Data: Context and Quality

Should we give Blase a loan?

Should we give Blase a loan?

Income	Default
33000	Y
42000	Y
55000	Y
37000	Y
103000	N
98000	N
70000	N
29000	Y
67000	N

Should we give Blase a loan?

Income	Default
33000	Y
42000	Y
55000	Y
37000	Y
103000	N
98000	N
70000	N
29000	Y
67000	N

Should we give Blase a loan?

Income	Default
33000	Y
42000	Y
55000	Y
37000	Y
103000	N
98000	N
70000	N
29000	Y
67000	N

Decision Method

```
if income >= 67000:
        grant loan
else:
        don't
```

Should we give Blase a loan?

Income	Default
33000	Y
42000	Y
55000	Y
37000	Y
103000	N
98000	N
70000	N
29000	Y
67000	N

family-income-2008.csv

Decision Method

```
if income >= 67000:
        grant loan
else:
        don't
```

Pitfalls of Data Repurposing

Pitfalls of Data Repurposing

- Data is continuously repurposed
  - That’s one of the reasons to keep accumulating it
    - ML Benchmarks
    - ML Training datasets
  - Improvements in technical capabilities means one can do more with data today than a few years ago.
- Beware what you are repurposing the data for though!
  - Why was that dataset created? What was its purpose?
- *How can we avoid these problems?*

Data Quality

- **Context**
  - Documentation
  - Provenance
  - Data formats
  - Assumptions
- Content
  - Errors
  - Missing Data
    - MAR, MNAR, MCAR
    - Disguised Missing Values

Why we need context

- Data acquisition (hunters/gatherers)
- Data stewards
- Data owners
- Data engineers
- Data analysts
- Data consumers

With multiple people involved in the process of transforming raw data into insights, some assumptions with downstream impact may end up buried in the complexity of organizations.
Context is important!

Documentation, Context, Semantics

- Provenance/Lineage
  - How was this data recorded/obtained/acquired/produced?
- Metadata
  - Is there documentation associated to the data? What do the attributes mean? What units do they use?
  - If not, find out that information before using the data. Then, include the assumptions you had to make
- E.g., Datasheets for Datasets

Common Metadata Questions

"We have extracted 155 DGIC questions data workers often face that would be addressed with access to the right MIs. These questions illustrate the metadata landscape we consider in this paper. We have synthesized the 155 questions into 27"

	Representatives of Common Data Questions	DGIC Category	5W1H+R Category
Q1	For what purpose was the dataset created?	G [21, 49]	Why
Q2	Are there tasks for which the dataset should not be used?	G,C [21]	Why
Q3	Who created the dataset?	G,C [21, 26]	Who
Q4	Who was involved in the data creation process?	G,C [21]	Who
Q5	How can the owner/curator/manager of the dataset be contacted?	G [21]	Who
Q6	What are the privacy and legal constraints on the accessibility of the dataset?	C [38]	Who
Q7	Is there an access control list for the dataset?	G,D [26]	Who
Q8	What is the reputation of the creator of a dataset?	G [24]	Who
Q9	What do the instances of the dataset represent?	D,G,I [21]	What
Q10	What is the size of the dataset?	D,G,I [26]	What
Q11	Are there errors in the dataset?	D,G,I [21, 24, 38]	What
Q12	Does the dataset have missing values?	D,G,I [24]	What
Q13	What is the domain of the values in this dataset?	D,G,I [30]	What
Q14	If the dataset is a sample of a larger dataset, what was the sampling strategy?	G,I [21]	How
Q15	Does the dataset contain personally identifiable information (PII)?	G,C [4, 49]	What
Q16	What is the quality of the dataset?	G [3, 4, 13, 39]	What
Q17	Was any preprocessing/cleaning/labeling of the dataset done?	G [21]	How
Q18	Was data collection randomized? Could it be biased in any way?	G [38]	How
Q19	Is there anything about dataset preprocessing/cleaning that could impact future uses?	G [21]	How
Q20	What is the dataset's release date?	D,G,I [30]	When
Q21	Is there an expiration date for this dataset?	D,G [3]	When
Q22	How often will the dataset be updated?	G,I [21]	When
Q23	When was the data last modified?	D,G,I [26]	When
Q24	How easy is it to download and explore this dataset?	D [24]	Where
Q25	What is the format of the dataset, and what type of repository is the dataset located in?	D [38]	Where
Q26	What is the provenance of this dataset?	I [54]	Relationship
Q27	What other datasets exist in this repository that are related to this dataset?	D,G,I [52]	Relationship

