Want more?

https://github.com/matthewbrems/missing-data-workshop

Good, Fast, Cheap: How to Do Data Science with Missing Data

Matt Brems

Introduction: Matt Brems (he/him)

Senior Manager, Data Science Product & Strategy, DataRobot

Managing Partner & Principal Data Scientist, BetaVector

Distinguished Faculty, General Assembly

Vice Chair, Statistics Without Borders

Previously:

- Growth + Computer Vision @ Roboflow
- R&D Fairness/Bias in Data @ FINRA
- Data Science Education @ General Assembly
- Data Science @ 0ptimus Consulting
- Enterprise Analytics @ Smucker's
- M.S. Statistics @ The Ohio State University

Recommended Reads:

- Data-Driven Thinking: "Factfulness"
- Data Visualization: "Storytelling with Data"
- Data Science: "Introduction to Statistical Learning with Applications in R"

Agenda

1. **Introduction to missing data.**

2. Strategies for doing data science with missing data.
    a. Avoid missing data.
    b. Ignore missing data.
    c. Account for missing data.
        i. Unit missingness.
        ii. Item missingness.

3. Practical considerations and warnings.

How big of a problem is missing data?

This is a difficult question to answer.

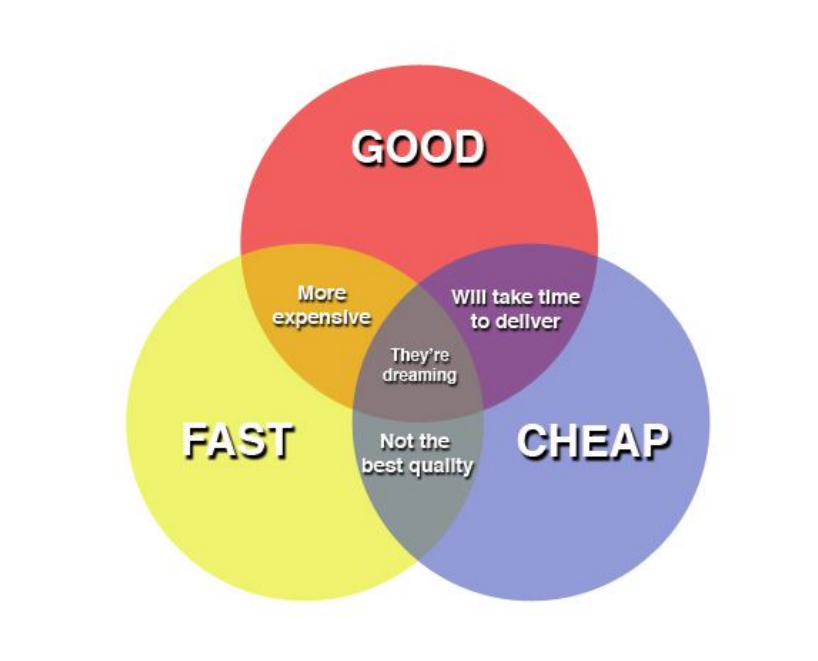
Practically, we can only see what we observe.

We can use **simulated data** to help answer this question.

To the notebook!

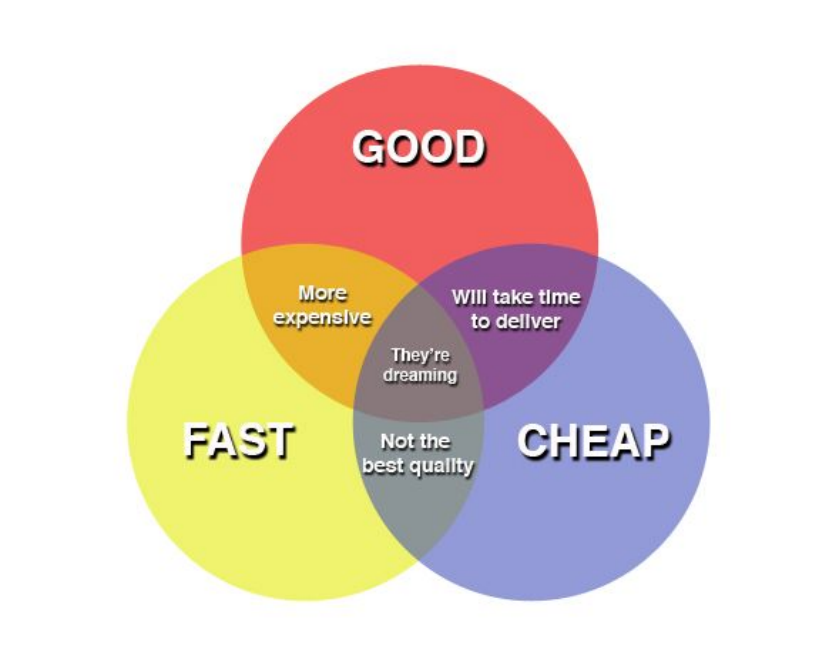
`00_interactive_plot.ipynb`

What is a realistic approach for us?



Source: http://www.pyragraph.com/2013/05/good-fast-cheap-you-can-only-pick-two/

Do you want your analysis to be…



Source: http://www.pyragraph.com/2013/05/good-fast-cheap-you-can-only-pick-two/

Fast and Cheap: Drop all missing values or single imputation.

Good and Cheap: Proper imputation or pattern submodel method.

Good and Fast: Gather data in a complete manner.

Agenda

1. Introduction to missing data.

2. **Strategies for doing data science with missing data.**
    a. Avoid missing data.
    b. Ignore missing data.
    c. Account for missing data.
        i. Unit missingness.
        ii. Item missingness.

3. Practical considerations and warnings.

Avoiding missing data.

- It's often more expensive up front but **cheaper in the long run to avoid missing data** than to make guesses about how to best handle our missing data.

  - Decrease burden on your respondent.
  - Change method of data collection.
  - Improve accessibility.
  - Change timing of your survey.
  - Minimize length of questionnaire.
  - Consider content of your survey.

Ignoring missing data.

- We **assume** that observations for which we've observed data are **similar to observations for which we're missing data**.

- One *general, **very rough** guideline* is that we *may* be OK ignoring missing data if less than 5% of our data is missing.
  - If we're doing supervised learning and we're missing a lot of our Y variable, this may be inadvisable.
  - If we're missing a lot from meaningful variables, this may be inadvisable.

Before “accounting for” missing data, let’s shift our mindset.

