STATISTICAL THINKING FOR DATA SCIENCE

CHRIS FONNESBECK
VANDERBILT UNIVERSITY

Department *of* Biostatistics

On Landing Like a Cat: It Is a Fact

Published: August 22, 1989

EVERY year, scores of cats fall from open windows in New York City. From June 4 through Nov. 4, 1984, for instance, 132 such victims were admitted to the Animal Medical Center on 62d Street in Manhattan.

Most of the cats landed on concrete. Most survived. Experts believe they were able to do so because of the laws of physics, superior balance and what might be called the flying-squirrel tactic.

In a study for the medical center, Dr. Wayne Whitney and Dr. Cheryl Mehlhaff recorded the distance of the fall for 129 of the 132 cats. The falls ranged from 2 to 32 stories, with an average distance of 5.5 stories. Two cats fell together. About a quarter fell during daylight hours, and about 40 percent at night. For the rest, the time of the fall was unknown. Surprising Data on Falls

Three cats were seen falling by their owners. Two were described as having fallen while turning on a narrow ledge, and the third had lunged for an insect.


21/22 FALLING 7+ STORIES
SURVIVED


2 FELL TOGETHER


40% AT NIGHT

HOME PAGE | TODAY'S PAPER | VIDEO | MOST POPULAR | U.S. Edition

cfonnesbeck | Help

Search All NYTimes.com

Go

WORLD | SPORTS | OPINION | ARTS | STYLE | TRAVEL | JOBS | REAL ESTATE | AUTOS

ENVIRONMENT | SPACE & COSMOS

On La

Published: A

EVERY y , s New
From Ju 4 t such
were adr treet
Manhattan.

Most of belie
they wer ior
balance and what might be called the flying-squirrel tactic.

In a stud Mehlhaf he dis to 32 stories, with an average distance of 5.5 stories. Two cats fell together. About a quarter fell during daylight hours, and about 40 percent at night. For the rest, the time of the fall was unknown. Surprising Data on Falls

Three cats were seen falling by their owners. Two were described as having fallen while turning on a narrow ledge, and the third had lunged for an insect.

MOST EMAILED | RECOMMENDED FOR YOU

articles viewe Chris
ecently

1. TH EEDIEST CASES
ping to

Beyond Comfort Zones

3. THE NEEDIEST CASES
Far From Home, a Young Man Refuses to Let One Bad Choice Determine His Fate

4. MODERN LOVE
Navigating Manila With a Map, a Bicycle and a Stormy Heart

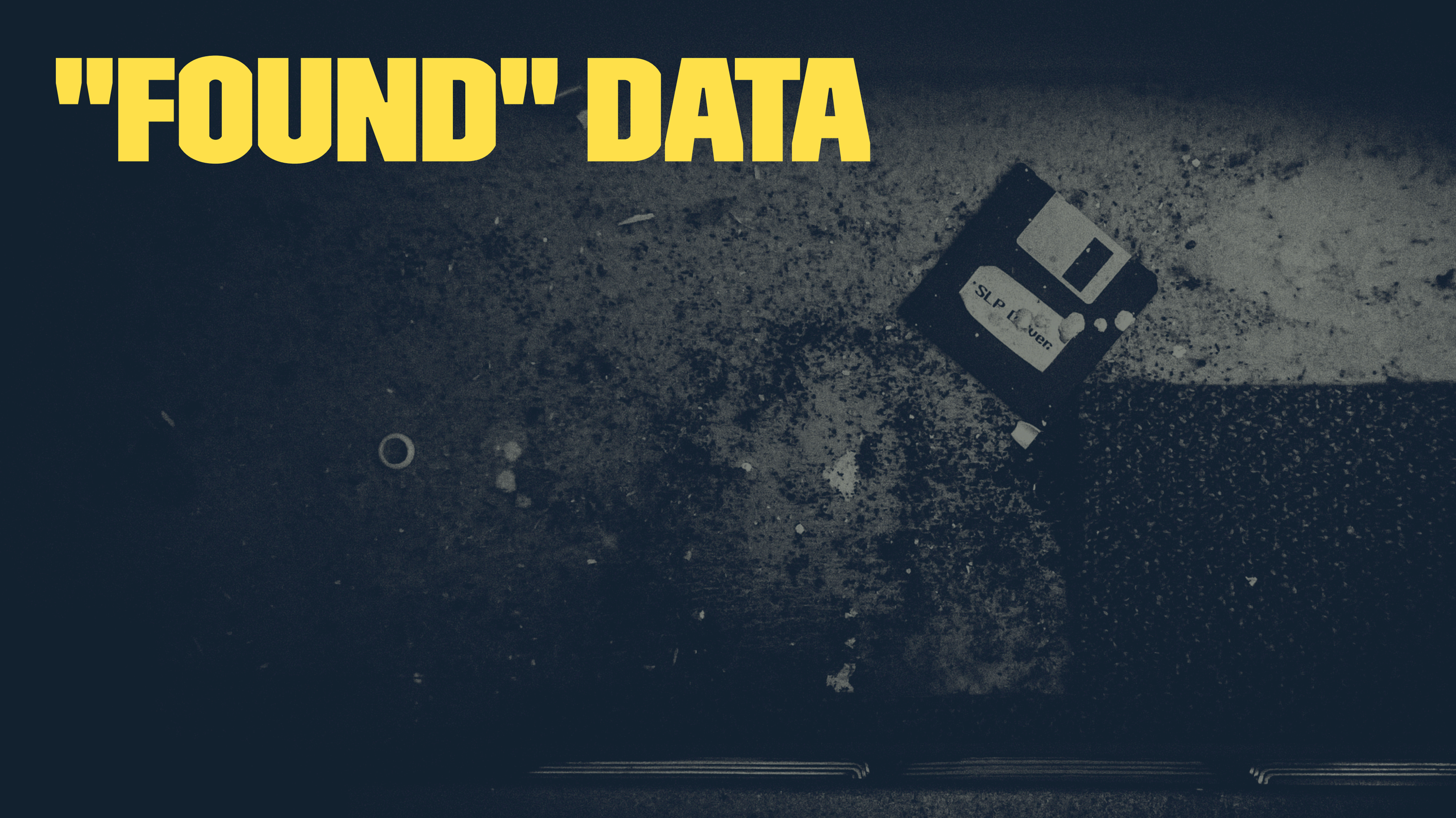
5. YOUR MONEY
A Gay Couple Awaits the Justices' Decision

"EVEN MORE SURPRISING, THE LONGER THE FALL, THE GREATER THE CHANCE OF SURVIVAL."


2 TO 32 STORIES
(AVERAGE = 5.5)

"... 132 SUCH VICTIMS WERE ADMITTED TO THE ANIMAL MEDICAL CENTER ON 62ND STREET IN MANHATTAN ..."

"FOUND" DATA



CONVENIENCE
SAMPLE

MISSING DATA

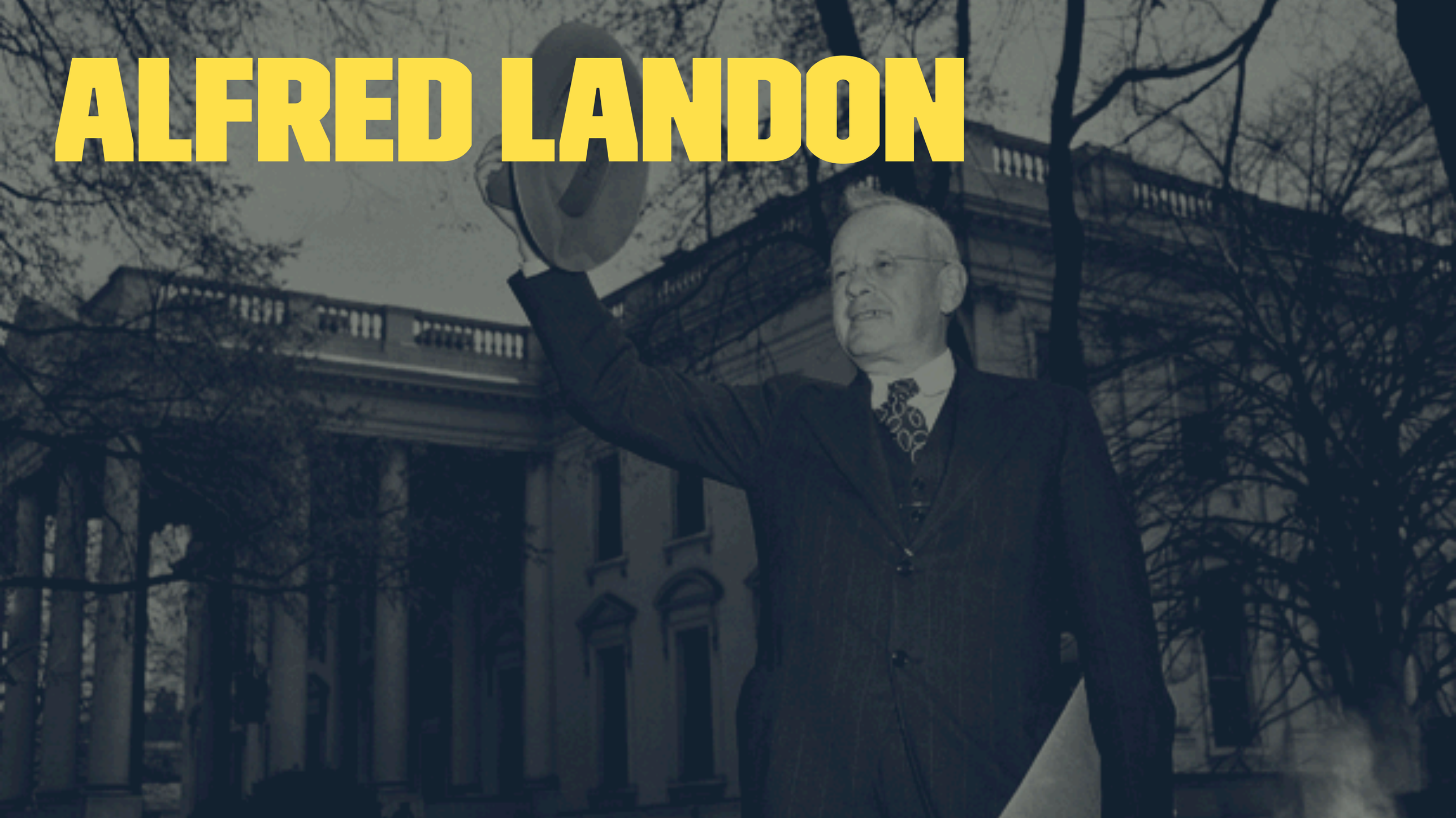
REPRESENTATIVE

STATISTICAL ISSUES

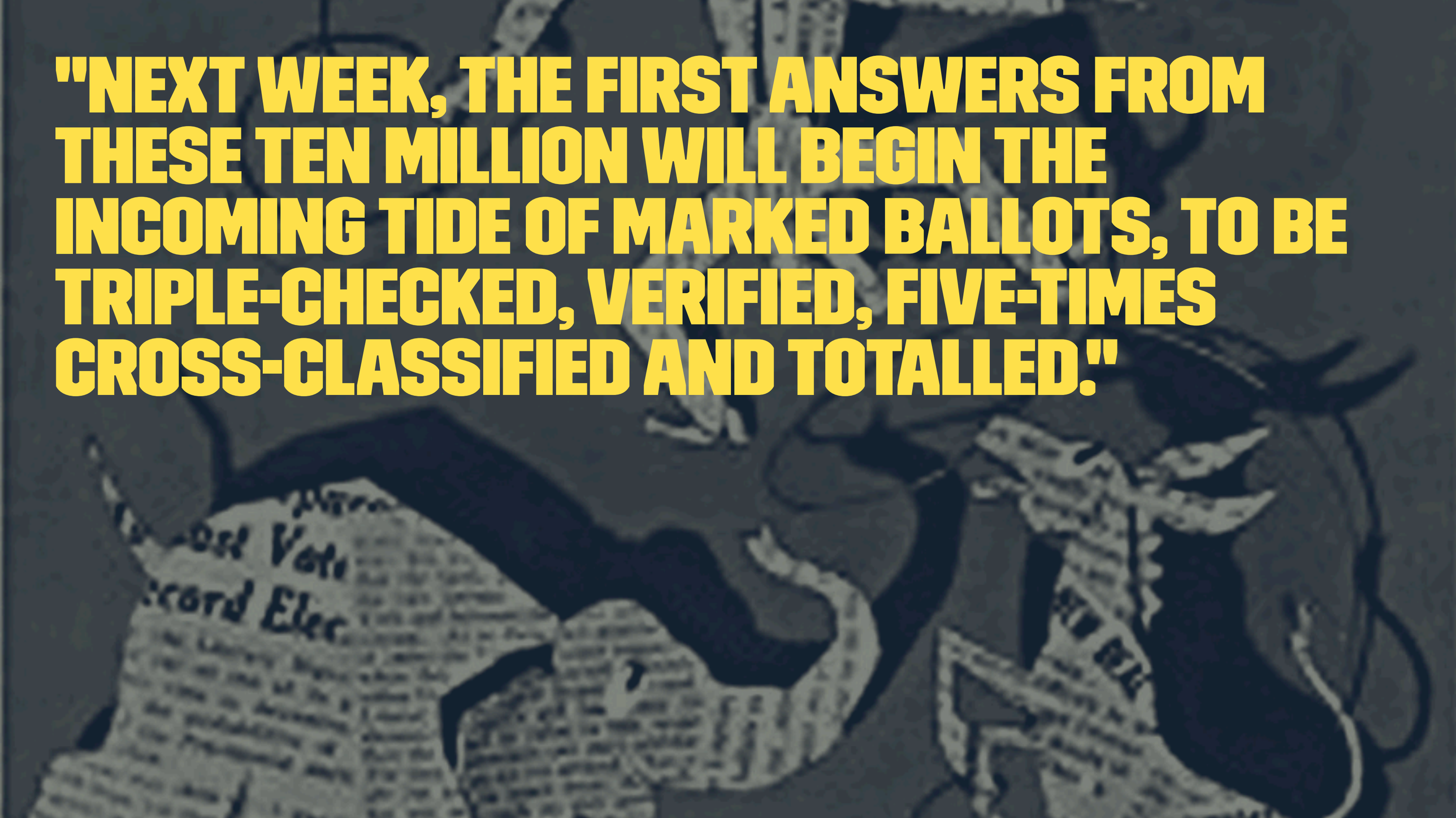

BIG DATA


“WITH ENOUGH DATA, THE NUMBERS SPEAK FOR THEMSELVES ”
Chris Anderson, Wired

ALFRED LANDON

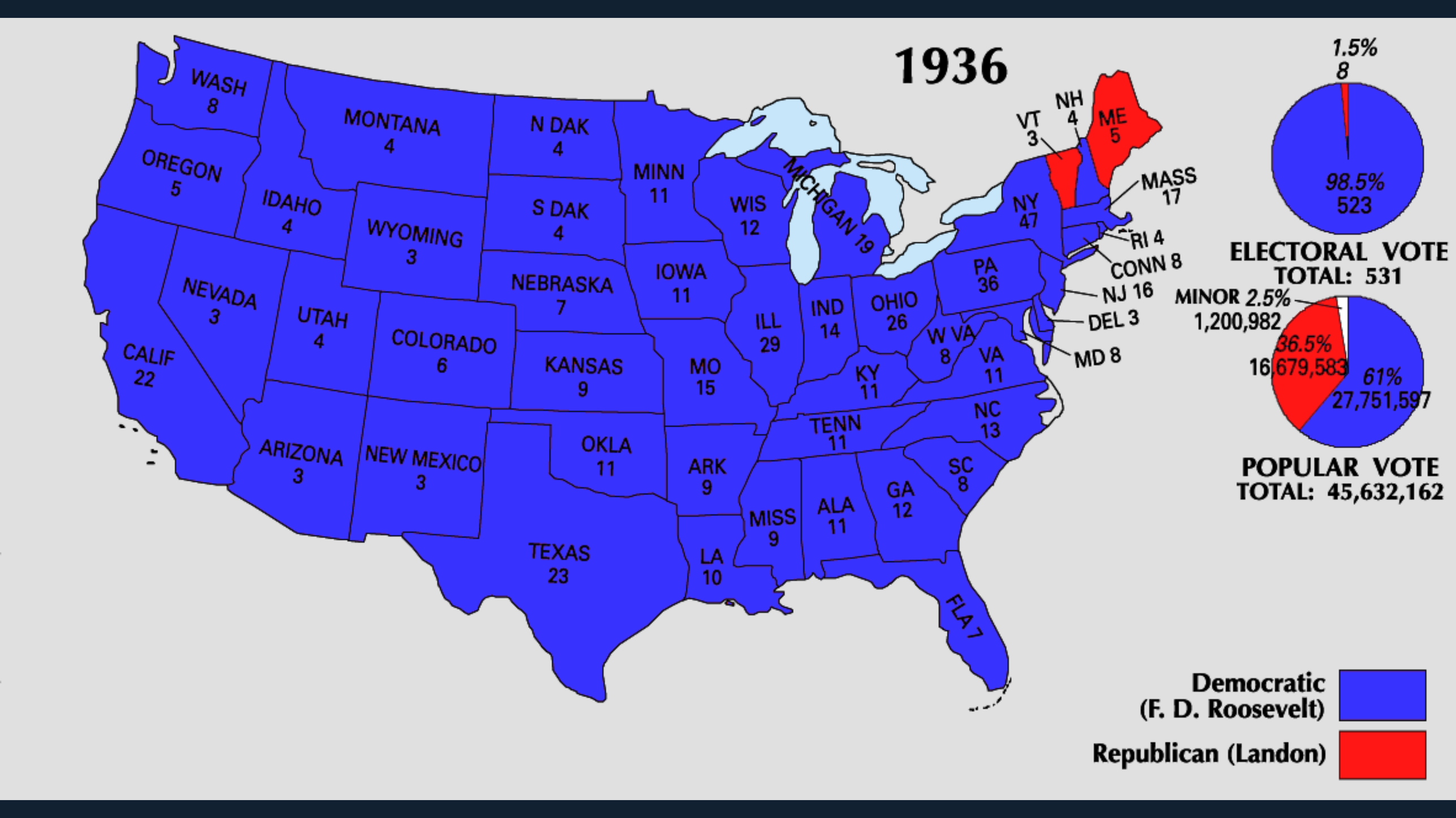


LITERARY DIGEST STRAW POLL

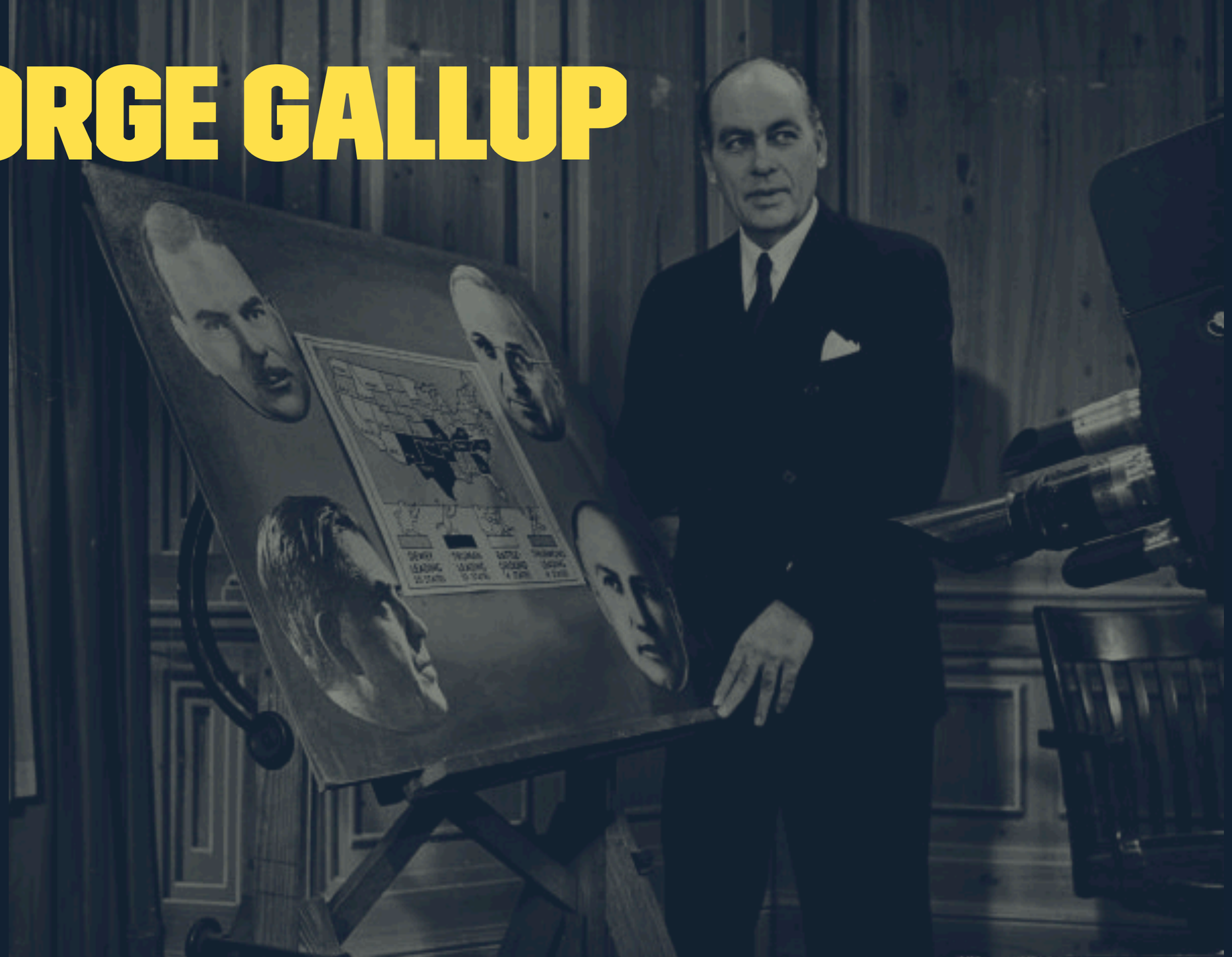

"NEXT WEEK, THE FIRST ANSWERS FROM THESE TEN MILLION WILL BEGIN THE INCOMING TIDE OF MARKED BALLOTS, TO BE TRIPLE-CHECKED, VERIFIED, FIVE-TIMES CROSS-CLASSIFIED AND TOTALLED."

2.4 MILLION RETURNS


41 - 55


1936
WASH 8
OREGON 5
CALIF 22
IDAHO 4
NEVADA 3
MONTANA 4
WYOMING 3
UTAH 4
ARIZONA 3
COLORADO 6
NEW MEXICO 3
N DAK 4
S DAK 4
NEBRASKA 7
KANSAS 9
OKLA 11
TEXAS 23
MINN 11
IOWA 11
MO 15
ARK 9
LA 10
WIS 12
ILL 29
MICHIGAN 19
IND 14
OHIO 26
KY 11
TENN 11
MISS 9
ALA 11
GA 12
FLA 7
SC 8
NC 13
VA 11
W VA 8
PA 36
NY 47
VT 3
NH 4
ME 5
MASS 17
RI 4
CONN 8
NJ 16
DEL 3
MD 8
1.5%
8
98.5%
523
ELECTORAL VOTE
TOTAL: 531
MINOR 2.5%
1,200,982
36.5%
16,679,583
61%
27,751,597
POPULAR VOTE
TOTAL: 45,632,162
Democratic
(F. D. Roosevelt)
Republican (Landon)

GEORGE GALLUP



SAMPLED

50,000


66%

RANDOM SAMPLING

Table 1. Demographics of HSCT by High Dose and Standard Group

Characteristic	HD Vaccine N=29	SD Vaccine N=15	P Value
Age at Enrollment	50	50	0.64
Gender			
Female	12/29 (41%)	5/15 (33%)	
Male	17/29 (59%)	10/15 (67%)	
Race			
White	29/29	15/15	
White blood count	6.4 (4.5, 8.3)	5.1 (3.8, 8.9)	0.69
Hemoglobin	12.3 (11.3, 13.4)	11.7 (10.7, 13.1)	0.58
Hematocrit	36.6 (34.0, 40.0)	36 (32.0, 38.5)	0.49
Platelets	206 (173, 244)	195 (134, 254)	0.92
Absolute Neutrophil Count	3.9 (2.6, 4.7)	4.3 (2.9, 5.5)	0.76
Absolute Lymphocyte Count	1.11 (0.59, 1.60)	0.80 (0.55, 1.83)	0.85
Quantitative IgG	676 (478, 838)	469 (404, 555)	0.025


BIAS

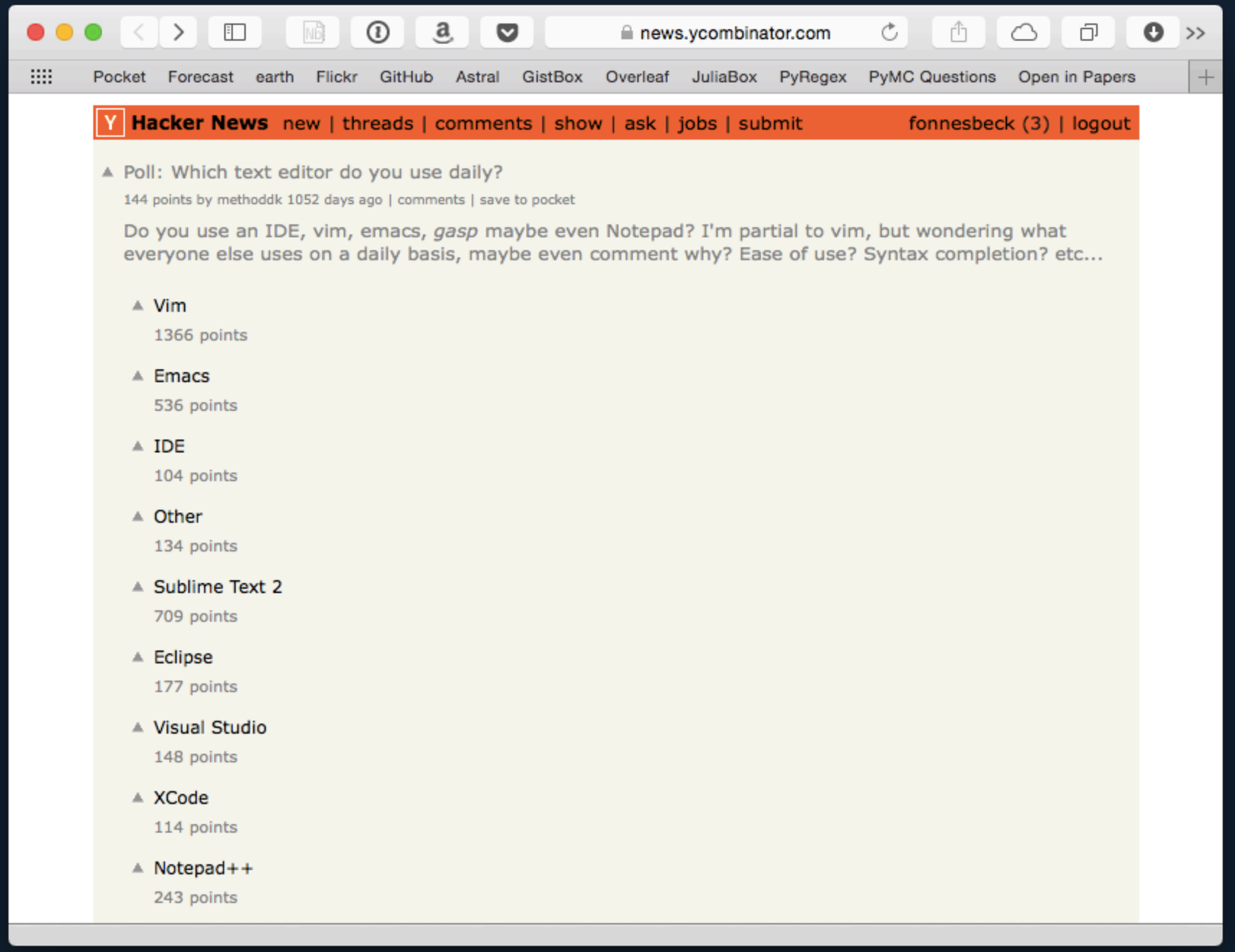
Presidential Vote	Received Poll	Not Receive Poll	Do Not Know
Roosevelt	55	71	73
Landon	44	27	25
Other	1	1	3
Total *N*	780	1339	149

SOURCE: American Institute of Public Opinion, 28 May 1937.

