



Underspecification
Presents Challenges
for Credibility in
Modern Machine
Learning



Underspecification Presents Challenges for Credibility in Modern Machine Learning

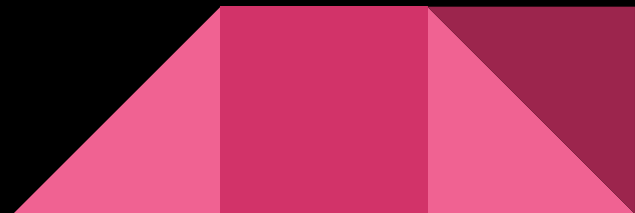
(and also maybe might have **killed** some folks)

Synopsis

Forty google researchers co-authored a paper about their investigation into “underspecification”, one of the issues that can plague ML models.

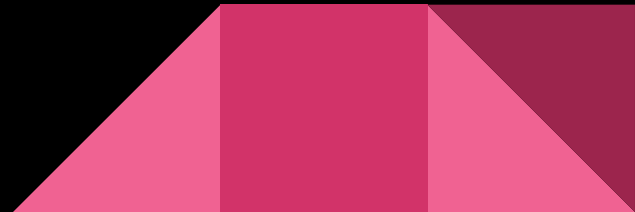
They conclude that underspecification is **widespread, underdiagnosed, and ill-addressed** across the ML production pipeline as a whole.

They examine and undertake several case studies exemplifying this issue, some of which demonstrate the issue has significant **real-world impact**.



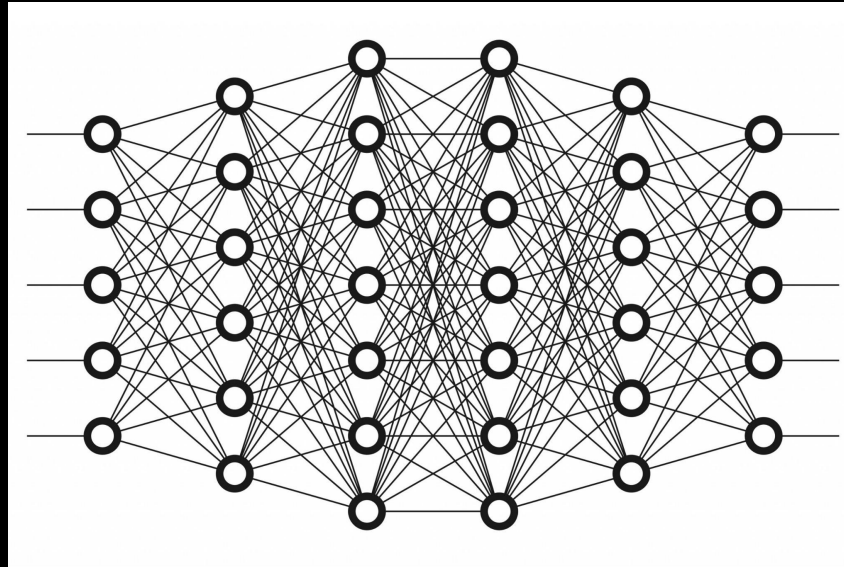
What is Machine Learning?

- Machine learning is a general term for when computer systems (machines) follow a repeated process of trial runs and error checking in order to improve their own performance against a set of success criteria (learning).
- An ML model is the product of this process: an algorithm or heuristic that takes in a wide variety of data inputs and produces a predictive output.



Machine Learning: In Brief

Neural Net



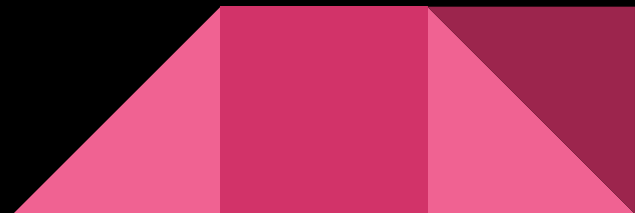
All the Data /
Prediction Factors

Results /
Predictive Output

Decision Layers

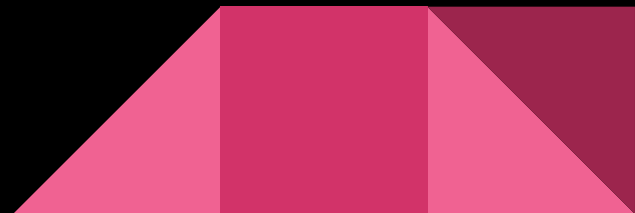
Machine Learning: In Brief

- Moreover, machine learning is incredibly powerful because there is no requirement for a human understanding of correlations between the prediction factors and the predictive output.
- The learning process will find these correlations on its own, meaning we can discover relations among data using machine learning that were opaque, unknown, or unknowable to us prior.



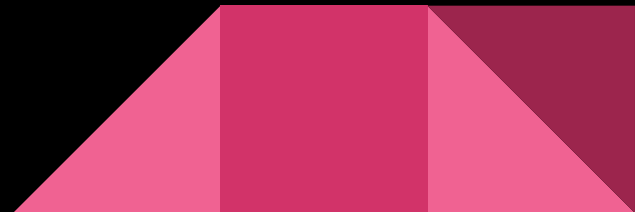
Machine Learning: In Brief

- A single ML process can produce many ML models from the same data. How can this be?
- Every time the ML process is initialized, it starts with random (or pseudorandom) values for its initial decisions. In the case of neural nets discussed before, random values of the weights on the edges.



