ICQCM

CRITICAL DATA SCIENCE FOR A DIVERSE WORLD

Computational Approaches III: Applications

Momin M. Malik

Thursday, 22 July 2021

ICQCM 2021 Seminar Series

Manzel Bowman, *Station No.99* (2018)

Overview

- The "computational approach" here: machine learning, applied to social data
- Won't discuss simulation modeling, the other computational approach
- I assume:
  - Familiarity with social statistics/econometrics
  - Some familiarity with R (for demo/tutorial)
- Focus on key conceptual and practical things, usually covered poorly
  - When should we use machine learning? How do we use it?

Road map

- Going from statistics to machine learning
- When is machine learning appropriate?
  - "Prediction" problems
- Model selection in machine learning
  - Cross-validation
- Model evaluation in machine learning
  - Setting aside a test set
- Demonstration in R

Going from statistics to machine learning

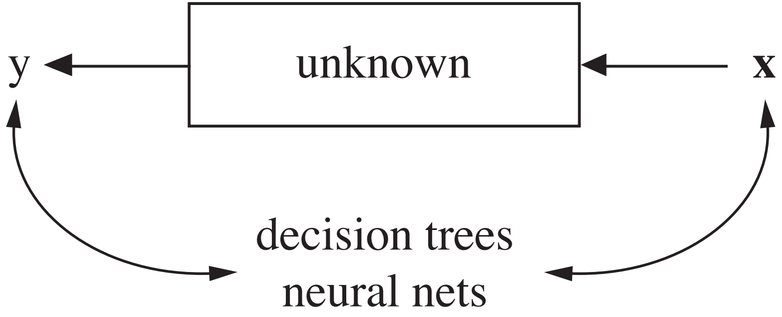
Machine learning is the instrumental use of correlations

Prediction and explanation are different goals and can be in conflict

Defining machine learning

Statistics:

Machine learning:



Machine learning: An instrumental use of statistical correlations to *mimic* the output of a target process, rather than understand the *relationship* between inputs and outputs. Involves finding expressions that maximize correlation.

Breiman 2001. See also Jones 2018.

Why are these different goals?

$$\widehat{y}$$

Spurious (non-causal) correlations may fit robustly

- Breiman 2001: Prediction problems
- Shmueli 2010: To predict
- Kleinberg et al. 2015: "Umbrella problems"
- Mullainathan and Spiess 2017: y-hat

$$\widehat{\beta}$$

Carefully built models that capture causality (or "pure" associations) may fit poorly overall

- Breiman 2001: Information
- Shmueli 2010: To explain
- Kleinberg et al. 2015: "Rain dance problems"
- Mullainathan and Spiess 2017: beta-hat

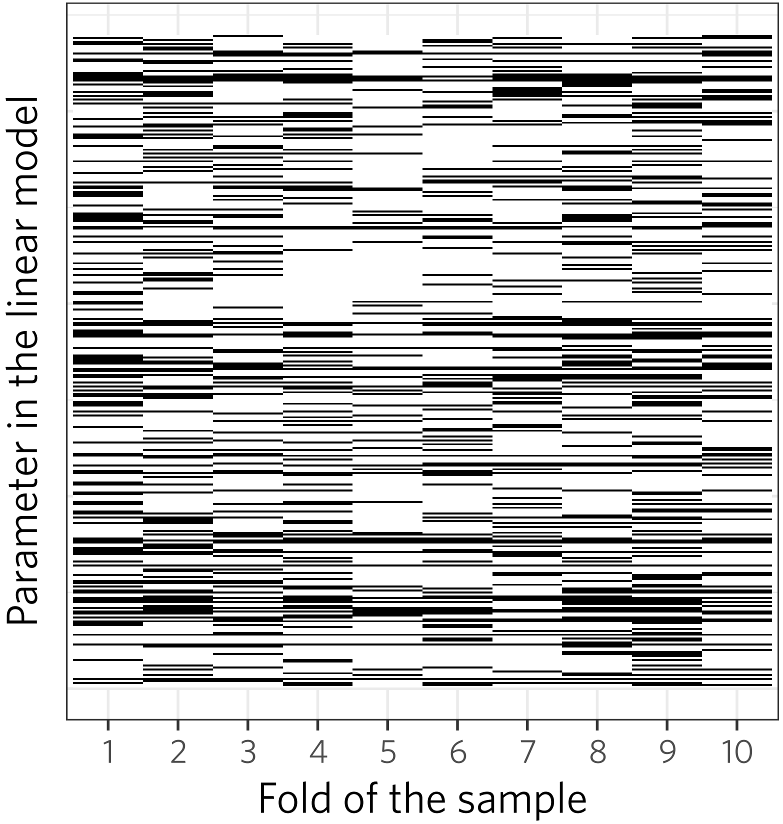
ICQCM

The surprising part

- *The best-fitting* (most accurate*) *model does not necessarily reflect how the world works*
- This has been shocking in statistics for decades (Stein's paradox, Leo Breiman's "two cultures"), but little known outside
- Why: one reason is the "bias-variance tradeoff"
  - Even when available, the "true" covariates may be noisy, in which case proxies (or even just going with the mean) sometimes does better
- Another reason: narrowing in to get one causal relationship "correct" might require sacrificing the rest of the model
- So: we can use correlations to "predict" without "explaining" (knowing causality)!

* Or other relevant metric of success

But: can't *intervene* based on correlations



- Very different sets of correlations can "predict" (fit) equally well (Mullainathan and Spiess 2017)
  - Leo Breiman (2001) called this the "Rashomon Effect"
- But different fits suggest very different interventions

So what is ML useful for? *Building systems*

- Recommend/narrow people's choices to "relevant" ones (friend connections, search results, products)
- Detection (facial, fraud)
- Anticipation (customer demand, equipment failure)
- It "works"…

How? Correlates *labels* and other data

"Source subject": Marquese Scott

Caroline Chan, "Everybody Dance Now: Motion Retargeting Video Subjects." https://youtu.be/PCBTZh41Ris

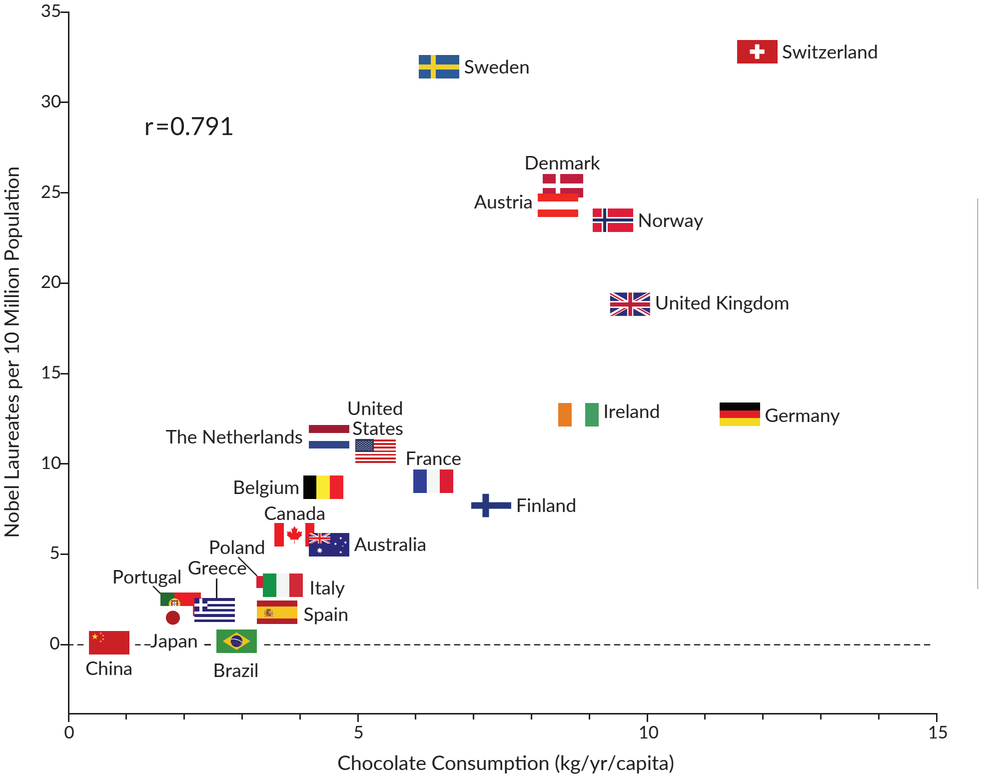
Key differences/comparisons

- ML overlaps lots with *nonparametric statistics*, which (for example) gets models by locally smoothing input data rather than doing global fits. But ML also uses parametric models, and lots of Bayesian models (although in decidedly non-Bayesian ways)
- ML: no statistical inference, and so doesn't need to calculate standard errors. Opens up modeling possibilities without that extra complication
- ML: focuses on *classification*, i.e. categorical responses. This is easier (only need to be on the 'correct' side of the 'true' underlying decision boundary)
- Even just in terms of pure model fit, does ML beat stats for social questions? Not always! (Junqué de Fortuny et al. 2013; Salganik et al. 2020; Garip 2020)
  - Note: deep learning only works for audio, images, and (sometimes) time series. For general forms of data, random forests are often the best (Caruana et al. 2008; Fernández-Delagado et al. 2014)
- Caution: statistical significance is not the same as feature importance!

ICQCM

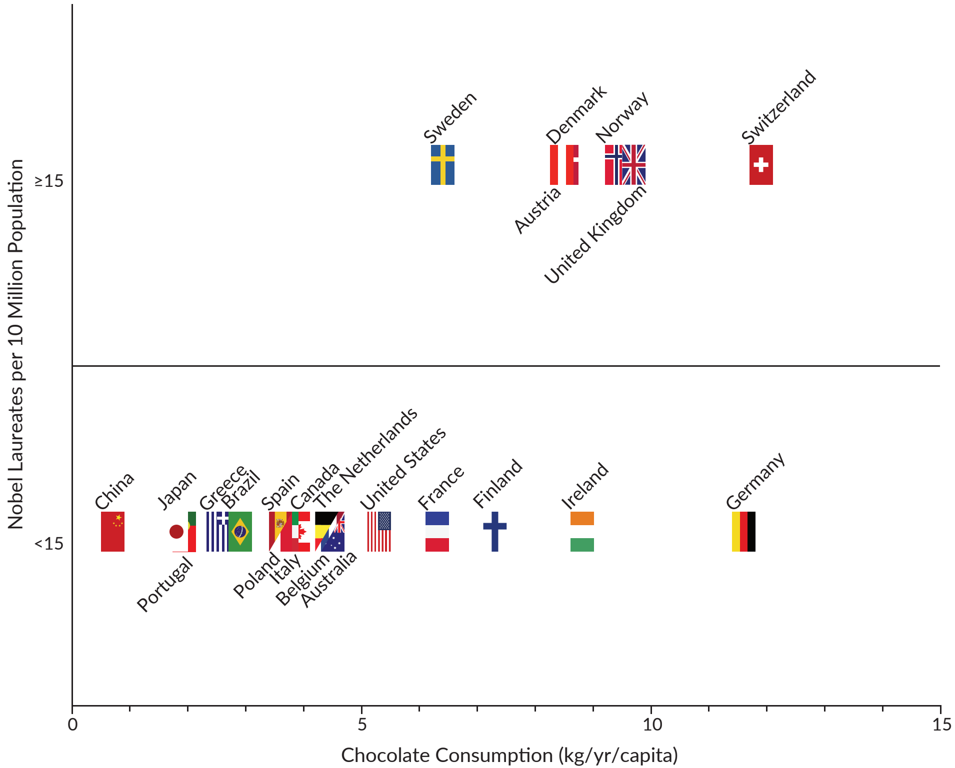
- Overview
- **Going from statistics to machine learning**
- When to use machine learning
- Key concepts
- Example for demo: *Titanic*
- Demo/ Tutorial
- Extra: Problems with explainability
- References

