Data Wrangling with OpenRefine

Jay Brodeur

Do More with Digital Scholarship Workshop Series February 1, 2021







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	В	С	A	
O Continuent 1 200 0	1	Data	Results	Formula		
The second sector	2	drapes	1	=countif(A2:A6, "drapes")		
		grapes	1	=countif(A2:A6, A2)	A REAL PROPERTY AND A REAL	
	4	grapeshot	2	=countif(A2:A6, "?rapes")		
y'y' Karmanan	5	grapefruit	3	=countif(A2:A6, "?rapes*")		
01101000101101001101	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)		
01101001010	11		4	=countif(B7:B10, "<="&D12		
	12	More Data:	1,000	100,000		
	10					





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
 - \rightarrow It helps inform you of the ways in which your data can be used
 - -- by you, your collaborators and others in your research community





Ideas Interests Questions Hypotheses Analyses / Applications Methods Tools Visual Textual Qualitative Quantitative

Dissemination

Visualization Multimedia Monograph Article Web Page





Ideas Interests Questions Hypotheses

Data

Analyses / Applications Methods Tools Visual Textual Qualitative Quantitative

Dissemination

Visualization Multimedia Monograph Article Web Page





Ideas Interests Questions Hypotheses

Data Preparation

Getting data

Analyses / Applications Methods Tools Visual Textual Qualitative Quantitative

Dissemination

Visualization Multimedia Monograph Article Web Page







Data Preparation Getting data Assessing it Cleaning it Transforming it Depositing it for use and sharing

Analyses / **Applications** Methods Tools Visual Textual Qualitative Quantitative

Dissemination

Visualization Multimedia Monograph Article

McMaster Library University 🔛

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.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 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.





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 \rightarrow

Data	
Your data	this can be a problem
Your data	
Your data	•
Your	
Data	

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

~ ~ P ~ ~ 0 0 ~ 0 ~ ~ ~ ~ 0 / ~ ~ ~ 0 / ~ ~ ~ ()		A	в	C		
O So Continues 1 20 0	1	Data	Results	Formula		
The second second	2	drapes	1	=countif(A2:A6, "drapes")		
	3	grapes	1	=countif(A2:A6, A2)	A Real Property in the local sector	
	4	grapeshot	2	=countif(A2:A6, "?rapes")		
y ' y Repairing	5	grapefruit	3	=countif(A2:A6, "?rapes*")		
011010011010101010	6	grapevine	4	=countif(A2:A6,"grape*")		
	7	100	1	=countif(A7:A10, "100")		
handamana	8	1,000	1	=countif(A7:A10, A7)	Dis AVSI	
	9	10,000	2	=countif(A7:A10, "<=1000")		
	10	100,000	3	=countif(B7:B10, "<>"&C12)		
01101001011			4	=countif(B7:B10, "<="&D12		
20120101000 B	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







...or 5?







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



- Downloading Files
- ► API access
- ➤ Web-scraping





Bulk Downloads

Common data formats

- ➤ ASCII text
- Delimited ASCII (.csv, .tsv)
- ► PDF
- JSON (<u>https://goo.gl/632RGb</u>)
- Markup languages (TEI, HTML, XML)
- Multimedia (.wav, .ogg, .mp4, .mkv)
- Other, program-specific and ad-hoc file formats (fixed width, S<language)</p>

"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> <languageTerm type="code" authority="iso639-2b">eng</languageTerm> </language> <physicalDescription> <extent>[1:37,000 approximately]</extent>





API Access

Lewis & Ruth

scds.ca

an

Centre

for Digital Scholarship

API = Application program interface

- set of protocols/tools for building software applications
- governs how software should interact with each other and user interfaces





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
 - \blacktriangleright Scrapes web page html to file \rightarrow



