Lecture I:
Course introduction

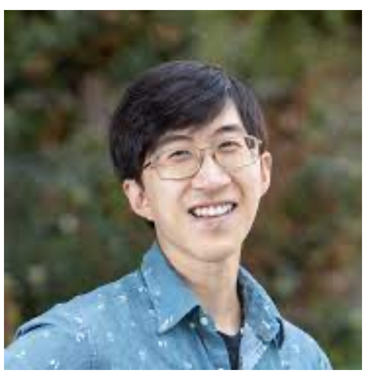
CSEP 590B: Explainable AI

Ian Covert & Su-In Lee
University of Washington

Teaching team

- **Su-In Lee**
  - Paul G. Allen Prof.
  - UW Allen School
  - Office hours: course website

- **Ian Covert**
  - Ph.D. candidate
  - UW Allen School
  - Office hours: course website



- **Hugh Chen**
  - Ph.D. student
  - UW Allen School
  - Office hours: course website

- **Chris Lin**
  - Ph.D. student
  - UW Allen School
  - Office hours: course website

Participation is important

- 50 people with diverse expertise
  - Microsoft 22
  - Amazon 6
  - Boeing 4
  - Google 3
  - Tableau 2
  - Apple, Avalara, Avnet, Best Buy, Comcast, D2K Technologies, Doctor.com, Facebook, Hulu, Petrin Technology Consulting LLC, Provoke Solutions, Salesforce, SAP Qualtrics, Twitter 1

FTE	Years
Min	2
Max	30
Mean	6.8
Median	4.5

- This course is a unique opportunity to share ideas about using XAI in industry
- Please do not hesitate to contact us with comments or feedback

Today

- Section 1
  - Motivation & aims
  - Course logistics
  - Examples in the healthcare space
- Section 2
  - Discussion: “Statistical Modeling: The Two Cultures”
- Section 3
  - Example scenario and ML review

Traditional algorithms

Find customers for rewards program

```
gold = []
platinum = []
for customer in customers:
  if customer.savings > amount1:
    platinum.append(customer)
  elif customer.savings > amount2:
    gold.append(customer)
return gold, platinum
```

Program based on simple rules, written by people

Machine learning algorithms

Decide if customer should be given a loan

```
default_risk = risk_prob(customer)
if default_risk < thresh:
  return True
else:
  return False
```

Based on many inputs (savings, age, income, job, homeownership…)

No simple rules → **learned from data**

ML now used for many problems

- Credit risk
- Medical diagnosis
- Biomedical discoveries
- Recidivism prediction
- Job candidate screening
- Content recommendation
- Ad targeting
- Search engines
- Text summarization

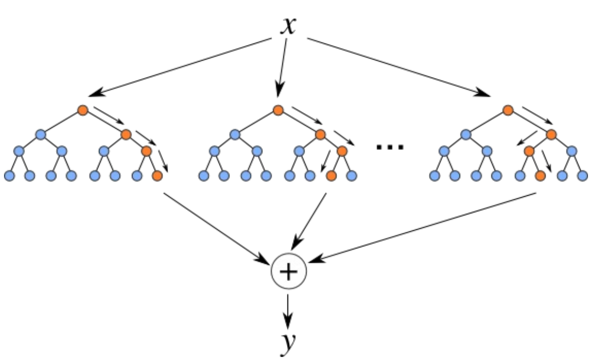
⋮

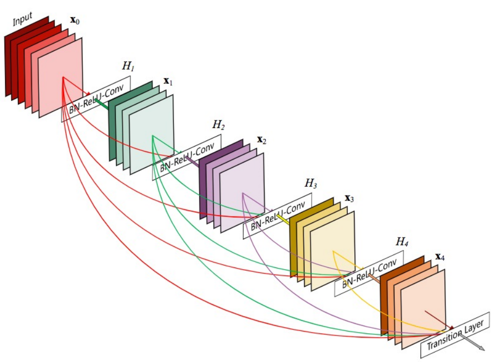
Technology has improved

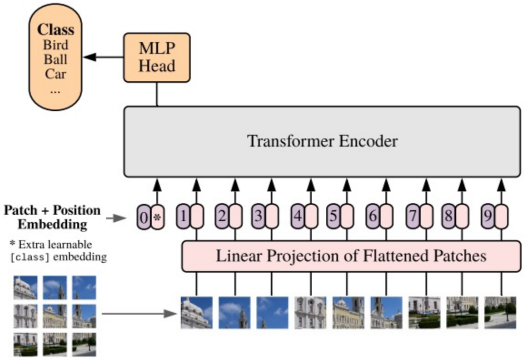
Compute

Data

Models



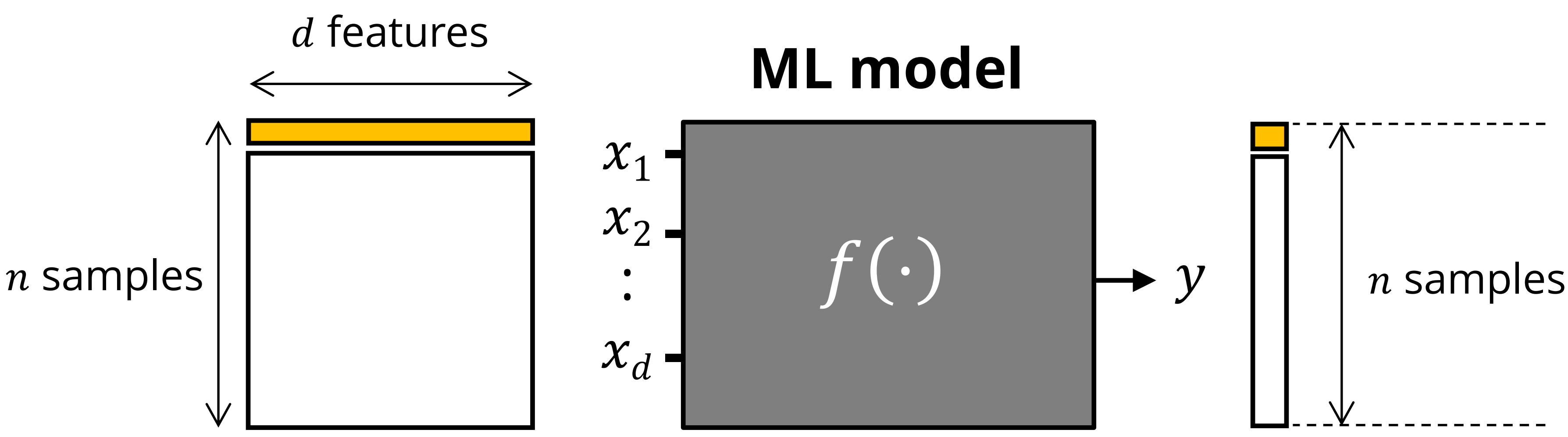




Learning ML models from data: supervised learning

ML ingredients

- Function class $f(x)$ to describe y based on x
- Training data with n samples of *features* x and *labels* y



