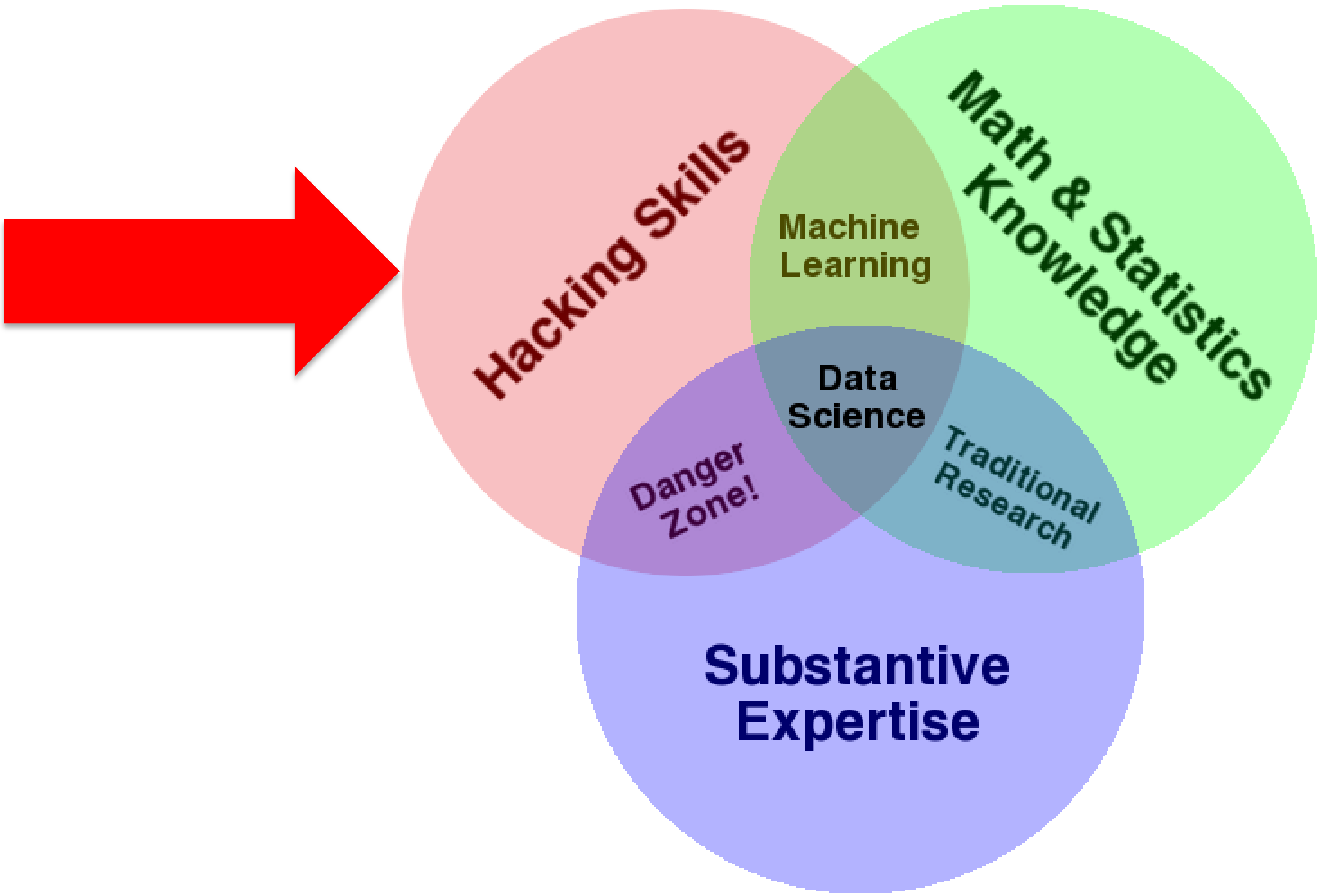
Introduction to Data Science
Lecture 4
Data Cleaning and Integration

CS 194 Fall 2014
John Canny
Based on notes by Michael Franklin, Dan Bruckner, Evan Sparks, Shivaram Venkataraman

Outline for this Evening

- Data Cleaning
  - Perspectives on “Dirty Data”
  - Perspectives on Data Quality
  - Some problems and solutions
- Data Integration
  - Item Similarity
  - Schema Matching

Data Science – One Definition



DB-hard Queries

Company_Name	Address	Market Cap
Google	Googleplex, Mtn. View, CA	$406Bn
Microsoft	Redmond, WA	$392Bn
Intl. Business Machines	Armonk, NY	$194Bn

```
SELECT Market_Cap
From Companies
Where Company_Name = “Apple”
```

```
Number of Rows: 0
```

Problem:
Missing Data

DB-hard Queries

Company_Name	Address	Market Cap
Google	Googleplex, Mtn. View, CA	$406Bn
Microsoft	Redmond, WA	$392Bn
Intl. Business Machines	Armonk, NY	$194Bn

```
SELECT Market_Cap
From Companies
Where Company_Name = “IBM”
```

Number of Rows: 0

Problem:
Entity Resolution

DB-hard Queries

Company_Name	Address	Market Cap
Google	Googleplex, Mtn. View, CA	$406
Microsoft	Redmond, WA	$392
Intl. Business Machines	Armonk, NY	$194
Sally's Lemonade Stand	Alameda,CA	$460

```
SELECT MAX(Market_Cap)
From Companies
```

Number of Rows: 1

Problem:
Unit Mismatch

WHO’S CALLING WHO’S DATA DIRTY?

Dirty Data

- The Statistics View:
  - There is a process that produces data
  - We want to model ideal samples of that process, but in practice we 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 (e.g. social networks)
  - You can add new models for each type of imperfection, but you can't model everything.
  - What's the best trade-off between accuracy and simplicity?

Dirty Data

- The 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 the quality of the values in your dataset

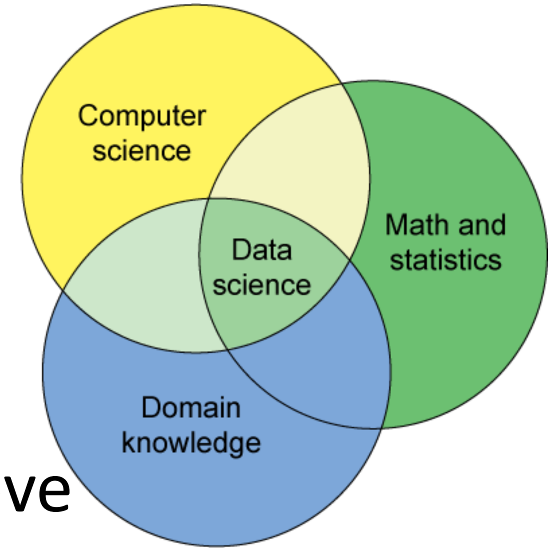
Dirty Data

- The Domain Expert's View:
  - This Data Doesn't look right
  - This Answer Doesn't look right
  - What happened?

