CPT-S 415

Big Data

Yinghui Wu

EME B45

CPT-S 415
Big Data

The veracity of big data

- ✓ Data quality management: An overview
- ✓ Central aspects of data quality
  - – Data consistency
  - – Entity resolution
  - – Information completeness
  - – Data currency
  - – Data accuracy
  - – Deducing the true values of objects in data fusion

The veracity of big data

When we talk about big data, we typically mean its quantity:

- ✓ What capacity of a system can cope with the size of the data?
- ✓ Is a query feasible on big data within our available resources?
- ✓ How can we make our queries tractable on big data?

Can we trust the answers to our queries in the data?

No, real-life data is typically dirty; you can't get correct answers to your queries in dirty data no matter how

- ✓ good your queries are, and
- ✓ how fast your system is

Big Data = Data Quantity + Data Quality

A real-life encounter

Mr. Smith, our database records indicate that you owe us an outstanding amount of £5,921 for council tax for 2016

NI#	name	AC	phone	street	city	zip
…	…	…	…	…	…	…
SC35621422	M. Smith	131	3456789	Crichton	EDI	EH8 9LE
SC35621422	M. Smith	020	6728593	Baker	LDN	NW1 6XE

- ✓ Mr. Smith already moved to London in 2015
- ✓ The council database had not been correctly updated
  - both old address and the new one are in the database

50% of bills have errors (phone bill reviews)

Customer records

country	AC	phone	street	city	zip
44	131	1234567	Mayfield	New York	EH8 9LE
44	131	3456789	Crichton	New York	EH8 9LE
01	908	3456789	Mountain Ave	New York	07974

Anything wrong?

- ✓ New York City is moved to the UK (country code: 44)
- ✓ Murray Hill (01-908) in New Jersey is moved to New York state

Error rates: 10% - 75% (telecommunication)

Dirty data are **costly**

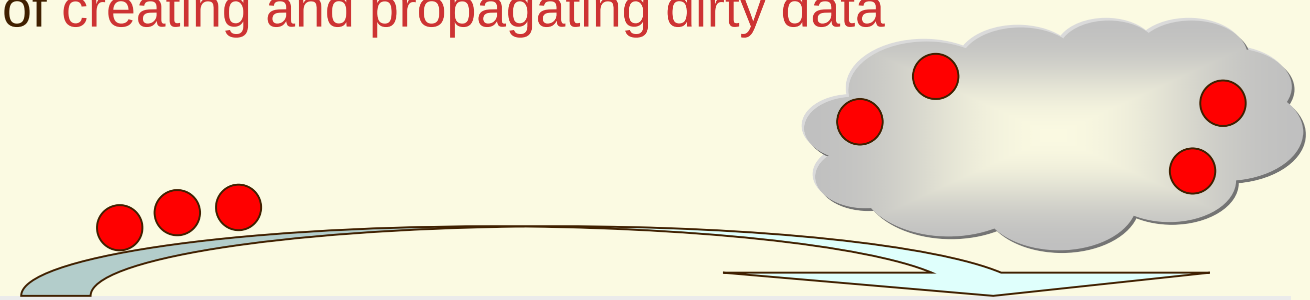
- ✓ Poor data cost US businesses $611 billion annually
- ✓ Erroneously priced data in retail databases cost US customers $2.5 billion each year 2000
- ✓ 1/3 of system development projects were forced to delay or cancel due to poor data quality 2001
- ✓ 30%-80% of the development time and budget for data warehousing are for data cleaning 1998
- ✓ CIA's World FactBook is extremely dirty!

The scale of the problem is even bigger in big data

Big data = quantity + quality!

