CS282BR: Topics in Machine Learning Interpretability and Explainability

Hima Lakkaraju

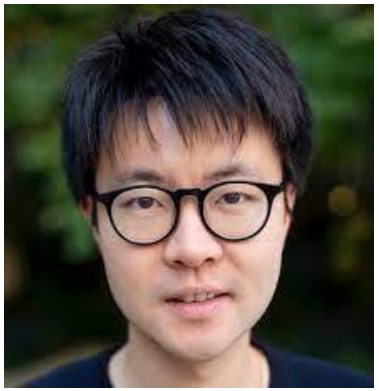
Assistant Professor
Harvard Business School + Computer Science

Agenda

- Course Goals & Logistics
- Model Understanding: Use Cases
- Inherently Interpretable Models vs. Post-hoc Explanations
- Defining and Understanding Interpretability

Course Staff

Hima Lakkaraju



Jiaqi Ma

Suraj Srinivas

Office Hours: Start from next week
Hima's office hours: Monday 1.30pm to 2.30pm
Jiaqi and Suraj's office hours: Thursday 1.00pm to 2.00pm
Location: Longeron Meeting Room, SEC 6th floor
and zoom (link on cavas)

Course Webpage: https://canvas.harvard.edu/courses/117650

Goals of this Course

Learn and improve upon the state-of-the-art literature on ML interpretability and explainability

- Understand where, when, and why is interpretability/explainability needed
- Read, present, and discuss research papers
  - Formulate, optimize, and evaluate algorithms
- Implement state-of-the-art algorithms; Do research!
- Understand, critique, and redefine literature
  - EMERGING FIELD!!

Course Overview

- Introduction & overview (Week 1)
- Evaluating interpretability (Week 2)
- Learning inherently interpretable models (Weeks 3 & 4)
- Post-hoc explanations of black-box models and their vulnerabilities (Weeks 5 - 9)
- Theory + Connections with adversarial robustness, fairness, DP (Weeks 10 – 11)
- Understanding and Reasoning in Large Language Models and Foundation Models (Weeks 12 – 14)

Course Readings

- Tentative readings and schedule already on canvas
- A calendar with all timelines/deadlines will be posted on canvas by the end of this week

Who should take this course?

- Course particularly tailored to students interested in research on interpretability/explainability
- Not a surface level course!
  - Not just applications!
- Goal is to push you to question existing work and make new contributions to the field
- 11 research papers came out of projects from previous iterations of this course!
  - NeurIPS, ICML, AIES

Class Format

- Course comprises of lectures by instructor, guests, and student presentations
- Each lecture will cover:
  - 2 papers (50 mins + 10 mins discussion)
  - 10 mins of (random) breakout groups
  - 5 mins of class discussion + Conclusion
- Students are expected to “at least” skim through the papers beforehand
- Each breakout group is expected to come up with:
  - a list of 2 to 3 weaknesses of each of the works discussed and
  - Strategies for addressing those weaknesses

Course Components

- Research project (60%)
  - 3 checkpoints (10% each) – Proposal + Baseline Implementation + Midterm Progress
  - Final Report (20%)
  - Final Presentation (10%)
  - Teams of 2 to 3
- Research Paper Presentation (30%)
  - Teams of 2 to 3; Each team presents two papers in the class
- Class Participation (10%)
  - Being active in discussions in class
  - Attending classes

Research Projects

- Biggest component of the course
- We will release:
  - a list of problem statements next week,
  - all the final project reports from last year's course
- You can either work on the problem statements we provide, or come up with your own
- Note that all problem statements need to be approved by our teaching team
  - Please talk to us before submitting your proposal and make sure we approve the problem
  - Otherwise, you may be asked to change your problem/rewrite your proposal

Project Milestones

- Proposal (10%)
  - 2 page overview of your project direction
  - What is the problem?
  - Why haven't prior works solved it?
  - How do you propose to solve it?
  - How will you measure success?

- Baseline Implementation (10%)
  - Implement a baseline for your project
  - (Preferably) one of the papers in the course
  - Implement the paper's core methodology, reproduce results, critique it and discuss how you would improve

Project Milestones

- Midterm Progress (10%)
  - 2 to 3 page update
  - Formal problem statement
  - Detailed solution
  - Preliminary results
- Final Report (20%)
  - 5 to 6 page expanded writeup
  - Formal problem statement
  - Detailed solution
  - Thorough empirical results
  - Findings and conclusions

Background

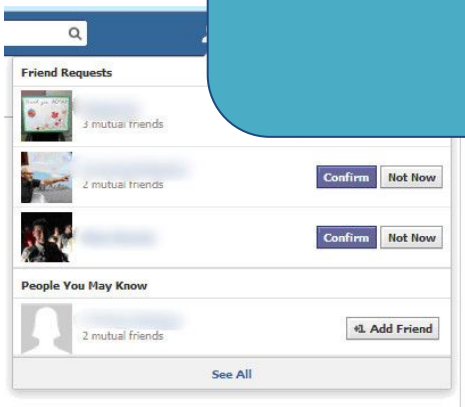
- Strong understanding: Linear algebra, probability, algorithms, machine learning (cs181 or equivalent), programming in python, numpy, sklearn;