ML today: black-box models

$$x \longrightarrow \boxed{f} \longrightarrow \hat{y}$$

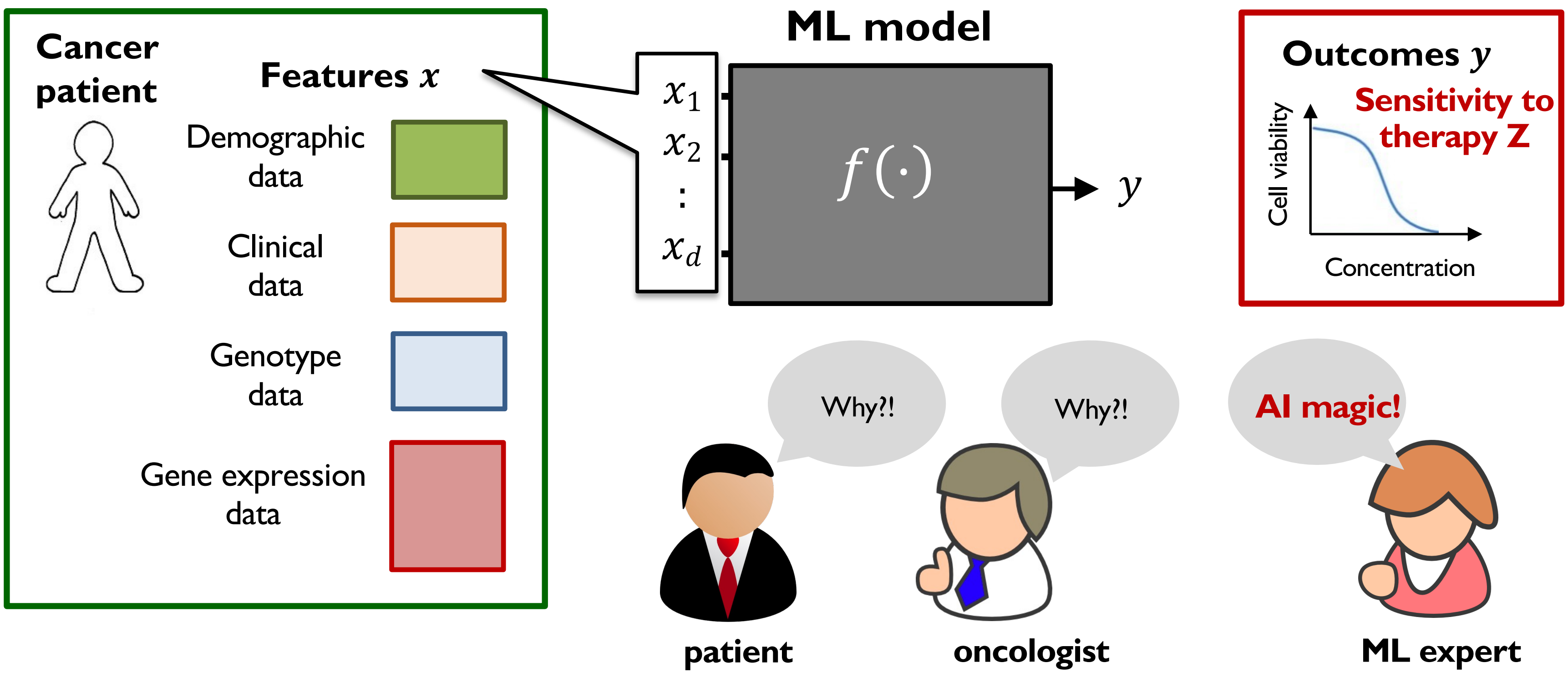
State-of-the-art accuracy

Performance tied to revenue, user experience, customer retention, etc.

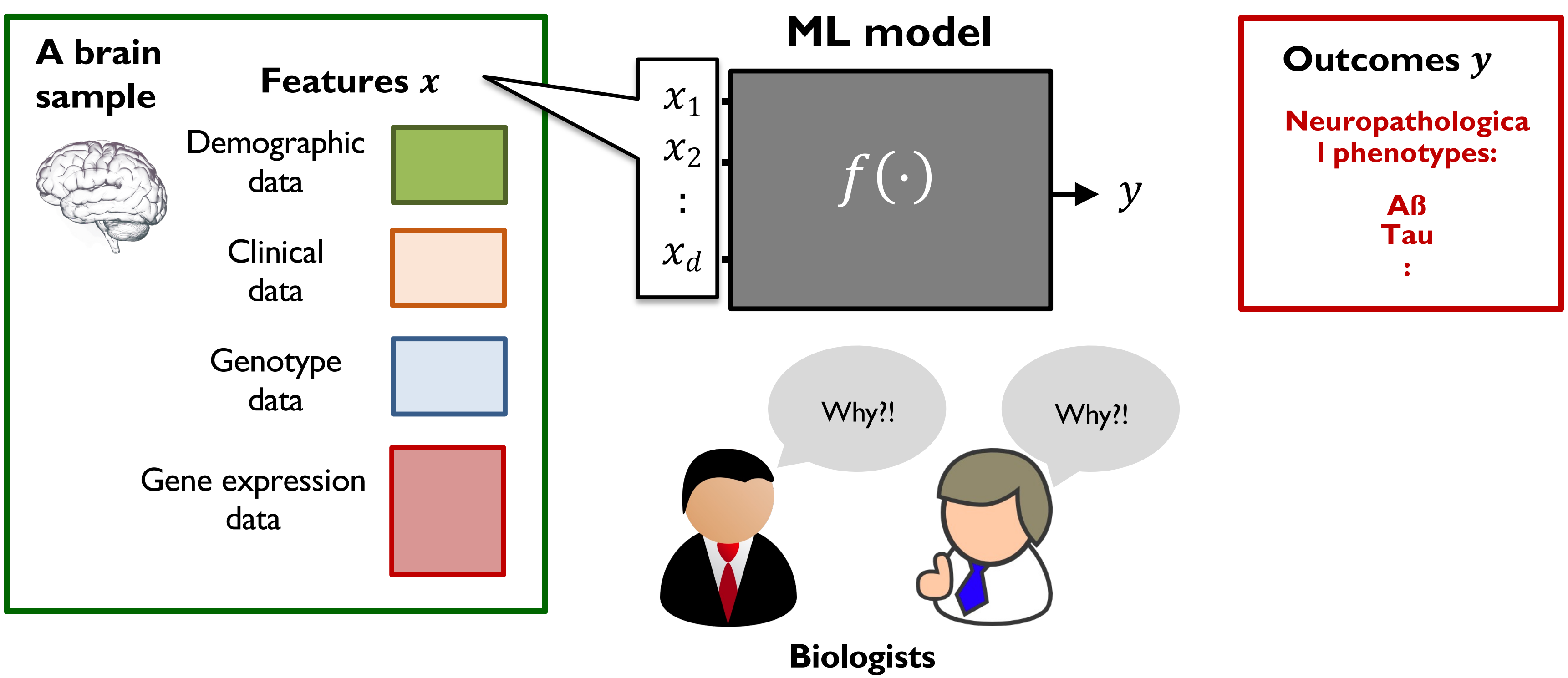
However, lack of transparency!

