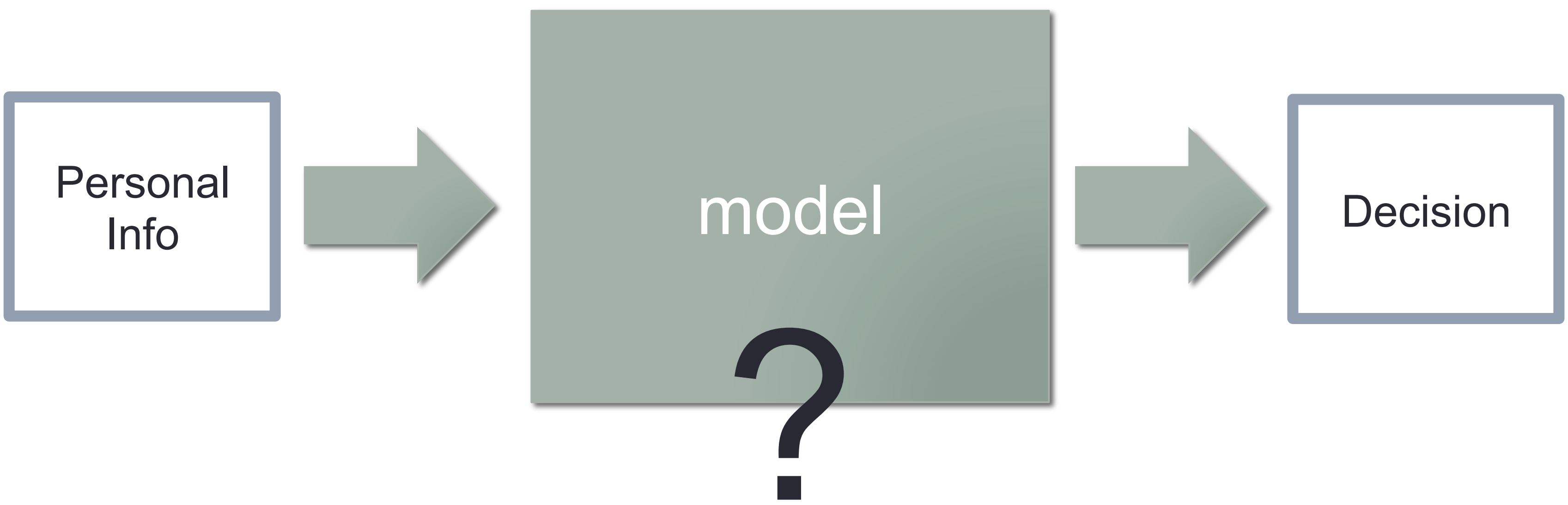
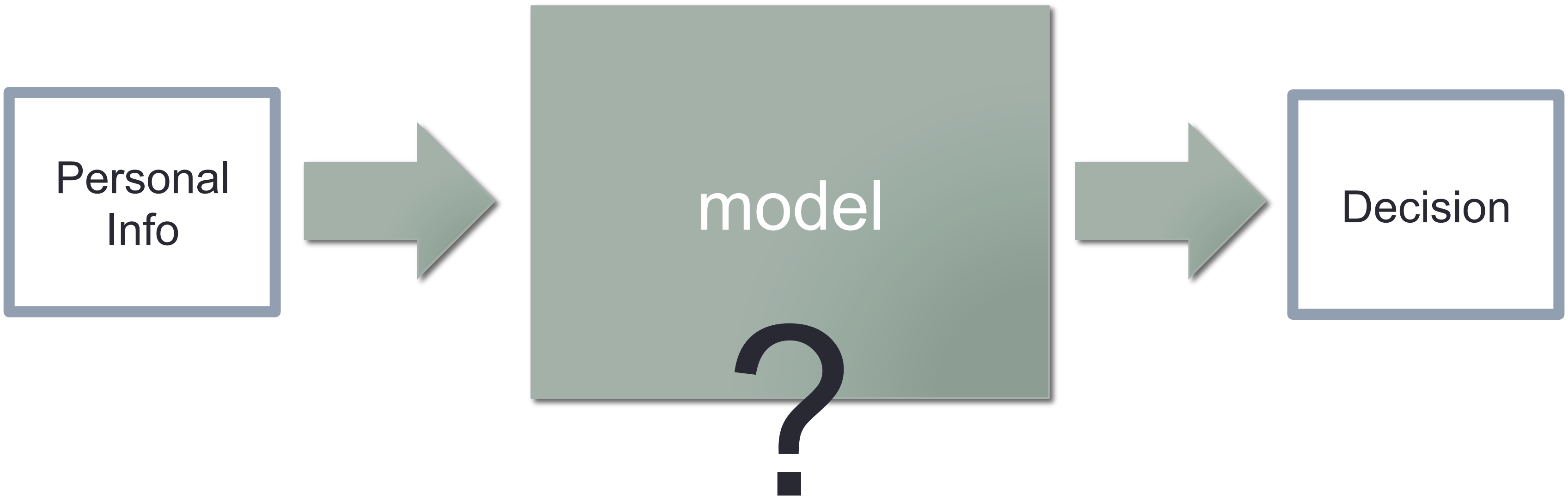
Interpreting and Auditing Machine-learning Algorithms

