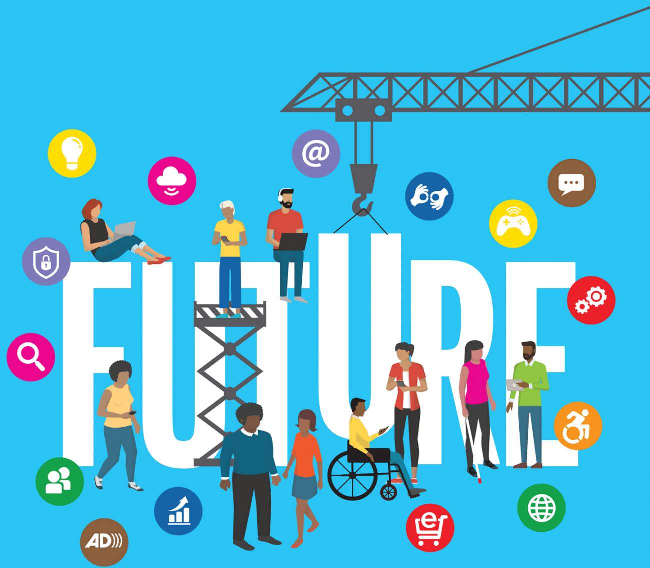
A Critical Introduction to Machine Learning

Momin M. Malik
Data Science Postdoctoral Fellow,
Berkman Klein Center for Internet & Society at Harvard

Slides: https://www.mominmalik.com/tapia2019.pdf

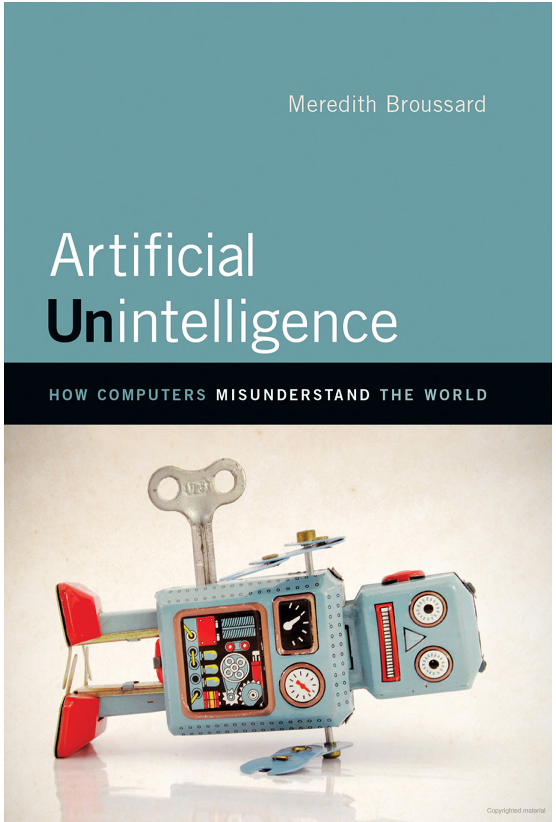
2019 ACM RICHARD TAPIA
CELEBRATION OF DIVERSITY IN COMPUTING CONFERENCE
THURSDAY, SEPTEMBER 19 | MARRIOTT 12



Basis: Meredith Broussard’s book

- *Artificial Unintelligence: How Computers Misunderstand the World* (MIT Press, 2018)
- Chapter 7 is the single best introduction to machine learning!
- Based on a datacamp tutorial, with commentary: I expand on this
- (One subtle but important mistake: see https://www.mominmalik.com/broussard)

"So, it's not real AI?"



- "So, it's not real AI?" he asked.
- "Oh, it's real," I said. "And it's spectacular. But you know, don't you, that there's no simulated person inside the machine? Nothing like that exists. It's computationally impossible."
- His face fell. "I thought that's what AI meant," he said. "I heard about IBM Watson, and the computer that beat the champion at Go, and self-driving cars. I thought they invented real AI."

Preliminaries

Install R + Rstudio

Introductions

Learning goals

Machine learning? Critical?

Outline

Prepare to follow along later!

- If you don't have it already, download and install R (search: "install R")
- Also install RStudio (search: "install RStudio")
- Installation should, at most, take about as long as the introduction

About me

- DEPARTMENT OF THE HISTORY OF SCIENCE, HARVARD UNIVERSITY
- VE RI TAS — Berkman: The Berkman Center for Internet & Society at Harvard University
- OXFORD INTERNET INSTITUTE — UNIVERSITY OF OXFORD
- Carnegie Mellon University School of Computer Science
- BERKMAN KLEIN CENTER FOR INTERNET & SOCIETY AT HARVARD UNIVERSITY

What about you?

- Undergrad student?
- Grad student?
- Academia?
- Industry?
- Public sector?

Learning goals by background

- No background in programming or statistics:
  - See what doing machine learning looks like in practice
  - Identify appropriateness of machine learning
- Linear regression (Excel, SPSS, Stata, Java):
  - Use cross-validation
- Logistic regression, and/or Python or R:
  - Build, evaluate, and critique a basic machine learning model

ML = *Finding correlations for prediction*

- Textbook definitions are aspirational.
- In practice, machine learning is about *finding correlations that we can use for prediction*
- Spurious correlations are fine, so long as they are robust
- Machine learning is not well suited for *understanding* (although people assume it is)

Critical = "See your glasses"

- Critical: To be able to see the glasses with which you see the world (Agre, 2000)
- A critical *theory*: identifies a *false consciousness*, and seeks to expose it to spur transformative action (Fay, 1987)
  - I think "Data positivism" (Jones, 2019) is the false consciousness of machine learning

Outline

1. Machine learning is correlations
2. When to use machine learning
3. Background needed
4. Key concepts
5. Live, interactive demo
6. Q & A

Reminder: prepare for later!

- If you don’t have it already, download and install R (search: “install R”)
- Also install RStudio (search: “install RStudio”)
- Installation should, at most, take as long as the talk portion

Machine learning is correlations

Machine learning is used to build systems

Takes *labels*, correlates with other data

"Predictions" are correlations

Correlations can go wrong

ML examples: Building systems

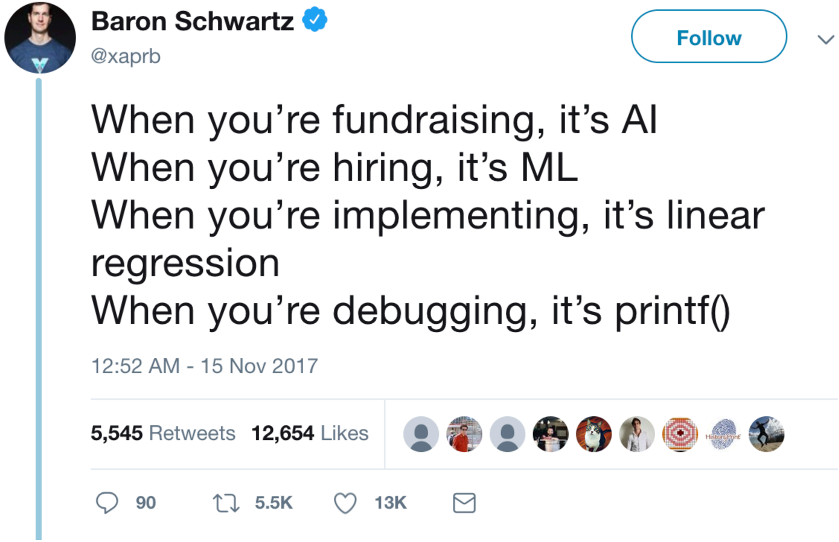
- Recommend/narrow people's choices to "relevant" ones (friend connections, search results, products)
- Detection (facial, fraud)
- Anticipation (customer demand, equipment failure)
- It "works"...

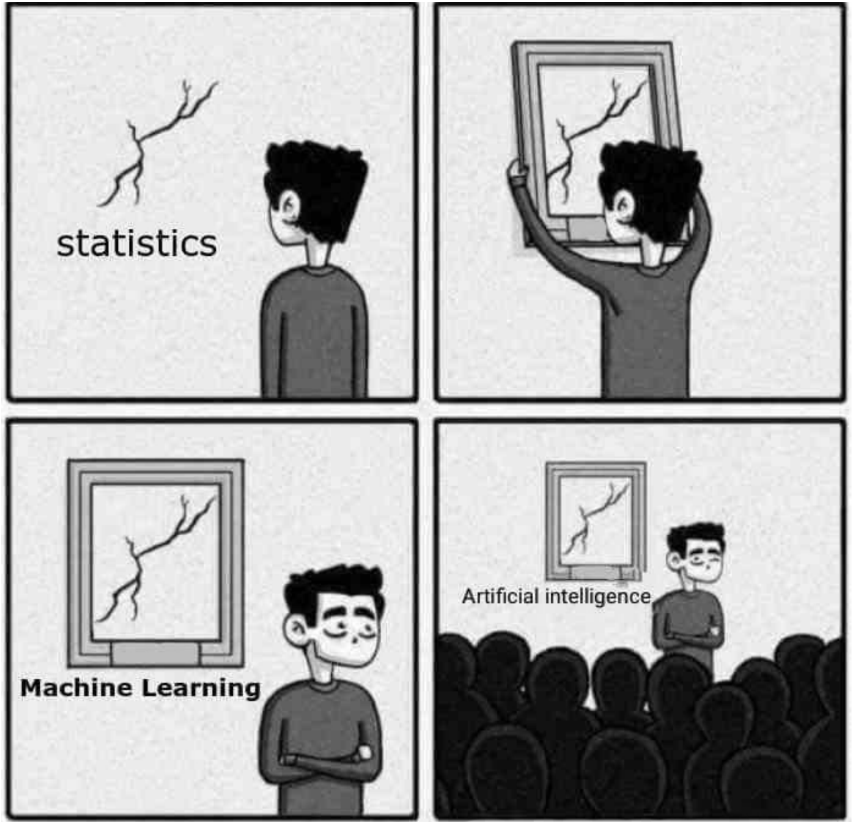
How? Correlates *labels* and other data

"Source subject": Marquese Scott

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

ML is all statistical





(Critiques of statistics apply!)

Source: Todd Rose. Illustration: Future for Learning.

(Critiques of statistics apply!)

