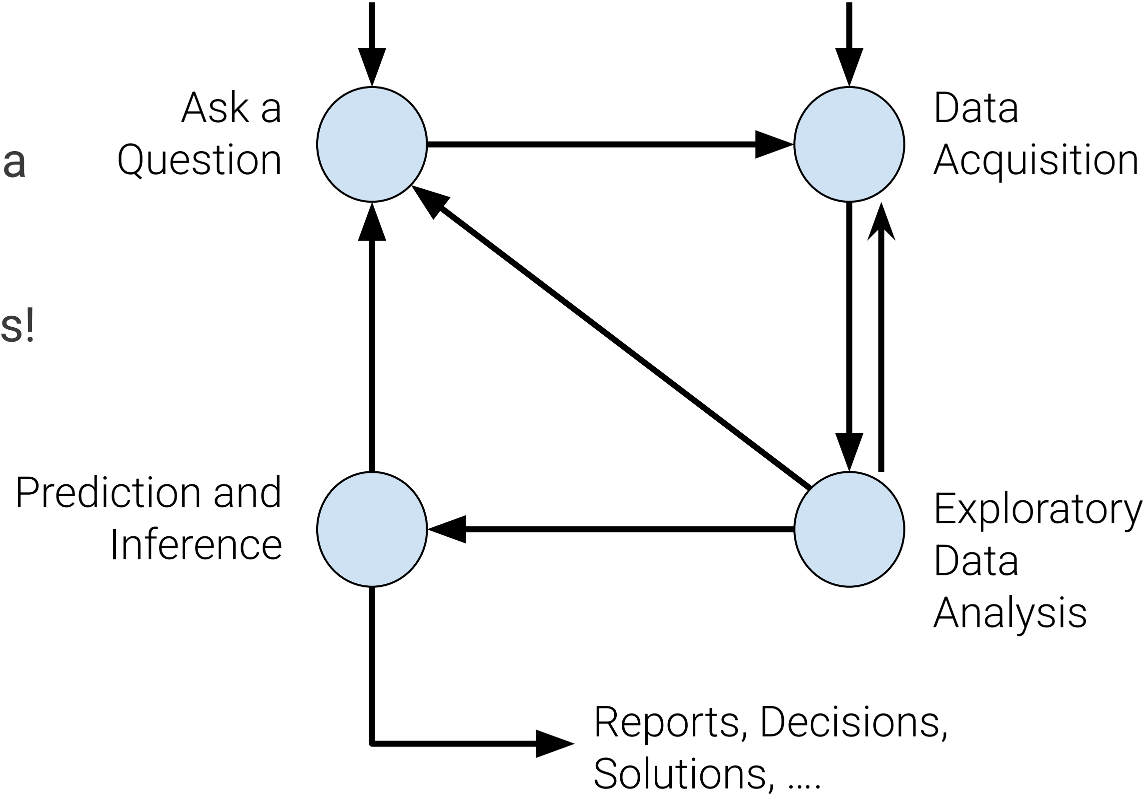
Data Science Lifecycle

The data science lifecycle is a **high-level description** of the data science workflow.

Note the two distinct entry points!

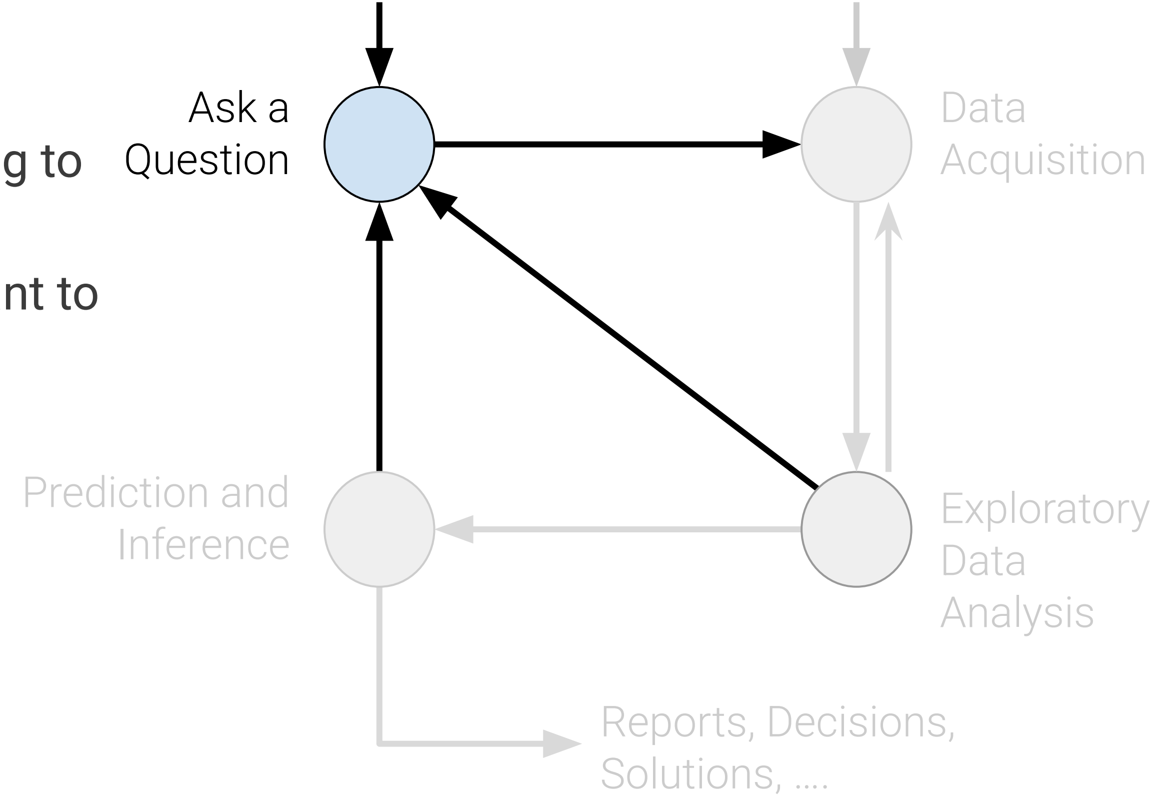


Credit: Learning Data Science

3531768

1. Problem Formulation

- What do we want to know?
- What problems are we trying to solve?
- What hypotheses do we want to test?
- What are our metrics for success?

