SCHARE

Tutorial Think-a-Thons

National Institutes of Health

Health Outcome Research Paradigm Shift: Understanding How Big Data Expands Knowledge

April 16, 2025

Deborah Duran, PhD · NIMHD
Elif Dede Yildirim, PhD · NIMHD
Mark Aronson, PhD · NIMHD

Part 2 of 3

SCHARE

Research Think-a-Thons

Novice **training webinars** for data science, cloud computing and research using Big Data

Generational career & discipline exchange

Think-a-Thons

3rd Wednesday of every month
2 pm

Goals:

- Upskill novice untrained users in data science and cloud computing
- Foster a research paradigm shift to use Big Data in health disparities/health outcomes research
- Promote use of Dark Data

1. TUTORIAL AND TARGETED THINK-A-THONS

- Monthly sessions (2 1/2 hours)
- Instructional/interactive
- Designed for new/experienced users
- Networking
- Mentoring and coaching
- Topics include:
  - Data Science 101
  - Terra
  - Social Determinants of Health analytics
  - Common Data Elements
  - AI readiness
  - Ethical and transparent AI

Launched April 2024

2. RESEARCH THINK-A-THONS

- Multi-career (students to senior investigators)
- Multi-discipline (data scientists and researchers)
- Featured datasets with guest experts leads
- Guest experts in topic areas, analytics, data sources etc. to provide guidance
- Generate research idea - decide design, datasets and analytics
- Learn Ethical AI
- Publications

Register:

bit.ly/think-a-thons

Think-a-Thon tutorials

bit.ly/think-a-thons

February	**Artificial Intelligence and Cloud Computing 101**
March	**SCHARE 1 – Accounts and Workspaces**
April	**SCHARE 2 – Terra Datasets**
May	**SCHARE 3 – Terra Google-hosted Datasets**
June	**SCHARE 4 – Terra SCHARE-hosted Datasets**
July	**An Introduction to Python for Data Science – Part 1**
August	**An Introduction to Python for Data Science – Part 2**
September	**SCHARE 5: A Review of the SCHARE Platform and Data Ecosystem**
October	**Preparing for AI 1: Common Data Elements and Data Aggregation**
November	**Preparing for AI 2: An Introduction to FAIR Data and AI-ready Datasets**
January	**Preparing for AI 3: Computational Data Science Strategies 101**
February/March	**Preparing for AI 4: Overview Prep for AI Summary with Transparency, Privacy, Ethics**
April	**Research Teams – SDoH and Health Disparities**
May	**Be a Part of the Future of Knowledge Generation 1: AI/Cloud Computing Basics and CDEs**
July	**Be a Part of the Future of Knowledge Generation 2: AI-Ready Datasets and Computations**

SPECIAL EVENTS

- SCHARE for **Educators** (Community Colleges and low-resource MSIs)
- SCHARE for **American Indian/ Alaska Native Researchers**
- SCHARE for **Coders and Programmers** to conduct research

SCHARE Research Think-a-Thon Teams

Team 1 investigates disparities in **HPV vaccine uptake** and cervical cancer burdens among underserved populations.

Team 2 examines potential factors impacting **cancer disparities**, including cancer survivor status, healthcare coverage, treatment and transportation availability, education, income, support systems, diet, and race/ethnicity.

Team 3 investigates **disparities in dental care access**, exploring key determinants such as out-of-pocket costs, preventive health behaviors, and socioeconomic factors.

Team 4 examines emergency department utilization, barriers to healthcare access, focusing on geographic disparities, demographics, insurance coverage, and transportation challenges affecting **access and utilization of health services**.

Team 5 explores the impact of environmental factors on multifaceted aspects of **breast cancer**.

SCHARE

Training pipeline

BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

Think-a-Thons training/mentoring pipeline

SCHARE Data Science Experts + **Think-a-Thons** (✓ Instructional ✓ Research) + **AIM-AHEAD** + **AnVil | HEAL | N3C | All of US | BioData Catalyst**

Using AI experts — *to train and mentor novice AI users* — *to upskill and mentor diverse perspectives in AI* — *to increase diverse perspectives in biomedical research*

Goal: "Upskilling"

- ✓ Data science specialists into health disparities and health outcomes research
- ✓ Health disparities/outcomes researchers into using big data and cloud computing

Target Audience:

- ✓ Novice untrained users in data science
- ✓ Data scientists with no or little research experience
- ✓ Resource and tool for Community Colleges and low-resource organizations