Sorelle Friedler
Haverford College

How is a model making its decision?



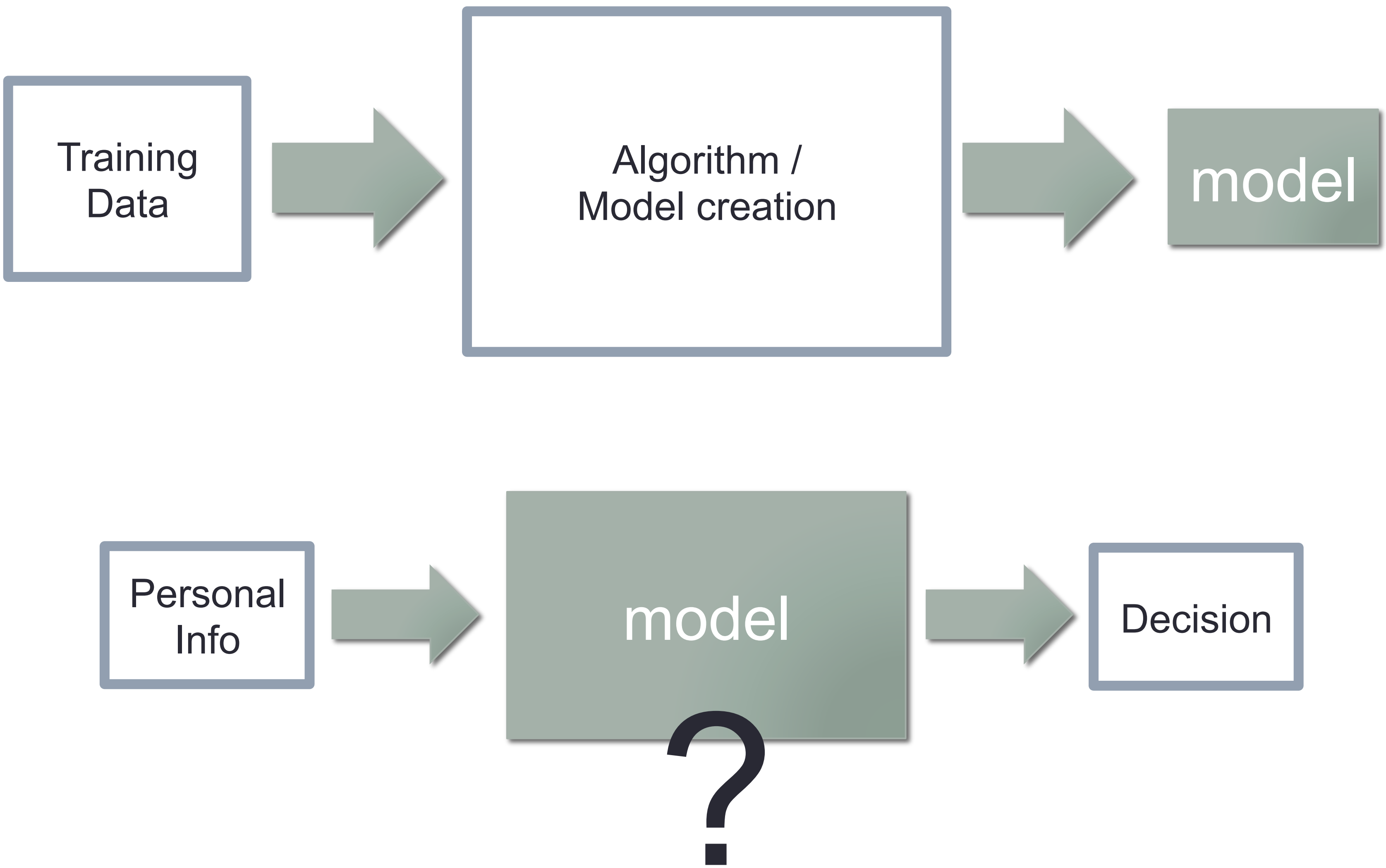
How is a model making its decision?



