Carnegie Mellon University
School of Computer Science

Ethical and policy issues in predictive modeling

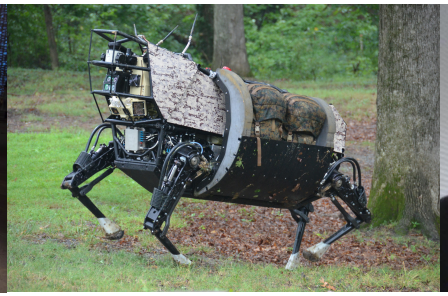
Momin M. Malik
Guest lecture given March 1, 2016, slides last revised February 13, 2017

08200/08630/19211: Ethics and Policy Issues in Computing S16
Professor James Herbsleb
Carnegie Mellon University

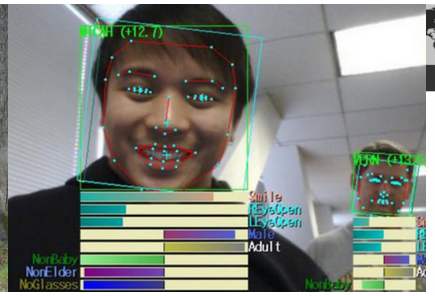
Modeling

- In many ethical dilemmas around technology, an issue exists in part because of a *model*

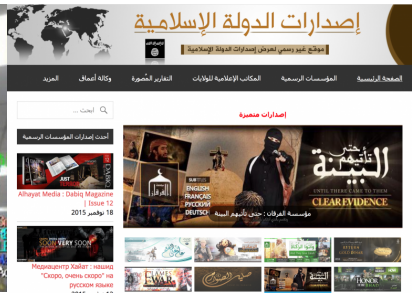
Self-driving cars: Use models of the environment



Autonomous military "killer robots"?



Facial recognition: Here, can see the model working



Recognizing terrorist websites

- *How* models work, and *how well* they work, is critical to everything downstream

Disclaimer: This is not about modeling generally, only about applications where models are used to make *predictions* that are then the basis for decisions or actions (potentially automatic)

This talk

- Previously: implications of when models work
  - Power, control (e.g., assassinations)
- What about when model predictions don't work/ fail?

Facial recognition: thwartable by face masks?

Goals

- To impress upon you that:
- "Prediction" is weird: what it means, and how it works
- With people, there often is no "truth"
- People, unlike physical systems, react to being modeled
  - They try to game, thwart, evade, and fool systems

