Data Wrangling

Session Objective

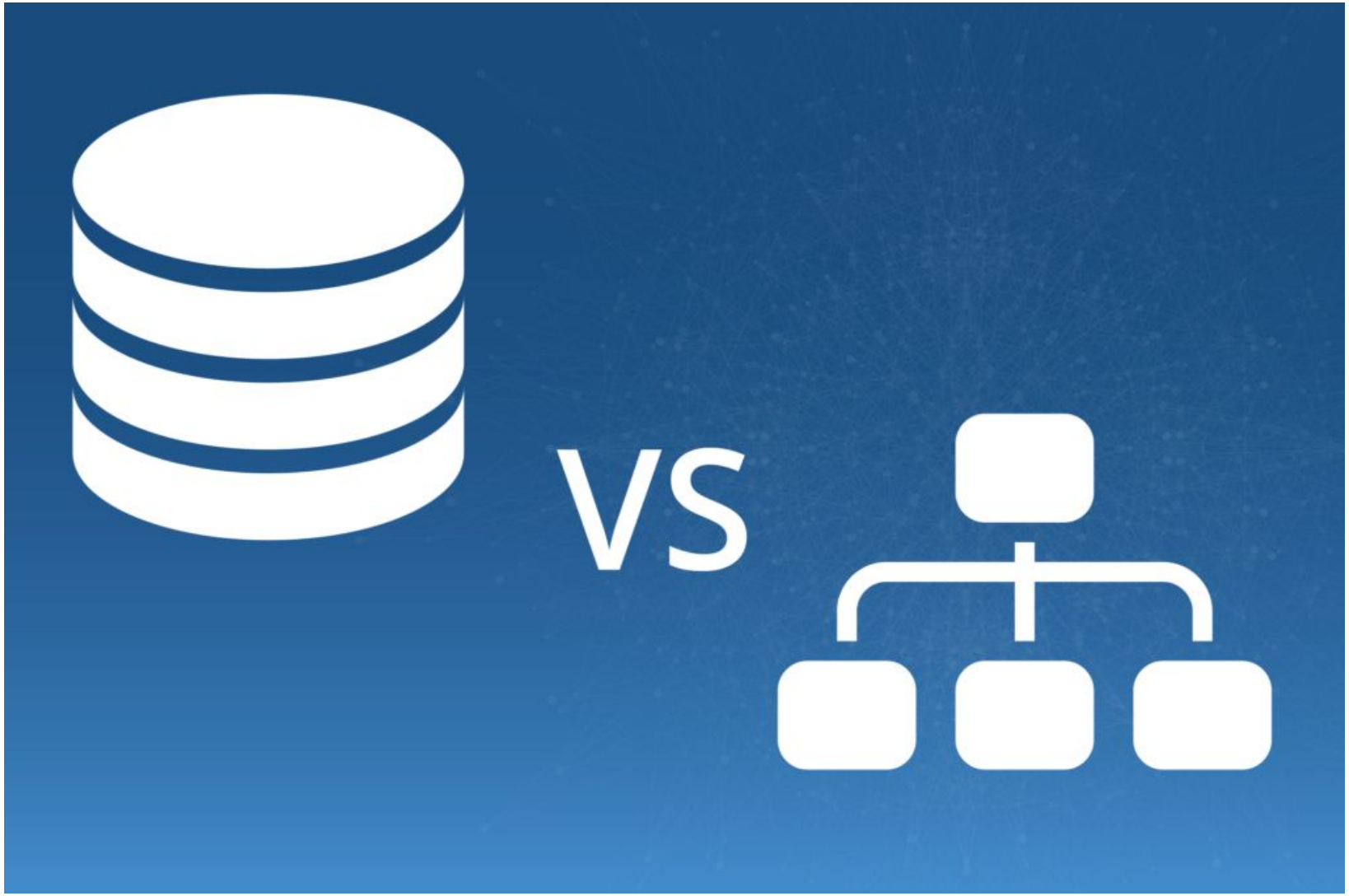
Understand why data wrangling is important

Understand common wrangling requirements and methods

Data Wrangling

Process of cleaning, structuring, and enriching raw data into a desired output for analysis

Difference between database system and file system?



Dirty Data...

- Data is dirty on its own
- Data sets are clean on their own but combining them introduces errors (e.g. duplicates, different naming conventions)
- Data doesn't "age well" (inflation, redistricting)
- Any combination of the above

Bad Data

All of these are commonly seen in the real-world:

- Zeros replace missing values
- Spelling inconsistent (esp with human-entered data)
- Rows are duplicated
- Inconsistent date formats (e.g. 10/4/20 vs. 4/10/20)
- Units not specified

https://github.com/Quartz/bad-data-guide

Big Data Borat
@BigDataBorat

Following

In Data Science, 80% of time spent prepare data, 20% of time spent complain about need for prepare data.

Data Wrangling

Process of cleaning, structuring, and enriching raw data into a desired output for analysis

What should we look for?

