Demystifying Digital Scholarship
An Introduction to Data Wrangling

Sherman Centre for Digital Scholarship
12-February, 2016
# Outline and Objectives

## Topics
- Let’s talk about data
- What is data ‘wrangling’?
- Why you should ‘wrangle’?
- How to ‘wrangle’?

## Objectives
- Present and discuss strategies for structuring data
Data & Data Wrangling
Every Project Has Data!

➢ Chances are that your project contains at least one (and likely more) data types:
  → Text, images, tags, geographical coordinates, categorical items, records, metadata, multimedia, etc.

➢ Understanding your data and your intended actions is a critical part of developing a DS/DH project
  → It guides your data activities
  → It helps inform you of the ways in which your data can be used
    -- by you, your collaborators and others in your research community
Defining data wrangling

**wrangle** *v.tr*

- To manage or herd
- To manage or control
- To grasp and maneuver (something); wrestle
- To win or obtain by argument
Defining data wrangling

**wran•gle v.tr**
- To manage or herd
- To manage or control
- To grasp and maneuver (something); wrestle
- To win or obtain by argument

**wran•gle v.intr**
- To attempt to deal with or understand something; contend or struggle
Defining data wrangling

**wran·gle**  *v.tr*

➢ To manage or herd
➢ To manage or control
➢ To grasp and maneuver (something); wrestle
➢ To win or obtain by argument

**wran·gle**  *v.intr*

➢ To attempt to deal with or understand something; contend or struggle

**wran·gle**  *n.*

➢ An angry, noisy argument or dispute.
Defining data wrangling

In the context of data, DATA WRANGLING:

➢ Is the process of cleaning and conditioning data into a usable format
➢ May be a manual, semi-automated or automated process
➢ Produces data that connects to tools, collaborators and communities
Why wrangle?

➢ Even if you understand and work well with your data, it doesn’t mean that a computer will be able to use it to the same extent.

➢ Computers (like people) are only as flexible/adaptable as far as they have been trained or instructed.

➢ Therefore, it often takes work to structure your information/data in a way that can be used in a computing environment.
Why wrangle?

Because this happens → , and this can be a problem

From School of Data’s Data Cleaning Module:
“the Invisible Man is in your spreadsheet, messing with your data”
Activity: Why wrangle?

All workshop materials are available in the following Google Drive folder: https://goo.gl/u53tkz

Check out Sample Copy - Crowdsourcing Exercise
(‘pop out’ into a new window; Make a copy and play with it, if you want)
➢ What’s not quite right?
➢ What could possibly go wrong?
➢ What needs to be cleaned?
➢ How best to clean this data?
What if it had 30000 rows? What if it amassed 30000 rows / day?
Understanding your needs
What is your ‘data’?

What kind of ‘wrangling’ is required?
Data wrangling: BIG impact, even for small data

Whether your data is small or “BIG”, conditioning it is critical.
Big data is defined by the 3 “V”s
...or the 4 “V”s

- Volume
  - Click stream
  - Active/passive sensor
  - Log
  - Event
  - Printed corpus
  - Speech
  - Social media
  - Traditional
- Variety
  - Unstructured
  - Semi-structured
  - Structured
- Velocity
  - Speed of generation
  - Rate of analysis
- Veracity
  - Untrusted
  - Uncleansed
...or 5?

**Volume**
- Terabytes
- Records/Arch
- Transactions
- Tables, Files

**Velocity**
- Batch
- Real/near-time
- Processes
- Streams

**Variety**
- Structured
- Unstructured
- Multi-factor
- Probabilistic

**Value**
- Statistical
- Events
- Correlations
- Hypothetical

**Veracity**
- Trustworthiness
- Authenticity
- Origin, Reputation
- Availability
- Accountability

5 Vs of Big Data
6? Really?

Big Spectrum Data

Volume
- Multi-domain data
- User/device data
- Geolocation data

Value
- Spectrum modeling
- Spectrum prediction
- Spectrum management

Velocity
- Data in motion
- Stream computing
- Batch algorithms
- Real-time algorithms

Variety
- Crowd sensing
- Geolocation database
- Heterogeneous sensors
- Different data types

Viability
- Variable selection
- Variable relevance
- Variable relationship

Veracity
- Spectrum data quality
- Data uncertainty
- Data security
Factors Influencing Your Wrangling Needs

Your data might...

