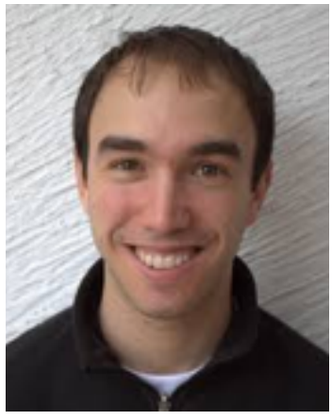
6.867 Machine Learning Fall 2017

Lecture 1. Introduction

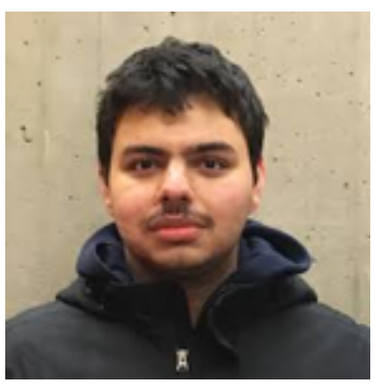
http://stellar.mit.edu/S/course/6/fa17/6.867/

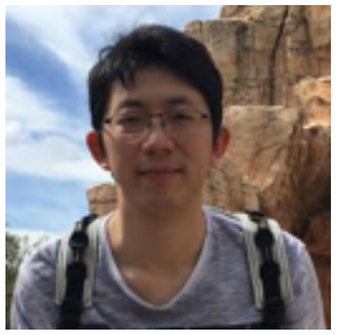
Staff

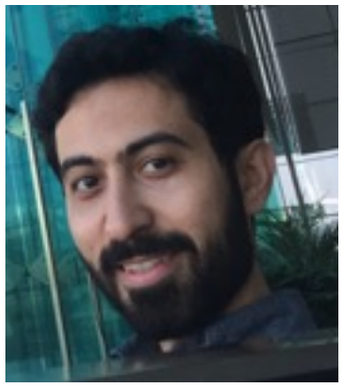
Instructors

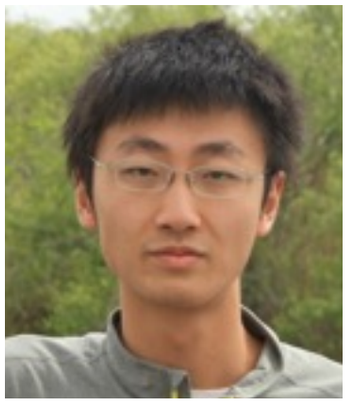


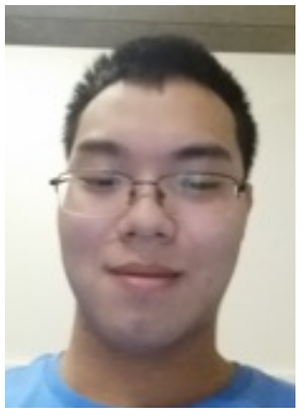
Teaching Assistants











Logistics

http://stellar.mit.edu/S/course/6/fa17/6.867/ > Materials > Information Sheet

- Structure
  - Lectures: T-Th 2.30-4 (here)
  - Recitations: Fridays (see assignment, if issues self-assign by 9/14)
  - Office Hours: M-W 4-6pm
  - Piazza: for all help. help each other. TAs will respond once / 24 hrs
  - Exercises: weekly set of questions (not graded, recitations around them)

- Grading
  - 3 HWs: 30% (HW0 is posted, for self-assessment)
  - Project: 20% (various milestones, first due on Sept 26 - find team of 3)
  - Two quizzes: 50% (Oct 19 and Nov 30: 7pm-9pm)

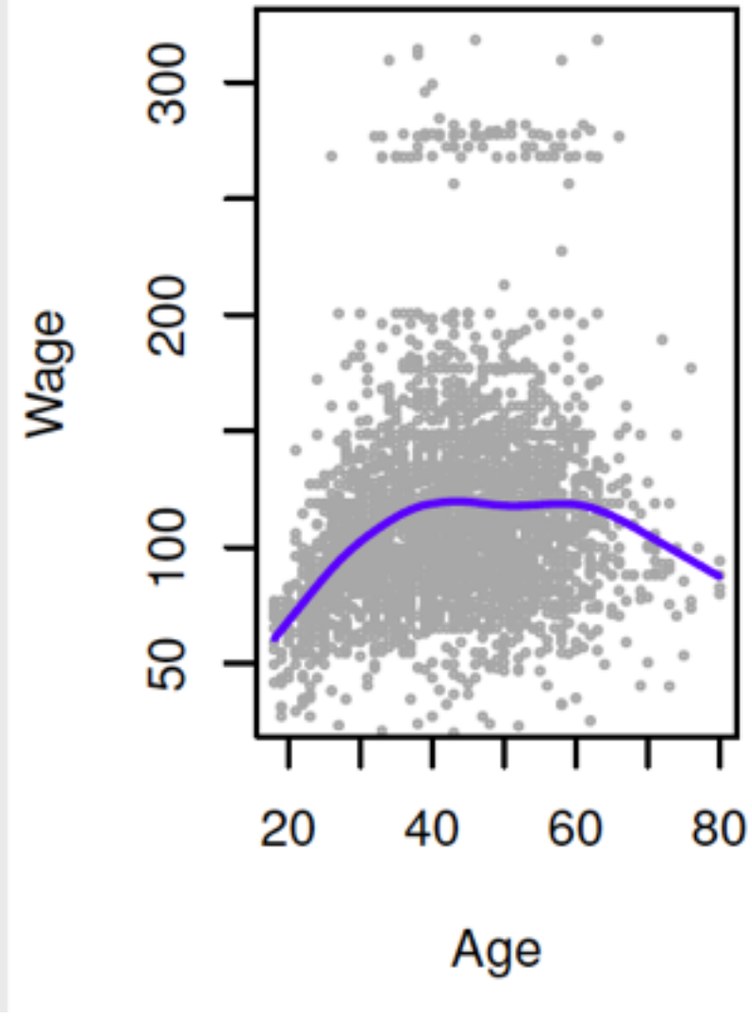
What is Machine Learning:
Models, Methods & Algorithms