Presidential Vote	Did Return	Did Not Return	Do Not Know
Roosevelt	48	69	56
Landon	51	30	40
Other	1	1	4
Total *N*	493	288	48

SOURCE: American Institute of Public Opinion, 28 May 1937.

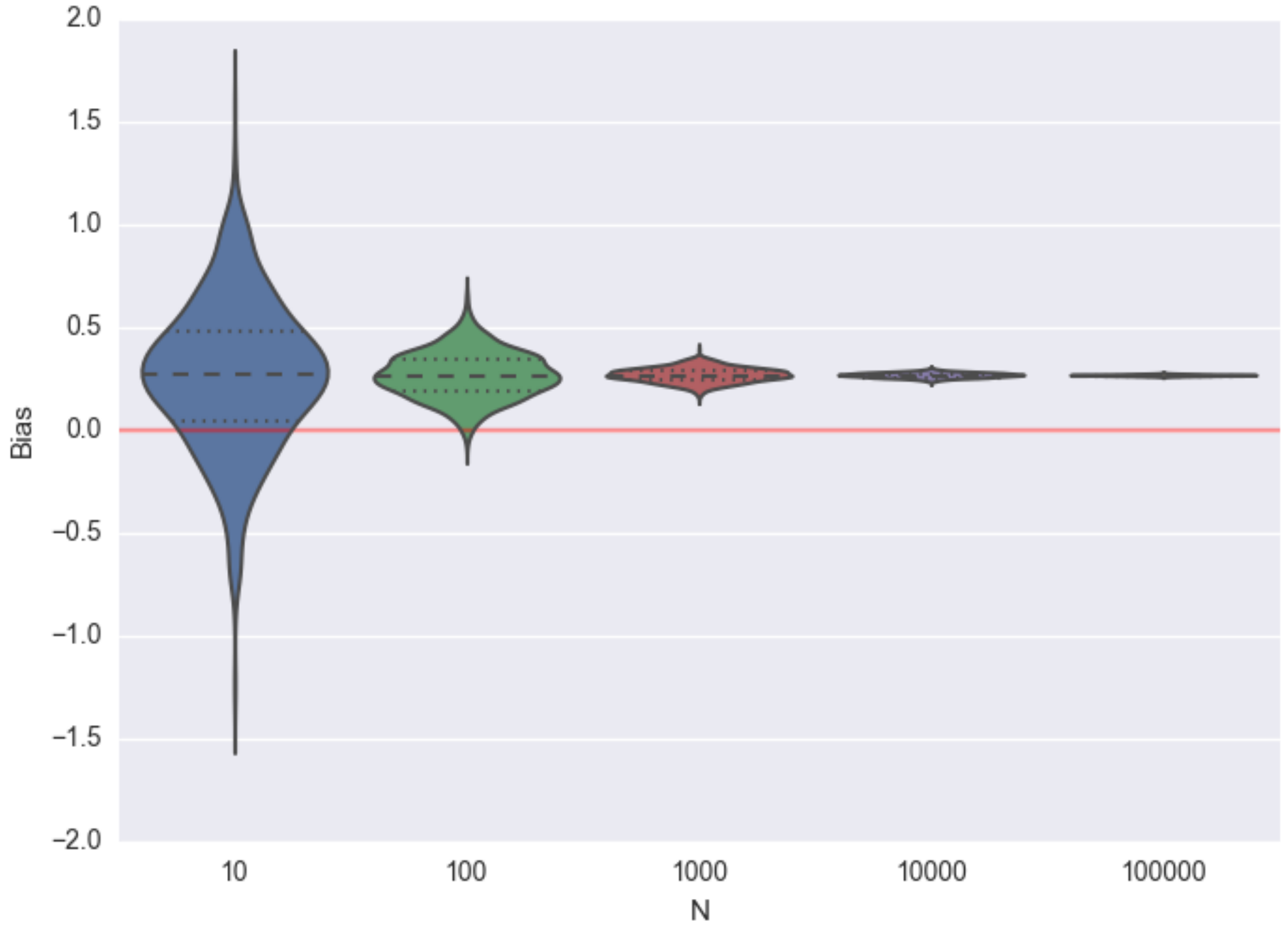
SELF-SELECTION BIAS


news.ycombinator.com
Pocket Forecast earth Flickr GitHub Astral GistBox Overleaf JuliaBox PyRegex PyMC Questions Open in Papers
Hacker News new | threads | comments | show | ask | jobs | submit fonnesbeck (3) | logout
Poll: Which text editor do you use daily?
144 points by methoddk 1052 days ago | comments | save to pocket
Do you use an IDE, vim, emacs, gasp maybe even Notepad? I'm partial to vim, but wondering what everyone else uses on a daily basis, maybe even comment why? Ease of use? Syntax completion? etc...
Vim
1366 points
Emacs
536 points
IDE
104 points
Other
134 points
Sublime Text 2
709 points
Eclipse
177 points
Visual Studio
148 points
XCode
114 points
Notepad++
243 points

For some estimate $\hat{\theta}$ of unknown quantity θ,

$$\text{Bias}[\hat{\theta}] = E_\theta[\hat{\theta}] - \theta$$

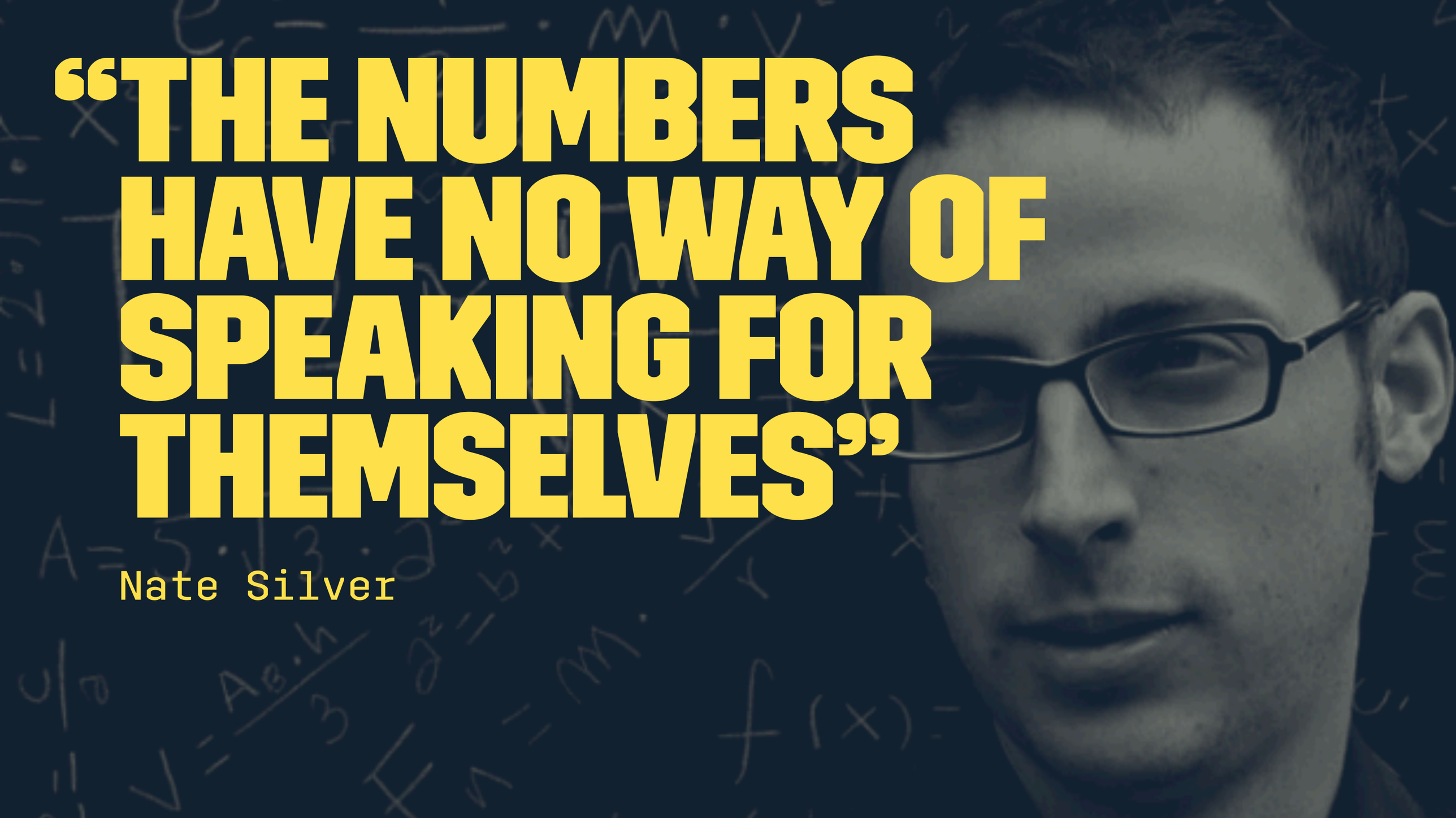
```
p = 0.5
sample_sizes = [10, 100, 1000, 10000, 100000]
replicates = 1000
biases = []

for n in sample_sizes:

    bias = np.empty(replicates)
    for i in range(replicates):

        true_sample = np.random.normal(size=n)
        negative_values = true_sample<0
        missing = np.random.binomial(1, p, n).astype(bool)
        observed_sample = true_sample[~(negative_values & missing)]

        bias[i] = observed_sample.mean()

    biases.append(bias)
```


2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
Bias
10
100
1000
10000
100000
N

ACCURACY

Mean Squared Error

$$E[(\hat{\theta} - \theta)^2] = \mathbf{Var}(\hat{\theta}) + \mathbf{Bias}(\hat{\theta})^2$$


“THE NUMBERS HAVE NO WAY OF SPEAKING FOR THEMSELVES”
Nate Silver

WHITE HOUSE BIG DATA PARTNERS WORKSHOP


WHITE HOUSE BIG DATA
PARTNERS WORKSHOP
19 PARTICIPANTS
0 STATISTICIANS

NSF WORKING GROUP ON BIG DATA

NSF WORKING GROUP ON BIG DATA

100 EXPERTS CONVENED

0 STATISTICIANS

MOORE FOUNDATION DATA SCIENCE ENVIRONMENTS

GORDON AN
MOO
FOUNDA

MOORE FOUNDATION DATA SCIENCE ENVIRONMENTS

0 DIRECTORS WITH STATISTICAL EXPERTISE

NIH BD2K EXECUTIVE COMMITTEE

NIH BD2K EXECUTIVE COMMITTEE

17 COMMITTEE MEMBERS

0 STATISTICIANS

National Institutes of Health

FEELING LEFT OUT?

IT'S OUR OWN FAULT

“ALMOST EVERYTHING YOU LEARNED IN YOUR COLLEGE STATISTICS COURSE WAS WRONG”

TYPICAL INTRODUCTORY STATISTICS SYLLABUS

1. Descriptive statistics and plotting

TYPICAL INTRODUCTORY STATISTICS SYLLABUS

1. Descriptive statistics and plotting
2. Basic probability

TYPICAL INTRODUCTORY STATISTICS SYLLABUS

1. Descriptive statistics and plotting
2. Basic probability
3. Hypothesis testing

TYPICAL INTRODUCTORY STATISTICS SYLLABUS

1. Descriptive statistics and plotting
2. Basic probability
3. Hypothesis testing
4. Experimental design

TYPICAL INTRODUCTORY STATISTICS SYLLABUS

1. Descriptive statistics and plotting
2. Basic probability
3. Hypothesis testing
4. Experimental design
5. ANOVA

STATISTICAL HYPOTHESIS TESTING

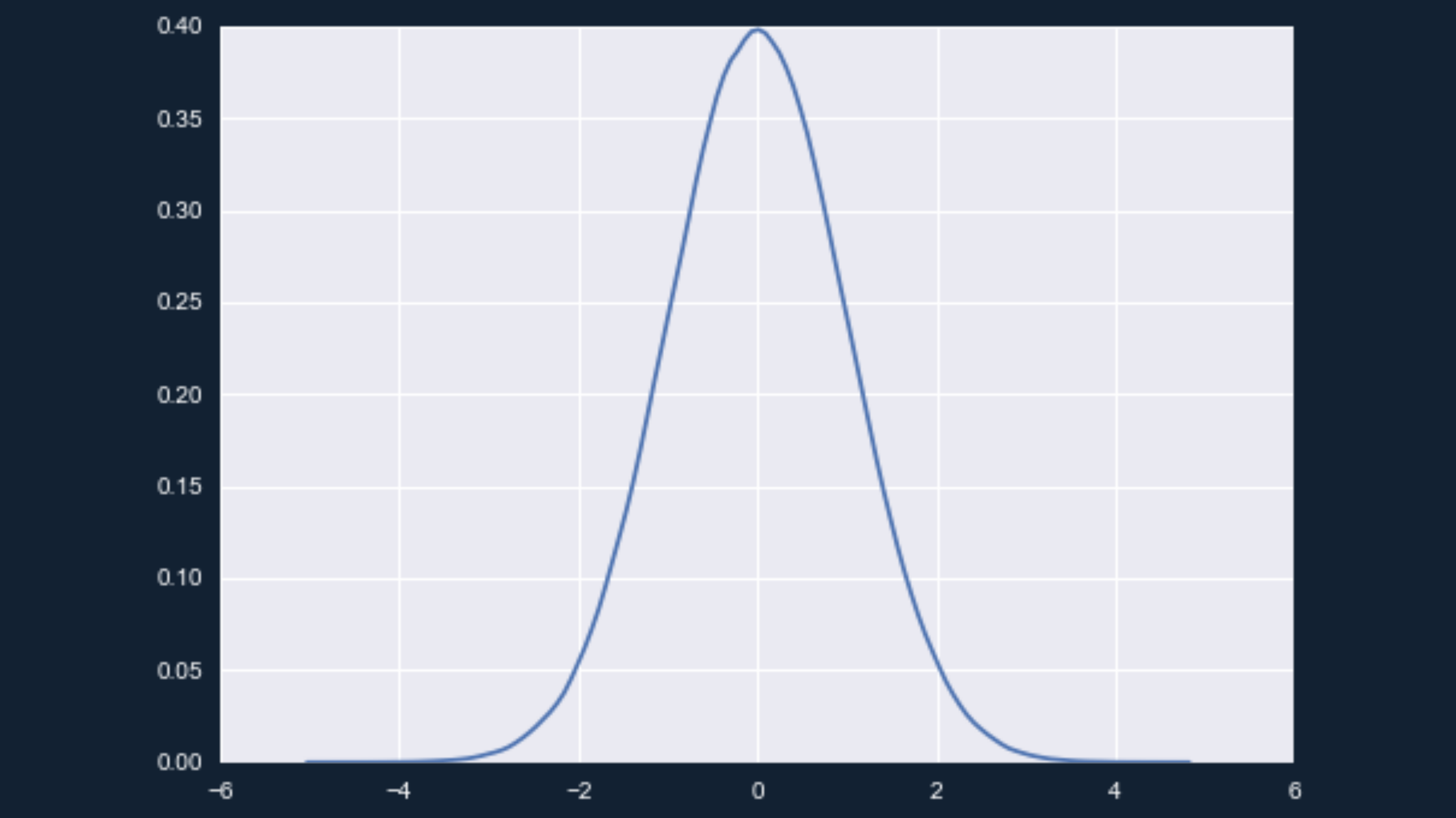
H_0

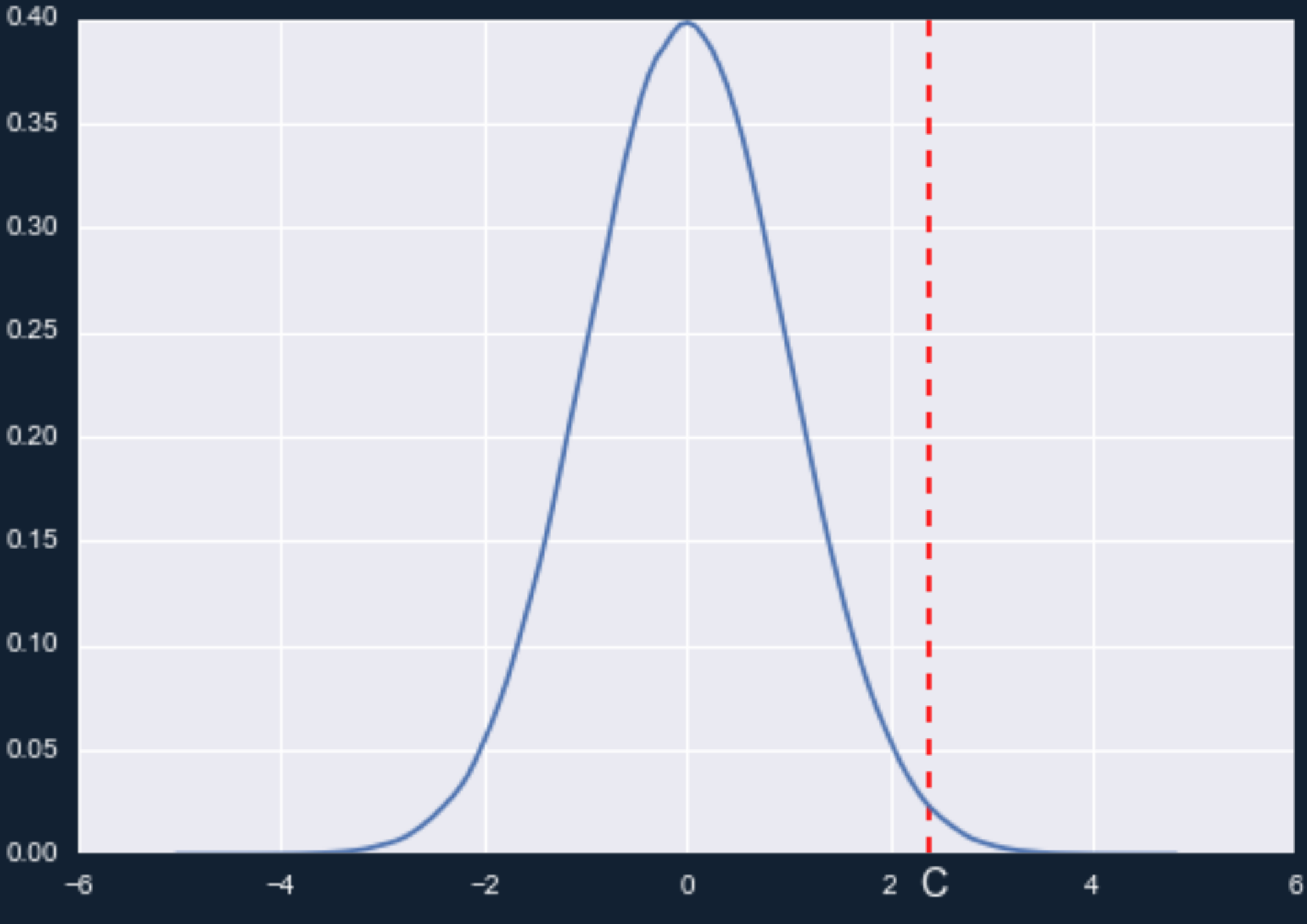
TEST STATISTIC

$$T$$

T-STATISTIC

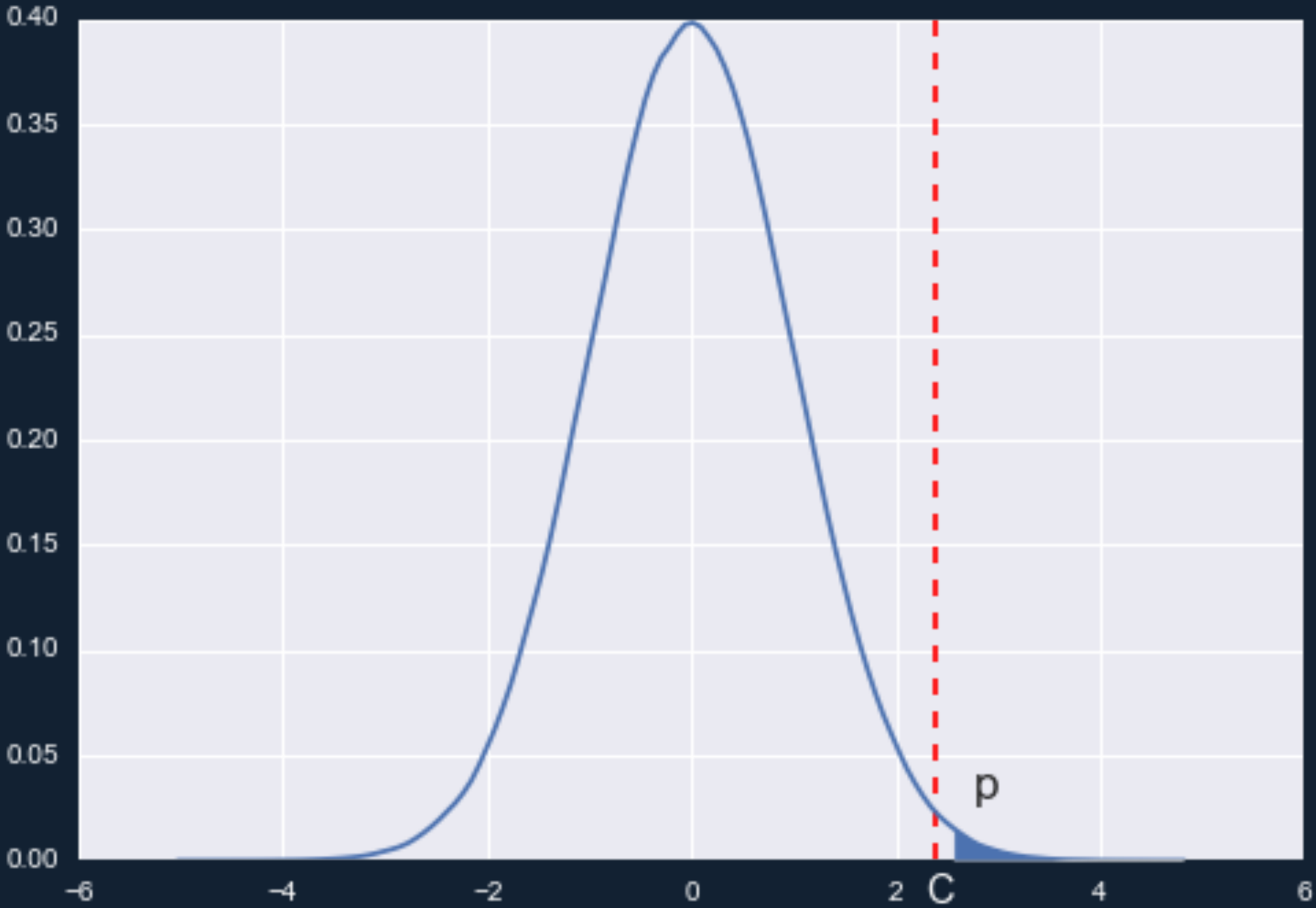
$$T = \sqrt{n}\,\frac{\bar{x} - \theta_0}{S}$$

