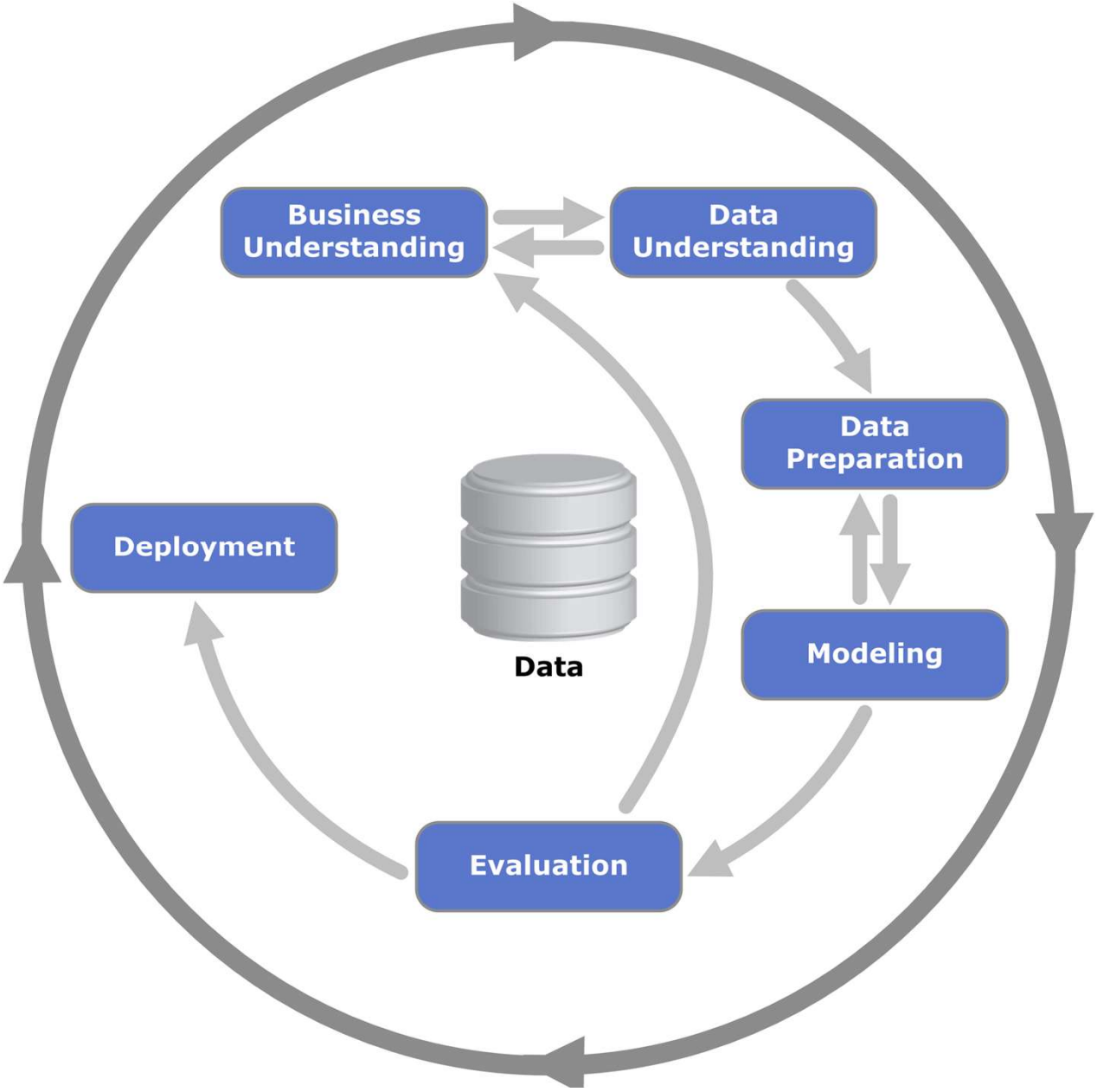
Statistical Data Mining
MATH 4720
Lecture 2
DATA UNDERSTANDING AND PREPARATION

Gaurav Gupta

GEH A401

Cross-industry standard process (CRISP-DM) for data mining



What kind of Data?

- **Relational database:** Database that stores and provides access to data points that are related to one another.
- **Data warehouse:** designed to enable and support business intelligence (BI) activities, especially analytics.
- **Transactional database:** it can reverse or scale back a database transaction or activity if it isn't performed correctly.
- **Data streaming:** Data streaming is the process of transmitting a continuous flow of data (also known as streams) typically fed into stream processing software to derive valuable insights.

- **Spatial data:** Spatial data comprise the relative geographic information about the earth and its features.
- **Spatiotemporal data:** A spatiotemporal database is a database that manages both space and time information.
- Multimedia data
- