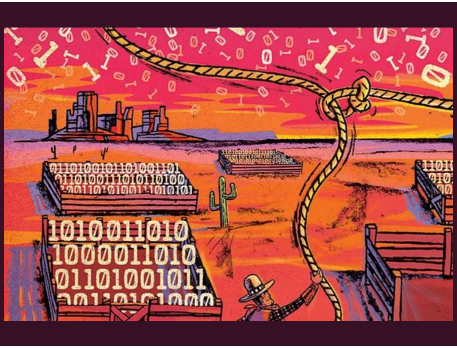
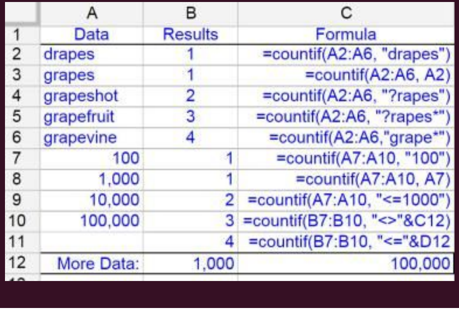
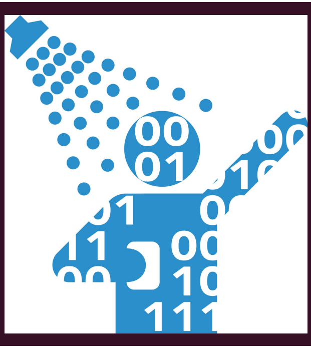
Demystifying Digital Scholarship

An Introduction to 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



Sherman Centre for Digital Scholarship

12-February, 2016

Outline and Objectives

Topics

Let’s talk about data

What is data ‘wrangling’?

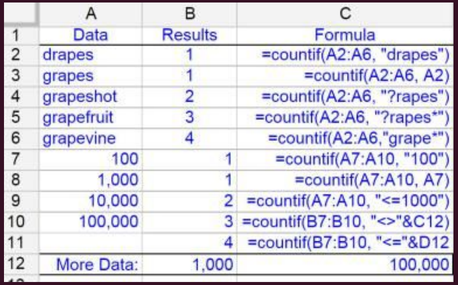
Why you should ‘wrangle’?

How to ‘wrangle’?

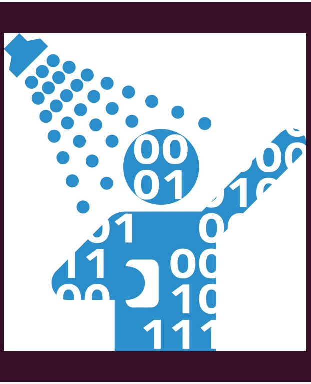
Objectives

Present and discuss strategies for structuring data

Data & Data Wrangling



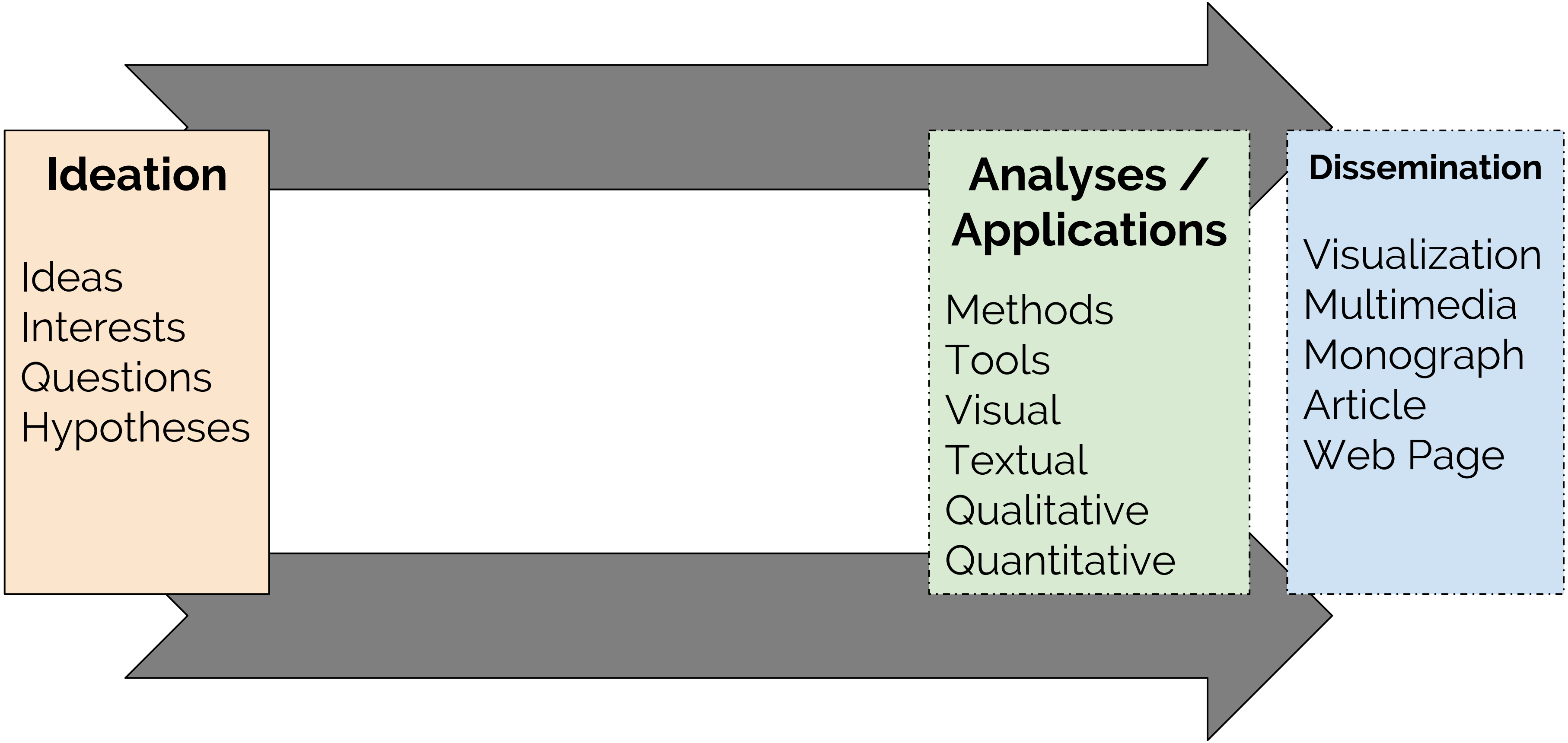
	A	B	C
1	Data	Results	Formula
2	drapes	1	=countif(A2:A6, "drapes")
3	grapes	1	=countif(A2:A6, A2)
4	grapeshot	2	=countif(A2:A6, "?rapes")
5	grapefruit	3	=countif(A2:A6, "?rapes*")
6	grapevine	4	=countif(A2:A6,"grape*")
7	100	1	=countif(A7:A10, "100")
8	1,000	1	=countif(A7:A10, A7)
9	10,000	2	=countif(A7:A10, "<=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

