Tutorial on Data Quality

Examples* of DQ problems for life sciences and medicine (MedClean Mastodons 2016)

*freely inspired/borrowed from scientific papers in top-most venues of the database community

Outline

Scenario 1: Missing Data (Angela)

Scenario 2: Uncertain Data (Laurent, presented by Radu)

Scenario 3: Inconsistent Data (Ioana)

Scenario 4: Data that cannot be repaired (Angela)

Scenario 5: Temporal Inconsistent Data (Marinette)

Scenario 1: Missing Data

The statistician's Viewpoint

Statisticians make a difference between ‘missing at random’ and ‘not missing at random’ data (the fact that the latter is missing is related to the actual missing data).

Possible options to deal with missing data:

- Imputing missing data with replacement values
- Imputing missing data with uncertainty
- Using statistical models to correlate missing values with the available data

Scenario 1: Missing Data

The database scientist's Viewpoint

In the relational model (relational tables), there is no distinction between the different semantics of missing data.

- Using a plain NULL value (distinct from the empty character string or a string of blank characters or any other number)
- Same NULL value for representing missing/inapplicable/not existing information (or undefined/empty set/not valid/not supplied etc.)
- Solutions:
  - replace null values with probability distributions or allowed intervals (when applicable)
  - replace null values with possible values as in probabilistic databases (example in the next

Scenario 1: Missing Data -->Probabilistic Databases

Table R:

R. SSN	R.NAME
{ 1 (p=.2) | 7 (p=.8) }	John
{ 4 (p=.3) | 7 (p=.7) }	Bill

Hypotheses:

- Assumption of independence of tuples (multiple worlds: probability of the world in which John has SSN 1 and Bill has SSN 7 is 0.2*0.7)

[KO08] Christoph Koch, Dan Olteanu: Conditioning probabilistic databases. PVLDB 1(1): 313-325 (2008)

Scenario 1: Missing Data -->Probabilistic Databases

- Four possible worlds (constraint enforcement may reduce the nr. of possible worlds)

SSN NAME R1 (P = .06)

1 John

4 Bill

SSN NAME R3 (P= .14)

1 John

7 Bill

SSN NAME R2 (P = .24)

7 John

4 Bill

SSN NAME R4 (P= .56)

7 John

7 Bill

Scenario 2: Uncertain Data

Table R:

R.SENSORID	R.TUPLEID	R.TEMP (°C)	R.PROBABILITY
s1	t0	36	0.6
	t1	39	0.4
s2	t2	40	0.7
	t3	36	0.3
S3	t4	37	1

X-tuples: current temperature captured by a sensor for a patient

Example: S1 is 36°C with a probability 0.6

[Mo13] L. Mo et al.: Cleaning uncertain data for top-k queries. ICDE 2013: 134-145

Scenario 2: Uncertain Data -> probabilistic DB

Possible Worlds Semantics (PWS)

Four possible worlds

R1: 0.42

R.SENSORID	R.TUPLEID	R.TEMP (°C)	R.PROBABILITY
s1	t0	36	0.6
s2	t2	40	0.7
S3	t4	37	1

Scenario 2: Uncertain Data -> probabilistic DB

Possible Worlds Semantics (PWS)

Four possible worlds

R2: 0.18

R.SENSORID	R.TUPLEID	R.TEMP (°C)	R.PROBABILITY
s1	t0	36	0.6
s2	t3	36	0.3
S3	t4	37	1

Scenario 2: Uncertain Data -> probabilistic DB

Possible Worlds Semantics (PWS)

Four possible worlds

R3: 0.28

R.SENSORID	R.TUPLEID	R.TEMP (°C)	R.PROBABILITY
s1	t1	39	0.4
s2	t2	40	0.7
S3	t4	37	1

Scenario 2: Uncertain Data -> probabilistic DB

Possible Worlds Semantics (PWS)

Four possible worlds

R4: 0.12

R.SENSORID	R.TUPLEID	R.TEMP (°C)	R.PROBABILITY
s1	t1	39	0.4
s2	t3	36	0.3
S3	t4	37	1

Scenario 2: Uncertain Data -> cleaning

Probe the sensor to get the latest reading

R.SENSORID	R.TUPLEID	R.TEMP (°C)	R.PROBABILITY
s1	t0	36	0.6
	t1	39	0.4
s2	**t3**	**36**	**1**
S3	t4	37	1

Scenario 2: Uncertain Data -> cleaning

Issues

- Cost vs limited resources: battery power, bandwidth, etc.
- Successfulness: operation may fail

Control x-tuples to be cleaned

Scenario 3: Inconsistent Data -> FD Violations

Table R:

R. SSN	R.NAME	R.PHONE
1	John	08
1	Bill	40
4	Cindy	03

- “Common sense” constraint : SSN uniquely determines name, i.e. for the same SSN the name should be exactly the same

Scenario 3: Inconsistent Data -> Repairs

- Remove tuples:

R. SSN	R.NAME	R.PHONE
1	John	08
4	Cindy	03

R. SSN	R.NAME	R.PHONE
1	Bill	40
4	Cindy	03

- Which should we prefer?