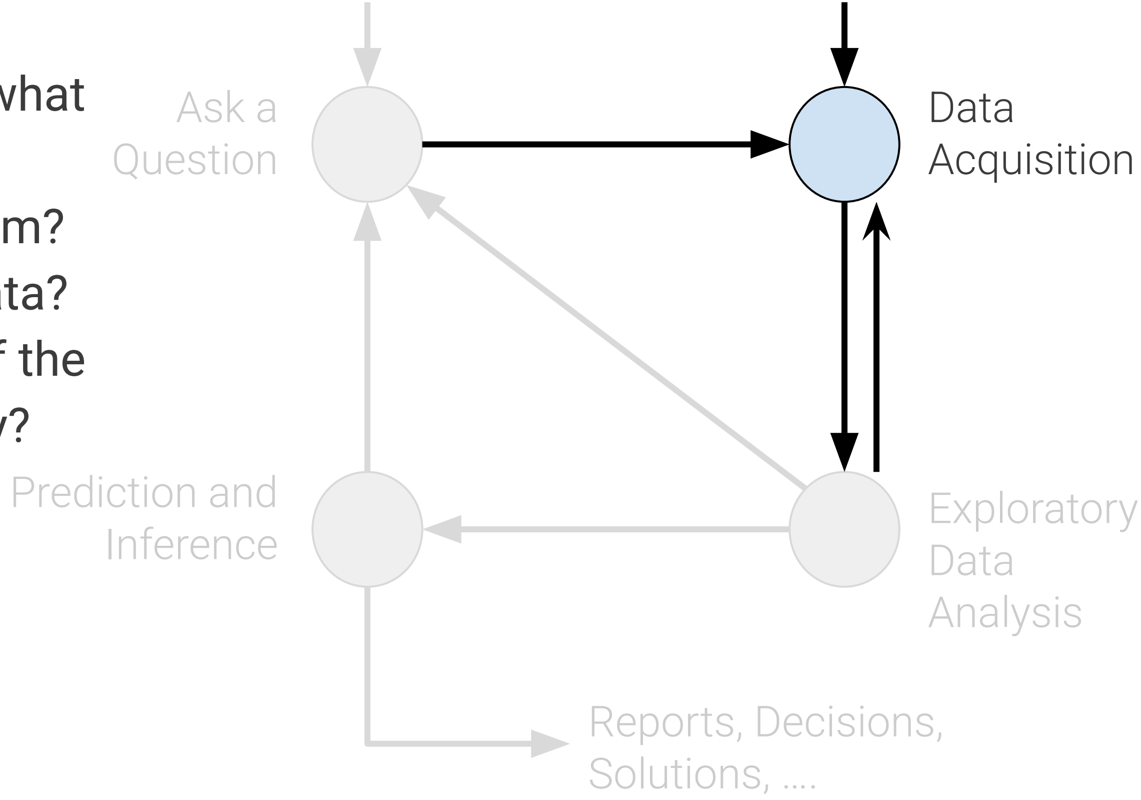


Credit: Learning Data Science

3531768

2. Data Acquisition and Cleaning

- What data do we have and what data do we need?
- Where do the data come from?
- How will we collect more data?
- Is our data representative of the population we want to study?

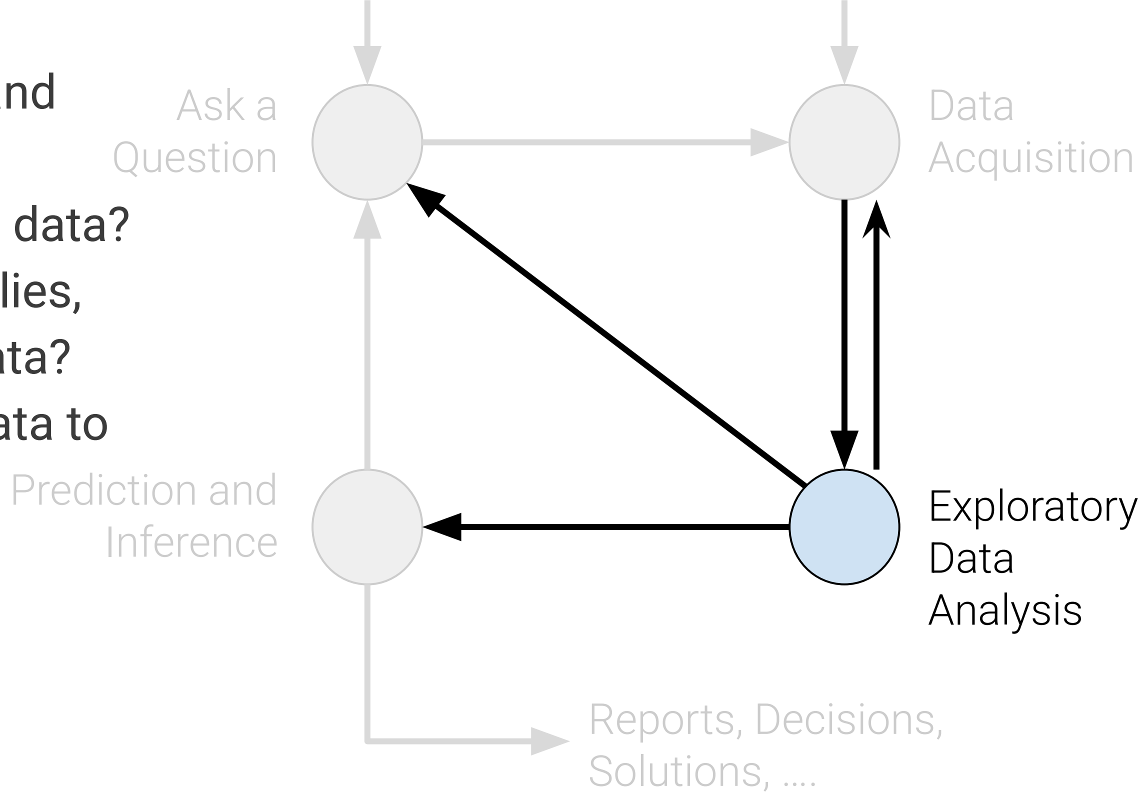


Credit: Learning Data Science

3. Exploratory Data Analysis (EDA) & Visualization

3531768

- How is our data organized and what does it contain?
- Do we already have relevant data?
- What are the biases, anomalies, and other issues with the data?
- How do we transform the data to enable effective analysis?

