Big data in business

Raghavendra Rau, University of Cambridge

Big data examples

Target

Amazon

Facebook

Netflix

Farecast

Google Flu

Cambridge Analytica

Data vs. big data

What is the difference between data and big data?

N= n vs. N = all

Handling data

- Clean data
- Sampling problems:
  - Precision increases with randomness, not sample size
    - Dewey defeats Truman, 1948; Trump election, 2016
  - Looking at finer samples decreases randomness
- Summarizing distributions: Mean, modes, medians, standard deviations
- Causation

Big data

- Analyzing the population vs. analyzing a sample
  - Domesday book
  - Deciannual censuses vs. real time data (Hollerith punched cards, 1890)
  - Cambridge Analytica, 2016
  - The population need not be large (Sumo wrestlers)

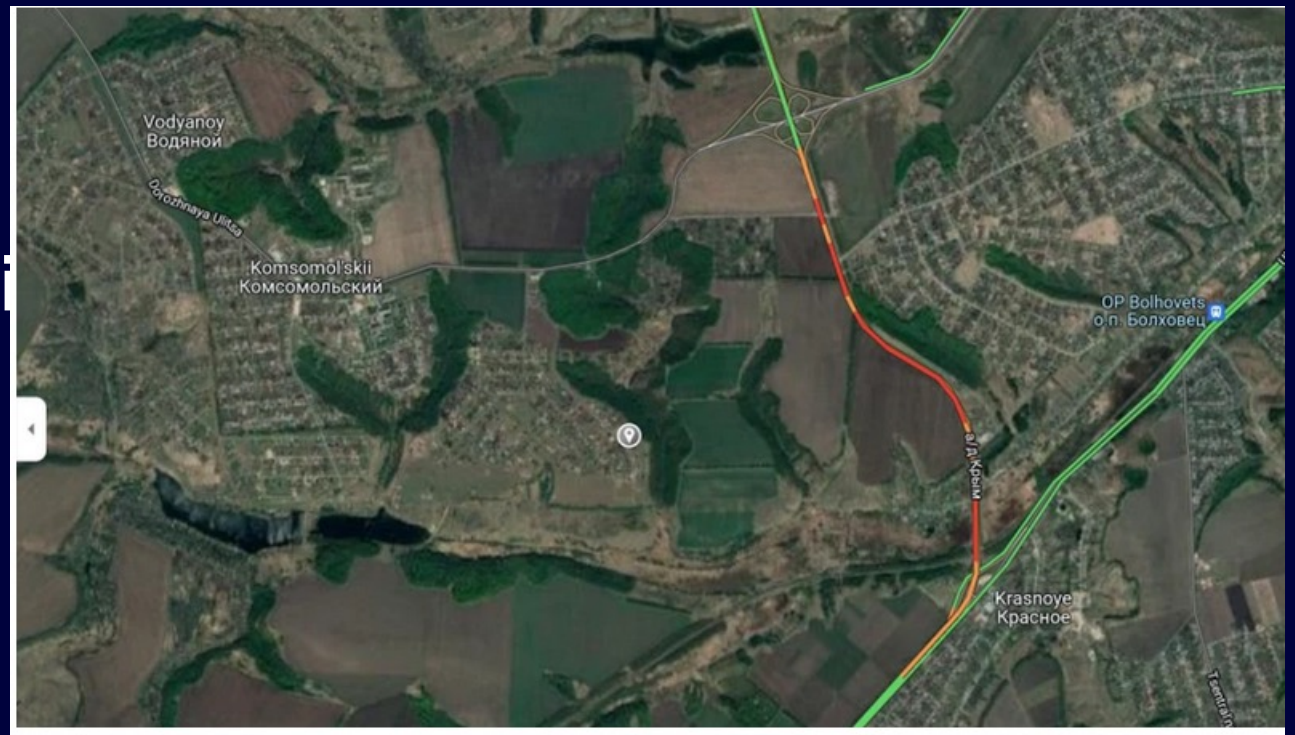
- More messy data (varies in quality, is collected at different times around the world, is kept in a wide variety of places)

- General direction vs. precise inferences

Where is the big data coming from?

- Reuse of data from other sources
  - Machine translation: Google translate
  - Billion Prices Project (Cavallo and Rigobon)

- New data:
  - Satellite photos
  - OS Int (Google maps i
  - Smart phones



CREDIT: GOOGLE MAPS

Smartphones

Smartphones can track what searches we carried out, what books we have bought, what vacations we shopped for, what dreams we have, and which diseases we fear to have.

Oura rings, Apple watches, Google Pixel watches, Fitbits: Add biometric information to the mix.

- These allow phones to capture additional information that we may not know ourselves.
- The smartphone knows when you are stressed, when you are low on sugar, and when you like a person of the same or opposite sex.

New forms of data

- Geospatial: Study of mobility and location-based proximity
- Sociometric: Study of social relationships, especially preferences, within groups
- Psychometric: Study of a person's mental state, personality and thought processes
- Biometric: Study of a person's biological characteristics

Why are we collecting more and more data?

- Business assume that they will get valuable insights. Regulators assume that it reduces asymmetric information.
  - Companies listed on the stock market are required to file quarterly financial reports with the stock market regulator, which are then made public.
  - Banks and investment funds must comply with stringent reporting obligations.
  - Companies operating in certain sectors (pharmaceuticals, health care, education, air travel) are required to provide additional information.
  - Problem: Most human beings do not have either the time or the ability to interpret the massive amount of data produced by smartphones.

We need to process the data to predict

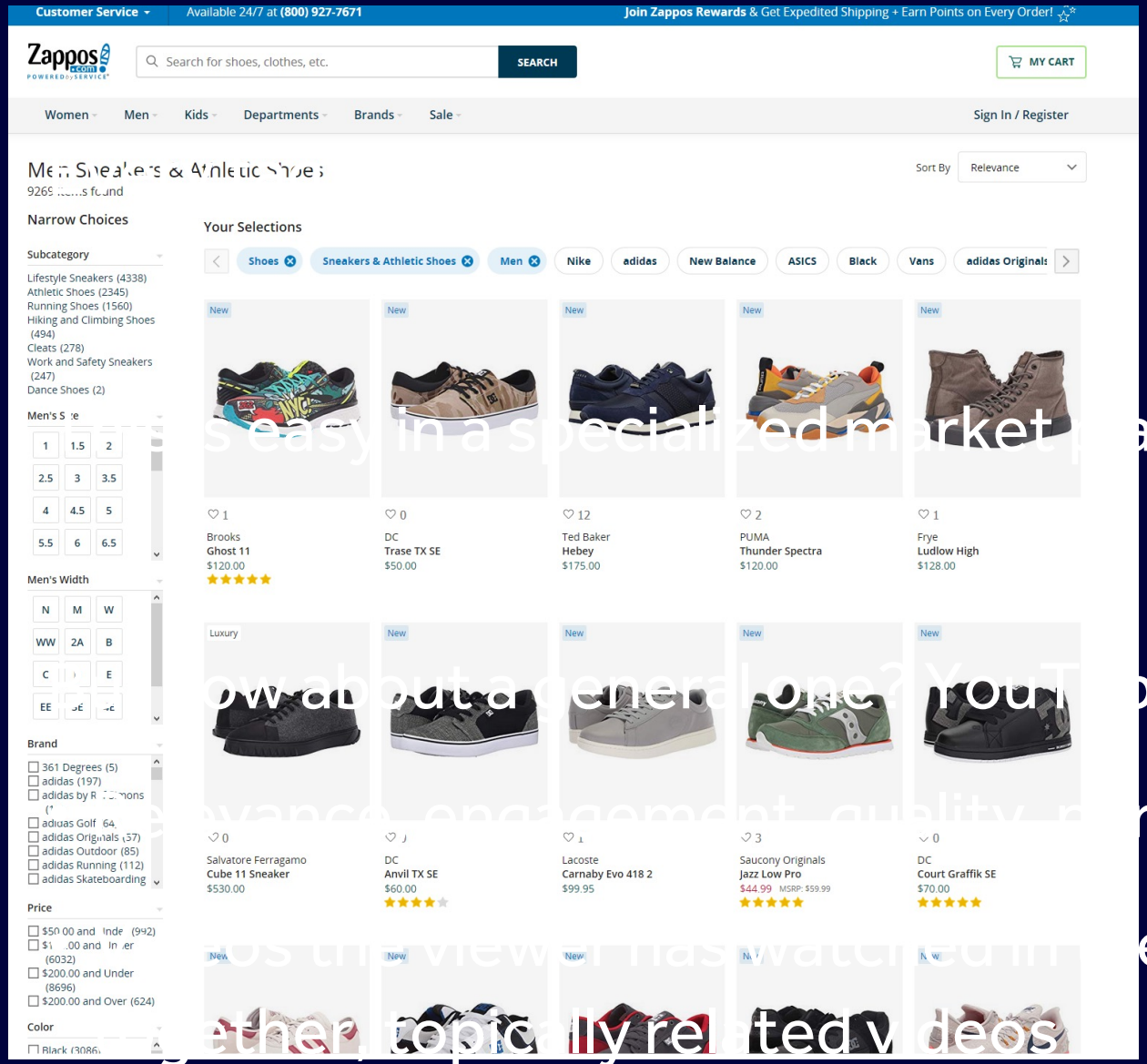
- We need a standard language to compare preferences
- We need to better match preferences along multiple dimensions
- We need an effective way to comprehensively capture our preferences.

