CSE 519: Data Science
Steven Skiena
Stony Brook University

Lecture 7: Data Cleaning

Cleaning Data: Garbage In, Garbage Out

Many issues arise in ensuring the sensible analysis of data from the field, including:

- Distinguishing errors from artifacts.
- Data compatibility / unification.
- Imputation of missing values.
- Estimating unobserved (zero) counts.
- Outlier detection.

Errors vs. Artifacts

- Data errors represent information that is fundamentally lost in acquisition.
- Artifacts are systematic problems arising from processing done to data.

The key to detecting artifacts is the sniff test, examining the product closely enough to get a whiff of something bad.

First-time Scientific Authors by Year?

In a bibliographic study, we analyzed PubMed data to identify the year of first publication for the 100,000 most frequently cited authors.

What *should* the distribution of new top authors by year look like?

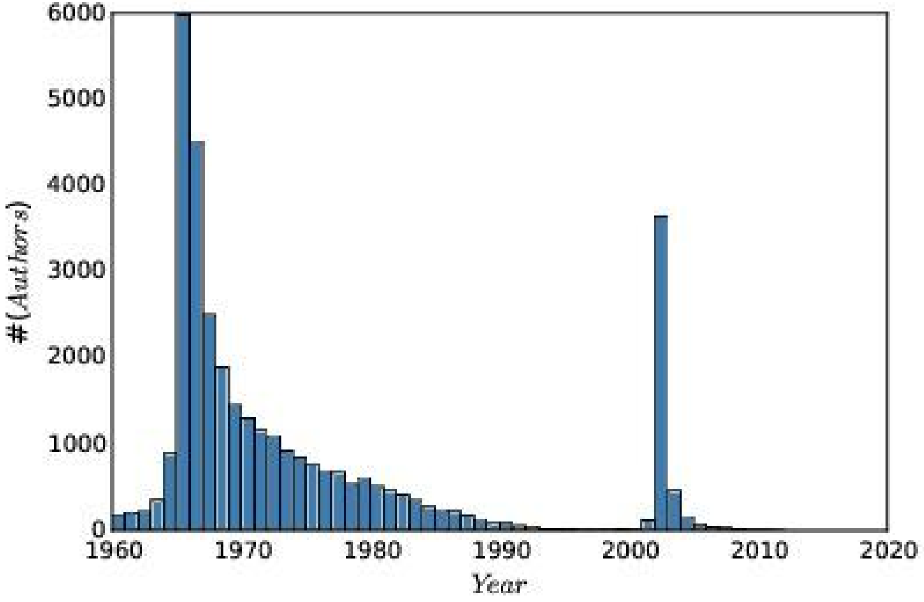
It is important to have a preconception of any result to help detect anomalies.

Might this be Right?

A student once tried to foist this off on me.

What artifacts do you see?

What possible explanations could cause them?



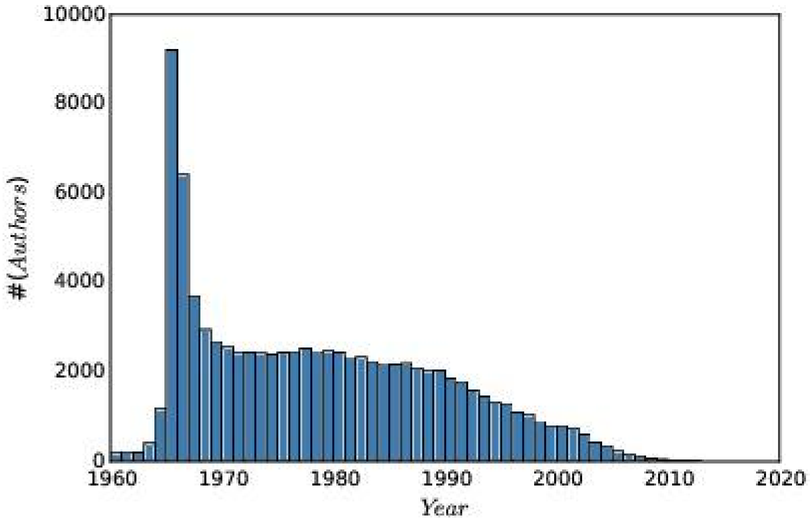
Mystery Solved!

Pubmed used author first names starting in 2002.

SS Skiena became

Steven S Skiena

Data cleaning gets rid of such artifacts.



Data Compatibility

Data needs to be carefully massaged to make ``apple to apple'' comparisons:

- Unit conversions
- Number / character code representations
- Name unification
- Time/date unification
- Financial unification

Unit Conversions

NASA's Mars Climate Orbiter exploded in 1999 due to a metric-to-English conversion issue.

- Even sticking to the metric system has potential inconsistencies: cm, m, km?
- Bimodal distributions can indicate trouble
- Z-scores are dimensionless quantities.

Vigilance in data integration is essential.

Number / Character Representations

The Ariane 5 rocket exploded in 1996 due to a bad 64-bit float to 16-bit integer conversion.

