Big Data Analysis with Apache Spark

APACHE Spark™

databricks™

amplab UC BERKELEY

Berkeley UNIVERSITY OF CALIFORNIA

BerkeleyX

This Lecture

Data Science Roles

Data Cleaning

Data Quality: Problems, Sources, and Continuum

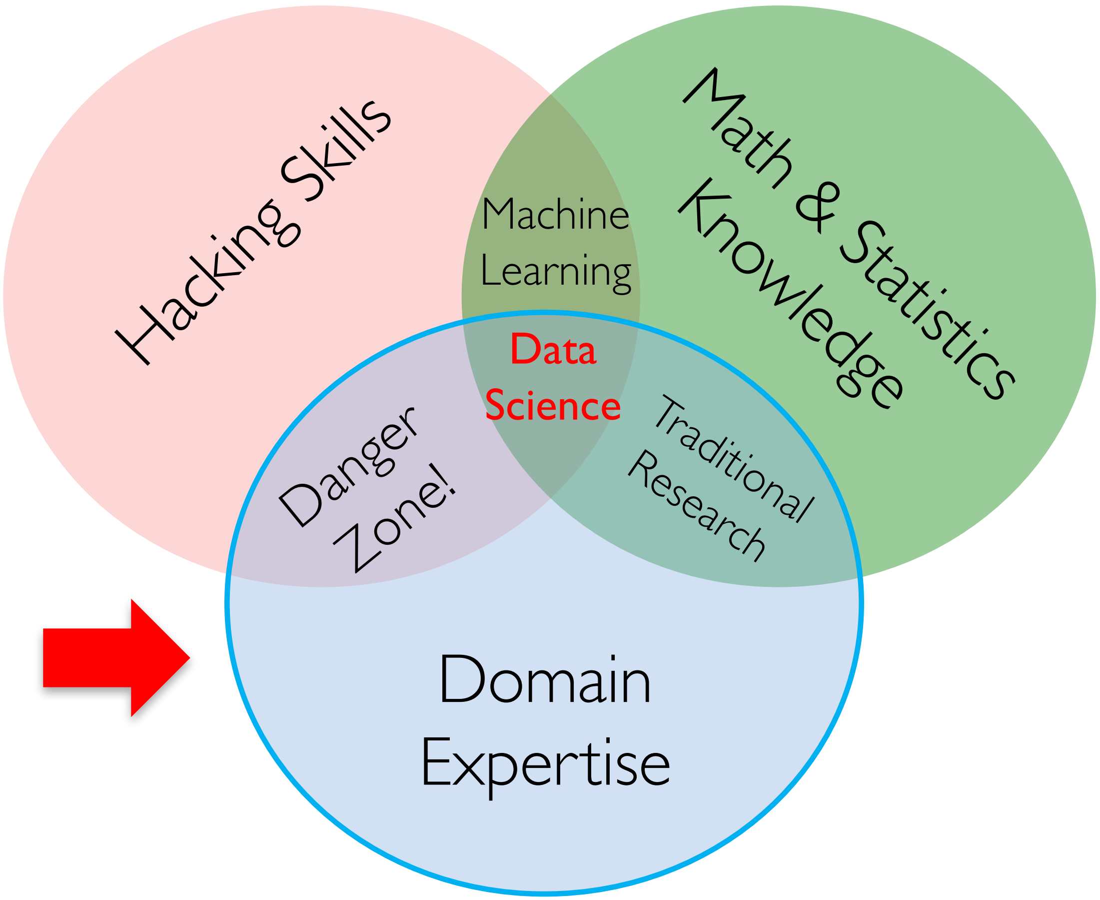
Data Gathering, Delivery, Storage, Retrieval, Mining/Analysis

Data Quality Constraints and Metrics

Data Integration

[redacted]

Data Science – One Definition



[redacted]

http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Data Science Roles

Businessperson

Programmer

Enterprise

Web Company

The Businessperson

Data Sources
- » Web pages
- » Excel

Extract-Transform-Load
- » Copy and paste

Data Warehouse
- » Excel

Business Intelligence and Analytics
- » Excel functions
- » Excel charts
- » Visual Basic

Image: http://www.fastcharacters.com/character-design/cartoon-business-man/

The Programmer

Data Sources

- Web scraping, web services API
- Excel spreadsheet exported as Comma Separated Values
- Database queries

Extract-Transform-Load

- wget, curl, Beautiful Soup, lxml

Data Warehouse

- Flat files

Business Intelligence and Analytics

- Numpy, Matplotlib, R, Matlab, Octave

Image: http://doctormo.deviantart.com/art/Computer-Programmer-Ink-346207753

The Enterprise

Data Sources
- » Application databases
- » Intranet files
- » Application server log files

Extract-Transform-Load
- » Informatica, IBM DataStage, Ab Initio, Talend

Data Warehouse
- » Teradata, Oracle, IBM DB2, Microsoft SQL Server

Business Intelligence and Analytics
- » SAP Business Objects, IBM Cognos, Microstrategy, SAS, SPSS, R

Image: http://www.publicdomainpictures.net/view-image.php?image=74743

The Web Company

Data Sources

- Application databases
- Logs from the services tier
- Web crawl data

Extract-Transform-Load

- Apache Flume, Apache Sqoop, Apache Pig, Apache Oozie, Apache Crunch

Data Warehouse

- Apache Hadoop/Apache Hive, Apache Spark/Spark SQL

Business Intelligence and Analytics

- Custom dashboards: Oracle Argus, Razorflow
- R, Apache Spark/Spark SQL

Image: http://www.future-web-net.com/2011/04/un-incendio-blocca-il-server-di-aruba.html

Data Cleaning

Helps deal with:

- » Missing data (ex: one dataset has humidity and other does not)
- » Entity resolution (ex: IBM vs. International Business Machines)
- » Unit mismatch (ex: $ versus £)
- » …

Dealing with Dirty Data – Statistics View

There is a ***process*** that produces data

- » Want to model ideal samples, but in practice have non-ideal samples
  - ***Distortion*** – some samples are corrupted by a process
  - ***Selection Bias*** - likelihood of a sample depends on its value
  - ***Left and Right Censorship*** - users come and go from our scrutiny
  - ***Dependence*** – samples are supposed to be independent, but are not (ex: social networks)

Add new models for each type of imperfection

- » Cannot model everything
- » What's the best trade-off between accuracy and simplicity?

Dirty Data – Database View

I got my hands on this data set

Some of the values are missing, corrupted, wrong, duplicated

Results are absolute (relational model)

You get a better answer by improving quality of values in dataset

Dirty Data – Domain Expert's View

This data doesn't look right

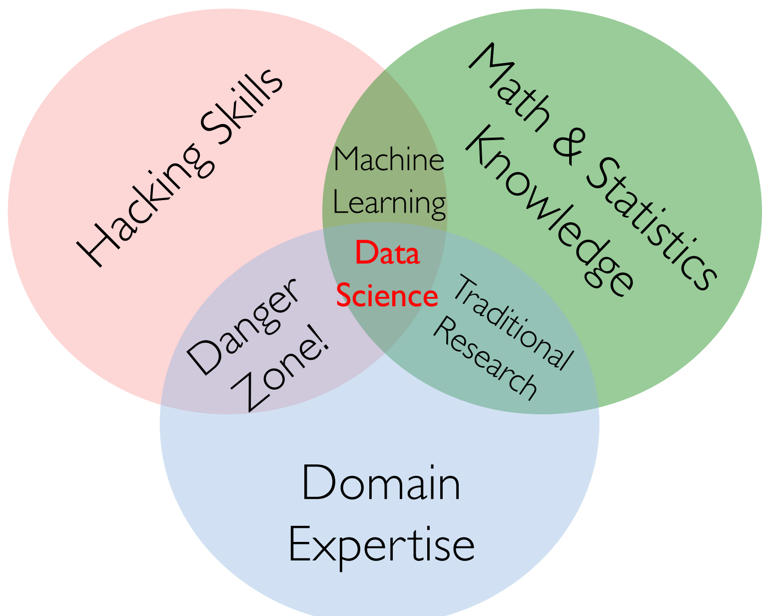
This answer doesn't look right

What happened?