Machine Learning: In Brief

- Ideally, there are strong and distinct predictive factors in the data, predictors with "credible inductive biases".
- Our final ML model may be "expected to encode some essential structure of the underlying system" due to having learned these factors.
- Each factor does not cover the same ground as the other factors, nor do they **cancel out, eclipse, and mask** important information from other factors.



We Have



~NEVER~

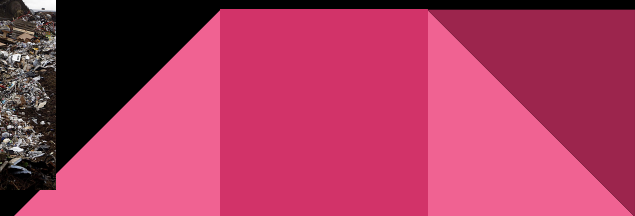
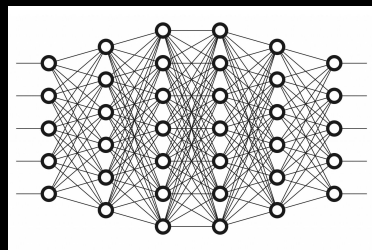


Lived a Good Life



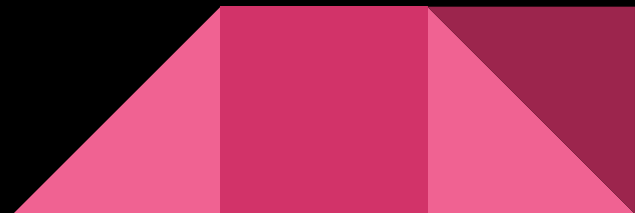
In Practice

- We dump **garbage** data into our ML processes, expecting them to sort through it for us.
- We **trust** that by achieving high accuracy on training and testing data (taken to be a representative sample of the real world info) our models are capturing some degree of true correlation and will therefore generalize.
- When **they don't**, we attribute the difference in performance to “structural conflict” and **don't follow up** with refinements of the model.



Underspecification

- Underspecification occurs when there are multiple, distinct, equally viable solutions that our ML process can find in its training data in order to meet its predictive goals.
- An ML pipeline is considered “underspecified” when “it can return many predictors with equivalently strong held-out performance in the training domain”

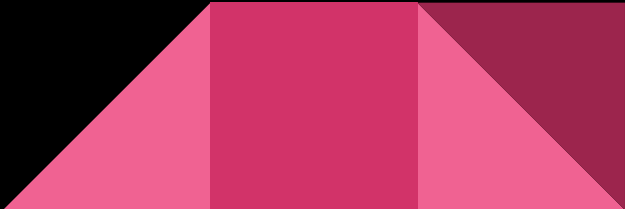




Underspecification

- Underspecification occurs when there are multiple, distinct, equally viable solutions that our ML process can find in its training data in order to meet its predictive goals.
- An ML pipeline is considered “underspecified” when “it can return many predictors with equivalently strong held-out performance in the training domain”

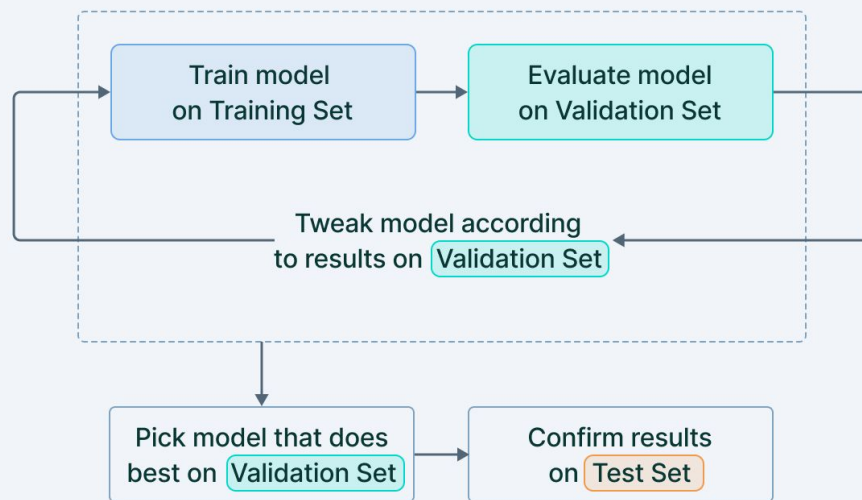
Underspecification: Example

- Consider a simple allegory: $X + Y = 2$
 - Possibly, $X = 1$ and $Y = 1$
 - However, $X = -1$ and $Y = 2$ is also possible.
 - We have no way of knowing which is better until we hit the real world and see another equation $X - Y = 0$
 - Now, we can tell which of two solutions was more appropriate, $X = 1$ and $Y = 1$.
 - Prior to encountering the new data in the wild, both solutions were equally valid.
- 

Underspecification

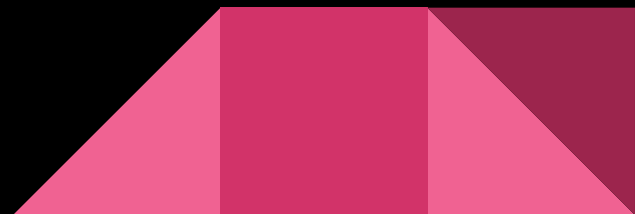
- It's the same for our ML models.
- In an underspecified scenario, the same process can generate **more than one model** that performs optimally on the validation and test sets and these models can **use different sets of predictors**.