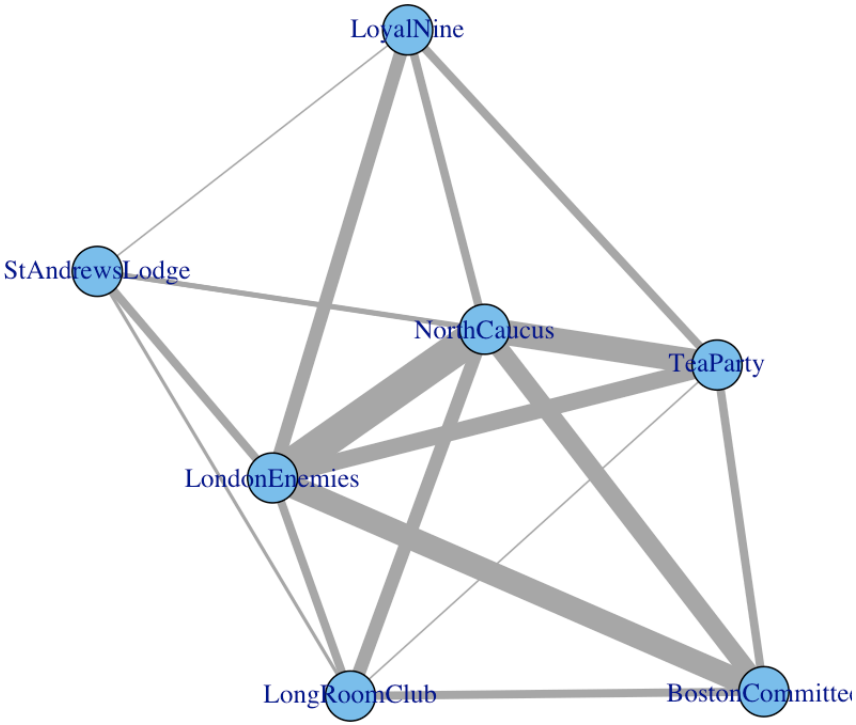
Outline

1. Intro: A paradox of prediction
2. What are models?
3. How do models work?
4. What can go wrong?

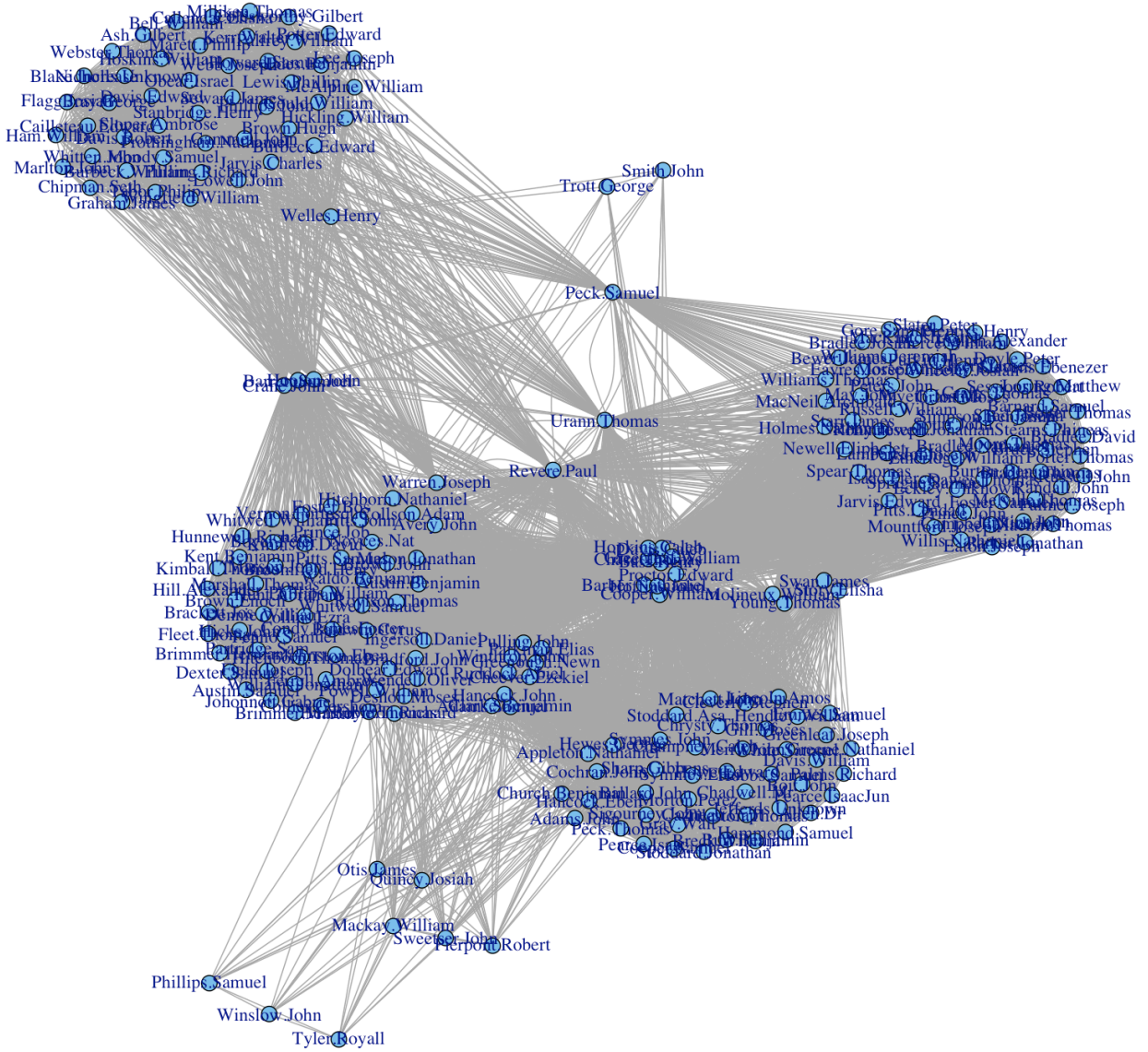
1. Intro: A paradox of prediction

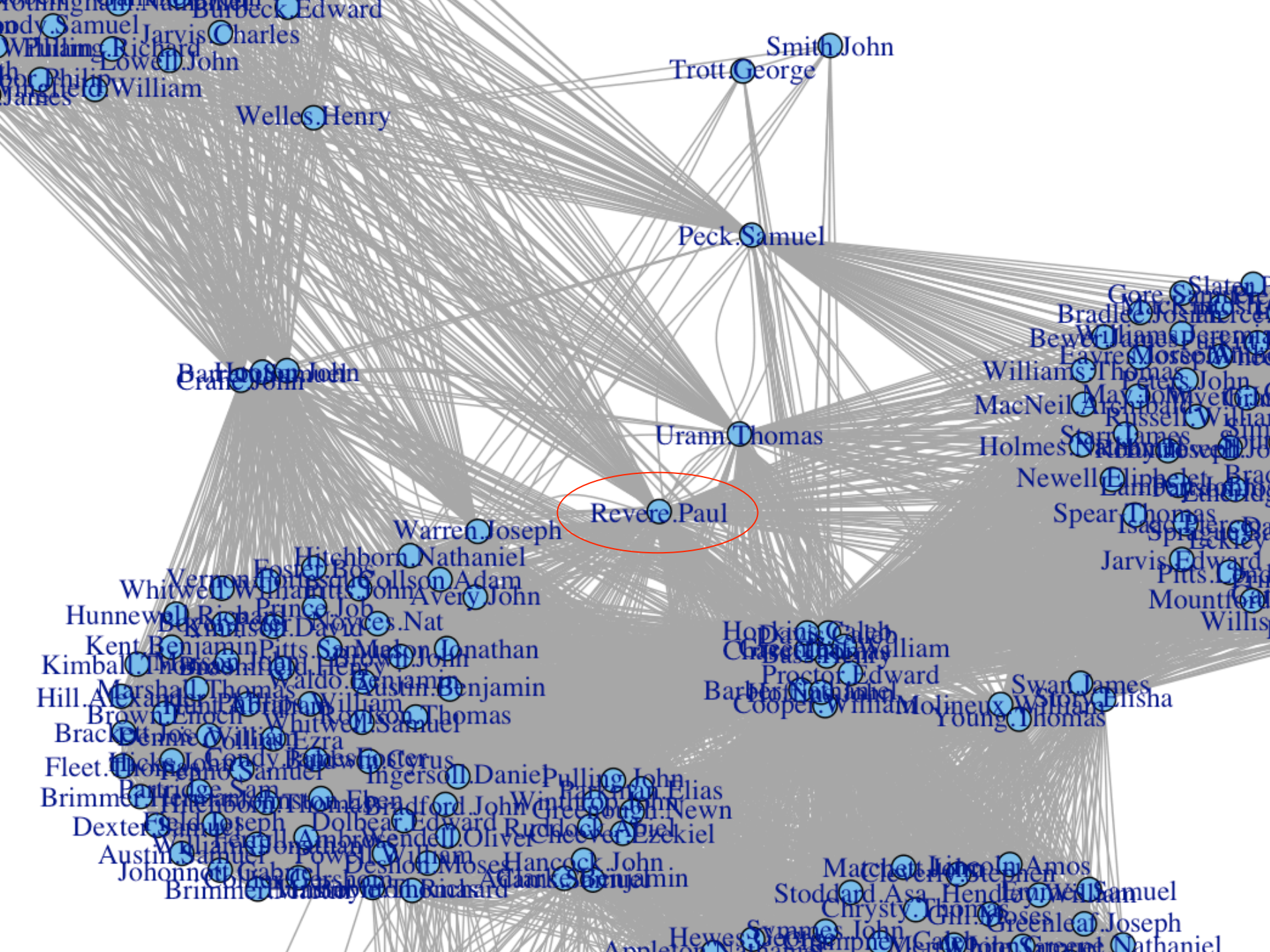
"Sniffing out Paul Revere"

- Data only about membership in seven organizations, circa 1772
- No communications data or content
- "Affiliation network" (2-mode/bipartite), it can be projected to connections between organizations or between individuals



http://kieranhealy.org/blog/archives/2013/06/09/using-metadata-to-find-paul-revere/


Smith John
Trott George
Peck Samuel
Urann Thomas
Revere Paul
Warren Joseph
Otis James
Quincy Josiah
Mackay William
Sweetser John
Pierpont Robert
Phillips Samuel
Winslow John
Tyler Royall
Welles Henry


Smith.John
Trott.George
Welles.Henry
Peck.Samuel
Urann.Thomas
Revere.Paul
Warren.Joseph
Hitchborn.Nathaniel
Avery.John
Prince.Job
Pitts.John
Hopkins.Caleb
Proctor.Edward
Swan.James
Molineux.William
Young.Thomas
Ingersoll.Daniel
Pulling.John
Hancock.John
Matchett.Lincoln
Stoddard.Asa
Greenleaf.Joseph

Centrality: A model of importance

Betweenness	
Revere, Paul	3839
Urann, Thomas	2185
Warren, Joseph	1817
Peck, Samuel	1150
Barber, Nathaniel	931
Cooper, William	931
Hoffins, John	931
Bass, Henry	852
Chase, Thomas	852
Davis, Caleb	852

