Data

Dr. Ab Mosca (they/them)

Slides based off slides courtesy of OpenIntro and John McGreedy of Johns Hopkins University

Reminder

- hw-01 is released today and due next week
- Check the course website for instructions, submit on PLATO

Plan for Today

- Data types
- Data tables
- Data ethics

Data Types

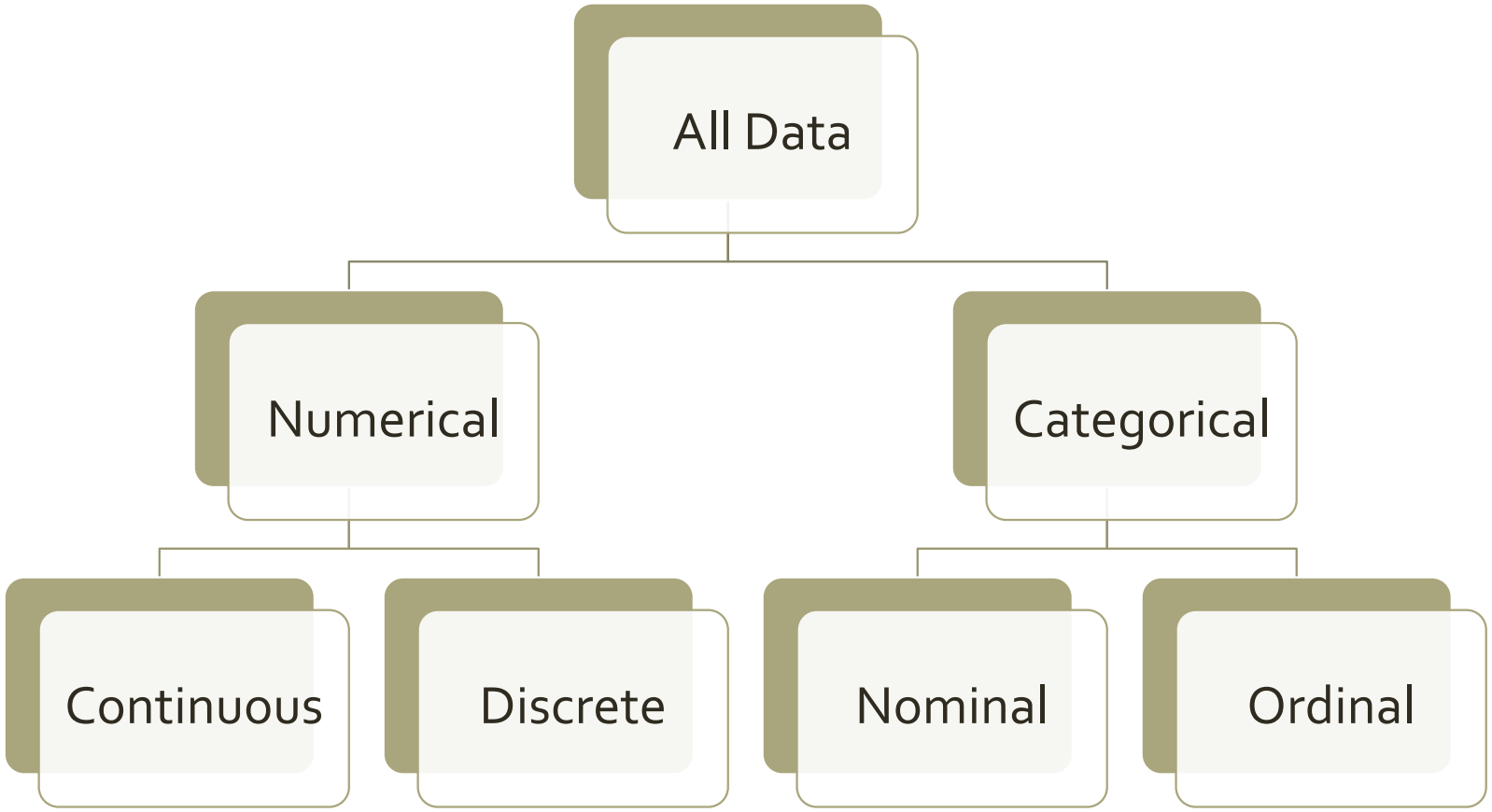
Brainstorm

- What information could we collect about folks in this class?