Regression: Continuous relationships

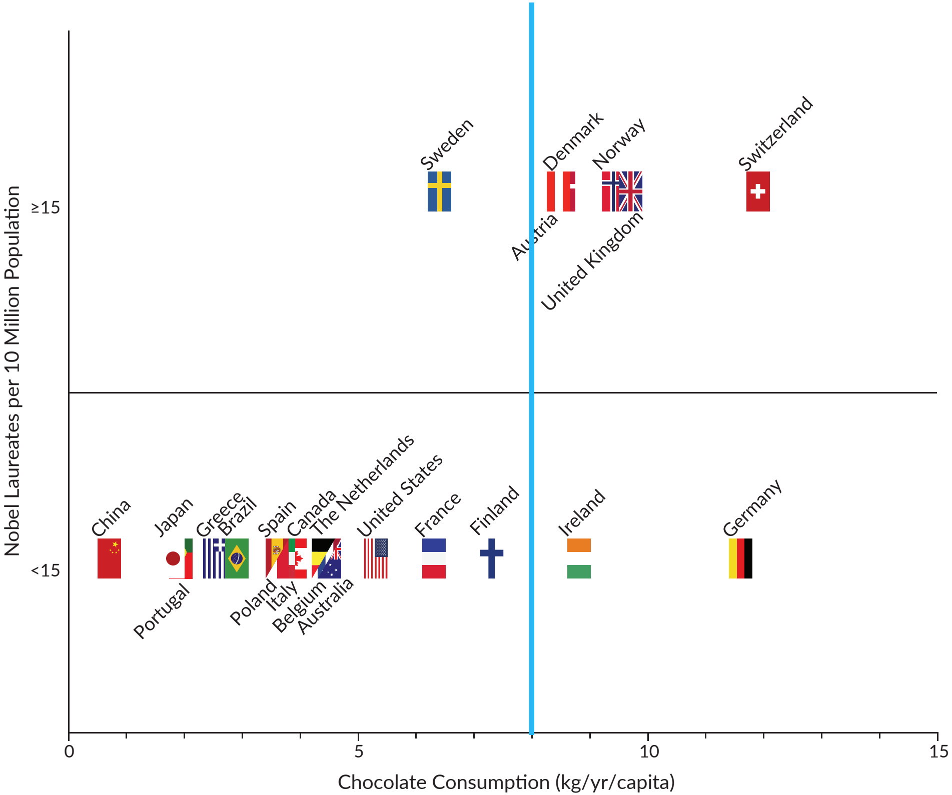


Messerli 2012

Classification: Discrete relationships

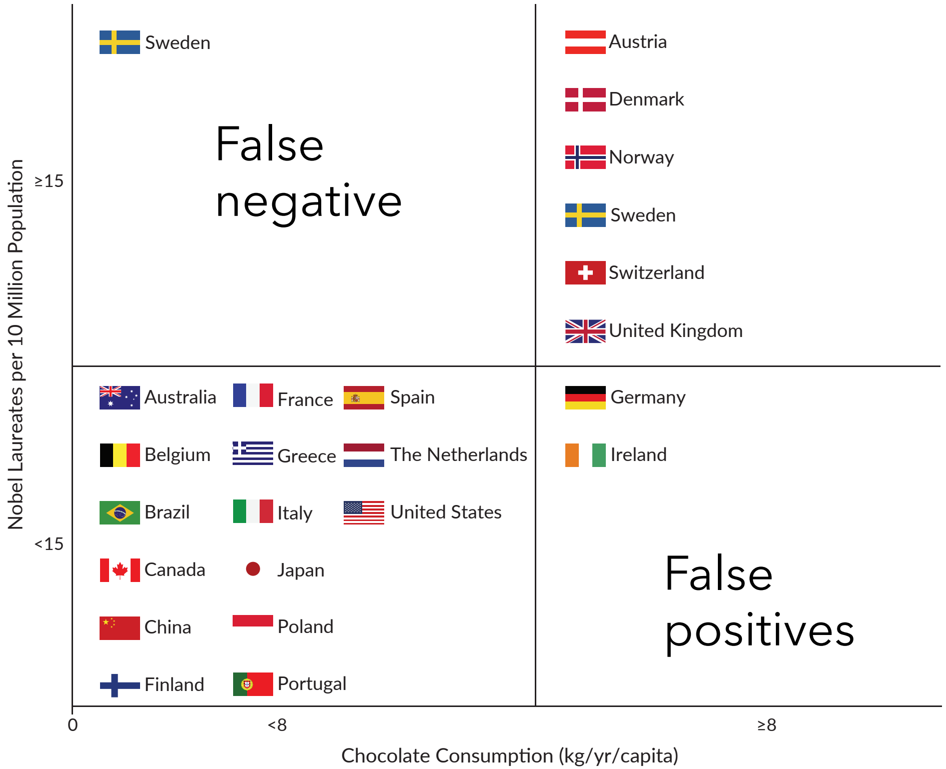


Fit a decision boundary



ICQCM

The prediction: the majority class



When to use machine learning

Key components of a good use case

Example of a “responsible” use case

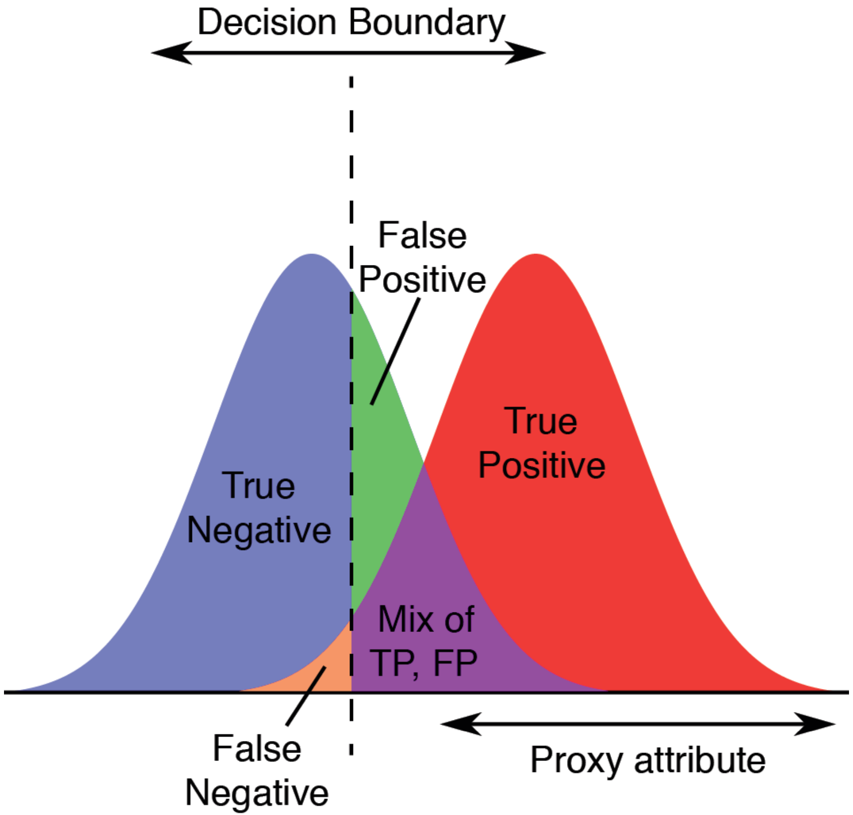
Key components of a good use case

1. We have reliable “ground truth” (e.g., human labels, previous failures/fraud);
2. “Ground truth” is hard to collect;
3. In the future or other contexts, “ground truth” is unknown but could be used if known;
4. We have some readily available proxy measure; and
5. *We don’t care how or what in the proxy recovers the “ground truth”, only that it does*

If we care about relationships between inputs and outputs, ML is useless (except for exploration)

ICQCM

ML model = "Ground truth" + proxy

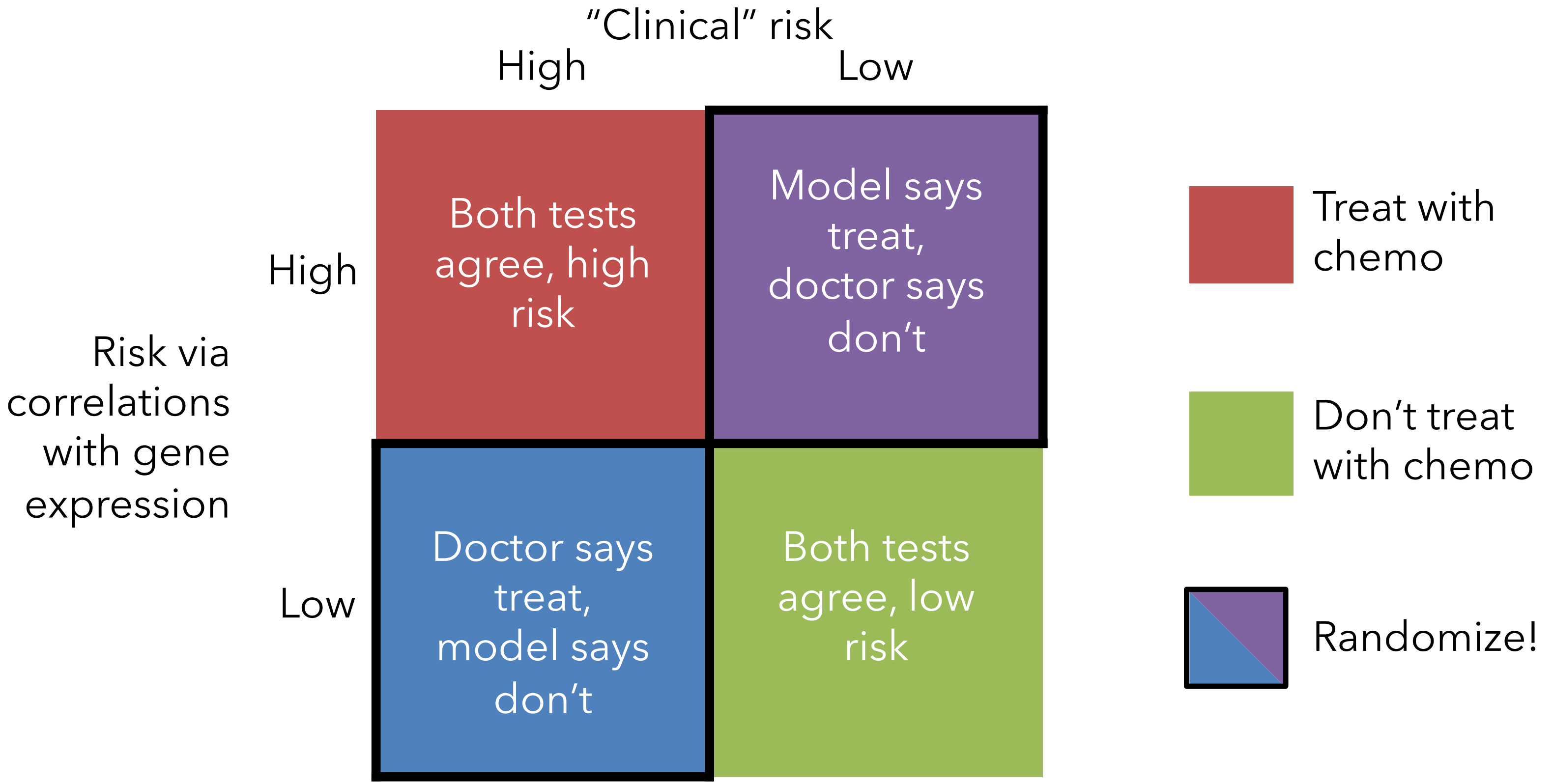


- Correlate known values/labels with available proxy for unknown values/labels
- Find *decision boundary/criterion/threshold*. Use this to treat new observations
- Shift that boundary to prioritize certain metrics
- Most ML is basically this!