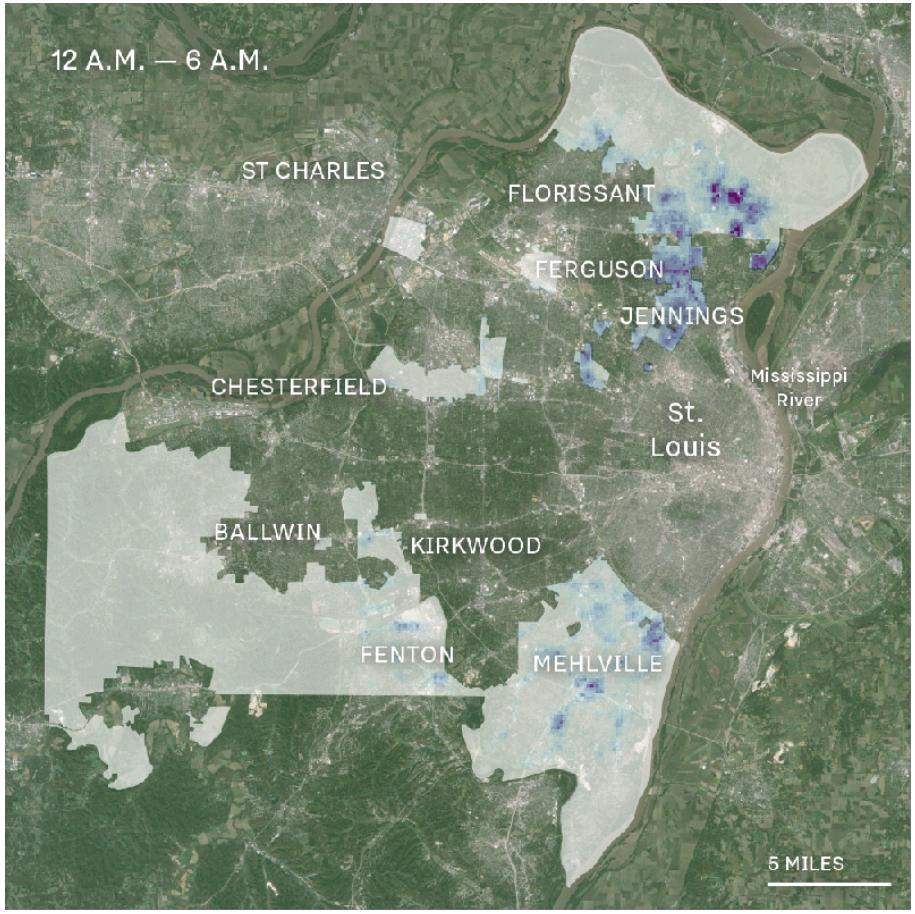
Eigenvector	
Barber, Nathaniel	1.00
Hoffins, John	1.00
Cooper, William	1.00
Revere, Paul	0.99
Bass, Henry	0.95
Davis, Caleb	0.95
Chase, Thomas	0.95
Greenleaf, William	0.95
Hopkins, Caleb	0.95
Proctor, Edward	0.90

Bonacich Power	
Revere, Paul	-1.51
Urann, Thomas	-1.44
Warren, Joseph	-1.42
Proctor, Edward	-1.40
Barber, Nathaniel	-1.36
Hoffins, John	-1.36
Cooper, William	-1.36
Peck, Samuel	-1.33
Davis, Caleb	-1.31
Chase, Thomas	-1.31

It worked! ...Or did it? Who are all those other people?

The paradox: We only know that model predictions are accurate when we already know the answer

And if we already know the answer, why do we need model predictions?



"[P]olice officers are often unsurprised by the locations of the boxes — police in Lincoln, Nebraska, started experimenting with the software in 2014 but have found it mostly tells them what they already know. 'When I look at the HunchLab maps,' said former police chief Tom Casady, 'I say, "Yep, it got that right!"'"

http://www.theverge.com/2016/2/3/10895804/st-louis-police-hunchlab-predictive-policing-marshall-project

Conversely…

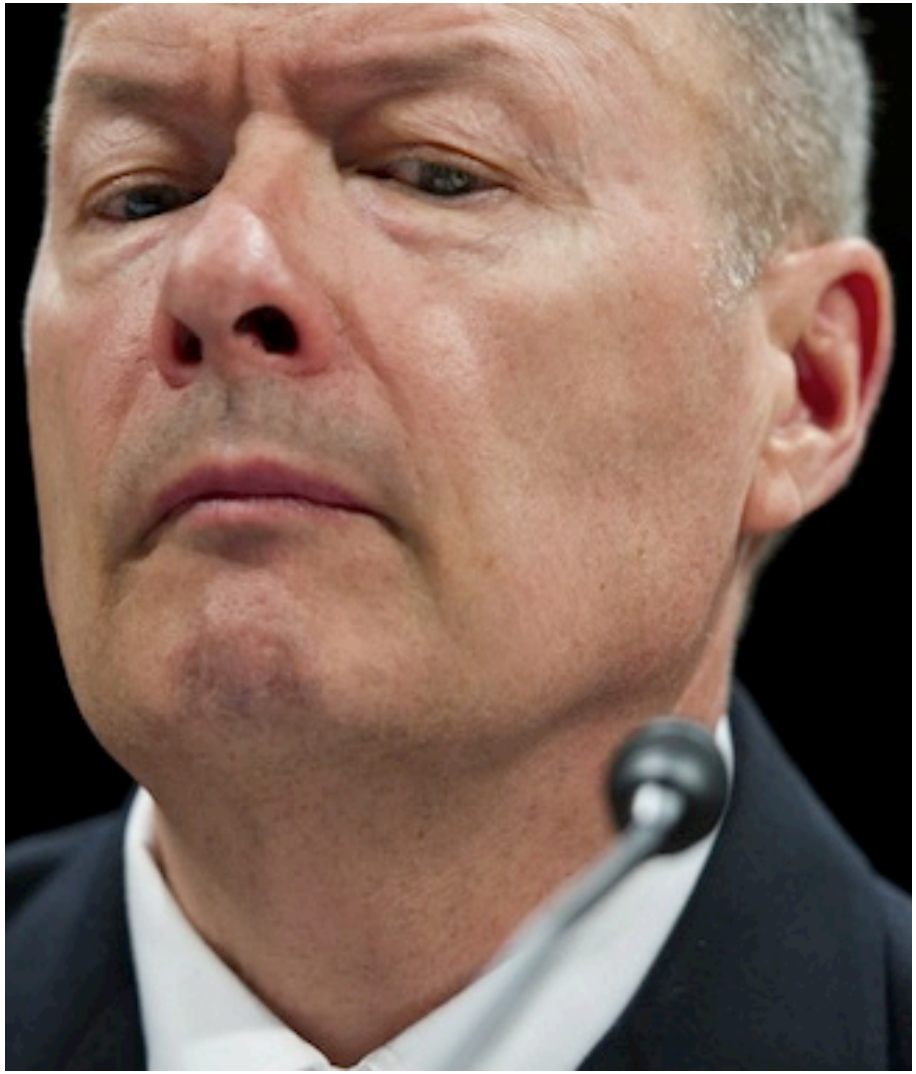
- If we get an unexpected result, did we make a mistake in modeling? Or have we discovered something?
- What do we do if we can't tell? And if we can't confirm through other means?
- (Paul Revere: Do you recognize even a single other name? Were these important people left out of the main historical narrative—or were these other people unimportant, and the model mistaken? What do we trust for the "truth" of the matter?)