- There’s a naive belief that we can just plug in the gaps in our data.
  - This is known as **imputation**.
  - We have to do this in a specific way, or we’re just making up data.

- **In most cases, we aren’t “fixing” data. We’re just learning how to cope with it!**

Agenda

1. Introduction to missing data.

2. Strategies for doing data science with missing data.
    a. Avoid missing data.
    b. Ignore missing data.
    c. **Account for missing data.**
        i. Unit missingness.
        ii. Item missingness.

3. Practical considerations and warnings.

What's the difference: Unit vs. Item Missingness?

Unit missingness has all values missing from an observation.

- Index 3.

Item missingness is where some, but not all, values are missing from an observation.

- Indices 1, 2, and 10,000.

Index	Age	Sex	Income
1	NA	M	NA
2	39	NA	75000
3	NA	NA	NA
4	28	F	50000
...	...	...	...
10000	18	F	NA

To the notebook!

`01_unit_missingness.ipynb`

Pulling the pieces together in one workflow...

1. I evaluate how much missing data I have during EDA. **Is it worth my time to try to address it?**

2. Is it reasonable to attempt **deductive imputation**?

3A. If my goal is to **generate predictions**, then use the **pattern submodel** method.

3B. If my goal is to **conduct inference**, then use the best imputation method available. If I'm doing proper imputation, combine output and results using **Rubin's rules**.

Scenario 1: Missing Completely at Random (MCAR)

I'm a grad student in a lab. While pipetting, I reach for my pen but accidentally knock a Petri dish off of the desk. From this Petri dish, I lose the data that I otherwise would have collected.

This is called **missing completely at random.**

- The data of interest is not systematically different between missing and observed.

bacteria on day 1	bacteria on day 2
10mm	15mm
12mm	12mm
9mm	11mm
10mm	11mm
15mm	19mm
13mm	15mm
11mm	~~16mm~~

Scenario 2: Missing at Random (MAR)

I work for the Department of Transportation. A sensor on the Pennsylvania Turnpike broke and did not gather information between 7:00am and 10:00am.

This is called **missing at random.**

- **Conditional on data we have observed**, the data of interest is not systematically different between missing and observed.
- Whether or not a data point is missing is dependent on observed data.

time	number of vehicles
4:00	206
5:00	519
6:00	934
7:00	~~1,650~~
8:00	~~1,921~~
9:00	~~1,010~~
10:00	889

Scenario 3: Not Missing at Random (NMAR)

I administer a survey with a question about income. Those who have lower incomes are less likely to reply to the income question.

This is called **not missing at random.**

- The data of interest are systematically different for missing and observed.
- Whether or not an observation is missing depends on the value of the unobserved data itself!

id	income
A	~~48,000~~
B	~~35,000~~
C	105,000
D	~~62,000~~
E	80,000
F	50,000
G	75,000

What type of missing data? (MCAR, MAR, NMAR)

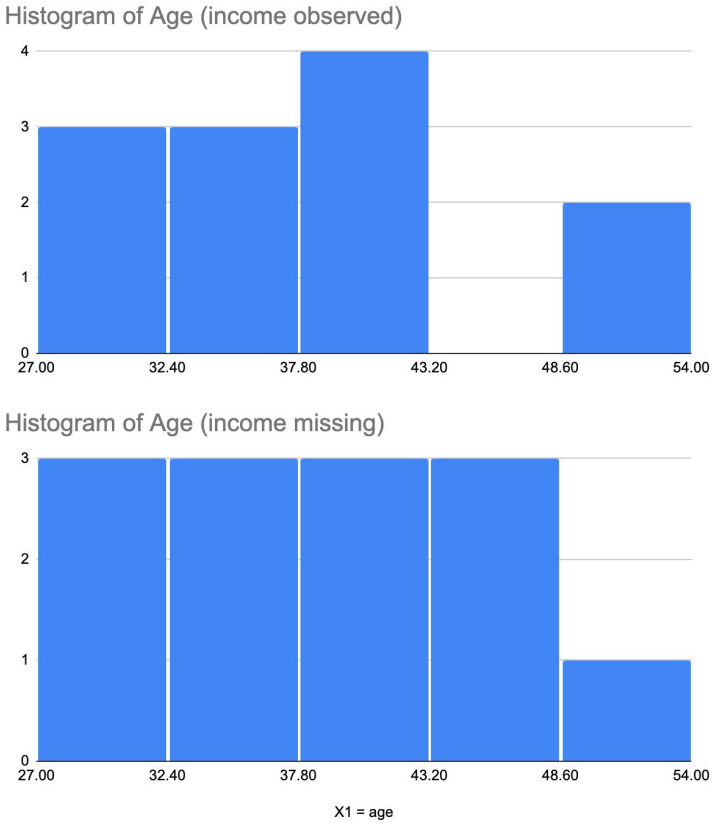
1. Little's Test for MCAR
   - $\mathbf{H_0}$: data is MCAR vs. $\mathbf{H_A}$: data is MAR
   - *No empirical test exists to evaluate whether or not data is NMAR.*

2. Split your data into two sets: one where your variable is observed, one where your variable is missing. **Can you spot patterns in other variables?** (e.g. is age significantly lower where income is missing?)

3. **Think about the missing data process.** Can you come up with a reasonable theory about how missing data come about?

Split your data into two sets.
Can you spot patterns in other variables?

X1 = age	X2 = income
38	48,000
59	35,000
55	105,000
56	62,000
51	80,000
30	50,000
60	75,000
40	98,000
44	100,000
33	103,000
40	110,000
25	54,000
42	~~NA~~
59	~~NA~~
28	~~NA~~
60	~~NA~~
43	~~NA~~
26	~~NA~~
27	~~NA~~



- **If the histograms are sufficiently different**, then we might think that the graphed variable (Age) helps explain whether or not the other variable (Income) is missing.

- **If the histograms are not sufficiently different**, then we might think that the graphed variable (Age) doesn't help explain whether or not the other variable (Income) is missing.

How do we account for item nonresponse?