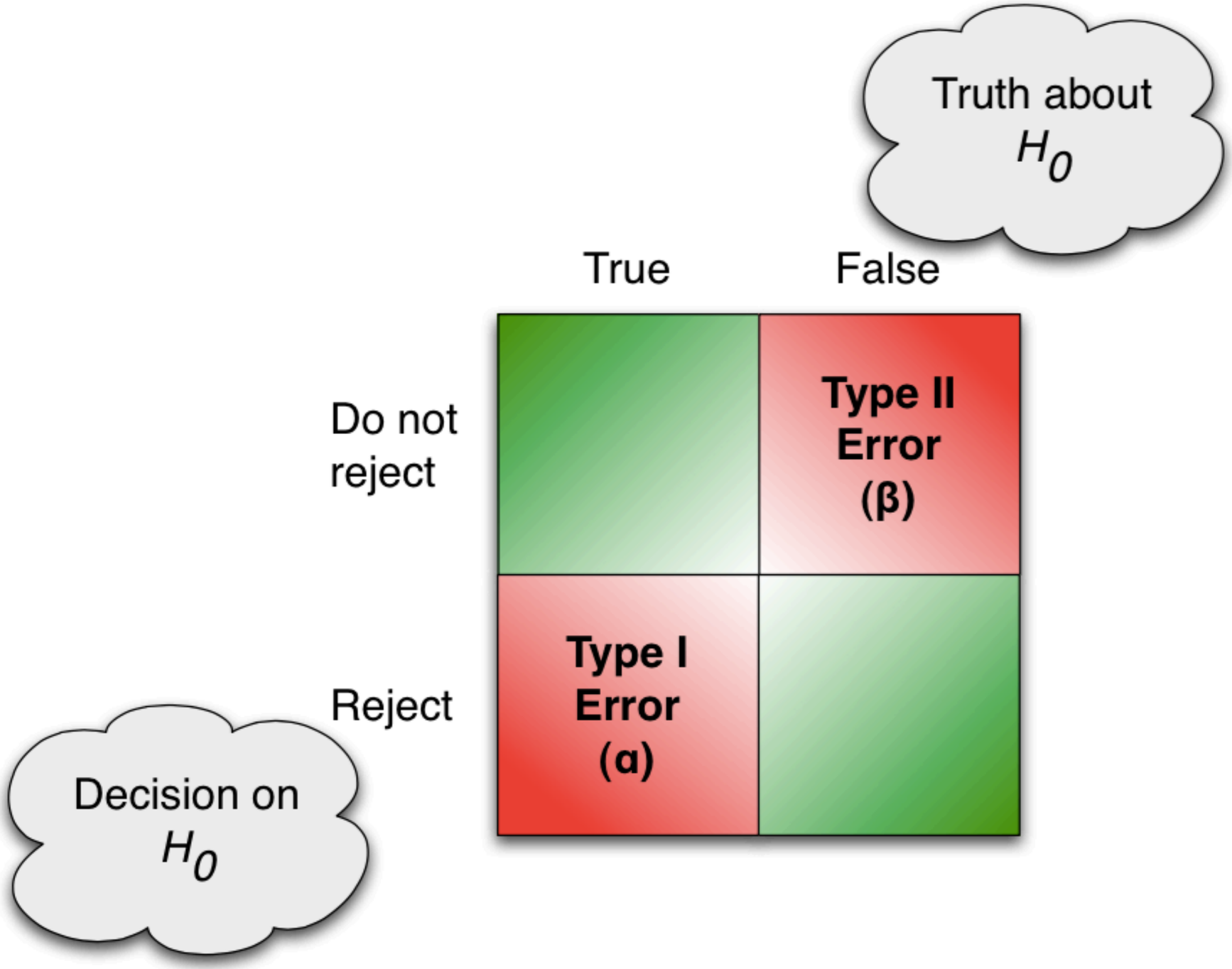

0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−6
−4
−2
0
2
4
6


0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−6
−4
−2
0
2
C
4
6

$$Pr(T > c|H_0)$$

P-VALUE


0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−6
−4
−2
0
2
C
4
6
p

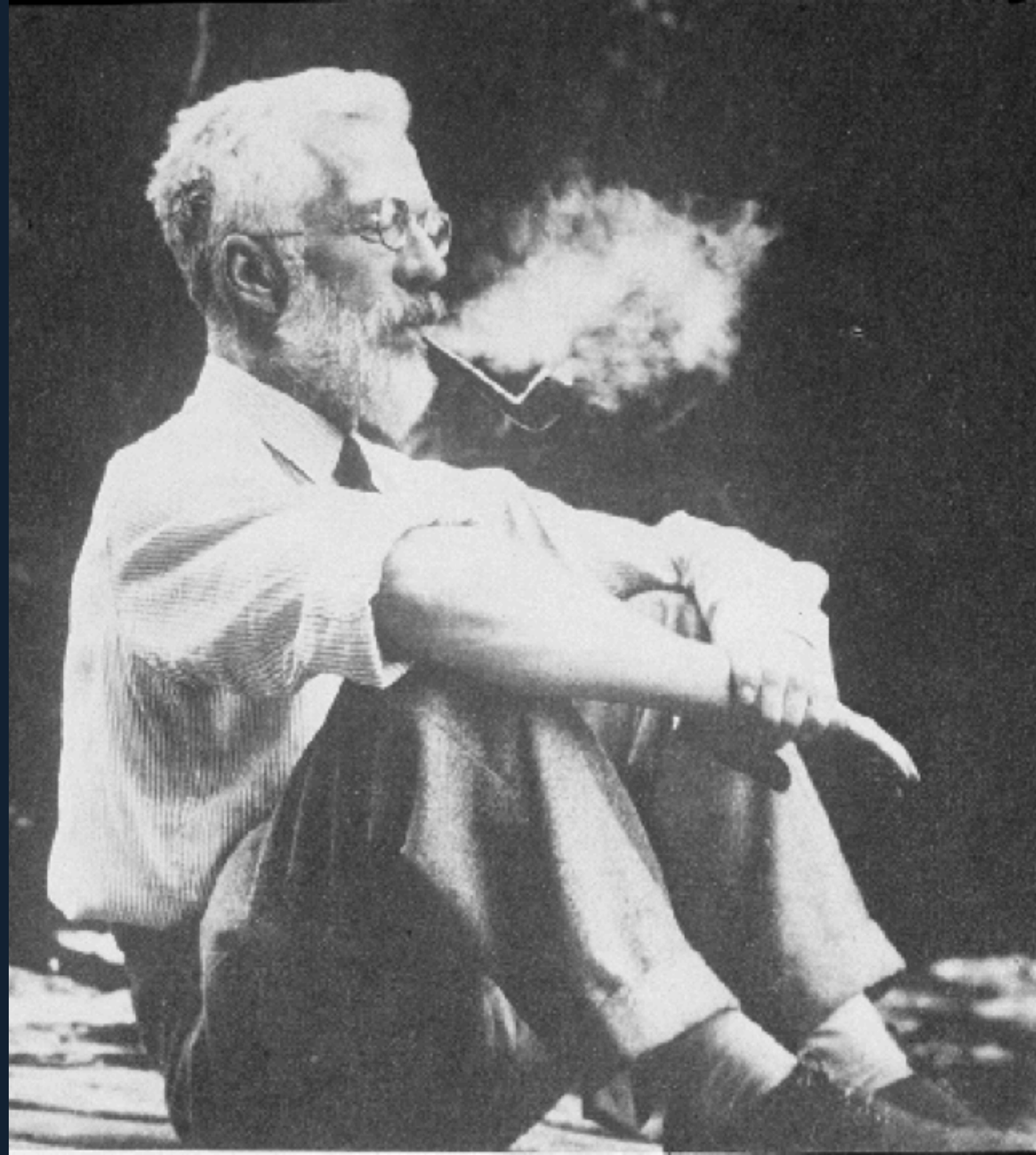

Truth about H_0
True
False
Do not reject
Reject
Type II Error (β)
Type I Error (α)
Decision on H_0

FALSE POSITIVE RATE

$$\alpha == 0.05$$

"THE VALUE FOR WHICH $p = 0.05$, OR 1 IN 20, IS 1.96 OR NEARLY 2; IT IS CONVENIENT TO TAKE THIS POINT AS A LIMIT IN JUDGING WHETHER A DEVIATION OUGHT TO BE CONSIDERED SIGNIFICANT OR NOT."

R.A. Fisher



P-VALUE

THE PROBABILITY THAT THE OBSERVED DIFFERENCES ARE DUE TO CHANCE

THE PROBABILITY THAT THE OBSERVED DIFFERENCES ARE DUE TO CHANCE

A MEASURE OF THE RELIABILITY OF THE RESULT

$$(1 - p)$$

~~A MEASURE OF THE~~
~~RELIABILITY~~ ~~OF THE~~
~~RESULT~~

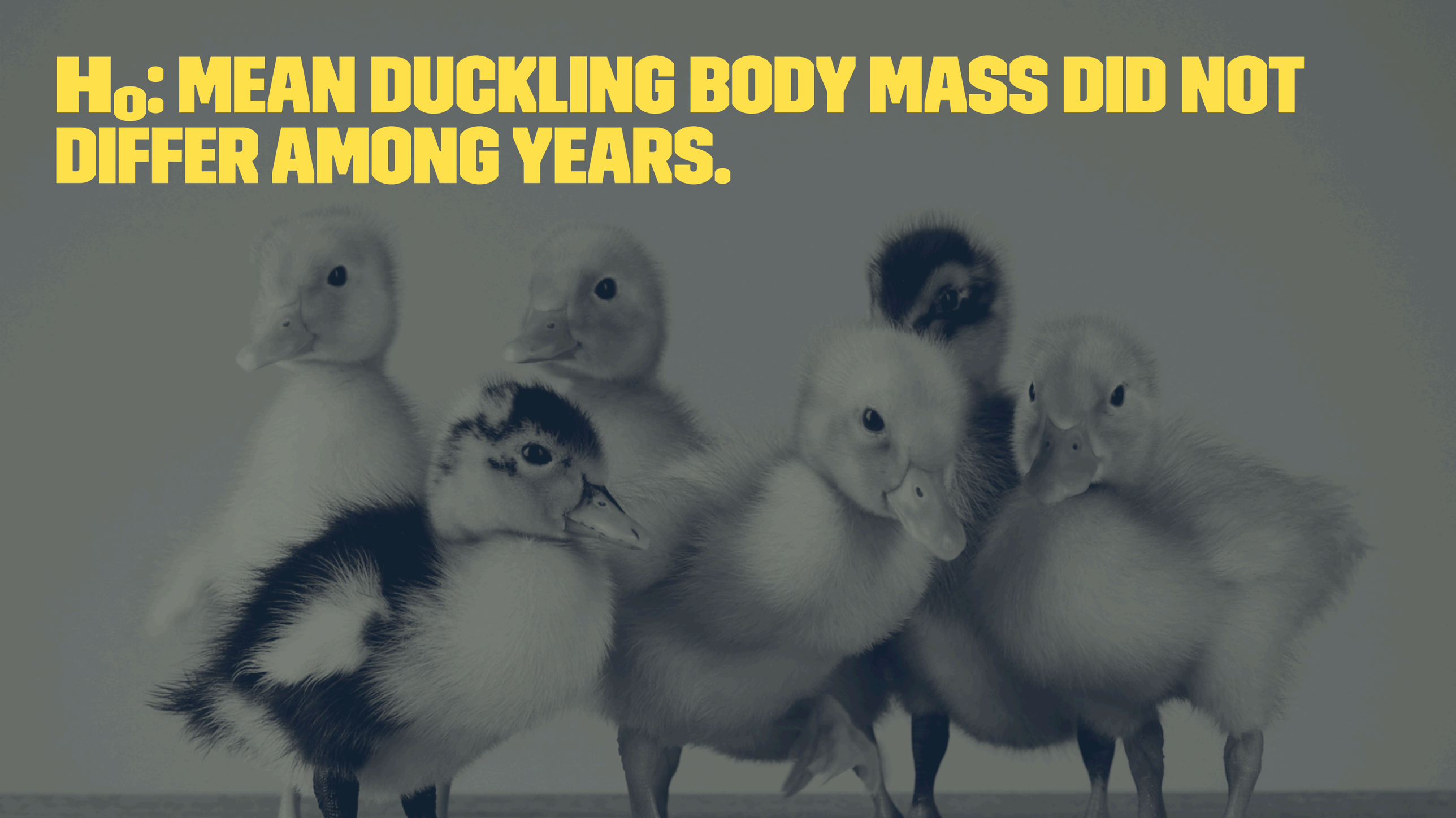
$$(1 - p)$$

THE PROBABILITY THAT THE NULL HYPOTHESIS IS TRUE

THE PROBABILITY THAT THE NULL HYPOTHESIS IS TRUE

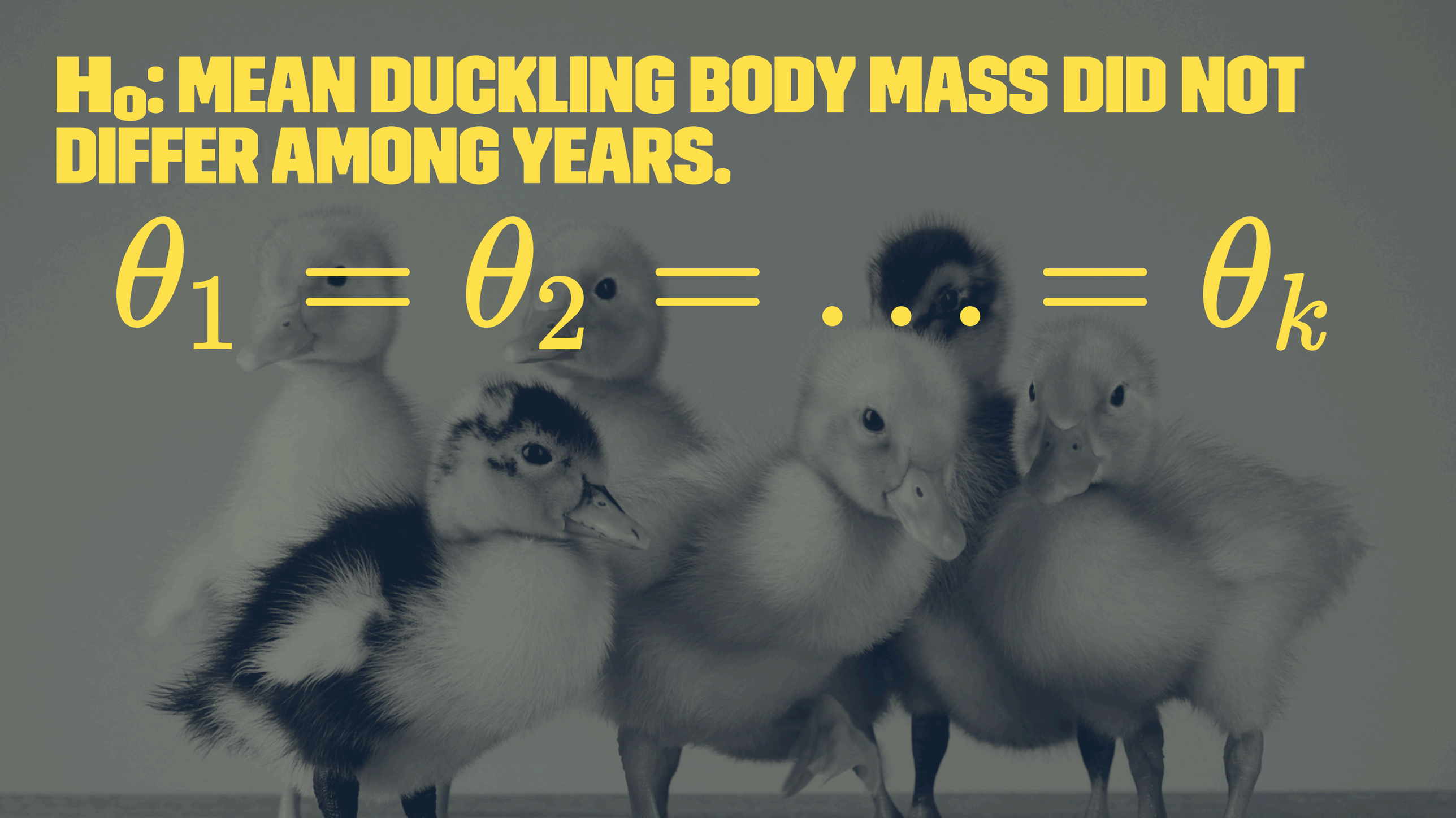
"IF AN EXPERIMENT WERE REPEATED INFINITELY, P REPRESENTS THE PROPORTION OF VALUES MORE EXTREME THAN THE OBSERVED VALUE, GIVEN THAT THE NULL HYPOTHESIS IS TRUE."

H_0: MEAN DUCKLING BODY MASS DID NOT DIFFER AMONG YEARS.



H_o: MEAN DUCKLING BODY MASS DID NOT DIFFER AMONG YEARS.

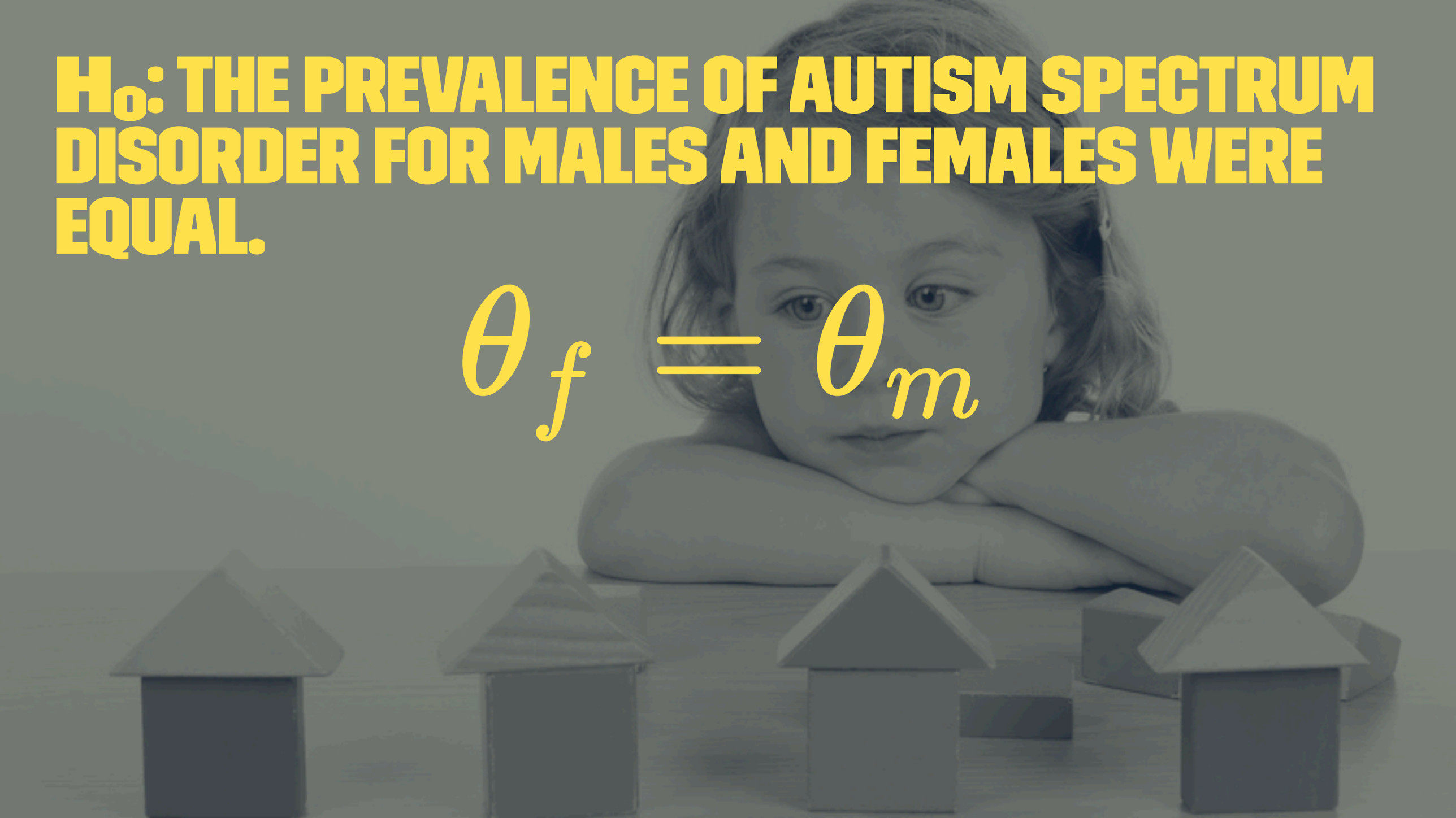
$$\theta_1 = \theta_2 = ... = \theta_k$$



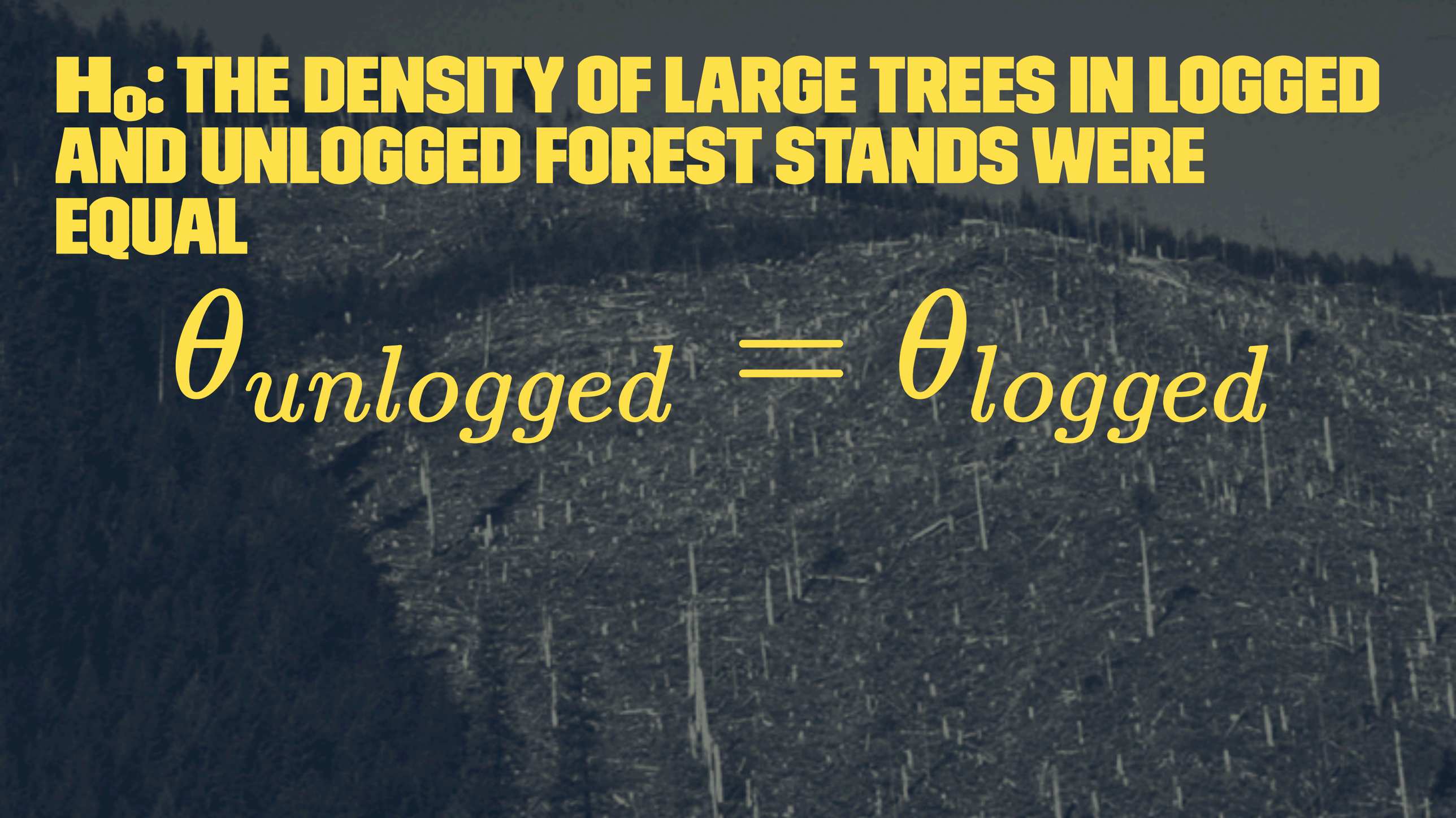
H_0: THE PREVALENCE OF AUTISM SPECTRUM DISORDER FOR MALES AND FEMALES WERE EQUAL.

