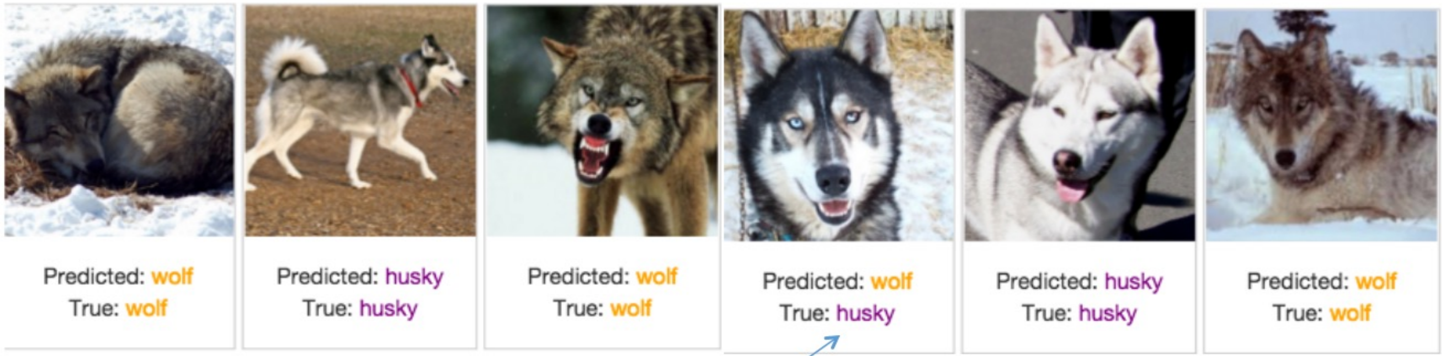
INTERPRETABILITY, EXPLAINABILITY, AND TRANSPARENCY

Dr. Savannah Thais
Columbia University
LSSTC
03/03/2023

Why should we care about interpretable/ explainable/ transparent AI?

Example Case: Doggos

Train a neural network to predict wolf v. husky



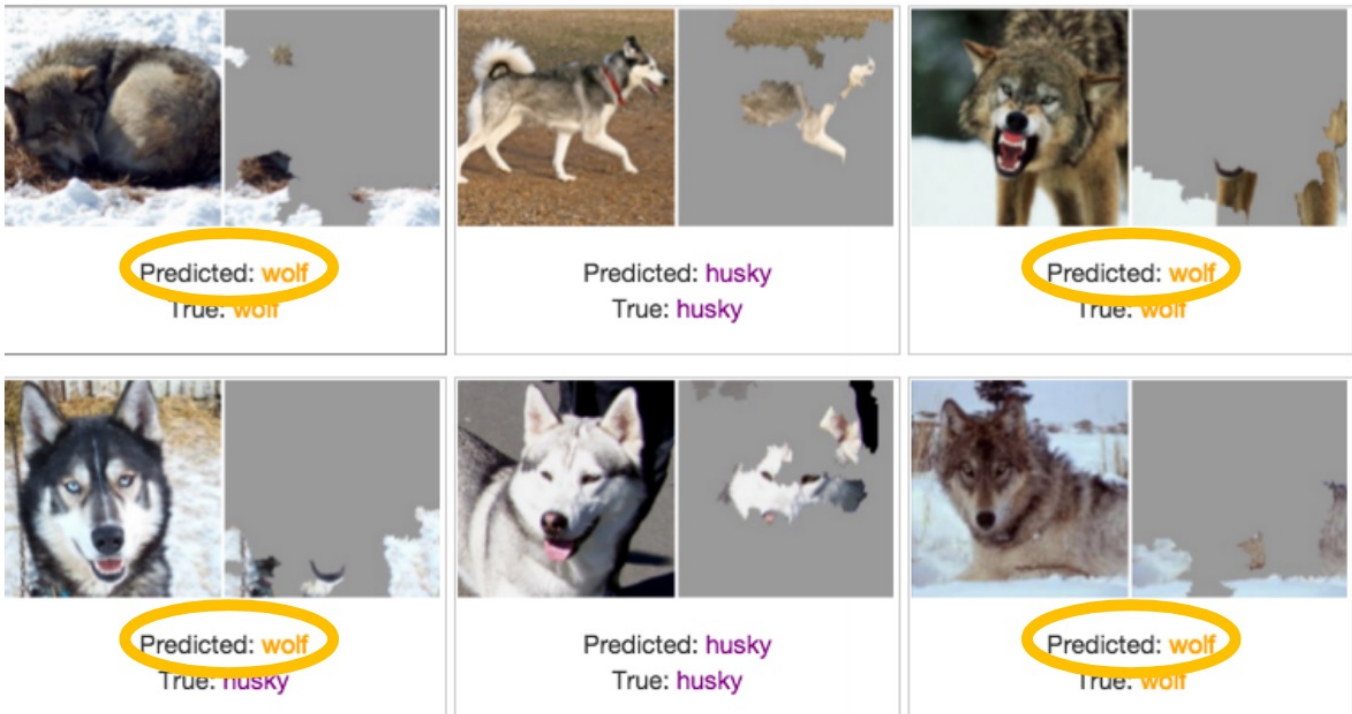
Only 1 mistake!!!

Do you trust this model?
How does it distinguish between huskies and wolves?

slide by Marco Tulio Ribeiro, KDD 2016

Example Case: Doggos

Explanations for neural network prediction



We’ve built a great snow detector… ☹

slide by Marco Tulio Ribeiro, KDD 2016

Example Case: Pnemonia

Predicting pneumonia risk from a set of reasonable variables:

Patient-history findings
Age (years)
Gender
A re-admission to the hospital
Admitted from a nursing home
Admitted through the ER
Has a chronic lung disease
Has asthma
Has diabetes mellitus
Has congestive heart failure
Has ischemic heart disease
Has cerebrovascular disease
Has chronic liver disease
Has chronic renal failure
Has history of seizures
Has cancer
Number of above disease conditions
Pleuritic of chest pain

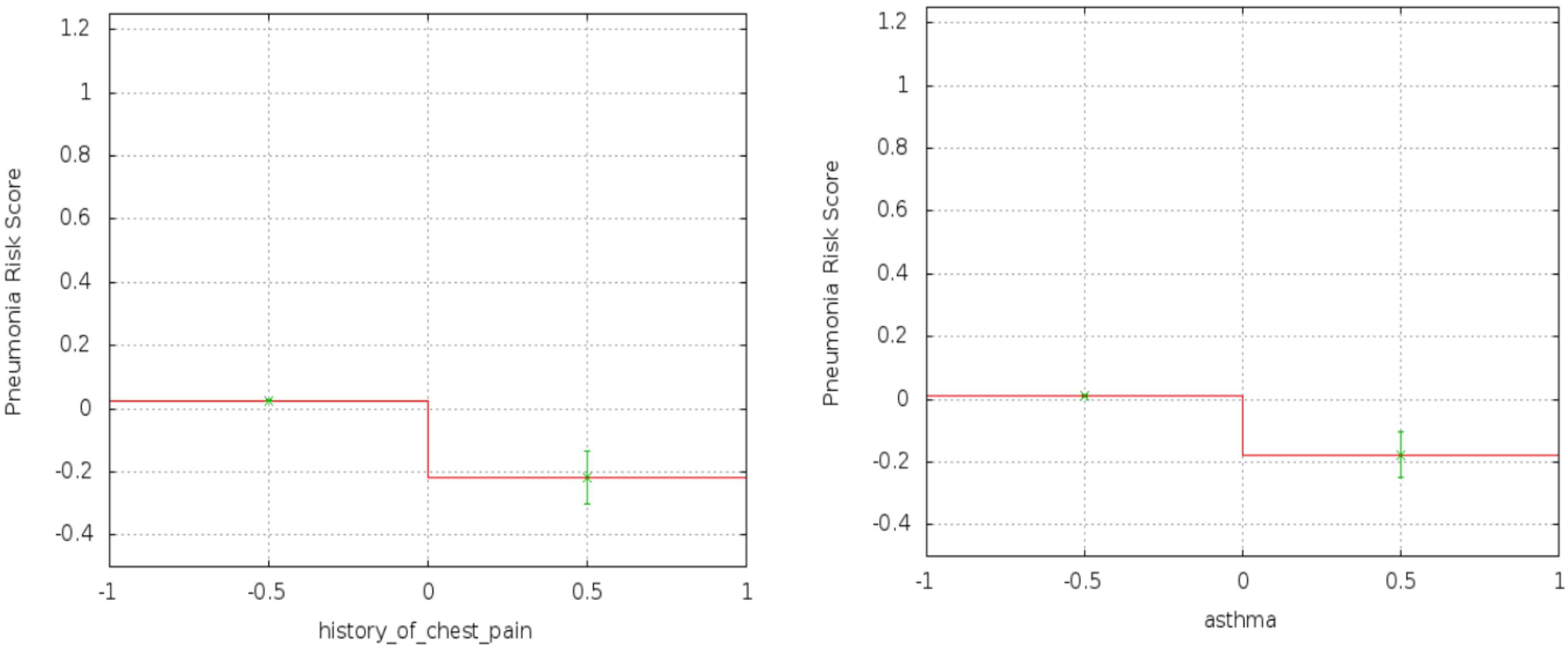
Physical examination findings
Respiration rate (resps/min)
Heart rate (beats/min)
Systolic blood pressure (mmHg)
Temperature (°C)
Altered mental status (disorientation, lethargy, or coma)
Wheezing
Stridor
Heart murmur
Gastrointestinal bleeding