Domain experts have implicit model of the data that they can test against…

Dirty Data – Data Scientist's View

Some Combination of all of the above



[redacted]

http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Data Quality Problems

(Source) Data is dirty on its own

Transformations corrupt data (complexity of software pipelines)

Clean datasets screwed up by integration (i.e., combining them)

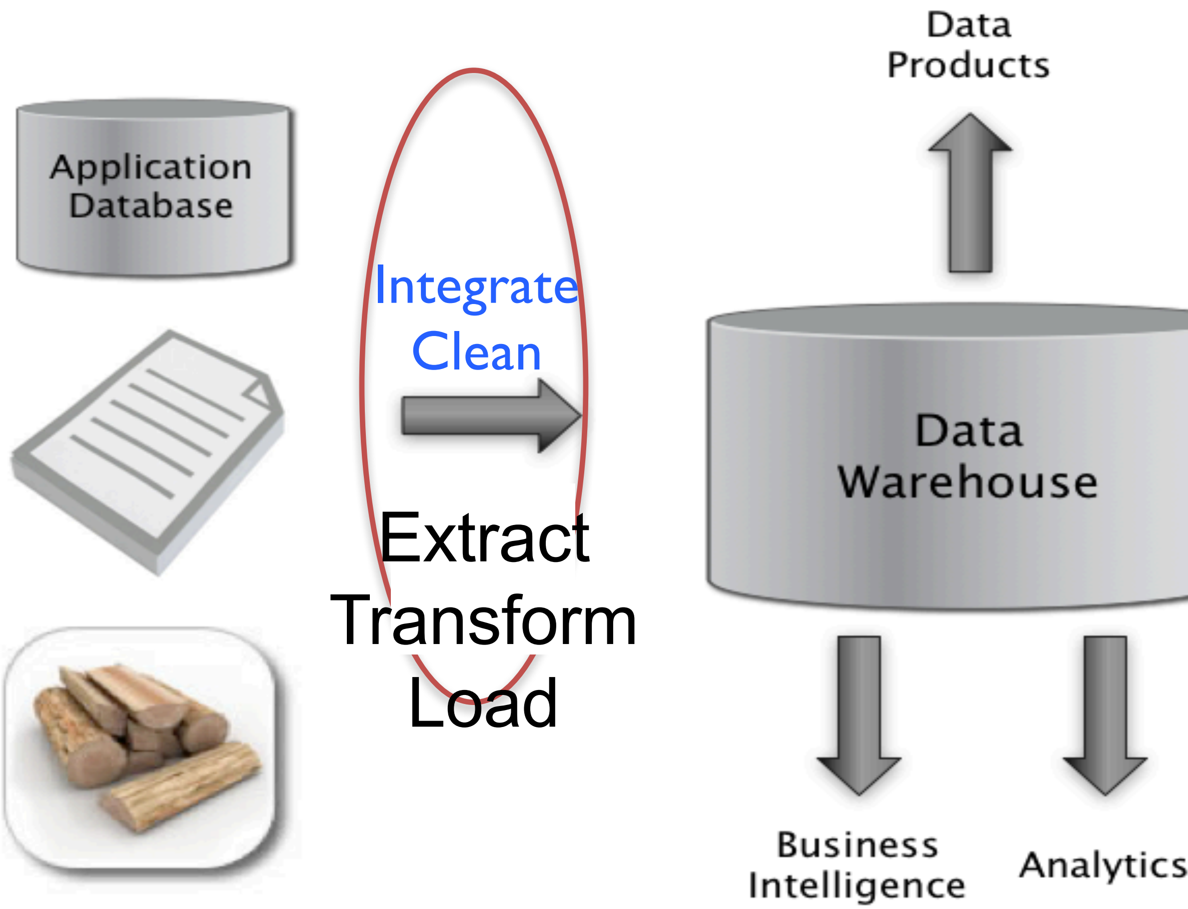
“Rare” errors can become frequent after transformation/integration

Clean datasets can suffer “bit rot”: data loses value/accuracy over time

Any combination of the above

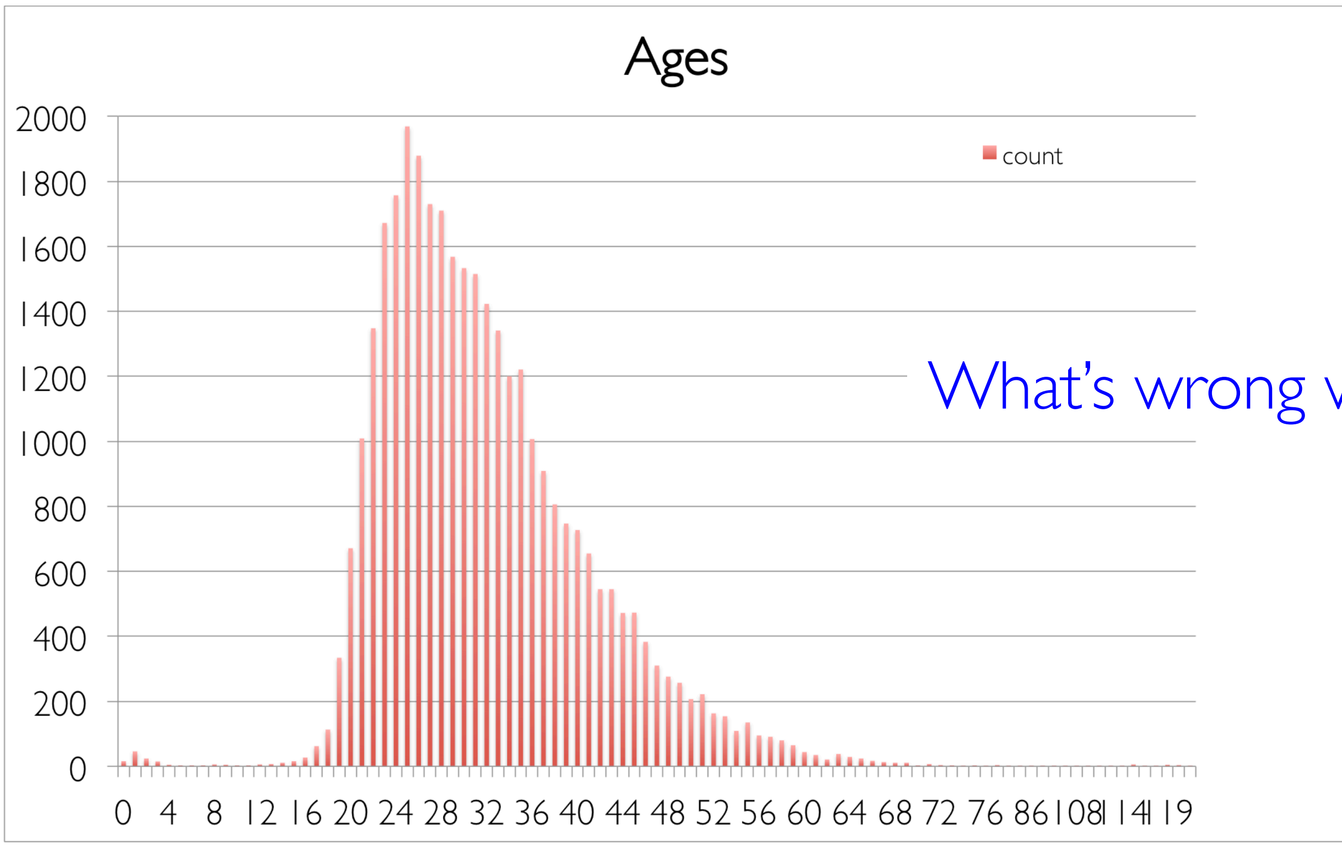
[redacted]

Where does Dirty Data Come from?



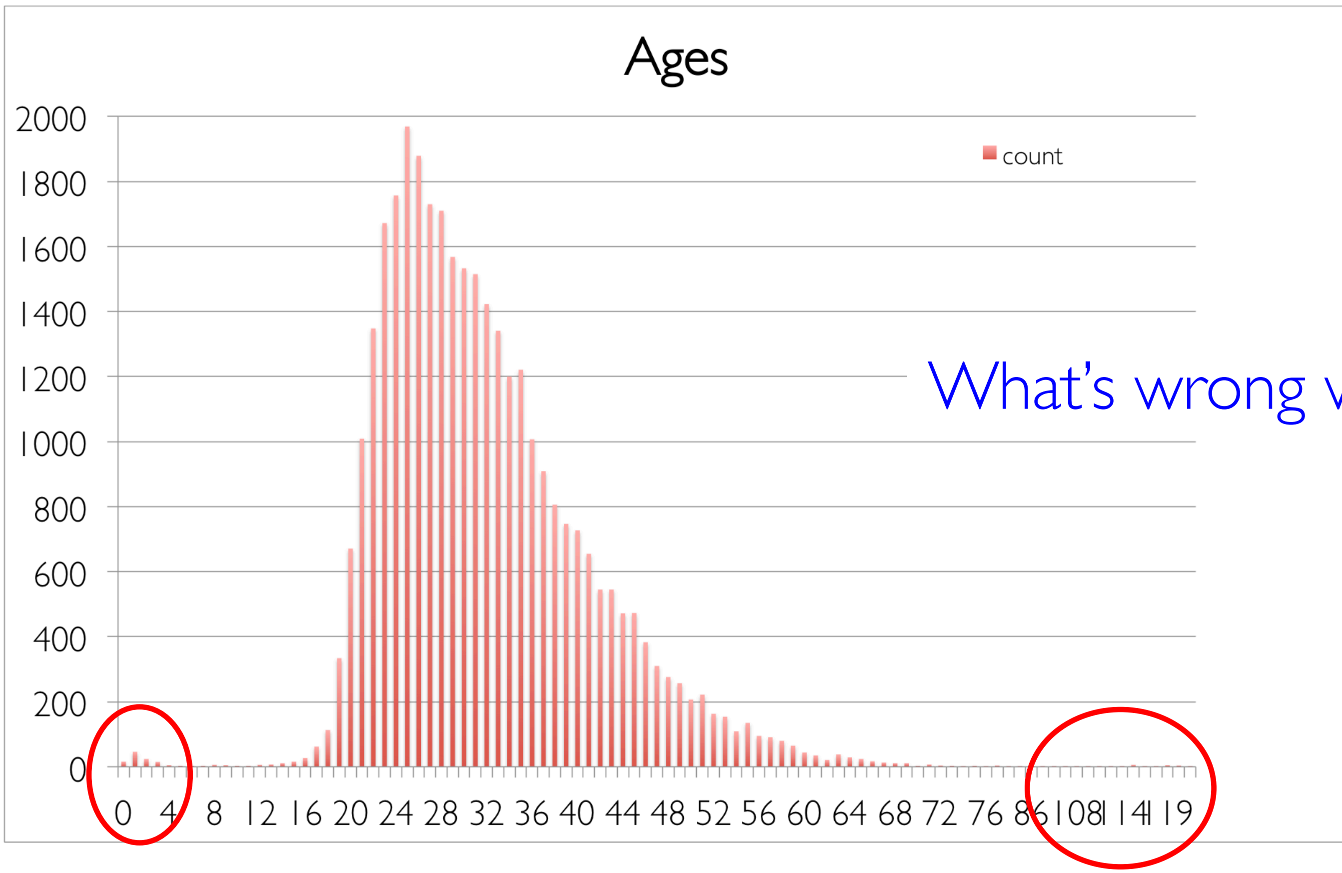
[redacted]

Ages of Students in an edX Course



What's wrong with this data?

Numeric Outliers



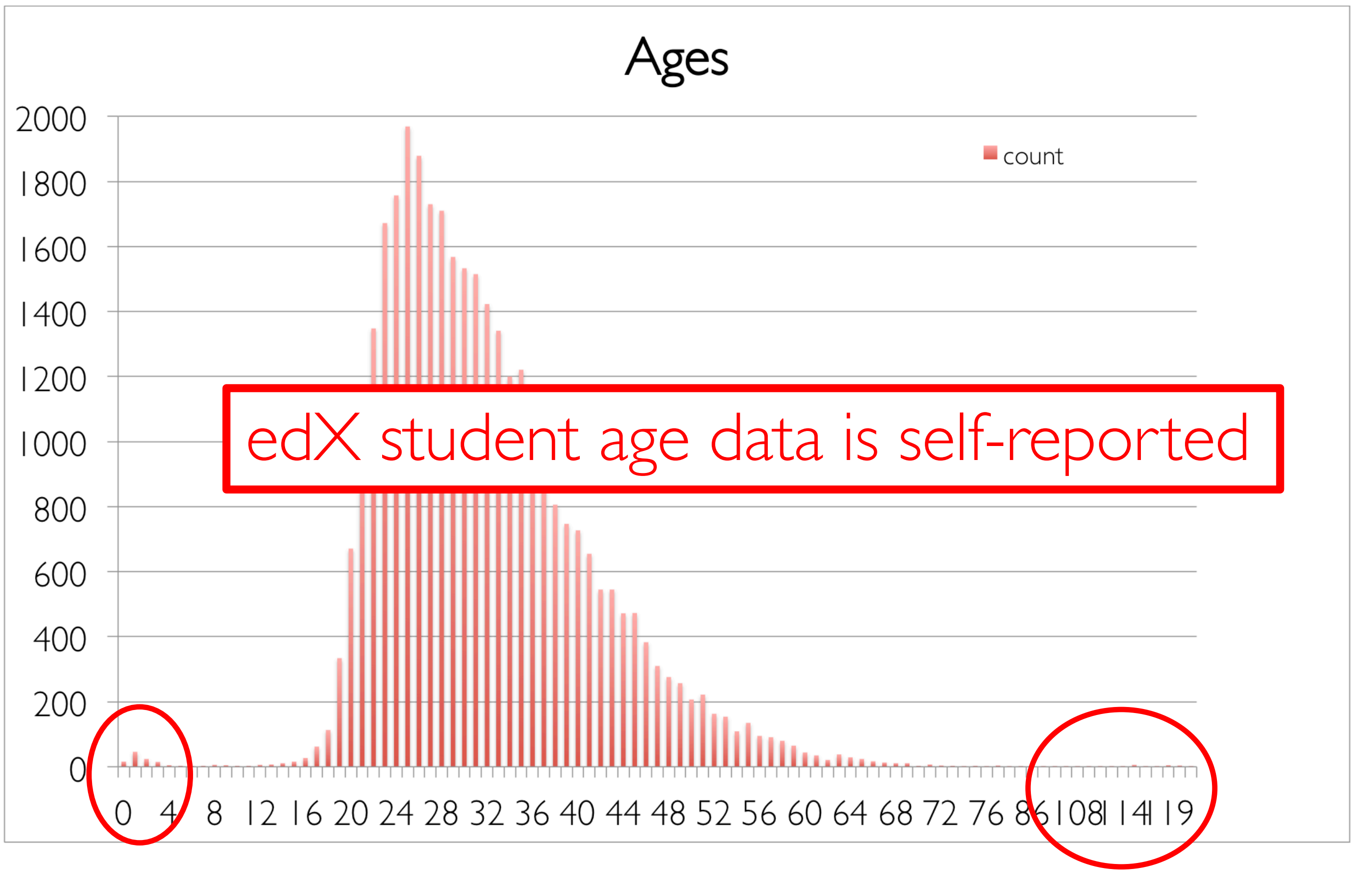
109 students ≤ 5

26 students > 100

What's wrong with this data?

[redacted]

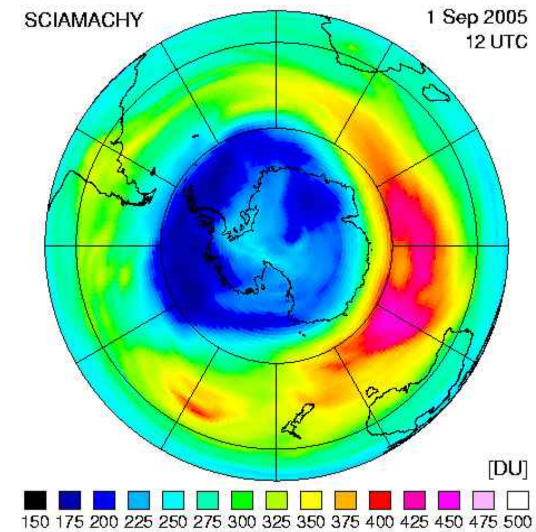
Numeric Outliers



109 students ≤ 5

26 students > 100

Data Cleaning Makes Everything Okay?

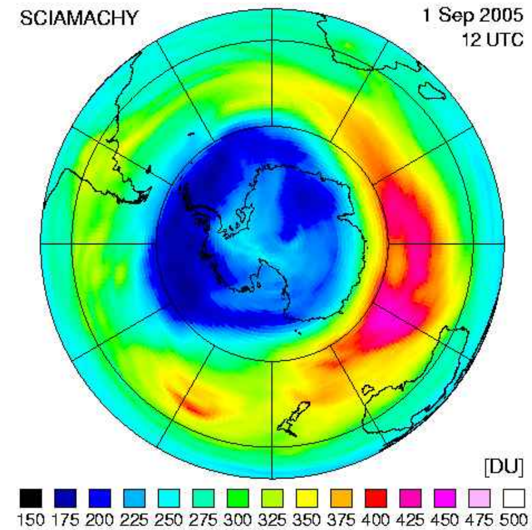


https://www.ucar.edu/learn/1_6_1.htm

Data Cleaning Makes Everything Okay?

"The appearance of a hole in the earth's ozone layer over Antarctica, first detected in 1976, was so unexpected that scientists didn't pay attention to what their instruments were telling them; they thought their instruments were malfunctioning."

