Trustworthy AI: Explainability

Destiny Pounds

M.S. Student, Biomedical Data Science

Advised by Vibhuti Gupta, Ph.D.

Assistant Professor, Computer Science and Data Science

School of Applied Computational Sciences

Meharry Medical College

Overview

- Defining Explainability
- Model Understanding
  - Examples
  - Benefits
- Approaches to Model Understanding
  - Interpretable Models
  - Post-hoc Explainability
    - Local Explanations
    - Global Explanations

Explainable Artificial Intelligence (XAI)

- Explainability of an AI model describes the extent to which human-users can comprehend and trust the results and output created by the model.

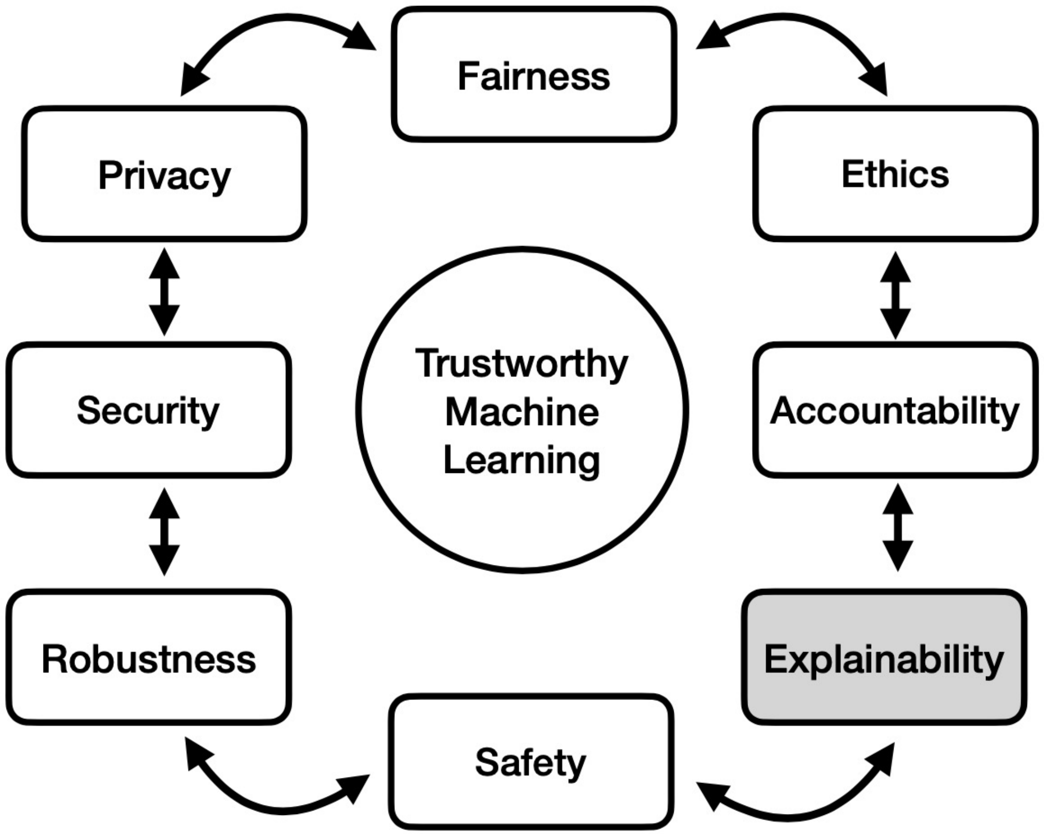


Image from Farnadi, G. (2022). *Trustworthy Machine Learning* [Slides]. HEC Montréal

Overview of Predictive Modeling Process

Input (Data) → Model → Output (Prediction)

Explainable AI requires model understanding.

Example: Why Model Understanding?

Input

Predictive Model

Prediction = Siberian Husky

Adapted from Farnadi, G. (2022). *Trustworthy Machine Learning* [Slides]. HEC Montréal

Example: Why Model Understanding?

Input

Predictive Model

Model Understanding

Prediction = Siberian Husky

This model is relying on incorrect features to make this prediction!! Let me fix the model

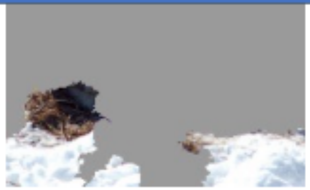
Adapted from Farnadi, G. (2022). *Trustworthy Machine Learning* [Slides]. HEC Montréal

Example: Why Model Understanding?

Input

This model is
incorrect
make
on!! Let
model

Model understanding facilitates debugging.



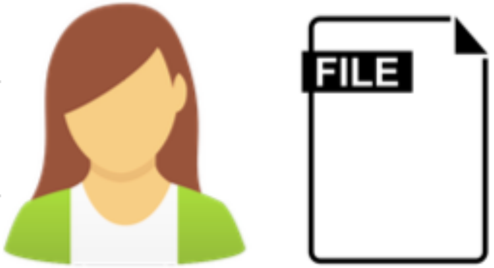
Predictive Model

Prediction = Siberian Husky

Adapted from Farnadi, G. (2022). *Trustworthy Machine Learning* [Slides]. HEC Montréal

Example: Why Model Understanding?

