer

Leslie Joe

Data Scientist

Claire (Zhihan) Li

M.S. in Data Science

Orion Tang

B.S. in Cognitive Science, Specialization in Design and Interaction

Mehri Sadri

B.S. in Economics, M.S. in Public Policy, Program Design & Evaluation

Eduardo Spiegel

B.S. in Data Science

Nived Neetha Sooraj

MSc in Data Science

Ana Caballero

B.S. in Cognitive Science, B.A. in International Business

Ty Albao

B.S. in Data Science

Jacie Littell

B.A. in Speculative Design

Purich Viwatkurkul

B.S. in Mathematics-Computer Science

Nicole Go

B.S. in Computer Science

Shay Samat

B.S. in Cognitive Science

Brenden Joseph Le

B.A. in Graphic Design

Diego Pereyra

B.S. in Computer Engineering

White Paper

Guiding Principles Of Responsible Data Science

How to Build a World Where Data Helps Without Harm

Authors

Adir Mancebo Jr., Ph.D.

Data Analyst, Data Science Alliance

Bradley Voytek, Ph.D.

Professor, Cognitive Science,
Data Science, and Neuroscience, UC San Diego
Founding Board Member, Data Science Alliance

Ilkay Altintas, Ph.D.

Chief Data Science Officer,
San Diego Supercomputer Center
Founding Board Member, Data Science Alliance

Di Le

Lead AI/ML Design Strategist, ServiceNow
Working Group Member, Data Science Alliance

Source: Data Science Alliance

Pledge for Responsible Data Science

- **I aim to provide a positive impact to our peers, community, and society as a whole, while minimizing harm, discrimination, or inequities that would result as a consequence of my work.**
- **I respect privacy and will adhere to localized laws and governance of personal information.**
- **I am rigorous, scientic, and strive to ensure my work is accurate and reflective of the truth to the best of my ability.**
- **I am transparent and will make my intentions clear with how I use personal information. I am proactive in communicating when there are risks to the above.**

Source: Data Science Alliance


DATA
SCIENCE
LLIANCE


DATA SCIENCE ALLIANCE
Pledge for Responsible Data Science
The Data Science Alliance is committed to building a community of responsible Data Science practitioners, and creating a framework to support responsible practices. The following values are what guide us.
I aim to provide a positive impact to our peers, community, and society as a whole, while minimizing harm, discrimination, or inequities that would result as a consequence of my work.
I respect privacy and will adhere to localized laws and governance of personal information.
I am rigorous, scientific, and strive to ensure my work is accurate and reflective of the truth to the best of my ability.
I am transparent and will make my intentions clear with how I use personal information. I am proactive in communicating when there are risks to the above.
In joining the Data Science Alliance, I expect our members to have the same principles and values, and provide role-model leadership that demonstrates responsible data science practice. Together, our members preserve these principles as we build a diverse, collaborative, and inclusive community of responsible data users.
Bill Nye
CEO of The Planetary Society
Todd Gloria
Mayor of San Diego
Patricia Jay R. Lopez
Executive Director of DSA
Date Signed — August 4, 2022

COGS 9
Introduction to Data Science

Biases & Ethics

UC San Diego

Today's Learning Objective

Determine how data scientists can address ethical concerns while maximizing privacy and minimizing harms.

Big data is *really* big

1. There is more data than ever in the history of data (Smolan and Erwitt 2012):

- Beginning of recorded history till 2003—5 billion gigabytes
- 2011—5 billion gigabytes every two days
- 2013—5 billion gigabytes every 10 min
- 2015—5 billion gigabytes every 10 s

- By the end of 2020, 44 zettabytes will make up the entire digital universe.
- Every day, 306.4 billion emails are sent, and 500 million Tweets are made.
- 463 exabytes of data will be generated each day by people as of 2025.

Source: Zwitter, *Big Data & Society* 2014

Big data ethics

"Big data and analytics technology can reap huge benefits to both individuals and organizations – bringing personalized service, detection of fraud and abuse, efficient use of resources and prevention of failure or accident. **So why are there questions being raised about the ethics of analytics, and its related technology, Big Data?**"

Source: IBM


AND, A SHORT DISTANCE AWAY...
MY FAULT--ALL MY FAULT! IF ONLY I HAD STOPPED HIM WHEN I COULD HAVE! BUT I DIDN'T--AND NOW --UNCLE BEN-- IS DEAD...
AND A LEAN, SILENT FIGURE SLOWLY FADES INTO THE GATHERING DARKNESS, AWARE AT LAST THAT IN THIS WORLD, WITH GREAT POWER THERE MUST ALSO COME--GREAT RESPONSIBILITY!
AND SO A LEGEND IS BORN AND A NEW NAME IS ADDED TO THE ROSTER OF THOSE WHO MAKE THE WORLD OF

Building a data-driven company

What about these data-driven organizations enables them to use data to gain a competitive advantage? In *Building Data Science Teams*, we said that a data-driven organization

> acquires, processes, and leverages data in a timely fashion to create efficiencies, iterate on and develop new products, and navigate the competitive landscape..

Source: Mason & Patil, *Data Driven* 2015

Data-driven

The analysis showed that Twitter needed to (a) teach new users what a tweet was, (b) suggest accounts that had high-quality content segmented by categories (e.g., NFL, NBA, news sites), and then (c) suggest other users who were highly likely to follow someone once they knew that person was on Twitter. Implementing these ideas adds friction to the onboarding process by teaching users about the tweet; it also puts people a user is likely to interact with last. However, the result wasn't a decrease in new users, but instead a 30% increase in people completing the experience and a 20% increase in long-term engagement!

Source: Mason & Patil, *Data Driven* 2015

Kinds of data

- **Experimental:**
- Data collected as part of a hypothesis-driven experiment.
  - examples include clinical interventions, brain imaging, chemical sampling, etc.

- **Observational (or retrospective or secondary):**
- Data just "sitting around"
  - examples include the human genome, web data, types of species, etc.

Source: *Smyth, UC Irvine*

Evidence-based decision making

- Experts analyze data, present the results to decision makers.
- Data analysts and decision makers operate in concert, discussing implications and refining the analysis/predictions together.
- Decision makers task experts to find a rationale for a decision they've already made.

Worst case scenario

Evidence-based decision making

- The commission found that the Challenger accident was caused by a failure in the O-rings sealing the aft field joint on the right solid rocket booster, causing pressurized hot gases and eventually flame to "blow by" the O-ring and contact the adjacent external tank, causing structural failure.
- The failure of the O-rings was attributed to a design flaw, as their performance could be too easily compromised by factors including the low temperature on the day of launch.


www.FeynmanPhysicsLectures

Evidence-based decision making

- Early tests resulted in some of the booster rocket's O-rings burning a third of the way through.
- These O-rings provided the gas-tight seal needed between the vertically stacked cylindrical sections that made up the solid fuel booster.
- NASA managers recorded this result as demonstrating that the O-rings had a "safety factor" of 3.

Evidence-based decision making

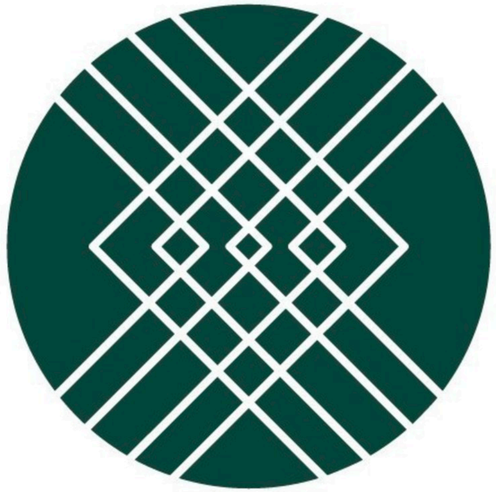
- Feynman incredulously explains the magnitude of this error: a "safety factor" refers to the practice of building an object to be capable of withstanding more force than the force to which it will conceivably be subjected.

Evidence-based decision making

- To paraphrase Feynman's example, if engineers built a bridge that could bear 3,000 pounds without any damage, even though it was never expected to bear more than 1,000 pounds in practice, the safety factor would be 3.
- If a 1,000 pound truck drove across the bridge and it cracked at all, even just a third of the way through a beam, the safety factor is now zero: the bridge is defective.
- But the "data-driven" approach to measuring the O-ring safety factor was "3", giving a false sense of security.