- Deductive Imputation
- Mean/Median/Mode Imputation
- Single Regression Imputation
- Proper Imputation
- Pattern Submodel Approach

Deductive Imputation

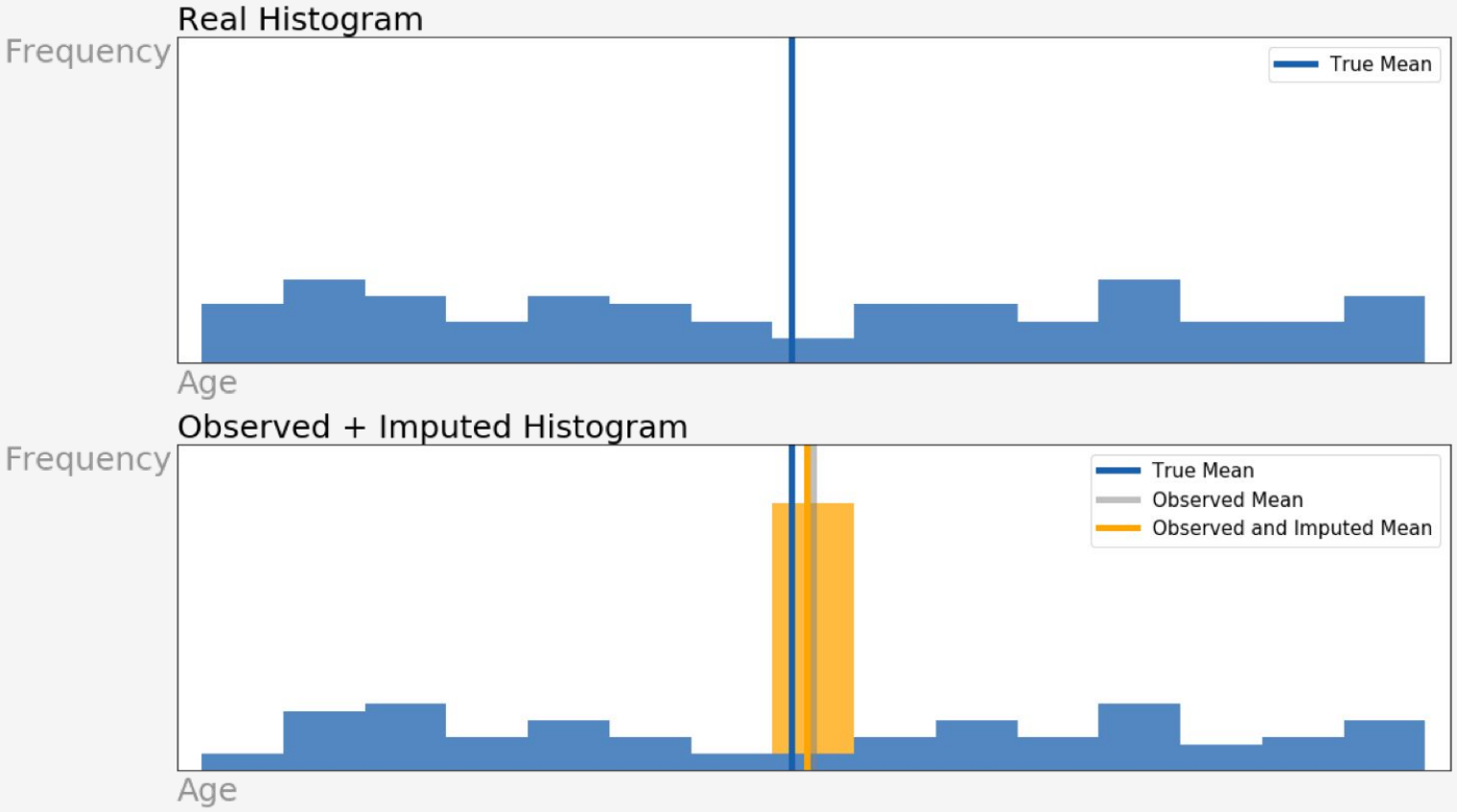
- **We use logical rules to fill in missing values**.
  - Survey asks if the respondent has logged into a Facebook-owned product in the previous month.
  - Respondent says no.
  - Survey then asks the respondent to select which Facebook product they've used the most in the previous month.
  - Respondent leaves this answer blank, including the "None of the above" option.

- Requires specific coding, can be time consuming, requires no inference, can be used regardless of missingness type.

Mean/Median/Mode Imputation

- For any “NA” value in a given column, replace “NA” with the mean, median, or mode.
- Quick, easy to implement, seems reasonable, can significantly distort histogram, underestimates variance, should only be considered if data is MCAR.

Mean Imputation



Goal: We want to calculate the standard deviation of X.

We have n rows, with k observed values and n-k missing.

Y = output	X = input
4	24
5	23
8	41
1	11
1	9
9	24
9	34
8	25
5	25
2	12
10	29
9	28
8	~~NA~~
4	~~NA~~
2	~~NA~~
9	~~NA~~
2	~~NA~~
3	~~NA~~
5	~~NA~~

n total rows in dataset

k observed values

n-k missing values

Why is underestimating variance a bad thing?

We have n rows, with k observed values and n-k missing.

$$s = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(x_i - \bar{x})^2}$$

Definition of SD for k observed values

Why is underestimating variance a bad thing?

We have n rows, with k observed values and n-k missing.

$$s = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(x_i - \bar{x})^2}$$

Definition of SD for k observed values

$$= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Definition of SD for n “observed” values

Goal: We want to calculate the standard deviation of X.

We have n rows, with k observed values and n-k missing.

n total rows in dataset

Y = output	X = input
4	24
5	23
8	41
1	11
1	9
9	24
9	34
8	25
5	25
2	12
10	29
9	28
8	~~NA~~
4	~~NA~~
2	~~NA~~
9	~~NA~~
2	~~NA~~
3	~~NA~~
5	~~NA~~

→

Y = output	X = input
4	24
5	23
8	41
1	11
1	9
9	24
9	34
8	25
5	25
2	12
10	29
9	28
8	23.75
4	23.75
2	23.75
9	23.75
2	23.75
3	23.75
5	23.75

k observed values

fill in mean of observed values

Why is underestimating variance a bad thing?

We have n rows, with k observed values and n-k missing.

$$s = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(x_i - \bar{x})^2}$$

Definition of SD for k observed values

$$= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Definition of SD for n “observed” values

$$= \sqrt{\frac{1}{n}\left[\sum_{i=1}^{k}(x_i - \bar{x})^2 + \sum_{i=k+1}^{n}(\bar{x} - \bar{x})^2\right]}$$

Rewrite definition for k observed values and n-k imputed values

Why is underestimating variance a bad thing?