Training data/validation/test



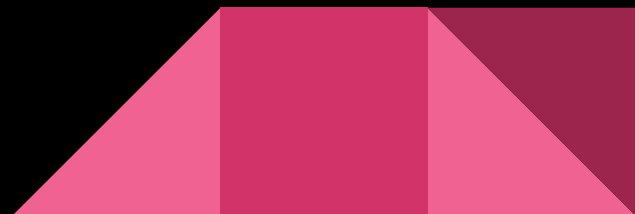
Reality Check

- We'd expect these models to show similar rates of performance on deployment (and under the **structural conflict** assumption, similar loss from the lab levels of performance).



Reality Check

- We'd expect these models to show similar rates of performance on deployment (and under the **structural conflict** assumption, similar loss from the lab levels of performance).
- However, this is not what the authors see.



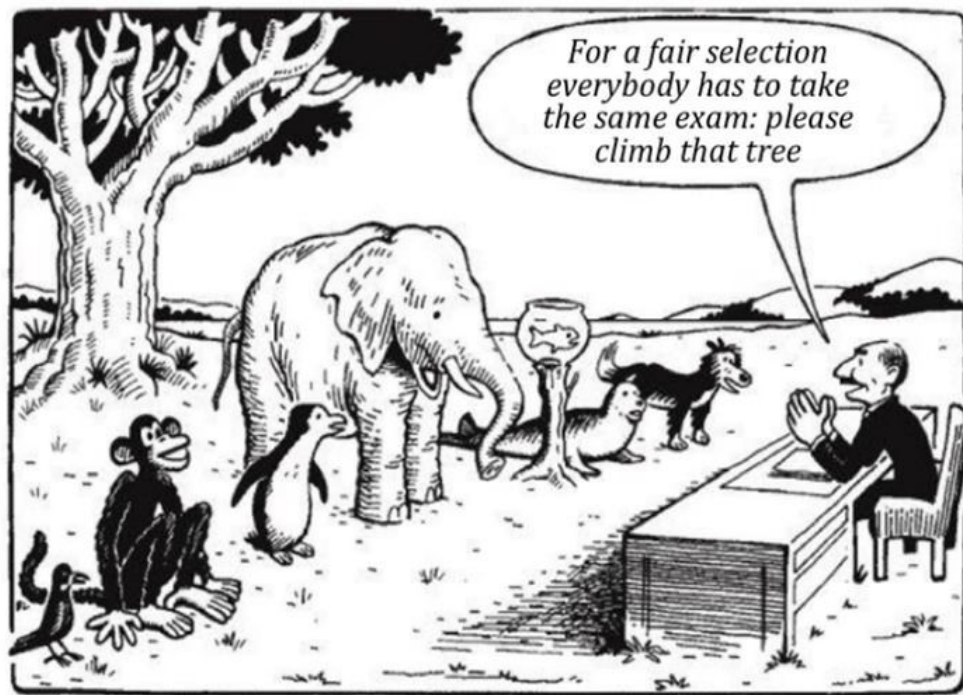
Reality Check

- We'd expect these models to show similar rates of performance on deployment (and under the **structural conflict** assumption, similar loss from the lab levels of performance).
- However, this is not what the authors see.



Reality Check

- Instead, models using some sets of predictors generalize better when they hit the field and some generalize **much worse**.
- Side note, training to too high of an accuracy level in the lab can also eliminate models that generalize well, a separate issue called **“overfitting”**.



Our Education System

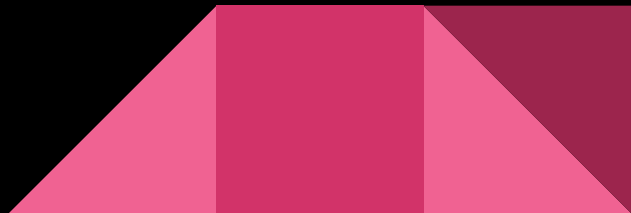
“Everybody is a genius. But if you judge a fish by its ability to climb a tree, it will live its whole life believing that it is stupid.”

- Albert Einstein

Central Claims

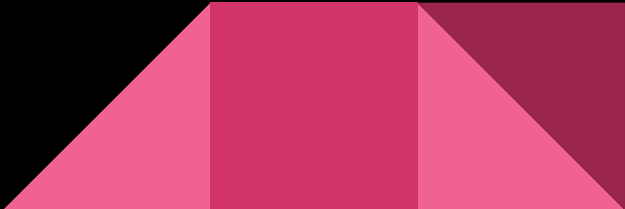
"...underspecification in ML pipelines is a key obstacle to reliably training models that behave as expected in deployment.

Specifically, when a training pipeline must choose between many predictors that yield near-optimal iid performance, if the pipeline is **only** sensitive to iid performance, it will return an **arbitrarily** chosen predictor from this class."



