



DAIA

Data Cleaning



AI project methodology: your roadmap



Target variable, model and domain requirements, data source(s), ...
What do you want to achieve (predict)?

What data (quality) is required?
Define a data dictionary

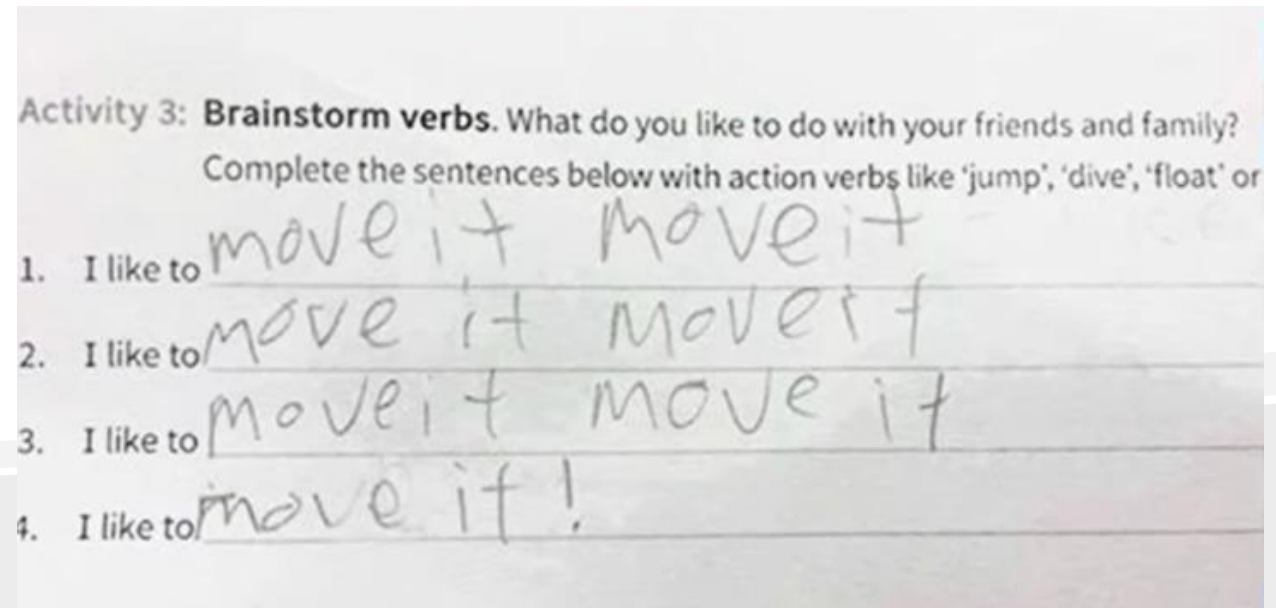
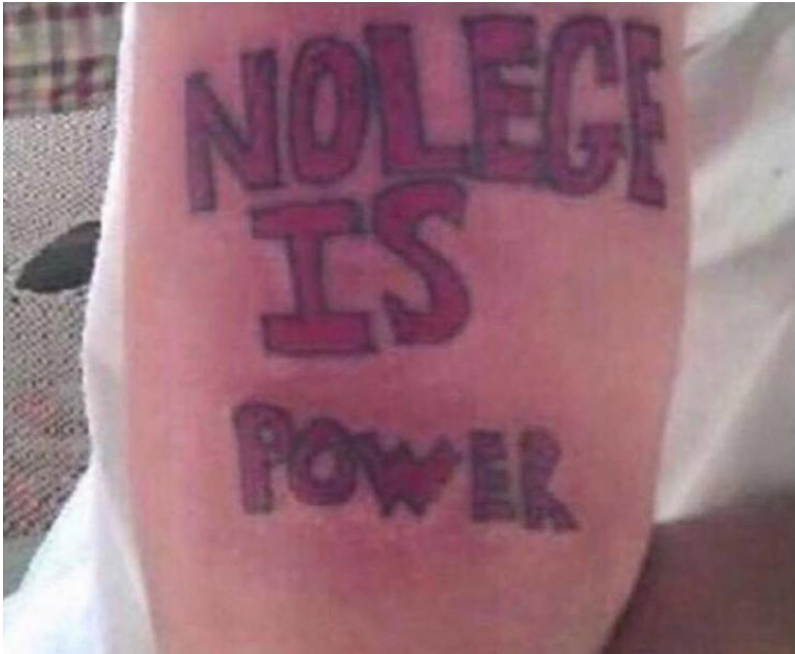
How do you get (generate) and combine your data? Capture the process.

Explore your data (EDA and EDV)

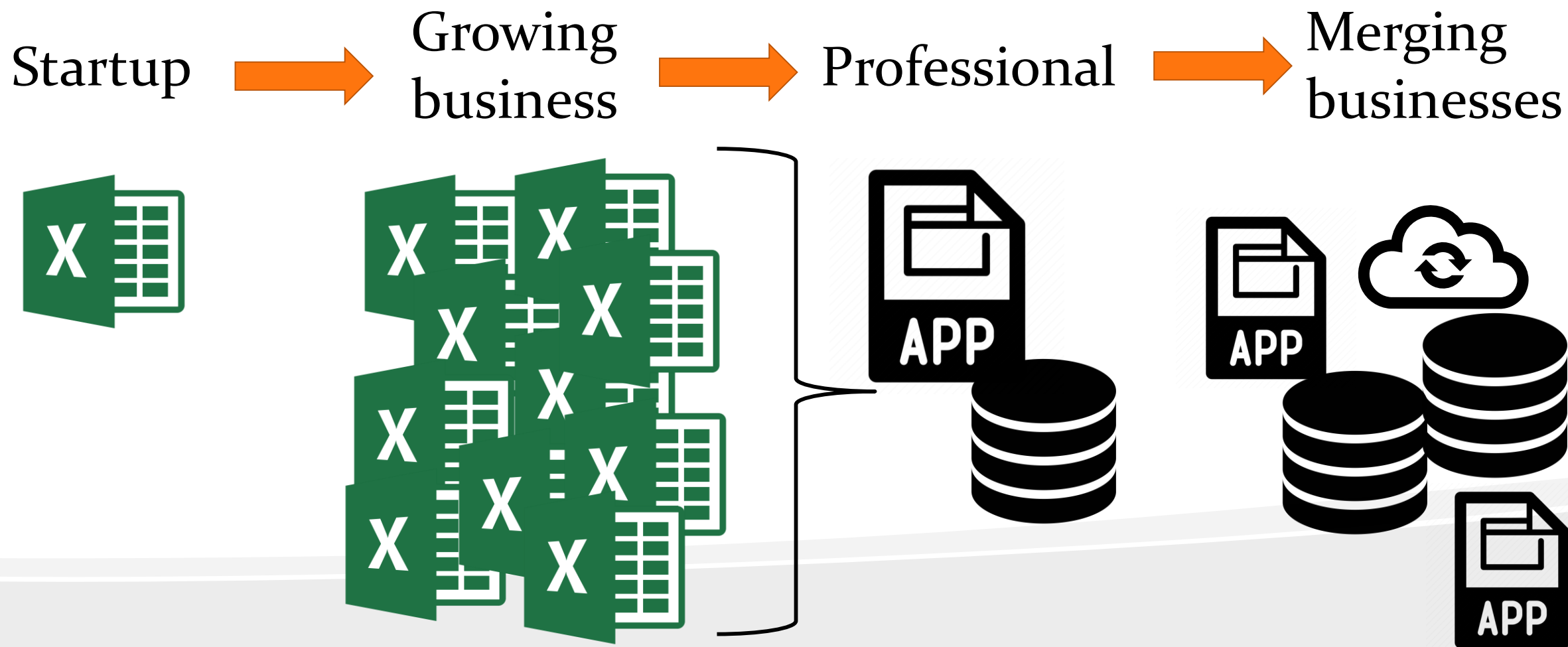
Prepare your data: meet the requirements

Why is data messy?