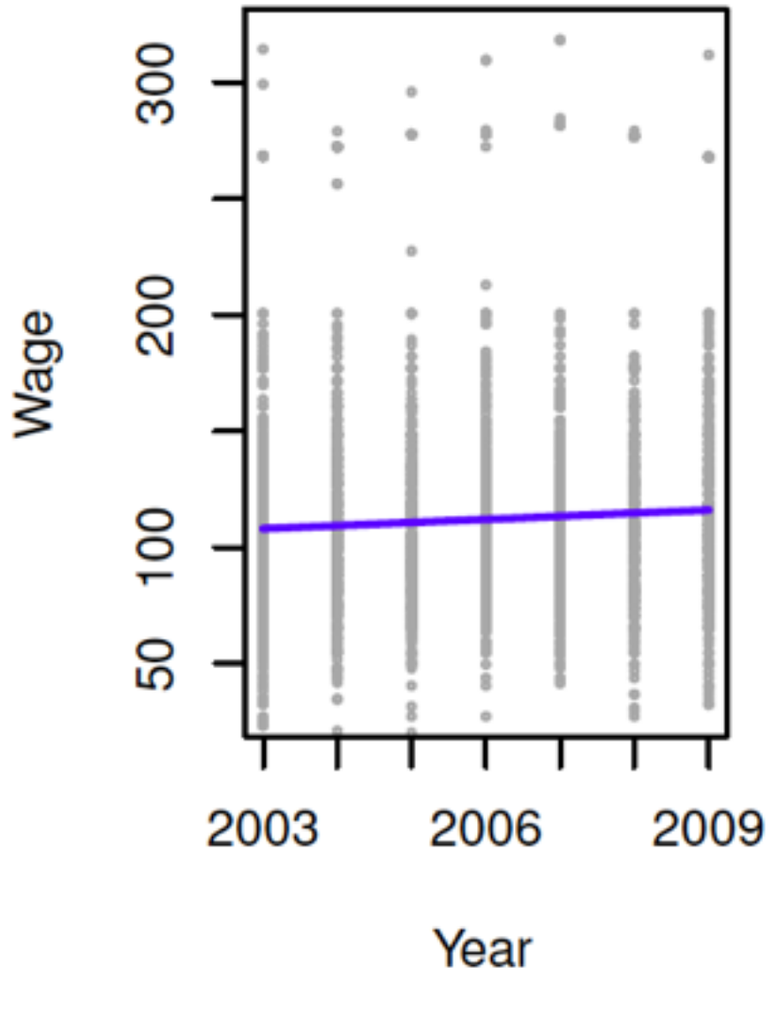
- Machine learning, Statistical learning, Pattern recognition, …
    - all about understanding data
    - they refer to a collection of tools, methods and algorithms to do so
- Classically, focus has been on "prediction" or "decisions" as the end goal
- But, "model learning" is an intermediate step
    - Many scientific disciplines, "model learning" is *the* step
- Prediction vs Interpretation
    - e.g. PV = nRT (ideal gas law) vs Using rest to infer P (pressure)
- Ability to compute determines what can be / can not be machine learnt

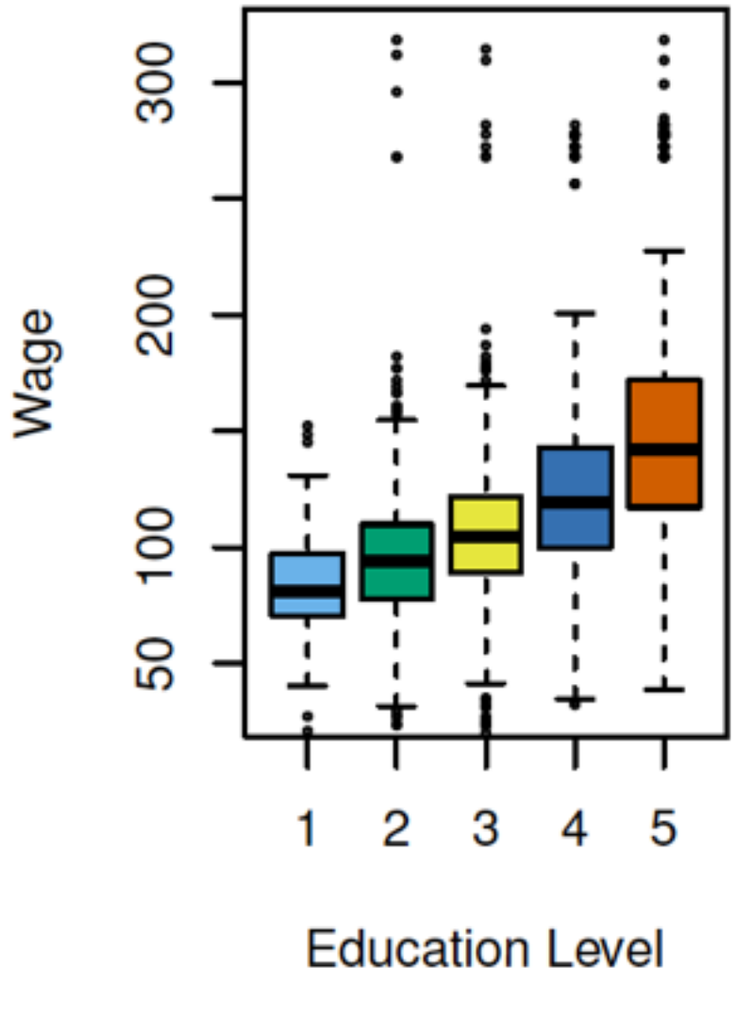
Problems / Methods: A Bird's Eye View (L2-14)

