THE UNIVERSITY OF MEMPHIS

COMP7/8150: Fundamentals of Data Science

Week 2: Data Management

Max Garzon

Data Life Cycle

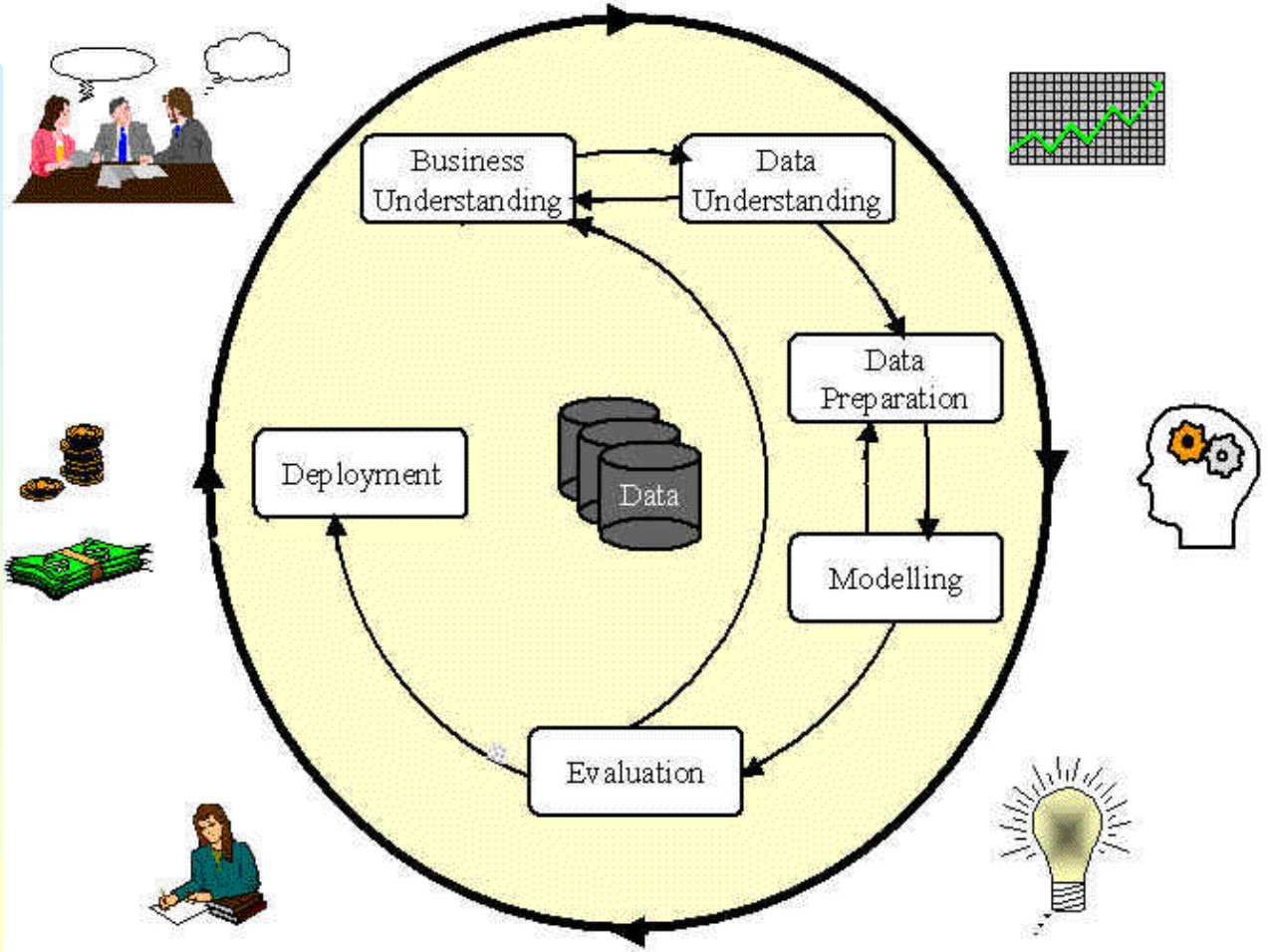
- **Problem Definition/goal**
  - Identify/specify goals of the data analysis
  - commit to specific deliverables
- **Data pre-processing**
  - Identify appropriate data
  - Acquire data (gather, lookup, understand)
- **Data processing**
  - Identify methods (gather, cleanse, store)
  - Carry out the analysis (patterns, trends, predictions?)
- **Data post-processing**
  - Visualize and present
  - Deploy and evaluate. Iterate, if necessary

