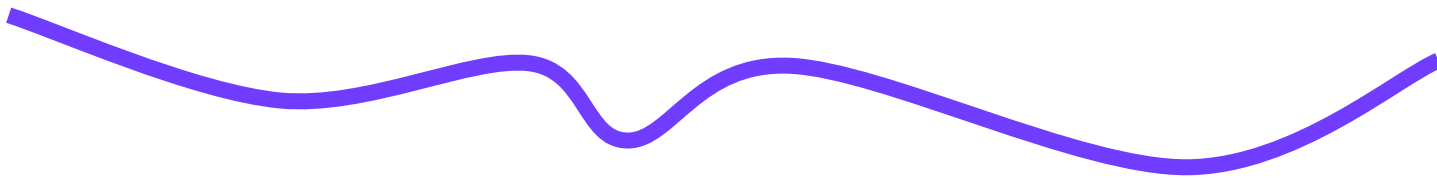




Introduction to Data Quality

Monica Scannapieco

Department of Computer Engineering
Università di Roma "La Sapienza", Italy



Who I am...

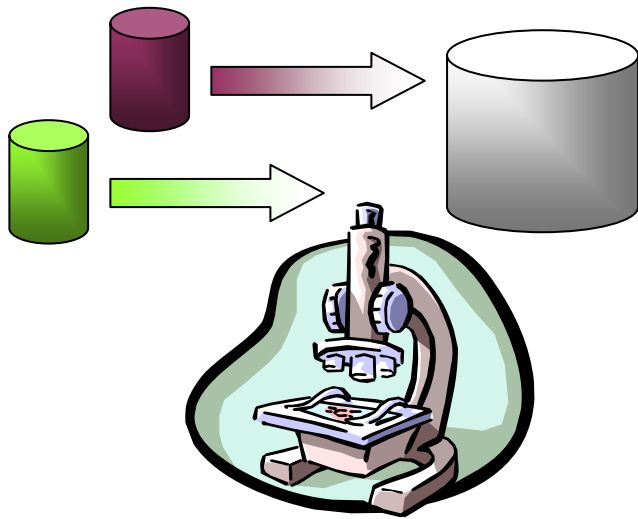
- Research Associate
Computer Engineering Department
Univ. Roma "La Sapienza", Italy
- Web page:
<http://www.dis.uniroma1.it/~monscan/>
- Email: monscan@dis.uniroma1.it

Outline

- Introduction
- What is data quality
- How to assess quality of data
- Data quality issues in modern ISs -
CISs
- Conclusions

Motivations

- In the last decade, increasing demand for :



- Integration

- Analysis

- Exchange

of very large data bases

*Critical Problem:
Data quality !!!*



Cont.

- "Current data quality problems cost U.S. businesses more than 600 billion dollars a year" [Data Warehouse Institute]
- "Between the 30% to 80% of the data analysis task is spent on cleaning and understanding the data" [Data mining practitioners survey]
- E-Government initiatives addressing data quality issues:
 - European directive 2003/98/CE on the reuse of public data
 - USA Data Quality Act

Data Quality

- Generically defined as “fitness for use”
- It is a complex concept, resulting from the “composition” of various *characteristics* or *dimensions*
- Standard set of dimensions not yet defined, though the community agrees on a common minimal set

Data Quality: Multidimensional Concept

- Accuracy

- Jhn vs. John

- Currency

- Residence (Permanent) Address: outdated vs. updated

- Consistency

- ZIP Code and City consistent

- Completeness

Prefix	StreetName	Number	ZipCode	City
Via	Salaria	113	00198	Roma

Prefix	StreetName	Number	ZipCode	City
	Salaria			Roma

Prefix	StreetName	Number	ZipCode	City
Via	Salaria	113	00198	Roma
Via	Gracchi	74	00193	Roma

Prefix	StreetName	Number	ZipCode	City
Via	Salaria	113	00198	Roma

Attribute
Completeness

Entity
Completeness

Example

ID	Title	Director	Year	#Remakes	LastRemakeYear
1	Casablanca	Weir	1942	3	1940
2	Dead Poets Society	Curtiz	1989	0	NULL
3	Rman Holiday	Wylder	1953	0	NULL
4	Sabrina	NULL	1964	0	1985

Cont.

- *Movies* table, with the cells with data quality problems that are shadowed
- At a quick look only the cell corresponding to the title of movie 3 is *wrong*, i.e., there is a misspelling in the title: **accuracy error**
- But many other errors are present!

Cont.

