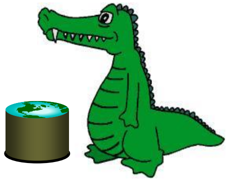
Logistics

- NIST pre-pilot project
  - Lab 4 + Final Project
  - First submission Nov. 2nd
  - Second submission Nov. 10th
- Pop Quiz
- Want to be added as Auditor? – please send me your UFID's in email

Review

- Schema-on-read vs. schema-on-write
- Examples and problems with dirty data
- Dirty data from the viewpoint of data scientists (statistics, database, domain expertise)
- How data quality issues occur in the data quality continuum (i.e., data analytical process) and possible solutions
- Metrics of data quality



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.
- Consistency
  - The data agrees with itself.

Adapted from Ted Johnson's SIGMOD 2003 Tutorial

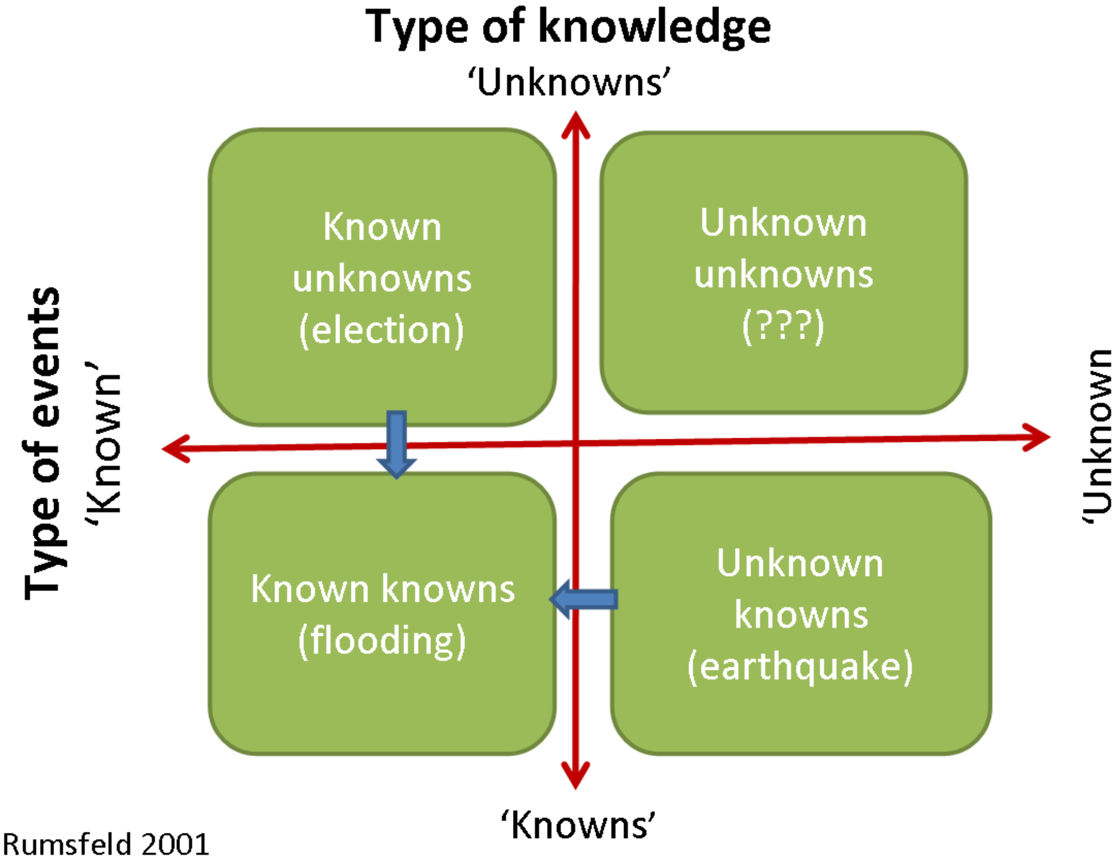
Problems …

- 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

In-completeness in knowledge

Types of risk



Type of knowledge

‘Unknowns’

Known unknowns (election)

Unknown unknowns (???)