Defendant Details

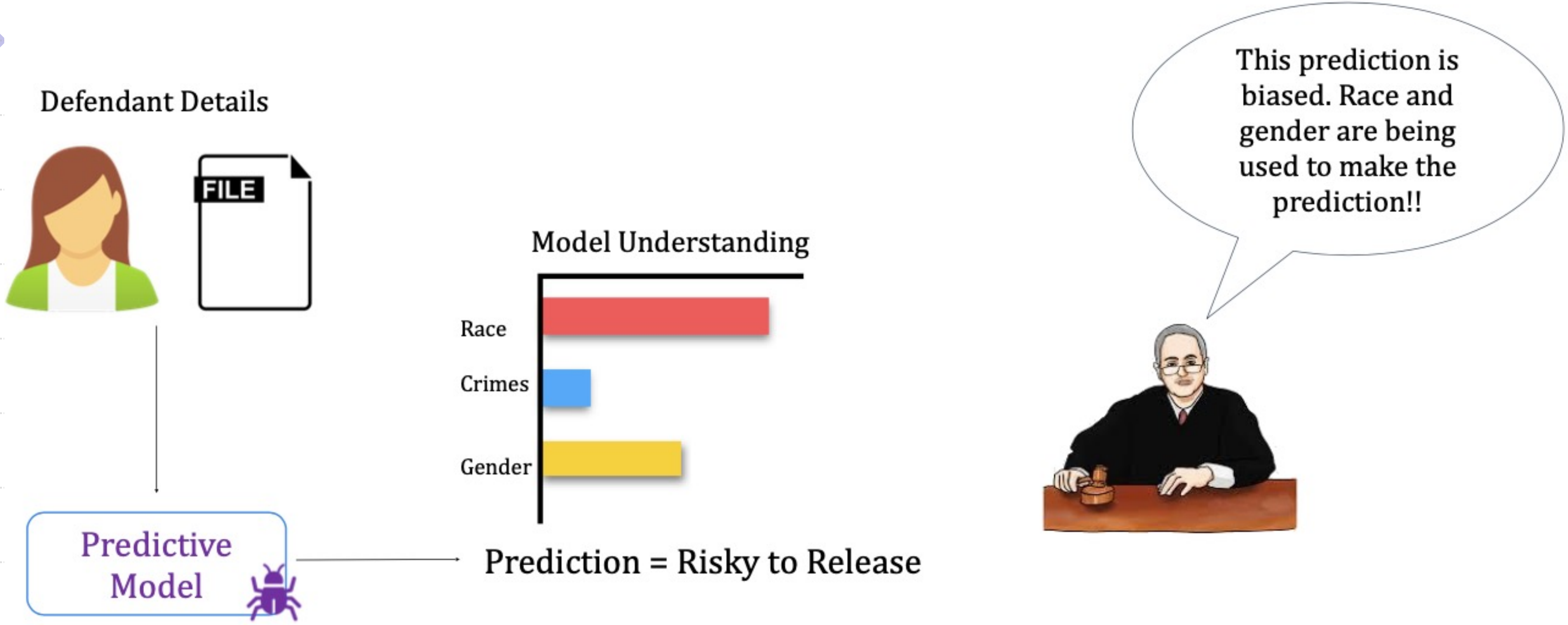


Predictive Model

Prediction = Risky to Release

Adapted from Farnadi, G. (2022). *Trustworthy Machine Learning* [Slides]. HEC Montréal

Example: Why Model Understanding?



Adapted from Farnadi, G. (2022). *Trustworthy Machine Learning* [Slides]. HEC Montréal

Example: Why Model Understanding?

This prediction is

Defendant Detai

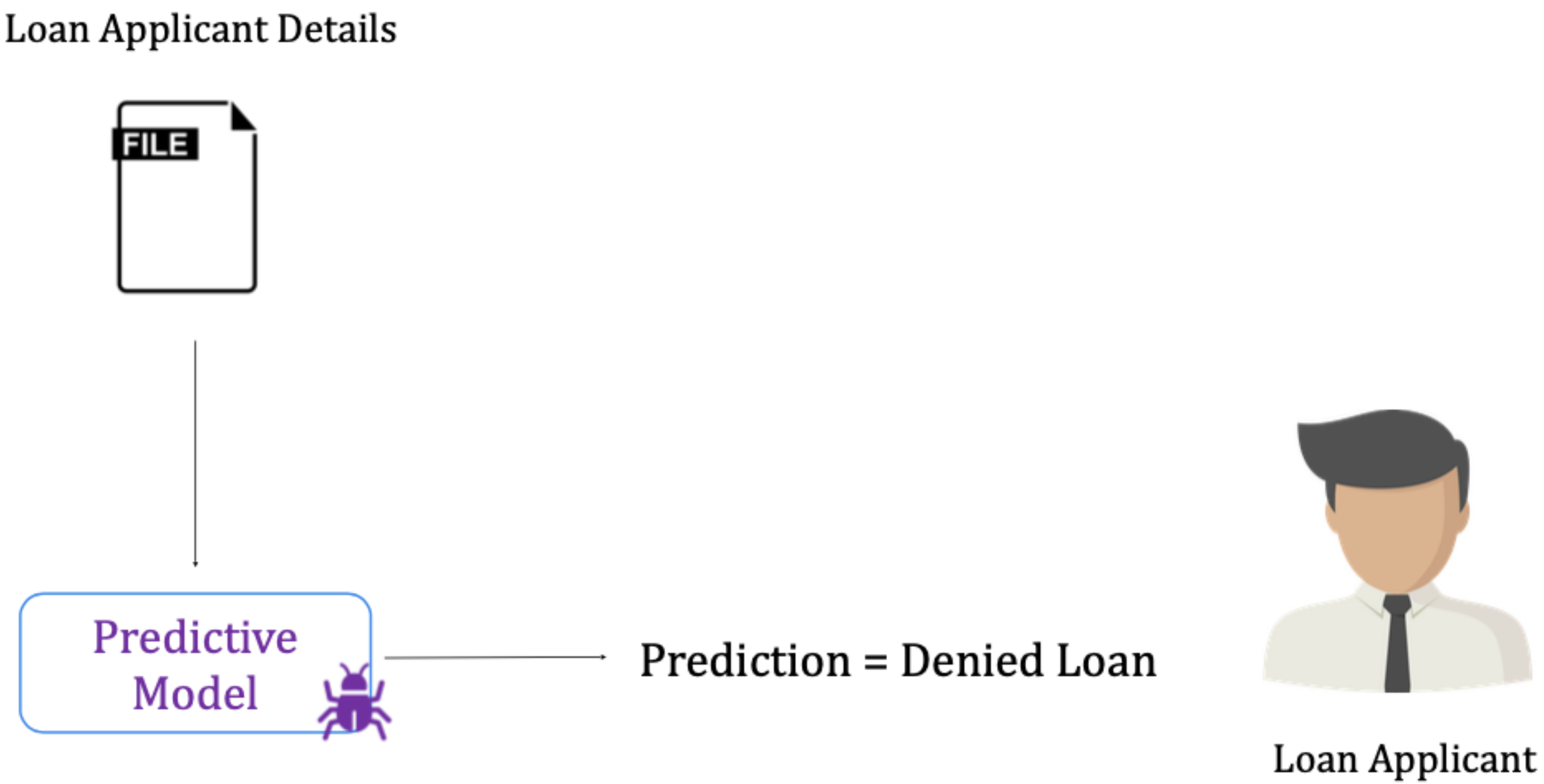
Model understanding facilitates bias detection.