- What information do we lose? (i.e. John may have SSN 1 and two phone

Scenario 3: Inconsistent Data -> Repairs

- Replace values:

R. SSN	R.NAME	R.PHONE	R. SSN	R.NAME	R.PHONE
1	John	08	1	Bill	08
1	John	40	1	Bill	40
4	Cindy	03	4	Cindy	03

Scenario 3: Inconsistent Data -> Curated data

R. SSN	R.NAME	R.PHONE
1	JOHN John	08
1	Bill	40
4	Cindy	03

- Value replacement: If we know that the value “John” is correct, we can replace “Bill” by “John”
- Tuple removal: If we know that the first entry (tuple) is correct, we can remove the second entry (i.e. Bill’s entry)

Scenario 3: Inconsistent Data -> Minimum repairs

R. SSN	R.CITY	R.COUNTRY	
1	LONDON	UK	
1	NEW YORK	UK	preferred change: replace by LONDON
4	NEW YORK	US	

- Functional dependencies: SSN -> CITY and CITY-> COUNTRY
- If we change in the second row NEW YORK into LONDON we obtain a correct table with 1 change

Scenario 3: Inconsistent Data -> Metric FDs

R.SSN	R.NAME	R.PHONE
1	John Jr	08
1	John Jr.	40
4	Cindy	03

- Functional dependency: SSN -> NAME: for the same SSN the names should be exactly the same!
- Is this instance really inconsistent?
- Metric functional dependency: SSN ~~> NAME: we only require that for the

Scenario 4: Data that cannot be repaired

- Numerical attributes: which value repairs does one choose? Type 0 (1, resp.) had maximal Flow equal to 1000 (1500, resp)

Traffic

Time	Link	Type	Flow
1.1	a	0	1100
1.1	b	1	900
1.3	b	1	850

[Be08] L.E. Bertossi et al.: The complexity and approximation of fixing numerical attributes in databases under integrity constraints. Inf. Syst. 33(4-5): 407-434 (2008)

Scenario 4: Data that cannot be repaired

- Numerical attributes: possible choices (delete measurement, or update Type or update Flow)

Traffic

Time	Link	Type	Flow
1.1	a	1	1100
1.1	b	1	900
1.3	b	1	850

Traffic

Time	Link	Type	Flow
1.1	a	0	1000
1.1	b	1	900
1.3	b	1	850

Sc4: Only numerical attributes are locally fixable

- New definition of repair, based on a quantitative distance function (overall variation of numerical values is small)
- A least squares repair (LS-repair) for D is a repair D0 that minimizes the square distance $\Delta_\alpha(D, D0)$ between D and D0 over all the instances D

Traffic

Time	Link	Type	Flow	
1.1	a	1	1100	$\Delta_\alpha(D, D1) = 100^2 \times 10^{-5} = 10^{-1}$
1.1	a	0	1000	$\Delta_\alpha(D, D2) = 1^2 \times 1.$

D1 is the only LS-repair.

Scenario 5: Inconsistent Temporal Data

The challenge

*- How facts across different sources are related to one another **over time** ?*

- It is referred to as the **temporal record linkage**

[Li15] F. Li et al. Linking Temporal Records for Profiling Entities. SIGMOD Conference 2015: 593-605

Scenario 5: Inconsistent Temporal Data

What is the problem with time ?

- In traditional record linkage problem, two facts refer to the same entity if the degree of similarity between the two records is high.

- These techniques are typically inadequate for identifying whether or not two records refer to the same entity at different times.

This is because an entity may change several of its attribute values over time (age, location, job…)

The solution

- Using temporal record linkage models

Scenario 5: Inconsistent Temporal Data

Example : Online recruitment system where organizations advertise positions available for job seekers.

The system wants more complete profiles of its users.

Employment history of a job seeker

Name	Organization	Title	Start	End
David Brown	S3 Xjek Aelita Quest Software	Engineer Engineer Manager Manager	2000 2000 2003 2006	2001 2002 2005 2009

F. Li, M. L. Lee, W. Hsu and W-C. Tan : Linking Temporal Records for Profiling Entities - SIGMOD, 2015.

Scenario 5: Inconsistent Temporal Data

Employment history of a job seeker

Name	Organization	Title	Start	End
David Brown	S3 Xjek Aelita Quest Software	Engineer Engineer Manager Manager	2000 2000 2003 2006	2001 2002 2005 2009

Records obtained from various sources

	Name	Organization	Title	Location	Interests	Time	Source
r1	David Brown	S3, Xjek	Engineer			2001	Google+
r2	David Brown		Engineer			2002	Google+
r3	David Brown	S3, Xjek	Engineer			2004	Facebook
r4	David Brown		Manager	Chicago		2004	Twitter
r5	David Brown	Quest Software	Director		Technology	2011	Google+
r6	David Brown	Quest Software	IT Contractor			2011	Google+
r7	David Brown		Engineer	Chicago	Sports, Politics	2012	Facebook
r8	David Brown		President	Chicago		2013	Twitter
r9	David Brown	WSO2	President		Technology	2013	Google+