- Swap in the directors of movies 1 and 2 also occurred: **accuracy error**
- Missing value for the director of movie 4 : **completeness error**
- For movie 4 a remake was actually made in 1985, therefore the 0 value for the *#Remakes* attribute is outdated: **currency error**
- For movie 1, the value of *LastRemakeYear* cannot be lower than *Year*, also, for movie 4, the value of *LastRemakeYear* and *#Remakes* are contradicting : **consistency errors**

Accuracy

- Can be evaluated for disparate **granularity levels** of a data model
 - ranging from single values to entire databases
- For single data values, accuracy measures the distance between a value v and a value v' which is considered correct
- Two kinds of accuracy can be identified: ***syntactic accuracy*** and ***semantic accuracy***

Syntactic Accuracy

- Syntactic accuracy is measured by means of *comparison functions* that evaluate the distance between v and v'
- Edit distance is a simple example of comparison function, taking into account *the cost of converting a string s to a string s' through a sequence of character insertions, deletions, and replacements*
- The accuracy error of movie 3 on the *Title* value is a syntactic accuracy error
 - As the correct value for *Rman Holidays* is *Roman Holidays*, the edit distance between the two values is equal to 1

Semantic Accuracy

- Semantic accuracy captures the cases in which v is a syntactically correct value, but it is different from v'
- In the movie example, swapping the directors' names for movies 1 and 2 results in a semantic accuracy error
 - Director named Weir for movie 1 is syntactically correct, but he is not the director of Casablanca

Completeness

- Completeness can be generically defined as “the extent to which data are of sufficient breadth, depth and scope for the task at hand” [Wang 1996]
- Three types of completeness:
 - *Schema completeness*, the degree to which entities and attributes are not missing from the schema
 - *Column completeness*, a function of the missing values in a column of a table
 - *Population completeness*, evaluates missing values with respect to a reference population

Cont.

- If considering a specific data model, more detailed characterization
- Example: relational data model with null values
 - Different possible meanings for null values
 - Different completeness characterizations

Cont. - Null Values Meaning

ID	Name	Surname	BirthDate	Email
1	John	Smith	03/17/1974	smith@abc.it
2	Edward	Monroe	02/03/1967	NULL
3	Anthony	White	01/01/1936	NULL
4	Marianne	Collins	11/20/1955	NULL

