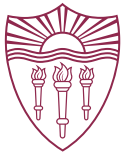


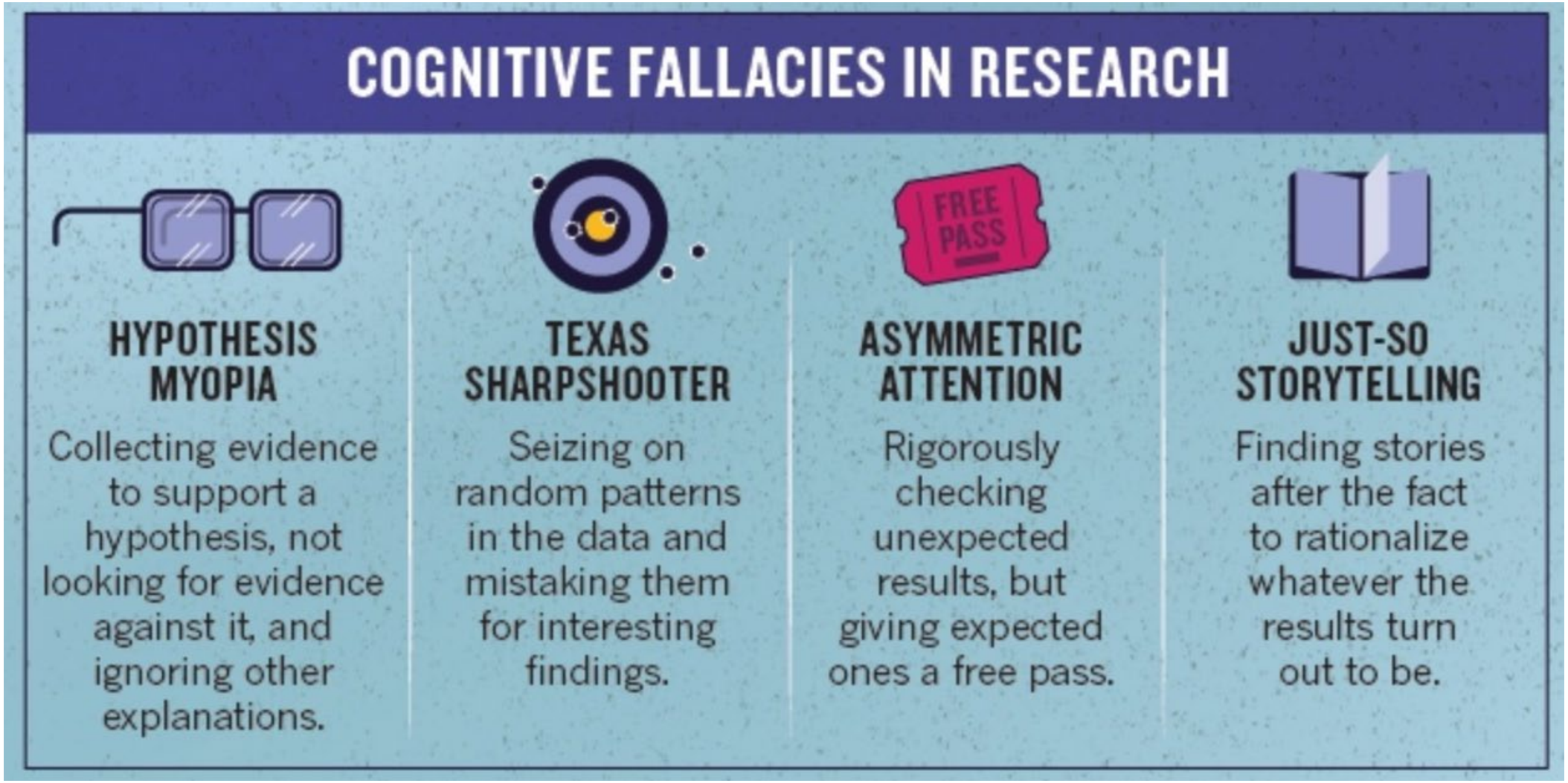
REPRODUCIBILITY IN DATA SCIENCE

OR HOW NOT TO SUCCUMB TO COGNITIVE BIASES

Kristina Lerman

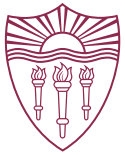
April 26, 2021

How data scientists fool themselves



https://www.nature.com/news/how-scientists-fool-themselves-and-how-they-can-stop-1.18517

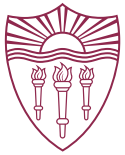
Hypothesis myopia



- Researchers tend to ask questions that give 'yes' answers if their favorite hypothesis is true

- They fixate on collecting evidence to support just one hypothesis
- Neglect to look for evidence against it
- Fail to consider other explanations.