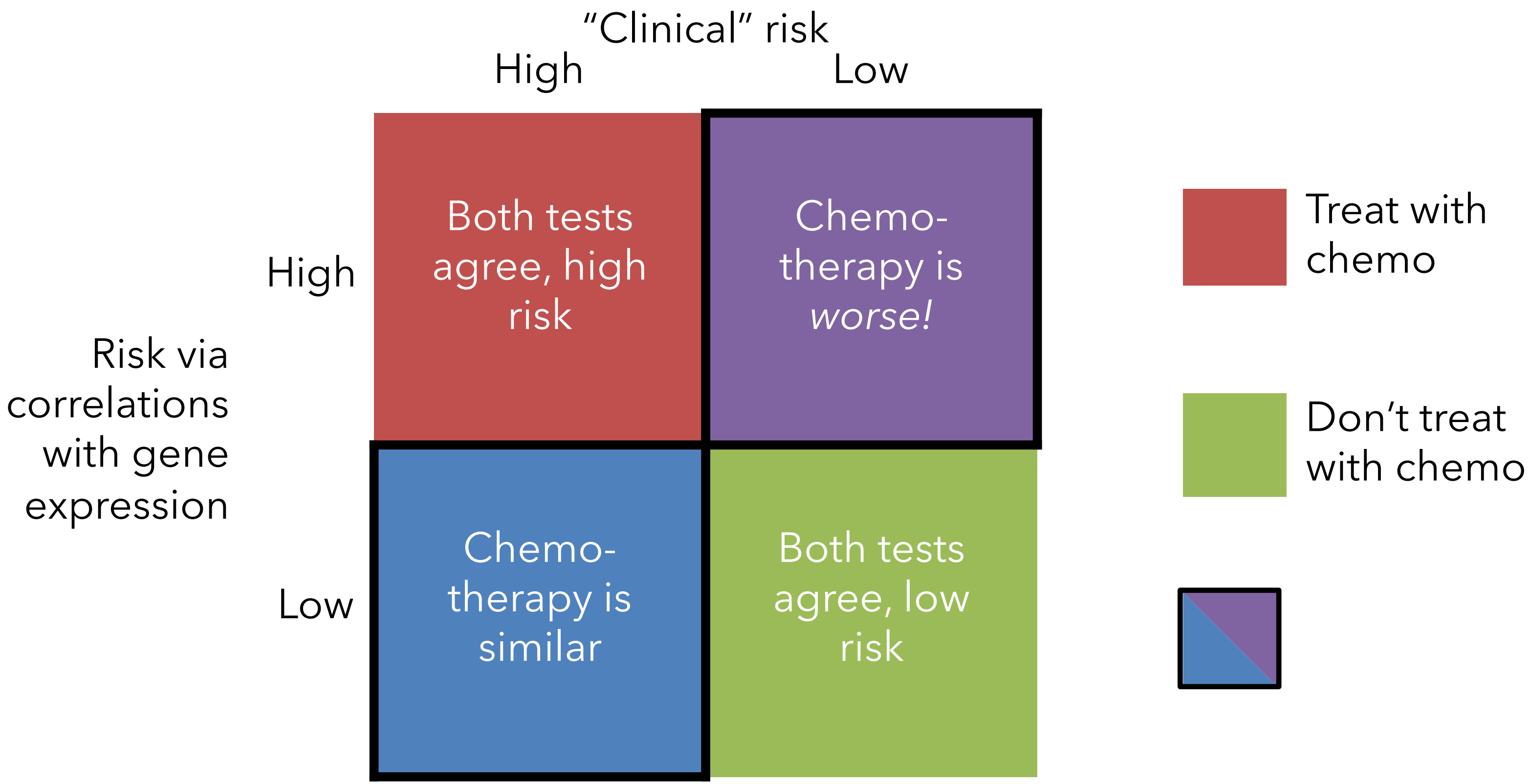
"Responsible" use case

- Baseline: Clinical diagnosis of breast cancer
- Researchers built a machine learning model that correlated gene expressions with developing breast cancer (van't Veer et al. 2002)
- Which is better? Experimentally test! (Cardoso et al. 2016)

Real-world testing



Real-world testing



Real-world testing

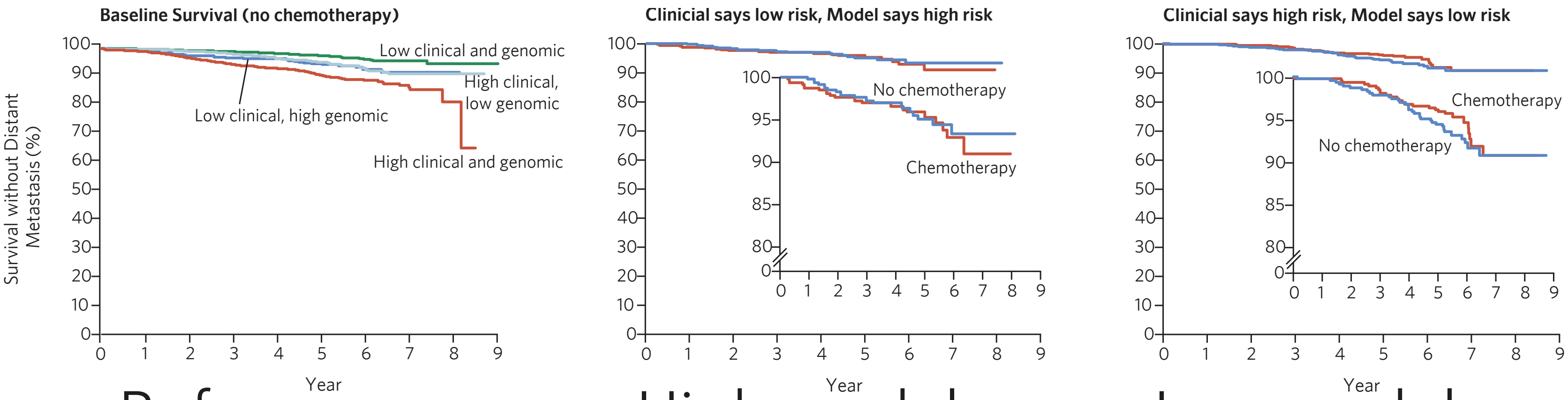
Risk via correlations with gene expression	"Clinical" risk: High	"Clinical" risk: Low
High	Both tests agree, high risk	Chemo-therapy is *worse!*
Low	Chemo-therapy is similar	Both tests agree, low risk

Treat with chemo (red)

Don't treat with chemo (green)

(Still: whose data went into the model? Who were the subjects in the experiment?)

Real-world testing: Details



- Before experiment (training data)
- High model risk, low clinical risk: randomize. Chemo *worse*!
- Low model risk, high clinical risk: chemo makes no difference

Cardoso et al. 2016

Key points

- Machine learning should not be used to say something about the way the world works; it should only be used instrumentally
- Machine learning is like training to win a race: it's only meaningful if we actually run the race! (see also Gayo-Avello 2012)
- (More on this later): *machine learning performance claims are always preliminary until we do real-world testing*

ICQCM

Overview

Going from statistics to machine learning

When to use machine learning

Key concepts

Example for demo: *Titanic*

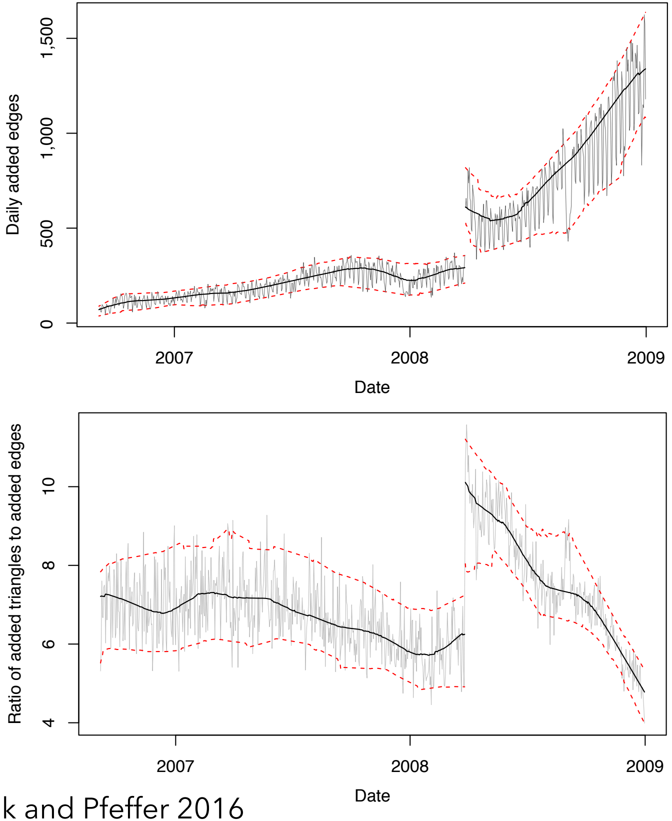
Demo/ Tutorial

Extra: Problems with explainability

References

What good is ML for social science? (1/3)

- For exploratory analysis, especially of "high-dimensional data"
  - Topic models for text corpuses
- Nonparametric models (which may be labeled as "machine learning" but, if they quantify uncertainty, I'd call them statistical) are useful for modeling complex bivariate relationships.
  - Substantive analysis and interpretation can only be done visually, so it's not really useful beyond bivariate relationships



Malik and Pfeffer 2016

Overview

Going from statistics to machine learning

When to use machine learning

Key concepts

Example for demo: *Titanic*

Demo/ Tutorial

Extra: Problems with explainability

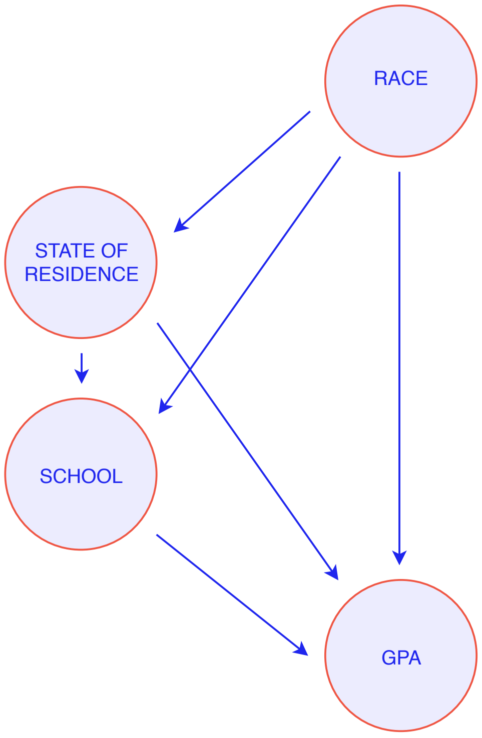
References

What good is ML for social science? (2/3)