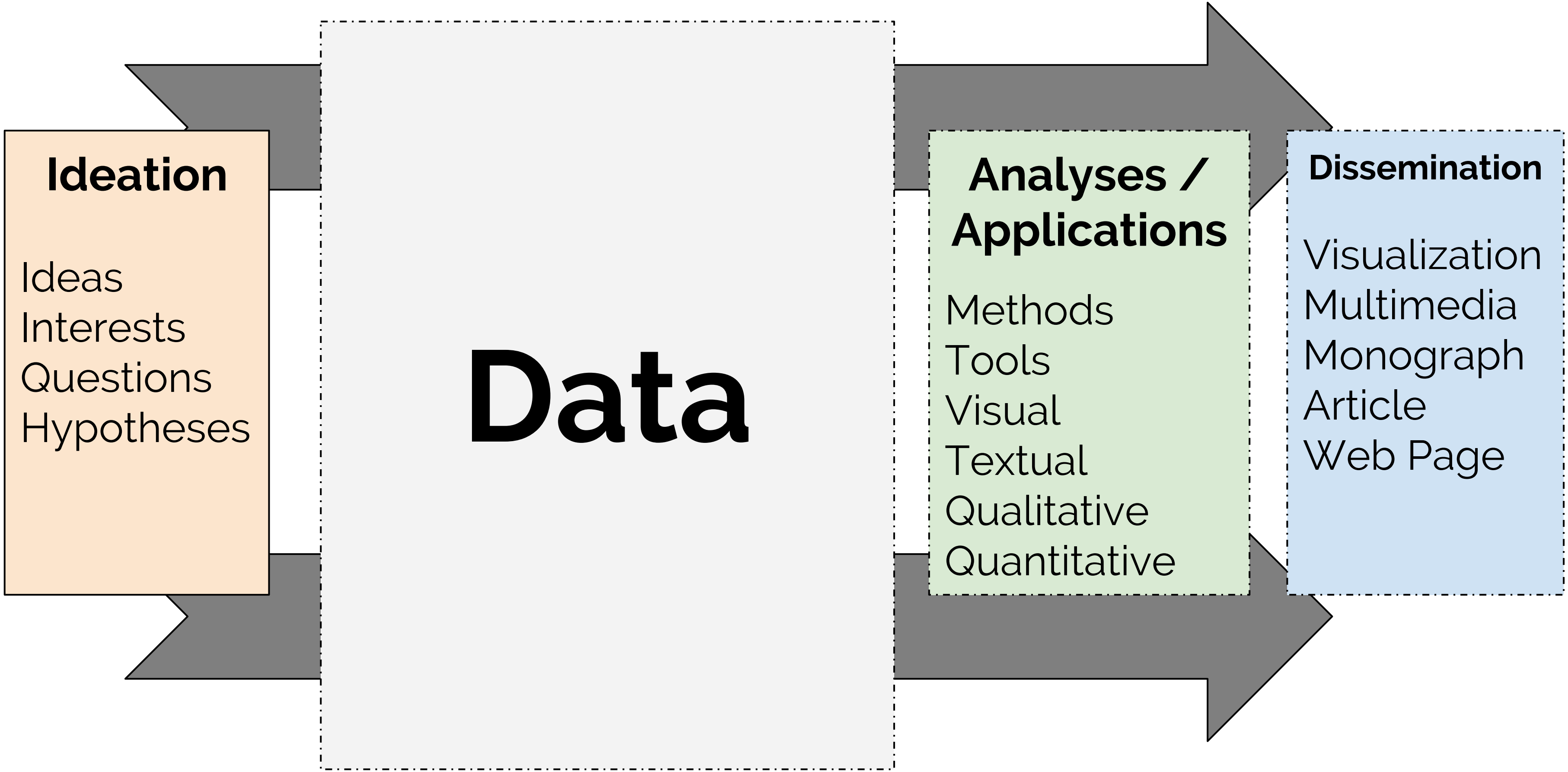
Ideas
Interests
Questions
Hypotheses

Analyses / Applications

Methods
Tools
Visual
Textual
Qualitative
Quantitative

Dissemination

Visualization
Multimedia
Monograph
Article
Web Page



Ideation

Ideas
Interests
Questions
Hypotheses

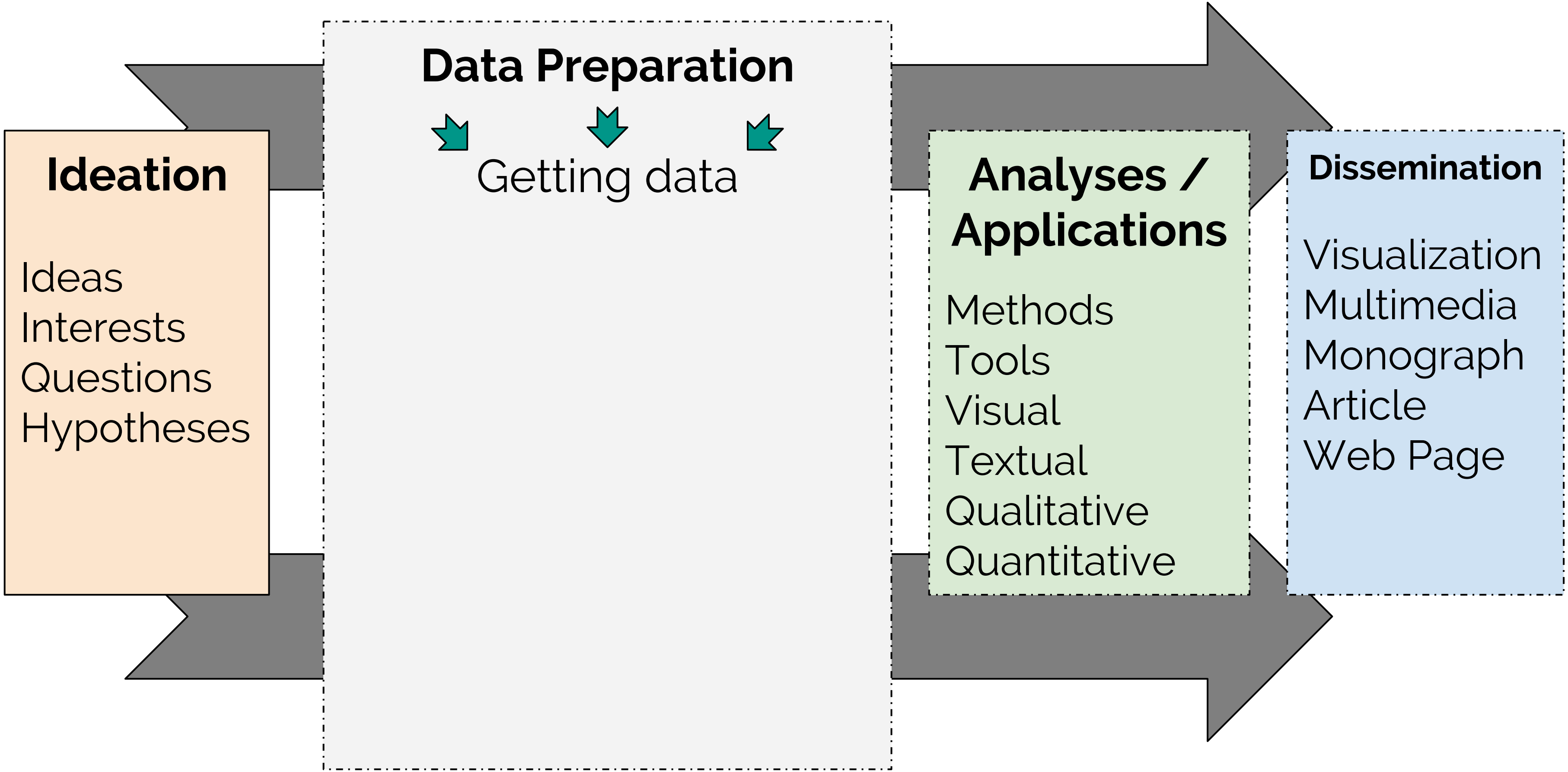
Data

Analyses / Applications

Methods
Tools
Visual
Textual
Qualitative
Quantitative

Dissemination

Visualization
Multimedia
Monograph
Article
Web Page



Ideation

Ideas
Interests
Questions
Hypotheses

Data Preparation

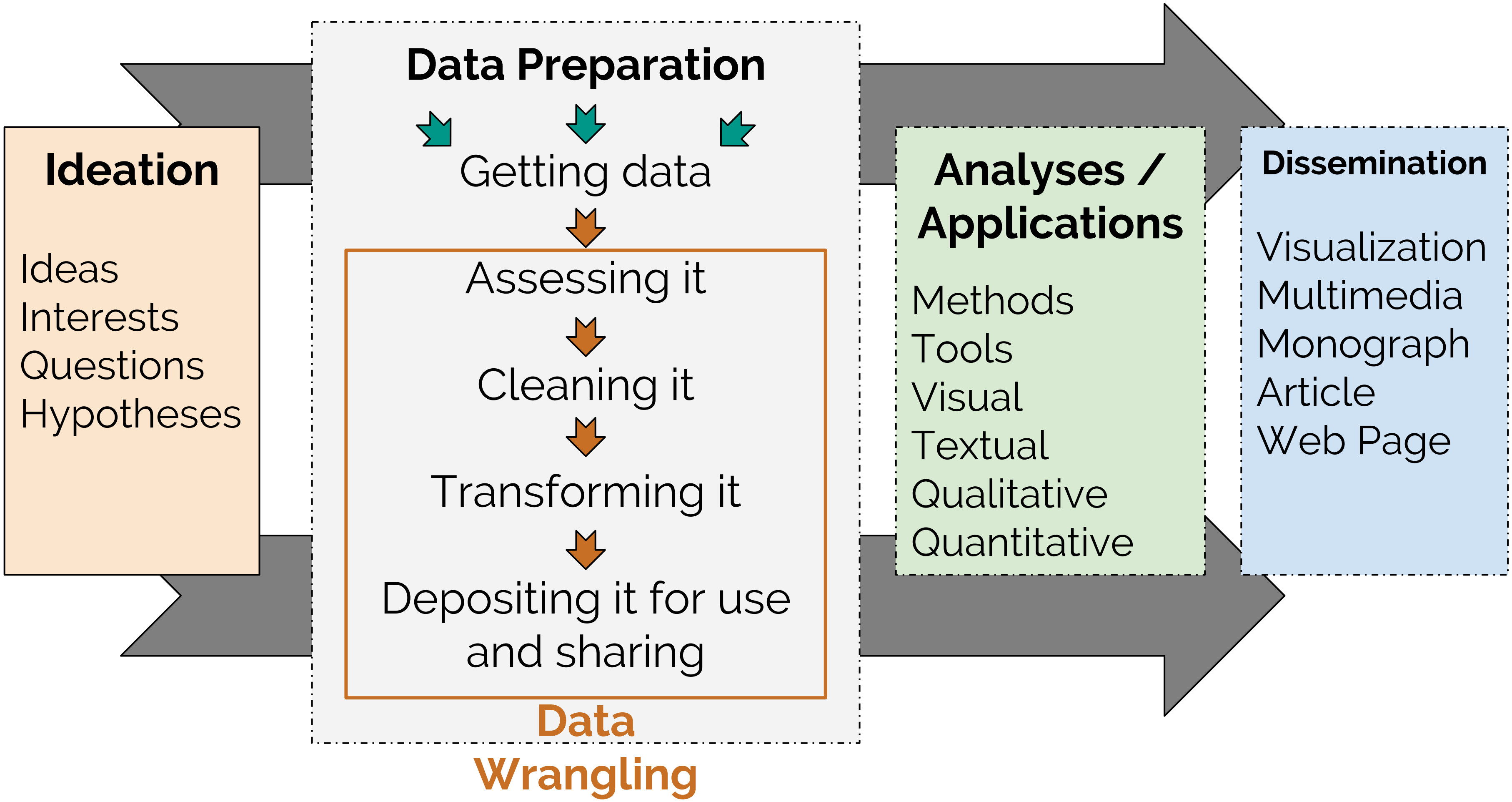
Getting data

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

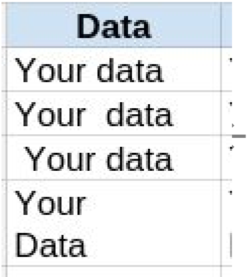
➢ Produces data that connects to tools, collaborators and communities

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 → – , and this can be a problem



Data
Your data
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/

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?

