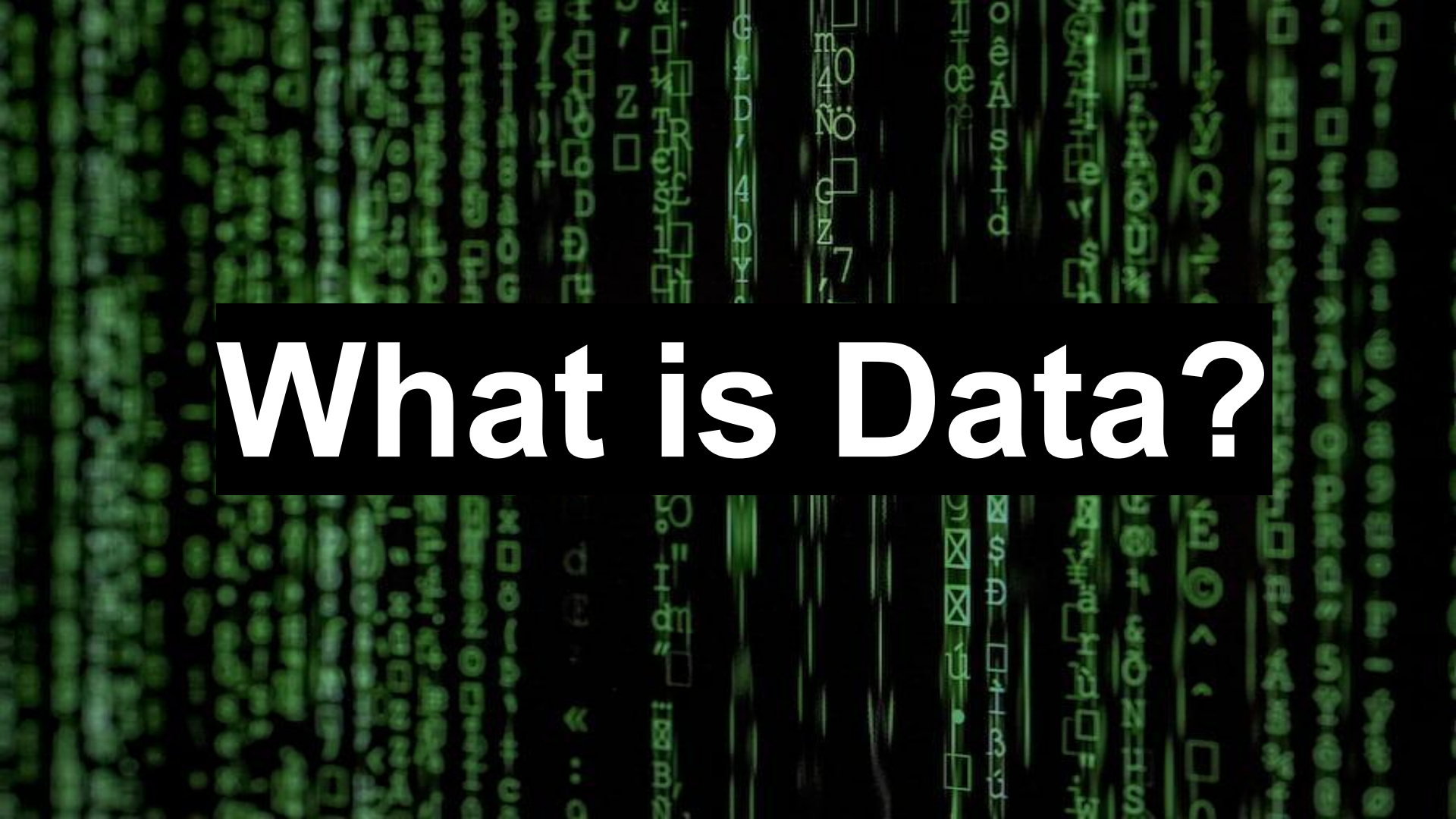
