# Lecture 1: What is Statistical Learning?

Monday, Sep 29

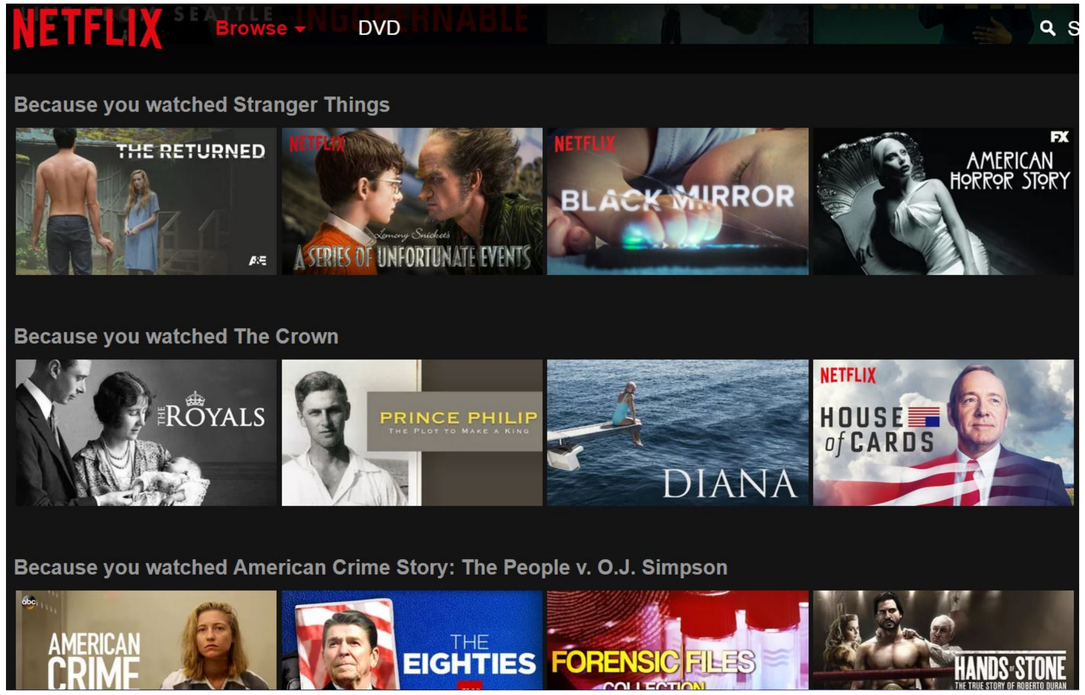
# Agenda

- The statistical learning framework
- Key concepts and terminology:
  - Prediction vs. Inference
  - Parametric vs. Non-parametric models
  - Flexibility vs. Interpretability
  - Supervised vs. Unsupervised learning
  - Regression vs. Classification

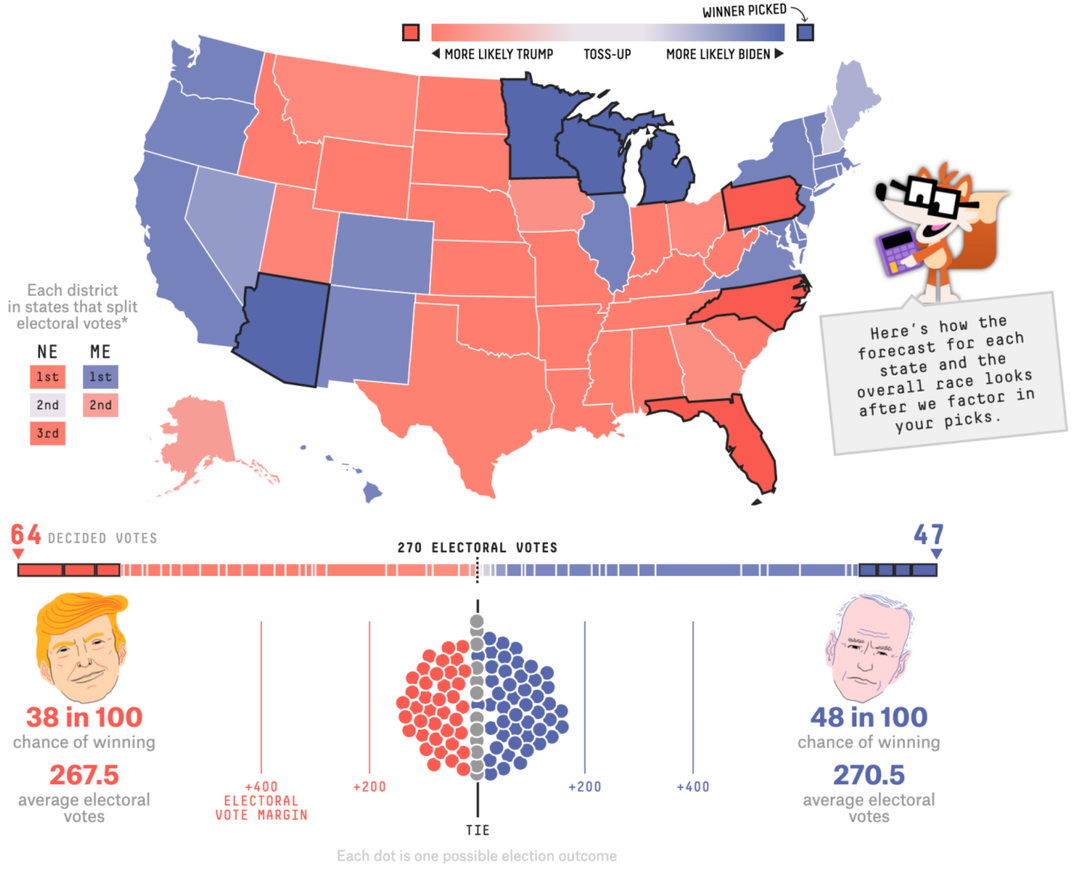
How can we use models and data to
understand and predict the world around us?