The Society Pages

CYBORGOLOGY

Fact Check: Your Demand for Statistical Proof is Racist

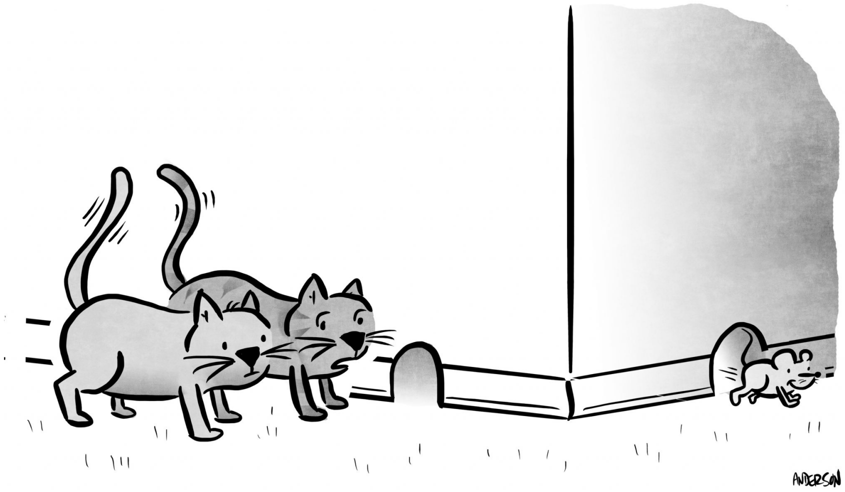
Candice Lanius on January 12, 2015

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."

"Predictions" are just correlations

- Spurious (non-causal) correlations/trends can be used for prediction!
- But this can break down…
- Google Flu Trends: half flu detector, half winter detector (Lazer et al., 2014)
- "X predicts Y" is really "X is correlated with Y"



"According to our current predictive analytics solution, the mouse should be exiting from this hole in 3… 2… 1…" #betterdata

Correlations can go wrong

- Do we know if a *specific* output is right or wrong?
- Treating people based on correlations denies agency and individuality
- Correlations are *proxies*, which can be gamed
- Correlations optimize to the average, leaving out those who are not "average" (as measured!) (Rose, 2014; Keyes, 2018)
- Mistakes can be unequally distributed across groups

Preliminaries

Machine learning is correlations

When to use machine learning

Background needed to do machine learning

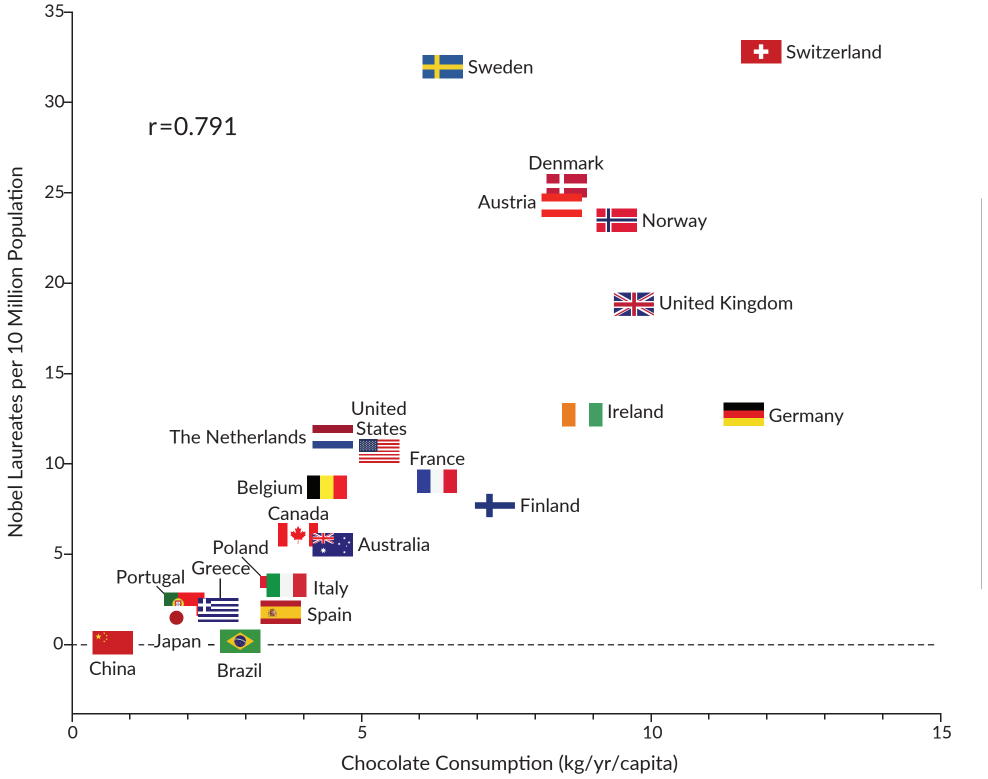
Key concepts

Example for demo: Titanic

Demo

Q & A

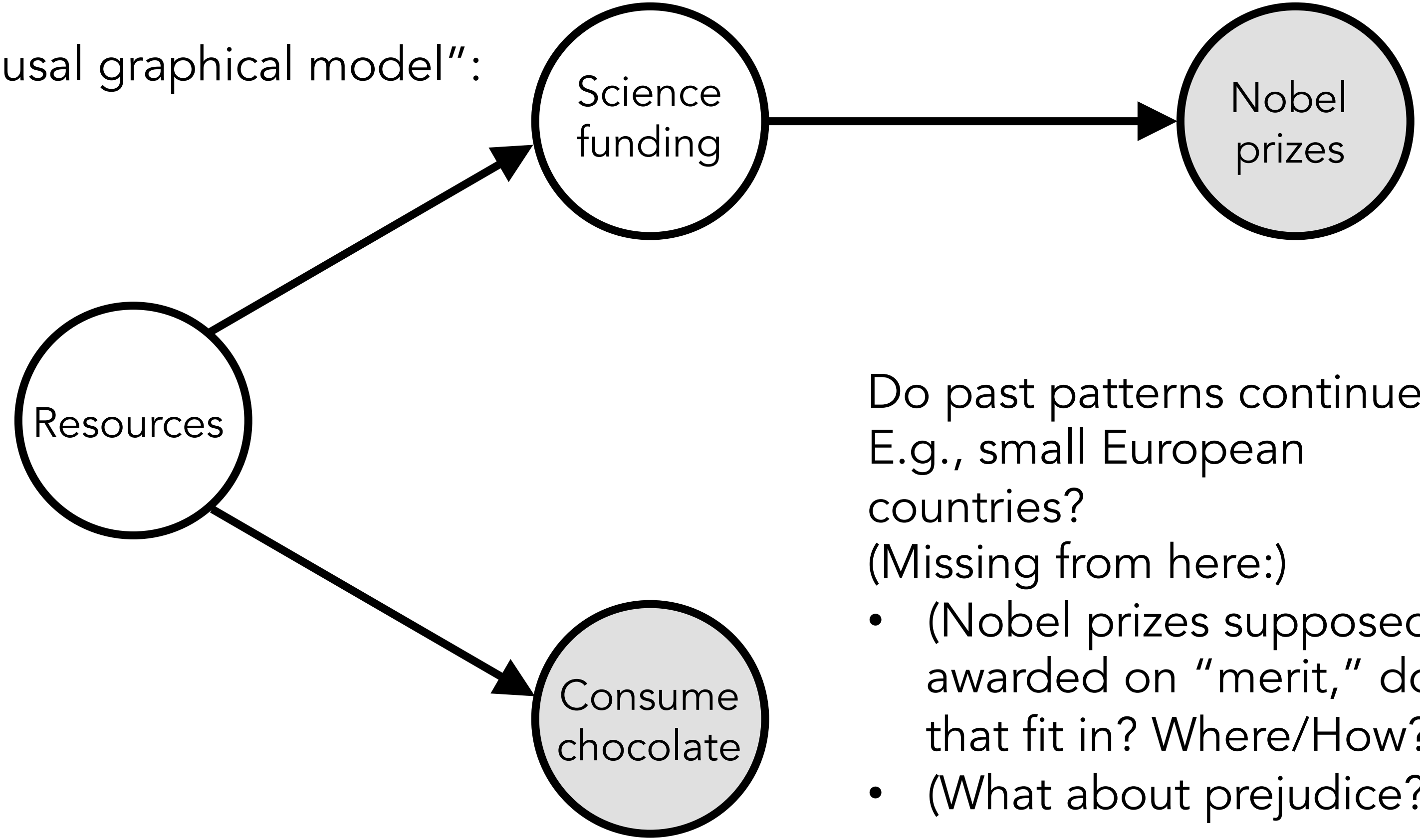
Ex: Chocolate and Nobel prizes



(Messerli, 2012)

Correlated, but *cause* is resources

A "causal graphical model":

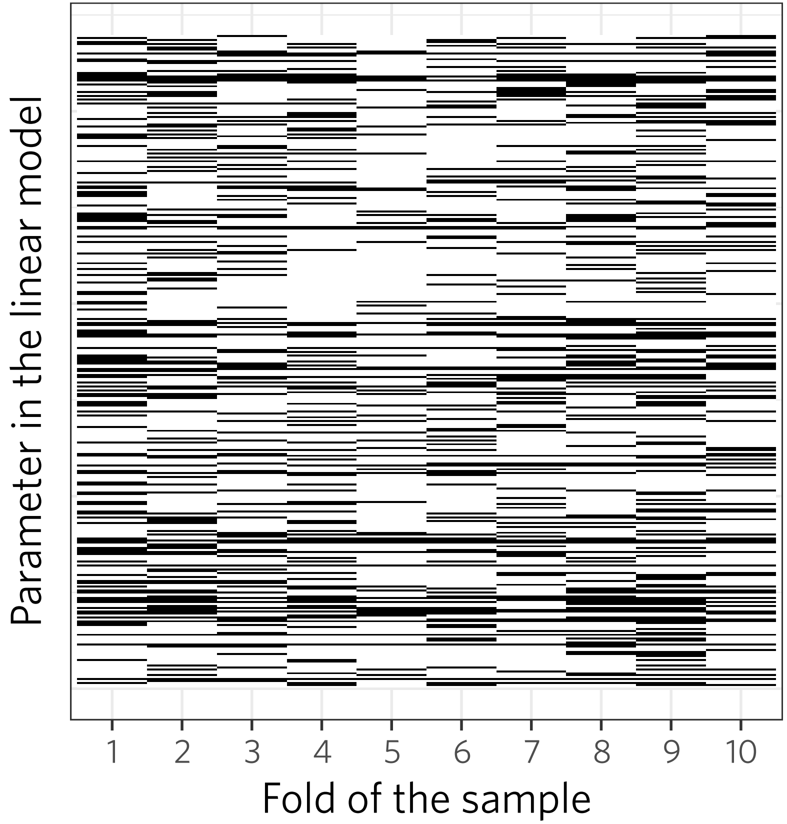


Do past patterns continue? E.g., small European countries?