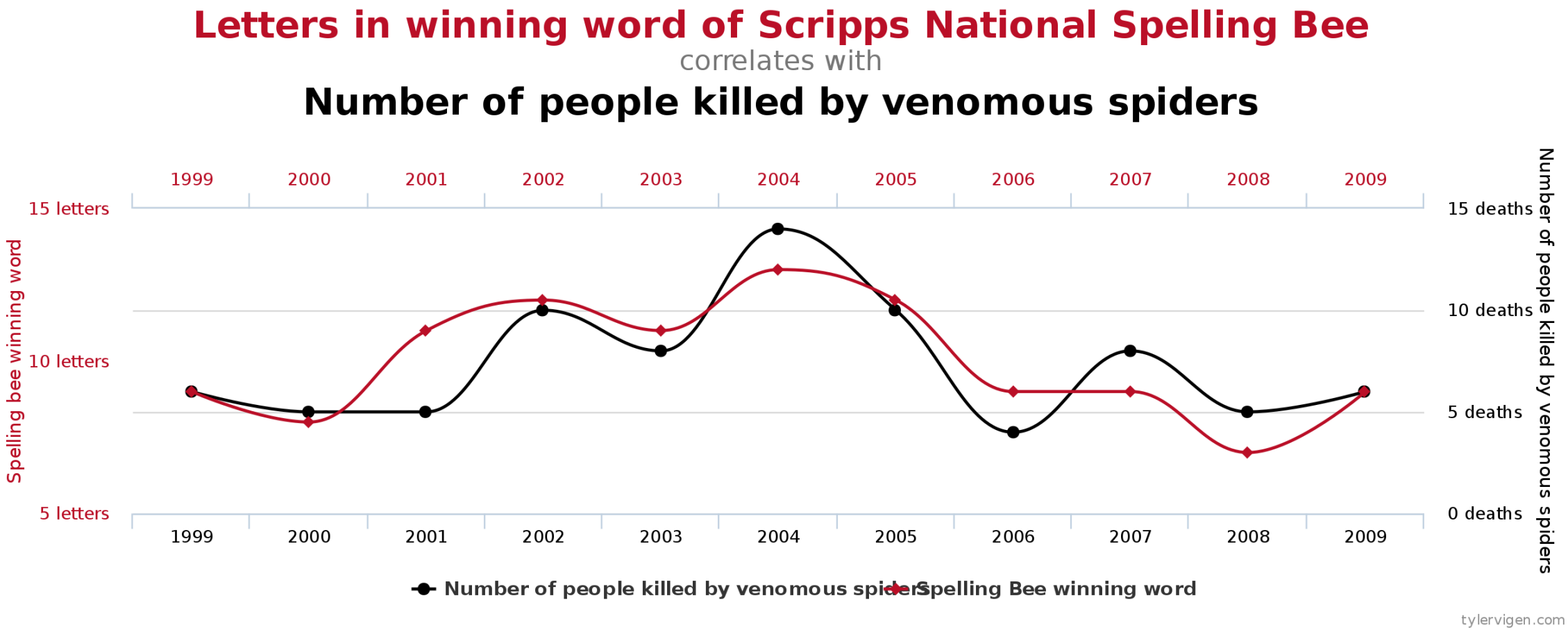
The Texas sharpshooter



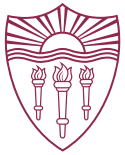
- A bad shooter who fires bullets at random at the side of a barn, then draws a “bull’s eye” around the biggest clump of bullet holes, and points proudly at his success.

- “p-hacking” - Exploiting researcher degrees of freedom until $p < 0.05$.
- misuse of data analysis to find patterns in data that can be presented as statistically significant, thus dramatically increasing and understating the risk of false positives.

Spurious correlations



Bonferroni correction



- A method to counteract the problem of multiple comparisons

- If you are testing m hypotheses for statistical significance at level α (e.g., α =0.05),
- The Bonferroni correction tests each hypothesis at α/m.

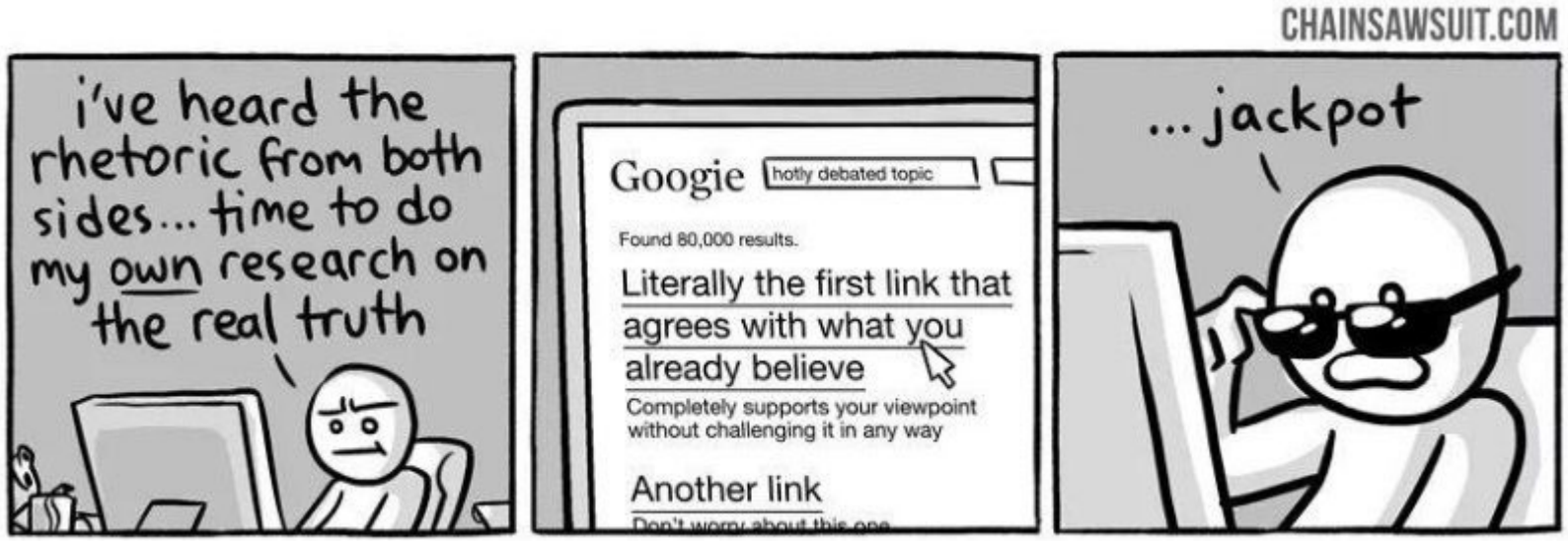
Asymmetric attention

- aka **disconfirmation bias** - when researchers accept the results they *expected*, but we rigorously check *unexpected* results
- A study of 165 laboratory experiments found that in 88% of cases in which results did not align with expectations, the scientists blamed the inconsistencies on how the experiments were conducted, rather than on their own theories. Consistent results, by contrast, were given little to no scrutiny.

- *cf* **confirmation bias**

Confirmation bias

Confirmation bias: tendency to search for, favor, and recall information that supports one's prior beliefs, and ignore contrary information. People also tend to interpret ambiguous evidence as supporting their pre-existing beliefs.

