Uncertainty visualizations for improving data science decision-making

How I learned to stop worrying and love uncertainty (by visualizing it)

Ryan Wesslen / ryanwesslen.com

April 22, 2021 / UNCC Analytics Frontiers

Introduction

Day job: Bank of America

- Lead natural language processing (NLP) team
- Chief Data Scientist Organization, Enterprise Data Strategy & Governance

This presentation: UNC Charlotte Ribarsky Center and School of Data Science

- Taught Spring/Fall 2019 DSBA5122 Visual Analytics (dsba5122.com)
- 20+ peer reviewed publications with 20+ UNCC research collaborators
- Human-computer interaction, visual analytics, information visualization, computational social science, cognitive science, psychology

POLITICS

‘Extraordinarily Uncertain,’ Powell Says of America’s Economic Future

Jerome H. Powell, the Federal Reserve chair, addressed the extent of the economic damage caused by the coronavirus pandemic.

By The Associated Press

Uncertainty In Congress Over Next Moves To Address Coronavirus Crisis

March 16, 2020 · 7:01 AM ET

 SUSAN DAVIS  KELSEY SNELL  CLAUDIA GRISALES

30th January

US stock market falls as Gamestop battle creates uncertainty for Wall Street

By Herald Scotland online

ECONOMY AND BUSINESS March 27 2021

The blockade of the Suez Canal continues to generate uncertainty and economic concern

12% of world trade transits by waterway

Why is uncertainty often ignored in data analysis?

Data journalism

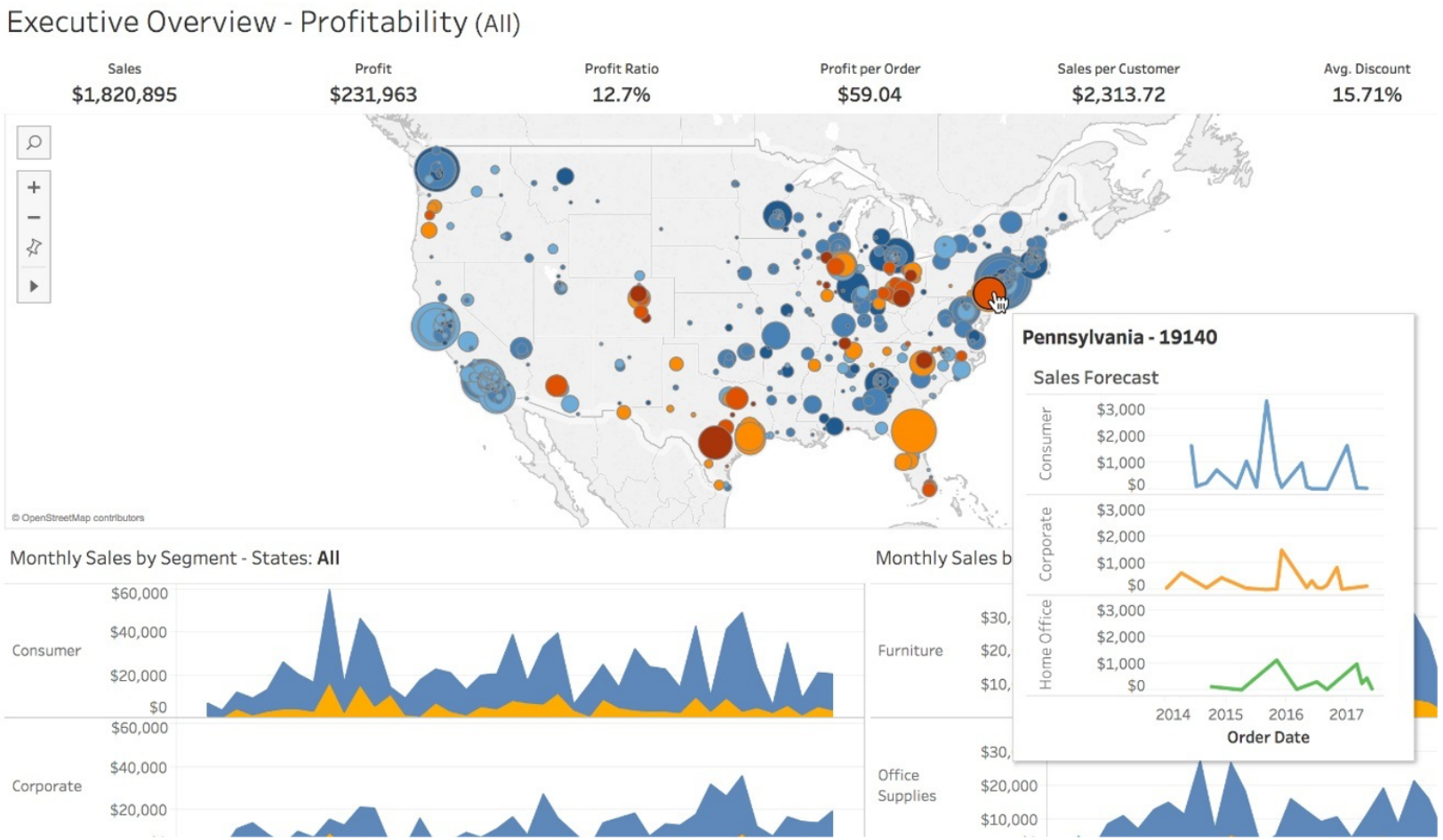
VACCINE · **Published** March 25

AstraZeneca COVID-19 vaccine shows 76% efficacy against symptomatic infection in updated data

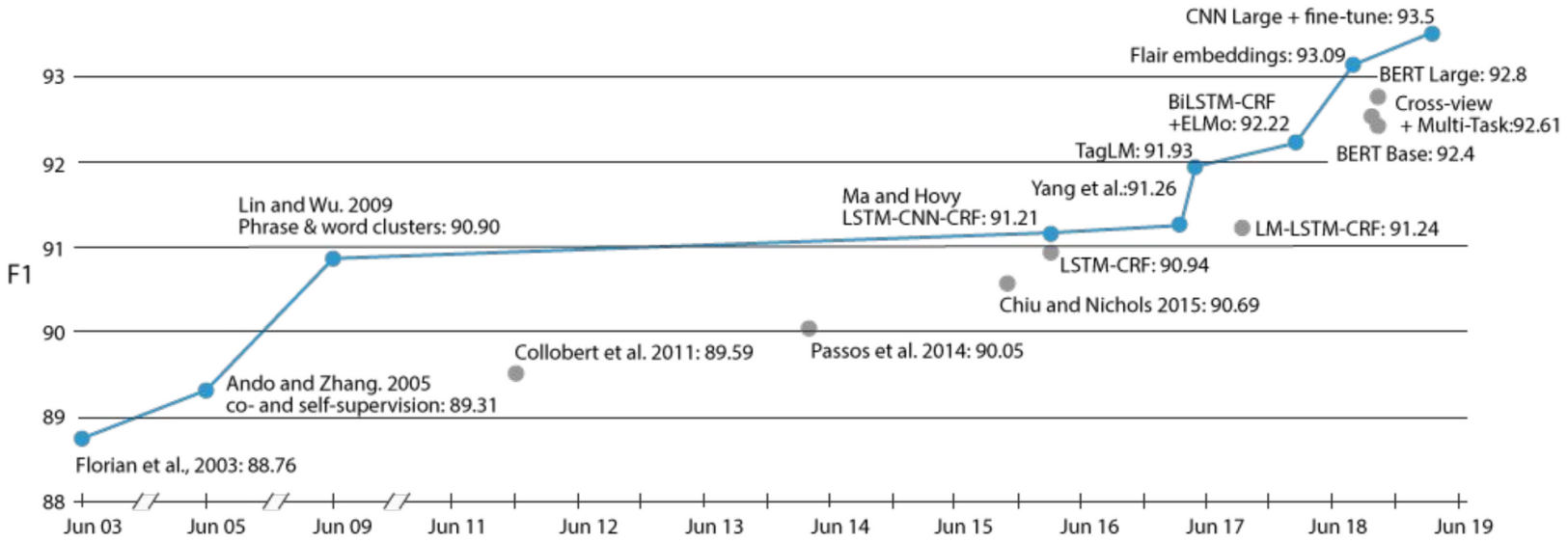
The new analysis also reported 100% efficacy 'against severe or critical disease and hospitalization'

By **Alexandria Hein | Fox News**

Data analytics



Artificial intelligence (e.g., NLP)



Accuracy on Named Entity Recognition (NER) on CoNLL-2003 over time: Sebastian Ruder

Why not visualize uncertainty?

Cognitively burdensome

User won't understand it

Implies inappropriate precision

Not integral to task

Undermines author's credibility

Hard to calculate

Lack of good visual techniques

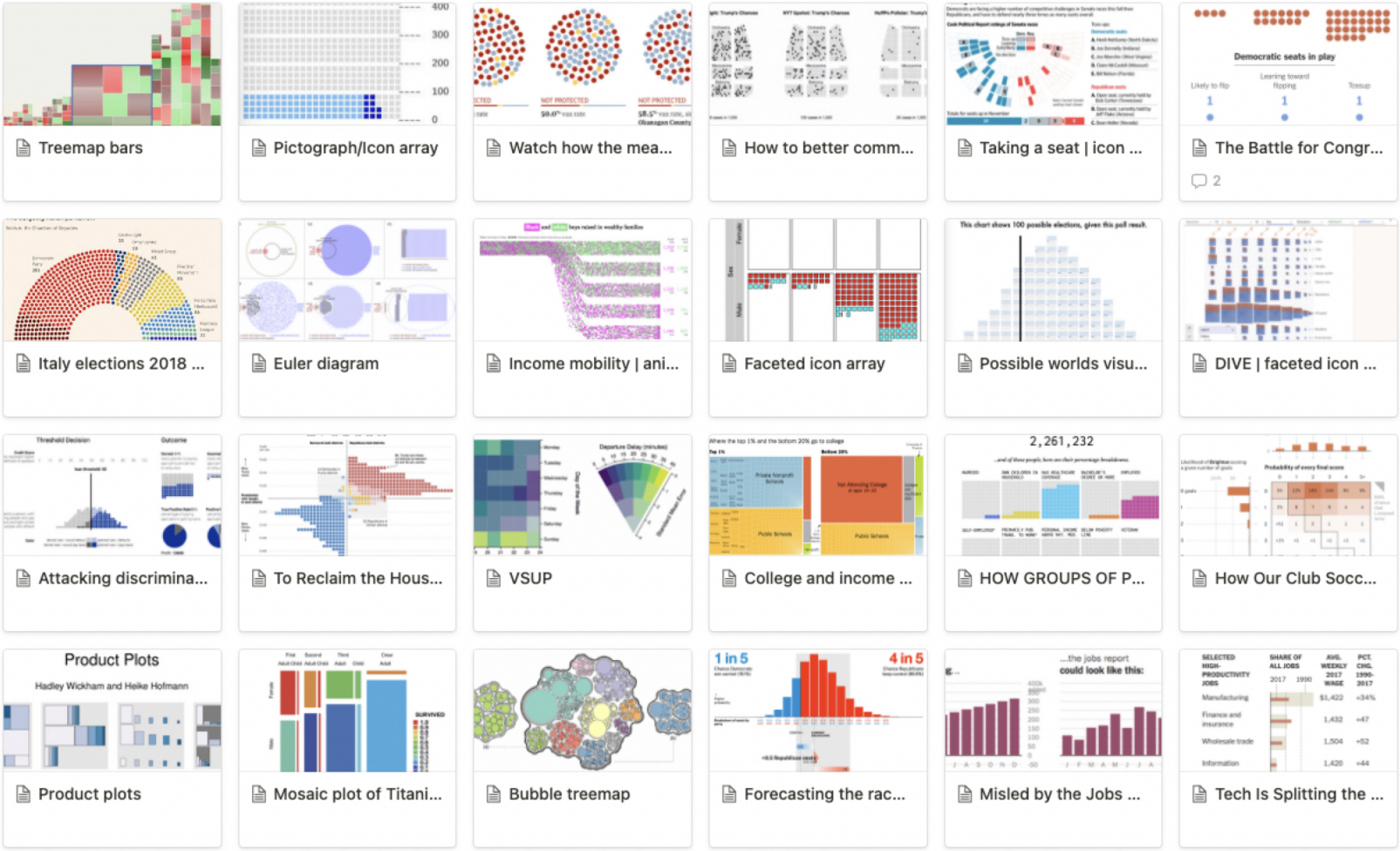
Hard to evaluate (unclear goals)

Unclear if error in data vs uncertainty

Jessica Hullman, Why Authors Don't Visualize Uncertainty (TVCG 2019)

Can better uncertainty representations enable better decision-making?

"People are very good at ignoring uncertainty... but it's especially true when we provide bad uncertainty representations" - Matthew Kay



Xiaoying Pu, Matthew Kay

Advances in communicating uncertainty in the fields **information visualization**, **visual analytics**, and **human-computer interaction** provide opportunities to better incorporate uncertainty into data science for better decisions.

Information visualization + Visual analytics + Human-computer interaction

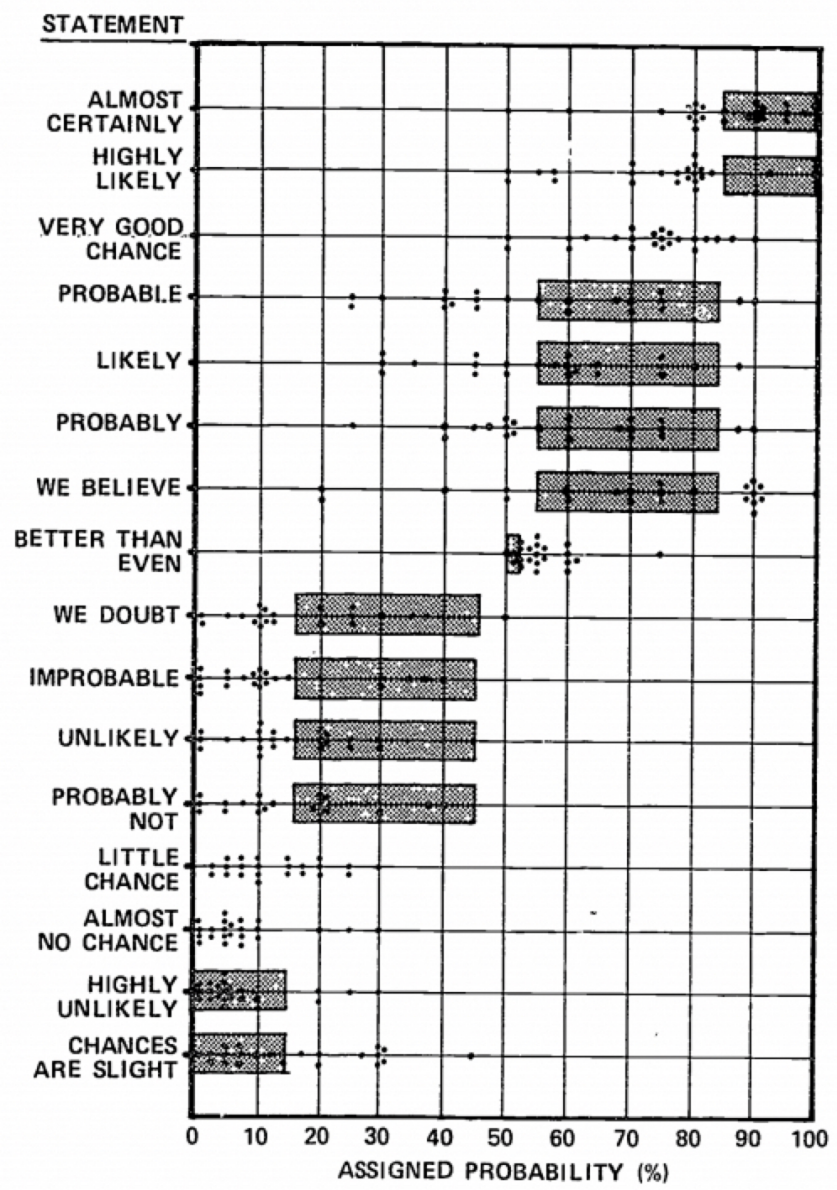
Psychology	UI/UX development	Experiment design
Cognitive science	Machine learning	Causal inference
Behavioral economics	Back end engineering	Qualitative feedback

