UNIVERSITY OF MARYLAND

INST447 -0101 Fall 2020 Lecture 3

Virtual

Instructor: Bill Farmer

TA: Jonathan Chen

Grader: Jeffrey Chen

September 15, 2020

01 Admin

02 Readings

03 Data Cleaning, transparency, etc.

04 Open Refine

05 Lab

06 Assignment

07 Next Class

This Week

Time: Tuesday virtual

- Admin
  - Office Hours Updates
  - Syllabus Updates
  - Piazza Added
- Readings
  - BadData
  - Scaling Data
- Data cleaning, transparency, etc.
- Lab will be cleaning up data with OpenRefine

Time: Thursday Virtual w/ optional live session

- Live session
  - OpenRefine examples
  - Jupyter Notebooks subjects from reading
    - Indexing
    - Concat & Append
- Lab & Assignment
- Projects - teams

If you are tired, stand up in the back of class.

Use of phone during class for non-class purposes is rude.

Admin

Admin

- Office Hours (need to schedule a time slot):
  - Monday 8-9 pm
  - Friday 8-10 am
  - Saturday 6-8 pm (changed from am to pm)
  - Sunday 6-7 pm (changed from 4-6 to 6-7)
  - By Appointment * Anytime
- Live class meetings - Thursdays 12:30-1:30
  - Class originally scheduled to start @ 12:30 so I figure this is a good time
  - We can add a couple of these at different times if/when needed
- Micro videos ? (e.g. running a notebook, Twitter account, other)
- Piazza vs. Canvas discussions

INST447 General Schedule

- Update 9/15/2020

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
					8-10am	
		Noon-ish		12:30-1:30pm		
6-7pm						
	8-9 pm				Lab 11:59pm	6-8pm

Office Hours Live

Office Hours By appointment

Video Ready

Class Live Sessions

Lab Due

UTA - Jonathan Chen

- Currently is a senior in Information Science
- Email: jonnyapple985@gmail.com
- Office hours:
  - Thursday's 4-5
  - https://umd.webex.com/umd/j.php?MTID=maa654cf56a69519872456151b7d2c073

Grader - Jeffrey Chen

- iSchool Alumni - 2019
- Currently in second semester for the Master's in Information Systems program at the R.H. Smith School of Business
- jeffrey.chen@rhsmith.umd.edu

