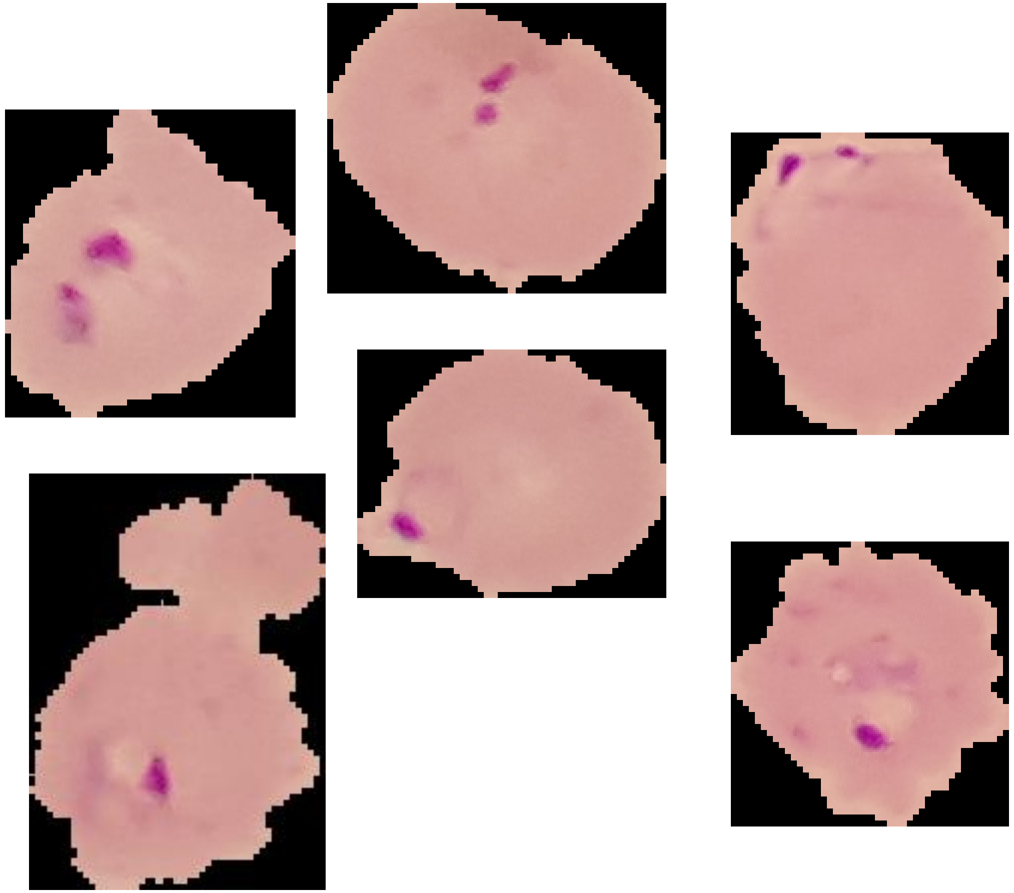
PyData London 2019

# Interpretable AI

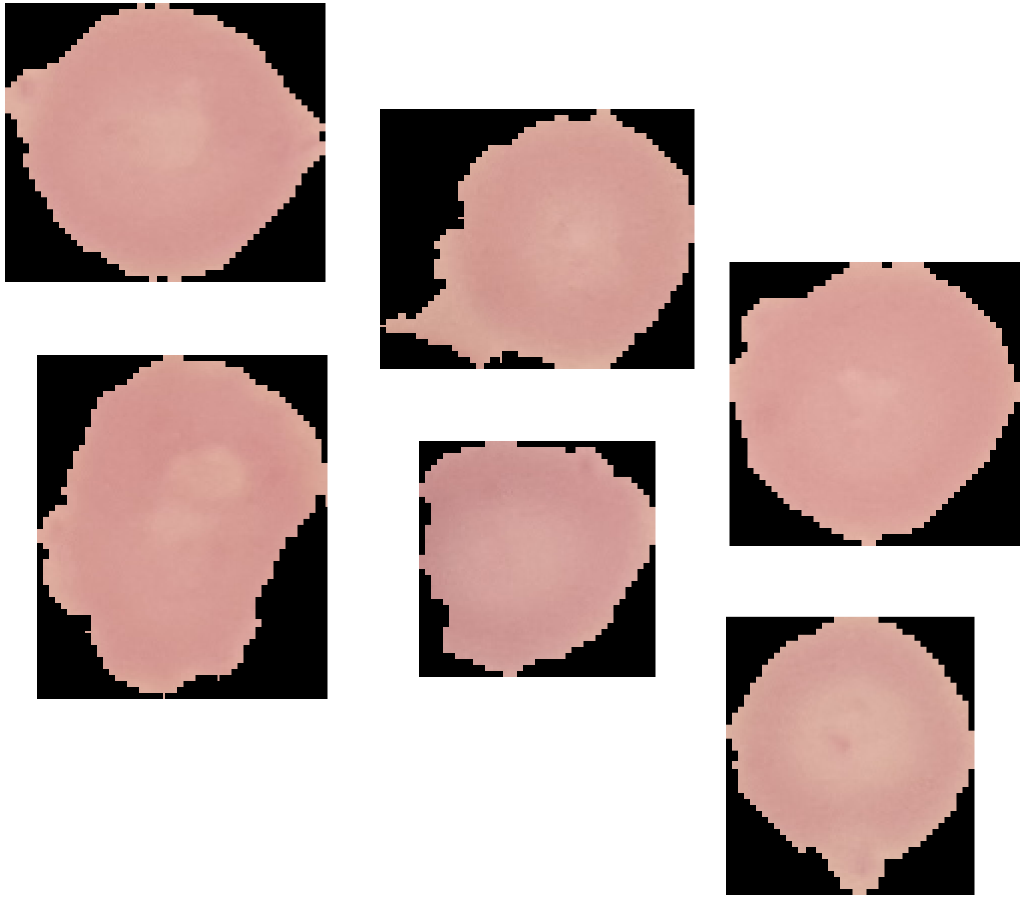
## How I Learned to Stop Worrying and Trust AI