Metadata Management and Catalogs

- Data Catalog:
  - A database for metadata
  - Centralizes tribal knowledge
  - Many challenges to make this work well
- Cultural and socio-technical as much as technical problem
  - Incentives to get people to insert metadata in the catalogs
  - Documenting datasets is not in their critical path
    - Except in regulated industries, or domains with strong auditors

Berkeley's Ground [28]
Microsoft Azure Data Catalog [31]
Apache Atlas [3]
Denodo platform [17]
SAP Data Intelligence platform [46]
Boomi Data platform [7]
WeWork's Marquez [50]
Lyft's Amundsen [41]
Linkedin's Datahub [37]

Metadata in the Sciences

- The FAIR principles
- Repositories for disseminating research data
  - With the appropriate context
  - ICPSR
  - Dataverse
- Still lots of work to do!

Open Access | Published: 15 March 2016

The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier, [...]

Scientific Data **3**, Article number: 160018 (2016) | Cite this article

An Addendum to this article was published on 19 March 2019

Abstract

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measureable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. This Comment is the first formal publication of the FAIR Principles, and includes the rationale behind them, and some exemplar implementations in the community.

Findability
Accessibility
Interoperability
Reusability

Data Quality

- Context
  - Documentation
  - Provenance
  - Data formats
  - Assumptions
- **Content**
  - Errors
  - Missing Data
    - MAR, MNAR, MCAR
    - Disguised Missing Values

Types of Data Errors 1/2

- Outliers
  - Values that deviate from the distribution (statistical sense)
  - 2, 3, 4, 5654545, 3, 2
- Duplicates
  - Distinct records that refer to same real-world entity
  - E.g., (first name, last name), (last name, first name)
- Rule Violations
  - Records that violate *integrity constraints*: not null, uniqueness, etc.
- Pattern Violations
  - Violate syntactic and semantic constraints: alignment, misspelling, semantic data types, etc.
    - Zip code -> State