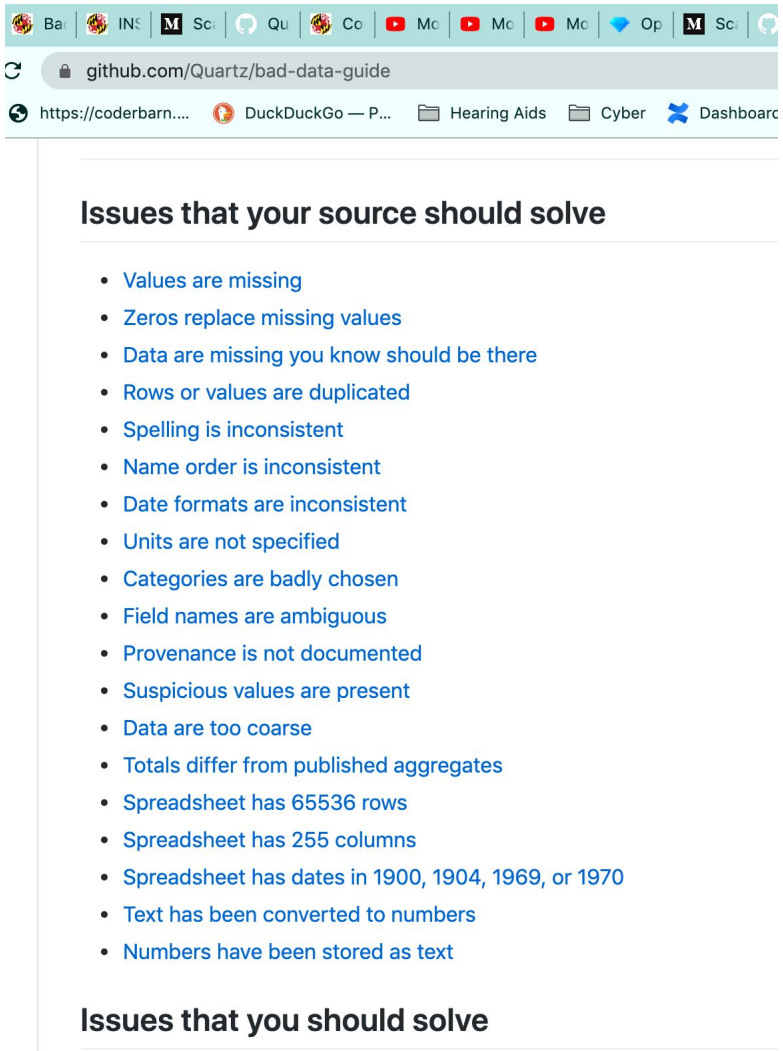
Syllabus Updates

- General syllabus schedule still applies
- Canvas -> Modules

Readings

BadData Guide

- https://github.com/Quartz/bad-data-guide
  - Issues that your source should solve
    - Missing values, duplicate rows, ambiguous field names, inconsistent date formats, etc.
  - Issues that you should solve
    - Seasonal variation skews the data, data entered by humans (error prone), non random data (time-of-day, native language, etc.)
  - Issues a third-party should help solve
    - Author is untrustworthy (get two or three sources), inexplicable outliers
  - Issues a programmer should help solve
    - Data are aggregated to the wrong categories or geographies
      - e.g. data aggregated by zip code rather than city neighborhoods



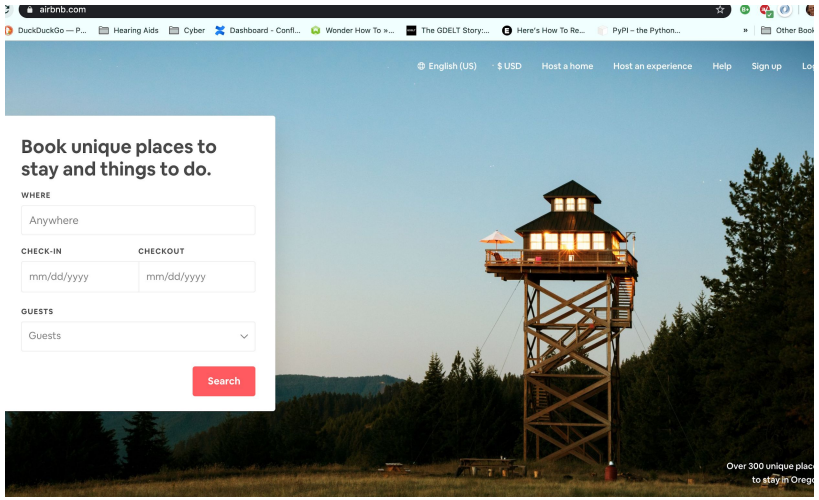
Scaling Knowledge at Airbnb

Data team

Problem: The data team at Airbnb has a responsibility to scale the ability to make decisions using data.

- @ Airbnb - “The democratize data access to empower all employees to make data-informed decisions”
- Give everybody the ability to use experiments to correctly measure the impact of decisions
- Turn those insights on user preferences into data products that improve the experience of using Airbnb.

- ***BUT….***How to make an insight discovered by one person transfer effectively beyond the target recipient **“scaling knowledge”**
- Issues, just like other DS/SWE teams
  - Disorganized knowledge repos (local, servers, emails)
  - Previous work doesn’t have up to date code.
  - Current version of code isn’t what generated the previous plots
  - General issue of trying to reproduce what someone else did and not being successful
  - She distributes her results in a presentation, email, or doc perpetuating the cycle
- All of this slows down analysis and speed of decision making
- A streamlined approach is needed. Realized that they could do better!

