New Directions for Data Quality Mining

Laure Berti-Équille

University of Rennes 1, France

berti@irisa.fr

Tamraparni Dasu

AT&T Labs-Research, NJ, USA

tamr@research.att.com

KDD 2009: Paris, France
June 28, 2009

Outline

Part I. Introduction to Data Quality Research

Part II. Data Quality Mining

Part III. Case Study

Part I. Introduction to Data Quality Research

1. Illustrative Examples
2. Definitions, concepts and motivation
3. Current solutions and their limits

What is Low Data Quality?

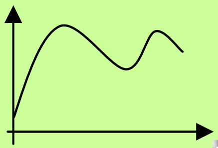
- Missing data
- Erroneous data
- Data anomalies
- Duplicates
- Inconsistent data
- Out-of-date data
- Undocumented data

Data Quality Problems

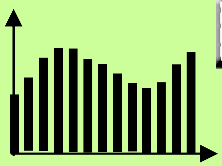
DATA

TYPE

Continuous



Discrete



Binary: 0101010101

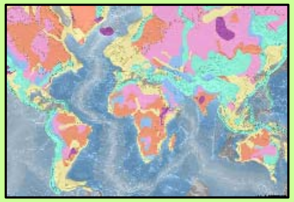
Sequence: ACACGTGT

Nominal: John Doe

Categorical: High / Medium / Low

Multimedia

Geomedia



RELATIONSHIP

- Structural (record)
- Sequential
- Graph-based
- Temporal
- Spatial
- Spatio-Temporal

DATA QUALITY PROBLEM

TYPE

- Missing data
- Anomalous data
- Duplicate data
- Inconsistent data
- Obsolete data

CARDINALITY

- Single-Point
- Collection

DETECTION MODE

- Model-based
- Data distribution-based
- Constraint-based
- Pattern-based

Part I. Introduction to Data Quality Research

1. Illustrative Examples
2. Definitions, concepts and motivation
3. Current solutions and their limits

Example 1

Data quality problems in a relational DB

SIGKDD Executive Committee

Non-standard representation

Name	*Affiliation*	*City, State, Zip, Country*	*Phone*
Piatetsky-Shapiro G.,PhD	KDDnuggets		617-264-9914
Usama Fayyad	Yahoo, Sunnyvale, CA	USA	
Jiawei Han	Univ. of Illinois	IL 61801, USA	(217) 333-6903
Usama Fayad			
Raghu Ramakrishnan	U. Wisconsin-Madison		
Sonata Sarawagi	IIT Bombay	Mumbai-400076 USA	
Robert Grossman	Open Data Group	Chicago IL, USA	999-999-9999
David Jensen	U. of Massachusetts	Amherst, MA, USA	111-111-1111

Duplicates

Typos

Misfielded Value

Inconsistency

Obsolete Value

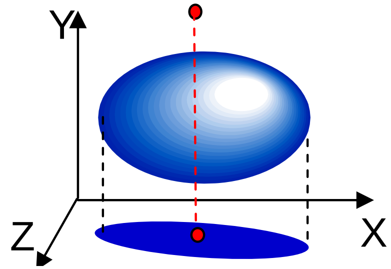
Missing Value

Incorrect Value

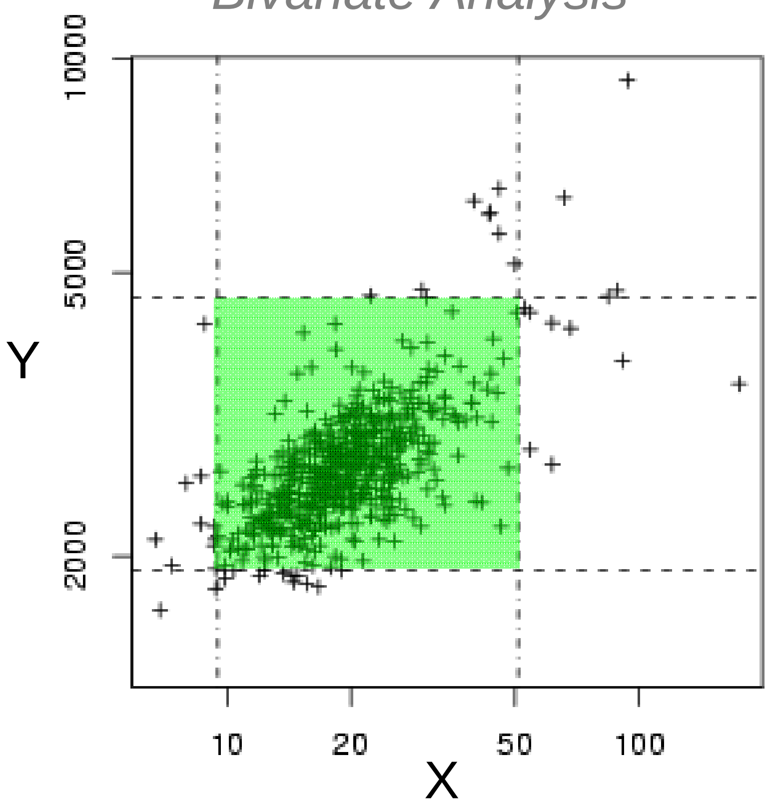
Incomplete Value

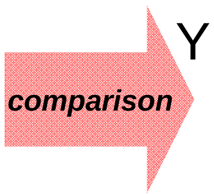
Example 2

Outliers

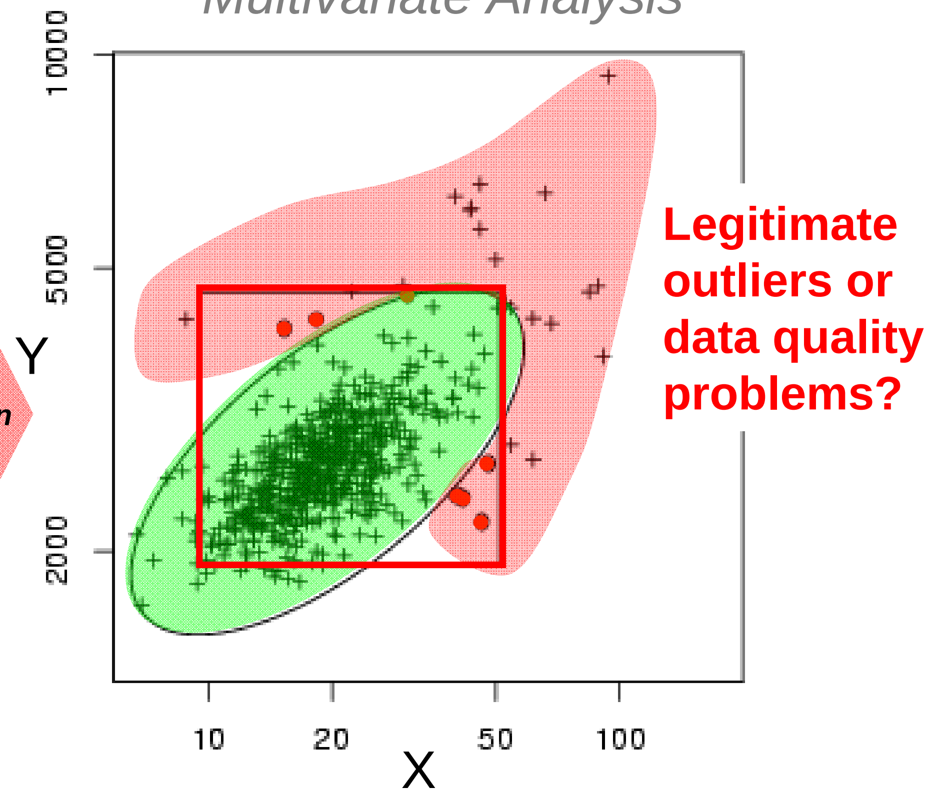


Bivariate Analysis





Multivariate Analysis



Rejection area: Data space excluding the area defined between 2% and 98% quantiles for X and Y

Rejection area based on:

Mahalanobis_dist(cov(X,Y)) > χ^2(.98,2)

Example 3

Disguised missing data

Some are obvious...

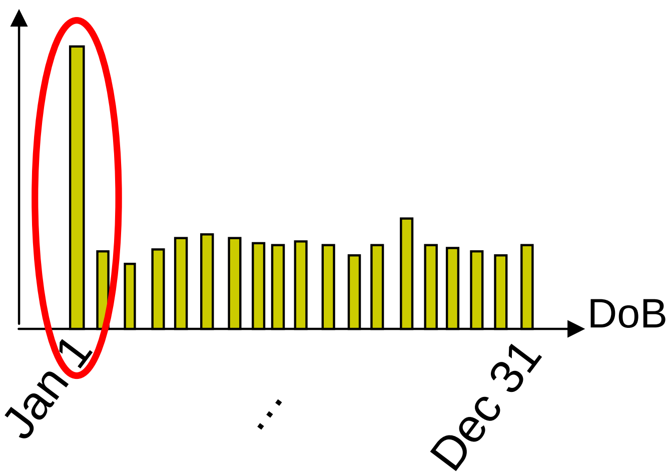
Detectable with syntactical or domain constraints

*Phone number: **999-999-9999***

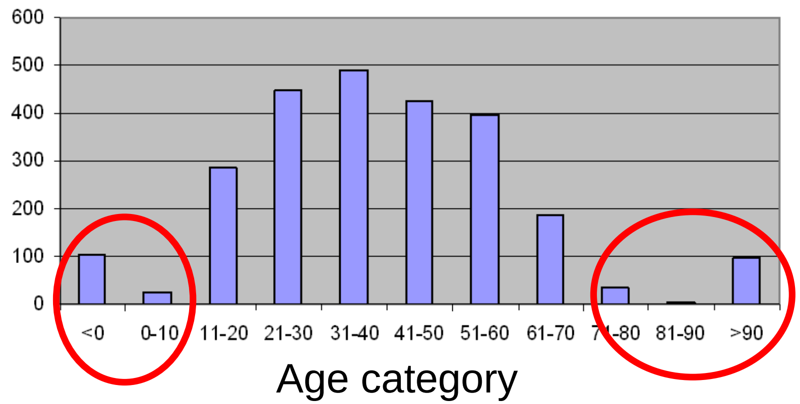
Others are not….

Could be suspected because the data distribution doesn’t conform to the expected model

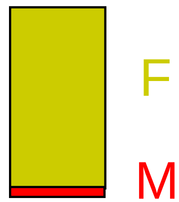
Histogram of DoBs per day of the year



Histogram of online shopping customers per age category



2% patients in the obstetrical emergency service are male…

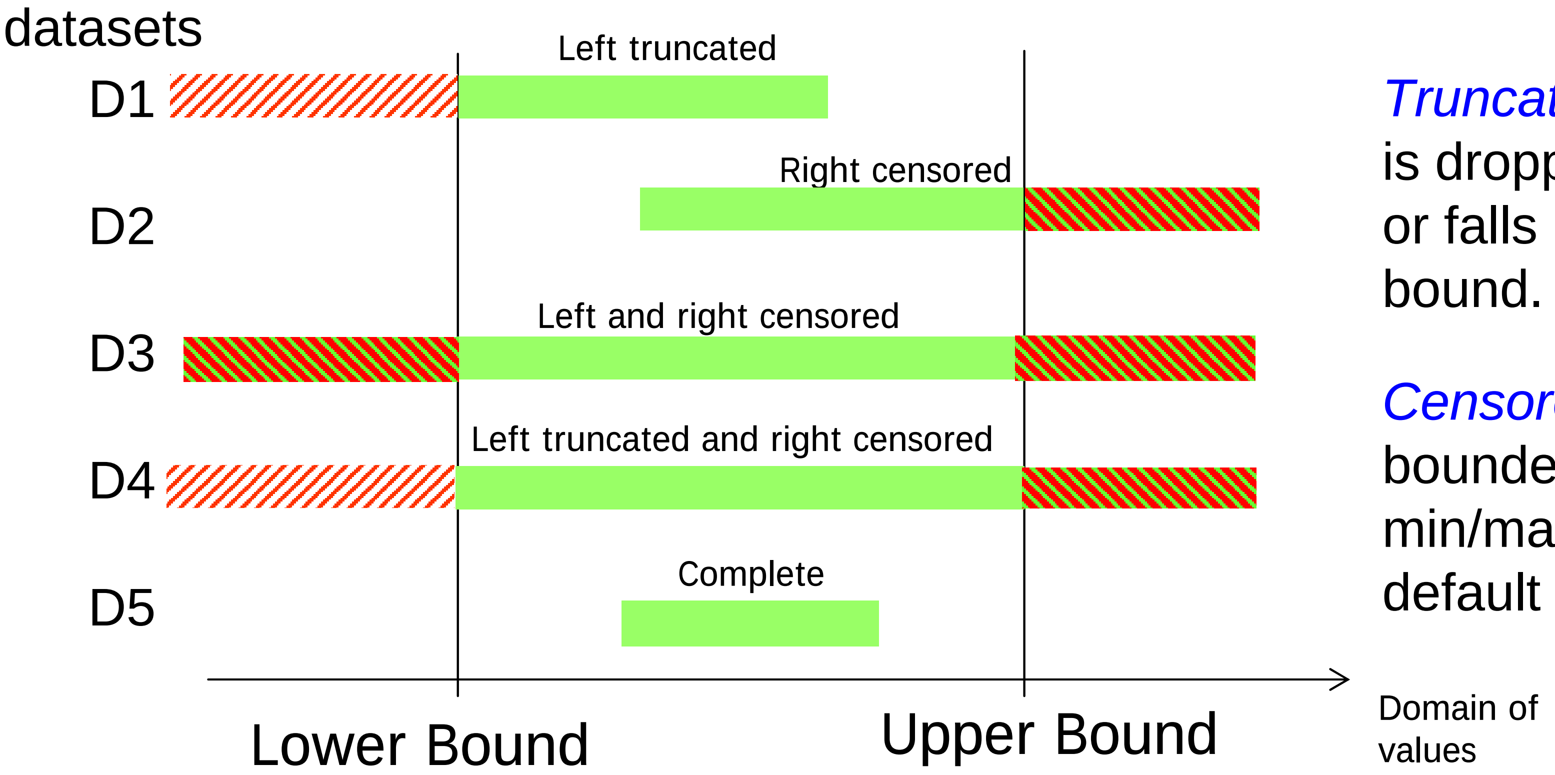


Example 4

Censored and Truncated Data

Phone call datasets

e.g., phone calls whose duration is < 1 second or > 6 hours



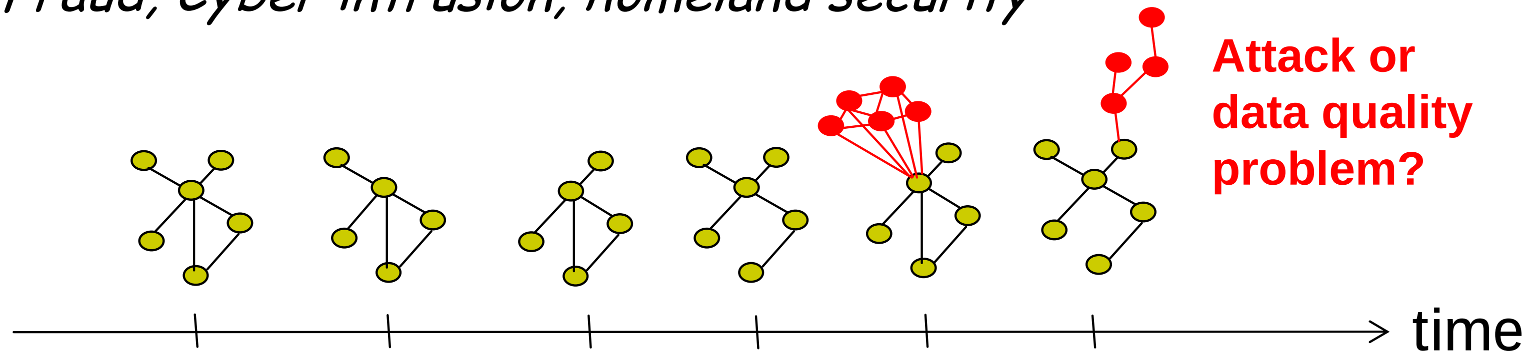
Truncated - Data point is dropped if it exceeds or falls below a certain bound.

Censored - Data is bounded to a fixed min/max limit or a default value.

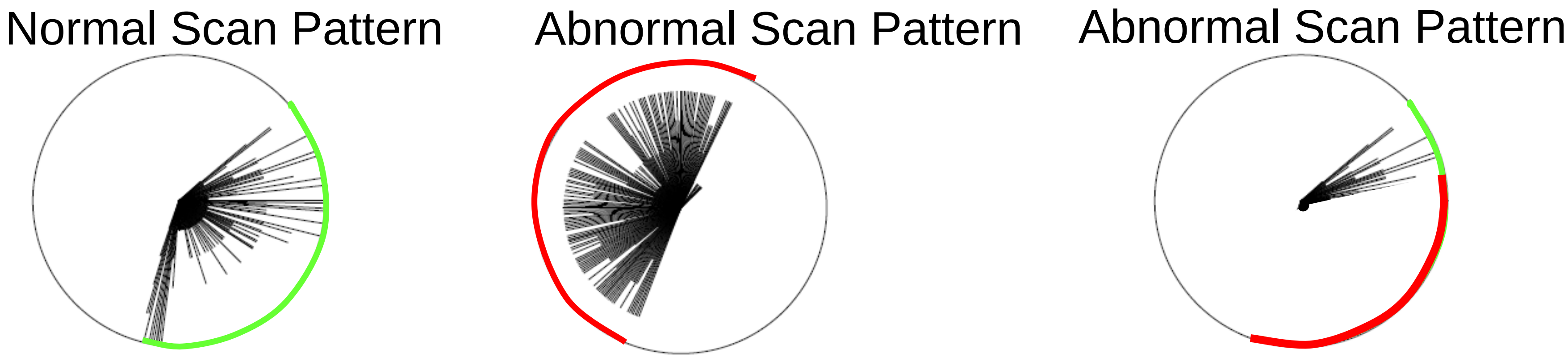
Example 5

Time-Dependent Anomalies : Unusual patterns in graph analysis

e.g., Fraud, Cyber intrusion, homeland security



e.g., IP Address Scan Patterns for a big server



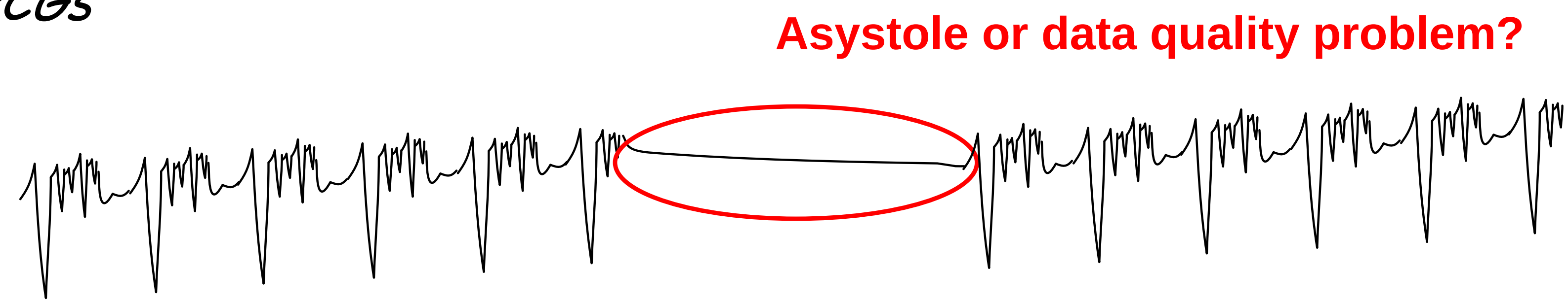
High volume communications with unknown IP addresses

Data loss due to transmission problems

Example 6

Anomalous subsequence

e.g., ECGs



Deviants in time-series

e.g, Sensors

Example 7

Contradictions between Images and Text

flickr **Abuse of tags**

Arbutus tree

ADD TO FAVES · BLOG THIS · ALL SIZES

Tags

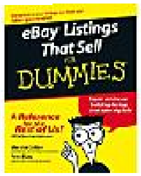
arbutus, tree, galiano, island, amsterdam, animal, animals, april, architecture, art

[...]

