Statistics and Machine Learning: Foundations, Limitations, and Ethics

Momin M. Malik, PhD <momin_malik@cyber.harvard.edu>
Data Science Postdoctoral Fellow
Berkman Klein Center for Internet & Society at Harvard University

Colby College, Mathematics and Statistics Colloquium 2019, 07 October 2019

Slides: https://mominmalik.com/colby2019.pdf

5-point summary

- Stats/ML divides the world into fixed entities with fixed properties; this is not natural
- The *central tendency* is the fundamental tool of stats/ML, and cannot consider individuality
- “Predictions” are correlations, and sometimes spurious correlations fit better than non-spurious ones
- Quantifying uncertainty (stats) has the blindspot of uncertainty in data choice, and cross-validation (ML) has the blindspots of dependencies and uncertainty
- Is it right to treat people as interchangeable? What about punishing/rewarding people based on correlations?

Goals

- For students, to clarify some things that confused me when I first started and didn’t find explained anywhere
- For future practitioners, to understand the nature of statements in stats/ML, and to understand resulting limitations
- Reflect on what statistics and machine learning *do* in the world, and if this is what we want in a given case

Outline

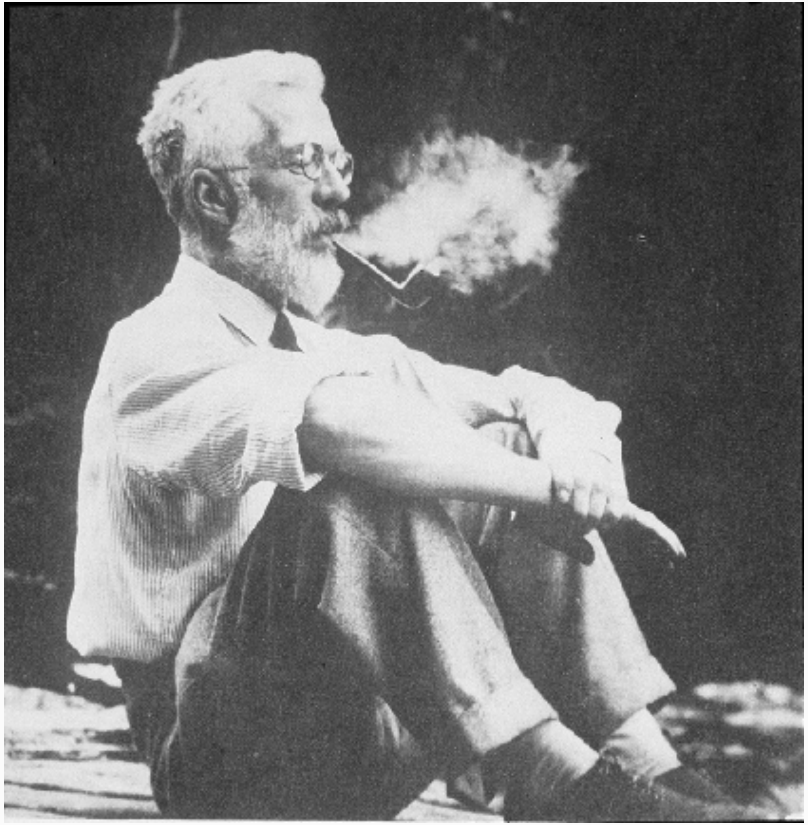
- Statistics
  - Reduction of data to central tendency
    - Limitations: Individuality, meaning, and experience
  - Quantifying uncertainty
- Machine learning
  - “Prediction”
    - Limitations: Causality, bias-variance
  - Cross-validation
- Ethics

What is statistics?

Statistics = “The reduction of data”

“briefly, and in its most concrete form, the object of statistical methods is the ***reduction of data****.”*

– R. A. Fisher, “On the mathematical foundations of theoretical statistics” (1922)



- Introduction
- What is statistics?
- Reduce data to “relevant information”
- Individuality, meaning, and experience
- Quantifying uncertainty
- What is machine learning?
- “Prediction”
- Causality
- Cross-validation
- Ethics
- References

Reduction to “relevant information”

“A quantity of data, which usually by its ***mere bulk*** *is* ***incapable of entering the mind****, is to be replaced by* ***relatively few quantities*** *which shall* ***adequately represent the whole****, or… as much as possible… of the* ***relevant information*** *contained in the original data.”*

– R. A. Fisher, “On the mathematical foundations of theoretical statistics” (1922)

Reduction to “relevant information”

A “statistic” (singular) is defined as *a function of the data*.

The discipline of Statistics is about *defining* “relevant information,” and finding functions to capture it.

How does it do so?

Relevant info defined via probability

I understand statistics as:

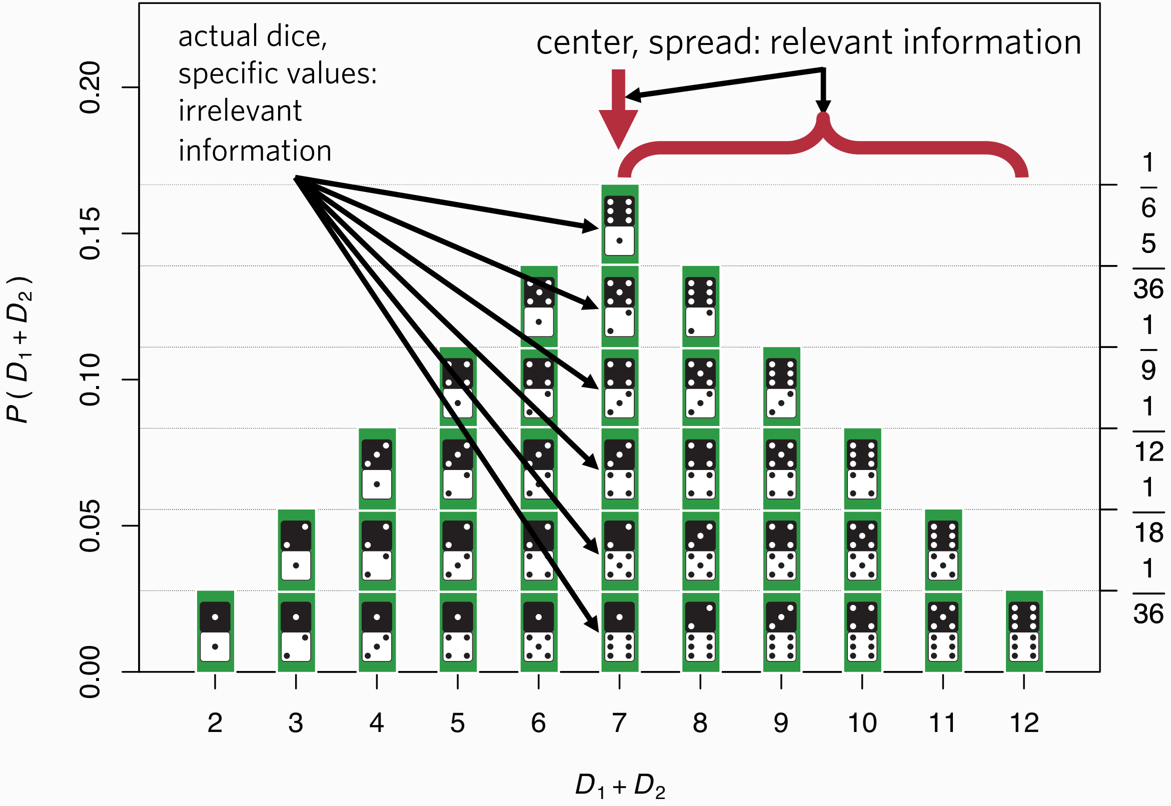
*The use of probability as a model for variability in the world.**

* Technically, *“Probability is used in two distinct, although interrelated, ways in statistics, phenomenologically to describe haphazard variability arising in the real world and epistemologically to represent uncertainty of knowledge.”* David R. Cox, “Role of models in statistical analysis” (1990). For now I focus only on the former.

BERKMAN KLEIN CENTER
FOR INTERNET & SOCIETY
AT HARVARD UNIVERSITY

- Introduction
- What is statistics?
- Reduce data to “relevant information”
- Individuality, meaning, and experience
- Quantifying uncertainty
- What is machine learning?
- “Prediction”
- Causality
- Cross-validation
- Ethics
- References

‘Relevant’ and ‘irrelevant’ information



Note: *the connection is not at all natural!*

*"It is remarkable that a science which began with the consideration of **games of chance** should have become the most important object of human knowledge."*

– Pierre-Simon Laplace, *Théorie Analytique des Probabilitiés* (1812)

