ICQCM

CRITICAL DATA SCIENCE
FOR A DIVERSE WORLD

Computational Approaches III: Applications

Momin M. Malik

Thursday, July 22nd, 2–3:30 PM EDT

ICQCM 2021 Seminar Series

Manzel Bowman,
Celestial Globe (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?

ICQCM

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

ICQCM

Going from statistics to machine learning

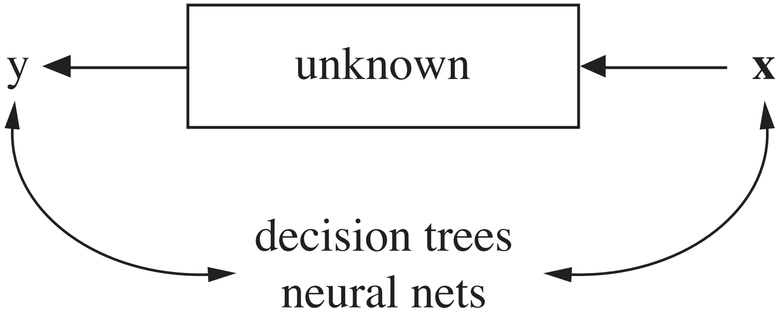
Machine learning is the instrumental use of correlations

Prediction and explanation are different goals and can be in conflict

Defining machine learning

Statistics:

Machine learning:



Breiman 2001. See also Jones 2018.

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.

ICQCM

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

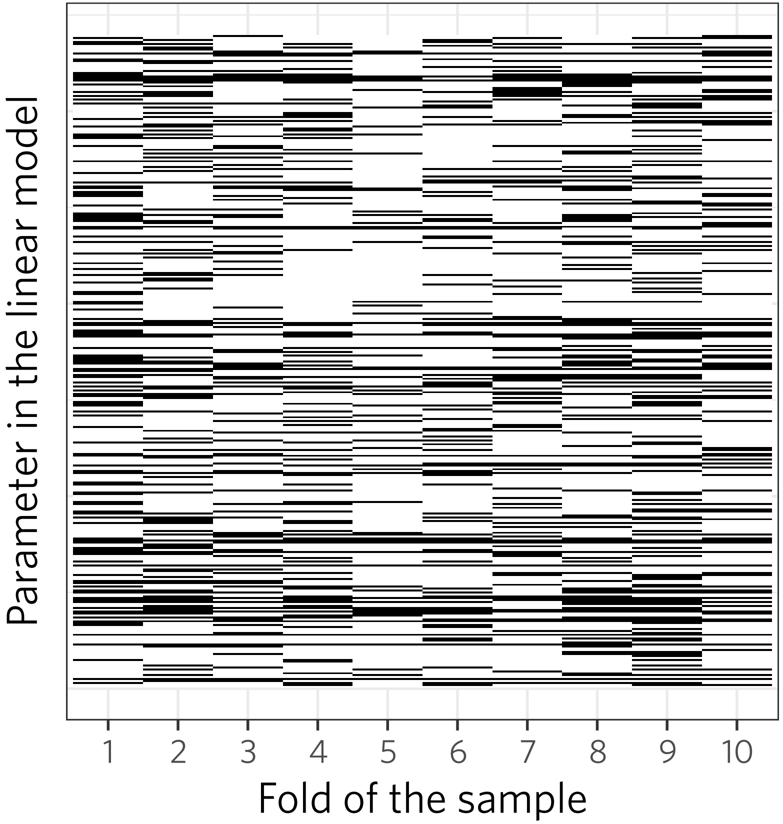
The surprising part

- *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
- We can use correlations to "predict" without knowing causality, or "explaining"!

* Or other relevant metric of success

ICQCM

But: can't *intervene* based on correlations



- Very different sets of correlations can "predict" equally well (Mullainathan & Spiess, 2017)
- But they would 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

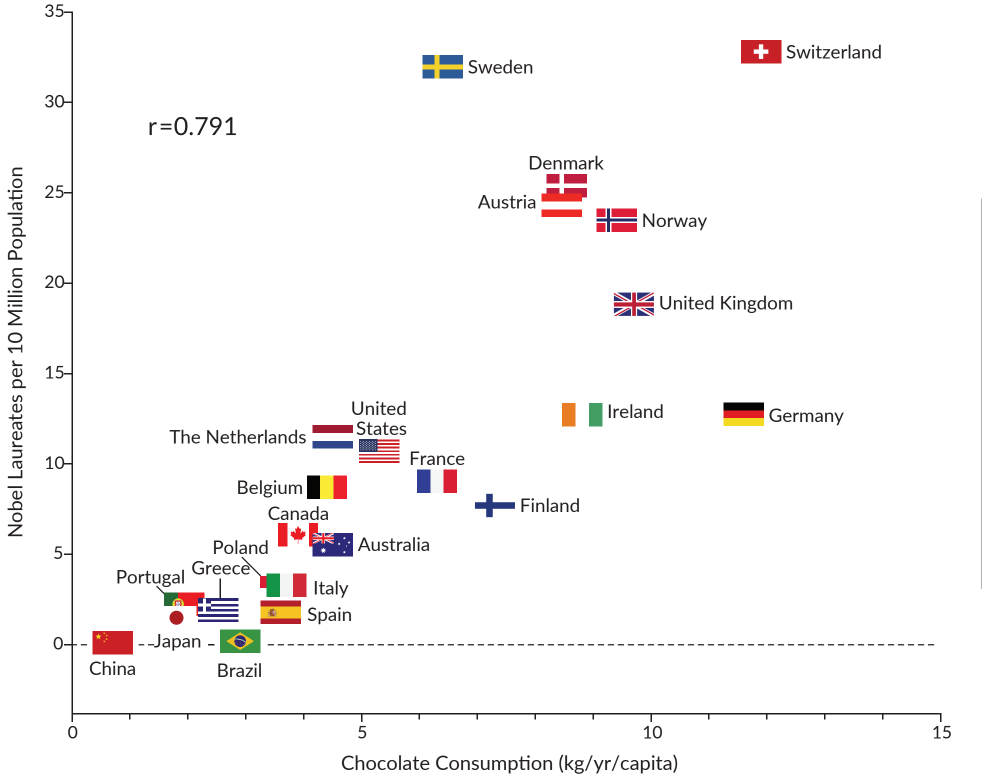
Caroline Chan, “Everybody Dance Now: Motion Retargeting Video Subjects.” https://youtu.be/PCBTZh41Ris

ICQCM

Going from stats to ML

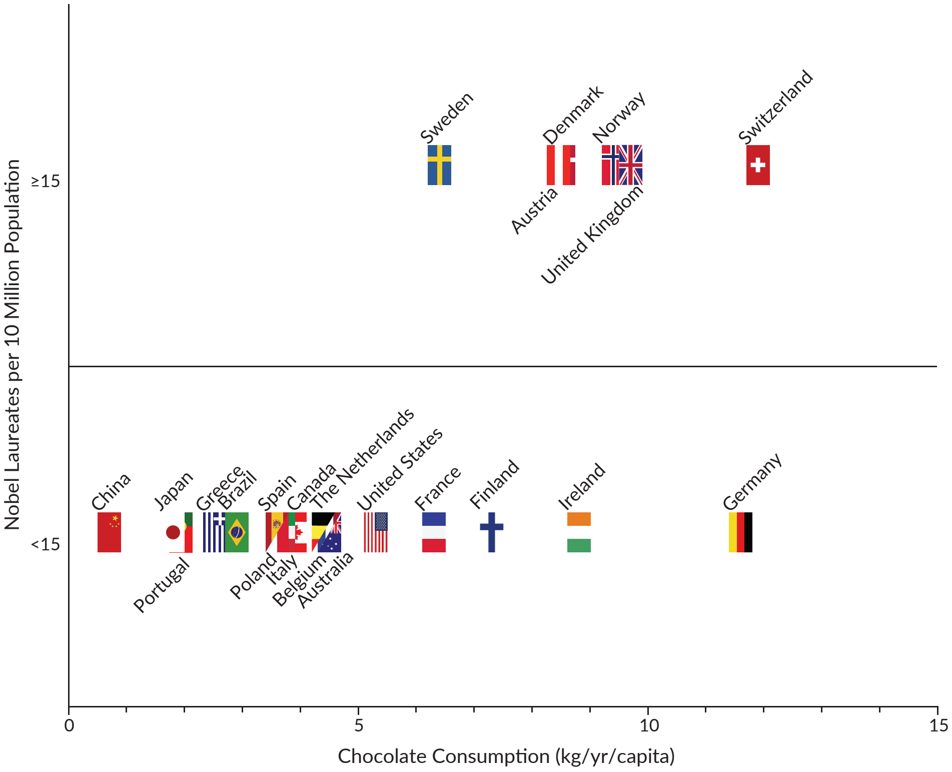
- ML overlaps lots with *nonparametric statistics*, which (for example) gets models by locally smoothing input data rather than getting global fits. But ML also uses parametric models, and lots of Bayesian models (although in decidedly non-Bayesian ways)
- ML: no statistical inference, and so doesn't need to calculate standard errors. Opens up modeling possibilities
- 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)
  - Note: deep learning only works for audio, images, and (sometimes) time series. For general modeling, random forests are usually best (Caruana et al. 2008; Fernández-Delagado et al., 2014)
- Caution: statistical significance is not the same as feature importance!

