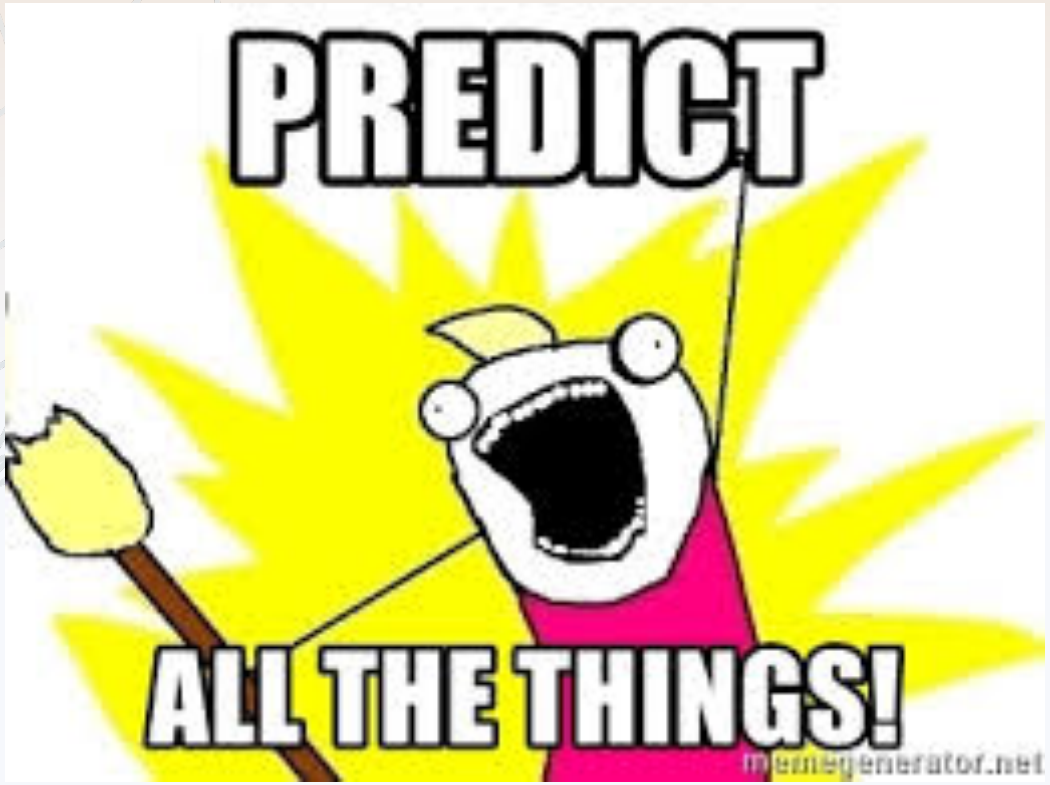
Machine Learning Beyond Prediction and the Role of Epidemiology

Eric Lofgren MSPH, PhD
Assistant Professor
Paul G. Allen School for Global Animal Health
Washington State University

Most of Machine Learning



What Role Does Epidemiology Play?

- What other uses are there for ML if you're an epidemiologist?
- What role is there for Epidemiology in the era of Big Data, Precision Medicine, etc.?
- Why does your field even exist now that we have neural networks?

Extensions of ML in Epidemiology

- Missing Data
- Anonymization and Synthetic Data
- Propensity Scores/IPTW
- Model Emulation
- There are definitely others, these are just the ones that interest me
- Many of these aren't "beyond prediction" in the sense that they aren't doing prediction, but where the predictions are an intermediate product toward some other aim