Key Data Properties to Consider for Wrangling

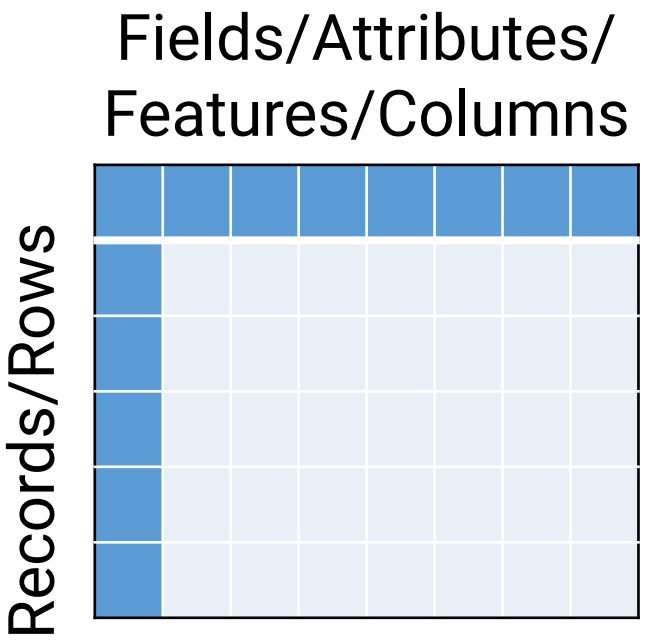
- **Structure --** *the “shape” of a data file*
- **Granularity –** *groupby/ Pivot table*
- **Faithfulness and Scope --** *how (in)complete is the data*

- **Structure --** *the “shape” of a data file*
- **Granularity –** *groupby/ pivot table*
- **Faithfulness and Scope --** *how (in)complete is the data*

Rectangular Data

We prefer rectangular data for data analysis (why?)

- Regular structures are easy manipulate and analyze
- A big part of data cleaning is about transforming data to be more rectangular



How are these data files formatted?

calls_for_service.tsv

```
CASENO  OFFENSE EVENTDT EVENTTM CVLEGEND    CVDOW   InDbDate    Block_Location
BLKADDR City    State
18000273    VEHICLE STOLEN  01/01/2018 12:00:00 AM  20:30   MOTOR VEHICLE THEFT
1   01/24/2018 03:30:18 AM  "1100 PARKER ST
Berkeley, CA
(37.859364, -122.288914)"   1100 PARKER ST  Berkeley    CA
17092476    BURGLARY AUTO   12/12/2017 12:00:00 AM  13:30   BURGLARY - VEHICLE
2   01/24/2018 03:30:17 AM  "2300 LE CONTE AVE
Berkeley
(37.8748
17092534
3   01/2
Berkeley
(37.8574
17091517
2   01/2
Berkeley
(37.8767
17048102
VEHICLE
```

Line 1, Column 1

TSV

Tab separated values

calls_for_service.csv — data

```
CASENO,OFFENSE,EVENTDT,EVENTTM,CVLEGEND,CVDOW,InDbDate,Block_Location,BLKADDR,City,Stat
e
18000273,VEHICLE STOLEN,01/01/2018 12:00:00 AM,20:30,MOTOR VEHICLE THEFT,1,01/24/2018
03:30:18 AM,"1100 PARKER ST
Berkeley, CA
(37.859364, -122.288914)",1100 PARKER ST,Berkeley,CA
17092476,BURGLARY AUTO,12/12/2017 12:00:00 AM,13:30,BURGLARY - VEHICLE,2,01/24/2018
03:30:17 AM,"2300 LE CONTE AVE
Berkeley, CA
(37.874867, -122.263689)",2300 LE CONTE AVE,Berkeley,CA
17092534,BURGLARY AUTO,
03:30:17 AM,"1700 STUAR
Berkeley, CA
(37.857495, -122.275256
17091517,THEFT MISD. (U
03:30:11 AM,"1600 CALIF
Berkeley, CA
(37.876791, -122.280472
17048102,THEFT FROM AUT
```

Line 1, Column 1

CSV

Comma separated values

Which is the best?

JSON

```
{
    "field1": "value1",
    "field2": ["list", "of", "values"],
    "myfield3": {"is_recursive": true, "a null value": null}
}
```

Line 5, Column 2 | 4 misspelled words | Spaces: 4 | JSON

Comma and Tab Separated Values Files

- Tabular data where
  - records are delimited by a *newline*: “\n”, “\r\n”
  - Fields are delimited by ‘,’ (comma) or ‘\t’ (tab)
- Very Common!
- Issues?
  - Commas, tabs in records
  - Quoting
  - …

calls_for_service.tsv

```
CASENO  OFFENSE EVENTDT EVENTTM CVLEGEND    CVDOW   InDbDate    Block_Location
BLKADDR City    State
18000273    VEHICLE STOLEN  01/01/2018 12:00:00 AM  20:30   MOTOR VEHICLE THEFT
1   01/24/2018 03:30:18 AM  "1100 PARKER ST
```

calls_for_service.csv — data

```
CASENO,OFFENSE,EVENTDT,EVENTTM,CVLEGEND,CVDOW,InDbDate,Block_Location,BLKADDR,City,State
18000273,VEHICLE STOLEN,01/01/2018 12:00:00 AM,20:30,MOTOR VEHICLE THEFT,1,01/24/2018 03:30:18 AM,"1100 PARKER ST
Berkeley, CA
(37.859364, -122.288914)",1100 PARKER ST,Berkeley,CA
17092476,BURGLARY AUTO,12/12/2017 12:00:00 AM,13:30,BURGLARY - VEHICLE,2,01/24/2018 03:30:17 AM,"2300 LE CONTE AVE
Berkeley, CA
(37.874867, -122.263689)",2300 LE CONTE AVE,Berkeley,CA
17092534,BURGLARY AUTO,12/20/2017 12:00:00 AM,05:00,BURGLARY - VEHICLE,3,01/24/2018 03:30:17 AM,"1700 STUART ST
Berkeley, CA
(37.857495, -122.275256)",1700 STUART ST,Berkeley,CA
17091517,THEFT MISD. (UNDER $950),08/01/2017 12:00:00 AM,00:30,LARCENY,2,01/24/2018 03:30:11 AM,"1600 CALIFORNIA ST
Berkeley, CA
(37.876791, -122.280472)",1600 CALIFORNIA ST,Berkeley,CA
17048102,THEFT FROM AUTO,08/13/2017 12:00:00 AM,00:40,LARCENY - FROM
```