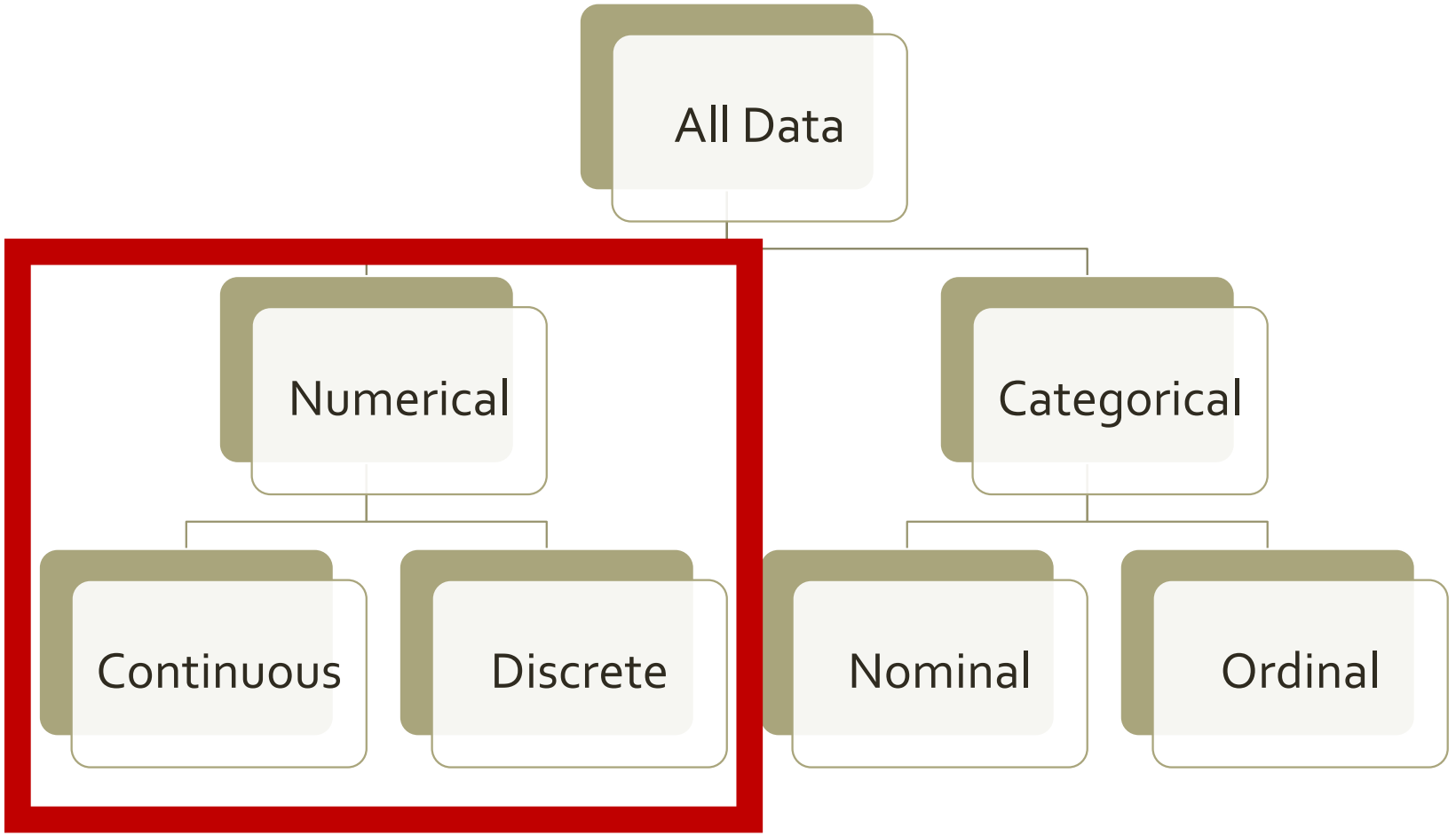
Brainstorm

- What information could we collect about folks in this class?
- Would all of the information we get be formatted similarly or differently?

Categorizing Data



Categorizing Data



Numerical - Continuous

- Continuous data *(incremental measurements)*
  - Blood pressure, mmHg
  - Weight, lbs (kgs, oz, etc.)
  - Height, ft (cm, in, etc.)
  - Age, years (months)
  - Income level, dollars/year (Euro by year, etc.)
- A defining characteristic of continuous data is that a one-unit change in the value means the same thing across the entire range of data values

Numerical - Continuous

- Continuous data *(incremental measurements)*
  - Blood pressure, mmHg
  - Weight, lbs (kgs, oz, etc.)
  - Height, ft (cm, in, etc.)
  - Age, years (months)
  - Income level, dollars/year (Euro by year, etc.)
- A defining characteristic of continuous data is that a one-unit change in the value means the same thing across the entire range of data values

What other examples can you think of?
Brainstorm with whoever is near you

Numerical – Discrete (Binary)

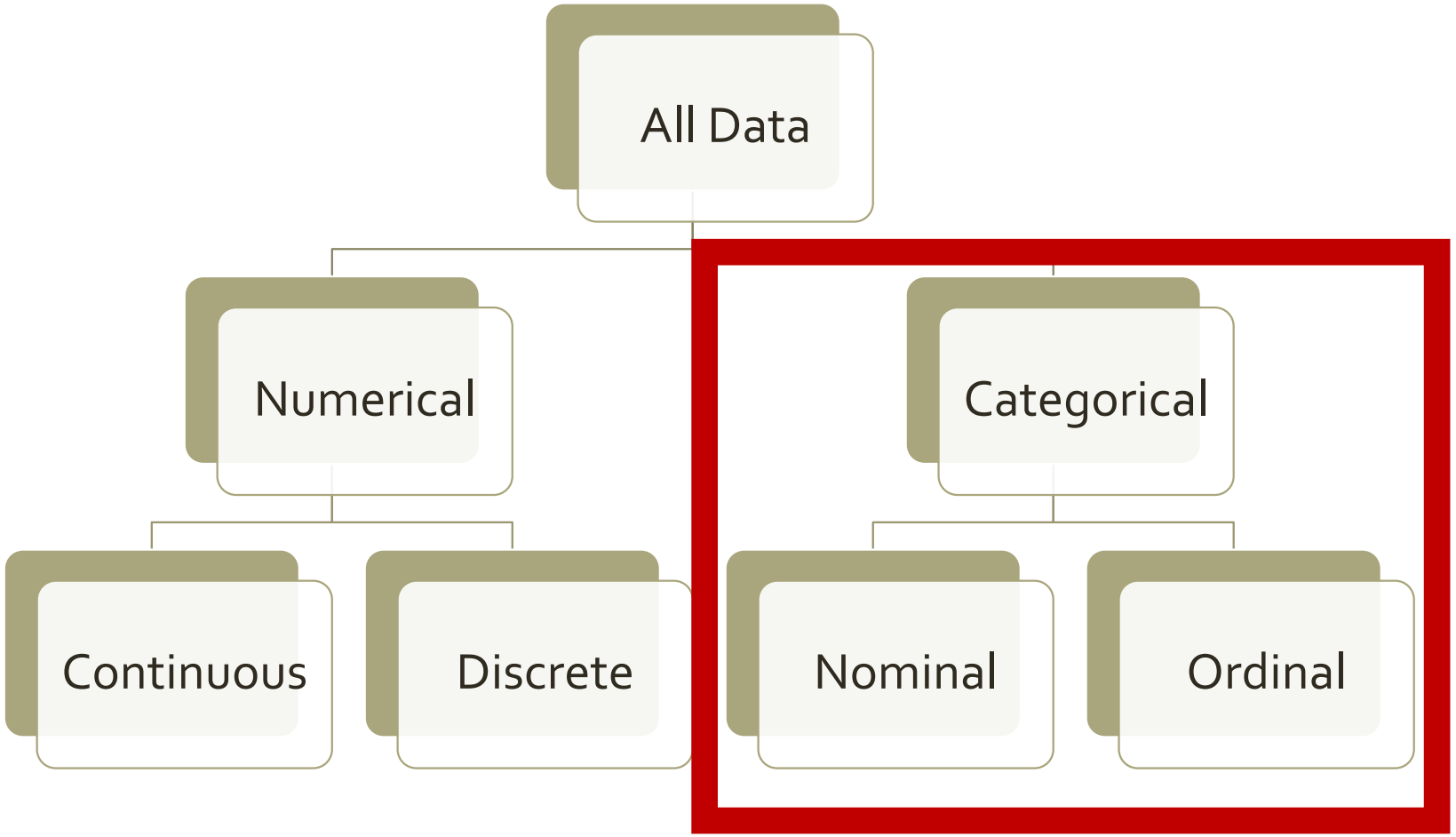
- Binary (dichotomous) data: takes on only two values, “yes” or “no”
- Binary (dichotomous) data (“yes/no” data)
  - Polio: Yes/No
  - Remission: Yes/No
  - Quit smoking: Yes/No
  - Etc.

