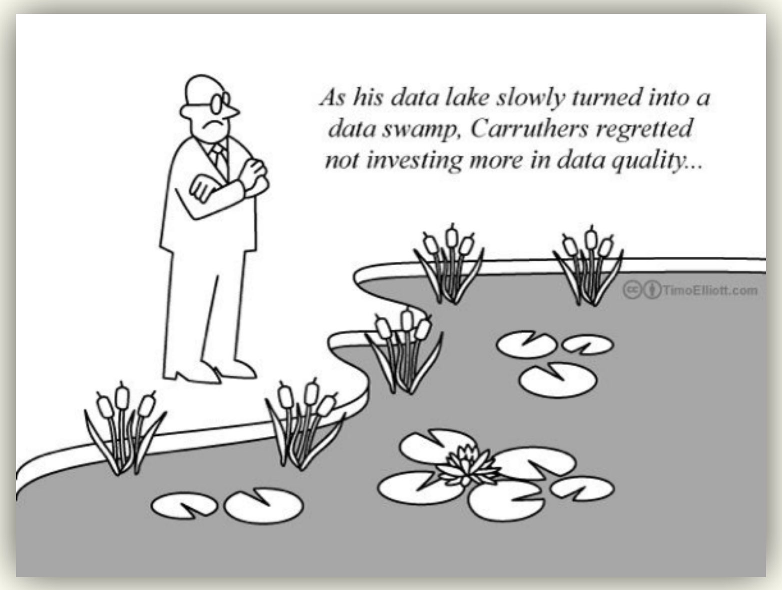
Data Quality I

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		6728593		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
- 1/3 of system development projects were forced to delay or cancel due to poor data quality
- 30%-80% of the development time and budget for data warehousing are for data cleaning

Far Reaching Impact

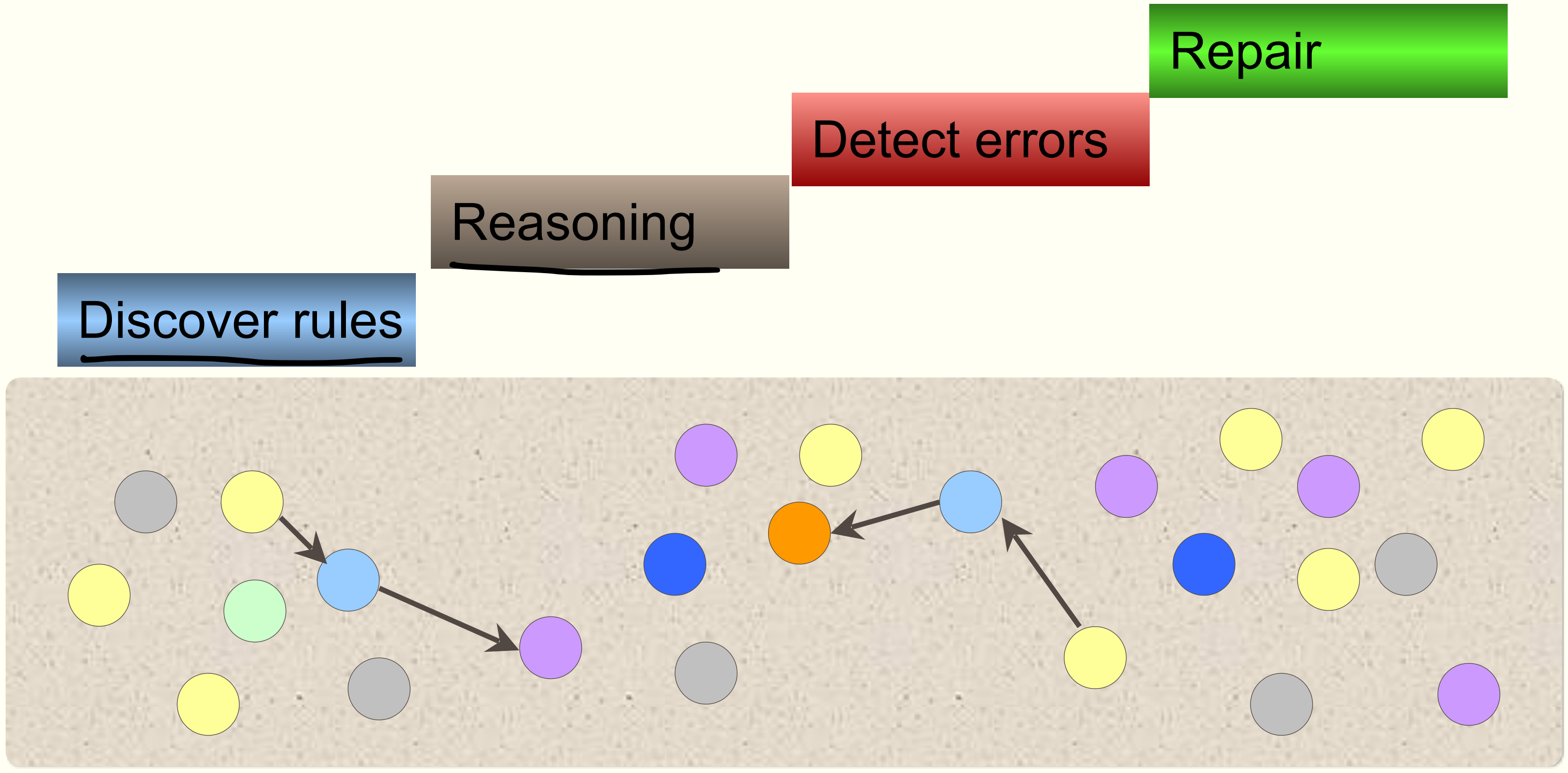
- Telecommunication: dirty data routinely lead to
  - failure to bill for services
  - delay in repairing network problems
  - unnecessary lease of equipment
  - misleading financial reports, strategic business planning decision
  - loss of revenue, credibility and customers
- Finance, life sciences, 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