(Missing from here:)

- (Nobel prizes supposedly awarded on "merit," does that fit in? Where/How?)
- (What about prejudice?)

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)

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
- We can "predict" without "explaining"!

* Or other relevant metric of success

When to use machine learning

Key components of a good use case

Example of a “responsible” use case

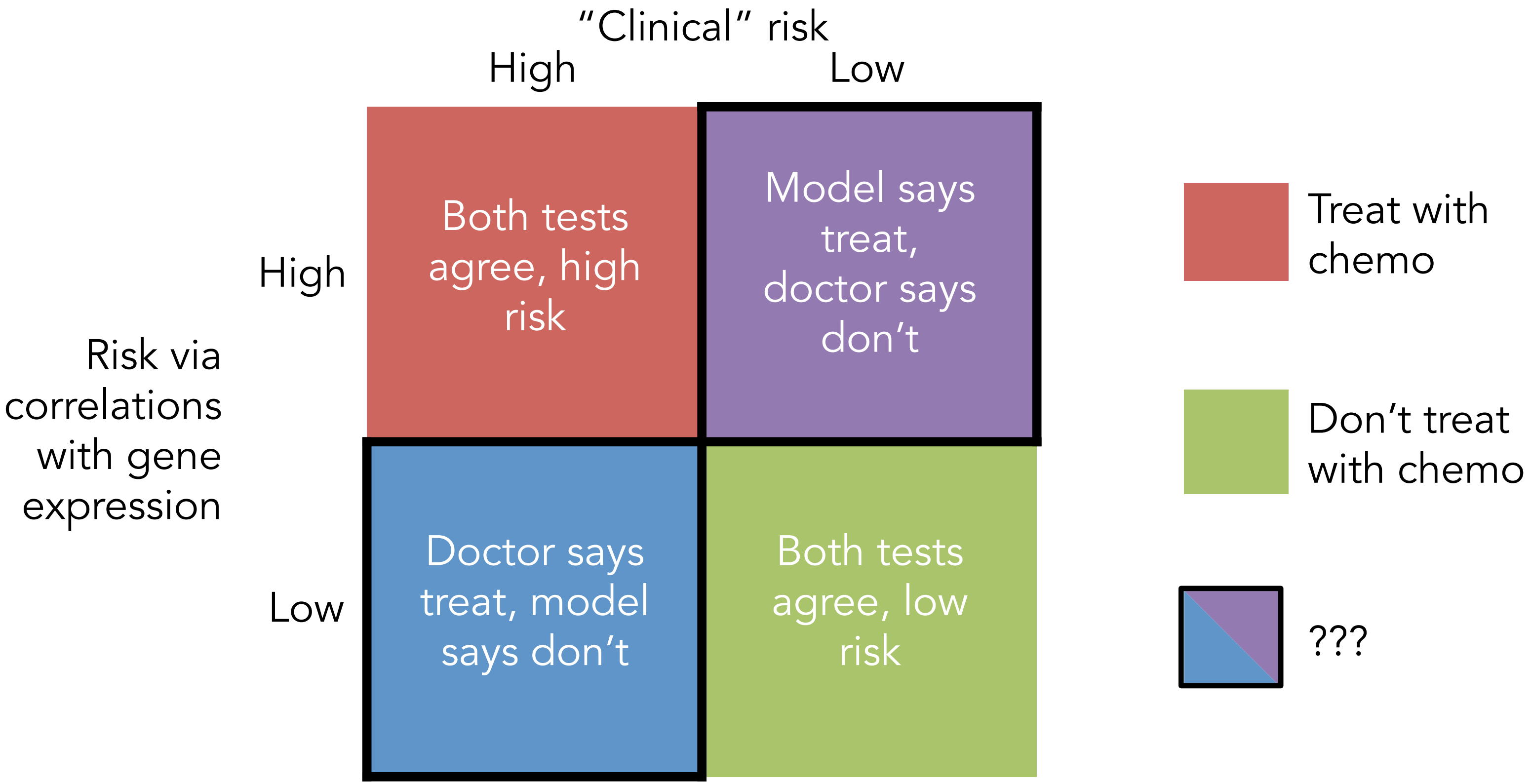
Key components of a good use case

1. We have "ground truth" (e.g., human labels, previous failures/fraud), and
2. Ground truth is hard to collect, and
3. We have some readily available proxy measure, and
4. *We don't care how or what in the proxy recovers the ground truth, only that it does*

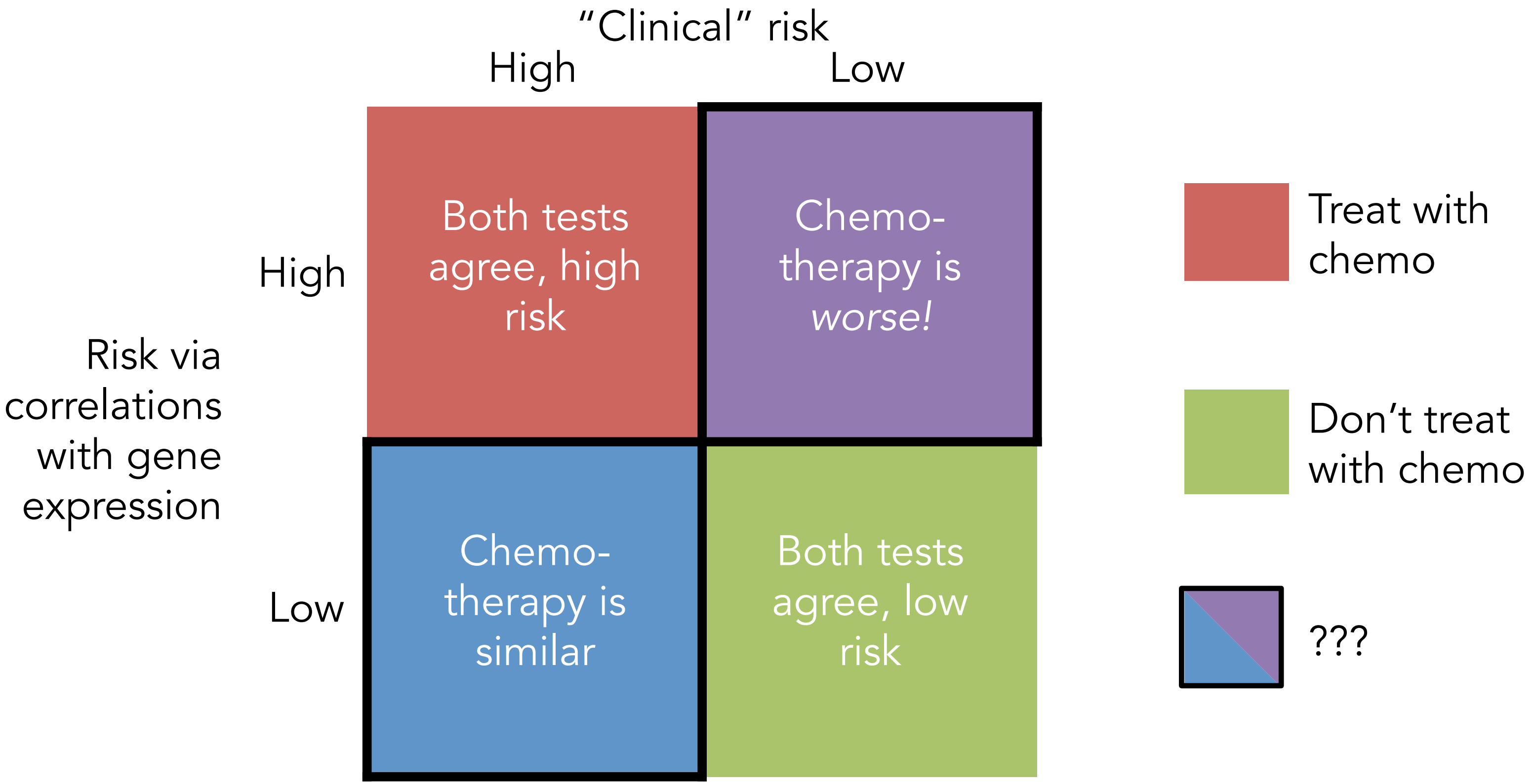
“Responsible” use case

- 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



Real-world testing



Real-world testing

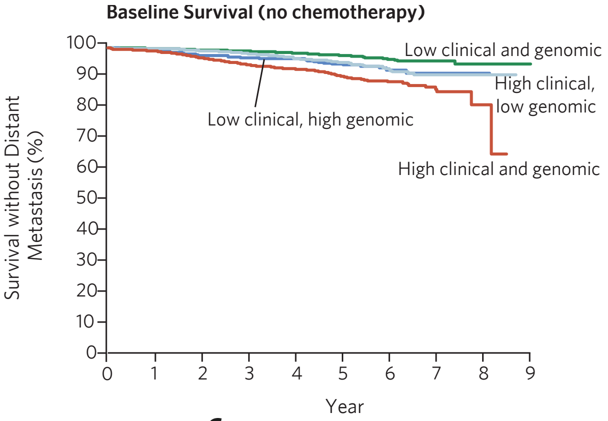
Risk via correlations with gene expression	"Clinical" risk: High	"Clinical" risk: Low
High	Both tests agree, high risk	Chemo-therapy is *worse!*
Low	Chemo-therapy is similar	Both tests agree, low risk

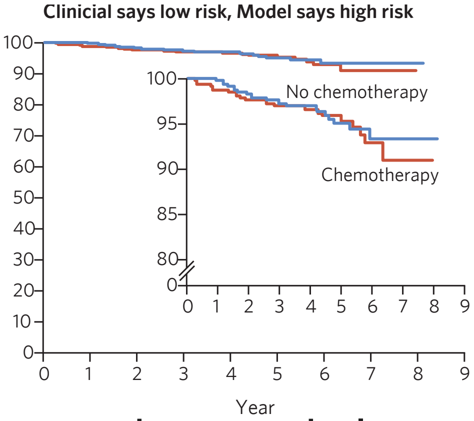
Treat with chemo

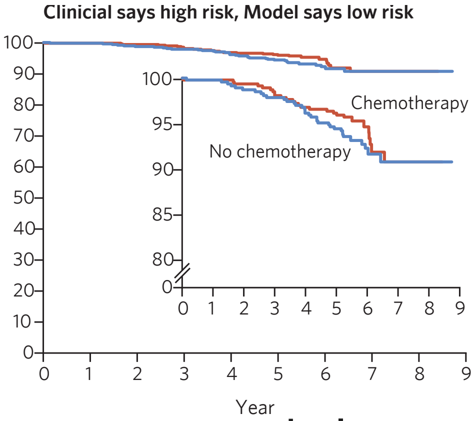
Don't treat with chemo

(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*

Preliminaries
Machine learning is correlations
When to use machine learning
Background needed to do machine learning
Key concepts
Example for demo: Titanic
Demo
Q & A

Background needed to *do* ML

How much programming/CS?