Numerical – Discrete (Binary)

- Binary (dichotomous) data: takes on only two values, "yes" or "no"
- Binary (dichotomous) data ("yes/no" data)
  - Polio: Yes/No
  - Remission: Yes/No
  - Quit smoking: Yes/No
  - Etc.

What other examples can you think of?
Brainstorm with whoever is near you

Categorizing Data



Categorical - Nominal

- Categorical data: an extension of binary data to include more than 2 possible values
- Nominal categorical data: no inherent order to categories
  - Gender Identity
  - Race/ethnicity
  - Country of birth
  - Religious affiliation

Categorical - Nominal

- Categorical data: an extension of binary data to include more than 2 possible values
- Nominal categorical data: no inherent order to categories
  - Gender Identity
  - Race/ethnicity
  - Country of birth
  - Religious affiliation

What other examples can you think of?
Brainstorm with whoever is near you

Categorical - Ordinal

- Categorical data: an extension of binary data to include more than 2 possible values
- Ordinal categorical data: order to categories
  - Income level categorized into four categories, least to greatest
  - Degree of agreement, five categories from strongly disagree to strongly agree

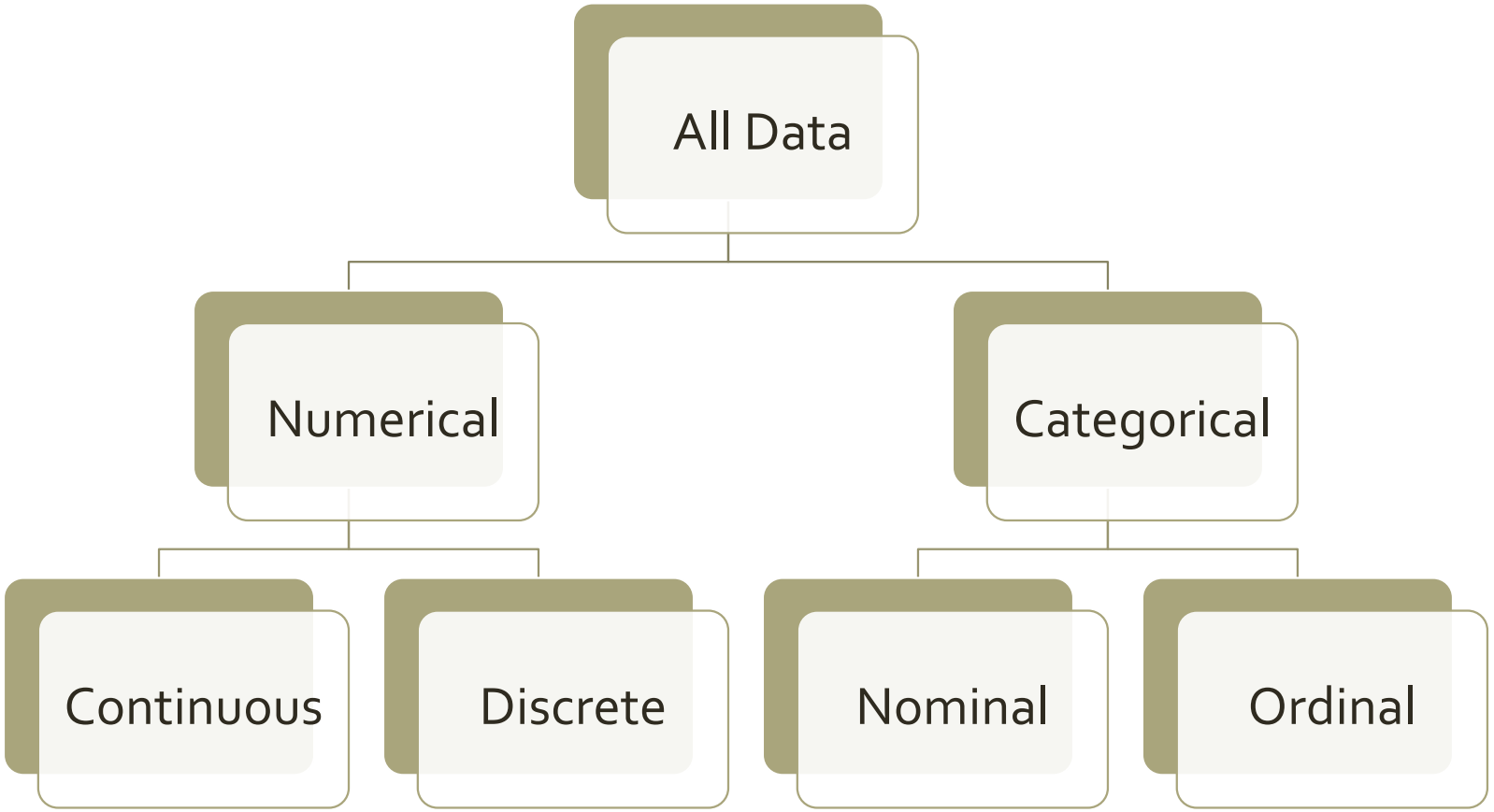
Categorical - Ordinal

- Categorical data: an extension of binary data to include more than 2 possible values
- Ordinal categorical data: order to categories
  - Income level categorized into four categories, least to greatest
  - Degree of agreement, five categories from strongly disagree to strongly agree

What other examples can you think of?
Brainstorm with whoever is near you

Practice

- Work with whoever is near you to categorize the data we decided to collect about the class into data types



Data Tables

Practice

Classroom survey

A survey was conducted on students in an introductory statistics course. Below are a few of the questions on the survey, and the corresponding variables the data from the responses were stored in:

- **gender**: What is your gender?
- **intro_extra**: Are you an introvert or an extrovert?
- **sleep**: How many hours do you sleep at night, on average?
- **bedtime**: What time do you usually go to bed?
- **countries**: How many countries have you visited?
- **dread**: On a scale of 1-5, how much do you dread being here?