…for one person?

…for all people?

How is a model made?



How does bias happen in model creation?

BBC Trending

Facebook challenges legitimacy of some Native names

By BBC Trending
What's popular and why

3 March 2015

Share

Dana Lone Hill had trouble using her real name on Facebook

When Lance Browneyes of the Oglala Lakota community in South Dakota was blocked from Facebook for using a "fake" name, he submitted proof of his identification. Facebook then changed his name to Lance Brown.

http://www.bbc.com/news/blogs-trending-31699618

A hypothetical case study

Training Data

Name	Top 1000 baby name	Dictionary word?	Real Name?
Sorelle Friedler	no	no	yes
Lady Gaga	no	yes	no
Big Bird	no	yes	no
Barack Obama	no	no	yes

Dana Lone Hill	yes	yes	???

BBC Trending

Facebook challenges legitimacy of some Native names

By BBC Trending
What's popular and why

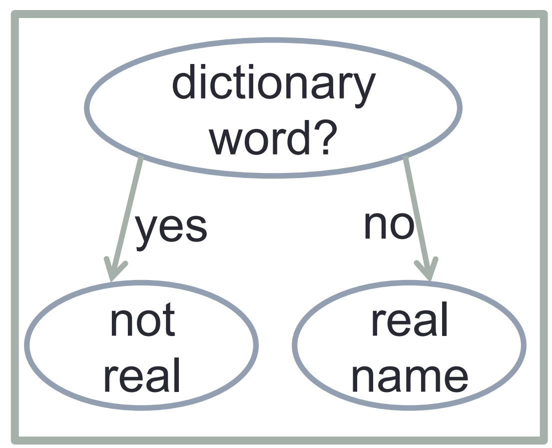
3 March 2015

Share

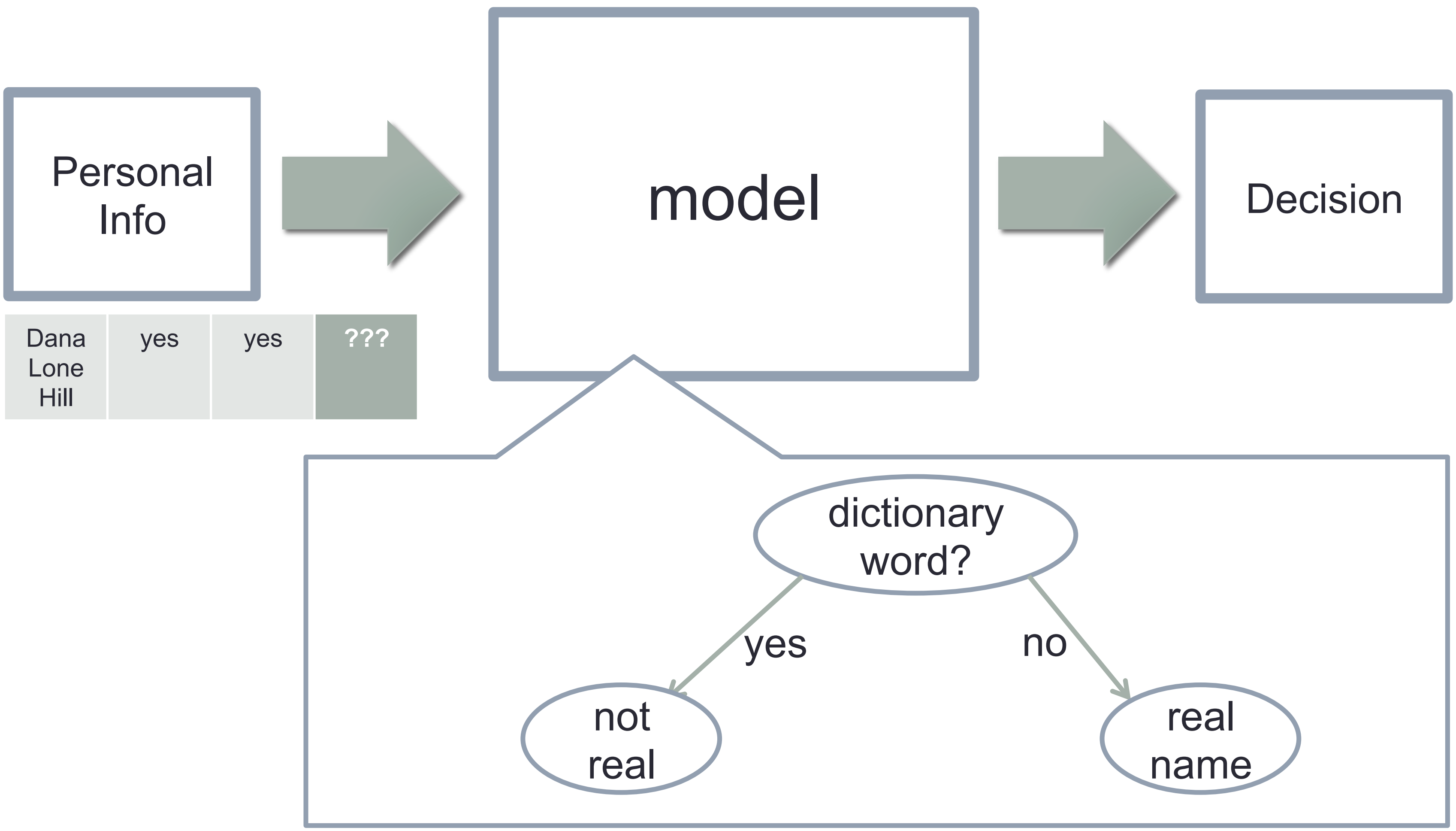
DANA LONE HILL

Dana Lone Hill had trouble using her real name on Facebook

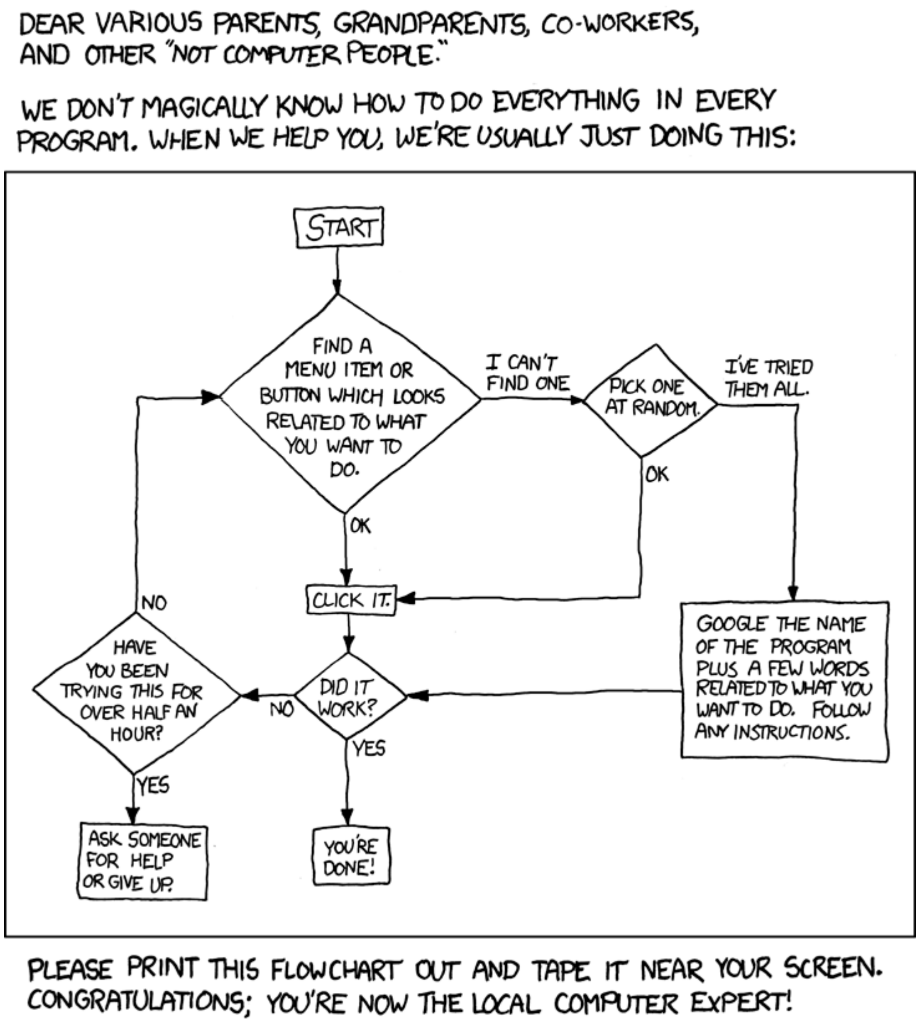
When Lance Browneyes of the Oglala Lakota community in South Dakota was blocked from Facebook for using a "fake" name, he submitted proof of his identification. Facebook then changed his name to Lance Brown.



