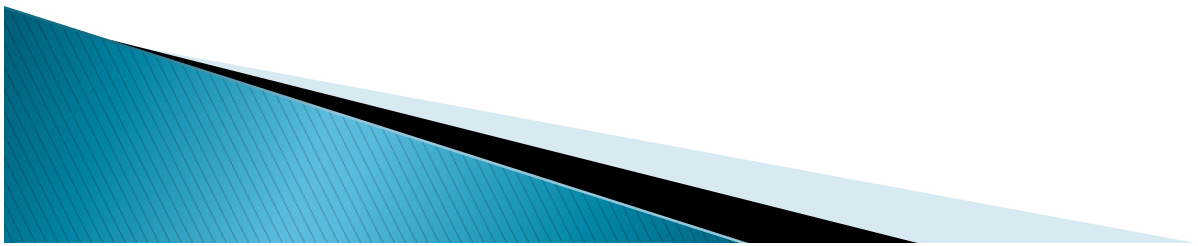
CSCI-UA.60-1 Database Design and Implementation

Garbage In, Garbage Out! Watching out for Bad Data

CSCI-UA.60-1
Prof Deena Engel
Department of Computer Science
deena.engel@nyu.edu

Introduction

- However … before we introduce working with data in SQLite, MySQL and other environments … it is important for you to learn to manipulate or “scrub” data!
- There are a number of data issues to be aware of … so that you are not caught using bad data.



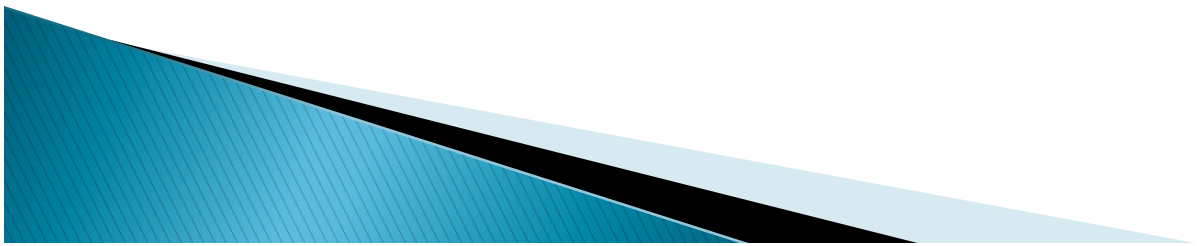
First, you need to understand your data structure and data values .

1. For example, in a CSV file, how many values are there per line? And is the number of fields consistent from one record or row to the next?
2. Is the delimiter character consistent
3. For fixed width columns … are the columns set up consistently?
   - Consider using python's **csv** module to create or read from .csv files.
   - *For class discussion: What are some of the approaches you might use to correct these problems?*

Field Validation

1. Are all of the values of a given field cast in the same data type?
2. Is each field limited to one value?

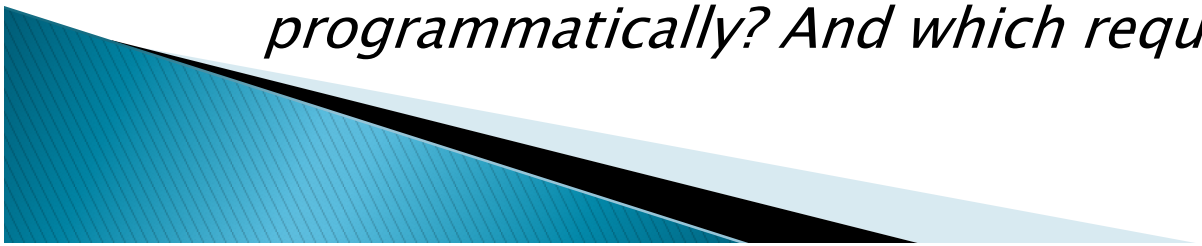
    - *For class discussion: which of the above can be checked programmatically? And which require human intervention?*