**Ajay Thampi (PhD)**
Lead Data Scientist
Microsoft


# Learning – Detecting Malaria

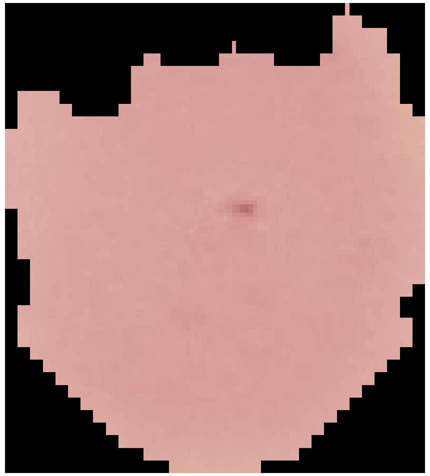


INFECTED



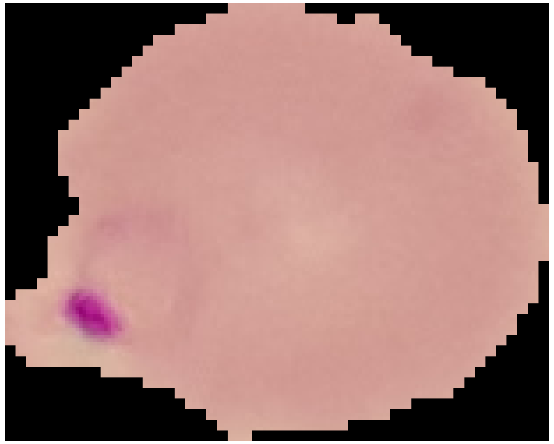
UNINFECTED

# Testing (1/3)



UNINFECTED

# Testing (2/3)

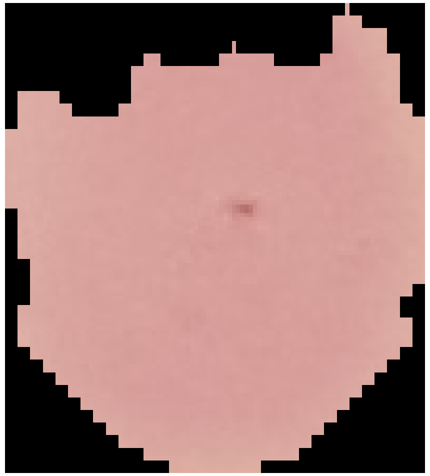


INFECTED

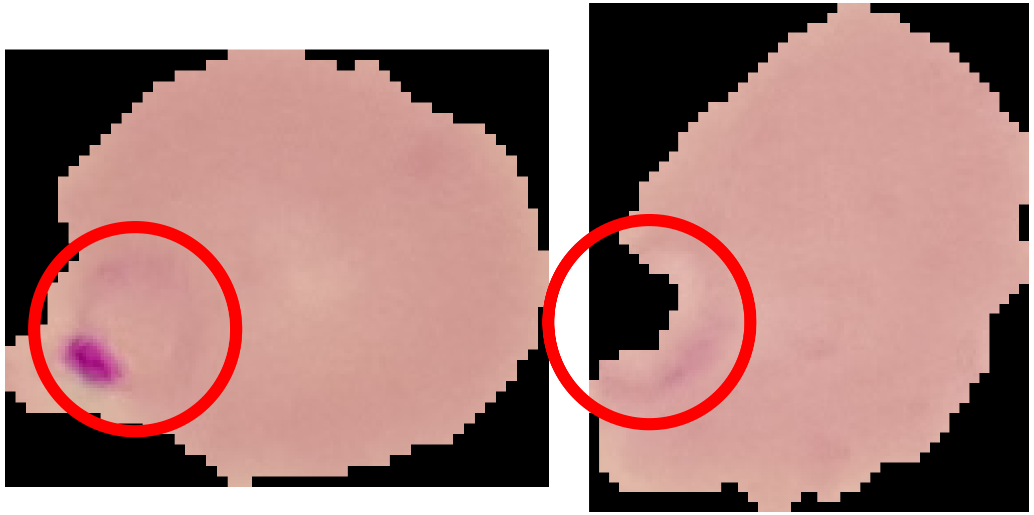
# Testing (3/3)

INFECTED

# Understanding

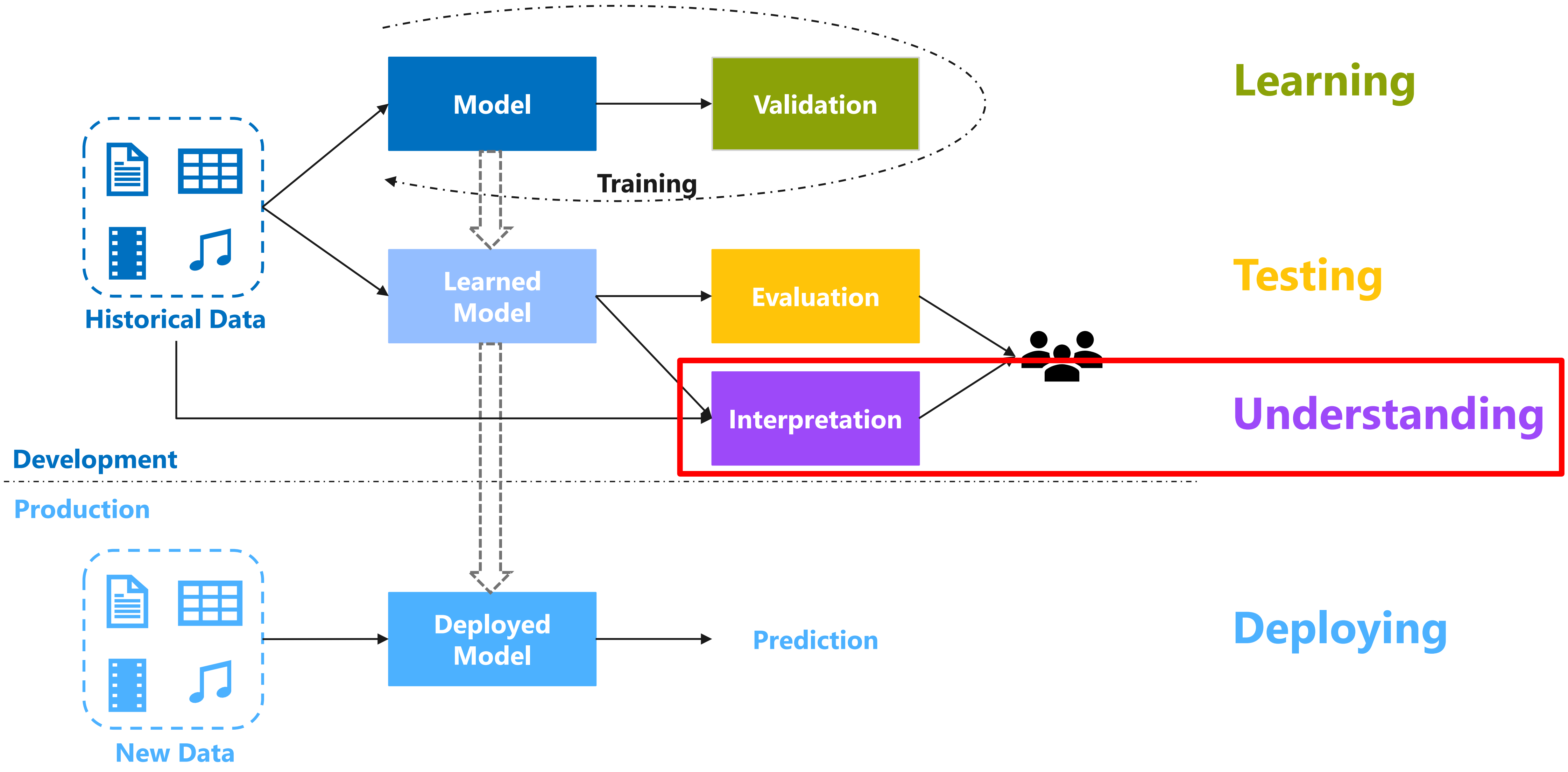


UNINFECTED



INFECTED

# Modelling and Deployment



# Transparent Models

# Decision Trees - Simple

## Iris Flower Dataset

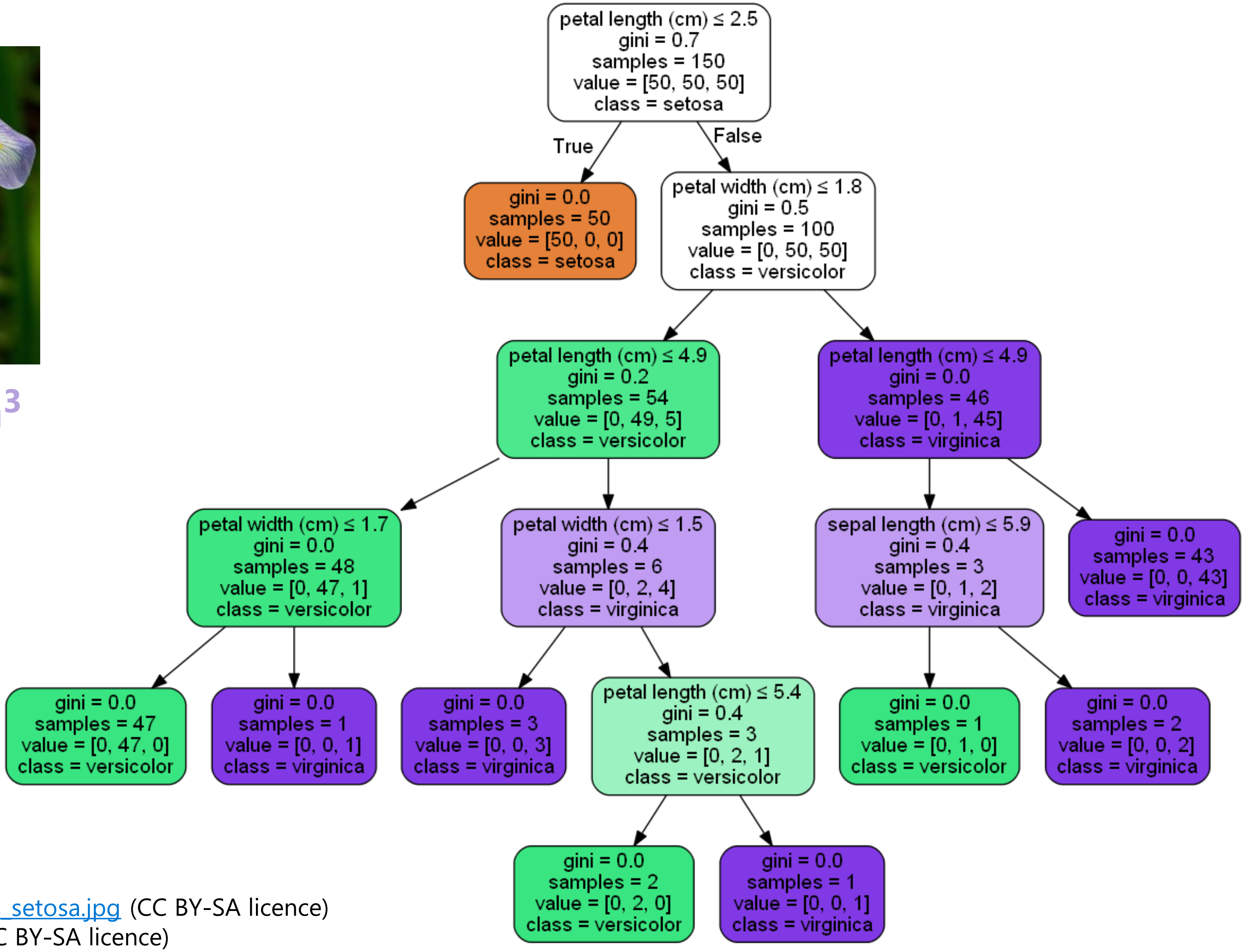
Setosa[1]