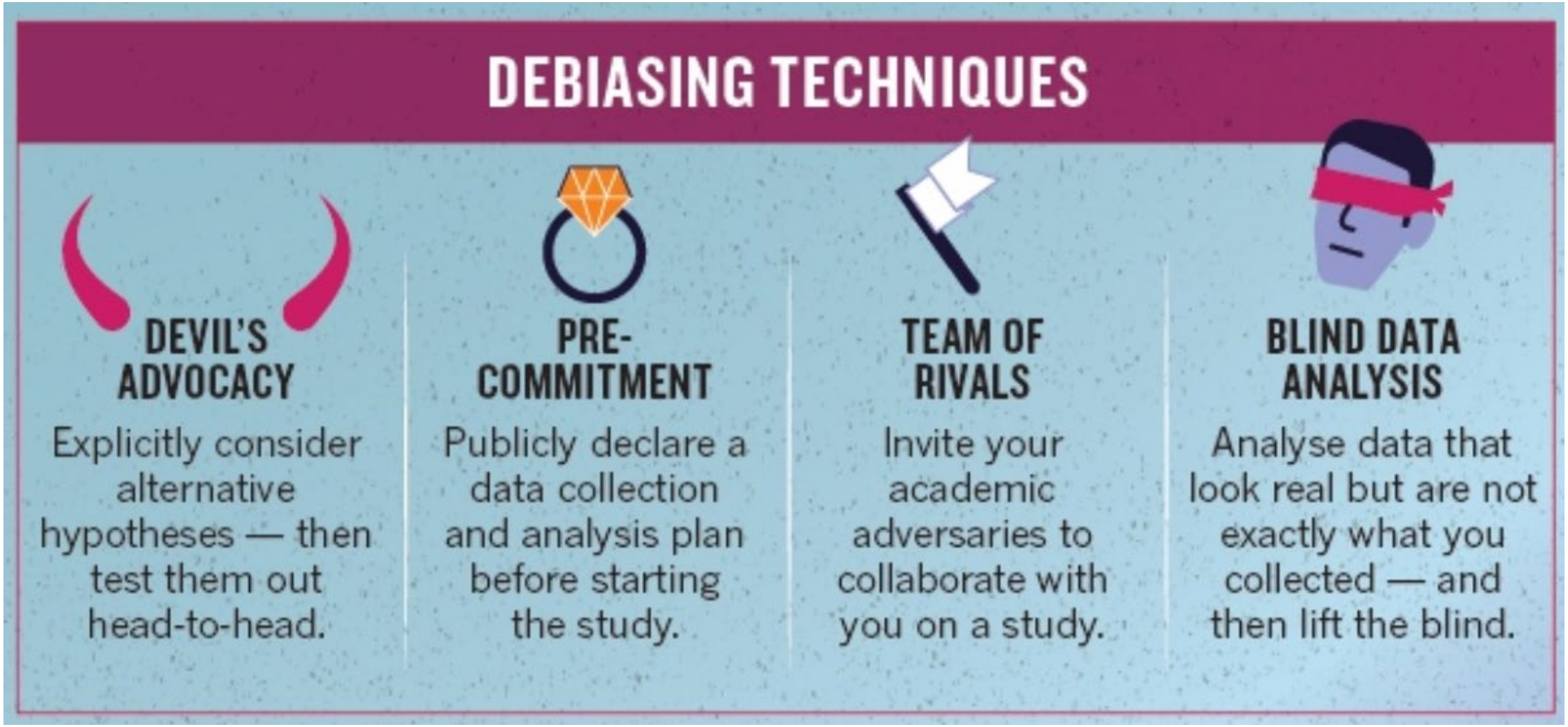


http://chainsawsuit.com/comic/2014/09/16/on-research/

Just-so storytelling

- post-hoc stories can be concocted to justify anything and everything
- ”JARK” - justifying after results are known

Resisting cognitive biases



https://www.nature.com/news/how-scientists-fool-themselves-and-how-they-can-stop-1.18517

Be your own devil's advocate

- explicitly consider competing hypotheses,
  - Develop and test alternate hypotheses
- develop experiments that can distinguish between them

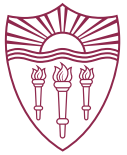
Transparency

- Open data science: always share your methods, (data,) computer code and results
- Researchers are migrating to the **Open Science Framework**
  - Hypothesis pre-registration: Precommit to analysis and report plan to mitigate the effect of cognitive biases
  - Registered report: publication in which scientists present their research plans for peer review before they even do the experiment.
  - Also avoids publication bias

Teams of rivals

- invite your rivals to work with you
- rivals will quickly spot flaws such as hypothesis myopia, asymmetric attention or just-so storytelling, and cancel them out with similar slants favouring the other side.

Blind data analysis



- Data scientists who do not know how close they are to desired results will be less likely to find what they are unconsciously looking for
- Data blinding approaches
  - Create alternative data sets, e.g., adding random noise or a hidden offset to data
  - Move participants to different experimental groups
  - Hiding demographic categories of data
- Data scientists analyze fake data set as usual — cleaning the data, handling outliers, running analyses
- In the background, a computer applies all of their actions to the real data.

An illustration “team of rivals”

[Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results, Silberzahn et al.]

- Are scientific results are highly contingent on subjective decisions at the analysis stage?
- 29 teams analyzed the same data to answer same research question.

Research question: Do soccer players with darker skin tones get more red cards?

Red carding

- Red card results in the player's ejection from the game and has severe consequences for his team → must play with one fewer player for the remainder of the match.
- Red cards are given for aggressive behavior, such as a tackling violently, fouling with the intent to deny an opponent a clear goal-scoring opportunity, hitting or spitting on an opposing player, or using threatening and abusive language.
- But, the evidence is often ambiguous, and referee decision may be biased

Data

- All teams were given the same large data set collected by a sports-statistics firm across four major football leagues, games, years, referees, players.
- Referee calls, counts of how often referees encountered each player, and player demographics including team position, height and weight.
- Includes a rating of players' skin color, coded manually:
  - two independent coders sorted photographs of players into five categories ranging from 'very light' to 'very dark' skin tone.

Some considerations of analysis

- Do you treat each red-card decision as an independent observation?
- How do you account for some referees giving more red cards than others?
- Would you take into account whether a referee's familiarity with a player affects the referee's likelihood of assigning a red card?
- Would you look at whether players in some leagues are more likely to receive red cards compared with players in other leagues, and whether the proportion of players with dark skin varies across leagues and player positions?