Types of Errors 2/2

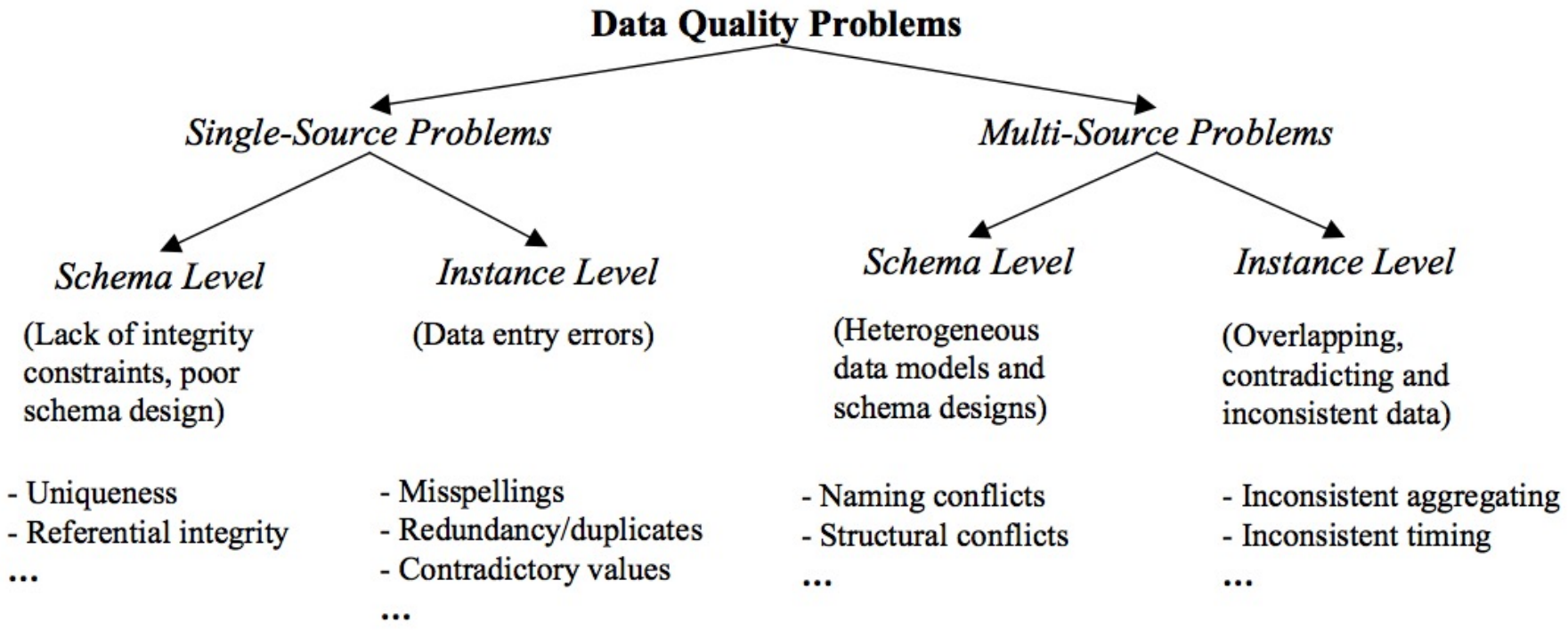


Figure 2. Classification of data quality problems in data sources

Erhard Rahm, Hong Hai Do: Data Cleaning: Problems and Current Approaches. IEEE Data Engineering Bulletin (23): 3-13 (2000)

How do errors affect analysis?

Dep. Name	Num Employees
Computer Science	41
Economics	112
Statistics	26
CS	41
Physics	33
Chemistry	31

- Duplicates and Outliers affect descriptive stats / aggr. functions
- Outliers are not always errors
  - They may indicate different measurement standards/methods
- Error or not?
  - Sometimes, this depends on what we are trying to achieve

The Art of Data Cleaning

As a large mass of raw information, Big Data is not self-explanatory. And yet the specific methodologies for interpreting the data are open to all sorts of philosophical debate. Can the data represent an 'objective truth' or is any interpretation necessarily biased by some subjective filter or the way that data is 'cleaned?'
The Promise and Peril of Big Data. Bollier, 2010, p. 13

Tooling

- OpenRefine
- Ad-hoc tools
- Most data cleaning remains in ad-hoc scripts prepared by data engineers and stewards
  - This is at odds with good documentation of datasets

- Trifacta, Tamr, and lots of other commercial offerings
  - Lots of VC money flowing into data quality

Data Quality

- Context
  - Documentation
  - Provenance
  - Data formats
  - Assumptions
- Content
  - Errors
  - **Missing Data**
    - MAR, MNAR, MCAR
    - Disguised Missing Values

Missing Data

- NULL values on data
  - Represented in many ways
    - e.g., NULL, null, “NULL”, “”, 0, -1, No, “nil”, , Nope, etc…
- NULL values on collection or because of data cleaning
  - How do people ‘repair’ dirty data?
    - A default strategy is to set the value = NULL
- Disguised missing values
  - These are the nastiest of all
- How do we handle these?
  - Common approach: Drop rows and hope for the best!