Type of events

‘Known’

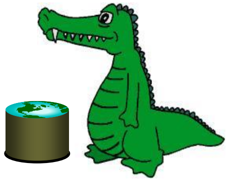
‘Unknown

Known knowns (flooding)

Unknown knowns (earthquake)

‘Knowns’

From Rumsfeld 2001



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)

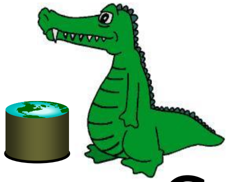
- With a better understanding of how and where data quality problems occur
  - The data quality continuum

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



Examples of Data Quality Metrics

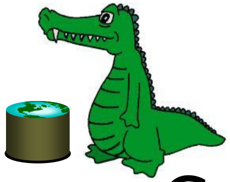
- Conformance to schema
  - Evaluate constraints on a snapshot.
- Conformance to business rules
  - Evaluate constraints on changes in the database.
- Accuracy
  - Perform inventory (expensive), or use proxy (track complaints). Audit samples?
- Accessibility
- Interpretability
- Glitches in analysis
- Successful completion of end-to-end process

Adapted from Ted Johnson's SIGMOD 2003 Tutorial

Technical Approaches for data cleaning

- 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



Data Integration

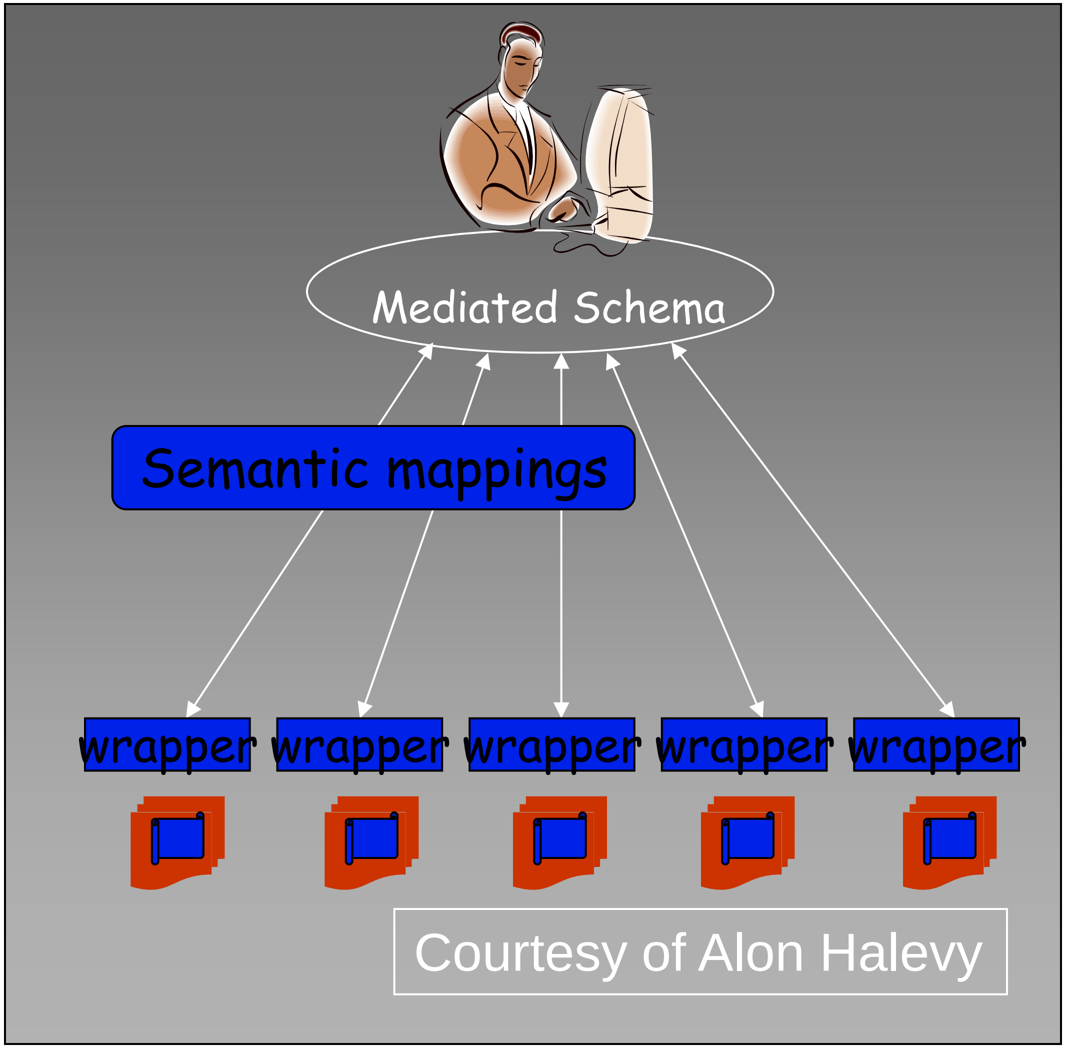
- Combine data sets (acquisitions, across departments).
- Common source of problems
  - Heterogenous data : no common key, different field formats
    - Approximate matching
  - Different definitions
    - What is a customer: an account, an individual, a family, …
  - Time synchronization
    - Does 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 and Data Integration

Which problems does Integration exacerbate?

Which problems does schema on write help?



Courtesy of Alon Halevy

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

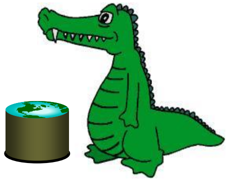
- ACSDb is useful for computing attribute probabilities
  - p("make"), p("model"), p("zipcode")
  - p("make" | "model"), p("make" | "zipcode")

ACSDb* Applications

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

*Attribute Correlation Statistics Database

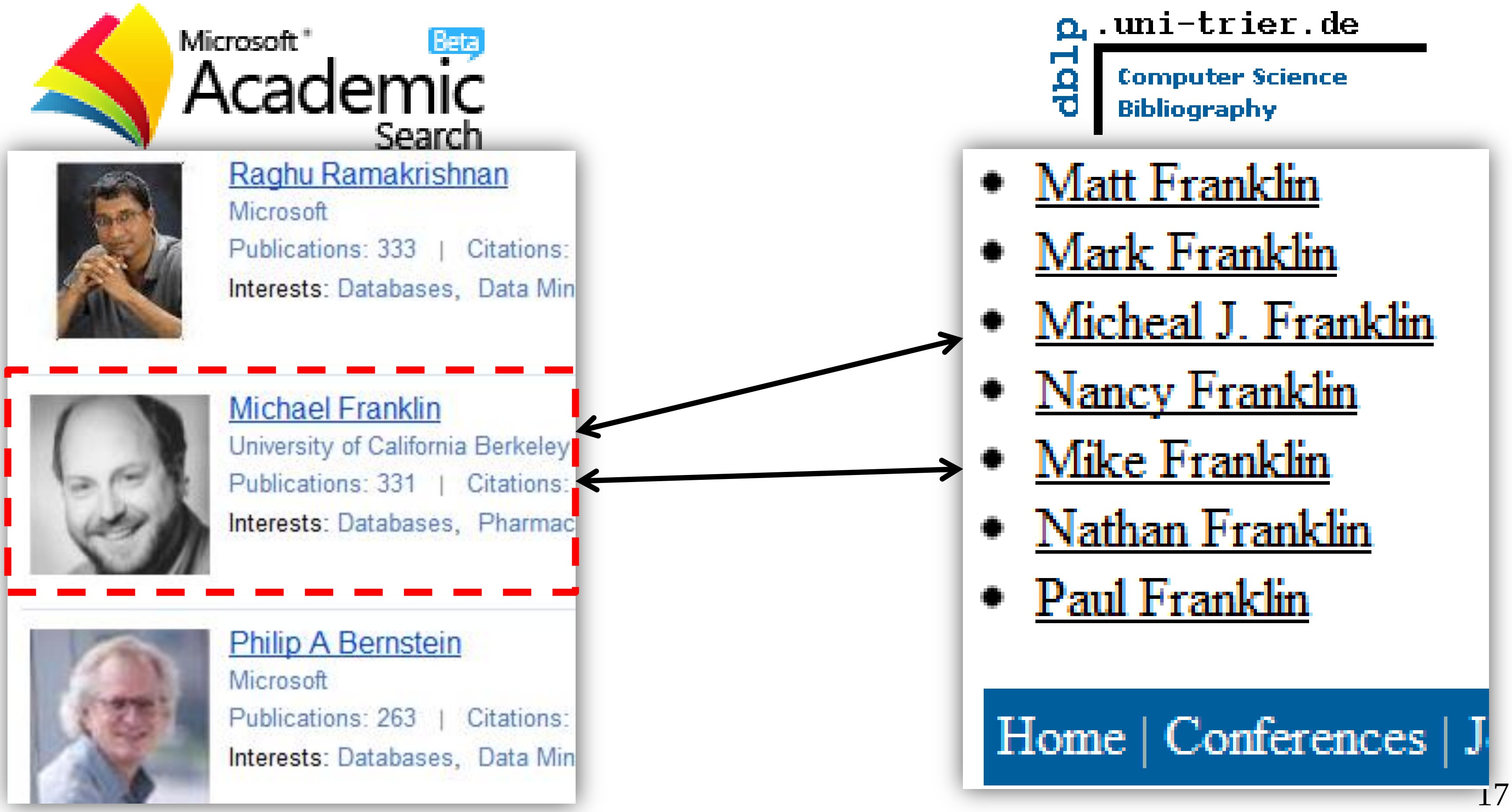
MATCHING: DATA AND STRUCTURE



Duplicate Record Detection needs DeDup!

- Step 1: Resolve multiple mentions:
  - Entity Resolution
  - Reference Reconciliation
  - Object Identification/Consolidation
- Step 2: Remove Duplicates
  - Merge/Purge
- Other variations:
  - – Record Linking (across data sources)
  - – Householding (interesting special case)
  - – Approximate Match (accept fuzziness)
  - – …

Example: Data Integration



Microsoft® Academic Search Beta

Raghu Ramakrishnan
Microsoft
Publications: 333 | Citations:
Interests: Databases, Data Min

Michael Franklin
University of California Berkeley
Publications: 331 | Citations:
Interests: Databases, Pharmac

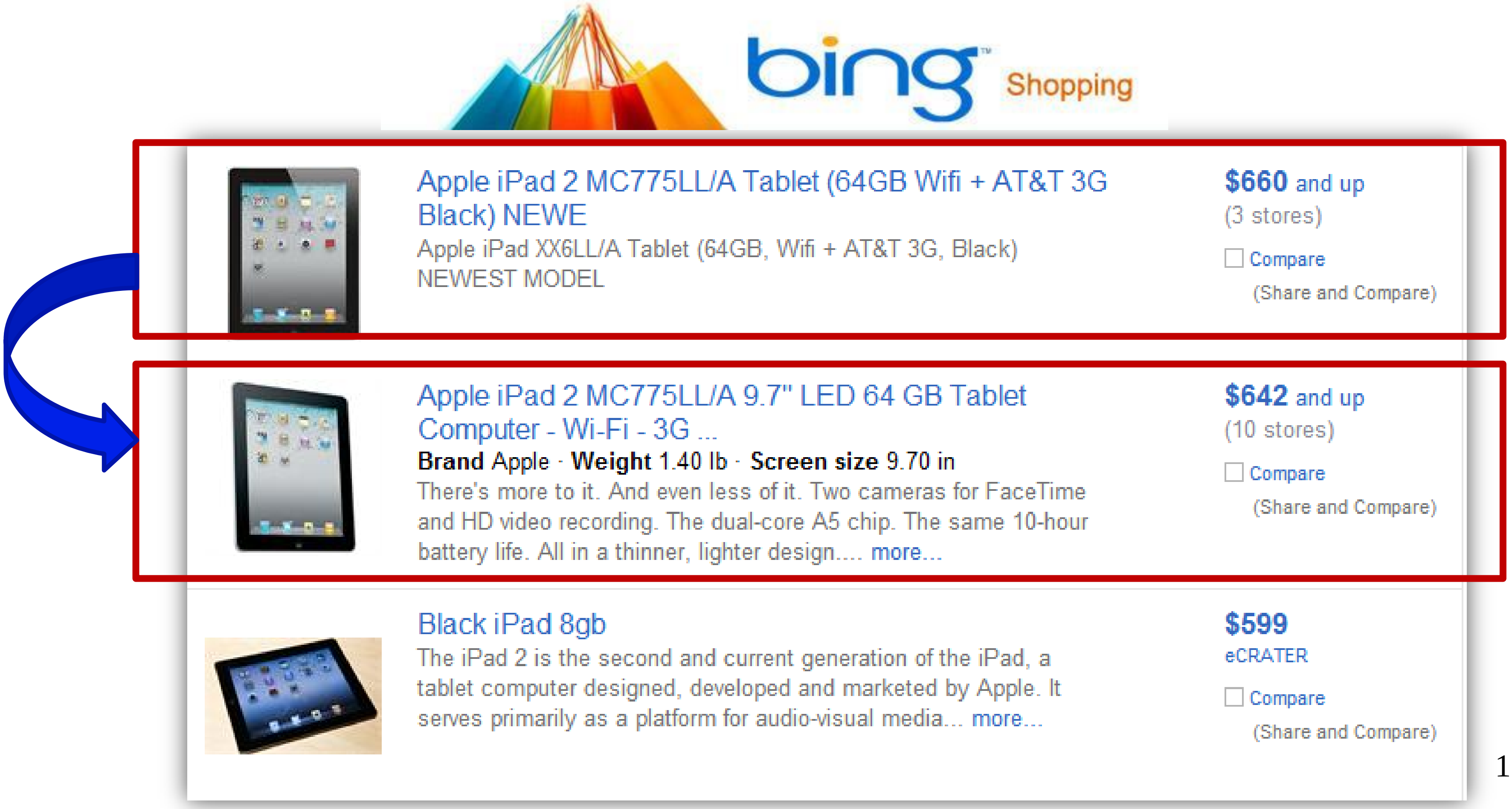
Philip A Bernstein
Microsoft
Publications: 263 | Citations:
Interests: Databases, Data Min

dblp.uni-trier.de
Computer Science Bibliography

- Matt Franklin
- Mark Franklin
- Micheal J. Franklin
- Nancy Franklin
- Mike Franklin
- Nathan Franklin
- Paul Franklin

Home | Conferences | J

Example: DeDup/Cleaning

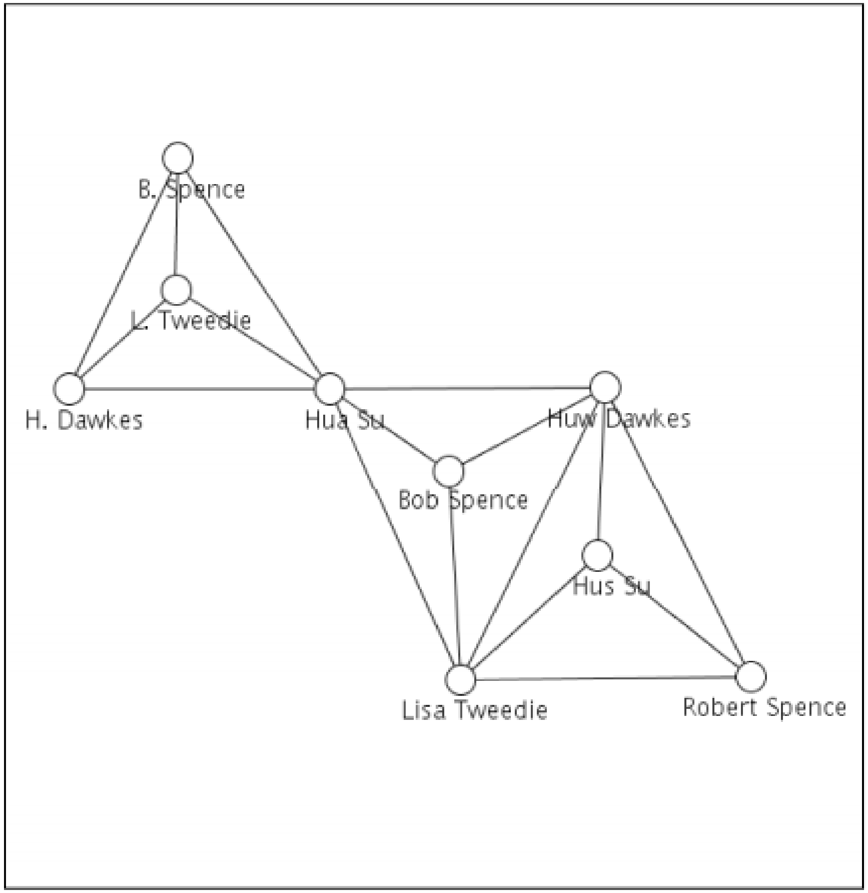


Apple iPad 2 MC775LL/A Tablet (64GB Wifi + AT&T 3G Black) NEWE
Apple iPad XX6LL/A Tablet (64GB, Wifi + AT&T 3G, Black) NEWEST MODEL
$660 and up
(3 stores)
□ Compare
(Share and Compare)

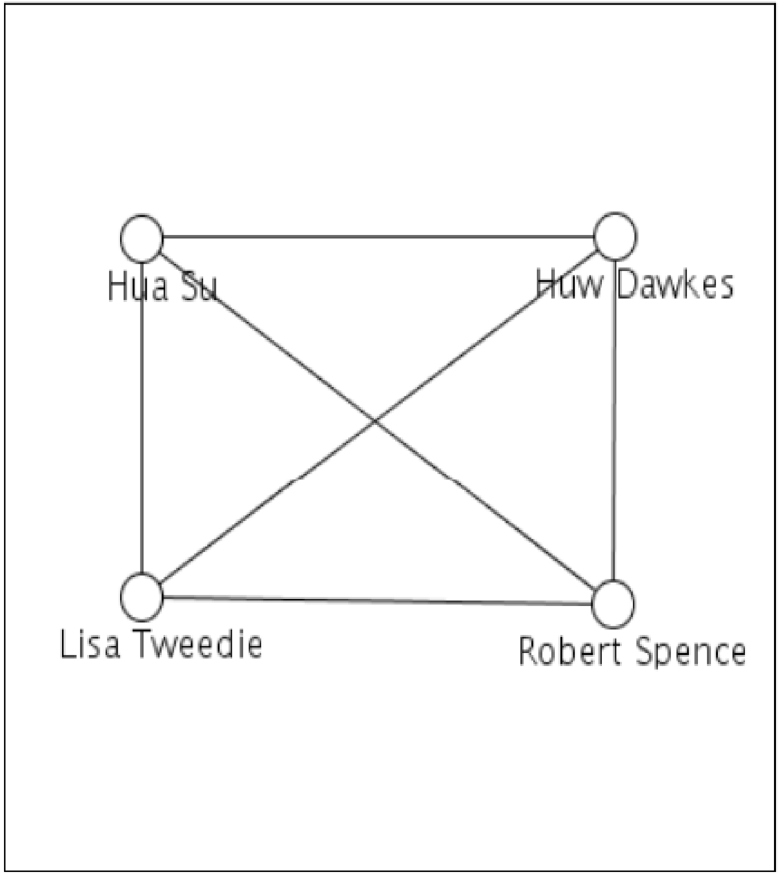
Apple iPad 2 MC775LL/A 9.7" LED 64 GB Tablet Computer - Wi-Fi - 3G ...
Brand Apple - **Weight** 1.40 lb - **Screen size** 9.70 in
There's more to it. And even less of it. Two cameras for FaceTime and HD video recording. The dual-core A5 chip. The same 10-hour battery life. All in a thinner, lighter design.... more...
$642 and up
(10 stores)
□ Compare
(Share and Compare)

Black iPad 8gb
The iPad 2 is the second and current generation of the iPad, a tablet computer designed, developed and marketed by Apple. It serves primarily as a platform for audio-visual media... more...
$599
eCRATER
□ Compare
(Share and Compare)

