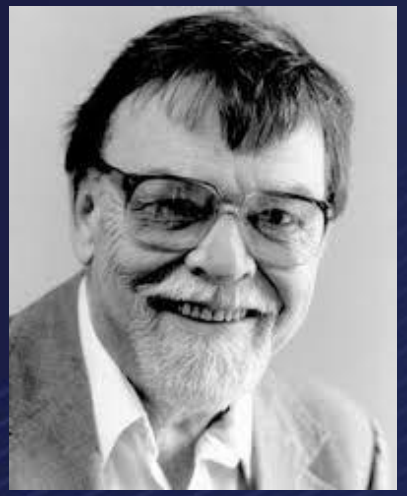
FAIR CONFERENCE 2019

How to Measure Risk with Limited and Messy Data: Overcoming the Myths

DOUGLAS HUBBARD
President,
Hubbard Decision Research

Historical Quantitative Models vs. Expert Intuition Alone

FAIR CONFERENCE 2019



Paul Meehl assessed 150 studies comparing experts to statistical models in many fields (sports, prognosis of liver disease, etc.).

"There is no controversy in social science which shows such a large body of qualitatively diverse studies coming out so uniformly in the same direction as this one."

Philip Tetlock tracked a total of over 82,000 forecasts from 284 political experts in a 20-year study covering elections, policy effects, wars, the economy and more.

"It is impossible to find any domain in which humans clearly outperformed crude extrapolation algorithms, less still sophisticated statistical ones."

So Why Don't We Use More Quantitative Methods?

FAIR CONFERENCE 2019

Have you heard (or said) any of these?

"We don't have sufficient data…"

"There is too much error and bias in the data for it to be worth the effort to gather it..."

"Each situation is too unique and complex to apply scientific analysis of historical data..."

"There are so many factors affecting this, this measurement alone tells us nothing..."

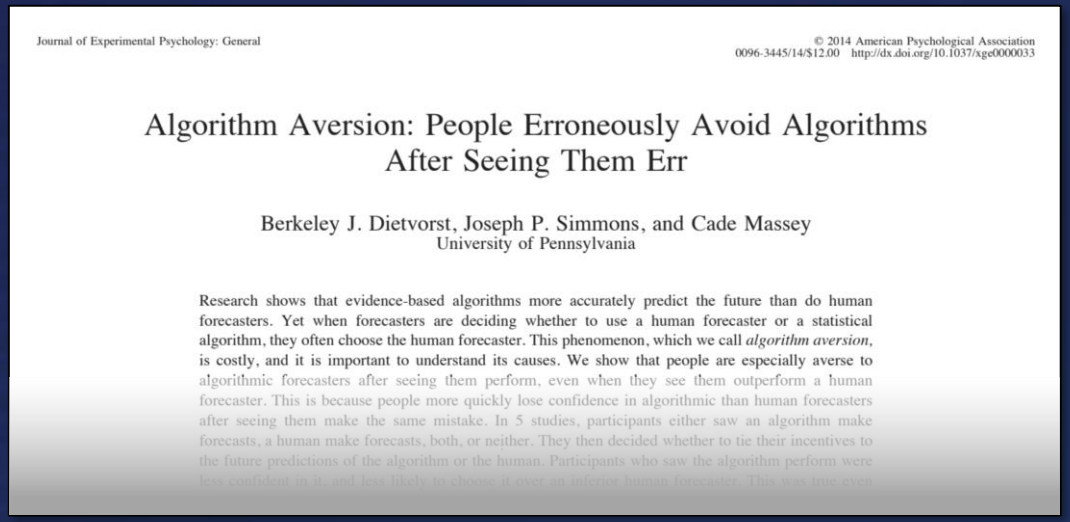
The implied (and unjustified) conclusion from each of these is….

"…therefore we are better off relying on our experience."

Aversion to Algorithms

FAIR CONFERENCE 2019

- There is a double standard when evaluating algorithms vs. human experts.
- Research shows that even when algorithms perform better than a human expert overall, people penalize the algorithm for an error more than the human.



Journal of Experimental Psychology: General

0096-3445/14/$12.00 http://dx.doi.org/10.1037/xge0000033

Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err

Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey
University of Pennsylvania