Probability as a model for variability

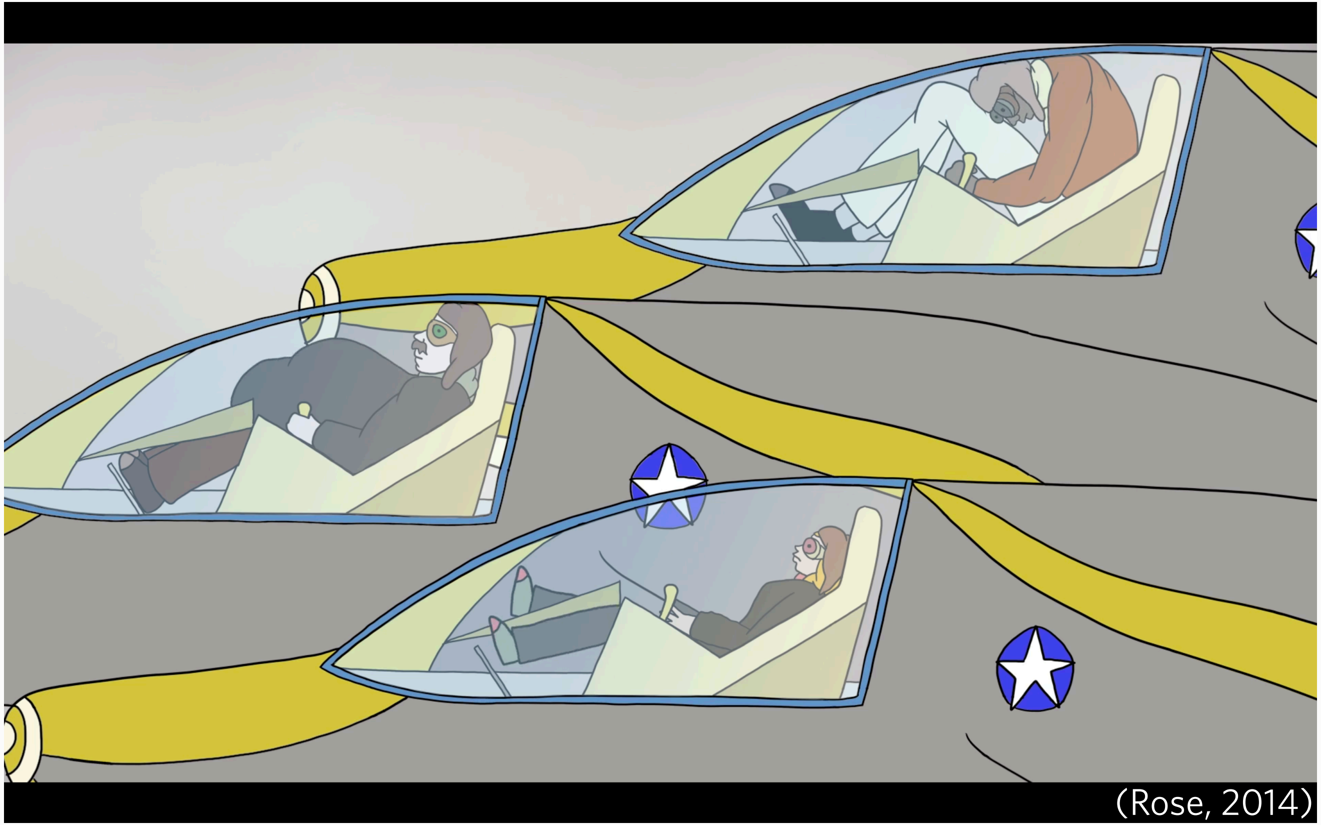
Makes a philosophical commitment:

- *There are distinct entities in the world that, despite being different, are similar in some way.*

Corollary: we can learn about one thing by studying other things (and eventually, make statements about not-yet-seen entities based on the study of seen entities).

Limitations

The "flaw of averages"



(Rose, 2014)

Must deal in aggregates

- Similarity/"relevant information" is in terms of some form of *central tendency*.
- Necessarily ignores individuality; can only say something if $n > 1$ (and, rule of thumb, at least $n > 30$)
- Ignoring individuality is a choice, not intrinsic.
- A true "average man" (Adolphe Quetelet's *l'homme moyen*, 1835), who is average in all aspects, would be quite peculiar!
- By choosing averages, we may actually end up imposing it on the world! (Treating people as interchangeable)

Meaning-making

"During the writing of this book, my first grandchild was born. The hospital records document her weight, height, health[;] the mother's condition, length of labor, time of birth, and hospital stay... These are physiological and institutional metrics. When aggregated across many babies and mothers, they provide trend data about the beginning of life—birthing." (Patton, 2015)

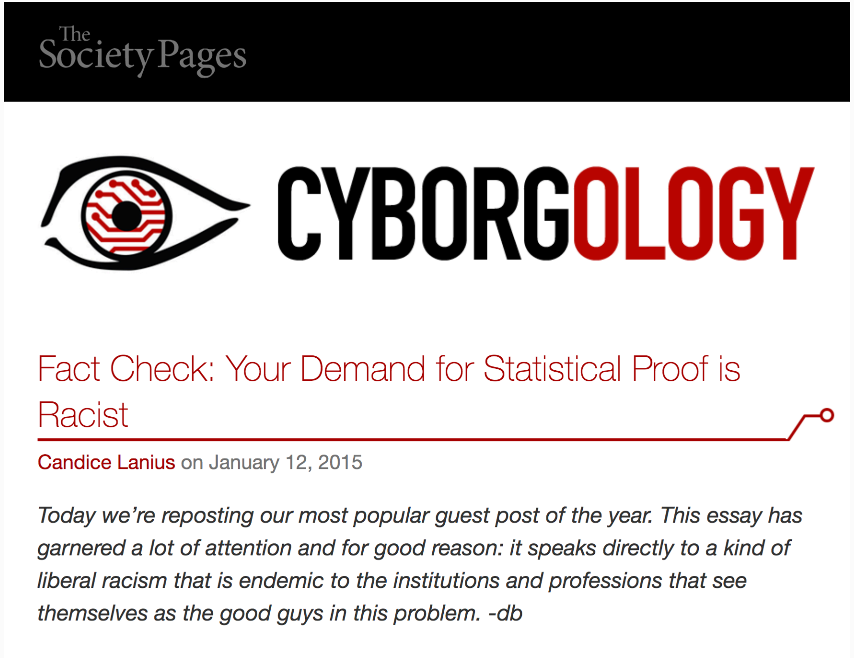
Meaning-making

“But nowhere in the hospital records will you find anything about what the birth of Calla Quinn *means.* Her existence is documented but not what she means to our family, what decision-making process led up to her birth, the experience and meaning of the pregnancy, the family experience of the birth process, and the familial, social, cultural, political, and economic context…” (Patton, 2015)

BERKMAN KLEIN CENTER
FOR INTERNET & SOCIETY
AT HARVARD UNIVERSITY

- Introduction
- What is statistics?
- Reduce data to “relevant information”
- Individuality, meaning, and experience
- Quantifying uncertainty
- What is machine learning?
- “Prediction”
- Causality
- Cross-validation
- Ethics
- References

Modeling vs. experience



The Society Pages

CYBORGOLOGY

Fact Check: Your Demand for Statistical Proof is Racist

Candice Lanius on January 12, 2015

Today we’re reposting our most popular guest post of the year. This essay has garnered a lot of attention and for good reason: it speaks directly to a kind of liberal racism that is endemic to the institutions and professions that see themselves as the good guys in this problem. -db

- “A white woman can say that a neighborhood is ‘sketchy’ and most people will smile and nod. She felt unsafe, and we automatically trust her opinion. A black man can tell the world that every day he lives in fear of the police, and suddenly everyone demands statistical evidence to prove that his life experience is real.”

Quantifying uncertainty

The other part of statistics: Uncertainty

“Probability is used in two distinct, although interrelated, ways in statistics, phenomenologically to describe haphazard variability arising in the real world and epistemologically to represent uncertainty of knowledge.”

– David R. Cox, “Role of models in statistical analysis” (1990)

Likelihood principle: Data to probability

- A probability distribution is a function,

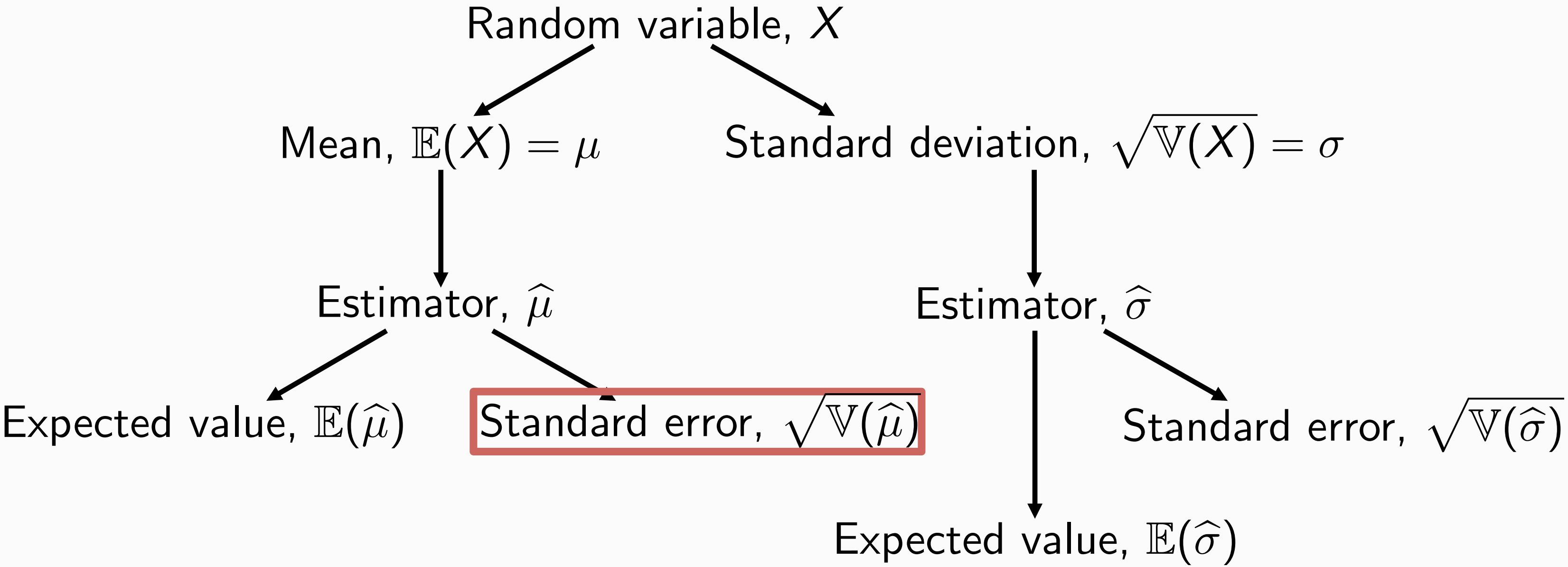
$$p(\mathbf{x}) \propto \exp\left\{\boldsymbol{\eta}(\boldsymbol{\theta})^T \mathbf{T}(\mathbf{x})\right\}$$