Regression: Continuous relationship

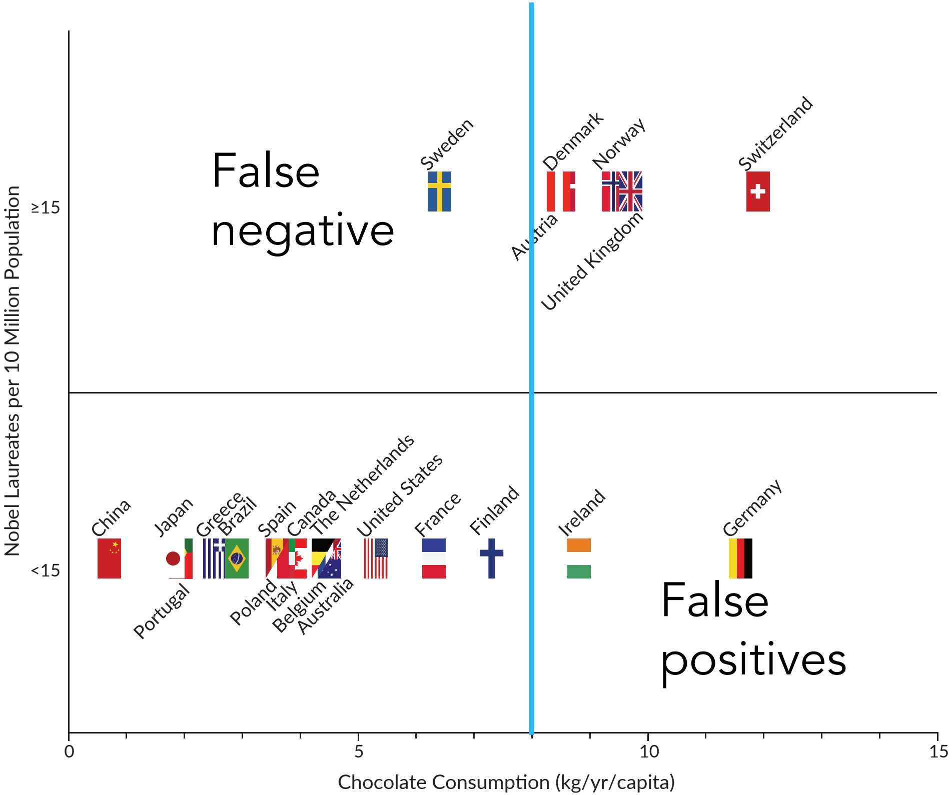


ICQCM

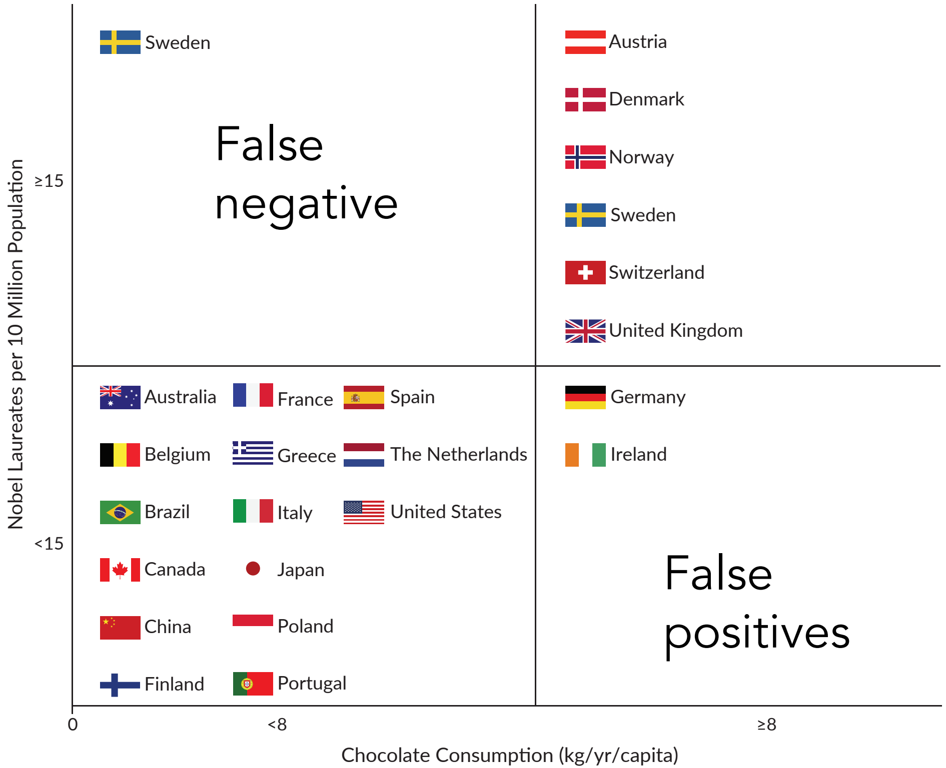
Classification: Discrete relationship



Fit the decision boundary



The prediction: the majority class



When to use machine learning

Key components of a good use case

Example of a “responsible” use case

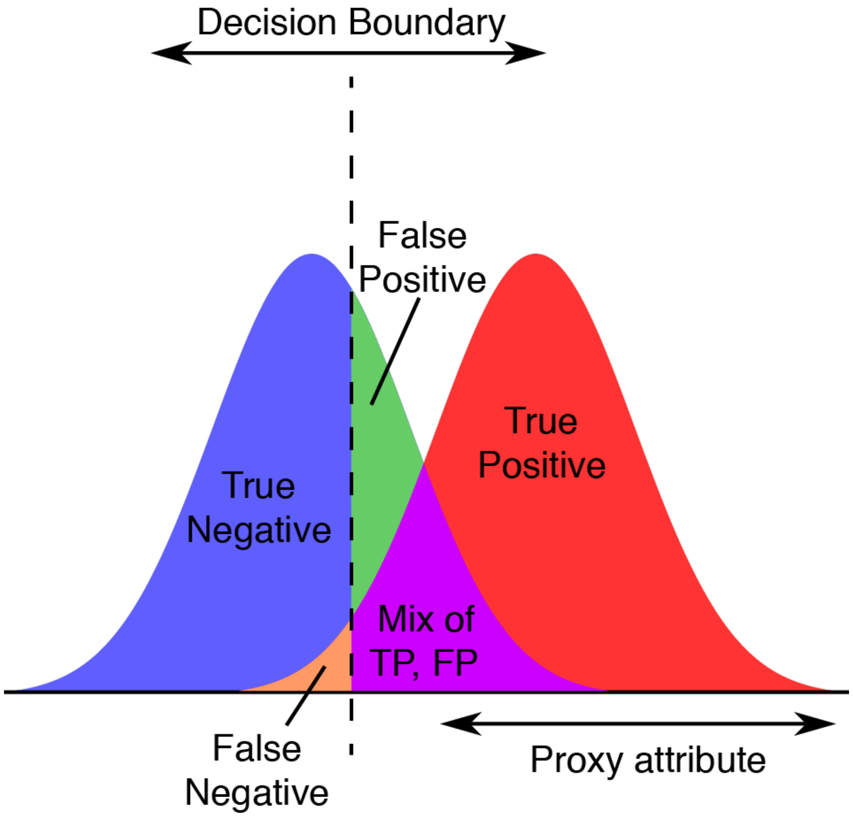
ICQCM

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)

