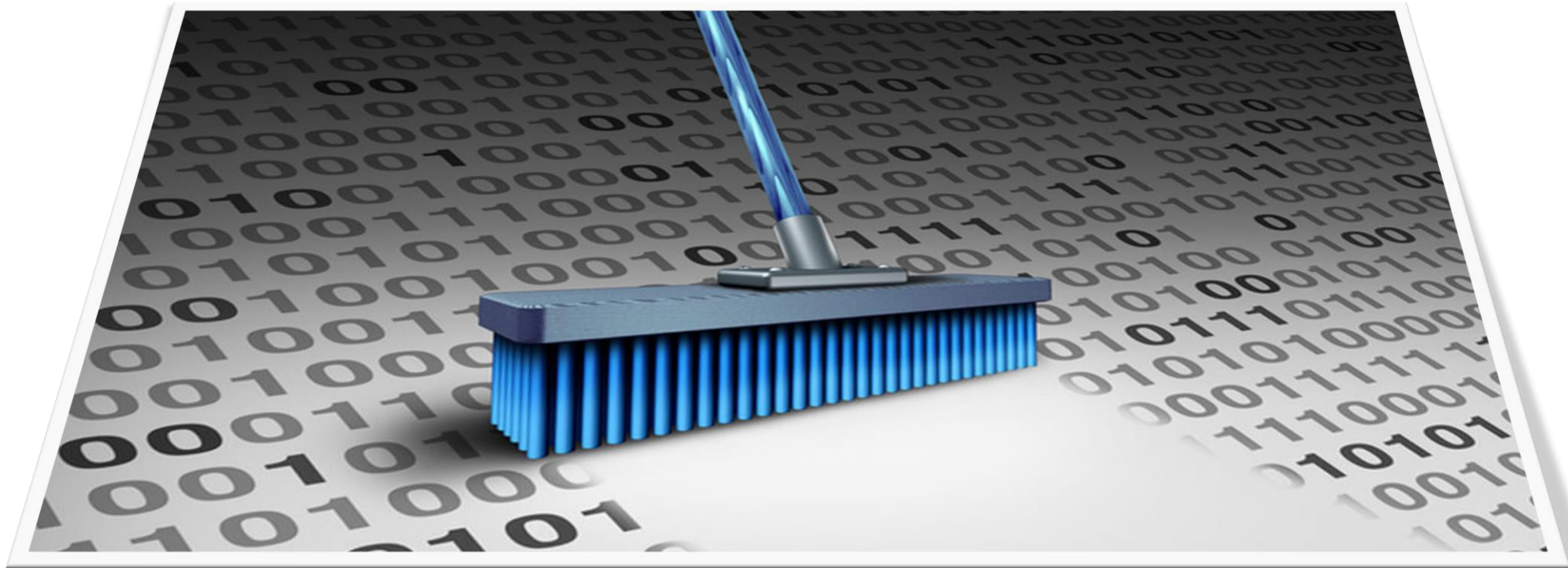


WQD7001



Data Cleaning

Prepared by Dr. Salimah Mokhtar

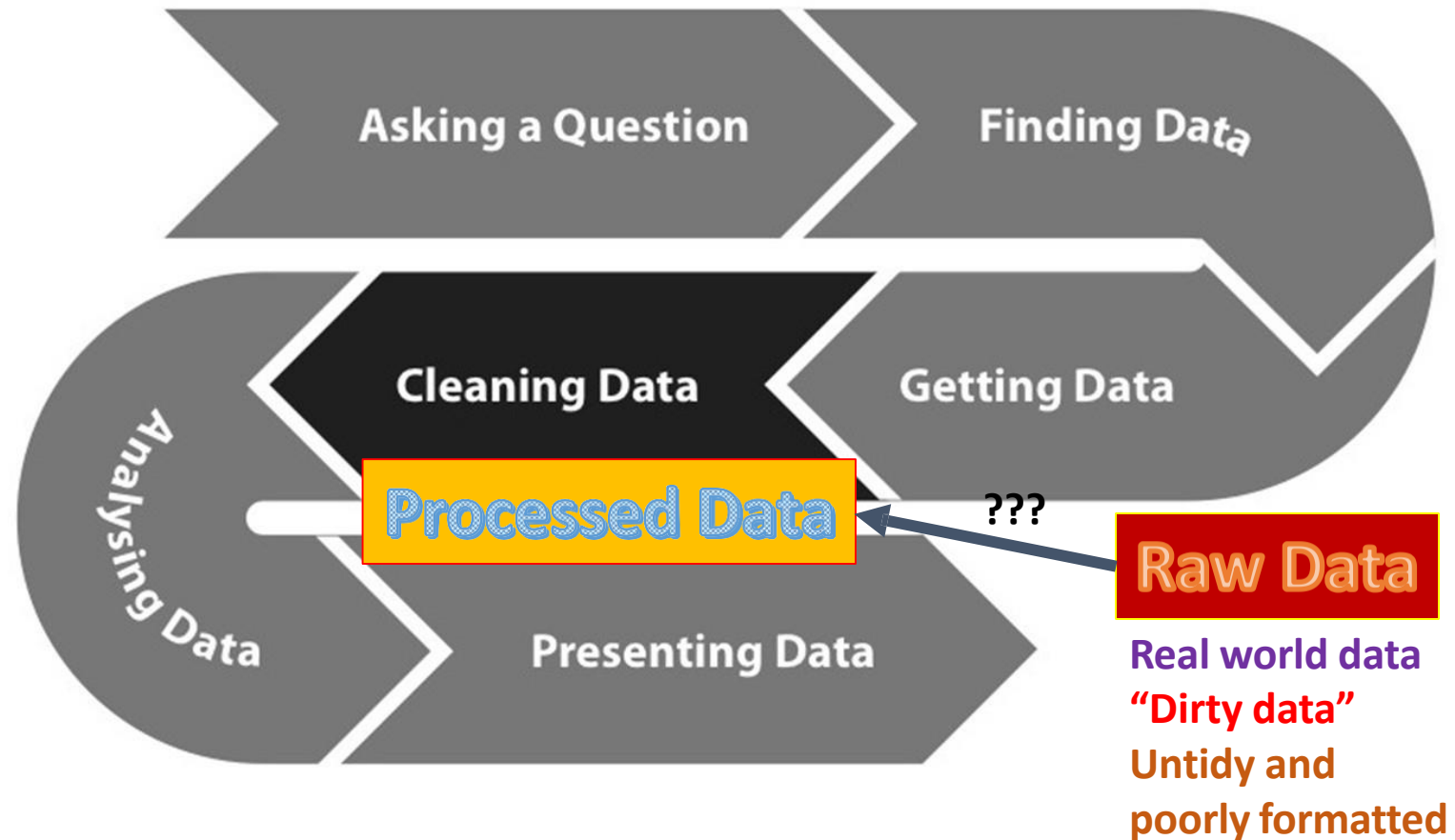
Data Cleaning

Learning Objectives:

1. To specify characteristics of **raw data** and **processed data**.
2. To determine the **problems** and **causes of dirty data**.
3. To categorize the **tasks in data pre-processing**.
4. To deal with **missing data**.
5. To identify **tools** for data pre-processing.
6. To describe **tidy data**.
7. To create a **code book**.

What Next?

Data cleaning is one those things that everyone does but no one really talks about.