- Supervised
  - Predict *target* using *input* / *features*
  - Learn a model to do so
    - *target* = f (*features*)
      - Regression
        - when *target* takes any real value (e.g. temperature)
      - Classification
        - when *target* takes one of few different values (e.g. hot or cold)

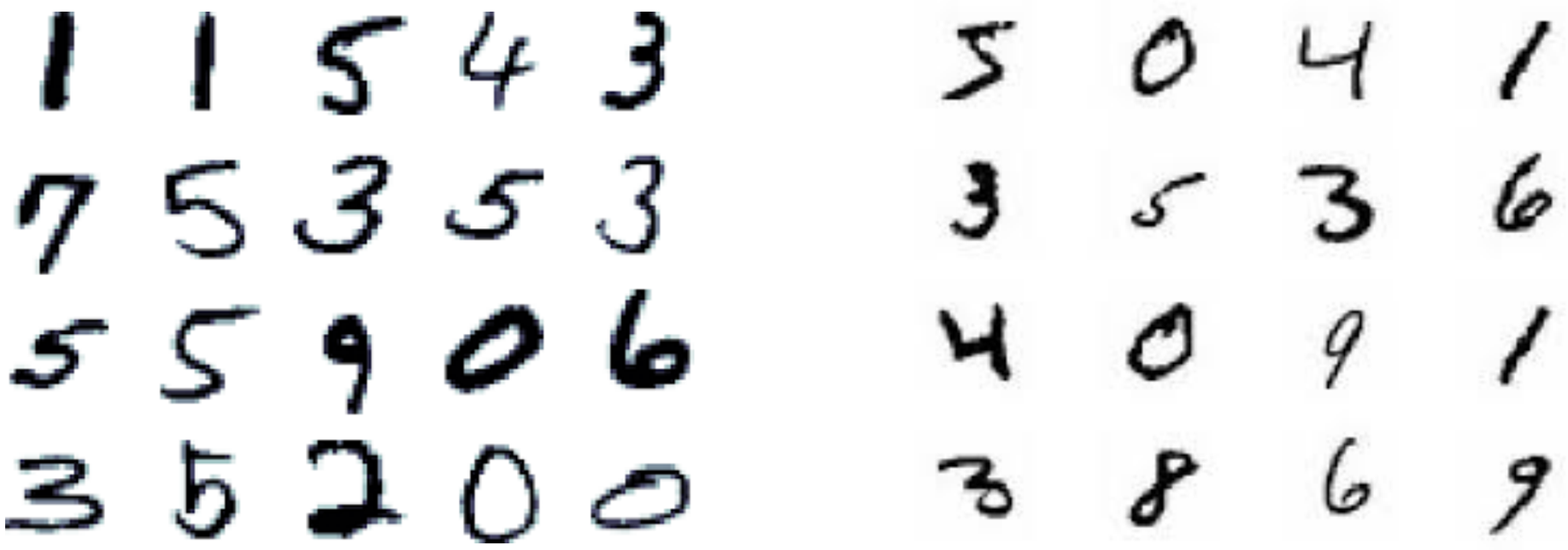
Regression: example







