Data Quality and Data Cleaning: An Overview

Theodore Johnson

johnsont@research.att.com

AT&T Labs – Research

(Lecture notes for CS541, 02/12/2004)

Based on:

- Recent book

Exploratory Data Mining and Data Quality
Dasu and Johnson
(Wiley, 2004)

- SIGMOD 2003 tutorial.

A unique, integrated approach to exploratory data mining and data quality

Data analysts at information-intensive businesses are frequently asked to analyze new data sets that are often dirty—composed of numerous tables possessing unknown properties. Prior to analysis, this data must be cleaned and explored—often a long and arduous task. Ensuring data quality is a notoriously messy problem that can only be addressed by drawing on methods from many disciplines, including statistics, exploratory data mining, database management, and metadata coding.

Where other books on data mining and analysis focus primarily on the last stage of the analysis procedure, *Exploratory Data Mining and Data Cleaning* uses a uniquely integrated approach to data exploration and data cleaning to develop a suitable modeling strategy that will help analysts to more effectively determine and implement the final technique.

The authors, both seasoned data analysts at a major corporation, draw on their own professional experience to:

- Present a brief overview of the main analytical techniques used in data mining practices, such as univariate and multivariate summaries of attributes and their interactions including Q-Q plots, fractal dimension and histograms, nonparametric approaches incorporating data depth, and more
- Provide numerous references to the related literature on clustering, classification, regression, and more
- Focus on developing an evolving modeling strategy through an iterative data exploration loop and incorporation of domain knowledge
- Address methods of detecting, quantifying (metrics), and correcting data quality issues that significantly impact findings and decisions, using commercially available tools as well as new algorithmic approaches
- Use case studies to illustrate applications in real-life scenarios
- Highlight new approaches and methodologies, such as the DataSphere space partitioning and summary-based analysis techniques

A groundbreaking addition to the existing literature, *Exploratory Data Mining and Data Cleaning* serves as an important reference for data analysts who need to analyze large amounts of unfamiliar data, operations managers, and students in undergraduate or graduate-level courses dealing with data analysis and data mining.

TAMRAPARNI DASU, PhD, and THEODORE JOHNSON, PhD, are both members of the technical staff at AT&T Labs-Research in Florham Park, New Jersey.

Subscribe to our free Statistics eNewsletter at
wiley.com/enewsletters

Visit wiley.com/statistics

WILEY-
INTERSCIENCE
wiley.com

ISBN 0-471-26851-8

DASU
JOHNSON

Exploratory Data Mining and Data Cleaning

WILEY
INTER-
SCIENCE

WILEY

Exploratory Data Mining and Data Cleaning

Tamraparni Dasu
Theodore Johnson

WILEY SERIES IN PROBABILITY AND STATISTICS

Tutorial Focus

- What research is relevant to Data Quality?
  - DQ is pervasive and expensive. It is an important problem.
  - But the problems are so messy and unstructured that research seems irrelevant.
- This tutorial will try to structure the problem to make research directions more clear.
- Overview
  - Data quality process
    - Where do problems come from
    - How can they be resolved
  - Disciplines
    - Management
    - Statistics
    - Database
    - Metadata

Overview

- The meaning of data quality (1)
- The data quality continuum
- The meaning of data quality (2)
- Data quality metrics
- Technical tools
  - Management
  - Statistical
  - Database
  - Metadata
- Case Study
- Research directions

The Meaning of Data Quality (1)

Meaning of Data Quality (1)

- Generally, you have a problem if the data doesn't mean what you think it does, or should
  - Data not up to spec : garbage in, glitches, etc.
  - You don't understand the spec : complexity, lack of metadata.
- Many sources and manifestations
  - As we will see.
- Data quality problems are expensive and pervasive
  - DQ problems cost hundreds of billion $$$ each year.
  - Resolving data quality problems is often the biggest effort in a data mining study.

Example

T.Das|97336o8327|24.95|Y|-|0.0|1000
Ted J.|973-360-8779|2000|N|M|NY|1000

- Can we interpret the data?
  - What do the fields mean?
  - What is the key? The measures?
- Data glitches
  - Typos, multiple formats, missing / default values
- Metadata and domain expertise
  - Field three is Revenue. In dollars or cents?
  - Field seven is Usage. Is it *censored*?
    - Field 4 is a censored flag. How to handle censored data?

Data Glitches

- Systemic changes to data which are external to the recorded process.
  - Changes in data layout / data types
    - Integer becomes string, fields swap positions, etc.
  - Changes in scale / format
    - Dollars vs. euros
  - Temporary reversion to defaults
    - Failure of a processing step
  - Missing and default values
    - Application programs do not handle NULL values well …
  - Gaps in time series
    - Especially when records represent incremental changes.

Conventional Definition of Data Quality

- Accuracy
  - The data was recorded correctly.
- Completeness
  - All relevant data was recorded.
- Uniqueness
  - Entities are recorded once.
- Timeliness
  - The data is kept up to date.
    - Special problems in federated data: time consistency.
- Consistency
  - The data agrees with itself.

Problems …

- Unmeasurable
  - Accuracy and completeness are extremely difficult, perhaps impossible to measure.
- Context independent
  - No accounting for what is important. E.g., if you are computing aggregates, you can tolerate a lot of inaccuracy.
- Incomplete
  - What about interpretability, accessibility, metadata, analysis, etc.
- Vague
  - The conventional definitions provide no guidance towards practical improvements of the data.

Finding a modern definition

- We need a definition of data quality which
  - Reflects the use of the data
  - Leads to improvements in processes
  - Is measurable (we can define metrics)

- First, we need a better understanding of how and where data quality problems occur
  - The data quality continuum

The Data Quality Continuum

The Data Quality Continuum

- Data and information is not static, it flows in a data collection and usage process
  - Data gathering
  - Data delivery
  - Data storage
  - Data integration
  - Data retrieval
  - Data mining/analysis

Data Gathering

- How does the data enter the system?
- Sources of problems:
  - Manual entry
  - No uniform standards for content and formats
  - Parallel data entry (duplicates)
  - Approximations, surrogates – SW/HW constraints
  - Measurement errors.

Solutions

- Potential Solutions:
  - Preemptive:
    - Process architecture (build in integrity checks)
    - Process management (reward accurate data entry, data sharing, data stewards)
  - Retrospective:
    - Cleaning focus (duplicate removal, merge/purge, name & address matching, field value standardization)
    - Diagnostic focus (automated detection of glitches).

Data Delivery

- Destroying or mutilating information by inappropriate pre-processing
  - Inappropriate aggregation
  - Nulls converted to default values
- Loss of data:
  - Buffer overflows
  - Transmission problems
  - No checks

Solutions

- Build reliable transmission protocols
  - Use a relay server
- Verification
  - Checksums, verification parser
  - Do the uploaded files fit an expected pattern?
- Relationships
  - Are there dependencies between data streams and processing steps
- Interface agreements
  - Data quality commitment from the data stream supplier.

Data Storage

- You get a data set. What do you do with it?
- Problems in physical storage
  - Can be an issue, but terabytes are cheap.
- Problems in logical storage (ER → relations)
  - Poor metadata.
    - Data feeds are often derived from application programs or legacy data sources. What does it mean?
  - Inappropriate data models.
    - Missing timestamps, incorrect normalization, etc.
  - Ad-hoc modifications.
    - Structure the data to fit the GUI.
  - Hardware / software constraints.
    - Data transmission via Excel spreadsheets, Y2K

Solutions

- Metadata
  - Document and publish data specifications.
- Planning
  - Assume that everything bad will happen.
  - Can be very difficult.
- Data exploration
  - Use data browsing and data mining tools to examine the data.
    - Does it meet the specifications you assumed?
    - Has something changed?

Data Integration

- Combine data sets (acquisitions, across departments).
- Common source of problems
  - Heterogenous data : no common key, different field formats
    - Approximate matching
  - Different definitions
    - What is a customer: an account, an individual, a family, …
  - Time synchronization
    - Does the data relate to the same time periods? Are the time windows compatible?
  - Legacy data
    - IMS, spreadsheets, ad-hoc structures
  - Sociological factors
    - Reluctance to share – loss of power.

Solutions

- Commercial Tools
  - Significant body of research in data integration
  - Many tools for address matching, schema mapping are available.
- Data browsing and exploration
  - Many hidden problems and meanings : must extract metadata.
  - View before and after results : did the integration go the way you thought?

Data Retrieval

- Exported data sets are often a view of the actual data. Problems occur because:
  - Source data not properly understood.
  - Need for derived data not understood.
  - Just plain mistakes.
    - Inner join vs. outer join
    - Understanding NULL values
- Computational constraints
  - E.g., too expensive to give a full history, we'll supply a snapshot.
- Incompatibility
  - Ebcdic?

Data Mining and Analysis

- What are you doing with all this data anyway?
- Problems in the analysis.
  - Scale and performance
  - Confidence bounds?
  - Black boxes and dart boards
    - “fire your Statisticians”
  - Attachment to models
  - Insufficient domain expertise
  - Casual empiricism

Solutions

- Data exploration
  - Determine which models and techniques are appropriate, find data bugs, develop domain expertise.
- Continuous analysis
  - Are the results stable? How do they change?
- Accountability
  - Make the analysis part of the feedback loop.

The Meaning of Data Quality (2)

Meaning of Data Quality (2)

- There are many types of data, which have different uses and typical quality problems
  - Federated data
  - High dimensional data
  - Descriptive data
  - Longitudinal data
  - Streaming data
  - Web (scraped) data
  - Numeric vs. categorical vs. text data

Meaning of Data Quality (2)

- There are many uses of data
  - Operations
  - Aggregate analysis
  - Customer relations …
- Data Interpretation : the data is useless if we don’t know all of the *rules* behind the data.
- Data Suitability : Can you get the answer from the available data
  - Use of proxy data
  - Relevant data is missing

Data Quality Constraints

- Many data quality problems can be captured by *static* constraints based on the schema.
  - Nulls not allowed, field domains, foreign key constraints, etc.
- Many others are due to problems in workflow, and can be captured by *dynamic* constraints
  - E.g., orders above $200 are processed by Biller 2
- The constraints follow an 80-20 rule
  - A few constraints capture most cases, thousands of constraints to capture the last few cases.
- Constraints are measurable. Data Quality Metrics?