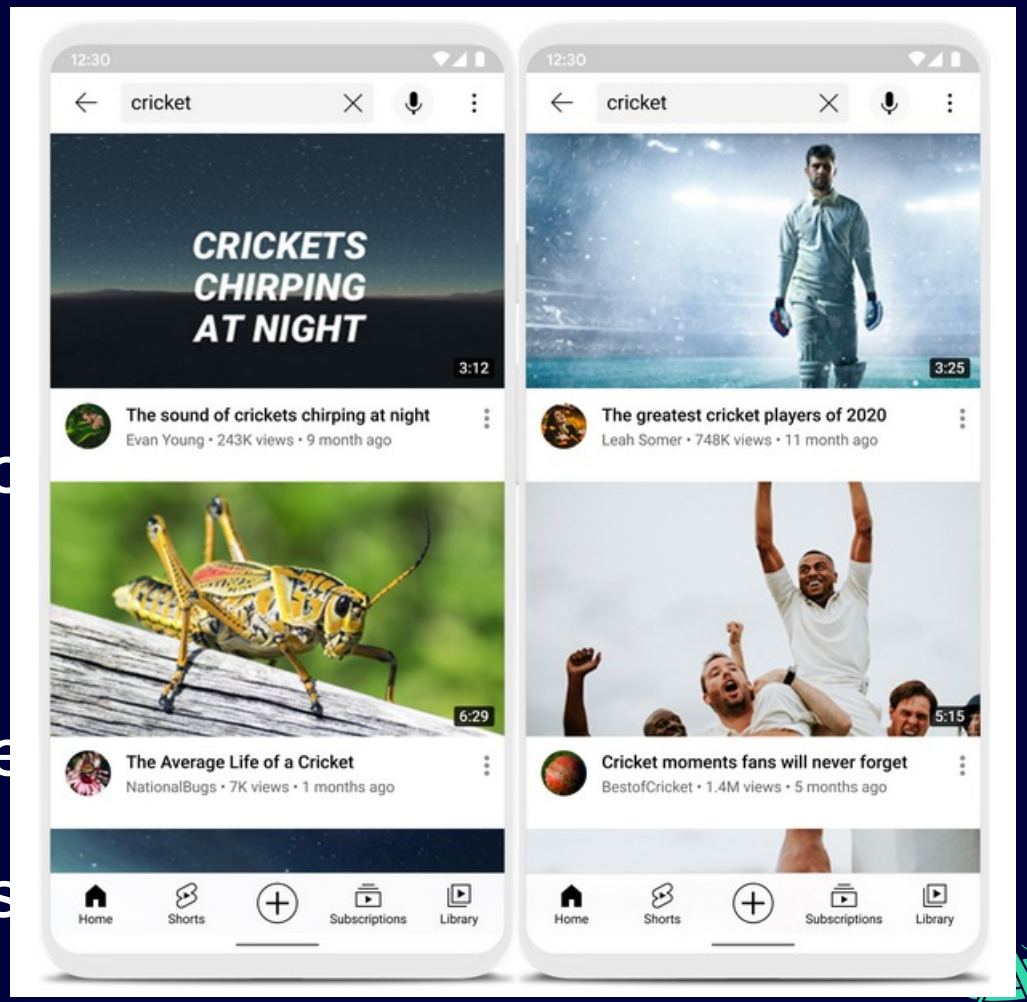
Why do we need a language?

- Multiple dimensions of data

- But all these dimensions need to be quantified

What does processing data mean?





s easy in a specialized market ac

owabout a friend one? YouT oe

dvance, engagement, quality rs

e past, videos typically watched

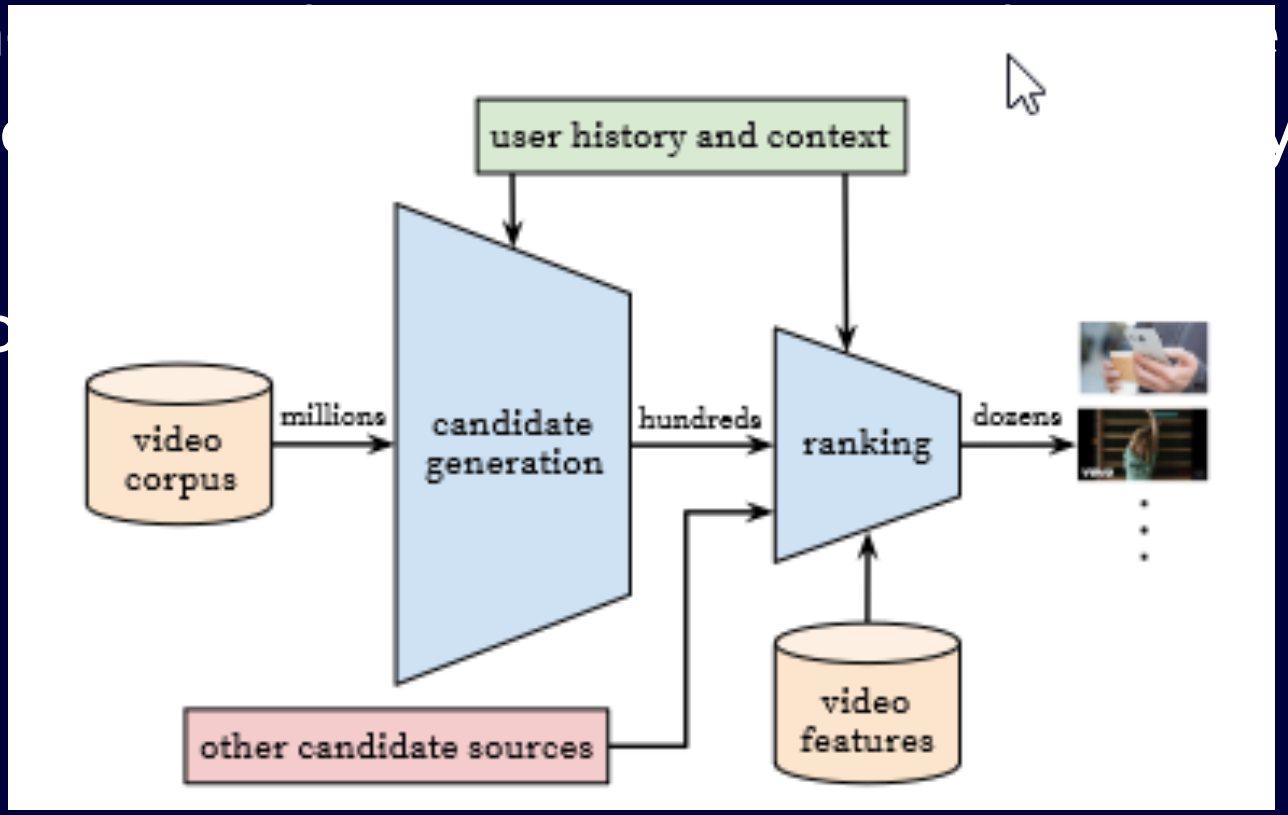
gether, topically related videos

How do you find anything?