Practice

Classroom survey

- **gender**: What is your gender?
- **intro_extra**: Are you an introvert or an extrovert?
- **sleep**: How many hours do you sleep at night, on average?
- **bedtime**: What time do you usually go to bed?
- **countries**: How many countries have you visited?
- **dread**: On a scale of 1-5, how much do you dread being here?

What is the data type for each of these?

Practice

Classroom survey

- **gender**: What is your gender?
  - categorical – nominal
- **intro_extra**: Are you an introvert or an extrovert?
  - numerical – discrete, OR categorical – nominal
- **sleep**: How many hours do you sleep at night, on average?
  - numerical – continuous
- **bedtime**: What time do you usually go to bed?
  - numerical – continuous
- **countries**: How many countries have you visited?
  - numerical – continuous
- **dread**: On a scale of 1-5, how much do you dread being here?
  - categorical – ordinal

Data Table

Classroom survey

Stu.	gender	intro_extra	···	dread
1	male	extravert	···	3
2	female	extravert	···	2
3	female	introvert	···	4
4	female	extravert	···	2
⋮	⋮	⋮	⋮	⋮
86	male	extravert	···	3

- Often, collected data is stored in one or more tables

Data Table

Classroom survey

variable ↓

Stu.	gender	intro_extra	···	dread
1	male	extravert	···	3
2	female	extravert	···	2
3	female	introvert	···	4
4	female	extravert	···	2
⋮	⋮	⋮	⋮	⋮
86	male	extravert	···	3

- Often, collected data is stored in one or more tables
- One column in a table represents one datum collected, and is called a **variable** (or attribute)

Data Table

Classroom survey

variable ↓

Stu.	gender	intro_extra	···	dread
1	male	extravert	···	3
2	female	extravert	···	2
3	female	introvert	···	4
4	female	extravert	···	2
⋮	⋮	⋮	⋮	⋮
86	male	extravert	···	3

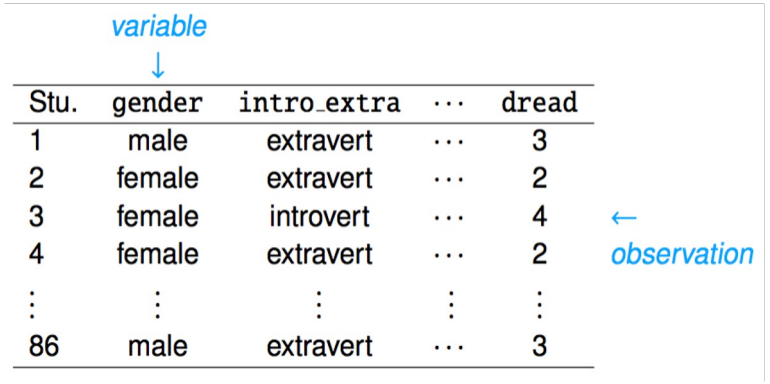
← *observation*

- Often, collected data is stored in one or more tables
- One column in a table represents one datum collected, and is called a **variable** (or attribute)
- One row in a table contains values for each datum collected about one item, and is called an **observation**

Practice

What are the variables in this data set?

What does one observation represent?



MA DCR Precipitation Database

STATION	YEAR	JAN	CITY	Region Composite Station	Region	Basin Name
GRA220	1908	3.05	Westfield	Yes	Connecticut River	WESTFIELD
GRA220	1909	4.06	Westfield	Yes	Connecticut River	WESTFIELD
GRA220	1910	6.4	Westfield	Yes	Connecticut River	WESTFIELD
GRA220	1911	2.06	Westfield	Yes	Connecticut River	WESTFIELD

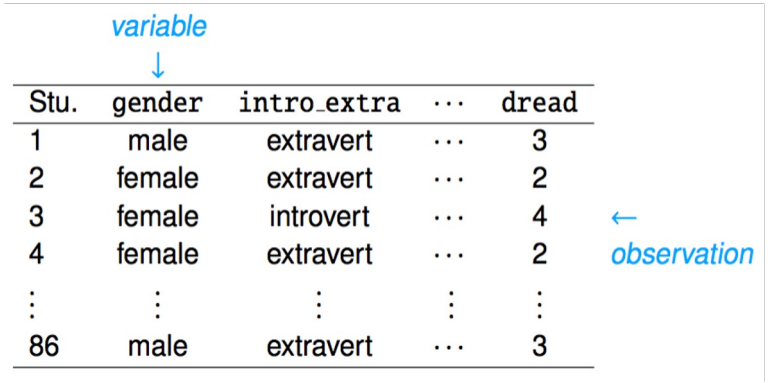
.
.
.

https://www.mass.gov/info-details/precipitation-data

Practice

What are the variables in this data set?

What does one observation represent?



MA DCR Precipitation Database

Variables →

STATION	YEAR	JAN	CITY	Region Composite Station	Region	Basin Name
GRA220	1908	3.05	Westfield	Yes	Connecticut River	WESTFIELD
GRA220	1909	4.06	Westfield	Yes	Connecticut River	WESTFIELD
GRA220	1910	6.4	Westfield	Yes	Connecticut River	WESTFIELD
GRA220	1911	2.06	Westfield	Yes	Connecticut River	WESTFIELD

row == one instance of rainfall collection

.

.

https://www.mass.gov/info-details/precipitation-data

Practice

What are the variables in this data set?