National Center for Atmospheric Research

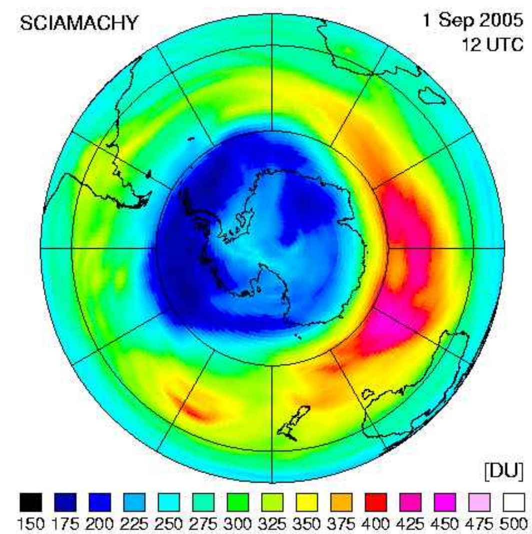


https://www.ucar.edu/learn/1_6_1.htm

Data Cleaning Makes Everything Okay?

“The appearance of a hole in the earth's ozone layer over Antarctica, first detected in 1976, was so unexpected that scientists didn't pay attention to what their instruments were telling them; they thought their instruments were malfunctioning.”

National Center for Atmospheric Research



In fact, the data were rejected as unreasonable by data quality control algorithms

https://www.ucar.edu/learn/1_6_1.htm

Dirty Data Problems

1. Parsing text into fields (separator issues)
2. Naming conventions (Entity Recognition: NYC vs. New York)
3. Missing required field (e.g., key field)
4. Primary key violation (from un- to structured or during integration
5. Licensing/Privacy issues prevent use of the data as you would like
6. Different representations (2 vs. Two)
7. Fields too long (get truncated)
8. Redundant Records (exact match or other)
9. Formatting issues – especially dates

From Stanford Data Integration Course

The Meaning of Data Quality

There are many uses of data

- » Operations, Aggregate analysis, Customer relations, …

Data Interpretation:

- » Data is useless if we don't know all of the *rules* behind the data

Data Suitability: Can you get answers from available data?

- » Use of proxy data
- » Relevant data is missing

[redacted]

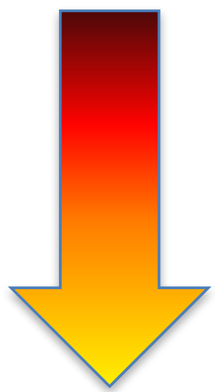
Adapted from [Ted Johnson's] SIGMOD 2003 Tutorial

The Data Quality Continuum

Data and information are not static

Flows in a data collection and usage process

» Data gathering
» Data delivery
» Data storage
» Data integration
» Data retrieval
» Data mining/analysis



Adapted from [Ted Johnson's]() SIGMOD 2003 Tutorial

Data Gathering

How does the data enter the system?

- » Experimentation, Observation, Collection

Sources of problems:

- » Manual entry
- » Approximations, surrogates – SW/HW constraints
- » No uniform standards for content and formats
- » Parallel data entry (duplicates)
- » Measurement or sensor errors

Adapted from [Ted Johnson's] SIGMOD 2003 Tutorial

Data Gathering – Potential Solutions

Preemptive:

» Process architecture (build in integrity checks)

» Process management (reward accurate data entry and sharing, provide data stewards)

Retrospective:

» Cleaning focus (duplicate removal, merge/purge, name/addr matching, field value standardization)

» Diagnostic focus (automated detection of glitches)

Adapted from Ted Johnson's SIGMOD 2003 Tutorial

Data Delivery

Destroying/mutilating information by bad pre-processing

- » Inappropriate aggregation
- » NULLs converted to default values

Loss of data:

- » Buffer overflows
- » Transmission problems
- » No checks

Adapted from [Ted Johnson's] SIGMOD 2003 Tutorial

Data Delivery – Potential Solutions

Build reliable transmission protocols: use a relay server

Verification: checksums, verification parser

» Do the uploaded files fit an expected pattern?

Relationships

» Dependencies between data streams and processing steps?

Interface agreements

» Data quality commitment from data supplier

Adapted from [Ted Johnson's] SIGMOD 2003 Tutorial

Data Storage

You get a data set – what do you do with it?

Problems in physical storage

» Potential issue but storage is cheap

Adapted from [Ted Johnson's](#) SIGMOD 2003 Tutorial

Data Storage

Problems in logical storage

» Poor metadata:

- Data feeds derived from programs or legacy sources – what does it mean?

» Inappropriate data models

- Missing timestamps, incorrect normalization, etc.

» Ad-hoc modifications.

- Structure the data to fit the GUI

» Hardware / software constraints.

- Data transmission via Excel spreadsheets, Y2K

Adapted from [Ted Johnson's](#) SIGMOD 2003 Tutorial

Data Storage – Potential Solutions

Metadata: document and publish data specifications

Planning: assume that everything bad will happen

» Can be very difficult to anticipate all problems

Data exploration

» Use data browsing and data mining tools to examine the data

- Does it meet the specifications you assumed?
- Has something changed?

Adapted from Ted Johnson's SIGMOD 2003 Tutorial

Data Retrieval

Exported data sets are often a view of the actual data

» Problems occur because:

- Source data or need for derived data not properly understood
- Just plain mistakes: inner join vs. outer join, not understanding NULL values

Computational constraints: Full history too expensive

» Supply limited snapshot instead

Incompatibility: ASCII? Unicode? UTF-8?

Adapted from [Ted Johnson's] SIGMOD 2003 Tutorial

Data Mining and Analysis

What are you doing with all this data anyway?

Problems in the analysis

- » Scale and performance
- » Confidence bounds?
- » Black boxes and dart boards
- » Attachment to models
- » Insufficient domain expertise
- » Casual empiricism (use arbitrary number to support a pre-conception)

Adapted from [Ted Johnson's] SIGMOD 2003 Tutorial

Retrieval and Mining – Potential Solutions

Data exploration

- » Determine which models and techniques are appropriate
- » Find data bugs
- » Develop domain expertise

Continuous analysis

- » Are the results stable? How do they change?

Accountability

- » Make the analysis part of the feedback loop

Adapted from Ted Johnson's SIGMOD 2003 Tutorial

Data Quality Constraints

Capture many data quality problems using schema's static constraints

» NULLs not allowed, field domains, foreign key constraints, etc.

Many other quality problems are due to problems in workflow

» Can be captured by *dynamic* constraints

» E.g., orders above $200 are processed by Biller 2

The constraints follow an 80-20 rule

» A few constraints capture most cases,

» Thousands of constraints to capture the last few cases

Constraints are measurable – data quality metrics?

Adapted from [Ted Johnson's](#) SIGMOD 2003 Tutorial

Data Quality Metrics

We want a measurable quantity

- » Indicates what is wrong and how to improve
- » Realize that DQ is a messy problem, no set of numbers will be perfect

Metrics should be directionally correct with improvement in data use

Types of metrics

- » Static vs. dynamic constraints
- » Operational vs. diagnostic

A very large number metrics are possible

- » Choose the most important ones

[redacted]

Adapted from [Ted Johnson's] SIGMOD 2003 Tutorial

Examples of Data Quality Metrics

Conformance to schema: evaluate constraints on a snapshot

Conformance to business rules: evaluate constraints on DB changes

Accuracy: perform expensive inventory or track complaints (proxy)

- » Audit samples?

Accessibility

Interpretability

Glitches in analysis

Successful completion of end-to-end process

[redacted]

Adapted from [Ted Johnson's]() SIGMOD 2003 Tutorial

Technical Approaches

Use multi-disciplinary approach to attack data quality problems

» *No one approach solves all problems*

Process Management: ensure proper procedures

Statistics: focus on analysis – find and repair anomalies in data

Database: focus on relationships – ensure consistency

Metadata / Domain Expertise

» What does data mean? How to interpret?

[redacted]

Adapted from [Ted Johnson's] SIGMOD 2003 Tutorial

Data Integration

Combine data sets (acquisitions, across departments)

Common source of problems

» Heterogeneous data : no common key, different field formats

- Approximate matching

» Different definitions: what is a customer – acct, individual, family?

» Time synchronization

- Does the data relate to the same time periods?
- Are the time windows compatible?