Central Claims

"...underspecification is **ubiquitous** in modern applications of ML, and has **substantial practical implications**. We support this claim with an empirical study...in computer vision, **medical imaging**, natural language processing (NLP), and **electronic health record (EHR) based prediction**"

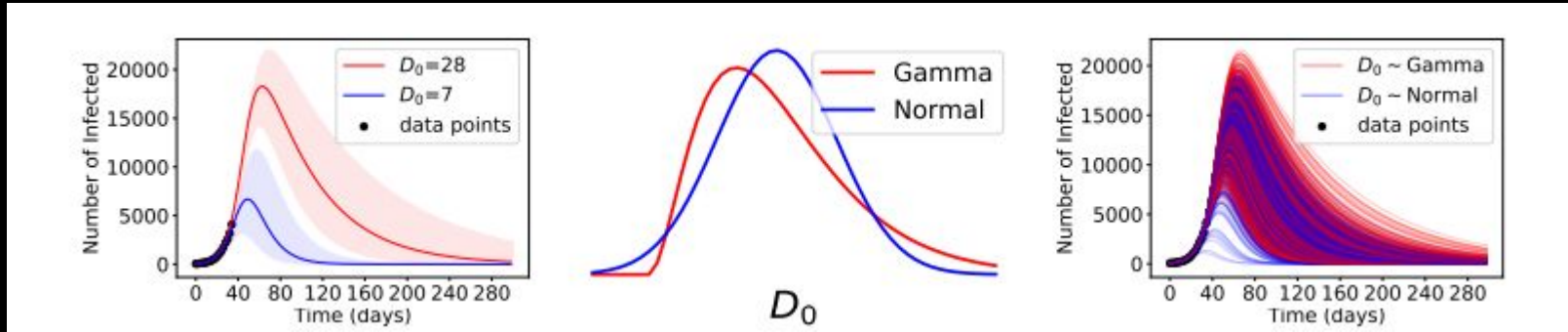


Theoretical Example: Epidemiology

- Epidemic simulation, the sort of thing you want to get right.
- They simulate using a Susceptible-Infected-Recovered model, where we have the rates at which S, I, and R change over time in a population of size N, β is the transmission rate of the disease and D is the average duration that an infected individual remains infectious.
- Rates of change for these values are given by

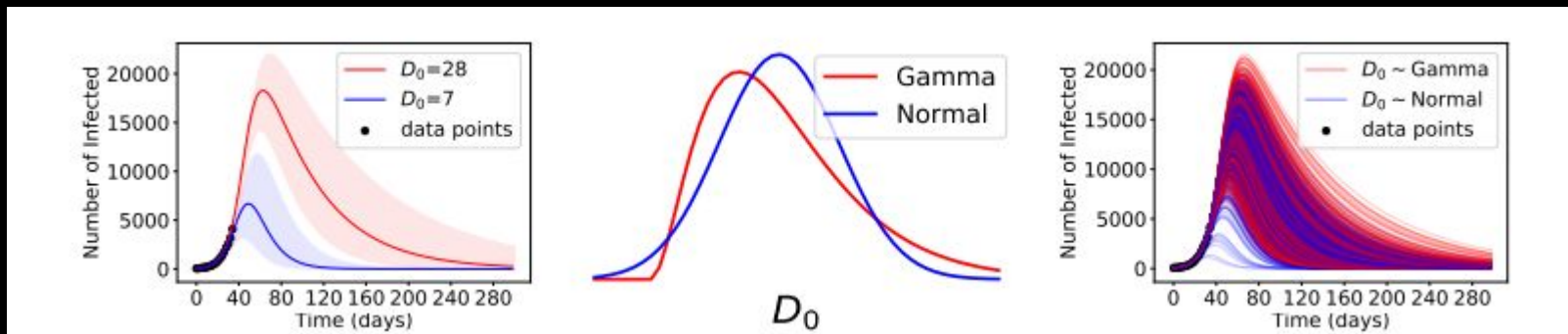
$$\frac{dS}{dt} = -\beta \left(\frac{I}{N} \right) S, \quad \frac{dI}{dt} = -\frac{I}{D} + \beta \left(\frac{I}{N} \right) S, \quad \frac{dR}{dt} = \frac{I}{D}.$$

Theoretical Example: Epidemiology



- The **fewer observations** we use, the more **unsure** our model is. Early on, many parameter values work equally well fitting our data.
- This is because S is nearly our entire N , so our rate of infection approximates $\beta - 1/D$. The equation is simple, and many values can fit.

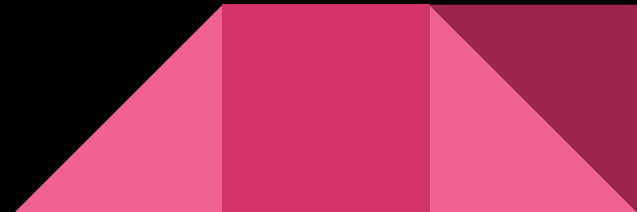
Theoretical Example: Epidemiology



- Key behaviors of the model are seen to be sensitive to the time at which we initialize our estimate of D , as well as the **arbitrary** random starting values or our process (**and the distribution they're drawn from**).
- All this while being fed the same data and meeting the same accuracy threshold in training.