Not Existing

Existing But Unknown

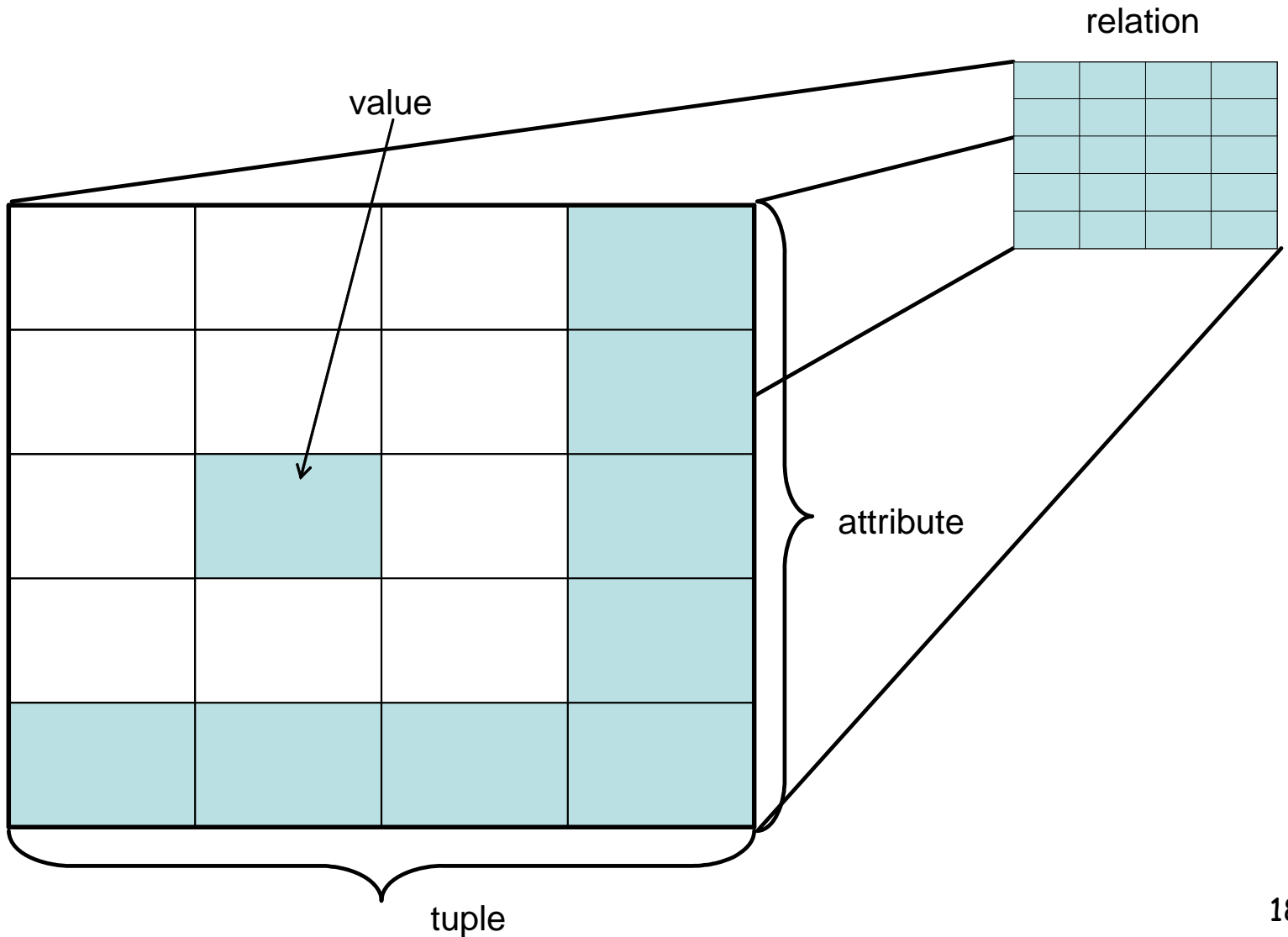
Not Known If Existing

- Tuple 2: no incompleteness
- Tuple 3: incompleteness
- Tuple 4: possible incompleteness

Cont. - Completeness of relational model elements

- *Value completeness*, captures the presence of null values for some attributes of tuples
- *Tuple completeness*, characterizes the completeness of a whole tuple with respect to the values of all attributes
- *Attribute completeness*, measures the number of null values of a specific attribute in a relation
- *Relation completeness*, captures the presence of null values in the whole relation
- But if more tuples should have been present in the relation (Open World Assumption), things are more difficult!!!

Cont. - Completeness of relational model elements



Time Related Dimensions

- Some data are stable in time: birth date, surnames, eye color, etc.
- Some data vary in time: ages, addresses, salaries, etc.
- Three time-related dimensions:
 - Currency
 - Timeliness
 - Volatility

Cont. - Currency

- Currency measures if (or the degree to which) data are updated
- In the movies example, *#Remakes* of movie 4 is not current because a remake of the movie 4 had been performed, but this information did not result in an increased value
- If a residence address of a person is updated, i.e. it actually corresponds to the address where the person lives, then it is current

Cont. - Volatility

- Volatility measures the frequency according to which data vary in time
- Stable data such as birth dates have the lowest value in a given metric scale for volatility, as they do not vary at all
- Stock quotes have a high volatility values

Cont. - Timeliness

- Timeliness measures how current data are, relative to a specific task
- If considering a timetable for university courses: it can be current, thus containing the most recent data, but it can be not timely, if it only becomes available after the start of lessons

Consistency

- The consistency dimension captures the violation of semantic rules defined over (a set of) data items
 - With reference to the relational theory, *integrity constraints* are an instantiation of such semantic rules

Cont. - Integrity Constraints

- Typically distinguished in: intra-relation constraints and inter-relation constraints
- Intra-relation integrity constraints can regard single attributes (also called domain constraints) or multiple attribute of a relation
- Example (intra-relation): in the *Movies* relation the *Year* attribute values must be lower than the *LastRemakeYear* attribute values
- Example (inter-relation): let us consider the *Movies* relation and a relation *OscarAwards*, specifying the oscar awards won by each movie, and including an attribute *Year*. *Movies.Year* must be equal to *OscarAwards.Year*.

Cont. - Consistency of non relational data

- In the statistical area, some data coming from census questionnaires have a structure corresponding to the questionnaire *schema*
- The semantic rules are thus defined over such a structure, very similarly to relational constraints
- Such rules, called *edits*, are less powerful than integrity constraints because they do not rely on a data model like the relational one

Cont. - Edit Example

- As an example, an inconsistent answer in a questionnaire can be to declare:
*marital status as married and
age as 5 years old*
- The rule to detect this kind of errors could be the following: *if marital status is married, age must be not less than 14*
- The rule must be put in form of an edit, which expresses the error condition, namely:
 $(\text{marital status} = \text{married}) \wedge (\text{age} < 14)$

Tradeoffs among dimensions

- Dimensions are not independent (orthogonal)
- Need to establish tradeoffs
- Examples of tradeoffs:
 1. timeliness and a dimension among accuracy, completeness, and consistency.
 - + Timeliness, - (accuracy, consistency, completeness): many Web applications
 - - Timeliness, + (accuracy, consistency, completeness): banking applications

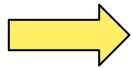
Cont.

2. Consistency and completeness. Is it better to have few but consistent data, i.e. poor completeness? Or is it better to have much more data but inconsistent i.e. poor consistency?
 - + completeness - consistency: statistical data analysis
 - - completeness + consistency: Applications computing list of student votes, salaries etc.