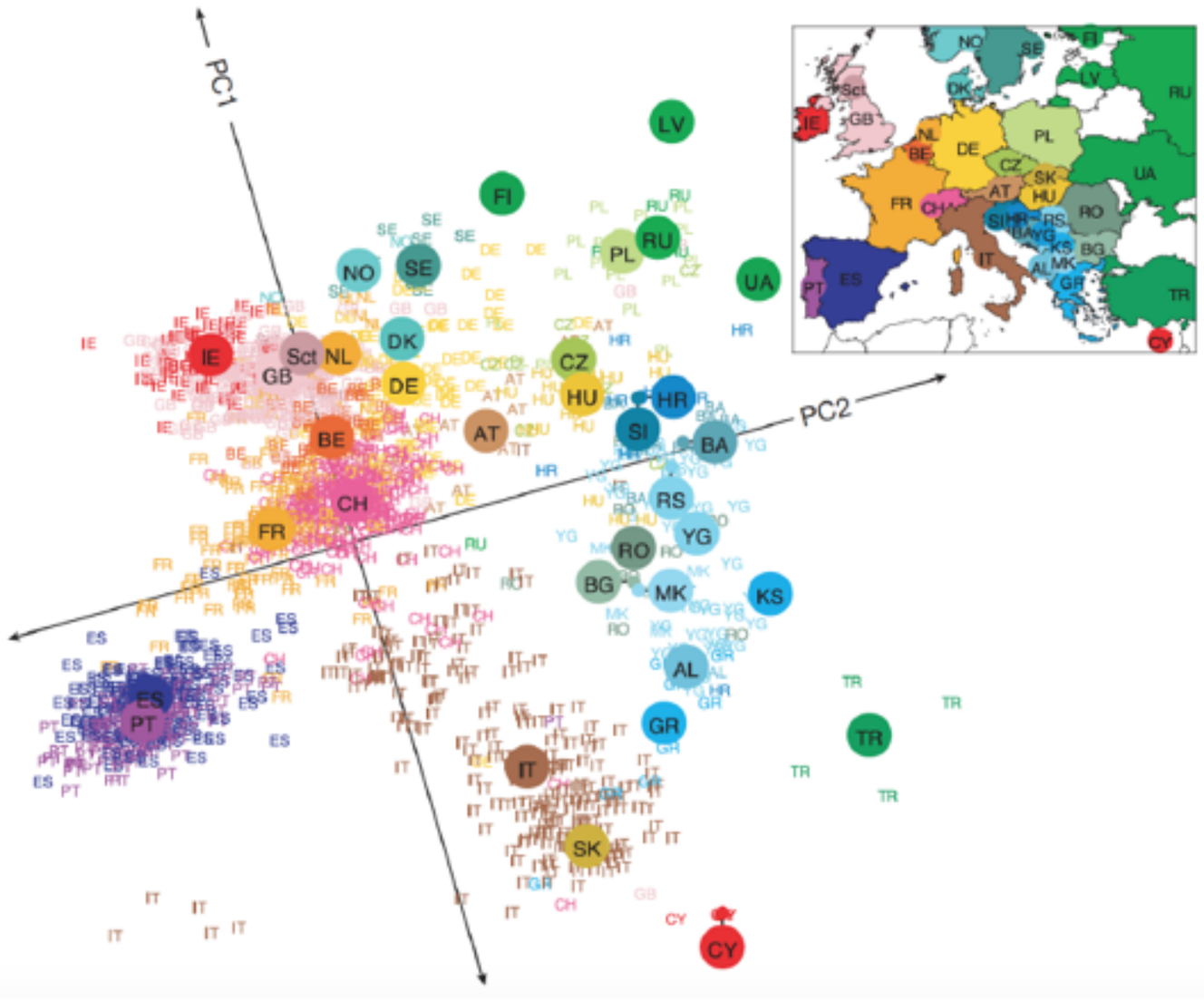
Classification: example



Problems / Methods: A Bird's Eye View (L15-19)

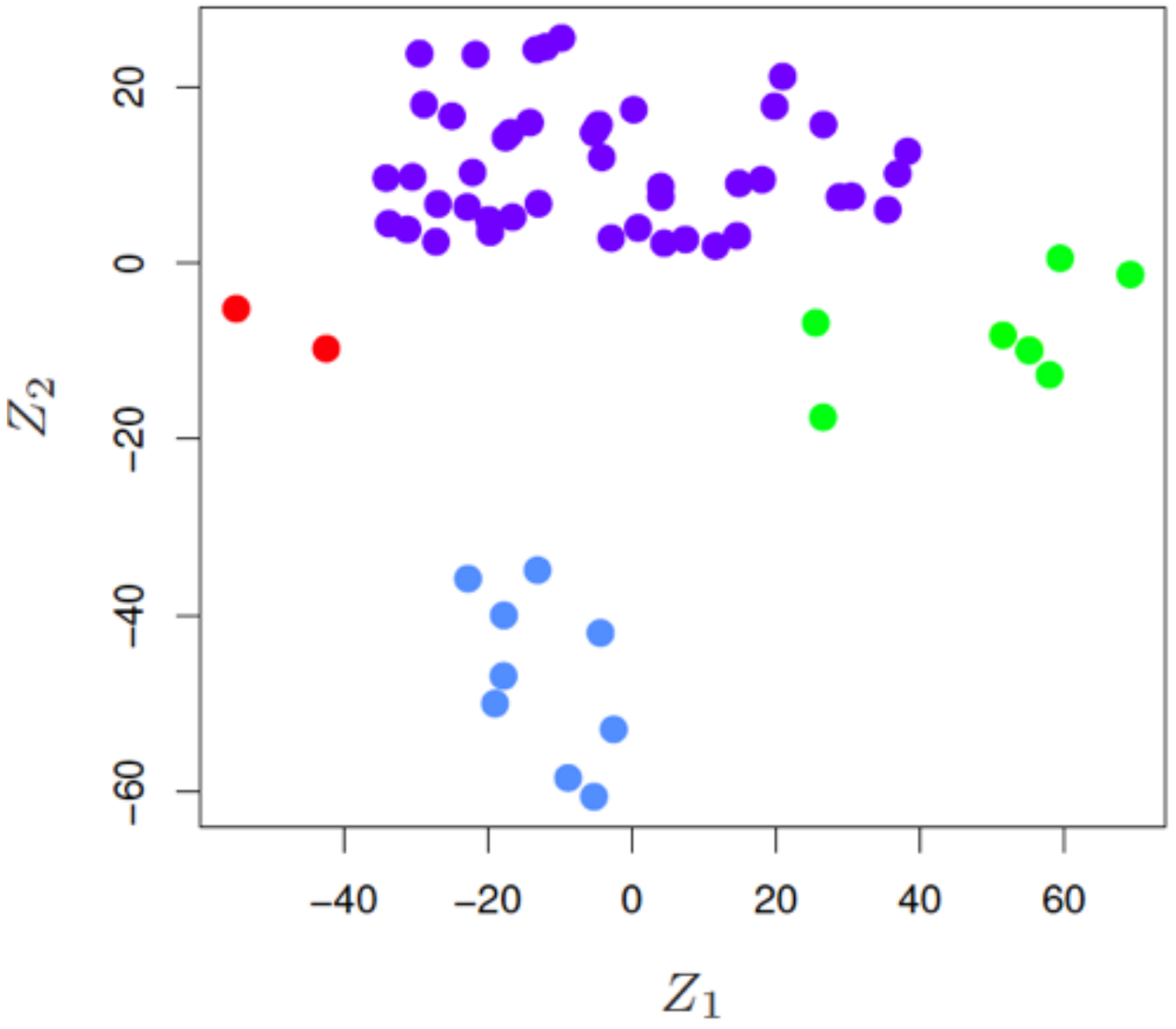
- Un-Supervised
  - No *target* only *input / features* given
  - Learning distribution
  - Clustering
  - Dimensionality reduction
  - Feature extraction
    - from unstructured data such as text, audio or image
    - or, for complexity reduction

