Data Wrangling with OpenRefine

Jay Brodeur

Do More with Digital Scholarship Workshop Series

February 1, 2021

Lewis & Ruth Sherman Centre for Digital Scholarship
scds.ca

McMaster University Library

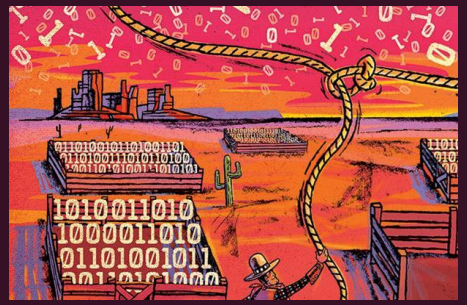
McMaster University sits on the traditional Territories of the Mississauga and Haudenosaunee Nations, and within the lands protected by the "Dish With One Spoon" wampum agreement.

Learning Objectives

By the conclusion of this workshop, you should be able to:

- Describe the basic principles of data wrangling and situations where it is useful
- Apply the basic functionality of OpenRefine to clean, explore, and analyze messy data
- Use OpenRefine to wrangle your own ‘messy’ data
- Identify where to find more information and support

Data & Data Wrangling



	A	B	C
1	Data	Results	Formula
2	drapes	1	=countif(A2:A6, "drapes")
3	grapes	1	=countif(A2:A6, A2)
4	grapeshot	2	=countif(A2:A6, "?rapes")
5	grapefruit	3	=countif(A2:A6, "?rapes*")
6	grapevine	4	=countif(A2:A6,"grape*")
7	100	1	=countif(A7:A10, "100")
8	1,000	1	=countif(A7:A10, A7)
9	10,000	2	=countif(A7:A10, "<=1000")
10	100,000	3	=countif(B7:B10, "<>"&C12)
11		4	=countif(B7:B10, "<="&D12
12	More Data:	1,000	100,000

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

Ideation

Ideas
Interests
Questions
Hypotheses

Analyses / Applications

Methods
Tools
Visual
Textual
Qualitative
Quantitative

Dissemination

Visualization
Multimedia
Monograph
Article
Web Page

Ideation

Ideas
Interests
Questions
Hypotheses

Data

Analyses / Applications

Methods
Tools
Visual
Textual
Qualitative
Quantitative

Dissemination

Visualization
Multimedia
Monograph
Article
Web Page

Ideation

Ideas
Interests
Questions
Hypotheses

Data Preparation

Getting data

Analyses / Applications

Methods
Tools
Visual
Textual
Qualitative
Quantitative

Dissemination

Visualization
Multimedia
Monograph
Article
Web Page

Ideation

Ideas
Interests
Questions
Hypotheses

Data Preparation

Getting data

Assessing it

Cleaning it

Transforming it

Depositing it for use and sharing

Data Wrangling

Analyses / Applications

Methods
Tools
Visual
Textual
Qualitative
Quantitative

Dissemination

Visualization
Multimedia
Monograph
Article
Web Page

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

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?

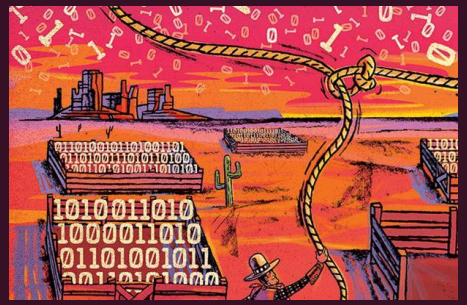
Because this happens →

Data
Your data
Your data
Your data
Your
Data

this can be a problem

From School of Data's Data Cleaning Module:
"the Invisible Man is in your spreadsheet, messing with your data"
http://schoolofdata.org/handbook/courses/data-cleaning-invisible-man-in-spreadsheets/

Understanding your needs



	A	B	C
1	Data	Results	Formula
2	drapes	1	=countif(A2:A6, "drapes")
3	grapes	1	=countif(A2:A6, A2)
4	grapeshot	2	=countif(A2:A6, "?rapes")
5	grapefruit	3	=countif(A2:A6, "?rapes*")
6	grapevine	4	=countif(A2:A6,"grape*")
7	100	1	=countif(A7:A10, "100")
8	1,000	1	=countif(A7:A10, A7)
9	10,000	2	=countif(A7:A10, "<=1000")
10	100,000	3	=countif(B7:B10, "<>"&C12)
11		4	=countif(B7:B10, "<="&D12
12	More Data:	1,000	100,000