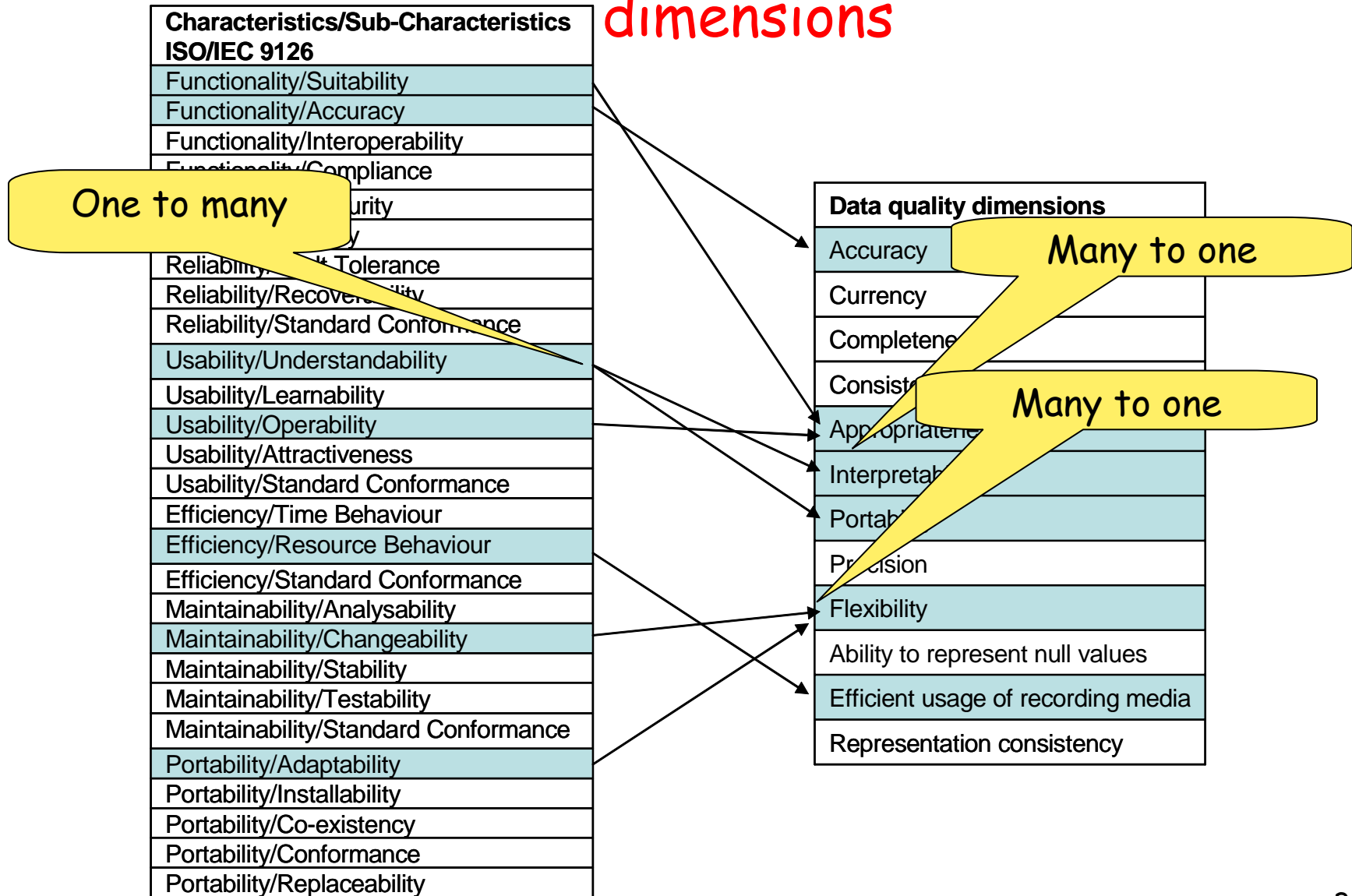
More on Proposals for Sets of DQ Dimensions

- The four dimensions discussed so far are a subset common to proposals in the literature
- Many other dimensions are proposed...
- ...even including security!
- Disagreement also on the exact definition of each dimensions
- Many examples of domain-specific definition: eg. biological domain, scientific data, etc.
- ISO is going to standardize the set

	WandWang 1996	WangStrong 1996	Redman 1996	Jarke 1999	Bovee 2001
Accuracy	X	X	X	X	X
Completeness	X	X	X	X	X
Consistency/ Representational Consistency	X	X	X	X	X
Time-related Dimensions: Currency, Timeliness, volatility	X	X	X	X	X
Interpretability		X	X	X	X
Ease of Understanding/ Understandability		X			
Reliability	X			X	
Credibility				X	X
Believability		X			
Reputation		X			
Objectivity		X			
Relevancy/Relevance		X	X		X
Accessibility		X		X	X
Security /Access Security		X		X	
Value-added		X			
Concise representation		X			
Appropriate amount of data/amount of data:		X	X		
Availability				X	
Portability			X	X	
Responsiveness/ Response Time				X	



DQ dimensions vs ISO 9126 software quality dimensions



How to assess quality of data?