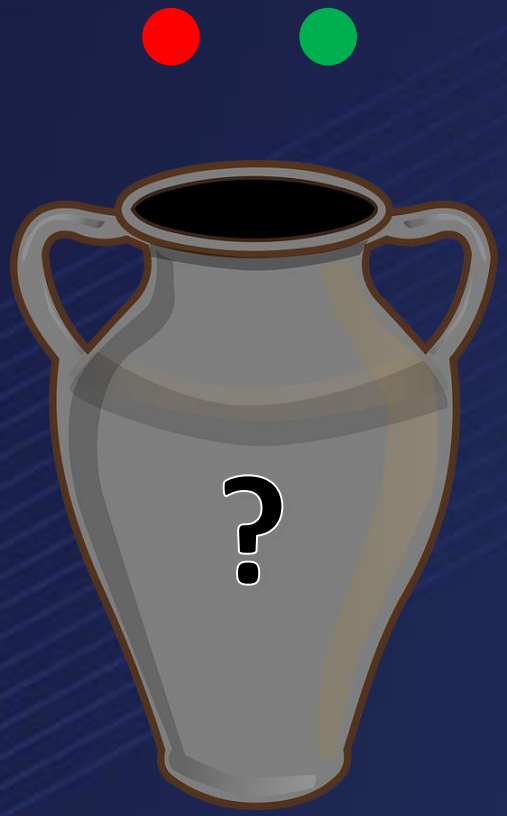
Research shows that evidence-based algorithms more accurately predict the future than do human forecasters. Yet when forecasters are deciding whether to use a human forecaster or a statistical algorithm, they often choose the human forecaster. This phenomenon, which we call *algorithm aversion*, is costly, and it is important to understand its causes. We show that people are especially averse to algorithmic forecasters after seeing them perform, even when they see them outperform a human forecaster. This is because people more quickly lose confidence in algorithmic than human forecasters after seeing them make the same mistake. In 5 studies, participants either saw an algorithm make forecasts, a human make forecasts, both, or neither. They then decided whether to tie their incentives to the future predictions of the algorithm or the human. Participants who saw the algorithm perform were less confident in it, and less likely to choose it over an inferior human forecaster. This was true even

Don’t make the classic “Beat the Bear” fallacy.

Exsupero Ursus

Methods of Measurement

FAIR CONFERENCE 2019



THE *URN OF MYSTERY* PROBLEM

- There is a warehouse full of urns
- Each urn is filled with over a million marbles, each of which are red or green
- The proportion of red marbles in each urn is unknown – it could be anything between 0% and 100% and all possibilities are equally likely.

Questions:

- If you randomly select a single marble from a randomly selected urn, what is the chance it is red?
- If the marble you draw is red, what is the chance the majority of marbles are red?
- If you draw 8 marbles and all are green, what is the chance that the next one you draw will be red?
- How is this like cybersecurity?

Intuition About Sample Information Is Often Wrong

FAIR CONFERENCE 2019

- Cybersecurity experts are not immune to widely held misconceptions about probabilities and statistics – especially if they vaguely remember some college stats.
- These misconceptions lead many experts to believe they lack data for assessing uncertainties or they need some ideal amount before anything can be inferred.

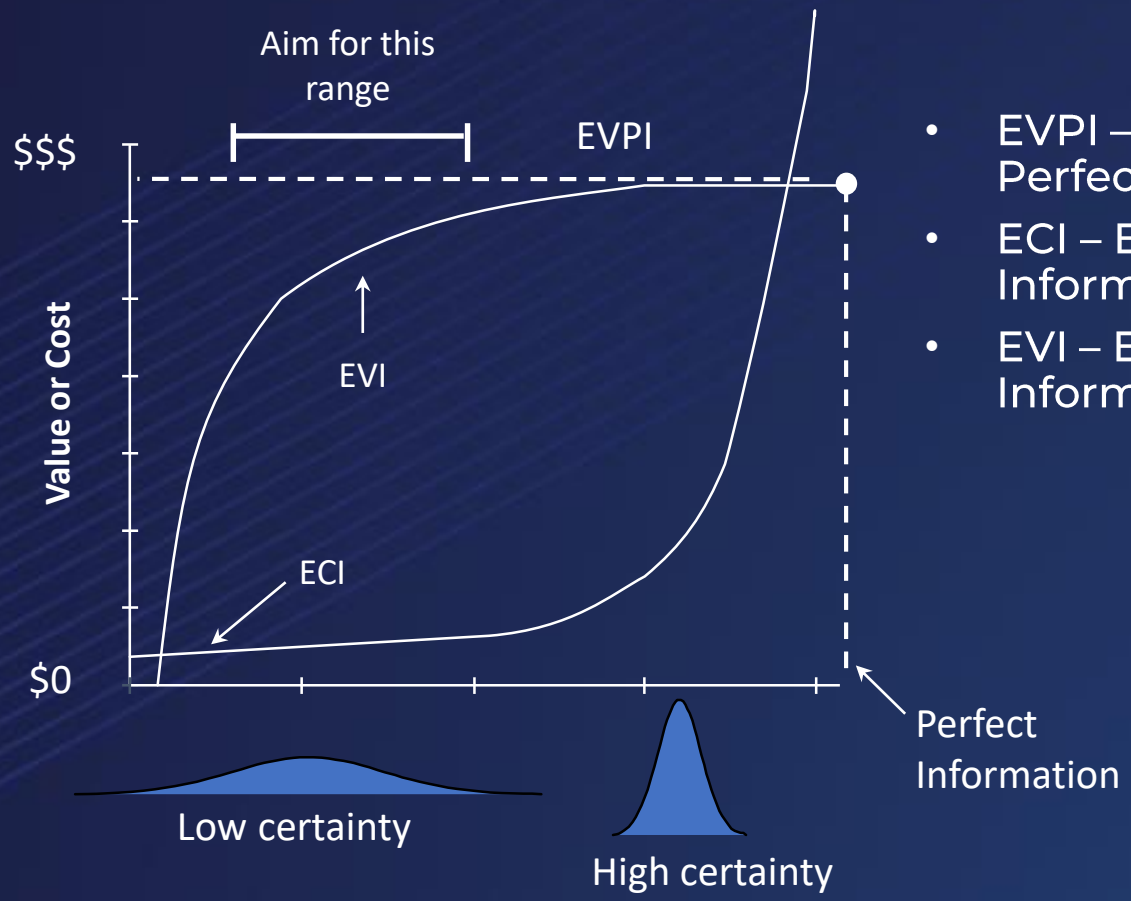
"Our thesis is that people have strong intuitions about random sampling…these intuitions are wrong in fundamental respects...[and] are shared by naive subjects and by trained scientists"