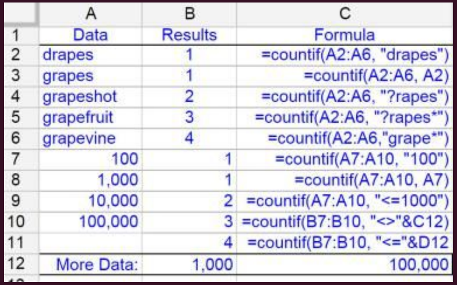
➢ How best to clean this data?

What if it had 30000 rows? What if it amassed 30000 rows / day?

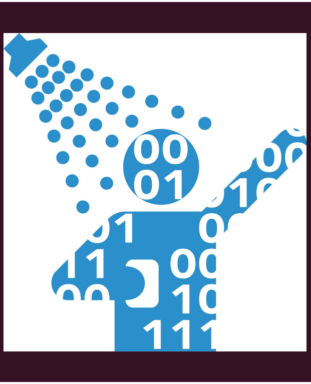
Secret Location

Eva's Map

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



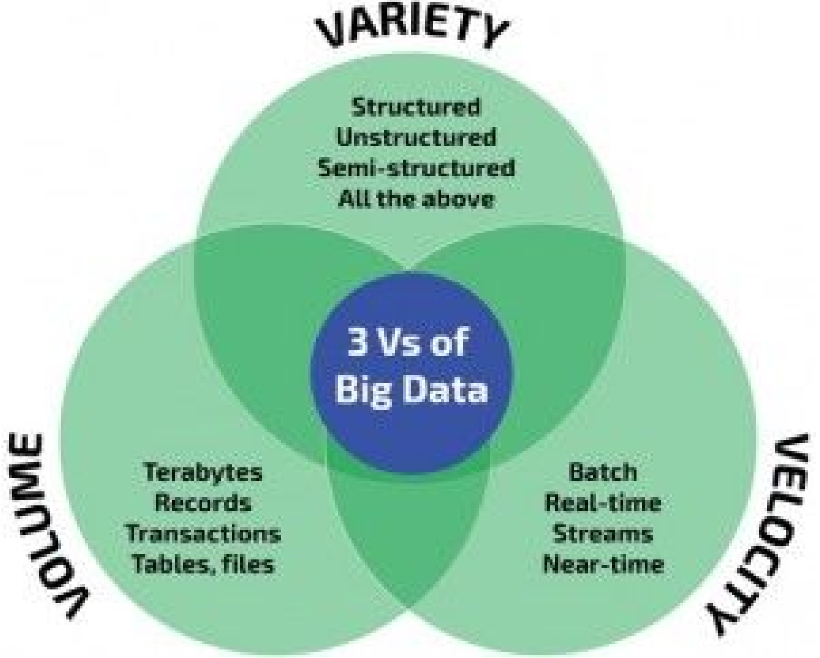
What is your ‘data’?

What kind of ‘wrangling’ is required?

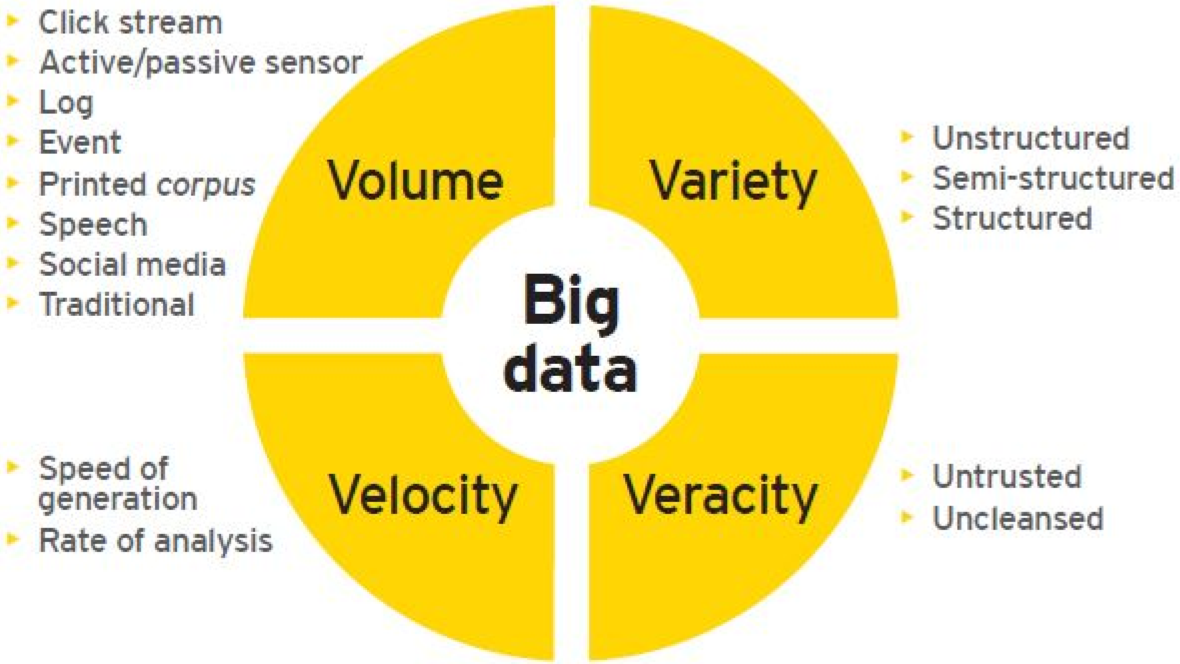
Data wrangling: BIG impact, even for small data

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



Big data

Volume

- Click stream
- Active/passive sensor
- Log
- Event
- Printed *corpus*
- Speech
- Social media
- Traditional

