

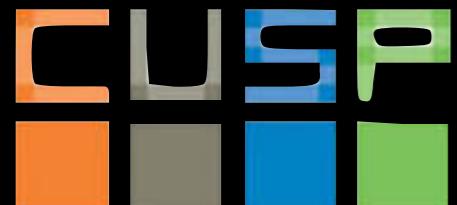
Urban Informatics

Fall 2015

dr. federica bianco fb55@nyu.edu

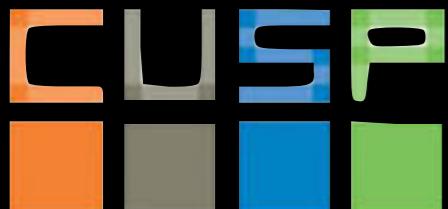


@fedhere



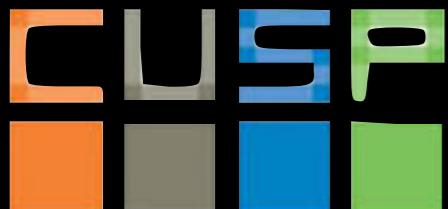
Summary:

- **Epistemological concepts:**
falsifiability, law of parsimony,
- **Good scientific practice:**
reproducibility of research



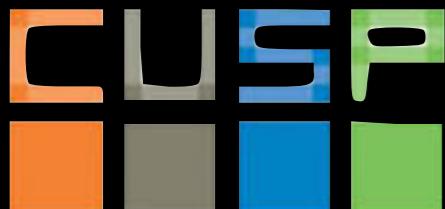
Summary:

- **Epistemological concepts:**
falsifiability, law of parsimony,
- **Good scientific practice:**
reproducibility of research
- **Gathering parsing data, API:**
data munging or wrangling, data jujitsu



Summary:

- **Epistemological concepts:**
falsifiability, law of parsimony,
- **Good scientific practice:**
reproducibility of research
- **Gathering parsing data, API:**
data munging or wrangling, data jujitsu
 - types of data
 - reporting data for reproducibility
 - cleaning and tabulating data
 - reading data from CSV files
 - reading data from JSON files
 - reading data from the CUSP dataHub
 - reading data from APIs



- IDEA
- dataset
 - define ideal data
 - figure out best data available
 - figure out if you can get new data
 - obtain data (including policy issues + technical issues)
- data handling
 - joining databases
 - formatting data
- exploratory data analysis
 - machine learning (clustering? dimensionality reduction?)
- statistics
 - models (regression)
 - prediction
 - validation (simulations)
- interpretation
- presentation
 - visualization
 - write a paper!



- IDEA
- dataset

PROBLEM IDENTIFICATION

FORMULATING HYPOTHESIS

DATA IDENTIFICATION

- figure out best data available
- figure out if you can get new data

- data handling
 - joining databases
 - formatting data
- exploratory data analysis
 - machine learning (clustering? dimensionality reduction?)
- statistics
 - models (regression)
 - prediction
 - validation (simulations)
- interpretation
- presentation
 - visualization
 - write a paper!



- IDEA
- dataset

PROBLEM IDENTIFICATION

FORMULATING HYPOTHESIS

DATA IDENTIFICATION

- define ideal data
- figure out best data available
- figure out if you can get new data

- obtain data (including policy issues + technical issues)

- data handling

- joining databases
- formatting data

DATA PREPARATION

- exploratory data analysis

- machine learning (clustering? dimensionality reduction?)

- statistics

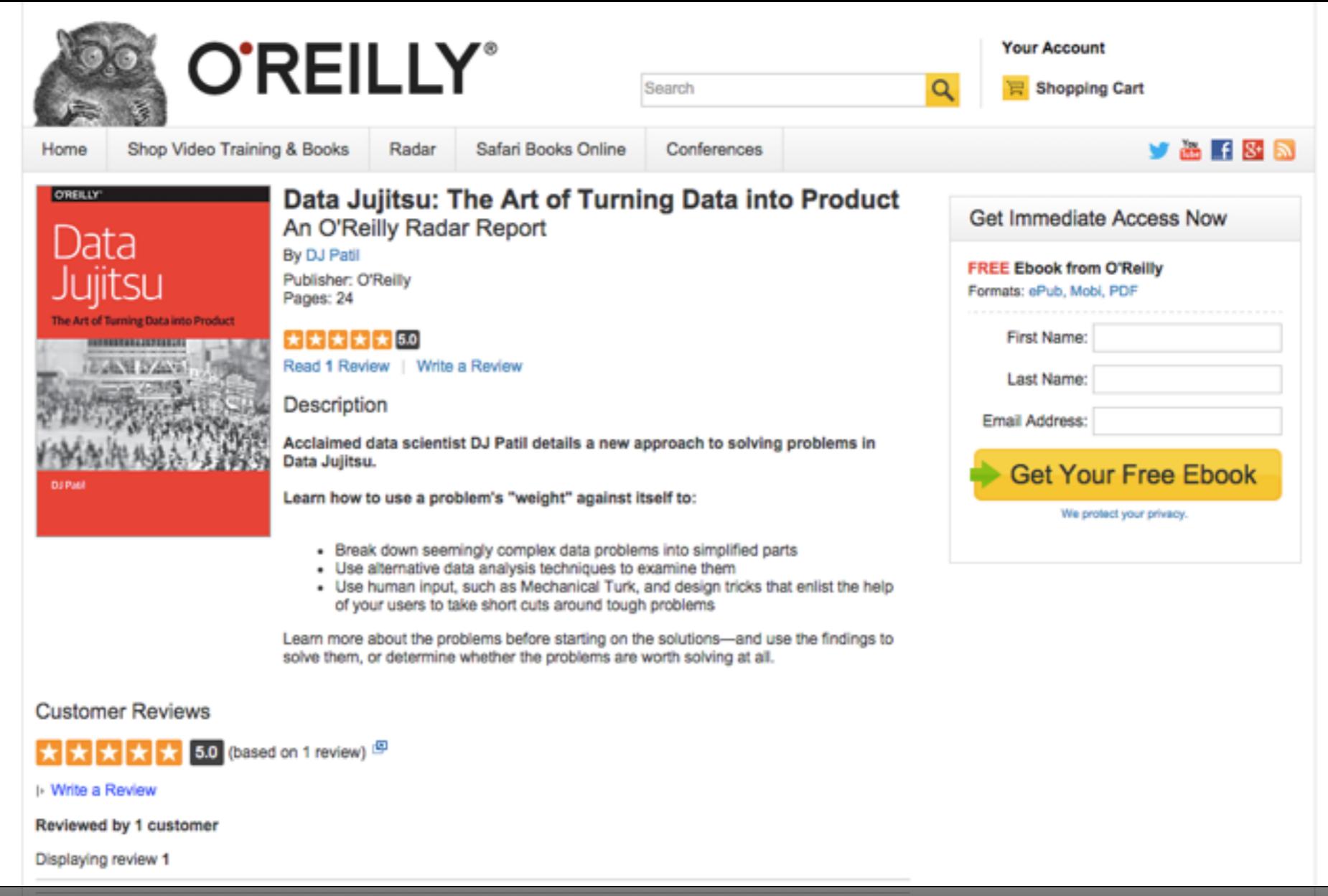
- models (regression)
- prediction
- validation (simulations)

- interpretation

- presentation

- visualization
- write a paper!