Develop / modify theories

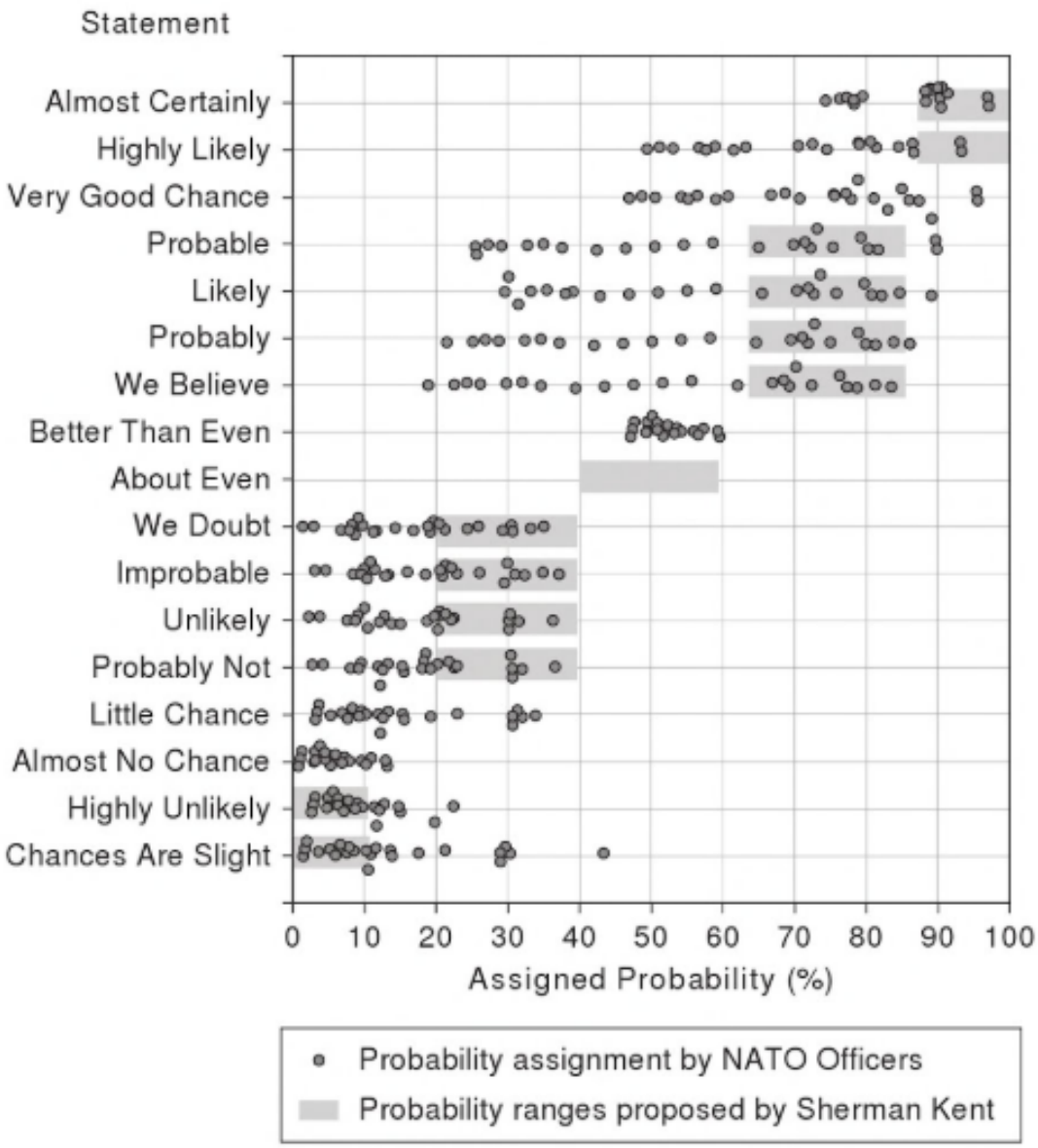
Probabilities are hard

Sherman Kent, 1903-1986.

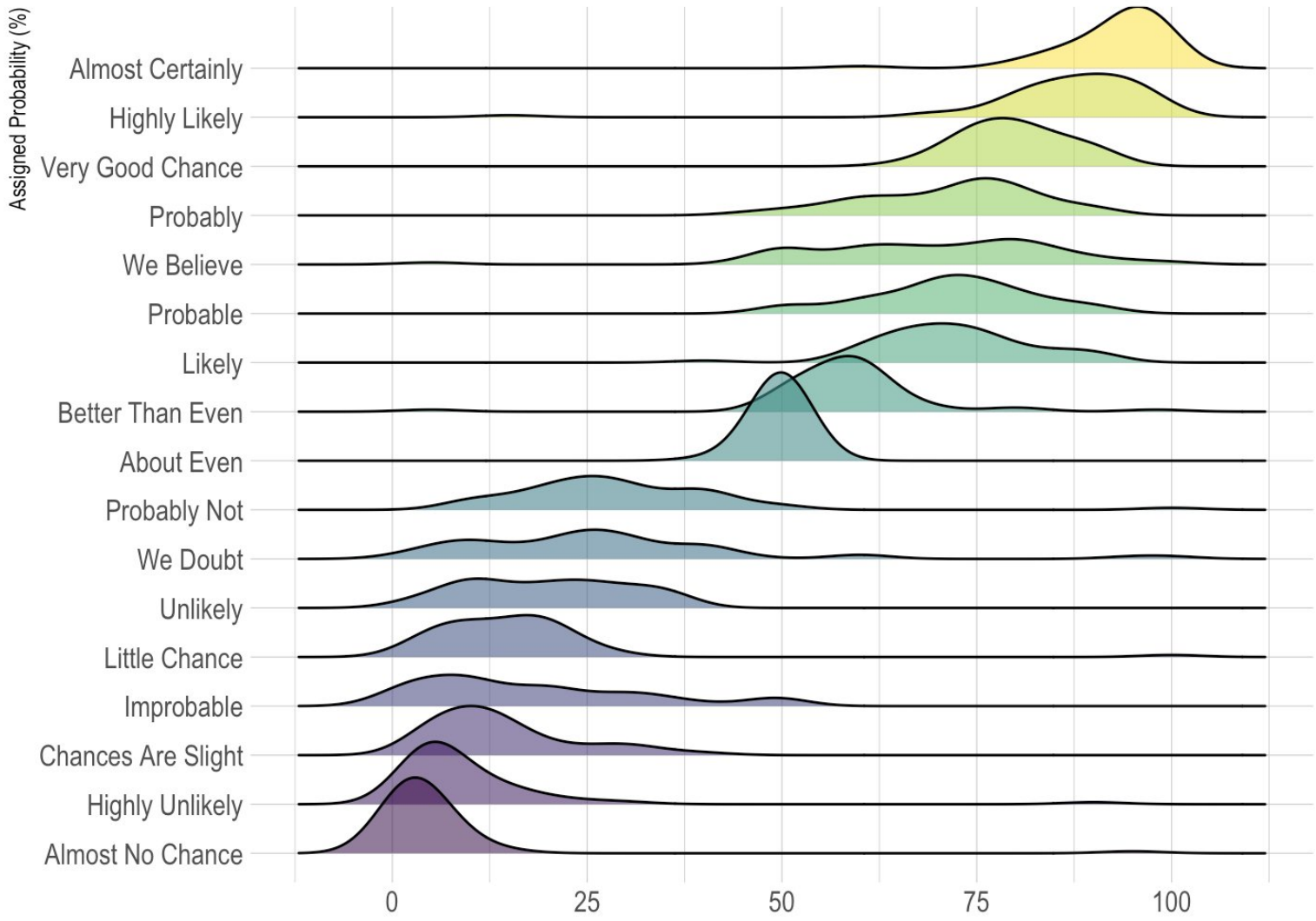
"Father of Intelligence Analysis"

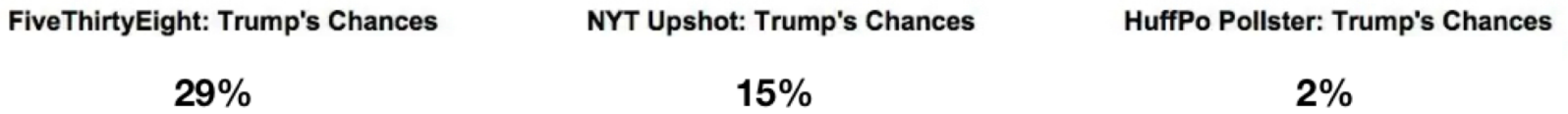


Barclay et al., 1977



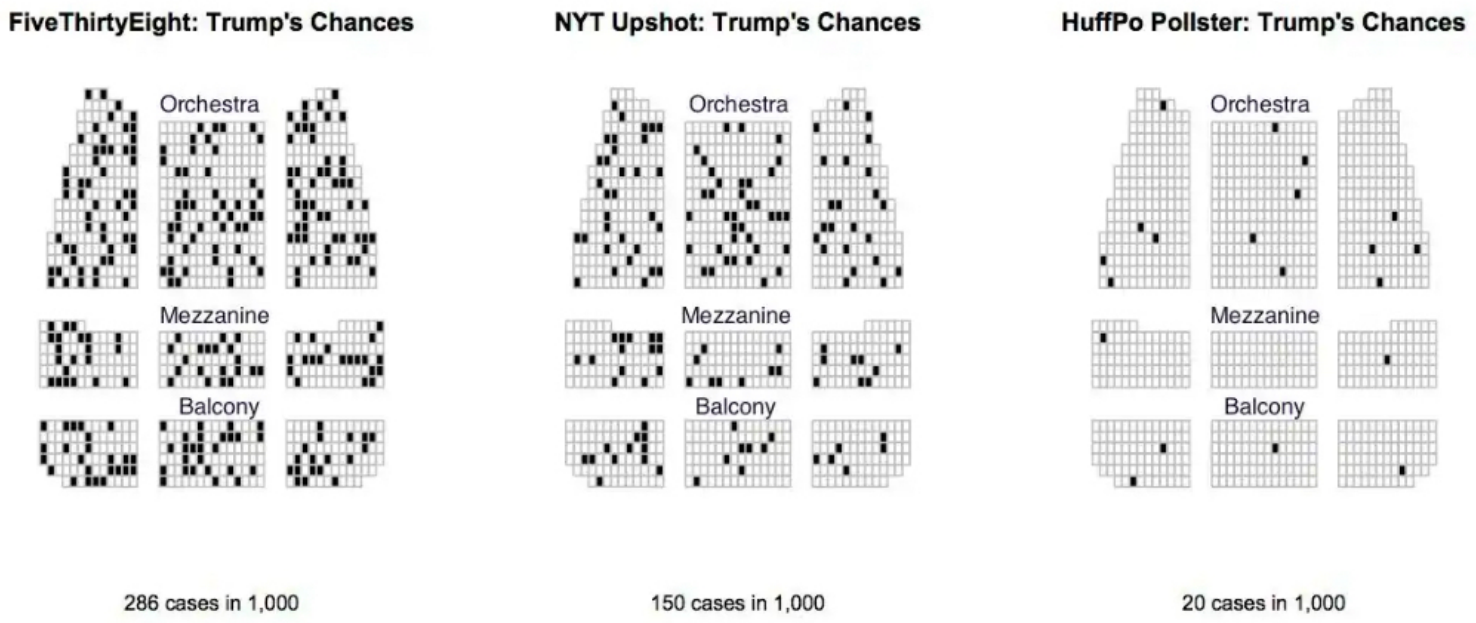
Pherson and Pherson, 2013


Assigned Probability (%)
Almost Certainly
Highly Likely
Very Good Chance
Probably
We Believe
Probable
Likely
Better Than Even
About Even
Probably Not
We Doubt
Unlikely
Little Chance
Improbable
Chances Are Slight
Highly Unlikely
Almost No Chance
0
25
50
75
100

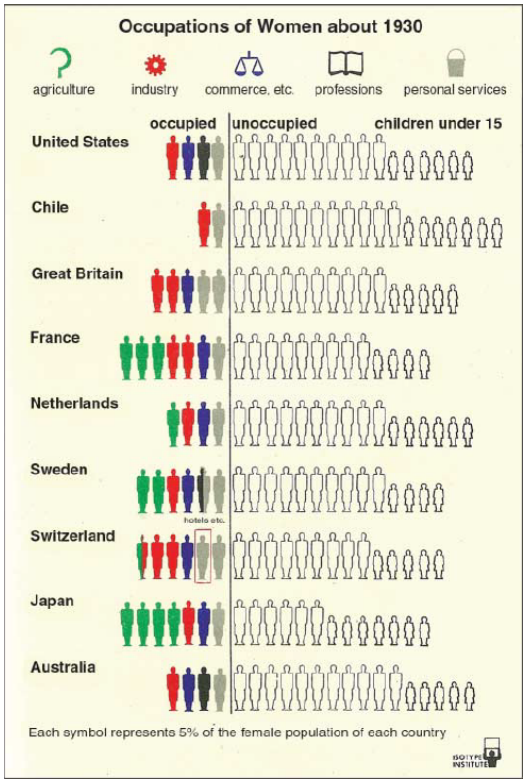


Justin Gross (Washington Post)

Frequency framing

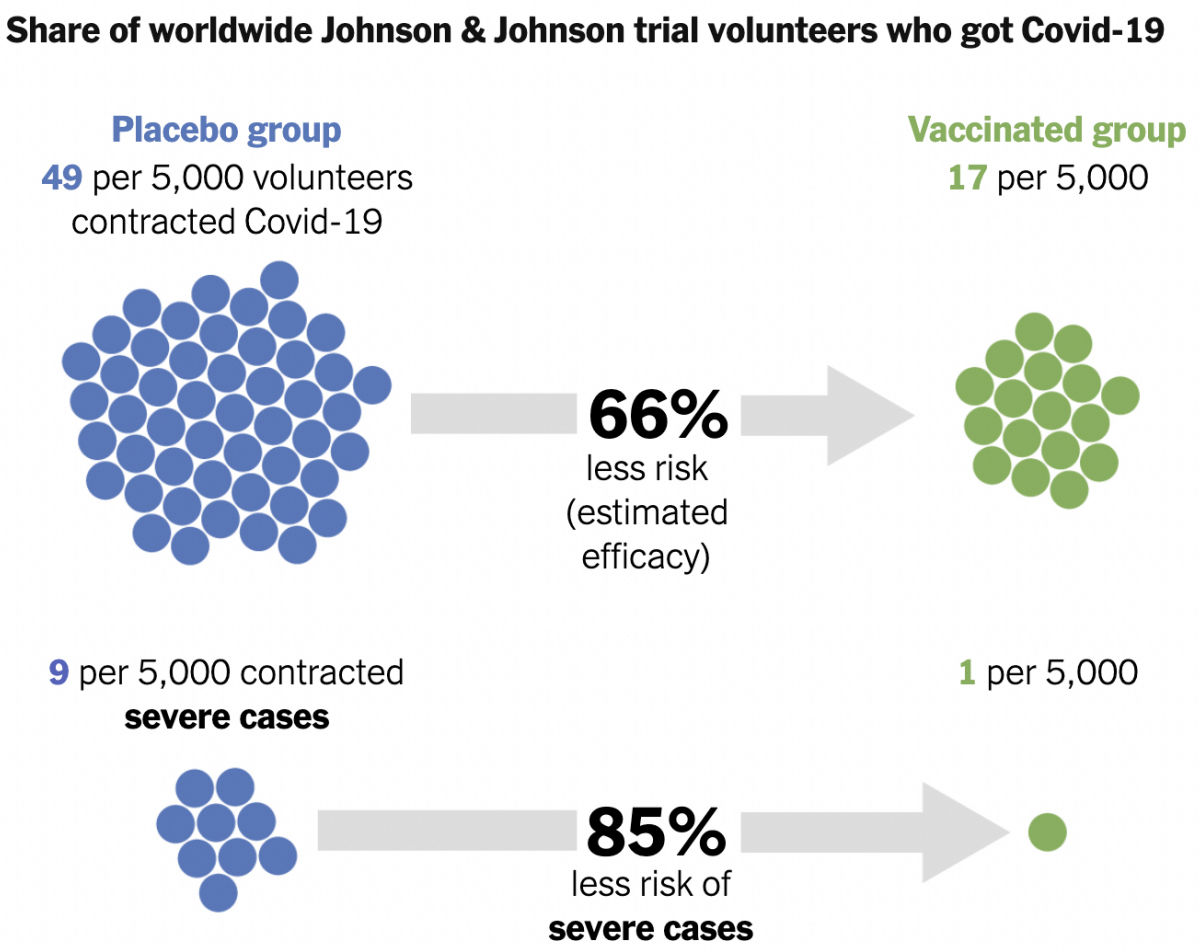


Justin Gross (Washington Post)