Learning Objectives

- **To characterize main objectives in pre-processing**
  - **Getting the data:**
    gathering, cleansing and storing
  - **understanding the data:**
    presentation/visualization

- **To identify and learn to use main techniques/tools to achieve them:**
  - **Data bases (storage)**
  - **Presenting (descriptive statistics, plots) and visualizing data (cognitive load, HCI concepts):**
    `R and Python`
  - **Understanding the data**
    (chunking, clustering, modeling)

Standard Methodology



Pre-processing in Data Life Cycle

Business Understanding:

- State Business Goal(s) (BI)
- Translate into Analytic requirements
- Specify success criteria consistent with BI […]
- Deploy/use results of data analyses

Data Understanding

- Gather data (appropriate? How?)
- Explore the data and verify the quality: ok? (Relevant? Accurate? Sufficient? Complete?)
- Cleanse the data (find outliers, incomplete data)
- Visualize/understand Data (render graphically?)

[1] G.J. Myatt, W.P. Johnson (2014). **Making Sense of Data I: A Practical Guide to Exploratory Data Analysis and Data Mining**. John Wiley & Sons.

Why pre-processing?

"A company's most important asset is information. A corporation's ability to compete, adapt, and grow in a business climate of rapid change is dependent in large measure on how well the company uses information to make decisions … Sharing information that isn't clean and consolidated to the fullest extent can substantially reduce the effectiveness of a system of significant investment and considerable pay-off potential."

Stoker, 1999

How to understand your data corpus?

- A data corpus consists of a finite number of data tables, possible of various dimensions, comprising all the data for the project
- *A data point (or record)* is an *n*Dimensional vector of features/predictors/variables/attributes/columns:
  - Each value is alphabetic or numeric
  - Usually $n >> 1$ (n is the dimensionality of the point)
- Organize your data by clustering points of the same dim into data tables (attributes are column headers; data points/records are the rows/tuples)
- If at all possible, design an ontology (a criterion to organize the data by, e.g., SSN for US taxpayers) to be able to understand the whole corpus across tables

Storage Methods

- Databases
  - Digital files, Spreadsheets (excel)
  - Databases (hierarchical, relational, nonrelational)
    MySQL, Oracle, SAS, SysBase, …
- Datawarehouses/farms
  Central repositories of entire data corpora of an organization (usually very large) data
- Data Banks
  Centralized organizations storing all kinds of data for creation, public access/transaction
- Clouds
  Repositories of heterogeneous from multiple sources data and means to crunch it

.. Storage Methods (RDBs)

- Relational Databases
  Primary currency is a table containing records
  - primary key(s): handle(s) to "address" the records
  - attributes (columns),
  - Records (tuples) consisting of a vector of values for each attribute in the table
- RDBSs provide a backend to manage the tables / records (create, edit, delete, query)
- Structured Query Language (SQL) standardizes the queries across RDBMs

.. Storage Methods (RDBs)

- RDB Normalization

  Codd [2] introduced rules to organize RDBs and optimize data searching and processing:

  - First Normal Form (create an ontology/schema)
    * Eliminate repeating groups in individual tables
    * Create a separate table for each kind of related data
    * Identify each record in a table with a primary key as an atom
  - Second Normal Form = 1NF + (no redundancy)
    * primary attributes are independent of all others candidate keys
  - Third Normal Form = 2NF + ...
    * all the attributes in a table are determined only by the primary keys and not any other attributes

  In summary: any attribute depends on the (primary) key, the whole key and nothing but the key.

[2] E.F. Codd (1970). A Relational Model of Data for Large-Shared Data Banks. *Comm ACM* 13:6, 377-387.