Far reaching impact

- ✓ Telecommunication: dirty data routinely lead to
  - – failure to bill for services
  - – delay in repairing network problems
  - – unnecessary lease of equipment
  - – misleading financial reports, strategic business planning decision
  - ⇒ loss of revenue, credibility and customers
- ✓ Finance, life sciences, e-government, …
- ✓ A longstanding issue for decades
- ✓ Internet has been increasing the risks, in an unprecedented scale, of creating and propagating dirty data

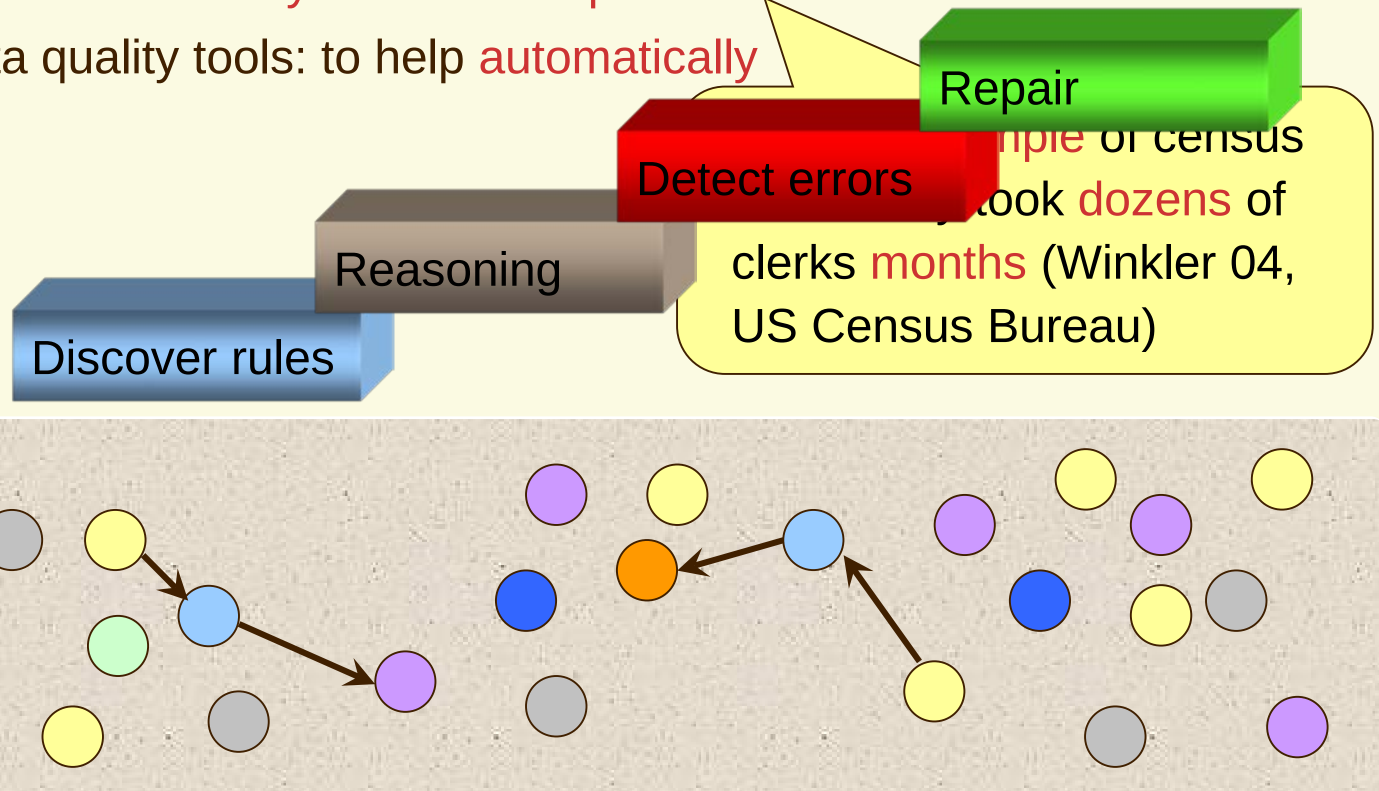


Data quality: The No. 1 problem for data management

The need for data quality tools

- Manual effort: beyond reach in practice
- Data quality tools: to help automatically