We have n rows, with k observed values and n-k missing.

$$s = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(x_i - \bar{x})^2}$$

Definition of SD for k observed values

$$= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Definition of SD for n “observed” values

$$= \sqrt{\frac{1}{n}\left[\sum_{i=1}^{k}(x_i - \bar{x})^2 + \sum_{i=k+1}^{n}(\bar{x} - \bar{x})^2\right]}$$

Rewrite definition for k observed values and n-k imputed values

$$= \sqrt{\frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^2}$$

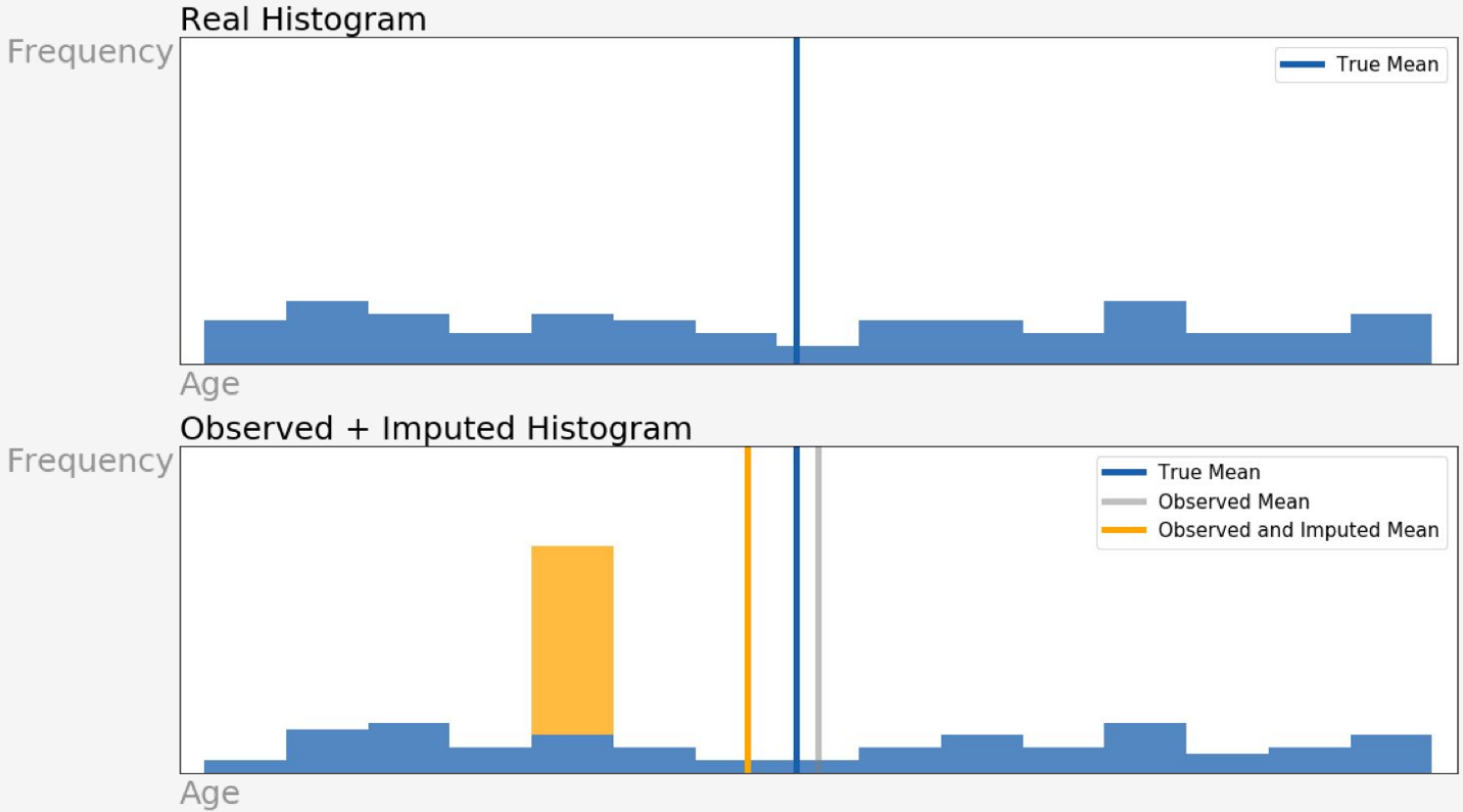
Simplify by removing all zeros

Why is underestimating variance a bad thing?

- If we underestimate variance in a confidence interval, our **confidence interval gets smaller for the same level of confidence**!
  - Our results ***artificially*** look more precise… but only because we imputed the mean!

- If we underestimate the variance in a hypothesis test, **our *p*-value will *artificially* get smaller**.
  - Our *p*-value may look significant… but only because we imputed the mean!

Mode Imputation



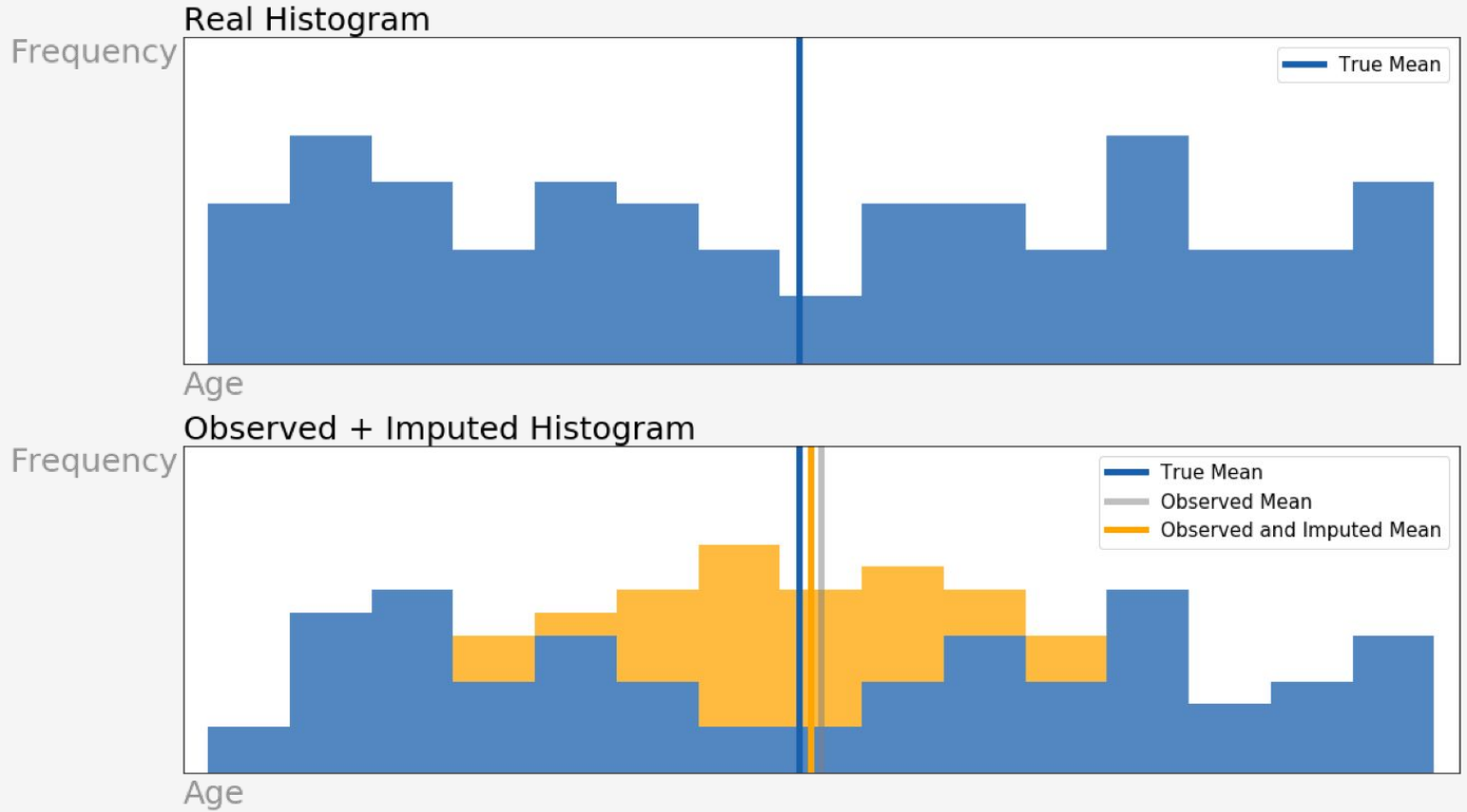
Single Regression Imputation

- Train a model on the rows of your data that are **fully observed**.
  - Suppose I'm missing income for some people, but have observed age and highest level of education for everybody.
  - X = age and highest level of education, Y = income.
  - Fit a model.

- For any "NA" value in a given column, replace "NA" with predicted value from that model.

- Seems reasonable, still distorts histogram, underestimates variance, should only be considered if data is MCAR or MAR.

Single Regression Imputation



Single Regression Imputation

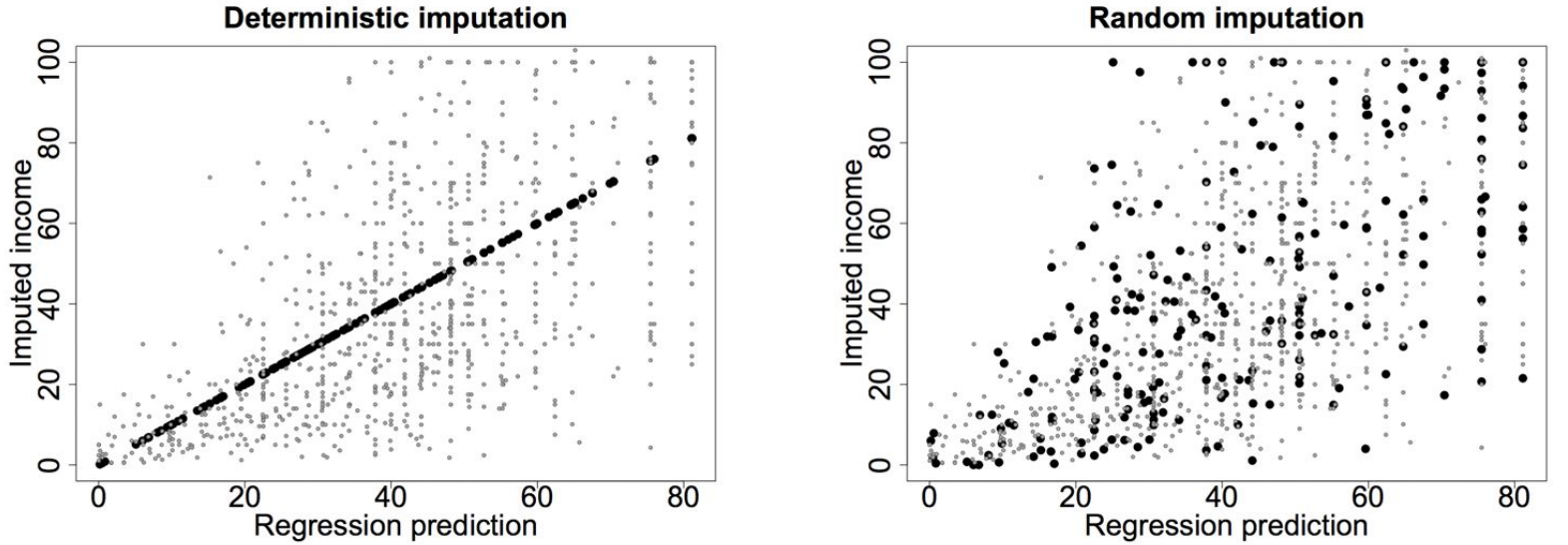


Figure 25.2 *Deterministic and random imputations for the 241 missing values of earnings in the Social Indicators Survey. The deterministic imputations are exactly at the regression predictions and ignore predictive uncertainty. In contrast, the random imputations are more variable and better capture the range of earnings in the data. See also Figure 25.1.*

Source: Andrew Gelman & Jennifer Hill, Data Analysis using Regression and Multi-level/Hierarchical Models: http://www.stat.columbia.edu/~gelman/arm/missing.pdf

