Data Wrangling and Data Analysis
Missing Data and Imputation

Daniel Oberski

Department of Methodology & Statistics

Utrecht University

VERSION: 2025-10-13

This week

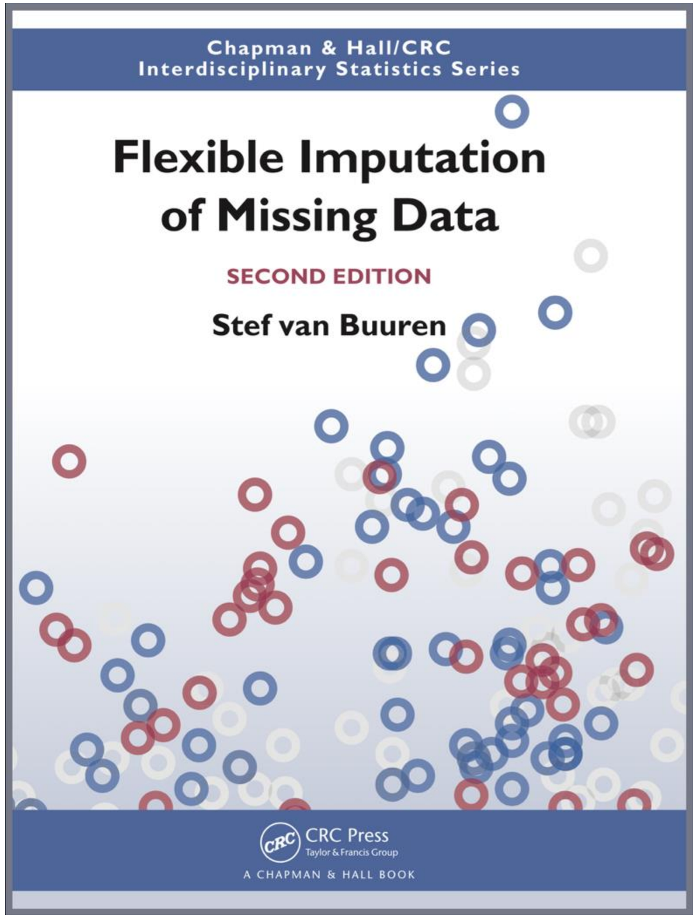
- **Lecture 1: Problems with missing data**
- Lecture 2: Solutions to problems with missing data
- Lecture 3: Clustering

Reading materials for this week

“Flexible Imputation of Missing Data”
https://stefvanbuuren.name/fimd

- Chapter 1, Sections 1.1, 1.2, 1.3, 1.4
- Optional:
  - Ch 3 (practical, recommended)
  - Ch 4 (practical, more technical)
  - Ch 2 and 6 (more theoretical)

Some of this week’s materials are adapted from Gerko Vink & Stef van Buuren’s courses on multiple imputation



Book recommendation

Dark Data: ***Why What You Don't Know Matters***
David J. Hand

A practical guide to making good decisions in a world of missing data

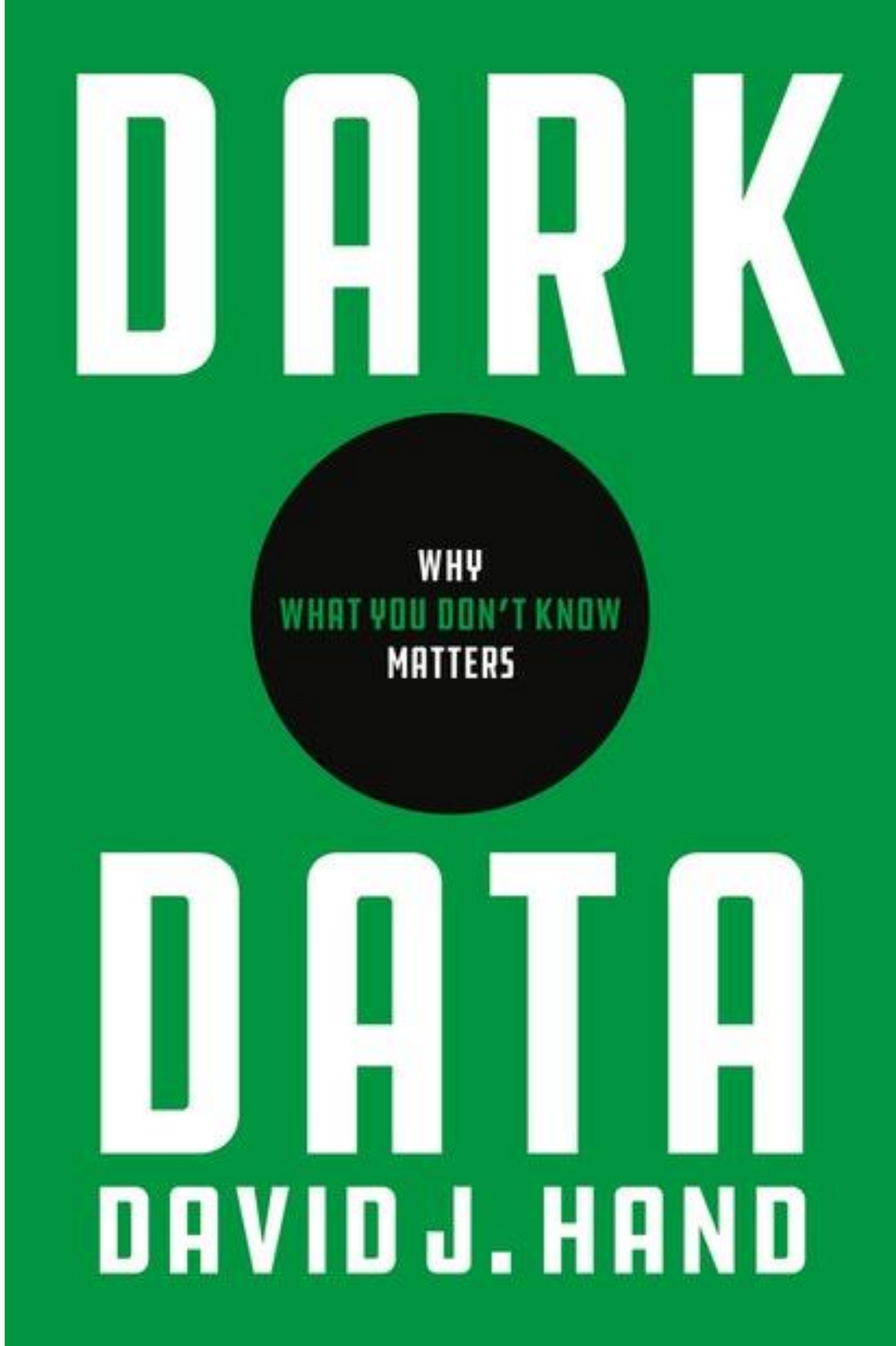


Image: [Sketchplanations]

"[missing data are] data you don't have – perhaps data you *wish* you had, or *hoped* to have, or *thought* you had, but nonetheless data you *don't* have. I argue, and illustrate with many examples, that **the missing data are at least as important as the data you do have**."

–DJH (emphases mine)

Why a week on missing data

- Missing data are everywhere
- Ad-hoc fixes do not (always) work
- A data scientist needs to understand when which methods do and do not work
- Goal of the week: get comfortable with solving missing data problems

Why is missing data important?