Icon arrays and Isotypes by Otto Neurath

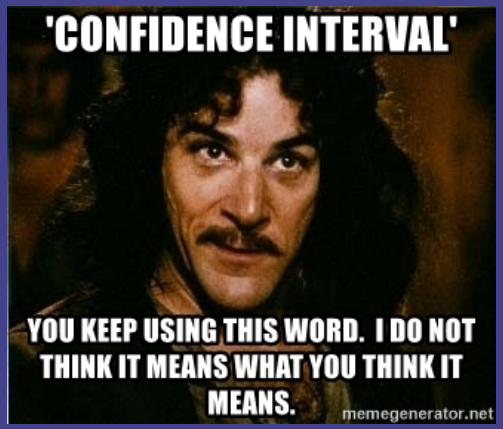
"Understanding uncertainty: Visualizing Probabilities" by Pearson & Short

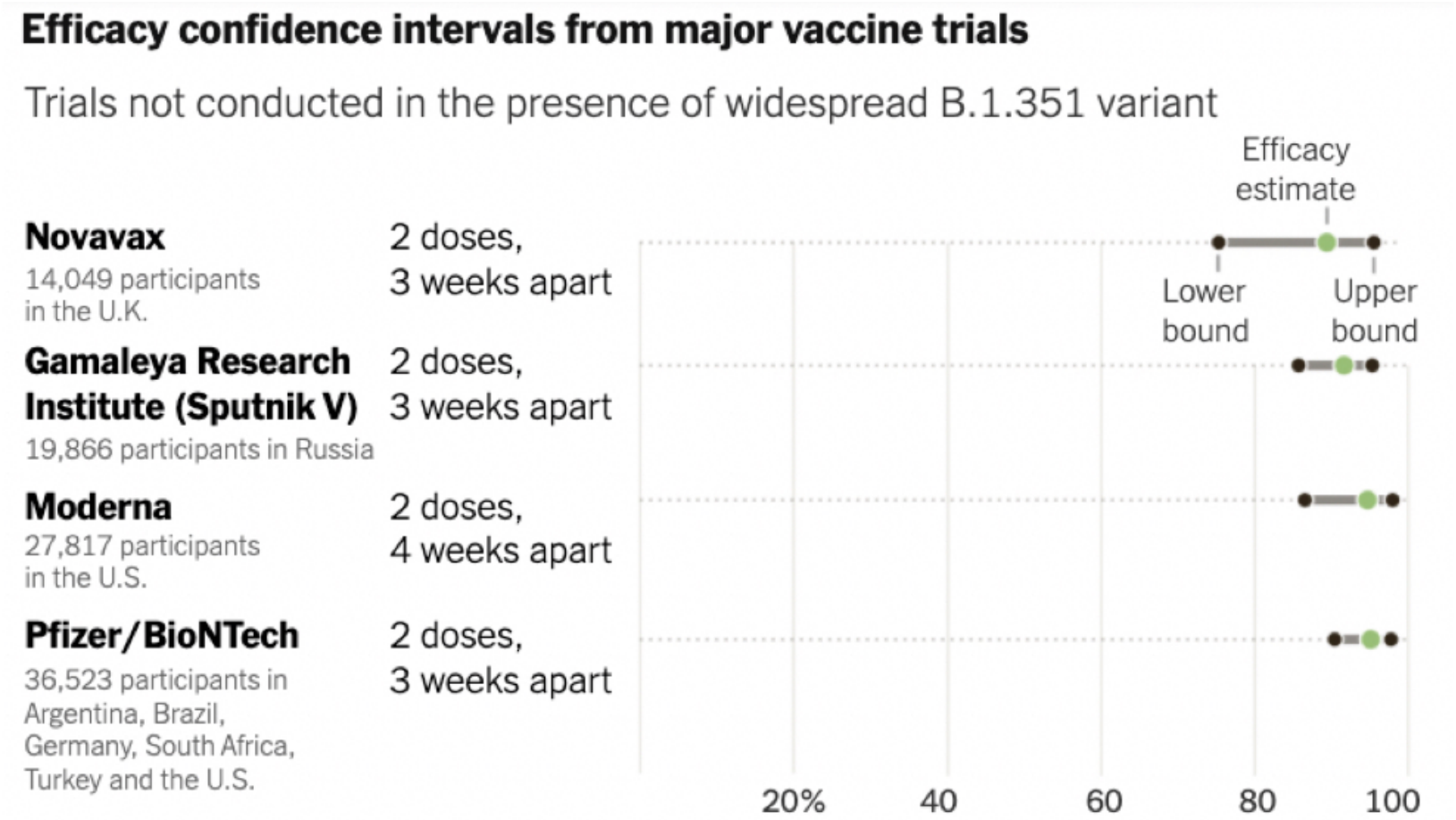


Notes: Numbers are rounded. There were 19,544 volunteers in the placebo group: 193 of them (0.99 percent) were infected with Covid-19, and 34 (0.17 percent) were severely infected. There were 19,514 volunteers in the vaccinated group: 66 of them (0.34 percent) were infected, and 5 (0.03 percent) were severely infected. ▪ Source: The Food and Drug Administration's analysis of clinical trials conducted by Johnson & Johnson

What Do Vaccine Efficacy Numbers Actually Mean? By Carl Zimmer and Keith Collins, March 3, 2021

Statistics can be confusing

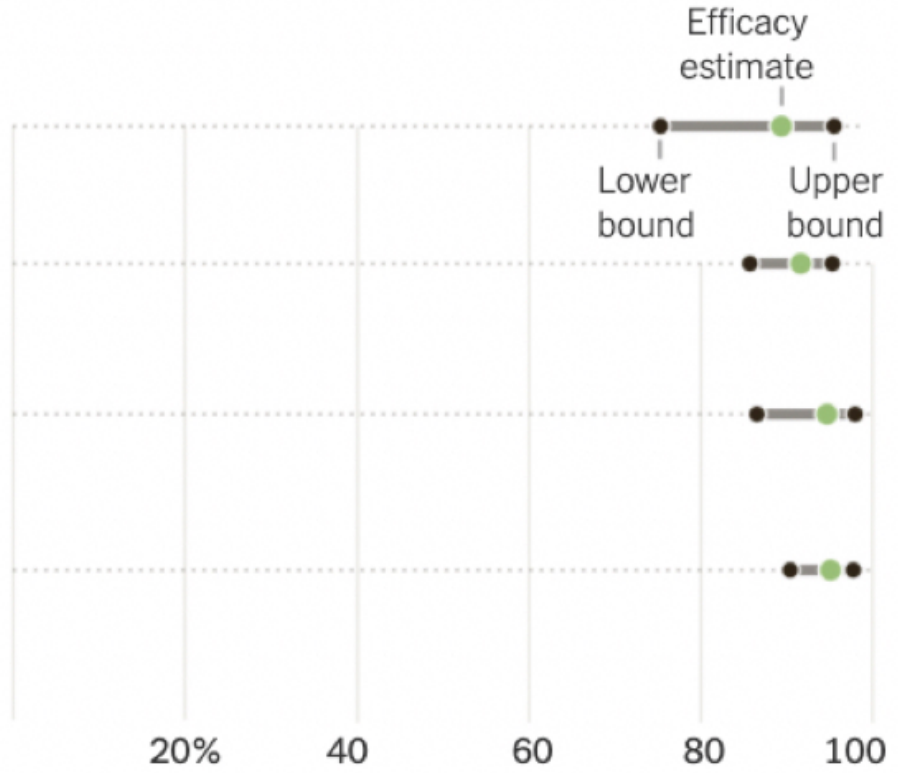




Does it mean: There is a 95% chance that the true efficacy falls within the confidence interval?

Efficacy confidence intervals from major vaccine trials

Trials not conducted in the presence of widespread B.1.351 variant



Novavax
14,049 participants in the U.K.
2 doses, 3 weeks apart

Gamaleya Research Institute (Sputnik V)
19,866 participants in Russia
2 doses, 3 weeks apart

Moderna
27,817 participants in the U.S.
2 doses, 4 weeks apart

Pfizer/BioNTech
36,523 participants in Argentina, Brazil, Germany, South Africa, Turkey and the U.S.
2 doses, 3 weeks apart

Does it mean: There is a 95% chance that the true efficacy falls within the confidence interval?

Lucy D'Agostino McGowan
@LucyStats

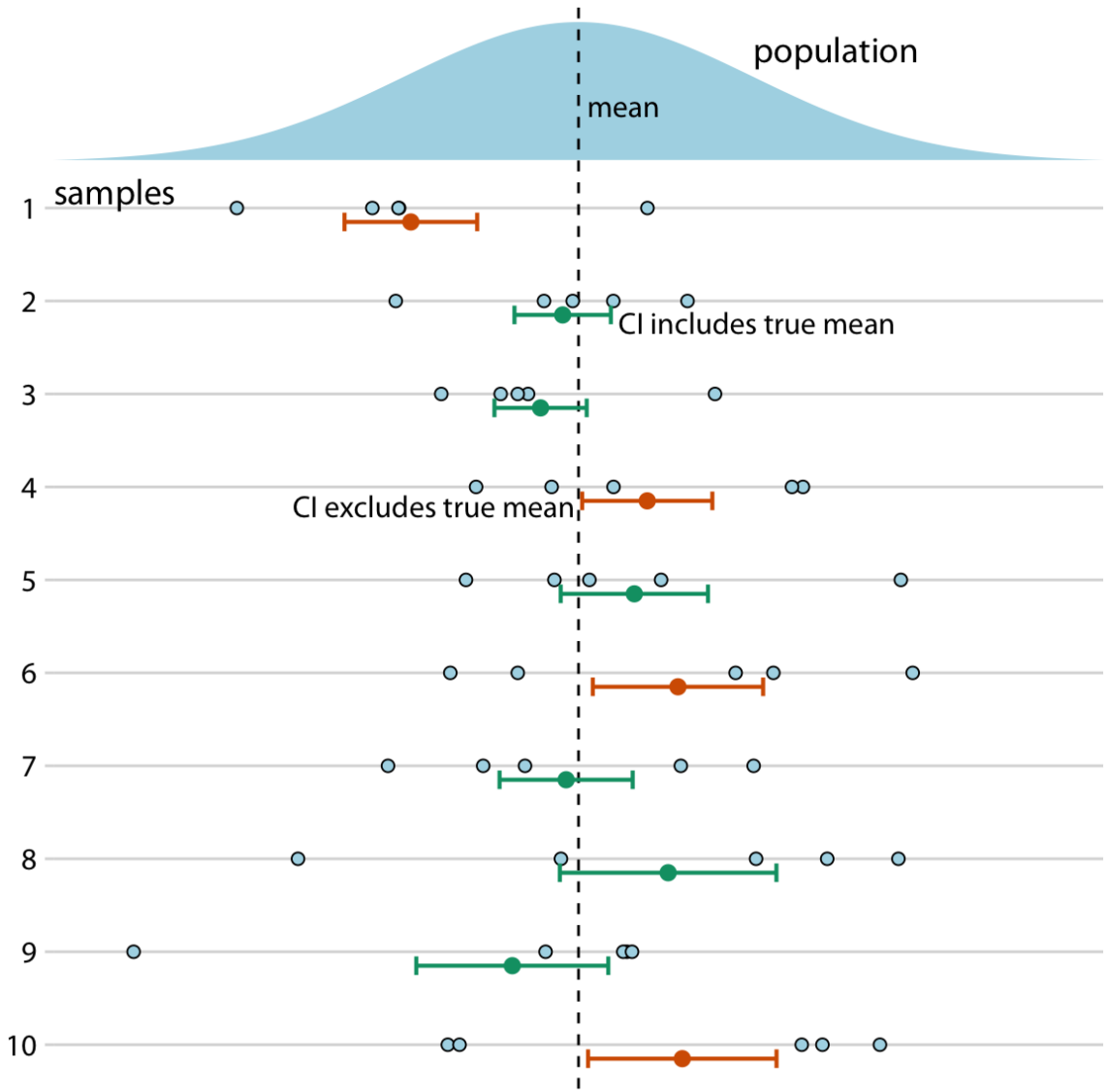
Proper confidence interval definition in the NYTimes, love to see it!

"If scientists came up with confidence intervals for 100 different samples using this method, the efficacy would fall inside the confidence intervals in 95 of them."

What Do Vaccine Efficacy Numbers Actually Mean?
All of the F.D.A.-authorized vaccines offer strong protection against Covid-19, and assessing their efficacy isn't as simple as a head-to-hea...
nytimes.com

7:12 PM · Mar 4, 2021

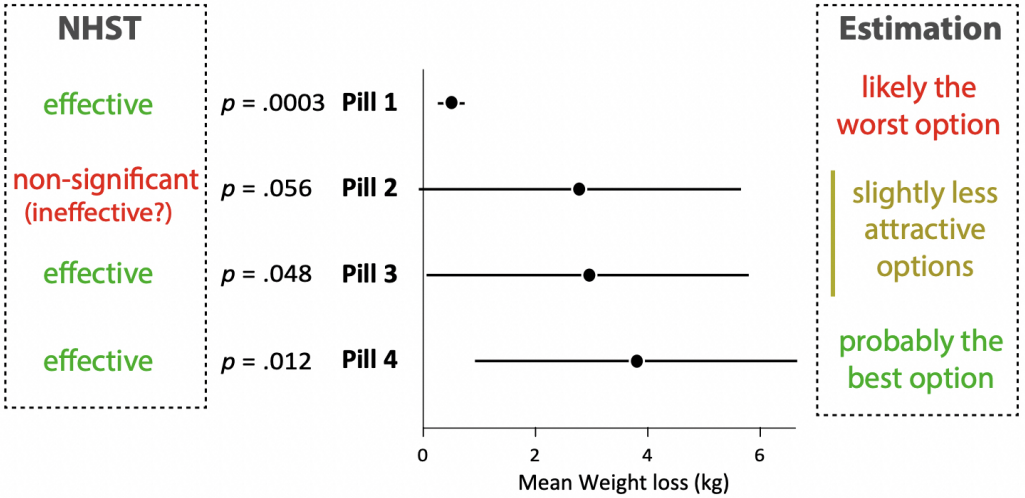
542 11 Copy link to Tweet


population
mean
samples
1
2
CI includes true mean
3
4
CI excludes true mean
5
6
7
8
9
10

Claus Wilke

Issue with null hypothesis testing

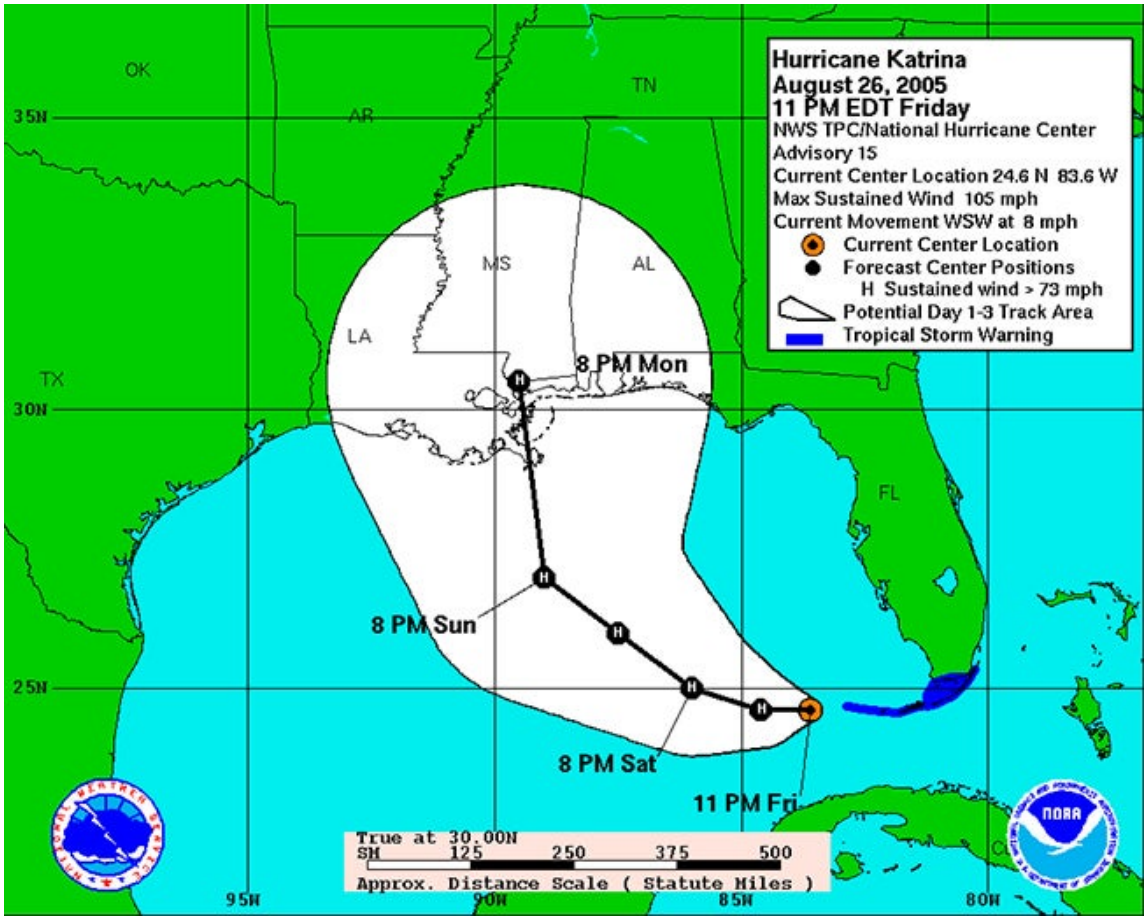
Fig. 8 The same four pills, ranked based on the outcome of statistical tests (left), and based on an examination of effect sizes and interval estimates (right).