Laboratory findings
Sodium level (mEq/l)
Potassium level (mEq/l)
Creatinine level (mg/dl)
Glucose level (mg/dl)
BUN level (mg/dl)
Liver function tests (coded only as normal* or abnormal)
Albumin level (gm/dl)
Hematocrit
White blood cell count (1000 cells/μl)
Percentage bands
Blood pH
Blood pO_2 (mmHg)
Blood pCO_2 (mmHg)

Chest X-ray findings
Positive chest X-ray
Lung infiltrate
Pleural effusion
Pneumothorax
Cavitation/empyema
Lobe or lung collapse
Chest mass

Example Case: Pneumonia

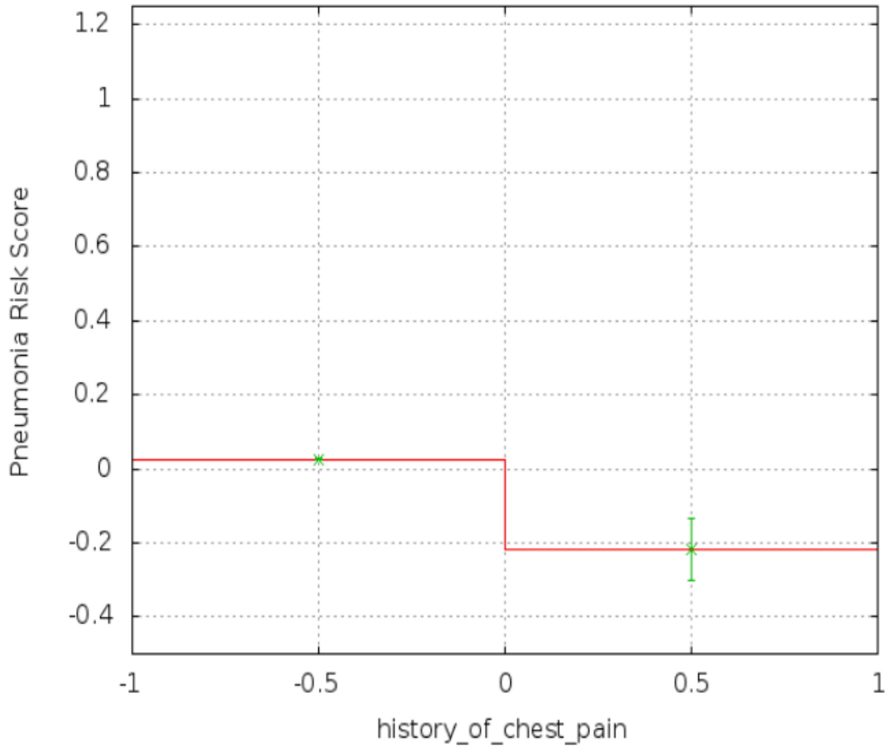
Predicting pneumonia risk from a set of reasonable variables:

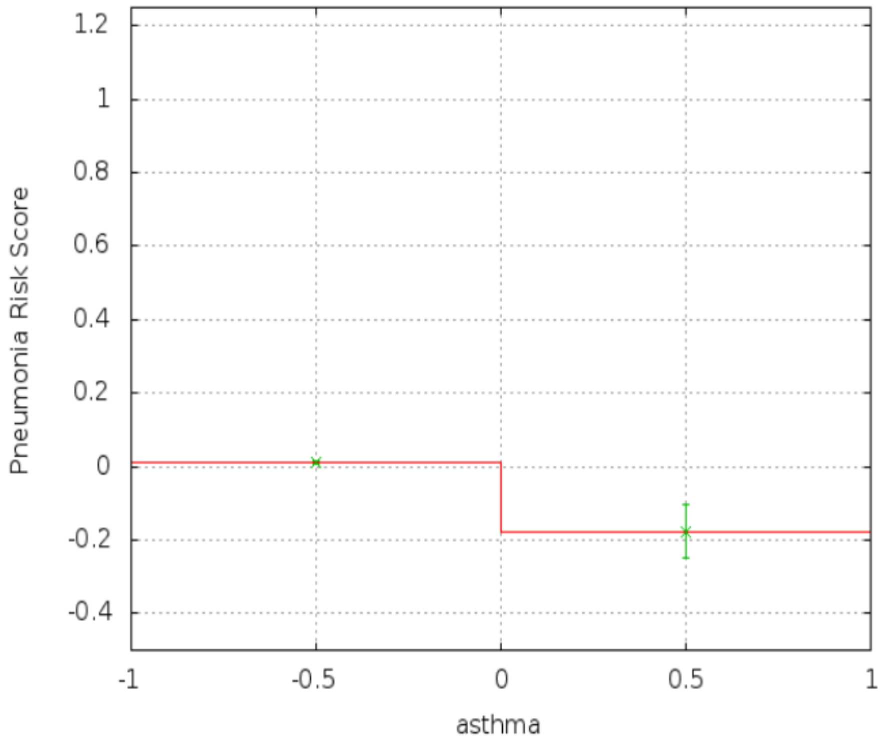


Applying interpretability method to the model revealed that both asthma and history of chest pain were associated with a LOWER risk of pneumonia!

Example Case: Pneumonia

Predicting pneumonia risk from a set of reasonable variables:





This is is obviously not the correct causal relationship. Discussing model outputs and explanations with doctors revealed the cause.

Interpretability/Explainability can help diagnose causal issues, but be careful that they don't create new ones!

Black Box Auditing Techniques

LIME

- The best explanation of a simple model is the model itself: the explanation is both accurate and interpretable. For complex models we must use a simpler explanation model — an interpretable approximation of the original model.

$$f : \mathbb{R}^d \to \mathbb{R}$$

model being explained

$$g \in G, dom(g) = \{0,1\}^{d'}$$

explanation model from a class of interpretable models, over a set of **simplified features**

- The overall goal of LIME is to identify an interpretable model that is locally faithful to the classier

$f(x)$ denotes the probability that ***x*** belongs to some class

π_x is a **proximity measure** relative to ***x***

we make no assumptions about ***f*** to remain model-agnostic: draw samples weighted by π_x

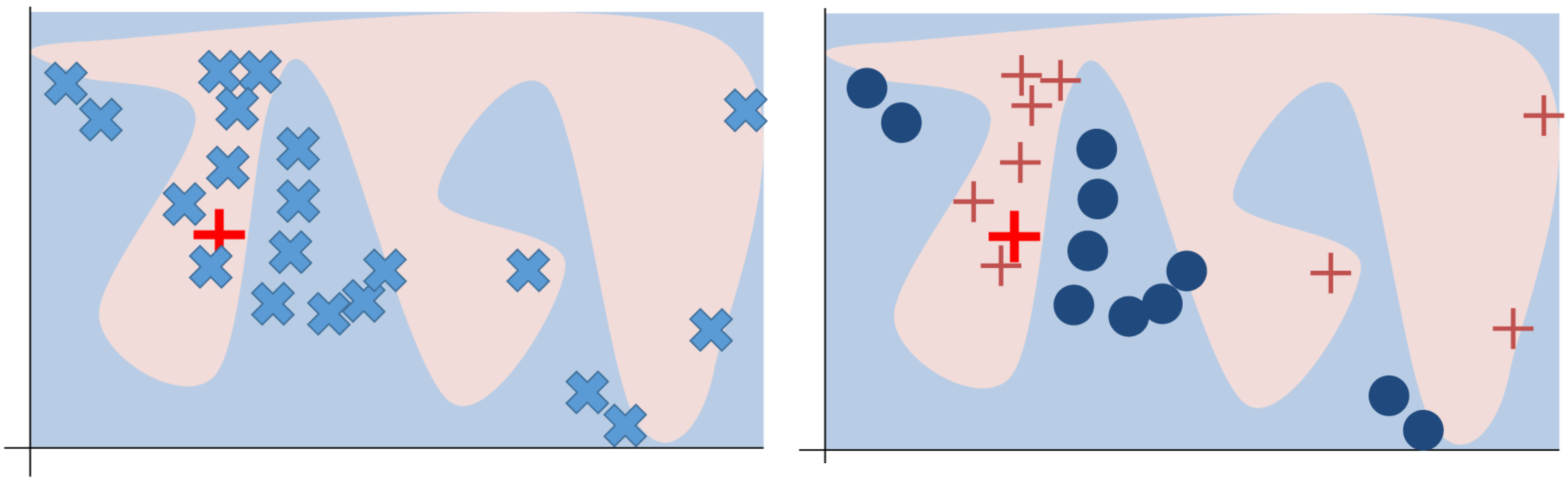
explanation

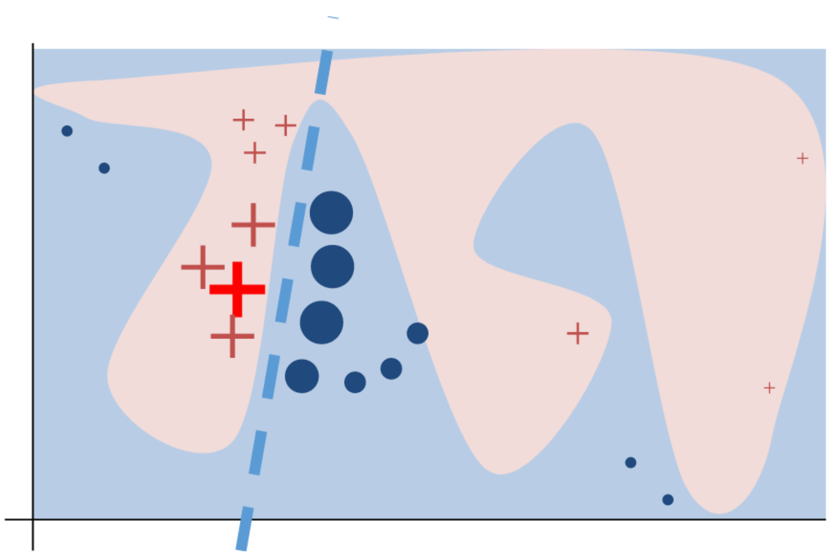
measures how unfaithful is ***g*** to ***f*** in the locality around ***x***

$$\xi(x) = \operatorname{argmin}_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