Amos Tversky and Daniel Kahneman, *Psychological Bulletin,* 1971

Increasing Cost and Value Information

FAIR CONFERENCE 2019

If we can model uncertainty about decisions, we can compute the value of information.



- EVPI – Expected Value of Perfect Information
- ECI – Expected Cost of Information
- EVI – Expected Value of Information

The Reference Class Fallacy

FAIR CONFERENCE 2019

Definition: The assumption that each situation is so unique nothing can be learned from other experiences. (i.e., the Mount St. Helens Fallacy)

- A reference class is a population from which you draw observations of events to determine their frequency.
- Your “reference class” is much larger than you.
- You can start by making as few assumptions as possible. A “robust prior” assumes as little prior knowledge as possible.

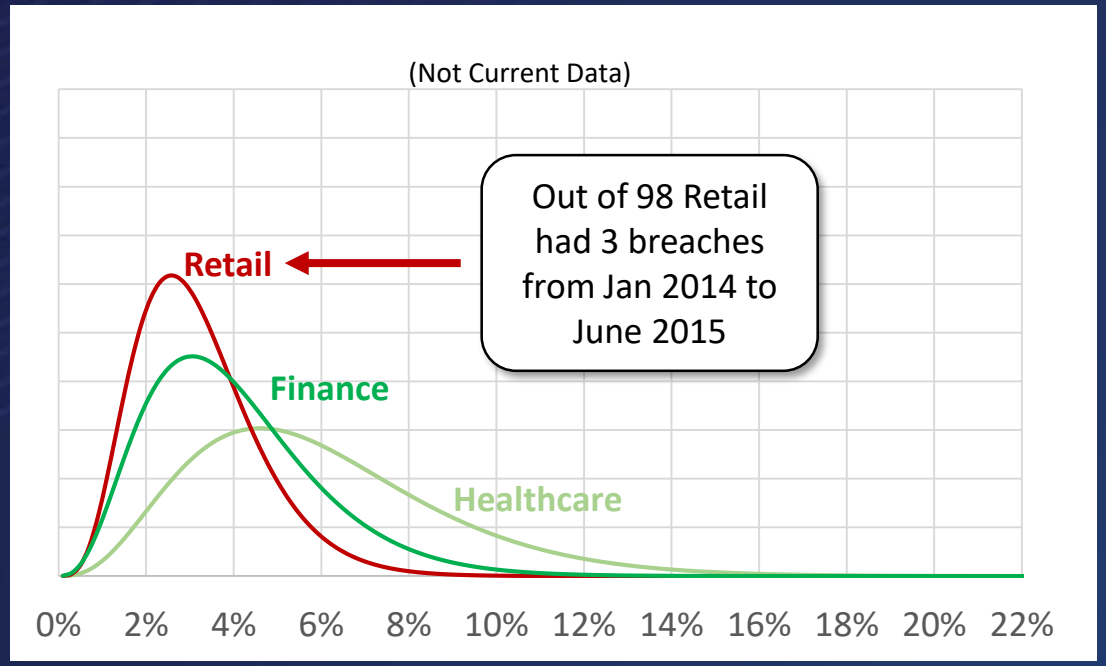
Given a population of reference class, like company-years, where some number of events occurred

Chance X will happen next year =(1+events)/(2+size of reference class)

Statistics Needs Less Data Than You Think

FAIR CONFERENCE 2019

- You have relatively few examples of major, reported breaches in each industry.
- There is a statistical method for estimating the frequency of breaches based on small samples. This is the “beta” distribution and it is provided in Excel as “=betadist(proportion, hits, misses)”.
- Spreadsheet for this is at www.howtomeasureanything.com/cybersecurity.