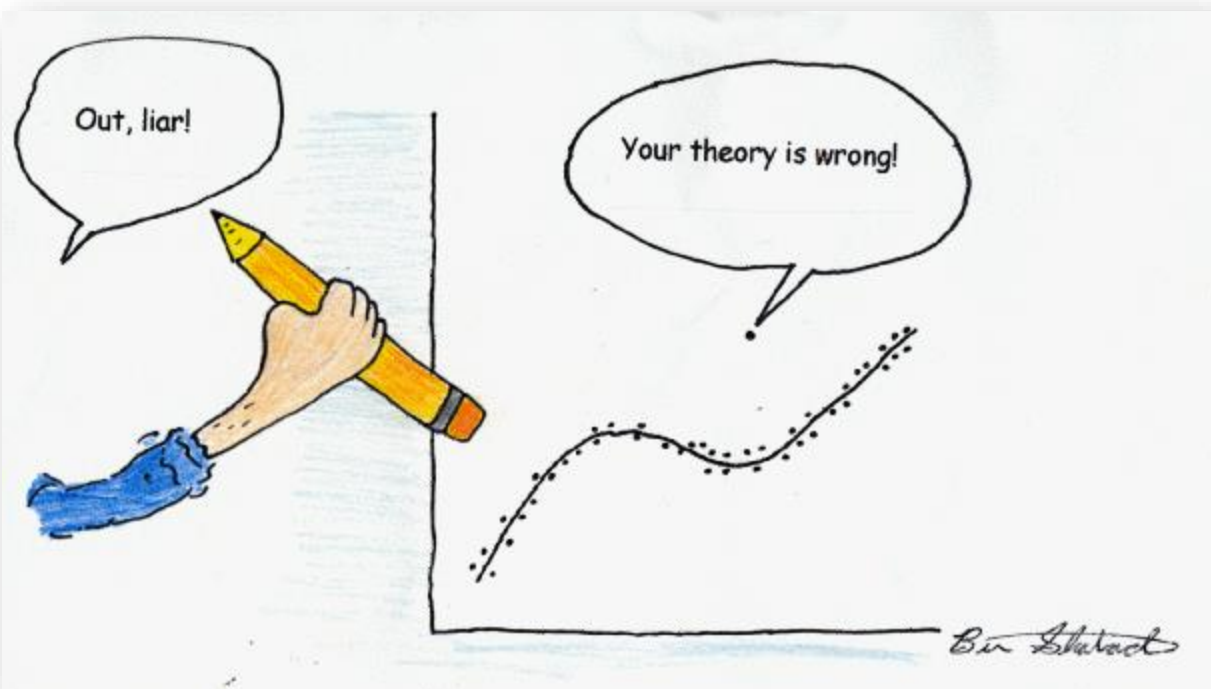
- What's your experience?