H_0: THE PREVALENCE OF AUTISM SPECTRUM DISORDER FOR MALES AND FEMALES WERE EQUAL.

$$\theta_f = \theta_m$$

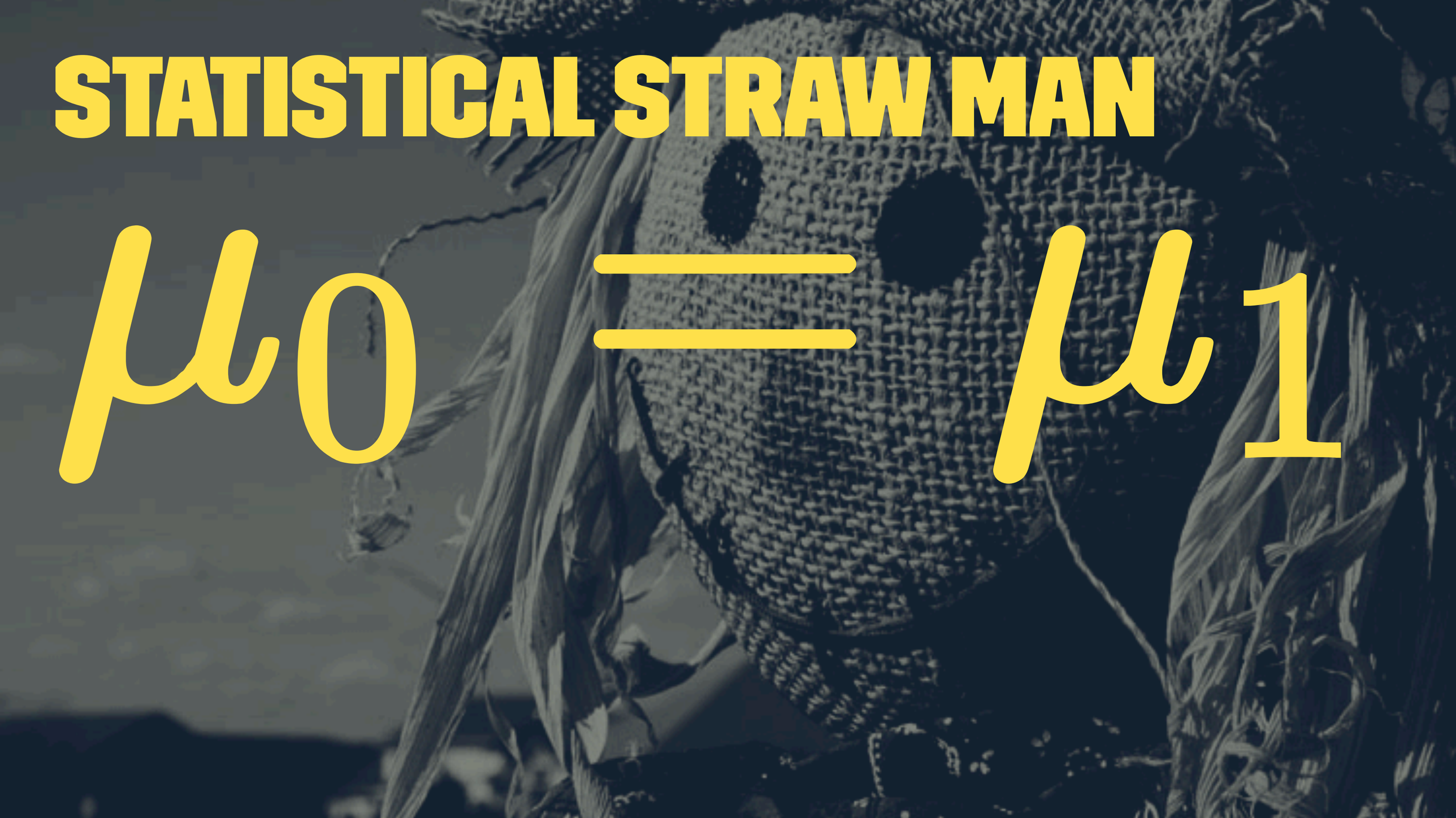


H_0: THE DENSITY OF LARGE TREES IN LOGGED AND UNLOGGED FOREST STANDS WERE EQUAL

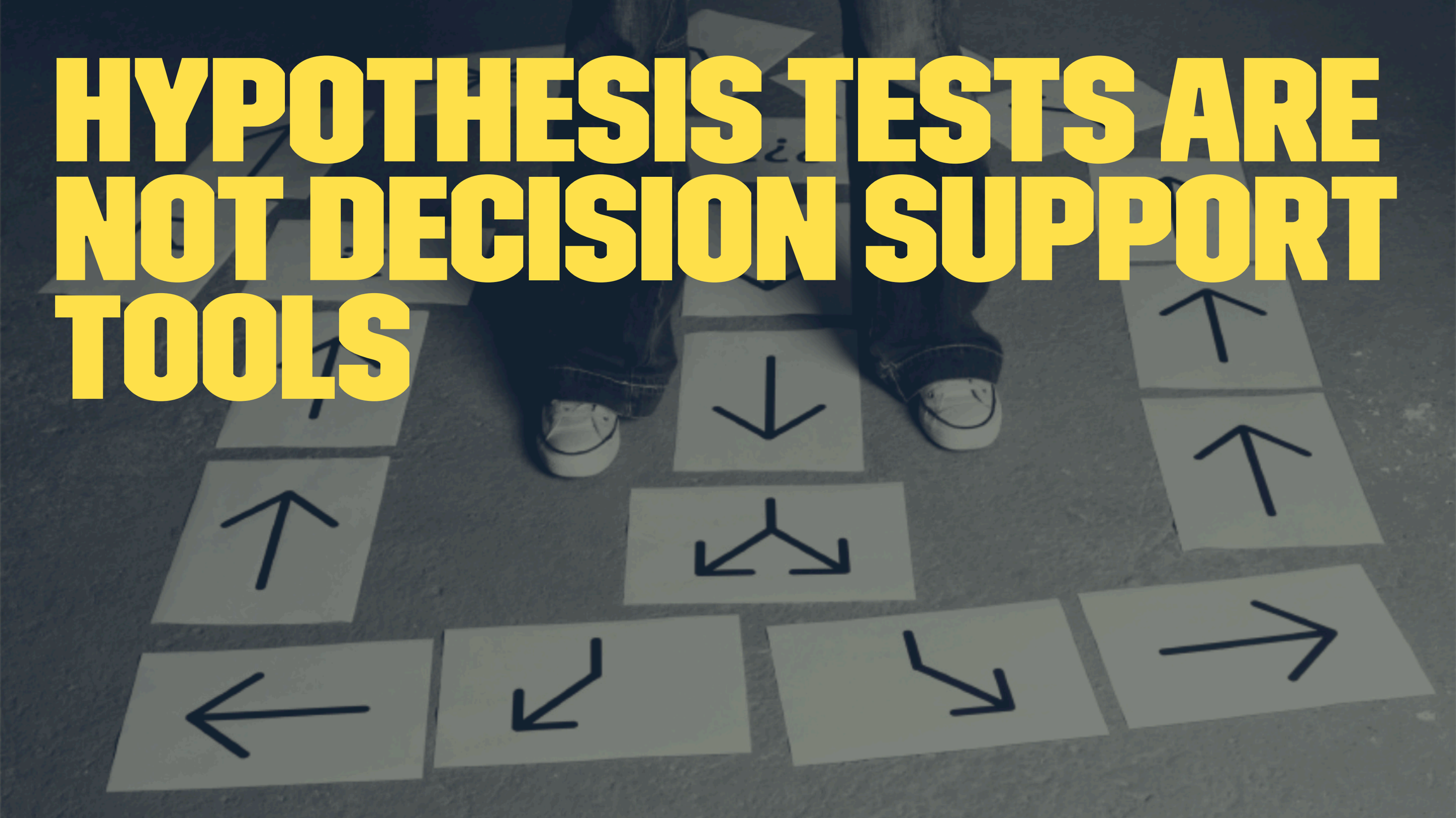


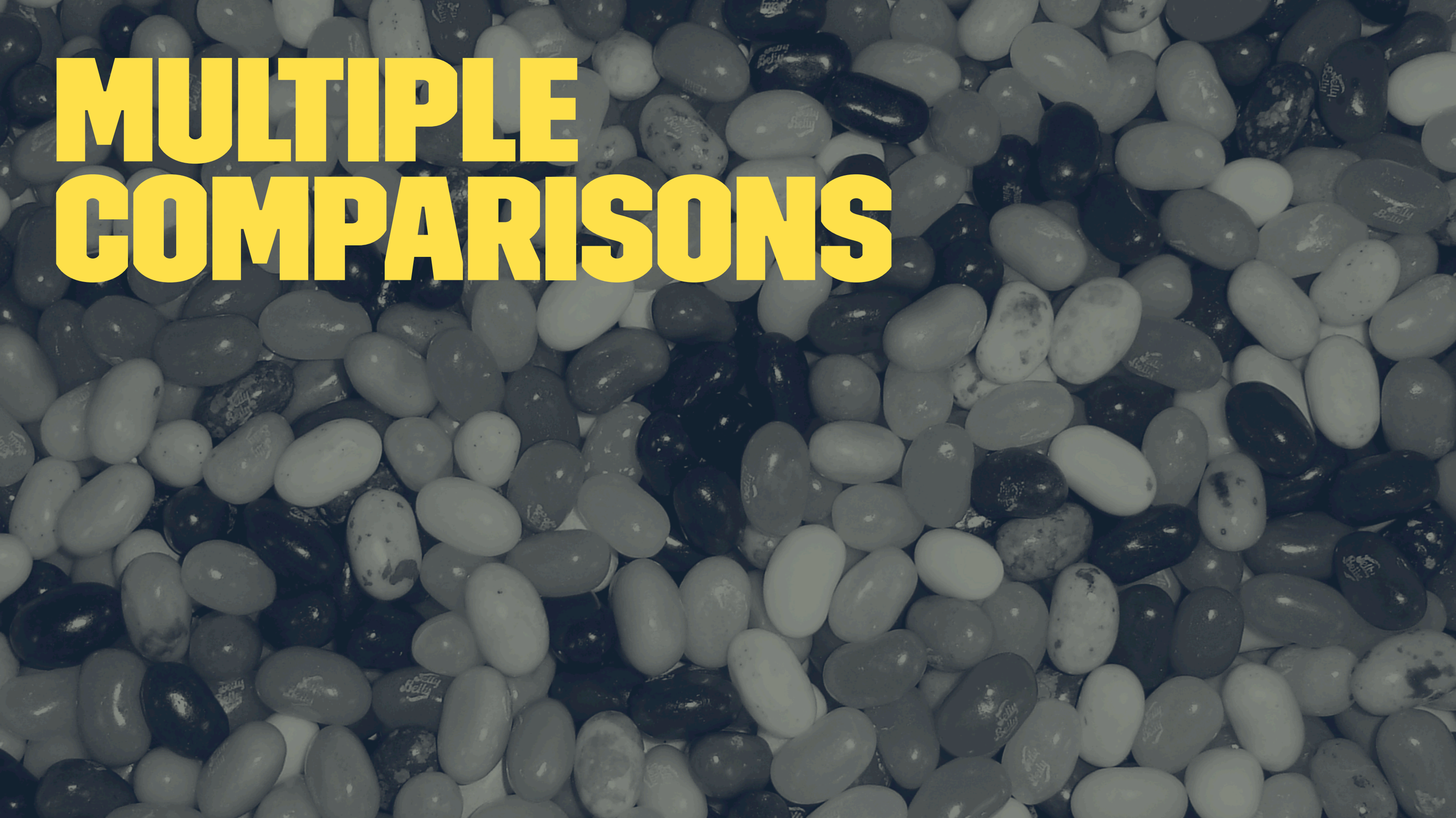
H_o: THE DENSITY OF LARGE TREES IN LOGGED AND UNLOGGED FOREST STANDS WERE EQUAL

$$\theta_{unlogged} = \theta_{logged}$$

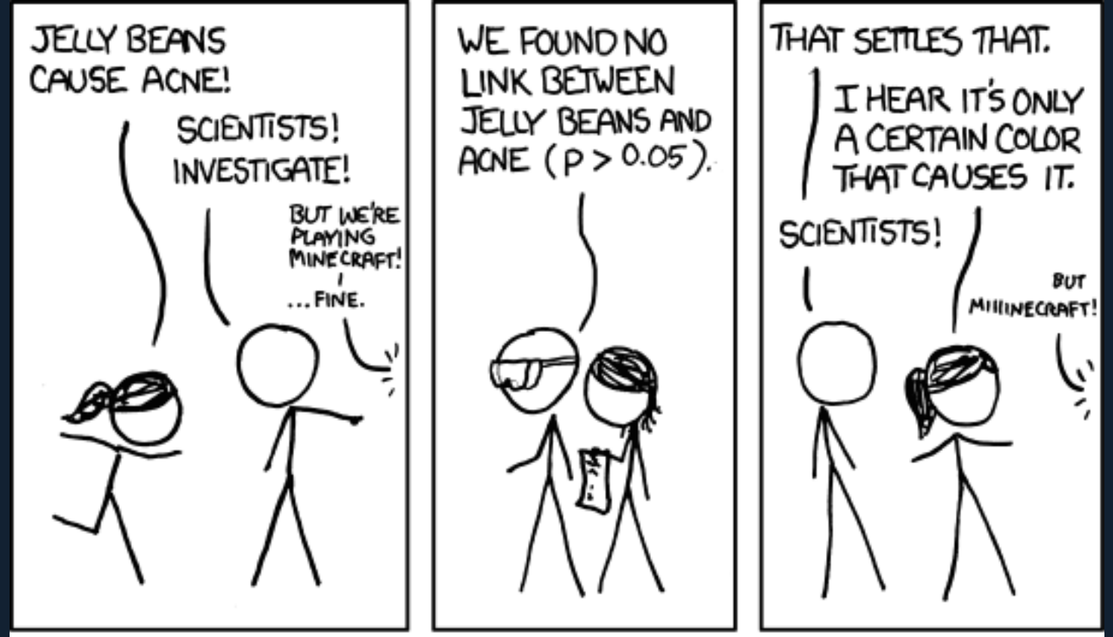

STATISTICAL STRAW MAN
$\mu_0 = \mu_1$

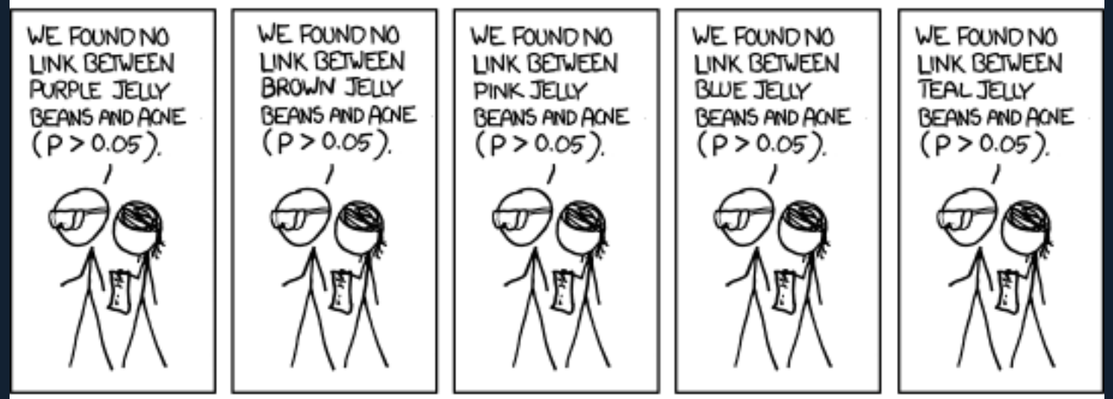

STATISTICAL HYPOTHESES ARE
NOT INTERESTING

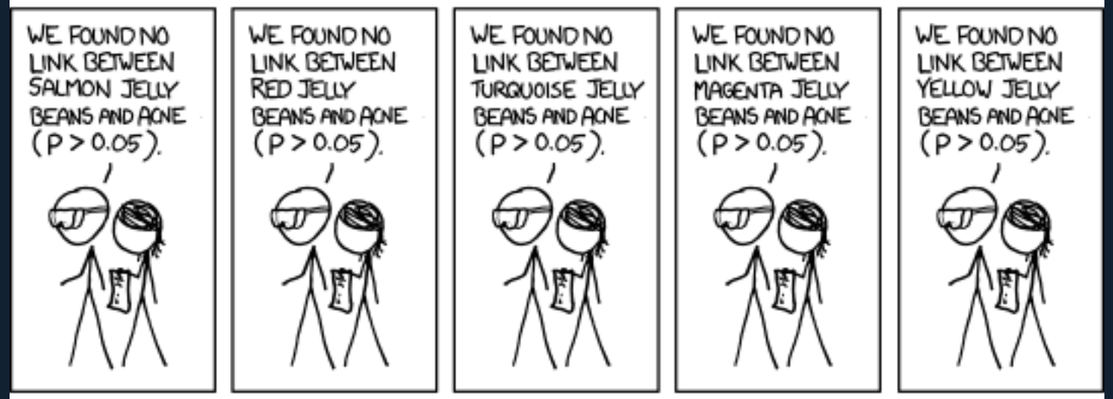

HYPOTHESIS TESTS ARE NOT DECISION SUPPORT TOOLS

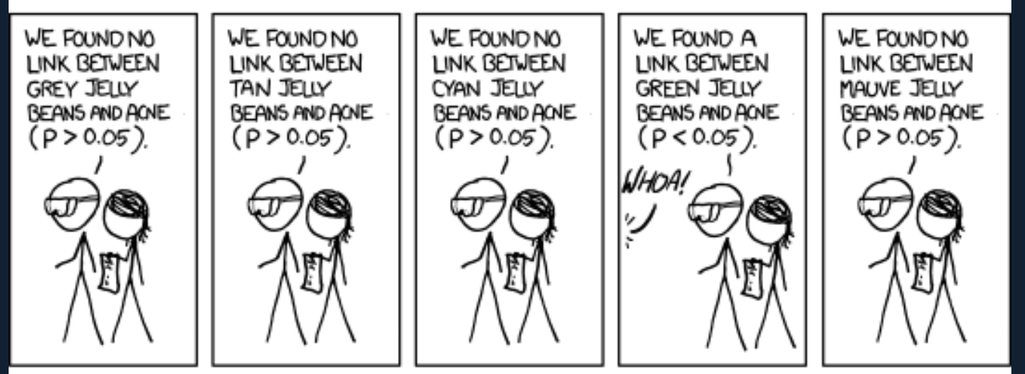


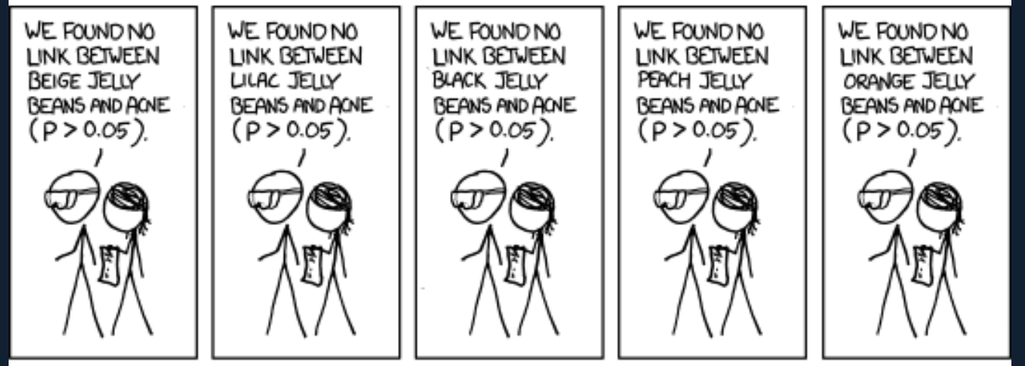
MULTIPLE COMPARISONS

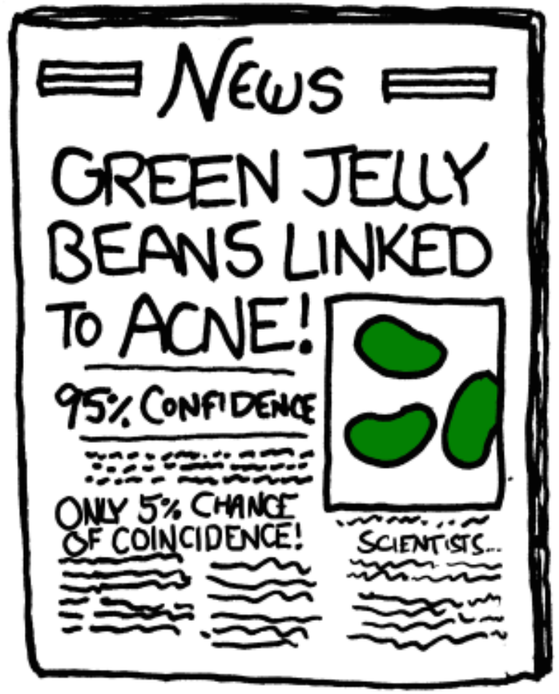

JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIINECRAFT!


WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).


WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).


WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).


WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).

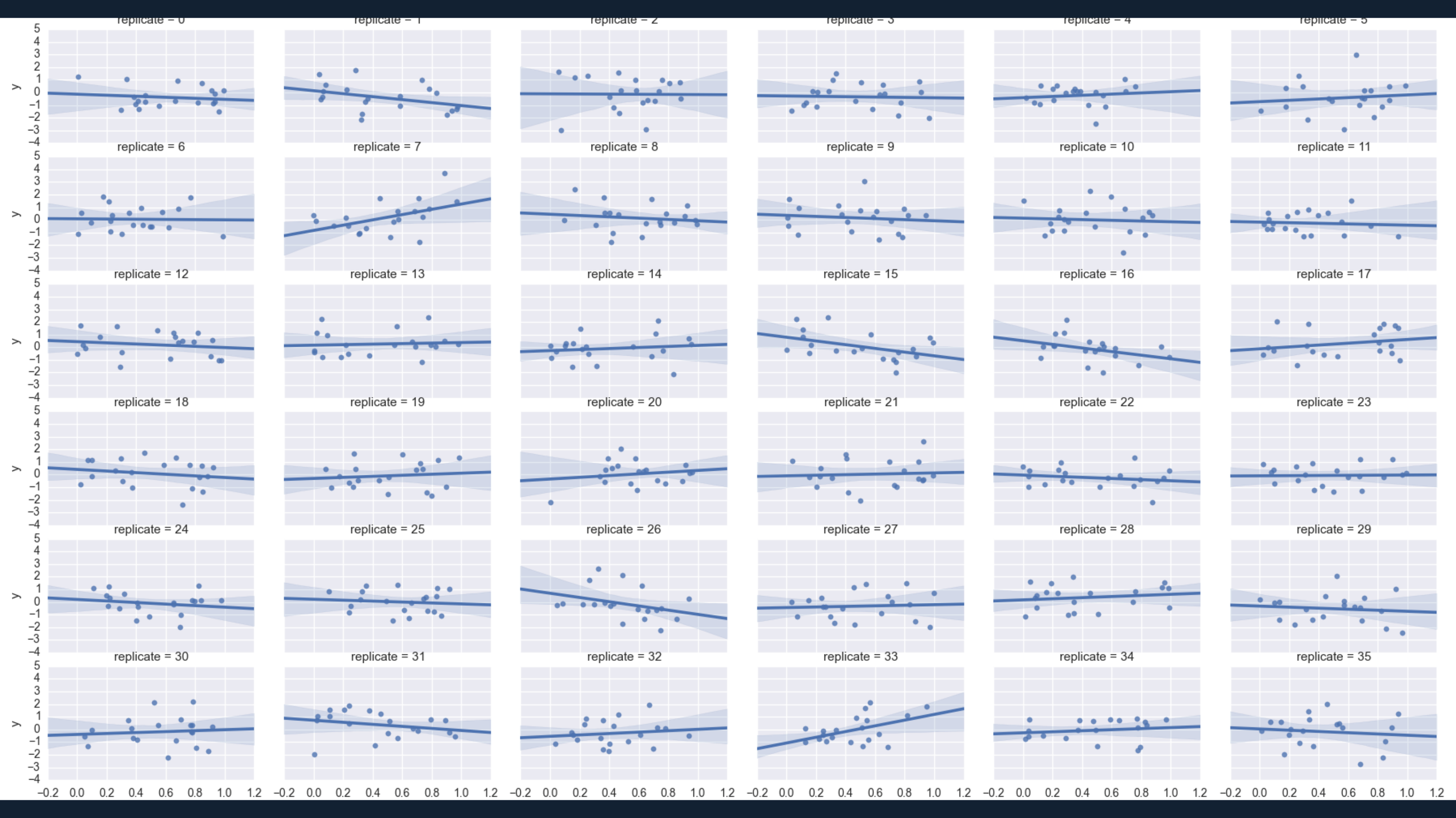

NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

FAMILY-WISE ERROR RATE

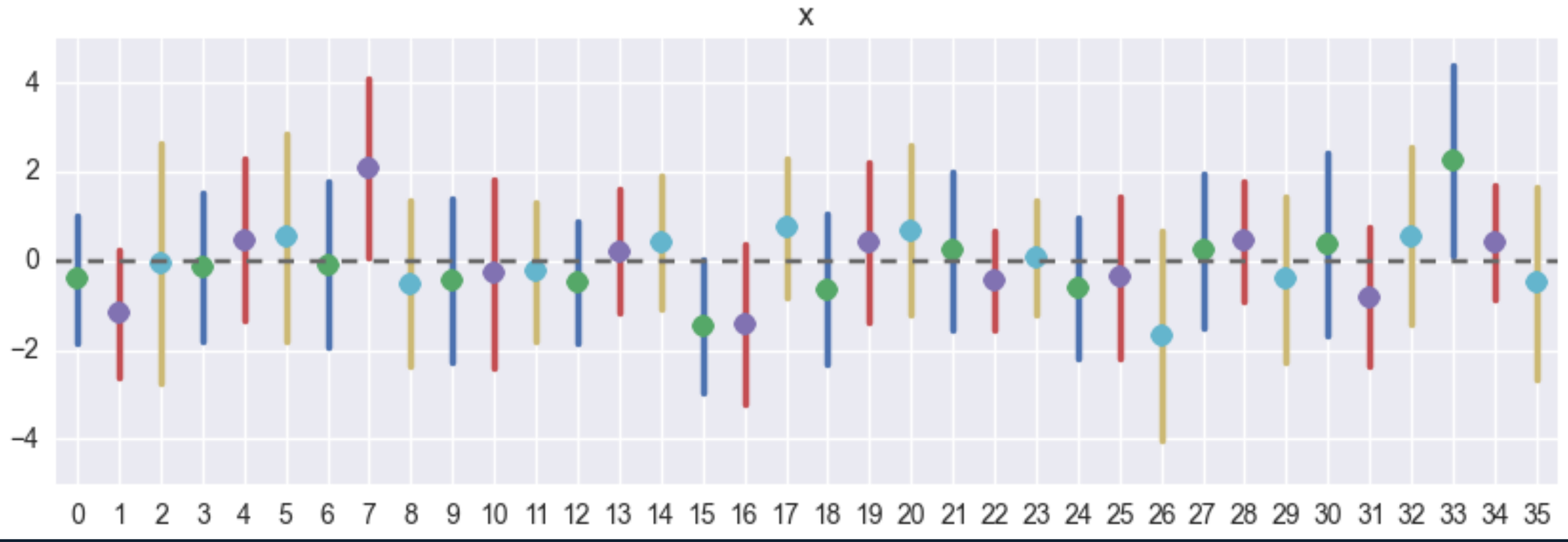
$$\alpha_{fw} \equiv 1 - (1 - \alpha)^n$$

```
>>> 1. - (1. - 0.05) ** 20
0.6415140775914581
```

```
import seaborn as sb
import pandas as pd


n = 20
r = 36

df = pd.concat([pd.DataFrame({'y':np.random.normal(size=n),
                              'x':np.random.random(n),
                              'replicate':[i]*n}) for i in range(r)])
sb.lmplot('x', 'y', df, col='replicate', col_wrap=6)
```


replicate = 0
replicate = 1
replicate = 2
replicate = 3
replicate = 4
replicate = 5
replicate = 6
replicate = 7
replicate = 8
replicate = 9
replicate = 10
replicate = 11
replicate = 12
replicate = 13
replicate = 14
replicate = 15
replicate = 16
replicate = 17
replicate = 18
replicate = 19
replicate = 20
replicate = 21
replicate = 22
replicate = 23
replicate = 24
replicate = 25
replicate = 26
replicate = 27
replicate = 28
replicate = 29
replicate = 30
replicate = 31
replicate = 32
replicate = 33
replicate = 34
replicate = 35
y
5 4 3 2 1 0 −1 −2 −3 −4
−0.2 0.0 0.2 0.4 0.6 0.8 1.0 1.2

STATISTICALLY SIGNIFICANT!



Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R+1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error

Essay

Why Most Published Research Findings Are False

false. The probability that a research claim

relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other

Positive Findings

of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles

true relationships among thousands hypotheses that may Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R+1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error

"DESPITE A LARGE STATISTICAL LITERATURE FOR MULTIPLE TESTING CORRECTIONS, USUALLY IT IS IMPOSSIBLE TO DECIPHER HOW MUCH DATA DREDGING BY THE REPORTING AUTHORS OR OTHER RESEARCH TEAMS HAS PRECEDED A REPORTED RESEARCH FINDING."