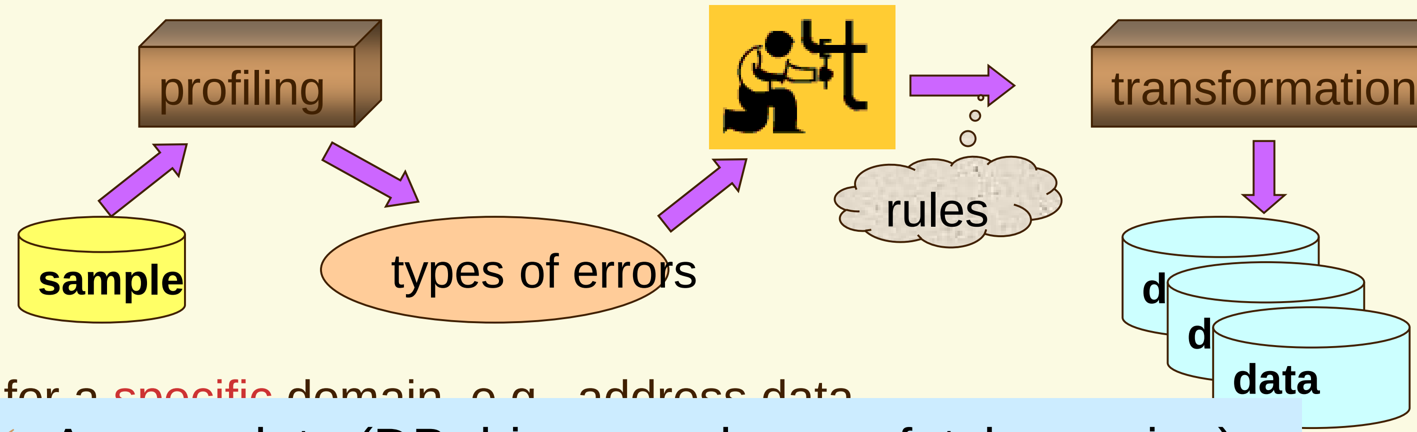
… of census
… took dozens of
clerks months (Winkler 04,
US Census Bureau)



The market for data quality tools is growing at 17% annually

>> *the 7% average of other IT segments* Gartner 2006

ETL (Extraction, Transformation, Loading)



- for a specific domain, e.g., address data

- Access data (DB drivers, web page fetch, parsing)
- Validate data (rules)
- Transform data (e.g. addresses, phone numbers)
- Load data

Not very helpful when processing data with rich semantics

Dependencies: A data cleaning approach

- Errors found in practice
  - Syntactic: a value not in the corresponding domain or range, e.g., name = 1.23, age = 250
  - Semantic: a value representing a real-world entity different from the true value of the entity

    Hard to detect and fix
  - Dependencies: for specifying the semantics of relational data
  - relation (table): a set of tuples (records)

NI#	name	AC	phone	street	city	zip
SC35621422	M. Smith	131	3456789	Crichton	EDI	EH8 9LE
SC35621422	M. Smith	020	6728593	Baker	LDN	NW1 6XE

How can dependencies help?

Data consistency

Data inconsistency

- ✓ The validity and integrity of data
  - – inconsistencies (conflicts, errors) are typically detected as violations of dependencies
- ✓ Inconsistencies in relational data
  - – in a single tuple
  - – across tuples in the same table
  - – across tuples in different (two or more relations)
- ✓ Fix data inconsistencies
  - – inconsistency detection: identifying errors
  - – data repairing: fixing the errors

Dependencies should logically become part of data cleaning process

Inconsistencies in a single tuple

country	area-code	phone	street	city	zip
44	131	1234567	Mayfield	NYC	EH8 9LE

- ✓ In the UK, if the area code is 131, then the city has to be EDI
- ✓ Inconsistency detection:
  - Find all inconsistent tuples
  - In each inconsistent tuple, locate the attributes with inconsistent values
- ✓ Data repairing: correct those inconsistent values such that the data satisfies the dependencies

Error localization and data imputation

Inconsistencies between two tuples

NI# → street, city, zip

- ✓ NI# determines address: for any two records, if they have the same NI#, then they must have the same address
- ✓ for each distinct NI#, there is a unique current address

NI#	name	AC	phone	street	city	zip
SC35621422	M. Smith	131	3456789	Crichton	EDI	EH8 9LE
SC35621422	M. Smith	020	6728593	Baker	LDN	NW1 6XE

- ✓ for SC35621422, at least one of the addresses is not up to date

A simple case of our familiar functional dependencies

Inconsistencies between tuples in different tables

book[asin, title, price] $\subseteq$ item[asin, title, price]

book

asin	isbn	title	price
a23	b32	Harry Potter	17.99
a56	b65	Snow white	7.94

item

asin	title	type	price
a23	Harry Potter	book	17.99
a12	J. Denver	CD	7.94

- Any book sold by a store must be an item carried by the store
  - for any book tuple, there must exist an item tuple such that their asin, title and price attributes pairwise agree with each other

Inclusion dependencies help us detect errors across relations

What dependencies should we use?

Dependencies: different expressive power, and different complexity

country	area-code	phone	street	city	zip
44	131	1234567	Mayfield	NYC	EH8 9LE
44	131	3456789	Crichton	NYC	EH8 9LE
01	908	3456789	Mountain Ave	NYC	07974

- ✓ functional dependencies (FDs)

  country, area-code, phone $\rightarrow$ *street, city, zip*

  country, area-code $\rightarrow$ *city*

The database satisfies the FDs, but the data is not clean!

A central problem is how to tell whether the data is dirty or clean

Conditional functional dependencies – new method

Record matching (entity resolution)

Record matching

To identify records from unreliable data sources that refer to the same real-world entity

FN	LN	address	tel	DOB	gender
Mark	Smith	10 Oak St, EDI, EH8 9LE	3256777	10/27/97	M

the same person?

FN	LN	post	phn	when	where	amount
M.	Smith	10 Oak St, EDI, EH8 9LE	null	1pm/7/7/09	EDI	*$3,500*
…	…	…	…	…	…	…
Max	Smith	PO Box 25, EDI	3256777	2pm/7/7/09	NYC	$6,300

Record linkage, entity resolution, data deduplication, merge/purge, …

Why bother?

Data quality, data integration, payment card fraud detection, …

Records for card holders

FN	LN	address	tel	DOB	gender
Mark	Smith	10 Oak St, EDI, EH8 9LE	3256777	10/27/97	M

fraud?

Transaction records

FN	LN	post	phn	when	where	amount
M.	Smith	10 Oak St, EDI, EH8 9LE	null	1pm/7/7/09	EDI	*$3,500*
…	…	…	…	…	…	…
Max	Smith	PO Box 25, EDI	3256777	2pm/7/7/09	NYC	$6,300

World-wide losses in 2006: $4.84 billion

Nontrivial: A longstanding problem

- ✓ Real-life data are often dirty: errors in the data sources
- ✓ Data are often represented differently in different sources

FN	LN	address	tel	DOB	gender
Mark	Smith	10 Oak St, EDI, EH8 9LE	3256777	10/27/97	M

FN	LN	post	phn	when	where	amount
M.	Smith	10 Oak St, EDI, EH8 9LE	null	1pm/7/7/09	EDI	*$3,500*
…	…	…	…	…	…	…
Max	Smith	PO Box 25, EDI	3256777	2pm/7/7/09	NYC	$6,300

Pairwise comparing attributes via equality only does not work!

Challenges

- ✓ Strike a balance between the efficiency and accuracy
  - – data files are often large, and quadratic time is too costly
    - blocking, windowing to speed up the process
  - – we want the result to be accurate
    - true positive, false positive, true negative, false negative
- ✓ real-life data is dirty
  - – We have to accommodate errors in data sources, and moreover, combine data repairing and record matching
- ✓ matching
  - – records in the same files
  - – records in different (even distributed files)

Record matching can also be done based on dependencies

Information completeness

Incomplete information: a central data quality issue

A database D of UK patients: *patient (name, street, city, zip, YoB)*

A simple query Q1: Find the streets of those patients who

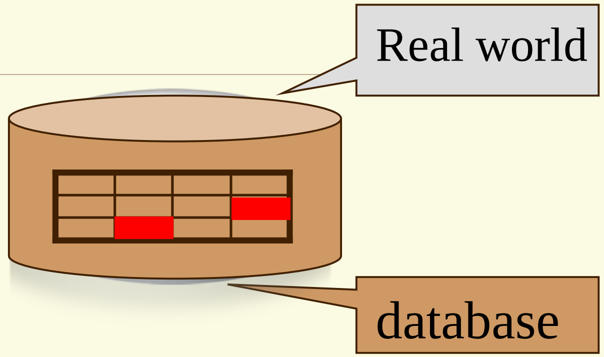
- ✓ were born in 2000 (*YoB*), and
- ✓ live in Edinburgh (Edi) with *zip = "EH8 9AB".*

Can we trust the query to find complete & accurate information?

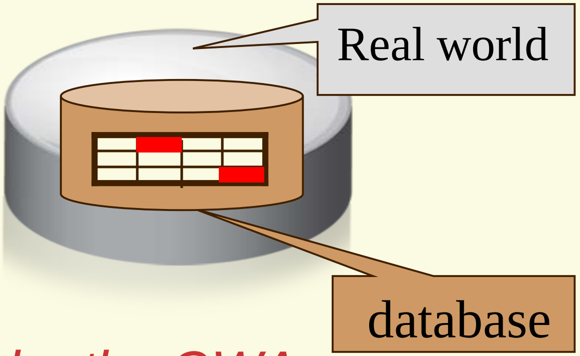
Both tuples and values may be missing from D!

"information perceived as being needed for clinical decisions was unavailable 13.6%--81% of the time" (2006)

Traditional approaches: The CWA vs. the OWA



- The Closed World Assumption (CWA)
  - all the real-world objects are already represented by tuples in the database
  - missing values only



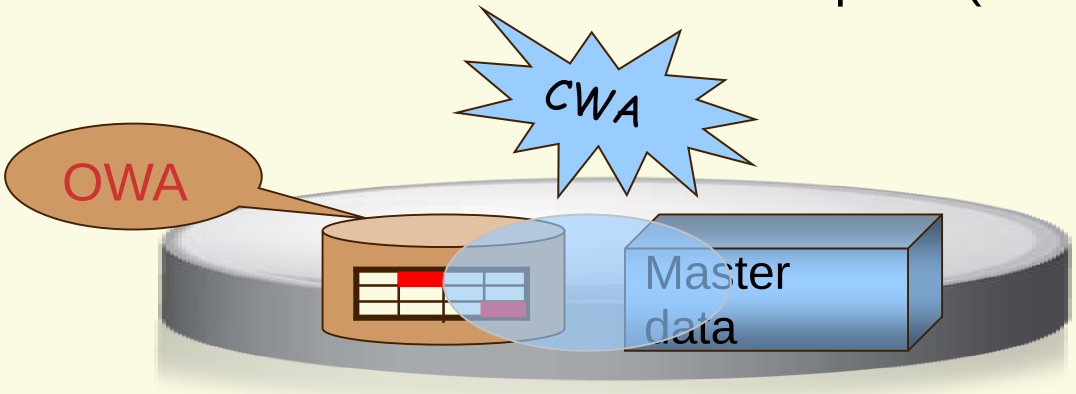
- The Open World Assumption (OWA)
  - the database is a subset of the tuples representing real-world objects
  - missing tuples and missing values

Few queries can find a complete answer under the OWA

None of the CWA or OWA is quite accurate in real life

In real-life applications

Master data (*reference data*)*:* a consistent and complete repository of *the core business entities* of an enterprise (certain categories)



- ✓ *The CWA: the master data – an upper bound of the part constrained*
- ✓ *The OWA: the part not covered by the master data*

Databases in real world are often
neither entirely closed-world, nor entirely open-world

Partially closed databases

- ✓ Master data D_m: *$patient_m$(name, street, zip, YoB)*
  - – *Complete* for Edinburgh patients with YoB > 1990
- ✓ Database D: *patient (name, street, city, zip, YoB)*

  Partially closed:
  - – D_m is an upper bound of Edi patients in D with YoB > 1990
- ✓ Query Q_1*:* Find the streets of all Edinburgh patients with YoB = 2000 and zip = "EH8 9AB".