» Legacy data: spreadsheets, ad-hoc structures

Adapted from Ted Johnson's SIGMOD 2003 Tutorial

Duplicate Record Detection (DeDup)

- Resolve multiple different entries:
  - » Entity resolution, reference reconciliation, object ID/consolidation
- Remove Duplicates: Merge/Purge
- Record Linking (across data sources)
- Approximate Match (accept fuzziness)
- Householding (special case)
  - » Different people in same house?
- …

Example: Entity Resolution

Web scrape of Google Shopping and Amazon product listings

Google listing:

- » clickart 950000 - premier image pack (dvd-rom) massive collection of images & fonts for all your design needs ondvd-rom!product informationinspire your creativity and perfect any creative project with thousands ofworld-class images in virtually every style. plus clickart 950000 makes iteasy for …

Amazon listing:

- » clickart 950 000 - premier image pack (dvd-rom)

Are these two listings the same product?

https://code.google.com/p/metric-learning/

Example: Entity Resolution

Web scrape Google Shopping and Amazon product listings

Google listing:

» clickart 950000 - premier image pack (dvd-rom) massive collection of images & fonts for all your design needs ondvd-rom!product informationinspire your creativity and perfect any creative project with thousands ofworld-class images in virtually every style. plus clickart 950000 makes iteasy for …

Amazon listing:

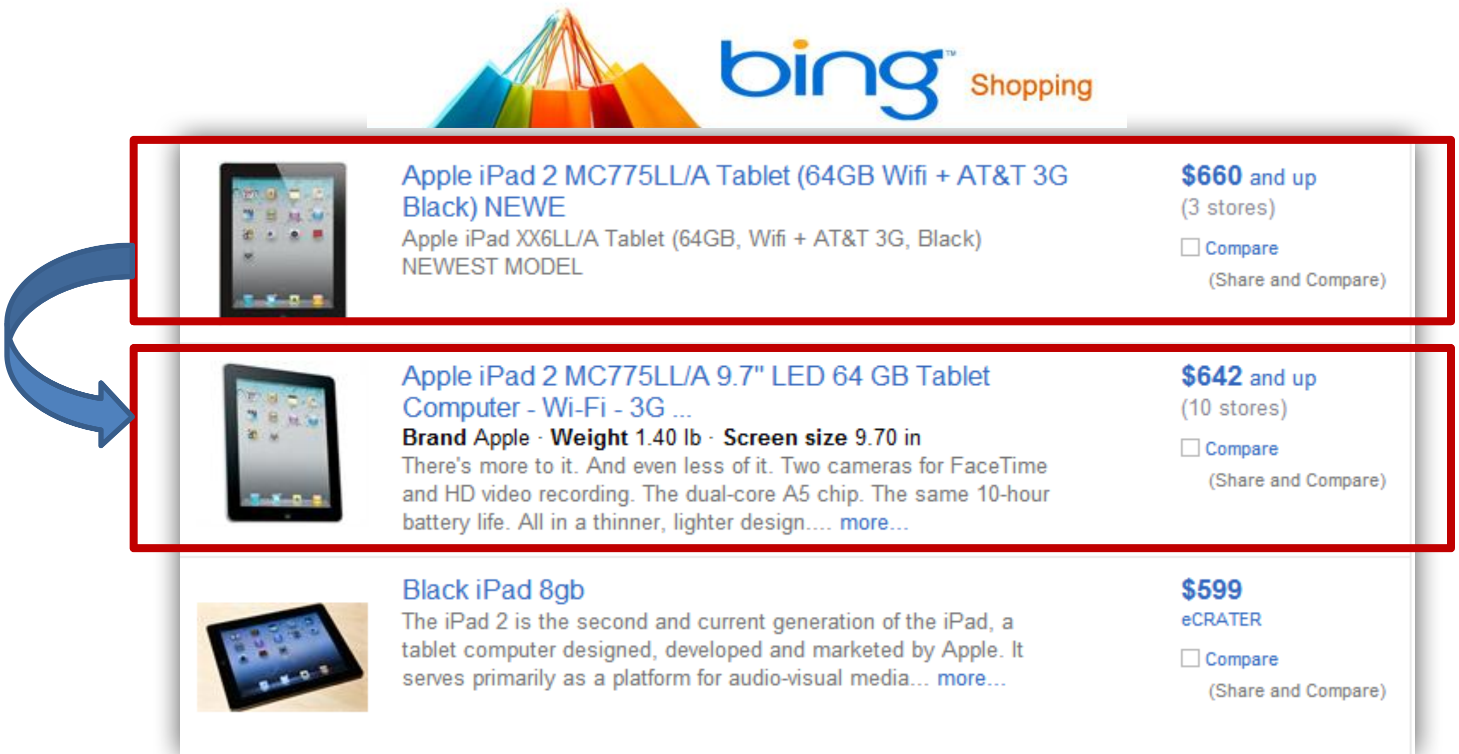
» clickart 950 000 - premier image pack (dvd-rom)

Are these two listings the same product?

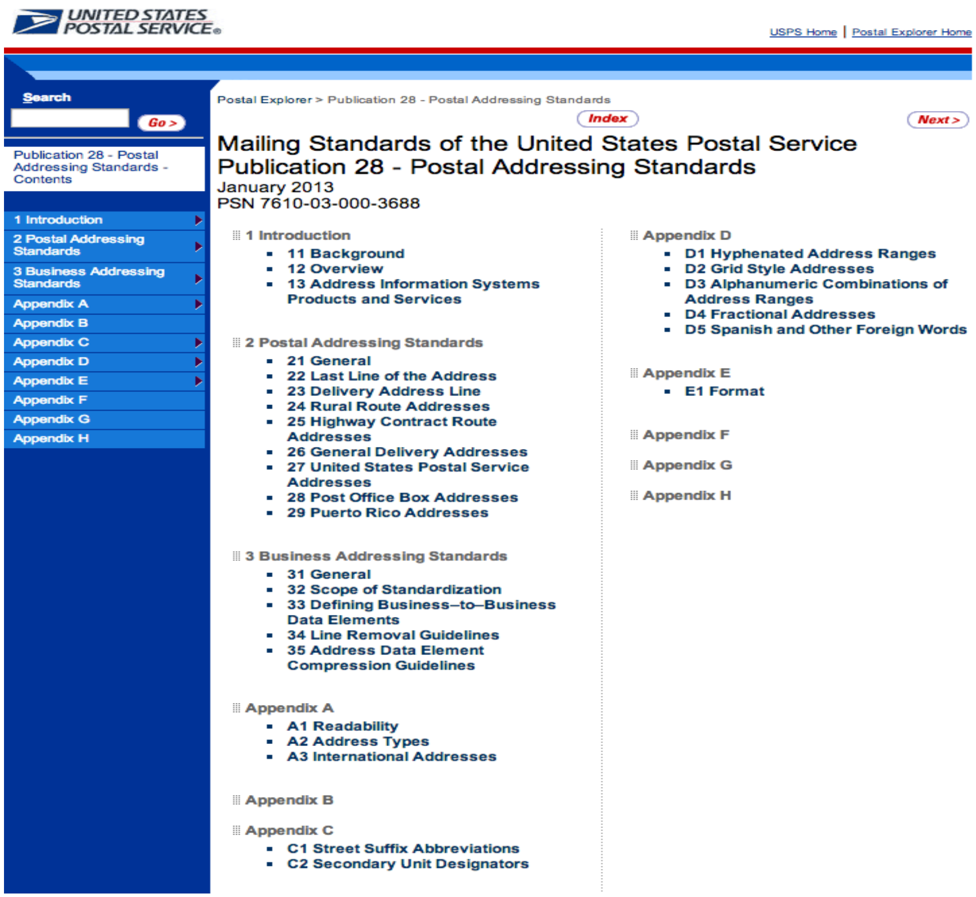
YES! Algorithmic approach in the Lab

https://code.google.com/p/metric-learning/

Example: DeDup/Cleaning



Preprocessing/Standardization


UNITED STATES POSTAL SERVICE®

USPS Home | Postal Explorer Home

Search

Go >

Publication 28 - Postal Addressing Standards - Contents

1 Introduction
2 Postal Addressing Standards
3 Business Addressing Standards
Appendix A
Appendix B
Appendix C
Appendix D
Appendix E
Appendix F
Appendix G
Appendix H

Postal Explorer > Publication 28 - Postal Addressing Standards

Index

Next >

Mailing Standards of the United States Postal Service
Publication 28 - Postal Addressing Standards
January 2013
PSN 7610-03-000-3688

Simple idea:
- » Convert to canonical form
- » Example: mailing addresses

More Sophisticated Techniques

Use evidence from multiple fields

» Positive and Negative instances are possible

Use evidence from linkage pattern with other records

Clustering-based approaches

…

[redacted]

Lots of Additional Problems

Address vs. Number, Street, City, …

Units

Differing Constraints

Multiple versions and schema evolution

Other Metadata

[redacted]

Data Integration – Solutions

Commercial Tools

- » Significant body of research in data integration
- » Many tools for address matching, schema mapping are available.

Data browsing and exploration

- » Many hidden problems and meanings: must extract metadata
- » View before and after results:
    - Did the integration go the way you thought?