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
- Which is better? Experimentally test! (Cardoso et al., 2016)

Real-world testing

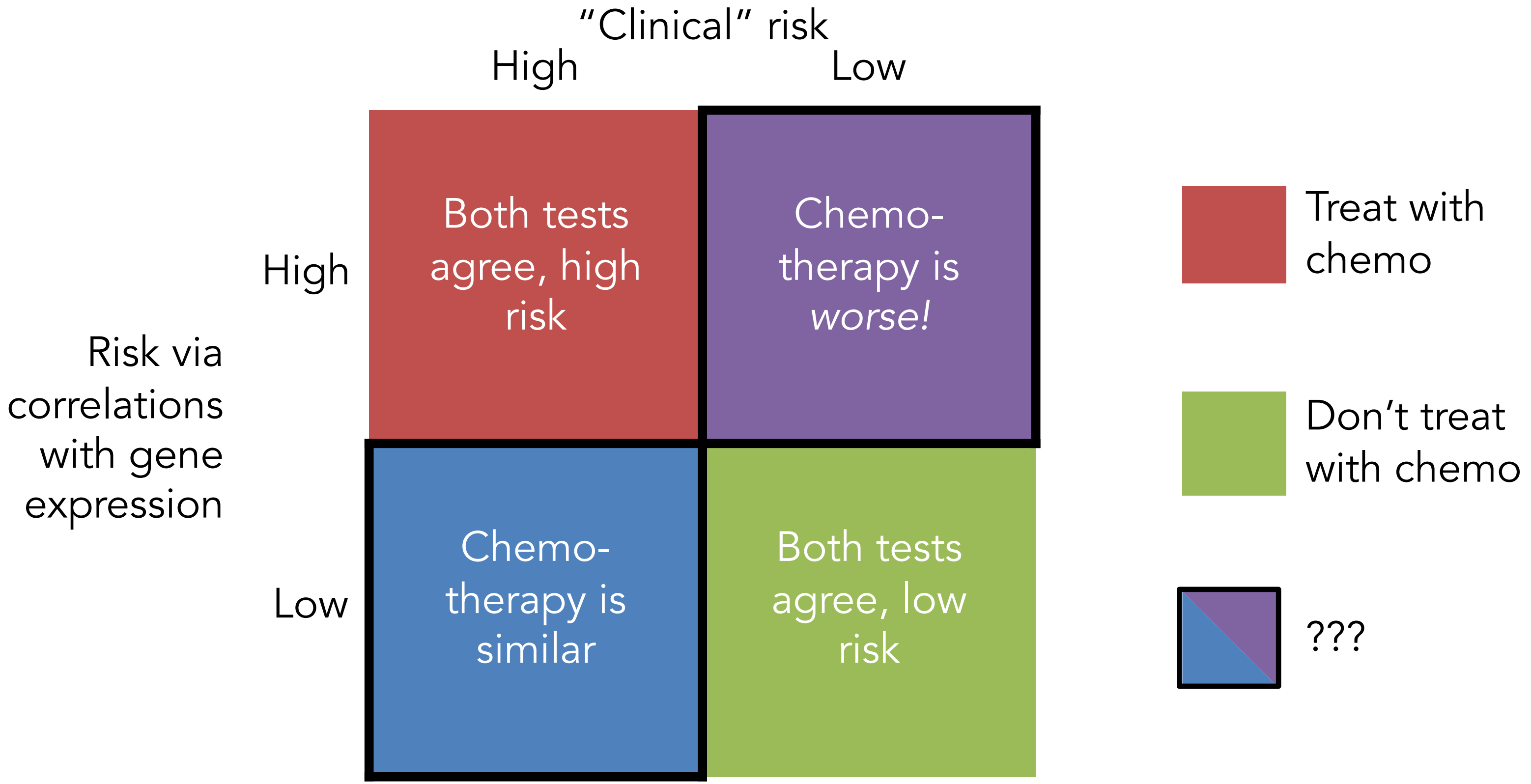
Risk via correlations with gene expression	"Clinical" risk: High	"Clinical" risk: Low
High	Both tests agree, high risk	Model says treat, doctor says don't
Low	Doctor says treat, model says don't	Both tests agree, low risk

Treat with chemo

Don't treat with chemo

???

Real-world testing



ICQCM

Real-world testing

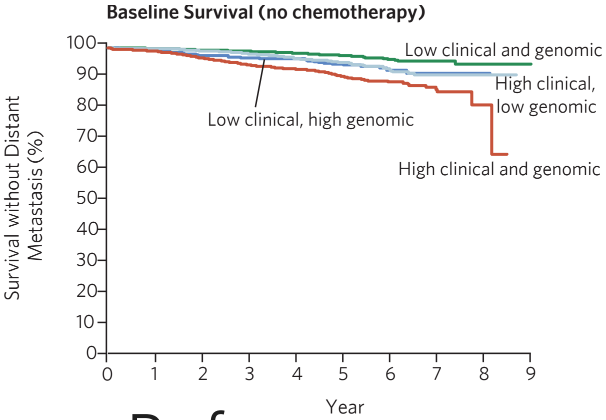
		“Clinical” risk: High	“Clinical” risk: Low
Risk via correlations with gene expression	High	Both tests agree, high risk	Chemo-therapy is *worse!*
	Low	Chemo-therapy is similar	Both tests agree, low risk

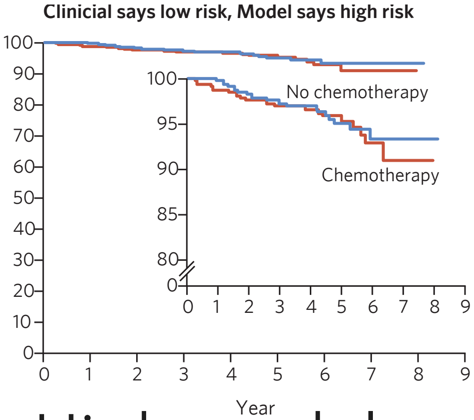
Red: Treat with chemo

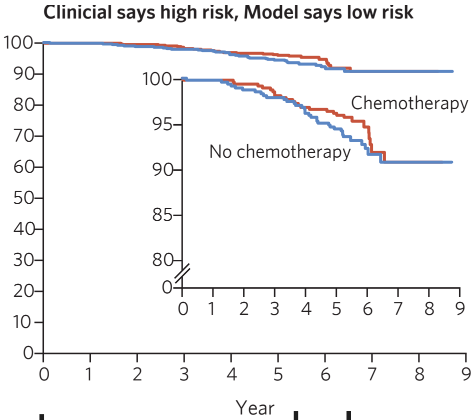
Green: Don't treat with chemo

(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

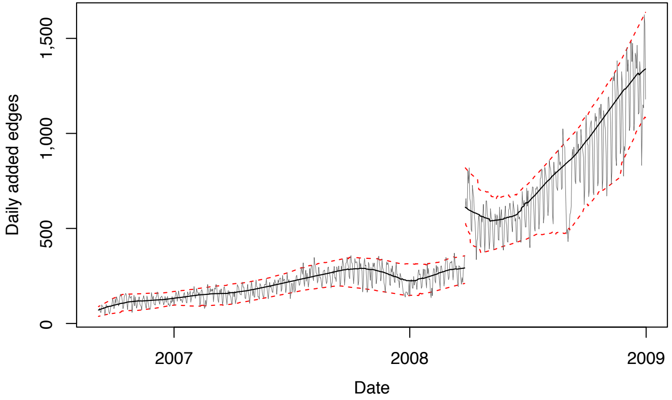
Cardoso et al., 2016, *NEJM*

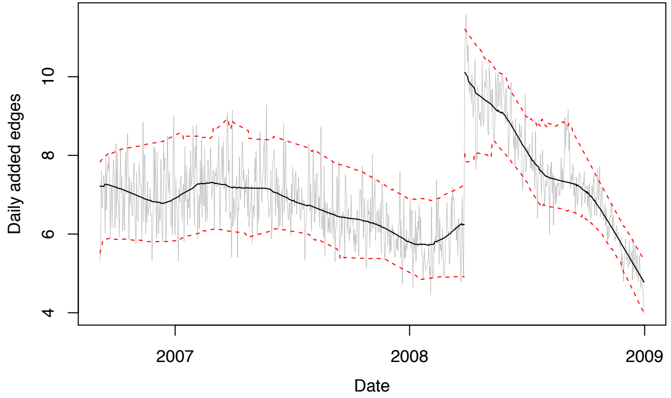
Key points

- 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, so it's only meaningful if we actually run the race
- (More on this later): *machine learning performance claims are always preliminary until we do real-world testing*

What good is machine learning for social science?