Variety

- Unstructured
- Semi-structured
- Structured

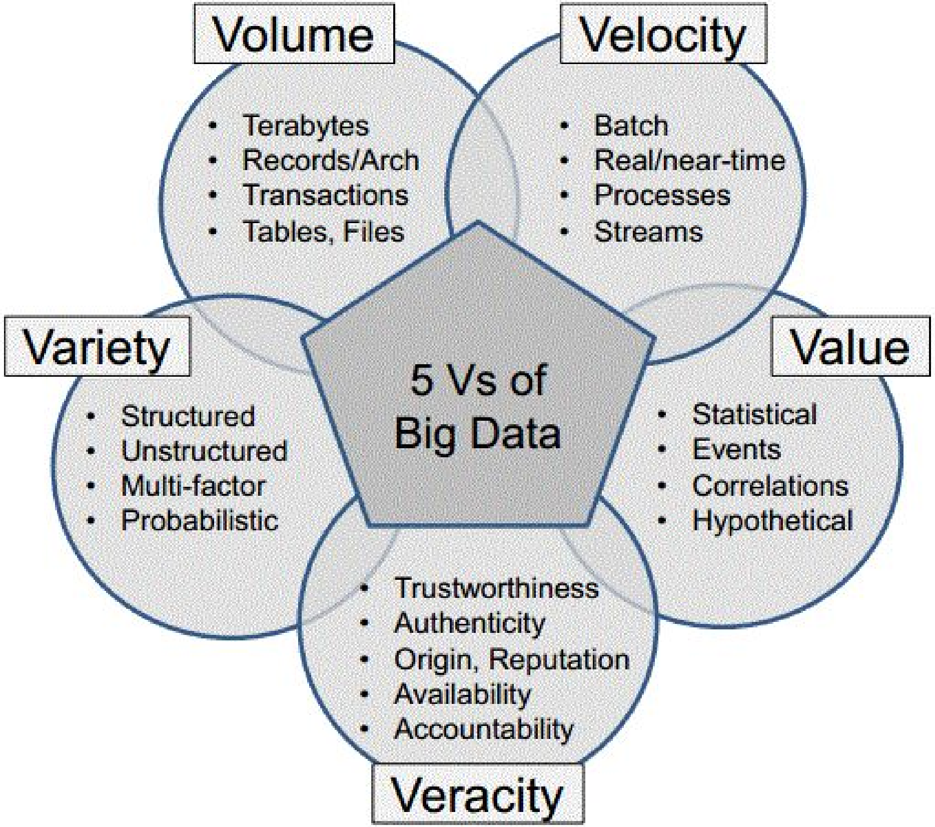
Velocity

- Speed of generation
- Rate of analysis

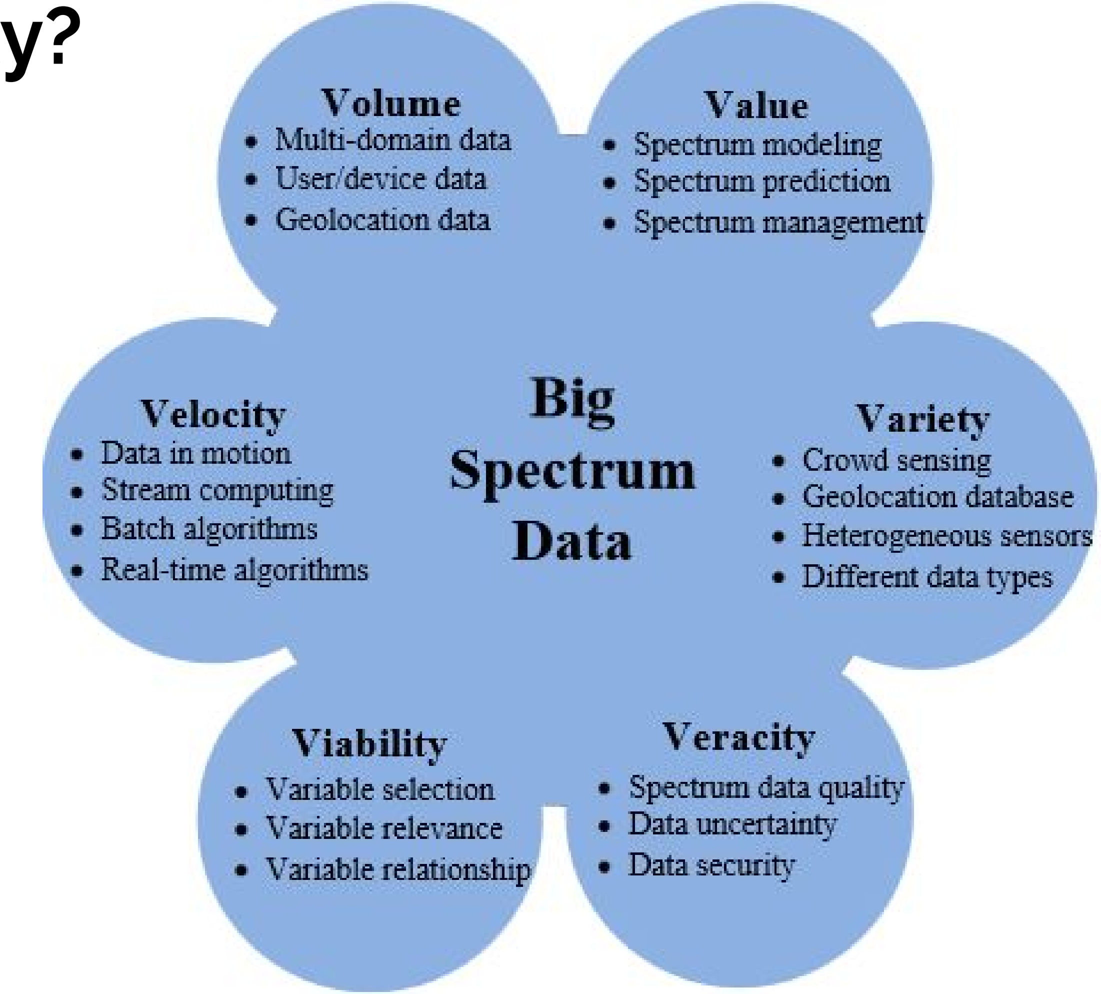
Veracity

- Untrusted
- Uncleansed

...or 5?


Volume
- Terabytes
- Records/Arch
- Transactions
- Tables, Files
Velocity
- Batch
- Real/near-time
- Processes
- Streams
Variety
- Structured
- Unstructured
- Multi-factor
- Probabilistic
5 Vs of Big Data
Value
- Statistical
- Events
- Correlations
- Hypothetical
- Trustworthiness
- Authenticity
- Origin, Reputation
- Availability
- Accountability
Veracity

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

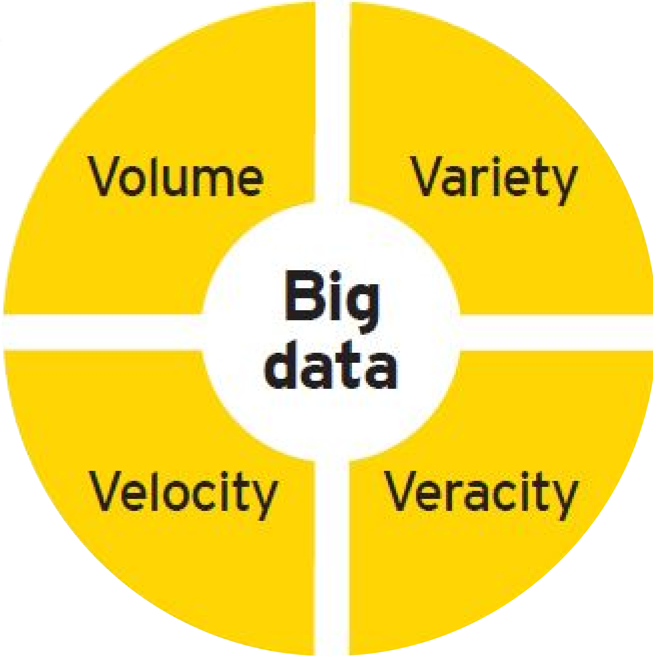
Veracity
- Spectrum data quality
- Data uncertainty
- Data security

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)

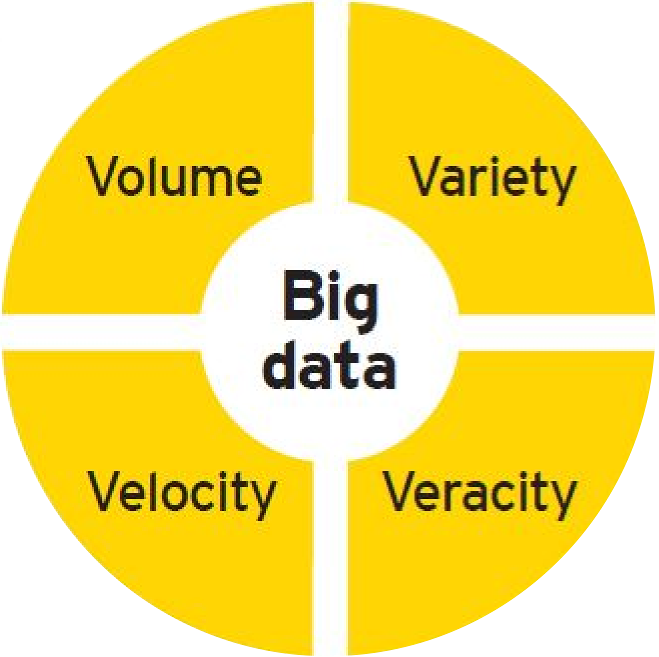
Or it might not be...
All of these factors influence your data wrangling needs and potential solutions!



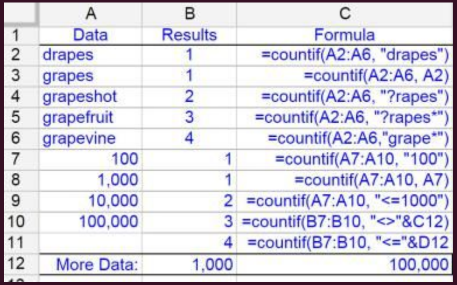
What else matters?

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

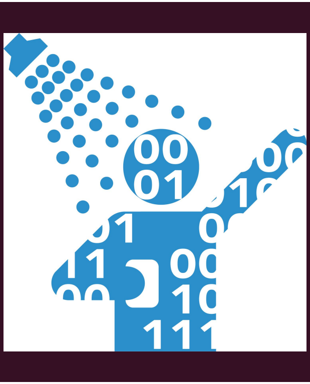
- What tools are you using?
- Are automated approaches available?
- What formats are required for later analyses?
- How tolerant are you to errors?