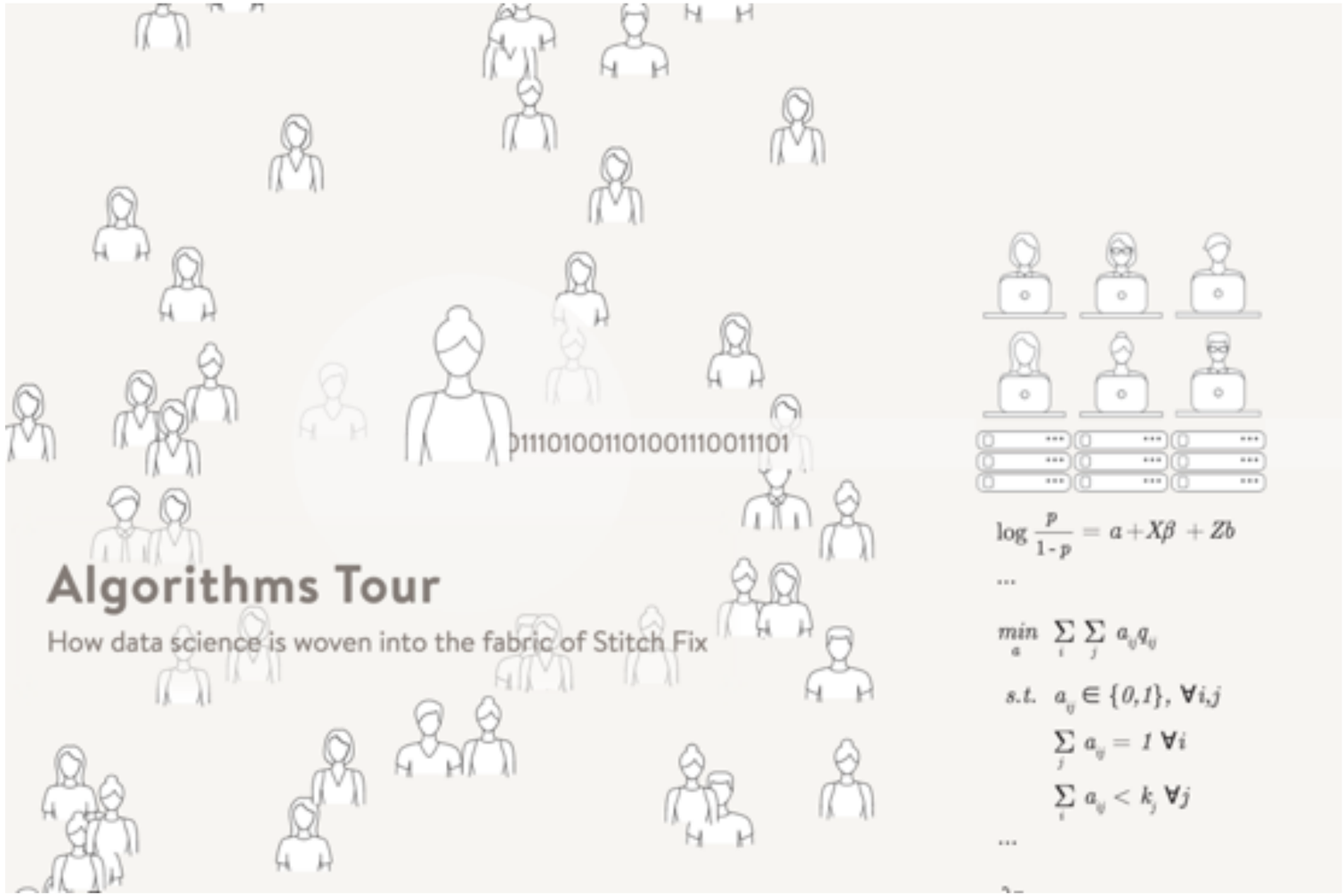
How algorithms drive our lives

Algorithms tour



STITCH FIX

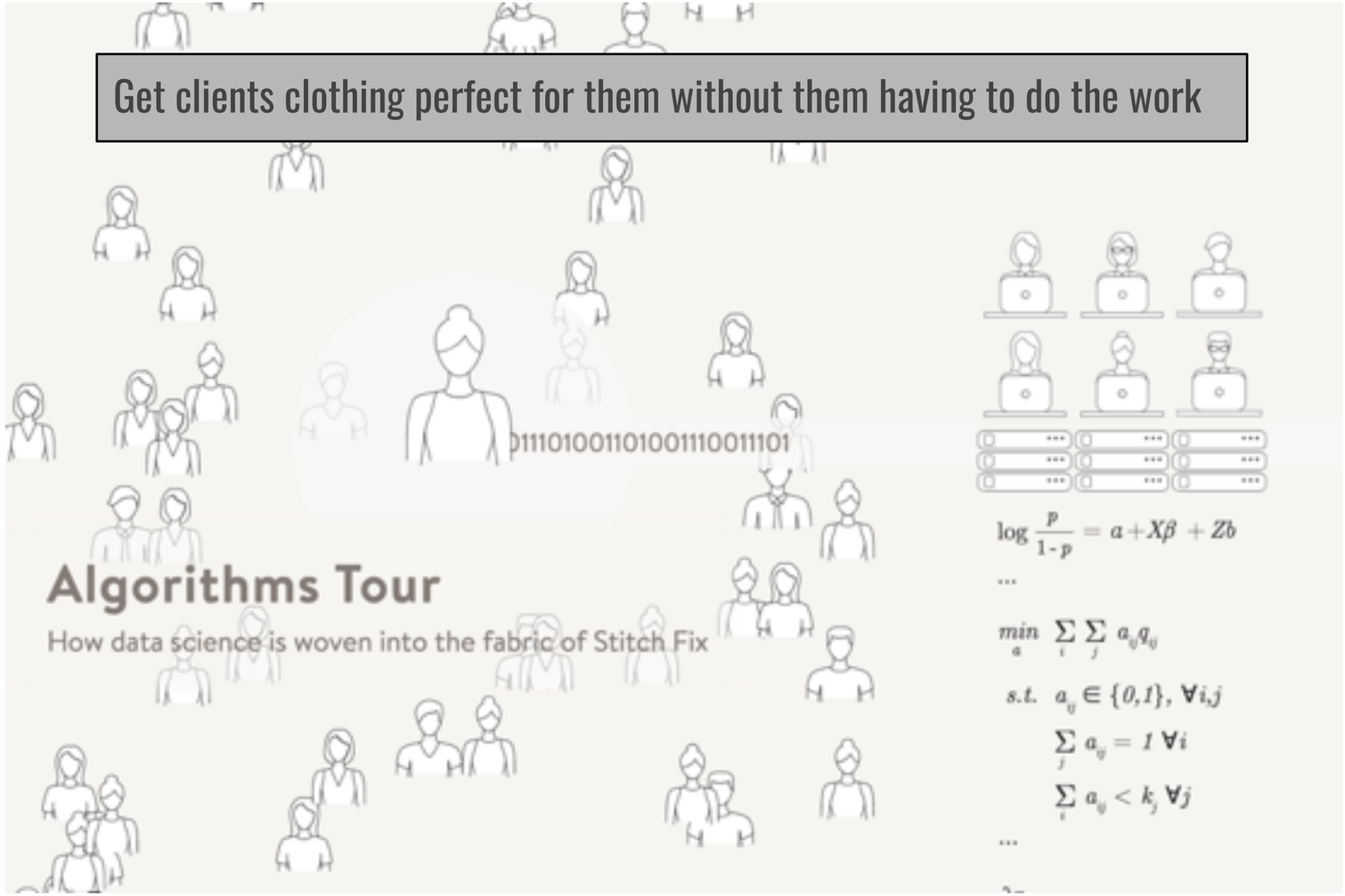
Algorithms tour



Source: https://algorithms-tour.stitchfix.com

UC San Diego

Algorithms tour



Source: https://algorithms-tour.stitchfix.com

Algorithms tour

> Our business model enables unprecedented data science, not only in recommendation systems, but also in human computation, resource management, inventory management, algorithmic fashion design and many other areas. Experimentation and algorithm development is deeply engrained in everything that Stitch Fix does. We'll describe a few examples in detail as you scroll along.

Algorithms come into play throughout the entire process at stitch fix

Source: https://algorithms-tour.stitchfix.com

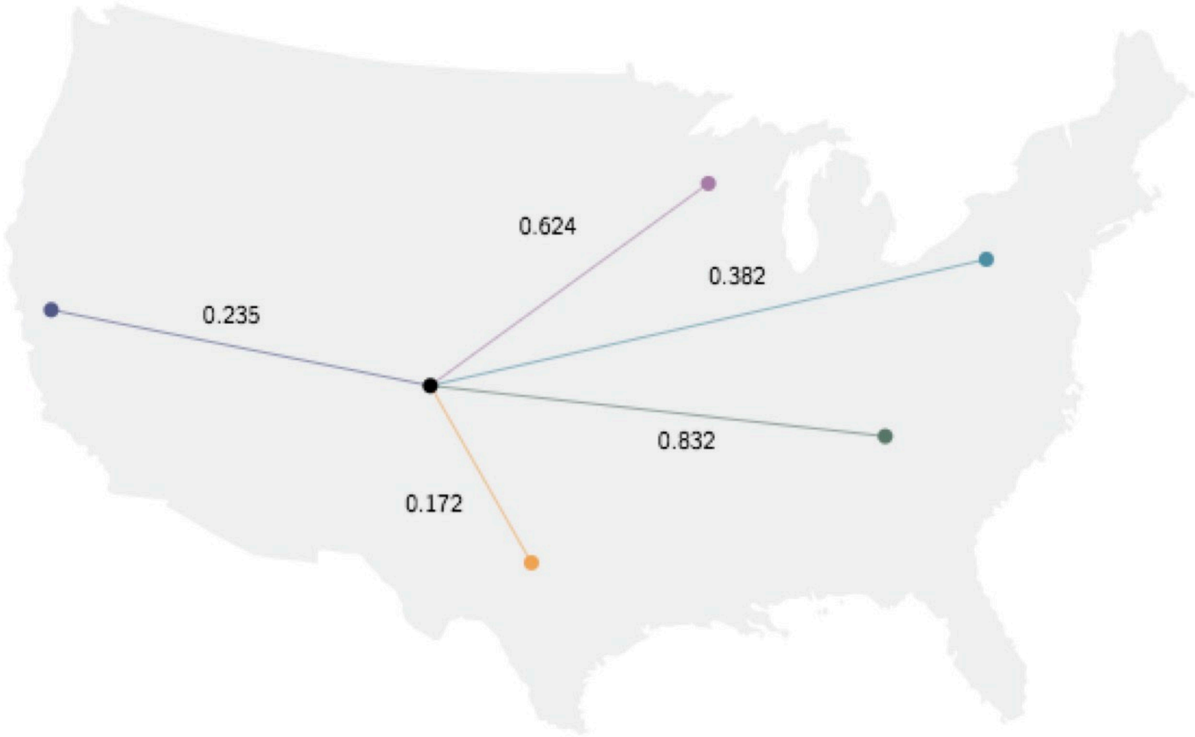
UC San Diego

Algorithms tour

Warehouse Assignment

$$q_{ij} = f\left(\text{shipping\_cost}_{ij},\ \text{shipping\_time}_{ij},\ \text{inventory\_match}_{ij},\ \ldots\right)$$

The shipment request is processed by an algorithm that assigns it to a warehouse. This algorithm calculates a cost function for each warehouse based on a combination of its location relative to the client and how well the inventories in the different warehouses match the client's needs.