SCHARE

Workshop setup

BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

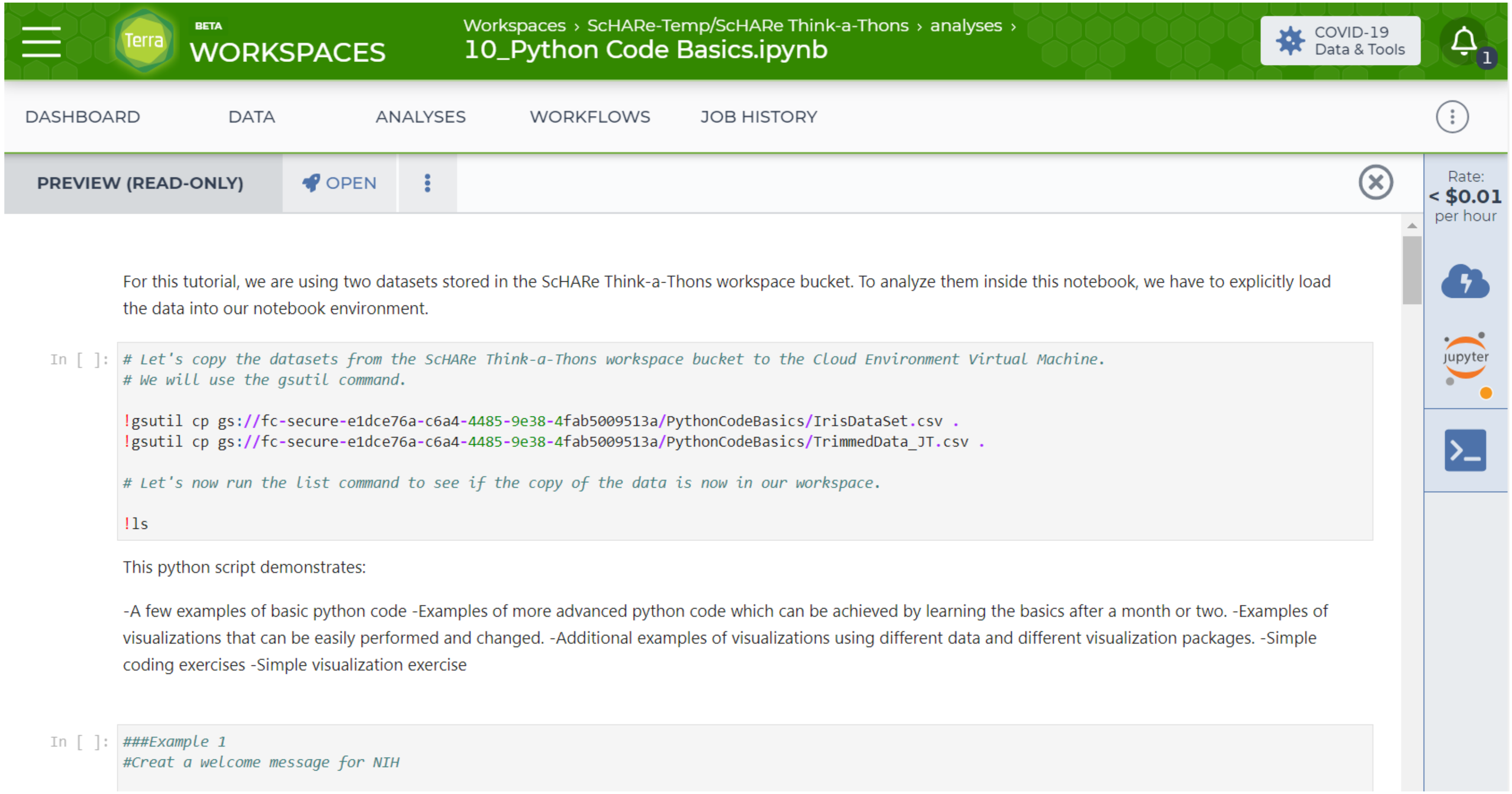
In preparation for the Think-a-Thon

Let's make sure that everyone:

1. **has provided their Gmail address and has been registered for SCHARE**
2. **has created a Terra account**
3. **can access the tutorial we will be using today at: bit.ly/schare-python-notebook**
4. **has configured their cloud environment**
5. **can run the tutorial in playground mode:**

If you have already created a Terra account and are logged in, you will see this:

bit.ly/schare-python-notebook



If you have not logged in, or have not yet created a Terra account, you will see this:

bit.ly/schare-python-notebook

If you have not yet created a Terra account or registered for SCHARE:

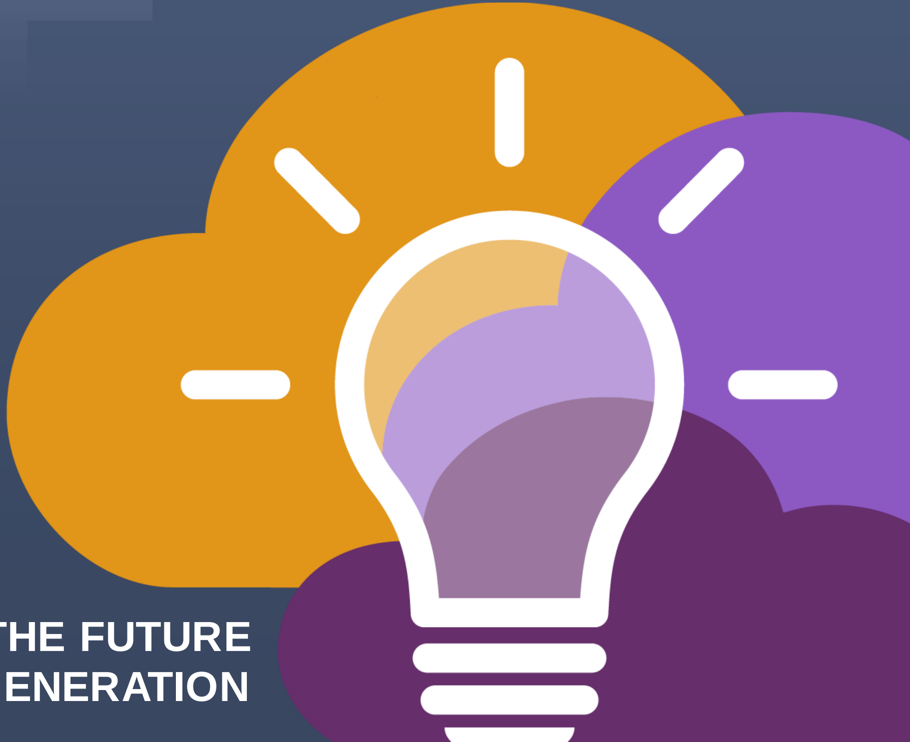
https://bit.ly/registerschare

All registered participants have been added to a free temporary billing project that will allow you to run the event materials with your instructors

- You will be active on this billing project for the duration of the Think-a-Thon. If you want to access work-in-progress after this time, you will need to set up your own billing and copy your workspaces to it