The raw data

- The strange **binary file** from a measurement machine
- The **unformatted Excel** file with 10 worksheets the company you contracted with sent you
- The **complicated JSON data** you got from scraping the Twitter API
- The **hand-entered numbers** you collected looking through a microscope

You know the raw data is in the right format if you

1. Ran no software on the data
2. Did not manipulate any of the numbers in the data
3. You did not remove any data from the data set
4. You did not summarize the data in any way.



Processed Data

- Data that is **ready for analysis.**
- Processing can include merging, subsetting, transforming etc.
- There may be **standards** for processing.
- All processing steps should be **recorded.**

How can we
make data
less ugly for
more
beautiful
insights?

DATA



SORTED



ARRANGED



PRESENTED
VISUALLY



Therefore, to get our dataset into tip-top shape
(from untidy to tidy data), it must go through
data cleaning.

Pause and Think!

What are the causes of dirty data? List TWO that you can think of. One word per cause.

Is this An Acceptable Data Set?

Id	Name	<u>DoB</u>	Age	Gender	Phone	Country
1801	Shah Rukh Khan	11/12/1984	34	Male	5551212	India
1802	<u>Roselinda</u>	14/13/1986	32	Female	4568765	Kuala Lumpur
1803	Muhammad Ali Jannah	31/08/1983	35	M	5678900	Jordan
1804	Lynda Carter	12/30/1980	38	Female	9999999	America
1805	Smith, Tracy	23/08/1981	37	2	6856262	UK
1806	Ng Chee Chin	3/10/1989	19	<u>Fenale</u>	3209876	Malaysia
1807	<u>Fatoush Olkan</u>	18/07/1982	36	-	2348765	Turkey
1808	John Doe	20/11/1987	31	Male	7735075	USA
1809	Tracy Smith	23/08/1981	37	2	8356753	UK
1809	<u>Ibrar Yaacob</u>	18/09/1974	44	Male	6544321	Pakistan

LO2 Dirty Data Problems

From Stanford Data Integration Course:

- 1) Parsing text into fields (separator issues)
- 2) Naming conventions: **NYC** vs **New York**
- 3) Missing required field (e.g. key field)
- 4) Different representations (**2** vs **Two**)
- 5) Fields too long (get truncated)
- 6) Primary key violation (from un- to structured or during integration)
- 7) Redundant Records (exact match or other)
- 8) Formatting issues – especially dates
- 9) Licensing issues/Privacy/ keep you from using the data as you would like?

Not Parsed Correctly

Name	LastName	FirstName
Smith, John	Smith	John

Unexpected Pattern

Email
john.doe@google.com
jmillier@hotmail
schow@yahoo.com

Extra Characters

Name	Name
"John Smith"	John Smith

Misspelled Entries

125 Main Street
125 Main Street

Duplicate Data Records

ID	Name	ID	Name
1	John Smith	1	John Smith
1	John Smith	2	Jane Doe
2	Jane Doe		

Incorrect Data

Date	Sales
2016-05-01	1000
1900-01-01	500
2016-04-28	830



Common Types of Dirty Data

1. **Incomplete data:** most common occurrence of dirty data. Important fields on master data records, useful to the business, are often left blank. For example, if you haven't classified your customers by industry, you cannot segment your sales and marketing initiatives by industry.
2. **Duplicate data:** very common. Most companies deal with issues with duplicate customer records, but duplicate materials are also very common. This can be costly to companies due to excess in inventory and sub-optimal procurement decisions.
3. **Incorrect data:** Incorrect data can occur when field values are created outside of the valid range of values. For example, the value in a month field should range from 1 to 12 or a street address should be a real address.
4. **Inaccurate data:** It is possible or data to be technically correct but inaccurate given the business context. Costly business interruptions are often rooted in inaccurate data. For example, minor errors in customer addresses can result in deliveries at the wrong locations even though the addresses are actual addresses.
5. **Business rule violations:** There are often large collections of poorly documented business rules associated with master data that are specific to the industry or business context. For example, beverage products should have a Unit of Measure in 'fl. oz.' or payment terms for a certain type of customers should always be 'Net 30.'
6. **Inconsistent data:** Data redundancy—i.e., the same field values stored in different places—often leads to inconsistencies. For example, most companies have customer information in multiple systems and the data is often not kept in sync.

Where Does the “Dirt” Come From?

Dirty data caused by **human error** can take multiple forms:

Incorrect – The value entered does not comply with the field’s valid values. For example, the value entered for month is likely to be a number from 1 to 12. This value can be enforced with lookup tables or edit checks.

In **violation of business rules** – The value is not valid or allowed, based on the business rules (e.g., An effective date must always come before an expiration date.)

Incomplete – The data has missing values. No data value is stored in a field. For example, the street address is missing in a customer record.

Inaccurate – The value entered is not accurate. Sometimes, the system can evaluate the data value for accuracy based on context. For most systems, accuracy validation requires a manual process.

Inconsistent – The value in one field is inconsistent with the value in a field that should have the same data. Particularly common with customer data, one source of data inconsistencies is manual or unchecked data redundancy.

Duplicate – The data appears more than once in a system of record. Common causes include repeat submissions, improper data joining or blending, and user error.

Where Does the “Dirt” Come From?

✧ **Incomplete data** comes from

- N/A data value when collected
- Different consideration between the time when the data was collected and when it is analyzed
- Human/hardware/software problems

✧ **Noisy data** comes from the processing of data

- Collection
- Entry
- Transmission

✧ **Inconsistent data** comes from

- Different data sources
- Functional dependency violation

According to [Technopedia](#), there are the following subsets of dirty data:

- Misleading data
- Duplicate data
- Incorrect data
- Inaccurate data
- Non-integrated data
- Data that violates business rules
- Data without a generalized formatting
- Incorrectly punctuated or spelled data

Identify these data problems

Email	Gender	Name1	Name2	Birthday
richard@example.com	M	Richard	Ware	
GUEST	F	Rhonda	Brown	1974
edwards@example.net		Mark	E	12/14/1996
	F	Sarah	Lyons	12/25/1956



Mr. Smith



(6/20/91)
123 Fake St.

Jon Smith



(6/20/91)
123 Fake St.

John Smith



(unknown)
123 Fake St.



Mark Horowi\$z
(m horowitz#gmail.com)
born December 33rd
Lives in zip code 8

Complete data looks like this:

Email	Gender	First Name	Last Name	Birthday
richard@example.com	M	Richard	Ware	06/22/1987
rbrown@example.com	F	Rhonda	Brown	10/10/1974
edwards@example.net	M	Mark	Edwards	12/14/1996
coolgrams@example.net	F	Sarah	Lyons	12/25/1956

Accurate data looks like this:



Unique data looks like this:

Jon Smith



(6/20/91)
123 Fake St.

John Smith



(4/18/87)
486 Fakerton
Blvd.

John Smithson



(8/19/90)
292 Faux St.

Valid data looks like this:



Mark Horowitz
(mhorowitz@gmail.com)
born December 3rd
Lives in zip code 48102

Dirty Data Is OK, How You Cleanse It Matters

Businesses can suffer analysis paralysis without quality data input, and they can never have clean data without the help of analytics to help them identify data errors.

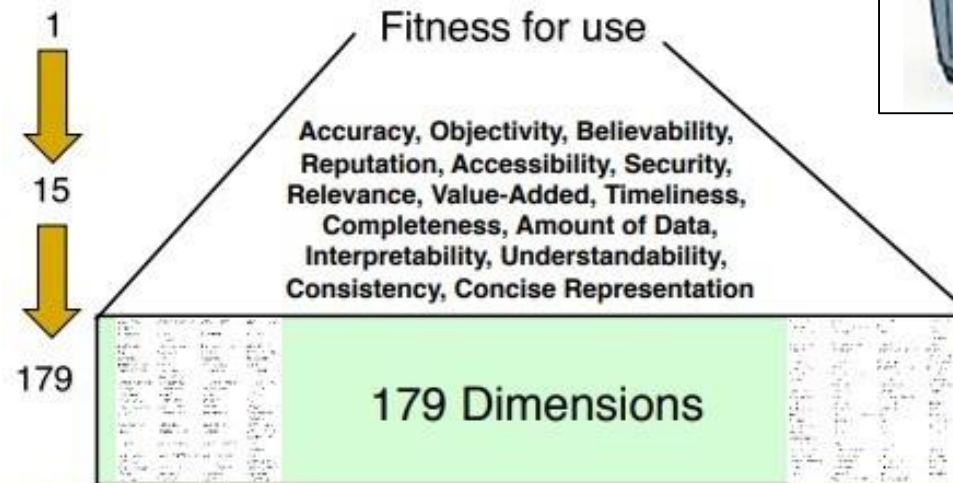


by Chirag Shivalker · Nov. 15, 17 · Big Data Zone · Opinion

Why is Data Cleaning Important?

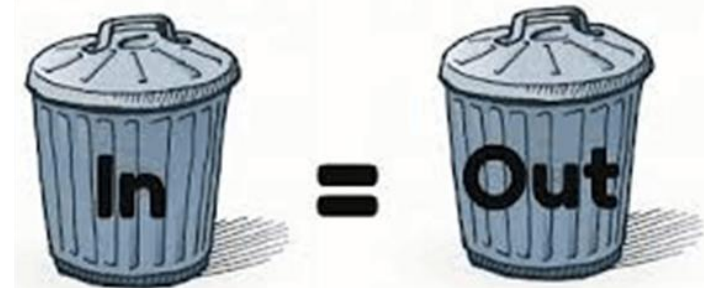
Cleaning and Transforming to get ... **High-quality data!**
No quality data, no quality decisions!

What is Data of Good Quality?

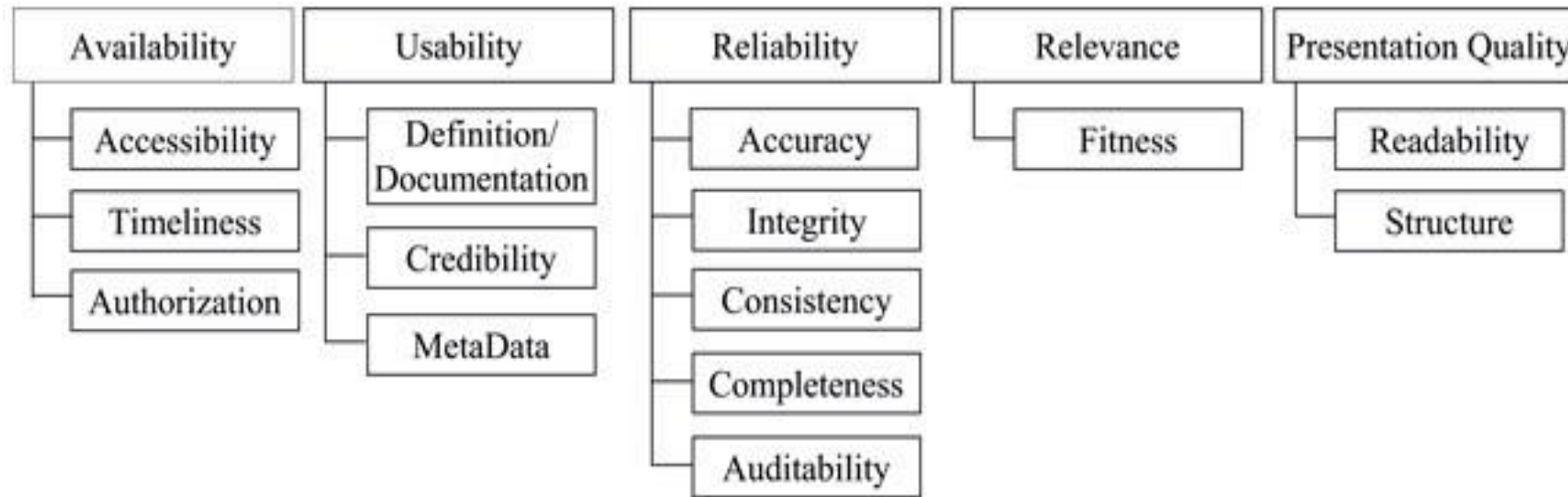


Problem using dirty data

GIGO



Data Quality Standard - composed of 5 dimensions of data quality



Availability is defined as the degree of convenience for users to obtain data and related information, which is divided into the three elements of accessibility, authorization, and timeliness.

Usability means whether the data are useful and meet users' needs, including data definition/documentation, reliability, and metadata.

Reliability refers to whether we can trust the data; this consists of accuracy, consistency, completeness, adequacy, and auditability elements.

Relevance is used to describe the degree of correlation between data content and users' expectations or demands; adaptability is its quality element.

Presentation quality refers to a valid description method for the data, which allows users to fully understand the data. Its dimensions are readability and structure.

Dimensions	Elements	Indicators	
1) Availability	1) Accessibility	■	Whether a data access interface is provided
		■	Data can be easily made public or easy to purchase
	2) Timeliness	■	Within a given time, whether the data arrive on time
		■	Whether data are regularly updated
		■	Whether the time interval from data collection and processing to release meets requirements

2) Usability	1) Credibility	■	Data come from specialized organizations of a country, field, or industry
		■	Experts or specialists regularly audit and check the correctness of the data content
		■	Data exist in the range of known or acceptable values

3) Reliability	1) Accuracy	■	Data provided are accurate
		■	Data representation (or value) well reflects the true state of the source information
		■	Information (data) representation will not cause ambiguity
	<hr/>		
	2) Consistency	■	After data have been processed, their concepts, value domains, and formats still match as before processing
		■	During a certain time, data remain consistent and verifiable
		■	Data and the data from other data sources are consistent or verifiable

3) Integrity	■	Data format is clear and meets the criteria
	■	Data are consistent with structural integrity
	■	Data are consistent with content integrity
<hr/>		
4) Completeness	■	Whether the deficiency of a component will impact use of the data for data with multi-components
	■	Whether the deficiency of a component will impact data accuracy and integrity

4) Relevance	1) Fitness	■	The data collected do not completely match the theme, but they expound one aspect
		■	Most datasets retrieved are within the retrieval theme users need
		■	Information theme provides matches with users' retrieval theme
<hr/>			
5) Presentation Quality	1) Readability	■	Data (content, format, etc.) are clear and understandable
		■	It is easy to judge that the data provided meet needs
		■	Data description, classification, and coding content satisfy specification and are easy to understand

Data science is the process
by which data becomes
understanding, knowledge
and insight

“Data is what you need to do analytics.
Information is what you need to do
business.”

“Your **results** are only
as good as your data.”

“Without data you’re
just another person with
an opinion.”