How much math?

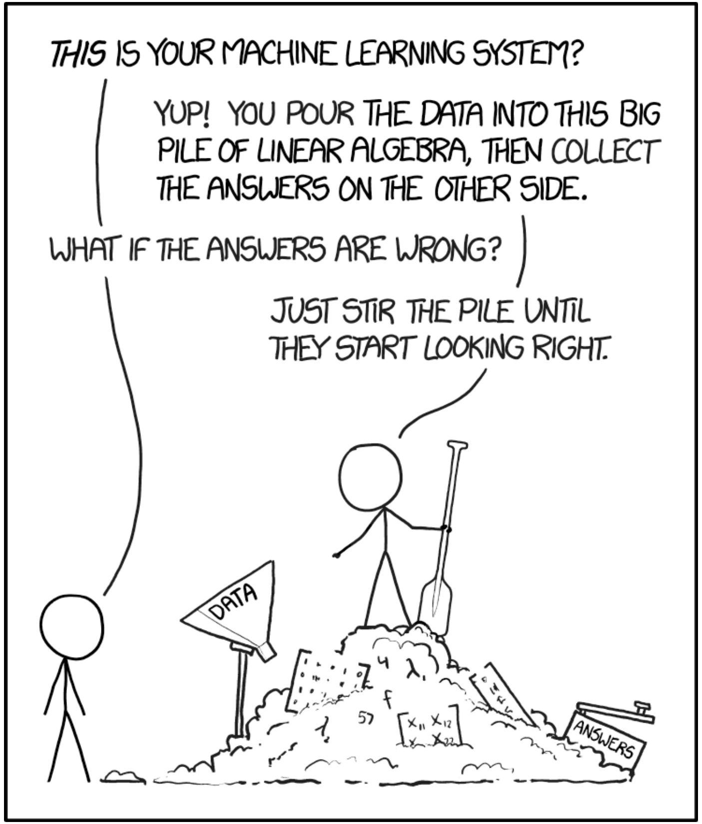
Which language/environment?

Resources

How much programming/CS?

- For personal use: at least be able to write loops and functions, and know up to sorting algorithms. Nothing more!
- For production: some software development principles.
- Alternatives: Weka and Rapid Miner have graphical interfaces, no programming or required

How much math?

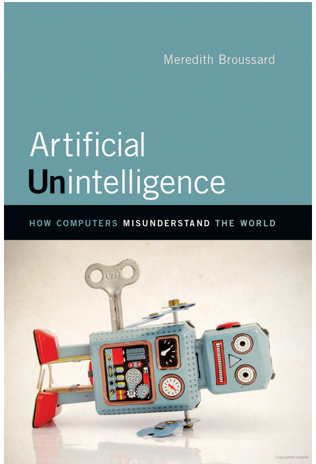


- To be a practitioner, same as what you need to do social statistics: algebra and a bit of calculus
- To understand and advance underlying *mechanics*: linear algebra, multivariate calculus
- To understand underling *principles*: learn probability and mathematical statistics

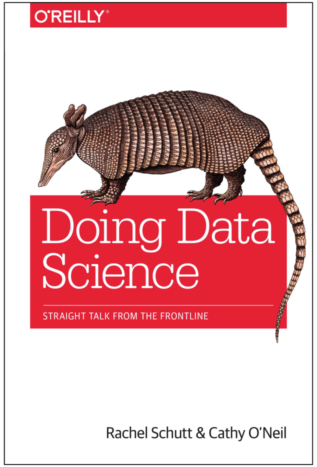
Which language/environment?

- Weka, Rapid Miner
  - Basic use
- Python (numpy, scipy, scikitlearn, pandas)
  - Scale, integrating into production, best visualizations (sometimes), all deep learning
- R
  - More flexibility in how to use techniques, a self-contained environment, and better integration with (social) statistics

Resources

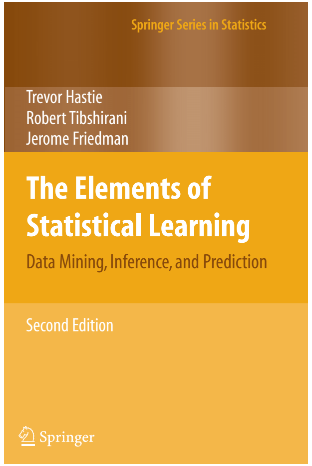


Chapter 7: ML in action



Basics

Machine learning without needing to know any programming



Theory

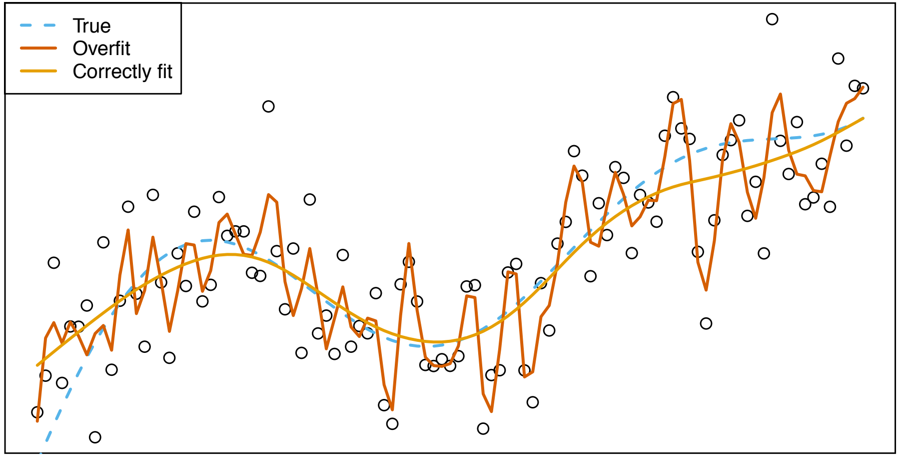
Unfortunately, I haven't spent time looking through online courses to have one I recommend.

Key concepts

Model "fit"

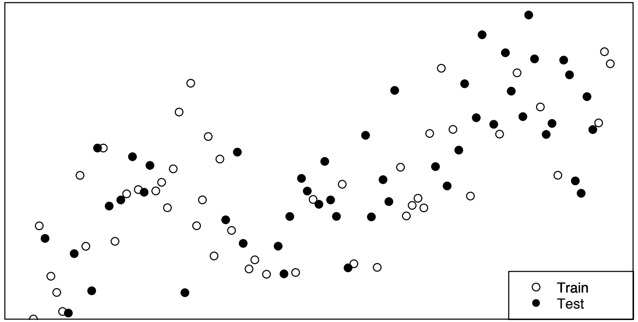
- All machine learning and statistics models take in data, process them via some assumptions, and then give out something: relationships, and/or likely future values.
- The processing is called "fitting", and the output is called a "fit." Machine learning uses "learning" or "training," but it's the same.

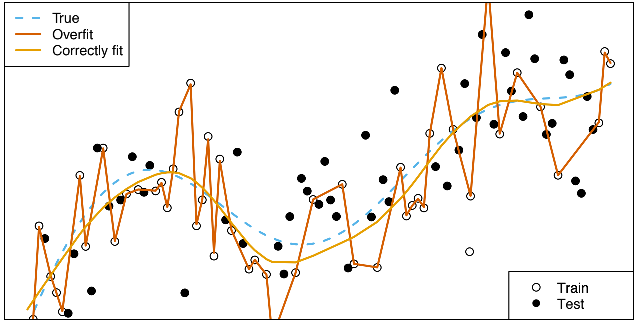
Overfitting: fit to noise



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

Data splitting: Catch overfitting

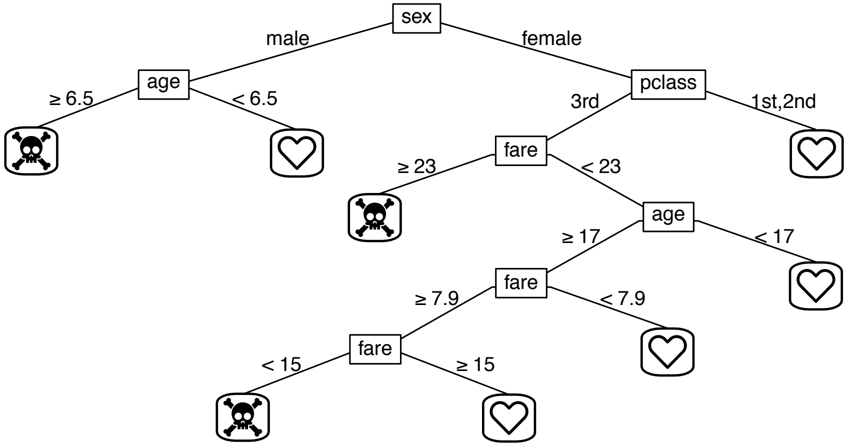


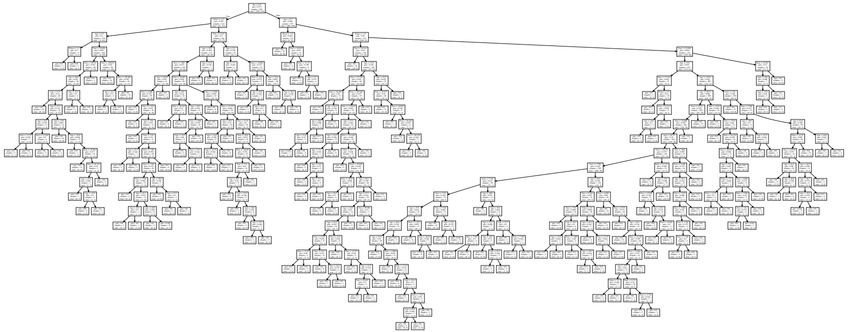


- Idea: if we split data into two parts, the signal should be the same but the noise would be different
- *Cross validation:* Fitting the model on one part of the data, and "testing" on the other