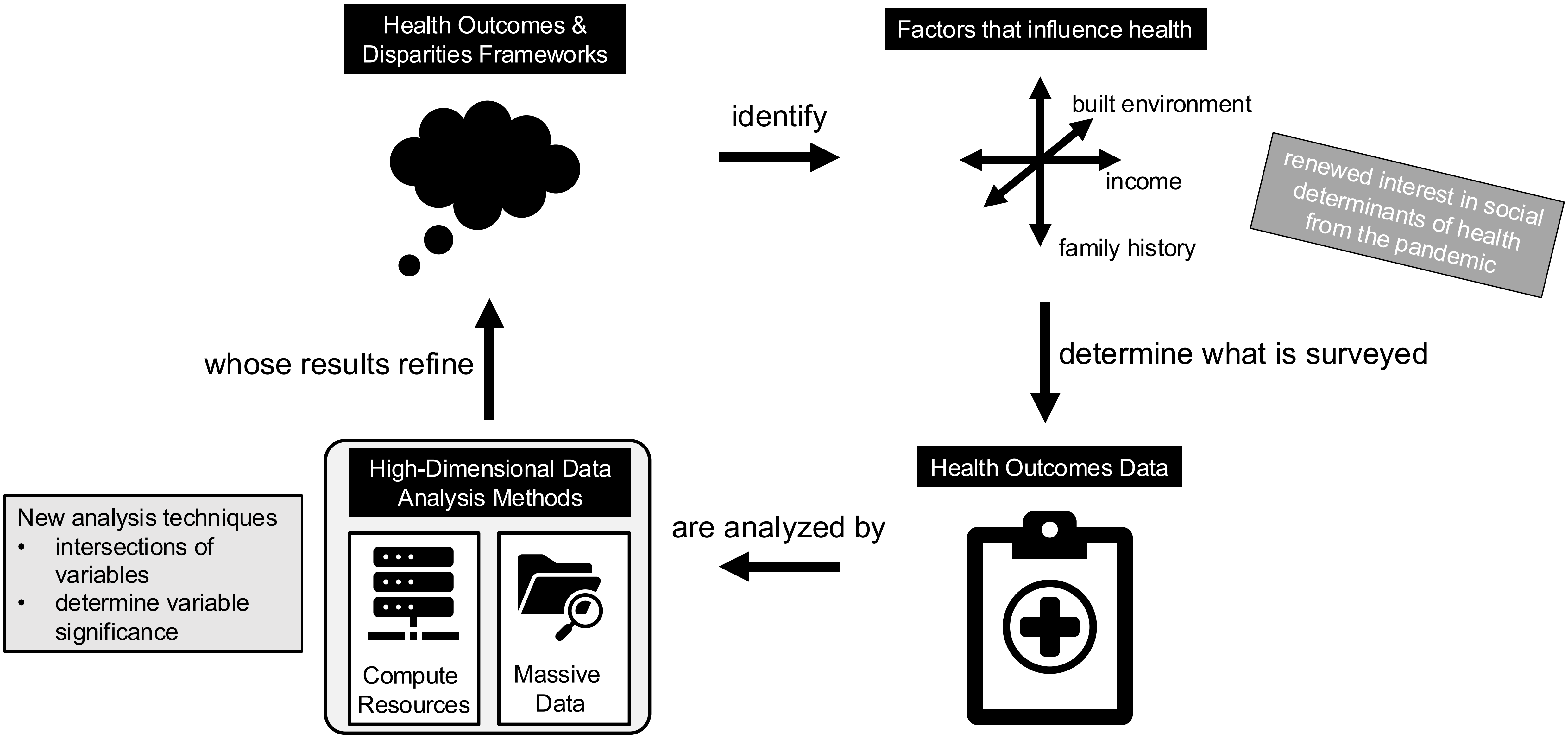
SCHARE

A conceptual model for using AI



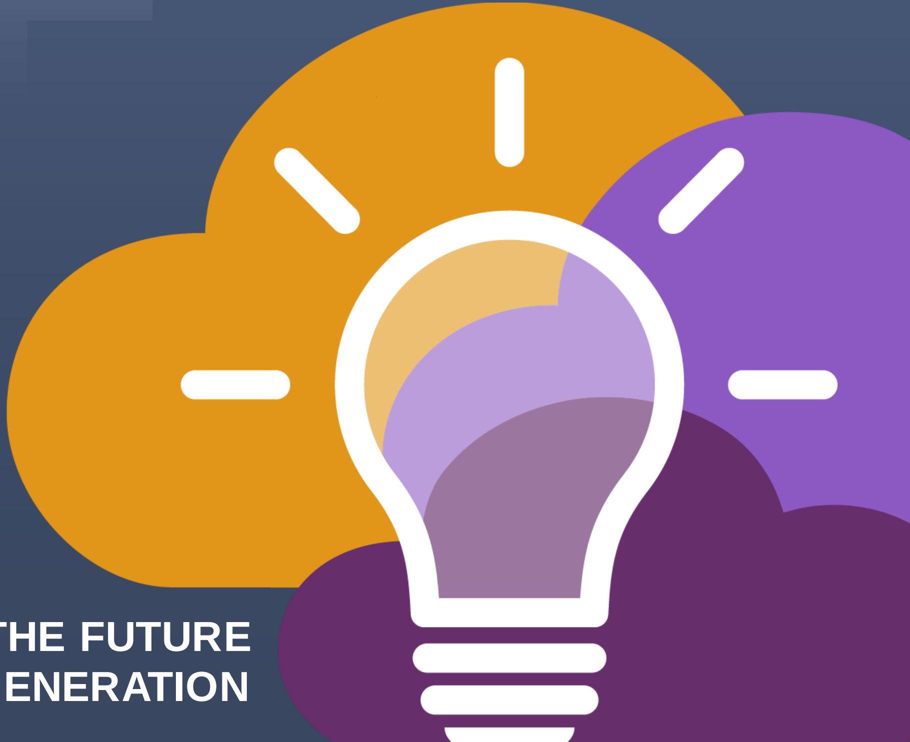
BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

Data science is poised to accelerate the health outcomes research cycle



SCHARE

Framing Questions for AI Approaches



BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

Analysis approaches are tailored to the question and data

	Statistical Models	Machine Learning Models
The Research Question		
The Data		

Research question drives the analysis strategy

The Research Question

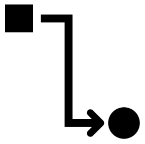
If your question is…

does changing X lead to a change in Y?

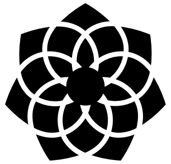
can we predict Y based on a bunch of X factors?