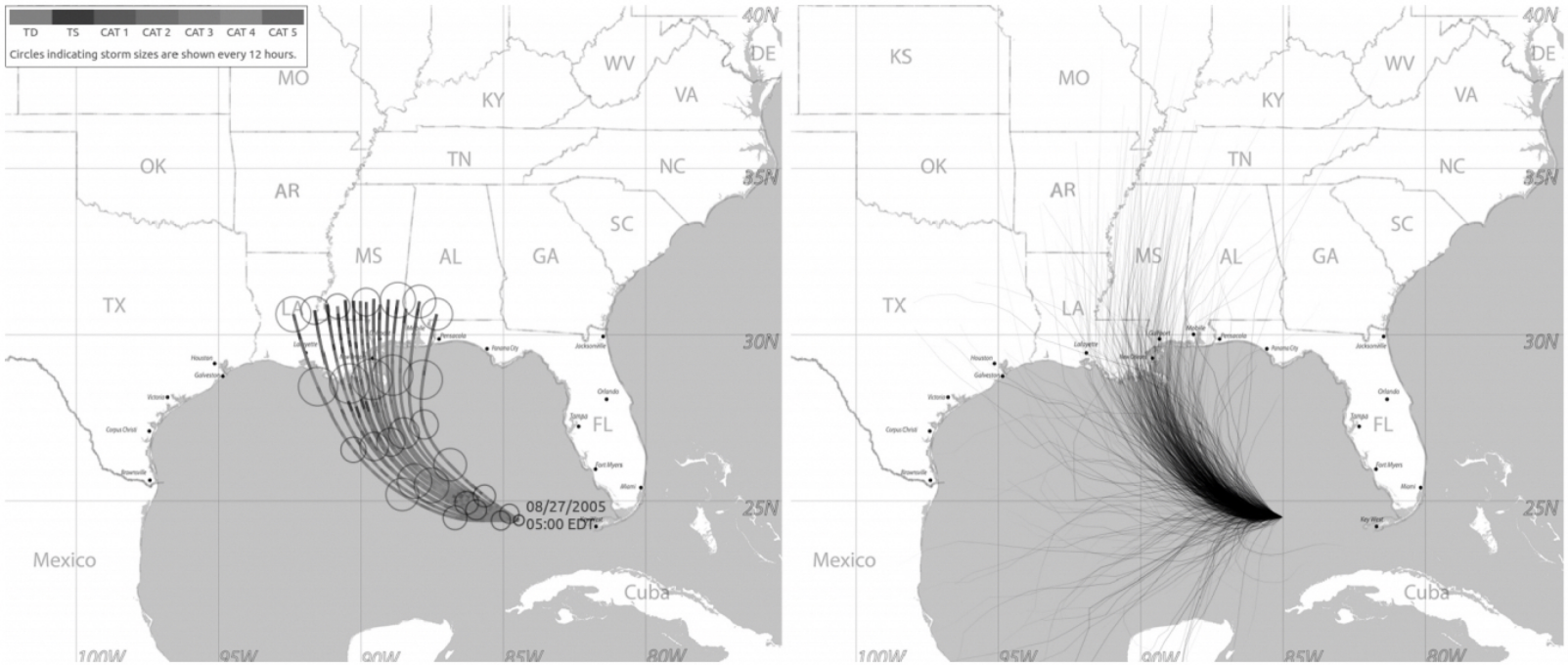
Pierre Dragicevic's Fair Statistical Communication in HCI

Szucs and Ioannidis, 2017: "When Null Hypothesis Significance Testing (NHST) Is Unsuitable for Research: A Reassessment"

Dichotomous thinking

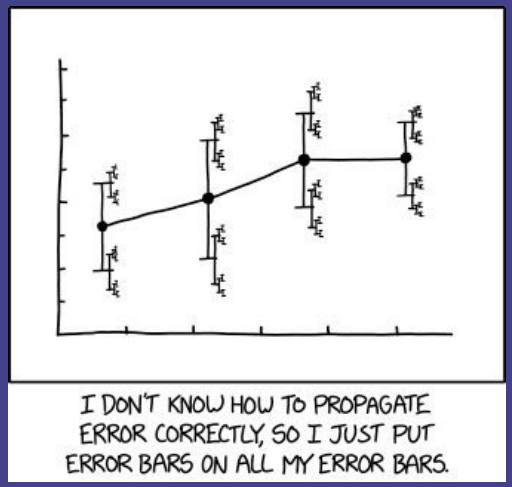


Ensemble hurricane paths



Liu et al. (2016); Padilla, Hullman, and Kay (2020)

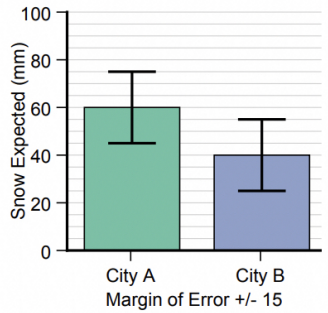
Error bars are not enough



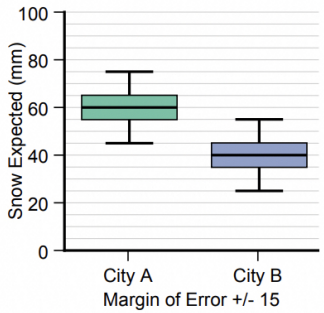
xkcd

Error Bars Considered Harmful: Exploring Alternate Encodings for Mean and Error

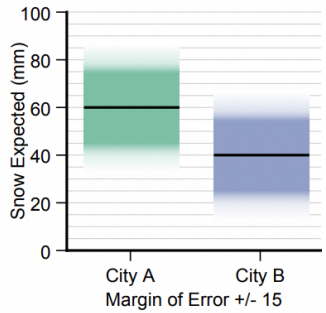
Michael Correll *Student Member, IEEE*, and Michael Gleicher *Member, IEEE*



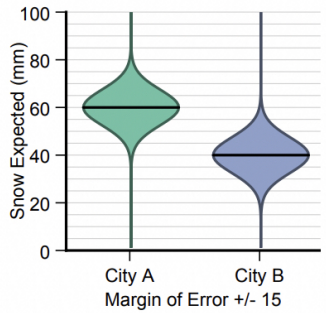
(a) **Bar chart** with error bars: the height of the bars encodes the sample mean, and the whiskers encode a 95% t-confidence interval.



(b) **Modified box plot**: The whiskers are the 95% t-confidence interval, the box is a 50% t-confidence interval.



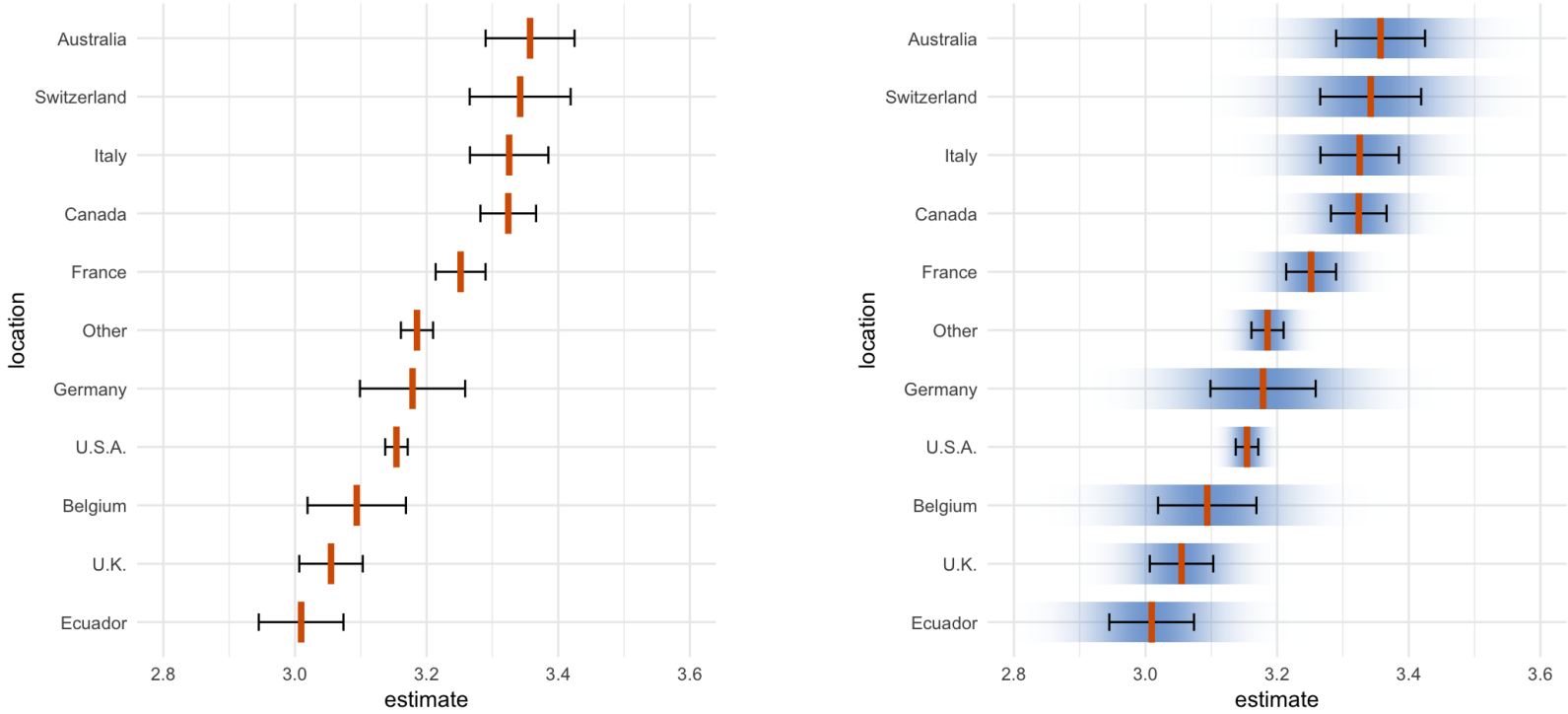
(c) **Gradient plot**: the transparency of the colored region corresponds to the cumulative density function of a t-distribution.



(d) **Violin plot**: the width of the colored region corresponds to the probability density function of a t-distribution.

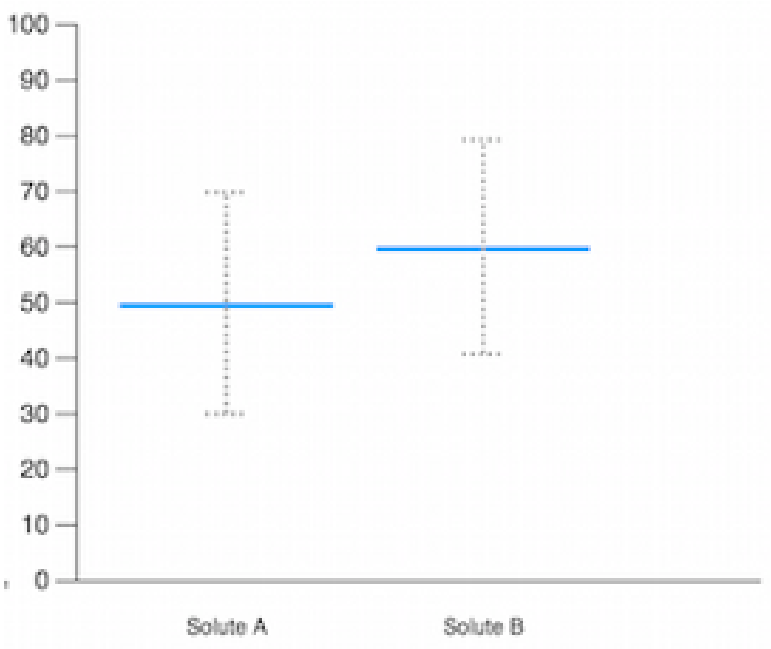
Correll and Gleicher, TVCG 2014

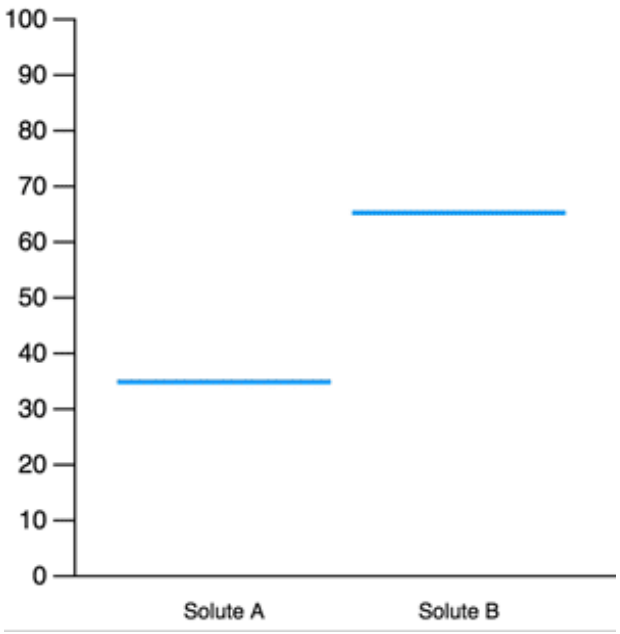
Chocolate ratings by country



Source: Claus Wilke's `ungeviz` package

How likely is it that B will be greater than A if many more draws are taken?