Adapted from [Ted Johnson's]() SIGMOD 2003 Tutorial

Estimation

Statistical Inference:

Making conclusions based on data in random samples

Estimation

Statistical Inference:

Making conclusions based on data in random samples

Example:

Use the data to guess the value of an unknown number

[redacted]

Estimation

Statistical Inference:

Making conclusions based on data in random samples

fixed

Example:

Use the data to guess the value of an <u>unknown number</u>

[redacted]

Estimation

Statistical Inference:

Making conclusions based on data in random samples

fixed

Example:

Use the data to guess the value of an <u>unknown number</u>

Create an **estimate** of the unknown quantity

[redacted]

Estimation

Statistical Inference:

Making conclusions based on data in random samples

Example:

fixed

Use the data to guess the value of an <u>unknown number</u>

[redacted]

Create an **estimate** of the unknown quantity

depends on the random sample

How Many Enemy Warplanes?

Example:

» During WWII, the allies wanted to determine the number of German fighter planes

https://www.flickr.com/photos/fun_flying/250148463

How can we determine the number of fighter planes?

Assumptions

Planes have serial numbers 1, 2, 3, …, N

We don't know N

Want estimate N based on serial numbers of planes we see

The main assumption

» The serial numbers of planes we see are a uniform random sample drawn with replacement from 1, 2, 3, …, N

[redacted]

Estimation

If you saw these serial numbers, what would be your estimate of N?

170	271	285	290	48
235	24	90	291	19

[redacted]

Estimation

If you saw these serial numbers, what would be your estimate of N?

```
170  271  285  290  48
235   24   90  291  19
```

One idea: `291` – just guess the largest one

[redacted]

The Largest Number Observed

Is it likely to be close to N?

- How likely?
- How close?

Some options:

- Could try to calculate probabilities and draw a probability histogram
- Could simulate and draw an empirical histogram

[redacted]

Verdict on the Estimate

The largest serial number observed is likely to be close to N

But, it is also likely to underestimate N

Another idea for an estimate:

» Average of the serial numbers observed ~ N/2

[redacted]

New estimate: 2 times the average of seen

Bias

Biased estimate: On average across all possible samples, the estimate is either too high or too low

Bias creates a systematic error in one direction

Good estimates typically have low bias

[redacted]

Variability

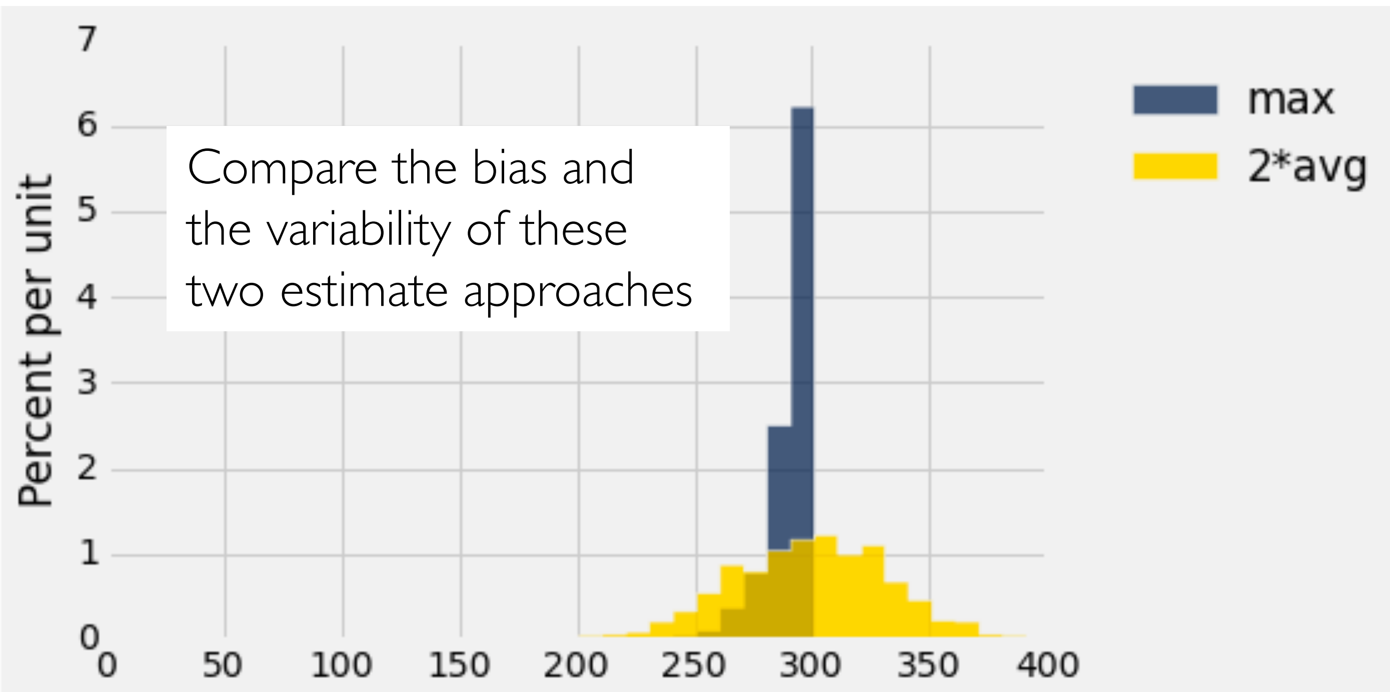
The value of an estimate **varies** from one sample to another

High variability makes it hard to estimate accurately

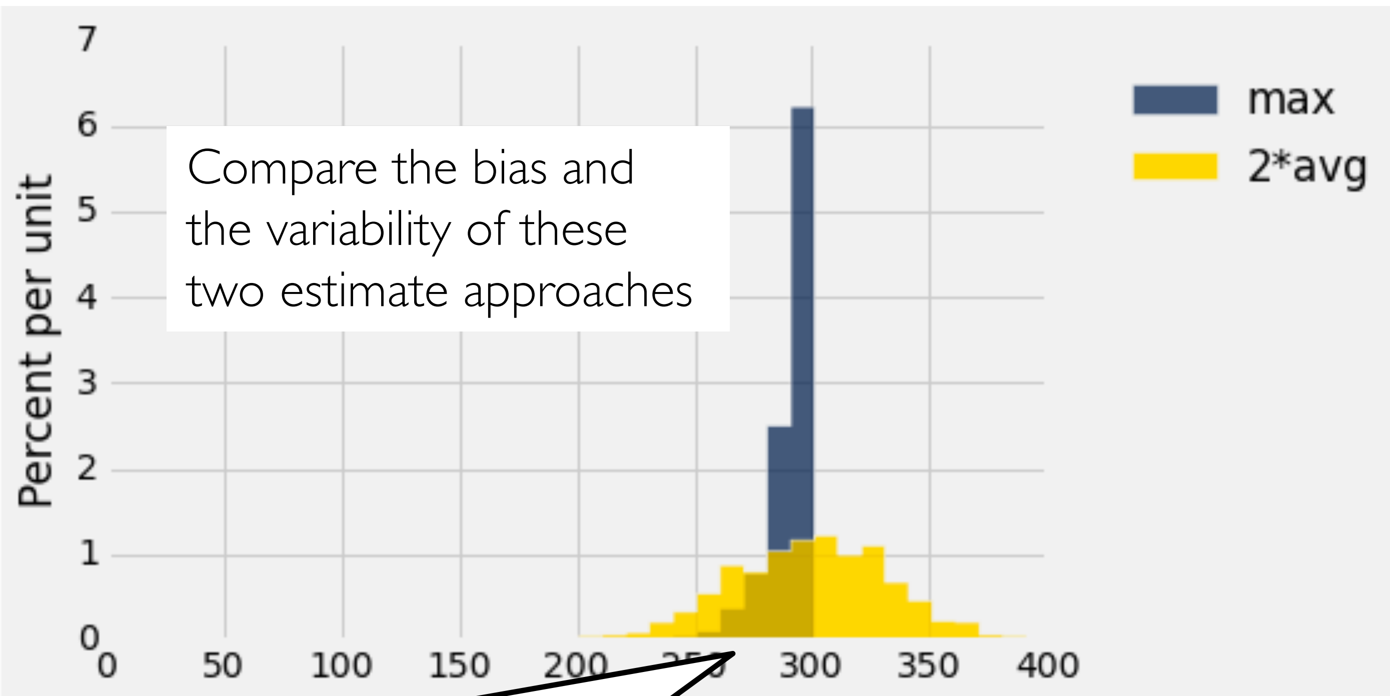
Good estimates typically have low variability

[redacted]