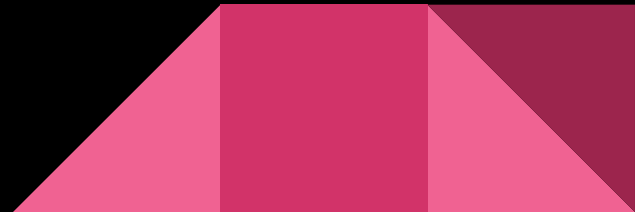
Case Studies in Medicine: Acute Kidney Injury

- The authors examine a model that uses Recurrent Neural Network (RNN) architecture with Electronic Health Record (EHR) data to predict acute kidney injury (AKI).
- "AKI is a common complication in hospitalized patients and is associated with increased morbidity, mortality, and healthcare costs (Khwaja, 2012). Early intervention can improve outcomes in AKI (National Institute for Health and Care Excellence (NICE), 2019), which has driven efforts to predict it in advance using machine learning."



Case Studies in Medicine: Acute Kidney Injury

- Program has "state of the art performance" up to 48 hours in advance of usual diagnosis with an average accuracy of 55.8%, (90.2% for episodes associated with dialysis administration).
- This is considered "strong discriminative performance".

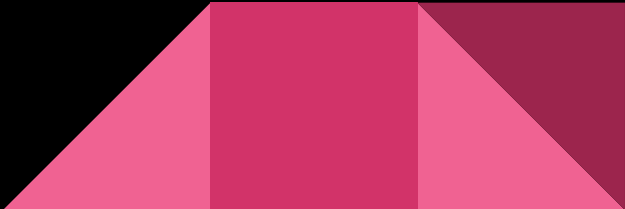


Case Studies in Medicine: Acute Kidney Injury

- Program has "state of the art performance" up to 48 hours in advance of usual diagnosis with an average accuracy of 55.8%, (90.2% for episodes associated with dialysis administration).
- This is considered "strong discriminative performance".



Case Studies in Medicine: Acute Kidney Injury

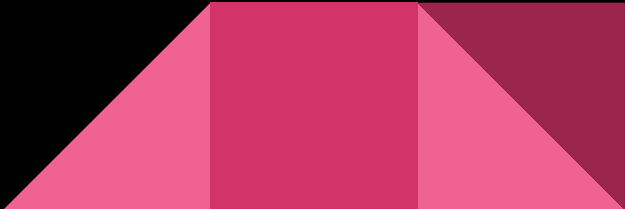
- The authors want to know though, is the model picking up on strong predictors with those credible inductive biases we mentioned?
 - They start by examining the data fed into the model, which includes labs, vital signs, diagnosis codes, etc in 6-hour time buckets.
 - Identifying factors such as free text notes and rare diagnoses were excluded and a small amount of noise added to all numerical values for patient privacy, times shifted to common scale.
- 

Case Studies in Medicine: Acute Kidney Injury

- The authors want to know though, is the model picking up on strong predictors with those credible inductive biases we mentioned?
- They start by examining the data fed into the model, which includes labs, vital signs, diagnosis codes, **etc** in 6-hour time buckets.
- Identifying factors such as free text notes and rare diagnoses are excluded and a small amount of noise added to all numerical variables to protect privacy, times shifted to common scale.

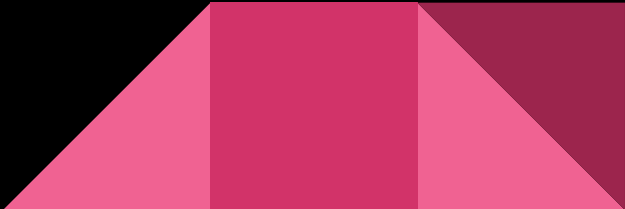


Case Studies in Medicine: Acute Kidney Injury

- The authors train a total fifteen models, five random seeds for each of the three RNN cell types: Simple Recursive Units, Long Short-Term Memory or Update Gate RNN.
 - They perform stress tests on operational predictors (fields tied to how the diagnosis was made, as opposed to physiological ones).
 - Specifically, the timing and number of labs recorded in the EHR
 - Three types of stress tests here:
 - - Stratified Performance Evaluations
 - - Shifted Performance Evaluations
 - - Contrastive Evaluations
- 

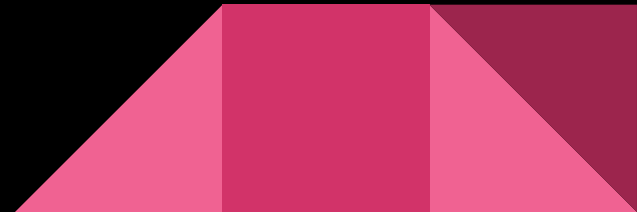
Case Studies in Medicine: Acute Kidney Injury

“AKI is diagnosed based on lab tests, and there are clear temporal patterns in how tests are ordered. For most patients, creatinine is measured in the morning as part of a ‘routine’, comprehensive panel of lab tests. Meanwhile, patients requiring closer monitoring may have creatinine samples taken at additional times, often ordered as part of an ‘acute’, limited panel (usually, the basic metabolic, panel 6). Thus, both the time of day that a test is ordered, and the panel of tests that accompany a given measurement may be considered primarily as **operational factors** correlated with AKI risk.”