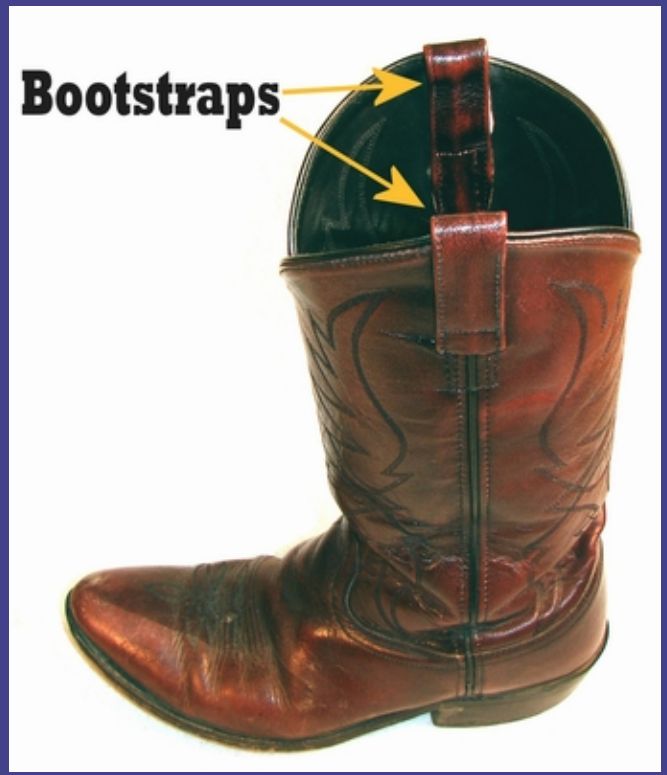




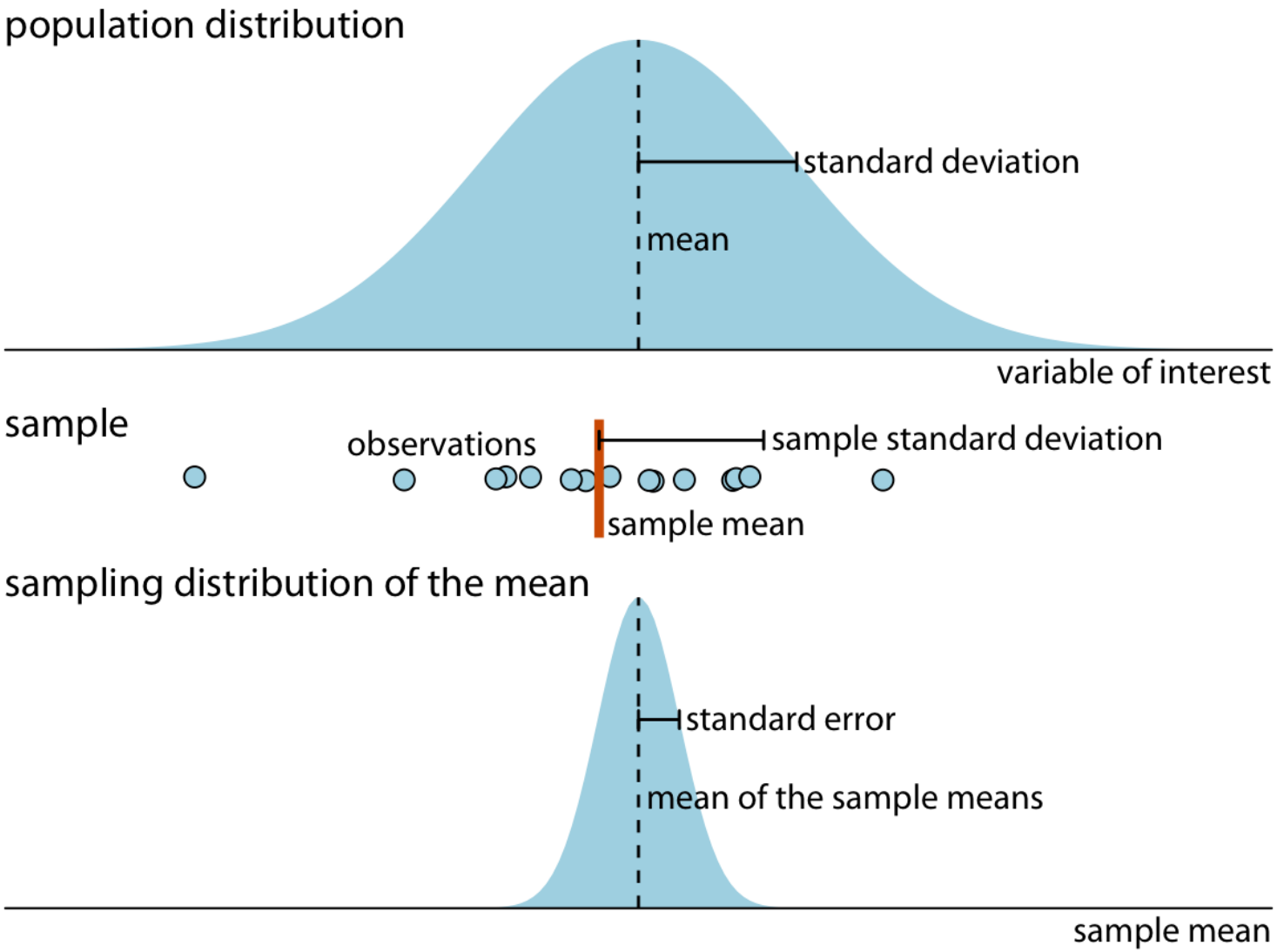
Hypothetical outcome plots

UW Interactive Data Lab / Hullman, Resnick and Adar, PLOS One 2015

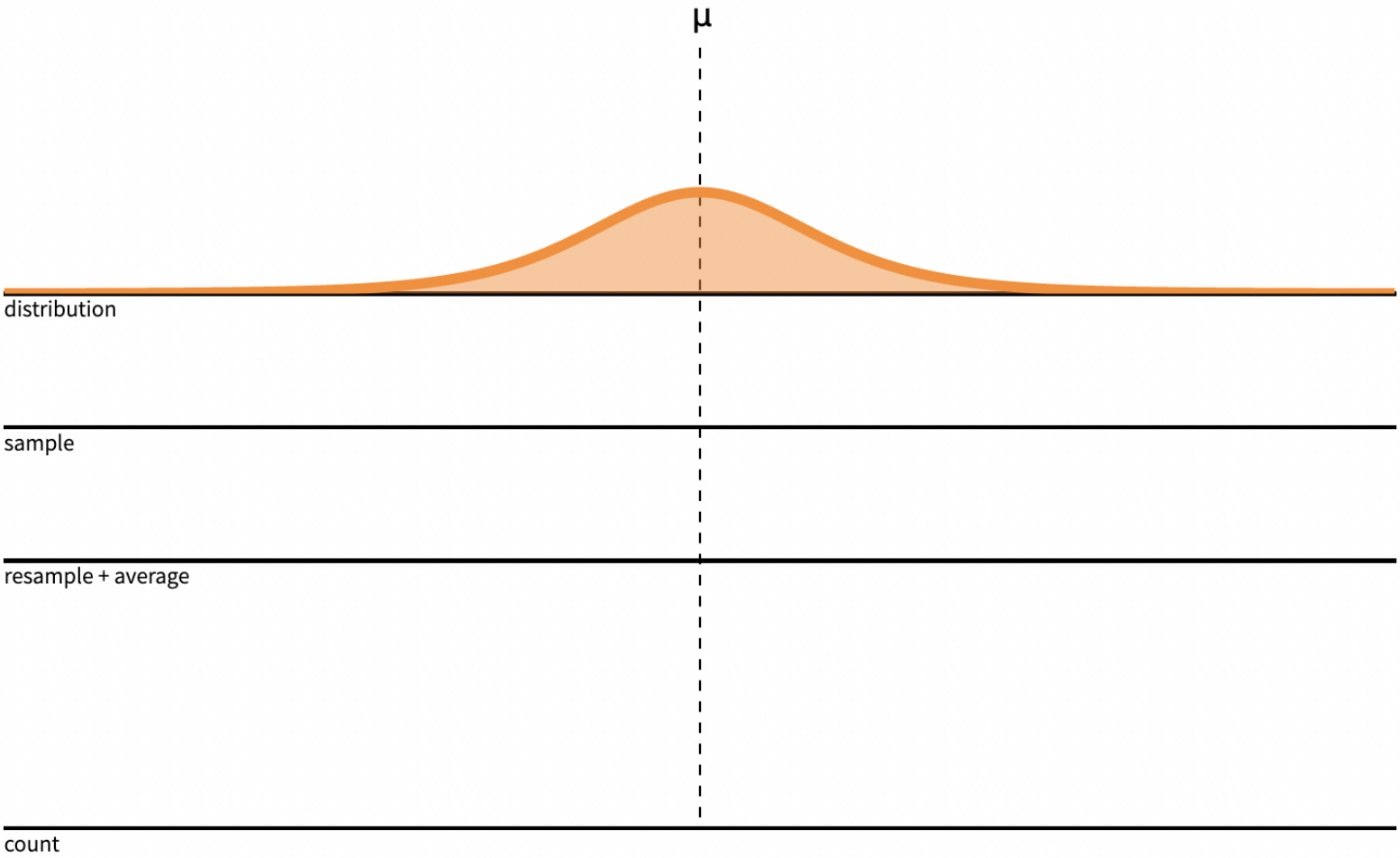
Bootstrapping



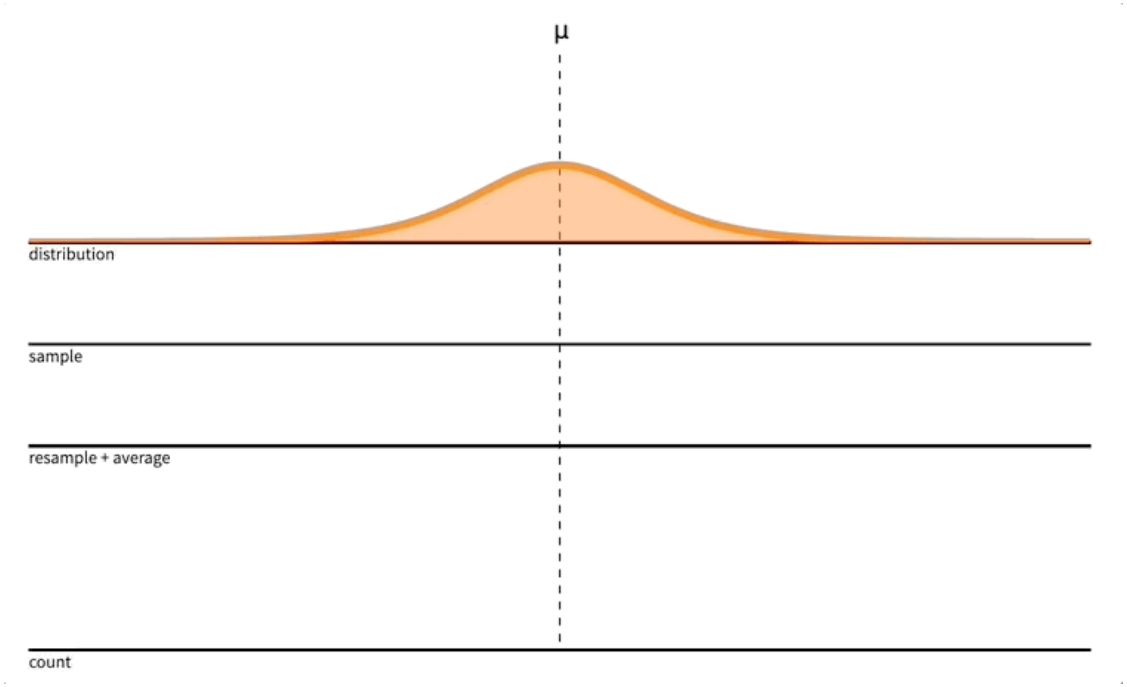
Bradley Efron's 1979 "Bootstrap Methods: Another Look at the Jackknife"


population distribution
standard deviation
mean
variable of interest
sample
observations
sample standard deviation
sample mean
sampling distribution of the mean
standard error
mean of the sample means
sample mean

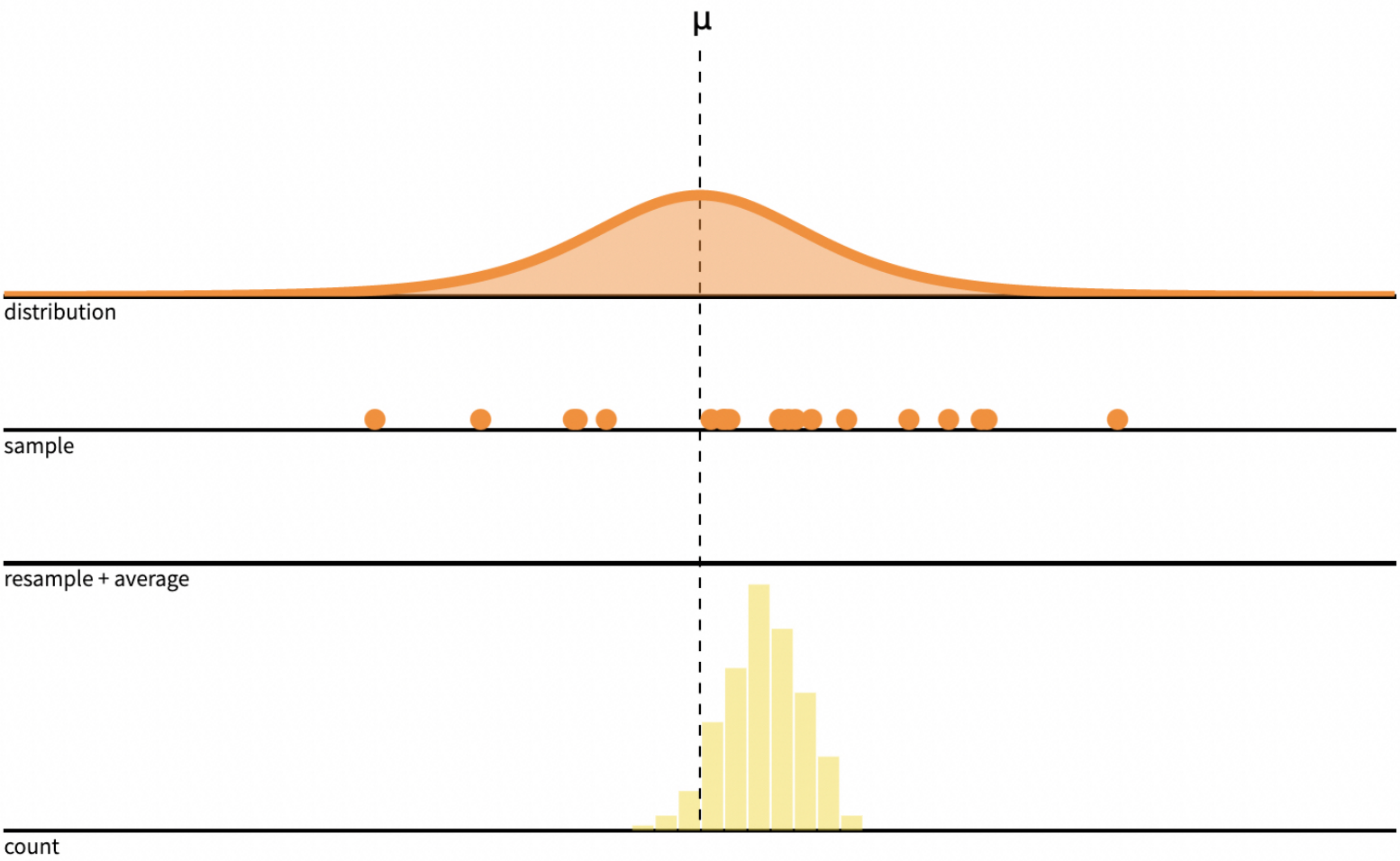
Claus Wilke



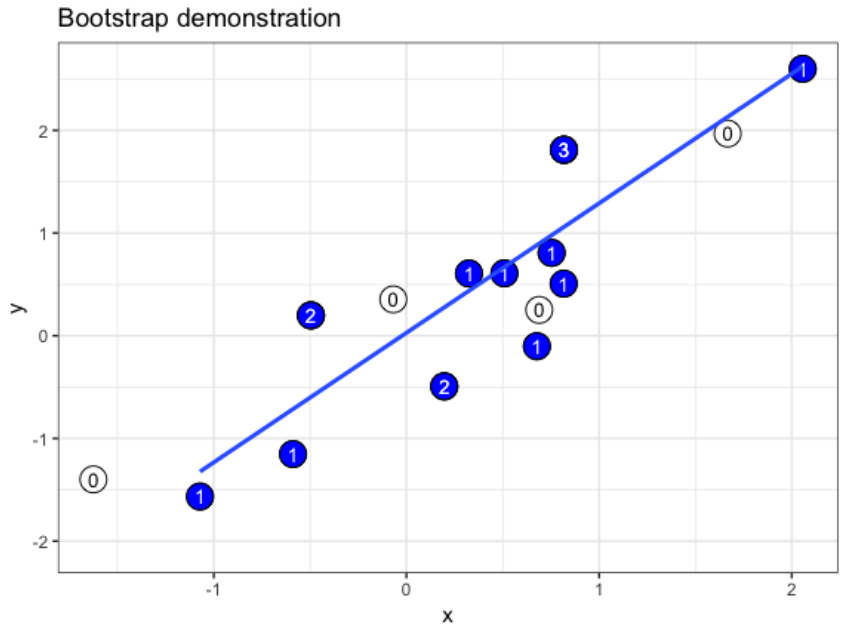
Source: Seeing Theory Website by Kunin, Guo, Devlin and Xiang