Versicolor[2]

Virginica[3]

**Features:**

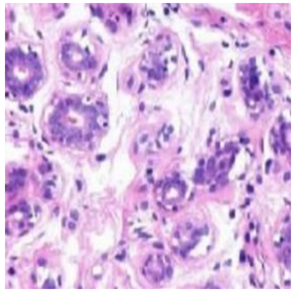
- Petal Length
- Petal Width
- Sepal Length
- Sepal Width



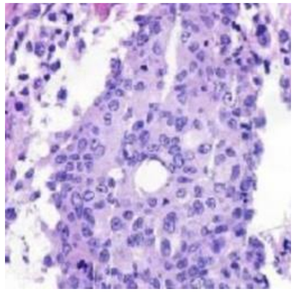

[1] https://en.wikipedia.org/wiki/Iris_setosa#/media/File:Kosaciec_szczecinkowaty_Iris_setosa.jpg (CC BY-SA licence)
[2] https://en.wikipedia.org/wiki/Iris_versicolor#/media/File:Blue_Flag,_Ottawa.jpg (CC BY-SA licence)
[3] https://en.wikipedia.org/wiki/Iris_virginica#/media/File:Iris_virginica_2.jpg (CC BY-SA licence)

# Decision Trees – More Complex

## Breast Cancer Dataset



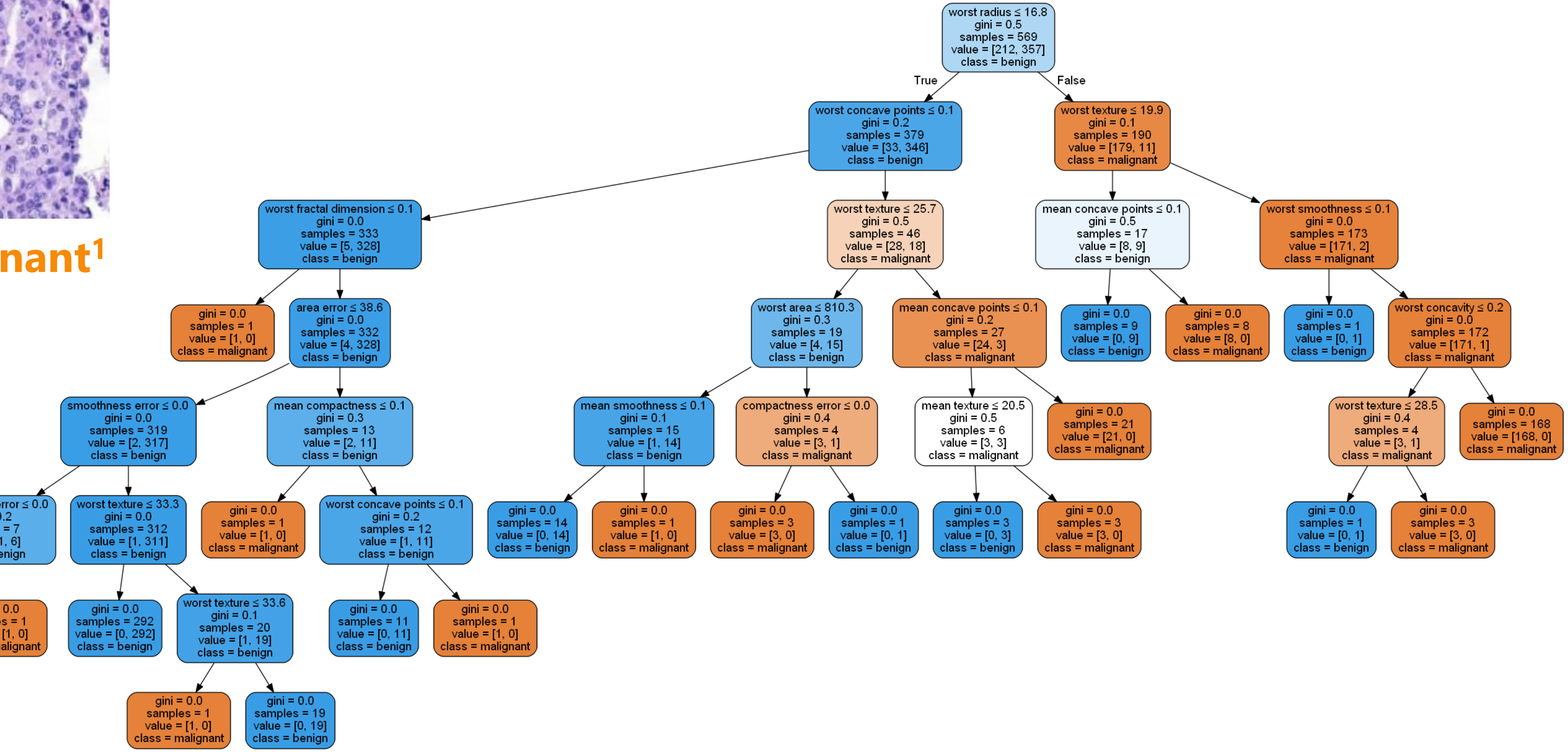
**Benign[1]**



**Malignant[1]**

### Features:

- Mean Radius
- Mean Perimeter
- …
- 30 in total!



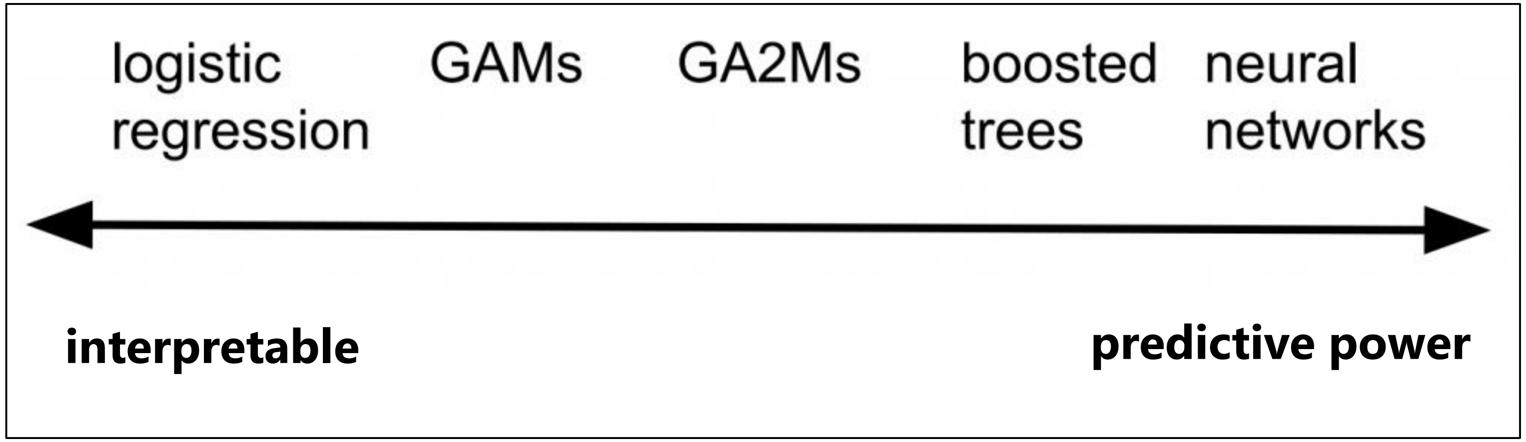

[1] UCSB Breast Cancer Cell Dataset https://bioimage.ucsb.edu/research/bio-segmentation