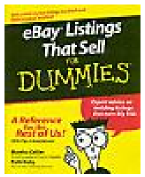
park, party, people, phone, photo, pink, portrait, red, reflection, river, roadtrip, rock, rome, sanfrancisco, school, scotland, sea, seattle, sign, sky, snow, spain, spring, street, summer, sun, sunset, taiwan, texas, thailand, tokyo, toronto, travel, trees, trip, uk, unfound, urban, usa, vacation, vancouver, washington, water, wedding, white, winter, yellow, zoo

Duplicates

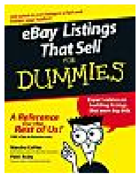
ebay **Fraud**



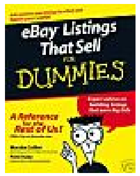
eBay Listings That Sell for Dummies - Collier, Mars New
Current Price: £11.72
View similar items...



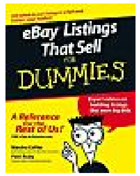
EBAY Listings That Sell Dummies E-Commerce B VALUE BOOK
Current Price: £12.89
View similar items...



EBay Listings That Sell for Dummies Book | Marsha Colli
Current Price: £13.19
View similar items...



Ebay Listings that sell for Dummies On CD- Cheap Book
Current Price: £2.99
View similar items...



eBay Listings That Sell For Dummies
Current Price: £13.19
View similar items...

Guide ID: 10000000001248134 created: 29/06/06 (updated 19/02/09) from http://reviews.ebay.co.uk/Item-Picture-Description-Theft-Prevent-eBay-Fraud_W0QQugidZ10000000001248134

Example 8

False information

Telegraph.co.uk

Home | News | Sport | Finance | Comment | Travel | Lifestyle | Culture | Fas

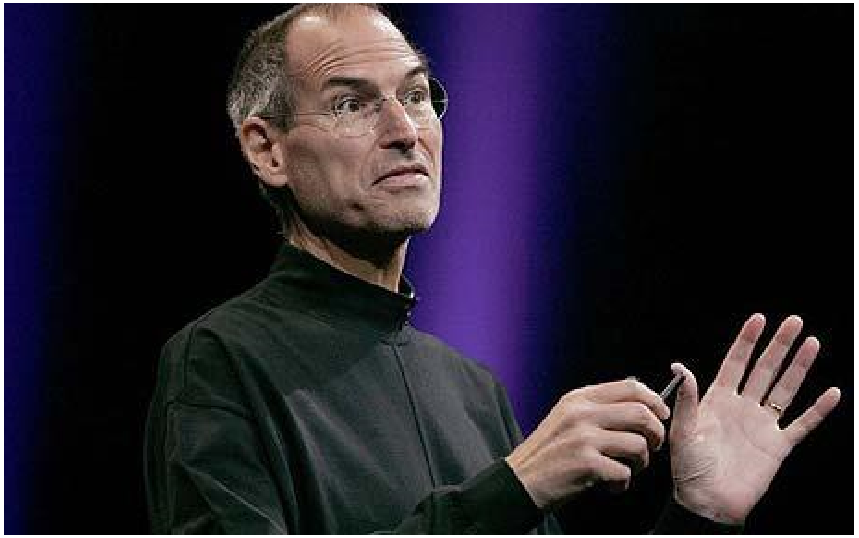
UK | World | Politics | Celebrities | Obituaries | Weird | Earth | Science | Health News | Educatio

HOME > NEWS > NEWS TOPICS > HOW ABOUT THAT?

Steve Jobs obituary published by Bloomberg

An obituary of very-much-alive Apple founder Steve Jobs has been accidentally published by the respected Bloomberg business news wire.

By Matthew Moore
Last Updated: 7:05PM BST 28 Aug 2008



Steve Jobs was described as the man who 'refashioned the mobile phone' in the erroneous obituary   Photo: REUTERS

The story, marked "Hold for release – Do not use", was sent in error to the news service's thousands of corporate clients.

Text Size + −
Email this article
Print this article
Share this article
91 diggs digg it

How about that?
USA
News

The week in pictures

Pictures of the day

Part I. Introduction to Data Quality Research

1. Illustrative Examples
2. Definitions, concepts and motivation
3. Current solutions and their limits

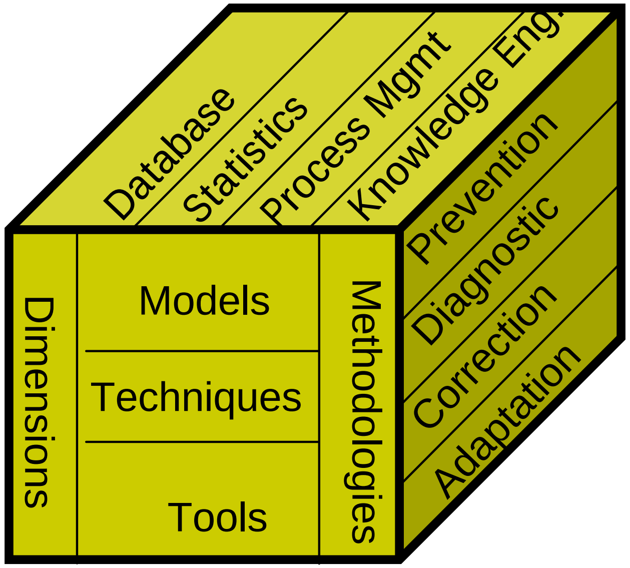
What is Data Quality?

A “subtle” combination of measurable dimensions:

- **Accuracy**
  - KDD’09 location is in Paris, France
- **Consistency**
  - Only one KDD conference per year
- **Completeness**
  - Every past KDD conference had a location
- **Freshness**
  - The location of the current KDD conference is in Paris
- **Uniqueness – no duplicate**
  - KDD is a conference, not the French hip-hop rap band
  - KDD’09, KDD 2009 and Knowledge Discovery and Data mining 2009 are the same conference edition

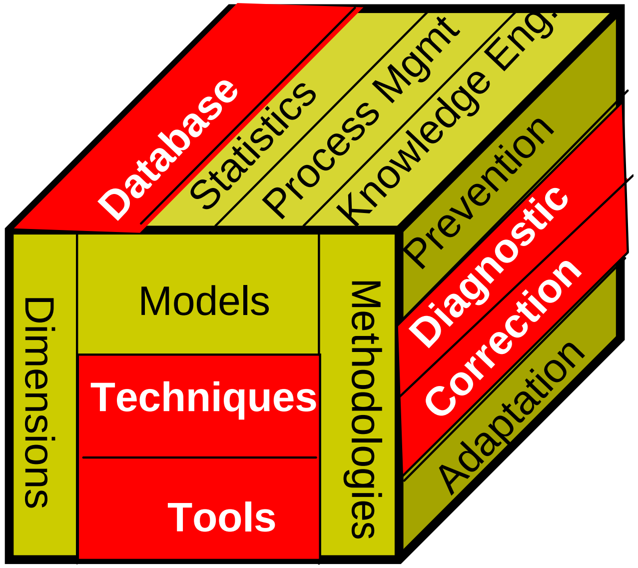
Data Quality Research:
A World of Possibilities

- 4 Disciplines
  - ✓ Statistics
  - ✓ Database
  - ✓ Knowledge Engineering
  - ✓ IT Process and Workflow Management
- 4 Types of approach
  - ✓ Prevention
  - ✓ Diagnostic
  - ✓ Correction
  - ✓ Adaptation
- 5 Levels
  - ✓ Dimensions
  - ✓ Models
  - ✓ Techniques
  - ✓ Tools
  - ✓ Methodologies



From the DB perspective

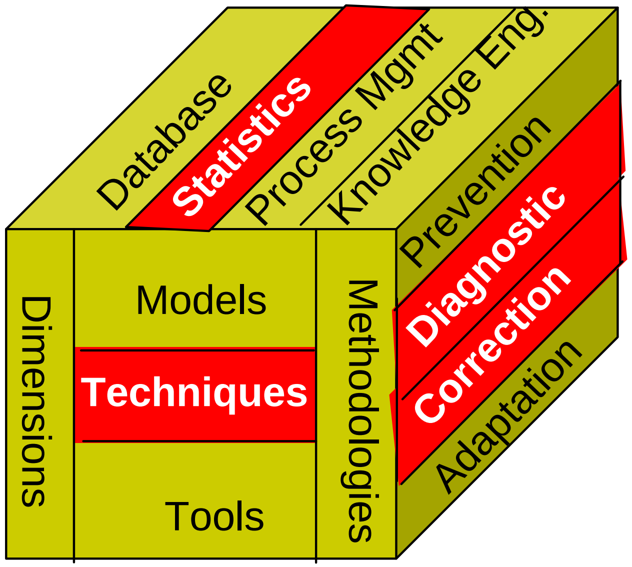
- **Data Quality Management**
  - ✓ Database profiling, data auditing
  - ✓ Integration of data
    - Source selection
    - Data cleaning, ETL
    - Schema and data mapping
    - Record linkage, deduplication
    - Conflict resolution, data fusion
  - ✓ Constraint and integrity checking
  - ✓ Data refreshment and synchronization policies
  - ✓ Metadata management



From the KDD perspective

- **Data Quality Mining is beyond data preparation**

  - ✓ Exploratory Data Analysis
  - ✓ Multivariate Statistics
  - ✓ Anomaly detection
  - ✓ Classification
    - Rule-based
    - Model-based
  - ✓ Clustering
    - Distance-based
    - Density-based
  - ✓ Visualization
  - ✓ Quantitative Data Cleaning
    - Distribution transformation
    - Treatment of missing values, inconsistencies, and outliers



What is Data Quality Mining?

“DQM can be defined as the deliberate application of data mining techniques for the purpose of data quality measurement and improvement. The goal of DQM is to detect, quantify, explain, and correct data quality deficiencies in very large databases.” *[Hipp, Güntzer, Grimmer, 2001]*

In addition,

Data Quality Mining (DQM) intends to be an iterative framework for creating, adapting, and applying data mining techniques for the discovery, explanation and quantitative cleaning of data glitches and their complex patterns in large and patchy datasets.

Motivation

Data quality problems are:

- Omnipresent in every application domain
- Interwoven and complex in any DB, DW or IS
- Critical to every data management, KDD and decision making project because of their massive financial impact

Limitations of current tools :

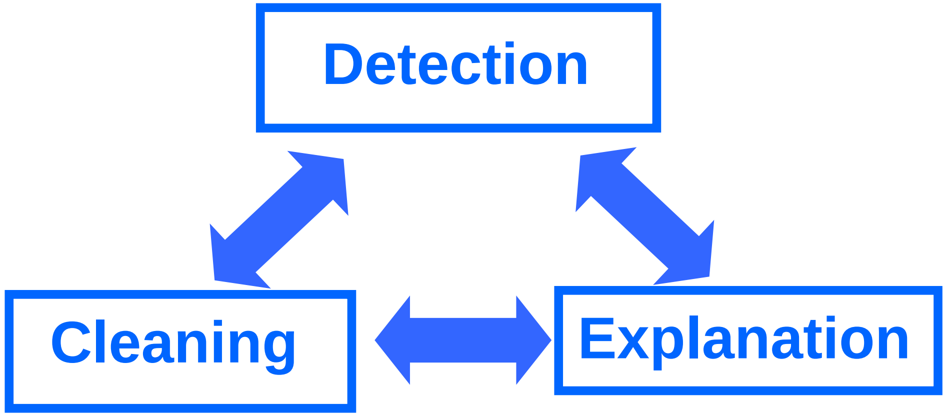
- They are *ad-hoc*, specialized, rule-based, and programmatic
- They are specific to a single-type of data quality problem
- They don’t catch interdependences between data quality problems
- Detection and cleaning tools are disconnected

Key Challenges

- Complexity and dimensionality
  - The exact notion of data quality is multidimensional and different from one application domain to another
  - Concomitant data quality problems increase the detection complexity
- Ambiguity
  - The boundary between quality and non-quality data is not precise
  - The boundary between a legitimate anomaly and a data quality problem is hard to define
- Change
  - Data and so data quality keep evolving
- Missing Metadata

Tutorial Overview

DQM: Discovering Complex Patterns of Data Glitches



Outliers

Missing Values

Inconsistencies

Duplicate Values

Complex Patterns

UV statistics
- Distributional techniques
- Skewness, Kurtosis
- Goodness of fit tests: normality, Chi-square tests, analysis of residulas, Kullback-Lieber divergence
- Control Charts: X-Bar, CUSUM, R

MV statistics
- Robust estimators

Model-based methods
- linear, logictic regression
- Probabilistic methods

Clustering
- Distance-based techniques
- Density-based techniques
- Subspace-based techniques

Classification
- Rule-based techniques
- SVM, Neural Networks, Bayesian Networks
- Information theoretic measures
- Kernel-based methods

Rule & Pattern Discovery

Visualization
- Graphics
- Q-Q plot
- Confusion Matrix
- Production Rules

Part I. Introduction to Data Quality Research

1. Illustrative Examples
2. Definitions, concepts and motivation
3. Current solutions and their limits

Current Solutions in Practice

- Diagnostic Approaches
  - Database profiling
  - Exploratory data analysis (EDA)

- Corrective Approaches
  - Extract-Load-Transform (ETL)
  - Record linkage (RL)
  - Quantitative Cleaning

Database Profiling

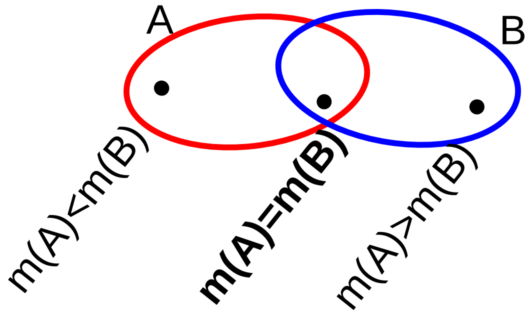
Include descriptive information

- Schema, table, domain, data sources definitions
- Business objects, rules and constraints
- Synonyms and available metadata

Systematically collect summaries of the dataset

- Number of tables, records, attributes
- Number of unique, null, distinct values for each attribute
- Skewness of data distributions
- Field Similarity (Bellman [Dasu et al., 2002])
  - By exact match
  - By substring similarity
    - Resemblance of Q-gram signatures
    - Resemblance of Q-gram min-hash distributions
- Finding Keys and FDs

Resemblance
of 2 sets A and B
$\rho(A,B) = |A \cap B| / |A \cup B|$



$\Pr[m(A) = m(B)] = \rho(A,B)$

Mainly applied to relational data

Exploratory Data Analysis (EDA)

EDA

- Use of simple statistical techniques for exploring and understanding the data (John Tukey)
- Usually for variable and model selection and for testing distributional assumptions

EDA for Data Quality

- Detect data glitches
  - Outliers and extremes
  - Missing values
  - High frequency values and duplicates
- Data transformation for model fitting
- Treatment of glitches
  - Selecting variables and records
  - Replacing using statistical models

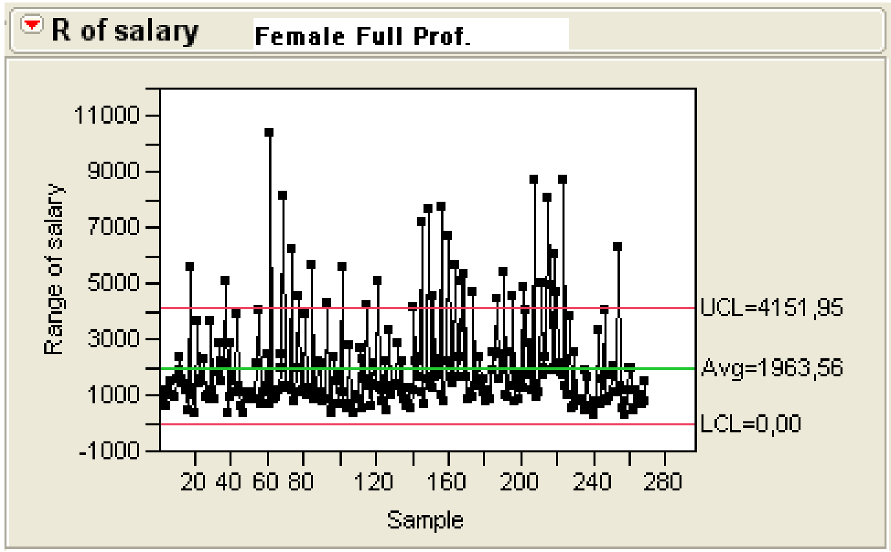
EDA – Outlier Detection

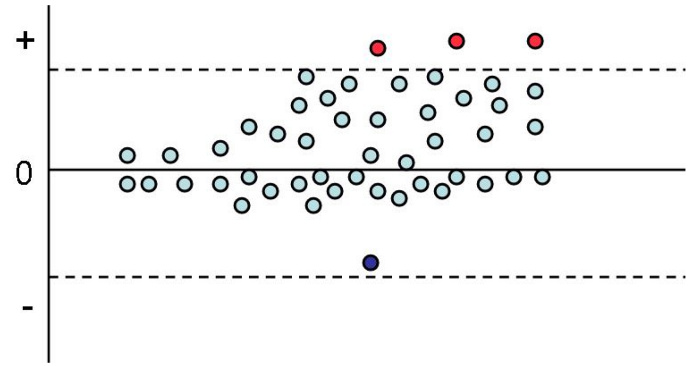
- Control chart/error bounds methods
  - e.g., expected value; confidence interval or error bounds; 3-Sigma, Hampel bounds, IQR
- Model-based outlier detection methods
  - e.g., regression model: outlyingness measured through residuals that capture deviation from the model
- Multivariate statistics for outlier detection
  - e.g., density-based and geometric or distance-based outlier detection

EDA - Control chart/error bounds

- Typical value (green) – arithmetic mean, median
- Error bounds (red) – standard deviation, IQR
- Underlying assumptions of normality and symmetry
- Simple, but potential for misleading conclusions
- Non trivial to extend to higher dimensional space

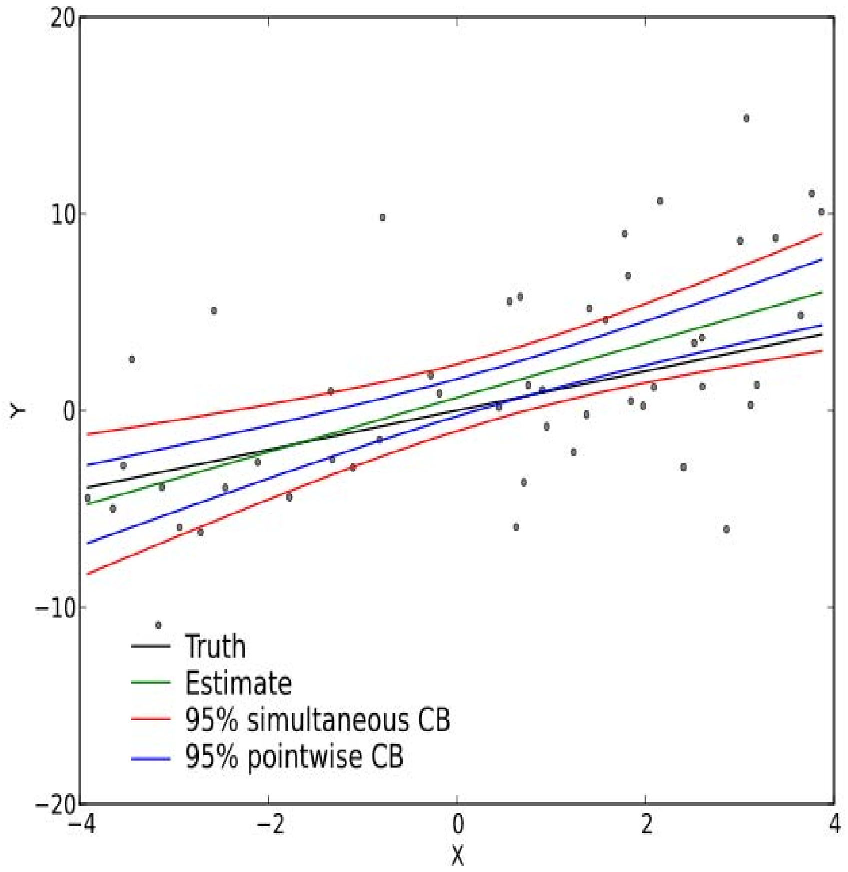
R chart





EDA - Model-based outlier detection

- Model captures relationships between variables
- Confidence bounds/bands capture variability
- Points that lie outside bounds
- The choice and correctness of the model are critical
- Expertise required for parameterization



http://en.wikipedia.org/wiki/File:Regression_confidence_band.svg

Finding Multivariate Outliers

INPUT: An d x d dataset

OUTPUT: Candidate Outliers

1. Calculate the mean μ and the variance–covariance matrix Σ
2. Let C be a column vector having length d, the square of the Mahalanobis distance to the mean μ is given by:

$$MD^2 = (x-\mu)'\Sigma^{-1}(x-\mu) = (x-\mu)' \begin{bmatrix} \sigma_{11} & \sigma_{21} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_{dd} \end{bmatrix}^{-1} (x-\mu)$$

3. Find points O in C whose value is greater than $inv\left(\sqrt{\chi_d^2(.975)}\right)$
4. Return O.

- Mean and standard deviation are extremely sensitive to outliers (Breakdown point=0%)

Robust estimators

Minimum Volume Ellipsoid (MVE) [Rousseeuw & Van Zomeren, 1990]

Let the column vector C with the length d ($d > 2$) be the estimate of location and let the d-by-d matrix $\mathbf{M}$ be the corresponding measure of scatter. The distance of the point $x_i = (x_{i1},\ldots,x_{id})$ from C is given by:

$$D_i = \sqrt{(x_i - C)' \mathrm{M}^{-1} (x_i - C)}$$

If $D_i > \sqrt{\chi^2_{.975,d}}$ then x_i *is declared an outlier.*

C is center of the minimum volume ellipsoid covering (at least) h points of the data set.

