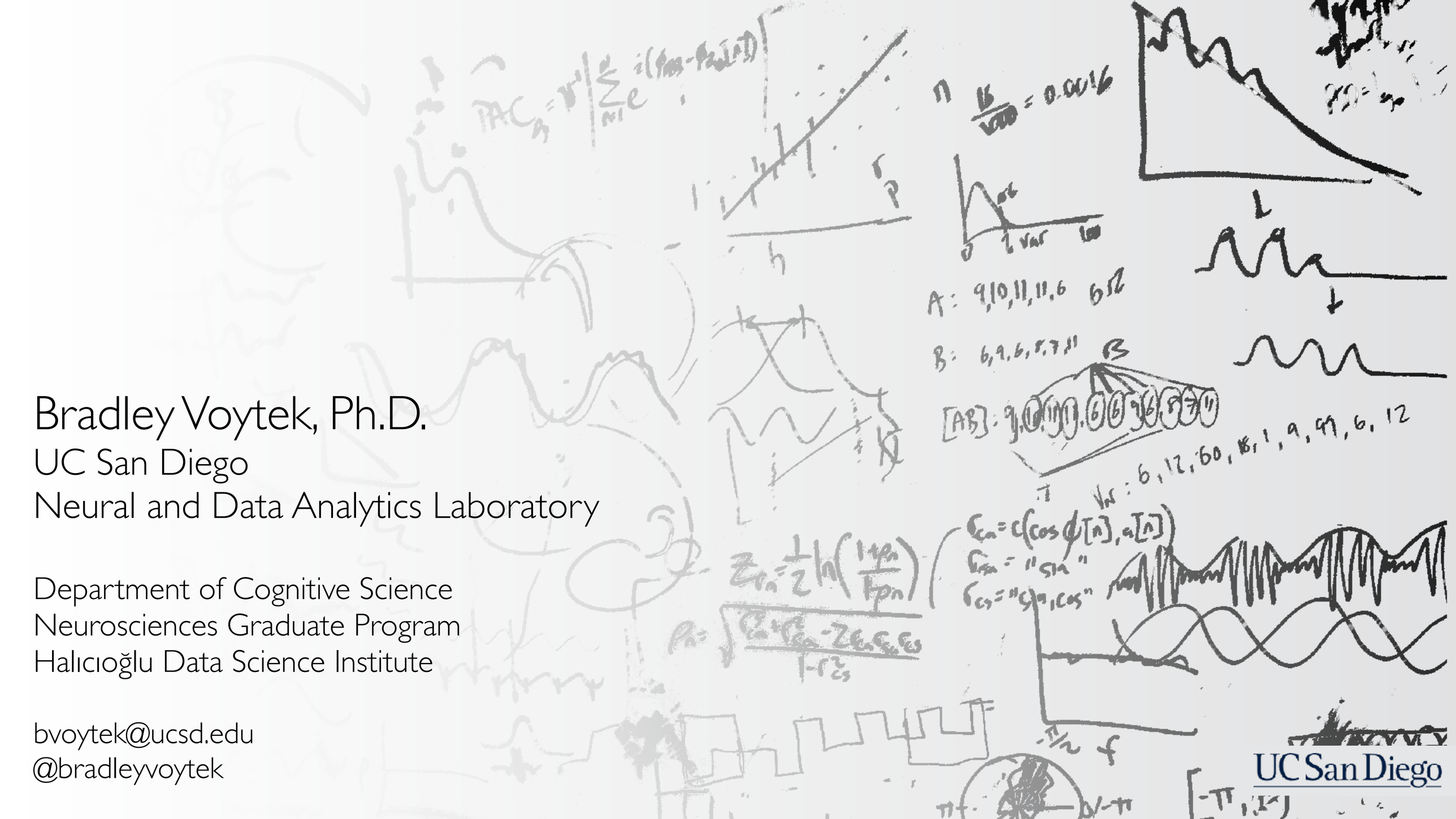


Bradley Voytek, Ph.D.

UC San Diego
Neural and Data Analytics Laboratory

Department of Cognitive Science
Neurosciences Graduate Program
Halıcıoğlu Data Science Institute

bvoytek@ucsd.edu
@bradleyvoytek

UC San Diego

COGS 108
Data Science in Practice

THE LAST LECTURE

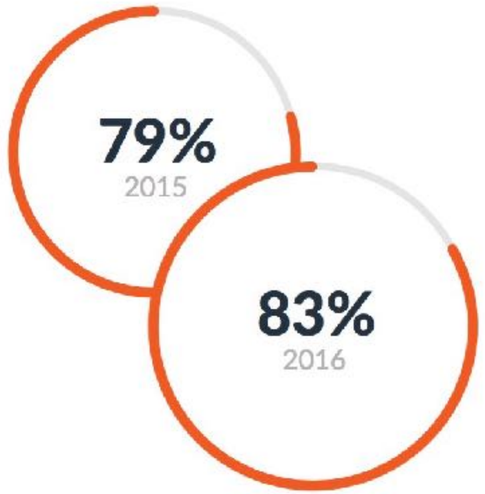
UC San Diego

State of Data Science

There's a Still a Shortage of Data Scientists (And it Might Be Getting Worse)

Last year, we found that 79 percent of respondents said that there were a shortage of data scientists in the field. And while that was staggering, our survey found that in 2016, things might be getting worse.

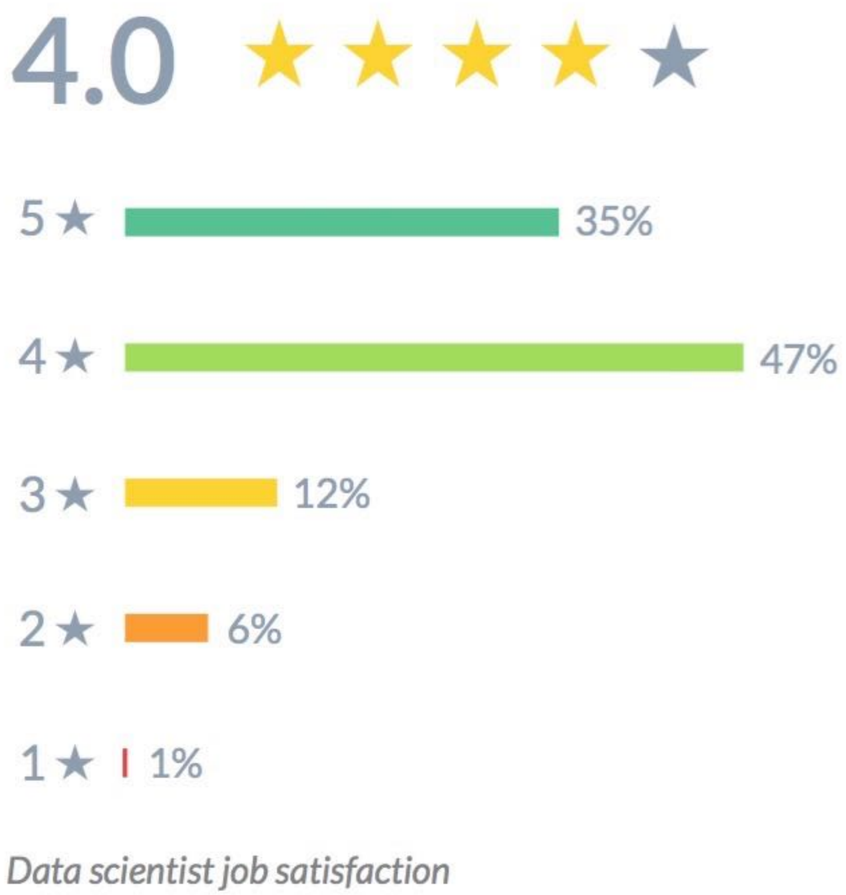
A full 83% of respondents said there weren't enough data scientists to go around. And with more and more enterprises and organization investing in data this trend is likely to continue.



Respondents who said there weren't enough data scientists to go around

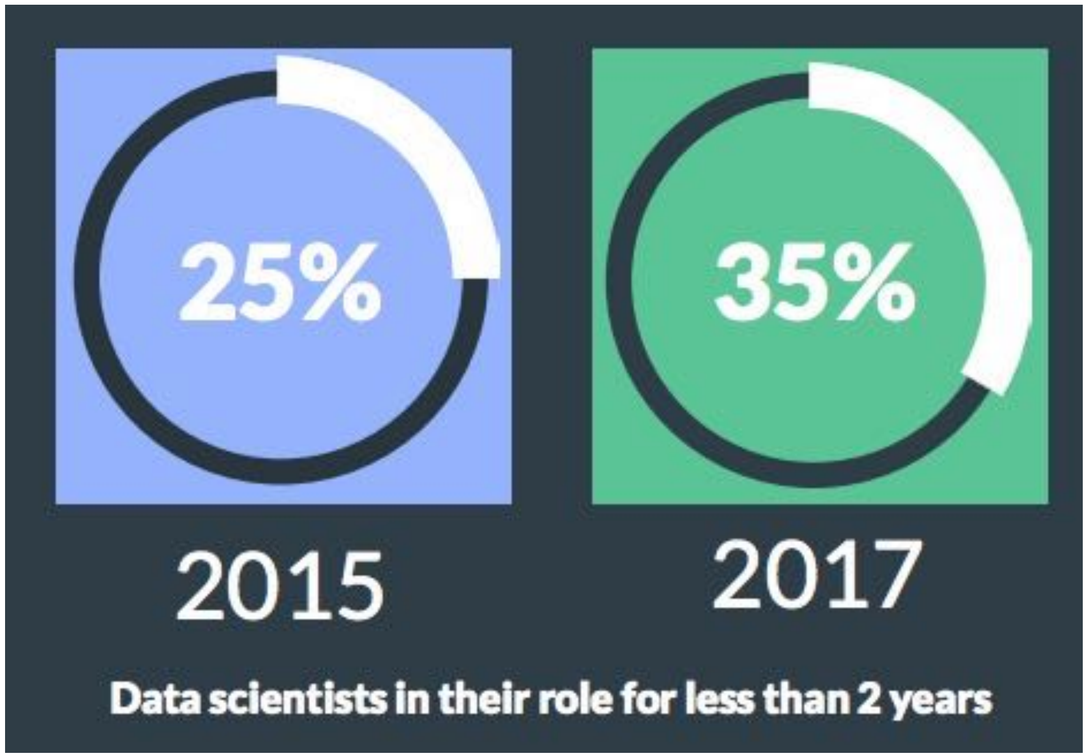
Source: CrowdFlower, *Data Science Report* 2016

