Data Literacies and E(e)thics

Defining Data

More than just numbers

- At its core, data is a collection of facts, figures, observations, or characteristics that are recorded and organized
- In social science, data represents aspects of human behavior, attitudes, beliefs, social structures, and interactions
- It can be qualitative (descriptive, non-numerical) or quantitative (numerical)
- Think of data as the raw material we use to understand and explain the social world around us

Data in Social Science

- Social scientists use data to:
    - Describe social phenomena (e.g., poverty rates, voting patterns)
    - Explain why certain social patterns exist (e.g., the impact of education on income)
    - Predict future social trends (e.g., the spread of social movements)
    - Evaluate the effectiveness of social programs and policies
- Data helps us move beyond anecdotal evidence and personal opinions towards more systematic and objective understandings
- It allows for the identification of patterns, relationships, and trends within and across societies

Structured vs. Unstructured Data

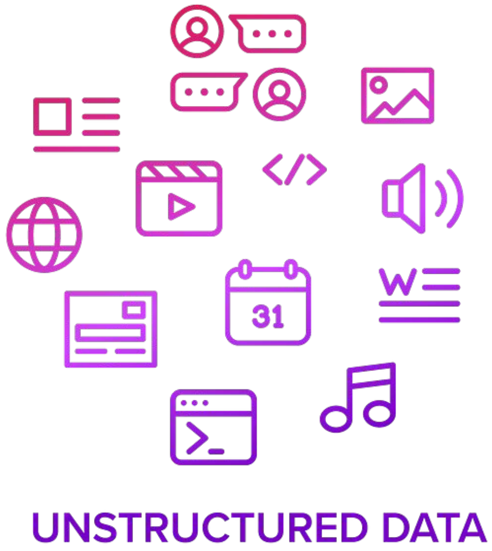
Structured data

- Data that is highly organized and fits neatly into predefined formats (like tables or databases)
- Easy to search, analyze, and manage; Often numerical or categorical with clear labels
- Examples: Survey data with fixed response options, census data, administrative records

Structured vs. Unstructured Data

Unstructured data

- Data that does not have a predefined format or organization
- More complex to analyze, requires specialized tools and techniques. Often rich in context and detail
- Examples: Interview transcripts, social media posts, open-ended survey responses, field notes



Basic data types

Quantitative Data: Numerical data that can be measured and statistically analyzed

- Examples: Age, income, test scores, frequency of an event

Qualitative Data: Non-numerical data that describes qualities or characteristics

- Examples: Interview excerpts, observational notes, textual documents

Basic data types

Big Data: Extremely large and complex datasets that are difficult to process with traditional data processing. Characterized by volume, velocity, variety, and veracity

- Social Science Examples: Analyzing social media trends, large-scale online survey data

Metadata: “Data about data" It provides information about the characteristics of a dataset

- Examples: Date of collection, source of data, variable definitions, data format. Crucial for understanding and using data effectively