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

Dirty Data

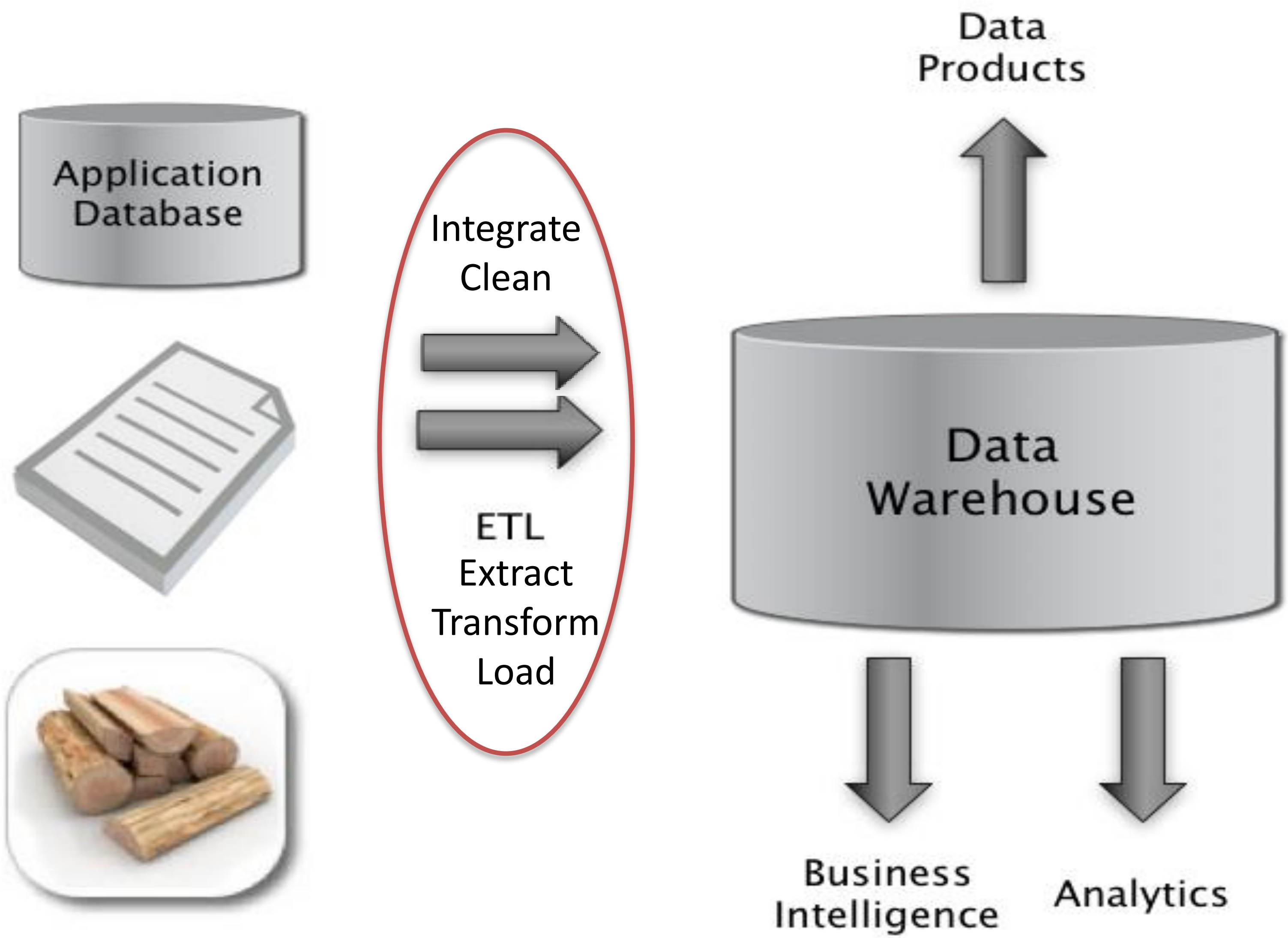
- The Data Scientist’s View:
    - Some Combination of all of the above



Data Quality Problems

- (Source) Data is dirty on its own.
- Transformations corrupt the data (complexity of software pipelines).
- Data sets are clean but integration (i.e., combining them) screws them up.
- “Rare” errors can become frequent after transformation or integration.
- Data sets are clean but suffer “bit rot”
  - Old data loses its value/accuracy over time
- Any combination of the above

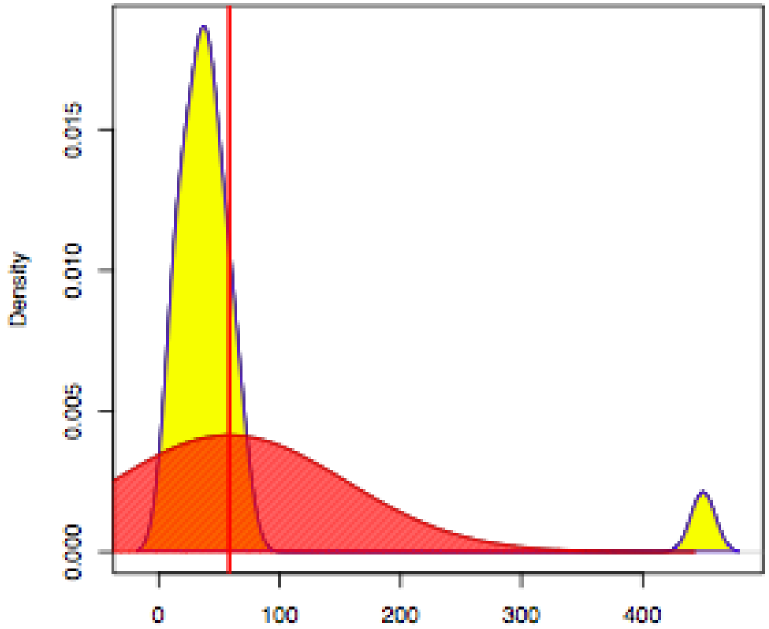
Big Picture: Where can Dirty Data Arise?