- Input possible value of data *x*, get back the probability
- If you instead have an observed values, take the same equation, but treat it as a function of the parameter(s)

$$\mathcal{L}(\boldsymbol{\theta}) \propto \exp\left\{\boldsymbol{\eta}(\boldsymbol{\theta})^T \mathbf{T}(\mathbf{x})\right\}$$

- Interpret as: what values of the parameters make the observed values *most likely*? Solve for parameters

Inference: Uncertainty of an estimator



Random variable, X

Mean, $\mathbb{E}(X) = \mu$

Standard deviation, $\sqrt{\mathbb{V}(X)} = \sigma$

Estimator, $\widehat{\mu}$

Estimator, $\widehat{\sigma}$

Expected value, $\mathbb{E}(\widehat{\mu})$

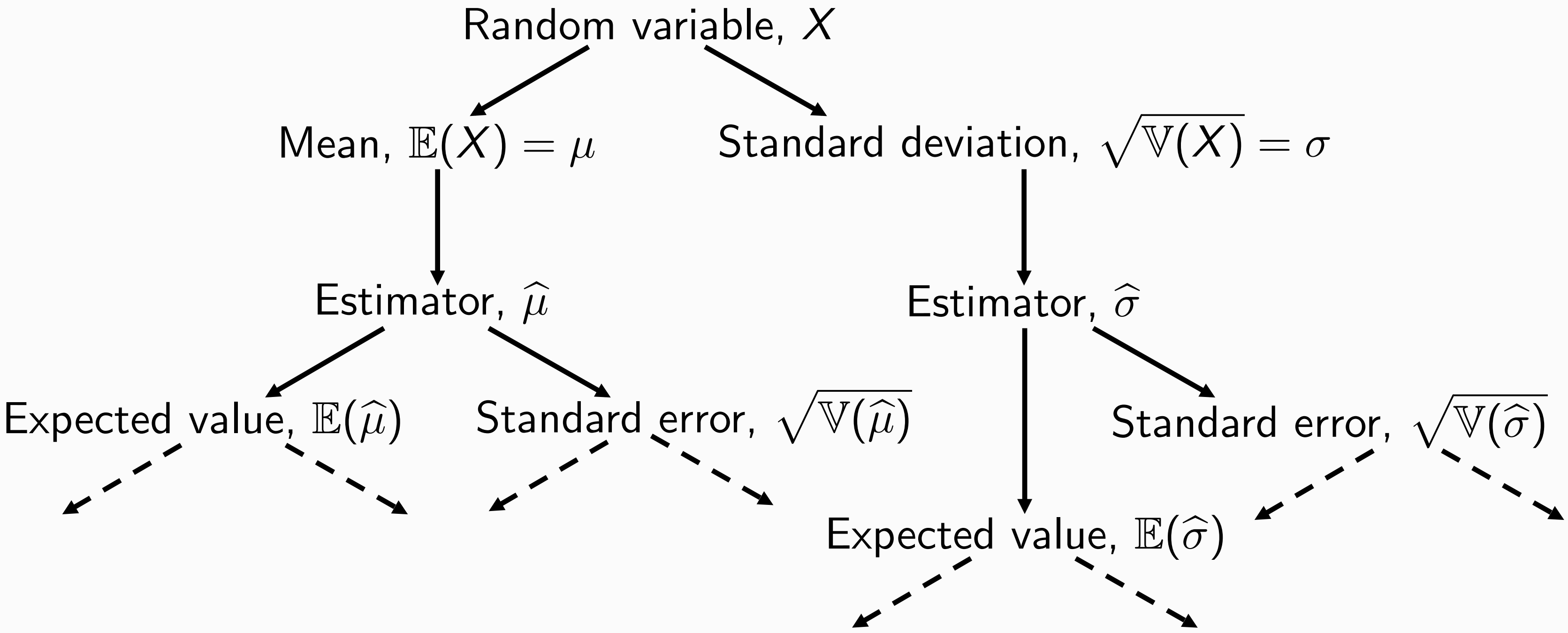
Standard error, $\sqrt{\mathbb{V}(\widehat{\mu})}$

Standard error, $\sqrt{\mathbb{V}(\widehat{\sigma})}$

Expected value, $\mathbb{E}(\widehat{\sigma})$

The *variance* of the *estimator* of the *mean* gives us the uncertainty of the estimate, and is given the special name of the *standard error*. If the uncertainty is small enough, we say we have made an *inference* to the underlying data-generating process.

Going on *ad infinitum*...



Random variable, X

Mean, $\mathbb{E}(X) = \mu$

Standard deviation, $\sqrt{\mathbb{V}(X)} = \sigma$

Estimator, $\widehat{\mu}$

Estimator, $\widehat{\sigma}$

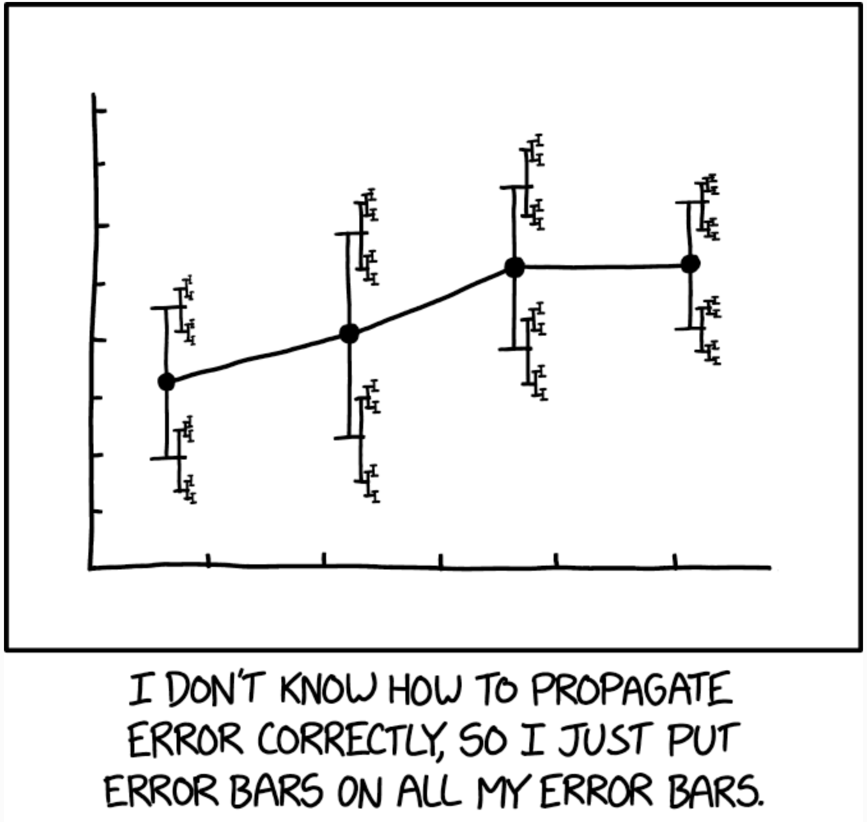
Expected value, $\mathbb{E}(\widehat{\mu})$

Standard error, $\sqrt{\mathbb{V}(\widehat{\mu})}$

Standard error, $\sqrt{\mathbb{V}(\widehat{\sigma})}$

Expected value, $\mathbb{E}(\widehat{\sigma})$

Going on *ad infinitum*...


I DON'T KNOW HOW TO PROPAGATE ERROR CORRECTLY, SO I JUST PUT ERROR BARS ON ALL MY ERROR BARS.

https://xkcd.com/2110/

Problems with quantifying uncertainty

- "*model uncertainty* is a fact of life and likely to be more serious than other sources of uncertainty which have received far more attention from statistician." (Chatfield, 1995)
- "the analyst will never know whether the inferences are good since the estimates cannot be compared directly with the truth."

What is machine learning?

- Introduction
- What is statistics?
- Reduce data to "relevant information"
- Individuality, meaning, and experience
- Quantifying uncertainty
- What is machine learning?
- "Prediction"
- Causality
- Cross-validation
- Ethics
- References

What is machine learning?

- *"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P* ***if its performance*** *at tasks in T, as measured by P,* ***improves with experience*** *E."*

  – Tom M. Mitchell (1997)

- This definition has nothing to do with statistics or probability!

“Learning” from data

“Statistics is the science of learning from data. Machine learning (ML) is the science of learning from data. These fields are identical in intent although they differ in their history, conventions, emphasis and culture.”

“At first, ML researchers developed expert systems that eschewed probability. But very quickly they adopted advanced statistical concepts like empirical process theory and concentration of measure. This transition happened in a matter of a few years.”