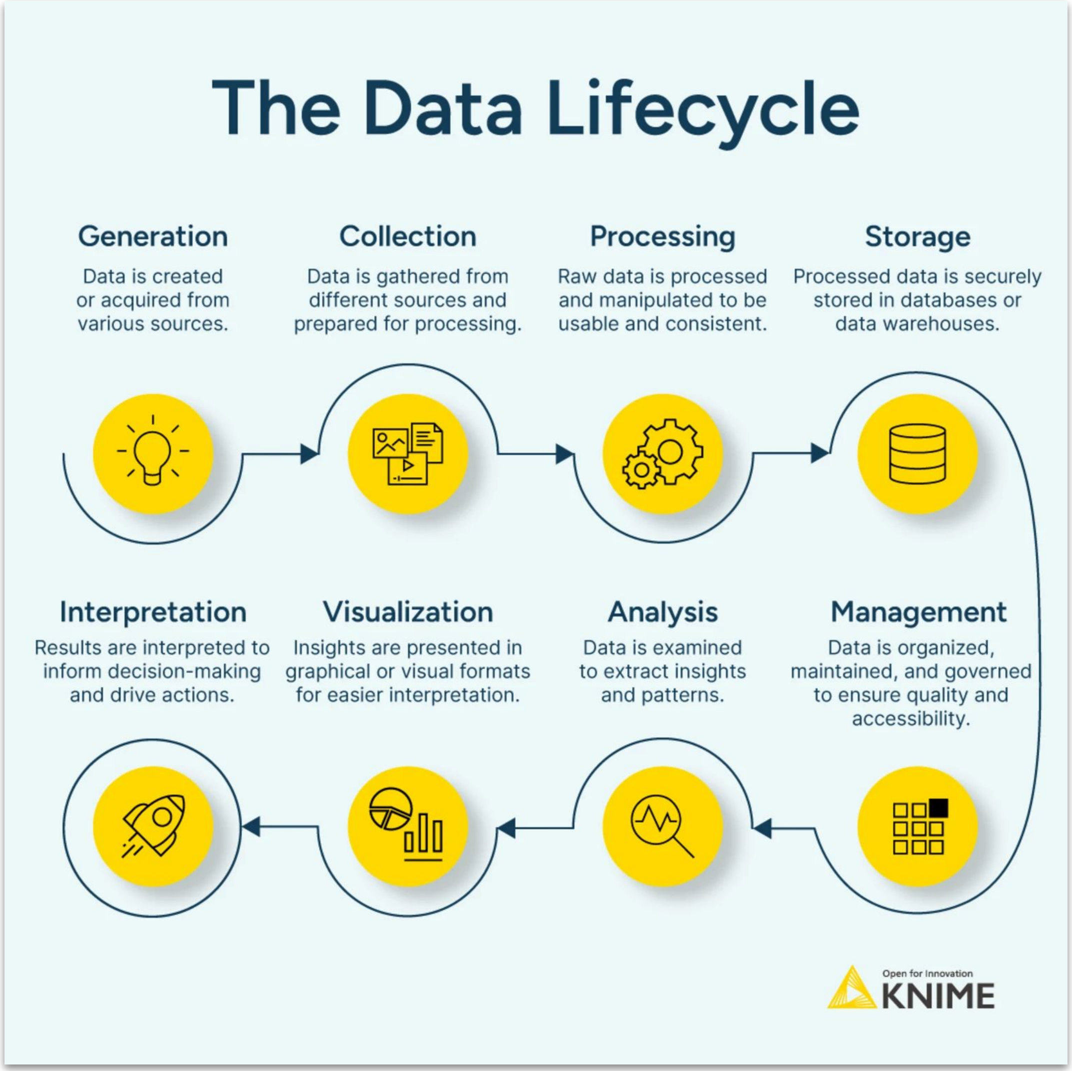
Using Data - From Collection to Analysis

Data Collection: Choosing appropriate methods (surveys, interviews, experiments, observations, existing datasets).

Data Cleaning and Preparation: Addressing errors, inconsistencies, and missing values. Organizing data for analysis.

Data Analysis: Applying statistical techniques (for quantitative data) or thematic analysis (for qualitative data) to identify patterns and insights.

Interpretation: Making sense of the findings and relating them back to the research question or social phenomenon being studied.



The Data Lifecycle

Generation

Data is created or acquired from various sources.

Collection

Data is gathered from different sources and prepared for processing.

Processing

Raw data is processed and manipulated to be usable and consistent.

Storage

Processed data is securely stored in databases or data warehouses.

Management

Data is organized, maintained, and governed to ensure quality and accessibility.

Analysis

Data is examined to extract insights and patterns.

Visualization

Insights are presented in graphical or visual formats for easier interpretation.

Interpretation

Results are interpreted to inform decision-making and drive actions.

Open for Innovation
KNIME

Sharing and communicating data

- Visualizations: Using charts, graphs, and other visual tools to communicate findings effectively (e.g., bar charts, scatter plots, infographics)
- Reports and Publications: Presenting detailed findings, methodologies, and interpretations in written form
- Presentations: Summarizing key findings and using visuals to engage an audience
- Data Repositories (Ethical Considerations): Sharing anonymized datasets (when appropriate and ethical) to allow for replication and further research

Effective and responsible communication of data insights is crucial for informing social understanding, policy, and action

Data everywhere?