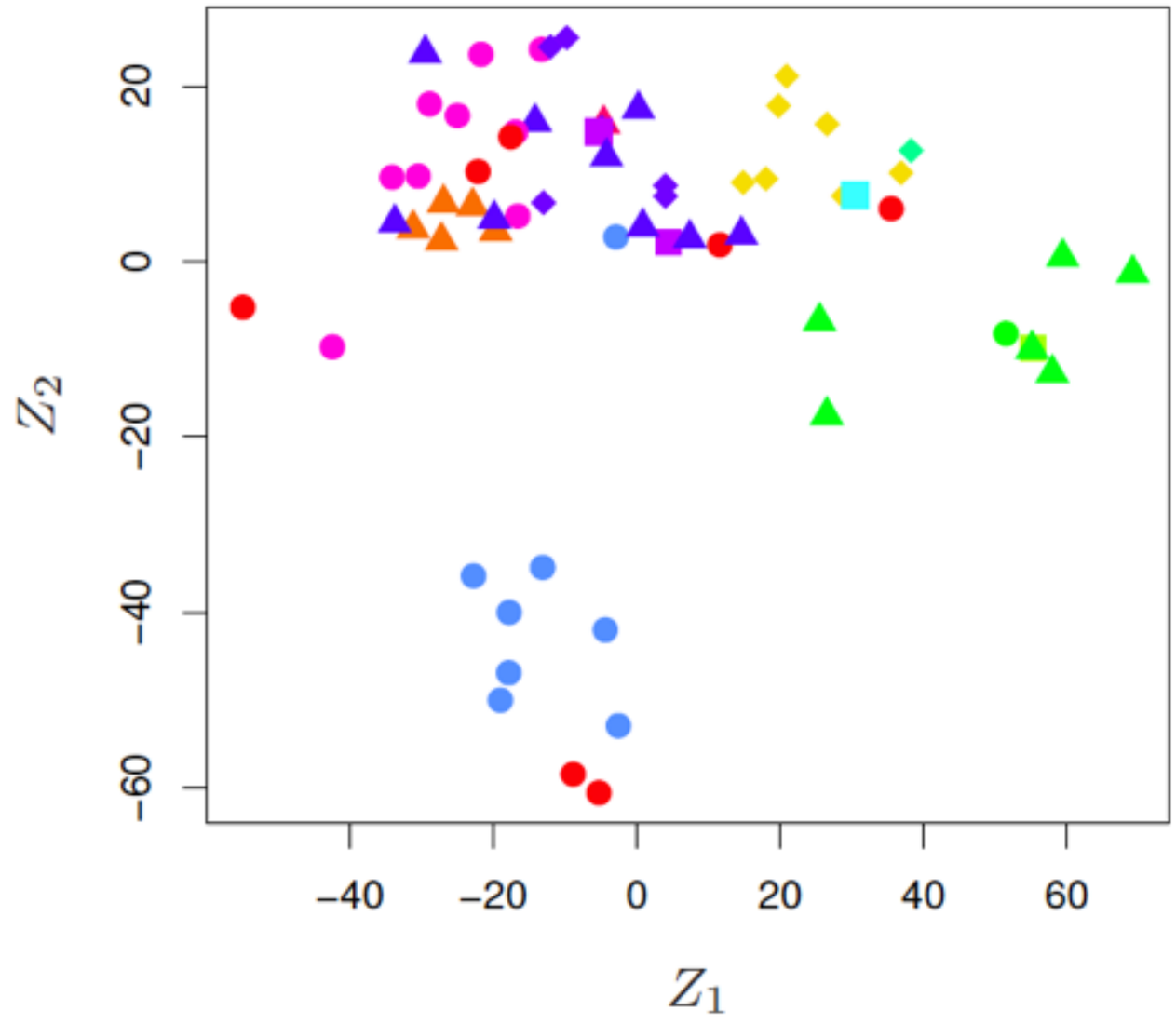
Dimensionality Reduction: Example



Genes Mirror Geography Within Europe, Novembre et al (2008)