Betweenness		**Eigenvector**		**Bonacich Power**	
Revere, Paul	3839	Barber, Nathaniel	1.00	Revere, Paul	-1.51
Urann, Thomas	2185	Hoffins, John	1.00	Urann, Thomas	-1.44
Warren, Joseph	1817	Cooper, William	1.00	Warren, Joseph	-1.42

Faith in data/models

Lt. Gen. Keith Alexander, 16th director of the NSA

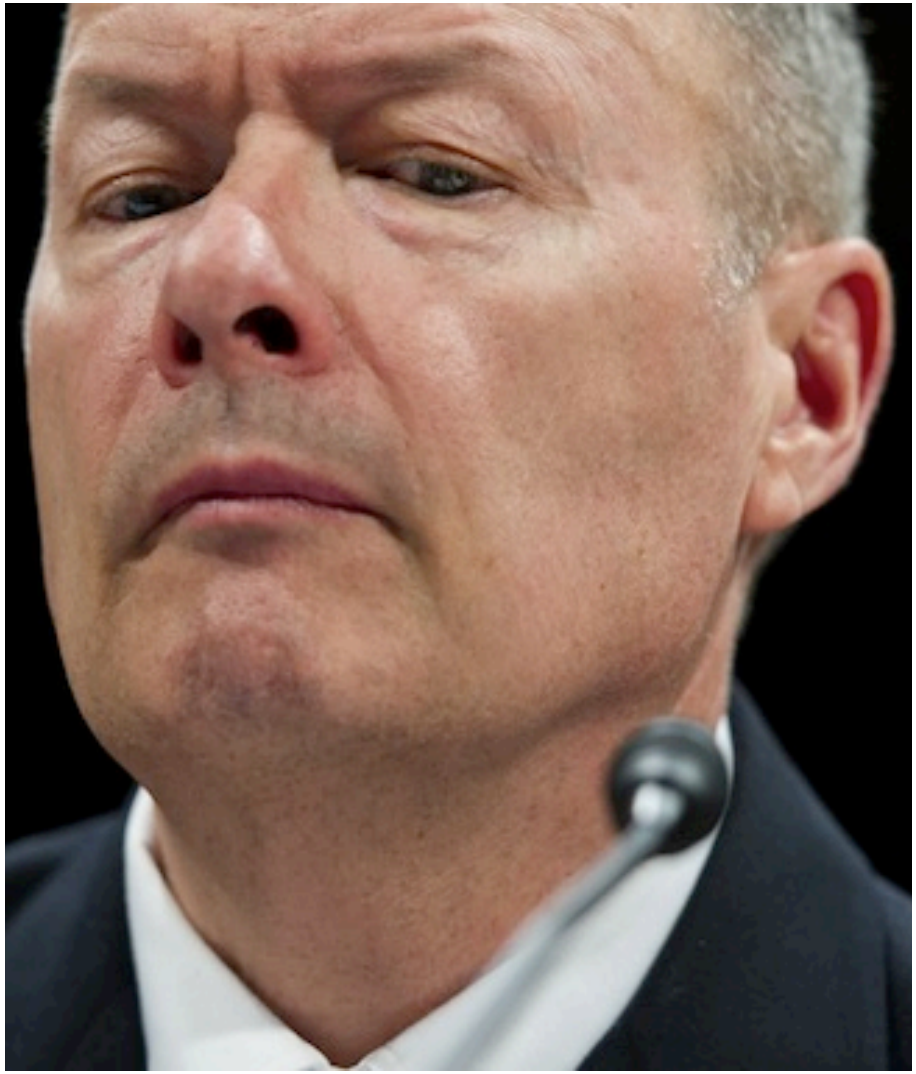
"He had all these diagrams showing how this guy was connected to that guy and to that guy," says a former NSA official who heard Alexander give briefings on the floor of the Information Dominance Center. "Some of my colleagues and I were skeptical. Later, we had a chance to review the information. It turns out that all [that] those guys were connected to were pizza shops."



http://foreignpolicy.com/2013/09/09/the-cowboy-of-the-nsa/

Lt. Gen. Keith Alexander, 16th director of the NSA

A retired military officer who worked with Alexander also describes a "massive network chart" that was purportedly about al Qaeda and its connections in Afghanistan. Upon closer examination, the retired officer says, "We found there was no data behind the links. No verifiable sources. We later found out that a quarter of the guys named on the chart had already been killed in Afghanistan."

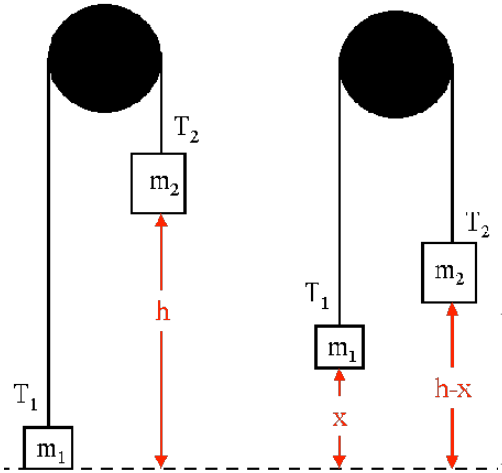


http://foreignpolicy.com/2013/09/09/the-cowboy-of-the-nsa/

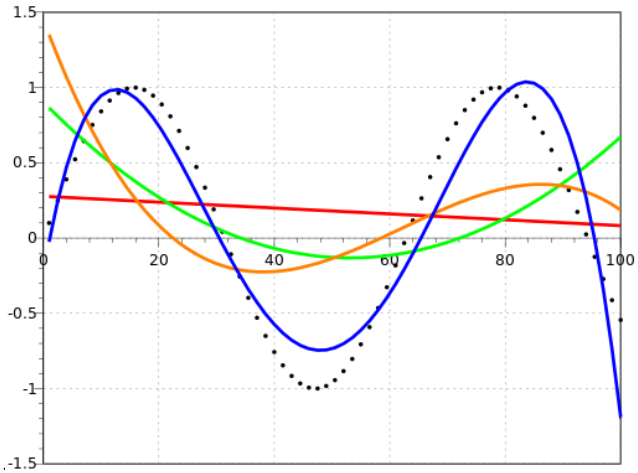
2. What are models?

What are models?

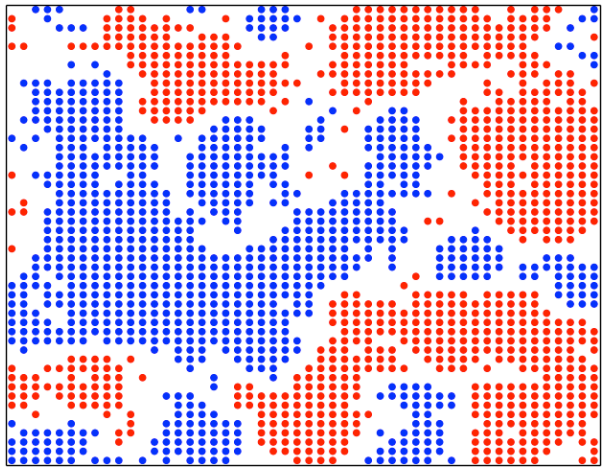
- An abstraction of what we think are the essential elements of a system
- We are dealing with *mathematical* models: and specifically, *data models* that take in numbers
- *Statistical* models can do inference, and use *past data* to try and *predict* out-of-sample (future) results
- Not all data models are statistical: e.g., deanonymization can be done by matching and heuristics. But usually, very quickly as a problem (a target phenomenon with a pattern and variability) scales, statistical approaches outperform heuristics, and are far more manageable (e.g., compared to hand-coded rules).
- (Machine Learning is all data modeling, and almost entirely statistical.)



