Topics in Machine Learning
Machine Learning for Healthcare

Rahul G. Krishnan
Assistant Professor
Computer science & Laboratory Medicine and Pathobiology

Slide credits to David Sontag & Uri Shalit

Announcements

- Last few weeks of classes – project presentations start next week
- Come by office hours
- TAs will reach out for practice presentations – note that each presentation should be ~25 minutes (including 3 mins for questions)

Outline

- Last week: Covariate adjustment for estimating causal effects
  - **Key idea**: Use machine learning to predict outcome given features and impute counterfactuals
- This lecture: Matching & Propensity score matching
- Missingness
- Next lecture: case studies in ML4H

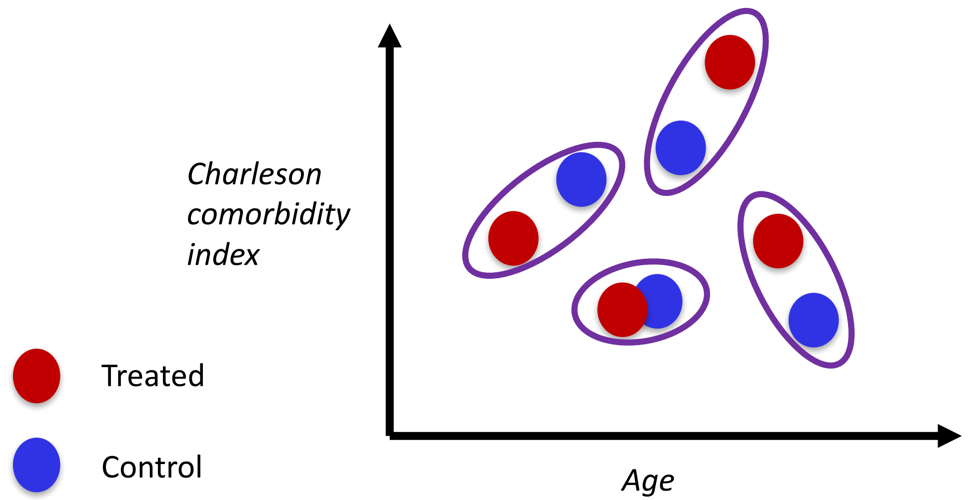
Recap: Covariate adjustment

- Explicitly model the relationship between treatment, confounders, and outcome
- Under ignorability, the expected causal effect of T on Y:
  $\mathbb{E}_{x \sim p(x)}\big[\, \mathbb{E}[Y_1 | T = 1, x] - \mathbb{E}[Y_0 | T = 0, x] \big]$
- Fit a model $f(x, t) \approx \mathbb{E}[Y_t | T = t, x]$

$$\widehat{CATE}(x_i) = f(x_i, 1) - f(x_i, 0)$$

Matching

- Find each individual's nearest counterfactual twin and use their outcome as a proxy for the individual's counterfactual



Identical to covariate adjustment when 1NN classifier is used as a classifier

Effect estimation with matching

$$\forall i \ \text{cfac(i)} = \text{argmin}_{j;\ t_i \neq t_j} d(x_j, x_i)$$

$$\text{CATE}(x_i) = \mathbb{I}[t_i == 1](y_i - y_{\text{cfac(i)}}) + \mathbb{I}[t_i == 0](y_{\text{cfac(i)}} - y_i)$$

$$\text{ATE} = \frac{1}{n}\sum_i \text{CATE}(x_i)$$

Interpretable!