The screenshot shows the O'Reilly website product page for "Data Jujitsu: The Art of Turning Data into Product".

Header: Features the O'Reilly logo with a cartoon owl icon, a search bar, and links for "Your Account" and "Shopping Cart". Social media icons for Twitter, YouTube, Facebook, Google+, and RSS are also present.

Product Summary: The title is "Data Jujitsu: The Art of Turning Data into Product" by DJ Patil. It's described as an "O'Reilly Radar Report". The publisher is O'Reilly, and it has 24 pages. A 5.0 star rating is shown with 1 review available.

Description: Acclaimed data scientist DJ Patil details a new approach to solving problems in Data Jujitsu. It includes a section on how to use a problem's "weight" against itself.

Key Features:

- Break down seemingly complex data problems into simplified parts
- Use alternative data analysis techniques to examine them
- Use human input, such as Mechanical Turk, and design tricks that enlist the help of your users to take short cuts around tough problems

Customer Reviews: 5.0 (based on 1 review). A link to "Write a Review" is provided.

Reviewed by 1 customer: Displaying review 1.

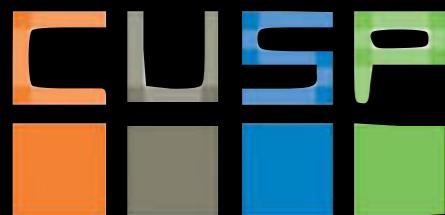
Review Content: The review is from DJ Patil, dated 10/10/2012, and reads: "I really enjoyed reading this book. It's a great introduction to the field of data science and provides practical advice for solving real-world problems." The review has 20 likes.

Call-to-Action: A box on the right offers a "FREE Ebook from O'Reilly" in ePUB, Mobi, and PDF formats. It includes fields for First Name, Last Name, and Email Address, and a large yellow "Get Your Free Ebook" button.

Footer: The CUSP logo is visible at the bottom left.

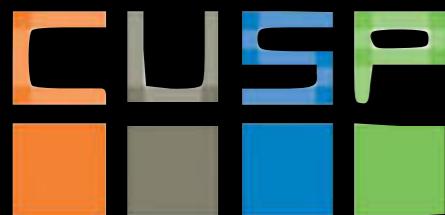
[...]data products are unique in that they are often extremely difficult, and seemingly intractable for small teams with limited funds. Yet, they get solved every day.

How? Are the people who solve them superhuman data scientists who can come up with better ideas in five minutes than most people can in a lifetime? Are they magicians of applied math who can cobble together millions of lines of code for high-performance machine learning in a few hours? No.
(continue in the next page...)



(...continued)

Many of them are incredibly smart, but meeting big problems head-on usually isn't the winning approach. There's a method to solving data problems that avoids the big, heavyweight solution, and instead, concentrates on building something quickly and iterating. Smart data scientists don't just solve big, hard problems; they also have an instinct for making big problems small.

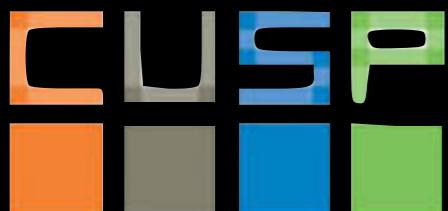


Reproducible research means:

code raw data

privacy concerns

in order for your research to be reproducible
the data you use must be accessible BUT
THAT IS NOT ALWAYS POSSIBLE:
CUSP has access to data that has restricted
access. Share your data when possible!



Reproducible research means:

code raw data

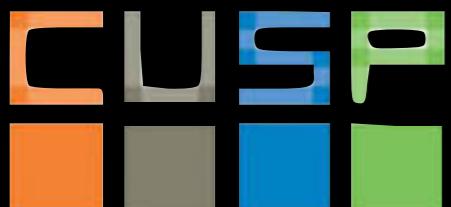
Remove sensitive data

Some day you or a collaborator may accidentally commit sensitive data, such as a password or SSH key, into a Git repository. Although you can remove the file from the latest commit with `git rm`, the file will still exist in the repository's history. Fortunately, there are other tools that can entirely remove unwanted files from a repository's history. This article will explain how to use two of them: `git filter-branch` and the [BFG Repo-Cleaner](#).

Danger: Once you have pushed a commit to GitHub, you should consider any data it contains to be compromised. If you committed a password, change it! If you committed a key, generate a new one.

This article tells you how to make commits with sensitive data unreachable from any branches or tags in your GitHub repository. However, it's important to note that those commits may still be accessible in any clones or forks of your repository, directly via their SHA-1 hashes in cached views on GitHub, and through any pull requests that reference them. You can't do anything about existing clones or forks of your repository, but you can permanently remove all of your repository's cached views and pull requests on GitHub by contacting [GitHub support](#).

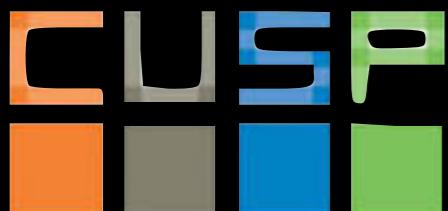
<https://help.github.com/articles/remove-sensitive-data/>



Reproducible research:

How to share your data

- Share/Reference the source of your raw data
- Share the “tidy” data
- Share the code used to process the data at each step

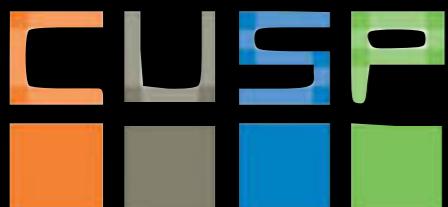


Reproducible research:

How to share your data

- Share/Reference the source of your raw data
- Share the “tidy” data
- Share the code used to process the data at each step

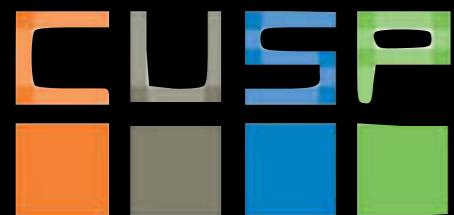
[https://github.com/jakevdp/PythonicPerambulations/blob/master/
content/downloads/notebooks/ProntoData.ipynb](https://github.com/jakevdp/PythonicPerambulations/blob/master/content/downloads/notebooks/ProntoData.ipynb)



Data are messy

...

(and you have to clean the mess!)



WHAT ARE THEIR BIGGEST CHALLENGES?

Dirty data is the #1 hurdle for data scientists.



66.7%

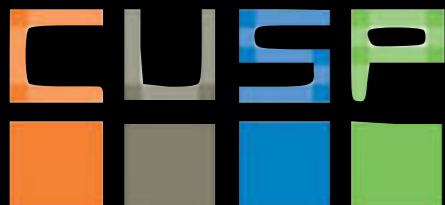
of data scientists say cleaning and organizing data is their most time-consuming task.

52.3%

of data scientists cite poor quality data as their biggest daily obstacle.