*Data will talk to you if you’re
willing to listen to it.*



“Torture the data, and it
will confess to anything.”

“Errors using inadequate
data are much less than
those using no data at
all.”

Tasks in Data Preparation / Data Pre-Processing

LO3

- ✓ **Data cleaning** - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies.
- ✓ **Data integration** - Integration of multiple databases, data cubes, or files.
- ✓ **Data transformation** - Normalization and aggregation.
- ✓ **Data reduction** - Obtains reduced representation in volume but produces the same or similar analytical results.

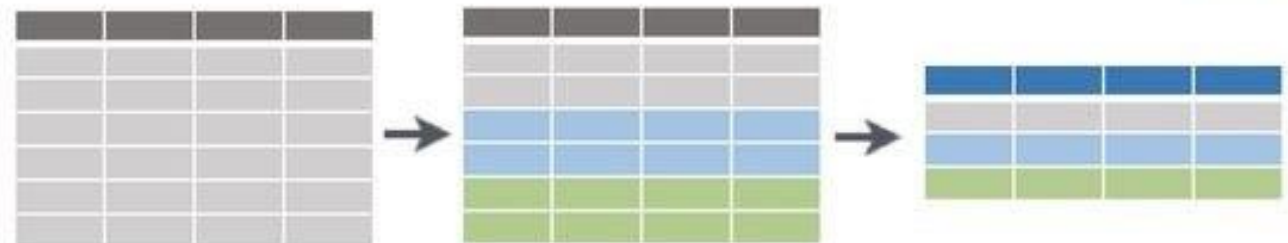
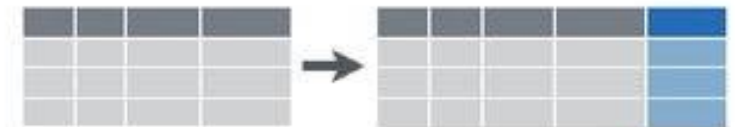
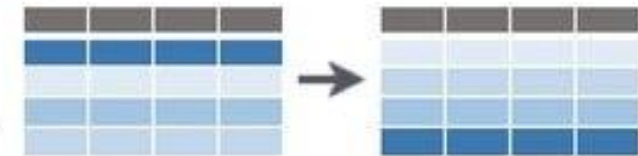
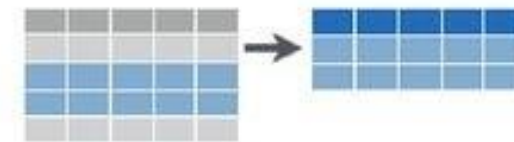
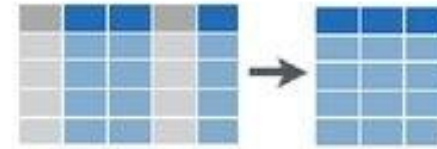
Data Manipulations

- **Filtering**, or subsetting: Remove observations based on some condition.
- **Aggregating**: collapse multiple values into a single value (e.g., count, average, summing or taking means).
- **Transforming**: add new variables or modify existing variables (e.g., log-transforming).
- **Sorting**: change the order of values.

**Data
Reduction**

Key Functions

- `select()`
 - Picks variables (columns) based on their names.
- `filter()`
 - Picks observations (rows) based on their values.
- `arrange()`
 - Changes the ordering of the rows based on their values.
- `summarise()`
 - Reduces multiple values down to a single summary value.
- `mutate()`
 - Adds new variables that are functions of existing variables.
- `group_by()`
 - Performs data operations on groups that are defined by variables.



Data cleansing or **data cleaning** is the process of detecting and correcting (or removing) corrupt or inaccurate [records](#) from a record set, [table](#), or [database](#) and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the [dirty](#) or coarse data.

Data cleansing may be performed [interactively](#) with [data wrangling](#) tools, or as [batch processing](#) through [scripting](#).

Wikipedia

Data Cleaning

Data cleaning is an important skill for data scientist.

“In our experience, the tasks of exploratory data mining and data cleaning constitute **80% of the effort** that determines 80% of the value of the ultimate data.”

T. Dasu and T. Johnson
Authors of Exploratory Data Mining

Revolutions

Daily news about using open source R for big data analysis, predictive modeling, data science, and visualization since 2008

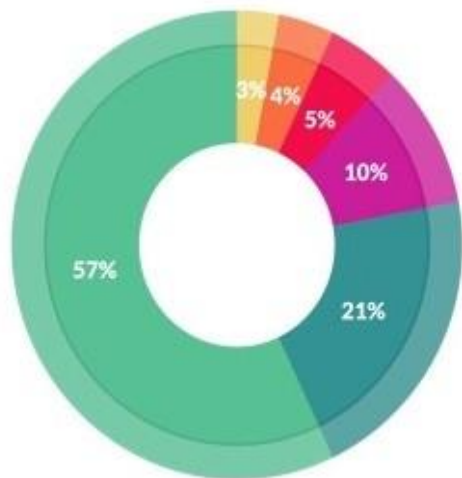
« [Because it's Friday: Speeding up time, keeping it smooth](#) | [Main](#) | [Integrating R with production systems using an HTTP API](#) »

August 18, 2014

Data Cleaning is a critical part of the Data Science process

A [New York Times](#) article yesterday discovers the 80-20 rule: that 80% of a typical data science project is sourcing cleaning and preparing the data, while the remaining 20% is actual data analysis. The article gives short shrift to this important task by calling it "janitorial work", but whether you call it data munging, data wrangling or anything else, it's a critical part of the data science. I'm in agreement with Jeffrey Heer, professor of computer science at the University of Washington and a co-founder of Trifacta, who is quoted in the article saying,

"It's an absolute myth that you can send an algorithm over raw data and have insights pop up."



What's the least enjoyable part of data science?

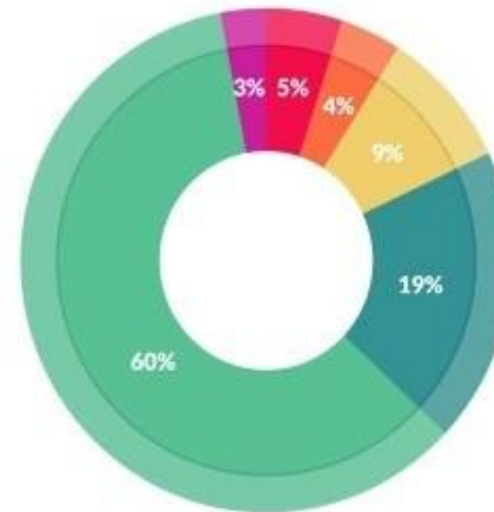
- Building training sets: 10%
- Cleaning and organizing data: 57%
- Collecting data sets: 21%
- Mining data for patterns: 3%
- Refining algorithms: 4%
- Other: 5%

Data cleaning
Data cleansing
Data munging
Data wrangling

80% of a typical data science project is sourcing cleaning and preparing the data, while the remaining 20% is actual data analysis.

Data Scientists Spend Most of Their Time Cleaning Data

Posted on [May 1, 2016](#)



What data scientists spend the most time doing

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets: 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%

Cleaning Data

Fixing up formats – Often when data is saved or translated from one format to another, some data may not be translated correctly. E.g. The timestamp_first_active column contained numbers like 20090609231247 instead of timestamps in the expected format: 2009-06-09 23:12:47. A typical job when it comes to cleaning data is correcting these types of issues.