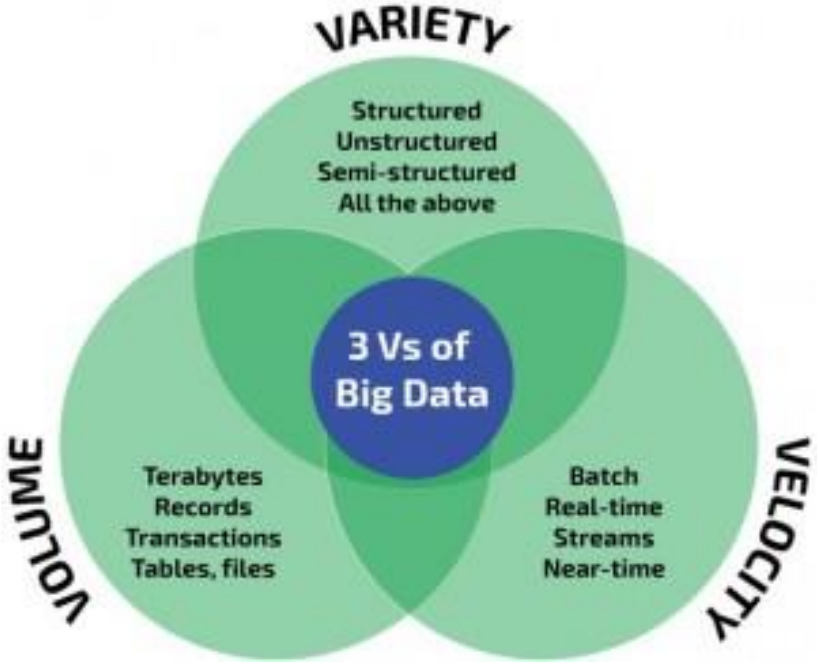
What is your ‘data’?

What kind of ‘wrangling’ is required?

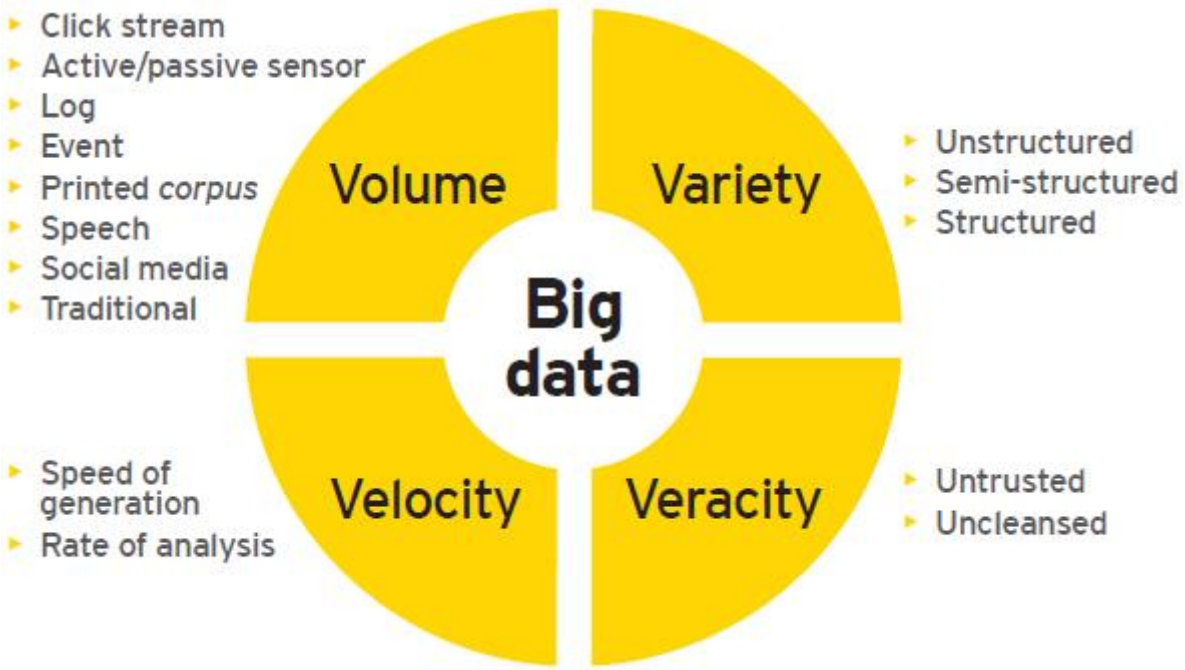
Data wrangling: BIG impact, even for small data

Whether your data is small or “BIG”, conditioning it is critical.