JavaScript Object Notation (JSON)

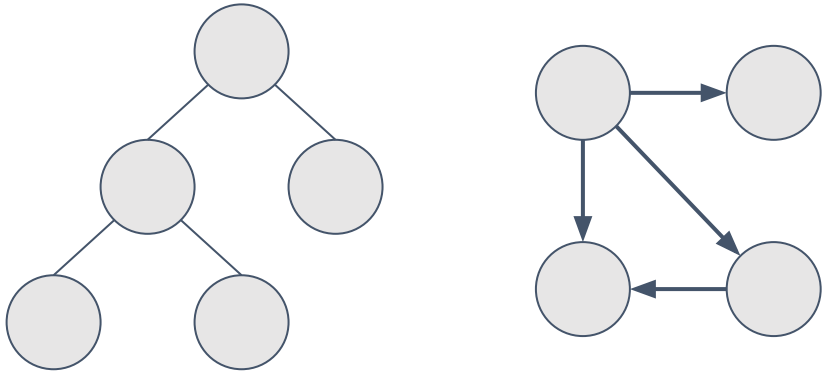
```
{
    "field1": "value1",
    "field2": ["list", "of", "values"],
    "myfield3": {"is_recursive": true, "a null value": null}
}
```

Line 5, Column 2 | 4 misspelled words | Spaces: 4 | JSON

- Widely used file format for nested data
  - Natural maps to python dictionaries (many tools for loading)
  - Strict formatting ”quoting” addresses some issues in CSV/TSV
- Issues
  - Each record can have different fields
  - Nesting means records can contain records complicated

JSON, XML, HTML, YAML, etc.

There are many formats to represent structured, nested data.