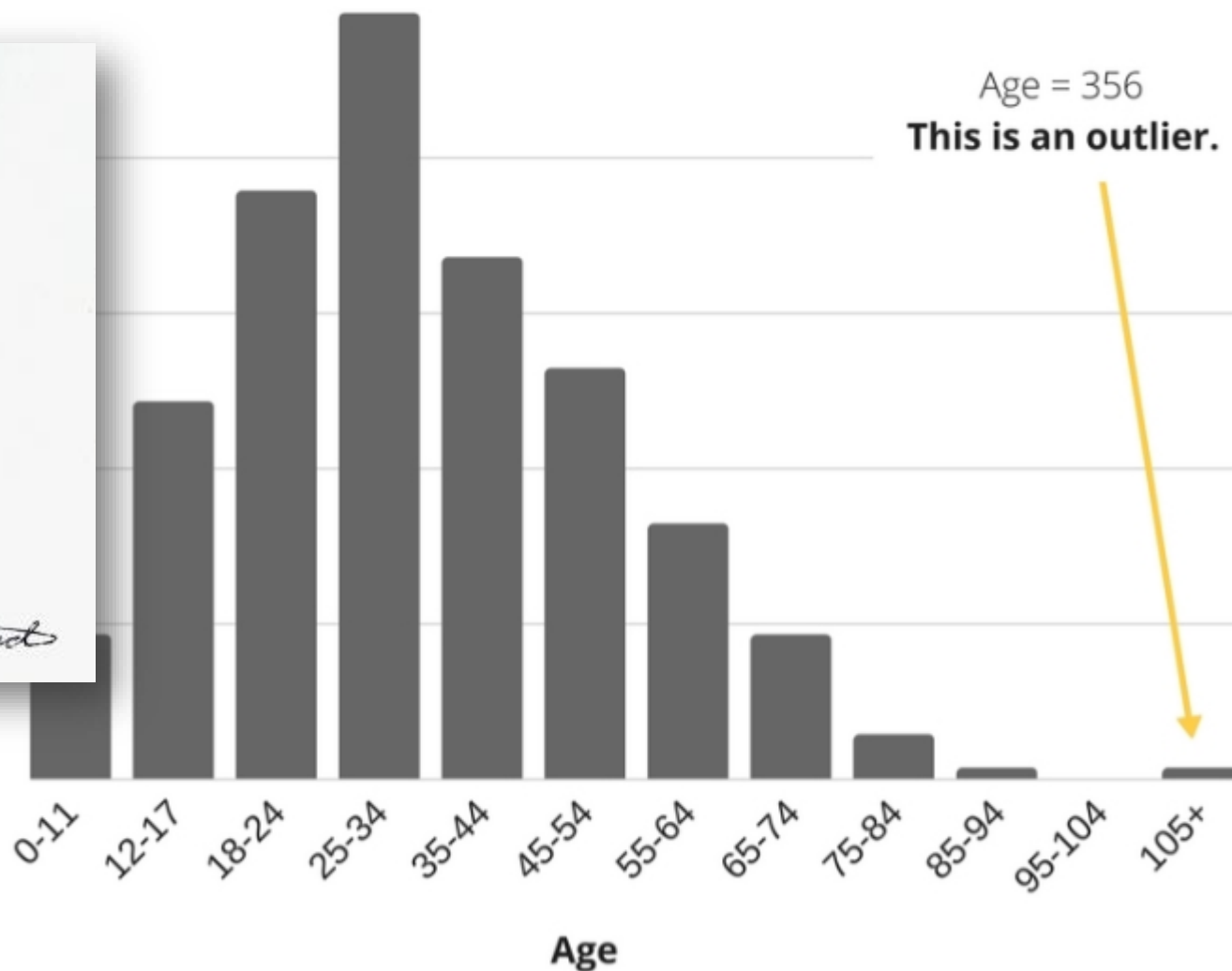
Why is data messy? → Data collection



... from wk2 Visualization basics

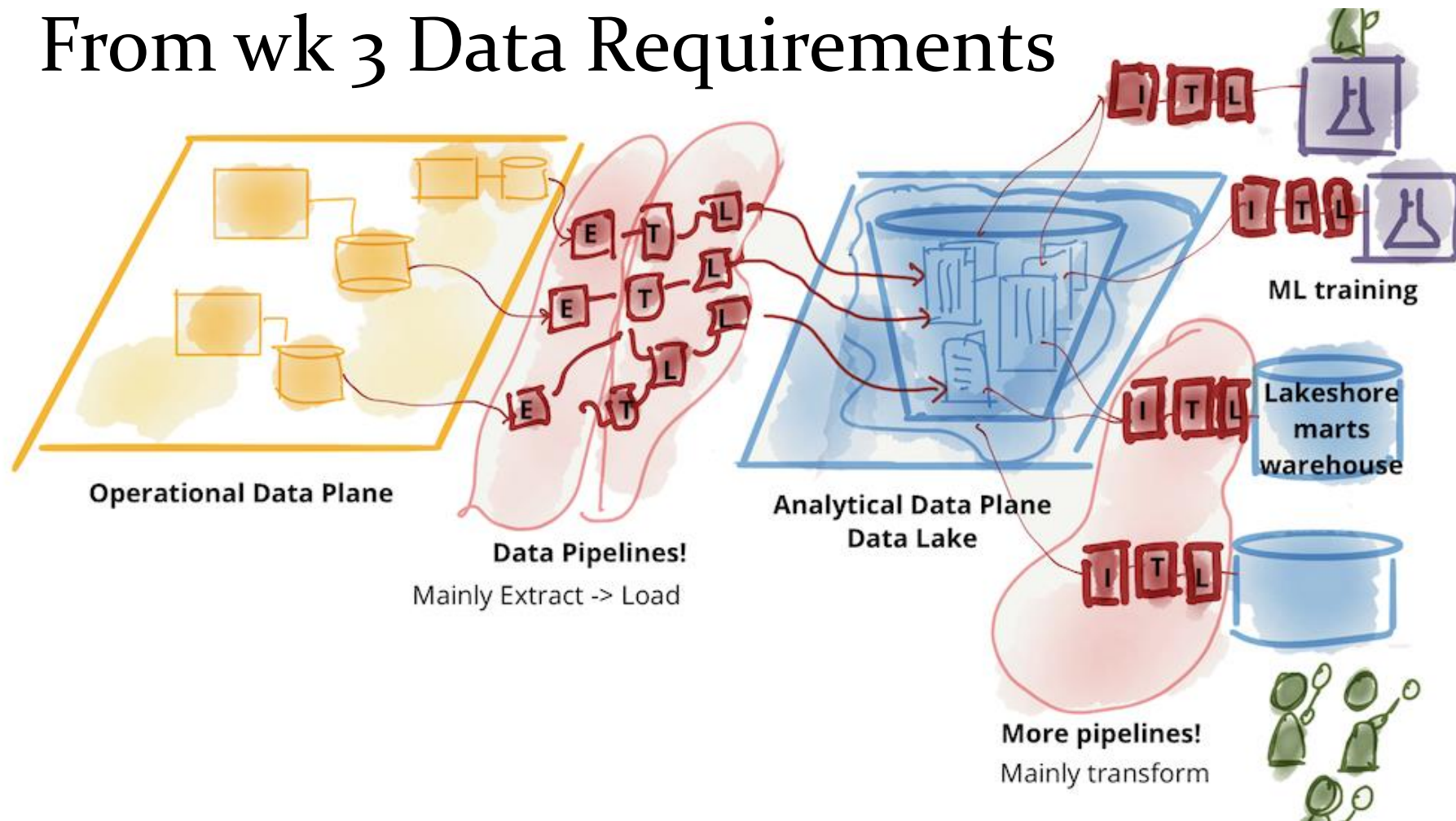


<https://psychab.wordpress.com/2011/10/28/27/>



Source: <https://medium.com/analytics-vidhya/its-all-about-outliers-cbe172aa1309>

From wk 3 Data Requirements



Data requirements (company perspective)



Data cleaning

- What **quality standards** are there?
- How to clean the data **systematically**?



Data quality criteria

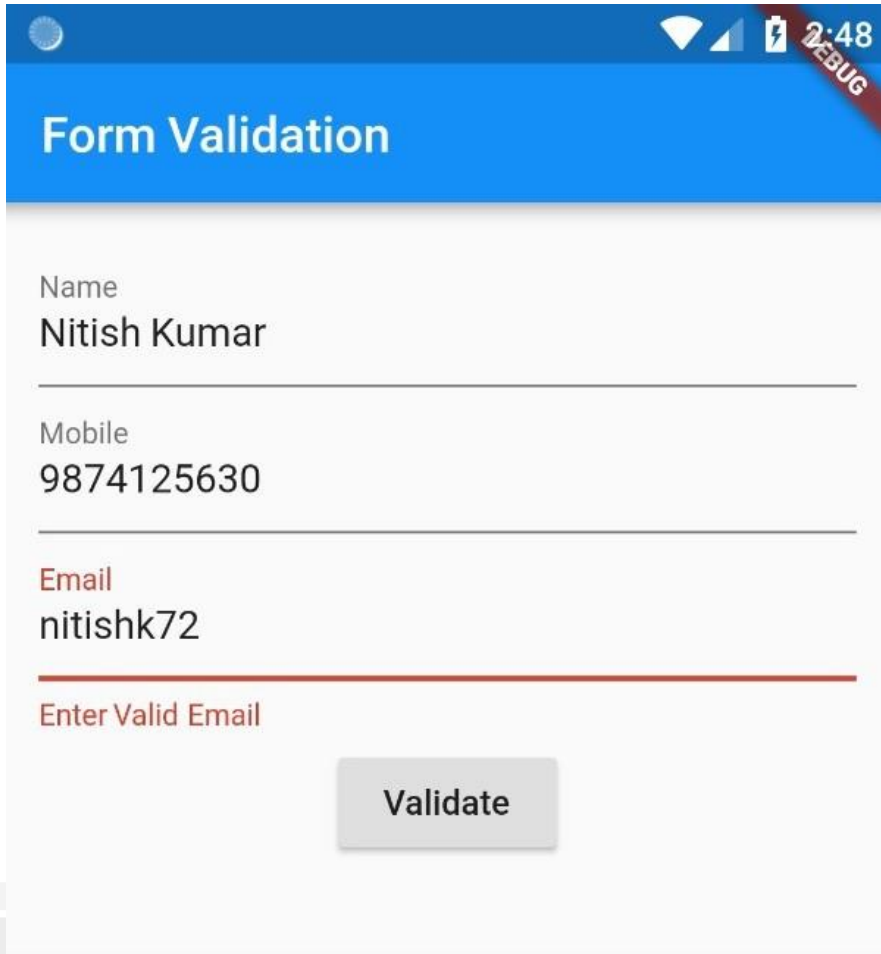


Data quality criteria

- **Validity**
degree to which conform to defined (business) rules or constraints
- Accuracy
- Consistency
- Completeness
- Uniqueness
- Timeliness