# Why we model

**Simplification**: models can draw conclusions from large amounts of data

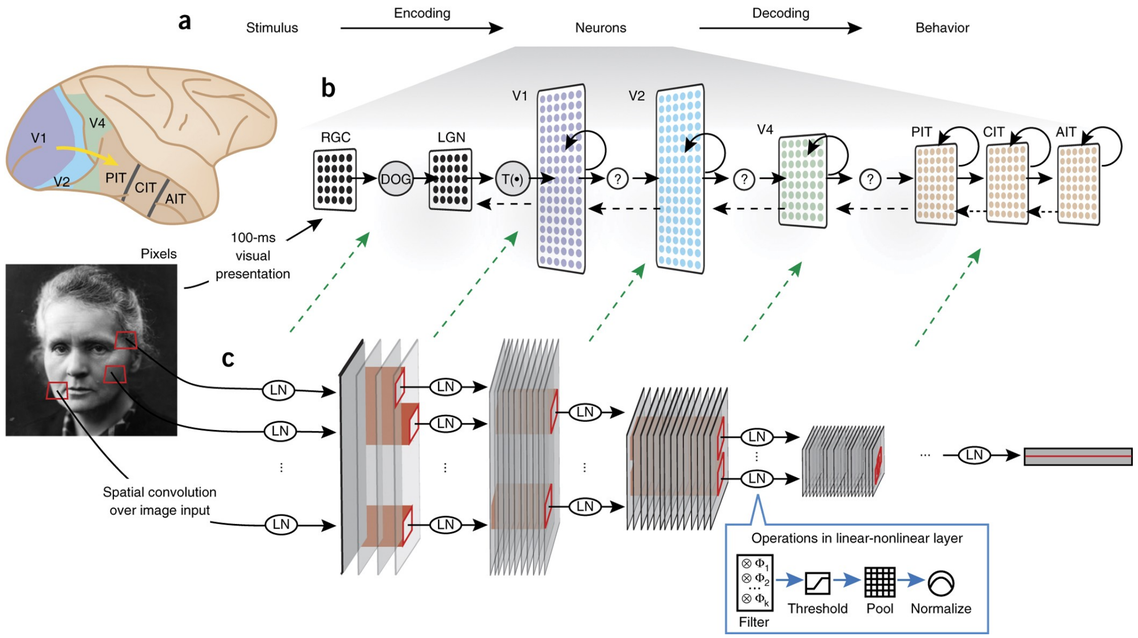



**Prediction**: models allow us to predict what might happen in the future

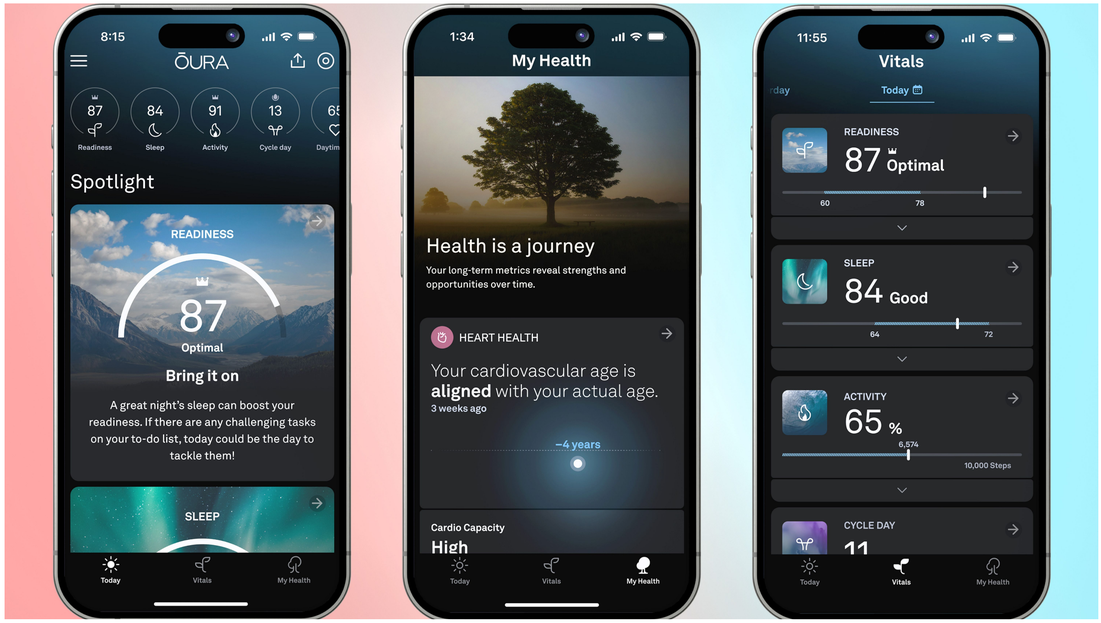


# Why we model

**Understanding complex systems:** models allow us to make sense of complex phenomena




**Inform our decision making:** models can help us make better choices



# Data set: organized information

- A **data set** can be organized as an *n x p* table of values
- *n* = the number of observations, or sample size
- *p* = the number of features
- Data values can be quantitative (continuous) or qualitative (discrete)

**College dataset**

*p* = 4 features

$X_1$ $X_2$ $X_3$ $X_4$

*n* = 12 data points

| | Private | Apps | Accept | Enroll |
|---|---|---|---|---|
| Abilene Christian University | Yes | 1660 | 1232 | 721 |
| Adelphi University | Yes | 2186 | 1924 | 512 |
| Adrian College | Yes | 1428 | 1097 | 336 |
| Agnes Scott College | Yes | 417 | 349 | 137 |
| Alaska Pacific University | Yes | 193 | 146 | 55 |
| Albertson College | Yes | 587 | 479 | 158 |
| Albertus Magnus College | Yes | 353 | 340 | 103 |
| Albion College | Yes | 1899 | 1720 | 489 |
| Albright College | Yes | 1038 | 839 | 227 |
| Alderson-Broaddus College | Yes | 582 | 498 | 172 |
| Alfred University | Yes | 1732 | 1425 | 472 |
| Allegheny College | Yes | 2652 | 1900 | 484 |