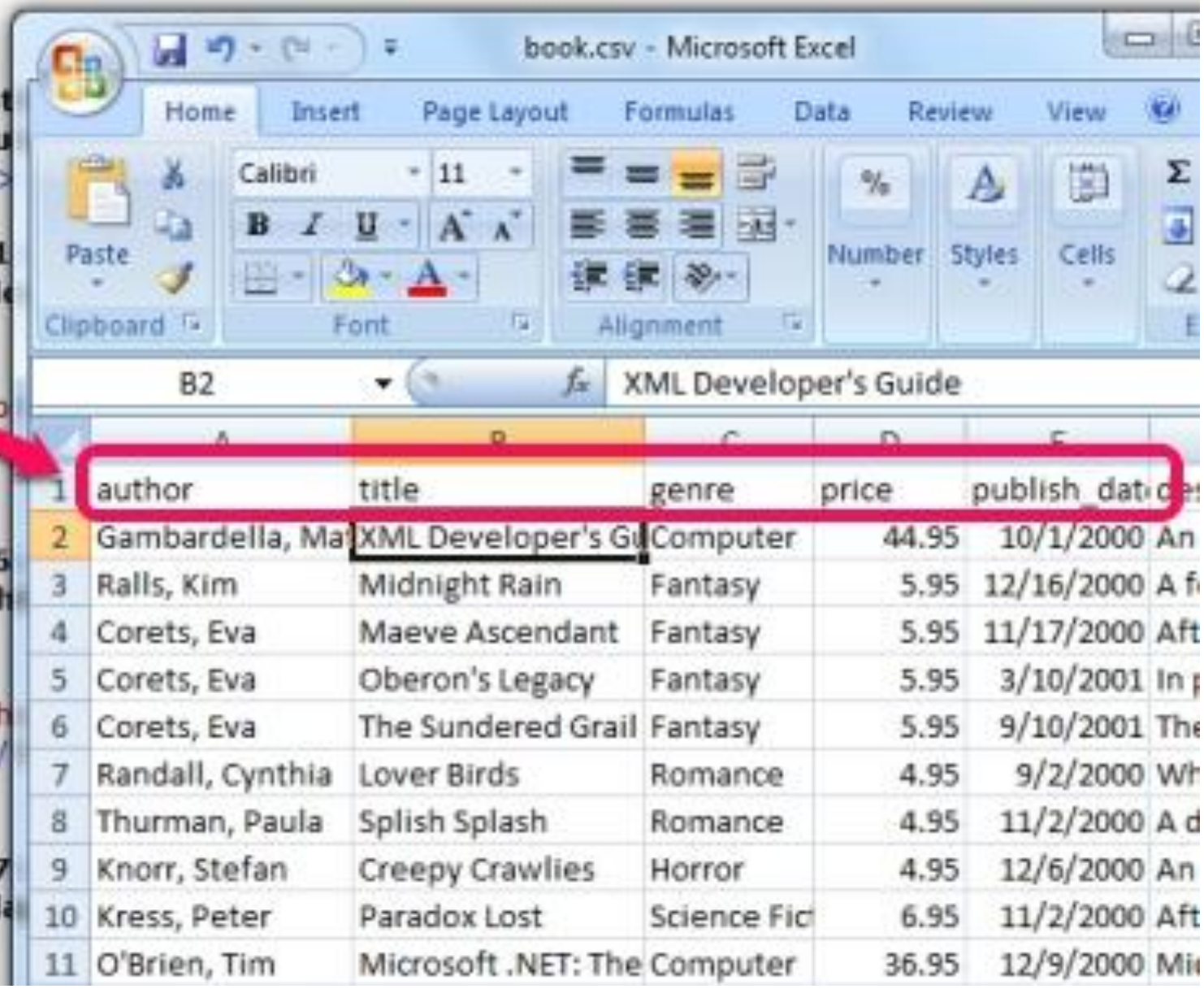


XML (kind of nested data)

```
<catalog>
  <plant type='a'>
    <common>Bloodroot</common>
    <botanical>Sanguinaria canadensis</botanical>
    <zone>4</zone>
    <light>Mostly Shady</light>
    <price>2.44</price>
    <availability>03/15/2006</availability>
    <description>
        <color>white</color>
        <petals>true</petals>
    </description>
    <indoor>true</indoor>
  </plant>
…
</catalog>
```

Nested structure

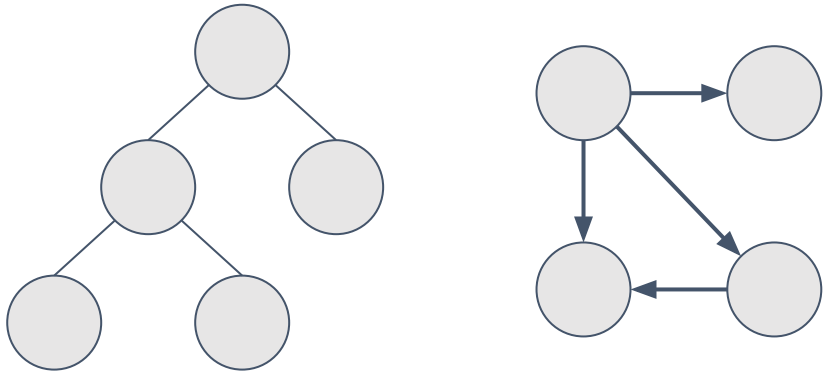
```
<?xml version="1.0"?>
<catalog>
  - <book id="bk101">
      <author>Gambardella, Mat
      <title>XML Developer's Gu
      <genre>Computer</genre>
      <price>44.95</price>
      <publish_date>2000-10-01
      <description>An in-depth l
    </book>
  - <book id="bk102">
      <author>Ralls, Kim</autho
      <title>Midnight Rain</ti
      <genre>Fantasy</genre>
      <price>5.95</price>
      <publish_date>2000-12-16
      <description>A former arch
    </book>
  - <book id="bk103">
      <author>Corets, Eva</auth
      <title>Maeve Ascendant</
      <genre>Fantasy</genre>
      <price>5.95</price>
      <publish_date>2000-11-17
      <description>After the colla
    </book>
  - <book id="bk104">
```



	author	title	genre	price	publish_dat	d
2	Gambardella, Ma	XML Developer's Gu	Computer	44.95	10/1/2000	An
3	Ralls, Kim	Midnight Rain	Fantasy	5.95	12/16/2000	A f
4	Corets, Eva	Maeve Ascendant	Fantasy	5.95	11/17/2000	Aft
5	Corets, Eva	Oberon's Legacy	Fantasy	5.95	3/10/2001	In p
6	Corets, Eva	The Sundered Grail	Fantasy	5.95	9/10/2001	The
7	Randall, Cynthia	Lover Birds	Romance	4.95	9/2/2000	Wh
8	Thurman, Paula	Splish Splash	Romance	4.95	11/2/2000	A d
9	Knorr, Stefan	Creepy Crawlies	Horror	4.95	12/6/2000	An
10	Kress, Peter	Paradox Lost	Science Fic	6.95	11/2/2000	Aft
11	O'Brien, Tim	Microsoft .NET: The	Computer	36.95	12/9/2000	Mi

JSON, XML, HTML, YAML, etc.

Converting hierarchical data to nested data often involves keys.



Log data

Is this a csv file? tsv? JSON/XML?

```
169.237.46.168 - - [26/Jan/2014:10:47:58 -0800] "GET
/stat141/Winter04 HTTP/1.1" 301 328
"http://anson.ucdavis.edu/courses/"  "Mozilla/4.0 (compatible; MSIE
6.0; Windows NT 5.0; .NET CLR 1.1.4322)”
```

```
169.237.6.168 - - [8/Jan/2014:10:47:58 -0800] "GET
/stat141/Winter04/ HTTP/1.1" 200 2585
"http://anson.ucdavis.edu/courses/" "Mozilla/4.0 (compatible; MSIE
6.0; Windows NT 5.0; .NET CLR 1.1.4322)"
```

Questions to ask about ***Structure***

- Are the data in a standard format?
  - **Tabular data:** CSV, TSV, Excel, SQL
  - **Nested data:** JSON or XML
- Are the data organized in “records”?
  - No: Can we define records by parsing the data?
- Are the data nested? (records contained within records…)
  - Yes: Can we reasonably un-nest the data?