Living in a data saturated world

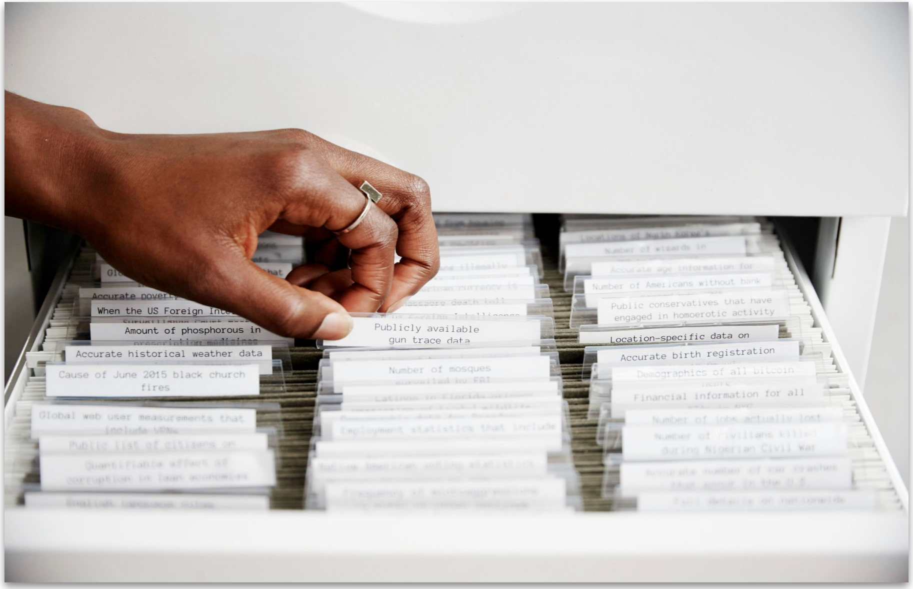
Context matters, local knowledge

yinz

Context matters, local knowledge

How *yinz* doing?

The Library of Missing Datasets (*GitHub*)


Accurate poverty
When the US Foreign Inte
Amount of phosphorous in
Accurate historical weather data
Cause of June 2015 black church fires
Global web user measurements that include VPNs
Public list of citizens on
Quantifiable affect of corruption in Iaan economies
Publicly available gun trace data
Number of mosques surveilled by FBI
Employment statistics that include
Locations of Boko Haram's
Number of wizards in
Accurate age information for
Number of Americans without bank
Public conservatives that have engaged in homoerotic activity
Location-specific data on
Accurate birth registration
Demographics of all bitcoin
Financial information for all
Number of jobs actually lost
Number of civilians killed during Nigerian Civil War
Accurate number of car crashes that occur in the U.S.
Full details on nationwide

Ethics = *IRB*

Ethics = Only *IRB?*

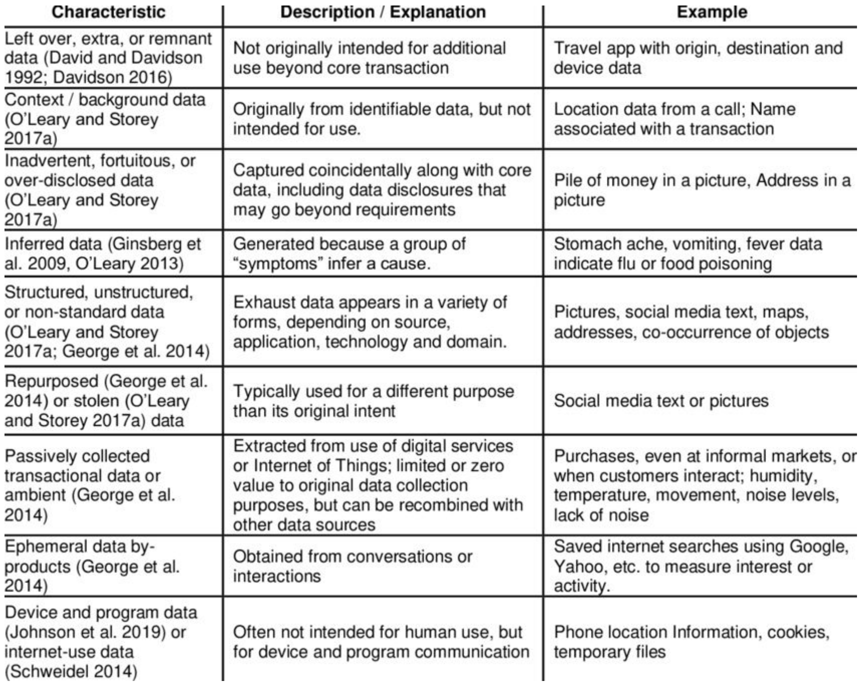
Public data… is it really free to use?

