PENNSTATE

Big Data Social Science
An Integrative Education and Research Program in Social Data Analytics

Approaches & Issues in Big Social Data

SoDA 501

"Ethics and scientific responsibility"

New ethics lessons hot off the presses

1. Don’t engage in sexual harassment.

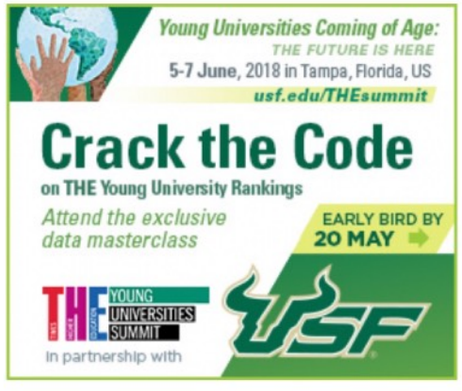
2. If you’re the editor of a major journal accused of quid pro quo sexual harassment, don’t use the journal’s website to call your accusers liars.

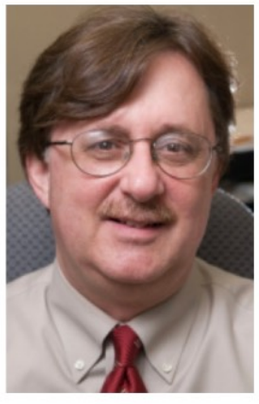
Editorial Malpractice?

The editor of a prestigious political science journal uses its website to deny harassment allegations against him, prompting the Midwest Political Science Association to take action critics say it should have taken months ago.

By Colleen Flaherty // April 19, 2018

0 COMMENTS





William Jacoby

The *American Journal of Political Science* is of one of the field’s most esteemed publications. So visitors to the journal’s main webpage were everything from incredulous to irate about what they saw there earlier this week: instead of just political science news, editor William G. Jacoby had posted a message denying the sexual harassment allegations he’s facing.

“It is apparently widely known that allegations related to sexual harassment have been made against me,” began the editorial note from Jacoby, a professor of political science at Michigan State University. “The allegations are untrue. I never engaged in the behaviors described in the allegations.”

Jacoby also used the highly visible space to announce that he’d be stepping down as editor of the journal at the end of December, of his own accord but due to “circumstances.”

In so doing, he continued to refute the allegations. While he is cooperating with several ongoing investigations into his conduct, he said, the charges are not going away, “despite their false nature.” Therefore, Jacoby wrote, “I do not want any questions about me as an individual (rather than as a scholar or editor) -- unfounded as these questions are -- to have any detrimental impact on the incredible, great things that have been accomplished at the journal so far.”

Some cases to consider …

TECH 2/16/2012 @ 11:02 上午 | 2,526,432 views

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

+ Comment Now + Follow Comments

Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy. Target, for example, has figured out how to data-mine its way into your womb, to figure out whether you have a baby on the way long

Target has got you in its aim

Source: Forbes


FACEBOOK SHOULDN'T CHOOSE WHAT STUFF THEY SHOW US TO CONDUCT UNETHICAL PSYCHOLOGICAL RESEARCH.
THEY SHOULD ONLY MAKE THOSE DECISIONS BASED ON, UH...
HOWEVER THEY WERE DOING IT BEFORE.
WHICH WAS PROBABLY ETHICAL, RIGHT?

Source: xkcd

The data is public.

Methods

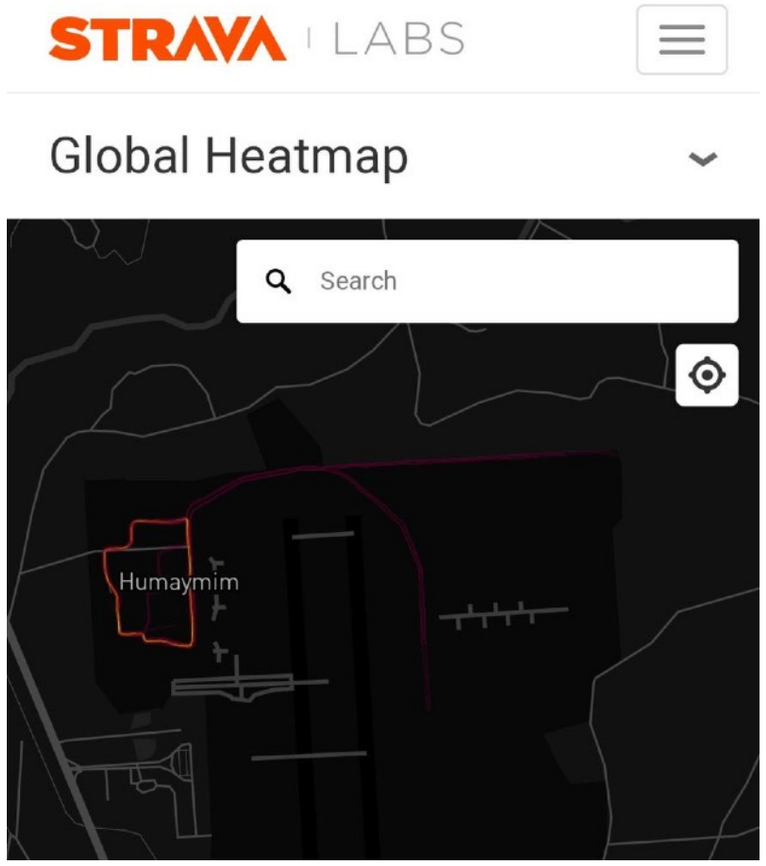
The complete details of the methodology are reviewed in the *SI Text* but are summarized briefly here. The research began by acquiring a list of registered voters in the state of New York that was accurate as of summer 2001. [Because all of the identifying information in this study stem from public records, the study was deemed exempt from Institutional Review Board review (Yale University Human Subjects Committee Protocol no. 1305011974).] The database contained personal information of all 9,995,513 New Yorkers registered at that time. Culling information from digital obituaries of 9/11 victims, I matched all 9/11 victims residing in New York to the statewide voter file (for reasons of cost and data availability, this study was restricted to 9/11 victims residing in New York at the time of the attack). Of the 1,729 victims from New York, I identified 1,181 (68%) as registered voters. This number matches the percentage of New York citizens overall who were registered to vote at the time.