Filling in missing values – Quite common for some values to be missing from datasets. This typically means that a piece of information was simply not collected. There are several options for handling missing data.

Correcting erroneous values – For some columns, there are values that can be identified as obviously incorrect. This may be a 'gender' column where someone has entered a number, or an 'age' column where someone has entered a value well over 100. These values either need to be corrected (if the correct value can be determined) or assumed to be missing.

Standardizing categories – More of a subcategory of 'correcting erroneous values', this type of data cleansing is so common it is worth special mention. In many (all?) cases where data is collected from users directly – particularly using free text fields – spelling mistakes, language differences or other factors will result in a given answer being provided in multiple ways. For example, when collecting data on country of birth, if users are not provided with a standardized list of countries, the data will inevitably contain multiple spellings of the same country (e.g., USA, United States, U.S. and so on). One of the main cleaning tasks often involves **standardizing** these values to ensure that there is only one version of each value.

Remove Unwanted Observations

The first step to data cleaning is **removing unwanted observations** from your dataset. This includes **duplicate observations** or **irrelevant observations**.

Duplicate observations

Duplicate observations most frequently arise during **data collection**, such as when you:

- Combine datasets from multiple places
- Scrape data
- Receive data from clients/other departments

Irrelevant observations

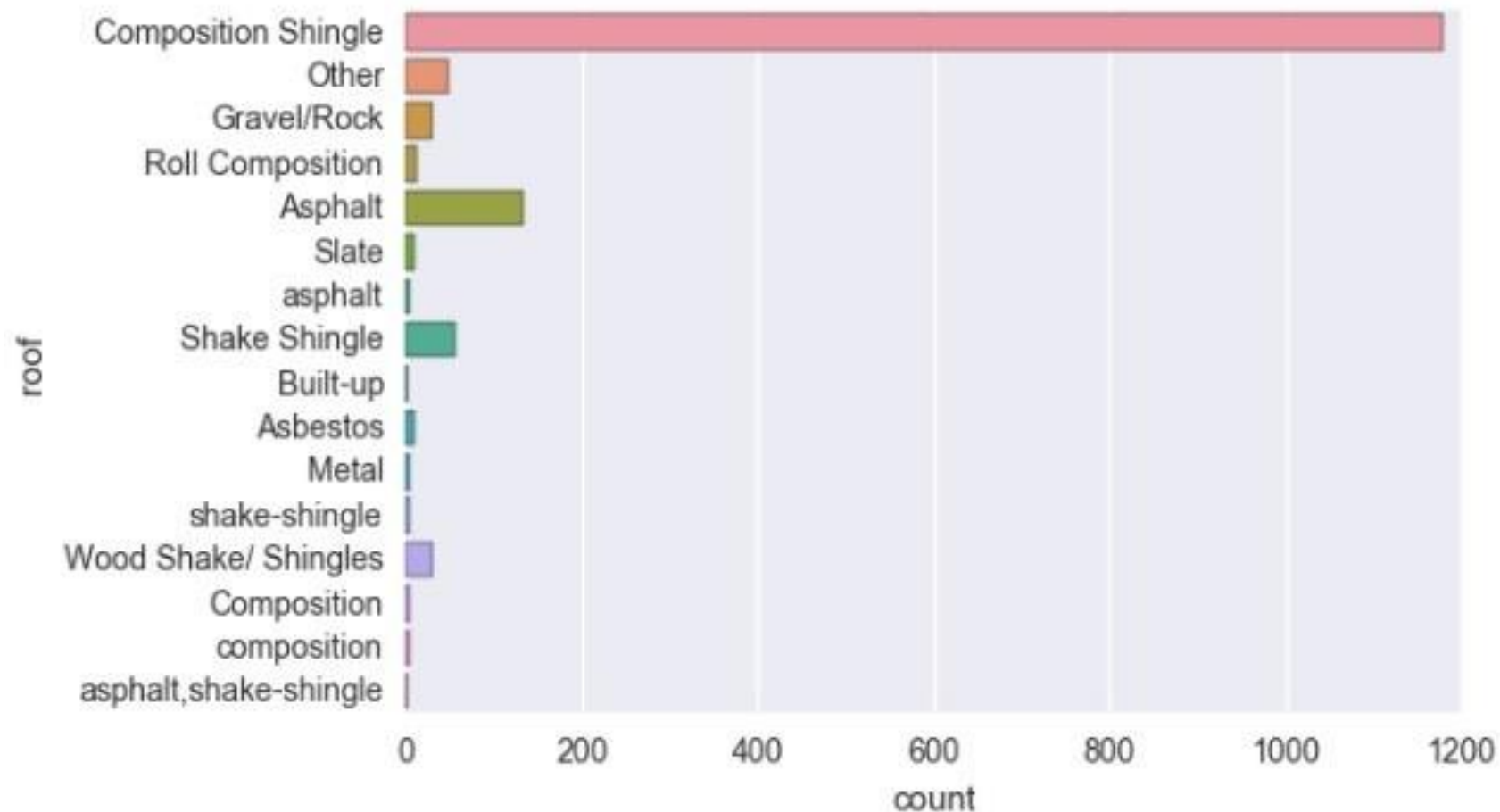
Irrelevant observations are those that don't actually fit the **specific problem** that you're trying to solve.

- For example, if you were building a model for Single-Family homes only, you wouldn't want observations for Apartments in there.
- This is also a great time to review your charts from Exploratory Analysis. You can look at the distribution charts for categorical features to see if there are any classes that shouldn't be there.
- Checking for irrelevant observations **before engineering features** can save you many headaches down the road.

Fix Structural Errors

Structural errors are those that arise during measurement, data transfer, or other types of "poor housekeeping."

Example shown below:



Check for typos or inconsistent capitalization, we will find:

'composition' is the same as 'Composition'

'asphalt' should be 'Asphalt'

'shake-shingle' should be 'Shake Shingle'

'asphalt,shake-shingle' could probably just be 'Shake Shingle'