qualitative quantitative

# Statistical learning framework (supervised learning)

$$Y = f(X) + \epsilon$$

# Statistical learning framework (supervised learning)

some unknown function that
we are trying to estimate

$$Y = f(X) + \epsilon$$

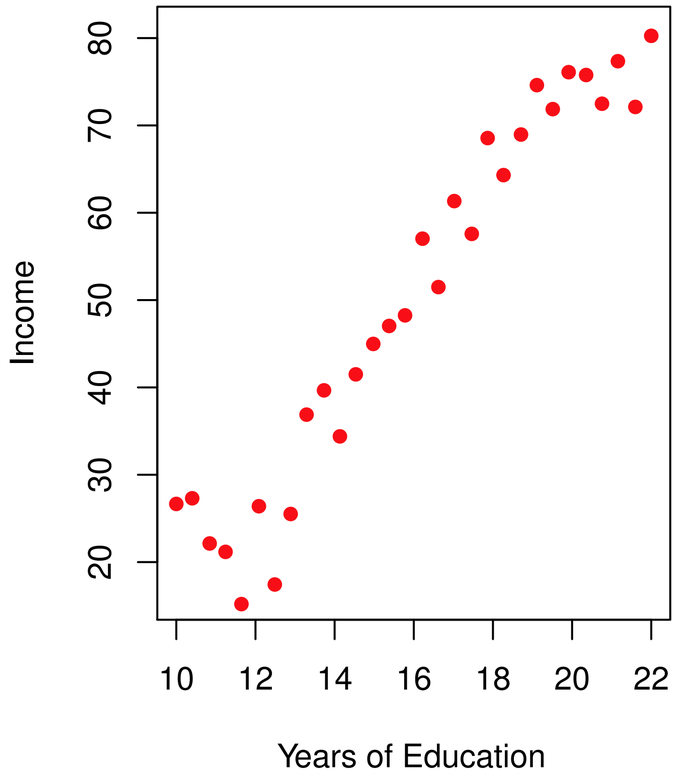
output,
response,
dependent variable

predictors,
features,
independent variables