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

(Discrete version of overfitting)





"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

Accuracy = (TP+TN)/N

↑ Overall correct

Confusion matrix

True label

		Positive	Negative
	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		

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

Example for demo: *Titanic*

Datacamp "Titanic" example

New York American

EDITION FOR 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. TUESDAY, APRIL 16, 1912. 16 PAGES PRICE ONE CENT

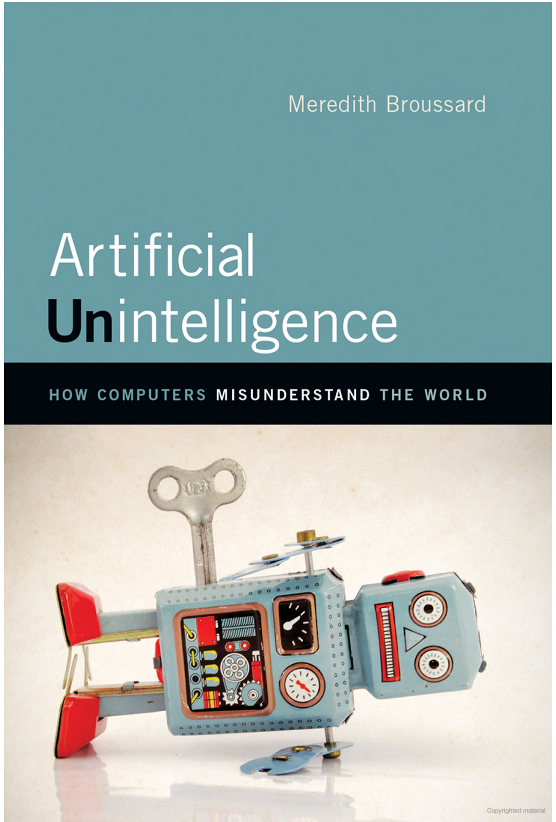
J. J. ASTOR LOST ON TITANIC

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.

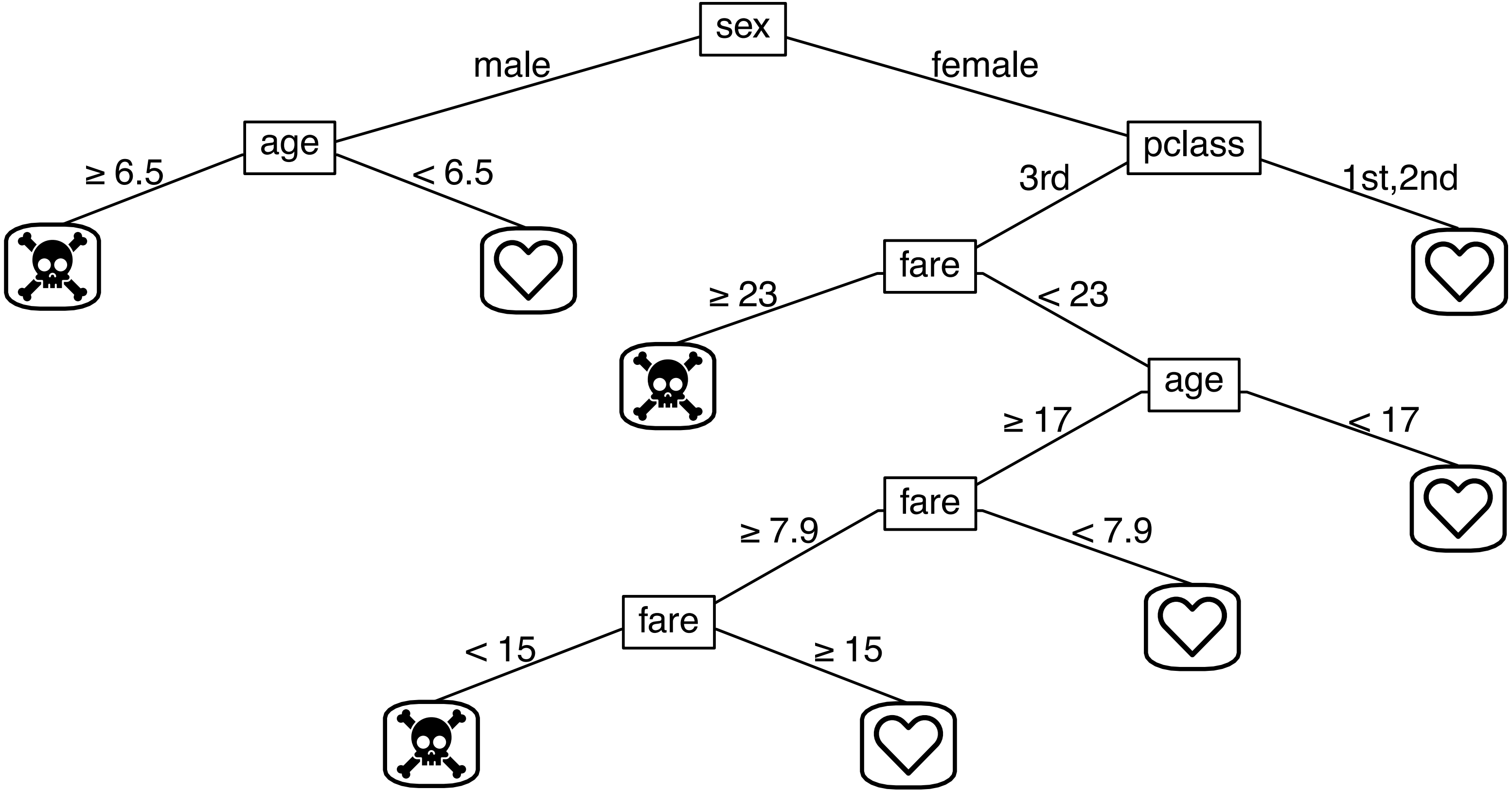
The Wireless Operator at Cape Race, Newfoundland, Flashes: "Eighteen Hundred Lives Have Been Lost in the Wreck of the Titanic."

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

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

End note: Decide how distracted to be

"The function, the very serious function of racism is distraction. It keeps you from doing your work. It keeps you explaining, over and over again, your reason for being. Somebody says you have no language and you spend twenty years proving that you do. Somebody says your head isn't shaped properly so you have scientists working on the fact that it is. Somebody says you have no art, so you dredge that up. Somebody says you have no kingdoms, so you dredge that up. None of this is necessary. There will always be one more thing."

–Toni Morrison, 1931-2019

(Thanks to my partner, Maya Randolph)

Thank you! Questions?

- Please send feedback! https://forms.gle/TrY7z6qivuVf2C8p7
- Contact me: momin_malik@cyber.harvard.edu
- Summary:
  - Machine learning is correlations
  - Can be powerful, but also can fail and (both in successes and failures) be oppressive
  - It leaves out a lot

References

Agre, Philip E. "Notes on critical thinking, Microsoft, and eBay, along with a bunch of recommendations and some URL's." *Red Rock Eater Newsletter*, 12 July 2000. https://pages.gseis.ucla.edu/faculty/agre/notes/00-7-12.html.

Breiman, Leo. "Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author)." *Statistical Science* 16, no. 3 (2001): 199–231. doi: 10.1214/ss/1009213726.

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

Cardoso, Fatima, Laura J. van't Veer, Jan Bogaerts, Leen Slaets, Giuseppe Viale, Suzette Delaloge, Jean-Yves Pierga, Etienne Brain, et al. "70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer." *The New England Journal of Medicine* 375, no. 8 (2016): 717–729. doi:10.1056/NEJMoa1602253

D'Ignazio, Catherine and Lauren Klein. *Data Feminism*. MIT Press, 2019.

Efron, Bradley, and Carl Morris. "Stein's Paradox in Statistics." *Scientific American* 236, no. 5 (1977): 119–127. doi:10.1038/scientificamerican0577-119.

Fay, Brian. *Critical Social Science: Liberation and its Limits*. Ithaca, New York: Cornell University Press, 1987.

Frey, Frey, Bruno S., David A. Savage, and Benno Torgler. "Behavior under Extreme Conditions: The *Titanic* Disaster." *Journal of Economic Perspectives* 25, no. 1 (2011): 209–222. doi:10.1257/jep.25.1.209.

Frey, Bruno S., David A. Savage, and Benno Torgler. "Noblesse Oblige? Determinants of Survival in a Life-or-Death Situation." *Journal of Economic Behavior & Organization* 74, nos. 1–2 (2010): 1–11. doi10.1016/j.jebo.2010.02.005.

Jones, Matthew L. "How We Became Instrumentalists (Again): Data Positivism since World War II." *Historical Studies in the Natural Sciences* 48, no. 5 (2018): 673–684. doi:10.1525/hsns.2018.48.5.673.

Keyes, Os. The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. In *Proceedings of the ACM on Human-Computer Interaction* (CSCW), 2, 88:1–88:22, 2018.

Lanius, Candice. *Fact Check: Your Demand for Statistical Proof is Racist*. *Cyborgology* [blog], January 2015. https: //thesocietypages.org/cyborgology/2015/01/12/fact-check-your-demand-for-statistic al-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.

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

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.

Rose, Todd. *The End of Average: How We Succeed in a World that Values Sameness*. New York: HarperCollins Publishers, 2015.

Rose, Todd. "When U.S. Air Force Discovered the Flaw of Averages." *The Star*, 16 January 2016. https://www.thestar.com/news/insight/2016/01/16/when-us-air-force-discovered-the-flaw-of-averages.html. Images: Screengrabs from Future for Learning, "Jet Cockpits and The End of Average." https://vimeo.com/237632676.

Schutt. Rachel and Cathy O'Neil. *Doing Data Science: Straight Talk from the Front Line*. O'Reilly Media, 2014.

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?