Data Quality Metrics

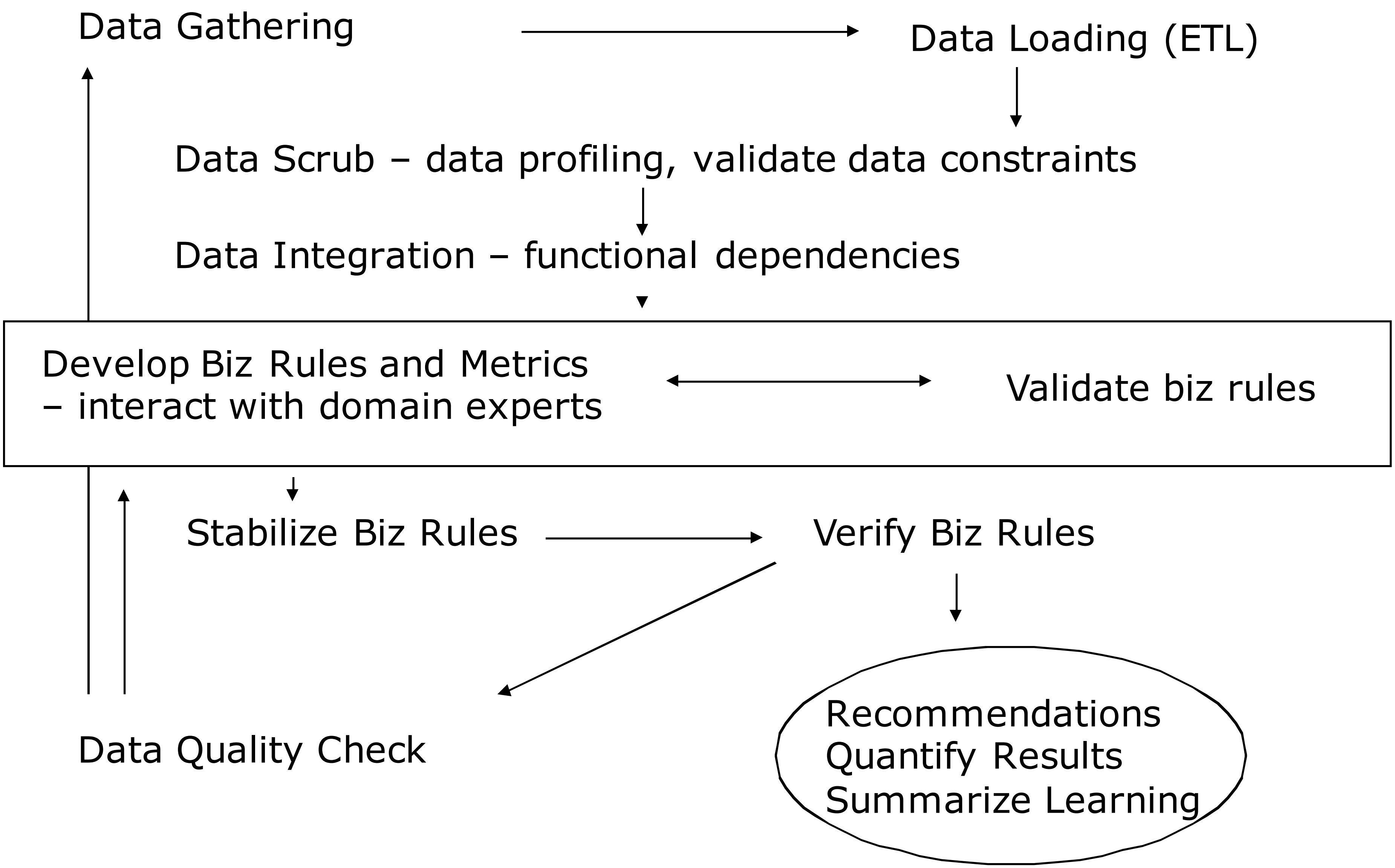
Data Quality Metrics

- We want a measurable quantity
  - Indicates what is wrong and how to improve
  - Realize that DQ is a messy problem, no set of numbers will be perfect
- Types of metrics
  - Static vs. dynamic constraints
  - Operational vs. diagnostic
- Metrics should be *directionally correct* with an improvement in use of the data.
- A very large number metrics are possible
  - Choose the most important ones.

Examples of Data Quality Metrics

- Conformance to schema
  - Evaluate constraints on a snapshot.
- Conformance to business rules
  - Evaluate constraints on changes in the database.
- Accuracy
  - Perform inventory (expensive), or use proxy (track complaints). Audit samples?
- Accessibility
- Interpretability
- Glitches in analysis
- Successful completion of end-to-end process

Data Quality Process



Data Gathering

Data Loading (ETL)

Data Scrub – data profiling, validate data constraints

Data Integration – functional dependencies

Develop Biz Rules and Metrics
– interact with domain experts

Validate biz rules

Stabilize Biz Rules

Verify Biz Rules

Data Quality Check

Recommendations
Quantify Results
Summarize Learning

Technical Tools

Technical Approaches

- We need a multi-disciplinary approach to attack data quality problems
  - No one approach solves all problem
- Process management
  - Ensure proper procedures
- Statistics
  - Focus on analysis: find and repair anomalies in data.
- Database
  - Focus on relationships: ensure consistency.
- Metadata / domain expertise
  - What does it mean? Interpretation

Process Management

- Business processes which encourage data quality.
  - Assign dollars to quality problems
  - Standardization of content and formats
  - Enter data once, enter it correctly (incentives for sales, customer care)
  - Automation
  - Assign responsibility : data stewards
  - End-to-end data audits and reviews
    - Transitions between organizations.
  - Data Monitoring
  - Data Publishing
  - Feedback loops

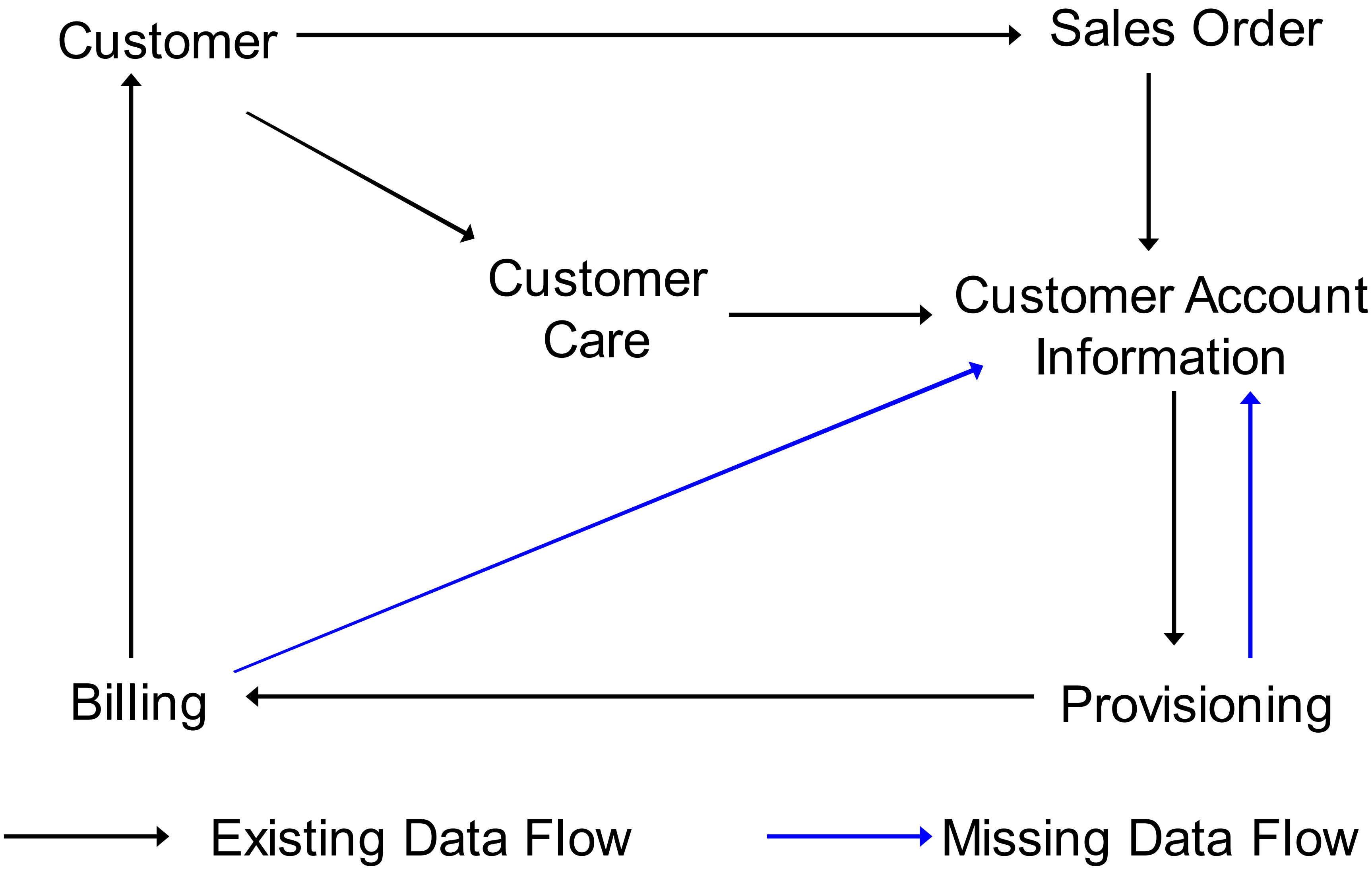
Feedback Loops

- Data processing systems are often thought of as open-loop systems.
  - Do your processing then throw the results over the fence.
  - Computers don’t make mistakes, do they?
- Analogy to control systems : *feedback loops*.
  - *Monitor* the system to detect difference between actual and intended
  - *Feedback loop* to correct the behavior of earlier components
  - Of course, data processing systems are much more complicated than linear control systems.

Example

- Sales, provisioning, and billing for telecommunications service
  - Many stages involving handoffs between organizations and databases
  - Simplified picture
- *Transition between organizational boundaries is a common cause of problems*.
- Natural feedback loops
  - Customer complains if the bill is to high
- Missing feedback loops
  - No complaints if we undercharge.

Example



Customer

Sales Order

Customer Care

Customer Account Information

Billing

Provisioning

Existing Data Flow

Missing Data Flow

Monitoring

- Use data monitoring to add missing feedback loops.
- Methods:
  - Data tracking / auditing
    - Follow a sample of transactions through the workflow.
    - Build secondary processing system to detect possible problems.
  - Reconciliation of incrementally updated databases with original sources.
  - Mandated consistency with a Database of Record (DBOR).
  - Feedback loop sync-up
  - Data Publishing

Data Publishing

- Make the contents of a database available in a readily accessible and digestible way
  - Web interface (universal client).
  - Data Squashing : Publish aggregates, cubes, samples, parametric representations.
  - Publish the metadata.
- Close feedback loops by getting a lot of people to look at the data.
- Surprisingly difficult sometimes.
  - Organizational boundaries, loss of control interpreted as loss of power, desire to hide problems.