Interpretable models

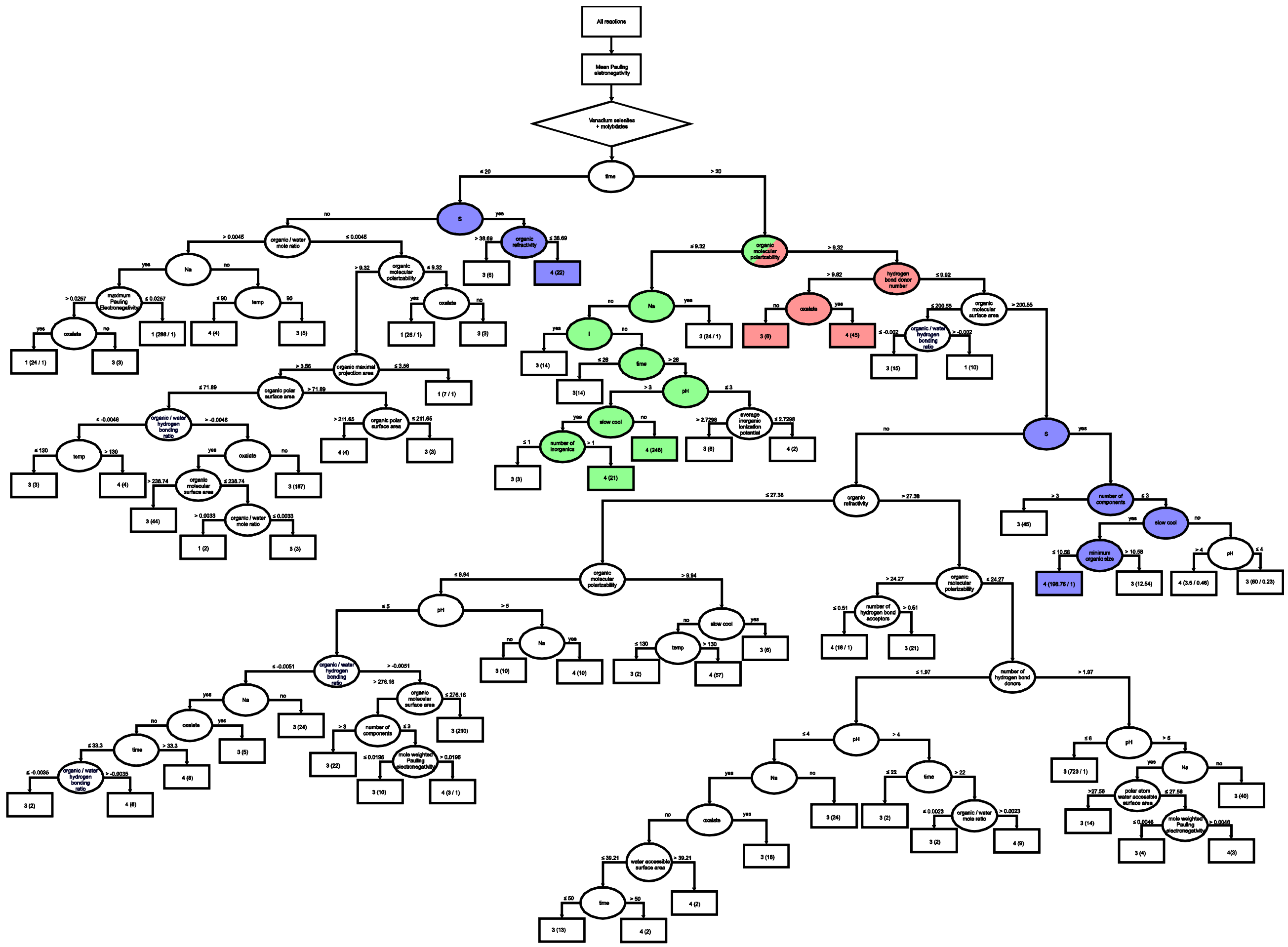


Interpretable models – decision trees



https://xkcd.com/627/

Interpretable models?



Paul Raccuglia, Katherine C. Elbert, Philip D. F. Adler, Casey Falk, Malia B. Wenny, Aurelio Mollo, Matthias Zeller, Sorelle A. Friedler, Joshua Schrier, and Alexander J. Norquist. Machine-learning-assisted materials discovery using failed experiments. Nature, 533: 73 - 76, May 5, 2016. http://dx.doi.org/10.1038/nature17439

Interpretable models

Cosmo Quiz

BY HOLLY EAGLESON

Is Your BFF Really on Your Side?

1. If you and your friend meet a cute guy you're both into, what would she do after you declare that you like him?

- a. Suggest you go for him and offer to be your wingwoman
- b. Flirt with him but back off if he's clearly more into you than her
- c. Say "Really? Hmm. He doesn't seem like your type."

2. When she bails on plans to hang out with you to get together with her boyfriend instead, you're:

- a. Hurt but not exactly surprised. It's happened many times before.
- b. Shocked. She's always made such a big deal about putting friends first.
- c. Disappointed yet aware that you've both flaked occasionally.

3. You just came up with the coolest idea for a blog. How would she react to it when you tell her?

- a. You don't usually share stuff like that with her because she can be critical of your ideas.
- b. Enthusiastically, and she'll point out possible snags that you haven't thought through yet.
- c. She'd be 100 percent behind it and offer to help, no questions asked.

4. When you two have a disagreement, how does it usually get resolved?

- a. You both speak your piece then back off and forget about it.
- b. You end up giving in—it's so much easier than arguing with her.
- c. Disagreement? You're such good friends, it never happens.

5. What role does your BFF usually take on when you two are in public together?

- a. Number one fan. She tends to stay by your side and let you be the topic of conversation.
- b. Comedian to your straight man. She'll take jabs at you to get a laugh, but it's all in good fun.
- c. Partner in crime. You are not glued at the hip but do check in with each other often.