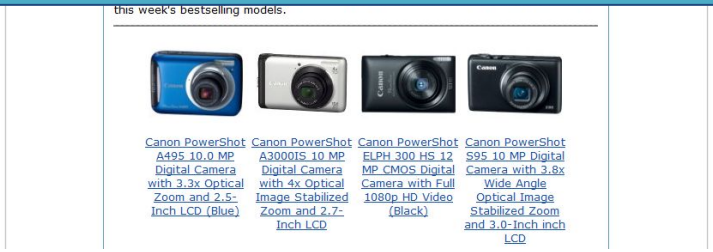
- Familiarity with statistics, optimization

Questions??

Motivation

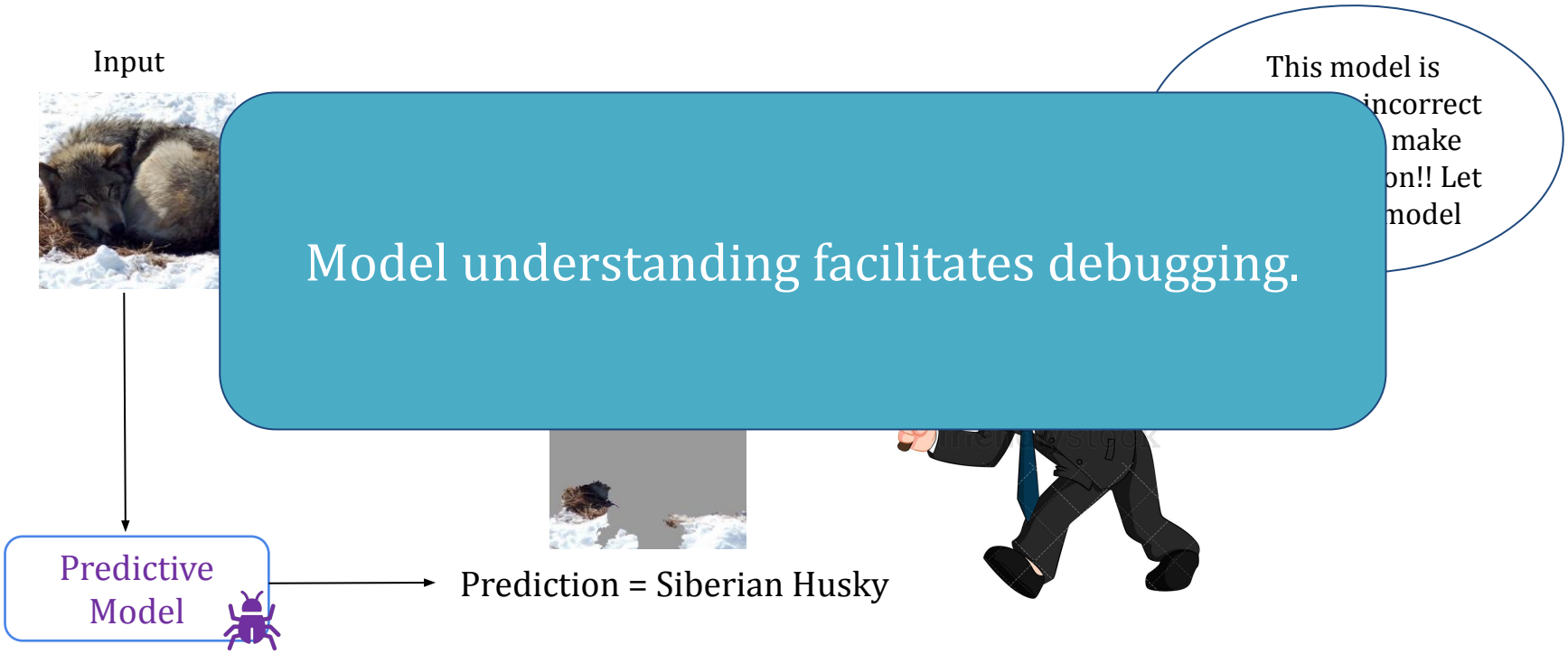
Machine Learning is EVERYWHERE!!