The *seemingly incomplete* D has complete information to answer Q_1

- ✓ if the answer to Q_1 in D return[s]
  with p[YoB] = 2000 and p[zip]

adding tuples to D does not change its answer to Q1

The database D is complete for Q_1 relative to D_m

Relative information completeness

- *Partially closed databases*: partially constrained by master data; neither CWA nor OWA
- *Relative completeness*: a partially closed database that has complete information to answer a query relative to master data
- The completeness and consistency taken together: containment constraints
- Fundamental problems:
  - Given a partially closed database D, master data D_m, and a query Q, decide whether D is complete Q for relatively to D_m
  - Given master data D_m and a query Q, decide whether there exists a partially closed database D that is complete for Q relatively to D_m

theory of relative information completeness

Data currency

Data currency: another central data quality issue

Data currency: *the state of the data being current*

Data get obsolete quickly: "In a customer file, within two years about 50% of record may become obsolete" (2002)

Multiple values pertaining to the same entity are present

- ✓ The values were once correct, but they have become stale and inaccurate
- ✓ Reliable timestamps are often not available

Identifying stale data is costly and difficult

How can we tell when the data are current or stale?

Determining the currency of data

FN	LN	address	salary	status
Mary	Smith	2 Small St	50k	single
Mary	Dupont	10 Elm St	50k	married
Mary	Dupont	6 Main St	80k	married

Entity: Mary

Identified via record matching

- ✓ Q1: what is Mary's current salary? 80k
- ✓ Temporal constraint: *salary is monotonically increasing*

Determining data currency in the absence of timestamps

Dependencies for determining the currency of data

FN	LN	address	salary	status
Mary	Smith	2 Small St	50k	single
Mary	Dupont	10 Elm St	50k	married
Mary	Dupont	6 Main St	80k	married

- ✓ Q1: what is Mary's current salary? 80k
- ✓ currency constraint: *salary is monotonically increasing*

For any tuples t and t' that refer to the same entity,

- *if t[salary] < t'[salary],*
- *then t'[salary] is more up-to-date (current) than t[salary]*

Reasoning about currency constraints to determine data currency

More on currency constraints

FN	LN	address	salary	status
Mary	Smith	2 Small St	50k	single
Mary	Dupont	10 Elm St	50k	married
Mary	Dupont	6 Main St	80k	married

- ✓ Q2: what is Mary's current last name? Dupont
- ✓ Marital status only changes from *single → married → divorced*

For any tuples t and t', if t[status] = "single" and t'[status] = "married", then t' [status] is more current than t[status]

- ✓ Tuples with the most *current marital status* also have the *most current last name*

if t'[status] is more current than t[status], then so is t'[LN] than t'[LN]

Specify the currency of correlated attributes

A data currency model

- *Data currency model*:
  - Partial temporal orders, currency constraints
- Fundamental problems: Given partial temporal orders, temporal constraints and a set of tuples pertaining to the same entity, to decide
  - whether a value is more current than another?

    Deduction based on constraints and partial temporal orders
  - whether a value is certainly more current than another?

    no matter how one completes the partial temporal orders, the value is always more current than the other

Deducing data currency using constraints and partial temporal orders

Certain current query answering

- *Certain current query answering*: answering queries with the current values of entities (over all possible “consistent completions” of the partial temporal orders)

- Fundamental problems: Given a query Q, partial temporal orders, temporal constraints, a set of tuples pertaining to the same entity, to decide
  - whether a tuple is a certain current answer to a query?

    No matter how we complete the partial temporal orders, the tuple is always in the certain current answers to Q

Fundamental problems have been studied; but efficient algorithms are not yet in place

There is much more to be done

Data accuracy

Data accuracy and relative accuracy

data may be consistent (no conflicts), but not accurate

id	FN	LN	age	job	city	zip
12653	Mary	Smith	25	retired	EDI	EH8 9LE

- ✓ Consistency rule: age < 120. The record is consistent. Is it accurate?

data accuracy: how close a value is to the true value of the entity that it represents?

