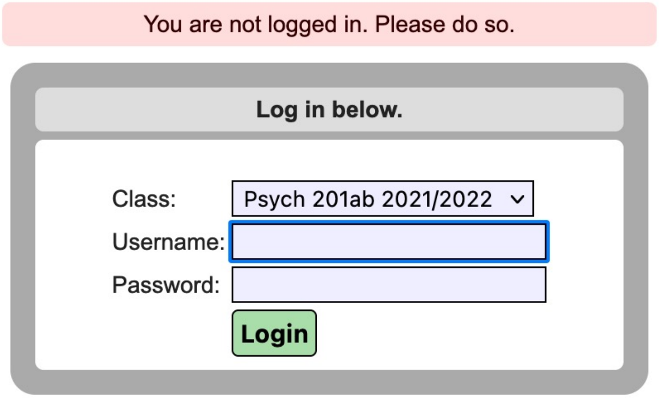
201a:
HW,
Data,
Projects,
and starting with R.

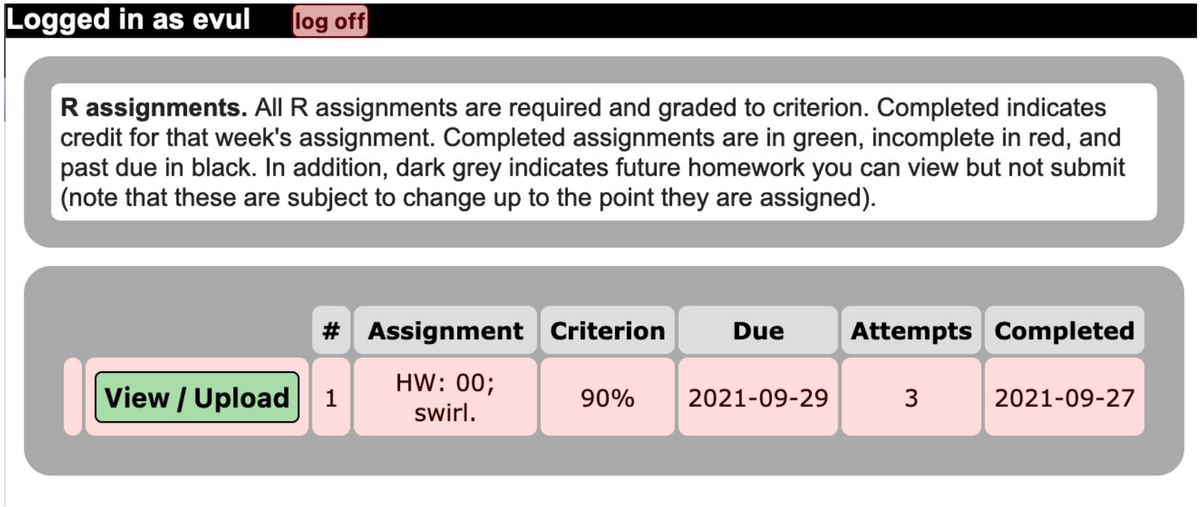
First: HW



You are not logged in. Please do so.

Log in below.

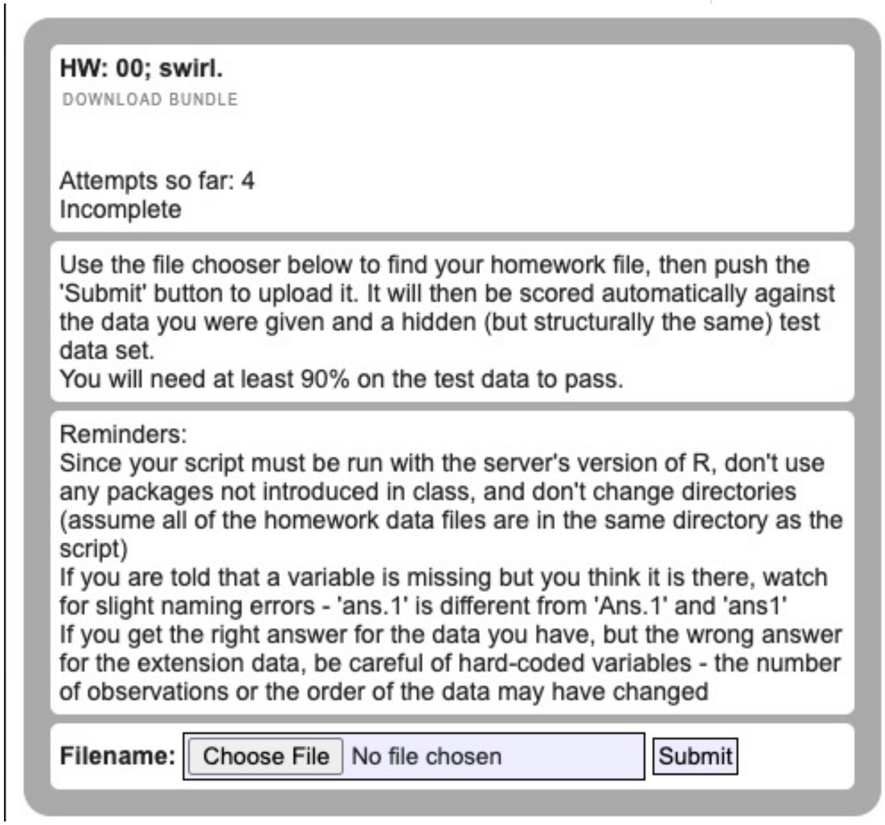
Class:	Psych 201ab 2021/2022
Username:	
Password:	



Logged in as evul

R assignments. All R assignments are required and graded to criterion. Completed indicates credit for that week's assignment. Completed assignments are in green, incomplete in red, and past due in black. In addition, dark grey indicates future homework you can view but not submit (note that these are subject to change up to the point they are assigned).

	#	Assignment	Criterion	Due	Attempts	Completed
View / Upload	1	HW: 00; swirl.	90%	2021-09-29	3	2021-09-27



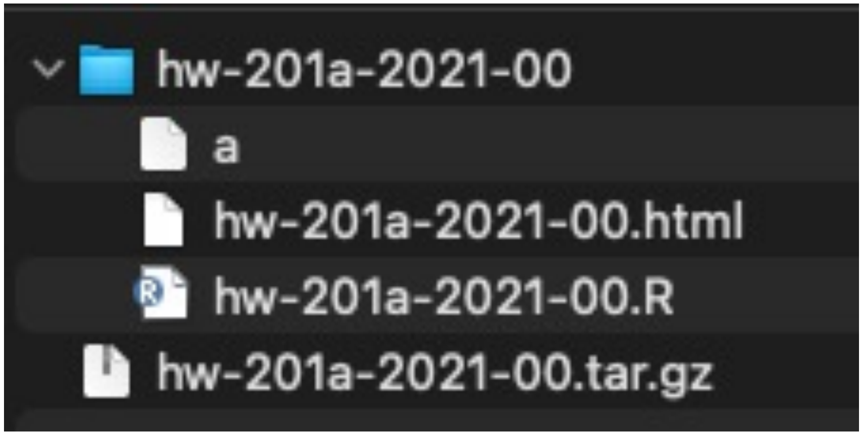
HW: 00; swirl.

Attempts so far: 4
Incomplete

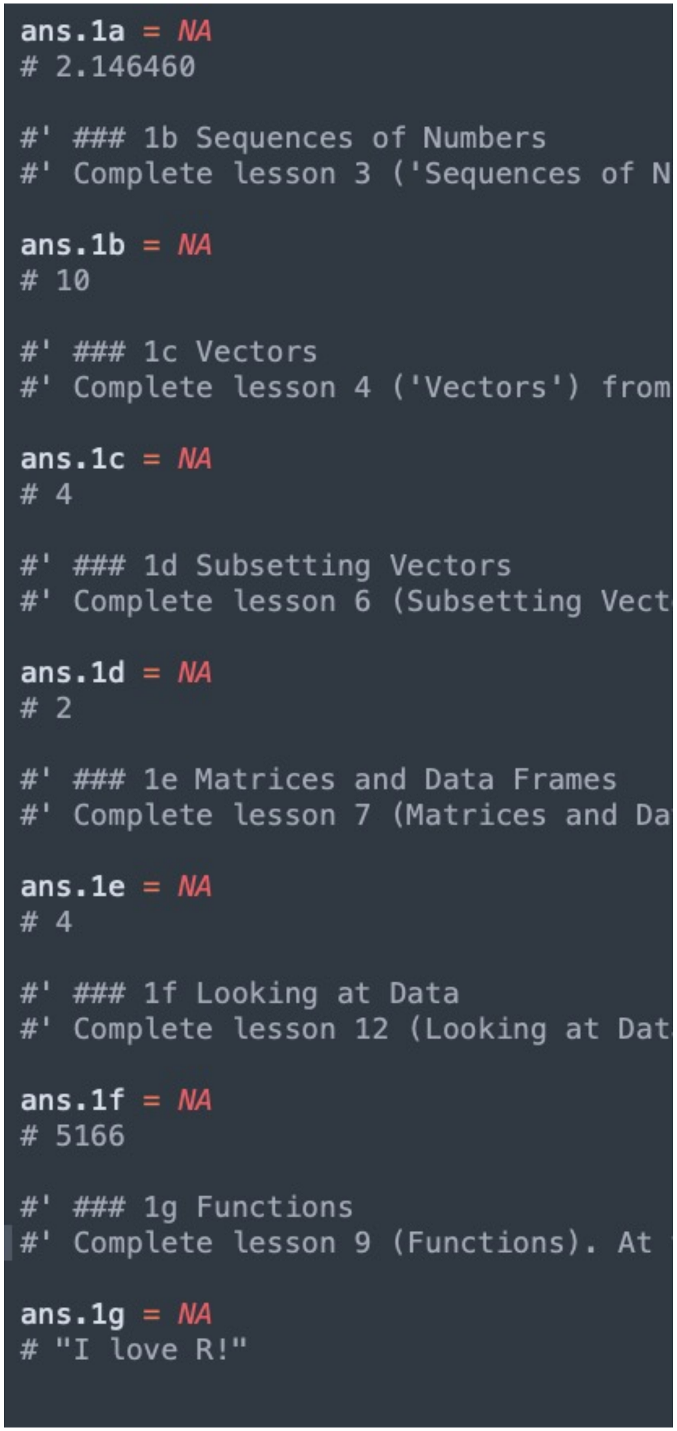
Use the file chooser below to find your homework file, then push the 'Submit' button to upload it. It will then be scored automatically against the data you were given and a hidden (but structurally the same) test data set.
You will need at least 90% on the test data to pass.

Reminders:
Since your script must be run with the server's version of R, don't use any packages not introduced in class, and don't change directories (assume all of the homework data files are in the same directory as the script)
If you are told that a variable is missing but you think it is there, watch for slight naming errors - 'ans.1' is different from 'Ans.1' and 'ans1'
If you get the right answer for the data you have, but the wrong answer for the extension data, be careful of hard-coded variables - the number of observations or the order of the data may have changed

Filename: Choose File No file chosen



```
hw-201a-2021-00
    a
    hw-201a-2021-00.html
    hw-201a-2021-00.R
hw-201a-2021-00.tar.gz
```



```
ans.1a = NA
# 2.146460

#' ### 1b Sequences of Numbers
#' Complete lesson 3 ('Sequences of N

ans.1b = NA
# 10

#' ### 1c Vectors
#' Complete lesson 4 ('Vectors') from

ans.1c = NA
# 4

#' ### 1d Subsetting Vectors
#' Complete lesson 6 (Subsetting Vect

ans.1d = NA
# 2

#' ### 1e Matrices and Data Frames
#' Complete lesson 7 (Matrices and Da

ans.1e = NA
# 4

#' ### 1f Looking at Data
#' Complete lesson 12 (Looking at Dat

ans.1f = NA
# 5166

#' ### 1g Functions
#' Complete lesson 9 (Functions). At

ans.1g = NA
# "I love R!"
```

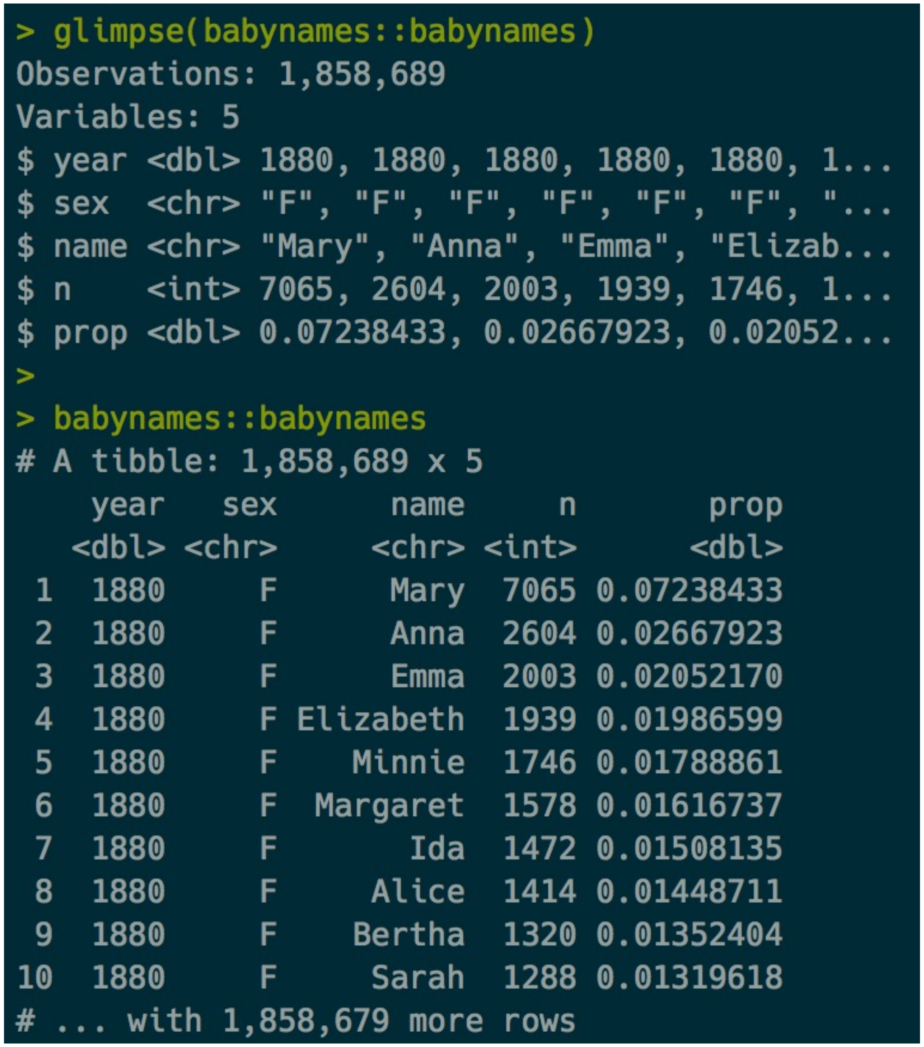
Understanding your data.

- Where did the data come from?
  i.e., what was measured? How?
  Sampling process?
  Experiment? Survey? Controls? Random assignment?

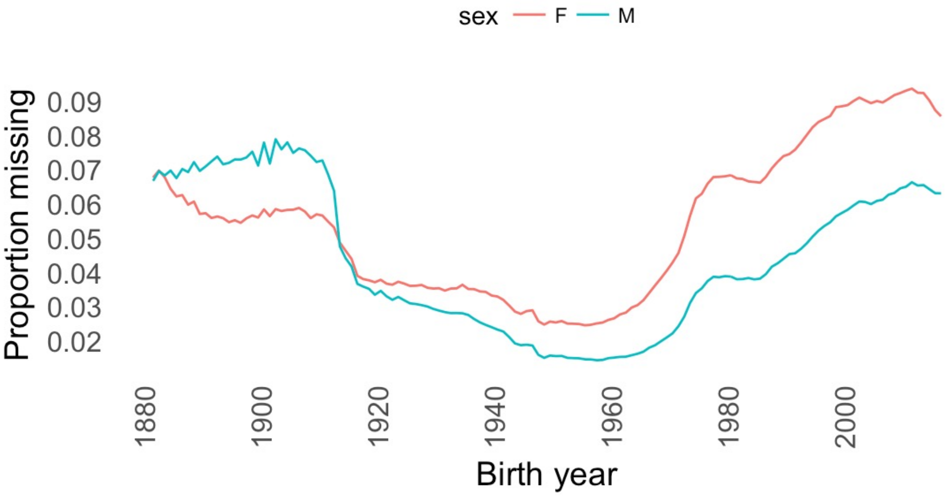
- What types of variables are you dealing with?

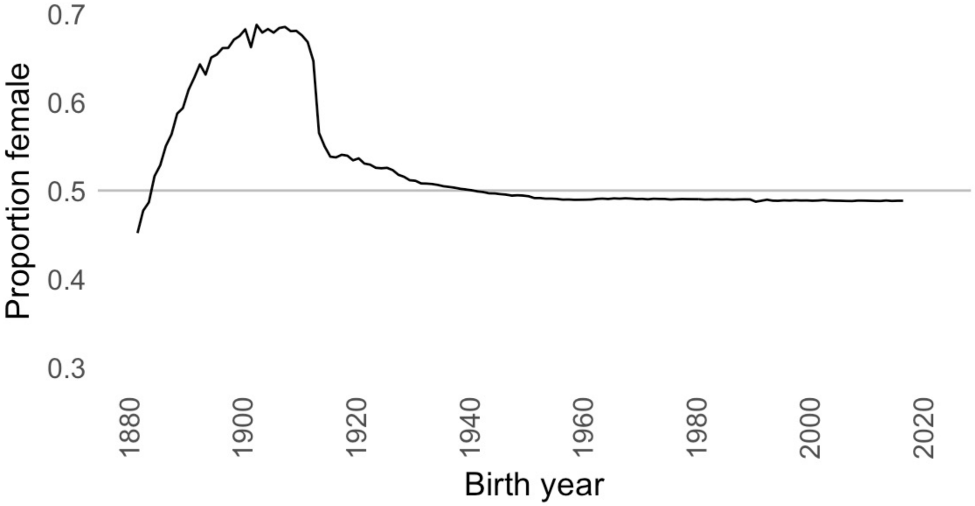
- What is the structure among measurements?

Understand where your data came from, or you will be confused.



```
> glimpse(babynames::babynames)
Observations: 1,858,689
Variables: 5
$ year <dbl> 1880, 1880, 1880, 1880, 1880, 1...
$ sex  <chr> "F", "F", "F", "F", "F", "F", "...
$ name <chr> "Mary", "Anna", "Emma", "Elizab...
$ n    <int> 7065, 2604, 2003, 1939, 1746, 1...
$ prop <dbl> 0.07238433, 0.02667923, 0.02052...
>
> babynames::babynames
# A tibble: 1,858,689 x 5
    year   sex      name     n       prop
   <dbl> <chr>     <chr> <int>      <dbl>
 1  1880     F      Mary  7065 0.07238433
 2  1880     F      Anna  2604 0.02667923
 3  1880     F      Emma  2003 0.02052170
 4  1880     F Elizabeth  1939 0.01986599
 5  1880     F    Minnie  1746 0.01788861
 6  1880     F  Margaret  1578 0.01616737
 7  1880     F       Ida  1472 0.01508135
 8  1880     F     Alice  1414 0.01448711
 9  1880     F    Bertha  1320 0.01352404
10  1880     F     Sarah  1288 0.01319618
# ... with 1,858,679 more rows
```





Measurement procedure

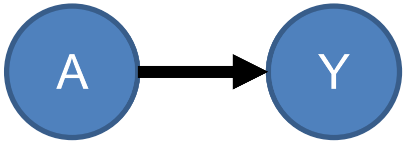
- Sampling:
  - Gold standard (Probability) vs. practice (convenience)
  - Non-representativeness (bias, variance, drop-out, demand)

- Intervention: Observation / Survey / Experiment
  - Control and randomization for causal inference.

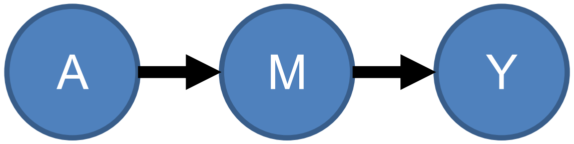
Causality

- Intuitions:
  - Counterfactual?
  - Intervention?

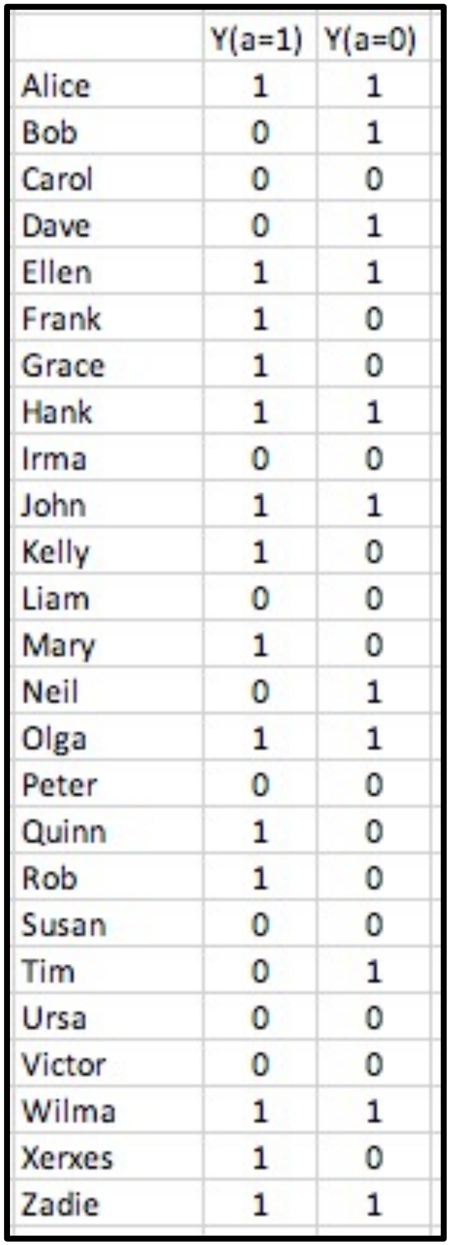
- Representing with Directed, Acyclic Graphs (DAGs, Bayes Nets)



“A causes Y”



“mediation”

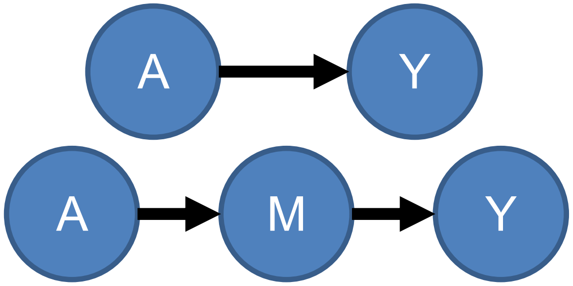


	Y(a=1)	Y(a=0)
Alice	1	1
Bob	0	1
Carol	0	0
Dave	0	1
Ellen	1	1
Frank	1	0
Grace	1	0
Hank	1	1
Irma	0	0
John	1	1
Kelly	1	0
Liam	0	0
Mary	1	0
Neil	0	1
Olga	1	1
Peter	0	0
Quinn	1	0
Rob	1	0
Susan	0	0
Tim	0	1
Ursa	0	0
Victor	0	0
Wilma	1	1
Xerxes	1	0
Zadie	1	1

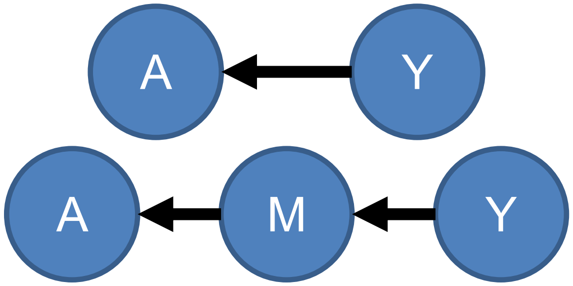
Association ≠ Causation

- A and Y co-occur.
- But why?

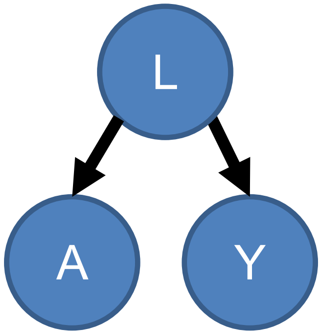
A causes Y somehow



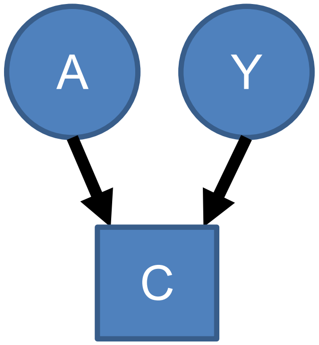
Y causes A somehow

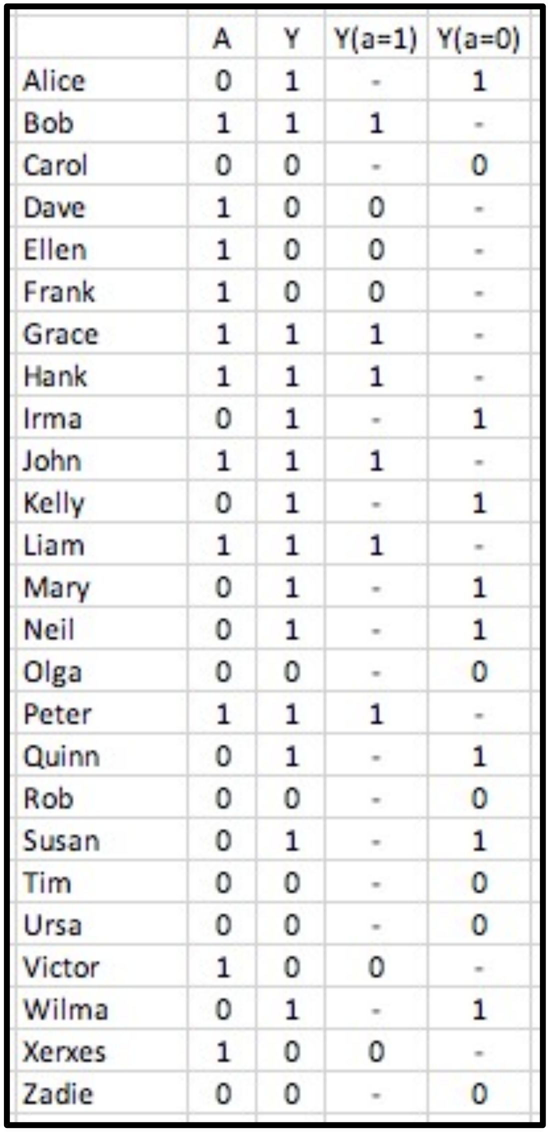


"Common cause"
L causes A and Y



Conditioning
on a collider



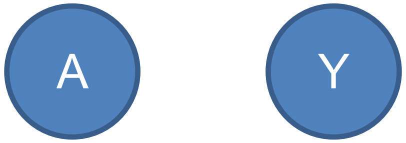


	A	Y	Y(a=1)	Y(a=0)
Alice	0	1	-	1
Bob	1	1	1	-
Carol	0	0	-	0
Dave	1	0	0	-
Ellen	1	0	0	-
Frank	1	0	0	-
Grace	1	1	1	-
Hank	1	1	1	-
Irma	0	1	-	1
John	1	1	1	-
Kelly	0	1	-	1
Liam	1	1	1	-
Mary	0	1	-	1
Neil	0	1	-	1
Olga	0	0	-	0
Peter	1	1	1	-
Quinn	0	1	-	1
Rob	0	0	-	0
Susan	0	1	-	1
Tim	0	0	-	0
Ursa	0	0	-	0
Victor	1	0	0	-
Wilma	0	1	-	1
Xerxes	1	0	0	-
Zadie	0	0	-	0

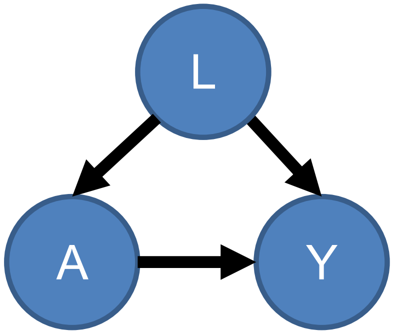
No Association ≠ No Causation

- A and Y do not co-occur.
- But why not?

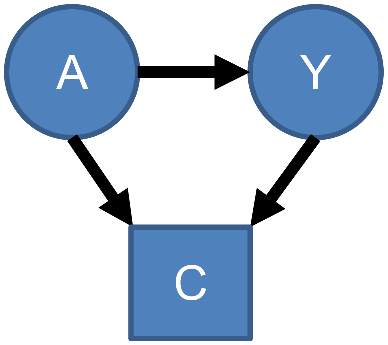
A and Y are independent



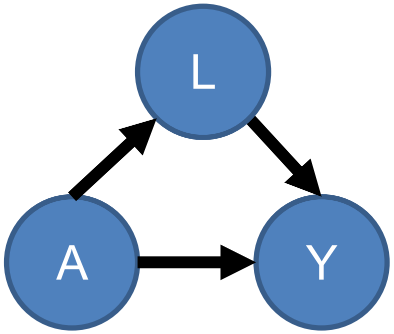
Masking common cause

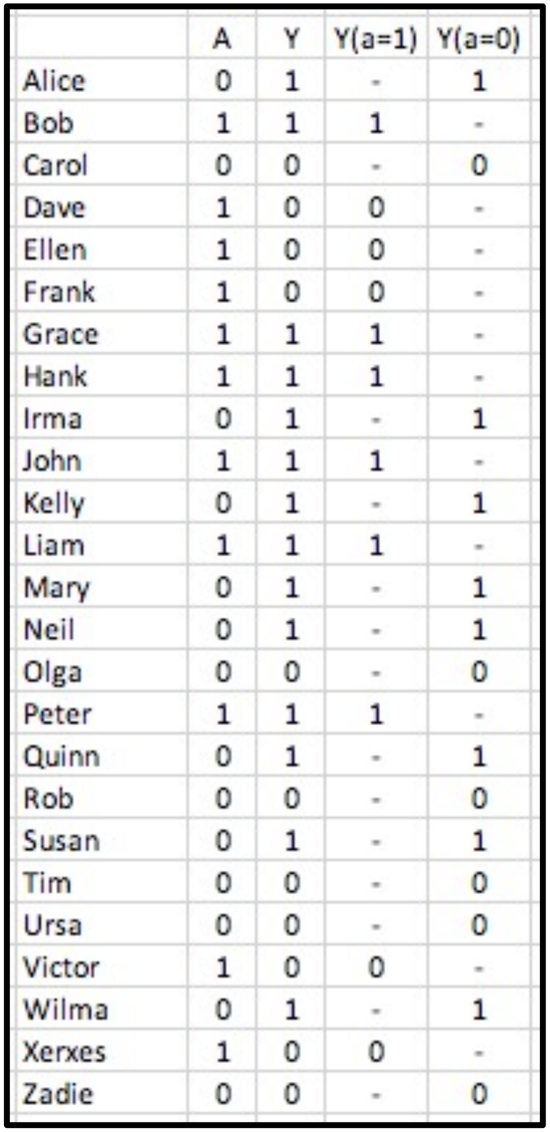


Conditioning on a collider



Confounder

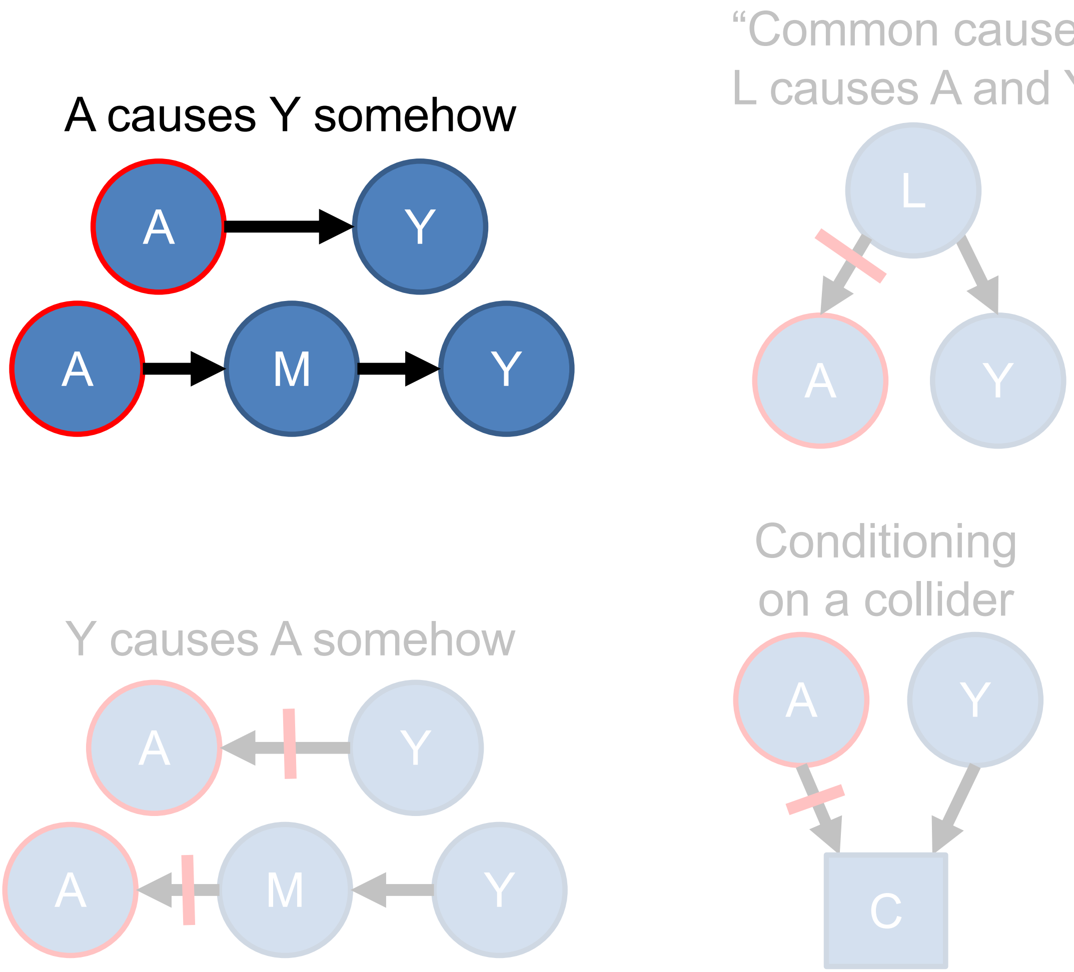




	A	Y	Y(a=1)	Y(a=0)
Alice	0	1	-	1
Bob	1	1	1	-
Carol	0	0	-	0
Dave	1	0	0	-
Ellen	1	0	0	-
Frank	1	0	0	-
Grace	1	1	1	-
Hank	1	1	1	-
Irma	0	1	-	1
John	1	1	1	-
Kelly	0	1	-	1
Liam	1	1	1	-
Mary	0	1	-	1
Neil	0	1	-	1
Olga	0	0	-	0
Peter	1	1	1	-
Quinn	0	1	-	1
Rob	0	0	-	0
Susan	0	1	-	1
Tim	0	0	-	0
Ursa	0	0	-	0
Victor	1	0	0	-
Wilma	0	1	-	1
Xerxes	1	0	0	-
Zadie	0	0	-	0

Randomized Experiment (idealized)

- We manipulate A; Y happens.



A causes Y somehow

“Common cause”
L causes A and Y

Y causes A somehow

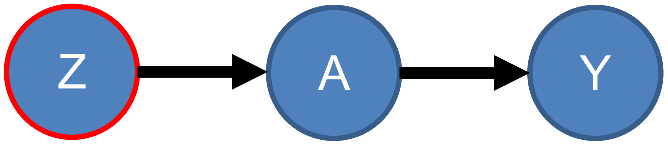
Conditioning
on a collider

Assuming we randomly assign A, and measure everyone, so no opportunity to condition on A.

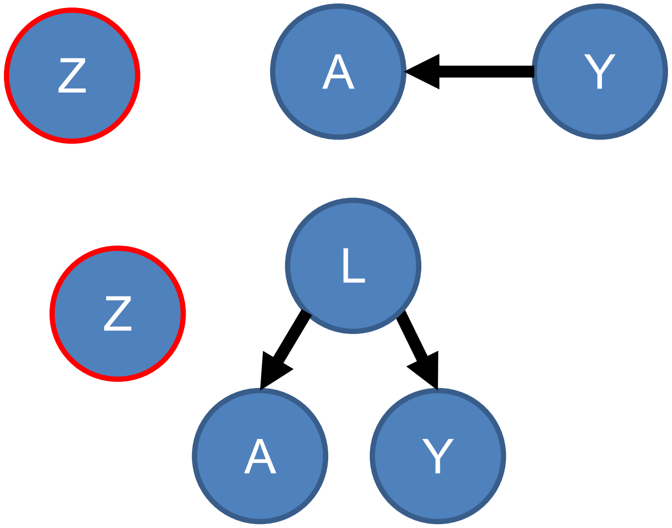
Randomized Experiment (realistic)

- We do Z to manipulate A; Y happens.

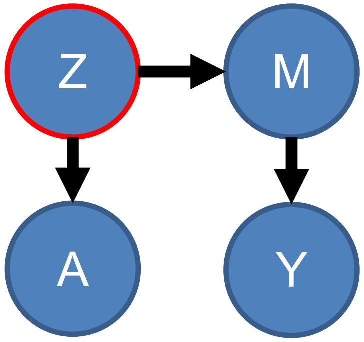
Z causes A, which causes Y



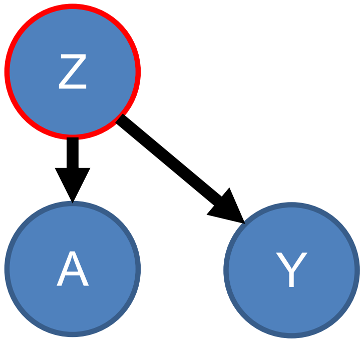
Z doesn't work. Oops.



"confound"



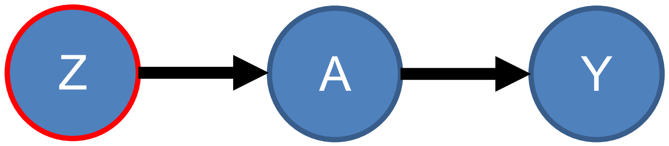
"demand effect"



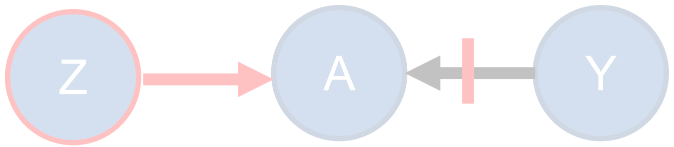
Randomized Experiment (realistic)

- We do Z to manipulate A; Y happens.

Z causes A, which causes Y



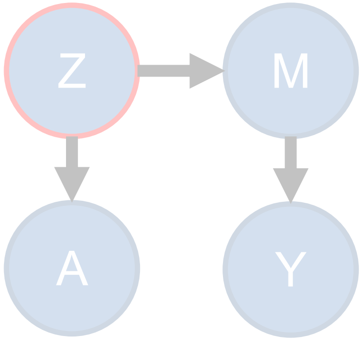
Z doesn’t work. Oops.



Measure A:
A manipulation check to ensure Z -> A.
Instrumental variable logic

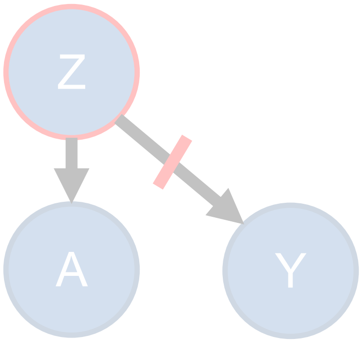
“confounds”

Lots of additional measures and controls.



“demand effect”

Double-blind design



Understanding your data.

- Where did the data come from?
  i.e., what was measured? How?
  Sampling process?
  Experiment? Survey? Controls? Random assignment?

- What types of variables are you dealing with?

- What is the structure among measurements?

Types of variables / measurements.

- SS Stevens' measurement scale
  - Categorical / qualitative / nominal
  - Numerical / quantitative: Ordinal? Interval? Ratio?

- Statistical / mathematical variable type
  - Support (what values can it take on)?
  - Distribution (what does variation look like)?

Canonical plots for scales

As a function of… (independent variable)

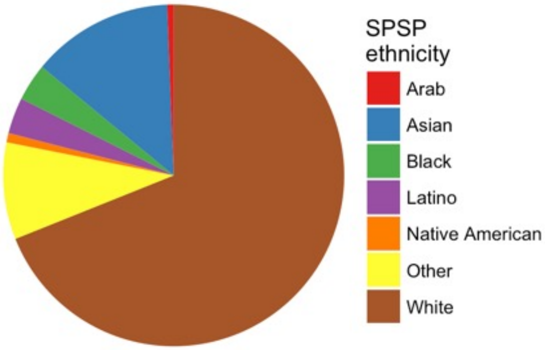
… ~ 0

… ~ categorical

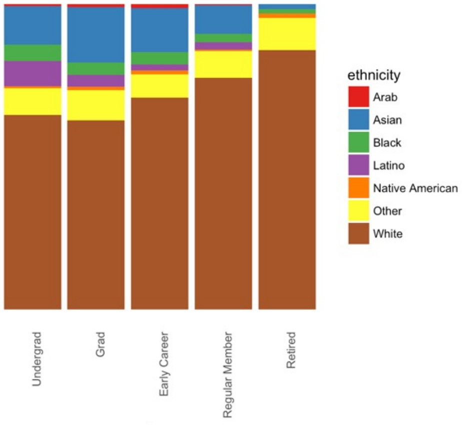
… ~ numerical

Response / dependent variable

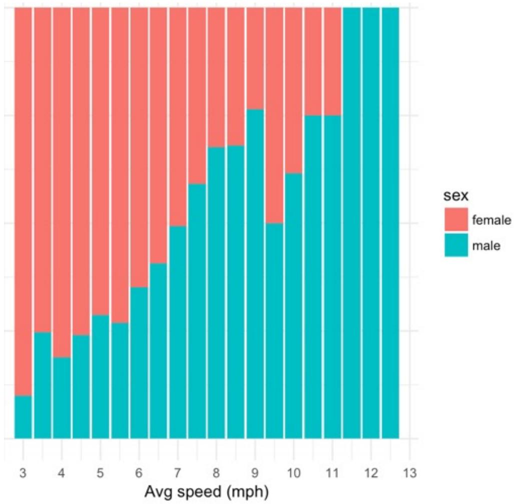
categorical ~ …



pie chart (ugh)

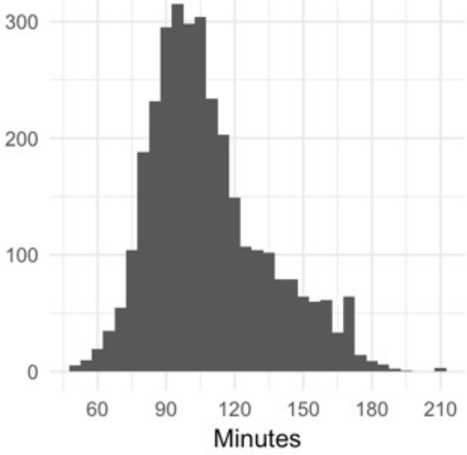


stacked area

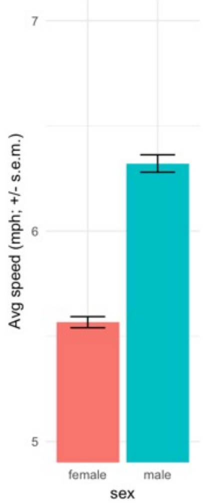


stacked area

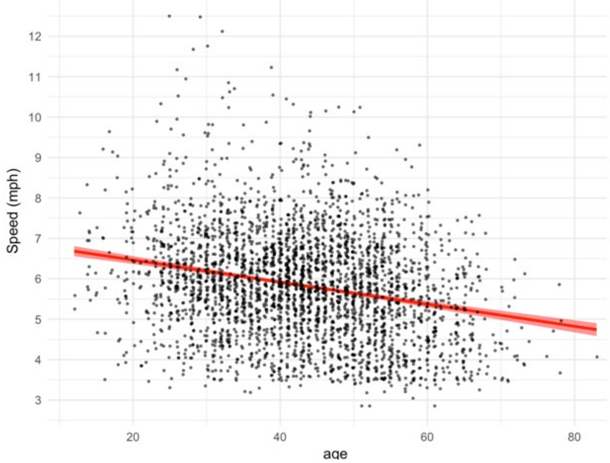
numerical ~ …



histogram



bar: mean + error

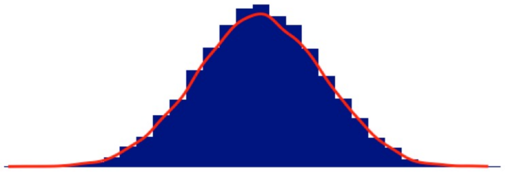
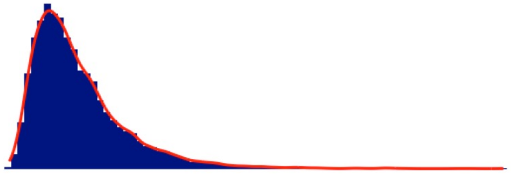
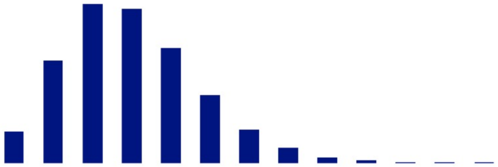
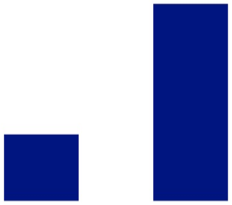
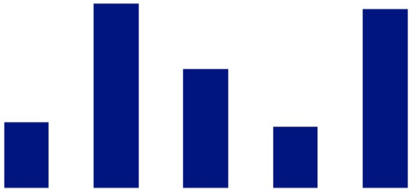


Scatterplot / conditional mean

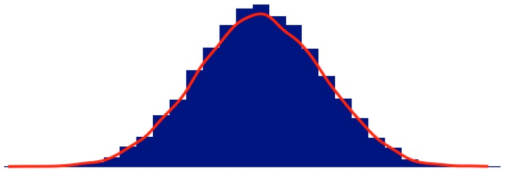
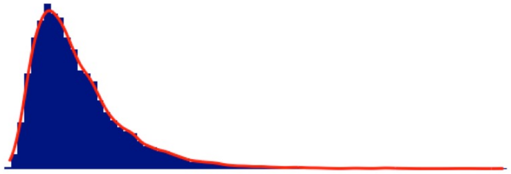
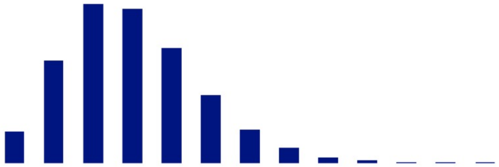
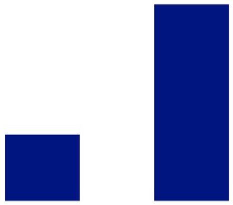
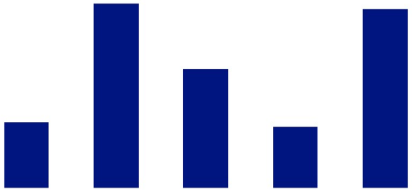
Overly specific named procedures

Response	~null	~binary	~category	~numerical	~numerical + category
Numerical	1-sample T-test	2-sample T-test	ANOVA	Regression, Pearson correlation	ANCOVA
Ranked-numerical		Mann-Whitney-U	Kruskall-Wallis	Spearman correlation	
2-category	Binomial test	Fisher's exact test	Chi-sq. indep.	Logistic regression	
k-category	Chi-sq. goodness of fit	Chi-squared independence			

Common types of variables.