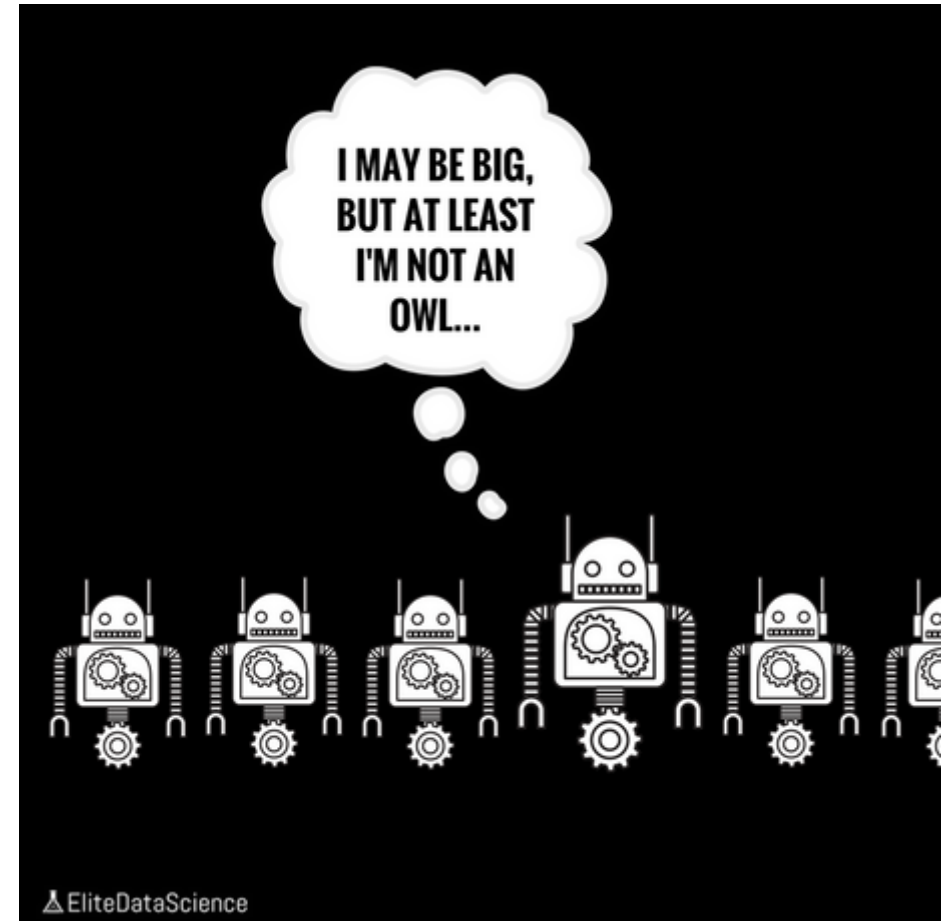
Filter Unwanted Outliers

Outliers can cause problems with certain types of models. For example, linear regression models are less robust to outliers than decision tree models.

In general, if you have a legitimate reason to remove an outlier, it will help your model's performance.

However, outliers are innocent until proven guilty. You should never remove an outlier just because it's a "big number." That big number could be very informative for your model.

MUST have a **good reason** for removing an outlier, such as suspicious measurements that are unlikely to be real data.



LO4

Missing Data

- There are many reasons why data could be missing, including:



Respondents forgot to answer questions.

Respondents refused to answer certain questions.

Respondents failed to complete the survey.



A sensor failed.

Someone purposefully turned off recording equipment.

There was a power cut.

The method of data capture was changed.



An internet connection was lost.

A network went down.

A hard drive became corrupt.

A data transfer was cut short.

Missing Data

An issue of almost every raw data set!

What is Missing Data?

Missing data (or missing values) is defined as the data value that is not stored for a variable in the observation of interest.

Why Bother?

- Some algorithm won't work.
- Can add **BIAS** to a model → overestimate or underestimate value

Respondent ID	Age	Income	Education	Homeowner
1	27	32000	HS	N
2	37	64000	BA	Y
3		44000	HS	N
4	55	78000	MA	Y
5	23		HS	N
6	25	42000		N
7	35	121000	PhD	Y
8	51	45000	BA	
9			MS	N
10	67	54000	MA	Y

Real Data Set

Respondent ID	Age	Income	Education	Homeowner
1	27	32000	HS	N
2	37	64000	BA	Y
3	70	44000	HS	N
4	55	78000	MA	Y
5	23	30000	HS	N
6	25	42000	HS	N
7	35	121000	PhD	Y
8	51	45000	BA	N
9	65	200000	MS	N
10	67	54000	MA	Y

Pattern of Missing Data

There are **3 types** of missing data:

- **MCAR** = missing completely at random
 - The distribution of missing data is unpredictable (i.e. the cases with missing data are indistinguishable from cases with complete data)
- **MAR** = missing at random (a.k.a. ignorable non-response)
 - The pattern is predictable from other variables in the data
- **MNAR** = missing not at random or non-ignorable
 - The pattern is related to the dependent variable and cannot be ignored



X
|
 Y

Z
|
 R

(a) MCAR

X Z
| |
 Y R

Diagram (b) shows a diagonal line connecting X to R , indicating that the missingness of Y depends on X .

(b) MAR

X Z
| |
 Y R

Diagram (c) shows a diagonal line connecting X to R and a horizontal line connecting Y to R , indicating that the missingness of Y depends on both X and Y .

(c) MNAR

Figure 2. Graphical representations of (a) missing completely at random (MCAR), (b) missing at random (MAR), and (c) missing not at random (MNAR) in a univariate missing-data pattern. X represents variables that are completely observed, Y represents a variable that is partly missing, Z represents the component of the causes of missingness unrelated to X and Y , and R represents the missingness.

Survey Data on Drug Abuse

- **Missing Completely at Random (MCAR):**
 - You removed 10% of the respondents data randomly.
- **Missing at Random (MAR):** (most common type)
 - People who come from poorer families might be less inclined to answer questions about drug use, and so the level of drug use is related to family income.
- **Not Missing at Random (NMAR):**
 - Students skipped the question on drug use because they feared that they would be expelled from school.

Options for Dealing With Missing Data

Missing data in general is one of the **trickier** issues that is dealt with when cleaning data. Broadly there are **TWO common solutions**:

1. Deleting/Ignoring rows with missing values

The **simplest solution** - is to not use the records with missing values when training your model.

Some issues to be aware of before you starting deleting masses of rows from your dataset.

- ❖ ONLY makes sense if the number of rows with missing data is relatively small compared to the dataset.

If deleting more than around **10%** of your dataset due to rows having missing values, you may need to reconsider.

- ❖ To delete the rows containing missing data, you have to be confident that the rows you are deleting do not contain information that is not contained in other rows.

2. Filling in the Values (Imputation)

Fill the missing values with a **value**. But what value to use? This depends on a range of factors, including the type of data you are trying to fill.

- ❖ For **categorical data** (i.e. countries, device types, etc.) - make sense to simply create a new category that will represent 'unknown'. Another option may be to fill the values with the most common value for that column (the mode).

- ❖ For **numerical values** (for example the age column) - use the mean or median or average based on some other criteria.

Mean

Add together all of the numbers in your data set. Then divide that sum by the number of addends.

Median

the middle number when listed in order from least to greatest

Mode

The number that occurred the most often in the data set.

Deleting rows with missing values

Respondent ID	Age	Income	Education	Homeowner
1	27	32000	HS	N
2	37	64000	BA	Y
3		44000	HS	N
4	55	78000	MA	Y
5	23		HS	N
6	25	42000		N
7	35	121000	PhD	Y
8	51	45000	BA	
9			MS	N
10	67	54000	MA	Y