Proper Imputation

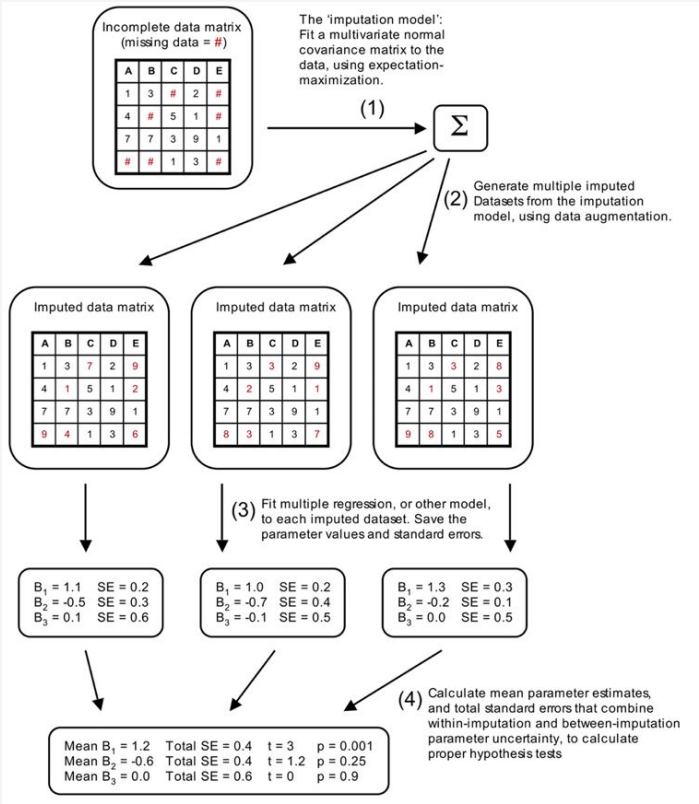
- In order to properly impute data, we need to fill in missing values with the proper amount of certainty/uncertainty.
- Replacing an NA with one value treats things like we know the true value.
- Therefore, **we need to impute multiple times**.

Proper Imputation

- We will make multiple copies (say 10) of our dataset.
- We will use random regression imputation to generate one value for each NA in each dataset.
- Once each of our 10 datasets are complete, we will do our “final model” or “final analysis” on each dataset.
- We will then combine the results of our multiple models together, just like we aggregate results in an ensemble model.

Proper Imputation

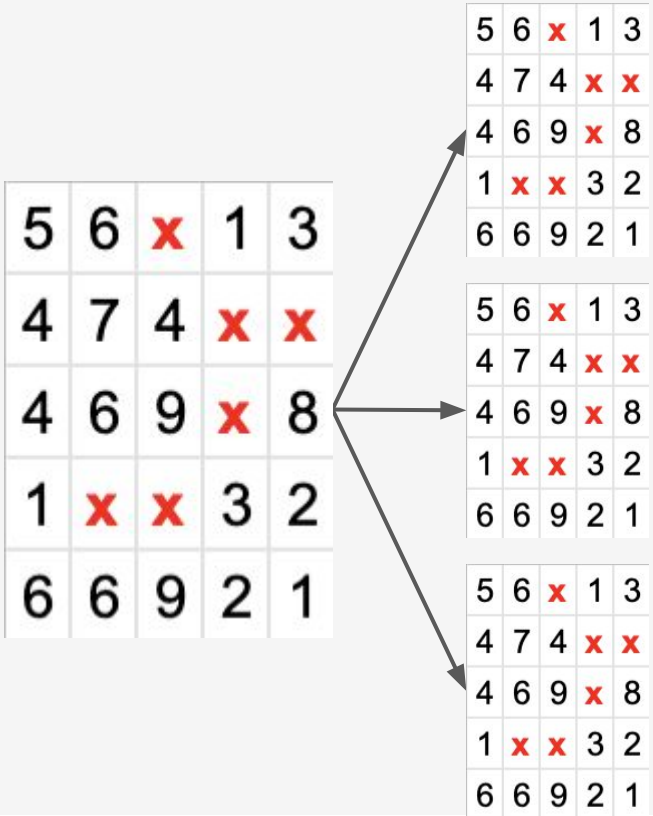
- We will make multiple copies (say 10) of our dataset.
- We will use random regression imputation to generate one value for each NA in each dataset.
- Once each of our 10 datasets are complete, we will do our “final model” or “final analysis” on each dataset.
- We will then combine the results of our multiple models together, just like we aggregate results in an ensemble model.