- Identify key factors in underlying process
- Generate scientific hypotheses

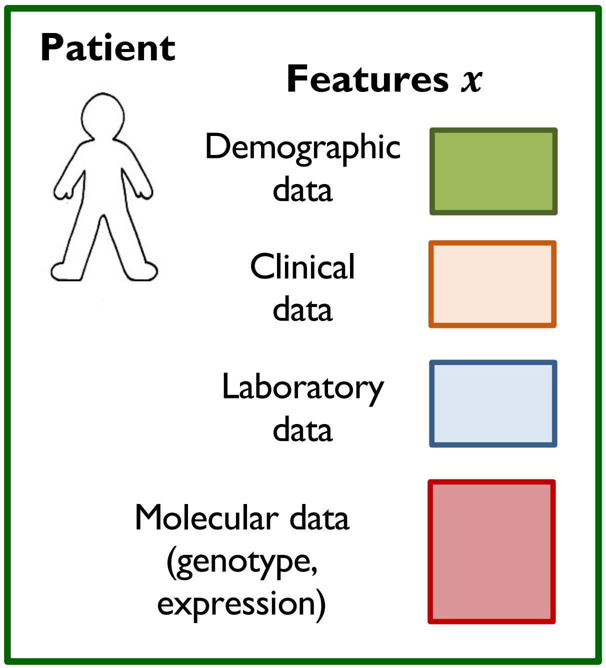
Accurately predicting clinical outcomes is important, but the key question is *why*

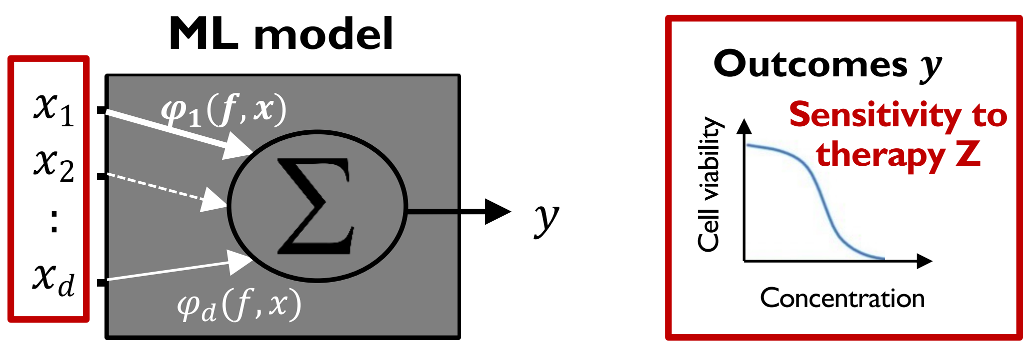


Accurately modeling biological phenotypes is important, but the key question is *why*



Explainable AI for biology and health





- **Explainability** can be as important as accuracy
  - Which features contributed to a certain prediction and how?
  - How to learn or select features that are most interpretable or informative?
  - How to make biological or clinical sense of a black-box model?

ML today: black-box models

State-of-the-art accuracy

Performance tied to revenue, user experience, customer retention, etc.

Transparency goals

- Identify key factors in underlying process
- Generate scientific hypotheses
- Diagnose model failures
- Audit for unwanted dependencies
- Enable regulation
- Improve dataset
- Build user, organizational trust
- Inform users of recourse options
- Pinpoint shortcuts, spurious signals

Transparency for any model

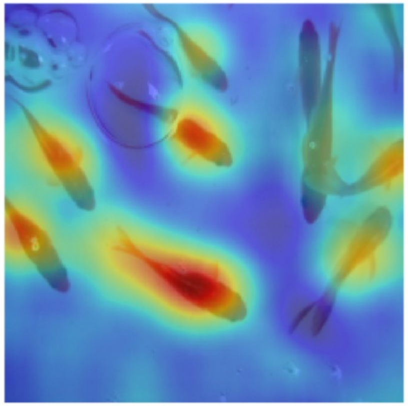
Explainable AI (XAI) ingredients

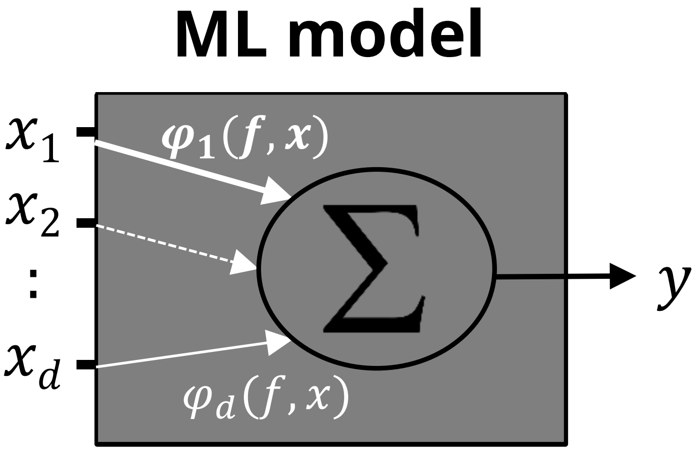
- **Model:** predictive model, possibly black-box (such as DNN, GBM, etc.)
- **Data:** individual data sample, or entire dataset
- **Question:** what do you need to understand?