Data Quality Metrics

- Most of the DQ metrics are domain-specific
- Nevertheless, some general techniques can be used for evaluating the dimensions described so far

Syntactic Accuracy Measurement

- Need of data dictionaries to be used as the correct reference sets
- Based on the usage of comparison functions
- Several types of comparison functions

Comparison functions

Type	Example
Identity	'Smith' = 'Smith'
Simple distance	'Smith' similar to 'Smth'
Complex distance	'Smith' similar to 'Smtih'
Error driven distance (Soundex)	Pain, Pane, Payn, Payne, etc. have a soundex code P500
Transformation	John Fitzgerald Kennedy Airport Acronym → JFK Airport

More on comparison functions

1. **Hamming distance** - counts the number of mismatches between two numbers or text strings (→ fixed length strings)
2. **Edit distance** - the minimum cost to convert one of the strings to the other by a sequence of character insertions, deletions and replacements. Each one of these modifications is assigned a cost value.
3. **Jaro's algorithm or distance** finds the number of **common characters** and the number of **transposed characters** in the two strings. J distance accounts for insertions, deletions, and transpositions.
 - A **common character** is a character that appears in both strings within a distance of half the length of the shorter string
 - A **transposed character** is a common character that appears in different positions
 - $F(S1,S2) = (Nc/|S1| + Nc/|S2| + 0.5 Nc/Nt)/3$, where Nc is the number of common characters and Nt is the number of transposed characters, divided by 3 for rescaling

More on comparison functions

4. **N-grams** comparison function forms the set of all the substrings of length n for each string. The distance between the two strings is defined as:
 - $\sqrt{(\sum \forall x |f_a - f_b|)}$, square root of the sum of differences among the number of occurrences of the substring x in the two strings a and b , respectively
5. **Soundex code** clusters together strings that have similar sounds. For example, the Soundex code of "Hilbert" and "Heilbpr" is similar

Main criteria of use

Hamming distance	Used primarily for numerical fixed size fields like Zip Code or SSN
Edit distance	Can be applied to variable length fields. To achieve reasonable accuracy, the modifications costs are tuned for each string data set.
Jaro Distance	The best one in several experiments
N - grams	Bigrams ($n = 2$) effective with minor typographical errors
Soundex code	Effective for dictation mistakes

Semantic Accuracy Measurement

- Needs comparison/matching with a reference source storing the same data
- This requires the ability to recognize that two records refer to the same *real world entity*
- This task is called as *Record Matching* (or also *Record Linkage* or *Object Identification*)

Record Matching

- Record Matching: How we decide that two records represent the same real world entity?
- Example: Is A2 the same record as B2?

Name	Lastname	BirthYear	SSN	ID
Mary	Gold	2000	332211 1004	A1
J. J.	Smith	1974	44335 56677	A2

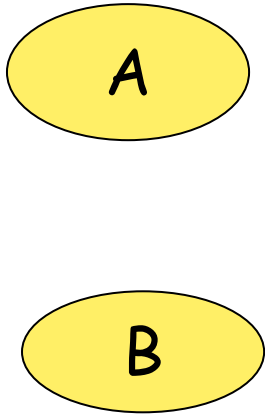
ID	Name	Lastname	Telephone_Nu m	SSN
B1	Marie	Gold	999555	332211 004
B2	John	Smith	222444	

Record Matching Problem Setting

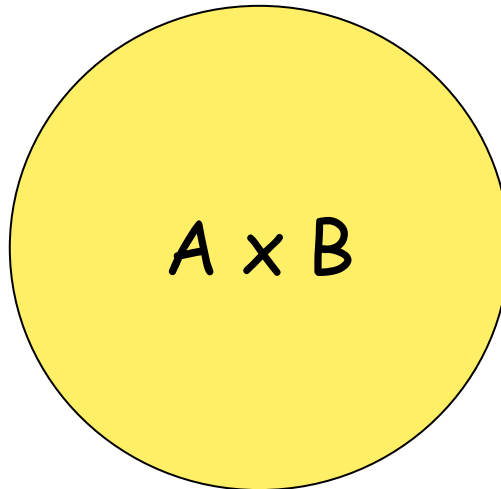
- Given two sets of records A and B , let us consider the cross product $A \times B = \{(a,b) | a \in A \text{ and } b \in B\}$
- Two disjoint sets M and U can be defined starting from $A \times B$, namely:
 - $M = \{(a,b) | a=b, a \in A \text{ and } b \in B\}$ and $U = \{(a,b) | a \neq b, a \in A \text{ and } b \in B\}$
 - M is named as the **matched set** and U is named as the **unmatched set**
- A third set P can be also introduced representing possible matches