Public data = Consent?

- Police surveillance and facial recognition

Data exhaust/digital trace data

- Pubic social media posts
- Location tracking during COVID-19 outbreak



Characteristic	Description / Explanation	Example
Left over, extra, or remnant data (David and Davidson 1992; Davidson 2016)	Not originally intended for additional use beyond core transaction	Travel app with origin, destination and device data
Context / background data (O'Leary and Storey 2017a)	Originally from identifiable data, but not intended for use.	Location data from a call; Name associated with a transaction
Inadvertent, fortuitous, or over-disclosed data (O'Leary and Storey 2017a)	Captured coincidentally along with core data, including data disclosures that may go beyond requirements	Pile of money in a picture, Address in a picture
Inferred data (Ginsberg et al. 2009, O'Leary 2013)	Generated because a group of "symptoms" infer a cause.	Stomach ache, vomiting, fever data indicate flu or food poisoning
Structured, unstructured, or non-standard data (O'Leary and Storey 2017a; George et al. 2014)	Exhaust data appears in a variety of forms, depending on source, application, technology and domain.	Pictures, social media text, maps, addresses, co-occurrence of objects
Repurposed (George et al. 2014) or stolen (O'Leary and Storey 2017a) data	Typically used for a different purpose than its original intent	Social media text or pictures
Passively collected transactional data or ambient (George et al. 2014)	Extracted from use of digital services or Internet of Things; limited or zero value to original data collection purposes, but can be recombined with other data sources	Purchases, even at informal markets, or when customers interact; humidity, temperature, movement, noise levels, lack of noise
Ephemeral data by-products (George et al. 2014)	Obtained from conversations or interactions	Saved internet searches using Google, Yahoo, etc. to measure interest or activity.
Device and program data (Johnson et al. 2019) or internet-use data (Schweidel 2014)	Often not intended for human use, but for device and program communication	Phone location Information, cookies, temporary files

Context matters, local knowledge

Small “e” ethics

- SAFELab at University of Pennsylvania
- Bronx Community Research Review Board

Think, Group, Share

Mental Health Study

You're interested in understanding how youth talk about their mental health online, especially how they engage in self-care and potential harmful practices. You have decided that you would be collecting ***public*** social media posts across a variety of social media platforms.

Community review board (5 mins)

- Why do you need this data?
- What are some of the ethical concerns?
  - How might you address them?
- How might you share your data and analysis with our communities?
- How might you deal with unintended consequences of your data, analyses, and/or visualization?

Group share out (10 mins)

Taking turns, each group shares one response

- Goal: Populate a guideline for the work we do
  - Focus on what has not been shared – perhaps new questions are emerging from the group share?

Minute madness style - try to keep your share-out under one minute!

Mental Health Study II

As you engaged mid-analysis, you realized that a large community of users on Instagram has made their posts private when following a particular tag.

a) How would this change your data analysis (if any)?

b) What are steps you might take to plan for such changes?

Mental Health Study III

As you engaged mid-analysis, you realized that a community of users on Reddit have been sharing strategies which have been coded to bypass censorship (you have gained knowledge of how users discuss in code).

a) What would you do as a researcher?

b) How does this impact your research study?

The fallacy of AI

Reducing bias?

Rise of predictive…

- policing with LAPD’s PredPol system
  - Ended amidst community concerns of reinforcing systematic bias
- child welfare screening with Allegheny Family Screening Tool
  - Designed to augment decision making
  - If purely automated, racial disparity rose to 20% compared to 11.3% pre-AFST
- hiring tools with HireVue
  - Who is the ideal candidate?

Lunch Hour

Lunch: Conversation Tables

Groups with guided conversation for part of today's lunch

In the room next door (145), there are four tables marked with various topics: 1) network analysis, 2) natural language processing, 3) data mining, 4) ethics & AI

- Grab your lunch and seat at a table with a topic you are interested in
  - These are informal conversations and it is okay if you're unsure what topic you're most interested in
- Have group conversations for ~20 minutes, then take a break, make a button, walk around outside, or do whatever you need until we begin again at 1:00pm