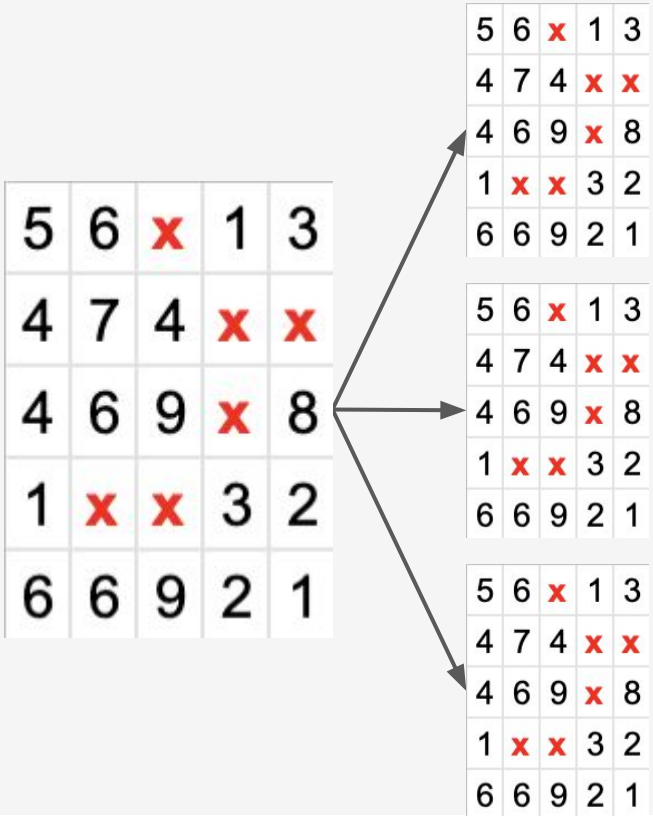
Proper Imputation

5	6	x	1	3
4	7	4	x	x
4	6	9	x	8
1	x	x	3	2
6	6	9	2	1

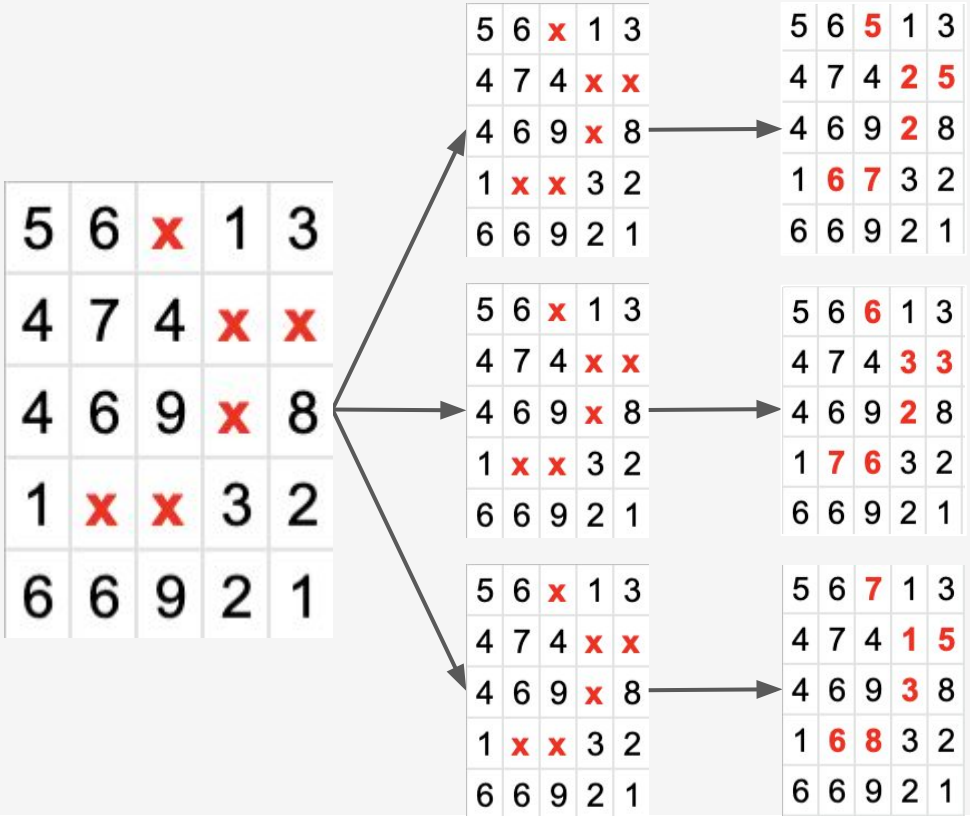
Proper Imputation -- Step 1: Make Copies



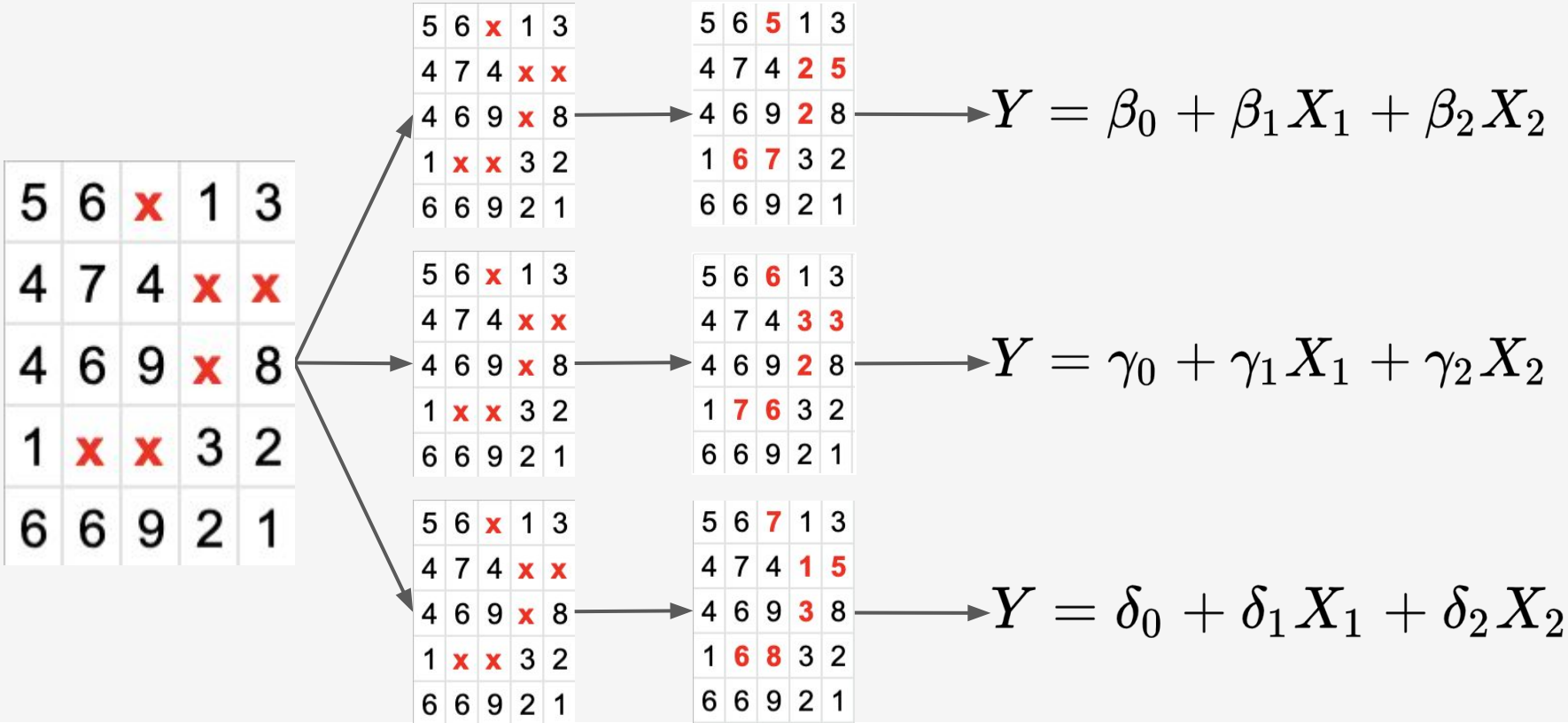
Proper Imputation -- Step 2: Build Model for Imputing



Proper Imputation -- Step 3: Impute values *(properly!)*



Proper Imputation -- Step 4: “Final analysis” on each set.



Proper Imputation -- Step 5: Combining "Models"

- **If you're generating predictions**, you can just:
  - average your predictions together in a regression problem, or
  - select the plurality class in a classification problem.

- **If your goal is to do inference** (e.g. understand how Y changes as X changes) and are fitting a linear model to each dataset, then you get a slope and y-intercept for each model.
  - **Rubin's rules**: a set of tools that allows you to take the slopes and y-intercepts from each model and combine them together.
  - Check out documentation in the repo if interested!

To the notebook!