Missing Data

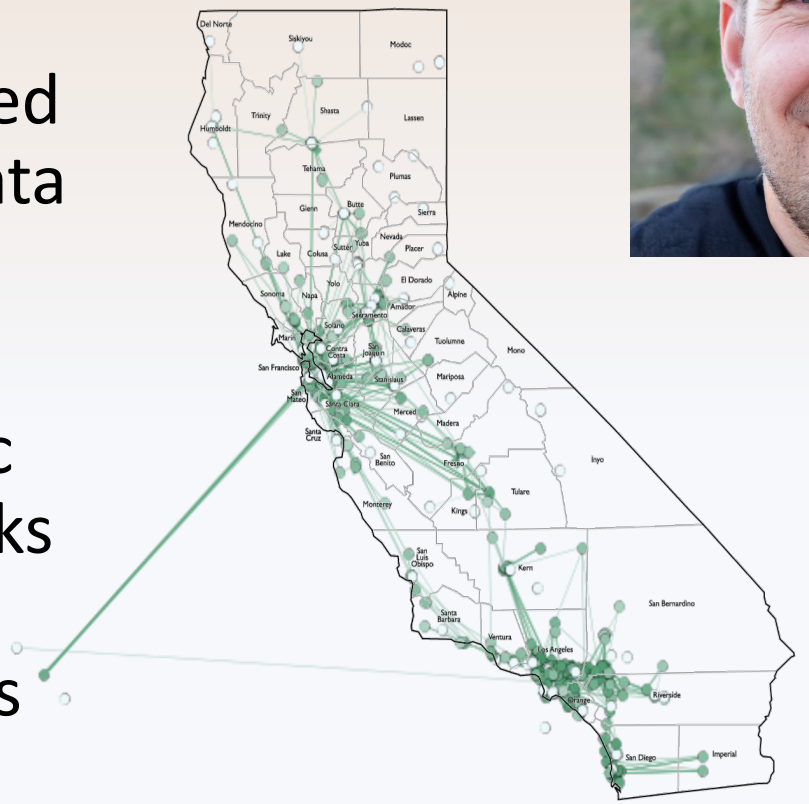
- Multiple imputation is inherently predicting, and then simulating the values of, missing data
- The normal models to predict missing data are fairly simple models assuming things are multivariate normal distributions
- There's no reason, for especially challenging problems in this space, that machine learning models can't be used to predict missing values

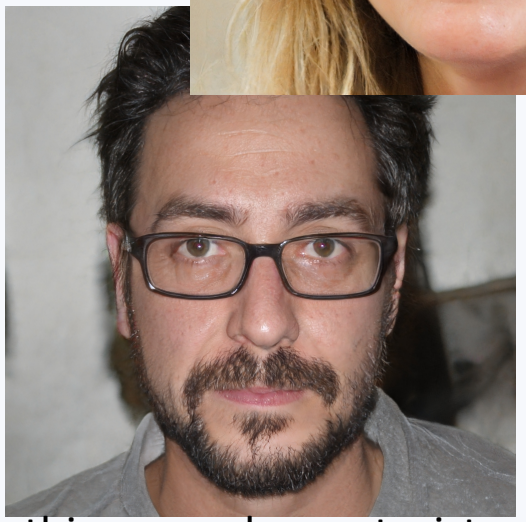
Anonymization

- There is a push in science generally to make more data freely available and accessible
- This often runs into problems in epidemiology for reasons of privacy, etc.
- Multiple imputation has been used for data anonymization – deliberately delete values, fill them in with predicted synthetic values until your dataset is realistic but not real
- You could potentially even use this, with some “seed” values inserted from the distribution of the variables in your data set, create entirely synthetic data
- There’s a lot of tricks in making sure things are actually anonymized – this is an interesting case of wanting good, but not great, prediction

Synthetic Data

- We can also use ML-based models to create new data whole-cloth
- Currently working on a project to create realistic hospital transfer networks via the same type of machine learning models used to produce fake images