What does one observation represent?

variable ↓

Stu.	gender	intro_extra	⋯	dread
1	male	extravert	⋯	3
2	female	extravert	⋯	2
3	female	introvert	⋯	4
4	female	extravert	⋯	2
⋮	⋮	⋮	⋮	⋮
86	male	extravert	⋯	3

← *observation*

MA Undergraduate Enrollment

Year	Segment	Unduplicated Headcount
2020	State Universities	36053
2020	University of Massachusetts	56857
2020	Community Colleges	67685

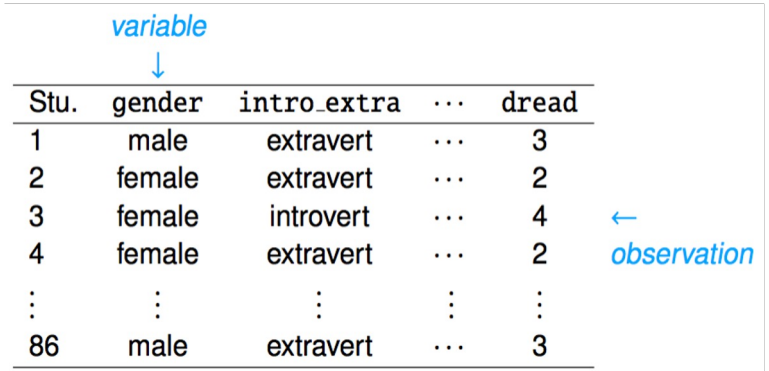
.
.
.

https://www.mass.edu/datacenter/2020enrollmentestimates.asp

Practice

What are the variables in this data set?

What does one observation represent?



MA Undergraduate Enrollment

Variables →

Year	Segment	Unduplicated Headcount
2020	State Universities	36053
2020	University of Massachusetts	56857
2020	Community Colleges	67685

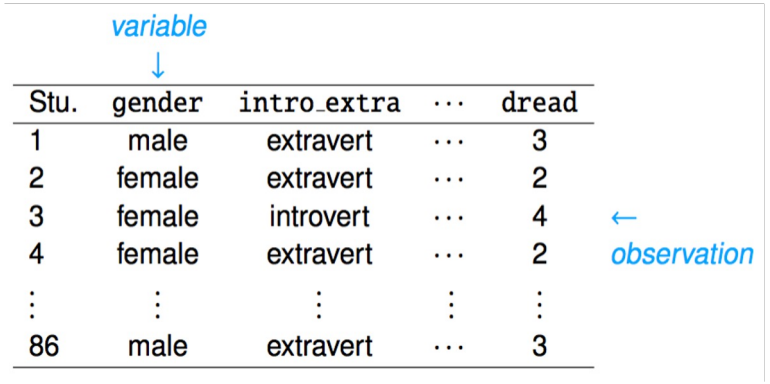
row == one segment in one year

https://www.mass.edu/datacenter/2020enrollmentestimates.asp

Practice

What are the variables in this data set?

What does one observation represent?



MA Outdoor Advertisement Signs

PermitHoldersName	SignCity	RoadIntendedToFace	Active	SignType
Clear Channel Outdoor	Worcester	Rt 122	1	Traditional
Clear Channel Outdoor	Worcester	Rt. 290	1	Digital
Murray Marketing, Inc	Worcester	I-290	1	Digital
Lamar Central Outdoor, LLC	Westfield	Southampton Rd	1	Traditional

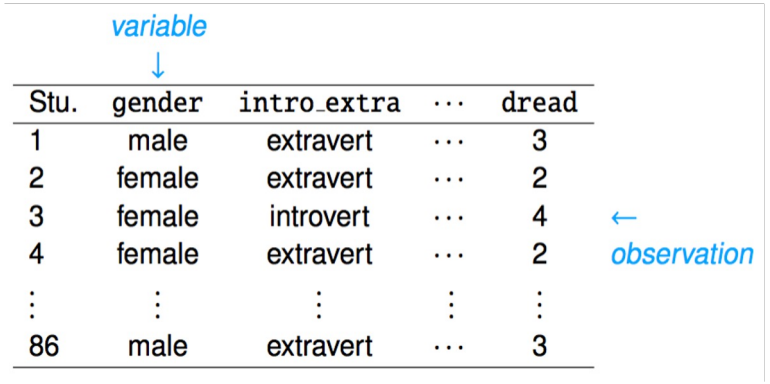
.
.
.

https://geo-massdot.opendata.arcgis.com/datasets/MassDOT::outdoor-advertising-outdoor-advertisement-signs/explore?location=42.046332%2C-71.718750%2C8.93

Practice

What are the variables in this data set?

What does one observation represent?