Statistical Approaches

- No explicit DQ methods
  - Traditional statistical data collected from carefully designed experiments, often tied to analysis
  - But, there are methods for finding anomalies and repairing data.
  - Existing methods can be adapted for DQ purposes.
- Four broad categories can be adapted for DQ
  - Missing, incomplete, ambiguous or damaged data e.g truncated, censored
  - Suspicious or abnormal data e.g. outliers
  - Testing for departure from models
  - Goodness-of-fit

Missing Data

- Missing data - values, attributes, entire records, entire sections
- Missing values and defaults are indistinguishable
- Truncation/censoring - not aware, mechanisms not known
- Problem: Misleading results, bias.

Detecting Missing Data

- Overtly missing data
  - Match data specifications against data - are all the attributes present?
  - Scan individual records - are there gaps?
  - Rough checks : number of files, file sizes, number of records, number of duplicates
  - Compare estimates (averages, frequencies, medians) with “expected” values and bounds; check at various levels of granularity since aggregates can be misleading.

Missing data detection (cont.)

- Hidden damage to data
  - Values are truncated or censored - check for spikes and dips in distributions and histograms
  - Missing values and defaults are indistinguishable - too many missing values? metadata or domain expertise can help
  - Errors of omission e.g. all calls from a particular area are missing - check if data are missing randomly or are localized in some way

Imputing Values to Missing Data

- In federated data, between 30%-70% of the data points will have at least one missing attribute - data wastage if we ignore all records with a missing value
- Remaining data is seriously biased
- Lack of confidence in results
- Understanding pattern of missing data unearths data integrity issues

Missing Value Imputation - 1

- Standalone imputation
  - Mean, median, other point estimates
  - Assume: Distribution of the missing values is the same as the non-missing values.
  - Does not take into account inter-relationships
  - Introduces bias
  - Convenient, easy to implement

Missing Value Imputation - 2

- Better imputation - use attribute relationships
- Assume : all prior attributes are populated
  - That is, *monotonicity* in missing values.

X1	X2	X3	X4	X5
1.0	20	3.5	4	.
1.1	18	4.0	2	.
1.9	22	2.2	.	.
0.9	15	.	.	.

- Two techniques
  - Regression (parametric),
  - Propensity score (nonparametric)

Missing Value Imputation –3

- Regression method
  - Use linear regression, sweep left-to-right

    $X3=a+b*X2+c*X1;$

    $X4=d+e*X3+f*X2+g*X1$, and so on
  - X3 in the second equation is estimated from the first equation if it is missing

Missing Value Imputation - 3

- Propensity Scores (nonparametric)
  - Let Y_j=1 if X_j is missing, 0 otherwise
  - Estimate P(Y_j =1) based on X_1 through $X_{(j-1)}$ using logistic regression
  - Group by propensity score P(Y_j =1)
  - Within each group, estimate missing X_js from known X_js using approximate Bayesian bootstrap.
  - Repeat until all attributes are populated.

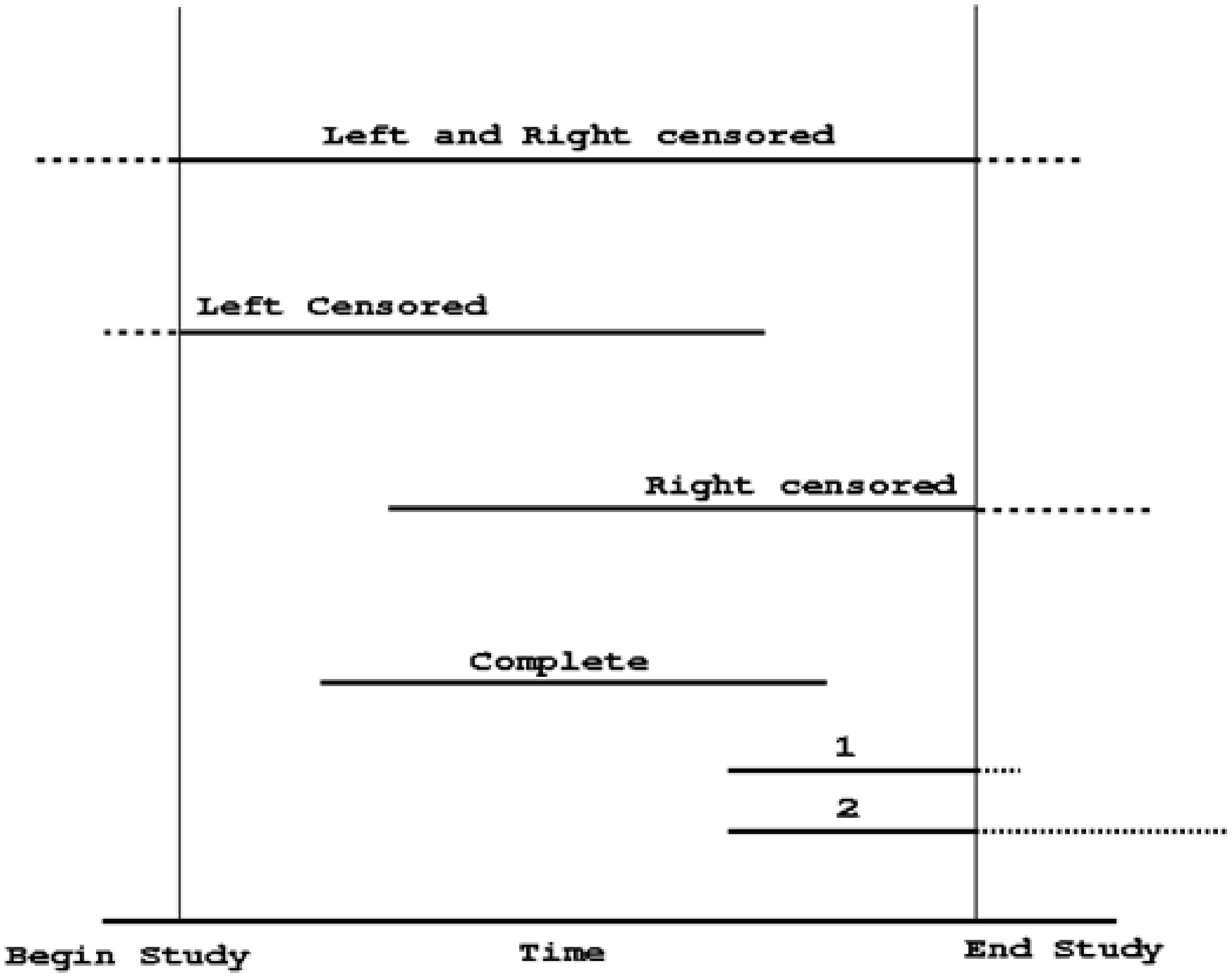
Missing Value Imputation - 4

- Arbitrary missing pattern
  - Markov Chain Monte Carlo (MCMC)
  - Assume data is multivariate Normal, with parameter Θ
  - (1) Simulate missing X, given Θ estimated from observed X ; (2) Re-compute Θ using filled in X
  - Repeat until stable.
  - Expensive: Used most often to induce monotonicity

- Note that imputed values are useful in aggregates but can’t be trusted individually

Censoring and Truncation

- Well studied in Biostatistics, relevant to time dependent data e.g. duration
- *Censored* - Measurement is bounded but not precise e.g. Call duration > 20 are recorded as 20
- *Truncated* - Data point dropped if it exceeds or falls below a certain bound e.g. customers with less than 2 minutes of calling per month

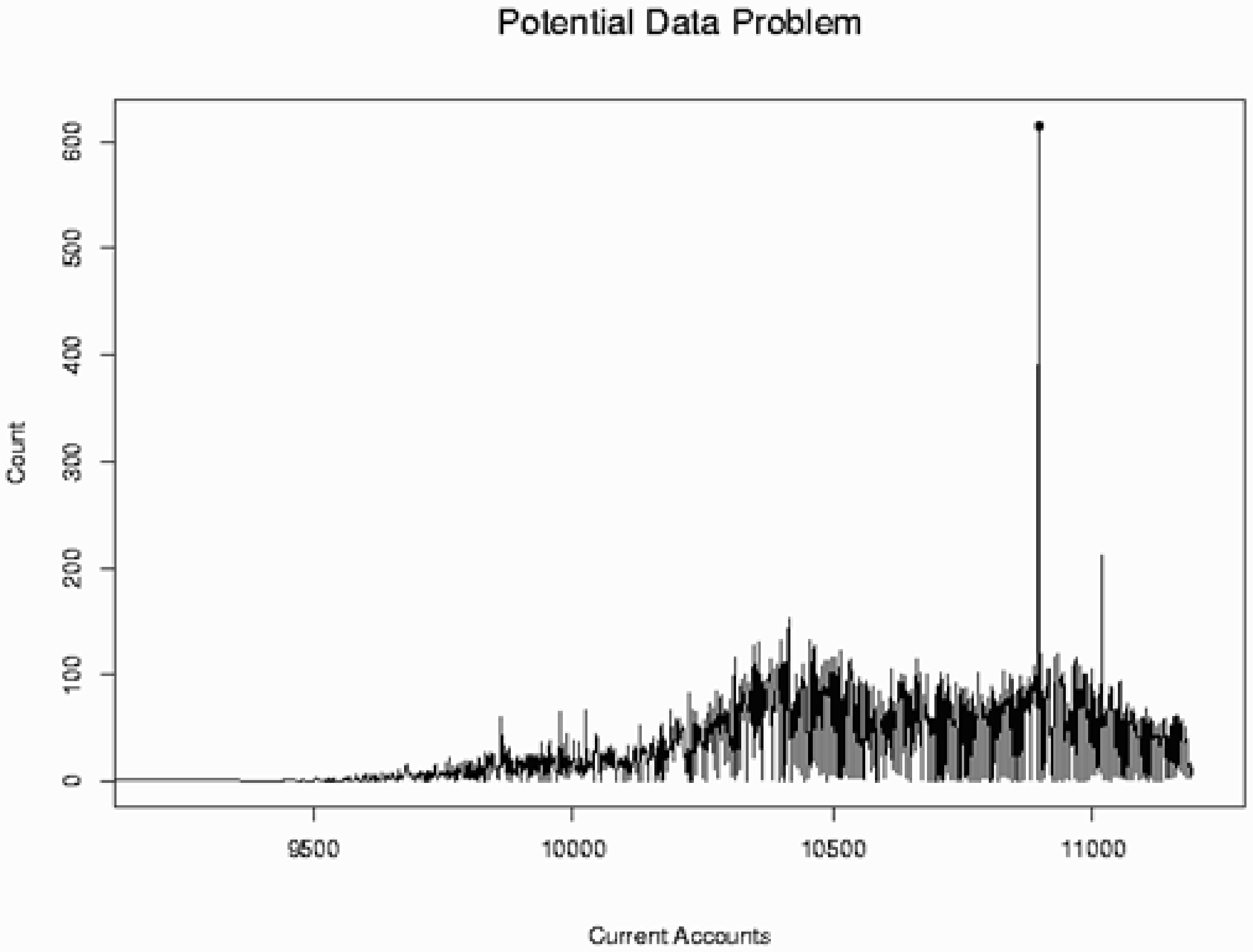


Censored time intervals

Censoring/Truncation (cont.)