Source: Seeing Theory Website by Kunin, Guo, Devlin and Xiang

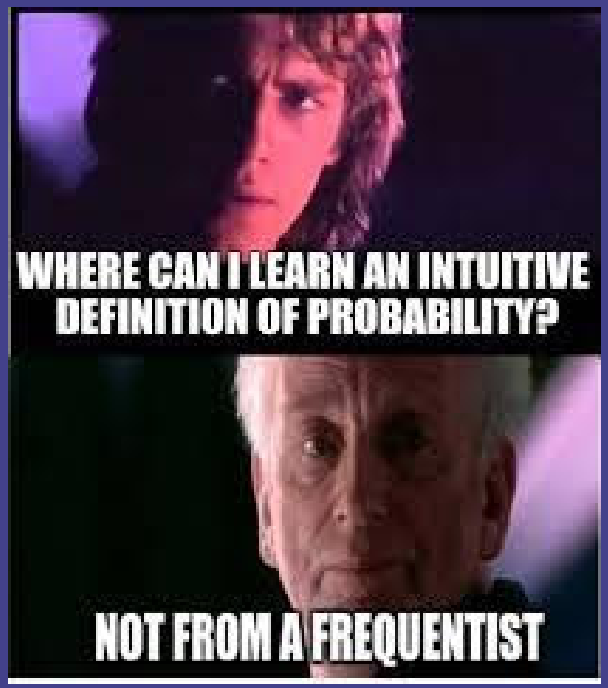


Source: Seeing Theory Website by Kunin, Guo, Devlin and Xiang

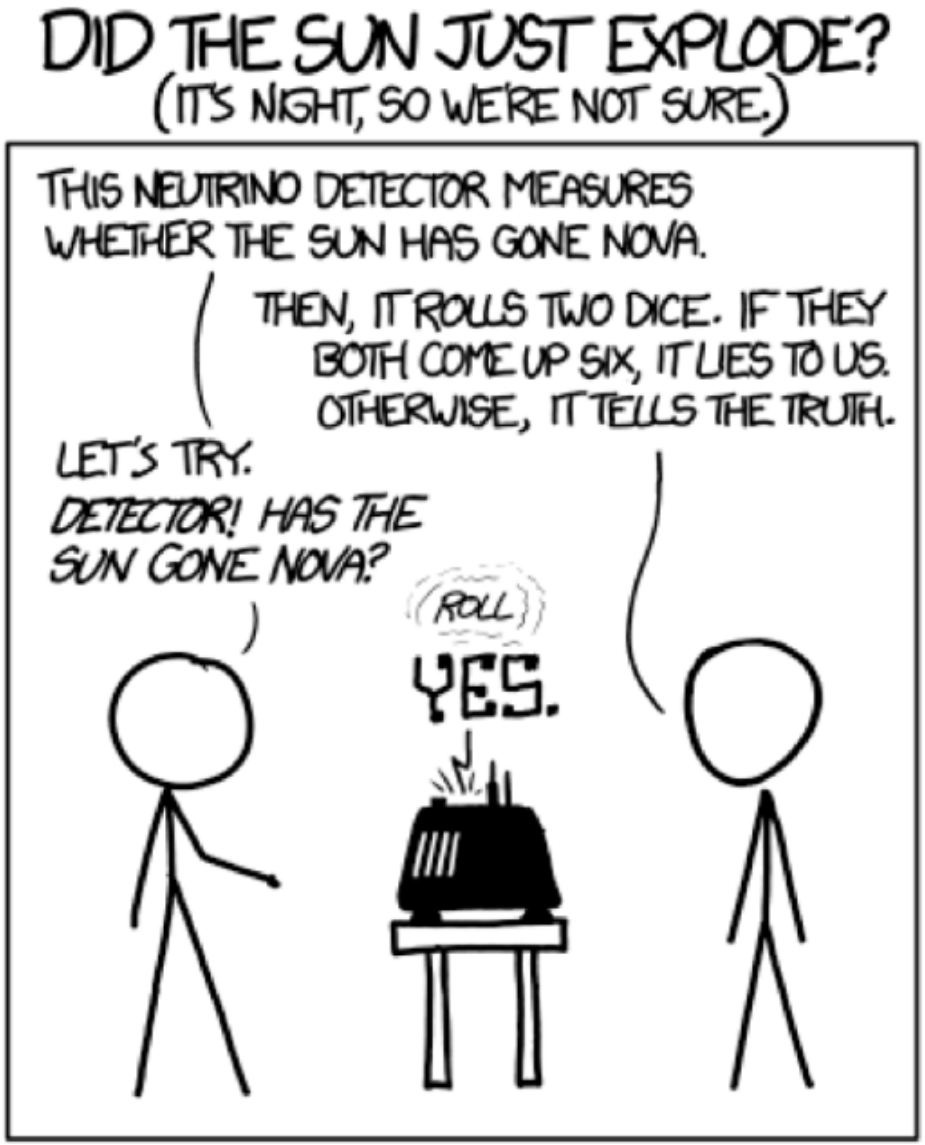


Source: Claus Wilke's `ungeviz` package

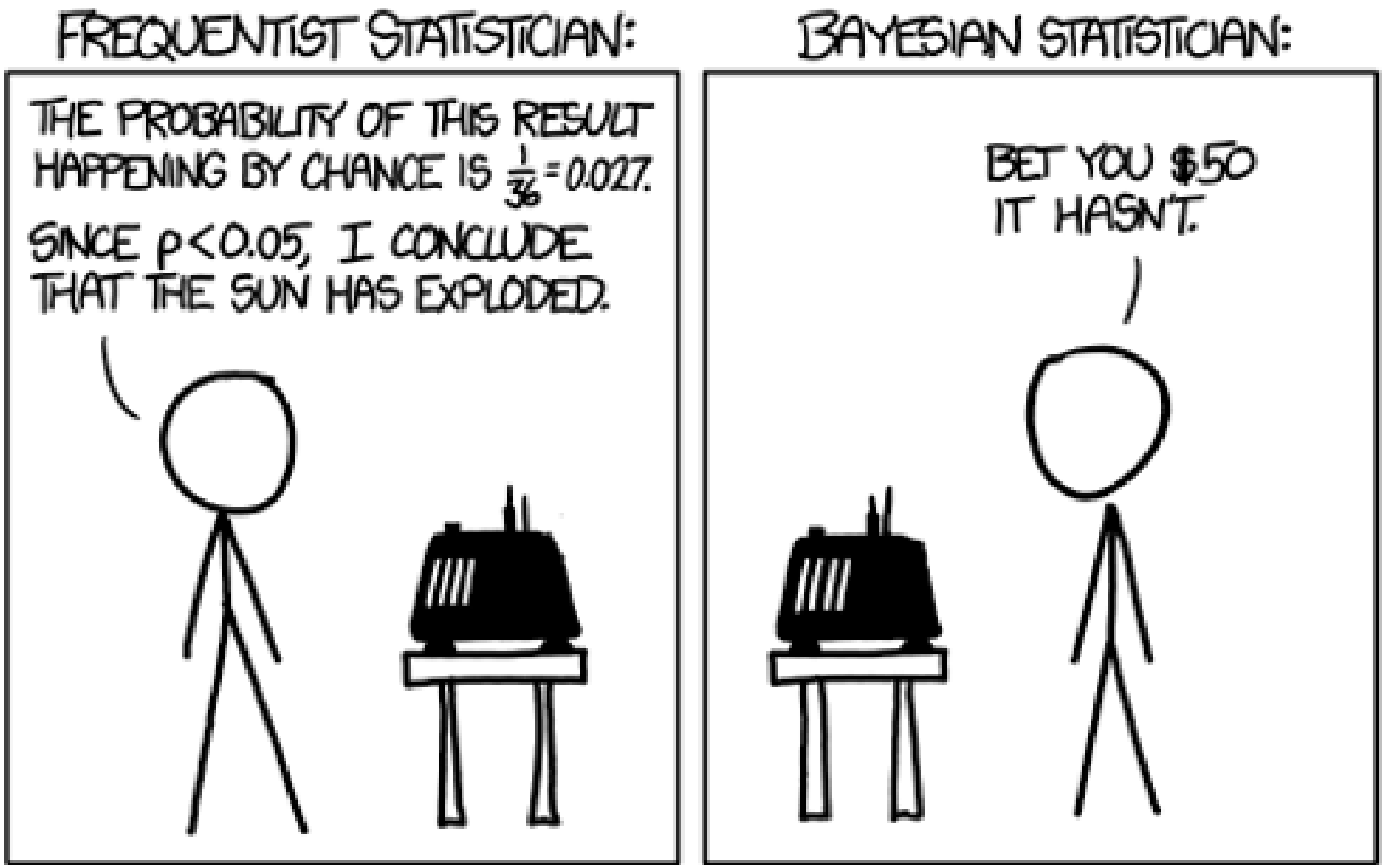
Bayesian analysis



@john_t_ormerod

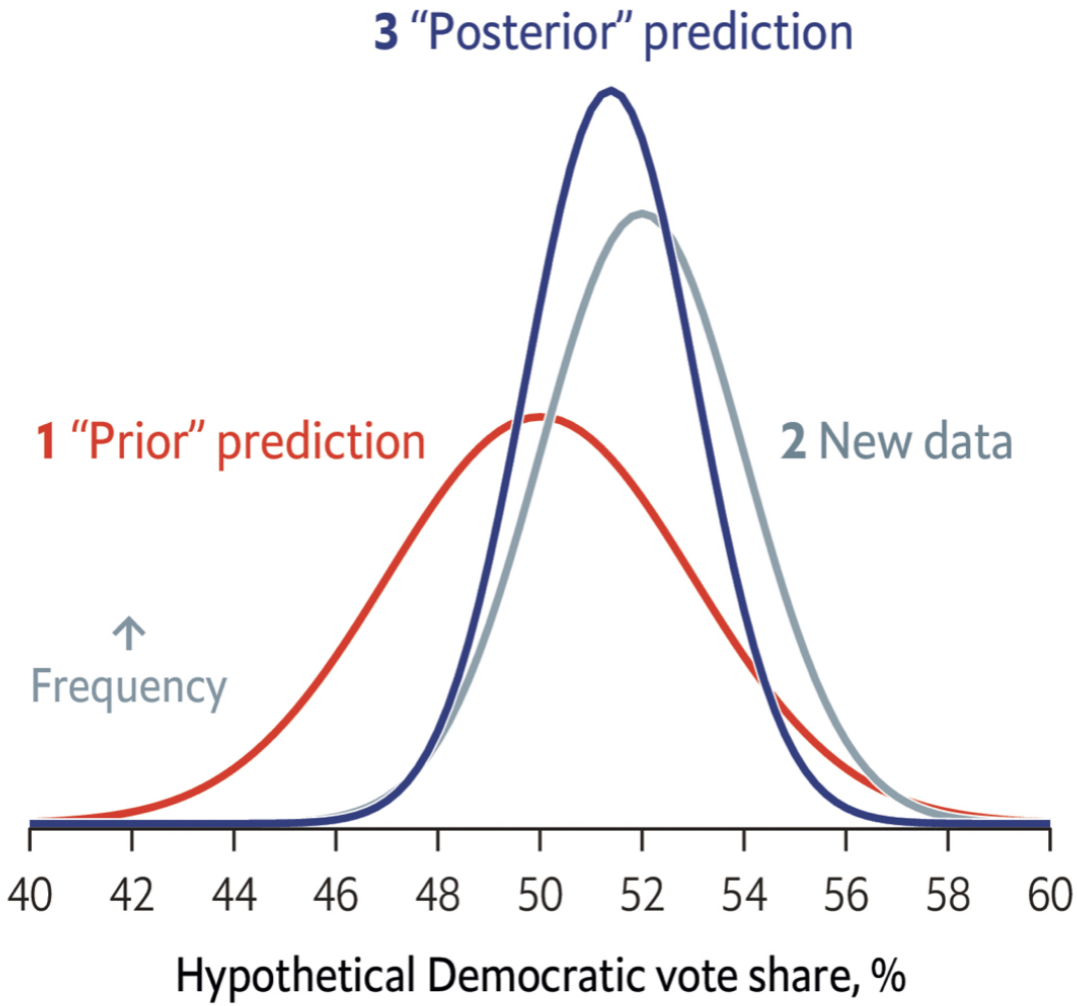

DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY.
DETECTOR! HAS THE SUN GONE NOVA?
(ROLL)
YES.

xkcd


FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS $\frac{1}{36} = 0.027$.
SINCE $p < 0.05$, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.

xkcd

Three steps of Bayesian inference



The Economist

Dr. William (Bill) Ribarsky, former UNCC Bank of America Endowed Chair

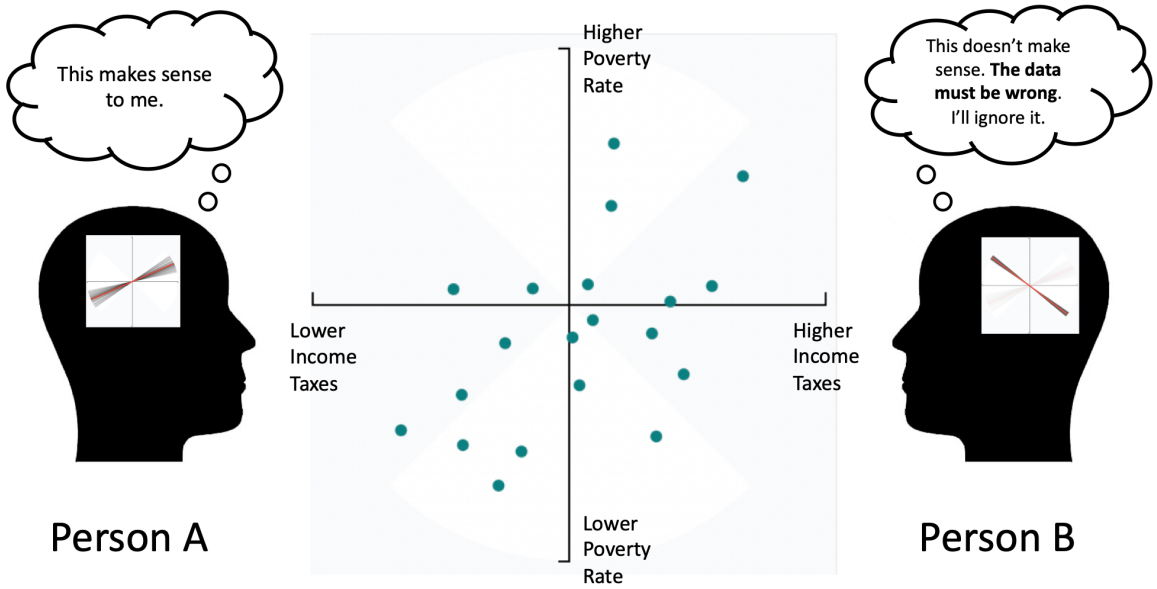
Research at UNCC Ribarsky Center for Visual Analytics (viscenter.uncc.edu)

Joint work with Wenwen Dou, Alireza Karduni, and Doug Markant

Correlation judgement

Karduni, Markant, Wesslen, and Dou (IEEE InfoVis 2020)

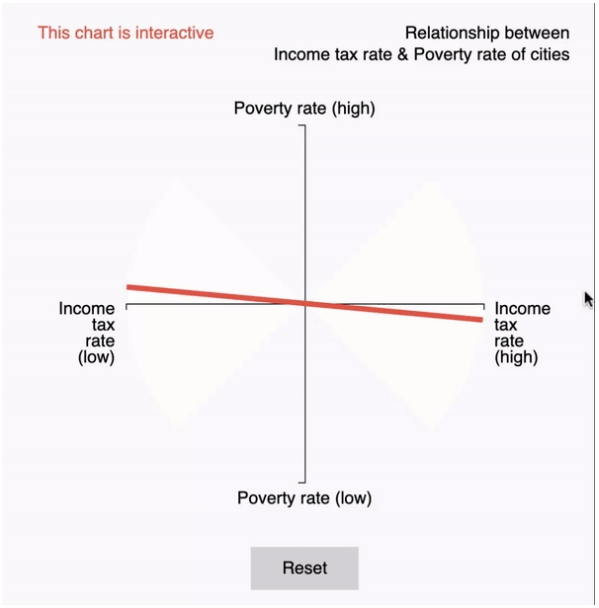
Prior beliefs



Karduni, Markant, Wesslen, and Dou (IEEE InfoVis 2020)

Line + Cone Elicitation

What is the relationship between income tax rate and poverty rate of cities?



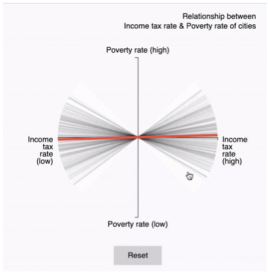
Built with D3.js (javascript)

Bayesian cognitive modeling

What is the relationship between income tax rate and poverty rate of cities?

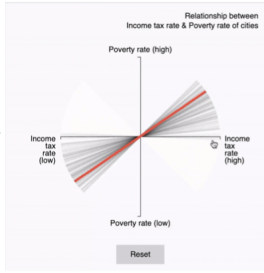
Experiments controlled for different visualizations, variable names, and data variance.