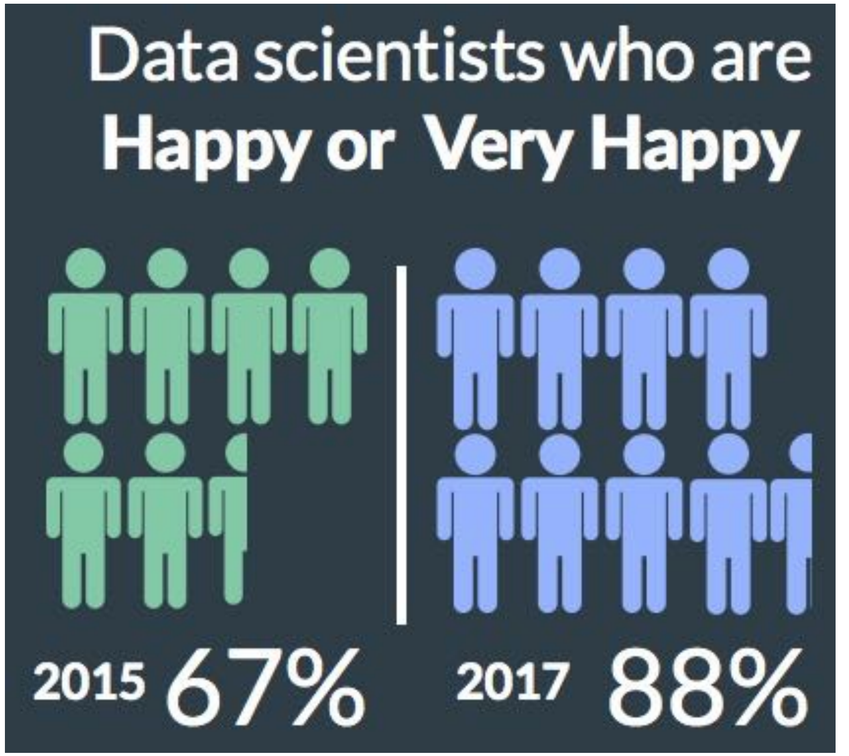
State of Data Science



Source: CrowdFlower, *Data Science Report* 2016

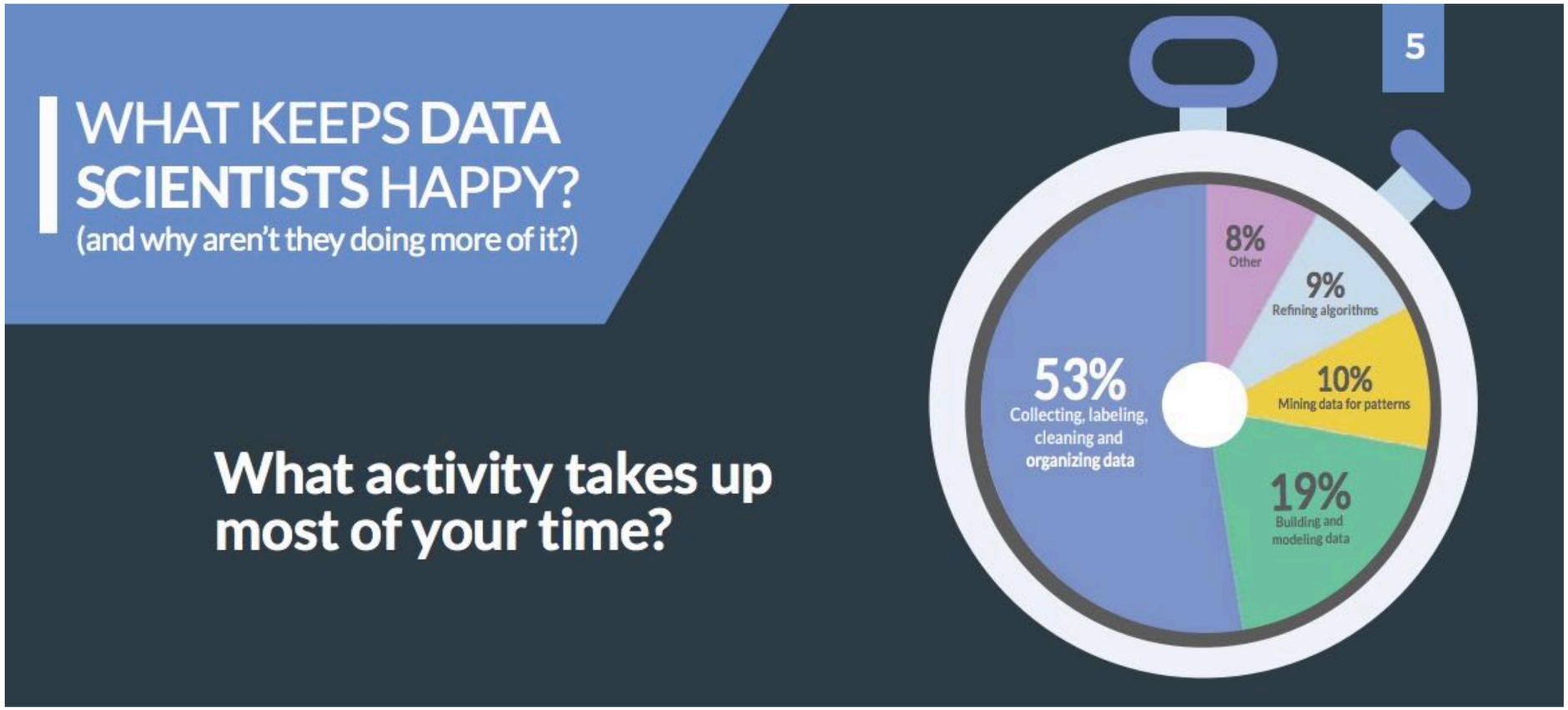
State of Data Science





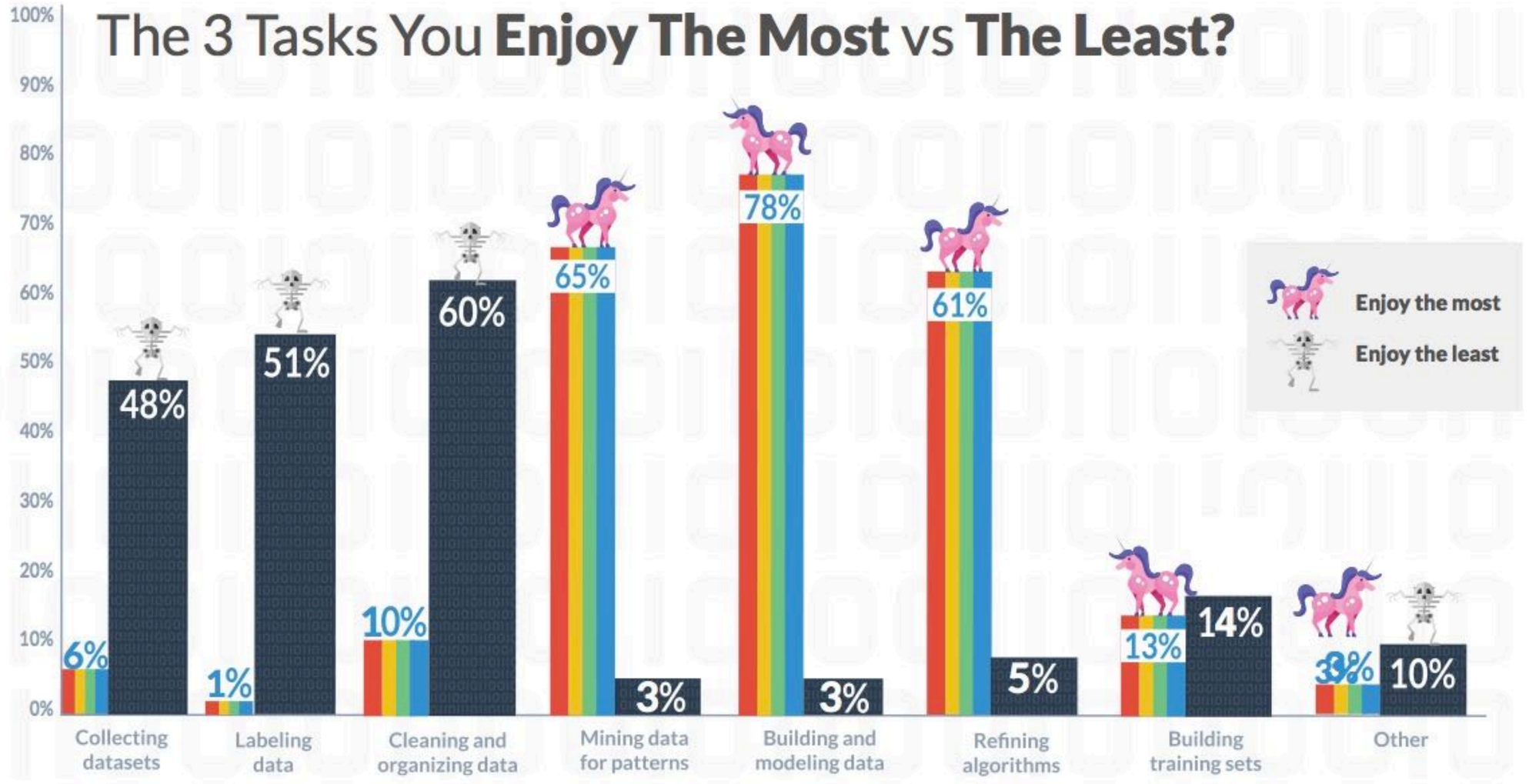
Source: CrowdFlower, *Data Science Report* 2017

State of Data Science



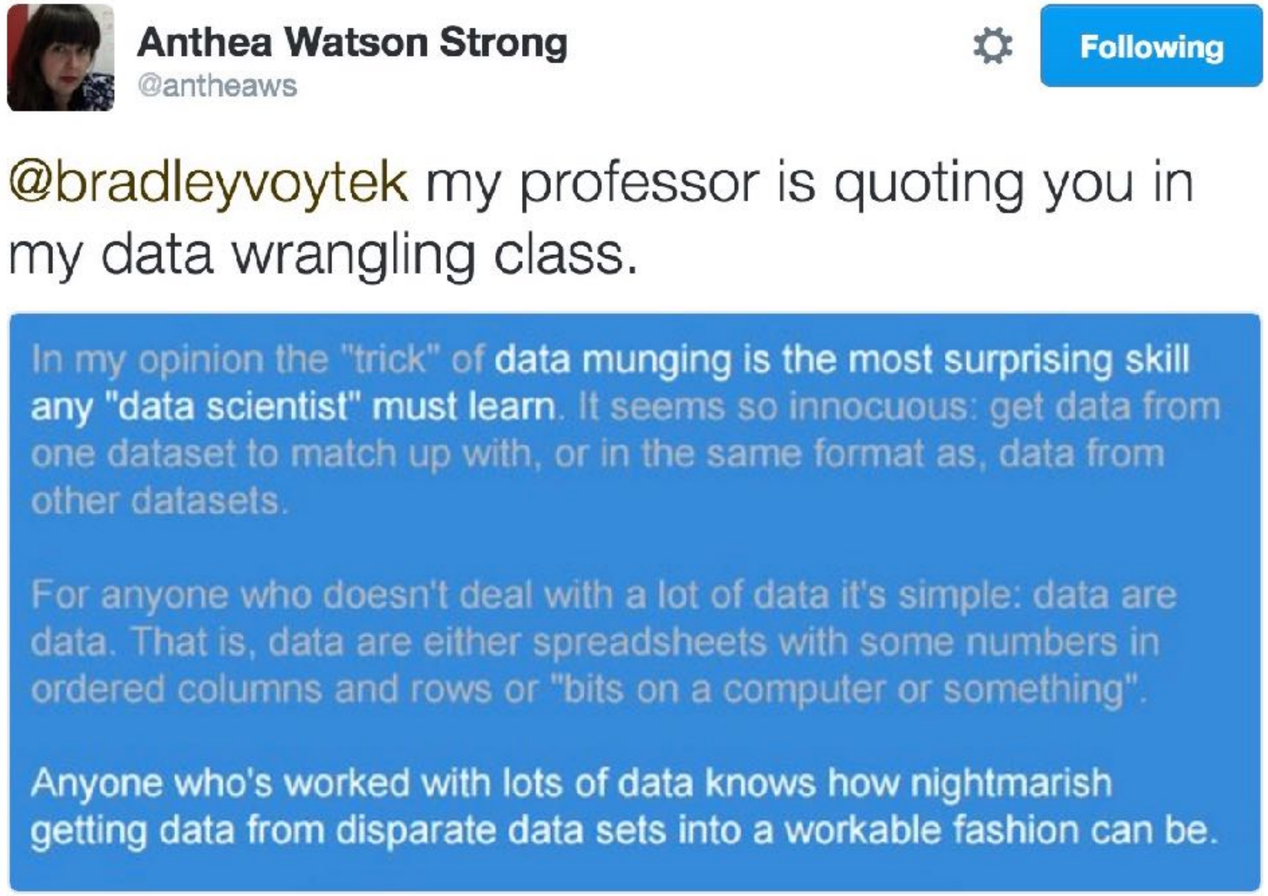
Source: CrowdFlower, *Data Science Report* 2017