Relevant steps

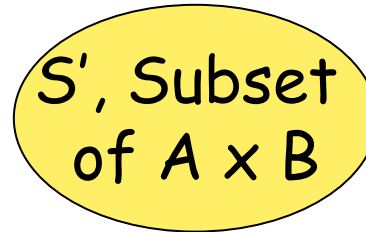
Input files



Initial Search space

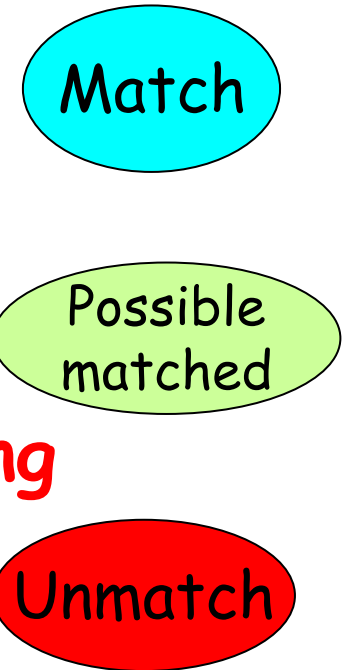


Reduced search Space S'



- Blocking/ Sorting
- Pruning

Assignment



Decision model

General strategy and phases

- 0. Preprocessing
 - Standardize fields to compare and correct simple errors
- 1. Establish blocking/searching method
 - Given the search space $S = A \times B$ of the two files, find a new search space S' contained in S , to apply further steps
- 2. Choose comparison function
 - Choose the function/set of rules that express the distance between pairs of records in S'
- 3. Choose decision model
 - Choose the method for assigning pairs in S' to M , the set of matching records, U the set of unmatching records, and P the set of possible matches
- 4. Check effectiveness of method

Completeness Measurement

- Easy to compute if only counting null values (taking into account their meaning)
- Concept of *domain coverage* for more complex metrics
- E.g.: completeness of CiteSeer (all on-line CS papers) vs. completeness of DBLP (DB, Logic Programming and Algorithmic papers)

Time-Related Dimension Measurement

- **Currency:**
 - *last update* metadata, i.e., the last time in which the specific data have been updated
- Data types changing with a fixed frequency: currency easily computed
- Data types changing with a variable frequency: a possibility is to calculate an average change frequency, and compute currency tolerating errors rates
 - E.g. if a data source stores residence addresses that are estimated to change each 5 years, then an address with a last update metadata which is 1 month before the observation time, can be estimated to be *current*

Cont.

- **Timeliness** measurement implies that not only data are current, but are also in time for a specific usage
- Easy measurement consists of:
 - currency measurement
 - check if data are available *before* the planned usage time
- **Volatility** is a dimension that inherently characterizes types of data, no need for specific metrics

Consistency Measurement

- Integrity constraints have been largely studied in the DB area, and the enforcement of dependencies (e.g. key dependency, functional dependency, etc.) is present in modern database systems
- The violation of integrity constraints in legacy database systems can be quite easily checked from an application layer encoding the consistency rules

Cont. - Edit Checking

- The problem of localizing errors by means of edits and imputing erroneous field is usually referred to as the **edit-imputation problem**
- The Fellegi-Holt method is a well-known theoretical model for editing with the following three main goals:
 - The data in each record should satisfy all edits by changing the fewest fields.
 - Imputation rules should be derived automatically from edits.
 - When imputation is necessary it is desirable to maintain the marginal and joint frequency distribution of variables.
- Many methods for practically solving the edit imputation problem [Winkler 2004]