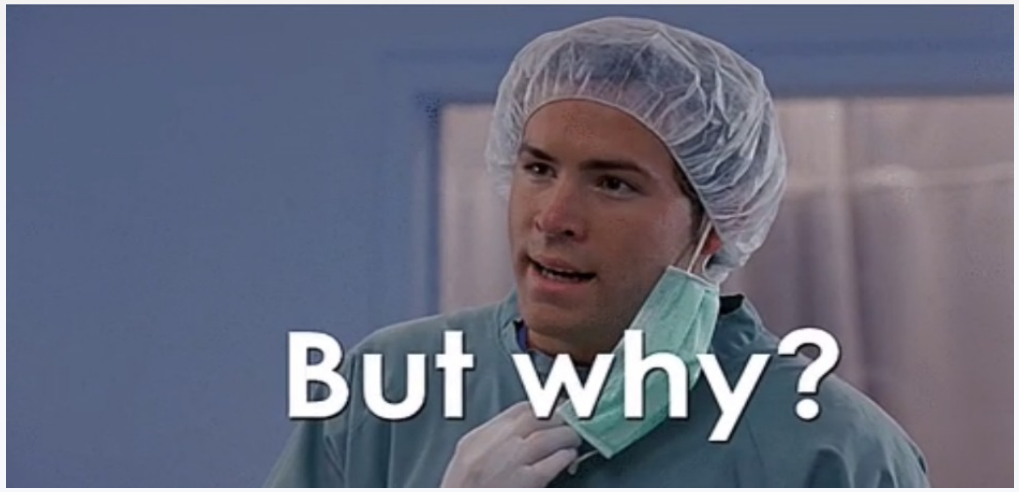
thispersondoesnotexist.com

Propensity Scores/IPTW

- The use of propensity scores or inverse probability of treatment weights involves the construction of a model predicting p(Exposure |**Z**)
- *In theory* these aren't purely predictive models – they should have the same causal backing as other adjustment methods
- But…machine learning models seem to be working fairly well in this space, even without that causal structure to them

Why?

- I think there's room here for figuring out why that is
- Personal theory: We don't collect random data. Most epidemiology studies have a sort of weak underlying DAG that makes even pure prediction models quasi-causal
  - I have nothing to back this up
  - It might also just be that things like conditioning on colliders + the use of machine learning for these approaches is sufficiently rare that no one has been bit hard yet