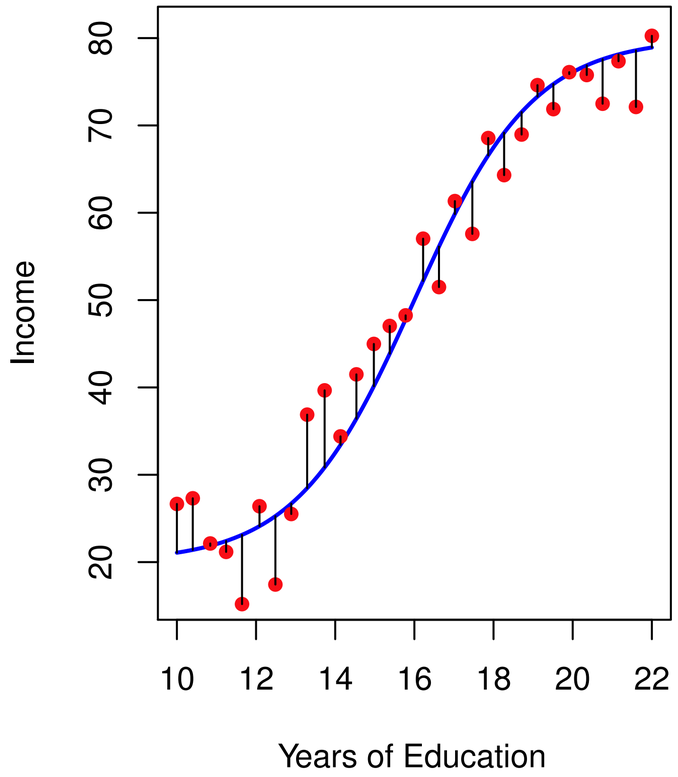
random or irreducible error,
has mean zero

# Statistical learning framework (supervised learning)

$$Y = f(X) + \epsilon$$




$$\hat{Y} = \hat{f}(X)$$



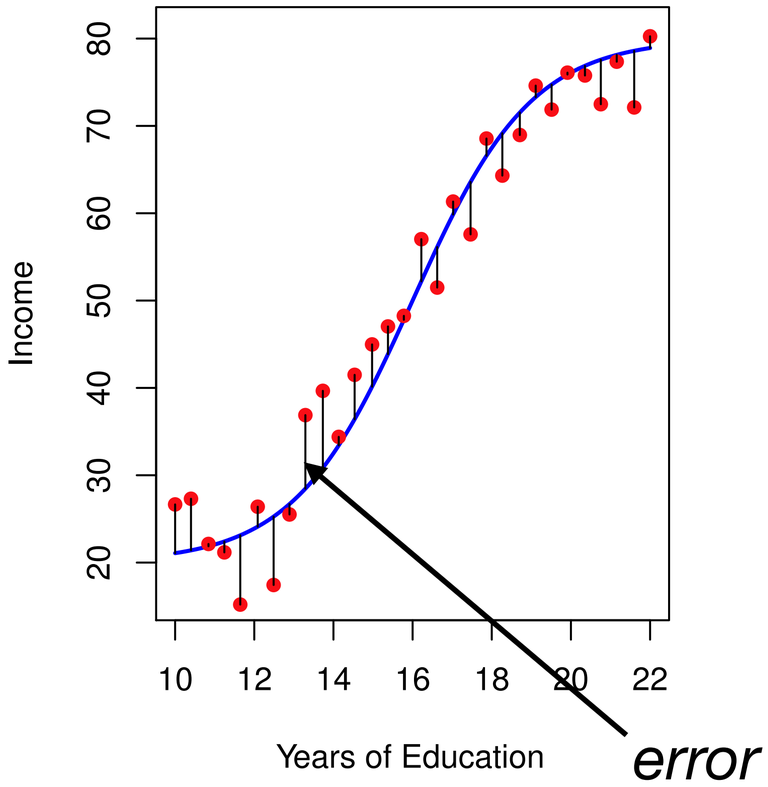

$f$ : true relationship between income and years of education

$\hat{f}$ : our best guess of $f$ estimated from data

statistical learning = estimating what $f$ is!