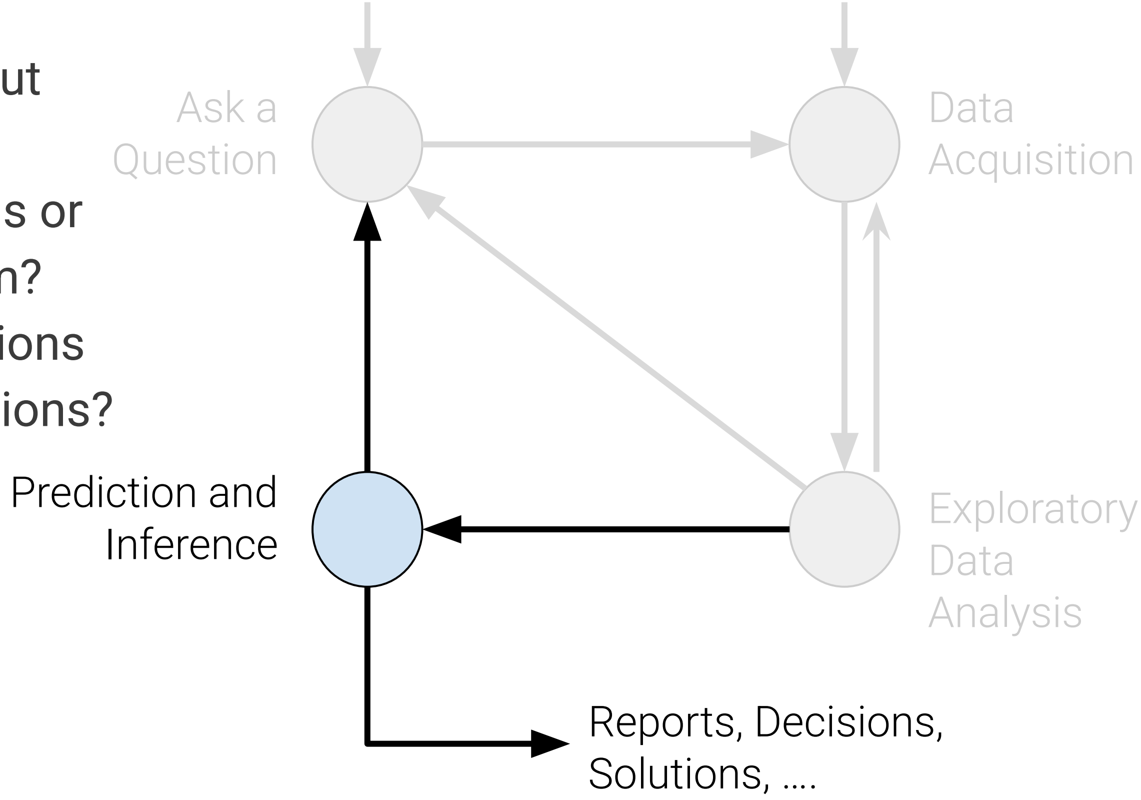


Credit: Learning Data Science

3531768

4. Prediction and Inference

- What does the data say about the world?
- Does it answer our questions or accurately solve the problem?
- How robust are our conclusions and can we trust the predictions?



Credit: Learning Data Science

Recall: Data Science Lifecycle

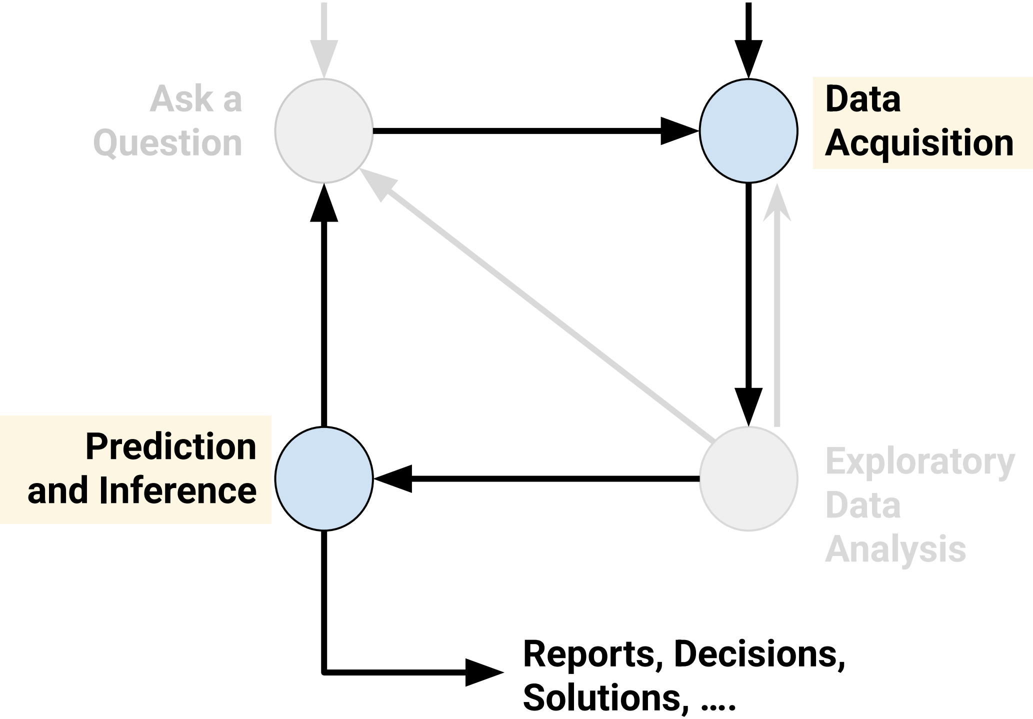
So far, we have focused on EDA.

But how do we collect data?

How does understanding data collection help us understand the world?

EDA: Understand the data.

Prediction/Inference: Understand the world.



Credit: Learning Data Science

3589828

Sampling from a finite population

A census is great, but expensive and difficult to execute.

- Would all voters be willing to participate in a voting census prior to an actual election?

A **sample** is (usually) a subset of the population.

- Samples are often used to make **inferences about the population.**
- How you draw the sample will affect your accuracy.
- Two common sources of error:
  - **Chance error:** random samples can vary from what is expected, in any direction.
  - **Bias:** a systematic error in one direction. Could come from our sampling scheme and survey methods.

Population, sample, and sampling frame

Population: The group that you want to learn something about.

Sampling Frame: The list from which the sample is drawn.

- If you’re sampling people, the sampling frame is the set of all people that could possibly end up in your sample.

Sample: Who you actually end up sampling.

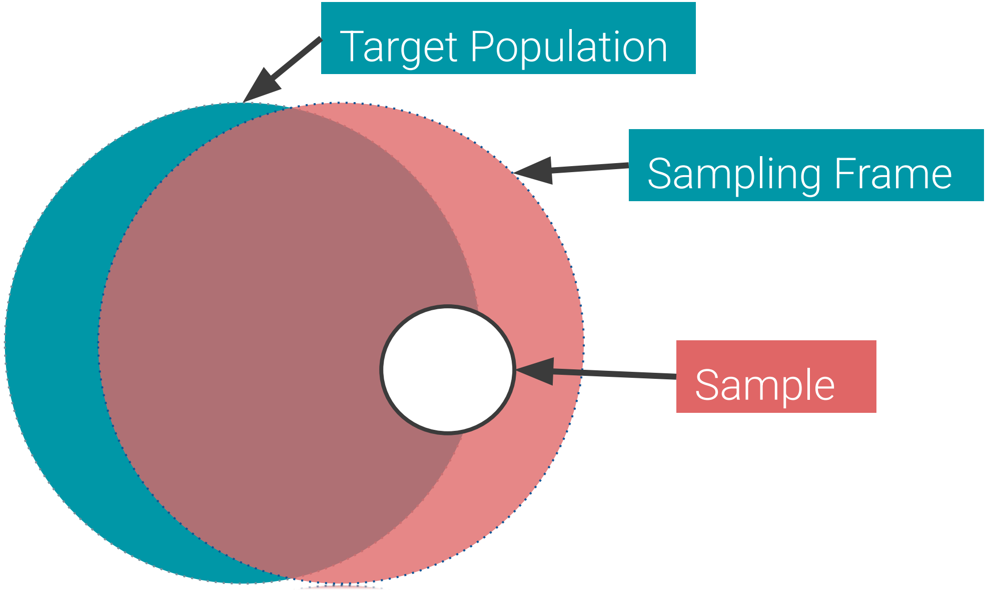
- A subset of your sampling frame.