Value Validation

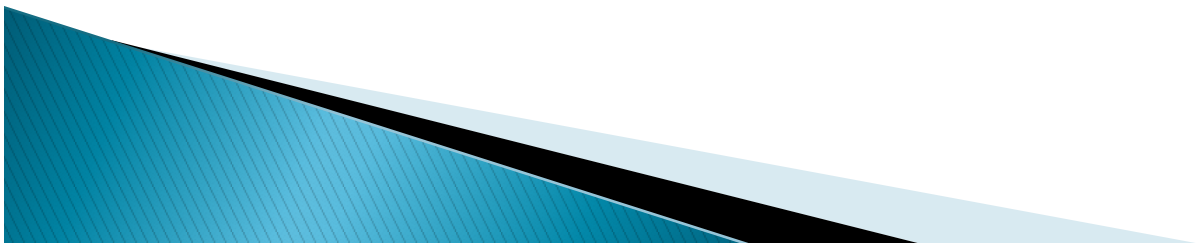
1. Do all of the values for a given field have the same meaning? (For example, in a field called “height”, are all of the values related to the height of each entity and not its weight?)
2. Are all of the values within a range that is reasonable for that field? (For example, 10 is not a reasonable value for the height of a person.)
3. Do all of the values use the same unit of measure (pounds, meters, years, etc)?
4. Do all of the values use an appropriate unit of measure (dollars for prices, meters for length, kilograms for weights, etc.)?

    - *For class discussion: which of the above can be checked programmatically? And which require human intervention?*



Using Simple Statistics to evaluate your data

- For numerical data, what are some of the statistics that could help to inform you if the values are meaningful?
  - For example, taking minimum and maximum values could highlight “outliers” or errors.
  - Checking both the mean and the median will help to clarify the data as well.
  - Use visualization (e.g. graphing your data) can give one a quick overview.



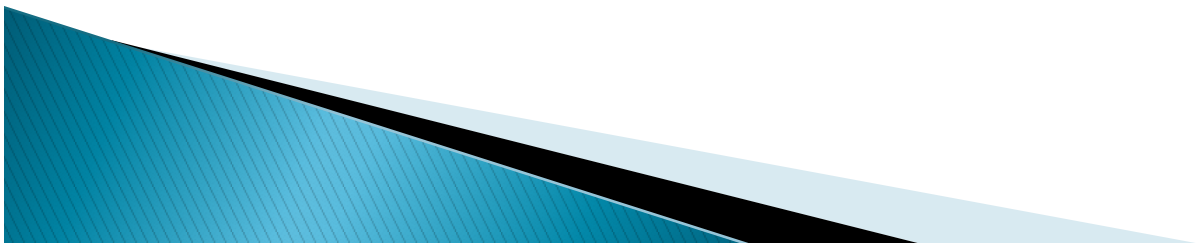
Data that is intended for human readers … not machines!

In some cases, data are widely spread out or organized in a logical and visual way for readers:

- The Bureau of Labor report series is one of many examples:
  - https://www.bls.gov/news.release/pdf/empsit.pdf
- It would be important to either find another source of these data which are pre-formatted for use by a database for analysis, or to write a program to modify this file.
- Another data issue you might encounter in this way are data sets that are written across several files.
- *The solution to this problem is by writing code!*

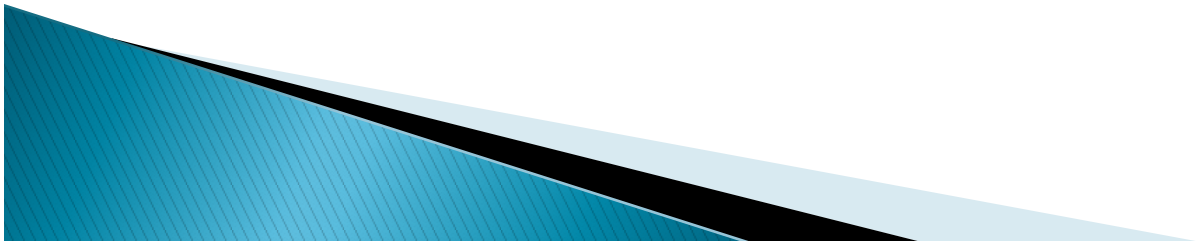
Another problem: Bad data in the text itself