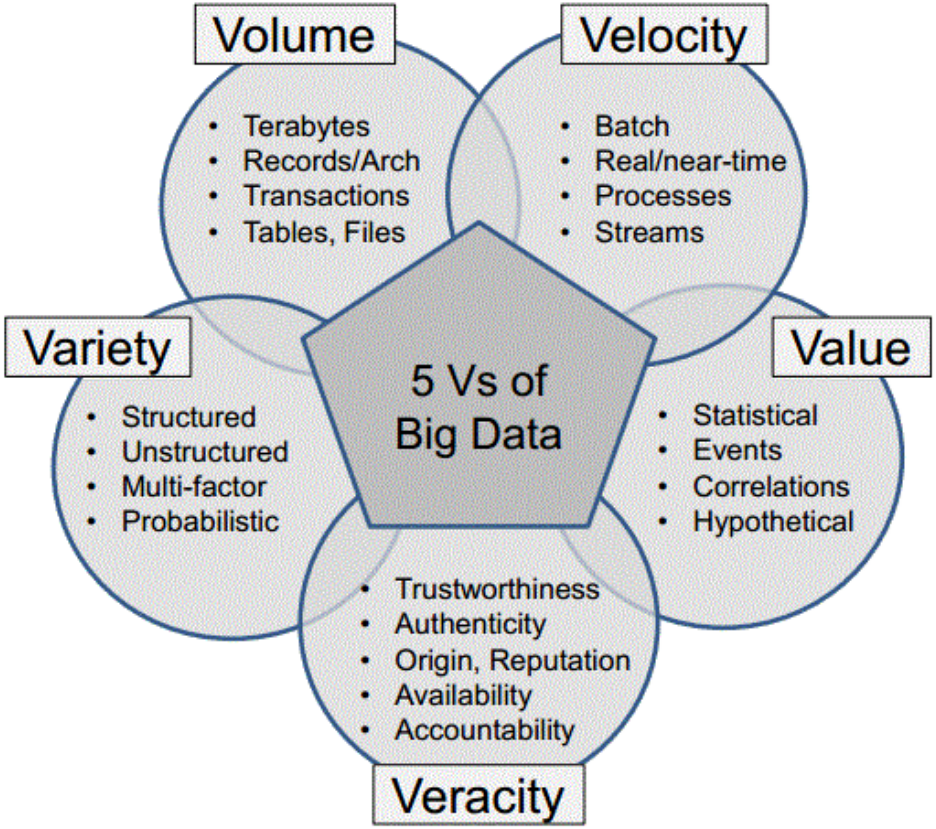
Big data is defined by the 3 “V”s



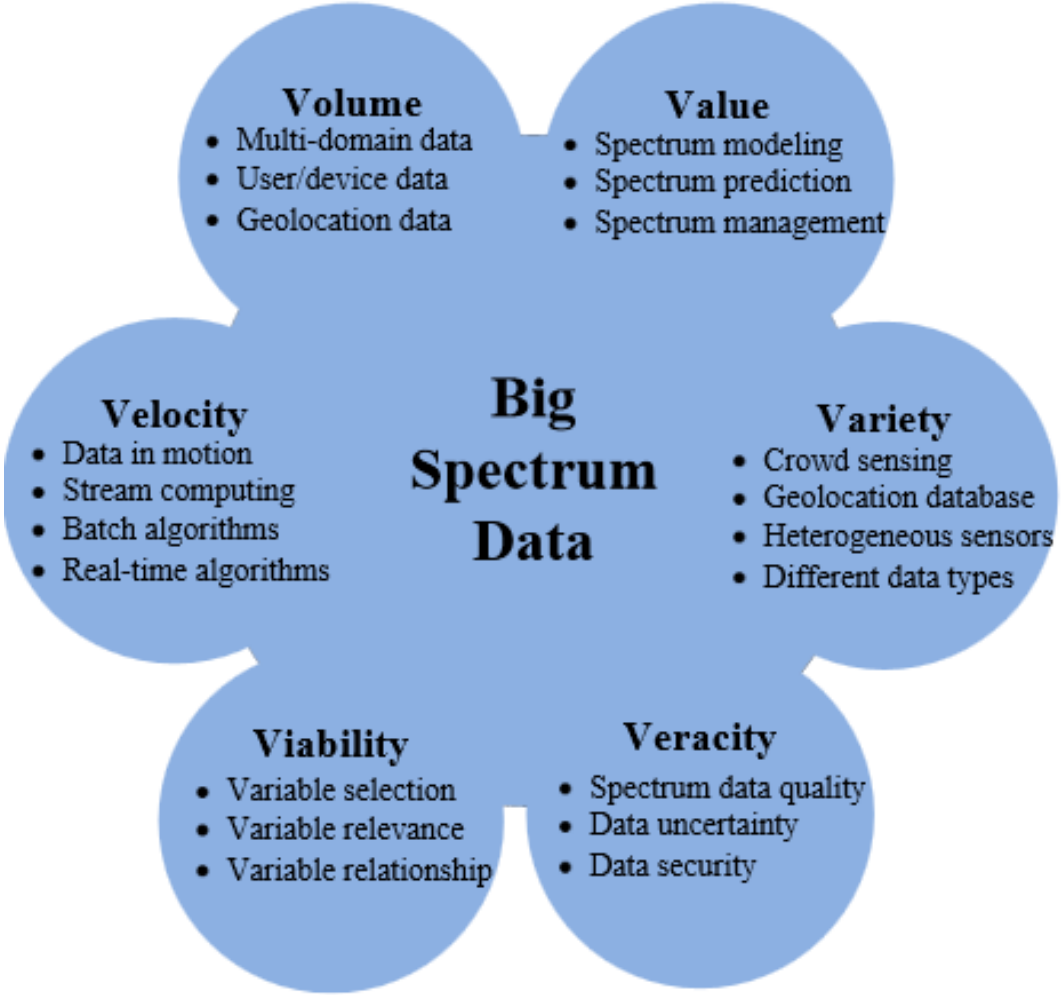
...or the 4 “V”s


Click stream
Active/passive sensor
Log
Event
Printed *corpus*
Speech
Social media
Traditional
Volume
Variety
Unstructured
Semi-structured
Structured
Big data
Speed of generation
Rate of analysis
Velocity
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
5 Vs of Big Data
Value
Statistical
Events
Correlations
Hypothetical
Trustworthiness
Authenticity
Origin, Reputation