# Other Transparent Models

- Linear Regression
- Logistic Regression
- Generalized Additive Models (GAMs)
  - GA2Ms

$$g(E[y]) = \sum f_i(x_i)$$

$$g(E[y]) = \sum f_i(x_i) + \sum f_{ij}(x_i, x_j)$$



# Data & Features

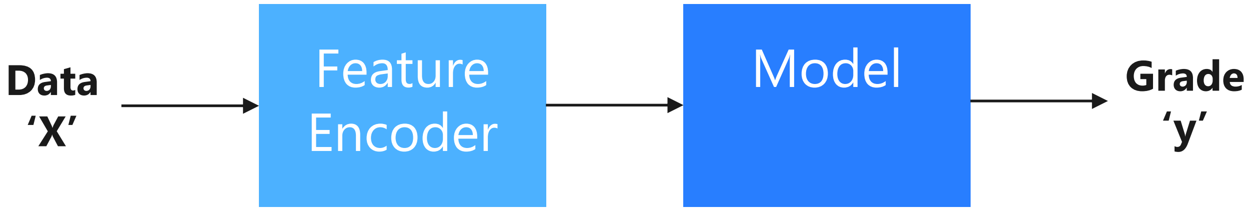
Going Beyond Feature Importance

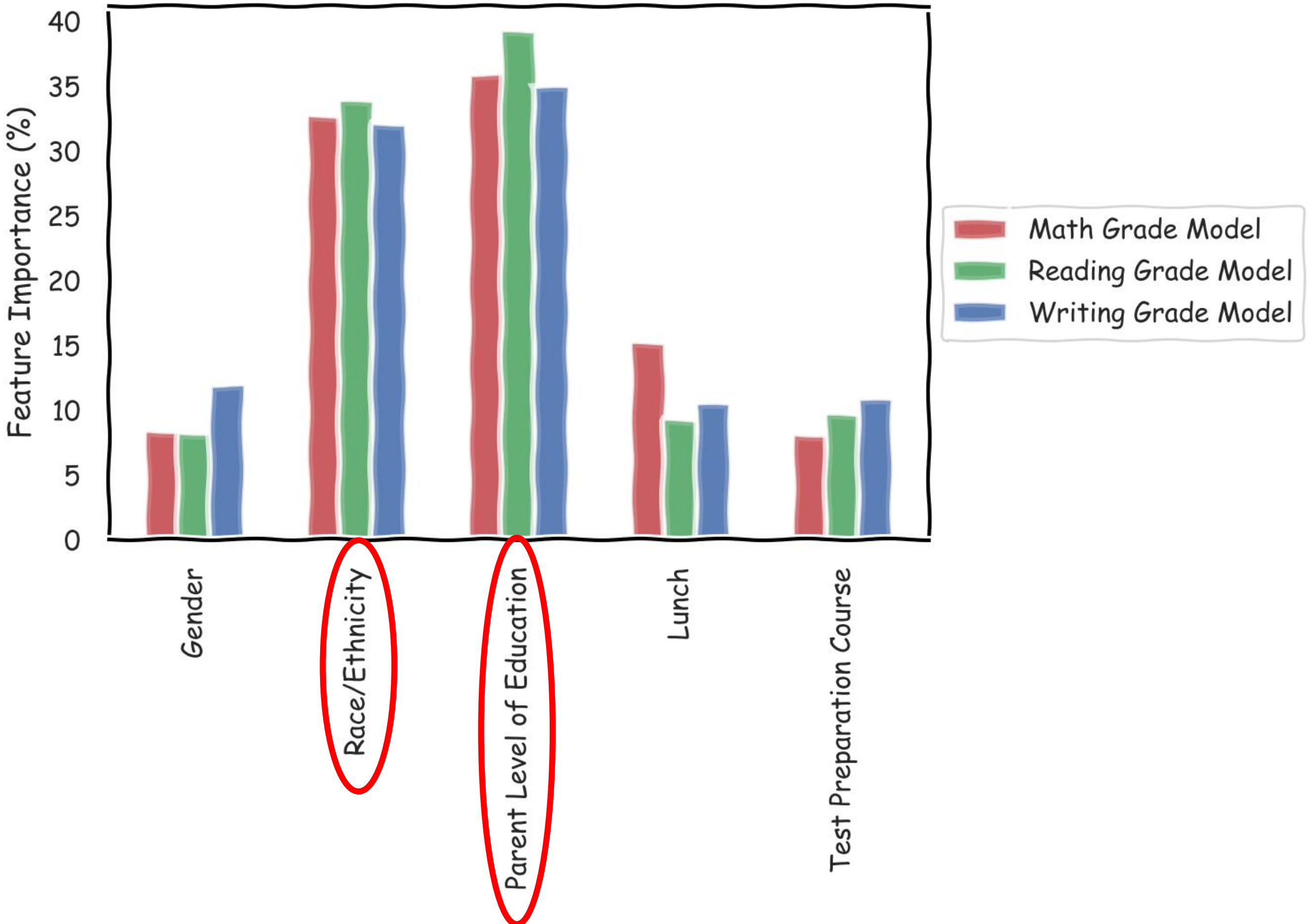
# Feature Importance

Problem: Predict high-school student grades for math, reading and writing

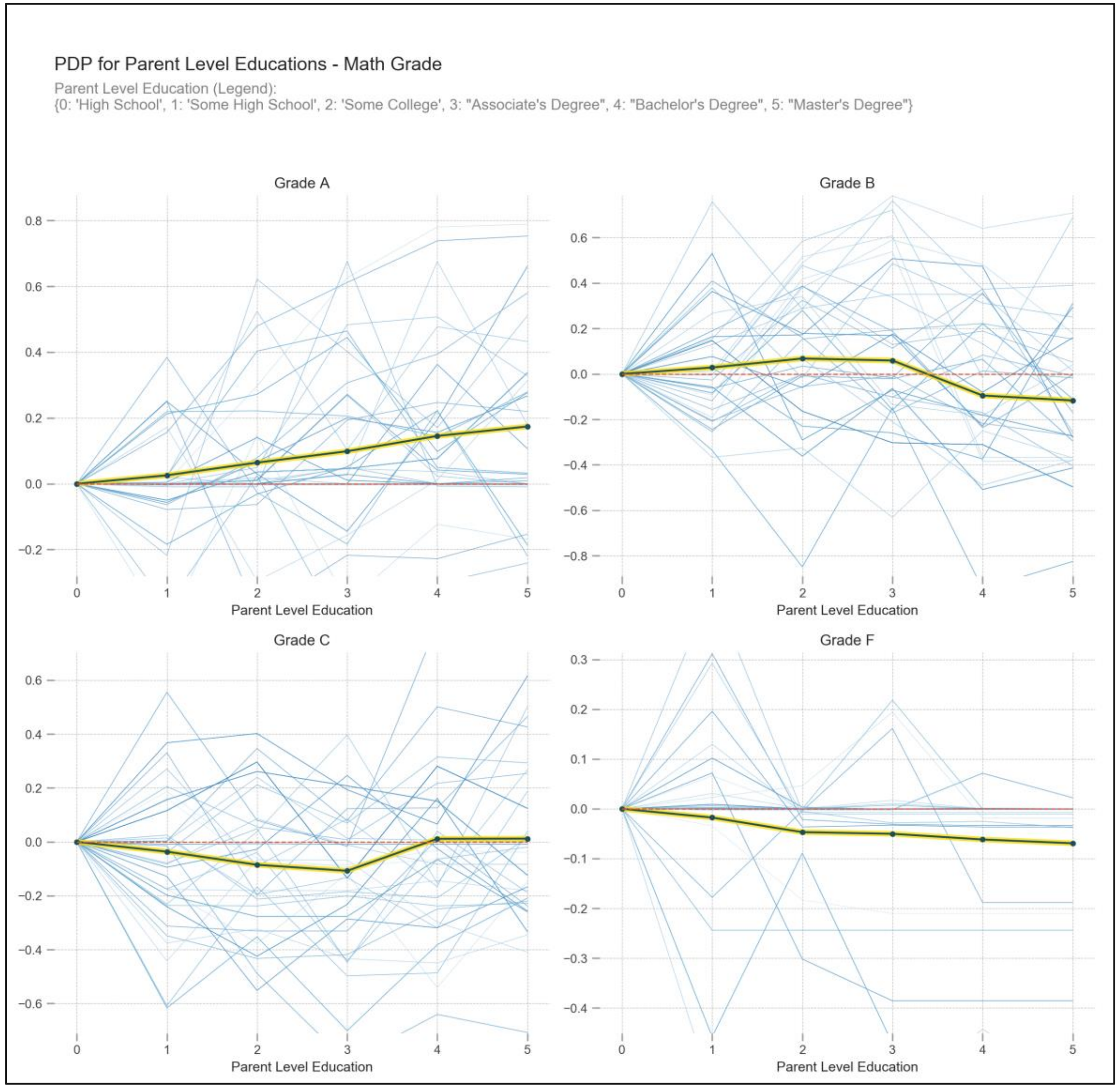
Features:

- Gender
- Race/Ethnicity
- Parent Level of Education
- Lunch
- Test Preparation

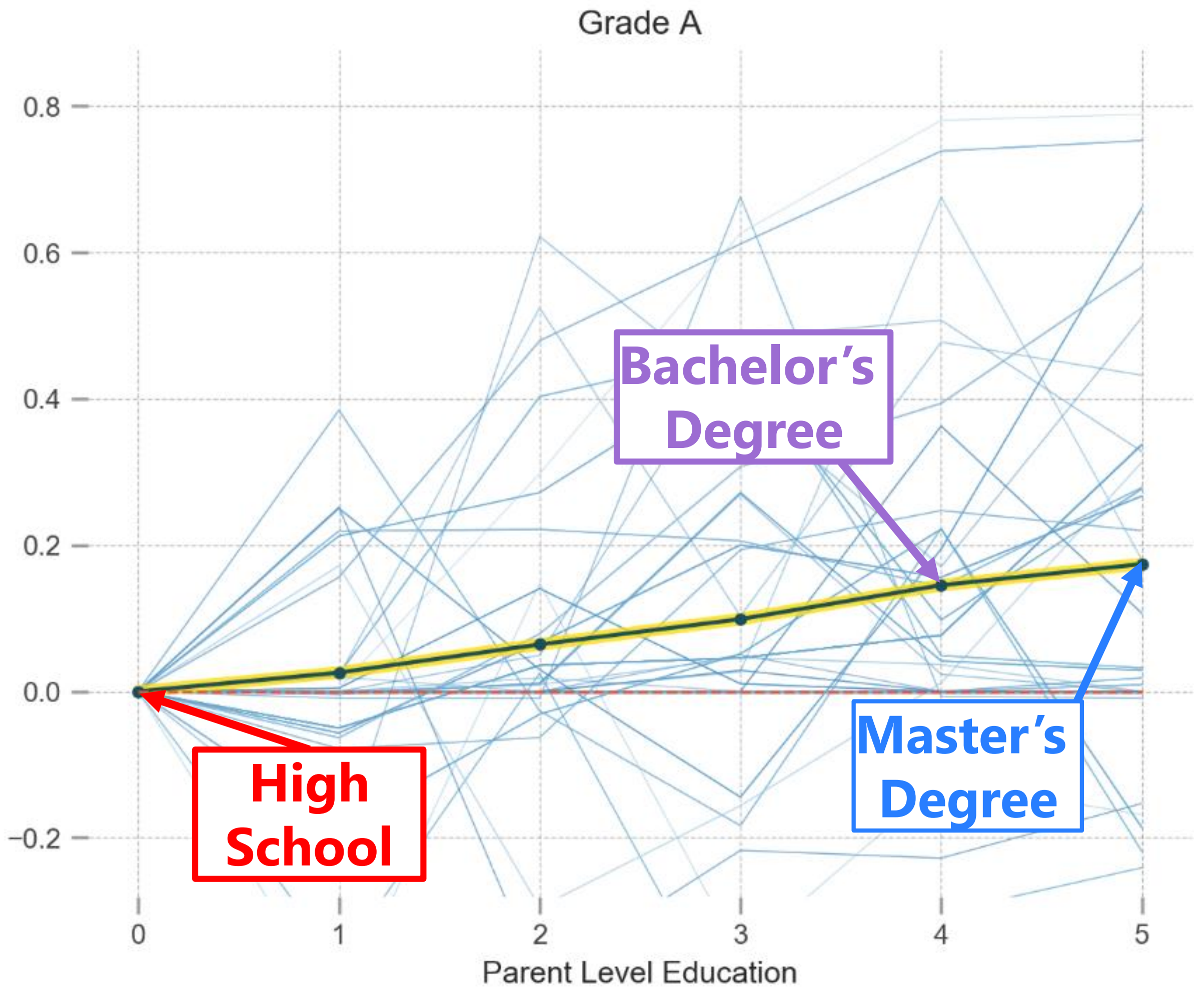


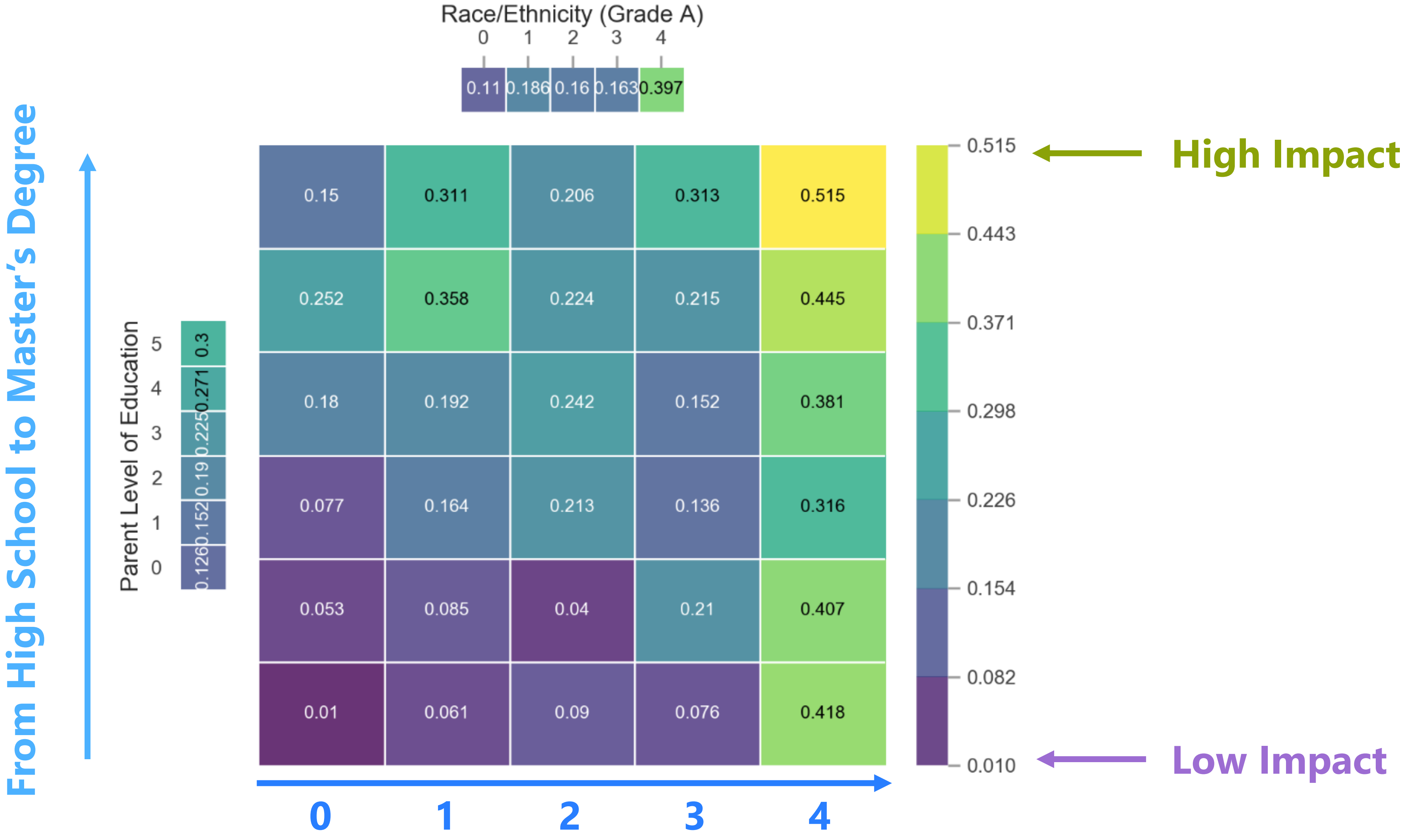
# Partial Dependency Plots (PDPs)