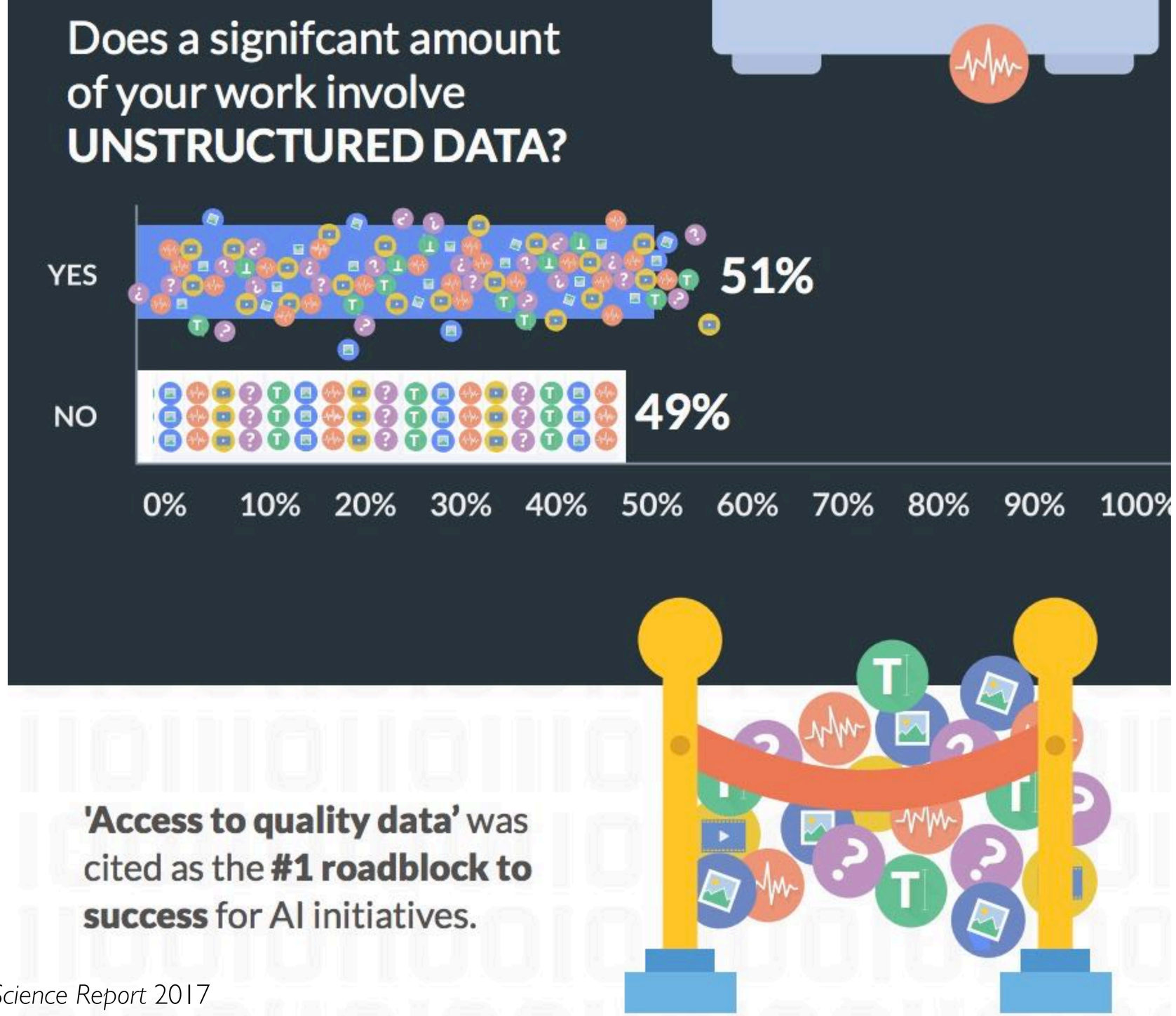
Data science reality



Source: CrowdFlower, *Data Science Report* 2017

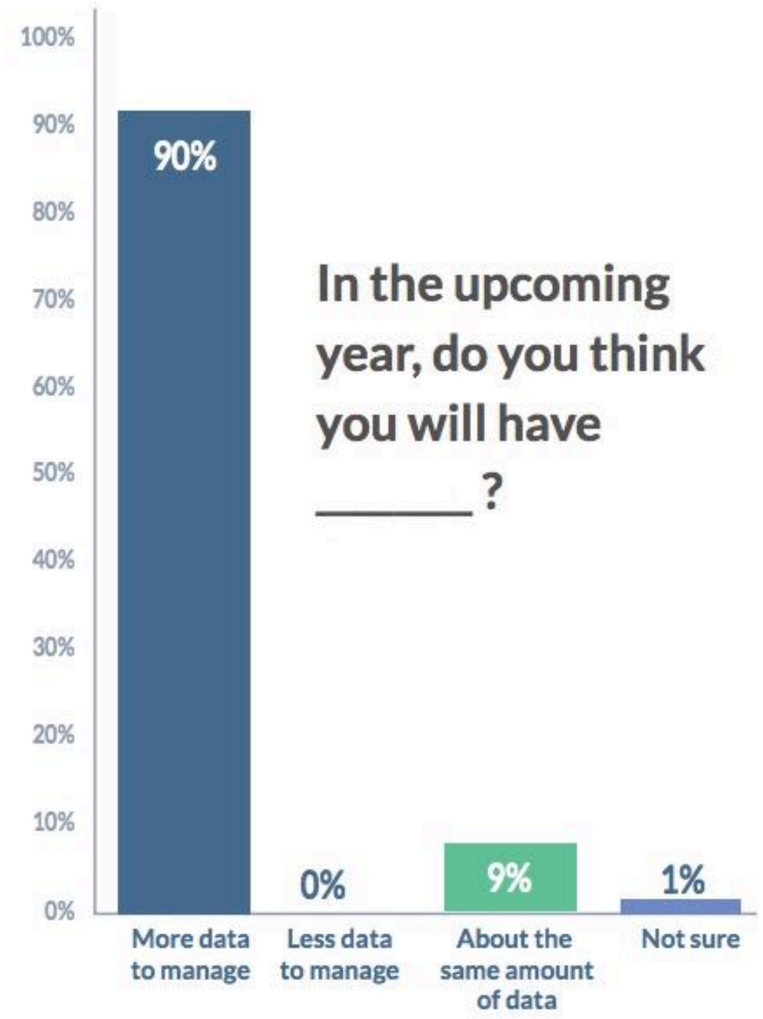
Data science reality





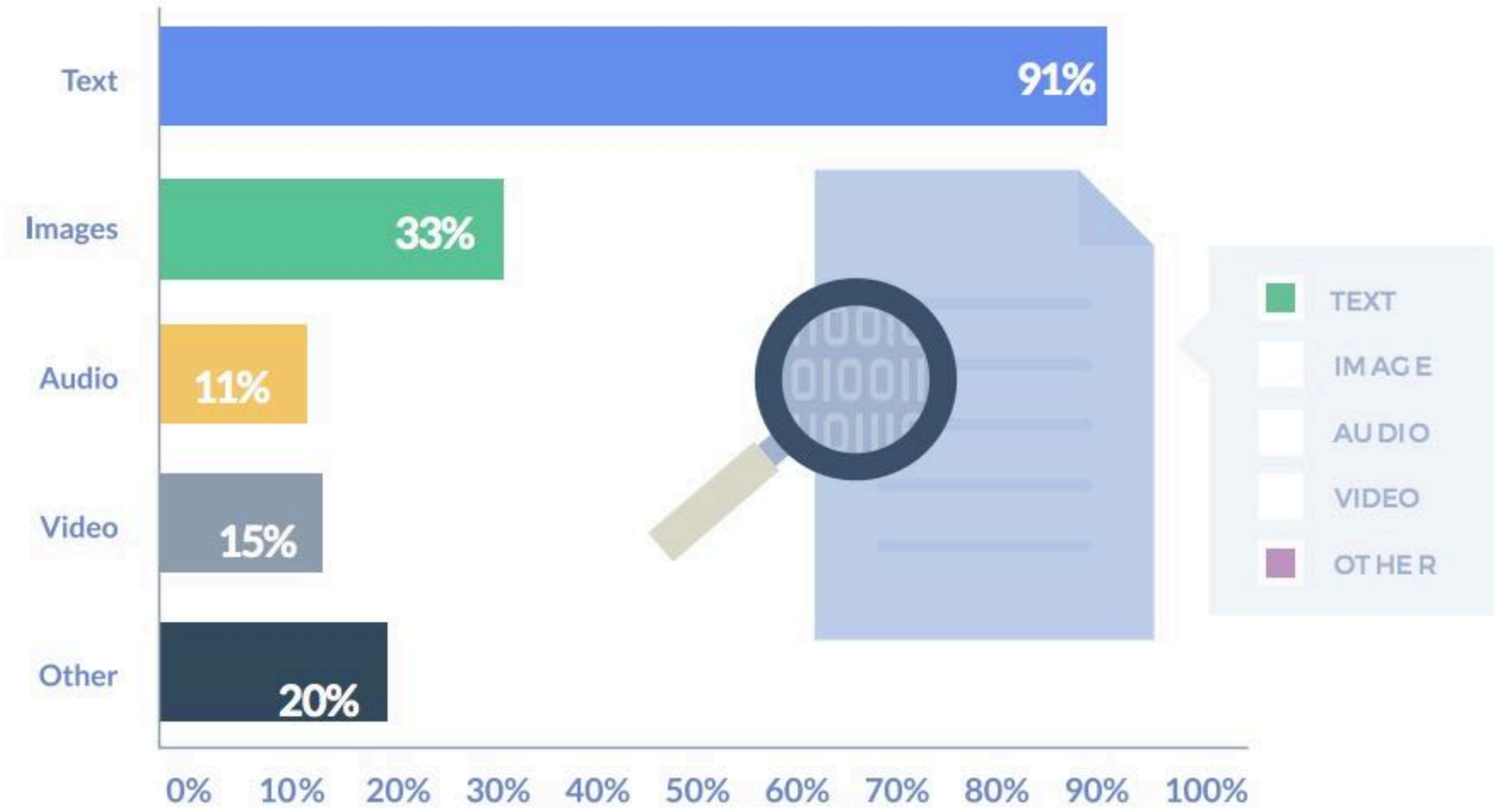
Source: CrowdFlower, *Data Science Report* 2017

State of Data Science



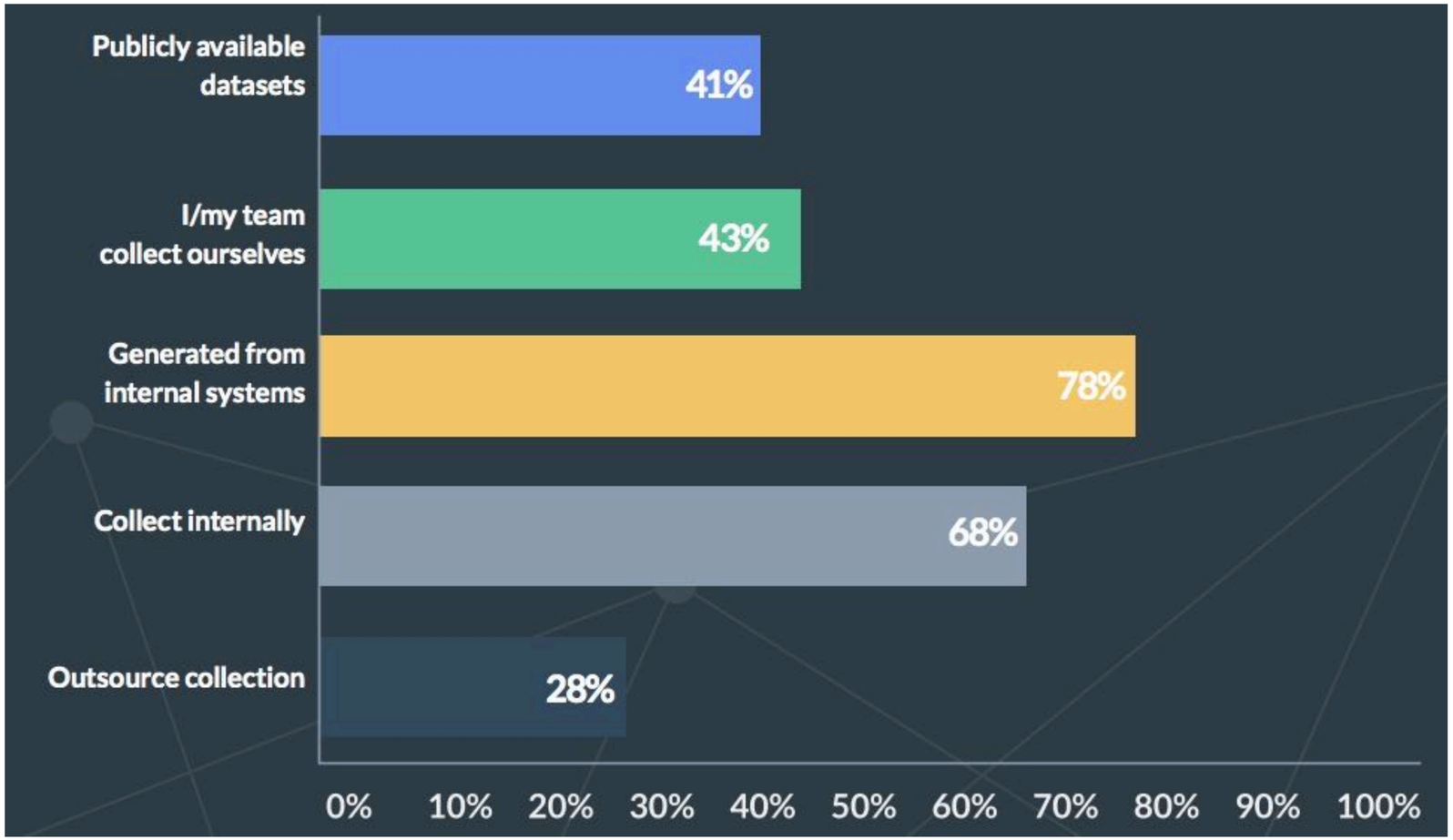
Source: CrowdFlower, *Data Science Report* 2017

State of Data Science



Source: CrowdFlower, *Data Science Report* 2017

State of Data Science



Source: CrowdFlower, *Data Science Report* 2017

Data Science

50 Best Jobs in America

This report ranks jobs according to each job's Glassdoor Job Score, determined by combining three factors: number of job openings, salary, and overall job satisfaction rating.

United States | 2018

0 Shares

1 Data Scientist

4.8 / 5
Job Score

4.2 / 5
Job Satisfaction

$110,000
Median Base Salary

4,524
Job Openings

View Jobs

Source: Glassdoor, 2018

Data Science skills

The Ten Most Common Data Science Skills in Job Postings

Skill	Percentage of Job Listings
Python	**72%**
R	**64%**
SQL	**51%**
Hadoop	**39%**
Java	**33%**
SAS	**30%**
Spark	**27%**
Matlab	**20%**
Hive	**17%**
Tableau	**14%**

Source: Glassdoor Economic Research.

glassdoor

Three Data Scientist Personas and What They Earn

	Skills Likely to Have	Percentage of Data Science Jobs	Average Estimated Salary
Core Data Scientist	Python, R, SQL	**71%**	**$116,203**
Researcher	SAS, Matlab, Java, Hadoop, Python, R	**15%**	**$112,346**
Big Data Specialist	Spark, Hive, Hadoop, Java, Python	**14%**	**$121,246**

Source: Glassdoor Economic Research.

glassdoor

UC San Diego

Source: Glassdoor, 2017

Data Science at UC San Diego

- **What might Data Science look like at UC San Diego?**

Data Science at UC San Diego

- Hiring full-time data visualization, machine learning, and software engineers to build out open source tools tutorials and intuitive visualizations for learning data science, to be shared openly and publicly for the world to access (this is to help establish UCSD as a leader in the field). Examples of wonderful interactive visualizations: Simpson's Paradox; Decision Trees. This would bring in in-house experts, and also bring good faith and legitimacy to UCSD, putting us on the map as the place to be for cutting-edge Data Science.

- Scholarships for undergrads and grads to work closely with data science faculty—with at least one scholarship awarded to students from each Division to ensure breadth of representation.

- Office space (and honorarium?) for a Data Scientist-in-residence. This is to bring in people from industry to spend several months to a year on site. This will help build industry relations, bring in new skills for students, and help faculty keep curricula up to date and fresh with relevant and topical tools and ideas.

Data Science at UC San Diego

- Data Journalist-in-residence, for the same reasons.
- The reverse of the Data Scientist/Journalist-in-residence where we can establish Data Science sabbaticals at various companies for our students, post-docs, staff, and faculty.
- A Data Science Incubator.
- Money for workshops, conferences, and guest speakers, recruiters, etc.
- Training: for example we could get someone trained on Software Carpentry to roam departments, offering workshops (coordinating with the library, etc.), or at least have some core admin to organize things like this.
- Data Science Summer Programs for local high school and community college students, with scholarships for underrepresented students.

COGS 108

- This is very much a work in progress.
- Thank you all for your patience, feedback, and time!

COGS 108

- **CAPE! http://cape.ucsd.edu/**

What is the point of this class?

- Where did we start from, 10 weeks ago?

What is the point of this class?

- Prediction
- Classification
- Knowledge discovery?
- **DOING USEFUL SHIT**

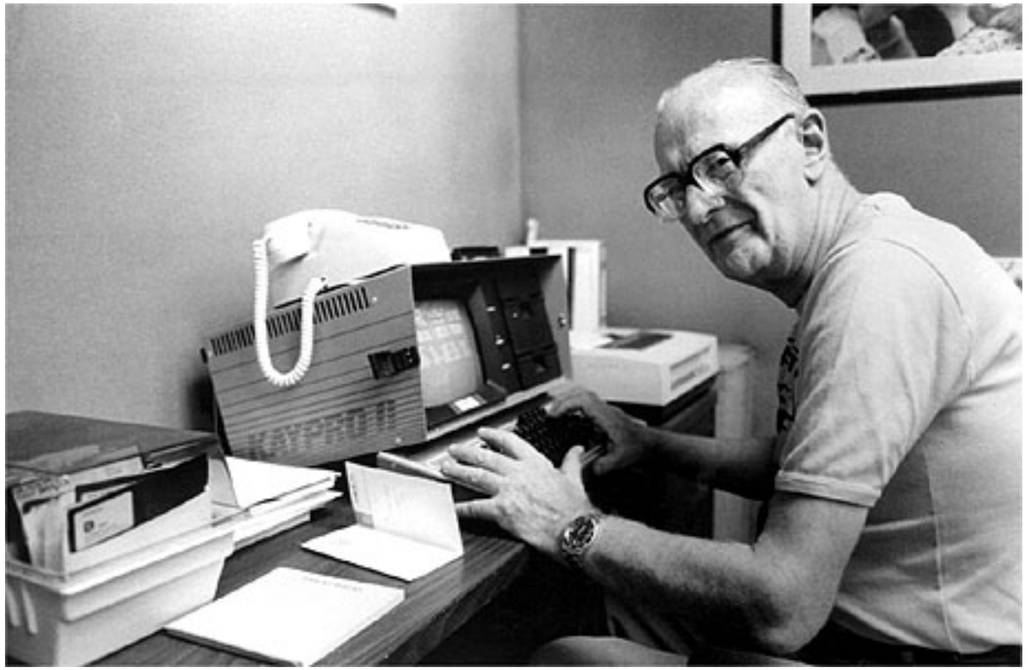
Proposed course order

1. Introduction: Why data analysis? (prediction and classification)
2. Python!
3. Data Science in Python (jupyter, pandas, numpy, scipy, scikit-learn, etc.)
4. Data gathering, wrangling, and cleaning (How do you find and clean data? (JSON, CSV, XML, SQL, APIs)
5. Data privacy, ethics, and HIPAA (anonymization)
6. **Jan 25** Guest lecture: Kevin Novak: Chief Data Officer, Tala (Formerly: Head of Data & Engineering, *Uber*)
7. Basic data visualization
8. Data intuition and the "sniff test" (Fermi estimation; distributions and outliers: histograms, CDF, PDFs)
9. Non-parametric statistics
10. Linear modeling
11. **Feb 13** Ilkay Altintas, PhD: Chief Data Science Officer, San Diego Supercomputer Center (SDSC)
12. NO CLASS!
13. OLS (optimization)
14. Multiple linear regression and collinearities
15. **Feb 27** Josh Wills: Director of Data Engineering, Slack (Formerly: Director of Data Science, Cloudera; Analytics, Google)
16. Model validation (bootstrapping, resampling, k-fold, leave-p-out, train/test)
17. Dimensionality reduction (PCA); clustering and classification (k-means, knn, SVM)
18. Feature selection
19. NLP and text-mining (bag of words, tf-idf, sentiment analysis)
20. Geospatial analysis

Where are we now?