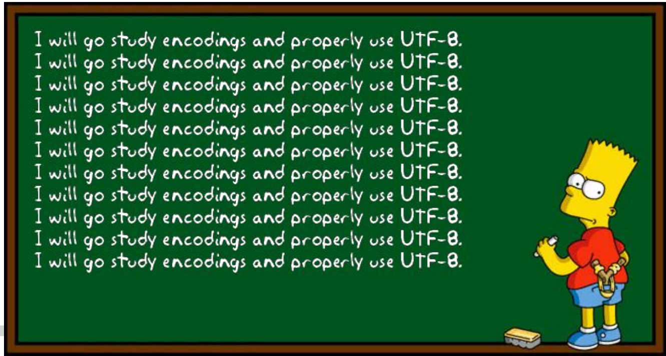
- Measurements should generally be decimal numbers
- Counts should be integers.
- Fractional quantities should be decimal, not (q,r) like (pounds,oz) or (feet,inches).

Character Representations

A particularly nasty cleaning issue in textual data is unifying character code representations:

- ISO 8859-1 is a single byte code for ASCII
- UTF-8 is a multibyte encoding for all Unicode characters.

Unicode font, UTF8 format	Unicode font, XXX... format
搜索简体中文网页	????????
Recherche avancée	Recherche avancée
網路畫廊，含中、港、澳參展作品	???????????????
โหรดˆดง่ายกว่า	??????????????
ウエブ全体から	???????
kehren Sie zur Suche zurück	kehren Sie zur Suche zurück
Сделайте Google стартово	???????? Google ????????
إخدوقت بحث أقل ووقت مطالعة أطول	?????? ??? ??? ???? ?????? ????



Name Unification

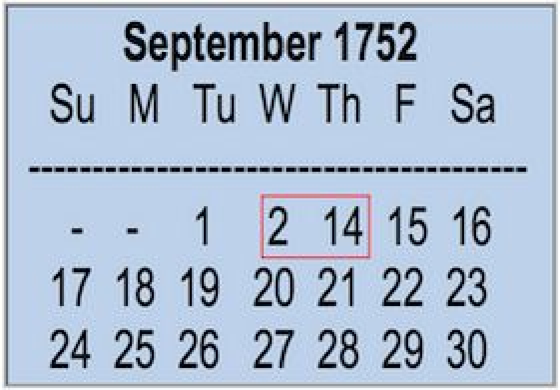
I appear on the web as:

(Steve|Steven|S.) (S.|Sol|_) (Skiena|Skeina|Skienna)

- Use simple transformations to unify names, like lower case, removing middle names, etc.
- Consider phonetic hashing methods like Soundex and Metaphone.

Tradeoff between false positives and negatives.

Time / Date Unification



September 1752

Su	M	Tu	W	Th	F	Sa
-	-	1	2	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

Aligning temporal events from different datasets/systems can be problematic.

- Use Coordinated Universal Time (UTC), a modern standard subsuming GMT.
- Financial time series are tricky because of weekends and holidays: how do you correlate stock prices and temperatures?

Financial Unification

- Currency conversion uses exchange rates.
- Correct stock prices for splits and dividends.
- Use returns / percentage change instead of absolute price changes.
- The time value of money needs correction for inflation for fair long-term comparisons.

Why do stock/oil prices correlate over 30 years?

Dealing with Missing Data

An important aspect of data cleaning is properly representing missing data:

- What is the year of death of a living person?
- What about a field left blank or filled with an obviously outlandish value?
- The frequency of events too rare to see?

Setting such values to zero is generally wrong

Imputing Missing Values

With enough training data, one might drop all records with missing values, but we may want to use the model on records with missing fields

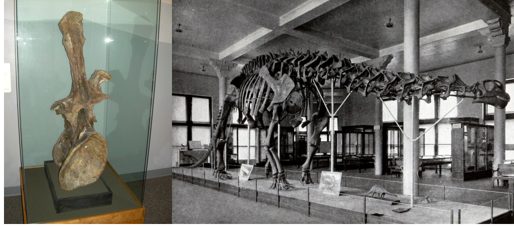
Often it is better to estimate or impute missing values instead of leaving them blank.

A good guess for your death year is birth+80.

Imputation Methods

- *Mean value imputation* - leaves mean same.
- *Random value imputation* - repeatedly selecting random values permits statistical evaluation of the impact of imputation.
- *Imputation by interpolation* - using linear regression to predict missing values works well if few fields are missing per record.

Outlier Detection



The largest reported dinosaur vertebra is 50% larger than all others: presumably a data error.

- Look critically at the maximum and minimum values for all variables.
- Normally distributed data should not have large outliers, *k sigma* from the mean.

Fix why you have an outlier. Don’t just delete.

Detecting Outliers

- Visually, it is easy to detect outliers, but only in low dimensional spaces.
- It can be thought of as an unsupervised learning problem, like clustering.
- Points which are far from their cluster center are good candidates for outliers

Delete Outliers Prior to Fitting?

- Deleting outliers prior to fitting can yield better models, e.g. if these points correspond to measurement error.
- Deleting outliers prior to fitting can yield worse models, e.g. if you are simply deleting points which are not explained by your simple model.