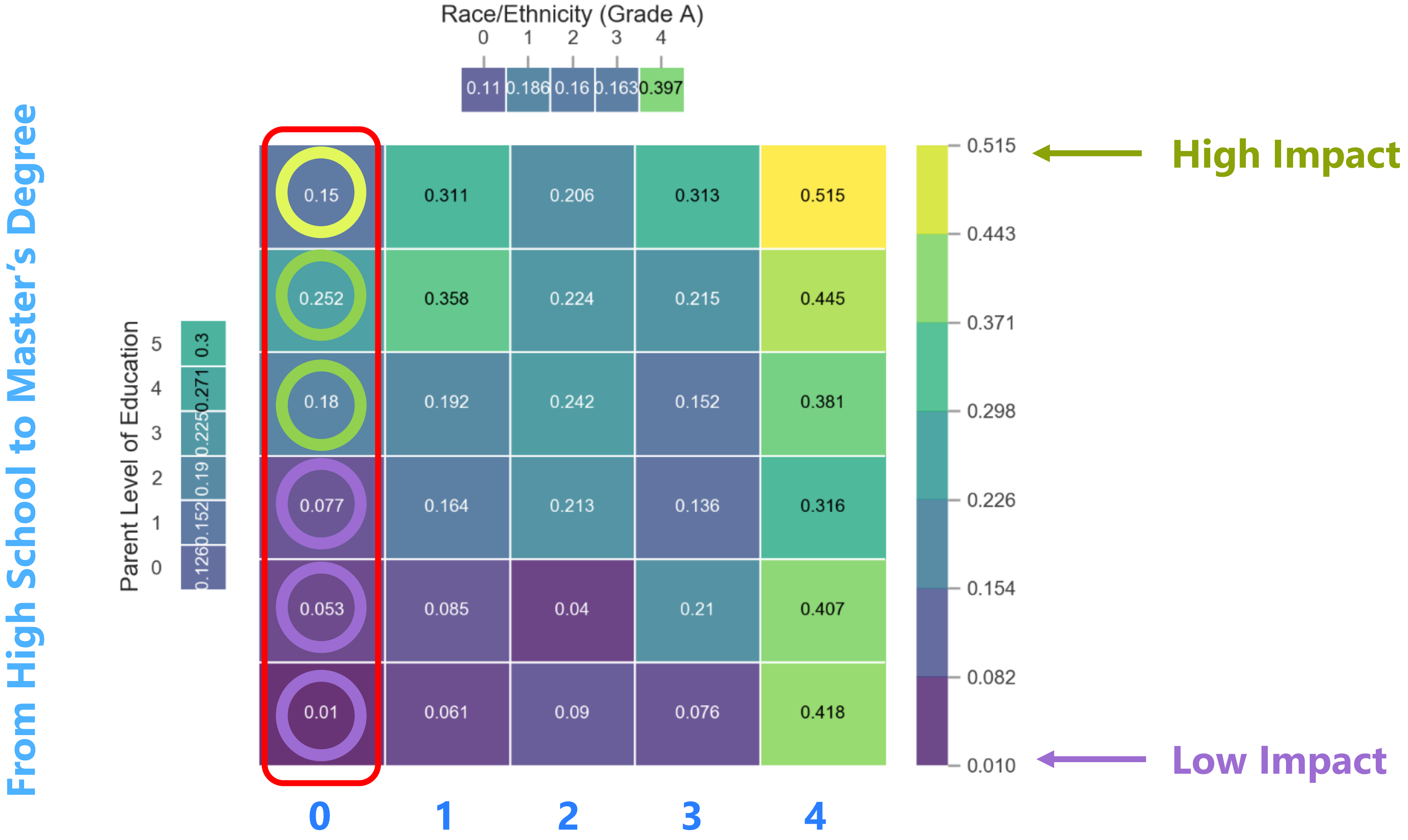

# Parent Level Education



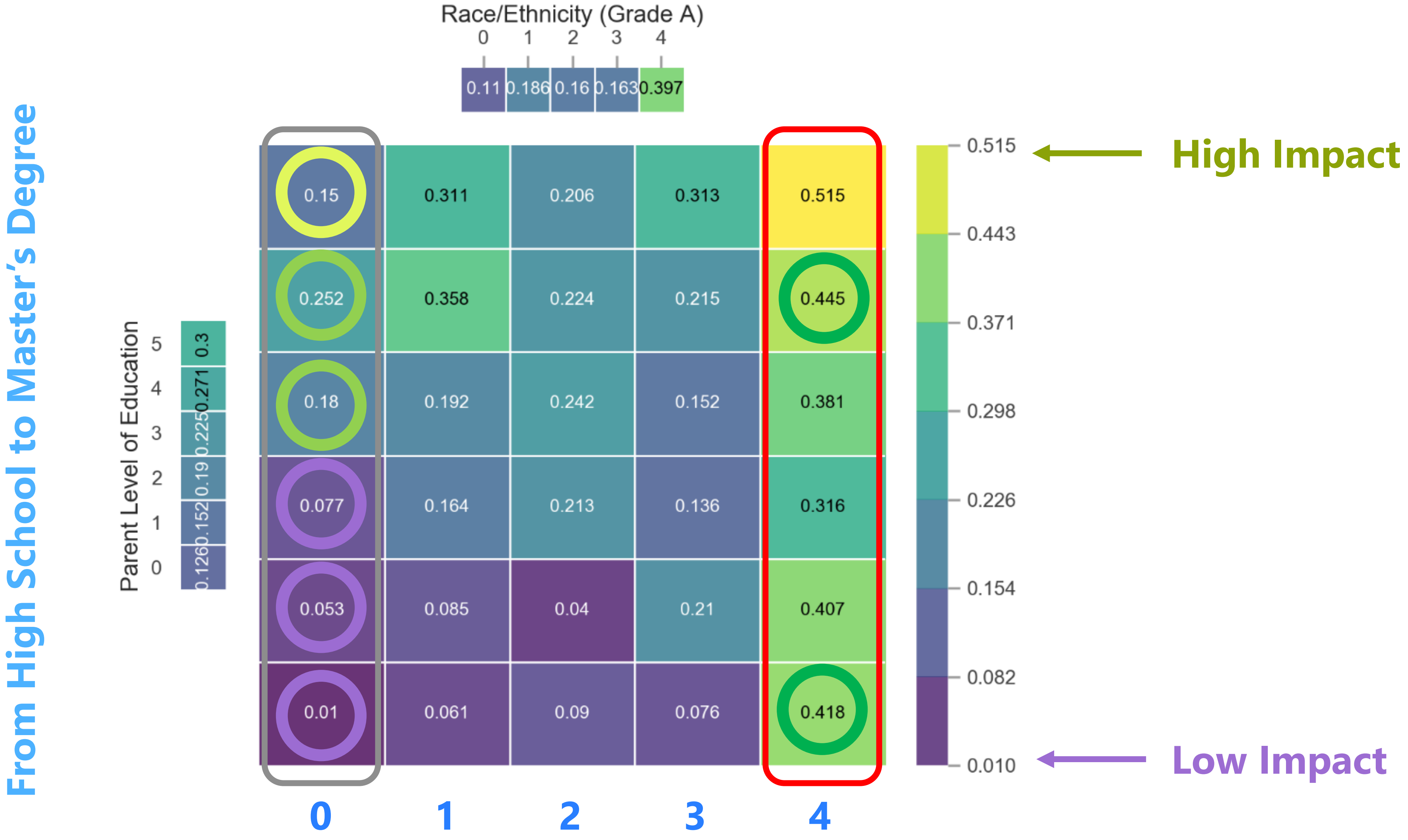
# Feature Interactions



# Feature Interactions



# Feature Interactions

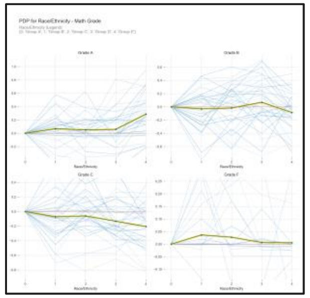


# PDP in Python

https://github.com/thampiman/interpretability

```
from pdpbox import pdp

pdp_race = pdp.pdp_isolate(model=math_model,
                          dataset=df,
                          model_features=features,
                          feature=‘race’)
```