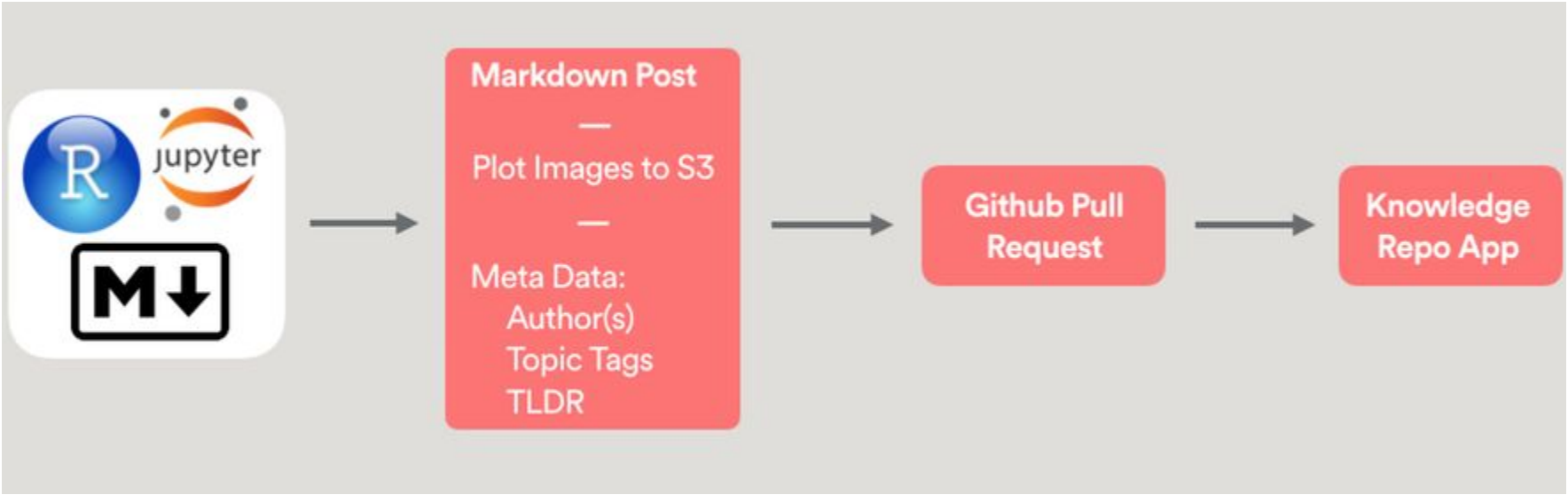


Scaling Knowledge at Airbnb

- Five Key Tenants for DS research going forward
  - *Reproducibility*
    - There should be no opportunity for code forks. The queries, transforms, visualizations and write-ups should be contained in each contribution and be up to date with the results.
  - *Quality*
    - Research should not be shared without being reviewed for correctness and precision (code/peer reviews)
  - *Consumability*
    - The results should be understandable to readers. Aesthetics should be consistent and on brand across research.
  - *Discoverability*
    - Anyone should be able to find, navigate, and stay up to date on the existing set of work on a topic
  - *Learning/Transparency*
    - Other researchers should be able to expand their abilities with tools and techniques from others' work

Scaling Knowledge at Airbnb

“Knowledge Repo”



Combined all of their ideas into one system. It combines a process around contributing and reviewing work, with a tool to present and distribute it. They call it a ‘Knowledge Repo’. Git repo. Posts are written in Markdown. Everything is committed. Templates for code - metadata (author, tags,and a TLDR). A Flask web-app renders the Repo’s contents as an internal blog, organized by time, topic, or contents.