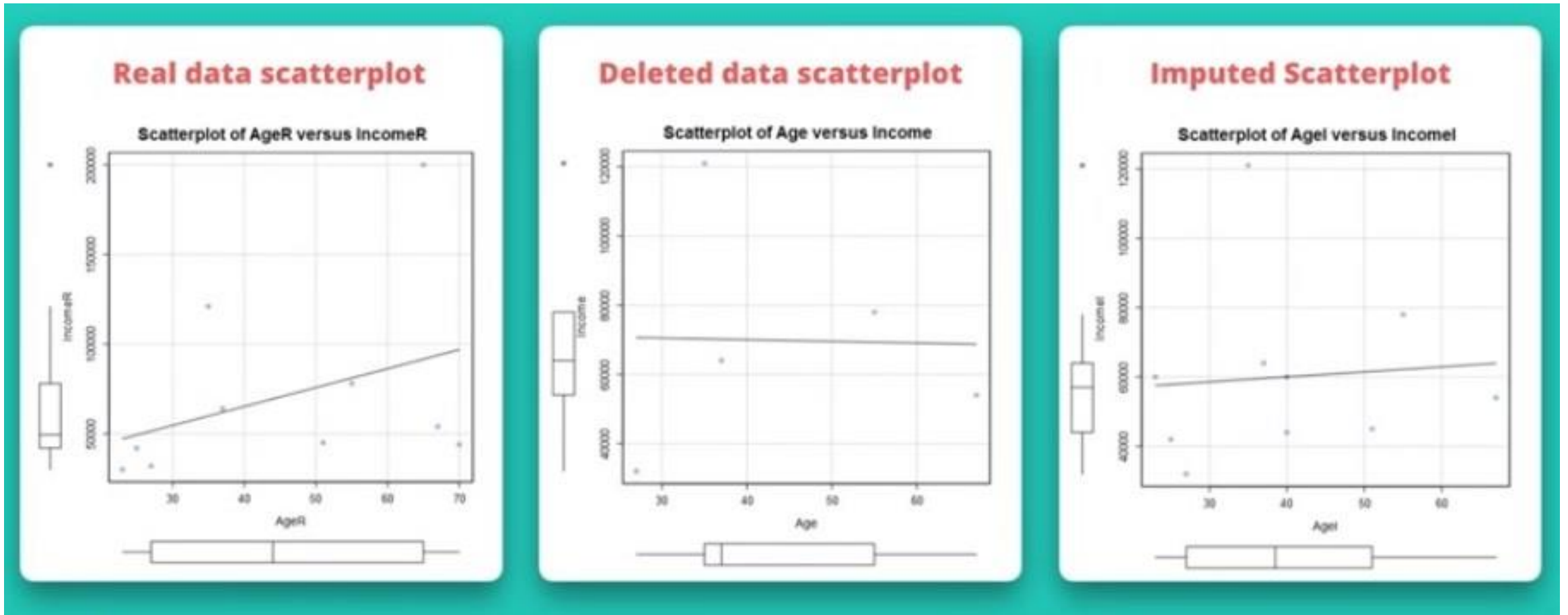
Respondent ID	Age	Income	Education	Homeowner
1	27	32000	HS	N
2	37	64000	BA	Y
4	55	78000	MA	Y
7	35	121000	PhD	Y
10	67	54000	MA	Y

Imputed Data (Mean)

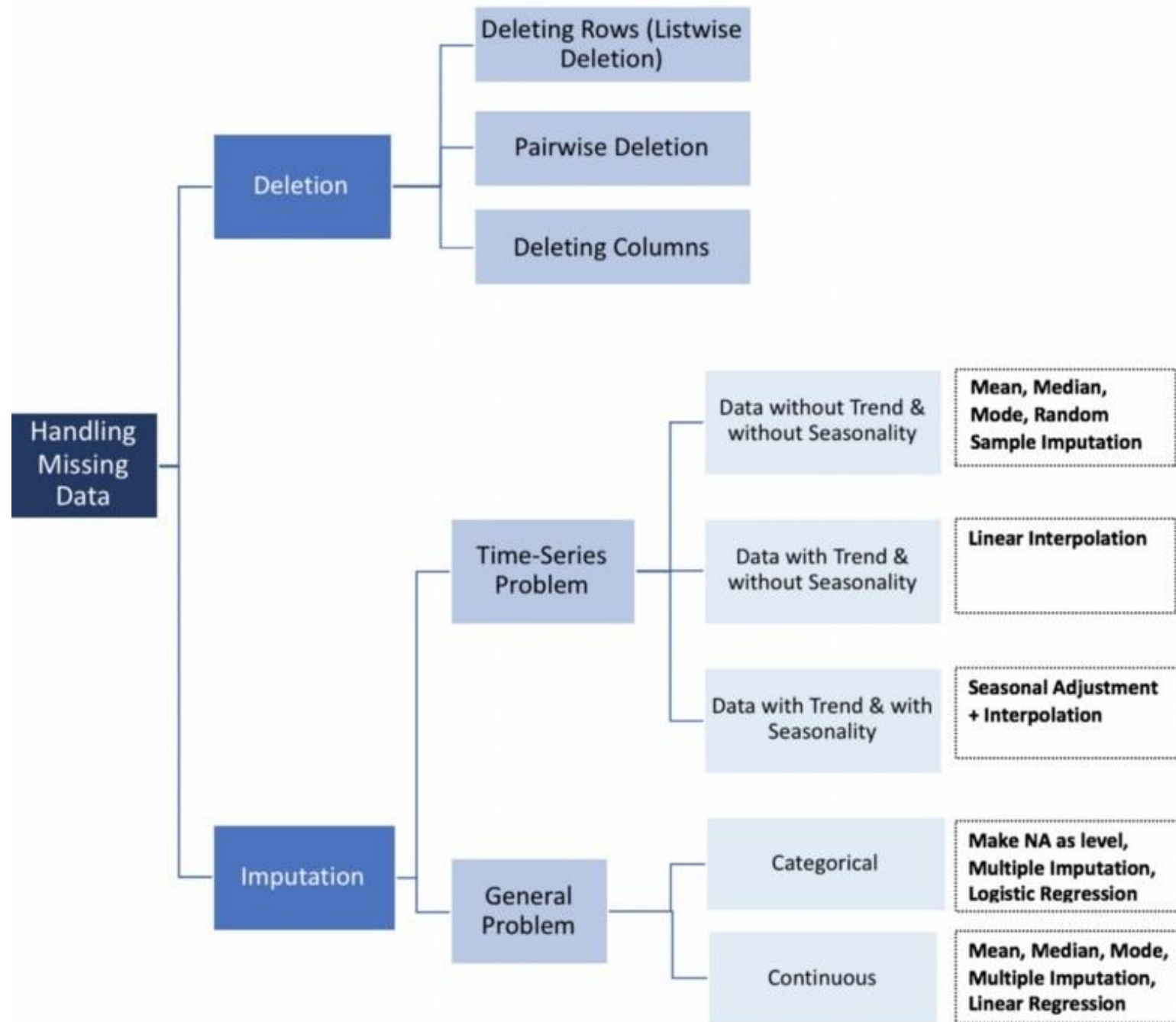
Respondent ID	Age	Income	Education	Homeowner
1	27	32000	HS	N
2	37	64000	BA	Y
3	40	44000	HS	N
4	55	78000	MA	Y
5	23	60000	HS	N
6	25	42000	HS	N
7	35	121000	PhD	Y
8	51	45000	BA	N
9	40	60000	MS	N
10	67	54000	MA	Y

Simple imputation methods have been used and may be sufficient in some settings, but have potential to introduce **bias** and **inaccuracy** into the statistical analysis. More complex methods such as **multiple imputation** potentially offer reduced risk of bias.

The Effect



Missing data can also lead to misleading results by introducing bias



a. Listwise: In this case, rows containing missing variables are deleted.

User	Device	OS	Transactions
A	Mobile	NA	5
B	Mobile	Android	3
C	NA	iOS	2
D	Tablet	Android	1
E	Mobile	iOS	4


b. Pairwise: In this case, only the missing observations are ignored and analysis is done on variables present.

User	Device	OS	Transactions
A	Mobile	NA	5
B	Mobile	Android	3
C	NA	iOS	2
D	Tablet	Android	1
E	Mobile	iOS	4

In the above case, 2 separate sample data will be analyzed, one with the combination of User, Device and Transaction and the other with the combination of User, OS and Transaction. In such a case, one won't be deleting any observation. Each of the samples will ignore the variable which has the missing value in it.