For each victim, I used a combination of exact matching and Mahalanobis distance matching to identify up to five "control victims." The variables used in the matching algorithm include demographics, prior political activities, and family and neighborhood characteristics. To be clear, the control victims, who were mostly residents of metro New York and similarly situated as the true victims, were themselves obviously affected by the events of 9/11. This only biases the study toward a null result and isolates the effect under investigation as being particular to families and neighbors of victims. The causal effect estimated here is not the general effect of the 9/11 attack, but rather the specific effect of the attacks on families and neighbors.

Hersh - 2013 *PNAS*

OKCupid data release

- Danish student researcher publicly released a dataset of nearly 70,000 users of the online dating site OkCupid, including usernames, age, gender, location, what kind of relationship (or sex) they're interested in, personality traits, and answers to thousands of profiling questions used by the site



posted by Tobias Schneider (Twitter)

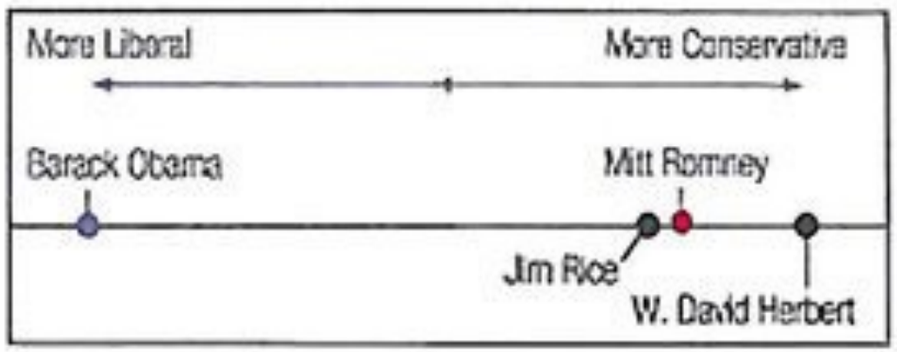
The IRB said it was fine.

VOTERS GUIDE

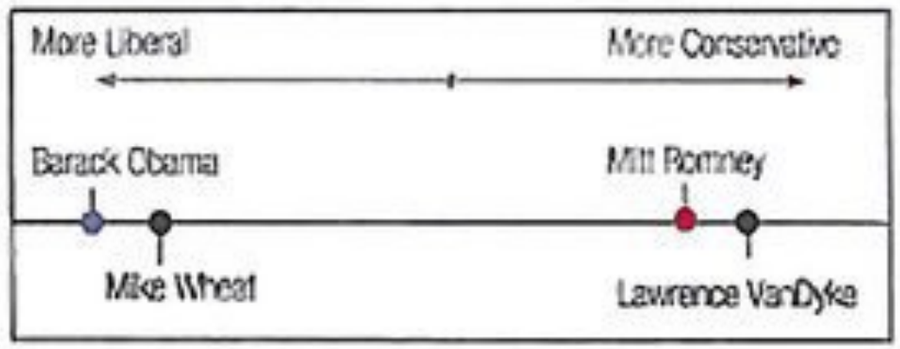
2014 Montana General Election Voter Information Guide

Election Date: November 4, 2014

Nonpartisan Supreme Court Justice #1 Race



Nonpartisan Supreme Court Justice #2 Race

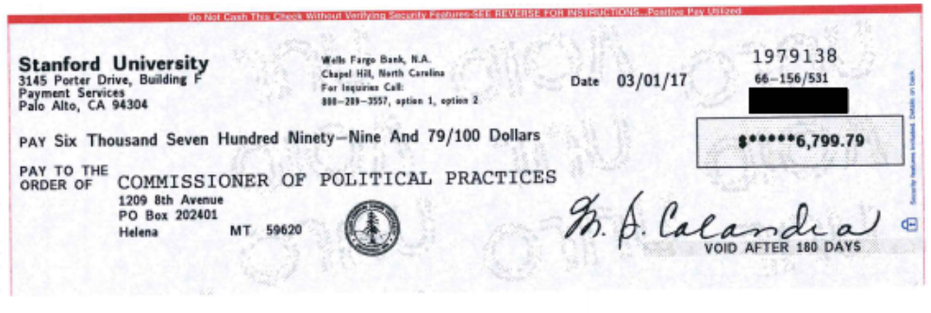


For more information on how these figures were created, please see http://data.stanford.edu/time. Please note that this guide is non-partisan and does not endorse any candidate or party. This guide was created as part of a joint research project at Stanford and Dartmouth.