Scaling Knowledge at Airbnb

Code Review of Software Engineering
+
Peer review of academia
=
Trusted, repeatable research going at
‘startup speed’

Constantly checking for improvements

Scaling Knowledge at Airbnb

- Results
  - *Reproducibility*
    - All of the work, from the query of the core ETL (Extract, Transform, Load) tables, to the transforms, visualizations, and write-up is contained in one Jupyter NB, RMarkdown, or markdown file.
  - *Quality*
    - Using GitHub's functionality of pull requests prior to publishing, peer review and version control is put directly into the work flow.
  - *Consumability*
    - Markdown served by the web-app hides code and uses their internal branded aesthetics, making the work more accessible to less technical readers. The peer review process provides feedback on writing and communication, improving the quality of work.
  - *Discoverability*
    - The structured metadata (author, tags, TLDR) allows for easy navigation through past work. Tags provide a many-to-one topic inheritance and searching. Users can subscribe to topics. Posts can be bookmarked, browsed by author.
  - *Learning*
    - By having previous work easily searchable, it becomes easier to learn from each other.

Other (optional interesting reads)

- "**Working with Data Across Services is Hard**" Billions of Messages a Day - Yelp's Real-time Data Pipeline
- news.ycombinator.com - search on data 'pipeline/sets/cleaning/etc.'
- engineering.salesforce.com - TLS Fingerprints w/ JA3 and JA3S (https://engineering.salesforce.com/tls-fingerprinting-with-ja3-and-ja3s-247362855967)
- https://towardsdatascience.com/
- https://www.reddit.com/r/datasets/

Open Refine

- Watch the 3 videos. They are self-explanatory.
- https://openrefine.org/
- “Google refine” -> Google open sourced it

Data Cleaning, Validation, Transparency and Reproducibility

Data sources

Types of Sources

- Primary Sources - Data you (or your organization) has created (** Interesting for project)
  - Collected customer data, survey data, interviews
  - Health data
  - IoT, instrument, or log data
  - etc.
- Secondary Sources - Data other people have created or aggregated
  - Government, FOIA requests,
  - Shared scientific data, national surveys
  - Social Media and other third party APIs
  - Finance
  - etc.

Common Data source issues

Secondary data sources

- Corrupted or Untrustworthy Data
    - social media, government provided, shared
- Poorly Documented
    - It’s not often that you find well documented data
- Difficult/Unfamiliar Format - KML, FITS

Data Provenance

- A historical record of where the data came from, how it was collected, and how it was handled.
    - Who collected it?
        - Some orgs are biased and may cherry pick data
    - How was the data collected?
        - Be as detailed as possible, some collection methods are flawed
    - Were there any modifications?
        - Human?
        - Computer?
        - Human & Computer?

Case Study

- Pro-Russia Ads Dataset
    - https://www.reddit.com/r/datasets/comments/8s0wrr/dataset_of_3500_ads_by_prorussia_group/

Follow the links backwards to reconstruct the data provenance

- Who collected the data?
- How was the data collected?
- Was it modified/preprocessed?
    - Computer?
    - Human?

Follow the links backwards to reconstruct the data provenance.

- Look at the summary on the Reddit page
- Description
    - Dems in the US House Intel Committee released 3500 pdfs with texts and images
    - Who collected it? beeeeeeers
    - Facebook and Instagram provided it? We assume.
    - USHIC provided them on their website

Follow the links backwards to reconstruct the data provenance.

- beeeeeeeeers applied OCR to turn pdf -> text
- beeeeeeeeers converted text to json/csv extracting key elements, eg. cost of ad
- Was it modified?
  - We know that it was modified (see previous steps)
  - Not sure exactly how. third party software? Python?
- Were there mistakes introduced?
  - scripting errors, convenience sampling, misunderstanding of fields

What are common Data Formats

- Structured data - organized and machine readable
    - csv, xml, json
    - database
    - spreadsheet
- Unstructured data
    - Images
    - Web pages
    - emails
    - audio files
    - pdfs