- “Obviously, the best way to treat missing data is not to have them.” (Orchard and Woodbury 1972)

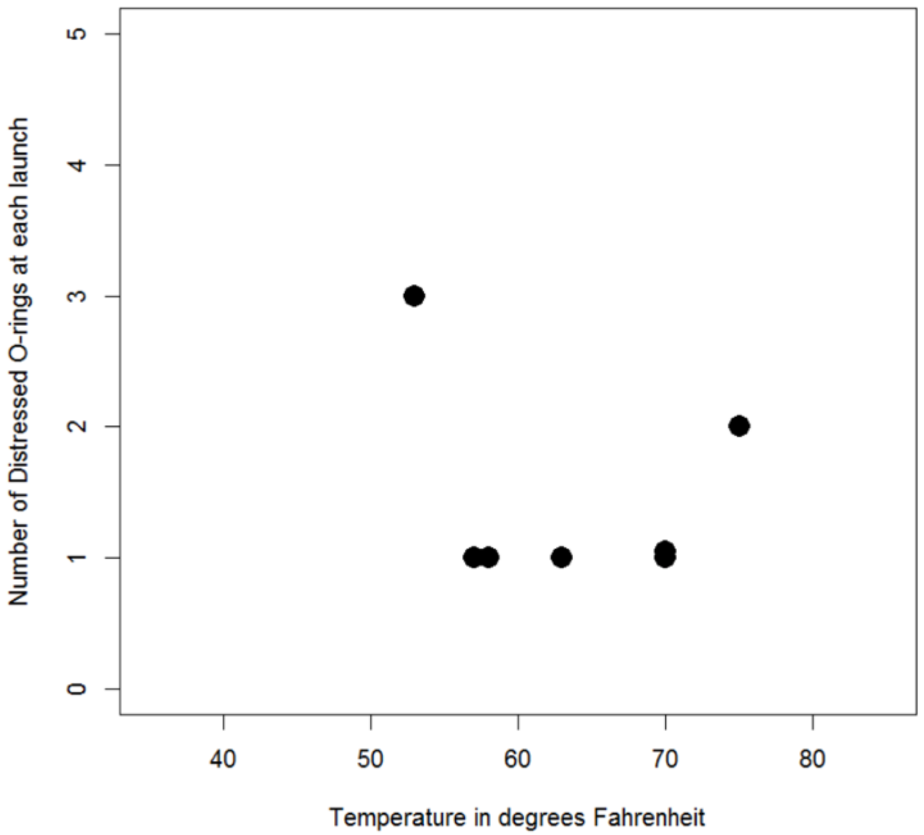
…but…

- “Sooner or later (usually sooner), anyone who does statistical analysis runs into problems with missing data” (Allison, 2002)
- Missing data problems are at the heart of data analysis

Example: Challenger (1986)

Figure 1.1 (a) Data examined in the pre-launch teleconference; **(b)** Complete data.

(a)



Why is missing data important?

1. **Just annoying**
   - Most procedures don’t deal with missings by default
2. **Less information than planned**:
   - Uncertainty of estimates (e.g. “standard errors”, “power”, “C.I”, etc.)
   - Accuracy of predictive models
3. **Systematic biases**:
   - Estimates of interest wrong on average
   - Prediction error seems better than it will be in reality

How to think about missing values

- Values do *exist*, but we are unable to *see* them
- Possible reasons:
  - Survey non-response
  - Person in medical study dies
  - Some users use high privacy settings in browser
  - Only high-income residents report potholes
  - Satellite experiences interference from sun
  - Etc.

Excel: Why using Microsoft's tool caused Covid-19 results to be lost

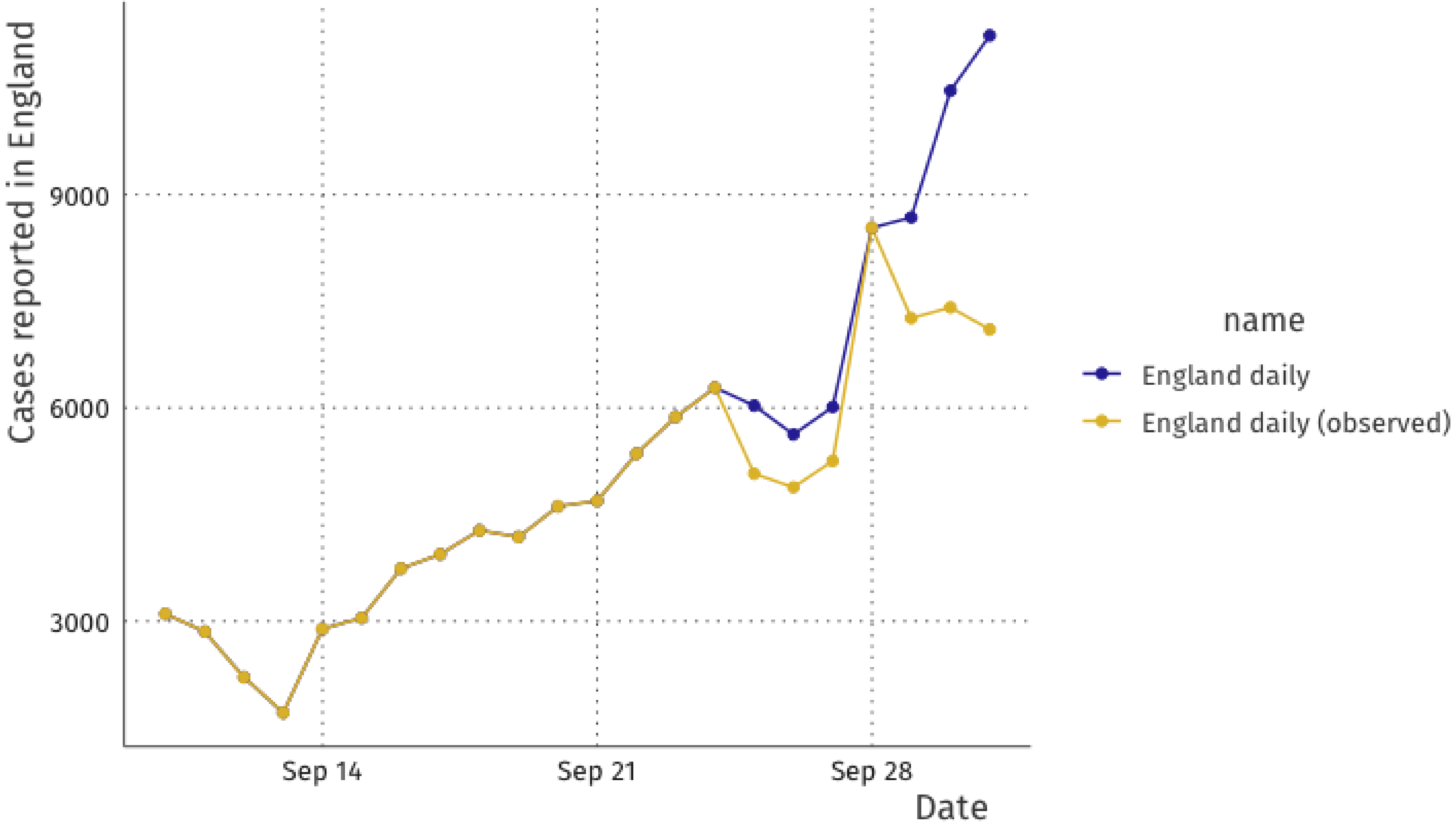
By Leo Kelion
Technology desk editor

5 days ago

Coronavirus pandemic

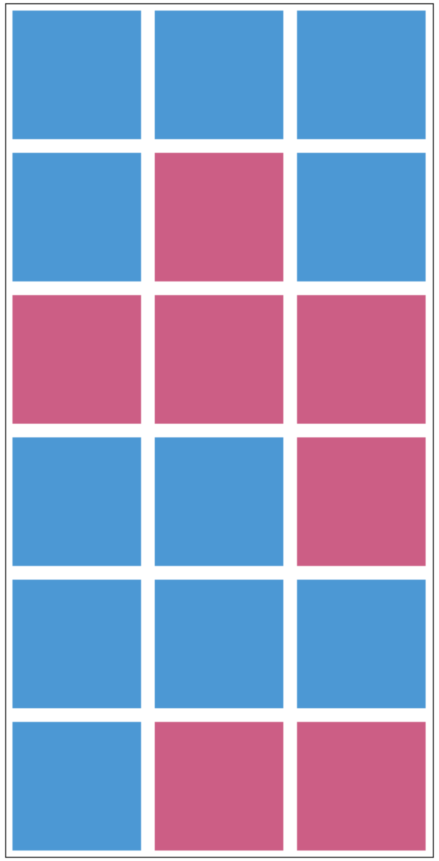
Date (recorded – flow though into following day's published numbers)	Expected reported date for GOV.UK	Cases that were not included on the expected data
24/09/2020	25/09/2020	957
25/09/2020	26/09/2020	744
26/09/2020	27/09/2020	757
27/09/2020	28/09/2020	0
28/09/2020	29/09/2020	1415
29/09/2020	30/09/2020	3049
30/09/2020	01/10/2020	4133
01/10/2020	02/10/2020	4786

Question: we would like to quantify the *trend in cases*. What is the effect of the Excel error on estimates of this trend?