Merging/Joining data across tables

Structure: Keys

- Often data will reference other pieces of data
- **Primary key:** *the column or set of columns in a table that determine the values of the remaining columns*
  - Primary keys are unique
  - Examples: SSN, ProductIDs, …
- **Foreign keys:** the column or sets of columns that reference primary keys in other tables.

Purchases.csv

OrderNum	ProdID	Quantity
1	42	3
1	999	2
2	42	1

Foreign Key

Orders.csv

OrderNum	CustID	Date
1	171345	8/21/2017
2	281139	8/30/2017

Products.csv

ProdID	Cost
42	3.14
999	2.72

Primary Key

Customers.csv

CustID	Addr
171345	Harmon..
281139	Main ..

Joining two tables

OrderNum	ProdID	Name
1	42	Gum
2	999	NullFood
2	42	Towel

X

OrderId	Cust Name	Date
1	Joe	8/21/2017
2	Arthur	8/14/2017

Left “key”

Right “key”

OrderNum	ProdID	Name	OrderId	Cust Name	Date
1	42	Gum	1	Joe	8/21/2017
~~1~~	~~42~~	~~Gum~~	~~2~~	~~Arthur~~	~~8/14/2017~~
~~2~~	~~999~~	~~NullFood~~	~~1~~	~~Joe~~	~~8/21/2017~~
2	999	NullFood	2	Arthur	8/14/2017
~~2~~	~~42~~	~~Towel~~	~~1~~	~~Joe~~	~~8/21/2017~~
2	42	Towel	2	Arthur	8/14/2017

Drop rows that don’t match on the key

OrderNum	ProdID	Name
1	42	Gum
2	999	NullFood
2	42	Towel

X

OrderId	Cust Name	Date
1	Joe	8/21/2017
2	Arthur	8/14/2017

Left “key”

Right “key”

OrderNum	ProdID	Name	OrderId	Cust Name	Date
1	42	Gum	1	Joe	8/21/2017
~~1~~	~~42~~	~~Gum~~	~~2~~	~~Arthur~~	~~8/14/2017~~
~~2~~	~~999~~	~~NullFood~~	~~1~~	~~Joe~~	~~8/21/2017~~
2	999	NullFood	2	Arthur	8/14/2017
~~2~~	~~42~~	~~Towel~~	~~1~~	~~Joe~~	~~8/21/2017~~
2	42	Towel	2	Arthur	8/14/2017

Drop rows that don’t match on the key

OrderNum	ProdID	Name	OrderId	Cust Name	Date
1	42	Gum	1	Joe	8/21/2017
2	999	NullFood	2	Arthur	8/14/2017
2	42	Towel	2	Arthur	8/14/2017

Pandas Merge

- **Structure --** *the "shape" of a data file*
- **Granularity –** *groupby/ pivot table*
- **Faithfulness and Scope --** *how (in)complete is the data*

Granularity

- What does each record represent?
  - Examples: a purchase, a person, a group of users
- Do all records capture granularity at the same level?
  - Some data will include summaries as records
- If the data are coarse how was it aggregated?
  - Sampling, averaging, …
- What kinds of aggregation is possible/desirable?
  - From individual people to demographic groups?
  - From individual events to totals across time or regions?
  - Hierarchies (city/county/state, second/minute/hour/days)
- Understanding and manipulating granularity can help reveal patterns.

Granularity and Keys

- The primary key defines what the record represents → Granularity
- What is the granularity of these example tables?
  - Purchases.csv: PK=(OrderNum + ProdID) → Each Item in an order
  - Orders.csv: PK = OrderNum → an order
- How might we adjust the granularity?
  - Aggregation: count, mean, median, var, groupby, pivot …