[Weller 2017]

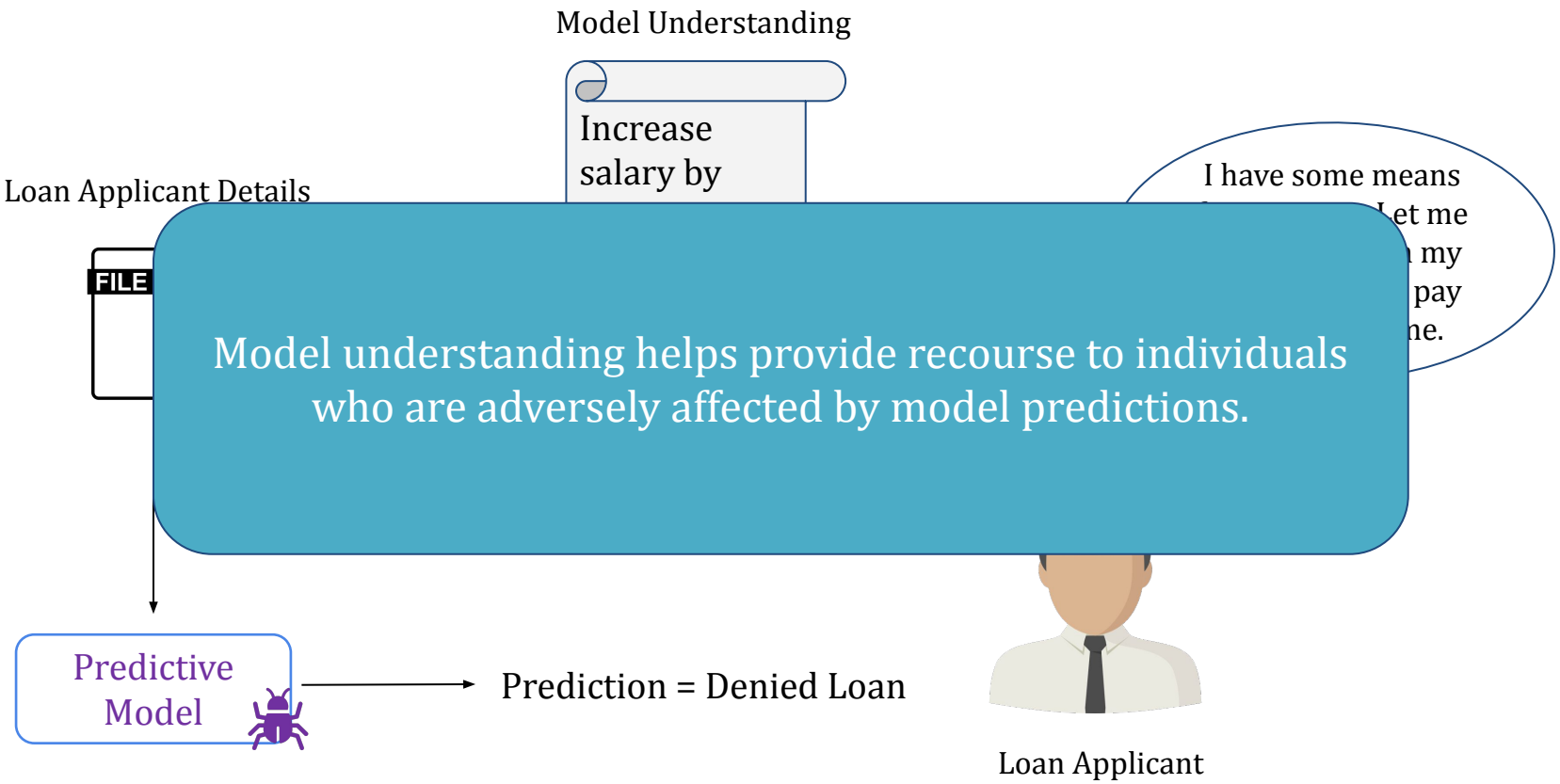
Motivation: Why Model Understanding?