– Larry Wasserman, “Rise of the Machines” (2014)

"Learning" is a metaphor

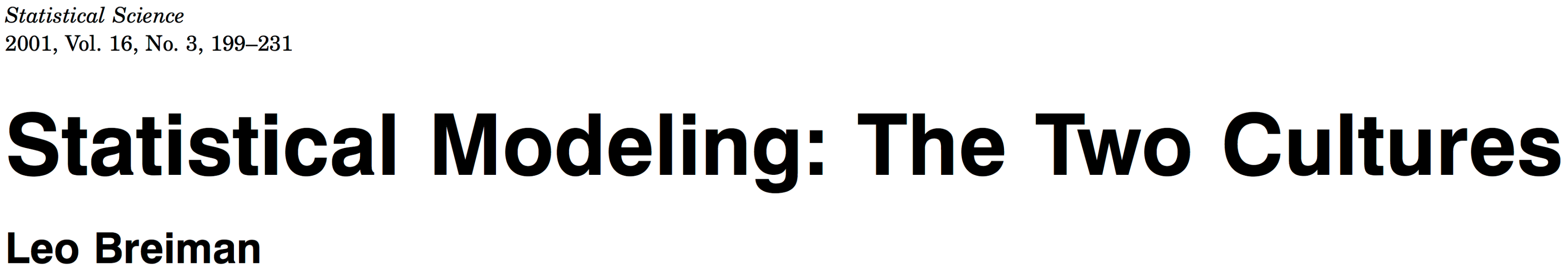
- The "learning" is a *metaphor*. The way in which machines "improve with data" has only a fleeting resemblance to human learning.
- "A.I. systems tend to be passive vessels, dredging through data in search of statistical correlations; humans are active engines for discovering how things work."
  – Gary Marcus, 2017, "Artificial Intelligence Is Stuck. Here's How to Move It Forward"
- (This perspective is not universal, it gets into heated philosophical debates and "hard" vs. "soft" artificial intelligence, Turing Test vs. the "Chinese Room," etc...)

Now, all statistical

- Surprising that statistical approaches, designed to uncover data-generating mechanisms with variation (and under uncertainty), could be applied to carry out operations resembling “intelligence”
- The original vision of AI, and ML, had to do with modeling (and thereby reproducing) *rules and reasoning*, but that failed; what worked was using statistics

- Introduction
- What is statistics?
- Reduce data to “relevant information”
- Individuality, meaning, and experience
- Quantifying uncertainty
- What is machine learning?
- “Prediction”
- Causality
- Cross-validation
- Ethics
- References

“Two cultures”



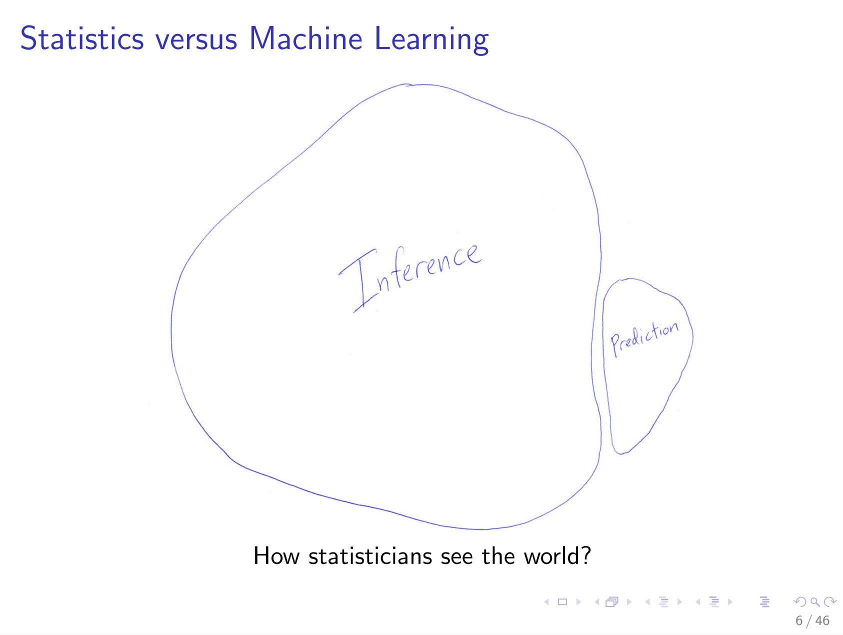
Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

“There are ***two cultures*** *in the use of statistical modeling to reach conclusions from data. One assumes that the* ***data are generated by a given stochastic data model****. The other uses* ***algorithmic models*** *and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to* ***irrelevant theory, questionable conclusions****, and has kept statisticians from working on a large range of interesting current problems.”*

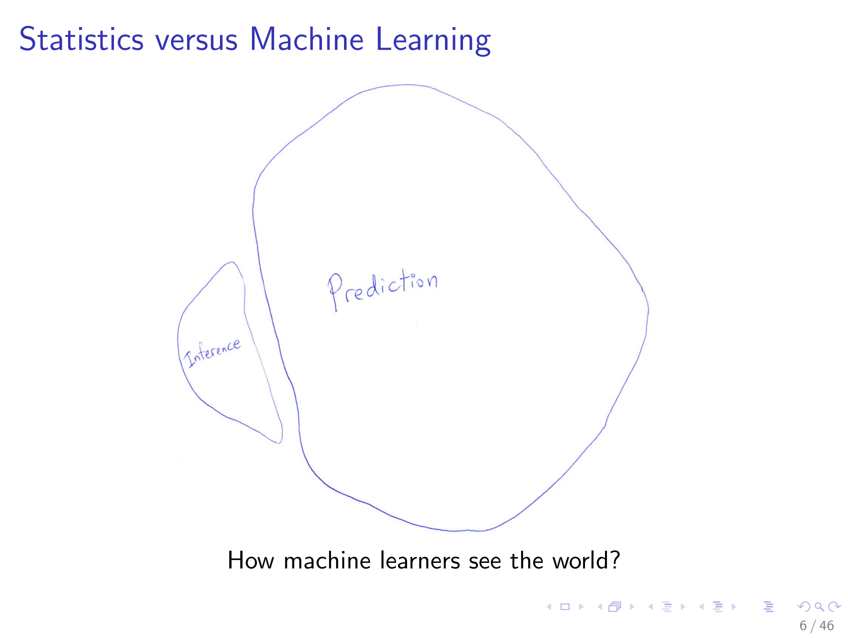
Statistics vs. machine learning



Statistics versus Machine Learning

How statisticians see the world?

6 / 46



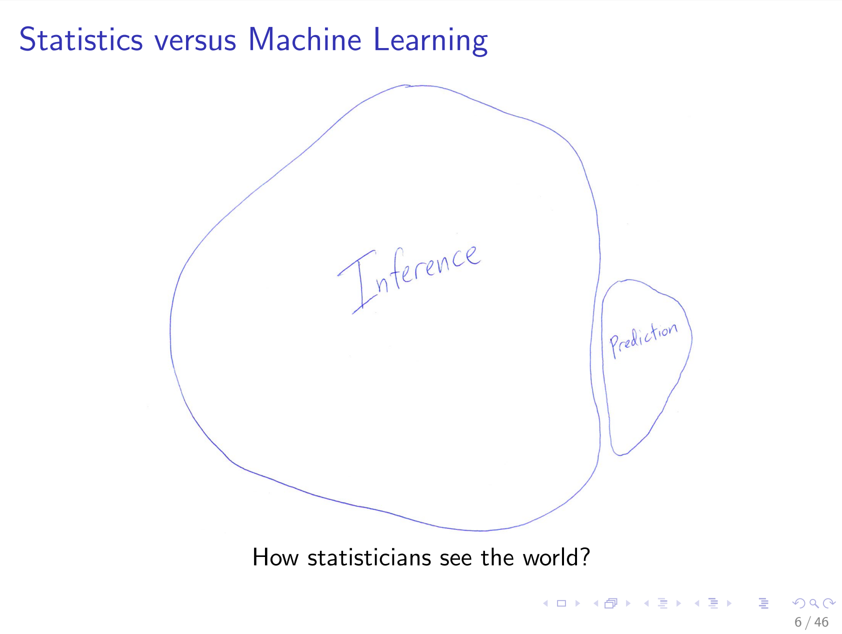
Statistics versus Machine Learning

How machine learners see the world?

6 / 46

Diagrams: Robert Tibshirani, “Recent Advances in Post-Selection Inference” (2015)

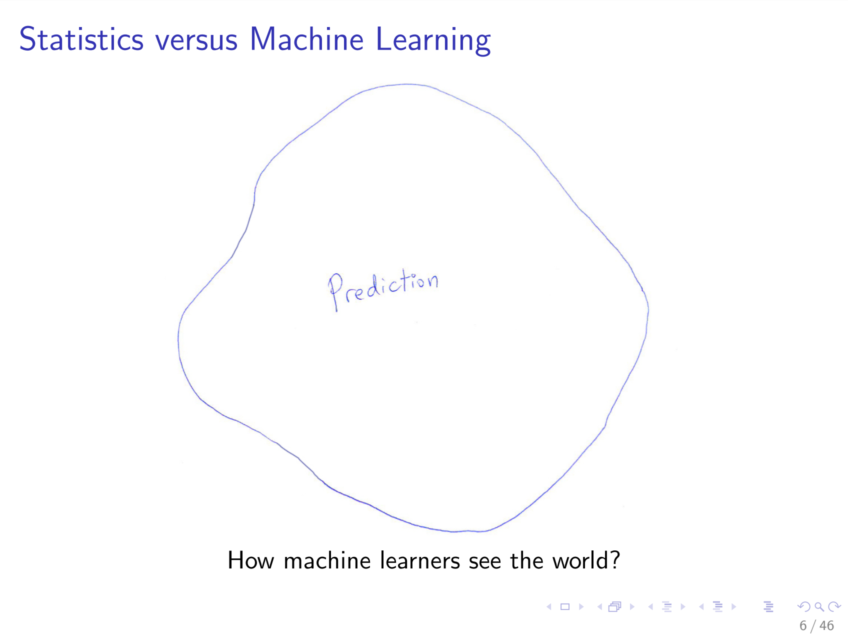
Statistics vs. machine learning



Statistics versus Machine Learning

How statisticians see the world?