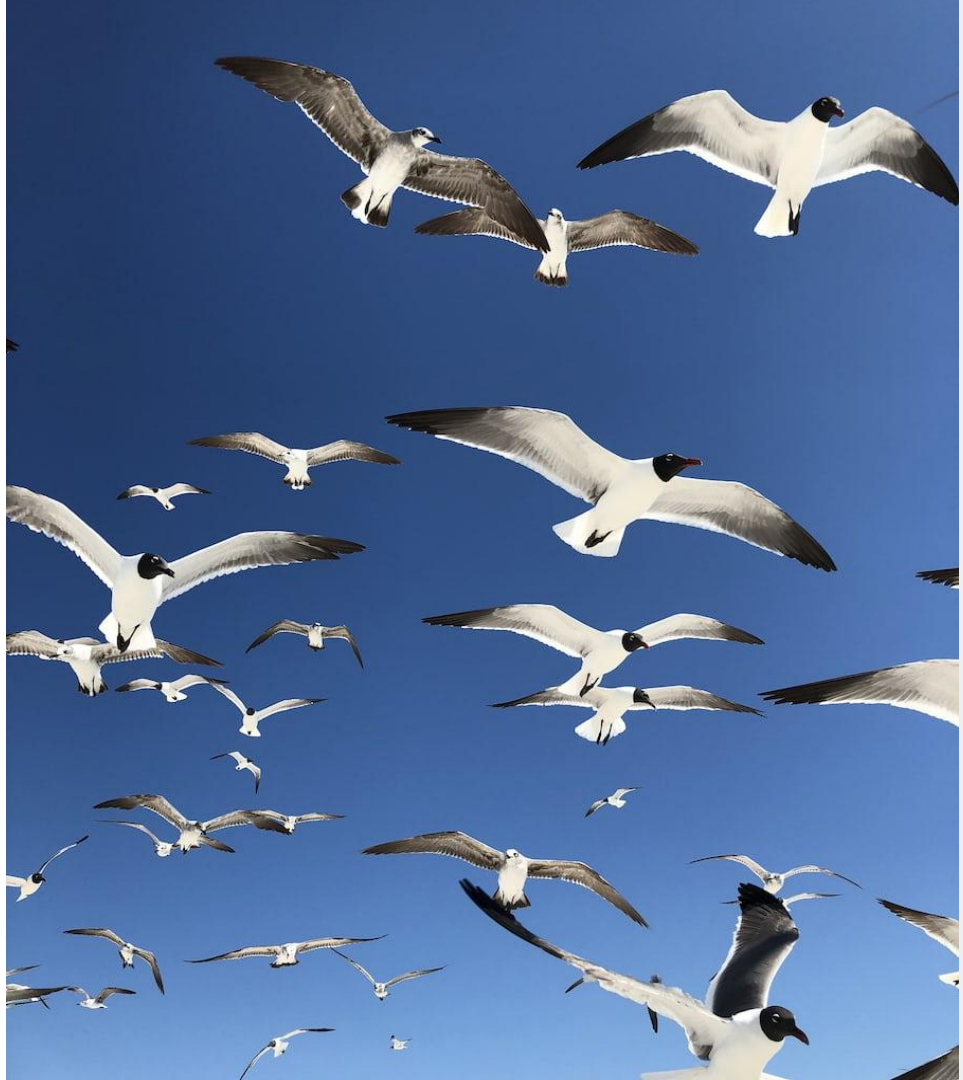
Finding and Evaluating Datasets

