Introduction to Big Data with Apache Spark

This Lecture

Data Cleaning

Data Quality: Problems, Sources, and Continuum

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

Data Quality Constraints and Metrics

Data Integration

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

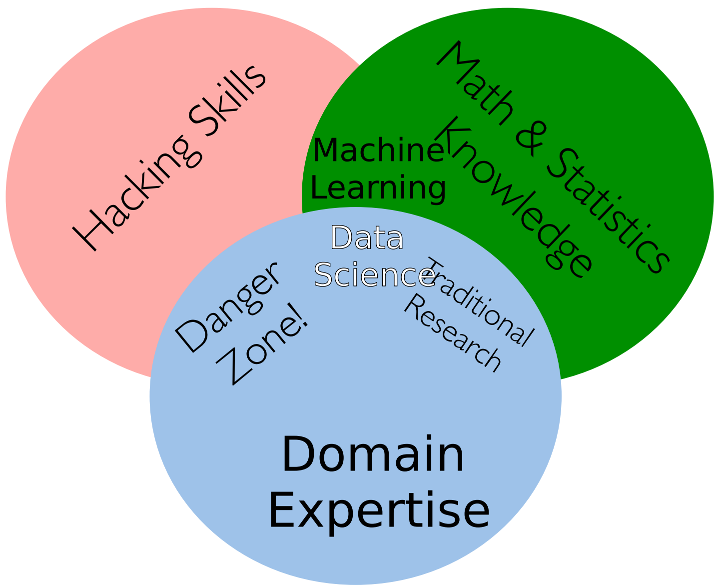