Availability
Accountability
Veracity

6? Really?


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
Big Spectrum Data
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

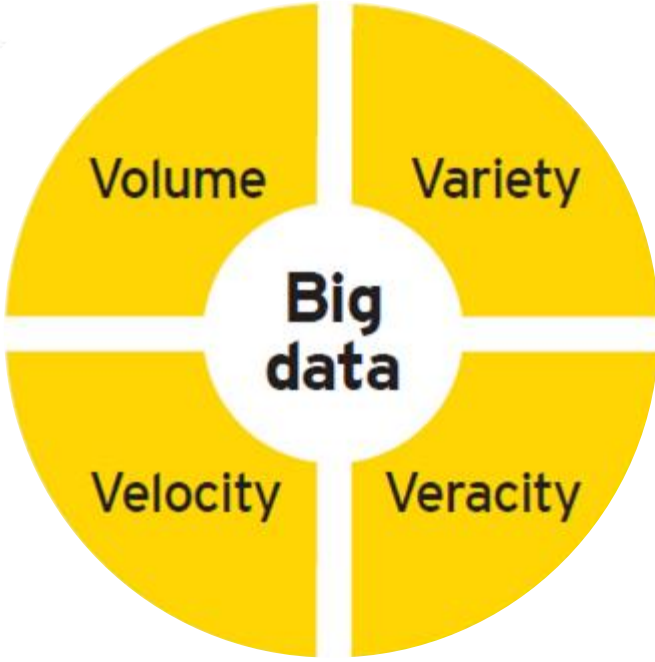
Factors Influencing Your Wrangling Needs

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...

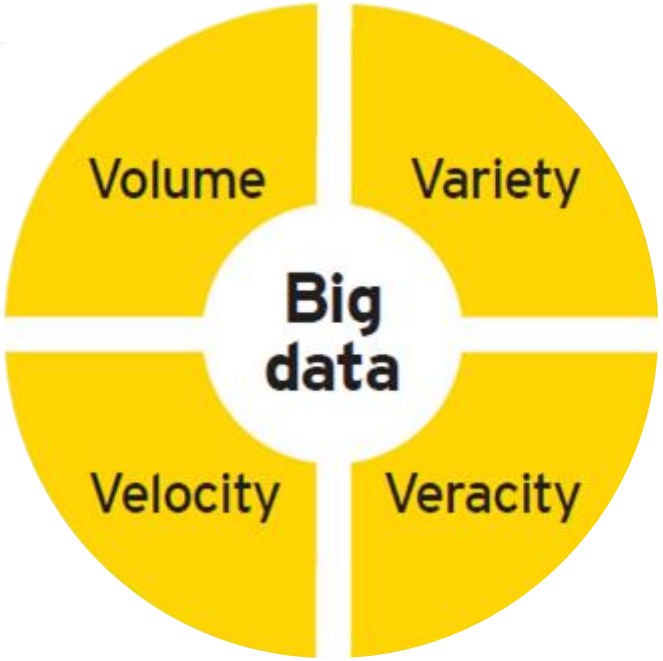
All of these factors influence your data wrangling needs and potential solutions!



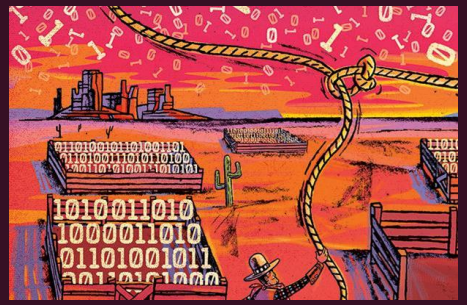
What else matters?