[Original paper]

LIME





1. sample points around +
2. use complex model ***f*** to assign class labels
3. weigh samples according to π_x
4. learn simple model ***g*** according to samples

based on a slide by Marco Tulio Ribeiro, KDD 2016

SHAP

- Additive feature attribution methods have an explanation model that is a linear function of binary variables (simplified features

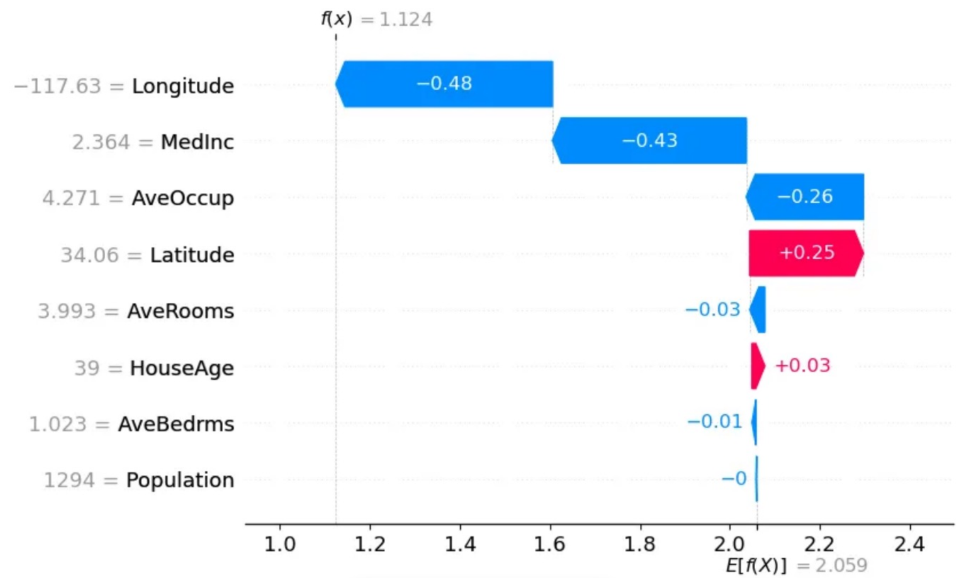
$$g(x') = \phi_0 + \sum_{i=1}^{d'} \phi_i x'_i \quad \text{where } x' \in \{0,1\}^{d'}, \text{and } \phi_i \in \mathbb{R}$$

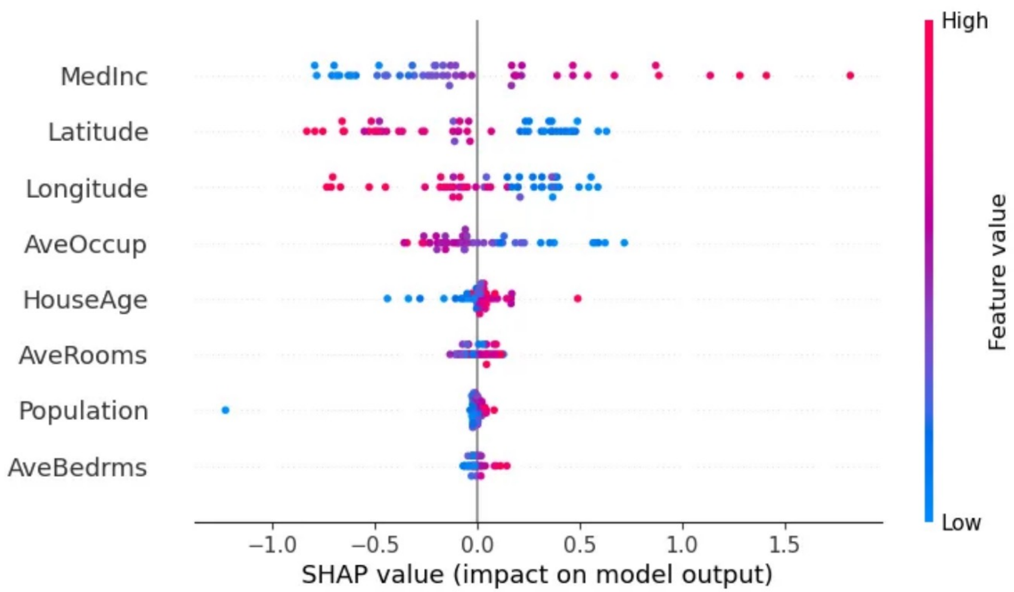
- SHAP estimates the impact of an individual feature on the prediction of an individual input datum by perturbing each feature and measuring how the prediction changes
  - The original formulation requires training new models with different combinations of input features

$$SHAP_{feature}(x) = \sum_{set:feature \in set} \left[|set| \times \binom{F}{|set|} \right]^{-1} \left[Predict_{set}(x) - Predict_{set \setminus feature}(x) \right]$$

Original paper

SHAP





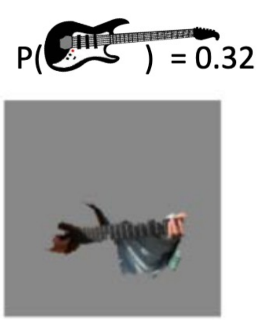
- Example: predicting housing prices (in 100k)
  - Individual SHAP ~shows why an individual decision was made (useful for redress)
  - Plotting all SHAPs gives some intuition for what is important to the model
- Neither LIME nor SHAP explains the entire model
  - Neither handle correlated features well
  - **Neither can actually measure causality**

Relevance Propagation

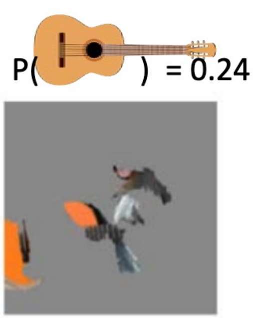
- Layerwise Relevance Propagation seeks to identify precisely which input features maximally contribute to a prediction
  - Uses the actual specific model and inputs, unlike LIME and SHAP
- Used specifically for NN based models
  - Uses network weights and neural activations to propagate output back through the network to input layer

$$R_j = \sum_k \frac{a_j w_{jk}}{\sum_{0,j} a_j w_{jk}} R_k$$

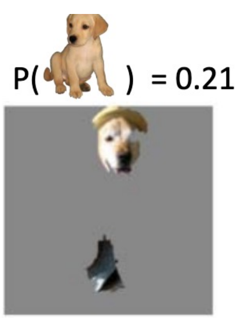
  - There are some tricks necessary to make this computationally feasible
- Tells you what part of input correlates most with a certain label



