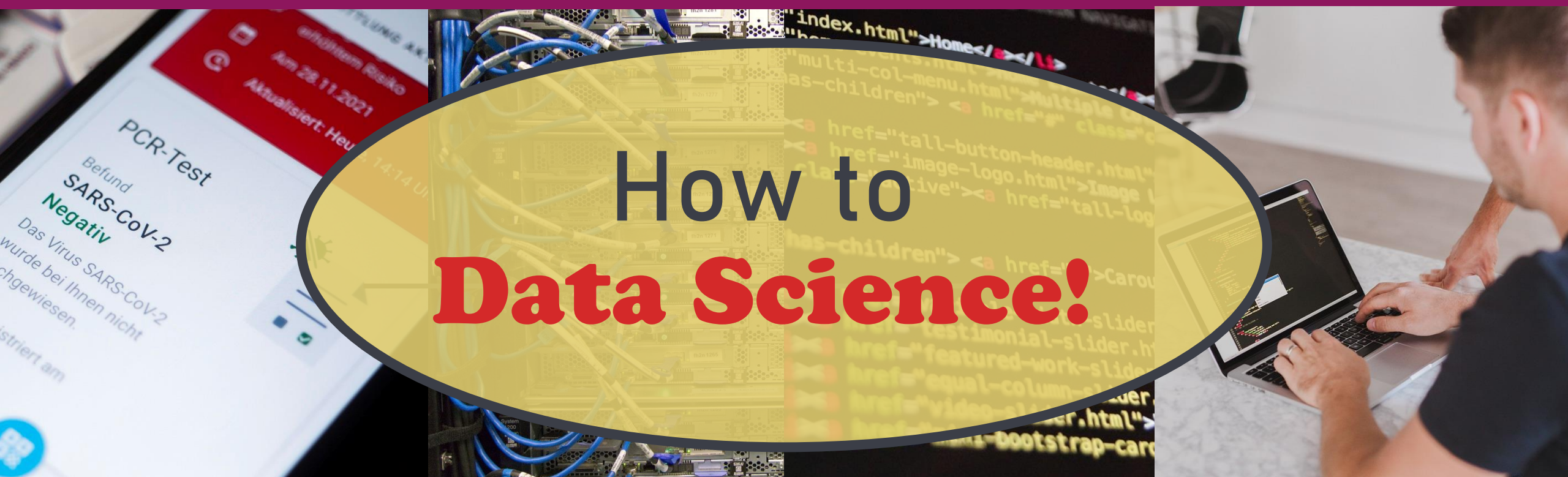


Topic: Root out Selection Bias with a Data Reduction Diagram!

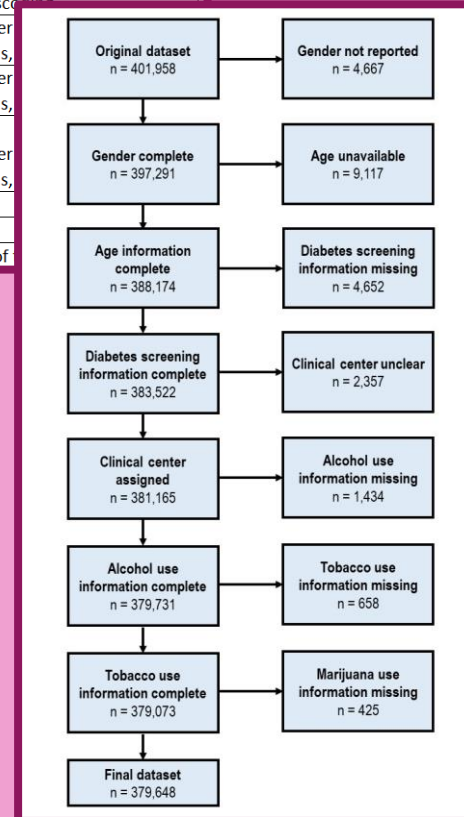


**How to
Data Science!**

Welcome to Today's Lecture

- **Topic:** How to make a data reduction diagram so you can evaluate **selection bias**
- How to assemble the data you need into a spreadsheet
- How to put those data into a diagram
- How to evaluate the results for selection bias