Example: Network Analysis



before

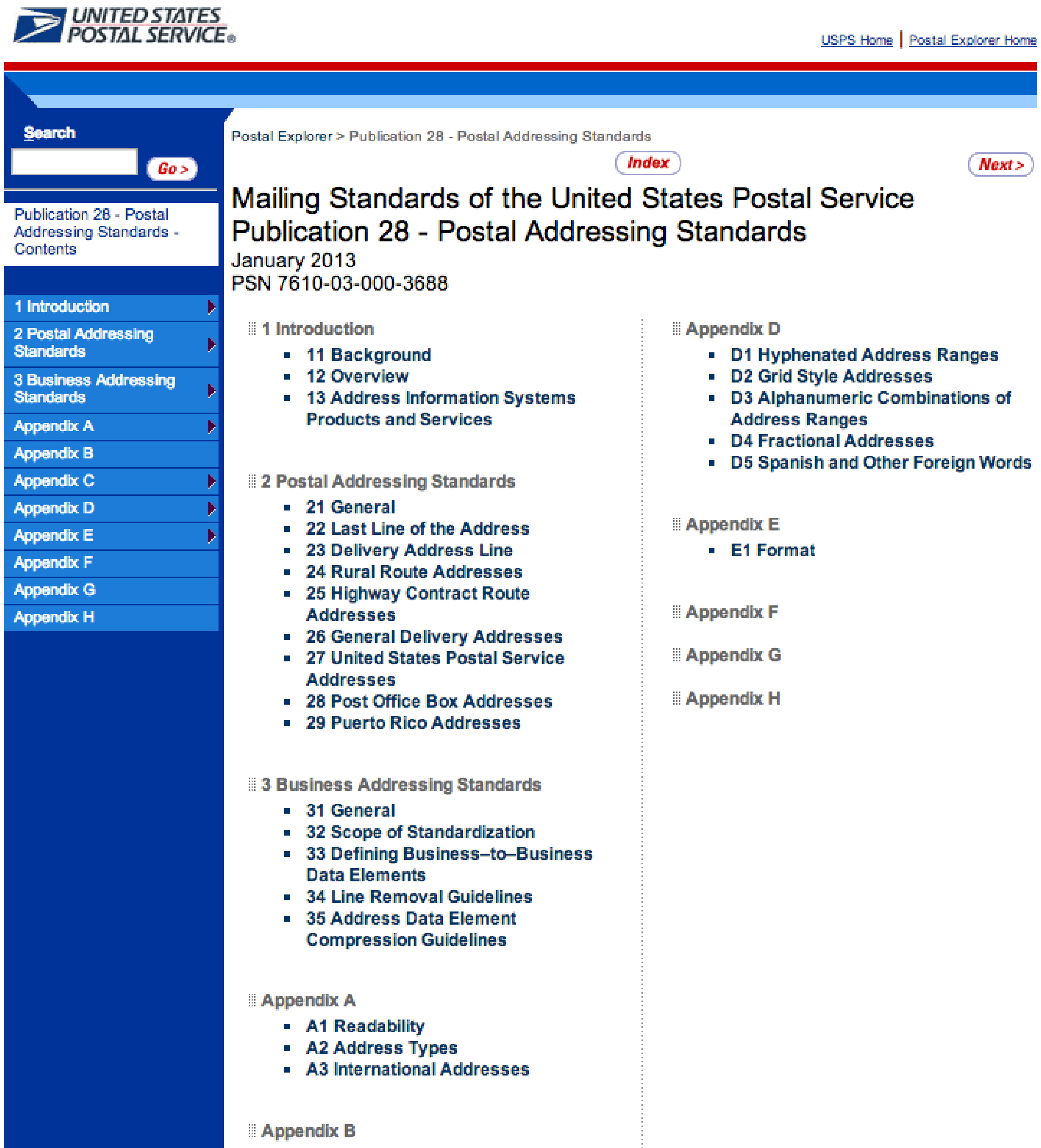


after

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

Preprocessing/Standardization

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



UNITED STATES POSTAL SERVICE®

USPS Home | Postal Explorer Home

Search

Go >

Publication 28 - Postal Addressing Standards - Contents

Postal Explorer > Publication 28 - Postal Addressing Standards

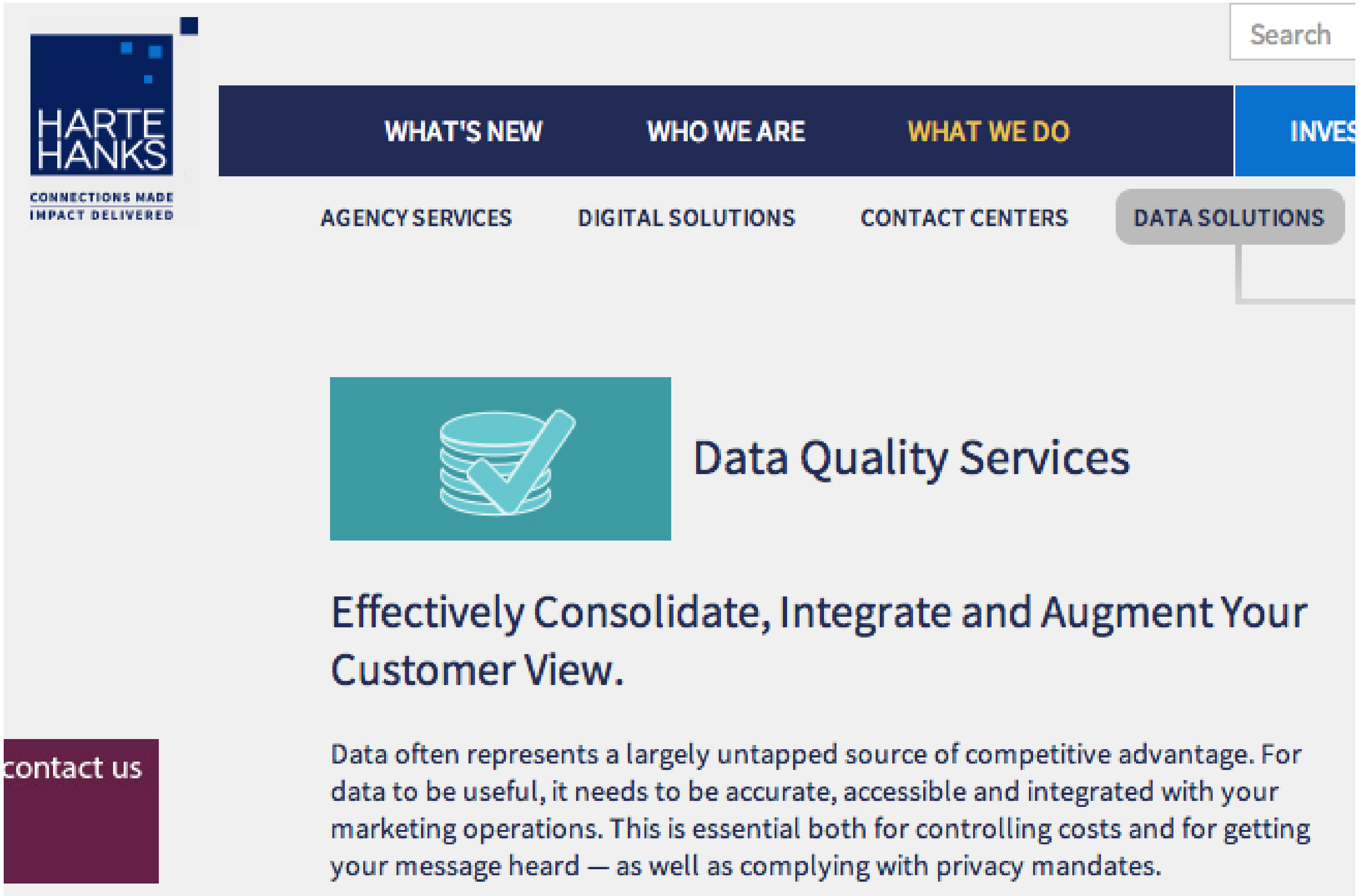
Index

Next >

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

More Complicated: Householding

- Different people in same house?



Data Quality Services

Effectively Consolidate, Integrate and Augment Your Customer View.

Data often represents a largely untapped source of competitive advantage. For data to be useful, it needs to be accurate, accessible and integrated with your marketing operations. This is essential both for controlling costs and for getting your message heard — as well as complying with privacy mandates.

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