Relative accuracy: given tuples t and t' pertaining to the same entity and attribute A, decide whether t[A] is more accurate than t'[A]

Challenge: the true value of the entity may be unknown

Determining relative accuracy

id	FN	LN	age	job	city	zip
12653	Mary	Smith	25	retired	EDI	EH8 9LE
12563	Mary	DuPont	65	retired	LDN	W11 2BQ

- Question: which age value is more accurate?

based on context:

- for any tuple t, if t[job] = "retired", then t[age] $\geq$ 60

65

If we know t[job] is accurate

Dependencies for deducing relative accuracy of attributes

Determining relative accuracy

id	FN	LN	age	job	city	zip
12653	Mary	Smith	25	retired	EDI	EH8 9LE
12563	Mary	DuPont	65	retired	LDN	W11 2BQ

- ✓ Question: which zip code is more accurate? W11 2BQ

based on master data:

- ✓ for any tuples t and master tuple s, if t[id] = s[id], then t[zip] should take the value of s[zip]

Id	zip	convict
12563	W11 2BQ	no

Master data

Determining relative accuracy

id	FN	LN	age	job	city	zip
12653	Mary	Smith	25	retired	EDI	EH8 9LE
12563	Mary	DuPont	65	retired	LDN	W11 2BQ

- ✓ Question: which city value is more accurate?

based on co-existence of attributes: LDN

we know that the 2nd zip code is more accurate

- ✓ for any tuples t and t',
  - if t'[zip] is more accurate than t[zip],
  - then t'[city] is more accurate than t[city]

Semantic rules: co-existence

Determining relative accuracy

id	FN	LN	age	status	city	zip
12653	Mary	Smith	25	single	EDI	EH8 9LE
12563	Mary	DuPont	65	married	LDN	W11 2BQ

- Question: which last name is more accurate?

DuPont

based on data currency:

We know "married" is more current than "single"

- for any tuples t and t',
  - if t'[status] is more current than t[status],
  - then t'[LN] is more accurate than t[LN]

Semantic rules: data currency

Computing relative accuracy

- *An accuracy model*: dependencies for deducing relative accuracy, and possibly a set of master data
- Fundamental problems: Given dependencies, master data, and a set of tuples pertaining to the same entity, to decide
  - whether an attribute is more accurate than another?
  - compute the most accurate values for the entity
  - . . .