Stevens'	Process	Support	Examples	Sample histogram
Numerical	Additive	[-∞, ∞]	Temperature Height IQ score	
Numerical	Multiplicative	[0, ∞]	Income GDP/capita Wealth	
Numerical	Counts	{0,1,2,…}	Calls received Crime count Spike count	
Categorical	2-values	{0,1}	Correct/incorrect Success/failure Win/lose	
Categorical	K-values	{a,b,c,…}	Shirt color Political party Major	

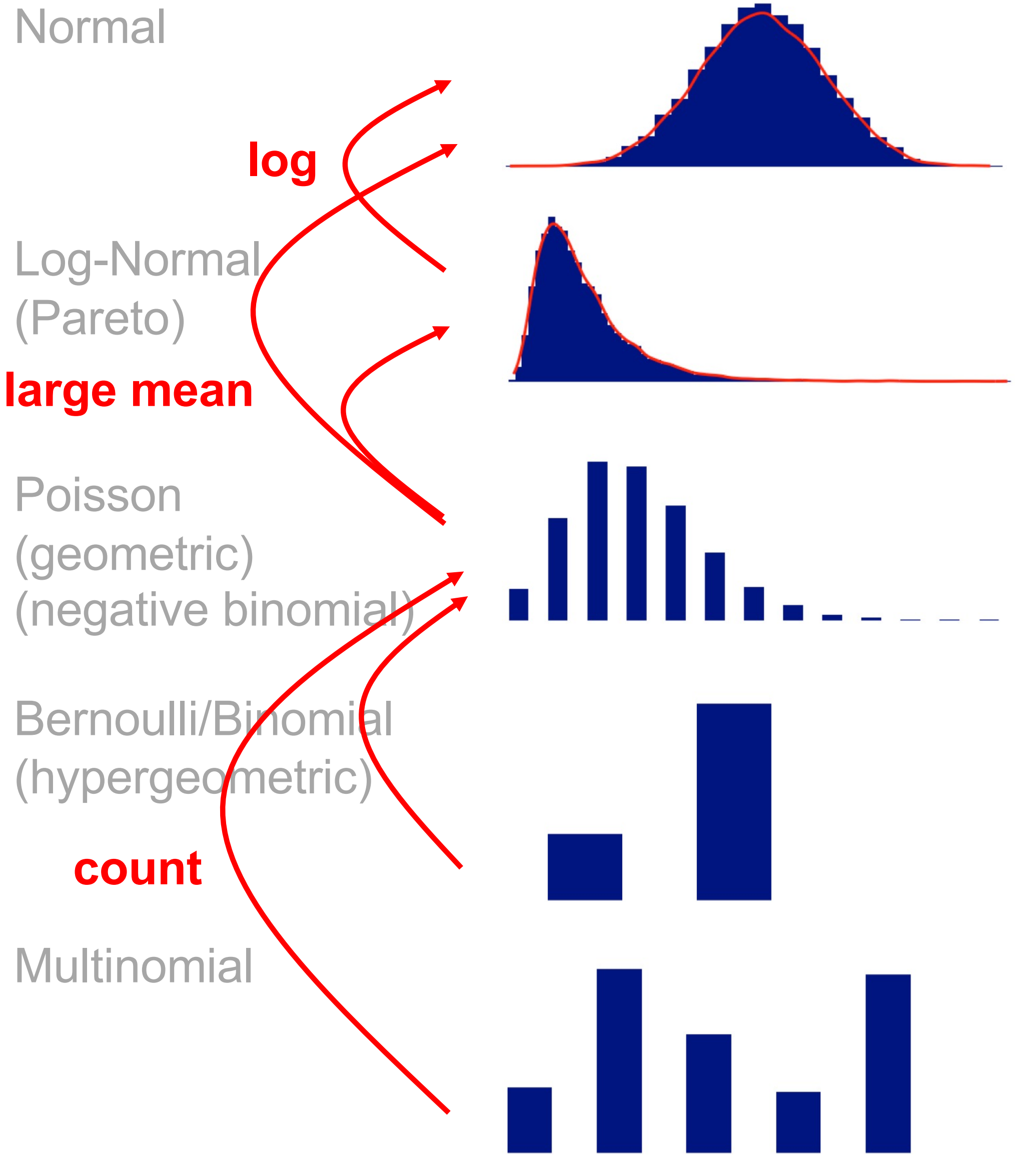
Common types of variables.

Stevens'	Process	Support	Distributions	Sample histogram
Numerical	Additive	[-∞, ∞]	Normal	
Numerical	Multiplicative	[0, ∞]	Log-Normal (Pareto)	
Numerical	Counts	{0,1,2,…}	Poisson (geometric) (negative binomial)	
Categorical	2-values	{0,1}	Bernoulli/Binomial (hypergeometric)	
Categorical	K-values	{a,b,c,…}	Multinomial	

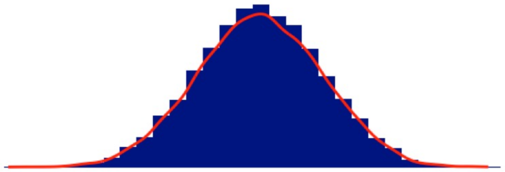
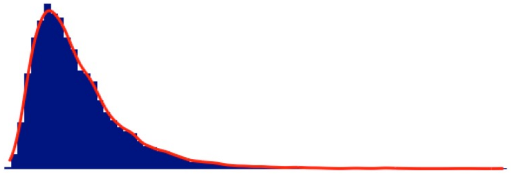
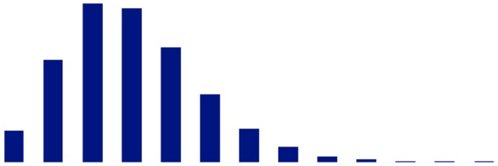
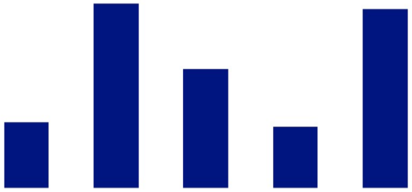
Common types of variables.

Stevens'	Process	Support	Distributions
Numerical	Additive	[-∞, ∞]	Normal
Numerical	Multiplicative	[0, ∞]	Log-Normal (Pareto)
Numerical	Counts	{0,1,2,…}	Poisson (geometric) (negative binomial)
Categorical	2-values	{0,1}	Bernoulli/Binomial (hypergeometric)
Categorical	K-values	{a,b,c,…}	Multinomial

Sample histogram



Canonical models

Distributions	Sample histogram	Model for this type of response.
Normal		Linear model lm(y ~ ...) *ranked: lm(rank(y) ~ …)*
Log-Normal		Log-linear model lm(log(y) ~ ...)
Poisson		Poisson-log-linear (generalized LM) glm(y~…, family=poisson())
Binomial		Logistic regression (generalized LM) glm(y~…, family=binomial())
Multinomial		Multinomial logistic regression Treat as counts -> poisson Chi-squared test

Measurements

- Psychometric properties:
  - *Fidelity/resolution*, Reliability, validity.

- SS Stevens' measurement scale
  - Qualitative / categorical / nominal
  - Quantitative / numerical: Ordinal? Interval? Ratio?

- Statistical / mathematical variable type
  - Support (what values can it take on)?
  - Distribution (what does variation look like)?

- Downgrading / upgrading scales?

Understanding your data.

- Where did the data come from?
  i.e., what was measured? How?
  Sampling process?
  Experiment? Survey? Controls? Random assignment?

- What types of variables are you dealing with?

- What is the structure among measurements?

Structure among measurements?

- What are the units being measured?
  - E.g., trials, individuals

- What are the relationships among units?
  - E.g., multiple trials per individual

- We will deal with flat structures for now.

Understanding your data.

- Where did the data come from?
  i.e., what was measured? How?
  Sampling process?
  Experiment? Survey? Controls? Random assignment?

- What types of variables are you dealing with?

- What is the structure among measurements?

Next: Projects

Projects

- Groups of ~4-5; undergrads disperse
- Find a question, and large-scale naturalistic dataset that could offer an answer. Ask the question.
- Goal: Something that could be publishable with a bit more refinement
- Examples: voting from google street view, anything by raj chetty, skill learning in online games, personality tests in blog posts, meme transmission in twitter, etc.
- Trickiness of naturalistic data.
  - Connecting theoretical constructs to real-world phenomena
  - Extracting relevant signals from all the variation.
  - Causality; exogenous vs endogenous variation
- Types of questions asked of naturalistic data.
  - Verifying lab conclusions via natural variation. Does X predict Y? What aspects of X relate to Y, how? Causality from exogenous variation in X? Picking apart structure of X.

How adoption speed affects the abandonment of cultural tastes

Jonah Berger[a,1,2] and Gaël Le Mens[b,c,1]

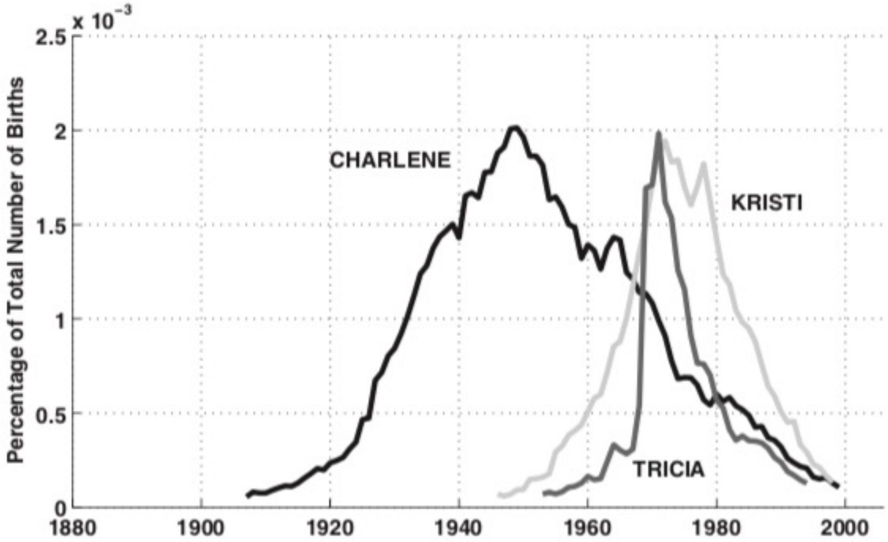


Fig. 1. A few trajectories of first-name popularity (in the U.S.). Most names show a period of almost consistent increase in popularity, followed by a decline that leads to abandonment, but names differ in how quickly their popularity rises and declines.

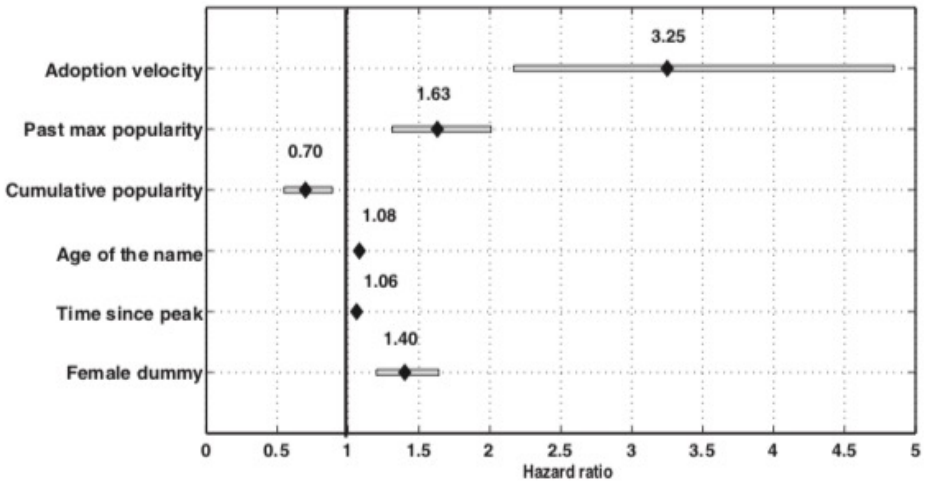


Fig. 2. Hazard ratios and 95% confidence intervals from hazard rate model estimation. The regression equation is: $r_i(y) = \exp(\gamma X_{i,y-1})$, where $r_i(y)$ refers to the instantaneous death rate of name i in year y, $X_{i,y-1}$ is a vector of time-varying covariates, and γ is the vector of estimated coefficients. For each name i and each year y, the past peak in popularity is defined as the past year $Y_{i,y} < y$ at which the contribution of i to all births of the same sex, $F_{i,y}$, was maximal over all past years. The adoption velocity is defined as the rate of change in adoption in the 5 years before $Y_{i,y}$: $\alpha_{i,y} = (F_{i,Y_{i,y}}-5/F_{i,Y_{i,y}})^{1/5} - 1$, where $F_{i,Y_{i,y}}$ is the contribution of name i to all births of the same sex at the past peak in popularity. The mean of the adoption velocity is 19.5%, and the standard deviation is 0.17. The age of a name is defined as the average number of years elapsed between births with name i and the focal year, computed over all past births with name i. The cumulative popularity is the contribution of a name to all births that occurred since it entered our dataset. Popularity and cumulative popularity are normalized for the estimations. The effect of a covariate is significant if the corresponding 95% confidence interval bar does not intersect with 1.0.

Decision contamination in the wild: Sequential dependencies in Yelp review ratings

David W. Vinson, Rick Dale
Cognitive and Information Sciences
University of California, Merced
[dvinson][rdale]@ucmerced.edu

Michael N. Jones
Departments of Psychology, Cognitive Science, and Informatics
Indiana University, Bloomington
jonesmn@indiana.edu

We used the most recent release of the Yelp Inc., dataset, part of Yelp's Dataset Challenge[2]. The dataset consists of just over 2.2 million reviews spanning 12 years from 2004-2016, with ratings between one (negative) and five (positive) stars, from approximately 552,000 reviewers on roughly 77,000 businesses. Reviews were provided from

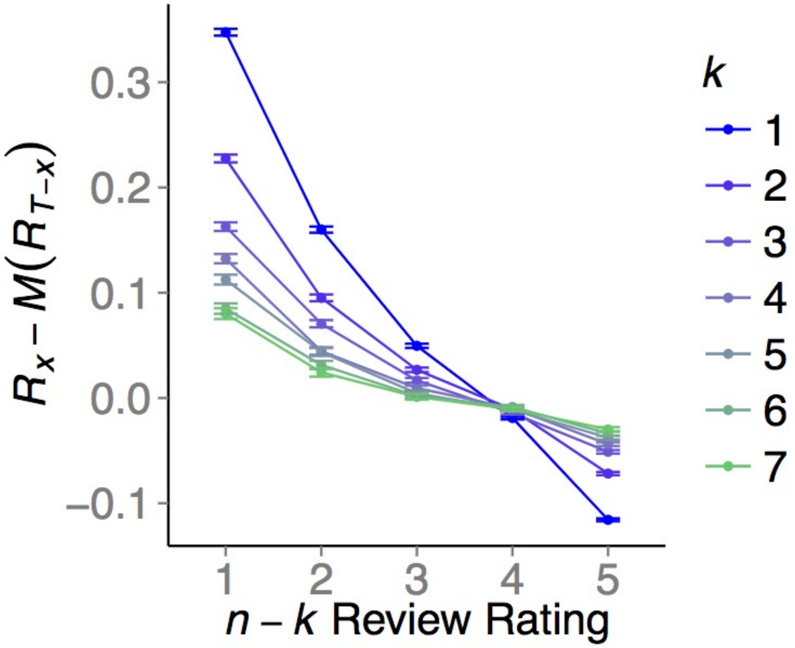


Figure 2: Within-reviewer contrast between previous and current review ratings at *k* Review Distances

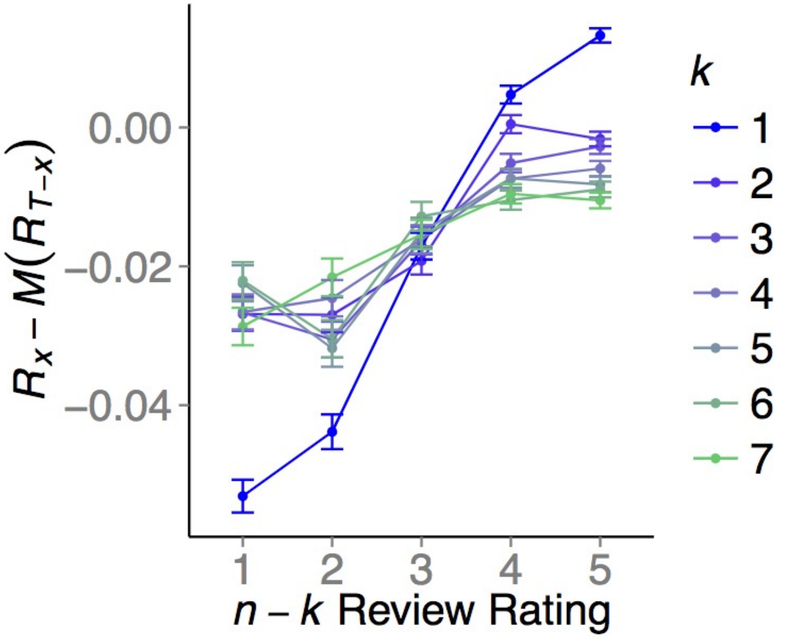


Figure 3: Within-business assimilation between previous and current review ratings at *k* Review Distances

Female chess players outperform expectations when playing men

Tom Stafford
Department of Psychology, University of Sheffielc
Sheffield, S1 2LT, United Kingdom

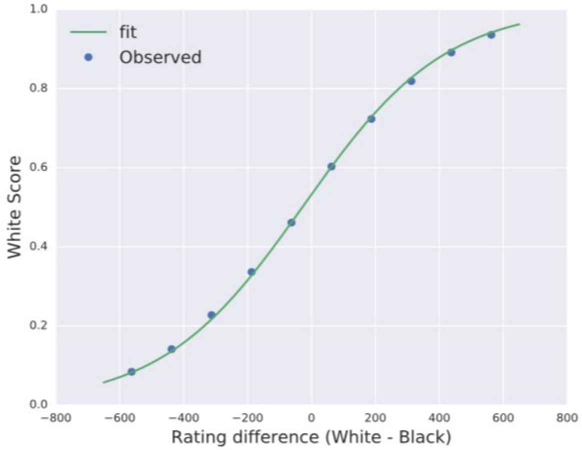


Figure 1. Difference in player rating against average game outcome (4,659,239 games from male-only competitors). 95% confidence intervals shown but not visible at this resolution.

