# CS513: Theory & Practice of Data Cleaning
# **Final Project Instructions (Summer 2021)**

Bertram Ludäscher
ludaesch@illinois.edu
University of Illinois, Urbana-Champaign

The goal of the group project is to conduct an end-to-end data cleaning project, using the various tools and techniques that we have covered throughout the course. In addition to the main tools that we used in class (i.e., RegEx, OpenRefine, Datalog, SQL, and Python), you are welcome to use other tools as well. For example, you may want to use any of the research prototypes mentioned (e.g., YesWorkflow or the OpenRefine companion tools such as or2yw), or commercial tools (e.g., Trifacta Wrangler, Tableau, etc.) In your report you will then document how you used these tools.

**Team Formation.** You are strongly encouraged to form project teams of **2 or 3 students**. If you would like to form a team of *4 students* or do an *individual project*, you need to submit a request (with explanation) to obtain approval by the instructor. This semester we will be using the Coursera Teams feature to simplify handling of project groups. If you still have questions regarding team formation, please contact the TAs (and instructor).

**Project Phases.** The project is organized into two phases, each with its own set of deliverables. **Phase-I** should typically be completed **within one or two weeks from the project start**, while **Phase-II** extends to the end of the semester (check campuswire for the specific target dates).

## Project Phase-I

During this phase your team needs to . . .

1. **Identify a dataset** $D$ of interest. This can be one of the provided datasets (NYPL menus; Airbnb; PPP loan applications; or US Farmers Markets), or a new dataset that your team would like to work with.

2. **Develop a target use case** $U_1$ for $D$ such that data cleaning is *necessary* and *sufficient* to support the data analysis use case. Thus, after performing data cleaning, your cleaned data $D'$ is *fit-for-purpose* (i.e., for $U_1$). In addition to your main use case $U_1$ you should also briefly describe two minor use cases: $U_0$ should be a use case that requires "zero data cleaning", i.e., $D$ is "good enough as it is". In contrast, $U_2$ is a use case for which the given dataset $D$ is "never (good) enough", i.e., no amount of data cleaning or wrangling will make $D$ suitable for $U_2$ (even though at first sight one might think so). The purpose of the corner cases $U_0$ (data cleaning is *not necessary*) and $U_2$ (data cleaning is *not sufficient*) is to reinforce the concept that **data cleaning should be done with a purpose in mind**, i.e., a use case such as your target use case $U_1$, where data cleaning really makes a difference.

3. **Describe the dataset** $D$. For example, you can provide a *conceptual model* (ER diagram) that depicts the entity types and relationship types, or an *ontology* that illustrates the main

classes and their relationships. Or you can provide a *database schema* that illustrates and explains the structure and contents of the dataset. You should also add a short *narrative*, i.e., one or more paragraphs in English to describe the origin of the data and any relevant metadata (e.g., a *temporal* or *spatial extent*). A dataset about farmers markets, e.g., can be described with a relational schema (e.g., CREATE TABLE statements); the narrative would then explain what the different columns (attributes) mean. Other metadata may describe, e.g., the spatial extent of the data (only Illinois markets? All of the Midwest? Or the US?), and the temporal extent (for which period is the data correct?), etc.

4. **List obvious data quality problems** (i.e., which are easy to spot). In order for your dataset $D$ and target use case $U_1$ to match, data cleaning must be necessary and sufficient to implement $U_1$. You need to support this claim by documenting data quality problems that your inspection of $D$ has revealed and that need to be addressed before $U_1$ can be tackeled.

5. **Devise an initial plan** that outlines how you intend to clean the dataset in Phase-II. A typical plan for the overall project will include the following steps: $S_1$: description of dataset $D$ and matching use case $U_1$; $S_2$: profiling of $D$ to identify the quality problems $P$ that need to be addressed to support $U_1$; $S_3$: performing the data cleaning process using one or more tools to address the problems $P$ (here you should describe which tools you are planning to use, e.g., OpenRefine; Python; etc.) $S_4$: checking that your new dataset $D'$ is an improved version of $D$, e.g., by documenting that certain problems $P$ are now absent and that $U_1$ is now supported; $S_5$: documenting the types and amount of changes that have been executed on $D$ to obtain $D'$.