Scenario 5: Inconsistent Temporal Data

Employment history of a job seeker

Name	Organization	Title	Start	End
David Brown	S3 Xjek Aelita Quest Software	Engineer Engineer Manager Manager	2000 2000 2003 2006	2001 2002 2005 2009

With traditional record linkage : r1-r4 match,
r5-r6 do not refer D. Brown

Records obtained from various sources

	Name	Organization	Title	Location	Interests	Time	Source
r1	David Brown	S3, Xjek	Engineer			2001	Google+
r2	David Brown		Engineer			2002	Google+
r3	David Brown	S3, Xjek	Engineer			2004	Facebook
r4	David Brown		Manager	Chicago		2004	Twitter
r5	David Brown	Quest Software	Director		Technology	2011	Google+
r6	David Brown	Quest Software	IT Contractor			2011	Google+
r7	David Brown		Engineer	Chicago	Sports, Politics	2012	Facebook
r8	David Brown		President	Chicago		2013	Twitter
r9	David Brown	WSO2	President		Technology	2013	Google+

Scenario 5: Inconsistent Temporal Data

Employment history of a job seeker

Name	Organization	Title	Start	End
David Brown	S3 Xjek Aelita Quest Software	Engineer Engineer Manager Manager	2000 2000 2003 2006	2001 2002 2005 2009

With traditional record linkage : r1-r4 match,
r5-r6 do not refer D. Brown

r5 and r6 fall outside the employment history.
They could describe how his job titles evolved in 2011 !

Records obtained from various sources

	Name	Organization	Title	Location	Interests	Time	Source
r1	David Brown	S3, Xjek	Engineer			2001	Google+
r2	David Brown		Engineer			2002	Google+
r3	David Brown	S3, Xjek	Engineer			2004	Facebook
r4	David Brown		Manager	Chicago		2004	Twitter
r5	David Brown	Quest Software	Director		Technology	2011	Google+
r6	David Brown	Quest Software	IT Contractor			2011	Google+
r7	David Brown		Engineer	Chicago	Sports, Politics	2012	Facebook
r8	David Brown		President	Chicago		2013	Twitter
r9	David Brown	WSO2	President		Technology	2013	Google+

Scenario 5: Inconsistent Temporal Data

Employment history of a job seeker

Name	Organization	Title	Start	End
David Brown	S3	Engineer	2000	2001
	Xjek	Engineer	2000	2002
	Aelita	Manager	2003	2005
	Quest Software	Manager	2006	2009

- When the attribute values of an entity change, they do not change arbitrarily (previous value + duration)

- Quality of sources (information published by a source is reliable and up-to-date ⇒ the freshness of sources)

Records obtained from variuos sources

	Name	Organization	Title	Location	Interests	Time	Source
r1	David Brown	S3, Xjek	Engineer			2001	Google+
r2	David Brown		Engineer			2002	Google+
r3	David Brown	S3, Xjek	Engineer			2004	Facebook
r4	David Brown		Manager	Chicago		2004	Twitter
r5	David Brown	Quest Software	Director		Technology	2011	Google+
r6	David Brown	Quest Software	IT Contractor			2011	Google+
r7	David Brown		Engineer	Chicago	Sports, Politics	2012	Facebook
r8	David Brown		President	Chicago		2013	Twitter
r9	David Brown	WSO2	President		Technology	2013	Google+

Scenario 5: Inconsistent Temporal Data

Example : Online recruitment system where organizations advertise positions available for job seekers.

The system wants more complete profiles of its users.

Employment history of a job seeker

Name	Organization	Title	Start	End
David Brown	S3	Engineer	2000	2001
	Xjek	Engineer	2000	2002
	Aelita	Manager	2003	2005
	Quest Software	Manager	2006	2009

Updated profile of David Brown

Organization	Title	Location	Interests	Start	End
S3	Engineer			2000	2001
Xjek	Engineer			2000	2002
Aelita	Manager	Chicago		2003	2005
Quest Software	Manager	Chicago		2006	2009
Quest Software	Director	Chicago	Technology, Sports, Politics	2011	-

Scenario 5: Inconsistent Temporal Data

In the medical domain

- Patients visit multiple medical professionals/organisms over the course of their lifetime, and often even simultaneously.

- Is it interesting

    - To have access to an integrated profile derived from the histories kept by each institution. Through the integrated profile, one could understand when a drug was administered and taken by a patient and for how long ?

    - To determine whether drugs with adverse interactions have been unintentionally prescribed to a patient by different institutions at the same time ?

    - *Discussion* : **meet you this type of problem in your field ?**

B. Alexe, M. Roth and and W-C. Tan : Preference-aware Integration of Temporal Data -PVLDB, 2014.