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

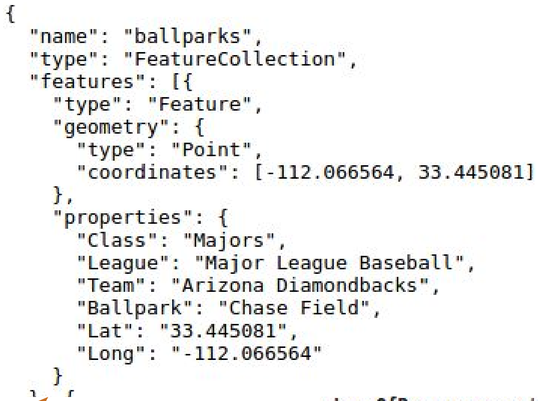


- ➢ Downloading Files
- ➢ API access
- ➢ Web-scraping

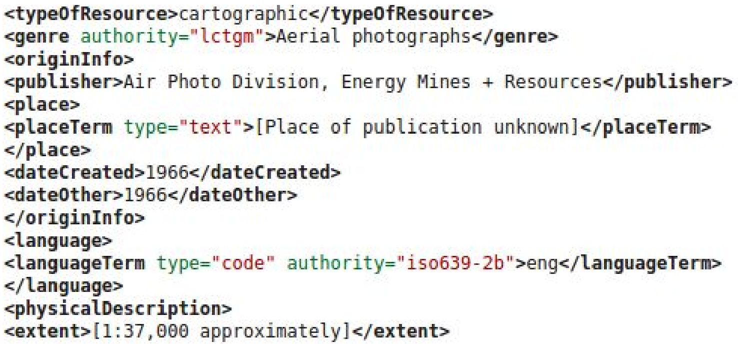
Bulk Downloads

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, SAS, xls)



```
{
  "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: https://www.reddit.com/dev/api

New York Times API: http://developer.nytimes.com/docs

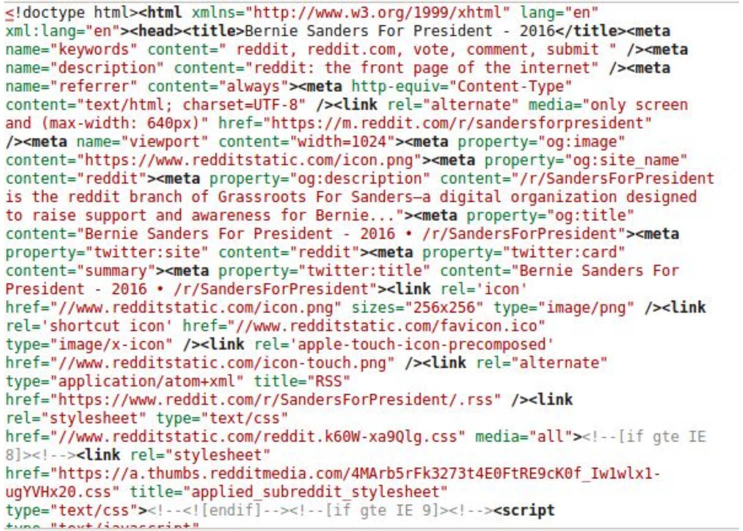
Web Scraping

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 →