YouTube:

The Youtube search engin
(i.e., laparoscopic-append
natural language like “Real
appendectomy”); in the Yo

How much of the video does the user watch?



Source: Covington, Adams, Sargin, Deep neural networks for YouTube recommendations

Element 1: Developing a data ontology

How do you find anything?

Ebay vs. Amazon

Element 2: Multi-dimensional Preference Matching

- Algorithms to evaluate sets of multiple preferences and their relative weights and to identify best matches.
- Same technology used to manage our photo collections to find pictures with certain faces/features, or to have Siri or Alexa “understand” voice commands, or to make our smart watches detect signs of a dangerous heart condition.
- Basically, preference data is just a data stream forming a particular pattern, so we can adapt pattern-matching algorithms.

Example: Netflix

Alternatives: Recommendations by employees or editors

Hacking

The original strategy: Cinematch system

Based on collaborative filtering

Slope One strategy

	Avengers	Spiderman
Pauline	2	2.5
Julien	3	?

The Cinematch system

Weighted average approach

	Avengers	Spiderman	Wonder Woman
Cesar	4	2	1
Blanche	2	3	?
Emma	?	1	4

- Combine Cesar and Blanche for Avengers vs. Spiderman: Emma = (2+(-1))/2=0.5; Emma for Avengers = 1 + 0.5 = 1.5
- Use Cesar for Avengers vs. Wonder Woman: Emma = 4 + (4-1)=7
- Weight 2/3-1/3 (2 sources when combining Avengers and Spiderman; 1 source for Avengers and WW)

Use correlation coefficients to see which customers' ratings are closely correlated

Cluster Analysis method: Use customers in the same cluster to make predictions

The Cinematch system

Ordinal logit model

Effective when the data set represents a rating scale satisfying the proportional odds assumption.

- If we have a rating scale from 1 to 5, we know 5 is greater than 4, 3 is greater than 2, and so on.

In a standard numeric series, the difference between each rating is 1. But for a ratings scale like the one used by Netflix, this is not always the case.

Individuals deciding whether a movie should receive a 4 or a 5 might set the difference at closer to 2 points, as 5 is a premium rating. The difference between 3 and 4 might be 1.5, and the consumer might see little difference among scores below 3.

Problems faced

Cold start problem: When a recommendation system lacks data to make predictions: Products newly added to the database and, more critically, for new users for whom no data is available.

Popularity bias: Popular movies are recommended frequently while less popular, niche films are recommended rarely or not at all.

Sparse data problem: The Netflix data set is both extremely large and relatively sparse, with only about 1% of the potential user-movie data points available.

Noisy data problem: Ratings that people report do not reflect actual behavior

The Cinematch system

The alternative least squares model

	Avengers	Frozen	Thor	Batman
Thomas		4.5	2.0	
Leandre	4.0		3.5	
Coppelia		5.0		2.0
Arthur		3.5	4.0	1.0

=

User matrix

Thomas	1.2	0.8
Leandre	1.4	0.9
Coppelia	1.5	1.0
Arthur	1.2	0.8

×

Movie matrix

Avengers	Frozen	Thor	Batman
1.5	1.2	1.0	0.8
1.7	0.6	1.1	0.4

What else did Netflix want to use big data for?

New shows to commission: House of Cards

How much to pay for new films

Challenge: Obtaining ratings (active rating vs. passive watching)

The Netflix prize

Is this new?

- 1418: The Duomo in Firenze by Brunelleschi
- 1741: The Longitude Act: John Harrison
- 1995: X-Prize:
  - $10 million for private, suborbital space flight; $10 million for creating safe, affordable, production-capable vehicles that exceeded 100 MPG energy equivalent; $2 million for advanced rocket development; $1.4 million for effective ways to counter oil spills

InnoCentive.com

The contest strategy

Own platform vs. outside host?

How much data to release

- Confidential data available to competitors
- Customer privacy

Intellectual property:

- Anonymous contest vs. open contest
- Hold-up problems from winners
- IP from losing solutions
- Algorithms leaked to competitors
- Lack of participation if you don’t share IP

Administrative and operational costs

The contest design

Problem specification: % increase in RMSE

How big an increase?

How long the contest should be kept open: Intermediate awards?

Multiple winners?

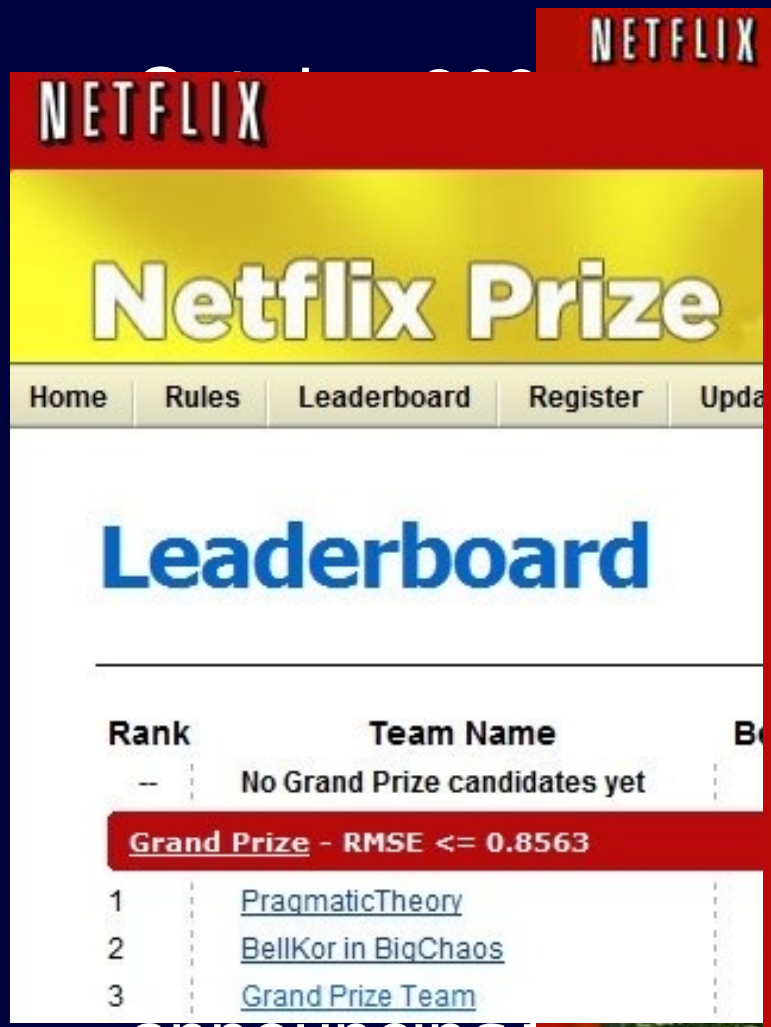
Size of award?

Type of contestant to attract?