➢ be voluminous and dynamic
➢ come from very diverse sources in a variety of formats and structures
➢ vary considerably in quality and consistency (i.e. may require heavy scrutiny)

Or it might not be...
All of these factors influence your data wrangling needs and potential solutions!
What else matters?

Other factors that will affect the nature of your wrangling activities:

➢ What tools are you using?
➢ Are automated approaches available?
➢ What formats are required for later analyses?
➢ How tolerant are you to errors?
Getting & Assessing Your Data

➢ Downloading Files
➢ API access
➢ Web-scraping
Bulk Downloads

Common data formats

➢ ASCII text
➢ Delimited ASCII (.csv, .tsv)
➢ PDF
➢ JSON (https://goo.gl/632RGb)
➢ Markup languages (TEI, HTML, XML)
➢ Multimedia (.wav, .ogg, .mp4, .mkv)
➢ Other, program-specific and ad-hoc file formats (fixed width, SAS, xls)
API Access

API = Application program interface
➢ set of protocols/tools for building software applications
➢ governs how software should interact with each other and user interfaces

Reddit API: https://www.reddit.com/dev/api
Web Scraping

Scraping vs. Parsing:

➢ Parsing: data being extracted is intended as input to another program
➢ Scraping: data being extracted is intended for display to an end user

e.g. > wget https://www.reddit.com/r/sandersforpresident

➢ Scrapes web page html to file →
Activity: Assessing Data

Wikipedia page exported to XML:
- [https://goo.gl/pCDg30](https://goo.gl/pCDg30)

Goal: Explore which Wikipedia articles are linked to Kanye?
- Can you extract the necessary information?
- Are there patterns/landmarks in the data that you can take advantage of?
- What is the consistency of the information?
Cleaning & Transforming Your Data
Data Cleaning: Goal & Activities

Goal: Create data sets that are **consistent** and **interoperable** with other data of interest

Data cleaning (scrubbing) may include:
- Detecting and remediating corrupt/inaccurate records
- Removing typographical errors
- Validating against a known list of possibilities (e.g. verify a string as a postal code)
- Eliminating duplicate entries
- Harmonizing and standardizing (e.g. represent ‘St’, ‘St.’, ‘Street’ as ‘street’)

Cleaning may be carried out
- Manually (interactive)
- Semi-automatically (guided)
- Automatically (scripted)
Approaches For Cleaning & Transformation

- Spreadsheets are often a good place to start
  - [http://schoolofdata.org/courses/#IntroDataCleaning](http://schoolofdata.org/courses/#IntroDataCleaning)
- Text Editors - Find and replace
- Customized applications
  - Trifacta: [https://www.trifacta.com/](https://www.trifacta.com/)
  - Open (Google) Refine: [http://openrefine.org/](http://openrefine.org/)
- Fully-automated scripts
ACCOUNTANT, AUDITOR AND ADJUSTER.

CITY OF HAMILTON

6 Jas Ramage, laborer
96 Sand Garrity
98 Geo lake, granite cutter
too Alex Martin, laborer
too Godfrey Auld, laborer
to6 Jas Flynn, laborer
110 Jas Furlong, laborer
112 11 A Broadbent, brakinn
to Henry N'ernon, directory
publish
16 Matthew Hunter, carprar
118 Herbert Dixon

Yoik 11 intersect(

36 A H McKeown, tnerchant
40 Thos II Jermyn. bookkpr
42 Jas McLaughrin, machist

Main st Intersects
68 Mrs Mary Liven
So, when to let the computer take over?

Whenever it works and will save you time!