Paid for by researchers at Stanford University and Dartmouth College, 616 Serra Street, Stanford, CA 94305

Take this to the polls!

Source: Bonica, Rodden, Dropp. 2014 Montana judicial elections experiment.


Do Not Cash This Check Without Verification Security Features-SEE REVERSE FOR INSTRUCTIONS - Positive Pay Utilized

Stanford University
3145 Porter Drive, Building F
Payment Services
Palo Alto, CA 94304

Wells Fargo Bank, N.A.
Chapel Hill, North Carolina
For Inquiries Call:
800-289-3557, option 1, option 2

1979138
Date 03/01/17
66-156/531
[redacted]

PAY Six Thousand Seven Hundred Ninety-Nine And 79/100 Dollars

$*****6,799.79

PAY TO THE ORDER OF COMMISSIONER OF POLITICAL PRACTICES
1209 8th Avenue
PO Box 202401
Helena MT 59620

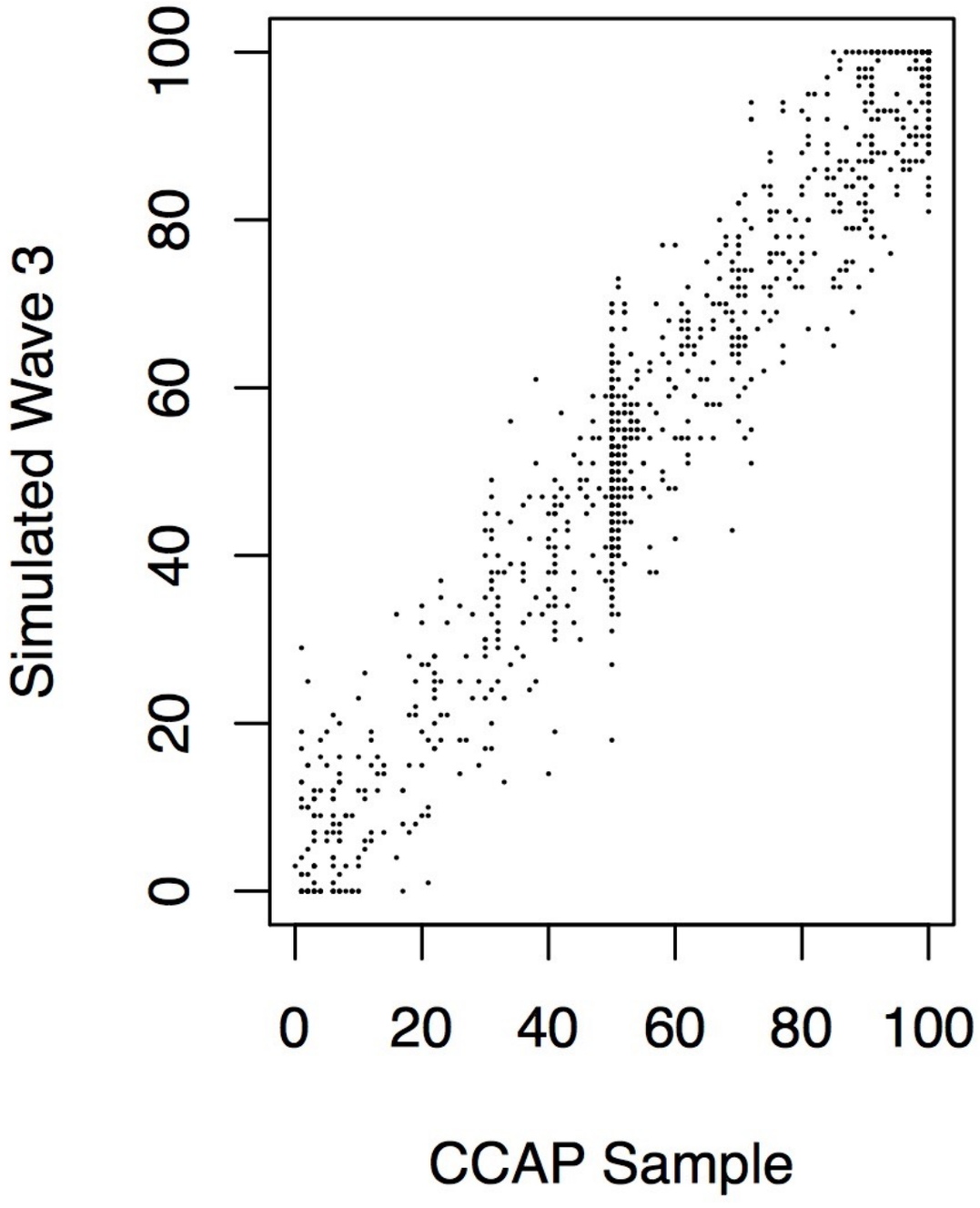
M. A. Calandra

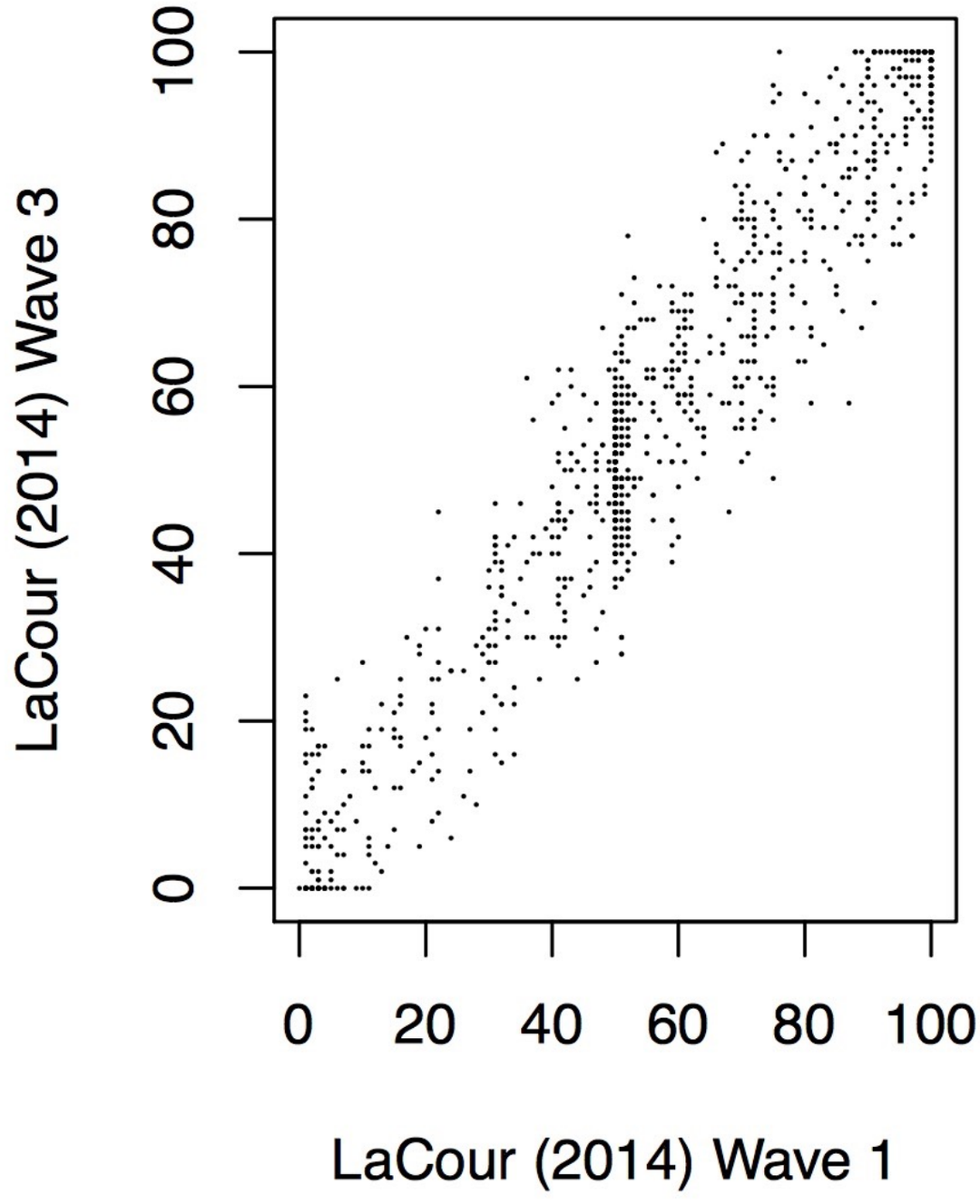
VOID AFTER 180 DAYS

[redacted]

Source: Bozeman Daily Chronicle

L'affaire Lacour …





Source: Broockman, Kalla, and Aranow. 2014. "Irregularities in LaCour (2014)."

“Human subjects research” ethics, principles, rules

- Tuskegee Syphilis Study (1939-1972)
  - → 1972: Exposed by *NYT*, shut down
  - → 1973: Congressional hearings
  - → 1979: Belmont Report
  - → 1981: *Common Rule* (and “the IRB”)
    2017: *Revised Common Rule* [effective 7/2018]

- Other social science touchstones
  - 1963: Stanley Milgram, Obedience to Authority
  - 1970: Laud Humphreys, Tearoom Trade

“Human subjects research” and data science

- The Common Rule and the IRB are …
  - designed to prevent repeats of Tuskegee, Milgram
    - Is digital different? Does the IRB scale?
  - designed for “human subjects research.” Data scientists are often not trained in this, often don’t think of their work as “human subjects research.”
- Proposed frameworks, digital science / data science
  - 2011: boyd & Crawford, “Critical questions”
  - 2012: Menlo Report
  - 2016: Association of Internet Researchers
  - 2016: Council for Big Data, Ethics, and Society

Belmont/Menlo Four Principles

- Respect for persons
- Beneficence
- Justice
- Respect for law and public interest (Menlo)

Respect for persons

- treating people as autonomous / in control
  - (and providing extra protections for those with diminished autonomy - like who?)
- → *informed consent* (“when possible”)

Informed consent

- for everything? some say yes.