- For *scaling up* human labels to a larger dataset (example in Malik 2018)
  - Let's say you have 1m tweets
  - Hand-code 1000 tweets between 3 coders, coding for whatever you care about, and make sure Cohen's kappa is sufficient as usual
  - Extract "*n*-gram" features
  - Fit a random forest to 500 observations; test on the remaining 500; report the accuracy, precision, and recall
  - Re-train a model with all 1000 labeled cases, use that to make "predicted" labels for the remainder of the data
  - Then you can make frequency statements about the presence of codes within the 1m tweets
  - (Ideally, also give confidence intervals on those frequency statements that take into account uncertainty from the imperfect model, and from the lack of perfect agreement among coders)

What good is ML for social science? (3/3)

- Caveat: Graphical models came out of machine learning, and can express complex causal structure (are equivalent to Structural Equation Models)
- Note: can express causal structure, not find/discover it (Malik 2020)
  - And ultimately "expresses" causality in a very limited way (Richardson 2020): e.g., graphical models with race counterfactuals are nonsensical from a constructivist view (Hu 2019a, 2019b, 2020)
- But within machine learning, graphical models are seldom used for causal modeling. So I don't count causal graphs as machine learning, despite origins



Hu 2019b

Key concepts

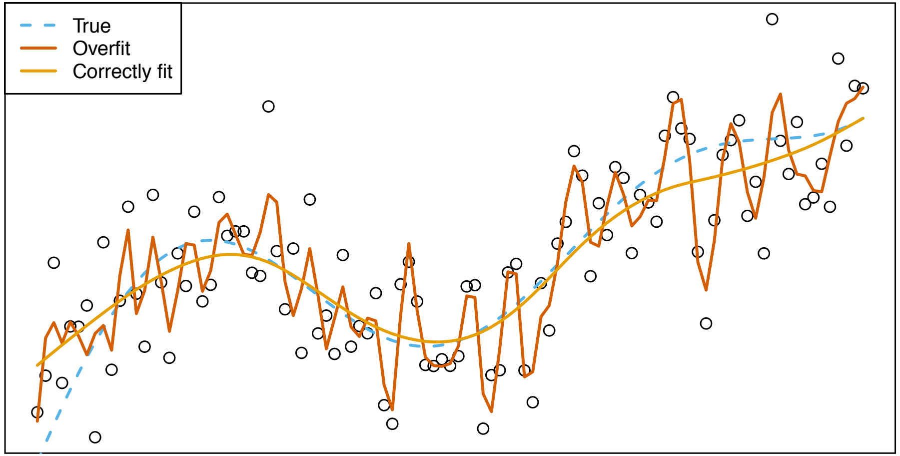
Model fit
Overfitting
Data splitting
Accuracy paradox
Confusion matrix

Model "fit"

- All machine learning and statistics models take in data, process them via some assumptions, and then give out something: relationships, and/or likely future values.
- The processing is called "fitting", and the output is called a "fit." Machine learning uses "learning" or "training," but it's the same.

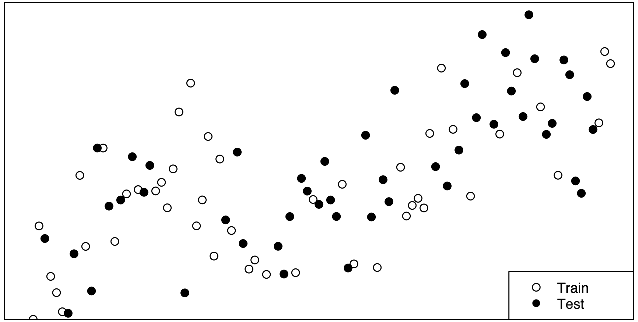
ICQCM

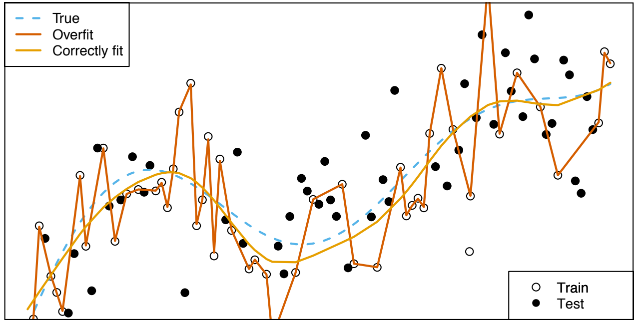
Overfitting: fit to noise



- If we are no longer guided by theory, and use flexible, automatic methods, we risk *overfitting*: fitting to the the noise, not the signal ("memorizing the data"). Applies to ML and nonparametric stats

Data splitting: Catch overfitting

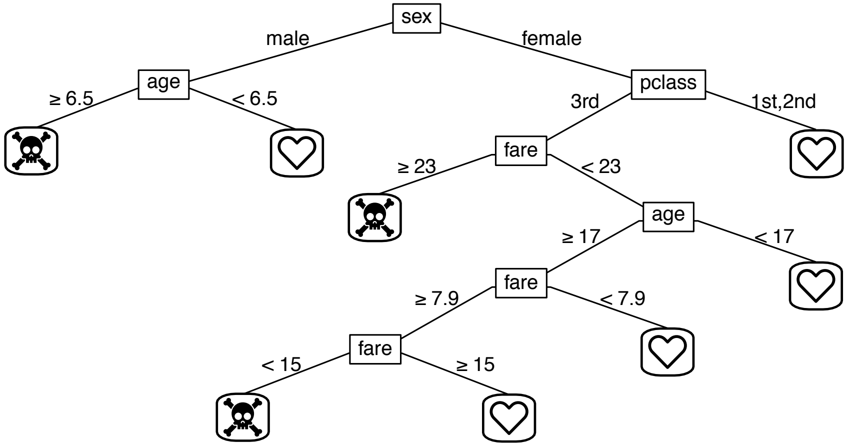




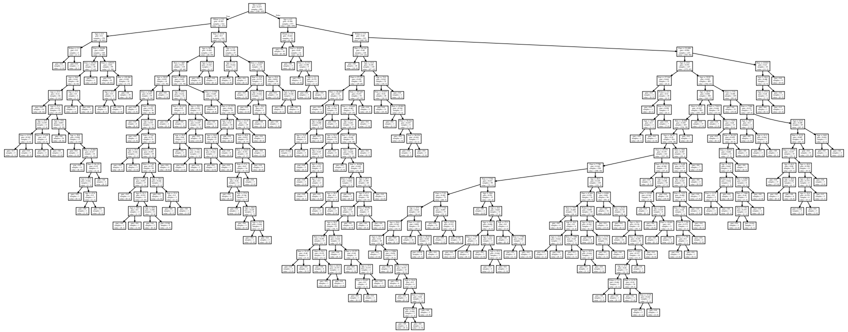
- Idea: if we split data into two parts, the signal should be the same but the noise would be different
- *Cross validation:* Fitting the model on one part of the data, and "testing" on the other to catch overfitting
- Or, fitting on one partition of the data, "tuning" on a second partition of the data such that we don't overfit, and then testing on a third partition of the data to make sure we succeeded

(Overfitting in a classification tree)

A non-overfitted classification tree (does almost as well on test data as on training)



An overfitted classification tree (does much worse on test data than on training): the default Python sklearn parameters produces this!!



Evaluation: "Accuracy paradox"

- Say, 5 out of 1000 observations are positive ("extreme class imbalance")
- A classifier that always predicts negative is 99.5% accurate, but useless
- Other metrics are more meaningful
- Use the *confusion matrix*

Confusion matrix

		True label	
	N	Positive	Negative
Predicted label	Predicted positive	True positive	False positive
	Predicted negative	False negative	True negative

Confusion matrix

		True label	
	N	Positive	Negative
Predicted label	Predicted positive	True positive	False positive
	Predicted negative	False negative	True negative

Accuracy = (TP+TN)/N

↑Overall correct

Confusion matrix

		True label	
	N	Positive	Negative
Predicted label	Predicted positive	True positive	False positive
	Predicted negative	False negative	True negative
		Recall/ sensitivity = TP/(TP+FN)	← How many you detect

Accuracy = (TP+TN)/N

↑Overall correct

Confusion matrix

True label

	N	Positive	Negative		Accuracy = (TP+TN)/N
Predicted label	Predicted positive	True positive	False positive	Precision = TP/(TP+FP)	↑Overall correct
	Predicted negative	False negative	True negative	↑How much is relevant	
		Recall/ sensitivity = TP/(TP+FN)	← How many you detect		

Confusion matrix

True label

	N	Positive	Negative		Accuracy = (TP+TN)/N
Predicted label	Predicted positive	True positive	False positive	Precision = TP/(TP+FP)	↑Overall correct
	Predicted negative	False negative	True negative	↑How much is relevant	
		Recall/ sensitivity = TP/(TP+FN)	← How many you detect		
		How many→ you correctly reject	Specificity = TN/(TF+TN)		

Confusion matrix

True label

	N = 165	Positive: 105	Negative: 60		Accuracy = 0.91
Predicted label	Predicted positive: 110	TP = 100	FP = 10	Precision = 0.91	↑Overall correct
	Predicted negative: 55	FN = 5	TN = 50	↑How much is relevant	
		Recall/ sensitivity = 0.95	← How many you detect		
		How many→ you correctly reject	Specificity = 0.83		

Confusion matrix

True label

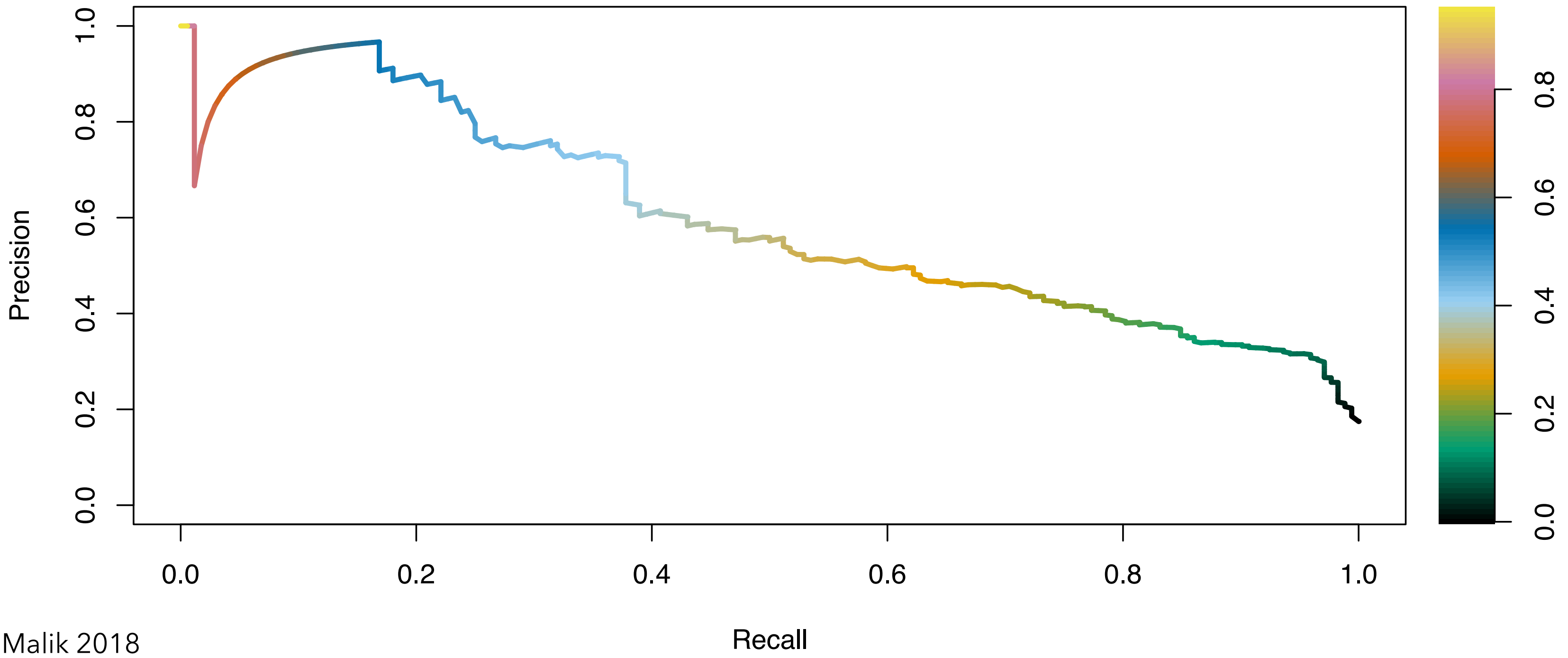
	N = 165	Positive: 105	Negative: 60		Accuracy = 0.91
Predicted label	Predicted positive: 110	TP = 100	FP = 10	Precision = 0.91	↑Overall correct
	Predicted negative: 55	FN = 5	TN = 50	↑How much is relevant	
		Recall/ sensitivity = 0.95	← How many you detect		
		How many→ you correctly reject	Specificity = 0.83		