6 / 46



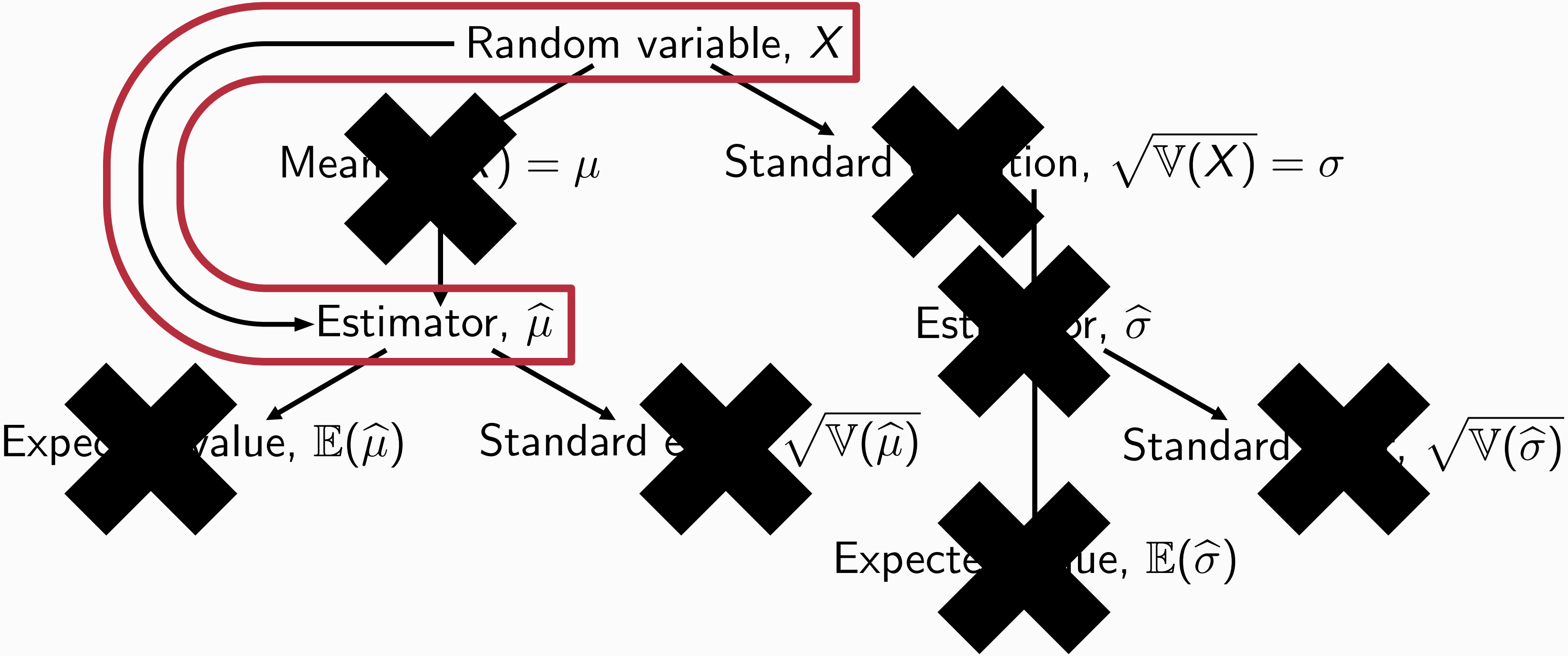
Statistics versus Machine Learning

How machine learners see the world?

6 / 46

Diagrams: Robert Tibshirani, “Recent Advances in Post-Selection Inference” (2015)

Machine learning: Instrumentalist



Random variable, X

Mea $) = \mu$

Standard tion, $\sqrt{\mathbb{V}(X)} = \sigma$

Estimator, $\widehat{\mu}$

Est r, $\widehat{\sigma}$

Expe value, $\mathbb{E}(\widehat{\mu})$

Standard e $\sqrt{\mathbb{V}(\widehat{\mu})}$

Standard , $\sqrt{\mathbb{V}(\widehat{\sigma})}$

Expecte ue, $\mathbb{E}(\widehat{\sigma})$

Machine learning skips over the entire machinery of inference, and creates estimators that can best recover some aspect of the data. (*Statistical machine learning* brings theory back in, but for the purpose of seeing what best predicts, not what recovers information.)

What is “prediction”?

Prediction is not what you think

- *“**It’s not prediction at all!** I have not found a single paper predicting a **future result**. All of them claim that a prediction **could have been** made; i.e. they are post-hoc analysis and, needless to say, negative results are rare to find.”*

  – Daniel Gayo-Avello, “I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper” (2012)

Prediction is "fitted values"

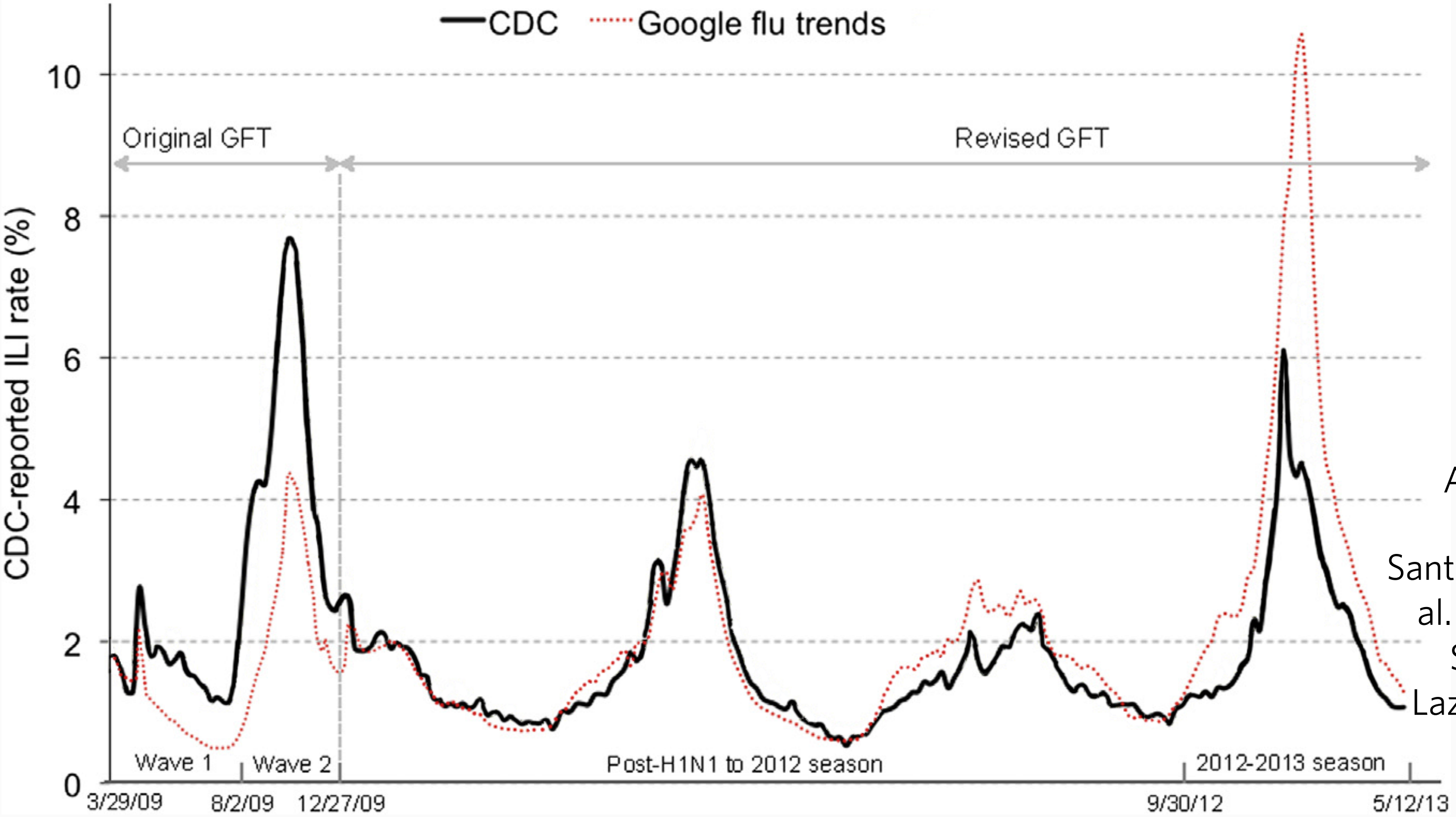
- "Predicted values" is a technical term synonymous with "fitted values," so in some sense Gayo-Avello is being unfair
- But defining predictions as fitted values only, and not accounting for change/intervention, is a confusing usage for laypeople and even for other scientists
- Read "We can predict X" instead as "We found a model that fits well"
- Fitting well is still an accomplishment, but it's quite different from actually being able to tell the future

Limitations

Correlation does not equal causation...