Validity



Form Validation

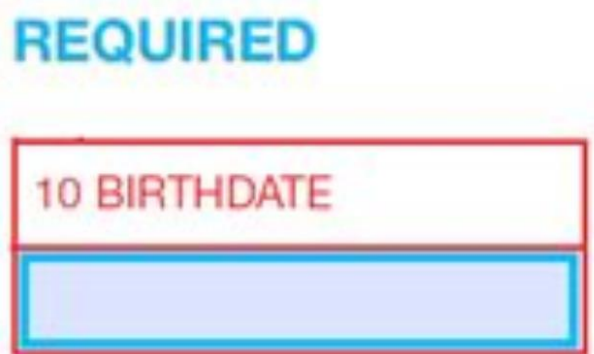
Name
Nitish Kumar

Mobile
9874125630

Email
nitishk72

Enter Valid Email

Validate



REQUIRED

10 BIRTHDATE

The date of birth of the patient
Use the format MMDDYYYY

Example: 11251964

Data quality criteria

- Validity
- **Accuracy**
degree to data represents the truth
- Consistency
- Completeness
- Uniqueness
- Timeliness





Accuracy

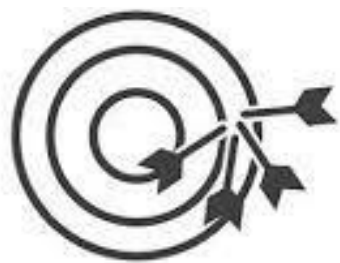
Submit Form

Username *

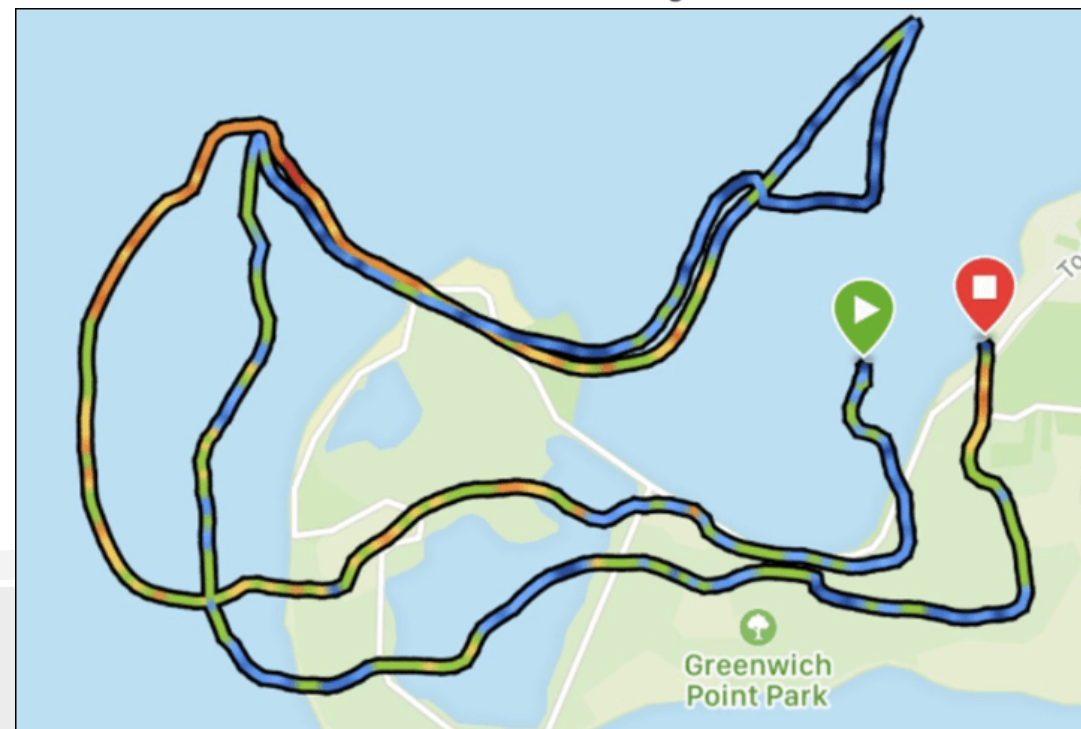
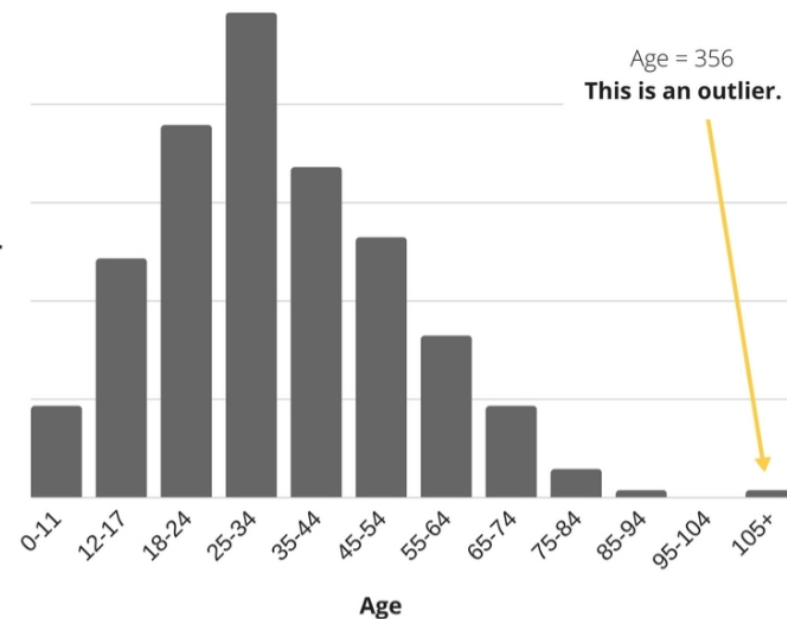
Password *

Confirm Password *

Age *



Accuracy Vs Precision



Data quality criteria

- Validity
- Accuracy
- **Consistency**
Degree to which data is equivalent across systems (shared understanding of definitions)
- Completeness
- Uniqueness
- Timeliness