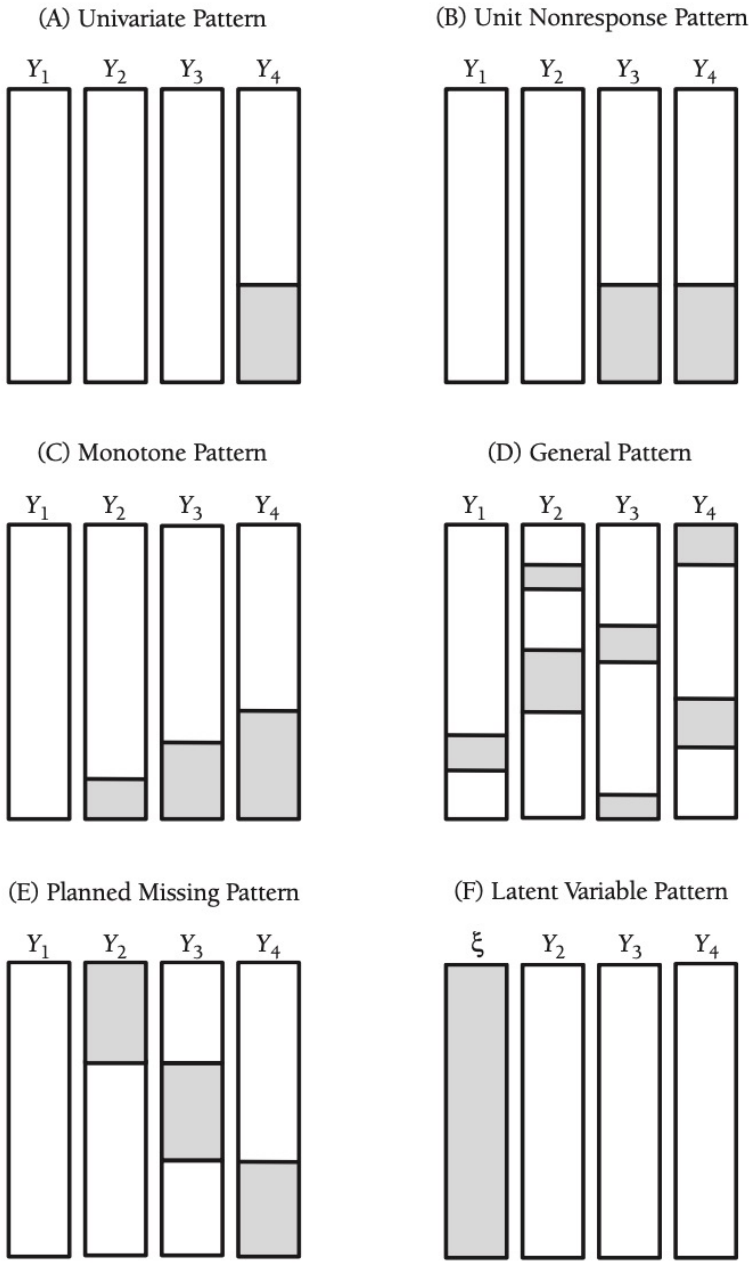
“…for most of our scientific history, we have approached missing data much like a doctor from the ancient world might use bloodletting to cure disease or amputation to stem infection (e.g, removing the infected parts of one’s data by using list-wise or pair-wise deletion). My metaphor should make you feel a bit squeamish, just as you should feel if you deal with missing data using the antediluvian and ill-advised approaches of old.“ Todd Little. Preface to Applied Missing Data Analysis, Craig Enders.

Missing Data. Some concepts

- Missing Data Patterns
  - What data is missing, e.g., what cells in a table
- Missing Data Mechanisms
  - Aims to find relationships between observed variables and missing data (not necessarily explain why data is missing though because correlation != causation)

Missing Data Patterns

- Patterns: location of missing values
- Does not explain why data is missing
- Certain patterns are associated with reasons
  - c) E.g., attrition rate in multi-phase study



Missing Data Mechanisms

- Missing at Random (MAR)
- Missing Completely at Random (MCAR)
- Missing Not at Random (MNAR)

Running example: hiring

- Consider a company's hiring procedure consists of two stages:
  - IQ test to determine who to hire
  - Followed by a job performance review 6 months in by a manager

TABLE 1.2. Job Performance Ratings with MCAR, MAR, and MNAR Missing Values

	Job performance ratings
IQ	MAR
78	—
84	—
84	—
85	—
87	—
91	7
92	9
94	9
94	11
96	7
99	7
105	10
105	11
106	15
108	10
112	10
113	12
115	14
118	16
134	12

From “Applied Missing Data Analysis”. Craig Enders

- Scenario 1
  - Why are those values be missing?

TABLE 1.2. Job Performance Ratings with MCAR, MAR, and MNAR Missing Values

	Job performance ratings
IQ	MAR
78	—
84	—
84	—
85	—
87	—
91	7
92	9
94	9
94	11
96	7
99	7
105	10
105	11
106	15
108	10
112	10
113	12
115	14
118	16
134	12

From “Applied Missing Data Analysis”. Craig Enders

- Scenario 1
  - Why are those values be missing?

- **Missing at Random (MAR)**. Probability of missing data in attribute X depends on some other attribute, Y, but not the values of X.