Design algorithms to answer specific questions about ML models

Types of XAI questions

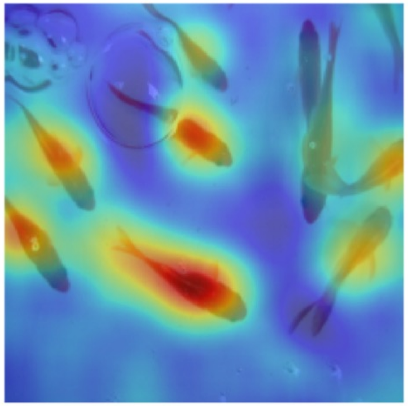
Feature importance



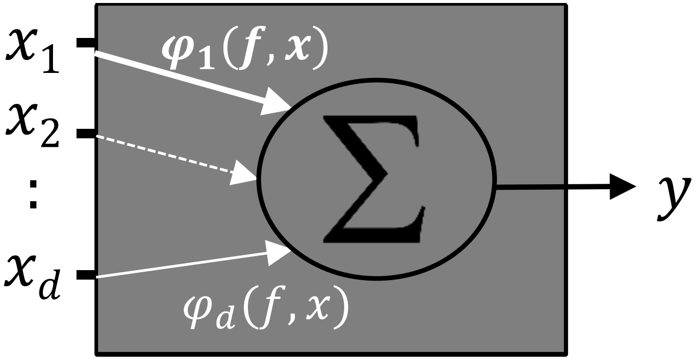


Types of XAI questions

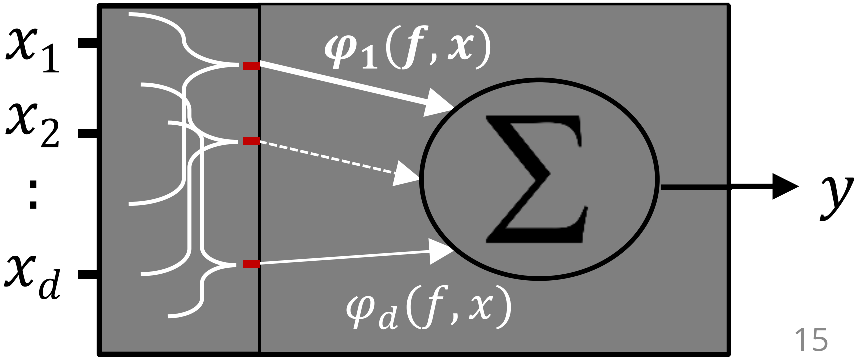
Feature importance



ML model

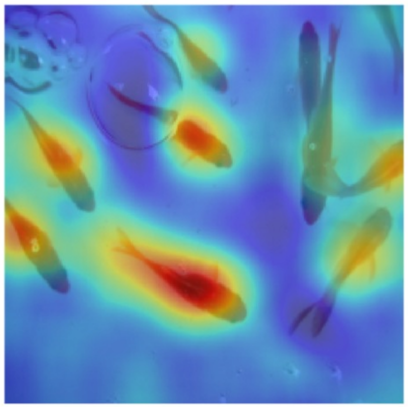


Concept

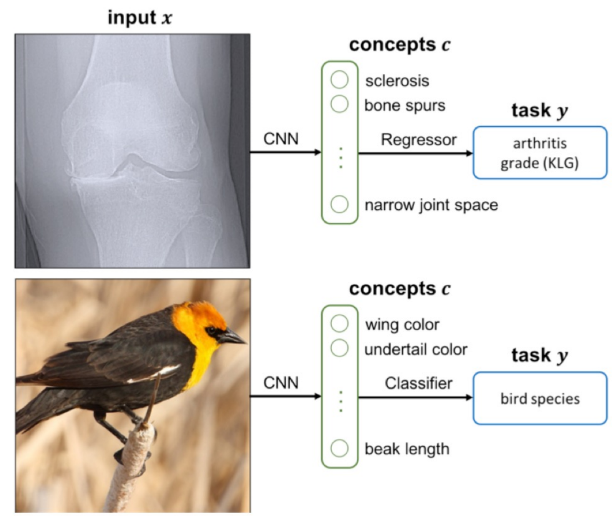


Types of XAI questions

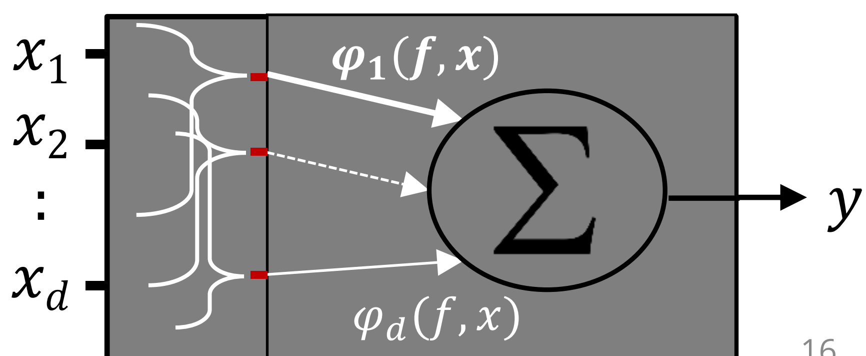
Feature importance



Important concepts

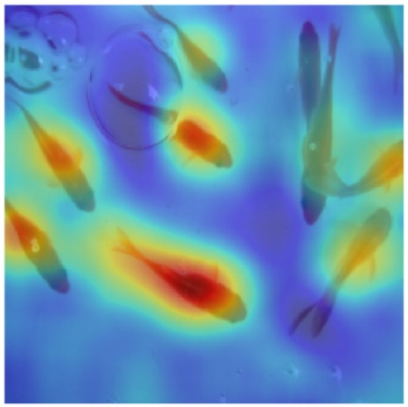


Concept

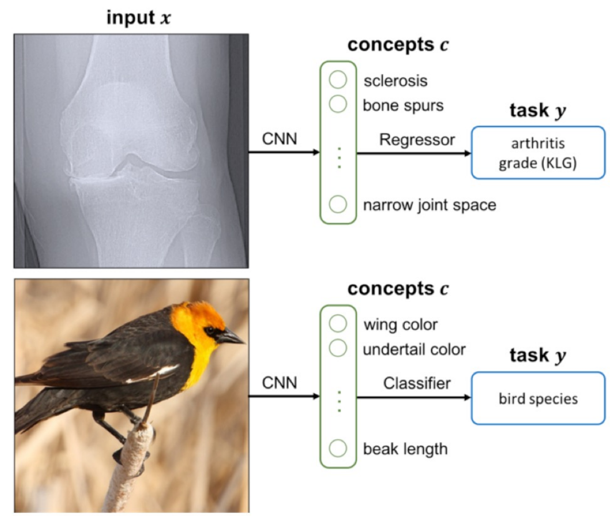


Types of XAI questions

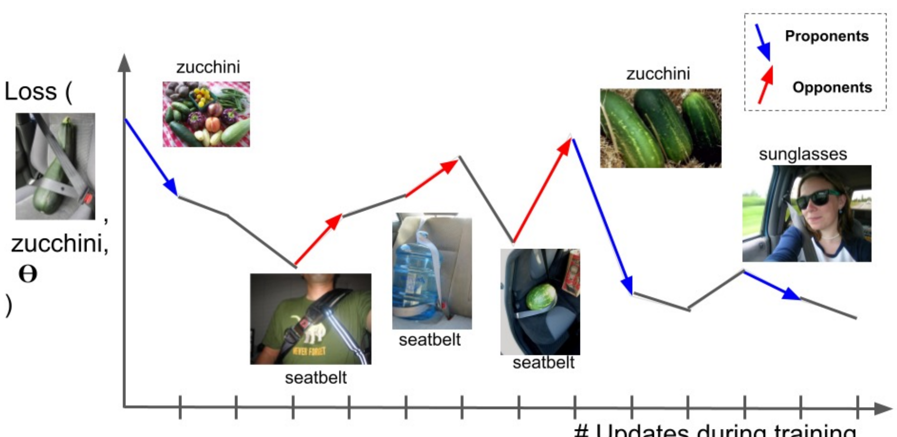
Feature importance



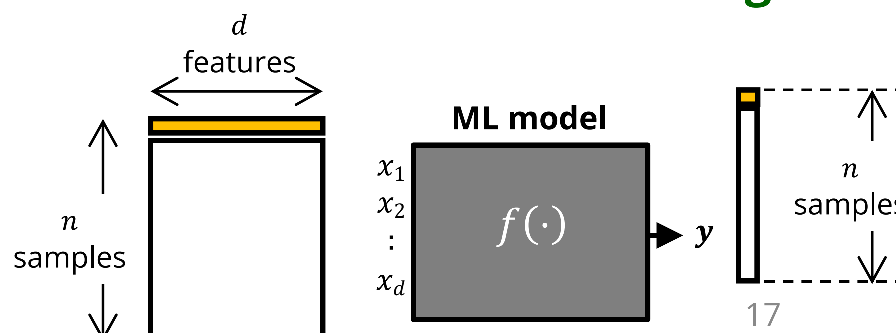
Important concepts



Data (instance) importance



How much does each sample contribute to model training?