.. Storage Methods (SQL: MySQL)

- Download/install MySQL from
  http://dev.mysql.com/downloads/
  (Make sure you enter and remember a pwd for root)
- Start it:
  Windows: > mysql –u root –p (from a RUN window)
  Mac: navigate to /usr/local/mysql/bin and enter
  > mysql –u root –p
  [Will need to enter userID/pwd you used to install it]
- Now you can play using SQL commands, e.g. to show all databases enter
  More (without the
  mysql >
  prompt) below

```
mysql> show databases;
+--------------------+
| Database           |
+--------------------+
| information_schema |
| menagerie          |
| mysql              |
| performance_schema |
| sys                |
+--------------------+
5 rows in set (0.04 sec)
```

.. Storage Methods (SQL: MySQL)

- TRUNCATE `Users`;
- `Users`; DROP TABLE IF EXISTS `Users`;
  CREATE TABLE `Users` (
  userID` mediumint(9) unsigned NOT NULL AUTO_INCREMENT,
  `userPwd` varchar(255) NOT NULL,
  `userAltPwd` varchar(255) NOT NULL,
  `userPrivileges` varchar(30) NOT NULL,
  `firstName` varchar(30) NOT NULL,
  `lastName` varchar(30) NOT NULL,
  `userRole` enum('PM','Developer','Tester','Designer') NOT NULL,
  PRIMARY KEY (`userID`)
) ENGINE=InnoDB DEFAULT CHARSET=latin1;
- SELECT * FROM Users WHERE userID='mgarzon' and userPwd= 'mymypwd' '';
- DROP TABLE IF EXISTS `SurveyQuestions;
  CREATE TABLE `SurveyQuestions` (
  `questionID` int(11) NOT NULL AUTO_INCREMENT,
  `questionText` varchar(255) DEFAULT NULL,
  `isActive` boolean DEFAULT TRUE,
  PRIMARY KEY (`questionID`)
) ENGINE=InnoDB
- INSERT INTO Users VALUES (100, 'myPwd', 'myOtherPwd', 'All', 'Max', 'Garzon', 'PM')
- SELECT * FROM Users WHERE userID='mgarzon' and userPwd= 'mymypwd' "
- DELETE FROM Users WHERE firstName = 'Max' and lastName = 'Garzon'

.. Why Preprocessing?

- Data in the real world is dirty, has a number of undesirable features:
  - Nonexistant
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
  - noisy: containing errors or outliers
  - inconsistent: containing discrepancies in codes or names
- Data analyses are only as good as data (Garbage in? Garbage out!)
  - Quality decisions can only come from quality data
  - Data warehouse needs consistent integration of quality data (ERP: Enterprise Resource Planning)
- A multi-dimensional measure of data quality:
  - A well-accepted multi-dimensional view:
    - Accuracy (external consistency with reality and internally), believability, accessibility, completeness, value added (welel organized, up-to-date, …)
  - Broad attributes:
    - intrinsic, contextual, representational, and accessible.

Data Gathering

- Data collection is very expensive (US 2010/20 census cost: $13/16B+)
  - Either get it from scratch or cleanse legacy data
- Data can be obtained from various sources
  - **Legacy data**
    census data, data banks (genebanks), organizations (Forrester)
    Databases, data warehouses
  - **Surveys**
    by others or your own
  - **Sensors**
    weather, health, surveillance cams
  - Others
    MoDaS: **Mob Data Sourcing** to develop a scientifically sound process of automatic crowd task generation, analysis, reasoning for data generation from the web.
- One of the most difficult but decisive steps
  - Quality decisions can only come from quality data
    - accurate, representational and accessible.

.. Preprocessing: Missing Data

- Data is not always available
  - For example, many points have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - unregistered history of or changes to the data
- Missing data may need to be inferred (e.g. by averages) or deleted altogether

.. Missing Data?

- Ignore the tuple: usually done when class label is missing (assuming the task is classification—not effective in certain cases)
- Fill in the missing value manually: tedious + infeasible?
- Use a global constant to fill in the missing value: e.g., “unknown”, a new class?
- Use the attribute mean to fill in the missing value
- Use the attribute mean for all samples of the same class to fill in the missing value (smarter)
- Use the most probable value to fill in the missing value: inference-based such as regression, Bayesian formula, decision trees, ...

