Understanding Model Predictions

Matthew Engelhard

Interpretable versus Explainable

Interpretable versus Explainable

An *interpretable* model:

It is easy to for us to understand why the model makes the predictions that it makes.

An *explainable* model:

One or more techniques can be used to provide a human-friendly explanation for each model prediction.

Interpretable

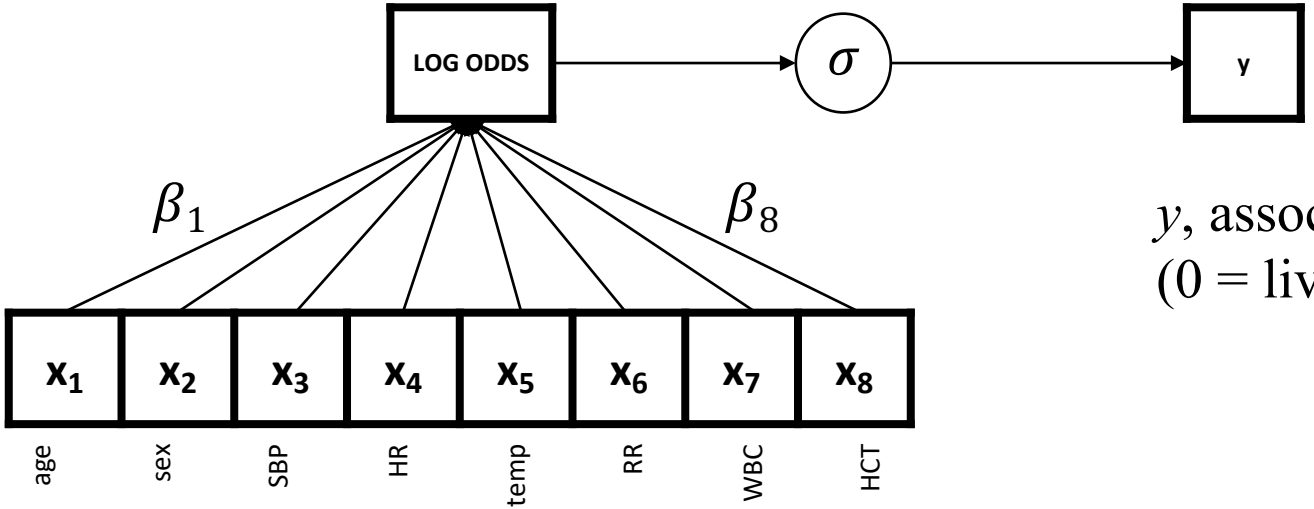
Explainable

We can *interpret* APACHE III

- Suppose a new patient is transferred to the ICU, and the (logistic regression) model predicts their mortality risk is high.
- **Q:** Is it hard to figure out why it made that prediction?

We can *interpret* APACHE III

- Suppose a new patient is transferred to the ICU, and the (logistic regression) model predicts their mortality risk is high.
- **Q:** Is it hard to figure out why it made that prediction?
- **A:** No. You can look at the coefficients to see which variables increased and decreased the predicted probability.



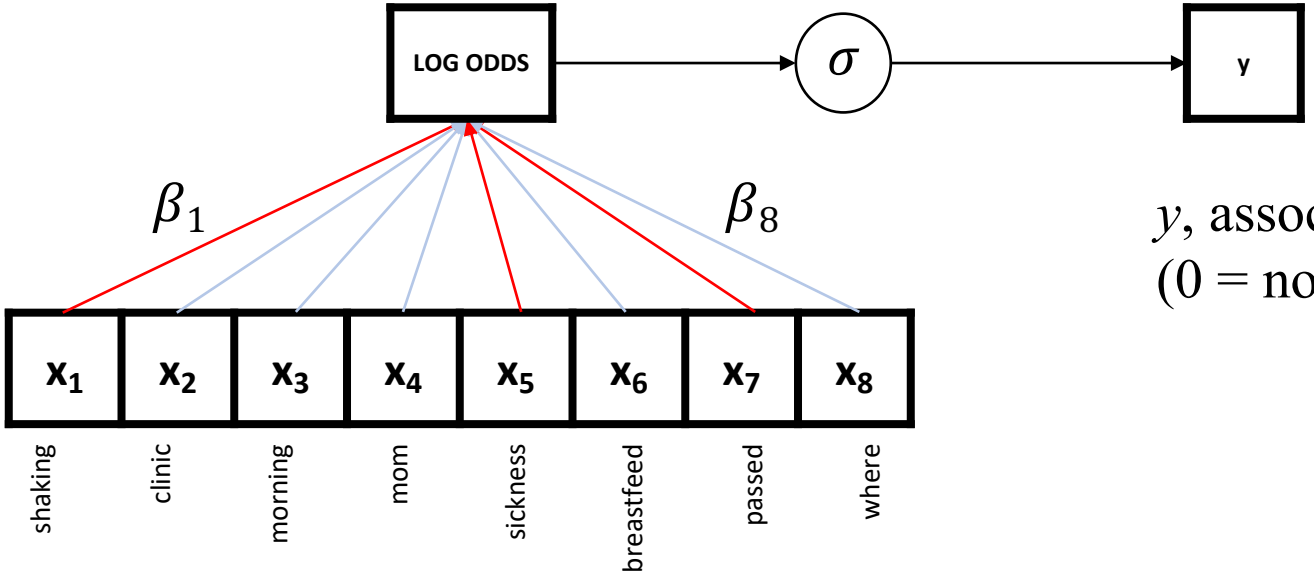
y, associated label:
(0 = live, 1 = die)

We can *interpret* a count-based NLP model

- Suppose you use logistic regression with count-based features, and your model predicts that that an SMS you receive is urgent.
- **Q:** Is it hard to figure out why it made that prediction?

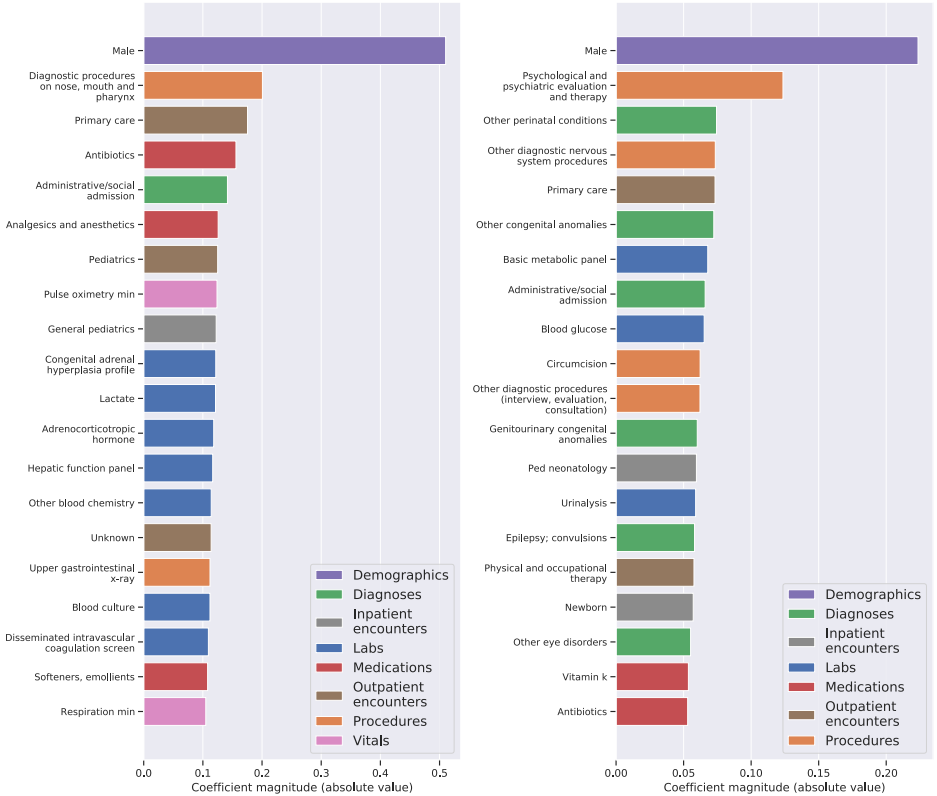
We can *interpret* a count-based NLP model

- Suppose you use logistic regression with count-based features, and your model predicts that that an SMS you receive is urgent.
- **Q:** Is it hard to figure out why it made that prediction?
- **A:** No. You can look at the coefficients to see which words increased and decreased the predicted probability.



y, associated label:
(0 = not urgent, 1 = urgent)

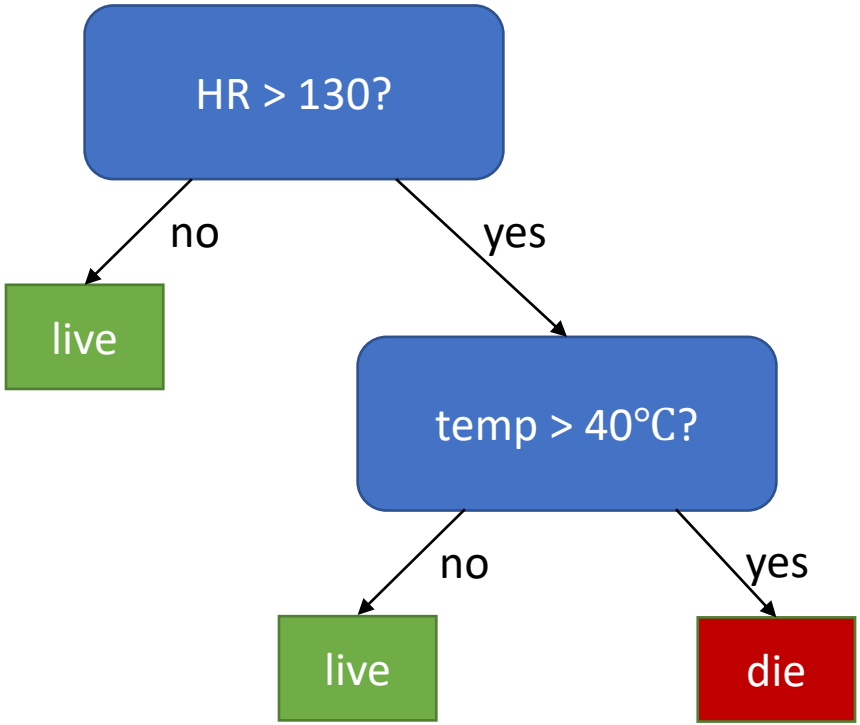
In linear models, parameter values show the effect of the corresponding feature.



Other Interpretable Models?

Other Interpretable Models?

- Decision Tree

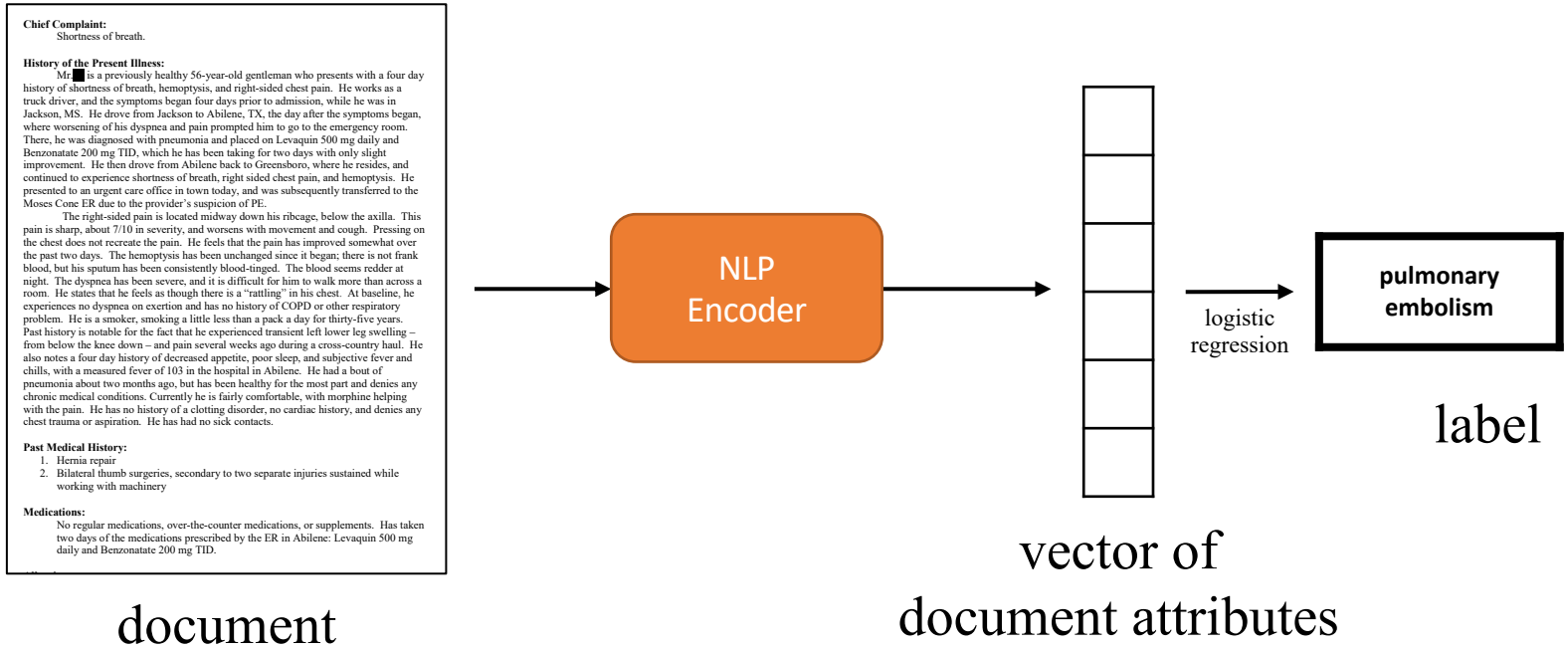


Can we *interpret* a deep learning NLP model?

- Suppose you apply a deep neural network to a sequence of word vectors, and your model predicts that that the SMS you receive is urgent.
- **Q:** Is it hard to figure out why it made that prediction?

Can we *interpret* a deep learning NLP model?

- Suppose you apply a deep neural network to a clinical note, and your model predicts it is describing a pulmonary embolism.
- **Q:** Is it hard to figure out why it made that prediction?



Can we *interpret* a deep learning NLP model?

- Suppose you apply a deep neural network to a sequence of word vectors, and your model predicts that that the SMS you receive is urgent.
- **Q:** Is it hard to figure out why it made that prediction?
- **A:** Yes, it's hard to figure out, because the effect of a given word on model predictions depends on the rest of the document.

- So, we can't directly interpret the model's predictions.
- **However, we have ways of asking the model to explain itself.**

Feature importance in complex models

- Permutation importance

  If I shuffle around the patients' ages, how much do model predictions change?

- Gradient-based methods

  As I slowly change the patients' ages, how quickly do model predictions change?

  (saliency maps)

  As I change the patient's age to a baseline value, how much do model predictions change?

  (integrated gradients)