Electric guitar - incorrect but reasonable, similar fretboard



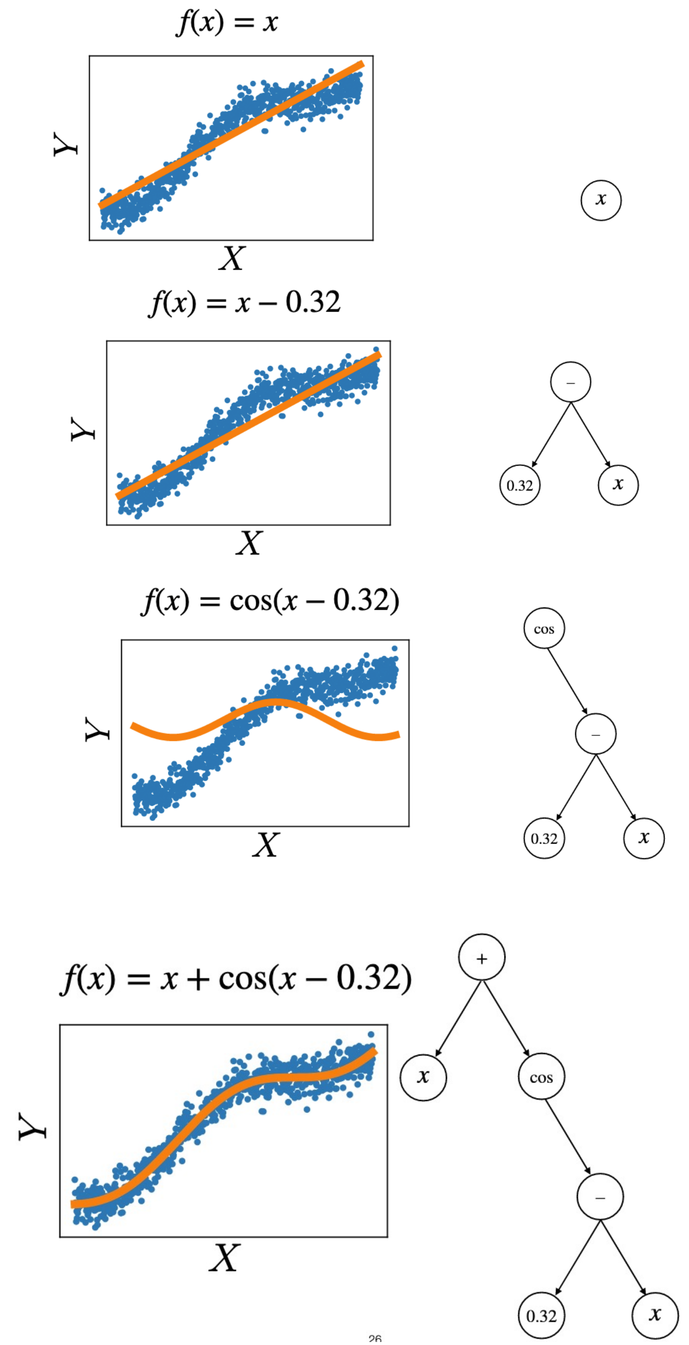
Acoustic guitar



Labrador

Symbolic Regression

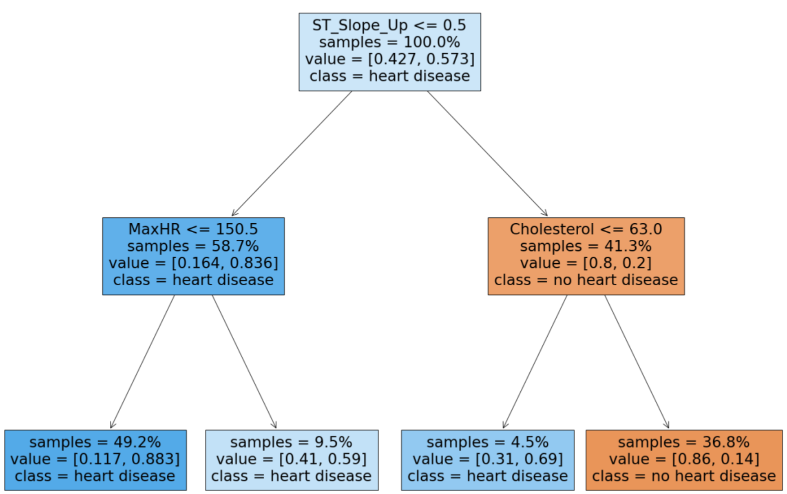
- Finds an analytic equation that mimics the predictions of a trained ML model
  - Find the analytic function that maps your inputs to the outputs of your model
  - By cleverly setting up your ML model you reduce the space of functions to search over
- Typically done with a genetic algorithm
  - Recursively build a function using basis space of input variables, operators, and constants (through crossover and mutation)
  - Minimize error between function and ML prediction
  - Result is a set of possible equations
  - Can enforce constraints like penalizing complexity

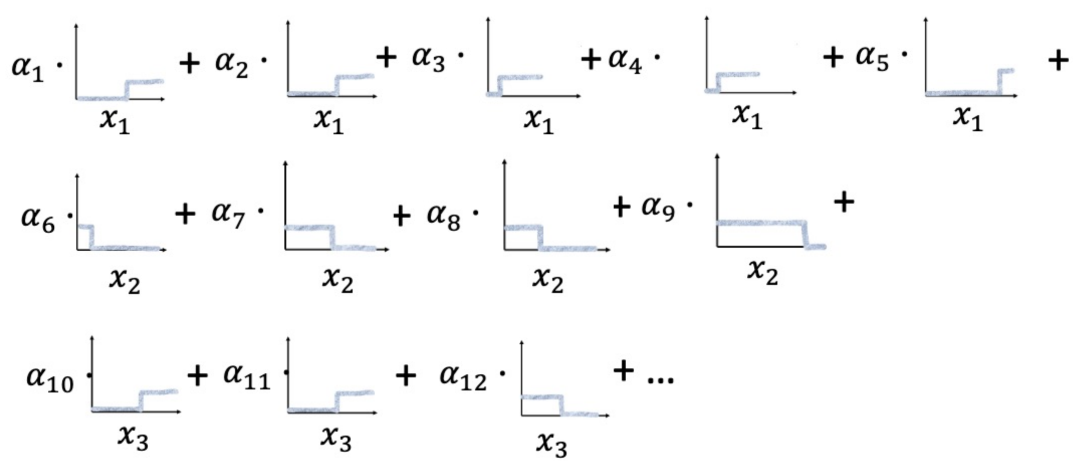


Alternative Methods

Inherently Explainable Models

- Some types of ML models are inherently interpretable
  - Linear and logistic regression, decision trees (~boosted decision trees)
- Also: Generalized additive models y= f1(x1)+f2(x2) +... +fn(xn)
  - Can include pairwise interactions, can be made sparse for easier computation
- In what cases does it make sense to use an interpretable model over a black box one?





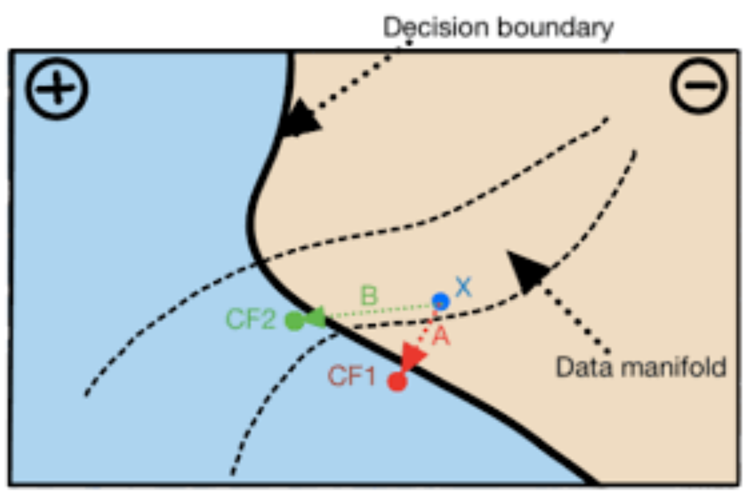
The Space of Good Models

- It's often said that there is a trade-off between interpretability and predictive accuracy
  - But there is some evidence that this is not always the case
- Interesting line of work seeks to characterize the space of good models

Model Adjustments + Input Variability

- How much do equally good models change particular predictions
  - Finding different models with equivalent accuracies can help estimate how reliably your explanations relate to the real world
  - Important implications for fairness: when errors happen to the same groups
- Counterfactual generation seeks to efficiently generate a related data point that would change a model's prediction
  - Various methods for computing 'nearby' points
  - Good option for providing actionable insight

Data	Delta	Ambiguity (Epsilon 0.5%)	Ambiguity (Epsilon 1%)	Ambiguity (Epsilon 2%)	Ambiguity (Epsilon 5%)	Discrepancy (Epsilon 0.5%)	Discrepancy (Epsilon 1%)	Discrepancy (Epsilon 2%)	Discrepancy (Epsilon 5%)
Breast Cancer	40%	0.0%	3.3%	11.7%	62.9%	0.4%	1.4%	1.8%	8.1%
	30%	3.1%	10.3%	38.3%	89.6%	1.4%	2.0%	4.3%	16.0%
	20%	10.5%	35.3%	77.1%	100.0%	1.7%	3.6%	12.6%	36.3%
	10%	55.9%	82.7%	100.0%	100.0%	13.8%	22.4%	51.8%	100.0%
Sleep Apnea	40%	2.5%	22.1%	84.0%	99.9%	0.0%	0.1%	0.7%	2.0%
	30%	16.9%	78.7%	97.7%	100.0%	0.0%	0.3%	1.2%	6.5%
	20%	78.4%	95.8%	99.9%	100.0%	0.2%	1.2%	4.6%	20.9%
	10%	97.8%	99.9%	100.0%	100.0%	4.7%	13.9%	33.1%	76.2%
Rearrest For crime	40%	2.6%	5.3%	15.3%	72.0%	0.9%	1.3%	3.1%	6.3%
	30%	5.1%	13.7%	41.3%	100.0%	1.4%	2.3%	5.7%	11.4%
	20%	16.3%	51.4%	100.0%	100.0%	2.5%	5.4%	11.4%	25.9%
	10%	94.0%	100.0%	100.0%	100.0%	5.4%	19.0%	37.8%	100.0%



Paper

How Do We Use Explanations?

Example: COVID Vulnerability

Based on diverse literature, we built complimentary suite of vulnerability measures to help quantitatively describe different types of community needs and understand how communities build resilience

Risk of Severe COVID Complications

Indicator	Weight
number of covid cases	1
% adults 65 and older*	4
diabetes	4
obesity	4
cardiovascular conditions	4
hypertension☆	4
respiratory conditions✦	3
smokers	1

Risk of Economic Harm

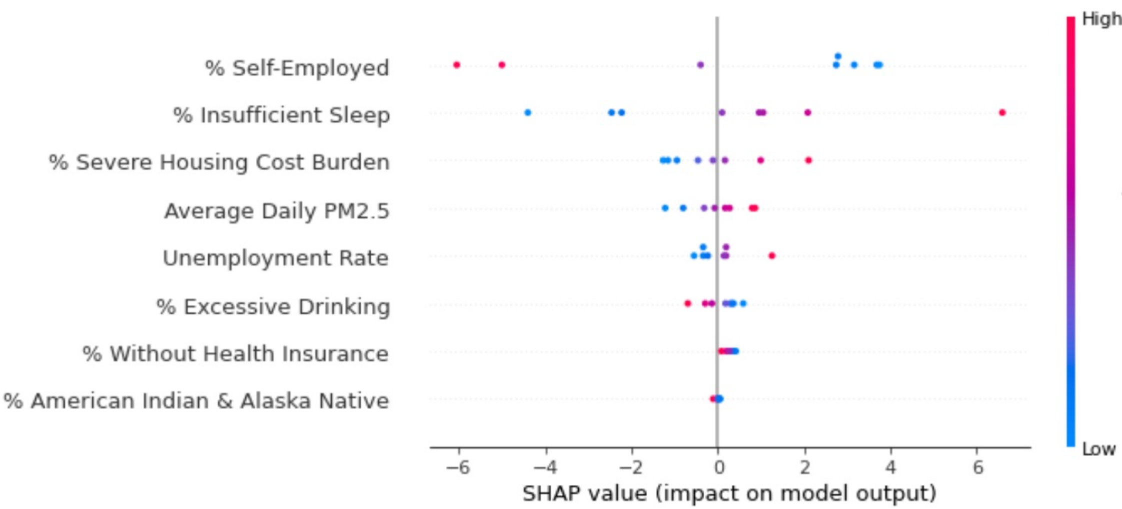
Indicator	Weight
Below poverty	1
Median income	1
No college degree (ages 25+)	1
Unemployed (16+)	1
Not in labor force but working age	1
% jobs in Tourism/Leisure/Hospitality	1
% part-time workers	1
% self-employed	1
Regional GDP per capita	1
Population change/migration	1