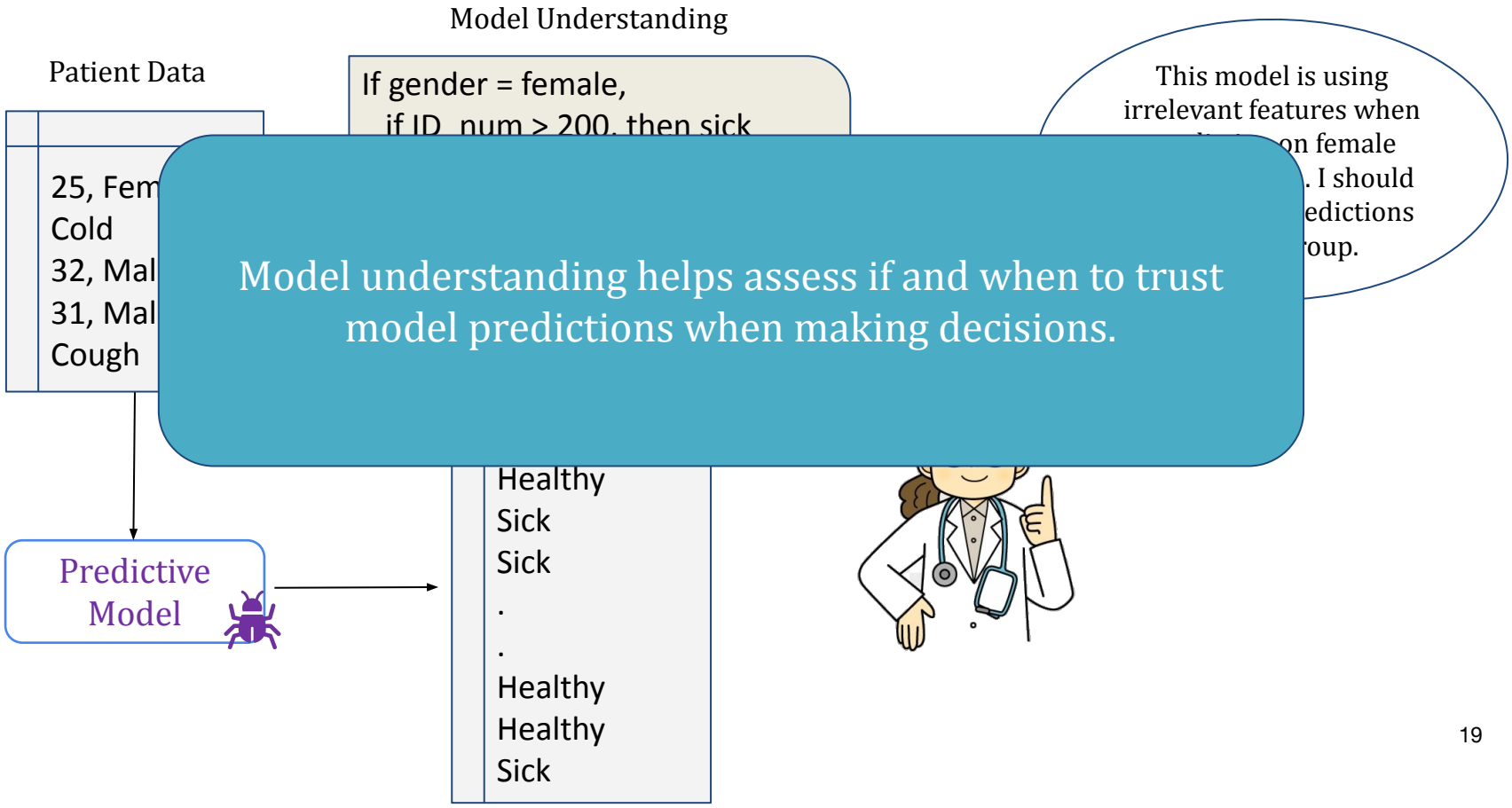
Motivation: Why Model Understanding?

[Larson et. al. 2016]

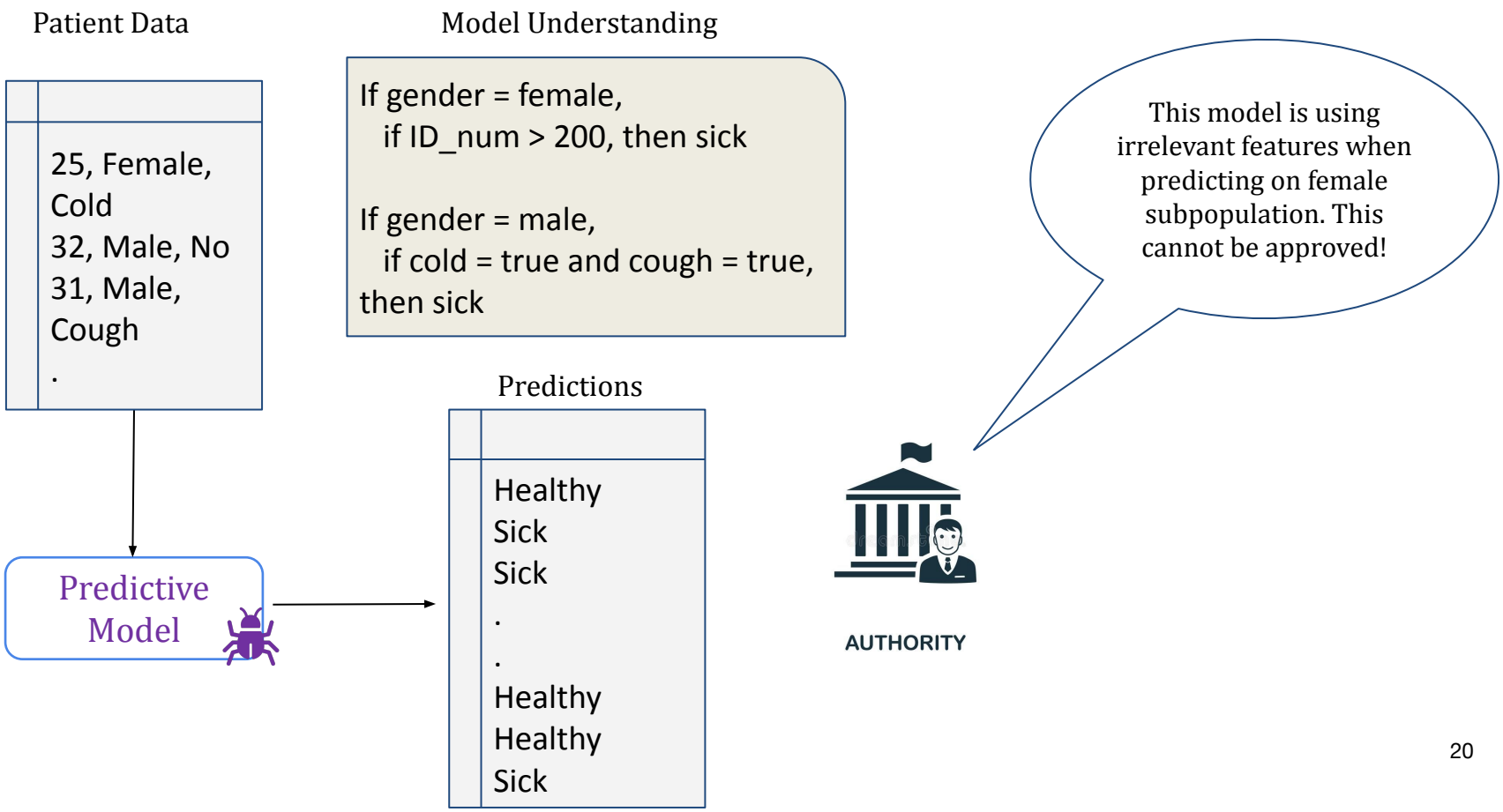
Motivation: Why Model Understanding?