.. Preprocessing: Noisy Data

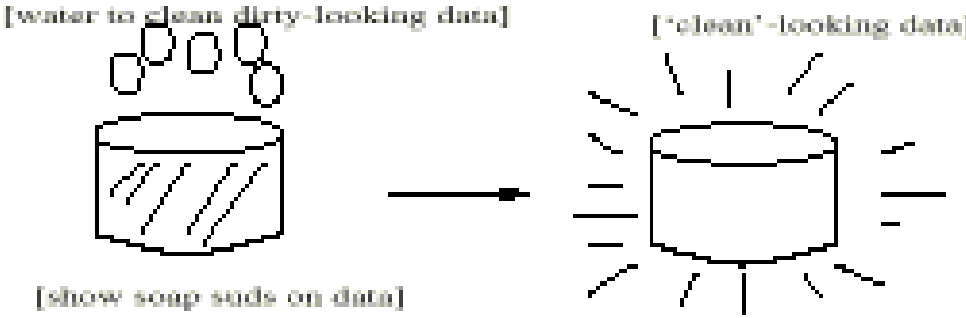
- Noise is a random error in a measurement
- Incorrect attribute values may be due to
  - faulty data collection instruments
  - data entry/transmission problems
  - technical limitations
  - inconsistencies in naming convention
- Other noise problems also require data cleansing:
  - duplicate records
  - incomplete data
  - inconsistent data

.. Preprocessing

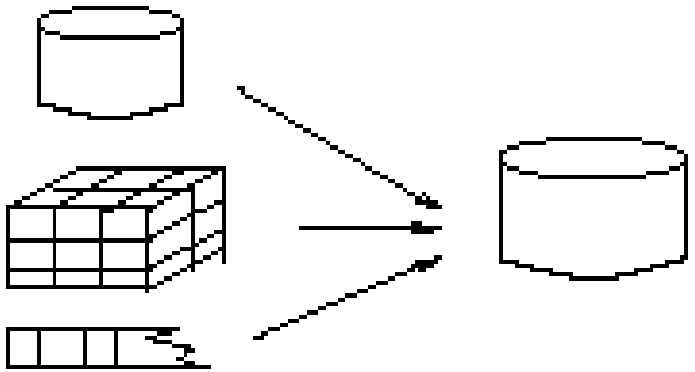
- Data preparation:
  - May take (and usually does) over 90% of the time
    - Collection
    - Assessment
    - Consolidation and Cleansing
      - table links, aggregation level, missing values, …
    - Data selection
      - active role in ignoring noncontributing data?
      - outliers?
      - *Use of samples*
      - visualization tools
    - Transformations - create new variables

Forms of data preprocessing

Data Cleaning



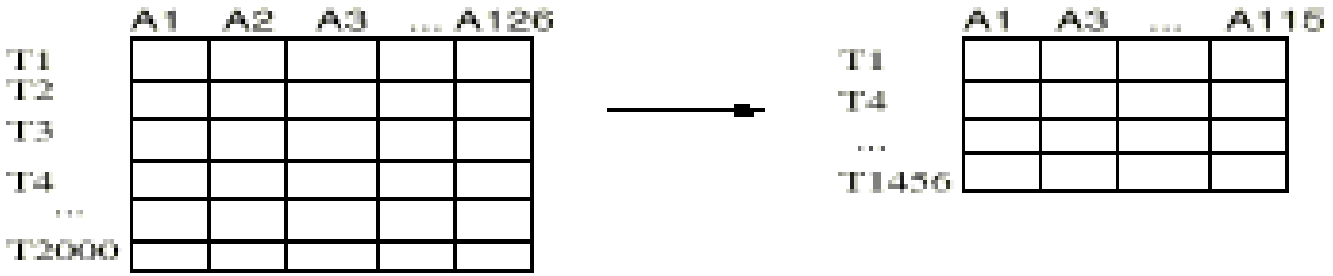
Data Integration



Data Transformation

-2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

Data Reduction



Major Tasks in Preprocessing

- Data cleansing
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data integration
  - Integration of multiple databases, data cubes, files, or notes
- Data transformation
  - Normalization (scaling to a specific standard range)
  - Aggregation (reduce whole clouds of data to representative value)
- Data reduction
  - Obtains reduced representation in volume but produces the same or similar analytical results
  - Data discretization: particularly important, especially for numerical data
  - Data aggregation, dimensionality reduction, data compression, generalization (rules)