Warehouse Assignment | Recommendation Systems | Matchmaking | Human Computation | Logistics Optimization

Source: https://algorithms-tour.stitchfix.com

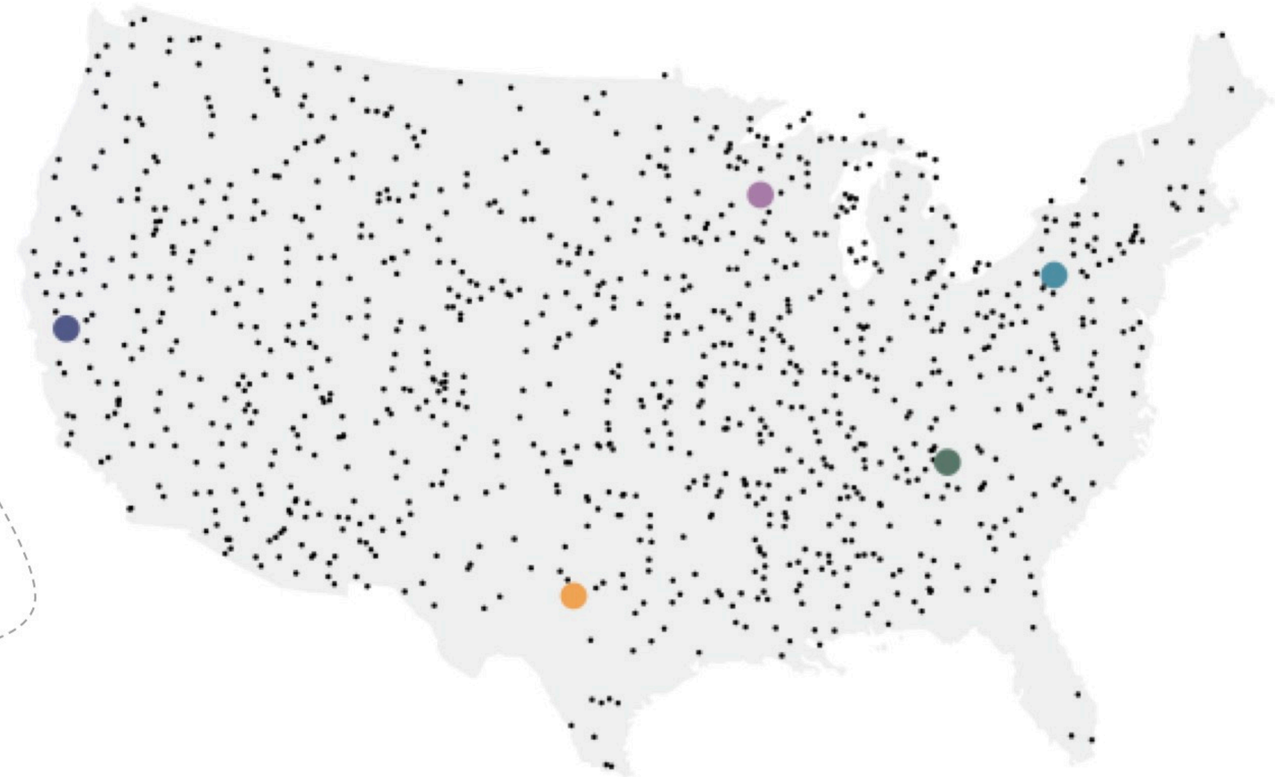
UC San Diego

Algorithms tour

warehouses j

costs q	CA	TX	WI	
clients i	0.235	0.172	0.624	...
	0.347	0.825	0.722	...
	0.533	0.874	0.193	...
	...			

$$q_{ij} = f\left(\text{shipping\_cost}_{ij},\ \text{shipping\_time}_{ij},\ \text{inventory\_match}_{ij},\ \ldots\right)$$



This set of cost calculations is carried out for each client to produce a cost matrix.

Warehouse Assignment — Recommendation Systems — Matchmaking — Human Computation — Logistics Optimization

Source: https://algorithms-tour.stitchfix.com

UC San Diego

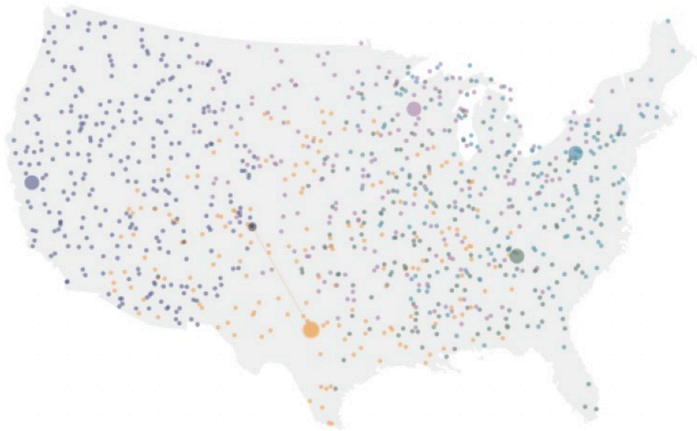
Algorithms tour

Which warehouse minimizes cost?

warehouses j

costs q	CA	TX	WI	
clients i	0.235	0.172	0.624	...
	0.347	0.825	0.722	...
	0.533	0.874	0.193	...
	...			

$$q_{ij} = f(\text{shipping\_cost}_{ij}, \text{shipping\_time}_{ij}, \text{inventory\_match}_{ij}, \ldots)$$



$$\min_{a} \sum_{i} \sum_{j} a_{ij} q_{ij}$$

$$s.t. \quad a_{ij} \in \{0,1\}, \forall i,j$$

$$\sum_{j} a_{ij} = 1 \; \forall i$$

$$\sum_{i} a_{ij} < k_j \; \forall j$$

...

> The assignment of clients to warehouses is then a binary optimization problem.

assignments a

	CA	TX	WI	
	0	1	0	...
	1	0	0	...
	0	0	1	...
	...			

feb 14 c_1

Warehouse Assignment — Recommendation Systems — Matchmaking — Human Computation — Logistics Optimization

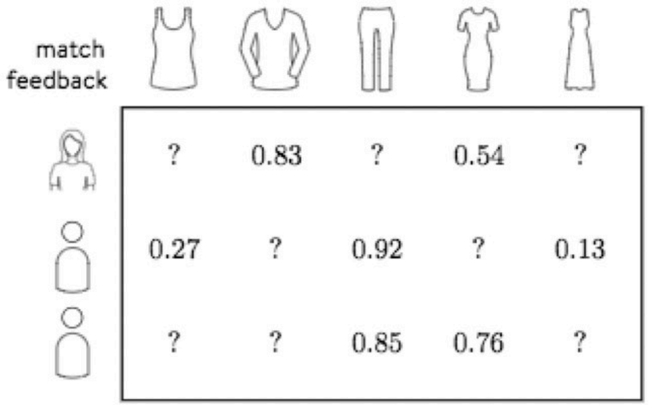
Source: https://algorithms-tour.stitchfix.com

UC San Diego

Algorithms tour

> In some ways, the problem is a classic collaborative filtering problem: given different clients' feedback on different styles, we must fill in the gaps in the (sparse) matrix to predict the result of sending a style to a client who has not yet received it. As such, we *do* use some standard collaborative filtering algorithms (e.g. those who have liked what you have liked have also liked ...).

For example: predict based on what others similar to you have liked historically



solve for missing elements in this client's row ...

Warehouse Assignment | Recommendation Systems | Matchmaking | Human Computation | Logistics Optimization

Source: https://algorithms-tour.stitchfix.com

UC San Diego

Algorithmic decision-making

ROBERT B. McKINLEY: Well, there's a gold mine of information residing out there in these databases by the consumer reporting agencies, the credit bureaus. They're collecting information about what kind of accounts you have open, the balances, whether or not you make those payments on time. And that's a huge reservoir of information there that they can tap into and be able to get a sense as to whether or not a consumer is a revolver, someone who doesn't pay the balance off in full each month. So they can kind of sift those out, and today, it's really become almost surgical.

Source: PBS Frontline, *Secret History of the Credit Card*, 2004

Algorithmic decision-making

NARRATOR: The ability to surgically target consumers and track their financial behavior has become a booming business dominated by three credit reporting agencies which gather information. All that data is then crunched by a little known company called Fair Isaac, which calculates a number called a FICO score for almost every American with a credit history.

Source: PBS Frontline, *Secret History of the Credit Card*, 2004

Algorithmic decision-making

TOM QUINN, Fair Isaac Corp.: We're not a credit-reporting agency like an Equifax, Trans-Union or Experian, that's gathering information daily on consumers and building up consumer records.

Source: PBS Frontline, *Secret History of the Credit Card*, 2004

Algorithmic decision-making

TOM QUINN: We simply work with the credit-reporting agencies, and they deploy their data onto our mathematical formula to create that score.