Project Stage	Work Package	Month																			
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	Building the Data Set																				
2	Recruitment and Initial Survey of Data Analysts																				
3	First Round of Data Analysis																				
4	Round-Robin Peer Evaluations																				
5	Second Round of Data Analysis																				
6a	Open Discussion and Debate, Further Analyses																				
6b	Write-Up of Manuscript																				
7	Internal Experts' Peer Review of Approaches																				
	Revision of Manuscript																				

Some data statistics

Height (cm)	$M = 181.74$ ($SD = 6.69$)
Weight (kg)	$M = 75.64$ ($SD = 7.10$)
Number of games	$M = 71.13$ ($SD = 36.17$)
Number of yellow cards	$M = 27.41$ ($SD = 24.08$)
Number of red cards	$M = 0.89$ ($SD = 1.26$)
League country	
England	$n = 564$ players
France	$n = 533$ players
Germany	$n = 489$ players
Spain	$n = 467$ players
Skin color	
0 (very light skin)	Rater 1: $n = 626$ players
	Rater 2: $n = 451$ players
.25	Rater 1: $n = 551$ players
	Rater 2: $n = 693$ players
.50	Rater 1: $n = 170$ players
	Rater 2: $n = 174$ players
.75	Rater 1: $n = 140$ players
	Rater 2: $n = 141$ players
1 (very dark skin)	Rater 1: $n = 98$ players
	Rater 2: $n = 126$ players

Some analytic approaches

Team	Distribution	Treatment of nonindependence	Number of covariates	Analytic approach
1	Linear	Clustered standard errors	7	Ordinary least squares regression with robust standard errors, logistic regression
6	Linear	Clustered standard errors	6	Linear probability model
14	Linear	Clustered standard errors	6	Weighted least squares regression with clustered standard errors
4	Linear	None	3	Spearman correlation
11	Linear	None	4	Multiple linear regression
10	Linear	Variance component	3	Multilevel regression and logistic regression
2	Logistic	Clustered standard errors	6	Linear probability model, logistic regression
30	Logistic	Clustered standard errors	3	Clustered robust binomial logistic regression
31	Logistic	Clustered standard errors	6	Logistic regression
32	Logistic	Clustered standard errors	1	Generalized linear models for binary data

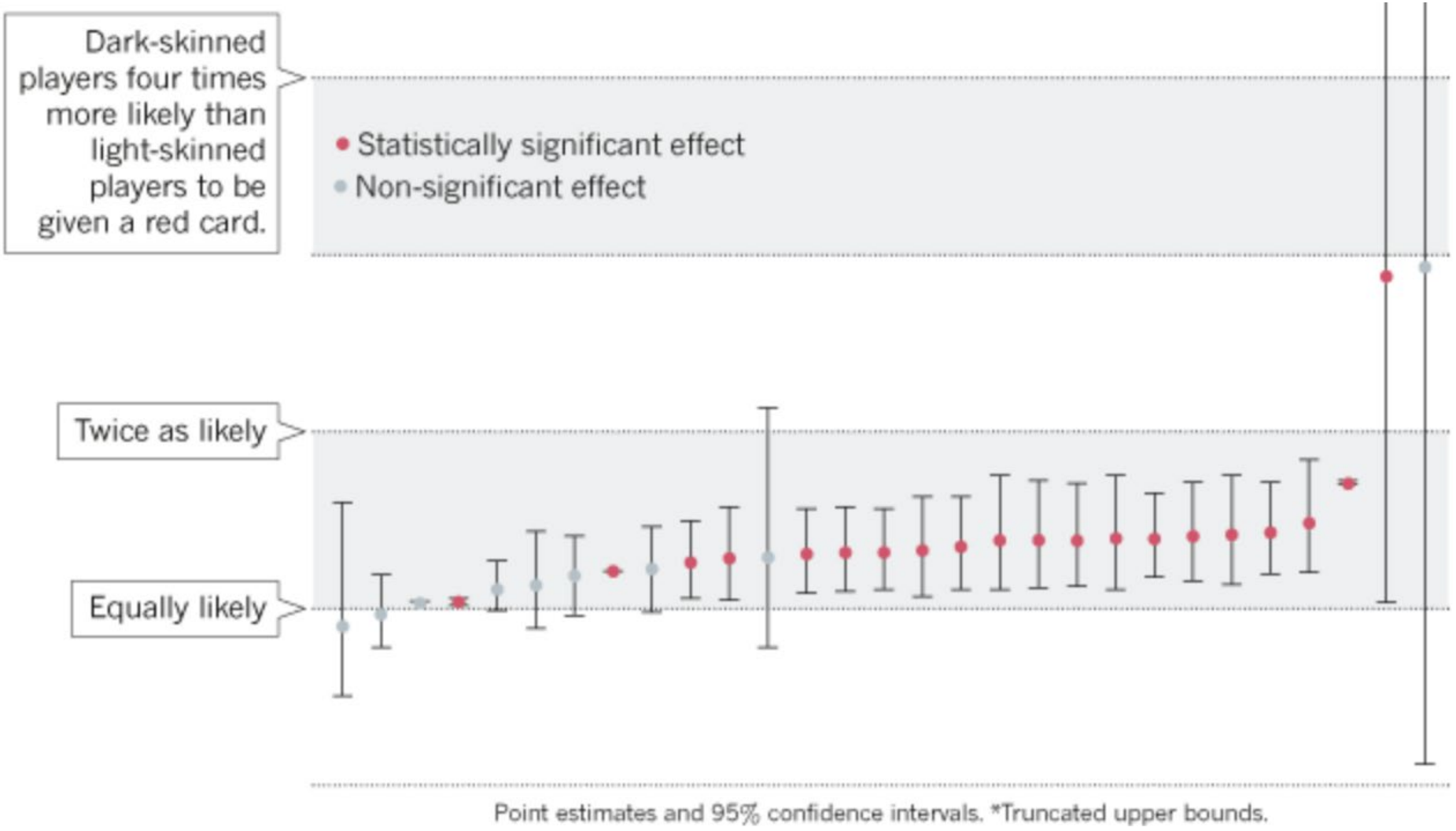
Covariates used by each team

Covariate	1	2	3	4	5	6	...
Player position	X	X	X			X	
Player's height	X	X		X		X	
Player's weight	X	X		X		X	
Player's league country[a]	X						
Player's age	X					X	
Goals scored by player		X					
Player's club	X					X	
Referee's country		X	X	X			
Referee	X					X	
Player's number of victories		X					
Number of cards received by player							
Player							
Number of cards awarded by referee							
Number of draws							
Number of covariates	7	6	2	3	0	6	