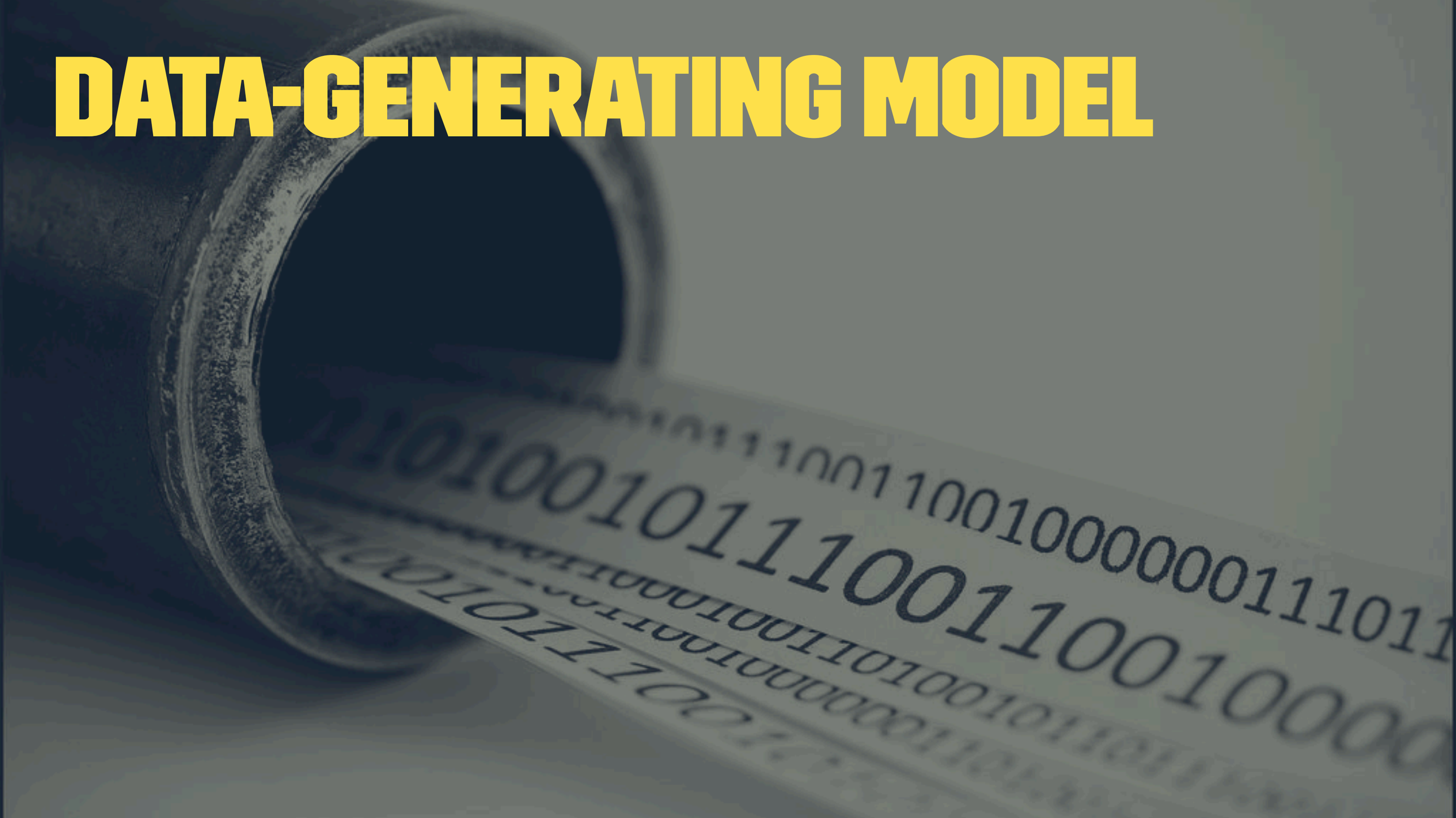
WHAT'S THE ALTERNATIVE?

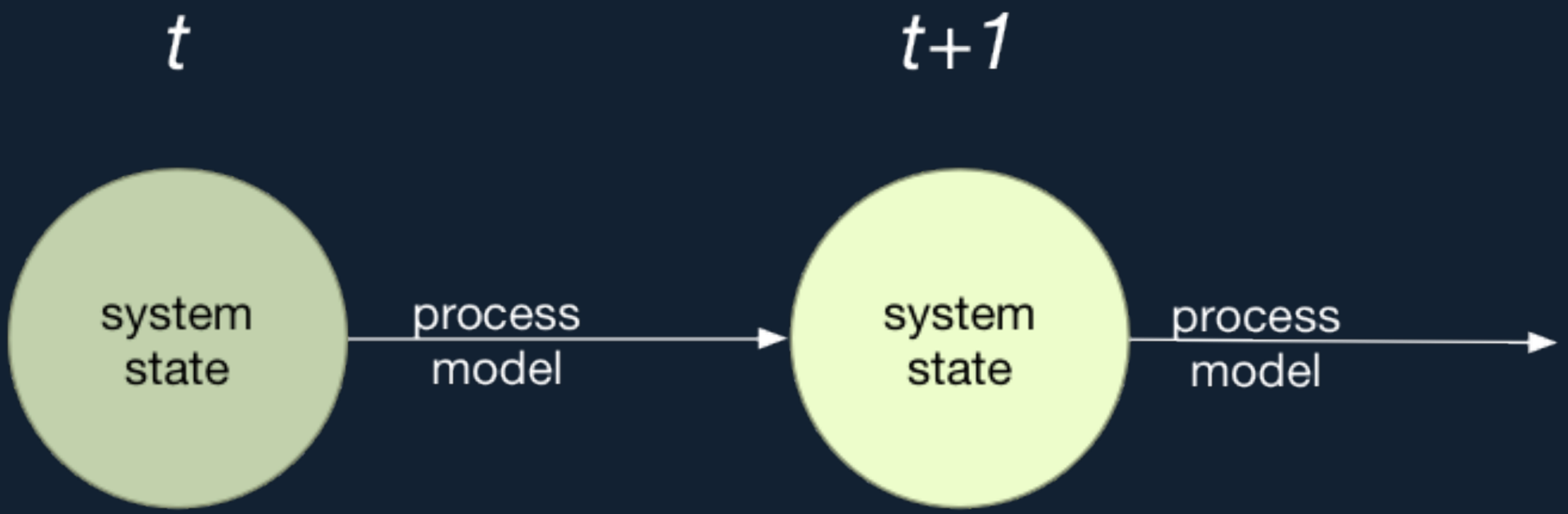
BUILD
MODELS
AND USE THEM TO
ESTIMATE
THINGS WE CARE
ABOUT

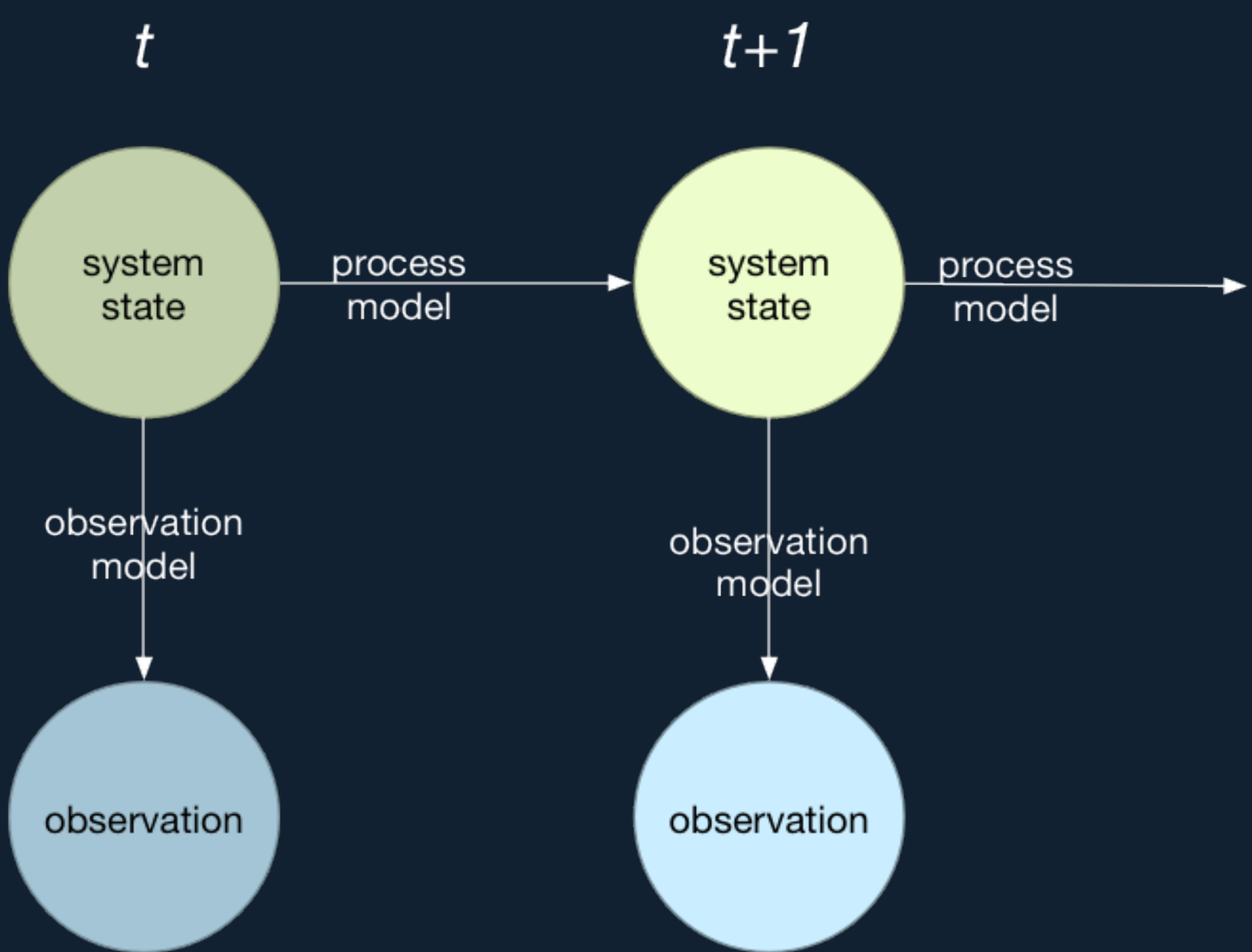
EFFECT SIZE ESTIMATION

$$d = \theta_1 - \theta_2$$

DATA-GENERATING MODEL




t
t+1
system state
process model
system state
process model


t
t+1
system state
process model
system state
process model
observation model
observation model
observation
observation

FLORIDA MANATEE

Trichechus manatus

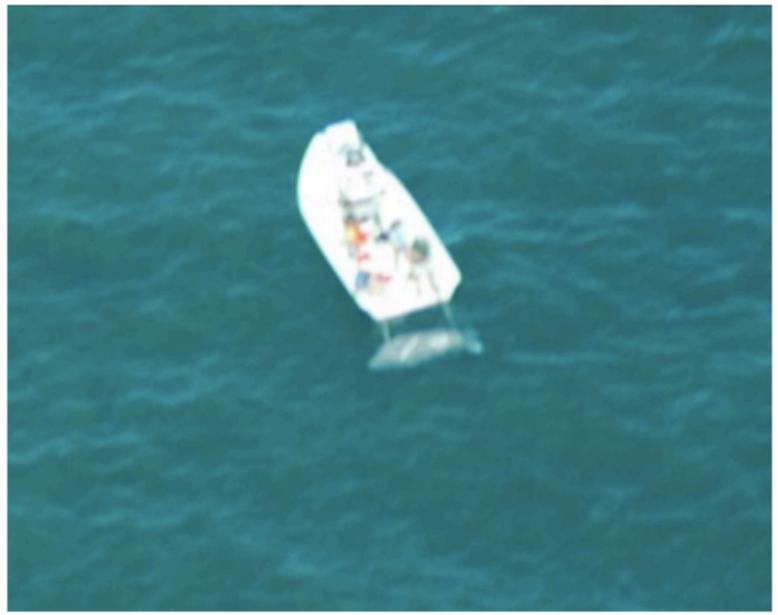

OCCUPIED?

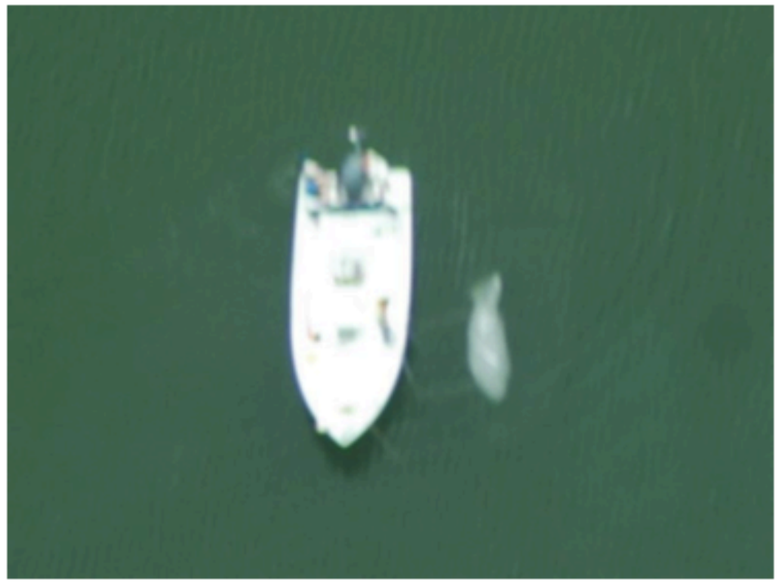

OCCUPIED?
AVAILABLE?

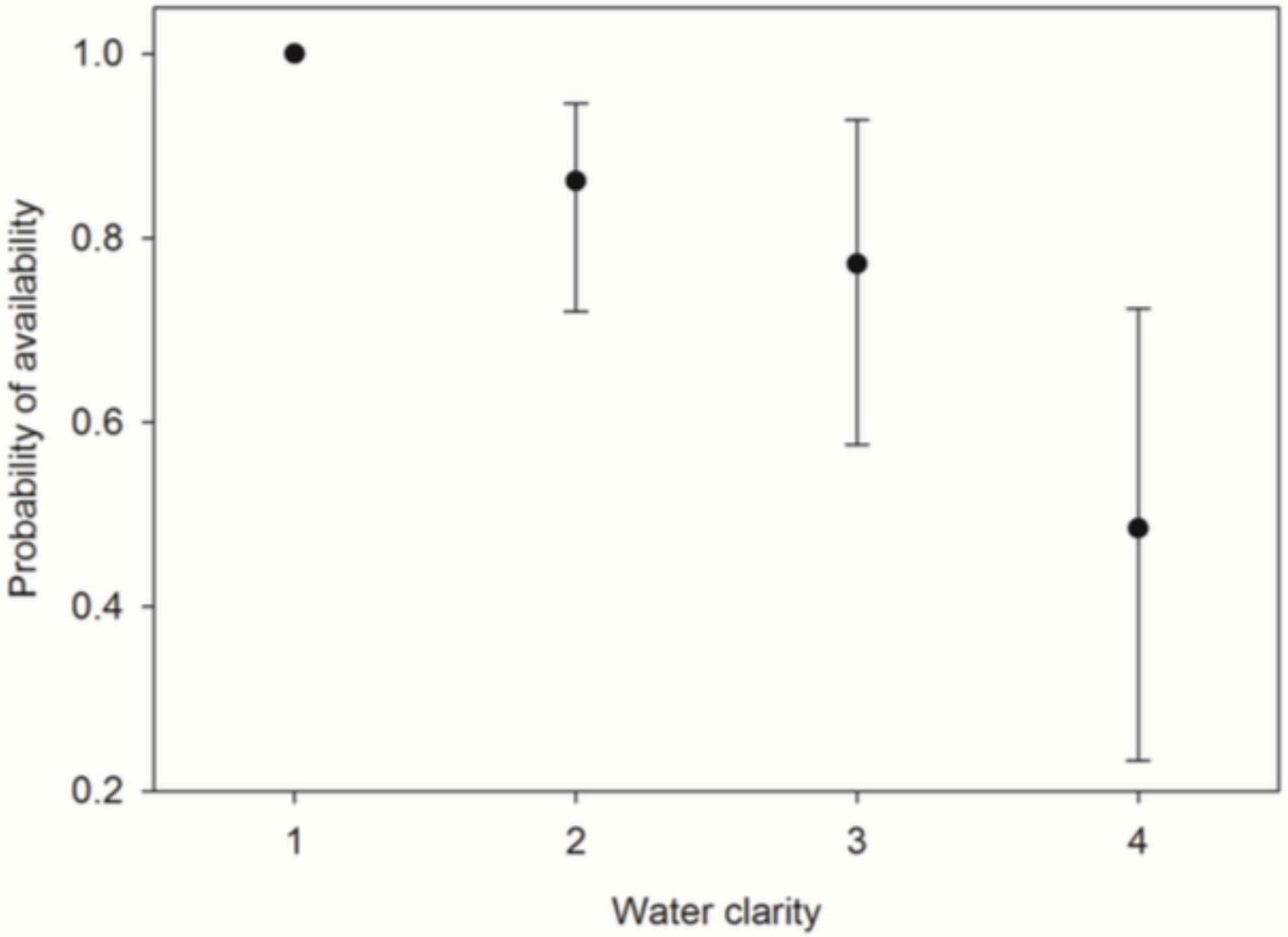

OCCUPIED?
AVAILABLE?
SEEN?

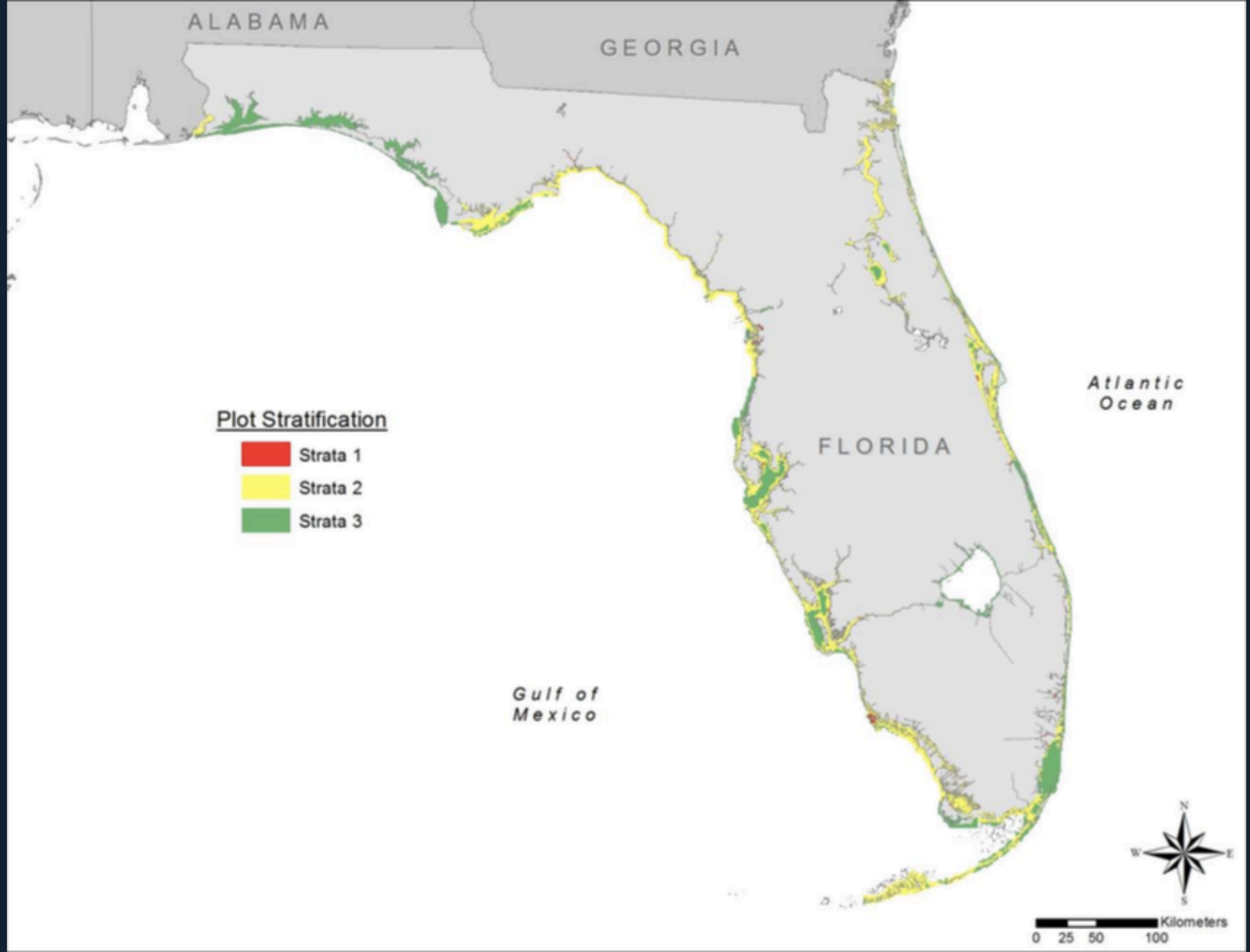

PTT 14610
MANATEE

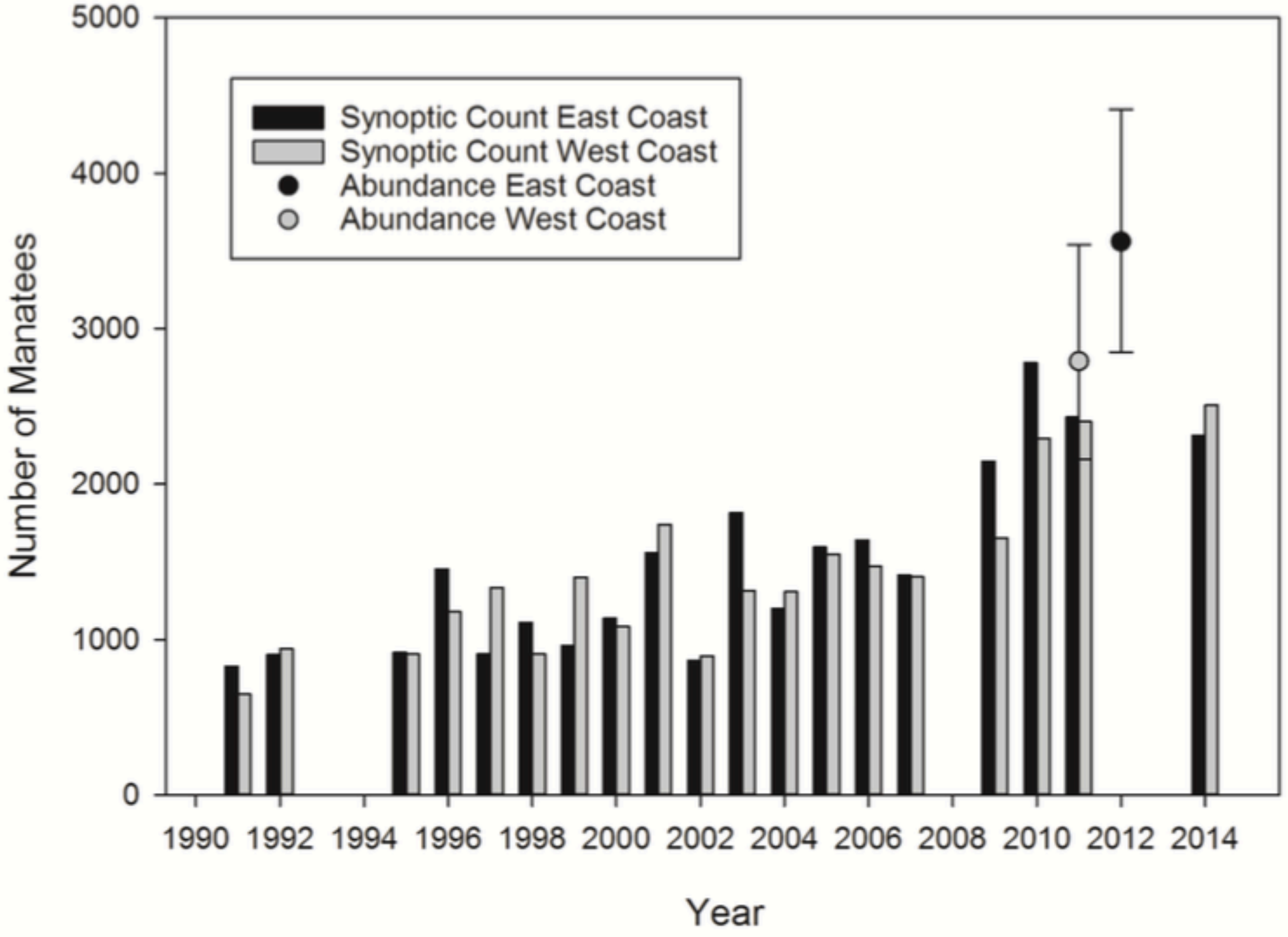
ESTIMATING VISIBILITY






1.0
0.8
0.6
0.4
0.2
Probability of availability
1
2
3
4
Water clarity


ALABAMA
GEORGIA
FLORIDA
Atlantic Ocean
Gulf of Mexico
Plot Stratification
Strata 1
Strata 2
Strata 3
N
W
E
S
Kilometers
0 25 50 100


Synoptic Count East Coast
Synoptic Count West Coast
Abundance East Coast
Abundance West Coast
Number of Manatees
5000
4000
3000
2000
1000
0
1990
1992
1994
1996
1998
2000
2002
2004
2006
2008
2010
2012
2014
Year

google.org Flu Trends

Google.org home

Dengue Trends

Flu Trends

Home

United States

Tennessee

Download data

How does this work?