"Mathematical" models, often deterministic, as in physics (other than thermo/statistical mechanics)



Statistical models. Most basic are "curve fitting" problems; treat data as noisy realizations of a trend, and "smooth" data to infer/"learn" trend.



Simulation models. Show a simple mechanism can account for "complex" behavior; but they are only a *conceptual demonstration*, not a proof or a way to predict outcomes in a specific system given data.

Data, Models, Algorithms, Software

- People conflate these, so I want to make the distinctions clear. E.g., Facebook newsfeed is an algorithm, but it implements a model.
- **Data/numbers/measurements** (also models! Choosing what to measure is tied up with models. E.g. SATs: the "measurement" of the raw scores is towards a model, and the number of an SAT score is also the product of a model)
- **Models** are mathematical statements with data. We focus here.
- **Algorithms** are abstract sets of instructions which describe how to *implement* models. E.g., when you have data, what operations do you take in which order to get a prediction?
- **Software** is the specific code, written in a specific programming language, that concretely implements algorithms
- Software can also implement the processes around the model
  - E.g., collect data on user behavior and plug them into a model to decide which ads to show, and then showing those ads

(Literature under "software studies" or "algorithmic governance" or "critical algorithm studies" is often about models—or at least I would say that the core issues they are address really have to do with modeling, and not algorithms or software *per se*.)

http://socialmediacollective.org/reading-lists/critical-algorithm-studies/

Warning 1: Machine Learning ≠ AI!

- Artificial Intelligence started off trying to model intelligence from first principles
- That reached a dead end… and the field switched to using statistical methods under the fanciful name "machine learning"
- "Learning" is a metaphor. The systems don't reason, they just optimize. It may look like learning, but it's far from intelligent in any real sense or anywhere close to the original goals of Artificial Intelligence.
- (Note: not everybody agrees, and there are strong opinions. Gets into philosophical issues, e.g., Turing Test vs. "Chinese Room.")

Don't just take it from me...

Video: Prof. Alex Smola, from Spring 2015's PhD ML intro course, talking about Deep Networks (arguably the most "human-like" machine learning):

https://www.youtube.com/watch?v=xZzZb7wZ6eE&t=20m28s

(start at 20m28s, end at 21m02s)

"It's not that the deep network understands or wants to learn, or some other nonsense that people tell you when they want to advertise and push deep networks: it's basically just math. You set up a suitable objective function and just optimize. And whether the network 'wants' to learn or not depends very much on whether you know how to do it, and not whether it wants it. The deep network wants nothing. It's just math. Sorry for saying that, but I've seen too many interviews and blog posts about such stuff."

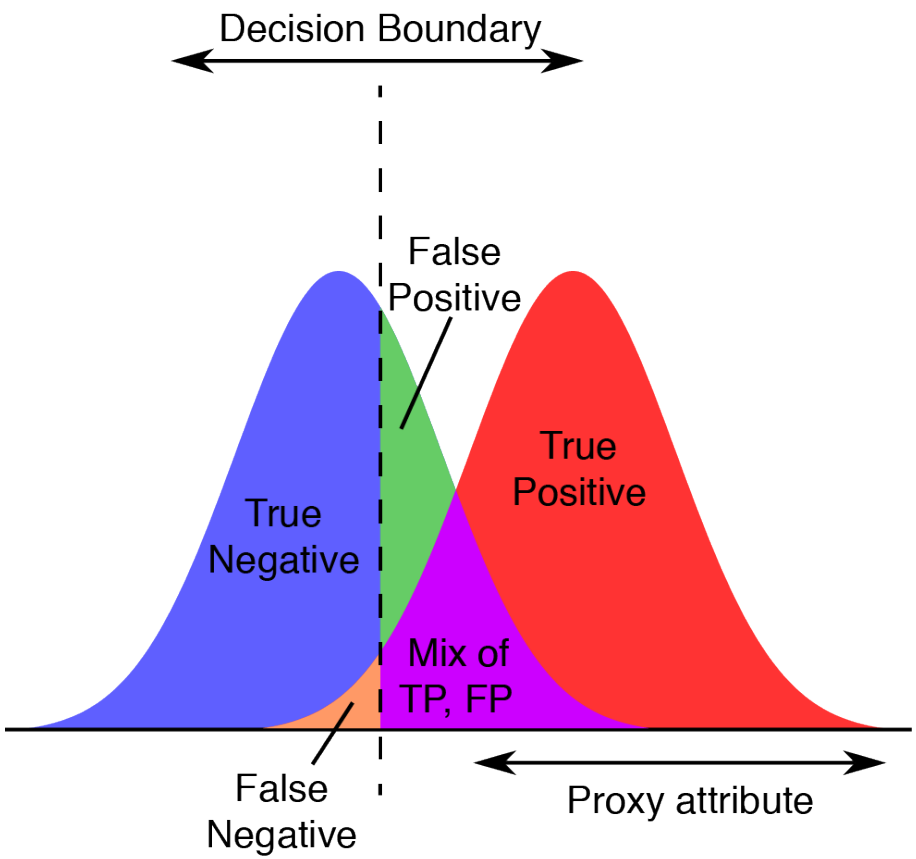
3. How do models work?

How do models work?

- With humans, almost always, the target is subjective, and/or difficult to observe/measure
- Need some initial way to measure (need some "ground truth" to train models!)
- Find *proxies*. These are often differences in distributions along various measurable attributes
- (Note: "ground truth" is an actual technical term, but "proxy" is just my choice. But "signaling theory," from evolutionary biology, is a relevant analog.)

Happiness: something we want to measure, but that doesn't have a physical reality or even a universal mental one

Ground truth + proxy = Model



- You want to find a *decision boundary*, or decision criterion, by which to place new observations into a class. You can shift that boundary to prioritize certain results. Everything is a variation on this basic theme. The advanced stuff is in how to chop and shuffle, splice and cross-reference data to find proxies (doing this too much causes statistical problems as well, but we won't go into them.)
- Note: most of machine learning is *classification*, trying to place an observation into one of two (or more) groups. Statistics does more *regression*, trying to predict the value of a continuous output (e.g., predict income), but ML does this as well.

The proxy is never the thing itself.
So long as this is true, we will get errors.

What is prediction?

Warning 2: Prediction means something different in stats/ML

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

Gayo-Avello 2012, "I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper"

- "Predicted values" is a technical term synonymous with "fitted values," so in some sense Gayo-Avello is being unfair
- But when the public reads press releases about scientists successfully "predicting X," they don't know that
- **Read "We can predict X" instead as "We found a model that fits well"**
- Fitting well is still an accomplishment, but it's quite different from actually being able to tell the future

Prediction is weird

- Correlation ⇏ Causation, and…
- **Prediction ⇏ Explanation**
- Counterintuitively, a "false" model may predict better than a "true" one
  - So, just because a model fits well doesn't mean we should take it as reflecting real associations!
- **We can successfully predict without knowing *how* or *why* something is a proxy**
  - "Algorithmic" or "black box" models working on superficial relationships often perform far, far better than attempts to model human behavior from first principles. This is something really weird.