Then your question is of type…

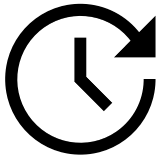
causal inference



pattern recognition



prediction



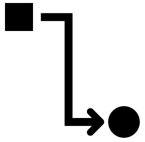
Analysis approaches are tailored to the question and data

Statistical Models

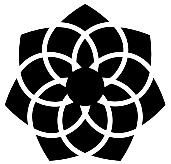
Machine Learning Models

The Research Question

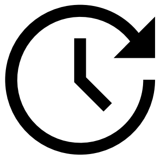
causal inference



pattern recognition



prediction



The Data

Machine learning models handle “wide” data better

The Data

number of variables

number of participants

Participant ID	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5	Variable 6
00001	0	0	2	2	23	0
00002	1	5	3	1	62	0
00003	1	3	1	5	72	0
00004	0	2	2	6	41	1

with “wide data” (number of variables >> number of observations), statistical models…

fit themselves to noise

do not generalize outside of the given data

does not distinguish between important and unimportant features

Analysis approaches are tailored to the question and data

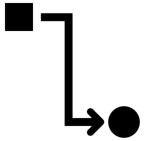
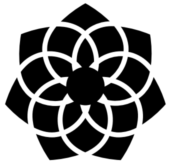
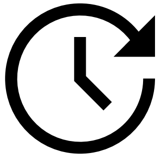
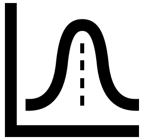
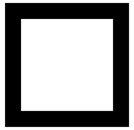
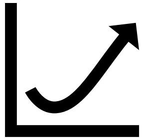
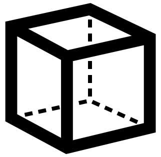
	Statistical Models	Machine Learning Models
The Research Question	causal inference	pattern recognition; prediction
The Data	“deep” data	“wide” data

A rule of thumb summary for data requirements

The Data

	Independence	Randomness	Normality	Linearity	Homoscedasticity	No multicollinearity
Statistical Models	✓	✓	✓	✓	✓	✓
Machine Learning Models	✓	✓	✗	✗	✗	✗

Analysis approaches are tailored to the question and data

	Statistical Models	Machine Learning Models
The Research Question 	causal inference 	pattern recognition  ; prediction 
The Data 	assumed distributions  ; “deep” data 	nonlinear relationships  ; “wide” data 

Example questions and suggested approaches

The Research Question	Determining Features	Suggested Approach
"How do socioeconomic status, race/ethnicity, and access to healthcare predict disparities in hypertension prevalence, and what are the estimated effect sizes of these factors?"	• defined variables of interest • desire to know effect size of each variable	Statistical Model (e.g. linear regression)
"Can we predict the risk of hospital readmission among patients from underserved communities using electronic health record (EHR) data, including demographic, clinical, and social determinants of health factors?"	• mix of many variable types • can handle correlated variables • desire to know personalized risk scores	Machine Learning Model (gradient boosting model, random forest)

Slido Poll

Why are traditional statistical models often not ideal for wide datasets (i.e., datasets with many more features than observations)?

a) They assume that all features are equally important, which leads to poor predictive performance
b) They tend to overfit the data and do not generalize well to new data
c) They perform automatic feature selection, which reduces model flexibility
d) While they handle missing data well, this is not that common for wide datasets, and therefore not a useful feature

SCHARE

Considerations for Datasets

BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

Data Categorizations

Data Format

Quantitative

Qualitative

Data Structure

Structured

Semi-Structured

Unstructured

Data Types

Tabular

Text

Time-Series

Geospatial

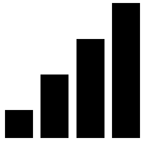
Image

Audio

Multimodal Data

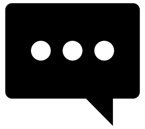
Data can be in quantitative or qualitative formats

Data Format



Quantitative

Quantitative data is any data that can be represented by a number, including numeric values, categorical data, binary data, and images



Qualitative

Qualitative data is any data representing information and concepts not captured by numbers, such as interview data

