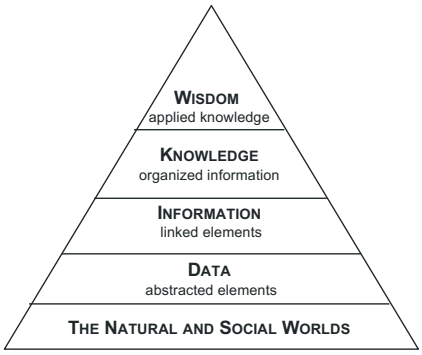
Data Science Project: Design Brief

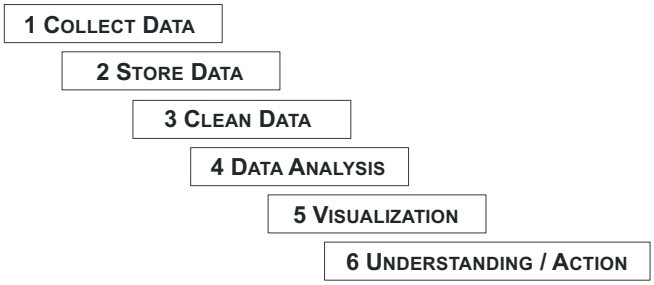
Consider …

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]



The Knowledge Pyramid

- 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 both at the same time. Also, data science projects often lead to unanticipated 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 Data Science Waterfall. Each of these stages takes skill. Responsible and thoughtful decisions at the early stages can compound and create more benefits and few harms overall.

A. The Project in Nutshell

Working in a group of four, you will investigate a problem or situation through data. You will write a professional report about your project and demonstrate knowledge for data science and R programming.

B. Introduction to Project

1. ***Project aim.*** Your project team will investigate a problem domain 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. ***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.
3. ***Problem domain.*** 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 **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**.
4. ***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**).

C. Project Requirements

1. ***Project structure: Dynamic reports and Interactive Visualization.*** The project is structured as three main components (see Canvas for details):

   P01: Project proposal

   P02: Exploratory Data Analysis: dynamic report, implemented with R Markdown

   P03: Final Project: interactive data visualization, implemented in Shiny.

2. ***Team size.*** Teams will be made of four students from your lab section, randomly chosen.

   - If have a conflict with a team member, please inform your Teaching Assistant.
   - If you would like to be on another team, propose a swap of team members from team A to B, and inform your Teaching Assistant.

3. ***Dataset size and complexity***. You should work with a dataset of reasonable complexity, and which gives you a way to investigate your problem domain. At a minimum, your data should meet these requirements:

   - ***Number of data files*****:** 3 or more
   - ***Total number of variables (attributes/features)*****:** 10 or more
   - ***Total number of observations***: 500 or more

   Some projects will employ much larger data sets (more than 30 variables and 50-100K observations). If you have questions about the size and complexity of your data set, please ask your Teaching Assistant.

4. ***Audience.*** Assume that prospective employers, open source developers, and knowledgeable people will visit your project website. Thus, it is important that your work demonstrates integrity and responsibility, along with careful coding, writing, and presentation.
5. ***Problem domain.*** You should investigate and report on your problem domain. Drawing on – and citing – research, you should address these topics:

   - *Framing the problem domain*. You should give background information on your problem domain, addressing this question:
     - What is the problem domain or design situation? What are the scientific, cultural, social, governmental, or economic issues or questions?

- *Human values*. What human values are within or connected to your problem domain? Where do the values 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?

Use the Envisioning Cards to identify issues and questions. Address those issues and questions when framing and describing your project.

6. ***Research questions.*** Given your problem domain, 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?
7. ***Data provenance.*** Related to your problem domain (#3), you should conduct a critical analysis of the origins of your project dataset. Drawing on D'Ignazio 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?

D. Where to find datasets?

There are many ways to find interesting data sets. Here are some 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/

4. 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. ProPublica Data Store
https://www.propublica.org/datastore/

7. 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.*** Groups will be announced at the beginning of week 4.
2. ***Contact information***. Meet your group members at labs. Share contact information.
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.

B. Consider your problem domain: Brainstorming and research

1. ***Project brief.*** Read the project brief and create a list of questions. Ask your TA and/or post questions on Teams.
2. ***Problem domain.*** 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.
3. ***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.
4. ***Find possible data sets.*** Begin searching for datasets. *Suggestion*: Ask TA for help.
5. ***Goto 1.*** These steps are iterative and integrative.

C. Project milestones (See Canvas for instructions)

P01: Project proposal (Week 4- 5):

1. ***Set-up 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.*
2. ***Write the project proposal.*** Follow the guidelines for writing the project proposal, on Canvas.

P02: Exploratory data analysis (Week 6 – 7)

Developing and publishing an R Markdown report

P03: Final project deliverable (Week 8 – 10)

Developing an interactive visualization in Shiny