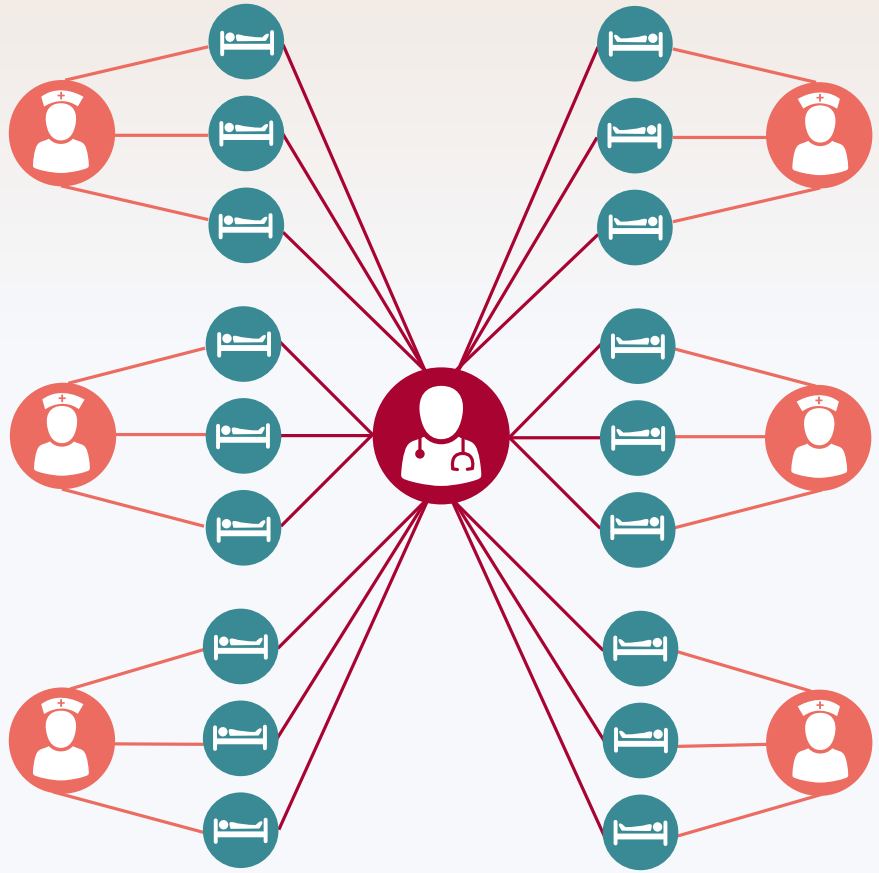


Model Emulation

- The bulk of my research is mathematical modeling
- Some models, especially complex agent-based models, or models of very large populations, are quite slow
- This is an even bigger deal for people modeling the climate, nuclear explosions, etc.
- Trying to cover every parameter combination, or generate new forecasts can be extremely time consuming
- Can we run a model to generate data, use a machine learning model to *predict* the outcomes of that model, and then run the ML model instead of the more expensive full model?