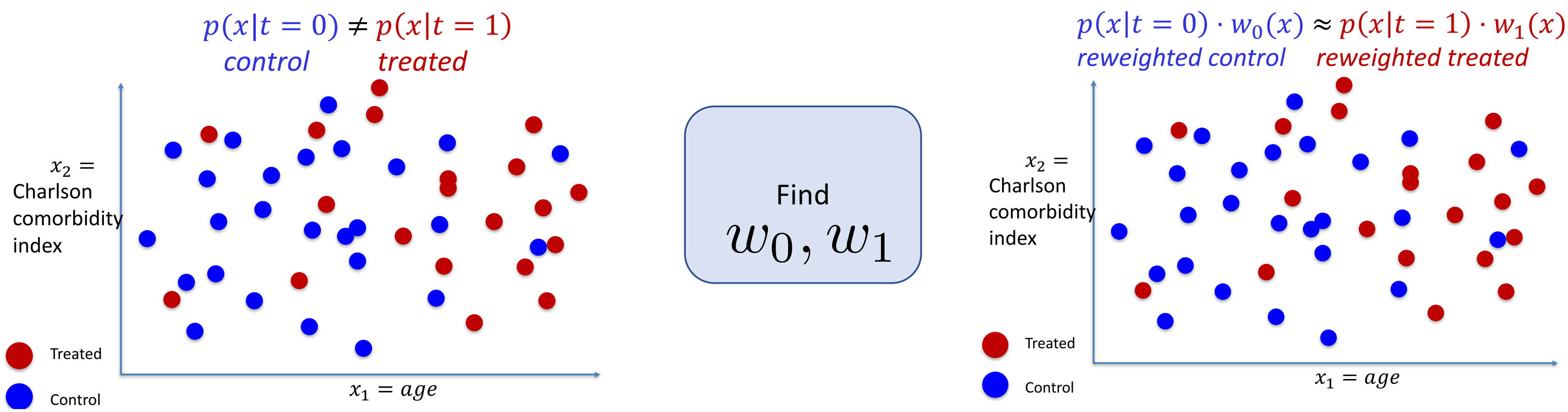
Non-parametric

Sensitive to the choice of metric d

Suffers from all the limitations of K-Nearest Neighbors

Propensity scores

- Reweight samples to turn an observational study into a pseudo-randomized trial



Propensity score (algorithm)

How to calculate ATE with propensity score

for sample $(x_1, t_1, y_1), \dots, (x_n, t_n, y_n)$

1. Use any ML method to estimate $\hat{p}(T = t|x)$

2. $$A\hat{T}E = \frac{1}{n} \sum_{i \text{ s.t. } t_i=1} \frac{y_i}{\hat{p}(t_i = 1|x_i)} - \frac{1}{n} \sum_{i \text{ s.t. } t_i=0} \frac{y_i}{\hat{p}(t_i = 0|x_i)}$$

Propensity score (for an RCT)

How to calculate ATE with propensity score

for sample $(x_1, t_1, y_1), \dots, (x_n, t_n, y_n)$

Sum over $\sim \frac{n}{2}$ terms

1. Randomized trial $p = 0.5$

2. $$\hat{ATE} = \frac{1}{n} \sum_{i \text{ s.t. } t_i=1} \frac{y_i}{0.5} - \frac{1}{n} \sum_{i \text{ s.t. } t_i=0} \frac{y_i}{0.5} =$$
$$\frac{2}{n} \sum_{i \text{ s.t. } t_i=1} y_i - \frac{2}{n} \sum_{i \text{ s.t. } t_i=0} y_i$$

Propensity score - derivation

- We want: $\mathbb{E}_{x\sim p(x)}[Y_1(x)]$

- We know that:

$$p(x|T=1)\cdot\frac{p(T=1)}{p(T=1|x)}=p(x)$$

- Thus:

$$\mathbb{E}_{x\sim p(x|T=1)}\left[\frac{p(T=1)}{p(T=1\mid x)}Y_1(x)\right]=\mathbb{E}_{x\sim p(x)}[Y_1(x)]$$

- We can approximate this empirically as:

$$\frac{1}{n_1}\sum_{i\ \text{s.t.}\ t_i=1}\left[\frac{n_1/n}{\hat{p}(t_i=1\mid x_i)}y_i\right]=\frac{1}{n}\sum_{i\ \text{s.t.}\ t_i=1}\frac{y_i}{\hat{p}(t_i=1\mid x_i)}$$

(similarly for t_i=0)

Propensity score - challenges

- If not much overlap, scores become non-informative
  - Think about what happens if $P(T=1|X) = 1$ or 0 vs 0.5
- When propensity scores are small, the estimators have a large variance

Natural experiments

- Does stress during pregnancy affect later child development?
- Confounding: genetic, mother personality, economic factors…
- Natural experiment: the Cuban missile crisis of October 1962. Many people were afraid a nuclear war is about to break out.
- Compare children who were in utero during the crisis with children from immediately before and after

Instrumental Variables

- Informally: a variable which affects treatment assignment but not the outcome
- Example: are private schools better than public schools?
- Can’t force people which school to go to
- Can *randomly* give out vouchers to some children, giving them an opportunity to attend private schools
- The voucher assignment is the instrumental variable