Minimum Covariance Determinant (MCD) [Rousseeuw & Driessen, 1999]

Given n data points, the MCD is the mean and covariance matrix based on the sample of size h ($h < n$) that minimizes the determinant of the covariance matrix.

- Masking the structure of the group of MV outliers (clustered vs scattered)

EDA - Distance-based outliers

Nearest Neighbour-based Approaches

- A point *O* in a dataset is an *DB*(*p*,*d*)-outlier if at least fraction *p* of the points in the data set lies greater than distance *d* from the point *O*. [Knorr, Ng, 1998]
- Outliers are the top n points whose distance to the k-th nearest neighbor is greatest. [Ramaswamy et al., 2000]

- Methods fails
  - When normal points do not have sufficient number of neighbours
  - In high dimensional spaces due to data sparseness
  - When datasets have modes with varying density
- Computationally expensive

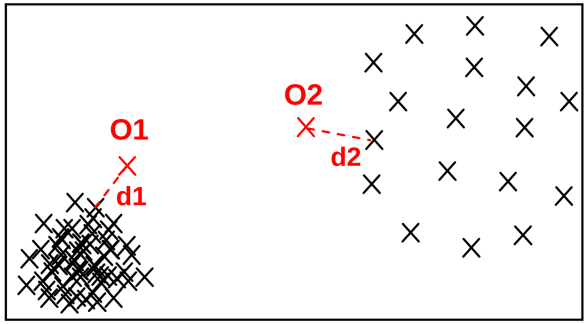
EDA - Density-based outliers

Method

Compute local densities of particular regions and declare data points in low density regions as potential anomalies

Approaches

- Local Outlier Factor (LOF) [Breunig et al., 2000]
- Connectivity Outlier Factor (COF) [Tang et al., 2002]
- Multi-Granularity Deviation Factor [Papadimitriou et al., 2003]



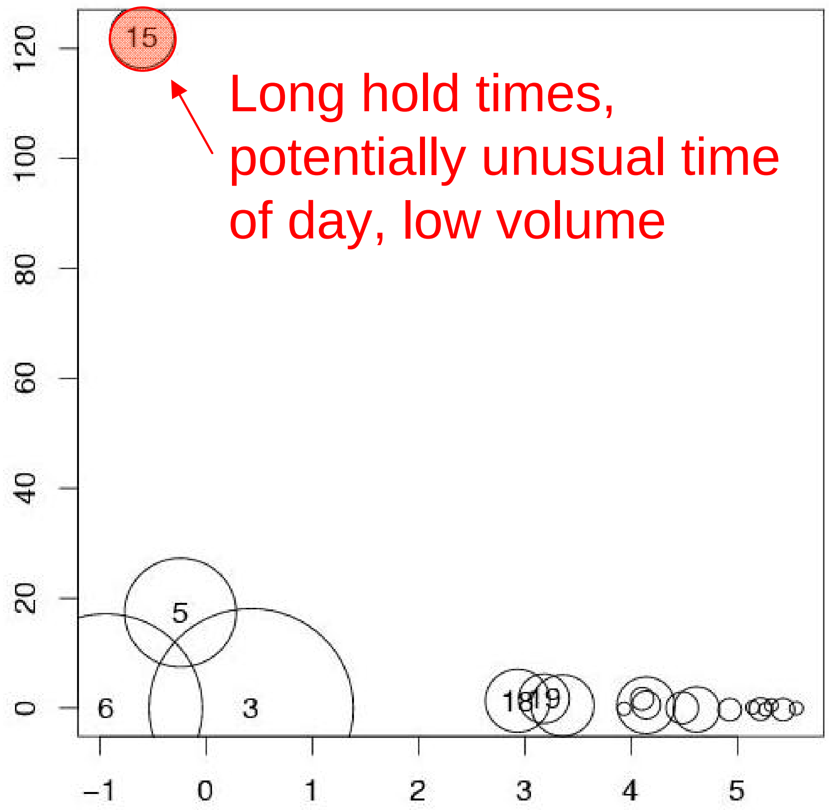
NN: O2 is outlier but O1 is not
LOF: O1 is outlier but O2 is not

In high dimensional spaces, LOF values will tend to cluster because density is defined in terms of distance.

Clustering for MV Outlier Detection

- AT&T Special service users ~ 1.67M multiple sessions
- Simple k-means clustering based on 7 variables in 2-D projection plot

- In general, computationally expensive and expert parameterization required
- Very sensitive to initial seeds and distance metrics
- Methods fails:
  - When normal points don't create any clusters
  - In high dimensional spaces, data is sparse and distances become similar



Current Practical Solutions

- Diagnostic Approaches
  - Database profiling
  - Exploratory data analysis (EDA)

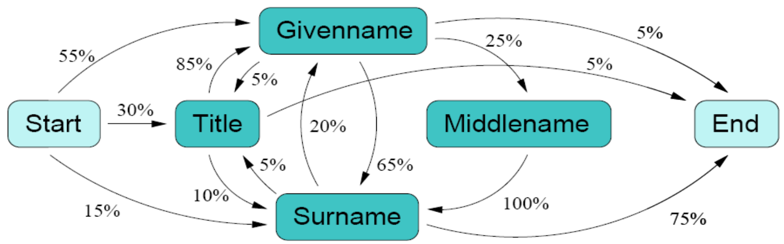
- Corrective Approaches
  - Extract-Load-Transform (ETL)
  - Record linkage (RL)
  - Quantitative Cleaning

Extract-Transform-Load and Cleaning

Goals

- Format conversion
- Standardization of values with loose or predictable structure
  - e.g., addresses, names, bibliographic entries
- Abbreviation enforcing
- Data consolidation based on dictionaries and constraints

Approaches



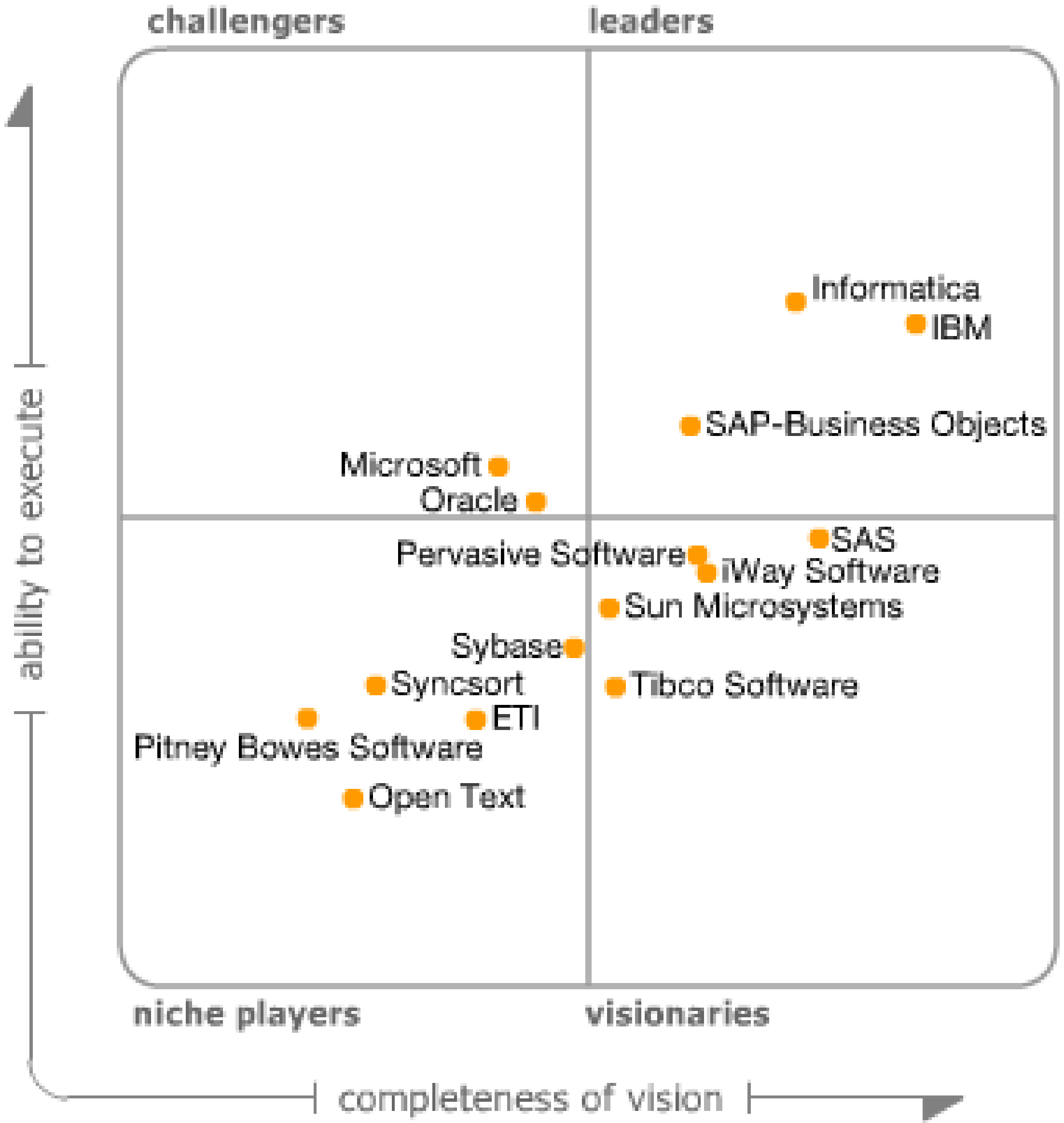
- Machine learning and HMM
  for field and record segmentation [Christen et al., 2002]
- Constraint-based method [Fan et al., 2008]

Performance and scalability issues of most ETL tools

Academic and Open Source ETL Tools

Name	Main characteristics	Data Transformation	Data Cleaning	Duplicate Detection	Data Enrichment	Data Profiling	Data Analysis
Potter's wheel [Raman *et al.* 2001]	Detection and correction of errors with data transformations: *add, drop, merge, split, divide, select, fold, format* Interactivity, inference of the data structure	X					X
Ajax [Galhardas *et al.* 2001]	Declarative language based on logical transformation operators: *mapping, view, matching, clustering, merging* 3 algorithms for record matching	X	X	X	X		
Arktos [Vassiliadis 2000]	Graphical and declarative (SQL-like and XML-like) facilities for the definition of data transformation and cleaning tasks, optimization, measures of quality factors	X	X				
Intelliclean [Low *et al.* 2001]	Detection and correction of anomalies using a set of rules (*duplicate identification, merge, purge, stemming, soundex, stemming, abbreviation*) - Not scalable			X			
Bellman [Dasu et al., 2002]	Data quality browser collecting database profiling summaries, implementing similarity search, set resemblance, Q-gram sketches for approximate string matching			X		X	X
Febrl [Christen, 2008]	Open source in Python, initially dedicated to data standardization and probabilistic record linkage in the biomedical domain, including Q-gram, sorted NN, TF-IDF methods for record linkage and HMM-based standardization	X	X	X		X	X
Pentaho-Kettle http://kettle.pentaho.org	Open source in Java for designing graphically ETL transformations and jobs such as reading, manipulating, and writing data to and from various data sources. Linked to Weka. Easily extensible via Java Plug-ins	X	X	(X)	(X)	(X)	(X)
Talend Open Studio http://www.talend.com	Open source based on Eclipse RCP including GUI and components for business process modeling, and technical implementations of ETL and data flows mappings. Script are generated in Perl and Java code.	X	X	(X)	(X)	(X)	(X)

Commercial ETL Tools



Source: Magic Quadrant for **Data Integration Tools**, Sept. 2008, Gartner RAS Core Research Note G00160825.

Criteria

Ability to execute

- Product/Service
- Overall Viability
- Sales Execution/Pricing
- Market Responsiveness
- Track Record
- Marketing Execution
- Customer Experience
- Operations

Completeness of vision

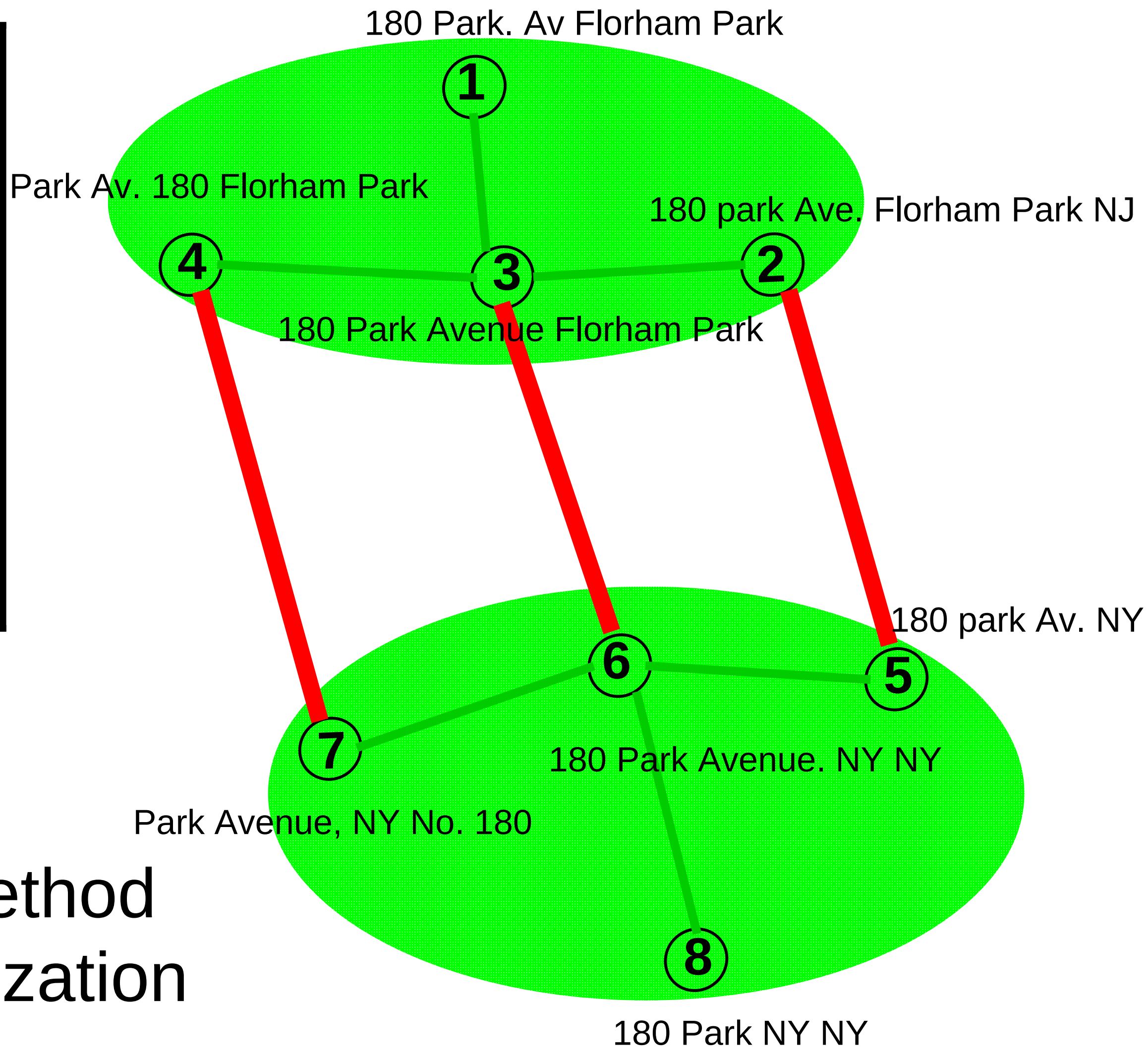
- Market Understanding
- Marketing Strategy
- Sales Strategy
- Offering (Product) Strategy
- Business Model
- Vertical/Industry Strategy
- Innovation
- Geographic Strategy

Record Linkage (RL) [Elmagarmid et al., 2007]

1. Pre-processing: transformation and standardization
2. Select a blocking method to reduce the search space partitioning the dataset into mutually exclusive blocks to compare
   - Hashing, sorted keys, sorted nearest neighbors
   - (Multiple) Windowing
   - Clustering
3. Select and compute a comparison function measuring the similarity distance between pairs of records
   - Token-based : N-grams comparison, Jaccard, TF-IDF, cosine similarity
   - Edit-based: Jaro distance, Edit distance, Levenshtein, Soundex
   - Domain-dependent: data types, ad-hoc rules, relationship-aware similarity measures
4. Select a decision model to classify pairs of records as matching, non-matching or potentially matching
5. Evaluation of the method (recall, precision, efficiency)

Chaining or Spurious Linkage

ID	*Name*	*Address*
1	AT&T	180 Park. Av Florham Park
2	ATT	180 park Ave. Florham Park NJ
3	AT&T Labs	180 Park Avenue Florham Park
4	ATT	Park Av. 180 Florham Park
5	TAT	180 park Av. NY
6	ATT	180 Park Avenue. NY NY
7	ATT	Park Avenue, NY No. 180
8	ATT	180 Park NY NY