Effect size measured as odds ratio

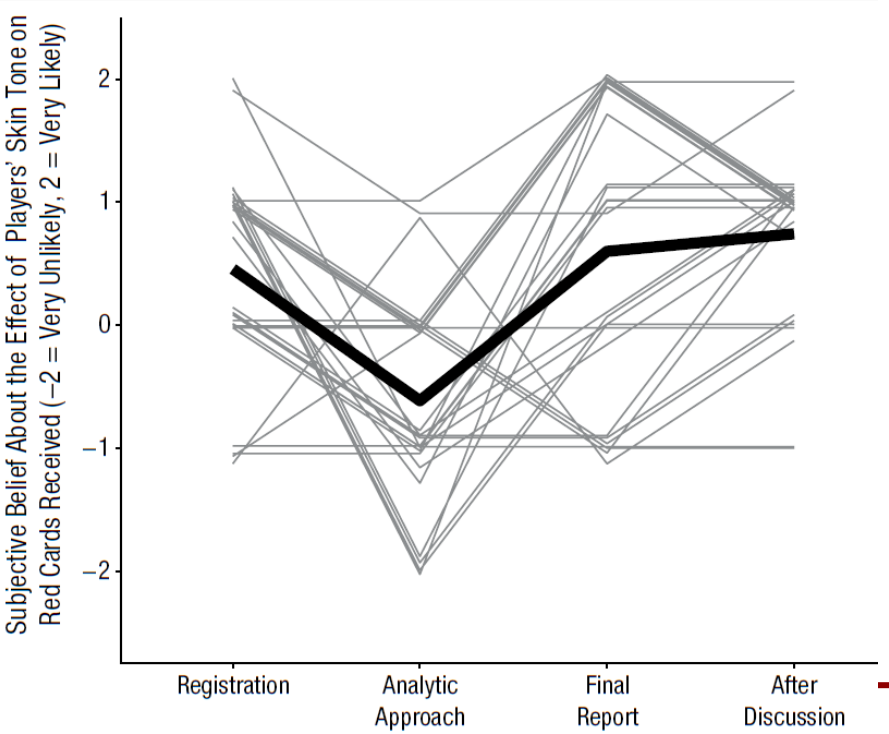
Analytic Approach	Odds Rat
Zero-Inflated Poisson Regression	0.89
Bayesian Logistic Regression	0.96
Hierarchical Log-Linear Modeling	1.02
Multilevel Regression and Logistic Regression	1.03
Hierarchical Bayes Model	1.10
Logistic Regression	1.12
OLS Regression With Robust Standard Errors, Logistic Regression	1.18
Spearman Correlation	1.21
WLS Regression With Clustered Standard Errors	1.21
Multiple Linear Regression	1.25
Clustered Robust Binomial Logistic Regression	1.28
Linear Probability Model	1.28
Hierarchical Generalized Linear Modeling With Poisson Sampling	1.30
Multilevel Logistic Regression Using Bayesian Inference	1.31
Mixed-Model Logistic Regression	1.31
Hierarchical Poisson Regression	1.32
Linear Probability Model, Logistic Regression	1.34
Generalized Linear Mixed Models	1.38
Multilevel Logistic Regression	1.38
Mixed-Effects Logistic Regression	1.38
Generalized Linear Models for Binary Data	1.39
Negative Binomial Regression With a Log Link	1.39
Cross-Classified Multilevel Negative Binomial Model	1.40
Poisson Multilevel Modeling	1.41
Multilevel Logistic Binomial Regression	1.42
Generalized Linear Mixed-Effects Models With a Logit Link	1.48
Dirichlet-Process Bayesian Clustering	1.71
Tobit Regression	2.88
Poisson Regression	2.93

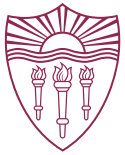
Odds ratio of red-carding a darker player



Summary

- Twenty teams (69%) found a significant positive relationship ($p < .05$), and nine teams (31%) found a nonsignificant relationship.
  - No team reported a significant negative relationship.
- *Did the analysts' beliefs regarding the hypothesis change over time?*
  - Analysts' beliefs at registration regarding whether players with darker skin tone were more likely to receive red cards were not significantly related to the reported effect sizes
  - Beliefs changed considerably throughout the project and were significantly related to the effect-size estimates
  - Not confirmation bias





SOME FINAL THOUGHTS

Do not oversell AI/DS capabilities

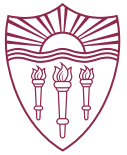
Slides courtesy of Arvind Narayanan [@random_walker]

- Some AI and Data Science technologies have made genuine, remarkable, widely-publicized progress

- Companies will exploit public confusion to slap the “AI” label on whatever they’re selling

- But, some problems are inherently hard, even for AI

Genuine and rapid progress



- Content identification (reverse music, image search)
- Face recognition*
- Medical diagnosis from scans
- Speech to text
- Deepfakes*

Automated perception

* Ethical concerns because of high accuracy

Far from perfect, but improving

- Spam detection
- Detection of copyrighted material
- Automated essay grading
- Hate speech detection
- Content recommendation

Automated judgment

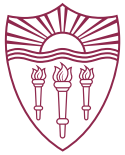
Ethical concerns in part because some error is inevitable

Fundamentally dubious

- Predicting criminal recidivism
- Predicting job performance
- Predictive policing
- Predicting terrorist risk
- Predicting at-risk kids

Predicting social outcomes

Ethical concerns amplified by inaccuracy



Genuine, rapid progress

- Content identification (reverse music, image search)
- Face recognition*
- Speech to text
- Deepfakes*

Imperfect, but improving

- Spam detection
- Copyright violation
- Automated grading
- Hate speech detect.
- Content recommend.