Observation Example



Observation Example

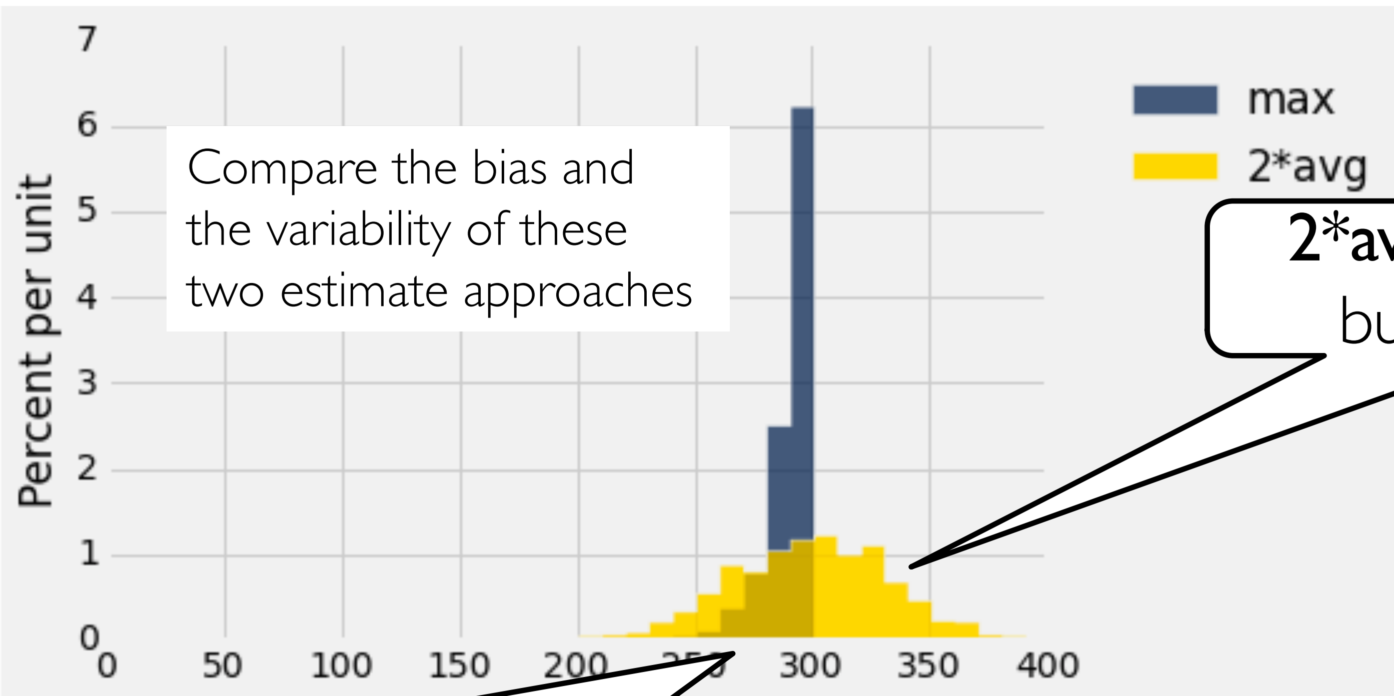


Compare the bias and the variability of these two estimate approaches

max has low variability, but is biased (left)

[redacted]

Observation Example



Compare the bias and the variability of these two estimate approaches

2*avg has little bias, but is highly variable

max has low variability, but is biased (left)

[redacted]

Bias-Variance Tradeoff

The max has low variability, but it is biased

» It under estimates the number of planes

2*average has little bias, but it is highly variable

» It under/over estimates the number of planes, typically more than max does

Which one to choose?

» Pick based on your utility?

[redacted]

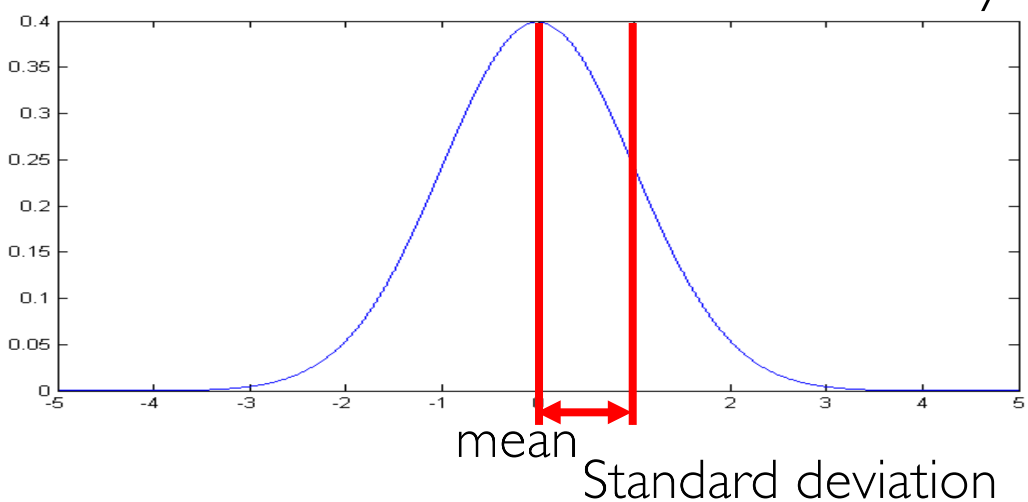
Normal Distributions, Mean, Variance

The **Mean** of a set of values is the average of the values

Variance is a measure of the width of a distribution

The **Standard Deviation** is the square root of variance

A **Normal Distribution** is characterized by mean and variance



Properties of the Mean

Balance point of the histogram

» **Not** the "half-way point" of the data (**median**)

If the histogram is skewed, then the **mean** is pulled away from the **median** in the direction of the tail

Defining Variability

Plan A: "biggest value - smallest value"

» Doesn't provide information about the shape of the distribution

Plan B:

» Measure variability around the mean

» Need to figure out a way to quantify this

[redacted]

Standard Deviation

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Root mean square of deviations from average

Measures roughly how far off the values are from average

[redacted]

The SD and the Histogram

Usually not easy to estimate SD by looking at a histogram

» But if the histogram has a special shape, then maybe

If a histogram is bell-shaped, then

» Average is at the center

» SD is distance between average and points of inflection on either side

How Large are Most of the Values?

No matter what the shape of the distribution

» The bulk of the data are in the range "average ± a few SDs"

If a histogram is bell-shaped

» SD is distance between avg and points of inflection on either side

» Almost all of the data are in the range

» "Average ± 3 SDs"

Central Limit Theorem

If the samples are

» A large set

» Drawn at random with replacement

Then, no matter what the distribution of the population:

» **Probability distribution of the sample average is roughly bell-shaped (normal distribution)**

[redacted]

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \quad -\infty < z < \infty$$

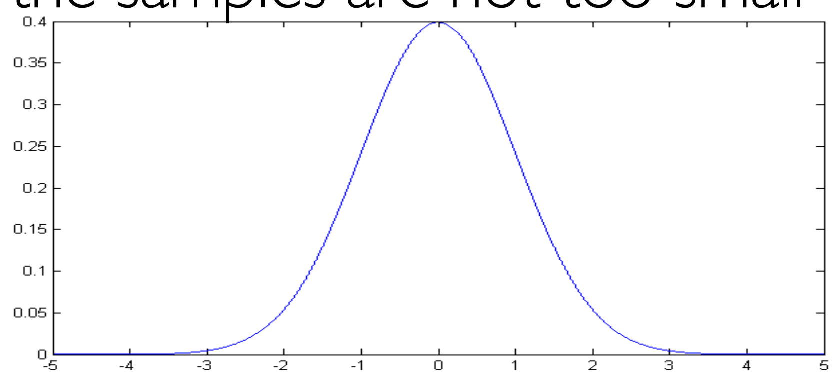
The standard normal curve

Central Limit Theorem

The distribution of sum (or mean) of n identically-distributed random variables X_i approaches a normal distribution as $n \to \infty$

Common parametric statistical tests (t-test & ANOVA) assume normally-distributed data, but depend on sample mean and variance

Tests work reasonably well for data that are not normally distributed as long as the samples are not too small



Bounds and Normal Approximations

Percent in Range	All Distributions	Normal Distribution
Average ± 1 SD	At least 0%	About 68%
Average ± 2 SDs	At least 75%	About 95%
Average ± 3 SDs	At least 88.888...%	About 99.73%