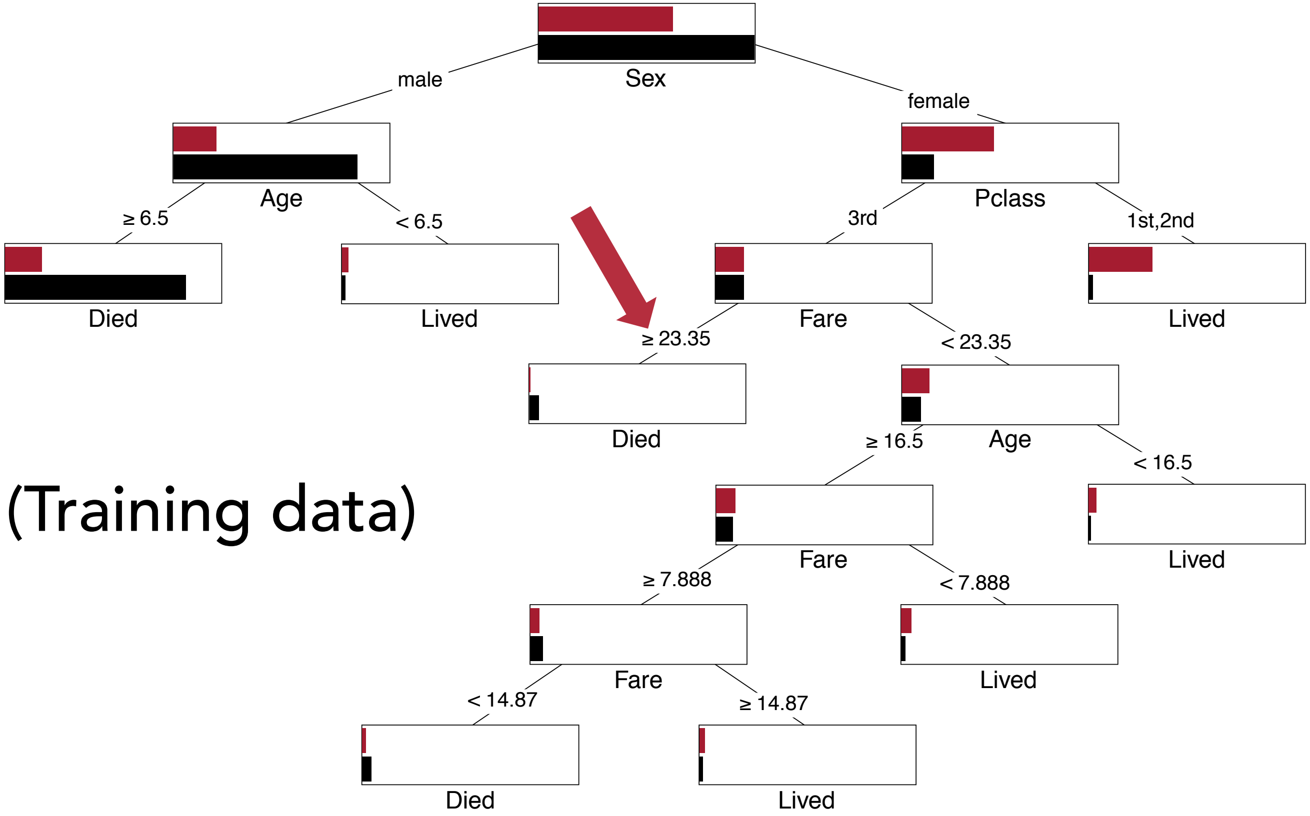
But ML is correlations, not causes

- 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

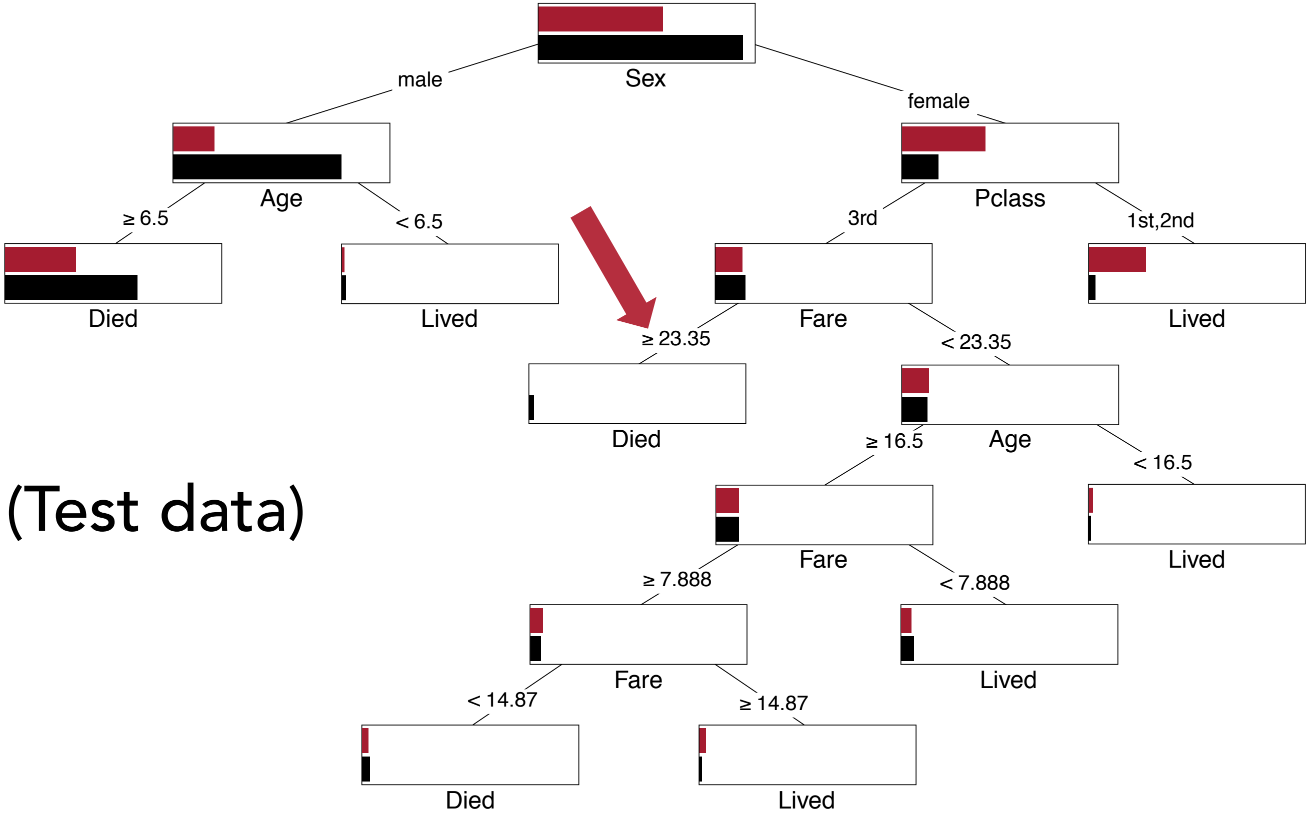
- Face validity as a way to check the model
- Anticipate where the model might break down (e.g., when it fails face validity)
- Use domain knowledge to 'fine-tune' the model

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



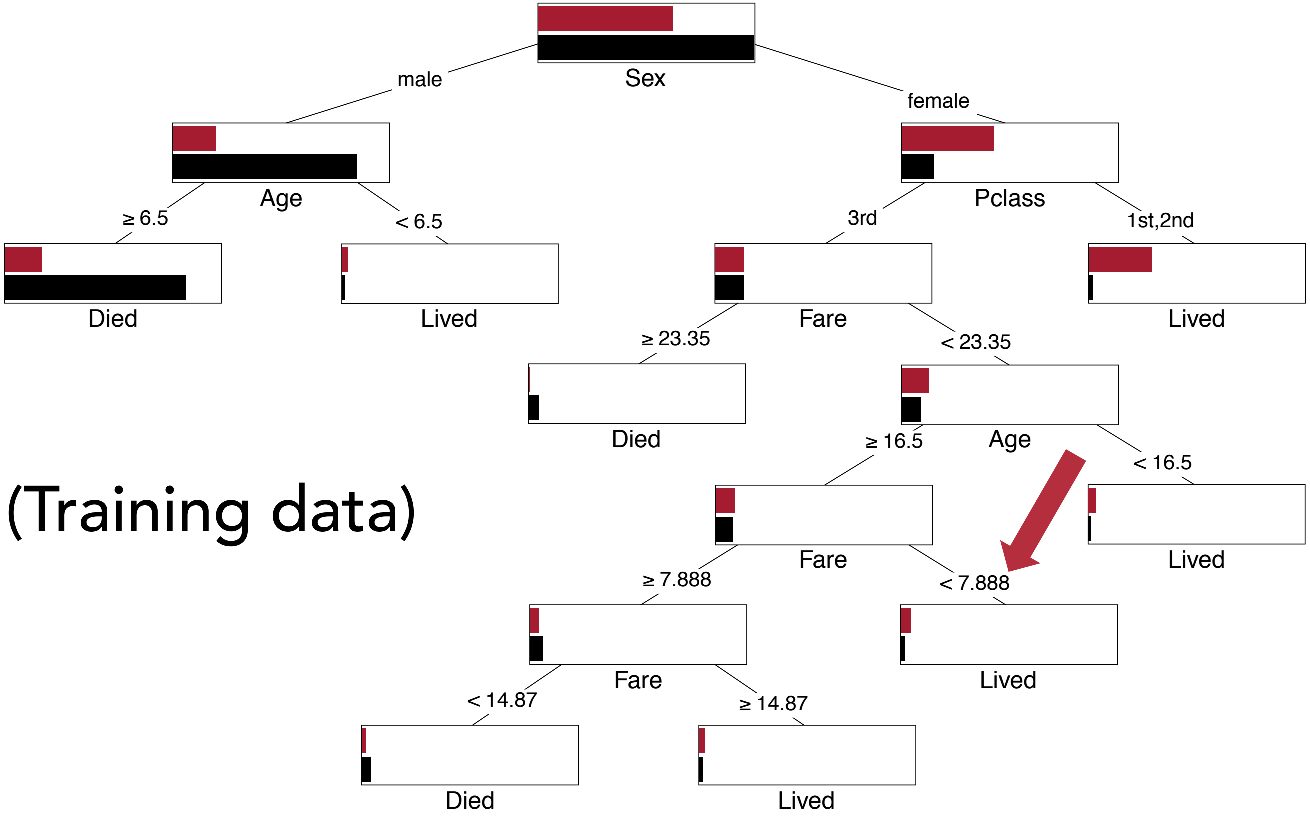
(Training data)