The Netflix prize

announcing the Net

Try yourself: https://grouplens.org/datasets/movielens/



Element 3: Understanding us

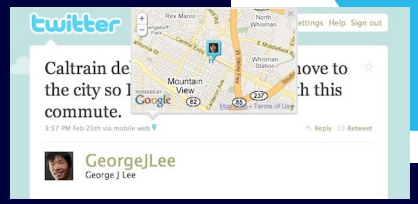
- Huge volumes of data
- Frequent feedback for the system
- Self-adjusting capacity
- Freshness

Can big data understand us?

- **Twitter/Foursquare data** (Hristova, D., Williams, M., Musolesi, M., Panzarasa, P., and Mascolo, C., 2016, Measuring Urban Social Diversity Using Interconnected Geo-Social Networks)
- **Facebook data** (Markovikj, D., Gievska, S., Kosinski, M., and Stillwell, D.J., 2021, Mining Facebook Data for Predictive Personality Modeling)
- **Loan application data**: Netzer, Oded, Lemaire, Alain and Herzenstein, Michal, 2019, When Words Sweat: Identifying Signals for Loan Default in the Text of Loan Applications. Columbia Business School Research Paper No. 16-83. Available at SSRN: https://ssrn.com/abstract=2865327

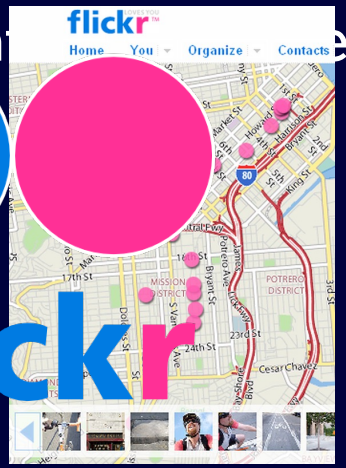
Geospatial and social data

- **Twitter/Foursquare data** (Hristova, D., Williams, M., Musolesi, M., Panzarasa, P., and Mascolo, C. Measuring Urban Social Diversity Using Interlinked Geo-Social Networks)

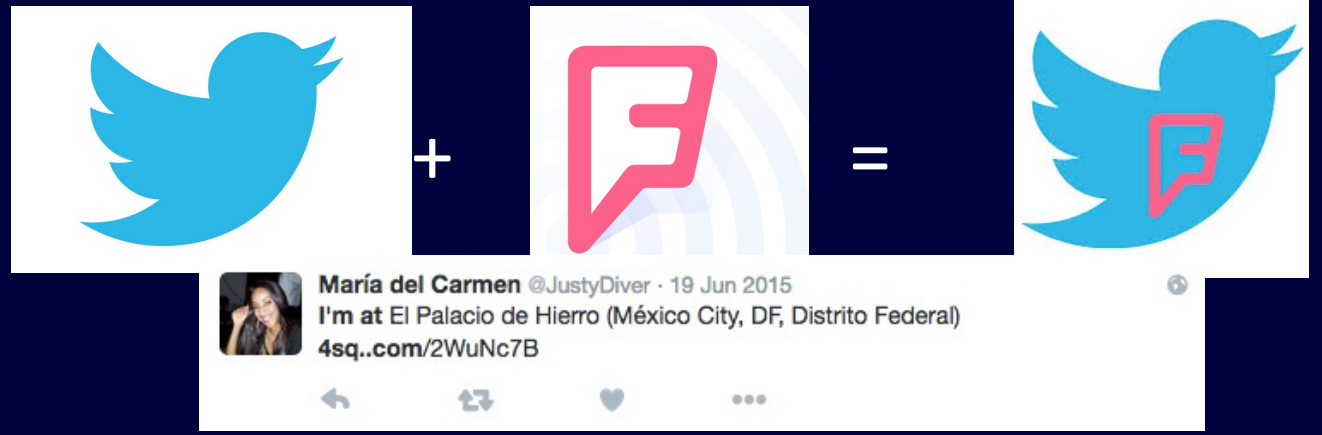


Geo-tagged tweets

Venue check-ins

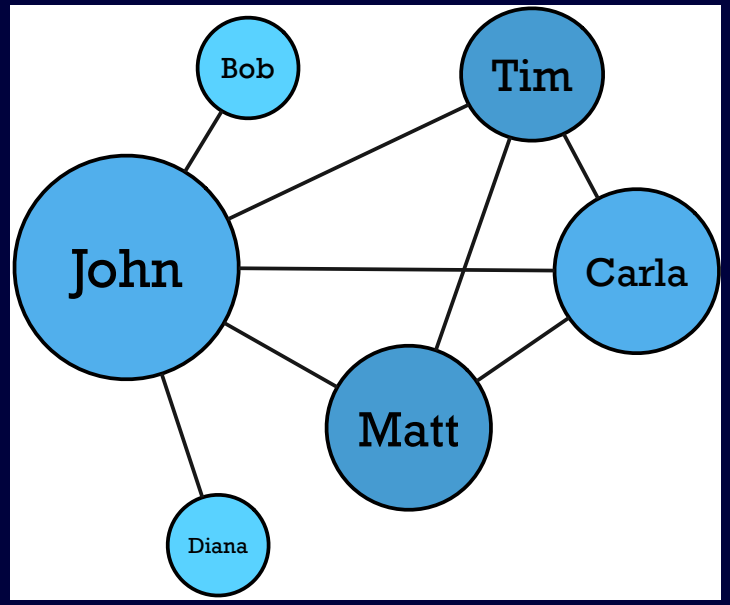


Geo-tagged photos



Geospatial and social data

- 38K users of Foursquare and Twitter
- 433K reciprocal connections on Twitter
- 550K check-ins to 42K venues in London
- 3M user transitions between venues on Foursquare