	A	B	C	D	E	F
	Starting Number	Exclusion	Number Removed	Number left	Variable	Comment
1						
2	401,958	Gender not reported	4,667	397,291	PTGENDER	
3	397,291	Age could not be calculated	9,117	388,174	PTDOB	DOB could be missing or miscoded
4	388,174	Diabetes screening variable not populated, or unclear code	4,652	383,522	DBSCREEN	Outcome variable (DV). We will make a binary version for logistic regression.
5	383,522	Single clinical center could not be assigned	2,357	381,165	PTCLINCTR	Exposure variable. Some had multiple clinical centers, and others had miscoded
6	381,165	Remove ALCGRP=9 (unknown) to eliminate small cells	1,434	379,731	ALCGRP	After cells,
7	379,731	Remove TOBGRP = 9 (unknown) to eliminate small cells	658	379,073	TOBGRP	After cells,
8	379,073	Remove MARJGRP = 9 (unknown) to eliminate small cells	425	378,648	MARJGRP	After cells,
9						
10						
11						94% % of



Welcome to Today's Lecture!

- In data science, we do **science**
 - We need to be transparent
 - Our analyses need to be replicable
- Part of this is documenting how you transformed any original dataset you are given into an analytic dataset
 - What rows did you remove and why?
 - How many rows were removed at each transformation step?
- I'll show you how to collect the data in Excel then make a diagram in PowerPoint that you can use to evaluate **selection bias**

Resources

Blog post with examples:

<https://dethwench.com/data-reduction-diagram-exposes-selection-bias-in-data-science/>

“How to Make a Data Reduction Diagram in PowerPoint” online course:

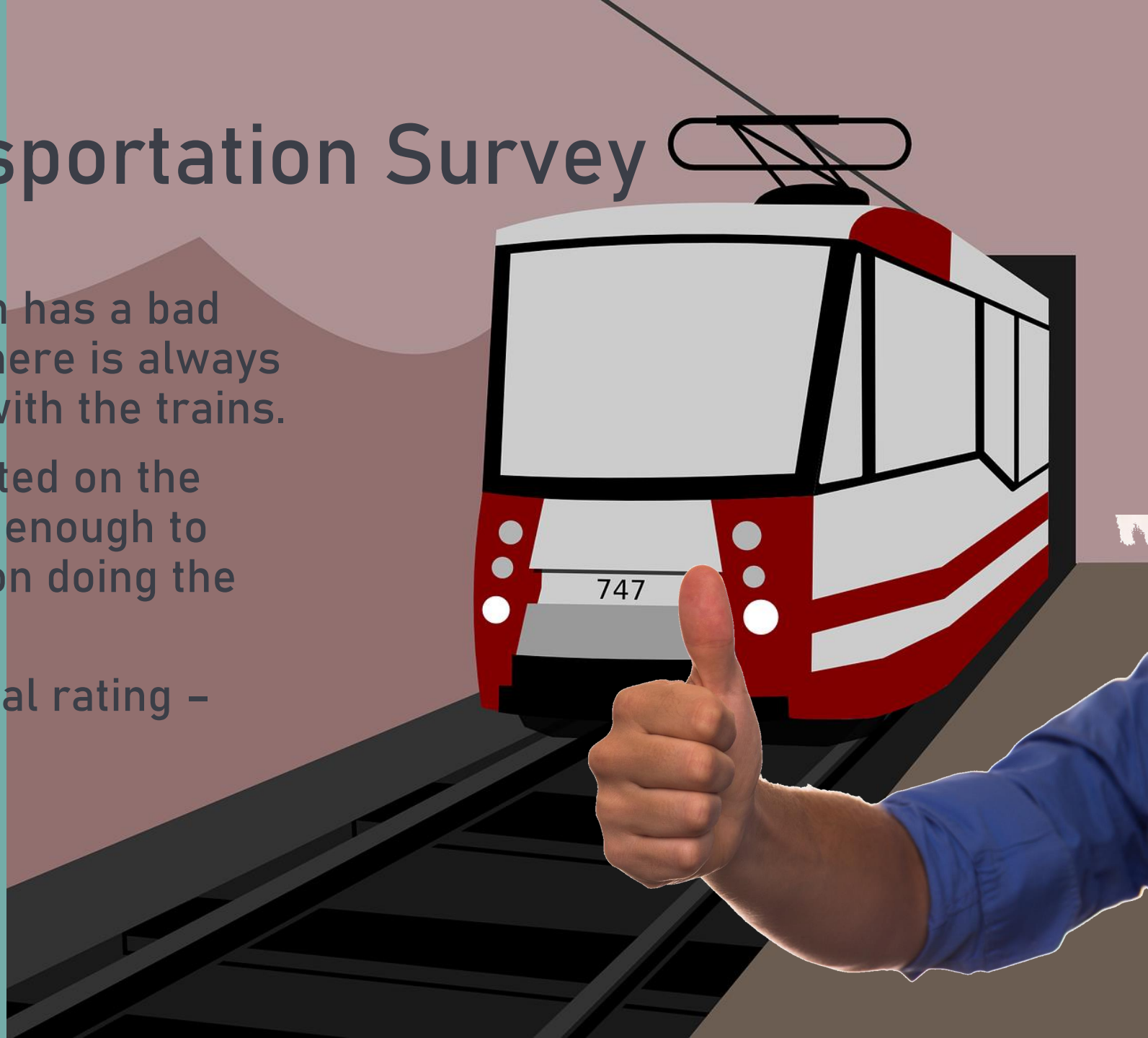
<https://dethwench.com/curated-datasets-for-data-science-portfolio-projects/>