Dirty Data – Domain Expert's View

- 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

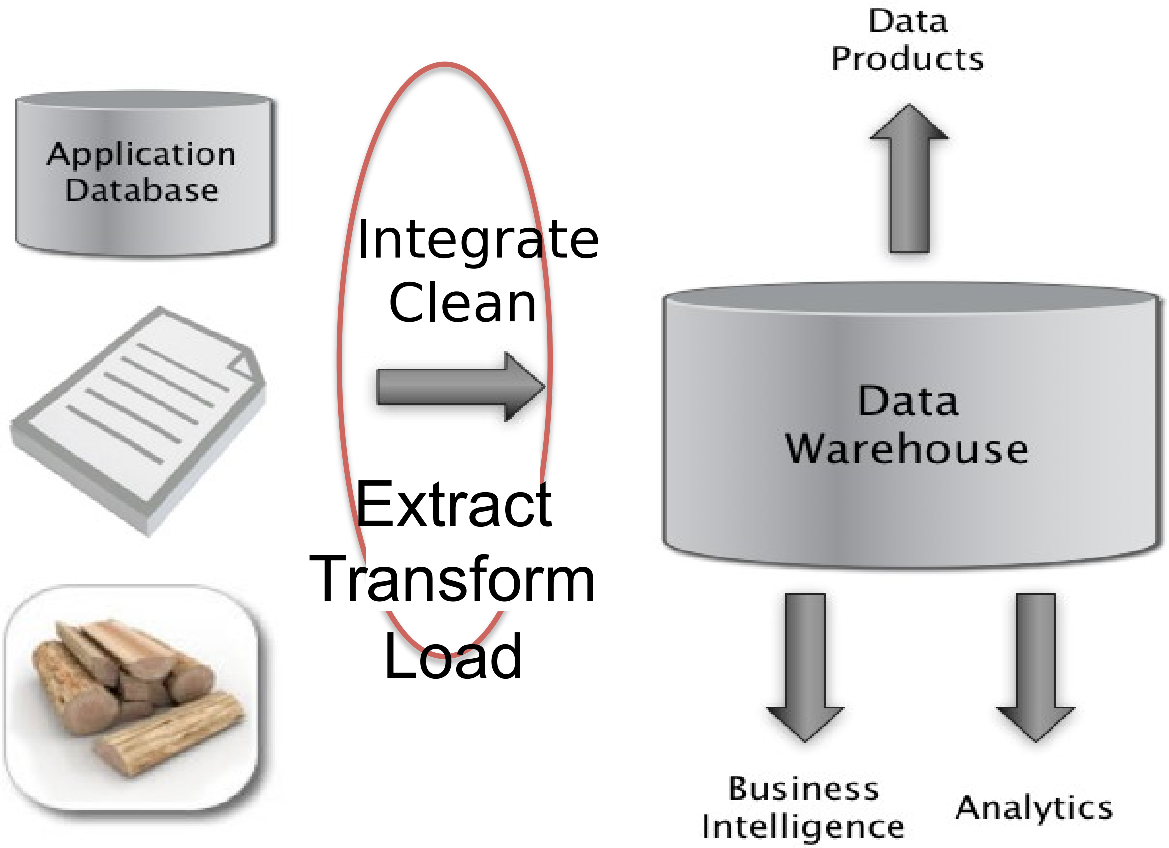


http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

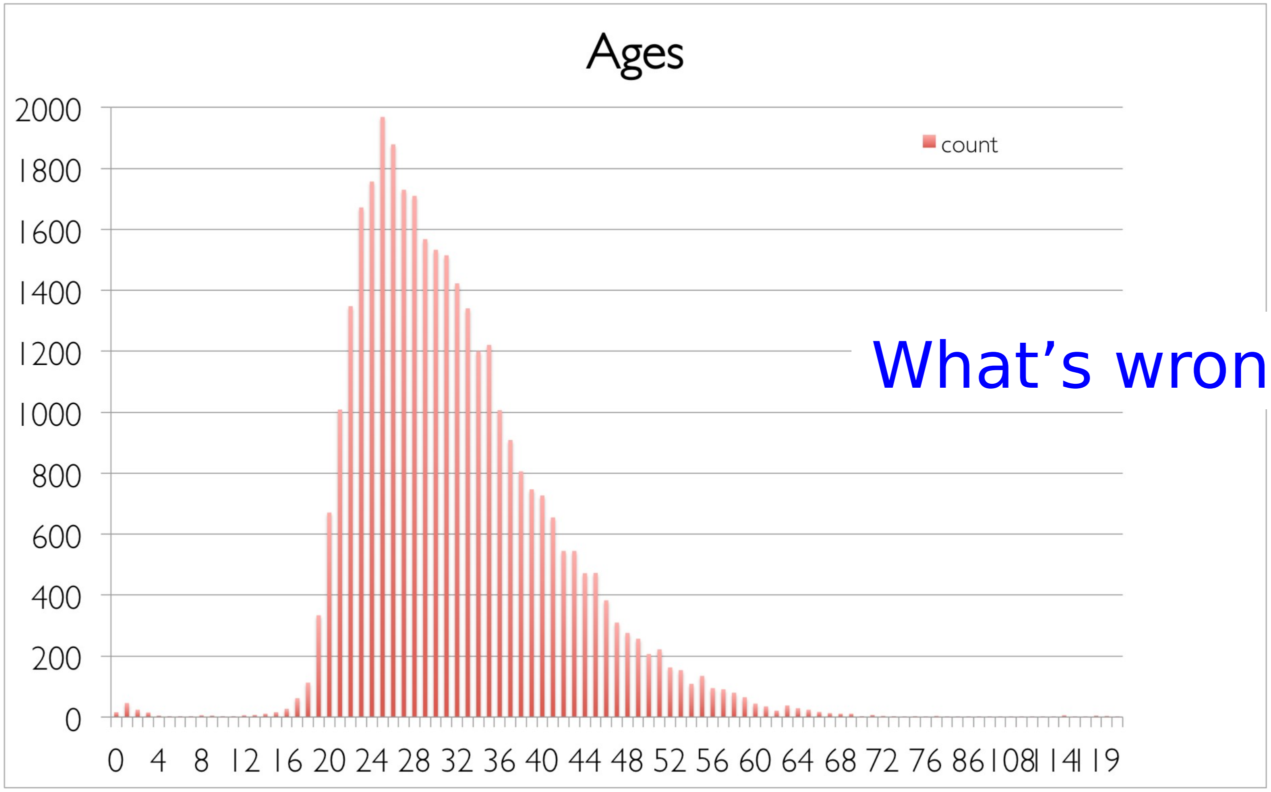
Data Quality Problems

- (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*

Where does Dirty Data Come from?