7 TO 10 POINTS

Frenemy in Disguise

You swear your BFF would never put a stiletto in your back, but guess what? She does it each time you let her walk all over you. "You might think you're being the bigger person by not getting upset when she blows off plans or takes a shot at you for laughs, but it makes your relationship unequal, and deep down, your resentment is building," says Anne-Renée Testa, PhD, author of *The Bully in Your Relationship*. Stop acquiescing to her needs, and stand up for yourself. A true best friend will appreciate the new, more assertive you.

4 TO 6 POINTS

Soul Sister

Like pearls and an LBD, you and she go together perfectly, and that isn't likely to change. "You have a supportive, complementary dynamic," says relationship expert Liz Pryor, author of *What Did I Do Wrong?* Even when things get a little tense—you compete over the same guy or have the occasional tiff—"you never bully each other or put your egos first, which keeps your friendship rock solid," says Testa.

0 TO 3 POINTS

Too-Faithful Friend

You call her your best friend, but c'mon, have you noticed that things are just a little too easy? While you might think it's great to have a pal who always nods and smiles, it hurts you in the end because she never lets you know what she's really thinking—something a genuine friend will do. Consider whether you're being too bossy or critical, preventing her from speaking out more. Your BFF should be your equal, not a sidekick, says Pryor.

COSMOPOLITAN.COM/QUIZ Log on to test yourself with more fun and insightful quizzes!

SCORING: **1.** a-0, b-1, c-2; **2.** a-2, b-0, c-1; **3.** a-2, b-1, c-0; **4.** a-1, b-2, c-0; **5.** a-0, b-2, c-1

Interpretable models - SLIM

PREDICT ARREST FOR ANY OFFENSE IF SCORE > 1

1.	*age_at_release_18_to_24*	2 points		......
2.	*prior_arrests≥5*	2 points	+	......
3.	*prior_arrest_for_misdemeanor*	1 point	+	......
4.	*no_prior_arrests*	-1 point	+	......
5.	*age_at_release≥40*	-1 point	+	......
	ADD POINTS FROM ROWS 1–5	**SCORE**	=	......

Jiaming Zeng, Berk Ustun, and Cynthia Rudin. Interpretable Classification Models for Recidivism Prediction. Accepted to Journal of the Royal Statistical Society, 2016.

What if using an interpretable model doesn't make sense?

Interpretable models don't always achieve the same level of accuracy as other models – there may be a tradeoff.

Revealing the model isn't always possible.

If we have access to run the model, we can still find out some information about how it's making decisions!

Audit options

Assumes access to appropriate input data and the ability to run the model and examine the outputs.

- Create an interpretable model of the model – use the predicted outputs as labels. Note: this is not the same model!
- Audit for direct influence - replace the feature with random noise and test the deterioration of the model.
- Audit for *indirect* influence – remove the feature and information about that feature contained in other features (e.g., proxy variables) and test the deterioration of the model.

A. Henelius, K. Puolamäki, H. Boström, L. Asker, and P. Papapetrou. A peek into the black box: exploring classifiers by randomization. Data Min Knowl Disc, 28:1503–1529, 2014.

A. Datta, S. Sen, and Y. Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In Proceedings of 37th IEEE Symposium on Security and Privacy, 2016.

P. Adler, C. Falk, S. Friedler, G. Rybeck, C. Scheidegger, B. Smith, and S. Venkatasubramanian. Auditing Black-box Models for Indirect Influence. In Proceedings of the IEEE International Conference on Data Mining (ICDM), 2016.