- Expertise required for method selection and parameterization

RL - Models and Prototypes

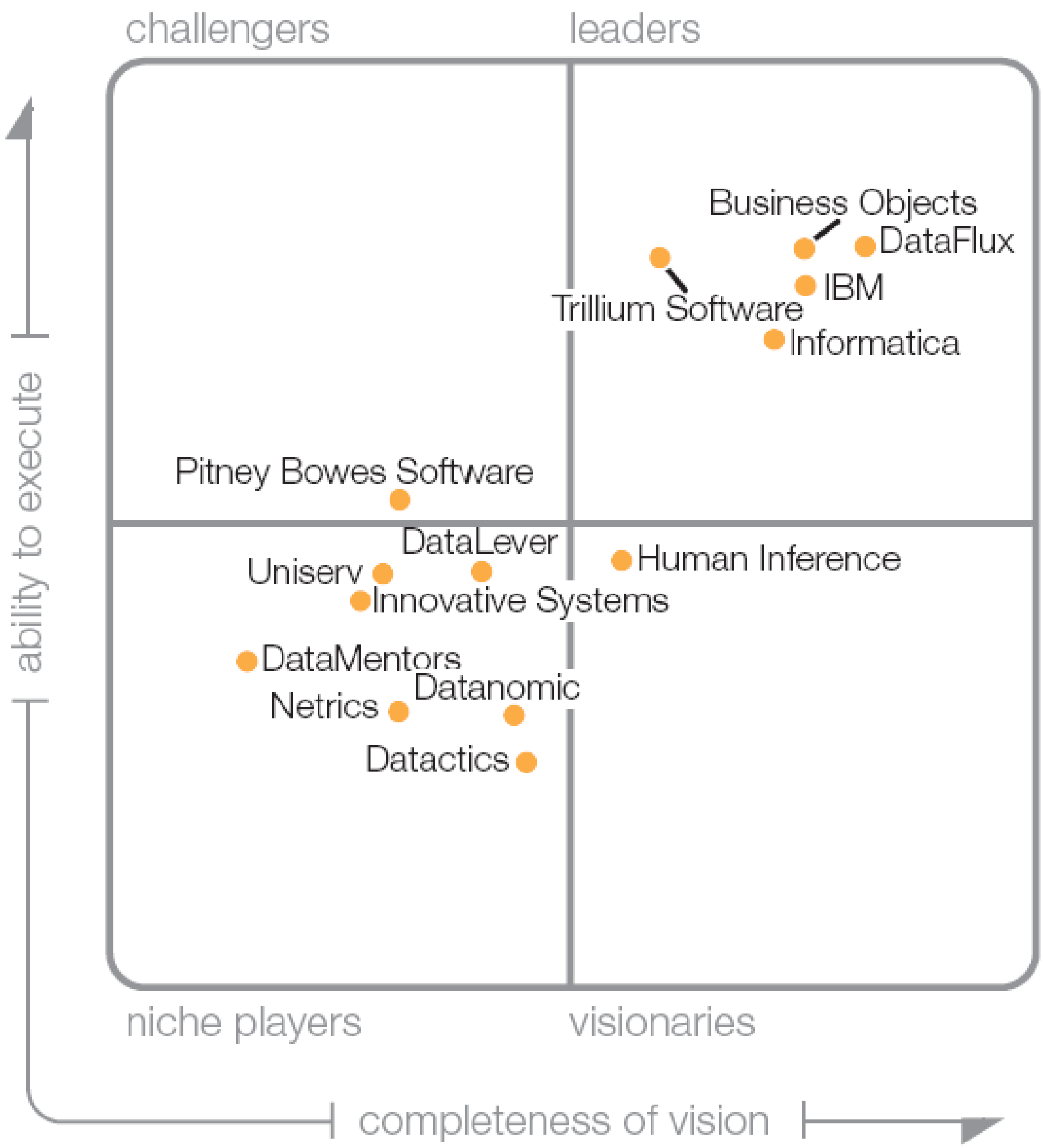
Decision Model (*Prototype*)	Authors	Type
Error-based Model	[Fellegi & Sunter 1969]	Probabilistic
EM-based Method	[Dempster *et al.* 1977]	
Induction Model Clustering Model (*Tailor*)	[Bilenko et Mooney 2003] [Elfeky *et al.* 2002]	
1-1 matching	[Winkler 2004]	
Bridging File	[Winkler 2004]	
Sorted Nearest Neighbors and variants		Empirical
XML object Matching	[Weiss, Naumann 2004]	
Hierarchical Structure (*Delphi*)	[Ananthakrishna *et al.* 2002]	
Matching Prediction based on clues	[Buechi *et al.* 2003]	Knowledge-based
Instance-based functional dependencies	[Lim *et* al. 1993]	
Transformation Fuctions (*Active Atlas*)	[Tejada *et al.* 2001]	
Variant of NN based on rules for identifying and merging duplicates (*Intelliclean*)	[Low *et al.* 2001]	

Interactive Data Cleaning

- **D-Dupe** [Kang et al., 2008] http://www.cs.umd.edu/projects/linqs/ddupe
  Duplicate search and visualization of cluster-wise relational context for entity resolution
- **Febrl** [Christen, 2008]: https://sourceforge.net/projects/febrl/
  Rule-based and HMM-based standardization and classification-based record linkage techniques
- **SEMANDAQ** [Fan et al., 2008]: CFD-based cleaning and exploration
- **HumMer** [Bilke et al., 2005]: Data fusion with various conflict resolution strategies
- **XClean** [Weis, Manolescu, 2007]: Declarative XML cleaning

Commercial Data Quality Tools

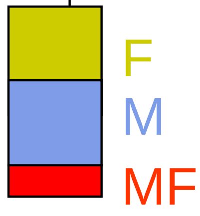
DQ Tools include:
- Profiling
- Improvement:
  - Standardization
  - Cleansing
  - Matching
  - Enrichment
- Monitoring



Source: Quadrant of the Magic Quadrant for **Data Quality Tools**, 2008.
Gartner RAS Core Research Note G00157464

Quantitative Data Cleaning

Methods



- **Inclusion** *(applicable for less than 15%)*
  - Anomalies are treated as a specific category
- ***Deletion***
  - List-wise deletion omits the complete record *(for less than 2%)*
  - Pair-wise deletion excludes only the anomaly value from a calculation
- ***Substitution*** *(applicable for less than 15%)*
  - Single imputation based on mean, mode or median replacement
  - Linear regression imputation
  - Multiple imputation (MI)
  - Full Information Maximum Likelihood (FIML)

Limits of Current Methods

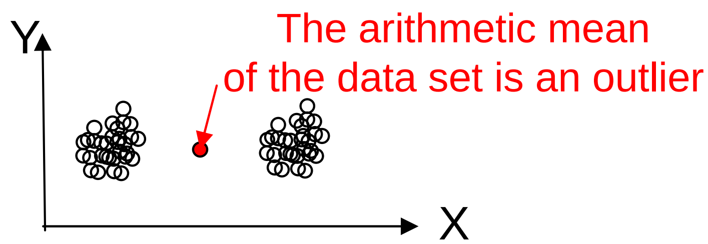
Non Realistic Assumptions

- Data quality problems may not be rare events
  $\rightarrow$ *rare class mining won't give the "complete picture"*
- Data quality problems don't occur at random
  $\rightarrow$ *MCAR/MAR assumptions are not applicable*
- Data quality problems are not uniformly distributed
  $\rightarrow$ model-based assumptions is hazardous
- Different types of data quality problems may co-occur and be (partially) correlated
  $\rightarrow$ mutual masking-effect of concomitant DQ problems
  $\rightarrow$ potential multicollinearity problem
- Their (co-)occurrence should be explainable
  $\rightarrow$ *explanatory variables/processes may be external and out of the scope of the analysis*
- They should be corrected with "predictable side-effects"
  $\rightarrow$ *Biases of imputation and regression methods*

In Particular

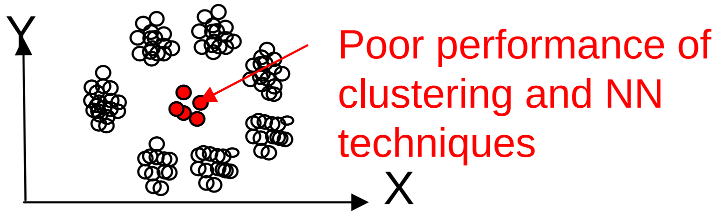
EDA - Statistics

- Methods tied to known distributions
- Parametric assumptions often do not hold for real datasets
- Bad points can completely skew the mean and standard deviation: Robustification of the estimators is required
- Statistical methods suffer the uni-modality and locality assumptions: they consider the data set globally



EDA - Clustering

- Magic numbers,
- complex parameterization and settings
- Locality and normality assumptions



Quantitative Cleaning

- Only applicable when the dataset has less than 15% of glitches
- Non negligible biases

Challenges of DQM

Detection of concomitant DQ problems

- Joint detection of fuzzy duplicates, disguised missing values, multivariate outliers, and deviants
- Detection of complex patterns of multivariate glitches
- Interactivity, rerunnability and recoverability of DQM processes

DQM in high dimensional data sets

Benchmarking

Outline

Part I. Introduction to Data Quality Research

Part II. Data Quality Mining

Part III. Case Study

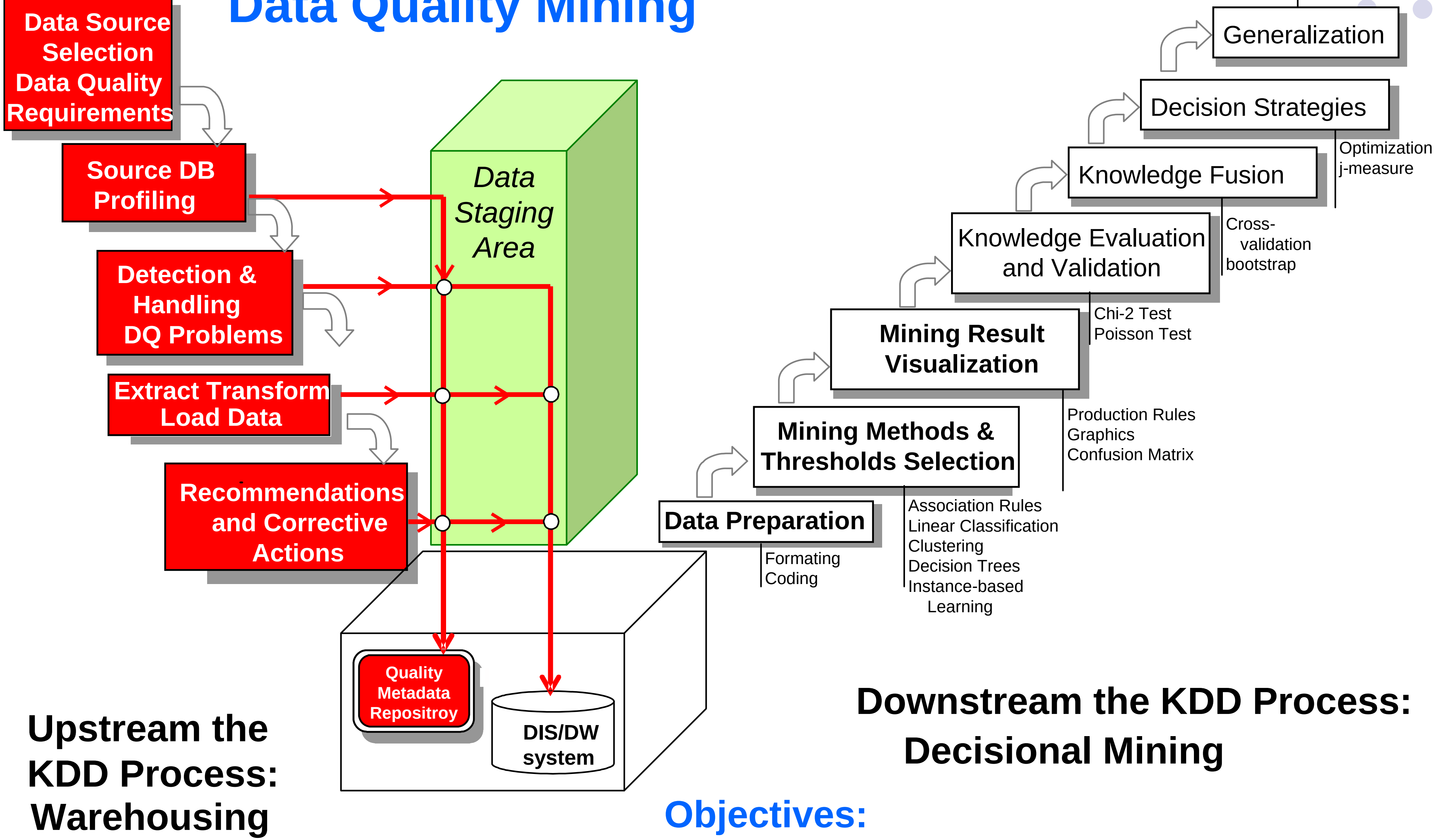
Part II. Data Quality Mining

1. DQM Framework
2. Advances on Single-Type DQ Problems
3. Discovery of Complex Glitches

Data Quality Mining: The Big Picture

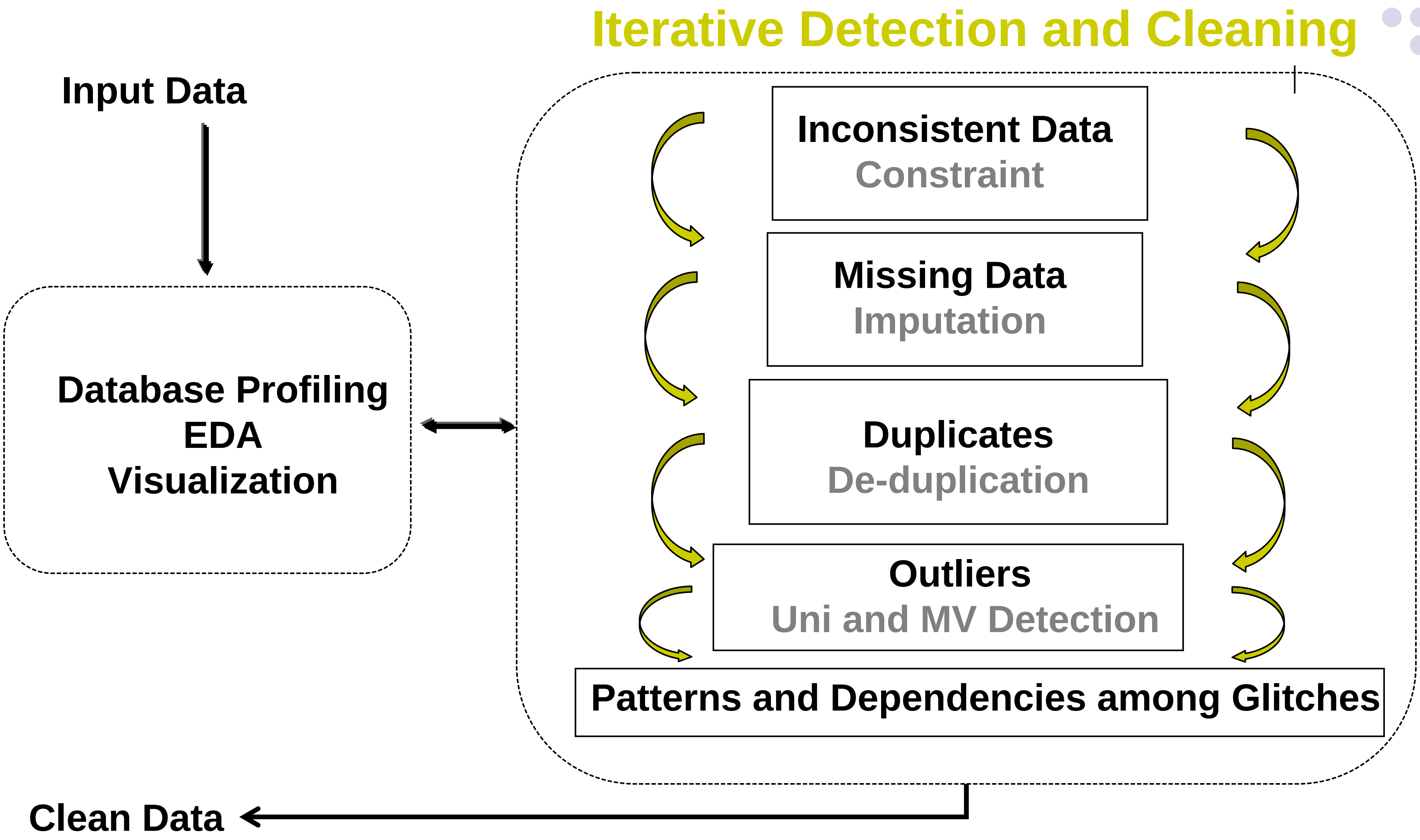
Data Quality Management

Data Quality Mining



Objectives:
Data Quality Measurement and Improvement

DQM Framework: The Process

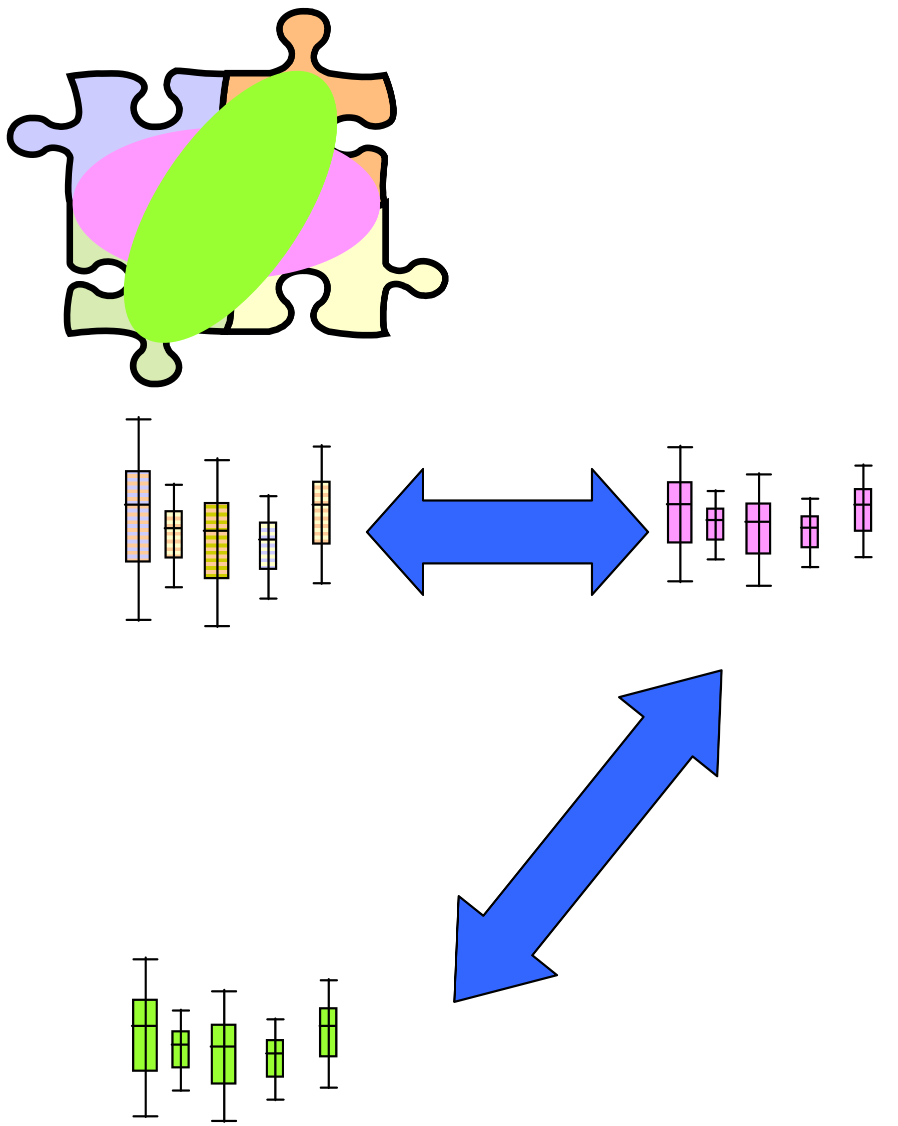


Best DQM Strategy?

