Data Munging

- Good data scientists spend most of their time cleaning and formatting data.
- The rest spend most of their time complaining there is no data available.
- *Data munging* or *data wrangling* is the art of acquiring data and preparing it for analysis.

Sources of Data

- Proprietary data sources
- Government data sets
- Academic data sets
- Web search
- Sensor data
- Crowdsourcing
- Digitization

Proprietary Data Sources

- Facebook, Google, Amazon, Blue Cross, etc. have exciting user/transaction/log data sets.
- Most organizations have/should have internal data sets of interest to their business.
- Getting outside access is usually impossible.
  - Business issues, and the fear of helping their competition.
  - Privacy issues, and the fear of offending their customers.
- Companies sometimes release rate-limited APIs, including Twitter and Google.
  - Providing customers and third parties with data that can increase sales
  - It is generally better for the company to provide well-behaved APIs

Government Data Sources

- Governments have made many data open.
- Data.gov has over 100,000 open data sets!
- The Right To Information (RTI) enables you to ask if something is not open.
- Preserving privacy is often the big issue in whether a data set can be released.

Academic Data Sets

- Making data available is now a requirement for publication in many fields.
- Expect to be able to find economic, medical, demographic, and meteorological data if you look hard enough.
- Track down from relevant papers, and ask.

Web Search/Scraping

- *Scraping* is the fine art of stripping text/data from a webpage.
- Libraries exist in Python to help parse/scrape the web, but first search:
  - Are APIs available from the source?
  - Did someone previously write a scraper?
- Terms of service limit what you can legally do.

Few Available Data Sources

- Bulk Downloads: e.g. Wikipedia, IMDB, Million Song Database.
- API access: e.g. New York Times, Twitter,
- Facebook, Google.

Be aware of limits and terms of use

Sensor Data Logging

- The “Internet of Things” can do amazing things:
  - Image/video data can do many things: e.g. measuring the weather using Flicker images.
  - Measure earthquakes using accelerometers in cell phones.
  - Identify traffic flows through GSP on taxis.

Build logging systems: storage is cheap!

Crowdsourcing

- Many amazing open data resources have been built up by teams of contributors:
  - Wikipedia/Freebase
  - IMDB
- Crowdsourcing platforms like Amazon Turk enable you to pay for armies of people to help you gather data, like human annotation.

Digitization

- But sometimes you must work for your data instead of stealing it.
- Much historical data still exists only on paper or PDF, requiring manual entry/curation.
- At one record per minute, you can enter 1,000 records in only two work days.

Cleaning Data: Garbage In, Garbage Out

- Data collected in raw form may not be in usable form for analysis.
- Before we start analysis, a proper cleaning is required.
- Cleaning of data may include
  - Distinguishing errors from artifacts
  - Data compatibility / unification
  - Imputation of missing values
  - Estimating unobserved (zero) counts
  - Outlier detection

Artifacts vs. Error

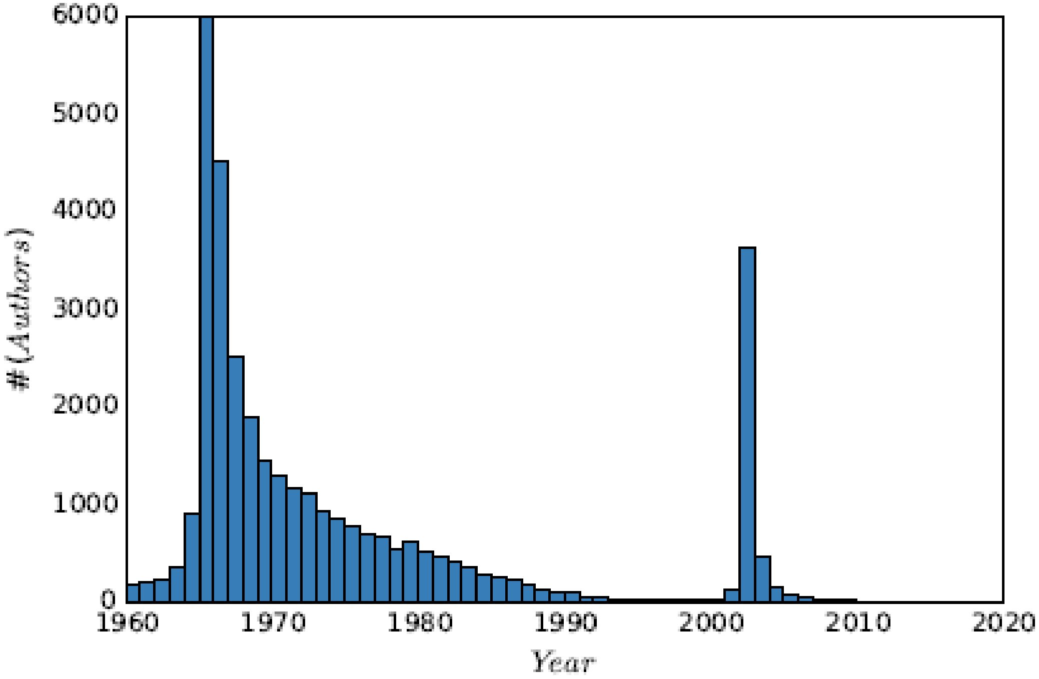
- data errors represent information that is fundamentally lost in acquisition.
  - The Gaussian noise blurring the resolution of our sensors represents error
  - The two hours of missing logs because the server crashed represents data error
- artifacts are generally systematic problems arising from processing done to the raw information