Purchases.csv

OrderNum	ProdID	Quantity
1	42	3
1	999	2
2	42	1

Orders.csv

OrderNum	CustID	Date
1	171345	8/21/2017
2	281139	8/30/2017

Products.csv

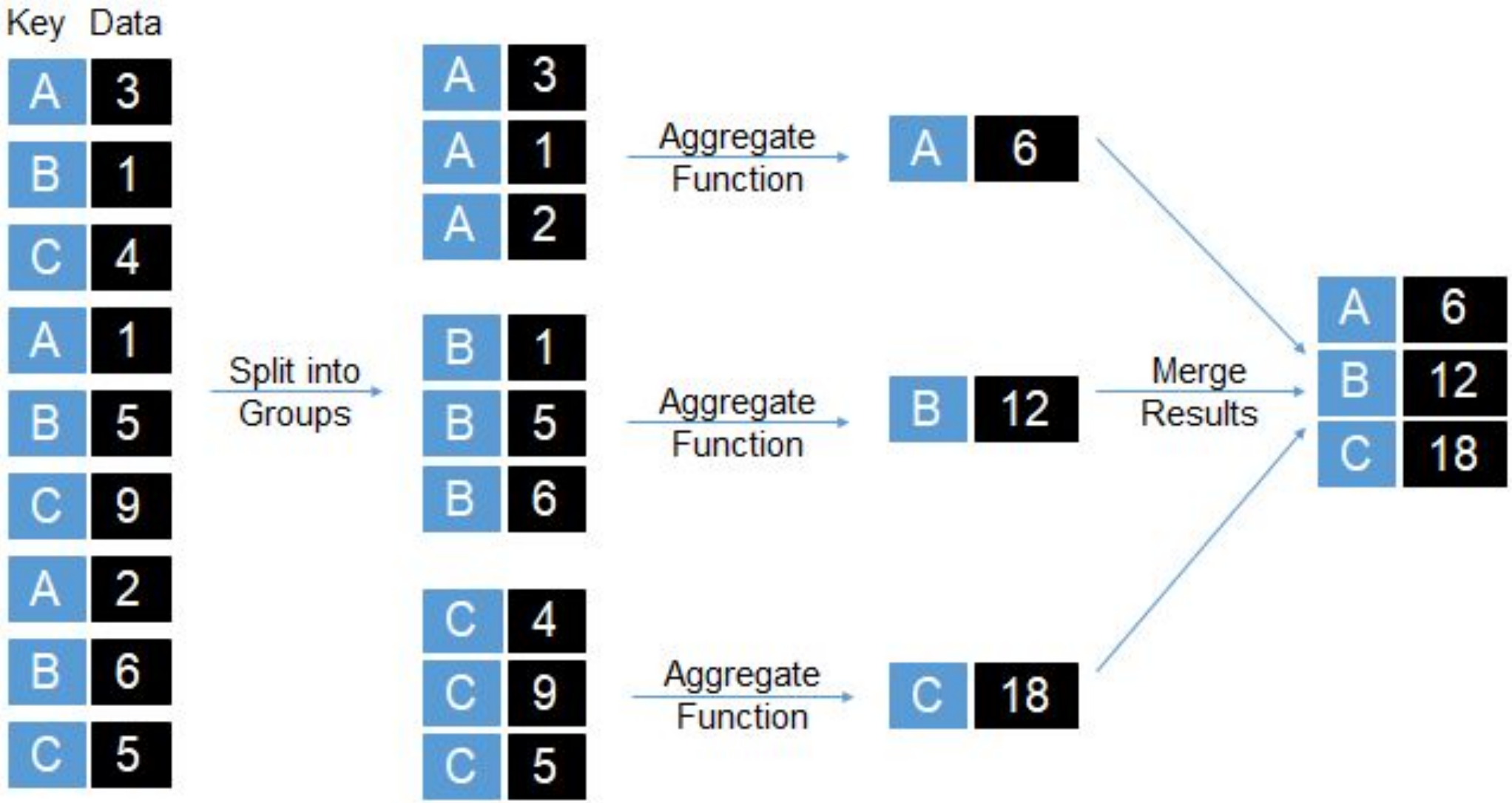
ProdID	Cost
42	3.14
999	2.72

Customers.csv

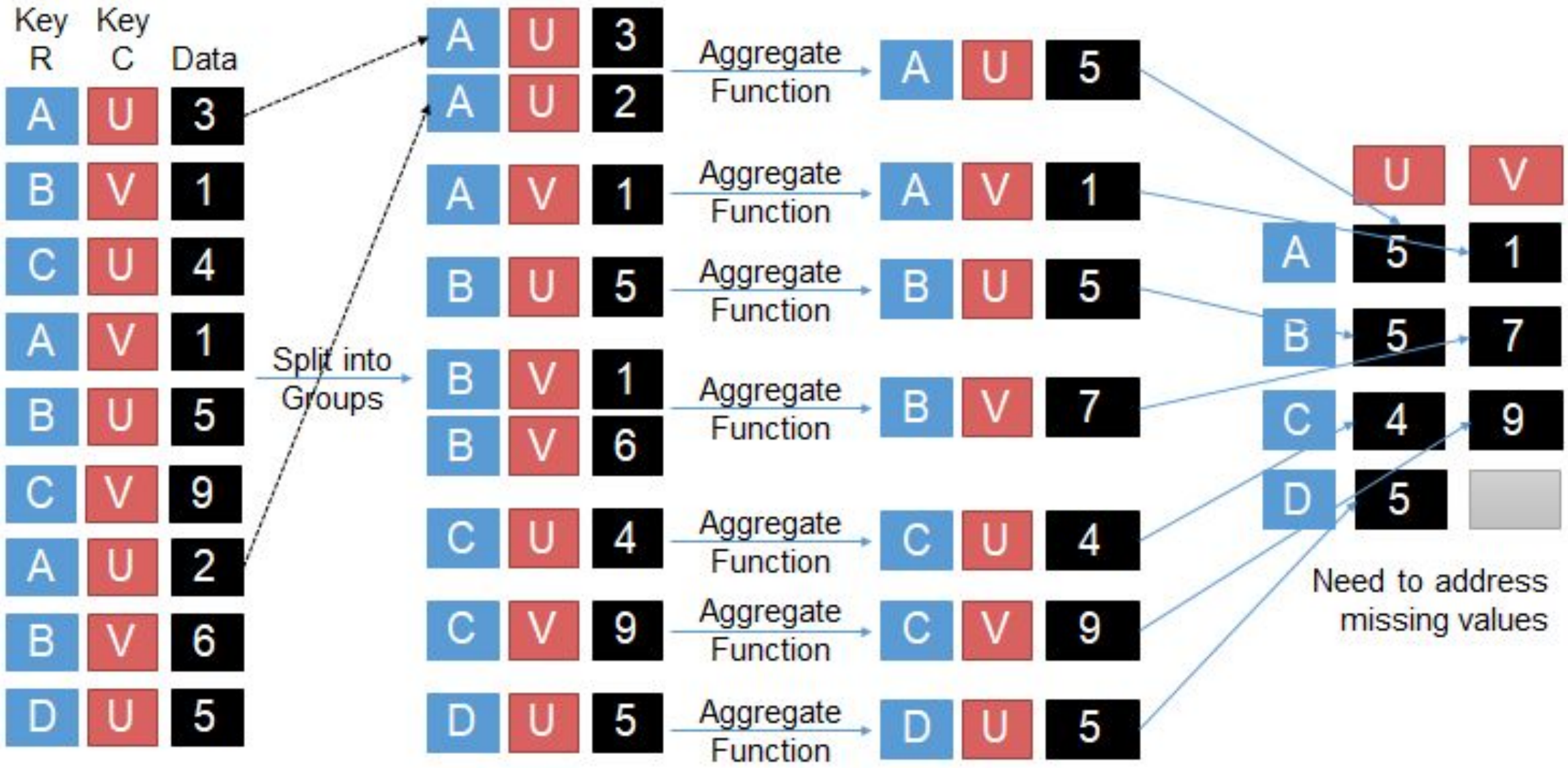
CustID	Addr
171345	Harmon..
281139	Main ..