3589828

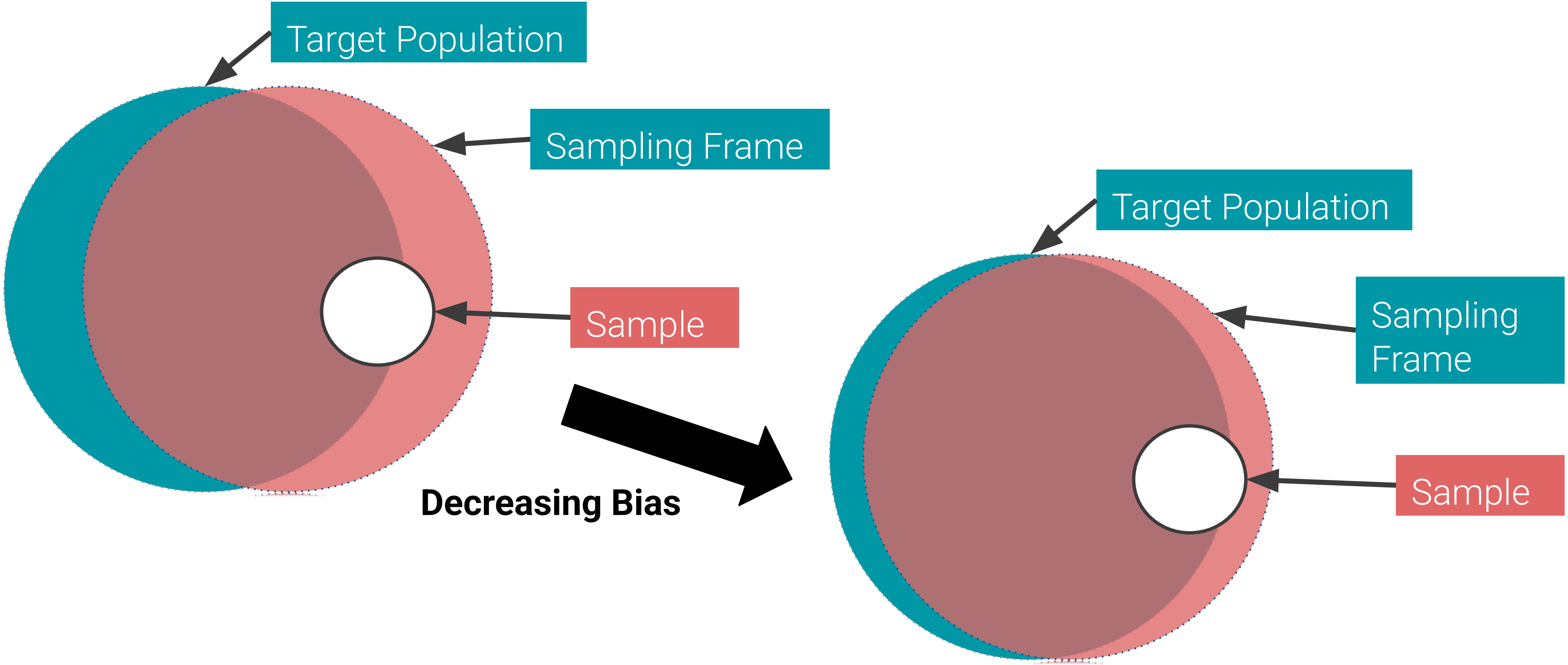
Population, sample, and sampling frame

There may be individuals in your **sampling frame** (and hence, your sample) that are **not** in your population!

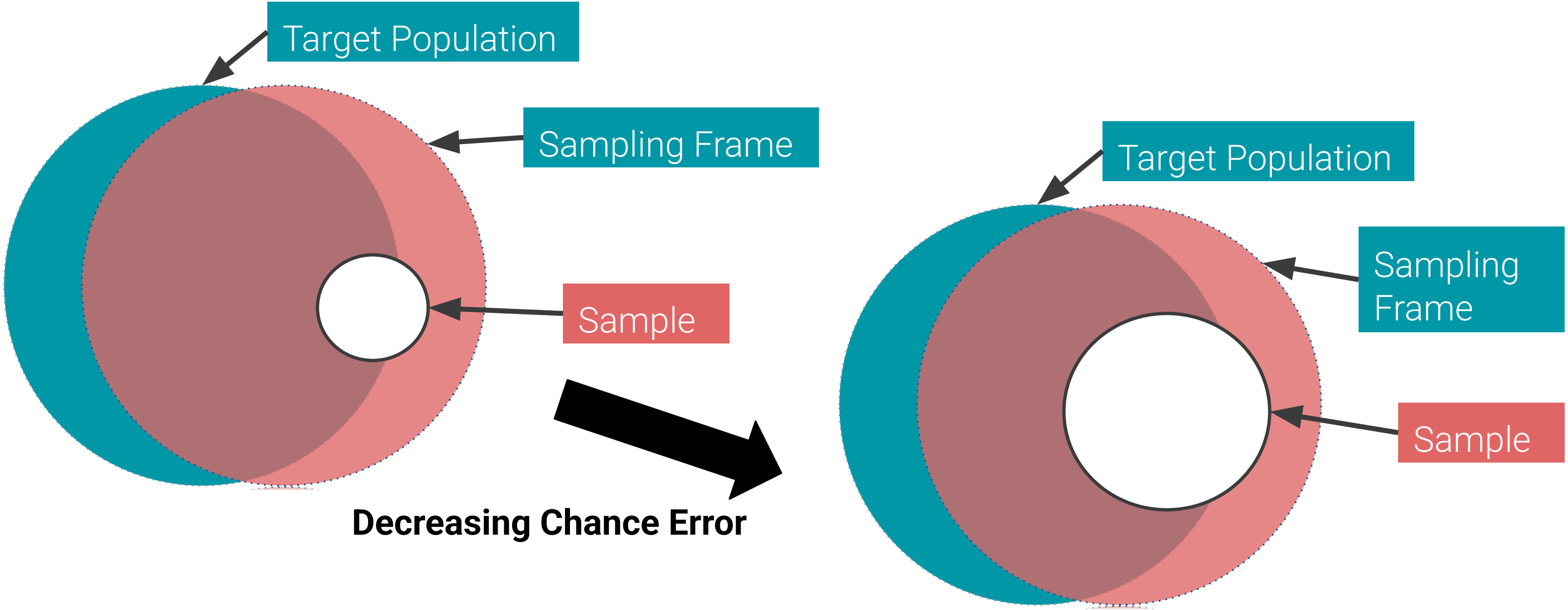
Similarly, there might be individuals in your target population that are not in your sampling frame.