Predictive Model

Prediction = Risky to Release

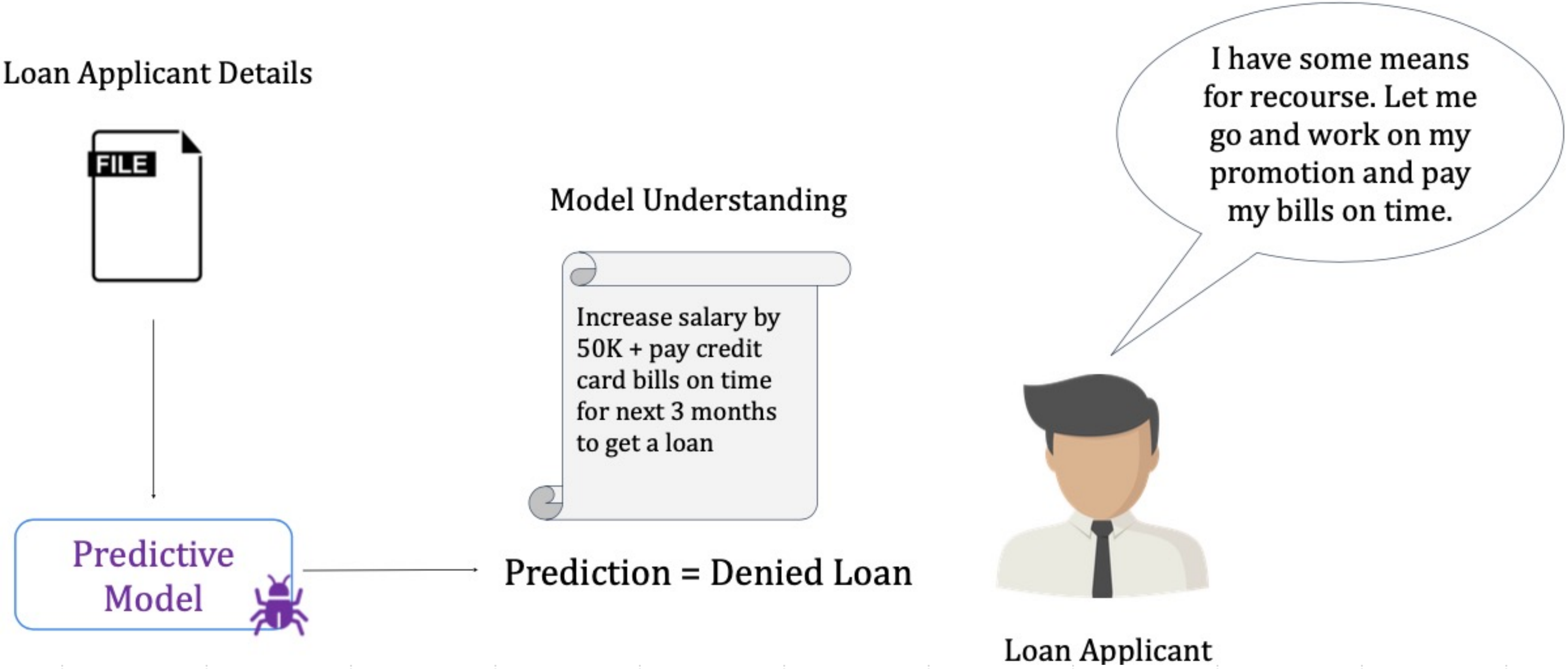
Adapted from Farnadi, G. (2022). *Trustworthy Machine Learning* [Slides]. HEC Montréal

Example: Why Model Understanding?



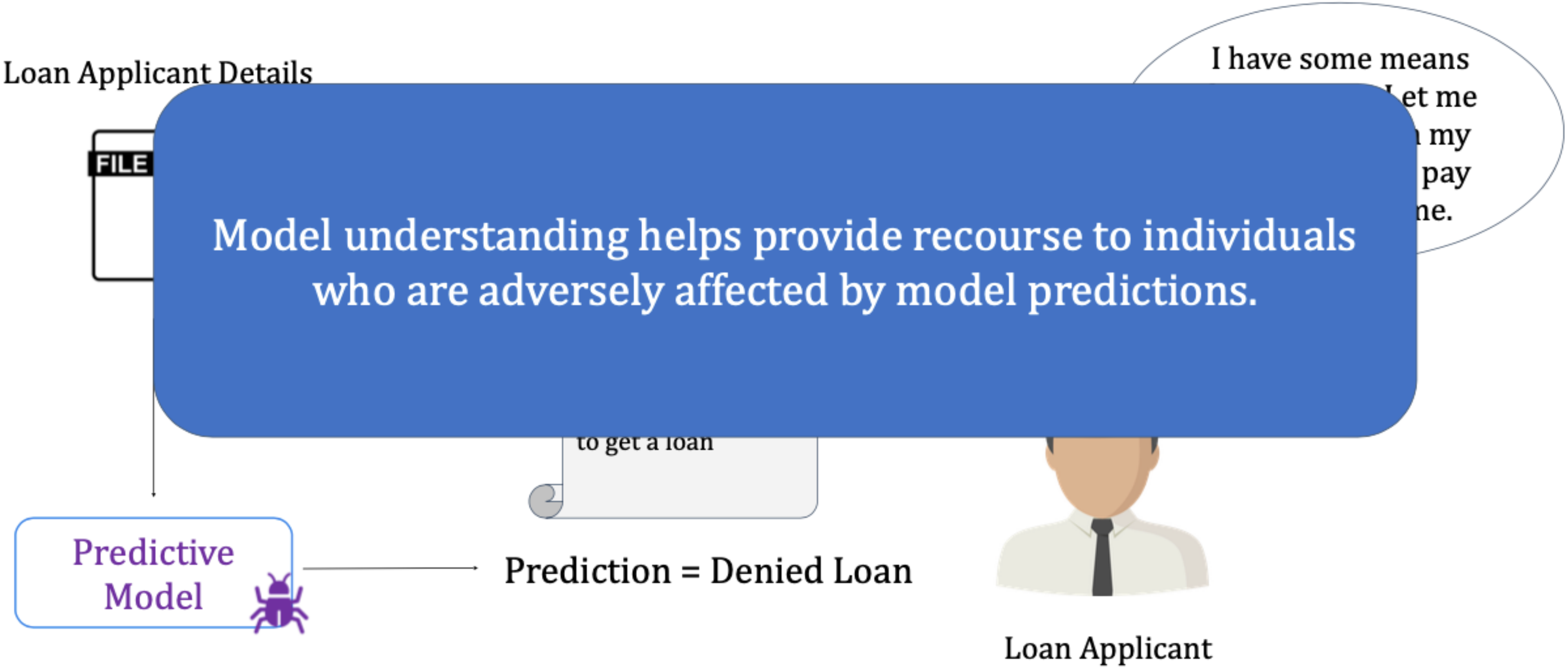
Adapted from Farnadi, G. (2022). *Trustworthy Machine Learning* [Slides]. HEC Montréal

Example: Why Model Understanding?