```
<!doctype html><html xmlns="http://www.w3.org/1999/xhtml" lang="en"
xml:lang="en"><head><title>Bernie Sanders For President - 2016</title><meta
name="keywords" content=" reddit, reddit.com, vote, comment, submit " /><meta
name="description" content="reddit: the front page of the internet" /><meta
name="referrer" content="always"><meta http-equiv="Content-Type"
content="text/html; charset=UTF-8" /><link rel="alternate" media="only screen
and (max-width: 640px)" href="https://m.reddit.com/r/sandersforpresident"
/><meta name="viewport" content="width=1024"><meta property="og:image"
content="https://www.redditstatic.com/icon.png"><meta property="og:site_name"
content="reddit"><meta property="og:description" content="/r/SandersForPresident
is the reddit branch of Grassroots For Sanders—a digital organization designed
to raise support and awareness for Bernie..."><meta property="og:title"
content="Bernie Sanders For President - 2016 • /r/SandersForPresident"><meta
property="twitter:site" content="reddit"><meta property="twitter:card"
content="summary"><meta property="twitter:title" content="Bernie Sanders For
President - 2016 • /r/SandersForPresident"><link rel='icon'
href="//www.redditstatic.com/icon.png" sizes="256x256" type="image/png" /><link
rel='shortcut icon' href="//www.redditstatic.com/favicon.ico"
type="image/x-icon" /><link rel='apple-touch-icon-precomposed'
href="//www.redditstatic.com/icon-touch.png" /><link rel="alternate"
type="application/atom+xml" title="RSS"
href="https://www.reddit.com/r/SandersForPresident/.rss" /><link
rel="stylesheet" type="text/css"
href="//www.redditstatic.com/reddit.k60W-xa9Qlg.css" media="all"><!--[if gte IE
8]><!--><link rel="stylesheet"
href="https://a.thumbs.redditmedia.com/4MArb5rFk3273t4E0FtRE9cK0f_Iw1wlx1-
ugYVHx20.css" title="applied_subreddit_stylesheet"
type="text/css"><!--<![endif]--><!--[if gte IE 9]><!--><script
type="text/javascript"
```

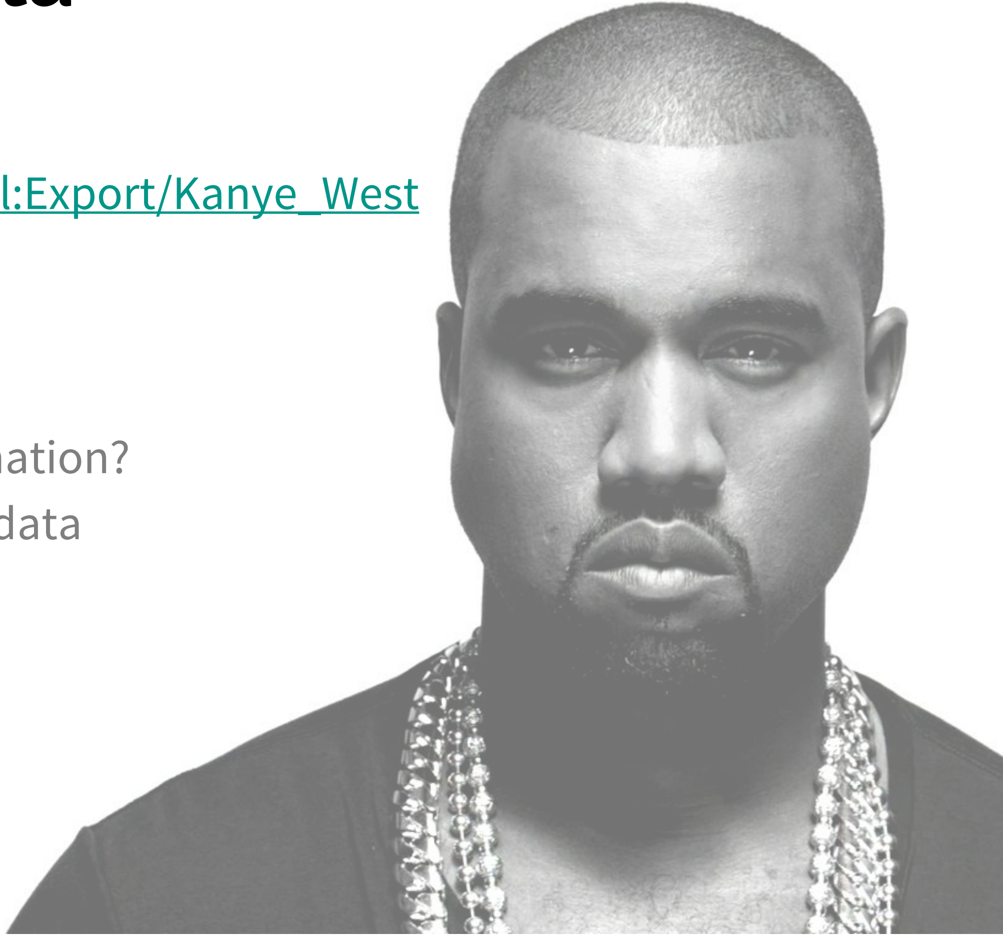
Activity: Assessing Data

Wikipedia page exported to XML:

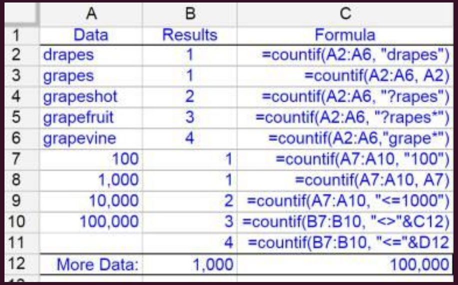
- https://en.wikipedia.org/wiki/Special:Export/Kanye_West
- https://goo.gl/pCDg30

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



	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:

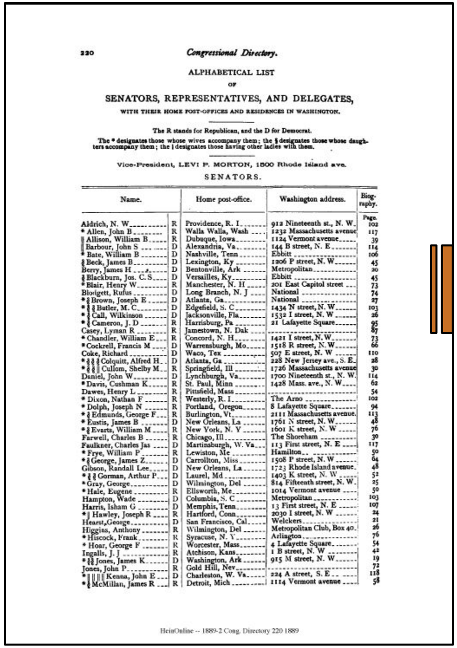
- ➢ Detecting and remediating corrupt/inaccurate records
- ➢ Removing typographical errors
- ➢ Validating against a known list of possibilities (e.g. verify a string as a postal code)
- ➢ Eliminating duplicate entries
- ➢ Harmonizing and standardizing (e.g. represent ‘St’, ‘St.’, ‘Street’ as ‘street’)

Cleaning may be carried out

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

Approaches For Cleaning & Transformation

- Spreadsheets are often a good place to start
  - http://schoolofdata.org/courses/#IntroDataCleaning
- Text Editors - Find and replace
- Command line - regular expressions (http://pythex.org/; http://www.regular-expressions.info/)
- Customized applications
  - Stanford data wrangler: http://vis.stanford.edu/wrangler/
  - Trifacta: https://www.trifacta.com/
  - Open (Google) Refine: http://openrefine.org/
- Fully-automated scripts



220 *Congressional Directory.*

ALPHABETICAL LIST
OF
SENATORS, REPRESENTATIVES, AND DELEGATES,
WITH THEIR HOME POST-OFFICES AND RESIDENCES IN WASHINGTON.

The R stands for Republican, and the D for Democrat.

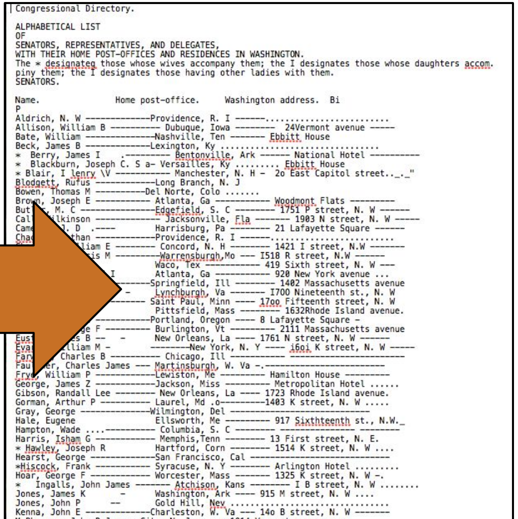
The * designates those whose wives accompany them; the § designates those whose daughters accompany them; the ‖ designates those having other ladies with them.

Vice-President, LEVI P. MORTON, 1800 Rhode Island ave.

SENATORS.

Name.		Home post-office.	Washington address.	Biography.
				Page.
Aldrich, N. W	R	Providence, R. I	912 Nineteenth st., N. W	102
* Allen, John B	R	Walla Walla, Wash	1232 Massachusetts avenue	117
‖ 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 B	D	Lexington, Ky	1206 P street, N. W	45
Berry, James H	D	Bentonville, Ark	Metropolitan	20
§ Blackburn, Jos. C. S	D	Versailles, Ky	Ebbitt	45
* Blair, Henry W	R	Manchester, N. H	201 East Capitol street	73
Blodgett, Rufus	D	Long Branch, N. J	National	74
*§ Brown, Joseph E	D	Atlanta, Ga	National	27
*§§ Butler, M. C	D	Edgefield, S. C	1434 N street, N. W	103
*§ 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	R	Concord, N. H	1421 I street, N. W	73
* Cockrell, Francis M	D	Warrensburgh, Mo	1518 R street, N. W	66
Coke, Richard	D	Waco, Tex	507 E street, N. W	110
*§§§ Colquitt, 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	D	Lynchburgh, Va	1700 Nineteenth st., N. W.	114
* Davis, Cushman K	R	St. Paul, Minn	1428 Mass. ave., N. W	62
Dawes, Henry L	R	Pittsfield, Mass		54
* Dixon, Nathan F	R	Westerly, R. I	The Arno	102
* Dolph, Joseph N	R	Portland, Oregon	8 Lafayette Square	94
*§ Edmunds, George F	R	Burlington, Vt	2111 Massachusetts avenue.	113
* Eustis, James B	D	New Orleans, La	1761 N street, N. W	48
*§ Evarts, William M	R	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	R	Lewiston, Me	Hamilton	50
*§ George, James Z	D	Carrollton, Miss	1508 P street, N. W	64
Gibson, Randall Lee	D	New Orleans, La	1723 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	R	Ellsworth, Me	1014 Vermont avenue	50
Hampton, Wade	D	Columbia, S. C	Metropolitan	103
Harris, Isham G	D	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	Metropolitan Club, Box 40.	26
* Hiscock, Frank	R	Syracuse, N. Y	Arlington	76
* Hoar, George F	R	Worcester, Mass	4 Lafayette Square	54
Ingalls, J. J	R	Atchison, Kans	1 B street, N. W	42
*§§ Jones, James K	D	Washington, Ark	915 M street, N. W	19
Jones, John P	R	Gold Hill, Nev		72
*‖‖‖ Kenna, John E	D	Charleston, W. Va	224 A street, S. E	118
*§ McMillan, James R	R	Detroit, Mich	1114 Vermont avenue	58

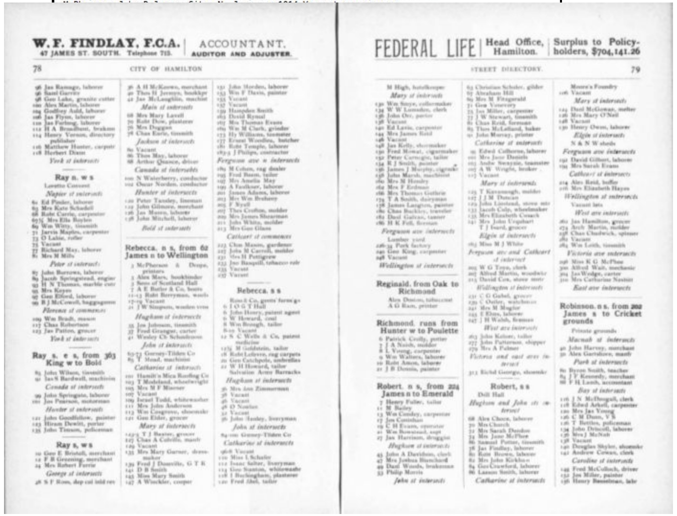
HeinOnline -- 1889-2 Cong. Directory 220 1889



```
| Congressional Directory.