Clustering: Example

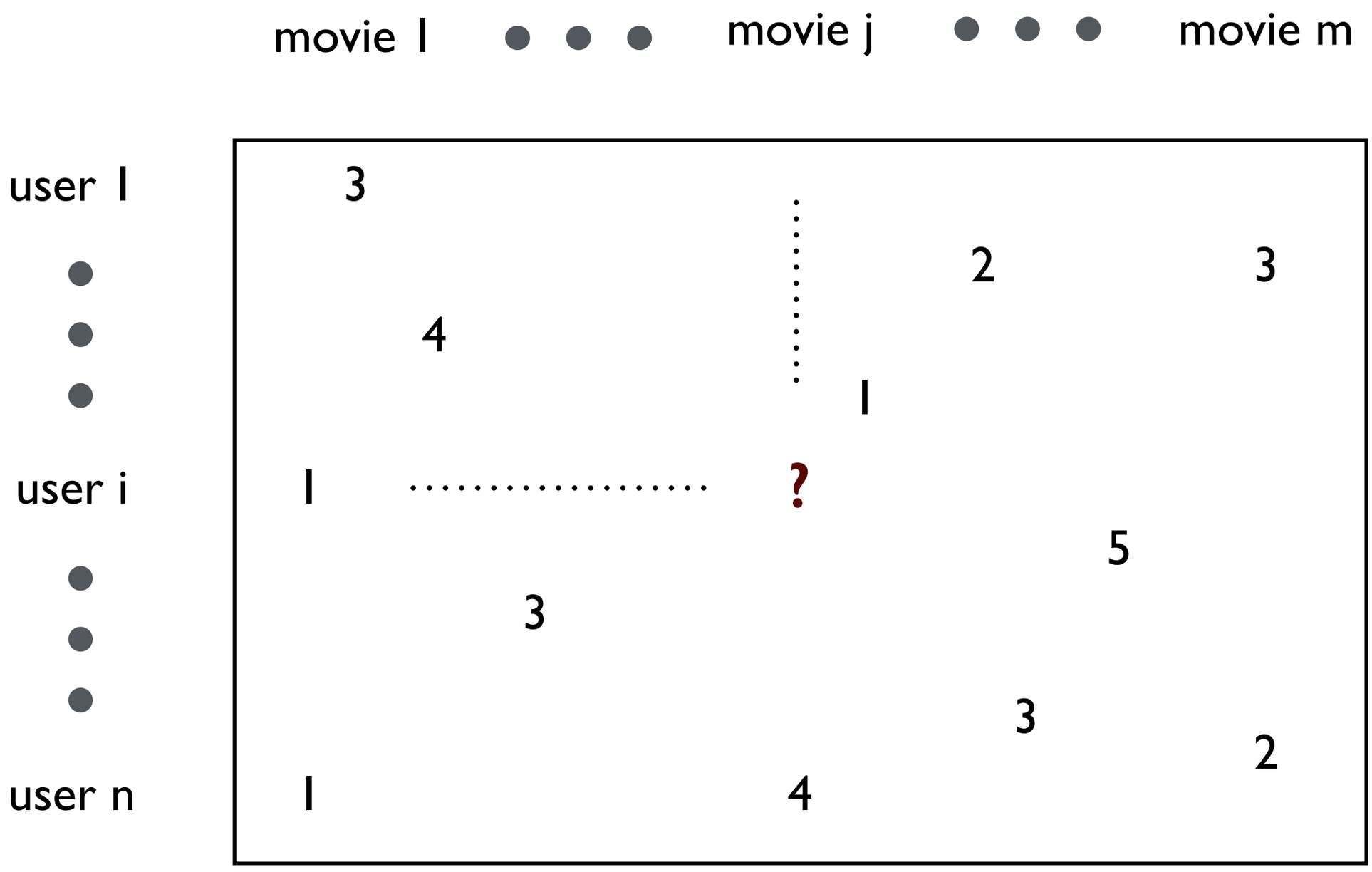




Problems / Methods: A Bird's Eye View (L20-22)

- Matrix-estimation
    - Matrix structured data
    - Goal is to predict missing information / de-noise available information
    - Information / correlation matrices
        - Dimensionality reduction, Structure learning
- Network structure
    - Community detection
- Two-party interaction data
    - recommendation systems