Best DQM Strategy

From your patchy raw data,
Define:

- An ideal dataset
  - Baseline; historical aggregate; pre-defined
- The representation
  - Signature; a statistical summary
- The comparison measure
- The function to optimize
  - Min, threshold



Plan the cleaning process to get the optimal dataset

Part II. Data Quality Mining

1. DQM Framework
2. Advances on Single-Type DQ Problems
   - Inconsistent Data
   - Deduplication
   - Missing Data
   - Outlier Mining
3. Discovery of Complex Glitches

Inconsistent Data

- **Probabilistic Approximate Constraints** [Korn et al., 2003]

  Given a legal ordered domain on an attribute,

  - A **domain PAC** specifies that all attribute values x fall within ε of D with at least probability δ, as $\Pr(x \in [D \pm \varepsilon]) \geq \delta$
  - A **functional dependency PAC** X → Y specifies that, if $|T_i.A_\ell - T_j.A_\ell| \leq \Delta_\ell \;\; \forall A_\ell \in X$ then $\Pr(|T_i.B_\ell - T_j.B_\ell| \leq \varepsilon_\ell) \geq \delta \;\; \forall B_\ell \in Y$

- **Pseudo-constraints** [Ceri et al., 2007]

  Pair <*P1,P2*> where *P1* and *P2* are predicates on the same domain *D* such that if *P1* holds, then usually *P2* also and therefore there are *few* rule violations. More formally, based on the probability contingency table,

$$\frac{p_{11}}{p_{11}+p_{21}} - \rho - (1-\rho).(p_{11}+p_{12}) > 0$$

	$P1$	$\overline{P1}$	
$P2$	p_{11}	p_{12}	$p_{1.}$
$\overline{P2}$	p_{21}	p_{22}	$p_{2.}$
	$p_{.1}$	$p_{.2}$	1

- **Pattern Tableaux for Conditional Functional Dependencies**

  [Bohannon et al. 2007, Bravo et al. 2007, Golab et al. 2008, Fan et al. 2009]

  A CFD is defined to be a pair $\varphi = R(\underbrace{A \rightarrow B}_{\text{Embedded FD}}, T_p)$, where $T_p =$

A	B
_	b_1
_	b_2

Duplicate Data: Learning Approaches for RL

Training examples

Customer 1	**D**	
Customer 2		
Customer 1	N	
Customer 3		
Customer 4	**D**	
Customer 5		

Similarity distance functions

f_1	$f_2 \ldots f_n$			
1.0	0.4	...	0.2	**1**
0.0	0.1	...	0.3	0
0.3	0.4	...	0.4	**1**

Unlabeled list

Customer 6
Customer 7
Customer 8
Customer 9
Customer 10
Customer 11

0.0	0.1	...	0.3	?
1.0	0.4	...	0.2	?
0.6	0.2	...	0.5	?
0.7	0.1	...	0.6	?
0.3	0.4	...	0.4	?
0.0	0.1	...	0.1	?

Classifier

Learnt Rule: All-Ngrams*0.4
+ CustomerAddressNgrams*0.2
– 0.3EnrollYearDifference
+ 1.0*CustomerNameEditDist
+ 0.2*NumberOfAccountsMatch – 3 > 0

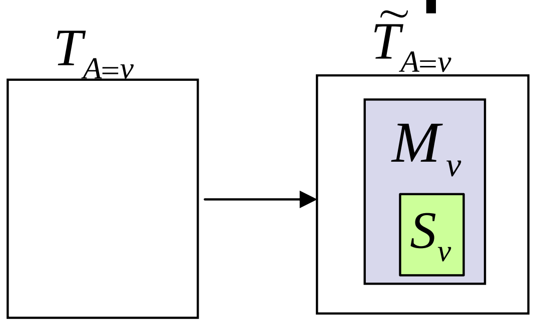
Learners:
SVMs: high accuracy with limited data [Christen, 2008]
Decision trees: interpretable, efficient to apply
Perceptrons: efficient incremental training
[Bilenko et al., 2005]

Disguised Missing Data [Hua, Pei, 2007]

Assumption: On an attribute, there often exist only a small number of inliers whose values are frequently used as disguised missing data, *e.g., popular DoB: Jan 1*

For a tuple t in the truth table T and $\tilde{t}$ in the recorded table $\tilde{T}$, the value $\tilde{t}.A$ is a disguised missing value if $t.A = \otimes$ but $\tilde{t}.A \neq \otimes$

Embedded Unbiased Sample Heuristic



Truth **Projected DB**

Goal: Find the largest set M_v embedding S_v and maximizing a correlation-based sample quality score.

Handling Missing Data

- **Completion Using Association Rules**
  - Based on a consensus from rules with high confidence and user interaction
  - Based on measures scoring the best rules to select the replacement value [Wu et al., 2004]
- **Imputation using NN, Clustering and SVM**
  - K-Nearest Neighbour Imputation [Batista, Monard, 2003]
  - K-means Clustering Imputation [Li et al., 2004]
  - Fuzzy K-means Clustering [Acuna, Rodriguez, 2004]
  - SVM [Feng et al. 2005]

Outlier Mining

- Multivariate techniques
  - Projection pursuit
  - Distance and depth based methods
  - Probability and kernel based methods
- Stream specific methods
- Too many outliers → Distributional shift?
  - Change detection
- Great tutorial on outliers [Kriegel et al., 2009]: http://www.dbs.informatik.uni-muenchen.de/Publikationen/Papers/tutorial_slides.pdf

Projection Based Methods

- Projection pursuit techniques are *applicable in diverse data situations* although at the expense of high computational cost.
  - No distributional assumptions, search for useful projections
- *Robust:* Filzmoser, Maronna, Werner (2008) propose a fast method based on robust PCA with differential weights to maximally separate outliers. Shyu et al. (2003) use a similar theme.
- *Time Series:* Galeano et al. (2006) extend the idea of projecting in directions of high and low kurtosis to multivariate time series.
- *Skewed Distributions:* Hubert and Van der Veeken (2007) extend the boxplot idea by defining adjusted outlyingness followed by random projections for detecting outliers in skewed data.

Outlier Mining - Robust PCA

[Shyu et al., 2003]

INPUT: An N x d dataset

OUTPUT: Candidate Outliers

1. Compute the principal components of the dataset
2. For each test point, compute its projection on these components
3. If y_i denotes the i^{th} component, then the following has a chi-square distribution

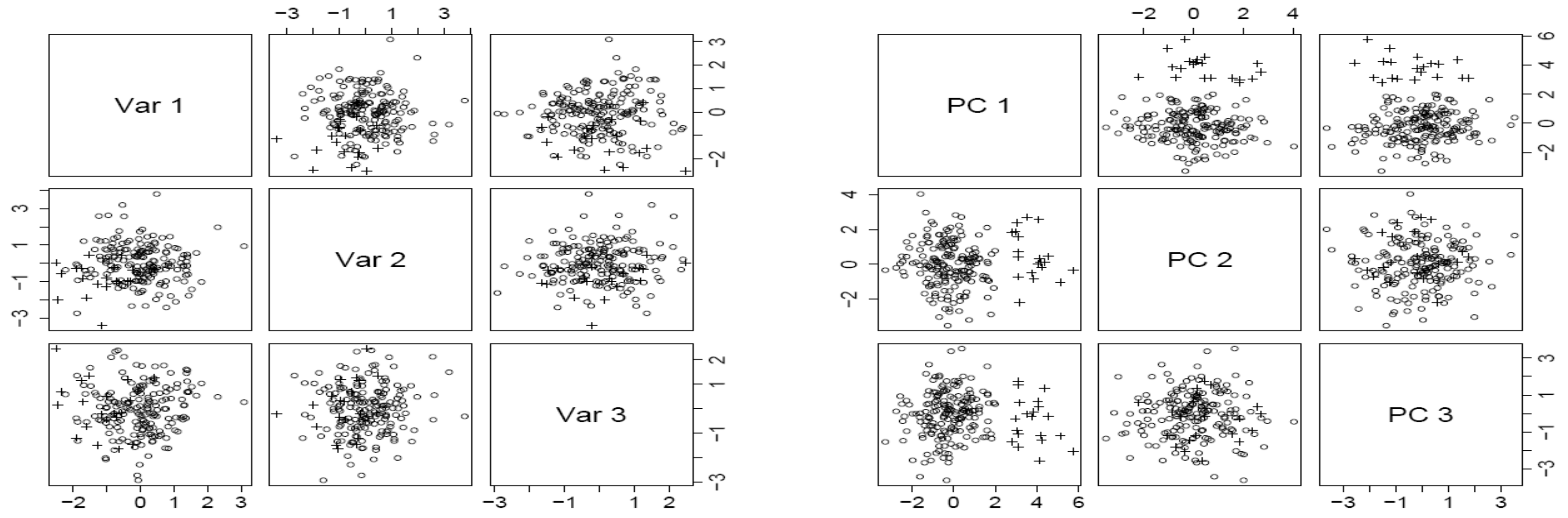
$$\sum_{i=1}^{q} \frac{y_i^2}{\lambda_i} = \frac{y_1^2}{\lambda_1} + \frac{y_2^2}{\lambda_2} + \cdots + \frac{y_q^2}{\lambda_q}, q \le p$$

3. For a given significance level α, an observation is an outlier if

$$\sum_{i=1}^{q} \frac{y_i^2}{\lambda_i} \ge \chi_q^2(\alpha)$$

Outlier Identification in High Dimensions

[Filzmoser, Maronna and Werner, 2008]

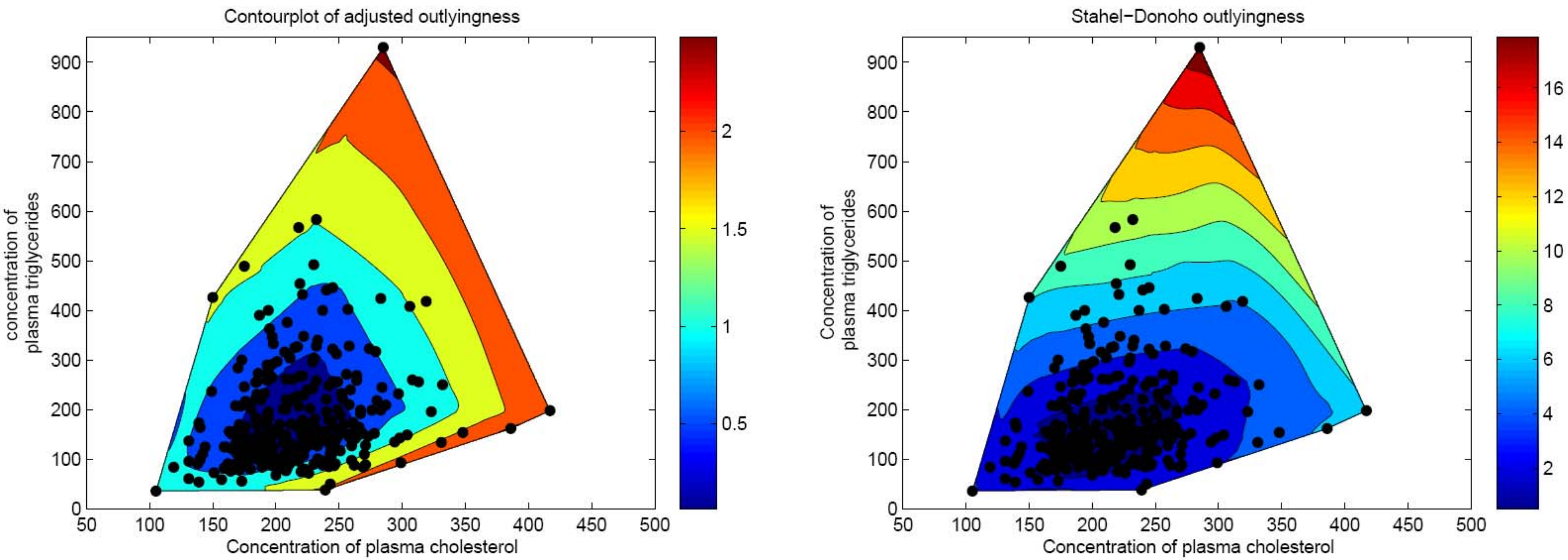


- Works in very high-D, where dimensions > samples, e.g., gene data
- Differential weights to detect location and scatter outliers; weights combined in final step
- Based on robust statistics

Outlier Detection for Skewed Data

[Hubert and Van der Veeken, 2007]

- For skewed distributions
- Key concepts
  - Adjusted outlyingness – different scaling on either side of median in boxplots.
  - MV equivalent, e.g., bagplot in 2-D
  - Random projections to identify outliers



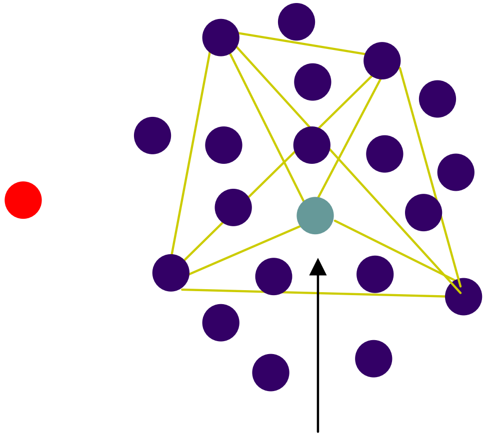
Distance and Depth Based Methods

- Distance-based methods aim to detect outliers by computing a measure of how far a particular point is from most of the data.
- *Robust methods*
  - Robust distance estimation in high-D [Maronna and Zamar, 2002] [Pena and Prieto, 2001]
- *Depth based nonparametric methods*
  - Nonparametric methods based on multivariate control charts [Liu et al, 2004]
  - Outlier detection with kernelized spatial depth function [Cheng, Dang, Peng and Bart, 2008]
- *Exotic methods*
  - Angle based detection [Kriegel, 2009]

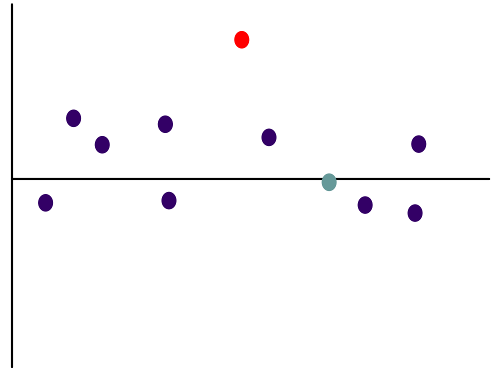
DDMA: Nonparametric Multivariate Moving Average Control Charts Based on Data Depth

[Liu, Singh and Teng, 2004]

- Extends simplicity of control charts to higher dimensions – relatively few assumptions
- Use any data depth, e.g., simplicial depth to map multidimensional data to a scalar and rank
- Apply moving average control chart techniques to data depth rank to identify outliers



$$X \in \mathrm{R}^{d} \rightarrow D \in \mathrm{R}$$



Deepest point, e.g., simplicial depth = contained in most triangles

Probability and Kernel based methods

- *Popular methods*: LOF, INFLO, LOCI
  see Tutorial of [Kriegel et al., 2009]
- *Mixture distribution:* Anomaly detection over noisy data using learned probability distributions [Eskin, 2000]
- *Entropy:* Discovering cluster-based local outliers [He, 2003]
- *Projection into higher dimensional space:* Kernel methods for pattern analysis [Shawne-Taylor, Cristiani, 2005]

Probability Based Methods

- **Probability distributions** [Eskin, 2000]

  Assumption: High probability to have the number of normal elements in a dataset *D* significantly larger than the number of outliers

  Approach:

  From the distribution for the dataset *D* given by: $D = (1-\lambda)\, M + \lambda\, A$

  with *M*: Majority distribution and λ: Anomaly distribution

  - Compute likelihood of *D* at time *t*: $L_t(D)$
  - Measure how likely each point p_t is outlier at time *t such as:*

  $$M_t = M_{t-1} \setminus \{p_t\} \text{ and } A_t = A_{t-1} \cup \{p_t\}$$

- **Entropy-based methods** [Lee et al., 2001][He 2005]

  Observations: large entropy $\rightarrow$ partition into regular subsets

  skewness of class distribution $\rightarrow$ small entropy $\rightarrow$ high redundancies

  Approach:

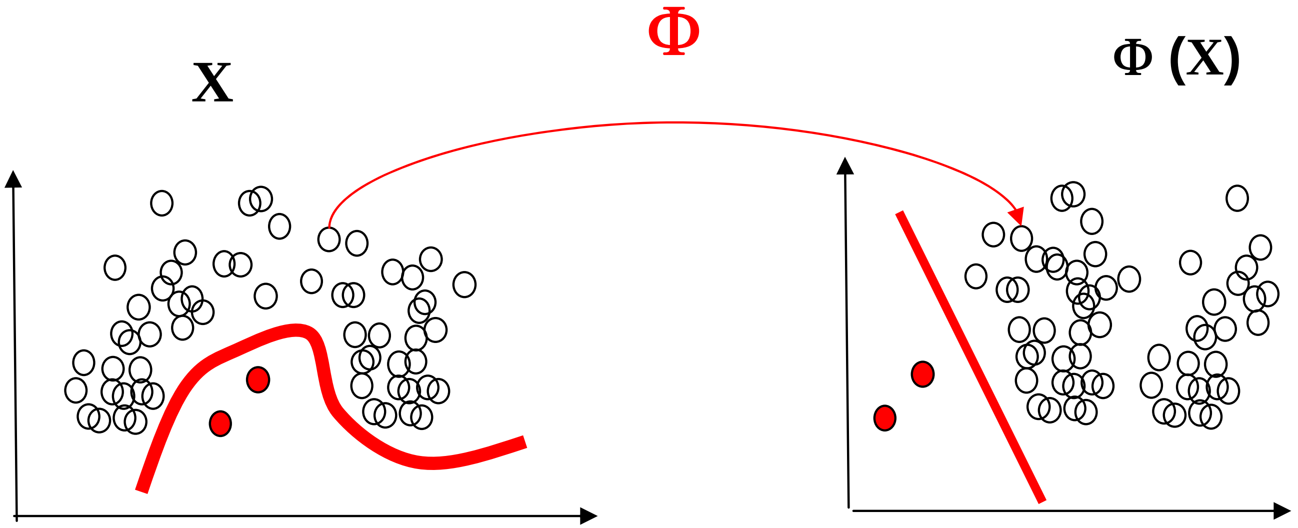
  Find a k-sized subset whose removal leads

  to the maximal decreasing of entropy

Kernel methods for pattern analysis

[Shawne-Taylor, Cristani, 2005]

Objective: Mapping data by a nonlinear mapping function Φ into a higher feature space where they are linearly separable into majority and outliers.