- Sometimes textual data includes characters that are specific to its presentation (e.g. markup)
  - If you are culling data from the web, consider using Beautiful Soup (http://www.crummy.com/software/BeautifulSoup/)
  - Consider using python's **urllib** to facilitate obtaining data directly from the web (https://docs.python.org/3.0/library/urllib.request.html)



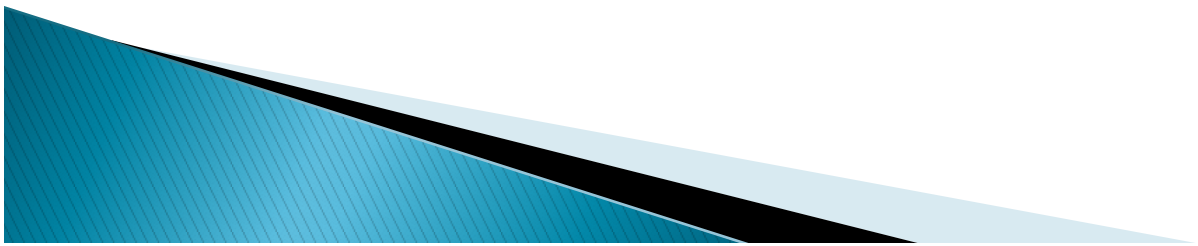
Liars!?

- Just because it is on the web … does that make the data valid? !!
- What are some of the ways that you might determine whether the provider of your data is lying to you? Or simply posting sloppy data?
- Many of the current techniques for evaluating textual data are outside of the scope of this course (e.g. sentiment classification, polarized language and other problems) but I would be happy to give you further resources if you are interested.



Spreadsheets

- Beware of data that is provided in or from spreadsheets.
  - Reinhart-Rogoff error in Excel (2010)
  - http://www.businessweek.com/articles/2013-04-18/faq-reinhart-rogoff-and-the-excel-error-that-changed-history
- *For class discussion: What are appropriate and constructive uses of spreadsheets for data analysis?*

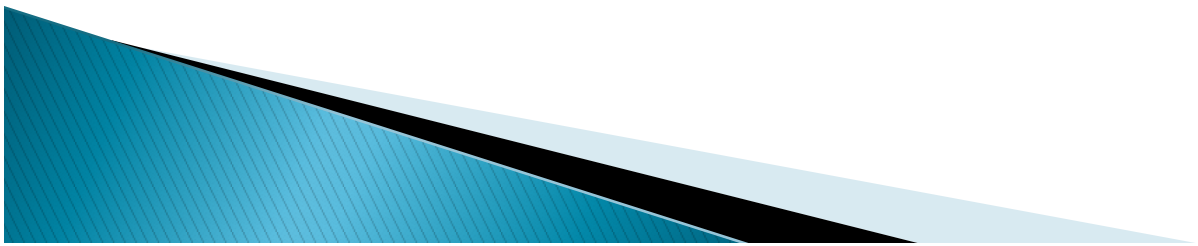


Other sources for errors

- It is important to evaluate your data in light of the context for common errors that can misshape your results:
  - For example, stock splits in a database of financial transactions
  - Political data that might be skewed by the provider
  - Survey data that are limited to specific socio-economic subgroups in society
- These considerations are outside of the scope of this course, but I would be happy to meet with you individually if you are concerned about the data in your field and also to assist you in finding one of the appropriate Subject Librarians at the Bobst to assist you (https://library.nyu.edu/subject-specialists/).

Good Practices

- Ensure data traceability: Keep careful notes on your data sources (e.g. URLs) as well as sample data files for future reference.
- Ensure reproducibility: Keep copies of your python programs that you write to work with your data so that you can reproduce the results with the same or more current datasets.



In conclusion …

- The four standards for evaluating and maintaining good data*:

1. Complete (Do you have all of the data that you need or if not, is this a representative sample?)
2. Coherent (do the data and the data structure "make sense"?)
3. Correct (are the values correct?)
4. Accountability (can you trace the data and reproduce any modifications?)

**Adapted from The Bad Data Handbook, edited by Q. Ethan McCallum, published by O'Reilly*