Argument

Dropping observations that have missing values. 

Imputing the missing values based on other observations. 

- Dropping missing values is sub-optimal because when you drop observations, **you drop information**.
- The fact that the value was missing may be informative in itself.
- “Missingness” is almost always informative in itself, and we should tell your algorithm if a value was missing.

Missing data puzzle pieces

Missing data is like missing a puzzle piece.

If you drop it, that’s like pretending the puzzle slot isn’t there.

If you impute it, that’s like trying to squeeze in a piece from somewhere else in the puzzle.

Always tell the algorithm that a value was missing because missingness is informative.

Suggestion

Label as missing.

- The key is to tell your algorithm that the value was originally missing.

Missing categorical data

- The best way to handle missing data for categorical features is to simply label them as 'Missing'!
- We're essentially adding a new class for the feature.
- This tells the algorithm that the value was missing.

Missing numeric data

- For missing numeric data, you should flag and fill the values.
- Flag the observation with an indicator variable of missingness.
- Then, fill the original missing value with 0 just to meet the technical requirement of no missing values.
- By using this technique of flagging and filling, we are essentially allowing the algorithm to estimate the optimal constant for missingness, instead of just filling it in with the mean.

L05

Tools of Data Preprocessing

- **R** - R a framework that consists of various packages that can be used for Data Preprocessing like dplyr etc
- **Weka** - Weka is a software that contains a collection of Machine Learning algorithms for Data Mining process. It consists of Data Preprocessing tools that are used before applying Machine Learning algorithms.
- **RapidMiner** - RapidMiner is an open-source Predictive Analytics Platform for Data Mining process. It provides the efficient tools for performing exact Data Preprocessing process.
- **Python** - Python is a programming language that provides various libraries that are used for Data Preprocessing.

Handling Missing Data in R

In R, **missing values** are represented by the symbol **NA** (not available).

Impossible values (e.g., dividing by zero) are represented by the symbol **NaN** (not a number).

Check for missing data in a data set

- When inputting data directly into R, '**NA**' is used to designate missing data. For example,

```
> xvar <- c(2,NA,3,4,5,8)
```

Creates a variable ('xvar') for a sample of 6 subjects, but the second subject is missing data for this variable. NA is also used to indicate missing data when R prints data:

```
> xvar
```

```
[1] 2 NA 3 4 5 8
```

```
> is.na(xvar)
```

```
1 FALSE TRUE FALSE FALSE FALSE FALSE
```

The Swirl logo, consisting of the word "Swirl" in a blue, stylized font with a reflection effect below it.

| Please choose a lesson, or type 0 to return to course menu.

1: Basic Building Blocks	2: Workspace and Files	3: Sequences of Numbers
4: Vectors	5: Missing values	6: Subsetting Vectors
7: Matrices and Data Frames	8: Logic	9: Functions
10: lapply and sapply	11: vapply and tapply	12: Looking at Data
13: Simulation	14: Dates and Times	15: Base Graphics

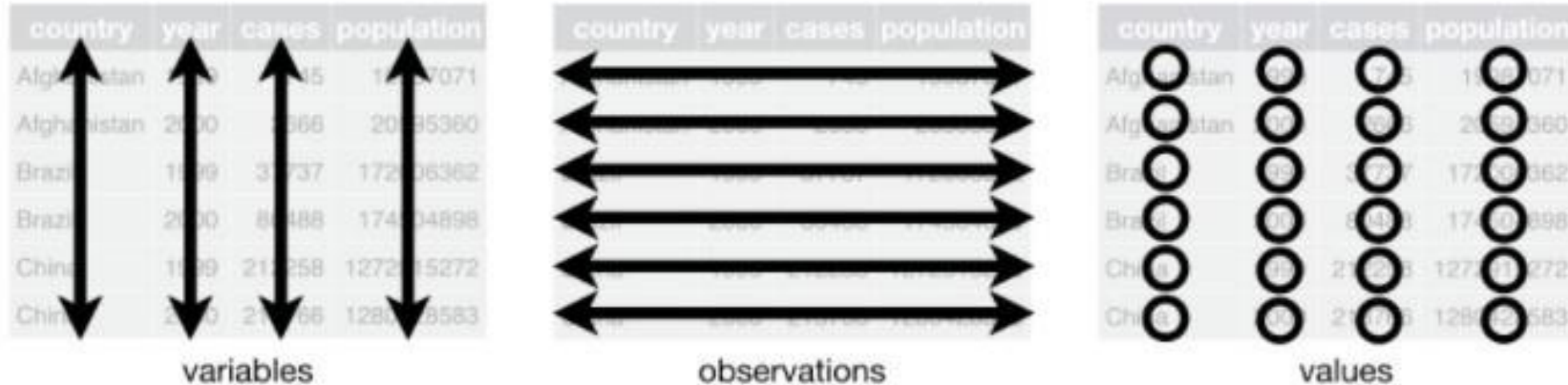
LO6 Tidy Data

- The **end goal** of cleaning data process.
- Each **variable** should be in one **column**.
- Each **observation** of that variable should be in a different **row**
- There should be **one table** for each kind of variable
 - if there are multiple tables, there should be a column to link them
- Include a row at the top of each file with variable names (variable names should make sense, human readable)
- In general data should be save in one file per table.

The point of creating a tidy data set is to get the data into a format that can be easily shared, computed on, and analyzed.



Data that satisfies these rules is known as *tidy data*. Notice that `table1` is tidy data.



In `table1`, each variable is placed in its own column, each observation in its own row, and each value in its own cell.

Variables = Columns
Observations = Rows
Data set = Table

Go here for further reading: <https://garrettgman.github.io/tidying/>

<https://jrnold.github.io/r4ds-exercise-solutions/tidy-data.html#non-tidy-data>

LO7

Code Book

For almost any data set, the **measurements** you calculate will need to be **described in more detail** than you will sneak into the spreadsheet. It is the **meta data**.

At minimum a code book should contain:

- **Information about the variables** (including units!) in the data set not contained in the tidy data.
- Information about the **summary choices** you made.
- Information about the **experimental study design** you used.

Data Cleaning Hands-on Examples

Please try out.

- Tricks for cleaning your data in R

<https://github.com/underthecurve/r-data-cleaning-tricks/blob/master/R-datacleaning-tricks.md>

- Checklist for Data Cleansing

https://sebastiansauer.github.io/data_cleansing/

- Cleaning Data in R

<https://www.youtube.com/watch?v=mGQvJ3FuNa8>

DataCleaningExample.csv

<https://github.com/davidcaughlin/R-Tutorial-Data-Files/blob/master/DataCleaningExample.csv>

References

The Elements of Data Analytic Style

<http://worldpece.org/sites/default/files/datastyle.pdf>

Data Cleaning - <https://elitedatascience.com/data-cleaning>

See **Cleaning Data in Excel** - <https://www.youtube.com/watch?v=plyMc-oqbv4>

A Taxonomy of Dirty Data -

[https://www.researchgate.net/publication/220451798 A Taxonomy of Dirty Data](https://www.researchgate.net/publication/220451798_A_Taxonomy_of_Dirty_Data)

Natural Language Processing (Part 2): Data Cleaning & Text Pre-Processing in Python https://www.youtube.com/watch?v=iQ1bfDMCv_c