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



	A	B	C
1	Data	Results	Formula
2	drapes	1	=countif(A2:A6, "drapes")
3	grapes	1	=countif(A2:A6, A2)
4	grapeshot	2	=countif(A2:A6, "?rapes")
5	grapefruit	3	=countif(A2:A6, "?rapes*")
6	grapevine	4	=countif(A2:A6,"grape*")
7	100	1	=countif(A7:A10, "100")
8	1,000	1	=countif(A7:A10, A7)
9	10,000	2	=countif(A7:A10, "<=1000")
10	100,000	3	=countif(B7:B10, "<>"&C12)
11		4	=countif(B7:B10, "<="&D12
12	More Data:	1,000	100,000

- 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, S


```
{
  "name": "ballparks",
  "type": "FeatureCollection",
  "features": [{
    "type": "Feature",
    "geometry": {
      "type": "Point",
      "coordinates": [-112.066564, 33.445081]
    },
    "properties": {
      "Class": "Majors",
      "League": "Major League Baseball",
      "Team": "Arizona Diamondbacks",
      "Ballpark": "Chase Field",
      "Lat": "33.445081",
      "Long": "-112.066564"
    }
```



```
<typeOfResource>cartographic</typeOfResource>
<genre authority="lctgm">Aerial photographs</genre>
<originInfo>
<publisher>Air Photo Division, Energy Mines + Resources</publisher>
<place>
<placeTerm type="text">[Place of publication unknown]</placeTerm>
</place>
<dateCreated>1966</dateCreated>
<dateOther>1966</dateOther>
</originInfo>
<language>
<languageTerm type="code" authority="iso639-2b">eng</languageTerm>
</language>
<physicalDescription>
<extent>[1:37,000 approximately]</extent>
```

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: http

New York Times

Web Scraping

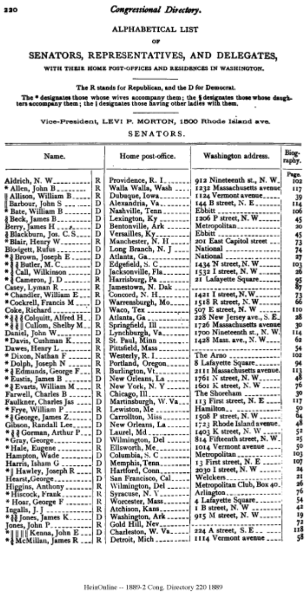
Scraping vs. Parsing:

- 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 →

Scanned / OCRed Data




```
| Congressional Directory.

ALPHABETICAL LIST
OF
SENATORS, REPRESENTATIVES, AND DELEGATES,
WITH THEIR HOME POST-OFFICES AND RESIDENCES IN WASHINGTON.
The * designateg those whose wives accompany them; the I designates those whose daughters accom.
piny them; the I designates those having other ladies with them.
SENATORS.

Name.                  Home post-office.     Washington address.  Bi
P
Aldrich, N. W -------------Providence, R. I ------.........................
Allison, William B ---------- Dubuque, Iowa --------  24Vermont avenue -----
Bate, William --------------Nashville, Ten ------- Ebbitt House
Beck, James B -------------Lexington, Ky ................................
*  Berry, James I    .---------- Bentonville, Ark ------ National Hotel ----------
*  Blackburn, Joseph C. S a- Versailles, Ky ......... Ebbitt House
* Blair, I lenry \V ----------- Manchester, N. H -  2o East Capitol street.._._"
Blodgett, Rufus ------------Long Branch, N. J
Bowen, Thomas M ----------Del Norte, Colo .......
Brown, Joseph E ----------- Atlanta, Ga ----------- Woodmont Flats ---------
Butl, M. C ---------------Edgefield, S. C --------- 1751 P street, N. W ------
Cal\ilkinson ------------- Jacksonville, Fla -------- 1903 N street, N. W -----
Came1. D  .----       Harrisburg, Pa --------- 21 Lafayette Square ------
Chacthan ------------Providence, R. I ------.........................
lliam E -------- Concord, N. H -------- 1421 I street, N.W -------
is M ---------Warrensburgh,Mo --- I518 R street, N.W ------
Waco, Tex ----------- 419 Sixth street, N. W ---
I       Atlanta, Ga ----------- 920 New York avenue ...
--------Springfield, Ill --------- 1402 Massachusetts avenue
-      Lynchburgh, Va -------- I700 Nineteenth st., N. W
-------- Saint Paul, Minn ---- 17oo Fifteenth street, N. W
Pittsfield, Mass -------- 1632Rhode Island avenue.
-----------Portland, Oregon ---- 8 Lafayette Square -
ge F --------- Burlington, Vt --------- 2111 Massachusetts avenue
Eustes B --   -     New Orleans, La ---- 1761 N street, N. W ------
Evailliam M -      ---------New York, N. Y ---- i6oi K street, N. W -----
Farv, Charles B ----------- Chicago, Ill ----------- -------------------
Faulrer, Charles James --- Martinsburgh, W. Va -.-----------------------
Frye, William P ------------Lewiston, Me ---------- Hamilton House ---------
George, James Z ------------Jackson, Miss ---------- Metropolitan Hotel ......
Gibson, Randall Lee -------- New Orleans, La ---- 1723 Rhode Island avenue.
Gorman, Arthur P ---------- Laurel, Md .o---------1403 K street, N. W .....
Gray, George --------------Wilmington, Del ---------------------------
Hale, Eugene               Ellsworth, Me --------- 917 Sixthteenth st., N.W._
Hampton, Wade ....---------- Columbia, S. C -------- -------------- --------
Harris, Isham G ------------ Memphis,Tenn -------- 13 First street, N. E.
* Hawley, Joseph R          Hartford, Corn -------- 1514 K street, N. W ....
Hearst, George ------------San Francisco, Cal -------------------------
*Hiscock, Frank ----------- Syracuse, N. Y -------- Arlington Hotel .........
Hoar, George F ------------ Worcester, Mass ------- 1325 K street, N. W -.
*   Ingalls, John James ------- Atchison, Kans -------- I B street, N. W ........
Jones, James K      -      Washington, Ark ---- 915 M street, N. W ....
Jones, John P      --      Gold Hill, Nev ................................
Kenna, John E -------------Charleston, W. Va --- 14o B street, N. W -------
```

ACCOU NTA NT,
AUDITOR AND ADJUSTER.

CITY OF HAMILTON

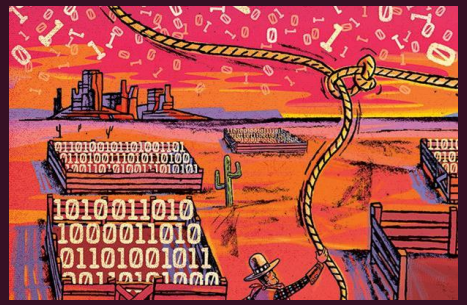
q6 Jas Ramage, laborer
96 Sand Garrity
98 Geo 1-ake, granite cutter
too Alex Martin, laborer
tog Godfrey Auld, laborer
to6 Jas Flynn, laborer
110 Jas Furlong, laborer
112 11 A Broadbent, brakinn
ta Henry N'ernon, directory
pulilisher
16 Matthew Hunter, carprar
118 Herbert Dixon

Yoik 11 intersect(

36 A H McKeown, tnerchant
40 Thos 11 Jermvn. bookkpr
42 Jas Mclaugrain, machist

Main st Intersects
68 Mrs Mary Liven

Cleaning & Transforming Your Data



	A	B	C
1	Data	Results	Formula
2	drapes	1	=countif(A2:A6, "drapes")
3	grapes	1	=countif(A2:A6, A2)
4	grapeshot	2	=countif(A2:A6, "?rapes")
5	grapefruit	3	=countif(A2:A6, "?rapes*")
6	grapevine	4	=countif(A2:A6,"grape*")
7	100	1	=countif(A7:A10, "100")
8	1,000	1	=countif(A7:A10, A7)
9	10,000	2	=countif(A7:A10, "<=1000")
10	100,000	3	=countif(B7:B10, "<>"&C12)
11		4	=countif(B7:B10, "<="&D12
12	More Data:	1,000	100,000

Data Cleaning: Goal & Activities

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

Data cleaning (scrubbing) may include:

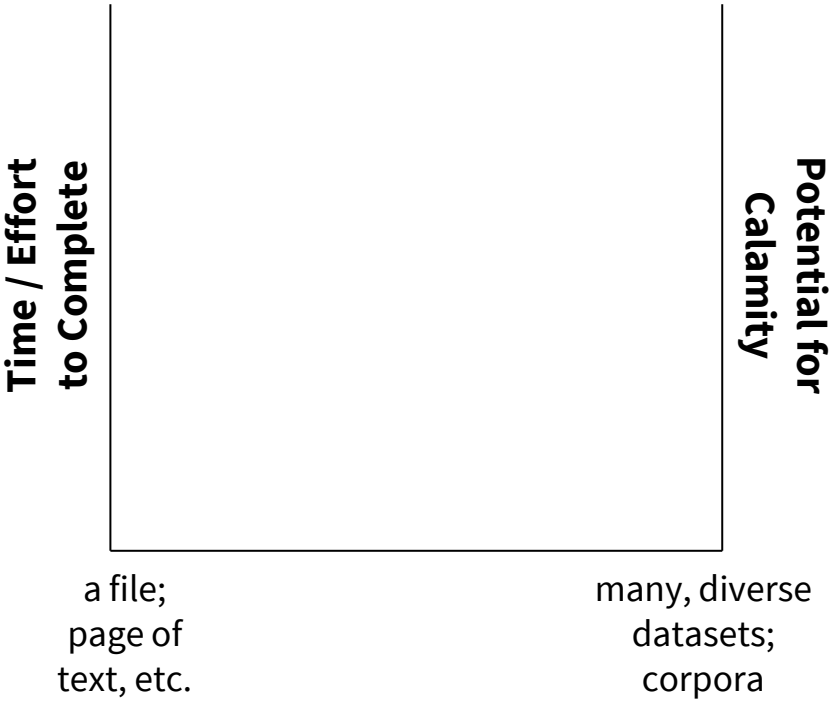
- Detecting and remediating corrupt/inaccurate records
- Removing typographical errors
- Validating against a known list of possibilities (e.g. verify a string as a postal code)
- Eliminating duplicate entries
- Harmonizing and standardizing (e.g. represent ‘St’, ‘St.’, ‘Street’ as ‘street’)