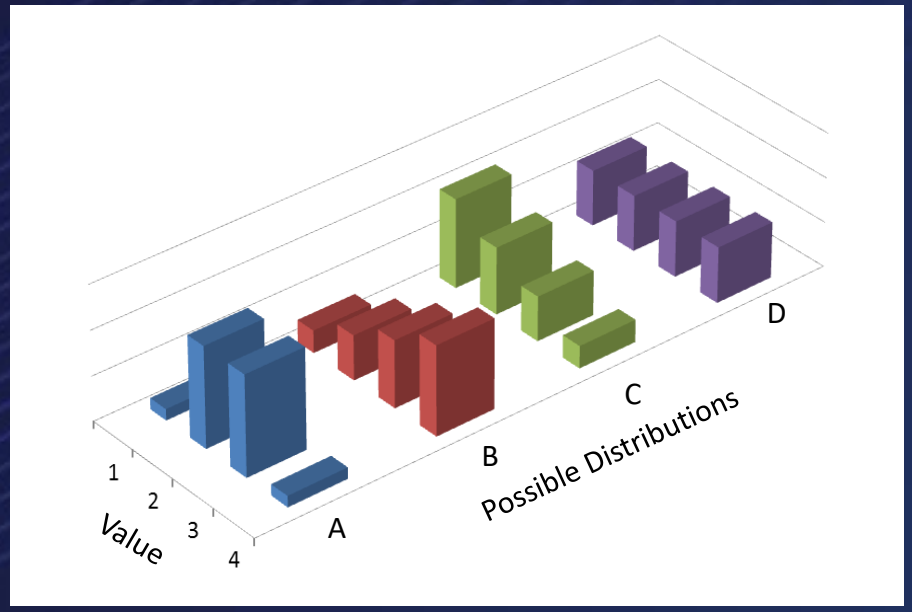


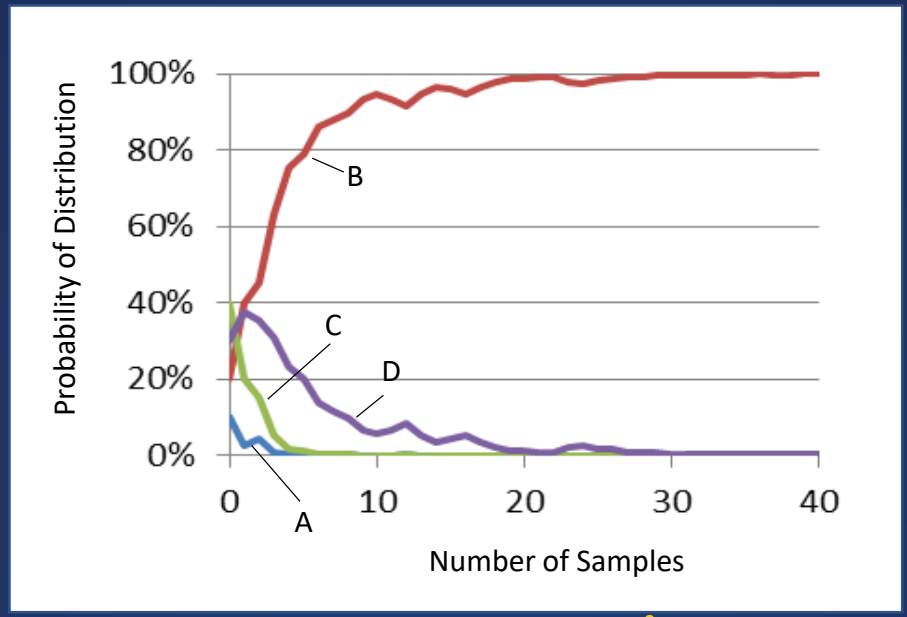
Annual Breach Frequency per Organization

“Low Resolution” Example of Bayes for Ranges

FAIR CONFERENCE 2019

- You can apply Bayes to estimating parameters (e.g. mean, median, etc.) of populations with continuous values (income, leisure time per week, etc.)
- Just identify possible population distribution types and the probability of each type – resolution can be as fine as you like.
- There are also similar solutions for regression models and controlled experiments.