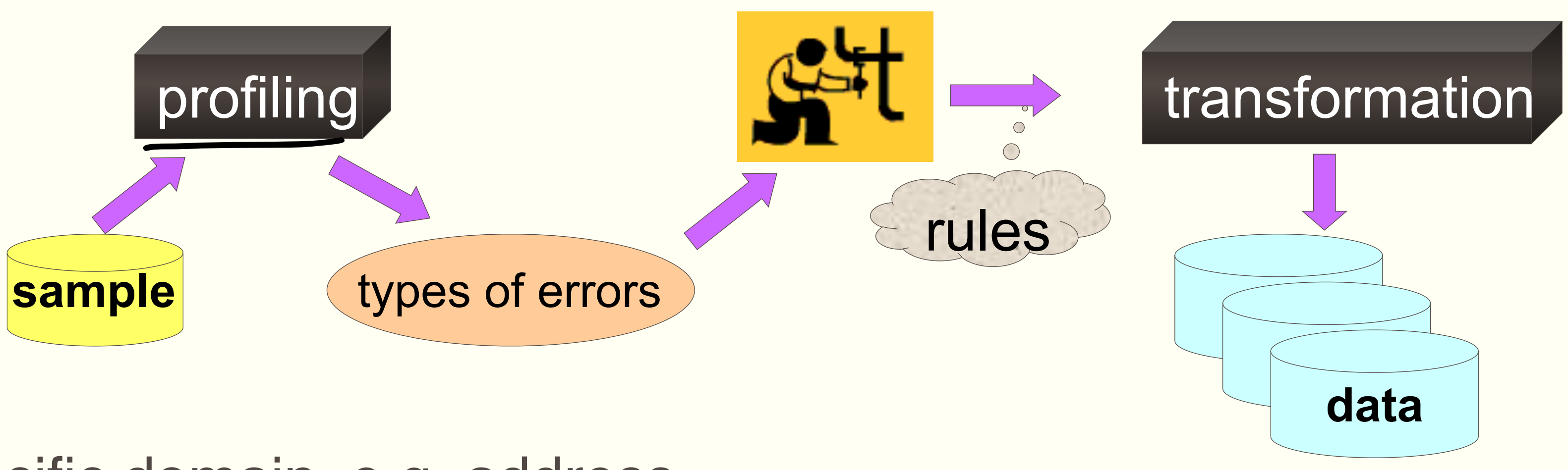
- Manual effort: beyond reach in practice

Editing a sample of census data easily took dozens of clerks months (Winkler 04, US Census Bureau)

- Data quality tools: to help automatically



ETL (Extraction, Transformation, Loading)



- For a specific domain, e.g. address
- Transform rules manually designed
- Low-level programs
  - Difficult to write
  - Difficult to maintain
- What if these rules are dirty?