MA Outdoor Advertisement Signs

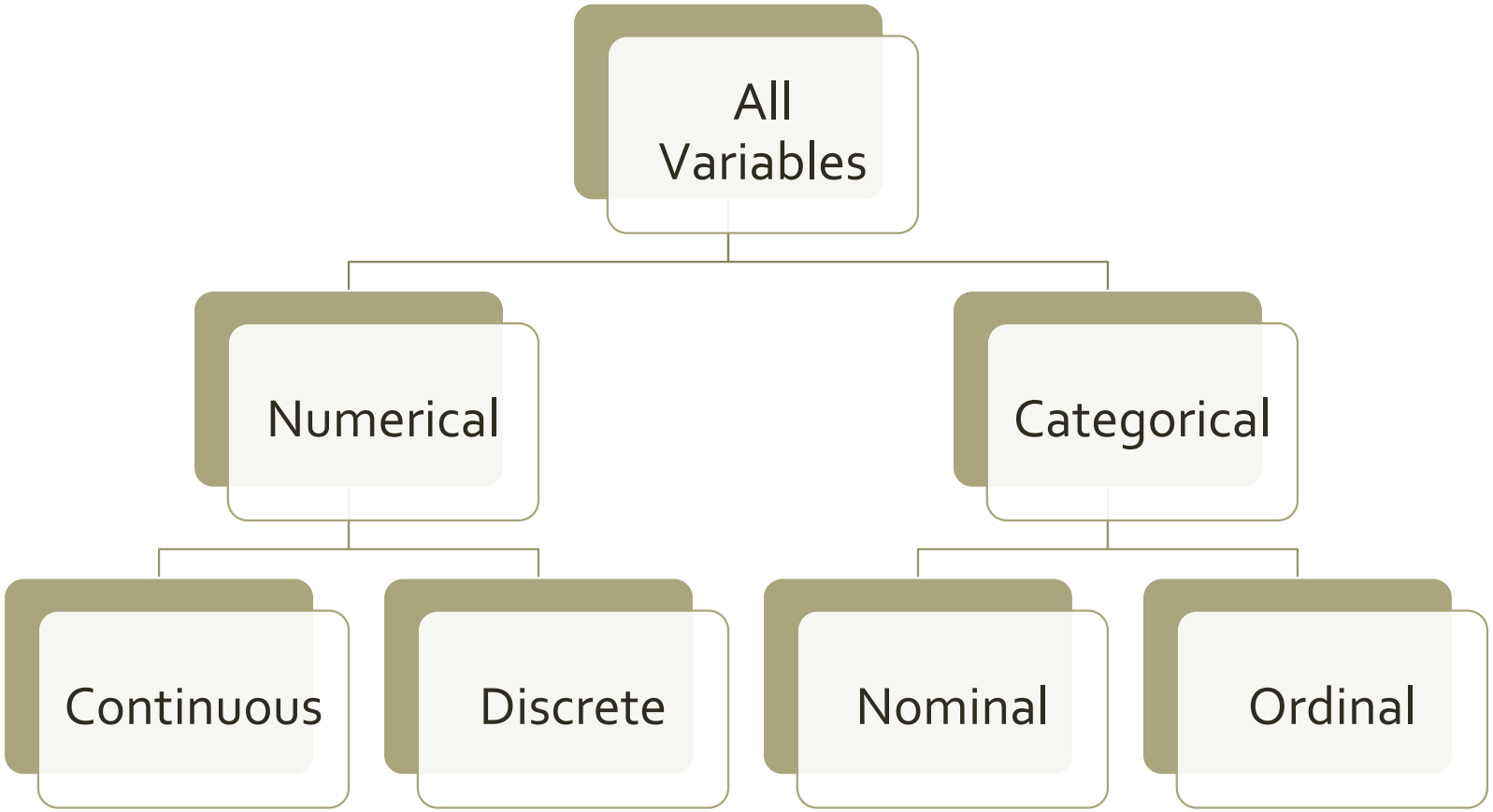
Variables →

PermitHoldersName	SignCity	RoadIntendedToFace	Active	SignType
Clear Channel Outdoor	Worcester	Rt 122	1	Traditional
Clear Channel Outdoor	Worcester	Rt. 290	1	Digital
Murray Marketing, Inc	Worcester	I-290	1	Digital
Lamar Central Outdoor, LLC	Westfield	Southampton Rd	1	Traditional

row == one outdoor advertisement sign

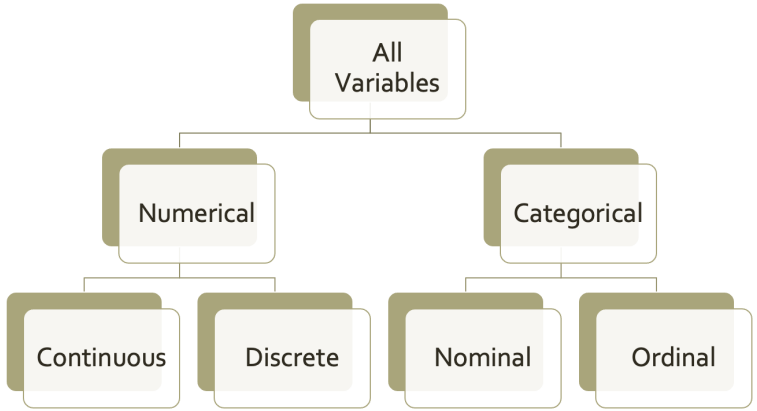
https://geo-massdot.opendata.arcgis.com/datasets/MassDOT::outdoor-advertising-outdoor-advertisement-signs/explore?location=42.046332%2C-71.718750%2C8.93

Types of Variables



Types of Variables

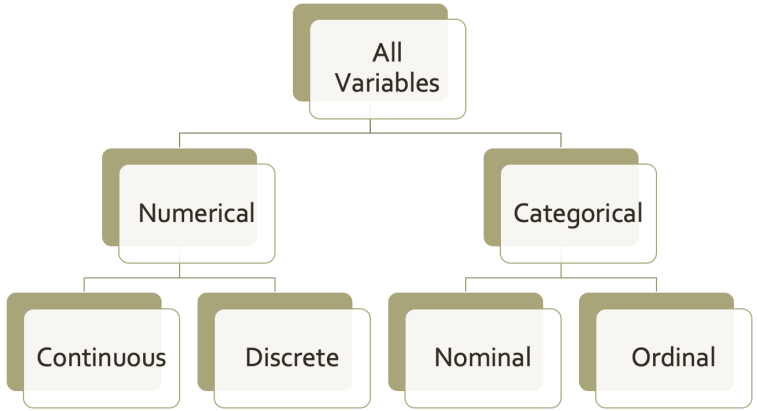
What are the types of these variables?



	gender	sleep	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

Types of Variables

What are the types of these variables?

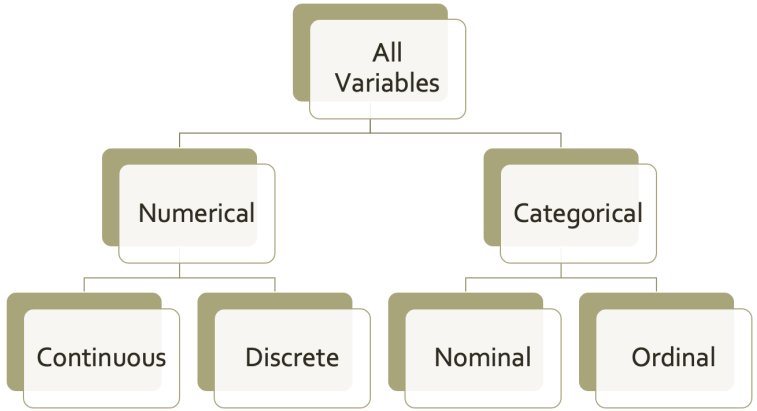


	gender	sleep	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

nominal
continuous
continuous
continuous
ordinal

Types of Variables

What are the types of these variables?