Data Bootcamp - Summer 2023

Finding Dataset Guide:

https://libguides.colorado.edu/findingdatasets/2023

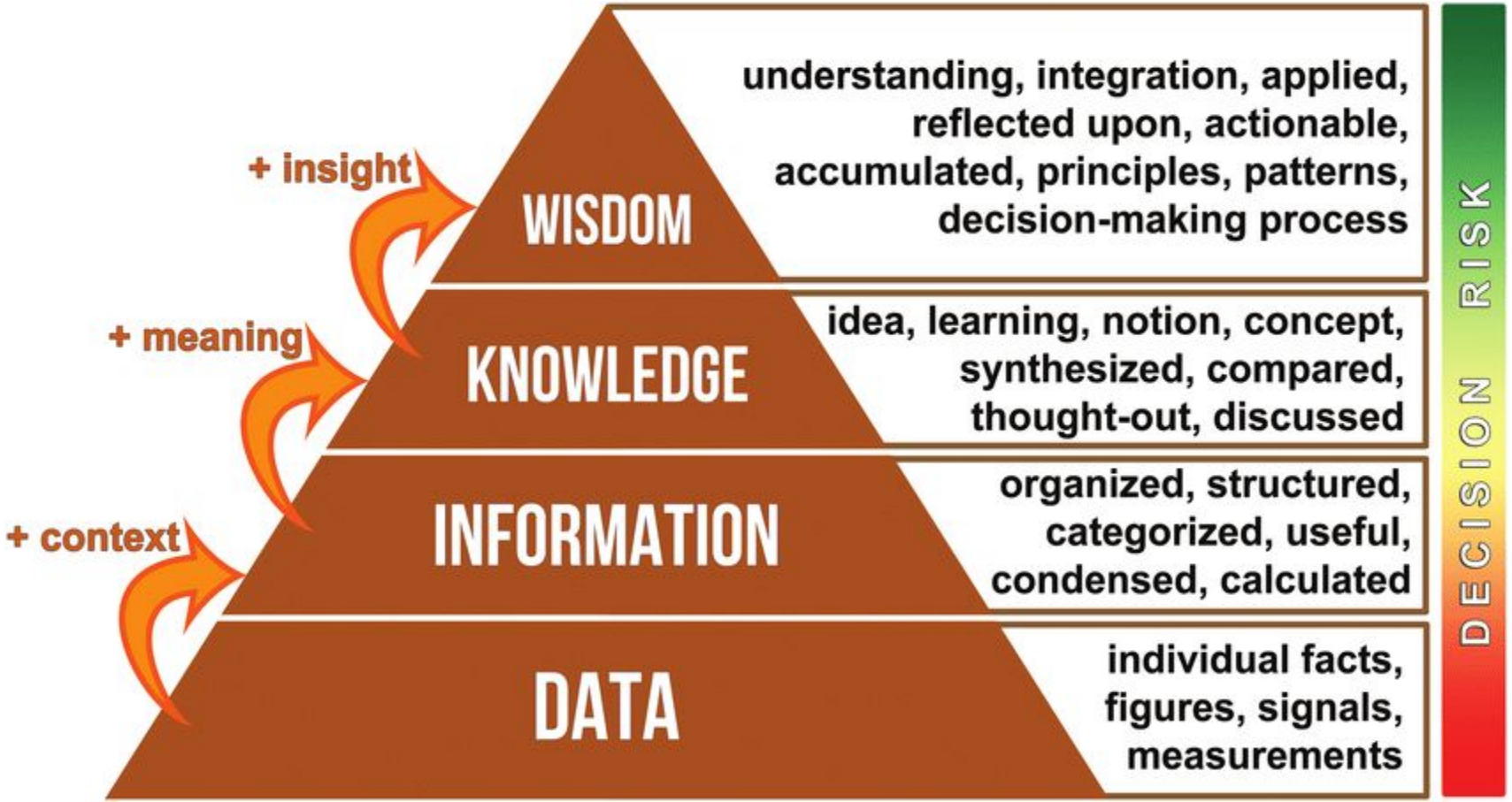

What is Data?



Data are…

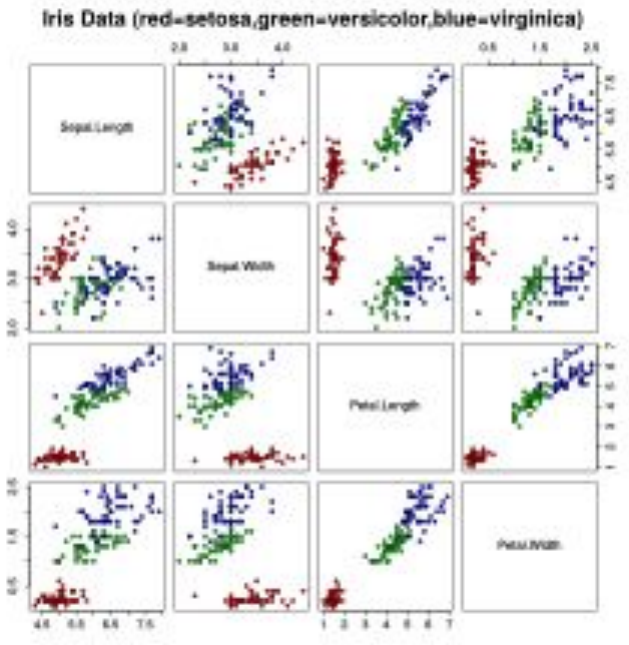
- …discrete values or units of information
  - numbers, words, characters, images, sound recordings, videos.
- ...anything that can be collected, stored, organized, and analyzed.

Image by Ryan Skjervem on Unsplash



The data–information–knowledge–wisdom (DIKW) hierarchy © Luis Tedeschi, CC BY-NC 4.0

What is a dataset?