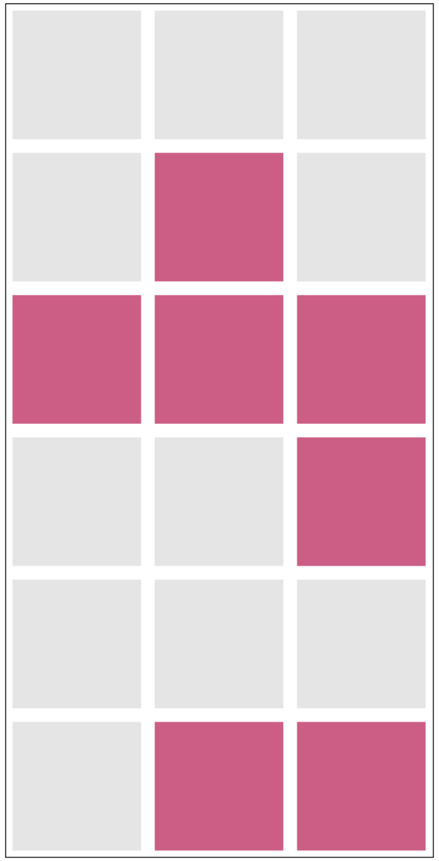


Cases reported in England
9000
6000
3000
Sep 14
Sep 21
Sep 28
Date
name
England daily
England daily (observed)

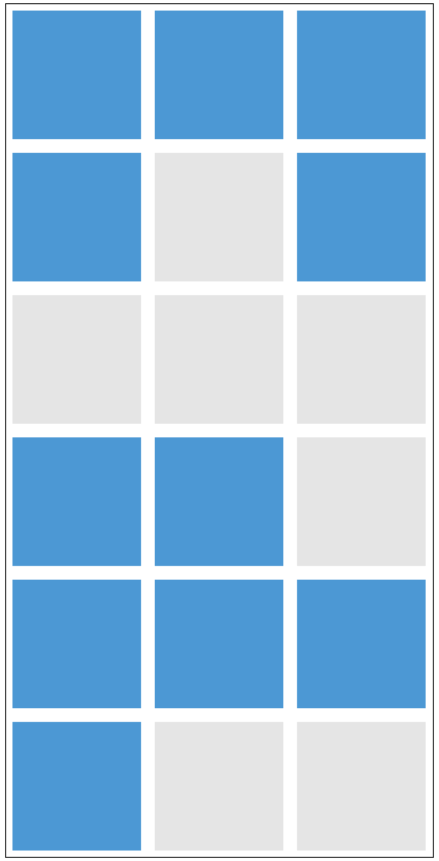
Complete data



Missing data



Observed data



How are missing data “coded”?

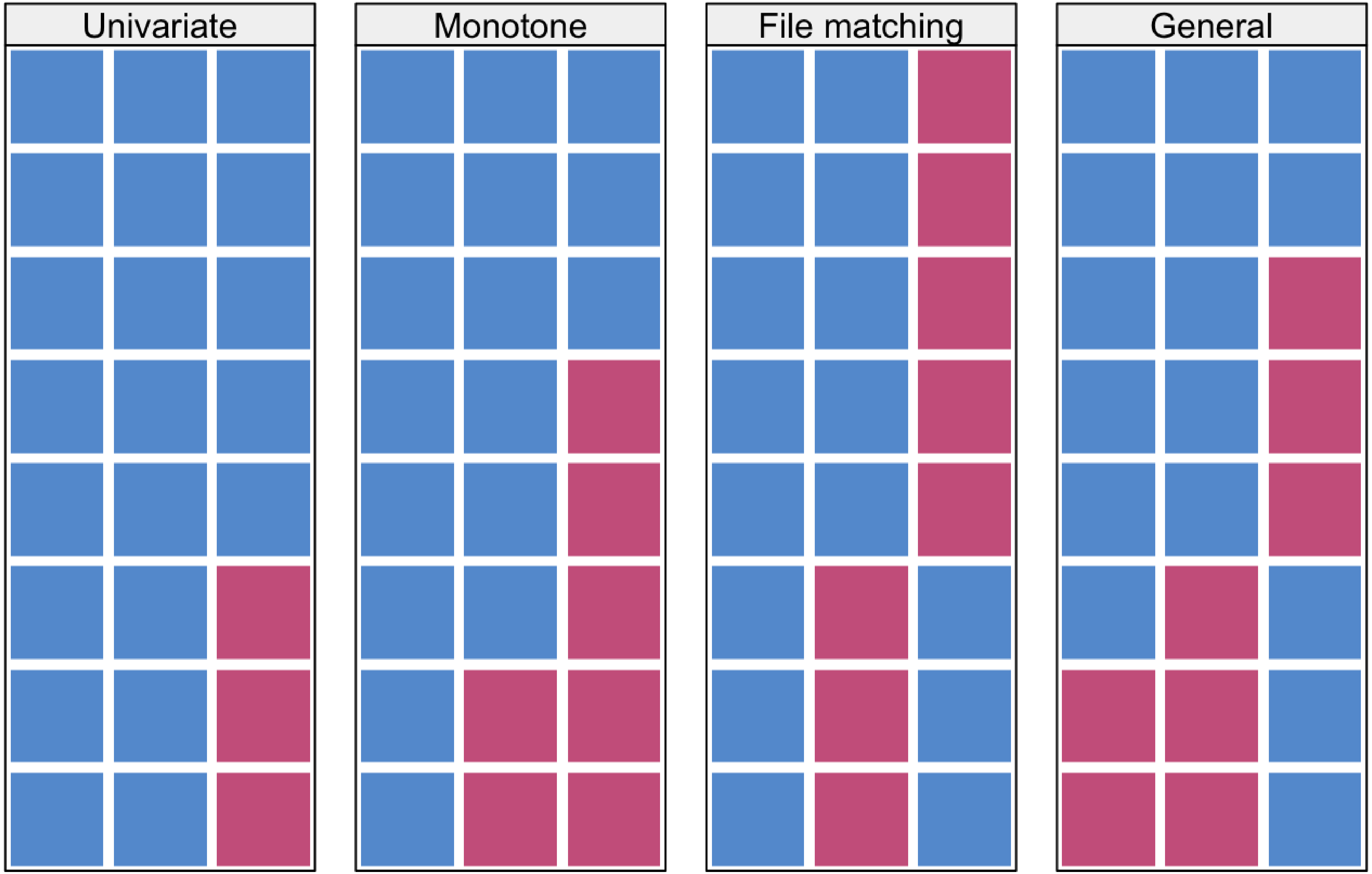
- `R` data frames: `NA`
- Python/Pandas: `NaN`
- SQL: `NULL`
- JavaScript/ json: `null`
- .csv files: `"NA"`, "", `"NULL"`
- SPSS: `-999, 99999,` … *sigh*.