Motivation: Why Model Understanding?



Motivation: Why Model Understanding?



Motivation: Why Model Understanding?

Utility

Debugging

Bias Detection

Recourse

If and when to trust model predictions

Vet models to assess suitability for deployment

Stakeholders

End users (e.g., loan applicants)

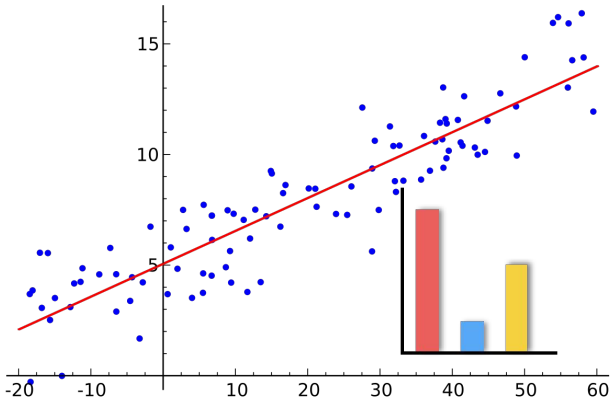
Decision makers (e.g., doctors, judges)

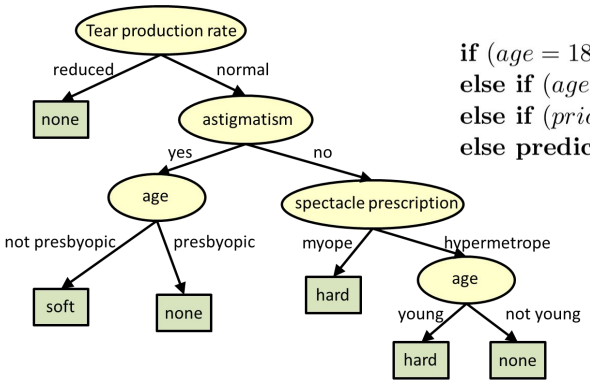
Regulatory agencies (e.g., FDA, European commission)

Researchers and engineers

Achieving Model Understanding

Take 1: Build *inherently interpretable* predictive models



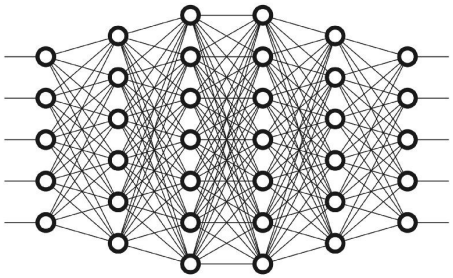


if $(age = 18 - 20)$ **and** $(sex = male)$ **then predict** yes
else if $(age = 21 - 23)$ **and** $(priors = 2 - 3)$ **then predict** yes
else if $(priors > 3)$ **then predict** yes
else predict no

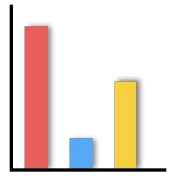
[Letham and Rudin 2015; Lakkaraju et. al. 2016]