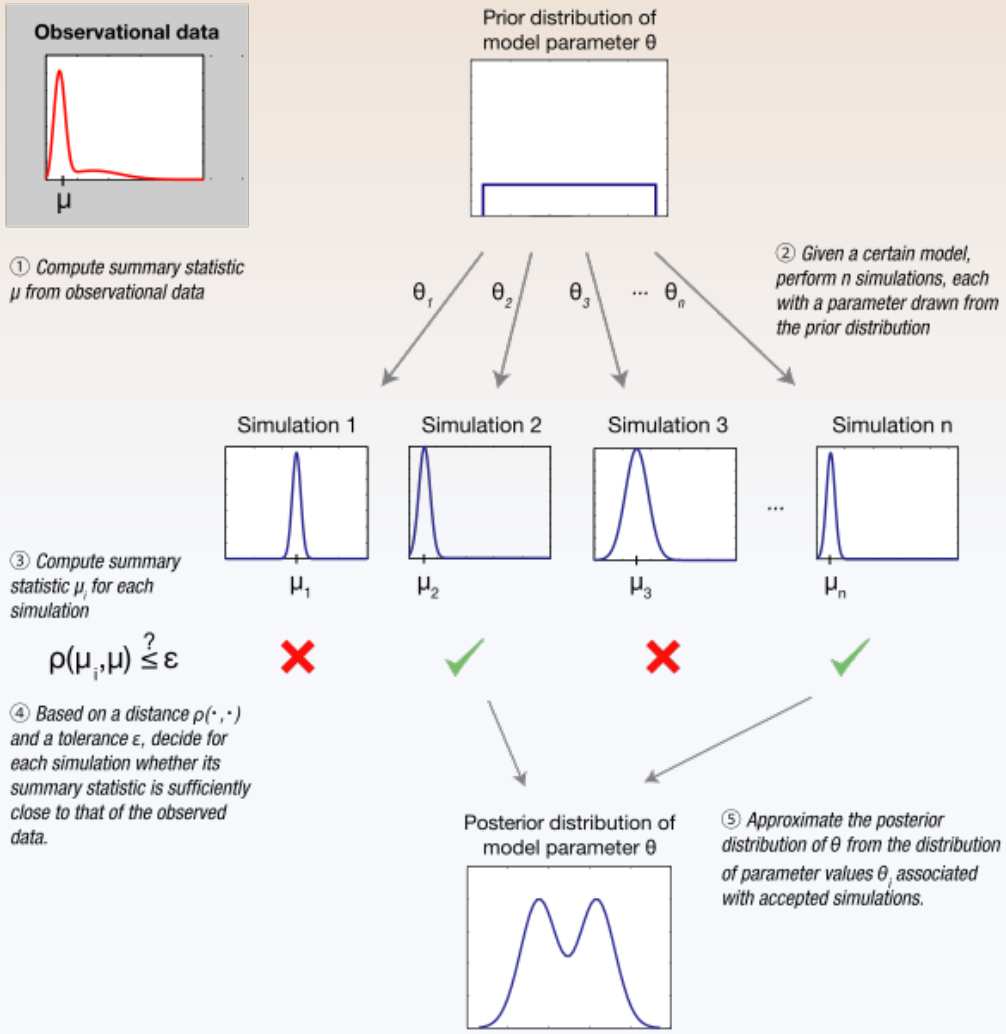
Another Hospital Epidemiology Example

- My group has a model of MRSA transmission in intensive care units
- It's a stochastic model, so we need to run it a lot to understand the variability in the system
- We'd like to let clinicians, policy makers, etc. work with the model behind a web-based interface
- But…



Slow Fitting is Slow

- Clinical audiences are very focused on parameter values – they want to model *their* hospital
- We can do that by letting them input variables, and then fitting a free parameter to their infection rate
- We use a technique called Approximate Bayesian Computation to do this
- ABC is *slow*



Solutions

- This clearly won't work. "The results will be done in three days" is fine for research, but isn't responsive to users
  - Especially not if we're talking about pandemic modeling…
- We could run the model over the entire parameter space to pre-compute the results for every combination a user could possibly ask for, store those in a database, and then just retrieve them when asked for
  - This seems…inelegant
- Can we use model emulation to solve this problem, and run a much faster ML model using the provided parameters?
- Visit Matt Mietchen's poster to find out!
  - Spoiler: Not yet

What About Epi in Machine Learning?

- While ML is unquestionably useful in some aspects of epidemiology, there's also a question of where epidemiology "fits" in the broader framework of health data science, machine learning, precision medicine and/or precision public health, etc.
- As a field, we have struggled with this a bit
- More broadly, I think epidemiology as a field hesitates to invite itself to the party

Applications of Epidemiology to ML

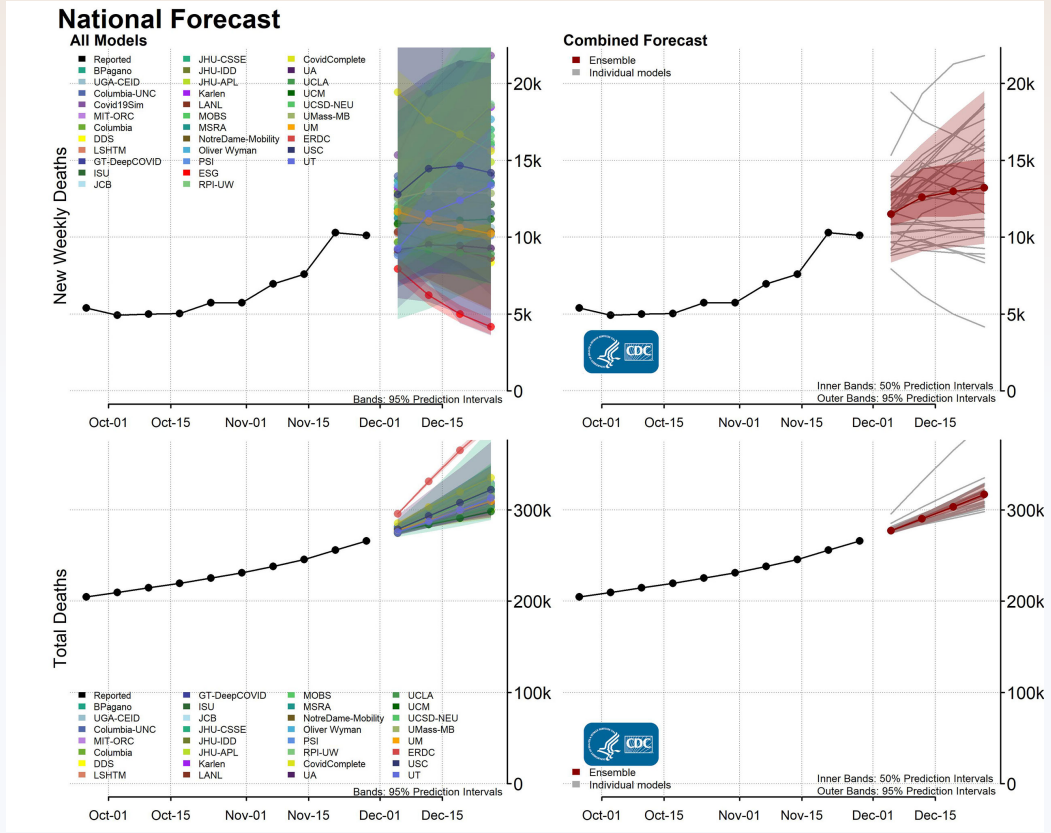
- Ensemble systems
- ML models as an outcome
- Epi's role as experts
- Fairness and Social Determinants of Health
- The inevitable causal questions
- Translation
- Again, there are certainly others