TABLE 1.2. Job Performance Ratings with MCAR, MAR, and MNAR Missing Values

	Job performance ratings
IQ	Complete
78	9
84	13
84	10
85	8
87	7
91	7
92	9
94	9
94	11
96	7
99	7
105	10
105	11
106	15
108	10
112	10
113	12
115	14
118	16
134	12

Ideal Scenario; ground truth

TABLE 1.2. Job Performance Ratings with MCAR, MAR, and MNAR Missing Values

	Job performance ratings	
IQ	Complete	MCAR
78	9	—
84	13	13
84	10	—
85	8	8
87	7	7
91	7	7
92	9	9
94	9	9
94	11	11
96	7	—
99	7	7
105	10	10
105	11	11
106	15	15
108	10	10
112	10	—
113	12	12
115	14	14
118	16	16
134	12	—

From “Applied Missing Data Analysis”. Craig Enders

- Scenario 2
  - Why are those values be missing?

TABLE 1.2. Job Performance Ratings with MCAR, MAR, and MNAR Missing Values

	Job performance ratings	
IQ	Complete	MCAR
78	9	—
84	13	13
84	10	—
85	8	8
87	7	7
91	7	7
92	9	9
94	9	9
94	11	11
96	7	—
99	7	7
105	10	10
105	11	11
106	15	15
108	10	10
112	10	—
113	12	12
115	14	14
118	16	16
134	12	—

From “Applied Missing Data Analysis”. Craig Enders

- Scenario 2
  - Why are those values be missing?

- **Missing Completely at Random (MCAR).** Probability of missing data in X is unrelated to values of X and unrelated to other attributes.

TABLE 1.2. Job Performance Ratings with MCAR, MAR, and MNAR Missing Values

	Job performance ratings	
IQ	Complete	MNAR
78	9	9
84	13	13
84	10	10
85	8	—
87	7	—
91	7	—
92	9	9
94	9	9
94	11	11
96	7	—
99	7	—
105	10	10
105	11	11
106	15	15
108	10	10
112	10	10
113	12	12
115	14	14
118	16	16
134	12	12

From “Applied Missing Data Analysis”. Craig Enders

- Scenario 3
  - Why are those values be missing?

TABLE 1.2. Job Performance Ratings with MCAR, MAR, and MNAR Missing Values

	Job performance ratings	
IQ	Complete	MNAR
78	9	9
84	13	13
84	10	10
85	8	—
87	7	—
91	7	—
92	9	9
94	9	9
94	11	11
96	7	—
99	7	—
105	10	10
105	11	11
106	15	15
108	10	10
112	10	10
113	12	12
115	14	14
118	16	16
134	12	12

From “Applied Missing Data Analysis”. Craig Enders

- Scenario 3
  - Why are those values be missing?

- **Missing Not at Random (MNAR).** Probability of missing data in attribute X is related to the values of X.

TABLE 1.2. Job Performance Ratings with MCAR, MAR, and MNAR Missing Values

	Job performance ratings			
IQ	Complete	MCAR	MAR	MNAR
78	9	—	—	9
84	13	13	—	13
84	10	—	—	10
85	8	8	—	—
87	7	7	—	—
91	7	7	7	—
92	9	9	9	9
94	9	9	9	9
94	11	11	11	11
96	7	—	7	—
99	7	7	7	—
105	10	10	10	10
105	11	11	11	11
106	15	15	15	15
108	10	10	10	10
112	10	—	10	10
113	12	12	12	12
115	14	14	14	14
118	16	16	16	16
134	12	—	12	12

From “Applied Missing Data Analysis”. Craig Enders

- **Missing at Random (MAR).** Probability of missing data in attribute X depends on some other attribute, Y, but not the values of X.
- **Missing Completely at Random (MCAR).** Probability of missing data in X is unrelated to values of X and unrelated to other attributes.
- **Missing Not at Random (MNAR).** Probability of missing data in attribute X is related to the values of X.