Librarv



Scanned / OCRed Data

ALPHABETICAL LIST									
OF									
SENATORS, REPRESENTATIVES, AND DELEGATES,									
WITH THEIR HOME POST-OFFICES AND RESIDENCES IN WASHINGTON									
The	R sta	ands for Republican, and the l	D for Democrat.						
The * designates those	whee	e wives accompany them; th	ac § designates those whose day	gh-					
ters accompany them; th	e i de	esignates those having other 1	adies with them.						
Vice-Preside	at. 1	EVI P. MORTON. UN	O Bhode Island ave						
		SENATORS.							
and the second se	4.C. 10								
Name.		Home post-office.	Washington address.	Biog- raphy.					
A14.1.1 N 10		Descriptions D. J.	and Minuteenth et N.W.	Page.					
Allen John B	R	Walla Walla, Wash	1232 Massachusetts avenue	103					
Allison, William B	R	Dubuque, Iowa	1124 Vermont avenue	39					
Barbour, John S	D	Alexandria, Va	144 B street, N. E	114					
* Bate, William B	D.	Nashville, Tenn	Ebbitt	106					
Beck, James D	15	Bentonville Ark	Meteropolitan	43					
Blackburn, Jos. C.S.	D.	Versailles, Ky	Ebbitt	45					
Blair, Henry W	R	Manchester, N. H	201 East Capitol street	73					
Blorigett, Rufus	D.	Long Branch, N. J	National	74					
Brown, Joseph E	15	Edosfald S C	rata N street N W	*7					
Call. Wilkinson	D	Jacksonville, Fla	1532 I street, N. W	26					
Cameron, J. D	R	Harrisburg, Pa	21 Lafayette Square	95					
Casey, Lyman R	R	Jamestown, N. Dak		87					
Chandler, William E	K	Concord, N. H.	1421 I street, N.W.	23					
Coke, Richard	Ď	Waco, Tex	507 E street, N. W	110					
Alfred H.	D	Atlanta, Ga	228 New Jersey ave., S. E.	28					
Cullom, Shelby M	R	Springfield, Ill	1726 Massachusetts avenue	30					
Daniel, John W	B	Lynchburgh, Va	1700 Nineteenth st., N. W.	62					
Dawes Heary L	12	Pittsfield, Mass	1410 1411. 1	54					
Dixon, Nathan F	R	Westerly, R. I.	The Arno	102					
Dolph, Joseph N	R	Portland, Oregon	8 Lafayette Square	94					
* 4 Edmunds, George F	18	New Orleans Lo	1761 N street N W	- 31					
#J Evarts, William M	1ñ	New York, N. Y	1601 K street, N.W	76					
Farwell, Charles B	R	Chicago, Ill	The Shoreham	30					
Faulkner, Charles Jas	D	Martinsburgh, W. Va	113 First street, N. E	117					
* Frye, William P	1R	Lewiston, Me	1608 P street, N.W	50					
Gibson, Randall Lee	16	New Orleans, La	1721 Rhode Island avenue.	- 48					
*]] Gorman, Arthur P	D	Laurel, Md	1403 K street, N. W	52					
Gray, George	D	Wilmington, Del	814 Fifteenth street, N. W.	25					
* Hale, Eugene	K	Ellsworth, Me	Naturnalitan	101					
Harris Isham G	16	Memphis Tenn	13 First street, N. E	107					
* Hawley, Joseph R	R	Hartford, Conn	2030 I street, N. W	24					
Hearst, George	D	San Francisco, Cal	Welckers	21					
Higgins, Anthony	R	Wilmington, Del	Metropontan Club, Box 40.	20					
* MISCOCK, FTREK	R	Worrester Mass	4 Lafavette Square	54					
Incalls, L.I.	R	Atchison, Kans	I B street, N. W	44					
* H Jones, James K	D	Washington, Ark	915 M street, N. W	19					
	I R	Gold Hill, Nev		72					
Jones, John P	12	117 12.	and A street C F	113					

HeisColine -- 1889-7 Cons. Directory 220 1889

Lewis & Ruth **Sherman Centre** for Digital Scholarship scds.ca

Congressional Directory.

Name

Ray s. e s. from 363 King w to Bold

Canada si intersecto

122 John Goodfithow, painter 123 Hiram Denkit, porter 124 John Timesen, performan

Ray s, w s su Gau E Bristall, merchant sz F B Generoing, merchant ag Mex Bahart Furrie

12 F.B. Grepering, myrchant 14 Mex Robert Farrie George st (storyset) 45 S.F. Kons, dag cal bill rev. 26 Subschler, samper

65-21 Garnes-Tilden Co 85 T Band, mathinin Catherins st intreach

Many at indersects

familt's Mics Reofing Co Modeland, whistweight fex M F Missner

Vocadi Jolio Hankey, Interpretate

John at intersects

qf-8 Vacuat von Miss L'Schafer 112 Isaac Soliar, Boaryman 114 Goo Startion, whitewash 126 I Buchingham, plantaren 226 Feed Abel, tailor

John it interacts

ALPHABETICAL LIST DBLANDS, REPRESENTATIVES, AND DELEGATES, WITH THEIR MOVE POST-OFFICES AN RESIDENCES IN MASHINGTON. The <u>sessingates</u> those whose vives accompany then; the I designates those whose daughters <u>ACCOM-</u> SeMANDS.''