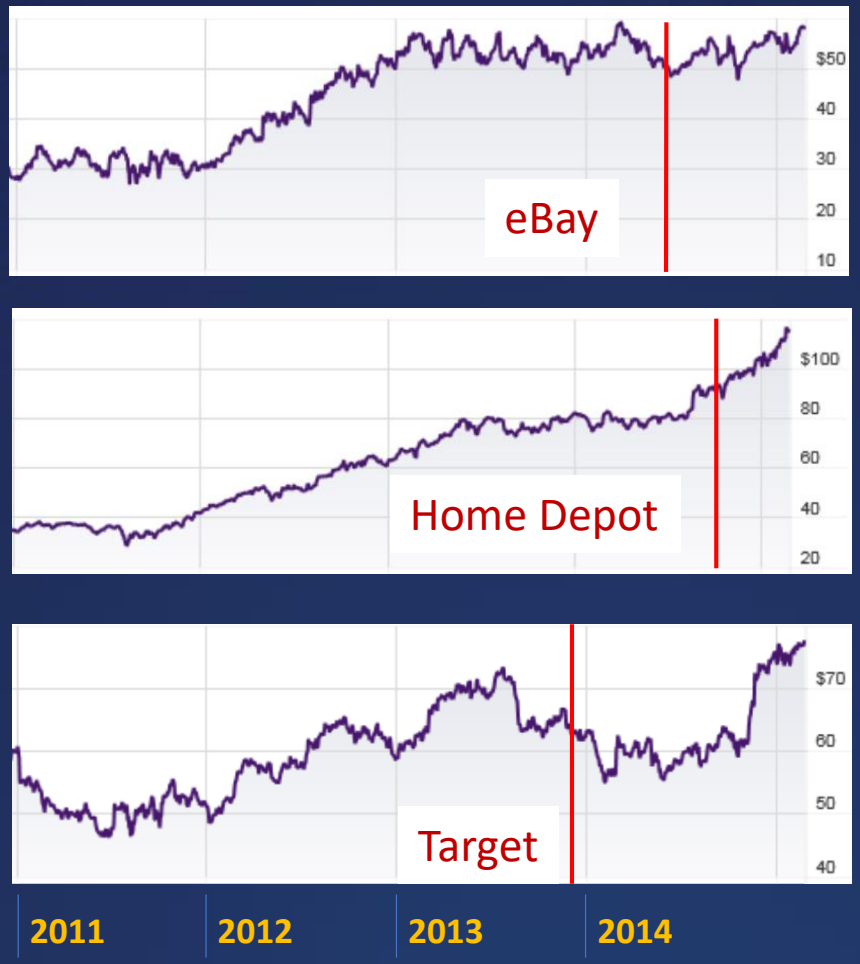




Measurement Challenge: Reputation Damage

FAIR CONFERENCE 2019

- One of the perceived most difficult measurements in cybersecurity is damage to reputation.
- Trick: *There is no such thing as a "secret" damage to reputation!*
- How about comparing stock prices after incidents? (That's all public!)
- So what is the *REAL* damage?
  - Legal liabilities
  - Customer outreach
  - "Penance" projects (security overkill)
- The upshot, damage to reputation actually has available information and easily observable measured costs incurred to *avoid* the bigger damages!



Bayesian Methods: Node Probability Tables

FAIR CONFERENCE 2019

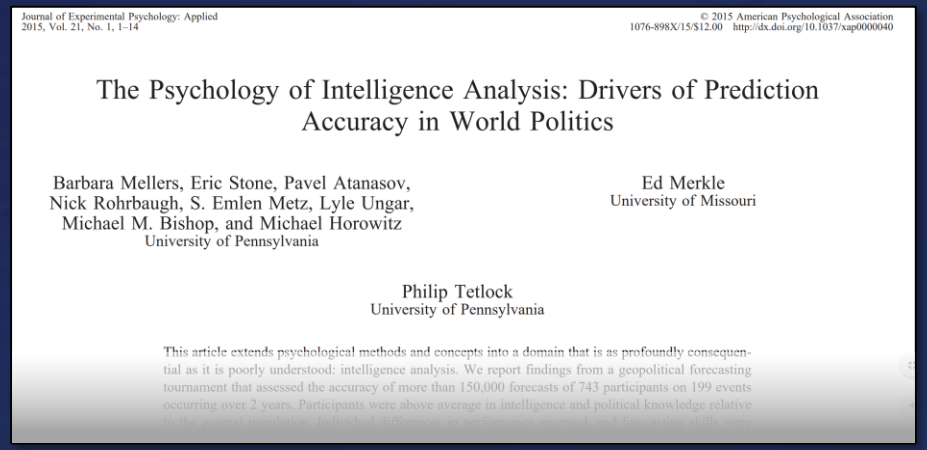
Node Probability Table				
Condition				
A	B	C	D	P(E|A,B,C,D)
Yes	Yes	Yes	Yes	86%
No	Yes	Yes	Yes	40%
Yes	No	Yes	Yes	1%
No	No	Yes	Yes	2%
Yes	Yes	No	Yes	75%
No	Yes	No	Yes	40%
Yes	No	No	Yes	2%
No	No	No	Yes	1%
Yes	Yes	Yes	No	90%
No	Yes	Yes	No	35%
Yes	No	Yes	No	2%
No	No	Yes	No	1%
Yes	Yes	No	No	80%
No	Yes	No	No	40%
Yes	No	No	No	2%
No	No	No	No	2%

- Conditional probabilities with combinations of conditions are recorded with an NPT.
- With more than a few conditions and conditions that are more than binary, it will become unwieldly.
- Recent models we created would have had thousands of rows.

Improving Expert Forecasts

FAIR CONFERENCE 2019

- Tetlock also looked at what improved *forecasting.*
- He tracked 743 individuals who made at least 30 forecasts each over a 2-year period.
- He determined factors that made the biggest difference in the performance of forecasting.