- most say no … deception sometimes considered ok
  - limited harm? (is risk worse *with* consent?)
  - social benefit?
  - other methods weak? (can you *debrief*?)
  - IC logistically impossible?

Beneficence

- do not harm
- maximize benefits and minimize harms

- → *risk-benefit analysis*
  - what are the risks? probability? severity?
    - to whom … subjects? others?
  - what are the possible benefits?
  - what would reduce risk? increase benefits?
- → *weigh them* - do benefits justify the risks?

Privacy / confidentiality / informational risk

- "Anonymization" ain't enough —
  - sparsity in big data, uniqueness of network structures, GPS patterns …
  - possibility of *auxiliary data*

- Tradeoff with scientific "utility"
  - *aggregation*, *row swapping*, noise, *synthetic data*
  - *differential privacy, walled gardens*

- Data sharing, open data, the "replication crisis"

Justice

- Are risks distributed fairly? Do vulnerable groups bear more risk than others?

- Are benefits distributed fairly? Do marginalized groups have access?

- Does the research reify or magnify existing biases and injustices? Could it be applied to do so?

Respect for law and public interest

- not just participants … all relevant stakeholders
- compliance - laws, contracts, TOS, robots.txt
- *transparency-based accountability*

Revisit …

TECH 2/16/2012 @ 11:02 上午 | 2,526,432 views

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

+ Comment Now + Follow Comments

Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy. Target, for example, has figured out how to data-mine its way into your womb, to figure out whether you have a baby on the way long

Target has got you in its aim

Source: Forbes


FACEBOOK SHOULDN'T CHOOSE WHAT STUFF THEY SHOW US TO CONDUCT UNETHICAL PSYCHOLOGICAL RESEARCH.
THEY SHOULD ONLY MAKE THOSE DECISIONS BASED ON, UH...
HOWEVER THEY WERE DOING IT BEFORE.
WHICH WAS PROBABLY ETHICAL, RIGHT?

Source: xkcd

The data is public.

Methods

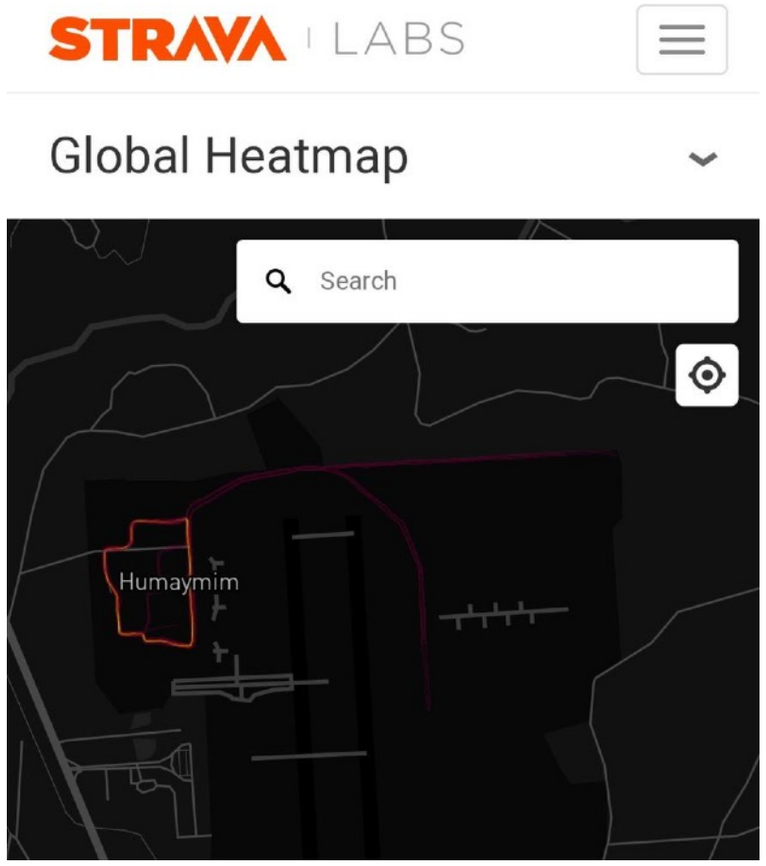
The complete details of the methodology are reviewed in the *SI Text* but are summarized briefly here. The research began by acquiring a list of registered voters in the state of New York that was accurate as of summer 2001. [Because all of the identifying information in this study stem from public records, the study was deemed exempt from Institutional Review Board review (Yale University Human Subjects Committee Protocol no. 1305011974).] The database contained personal information of all 9,995,513 New Yorkers registered at that time. Culling information from digital obituaries of 9/11 victims, I matched all 9/11 victims residing in New York to the statewide voter file (for reasons of cost and data availability, this study was restricted to 9/11 victims residing in New York at the time of the attack). Of the 1,729 victims from New York, I identified 1,181 (68%) as registered voters. This number matches the percentage of New York citizens overall who were registered to vote at the time.