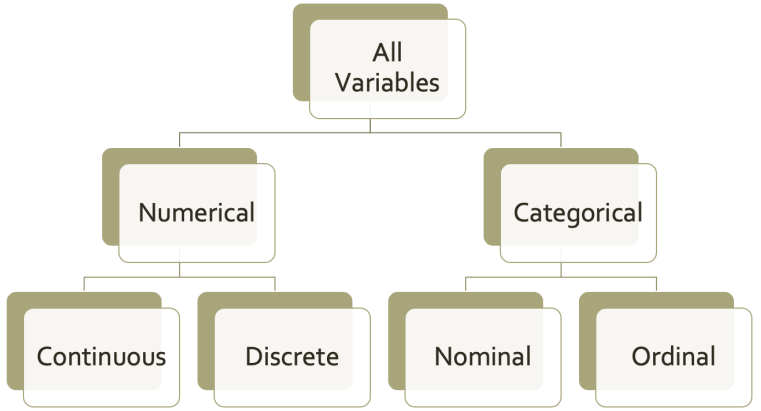
MA DCR Precipitation Database

STATION	YEAR	JAN	CITY	Region Composite Station	Region	Basin Name
GRA220	1908	3.05	Westfield	Yes	Connecticut River	WESTFIELD
GRA220	1909	4.06	Westfield	Yes	Connecticut River	WESTFIELD
GRA220	1910	6.4	Westfield	Yes	Connecticut River	WESTFIELD
GRA220	1911	2.06	Westfield	Yes	Connecticut River	WESTFIELD

.
.
.

Types of Variables

What are the types of these variables?



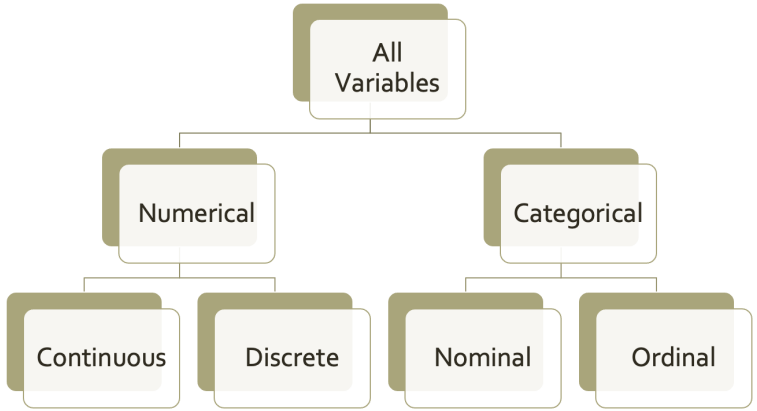
MA DCR Precipitation Database

STATION	YEAR	JAN	CITY	Region Composite Station	Region	Basin Name
GRA220	1908	3.05	Westfield	Yes	Connecticut River	WESTFIELD
GRA220	1909	4.06	Westfield	Yes	Connecticut River	WESTFIELD
GRA220	1910	6.4	Westfield	Yes	Connecticut River	WESTFIELD
GRA220	1911	2.06	Westfield	Yes	Connecticut River	WESTFIELD
.	.	.	.	.	.	.

nominal | continuous | continuous | nominal | discrete | nominal | nominal

Types of Variables

What are the types of these variables?



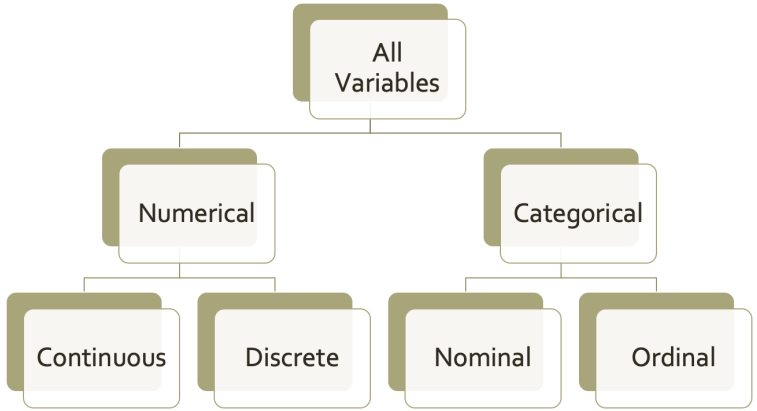
MA Undergraduate Enrollment

Year	Segment	Unduplicated Headcount
2020	State Universities	36053
2020	University of Massachusetts	56857
2020	Community Colleges	67685

.
.
.

Types of Variables

What are the types of these variables?



MA Undergraduate Enrollment

Year	Segment	Unduplicated Headcount
2020	State Universities	36053
2020	University of Massachusetts	56857
2020	Community Colleges	67685

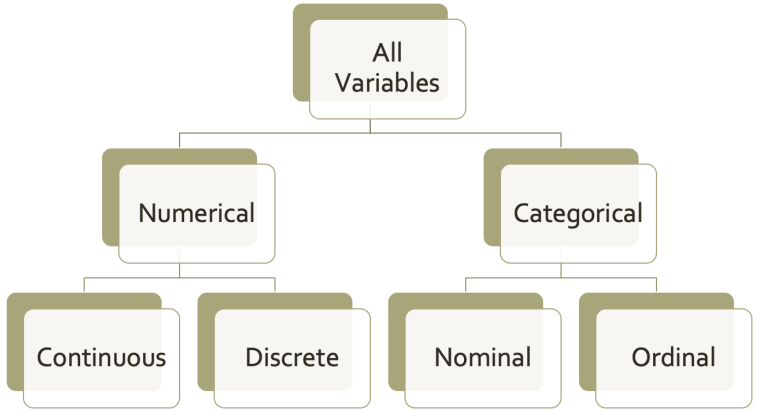
continuous

nominal

continuous

Types of Variables

What are the types of these variables?



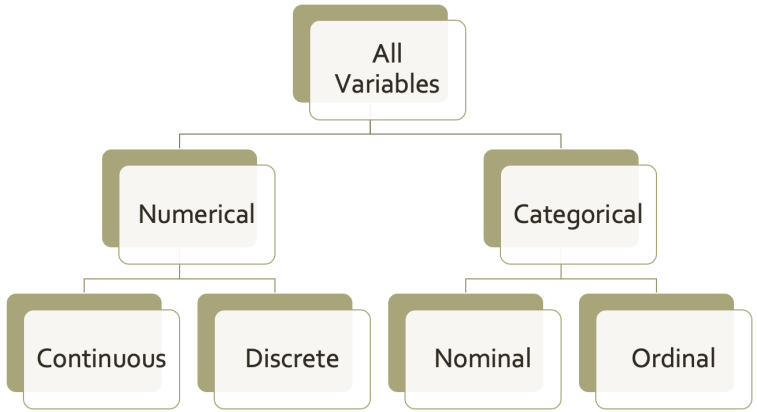
MA Outdoor Advertisement Signs

PermitHoldersName	SignCity	RoadIntendedToFace	Active	SignType
Clear Channel Outdoor	Worcester	Rt 122	1	Traditional
Clear Channel Outdoor	Worcester	Rt. 290	1	Digital
Murray Marketing, Inc	Worcester	I-290	1	Digital
Lamar Central Outdoor, LLC	Westfield	Southampton Rd	1	Traditional

.
.
.