- SHAP Values

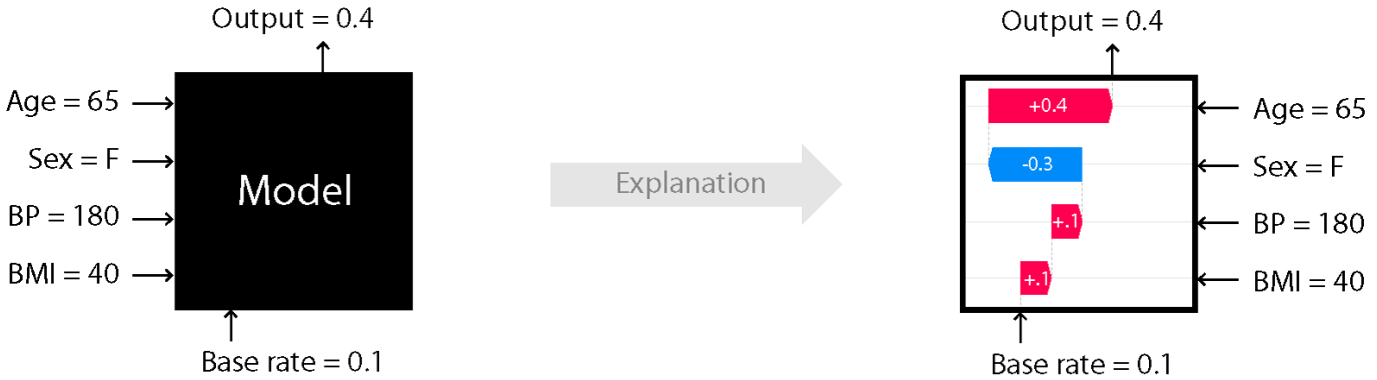
  If I excluded age from the model, how much would the prediction change?

Most common: SHAP Values

Key differences compared to interpretation:

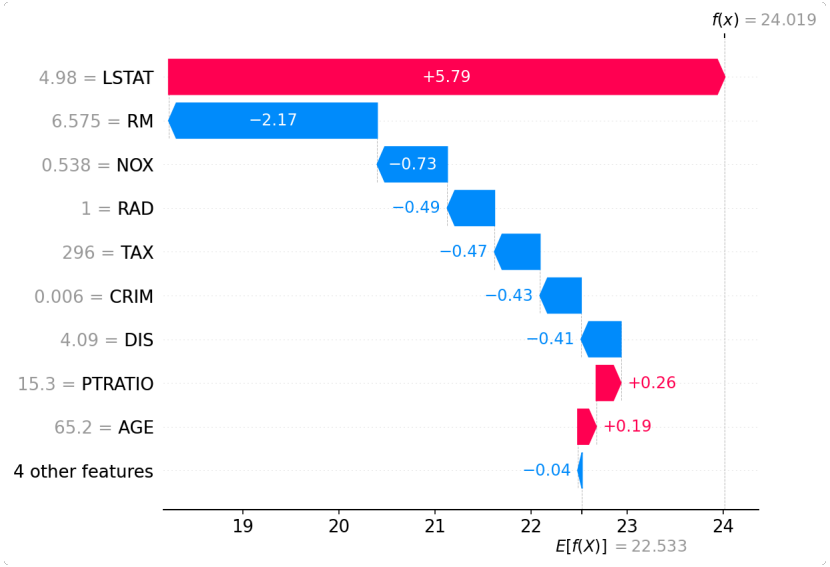
1. Explanations apply to individual predictions, not the model as a whole
2. Computing them can be computationally expensive