Consistency

	A	B
1	Name	Age
2	Abhay	37
3	Arinjay	29
4	Balram	27
5	Chetan	28
6	Tarun	Thirty
7		

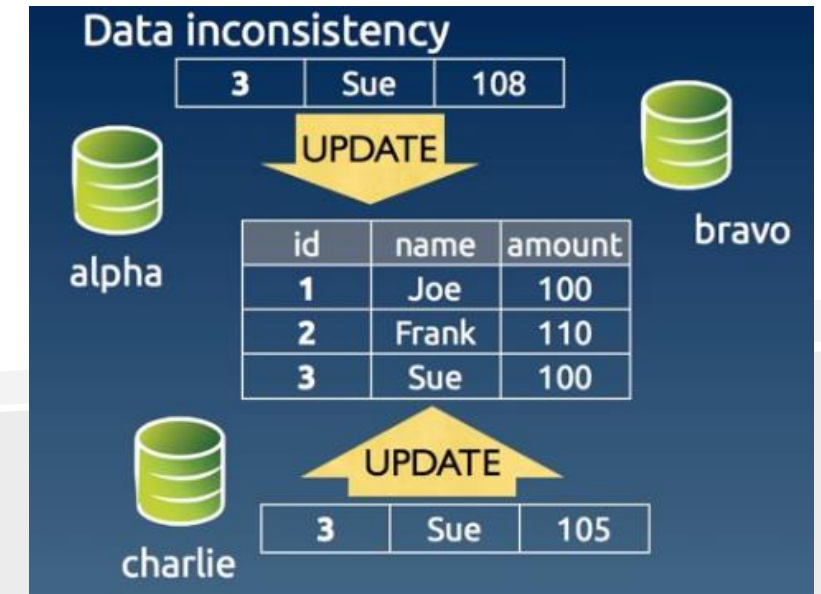
Valid dates:

12-11-2014 = 11. November '14?

Remember Data Dictionary:

Variable name	Variable description	Variable type	Variable width	Values/notes
MRN	Medical record number	Numeric	6.0	000001-900000
Treatment	Treatment group	Numeric	1.0	1=treated, 2=control
AGE	Age in months	Numeric	2.1	6.0-59.9
DOB	Date of birth	Date	-	MM/DD/YYYY
SEX	Sex	Numeric	1.0	1=male, 2=female
Height	Height (cm)	Numeric	3.1	-
Weight	Weight (kg)	Numeric	3.1	-
Proc_date	Procedure date	Date	-	MM/DD/YYYY
systolic_BP	Systolic blood pressure	Numeric	3.0	-

Also consider the process



Data quality criteria

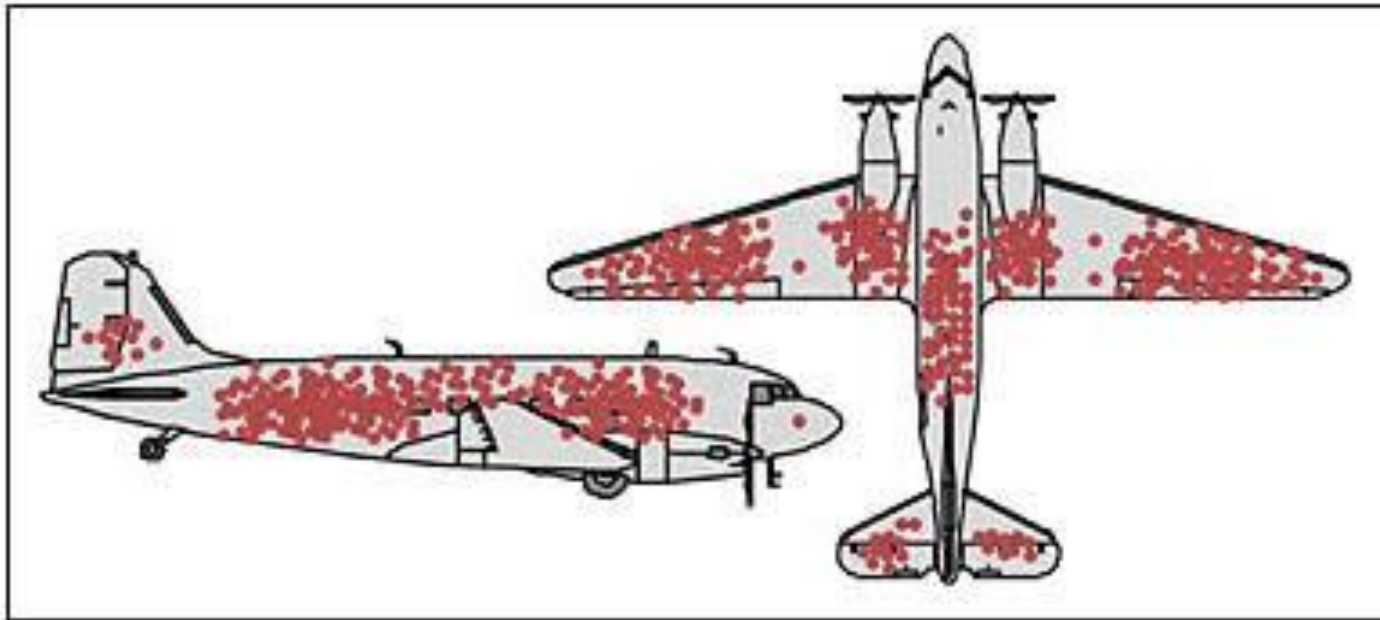
- Validity
- Accuracy
- Consistency
- **Completeness**
Degree to which all required data is known/present
- Uniqueness
- Timeliness



Completeness

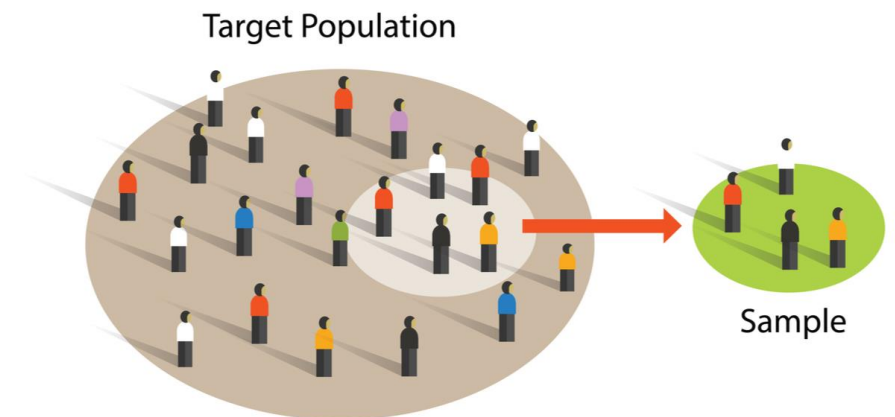
Missing values:

- Missing values within observations (NA's / nulls)
- Missing entire observations



Credit: Cameron Moll

Sample of the real world



Data quality criteria

- Validity
- Accuracy
- Completeness
- Consistency
- **Uniqueness**
Degree to which data is unique (unambiguous)
- Timeliness



Uniqueness

Over time, data can change:

- phone numbers
- addresses
- names
- etc.