Types of Variables

What are the types of these variables?



MA Outdoor Advertisement Signs

PermitHoldersName	SignCity	RoadIntendedToFace	Active	SignType
Clear Channel Outdoor	Worcester	Rt 122	1	Traditional
Clear Channel Outdoor	Worcester	Rt. 290	1	Digital
Murray Marketing, Inc	Worcester	I-290	1	Digital
Lamar Central Outdoor, LLC	Westfield	Southampton Rd	1	Traditional

nominal nominal nominal discrete nominal

Data Ethics

Responsible Consumption

- What details should we look for when evaluating a data set?
- Brainstorm with the person next to you and come up with at least 3 things

Responsible Consumption

- What details should we look for when evaluating a data set?
  - Who published the data?
  - Who collected the data?
  - Who funded collection of the data?
  - Who (or what) is included in the dataset?
  - Who (or what) is missing from the dataset?
  - Was the data transparently collected?
  - Was the data legally collected?
  - Are there any privacy issues?

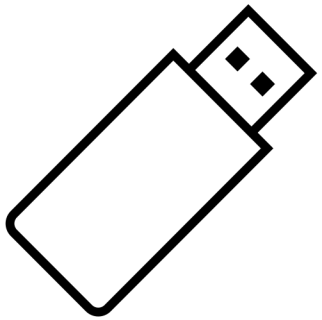
Responsible Collection and Use

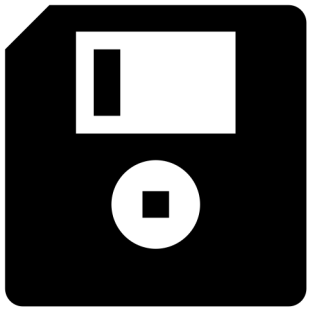
- What can go wrong?
  - Disclosure and Reidentification

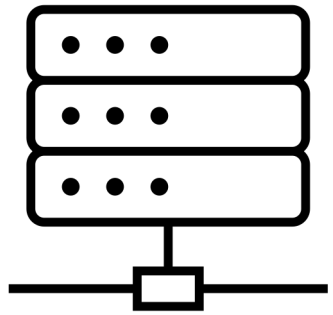
https://mdsr-book.github.io/mdsr2e/ch-ethics.html

Responsible Collection and Use

- What can go wrong?
  - Unsafe storage



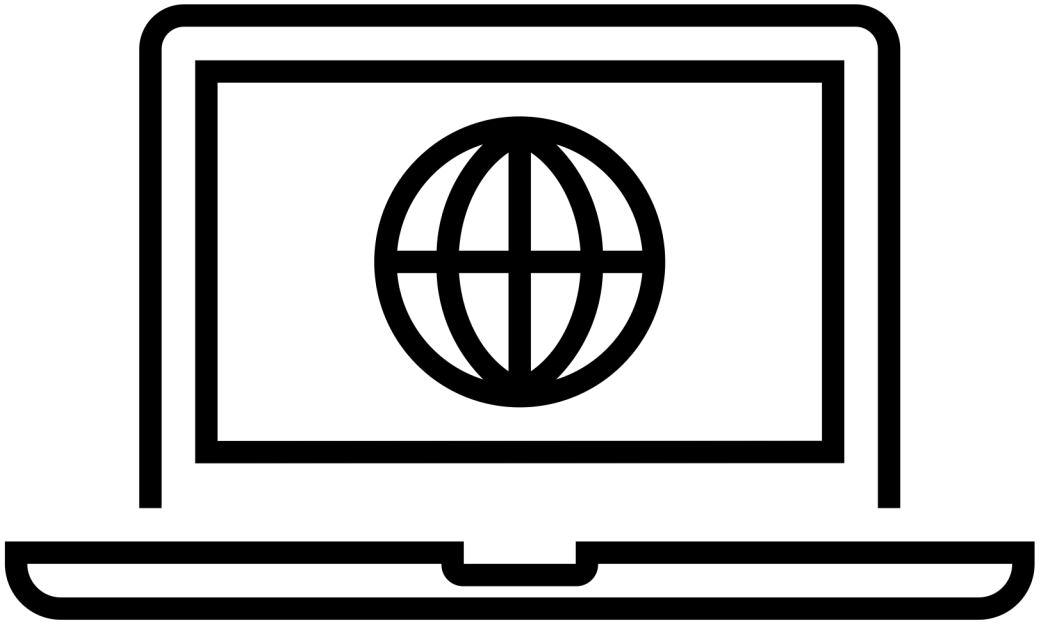




https://mdsr-book.github.io/mdsr2e/ch-ethics.html

Responsible Collection and Use

- What can go wrong?
  - Scraping and Terms of Use



https://mdsr-book.github.io/mdsr2e/ch-ethics.html

Responsible Collection and Use

- Reproducibility
- A reproducible analysis records each and every step, no matter how trivial seeming, in a data analysis. The main elements of a reproducible analysis plan (as described by Project TIER include:
  - **Data**: all original data files in the form in which they originated,
  - **Metadata**: codebooks and other information needed to understand the data,
  - **Commands**: the computer code needed to extract, transform, and load the data—then run analyses, fit models, generate graphical displays, and
  - **Map**: a file that maps between the output and the results in the report.

https://mdsr-book.github.io/mdsr2e/ch-ethics.html

Investigate

- Work with a few people
- Find an article from a news source (ex. New York Times, Washington Post, FiveThirtyEight) that includes data
- See if you can
  - Find who collected the data
  - Find who funded the data collection
  - Find the original data set
- How difficult was it to find these things?
- Be prepared to share