SHAP

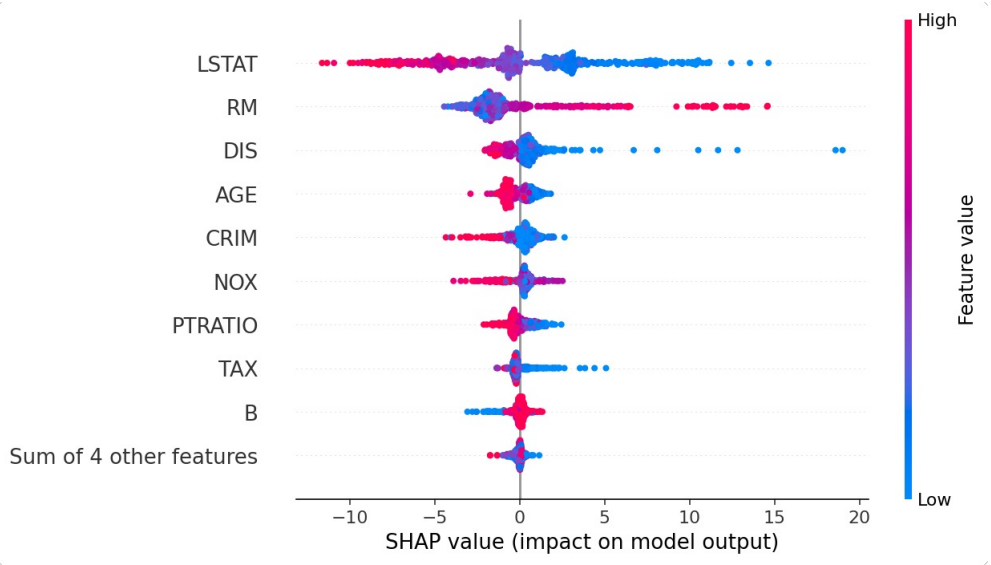


SHAP Values

For one prediction



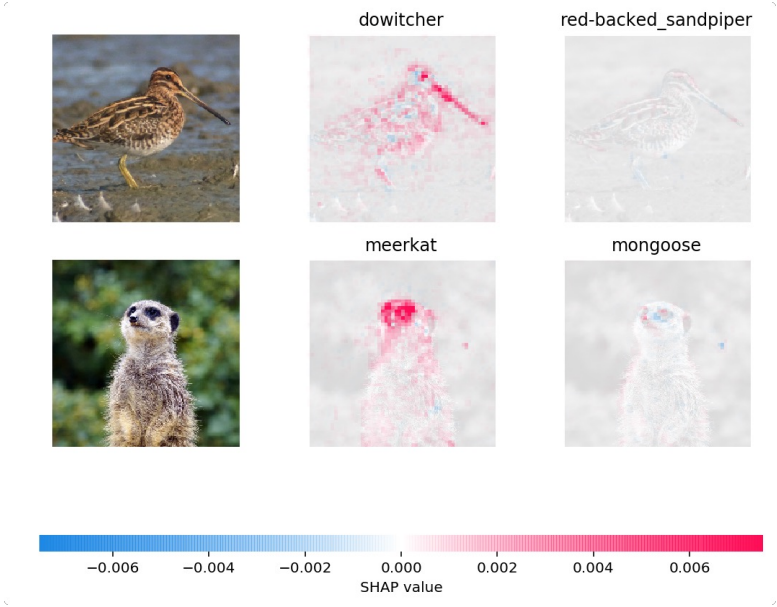
For all predictions



https://github.com/slundberg/shap

We can *explain* a CNN's predictions

- Suppose you apply a CNN to images of animals, and the model predicts that an animal is a meerkat.
- **Q:** Is it hard to figure out why it made that prediction?
- **A:** Yes, an explanation can be provided:



- Similarly, we can identify passages within a text document that would change predictions most if they were removed.

Caveats:

- This is for a single example; not clear how to summarize across all examples for images and text
- Can be very computationally expensive

Natural Language Processing: Which Text is Predictive?