Comma Separated Values . csv

File

Animals,Color

Elephant,Grey

Giraffe,Yellow

Dolphin,Blue

Representation

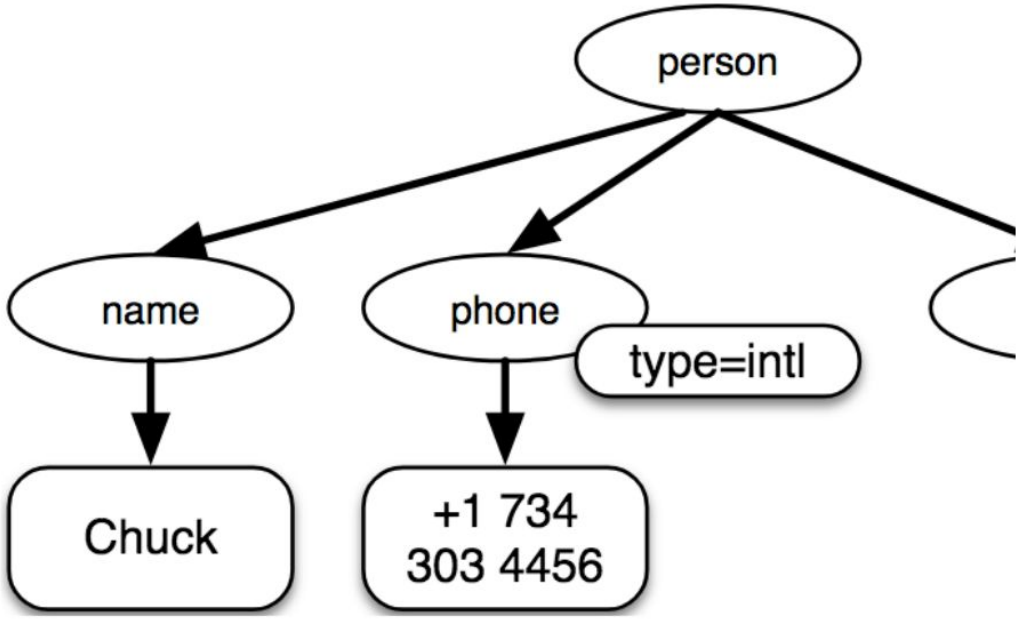
	A	B
1	Animals	Color
2	Elephant	Grey
3	Giraffe	Yellow
4	Dolphin	Blue

JavaScript Object Notation (JSON)

File

```
{
  "name" : "Chuck",
  "phone" : {
    "type" : "intl",
    "number" : "+1 734 303 4456"
   },
   "email" : {
     "hide" : "yes"
   }
}
```

Representation



Unstructured Data

- Data values are not organized in a standardized way.
- This makes it difficult to read and extract values with a computer, humans can read easily
- Examples
  - webpages
  - emails
  - audio files
  - pdfs

Data Orientation

- Get oriented with your data before you start your data cleaning
  - data provenance
  - what is the format?
  - how many observations?
  - how many variables?
  - what are the variables?
  - what do the values mean?
  - what are typical values for each variable?

Bias

Slides adapted from Nekabari Sigalo

How to handle data issues & biases

- Identify issues and biases
  - e.g. misspellings
- Address issues and biases if possible
  - e.g. fix misspellings
- Document issues and biases
  - e.g. record that misspellings were fixed in cells X & Y
- Decide whether results are valid despite issues and biases
  - e.g. results are more accurate because misspellings were fixed

Types of data issues & biases

- 4 Common Types
  - Biases due to data collection methods
  - Missing data
  - Inconsistent data
  - Data errors

Biases due to data collection

- Random Sampling - each entity in the population has an equal chance of being selected
  - e.g. randomly select 20 students from registry
- Convenience Sampling - A set of entities in that population that were easy to gather data from
  - e.g. select 20 students from physics 100
- Most statistical techniques assume data is collected through random sampling. In reality, sampling is almost always collected through convenience sampling which creates bias.