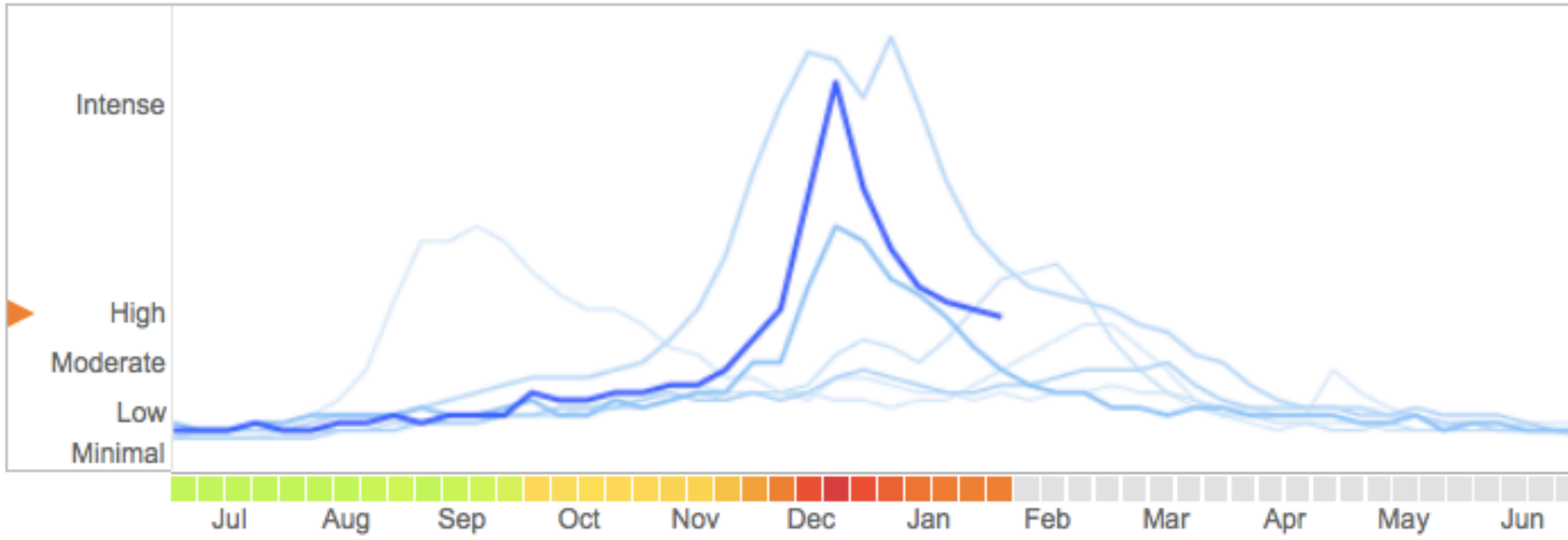
FAQ

Explore flu trends - United States

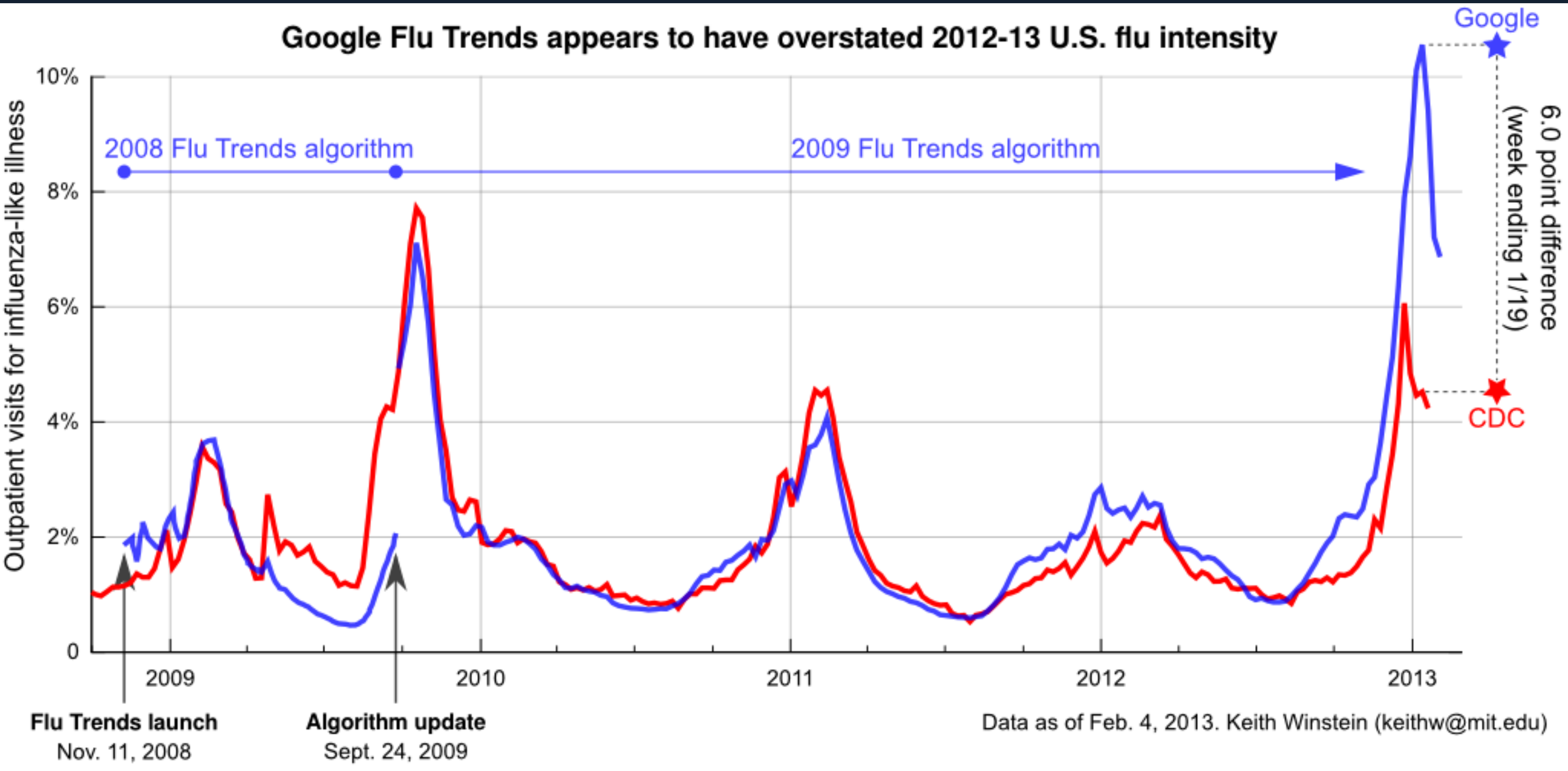
We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. Learn more »

United States > Tennessee

2014-2015 Past years



States | Cities (Experimental)



Sources: http://www.google.org/flutrends/us, CDC ILInet data from http://gis.cdc.gov/grasp/fluview/fluportaldashboard.html,
Cook et al. (2011) Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic.

OPEN ACCESS Freely available online

PLOS | ONE

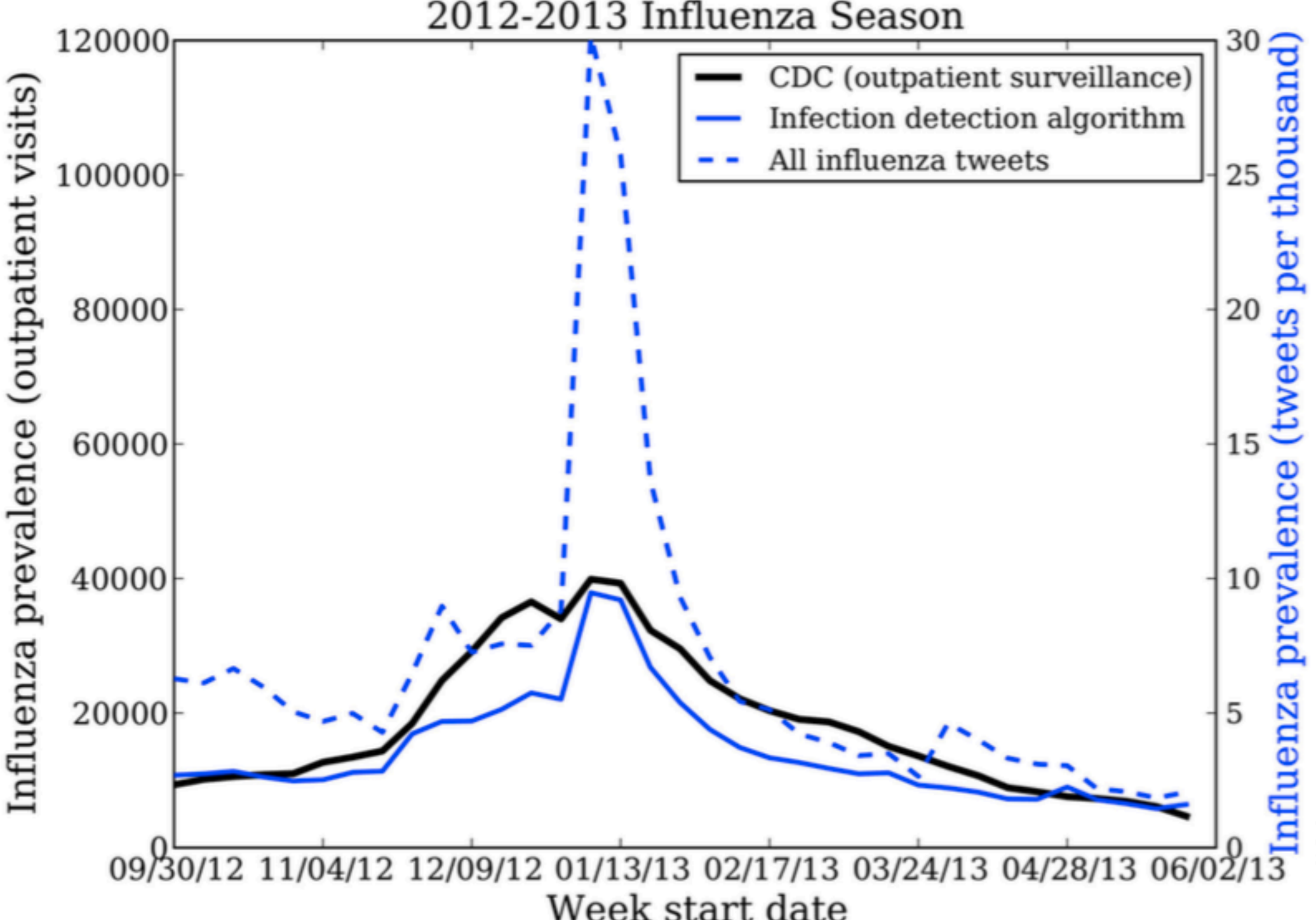
National and Local Influenza Surveillance through Twitter: An Analysis of the 2012-2013 Influenza Epidemic

David A. Broniatowski[1,2], Michael J. Paul[3*], Mark Dredze[4]

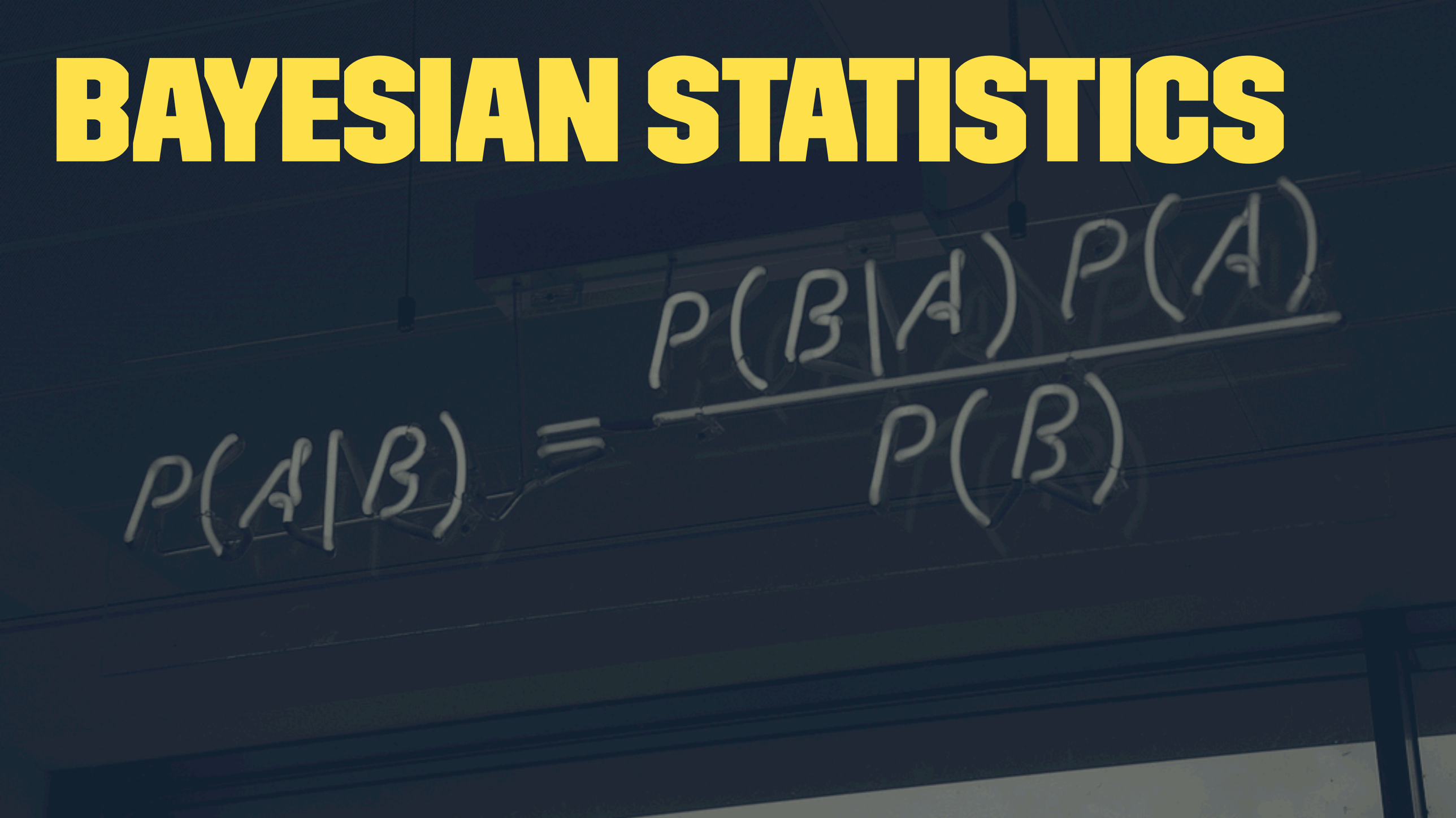
1 Department of Engineering Management and Systems Engineering, The George Washington University, Washington, District of Columbia, United States of America, **2** Center for Advanced Modeling in the Social, Behavioral, and Health Sciences, Department of Emergency Medicine, School of Medicine, Johns Hopkins University, Baltimore, Maryland, United States of America, **3** Department of Computer Science and Center for Language and Speech Processing, Johns Hopkins University, Baltimore, Maryland, United States of America, **4** Human Language Technology Center of Excellence and Department of Computer Science, Johns Hopkins University, Baltimore, Maryland, United States of America

Abstract

Social media have been proposed as a data source for influenza surveillance because they have the potential to offer real-time access to millions of short, geographically localized messages containing information regarding personal well-being. However, accuracy of social media surveillance systems declines with media attention because media attention increases "chatter" – messages that are about influenza but that do not pertain to an actual infection – masking signs of true influenza prevalence. This paper summarizes our recently developed influenza infection detection algorithm that automatically distinguishes relevant tweets from other chatter, and we describe our current influenza surveillance system which was actively deployed during the full 2012-2013 influenza season. Our objective was to analyze the performance of this system during the most recent 2012–2013 influenza season and to analyze the performance at multiple levels of geographic granularity, unlike past studies that focused on national or regional surveillance. Our system's influenza prevalence estimates were strongly correlated with surveillance data from the Centers for Disease Control and Prevention for the United States ($r = 0.93$, $p < 0.001$) as well as surveillance data from the Department of Health and Mental Hygiene of New York City ($r = 0.88$, $p < 0.001$). Our system detected the


2012-2013 Influenza Season
CDC (outpatient surveillance)
Infection detection algorithm
All influenza tweets
Influenza prevalence (outpatient visits)
Influenza prevalence (tweets per thousand)
Week start date
0
20000
40000
60000
80000
100000
120000
0
5
10
15
20
25
30
09/30/12
11/04/12
12/09/12
01/13/13
02/17/13
03/24/13
04/28/13
06/02/13