- **Grouped information collected, assembled, and organized**
- **Used for analysis of an issue, phenomenon, or subject**
- **Can be textual, numeric, images, sounds, video, code, geospatial**
- **Many formats**
  - **CSV, XML, TIFF, PDF, etc,**

“Traditional” vs. “Big” Data

- structured format
- stored in relational databases (uses tables, each row has a unique identity or ‘key’)
- Uses “traditional” storage methods and tools
- Majority of datasets you’ll find to work with

- Large volume data (too big for relational databases)
- Variety of formats/types of data
- Receives data at high velocity (is constantly being created and stored)
- Stored in data warehouses (w/software framework such as Hadoop) or in the cloud

Good Data vs. Bad Data

Good Data has…

1. **Documentation:**
   a. Method of data collection/ generation/manipulation
   b. Clear variables/column headers
   c. COIs/funding
2. **License:**
   a. Does it have one? Is it permissive enough for your use case?
3. **Usable Format:**
   a. Is it *.stata that you can only use with STATA, or is it a *.csv that you can use with anything?
   b. Do you need *streaming* data, or is *static* data ok?

Also…

1. **Complete data:**
    a. No gaps
2. **Good Metadata:**
    a. Accurate labels
3. **Minimal Cleaning**
    a. You can use the dataset with relatively little manual correction or improvements or adjustment

Image by Setyaki Irham on Unsplash

"bad" data

- **Hinders/delays analysis**
- **Can lead to: inaccurate/harmful conclusions**

Image from this post

1. Incomplete
2. Unknown sources
3. Poorly documented methodology
4. Irregular entries
    a. Happens easily with multiple contributors or sources of data
5. Contains errors
6. Inaccurate/sloppy metadata/labels
    a. Date format (and timestamp)
    b. Fractions, measurements
7. Old, expired, not relevant

Examples: “Dirty” or “Unclean” Data

Misspellings or alternate spellings

Non-standard dates, names, etc

Missing field values

Duplicate records

Different data formats

Outdated data

Metadata

- Provides information about contents, format, and accessibility:
  - Creator/author(s), purpose, time, location, type of data, origin, purpose, time, geographic location, creator, access, and terms of use
- Used for retrieval, indexing and citation

Examples of metadata standards:

- Astronomy Visualization Metadata
- Darwin Core
- Data Documentation Initiative (DDI) to document numeric data files
- Dublin Core, a general purpose metadata standard
- ISO 19115 or FGDC's Content Standard for Digital Geospatial Metadata for geospatial data
- Ecological Metadata Language

Title
Name of the project or collection of datasets

Creator
Names and institutions of the people who created the data

Date
Key dates associated with the data, such as dates covered by the data or date of creation

Description
Description of the resource

Keywords or Subjects
Keywords or subjects describing the content of the data

Identifier
Unique number or alphanumeric string used to identify the data like a DOI. Many repositories provide DOIs for deposited datasets.

Coverage (if applicable)
Geographic coverage

Language
Language of the resource

Publisher
Entity responsible for making the dataset available

Funding Agencies
Organization or agency who funded the research

Access restrictions
Where and how your data can be accessed by other researchers

License
E.g., CC0, CC By 4.0, MIT, etc. See the ReDATA license matrix for help selecting a license.

Format

Title
Name of the dataset or research project that produced it

Creator
Names and addresses of the organization or people who created the data

Identifier
Number used to identify the data, even if it is just an internal project reference number

Dates
Key dates associated with the data, including project start and end date, data modification data release date, and time period covered by the data

Subject
Keywords or phrases describing the subject or content of the data

Funders
Organizations or agencies who funded the research

Rights
Any known intellectual property rights held for the data

Language
Language(s) of the intellectual content of the resource, when applicable

Location
Where the data relates to a physical location, record information about its spatial coverage