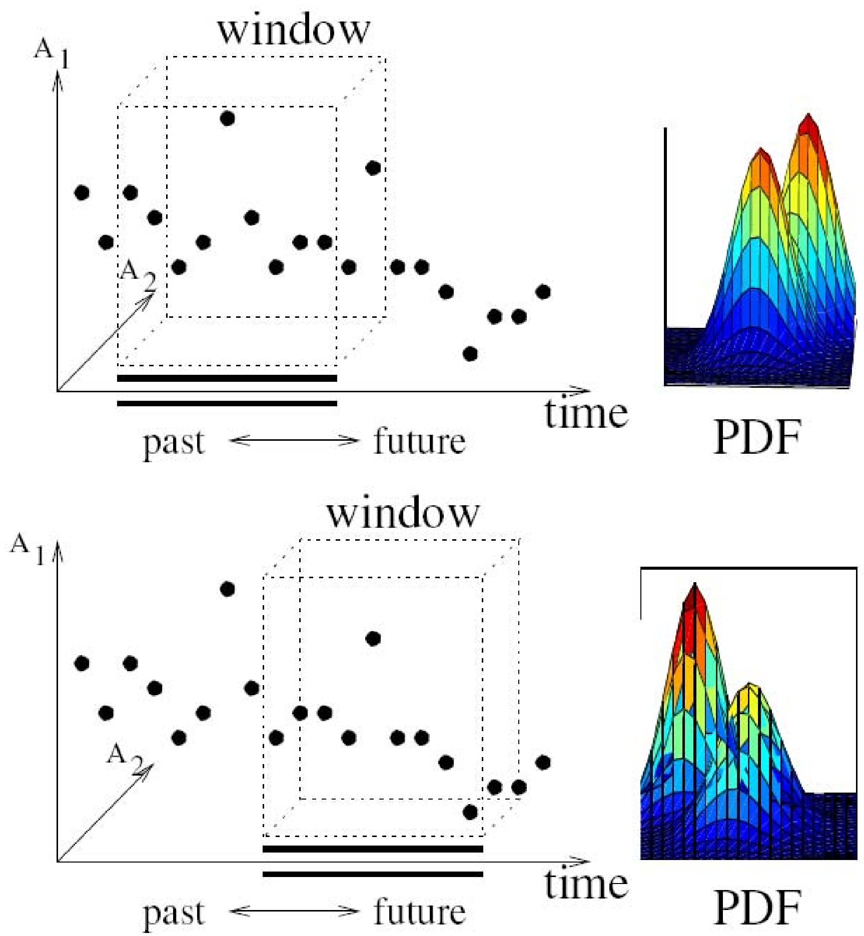
Stream Specific Methods

- *Distance based outliers:* Detecting distance based outliers in streams of data. [Anguilli and Fassetti, 2007]
- *Distributed streams:* Adaptive Outlier Detection in Distributed Streams [Su, Han, Yang, Zou, Jia, 2007]
- *A general density estimation scheme:* Online outlier detection in sensor streams [Subramaniam et al , 2006]
- *Projections and high dimensions*: Projected outliers in High-D data streams [Zhang, Gao, Wang, 2008]
- *Items of interest:* Finding frequent items in data streams [Cormode and Hadjieleftheriou, 2008]

Online Outlier Detection in Sensor Data Using Non-Parametric Models

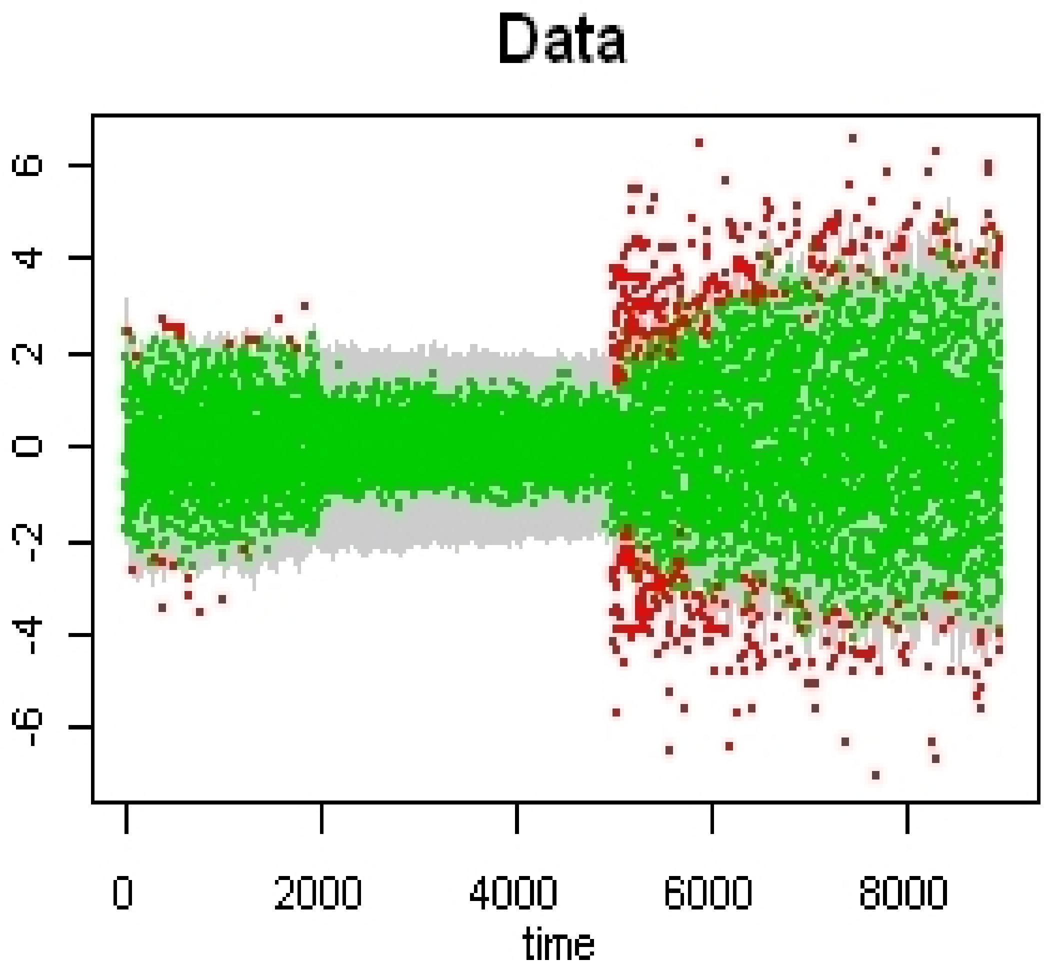
[Subramaniam et al., 2006]

- Online outlier detection in hierarchical sensor networks
- Solve the more general problem of estimating the multidimensional data distribution
  - Chain sampling
  - Epanechnikov kernel



Outliers and Change Detection

- Often, an increase or decrease in outliers is the first sign of a distributional shift
- Serious implications for data quality – recalibrate anomaly detection methods
- Change detection methods are critical



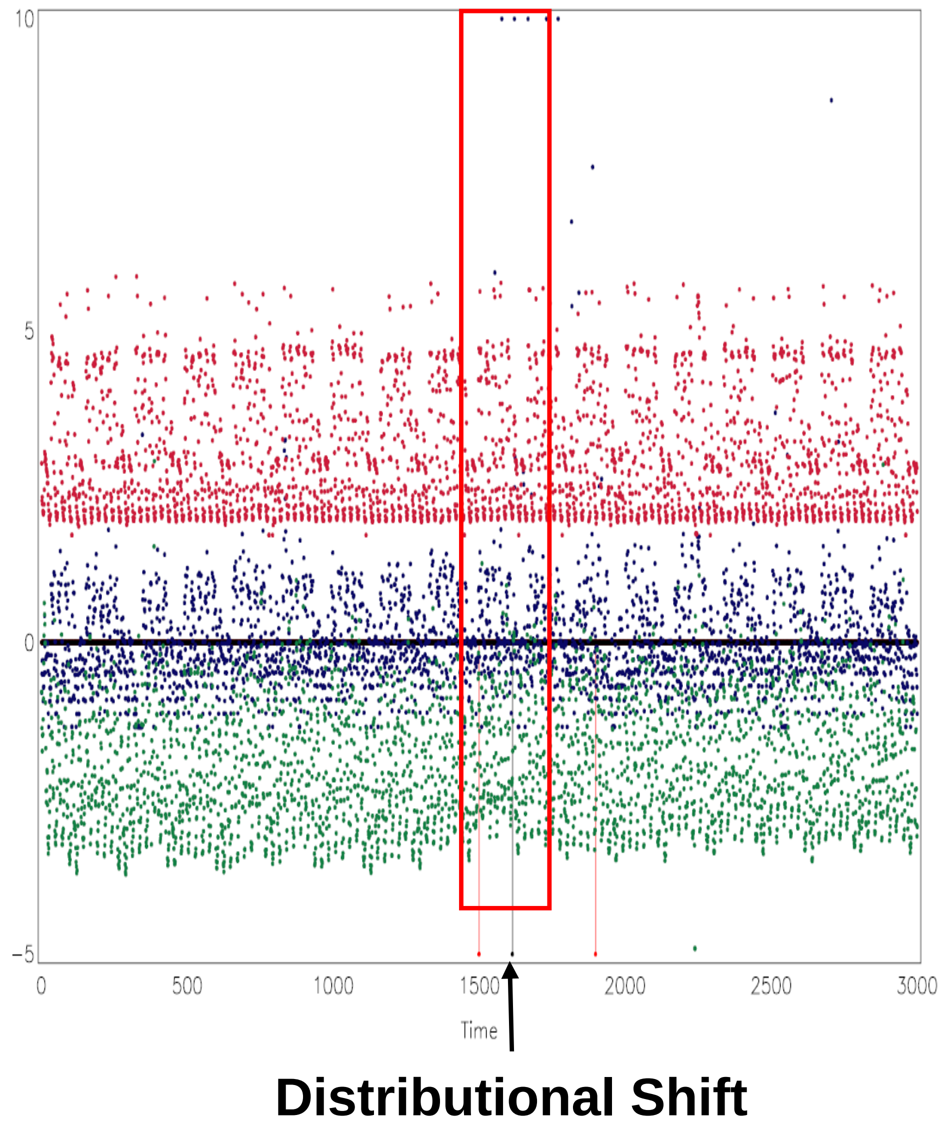
Change Detection Schemes

- *Comprehensive framework:* Detecting Changes in Data Streams. [Kifer et al., 2004]
- *Kernel based:* Statistical Change Detection in Multi-dimensional Data. [Song et al., 2007]
- *Nonparametric, fast, high-D:* Change Detection you can believe in: Finding Distributional Shifts in Data Streams. [Dasu et al., 2006, 2009]

Change (Detection) you can believe in: Finding Distributional Shifts in Data Streams

[Dasu, Krishnan, Li, Venkatasubramanian, Yi, 2009]

- Compare data distributions in two windows
  - Kdq-tree partitioning
  - Kullback-Leibler distance of histograms
    - Counts
    - Referential distance
  - Bootstrap to determine threshold
  - File descriptor data stream
    - 3 variables shown
    - Change detection led to improvement in process and cycle times



Part II. Data Quality Mining

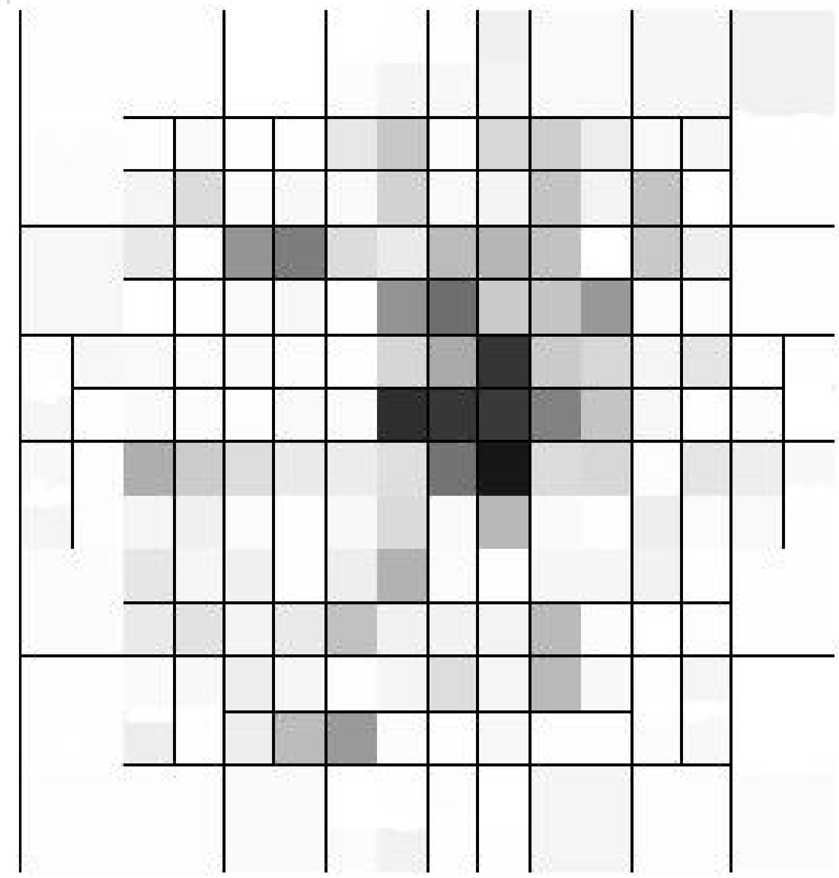
1. DQM Framework
2. Advances on Single-Type DQ Problems
3. Discovery of Complex Glitches

Changes in Distributions Caused by Missing/Duplicate Data

- Subtle cases of duplication/missing data
  - Result in changes in distributions
  - Missing → "lower" density regions
  - Duplicates → "higher" density regions
- Multinomial tests
  - Contingency tables (Chi-square test)
  - Difference in proportions (e.g., counts)
- Difference in Distributions
  - Histogram distances (Kullback Leibler)
  - Rank based (Wilcoxon)
  - Cumulative distribution based (Kolmogorov-Smirnov)

Missing Data Example

- Comparison of telecommunications data sets
- Anomalous months
  - Missing data
  - Kdq tree partition
  - Darker → greater density difference
- Automatic detection is speedy, provides an opportunity to recover and replace data before it is archived



Outline

Part I. Introduction to Data Quality Research

Part II. Data Quality Mining

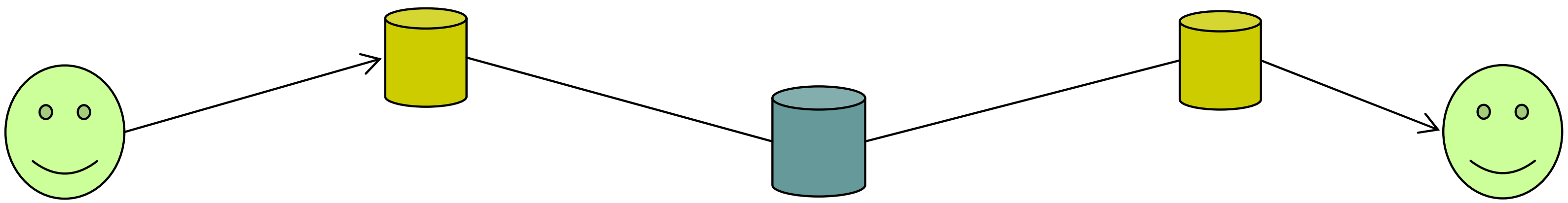
Part III. Case Study

Case Study: DQ Patterns in networking data streams

- Domain knowledge and goal
- Data set description
- DQM tasks
- Analysis
- Best DQM strategy

IP Network Data Streams

- Thousands of network elements on an IP network
- Transmit data streams as they communicate to verify availability and transmit data
- Monitor the data streams to measure performance, and to optimize the network

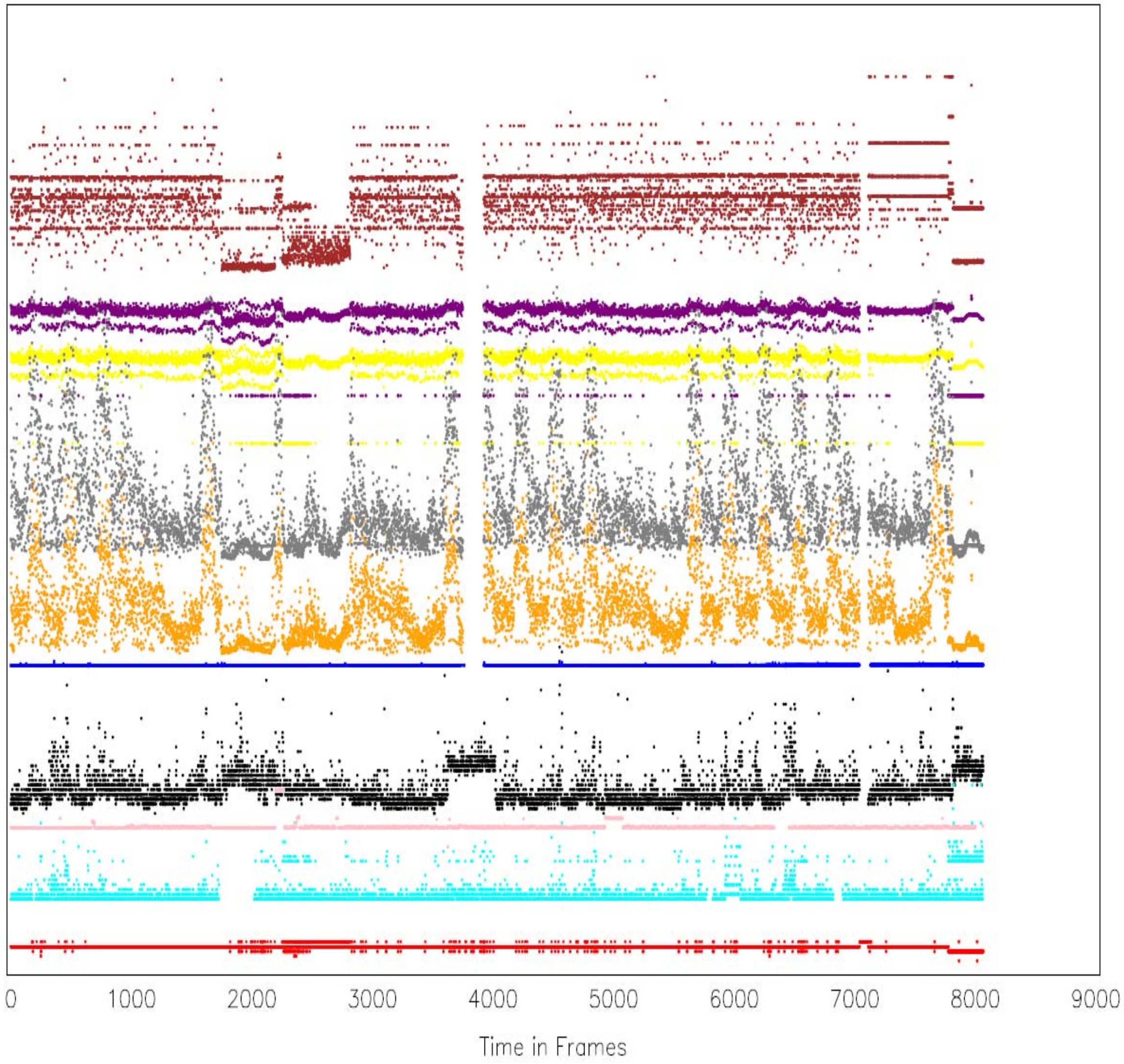


IP Data Streams: Attributes

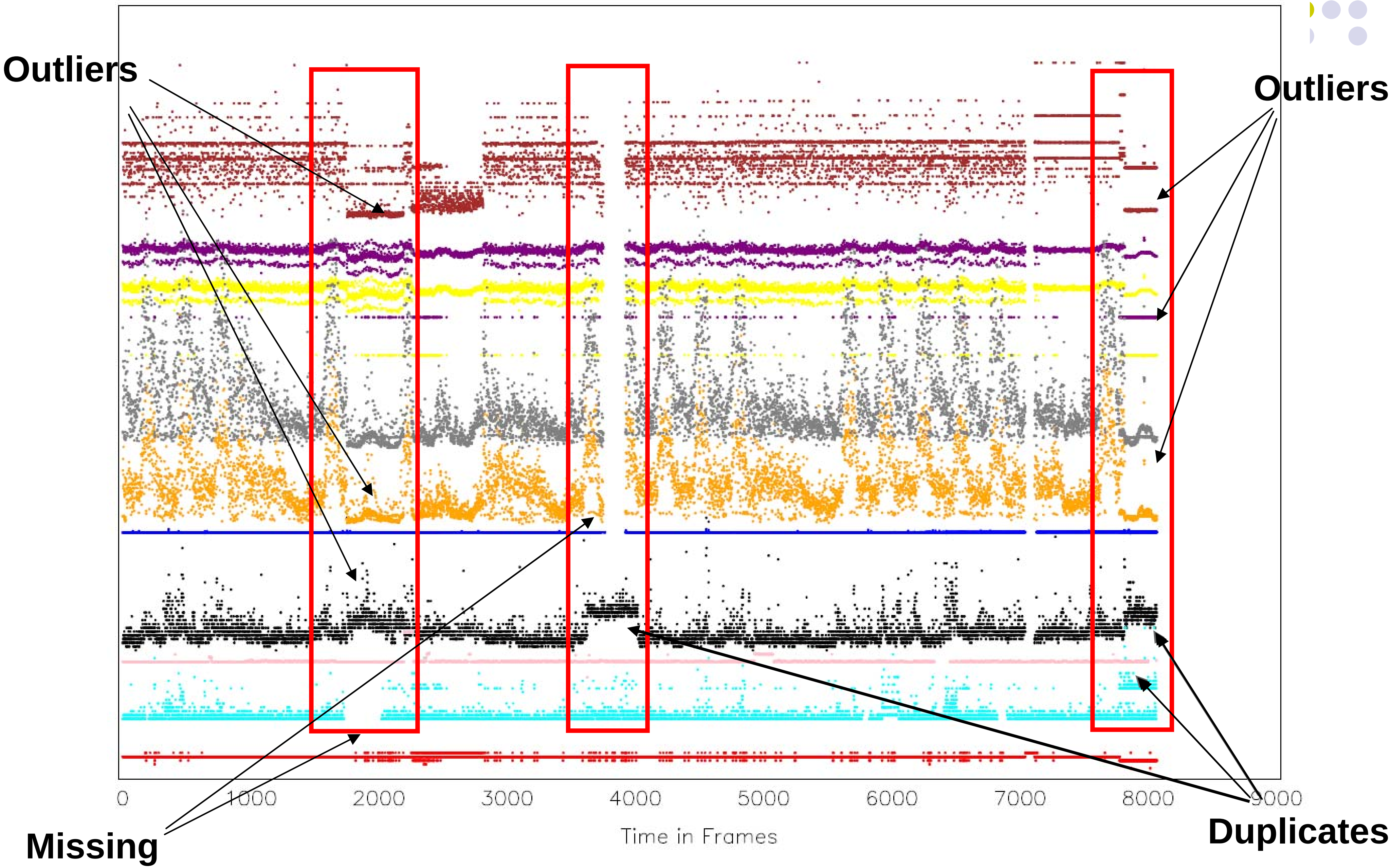
- Data measured at desired frequency
  - Real-time, minutes, hours
- Massive amounts of data!
- Attributes
  - Resource usage
  - Traffic measurements
  - Performance metrics
  - Alarms
- Gathered from multiple, disparate sources