Ensemble systems

- Ensembles are collections of models, where the prediction is some weighted results of the various model outputs
  - Random forest models are an ensemble of decision trees
- Ensembles help protect you from structural misspecification – it's not that you picked the wrong variable, its that you picked the wrong *model*
- Or that some models work better in certain circumstances than others
- "Early in an epidemic, a ruler and a piece of log-scale graph paper is all you need" – A colleague of mine who will remain nameless

COVID-19 Forecasting Hub

- This is being used heavily in the CDC's COVID-19 Forecasting Hub, which uses an ensemble of submitted models
- These can include mathematical models of transmission, machine learning models, more conventional statistical models



Less Intensive Extensions

- Could you make a “doubly robust” exposure weighting model ala IPTW or a propensity score, based on an ensemble of a machine learning model and a more causally focused epi model?
- Comparing approaches – think back to the antibiotic use example from earlier
  - Is the more epidemiology-based model sufficiently accurate, while preserving causal explanations?

Machine Learning as an Outcome

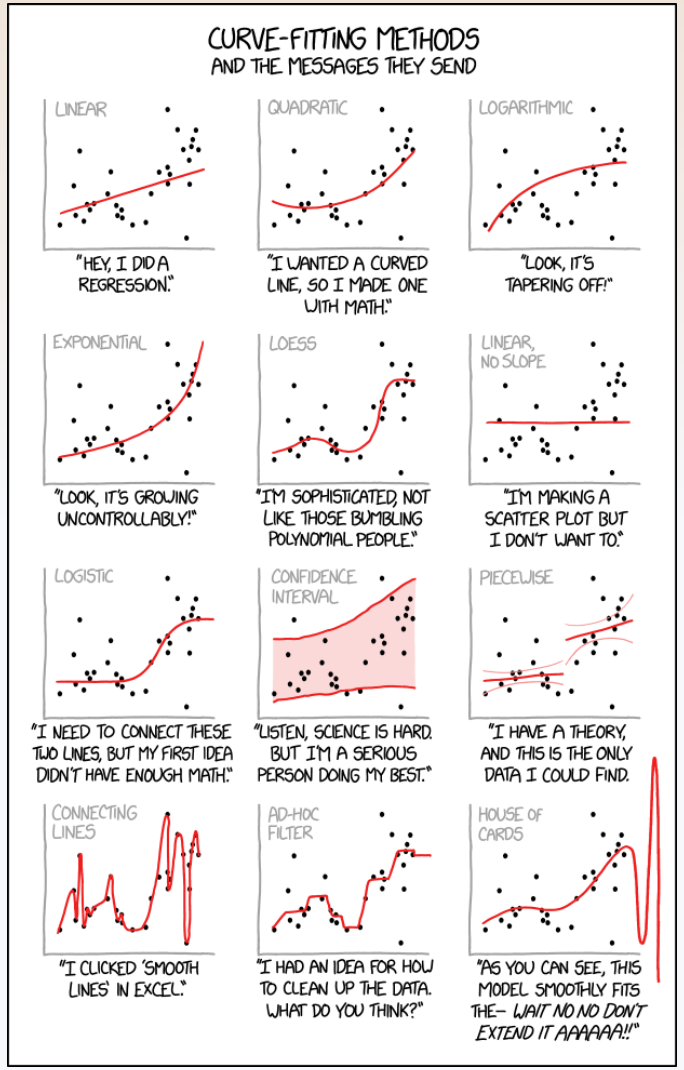
- The introduction of a machine learning system can be thought of as an intervention
- The answer is often “We built it, it predicts well, therefore it works.”
- But does it?
- Enter the Epidemiologists
- Plenty of room for study designs like interrupted time-series, difference-in-difference, etc. evaluating these outcomes
- Consider a predictive model that suggests particular antibiotics – does the usage of that system improve prescribing? Patient outcomes?
  - The adoption of a system like that isn’t random – is there facility-level confounding by indication? Or resources?
- As we use a ML model, does the quality of its predictions begin to degrade?