Missing at Random (MAR)

- Probability of missing data in attribute X is related to other measured variable, Y, but not to the values of X
- Con: Impossible to verify that the relationship depends on the observed data only (and not other variables) without access to the missing values

TABLE 1.2. Job Performance Ratings with MCAR, MAR, and MNAR Missing Values

	Job performance ratings
IQ	MAR
78	—
84	—
84	—
85	—
87	—
91	7
92	9
94	9
94	11
96	7
99	7
105	10
105	11
106	15
108	10
112	10
113	12
115	14
118	16
134	12

From “Applied Missing Data Analysis”. Craig Enders

- Imagine we did not know the underlying hiring scenario
  - How can we verify if the missing values depend on IQ?
  - Consider another Boolean variable, hired?
    - And imagine that other factors than IQ play a role in the hiring decision

Missing Completely at Random (MCAR)

- Probability of missing data in attribute X is unrelated to other variables and unrelated to values of X.
- The observed data is a random sample of the ground truth data.
- In principle, it's possible to verify data is MCAR by comparing descriptive statistics for missing data group vs non-missing data group
  - E.g., comparing means: they should be similar.

Missing not at Random Data (MNAR)

- Probability of missing data in attribute X is related to the values of X
- Con: No way to verify data is MNAR without access to the missing values

Handling Missing Data

- Drop rows
- Fill in the blanks with the mean
- Maximum Likelihood Estimation
- Multiple Imputation

Handling Missing Data: Deletion

- Remove tuples that have at least one missing value

Handling Missing Data: Deletion

- Remove tuples that have at least one missing value
  - Assumes MCAR. Otherwise this will bias the data!
    - MCAR is a simple random sample of the data
  - May reduce sample size a lot!
    - E.g., 3% missing values, sprinkled throughout many rows
- Widespread use because it's very easy to implement
  - Lots of software packages include a 'drop_null' function
    - Pandas documentation on 'Working with missing data'
- Works with any kind of missing data

Handling Missing Data: Imputation

- Generates 1 value for each missing data point
  - Yields a complete dataset (unlike deletion methods)
- They produce biased datasets (sometimes even when data is MCAR)
- Arithmetic Mean Imputation/Mean substitution:
  - Reduces the variability of the data -> attenuates standard error/deviation
- Regression Imputation:
  - Use a regression line fit using some other (correlated) variable
  - Overestimates correlations
- Stochastic Regression Imputation: Augments regression imputation with a normally distributed residual term, i.e., adds normal noise.
  - Gives unbiased parameter estimates under MAR
- Requires numerical data
  - More advanced techniques for filling categorical data (augmentation, enrichment techniques)

Handling Missing Data: Other techniques

- Hot-Deck Imputation: Fill in missing value with a non-missing value
  - Variation: cluster other observations based on variables first.
  - Think about the many assumptions this method is making!
- Many other methods:
  - Similar response pattern imputation (similar to hot deck)
  - Averaging available items
  - Last observation carried forward

SOTA Missing Data Handling

- All previous methods assume MCAR and will bias data when data is MAR or MNAR (sometimes even when it’s MCAR).
- **Maximum Likelihood Estimation** (MLE) and **multiple imputation** produce unbiased estimates with MCAR and MAR data but not with MNAR.
- But we cannot test for MAR! Then why is this useful?

SOTA Missing Data Handling

- All previous methods assume MCAR and will bias data when data is MAR or MNAR (sometimes even when it's MCAR).
- **Maximum Likelihood Estimation** (MLE) and **multiple imputation** produce unbiased estimates with MCAR and MAR data but not with MNAR.
- But we cannot test for MAR! Then why is this useful?
  - Sometimes data is missed on purpose (planned missing data designs) in order to make its collection cheaper. Sometimes, in these cases one can assume MAR and then those methods are useful.
    - E.g., Questionnaire burden, time, money

Disguised Missing Values

- Phone number:
  - (999)999999
- Email address:
  - nope@nope.com
- Age:
  - 666

Source	Table	Column	DMVs
UCI ML	Diabetes	Blood Pressurse	0
	adult	workclass	?
		education	Some College
U.S. FDA	Even Reports	EVENT_DT	20010101, 20030101
data.gov	Vendor Location	Ref_ID	-1
data.gov	Graduation	Regents Num	s, -
data.gov.uk	Accidents 2015	Junction Control	-1

Table 1: Sample DMVs.

From “FAHES: A Disguised Missing Value Detector.” KDD 2018

- Practically, these are missing values.
  - But much harder to identify

Summary – Take away message

- Beware data repurposing
  - Always aim to understand well the data before doing anything with it
- Data Errors / Data Cleaning
  - Many types of errors, *many tools*. Understand their assumptions
- Missing data
  - Careful when handling missing data
    - Dropping values can easily bias your sample
  - Disguised missing values
- Data: Context and Quality