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

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
  - 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

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

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

Inconsistencies Between Tuples Across Different Tables

- $book[\text{asin}, \ title, \ price] \subseteq item[\text{asin}, \ title, \ price]$

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
t1	44	131	1234567	Mayfield	NYC	EH8 9LE
t2	44	131	3456789	Crichton	NYC	EH8 9LE
t2	01	908	3456789	Mountain Ave	NYC	07974

- functional dependencies (FDs)

  country, area-code, phone → street, city, zip

  country, area-code → city

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

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

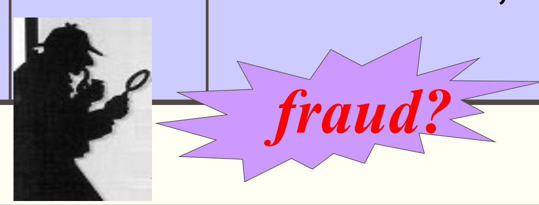
- Record linkage, entity resolution, data deduplication, merge/purge, …

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



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

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

Challenges

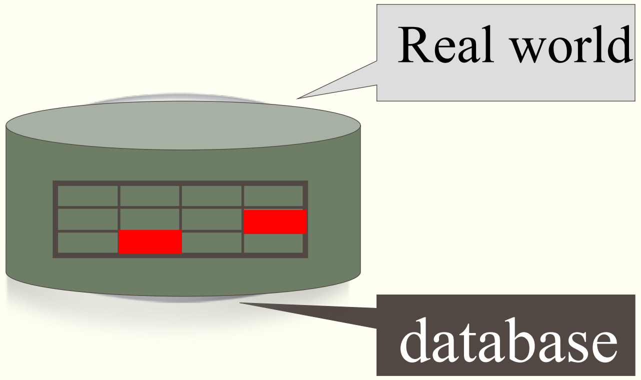
- Strike a balance between the efficiency and accuracy
  - data files are often large, and quadratic time is too costly
    - blocking, windowing to speed up the process
  - we want the result to be accurate
    - true positive, false positive, true negative, false negative
- real-life data is dirty
  - 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)

Incomplete Information: A Central Data Quality Issue

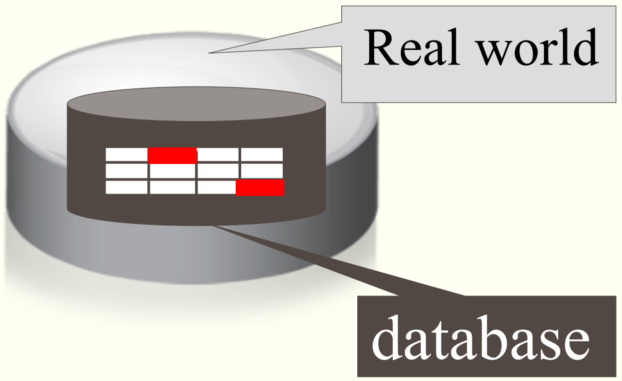
- A database D of UK patients: patient (name, street, city, zip, YoB)
- A simple query Q1: Find the streets of those patients who
  - were born in 2000 (YoB), and
  - live in Edinburgh (Edi) with zip = “EH8 9AB”.
- Can we trust the query to find complete & accurate information?
- Both tuples and values 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

Reading List

- W. Fan, X Jia, J Li and S Ma. Reasoning about record matching rules, VLDB, 2009.
- F. Chiang and M. Miller, Discovering data quality rules, VLDB 2008. http://dblab.cs.toronto.edu/~fchiang/docs/vldb08.pdf
- Leonid Libkin and Christina Sirangelo, Open and Closed World Assumptions in Data Exchange, DL 2009.