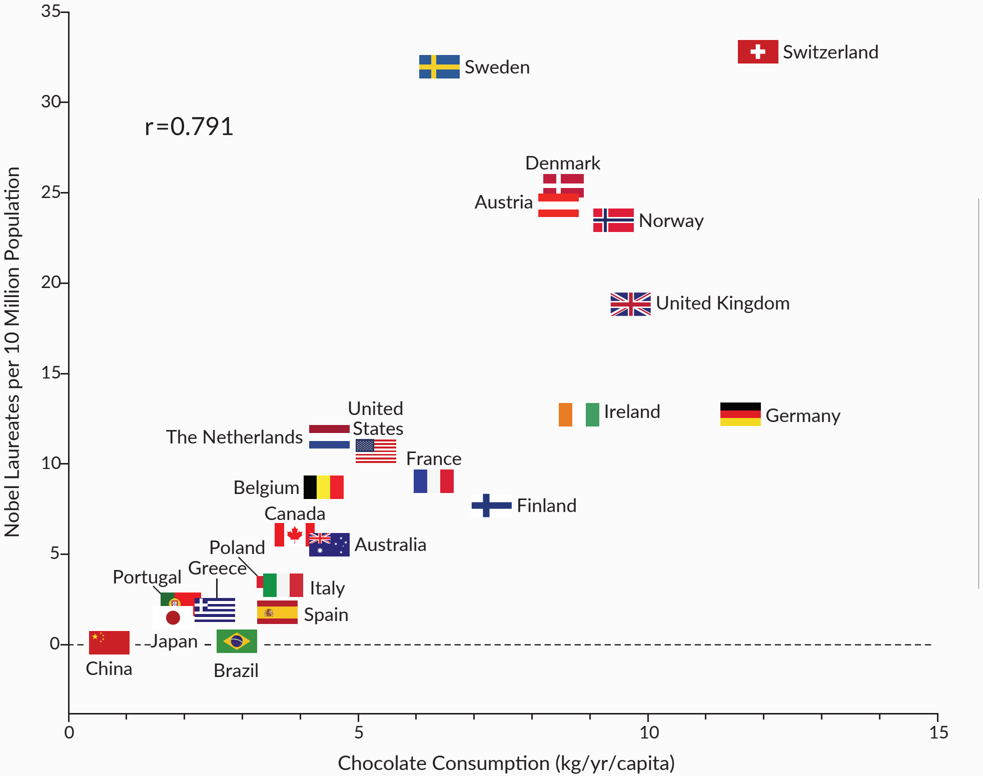
- And prediction does not mean explanation.
- Two issues: spurious correlations, and bias-variance tradeoff.
- Spurious correlations: A spurious (non-causal) correlation can be fairly robust, and a good basis for making “predictions”
- But this can be fragile

But correlations can be fragile



Adapted from Santillana et al. (2014). See also Lazer et al. (2014).

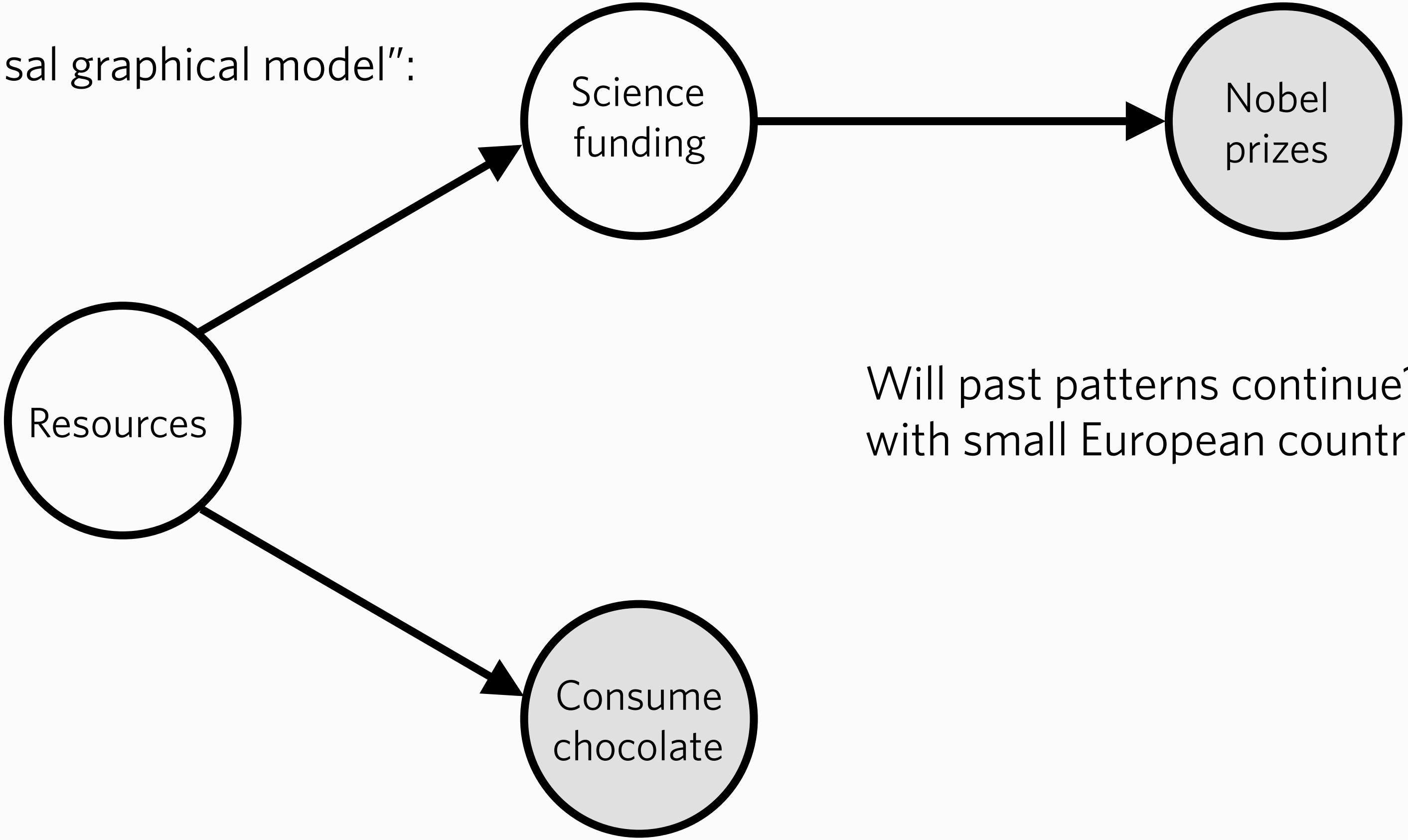
Ex: Chocolate and Nobel prizes



(Messerli, 2012)

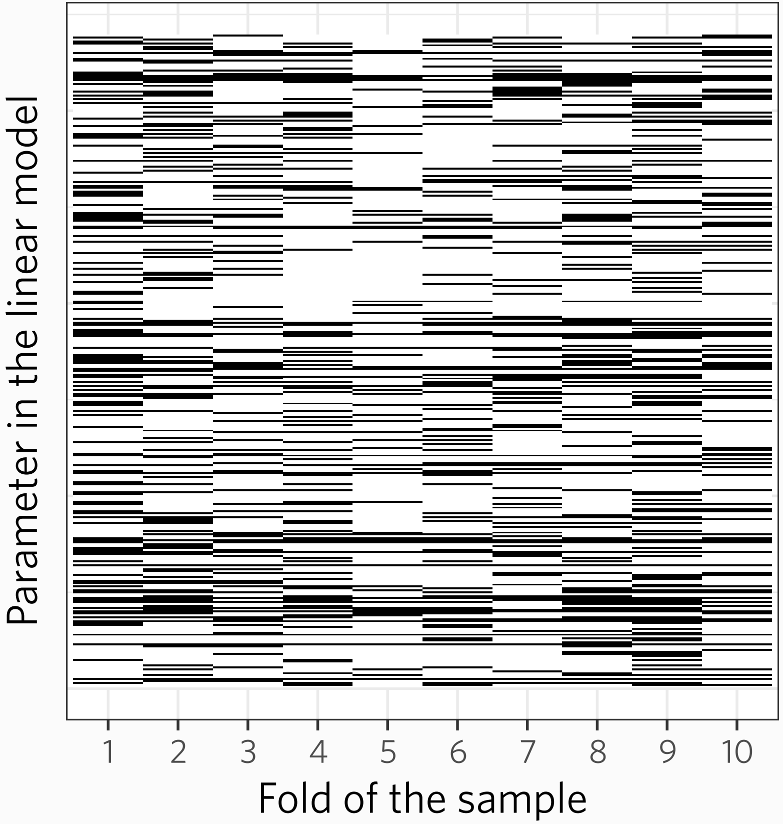
Underling *cause* can lead to failure

A "causal graphical model":



Will past patterns continue? E.g., with small European countries?

Can't *intervene* based on correlations



- Probably won’t win more Nobel prizes by feeding population more chocolate
- Very different sets of correlations can “predict” equally well (Mullainathan & Spiess, 2017)

Often, what we want is causality

- “The optimization objective for most supervised learning models... is simply to minimize error, a feat that might be achieved in a purely correlative fashion.” (Lipton & Steinhardt, 2018)
- “Because the models in this paper are intelligible, it is tempting to interpret them causally. Although the models accurately explain the predictions they make, they are still based on correlation.” (Caruana et al., 2015)
- “one can provide a feasible explanation that fails to correspond to a causal structure, exposing a potential concern.” (Doshi-Velez & Kim, 2017)
- “interpretation might explain the behavior of the model but not give deep insight into the causal associations in the underlying data... The real goal may be to discover potentially causal associations that can guide interventions.” (Lipton, 2015)