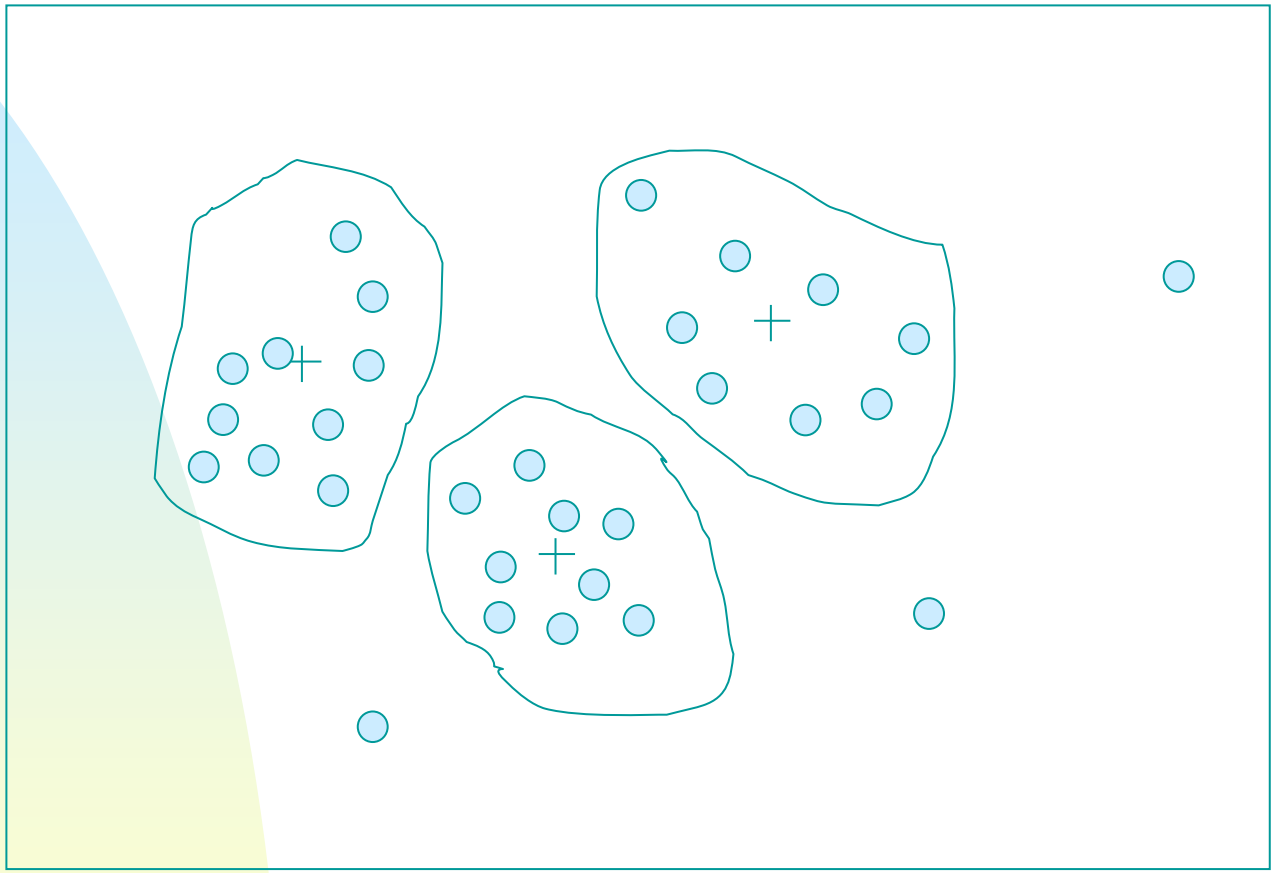
.. Preprocessing: Data Integration

- Data integration
  - combines data from multiple sources into a single coherent corpus
- Schema integration
  - integrate metadata from different sources
  - Entity identification problem: identify real world entities from multiple data sources, e.g.,
    A.cust-id ≡ B.cust-#
- Detecting/resolving data value conflicts
  - for the same real world entity, attribute values from different sources are different (due to different representations or scales, e.g., metric vs. British units, different currency)