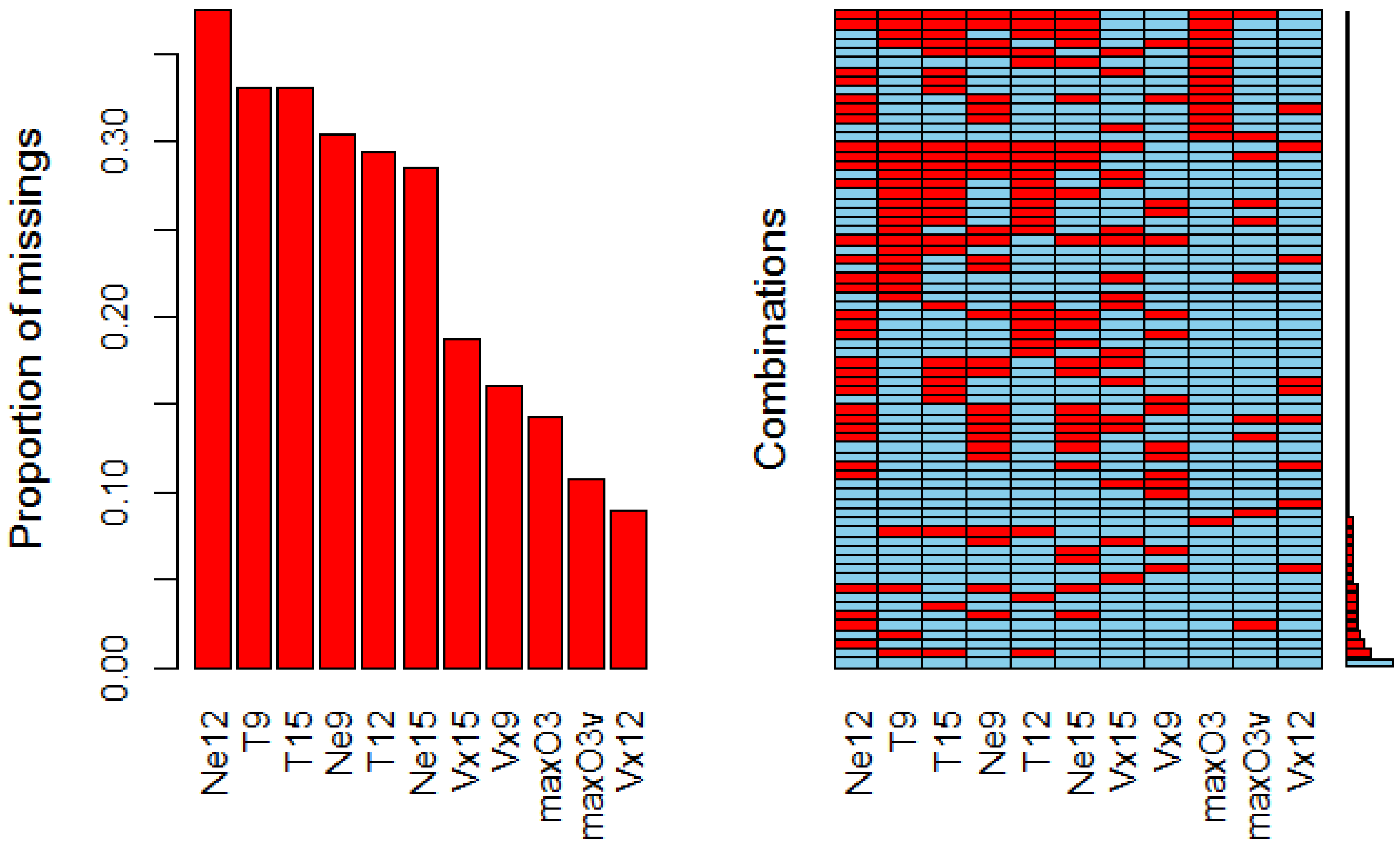
So pay attention during data wrangling!

Visualizing missing data

- It can be helpful to visualize the missing data
- Summarizes & provides insight into the amount and pattern of missingness
- In the assignment this afternoon you will work in R with missing data and create visualizations

- Visualization can help in understanding the missing data mechanism


Univariate
Monotone
File matching
General



Proportion of missings
0.00
0.10
0.20
0.30
Ne12
T9
T15
Ne9
T12
Ne15
Vx15
Vx9
maxO3
maxO3v
Vx12
Combinations
Ne12
T9
T15
Ne9
T12
Ne15
Vx15
Vx9
maxO3
maxO3v
Vx12

BMI example

- BMI measured at two time points
- **Research question**: does BMI decrease?

person_id	bmi_1	bmi_2
1	18	19
2	25	22
3	24	?
4	34	?
5	17	?

BMI example

- What is the average of `bmi_2`?
- What is the correlation with `bmi_1`?
- What is the average difference?

person_id	bmi_1	bmi_2
1	18	19
2	25	22
3	24	?
4	34	?
5	17	?

Different kinds of people

Suppose there are **three kinds of missing** people:

1. People who just moved away or dropped out for other reasons unrelated to the study;
2. People who not dramatically overweight at the start and who dropped out because of loss of motivation;
3. People who drop out because they failed to stick to the program and felt too embarrassed to return to the group.

Shining a light

*Use **what you do know** about **what you don't know***

NDD: It's missing for reasons unrelated to the data (MCAR)

Sickness preventing students from sitting exam

SDD: It's missing for reasons related to data you have got (MAR)

School discouraging less able students from sitting exam

UDD: missing because of the values you *would have* obtained (MNAR)

Student realised revised wrong material, so didn't turn up to sit exam

Source: David Hand

Different kinds of people

1. **NDD**: Missing reasons **unrelated** to the study:

$$\Pr(M = 1 \mid \texttt{bmi\_1}, \texttt{bmi\_2}) = \Pr(M = 1)\,,$$

   where M indicates whether `bmi_2` is missing (1) or not (0).

2. **SDD**: Missing completely **explained by low BMI at the start** (`bmi_1`):

$$\Pr(M = 1 \mid \texttt{bmi\_1}, \texttt{bmi\_2}) = \Pr(M = 1 \mid \texttt{bmi\_1})$$

3. **UDD**: Missing **depends on unseen values** themselves:

$$\Pr(M = 1 \mid \texttt{bmi\_1}, \texttt{bmi\_2}) \text{ can't be reduced any further.}$$

NDD, SDD, UDD

- **NDD**: The probability to be missing is **constant** for all units
- **SDD**: The probability to be missing **depends** on **observed** data
- **UDD**: The probability to be missing **depends** on ***un*observed** factors

MCAR, MAR, MNAR ????

MCAR: Missing Completely At Random (NDD)

MAR: Missing At Random (SDD)

MNAR: Missing Not At Random (UDD)

WARNING

- This nomenclature is due to Rubin and has confused and confounded many generations of students and practictioners.
- Hand’s (2020) terms refer to the exact same concepts, but have the advantage of being intelligible.
- *I advise you to switch to Hand’s terms.*

Example:

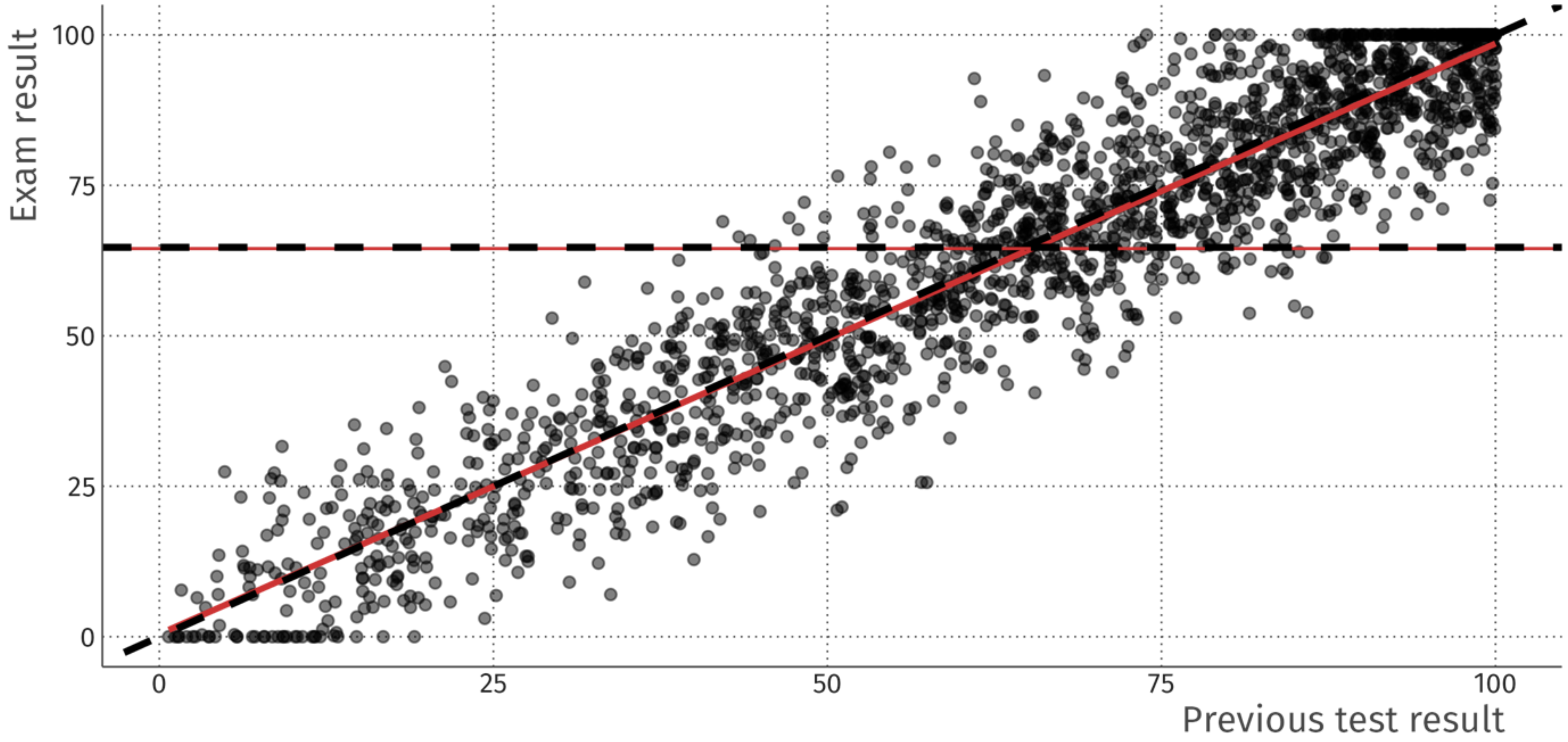
Estimate what would be the ***average final exam*** score for a school if every student had been examined

The data are:

- previous test results for all students
- final exam score for those who sat it

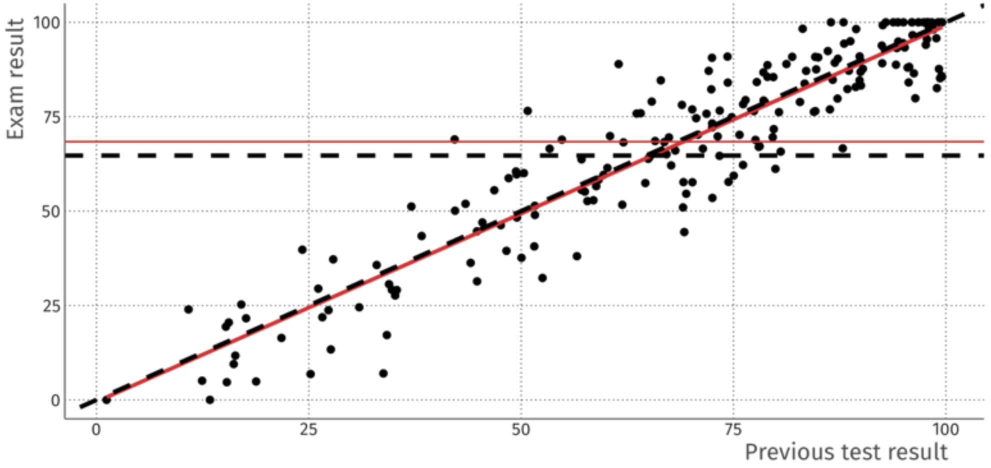
Source: David Hand

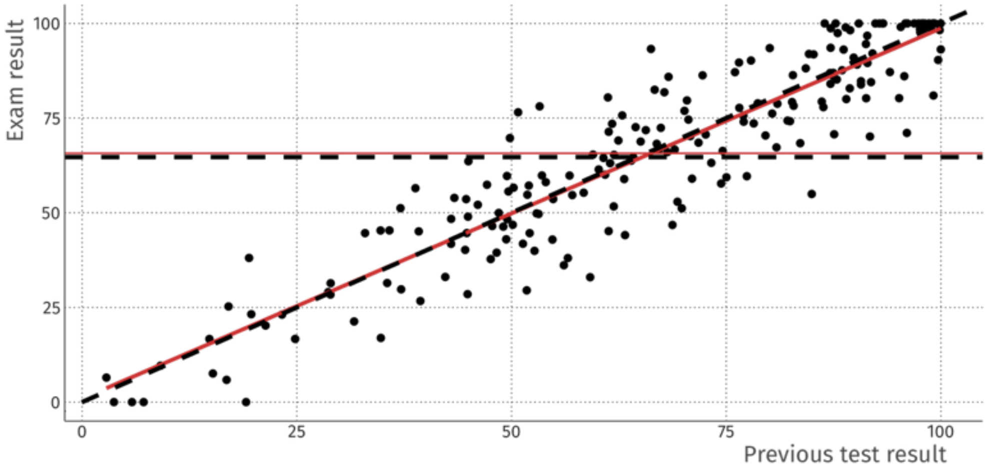
Complete data

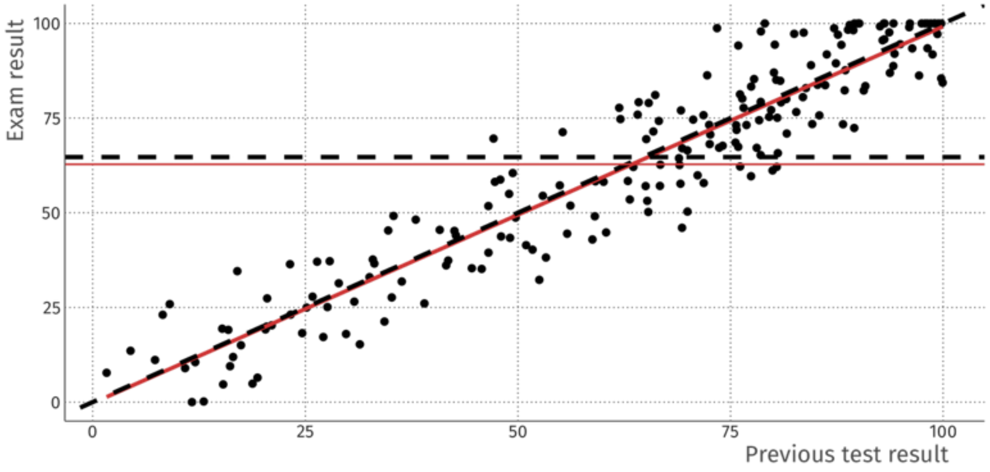


NDD Not Data-Dependent ("MCAR")