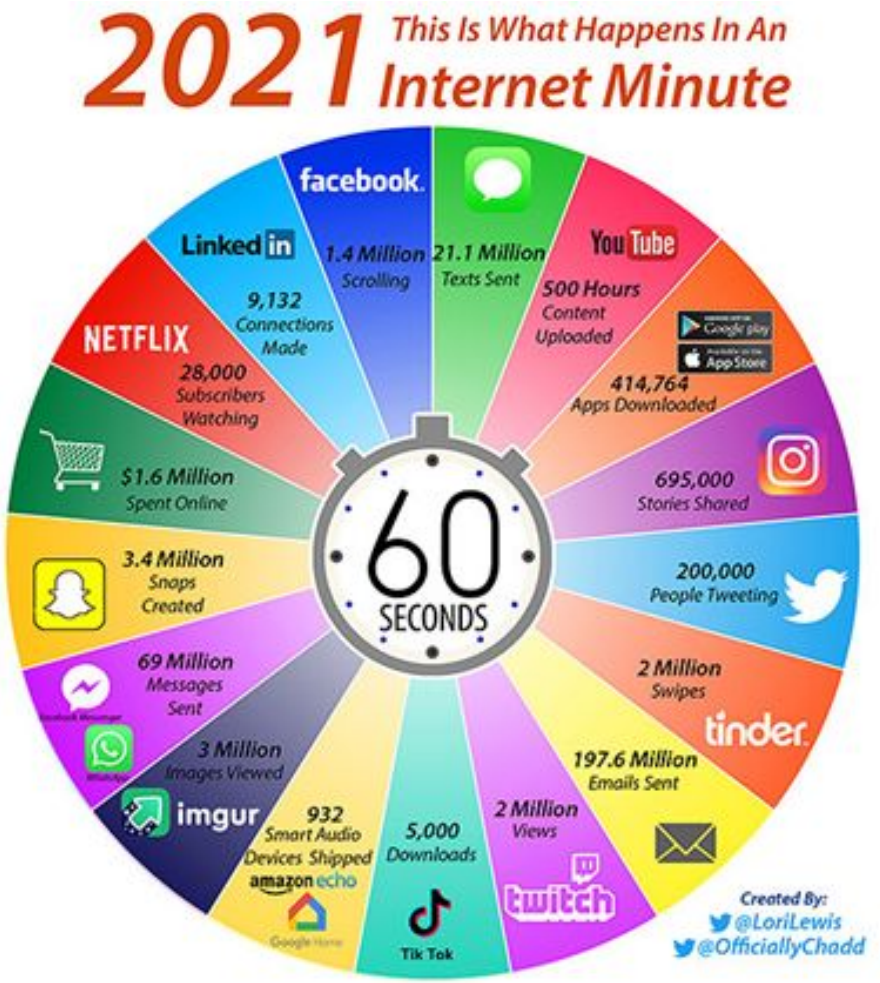
Methodology
How the data was generated, including equipment or software used, experimental protocol, other things you might include in a lab notebook