2015 survey of data scientists

<https://www.crowdflower.com/the-data-behind-todays-data-scientists-an-infographic/>
II: Data Wrangling



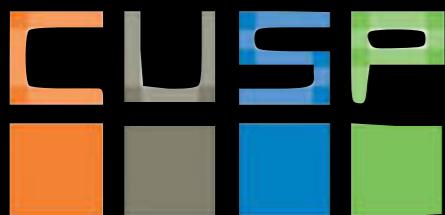
Reproducible research means:

code raw data

Raw data formats: API, Json, binary, paper...

Tidy data: tabulated data, binary, csv, tsv, ascii,
Excel (ugh!) or (object oriented) Json,...

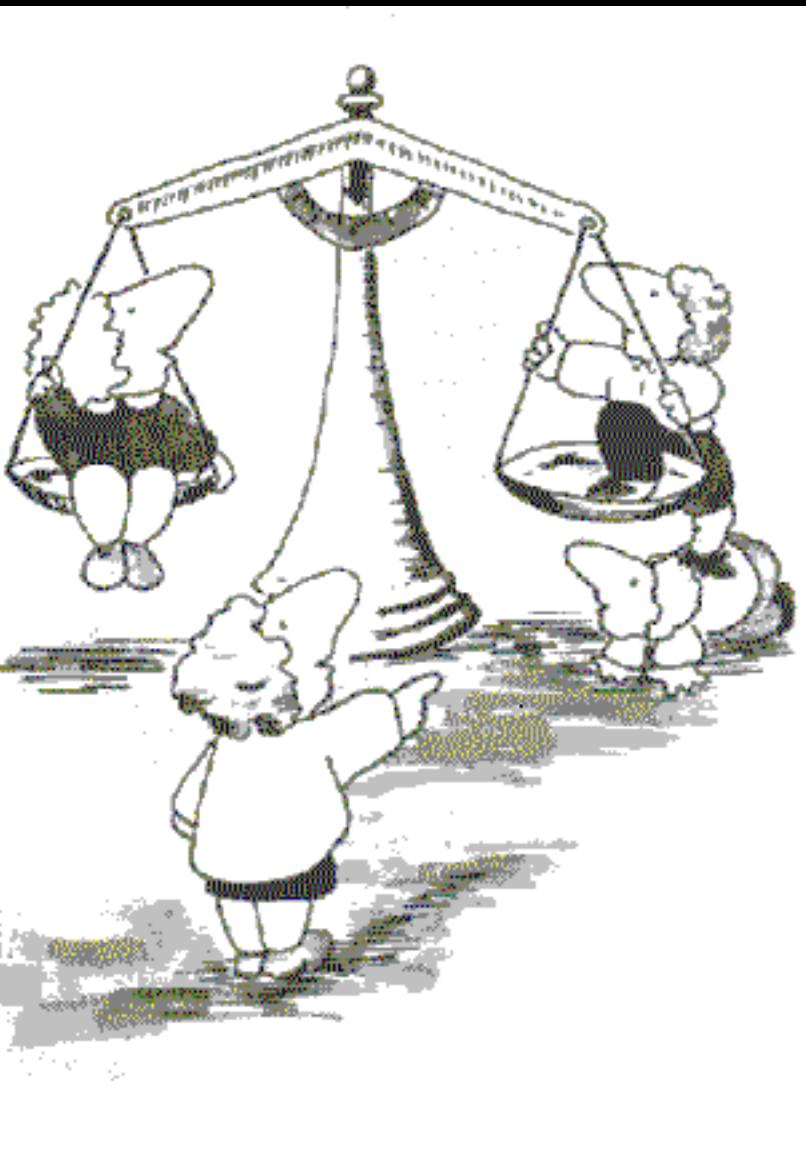
	var 1	var 2
obs 1	7.4	9.2
obs 2	NaN	13.1



Reproducible research means:

Data are noisy

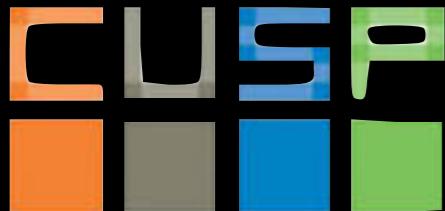
- Data entry errors
- Measurement errors
- Extraction errors



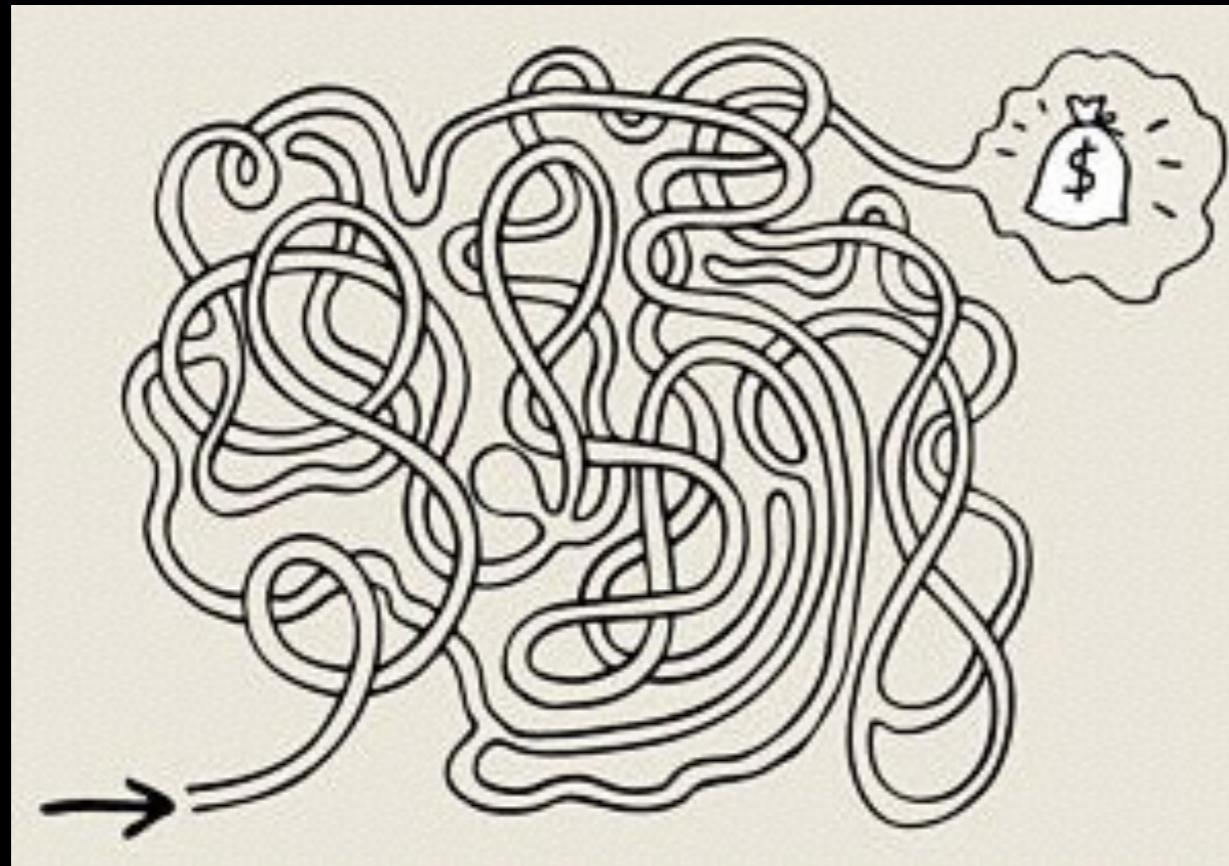
Make sure you do not make them noisier...