Lacks face validity, but holds on test data



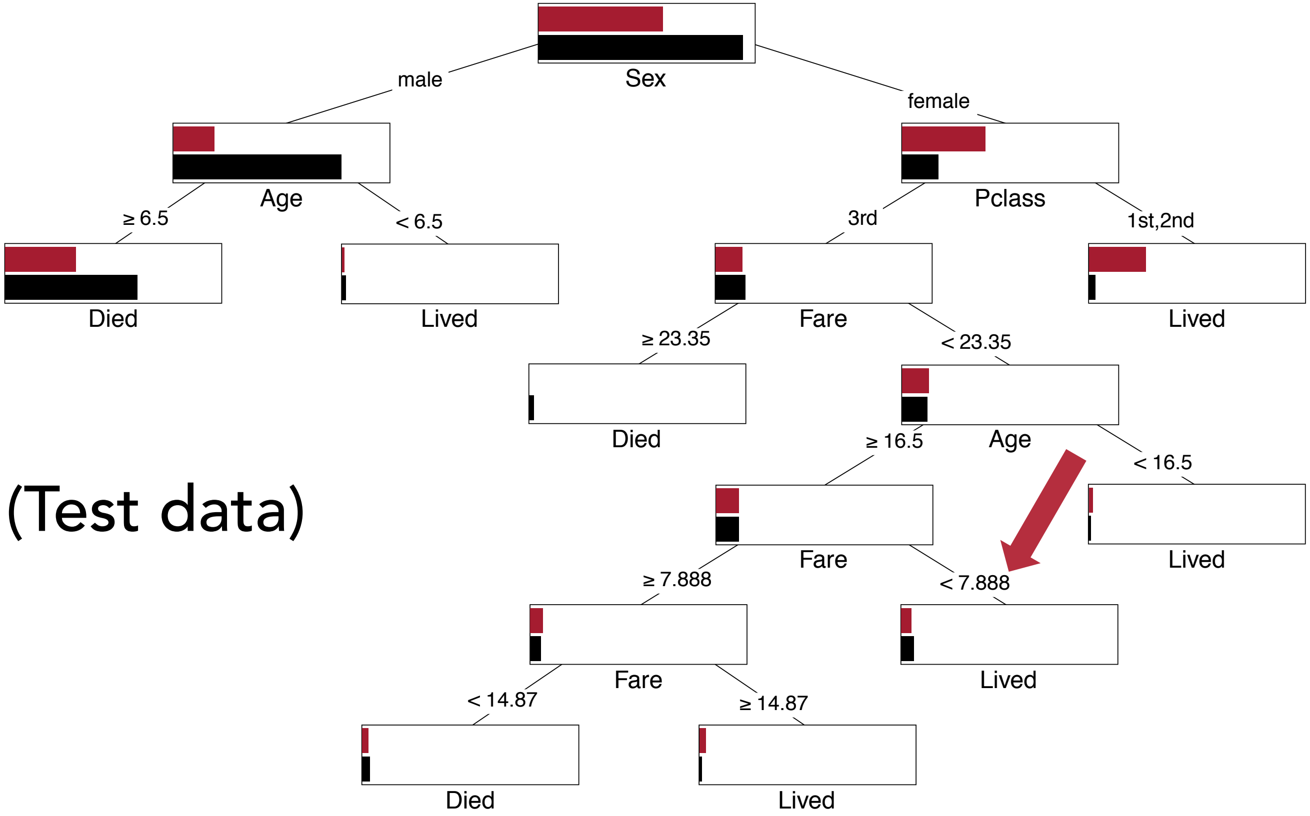
(Test data)

Converse: has face validity, but fails to generalize?



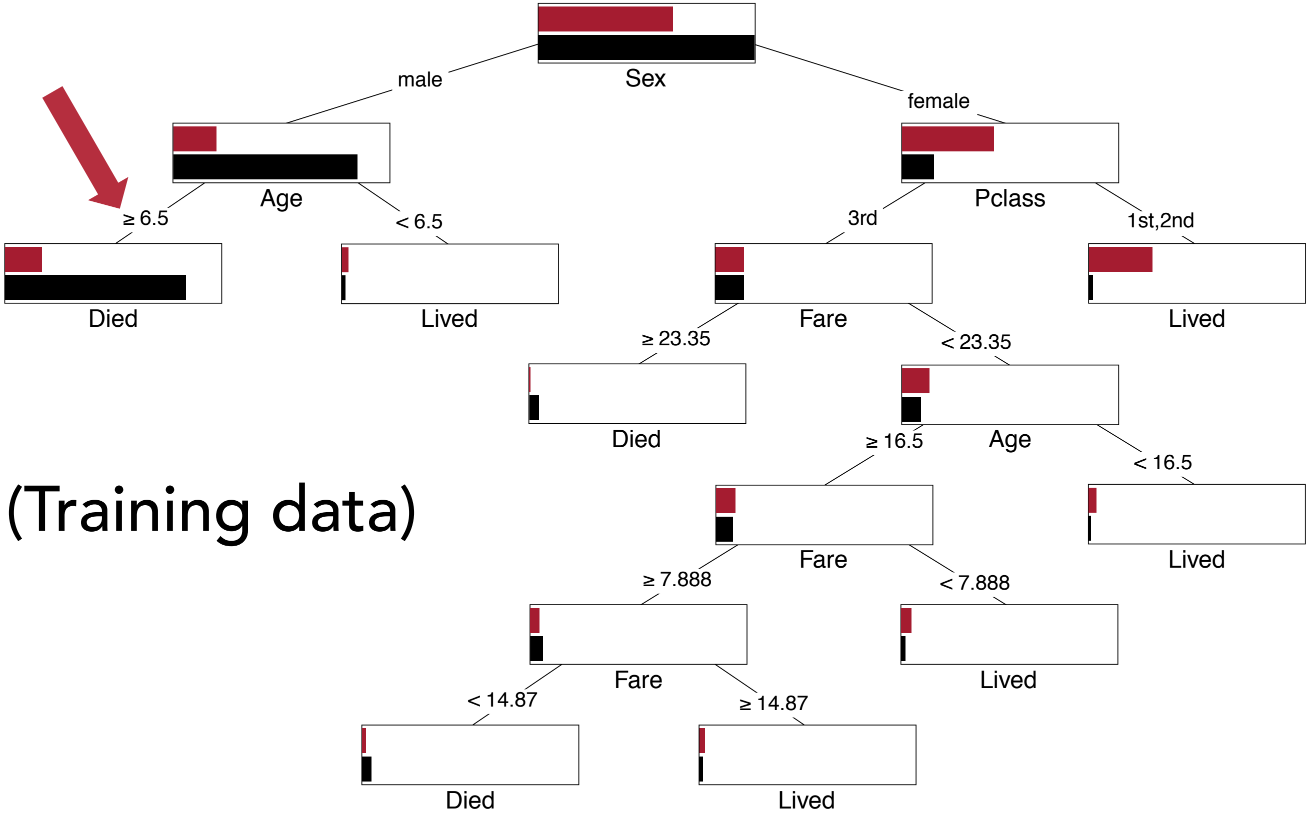
(Training data)

Yes. Interpretability doesn't help anticipate breakdowns



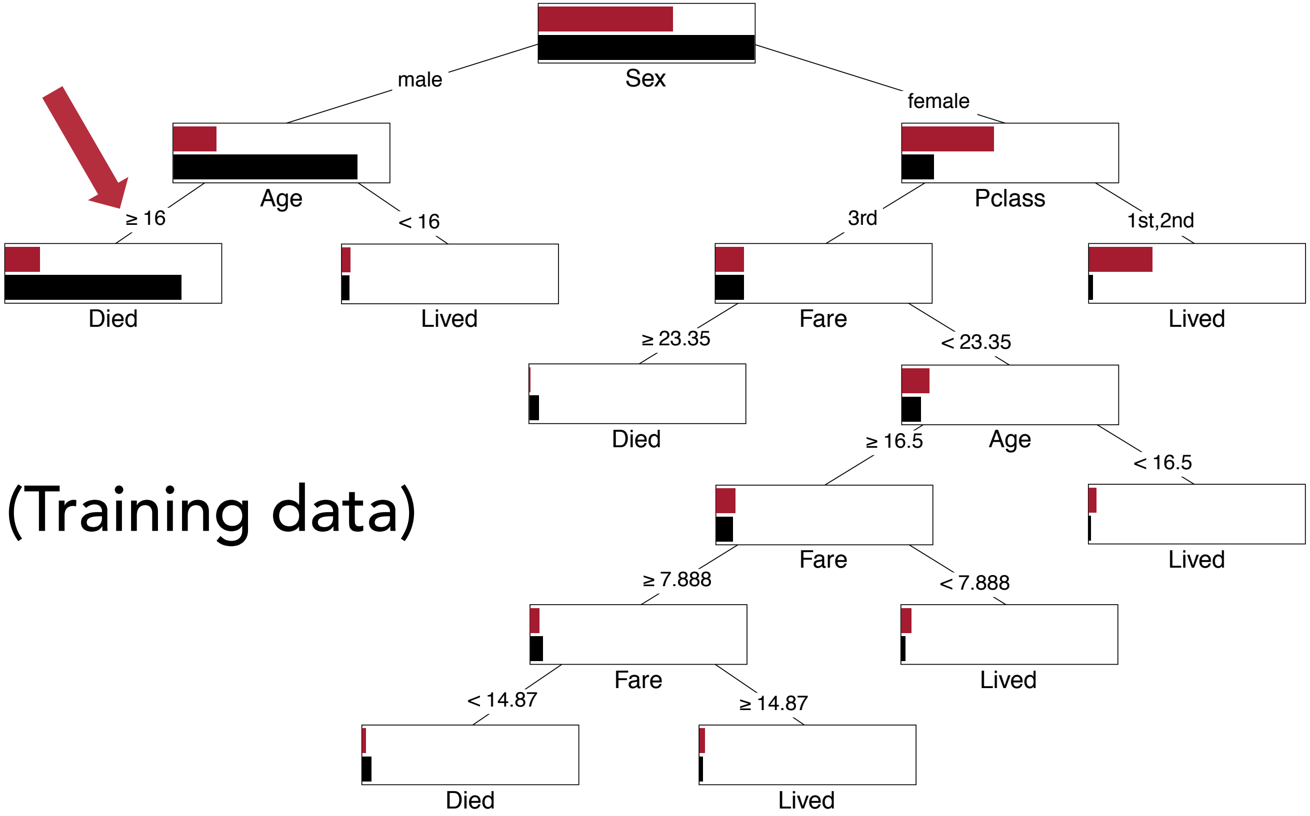
(Test data)

Interpretations to 'fine-tune' model?