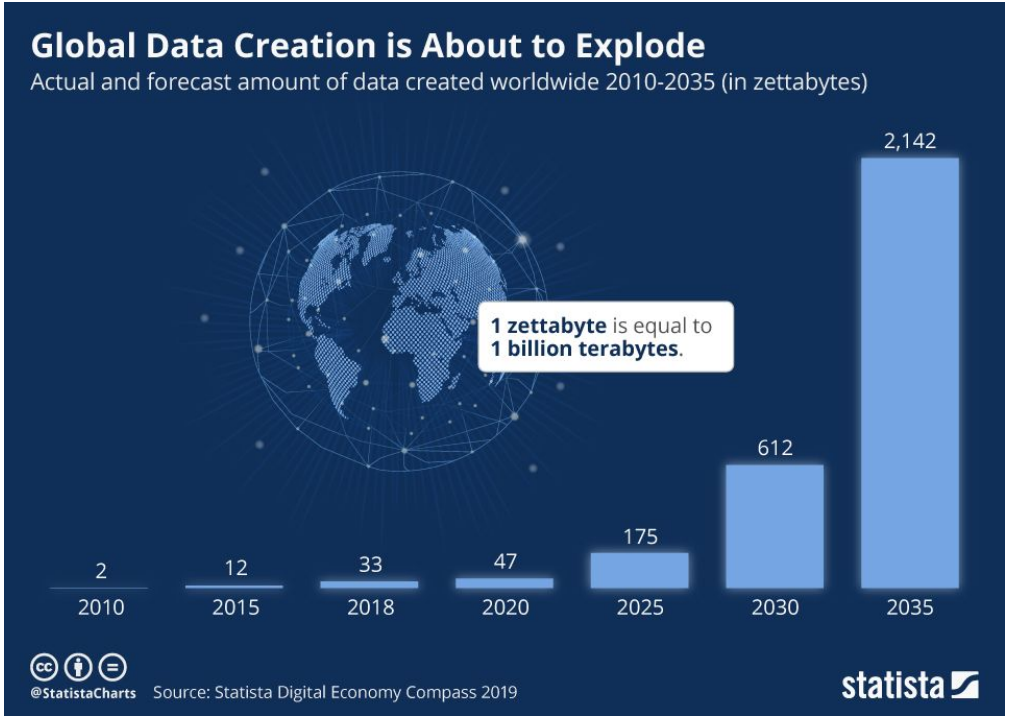
Finding Datasets

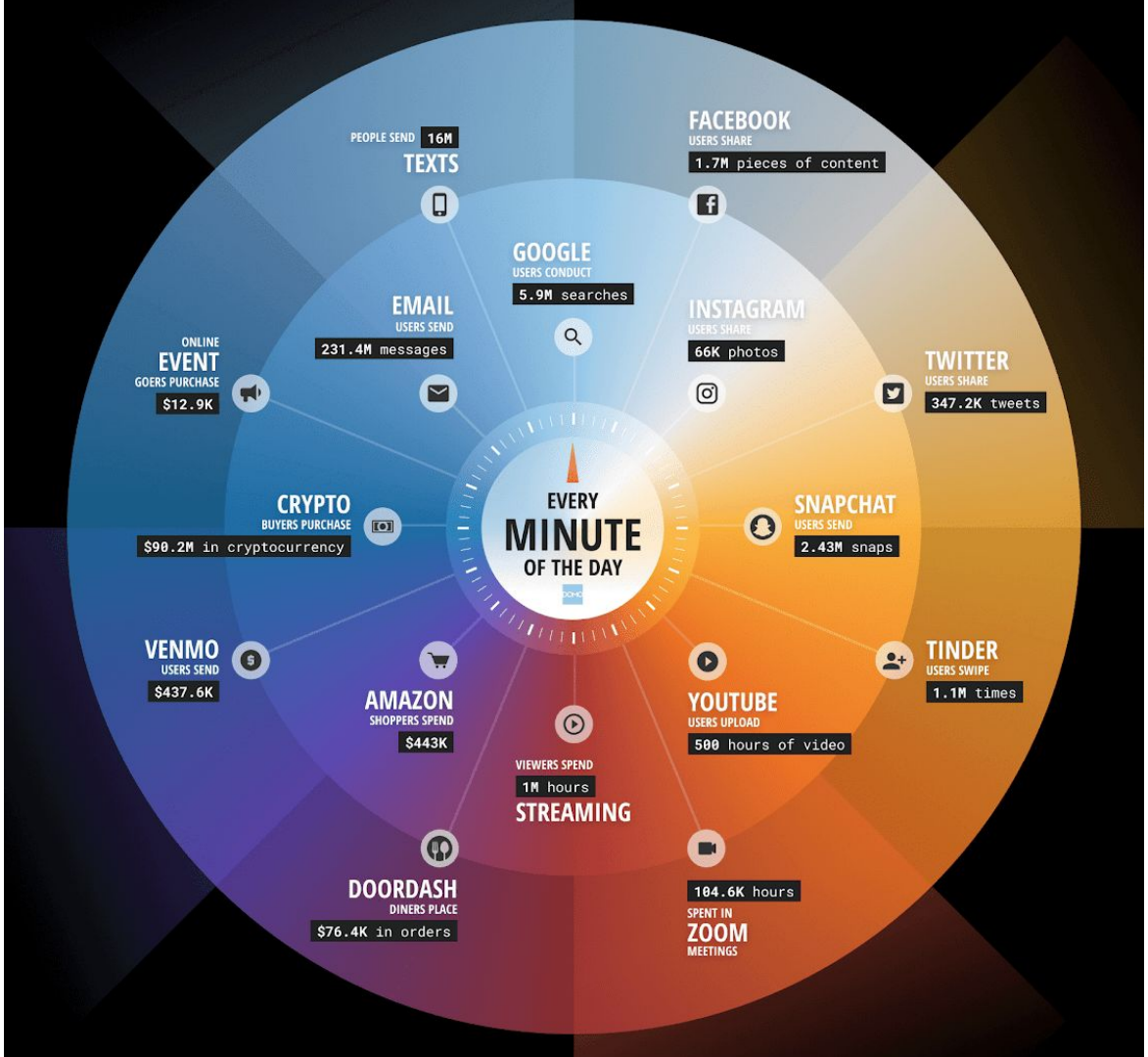
Managing Expectations…

Image by Charlesdeluvio on Unsplash

1. The dataset you want may not exist
2. You may have to “clean” your data
    a. Format
    b. Errors or other issues
3. Not all datasets are free