Numeric Outliers

12	13	14	21	22	26	33	35	36	37	39	42	45	47	54	57	61	68	450

ages of employees (US)



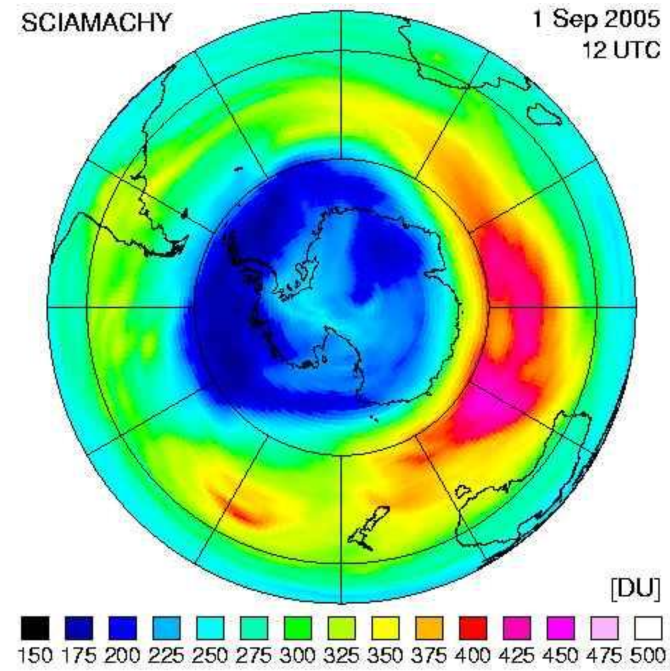
- median 37
- mean 58.52632
- variance 9252.041

Adapted from Joe Hellerstein's 2012 CS 194 Guest Lecture

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

- From Stanford Data Integration Course:
  1) parsing text into fields (separator issues)
  2) Naming conventions: ER: NYC vs New York
  3) Missing required field (e.g. key field)
  4) Different representations (2 vs Two)
  5) Fields too long (get truncated)
  6) Primary key violation (from un- to structured or during integration
  7) Redundant Records (exact match or other)
  8) Formatting issues – especially dates
  9) Licensing issues/Privacy/ keep you from using the data as you would like?

Conventional Definition of Data Quality

- Accuracy
  - The data was recorded correctly.
- Completeness
  - All relevant data was recorded.
- Uniqueness
  - Entities are recorded once.
- Timeliness
  - The data is kept up to date.
    - Special problems in federated data: time consistency.
- Consistency
  - The data agrees with itself.

Adapted from Ted Johnson's SIGMOD 2003 Tutorial

Problems …

- Unmeasurable
  - Accuracy and completeness are extremely difficult, perhaps impossible to measure.
- Context independent
  - No accounting for what is important. E.g., if you are computing aggregates, you can tolerate a lot of inaccuracy.
- Incomplete
  - What about interpretability, accessibility, metadata, analysis, etc.
- Vague
  - The conventional definitions provide no guidance towards practical improvements of the data.

Adapted from Ted Johnson's SIGMOD 2003 Tutorial

Finding a modern definition

- We need a definition of data quality which
  - Reflects the **use** of the data
  - Leads to **improvements in processes**
  - Is **measurable** (we can define metrics)

- First, we need a better understanding of how and where data quality problems occur
  - The data quality continuum

Adapted from Ted Johnson's SIGMOD 2003 Tutorial

Meaning of Data Quality (2)

- There are many types of data, which have different uses and typical quality problems
  - Federated data
  - High dimensional data
  - Descriptive data
  - Longitudinal data
  - Streaming data
  - Web (scraped) data
  - Numeric vs. categorical vs. text data

Adapted from Ted Johnson's SIGMOD 2003 Tutorial

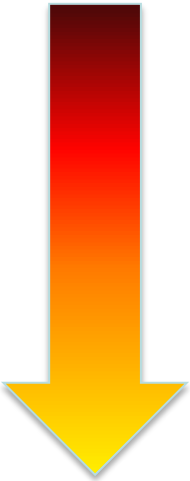
Meaning of Data Quality (2)

- There are many uses of data
  - Operations
  - Aggregate analysis
  - Customer relations …
- Data Interpretation : the data is useless if we don’t know all of the *rules* behind the data.
- Data Suitability : Can you get the answer from the 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 is not static, it 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?
- Sources of problems:
  - Manual entry
  - No uniform standards for content and formats
  - Parallel data entry (duplicates)
  - Approximations, surrogates – SW/HW constraints
  - Measurement or sensor errors.

Adapted from Ted Johnson's SIGMOD 2003 Tutorial

Data Gathering - Solutions

- Potential Solutions:
  - Preemptive:
    - Process architecture (build in integrity checks)
    - Process management (reward accurate data entry, data sharing, data stewards)
  - Retrospective:
    - Cleaning focus (duplicate removal, merge/purge, name & address matching, field value standardization)
    - Diagnostic focus (automated detection of glitches).

Adapted from Ted Johnson's SIGMOD 2003 Tutorial

Data Delivery

- Destroying or mutilating information by inappropriate 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 - Solutions

- Build reliable transmission protocols
  - Use a relay server
- Verification
  - Checksums, verification parser
  - Do the uploaded files fit an expected pattern?
- Relationships
  - Are there dependencies between data streams and processing steps
- Interface agreements
  - Data quality commitment from the data stream 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
  - Can be an issue, but terabytes are cheap.
- Problems in logical storage
  - Poor metadata.
    - Data feeds are often derived from application programs or legacy data 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 - Solutions

- Metadata
  - Document and publish data specifications.
- Planning
  - Assume that everything bad will happen.
  - Can be very difficult.
- 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 not properly understood.
  - Need for derived data not understood.
  - Just plain mistakes.
    - Inner join vs. outer join
    - Understanding NULL values
- Computational constraints
  - E.g., too expensive to give a full history, we'll supply a snapshot.
- Incompatibility
  - Ebcdic? Unicode?

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

Adapted from Ted Johnson's SIGMOD 2003 Tutorial

Retrieval and Mining - 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

- Many data quality problems can be captured by *static* constraints based on the schema.
  - Nulls not allowed, field domains, foreign key constraints, etc.
- Many others are due to problems in workflow, and 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
- Types of metrics
  - Static vs. dynamic constraints
  - Operational vs. diagnostic
- Metrics should be *directionally correct* with an improvement in use of the data.
- 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 changes in the database.
- Accuracy
  – Perform inventory (expensive), or use proxy (track complaints). Audit samples?
- Accessibility
- Interpretability
- Glitches in analysis
- Successful completion of end-to-end process

Adapted from Ted Johnson's SIGMOD 2003 Tutorial

Technical Approaches

- We need a multi-disciplinary approach to attack data quality problems
  - No one approach solves all problem
- 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 it mean? Interpretation