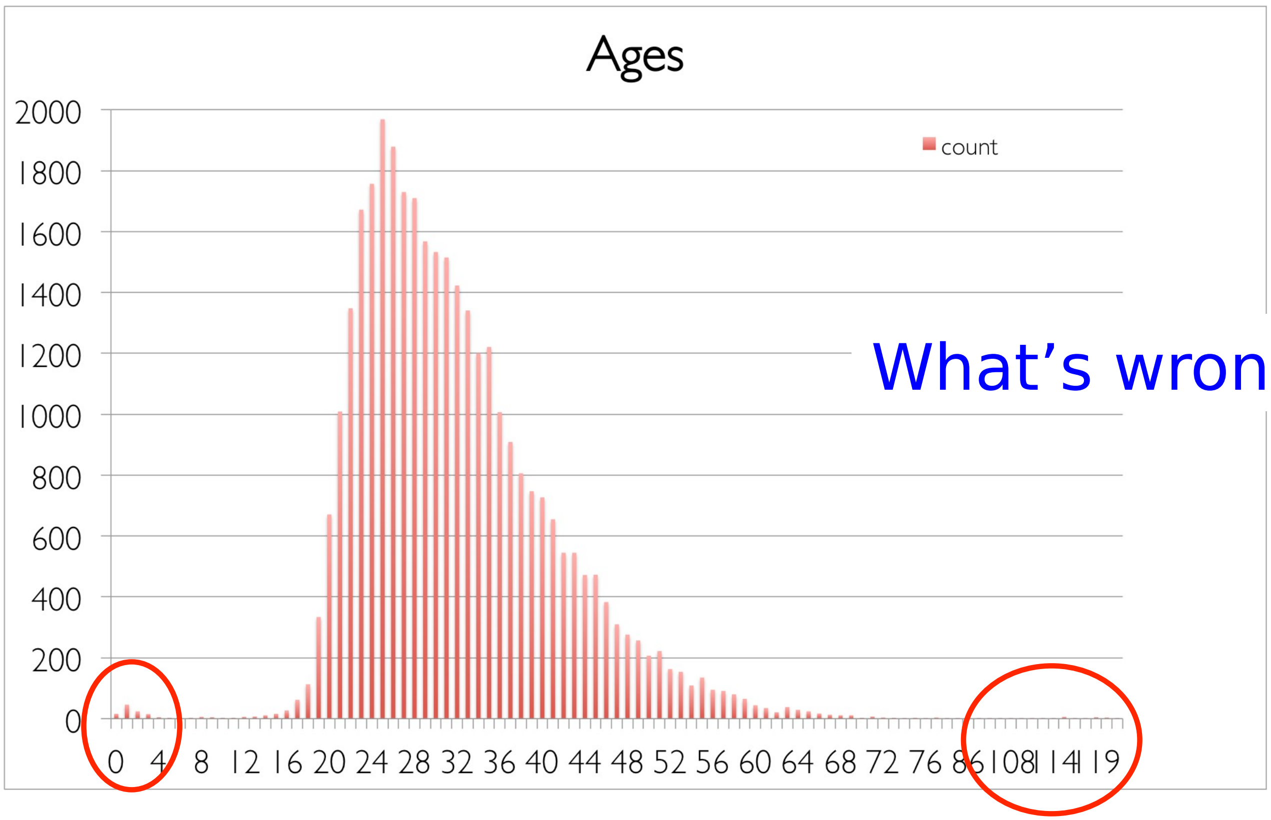


Ages of Students in a online course



What’s wrong with this data?