Fundamentally dubious

- Predicting recidivism
- Predicting job success
- Predictive policing
- Predicting terrorism
- Predicting at-risk kids

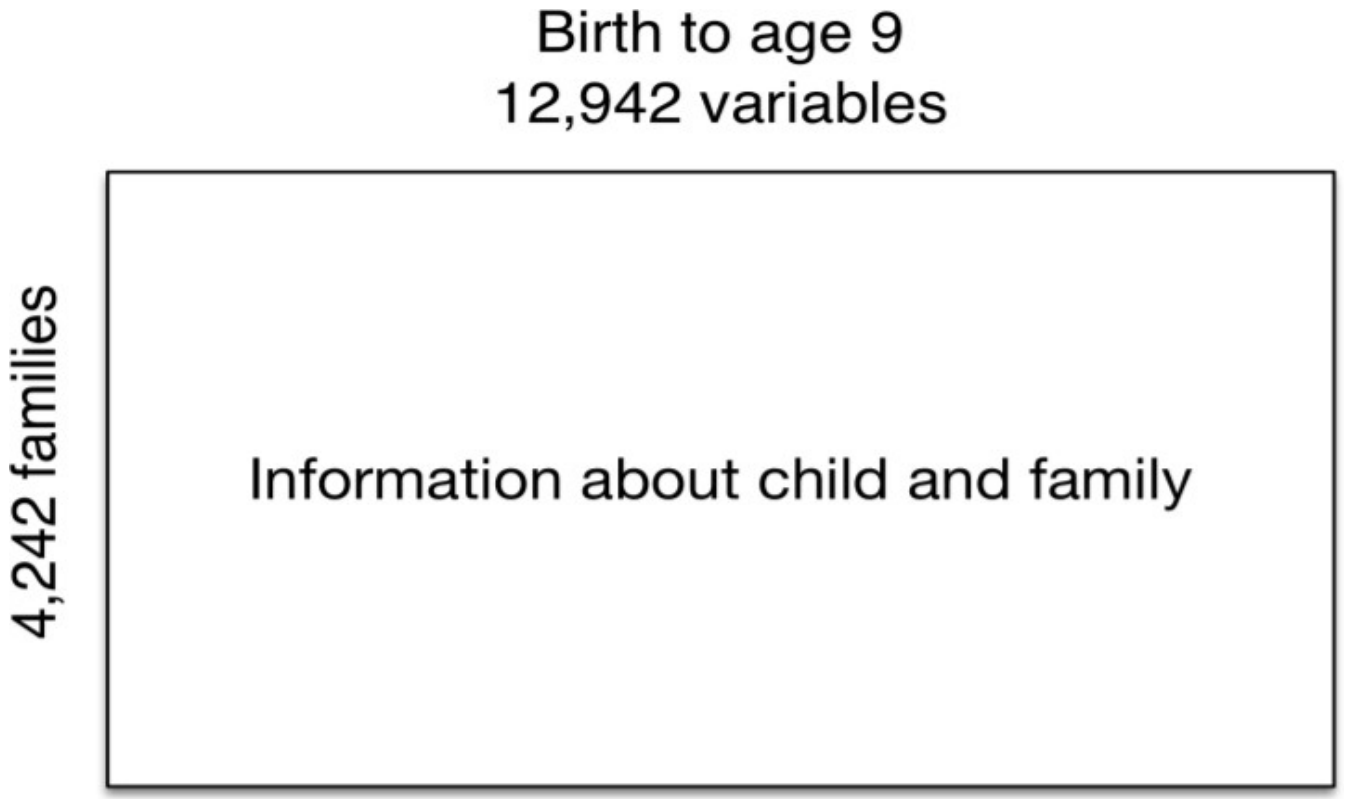