Course goals

What to expect

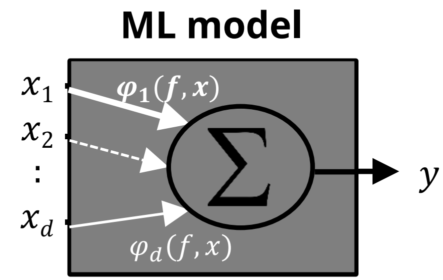
- **Broad overview:** learn about many areas of XAI, an emerging area of ML research
- **Preparation for learning:** the field is fast-moving, and we'll cover principles that help learn new techniques on your own
- **Usage experience:** get experience working with several popular tools and libraries (e.g., shap)

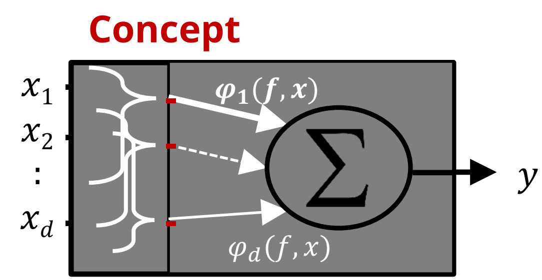
Prerequisites

- Background in calculus, probability, linear algebra
- ≥1 previous ML courses (ugrad or grad level)
  - Most PMP students with a B.S. in Computer Science should fulfill the prerequisites. If you're unsure, please contact us

Course overview

- **Course introduction** (1 lecture)
- **Feature importance explanations** (3 lectures)
  - Removal-based explanations
  - Shapley values
  - Propagation-based explanations
- **Evaluating explanations** (1 lecture)
- **Inherently interpretable models** (1 lecture)
- **Other approaches** (2 lectures)
  - Concept-based explanations, neuron interpretation
  - Counterfactual explanations, instance explanations
- **Enhancing human-AI collaboration** (1 lecture)
- **XAI in industry, model improvement** (1 lecture)





Today

- Section 1
  - Motivation & aims
  - Course logistics
  - Examples in the healthcare space
- Section 2
  - Discussion: “Statistical Modeling: The Two Cultures”
- Section 3
  - Example scenario and ML review

Lecture format

Before: read a research paper, write discussion post

During (2 hrs 50 mins):

- **Discussion** (50 mins) followed by 10 min break
- **Lecture part 1** (50 mins) followed by 10 min break
- **Lecture part 2** (50 mins)

Textbooks

- The field is young and rapidly changing
  - No single textbook covers all relevant content
  - Christoph Molnar's e-book is pretty good
- Instead, we'll read and discuss recent papers
  - Active vs. passive learning
  - Keep up with new content, practice reading papers
  - **Student-led discussions**

Grading basis

- 50% Homework
- 40% Paper discussions
  - Discussion board posts and in-class discussion
  - Bonus: leading the discussion
- 10% In-class participation

Homework assignments

- HW 0 (30 points, **due next Monday 11:59 pm**)
  - Refresher on probability, calculus, ML models

- HW 1, 2, 3 (100 points each)
  - Several problems, including math and programming
  - You'll have several weeks for each assignment, but start early

Homework policies

- **Collaboration:**
  - Students must submit their own answers and their own code for programming problems
  - Limited collaboration is allowed, but you must indicate on the homework with whom you collaborated
- **Late policy:**
  - Homeworks must be submitted online on Canvas by the posted due date
  - The penalty for late work is 20 points per day, and each student gets 3 free late days for the quarter

Homework tools

- Submitted electronically on Canvas, PDF format
  - Preference: Latex > Word > handwritten
- Programming in Python only
- We'll use open-source software:
  - numpy, pandas, matplotlib
  - sklearn, Pytorch/Tensorflow
  - shap, lime

Discussion posts

- Read the paper prior to class
- Discussion post due the night before
  - What is the paper about?
  - How does the method work?
  - What questions does it answer about a model?
  - How does it differ from other methods we've discussed?
  - Does it seem technically sound? What concerns do you have?
  - Could you use it in your job? How?
- Graded on 0-2 scale

In-class discussion

- Led by two student volunteers
  - Prepare slides, a short summary of the paper
  - Suggest topics to discuss about the method, its evaluation, etc.
  - Beyond the paper: broader questions, utility in business or scientific research, relationship with other methods
- **We need volunteers!**
  - Bonus points towards final grade
  - We'll email a link to schedule spreadsheet

Course information

- Course website (link)
  - Required readings, course information
- Ed discussion board (link)
  - Discussion post submission, HW questions, etc.
- Canvas (link)
  - Lecture slides, HW submission
- Mailing list: csep590b_sp22@cs.washington.edu

Course schedule

Student-led discussions begin →

1	3/29
2	4/5
3	4/12
4	4/19
5	4/26
6	5/3
7	5/10
8	5/17
9	5/24
10	5/31

- 4/4 (Mon) HW0: Warm-up (30 points)
- 4/25 (Mon) HW1: Feature importance (100 points)
- 5/16 (Mon) HW2: TBD (100 points)
- 5/30 (Mon) HW3: TBD (100 points)

Course overview

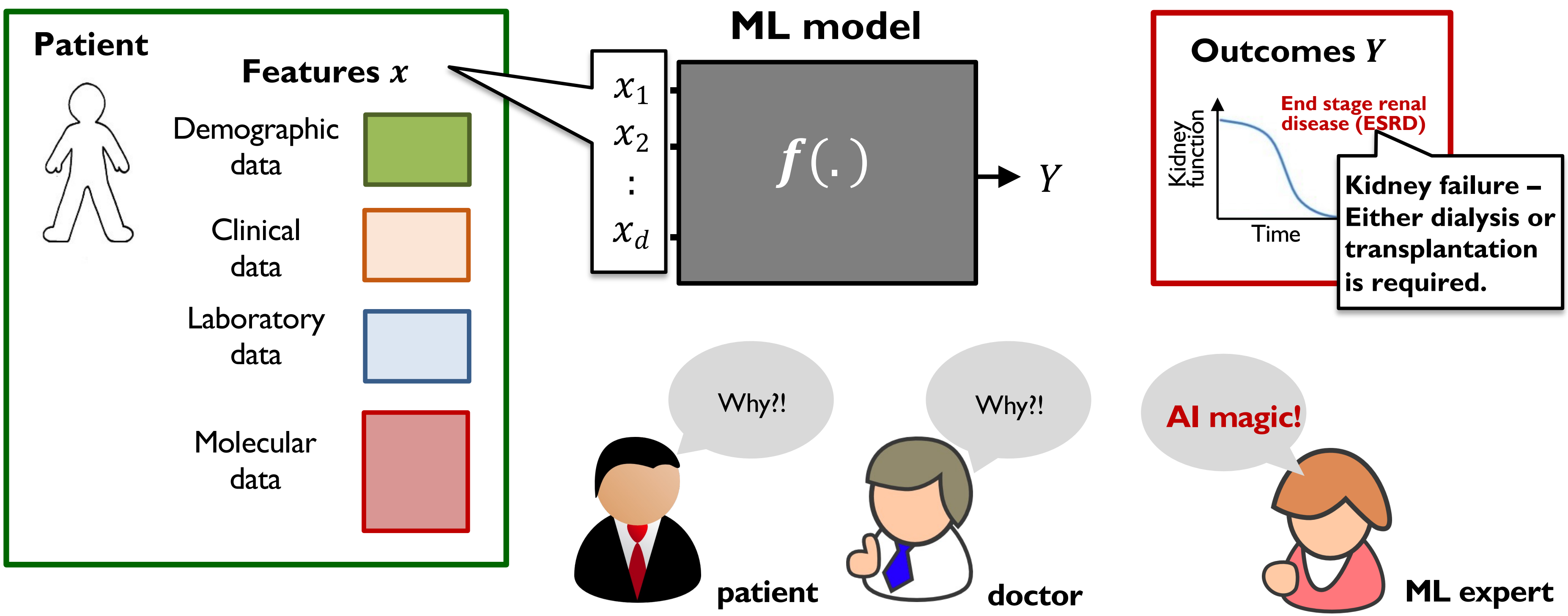
Student-led discussions begin

1	3/29	▪ **Course introduction** (1 lecture)
2	4/5	▪ **Feature importance explanations** (3 lectures) ▪ Removal-based explanations; Shapley values; Propagation-based explanations
3	4/12	
4	4/19	
5	4/26	▪ **Evaluating interpretability** (1 lecture)
6	5/3	▪ **Inherently interpretable models** (1 lecture)
7	5/10	▪ **Other approaches** (2 lectures) ▪ Concept-based explanations; instance explanations
8	5/17	
9	5/24	▪ **Enhancing human-AI collaboration** (1 lecture)
10	5/31	▪ **XAI in industry, model improvement** (1 lecture)