Data Categorizations

Data Format

Quantitative

Qualitative

Data Structure

Structured

Semi-Structured

Unstructured

Data Types

Tabular

Text

Time-Series

Geospatial

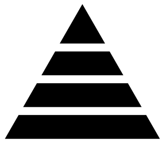
Image

Audio

Multimodal Data

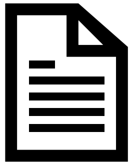
Data can be categorized by structure

Data Structure



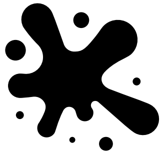
Structured

Structured data has a fixed schema and fits neatly into rows and columns



Semi-Structured

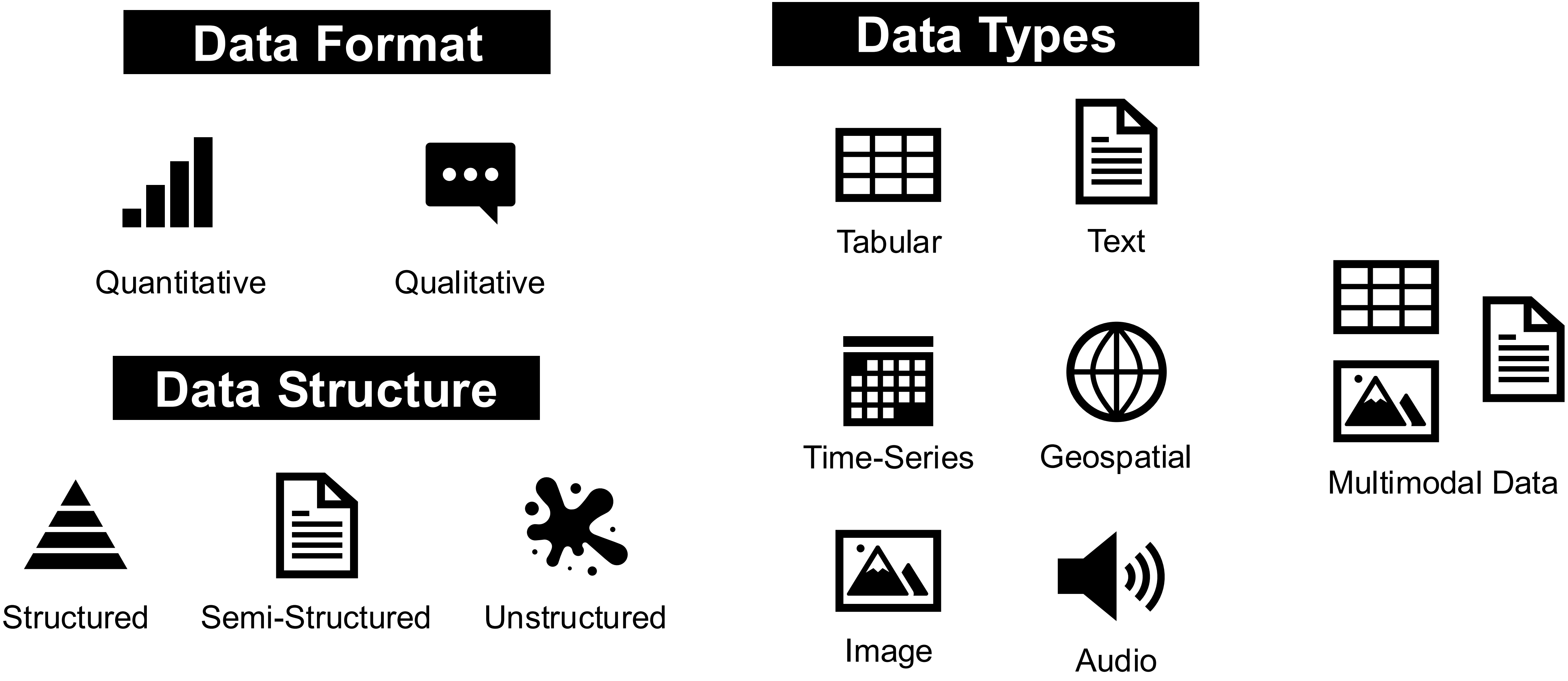
Semi-structured data contains elements of both structured and unstructured data, with some data fitting into rows and columns and some data that has open-ended data containing unstructured elements



Unstructured

Unstructured data doesn't fit neatly into a data table because its size or nature: for example, audio and video files and large text documents.

Tabular data is the dominant data type for health outcomes research



Data exists in many different types

Data Types

Tabular
Data represented in tables

Time-Series
Data representing events happening over time

Image
Data representing images using pixel values

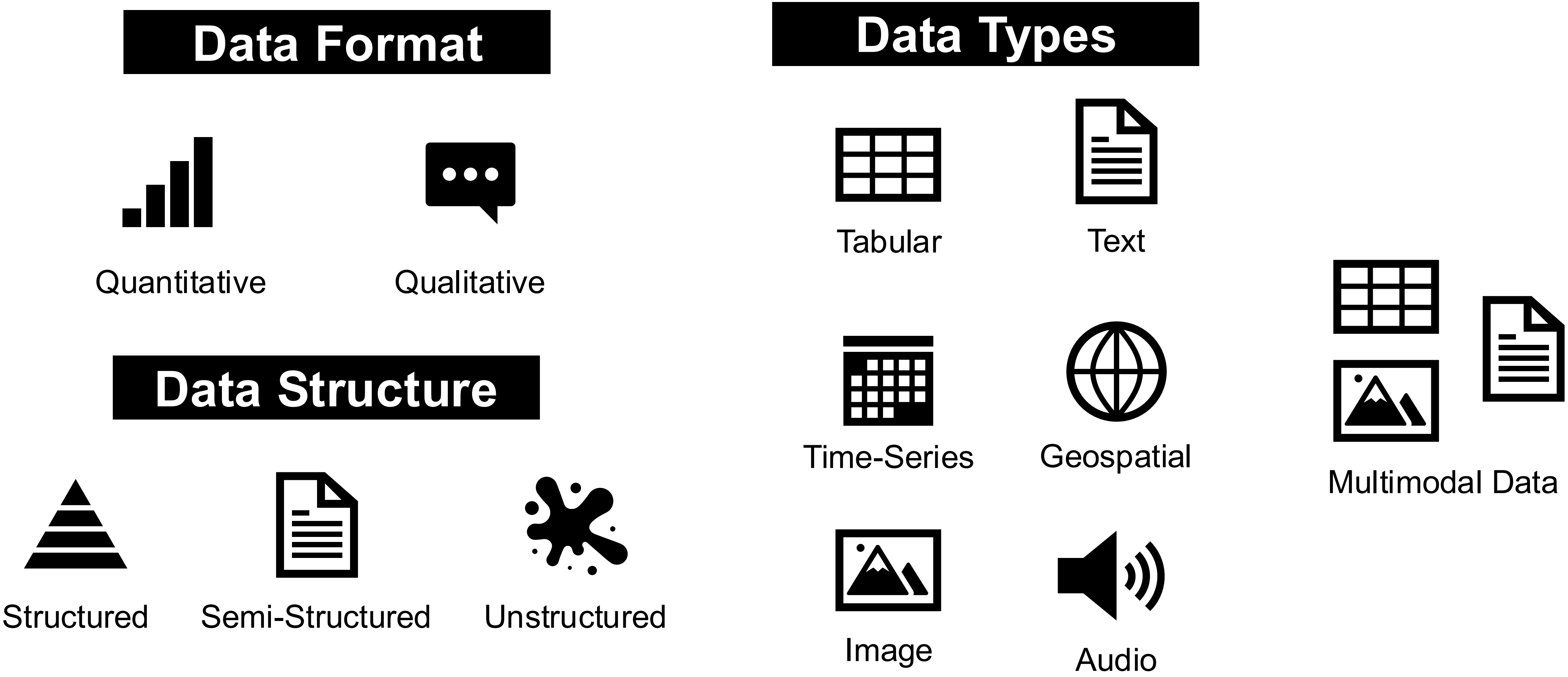
Text
Data composed of text

Geospatial
Data representing geospatial coordinates

Audio
Data representing audio

Multimodal Data
Data composed of a mix of data types

Tabular data is the dominant data type for health outcomes research



Example: SAMHSA National Mental Health Services Survey

SAMHSA
Substance Abuse and Mental Health Services Administration

National Mental Health Services Survey (N-MHSS)

The National Mental Health Services Survey (N-MHSS) is a source of national- and state-level data on the mental health services delivery system reported by both publicly and privately operated specialty mental health treatment facilities.

The Data (2020)

CASEID	LST	MHINTAKE	MHDIAGEVAL	MHREFERRAL	SMISEDSUD	TREATMT	ADMINSERV	SETTINGIP	SETTINGRC	SETTINGDTPH	SETTINGOP	FACILITYTYPE	FOCUS	OWNERSHP	PUBLICAGENCY	RELIG
202000001	AK	1	1	1	1	1	0	0	0	0	1	10	3	3	4	0
202000002	AK	1	1	1	1	1	1	0	0	0	1	10	3	3	4	0
202000003	AK	1	1	1	1	1	0	0	0	0	1	10	3	3	5	0
202000004	AK	1	1	1	1	1	0	0	0	0	1	7	3	3	5	0
202000005	AK	1	1	1	1	1	0	0	0	0	1	10	3	2	-2	0
202000006	AK	1	1	1	1	1	1	0	0	0	1	8	1	2	-2	0
202000007	AK	1	1	1	1	0	1	0	1	0	0	3	1	2	-2	1
202000008	AK	1	1	1	0	0	0	0	0	0	1	10	1	1	-2	0
202000009	AK	1	1	1	1	1	1	0	0	0	1	6	4	3	6	0
202000010	AK	1	1	1	0	0	1	0	0	0	1	7	1	2	-2	0
202000011	AK	1	1	1	1	1	1	0	1	0	0	5	1	2	-2	0
202000012	AK	1	1	1	1	1	1	0	0	0	1	10	3	3	4	0
202000013	AK	1	1	1	0	1	1	0	1	0	0	3	1	1	-2	0
202000014	AK	1	1	1	1	0	0	1	0	0	0	1	1	3	2	0
202000015	AK	1	1	1	0	0	0	0	1	0	0	5	1	2	-2	1
202000016	AK	1	1	1	1	0	0	0	0	0	1	10	1	2	-2	1
202000017	AK	1	1	1	1	0	1	0	0	0	1	10	1	2	-2	0
202000018	AK	1	1	1	1	1	1	0	0	0	1	7	3	2	-2	0
202000019	AK	1	1	1	1	1	1	0	0	1	1	8	1	2	-2	0
202000020	AK	1	1	1	0	0	1	0	0	0	1	8	1	2	-2	0
202000021	AK	1	1	1	1	1	1	0	0	0	1	10	3	3	5	0
202000022	AK	1	1	1	0	0	1	0	0	0	1	10	1	2	-2	0
202000023	AK	1	1	1	0	0	1	0	0	0	1	7	1	2	-2	0
202000024	AK	1	1	0	0	0	0	1	0	0	0	2	1	2	-2	1
202000025	AK	1	1	1	1	0	0	0	0	0	1	10	1	1	-2	0