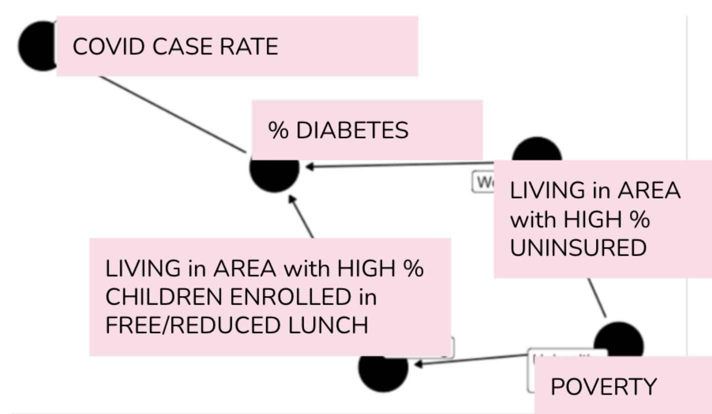
Need for Mobile Health Resources

Indicator	Weight
% rural / rurality	3
% household without a car	2
% using public transportation	-2
ratio of primary care providers	-3
% uninsured residents	2
% non-white residents	1
% non-English speaking residents	2
% veterans	1
% adults 65 and older*	2
% disabled residents	2
% opioid use	1
% fair or poor health	1
number of hospitals	-3

Validation Study

- How do you understand how accurate these models are when you can't directly measure community risk or need?
- One option is testing the ability of our model to predict a related proxy outcome
  - Can test if the metric features are predictive of the proxy outcome and investigate other predictive features
- Used SHAP values to study feature importance
- Requires domain knowledge at every step!

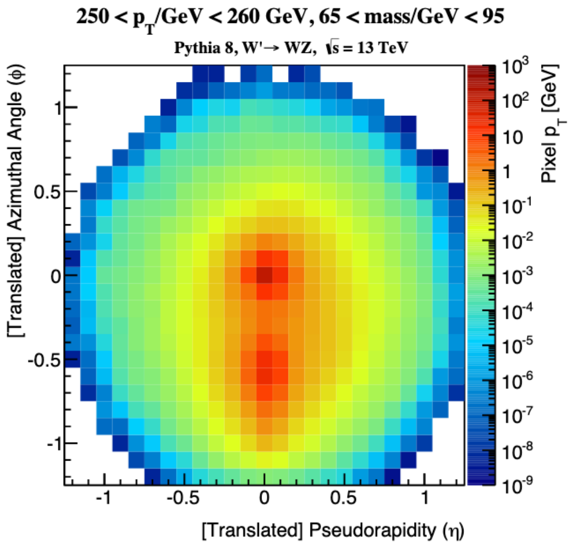


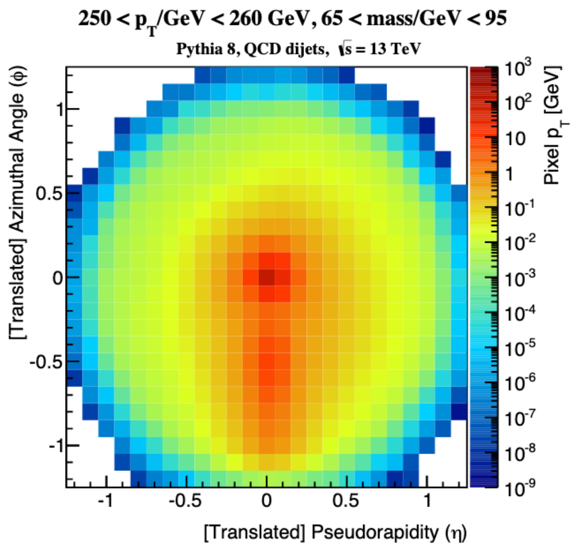


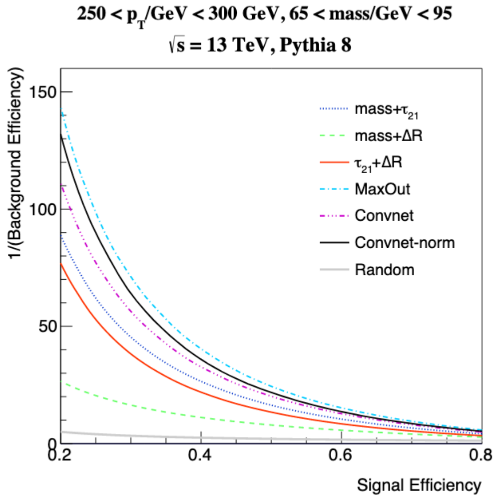
[Paper]

Particles as Images

- Heavy particles hadronize into collimated sprays (jets) and are absorbed in the granular calorimeter
  - Want to distinguish different types of jets based on their energy patterns
- Can achieve higher classification accuracy using CV
  - Standard approach uses cuts on physics-inspired features
- 'Unroll' the detector and map each cell to an image pixel
  - Apply preprocessing (normalization, rotation, translation) to standardize
  - Train CNN to classify jets (simplify to binary classification)



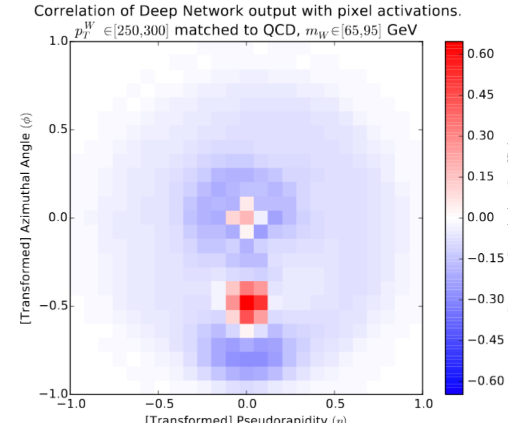


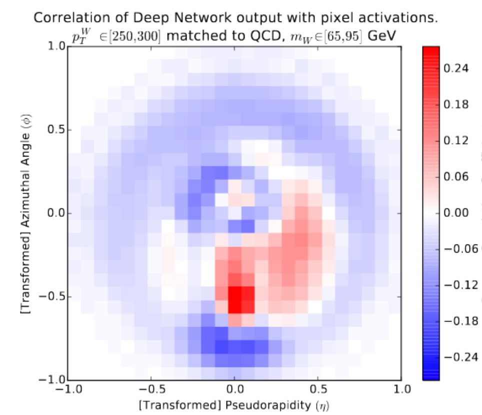


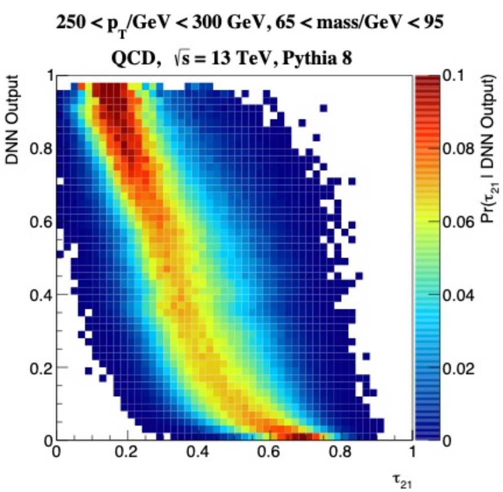
paper

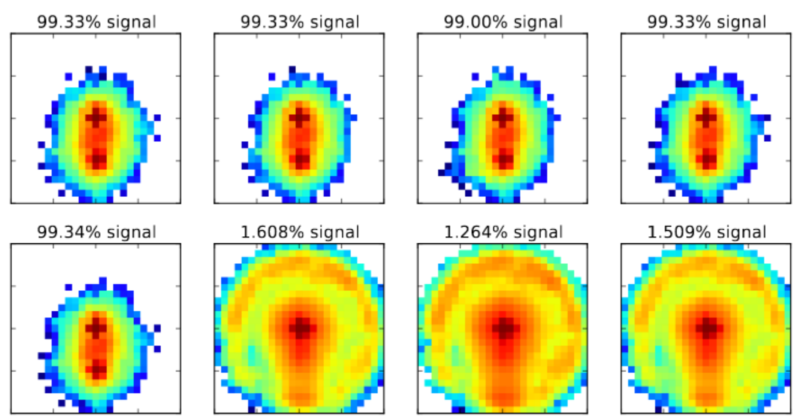
CNN Relevance

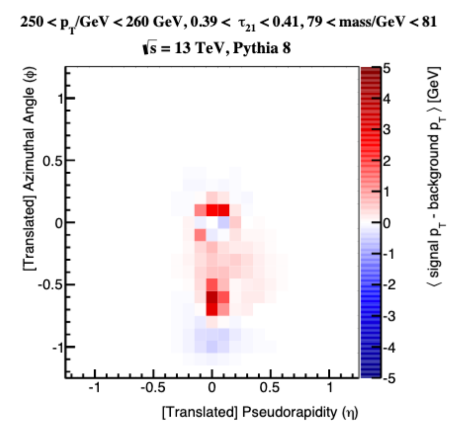
- Look at correlation of CNN output with standard physics features → it's learning thing we expect to be important
- Look at average of images with highest activations for last hidden layer → presence of secondary core is informative
- Look at per pixel correlation with CNN output (doesn't map to a known physical function)
  - Reweight samples to remove known physical variables → the radiation around the second core seems to matter
  - Look at only jets with W-like mass → radiation between cores seems to matter → learning about color flow?