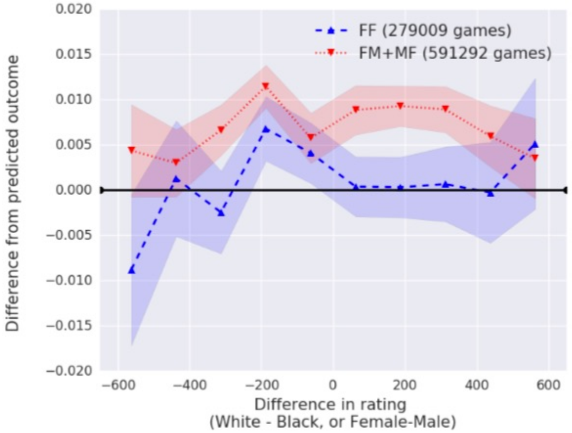


Figure 2. How player gender pairing affects game outcome (5,558,110 games total). Baseline expectation, from analysis of MM games, shown in black. Shaded regions show 95% confidence intervals.

The data comprise records of $9,662,202$ games of standard tournament chess, played between January 2008 and August 2015. There are also records of $461,637$ FIDE rated players ($56,474$, 12.2%, women. The average birth year for these players was 1983, with an average age of 31.5 years (standard deviation 19.28) at the time the games were played.

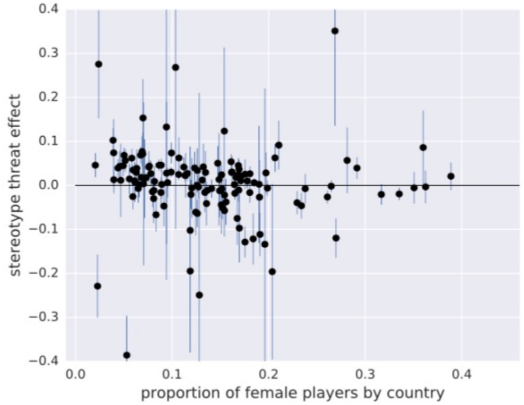


Figure 3. Stereotype threat effect, average by country. 95% confidence intervals shown.

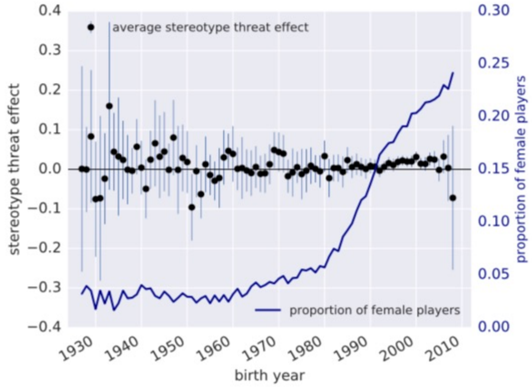


Figure 4. Stereotype threat effect, average by birth year (dots, left axis). 95% confidence intervals shown. Right axis shows proportion of female players in dataset for that birth year (continuous line).

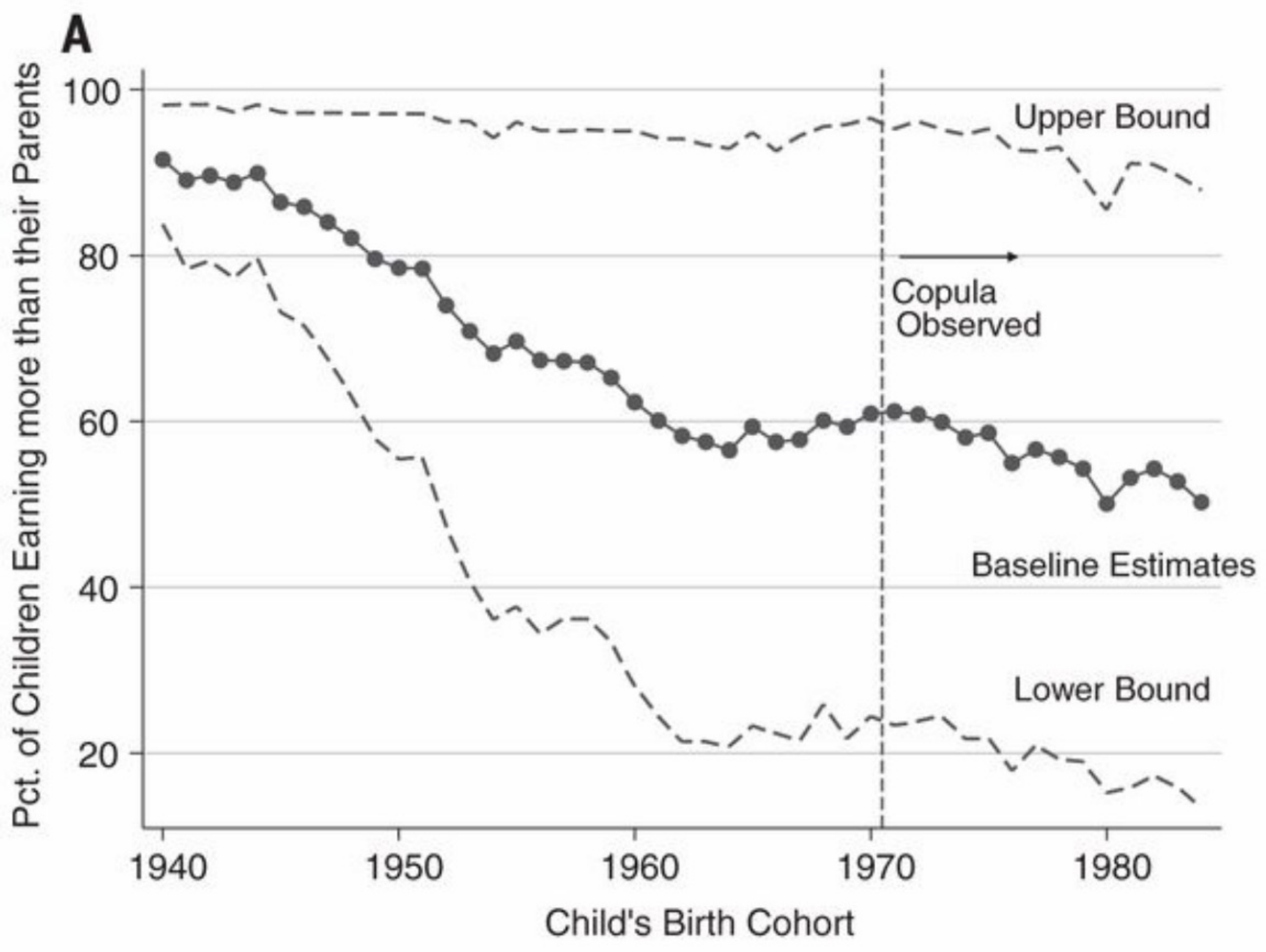
The fading American dream: Trends in absolute income mobility since 1940

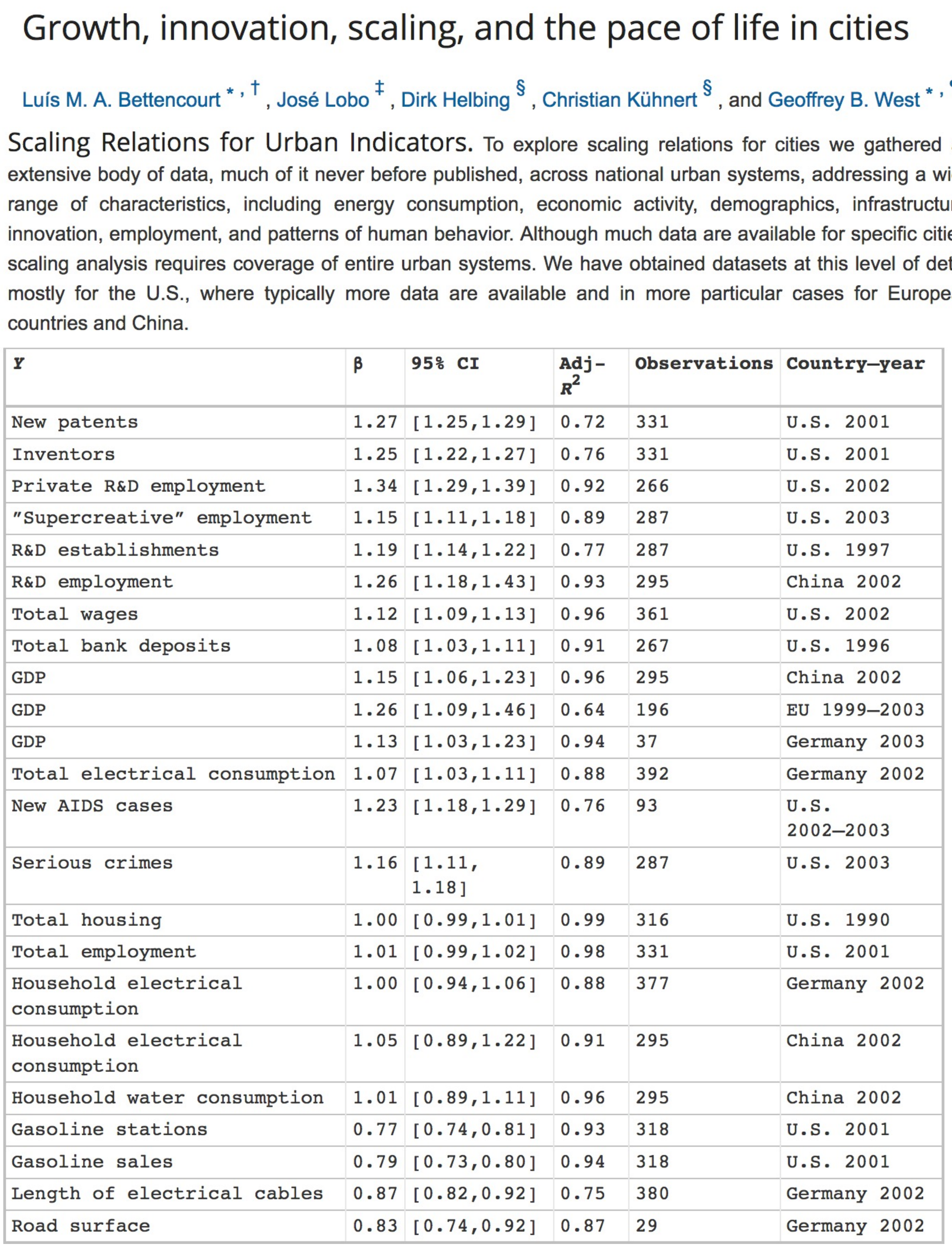
Raj Chetty[1,*], David Grusky[2,*], Maximilian Hell[2], Nathaniel Hendren[3,*], Robert Manduca[4], Jimmy Narang[5]

+ See all authors and affiliations

We used cross-sectional data from the decennial U.S. Census and Current Population Surveys (CPS) to estimate marginal income distributions for children in the 1940 to 1984 birth cohorts and their parents. The census data sets cover between 1% and 5% of the U.S. population, yielding samples of 20,000 to 35,000 families per cohort, whereas the CPS samples include approximately 1500 to 3000 people per cohort. In our baseline analysis, we measured income in pretax dollars at

We estimated the fraction of children who earn more than their parents in each birth cohort by combining the marginal income distributions with the copula in each cohort. For children born in or after 1980, we followed Chetty *et al.* ([12]) and directly estimated the joint distribution of parent and child ranks, using information from de-identified federal income tax returns covering more than 10 million parent-child pairs. For cohorts born before 1980, such population-level panel data are not





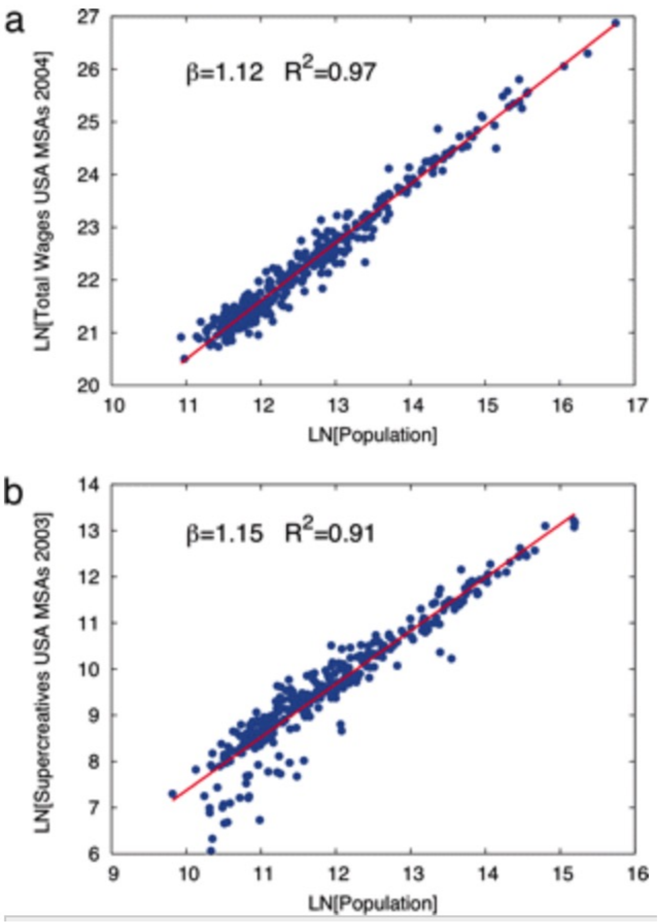
Growth, innovation, scaling, and the pace of life in cities

Luís M. A. Bettencourt [*, †], José Lobo [‡], Dirk Helbing [§], Christian Kühnert [§], and Geoffrey B. West [*, ¶]

Scaling Relations for Urban Indicators.

To explore scaling relations for cities we gathered an extensive body of data, much of it never before published, across national urban systems, addressing a wide range of characteristics, including energy consumption, economic activity, demographics, infrastructure, innovation, employment, and patterns of human behavior. Although much data are available for specific cities, scaling analysis requires coverage of entire urban systems. We have obtained datasets at this level of detail mostly for the U.S., where typically more data are available and in more particular cases for European countries and China.