IP Data Streams: A Picture

- 10 Attributes, every 5 minutes, over four weeks
- Axes transformed for plotting
- Glitches
  - Automatic detection



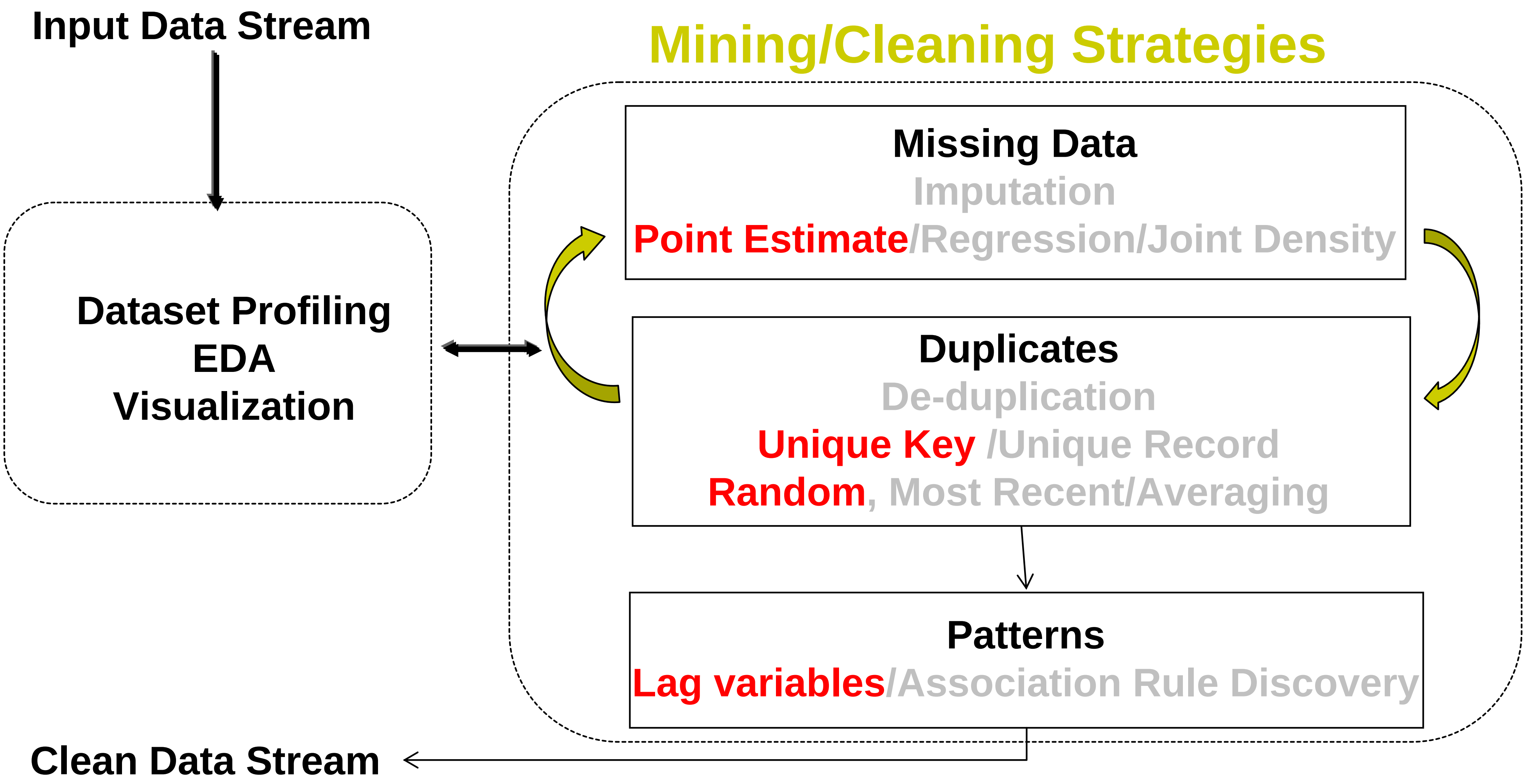
IP Data: Multivariate Glitches



DQM Tasks: IP data streams

- Goal
  - Extract the "cleanest" dataset
  - Discover patterns and interactions between glitches
- Challenges
  - Domain specific: dynamic nature of the network
  - Data integration
- Focus
  - Missing Data
  - Duplicates
  - Outliers
  - Others are not considered in the case study

IP Data: Data Quality Mining



Best DQM Strategy?

Best Strategy Definition

- *Ideal data*: Super clean data, any record that has no ambiguity or imperfection
- *Representation*: A histogram (MV histogram using kdq-tree)
- *Distance*: Kullback Leibler distance
- *Best strategy*: Smallest KL distance to the Ideal

Finding the optimal dataset

MV strategy shown with Univariate projection

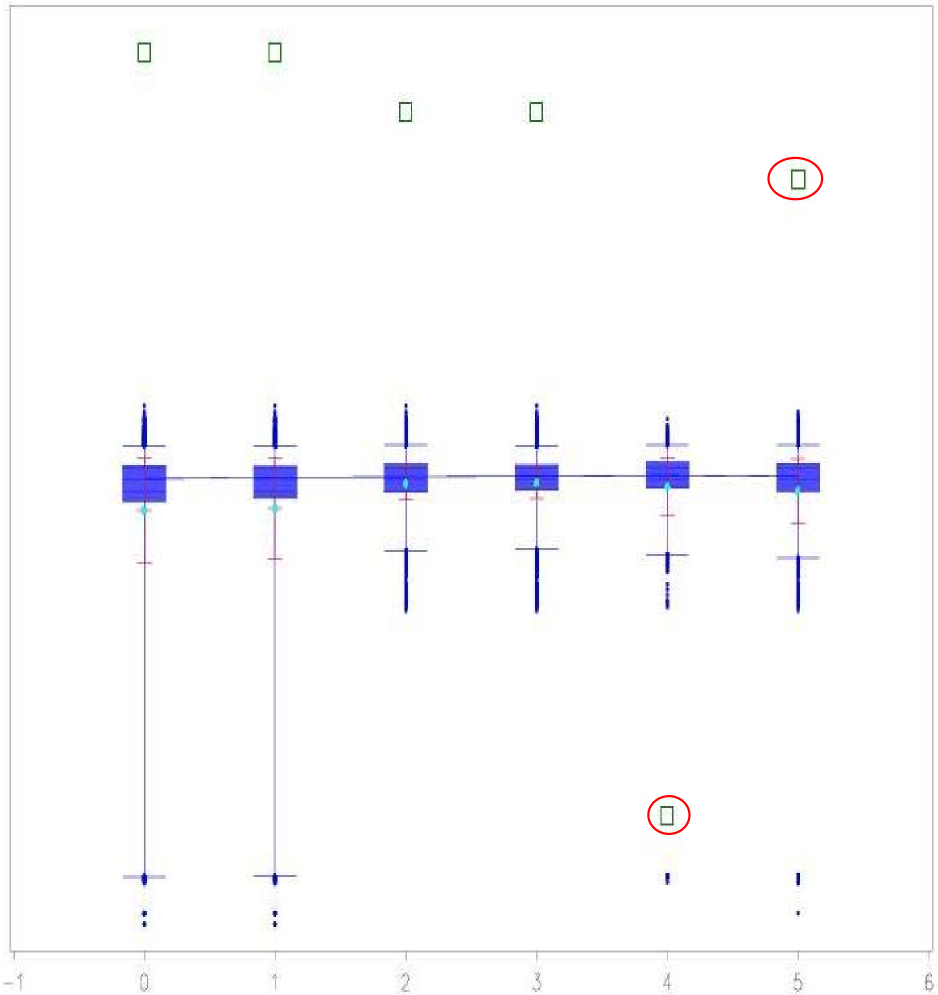
0= Raw data 1= Median-based imputation from 4

2= De-duplicated data

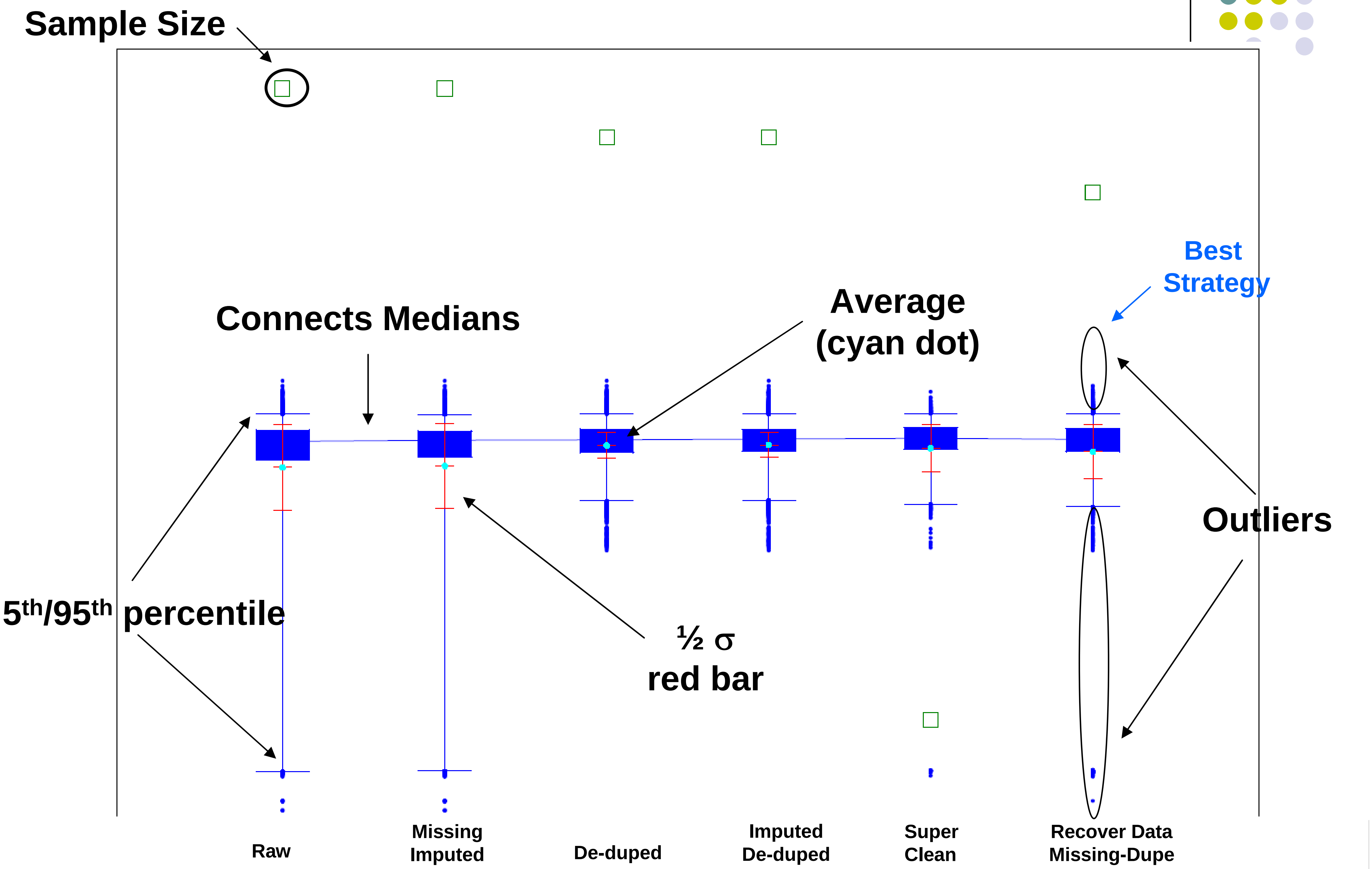
3= 1+2

4= Super clean (Full deletion)

5= Strategy driven by the discovery of patterns of glitches

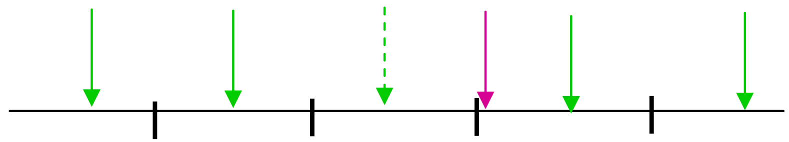


IP Data Cleaning Strategies
Boxplot comparison



What are the patterns?

Missing-Duplicate pairs



Types of Patterns

- Univariate/Multivariate
- Discovered Patterns
  - Missing-duplicate pairs: the responses arrive at the bin boundary resulting in the pattern
  - Complex patterns:
    - co-occurrence or lagged occurrence of outliers and missing,
    - outliers and duplicates
    - missing and duplicates
- Cleaning strategies
  - Quantitative cleaning, e.g., blind imputation
  - Domain knowledge-driven replacement of missing values with adjacent duplicates
  - Additional iterations needed because cleaning reveals new glitches

Case Study: Conclusion

- IP data stream – multivariate, massive, glitchy
- Critical for network monitoring
- Patterns and dependencies in glitches are used to recover much of the data such that the treated dataset is close to the ideal dataset
- Discovery of explanatory variables is useful for understanding recurrent DQ problems

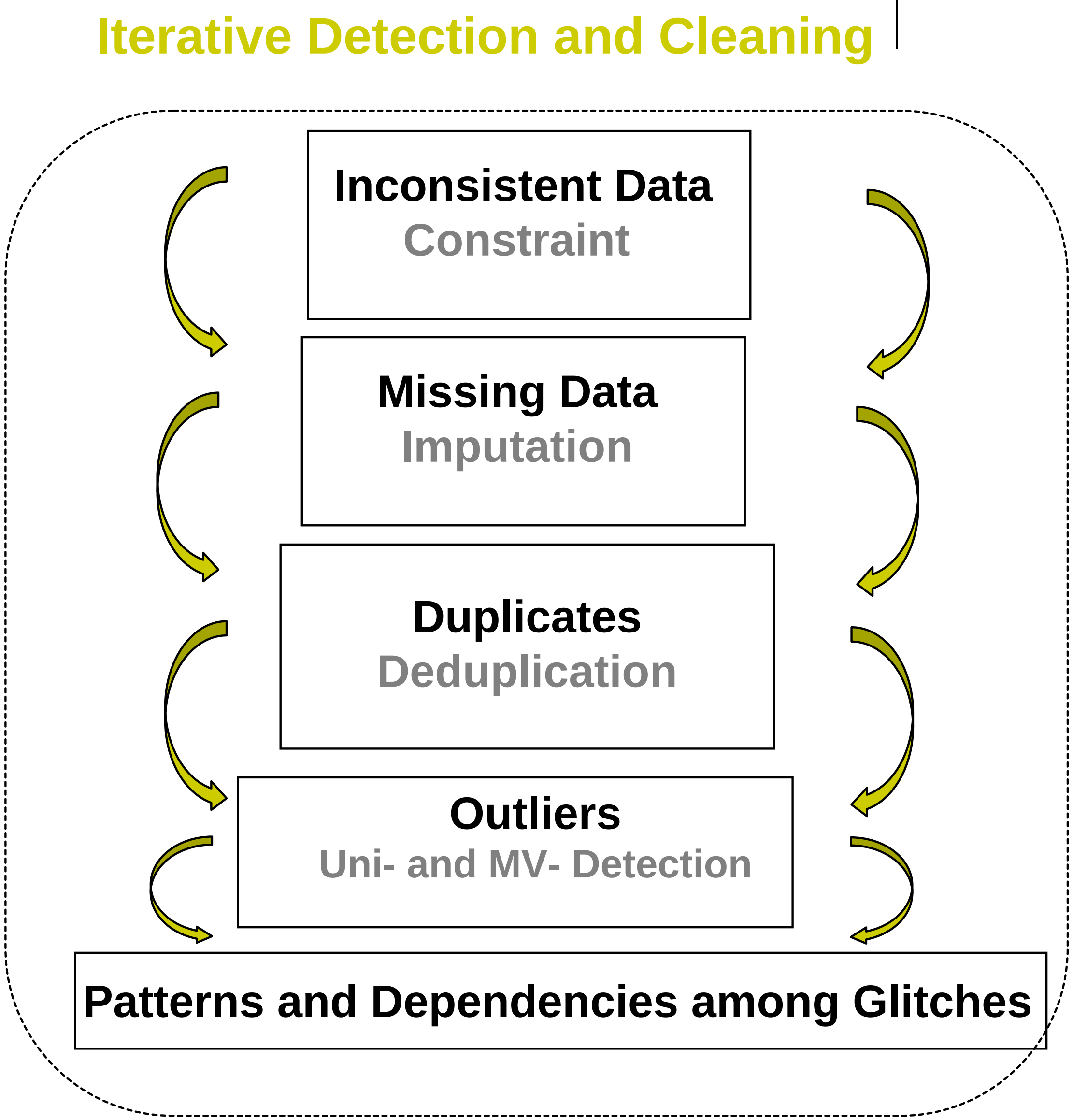
In Summary

DQM Summary:
Multivariate Glitches

- Glitches are multivariate with strong interdependencies
  - Static & temporal
  - Domain and application dependent
- DQM framework is important
  - Extant approaches tend to treat each class of glitches separately – misleading.
- Patterns and distribution of glitches are crucial in formulating cleaning strategies

DQM Summary: Process and Strategies

- Iterative and complementary cleaning strategies
- Best DQM strategies
  - Quantitative criteria
  - Resource-dependent
  - Domain, user and operational needs



Thanks

Any questions?

Data Quality: Up-coming Events

August 24, 2009: **QDB (Quality in Databases)**
Workshop in conjunction with VLDB 2009 in Lyon, France
http://qdb09.irisa.fr

November 7-9, 2009: **ICIQ (International Conference on Information Quality)**
Hasso-Plattner Institut, Potsdam, Germany
http://www.hpi.uni-potsdam.de/naumann/iciq2009/

Limited Bibliography

References

Books

- BATINI, Carlo, SCANNAPIECO, Monica. Data Quality Concepts, Methodologies and Techniques. Data-Centric Systems and Applications. Springer-Verlag, 2006.
- BARNETT, V., LEWIS, T., Outliers in statistical data. John Wiley, Chichester, 1994.
- DASU, Tamraparni, JOHNSON, Theodore. Exploratory Data Mining and Data Cleaning. John Wiley, 2003.
- HAWKINS, D., Identification of Outliers. Chapman and Hall, London, 1980.
- HERZOG, Thomas N., SCHEUREN, Fritz J., WINKLER, William E., Data Quality and Record Linkage Techniques, Springer, May 2007.
- KIMBALL, Ralph, CASERTA, Joe. The Data Warehouse ETL Toolkit, Wiley, 2004.
- NAUMANN, Felix Quality-Driven Query Answering for Integrated Information Systems. Lecture Notes in Computer Science, vol. 2261. Springer-Verlag, 2002.
- Tukey, John Wilder. Exploratory Data Analysis. Addison-Wesley, 1977
- WANG, Richard Y., ZIAD, Mostapha, LEE, Yang W. Data Quality. Advances in Database Systems, vol. 23. Kluwer Academic Publishers, 2002.