Groupby and Pivot

Group By – manipulating granularity



Pivot – A kind of Group By Operation



- **Structure --** *the "shape" of a data file*
- **Granularity --** *how fine/coarse is each datum*
- **Faithfulness and Scope --** *how (in)complete is the data*

What to do with Missing Values

Often, rectangular data sets have missing values:

- Field lost, hidden, removed, replaced, or never entered.
- Or, perhaps the entity described by a record does not have a particular attribute.
  E.g., some people don’t have a permanent address.

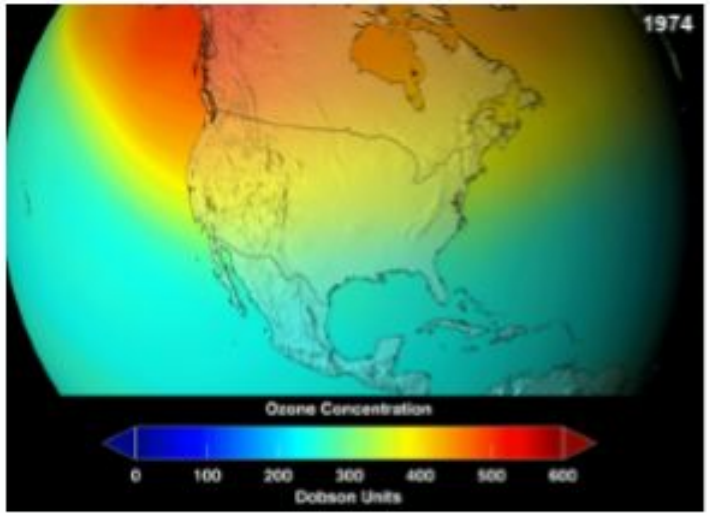
What to do with Missing Values

How to treat missing values depends on goals and context.

Discarding whole records (because of missing values) results in a sample.

- That sample isn’t a random sample, so statistical inference is inappropriate.
- The sample will often be biased — not representative of the population.

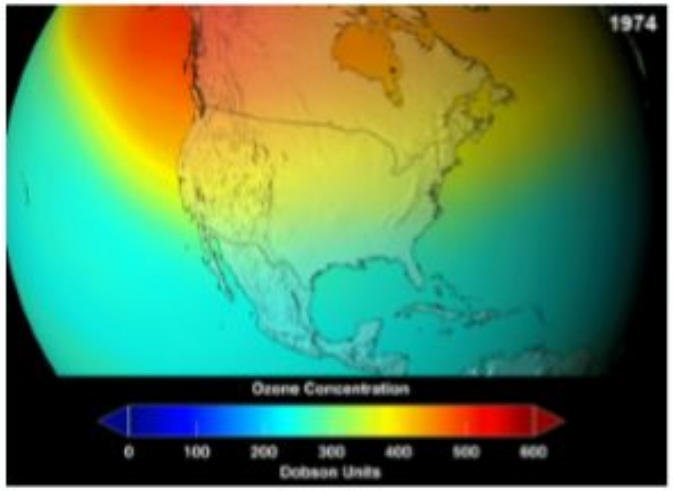
Set Defaults/Remove Outliers



The discovery of the Antarctic "ozone hole" by British Antarctic Survey scientists Farman, Gardiner and Shanklin…came as a shock to the scientific community…[The data] were initially rejected as unreasonable by data quality control algorithms (they were filtered out as errors since the values were unexpectedly low); the ozone hole was detected only in satellite data when the raw data was reprocessed following evidence of ozone depletion in in situ observations. When the software was rerun without the flags, the ozone hole was seen as far back as 1976.

https://en.wikipedia.org/wiki/Ozone_depletion#Antarctic_ozone_hole

Set Defaults/Remove Outliers



The discovery of the Antarctic "ozone hole" by British Antarctic Survey scientists Farman, Gardiner and Shanklin…came as a shock to the scientific community…[The data] were initially rejected as unreasonable by data quality control algorithms (they were filtered out as errors since the values were unexpectedly low); the ozone hole was detected only in satellite data when the raw data was reprocessed following evidence of ozone depletion in in situ observations. When the software was rerun without the flags, the ozone hole was seen as far back as 1976.