# of files/items to wrangle

<table>
<thead>
<tr>
<th>Time / Effort to Complete</th>
<th>Potential for Calamity</th>
</tr>
</thead>
<tbody>
<tr>
<td>a file; page of text, etc.</td>
<td>many, diverse datasets; corpora</td>
</tr>
</tbody>
</table>
Spreadsheets: The frenemy of research

Doing it manually

# of files/items to wrangle

- a file;
- page of text, etc.
- many, diverse datasets;
- corpora

Potential for Calamity

Time / Effort to Complete
Spreadsheets: The frenemy of research

- Doing it manually
- Doing it automatically

<table>
<thead>
<tr>
<th>Time / Effort to Complete</th>
<th>Potential for Calamity</th>
</tr>
</thead>
<tbody>
<tr>
<td>a file; page of text, etc.</td>
<td></td>
</tr>
<tr>
<td>many, diverse datasets; corpora</td>
<td></td>
</tr>
</tbody>
</table>

# of files/items to wrangle

Also see this: https://xkcd.com/1205/ (thanks Chris!)
Flaws Found in Study Favored by Backers of Austerity

By MATTHEW DALTON and GEOFFREY T. SMITH
April 18, 2013 4:29 p.m. ET

But a study published this week by three economists at the University of Massachusetts says the Reinhart-Rogoff findings contain basic errors, including one involving their spreadsheets that omitted five countries from the result.

Correcting for these errors largely causes the Reinhart-Rogoff finding to disappear, according to the new study by the economists Thomas Herndon, Michael Ash and Robert Pollin.
Thursday, Ortiz ultimately agreed with Geissler that the state database showed 1,791 inactive ballots were returned in Pueblo County. Ortiz blamed simple counting errors in his office for his inaccurate number. The state computer data is cumbersome to read, so Ortiz’s office converted its display to a different form of spreadsheet. That’s where the counting error occurred, the clerk said. Anytime there is a dispute over tallying votes, it always raises suspicions about possible voter fraud.

The town mistakenly reckoned it had $1.5 million more in this year’s budget than it actually has and must now use $600,000 in unexpected state aid to help fill that gap, officials said yesterday.

Chief Financial Officer Mary Ellen Kelley said she takes responsibility for the mistake, which she found Wednesday night on the debt services line item in the $208.6 million fiscal 2012 operating budget. She said a figure went missing as staff managed “monstrous spreadsheets.”

“[T]he town is frustrated,” Kelley said yesterday. “I hate when we make mistakes. People are human and they do make mistakes, but I hate it.”
Activity: Why wrangle?

All workshop materials are available in the following Google Drive folder: https://goo.gl/u53tkz

Open one of the folders in the Google Drive; explore the contents

➢ Scraped reddit subreddit pages in JSON and HTML
  ➢ Google sheet, transformed from html page (import-reddit-xxxxxxx) - created with www.import.io/

➢ Google sheet of structured twitter data (#xxxxxxxx) - created with https://tags.hawksey.info/

Devise a plan of how you might use these data to explore a research question.

➢ How will you analyze the data?
➢ In what ways will you need to clean | transform the data?
➢ What tools might work for this?
Trump:
#Trump2016
https://www.reddit.com/r/the_donald

Bernie:
#FeelTheBern
https://www.reddit.com/r/sandersforpresident

Hillary:
#ImWithHer
https://www.reddit.com/r/hillaryclinton/
Depositing for Use & Sharing
Depositing for Use & Sharing

Your data can (and should!) outlive your research project

Consider how you can make your data more:
➢ adaptable, flexible, extensible to other uses
➢ secure and sustainable
➢ collaborative / social
Depositing for Use & Sharing

How might you need to package/adjust your data in order to:

➢ be better able to show (demonstrate) it to people?

➢ make it work with different tools?
  → Keep data in translatable formats (or formats that work well with visualization)
  → Explore formatting requirements/compatibility for other community tools

➢ have it be discoverable and understandable to others?

➢ maximize its lifespan?
  → Document it!
  → Organize it.
  → Archive it to an appropriate repository (IR, data repository)
Final Thoughts and Strategies
Often, experimentation and iteration are important in establishing the best way to get your data into ‘shape’.

Starting with a sample of data is a good approach.

It’s important to document your outcomes and your process.
How to save time in the long run…

➢ Look for tools that exist to help you wrangle your data
  → automated or semi-automated (guided) cleaning
  → data transformation
  → converting between data formats

➢ Seek out tutorials / instruction for the tools you’re using

➢ Control your data at the point of collection - refine your process to reduce ‘garbage in’
  → e.g. when crowdsourcing data -- use controlled fields and vocabularies; insert data validation processes