Identify if data collection introduced bias

- Common sources of bias based on flawed data collection methods
  - Sample size is too small
    - e.g. only 20 observations
  - Stopping procedure is based on data
    - e.g. tricking the data into giving you the result. You collect data until you get the result you want.
  - Convenience sampling collects a non-representative group
    - e.g. survey about exercise recruit people at a gym
  - Only capturing data for part of a “season”
    - e.g. collect uber traffic data on M-Th when most traffic happens on weekend

Address bias from data collection methods and decide if problematic

- Address bias from data collection methods
  - Use good data collection methods if you are collecting yourself
  - Only use data sets collected by others that were collected using good methods that minimize bias
- Decide if results are valid given data collection methods
  - Evaluate how much bias is introduced on data collection methods. Evaluate how it affects your results

Missing Data

- Ideally you would have no missing data. There are several reasons why missing data occurs
  - Data was never recorded
    - e.g. sensor failure, human did not respond to question
  - Data was lost or corrupted
    - e.g. value out of range for database, human error

Missing Data can create bias

- Missing data can create biased results because data is rarely missing at random.
  - Extreme values are more likely to be incompatible with database
  - Humans often choose not to answer sensitive questions
    - e.g. low income individuals may not want to answer a question about income
    - e.g. heavy drug users may not want to answer a question about drug use

Identify missing data

- Identify values used to encode missing data (e.g. N/A, blank, null)
  - Check documentation
  - How is it encoded
    - NA, N/A, NULL, Nan, “”, 0, -1
    - 1970-01-01T00:00:00Z
  - Inspect values, there may be a mixture
- Did you receive all of the rows and columns that you expected?
- Get a count of missing values per column

Addressing missing data

- Find out why you don’t have it
  - May not be able to
- Imputation methods
  - Infer it from data around it, make a guess, assume the average
- Exclude it
  - Ignore the entire row.
- Decide whether or not you can use the data if you are missing too many values!

Inconsistent data

- Dates
  - e.g. “2019-03-16” vs “March 16, 2019”
- Variants to represent the same values
  - “USA” vs “United States” vs “United States of America”
- Different units are used
  - e.g. height in cm and inches

Inconsistent data

- Standardize the values
  - e.g. reformat dates
  - e.g. combine variants by recoding values (e.g. “USA”)
  - e.g. use consistent units (inches vs cm)

Identifying and addressing data errors

- Unusual or unexpected values may be errors
  - e.g. out of range date “2070-03-23”
  - e.g. very large integer “99999999999”
  - e.g. negative number for something that should only be positive like height, weight, age
- Outliers may be errors
- How to address?
  - Try to fix
  - Exclude

Transparency and Reproducibility

Transparent & Reproducible Projects

- Transparency – It is easy for you and others to understand the data and how it has been analyzed
  - Documentation
    - Source of data & collection methods
    - Known issues & how they were addressed n Transformations & calculations
  - Justification
    - Why were these methods used

Transparent & Reproducible Projects

- Reproducibility – It is easy for you and others to repeat your analysis.
  - Create copies of data
    - Don’t overwrite your raw data create intermediate data sets
    - Save a final version of your data before analysis ¨ Create an automated pipeline
  - You will want to run your pipeline multiple times
    - Preferably included in only one script
  - OpenRefine automatically creates a list of cleaning steps taken

Design embodies good cleaning practices

- Reproducibility
  - Saves data to new file.
  - Built in versioning that allows you to reverse any action.
  - Can export and apply the same script to a new data set (or same data set again).

Documentation

- Metadata: Structured information describing a dataset.
  - e.g. dates collected, creators of dataset, variables
- Codebook: Description of variables and data values in dataset.
  - e.g. units, missing values, transformations etc.

Reproducibility

- Reproducibility - It is easy for you or others to repeat the steps of your analysis
  - Create copies of data
    - Don’t overwrite your raw data, create intermediate data sets
    - Save final version of your data before analysis
  - Create an automated pipeline
    - You will want to run your pipeline multiple times
    - Preferably included in only one script (often not realistic). Make sure documentation is up to speed!