Bias and sampling frames



Bias and sampling frames

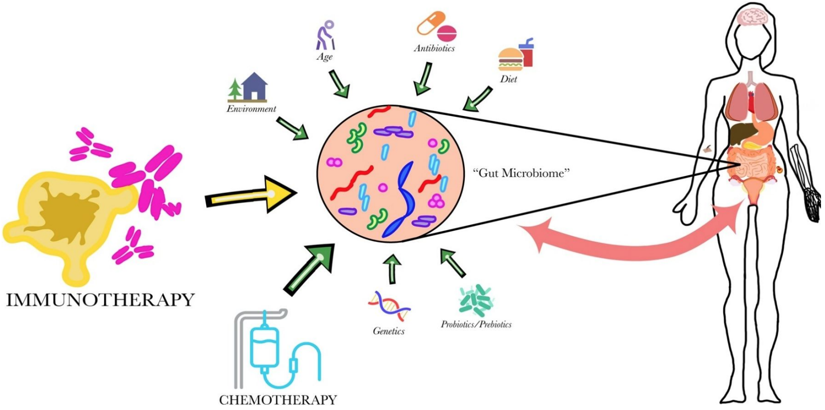


3589828

Other kinds of populations

Individuals in a population are not always people! Could be:

- Bacteria in your gut (sampled using DNA sequencing)
- Trees of a certain species
- Small businesses receiving a microloan
- Published results in a journal / field (example)



In any of these cases we might examine a sample and try to draw an **inference** about the population it came from.

- Simplest example: what % have some property (voting intention for candidate A)?

Probability Samples

Quality, not quantity!

A **huge sample size** does not fix a **bad sampling method**!

We want the sample to be **representative** of the population.

Think about **tasting soup**: if it’s **well-stirred**, a spoonful is all you need!

- Don’t just try to get a BIG sample. If your method of sampling is BAD, and your sample is BIG, what you’ll have is a BIG BAD sample

Easiest way to to get a representative sample is by using **randomness**.

Convenience sampling

Example: stand at UM breezeway and take first ten people who pass by

We call this a **convenience sample**. It’s whoever we can get ahold of.

Question: Is this a random sample?

Convenience sampling

Example: stand at UM breezeway and take first ten people who pass by

We call this a **convenience sample**. It’s whoever we can get ahold of.

Question: Is this a random sample? **No!**

Just because you think you’re sampling “randomly” doesn’t mean you have a random sample.

Probability sample (aka random sample)

For a **probability sample**:

- We have to be able to provide…
  1. the population
  2. the **chance** of selection, for each **group** in the population
- All individuals in the population **need not** have the same chance of being selected.
- Because we know all the probabilities, we will be able to **measure the errors**.

Probability sample (aka random sample)

For a **probability sample**:

- We have to be able to provide…
  1. the population
  2. the **chance** of selection, for each **group** in the population
- All individuals in the population **need not** have the same chance of being selected.
- Because we know all the probabilities, we will be able to **measure the errors**.

Why are we doing this?

- Get more representative samples → **reduce bias**
  - Random samples **can** produce biased estimates of population quantities.
- We can **estimate** the bias and chance error → **quantify uncertainty!**

3589828

Example Scheme 1: Probability Sample

Suppose I have 3 TA's (**A**lan, **B**ennett, **C**eline):
I decide to sample 2 of them as follows:

- I choose **A** with probability 1.0
- I choose either **B** or **C**, each with probability 0.5.

All subsets of 2:	{**A**, **B**}	{**A**, **C**}	{**B**, **C**}
Probabilities:	0.5	0.5	0

This is a **probability sample** (though not a great one).

- Of the 3 people in the population, I know the chance of getting each subset.
  - This scheme does not see the entire population!
  - My estimate using the single sample I take has some **chance error** depending on if I see AB or AC.
  - This scheme **biases** towards A's response

Common random sampling schemes

A **random sample with replacement** is a sample drawn **uniformly** at random **WITH** replacement.

- Random doesn’t always mean “uniformly at random,” but in this specific context, it does.
- Some individuals in the population might get picked more than once.

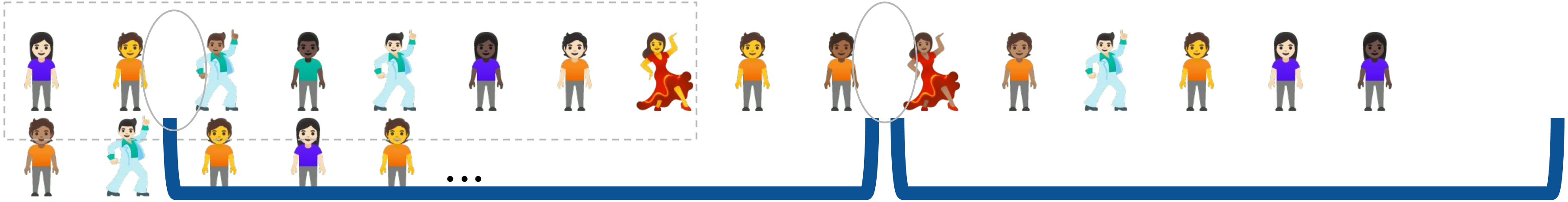
A **simple random sample (SRS)** is a sample drawn **uniformly** at random **WITHOUT** replacement.

- Every individual (and subset of individuals) has the same chance of being selected.
- Every pair has the same chance as every other pair.
- Every triple has the same chance as every other triple.
- And so on.

Example: Simple Random Sample?

We have the following sampling scheme:

- Computer Science (CSC) roster has 800 students listed alphabetically.
- Pick one of the first 10 students on the list at random (e.g. Student 8).
- To create your sample, take that student and every 10th student listed after that (e.g. Students 8, 18, 28, 38, etc).



Example: Simple Random Sample?

We have the following sampling scheme:

- Computer Science (CSC) roster has 800 students listed alphabetically.
- Pick one of the first 10 students on the list at random (e.g. Student 8).
- To create your sample, take that student and every 10th student listed after that (e.g. Students 8, 18, 28, 38, etc).

Questions:

1. Is this a **probability sample**?
2. Does each student have the same probability of being selected?
3. Is this a **simple random sample**?

Example: Simple Random Sample?

We have the following sampling scheme:

- Computer Science (CSC) roster has 800 students listed alphabetically.
- Pick one of the first 10 students on the list at random (e.g. Student 8).
- To create your sample, take that student and every 10th student listed after that (e.g. Students 8, 18, 28, 38, etc).

1. Is this a **probability sample**?
   - **Yes.** For a sample [n, n + 10, n + 20, …, n + 790], where n is between 1 and 10, the probability of that sample is 1/10. Otherwise, 0. Only 10 possible samples!
2. Does each student have the same probability of being selected?
   - **Yes.** Each student is chosen with same probability (1/10).
3. Is this a **simple random sample**?
   - **No.** Chance of selecting (8, 18) is not the same as the chance of selecting (8, 9).