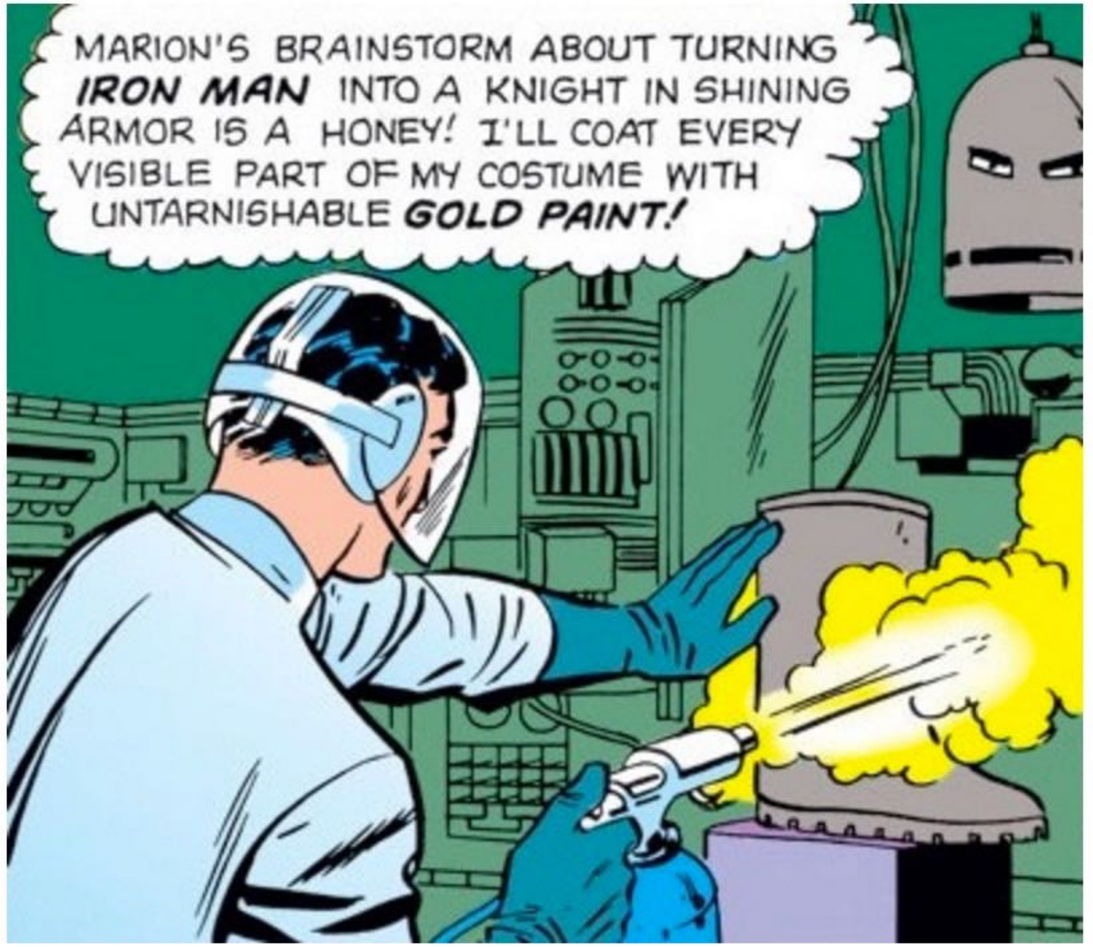
Source: NPR

Clarke's Laws



3. Any sufficiently advanced technology is indistinguishable from magic.

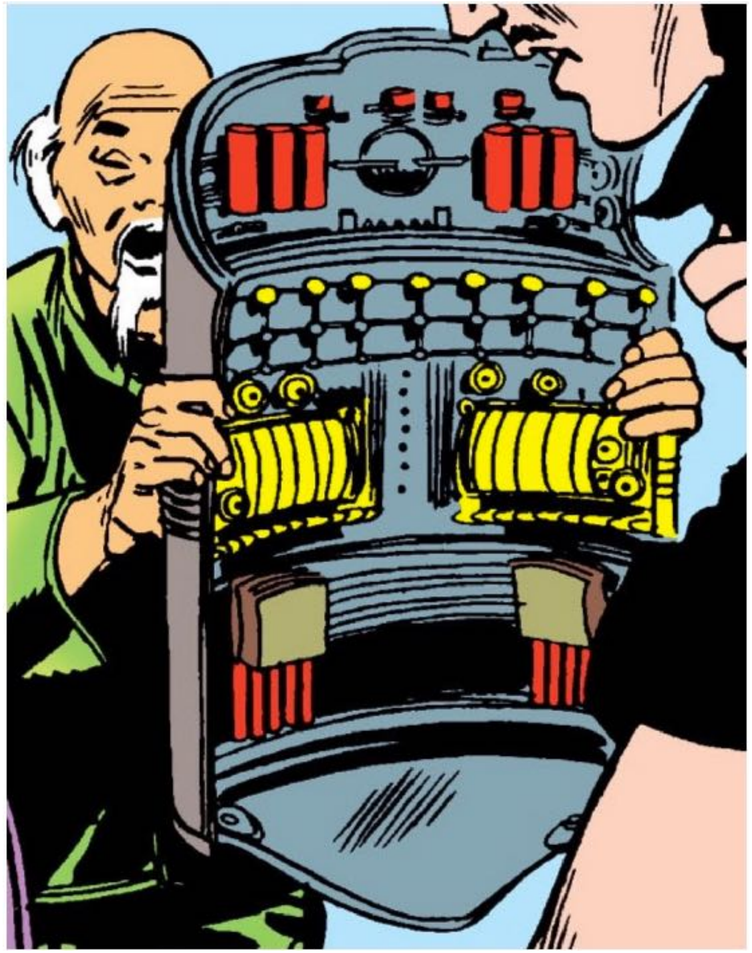
Iron Man!

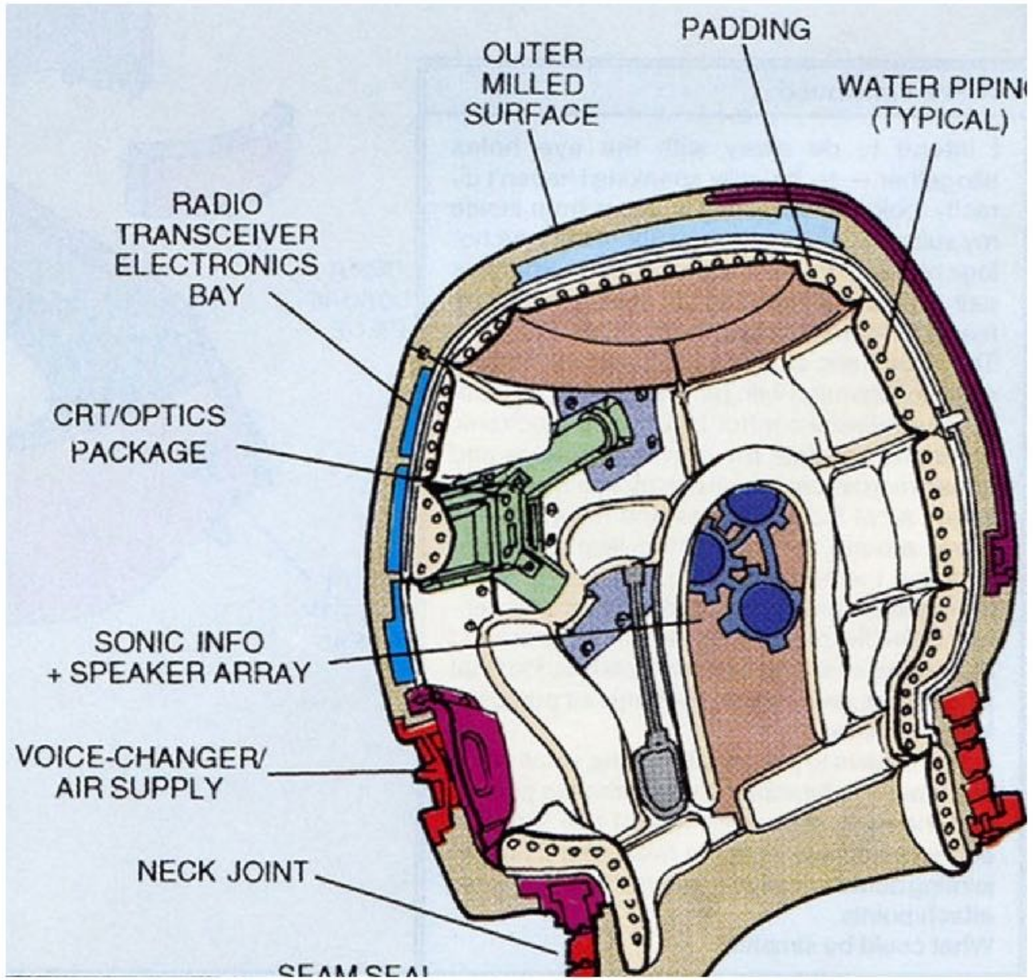


UC San Diego

Source: "Iron Man Armor Model 1", *Marvel Wikia*

Iron Man!



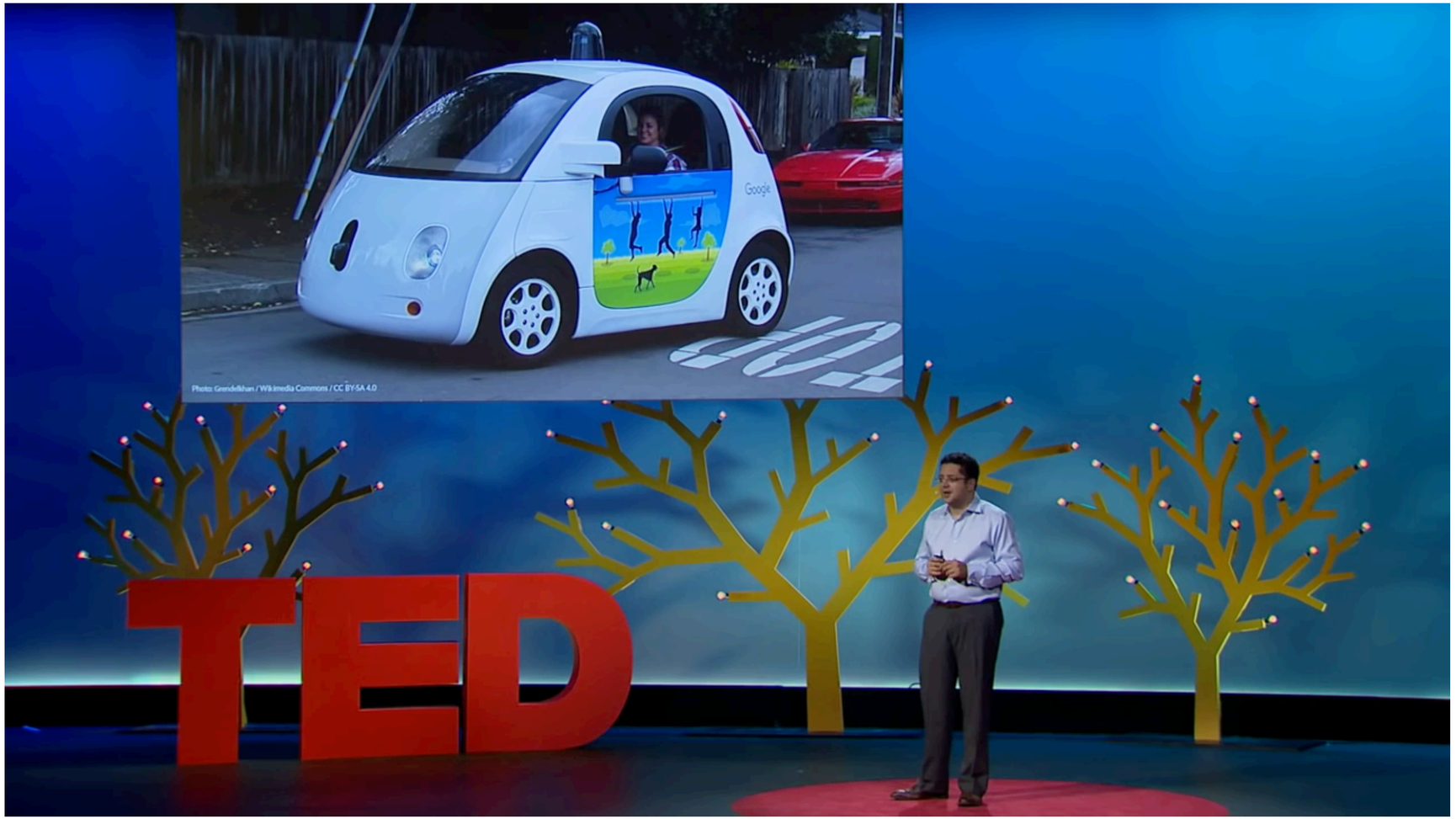


Source: "Iron Man Armor Model 1", *Marvel Wikia*

Iron Man!