Adapted from Farnadi, G. (2022). *Trustworthy Machine Learning* [Slides]. HEC Montréal

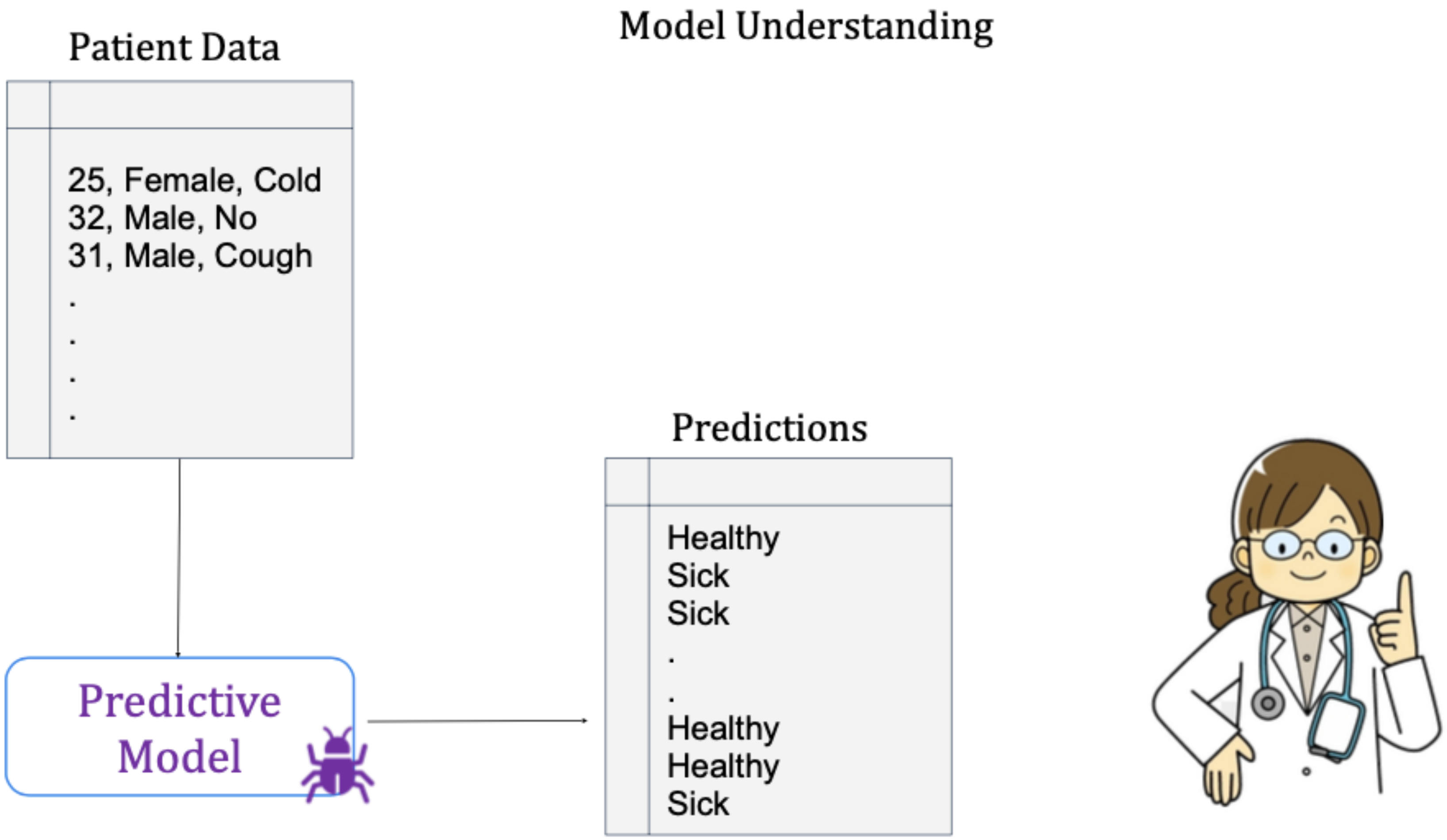
Example: Why Model Understanding?



Model understanding helps provide recourse to individuals who are adversely affected by model predictions.