Causal inference - Overview

- The last two lectures give us an overview of two techniques to estimate causal effects:
  - Assumptions necessary for causal inference
    - Positivity
    - Common support
    - No unobserved confounding
  - Covariate adjustment
  - Propensity score matching
  - **Key idea:** Using assumptions to write down estimators of causal effects using observational data
- Many more extensions to more complex regimes where data are time-varying

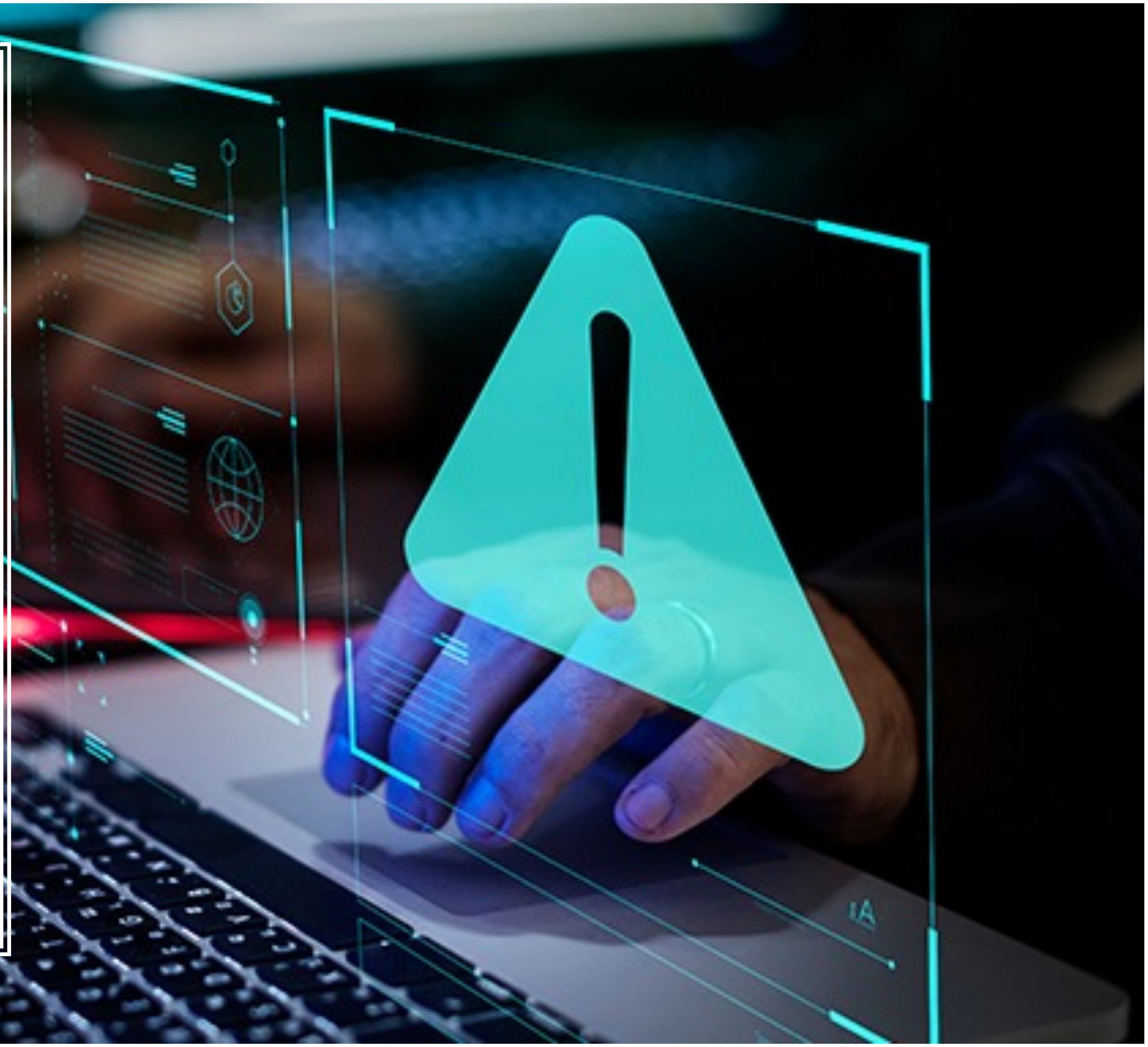
Missingness

- Missingness is a common aspect of many kinds of data in healthcare
- Unlike domains like computer vision and natural language processing
- Important to know how to:
  - Identifying and categorize the different kinds of missingness
  - Know about common techniques used to handle missing data and their limitations/strengths

Why does missingness occur in clinical data?