Automated perception

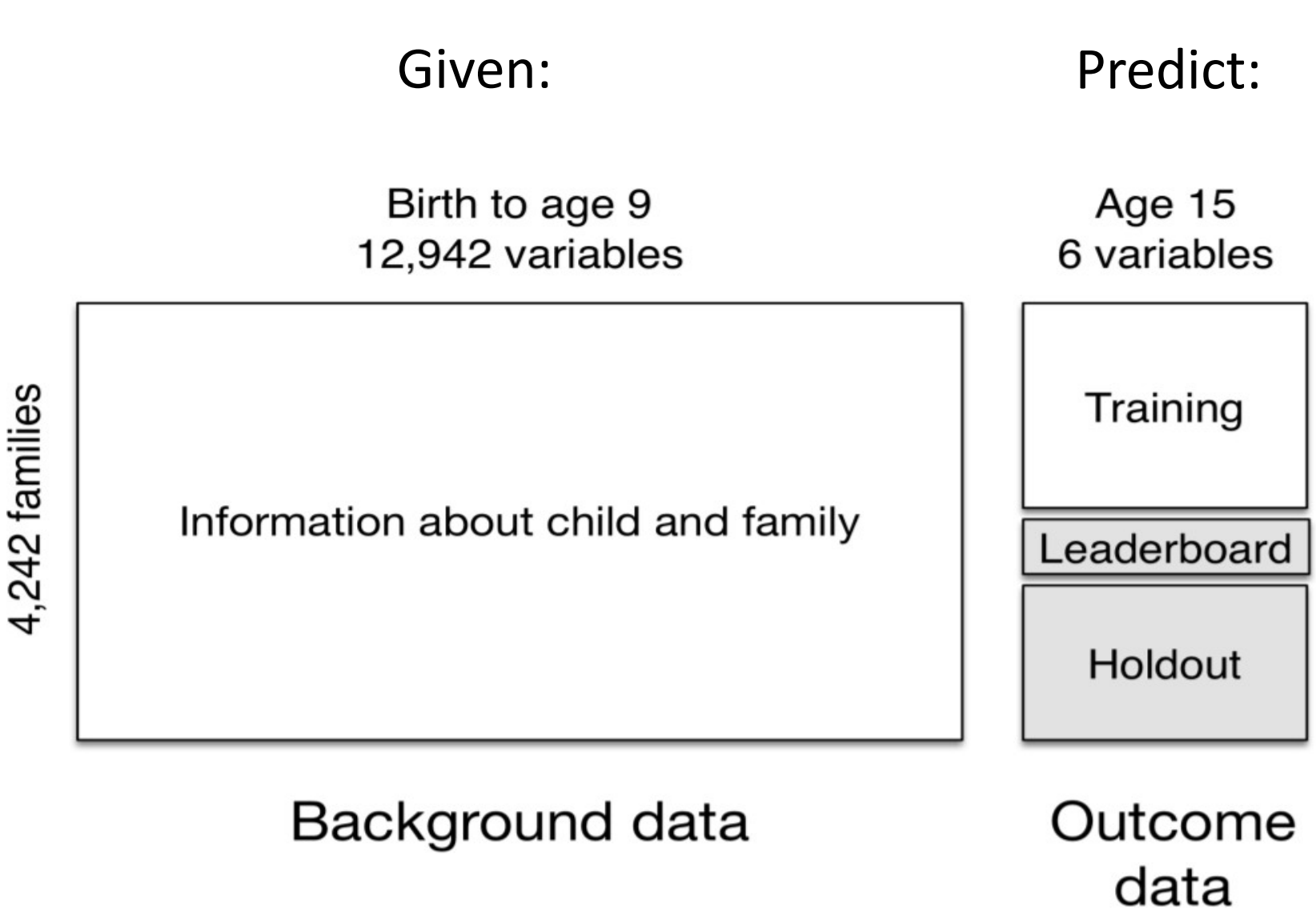
Automated judgment

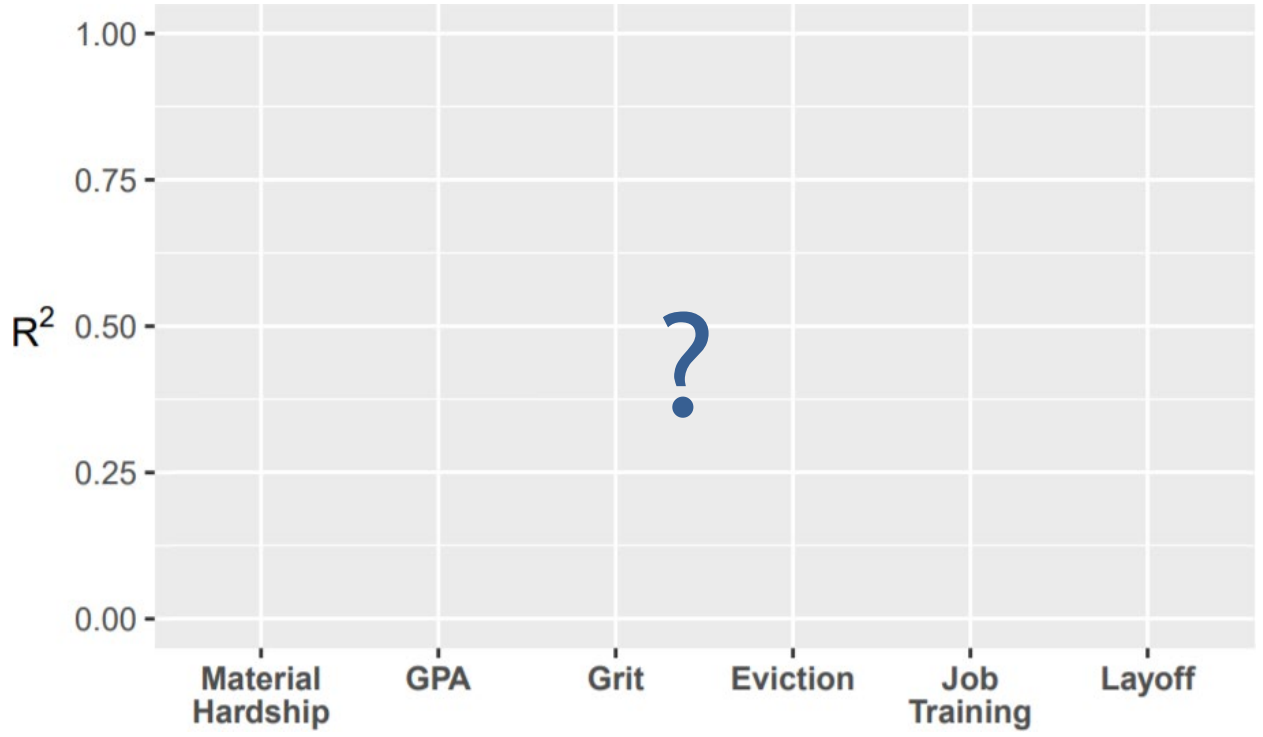
Predicting social outcomes

Can social outcomes be predicted?