ALPHABETICAL LIST
OF
SENATORS, REPRESENTATIVES, AND DELEGATES,
WITH THEIR HOME POST-OFFICES AND RESIDENCES IN WASHINGTON.
The * designateq 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 ---------
But   , M. C ---------------Edgefield, S. C -------- 1751 P street, N. W ------
Call  ilkinson ------------- Jacksonville, Fla ------- 1903 N street, N. W -----
Came  . D  .----        Harrisburg, Pa -------- 21 Lafayette Square ------
Chac     than ------------Providence, R. I ------.........................
          liam 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
Eust     es B --   -      New Orleans, La ---- 1761 N street, N. W ------
Evar     lliam M -        --------New York, N. Y ---- i6oi K street, N. W -----
Farw  , Charles B ---------- Chicago, Ill ----------- ----------------------
Faul  er, 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 -------
```



W. F. FINDLAY, F.C.A. | ACCOUNTANT.
47 JAMES ST. SOUTH. Telephone 715. | AUDITOR AND ADJUSTER.

78 CITY OF HAMILTON

FEDERAL LIFE | Head Office, Hamilton. | Surplus to Policy-holders, $704,141.26

STREET DIRECTORY. 79

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

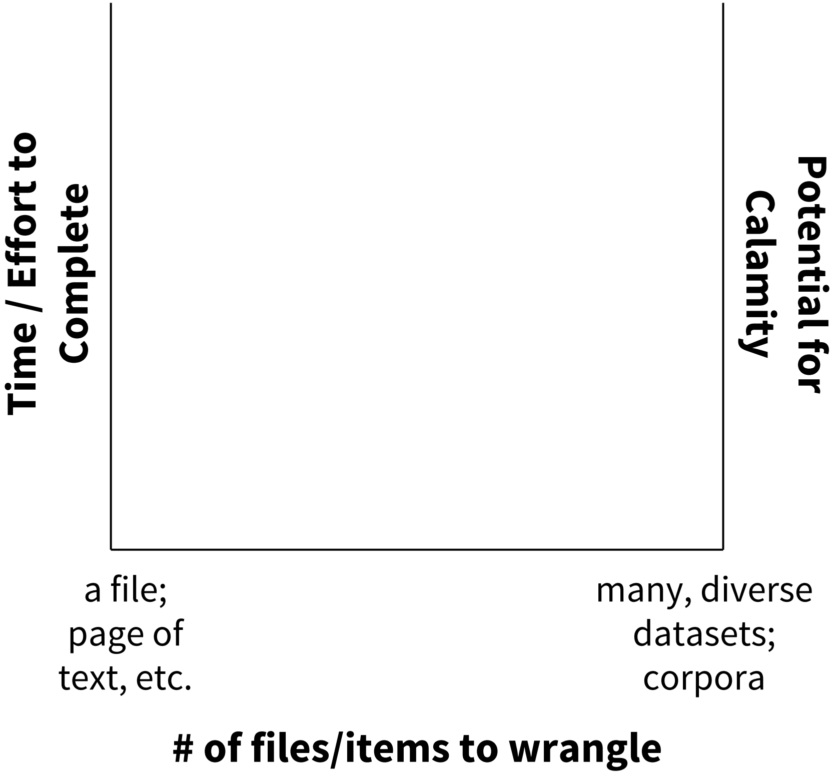
Yoik 11 intersect(

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

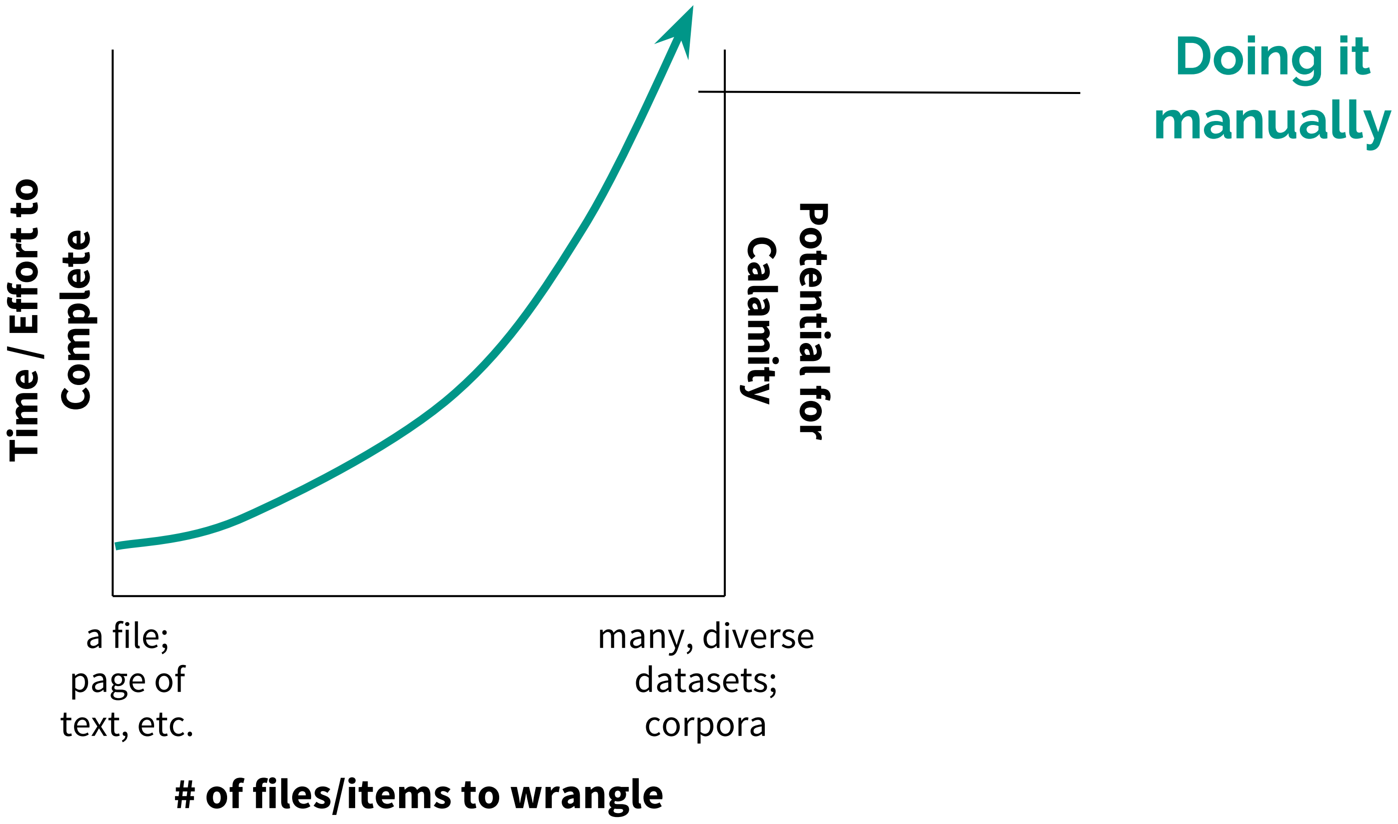
Main st Intersects
68 Mrs Mary Liven

So, when to let the computer take over?

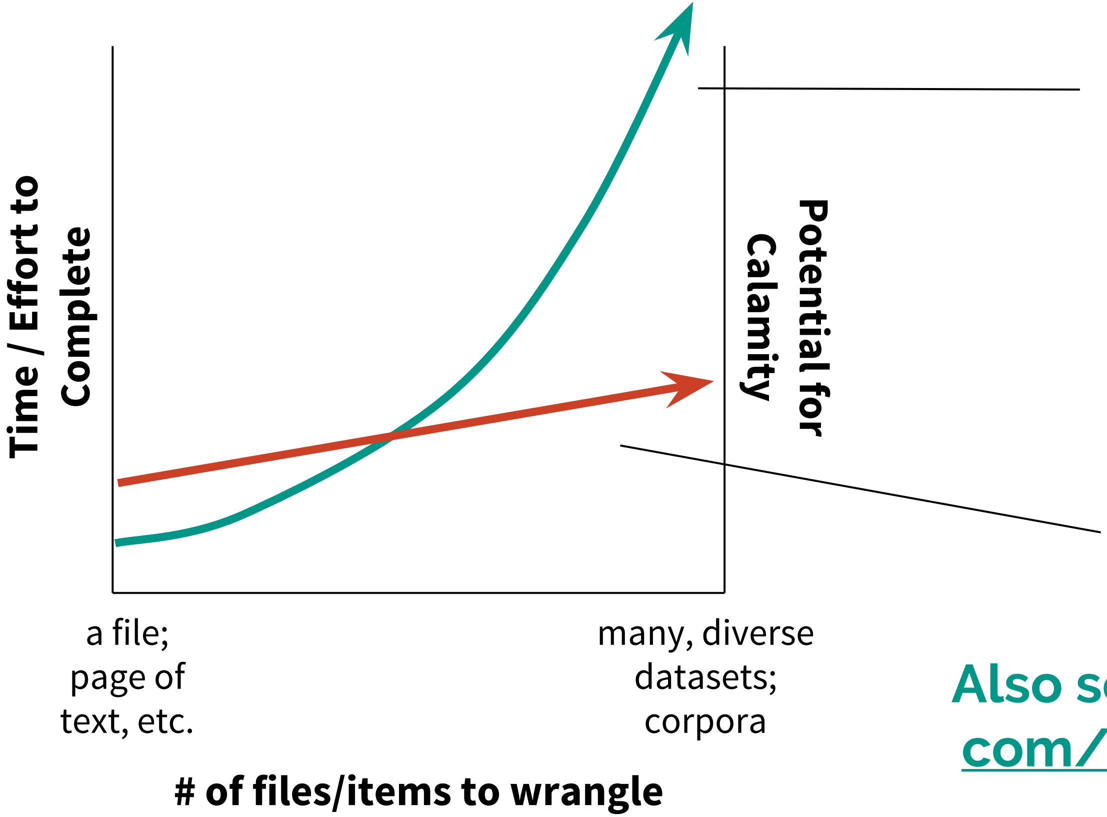
Whenever it works and will save you time!



Spreadsheets: The frenemy of research



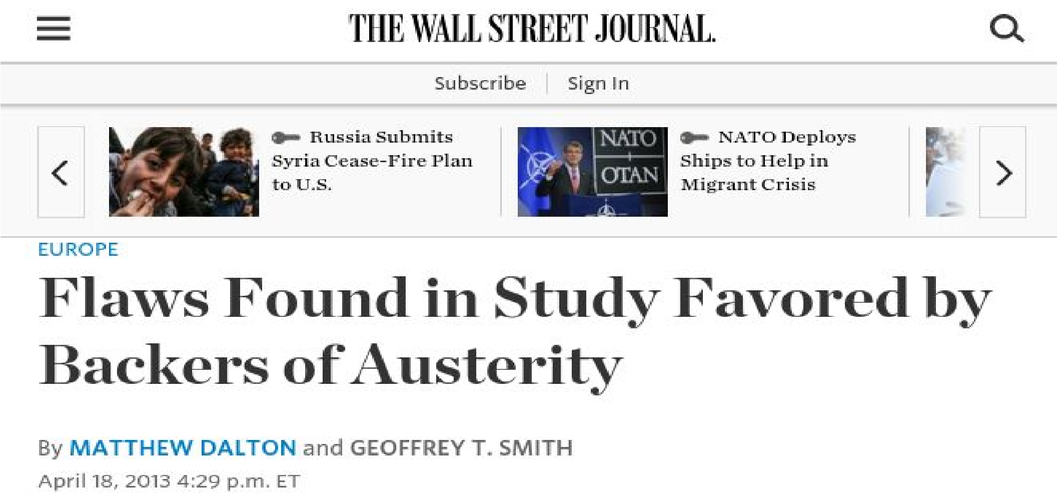
Spreadsheets: The frenemy of research



Doing it manually

Doing it automatically

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

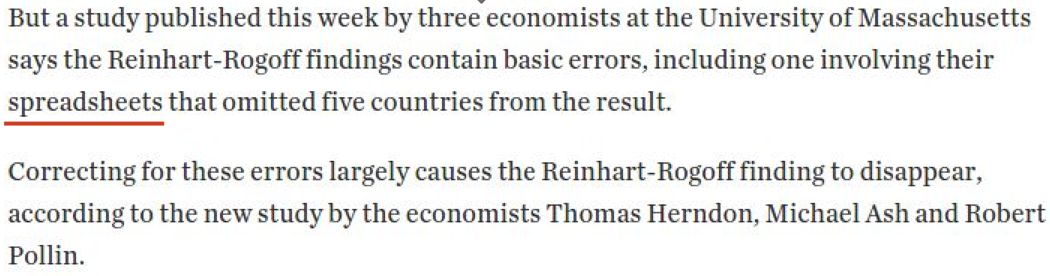

THE WALL STREET JOURNAL.
Subscribe
Sign In
Russia Submits Syria Cease-Fire Plan to U.S.
NATO Deploys Ships to Help in Migrant Crisis
NATO OTAN

EUROPE

Flaws Found in Study Favored by Backers of Austerity

By MATTHEW DALTON and GEOFFREY T. SMITH

April 18, 2013 4:29 p.m. ET

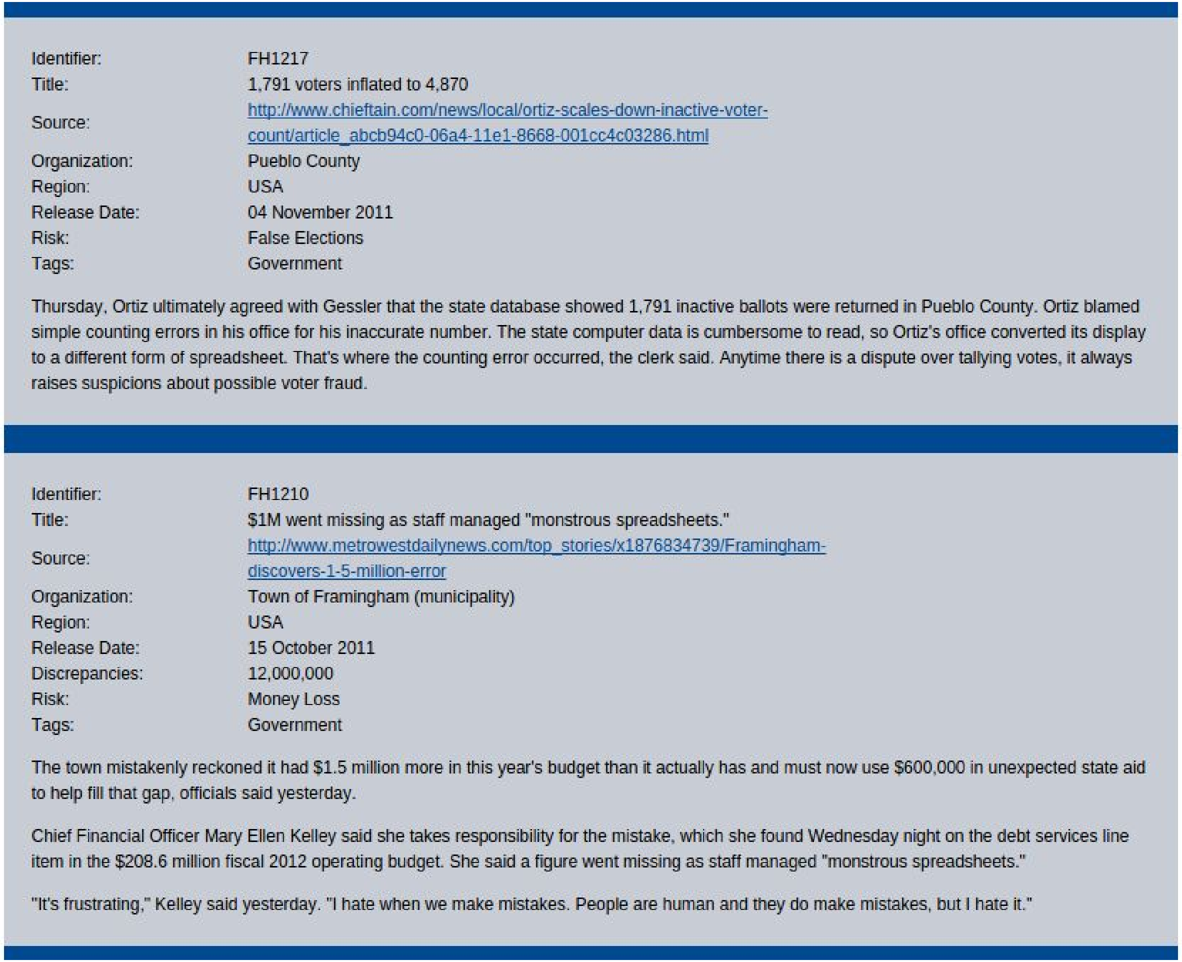


But a study published this week by three economists at the University of Massachusetts says the Reinhart-Rogoff findings contain basic errors, including one involving their spreadsheets that omitted five countries from the result.