Other metrics: F1 score, Area Under the [ROC] curve, Matthews correlation coefficient…

ICQCM

Trade-offs between metrics

Most models give fitted probabilities between 0 and 1 (or something that can be converted into fitted probabilities, like odds ratios). The *accuracy* is maximized if we take the decision boundary of 0.5, but if we care more about precision or recall, we can shift that boundary to prioritize one or the other. Precision-recall curves capture one such tradeoff.



Malik 2018

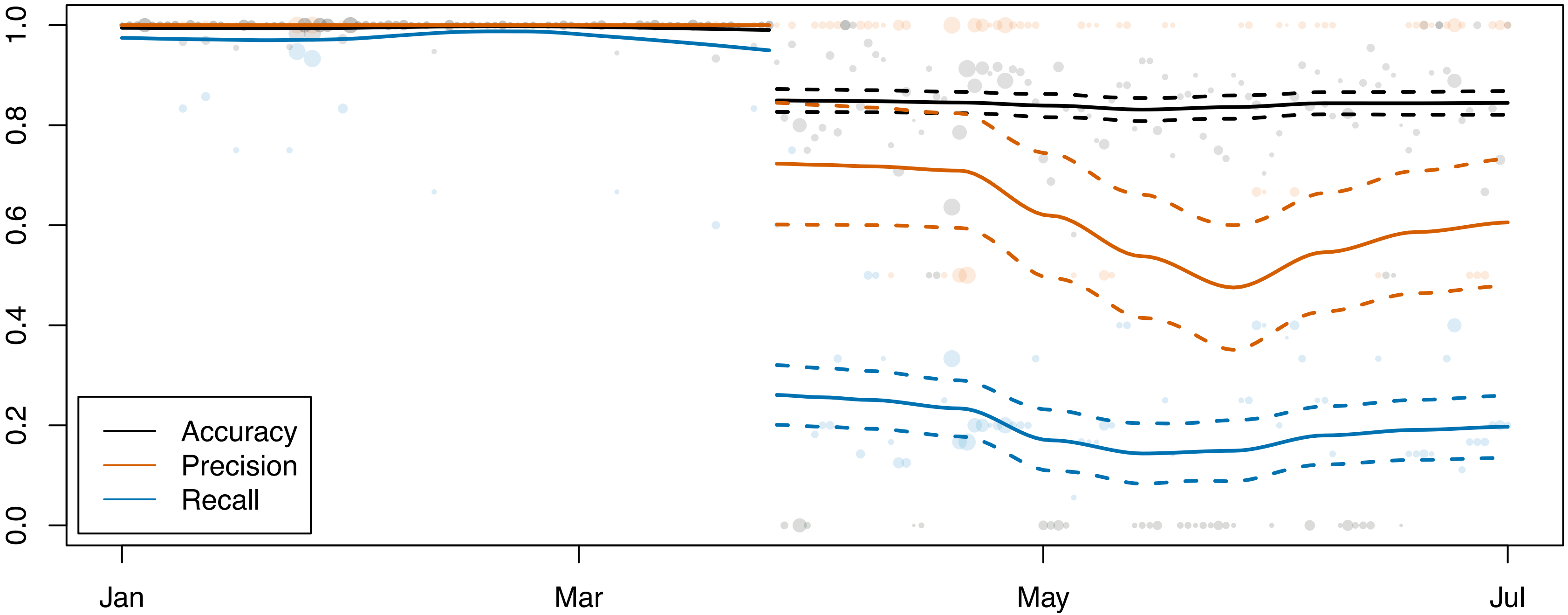
ICQCM

Doing data splitting correctly

- Data splitting is used for two distinct things in machine learning: model selection and model evaluation
  - Selection could be between *model class*, like between a logistic regression and a decision tree; or it could be selection of *tuning parameters*, like the bandwidth of local polynomial regression
- Data splitting for selection has very different theoretical properties than for evaluation
  - *k*-fold cross validation is only valid for model selection; for model evaluation, completely set aside some data for testing at the very end
- For both: *want to split in a way that respects dependencies*
- E.g., random splits of a time series (versus training only on the past) means you use future values to “predict” past ones
  - “Time-traveling”
  - Not a realistic test of out-of-sample performance

ICQCM

Doing data splitting correctly



Malik 2018

ICQCM

Feature engineering

- In social science, we have the variables (e.g., the survey responses)
- In machine learning, you might have lots of text data, or lots of sensor data, for a single outcome
- “Feature engineering”: heuristics to extract variables to summarize the data. Huge part of ML, no systematic solution for every data type
- Deep learning exciting because it does “automatically”, but only for very specific data types

ICQCM

Statistics on machine learning results

- Test error is an *estimator* of the generalizability error
- We can get a confidence interval around it! Can do significance testing!
  - McNemar's test: can be applied to the confusion matrix
  - When in doubt, can always try bootstrapping
- It can be biased! E.g., by selection bias, endogeneity…
- Kleinberg et al. (2017) use an instrumental variable (judge leniency) to try get and eliminate bias in test error caused by selection effects: i.e., econometrics techniques can be applied to ML performance!

Example for demo: *Titanic*

I put together multiple versions of this dataset to get something complete, and to get the test cases that Datacamp/Kaggle exclude, at https://www.mominmalik.com/titanic.csv

ICQCM

Datacamp "Titanic" example

WHAT'S THE MATTER WITH OUR BALL TEAM?

Unsettled Tuesday, Wednesday fair and cooler; moderate southerly winds, becoming variable.

Average temperature yesterday 54
Average temperature for corresponding date last year..53
Average temperature for corresponding date last thirty-three years 48

New York American

EDITION FOR GREATER NEW YORK.

THE TWENTIETH CENTURY NEWSPAPER

BUSINESS PROPERTY

TO LET. The best Variety in Every Section of City. Advertised Every Day and Sunday in AMERICAN "WANT AD" PAGES

NEW YORK AMERICAN, Published by Star Company, at 232-238 William st. President, S. S. Carvalho; Treasurer, Bradford Merrill; Secretary, Caleb Van Hamm; address, 238 William st. Main Advertising Office, 1 Park pl., cor. B'way.

No. 10,499. Copyright, 1912, by the Star Company. Registered in U. S. Patent Office. TUESDAY, APRIL 16, 1912. 16 PAGES PRICE ONE CENT In Greater New York and Jersey City. Elsewhere TWO CENTS.

J. J. ASTOR LOST ON TITANIC

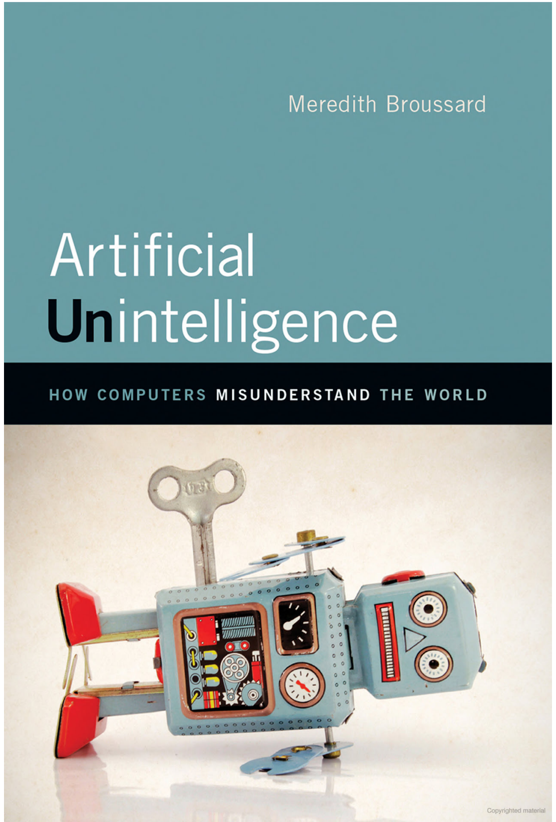
1,500 TO 1,800 DEAD

John Jacob Astor was among the passengers who went down with the ship, according to a wireless dispatch received by Bradstreets last night from the liner Olympic. Mrs. Astor was saved and is being brought to shore by the Carpathia.

The Wireless Operator at Cape Race, Newfoundland, Flashes: "Eighteen Hundred Lives Have Been Lost in the Wreck of the Titanic."