NARRATOR: The median FICO score is 720 out of a possible 850. The riskiest customers have scores below 600. The score is an indication of how likely you are to pay your bills.

TOM QUINN: Lenders use that score almost like a thermometer to determine if they're going to grant credit or not. So the algorithm is an indication of that consumer's future risk, in terms of credit behavior.

Source: PBS Frontline, *Secret History of the Credit Card*, 2004

The New York Times Magazine

How Companies Learn Your Secrets

"My daughter got this in the mail!" he said. "She's still in high school, and you're sending her coupons for baby clothes and cribs? Are you trying to encourage her to get pregnant?"

Source: https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html

The New York Times Magazine

How Companies Learn Your Secrets

The manager didn't have any idea what the man was talking about. He looked at the mailer. Sure enough, it was addressed to the man's daughter and contained advertisements for maternity clothing, nursery furniture and pictures of smiling infants. The manager apologized and then called a few days later to apologize again.

Source: https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html

The New York Times Magazine

How Companies Learn Your Secrets

On the phone, though, the father was somewhat abashed. “I had a talk with my daughter,” he said. “It turns out there’s been some activities in my house I haven’t been completely aware of. She’s due in August. I owe you an apology.”

Source: https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html

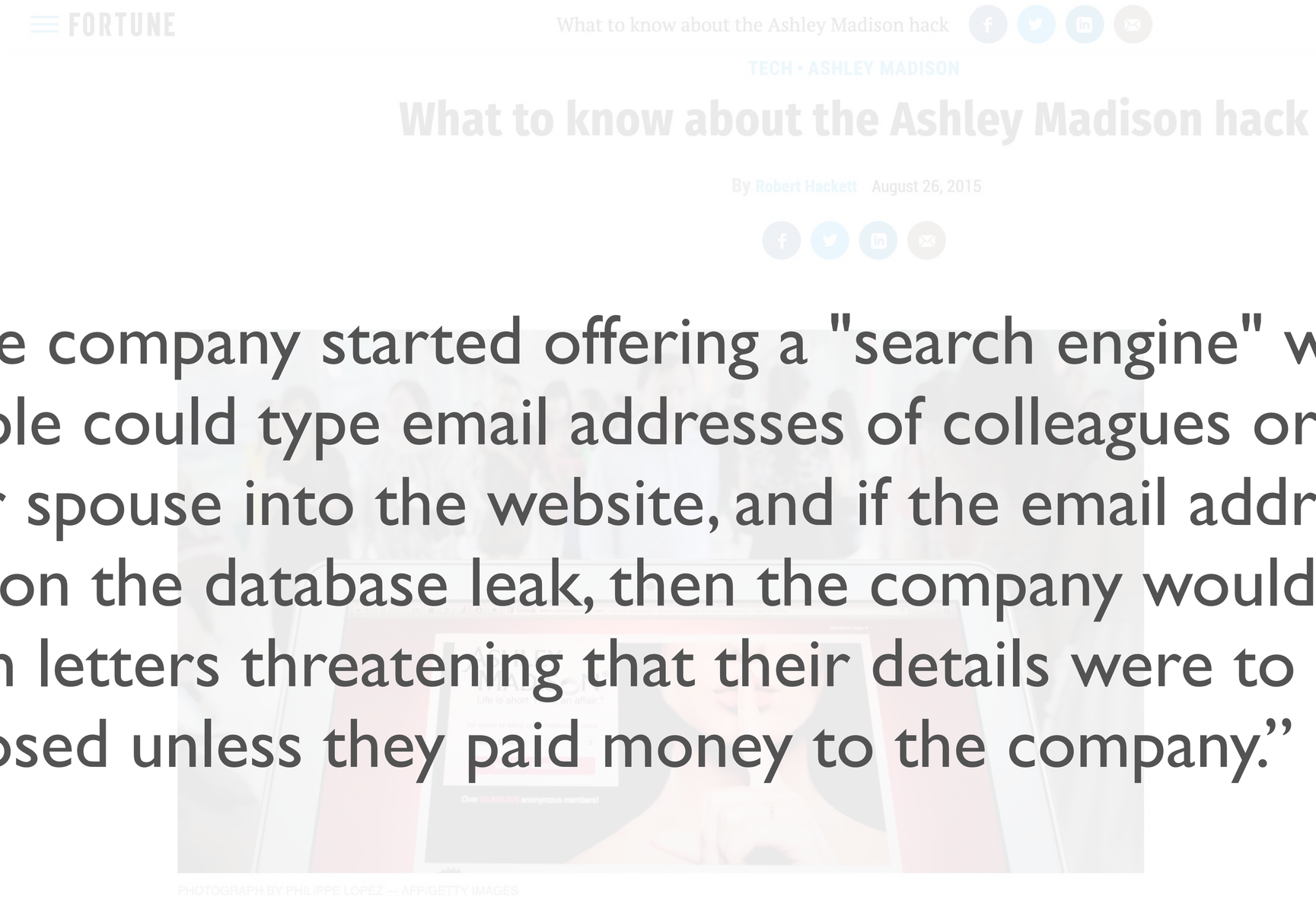
TECH • ASHLEY MADISON

What to know about the Ashley Madison hack

By Robert Hackett August 26, 2015

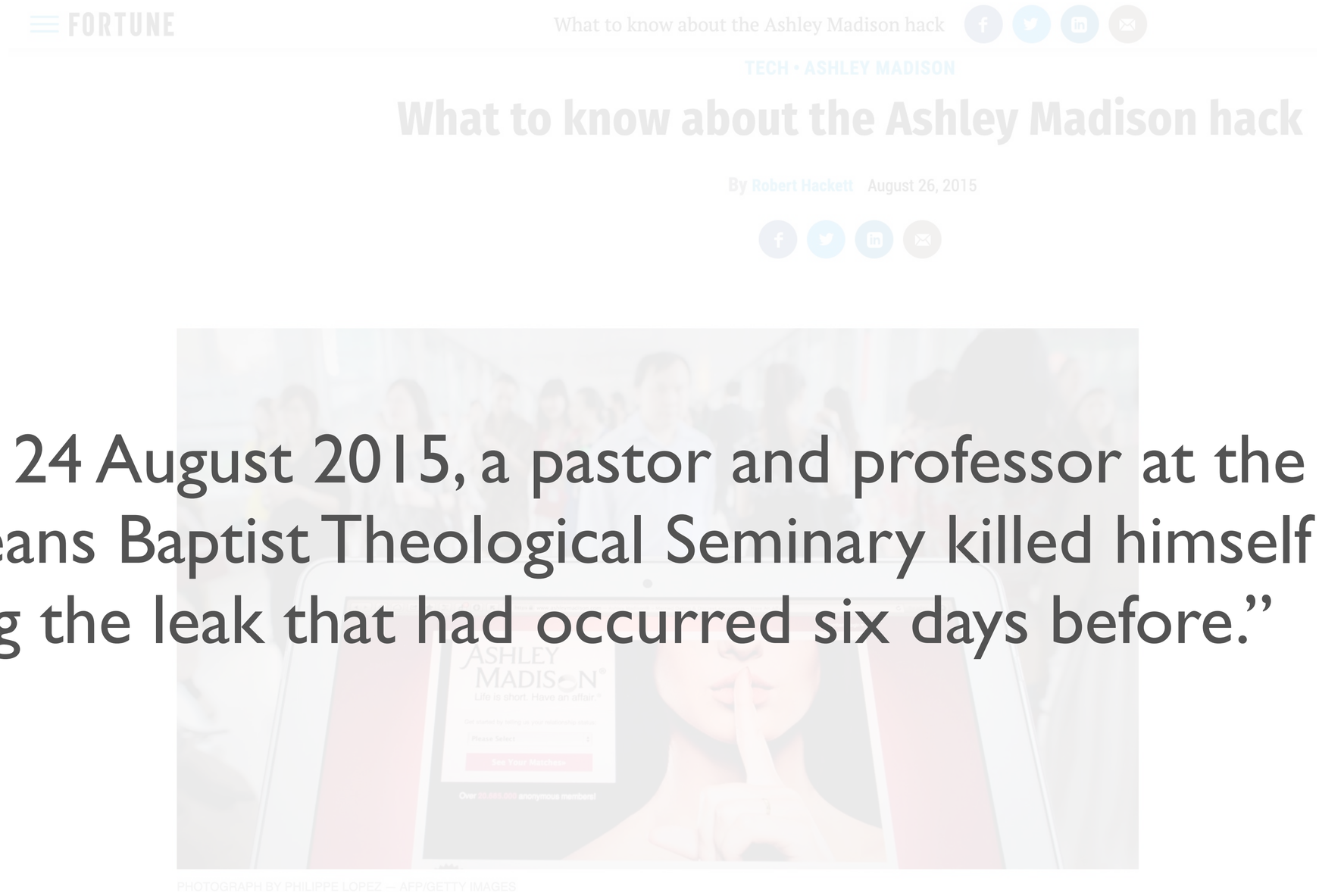
PHOTOGRAPH BY PHILIPPE LOPEZ — AFP/GETTY IMAGES

Source: https://fortune.com/2015/08/26/ashley-madison-hack/



"One company started offering a "search engine" where people could type email addresses of colleagues or their spouse into the website, and if the email address was on the database leak, then the company would send them letters threatening that their details were to be exposed unless they paid money to the company."

Source: https://en.wikipedia.org/wiki/Ashley_Madison_data_breach



"On 24 August 2015, a pastor and professor at the New Orleans Baptist Theological Seminary killed himself citing the leak that had occurred six days before."