.. Preprocessing: Discretize by Binning

- Equal-width (distance) partitioning:
  - most straightforward: it divides the range into *N* intervals of equal size: histograms, uniform grid
    if *A* and *B* are the lowest and highest values of the attribute, the width of intervals will be: *W* = (*B*-*A*)/*N*.
  - However, outliers may dominate presentation ...
  - and skewed data is not handled well.
- Equal-depth (frequency) partitioning:
  - It divides the range into *N* intervals, each containing approximately same number of samples
  - Good data scaling
  - Managing categorical attributes can be tricky.

.. Preprocessing: Clustering



.. Preprocessing: Redundancy

- Redundant data occur often when integrating multiple DBs
  - The same attribute may have different names in different databases
  - One explicit attribute may be a “derived” attribute in another table, e.g., annual revenue
- Redundant data may be detected by correlation analysis

$$r_{A,B} = \frac{\Sigma(A-\bar{A})(B-\bar{B})}{(n-1)\sigma_A\sigma_B}$$

- Careful integration can help reduce/avoid redundancies and inconsistencies and improve data mining speed and quality

.. Preprocessing: Data Transformations

- Smoothing: remove noise from data (by binning, clustering, regression)
- Aggregation: summarization, smoothing
- Generalization: concept hierarchy climbing
- Normalization: scaled to fall within a small, specified range
  - min-max normalization
  - *z*-score normalization
  - normalization by decimal scaling
- Attribute/feature construction
  - New attributes constructed from the given ones

.. Transformation: Normalization

- min-max normalization

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

- z-score normalization

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

- normalization by decimal scaling

$$v' = \frac{v}{10^j}$$, where j is the smallest integer such that Max(| * |)<1

v'

.. Cleansing and Data Quality

- Data is a product that can be characterized as either “quality” or “nonquality.” The ability to make quality decisions depends in part on the decision-maker’s ability to access quality data.
- Data cleansing is the process that insures that the same piece of information is referred to in only ONE way. When data is clean, its users can focus on its use and not its credibility.

Why is "Dirty" Data a Problem?

- Simply put, dirty data for data warehouses is the product of relying on data from legacy systems.
- But if company's have relied on this data for decades, why is it a problem today?
- Because a data warehouse "promises" to deliver "a single version of the truth." Unfortunately integrating data from different sources magnifies the problems.

Why is Legacy Data "Dirty" ?

- Dummy Values,
- Absence of Data,
- Multipurpose Fields,
- Cryptic Data,
- (self-)Contradicting Data,
- Inappropriate Use of Address Lines,
- Violation of Business Rules,
- Normalization issues (e.g.,no primary key)
- Integration Problems

To Cleanse or Not to Cleanse

- CAN the legacy data be cleansed?
- Sometimes the answer is “NO”
- Then, SHOULD it be cleansed?
- Again, sometimes “NO”
- Next, WHERE should it be cleansed?
- Finally, HOW should it be cleansed?

Steps in Cleansing Data

http://www.youtube.com/watch?v=06jBp-vfqqs

- Parsing
- Correcting
- Standardizing
- Matching
- Consolidating

.. Cleansing: Parsing

Parsing locates and identifies individual data elements in the source files and then isolates these data elements in the target files.

.. Cleansing: Parsing

Example:

Input Data from Source File
Beth Christine Parker, SLS MGR
Regional Port Authority
Federal Building
12800 Lake Calumet
Hedgewisch, IL

Parsed Data in Target File

First Name:	Beth
Middle Name:	Christine
Last Name:	Parker
Title:	SLS MGR
Firm:	Regional Port Authority
Location:	Federal Building
Number:	12800
Street:	Lake Calumet
City:	Hedgewisch
State:	IL

.. Cleansing: Correcting

May have to correct parsed individual data components using sophisticated data algorithms and secondary data sources.