- If censoring/truncation mechanism not known, analysis can be inaccurate and biased.
- But if you know the mechanism, you can mitigate the bias from the analysis.
- Metadata should record the existence as well as the nature of censoring/truncation

Spikes usually indicate censored time intervals caused by resetting of timestamps to defaults



Suspicious Data

- Consider the data points

  3, 4, 7, 4, 8, 3, 9, 5, 7, 6, 92

- “92” is suspicious - an *outlier*
- Outliers are potentially legitimate
- Often, they are data or model glitches
- Or, they could be a data miner’s dream, e.g. highly profitable customers

Outliers

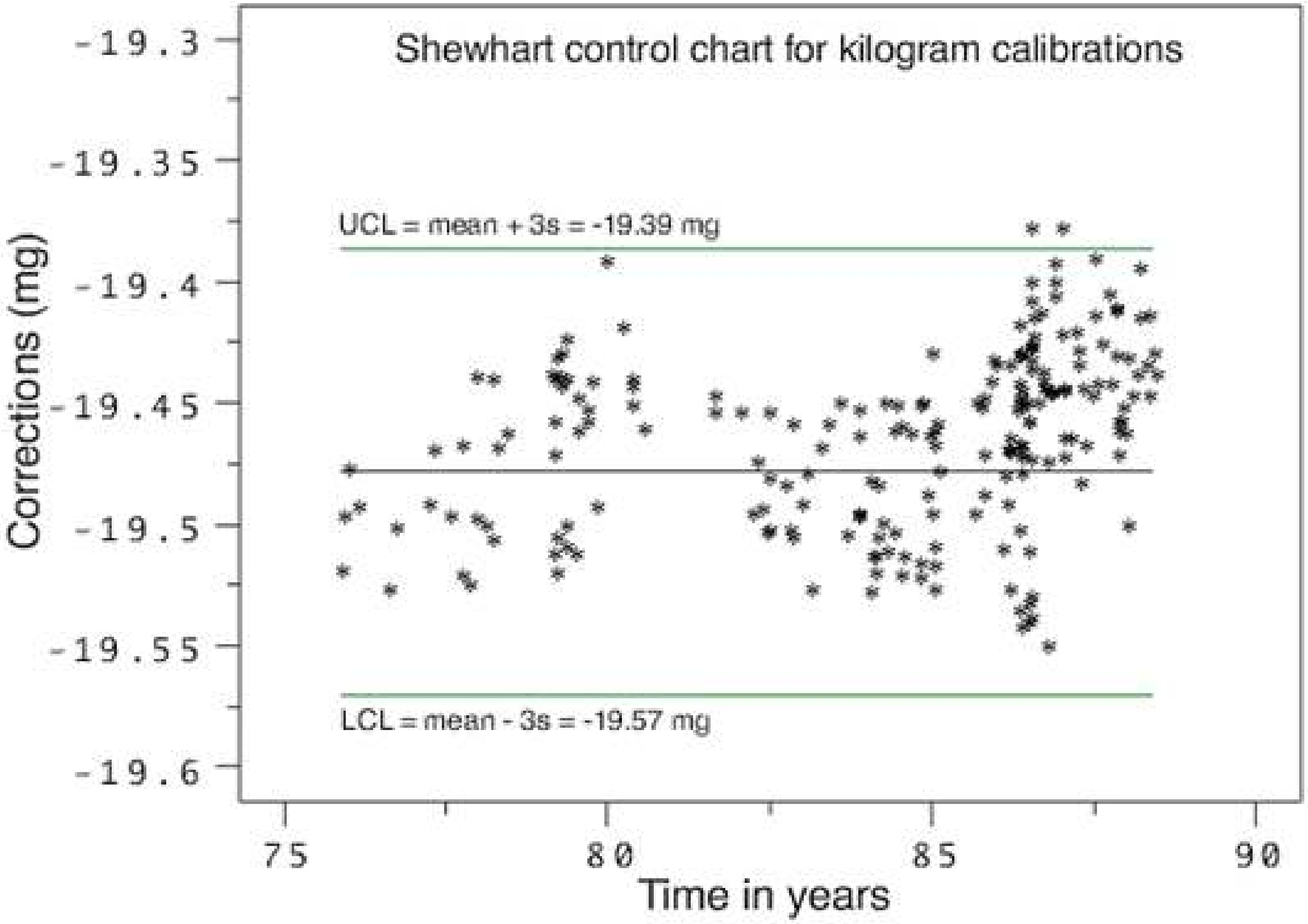
- Outlier – “departure from the expected”
- Types of outliers – defining “expected”
- Many approaches
  - Error bounds, tolerance limits – control charts
  - Model based – regression depth, analysis of residuals
  - Geometric
  - Distributional
  - Time Series outliers

Control Charts

- Quality control of production lots
- Typically univariate: X-Bar, R, CUSUM
- Distributional assumptions for charts not based on means e.g. R–charts
- Main steps (based on statistical inference)
  - Define “expected” and “departure” e.g. Mean and standard error based on sampling distribution of sample mean (aggregate);
  - Compute aggregate each sample
  - Plot aggregates vs expected and error bounds
  - “Out of Control” if aggregates fall outside bounds

An Example

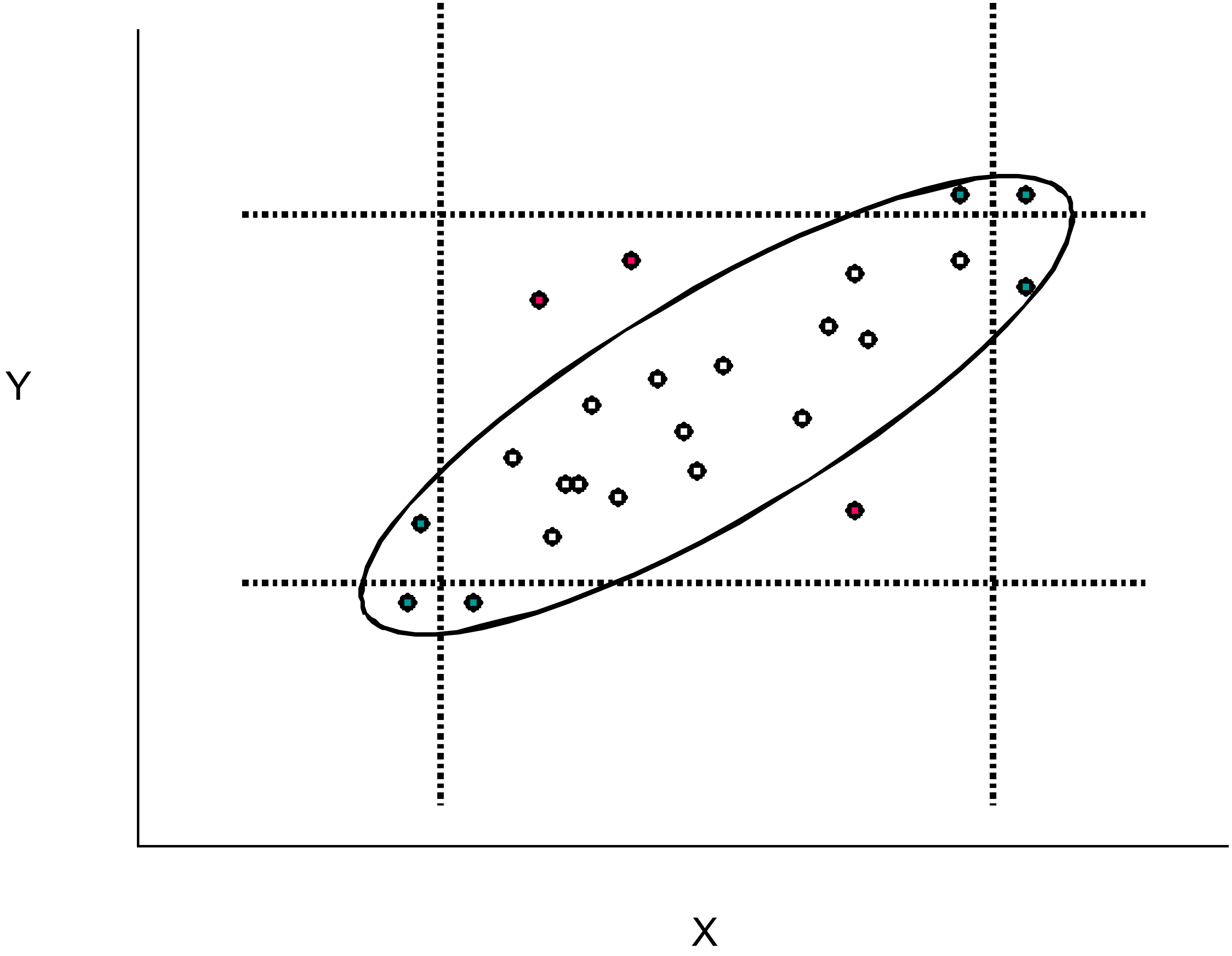
(http://www.itl.nist.gov/div898/handbook/mpc/section3/mpc3521.htm)



Multivariate Control Charts - 1

- Bivariate charts:
  - based on bivariate Normal assumptions
  - component-wise limits lead to Type I, II errors
- Depth based control charts (nonparametric):
  - map n-dimensional data to one dimension using depth e.g. Mahalanobis
  - Build control charts for depth
  - Compare against benchmark using depth e.g. Q-Q plots of depth of each data set

Bivariate Control Chart



Multivariate Control Charts - 2

- Multiscale process control with wavelets:
  - Detects abnormalities at multiple scales as large wavelet coefficients.
  - Useful for data with heteroscedasticity
  - Applied in chemical process control

Model Fitting and Outliers

- Models summarize general trends in data
  - more complex than simple aggregates
  - e.g. linear regression, logistic regression focus on attribute relationships
- Data points that do not conform to well fitting models are *potential outliers*
- Goodness of fit tests (DQ for analysis/mining)
  - check suitableness of model to data
  - verify validity of assumptions
  - data rich enough to answer analysis/business question?