Selection Bias

- Imagine you have a survey of 100 individuals, and you remove everyone who says “unknown” to a certain question.
- If you remove 2 respondents, no problem. But what if you remove 30 respondents – almost one third of the dataset?
- You need to troubleshoot this unusual situation. Why is the unknown category so popular?
- What do all those 30 people have in common?
 - Are they all uncomfortable with the question?
 - Maybe the question is confusing, so they don't know what to say
 - Or maybe they are all mad at the question!
- With almost one third of people missing, it's hard to believe that is random

Transportation Survey

- Subway system in Boston has a bad reputation. It is old and there is always something going wrong with the trains.
- In-person survey conducted on the platform with riders nice enough to stop and talk to the person doing the survey
- Found a very high approval rating – above 90%!



Use-Case #1 Scenario

- Fictitious Community Health and Nutrition Assessment (CHNA) taking place in Florida, United States
- Based on a true scenario using data from a large-scale health survey
- Extremely large dataset accrued from interviewing patients at multiple clinical centers in Florida
- Want develop an analytic dataset to summarize:
 - Gender and age demographics
 - Rates of Type II diabetes screening
 - Rates of behavioral risk factors: Alcohol use, tobacco use, and marijuana use

Use-Case #2 Scenario

- Fictitious health research institute called the Institute for Reverse Aging
- Trend among rich tech entrepreneurs
- Based on a true scenario using data from an ICU setting in outside the United States
- Families of patients in the ICU who are likely to pass away are contacted for consent for organ donation
- Not all patients will qualify for organ donation
 - If brain dead, cannot donate
 - There may be objections to even approaching the family socially (e.g., crime victims)
 - Cannot donate if organ harvesting plans not in place before death
 - Families may also not consent
- ICU wanted to increase the number consenting, but wanted to evaluate where they were dropping out in the donation process

Questions?

Facilitator:

Monika M. Wahi, MPH, CPH

Email: dethwench@gmail.com

Blog:
<https://dethwench.com/blog/>

LinkedIn:
<https://www.linkedin.com/in/dethwench/>

Follow our
company page
to make sure you don't
miss our
new and exciting
data science
educational events
and resources!

<https://www.linkedin.com/company/dethwench-professional-services/>