Y	β	95% CI	Adj-R^2	Observations	Country–year
New patents	1.27	[1.25,1.29]	0.72	331	U.S. 2001
Inventors	1.25	[1.22,1.27]	0.76	331	U.S. 2001
Private R&D employment	1.34	[1.29,1.39]	0.92	266	U.S. 2002
”Supercreative” employment	1.15	[1.11,1.18]	0.89	287	U.S. 2003
R&D establishments	1.19	[1.14,1.22]	0.77	287	U.S. 1997
R&D employment	1.26	[1.18,1.43]	0.93	295	China 2002
Total wages	1.12	[1.09,1.13]	0.96	361	U.S. 2002
Total bank deposits	1.08	[1.03,1.11]	0.91	267	U.S. 1996
GDP	1.15	[1.06,1.23]	0.96	295	China 2002
GDP	1.26	[1.09,1.46]	0.64	196	EU 1999–2003
GDP	1.13	[1.03,1.23]	0.94	37	Germany 2003
Total electrical consumption	1.07	[1.03,1.11]	0.88	392	Germany 2002
New AIDS cases	1.23	[1.18,1.29]	0.76	93	U.S. 2002–2003
Serious crimes	1.16	[1.11, 1.18]	0.89	287	U.S. 2003
Total housing	1.00	[0.99,1.01]	0.99	316	U.S. 1990
Total employment	1.01	[0.99,1.02]	0.98	331	U.S. 2001
Household electrical consumption	1.00	[0.94,1.06]	0.88	377	Germany 2002
Household electrical consumption	1.05	[0.89,1.22]	0.91	295	China 2002
Household water consumption	1.01	[0.89,1.11]	0.96	295	China 2002
Gasoline stations	0.77	[0.74,0.81]	0.93	318	U.S. 2001
Gasoline sales	0.79	[0.73,0.80]	0.94	318	U.S. 2001
Length of electrical cables	0.87	[0.82,0.92]	0.75	380	Germany 2002
Road surface	0.83	[0.74,0.92]	0.87	29	Germany 2002

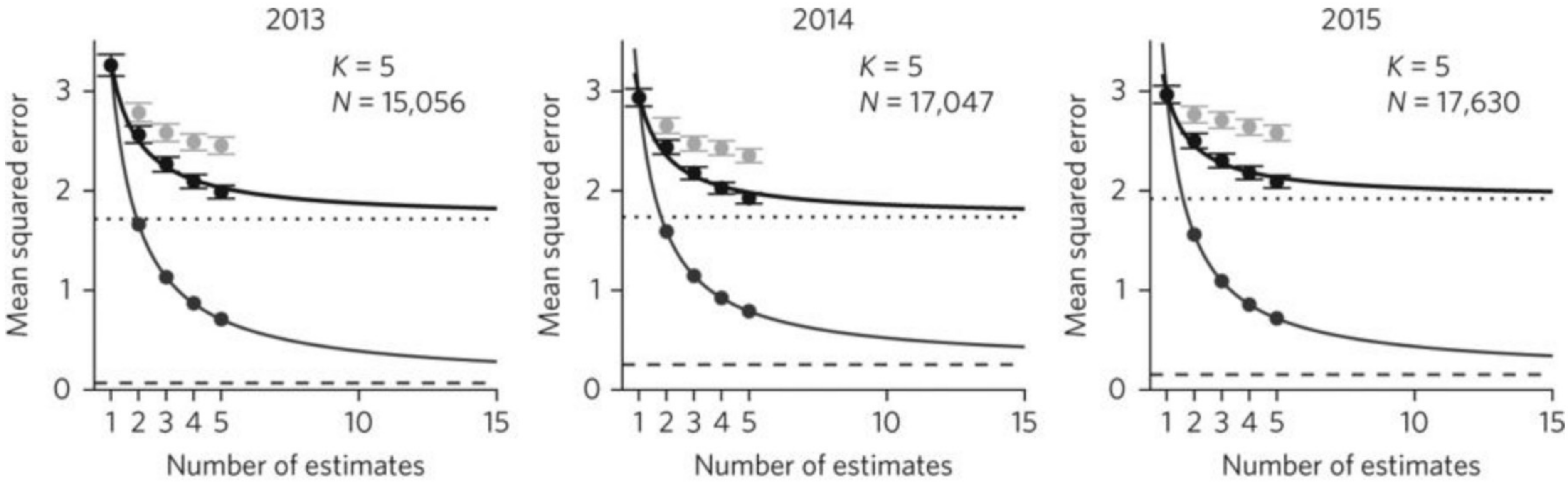


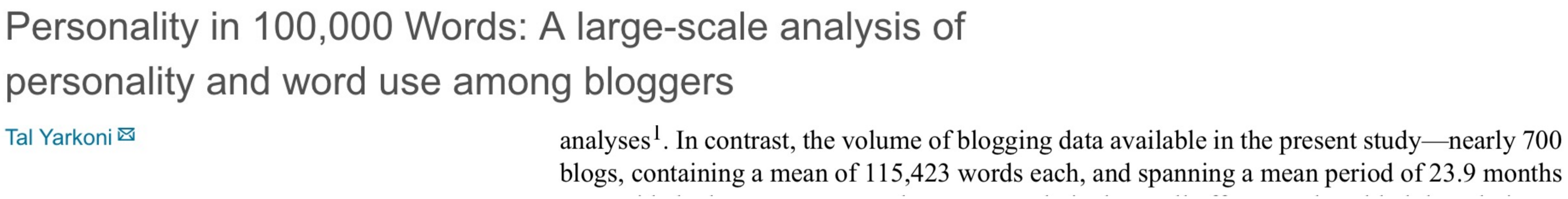
The wisdom of the inner crowd in three large natural experiments

Dennie van Dolder & Martijn J. van den Assem

Our data are from three promotional events organized by the Dutch state-owned casino chain Holland Casino. During the last 7 weeks of 2013, 2014 and 2015, anybody who visited one of the casinos received a voucher with a login code. Via a terminal inside the casino and via the Internet, this code granted access to a competition in which participants were asked to estimate the number of objects in a transparent plastic container located just inside the entrance. This

Our pseudonymized data sets contain all entries for the three years: a total of 369,260 estimates from 163,719 different players in 2013, 388,352 estimates from 154,790 players in 2014, and 407,622 estimates from 162,275 players in 2015. Many players submitted multiple estimates

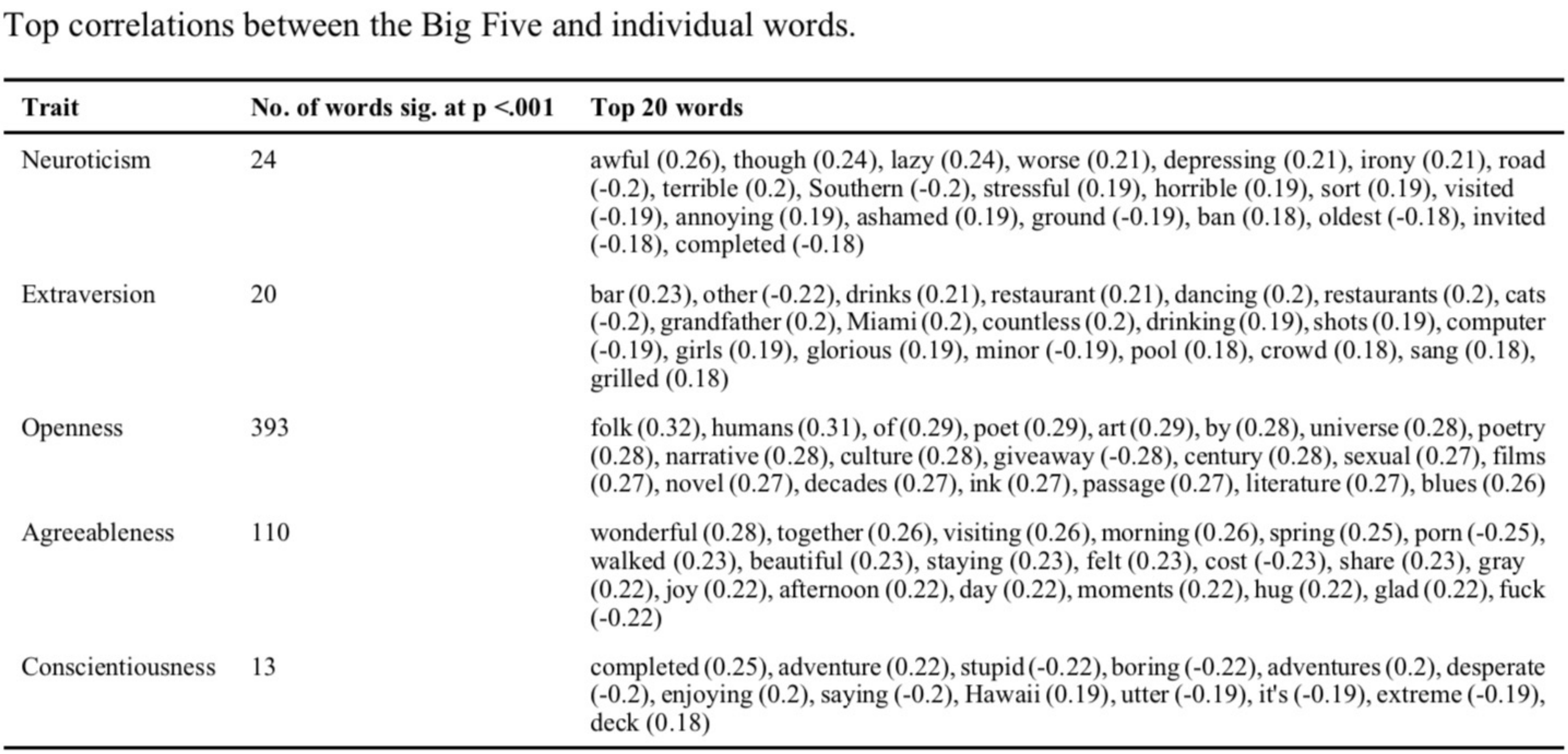




Personality in 100,000 Words: A large-scale analysis of personality and word use among bloggers

Tal Yarkoni ✉

analyses[1]. In contrast, the volume of blogging data available in the present study—nearly 700 blogs, containing a mean of 115,423 words each, and spanning a mean period of 23.9 months

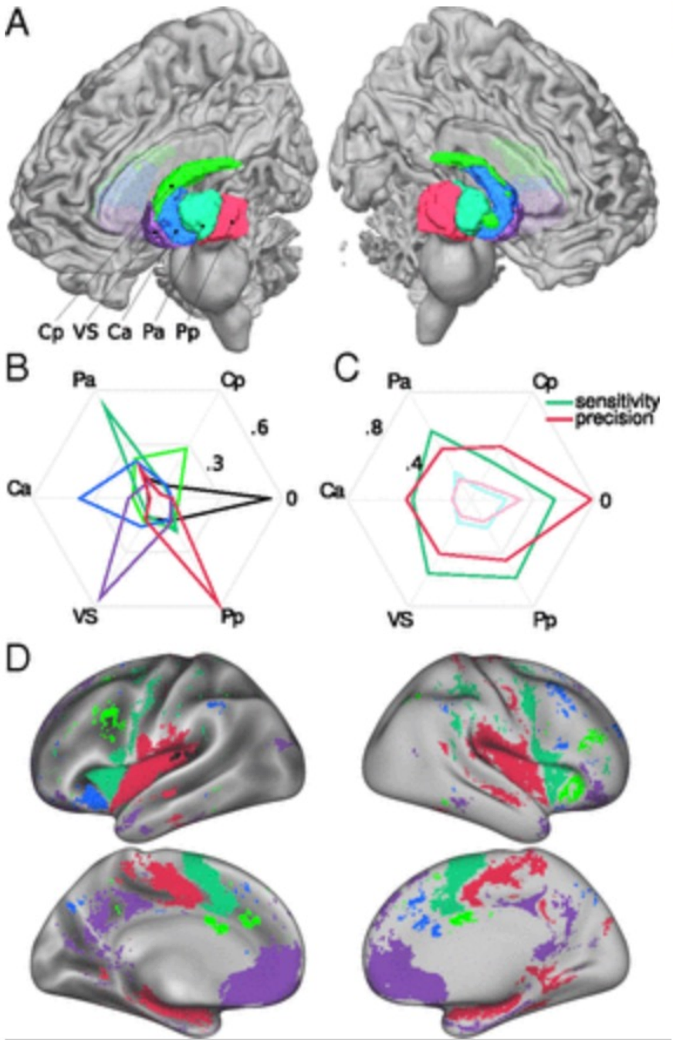


Top correlations between the Big Five and individual words.

Trait	No. of words sig. at p <.001	Top 20 words
Neuroticism	24	awful (0.26), though (0.24), lazy (0.24), worse (0.21), depressing (0.21), irony (0.21), road (-0.2), terrible (0.2), Southern (-0.2), stressful (0.19), horrible (0.19), sort (0.19), visited (-0.19), annoying (0.19), ashamed (0.19), ground (-0.19), ban (0.18), oldest (-0.18), invited (-0.18), completed (-0.18)
Extraversion	20	bar (0.23), other (-0.22), drinks (0.21), restaurant (0.21), dancing (0.2), restaurants (0.2), cats (-0.2), grandfather (0.2), Miami (0.2), countless (0.2), drinking (0.19), shots (0.19), computer (-0.19), girls (0.19), glorious (0.19), minor (-0.19), pool (0.18), crowd (0.18), sang (0.18), grilled (0.18)
Openness	393	folk (0.32), humans (0.31), of (0.29), poet (0.29), art (0.29), by (0.28), universe (0.28), poetry (0.28), narrative (0.28), culture (0.28), giveaway (-0.28), century (0.28), sexual (0.27), films (0.27), novel (0.27), decades (0.27), ink (0.27), passage (0.27), literature (0.27), blues (0.26)
Agreeableness	110	wonderful (0.28), together (0.26), visiting (0.26), morning (0.26), spring (0.25), porn (-0.25), walked (0.23), beautiful (0.23), staying (0.23), felt (0.23), cost (-0.23), share (0.23), gray (0.22), joy (0.22), afternoon (0.22), day (0.22), moments (0.22), hug (0.22), glad (0.22), fuck (-0.22)
Conscientiousness	13	completed (0.25), adventure (0.22), stupid (-0.22), boring (-0.22), adventures (0.2), desperate (-0.2), enjoying (0.2), saying (-0.2), Hawaii (0.19), utter (-0.19), it's (-0.19), extreme (-0.19), deck (0.18)

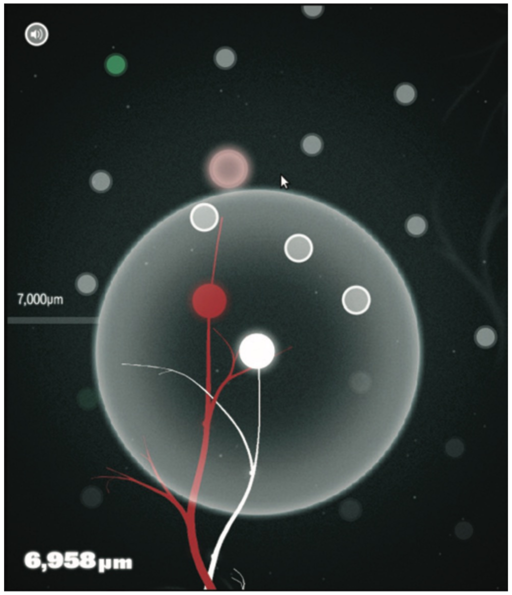
Regional specialization within the human striatum for diverse psychological functions

Wolfgang M. Pauli[a,1], Randall C. O'Reilly[b], Tal Yarkoni[c], and Tor D. Wager[b,d]



Metaanalytic Coactivation Analysis. We relied on the NeuroSynth database to gain a comprehensive and unbiased window into coactivation of the striatum with other regions. The NeuroSynth database contains activation coordinates for 5,809 functional MRI (fMRI) studies that were not selected for specific criteria, or with regard to the psychological processes under investigation, but only for the presence of reported brain activations; hence, it is highly representative of the broader neuroimaging field (26). In this database, 10-mm

Tracing the Trajectory of Skill Learning With a Very Large Sample of Online Game Players



Tom Stafford[1] and Michael Dewar[2]

[1]Department of Psychology, University of Sheffield, and [2]The New York Times Research & Development Lab, New York, New York

In the present study, we analyzed data from a very large sample (*N* = 854,064) of players of an online game involving rapid perception, decision making, and motor responding. Use of game data allowed us to connect, for the first time,

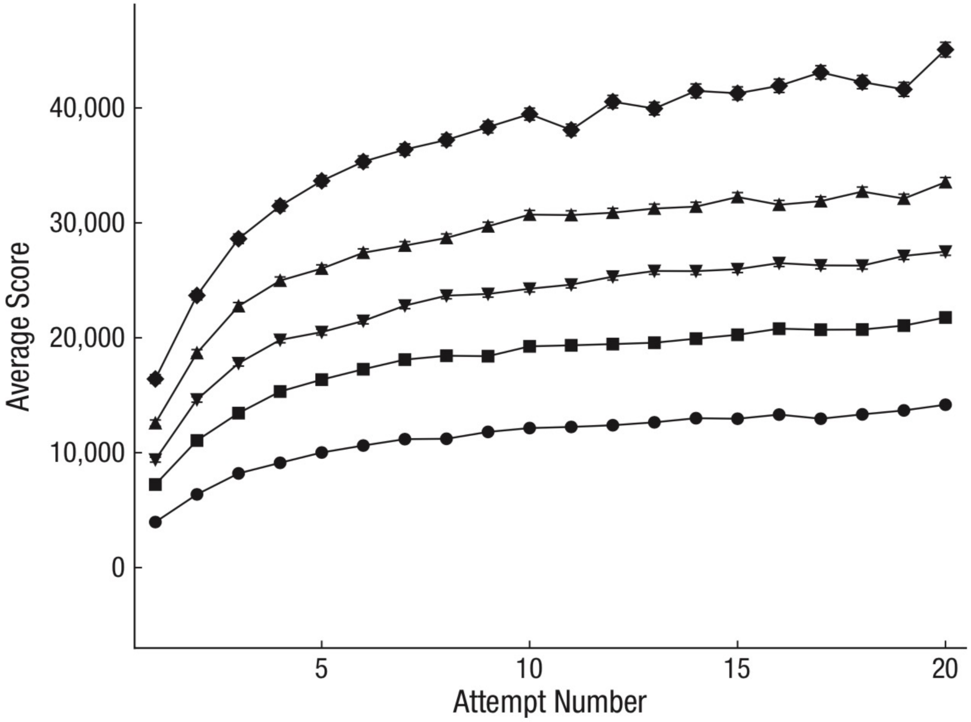


Fig. 2. Average score as a function of attempt number and percentile ranking based on players' highest score. Error bars show standard errors.