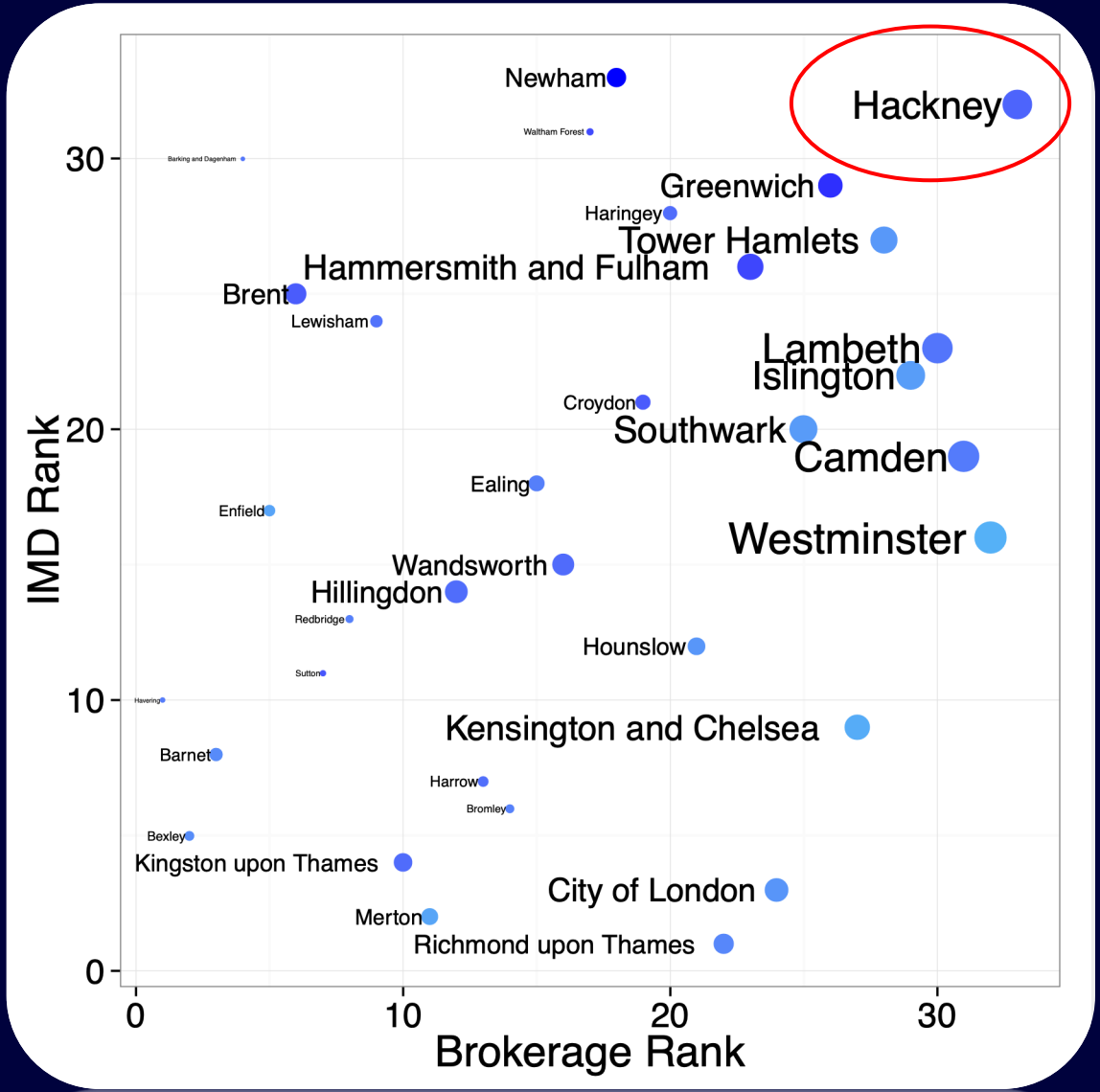


Social Network

check-in

Place Network

Applying to real estate predictions



Hackney, 2011
Population: 246K
Age: 25-34
Avg. House Price: £326K
2nd most deprived

Hackney, 2016
Population: 263K
Age: 25-34
Avg. House Price: £546K
11th most deprived

Diversity Rank

Psychometric data

- Facebook data (Markovikj, D., Gievska, S., Kosinski, M., and Stillwell, D.J. (2021) Mining Facebook Data for Predictive Personality Modeling)

A

or

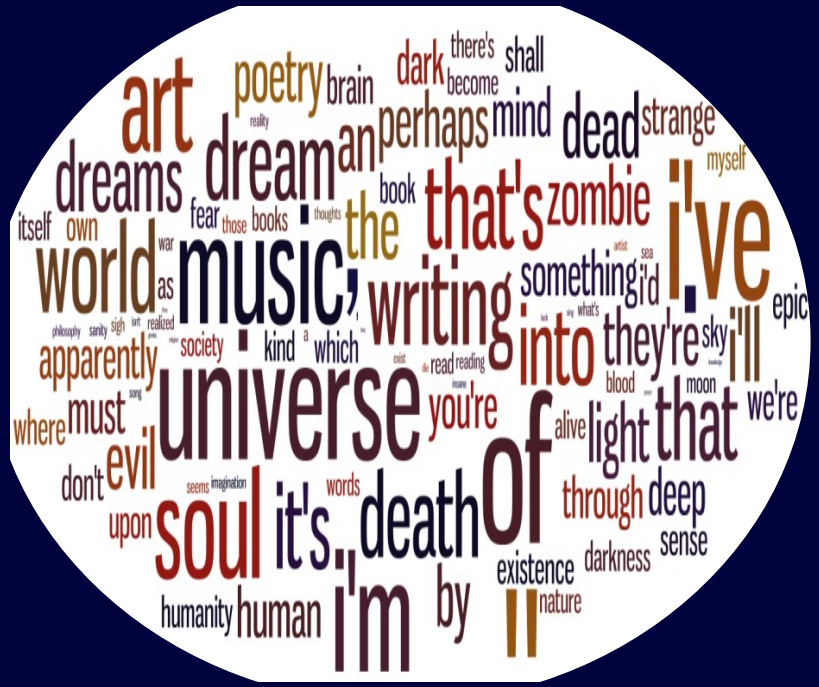
B

Psychometric data

- Which word cloud is more you?