Typical data wrangling issues

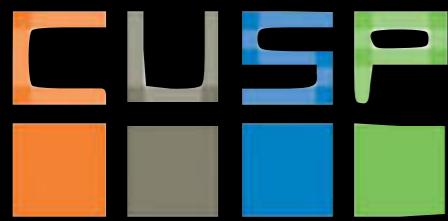
- Licensing issues/Privacy (some data unavailable)
- Missing required field
- Primary key violation (two people with the same social security number)
- Parsing text into fields (separator issues)
- Naming conventions: NYC vs New York
- Different representations (2 vs Two)
- Fields corrupted (too long get truncated)
- Formatting issues – especially dates
- Redundant Records (exact match or other)



• noisy data in



• tidy data out



II: Data Wrangling

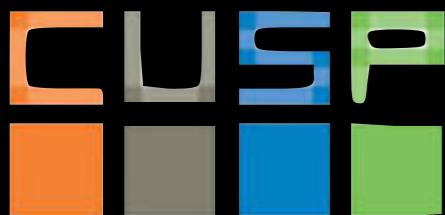
Reproducible research means:

code raw data

Raw data formats: API, Json, binary, paper...

Tidy data: tabulated data, binary, csv, tsv, ascii,
Excel (ugh!) or (object oriented) Json,...

	var 1	var 2
obs 1	7.4	9.2
obs 2	NaN	13.1



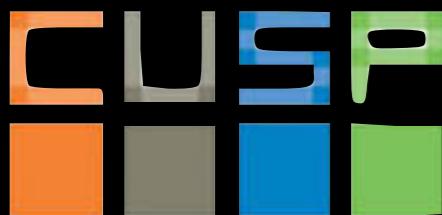
Reproducible research means:

code raw data

Raw data formats: API, Json, binary, paper...

Tidy data: tabulated data, binary, csv, tsv, ascii,
Excel (ugh!) or (object oriented) Json,...

	var 1	var 2
obs 1	7.4	9.2
obs 2	NaN	13.1



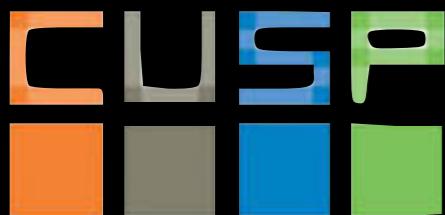
Reproducible research means:

code raw data

Raw data formats: API, Json, binary, paper...

Tidy data: tabulated data, binary, csv, tsv, ascii,
Excel (ugh!) or (object oriented) Json,...

	var 1	var 2	link
obs 1	7.4	9.2	URL 1
obs 2	NaN	13.1	URL 1



Reproducible research means:

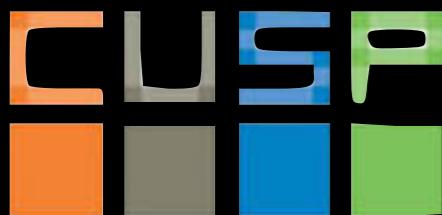
code raw data

Raw data formats: API, Json, binary, paper...

Tidy data: tabulated data, binary, csv, tsv, ascii,
Excel (ugh!) or (object oriented) Json,...

	var 1	var 2	link
obs 1	7.4	9.2	URL 1
obs 2	NaN	13.1	URL 1

never use
moc values
for missing
data!



Reproducible research means:

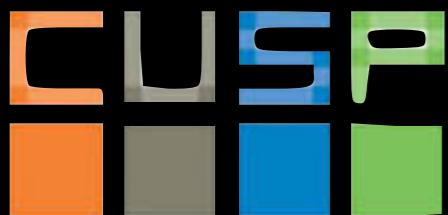
code raw data

Raw data formats: API, Json, binary, paper...

Tidy data: tabulated data, binary, csv, tsv, ascii, Excel (ugh!) or (object oriented) Json,...

Report separately :

- **Details about variables:** e.g. date: “date when test was performed”
- **Units:** remember the Mars Rover that was lost at see cause scientists got Metric and Imperial units mixed up?!?!



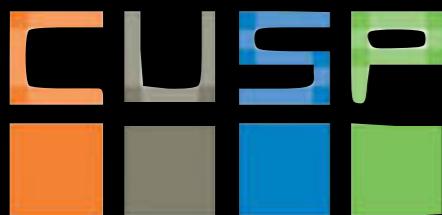
Reproducible research means:

code raw data

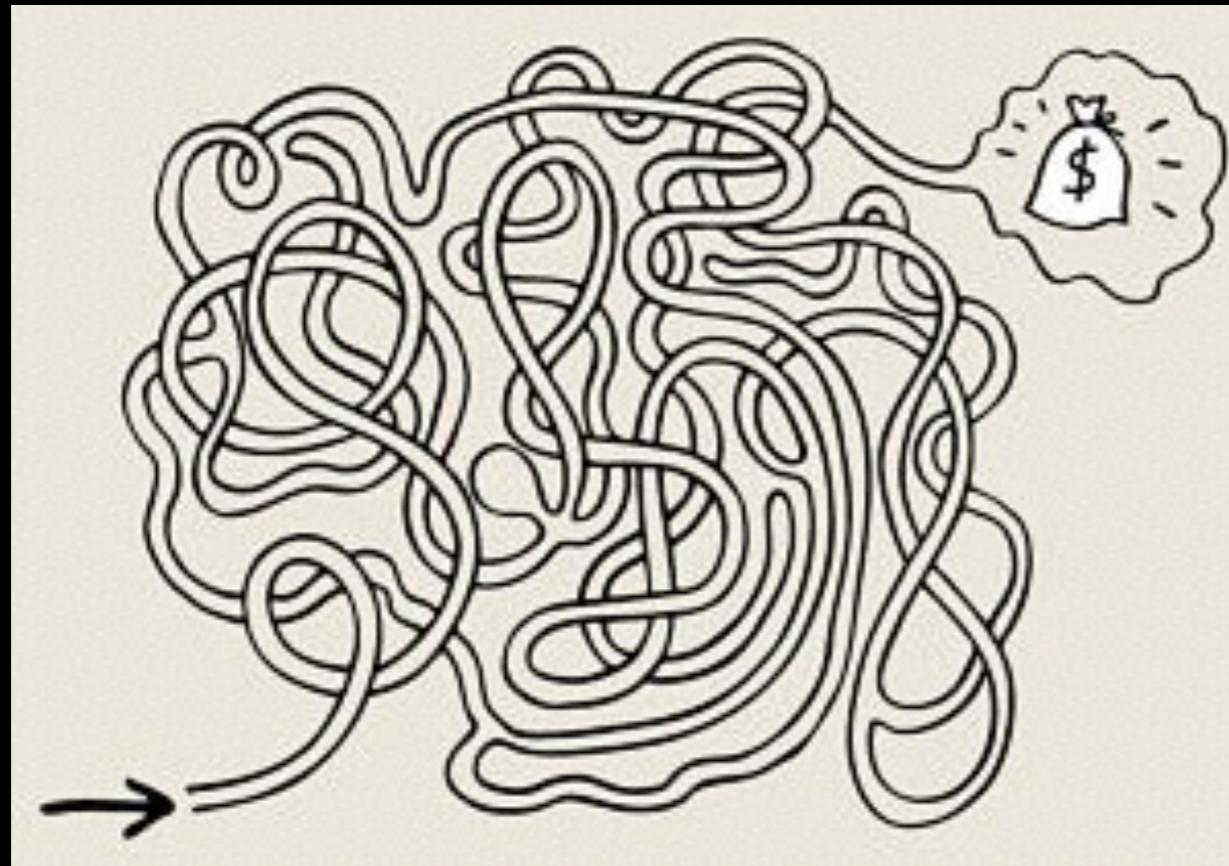
Raw data formats: API, Json, binary, paper...