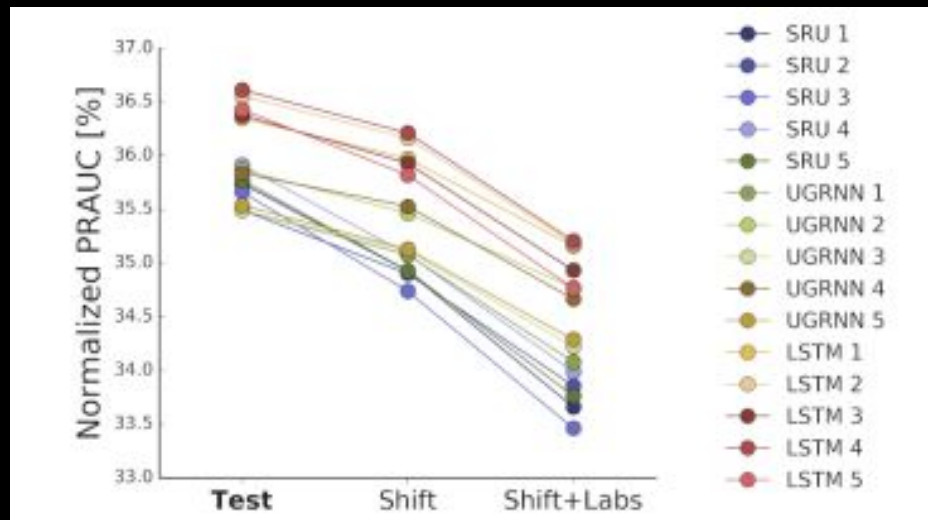
Case Studies in Medicine: Acute Kidney Injury

- They shift the timings of the lab tests by a consistent factor, so they fall in different 6h buckets. Morning can be noon, noon can be evening, etc.
- They remove from the panels any tests that have nothing to do with AKI diagnosis. In effect, that 'routine' panel now looks like the 'limited' panel.

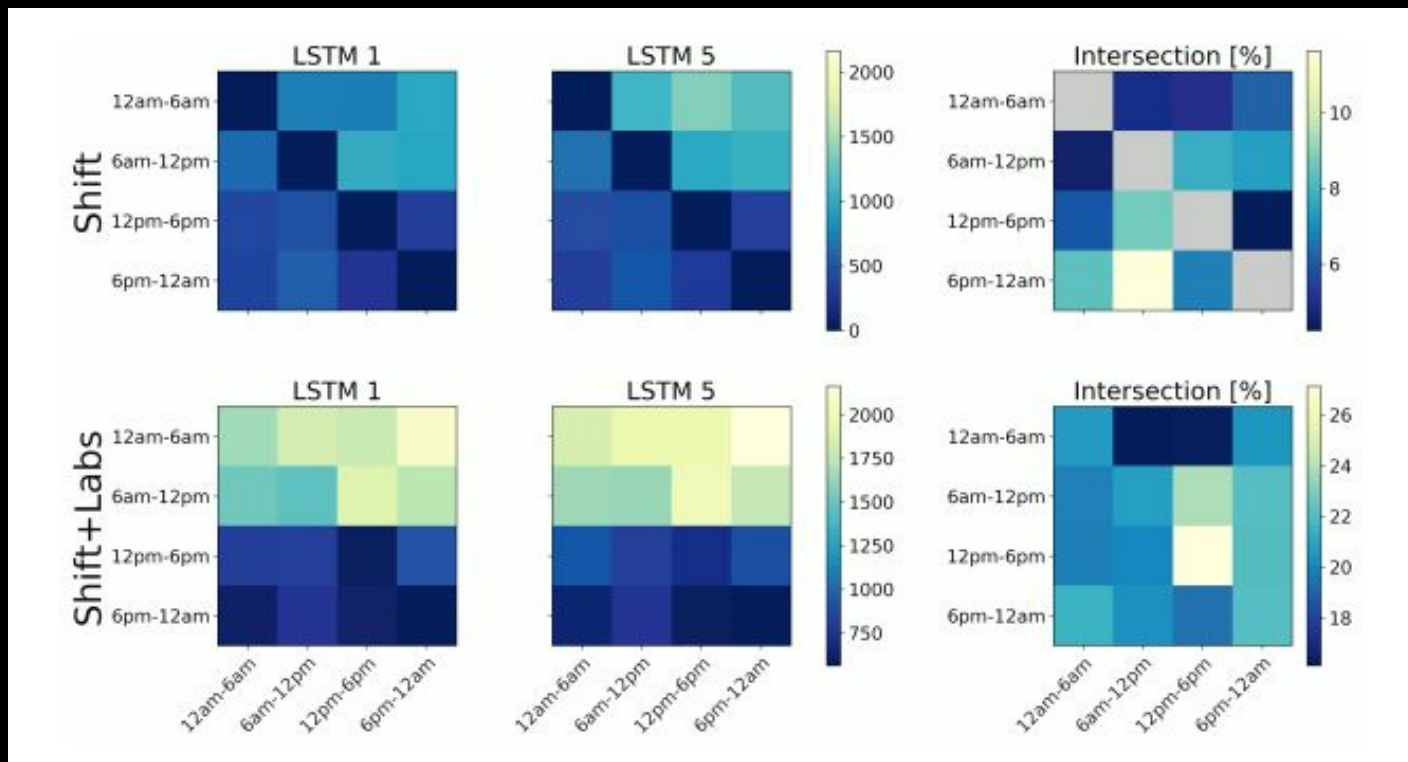


Case Studies in Medicine: Acute Kidney Injury

- They shift the timings of the lab tests by a consistent factor, so they fall in different 6h buckets. Morning can be noon, noon can be evening, etc.
- They remove from the panels any tests that have nothing to do with AKI diagnosis. In effect, that 'routine' panel now looks like the 'limited' panel.