First-time Scientific Authors by Year?

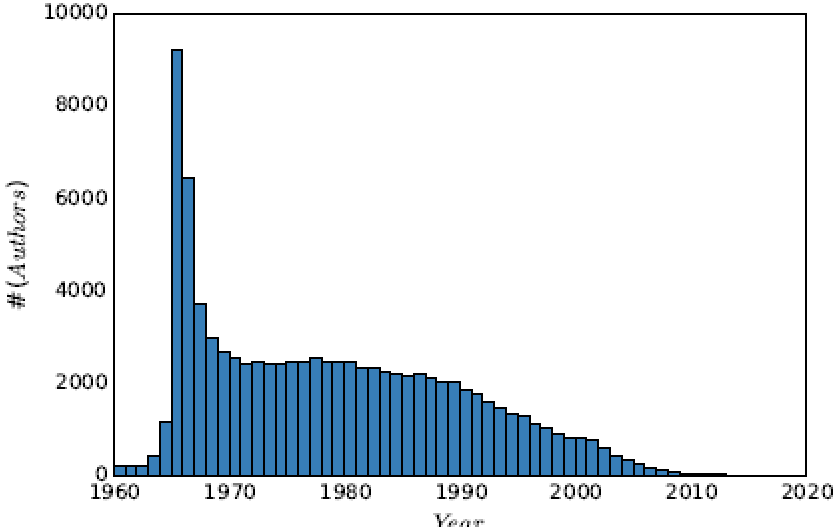
- In a bibliographic study, Skiena analyzed PubMed data to identify the year of first publication for the 100,000 most frequently cited authors.
- What *should* the distribution of new top authors by year look like?
- It is important to have a preconception of any result to help detect anomalies.

Might this be Right?

- What artifacts do you see?
- What possible explanations could cause them?



- 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 handled to make ``apple to apple’’ comparisons:

- Unit conversions
- Numerical Representation Conversions
- Name unification
- Time/date unification
- Financial unification

Unit Conversions

- It makes no sense to compare weights of 123.5 against 78.9, when one is in pounds and the other is in kilograms.
- It makes no sense to directly compare the movie gross of Gone with the Wind against that of Avatar, because 1939 dollars are 15.43 times more valuable than 2009 dollars.
- It makes no sense to compare the price of gold at noon today in New York and London, because the time zones are five hours off, and the prices affected by intervening events.
- It makes no sense to compare the stock price of Microsoft on February 17, 2003 to that of February 18, 2003, because the intervening 2-for-1 stock split cut the price in half, but reflects no change in real value.
- NASA lost the $125 million on September 23, 1999 due to a metric conversion issue.

Numerical Representation Conversions

- Numerical features are the easiest to incorporate into mathematical models
- But even turning numbers into numbers can have issue.
- Numerical fields might be represented in different ways: as integers (123), as decimals (123.5), or even as fractions (123 1/2). Numbers can even be represented as text, requiring the conversion from “ten million" to 10000000 for numerical processing.
- The Ariane 5 rocket exploded in 1996 due to a bad 64-bit float to 16-bit integer conversion.

Name Unification

- Database show Skiena’s publications as authored by the Cartesian product of his first (Steve, Steven, or S.), middle (Sol, S., or blank), and last (Skiena) names, allowing for nine different variations.
- And things get worse if we include misspellings (Skienna and Skeina)

- Use simple transformations to unify names, like lower case, removing middle names, etc
- Unify records by other parameters like co-authors, affiliation, nature of publication, field of research
- Tradeoff between false positives and negatives.

Date/Time Unification

- align all time measurements in same time zone.
- The Gregorian calendar is common throughout the technology world.
- 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.
- The other important correction is for inflation.
- A meaningful way to represent price changes over time is probably not differences but returns (in percentage).

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?

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.

- *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

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