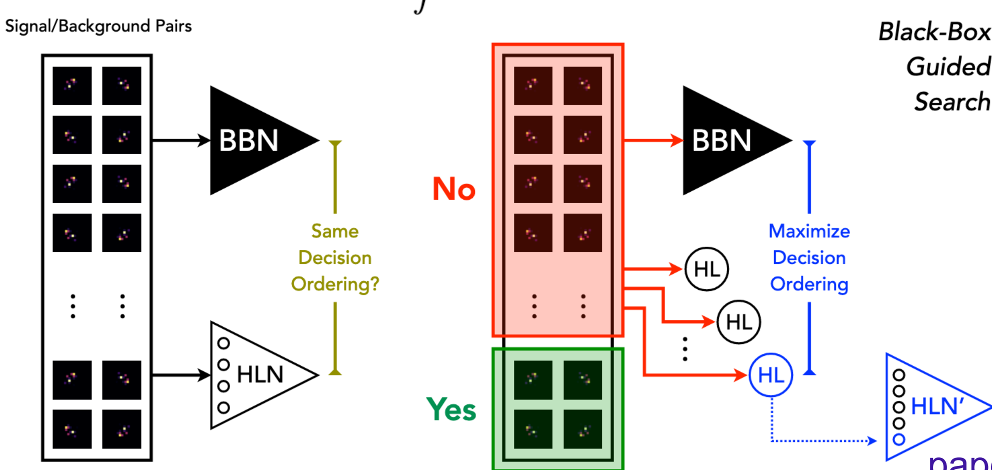


Local Approximators

- Use a CNN trained on low level information (jet images) to guide the construction of a simplified classifier based on high level interpretable features
  - Use average decision ordering to maximize the similarity between the decision boundaries of the two models
  - Use a black box guided search: iteratively selecting HL features that maximize ADO with the LL classifier
  - At each search step separate samples where HL and LL classifiers disagree
- The bulk of the CNN's power can be captured by 6 known jet features

$$\mathrm{DO}[f,g](x,x') = \Theta\Big(\big(f(x)-f(x')\big)\big(g(x)-g(x')\big)\Big)$$

$$\mathrm{ADO}[f,g] = \int dx\, dx'\, p_{\mathrm{sig}}(x)\, p_{\mathrm{bkg}}(x')\, \mathrm{DO}[f,g](x,x').$$

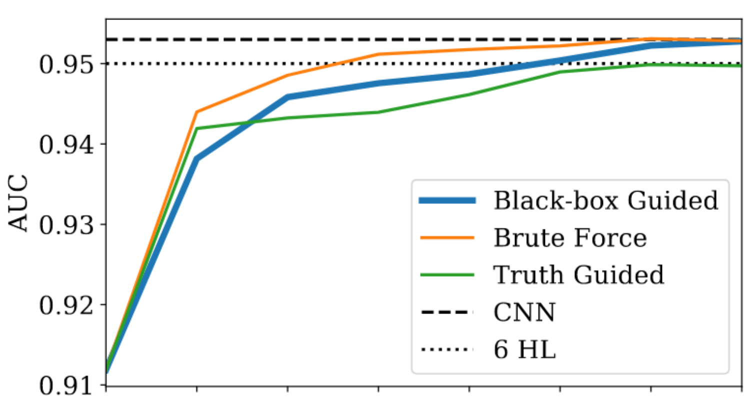


paper

Observable	AUC	ADO[CNN, Obs.]
M_{jet}	0.898 ± 0.004	0.807
$C_2^{\beta=1}$	0.660 ± 0.006	0.584
$C_2^{\beta=2}$	0.604 ± 0.007	0.548
$D_2^{\beta=1}$	0.790 ± 0.005	0.743
$D_2^{\beta=2}$	0.807 ± 0.005	0.762
$\tau_2^{\beta=1}$	0.662 ± 0.006	0.600
6HL	0.9504 ± 0.0002	0.971
CNN	0.9531 ± 0.0002	1.000
488HL	0.9535 ± 0.0002	0.978
$7\mathrm{HL}_{\text{black-box}}$	0.9528 ± 0.0003	0.971

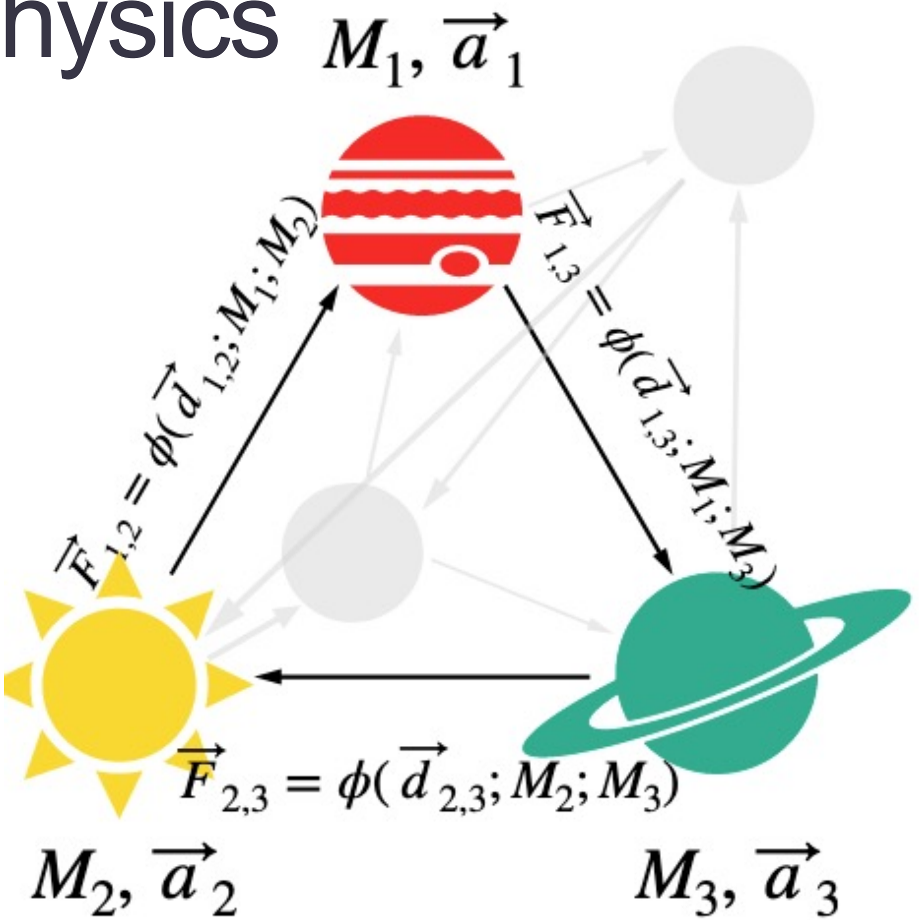
Even Better Local Approximators

- Define a basis space that captures a broad spectrum of physically interpretable information
  - Energy Flow Polynomials (EFPs): functions of momentum fraction of calorimeter cell and pairwise angular distance between cells
- Define a subspace of samples where 6-feature NN did not match CNN performance and search for EFP with max ADO
  - Identifies a new EFP that seems to help on edge cases
- Can use black box guided search directly on space of EFPs
  - Some EFPs identified are substantially different than traditional jet features



Iteration (n)	EFP	κ	β	Chrom #	$\text{ADO}[\text{EFP},\text{CNN}]_{X_{n-1}}$	AUC[EFP]	$\text{ADO}[\text{HLN}_n,\text{CNN}]_{X_{\text{all}}}$	$\text{AUC}[\text{HLN}_n]$
0	$M_{\text{jet}} + p_{\text{T}}$	–	–	–	–	–	0.9259	0.9119
1		2	$\frac{1}{2}$	2	0.8144	0.8190	0.9570	0.9382
2		0	2	2	0.6377	0.8106	0.9673	0.9458
3		0	–	1	0.5460	0.6737	0.9692	0.9476
4		1	$\frac{1}{2}$	2	0.5274	0.8464	0.9712	0.9487
5		−1	–	1	0.5450	0.5882	0.9714	0.9504
6		1	$\frac{1}{2}$	4	0.5382	0.7678	0.9734	0.9523
7		−1	$\frac{1}{2}$	2	0.5561	0.5957	0.9741	0.9528