Sickness preventing students from sitting exam

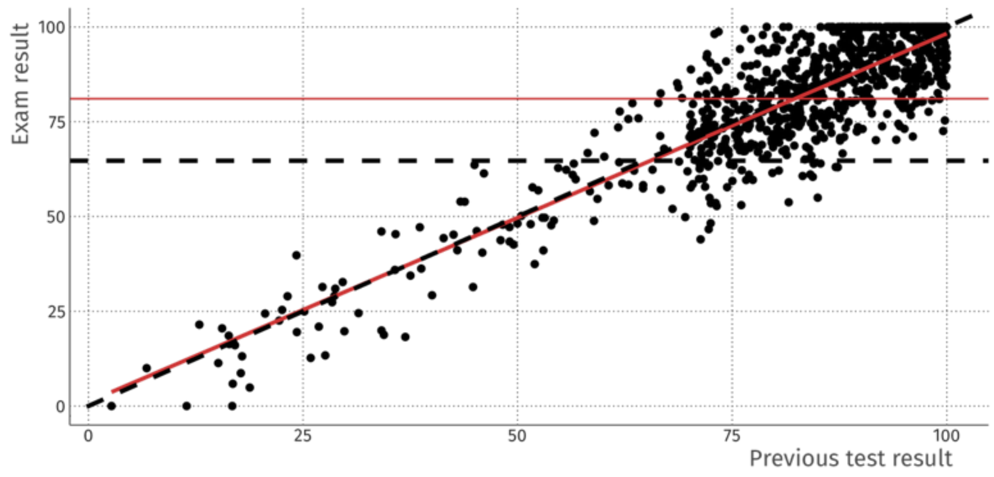


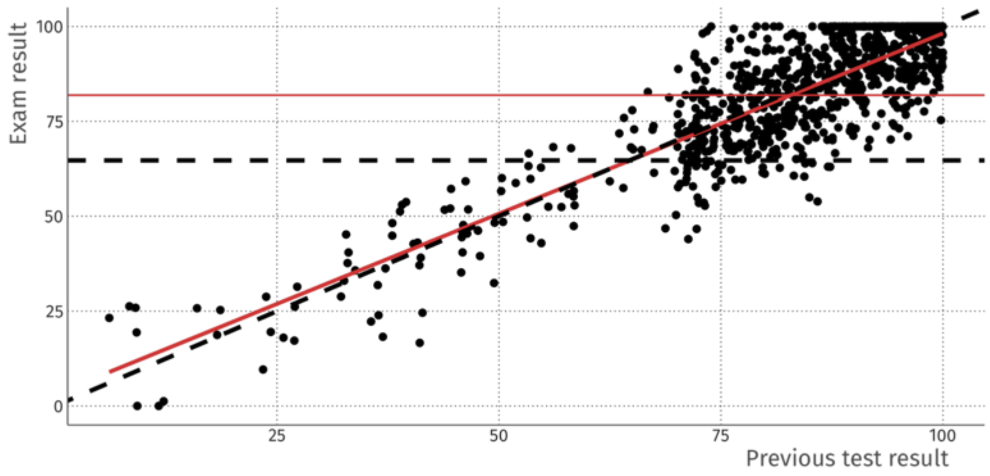


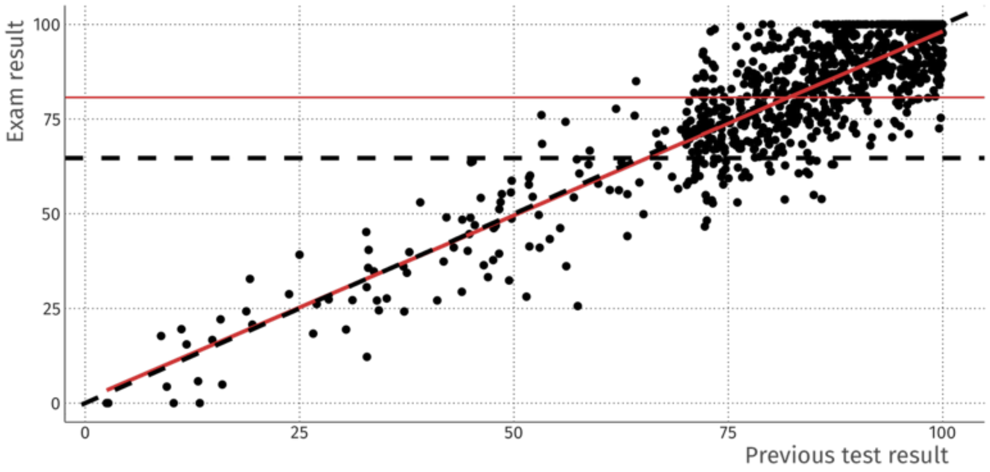


SDD *Seen* Data-Dependent ("MAR")

School discouraging students from sitting exam based on previous test

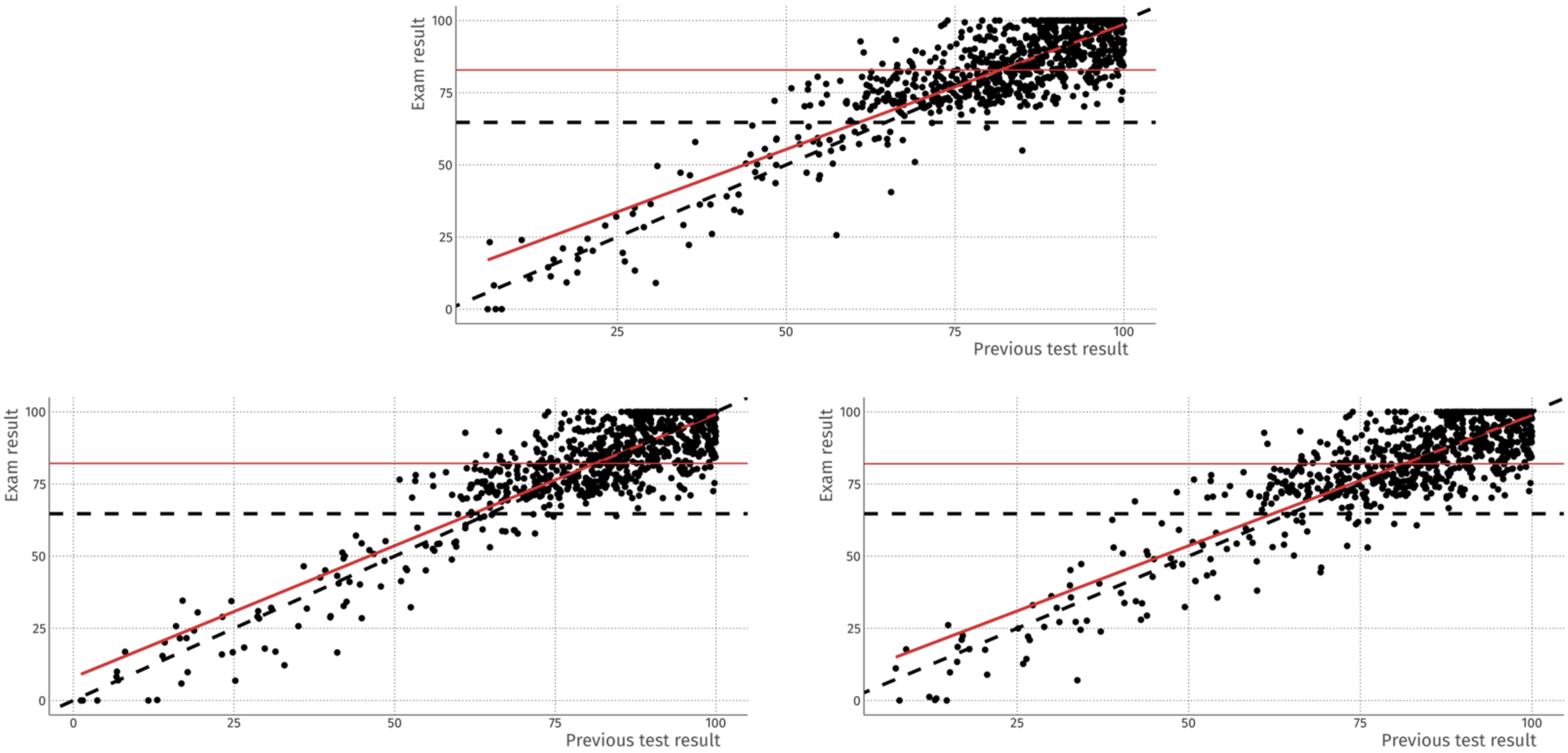






UDD *Unseen* Data-Dependent ("MNAR")

Students realised they revised wrong material, so didn't turn up to sit exam



Practice

Question: Is this (probably) NDD, SDD, or UDD:

1. We asked people’s income, and everybody answered. But then a server log file was accidentally deleted;
2. We asked people’s income, but the rich people did not want to say;
3. We asked people’s income, but those with a university education did not want to say. We only observe income.
4. Same as 3, but now we also observe education.