Journal of Experimental Psychology: Applied
2015, Vol. 21, No. 1, 1–14

1076-898X/15/$12.00 http://dx.doi.org/10.1037/xap0000040

The Psychology of Intelligence Analysis: Drivers of Prediction Accuracy in World Politics

Barbara Mellers, Eric Stone, Pavel Atanasov, Nick Rohrbaugh, S. Emlen Metz, Lyle Ungar, Michael M. Bishop, and Michael Horowitz
University of Pennsylvania

Ed Merkle
University of Missouri

Philip Tetlock
University of Pennsylvania

This article extends psychological methods and concepts into a domain that is as profoundly consequential as it is poorly understood: intelligence analysis. We report findings from a geopolitical forecasting tournament that assessed the accuracy of more than 150,000 forecasts of 743 participants on 199 events occurring over 2 years. Participants were above average in intelligence and political knowledge relative

Probabilistic Training

- Subjects were trained in basic inference methods, using reference classes, and avoiding common errors and biases.

Teams and Belief Updating

- Teams deliberated more and individuals were willing to update beliefs based on new information.

Selecting the Best

- Brains matter. Both topic expertise and overall IQ were the best predictors of performance.

Quantifying Your Current Uncertainty

FAIR CONFERENCE 2019

- Decades of studies show that most managers are statistically "overconfident" when assessing their own uncertainty.
- Studies also show that measuring *your own* uncertainty about a quantity is a general skill that can be taught with a ***measurable*** improvement.
- Training can "calibrate" people so that of all the times they say they are 90% confident, they will be right 90% of the time.
- HDR has calibrated over 1,000 people in the last 20 years – 85% of participants reach calibration within a half-day of training.

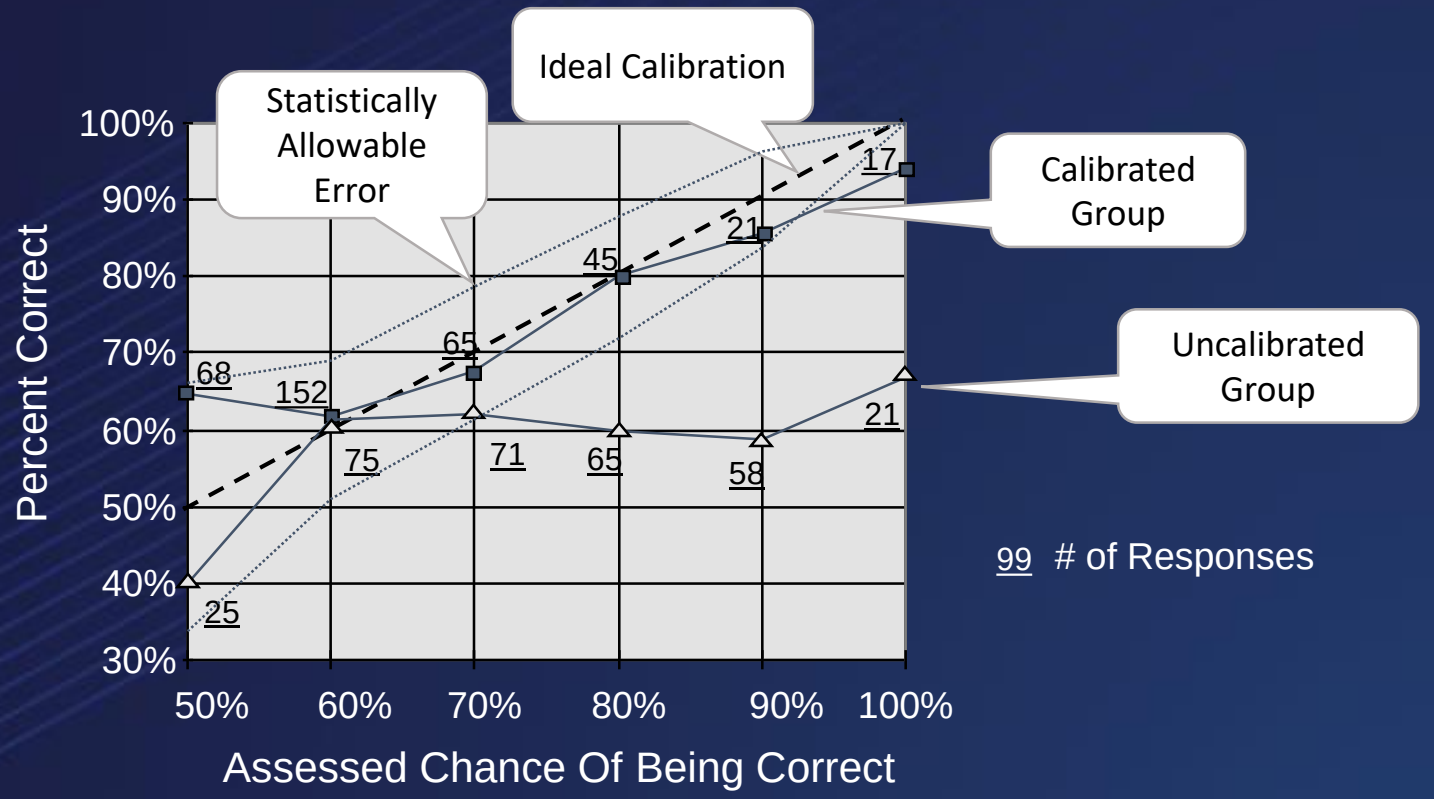
"Overconfident professionals sincerely believe they have expertise, act as experts and look like experts. You will have to struggle to remind yourself that they may be in the grip of an illusion."