Tidy data: tabulated data, binary, csv, tsv, ascii, Excel (ugh!) or (object oriented) Json,...

	speed (m/s)	distance (m)	link (accessed date)
obs 1	7.4	9.2	URL 1
obs 2	NaN	13.1	URL 1



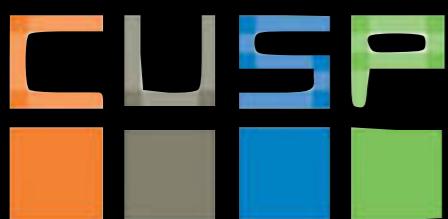
noisy data in



tidy data out

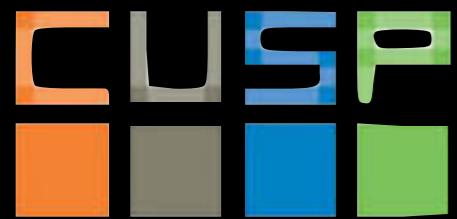


(urban) science!



II: Data Wrangling

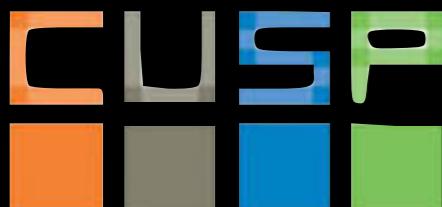
data ingestion



Types of Data Files:

- **CSV** comma separated values (also TSV - tab)
- **JSON:** corresponds to a Python data dictionary
- **XML:** similar to HTML

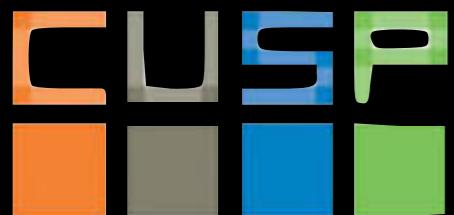
support nested structures





loading data on a computer

[https://github.com/fedhere/UInotebooks/blob/master/
dataWrangling/readingData.ipynb](https://github.com/fedhere/UInotebooks/blob/master/dataWrangling/readingData.ipynb)



II: Data Wrangling



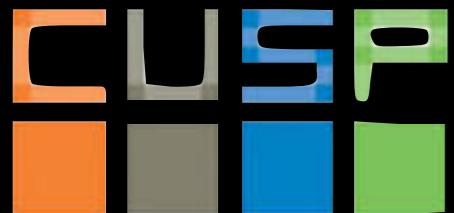
Python for Data Analysis (Pandas manual) Chapter 7

[https://github.com/fedhere/UInotebooks/blob/master/dataWrangling/
PandasDataWrangling-Chap7.ipynb](https://github.com/fedhere/UInotebooks/blob/master/dataWrangling/PandasDataWrangling-Chap7.ipynb)



II: Data Wrangling

data types



Introduction to Statistics

Online Edition

Primary author and editor: David M. Lanel

Introduction to Statistics

Online Edition

Primary author and editor:
David M. Lane¹

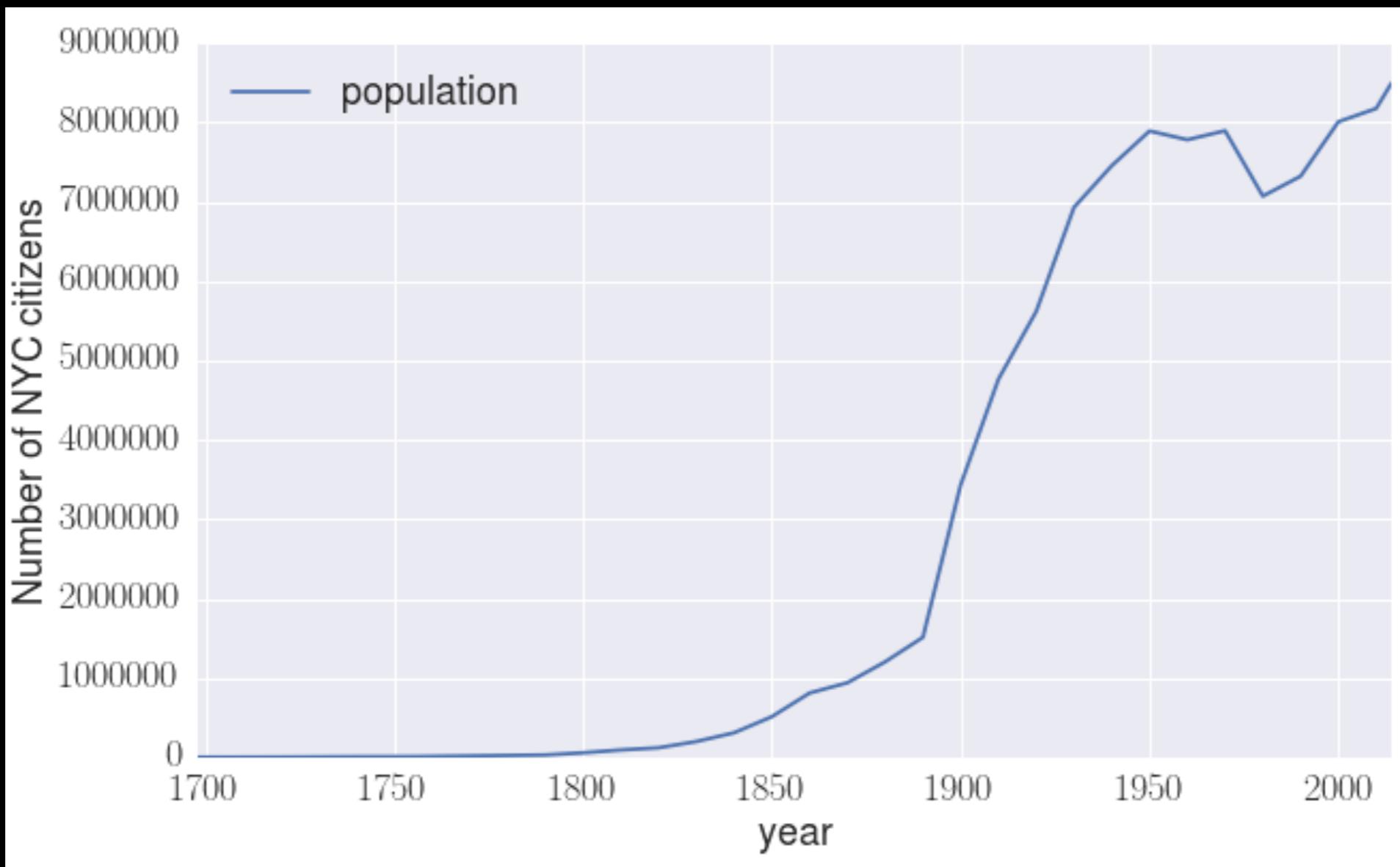
Other authors:

David Scott¹, Mikki Hebl¹, Rudy Guerra¹, Dan Osherson¹, and Heidi Zimmer²

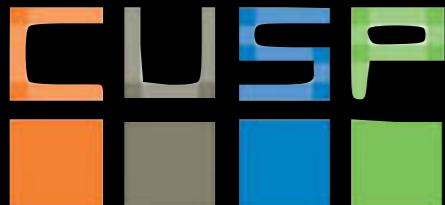
¹Rice University; ²University of Houston, Downtown Campus



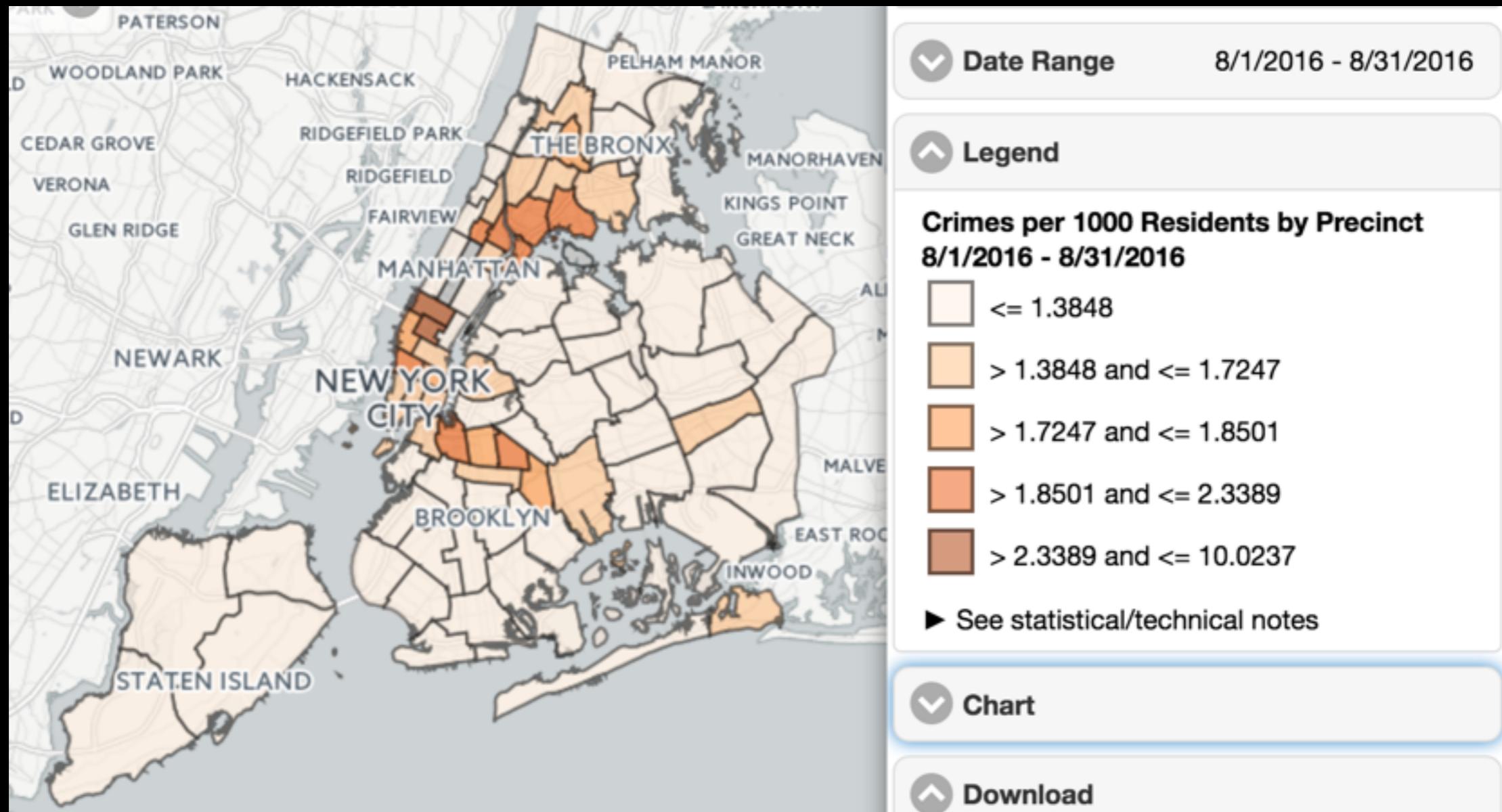
Dependent vs Independent Variable



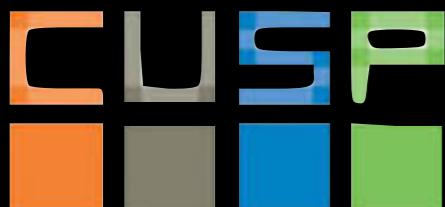
<http://flowingdata.com/2016/09/12/watch-bacteria-evolve-resistance-to-antibiotic/>



Dependent vs Independent Variables



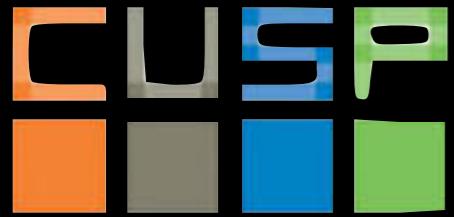
<https://maps.nyc.gov/crime/>



Dependent vs Independent Variable

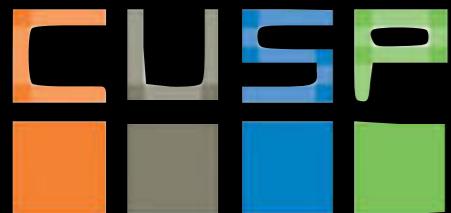


<http://flowingdata.com/2016/09/12/watch-bacteria-evolve-resistance-to-antibiotic/>



Dependent vs Independent Variable

Levels of an independent variable:
the number of experimental conditions.
e.g.: control test and experiment are 2 levels



Types of Data: *Qualitative* variables

- No ordering

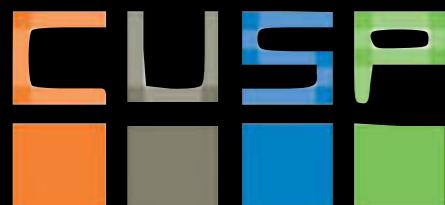
UrbanScience e.g. precinct, state, gender

Also called *Nominal, Categorical*

Types of Data: *Quantitative* variables

- Ordering is meaningful

Time, Distance, Age, Length, Intensity, Satisfaction



Types of Data: *Quantitative* variables

- Continuous: _____
- Discrete: ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ●