Today

- Section 1
  - Motivation & aims
  - Course logistics
  - Examples in the healthcare space
- Section 2
  - Discussion: “Statistical Modeling: The Two Cultures”
- Section 3
  - Example scenario and ML review

Accurately predicting clinical outcomes is important, but the key question is *why*



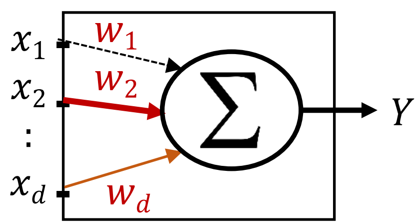
Scott Lundberg, […], and Su-In Lee. **Explainable AI for Trees: From Local Explanations to Global Understanding.** ***Nature Machine Intelligence (2020)*** **– Cover article**

Our solution: a technique that can explain any prediction

- Accuracy vs. interpretability
  - Simple models often lead to worse performance
  - Complex models are often considered to be a black box

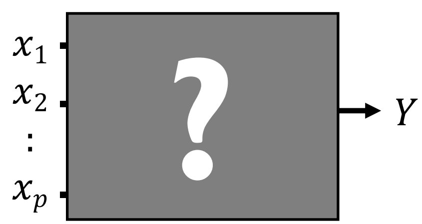
Linear model

x: Features Y: Outcome



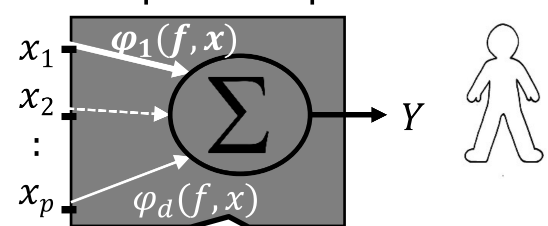
Complex model $f(\cdot)$

Black Box



Our approach, SHAP

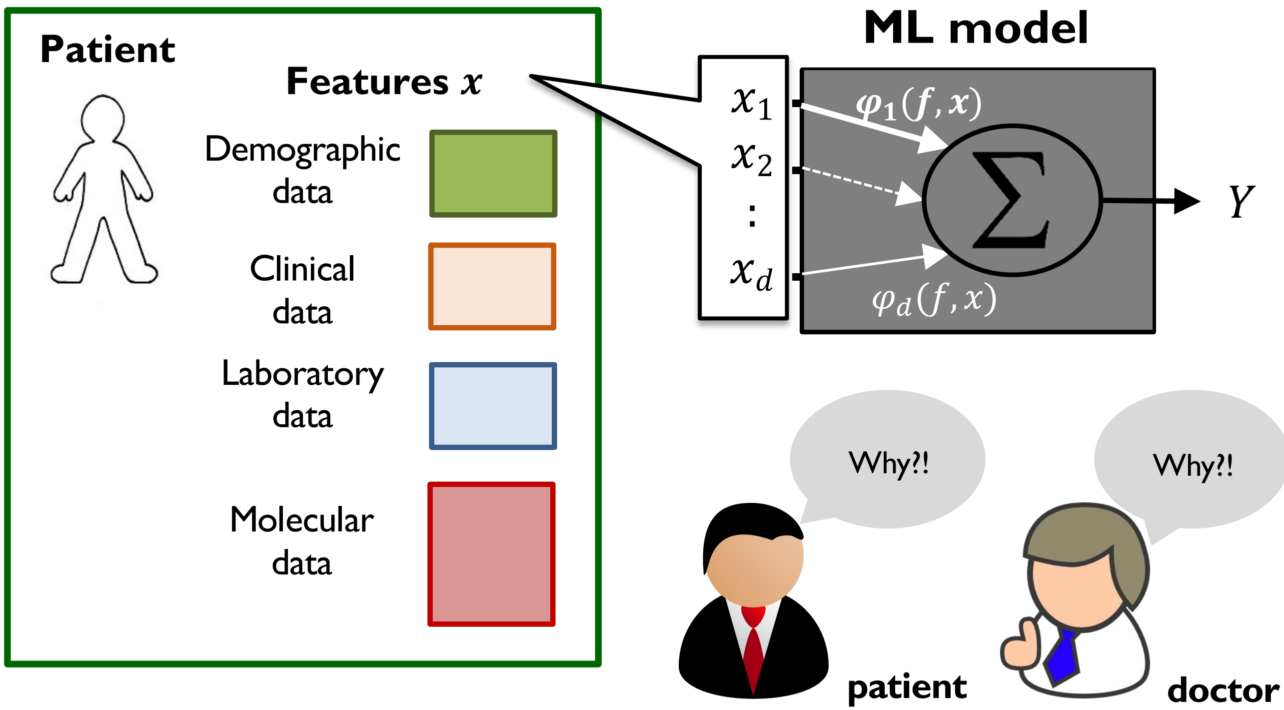
For a particular prediction



- SHAP can calculate feature importance for a particular prediction for any model