Daniel Kahneman, Psychologist, Economics Nobel

Training Experts to Give Calibrated Probabilities

FAIR CONFERENCE 2019

Training can "calibrate" people so that of all the times they say they are 90% confident, they will be right 90% of the time.



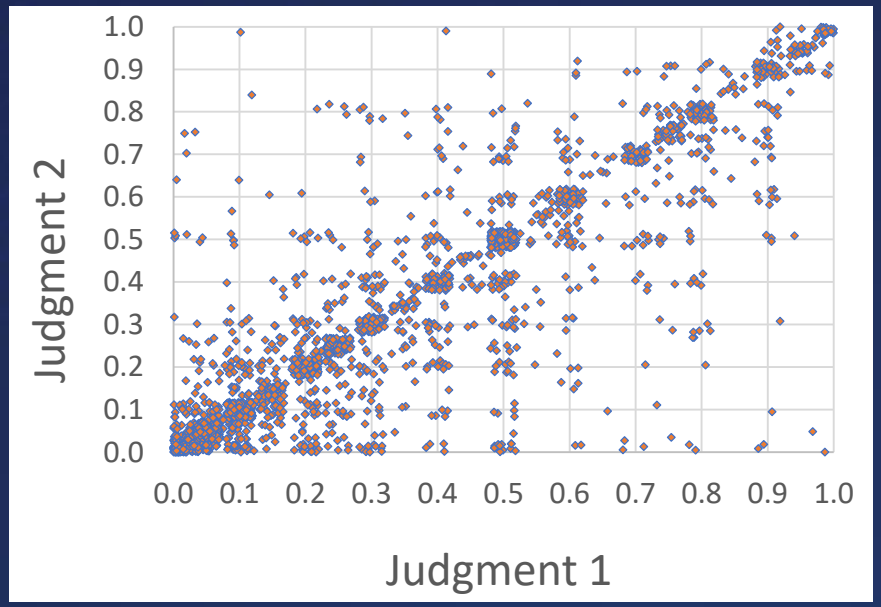
Source: Hubbard Decision Research, Giga Information Group

Calibrating Expert Consistency

FAIR CONFERENCE 2019

- We have gathered over 30,000 individual estimates of probabilities of cyber events from Subject Matter Experts (SMEs).
- Unknown to the SMEs, these estimates included over 2,000 duplicate scenarios pairs.

Comparison of 1st to 2nd Estimates of Cyber risk judgements by same SME



21% of variation in expert responses are explained by *inconsistency.*

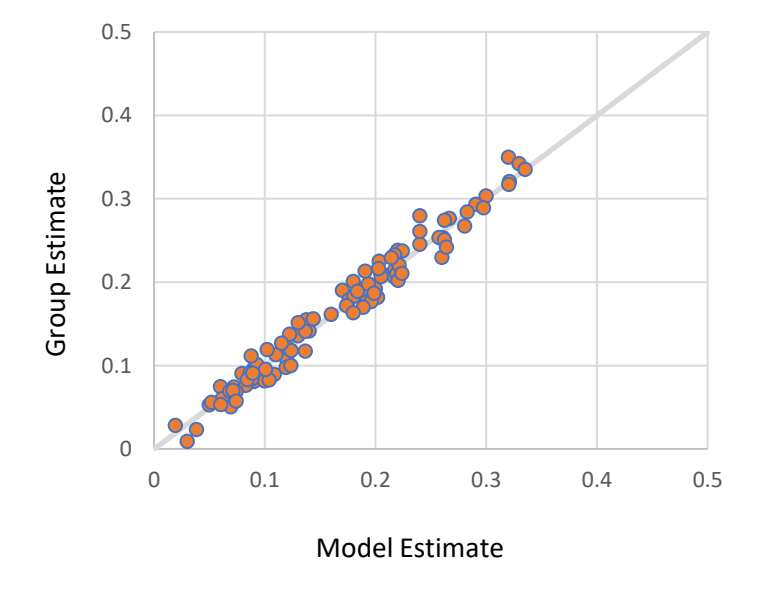
(79% are explained by the actual information they were given)