.. Cleansing: Correcting

.. Example:

Parsed Data

First Name:	Beth
Middle Name:	Christine
Last Name:	Parker
Title:	SLS MGR
Firm:	Regional Port Authority
Location:	Federal Building
Number:	12800
Street:	Lake Calumet
City:	Hedgewisch
State:	IL

Corrected Data

First Name:	Beth
Middle Name:	Christine
Last Name:	Parker
Title:	SLS MGR
Firm:	Regional Port Authority
Location:	Federal Building
Number:	12800
Street:	South Butler Drive
City:	Chicago
State:	IL
Zip:	60633
Zip+Four:	2398

.. Cleansing: Standardizing

Standardizing applies conversion routines to transform data into its preferred (and consistent) format using both standard and custom business rules.

.. Cleansing: Standardizing

.. Example:

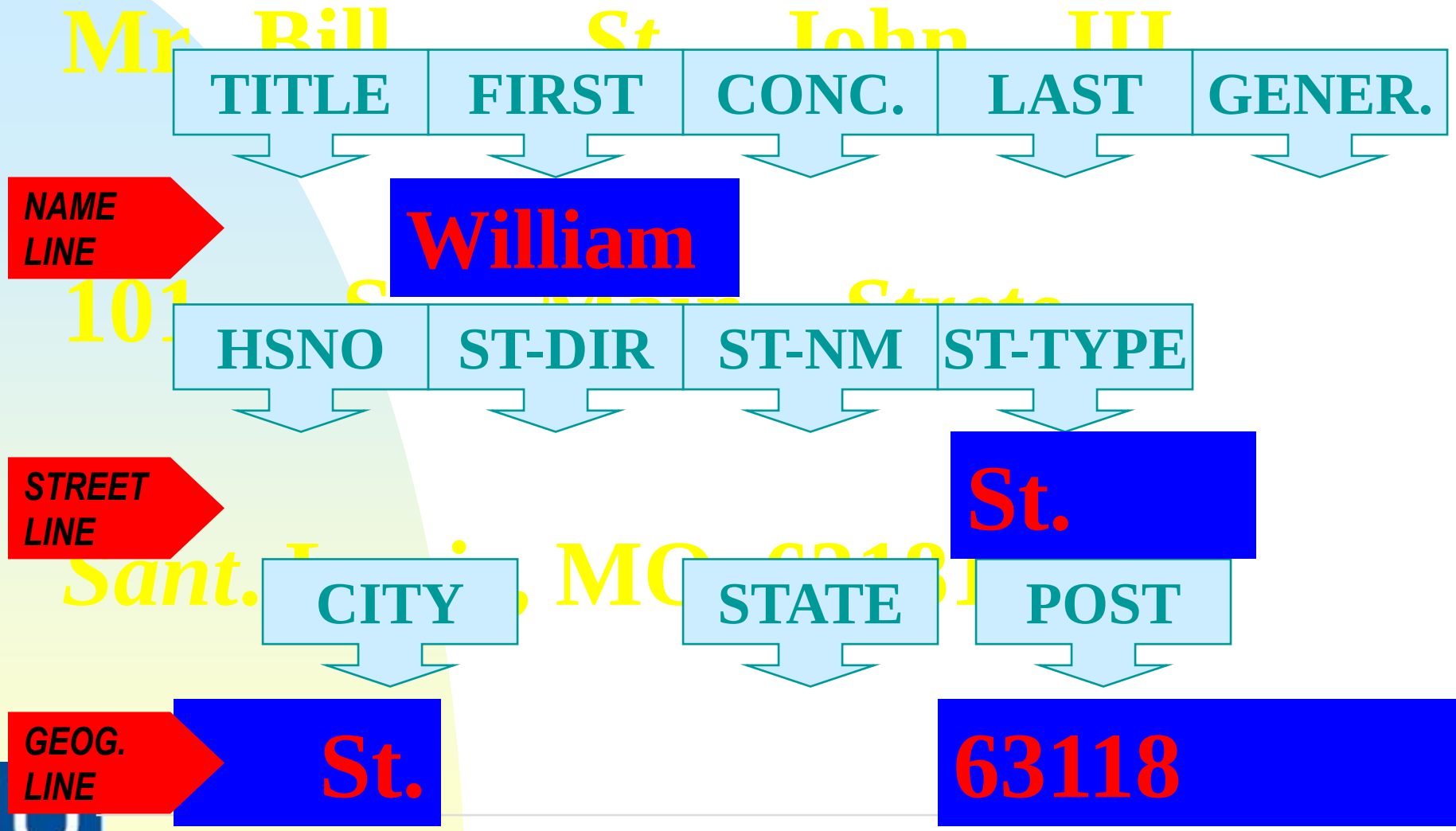
Corrected Data

First Name:	Beth
Middle Name:	Christine
Last Name:	Parker
Title:	SLS MGR
Firm:	Regional Port Authority
Location:	Federal Building
Number:	12800
Street:	South Butler Drive
City:	Chicago
State:	IL
Zip:	60633
Zip+Four:	2398

Corrected Data

Pre-name:	Ms.
First Name:	Beth
1st Name Match Standards:	Elizabeth, Bethany, Bethel
Middle Name:	Christine
Last Name:	Parker
Title:	Sales Mgr.
Firm:	Regional Port Authority
Location:	Federal Building
Number:	12800
Street:	S. Butler Dr.
City:	Chicago
State:	IL
Zip:	60633
Zip+Four:	2398

Parsing, Correcting, Standardizing



.. Cleansing: Matching

Searching and matching records within and across the parsed, corrected and standardized data based on predefined business rules to eliminate duplications.

.. Cleansing: Match Patterns

Business Name	Street	*Branch Type*	*Customer #/Tax ID*	City	*Vendor Code*	Pattern	Pattern I.D.
Exact	Exact	Exact	Exact	Exact	Exact	AAAAAA	P110
Exact	VClose	Exact	VClose	Exact	Blanks	ABAAA-	P115
Exact	VClose	Exact	Blanks	Exact	Exact	ABA-AA	P120
Exact	VClose	Close	Close	Exact	Exact	ABCCAA	S300