- 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



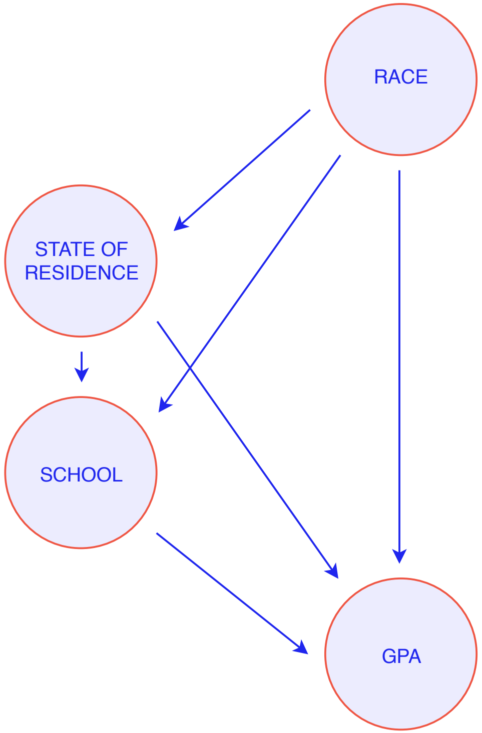


What good is machine learning in social science?

- For *scaling up* human labels to a larger dataset
  - 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 that take into account the uncertainty of the imperfect model, and lack of perfect agreement among coders

What good is machine learning in social science?

- Graphical models are part of machine learning, and can express complex causal structure (are equivalent to Structural Equation Models)
- Note: can <u>express</u> causal structure, not find/discover it
  - And ultimately "expresses" causality in a very limited way (Hu 2019a, 2019b, 2020; Richardson 2021)
- But within machine learning, graphical models are seldom used for causal modeling



ICQCM

Key concepts

Model fit

Overfitting

Data splitting

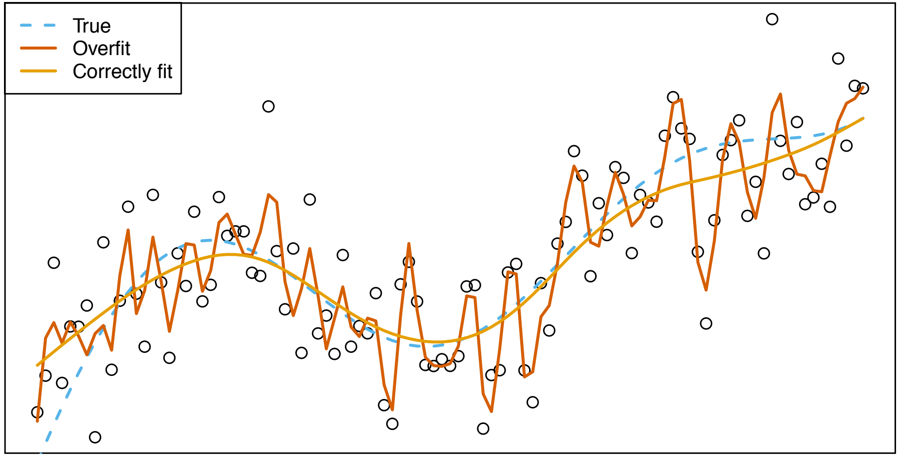
Accuracy paradox

Confusion matrix

Model "fit"

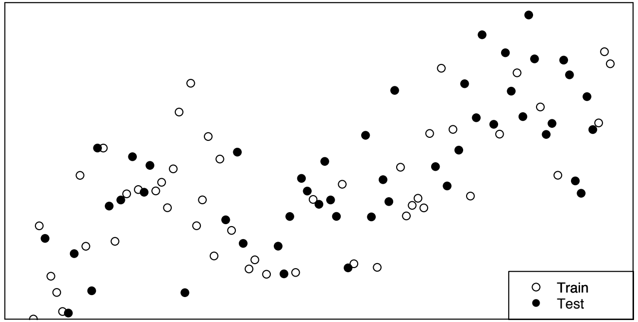
- All machine learning and statistics models take in data, process them via some assumptions, and then give out something: relationships, and/or likely future values.
- The processing is called "fitting", and the output is called a "fit." Machine learning uses "learning" or "training," but it's the same.

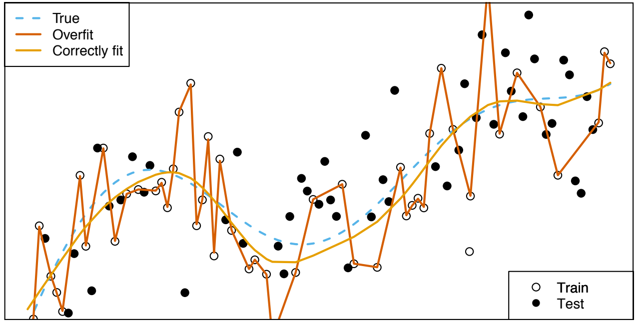
Overfitting: fit to noise



- 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

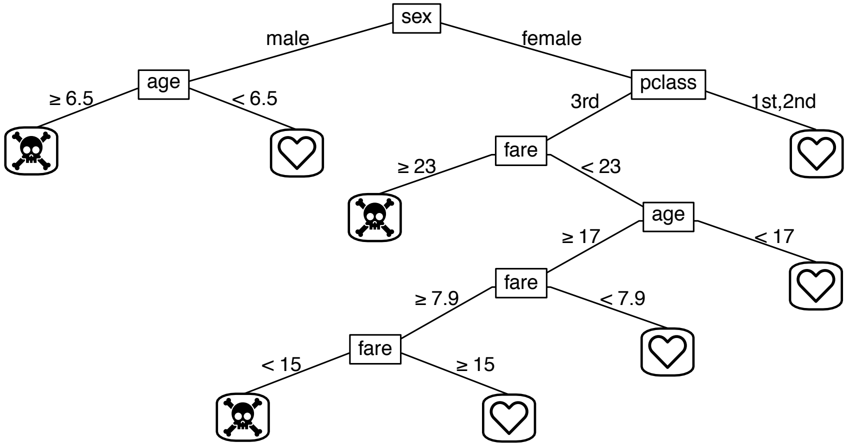


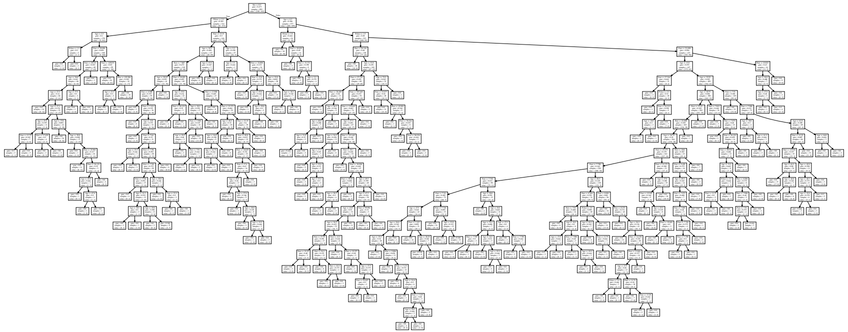


- Idea: if we split data into two parts, the signal should be the same but the noise would be different
- *Cross validation:* Fitting the model on one part of the data, and "testing" on the other