Epidemiologists as Experts

- We *are* subject matter experts
- Provide relevant problems – what is an interesting *computer science* question, and what is an impactful *public health* question aren't the same
  - Again, the antibiotic example – that prediction model is used *a lot*, but at it's core, it's a straightforward classification problem
- Understanding factors – what does this variable mean? How was it collected? How does missingness arise (ML people are bad at missing data in my experience)?

Epidemiologists as Data Scientists

- "Data Science" is a fuzzy term, but is often defined as someone at the intersection of a specific subject, statistics and computer science
- Epidemiology is arguably sitting on at least two of those three already, and extending into the aspects of computer science needed to meaningfully contribute the ML is straightforward
- There are a lot of techniques that fail to outperform our good friend logistic regression



Fairness

- Epidemiology, through movements like #blackepimatters and an increased focus on how we think about both the consequences of our work and the appropriateness of how we model things like race and gender, is asking many of the same questions as the people working on fair ML
- Many of the questions around fairness are inherently questions of counterfactuals and causal arguments – things we're familiar with
- The "why" of bias, and its impact

More on ”Why?”

- As much as ML is often said to be just prediction and there’s not a causal argument (and often it’s hard to untangle just what’s going into a prediction), the question “Why?” is still asked
- Is whatever variable is important just correlated with something?
- Is there a causal argument there?
- This often involves returning to the data and doing more digging, combined with the subject matter expertise that comes along with being an epidemiologist
- There’s also a whole subfield of ML interested in the same causal inference questions we are

Epidemiologists as Translators

- Epidemiology exists at the intersection of a number of different fields, and is “one adjacent” to a huge number of disciplines
- That means we can bridge gaps between fields that are further apart, don’t necessarily speak the same “language” or have a good intuition about what problems, data, tools, etc. are available
- “I know someone who can fix that…” or “I think we could find the data to explore that question…”

Why This Is Worth Exploring

- There are benefits to this, even if you don't end up wanting to do ML
- Machine learning has a lot of its roots in computer science and is used heavily in tech, which means dabbling in ML invariably means exposure to software engineering concepts
- Epidemiologists are coders – but often not very good ones
- Public health would benefit greatly from more people versed in things like version control, test driven development, etc.
- Epidemiologists working in those spaces also ensure those tools remain relevant to us
- They're also another career path

Thank You

- Shameless Plug:

I’m very likely to be in the market for a postdoc quite soon, with some geographic flexibility (as much as Pullman, WA is a beautiful place). If you’re interested in applications of network science to disease surveillance in communities with limited access to health resources, drop me an email at Eric.Lofgren@wsu.edu or on Twitter @GermsAndNumbers