For each victim, I used a combination of exact matching and Mahalanobis distance matching to identify up to five "control victims." The variables used in the matching algorithm include demographics, prior political activities, and family and neighborhood characteristics. To be clear, the control victims, who were mostly residents of metro New York and similarly situated as the true victims, were themselves obviously affected by the events of 9/11. This only biases the study toward a null result and isolates the effect under investigation as being particular to families and neighbors of victims. The causal effect estimated here is not the general effect of the 9/11 attack, but rather the specific effect of the attacks on families and neighbors.

Hersh - 2013 *PNAS*

OKCupid data release

- Danish student researcher publicly released a dataset of nearly 70,000 users of the online dating site OkCupid, including usernames, age, gender, location, what kind of relationship (or sex) they're interested in, personality traits, and answers to thousands of profiling questions used by the site



posted by Tobias Schneider (Twitter)

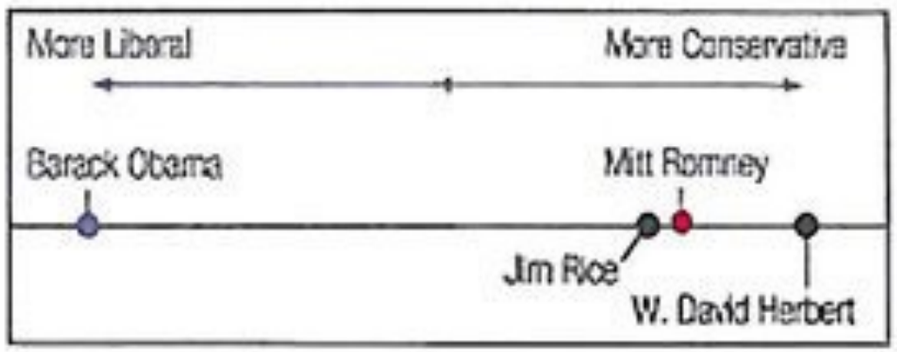
The IRB said it was fine.

VOTERS GUIDE

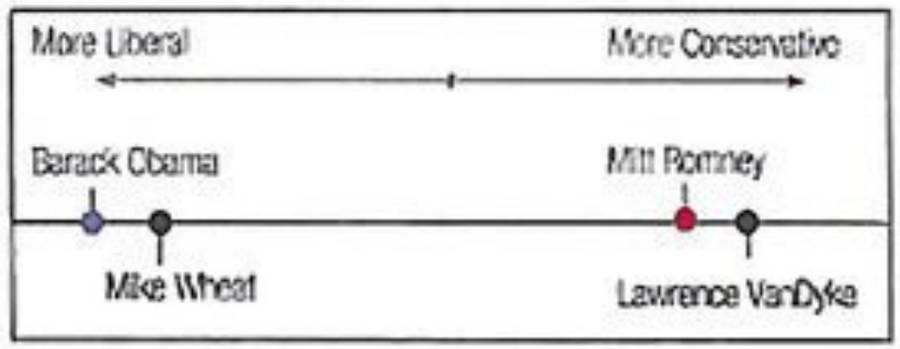
2014 Montana General Election Voter Information Guide

Election Date: November 4, 2014

Nonpartisan Supreme Court Justice #1 Race



Nonpartisan Supreme Court Justice #2 Race

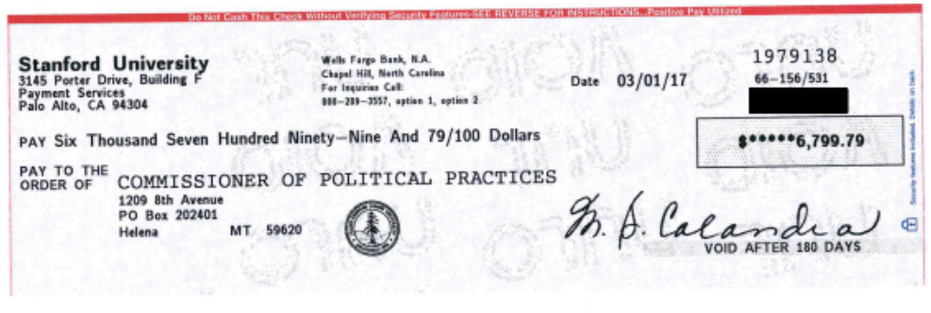


For more information on how these figures were created, please see http://data.stanford.edu/time. Please note that this guide is non-partisan and does not endorse any candidate or party. This guide was created as part of a joint research project at Stanford and Dartmouth.

Paid for by researchers at Stanford University and Dartmouth College, 616 Serra Street, Stanford, CA 94305

Take this to the polls!

Source: Bonica, Rodden, Dropp. 2014 Montana judicial elections experiment.