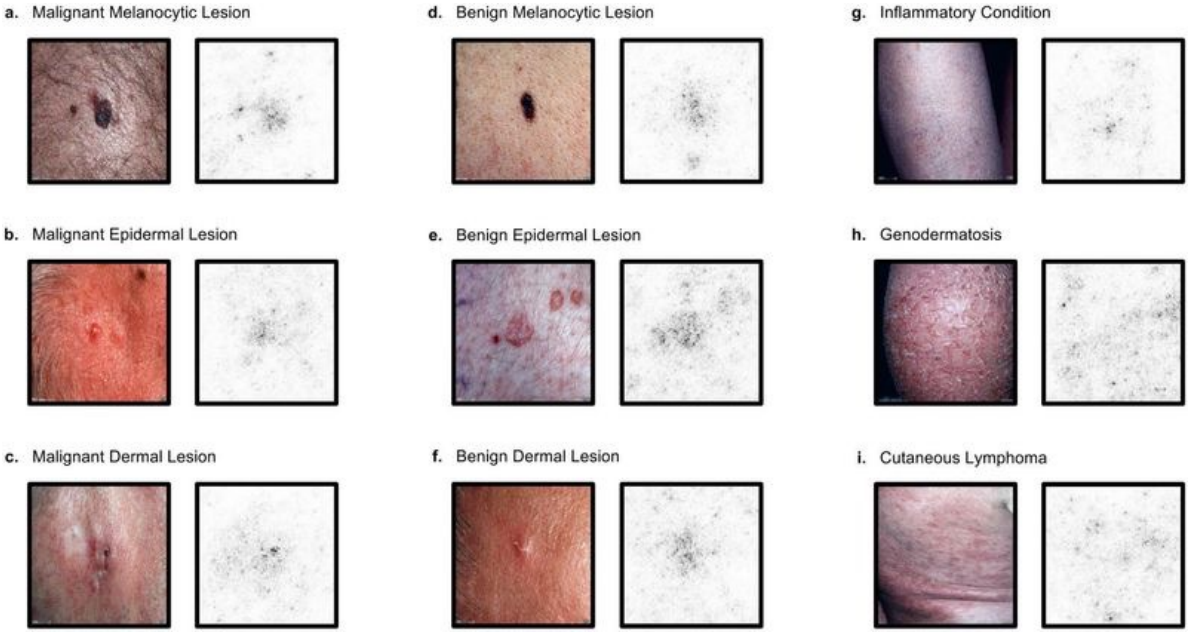
Passage (from note)	Change in predicted autism dx log-odds
subjective intake chief complaint problems with sleep, inattention, and behavioral concerns both in the home and school setting. DATE, recently more anger and recent tic like behavior	+6.95
psychologist presenting problem NAME is a 3 year, 4 month old female who was referred for a neurodevelopmental assessment due to concerns regarding her overall development, behavior, and social emotional functioning and to assess for autism spectrum disorder	+6.82
problem list diagnosis • disruptive behavior disorder • impaired speech articulation • daytime enuresis • other subjective visual disturbances • hypermetropia of both eyes • adhd attention deficit	+6.81
problem list diagnosis • anemia of prematurity • history of colitis • meconium tox for thc • extreme immaturity of newborn, 27 completed weeks • nasal congestion of newborn • presumed	+6.78
motor delay DATE • hypotonia DATE • clasped thumb DATE • polydactyly DATE • developmental	+6.74
therapy NAME was seen for developmental support during rop eye exam today. the	+6.65

Developmental and behavioral concerns are highly predictive

Premature birth and perinatal complications are also highly predictive

Subramanian V, Engelhard MM, Berchuck SI, Chen L, Carin L. SpanPredict: Extraction of Predictive Document Spans with Neural Attention. NAACL 2021

Saliency maps from Esteva et al.



Saliency maps show gradients for each pixel with respect to the CNN's loss function. Darker pixels represent those with more influence.

How would this compare to a dermatologist's explanation?