Source: "Iron Man Armor Model 37", *Marvel Wikia*

Self-driving cars

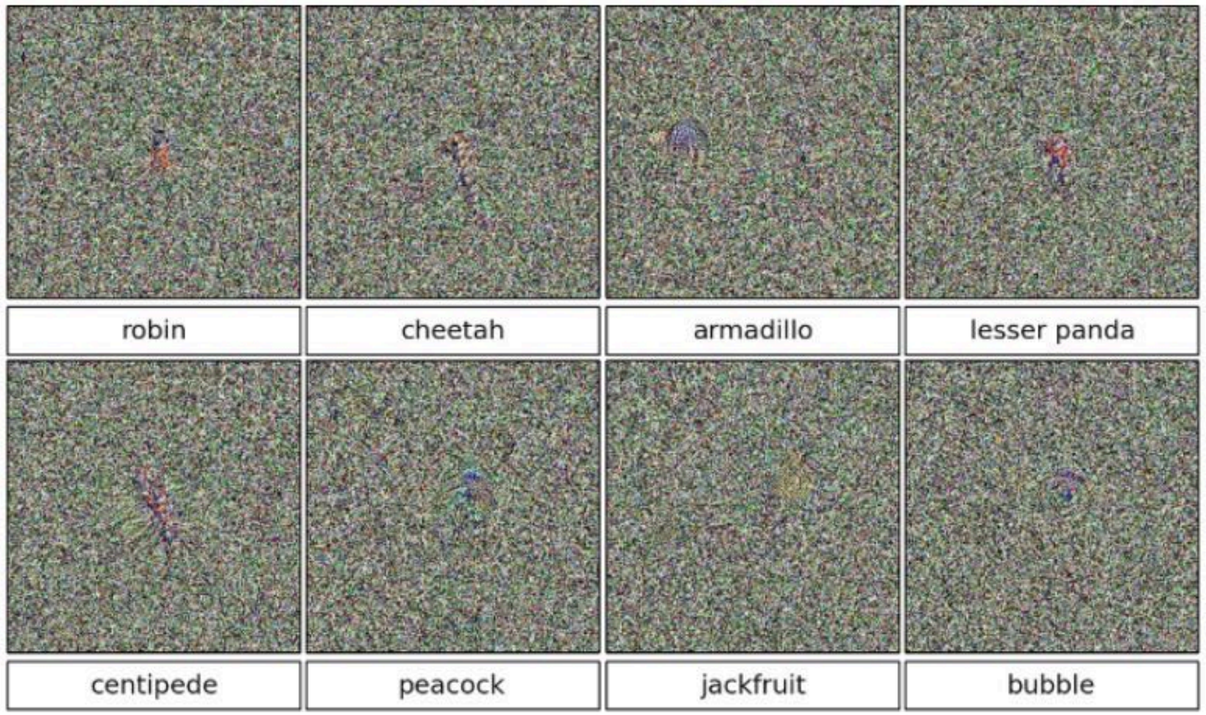


Source: Wanis Kabbaj, TED

Self-driving cars

Source: James Bridle (artist)

Data and Deep Learning



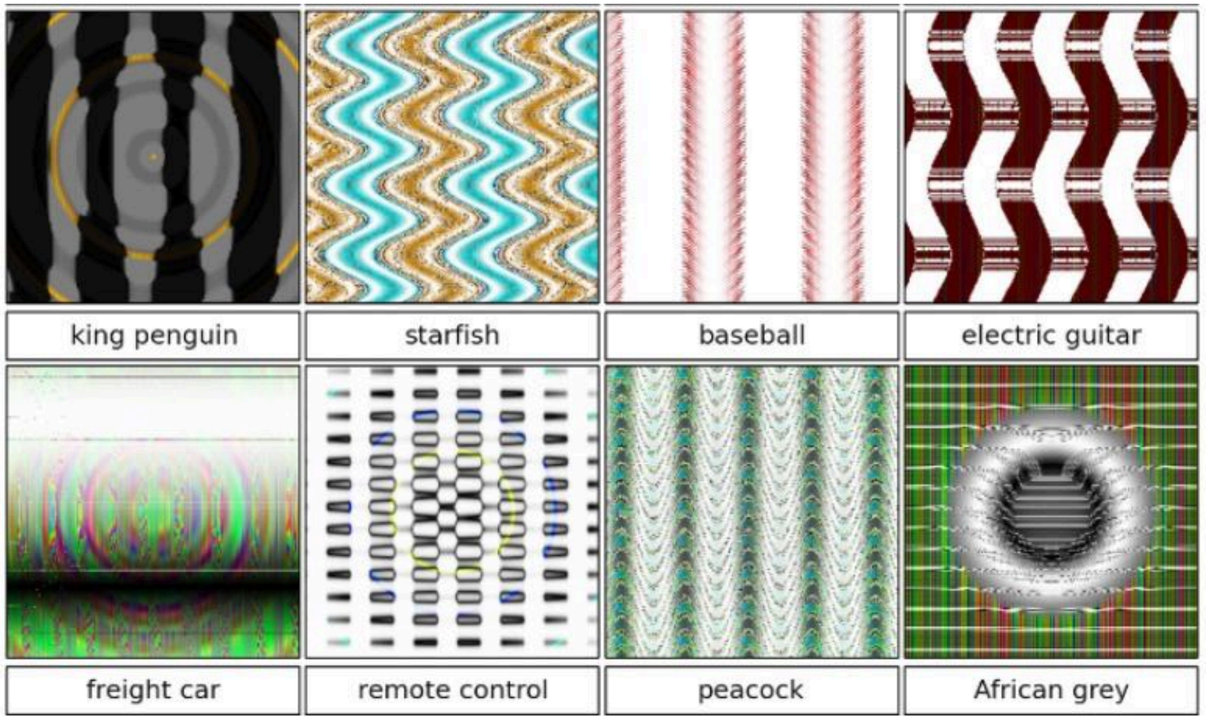


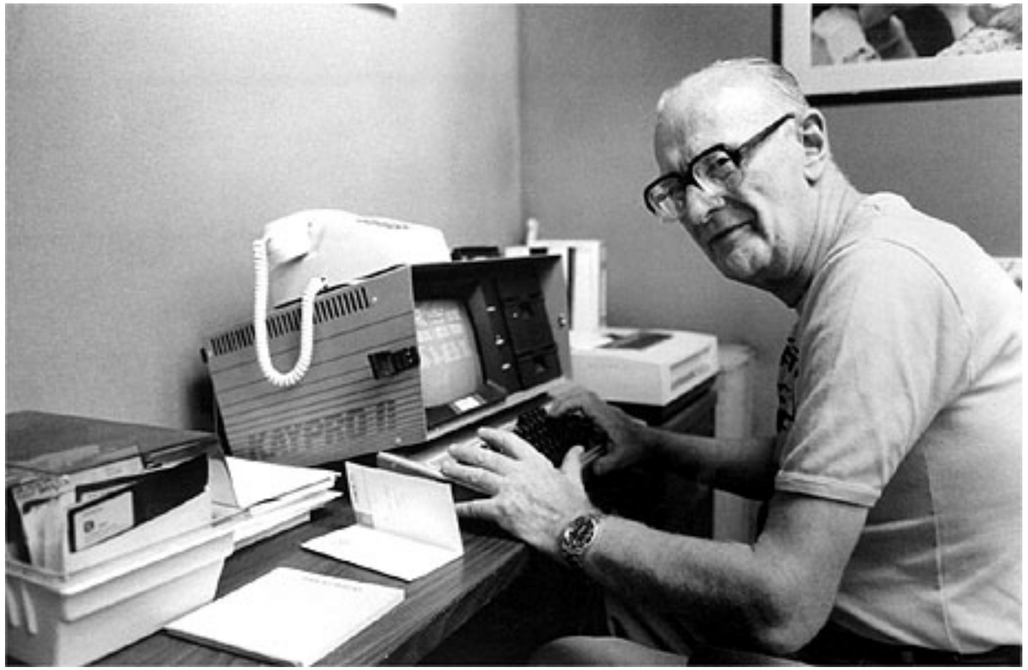
Figure 1. Evolved images that are unrecognizable to humans, but that state-of-the-art DNNs trained on ImageNet believe with $\geq$ 99.6% certainty to be a familiar object. This result highlights differences between how DNNs and humans recognize objects. Images are either directly (*top*) or indirectly (*bottom*) encoded.

Source: Nguyen, Yosinski, Clune, *IEEE CVPR* 2015

Data and Deep Learning

Source: Goodfellow, Shlens, Szegedy, *ICLR* 2015

“…indistinguishable from magic.”



Algorithms are fragile

Algorithms

The Making of a Fly: The Genetics of Animal Design (Paperback)
by Peter A. Lawrence

‹ Return to product information

Always pay through Amazon.com's Shopping Cart or 1-Click.
Learn more about Safe Online Shopping and our safe buying guarantee.

Price at a Glance

List Price: ~~$70.00~~
Used: from **$35.54**
New: from **$1,730,045.91**

Have one to sell? Sell yours here

All | New (2 from $1,730,045.91) | Used (15 from $35.54)

Show ⊙ New ○ Prime offers only (0) — Sorted by Price + Shipping

New 1-2 of 2 offers

Price + Shipping	Condition	Seller Information	Buying Options
$1,730,045.91 + $3.99 shipping	**New**	Seller: **profnath** Seller Rating: 93% positive over the past 12 months. (8,193 total ratings) In Stock. Ships from NJ, United States. Domestic shipping rates and return policy. Brand new, Perfect condition, Satisfaction Guaranteed.	Add to Cart or Sign in to turn on 1-Click ordering.
$2,198,177.95 + $3.99 shipping	**New**	Seller: **bordeebook** Seller Rating: 93% positive over the past 12 months. (125,891 total ratings) In Stock. Ships from United States. Domestic shipping rates and return policy. New item in excellent condition. Not used. May be a publisher overstock or have slight shelf wear. Satisfaction guaranteed!	Add to Cart or Sign in to turn on 1-Click ordering.

Source: Michael Eisen, *UC Berkeley*

Algorithms

	profnath	bordeebook	profnath over previous bordeebook	bordeebook over profnath
8-Apr	$1,730,045.91	$2,198,177.95		1.27059
9-Apr	$2,194,443.04	$2,788,233.00	0.99830	1.27059
10-Apr	$2,783,493.00	$3,536,675.57	0.99830	1.27059
11-Apr	$3,530,663.65	$4,486,021.69	0.99830	1.27059
12-Apr	$4,478,395.76	$5,690,199.43	0.99830	1.27059
13-Apr	$5,680,526.66	$7,217,612.38	0.99830	1.27059

Source: Michael Eisen, *UC Berkeley*

Algorithms

The Making of a Fly: The Genetics of Animal Design (Paperback)
by Peter A. Lawrence

‹ Return to product information

Always pay through Amazon.com's Shopping Cart or 1-Click.
Learn more about Safe Online Shopping and our safe buying guarantee.

Price at a Glance

List Price: ~~$70.00~~
Used: from **$42.56**
New: from **$18,651,718.08**

Have one to sell? Sell yours here

All | New (2 from $18,651,718.08) | Used (11 from $42.56)

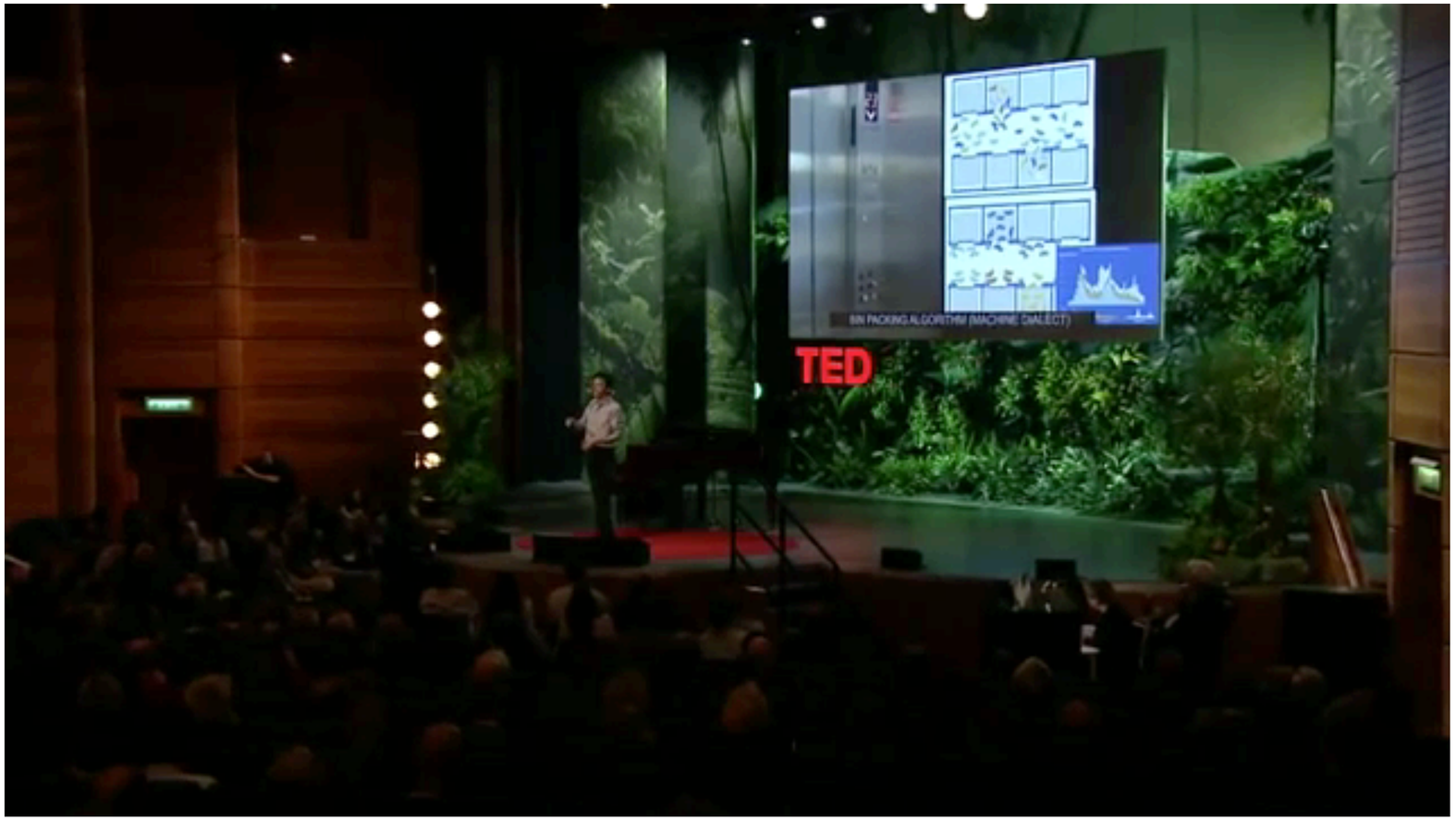
Show: New | Prime offers only (0) — Sorted by: Price + Shipping

New 1-2 of 2 offers

Price + Shipping	Condition	Seller Information	Buying Options
$18,651,718.08 + $3.99 shipping	**New**	Seller: **profnath** Seller Rating: ★★★★★ 93% positive over the past 12 months. (8,278 total ratings) In Stock. Ships from NJ, United States. Domestic shipping rates and return policy. Brand new, Perfect condition, Satisfaction Guaranteed.	Add to Cart or Sign in to turn on 1-Click ordering.
$23,698,655.93 + $3.99 shipping	**New**	Seller: **bordeebook** Seller Rating: ★★★★★ 93% positive over the past 12 months. (127,332 total ratings) In Stock. Ships from United States. Domestic shipping rates and return policy. New item in excellent condition. Not used. May be a publisher overstock or have slight shelf wear. Satisfaction guaranteed!	Add to Cart or Sign in to turn on 1-Click ordering.

