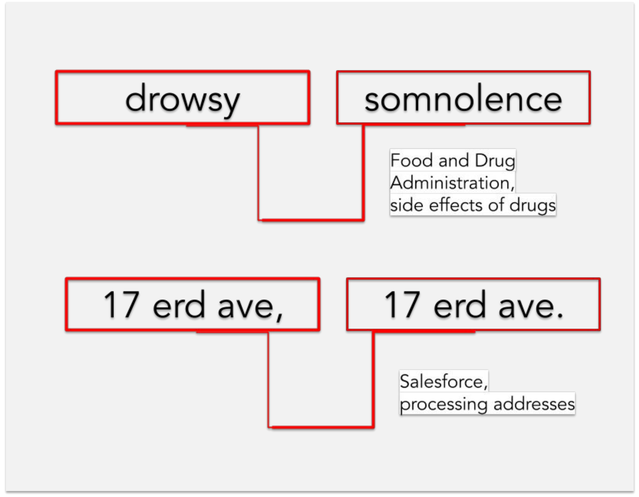
Wrangling data without the right people or tools

Allison Huang, History Department, ah25@princeton.edu

SUMMER RESEARCH COLLOQUIUM
PRINCETON 2019

"Data Wrangling" is necessary

Data is messy. Part of the problem is that computers don't process data the way humans do. For example, a computer sees "17 erd ave" and "17 erd avenue" differently. Data scientists spend 50-80% of their time "data wrangling," that is, collecting and preparing unruly data.[1]



Two words may mean the same thing, but computers cannot tell the difference.

Who tends data in Nonprofits?

I explored how the data management concerns of a nonprofit like Isles (my partner organization) overlap and depart from the concerns of well-staffed organizations working with Big Data.

Who I Consulted

- Scientific articles on best data management practices

I found three overarching data management principles, two of which are discussed below.

- A comprehensive survey of 467 nonprofit professionals in 2016 on the use of data in nonprofits (nonprofithub.org)
- My own experience as a "data consultant" for Isles

Lacking the people for the job

Scholars recommended cleaning data with computer programs (rather than manual corrections) and storing raw data with metadata (so anyone can understand the data).[2] Nonprofits lack trained personnel to operate these computer programs and metadata formats. According to the aforementioned survey by nonprofithub, the top two factors preventing nonprofits from using their data are: "Not enough time or personnel to focus on data" or "Personnel doesn't have experience using data."[3]

Who does data wrangling? In corporations and research centers, data scientists do. Nonprofits lack these personnel.

Employees use data differently

There's a disconnect in how people use data. Nonprofithub found that "87 percent of nonprofit professionals indicated that data was moderately to extremely important to operations and decision making...organizations as a whole placed much less importance (57 percent) on data in those circumstances, showing a disconnect between the way staff and organizations are using data as a whole."[4]

Imagine these three different audiences that use data. Each views the purposes of the data differently.

A nonprofit specific database?

With these findings, one might create an nonprofit-specific relational database (data management system). It would anticipate the lack of technically trained personnel and have special features that allow the diverse groups of people using the same data to communicate.

Acknowledgements

A thank you to the Program for Community Engaged Scholarship for sponsoring my work, especially Maria Lockwood. A thank you to Isles, inc for having me as an intern, especially Peter Rose for his guidance this summer.

References

1 Lohr, S. (2014, August 18). For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights. New York Times.
2 Hart, E. M., Barmby, P., LeBauer, D., Michonneau, F., Mount, S., Mulrooney, P., . . . Hollister, J. W. (2016, October 20). Ten Simple Rules for Digital Data Storage.
3-4 The State of Data in the Nonprofit Sector. (2016, March 16).