Do Not Cash This Check Without Verification Security Features-SEE REVERSE FOR INSTRUCTIONS - Positive Pay Utilized

Stanford University
3145 Porter Drive, Building F
Payment Services
Palo Alto, CA 94304

Wells Fargo Bank, N.A.
Chapel Hill, North Carolina
For Inquiries Call:
800-289-3557, option 1, option 2

1979138
Date 03/01/17
66-156/531
[redacted]

PAY Six Thousand Seven Hundred Ninety-Nine And 79/100 Dollars

$*****6,799.79

PAY TO THE ORDER OF COMMISSIONER OF POLITICAL PRACTICES
1209 8th Avenue
PO Box 202401
Helena MT 59620

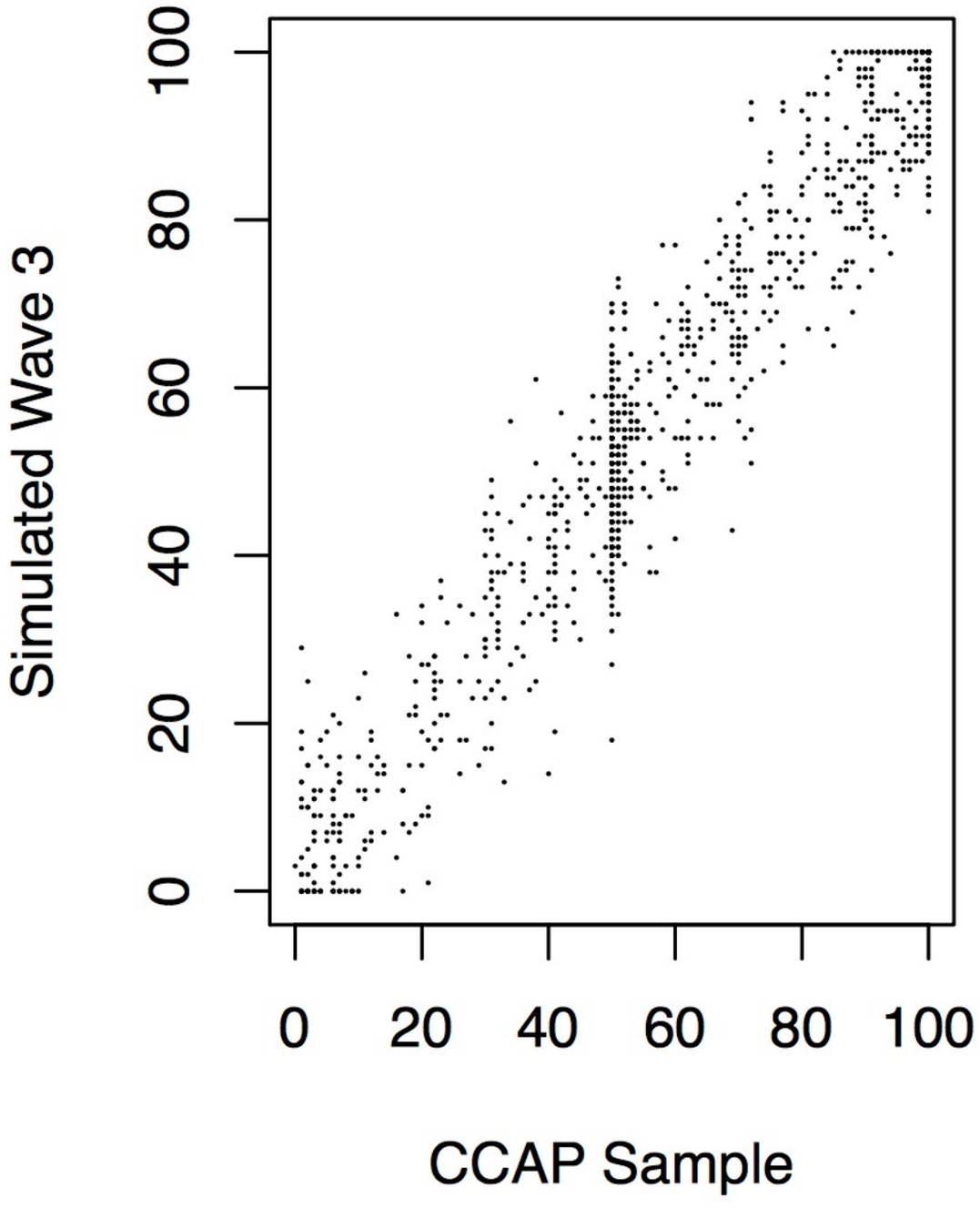
M. A. Calandra

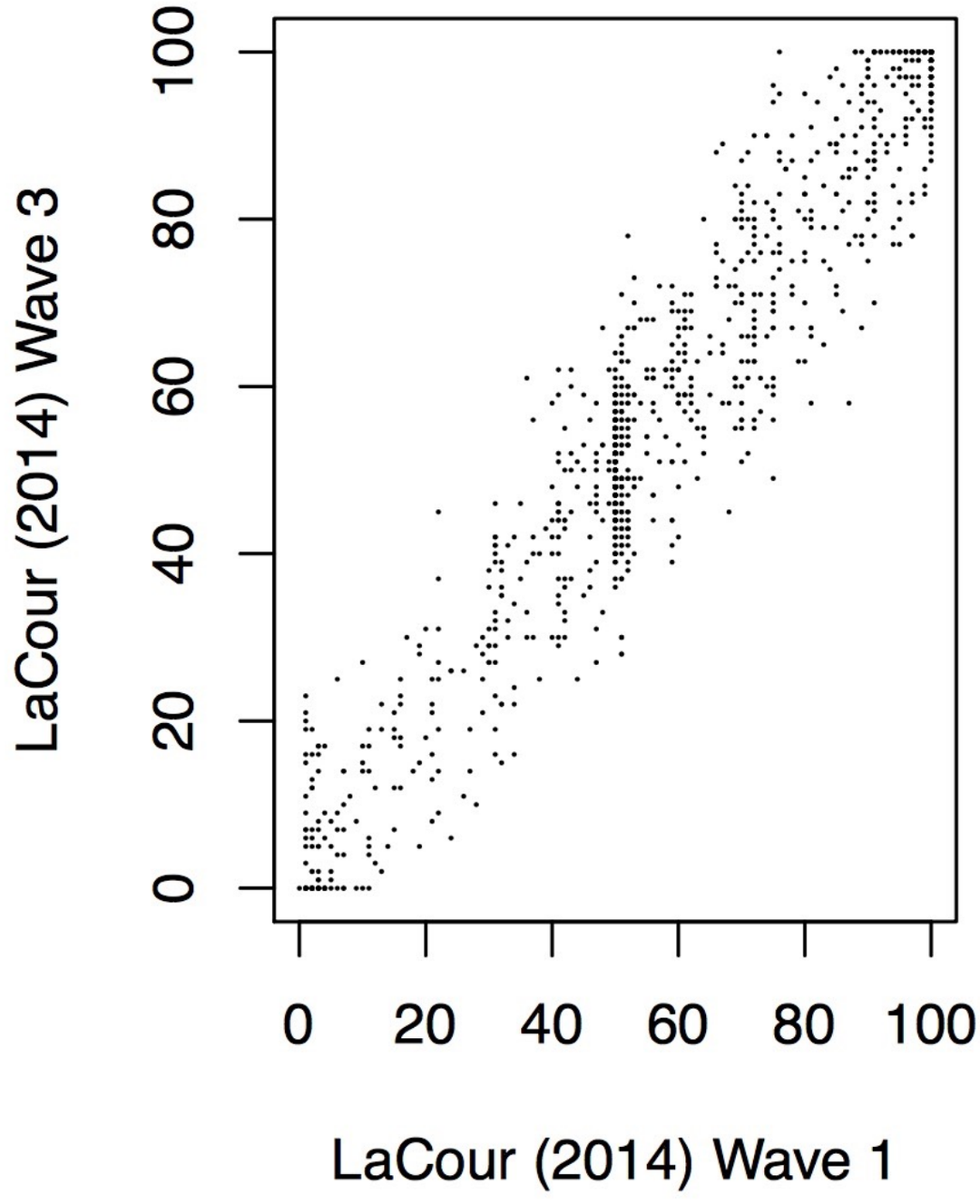
VOID AFTER 180 DAYS

[redacted]

Source: Bozeman Daily Chronicle

L'affaire Lacour …





Source: Broockman, Kalla, and Aranow. 2014. "Irregularities in LaCour (2014)."

Some extra guidelines

- “Legal” is insufficient for “ethical.” (Are there cases where “illegal” is ethical?)

- “IRB-approved” or “IRB-exempt” is insufficient for ethical (or legal).

- “These data are already public” is insufficient for “ethical.”

- Abstractly “ethical” is insufficient for “advisable.”

Extra topics …

Consequentialism and deontology

- what the what now?
- what's the difference?
- what are the tradeoffs?

Metcalf, Keller, & boyd - Policy recommendations

- Ensure the Common Rule addresses data science.

- New approaches, ethics review, academia & industry.

- Ethical assessment calibrated to big data practices.

- Integrate data ethics into NSF program solicitations.

Metcalf, Keller, & boyd

- Pedagogical interventions
  - Distribute high quality data ethics case studies.
  - Data science curricula w integrative approaches to ethics education.
  - Train librarians to achieve / promulgate data science literacy
  - Strengthen ethics-oriented activities in professional associations.

- Developing networks & cultures
  - Create hybrid spaces for ethics engagement.
  - Build models of internal/external ethics regulation bodies in industry.
  - Set standards for responsible cross-sector data sharing.

Metcalf, Keller, & boyd - Future research?

- Should human data science be regarded as human-subjects research?
- What are the quantifiable risks in correlative &/or predictive data research?
- How should we account for data sharing risk w/ unknown auxiliary data?
- How is big data redefining when/how public benefits from research? More precise ways of assessing public benefit or justice considerations?
- How should data privacy & security scientists approach illicitly gained data?
- Options for self-regulation in data science?
- What resources are needed in the university context, outside IRB?
- How can integrative approaches to data ethics be fostered in the classroom?
- What are the ecological & environmental impacts of a rise in big data research?
- How can ethical issues be integrated into core technical research?
- What motivates data scientists in industry to establish ethics processes? What ethics review structures work?
- What is the proper purview of “research ethics” as a topic in the age of big data?