e.g. Weekend, Home, Happy

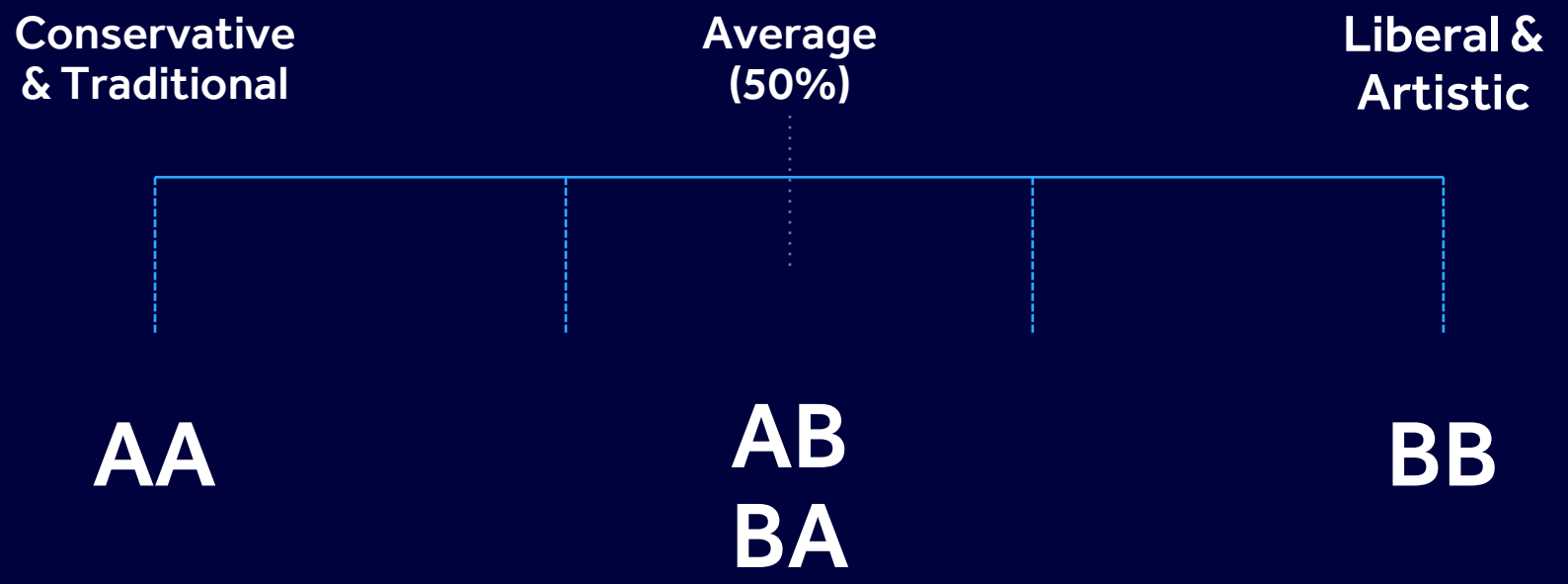
A



e.g. Universe, Music, Dream

B

Results: Openness scale



Mypersonality App launched in 2007

6mil individual psych & social profiles

25 validated psychometric tests

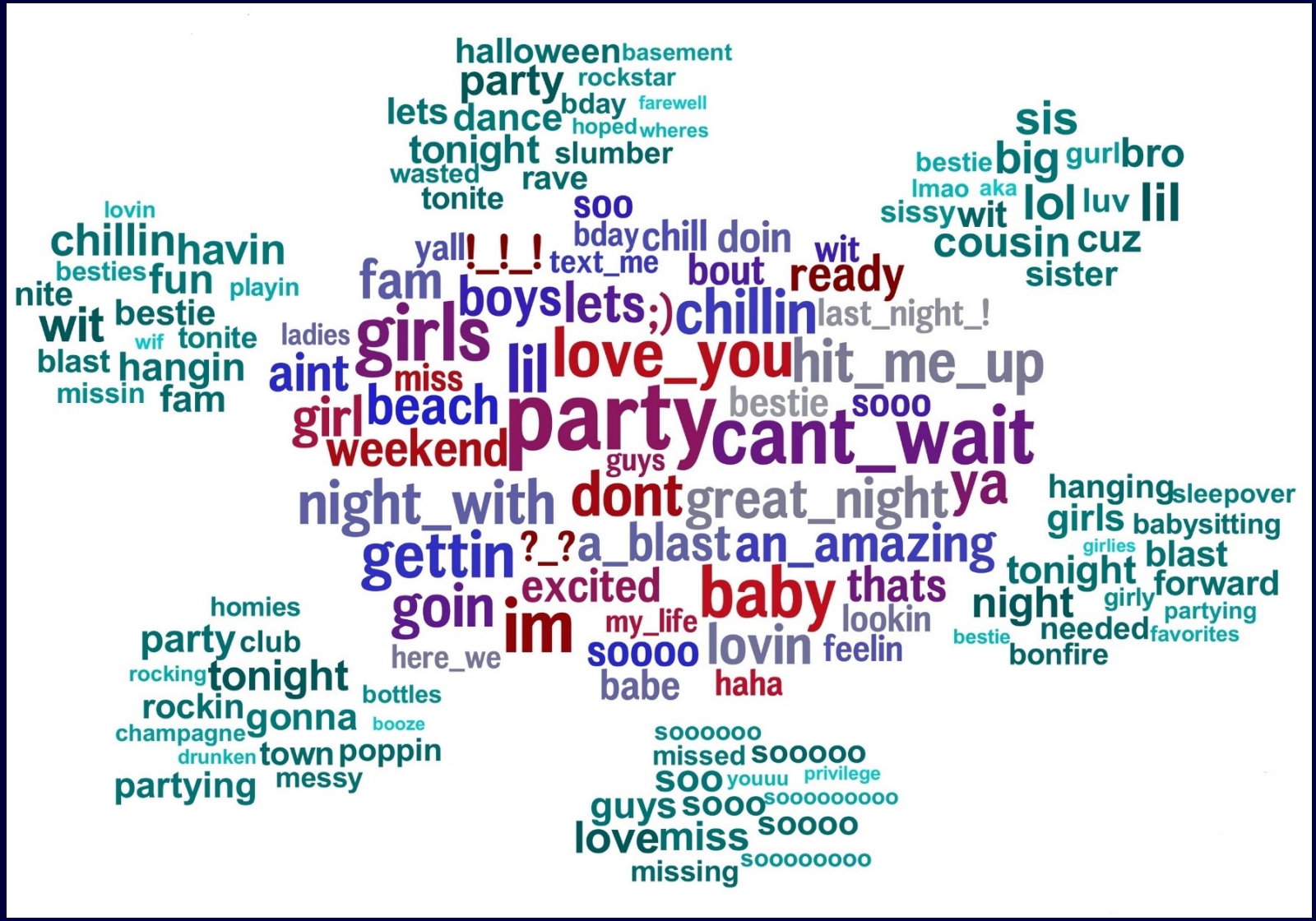
All data collected through opt-in

Data shared with >100 Universities

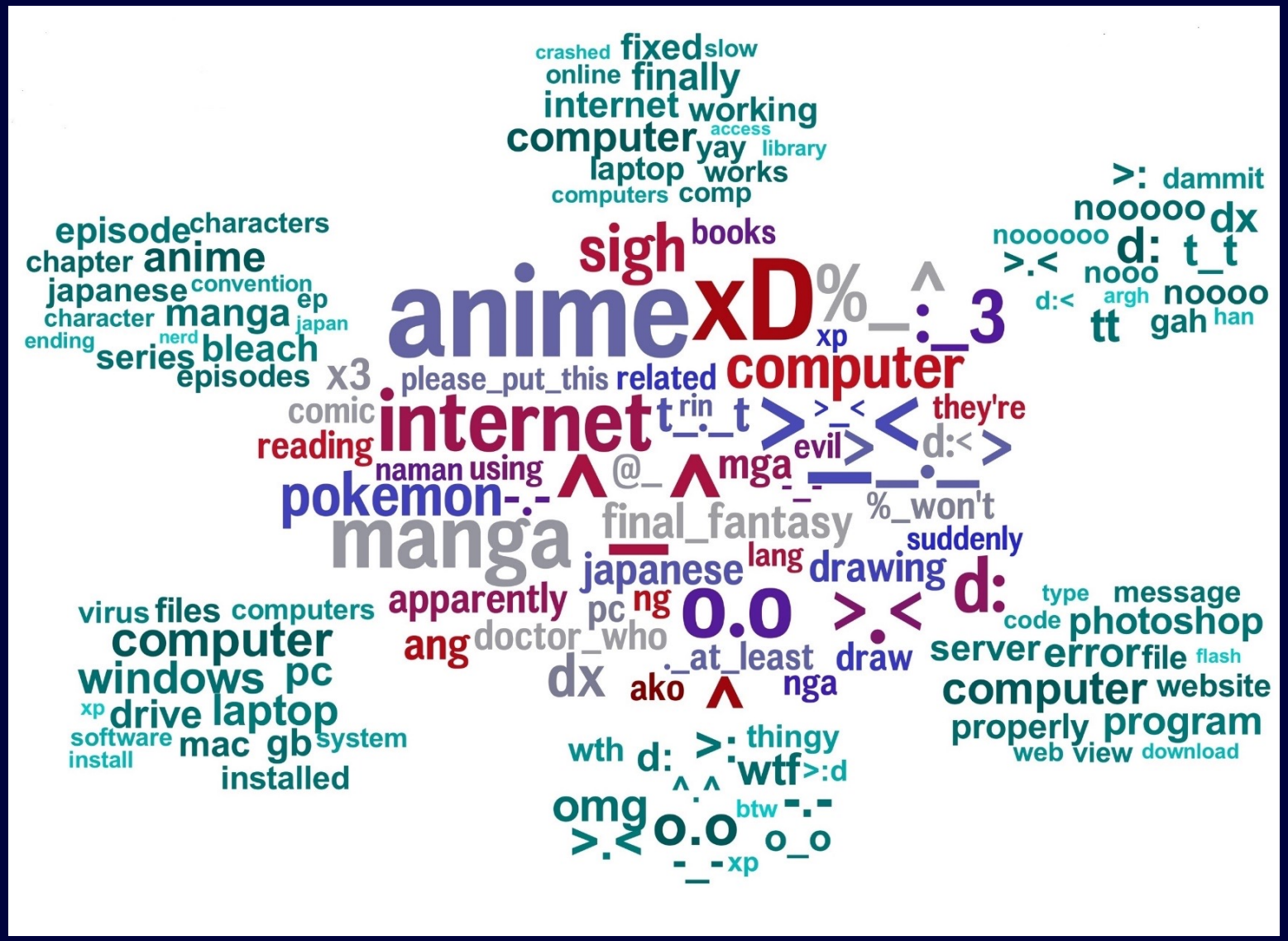
Feedback was the only incentive

40 journal articles since 2011

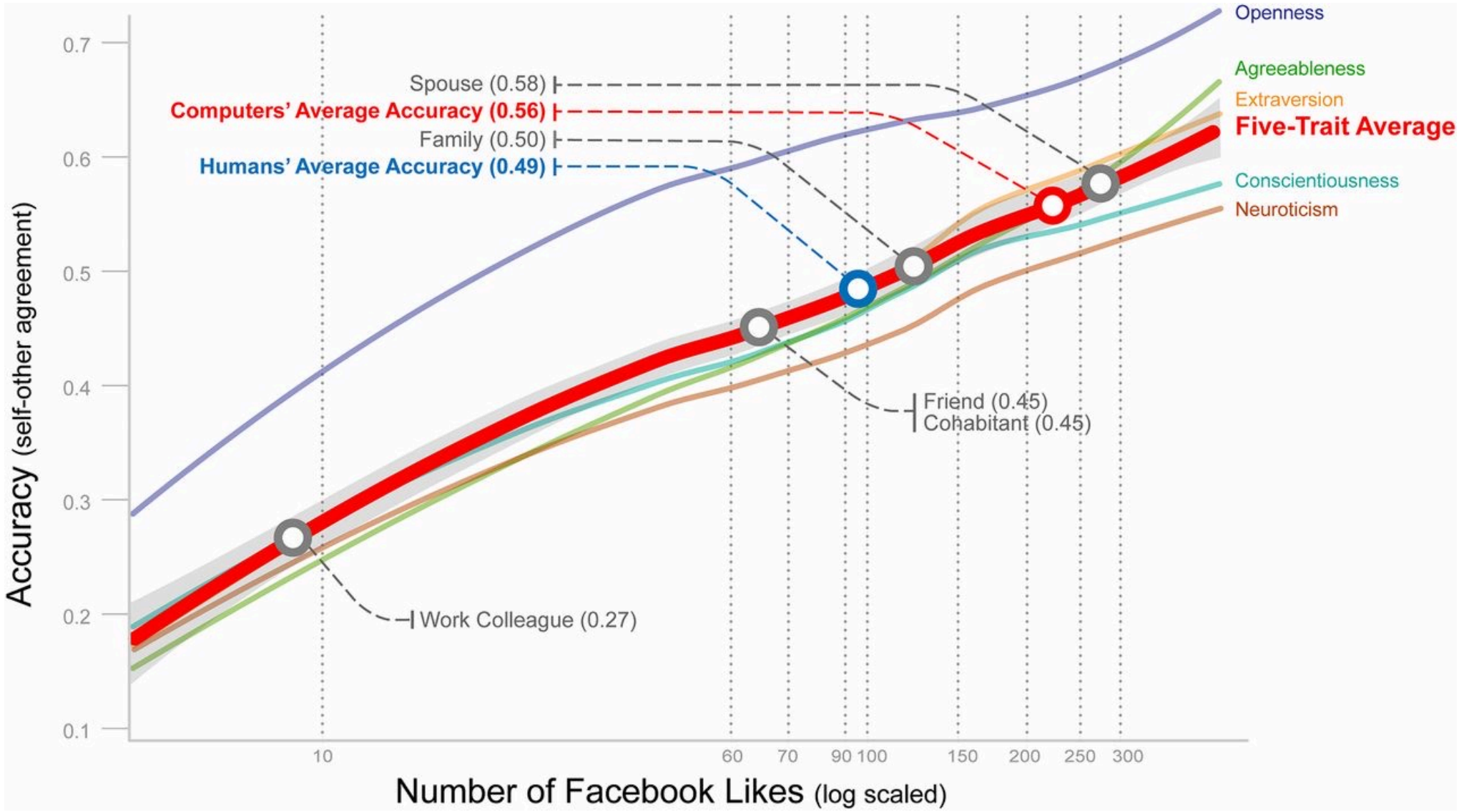
Are you an extrovert?



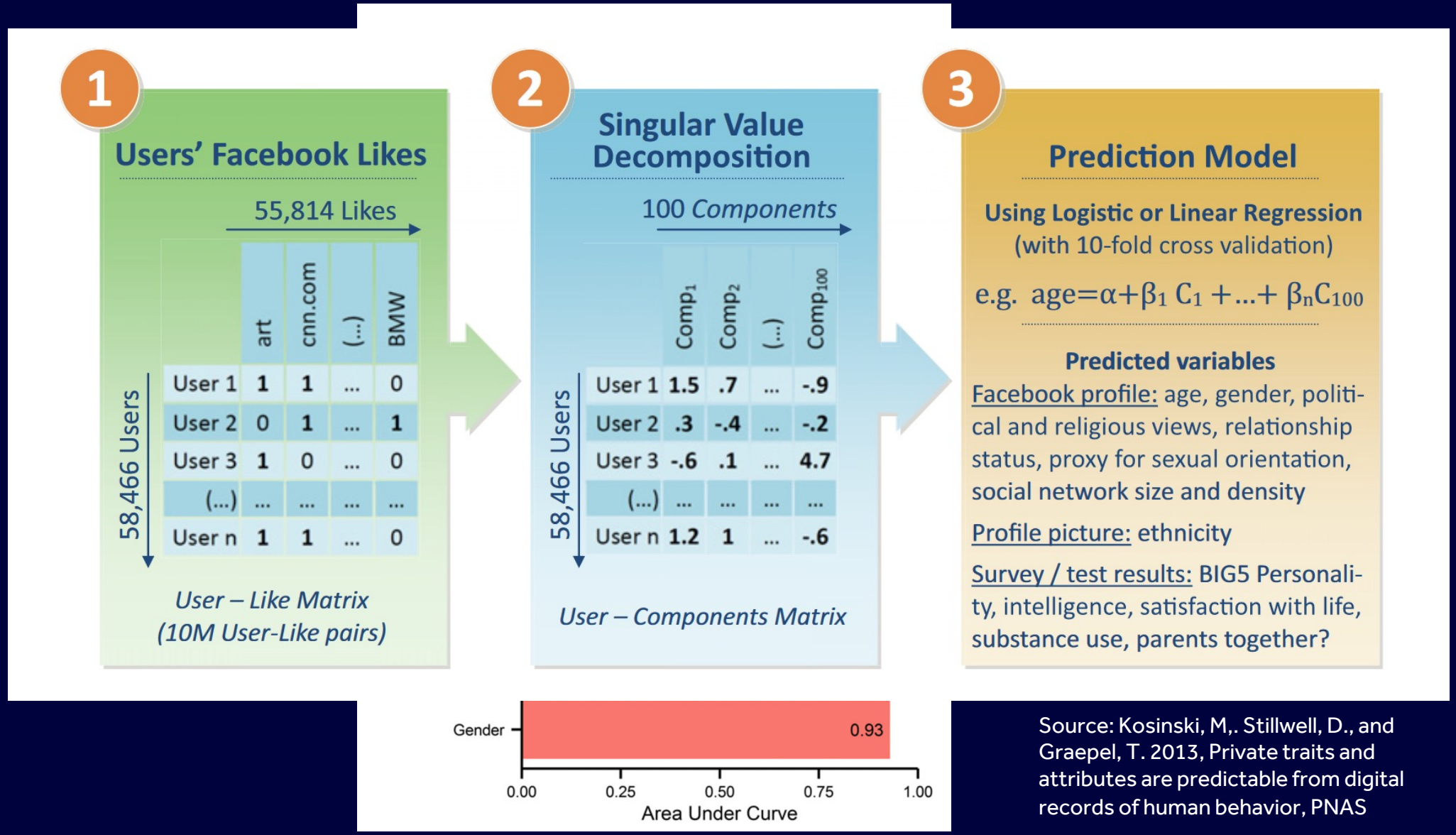
Or are you an introvert?



How do these predictions compare to humans?



They could go further



Source: Kosinski, M,. Stillwell, D., and Graepel, T. 2013, Private traits and attributes are predictable from digital records of human behavior, PNAS

Loan application data

- Netzer, Oded, Lemaire, Alain and Herzenstein, Michal, 2019, When Words Sweat: Identifying Signals for Loan Default in the Text of Loan Applications. Columbia Business School Research Paper No. 16-83. Available at SSRN: https://ssrn.com/abstract=2865327

Examine Prosper – a peer-to-peer lending site.