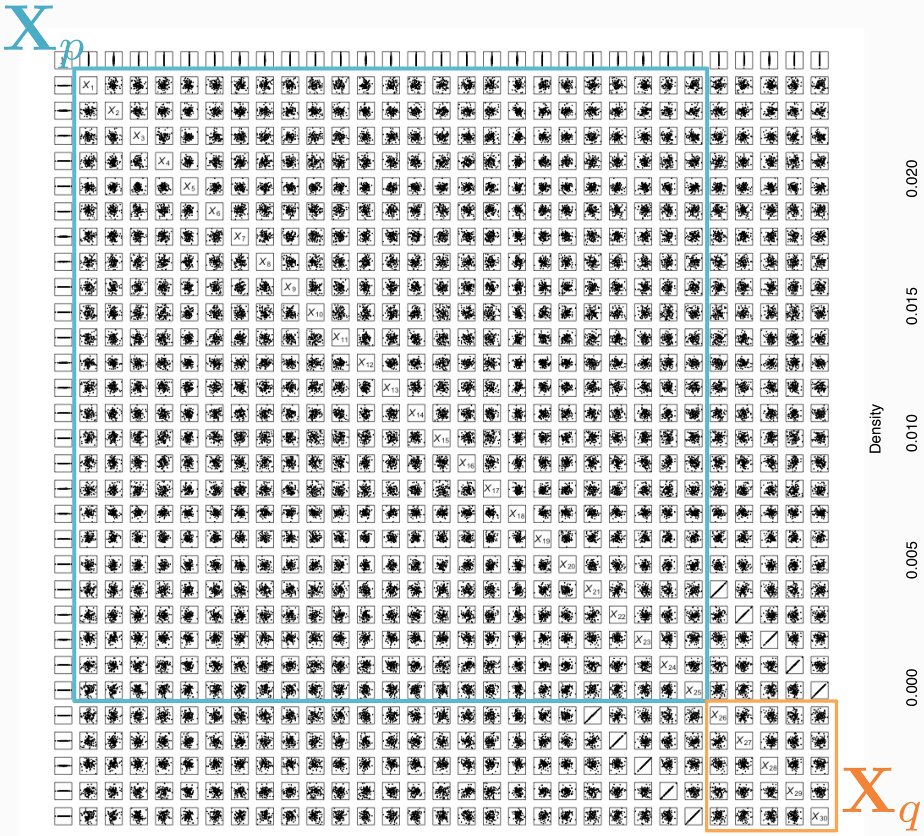
Bias-variance tradeoff

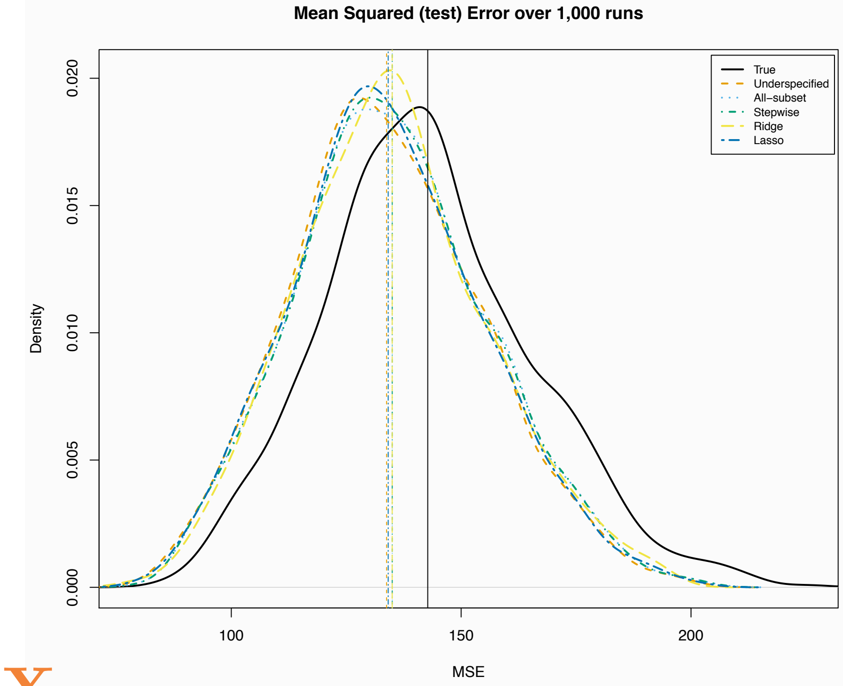
- Squared error loss decomposes into irreducible error (phenomenon) + bias squared + variance (of the estimator):

$$
\begin{aligned}
\text{EPE}(x) &= \mathbb{E}\left[\left(Y - \hat{f}(x)\right)^2 | X = x\right] \\
&= \mathbb{V}(Y) + \mathbb{E}\left[\left(\hat{f}(x) - f(x)\right)^2 | X = x\right] + \mathbb{E}\left[\left(\hat{f}(x) - \mathbb{E}[\hat{f}(x)]\right)^2 | X = x\right] \\
&= \sigma^2 + \mathrm{bias}^2\left(\hat{f}(x)\right) + \mathbb{V}\left(\hat{f}(x)\right)
\end{aligned}
$$

- σ^2 is the irreducible error (the variance of Y, beyond any signal from X)
- Bias: how far the estimator is from the true signal of X
- Variance: how noisy the estimator is (term has nothing to do with Y!)
- Turns out: bias that decreases variance can improve prediction!

The 'true' model can predict worse!

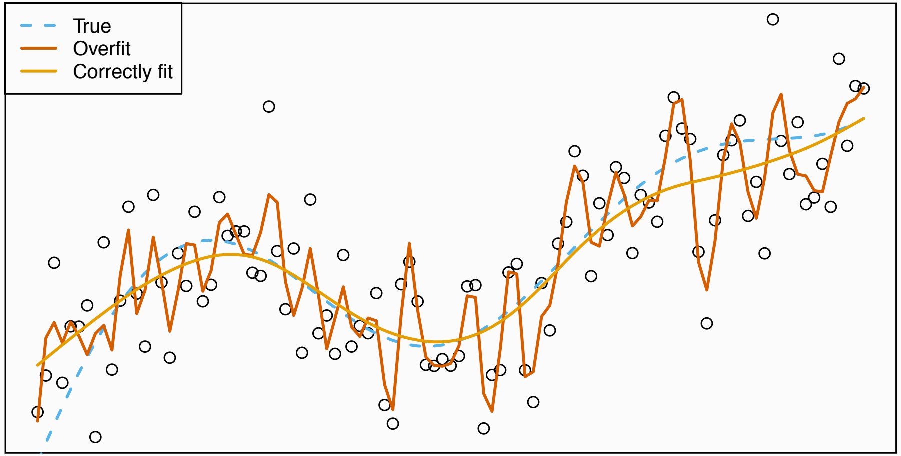




Simulation based on Shmueli (2010)

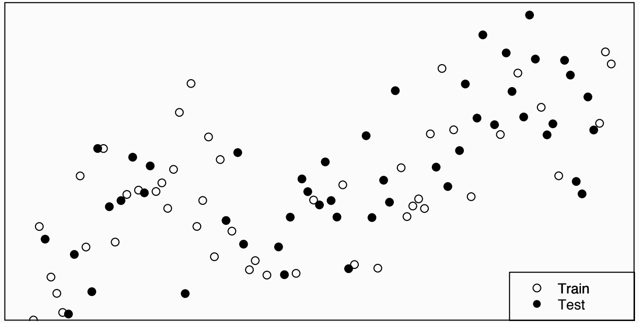
Cross-validation

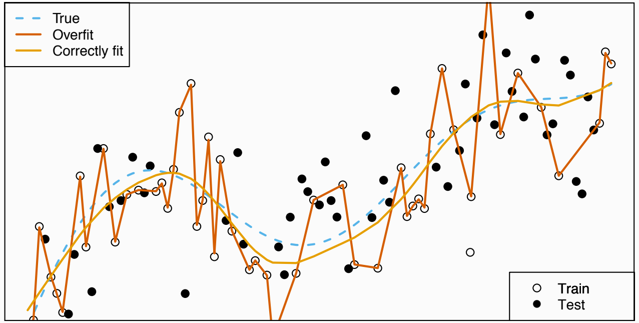
Overfitting: fit to noise



- If we are no longer guided by theory, and use automatic methods, we risk *overfitting*: fitting to the the noise, not the signal (“memorize the data”)

Data splitting: Catch overfitting





- Idea: if we split data into two parts, the signal should be the same but the noise would be different
- *Cross validation:* Fitting the model on one part of the data, and “testing” on the other
- ML lives and dies by cross-validation

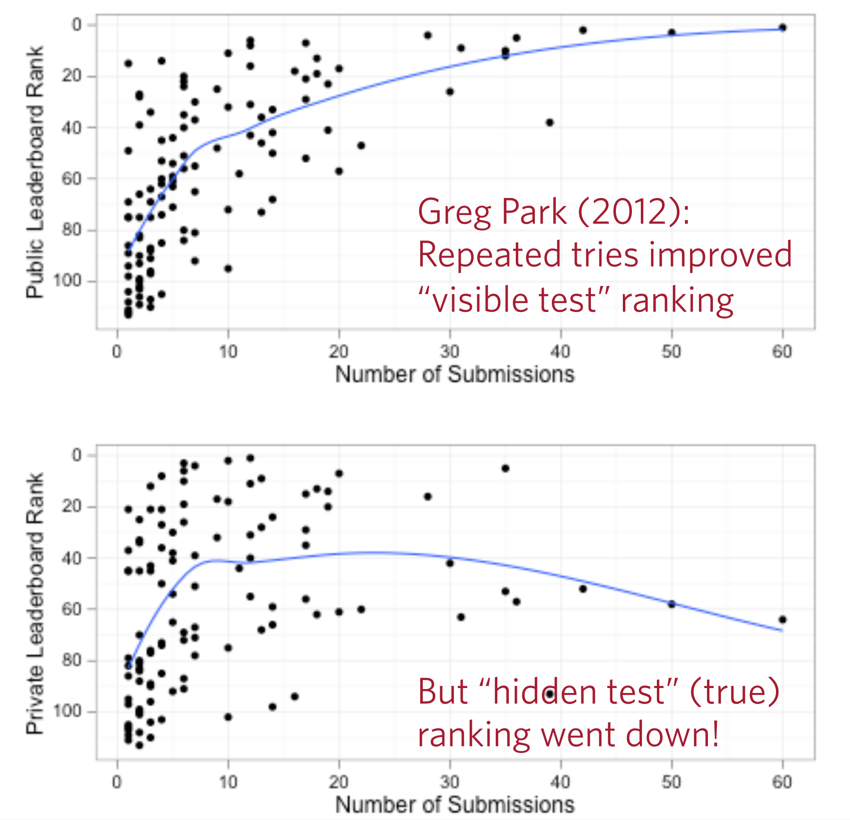
https://medium.com/greyatom/what-is-underfitting-and-overfitting-in-machine-learning-and-how-to-deal-with-it-6803a989c76

BERKMAN KLEIN CENTER
FOR INTERNET & SOCIETY AT HARVARD UNIVERSITY

- Introduction
- What is statistics?
- Reduce data to “relevant information”
- Individuality, meaning, and experience
- Quantifying uncertainty
- What is machine learning?
- “Prediction”
- Causality
- Cross-validation
- Ethics
- References

But cross-validation can fail

- Re-using a test set can overfit to the test set! Happens in Kaggle
- Or, if there are dependencies (temporal, network, group) between data splits, it “shares” information
- E.g., temporal: Fitting on values that come after test values is “time traveling”!