Adapted from Ted Johnson's SIGMOD 2003 Tutorial

Some Notes on the Class

- HW1 due Thursday

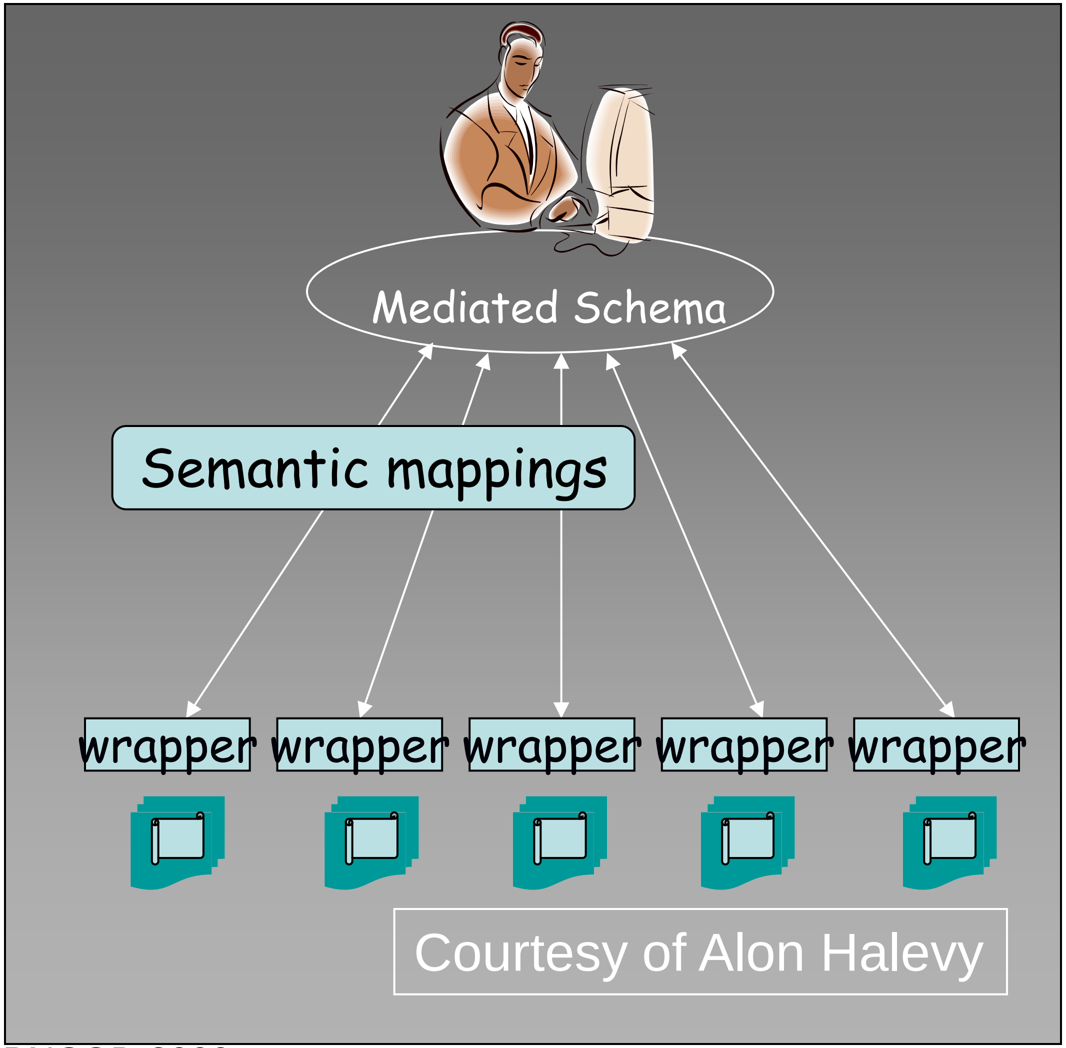
- FINAL PROJECTS
  - **Project Teams (due Friday 9/26/14)**
  - **Project Preferences (due Wednesday 10/1/14)**
  - **Project Assignments (Friday 10/3/14)**
  - **Project Proposals (due Friday 10/10/14)**

Break – 5 min

Schema and Data Integration

Which problems does Integration exacerbate?

Which problems does schema on write help?



Data Integration

- Combine data sets (acquisitions, across departments).
- Common source of problems
  - Heterogenous data : no common key, different field formats
    - Approximate matching
  - Different definitions
    - What is a customer: an account, an individual, a family, …
  - Time synchronization
    - Does the data relate to the same time periods? Are the time windows compatible?
  - Legacy data
    - IMS, spreadsheets, ad-hoc structures

Adapted from Ted Johnson's SIGMOD 2003 Tutorial

Schema Matching

- Original Problem: merge structured databases
  - But, even in a looser schema (e.g. NoSQL) world structural matching matters
- WebTables paper shows an extreme version of this
  - 2.6M Unique schemas (appear >1 time)
  - 5.4M Unique attribute (field) names (>1 time)
  - Found by web crawling/scraping

WebTables Extracted Tables

make	model	year
Toyota	Camry	1984

make	model	year
Mazda	Protégé	2003
Chevrolet	Impala	1979

make	model	year	color
Chrysler	Volare	1974	yellow
Nissan	Sentra	1994	red

name	addr	city	state	zip
Dan S	16 Park	Seattle	WA	98195
Alon H	129 Elm	Belmont	CA	94011

name	size	last-modified
Readme.txt	182	Apr 26, 2005
cac.xml	813	Jul 23, 2008

Schema	Freq
{make, model, year}	2
{make, model, year, color}	1
{name, addr, city, state, zip}	1
{name, size, last-modified}	1

- ACSDb is useful for computing attribute probabilities
  - p("make"), p("model"), p("zipcode")
  - p("make" | "model"), p("make" | "zipcode")

ACSDb* Applications

- Schema Auto Complete
- Attribute Synonym-Finding
- Join Graph Traversal

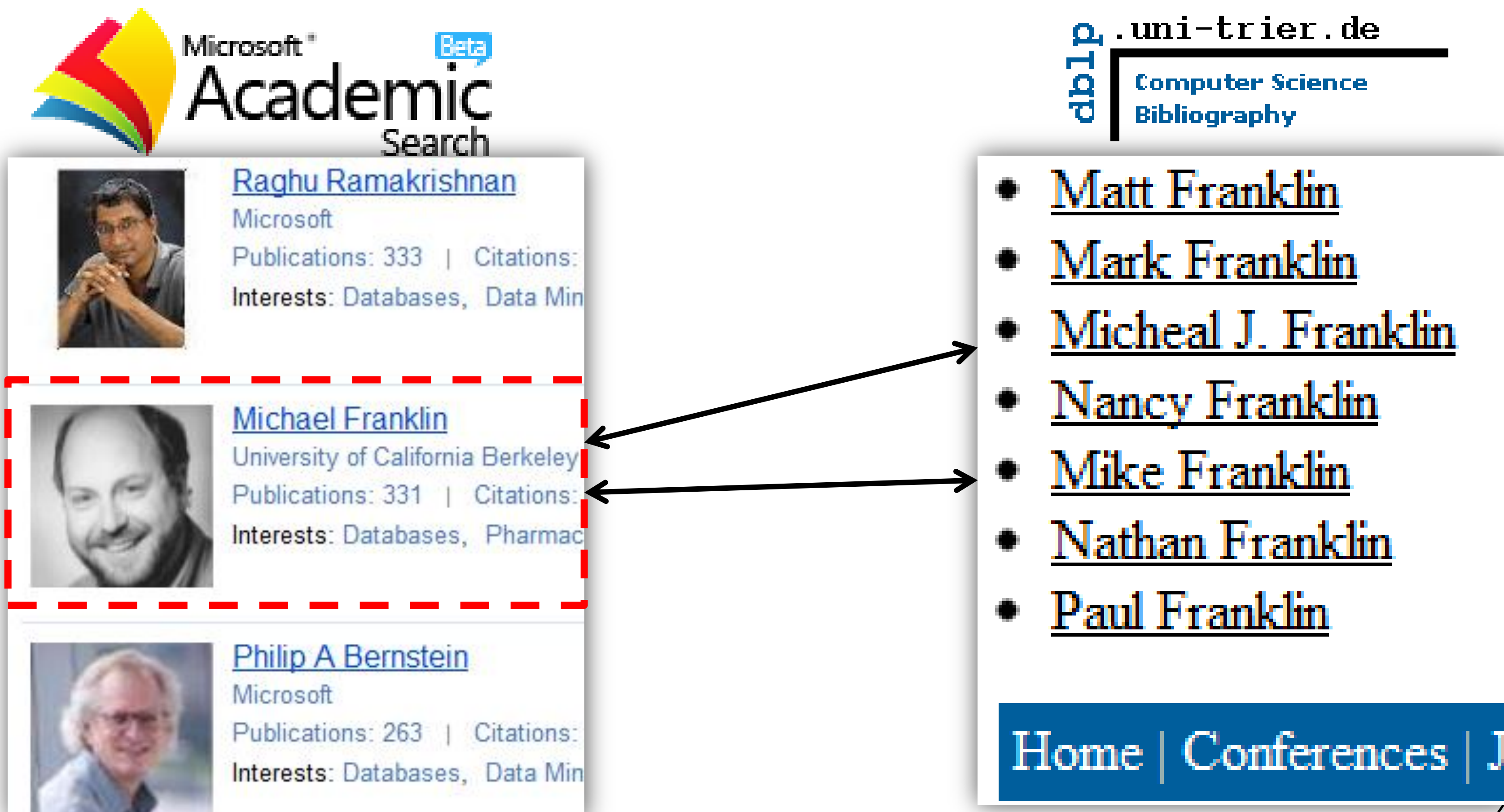
*Attribute Correlation Statistics Database