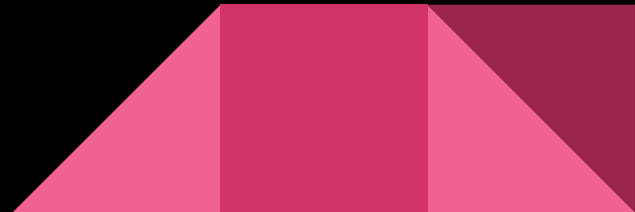


Case Studies in Medicine: Acute Kidney Injury



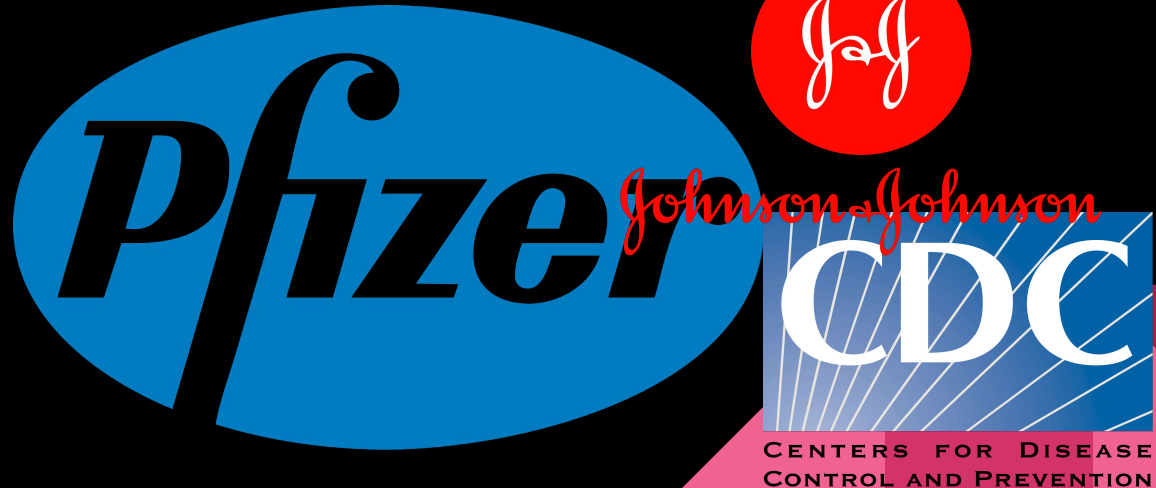
Case Studies in Medicine: Dermatology

- The authors find a **deployed** ML Model for identifying skin cancer (Liu et al. 2020b)
- The authors create 10 ML models after the pattern of this one, differing only in "random initialization at the fine tuning stage" and subject them to stress tests using a test set with different skin tones than the training data.



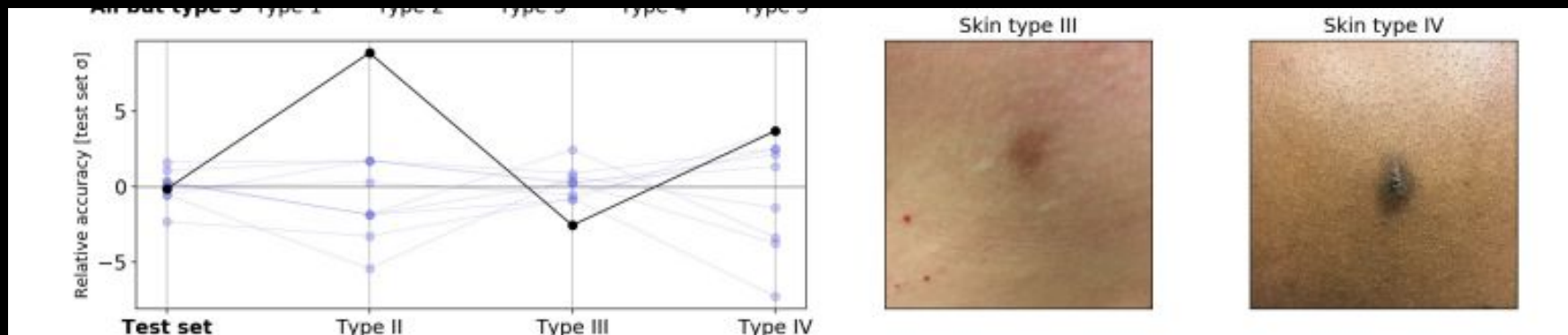
Why Skin Types?

- "Given the social salience of skin type, this concern is aligned with broader concerns about ensuring that machine learning does not **amplify existing healthcare disparities**".
- Hello, ethics.