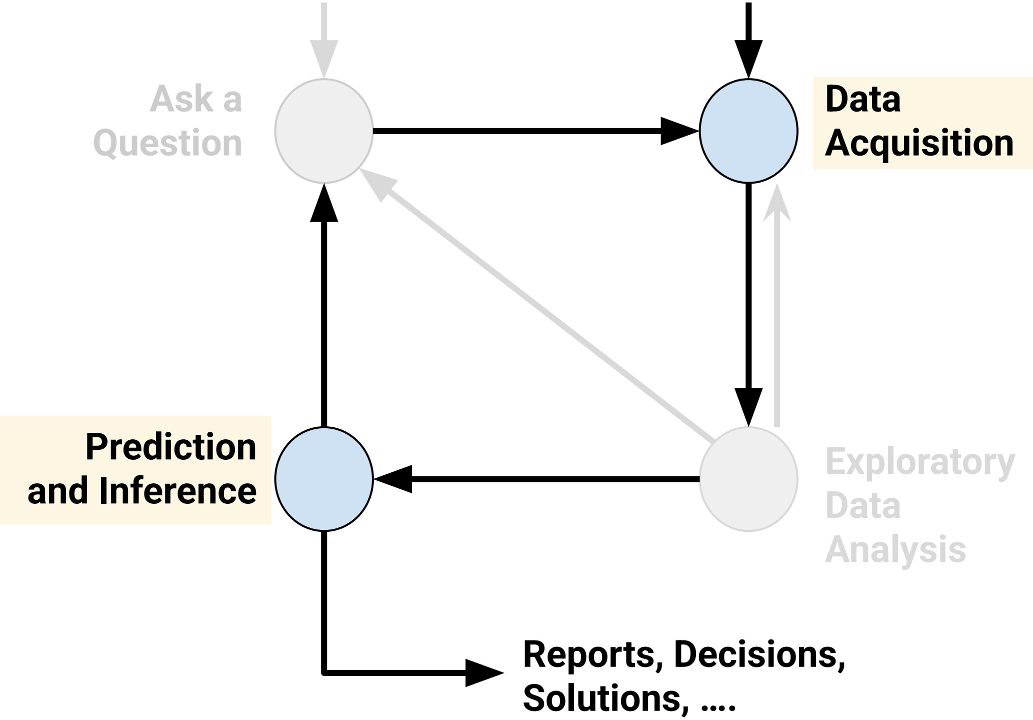
Summary

53372784

Understanding the sampling process is what lets us go from **describing the data** to **understanding the world**.

Without knowing / assuming something about how the data were collected:

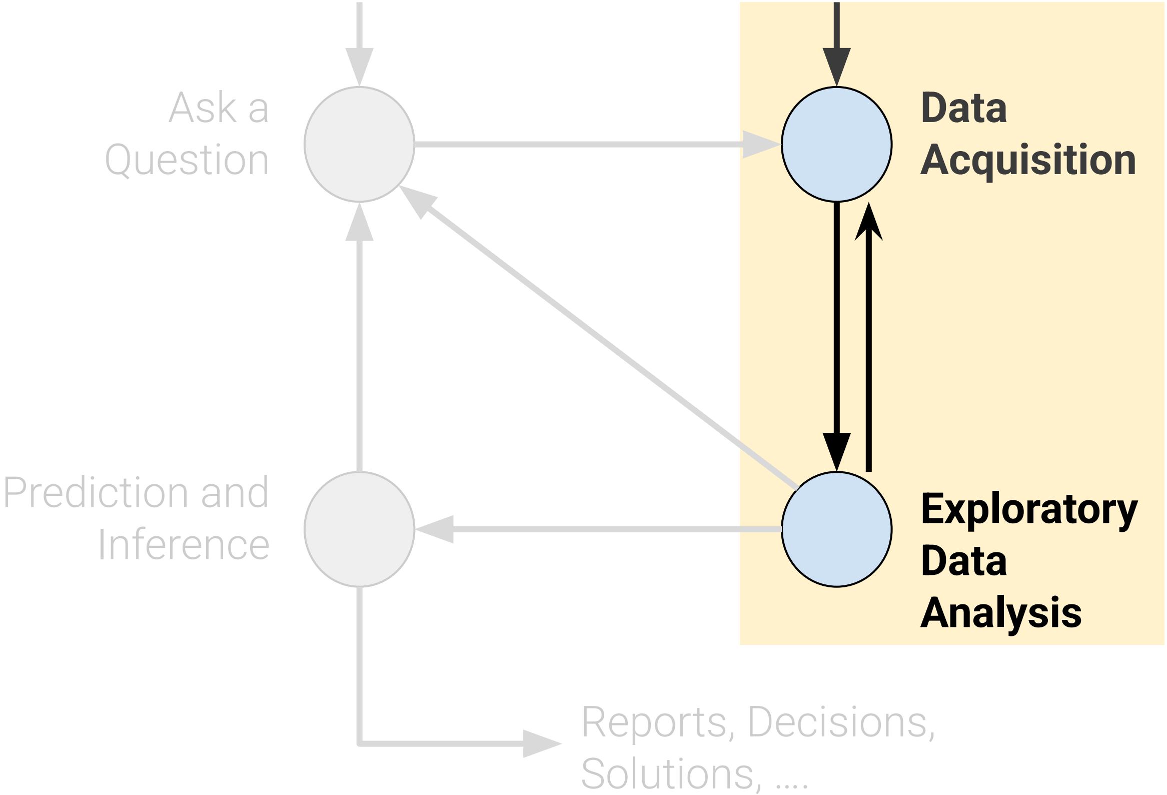
- There is no connection between the **sample** and the **population.**
- The **dataset** doesn't tell us about the **world behind the data**.



Credit: Learning Data Science

Data Transformation

Big Picture



2482612

How do you do EDA/data wrangling?

- Examine the **dataset**
  - What is the date, size, organization, and structure of the data?
- Examine each **variable/attribute/feature** individually
- Examine **pairs of related attributes**
  - Stratify earlier analysis: break down grades by major …

Key Data Properties to Consider in EDA

- **Structure** -- the "shape" of a data file
- **Scope** -- how (in)complete is the data
- **Temporality** -- how is the data situated in time
- **Faithfulness** -- how well does the data capture "reality"

This week: Identifying problems along these 4 properties.

Key Data Properties to Consider: Structure

- Are the data in a standard format or encoding? Is the data organized as records or something else?
    - Tabular data: CSV, TSV, Excel
    - “Nested” data: JSON or XML
- Does the data reference other data?
    - Can the data be joined? Do we need to?
- What are the **variables** in each record?
    - What is the type of the data?

Key Data Properties to Consider: Structure

- What is the **observational unit** being measured? Does each row in the data form a single **observation**?
  - Do we need to simplify the structure? (e.g., select, filter)
  - Do we need to adjust the granularity of the data? (e.g., group_by and summarize)
  - Does the dataset have mixed granularity? Are there records at different levels of detail within the same data file?
  - Does the data need to be reshaped?

3185346

Structure: multiple data files

Sometimes your data comes in multiple files:

- Often data will reference other pieces of data.
- Alternatively, you will collect multiple pieces of related data.

Solution: **join** the data on **keys**.