Direct and Indirect Influence Audits
Synthetic Data (decision tree)

<u>Synthetic Data:</u>

A: item i number

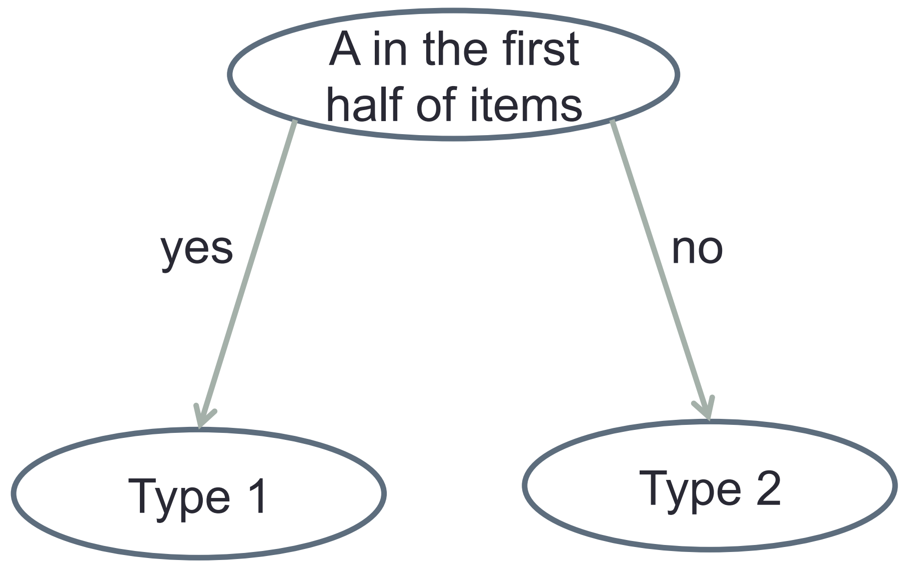
B: 2i, C: -i

Constant, Random

Outcome:

first half of items 1

second half 2



<u>Direct Influence:</u>

A: 0.5

B: 0

Constant: 0

<u>Indirect Influence:</u>

A: 0.5

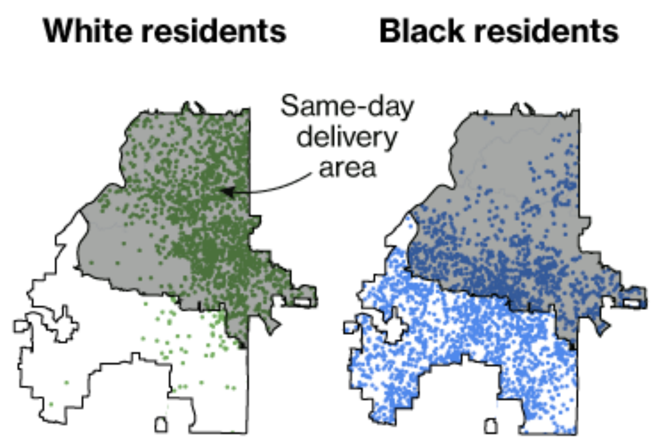
B: 0.5

Constant: 0

P. Adler, C. Falk, S. Friedler, G. Rybeck, C. Scheidegger, B. Smith, and S. Venkatasubramanian. Auditing Black-box Models for Indirect Influence. In Proceedings of the IEEE International Conference on Data Mining (ICDM), 2016.

Direct vs. Indirect Influence Audits

The northern half of Atlanta, home to 96% of the city's white residents, has same-day delivery. The southern half, where 90% of the residents are black, is excluded.



Zip code is a proxy for race.

http://www.bloomberg.com/graphics/2016-amazon-same-day/

Audit specifically for non-discrimination

Theorem:

the information content of a feature can be estimated by trying to predict it from the remaining features

If a protected feature can’t be predicted from the remaining features, then the information from that feature can’t influence the outcome of the model.

Audit: Build a classifier to try to predict the protected feature from the remaining training data. If the error is high, any trained model is non-discriminatory.

Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and Removing Disparate Impact. *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015.

Policy points

- It’s possible to create interpretable models!
- Choosing interpretable models restricts model design choice, which *may* lower accuracy.
- We can audit a model even if it’s not interpretable:
  - model the model
  - direct influence
  - indirect influence
  - goal-specific audit (e.g., non-discrimination)

THANKS!

find me at:

sorelle@cs.haverford.edu

sorelle.friedler.net