Example Data: Binary Values

Definition

Variables that take on either a value of 0 or 1

Codebook

MHINTAKE: Facility offers mental health intake (Q.A1)

Value	Label	Frequency	%
0	No	990	8.1%
1	Yes	11,285	91.9%
	Total	*12,275*	*100%*

Variable Type: numeric

Raw Data

CASEID	LST	MHINTAKE
202000028	AK	1
202000029	AK	1
202000030	AK	1
202000031	AK	0
202000032	AK	1
202000033	AK	1
202000034	AK	1
202000035	AK	1
202000036	AK	1
202000037	AK	1
202000038	AK	1
202000039	AK	0
202000040	AK	0
202000041	AK	1
202000042	AK	0
202000043	AK	0
202000044	AK	0
202000045	AK	0
202000046	AK	1

Example Data: Categorical Values

Definition

Variables that take on one of a defined preset range of values

Codebook

FACILITYTYPE: Facility type (Q.A4)

Value	Label	Frequency	%
1	Psychiatric hospital	668	5.4%
2	Separate inpatient psychiatric unit of a general hospital	967	7.9%
3	Residential treatment center for children	592	4.8%
4	Residential treatment center for adults	807	6.6%
5	Other type of residential treatment facility	63	0.5%
6	Veterans Administration Medical Center (VAMC)	552	4.5%
7	Community Mental Health Center (CMHC)	2,548	20.8%
8	Certified Community Behavioral Health Clinic (CCBHC)	336	2.7%
9	Partial hospitalization/day treatment facility	429	3.5%
10	Outpatient mental health facility	4,941	40.3%
11	Multi-setting mental health facility	369	3.0%
12	Other	3	0.0%
	Total	*12,275*	*100%*

Variable Type: numeric

Raw Data

CASEID	FACILITYTYPE
202000001	10
202000002	10
202000003	10
202000004	7
202000005	10
202000006	8
202000007	3
202000008	10
202000009	6
202000010	7
202000011	5
202000012	10
202000013	3
202000014	1
202000015	5
202000016	10
202000017	10
202000018	7
202000019	8
202000020	8
202000021	10
202000022	10

Example Data: Numeric Data

Definition

Variables that have a numeric variable

Codebook

TOTADMIS: Number of mental health treatment admissions in previous 12-month period (Q.B7)

Value	Label	Frequency	%
0	None	394	3.2%
1	1 to 10	581	4.7%
2	11 to 20	352	2.9%
3	21 to 30	244	2.0%
4	31 to 40	206	1.7%
5	41 to 50	262	2.1%
6	51 to 75	452	3.7%
7	76 to 100	364	3.0%
8	101 to 250	1,244	10.1%
9	251 to 500	1,227	10.0%
10	501 to 1000	1,224	10.0%
11	1001 to 1500	529	4.3%
12	More than 1500	1,026	8.4%
-1	Missing	1,395	11.4%
-2	Logical skip	2,775	22.6%
	Total	*12,275*	*100%*

Variable Type: numeric

Raw Data

CASEID	TOTADMIS
202000001	2
202000002	1
202000003	6
202000004	5
202000005	-1
202000006	-1
202000007	1
202000008	0
202000009	-1
202000010	-2
202000011	6
202000012	1
202000013	5
202000014	10
202000015	9
202000016	-1
202000017	2
202000018	6
202000019	-2
202000020	-2
202000021	8
202000022	3
202000023	-2

Slido Poll

Which of the following best describes the difference between structured and unstructured data?

a) Structured data resides in fixed fields and formats, whereas unstructured data lacks a predefined data model.
b) Unstructured data can be stored in relational databases, while structured data cannot.
c) Structured data includes formats like images and audio, while unstructured data includes spreadsheets and SQL tables.
d) Structured data is always numerical, while unstructured data is always textual.