Questions

1. In practice, how can we tell whether we are in the **not**-data-dependent versus **seen**-data-dependent situation?

2. In practice, how can we tell whether we are in the **seen**-data dependent versus **unseen**-data-dependent situation?

Bias in machine learning due to missing data

Source: Shankar et al. (2017). No classification without representation: Assessing geodiversity issues in open data sets for the developing world.

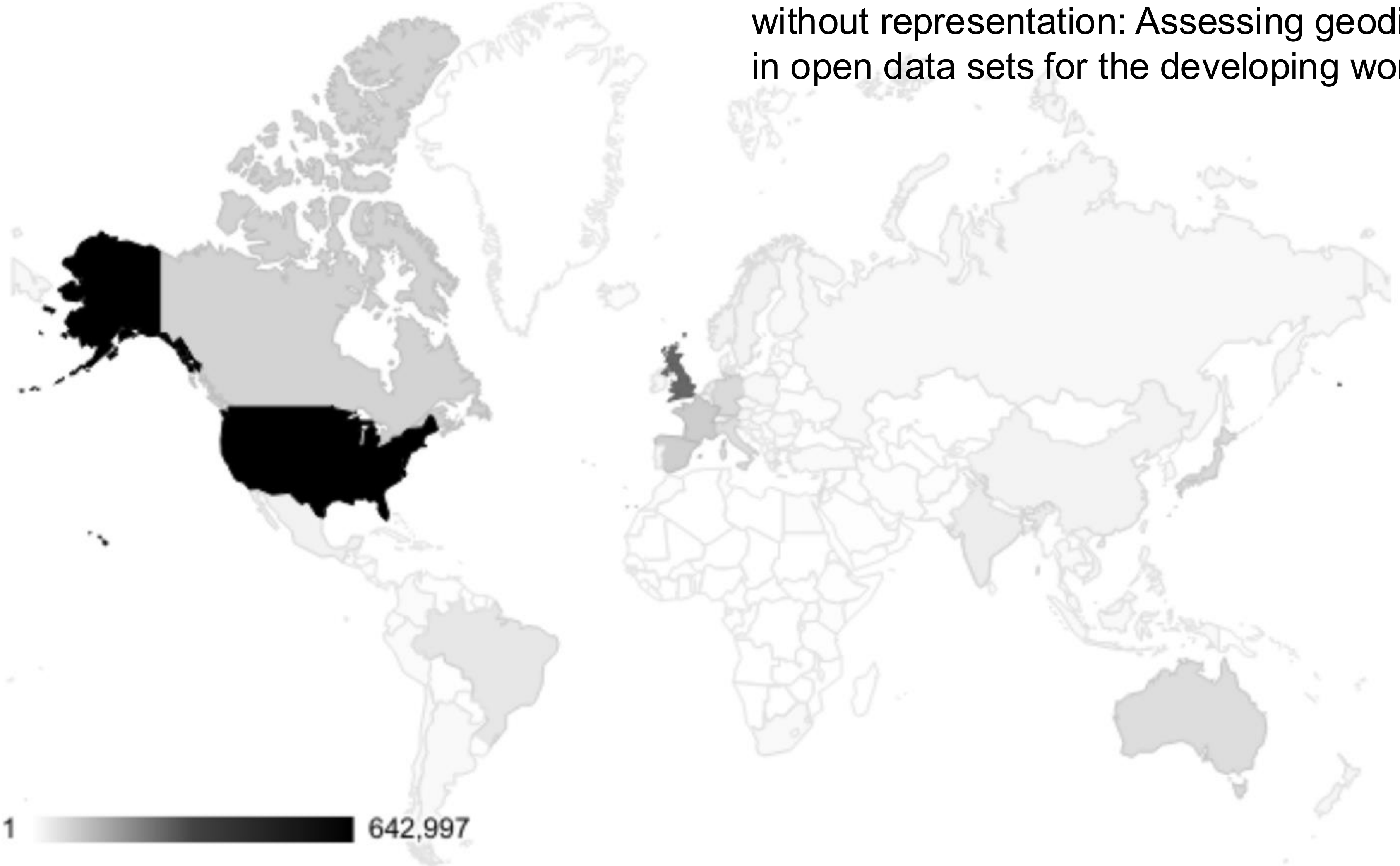
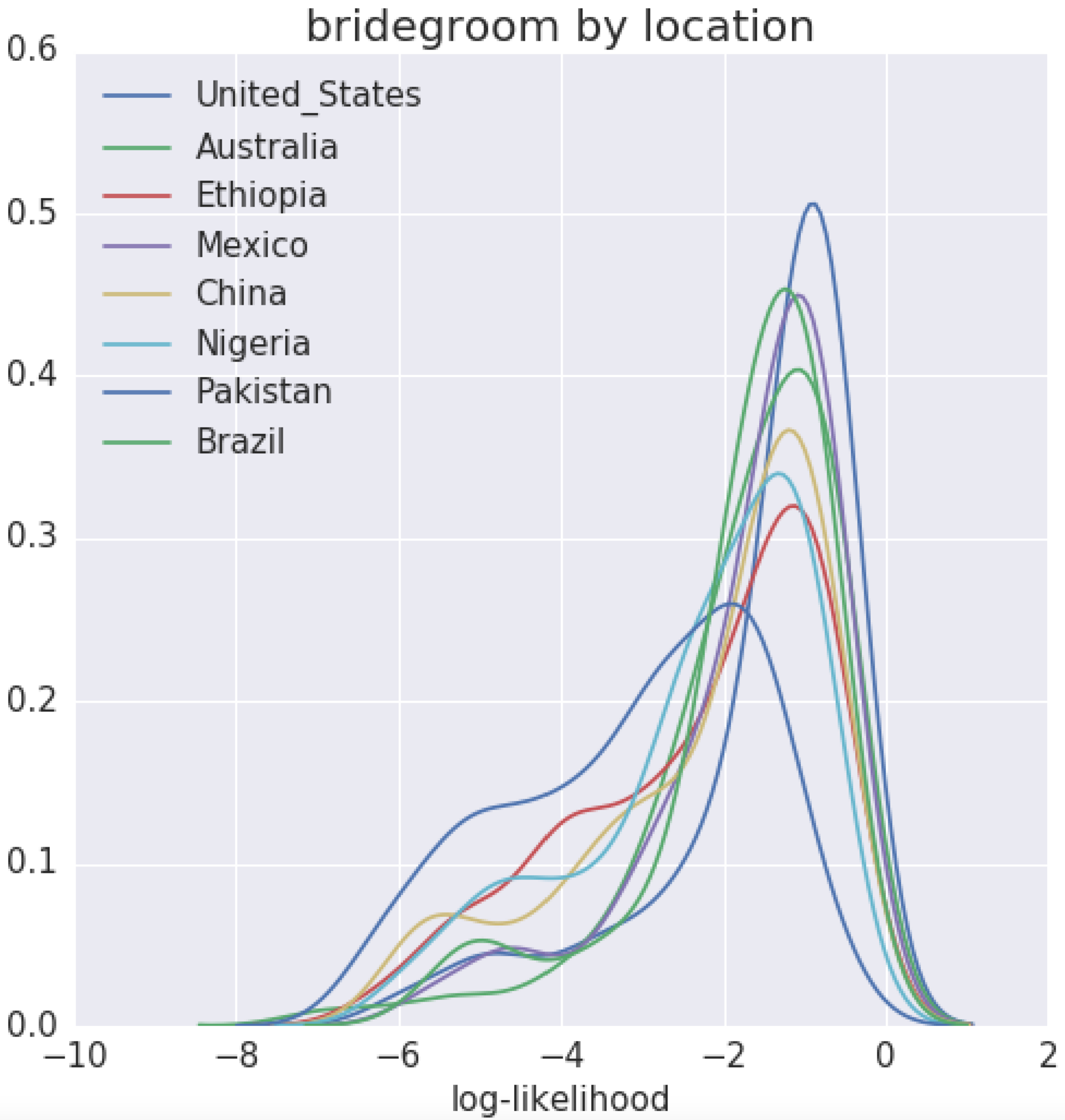