Achieving Model Understanding

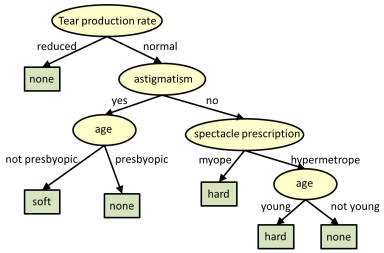
Take 2: *Explain* pre-built models *in a post-hoc manner*



Explainer



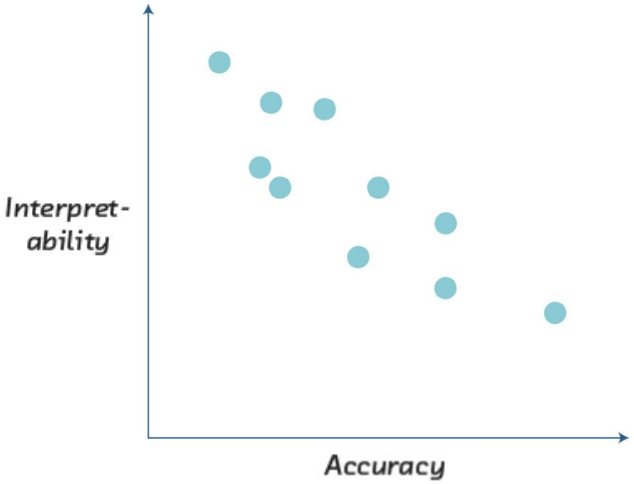
if $(age = 18 - 20)$ **and** $(sex = male)$ **then predict** yes
else if $(age = 21 - 23)$ **and** $(priors = 2 - 3)$ **then predict** yes
else if $(priors > 3)$ **then predict** yes
else predict no

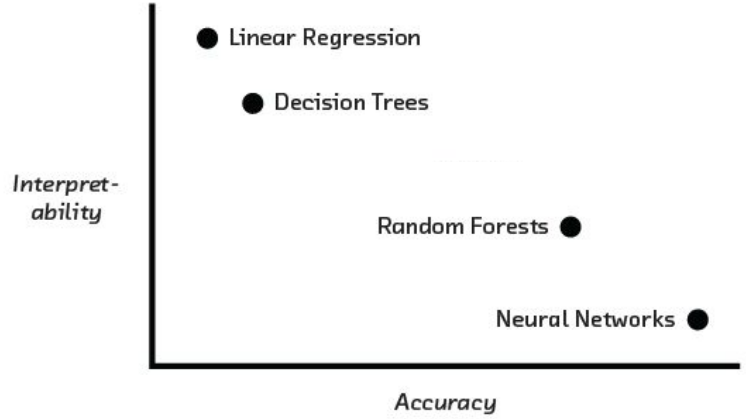


[Ribeiro et. al. 2016, 2018; Lakkaraju et. al. 2019]

Inherently Interpretable Models vs. Post hoc Explanations

Example

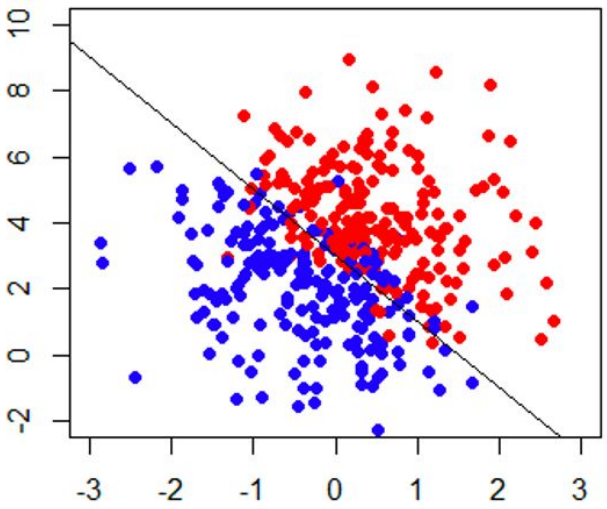




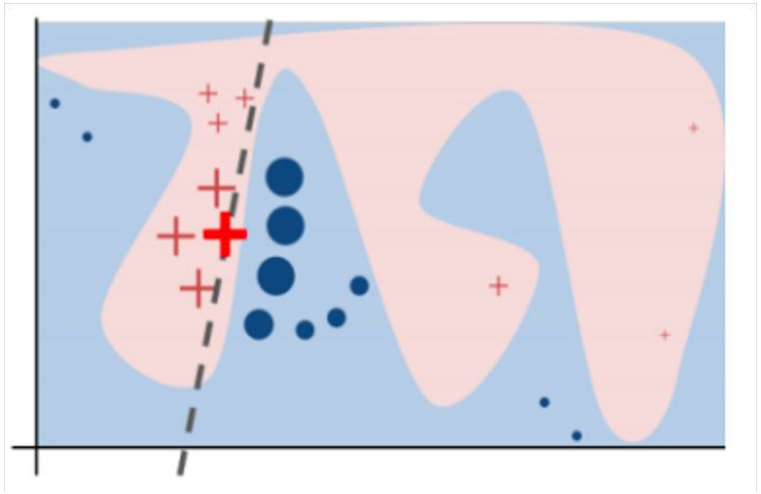
In ***certain*** settings, *accuracy-interpretability trade offs* may exist.