Job interview

studentno	Name	Surname	dob
9314098	Chris	Kellen	14/02/1998

Administration

studentno	Name	Surname	dob
9314098	Iris	Kellen	14/02/1998



Data quality criteria

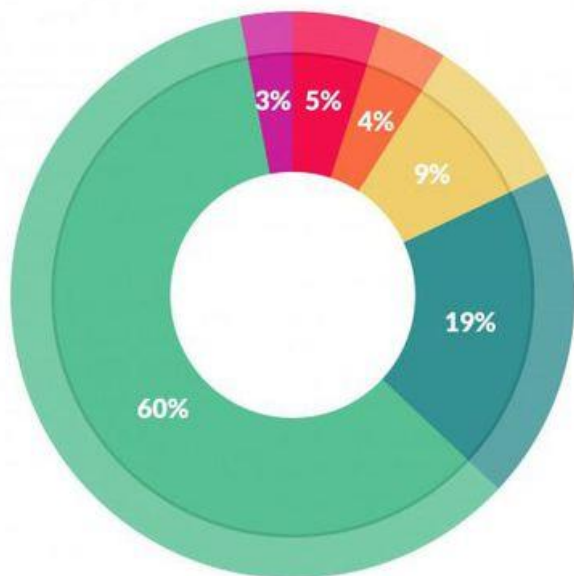
- Validity
- Accuracy
- Completeness
- Consistency
- Uniqueness
- **Timeliness**
Degree to which data is available when needed



Who likes to clean anyway...



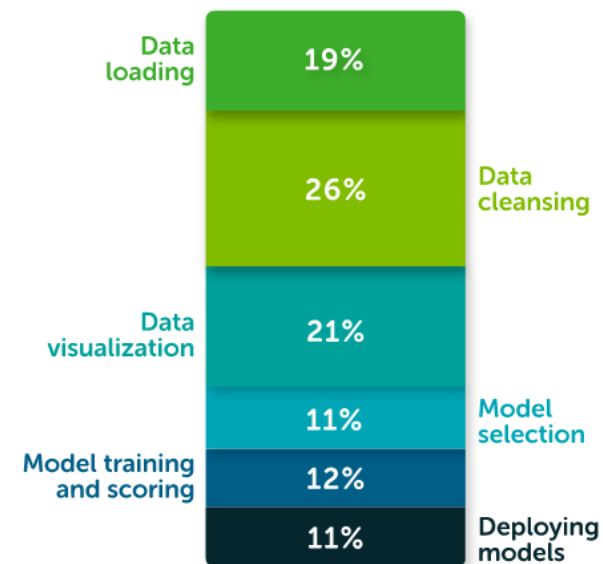
Who likes to clean anyway...



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%

THINKING ABOUT YOUR CURRENT JOB, HOW MUCH OF YOUR TIME IS SPENT IN EACH OF THE FOLLOWING TASKS?

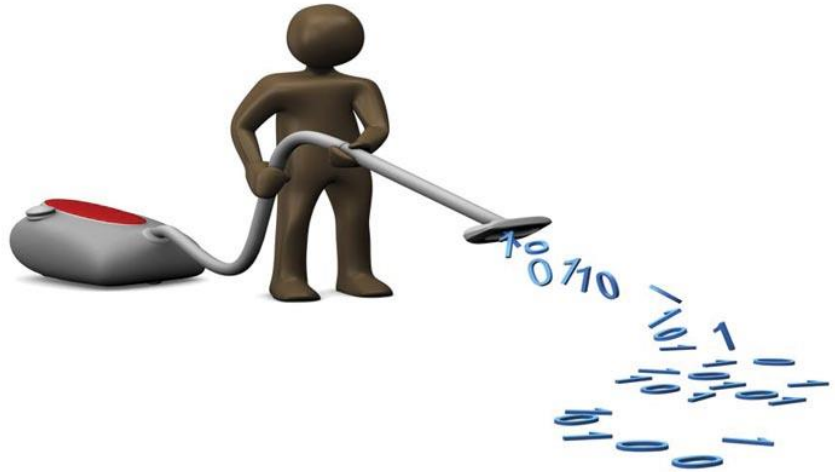


For most respondents, data management tasks still consume a disproportionate amount of work time.

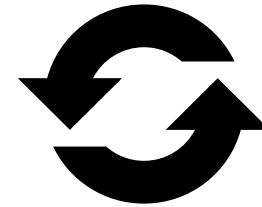
Source: Anaconda, State of Data Science 2020

Source: <https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/?sh=1b7a729a6f63>

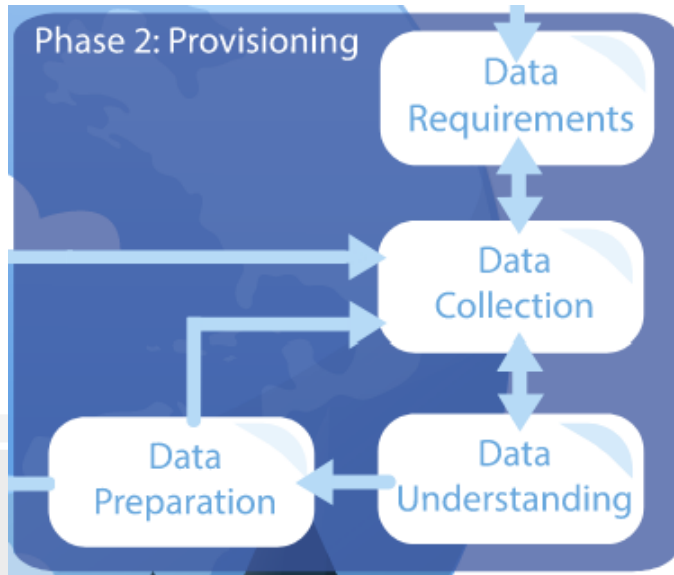
How to determine how 'clean' your data is...



How do you check your data?

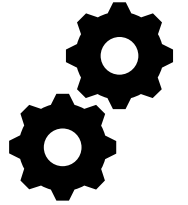


What (quality) criteria do you use?





Practise small: Be a lean clean machine...



- (2 min) Explore the dataset provided (see chat), use a tool of your liking...
- (5 min) Find & label data to clean using the quality criteria
- (1 min) Post your findings (labelled data which needs cleaning) in the chat.



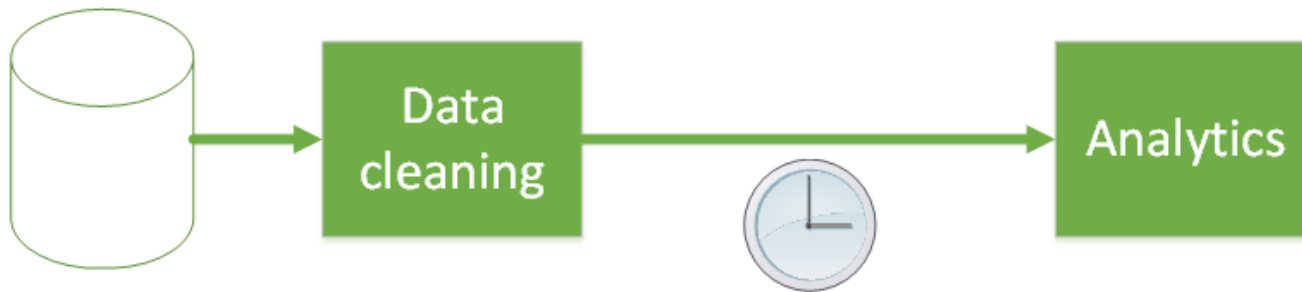


Found these?