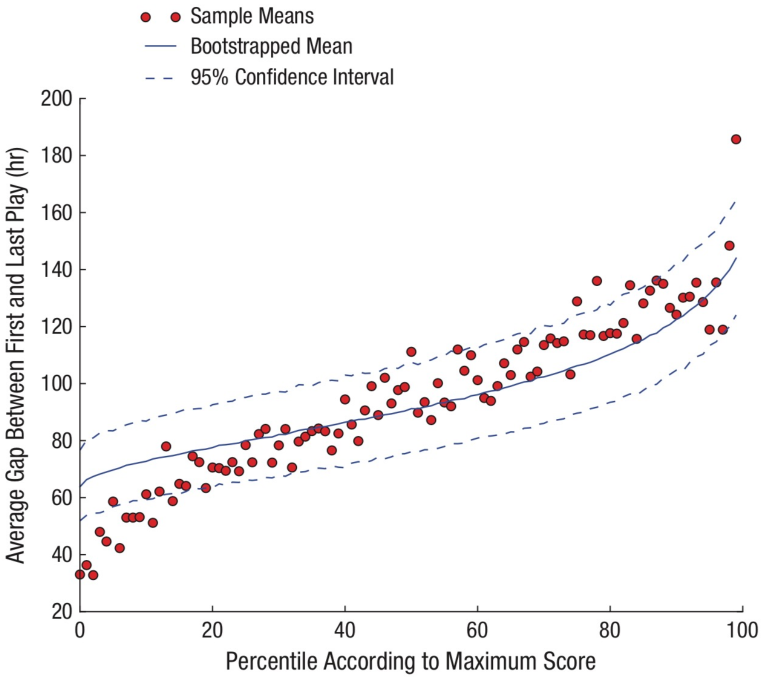


Fig. 3. Average time delay between players' first and last plays, for each percentile rank (based on players' maximum scores). Rankings are based on data from all players who played more than nine times. Bootstrapped means (with 95% confidence interval) are also shown.

Predicting scientific success

Daniel E. Acuna, **Stefano Allesina** and **Konrad P. Kording** present a formula to estimate the future *h*-index of life scientists.

METRICS

Predict your future h*-index*

These are approximate equations for predicting the *h*-index of neuroscientists in the future. They are probably reasonably precise for life scientists, but likely to be less meaningful for the other sciences. Try it for yourself online at go.nature.com/z4rroc.

- **Predicting next year** ($R^2=0.92$):

$$h_{+1}=0.76+0.37\sqrt{n}+0.97h-0.07y+0.02j+0.03q$$

- **Predicting 5 years into the future** ($R^2=0.67$):

$$h_{+5}=4+1.58\sqrt{n}+0.86h-0.35y+0.06j+0.2q$$

- **Predicting 10 years into the future** ($R^2=0.48$):

$$h_{+10}=8.73+1.33\sqrt{n}+0.48h-0.41y+0.52j+0.82q$$

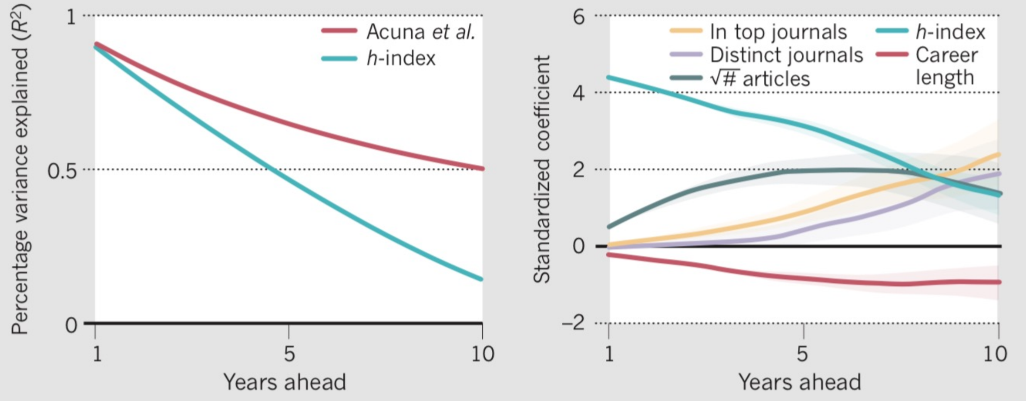
Key: *n*, number of articles written; *h*, current *h*-index; *y*, years since publishing first a *j*, number of distinct journals published in; *q*, number of articles in *Nature*, *Science*, *Neuroscience*, *Proceedings of the National Academy of Sciences* and *Neuron*.

To construct a formula to predict future *h*-index, we assembled a large data set and analysed it using machine-learning techniques. Our initial sample from academictree.org — a crowd-sourced website listing scientists' mentors, trainees and collaborators — contains the names and institutions of about 34,800 neuroscientists, 2,000 scientists studying the fruitfly *Drosophila* and 1,300 evolutionary researchers. We matched these authors to records in Scopus, an online database of academic papers and citation data. We restricted our analysis to authors who had accrued an *h*-index greater than 4 (to exclude inactive scientists); to publications after 1995 (because electronic records are sparse before then); to authors who had published their first manuscript in the past 5–12 years; and to authors who were identifiable in Scopus.

That left us with 3,085 neuroscientists, 57 *Drosophila* researchers and 151 evolutionary scientists for whom we constructed a history of publication, citation and funding.

PATHS TO SUCCESS

The accuracy of future *h*-index prediction decreases over time, but the Acuna *et al*. formula predicts future *h*-index better than does current *h*-index alone (left). The contribution of each factor to the formula accuracy also changes over time (right). Shading indicates 95% confidence error bars.

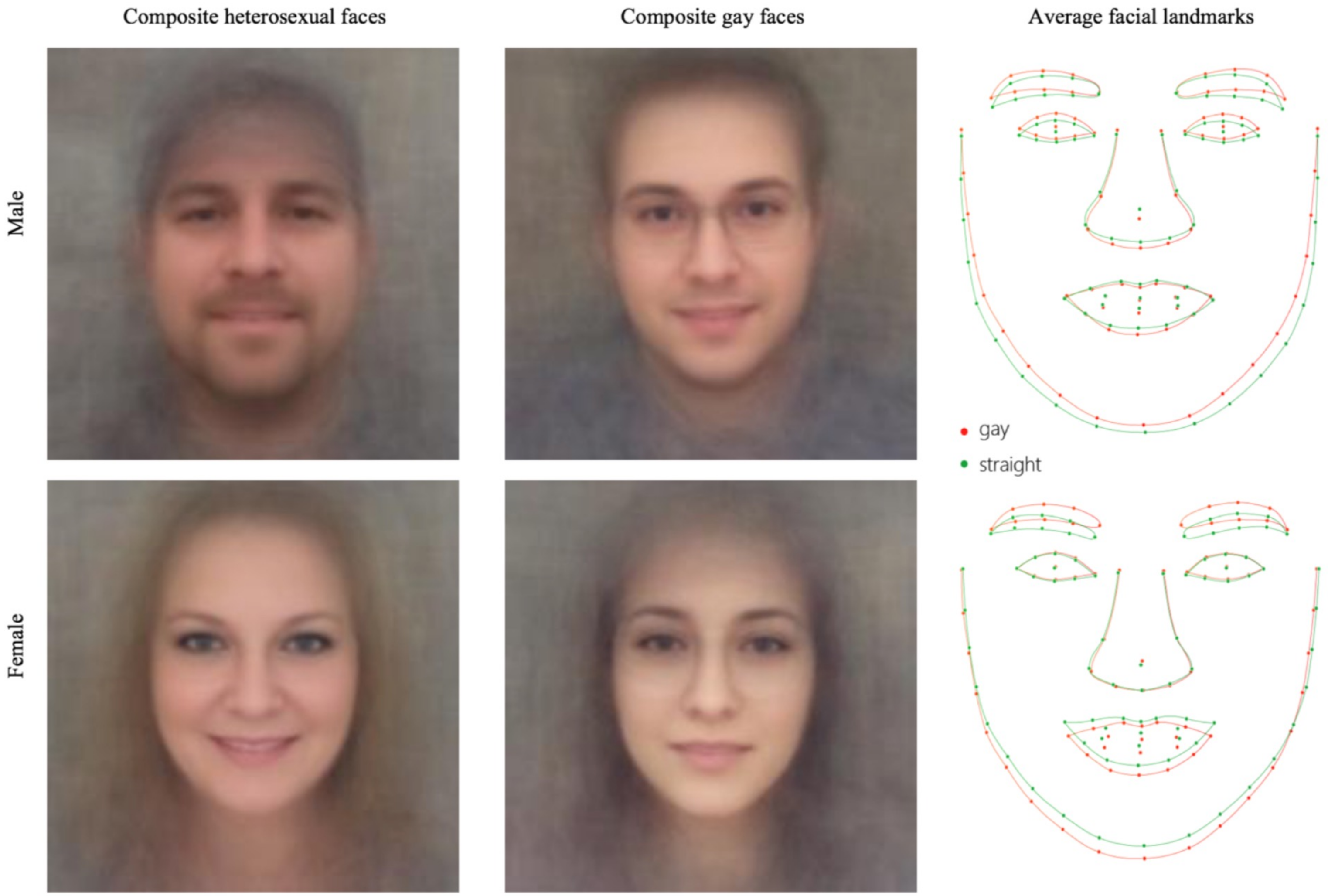


DEEP NEURAL NETWORKS CAN DETECT SEXUAL ORIENTATION FROM FACES

Deep neural networks are more accurate than humans at detecting sexual orientation from facial images

Yilun Wang, Michal Kosinski

perceived and interpreted by the human brain. We used deep neural networks to extract features from 35,326 facial images. These features were entered into a logistic regression aimed at classifying sexual orientation. Given a single facial image, a classifier could correctly distinguish between gay and heterosexual men in 81% of cases, and in 71% of cases for women. Human


Composite heterosexual faces

Composite gay faces

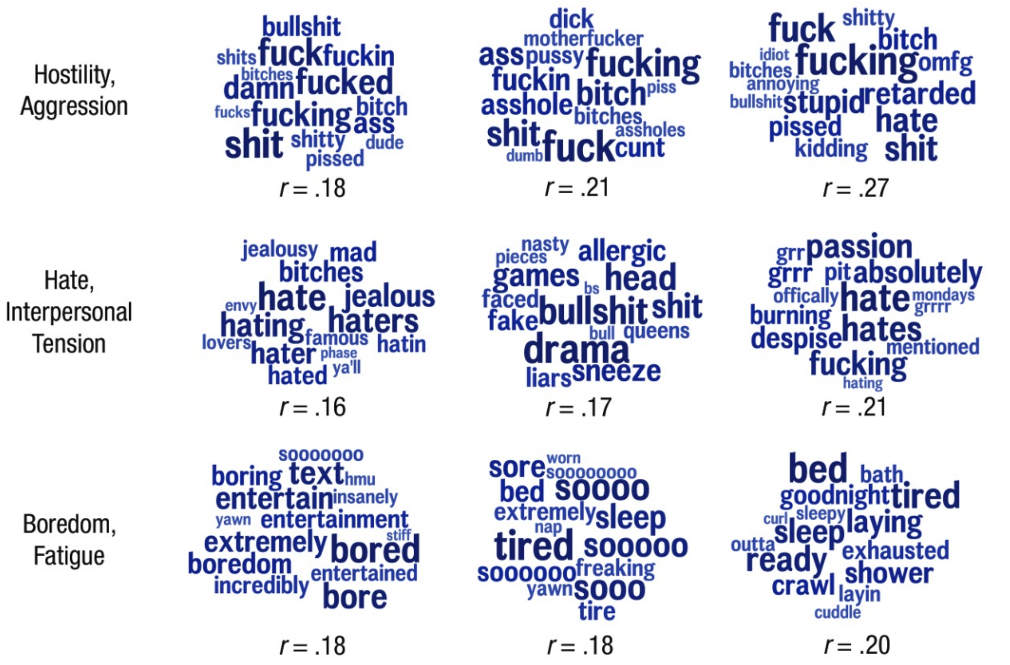
Average facial landmarks

Psychological Language on Twitter Predicts County-Level Heart Disease Mortality

Johannes C. Eichstaedt[1], Hansen Andrew Schwartz[1,2], Margaret L. Kern[1,3], Gregory Park[1], Darwin R. Labarthe[4], Raina M. Merchant[5], Sneha Jha[2], Megha Agrawal[2], Lukasz A. Dziurzynski[1], Maarten Sap[1], Christopher Weeg[1], Emily E. Larson[1], Lyle H. Ungar[1,2], and Martin E. P. Seligman[1]

[1]Department of Psychology, University of Pennsylvania; [2]Department of Computer and Information Science, University of Pennsylvania; [3]Graduate School of Education, University of Melbourne; [4]School of Medicine, Northwestern University; and [5]Department of Emergency Medicine, University of Pennsylvania

Twitter Topics Positively Correlated With County-Level AHD Mortality

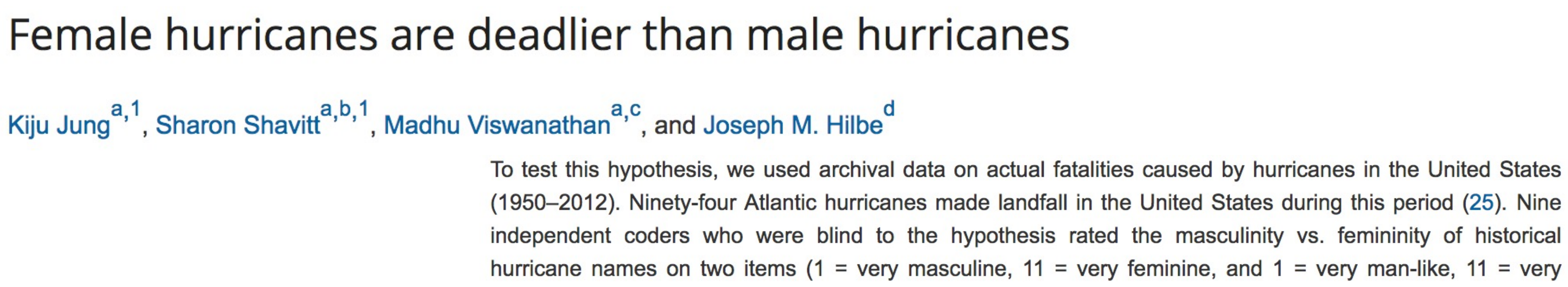


Data sources

We used data from 1,347 U.S. counties for which AHD mortality rates; county-level socioeconomic, demographic, and health variables; and at least 50,000 tweeted words were available. More than 88% of the U.S. population lives in the included counties (U.S. Census Bureau, 2010).[1]

Twitter data. Tweets are brief messages (no more than 140 characters) containing information about emotions, thoughts, behaviors, and other personally salient information. In 2009 and 2010, Twitter made a 10% random sample of tweets (the "Garden Hose") available for researchers through direct access to its servers. We obtained a sample of 826 million tweets collected between June 2009 and March 2010. Many Twitter users self-reported their locations in their user profiles, and we used this information to map tweets to counties (for details, see the Mapping Tweets to Counties section of the Supplemental Method in the Supplemental Material available online). This resulted in 148 million county-mapped tweets across 1,347 counties.

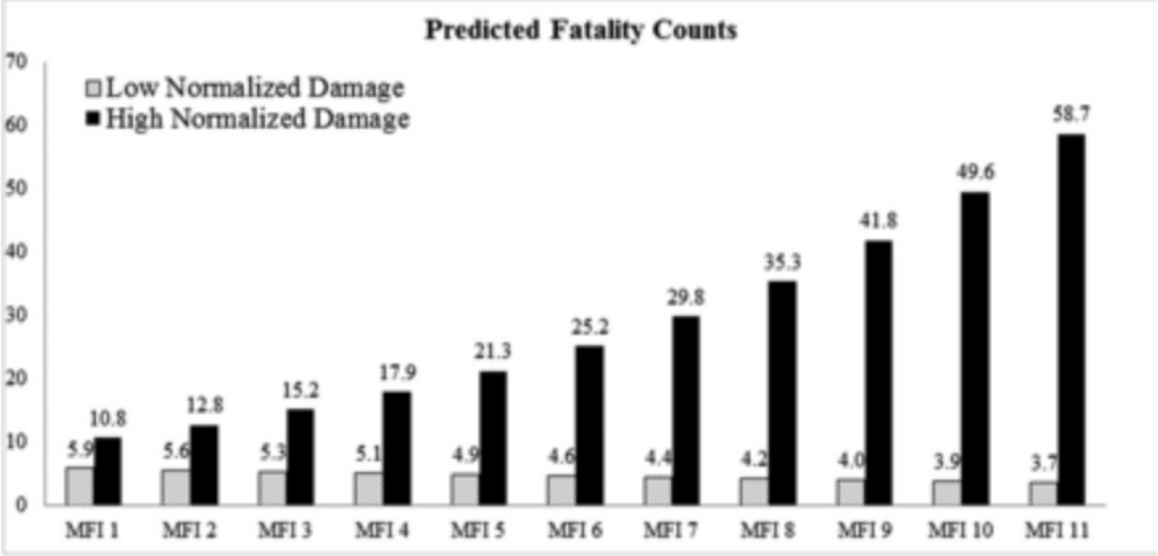
Heart disease data. Counties are the smallest socio-ecological level for which most CDC health variables and U.S. Census information are available. From the Centers for Disease Control and Prevention (2010b) we obtained county-level age-adjusted mortality rates for AHD, which is represented by code I25.1 in the International Classification of Disease, 10th edition (ICD 10; World Health Organization, 1992). This code has the highest overall mortality rate in the United States (prevalence = 51.5 deaths per 100,000 in 2010). We averaged AHD mortality rates across 2009 and 2010 to match the time period of the Twitter-language data set.