Set Comparison and Outlier Detection

- "Model" consists of partition based summaries
- Perform nonparametric statistical tests for a rapid section-wise comparison of two or more massive data sets
- If there exists a baseline "good'' data set, this technique can detect potentially corrupt sections in the test data set

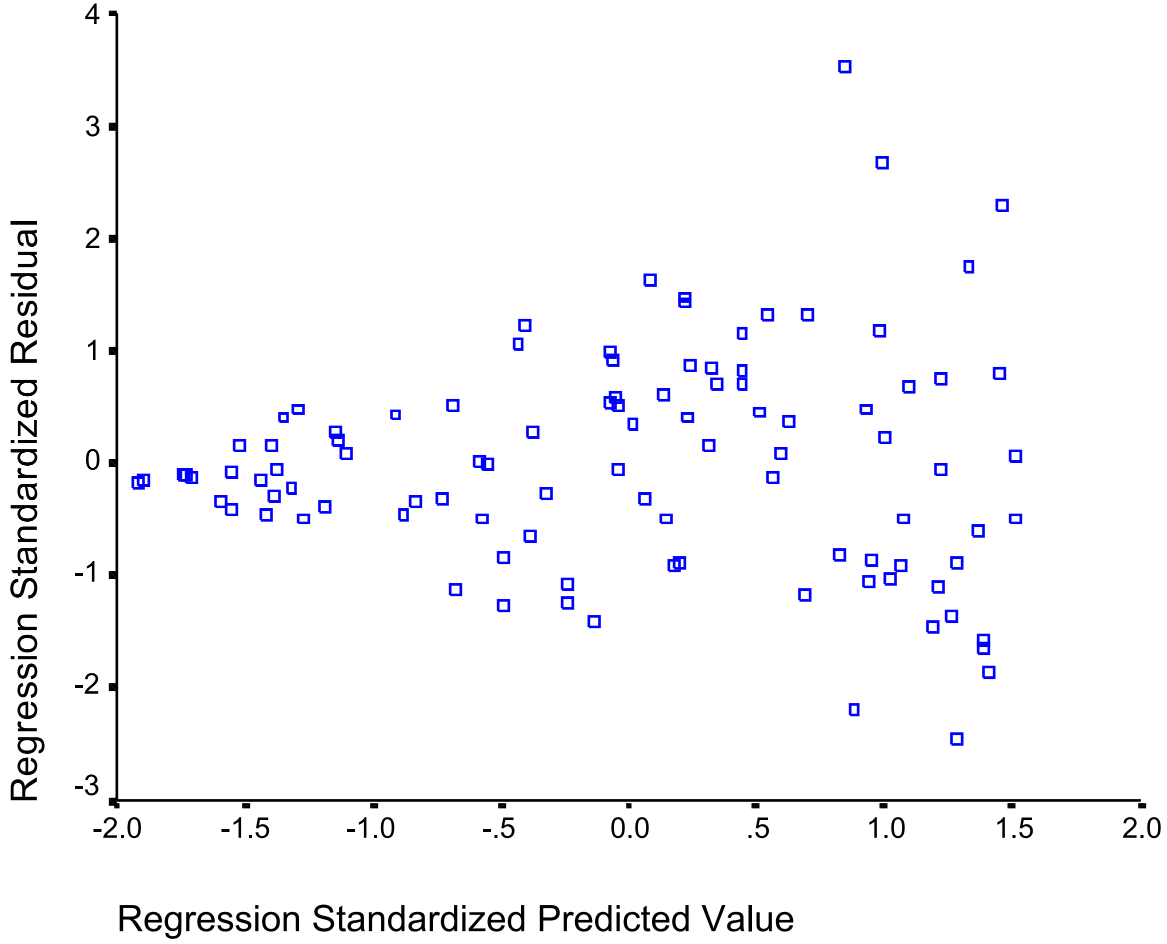
Goodness of Fit - 1

- Chi-square test
  - Are the attributes independent?
  - Does the observed (discrete) distribution match the assumed distribution?
- Tests for Normality
- Q-Q plots (visual)
- Kolmogorov-Smirnov test
- Kullback-Liebler divergence

Goodness of Fit - 2

- Analysis of residuals
  - Departure of individual points from model
  - Patterns in residuals reveal inadequacies of model or violations of assumptions
  - Reveals bias (data are non-linear) and peculiarities in data (variance of one attribute is a function of other attributes)
  - Residual plots

Detecting heteroscedasticity



http://www.socstats.soton.ac.uk/courses/st207307/lecture_slides/l4.doc

Goodness of Fit -3

- Regression depth
  - measures the “outlyingness” of a model, not an individual data point
  - indicates how well a regression plane represents the data
  - If a regression plane needs to pass through many points to rotate to the vertical (non-fit) position, it has high regression depth

Geometric Outliers

- Define outliers as those points at the periphery of the data set.
- *Peeling* : define layers of increasing depth, outer layers contain the outlying points
  - *Convex Hull:* peel off successive convex hull points.
  - *Depth Contours*: layers are the *data depth* layers.
- Efficient algorithms for 2-D, 3-D.
- Computational complexity increases rapidly with dimension.
  - $\Omega(N^{ceil(d/2)})$ complexity for N points, d dimensions

Distributional Outliers

- For each point, compute the maximum distance to its k nearest neighbors.
  - *DB(p,D)-outlier* : at least fraction *p* of the points in the database lie at distance greater than *D*.
- Fast algorithms
  - One is $O(dN^2)$, one is $O(c^d+N)$
- *Local Outliers* : adjust definition of outlier based on density of nearest data clusters.

Time Series Outliers

- Data is a time series of measurements of a large collection of entities (e.g. customer usage).
- Vector of measurements define a trajectory for an entity.
- A trajectory can be glitched, and it can make make radical but valid changes.
- Approach: develop models based on entity's past behavior (*within*) and all entity behavior (*relative*).
- Find potential glitches:
  - Common glitch trajectories
  - Deviations from within and relative behavior.

Database Tools

- Most DBMS’s provide many data consistency tools
  - Transactions
  - Data types
  - Domains (restricted set of field values)
  - Constraints
    - Column Constraints
      - Not Null, Unique, Restriction of values
    - Table constraints
      - Primary and foreign key constraints
  - Powerful query language
  - Triggers
  - Timestamps, temporal DBMS

Then why is every DB dirty?

- Consistency constraints are often not used
  - Cost of enforcing the constraint
    - E.g., foreign key constraints, triggers.
  - Loss of flexibility
  - Constraints not understood
    - E.g., large, complex databases with rapidly changing requirements
  - DBA does not know / does not care.
- Garbage in
  - Merged, federated, web-scraped DBs.
- Undetectable problems
  - Incorrect values, missing data
- Metadata not maintained
- Database is too complex to understand

Too complex to understand …

- Recall Lecture 2 : ER diagrams
  - Modeling even toy problems gets complicated
- Unintended consequences
  - Best example: cascading deletes to enforce participation constraints
    - Consider salesforce table and sales table. Participation constraint of salesforce in sales. Then you fire a salesman …
- Real life is complicated. Hard to anticipate special situations
  - Textbook example of functional dependencies: zip code determines state. Except for a few zip codes in sparsely populated regions that straddle states.

Tools

- Extraction, Transformation, Loading
- Approximate joins
- Duplicate finding
- Database exploration

Data Loading

- Extraction, Transformation, Loading (ETL)
- The data might be derived from a questionable source.
  - Federated database, Merged databases
  - Text files, log records
  - Web scraping
- The source database might admit a limited set of queries
- The data might need restructuring
  - Field value transformation
  - Transform tables (e.g. denormalize, pivot, fold)

(example of pivot)

unpivot

Customer	Part	Sales
Bob	bolt	32
Bob	nail	112
Bob	rivet	44
Sue	glue	12
Sue	nail	8
Pete	bolt	421
Pete	glue	6

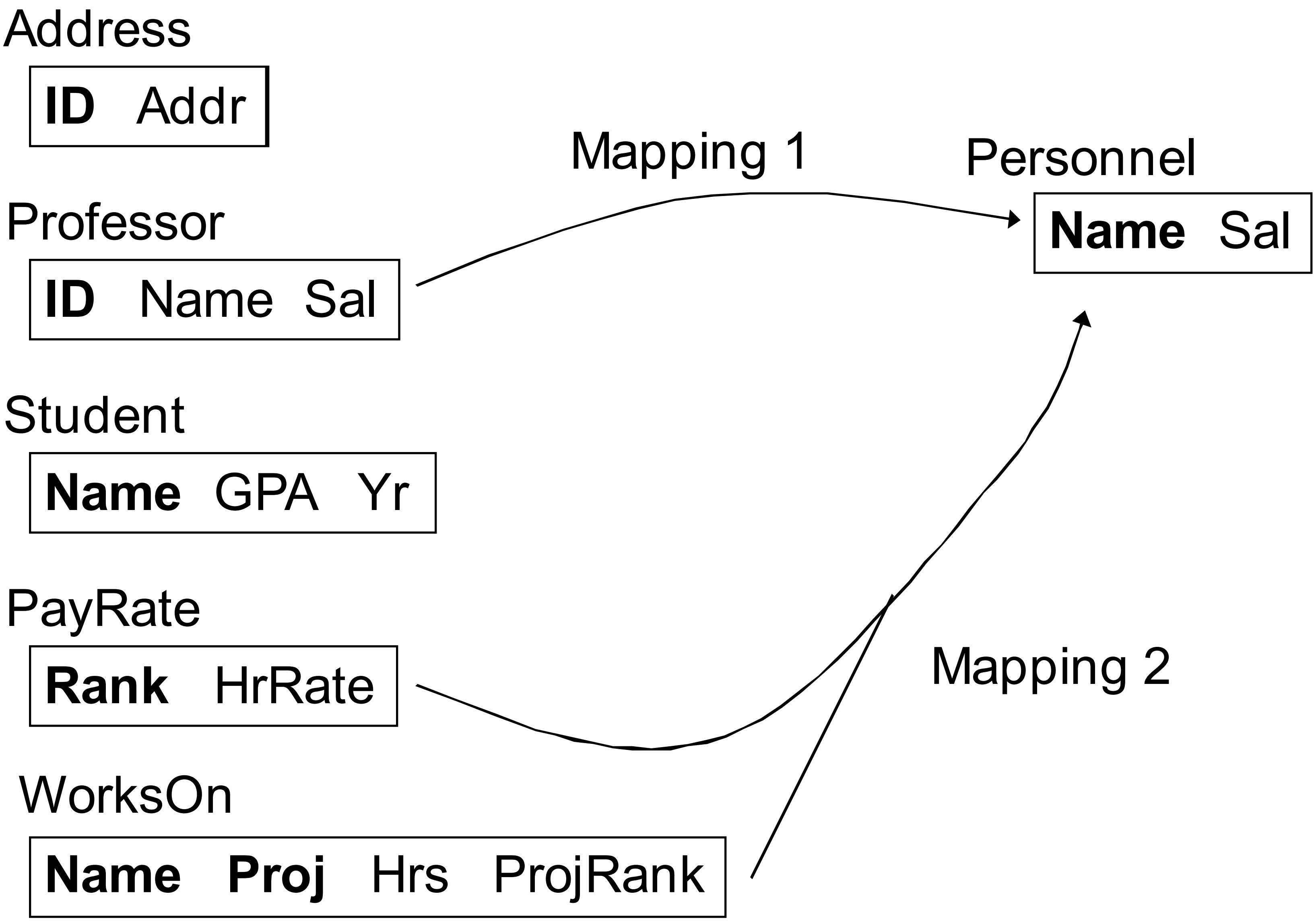
pivot

Customer	bolt	nail	rivet	glue
Bob	32	112	44	0
Sue	0	8	0	12
Pete	421	0	0	6

ETL

- Provides tools to
  - Access data (DB drivers, web page fetch, parse tools)
  - Validate data (ensure constraints)
  - Transform data (e.g. addresses, phone numbers)
  - Load data
- Design automation
  - Schema mapping
  - Queries to data sets with limited query interfaces (web queries)

