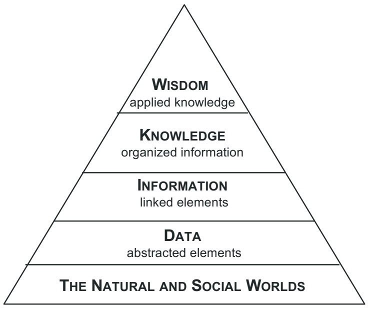
Data Science Project: Design Brief

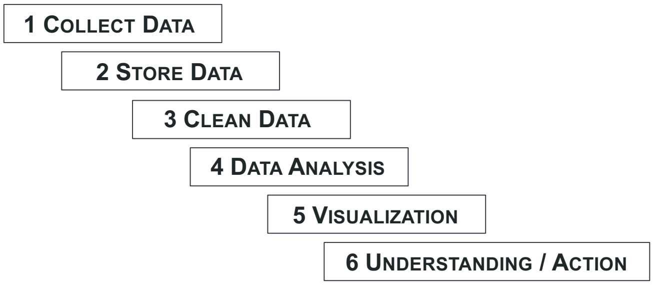
Consider …

> **Data science** (noun). Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data, *with a focus on human well-being*. [INFO-201 Syllabus]

- As we have seen in this class, data science is a technical process, where scientific knowledge and technical skill are used to understand the past, to make sense of the present, or to shape the future.
- Data science projects can lead to benefits or to harms. Often, a data science project will lead to both benefits and harms, along with unanticipated negative consequences.
- Thus, to the standard definition of data science, we added the clause, *with a focus on human well-being*. We did so to make the point that we, as designers, have a responsibility to use data science for good.
- *Human well-being* is a very broad value, which might be related to quality of life, happiness (fun!), human dignity and justice, and human flourishing, that is, a good life.



The Knowledge Pyramid. From sensing and collecting data in the world to wisdom and taking responsible action.



The Data Science Waterfall. Each of these stages takes skill. Responsible and thoughtful decisions at the early stages can compound and create more benefit than harm overall.

A. Introduction to Project

1. ***Project aim.*** Your project team will investigate a topic of concern of your own choosing and find a relevant dataset (steps: 1 collect and 2 store data). You will clean and organize the data (step 3) and you will conduct a process of exploratory data analysis (step 4) and visualization (step 5). Finally, you will seek to answer of set of 3–5 research questions and draw out implications for making recommendations to technologists, designers, or policymakers (step 6).
2. ***Possible topics of concern (aka problem domain/design situation).*** The topic of your research can be anything that you care about. It should be narrow, well-defined, and related in some way to human well-being. It might concern ocean acidity or sea-level rise or some other aspect of the **climate crisis**. It might concern **sleep** and saccadic rhythms. It might concern **mental health** or the **physiology** of high-performance bicycle racing. It might concern music or movie recommendations or more broadly **culture and media production**. It might concern workforce prediction algorithms, precision agricultural systems, supply chains or other **global systems**. It might concern fair treatment before the law, police shootings, food insecurity, or matters **of social justice**.
3. ***Your learning goal – Be curious; Be creative; Seek to make a difference***. Broadly, your goal is to develop your skills for coding (being a **Coder**), your skills for team work and responsible innovation (being a **Responsible Innovator**), and for critically thinking about data and code (being a **Critical Thinker**).

B. General Project Requirements

1. ***Team size***. Teams will be made of four students (perhaps three students) from the same lab section. Your Teaching Assistant will organize the formation of teams.
2. ***Audience.*** Assume that prospective employers, open source developers, and thousands of people will visit your project website. Thus, it is important that your work demonstrates integrity and responsibility, along with careful coding, writing, and presentation.
3. ***Topic of concern.*** You should investigate and report on your topic of concern (aka problem domain or design situation). Drawing on – and citing – research, you should address these kinds of questions:
    - *Framing the topic of concern*. What is the topic of concern, problem domain or design situation? What are the key elements of the topic? What are the key scientific, cultural, social, governmental, of economic issues or questions?
    - *Human values*. What human values are within or connected to your topic of concern? Where do the values seem to originate? What value tensions among different values are present?
    - *Stakeholders*. Who are the direct stakeholders of your topic of interest? What skills are assumed? What motivations and values do they hold? And, who are the indirect stakeholders.
    - *Benefits and harms*. If interventions are taken with data and technology, what are the potential benefits and harms? Which stakeholders are likely to be benefit and be harmed? What unanticipated consequences might occur?
4. ***Research questions.*** Given your topic of concern, what are 3–5 research questions will you address in your project? What motivates these questions? Why are these questions important? Generally, how will you address them?
5. ***Dataset size and complexity***. You should work with a dataset of reasonable complexity, and which gives you a way to investigate your topic of concern. While there are no strict rules, as a general guideline, the minimal requirement is:
    - *One data file of more than 200 observations (rows/records) and 6 variables (attributes/features).*