MATCHING: DATA AND STRUCTURE

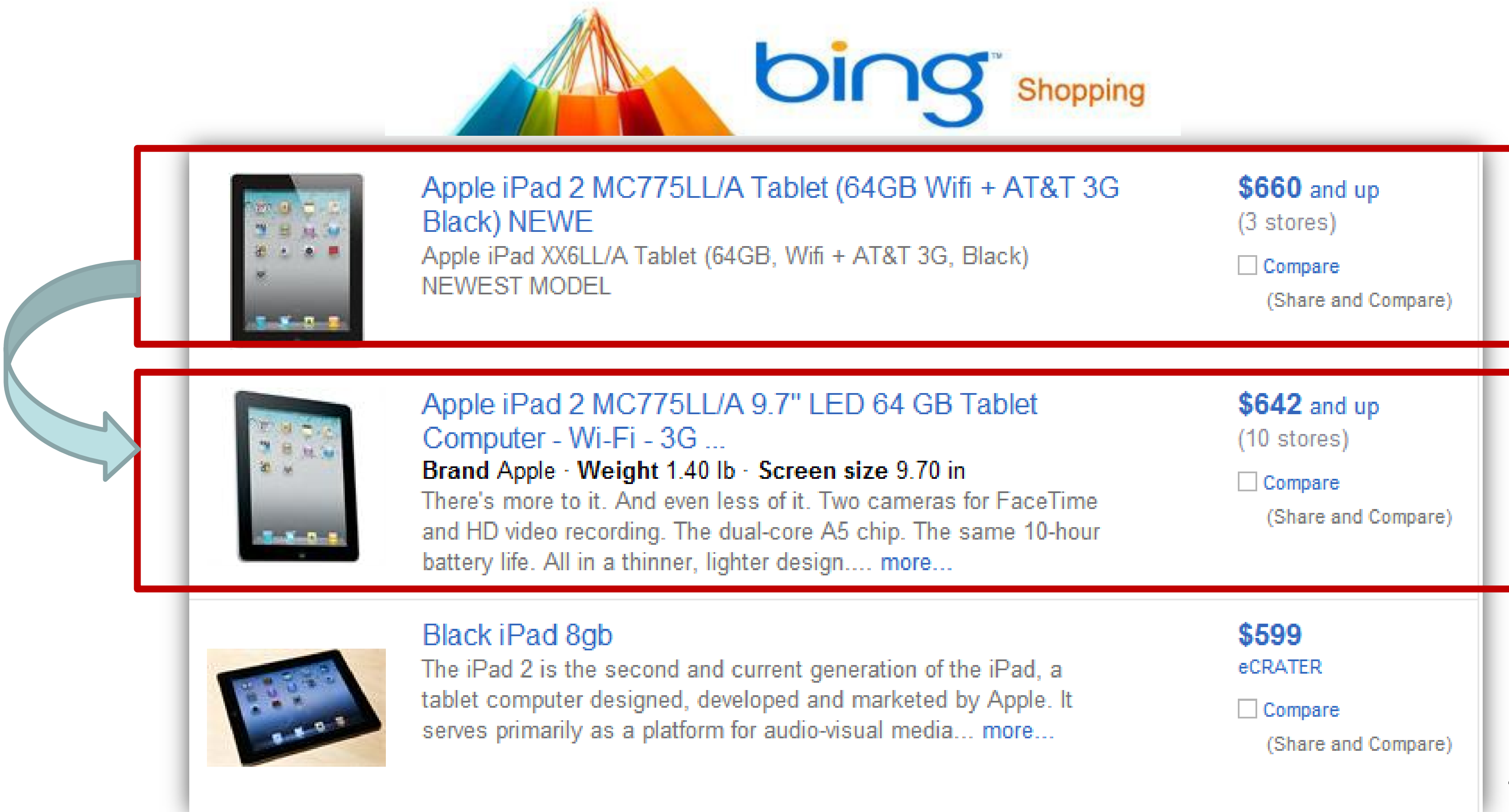
Duplicate Record Detection needs DeDup!

- Resolve multiple different mentions:
  - Entity Resolution
  - Reference Reconciliation
  - Object Identification/Consolidation
- Remove Duplicates
  - Merge/Purge
- Record Linking (across data sources)
- Householding (interesting special case)
- Approximate Match (accept fuzziness)
- …