2021 This Is What Happens In An Internet Minute
facebook
1.4 Million Scrolling
21.1 Million Texts Sent
YouTube
500 Hours Content Uploaded
414,764 Apps Downloaded
695,000 Stories Shared
200,000 People Tweeting
2 Million Swipes
tinder
197.6 Million Emails Sent
2 Million Views
twitch
5,000 Downloads
Tik Tok
932 Smart Audio Devices Shipped
amazon echo
3 Million Images Viewed
imgur
69 Million Messages Sent
3.4 Million Snaps Created
$1.6 Million Spent Online
NETFLIX
28,000 Subscribers Watching
LinkedIn
9,132 Connections Made
60 SECONDS
Created By:
@LoriLewis
@OfficiallyChadd


Global Data Creation is About to Explode
Actual and forecast amount of data created worldwide 2010-2035 (in zettabytes)
1 zettabyte is equal to 1 billion terabytes.
2 2010
12 2015
33 2018
47 2020
175 2025
612 2030
2,142 2035
@StatistaCharts
Source: Statista Digital Economy Compass 2019
statista



In 2022:

- 5 billion internet users (DOMO)
- 397 minutes (6 hrs 37 min) average internet time per day (OBERLO)

What are the barriers to finding the perfect dataset?

Open vs. Proprietary Data

- Open data: available to all
    a. Data sets are often required by grant issuing agencies
        i. Publicly-funded research is already required to be available in an open-access repository (see OSTP memo)
        ii. New changes will eliminate the embargo period (during which it is not available so only those with paid access can read at first)
- Proprietary data: privately owned and funded; protected by copyright, patents, contracts, privacy protected
        i. May be related to computer software, business/financial information, or unpublished research (insurance data, health data, financial data, data protected by court order, recipes, designs, patterns)
        ii. The University Libraries gives you access to some proprietary data
- **If you can't get a specific data set, what else is available?** You may need to be creative!

If you can’t find or access a dataset:

- Your advisor, instructor, research team, and subject librarian can give you advice and assistance
- You can ask the researchers of a project for their data
  - (But they might not be willing to share)
    - In a recent article, researchers gave reasons for not sharing data:
      - lack of time to search for data (29.2%)
      - loss of data (27.7%)
      - privacy or legal concerns (23.1%)

Steps for Finding Data

1. Define your question or the problem you are investigating
2. Search background information
3. Identify possible sources/repositories
4. Search for data set
5. Evaluate data set

Photo by Suket Dedhia on Pexels

Find Background Information

1. General background information provides you with:
    a. Context
    b. Knowledge of **terms used by experts** in the field you are investigating
    c. Can help you **identify places to search**
2. You can see what research has already been done on your topic
    a. See the methods used to collect and analyze data
    b. See what issues investigated (so you don't repeat what has been done)
3. Sometimes you can find data sets in **articles** or in databases
        i. Web of Science allows you to search for articles with data included
        ii. ACM (calls data "artifacts")

***To search background information, use **popular sources** (internet/social media search) and **scholarly sources** (check out library guides about where to look in various subject areas)

Identify Possible Sources

- Data sets must be created: collected, organized, stored, made accessible
- This takes time, money and effort
  - Who has the resources, responsibility/mandate, interest, in collecting this data?
    - Person? (Researcher? Scientist?)
    - Research organizations/labs?
    - Government departments and agencies? (EPA, Census Bureau?)
    - International organizations? (World Bank, World Health Organization)
    - Companies? (Facebook, Amazon, Pfizer)
    - **Other???**

Searching for Data: Internet Search

1. **Internet search**
    1. Find and search the sources you identified as places to start
    2. Do a general search; try lots of different search terms (synonyms, terms you collected during background research)
        i. Try specific words such as "data set," "survey," "statistics," "poll"
        ii. Use quotation marks to keep words together: "environmental impact," "college student"
        iii. Use boolean operators to search your terms "environmental impact" AND "cryptocurrency mining"
        iv. Use a site search: "environmental impact" site:reddit.com
        v. Search for a data repository for your topic: "open data repository for ornithology"

4. Library Resources

1. **Use Library Guides by discipline to find data**:
    1. Government Information (has TONS of links to local and federal government websites)
        i. https://libguides.colorado.edu/strategies/government
    2. Business Information (has TONS of guides about how to find various kinds of data at the bottom of the page)
        i. https://libguides.colorado.edu/portal/business
    3. Geospatial Data: https://libguides.colorado.edu/geospatialdata
2. **Use Specific Library Databases to Find Data**:
    1. Statista
    2. Statistical Insight
    3. Proquest Statistical Abstracts of the World
    4. Statistical Abstract of the United States
    5. Data Planet Statistical Datasets
    6. Passport
    7. Web of Science
        i. use filter on the left side of the screen to limit to "associated data"