Source: Michael Eisen, *UC Berkeley*

Algorithms



Source: Kevin Slavin, *TED*

Algorithms

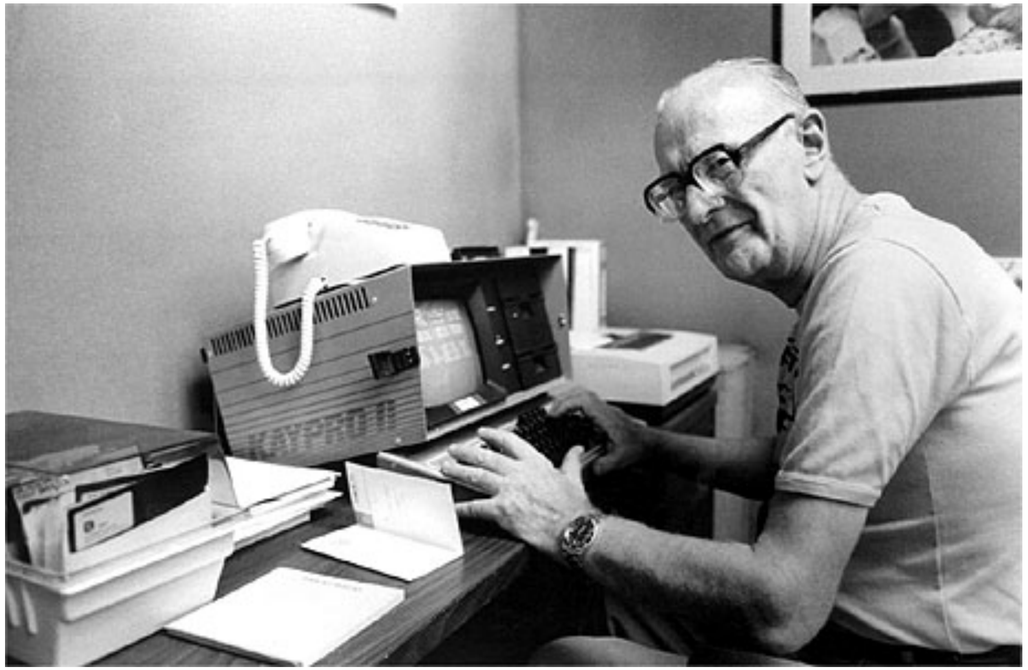
Trading program sparked May 'flash crash'

Government regulators say a trading program was behind the massive stock slide on May 6.

Automatic computerized traders on the stock market shut down as they detected the sharp rise in buying and selling. Altogether, this led to the abrupt drop in prices of individual stocks and other financial instruments like exchange-traded funds, and caused shares of some prominent companies like Procter & Gamble and Accenture to trade down as low as a penny or as high as $100,000.

Source: *CNN Money* —— *The New York Times*

“…indistinguishable from magic.”



Algorithms are powerful

Language analysis

Table 1 | The 177 irregular verbs studied

Frequency	Verbs	Regularization (%)	Half-life (yr)
10^{-1}–1	be, have	0	38,800
10^{-2}–10^{-1}	come, do, find, get, give, go, know, say, see, take, think	0	14,400
10^{-3}–10^{-2}	begin, break, bring, buy, choose, draw, drink, drive, eat, fall, fight, forget, grow, hang, help, hold, leave, let, lie, lose, reach, rise, run, seek, set, shake, sit, sleep, speak, stand, teach, throw, understand, walk, win, work, write	10	5,400
10^{-4}–10^{-3}	arise, bake, bear, beat, bind, bite, blow, bow, burn, burst, carve, chew, climb, cling, creep, dare, dig, drag, flee, float, flow, fly, fold, freeze, grind, leap, lend, lock, melt, reckon, ride, rush, shape, shine, shoot, shrink, sigh, sing, sink, slide, slip, smoke, spin, spring, starve, steal, step, stretch, strike, stroke, suck, swallow, swear, sweep, swim, swing, tear, wake, wash, weave, weep, weigh, wind, yell, yield	43	2,000
10^{-5}–10^{-4}	bark, bellow, bid, blend, braid, brew, cleave, cringe, crow, dive, drip, fare, fret, glide, gnaw, grip, heave, knead, low, milk, mourn, mow, prescribe, redden, reek, row, scrape, seethe, shear, shed, shove, slay, slit, smite, sow, span, spurn, sting, stink, strew, stride, swell, tread, uproot, wade, warp, wax, wield, wring, writhe	72	700
10^{-6}–10^{-5}	bide, chide, delve, flay, hew, rue, shrive, slink, snip, spew, sup, wreak	91	300

177 Old English irregular verbs were compiled for this study. These are arranged according to frequency bin, and in alphabetical order within each bin. Also shown is the percentage of verbs in each bin that have regularized. The half-life is shown in years. Verbs that have regularized are indicated in red. As we move down the list, an increasingly large fraction of the verbs are red; the frequency-dependent regularization of irregular verbs becomes immediately apparent.

UC San Diego

Source: Lieberman *et al.*, *Nature* 2007

Human-based computation

Human-based computation

Human-based computation

Culturomics

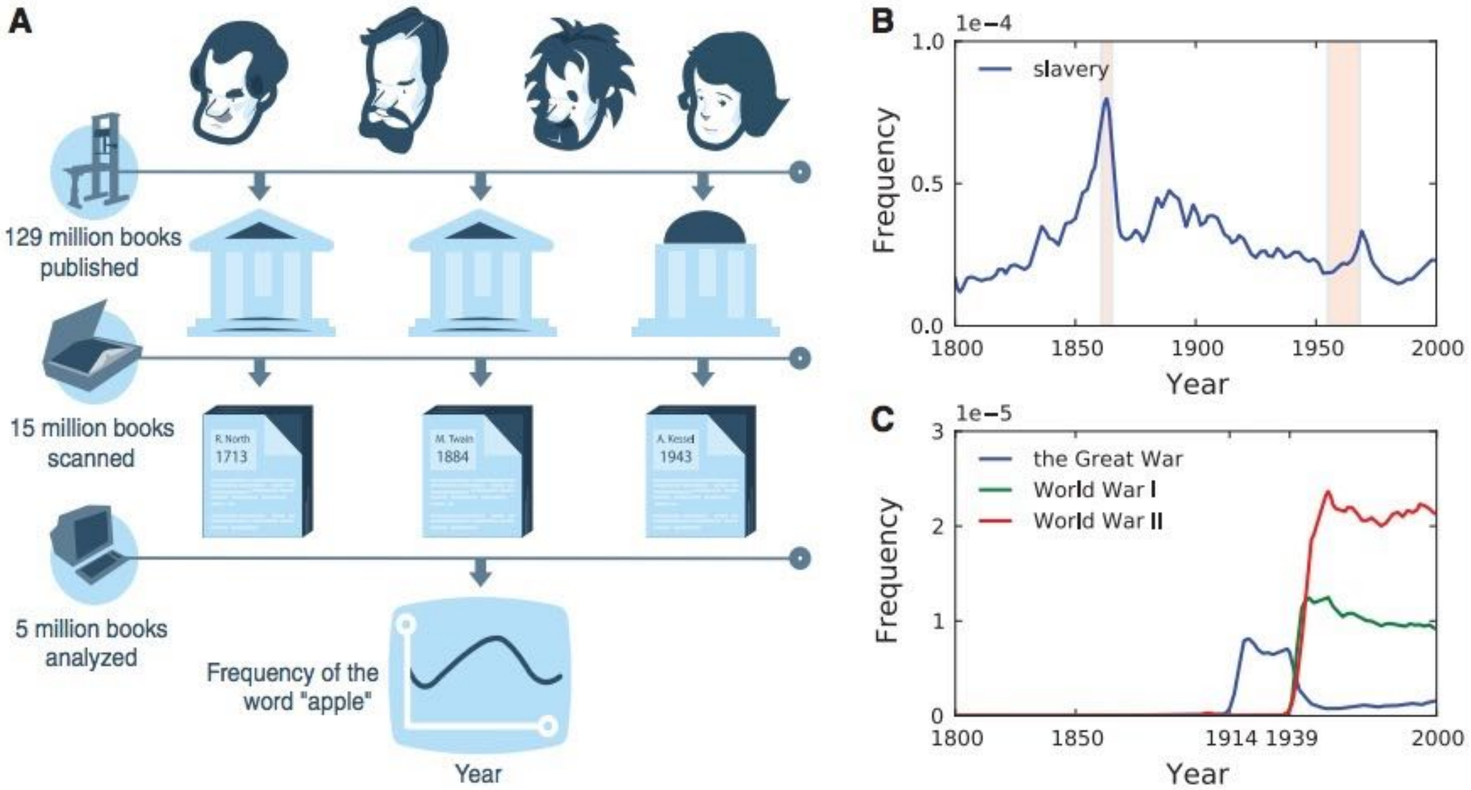
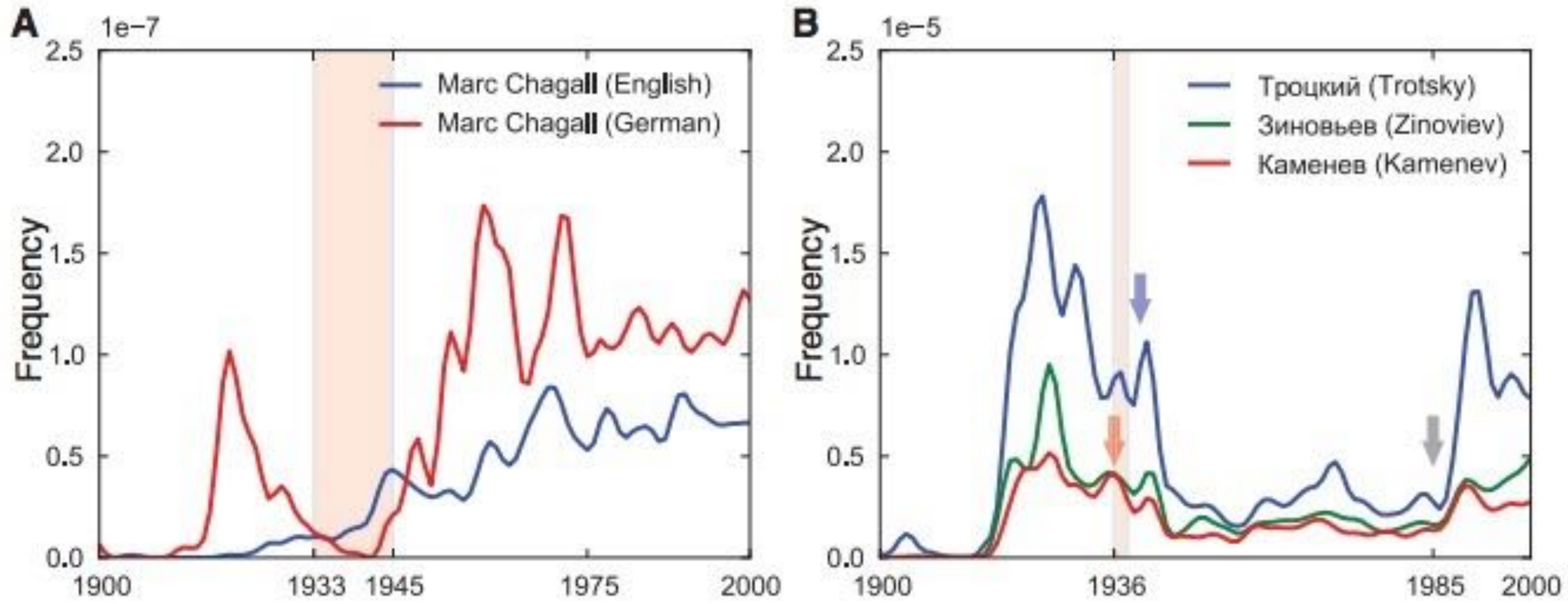


Fig. 1. Culturomic analyses study millions of books at once. **(A)** Top row: Authors have been writing for millennia; ~129 million book editions have been published since the advent of the printing press (upper left). Second row: Libraries and publishing houses provide books to Google for scanning (middle left). Over 15 million books have been digitized. Third row: Each book is associated with metadata. Five million books are chosen for computational analysis (bottom left). Bottom row: A culturomic time line shows the frequency of "apple" in English books over time (1800–2000). **(B)** Usage frequency of "slavery". The Civil War (1861–1865) and the civil rights movement (1955–1968) are highlighted in red. The number in the upper left (1e-4 = 10^{-4}) is the unit of frequency. **(C)** Usage frequency over time for "the Great War" (blue), "World War I" (green), and "World War II" (red).