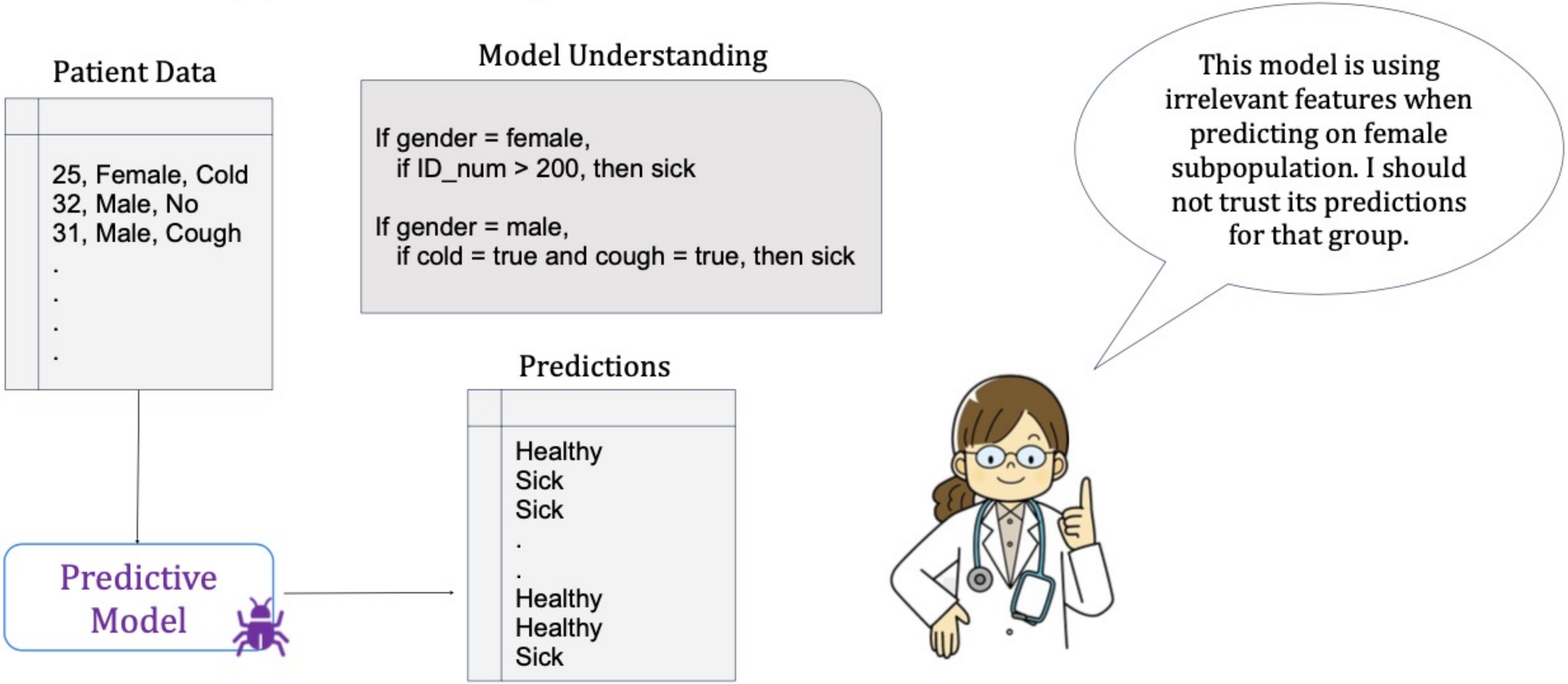
Adapted from Farnadi, G. (2022). *Trustworthy Machine Learning* [Slides]. HEC Montréal

Example: Why Model Understanding?



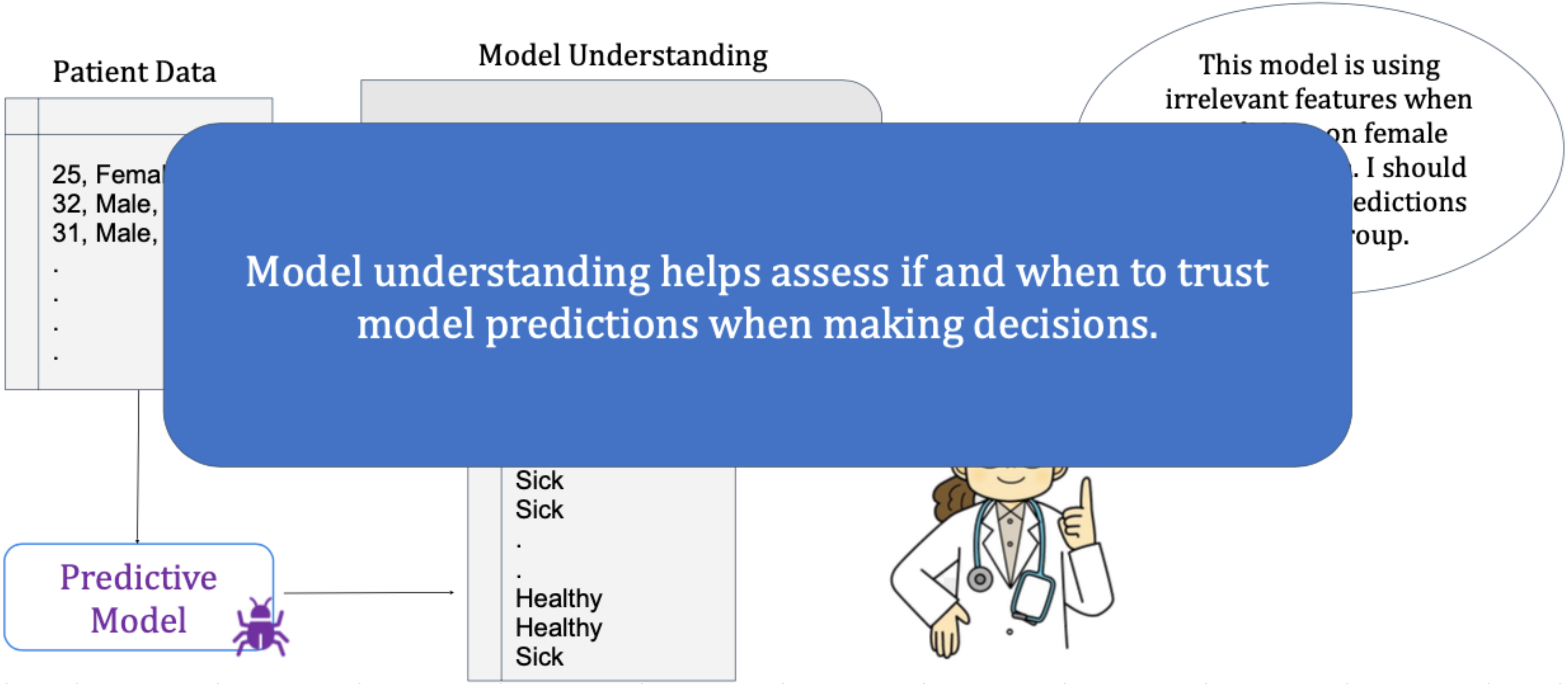
Adapted from Farnadi, G. (2022). *Trustworthy Machine Learning* [Slides]. HEC Montréal

Example: Why Model Understanding?