Searching for Data: Use a Data Search Engine

- Use a data repository database to search for data sets
  a. Google Data Search: Dataset Search is a search engine for datasets. Using a simple keyword search, users can discover datasets hosted in thousands of repositories across the Web.
  b. Re3data.org: global registry of research data repositories that covers research data repositories from different academic disciplines.
  c. Open Access Directory's List of Open Repositories
  d. Open Access Directory's List of Repositorties by Discipline
  e. Nature's List of Scientific Data Repositories

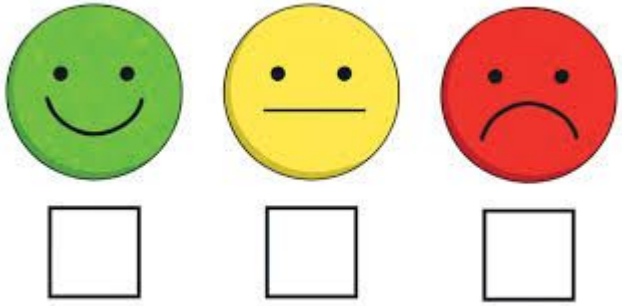
Some Databases to Check Out…

- Data.gov
- Kaggle
- Microsoft Research Open Data
- Reddit Datasets
- ICPSR (Inter-university Consortium for Political and Social Research)
- World Bank Open Data World Health Organization Data
- Amazon Web Services (AWS) Data Exchange
- Data.europa.eu
- Figshare
- Zenodo

Evaluate Your Data Set

1. Return to the questions, is the data:
    a. Usable: readable, well-documented, and available
    b. Appropriate to your task/question
    c. Functional format for your software/analysis
    d. Complete, has good metadata
    e. Minimal “cleaning” or “wrangling” needed
2. Consider who (people and organizations) that have made the data
    a. Currency
    b. Relevance
    c. Authority
    d. Accuracy
    e. Purpose
3. How was the data set created? What kinds of flaws (sample size, bias) might it have?

How to evaluate data sets?



Consider:

- Currency (is it still relevant? Newer, better data?)
- Relevance (is this the right data for your question?)
- Authority (who put together the data? Are they qualified?)
- Accuracy (How detailed is the data? How precise? Are there inconsistencies?)
- Purpose (Why was it collected?)

Also consider:

- Funding source?
- Sample size?
- Bias? (look at methodology behind collection)
- Is it available? Accessible to you?

Need Help Using Data?

- CRDDS Learning Materials and Classes/Consultations
- Coursera
- LinkedIn Learning
- YouTube
- ICPSR training guide on Youtube
- SAGE Research Methods Online

Traits of a “good” data set

1. Usable: readable, well-documented, and available
2. Appropriate to your task/question
3. Functional format for your software/analysis
4. Complete, has good metadata
5. Minimal “cleaning” or “wrangling” needed

- **Provides information about your topic**
- **Helps prove/disprove theories**

Some Data Formats

From Axiom Data Science:

- Containers: TAR, GZIP, ZIP
- Databases: CSV, XML
- Tabular data: CSV
- Geospatial vector data: SHP, GeoJSON, KML, DBF, NetCDF
- Geospatial raster data: GeoTIFF/TIFF, NetCDF, HDF-EOS
- Moving images: MOV, MPEG, AVI, MXF
- Sounds: WAVE, AIFF, MP3, MXF
- Statistics: ASCII, DTA, POR, SAS, SAV
- Still images: TIFF, JPEG 2000, PDF, PNG, GIF, BMP
- Text: XML, PDF/A, HTML, ASCII, UTF-8
- Web archive: WARC

1. Define your question

1. What is your topic?
2. How can you think about this topic empirically?
3. What kind of data do you want to find?
    a. You can always search around to see what data is available…

Keep in mind:

- Your research question or focus may change as you investigate, be flexible
- Not all data is available, you might have to come up with alternative kinds of data to complete your research

What is a data set?

- **It is information that is:**
  - **collected, assembled, and organized (by someone)**
  - **for analysis of an issue, phenomenon, or subject.**
  - Data can be: textual, numeric, images, sound, video, code, geospatial
  - Some formats (CSV, XML, TIFF, PDF, etc.)