VClose	VClose	Exact	Close	Exact	Exact	BBACAA	S310

.. Preprocessing: Visualization

http://www.youtube.com/watch?v=pLqjQ55t
http://www.youtube.com/watch?v=qJ_mG_ucyY

Visualization
The use of computer-supported, interactive, visual representations of data to amplify cognition ("mental activity") and render data comprehensible.

Information Visualization
The use of computer-supported, interactive visual representations of abstract data to facilitate understanding of the data.

S. Card

.. Preprocessing: MDS

- Multi-dimensionality scaling enables to visualize 40 dimensional data on a 2D display (e.g. scatter plots of a few variables at a time)
- The idea is to keep distance relations between nodes, proportionally consistent as you reduce dimensions of the space
- iterative and very costly
- The distance in 40D space is *d*, then distance in 2D space should be λd , where λ is a constant for all elements

.. Preprocessing: nonmetric MDS

- Nonmetric type introduced by Kruskal
- defines a stress function to place data points in lower dimensional space

$$Stress = \frac{\sum_{i<j} (d_{ij} - g_{ij})^2}{\sum_{i<j} g_{ij}^2}$$

where d_{ij} is the distance in high dimensional space and g_{ij} is the distance in lowD space

- Points are displaced to lower stress and iterations are stopped when overall stress reaches below a certain threshold
- The distance function is generally the ordinary Euclidean distance

.. Preprocessing Tools

- Software packages
  R
  Python
  MathLab + hundreds of others
- Statistical Methods
  What is the distribution of the cloud of data?
- Associative Memory Methods (later)
  a topological (distance) map of grouping data by the most salient abstract relationships among the records
- Machine Learning Methods (later)
  Decision trees, SVMs, Neural Networks