Adapted from Farnadi, G. (2022). *Trustworthy Machine Learning* [Slides]. HEC Montréal

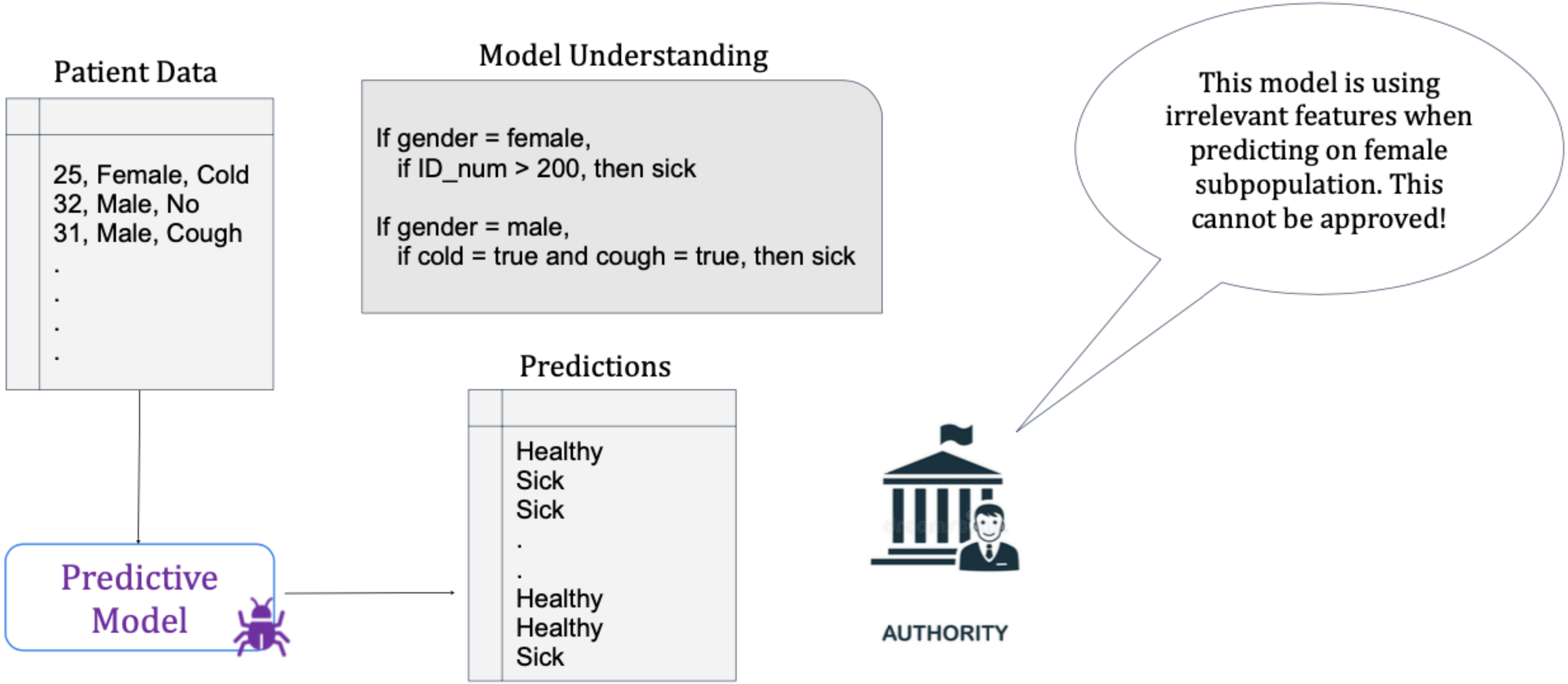
Example: Why Model Understanding?



Model understanding helps assess if and when to trust model predictions when making decisions.

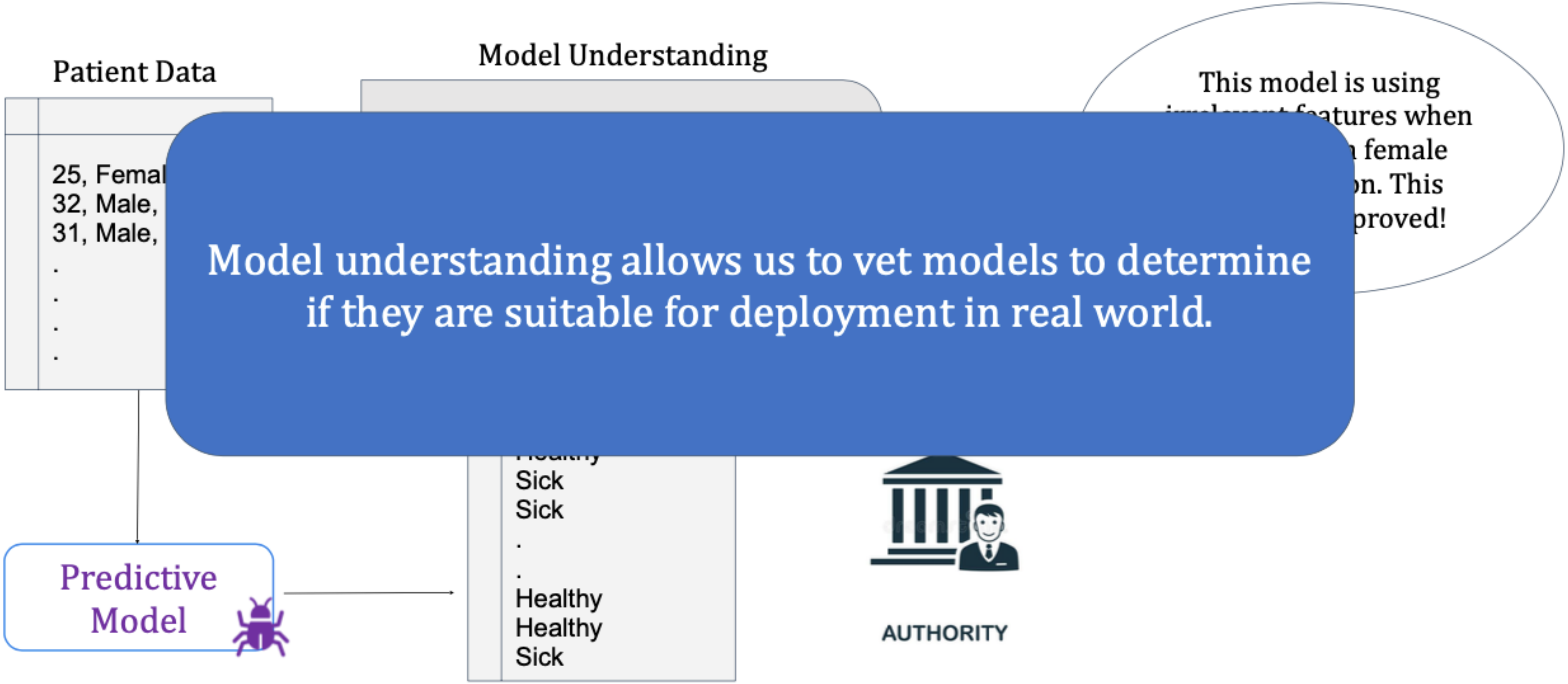
Adapted from Farnadi, G. (2022). *Trustworthy Machine Learning* [Slides]. HEC Montréal

Example: Why Model Understanding?