Figure 2: Distribution of the geographically identifiable images in the Open Images data set, by country. Almost a third of the data in our sample was US-based, and 60% of the data was from the six most represented countries across North America and Europe.



Source: Shankar et al. (2017).

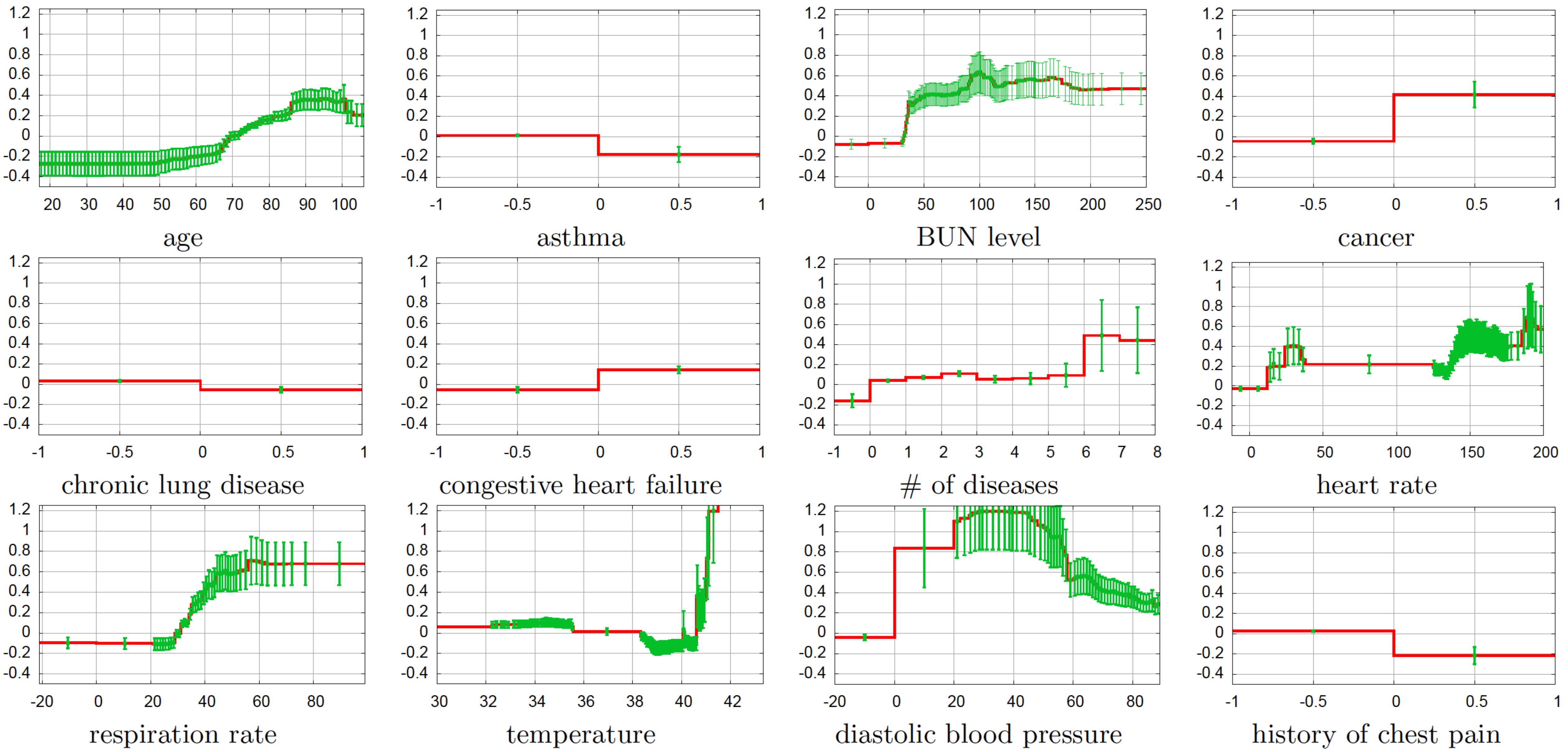
Predicting mortality in pneumonia patients coming into the ICU

Caruana et al. (2015), KDD

Target: mortality, i.e. death (1) or not (0)

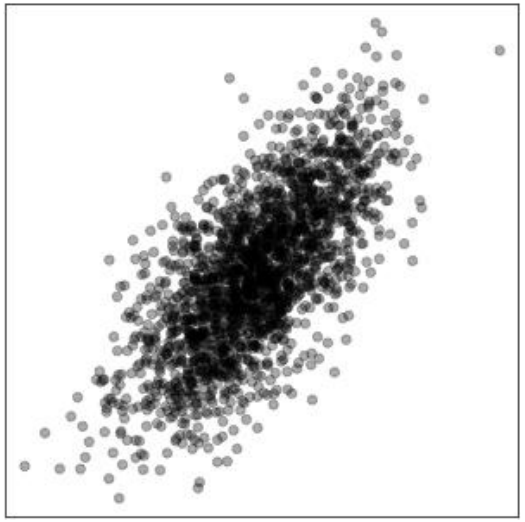
Data: patients at entry into ICU

Question: do you see anything surprising?



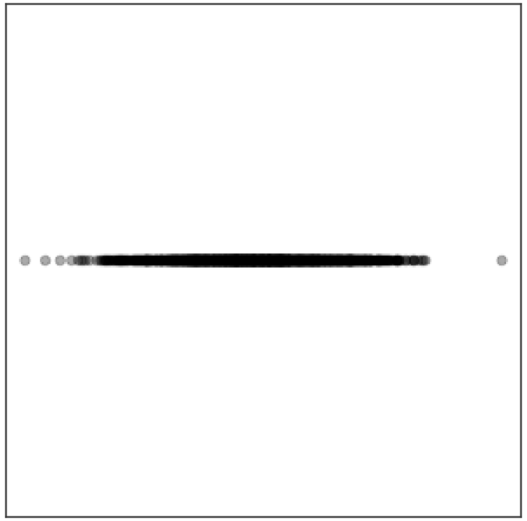
Why *mean imputation* is a bad idea

Before missingness



$$\mathrm{var}(y) = 2$$
$$\mathrm{MSE}(y|x) = 1$$

After mean imputation



1. What is the **cross-validated MSE** of the best model after mean imputation?
2. What is the **true MSE** of this model for new data?

Image: [Sketchplanations]

Conclusion

- Missing data are not just annoying, but can also **bias your analysis**
  - *Means, Trends, Covariances, Prediction models, …*
- We have seen that just ignoring (removing) missings doesn't usually solve the problem
- Mean imputation also does not usually solve the problem
- What we *can* do:

Use ***what you do know*** *about* ***what you don't know***

- This will be discussed tomorrow