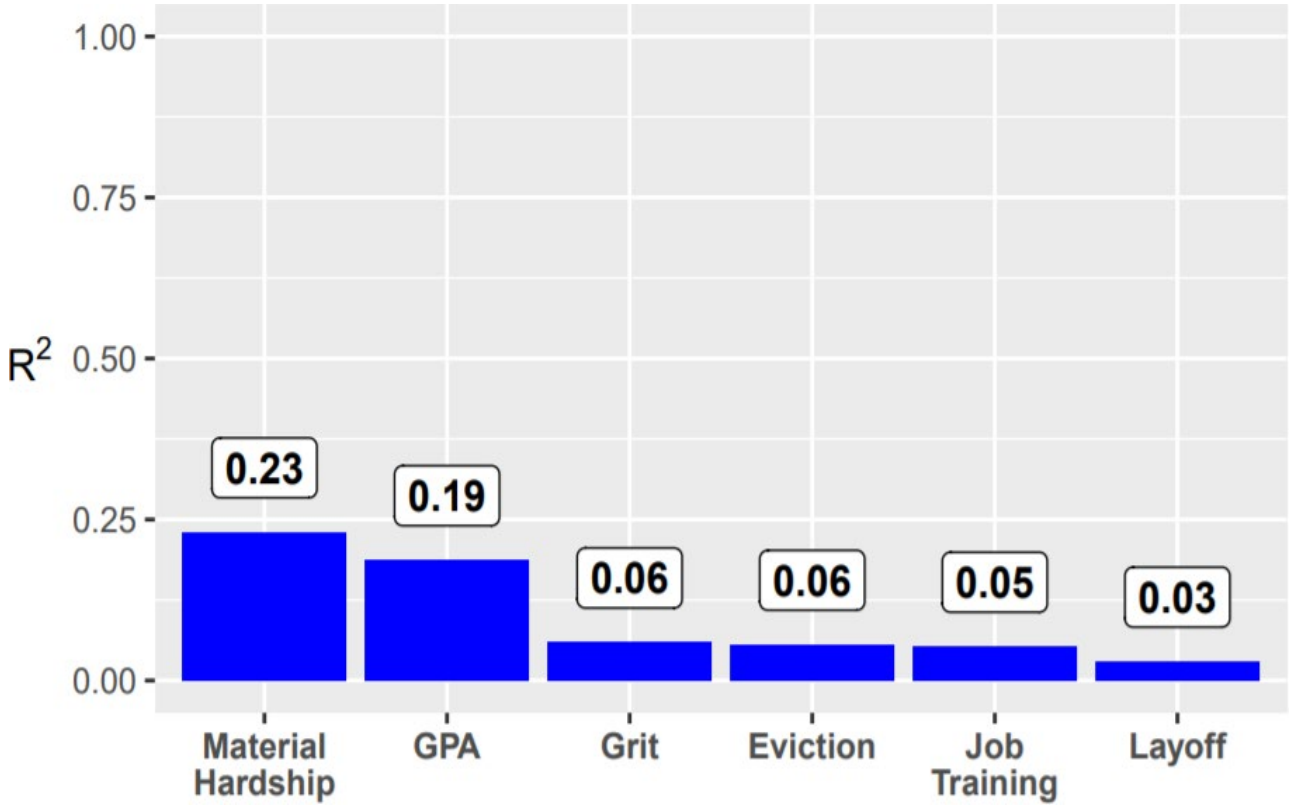
- Matthew Salganik, Ian Lundberg, Alex Kindel, Sara McLanahan, et al
- Mass collaboration involving 457 researchers.


Birth to age 9
12,942 variables
4,242 families
Information about child and family

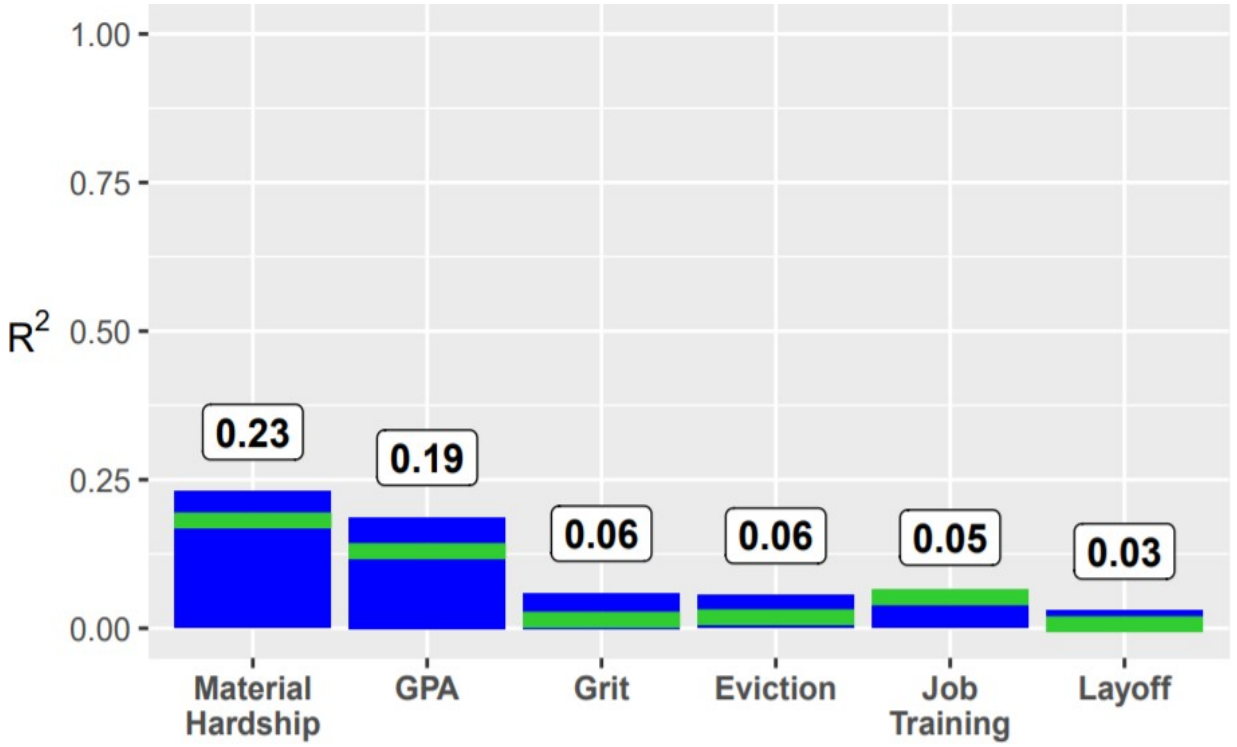

Given:
Predict:
Birth to age 9
12,942 variables
Age 15
6 variables
4,242 families
Information about child and family
Training
Leaderboard
Holdout
Background data
Outcome data


1.00
0.75
R^2
0.50
0.25
0.00
?
Material Hardship
GPA
Grit
Eviction
Job Training
Layoff

Best performing models



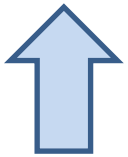
Green line: 4-variable linear regression



- 13,000 features hardly better than 4 features!
- "AI" hardly better than simple linear formula

Accuracy of recidivism prediction

COMPAS tool (137 features):	65% ± 1%	(slightly better than random)
Logistic regression (2 features):	67% ± 2%	



Age and number of priors

[Dressel & Farid. *The accuracy, fairness, and limits of predicting recidivism*. Science Advances 2018]

Potential harms

- Hunger for personal data
- Massive transfer of power from domain experts & workers to unaccountable tech companies
- Lack of explainability
- Distracts from interventions
- Veneer of accuracy
- …