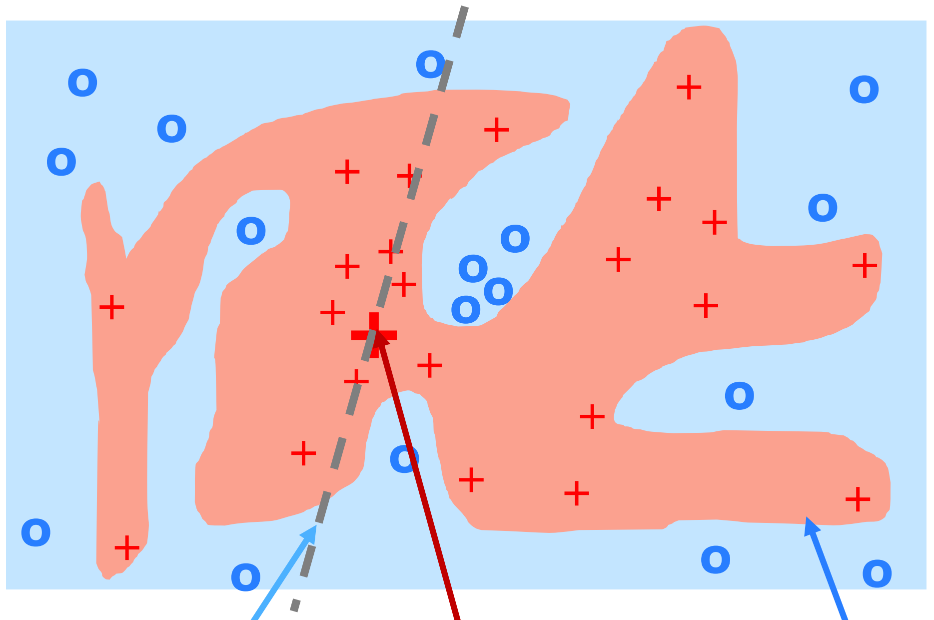


```
pdp_race_parent = pdp.pdp_interact(model=math_model,
                          dataset=df,
                          model_features=features,
                          features=[‘race’, ‘parent’])
```

# Black-Box Models

Post-Hoc Explanations

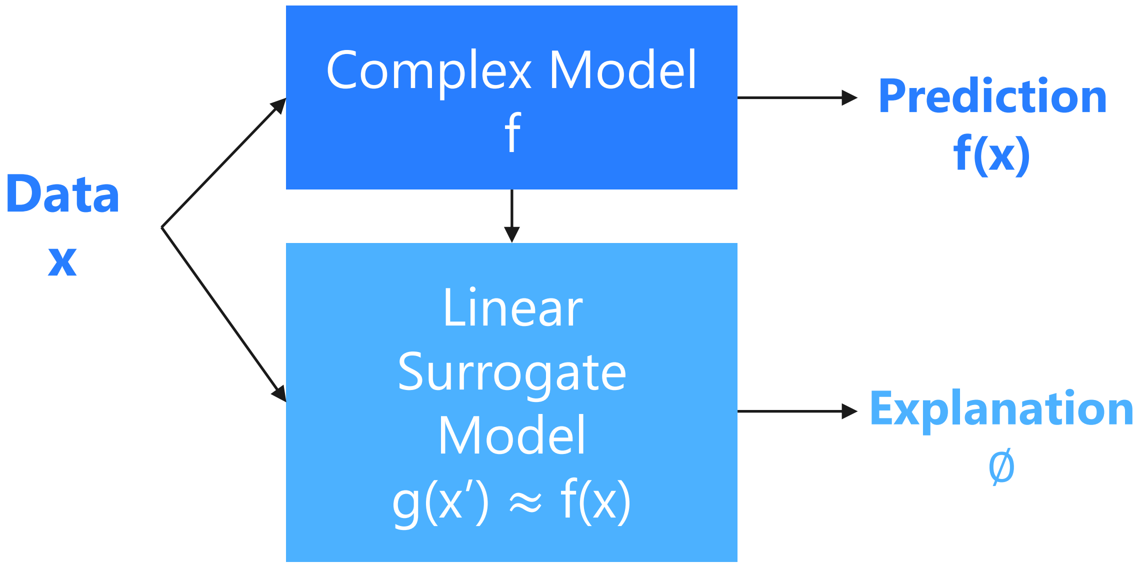
# Post-Hoc Explanations



Learned representation that is locally faithful but not globally

Pick an instance to explain

Complex decision function - hard to explain




$$g(x') = \emptyset_0 + \sum_{i=1}^{M} \emptyset_i x_i' \approx f(x)$$

**Explanation Parameters**

**LIME** = **L**ocal **I**nterpretable **M**odel-agnostic **E**xplanations

**2016**

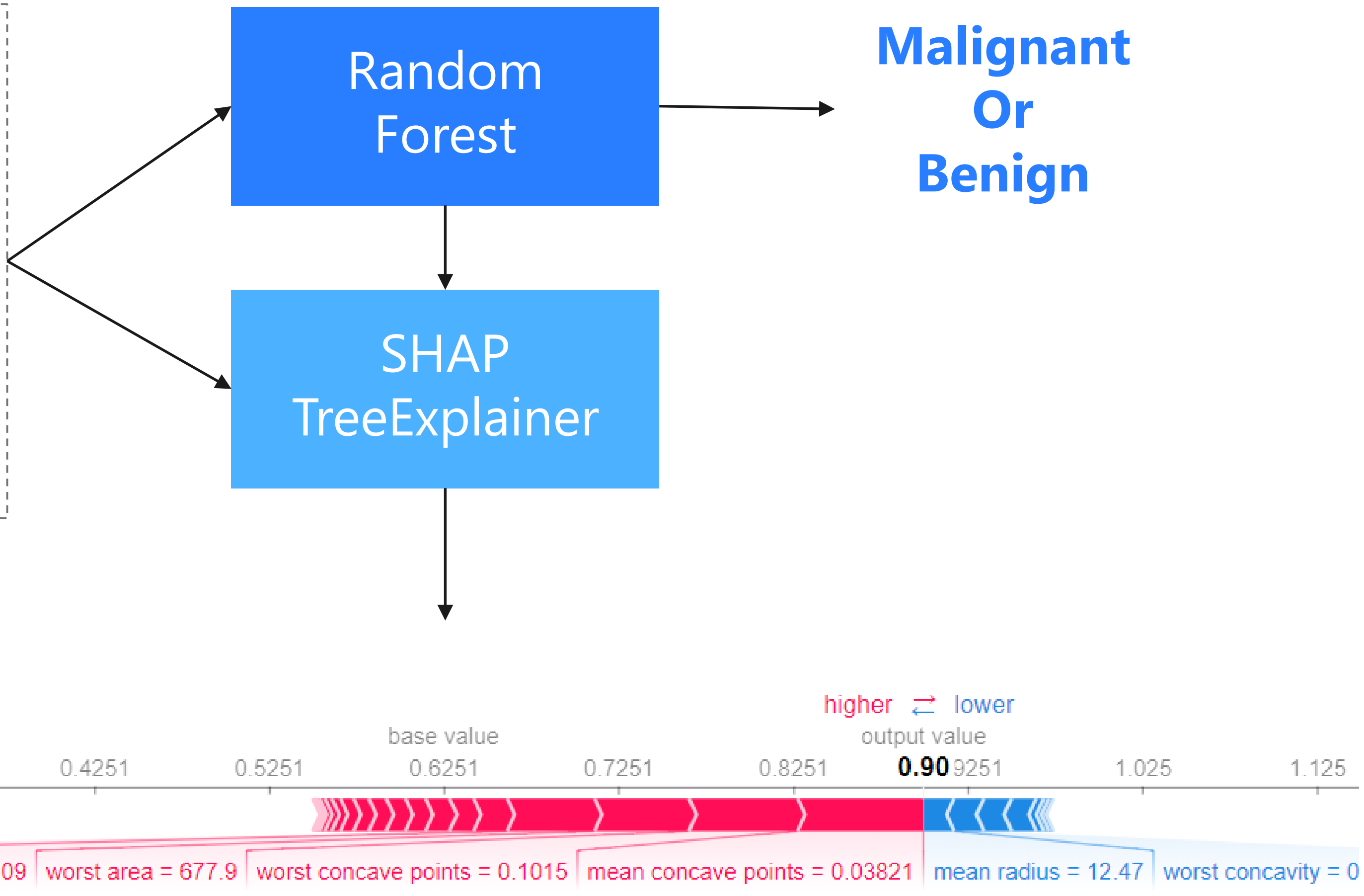
**SHAP** = **Shapley Additive exPlanations**