When Data Quality are critical? Issues in CISSs

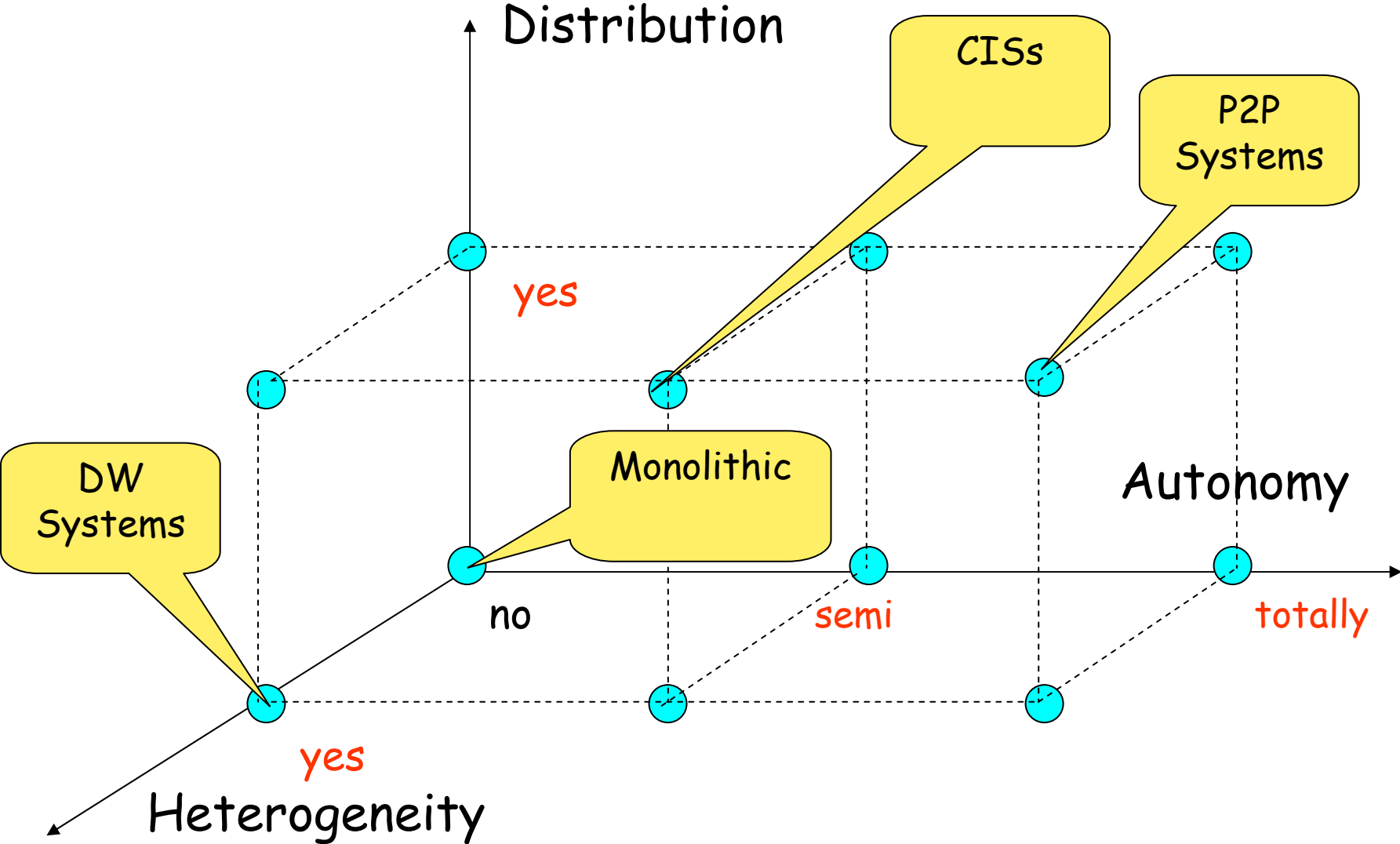
Cooperative Information Systems (CISs)

- *"... Distributed and heterogeneous information systems that cooperate requesting and sharing information, constraints, and goals ..."*

[Mylopoulos et al 1997] J. Mylopoulos, M. Papazoglou (eds.): Cooperative Information Systems. *IEEE Expert Intelligent Systems & Their Applications*, vol. 12, no. 5, September/October 1997

- CISs include:
 - Data integration systems
 - Cross-organization workflow management systems
 - ...
- Examples of contexts requiring CISs
 - Set of public administrations in an e-Government scenario
 - Set of companies of a virtual enterprise
 - ...

Types of information systems



Data Quality & CISs

- CISs features:
 - Data sharing to accomplish cooperative tasks
 - High data replication