(Example of schema mapping [MHH00])



Address

ID Addr

Professor

ID Name Sal

Student

Name GPA Yr

PayRate

Rank HrRate

WorksOn

Name **Proj** Hrs ProjRank

Personnel

Name Sal

Web Scraping

- Lots of data in the web, but its mixed up with a lot of junk.
- Problems:
  - Limited query interfaces
    - Fill in forms
  - “Free text” fields
    - E.g. addresses
  - Inconsistent output
    - I.e., html tags which mark interesting fields might be different on different pages.
  - Rapid change without notice.

Tools

- Automated generation of web scrapers
  - Excel will load html tables
- Automatic translation of queries
  - Given a description of allowable queries on a particular source
- Monitor results to detect quality deterioration
- Extraction of data from free-form text
  - E.g. addresses, names, phone numbers
  - Auto-detect field domain

Approximate Matching

- Relate tuples whose fields are “close”
  - Approximate string matching
    - Generally, based on edit distance.
    - Fast SQL expression using a *q-gram* index
  - Approximate tree matching
    - For XML
    - Much more expensive than string matching
    - Recent research in fast approximations
  - Feature vector matching
    - Similarity search
    - Many techniques discussed in the data mining literature.
  - Ad-hoc matching
    - Look for a clever trick.

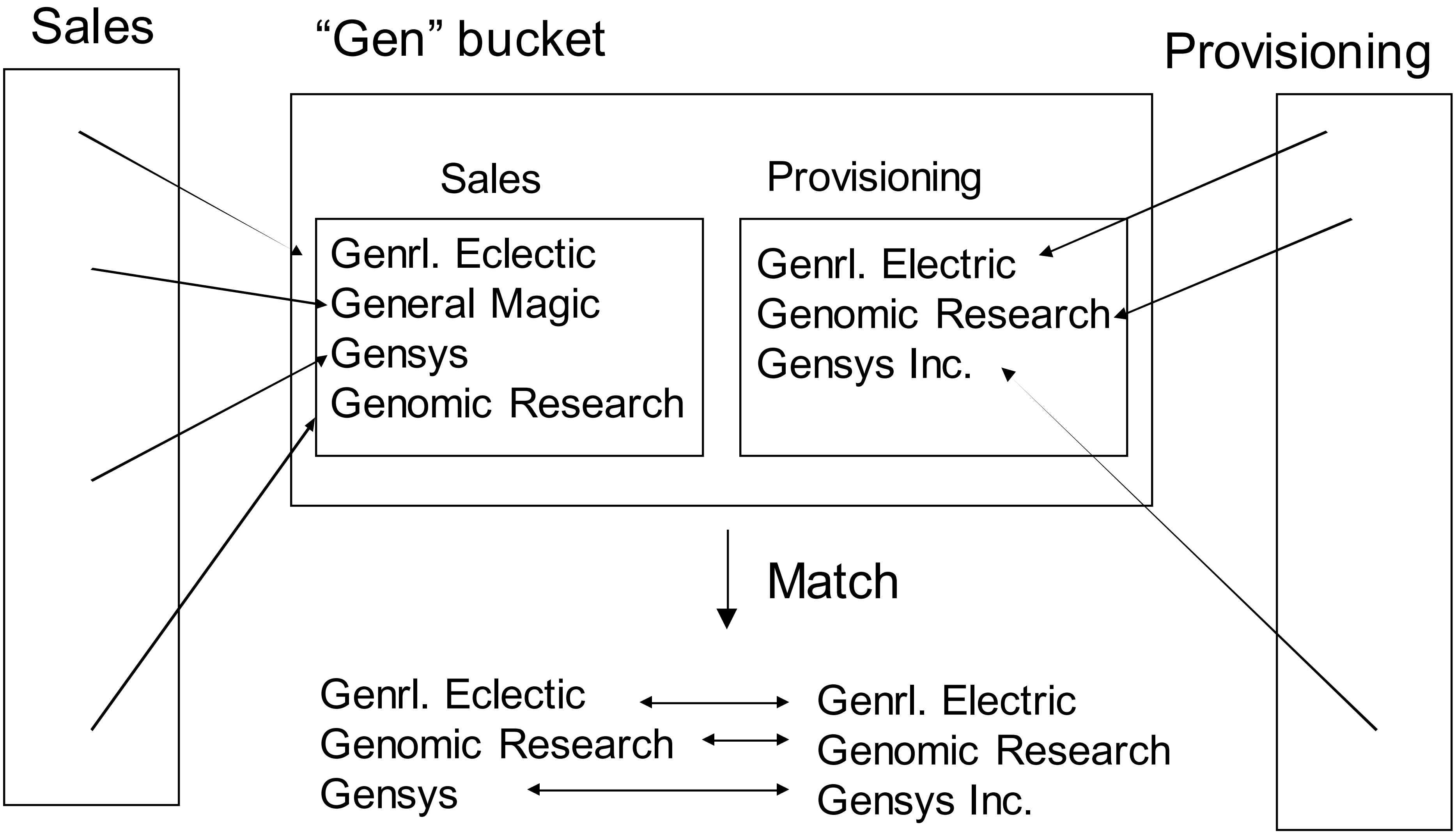
Approximate Joins and Duplicate Elimination

- Perform joins based on incomplete or corrupted information.
  - Approximate join : between two different tables
  - Duplicate elimination : within the same table
- More general than approximate matching.
  - **Semantics** : Need to use special transforms and scoring functions.
  - **Correlating information** : verification from other sources, e.g. usage correlates with billing.
  - **Missing data** : Need to use several orthogonal search and scoring criteria.
- But approximate matching is a valuable tool …

Algorithm

- Partition data set
  - By hash on computed key
  - By sort order on computed key
  - By similarity search / approximate match on computed key
- Perform scoring within the partition
  - Hash : all pairs
  - Sort order, similarity search : target record to retrieved records
- Record pairs with high scores are matches
- Use multiple computed keys / hash functions
- Duplicate elimination : duplicate records form an equivalence class.

(Approximate Join Example)



Sales

“Gen” bucket

Provisioning

Sales

Genrl. Eclectic
General Magic
Gensys
Genomic Research

Provisioning

Genrl. Electric
Genomic Research
Gensys Inc.

Match

Genrl. Eclectic ↔ Genrl. Electric
Genomic Research ↔ Genomic Research
Gensys ↔ Gensys Inc.

Database Exploration

- Tools for finding problems in a database
  - Opposite of ETL
  - Similar to data quality mining
- Simple queries are effective:

```
Select Field, count(*) as Cnt
from Table
Group by Field
Order by Cnt Desc
```

  - Hidden NULL values at the head of the list, typos at the end of the list
- Just look at a sample of the data in the table.

Database Profiling

- Systematically collect summaries of the data in the database
  - Number of rows in each table
  - Number of unique, null values of each field
  - Skewness of distribution of field values
  - Data type, length of the field
    - Use free-text field extraction to guess field types (address, name, zip code, etc.)
  - Functional dependencies, keys
  - Join paths
- Does the database contain what you think it contains?
  - Usually not.

Finding Keys and Functional Dependencies

- **Key**: set of fields whose value is unique in every row
- **Functional Dependency:** A set of fields which determine the value of another field
  - E.g., ZipCode determines the value of State
    - But not really …
- Problems: keys not identified, uniqueness not enforced, hidden keys and functional dependencies.
- Key finding is expensive: O(f^k) **`Count Distinct`** queries to find all keys of up to k fields.
- Fortunately, we can prune a lot of this search space if we search only for *minimal* keys and FDs
- **Approximate keys** : almost but not quite unique.
- **Approximate FD** : similar idea

Effective Algorithm

- Eliminate “bad” fields
  - Float data type, mostly NULL, etc.
- Collect an in-memory sample
  - Perhaps storing a hash of the field value
- Compute *count distinct* on the sample
  - High count : verify by *count distinct* on database table.
- Use *Tane* style level-wise pruning
- Stop after examining 3-way or 4-way keys
  - False keys with enough attributes.

Finding Join Paths

- How do I correlate this information?
- In large databases, hundreds of tables, thousands of fields.
- Our experience: field names are *very* unreliable.
  - Natural join does not exist outside the laboratory.
- Use data types and field characterization to narrow the search space.

Min Hash Sampling

- Special type of sampling which can estimate the *resemblance* of two sets
  - Size of intersection / size of union
- Apply to set of values in a field, store the min hash sample in a database
  - Use an SQL query to find all fields with high resemblance to a given field
  - Small sample sizes suffice.
- Problem: fields which join after a small data transformation
  - E.g "SS123-45-6789" vs. "123-45-6789"
- Solution: collect min hash samples on the *qgrams* of a field
  - Alternative: collect *sketches* of qgram frequency vectors

Domain Expertise

- *Data quality gurus*: “We found these peculiar records in your database after running sophisticated algorithms!”

  Domain Experts: “Oh, those apples - we put them in the same baskets as oranges because there are too few apples to bother. Not a big deal. We knew that already.”

Why Domain Expertise?

- DE is important for understanding the data, the problem and interpreting the results
  - “The counter resets to 0 if the number of calls exceeds N”.
  - “The missing values are represented by 0, but the default billed amount is 0 too.”
- Insufficient DE is a primary cause of poor DQ – data are unusable
- DE should be documented as metadata

Where is the Domain Expertise?

- Usually in people’s heads – seldom documented
- Fragmented across organizations
  - Often experts don’t agree. Force consensus.
- Lost during personnel and project transitions
- If undocumented, deteriorates and becomes fuzzy over time

Metadata

- Data about the data
- Data types, domains, and constraints help, but are often not enough
- Interpretation of values
  - Scale, units of measurement, meaning of labels
- Interpretation of tables
  - Frequency of refresh, associations, view definitions
- Most work done for scientific databases
  - Metadata can include programs for interpreting the data set.

XML

- Data interchange format, based on SGML
- Tree structured
  - Multiple field values, complex structure, etc.
- “Self-describing” : schema is part of the record
  - Field attributes
- DTD : minimal schema in an XML record.

```
<tutorial>
 <title> Data Quality and Data Cleaning: An Overview <\title>
 <Conference area=“database”> SIGMOD <\Conference>
 <author> T. Dasu
           <bio>  Statistician <\bio> <\author>
 <author> T. Johnson
           <institution> AT&T Labs <\institution> <\author>
 <\tutorial>
```

What’s Missing?

- Most metadata relates to static properties
  - Database schema
  - Field interpretation
- Data use and interpretation requires *dynamic* properties as well
  - What is the business process?
  - 80-20 rule

Lineage Tracing

- Record the processing used to create data
  - Coarse grained: record processing of a table
  - Fine grained: record processing of a record
- Record graph of data transformation steps.
- Used for analysis, debugging, feedback loops

Case Study

Case Study

- Provisioning inventory database
  - Identify equipment needed to satisfy customer order.
    - False positives : provisioning delay
    - False negatives : decline the order, purchase unnecessary equipment
- The initiative
  - Validate the corporate inventory
  - Build a database of record.
  - Has top management support.

Task Description

- OPED : operations database
  - Components available in each local warehouse
- IOWA : information warehouse
  - Machine descriptions, owner descriptions
- SAPDB : Sales and provisioning database
  - Used by sales force when dealing with clients.
- Data flow

  OPED → IOWA → SAPDB

Data Audits

- Analyze databases and data flow to verify metadata / find problems
  - Documented metadata was insufficient
    - OPED.warehouseid is corrupted, workaround process used
    - 70 machine types in OPED, only 10 defined.
    - SAPDB contains only 15% of the records in OPED or IOWA
  - "Satellite" databases at local sites not integrated with main databases
  - Numerous workaround processes.

Data Improvements

- Satellite databases integrated into main databases.
- Address mismatches cleaned up.
  - And so was the process which caused the mismatches
- Static and dynamic data constraints defined.
  - Automated auditing process
  - Regular checks and cleanups

What did we learn?

- Take nothing for granted
  - Metadata is always wrong, every bad thing happens.
- Manual entry and intervention causes problems
  - Automate processes.
  - Remove the need for manual intervention.
    - Make the regular process reflect practice.
- Defining data quality metrics is key
  - Defines and measures the problem.
  - Creates metadata.
- Organization-wide data quality
  - Data steward for the end-to-end process.
  - Data publishing to establish feedback loops.

Research Directions

Challenges in Data Quality

- Multifaceted nature
  - Problems are introduced at all stages of the process.
    - but especially at organization boundaries.
  - Many types of data and applications.
- Highly complex and context-dependent
  - The processes and entities are complex.
  - Many problems in many forms.
- No silver bullet
  - Need an array of tools.
  - And the discipline to use them.

Data Quality Research

- Burning issues
  - Data quality mining
  - Advanced browsing / exploratory data mining
  - Reducing complexity
  - Data quality metrics

“Interesting” Data Quality Research

- Recent research that I think is interesting and important for an aspect of data quality.
- CAVEAT
  - This list is meant to be an example, it is not exhaustive.
  - It contains research that I’ve read recently.
  - I’m not listing many interesting papers, including yours.

Bellman

- T. Dasu, T. Johnson, S. Muthukrishnan, V. Shkapenyuk, *Mining database structure; or, how to build a data quality browser*, SIGMOD 2002 pg 240-251
- “Data quality” browser.
- Perform *profiling* on the database
  - Counts, keys, join paths, substring associations
- Use to explore large databases.
  - Extract missing metadata.

DBXplorer

- S. Agrawal, S. Chaudhuri, G. Das, *DBXplorer: A System for Keyword-Based Search over Relational Databases*, ICDE 2002.
- Keyword search in a relational database, independent of the schema.
- Pre-processing to build inverted list indices (profiling).
- Build join queries for multiple keyword search.

Potters Wheel

- V. Raman, J.M. Hellerstein, *Potter's Wheel: An Interactive Data Cleaning System*, VLDB 2001 pg. 381-390
- ETL tool, especially for web scraped data.
- Two interesting features:
  - Scalable spreadsheet : interactive view of the results of applying a data transformation.
  - Field domain determination
    - Apply domain patterns to fields, see which ones fit best.
    - Report exceptions.

OLAP Exploration

- S. Sarawagi, G. Sathe, *i^3: Intelligent, Interactive Investigation of OLAP data cubes*, SIGMOD 2000 pg. 589
- Suite of tools (operators) to automate the browsing of a data cube.
  - *Find “interesting” regions*

Data Quality Mining

Contaminated Data

- Pearson, R. (2001) “Data Mining in the Face of Contaminated and Incomplete Records”, tutorial at SDM 2002
- **Outliers in process modeling and identification** *Pearson, R.K.;* Control Systems Technology, IEEE Transactions on , Volume: 10 Issue: 1 , Jan 2002 Page(s): 55 -63
- Methods
  - identifying outliers (Hampel limits),
  - missing value imputation,
  - compare results of fixed analysis on similar data subsets
  - others

Data Quality Mining : Deviants

- H.V. Jagadish, N. Koudas, S. Muthukrishnan, *Mining Deviants in a Time Series Database*, VLDB 1999 102-112.
- *Deviants* : points in a time series which, when removed, yield best accuracy improvement in a histogram.
- Use deviants to find glitches in time series data.

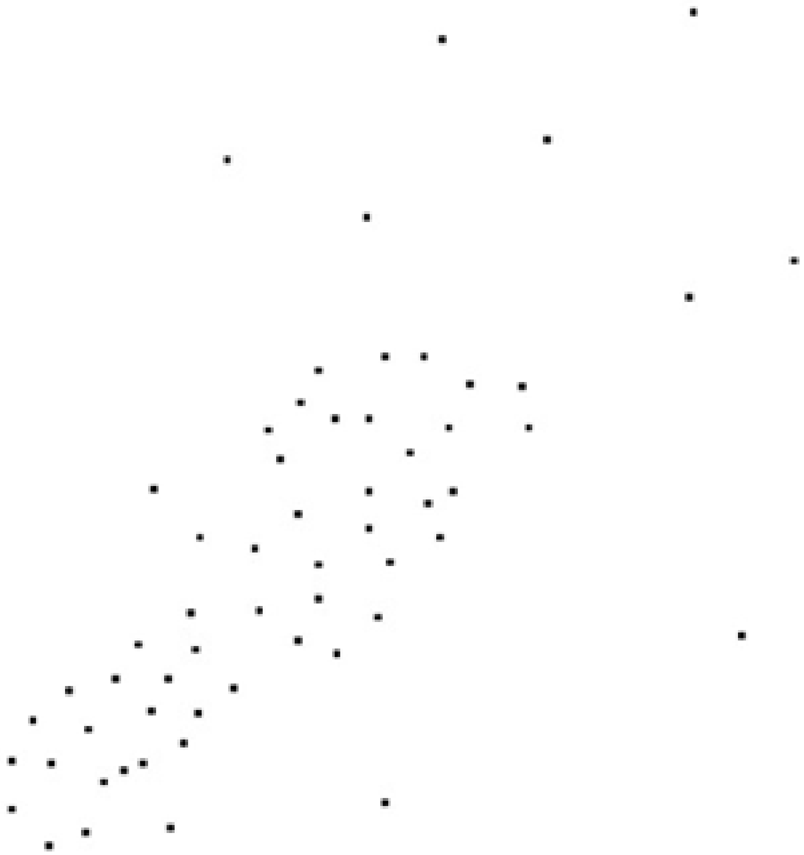
Data Quality Mining

- F. Korn, S. Muthukrishnan, Y. Zhu, *Monitoring Data Quality Problems in Network Databases*, VLDB 2003
- Define *probably approximately correct* constraints for a data feed (network performance data)
  - Range, smoothness, balance, functional dependence, unique keys
- Automation of constraint selection and threshold setting
- Raise alarm when constraints fail above tolerable level.

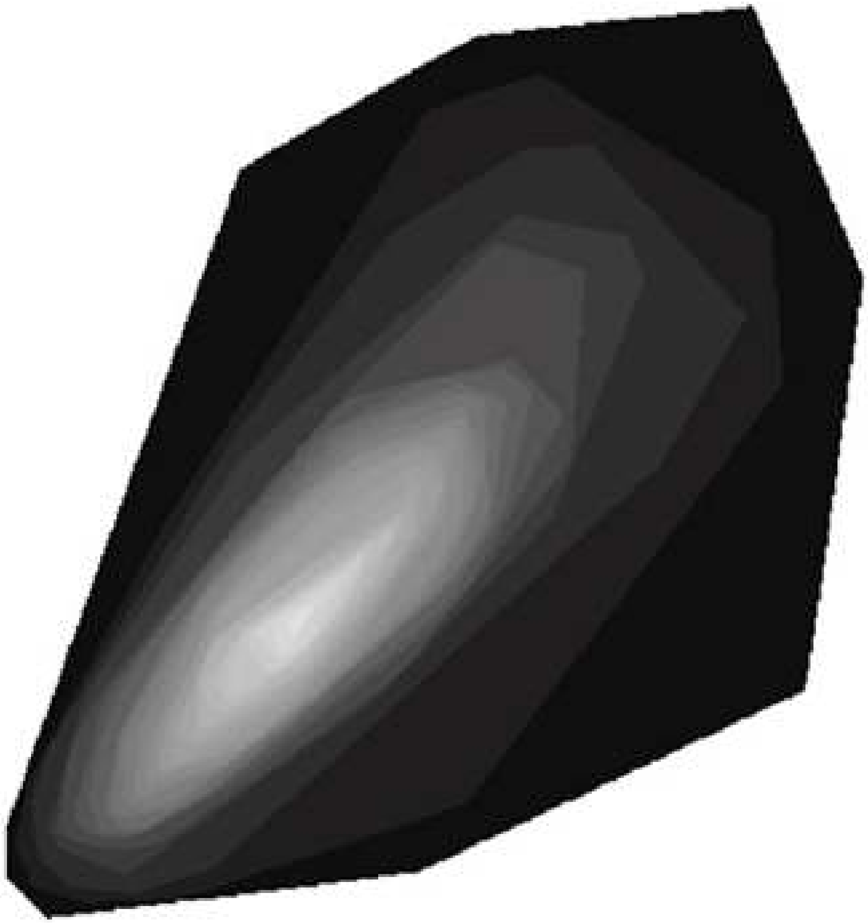
Data Quality Mining: Depth Contours

- S. Krishnan, N. Mustafa, S. Venkatasubramanian, *Hardware-Assisted Computation of Depth Contours*. SODA 2002 558-567.
- Parallel computation of *depth contours* using graphics card hardware.
  - Cheap parallel processor
  - Depth contours :
    - Multidimensional analog of the median
    - Used for nonparametric statistics