Rationales for missingness

- Patients do not consistently interact with the healthcare system
- Errors in data entry
- Errors in extracting information from the Electronic Medical Record (typically a set of SQL tables inside a hospital database)



Handling missingness

- How well does imputation do?
- In general, it depends on the *kind* of missingness
- Lets see a specific example of how things can go wrong

Ramifications of improperly handling missingness

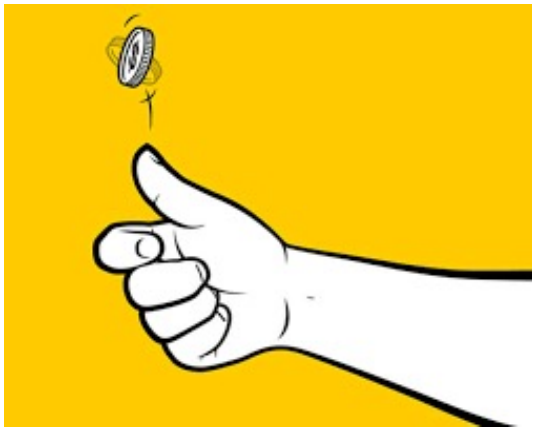
1. Generate synthetic data matrix and learn a linear regression model. Our goal is to estimate the first feature of vector w

2.
$$y = 4 + w^T x + \epsilon$$
$$w = [1, 2, -1, -2]$$

3. Assume that we have access to a large number of samples (~100K) of x under ***two*** kinds of missingness

4. Assess the effect of learning regression function
when data are missing and imputed with 0

Missingness mechanism for feature (1)



- Missingness mechanism:
  - Flip a coin, if heads, impute second feature with 0, if tails do not impute
- Train linear regression with imputed features
- Look at learned coefficients

Missingness mechanism for feature (2)

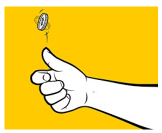
- Missingness mechanism:
  - If second feature is greater than 4, then impute to 0, otherwise do not impute
- Train linear regression with imputed features
- Look at learned coefficients

Question: Will we recover the true regression coefficients? In which case is the recovery easier/harder?

Results

Question: Which missingness is harder to recover parameters from?

Scenario (1)



Scenario (2)

w* = [1.0025 1.995 -0.9999 -1.997]

w* = [1.622 -0.372 -0.375 -1.369]

Ground truth

w* = [1. 2. -1. -2.]

Not all missingness is created equal – some are harder to recover from and naïve methods for imputation can result in biased parameters

A taxonomy of missingness

- Graphical Models for Inference with Missing Data, Mohan et. al, 2013
- Addresses the problem of recoverability – deciding when there exists a consistent estimator for a probabilistic query
- **Key idea:**
  - Derive a causal graph that characterizes the missingness process
  - Use this representation to derive conditions under which query can be answered

Missingness graphs

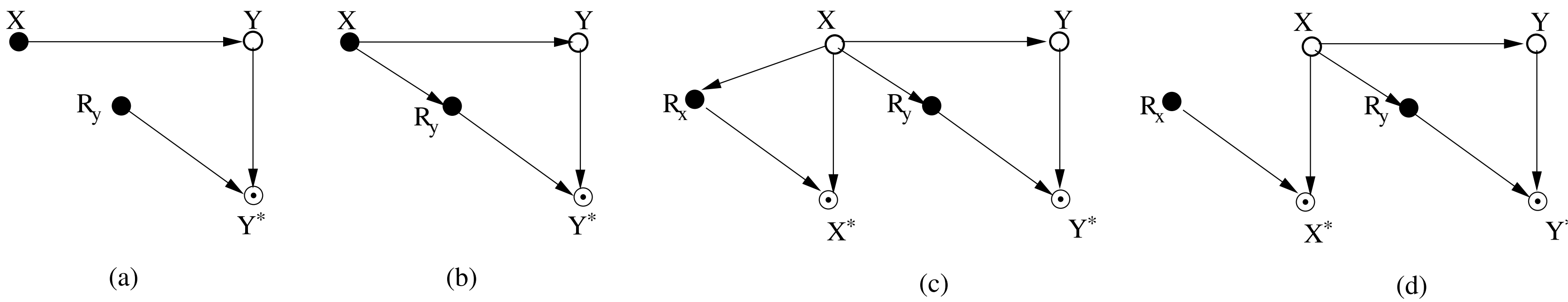
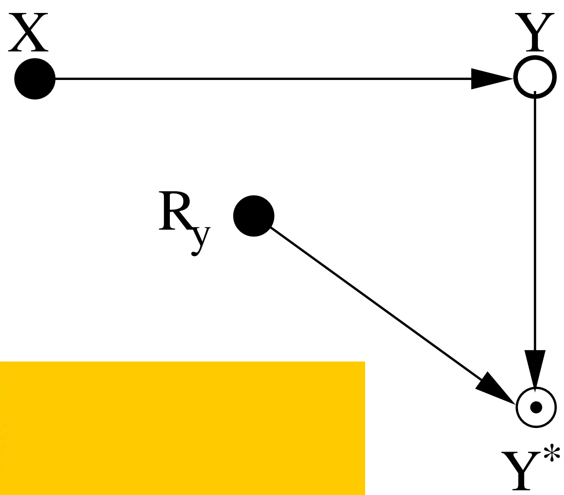