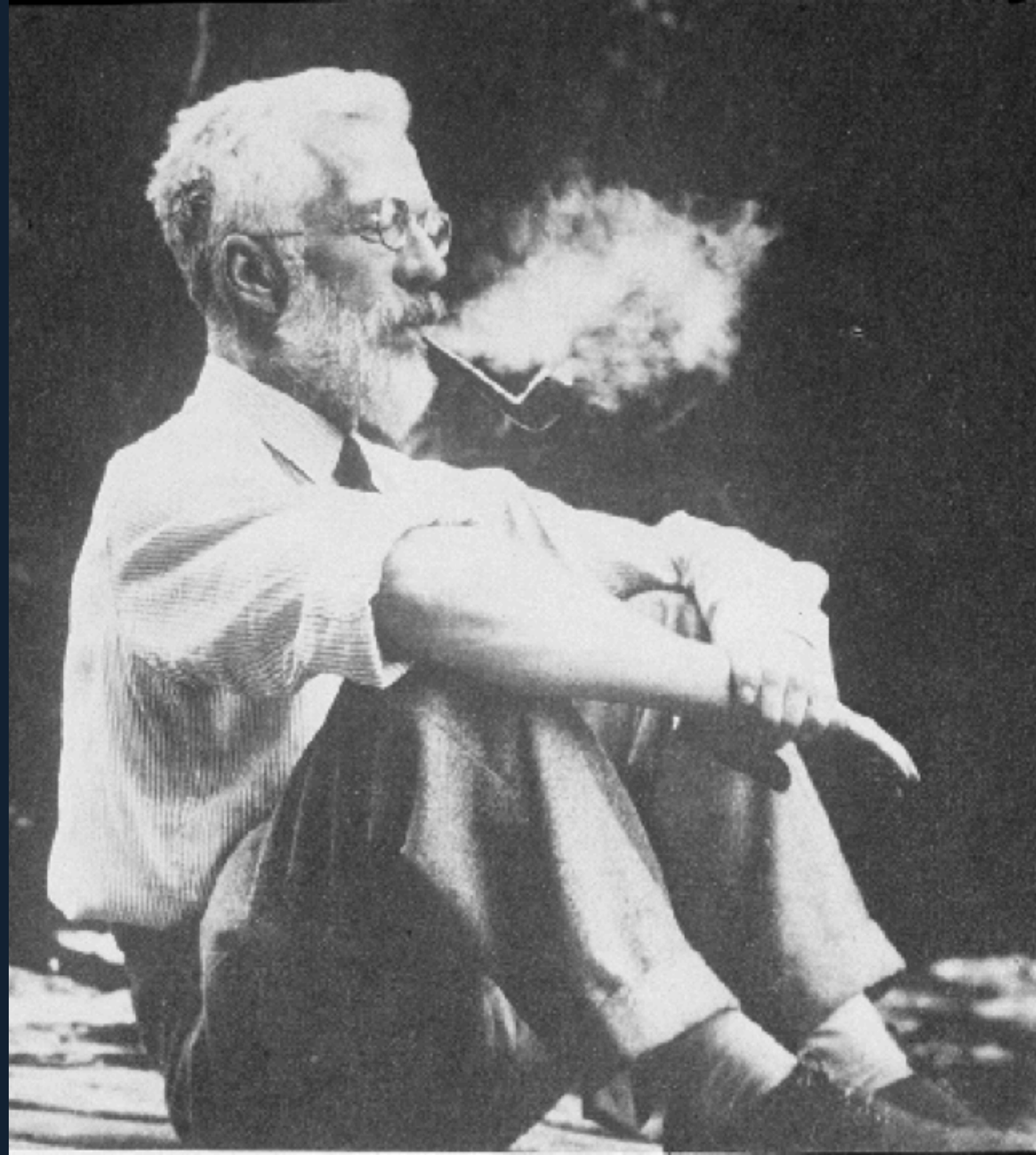
BAYESIAN STATISTICS



Model

$$Pr(y|\theta)$$

Data



Model

$$Pr(\theta|y)$$

Data

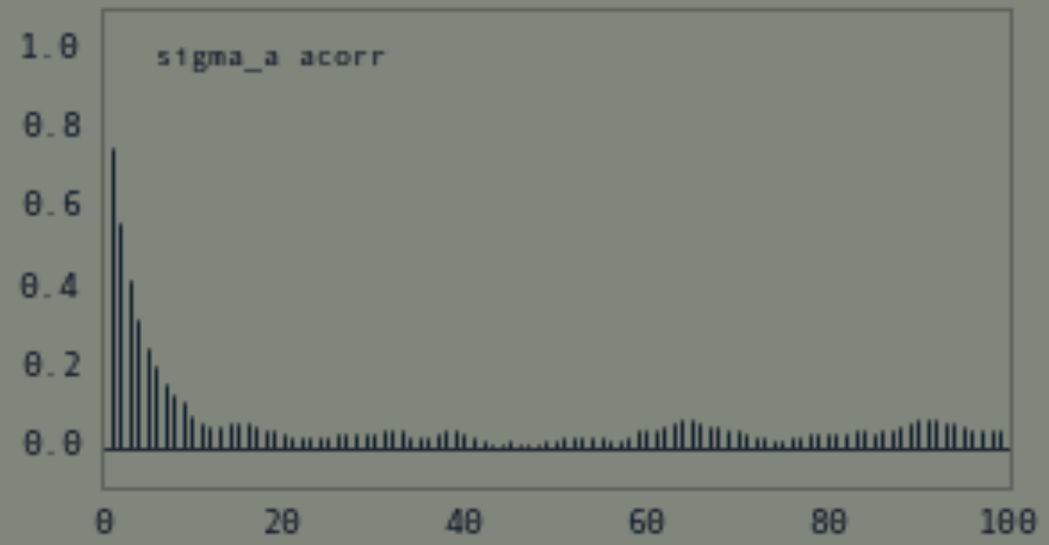
Rev. T. Bayes

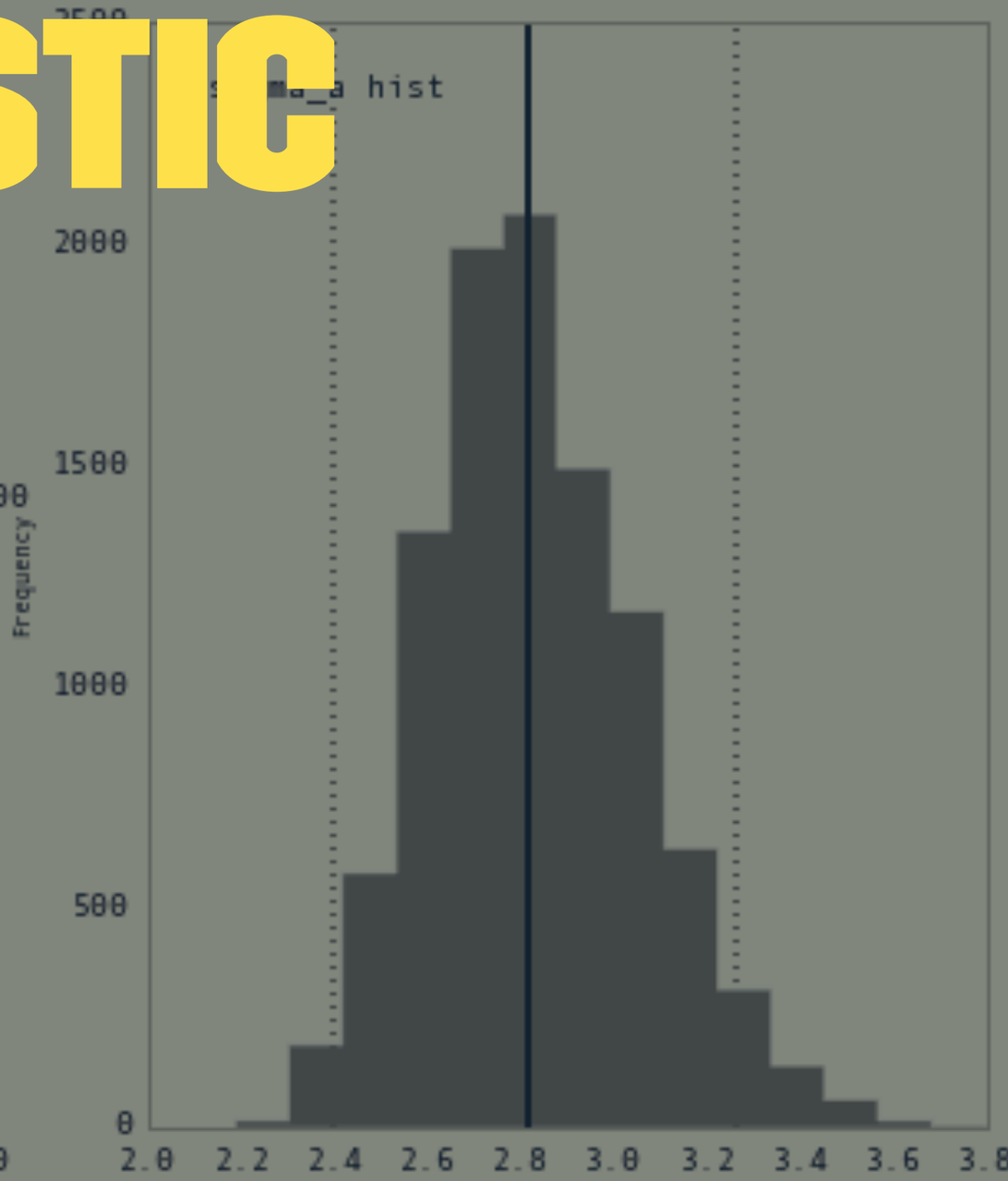
BAYES' FORMULA

Posterior Likelihood Prior

$$Pr(\theta|y) = \frac{Pr(y|\theta)\,Pr(\theta)}{Pr(y)}$$

PROBABILISTIC MODELING



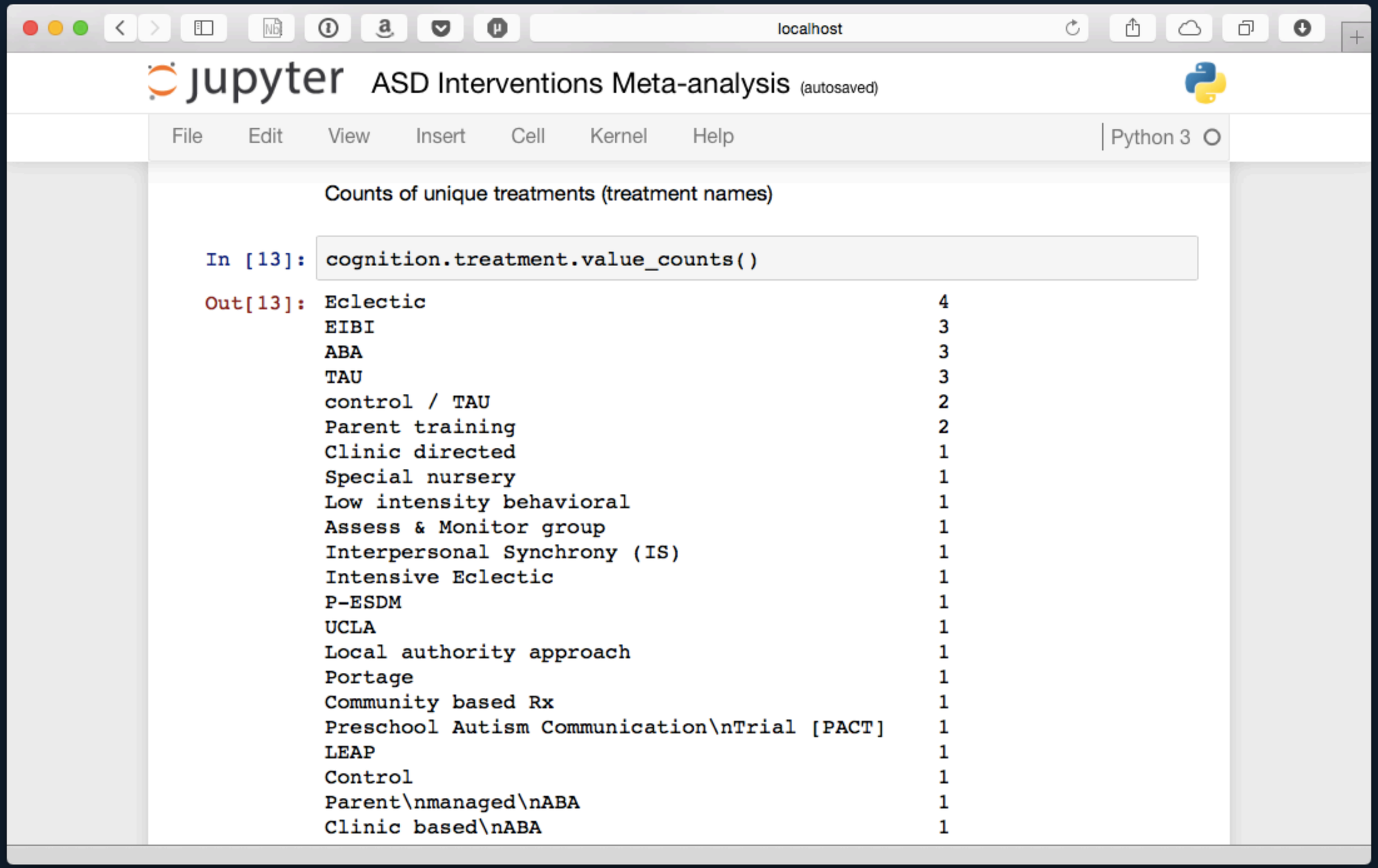


EVIDENCE-BASED MEDICINE

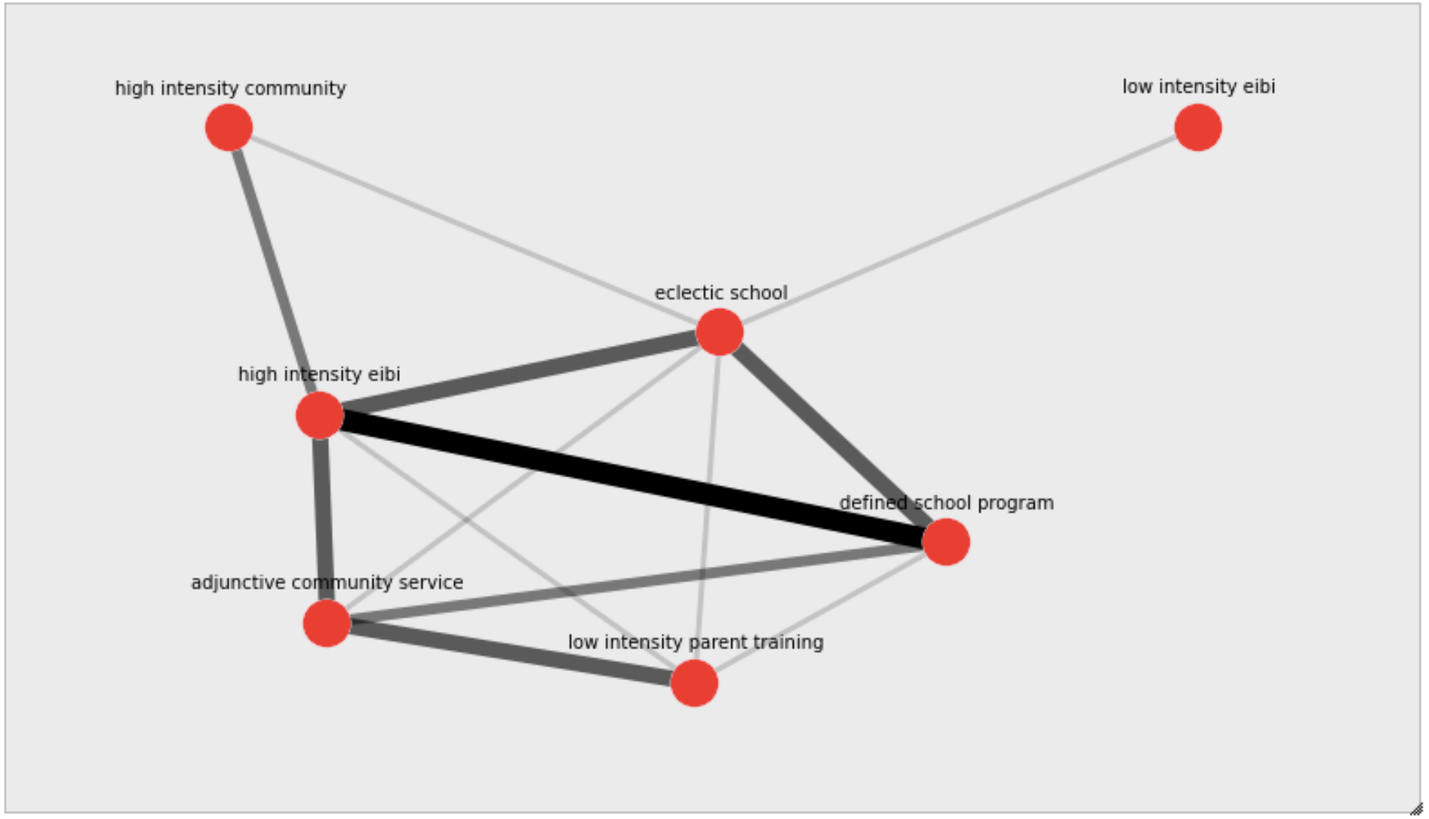
ASD INTERVENTIONS RESEARCH

19 INDEPENDENT STUDIES

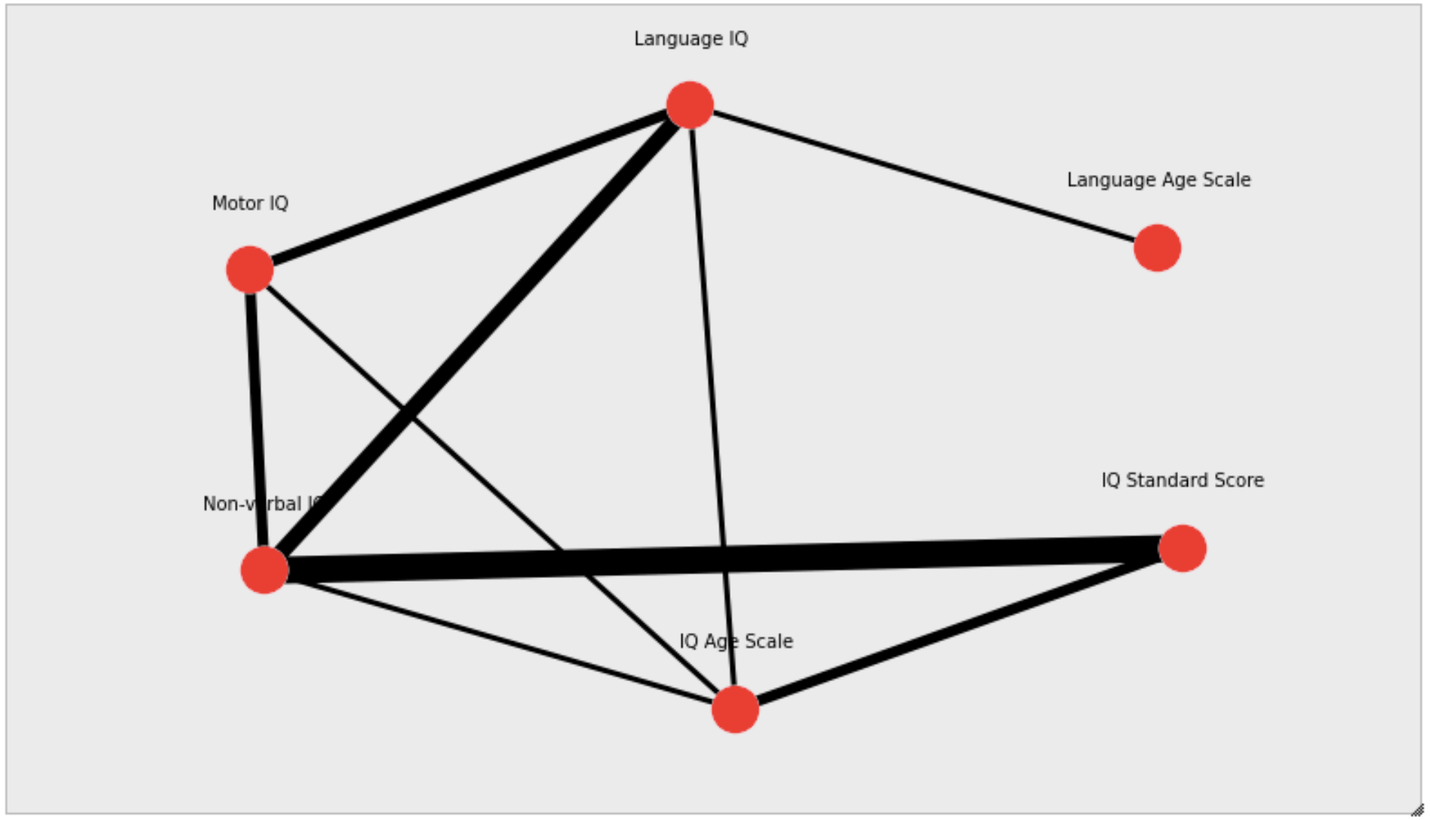
27 DIFFERENT INTERVENTIONS

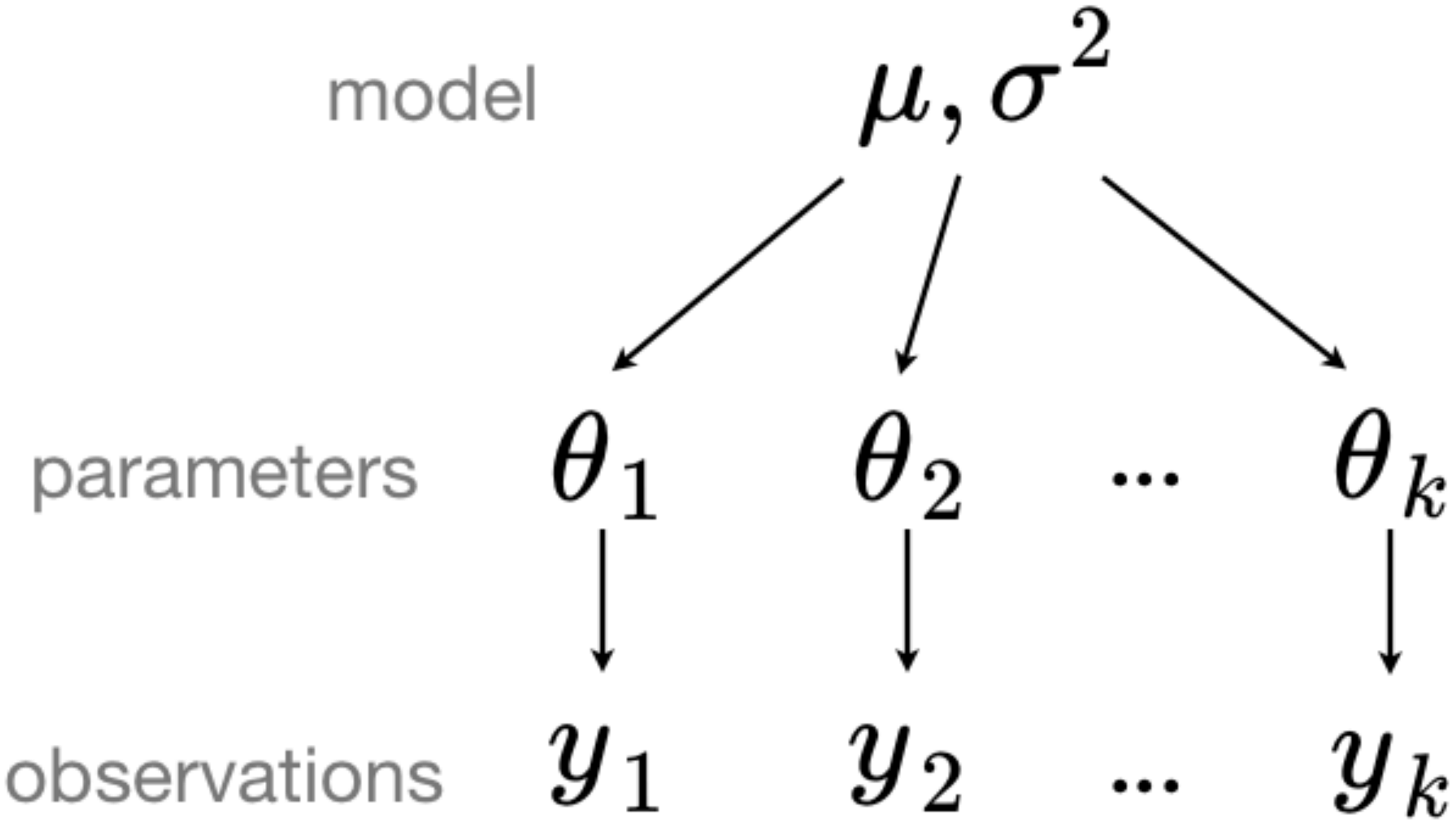

localhost
jupyter ASD Interventions Meta-analysis (autosaved)
File Edit View Insert Cell Kernel Help
Python 3
Counts of unique treatments (treatment names)
In [13]: cognition.treatment.value_counts()
Out[13]: Eclectic 4
EIBI 3
ABA 3
TAU 3
control / TAU 2
Parent training 2
Clinic directed 1
Special nursery 1
Low intensity behavioral 1
Assess & Monitor group 1
Interpersonal Synchrony (IS) 1
Intensive Eclectic 1
P-ESDM 1
UCLA 1
Local authority approach 1
Portage 1
Community based Rx 1
Preschool Autism Communication\nTrial [PACT] 1
LEAP 1
Control 1
Parent\nmanaged\nABA 1
Clinic based\nABA 1

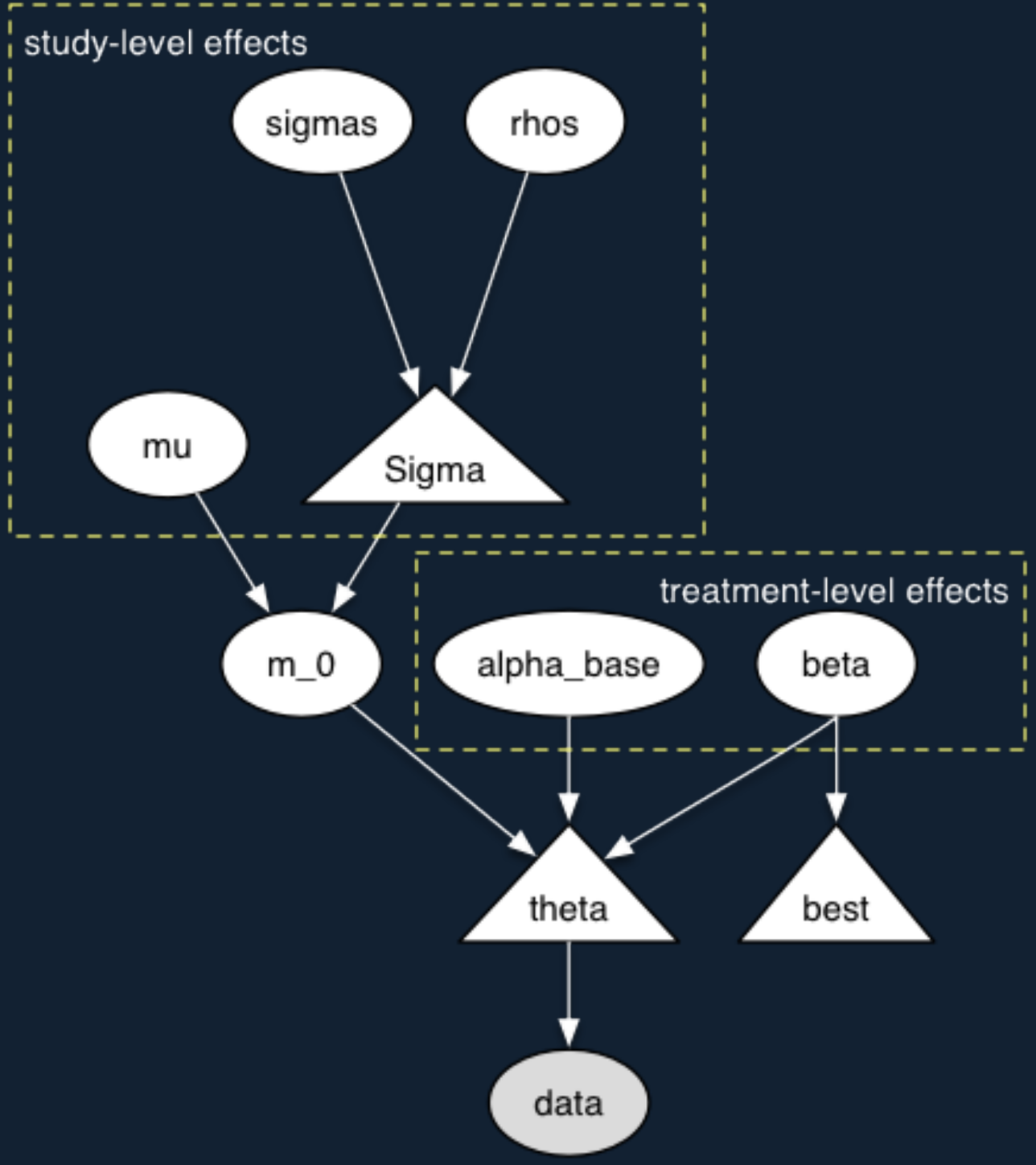
```
    nx.draw_networkx_edges(G, pos, edgelist=edge_1, width=(1+1)*3, alpha=alpha)
pos_shifted = {p: (pos[p][0], pos[p][1]+10) for p in pos}
nx.draw_networkx_labels(G, pos_shifted, font_size=10, font_family='sans-serif')
plt.gca().xaxis.set_major_locator(plt.NullLocator())
plt.gca().yaxis.set_major_locator(plt.NullLocator());
```



```
pos_shifted2 = {p: (pos2[p][0], pos2[p][1]+10) for p in pos2}
nx.draw_networkx_labels(G2, pos_shifted2, font_size=10, font_family='sans-serif')
plt.gca().xaxis.set_major_locator(plt.NullLocator())
plt.gca().yaxis.set_major_locator(plt.NullLocator());
```




model
μ, σ^2
parameters
θ_1
θ_2
...
θ_k
observations
y_1
y_2
...
y_k

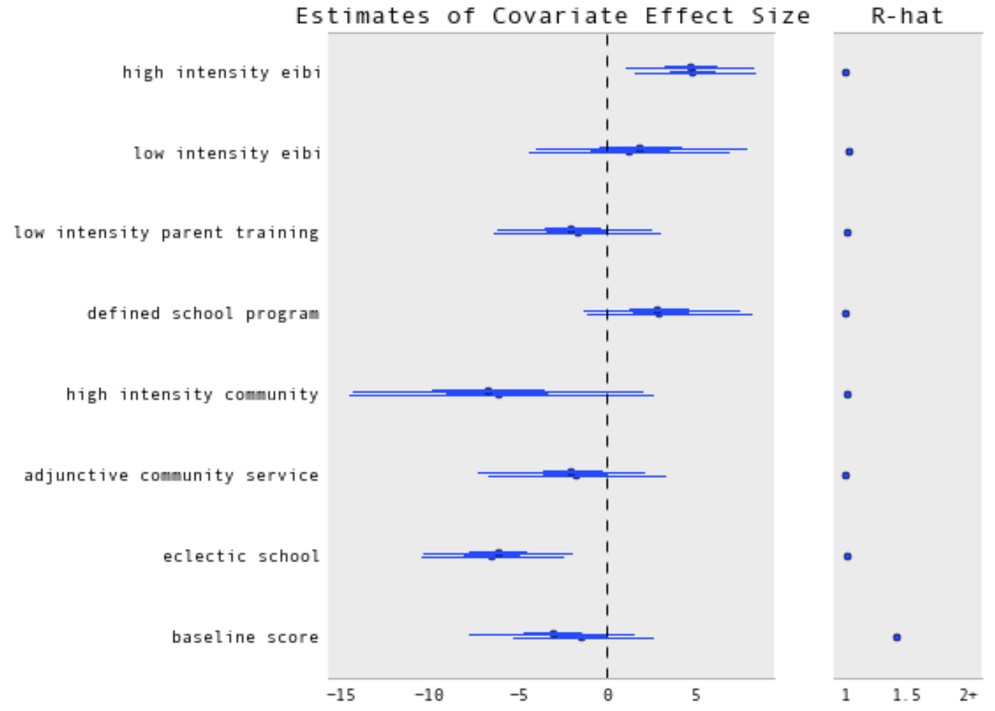

study-level effects
sigmas
rhos
mu
Sigma
treatment-level effects
m_0
alpha_base
beta
theta
best
data

localhost

jupyter ASD Interventions Meta-analysis Last Checkpoint: 11/12/2014 (unsaved changes)

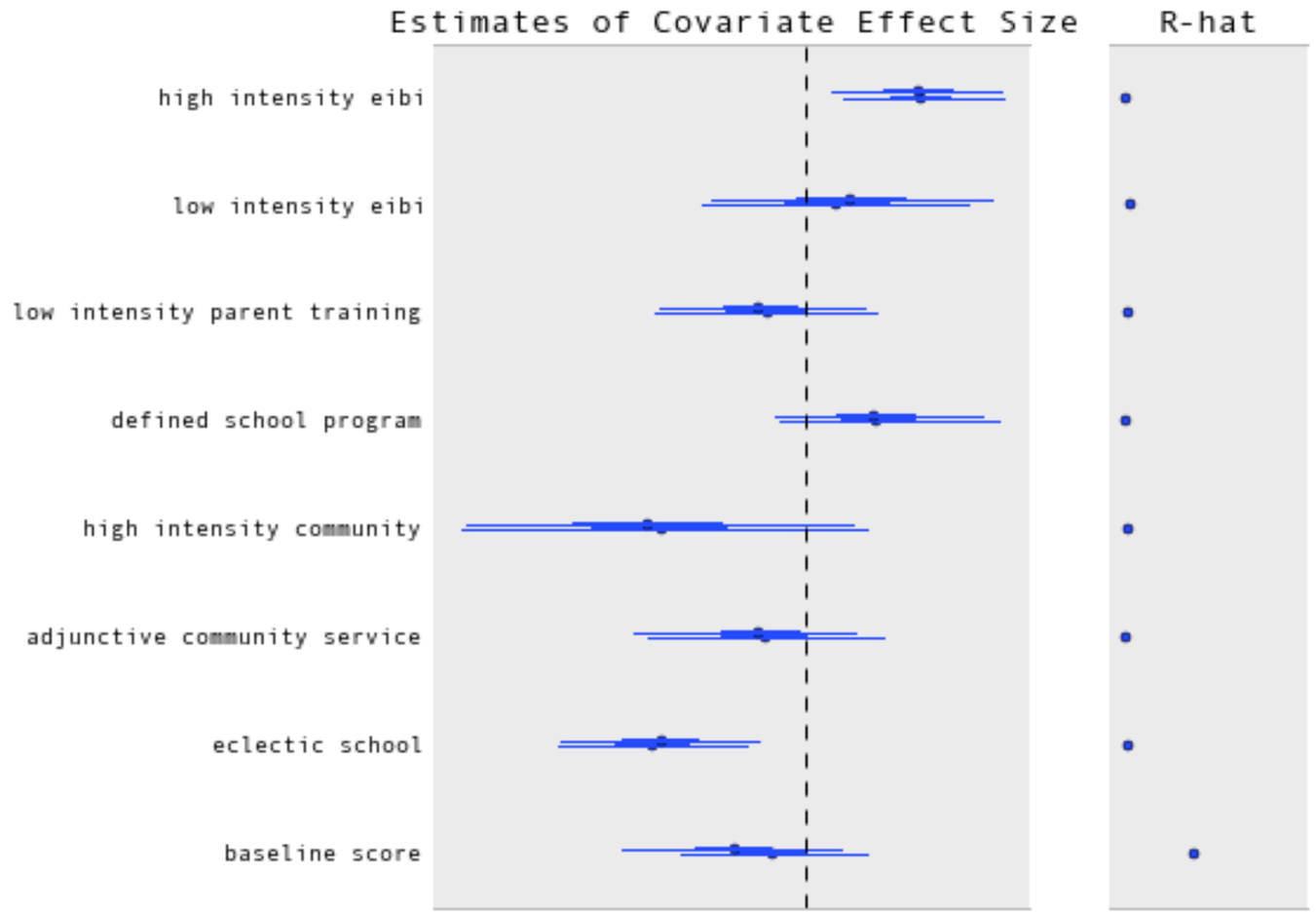
File Edit View Insert Cell Kernel Help Python 3

In [85]:
```
plt.figure(figsize=(10,6))
pm.Matplot.summary_plot([M.beta, M.alpha_base], custom_labels=['high intensity eibi',
                                'low intensity eibi',
                                'low intensity parent training',
                                'defined school program',
                                'high intensity community',
                                'adjunctive community service',
                                'eclectic school',
                                'baseline score'], main='Estimates of Covariate Effect Size')
```



File Edit View Insert Cell Kernel Help Python 3

In [85]:
```
plt.figure(figsize=(10,6))
pm.Matplot.summary_plot([M.beta, M.alpha_base], custom_labels=['high intensity eibi',
                                    'low intensity eibi',
                                    'low intensity parent training',
                                    'defined school program',
                                    'high intensity community',
                                    'adjunctive community service',
                                    'eclectic school',
                                    'baseline score'], main='Estimates of Covariate Effect Size')
```



September 11, 2012, 12:18 PM | 109 Comments

Sept. 10: Will Obama's Bounce Hold?