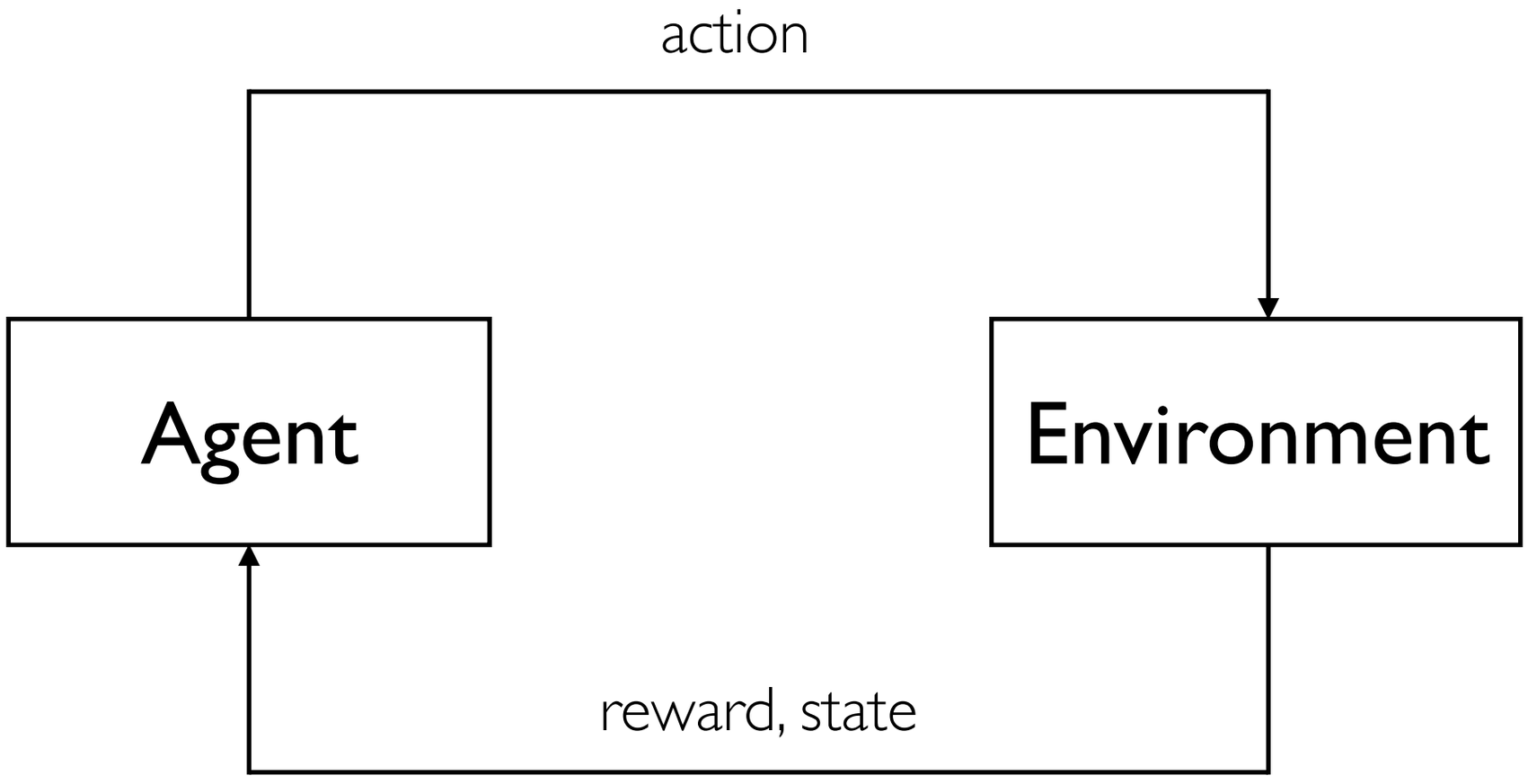
Matrix-Estimation: example



Rating Matrix A

Problems / Methods: A Bird's Eye View (L23-24)

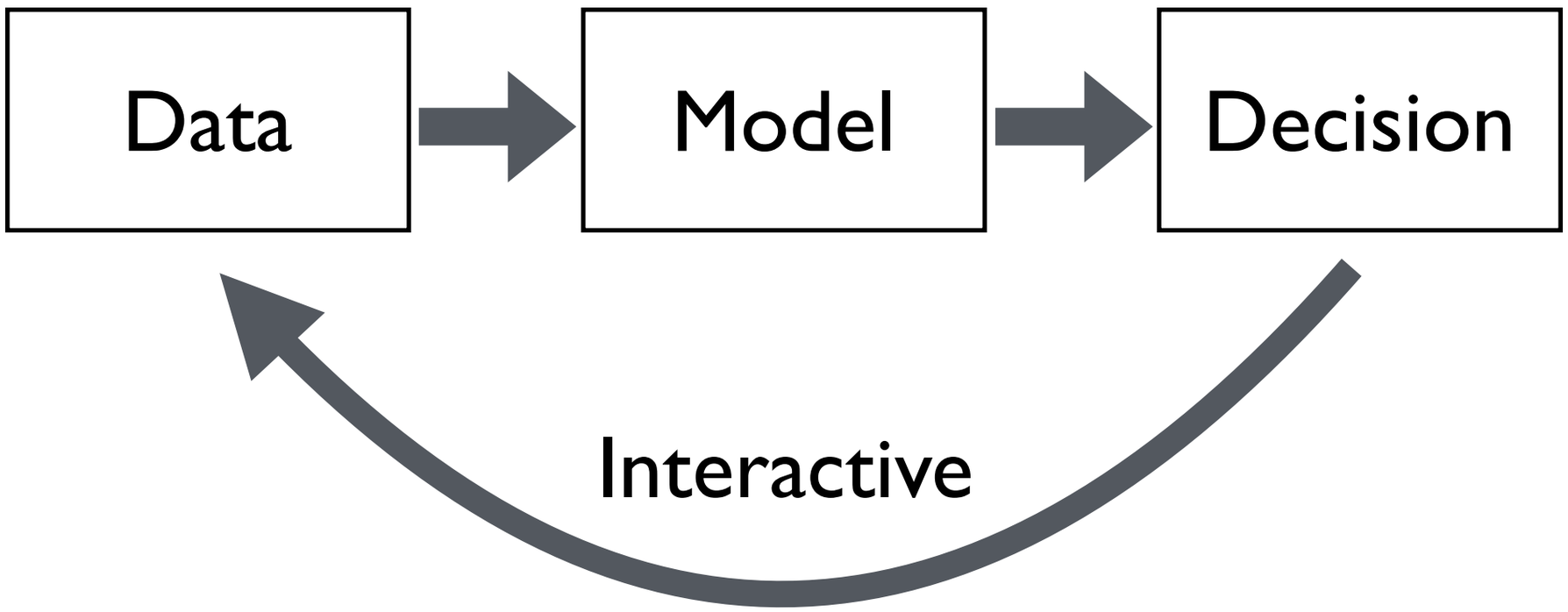
- Re-inforcement learning or Sequential learning



- Examples
  - Playing Go or BackGammon or Chess
  - Self-driving car or self-driving X
  - Control Systems

From The Lens of Decision Making

supervised learning



reinforcement learning

explore vs exploit

And Some

- Active Learning
    - *actively* obtain data as each data point is *expensive*
- Transfer Learning
    - *transfer* data collected for one task to other learning task
- Semi-supervised Learning
    - Supervised setting with (additional) unsupervised data

Framework: Probability

- Axioms of Probability:
  - Probability Space: Ω Events: $\mathcal{F}$ Probability Function: $\mathbb{P} : \mathcal{F} \to [0, 1]$
  - Axiom 1.

$$\mathbb{P}(A) \geq 0, \text{ for all } A \in \mathcal{F}$$

  - Axiom 2.

$$\mathbb{P}(\Omega) = 1$$

  - Axiom 3.

$$\mathbb{P}(\cup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} \mathbb{P}(E_i), \quad \text{if } E_i \cap E_j = \emptyset, \text{for all } i \neq j.$$

- All reasonable axioms about decision making / belief operations lead to the above axioms, one way or another (Cox 40s, Savage 60s, de Finetti 70s,…)

Model Selection: an Example

- Data
  - Samples from a distribution: $x_1, \ldots, x_N$
- Goal
  - Learn the distribution
- Assumption
  - Data is generated from a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$
- Refined Goal
  - Learn the mean and variance
- Question
  - How to learn (parameters, mean and variance) ?