Figure 1: m-graphs for data that are: (a) MCAR, (b) MAR, (c) & (d) MNAR; Hollow and solid circles denote partially and fully observed variables respectively.

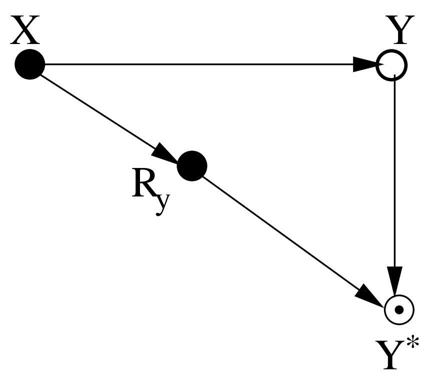
$$v_i^* = f(r_{v_i}, v_i) = \begin{cases} v_i & \text{if } r_{v_i} = 0 \\ m & \text{if } r_{v_i} = 1 \end{cases}$$

MCAR [Missing completely at random]



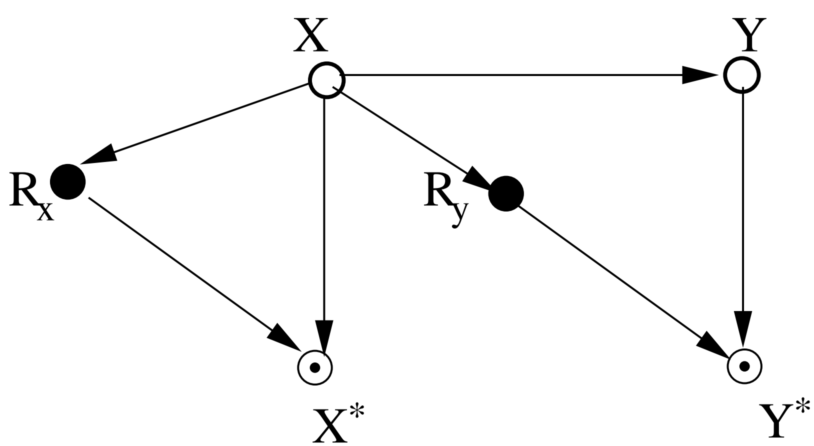
- The indicator (R) that decides whether or not we observe the value of Y is marginally independent
- Often leads to consistent estimates for probabilistic queries
- **Example**: Tabular data corrupted randomly during transmission due to a noisy channel

MAR [Missing at random]