Modeling Group Estimates of IT Security Event Likelihood

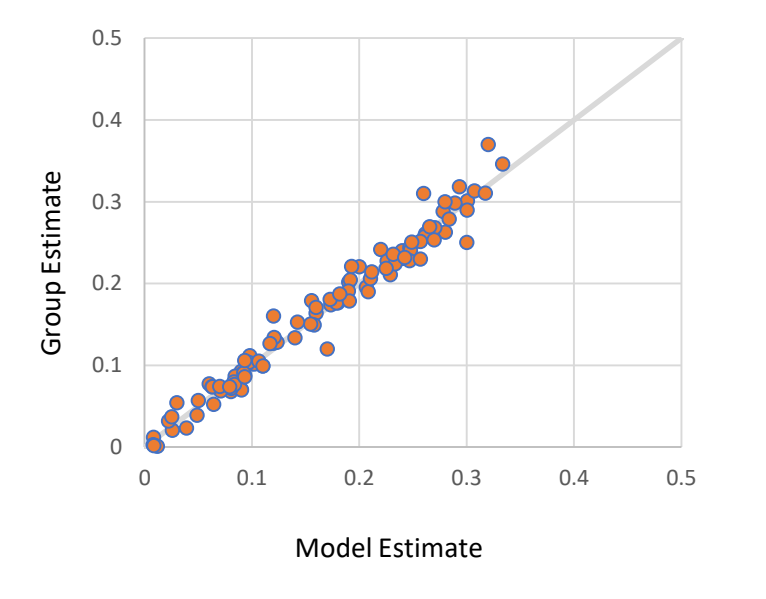
FAIR CONFERENCE 2019

Examples of Models vs. Group Averages: Probabilities of different security events happening in the next 12 months for various systems prior to applying particular controls.

Confidentiality Breach Resulting In Fines

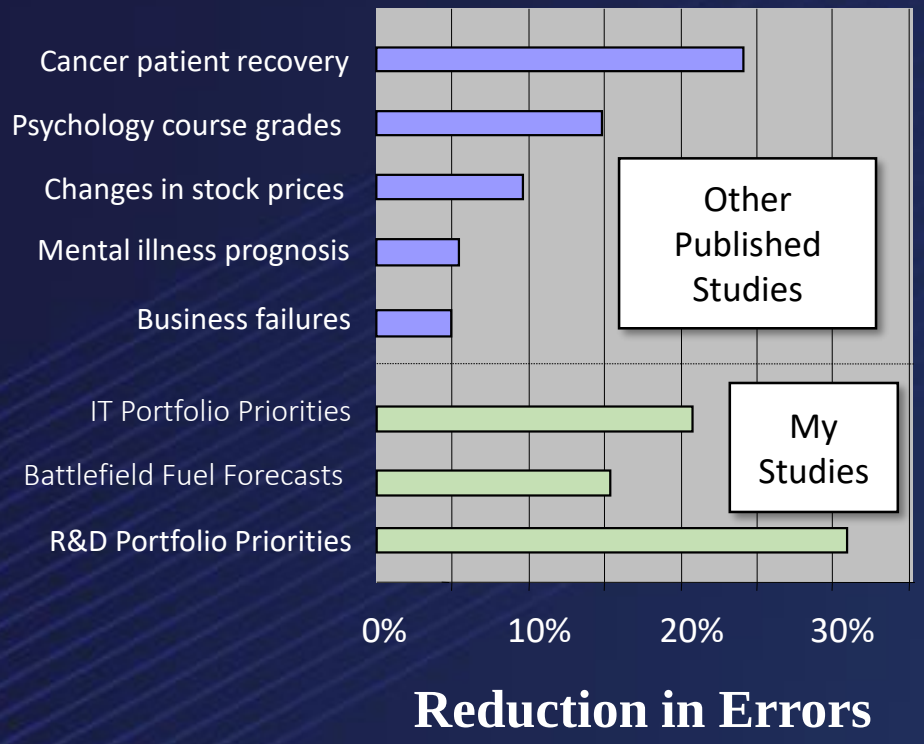


Internal Unauthorized Access Resulting In Productivity Loss



- The models created produce results which closely match the group's average.
- A large portion of the model error is due to judge inconsistency.
- This nearly eliminates the inconsistency error.

Effects of Removing Inconsistency Alone



- A method of improving expert estimates of various quantities was developed in the 1950's by Egon Brunswik.
- He called it the "Lens Method".
- It has been applied to several types of problems, including expert systems, with consistently beneficial results.

Summary

FAIR CONFERENCE 2019

It's Been Measured Before	• Don't forget publically available data. All "reputation" data is public by definition. Important topics have often been measured already.
You Have More Data Than You Think	• Define a reference class – don't commit the reference class fallacy.
You Need Less Data Than You Think	• Question your intuition about how and whether messy and incomplete data is useful – you may be surprised.

Questions?

FAIR CONFERENCE 2019

Contact:

Doug Hubbard
Hubbard Decision Research
dwhubbard@hubbardresearch.com
www.hubbardresearch.com
630-858-2788