By NATE SILVER

Has President Obama's convention bounce reached its peak? On Monday, his position declined slightly in our forecast for the first time since Aug. 27.

To be sure, Mr. Obama had a fairly strong day of polling on Monday relative to the long-term baseline. But the data was a little bit more equivocal than in polls released over the weekend — which may suggest, at least, that he will make few further gains in the polls.

Mr. Obama gained a single point in the Rasmussen Reports and Ipsos tracking polls on Monday, but held steady in two others from Gallup and the RAND Corporation. (His lead also declined in the Rasmussen Reports tracking poll on Tuesday, which was published after our forecast was updated late Monday night.)

Some of the surveys apart from the tracking polls were erratic and not necessarily so strong for Mr. Obama. He actually declined by several points in one national poll, conducted by TIPP, although a considerable amount of their data was from early last week and did not reflect the key convention speeches.

FiveThirtyEight Forecast

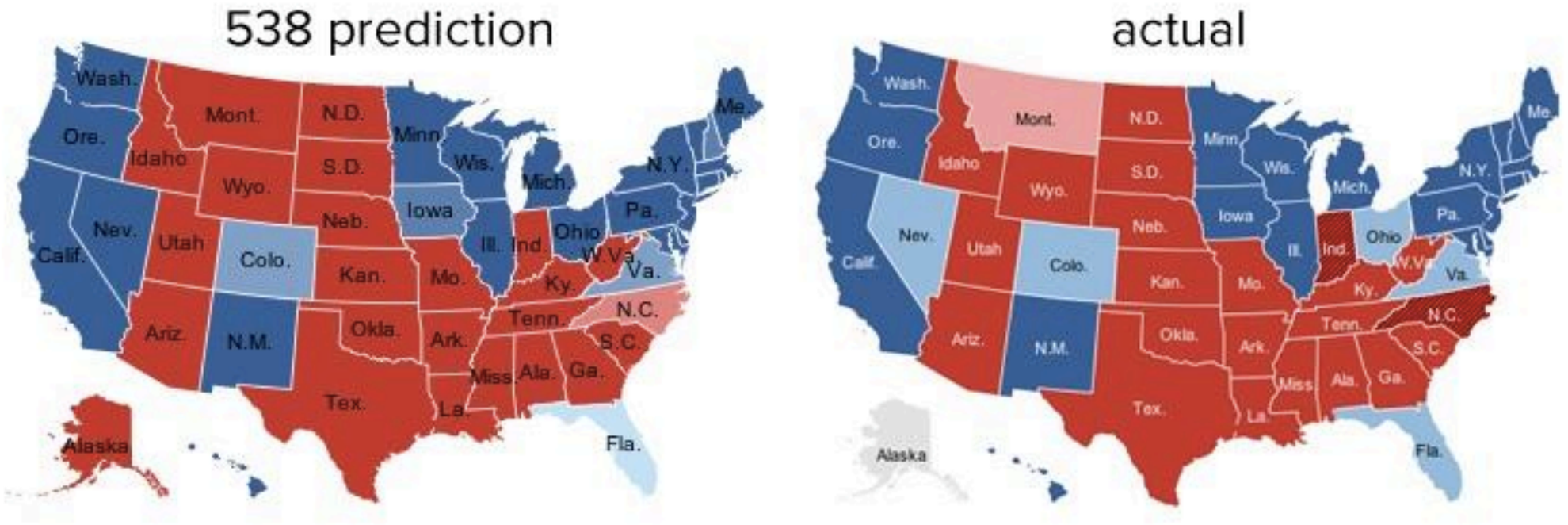
Updated 1:05 AM ET on Sept. 11

Nov. 6 Forecast | Now-cast

Barack Obama		Mitt Romney
314.9	Electoral vote	223.1
+6.7 since Sept. 3		-6.7 since Sept. 3

270 to win

For the Nate-haters, here's the 538



Michael Cosentino @cosentino · 6 Nov 2012

For the Nate-haters, here's the 538 prediction and actual results side by side

3.1K 716

Reply to @cosentino

POLITICS

More: Election 2012 Barack Obama Mitt Romney Wall Street Journal

Peggy Noonan Predicts A Romney Victory In The Most Anti-Nate Silver Column Imaginable

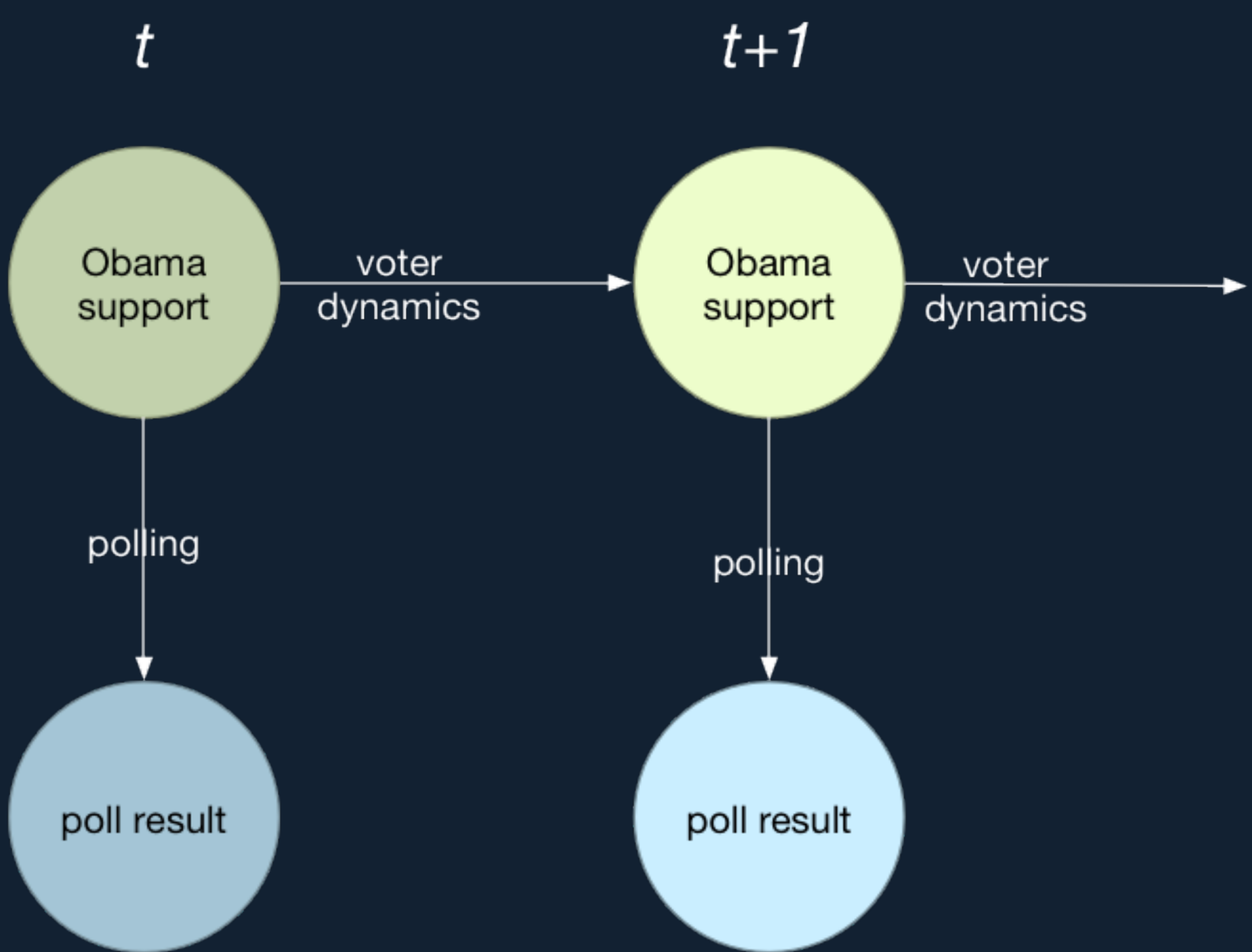
BRETT LOGIURATO

NOV. 5, 2012, 3:19 PM | 38,147 | 87

Wall Street Journal conservative columnist Peggy Noonan predicts that Republican nominee Mitt Romney will beat President Barack Obama tomorrow.

Why? Because it feels like it.

“WHILE EVERYONE IS LOOKING AT THE POLLS AND THE STORM, ROMNEY’S SLIPPING INTO THE PRESIDENCY.”


t
t+1
Obama support
voter dynamics
Obama support
voter dynamics
polling
polling
poll result
poll result

HEIRARCHICAL MODELING

CO	Purple Strategies	Obama +1.0		Romney +2.0
FL	[illegible]	[illegible]		
FL	[illegible]		Obama +3.0	
IA	[illegible]	Obama +4.0		
NC	[illegible]	Romney +1.0		Romney +5.0
NC	Gravis Marketing	Romney +8.0	Obama +0.7	
NC	Grove Insight	Obama +3.0	unchanged	
NH	New England College	Obama +3.0		
NH	Grove Insight	Obama +3.0		
NM	PPP	Obama +9.0		unchanged
NV	Gravis Marketing	Obama +1.0	Romney +0.1	
OH	CNN*	Obama +4.0	Obama +0.5	
OH	American Research Group	Obama +2.0	Obama +3.0	
OH	Purple Strategies	Obama +2.0		Romney +2.0

POLLSTER EFFECTS

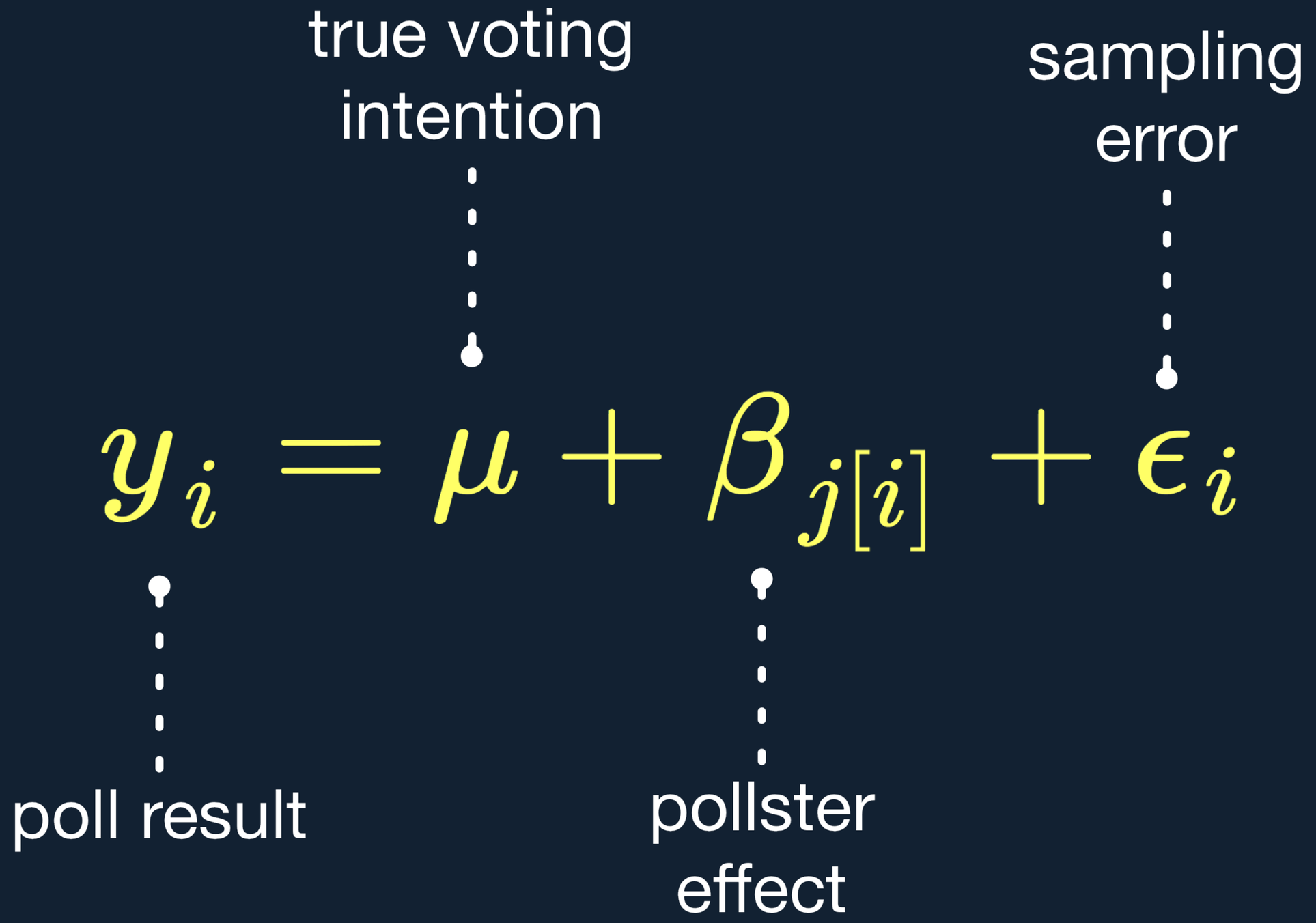
Pollster Accuracy and Bias, 2012 Presidential Election

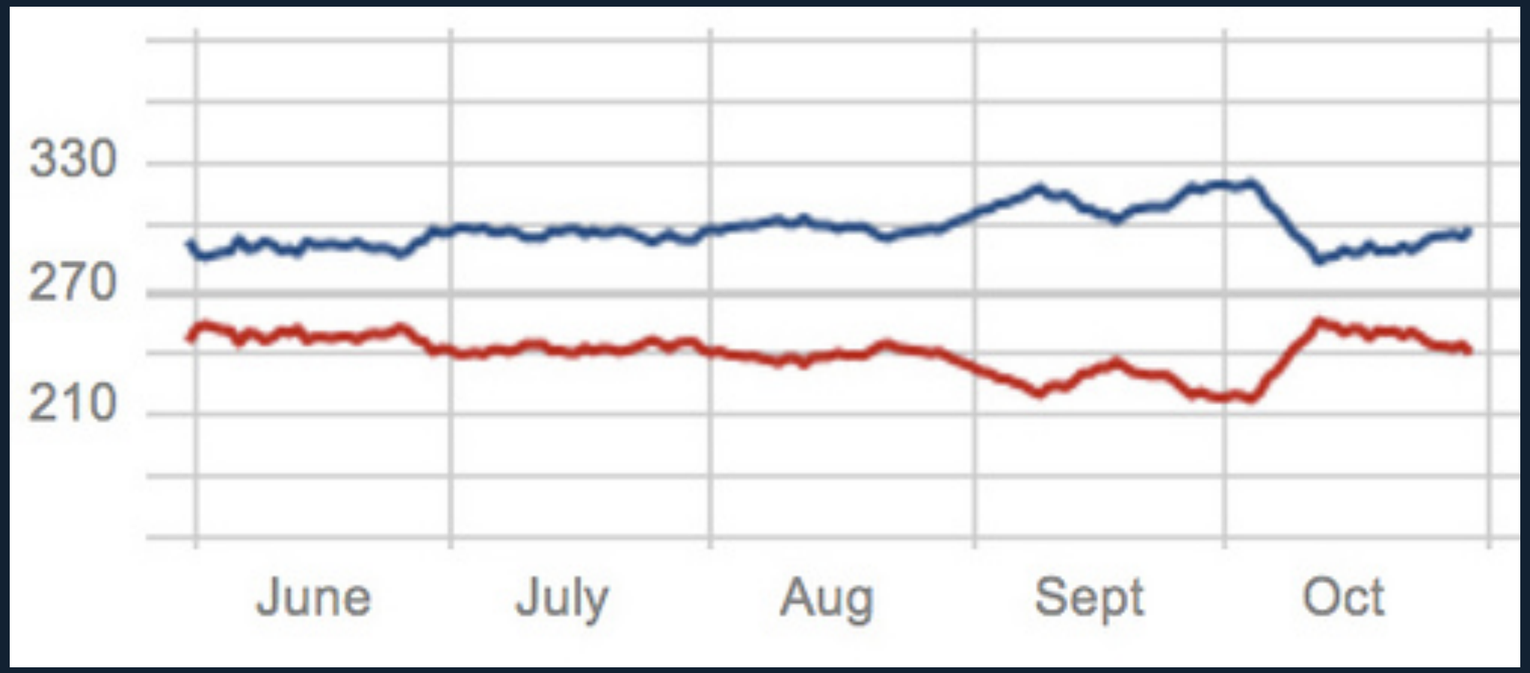
Likely Voters Polls in Last 21 Days of Campaign

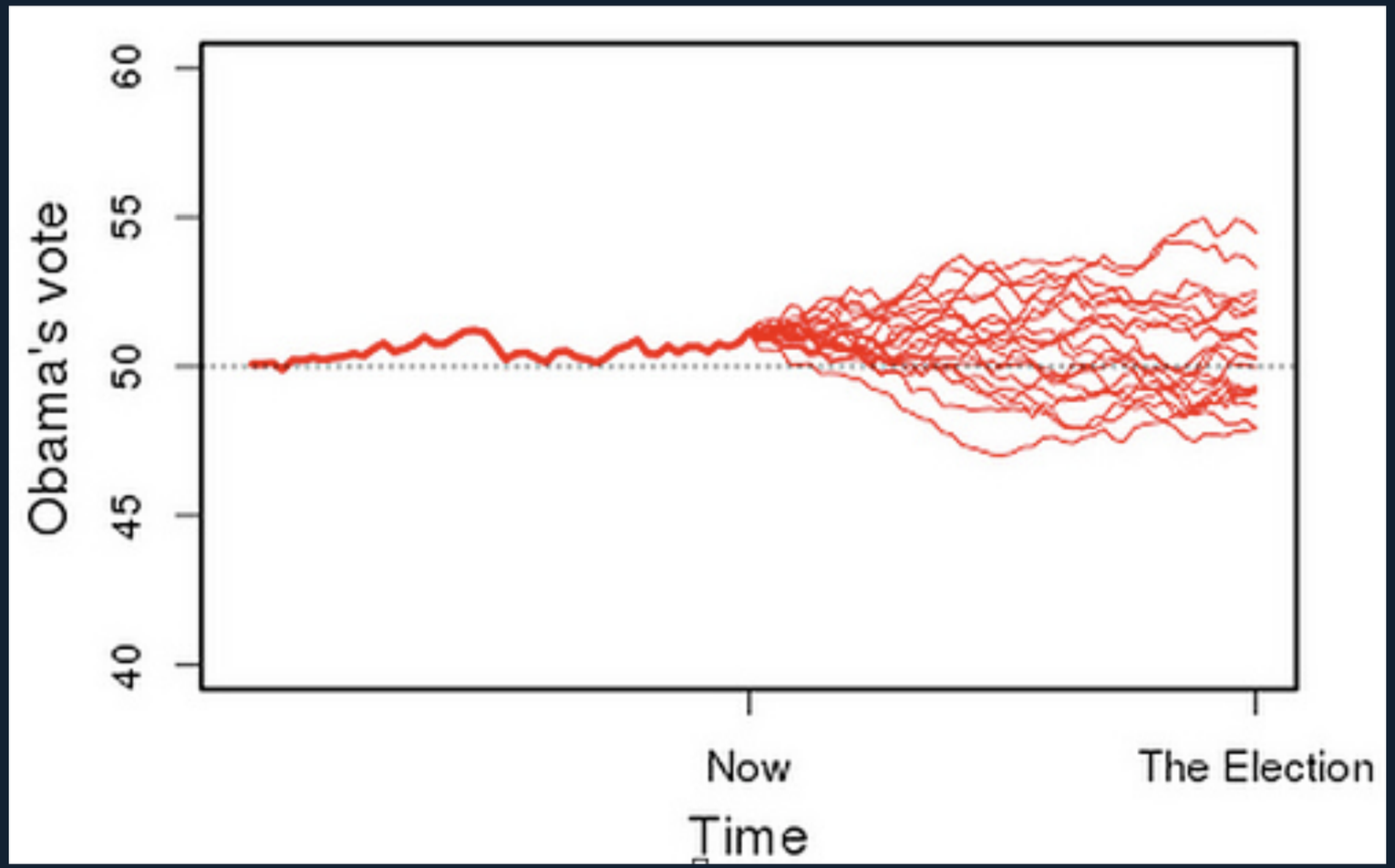
Minimum 5 Polls

Pollster	# Polls	Avg. Error	Bias	Mode	Cell?
IBD / TIPP	11	0.9	R +0.1	Live Phone	Yes
Google Consumer Surveys	12	1.6	R +1.0	Internet	N/A
Mellman	9	1.6	R +0.0	Live Phone	Yes
RAND Corporation	17	1.8	D +1.5	Internet	N/A
CNN / Opinion Research	10	1.9	R +0.6	Live Phone	Yes
Ipsos / Reuters (online)	42	1.9	R +1.4	Internet	N/A
Angus Reid	11	1.9	R +0.8	Internet	N/A
CVOTER International / UPI	13	2.0	R +2.0	Live Phone	??
Grove Insight	18	2.0	R +0.1	Live Phone	Yes
SurveyUSA	17	2.2	R +0.5	Robodial	Yes
Quinnipiac	5	2.3	D +0.3	Live Phone	Yes
Marist	11	2.5	R +1.0	Live Phone	Yes
YouGov	30	2.6	R +1.1	Internet	N/A
We Ask America	9	2.6	D +0.1	Robodial	No
Public Policy Polling	71	2.7	R +1.6	Robodial	No
Gravis Marketing	16	2.7	R +2.7	Robodial	No
JZ Analytics*	17	2.8	R +0.1	Internet	N/A
Washington Post / ABC News	16	2.8	R +2.7	Live Phone	Yes
Pharos Research Group*	14	4.0	D +2.5	Live Phone	No
Rasmussen Reports	60	4.2	R +3.7	Robo + Internet	No
American Research Group	9	4.5	R +4.5	Live Phone	Yes
Mason-Dixon	8	5.4	R +2.2	Live Phone	Yes
Gallup	11	7.2	R +7.2	Live Phone	Yes

* Not used in FiveThirtyEight forecast.


true voting intention
sampling error
$y_i = \mu + \beta_{j[i]} + \epsilon_i$
poll result
pollster effect


330
270
210
June
July
Aug
Sept
Oct

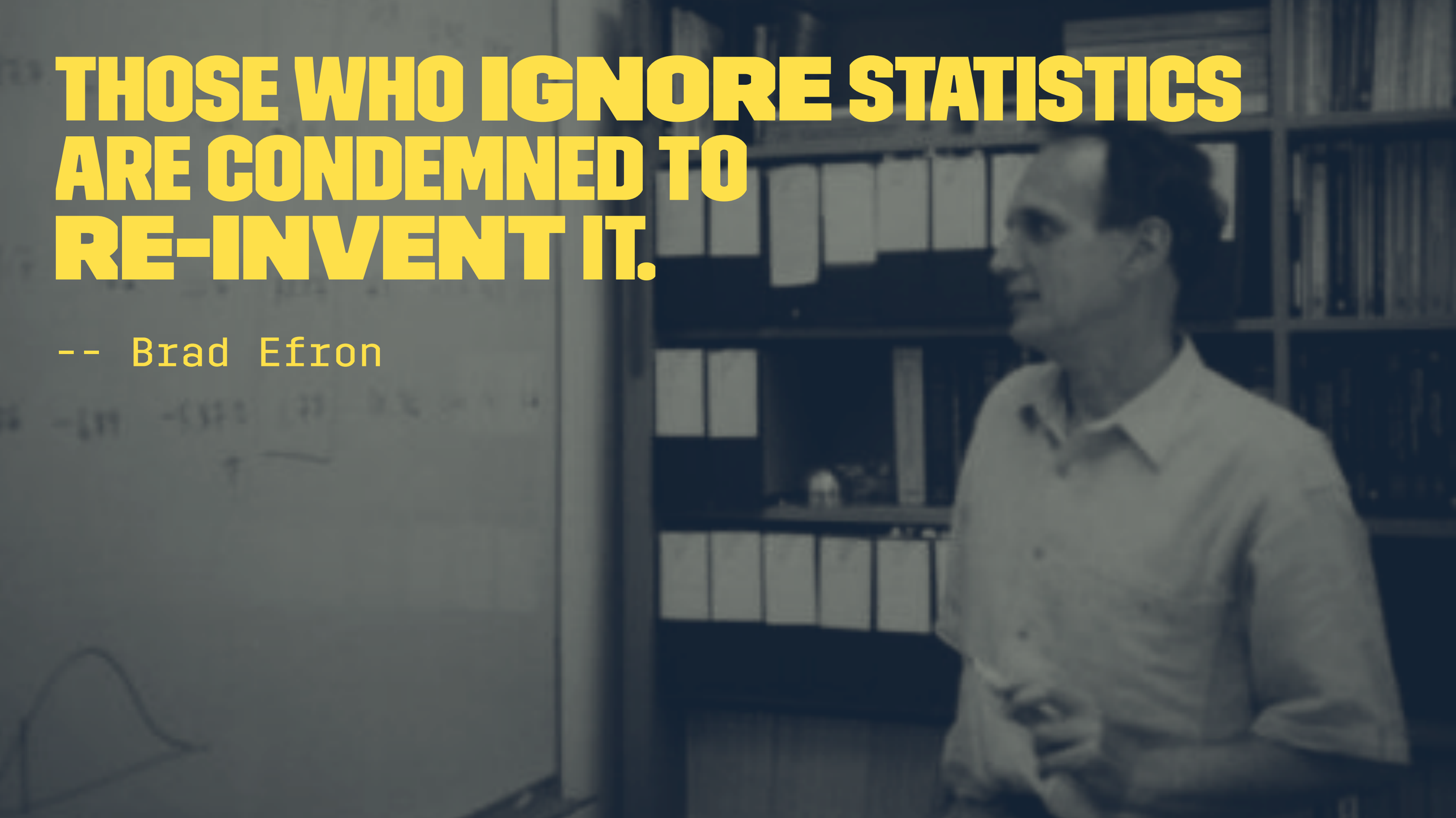

Obama's vote
60
55
50
45
40
Now
The Election
Time


90.9%
+13.5 since Oct. 30
Chance of Winning
9.1%
-13.5 since Oct. 30
50%
100%
75%
50%
25%
0%
June
July
Aug
Sept
Oct
Nov

DATA SCIENCE


DATA

SCIENCE


THOSE WHO IGNORE STATISTICS ARE CONDEMNED TO RE-INVENT IT.
-- Brad Efron