Surveys

- CHANDOLA, Varun, BANERJEE, Arindam, KUMAR, Vipin, Anomaly Detection A Survey. ACM Computing Surveys, September 2009.
- ELMAGARMID, Ahmed K., IPEIROTIS, Panagiotis G., VERYKIOS, Vassilios S., Duplicate Record Detection A Survey, IEEE Transations on knowledge and Data Engineering (TKDE) Vol. 19 No. 1 January 2007, pp. 1-16.
- HELLERSTEIN, Joseph, Quantitative Data Cleaning for Large Databases. White paper, United Nations Economic Commission for Europe, February, 2008. http://db.cs.berkeley.edu/jmh/cleaning-unece.pdf
- NAVARRO, Gonzalo. A Guided Tour to Approximate String Matching. ACM Comput. Surv., 33(1), pp. 31–88, 2001.
- WINKLER, William E., Overview of Record Linkage and Current Research Directions, Tech. Rep. of U.S. Census Bureau, February. 2006 http://www.census.gov/srd/papers/pdf/rrs2006-02.pdf

References

Tutorials

- BATINI, Carlo, CATARCI, Tiziana, SCANNAPIECO, Monica. A Survey of Data Quality Issues in Cooperative Systems. Tutorial ER 2004.
- KOUDAS, Nick, SARAWAGI, Sunita, SRIVASTAVA, Divesh. Record Linkage Similarity Measures and Algorithms. Tutorial SIGMOD 2006.
- BANERJEE, Arindam, CHANDOLA, Varun, KUMAR, Vipin, SRIVASTAVA Jaideep, LAZAREVIC, Aleksandar. Anomaly Detection A Tutorial. Tutorial SIAM Conf. on Data Mining 2008.
- KRIEGEL, Hans-Peter, KROGER, Peer, ZIMEK, Arthur. Outlier Detection Techniques. Tutorial, PAKDD 2009. http://www.dbs.informatik.uni-muenchen.de/Publikationen/Papers/tutorial_slides.pdf

Data Profiling

- CARUSO, FRANCESCO, COCHINWALA, MUNIR, GANAPATHY, UMA, LALK, GAIL, MISSIER, PAOLO. 2000. Telcordia’s Database Reconciliation and Data Quality Analysis Tool. Proc. VLDB 2000, pp. 615–618, 2000.
- DASU, TAMRAPARNI, JOHNSON, THEODORE, S. Muthukrishnan, V. Shkapenyuk, Mining Database Structure; Or, How to Build a Data Quality Browser, Proc. SIGMOD 2002.

Data Preparation and Data Quality Mining

- HIPP, J., GUNTZER, U., GRIMMER, U. Data Quality Mining - Making a Virtue of Necessity. Proc. Workshop DMKD 2001.
- LUBBERS, D., GRIMMER, U., JARKE, M. Systematic Development of Data Mining-Based Data Quality Tools. Proc. VLDB 2003, pp. 548-559, 2003.
- KLINE, R.B., Data Preparation and Screening, Chapter 3. in Principles and Practice of Structural Equation Modeling, NY Guilford Press, pp. 45-62, 2005.
- PEARSON, Ronald K. Surveying Data for Patchy Structure. SDM 2005.
- STATNOTES Topics in Multivariate Analysis. Retrieved 10/17/2008 from http://www2.chass.ncsu.edu/garson/pa765/statnote.htm

References

Data Cleaning – ETL

- BILKE, Alexander, BEIHOLDER, Jens, BOHM, Christoph, DRABA Karsten, NAUMANN, Felix, WEIS, Melanie. Automatic Data Fusion with HumMer. Proc. VLDB 2005 1251-1254, 2005.
- CHAUDHURI, Surajit, GANTI, Venkateh, KAUSHIK, Raghav. A Primitive Operator for Similarity Joins in Data Cleaning. Proc. ICDE 2006.
- CHRISTEN, Peter. Febrl an open source data cleaning, deduplication and record linkage system with a graphical user interface. KDD 2008, pp. 1065-1068, 2008.
- CHRISTEN, Peter, CHURCHES, Tim, ZHU, Xi. Probabilistic name and address cleaning and standardization. Proc. Australasian Data Mining Workshop 2002. http://cs.anu.edu.au/~Peter.Christen/publications/adm2002-cleaning.pdf
- GALHARDAS, Helena, FLORESCU, Daniela, SHASHA, Dennis, SIMON, Eric, SAITA, Cristian-Augustin. Declarative Data Cleaning Language, Model, and Algorithms, Proc. VLDB Conf., pp. 371-380, 2001.
- HERNANDEZ, M., STOLFO, S., Real-World Data is Dirty Data Cleansing and the Merge/Purge Problem, Data Mining and Knowledge Discovery, 2(1)9-37, 1998.
- RAHM, E., DO, H.H., Data Cleaning Problems and Current Approaches, Data Engineering Bulletin, 23(4) 3-13, 2000.
- RAMAN, V., HELLERSTEIN, J.M. Potter's Wheel: An Interactive Data Cleaning System. Proc. VLDB 2001, pp. 381-390, 2001.
- VASSILIADIS, P., VAGENA, Z., SKIADOPOULOS, S., KARAYANNIDIS, N., SELLIS, T. ARKTOS A Tool For Data Cleaning and Transformation in Data Warehouse Environments. Bulletin of the Technical Committee on Data Engineering, 23(4), pp. 42-47, 2000.
- VASSILIADIS, P., KARAGIANNIS A., TZIOVARA, V., SIMITSIS, A. Towards a Benchmark for ETL Workflows. Proc. QDB 2007, pp. 49-60, 2007.
- WEIS, Melanie, MANOLESCU, Ioana. XClean in Action (Demo). CIDR 2007, pp. 259-262, 2007.

References

Record Linkage and duplicate detection (1/2)

- ANANTHAKRISHNA, ROHIT, CHAUDHURI, SURAJIT, GANTI, VENKATESH. Eliminating Fuzzy Duplicates in Data Warehouses. pp. 586–597, Proc. of VLDB 2002.
- BANSAL, NIKHIL, BLUM, AVRIM, CHAWLA, SHUCHI. Correlation clustering. Machine Learning, 56(1-3):89–113, 2004.
- BAXTER, ROHAN A., CHRISTEN, PETER, CHURCHES, TIM. A Comparison of Fast Blocking Methods for Record Linkage. pp. 27–29 Proc. of the KDD'03 Workshop on Data Cleaning, Record Linkage and Object Consolidation, 2003.
- BHATTACHARYA, INDRAJIT, GETOOR, LISE. Iterative Record Linkage for Cleaning and Integration. pp. 11–18 Proc. of the 9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, DMKD, 2004.
- BHATTACHARYA, INDRAJIT, GETOOR, LISE. Collective entity resolution in relational data. TKDD, 1(1), 2007.
- BILENKO, MIKHAIL, MOONEY, RAYMOND J. Adaptive Duplicate Detection Using Learnable String Similarity Measures. Proc. KDD 2003, pp. 39–48, 2003.
- BILENKO, MIKHAIL, BASU, SUGATO, SAHAMI, MEHRAN. 2005. Adaptive Product Normalization Using Online Learning for Record Linkage in Comparison Shopping. Proc. ICDM 2005, pp. 58–65, 2005.
- CHRISTEN, Peter, Automatic Record Linkage using Seeded Nearest Neighbour and Support Vector Machine Classification, ACM SIGKDD 2008 Conf., Las Vegas, August 2008.
- ELFEKY, MOHAMED G., ELMAGARMID, AHMED K., VERYKIOS, VASSILIOS S. TAILOR A Record Linkage Tool Box. pp. 17–28 Proc. of the 18th International Conf. on Data Engineering, ICDE 2002. San Jose, CA, USA, 2002.
- ELMAGARMID, AHMED K., IPEIROTIS, PANAGIOTIS G., VERYKIOS, VASSILIOS S. Duplicate Record Detection A Survey. IEEE Trans. Know. Data Eng., 19(1), 1–16, 2007.
- FELLEGI, IVAN P., SUNTER, A.B. A Theory for Record Linkage. Journal of the American Statistical Association, 64, 1183–1210, 1969.

References

Record Linkage and duplicate detection (2/2)

- GRAVANO, Luis, IPEIROTIS, Panagiotis G., JAGADISH, H. V., KOUDAS, Nick, MUTHUKRISHNAN, S., PIETARINEN, Lauri, SRIVASTAVA, Divesh. Using q-grams in a DBMS for Approximate String Processing. IEEE Data Eng. Bull., 24(4), 28–34, 2001.
- GRAVANO, LUIS, IPEIROTIS, PANAGIOTIS G., KOUDAS, NICK, SRIVASTAVA, DIVESH. Text Joins for Data Cleansing and Integration in an RDBMS. Proc. ICDE 2003, pp. 729–731, Bangalore, India, 2003.
- HERNANDEZ, M., STOLFO, S., The Merge/Purge Problem for Large Databases, Proc. SIGMOD Conf pg 127-135, 1995.
- LOW, WAI LUP, LEE, MONG-LI, LING, TOK WANG. A Knowledge-Based Approach for Duplicate Elimination in Data Cleaning. Inf. Syst., 26(8), 585–606, 2001.
- KANG, Hyunmo, GETOOR, Lise, SHNEIDERMAN, Ben, BILGIC, Mustafa, LICAMELE, Louis. Interactive Entity Resolution in Relational Data: A Visual Analytic Tool and Its Evaluation. IEEE Trans. Vis. Comput. Graph. 14(5), pp. 999-1014, 2008.
- MCCALLUM, ANDREW, NIGAM, KAMAL, UNGAR, LYLE H. 2000. Efficient Clustering of High-Dimensional Data Sets with Application to Reference Matching. Proc. KDD 2000, pp. 169–178. Boston, MA, USA.
- MONGE, ALVARO E. 2000. Matching Algorithms within a Duplicate Detection System. IEEE Data Eng. Bull., 23(4), 14–20.
- TEJADA, SHEILA, KNOBLOCK, CRAIG A., MINTON, STEVEN. 2002. Learning Domain-Independent String Transformation Weights for High Accuracy Object Identification. Proc. KDD 2002, pp. 350–359, 2002.
- WEIS, MELANIE, NAUMANN, FELIX, BROSY, FRANZISKA. 2006. A Duplicate Detection Benchmark for XML (and Relational) Data. Proc. ACM SIGMOD 2006 Workshop on Information Quality in Information Systems, IQIS 2006. Chicago, IL, USA.
- WINKLER, WILLIAM E. Methods for Evaluating and Creating Data Quality. Inf. Syst., 29(7), 531–550, 2004.
- WINKLER, WILLIAM E., THIBAUDEAU, YVES. An Application of the Fellegi-Sunter Model of Record Linkage to the 1990 U.S. Decennial Census. Tech. Rept. Statistical Research Report Series RR91/09. U.S. Bureau of the Census, Washington, DC, USA, 1991.

References

Inconsistencies

- BOHANNON, Philip, FAN Wenfei, GEERTS, Floris, JIA, Xibei, KEMENTSIETSIDIS, Anastasios Conditional Functional Dependencies for Data Cleaning. Proc. ICDE 2007, pp. 746-755.
- BRAVO, Loreto, FAN, Wenfei, MA, Shuai. Extending Dependencies with Conditions. Proc. VLDB 2007, pp. 243-254.
- CERI, Stefano, Di GIUNTA, Francesco, LANZI, Pier Luca. Mining constraint violations. ACM Trans. Database Syst., 32(1): 6, 2007.
- CHANDEL, A., KOUDAS, Nick, PU, K., SRIVASTAVA Divesh. Fast Identication of Relational Constraint Violations. Proc. ICDE 2007.
- FAN, Wenfei, GEERTS, Floris, KEMENTSIETSIDIS, Anastasios Conditional functional dependencies for capturing data inconsistencies. TODS:33(2), June 2008.
- FAN, Wenfei, GEERTS, Floris, JIA, Xibei Semandaq A Data Quality System Based on Conditional Functional Dependencies, VLDB'08, (demo), 2008.
- FAN, Wenfei, GEERTS, Floris, LAKSHMANAN, Laks V. S., XIONG, Ming. Discovering Conditional Functional Dependencies. Proc. ICDE 2009, pp. 1231-1234.
- GOLAB, Lukasz, KARLOFF, Howard J., KORN, Flip, SRIVASTAVA Divesh, YU, Bei. On generating near-optimal tableaux for conditional functional dependencies. PVLDB 1(1) 376-390, 2008.
- KORN, Flip, MUTHUKRISHNAN S., ZHU, Yunyue Checks and Balances Monitoring Data Quality Problems in Network Traffic Databases. Proc. VLDB 2003, pp. 536-547.

Change Detection

- AGGARWAL, C. C. A framework for diagnosing changes in evolving data streams. Proc. ACM SIGMOD 2003.
- DASU, T., KRISHNAN S., LIN, D., VENKATASUBRAMANIAN, S., YI, K. Change (Detection) you can believe in Finding Distributional Shifts in Data streams. Proc. IDA'09, 2009.
- DASU, T., KRISHNAN S., VENKATASUBRAMANIAN, S., YI, K. An information-theoretic approach to detecting changes in multi-dimensional data streams. Proc. Interface'06, 2006.
- SONG, X., WU, M., JERMAINE, C., RANKA S. Statistical change detection for multidimensional data. Proc. ACM SIGKDD'07, pp. 667-676, 2007.

References

Outlier Detection (1/2)

- AGARWAL, D., Detecting anomalies in cross-classified streams a Bayesian approach. Know. Inf. Syst., 11(1):29–44, 2006.
- ANGIULLI, F., PRIZZUTI, C., Fast Outlier Detection in High Dimensional Spaces. Proc. Conf. on Principles of Data Mining and Knowledge Discovery, pp. 15–26, 2002.
- BAY, D.S., SCHWABACHER, M., Mining distance-based outliers in near linear time with randomization and a simple pruning rule. Proc. KDD 2003.
- BREUNIG, M., KRIEGEL, H-P., NG, R.T., SANDER, J., LOF Identifying Density-Based Local Outliers. Proc. of the 2000 ACM SIGMOD International Conf. on Management of Data, pp. 93–104. Dallas, TX, USA, 2000.
- CHEN, Y., DANG, X., PENG, H., and BART, H., Outlier detection with the kernelized spatial depth function. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008.
- CORMODE, G., HADJIELEFTHERIOU, M., Finding frequent items in data streams. Proc. VLDB 2008.
- ESKIN, E., Anomaly detection over noisy data using learned probability distributions. Proc. ICML 2000, pp. 255-262, 2000.
- FILZMOSER, P., MARONNA, R., WERNER, M. Outlier detection in high dimensions. Computational Statistics and Data Analysis, 52, pp. 1694–1711, 2008.
- GALEANO, P., PENA, D., TSAY, R. S. Outlier detection in multivariate time series by projection pursuit. Journal of American Statistical Association, 101(474):654–669, 2006.
- HAN, F., WANG, Y., WANG H., Odabk: An effective approach to detecting outlier in data stream. Proc. Intl. Conf. on Mach. Learn. and Cybernetics, pp. 1036–1041, 2006.
- HE, Z., XU, X., DENG, S., Discovering cluster-based local outliers. Pattern Recognition Letters, 24(9-10), pp. 1641–1650, 2003.
- HUBERT, M., VADER VEEKEN, S., Outlier detection for skewed data. Journal of Chemometrics, 22, pp. 235–246, 2007.
- JIANG, S.-Y., LI, Q.-H., LI, K.-L., WANG, H., MENG, Z.-L., GLOF a new approach for mining local outlier. Proc. Int. Conf. Mach. Learn. Cybernetics, vol. 11, pp. 157–162, 2003.

References

Outlier Detection (2/2)

- JIN, W., TUNG, A.K.H., HAN, J., Mining Top-n Local Outliers in Large Databases. Proc. KDD 2001, pp. 293–298, 2001.
- KIFER, D., BEN-DAVID, S., GEHRKE, J., Detecting changes in data streams. Proc. VLDB 2004, pages 180–191, 2004.
- KNORR, Edwin M., NG, Raymond T., Algorithms for Mining Distance-Based Outliers in Large Datasets. Proc. VLDB 1998, pp. 392–403, 1998.
- LIU, R., SINGH, K., TENG, J., Ddma-charts: Nonparametric multivariate moving average control charts based on data depth. Advances in Statistical Analysis, 88, pp. 235–258, 2004.
- MARONNA, R., ZAMAR, R., Robust estimates of location and dispersion for high-dimensional data sets. Technometrics, 44(4), pp. 307–317, 2002.
- PAPADIMITRIOU, S., KITAGAWA, H., GIBBONS, P.B., FALOUTSOS, C., LOCI: Fast outlier detection using the local correlation integral. Tech. Rep. Intel Research Lab, IRP-TR-02-09, July 2002.
- PENA, D., PRIETO, F., Multivariate outlier detection and robust covariance matrix estimation. Technometrics, 43(3):286–310, 2001.
- RAMASWAMY, S., RASTOGI, R., KYUSEOK, S., Efficient algorithms for mining outliers from large data sets. Proc. ACM SIGMOD 2000, pp. 427–438, 2000.
- ROUSSEEUW, P.J., DRIESSEN, K.V., A fast algorithm for the minimum covariance determinant estimator. Technometrics, 41(3), pp. 212–223, 1999.
- ROUSSEEUW, P.J., Van ZOMEREN, B.C., Unmasking Multivariate Outliers and Leverage Points, Journal of the American Statistical Association, 85, pp. 633-639, 1990.
- SHAWNE-TAYLOR J., CRISTIANI N., Kernel methods for pattern analysis. Cambridge, 2005.
- SHYU, M.-L., CHEN, S.-C., SARINNAPAKORN, K., CHANG, L., A novel anomaly detection scheme based on principal component classifier. Proc. ICDM 20003, pp. 353-365, 2003.
- SU, L., HAN, W., YANG, S., ZOU, P., JIA, Y., Continuous adaptive outlier detection on distributed data streams. In HPCC, LNCS 4782, pp. 74–85, 2007.
- SUBRAMANIAM, S., PALPANAS, T., PAPADOPOULOS, D., KALOGERAKI, V., GUNOPULOS, D., Online outlier detection in sensor data using non-parametric models. Proc. VLDB 2006, pp. 187–198, 2006.
- TANG, J., CHEN, Z., FU, A.W.-C., CHEUNG, D.W.-L., Enhancing Effectiveness of Outlier Detections for Low Density Patterns. Proc. PAKDD 2002. LNAI 2336, 2002.
- ZHANG, J., GAO, Q., WANG, H., Spot: A system for detecting projected outliers from high-dimensional data streams. Proc. ICDE 2008, pp. 1628–1631, 2008.

References

Missing Values

- ACUNA, E., RODRIGUEZ, C., The treatment of missing values and its effect in the classifier accuracy. Classification, Clustering and Data Mining Applications, Springer-Verlag, pp. 639-648, 2004.
- BATISTA G., MONARD, M.C., An analysis of four missing data treatment methods for supervised learning. Applied Artificial Intelligence 17, pp. 519-533, 2003.
- DEMPSTER, Arthur P., LAIRD, Nan M., RUBIN, Donald B., Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, 39, 1–38, 1977.
- FAN, Wenfei, GEERTS, Floris. Relative Information Completeness, PODS'09, 2009.
- FARHANGFAR, A., KURGAN, L., DY, J., Impact of imputation of missing values on classification error for discrete data. Pattern Recognition, 41, 3692-3705, 2008.
- FENG, H.A.B., Chen, G.C., Yin, C.D., Yang, B.B., Chen, Y.E., A SVM regression based approach to filling in missing values. Knowledge-Based Intelligent Information and Engineering Systems (KES05). LNCS 3683, pp. 581-587, 2005.
- HUA, Ming, PEI, Jian. Cleaning Disguised Missing Data A Heuristic Approach, Proc. KDD 2007.
- LI, D., DEOGUN, J., SPAULDING, W. Towards Missing Data Imputation: A Study of Fuzzy K-means Clustering Method. Rough Sets and Current Trends in Computing. LNCS 3066, 2004.
- LITTLE, R. J. A., RUBIN, D. B., Statistical Analysis with Missing Data. New York John Wiley Sons, 1987.
- Mc KNIGHT, P. E., FIGUEREDO, A. J., SIDANI, S., Missing Data A Gentle Introduction. Guilford Press, 2007.
- PEARSON, RONALD K., The problem of disguised missing data. SIGKDD Explorations 8(1) 83-92, 2006.
- SCHAFER, J. L., Analysis of Incomplete Multivariate Data, New York Chapman and Hall, 1997.
- TIMM, H., DORING, C., KRUSE, R., Different approaches to fuzzy clustering of incomplete datasets. International Journal of Approximate Reasoning, 35, 2003.
- WU, C.-H., WUN, C.-H., CHOU, H.-J., Using association rules for completing missing data. Proc. Hybrid Intelligent Systems (HIS’04), pp. 236–241, 2004.