Source: https://en.wikipedia.org/wiki/Ashley_Madison_data_breach

GPS

This article is more than **1 year old**

Fitness tracking app Strava gives away location of secret US army bases

Data about exercise routes shared online by soldiers can be used to pinpoint overseas facilities

Latest: Strava suggests military users 'opt out' of heatmap as row deepens

Alex Hern
@alexhern
Sun 28 Jan 2018 16.51 EST

6,895

▲ A military base in Helmand Province, Afghanistan with route taken by joggers highlighted by Strava. Photograph: Strava Heatmap

Source: https://www.theguardian.com/world/2018/jan/28/fitness-tracking-app-gives-away-location-of-secret-us-army-bases

FORTUNE

MPW • AMAZON

Amazon Reportedly Killed an AI Recruitment System Because It Couldn't Stop the Tool from Discriminating Against Women

By David Meyer October 10, 2018

"The tech industry is famously male-dominated and, accordingly, most of those resumes came from men. So, trained on that selection of information, the recruitment system began to favor men over women…"

Source: https://fortune.com/2018/10/10/amazon-ai-recruitment-bias-women-sexist/

"…Amazon's system taught itself to downgrade resumes with the word "women's" in them, and to assign lower scores to graduates of two women-only colleges."

Source: https://fortune.com/2018/10/10/amazon-ai-recruitment-bias-women-sexist/

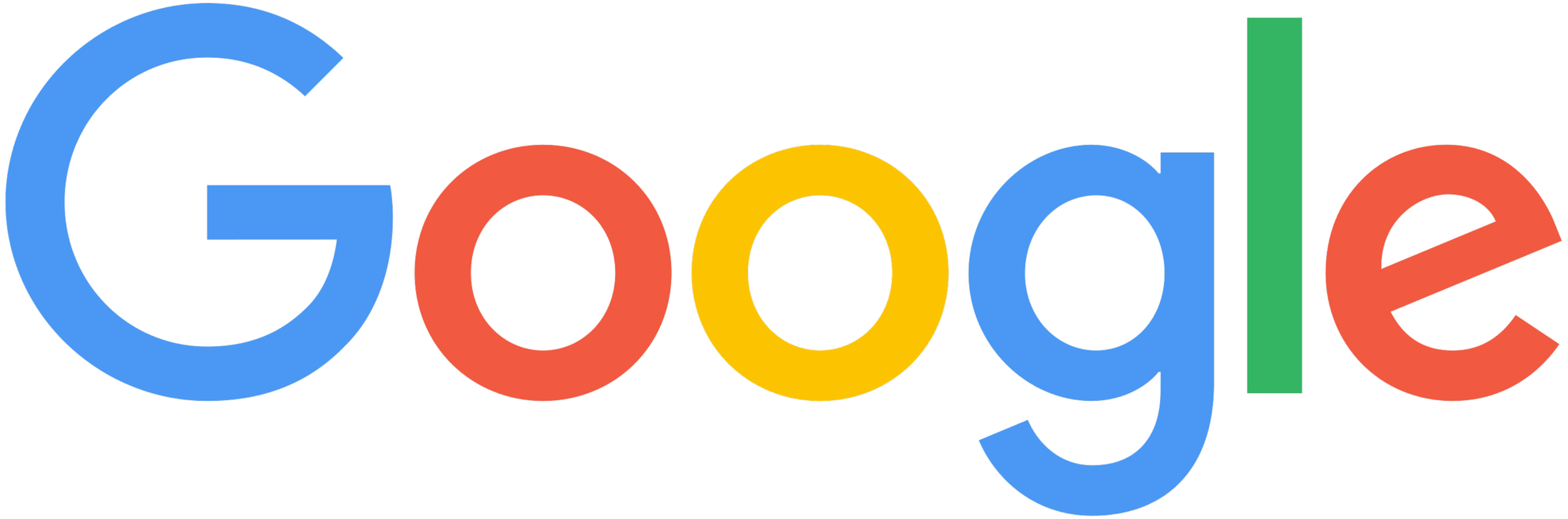
Ethics

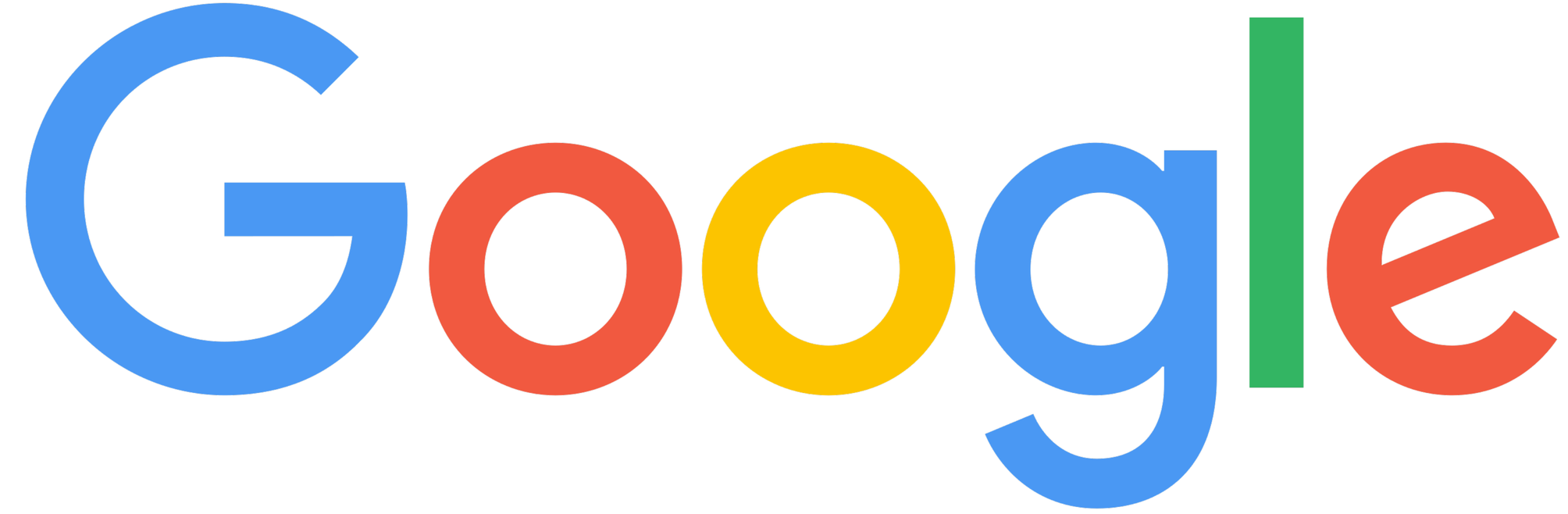
"Moral principles that govern a person's behaviour or the conducting of an activity."