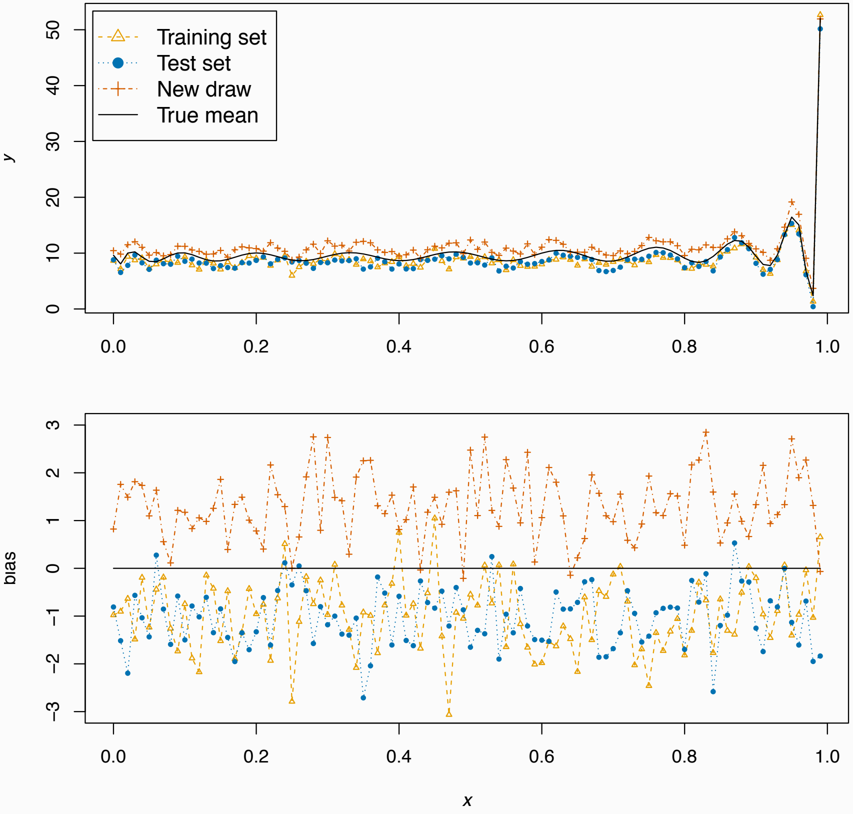
Greg Park (2012): Repeated tries improved “visible test” ranking

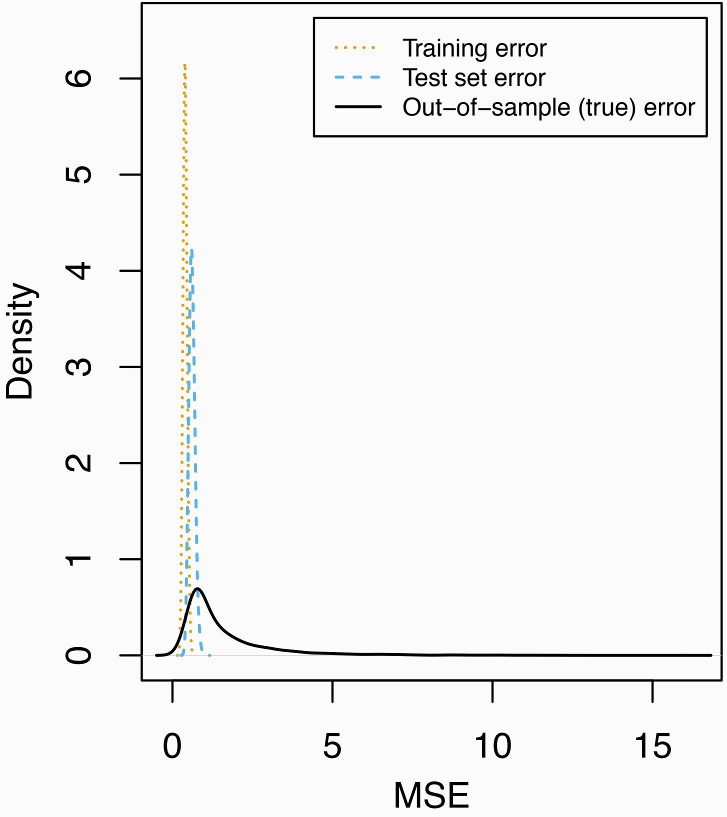
But “hidden test” (true) ranking went down!

BERKMAN KLEIN CENTER
FOR INTERNET & SOCIETY AT HARVARD UNIVERSITY

- Introduction
- What is statistics?
- Reduce data to “relevant information”
- Individuality, meaning, and experience
- Quantifying uncertainty
- What is machine learning?
- “Prediction”
- Causality
- Cross-validation
- Ethics
- References

Dependencies affect CV, too!





- Introduction
- What is statistics?
- Reduce data to “relevant information”
- Individuality, meaning, and experience
- Quantifying uncertainty
- What is machine learning?
- “Prediction”
- Causality
- Cross-validation
- Ethics
- References

Ethics

- Is it fair to determine effective rewards or punishments, e.g. insurance premiums, by correlations?
- That’s what the entire insurance industry is based on, and with ML, it’s only intensifying

References (1/3)

Breiman, Leo. "Statistical Modeling: The Two Cultures." *Statistical Science* 16, no. 3 (2001): 199–231. doi:10.1214/ss/1009213726

Caruana, Rich, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. "Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-day Readmission." In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD '15), 1721–1730. 2015. doi:10.1145/2783258.2788613.

Chatfield, Chris. "Model Uncertainty, Data Mining and Statistical Inference." *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 158, no. 3 (1995): 419–466. doi:10.2307/2983440.

Cox, David R. "Role of Models in Statistical Analysis." *Statistical Science* 5, no. 2 (May 1990): 169–174. doi:10.1214/ss/1177012165

Doshi-Velez, Finale and Been Kim. Towards a Rigorous Science of Interpretable Machine Learning. 2017. arXiv:1702.08608.

Fisher, R. A. "On the Mathematical Foundations of Theoretical Statistics." *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 222 (1922): 309–368. doi:10.1098/rsta.1922.0009

Gayo-Avello, Daniel. "'I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper': A Balanced Survey on Election Prediction using Twitter Data." 2012. *arXiv*:1204.6441v1.

References (2/3)

Lanius, Candice. “Fact Check: Your Demand for Statistical Proof is Racist.” Cyborgology blog, January 15, 2015. https://thesocietypages.org/cyborgology/2015/01/12/fact-check-your-demand-for-statistical-proof-is-racist/.

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. “The Parable of Google Flu: Traps in Big Data Analysis.” *Science* 343, no. 6176 (2014): 1203–1205. doi:10.1126/science.1248506.

Lipton, Zachary C. “The Myth of Model Interpretability.” *KDnuggets* 15, no. 13 (April 2015). https://www.kdnuggets.com/2015/04/model-interpretability-neural-networks-deep-learning.html.

Lipton, Zachary C. and Jacob Steinhardt. Troubling trends in machine learning scholarship. 2018. arXiv:1807.03341.

Marcus, Gary. “Artificial Intelligence Is Stuck. Here’s How to Move It Forward.” *New York Times* (29 July 2017). https://nyti.ms/2hav95E.

Messerli, Franz H. “Chocolate Consumption, Cognitive Function, and Nobel Laureates.” *The New England Journal of Medicine*, 367 (2012): 1562–1564. doi:10.1056/NEJMon1211064.

Mitchell, Tom M. *Machine Learning*. New York, NY: McGraw Hill, 1997.

Mullainathan, Sendhil and Jann Spiess. “Machine Learning: An Applied Econometric Approach.” *Journal of Economic Perspectives* 31, no. 2 (2017): 87–106. doi:10.1257/jep.31.2.87.

References (3/3)

Patton, Michael Quinn. "The Nature, Niche, Value, and Fruit of Qualitative Inquiry." In *Qualitative Research & Evaluation Methods: Integrating Theory and Practice*, 4th edition, 2–44. SAGE Publications, Inc., 2014. https://uk.sagepub.com/sites/default/files/upm-binaries/64990_Patton_Ch_01.pdf.

Rose, Todd. *The End of Average: How We Succeed in a World That Values Sameness*. New York: HarperOne, 2016. See excerpt at https://www.thestar.com/news/insight/2016/01/16/when-us-air-force-discovered-the-flaw-of-averages.html. Animated video: https://vimeo.com/237632676.

Santillana, Mauricio, Wendong Zhang, Benjamin M. Althouse, and John W. Ayers. "What Can Digital Disease Detection Learn from (an External Revision to) Google Flu Trends?" *American Journal of Preventive Medicine* 47, no. 3 (2014): 341–347. doi:10.1016/j.amepre.2014.05.020.

Shmueli, Galit. "To Explain or to Predict?" *Statistical Science* 25, no. 3 (2010): 289–310. doi:10.1214/10-STS330

Tibshirani, Robert. "Recent Advances in Post-Selection Inference." Breiman Lecture, NIPS 2015 (9 December 2015) http://statweb.stanford.edu/~tibs/ftp/nips2015.pdf

Wasserman, Larry A. "Rise of the Machines." In *Past, Present, and Future of Statistical Science*, 525–536. Boca Raton, FL: Chapman and Hall/CRC, 2013. http://www.stat.cmu.edu/~larry/Wasserman.pdf