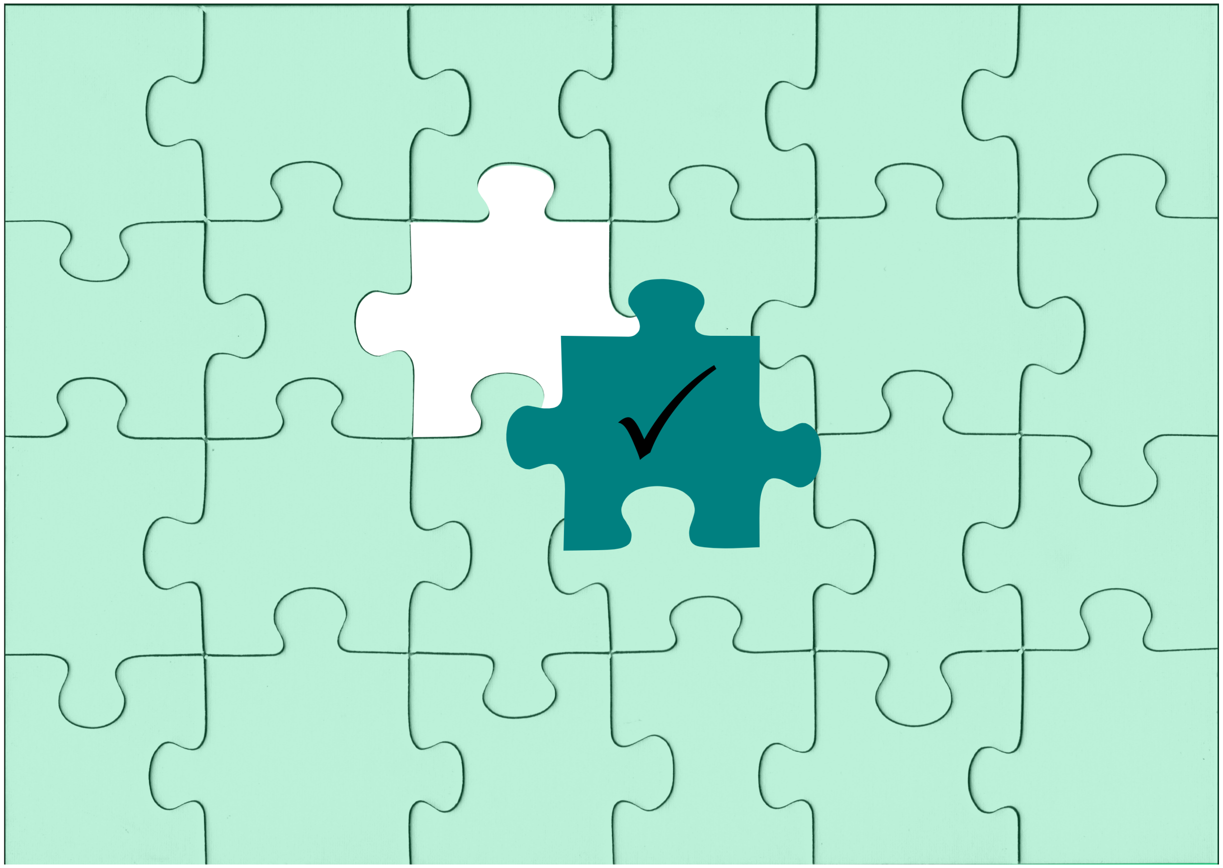


- Whether or not a value is missing here depends on the value of other (observed) features
- Techniques like MICE which build predictive models of features for imputation will work well here
- **Example**: Childhood data handled by different hospital team than adult data which could affect missingness pattern

MNAR (Missing not at random)



- Whether or not a feature is missing depends on the value the feature takes.
- Often not possible to answer probabilistic queries without assumptions/domain knowledge
- **Example**: The patient’s blood pressure is not recorded if it is above 127



Missingness in tabular data

- Impute missing data with zero
- Impute missing data with mean of observed data in that column
- Impute missing data using prior knowledge
- Use ideas from machine learning to impute missing data

Multiple Imputation with Chained Equations (MIC)

- MICE [van Buuren et. al, JSS 2011]
  - Also known as sequential regression multiple imputation
  - Extension to multiple imputation [Rubin, 1987]
- Algorithm:
  - Learn a parametric model of each feature given all the others
  - Impute the missing data
  - Retraining models
  - Repeat
- Found to work well in practice with many open source packages

$$P(Y_1|Y_{-1},\theta_1)$$

$$\vdots$$

$$P(Y_p|Y_{-p},\theta_p).$$

Missingness in longitudinal data

- More challenging to impute missingness in longitudinal data since imputations have to be *consistent* with observed dynamics
- Zero imputation – generally not a good idea
- Forward-fill imputation
  - Carry forward the previous value
- Model-based imputation

Model-based handling of missingness in longitudinal data [supervised learning]

- Use forward fill imputation and append missingness vectors into the model
- Example – RNNs for multivariate time-series with missing data

Model-based handling of missingness in longitudinal data [unsupervised learning]

- Learn a statistical model of the observed data and use it to impute the unobserved values
- Maximum likelihood estimation for state space models is feasible even when data is missing

Questions?