    Some projects will employ:

    - *3–5 different data files, comprising more than 50-100K observations and more that 30 variables.*

    Most projects, however, will lie somewhere between these two levels of complexity.
6. ***Data provenance.*** Related to your topic of concern (#3), you should conduct a critical analysis of the origins of your project dataset. Drawing on D'lgnazio and Klein (2020), you should address these and similar questions:
    - Who or what is represented in the data?
    - What is an observation? What variables are included (and excluded)?
    - Who collected the data? When? For what purpose? How was the data collection effort funded? Who is likely to benefit from the data or make money?
    - How was the data validated and held secure? Is it credible and trustworthy?
    - How did you obtain the data? Do you credit the source of the data?
7. ***Dynamic reports and interactive visualization.*** Using R, you will design, implement, and test two specific technological deliverables:
    - *P02: Exploratory Data Analysis:* dynamic report, implemented with R Markdown;
    - *P03: Final Project*: interactive data visualization, implemented in Shiny.
8. ***Scientific and design-based inquiry.*** You should strive to demonstrate a nuanced understanding of the important features of the dataset, demonstrating the knowledge pyramid in action, from data to wise action. You'll uncover high-level insights – important descriptive information, major trends, notable outliers, and so on. Importantly, you will discuss the implications of your work, showing how it can be applied to make decisions, make policy, or to better understand the problem situation.

C. Where to find datasets?

There are many ways to find interesting data sets. Here are six suggestions, in alphabetic order:

1. Earth Data from NASA
Large number of datasets about the Earth
https://earthdata.nasa.gov/

2. FBI Crime Data Explorer
https://crime-data-explorer.fr.cloud.gov/pages/home

3. Google dataset search
For example, type "sleep," "ocean acidification," "music," and so on.
https://datasetsearch.research.google.com/

5. Kaggle
A community hub with many, many datasets organized into categories
https://www.kaggle.com/datasets

5. NYTimes Developers
APIs to access data from the NYTimes
https://developer.nytimes.com/

6. World Health Organization: Global Health Observatory data repository
https://apps.who.int/gho/data/node.home

You will find that some of the datasets at these sites of are complex. If you experience difficulties, please ask your Teaching Assistant or post on Teams!

D. How to get started?

A. Project team organization

1. ***Group formation.*** Work with your teaching assistant to form a project group. Groups will be formed during labs for week 4 and 5.
2. ***Contact information***. Share contact information among your group members.
3. ***Schedule a weekly meeting***. Find 60-90 minutes when you can meet together to work on the project. This weekly meeting time is essential for brainstorming, writing, coding, and coordinating work on GitHub.
4. ***Google docs***. Set-up a google document and use that to keep notes project plans and links to key resources and documents.
5. ***Canvas group***. Ensure your group is entered in Canvas. Please check with your TA.

B. Consider topic of concern: Brainstorming and research

6. ***Project brief.*** Read the project brief and create a list of questions. Ask your TA and/or post questions on Teams.
7. ***Topic of concern.*** What you are interested in? What do you care about? Brainstorm some topics. Try to identify five or more possible topics. Then, choose the one you like best.
8. ***Consider the general project requirements.*** Start taking notes and begin research on your topic of interest. *Suggestion*: Go to the library and ask for research assistance.
9. ***Find possible data sets.*** Begin searching for datasets. *Suggestion*: Ask TA for help.
10. ***Goto 6.*** These steps are iterative and integrative.

C. P01: Project proposal (Week 4- 5): See Canvas for instructions

11. ***Final Project Repository.*** On Canvas, follow the instructions for setting up your GitHub repository. *Note: Because this is a group project, the set-up is different than individual assignments.*
12. ***Write the project proposal.*** Follow the guidelines for writing the project proposal, on Canvas.

D. P02: Exploratory data analysis (Week 6 – 7)
Developing and publishing an R Markdown report

E. P03: Final project deliverable (Week 8 – 10)
Developing aa interactive visualization in Shiny