Open Refine

OpenRefine / Google Refine

- Open source source tool to clean messy data originally created by Google.
  - Graphical User Interface
  - Scripting Language

- Using OpenRefine helps you to see the big picture of your data, discover inconsistencies, and fix them.

- Not sure about extremely large data sets; however, you can use a subset of data in combination with OpenRefine to find general problems.
-

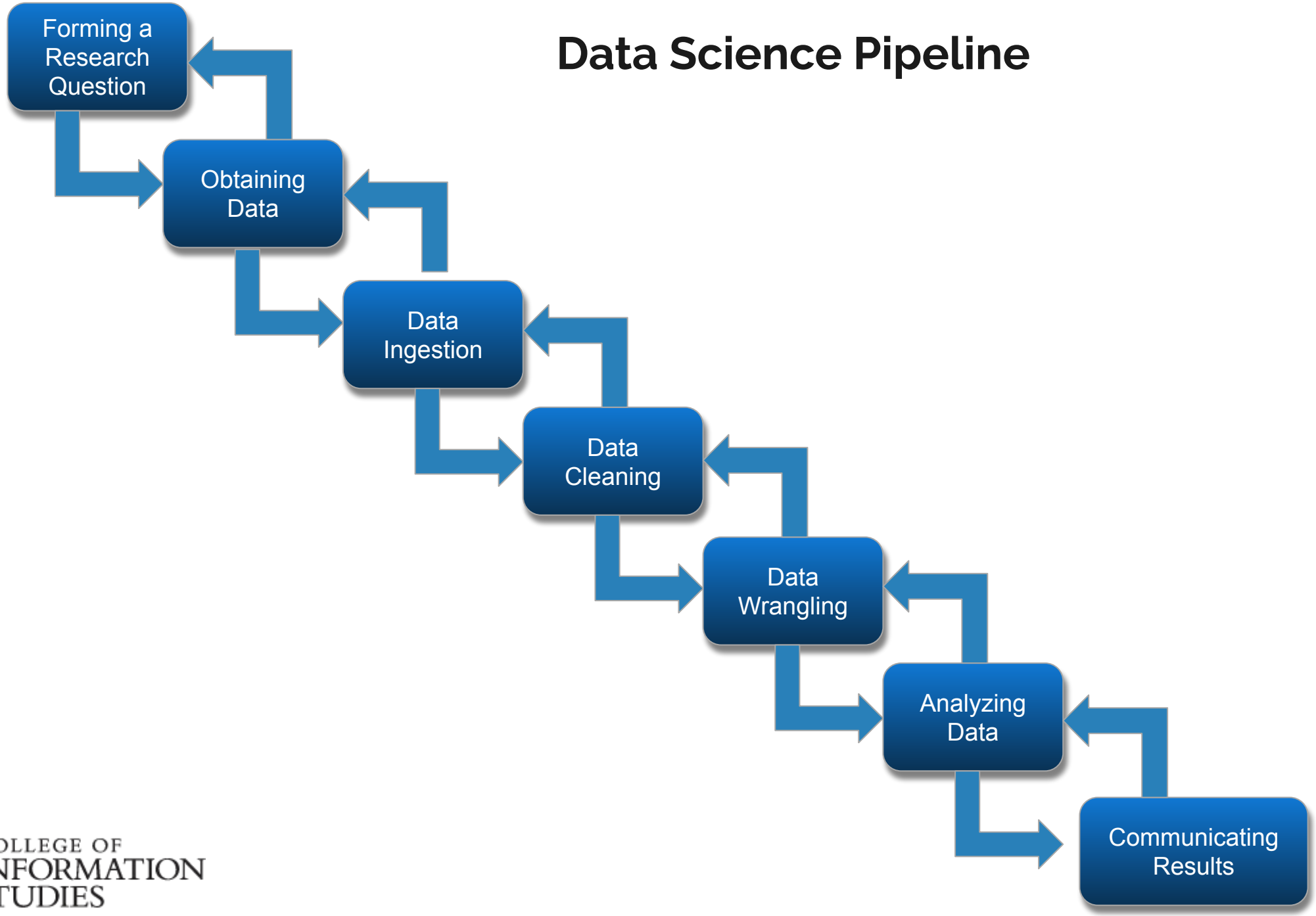
Keep Track Changes OpenRefine

- Use the Facets to fix up standardization issues in the variable EMPLOYER
  - What kind of non-standardization issues do you notice?
- View a list of edits that you've made using the "Undo/Redo" tab
- Export a script with your list of changes using the "Extract" option in the "Undo/Redo" tab

Lab

Projects

Data Science Pipeline



COLLEGE OF INFORMATION STUDIES

Projects

- Teams of 2, 3, or 4
  - Class is small so I prefer sizes of 2 or 3
  - Project proposals due 2/28, so you have time
- API Keys
  - Twitter (see my submission process)
- Scraping
  - Reddit example (json)
  - Reddit group on data sets https://www.reddit.com/r/datasets/

Data Science Projects

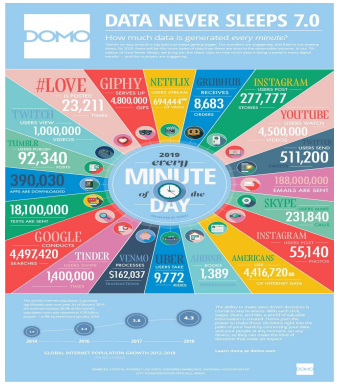
Data science - the ability to take large amounts of data in many different formats and be able to understand it, to process it, to extract value from it, to summarize it, to visualize it, and to communicate it to others.

Data science is the field of study that combines domain expertise, programming skills, and knowledge of math and statistics to extract meaningful insights from data. Data science practitioners apply machine learning algorithms to numbers, text, images, video, audio, and more to produce artificial intelligence (AI) systems that perform tasks which ordinarily require human intelligence. In turn, these systems generate insights that analysts and business users translate into tangible business value. -datarobot.com