Types of Data: *Quantitative* variables

- **Continuous:** distance to the closest park
- **Discrete:** *any* countable, e.g. number of crimes

Discrete data may be:

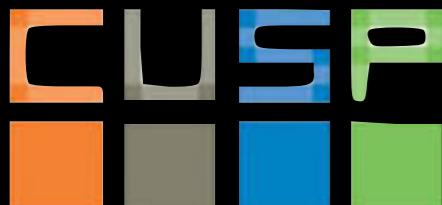
- **Counts:** number of bacteria at time t in section A
- **Ordinal:** survey response Good/Fair/Poor

Continuous data may be:

- **Continuous Ordinal:** Earthquakes (not linear scale)
- **Interval:** F temperature - interval size preserved
- **Ratio:** Car speed - 0 is naturally defined

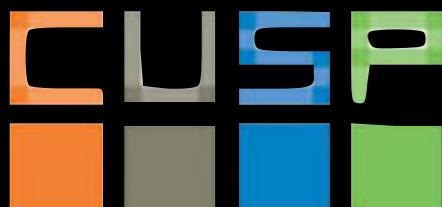
Data may also be:

- **Censored:** age > 90
- **Missing:** “Prefer not to answer” (NA / NaN)



Data Definitions

- **Data:** observations that have been collected
- **Population:** the complete body of subjects we want to infer about
- **Sample:** the subset of the population we actually studied
- **Census:** collection of data from the entire population
- **Parameter:** numerical value describing an attribute of the *population*
- **Statistics:** numerical value describing an attribute of the *sample*



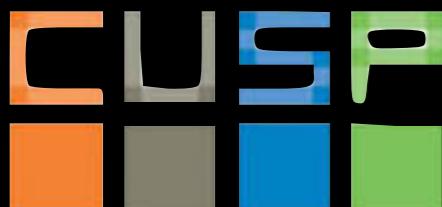
Data Definitions

The analysis of our _____
showed that for our 10 _____ the mean height is 6.4 ft.
The standard deviation of the _____ means is 0.5 ft.
From this _____ we infer for the _____ a mean
height _____ $6.4 +/ - 5$ ft

Data Definitions

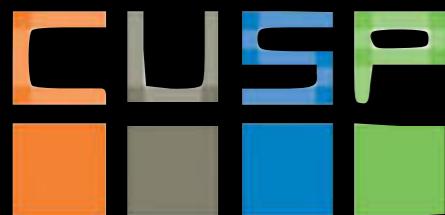
The analysis of our _____ showed that for our 10 _____ the mean height is 6.4 ft. The standard deviation of the _____ means is 0.5 ft. From this _____ we infer for the _____ a mean height _____ 6.4 ± 0.5 ft

sample parameter data population statistics



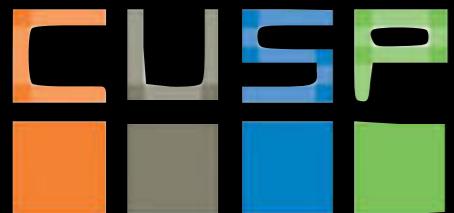
Data Definitions

The analysis of our **data** showed that for our 10 **samples** the mean height is 6.4 ft. The standard deviation of the **sample** means is 0.5 ft. From this **statistic** we infer for the **population** a mean height **parameter** 6.4 ± 0.5 ft





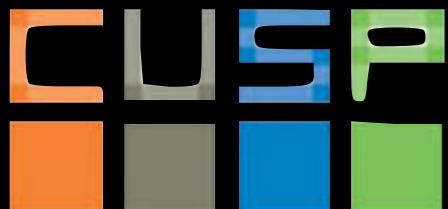
basic data inputation and
manipulation Lab:
FormattingTables.ipynb



II: Data Wrangling

How to acquire data

- Downloadable files
- API *Application Program Interface*
- Databases (MySQL, MongoDB...
Prof. Huy Vo guest lecture)



How to acquire data

- **Downloadable files**

PROS: easy to do
reproducibility: you can provide the data

CONS: store the data locally
(bad if large)

- **API Application Program Interface**

PROS: works directly with the source

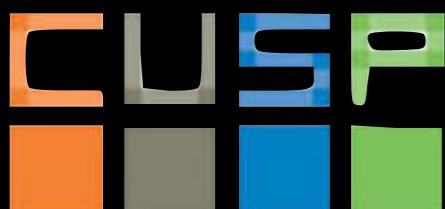
CONS: store the data locally
reproducibility: relies on data source to be permanent & API stable and accessible

- **Databases (MySQL, MongoDB...)**

Prof. Huy Vo guest lecture)

PROS: fast and flexible
can query only data of interest
reproducibility : can host remotely & provide access

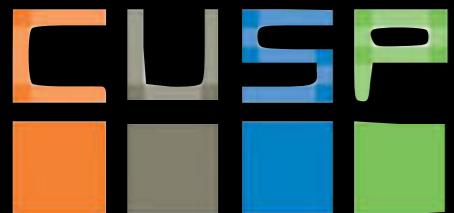
CONS: must build a database





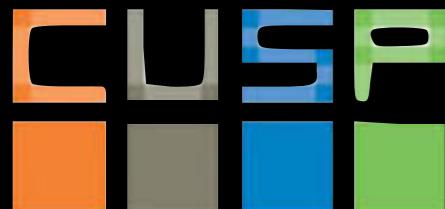
Downolading files directly
in your notebook

[https://github.com/fedhere/UInotebooks/blob/master/
dataWrangling/acquiringData.ipynb](https://github.com/fedhere/UInotebooks/blob/master/dataWrangling/acquiringData.ipynb)



Types of Data Files:

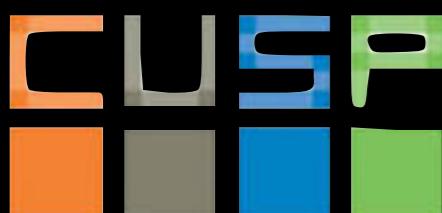
- CSV comma separated values (also TSV - tab)
- JSON: corresponds to a Python data dictionary
- XML: symilar to HTML



Types of Data Files:

- CSV comma separated values (also TSV - tab)
- JSON: corresponds to a Python data dictionary
- XML: similar to HTML

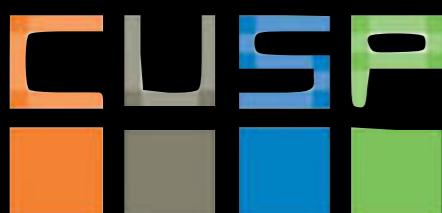
support nested structures



Types of Data Files:

- CSV comma separated values (also TSV - tab)
- JSON: corresponds to a Python data dictionary
- XML: similar to HTML

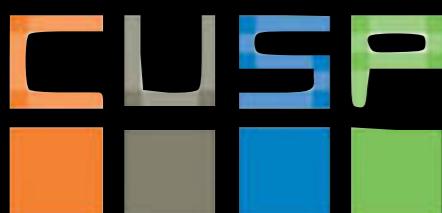
support nested structures



Types of Data Files:

- CSV comma separated values (also TSV - tab)
- JSON: corresponds to a Python data dictionary
- XML: similar to HTML