[Cireşan et. al. 2012, Caruana et. al. 2006, Frosst et. al. 2017, Stewart 2020]

Inherently Interpretable Models vs. Post hoc Explanations



can build interpretable + accurate models



complex models might achieve higher accuracy

Inherently Interpretable Models vs. Post hoc Explanations

Sometimes, you don't have enough data to build your model from scratch.

And, all you have is a (proprietary) black box!

[Ribeiro et. al. 2016]

Inherently Interpretable Models vs. Post hoc Explanations

If you *can build* an interpretable model which is also adequately accurate for your setting, DO IT!

Otherwise, ***post hoc explanations*** come to the rescue!

Let’s get into some details!

Next Up!

- Define and evaluate interpretability
  - somewhat! ☺️
- Taxonomy of interpretability evaluation
- Taxonomy of interpretability based on applications/tasks
- Taxonomy of interpretability based on methods

[Doshi-Velez & Kim, 2017]

Motivation for Interpretability

- ML systems are being deployed in complex high-stakes settings
- Accuracy alone is no longer enough
- Auxiliary criteria are important:
  - Safety
  - Nondiscrimination
  - Right to explanation

Motivation for Interpretability

- Auxiliary criteria are often hard to quantify (completely)
  - E.g.: Impossible to enumerate all scenarios violating safety of an autonomous car

- Fallback option: interpretability
  - *If the system can explain its reasoning, we can verify if that reasoning is sound w.r.t. auxiliary criteria*

Prior Work: Defining and Measuring Interpretability

- Little consensus on what interpretability is and how to evaluate it
- Interpretability evaluation typically falls into:
  - Evaluate in the context of an application
  - Evaluate via a quantifiable proxy

Prior Work: Defining and Measuring Interpretability

- Evaluate in the context of an application
  - If a system is useful in a practical application or a simplified version, it must be interpretable
- Evaluate via a quantifiable proxy
  - Claim some model class is interpretable and present algorithms to optimize within that class
  - E.g. rule lists

You will know it when you see it!

Lack of Rigor?

Important to formalize these notions!!!

- Are all models in all "interpretable" model classes equally interpretable?
  - Model sparsity allows for comparison
- How to compare a linear model with a decision tree?
- Do all applications have same interpretability needs?

What is Interpretability?

- Defn: Ability to explain or to present in understandable terms to a human

- No clear answers in psychology to:
  - What constitutes an explanation?
  - What makes some explanations better than the others?
  - When are explanations sought?

When and Why Interpretability?

- Not all ML systems require interpretability
  - E.g., ad servers, postal code sorting
  - No human intervention

- No explanation needed because:
  - No consequences for unacceptable results
  - Problem is well studied and validated well in real-world applications trust system's decision

When do we need explanation then?

When and Why Interpretability?

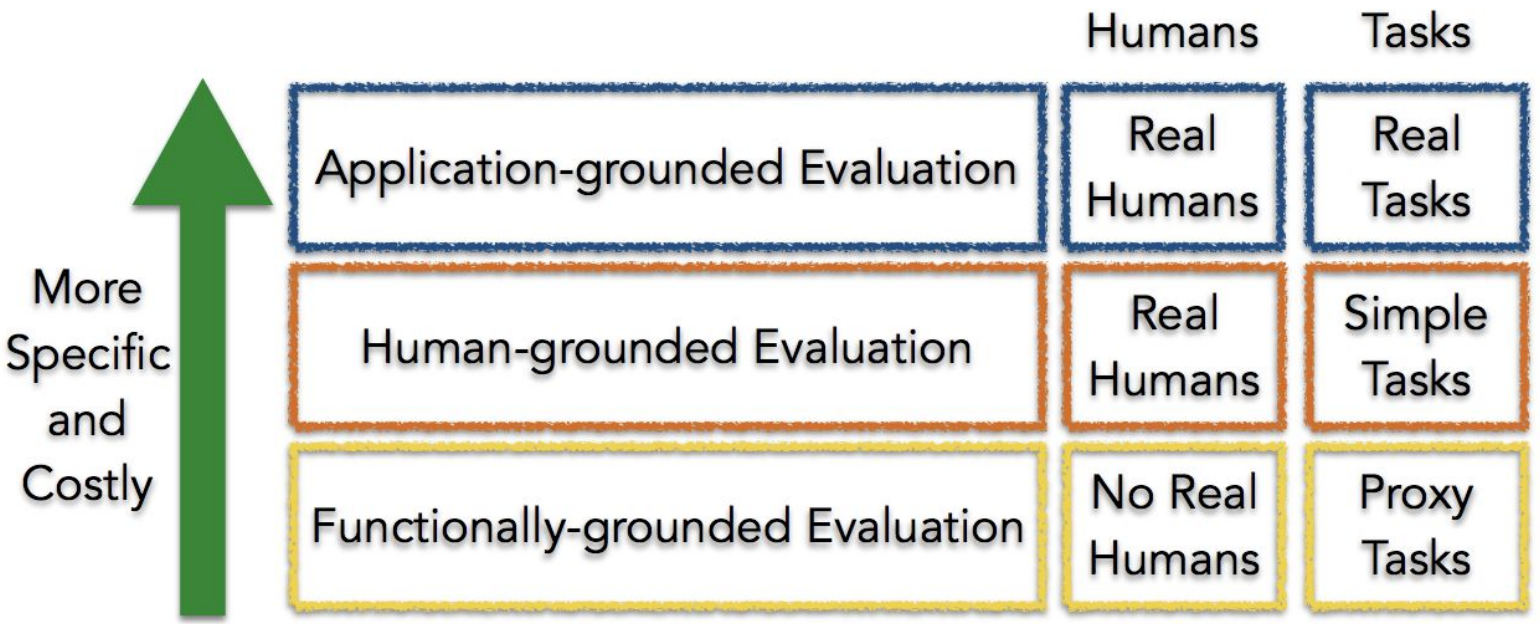
- *Incompleteness* in problem formalization
  - Hinders optimization and evaluation