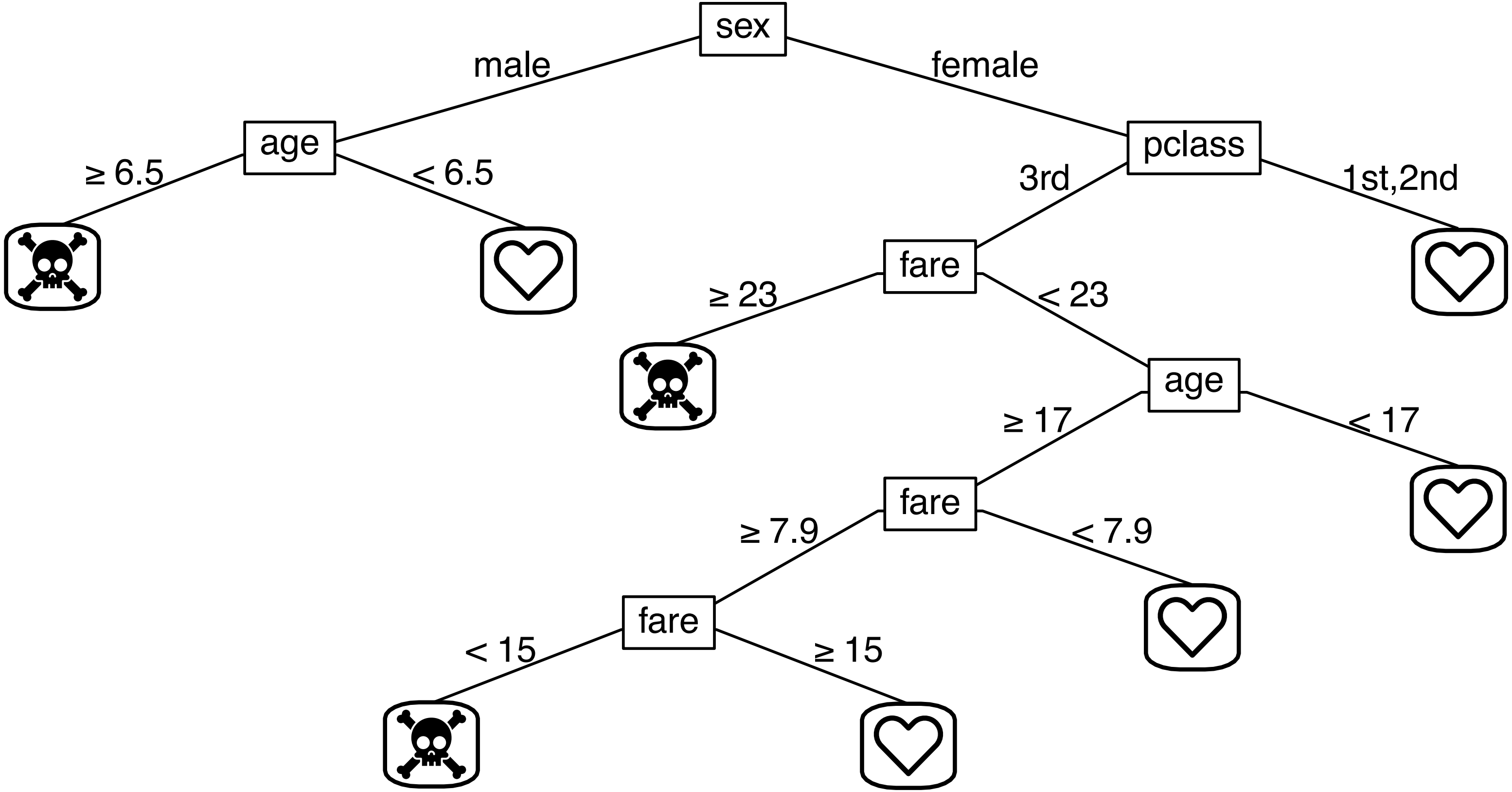
ICQCM

Broussard's Commentary



- Captain: "Put the women and children in and lower away."
- First Officer (starboard): women and children *first*
- Second Officer (port): women and children *only*
- "the lifeboat number isn't in the data. This is a profound and insurmountable problem. Unless a factor is loaded into the model and represented in a manner a computer can calculate, it won't count… The computer can't reach out and find out the extra information that might matter. A human can."
- (Original dataset does have lifeboat number; but even if we did feature engineering for odd/even, we don't know who wasn't allowed into a lifeboat!)

Fit a "decision tree" for survival



ICQCM

Social science baseline for comparison

- 5 econometrics papers from Frey, Savage, and Torgler (2009-2011) give a comparative "social statistics" approach

Compare: narrative and "prediction"

- Joseph Philippe Lemercier Laroche
- Haitian engineer
- Married French woman, Juliette Lafargue
- Denied jobs in France
- Was returning to Haiti where his uncle was president (!) with Juliette, pregnant, and their two children, Simonne and Louise
- 2003 opera by Sharon J. Willis

Demo/Tutorial time!

Data: https://www.mominmalik.com/titanic.csv

Direct link: https://github.com/momin-malik/guides/raw/master/titanic.csv

Lessons

- Machine learning modeling is structured very similarly to statistics (in R, this is deliberate)
- But we only care about something very narrow: predictive performance. Different from carefully building a theoretically-motivated model and interpreting its estimated coefficients
- Test performance is almost certainly worse than training performance—and out-of-sample performance is almost certainly always worse than test performance

(Meta-)commentary

- Machine learning is like training to win a race: what race could we try to run and win from this training?
  - I.e., what would this even generalize to? Predicting death from sinkings of other 19th-century cruise liners? But we already know the outcomes for those too…
  - In contrast, the social statistical/econometric approach lets us draw conclusions about the system from which the data were drawn
- Meta: this dataset exists because of the enormous effort put in by people in the cultural wake of a movie!
- Meta: I would speculate this is used as an exercise because it gives a feeling of power over life and death. Take that as you will…
- I highly recommend Matt Jones' (2018) historical work on Leo Breiman and decision trees/random forests

ICQCM

Effect of dependencies

- One thing I didn't do in the demonstration is examine the effect of dependencies on data splitting.
- The split given by Kaggle/Datacamp puts siblings on opposite sides of the training/test split!
  - One example: 1st class passengers Miss. Alice Elizabeth Fortune (24 years old) and Miss. Mabel Helen Fortune (23) are in the training set; their sister, Miss. Ethel Flora Fortune (28), is in the test set.
- If siblings (perhaps conditioned on sex, age, and passenger class) tended to survive together or perish together, then we have a problem:
  - One way to see: we are sharing information across training and test split, making our accuracy higher in testing than it would otherwise be
  - Another way to see: our "effective sample size" is lower (see: "Galton's problem"), so our estimates of accuracy are inflated

Extra: problems with “explainability”

(Or “interpretability”)

If a model’s “explainability” is not the way in which it captures causality in the world, then what good is it?

ICQCM

Explanations of models seem to be about the world

if male **and** adult **then** *survival probability* 21% (19%–23%)
else if 3rd class **then** *survival probability* 44% (38%–51%)
else if 1st class **then** *survival probability* 96% (92%–99%)
else *survival probability* 88% (82%–94%)

- Decision list: interpretable and explainable
- Letham, Rudin et al. (2015): "For example, we predict that a passenger is less likely to survive than not *because* he or she was in the 3rd class."
- "Because" the model, or "because" the world?

ICQCM

Overview

Going from statistics to machine learning

When to use machine learning

Key concepts

Example for demo: *Titanic*

Demo/ Tutorial

Extra: Prob- lems with explainability

References

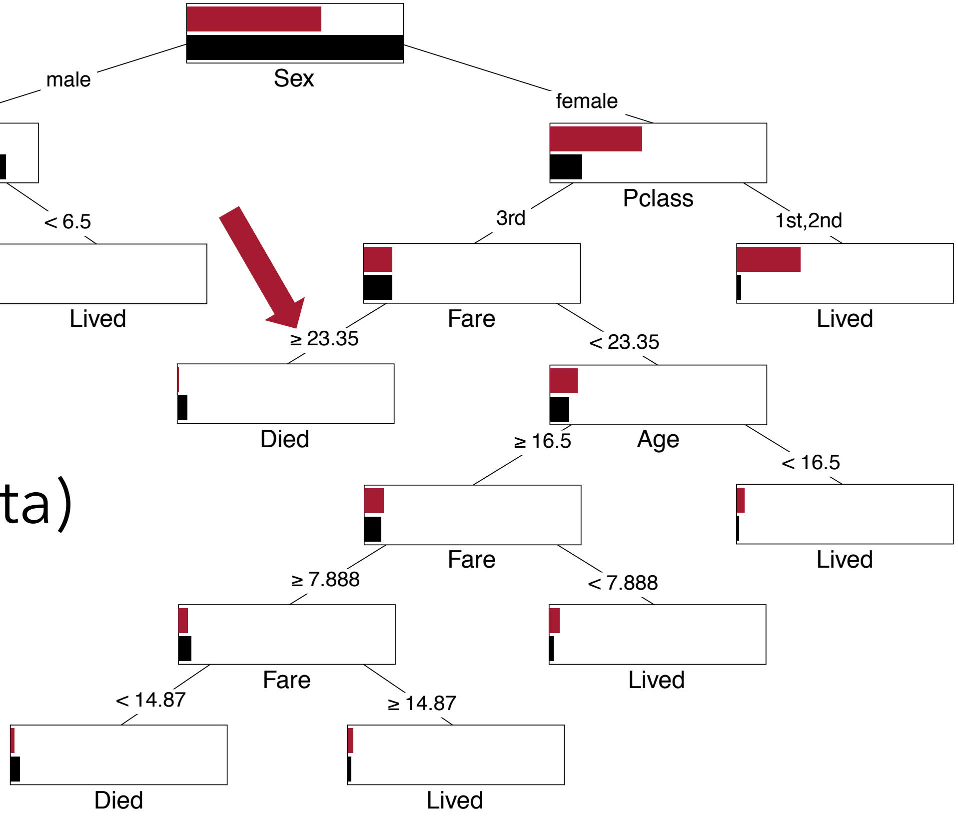
But ML is correlations, not causes

- Finale Doshi-Velez & Been Kim: "one can provide a feasible explanation that fails to correspond to a causal structure, exposing a potential concern."
- Rich Caruana et al. (2015): "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."
- Zachary Lipton: "Another problem is that such an 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."

Wish list for interpretability

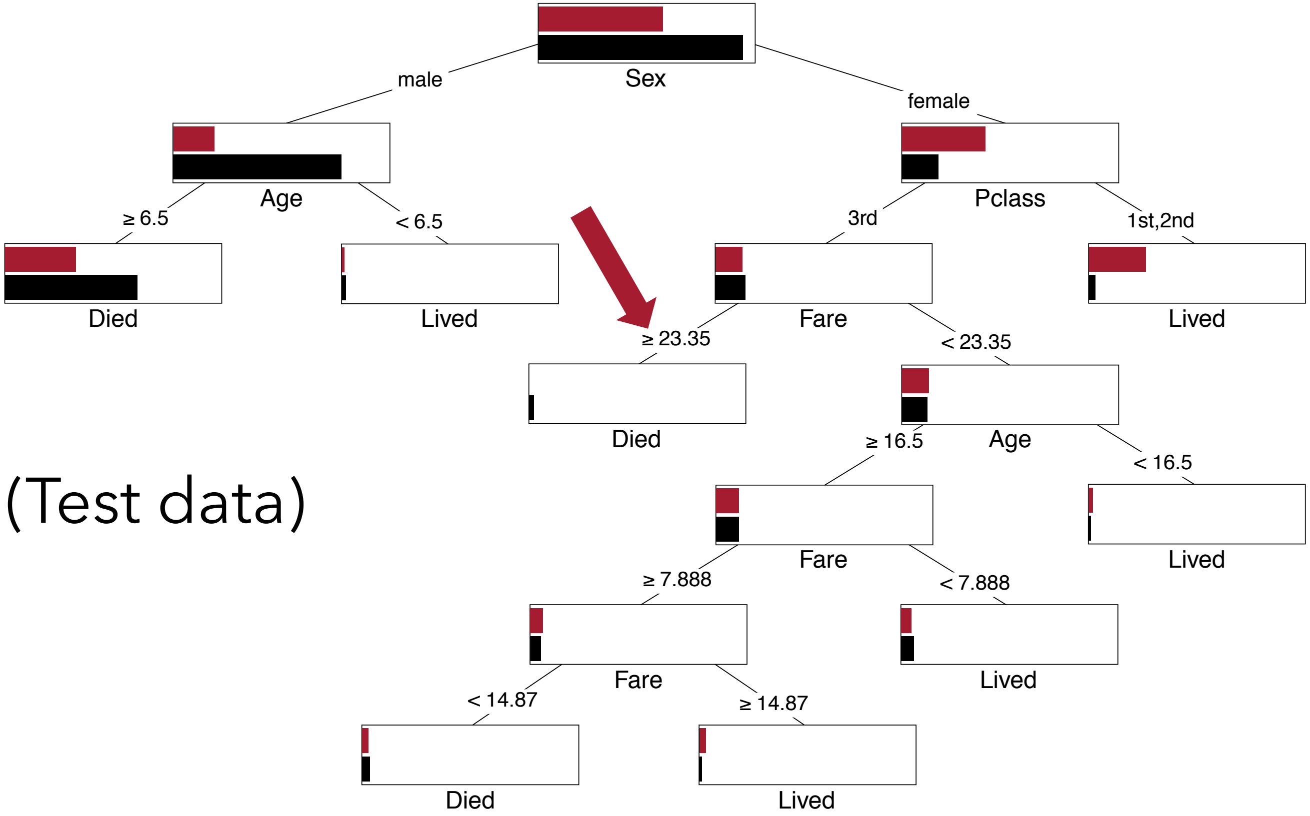
- Face validity as a way to check the model;
- Anticipate where the model might break down (e.g., when it fails face validity);
- Use domain knowledge to 'fine-tune' the model.
- (For my full argument, see https://www.mominmalik.com/ier2019.pdf)

Female, 3rd class less likely to survive *because* of higher fare?