Potential borrowers write a brief description of **why they need a loan and why they are likely to pay the lender back.** Potential lenders decide whether to lend them the money.

13% of borrowers defaulted on their loan.

The authors use text-mining and machine-learning tools to automatically process and analyze the raw text in over 120,000 loan requests

Which borrower will pay you back?

Borrower #1 writes “I am a hard working person, married for 25 years, and have two wonderful boys. Please let me explain why I need help. I would use the $2,000 loan to fix our roof. Thank you, God bless you, and I promise to pay you back.”

Borrower #2 writes “While the past year in our new place has been more than great, the roof is now leaking and I need to borrow $2,000 to cover the cost of the repair. I pay all bills (e.g., car loans, cable, utilities) on time.”

Will a borrower pay you back?

Generally, **if someone tells you he will pay you back, he will not pay you back.** The more assertive the promise, the more likely he will break it. If someone writes “I promise I will pay you back, so help me God,” he is among the least likely to pay you back.

Religion does matter: Someone who mentions god is 2.2 times more likely to default.

The ultimate goal of big data

The pitfalls of big data

- Correlations
  - PCA or discriminant analysis
- 2012: Kaggle and the used car quality problem
- 1854: John Snow vs. William Farr

Making inferences

The economist cannot see what you think, only what you do.

If I offer you an apple or a banana at the same price, and you pick the banana what do I infer?

But what are you reacting to? Suppose you prefer:

- apples > bananas
- organic > regular and
- ripe > green.

- Are you choosing between ripe organic bananas and regular green apples? What dimension is more important to you? What happens if there are more dimensions (how and where it was grown, its sugar content, nutritional value, and shelf life)?

We are bad at processing more than a few dimensions

- Except for visual patterns, the human brain isn't very good at processing huge amounts of information.
  - •Huma gle about half a dozen distinct pieces time—we can't even compare three erent products.

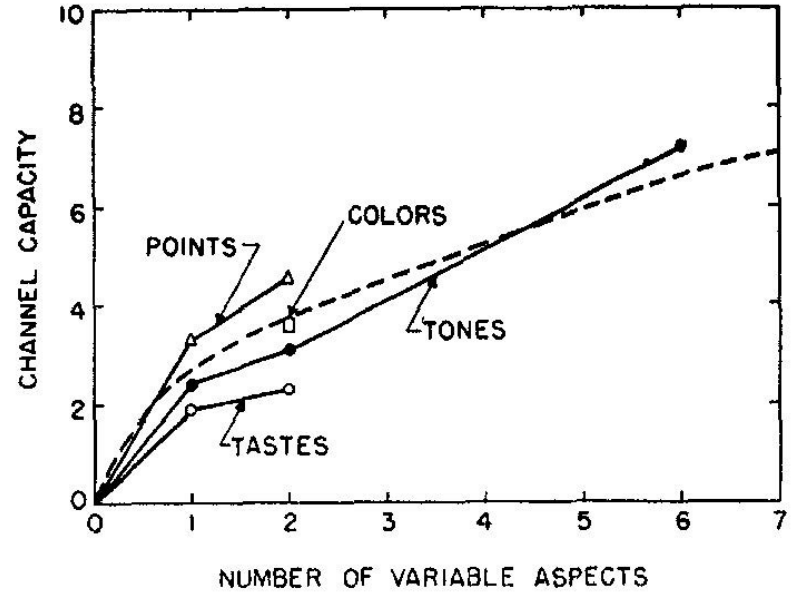
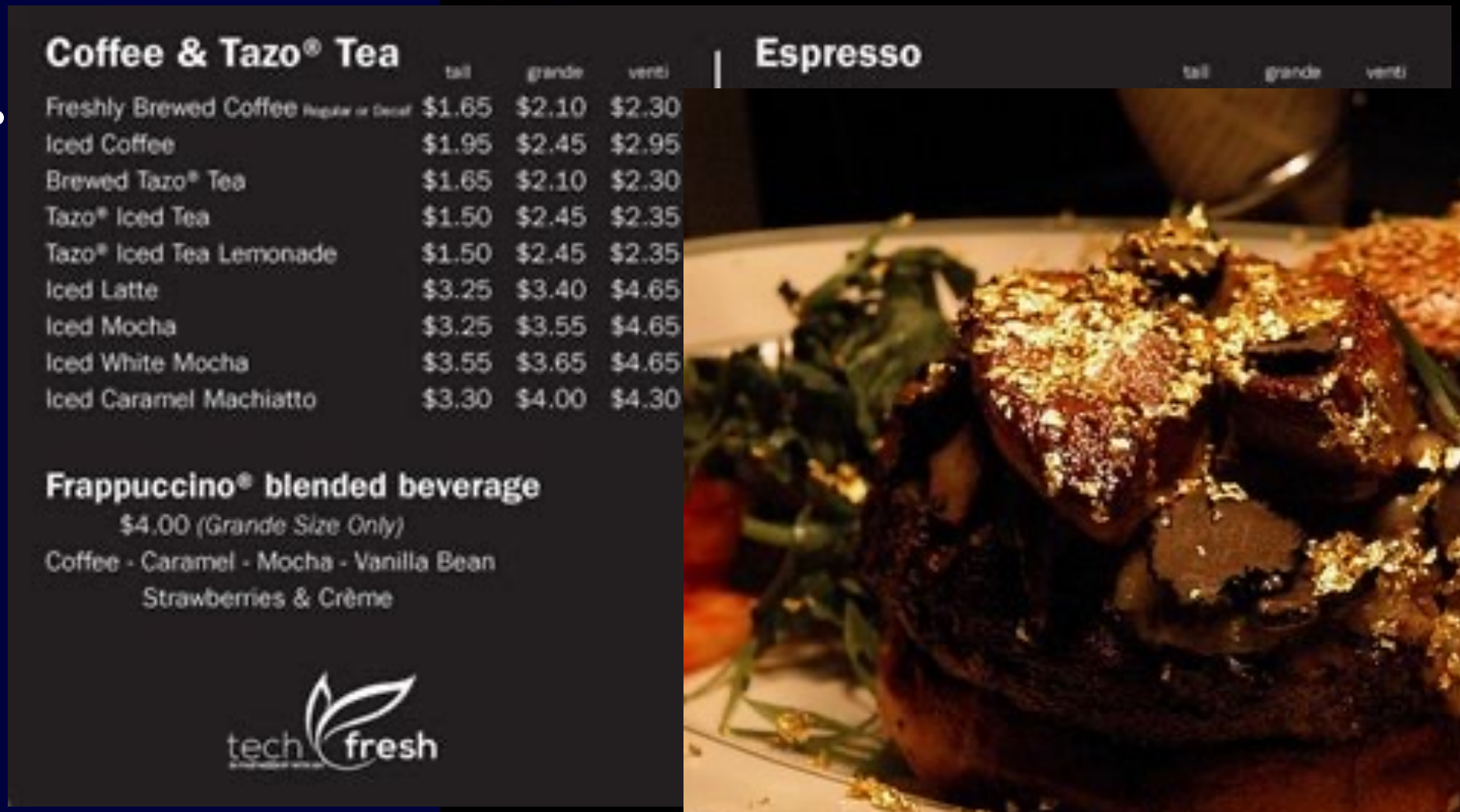


FIG. 6. The general form of the relation between channel capacity and the number of independently variable attributes of the stimuli.

George A. Miller, "The Magical Number Seven Plus or Minus Two: Some Limits on Our Capacity for Processing Information, Psy Review, 63 (2), 1956

And we can easily be manipulated

- 

FOR THE LOVE OF LEARNING SINCE 1597