Final Project Information

Learning Objectives

- Practice data integration and gain familiarity with the concept of identity.
- Gain experience canonicalizing data and implementing strategies for ensuring equivalence.
- Construct convincing arguments about the importance of data curation.

Scenario

Recently, you were hired as a data scientist for a government agency. The agency tracks complaints made by customers against financial institutions, like banks and lending companies. The agency has switched systems for managing its complaint data, and this requires a transfer of complaint data from the old system into the new system.

You've been asked to assess the quality of the data transfer, especially to ensure equivalence of the data in the old and new systems. Given what you learned in your data curation course, you recognize this task requires canonicalizing two files.

In addition, the agency recently appointed a new director of the data science department, and the new director has de-prioritized data curation work. You have been asked to write a memo making a case for data curation.

You are given the data files

- File A (Old system)

```
Consumer_Complaints_FileA.xml
```

- File B (New system)