Note: There are many kinds of join. We won't cover them in this course.

businesses.csv

business_id	name	address
19	NRGIZE...	1200 VAN..
24	OMNI S.F...	500 Califor...
31	NORMAN'S...	2801 Leave..

violations.csv

business_id	date	description
19	20171211	Inadequate food...
19	20171211	Unclean or degrade...
24	20171101	Improper food stor...

inspections.csv

business_id	date	score	type
19	20160513	94	routine
19	20171211	94	routine
24	20171101	98	routine

Primary Key

Primary key: the column or set of columns in a table that *uniquely* determine the values in the remaining columns

- A primary key column is unique, but could be composed of more than one column.
- Examples: SSN, Product ID, Cane ID, …

Primary Key

businesses.csv

business_id	name	address
19	NRGIZE…	1200 VAN..
24	OMNI S.F…	500 Califor…
31	NORMAN'S…	2801 Leave..

No Primary Key!

violations.csv

business_id	date	description
19	20171211	Inadequate food…
19	20171211	Unclean or degrade…
24	20171101	Improper food stor…

Primary Key

inspections.csv

business_id	date	score	type
19	20160513	94	routine
19	20171211	94	routine
24	20171101	98	routine

8065274

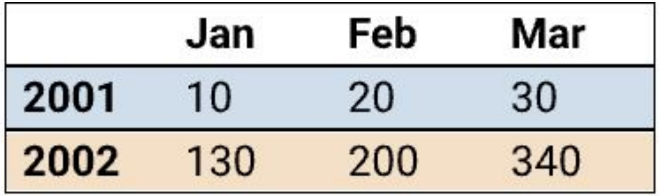
More Structure Problems: Reshaping Data

Observational unit: Number of seats available in a particular month.

Both tables contain the same data, but presented in a different way.

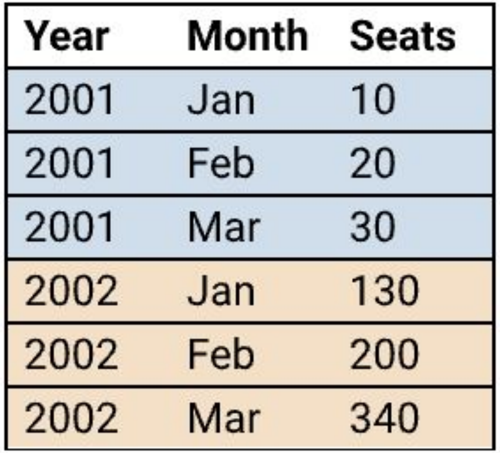
Question: In which format can you identify the observational unit from a single row?

Wide Format



	Jan	Feb	Mar
2001	10	20	30
2002	130	200	340

Long Format

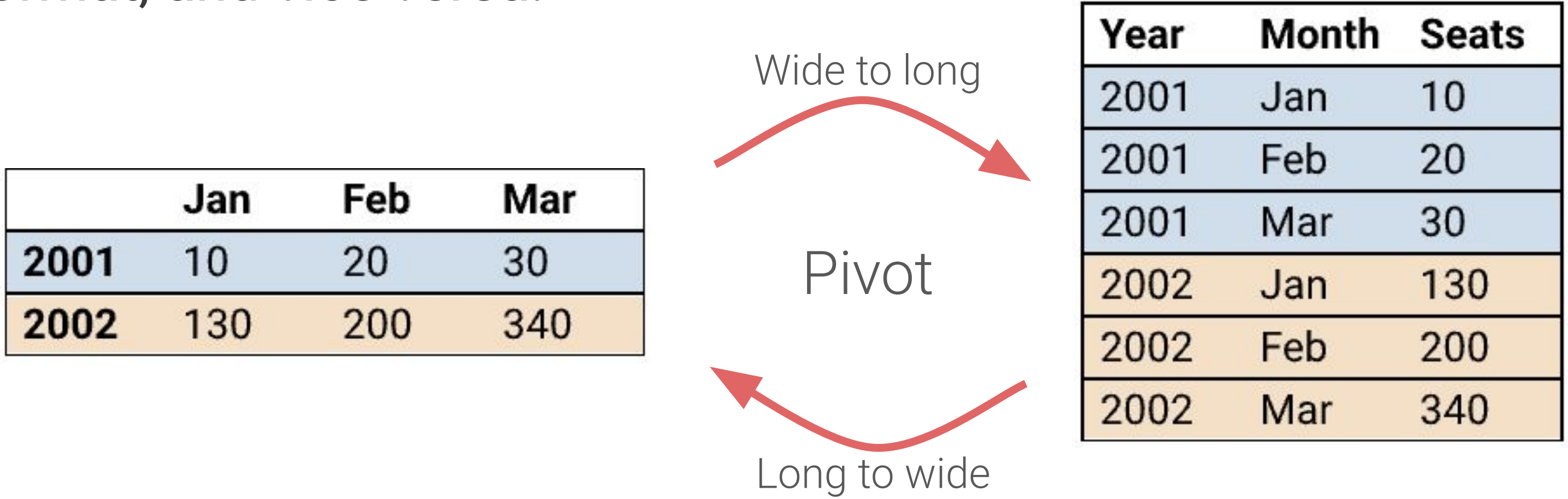


Year	Month	Seats
2001	Jan	10
2001	Feb	20
2001	Mar	30
2002	Jan	130
2002	Feb	200
2002	Mar	340

8065274

Structure: Reshaping Data

Pivot transforms allow us to reshape data from wide to long format, and vice versa.



	Jan	Feb	Mar
2001	10	20	30
2002	130	200	340

Wide to long

Pivot

Long to wide

Year	Month	Seats
2001	Jan	10
2001	Feb	20
2001	Mar	30
2002	Jan	130
2002	Feb	200
2002	Mar	340

FYI: When each observational unit forms a row (among a few other rules), the data is said to be tidy data.

Key Data Properties to Consider: Scope

Will my data be enough to answer my question?

- **Example:** I am interested in studying crime in Florida but I only have Miami crime data.
- **Solution:** collect more data/change research question

Is my data too expansive?

- **Example:** I am interested in student grades for CSC198 but have student grades for all UM Computer Science classes.
- **Solution:** Filter the data (implications on sample?)

“Scope” questions are defined by your question/problem and inform if you need better-scoped data.

Key Data Properties to Consider: Temporality

- Does my data cover the right time frame?
- What is the meaning of the time and date fields? A few options:
  - When the “event” happened?
  - When the data was collected or was entered into the system?
  - Date the data was copied into a database?
- Time depends on **where**! (**time zones** & daylight savings)
  - Regions have different date representations: what does **07/08/09** mean? July 8 2009? August 7 2009…?
- Are there strange values that indicate “missingness”?
  - e.g., January 1st 1970, January 1st 1900…?