- Whitespaces, new lines,
- Blank cells
- Fixing numbers that aren't numbers (excluding pivot tables)
- And many more problems:
 - multi-value fields
 - Various ways to express currencies, units, measures (USD, \$, M€)
 - Differences (mistake) in spelling, abbreviations, grammar (US/UK), capitals, notations/standards, ...
- Check: <https://schoolofdata.org/handbook/recipes/cleaning-data-with-spreadsheets/> (also on Canvas)

Why consider the process of data cleaning?

► *# try to 'fix' US\$77 million in the Projected Investment column*
`df['Projected investment'].replace('US$77 million', '77000000')`



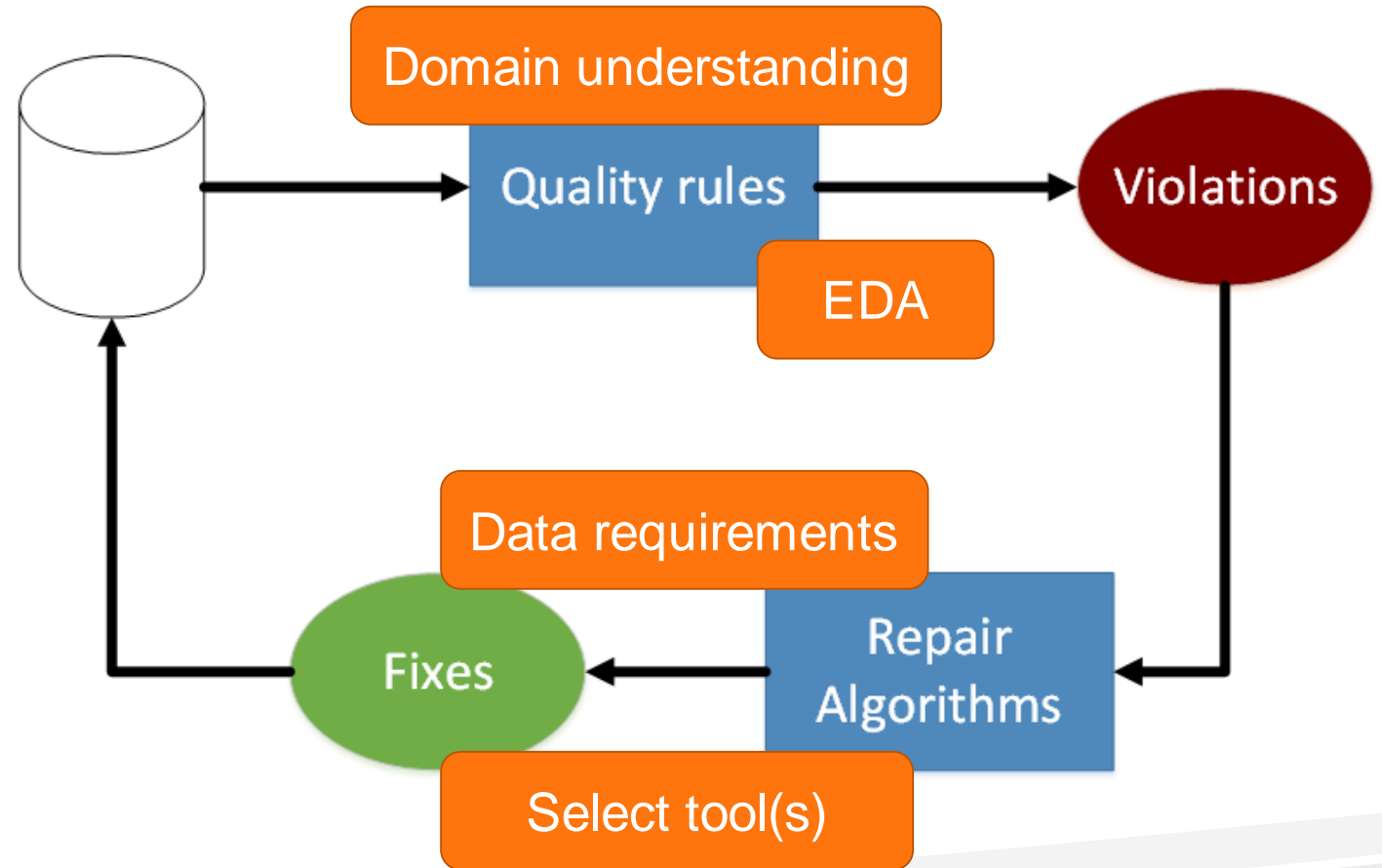
Data can change in:

- amount/size
- structure
- definitions/rules
- ...

Why consider the process of data cleaning?

Data can change in:


- amount/size
- structure
- definitions/rules
- ...





From violations to fixes

- Search for violations
- Check for patterns, (sub)sets, regularities
- Try to explain the violations (find the cause, not the effect)
- Create a fix (genericly)
- Assess/imagine the impact of the fix (for future cases)



Example: duplicates

- Search for violations:
- Check for patterns, (sub)sets, regularities
- Try to explain the violations (find the cause, not the effect)
- Create a fix (genericly)
- Assess/imagine the impact of the fix (for future cases)

```
df.duplicated()
```

```
df.duplicated().value_counts()
```

```
df.drop_duplicates(subset=None,  
                  keep='first',  
                  inplace=False,  
                  ignore_index=False)
```

Example: remove empty/white spaces

```
In [15]: ▶ # By checking the unique values we actually find duplicated values in 'Status of deal'
df['Status of deal'].unique()
```

```
Out[15]: array(['Done', 'Done ', 'In process', 'Done (50-yr lease)', 'Suspended',
               'Proposed', 'MoU signed (2009)', 'Done\n', 'Suspended ',
               'Suspended (October 2011)'], dtype=object)
```

```
In [20]: ▶ # among the string functions is strip() to remove all white spaces (and other 'invisible characters' )
df['Status of deal'] = df['Status of deal'].str.strip()
df['Status of deal'].unique()
```

```
Out[20]: array(['Done', 'In process', 'Done (50-yr lease)', 'Suspended',
               'Proposed', 'MoU signed (2009)', 'Suspended (October 2011)'],
               dtype=object)
```

How to deal with () in this case?