UC San Diego

Source: Michel *et al.*, *Science* 2011

Culturomics



Source: Michel *et al.*, *Science* 2011

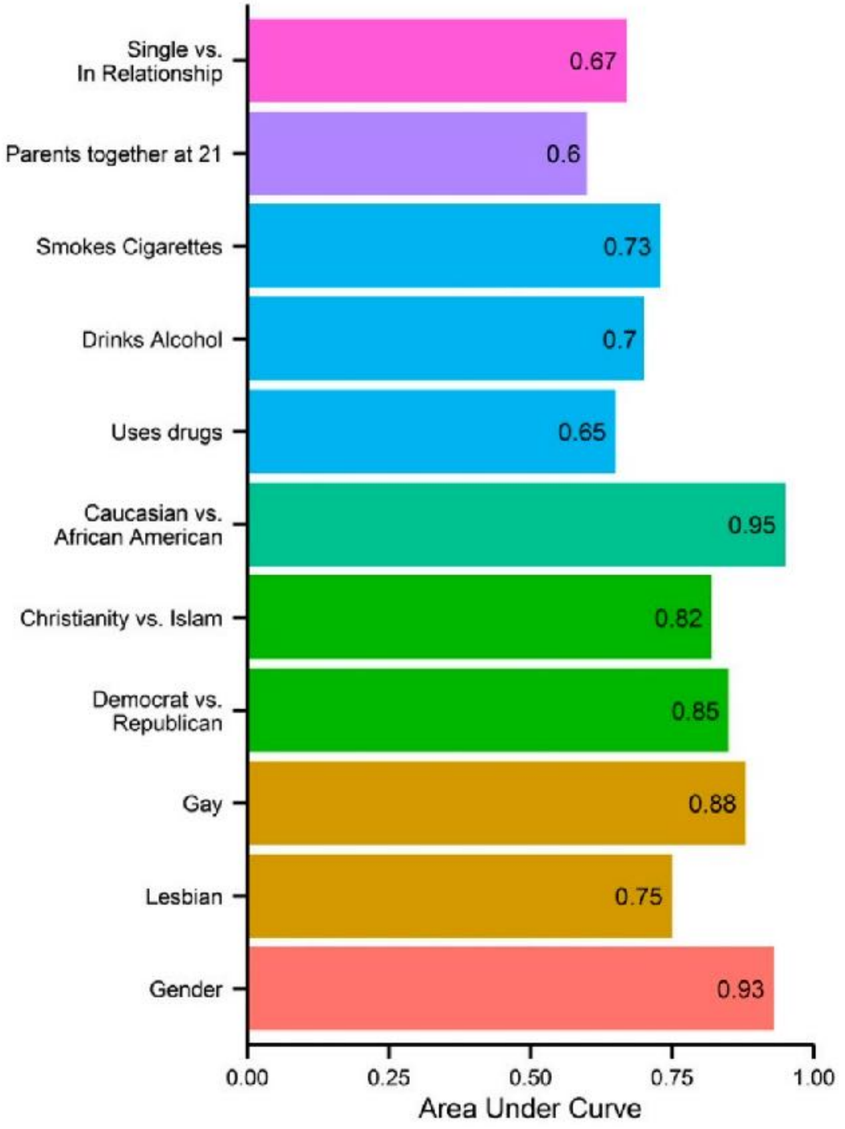


Fig. 2. Prediction accuracy of classification for dichotomous/dichotomized attributes expressed by the AUC.

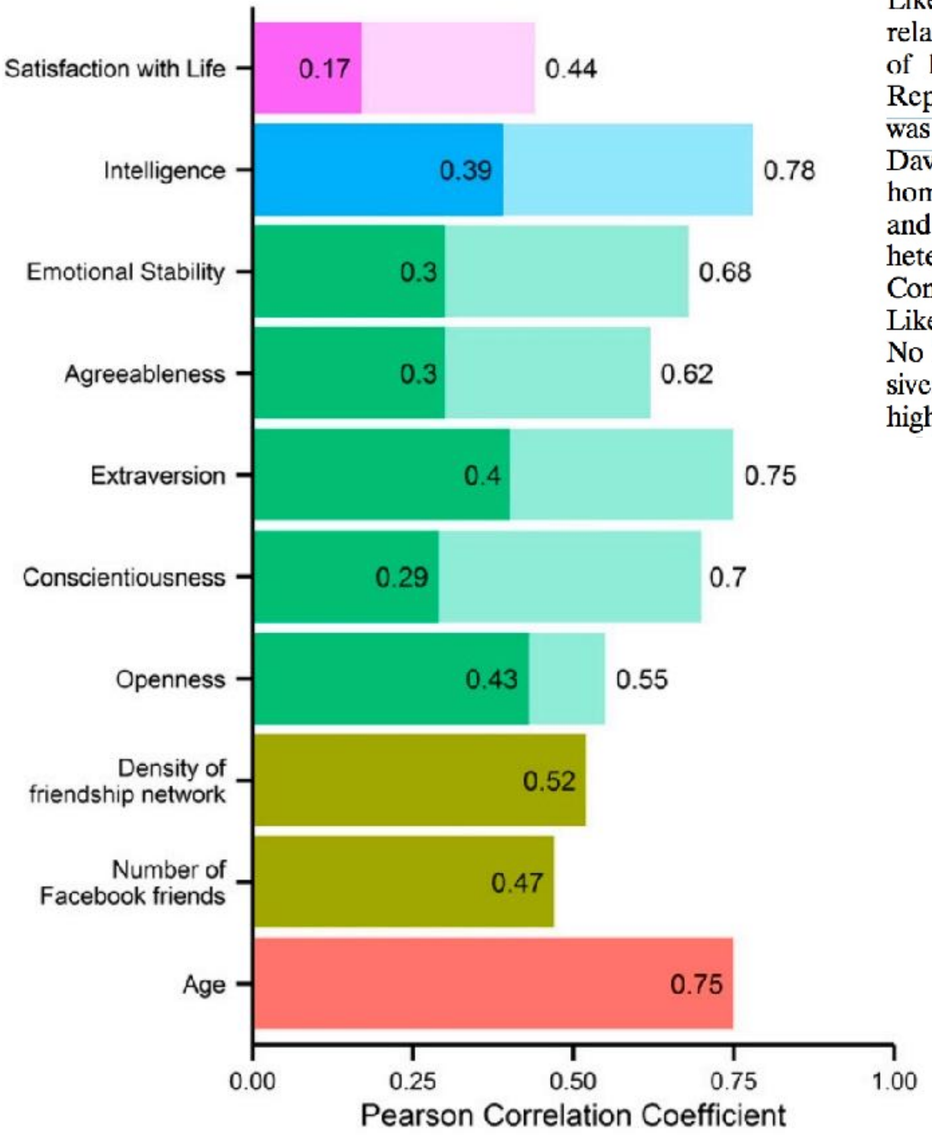
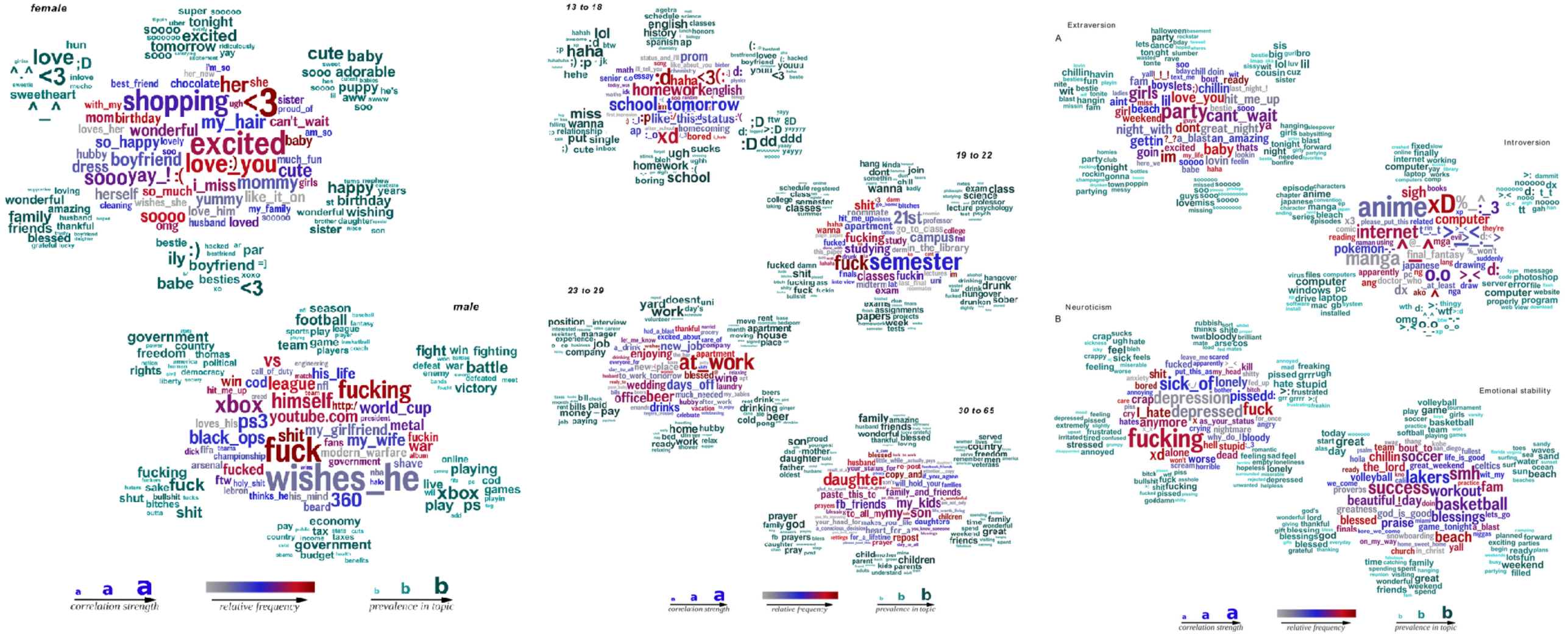


Fig. 3. Prediction accuracy of regression for numeric attributes and traits expressed by the Pearson correlation coefficient between predicted and actual attribute values; all correlations are significant at the $P < 0.001$ level. The transparent bars indicate the questionnaire's baseline accuracy, expressed in terms of test–retest reliability.

Predictive Power of Likes. Individual traits and attributes can be predicted to a high degree of accuracy based on records of users' Likes. Table S1 presents a sample of highly predictive Likes related to each of the attributes. For example, the best predictors of high intelligence include "Thunderstorms," "The Colbert Report," "Science," and "Curly Fries," whereas low intelligence was indicated by "Sephora," "I Love Being A Mom," "Harley Davidson," and "Lady Antebellum." Good predictors of male homosexuality included "No H8 Campaign," "Mac Cosmetics," and "Wicked The Musical," whereas strong predictors of male heterosexuality included "Wu-Tang Clan," "Shaq," and "Being Confused After Waking Up From Naps." Although some of the Likes clearly relate to their predicted attribute, as in the case of No H8 Campaign and homosexuality, other pairs are more elusive; there is no obvious connection between Curly Fries and high intelligence.

Source: Kosinski *et al.*, *PNAS* 2013

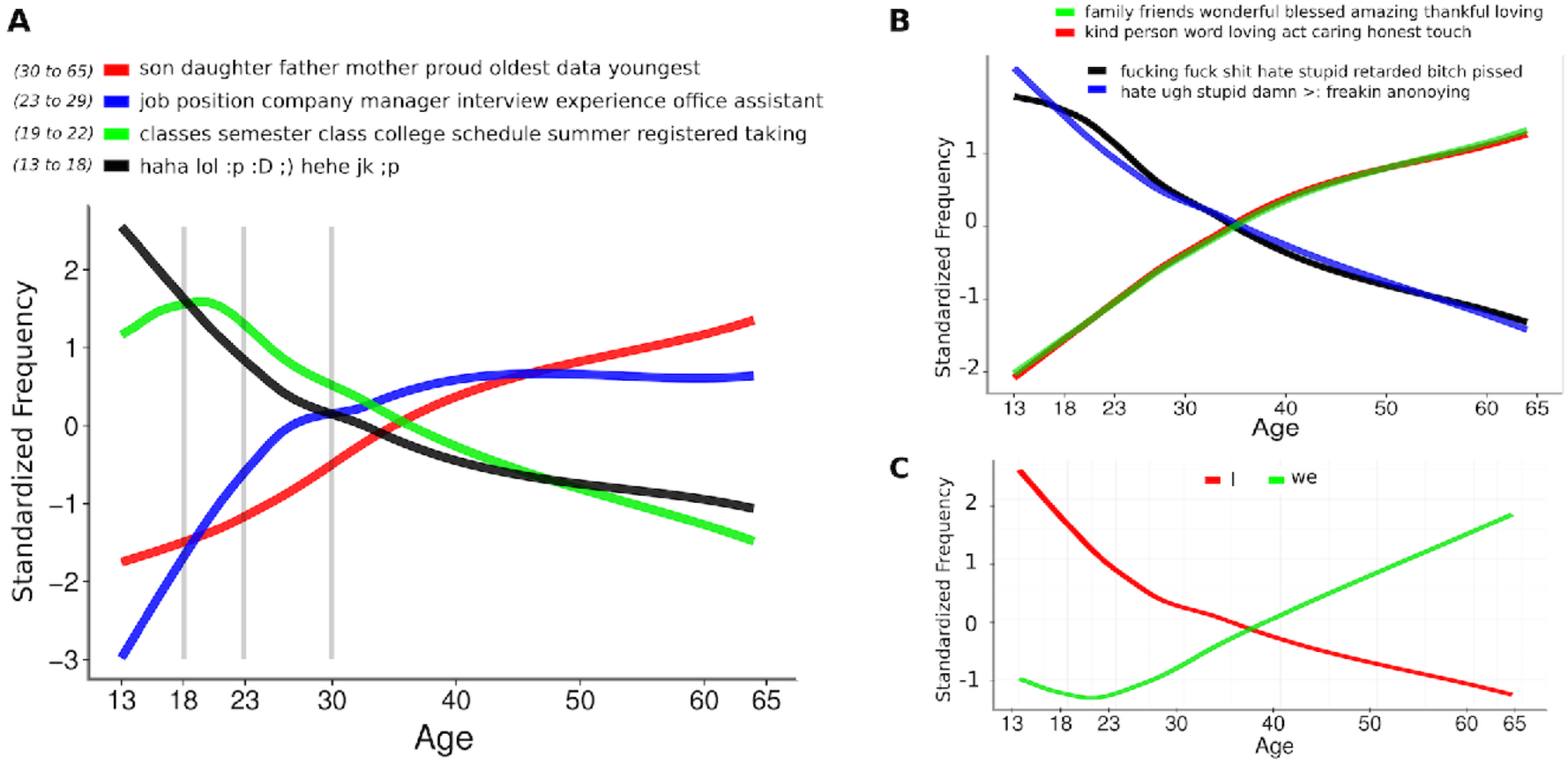
Doing the amusing



Source: Schwartz *et al.*, PLoSONE 2013

UC San Diego

Doing the amusing



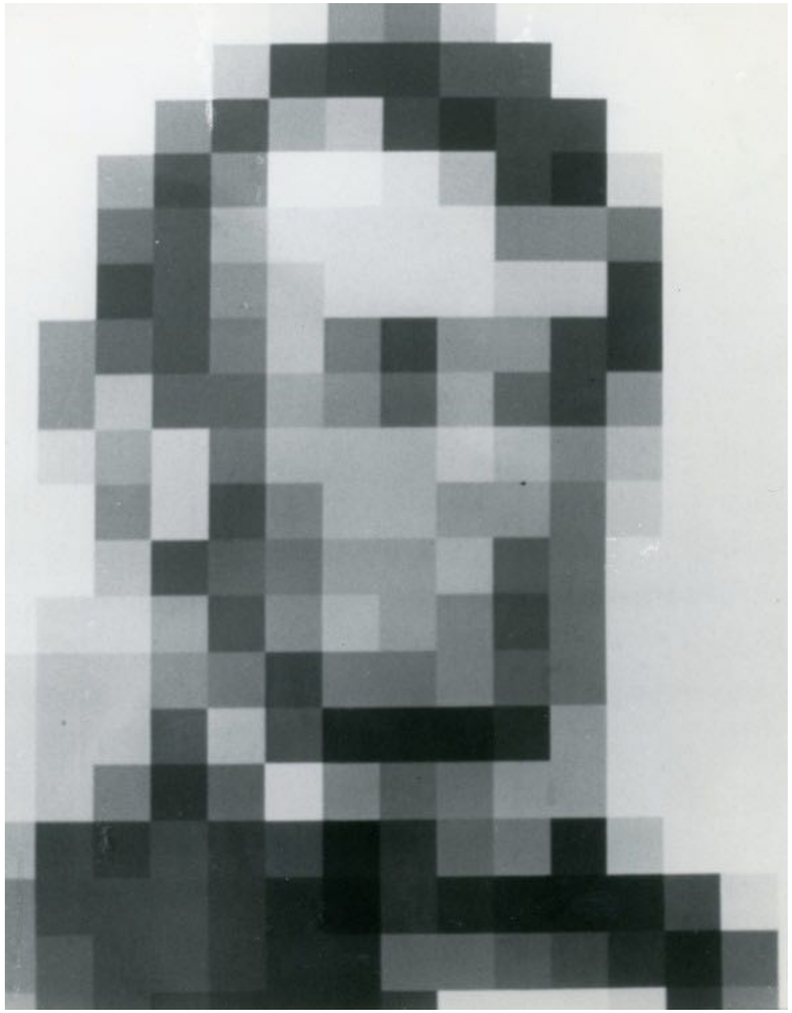
Source: Schwartz *et al.*, *PLoSONE* 2013