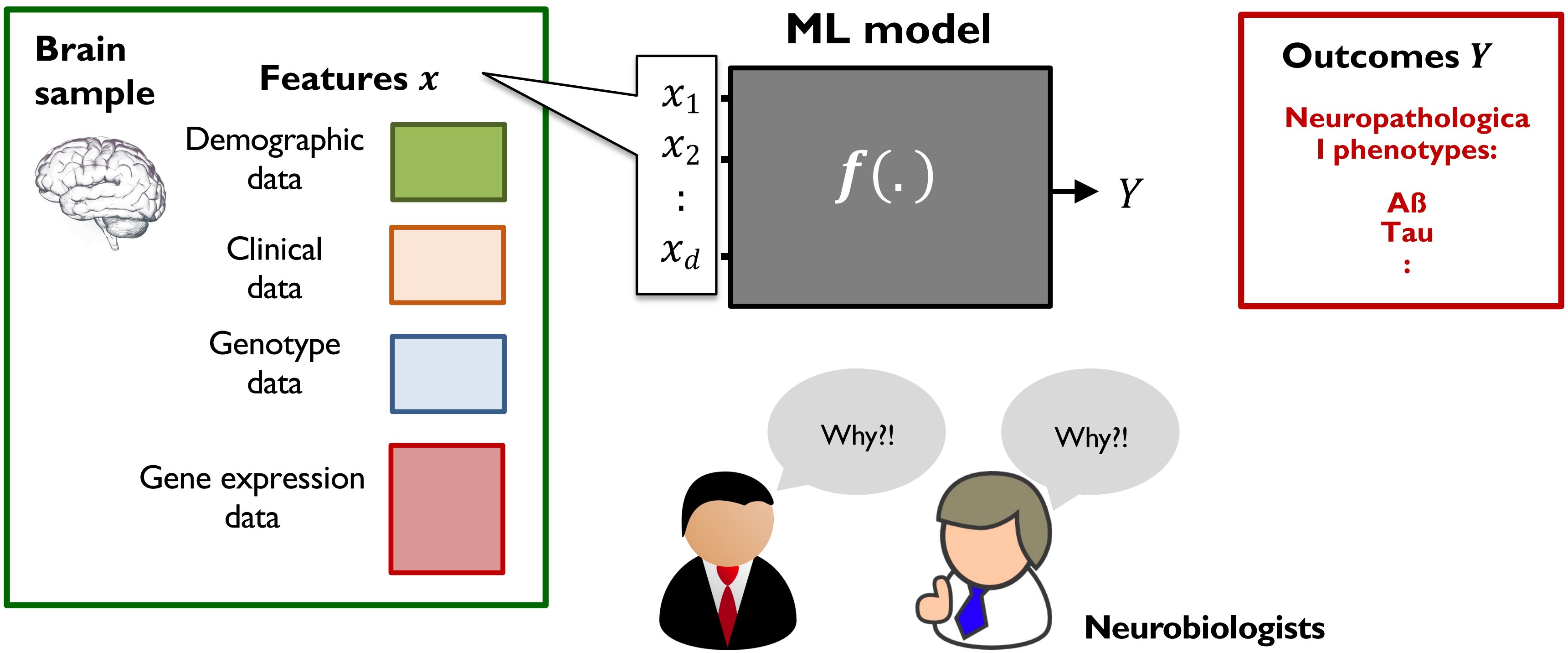
Scott Lundberg and Su-In Lee. **A Unified Approach to Interpreting Model Predictions**. ***NeurIPS (2017)*** – **Oral presentation**
NeurIPS workshop on Interpretable ML (2016) – **Best paper award**

Accurately predicting clinical outcomes is important, but the key question is *why*



Scott Lundberg, et al. **Explainable AI for Trees: From Local Explanations to Global Understanding.** ***Nature Machine Intelligence (2020)*** **– Cover article**

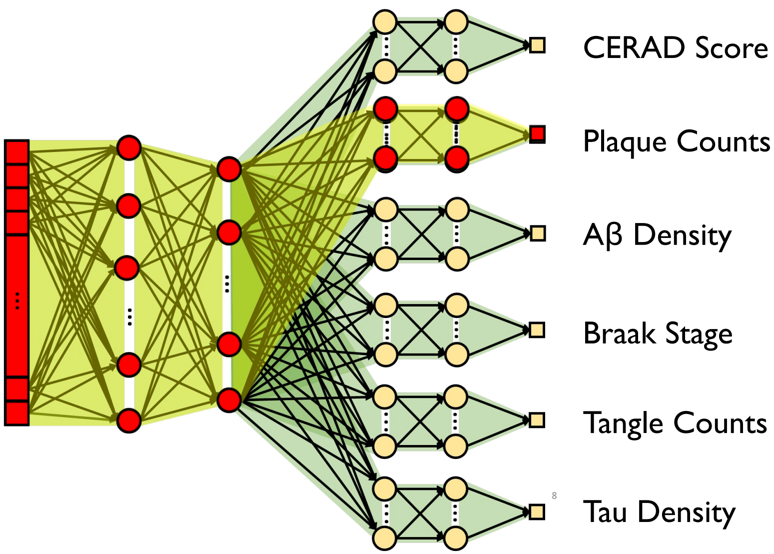
Identifying expression markers for phenotypes is important, but the key question is the *mechanistic explanation*

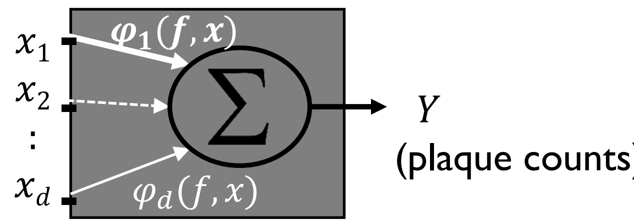


Nicasia Beebe-Wang, […], Sara Mostafavi* and Su-In Lee.* **Unified AI framework to uncover deep interrelationships between gene expression and Alzheimer's disease neuropathologies.** ***Nature Communications (2021)***

Identifying genes that are important to neuropathological phenotypes

- XAI methods can uncover each gene's contribution to the output variables

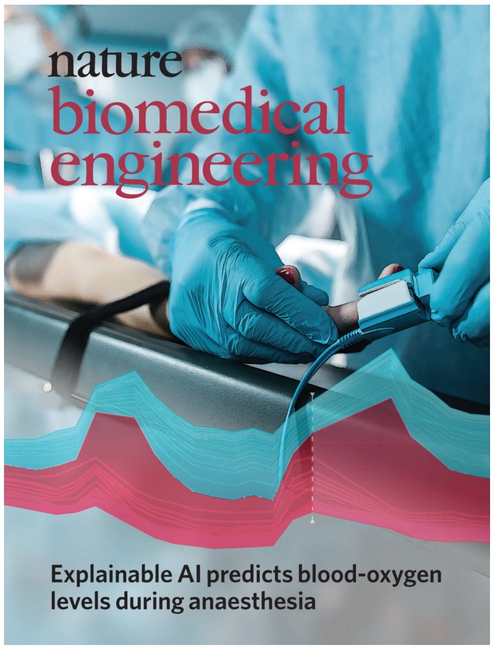




- Previously unknown sex-specific associations between immune response genes and AD neuropathological phenotypes

Nicasia Beebe-Wang, […], Sara Mostafavi* and Su-In Lee.* **Unified AI framework to uncover deep interrelationships between gene expression and Alzheimer's disease neuropathologies.** ***Nature Communications (2021)***

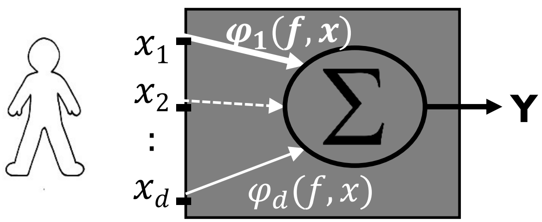
Providing *explainable predictions* improves healthcare provider's ability to predict clinical outcomes