FEC.gov

YAHOO! FINANCE

United States® Census Bureau



- Sentiment Analysis
- Customer Segmentation
- Recommending products
- Public Health Issues
- Manufacturing - predicting faults
- Financial Risk Analysis

Illustration from: https://www.socialmediatoday.com/news/what-happens-on-the-internet-every-minute-2019-version-infographic/558793/

Projects

salesforce

https://engineering.salesforce.com/ & TLS Fingerprints JA3 and JA3S

Berkeley SETI Research Center

FITS file handling
https://docs.astropy.org/en/stable/io/fits/index.html

astropy:docs

COLLEGE OF INFORMATION STUDIES

Python and Jupyter Notebooks

Python and Jupyter Notebooks review

- Examples and Pandas

Next Week

Next Week

- Assignment 1 due
  - Size of data
- Biases
- Proposal Brainstorm
- More Python/Pandas
  - Data frames
  - Visualizing
  - Summarizing
-

I appreciate your attention

Hope to see you on Thursday!

Reference Material
Install Software

4 Programming Assignments

- Work independently
- Deeper investigation into a data set and research question
- Turn in a well-structured and written report using Jupyter notebooks

Software Tools

- Python & Jupyter Notebook
  - Method 1
    - Python 3 (https://www.python.org/downloads)
      - Pandas Data Analysis Library (pandas)
      - Other modules (e.g. numpy, plotnine)
    - Jupyter Notebooks (aka ipython) (https://jupyter.org/install)
      - blend narrative text
      - code
      - output
      - visualizations
  - Method 2
    - Install Anaconda (includes both) (https://www.anaconda.com/distribution)
- Open Refine
  - http://openrefine.org/download.html
- Data sets
  - https://www.reddit.com/r/datasets/
  - https://opendata.dc.gov/
  - https://datasetsearch.research.google.com/
  - https://www.kaggle.com/datasets