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.

EuSpRiG

European Spreadsheet
Risks Interest Group



Identifier:	FH1217
Title:	1,791 voters inflated to 4,870
Source:	http://www.chieftain.com/news/local/ortiz-scales-down-inactive-voter-count/article_abcb94c0-06a4-11e1-8668-001cc4c03286.html
Organization:	Pueblo County
Region:	USA
Release Date:	04 November 2011
Risk:	False Elections
Tags:	Government

Thursday, Ortiz ultimately agreed with Gessler 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.

Identifier:	FH1210
Title:	$1M went missing as staff managed "monstrous spreadsheets."
Source:	http://www.metrowestdailynews.com/top_stories/x1876834739/Framingham-discovers-1-5-million-error
Organization:	Town of Framingham (municipality)
Region:	USA
Release Date:	15 October 2011
Discrepancies:	12,000,000
Risk:	Money Loss
Tags:	Government

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."

"It's frustrating," Kelley said yesterday. "I hate when we make mistakes. People are human and they do make mistakes, but I hate it."

http://www.eusprig.org/horror-stories.htm

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-xxxxxx) - 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.

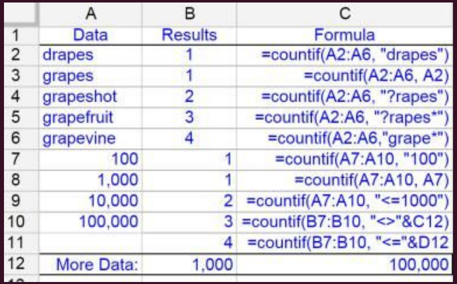
- How will you analyze the data?
- In what ways will you need to clean | transform the data?
- What tools might work for this?

Trump:
#Trump2016
https://www.reddit.com/r/the_donald

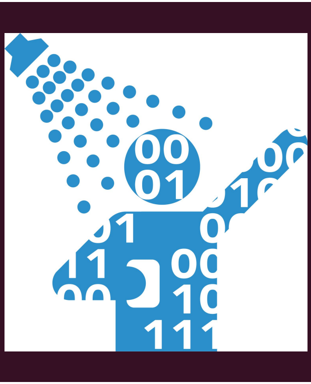
Bernie:
#FeelTheBern
https://www.reddit.com/r/sandersforpresident

Hillary:
#ImWithHer
https://www.reddit.com/r/hillaryclinton/

Depositing for Use & Sharing



	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



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

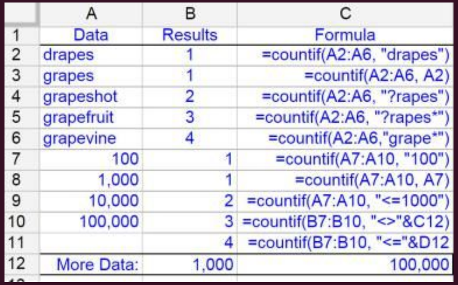
Depositing for Use & Sharing

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

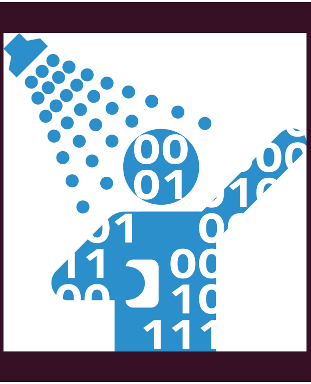
- ➢ be better able to show (demonstrate) it to people?
- ➢ make it work with different tools?
  - → Keep data is translatable formats (or formats that work well with visualization)
  - → Explore formatting requirements/compatibility for other community tools

- ➢ 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



	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