Shmueli 2010, "To Explain or to Predict?" https://projecteuclid.org/euclid.ss/1294167961
Breiman 2001, "Statistical Modeling: The Two Cultures." http://projecteuclid.org/euclid.ss/1009213726
Gayo-Avello 2012, "I Wanted to Predict Elections with Twitter and All I Got was this Lousy Paper." http://arxiv.org/abs/1204.6441
Halevy, Norvig, and Pereira, "The Unreasonable Effectiveness of Data." http://doi.ieeecomputersociety.org/10.1109/MIS.2009.36

Do models work?

AI Magazine Volume 18 Number 3 (1997) (© AAAI)

Does Machine Learning Really Work?

Tom M. Mitchell

"Yes."

(By this I broadly mean predictive statistical models. Examples of things that we know from personal experience work pretty well: all large-scale recommender systems. Internet search. Your location as determined by your phone. Speech recognition. Google translate.)

Or does it?

"performance claims can easily be taken to be an assertion about the performance of the system under general conditions. In fact, we suspect that most authors of these works had similar assumptions in mind (author's note: we did!)."

Cohen & Ruths 2013, "Classifying Political Orientation on Twitter: It's Not Easy!"

The real test of whether a model works is testing it on out-of-sample data. **For physical systems, this distinction doesn't matter. For human beings, it really does.**

Why *would* a model work?

"The miracle of the appropriateness of the language of mathematics for the formulation of the laws of physics is a wonderful gift which we neither understand nor deserve."

Wigner 1960, "The Unreasonable Effectiveness of Mathematics in the Natural Science"

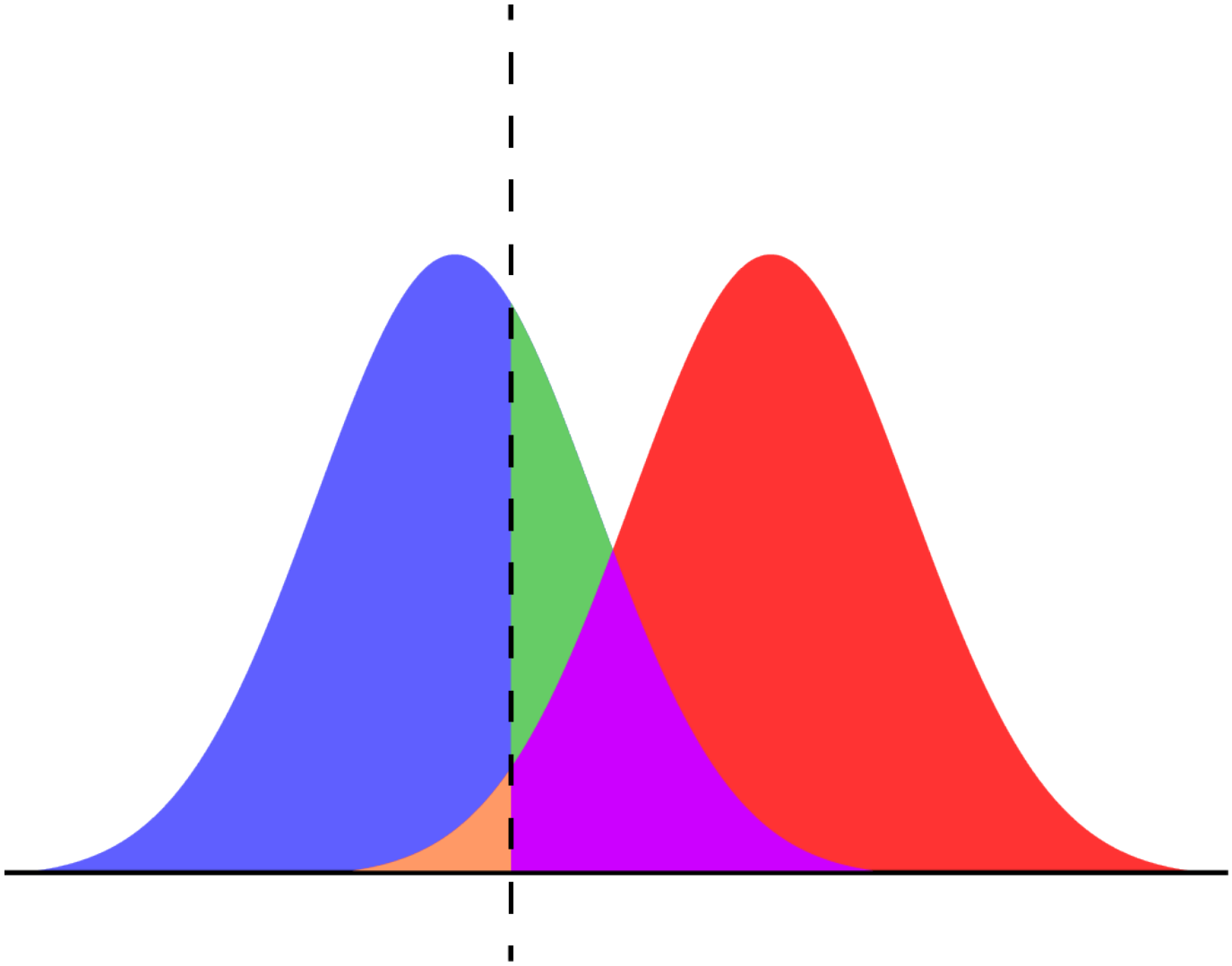
"There is general agreement that, in [social science], all models in use are wrong – not merely falsifiable, but actually false. With enough data – and often only a fairly moderate amount – any analyst could reject any model now in use to any desired level of confidence."

Gelman & Shalizi 2012, "Philosophy and the Practice of Bayesian Statistics"

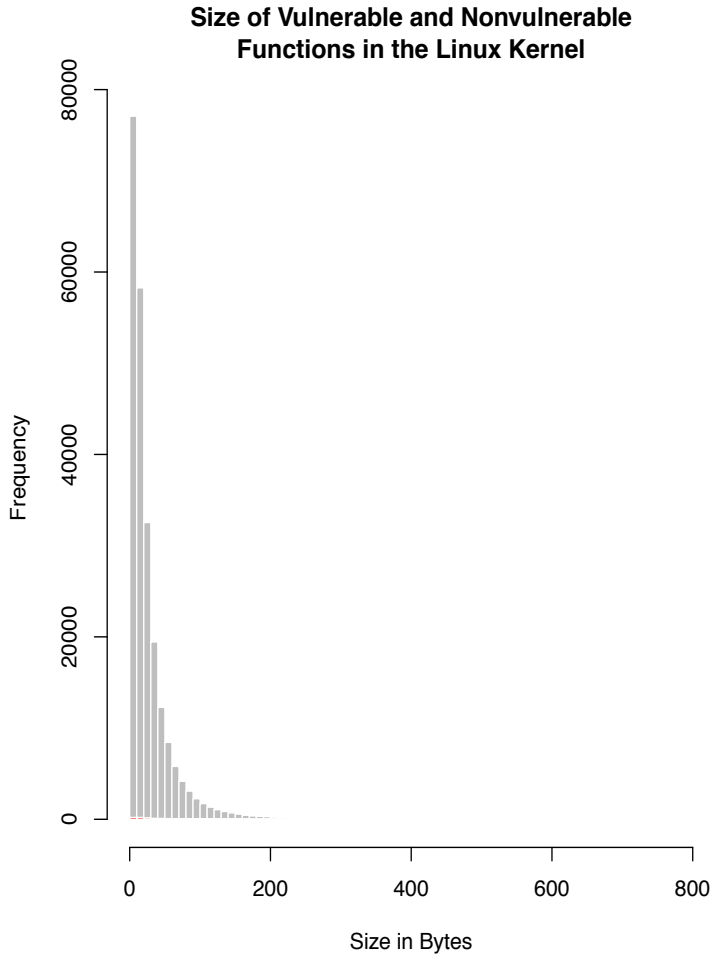
Short answer: philosophically, we have no idea.