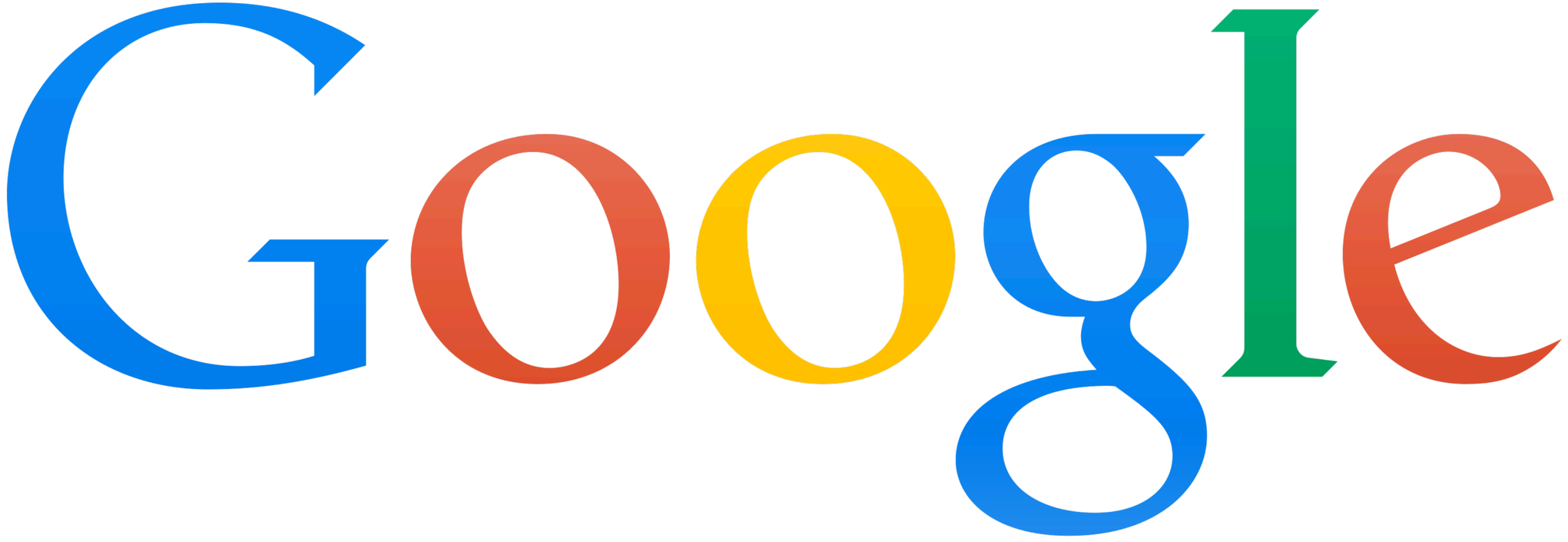
UC San Diego

Source: Oxford Dictionary

Data-driven experiments

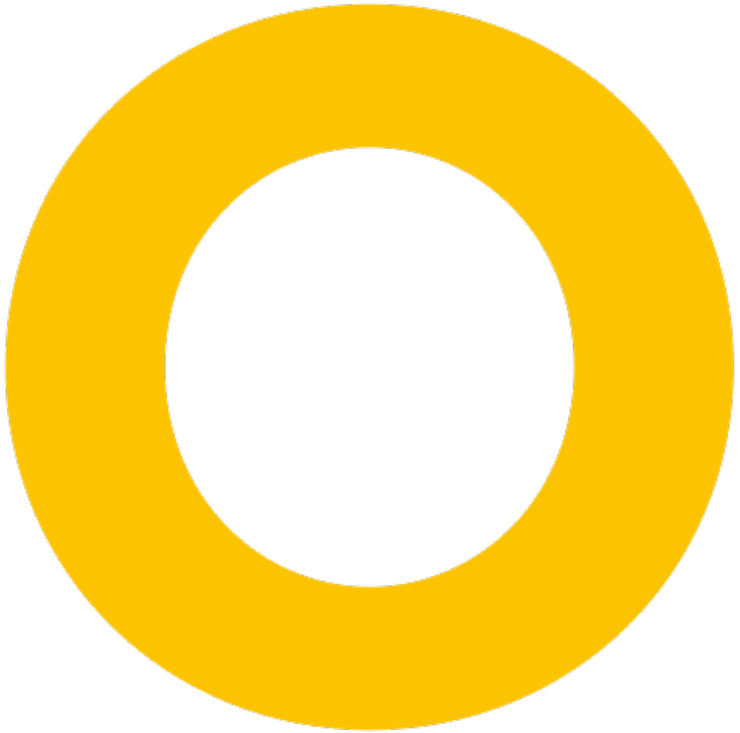

Google

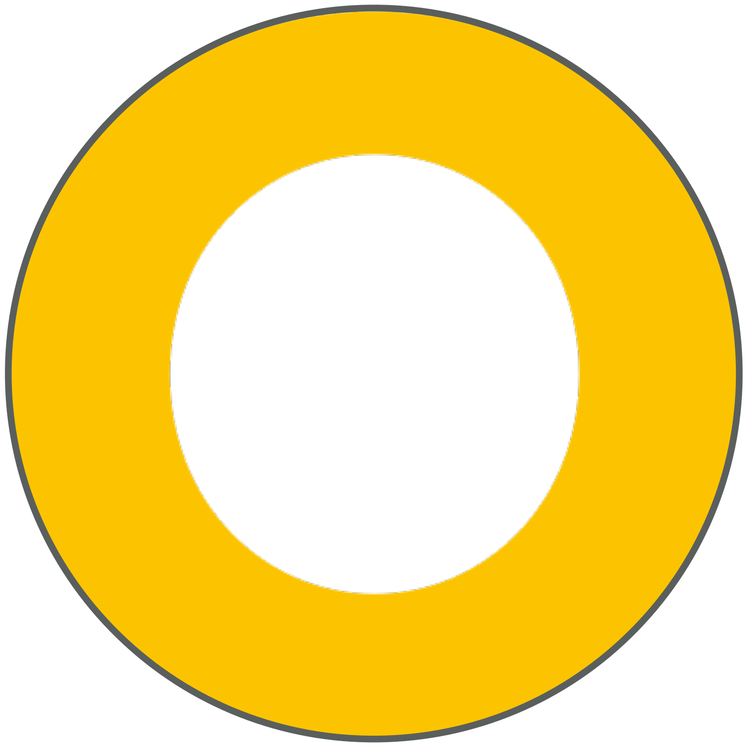

Google

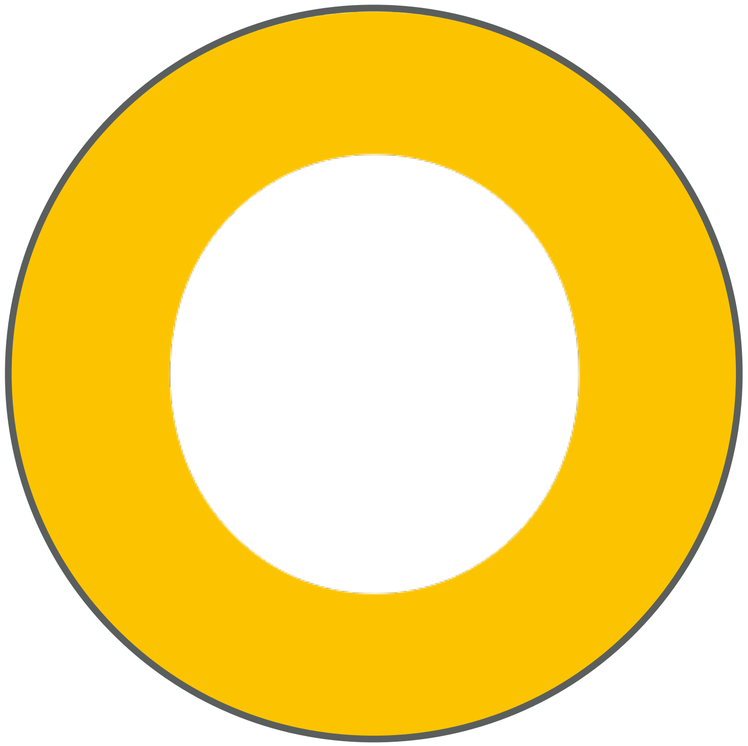

Google

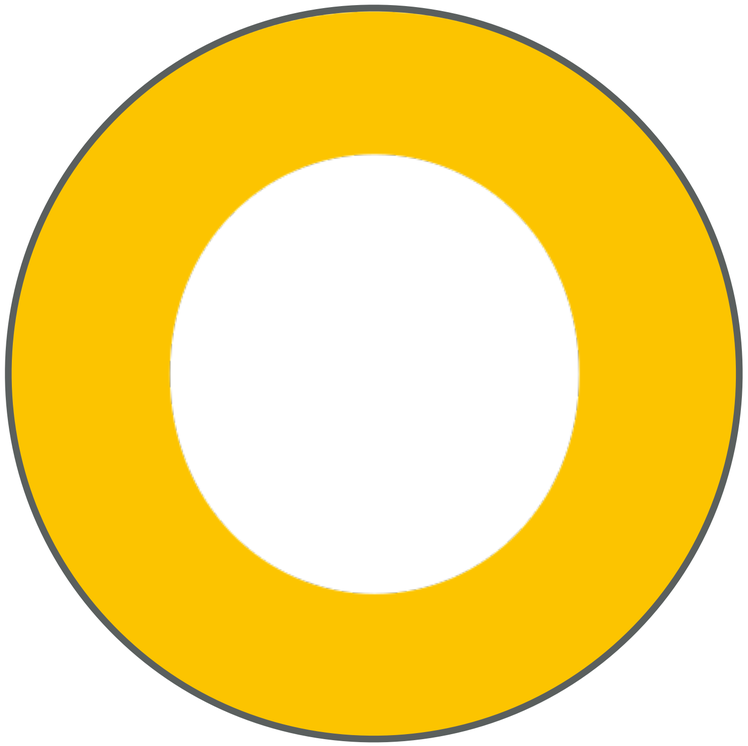
A | B testing

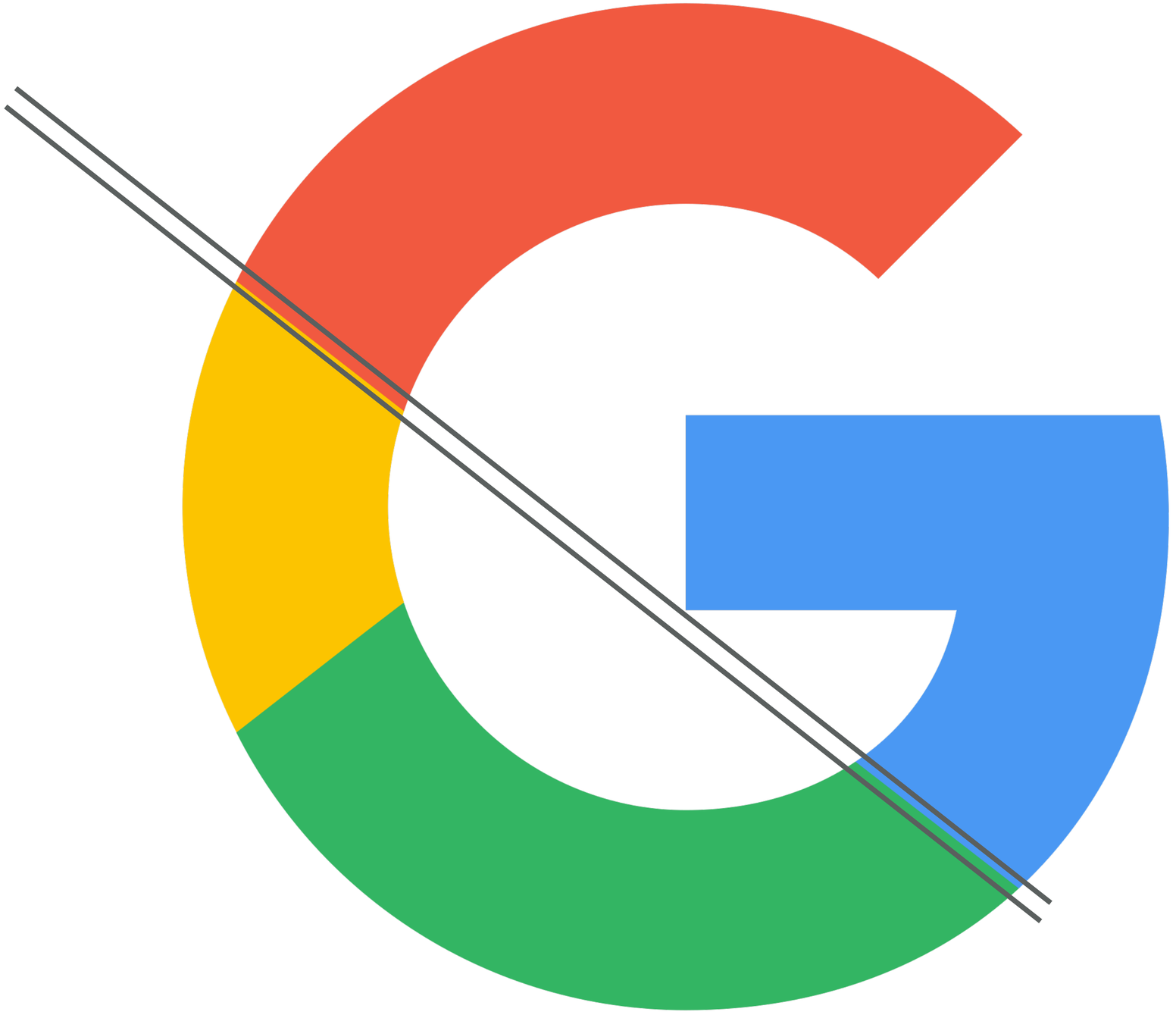
Before

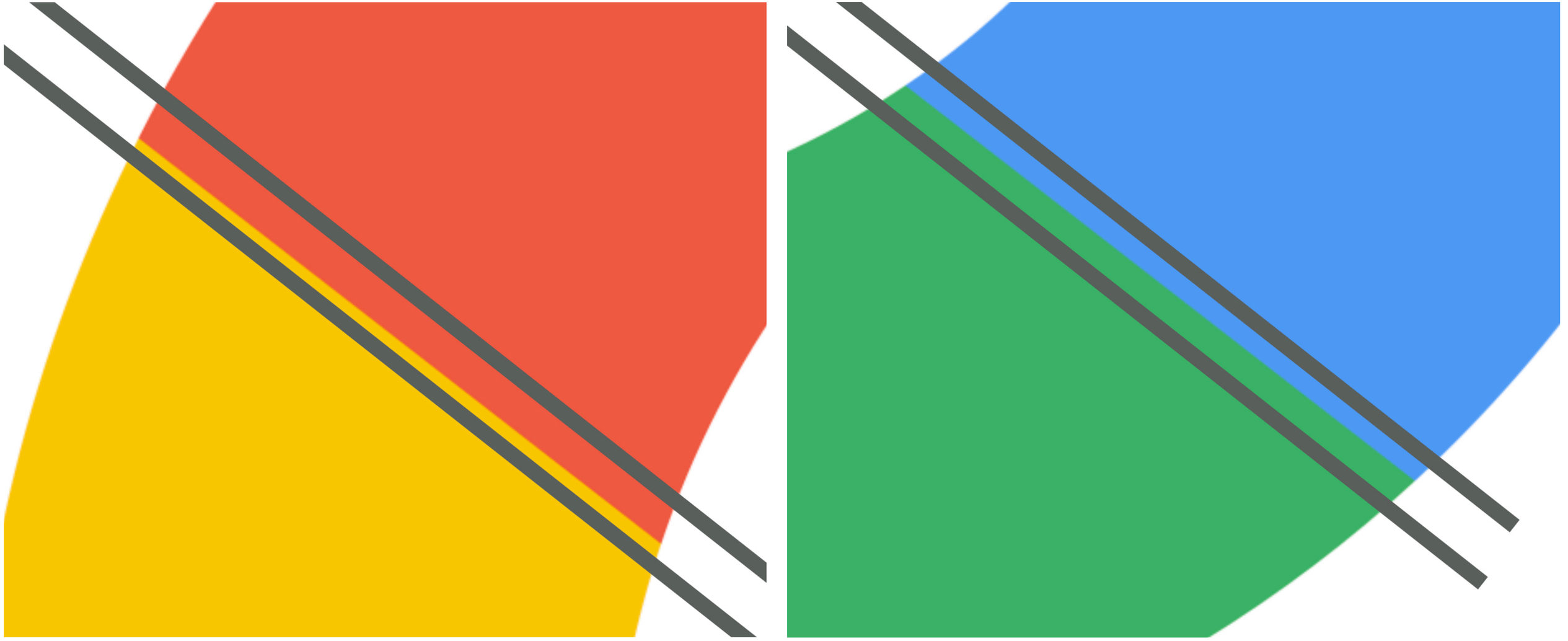


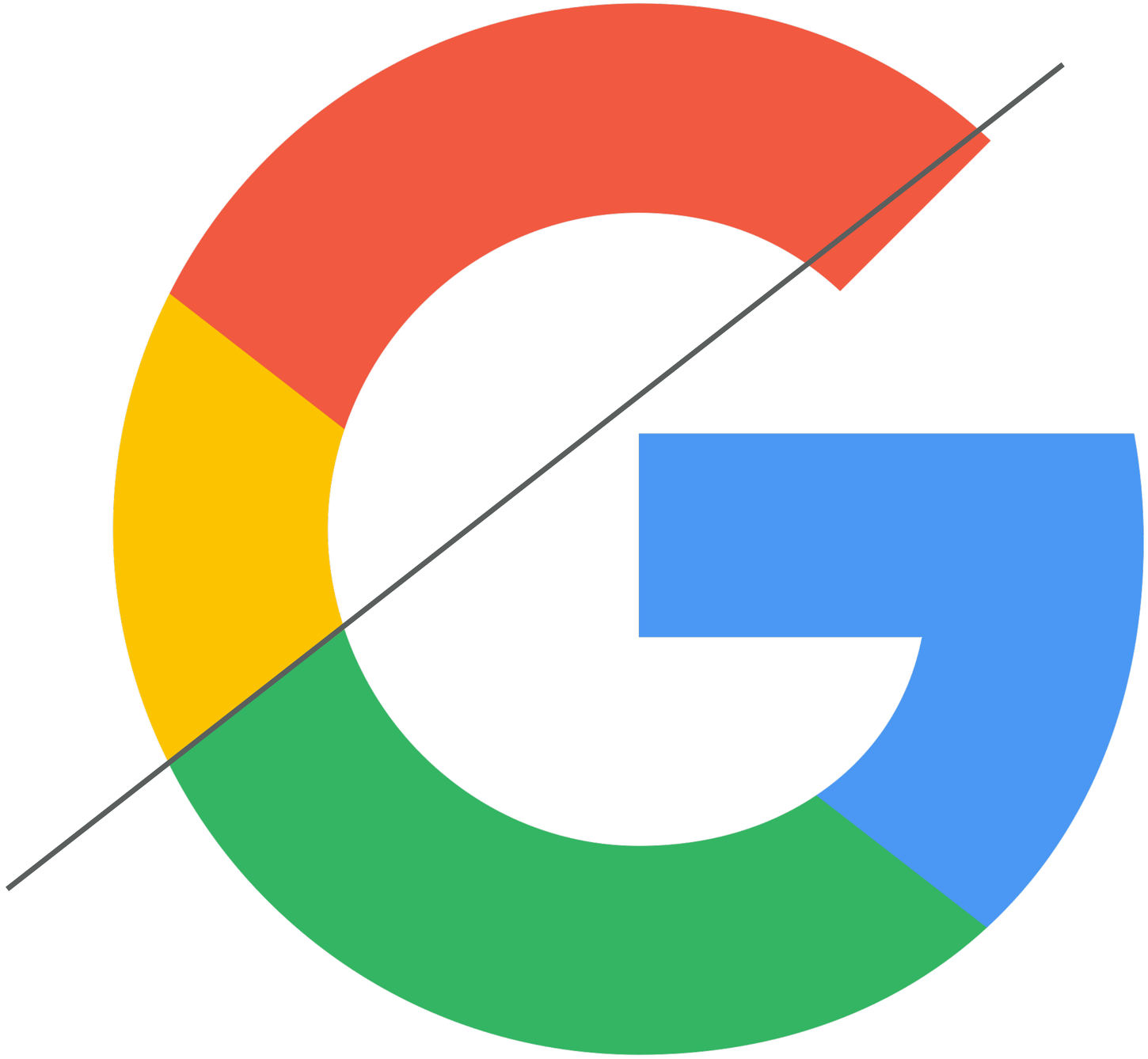












Large-scale experiments

The Facebook Study

PNAS
Proceedings of the National Academy of Sciences of the United States of America

Experimental evidence of massive-scale emotional contagion through social networks

Adam D. I. Kramer[a,1], Jamie E. Guillory[b,2], and Jeffrey T. Hancock[b,c]

[a]Core Data Science Team, Facebook, Inc., Menlo Park, CA 94025; and Departments of [b]Communication and [c]Information Science, Cornell University, Ithaca, NY 14853

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved March 25, 2014 (received for review October 23, 2013)

The Facebook Study

- Over 1B served/mo worldwide (in U.S., 128m users/day)
- Average user: ~1500 items eligible for news feed
- Algorithm selects & shows ~300 items (to see all, go to friend's wall)
- Has always existed
- FB constantly changes it
- Proprietary
- Now based on ~100K criteria
- Almost certainly affects how much emotionally charged content users see; very likely prioritizes ☺ posts