   You should also include a tentative assignments of tasks to team members (who does what).

## Additional Information

Regarding (2): How do you specify data analysis **use cases**? Generally speaking, you can simply explain the use case in a short paragraph. You might also want to be more specific and phrase a use case as one or more **questions**: *What is it that we want to know from or about the data?*

In particular, a use case may be a set of database **queries** $Q_1, \ldots, Q_n$ against the dataset $D$ (e.g., how many farmers markets offer bakery goods in addition to vegetables and fruits?) On the other hand, use cases may also be more general, e.g., you could state that you'd like to develop a web application that serves a particular purpose.

The advantage of specifying a use case $U$ as one or more queries $Q_U$ is that you can be very precise about when data cleaning is necessary and sufficient for $U$: if running $Q_U$ on the original ("dirty") data $D$ would result in an answer $A = Q_U(D)$ that is incorrect and/or misleading, then data cleaning is **necessary**. Conversely, data cleaning is **sufficient** if the answer $A = Q_($D'$)$ on the cleaned dataset $D'$ is correct (and not misleading).

In (4) above, how do you document data quality problems? One simple way is to include snippets of "dirty data" in your Phase-I report (you can also use screenshots for illustration) and then explain what the problem is.copy-pasting

How do you describe your **plan** in (5)? A short list of your planned steps $S_1, \ldots, S_5$ will do during Phase-I. In Phase-II, you should also include a **workflow diagram** for the actual data cleaning steps that you performed (e.g., with YesWorkflow or any other diagramming tool). Of course your Phase-I *plan* and your *actual* Phase-II *workflow* might be different.

## Project Phase-II

During this phase you will execute the plans you've come up with in Phase-I, possibly adjusting course based on what you find when actually working with the data.

## What to Submit

**Phase-I:**

- **A single PDF file with your Phase-I report**, having the 5 elements described above.

**Phase-II:**

- **A single PDF file with your Phase-II report**. This report should include:
  - An updated and if necessary expanded version of the items from the Phase-I report: (1) dataset $D$, (2) description of use case $U_1$ (and brief descriptions of $U_0, U_2$), (3) dataset description, (4) list of quality problems (expanded from Phase-I, as you now have worked with the data); (5) workflow diagram $W$ describing actual steps performed.
  - A narrative that ties all steps together and explains the motivation (use case $U_1$), the rationale for the design of the overall workflow $W$ and the tools used.
  - Documentation that data quality was improved, e.g., through running "before queries" $Q_U(D)$ and "after queries" $Q_U(D')$ on $D$ (original) and $D'$ (cleaned), respectively.
  - A summary of the data changes $\Delta D$ resulting from the overall workflow $W$: $D \rightsquigarrow D'$.
  - A summary of findings, problems encountered, and lessons learned, including possible next steps (e.g., implementation of $U_1$).
- **Supplementary Materials**. In addition to the project report, you need to provide the following supplementary materials (as a single ZIP file):
  1. **Operation History**: A copy of the OpenRefine *operation history* (copy-paste it into a json file named OpenRefineHistory.json). If you are using an alternative tool instead of OpenRefine, please provide an analogous history file (OtherToolHistory.json) and other provenance information (as available for that tool).
  2. **Queries**: A copy of the queries written in SQL or Datalog to profile the dataset and check the integrity constraints (copy-paste them into a text file named Queries.txt).
  3. **Workflow Model**: For the overall workflow model $W$ (using YesWorkflow or other diagramming tools), provide the file that has the *annotations* (e.g., OverallWorkflow.txt), and the generated Graphviz or DOT file (e.g., OverallWorkflow.gv). For the OpenRefine workflow, provide similar files.
  4. **Raw and Cleaned Datasets**: Please **do not** provide the datasets in the ZIP file. Rather, upload the raw and cleaned datasets in a Box folder and **share the link** in a plain text file (DataLinks.txt).