# Reducible vs. Irreducible Error

$$\hat{Y} = \hat{f}(X)$$




## Reducible Error: $[f(X) - \hat{f}(X)]^2$

- Error due to model limitations; can be reduced with a better model $f$, or adding features $X$
- **Examples**: using a linear model when the true relationship is nonlinear; leaving out an important predictor.

## Irreducible Error: $Var(\epsilon)$

- Random noise from unmeasured or unpredictable factors; cannot be eliminated.
- **Examples:** trial-to-trial variability in reaction times; random distractions affecting exam performance.

# Two goals: Prediction vs. Inference

**Prediction**: estimating what $f$ is, in order to make the *best prediction* about $Y$

$$\hat{Y} = \hat{f}(X)$$

- Only care about making accurate predictions about $Y$
- Not concerned with the exact form of $f$
- **Example:** predicting whether a patient will have an allergic reaction to a drug ($Y$)
- $\hat{f}$ can be treated as a **black box**: "*don't care how you got the answer…just that it's the right answer*"

# Two goals: Prediction vs. Inference

**Prediction**: estimating what $f$ is, in order to make the *best prediction* about $Y$

$$\hat{Y} = \hat{f}(X)$$

- Only care about making accurate predictions about $Y$
- Not concerned with the exact form of $f$
- **Example:** predicting whether a patient will have an allergic reaction to a drug ($Y$)
- $\hat{f}$ can be treated as a **black box**: "*don't care how you got the answer…just that it's the right answer*"

**Inference**: estimating what $f$ is, in order to understand the *relationship* between $X$ and $Y$

$$\hat{Y} = \hat{f}(X)$$

- Which $X$'s are associated with $Y$?
- How much does $Y$ change if I change $X_1, \ldots, X_n$?
- Not concerned with the exact form of $f$
- **Example:** which blood tests ($X$) are most predictive of whether the patient will have an allergic reaction ($Y$)

# Estimating $f$: Parametric vs. Non-parametric models

**Parametric models**: assume a functional form (or "shape") of $f$

linear models $Y = mx + b$

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots$$

- After model is selected, use data to train or fit the model
- **Advantage:** estimating $f$ is simplified to estimating parameters (e.g., $\beta_0, \beta_1, \beta_2, \dots$), model is interpretable
- **Disadvantage:** chosen form may not capture data well—it is *not* flexible

**Non-parametric models**: no assumption about the form of $f$

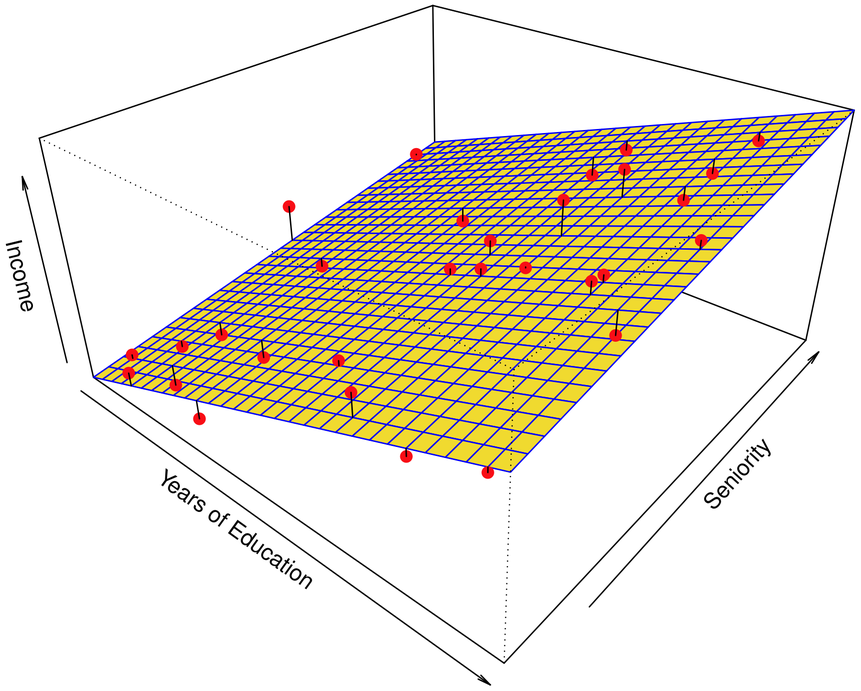
neural networks
"black box"

$$f(X) = ?$$

- **Advantage:** Can fit many "shapes" of $f$—model is flexible
- **Disadvantage:** Requires large $n$ to get accurate estimate of $f$, it is *not* interpretable