(Training data)

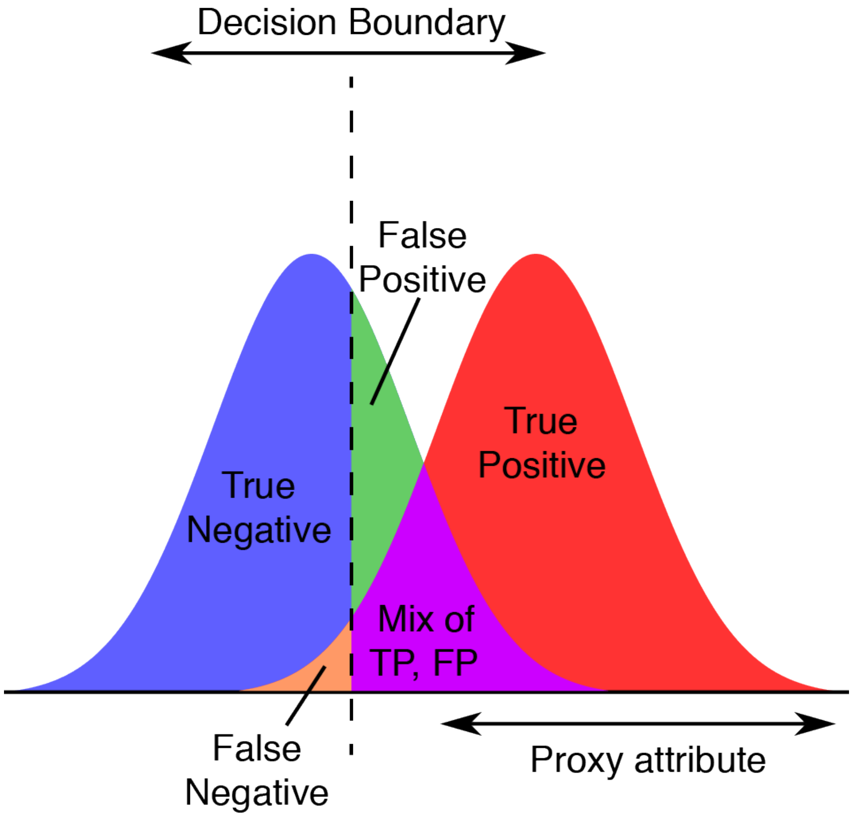
Model is already optimally tuned



(Training data)

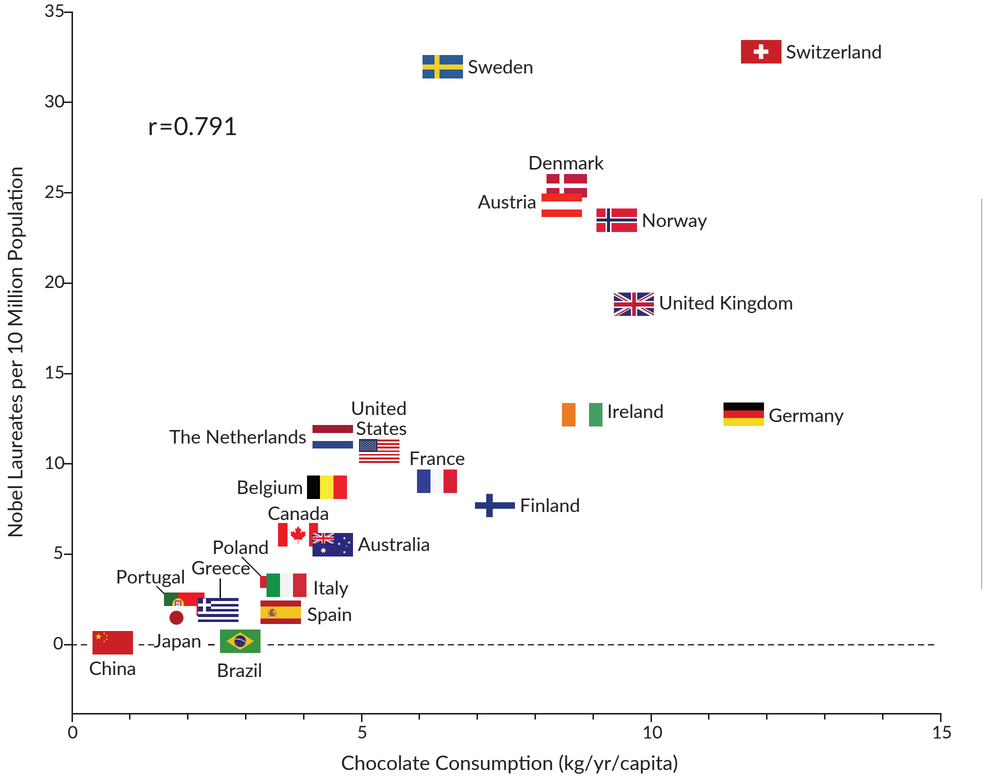
Extra: Discrete "correlations"

ML model = "Ground truth" + proxy

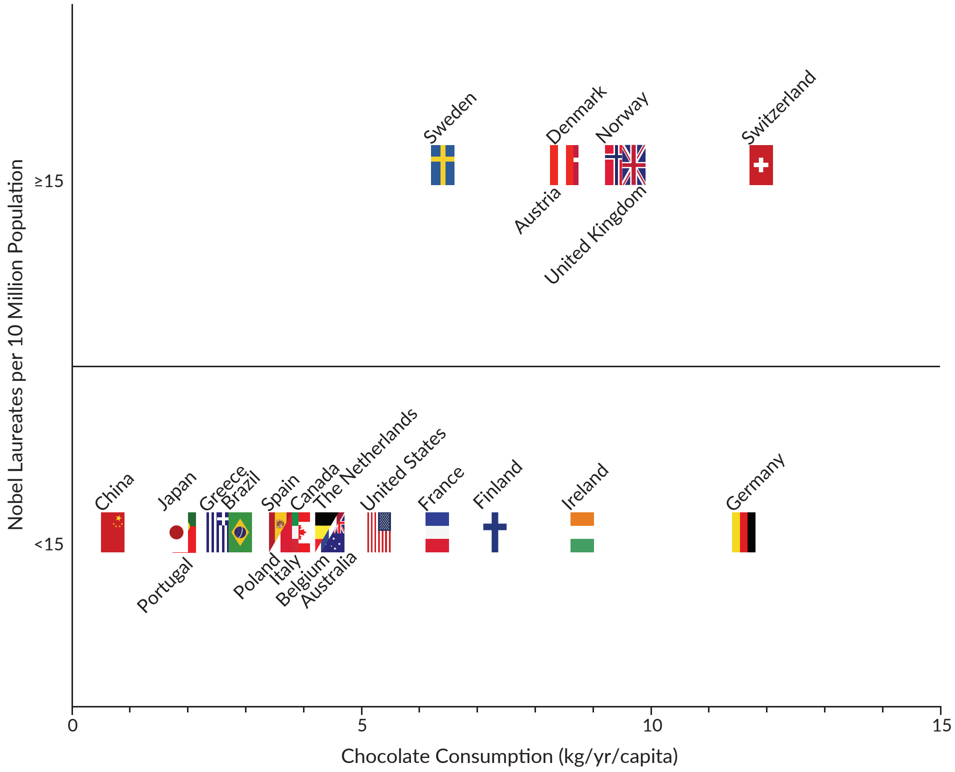


- Correlate known values/ labels with available proxy for unknown values/labels
- Find *decision boundary/ criterion/threshold*. Use this to treat new observations
- Shift that boundary to prioritize certain metrics
- Most ML is basically this!

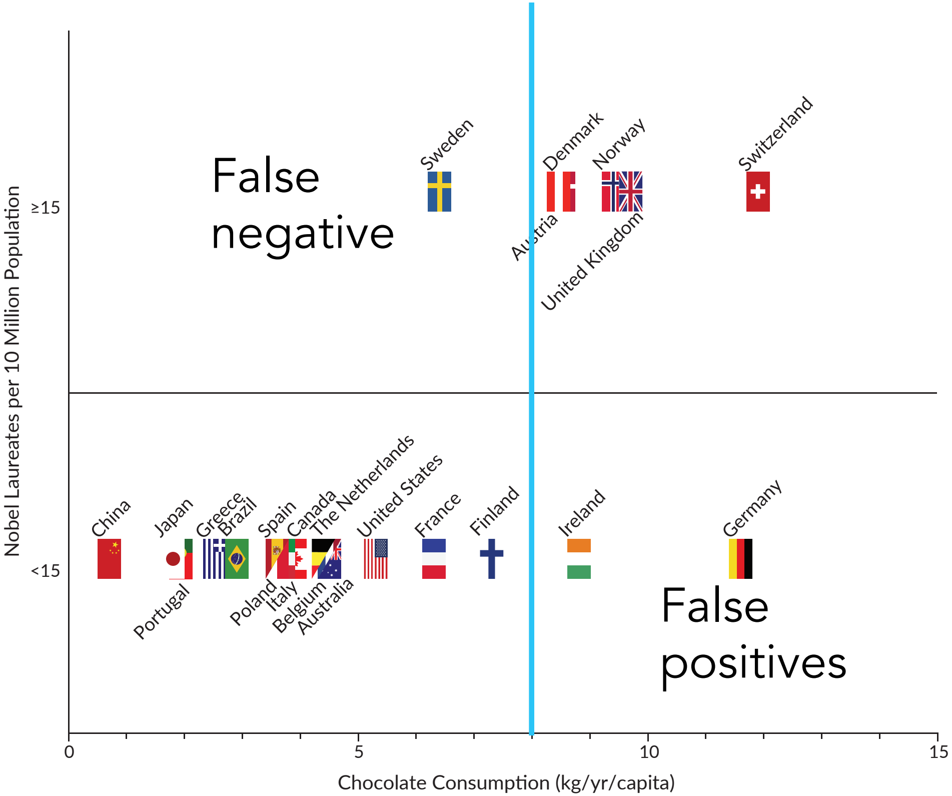
Regression: Continuous relationship



Classification: Discrete relationship



Fit the decision boundary



The prediction: the majority class