Numeric Outliers

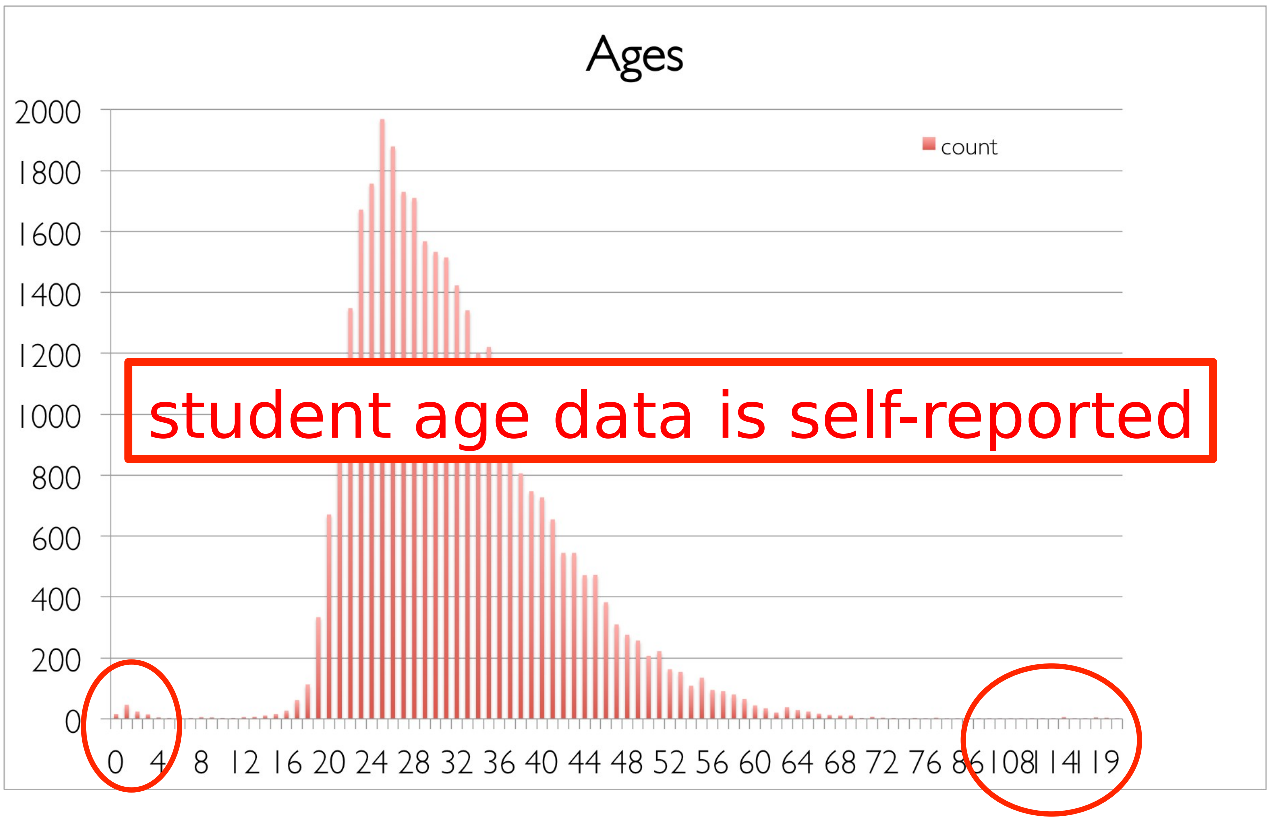


109 students ≤ 5

26 students > 100

What’s wrong with this data?

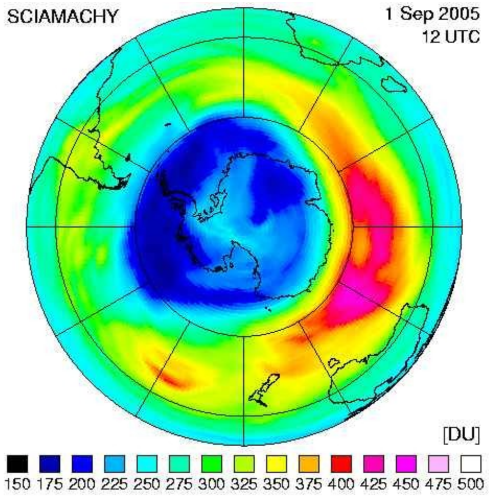
Numeric Outliers



109 students ≤ 5

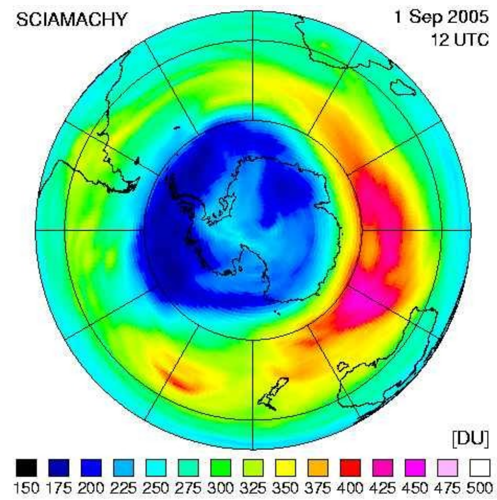
26 students > 100

Data Cleaning Makes Everything Okay?