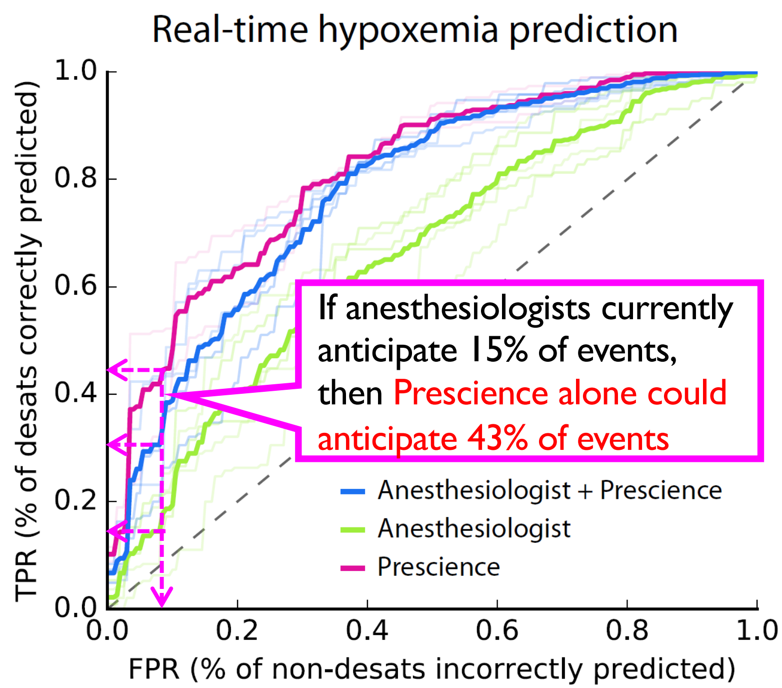


- Our *Prescience* method predicts hypoxemia in the next 5 minutes, provides explanations in real time

Our approach, SHAP

For a particular prediction

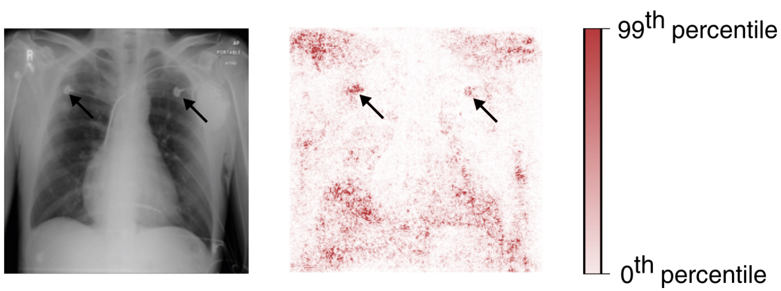


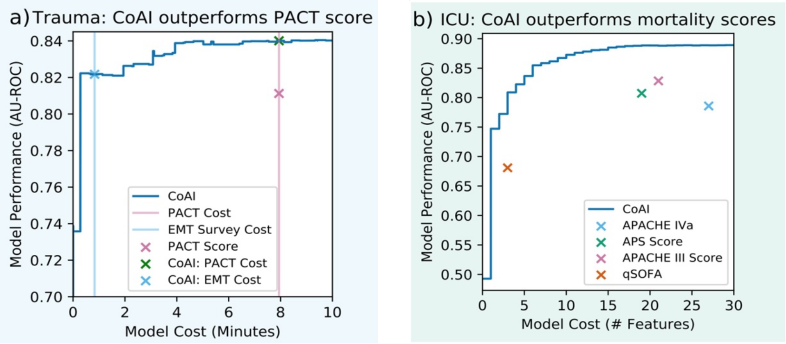


Scott M. Lundberg, […], and Su-In Lee. **Explainable machine-learning predictions for the prevention of hypoxaemia during surgery.** ***Nature Biomedical Engineering*** **2**, 749–760 (2018) – **Cover article**

Explainable AI enables model auditing and cost-aware AI

- We revealed that many published AI systems to detect COVID-19 rely on “shortcuts” rather than genuine pathology
- CoAI enables drastic reduction in feature acquisition cost (e.g., time) to help emergency medicine or ICU patients

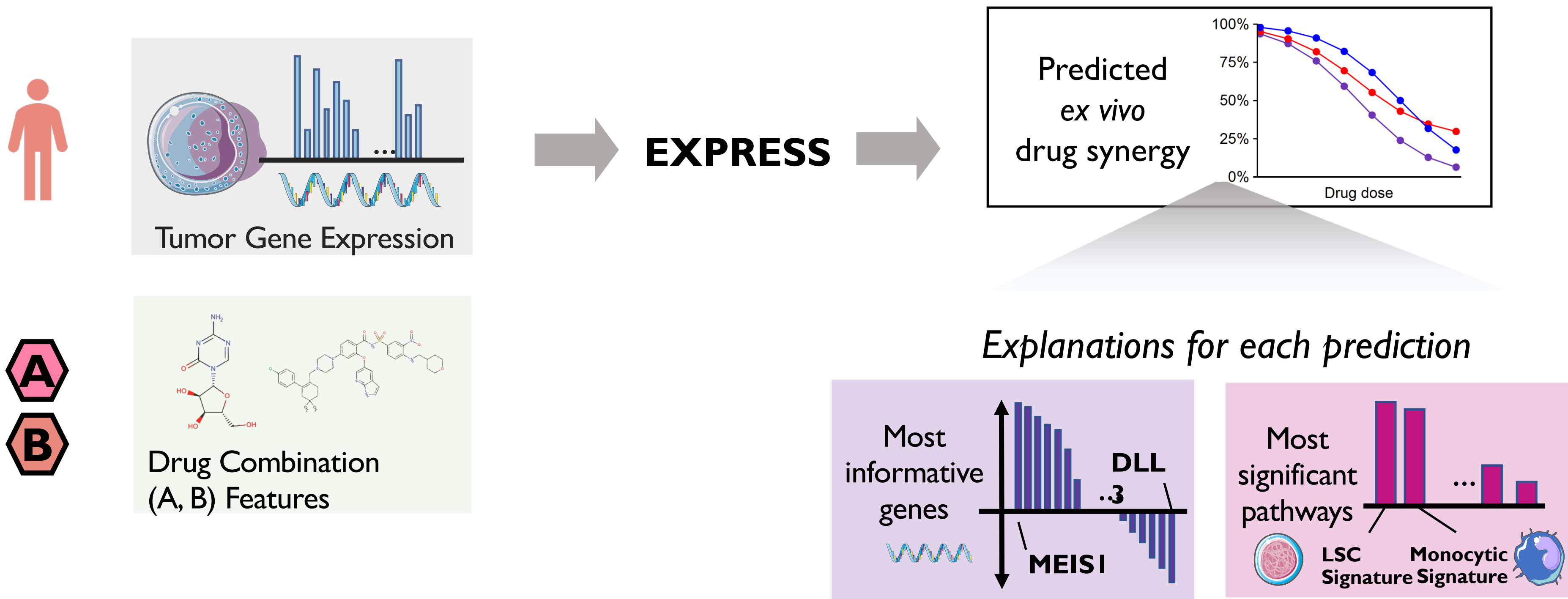




Alex DeGrave*, Joe Janizek*, and Su-In Lee. **AI for radiographic COVID-19 detection selects shortcuts over signal.** ***Nature Machine Intelligence*** *(2021)*

Gabe Erion, Joe Janizek […] Nathan White*, and Su-In Lee*
CoAI: Cost-Aware Artificial Intelligence for Health Care.
In Press ***Nature Biomedical Engineering***

Explainable prediction of drug synergy in AML (EXPRESS)



Joseph Janizek, […], Kamila Naxeriva*, and Su-n Lee*. **Uncovering expression signatures of synergistic drug response using an ensemble of explainable AI models.** In Revision *Nature Biomedical Engineering*

Today

- Section 1
  - Motivation & aims
  - Course logistics
  - Examples in the healthcare space
  - **10 min break: office hours poll (see your email)**
- Section 2
  - Discussion: “Statistical Modeling: The Two Cultures”
- Section 3
  - Example scenario and ML review