UC San Diego

Doing the amazing

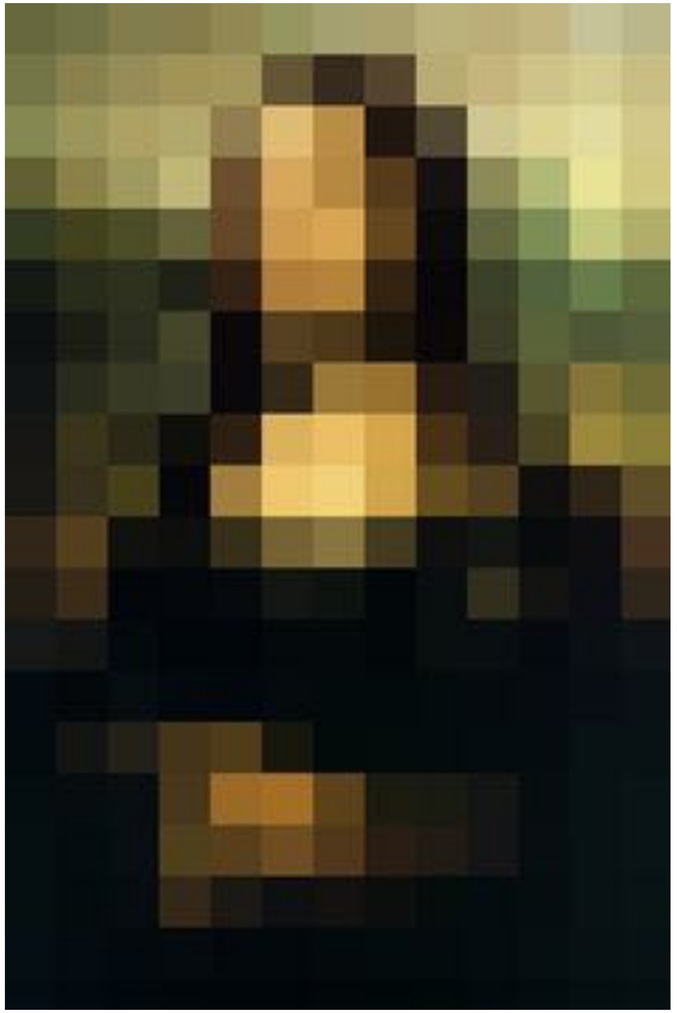
Source: Andrew Ochoa, Waverly Labs

ENHANCE!



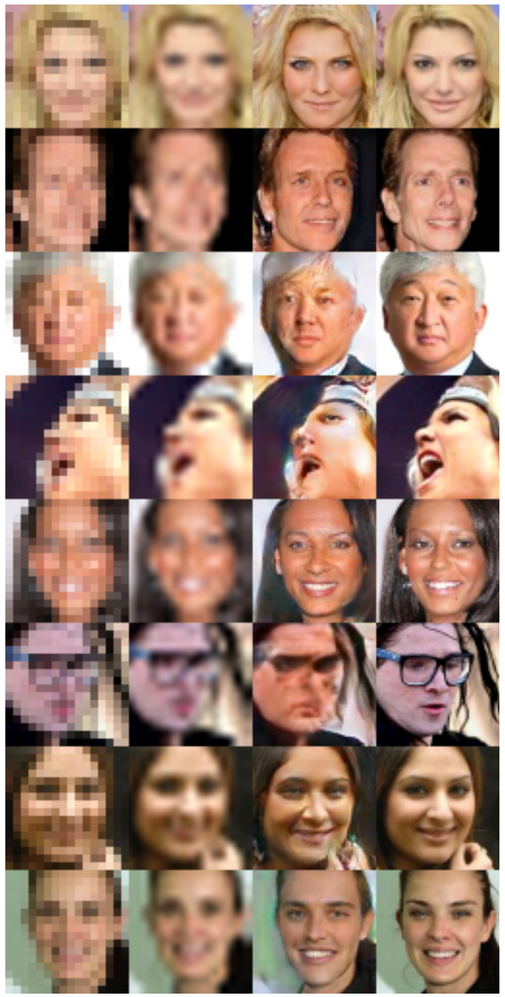
c.f. Harmon & Julesz, *Science* 1973

ENHANCE!



c.f. Harmon & Julesz, *Science* 1973

ENHANCE!



"…the first column is the 16x16 input image, the second one is what you would get from a standard bicubic interpolation, the third is the output generated by the neural net, and on the right is the ground truth."

Source: David Garcia, Qualcomm

Reality manipulation

Source: *Adobe*

Reality manipulation

Source: Theis *et al.*, *CVPR* 2016

Personal privacy

Source: Deb Roy, *TED* 2011

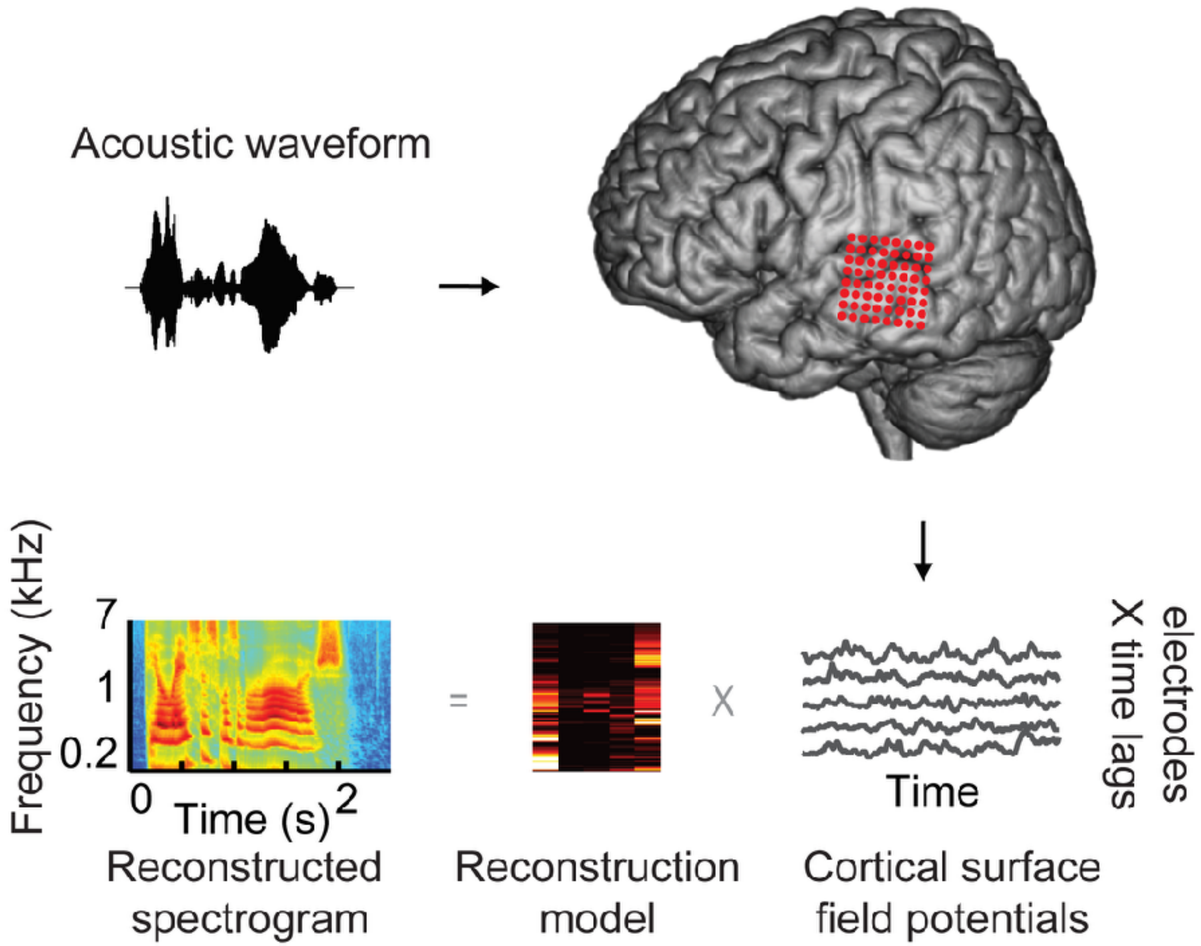
Espionage

Source: Davis A *et al.*, *SIGGRAPH* 2014

Espionage

Source: Davis A *et al.*, *SIGGRAPH* 2014

Neural decoding



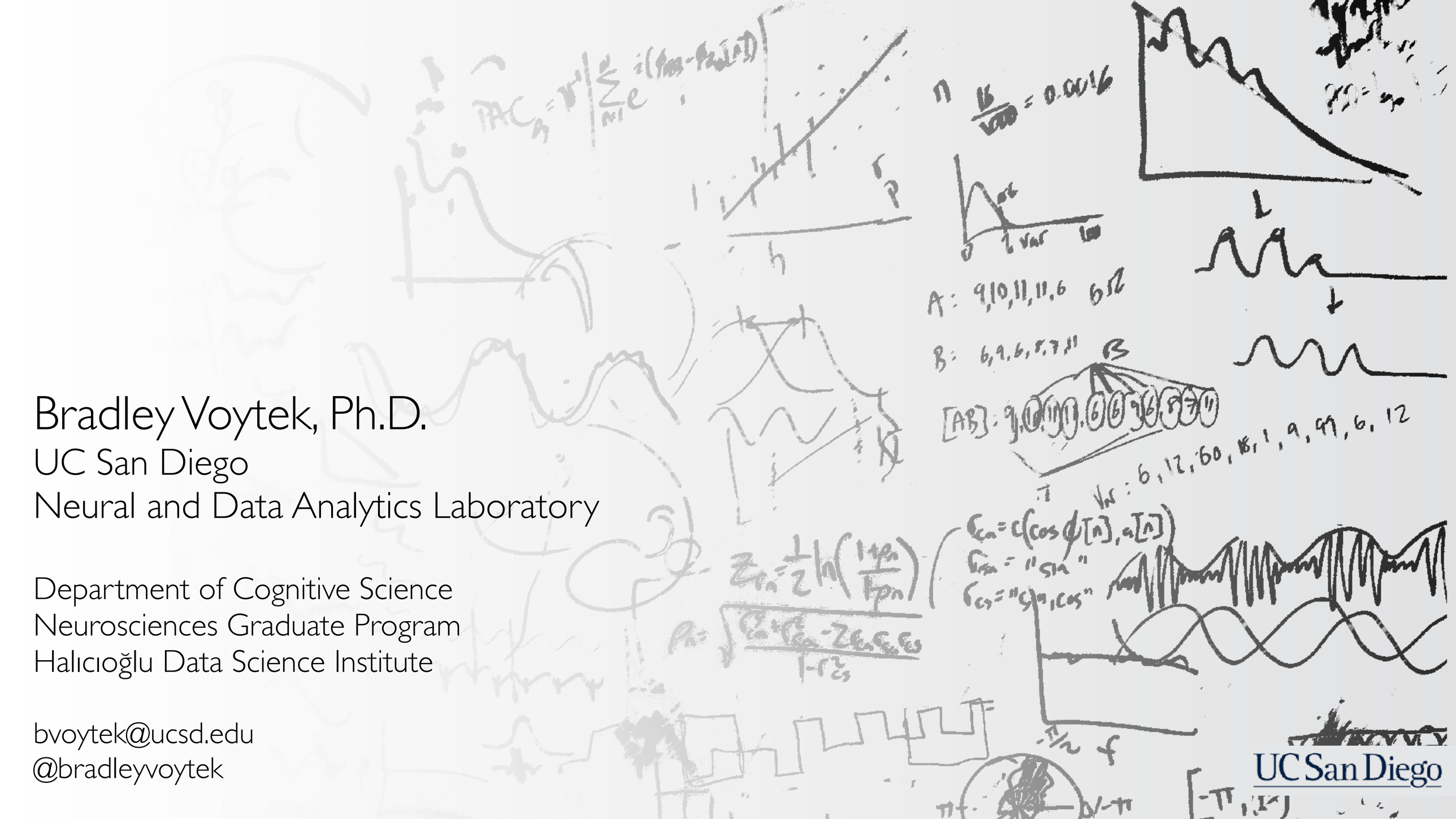
Source: Pasley *et al.*, *PLoS Biol* 2012

Neural decoding

Presented clip

Clip reconstructed from brain activity

Source: Nishimoto et al., *Current Biol* 2011



Bradley Voytek, Ph.D.

UC San Diego
Neural and Data Analytics Laboratory

Department of Cognitive Science
Neurosciences Graduate Program
Halıcıoğlu Data Science Institute

bvoytek@ucsd.edu
@bradleyvoytek

UC San Diego