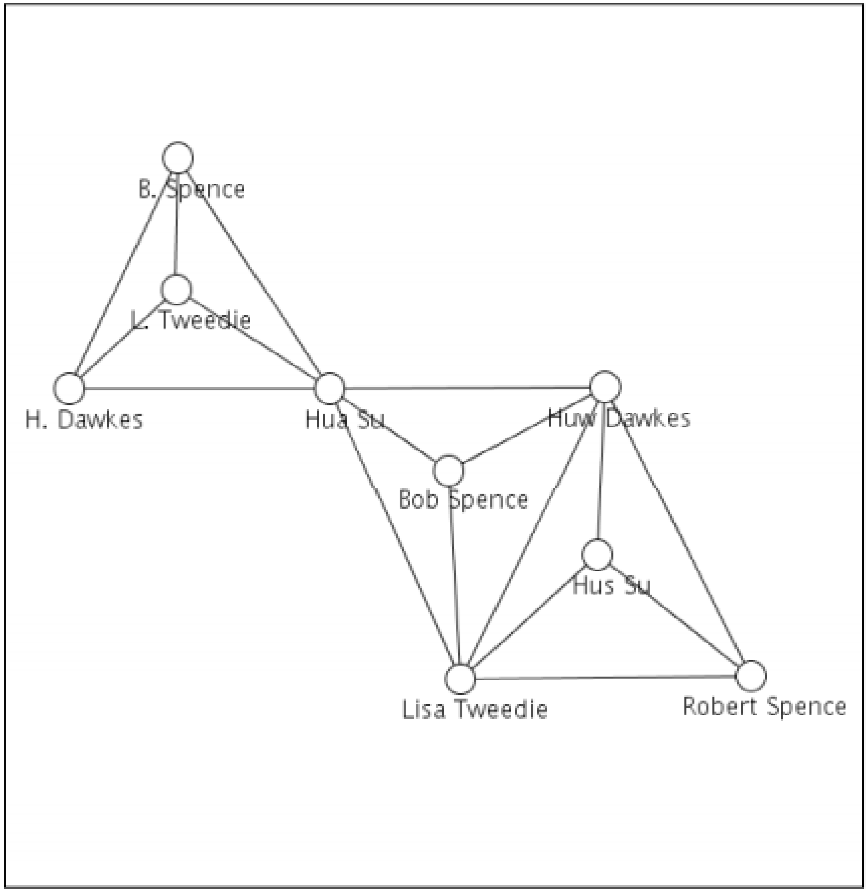
Example: Data Integration



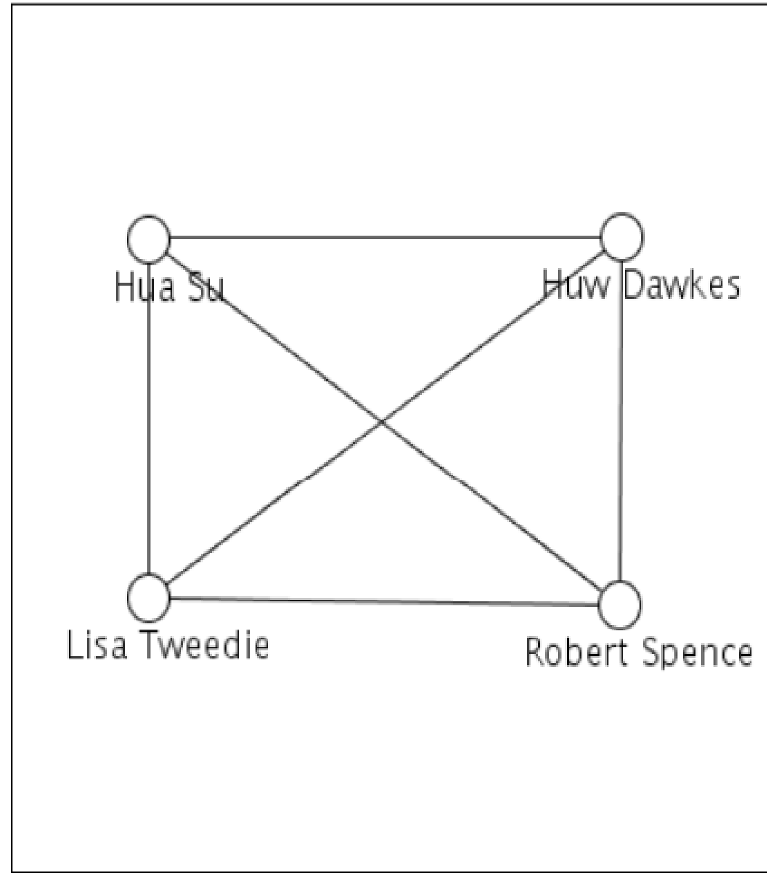
Example: DeDup/Cleaning



Example: Network Analysis



before

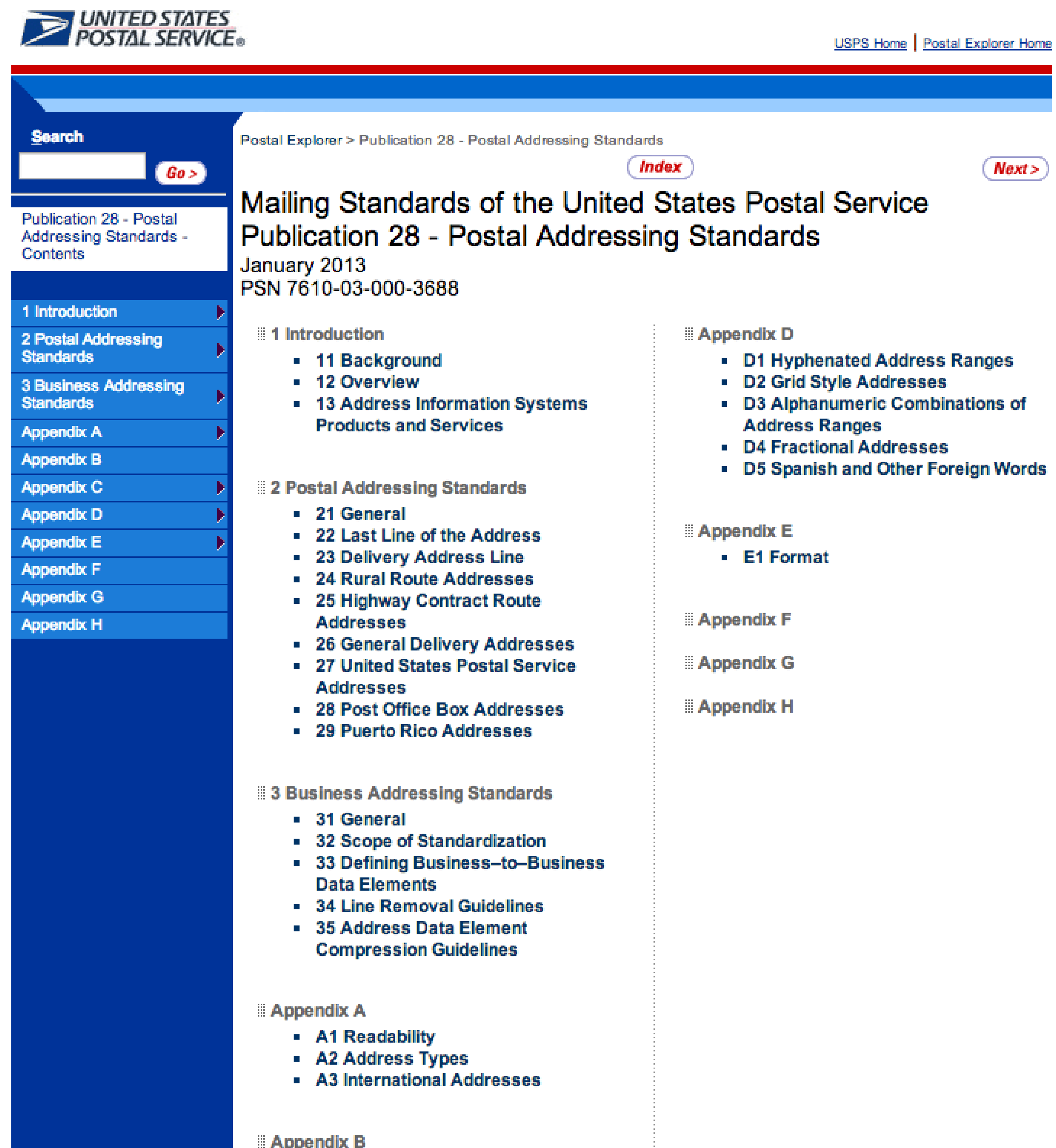


after

From: Getoor & Machanavajjhala: “Entity Resolution Tutorial”, VLDB 2012

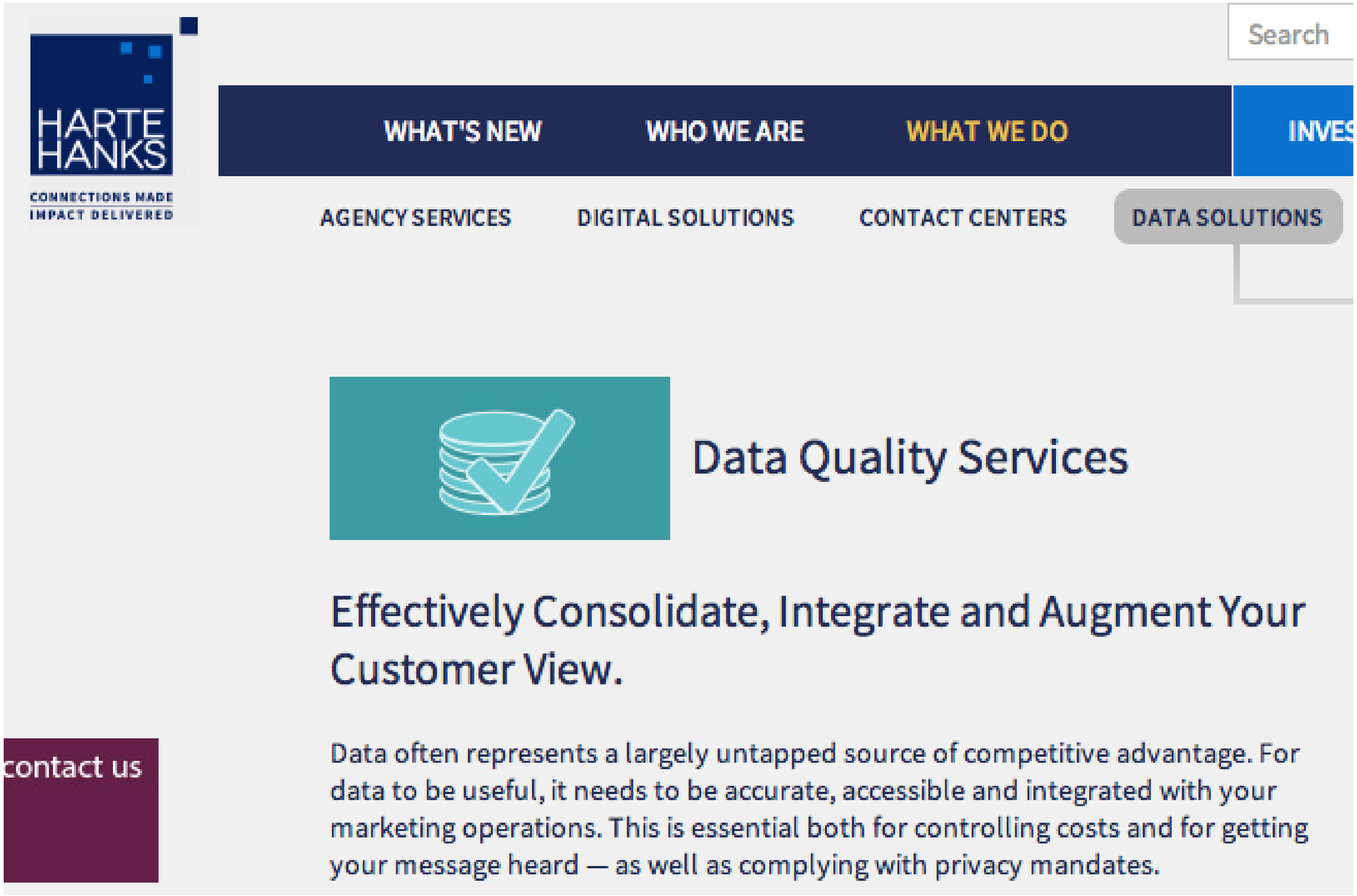
Preprocessing/Standardization

- Simple idea:
- Convert to canonical form
- e.g. addresses



More Complicated: Householding

- Different people in same house?



Approximate Matching

- Relate tuples whose fields are “close”
  - Approximate string matching
    - Generally, based on edit distance.
    - Fast SQL expression using a *q-gram* index (a q-gram is like an n-gram on syllables)
  - Approximate tree matching
    - For Nested Data Structures (or flattened ones)
    - Much more expensive than string matching
    - Recent research in fast approximations
  - Feature vector matching
    - Similarity search
    - Many techniques discussed in the data mining literature.
  - Ad-hoc or Domain-focused matching
    - Use domain insights and/or clever tricks.

Some Similarity Measures

Handle Typographical errors

- Equality on a boolean predicate
- Edit distance
  - Levenstein, Smith-Waterman, Affine
- Set similarity
  - Jaccard, Dice
- Vector Based
  - Cosine similarity, TFIDF

Good for Text like reviews/ tweets

Good for Names

- Alignment-based or Two-tiered
  - Jaro-Winkler, Soft-TFIDF, Monge-Elkan
- Phonetic Similarity
  - Soundex
- Translation-based
- Numeric distance between values
- Domain-specific

Useful for abbreviations, alternate names.

From: Getoor & Machanavajjhala: “Entity Resolution Tutorial”, VLDB 2012

Soundex Encoding

A phonetic algorithm that indexes names by their sounds when pronounced in english.

Consists of the first letter of the name followed by three numbers. Numbers encode similar sounding consonants.

- Remove all W, H
- B, F, P, V encoded as 1, C,G,J,K,Q,S,X,Z as 2
- D,T as 3, L as 4, M,N as 5, R as 6, Remove vowels
- Concatenate first letter of string with first 3 numerals

Ex: great and grate become 6EA3 and 6A3E and then G63

More recent, metaphone, double metaphone etc.