Check regular expressions to create ‘tailor made’ expressions to manipulate or search strings (e.g. <https://realpython.com/regex-python/>)

Example: multi-value columns

```
In [5]: ▶ # and check the first lines of data to have a first glance
df.head()
```

Out[5]:

	Landgrabbed	Landgrabber	Base	Sector	Hectares	Production	Projected investment	Status of deal	Summary
0	Algeria	Al Qudra	UAE	Finance, real estate	31000.0	Milk, olive oil, potatoes	NaN	Done	Al Qudra Holding is a joint-stock company esta...
1	Angola	CAMC Engineering Co. Ltd	China	Construction	1500.0	Rice	US\$77 million	Done	CAMCE is a subsidiary of the China National Ma...
2	Angola	ENI	Italy	Energy	12000.0	Oil palm	NaN	In process	The project is a joint venture between Sonango...
3	Angola	AfriAgro	Portugal	Finance, real estate	5000.0	Oil palm	US\$30-35 million	Done	AfriAgro is a subsidiary of the Portugal-based...
4	Angola	Eurico Ferreira	Portugal	Energy, telecommunications\n	30000.0	Sugar cane	US\$200 million	Done	In 2008, Portuguese conglomerate Eurico Ferrei...

Example: multivalue columns

```
In [22]: # let's try to split those values (first in a separate data frame)
subset = df['Sector'].str.split(',', expand = True)
subset.head(10)
```

Out[22]:

	0	1	2
0	Finance	real estate	None
1	Construction	None	None
2	Energy	None	None
3	Finance	real estate	None
4	Energy telecommunications		None
5	Agribusiness	energy	None
6	Agribusiness	None	None
7	Agribusiness	None	None
8	Agribusiness	None	None
9	Finance	None	None

Example: multivalue columns

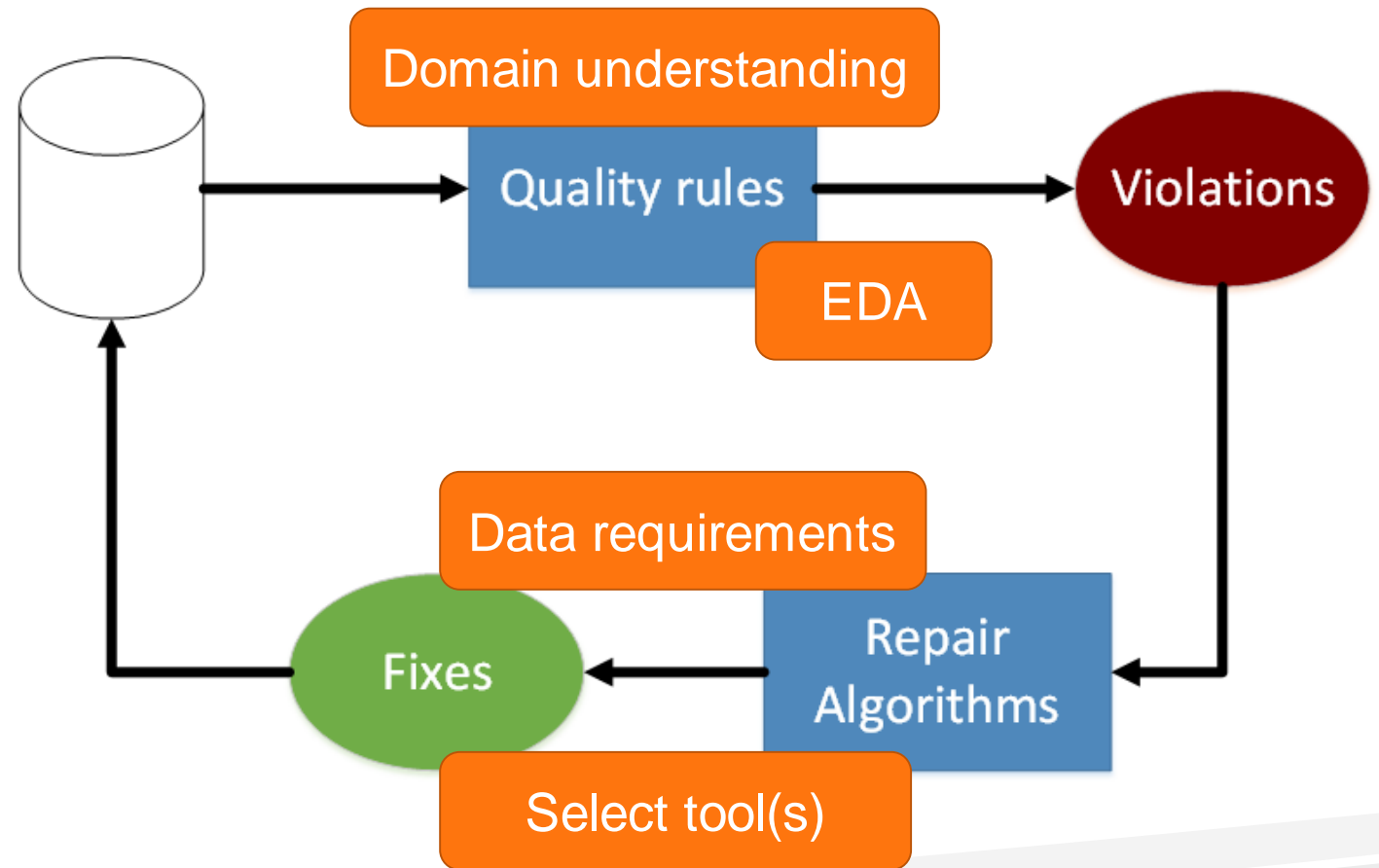
```
In [9]: # we can also include the newly split columns in the dataframe itself
df[['sector1', 'sector2', 'sector3']] = df['Sector'].str.split(',', expand = True)
# and delete the original column to clean the set (and prevent duplicate values)
df.drop('Sector', axis='columns', inplace=True)
df.head()
```

Out[9]:

	Landgrabbed	Landgrabber	Base	Hectares	Production	Projected investment	Status of deal	Summary	sector1	sector2	sector3
0	Algeria	Al Qudra	UAE	31000.0	Milk, olive oil, potatoes	NaN	Done	Al Qudra Holding is a joint-stock company esta...	Finance	real estate	None
1	Angola	CAMC Engineering Co. Ltd	China	1500.0	Rice	US\$77 million	Done	CAMCE is a subsidiary of the China National Ma...	Construction	None	None
2	Angola	ENI	Italy	12000.0	Oil palm	NaN	In process	The project is a joint venture between Sonango...	Energy	None	None
3	Angola	AfriAgro	Portugal	5000.0	Oil palm	US\$30-35 million	Done	AfriAgro is a subsidiary of the Portugal-based...	Finance	real estate	None
4	Angola	Eurico Ferreira	Portugal	30000.0	Sugar cane	US\$200 million	Done	In 2008, Portuguese conglomerate Eurico Ferrei...	Energy telecommunications\n		None

Always record/describe the process of data cleaning

- Data can change in:
- amount/size
 - structure
 - definitions/rules
 - ...





Some *don'ts* from practice...

- Don't delete (entire) rows before checking the effect on the data set
- Don't base your cleaning/fixes on the first n rows...
- Don't apply a standard set of 'fixes'
- Don't forget to explain your cleaning: include the quality criteria and violation/fix reasoning in your text/comments.
- Cleaning data can easily be confused with changing data
- ...

More on clean and fixing...

Check Canvas materials under 'Data Preparation'

Continue with the exercise, results will be discussed Thursday.