Temporality: Unix / POSIX Time

Time measured in seconds since **January 1 1970 UTC**

- Minus leap seconds …

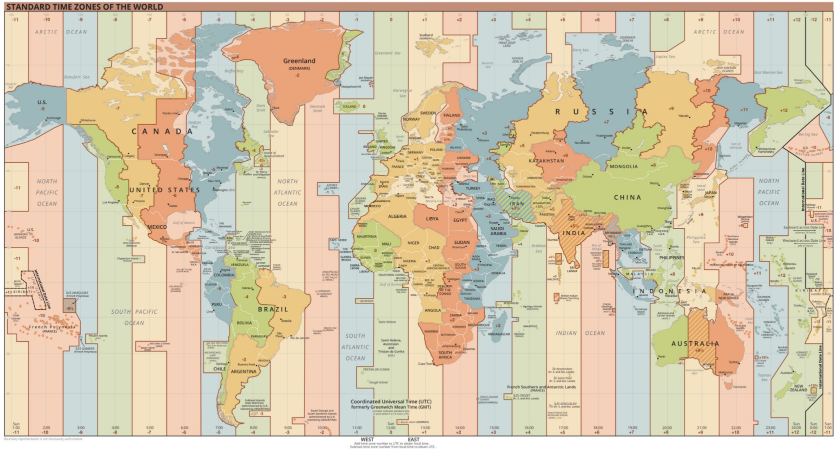
Feb 5, 2025 11:15am EST
`1738772100`

UTC is Coordinated Universal Time

- International time standard
- Measured at 0 degrees latitude
- No daylight savings

Time Zones:

- Miami (UTC-5) without daylight savings



https://en.wikipedia.org/wiki/Coordinated_Universal_Time

8065274

Faithfulness: Do I trust this data?

Does my data contain **unrealistic or “incorrect” values**?

- Dates in the future for events in the past
- Locations that don’t exist
- Negative counts
- Misspellings of names
- Large outliers

Does my data violate **obvious dependencies**?

- E.g., values in two columns “age” and “birthday” don’t match

Faithfulness: Do I trust this data?

Was the data **entered by hand**?

- Spelling errors, fields shifted …
- Did the form require all fields or provide default values?

Are there obvious signs of **data falsification**?

- Repeated names, fake looking email addresses, repeated use of uncommon names or fields.

Signs that your data may not be faithful

Truncated data

Early Microsoft Excel limits: 65536 Rows, 255 Columns

Spelling Errors

Apply corrections or delete records not in a dictionary

Time Zone Inconsistencies

Convert to a common timezone (e.g., UTC)

Duplicated Records or Fields

Identify and eliminate (use primary key).

Units not specified or consistent

Infer units, check values are in reasonable ranges for data

Be aware of consequences in analysis when using data with inconsistencies.

Signs of Missing Values

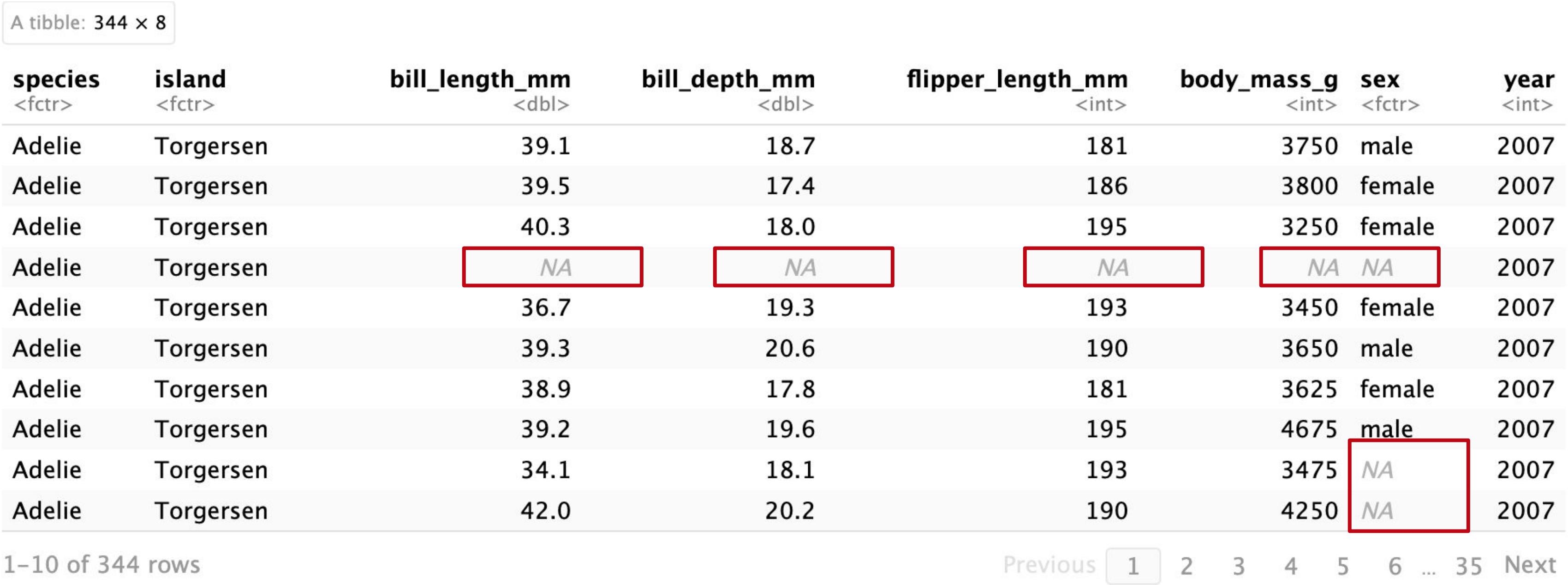
Examples

```
"  “             1970, 1900
0, -1            NaN
999, 12345       Null
```

NaN: “Not a Number”

8065274

Faithfulness: Missing Values



A tibble: 344 × 8

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
<fctr>	<fctr>	<dbl>	<dbl>	<int>	<int>	<fctr>	<int>
Adelie	Torgersen	39.1	18.7	181	3750	male	2007
Adelie	Torgersen	39.5	17.4	186	3800	female	2007
Adelie	Torgersen	40.3	18.0	195	3250	female	2007
Adelie	Torgersen	NA	NA	NA	NA	NA	2007
Adelie	Torgersen	36.7	19.3	193	3450	female	2007
Adelie	Torgersen	39.3	20.6	190	3650	male	2007
Adelie	Torgersen	38.9	17.8	181	3625	female	2007
Adelie	Torgersen	39.2	19.6	195	4675	male	2007
Adelie	Torgersen	34.1	18.1	193	3475	NA	2007
Adelie	Torgersen	42.0	20.2	190	4250	NA	2007

1–10 of 344 rows

Problem: “Holes” in the penguins data!

Missing Values: Popular Solutions

A. **Delete records** with missing values
   - Probably most common; called a **complete cases analysis**

B. **Do nothing**; keep as missing

C. **Imputation/Interpolation:** Inferring missing values somehow
   - **Average Imputation:** replace with an average value
   - **Hot deck imputation:** replace with a random value
   - **Arbitrary value imputation:** replace with some arbitrary value

Caution: approaches in (A) and (C) can induce biases; missing records might be related to something of interest.