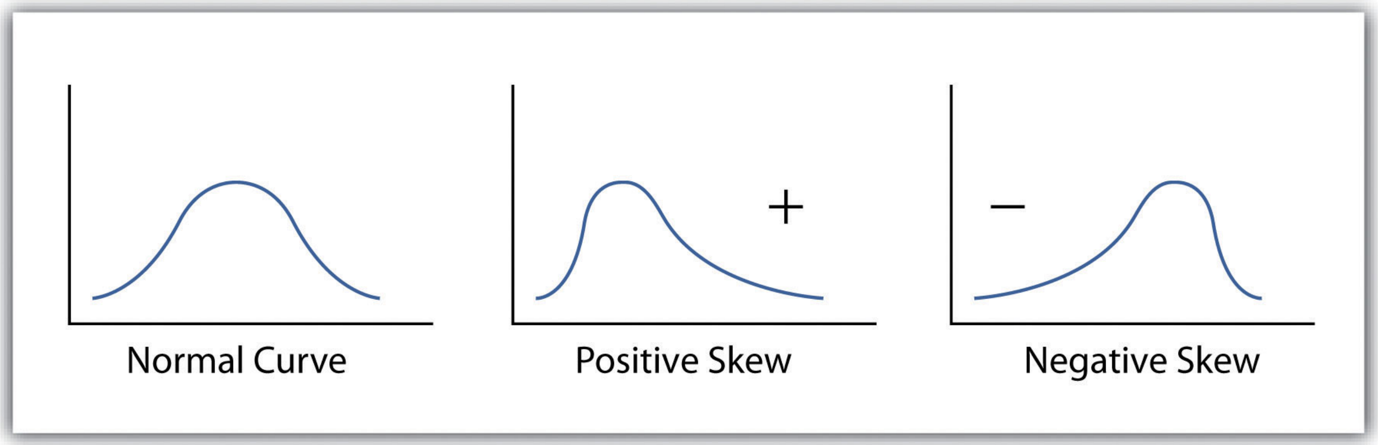
Chebychev's Inequality

No matter what the shape of the distribution, the proportion of values in the range "average ± z SDs" is at least $1 - 1/z^2$

Correcting Distributions

Many statistical tools (mean, variance, t-test, ANOVA) assume data are normally distributed

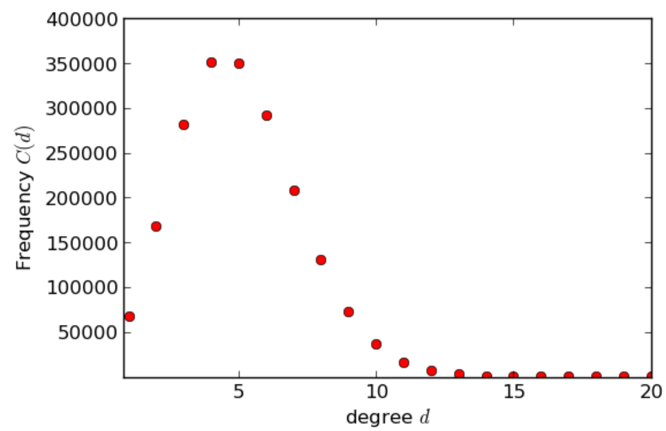
Very often this is not true – examine the histogram



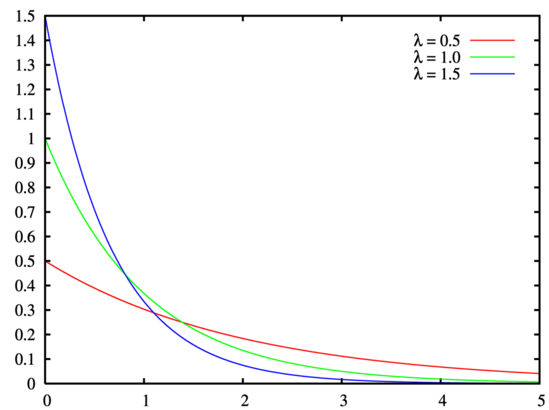
Other Important Distributions

Poisson: distribution of counts that occur at a certain "rate"

- Observed frequency of a given term in a corpus
- Number of visits to web site in a fixed time interval
- Number of web site clicks in an hour



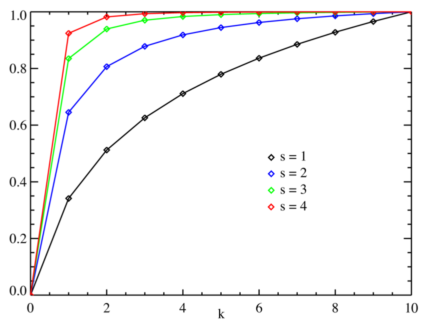
Exponential: interval between two such events

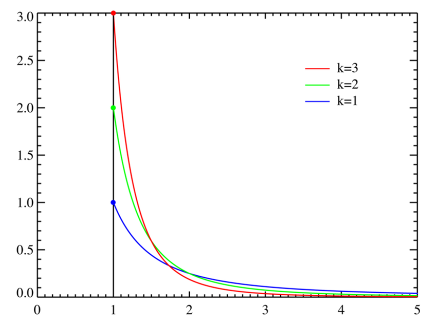


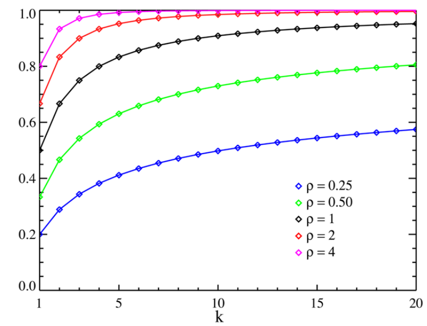
Other Important Distributions

Zipf/Pareto/Yule distributions:

» Govern frequencies of different terms in a document, or web site visits



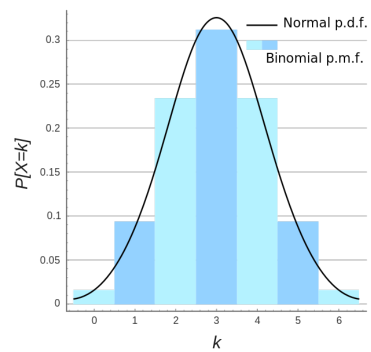




Binomial/Multinomial:

» Number of counts of events

» Example: 6 die tosses out of n trials



[Understand your data's distribution before applying any model]()

Spark's Machine Learning Toolkit

mllib and ML Pipelines: scalable, distributed ML libraries

» Scikit-learn like ML toolkit, Interoperates with NumPy

Classification:

» SVM, Logistic Regression, Decision Trees, Naive Bayes, …

Regression: Linear, Lasso, Ridge, …

Miscellaneous:

» Alternating Least Squares, K-Means, SVD

» Optimization primitives (SGD, L-BGFS)

» …

[redacted]

Lab: Collaborative Filtering

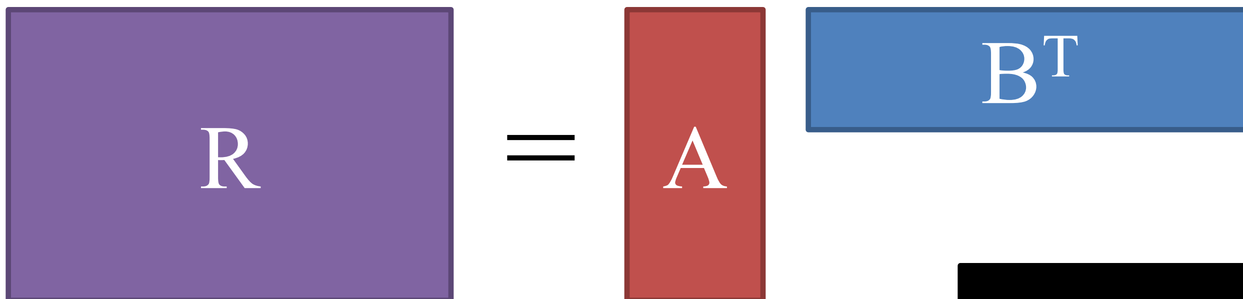
Goal: predict users' movie ratings based on past ratings of other movies

$$\textbf{\textit{Ratings}} = \begin{array}{cc} \text{Movies} & \\ \begin{bmatrix} 1 & ? & ? & 4 & 5 & ? & 3 \\ ? & ? & 3 & 5 & ? & ? & 3 \\ 5 & ? & 5 & ? & ? & ? & 1 \\ 4 & ? & ? & ? & ? & 2 & ? \end{bmatrix} & \text{Users} \end{array}$$

[redacted]

Model and Algorithm

Model ***Ratings*** as product of ***User (A)*** and ***Movie Feature (B)*** matrices of size U × K and M × K



K: ***rank***

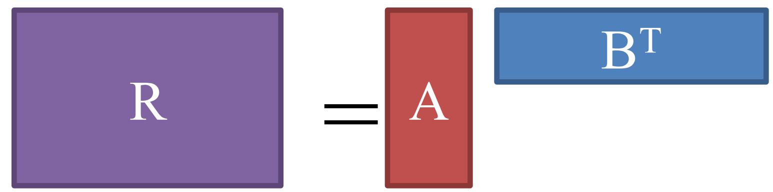
Learn *K* factors for each user

Learn *K* factors for each movie

[redacted]

Model and Algorithm

Model ***Ratings*** as product of ***User (A)*** and ***Movie Feature (B)*** matrices of size U × K and M × K



Alternating Least Squares (ALS)

» Start with random A and B vectors

» Optimize user vectors (A) based on movies

» Optimize movie vectors (B) based on users

» Repeat until converged