1. Elicited prior



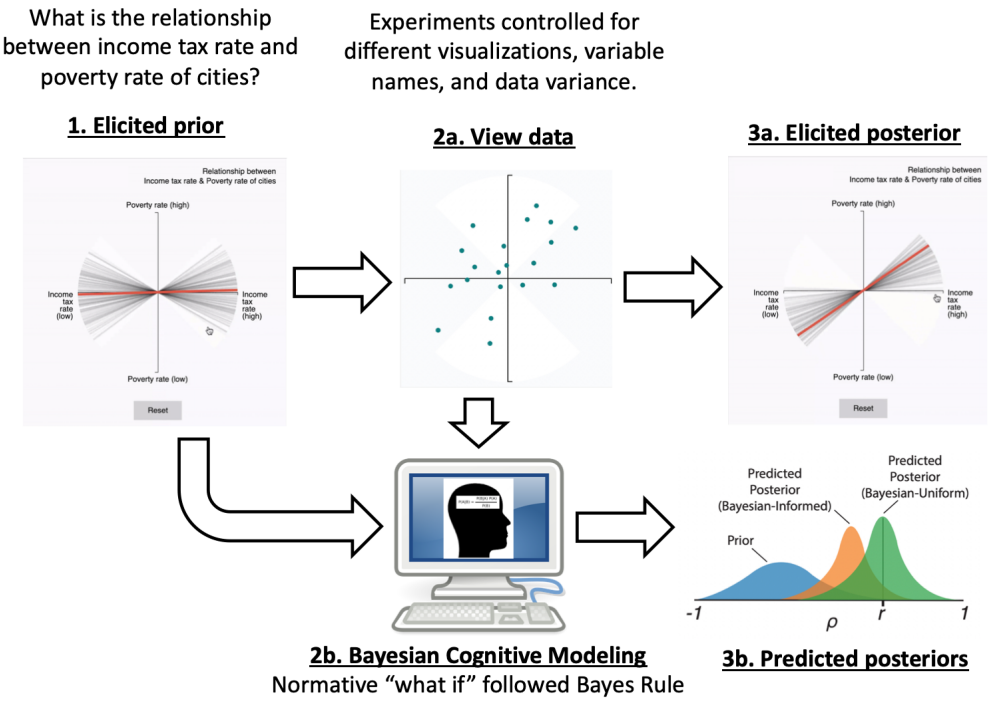
2a. View data

3a. Elicited posterior



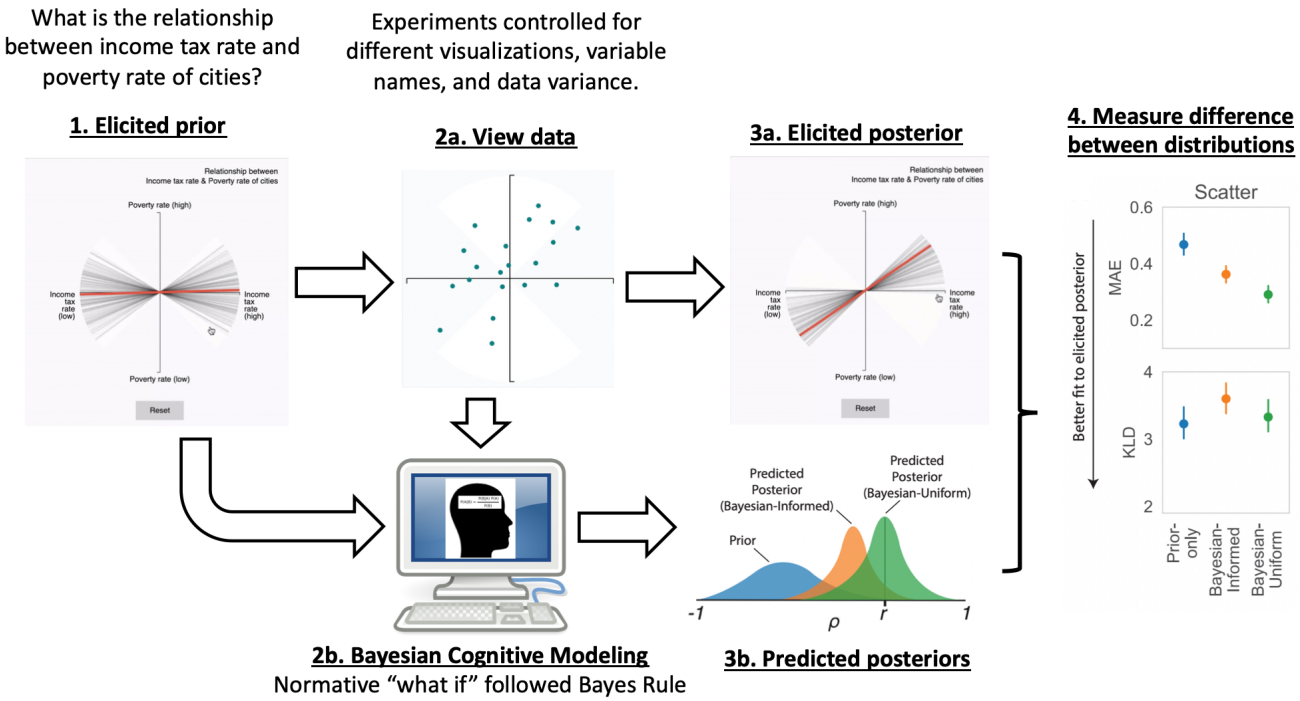
Karduni, Markant, Wesslen, and Dou (IEEE InfoVis 2020)

Bayesian cognitive modeling



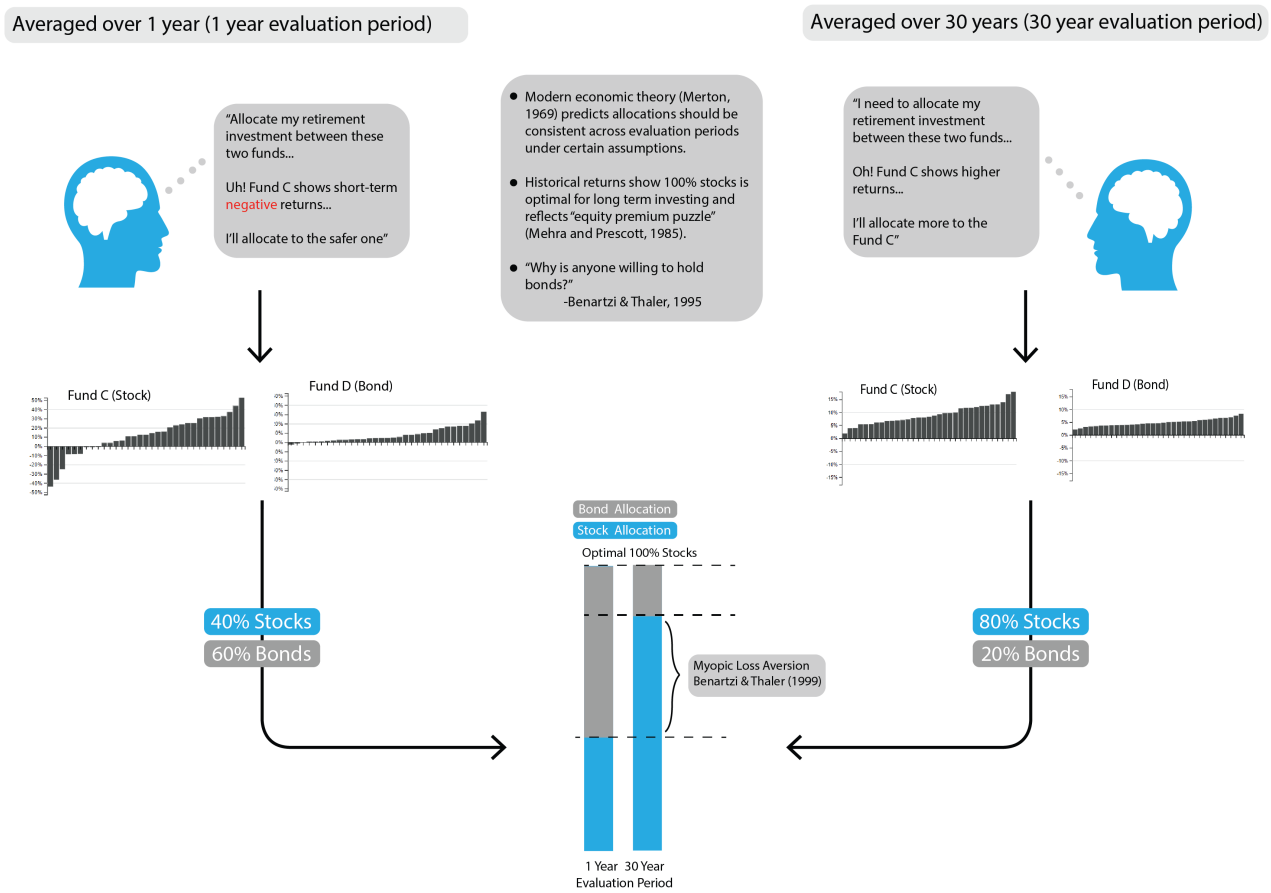
Karduni, Markant, Wesslen, and Dou (IEEE InfoVis 2020)

Bayesian cognitive modeling

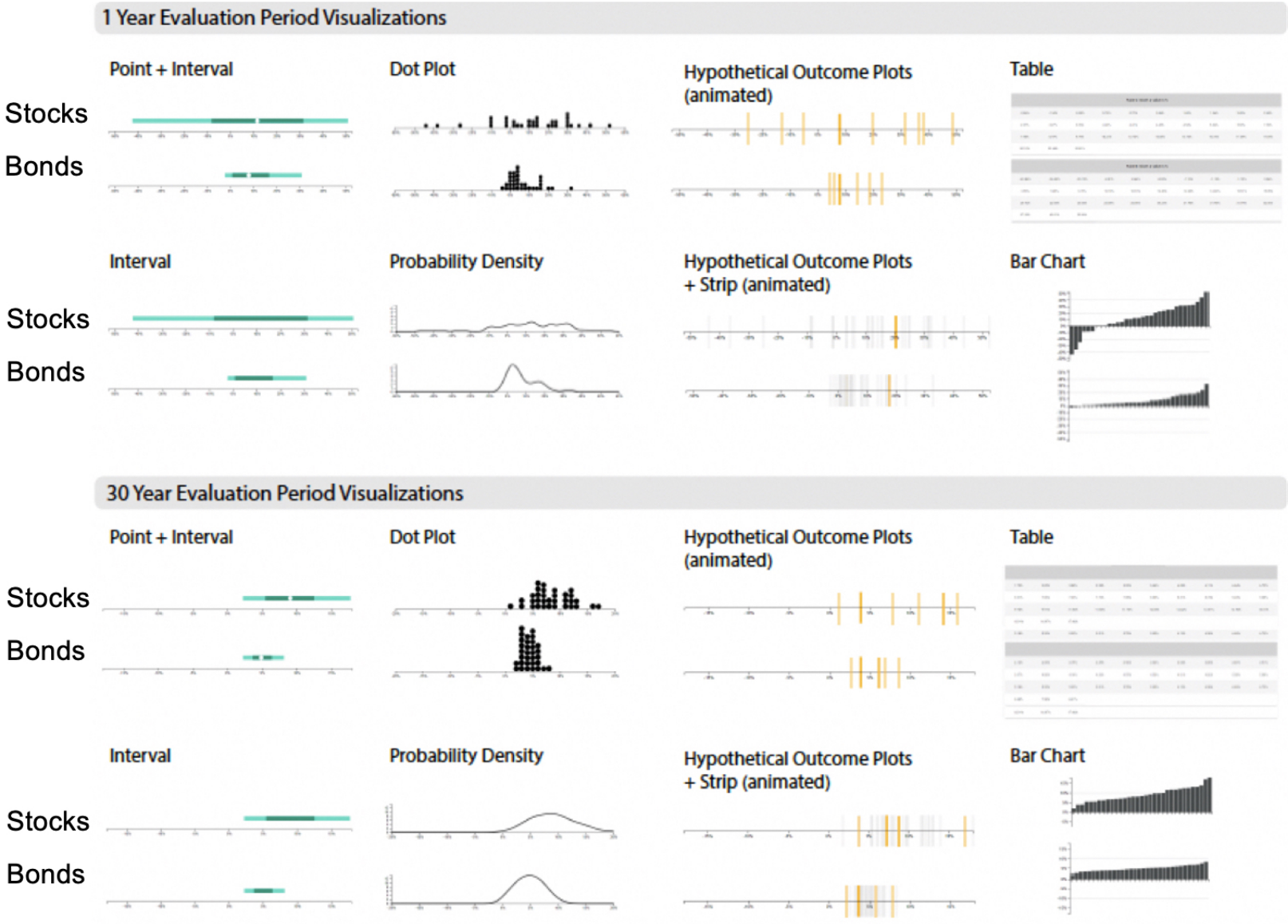


Karduni, Markant, Wesslen, and Dou (IEEE InfoVis 2020)

Investing for Retirement

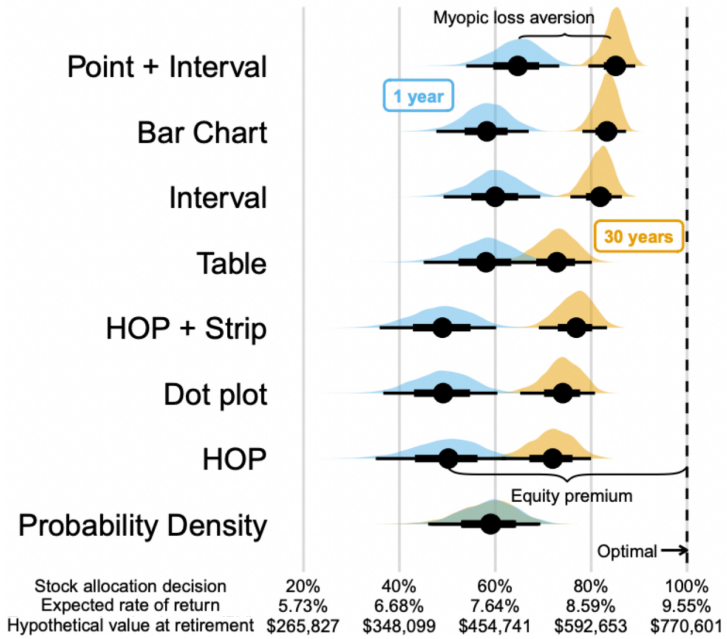


Wesslen, Karduni, Markant, and Dou (In Review)



Wesslen, Karduni, Markant, and Dou (In Review)

MTurk stock allocation decisions by visualization and evaluation period. Bayesian regression posterior means and 66% / 95% credible intervals



Wesslen, Karduni, Markant, and Dou (In Review)

R packages used `ggdist`, `tidybayes`, and `brms` packages

Final thoughts

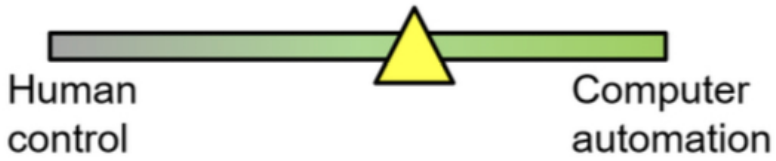


Figure 1: One-dimensional thinking suggest that designers must choose between human control and computer automation

Ben Schneiderman: Human-Centered Artificial Intelligence

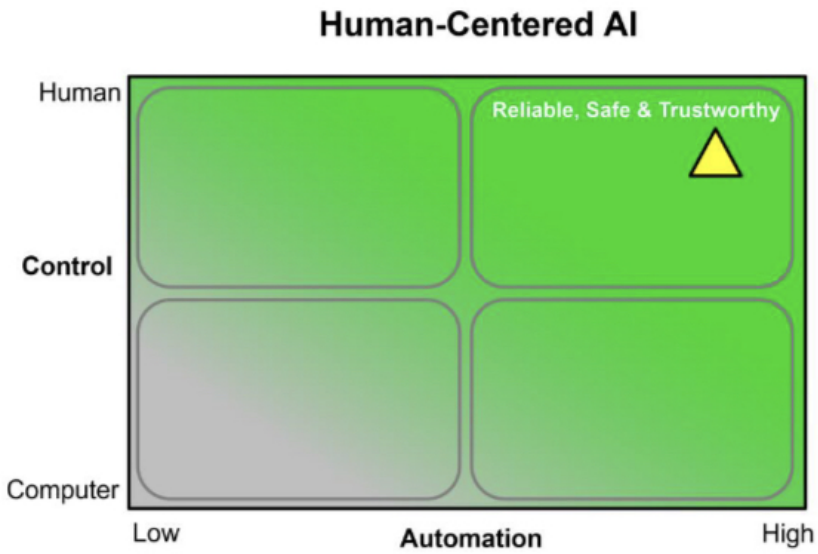


Figure 2: Two-dimensional framework with the goal of Reliable, Safe & Trustworthy, which is achieved by a high level of human control and high level of automation (yellow triangle).