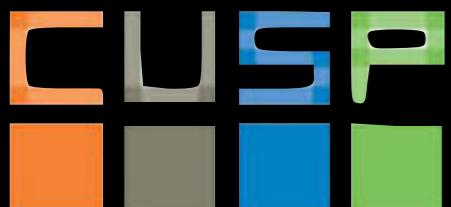
support nested structures



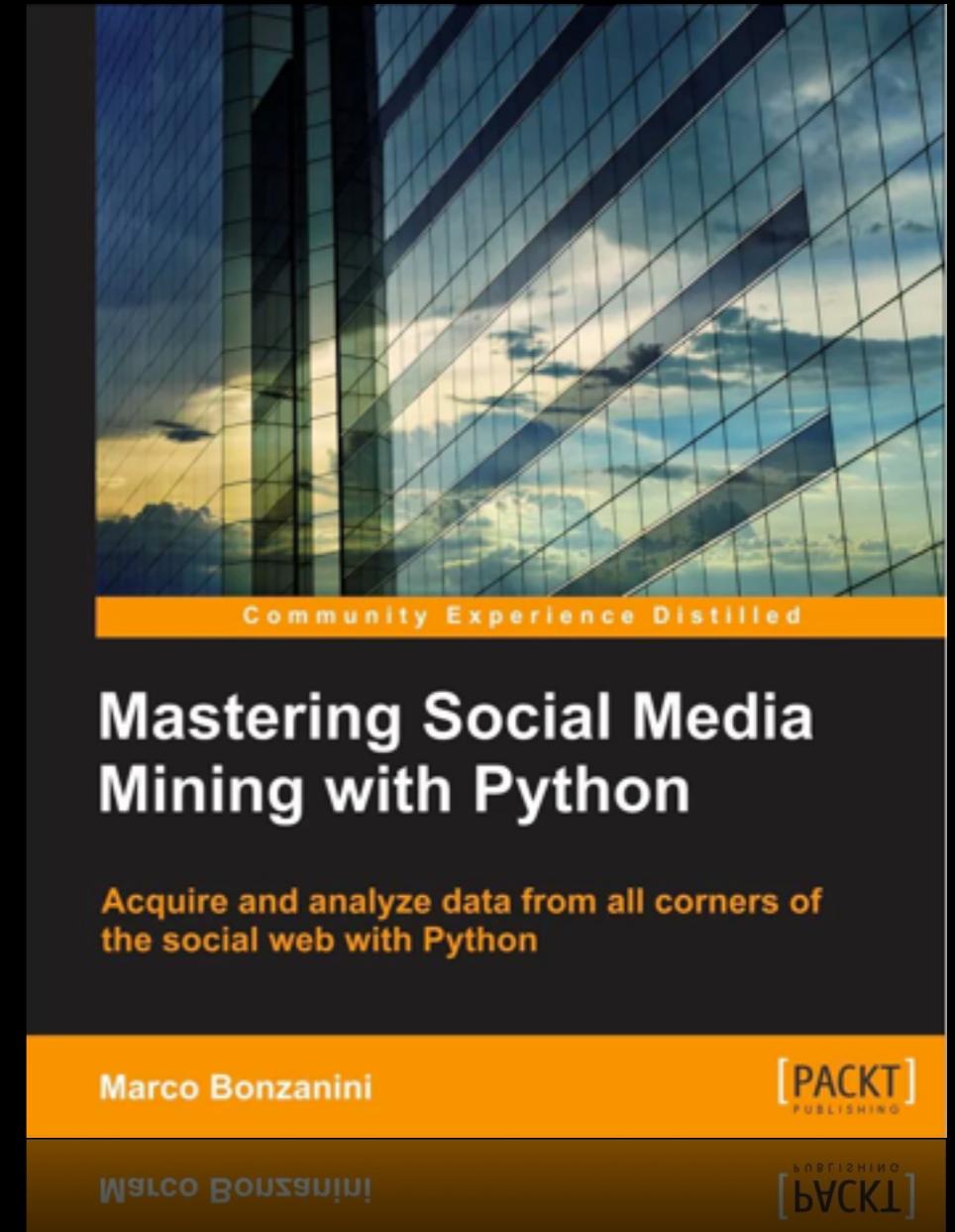
[https://github.com/fedhere/
PUI2016_fb55/blob/master/Lab2_fb55/
twitterJson.py](https://github.com/fedhere/PUI2016_fb55/blob/master/Lab2_fb55/twitterJson.py)



download an
online, nested
JSON file



<http://json.parser.online.fr/>

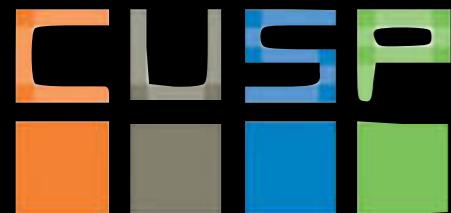


[https://marcobonzanini.com/
2015/03/02/mining-twitter-data-
with-python-part-1/](https://marcobonzanini.com/2015/03/02/mining-twitter-data-with-python-part-1/)

II: Data Wrangling

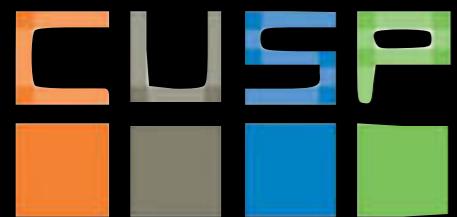
Access data through API

<http://openweathermap.org/api>





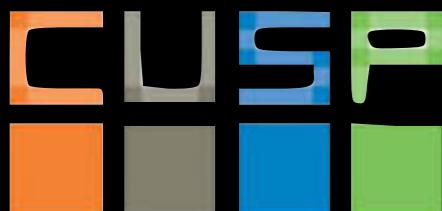
API scraping and json manipulation lab



II: Data Wrangling

Key Concepts:

- importance of reproducibility
 - data ethics and data curation
 - types of variables
 - CSV, JSON file formats properties
-
- DATA Facility data access
 - API data access
 - JSON data manipulation
 - working with Pandas Dataframes



Resources:

D.Lane 2014

Introduction to Statistics Chapters 1.6 -1.8 & 6.1-6.3

<http://onlinestatbook.com/>

Mohit Sharma 2016

CUSP UCSL tutorials Python Dictionaries

<https://sharmamohit.com/tutorials/ucsl/Python-Dictionaries/>

Allen B. Downey 2012

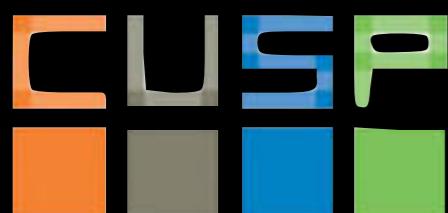
Thinking Python (free online) Appendix A ASSIGNED READING

<http://greenteapress.com/thinkpython/html/thinkpython021.html>

Wes McKinney 2013

Pandas for Data Analysis Chapter 7

<http://tinyurl.com/hydko7p>



[https://github.com/fedhere/UInotebooks/blob/master/
dataWrangling](https://github.com/fedhere/UInotebooks/blob/master/dataWrangling)

II: Data Wrangling