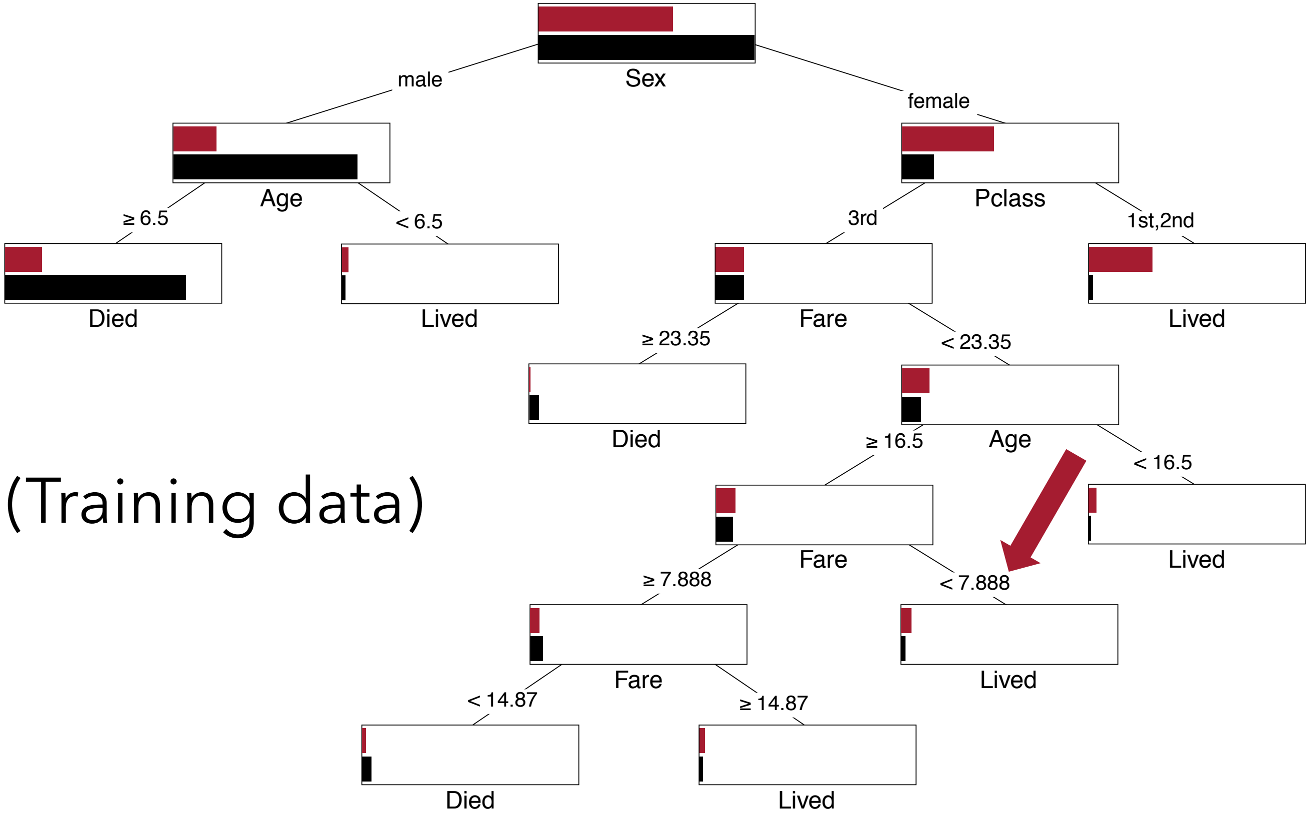
(Training data)

Lacks face validity, but holds on test data



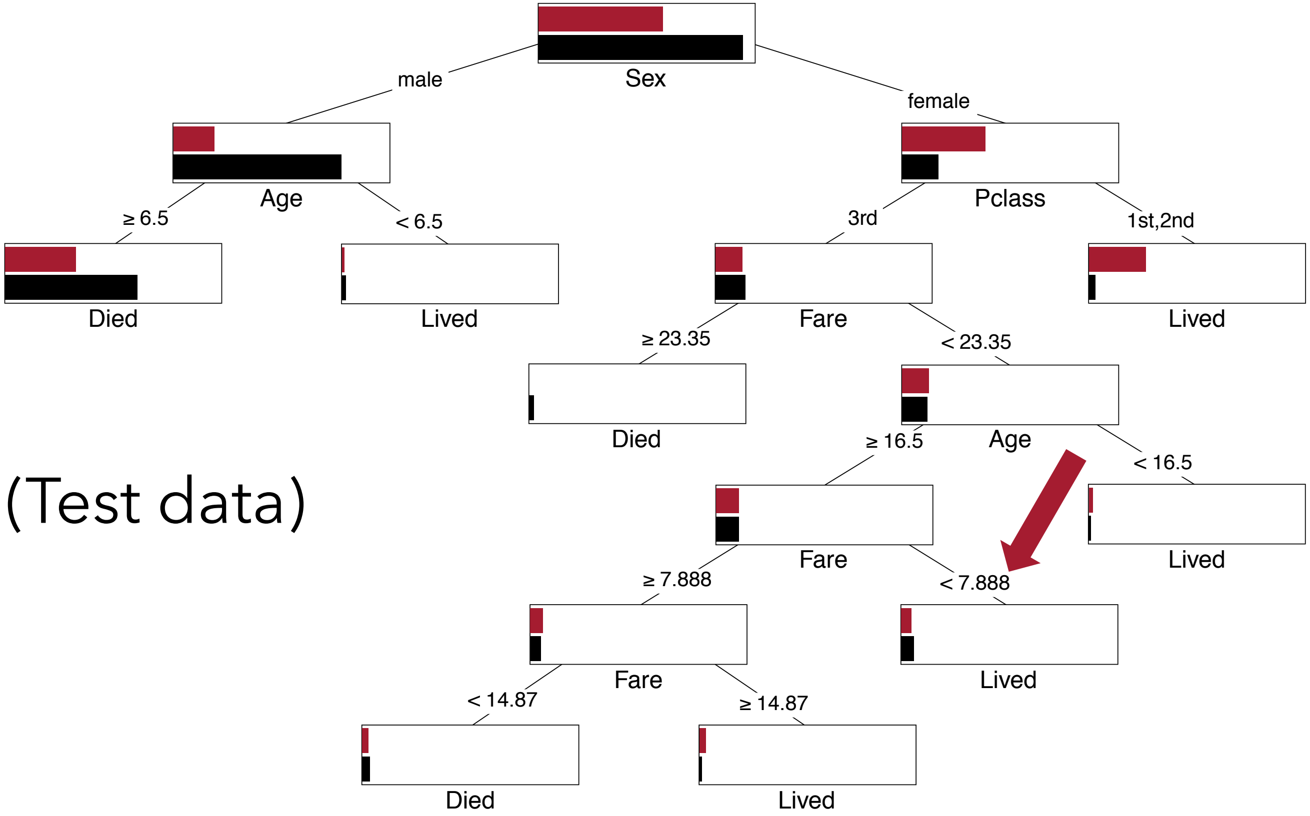
(Test data)

Converse: has face validity, but fails to generalize?



(Training data)

Yes. Interpretability doesn't help anticipate breakdowns

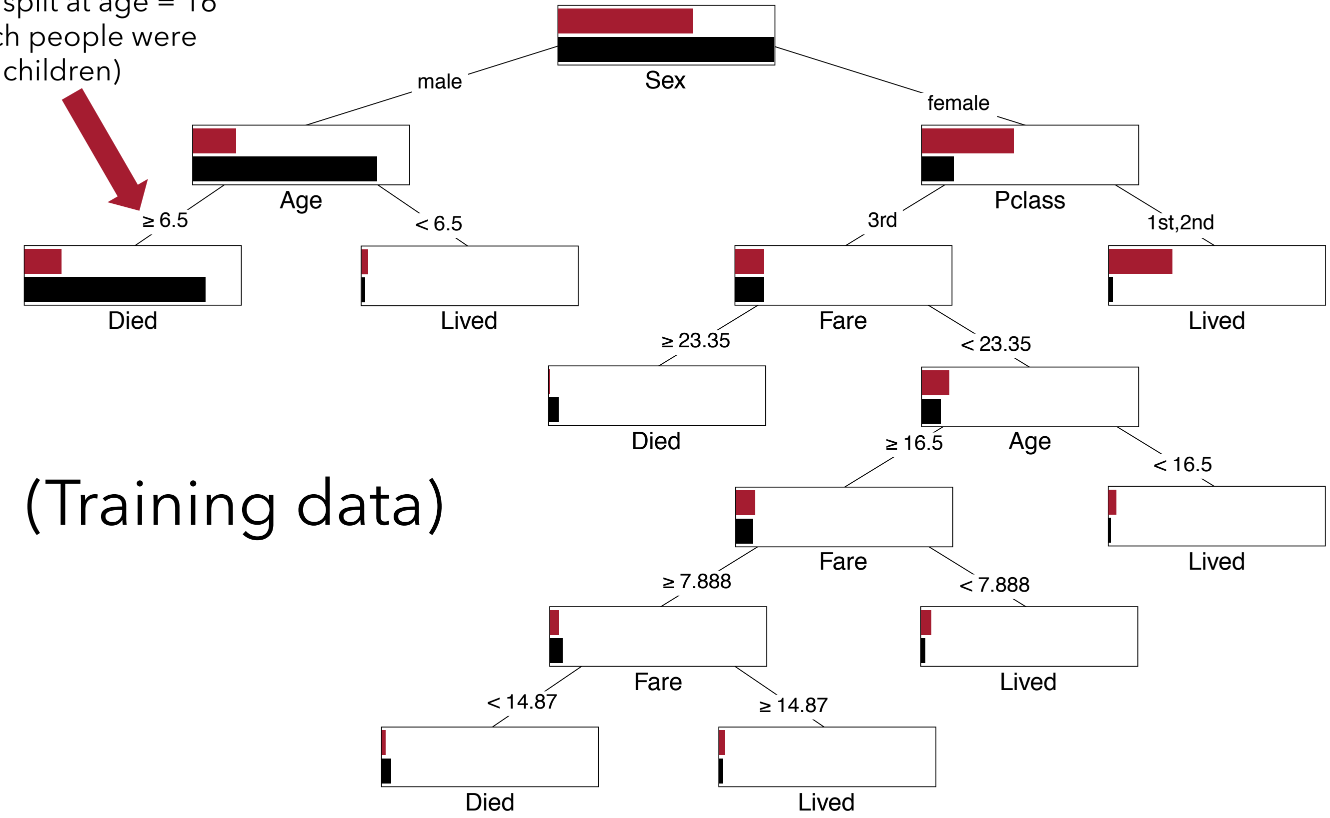


(Test data)

