- A **dataset** of moderate size and complexity.
    - It should be large enough to have interesting complexity, but not so big as to be unwieldy. As a rough guideline, your dataset should be longer than you could print on a single page in standard spreadsheet format, but smaller than 20 MB.
    - You may use existing datasets, combine data from APIs, or create entirely new data through instruments or surveys.
    - How you collect your data will affect your degree-of-difficulty rating. For instance, if you simply download a ready-made dataset from Kaggle, that would constitute a lower degree of difficulty than web-scraping raw data yourself. Thatâ€™s OK, as long as you are aware and perhaps compensate for it in other sections (like in data analysis).
    - The dataset you turn in does *not* have to be the dataset that you initially collected. For example, you might download 50 MB of raw logs, but use filtering and aggregation to reduce the dataset to 100 kB for your actual analysis. We want you to submit your *analysis-ready data*, but you should describe your full data-collection protocol and any preprocessing done in the data description section of your final report (see below). All source code use for data collection and preprocessing should also be linked to in the source code section of your final report.Â Â
    - If your final, curated dataset is larger than 10MB, share a copy in Cornell Box and include a link to it in your final report.
- A **final report**, as a Jupyter notebook with executed cells, containing the following sections:Â

- **Introduction**. What is the context of the work? What research question are you trying to answer? What are your main findings?Â

- **Data description.** This should be inspired by the format presented in Gebru et al, 2018. Answer the following questions:

    - What are the observations (rows) and the attributes (columns)?

    - Why was this dataset created?

    - Who funded the creation of the dataset?

    - What processes might have influenced what data was observed and recorded and what was not?

    - What preprocessing was done, and how did the data come to be in the form that you are using?

    - If people are involved, were they aware of the data collection and if so, what purpose did they expect the data to be used for?

    - Where can your raw source data be found, if applicable? Provide a link to the raw data (hosted in a Cornell Google Drive or Cornell Box).Â

- **Data analysis.**

    - Use summary functions like mean and standard deviation along with visual displays like scatter plots and histograms to describe data.

    - Provide at least one model showing patterns or relationships between variables that addresses your research question. This could be a regression or clustering, or something else that measures some property of the dataset.

- **Evaluation of significance.** Use hypothesis tests, simulation, randomization, or any other techniques we have learned to compare the patterns you observe in the dataset to simple randomness.Â

- **Conclusion.** What did you find over the course of your data analysis, and how confident are you in these conclusions? Interpret these results in the wider context of the real-life application from where your data hails.

- **Source code**. Provide a link to your Github repository (or other file hosting site) that has all of your project code (if applicable). For example, you might include web scraping code or data filtering and aggregation code.

- **Acknowledgments**. Recognize any people or online resources that you found helpful. These can be tutorials, software packages, Stack Overflow questions, peers, and data sources. Showing gratitude is a great way to feel happier! But it also has the nice side-effect of reassuring us that you're not passing off someone else's work as your own. Crossover with other courses is permitted and encouraged, but it must be clearly stated, and it must be obvious what parts were and were not done for 2950. Copying without attribution robs you of the chance to learn, and wastes our time investigating.