Adapted from Farnadi, G. (2022). *Trustworthy Machine Learning* [Slides]. HEC Montréal

Example: Why Model Understanding?



Model understanding allows us to vet models to determine if they are suitable for deployment in real world.

Adapted from Farnadi, G. (2022). *Trustworthy Machine Learning* [Slides]. HEC Montréal

Summary: Benefits of Model Understanding

- Debugging
- Bias Detection
- Recourse
- Assess Trustworthiness of Predictions
- Vet Models for Deployment

Achieving Model Understanding

Approach #1:

Build inherently interpretable (white box) models.

Approach #2:

Explain pre-built (black-box) models in a post-hoc manner.

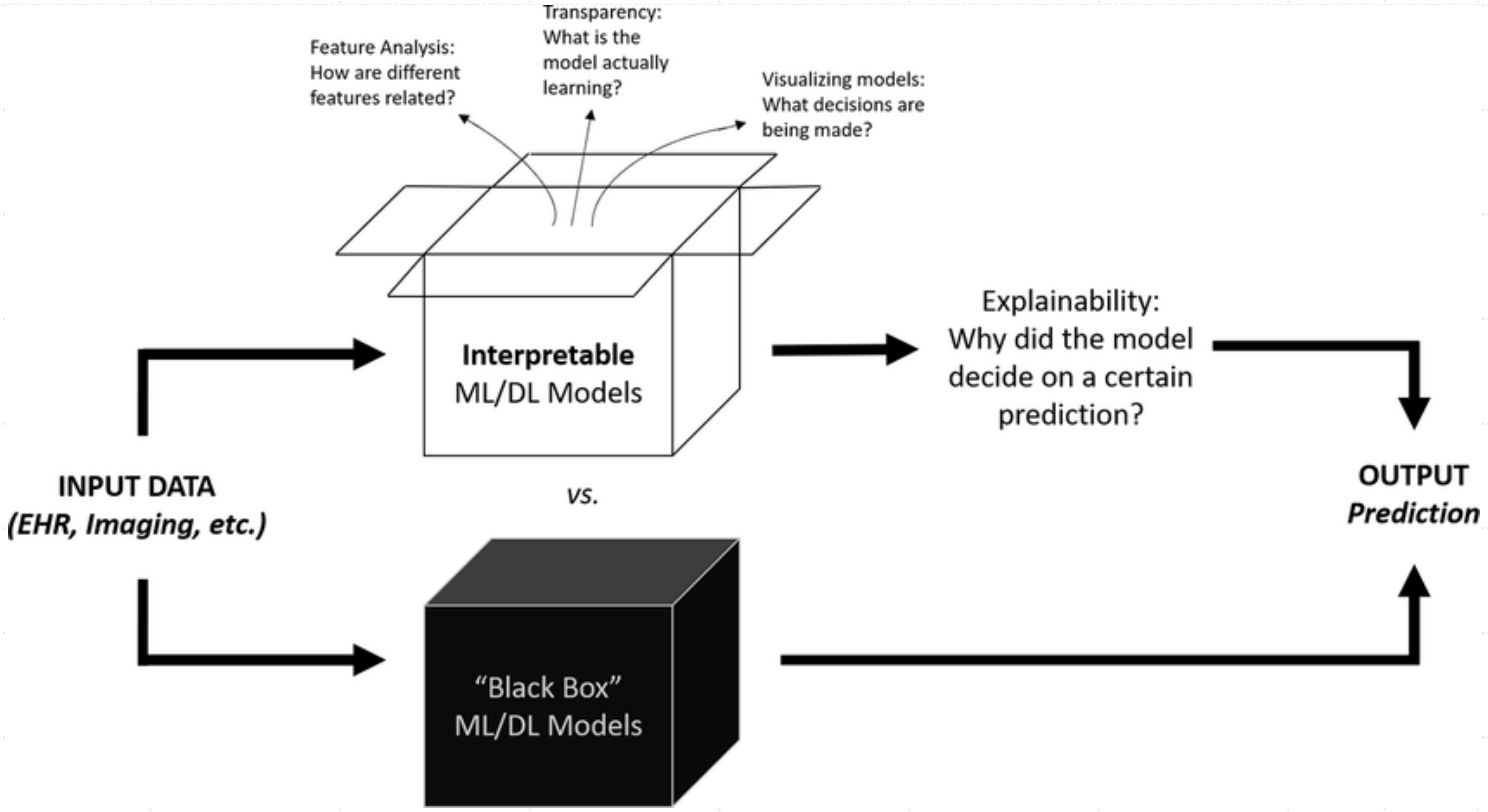
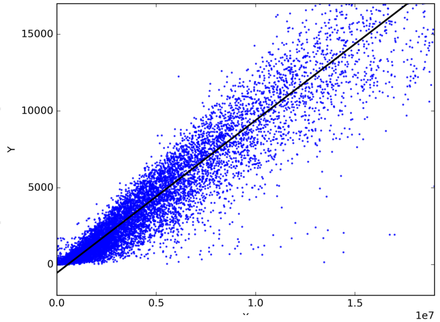
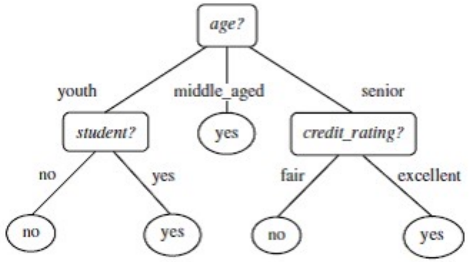


Image from Hui, Aaron & Ahn, Shawn & Lye, Carolyn & Deng, Jun. (2022). Ethical Challenges of Artificial Intelligence in Health Care: A Narrative Review. Ethics in Biology, Engineering and Medicine: An International Journal. 12. 10.1615/EthicsBiologyEngMed.2022041580.

Inherently Interpretable Models

- These models are interpretable by design. Understanding of the model is clear before receiving results.