Cleaning may be carried out

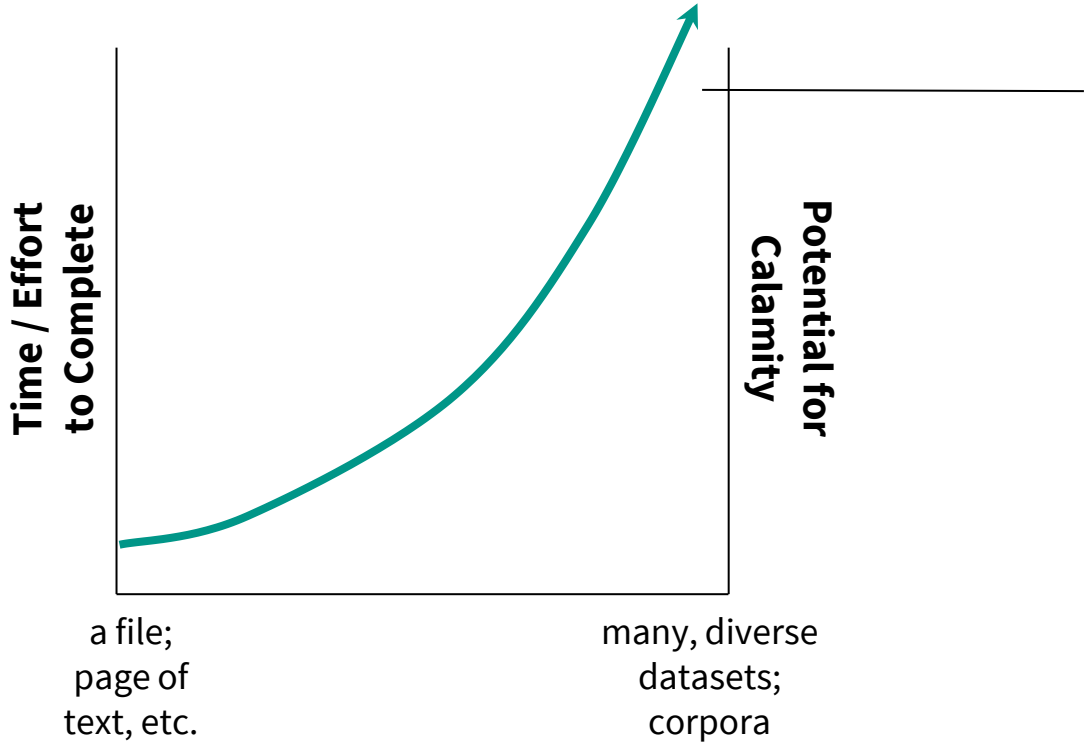
- Manually (interactive)
- Semi-automatically (guided)
- Automatically (scripted)

So, when to let the computer take over?

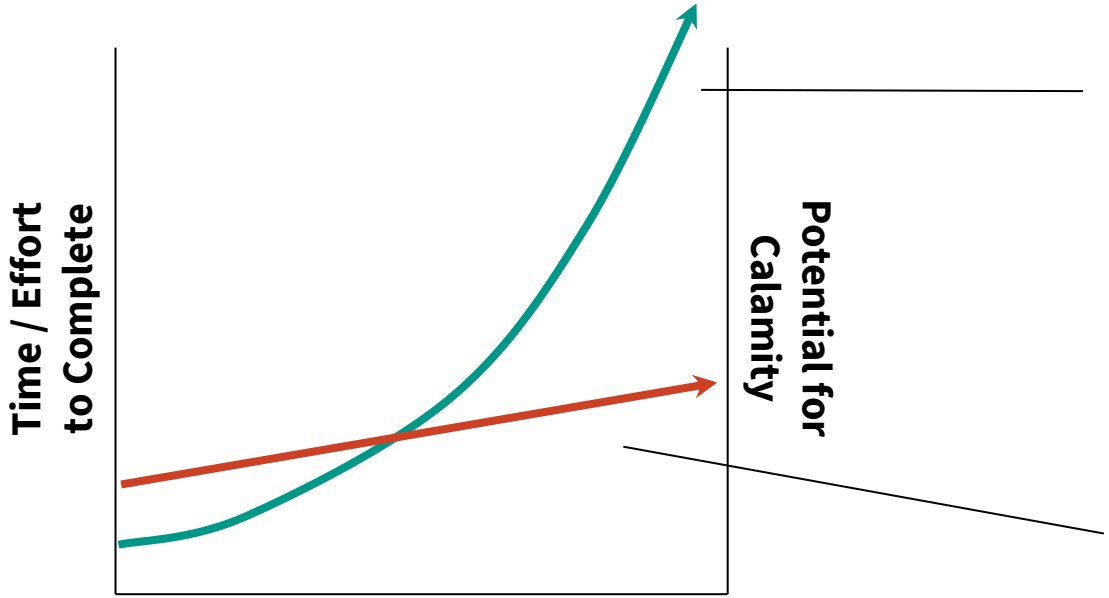
Whenever it works and will save you time!