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

(Overfitting in a classification tree)





ICQCM

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

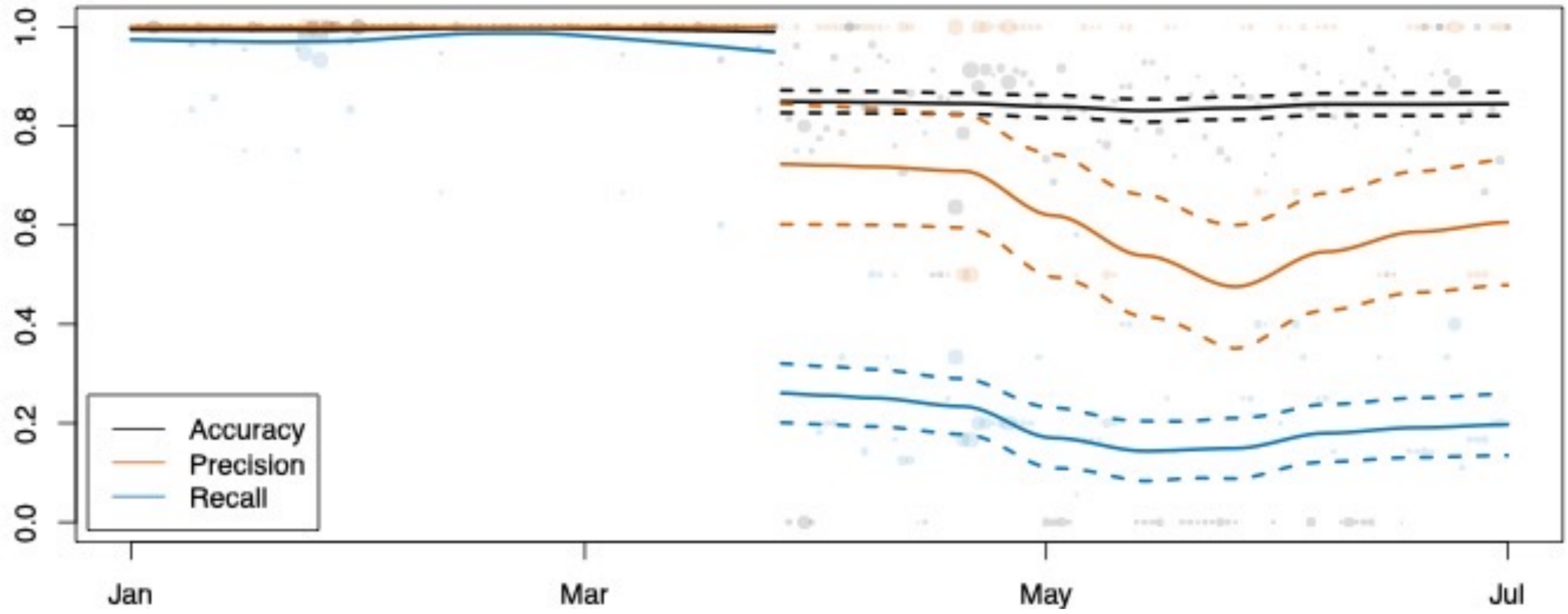
True label

	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		

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
- They have very different theoretical properties
- For both: *want to split in a way that respects dependencies*
- E.g., random splits of a time series means you use future values to "predict" past ones
  - "Time-traveling"
  - Not a realistic test of out-of-sample performance

Doing data splitting correctly



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

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: they use an instrumental variable (judge leniency) to try and get unbiased estimates of generalizability error

Example for demo: *Titanic*

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 THE TWENTIETH CENTURY NEWSPAPER GREATER NEW YORK.

BUSINESS PROPERTY

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

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.

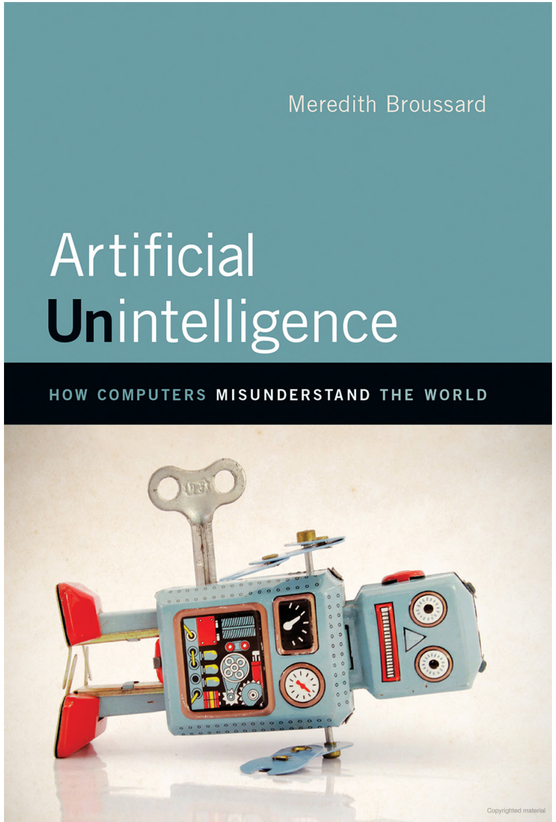
J. J. ASTOR LOST ON TITANIC

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.

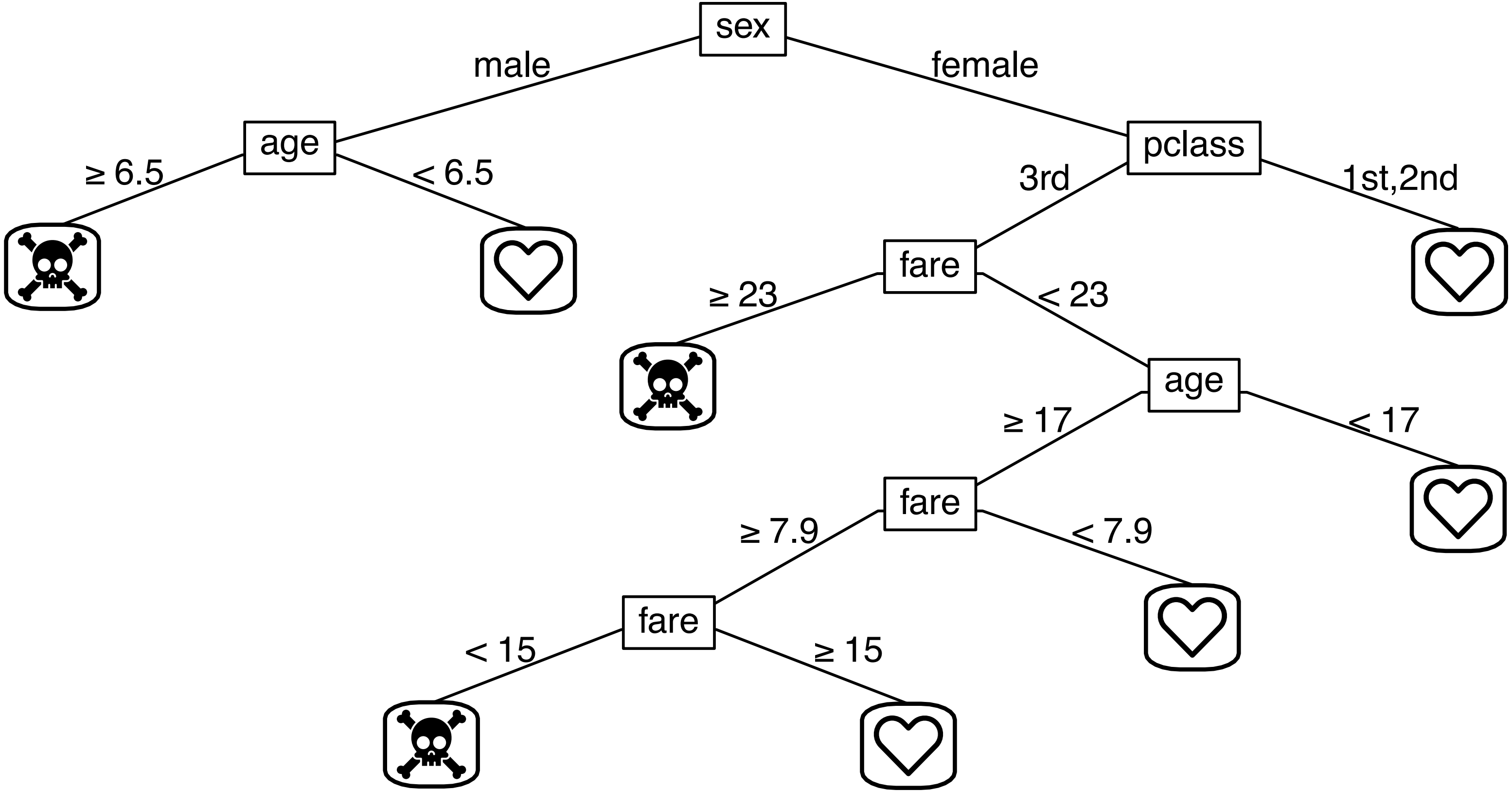
The Wireless Operator at Cape Race, Newfoundland, Flashes: "Eighteen Hundred Lives Have Been Lost in the Wreck of the Titanic."