From: Koudas, Sarawagi, Strivastava, "Record Linkage: Similarity Measures and Algorithms", VLDB 2006

Edit Distance

- Character Operations: I (insert), D (delete), R (Replace).
- Unit costs.
- Given two strings, s,t, edit(s,t):
  - Minimum cost sequence of operations to transform s to t.
  - Example: edit(Error,Eror) = 1, edit(great,grate) = 2
- Folklore dynamic programming algorithm to compute edit();
- Computation and decision problem: quadratic (on string length) in the worst case.
  - May be costly operation for large strings
  - Suitable for common typing mistakes
    - Comprehensive vs Comprenhensive
  - Problematic for specific domains
    - AT&T Corporation vs AT&T Corp
    - IBM Corporation vs AT&T Corporation

From: Koudas, Sarawagi, Strivastava, “Record Linkage: Similarity Measures and Algorithms”, VLDB 2006

Overlap Metrics

- Given two sets of terms S, T
  - Jaccard coef.: Jaccard(S,T) = $|S \cap T| / |S \cup T|$
  - Variants
    - If scores (weights) available for each term (element in the set) compute Jaccard() only for terms with weight above a specific threshold.
- What constitutes a good choice of a term score?

- Terms can be words or “q-grams” (sequence of q characters in a field:
  - e.g., {‘AT&’, ‘T&T’, ‘&T ‘, ‘T C’, …} for AT&T Corp.

From: Koudas, Sarawagi, Strivastava, “Record Linkage: Similarity Measures and Algorithms”, VLDB 2006

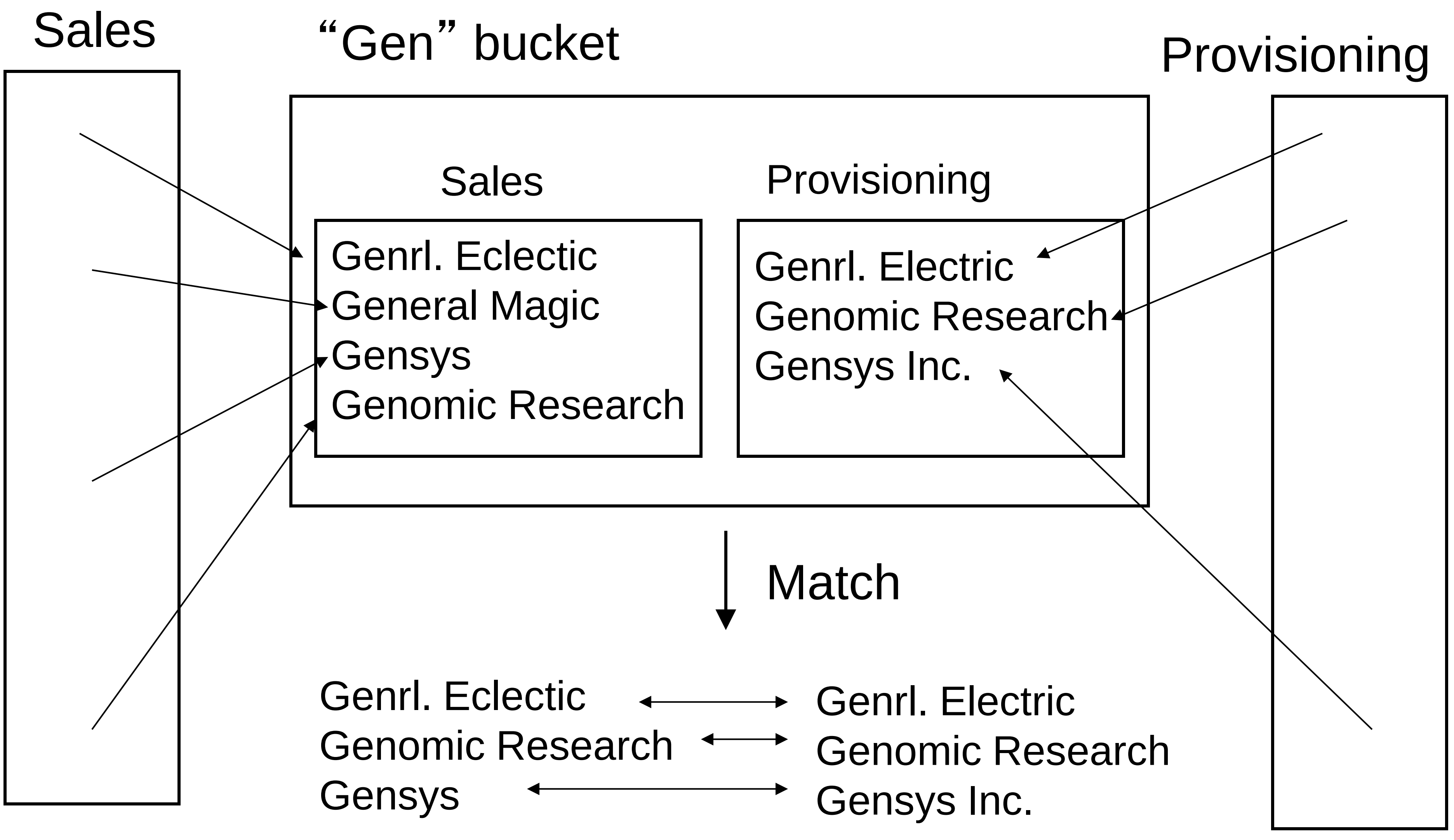
More Sophisticated Techniques

- Evidence from multiple fields
  - Positive and Negative are possible
- Evidence from linkage pattern with other records
- Clustering-based approaches
- …

Approximate Joins and Duplicate Elimination

- Perform joins based on incomplete or corrupted information.
  - Approximate join : between two different tables
  - Duplicate elimination : within the same table
- More general than approximate matching.
  - **Semantics** : Need to use special transforms and scoring functions.
  - **Correlating information** : verification from other sources, e.g. usage correlates with billing.
  - **Missing data** : Need to use several orthogonal search and scoring criteria.
- But approximate matching is a valuable tool …

(Approximate Join Example)



Algorithm (for scalability)

- Partition data set
  - By hash on computed key
  - By sort order on computed key
  - By similarity search / approximate match on computed key
- Perform scoring within the partition
  - Hash : all pairs
  - Sort order, similarity search : target record to retrieved records
- Record pairs with high scores are matches
- Use multiple computed keys / hash functions
- Duplicate elimination : duplicate records form an equivalence class.

Schema Matching

- Use similarity measures and structural cues (e.g. column names, data types, etc.) to match data definitions
- Looking at data instances (or examples of them can help)
- Constraints in the schema (if you have them) can also help.
- Auxiliary Information: dictionaries, documentation, usage… ditto

Lots of Additional Problems

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

Data Integration

- Combine data sets (acquisitions, across departments).
- Common source of problems
  - Heterogenous data : no common key, different field formats
    - Approximate matching
  - Different definitions
    - What is a customer: an account, an individual, a family, …
  - Time synchronization
    - Does the data relate to the same time periods? Are the time windows compatible?
  - Legacy data
    - IMS, spreadsheets, ad-hoc structures
  - Sociological factors
    - Reluctance to share – loss of power.

Adapted from Ted Johnson's SIGMOD 2003 Tutorial

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

Summary

- Data Cleaning
  - Perspectives on “Dirty Data”
  - Perspectives on Data Quality
  - Some problems and solutions
- Data Integration
  - Item Similarity
  - Schema Matching