https://en.wikipedia.org/wiki/Ozone_depletion#Antarctic_ozone_hole

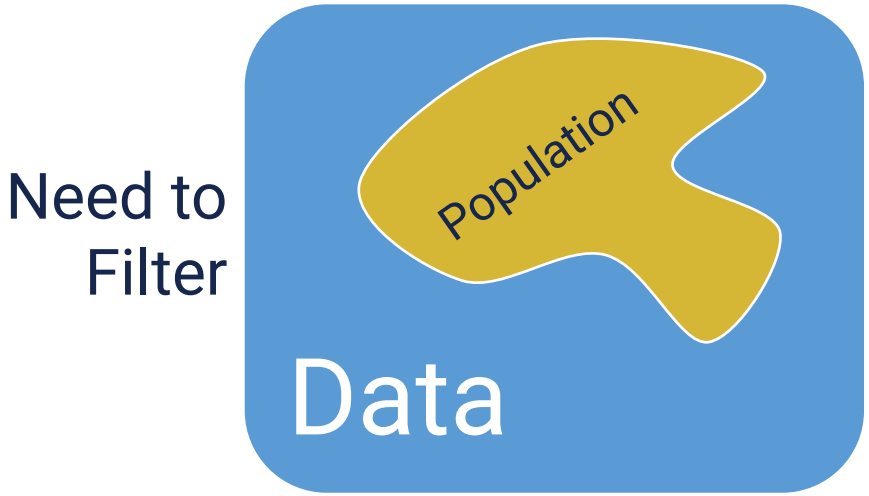
Signs that your data may not be faithful

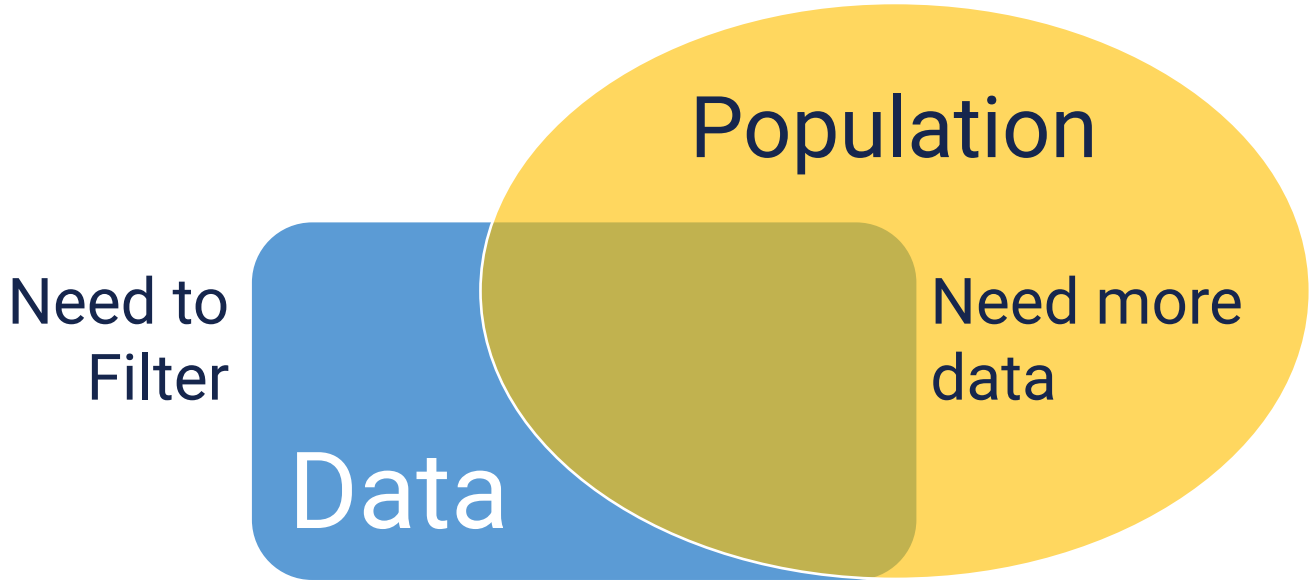
- Missing Values/Default values: (0, -1, 999, 12345, NaN, Null, 1970, 1900, … others?)
  - **Soln 1:** Drop records with missing values implications on your sample!
  - **Soln 2:** Impute missing values Bias your conclusions
- Time Zone Inconsistencies
  - **Soln 1:** convert to a common timezone (e.g., UTC)
  - **Soln 2:** convert to the timezone of the location – useful in modeling behavior.
- Duplicated Records or Fields
  - **Soln:** identify and eliminate (use primary key) implications on sample?
- Spelling Errors
  - **Soln:** Apply corrections or drop records not in a dictionary implications on sample?
- Units not specified or consistent
  - **Solns**: Infer units, check values are in reasonable ranges for data
- Truncated data (early excel limits: 65536 Rows, 255 Columns)
  - **Soln:** be aware of consequences in analysis how did truncation affect sample?
- Others…

Scope

- Does my data cover my area of interest?
  - **Example:** *I am interested in studying crime in California but I only have Berkeley crime data.*
- Is my data too expansive?
  - **Example:** *I am interested in student grades for Data but have student grades for all apprentice classes.*
  - **Solution:** *Filtering → Implications on sample?*
    - *If the data is a sample I may have poor coverage after filtering …*

Scope


Need to
Filter
Population
Data


Population
Need to
Filter
Need more
data
Data