Broussard's Commentary



- Captain: "Put the women and children in and lower away."
- First Officer: women and children *first*
- Second Officer: 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."

Fit a “decision tree” for survival



Social science baseline for comparison

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

ICQCM

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 time!

Data:
https://www.mominmalik.com/titanic.csv
https://github.com/momin-malik/guides/raw/master/titanic.csv

ICQCM

Lessons

- Machine learning is structured very similarly to statistics
- But we only care about something very narrow, and something different from what we care about when doing statistics (or narrative)
  - What would this even generalize to? Predicting death from sinkings of other 19th-century cruise liners?
  - Meta-commentary: this dataset exists because of effort put in by people in the cultural wake of a movie!
  - I would speculate this is used as an exercise because it gives a feeling of power over life and death
- Data splitting shouldn't be done lightly
  - Siblings fall along different sides of test splits, and they may have survived together
- Test performance is almost certainly than training performance—and out-of-sample performance is almost certainly always worse than test performance

Extra: problems with “explainability”

Or “interpretability”

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
- Lethan, Rudin et al.: "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?

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.: "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

ICQCM

Overview

Going from statistics to machine learning

When to use machine learning

Key concepts

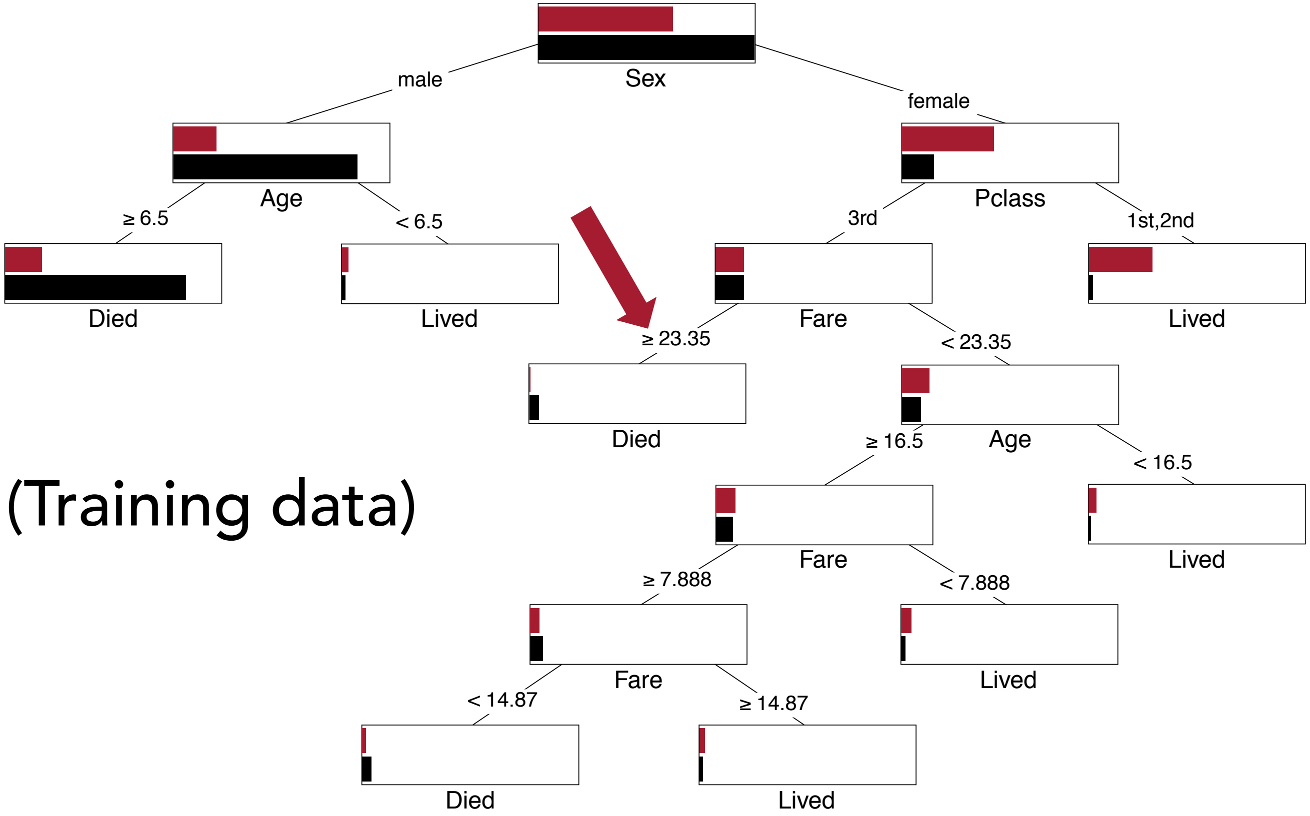
Example for demo: *Titanic*

Demo

Extra: Problems with explainability

References

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



(Training data)