McGee M, "EdgeRank Is Dead: Facebook's News Feed Algorithm Now Has Close To 100K Weight Factors," *Marketing Land*, Aug 2013
http://marketingland.com/edgerank-is-dead-facebooks-news-feed-algorithm-now-has-close-to-100k-weight-factors-55908

UC San Diego

Source: Michelle Meyer

The Facebook Study: Three hypotheses

1. **Social Comparison** (less positive posts → more people sad)
   - Small, observational studies
   - Correlations b/w FB use and stress, jealously, loneliness, and depression
   - FB creates "self promotion-envy spiral"

2. **Emotional Contagion** (less negative posts → more people happy)
   - Lab experiments and field data found happiness and depression spread via in-person social networks

3. **Null Hypothesis**

Source: Michelle Meyer; Fowler & Christakis, "Dynamic spread of happiness in a large social network," *BMJ* 2008

The Facebook Study: Design

Social Comparison Arm
Experiment 1

Positive Posts Reduced

Emotional Contagion Arm
Experiment 2

Negative Posts Reduced

689,003
Users Split
into 4 groups
~155k per group

Control Group

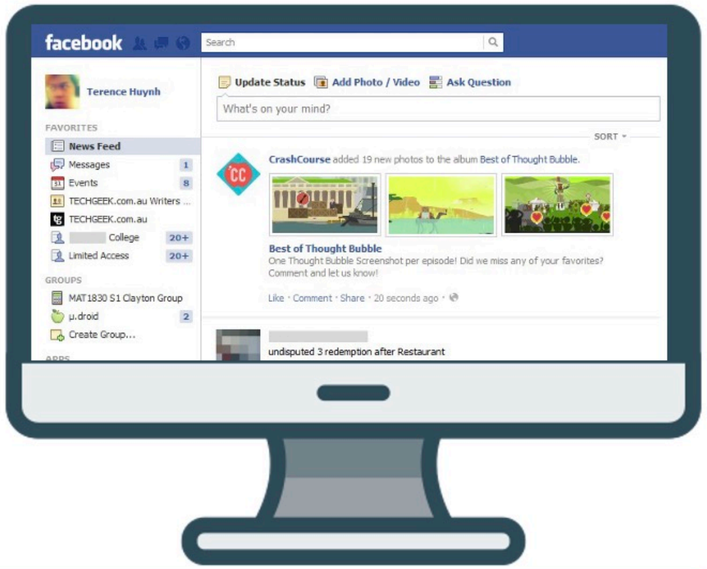
Control Group

Random Posts Removed

Random Posts Removed

The Facebook Study: Design

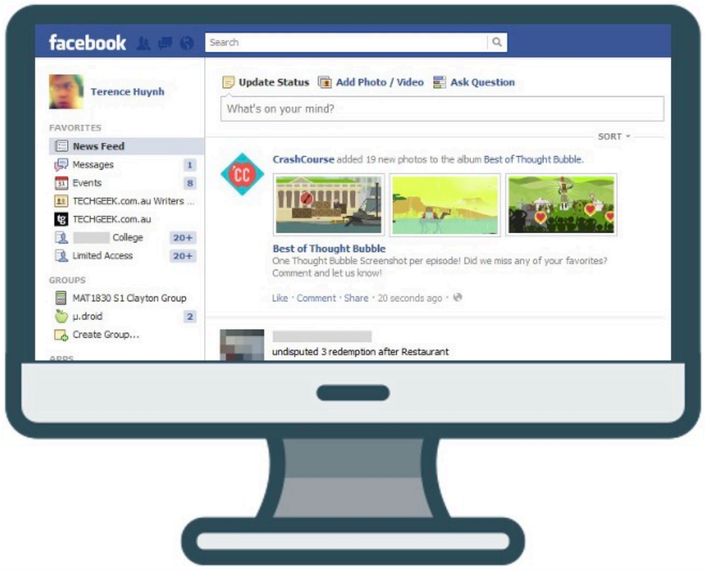
Social Comparison Arm



Positive Posts Reduced

~10%

Emotional Contagion Arm



Negative Posts Reduced

~10%

- 1 week (Jan. 11-18, 2012)
- Subjects: randomly selected by userID, from population of users who viewed FB in 'english'
- N = 689,003

The Raw Data

Post Coded
+/- for words

positive post

negative post

neutral post

The Facebook Study: Design

Two parallel experiments; 4 conditions (for each, N = ~155,000)

Experiment 1: Removing ☺ posts

- Treatment: Each ☺ post "had between a 10% & 90% chance (based on user ID) of being omitted…for that specific viewing."
- Control: 10%–90% of 46.8% (i.e., 4.68%-42.12%) of eligible posts randomly removed w/o regard to emotional content

Experiment 2: Removing ☹ posts:

- Treatment: Each ☹ post "had between a 10% & 90% chance (based on user ID) of being omitted…for that specific viewing."
- Control: 10%–90% of 22.4% (i.e., 2.24%-20.16%) of eligible posts randomly removed w/o regard to emotional content

Source: Michelle Meyer; Kramer, Guillory, Hancock, "Experimental evidence of massive-scale emotional contagion through social networks," *PNAS* 2014

The Facebook Study: Results

Compared to control subjects, subjects exposed to fewer ☺ posts subsequently in their own posts:

- Used 0.1% fewer ☺ words (Cohen's d = 0.02)
- Used 0.04% more ☹ words (Cohen's d = 0.001)
- Produced only 96.7% as many words overall

Compared to control subjects, subjects exposed to fewer ☹ posts subsequently in their own posts:

- Used 0.07% fewer ☹ words (Cohen's d = 0.02)
- Used 0.06% more ☺ words (Cohen's d = 0.008)
- Produced only 99.7% as many words overall

Source: Michelle Meyer; Kramer, Guillory, Hancock, "Experimental evidence of massive-scale emotional contagion through social networks," *PNAS* 2014

The Facebook Study: Results

Issues of methodology & interpretation:

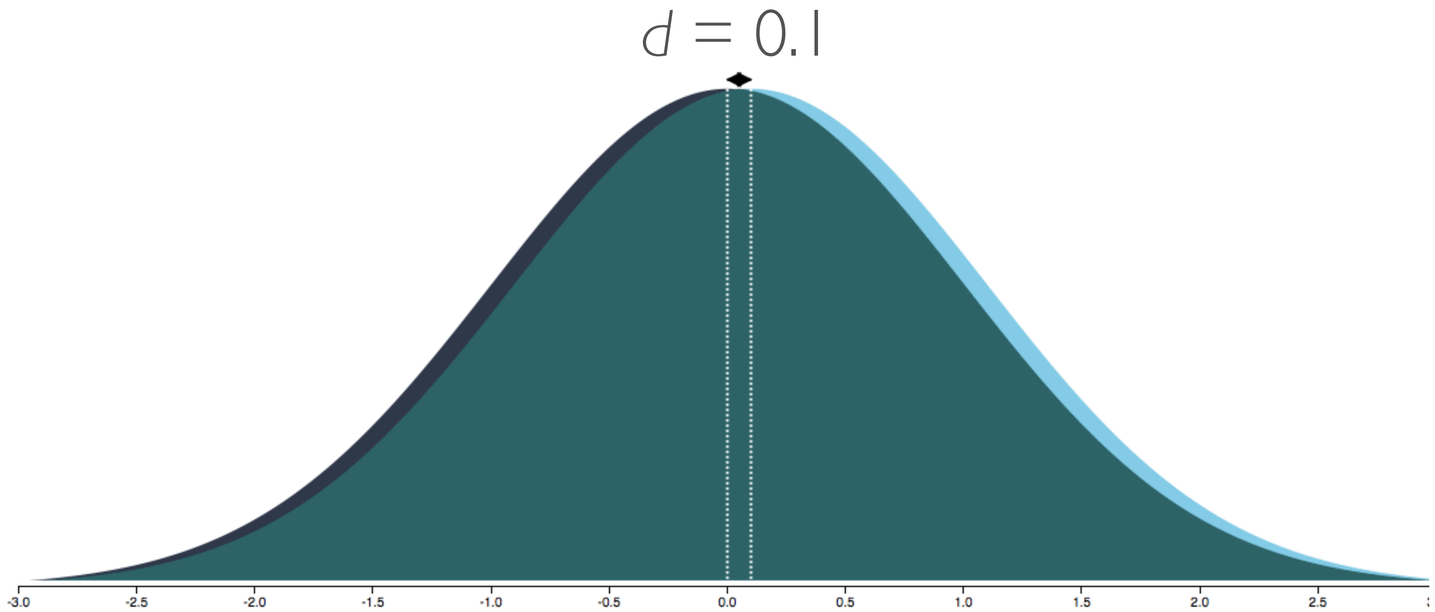
- Questionable instrument: LIWC 2007
    - not intended for lengthy text
- Questionable coding: "I'm not having a great day." "Oh great."

UC San Diego

Source: Michelle Meyer

The Facebook Study: Results

"First, these effects, while highly statistically significant, are tiny. The largest effect size reported had a Cohen's d of 0.02–meaning that eliminating a substantial proportion of emotional content from a user's feed had the monumental effect of shifting that user's own emotional word use by two hundredths of a standard deviation."



Source: Tal Yarkoni

UC San Diego

The Facebook Study: Results

"To put it in intuitive terms, the effect of condition in the Facebook study is roughly comparable to a hypothetical treatment that increased the average height of the male population in the United States by about one twentieth of an inch (given a standard deviation of ~2.8 inches). Theoretically interesting, perhaps, but not very meaningful in practice."

Source: Tal Yarkoni