`02_item_missingness.ipynb`

How do we account for item nonresponse?

- Deductive Imputation
- Mean/Median/Mode Imputation
- Single Regression Imputation
- Proper Imputation
- Pattern Submodel Approach

Pattern Submodel Approach

- Big Picture: We will break our dataset into subsets based on missingness pattern. We will then build one model on each subset, creating many different models.

Pattern Submodel Approach

Y	X1	X2
obs	obs	obs
obs	obs	obs
obs	obs	~~NA~~
obs	obs	~~NA~~
obs	~~NA~~	obs
obs	~~NA~~	obs
obs	~~NA~~	~~NA~~

1. We will break our dataset into subsets based on missingness pattern.

2. We will then build one model on each subset, creating many different models.

Pattern Submodel Approach

Y	X1	X2
obs	obs	obs
obs	obs	obs
obs	obs	~~NA~~
obs	obs	~~NA~~
obs	~~NA~~	obs
obs	~~NA~~	obs
obs	~~NA~~	~~NA~~

1. **We will break our dataset into subsets based on missingness pattern.**

2. We will then build one model on each subset, creating many different models.

Pattern Submodel Approach

Y	X1	X2
obs	obs	obs
obs	obs	obs
obs	obs	~~NA~~
obs	obs	~~NA~~
obs	~~NA~~	obs
obs	~~NA~~	obs
obs	~~NA~~	~~NA~~

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

1. We will break our dataset into subsets based on missingness pattern.

2. **We will then build one model on each subset, creating many different models.**

Pattern Submodel Approach

Y	X1	X2
obs	obs	obs
obs	obs	obs
obs	obs	~~NA~~
obs	obs	~~NA~~
obs	~~NA~~	obs
obs	~~NA~~	obs
obs	~~NA~~	~~NA~~

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

$$Y = \gamma_0 + \gamma_1 X_1$$

1. We will break our dataset into subsets based on missingness pattern.
2. **We will then build one model on each subset, creating many different models.**

Pattern Submodel Approach

Y	X1	X2	
obs	obs	obs	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
obs	obs	obs	
obs	obs	~~NA~~	$Y = \gamma_0 + \gamma_1 X_1$
obs	obs	~~NA~~	
obs	~~NA~~	obs	$Y = \delta_0 + \delta_1 X_2$
obs	~~NA~~	obs	
obs	~~NA~~	~~NA~~	$Y = \varepsilon_0$

1. We will break our dataset into subsets based on missingness pattern.

2. **We will then build one model on each subset, creating many different models.**

Pattern Submodel Approach

- When data are NMAR, the pattern submodel method is expected to outperform imputation methods. When data are MCAR or MAR, the pattern submodel method is expected to perform about as well as imputation methods.
- **It does not require missingness assumptions**!
- You can generate predictions for test observations containing missing data.
- This is not a well-understood method for inference.

Pulling the pieces together in one workflow…

1. I evaluate how much missing data I have during EDA. **Is it worth my time to try to address it?**

2. Is it reasonable to attempt **deductive imputation**?

3A. If my goal is to **generate predictions**, then use the **pattern submodel** method.

3B. If my goal is to **conduct inference**, then use the best imputation method available. If I'm doing proper imputation, combine output and results using **Rubin's rules**.

Agenda

1. Introduction to missing data.

2. Strategies for doing data science with missing data.
    a. Avoid missing data.
    b. Ignore missing data.
    c. Account for missing data.
        i. Unit missingness.
        ii. Item missingness.

3. **Practical considerations and warnings.**

Warning 1

- If your goal is just to have a “complete” data set for further analysis, **be very careful**!
  - After you construct this dataset, **nobody will know the difference between observed and imputed data**.

Warning 2

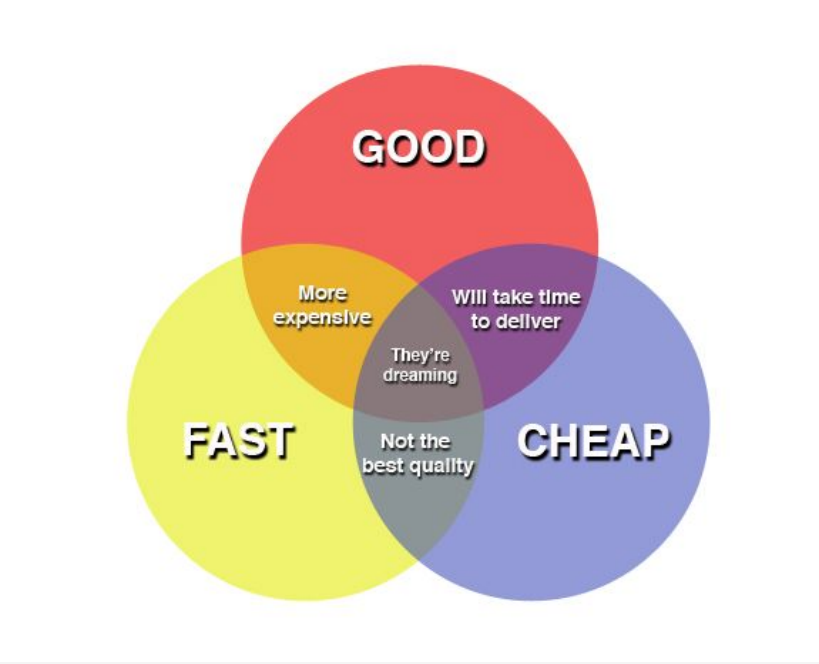
- **Improper imputation is making up data.**
  - This is an **unethical** thing to do.
  - Avoid doing this!

- **Proper imputation** is not making up data.
  - Proper imputation will impute data while properly estimating the amount of variance in your data.

Warning 3

- There are **`sklearn`** methods (**`SingleImputer`**, **`IterativeImputer`**) that are new-ish and that can be used.
- However, **`IterativeImputer`** was recently experimental!
- **Proceed with caution**: these may not work as expected**!**

What is a realistic approach for us?



Source: http://www.pyragraph.com/2013/05/good-fast-cheap-you-can-only-pick-two/

Fast and Cheap Analysis: Drop all missing values or single imputation.

Good and Cheap Analysis: Proper imputation or pattern submodel method.

Good and Fast Analysis: Gather data in a complete manner.

Thank you!

LinkedIn: Matthew Brems

Twitter: @MatthewBrems

Github: MatthewBrems

Email: matt@betavector.com