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

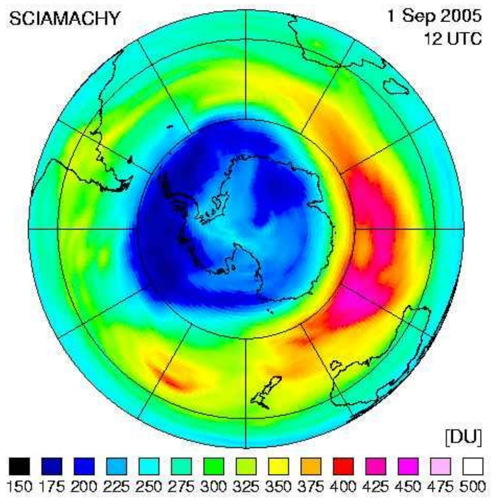
Data Cleaning Makes Everything Okay?



"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

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

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

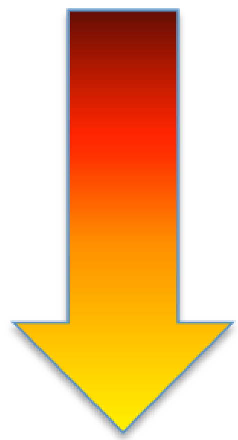
The Meaning of Data Quality

- 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 answer from available data
  - » Use of proxy data
  - » Relevant data is missing

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



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, sharing, 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 others 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

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

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?

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)

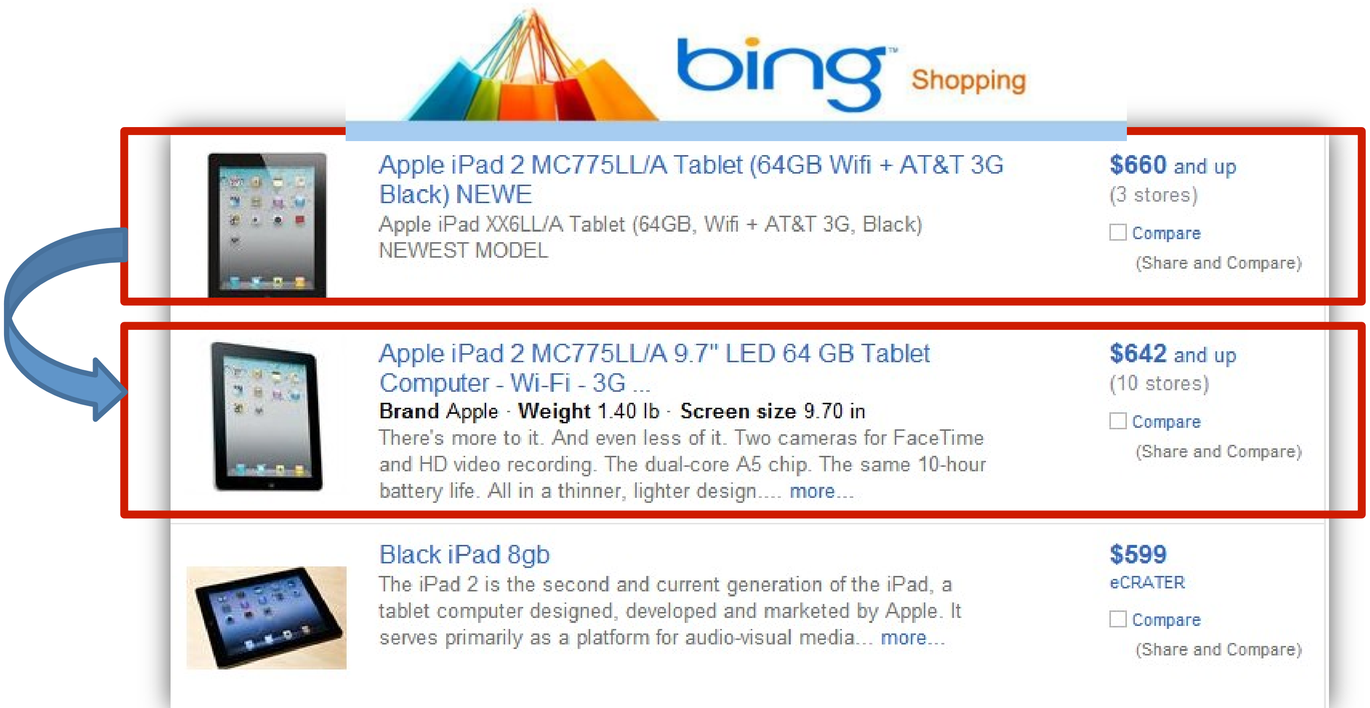
- Resolve multiple different entries:
  - » Entity resolution, reference reconciliation, object ID/consolidation
- Remove Duplicates: Merge/Purge
- Record Linking (across data sources)
- Approximate Match (accept fuzziness)
- Householding (special case)
  - » Different people in same house?
- …