- Incompleteness ≠ Uncertainty
  - Uncertainty can be quantified
  - E.g., trying to learn from a small dataset (uncertainty)

Incompleteness: Illustrative Examples

- Scientific Knowledge
  - E.g., understanding the characteristics of a large dataset
  - Goal is abstract
- Safety
  - End to end system is never completely testable
  - Not possible to check all possible inputs
- Ethics
  - Guard against certain kinds of discrimination which are too abstract to be encoded
  - No idea about the nature of discrimination beforehand

Taxonomy of Interpretability Evaluation



Claim of the research should match the type of the evaluation!

Application-grounded evaluation

- Real humans (domain experts), real tasks
- Domain experts experiment with exact application task
- Domain experts experiment with a simpler or partial task
  - Shorten experiment time
  - Increases number of potential subjects
- Typical in HCI and visualization communities

Human-grounded evaluation

- Real humans, simplified tasks
  - Can be completed with lay humans
  - Larger pool, less expensive

- Potential experiments
  - Pairwise comparisons
  - Simulate the model output
  - What changes should be made to input to change the output?

Functionally-grounded evaluation

- No humans, just proxies
  - Appropriate for a class of models already validated
    - E.g., decision trees
  - A method is not yet mature
  - Human subject experiments are unethical
  - What proxies to use?

- Potential experiments
  - Complexity (of a decision tree) compared to other other models of the same (similar) class
    - How many levels? How many rules?

Open Problems: Design Issues

- What proxies are best for what real world applications?
- What factors to consider when designing simpler tasks in place of real world tasks?

Taxonomy based on applications/tasks

- Global vs. Local
  - High level patterns vs. specific decisions
- Degree of Incompleteness
  - What part of the problem is incomplete? How incomplete is it?
  - Incomplete inputs or constraints or costs?
- Time Constraints
  - How much time can the user spend to understand explanation?

Taxonomy based on applications/tasks

- Nature of User Expertise
  - How experienced is end user?
  - Experience affects how users process information
  - E.g., domain experts can handle detailed, complex explanations compared to opaque, smaller ones

- Note: These taxonomies are constructed based on intuition and are not data or evidence driven. They must be treated as hypotheses.

Taxonomy based on methods

- Basic units of explanation:
  - Raw features? E.g., pixel values
  - Semantically meaningful? E.g., objects in an image
  - Prototypes?

- Number of basic units of explanation:
  - How many does the explanation contain?
  - How do various types of basic units interact?
  - E.g., prototype vs. feature

Taxonomy based on methods

- Level of compositionality:
  - Are the basic units organized in a structured way?
  - How do the basic units compose to form higher order units?
- Interactions between basic units:
  - Combined in linear or non-linear ways?
  - Are some combinations easier to understand?
- Uncertainty:
  - What kind of uncertainty is captured by the methods?
  - How easy is it for humans to process uncertainty?

Questions??

Relevant Conferences to Explore

- ICML
- NeurIPS
- ICLR
- UAI
- AISTATS
- KDD
- AAAI
- FAccT
- AIES
- CHI
- CSCW
- HCOMP

Breakout Groups

- Say hi to your neighbors! Introduce yourselves!
- What topics are you most excited about learning as part of this course?
- Are you convinced that model interpretability/explainability is important?
- Do you think we can really interpret/explain models (correctly)?
- What is your take on inherently interpretable models vs. post hoc explanations? Would you favor one over the other? Why?


Thank you