Overview

Going from statistics to machine learning

When to use machine learning

Key concepts

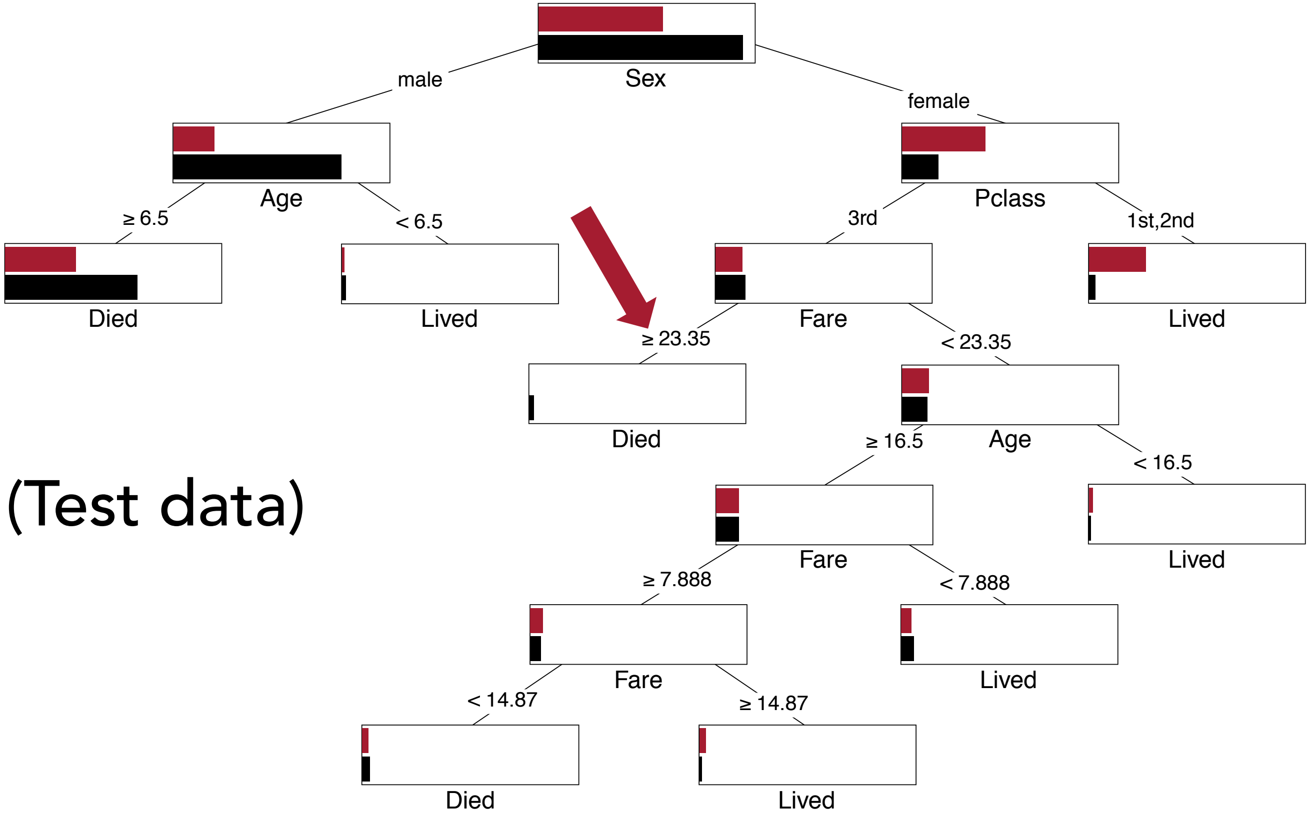
Example for demo: *Titanic*

Demo

Extra: Problems with explainability

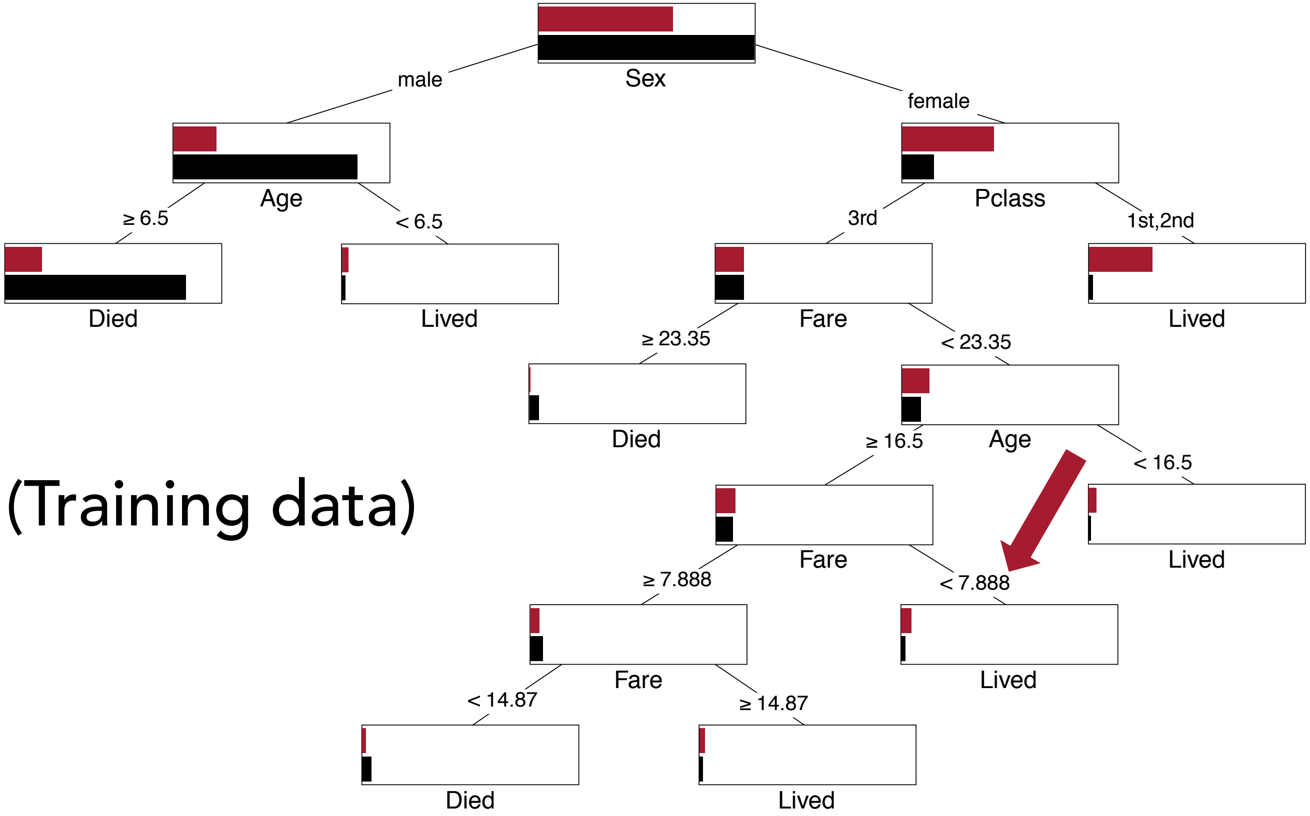
References

Lacks face validity, but holds on test data



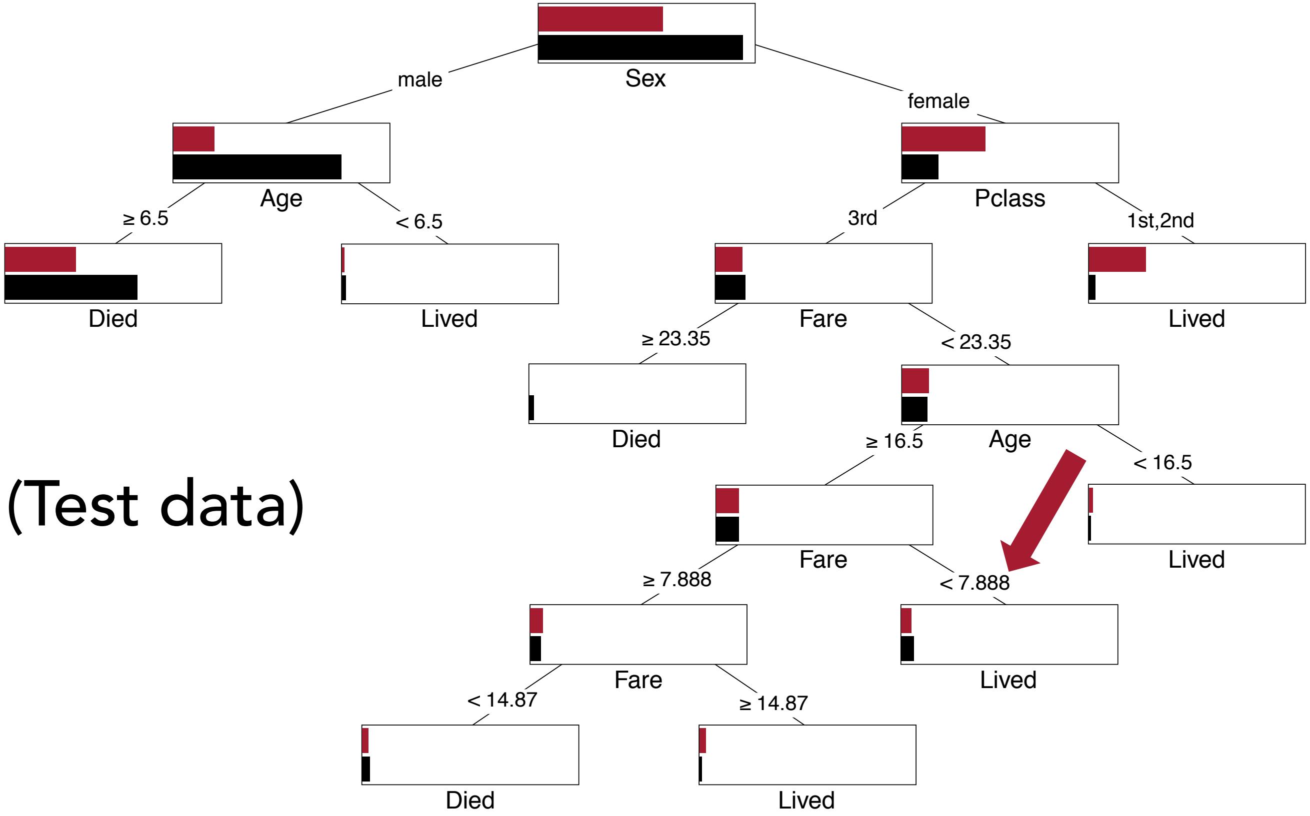
(Test data)

Converse: has face validity, but fails to generalize?