Ben Schneiderman: Human-Centered Artificial Intelligence

Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty

Umang Bhatt[1,2], **Javier Antorán**[1], **Yunfeng Zhang**[3], **Q. Vera Liao**[3], **Prasanna Sattigeri**[3], **Riccardo Fogliato**[2,4], **Gabrielle Gauthier Melançon**[5], **Ranganath Krishnan**[6], **Jason Stanley**[5], **Omesh Tickoo**[6], **Lama Nachman**[6], **Rumi Chunara**[7], **Madhulika Srikumar**[2], **Adrian Weller**[1,8], **Alice Xiang**[2,9]
[1]University of Cambridge, [2]Partnership on AI, [3]IBM Research, [4]Carnegie Mellon University, [5]Element AI, [6]Intel Labs, [7]New York University, [8]The Alan Turing Institute, [9]Sony AI
usb20@cam.ac.uk

Abstract

Algorithmic transparency entails exposing system properties to various stakeholders for purposes that include understanding, improving, and contesting predictions. Until now, most research into algorithmic transparency has predominantly focused on explainability. Explainability attempts to provide reasons for a machine learning model's behavior to stakeholders. However, understanding a model's specific behavior alone might not be enough for stakeholders to gauge whether the model is wrong or lacks sufficient knowledge to solve the task at hand. In this paper, we argue for considering a complementary form of transparency by estimating and communicating the uncertainty associated with model predictions. First, we discuss methods for assessing uncertainty. Then, we characterize how uncertainty can be used to mitigate model unfairness, augment decision-making, and build trustworthy systems. Finally, we outline methods for displaying uncertainty to stakeholders and recommend how to collect information required for incorporating uncertainty into existing ML pipelines. This work constitutes an interdisciplinary review drawn from literature spanning machine learning, visualization/HCI, decision-making and fairness. We aim to encourage researchers and practitioners to measure, communicate, and use uncertainty as a form of transparency.

1 Introduction

Transparency in machine learning (ML) encompasses a wide variety of efforts to provide stakeholders, such as model developers and end users, with relevant information about how a ML model works (O'Neill 2018; Weller 2019; Bhatt et al. 2020). One form of transparency is procedural transparency, which provides information about model development (e.g., code release, model cards, dataset details) (Gebru et al. 2018; Raji and Yang 2019; Arnold et al. 2019; Mitchell et al. 2019). Another form is algorithmic transparency, which exposes information about a model's behavior to various stakeholders (Ribeiro, Singh, and Guestrin 2016; Sundararajan, Taly, and Yan 2017; Koh and Liang 2017). The ML community has mostly considered explainability, which attempts to provide reasoning for a model's behavior to stakeholders, as a proxy for algorithmic transparency. With this work, we seek to encourage researchers to study uncertainty as an alternative form of algorithmic transparency and practitioners to communicate uncertainty estimates to stakeholders. Uncertainty is crucial yet often overlooked in the context of ML-assisted, or automated, decision-making (Schum et al. 2014; Kochenderfer 2015). If well-calibrated and effectively communicated, uncertainty can help stakeholders understand when they should trust model predictions and help developers address fairness issues in their models (Zhang, Liao, and Bellamy 2020).

Uncertainty refers to our lack of knowledge about some outcome. As such, uncertainty will be characterized differently depending on the task at hand. In regression tasks, uncertainty is often expressed in terms of error bars, also known as confidence intervals. For example, when predicting the number of crimes in a given city, we could report that the number of predicted crimes is 943 ± 10, where "± 10" represents a 95% *confidence interval* (capturing two standard deviations on either side of the central, mean estimate). The smaller the interval, the more certain the model. In classification tasks, probability scores are often used to capture how confident a model is in a specific prediction. For example, a classification model may predict that a person is at a high risk for developing diabetes given a prediction of 85% chance of diabetes. Broadly, uncertainty in data-driven decision-making systems may stem from different sources and thus communicate different information to stakeholders (Hora 1996; Gal 2016). Aleatoric uncertainty is induced by inherent randomness (or noise) in the quantity to predict given input variables. Epistemic uncertainty can arise due to lack of sufficient data to learn our model precisely.

Why do we care?

We posit uncertainty can be useful for obtaining fairer models, improving decision-making, and building trust in automated systems. Throughout this work, we will use the following cancer diagnostics scenario for illustrative purposes: Suppose we are tasked with diagnosing individuals as having breast cancer or not, as in (Curtis et al. 2012; Dua and Graff 2017). Given categorical and continuous characteristics about an individual (medical test results, family medical history, etc.), we estimate the probability of an individual having breast cancer. We can then apply a threshold to classify them into high- or low-risk groups. Specifically, we have been tasked with building ML-powered tools to help three distinct audiences: doctors, who will be assisted in

Bhatt et al., 2021

"Uncertainty is **crucial** yet **often overlooked** in the context of ML-assisted, or automated, decision-making (Schum et al. 2014; Kochenderfer 2015)"

"We seek to **encourage researchers to study uncertainty** as an alternative form of algorithmic **transparency** and practitioners to communicate **uncertainty estimates** to stakeholders."

"We posit uncertainty can be useful for obtaining **fairer models**, **improving decision-making**, and **building trust** in automated systems."

Be thoughtful about uncertainty

- Speed vs precision

- Don't automatically sweep under the rug; learn and experiment!

Think beyond your college (Frequentist) stats class

- Computational methods (bootstrap and Bayesian) can be very helpful

- New uncertainty open source packages emerging and easier to use

Better communicate uncertainty, better human-AI collaboration?

- Uncertainty is typically overlooked in AI/ML

- Likely very important in future for trust and transparency in AI

Thank you and questions!

- UNCC Collaborators: Wenwen Dou, Alireza Karduni, Doug Markant
- Vis Researchers: Matthew Kay, Jessica Hullman, Lace Padilla, Alex Kale, Michael Correll, among others!

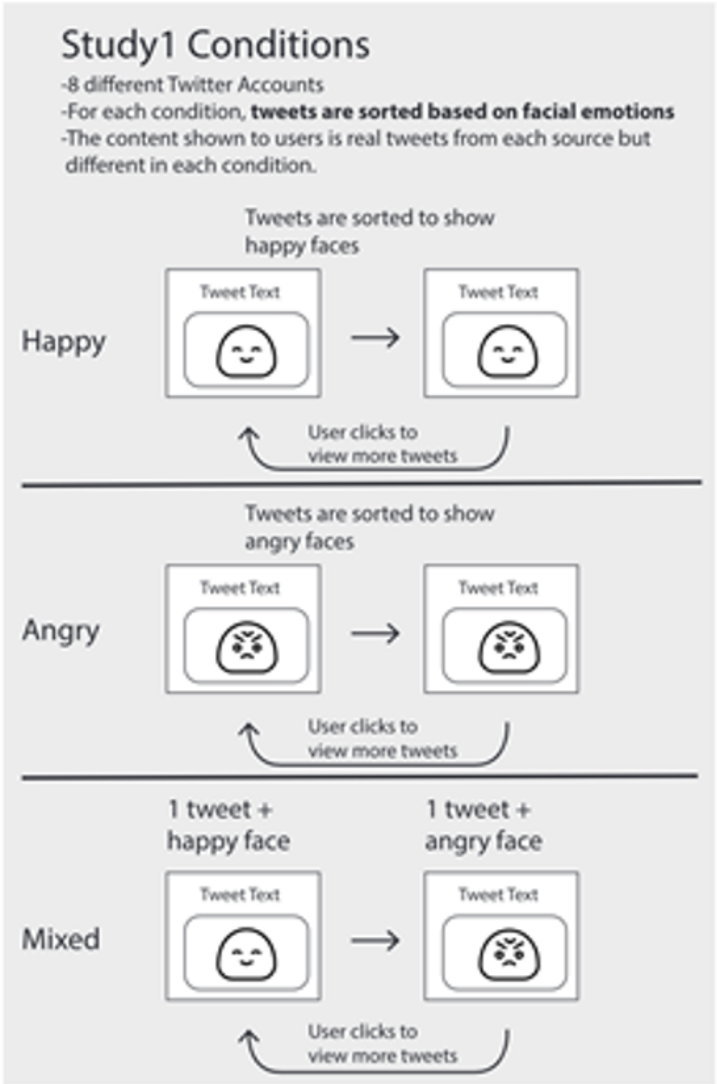
Packages / tools

	R	Python
Bootstrap	`rsample`	`bootstrapped` `DABEST`
Bayesian Statistics	`rstan` `brms` `tidybayes`	`PyMC3`
Uncertainty Visualizations	`ggdist`	`uncertainty-toolbox`

Helpful references

- Padilla, Kay, and Hullman (2020): Uncertainty Visualization
- Hullman et al. (2019): In Pursuit of Error: A Survey of Uncertainty Visualization Evaluation
- Bhatt et al. (2021): Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty

Appendix



Karduni, Wesslen, Markant, and Dou (In Review)

Alireza Karduni
@don_kordeone

Replying to @don_kordeone

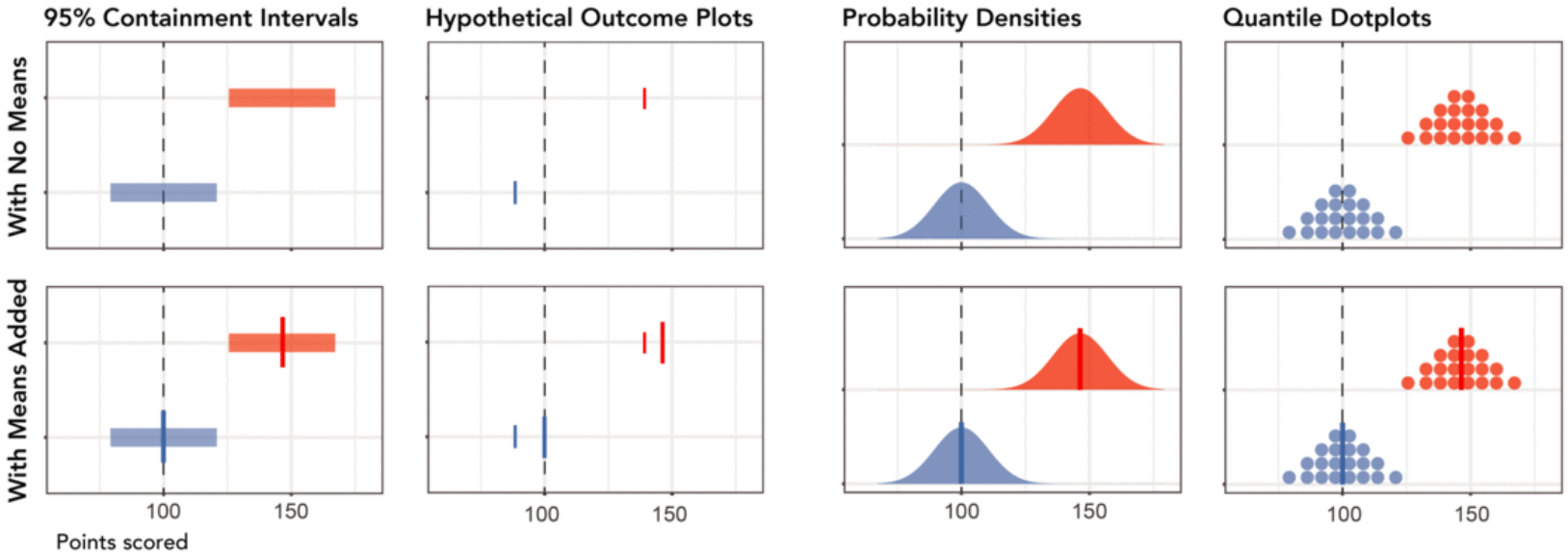
We created an interface to elicit users' judgments and uncertainties using a visual technique. Users would 1) assess how biased tweets from a specific (anonymized) source are and 2) rate how credible that source is. They would also write down their rationale

5/16

GIF

8:52 PM · Mar 1, 2021

2 1 Copy link to Tweet

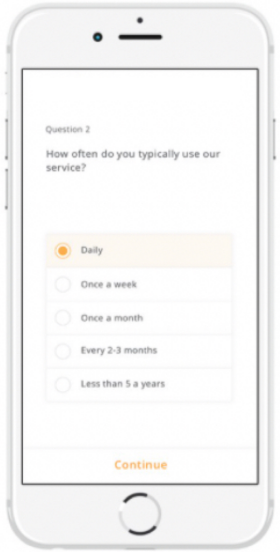


Kale, Kay, and Hullman, "Visual Reasoning Strategies for Effect Size Judgments and Decisions", IEEE Vis 2020 Best Paper

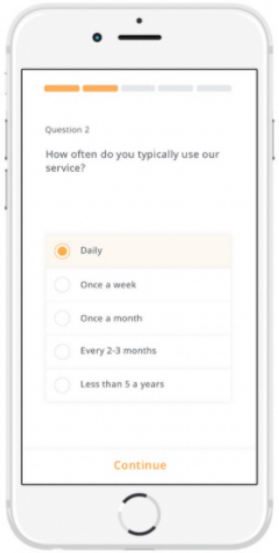
A / B Evaluation in UI/UX

Which design reduces survey "drop-off"? Villar, Callegaro, and Yang (2013)

No progress bar (control)



"Fast-to-slow" progress bar



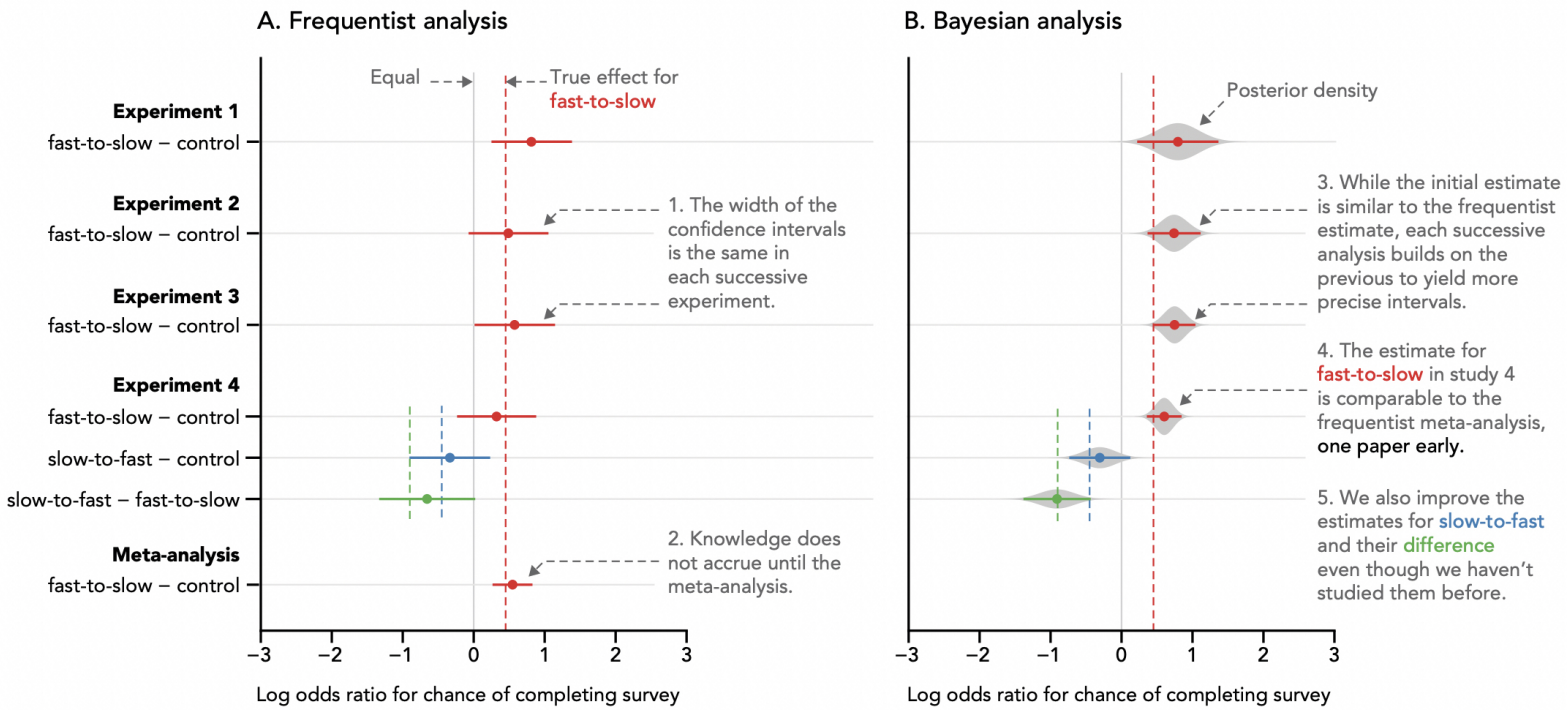


Figure 1. Forest plots of effects from the frequentist (A) and Bayesian (B) analyses applied to one of our simulated worlds with 100 participants per condition.

Kay, Nelson, and Hekler, "Research-Centered Design of Statistics: Why Bayesian Statistics Better Fit the Culture and Incentives of HCI", ACM CHI 2016 Best Honorable Mention