**2017**

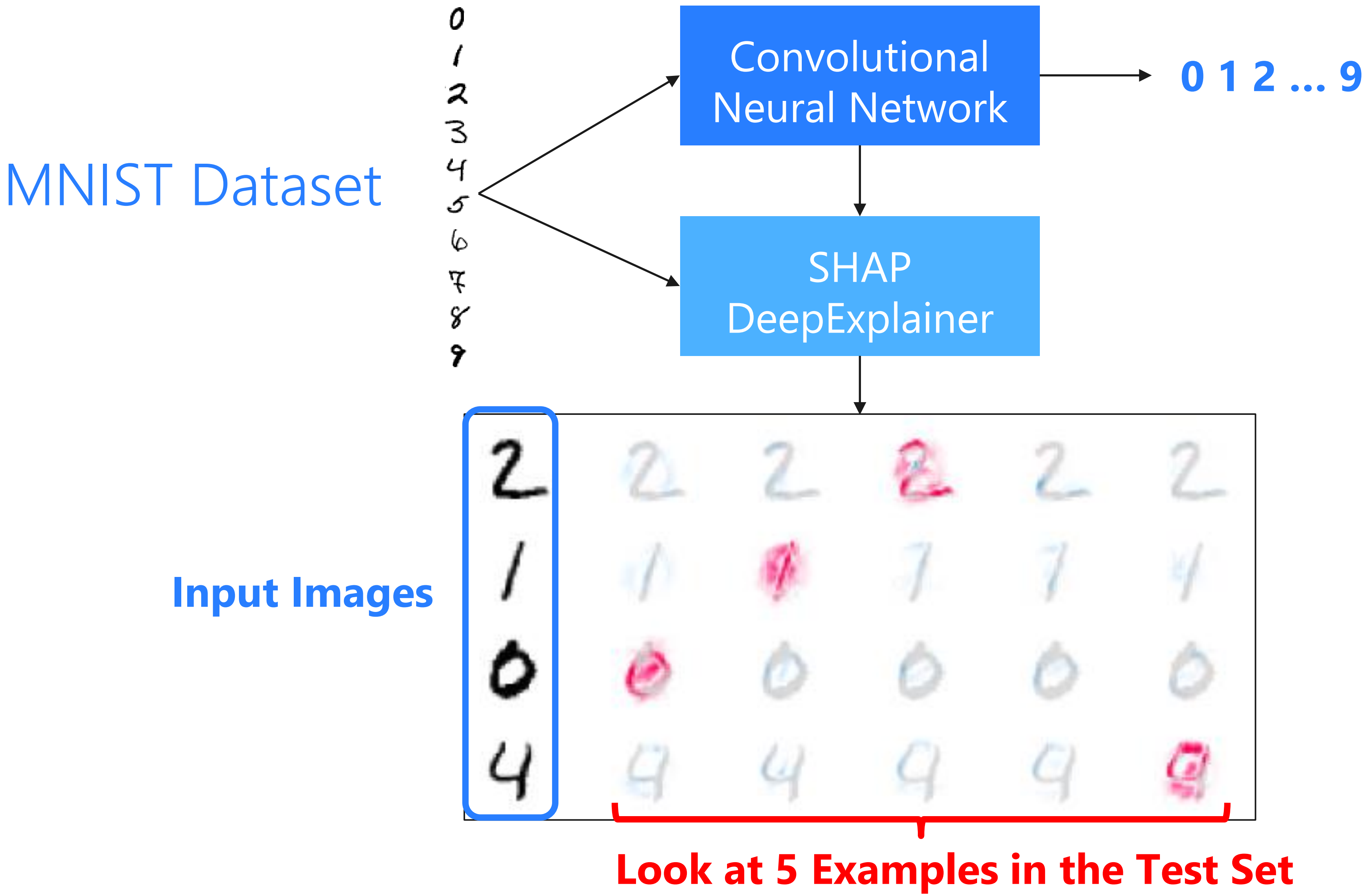
# SHAP Tree Ensemble Explainer

## Breast Cancer Detection

**Data**
- Mean Radius
- Mean Perimeter
- Worst Concave Points
- Mean Concave Points
- …



# SHAP Deep Learning Explainer

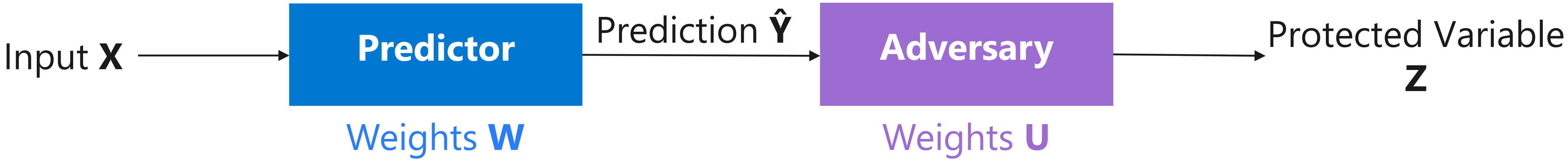


# SHAP Deep Learning Explainer – Explained!

# Mitigating Bias

Algorithmic Debiasing

# Adversarial Debiasing




- Demographic Parity: $\hat{Y} \perp Z$
- Equality of Odds: $\hat{Y} \perp Z \mid Y$
- Equality of Opportunity: $\hat{Y} \perp Z \mid Y = y$

***Source****: Brian Zhang, Blake Lemoine and Margaret Mitchell. Mitigating Unwanted Biases with Adversarial Learning. AAAI Conference on AI, Ethics and Society, 2018.*

# Adversarial Debiasing Demo

Word Embedding Analogy Task

**He : She :: Doctor : ?**

| Word | Score |
|---|---|
| Nurse | 0.62 |
| Her | 0.60 |
| Woman | 0.58 |
| Mother | 0.57 |
| Doctors | 0.55 |
| Physician | 0.53 |
| Pregnant | 0.51 |

# Summary

- Include model understanding in your data science process
- Be mindful of your audience – interpretability means different things to different people
- Apply interpretability techniques (like PDPs, LIME, SHAP, etc.) to improve model understanding
- Build fair models by mitigating bias

# Additional Resources (1/2)

- Source code of demos: https://github.com/thampiman/interpretability
- Blog post on interpretability: https://towardsdatascience.com/interpretable-ai-or-how-i-learned-to-stop-worrying-and-trust-ai-e61f9e8ee2c2
- Saliency Maps: https://distill.pub/2018/building-blocks/
- Representational Learning: https://www.cl.uni-heidelberg.de/courses/ws14/deepl/BengioETAL12.pdf
- t-SNE: https://lvdmaaten.github.io/tsne/
- PDP Box: https://github.com/SauceCat/PDPbox
- LIME: https://arxiv.org/pdf/1602.04938.pdf
- Kernel SHAP: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions

# Additional Resources (2/2)

- SandDance: https://www.microsoft.com/en-us/research/project/sanddance/
- GAMut: https://www.microsoft.com/en-us/research/uploads/prod/2019/01/19_gamut_chi.pdf
- Datasheets for Datasets: https://arxiv.org/pdf/1803.09010.pdf
- Challenges for Transparency: https://arxiv.org/pdf/1708.01870.pdf
- Synthetic Data (MSR): https://arxiv.org/pdf/1810.00471.pdf
- Counterfactual Explanations: https://arxiv.org/abs/1711.00399
- Noise Audit: https://hbr.org/2016/10/noise
- Interpretable ML: https://christophm.github.io/interpretable-ml-book/limo.html

# Thank You

Q&A

@thampiman