Learning Astrophysics

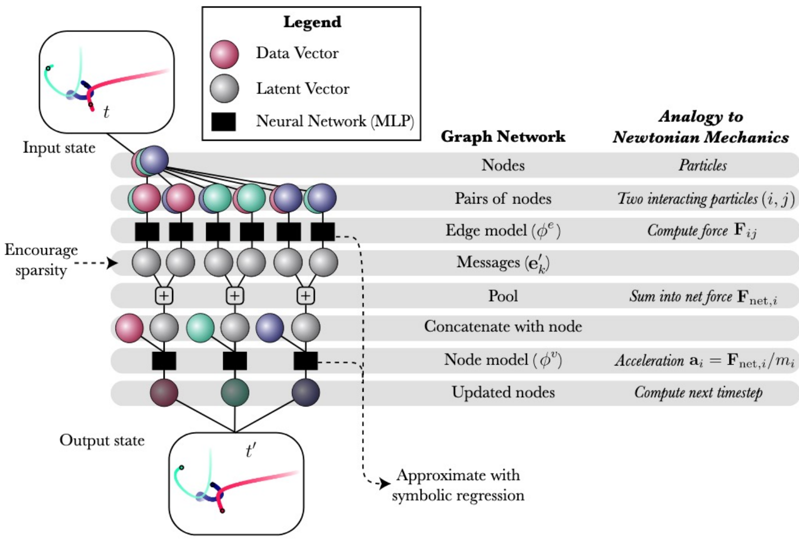


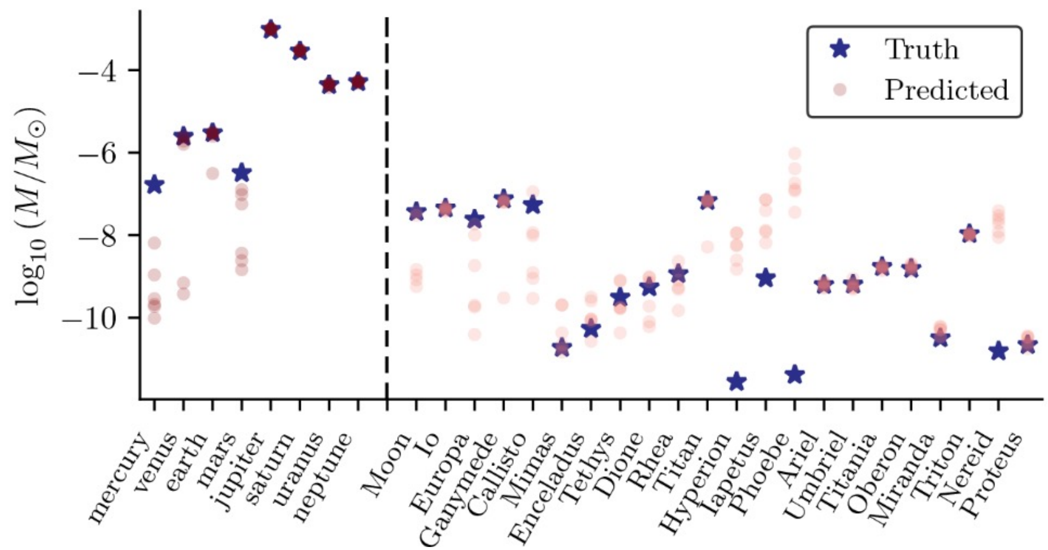
1. Our inputs are the positions of the bodies
2. They are converted into pairwise distances
3. **Our model tries to guess a mass for each body**
4. **It then also guesses a force, that is a function of distance and masses**
5. **Using Newton's laws of motion ($\sum \vec{F} = M\vec{a}$) it converts the forces into accelerations**
6. **Finally, it compares this predicted acceleration, with the true acceleration from the data**

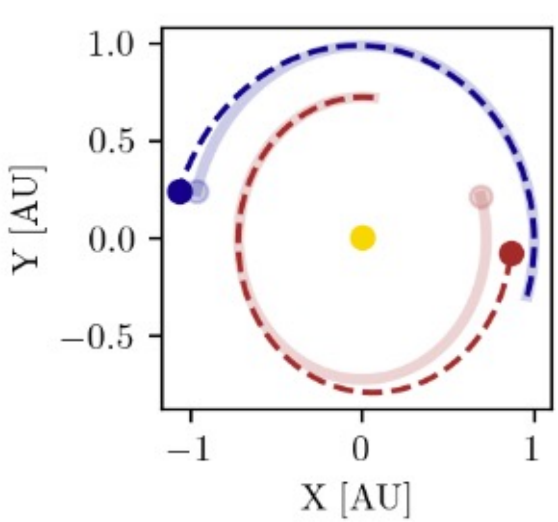
Minimize
$$\left| \vec{a}(\text{pred}) - \vec{a}(\text{true}) \right|^2$$

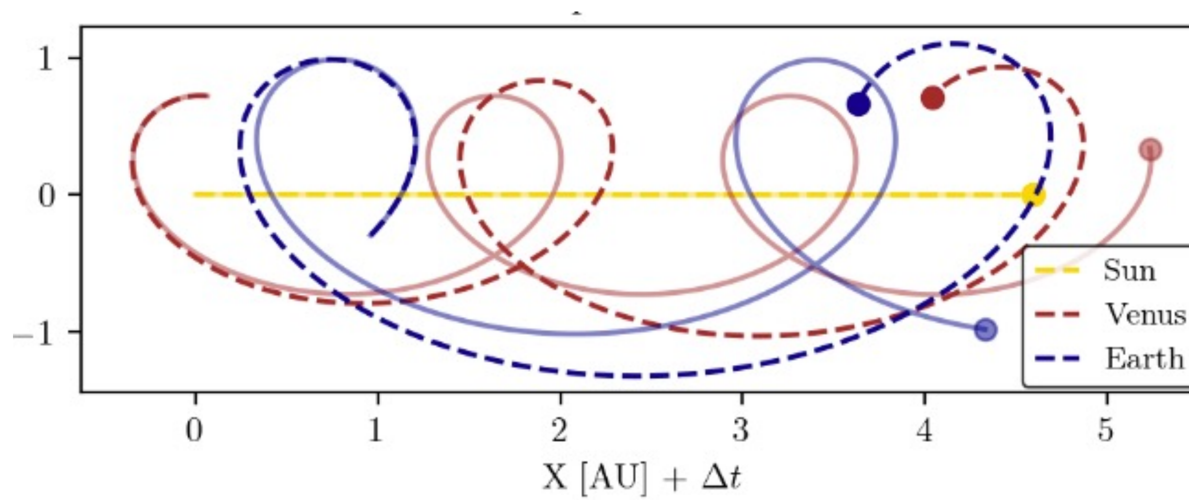
paper

Learning Astrophysics

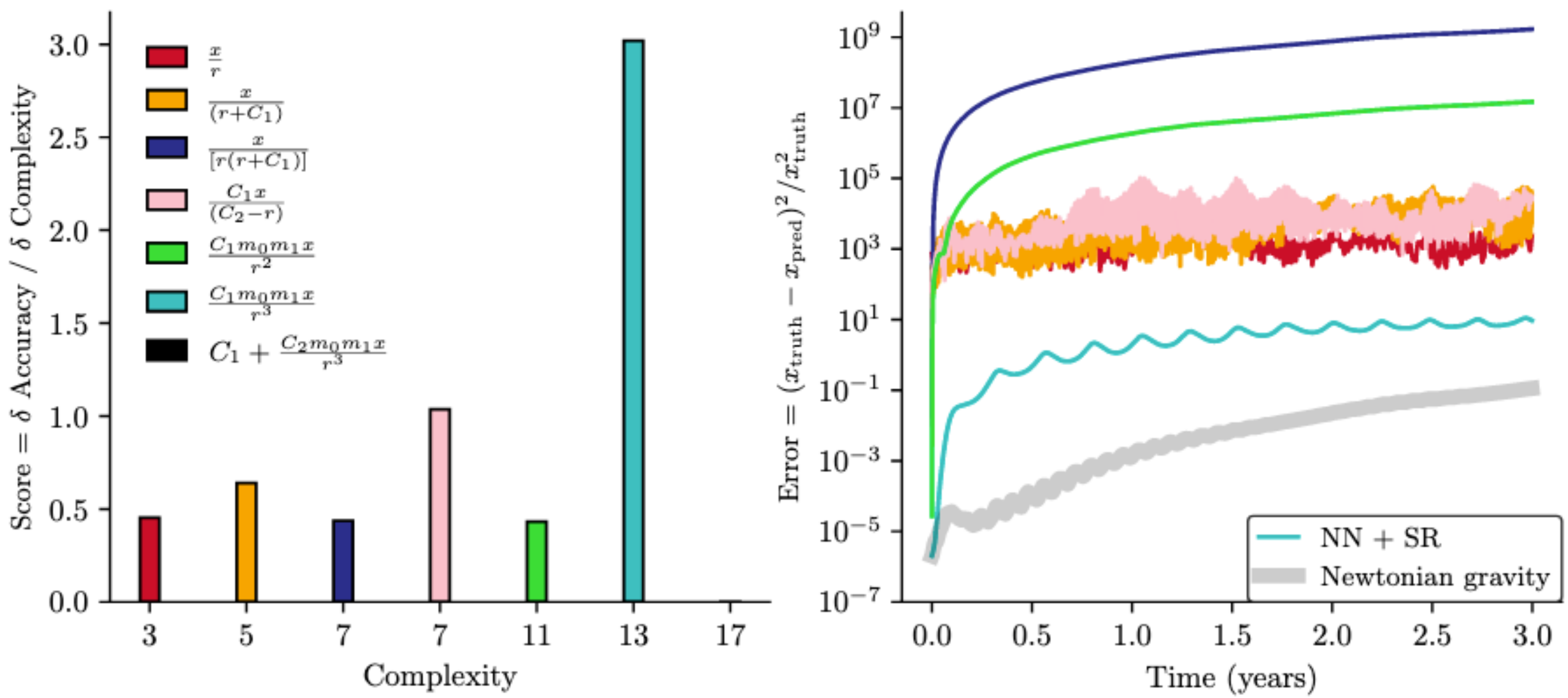








Symbolic Regression



- Apply symbolic regression to the GNN messages (forces) with a constraint to balance accuracy and equation complexity
- Can substitute learned equation for the force guess to improve the simulator and the mass predictions (node predictions)