Points



Depth Contours



Approximate Matching

- L. Gravano, P.G. Ipeirotis, N. Koudas, D. Srivastava, *Text Joins in a RDBMS for Web Data Integration*, WWW2003
- Approximate string matching using IR techniques
  - Weight edit distance by inverse frequency of differing tokens (words or q-grams)
    - If "Corp." appears often, its presence or absence carries little weight. "IBM Corp." close to "IBM", far from "AT&T Corp."
- Define an SQL-queryable index

Exploratory Data Mining

- J.D. Becher, P. Berkhin, E. Freeman, *Automating Exploratory Data Analysis for Efficient Data Mining*, KDD 2000
- Use data mining and analysis tools to determine appropriate data models.
- In this paper, attribute selection for classification.

Exploratory Data Mining

- R.T. Ng, L.V.S. Lakshmanan, J. Han, A. Pang, *Exploratory Mining and Pruning Optimizations of Constrained Association Rules*, SIGMOD 1998 pg 13-24
- Interactive exploration of data mining (association rule) results through constraint specification.

Exploratory Schema Mapping

- M.A. Hernandez, R.J. Miller, L.M. Haas, *Clio: A Semi-Automatic Tool for Schema Mapping*, SIGMOD 2001
- Automatic generation and ranking of schema mapping queries
- Tool for suggesting field mappings
- Interactive display of alternate query results.

Conclusions

- Now that processing is cheap and access is easy, the big problem is data quality.
- Considerable research, but highly fragmented
- Lots of opportunities for applied research, once you understand the problem domain.

- Any questions?

Bibliography

Note: these references are an *introductory sample* of the literature.

References

- Process Management
  - http://web.mit.edu/tdqm/www/about.html
- Missing Value Imputation
  - Schafer, J. L. (1997), *Analysis of Incomplete Multivariate Data,* New York: Chapman and Hall
  - Little, R. J. A. and D. B. Rubin. 1987. "*Statistical Analysis with Missing Data*." New York: John Wiley & Sons.
  - Release 8.2 of SAS/STAT - PROCs MI, MIANALYZE
  - “Learning from incomplete data”. Z. Ghahramani and M. I. Jordan. AI Memo 1509, CBCL Paper 108, January 1995, 11 pages.

References

- Censoring / Truncation
  - Survival Analysis: Techniques for Censored and Truncated Data”. John P. Klein and Melvin L. Moeschberger
  - "Empirical Processes With Applications to Statistics”. Galen R. Shorack and Jon A. Wellner; Wiley, New York; 1986.
- Control Charts
  - A.J. Duncan, *Quality Control and Industrial Statistics.* Richard D. Irwin, Inc., Ill, 1974.
  - Liu, R. Y. and Singh, K. (1993). A quality index based on data depth and multivariate rank tests. J. Amer. Statist. Assoc. 88 252-260. 13
  - Aradhye, H. B., B. R. Bakshi, R. A. Strauss,and J. F. Davis (2001). Multiscale Statistical Process Control Using Wavelets - Theoretical Analysis and Properties. Technical Report, Ohio State University

References

- Set comparison
  - Theodore Johnson, Tamraparni Dasu: Comparing Massive High-Dimensional Data Sets. KDD 1998: 229-233
  - Venkatesh Ganti, Johannes Gehrke, Raghu Ramakrishnan: A Framework for Measuring Changes in Data Characteristics. PODS 1999, 126-137
- Goodness of fit
  - Computing location depth and regression depth in higher dimensions. Statistics and Computing 8:193-203. Rousseeuw P.J. and Struyf A. 1998.
  - Belsley, D.A., Kuh, E., and Welsch, R.E. (1980), *Regression Diagnostics*, New York: John Wiley and Sons, Inc.

References

- Geometric Outliers
  - Computational Geometry: An Introduction", Preparata, Shamos, *Springer-Verlag 1988*
  - "Fast Computation of 2-Dimensional Depth Contours", T. Johnson, I. Kwok, R. Ng, *Proc. Conf. Knowledge Discovery and Data Mining pg 224-228 1988*
- Distributional Outliers
  - "Algorithms for Mining Distance-Based Outliers in Large Datasets", E.M. Knorr, R. Ng, *Proc. VLDB Conf. 1998*
  - "LOF: Identifying Density-Based Local Outliers", M.M. Breunig, H.-P. Kriegel, R. Ng, J. Sander, *Proc. SIGMOD Conf. 2000*
- Time Series Outliers
  - "Hunting data glitches in massive time series data", T. Dasu, T. Johnson, MIT Workshop on Information Quality 2000.

References

- ETL
  - “Data Cleaning: Problems and Current Approaches”, E. Rahm, H.H. Do, *Data Engineering Bulletin 23(4) 3-13, 2000*
  - “Declarative Data Cleaning: Language, Model, and Algorithms”, H. Galhardas, D. Florescu, D. Shasha, E. Simon, C.-A. Saita, *Proc. VLDB Conf. 2001*
  - “Schema Mapping as Query Discovery”, R.J. Miller, L.M. Haas, M.A. Hernandez, *Proc. 26th VLDB Conf. Pg 77-88 2000*
  - “Answering Queries Using Views: A Survey”, A. Halevy, VLDB Journal, 2001
  - “A Foundation for Multi-dimensional Databases”, M. Gyssens, L.V.S. Lakshmanan, VLDB 1997 pg. 106-115
  - “SchemaSQL – An Extension to SQL for Multidatabase Interoperability”, L.V.S. Lakshmanan, F. Sadri, S.N. Subramanian, ACM Transactions on Database Systems 26(4) 476-519 2001
  - “Don't Scrap It, Wrap It! A Wrapper Architecture for Legacy Data Sources”, M.T. Roth, P.M. Schwarz, Proc. VLDB Conf. 266-275 1997
  - “Declarative Data Cleaning: Language, Model, and Algorithms
  - ”, H. Galhardas, D. Florescu, D. Shasha, E. Simon, C. Saita, *Proc. VLDB Conf. Pg 371-380 2001*

References

- Web Scraping
  - “Automatically Extracting Structure from Free Text Addresses”, V.R. Borkar, K. Deshmukh, S. Sarawagi, *Data Engineering Bulletin 23(4) 27-32, 2000*
  - “Potters Wheel: An Interactive Data Cleaning System”, V. Raman and J.M. Hellerstein, *Proc. VLDB 2001*
  - “Accurately and Reliably Extracting Data From the Web”, C.A. Knoblock, K. Lerman, S. Minton, I. Muslea, *Data Engineering Bulletin 23(4) 33-41, 2000*
- Approximate String Matching
  - “A Guided Tour to Approximate String Matching”, G. Navarro, *ACM Computer Surveys 33(1):31-88, 2001*
  - “Using q-grams in a DBMS for Approximate String Processing”, L. Gravano, P.G. Ipeirotis, H.V. Jagadish, N. Koudas, S. Muthukrishnan, L. Pietarinen, D. Srivastava, *Data Engineering Bulletin 24(4):28-37,2001.*

References

- Other Approximate Matching
  - “Approximate XML Joins”, N. Koudas, D. Srivastava, H.V. Jagadish, S. Guha, T. Yu, *SIGMOD 2002*
  - “Searching Multimedia Databases by Content”, C. Faloutsos, Klewer, 1996.
- Approximate Joins and Duplicate Detection
  - “The Merge/Purge Problem for Large Databases”, M. Hernandez, S. Stolfo, Proc. SIGMOD Conf pg 127-135 1995
  - “Real-World Data is Dirty: Data Cleansing and the Merge/Purge Problem”, M. Hernandez, S. Stolfo, *Data Mining and Knowledge Discovery 2(1)9-37, 1998*
  - “Telcordia’s Database Reconciliation and Data Quality Analysis Tool”, F. Caruso, M. Cochinwala, U. Ganapathy, G. Lalk, P. Missier, *Proc. VLDB Conf. Pg 615-618 2000*
  - “Hardening Soft Information Sources”, W.W. Cohen, H. Kautz, D. McAllester, *Proc. KDD Conf., 255-259 2000*

References

- Data Profiling

  - "Data Profiling and Mapping, The Essential First Step in Data Migration and Integration Projects", Evoke Software, *http://www.evokesoftware.com/pdf/wtpprDPM.pdf*
  - "TANE: An Efficient Algorithm for Discovering Functional and Approximate Dependencies", Y. Huhtala, J. K., P. Porkka, H. Toivonen, *The Computer Journal 42(2): 100-111 (1999)*
  - "Mining Database Structure; Or, How to Build a Data Quality Browser", T.Dasu, T. Johnson, S. Muthukrishnan, V. Shkapenyuk, Proc. SIGMOD Conf. 2002
  - "Data-Driven Understanding and Refinement of Schema Mappings", L.-L. Yan, R. Miller, L.M. Haas, R. Fagin, Proc. SIGMOD Conf. 2001

References

- Metadata
  - “A Metadata Resource to Promote Data Integration”, L. Seligman, A. Rosenthal, *IEEE Metadata Workshop, 1996*
  - “Using Semantic Values to Facilitate Interoperability Among Heterogenous Information Sources”, E. Sciore, M. Siegel, A. Rosenthal, *ACM Trans. On Database Systems 19(2) 255-190 1994*
  - “XML Data: From Research to Standards”, D. Florescu, J. Simeon, *VLDB 2000 Tutorial*, http://www-db.research.bell-labs.com/user/simeon/vldb2000.ppt
  - “XML’s Impact on Databases and Data Sharing”, A. Rosenthal, *IEEE Computer 59-67 2000*
  - “Lineage Tracing for General Data Warehouse Transformations”, Y. Cui, J. Widom, Proc. VLDB Conf. 471-480 2001