Spreadsheets: The frenemy of research



Spreadsheets: The frenemy of research

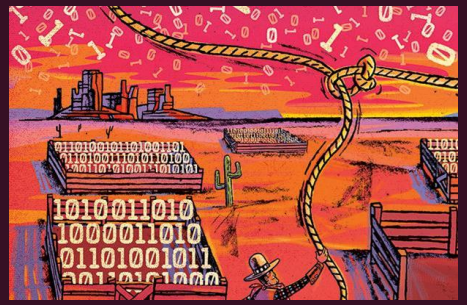


Doing it manually

Doing it automatically

Also see this:
https://xkcd.com/1205/
(thanks Chris!)

Final Thoughts and Strategies



	A	B	C
1	Data	Results	Formula
2	drapes	1	=countif(A2:A6, "drapes")
3	grapes	1	=countif(A2:A6, A2)
4	grapeshot	2	=countif(A2:A6, "?rapes")
5	grapefruit	3	=countif(A2:A6, "?rapes*")
6	grapevine	4	=countif(A2:A6,"grape*")
7	100	1	=countif(A7:A10, "100")
8	1,000	1	=countif(A7:A10, A7)
9	10,000	2	=countif(A7:A10, "<=1000")
10	100,000	3	=countif(B7:B10, "<>"&C12)
11		4	=countif(B7:B10, "<="&D12
12	More Data:	1,000	100,000

Experimentation and Documentation

Often, experimentation and iteration are important in establishing the best way to get your data into ‘shape’.

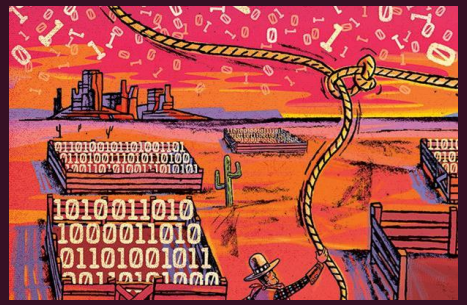
Starting with a sample of data is a good approach.

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

Data Wrangling with OpenRefine



	A	B	C
1	Data	Results	Formula
2	drapes	1	=countif(A2:A6, "drapes")
3	grapes	1	=countif(A2:A6, A2)
4	grapeshot	2	=countif(A2:A6, "?rapes")
5	grapefruit	3	=countif(A2:A6, "?rapes*")
6	grapevine	4	=countif(A2:A6,"grape*")
7	100	1	=countif(A7:A10, "100")
8	1,000	1	=countif(A7:A10, A7)
9	10,000	2	=countif(A7:A10, "<=1000")
10	100,000	3	=countif(B7:B10, "<>"&C12)
11		4	=countif(B7:B10, "<="&D12
12	More Data:	1,000	100,000

What is OpenRefine?

- Formerly known as “Google Refine”
- Free & open-source
- Tool for cleaning messy, large-ish datasets (100,000 rows)
- Desktop application – runs locally through your (modern) browser, e.g., Firefox or Chrome

Why OpenRefine?

- Powerful, but ‘easy’ to use – low learning curve
- Can import data from variety of sources, including directly from the web
- Retains log of all data cleaning actions (i.e. you can revert your changes)
- Does not modify original data files
- Browser-based, so works on most computers (can even run it off of a flash drive)
- Can track, export, and re-apply actions across datasets
- Apply powerful clustering algorithms to help clean your data

https://openrefine.org

Time to get started

Go to scds.github.io/data-wrangling/ for a introduction and preparation

Then, complete the Data Carpentries self-serve OpenRefine Workshop
https://datacarpentry.org/openrefine-socialsci/

If you have questions, register for the OpenRefine drop-in session
on February 11, 2021:
u.mcmaster.ca/dmds-data-wrangling-drop-in