Example: Entity Resolution

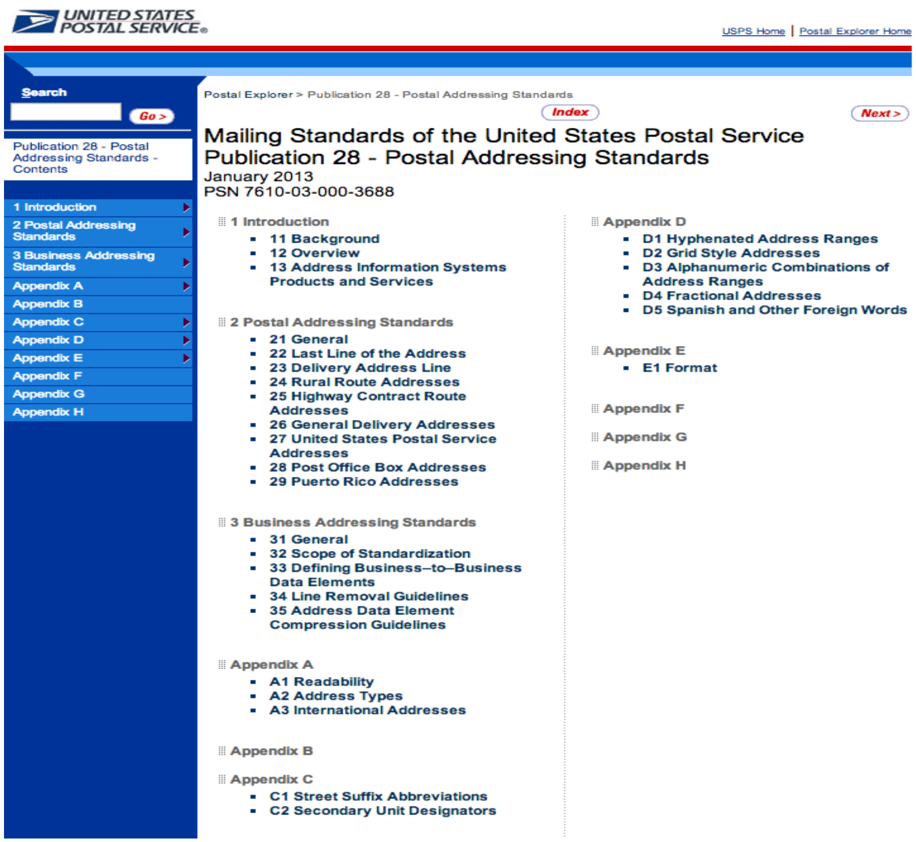
- Web scrape Google Shopping and Amazon product listings
- Google listing:
  - » clickart 950000 - premier image pack (dvd-rom) massive collection of images & fonts for all your design needs ondvd-rom!product informationinspire your creativity and perfect any creative project with thousands ofworld-class images in virtually every style. plus clickart 950000 makes iteasy for …
- Amazon listing:
  - » clickart 950 000 - premier image pack (dvd-rom)
- Are they these two listings the same product?

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
- …

Lots of Additional Problems

- Address vs. Number, Street, City, …
- Units
- Differing Constraints
- Multiple versions and schema evolution
- Other Metadata

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