# Estimating $f$: Parametric vs. Non-parametric models

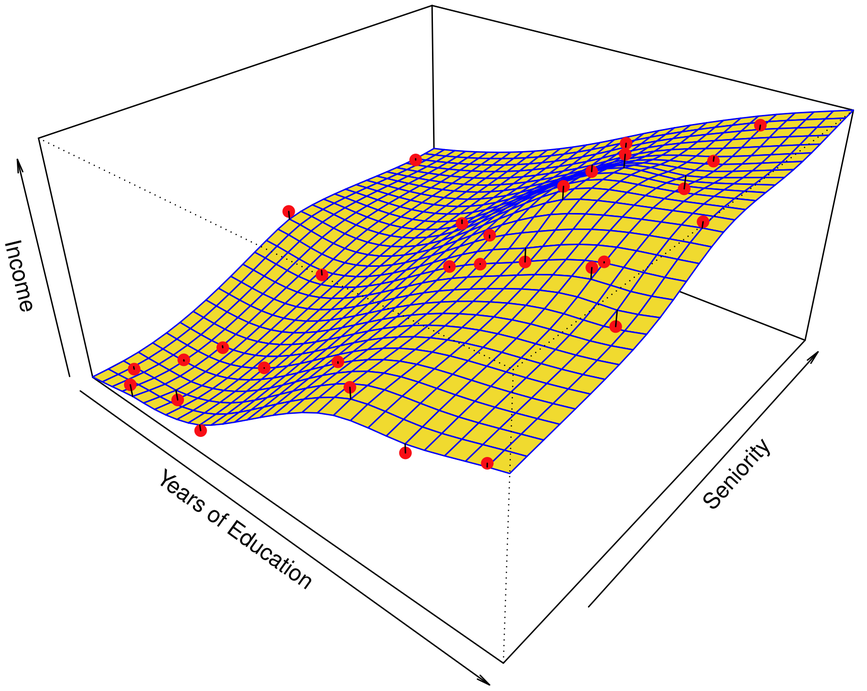
## parametric: linear model




$$\text{income} = \beta_0 + \beta_1 \times \text{education} + \beta_2 \times \text{seniority}$$

- rigid shape that may be underfitting

## non-parametric: thin plate spline




- fits the data “as close as possible subject to the fit being smooth”
- may run the risk of overfitting

# Flexibility vs. Interpretability


- The more flexible a model, the less interpretable it generally is
- We will encounter a range of models along these axes
- Why would you ever want to use a less flexible model?
  - Depends on what your question or goal is

# Flexibility vs. Interpretability

## Flexible / less interpretable models

perform well when:

- There is a large amount of training data (large *n*) compared to the number of predictors (*p*)
- There is relatively low noise, i.e. Var($\epsilon$) is small
- The true relationship $Y = f(X) + \epsilon$ is complex or nonlinear

## Less flexible / interpretable models

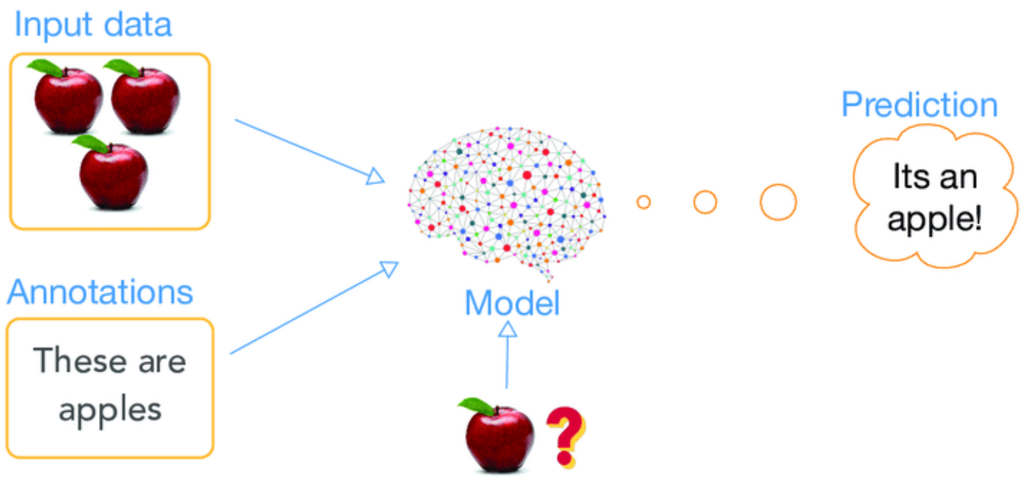
perform well when:

- There is a limited amount of training data (low *n*) compared to *p*
- The data are very noisy, i.e. Var($\epsilon$) is large
- We care more about inference or interpretability rather than predictive accuracy

# Supervised vs. Unsupervised Learning

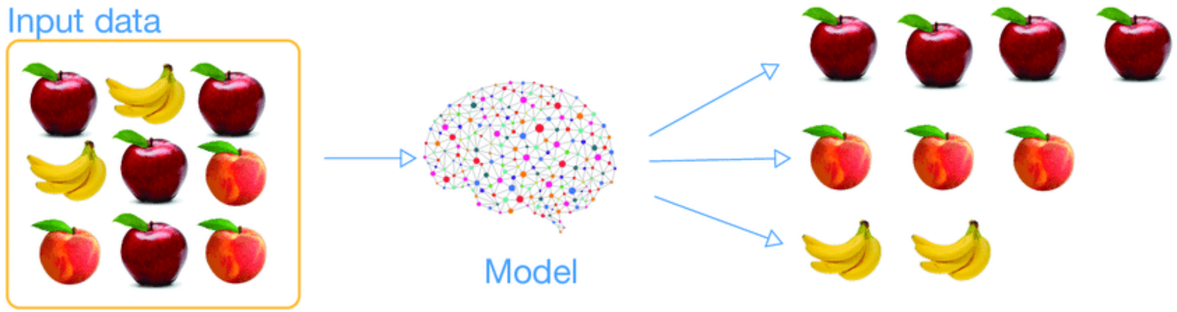
**Supervised**: there is a correct answer or label to be predicted

- For each $x_1, x_2, x_3, \ldots$ there is a corresponding $y_1, y_2, y_3, \ldots$
- **Goal:** correctly predict $y$ for every $x$; inference problems




**Unsupervised**: no labels or correct answers

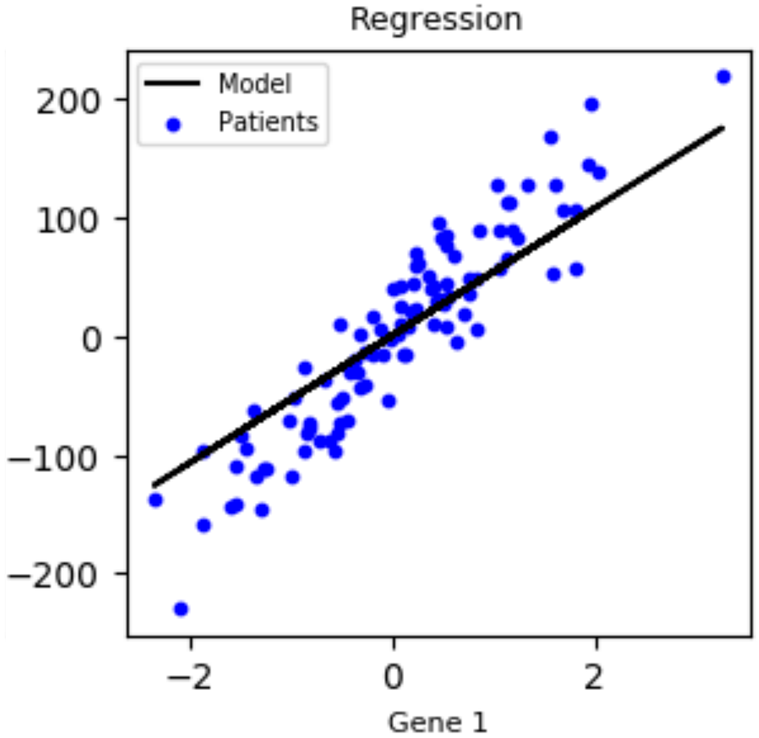
- Many features $x_1, x_2, x_3, \ldots$
- **Goal:** find patterns and structure in unlabeled data (e.g., by grouping or clustering data by similarity of features)