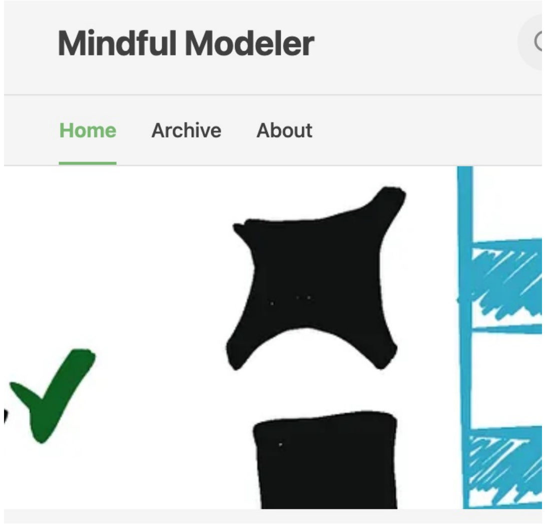
Implications/Limitations of Explainability in Physics

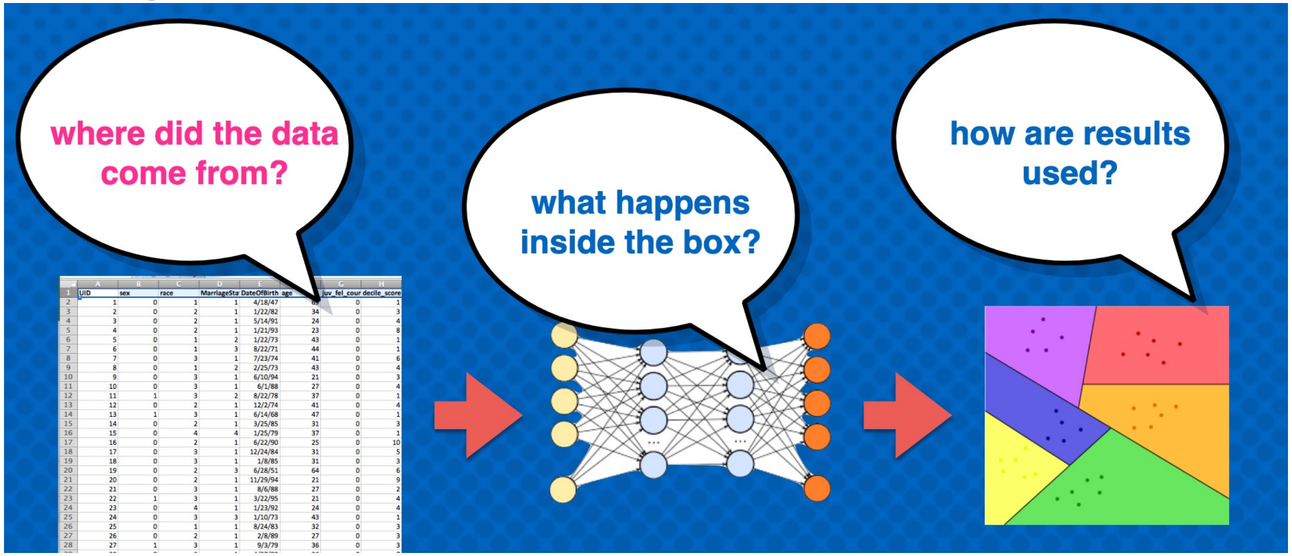
- For relevance propagation now way to know if relevances are due to true physics or statistical artifacts
  - Difficult to map relevances to mathematical information
- These methods don't characterize performance on edge cases or difficult samples
- Many problems don't have a nice basis space of features to search over for local approximators
  - These bases don't provide full coverage, unable to characterize other learned information
- For symbolic regression how do we decide which equation to pick?
  - Simplicity of an equation as a decision factor is a big assumption
- How do you account for uncertainties/mismodelings in the synthetic data or reconstruction software
  - Is the ML model decision actually describing nature

Explainability Isn’t Enough

Limitations

- Even with a fully explainable model, things can still go wrong:
  - Underlying statistical assumptions that may not be met
  - There is no promise that your model reflects the real world
  - Explainability techniques cannot fix issues/biases in training data
  - Correlation doesn't imply causation
- In order for domain experts to help you evaluate models and for the public to develop trust (when founded) in models, **we need transparency**





Documentation: Datasheets and Nutrition Labels

- Meaningfully documenting datasets promotes reflection about how datasets affect models
  - Helps creators think through assumptions, risks, and implications
  - Helps consumers make informed decisions about limitations
  - Maintains information about where data came from

paper

paper

Questions

Motivation

- **For what purpose was the dataset created?** Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.
- **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**
- **Who funded the creation of the dataset?** If there is an associated grant, please provide the name of the grantor and the grant name and number.
- **Any other comments?**

Composition

- **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?** Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

Collection Process

- **How was the data associated with each instance acquired?** Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
- **What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?** How were these mechanisms or procedures validated?
- **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**
- **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**
- **Over what timeframe was the data collected?** Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.
- **Were any ethical review processes conducted (e.g., by an institutional review board)?** If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.
- **Does the dataset relate to people?** If not, you may skip the remaining questions in this section.
- **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**

Documentation: Model Cards

- Documenting models promotes reflection about model design choices, limitations, and evaluations
  - Describes under what conditions and model is expected to perform well
  - Provides guidance and caution for responsible use
  - Can help build technical literacy in the public and media

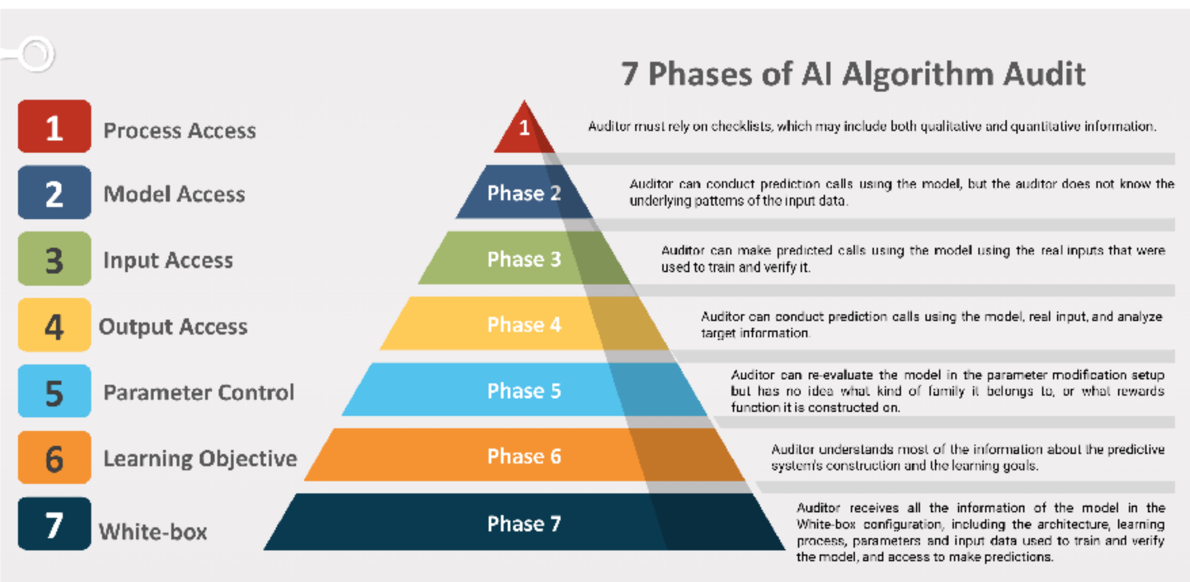
Model Card

- **Model Details**. Basic information about the model.
  - Person or organization developing model
  - Model date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more information
  - Citation details
  - License
  - Where to send questions or comments about the model
- **Intended Use**. Use cases that were envisioned during development.
  - Primary intended uses
  - Primary intended users
  - Out-of-scope use cases
- **Factors**. Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3.
  - Relevant factors
  - Evaluation factors
- **Metrics**. Metrics should be chosen to reflect potential real-world impacts of the model.
  - Model performance measures
  - Decision thresholds
  - Variation approaches
- **Evaluation Data**. Details on the dataset(s) used for the quantitative analyses in the card.
  - Datasets
  - Motivation
  - Preprocessing
- **Training Data**. May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.
- **Quantitative Analyses**
  - Unitary results
  - Intersectional results
- **Ethical Considerations**
- **Caveats and Recommendations**

paper

Model Auditing/Impact Assessments

- Like we saw in the first hands on exercise, it's important to exhaustively test models before they're deployed
  - Lots of scholarship in AI Ethics promotes external model audits
  - Requires difficult conversations about regulation and trade secrecy
- Algorithmic impact statements aim to force consideration of the problem before design, construction, and deployment
  - Impact assessments seek to measure the effect of an algorithm



Algorithmic impact assessments

Algorithmic risk assessment	Algorithmic impact evaluation
Assessing possible societal impacts of an algorithmic system before the system is in use (with ongoing monitoring advised)	Assessing possible societal impacts of an algorithmic system on the users or population it affects after it is in use

Discussion

- What are the benefits of interpretability/explainability/transparency?
- What are the limitations or draw backs?
- Which techniques do you find the most compelling or useful?
- How should this kind of work inform regulation of AI?

Hands on Exercise Two