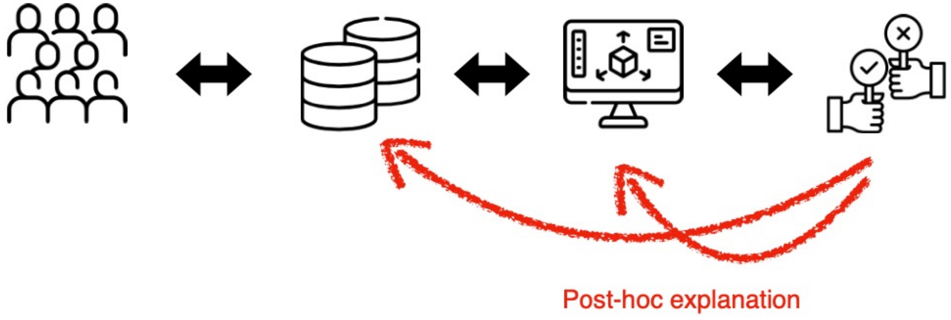


The HEART Score for Chest Pain Patients in the ED		
History	• Highly Suspicious • Moderately Suspicious • Slightly or Non-Suspicious	• 2 points • 1 point • 0 points
ECG	• Significant ST-Depression • Nonspecific Repolarization • Normal	• 2 points • 1 point • 0 points
Age	• ≥ 65 years • > 45 - < 65 years • ≤ 45 years	• 2 points • 1 point • 0 points
Risk Factors	• ≥ 3 Risk Factors or History of CAD • 1 or 2 Risk Factors • No Risk Factors	• 2 points • 1 point • 0 points
Troponin	• ≥ 3 x Normal Limit • > 1 - < 3 x Normal Limit • ≤ Normal Limit	• 2 points • 1 point • 0 points
Risk Factors: DM, current or recent (<one month) smoker, HTN, HLP, family history of CAD, & obesity		
Score 0 – 3: 2.5% MACE over next 6 weeks → Discharge Home **Score 4 – 6:** 20.3% MACE over next 6 weeks → Admit for Clinical Observation **Score 7 – 10:** 72.7% MACE over next 6 weeks → Early Invasive Strategies		

Examples

- Linear Models
- Shallow Decision Tree
- Rule Based Models
- Risk Scores

Post-hoc Explainability

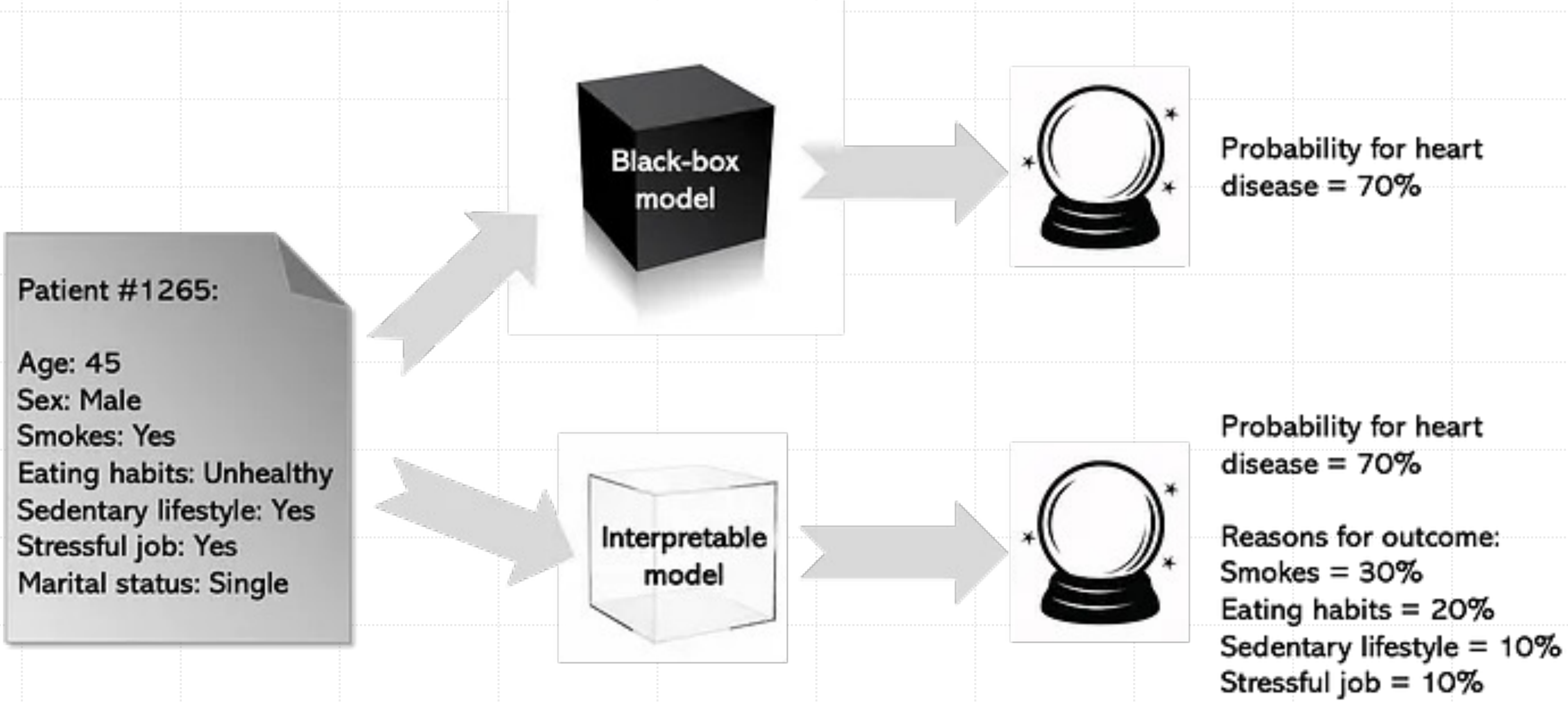


Building an inherently interpretable model is not always possible and we are left with a black box. In that case, models can be interpreted post-hoc.

Post-hoc explanation occurs after execution of the model.

Adapted from Farnadi, G. (2022). *Trustworthy Machine Learning* [Slides]. HEC Montréal

White Box vs. Black Box Example



Approaches for Post-hoc Explainability

Local Explanations

Local explanations focus on the data and provide explanations for individual outcomes. Therefore, they provide trust for model outcomes.

Global Explanations

Global explanations focus on the model and provide an understanding of the decision process. Therefore, they provide a sense of understanding to how the model works.

Thank You

Please send us your questions at:
vgupta@mmc.edu and
dpounds24@email.mmc.edu