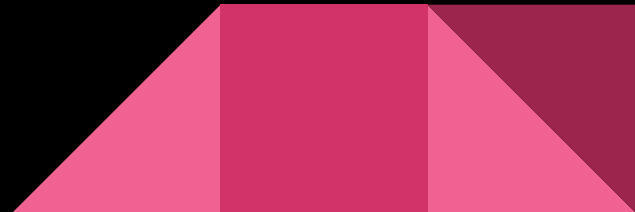
Case Studies in Medicine: Dermatology

- “These results are exploratory, but they suggest a need to **pay special attention** to this dimension of underspecification in ML models for dermatology”



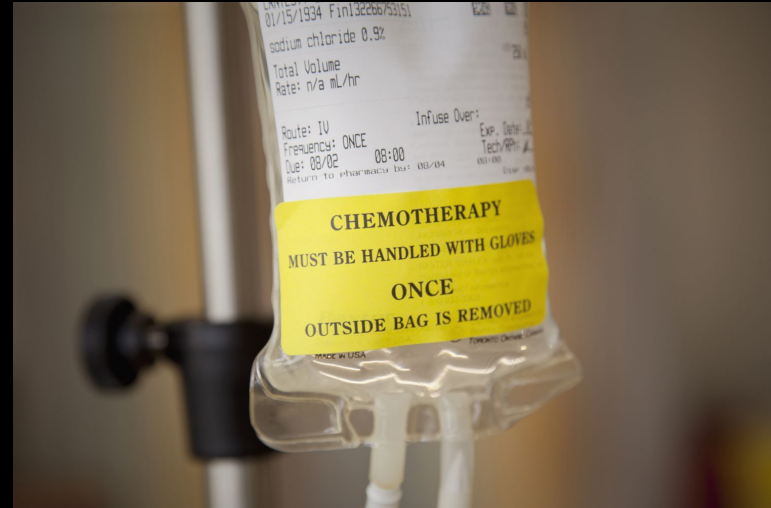
Case Studies in Medicine: Dermatology

- Some of the created models perform better, some worse, but the main point is that if you don't generate many of them, measure their differences, and select the most equitable one, if you just go with the first one generated and take it into your practice, you can be do many of your patients some seriously extreme disservice.



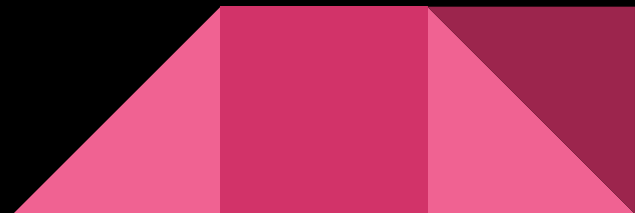
Case Studies in Medicine: Dermatology

- Misidentifying cancer and subjecting patients to unnecessary treatment
- Failing to correctly identify that they have cancer
- All due to arbitrary, random starting values of your ML.



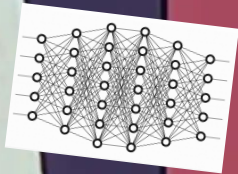
Honorable Mention

- You can achieve this and other, even more harrowing results by having an **untidy, unrepresentative dataset**.
- "For example, Winkler et al. (2019) report on a CNN model used to diagnose skin lesions, which exhibited strong reliance on **surgical ink markings** around skin lesions that **doctors had deemed to be cancerous**."
- "...but these markings would **not** be expected to be **present in deployment**, where **the predictor would itself be part of the workflow for making a diagnostic judgment**."





NYPD



Case Studies in Medicine

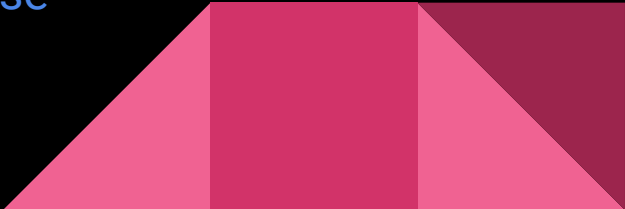
- "Medical imaging is one of the primary high-stakes domains where deep image classification models are directly applicable."
- "A key use case for these models is to augment human clinical expertise in underserved settings, where doctor capacity may be stretched thin."

Hello, ethics.



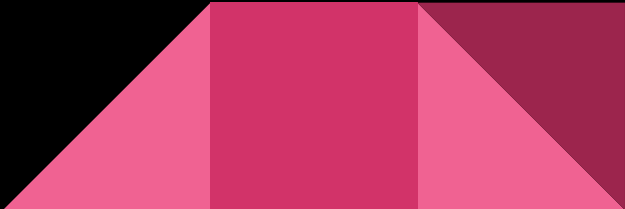
The Paper's Conclusion

"...our findings underscore the need to thoroughly test models on application-specific tasks, and in particular to check that the performance on these tasks is **stable**. The extreme complexity of modern ML models ensures that **some aspect of the model will almost certainly be underspecified**; thus, the challenge is to **ensure that this underspecification does not jeopardize the inductive biases that are required by an application.**"

- Bootstrap MLs, check their performance, pick your horse
 - Stress tests for predictors, make sure they make sense
- 

The Paper's Conclusion

"Finally, these results suggest a need for training and evaluation techniques tailored to address underspecification, such as flexible methods to constrain ML pipelines toward the credible inductive biases for each specific application."

- Hard to generalize though.
 - Getting a human involved in deciding which data to use has impacts on bias and fairness.
- 

My Conclusion

ML (by itself) has a much greater propensity for damaging human lives through **misinterpretation, misapplication, and arbitrary error** than **malign agency**.