ICQCM

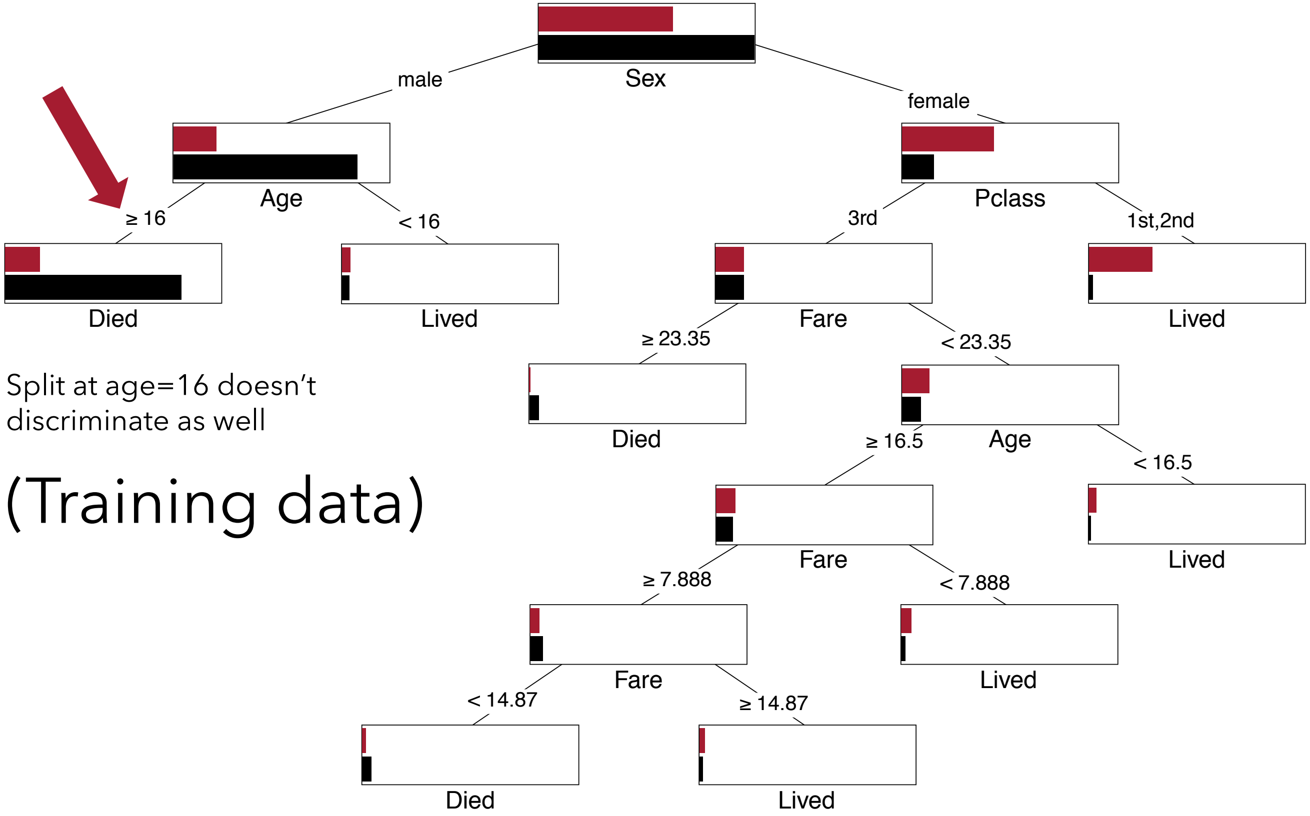
Interpretations to 'fine-tune' model?

Frey et al. say there is a substantive split at age = 16 (below which people were considered children)



(Training data)

Model is already optimally tuned



Split at age=16 doesn't discriminate as well

(Training data)

ICQCM

Overview

Going from statistics to machine learning

When to use machine learning

Key concepts

Example for demo: *Titanic*

Demo/ Tutorial

Extra: Problems with explainability

References

References (1/2)

Breiman, Leo. 2001. "Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author)." *Statistical Science* 16 (3): 199–231. https://doi.org/10.1214/ss/1009213726.

Broussard, Meredith. 2018. *Artificial Unintelligence: How Computers Misunderstand the World*. MIT Press.

Cardoso, Fatima, Laura J. van't Veer, Jan Bogaerts, Leen Slaets, Giuseppe Viale, Suzette Delaloge, Jean-Yves Pierga, Etienne Brain, et al. 2016. "70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer." *The New England Journal of Medicine* 375 (8): 717–729. https://doi.org/10.1056/NEJMoa1602253

Caruana, Rich, Nikos Karampatziakis, and Ainur Yessenalina. 2008. "An Empirical Evaluation of Supervised Learning in High Dimensions." In *Proceedings of the 25th International Conference on Machine Learning (ICML '08)*, 96–103. https://doi.org/10.1145/1390156.1390169

Caruana, Rich, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. "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. https://doi.org/10.1145/2783258.2788613

Doshi-Velez, Finale, and Been Kim. 2017. "Towards A Rigorous Science of Interpretable Machine Learning." https://arxiv.org/abs/1702.08608

Efron, Bradley, and Carl Morris. 1977. "Stein's Paradox in Statistics." *Scientific American* 236 (5): 119–127. https://doi.org/10.1038/scientificamerican0577-119.

Fernández-Delgado, Manuel, Eva Cernadas, Senén Barro, and Dinani Amorim. 2014. "Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?" *Journal of Machine Learning Research* 15 (90): 3133–3181. https://jmlr.org/papers/v15/delgado14a.html

Frey, Bruno S., David A. Savage, and Benno Torgler. 2011. "Behavior under Extreme Conditions: The *Titanic* Disaster." *Journal of Economic Perspectives* 25 (1): 209–222. https://doi.org/10.1257/jep.25.1.209

Frey, Bruno S., David A. Savage, and Benno Torgler. 2010. "Noblesse Oblige? Determinants of Survival in a Life-or-Death Situation." *Journal of Economic Behavior & Organization* 74 (1–2): 1–11. https://doi.org/10.1016/j.jebo.2010.02.005

Garip, Filiz. 2020. "What Failure to Predict Life Outcomes can Teach Us." *PNAS* 117 (15): 8234–8235. https://doi.org/10.1073/pnas.2003390117

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

Gayo-Avello, Daniel. 2012. "No, You Cannot Predict Elections with Twitter." *IEEE Internet Computing* 16 (6): 91–94. https://doi.org/10. 1109/MIC.2012.137

Hu, Lily. 2019a. "Disparate Causes, pt. I". *Phenomenal World*, October 11. https://phenomenalworld.org/analysis/disparate-causes-i

Hu, Lily. 2019b. "Disparate Causes, pt. II". *Phenomenal World*, October 17. https://phenomenalworld.org/analysis/disparate-causes-pt-ii

Hu, Lily. 2020. "Direct Effects." *Phenomenal World*, September 25. https://phenomenalworld.org/analysis/direct-effects

Jones, Matthew L. 2018. "How We Became Instrumentalists (Again): Data Positivism since World War II." *Historical Studies in the Natural Sciences* 48 (5): 673–684. https://doi.org/10.1525/hsns.2018.48.5.673.

References (2/2)

Junqué de Fortuny, Enric, David Martens, and Foster Provost. 2013. "Predictive Modeling With Big Data: *Is Bigger Really Better*?" *Big Data* 1 (4): 215-226. https://doi.org/10.1089/big.2013.0037

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. "Human Decisions and Machine Predictions." *The Quarterly Journal of Economics* 133 (1): 237–293. https://doi.org/10.1093/qje/qjx032

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. "Prediction Policy Problems." *American Economic Review* 105 (5): 491–95. https://doi.org/10.1257/aer.p20151023

Letham, Benjamin, Cynthia Rudin, Tyler H. McCormick, and David Madigan. 2015. "Interpretable Classifiers Using Rules and Bayesian Analysis: Building a Better Stroke Prediction Model." *The Annals of Applied Statistics* 9 (3): 1350–1371. https://doi.org/10.1214/15-AOAS848

Lipton, Zachary C. 2015. "The Myth of Model Interpretability." *KDnuggets* 15 (13) https://www.kdnuggets.com/2015/04/model-interpretability-neural-networks-deep-learning.html

Lipton, Zachary C. 2018. "The Mythos of Model Interpretability." *ACM Queue* 16 (3): 31–57. https://doi.org/10.1145/32363 86.3241340

Malik, Momin M. 2018. *Bias and Beyond in Digital Trace Data*. PhD diss., Carnegie Mellon University. SCS Technical Report Collection CMU-ISR-18-105. http://ra.adm.cs.cmu.edu/anon/isr2018/abstracts/18-105.html

Malik, Momin M. 2020. "A Hierarchy of Limitations in Machine Learning." https://www.arxiv.org/abs/2002.05193

Malik, Momin M., and Jürgen Pfeffer. 2016. "Identifying Platform Effects in Social Media Data." In *Proceedings of the Tenth International AAAI Conference on Web and Social Media* (ICWSM-16), 241–249. Updated version at http://mominmalik.com/malik_chapter2.pdf.

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

Mullainathan, Sendhil, and Jann Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives* 31 (2): 87–106. https://doi.org/10.1257/jep.31.2.87.

Richardson, Eugene T. 2020. *Epidemic Illusions: On the Coloniality of Global Public Health*. MIT Press.

Rudin, Cynthia. 2019. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." *Nature Machine Intelligence* 1: 206–2015. https://doi.org/10.1038/s42256-019-0048-x

Salganik, Matthew J., Ian Lundberg, Alexander T. Kindel, Caitlin E. Ahearn, Khaled Al-Ghoneim, Abdullah Almaatouq, Drew M. Altschul, Jennie E. Brand, Nicole Bohme Carnegie, Ryan James Compton, …, Barbara E. Engelhardt, Moritz Hardt, Dean Knox, Karen Levy, Arvind Narayanan, Brandon M. Stewart, Duncan J. Watts, and Sara McLanahan. 2020. "Measuring the Predictability of Life Outcomes with a Scientific Mass Collaboration." *PNAS* 117 (15): 8398-8403. https://doi.org/10.1073/pnas.1915006117

Shmueli, Galit. 2010. "To Explain or to Predict?" *Statistical Science* 25 (3): 289–310. https://doi.org/10.1214/10-STS330

van't Veer, Laura J., Hongyue Dai, Marc J. van de Vijver, Yudong D. He, Augustinus A. M. Hart, Mao Mao, Hans L. Peterse, Karin van der Kooy, Matthew J. Marton, Anke T. Witteveen, George J. Schreiber, Ron M. Kerkhoven, Chris Roberts, Peter S. Linsley, René Bernards, and Stephen H. Friend. 2002. "Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer." *Nature* 415 (6871): 530–536. https://doi.org/10.1038/415530a