Maximum Likelihood

- Choose the parameters that maximize

$$\mathbb{P}(\text{data}|\text{parameters})$$

- To choose mean, variance from samples:

$$\mathbb{P}(x_1, \ldots, x_N|\mu, \sigma^2) = \prod_{i=1}^{N} \mathbb{P}(x_i|\mu, \sigma^2)$$

$$= \prod_{i=1}^{N} \exp\Big(-\frac{(x_i-\mu)^2}{2\sigma^2}\Big) \times (2\pi\sigma^2)^{-1/2}$$

Maximum Likelihood

- Maximizing likelihood is same as maximizing logarithm of likelihood
- This leads to

$$\max_{\mu,\sigma^2} g(\mu, \sigma^2)$$

- where

$$g(\mu, \sigma^2) = -\frac{(x_i - \mu)^2}{2\sigma^2} - N \log \sigma - N \log \sqrt{2\pi}$$

- This is an *optimization* problem and its solution is what we desire
    - For such reasons, optimization is an integral part of Machine Learning

Maximum Likelihood

- Solution

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_i x_i$$

$$\sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_i (x_i - \mu_{\mathrm{ML}})^2$$

- Bessel's correction

$$\tilde{\sigma}^2 = \frac{1}{N-1}\sum_i (x_i - \mu_{\mathrm{ML}})^2$$

Bayesian Model Selection

- Choose the parameters that maximize

$$\mathbb{P}(\text{parameters}|\text{data})$$

- where, per Bayes' rule

$$\mathbb{P}(\text{parameters}|\text{data}) \propto \mathbb{P}(\text{data}|\text{parameters}) \times \mathbb{P}(\text{parameters})$$

posterior likelihood prior

- How to select *prior?*

Decision Driven Model Selection

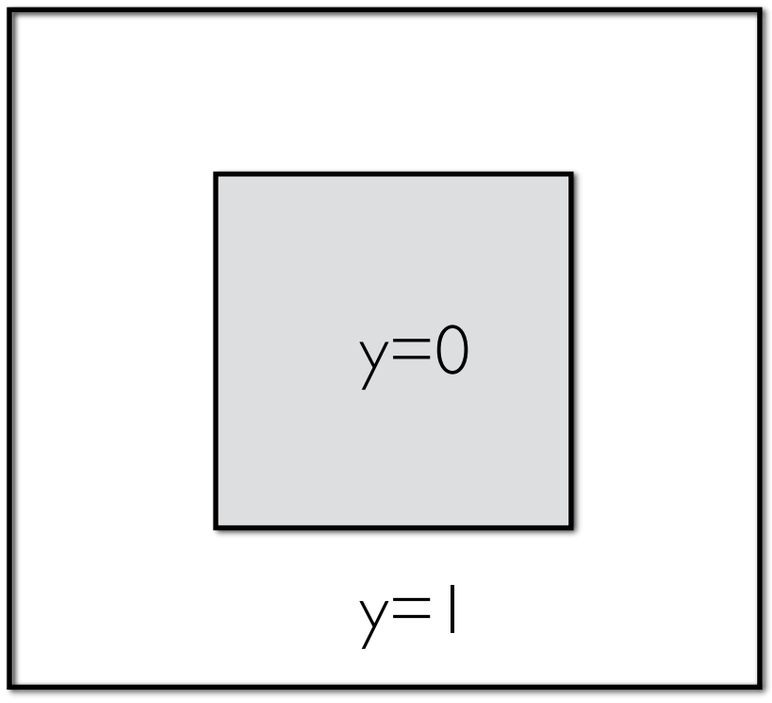
- Consider setting of regression
  - *features* X, and *target* Y
  - goal is to predict Y given X: let $\hat{Y}(X)$ be prediction
  - so as to minimize *error* or *loss*
  - consider average squared error: $\mathbb{E}[(Y - \hat{Y}(X))^2]$
- That is, goal is to $\min_{\hat{Y}} \mathbb{E}[(Y - \hat{Y}(X))^2]$
- Solution turns out to be: $\mathbb{E}[Y|X]$
- Different way to "fit" $\mathbb{E}[Y|X]$
  - *Parametric*: Linear, Neural Networks, …
  - *Non-parametric*: Nearest-neighbor, …

Overfitting

- Trust your data, but only so much…
- An example:
  - observations: $(x_i, y_i), 1 \leq i \leq n$
  - function fit

$$f(x) = \begin{cases} y_i & \text{if } x = x_i \\ 0 & \text{otherwise} \end{cases}$$

  - perfect fit for observation
  - but, as bad as "random" function!



x distributed uniformly in the unit square

Preventing Overfitting

- Cross-validation:
    - data split into two parts: *train* and *test*
    - *k-fold* cross-validation

- Regularization:
    - Model complexity as a penalty (a la Occam's Razor)

- Information criteria:
    - Akaike information criteria (AIC)
    - Bayesian information criteria (BIC)

Bias-Variance Tradeoff

- Best model fit tries to balance between
  - model inaccuracy in terms of "bias"
  - model complexity in terms of "variance"
- In the context of regression with squared loss:
  - for any estimator $\hat{Y}(X)$, it can be shown that

$$\mathbb{E}[(Y - \hat{Y}(X))^2] = \mathbb{E}[(Y - \mathbb{E}[Y|X])^2] + \mathbb{E}[(\mathbb{E}[Y|X] - \mathbb{E}[\hat{Y}(X)])^2] + \mathbb{E}[(\hat{Y}(X) - \mathbb{E}[\hat{Y}(X)])^2]$$

loss | inherent loss | $(\text{bias})^2$ | variance