Female hurricanes are deadlier than male hurricanes

Kiju Jung[a,1], Sharon Shavitt[a,b,1], Madhu Viswanathan[a,c], and Joseph M. Hilbe[d]

To test this hypothesis, we used archival data on actual fatalities caused by hurricanes in the United States (1950–2012). Ninety-four Atlantic hurricanes made landfall in the United States during this period (25). Nine independent coders who were blind to the hypothesis rated the masculinity vs. femininity of historical hurricane names on two items (1 = very masculine, 11 = very feminine, and 1 = very man-like, 11 = very



…I *really* don't endorse this paper, but it is yet another example of a question + natural data suitable for this project.

May yield publishable product.

Formalizing Interdisciplinary Collaboration in the CogSci Community

Lauren Oey* (loey@ucsd.edu)
Isabella DeStefano* (idestefa@ucsd.edu)
Erik Brockbank (ebrockba@ucsd.edu)
Edward Vul (evul@ucsd.edu)
University of California, San Diego, Department of Psychology
9500 Gilman Dr., La Jolla, CA 92093 USA

Abstract

Is cognitive science interdisciplinary or multidisciplinary? We contribute to this debate by examining the authorship structure and topic similarity of contributions to the Cognitive Science Society from 2000 to 2019. We compare findings from CogSci to abstracts from the Vision Science Society over the same time frame. Our analysis focuses on graph theoretic features of the co-authorship network—edge density, transitivity, and maximum subgraph size—as well as clustering within the topic space of CogSci contributions. We also combine structural and semantic information with an analysis of homophily. We validate this approach by predicting new collaborations in this year's CogSci proceedings. Our results suggest that cognitive science has become increasingly interdisciplinary in the last 19 years. More broadly, we argue that a formal quantitative approach which combines structural co-authorship information and semantic topic analysis provides inroads to questions about the level of interdisciplinary collaboration in the cognitive science community.

Keywords: co-authorship networks; topic modeling; interdisciplinarity; multidisciplinarity; scientometrics

Introduction

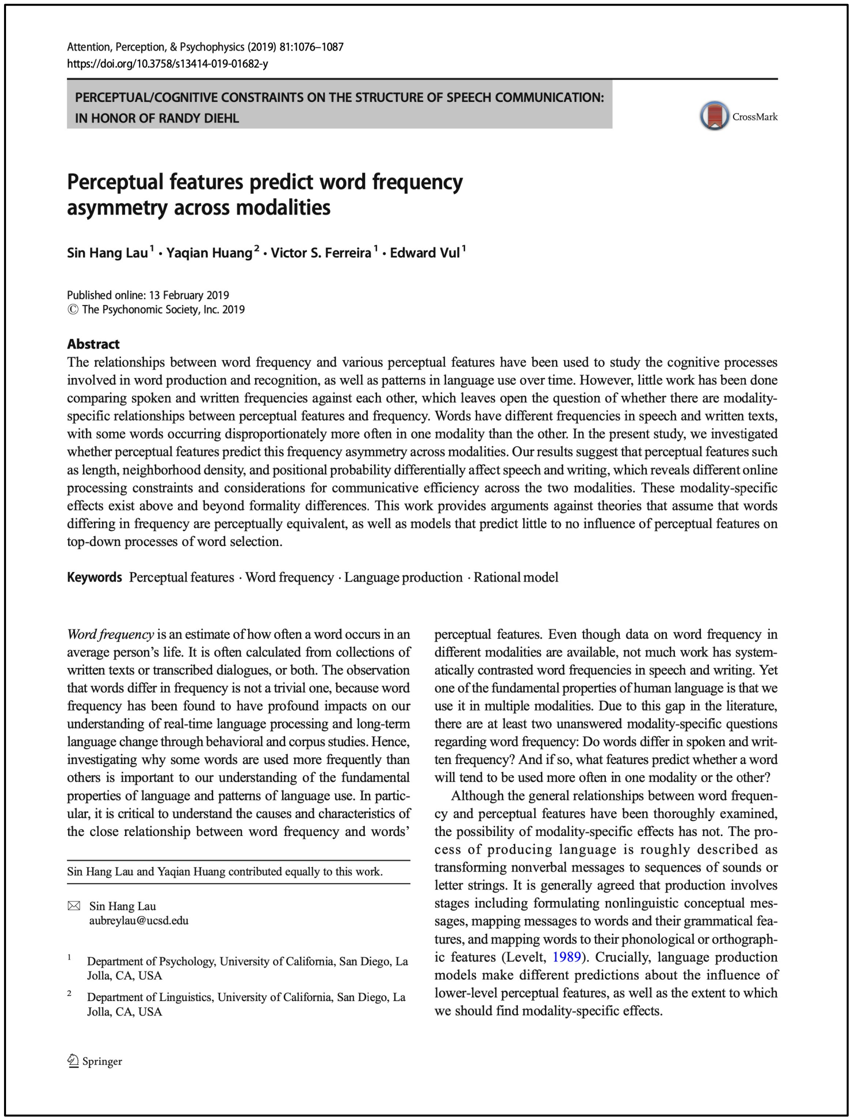
Since its foundation, the Cognitive Science Society sought to unify various disciplines of study under one interdisciplinary research field. Recently, criticism of the success of this mission has sparked a debate about whether cognitive science is fundamentally multidisciplinary rather than interdisciplinary (Núñez et al., 2019; Gray, 2019). The distinction between these community structures is subtle, making any claims favoring one or the other difficult to evaluate. Broadly, a research community might be considered to be more *multidisciplinary* if collaborations happen mostly within small groups and there is greater topical isolation of each group from the rest. On the other hand, a more *interdisciplinary* research community will show fewer isolated groups structurally and less separation of research interests across groups.

But how do we measure interdisciplinarity in a way that captures meaningful differences within diverse communities? Currently, there is no consensus on a single measure that best aligns with this abstract concept. Previous studies quantified interdisciplinarity by considering the journals as tags for different disciplines. Some of these studies have examined the distribution of journals cited (Goldstone & Leydesdorff, 2006; Porter, Cohen, Roessner, & Perreault, 2007; Núñez et al., 2019), the citation networks (Rafols & Meyer, 2010), and the journals that authors previously published in (Bergmann, Dale, Sattari, Heit, & Bhat, 2017). But this earlier research aiming to quantify interdisciplinarity was primarily targeted at the categorization of disciplines. These measures suffer from inconsistencies across classification systems, leading to variable conclusions (Wagner et al., 2011). Others have used departmental affiliation and educational background (Núñez et al., 2019; Schunn, Crowley, & Okada, 1998), but research interests often shift over the course of a lifetime which makes the affiliation label a transient indicator (Porter et al., 2007).

In the present work, we address the challenges of defining and measuring interdisciplinarity through a combination of co-authorship network features, topic analysis, and assessment of graph homophily that unifies both structure and content of publications. We validate our measures using full papers from the Cognitive Science Society proceedings between 2000 and 2019 and abstracts from the Vision Science Society (only abstracts are submitted) over a similar time frame (2001 to 2019).

First, the degree to which a community is interdisciplinary or multidisciplinary may in large part be revealed by who collaborates with whom. Scientific collaboration can be represented as an undirected graph, in which nodes correspond to individual authors and edges between nodes indicate whether any two authors co-authored a paper together (Newman, 2001, 2004; Barabási et al., 2002). Co-authorships within a community containing multiple areas of study can range from highly integrated to highly modular, and the structure of the resulting co-authorship network will reflect this spectrum of possibilities.

Second, while the collaboration structure of a community no doubt reveals something about the modularity of interdisciplinary work that occurs within it, the ways in which research interests combine must play a role as well. To better understand how the *content* of collaborations informs the interdisciplinarity of the field, we use a topic model (Griffiths & Steyvers, 2004) to extract high level patterns in cognitive science research over the last 19 years. Topic models have been used in previous research to capture trends in the published work within a discipline, including within cognitive science (Cohen Priva & Austerweil, 2015; Rothe, Rich, & Zhi-Wei, 2018). Studies specifically addressing interdisciplinarity have used topic models to complement pre-defined discipline tagging (Nichols, 2014). In the present work, we apply clustering algorithms to the topics that authors study, addressing the separability of the interests and methods of researchers in



Attention, Perception, & Psychophysics (2019) 81:1076–1087
https://doi.org/10.3758/s13414-019-01682-y

PERCEPTUAL/COGNITIVE CONSTRAINTS ON THE STRUCTURE OF SPEECH COMMUNICATION: IN HONOR OF RANDY DIEHL

Perceptual features predict word frequency asymmetry across modalities

Sin Hang Lau[1] · Yaqian Huang[2] · Victor S. Ferreira[1] · Edward Vul[1]

Published online: 13 February 2019

Abstract

The relationships between word frequency and various perceptual features have been used to study the cognitive processes involved in word production and recognition, as well as patterns in language use over time. However, little work has been done comparing spoken and written frequencies against each other, which leaves open the question of whether there are modality-specific relationships between perceptual features and frequency. Words have different frequencies in speech and written texts, with some words occurring disproportionately more often in one modality than the other. In the present study, we investigated whether perceptual features predict this frequency asymmetry across modalities. Our results suggest that perceptual features such as length, neighborhood density, and positional probability differentially affect speech and writing, which reveals different online processing constraints and considerations for communicative efficiency across the two modalities. These modality-specific effects exist above and beyond formality differences. This work provides arguments against theories that assume that words differing in frequency are perceptually equivalent, as well as models that predict little to no influence of perceptual features on top-down processes of word selection.

Keywords Perceptual features · Word frequency · Language production · Rational model

Word frequency is an estimate of how often a word occurs in an average person's life. It is often calculated from collections of written texts or transcribed dialogues, or both. The observation that words differ in frequency is not a trivial one, because word frequency has been found to have profound impacts on our understanding of real-time language processing and long-term language change through behavioral and corpus studies. Hence, investigating why some words are used more frequently than others is important to our understanding of the fundamental properties of language and patterns of language use. In particular, it is critical to understand the causes and characteristics of the close relationship between word frequency and words' perceptual features. Even though data on word frequency in different modalities are available, not much work has systematically contrasted word frequencies in speech and writing. Yet one of the fundamental properties of human language is that we use it in multiple modalities. Due to this gap in the literature, there are at least two unanswered modality-specific questions regarding word frequency: Do words differ in spoken and written frequency? And if so, what features predict whether a word will tend to be used more often in one modality or the other?

Although the general relationships between word frequency and perceptual features have been thoroughly examined, the possibility of modality-specific effects has not. The process of producing language is roughly described as transforming nonverbal messages to sequences of sounds or letter strings. It is generally agreed that production involves stages including formulating nonlinguistic conceptual messages, mapping messages to words and their grammatical features, and mapping words to their phonological or orthographic features (Levelt, 1989). Crucially, language production models make different predictions about the influence of lower-level perceptual features, as well as the extent to which we should find modality-specific effects.

Sin Hang Lau and Yaqian Huang contributed equally to this work.

✉ Sin Hang Lau
aubreylau@ucsd.edu

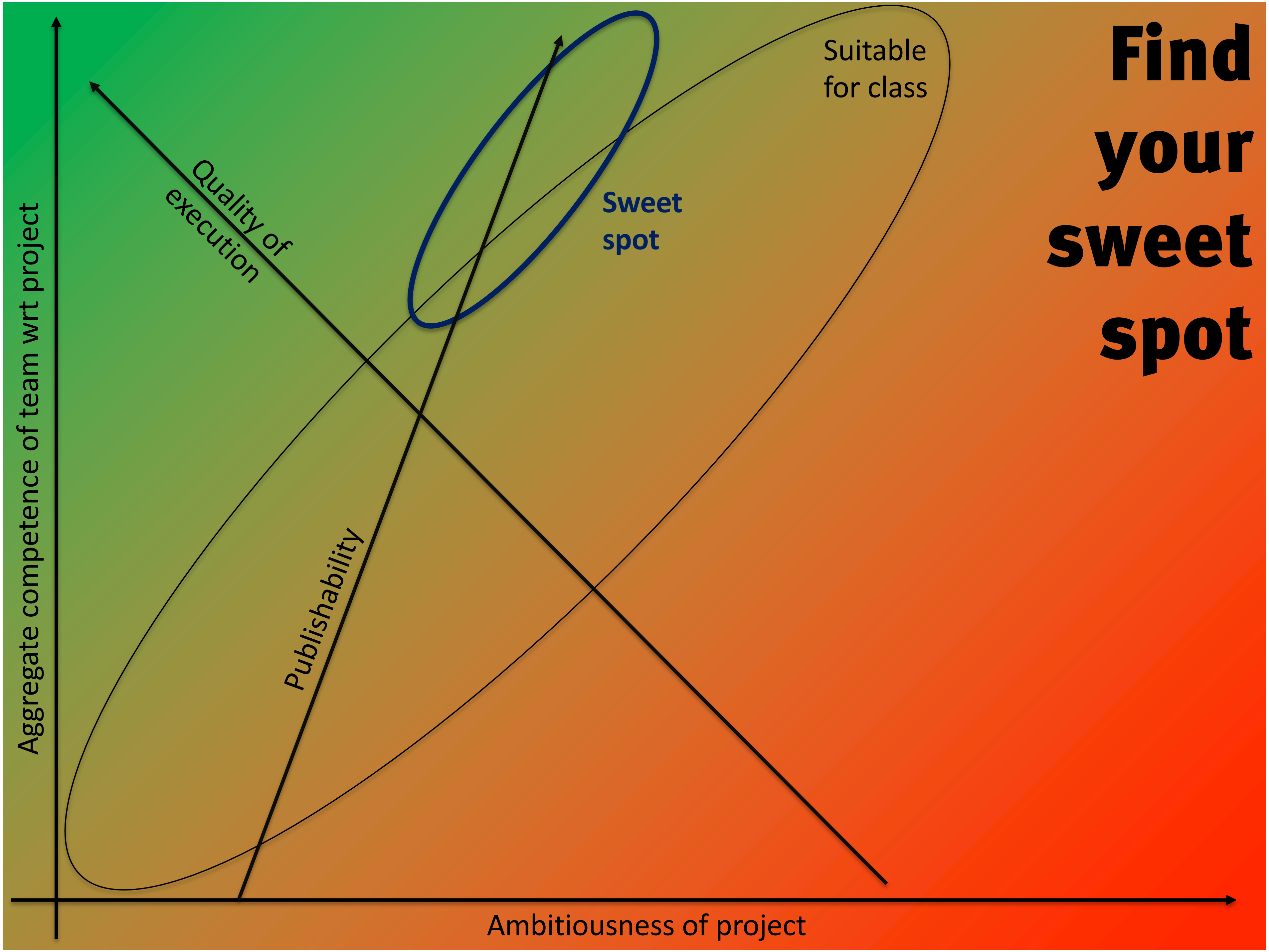
[1] Department of Psychology, University of California, San Diego, La Jolla, CA, USA

[2] Department of Linguistics, University of California, San Diego, La Jolla, CA, USA

Sources of data.

- User behavior from web services (netflix, yelp, okcupid, twitter, blogger, etc.)
- Government statistical agencies (Census, BLS, CDC, FBI, FAA, FDIC, BBB, SS records etc)
- Climate, earthquake, fire, etc. data.
- Language/text corpora
- Explicitly gathered datasets for this sort of thing: GSS, framingham, dolphin co-occurrence, etc.
- Field specific databases: sports statistics/events, citation statistics, imdb for movies, university ratings, etc.
- Meta-analytic datasets: neurosynth, wordbank, etc.
- Dataset Databases: dataverse, datahub, dataportals, icpsr, ssds, Kaggle, etc.
- Scraping something somewhere.

Find your sweet spot



Class project deliverables.

- 10/11: Groups due.
- 10/18: Project plan due
- 11/08: Preliminary data summaries due
- 12/09 (before final time): write-ups due
- 12/09 (during final time): project presentations
- 12/10: Group evaluations due.

Next: R live