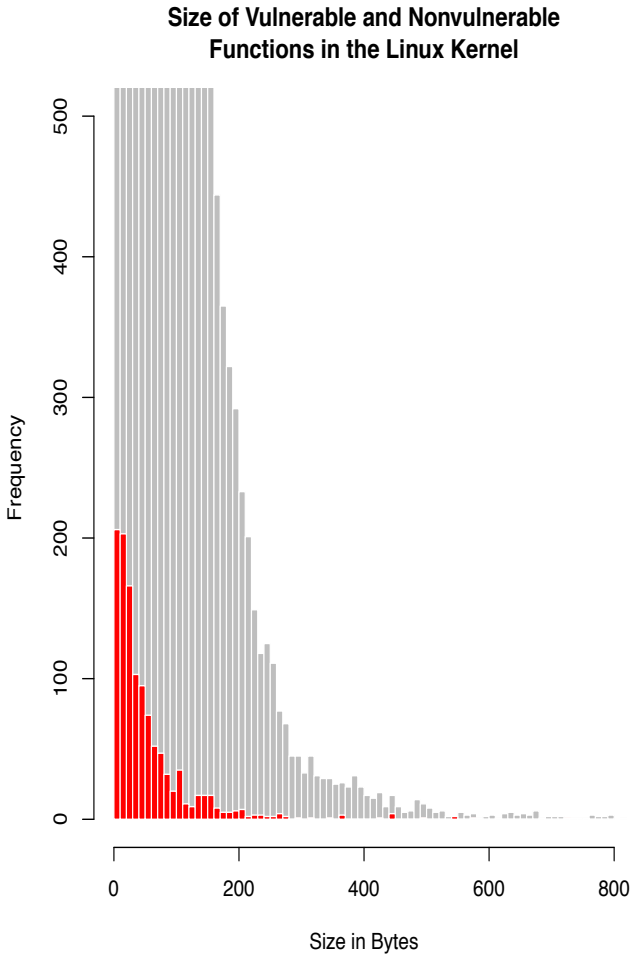
What can go wrong?

Nice picture, but...



...what if it looks like this?





Example from my research. Here, there is simply no signal, this variable has no information about whether a function in the Linux Kernel is likely to be vulnerable. No transformation can change that (maybe combining with other variables will give something.)

What goes wrong in *making* models

- Can't get a good fit: no "signal"
  - "Accuracy paradox"
  - Maybe we're not good enough modelers
  - Maybe the phenomenon has too much variability
  - Maybe we're not measuring the right thing
- No or not enough "ground truth" means we have nothing to train/fit on (can do "clustering", or "unsupervised learning" but no way to verify: the example of centrality scores from the beginning is actually not statistical, but descriptive statistics and sometimes rankings are also cases of no ground truth)
- Practical problems of scale: the model is not tractable
  - Takes too long to solve for needed size
  - Needs too much "training" data (e.g., hand-labeling)

https://theintercept.com/2016/01/20/the-white-house-asked-social-media-companies-to-look-for-terrorists-heres-why-theyd-fail/

What goes wrong in *application*

- Mistakes in model-building
- Out-of-sample data is different
  - The training data didn't represent the phenomenon or population
  - We "overfit," even with attempts to not do so (this is incredibly, incredibly profound, and appears everywhere: e.g., [technical point] if we try too many models on training data, we get a distribution over models and *this* overfits even with cross-validation)
  - Without knowing *why* a model works, we don't know when it might stop working. Where knowing "causation," or more precisely the "data-generating process," is critical.
- **Again: there will always be mistakes, since proxies are still never the thing itself**

Using models

Using models

- Given what we've covered above, how should we be using models?
- What can go wrong?
- What has gone wrong?

Key questions

- If we had perfect models, what would the ethical and policy questions be?
  - Similar to the hypothetical of perfect DRM, but more profound
  - Free will
  - Power, domination and control
  - Cases where *not* using the models is unethical (e.g., saving soldier's lives)
- Still: what should we be trying to model? When do we institutionalize models?
  - Nicholas Carr: we end up "optimizing the status quo" rather than challenging it

Key questions

- How do we deal with models always having errors? …How do we deal with not knowing when we have misclassified?
  - Keywords:
  - Accountability
  - Transparency
  - Due process

Yaseen Kadura

- US medical student, online anti-Gaddafi activist
- Spent six months in Libya helping run aid convoys and being a fixer for international journalists
- Four months after return, "flagged by DHS' Automated Targeting System, or ATS, a program originally created to identify potentially risky container cargo but later expanded to assign risk ratings to individual people."
- Years of humiliating searches and interrogations at US/Canada border, banned from flying
- Spent three years trying to clear his name without success

https://theintercept.com/2016/01/21/how-a-young-american-escaped-the-no-fly-list/

Second-order effects

- Gaming the system: proxies can be manipulated
- When we make decisions based on models, we create incentives to manipulate
- (E.g., as soon as people find a way to model computer security, hackers find a way around it.)

SATs: why is it possible to study for them? If they did what they are supposed to do, studying would not make a difference. But these tests are a proxy, with a lot at stake.

Vietnam War: Robert McNamara and his obsession with "body counts" as a metric of success without verifying. Eventually, people started lying to make him happy.

Adrian Schoolcraft, whistleblower for the NYPD making arrests to meet quotas, since that's what their incentives were built around

http://fivethirtyeight.com/features/how-the-nypd-abused-citizens-in-the-name-of-data-and-how-one-cop-exposed-it-all/

Second-order effects

- Maybe models work because of disparities
- Feedback loops: If decisions based on the model change the measures, models become self-reinforcing

Existing example: credit scores. So long as having money is a predictor of ability to pay back, the system self-reinforces

Predictive policing: if it is trained on previous data, which includes racial bias by police, then police will focus on the same groups, catching more of crime there and missing others

- When we use models in this way, what sort of society does it create?
  - Disparate impact
  - Fairness

What will we try to do?