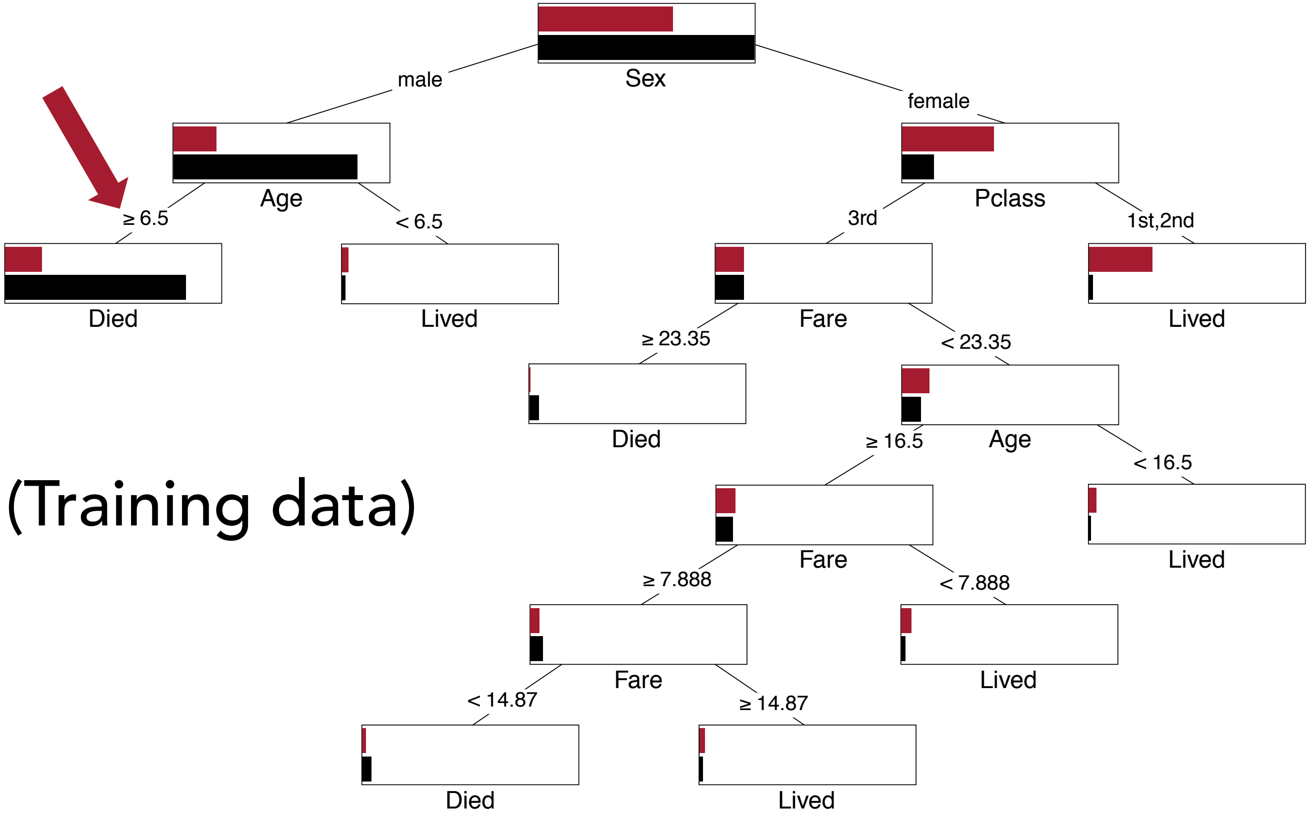
(Training data)

Yes. Interpretability doesn't help anticipate breakdowns



(Test data)

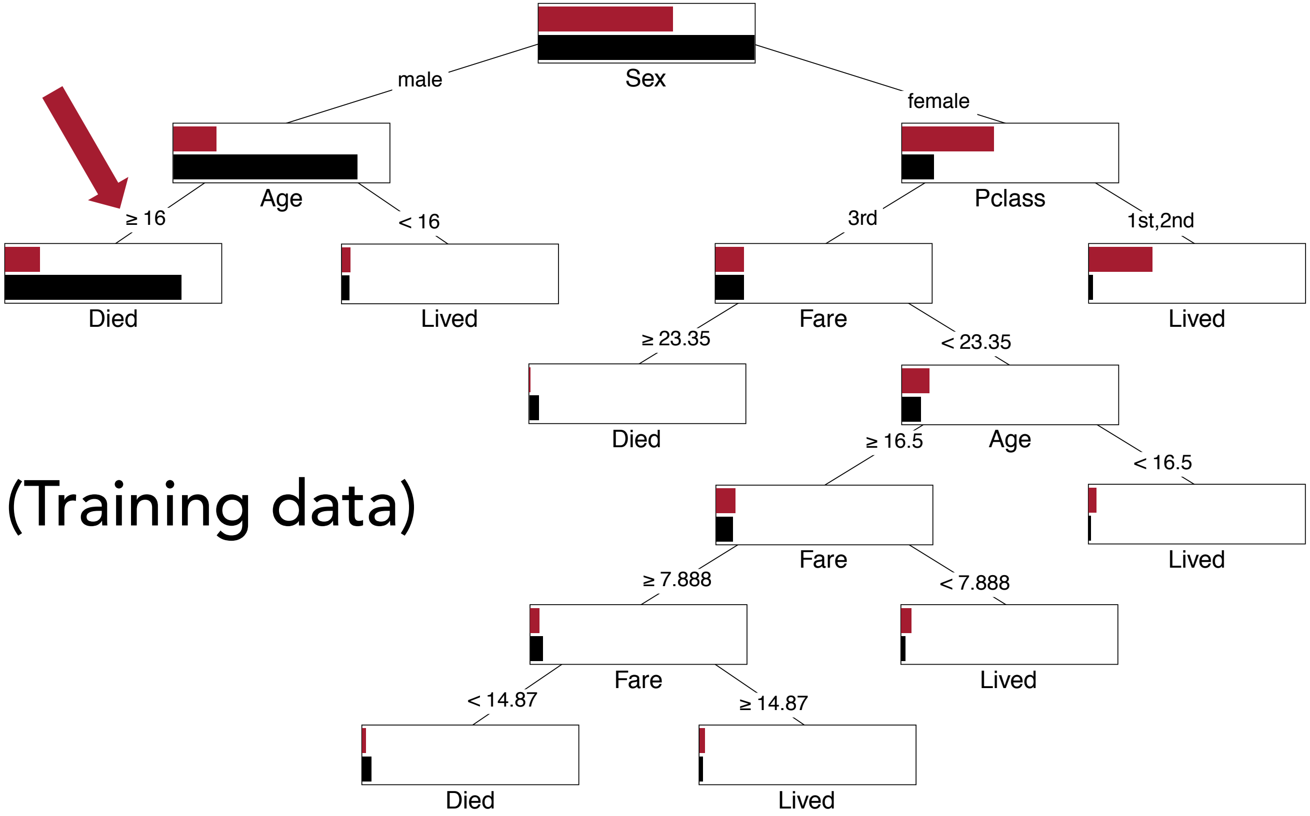
Interpretations to 'fine-tune' model?



(Training data)

ICQCM

Model is already optimally tuned



(Training data)