# Regression vs. Classification (supervised learning)

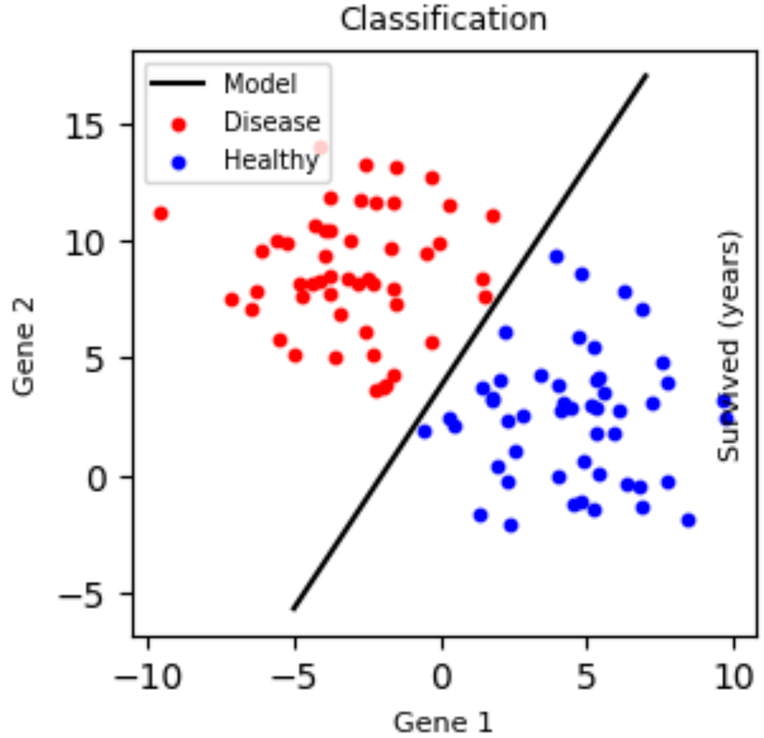
**Regression**: prediction for quantitative (numerical, continuous) responses

- **Example $Y$s:** predicting age, height, income, neural activity, heart rate, etc.




**Classification**: prediction for qualitative (categorical) responses

- **Example $Y$s:** predict pass vs. fail, spam vs. not spam, tumor category, image labels, etc.



# Homework 1

## COGS 109: Homework 1

Due Friday, Oct 3 @ 11:59pm PT

*Total estimated time to complete: 4-5 hours*

Instructions (PLEASE READ)

1. Please copy and paste this entire assignment in your own document and write your answers below each question (grab the template from Canvas/course website). Leave unanswered questions blank (don't delete them).
2. For full competition credit (3pts), please choose one of these two options:
    a. Complete Part A (1.5pt) + Part B (1pt) + Part D (0.5pt) = 3 points
    b. Complete 8 questions from Part A (1pt) + Part B (1pt) + Part C (0.5pt) + Part D (0.5pt) = 3 points
3. Overachiever option! If you choose to complete all parts in their entirety, you can earn the full 3.5 points.

💡 **Tips: (1) Start early! (2) Follow the lectures / textbook**

# Office hours

## Instructor

**Dr. Lucy Lai**

CSB 244 or Zoom

- Wed, 3-4:30pm (book / walk-in)
- Thurs, 4:30-5:30pm (book only)
- Fri, 4-6p in WLH 2207 (occasionally)


## TA

**Jiesen Zhang**

During section

## PLAs

**Johny Nguyen**

During W section & W, 1-2pm

**Parinita Saha**

During section & TBD

# Upcoming + Reminders

**Reminders:**

- Sections begin this week!
- Exam Q's → OH / group work

- Wed, 2-2:50p @ WLH 2113
- Fri, 4-4:50p @ WLH 2207
- Fri, 5-5:50p @ WLH 2207

**Assignments:**

- Pre-course survey (DUE: Fri, Oct 3) — make sure to **submit** on Canvas
- Syllabus quiz (DUE: Fri, Oct 3)
- HW 1 (DUE: Fri, Oct 3)

**Wednesday's topic:** ***Assessing Model Accuracy***

- **Read:** ISLP Ch. 2.2

# COGS 109 *Weekly* Rhythm

Here's how you can expect to spend your time in this course:

Fall 2025
**Instructor:** Prof. Lucy Lai

| Component | Monday | Tuesday | Wednesday | Thursday | Friday |
|---|---|---|---|---|---|
| **Lecture** (9-9:50 AM @ CNTR 113) | Lecture | | Lecture | | Lecture |
| **Discussion section** (Various times @ WLH) | | | S1 @ WLH 2113 (2-2:50PM) | | S2 @ WLH 2207 (4-4:50PM) S3 @ WLH 2207 (5-5:50PM) |
| **Assignments** | **Quiz DUE @ 11:59PM** | | | Quiz released | Work on quiz |
| | Homework released | Work on homework | | | **Homework DUE @ 11:59PM** |
| **In-class exams** | **Exam dates** Oct 13 Oct 27 Nov 10 Dec 1 | | | | |
| **Office hours** | | | **Prof. Lai** 3-4:30PM @ CSB 244 Bookable | **Prof. Lai** 4:30-5:40 @ Zoom Bookable | **Prof. Lai** 4-6PM @ WLH 2207 (occasionally) |
| **Other** | Readings, study for exams, work on group project | | | | |