Home post-office, Washington address, Bi

Aldrich, N. W ------Allison, William B ---Providence, R. I ----- 24Vermont avenue ----* Blair, I lenry \V -Blodgett, Rufus -----Bowen, Thomas M -------Del Norte, Colo Brown, But Call Cha -Portland, Va. ----- 1700 Hassachusetts avenue Saint Paul, Minn ---- 1700 Fifteenth street, N. W Pittsfield, Mass ----- 1702 SizRhode Island avenue. --Portland, Oregon ---- 8 Lafayette Square -Burlington, Vt ----- 2111 Massachusett New Orleans, La ---- 1761 N street, N. W -2111 Massachusetts avenue ge F -----------New York, N. Y ---- 1601 K street, N. W ---Fryt, William P -----George, James Z -----Gibson, Randall Lee ------ Hamilton House -------Jackson, Miss --Gorman, Arthur P ----Wilmington, Del --Ellsworth, Me ------ 917 Sixthteenth st., N.W. - Columbia, S. C ----- Memphis, Tenn
Memphis, Tenn
13 First street, N. E.
Hartford, Corn
1514 K street, N. W Hartford, Corn ___________ 134 K street, N. W ... San Francisco, Cal ______ Syracuse, N. Y ______ Arlington Hotel ______ Morcester, Mass _______I25 K street, N. W -______ Assington, Ark ______I5 Street, N. W Gold Hill, Nev Charleston, W. Ya ______I40 B street, N. W ______ Jones, John P Kenna, John E ---W. F. FINDLAY, F.C.A. | ACCOUNTANT. AUDITOR AND ADJUSTER CITY OF HAMILTON STREET DIRECTOR 1) Christian Scholer, gild Dy Abrafaan Hill Mrs M. Fragarald Di Mrs M. Fragarald Gan Miller, carpaning J. W. Stewart, foremitte Di Chas Hold, Streesan D. Thus Mol. Alland, bake it June Morcuy, printer it June Morcuy, printer Um F Davis, Jahavar
 Bit Hagis, Santharappe

 May at stream

 Bit Stream

 M High, hut-ficepar Ven F Dawin, pulieter locati accest accest Swith Rysmal for Therenan Rosana Ven II Charle, gebieder J Williams, teammate rosat Woodbas, heber rosat Woodbas, heber J Philips, constructor at Jan Wellangtiir, much Main if entresets in Men Mary Lavel 50 Role Dow, plaateeu 56 Men Deggan 56 Chan Earle, Steenith tim Jas Parling, laborer ter H A Broadbast, broken tig Henry Verson, directory publisher off Matthew Hoster, carport off Matthew Dises Cathorine of subsyste in Vacuati in These May, laborer at Arthur Quance, driver internet Ford of interact tune any a intersted tone allo a JARPA E Cohen, sing douber rod Breen, taiher In Arentin May Easthoure, laborar souce Adams, laborar for West Brotheery Wyst Brotheery Wyst Des Crothag, weidder Ray n. w s Mary of interarms 17 Tex-remengib, unider 17 Tex-remengib, unider 19 J.M. Dominio 19 Joint Linetand, stores etter 20 Joint Colly, obserimsker 20 Mer Elizebeth Conach 18 Mer John Ucaphant T.J. Isand, grucer Napier st saleron Hunter st intersects August is process in Ed Poster, informe in Mre Kato Schadoll il Baht Carrie, carper legit, Mre Elat Maples, Wie Witt, Unsmith y Jarvis Maples, carp 23 O Lakis, refler y Yacust Bold at under sects Ferguson sor interacts Elpin at internets Lumber yard aning Park factory an Gen King, colposter and Vacion Elgin at indernets ong Mun N J White Perguna any and Catheort at indering? ang W G Teyn, shork ang Albud Martin, wandwise etj David Cen, store more S Vacual 17 Richard May, inhere 8; Mrs H Mills Rebecca. n s, from 62 James n to Wellington Jose Massen, gardanar ola M Carroll, mulder fro B Polligrew to: Bosquill, tobarco rol Poter si catoras) McPharmon & Drope, Wellington at intersects Fig. Judis Burrows, induced fig. Jacob Springstend, et up H. N. Tusseare, marble in Mrs. Keysm iff. Gen Elford, inherer og B.J.McCowoll, hugging Reginald, from Oak to Richmond Rebecca, 8.8 on 2 Co. gents' farmig-O G Y Hall Alex Desires, initsecond A G Bass, printer (p Vacard | W Neupern, worders (Plesence at commonly Const Hall teles Honey, patent ager V Honard, coul Kos Brough, tailor Hugham st interac 100 Wei Bradt, mason 112 Chas Robertson 123 Jas Patter, grocer Ford at anterory West are interest n Vacuett & Co, patent N. C. Wellin & Cu, patient multiclear (§ M. Goldsteine, tailler Robit Lefferere, rag carpete Gross Catchpole, understate W. H. Howards, tailler Malvation. Army Barracke

FEDERAL LIFE Head Office, Surplus to Policy-Hamilton. holders, \$704,141.26 Norra's Foundry Vacant Mary at indexacts say David McGioman, meller ado Mrs Mary O'Nell tall Vacsati tyt: Henry Owen, laborer Elpis of interacts N B. N W chedu Forguan anv inheraceli Cathrors at intersects eng Aleo Reid, Indfar oth Mey Elizabeth Report Wellington at auteroscie Jan Hamilton, grocer Arch Martin, molder Vacute Wes Leith, Gaussith East ave internets Robinson.ns. from 202 James s to Cricket grounds Richmond, runs from Hunter w to Poulette 6 Parisk Cody, petter 7 A Nuch, nubber 8 U. Voerg, corporate 9 Wei Walters, labout 10 Rate Anon, labout 11 Biberain, palmar Marmah at intermets Pideria and said area in Brind Park of intermeda pra Birbd George, shoend Robert, s s Robert, n n, from 224 Bay at interacts James is to groera Henry Fuller, taller M Balley We Corpley, carports Hughum and John its in Breact Jon Cattisfian C H Exam, spotstor Was Borrstond, augt Jan Harrison, druggin 387 Adea Chucce, Laboury yu Miw Church 14 Miw Jazath Davadour 14 Miw Jazath Davadour 95 Jan Tanllay, Lehomer 96 Jan Findlay, Lehomer 16 Hani Farowa, Johowen 16 Hani Farowa, Johowen 16 Gan Cowsingth, Jahnerge Wi Lanseen Smith, Jahnerge Kangan Santan Santha Jahnerge Kangan Santha Santha Santha Santharya Kangan Santha Santharya Kangan Kangan Santharya Santharya Kangan Santharya Kanga D. Bertlen, policeman 14 John Dhinoll, laborer 15 Anna J. McNub 17 Yacaut 14 Dauglas Wayles, shored 14 Andrew Corean, clark Caroline st interacty

Calharine st interacti

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

Potential for **Fime / Effort** to Complete Calamity a file; many, diverse page of datasets; text, etc. corpora # of files/items to wrangle

Spreadsheets: The frenemy of research



of files/items to wrangle

Spreadsheets: The frenemy of research



Final Thoughts and Strategies

	A	В	с		
	Data	Results	Formula		
	drapes	1	=countif(A2:A6, "drapes")		
	grapes	1	=countif(A2:A6, A2)		
	grapeshot	2	=countif(A2:A6, "?rapes")		
1 / Y NORMAN / 5	grapefruit	3	=countif(A2:A6, "?rapes*")		
	grapevine	4	=countif(A2:A6,"grape*")		
	100	1	=countif(A7:A10, "100")		
8	1,000	1	=countif(A7:A10, A7)	Mars NOV	
	10,000	2	=countif(A7:A10, "<=1000")		
	100,000	3	=countif(B7:B10, "<>"&C12)		
		4	=countif(B7:B10, "<="&D12		
	More Data:	1,000	100,000		444

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

V O WO WE TO W O TO O O TO I VI UP TO						
NO PRODUCTION PLONDO		A	В	С		
O So Carrieron 1 201 O	1	Data	Results	Formula		
The second sector	2	drapes	1	=countif(A2:A6, "drapes")	Aller Aller	
		grapes	1	=countif(A2:A6, A2)		
	4	grapeshot	2	=countif(A2:A6, "?rapes")		
y' y grander	5	grapefruit	3	=countif(A2:A6, "?rapes*")		
01101001011010001101 0110100111010101000	6	grapevine	4	=countif(A2:A6,"grape*")		
200110101001010101010	7	100	1	=countif(A7:A10, "100")		
handania -	8	1,000	1	=countif(A7:A10, A7)	ANS AVOI	
	9	10,000	2	=countif(A7:A10, "<=1000")	20	
	10	100,000	3 :	=countif(B7:B10, "<>"&C12)	1 Charles and the second second	
01101001011	11		4	=countif(B7:B10, "<="&D12		
anizoicinna di H	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





39

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 <u>scds.github.io/data-wrangling/</u> for a introduction and preparation

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

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