Instance level
heterogeneities, not
trusted data



CISs need data quality

High data replication,
different available
sources

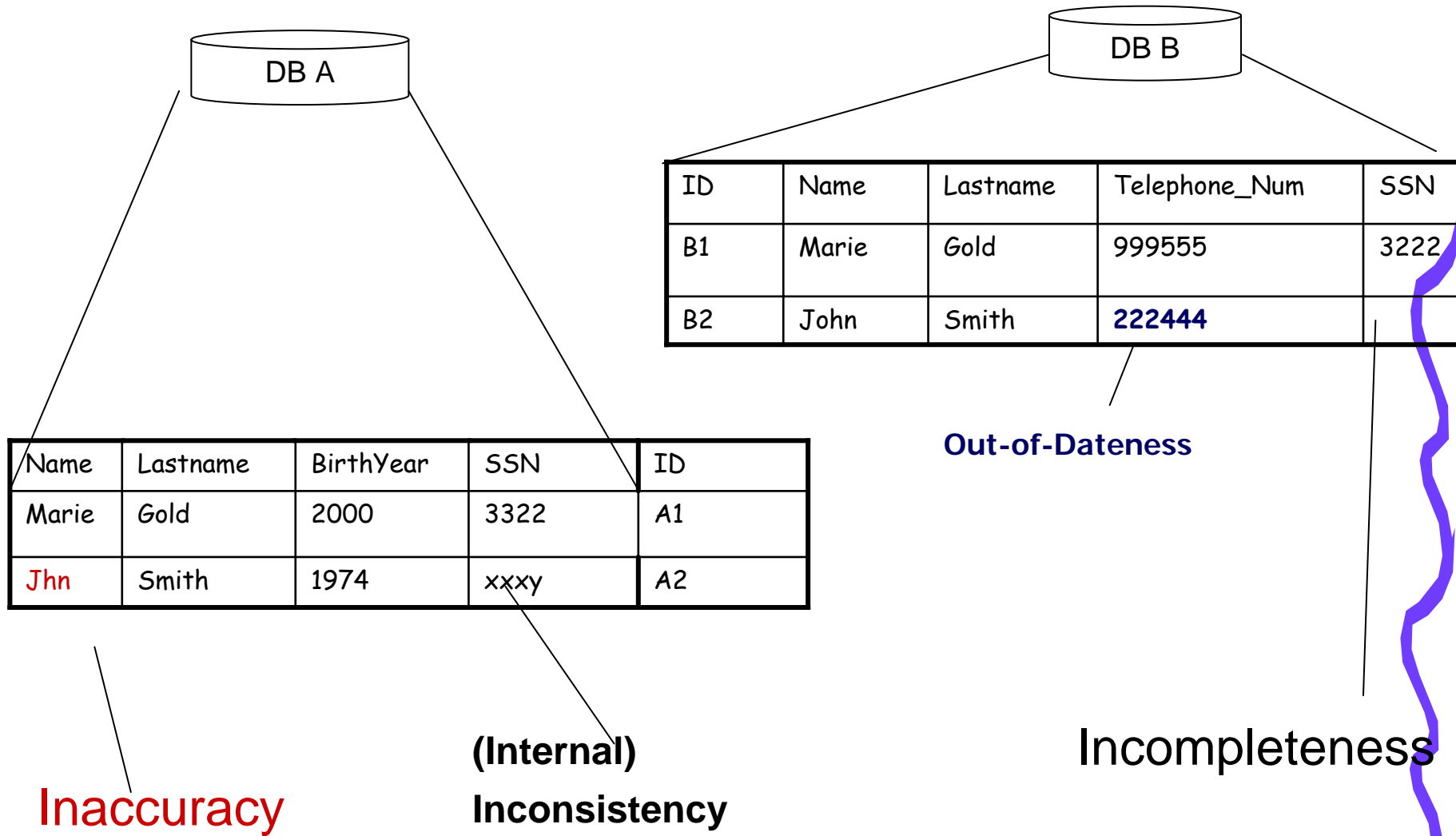


Data quality needs CISs

CISs need Data Quality-Data Inconsistency

- Copies of the same data are stored by different organizations (or within a single organization)
- Problem: data inconsistency!!!

Data Inconsistency



Data Quality needs CISs

- Replication can be exploited to correct data by means of Record Matching
- More details in [Scannapieco-2004]

Conclusions

- We have learnt:
 - **what is data quality** in terms of its major dimensions ...
 - ... that are however a subset of larger possible sets that can take into account domain specific features
 - some elements on **how assessing quality of data** in general contexts
 - **data quality is critical in** modern information systems, like **CISs**

References

- For a system implementing DQ issues in CISs: [Scannapieco et al 2004] M. Scannapieco, A. Virgillito, M. Marchetti, M. Mecella, R. Baldoni: The DaQuinCIS Architecture: a Platform for Exchanging and Improving Data Quality in Cooperative Information Systems. *Information Systems*, vol. 29, no. 7, 2004.
- Introduction to DQ: [Scannapieco et al 2005] M. Scannapieco, P. Missier, C. Batini: Data Quality at a Glance, To appear on Databank Spektrum 2005.
- Introduction to DQ: [Wang-1996] Wang R.Y., Strong D.M.: "Beyond Accuracy: What Data Quality Means to Data Consumers". *Journal of Management Information Systems*, 12(4), 1996.
- Statistical perspective on DQ: [Winkler-2004] Winkler, W.E. , Methods for Evaluating and Creating Data Quality, *Information Systems*, vol. 29, no. 7, 2004.

Write me an email for references
on specific topics you are interested in !!!