Models, with their allure of objectivity and mathematical rigor, can be a distraction from what may be the real issue.

http://www.theverge.com/2016/2/3/10895804/st-louis-police-hunchlab-predictive-policing-marshall-project

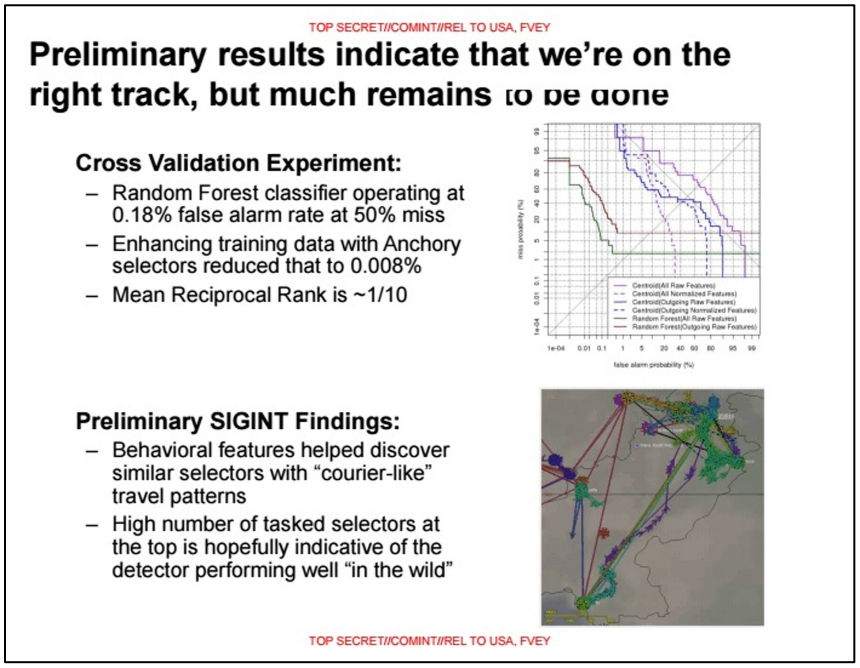
"We kill people based on metadata."

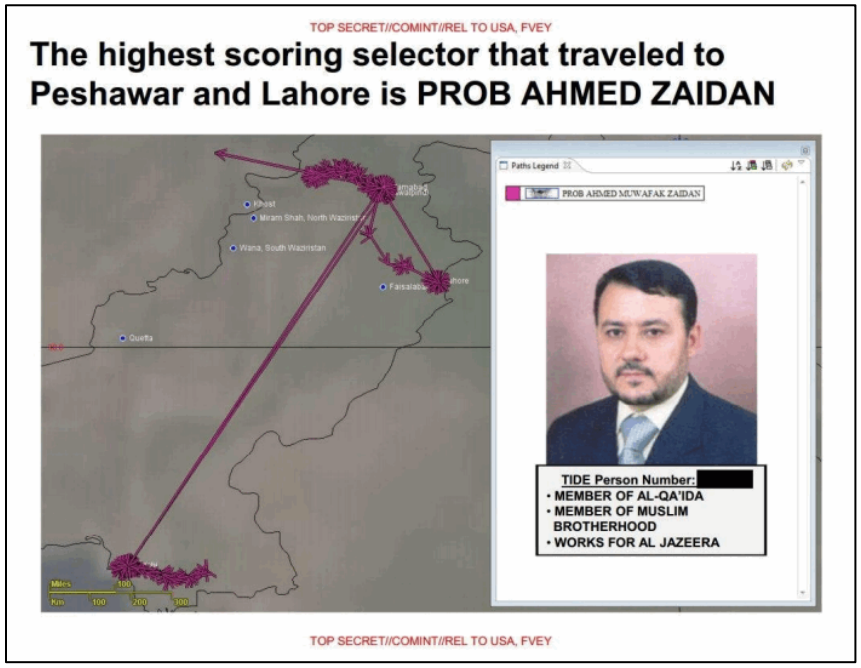
General Michael Hayden
(former director, NSA and CIA)

http://www.nybooks.com/daily/2014/05/10/we-kill-people-based-metadata/

...should we kill people based on *models*, trained on metadata?

Revere, revisited: Find a ~~terrorist~~ journalist





(The journalist in question denies being a member of Al Qaeda or the Muslim Brotherhood. So this "confirmation" points to shoddy, or possibly politically biased, human analysis: but as the Guardian article points out, it's otherwise quite reasonable to identify a journalist as having "courier-like patterns." And, also pointed out there, the machine learning in question hasn't actually been used in the field, let alone as a basis for deciding who to kill [or automatically killing]... yet.)

http://arstechnica.co.uk/security/2016/02/the-nsas-skynet-program-may-be-killing-thousands-of-innocent-people/
https://www.theguardian.com/science/the-lay-scientist/2016/feb/18/has-a-rampaging-ai-algorithm-really-killed-thousands-in-pakistan

(Sidenote: Automatic skepticism for models only when they challenge power)

- Consumer Financial Protection Bureau, "Using publicly available information to proxy for unidentified race and ethnicity: A methodology and assessment," 2014
  - A model for discrimination by indirect auto lenders
  - Preferred overestimating discrimination to underestimating it
- This model was criticized as biased by the *American Banker*, who called it "sophisticated guesswork"
- But all predictive models are sophisticated guesswork
- Lesson: When models support structures of power, their value/objectivity/rigor is treated as a *fait accompli.* But when they challenge power, they are not treated the same way.

http://www.americanbanker.com/news/law-regulation/cfpb-overestimates-potential-discrimination-documents-show-1076742-1.html

Take-aways

- "Prediction" doesn't actually mean magic powers
- Modeling physical systems (which works amazingly well) is very different from modeling humans (which doesn't work so well)
- Humans can—and do—react to modeling
- There is a danger in the power of models that work—but also a danger in assuming models work and not knowing (and of models being used by power to avoid accountability, regardless of whether they work or not)

Discussion. What should we model? How should we act on model results?

Real Examples:

- Movie studios predicting which scripts are likely to succeed
- Banks predicting who is likely to default on a loan
  - …based on Facebook data
- Governments predicting who will become radicalized
- Governments predicting who is a terrorist
- Judges predicting who won't (re)offend when out on bail
- Google predicting to whom to display ads for high-paying executive jobs
- Employers predicting which employees will get sick
- Employers predicting which hires will be the best
- Police predicting where crime will be
- Police predicting who will commit crimes
- Governments predicting which cops are will commit misconduct

http://www.nytimes.com/2013/05/06/business/media/solving-equation-of-a-hit-film-script-with-data.html
http://venturebeat.com/2015/08/04/facebook-patents-technology-to-help-lenders-discriminate-against-borrowers-based-on-social-connections/
https://www.washingtonpost.com/news/the-intersect/wp/2015/07/06/googles-algorithm-shows-prestigious-job-ads-to-men-but-not-to-women-heres-why-that-should-worry-you/
http://www.wsj.com/articles/bosses-harness-big-data-to-predict-which-workers-might-get-sick-1455664940
http://www.ft.com/cms/s/2/e3561cd0-dd11-11e3-8546-00144feabdc0.html
http://www.bloomberg.com/news/articles/2015-11-17/machines-are-better-than-humans-at-hiring-top-employees
http://fivethirtyeight.com/features/how-to-predict-which-chicago-cops-will-commit-misconduct/