Interpretable ? Explainable

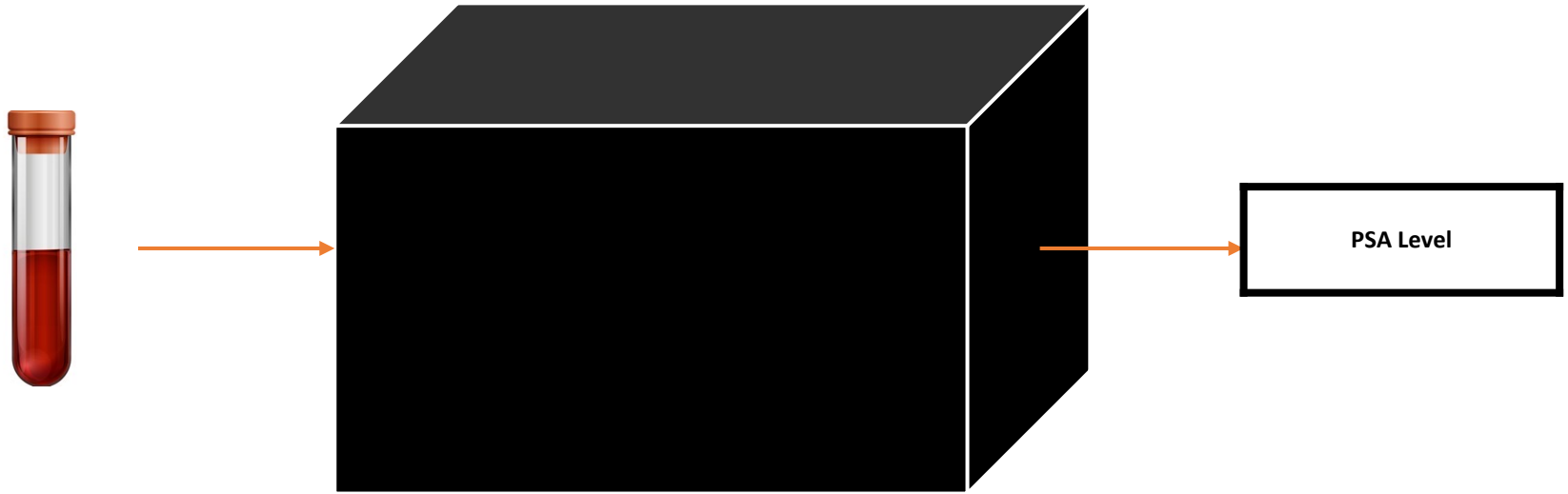
where does clinician decision-making fall on this spectrum?

Perspectives on Interpretability

Compared to simpler models, DNNs are a black box.

We can explain model predictions, but it's case by case.

Is PSA measurement a black box?

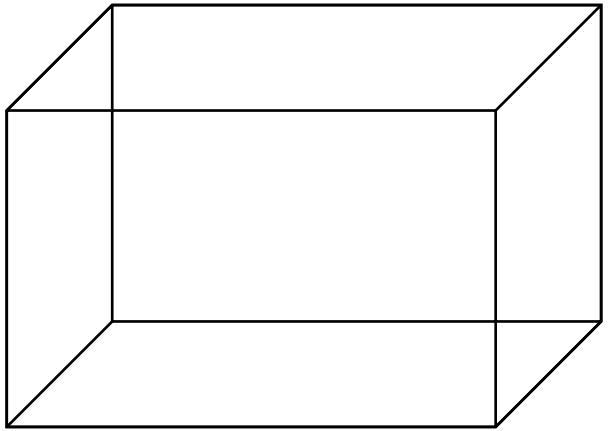


Is clinician decision-making a black box?

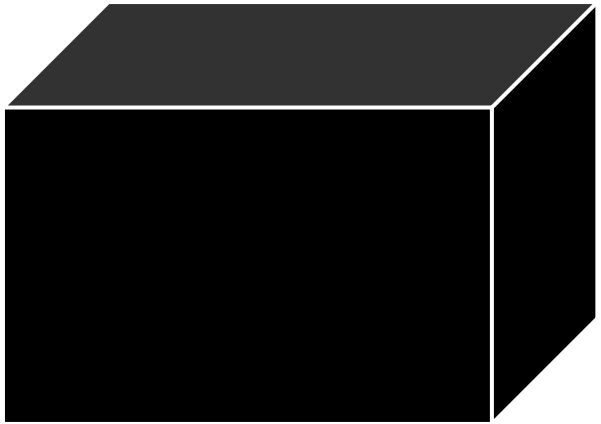
Two competing perspectives

Interpretability will be central to FDA regulation of clinical decision support.

We can only use tools if we fully understand how they work



We just need to make sure our tools are *valid* and *reliable*



Conclusions

- A predictive model is *interpretable* if it's easy to understand how it works – in other words, the effect of features on model predictions

- Neural networks are not *interpretable*, but we have techniques that can quantify the effect of features on each individual prediction – in other words, *explain* the prediction

- There are a wide range of perspectives on the importance of interpretability and explainability. However, it appears the FDA regulation of clinical decision support will turn on whether the tool is interpretable.

- Interpretable models still do not tell us anything about causation.