Reading: Determining the relative accuracy of attributes, SIGMOD 2013

Deducing the true values of entities

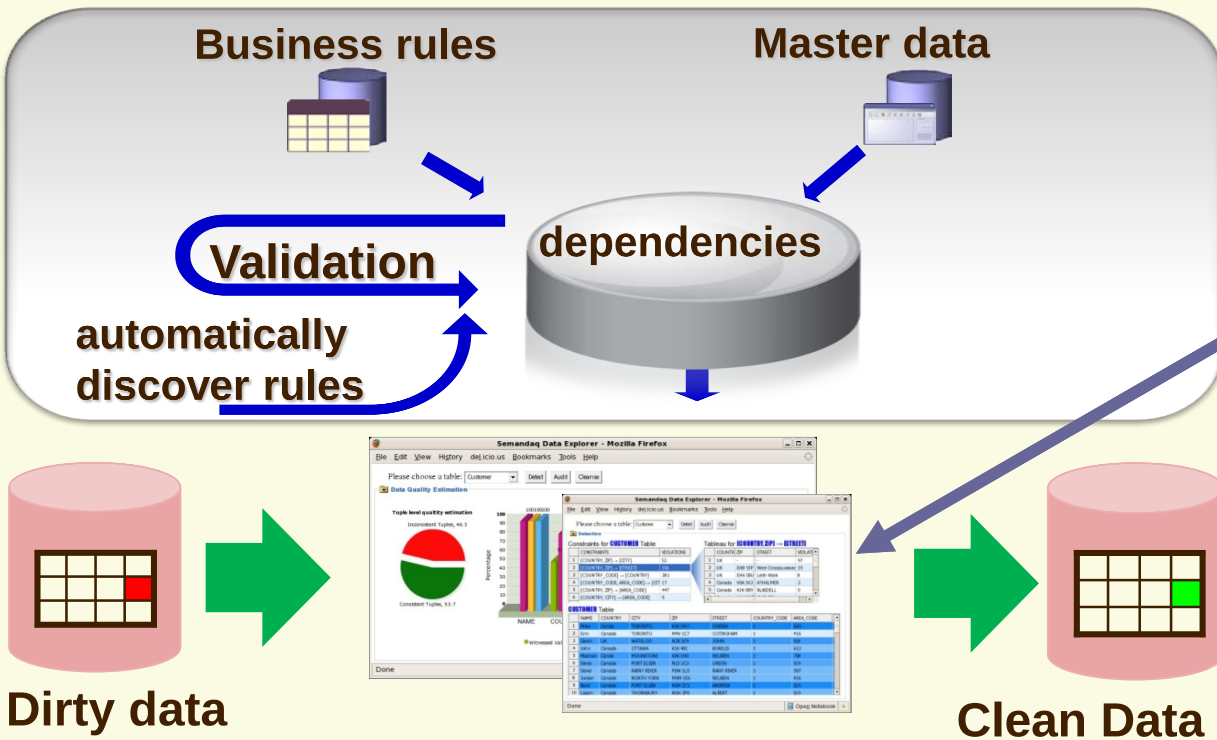
Putting things together

Dependencies for improving data quality

- ✓ The five central issues of data quality can all be modeled in terms of dependencies as data quality rules
- ✓ We can study the interaction of these central issues in the same logic framework
  - – we have to take all five central issues together
  - – These issues interact with each other
    - data repairing and record matching
    - data currency, record matching, data accuracy,
    - …
- ✓ More needs to be done: data beyond relational, distributed data, big data, effective algorithms, …

A uniform logic framework for improving data quality

Improving data quality with dependencies



- Profiling
- Cleaning
- Record matching

- standardization
- data currency
- data enrichment
- data accuracy
- monitoring
- data explorer

Duplicate Payment Detection

Customer Account Consolidation

Credit Card Fraud Detection …

example applications

Opportunities

Look ahead: 2-3 years from now

✓ Big data collection: to accumulate data

Data quality

Assumption: the data collected must be of high quality!

✓ Applications on big data – to make use of big data

Without data quality systems, big data is not much of practical use!

"After 2-3 years, we will see the need for data quality systems substantially increasing, in an unprecedented scale!"

Big challenges, and great opportunities

Challenges

- ✓ Data quality: The No.1 problem for data management

> ✓ *dirty data is everywhere:* telecommunication, life sciences, finance, e-government, …; and *dirty data is costly!*
>
> ✓ *data quality management is a must for coping with big data*

- ✓
  - How to identify entities represented by graphs?
  - How to detect errors from data that comes from a large number of heterogeneous sources?
  - Can we still detect errors in a dataset that is too large even for a linear scan?
  - After we identify errors in big data, can we efficiently repair the data?

The study of data quality is still in its infancy

Summary and Review

- ✓ Why do we have to worry about data quality?
- ✓ What is data consistency? Give an example
- ✓ What is data accuracy?
- ✓ What does information completeness mean?
- ✓ What is data currency (timeliness)?
- ✓ What is entity resolution? Record matching? Data deduplication?
- ✓ What are central issues for data quality? How should we handle these issues?
- ✓ What are new challenges introduced by big data to data quality management?

Reading list

1. W. Fan, F. Geerts, X. Jia and A. Kementsietsidis. Conditional Functional Dependencies for Capturing Data Inconsistencies, TODS, 33(2), 2008.
2. L. Bravo, W. Fan. S. Ma. Extending dependencies with conditions. VLDB 2007.
3. W. Fan, J. Li, X. Jia, and S. Ma. Dynamic constraints for record matching, VLDB, 2009.
4. L. E. Bertossi, S. Kolahi, L.Lakshmanan: Data cleaning and query answering with matching dependencies and matching functions, ICDT 2011. http://people.scs.carleton.ca/~bertossi/papers/matchingDC-full.pdf
5. F. Chiang and M. Miller, Discovering data quality rules, VLDB 2008. http://dblab.cs.toronto.edu/~fchiang/docs/vldb08.pdf
6. L. Golab, H. J. Karloff, F. Korn, D. Srivastava, and B. Yu, On generating near-optimal tableaux for conditional functional dependencies, VLDB 2008. http://www.vldb.org/pvldb/1/1453900.pdf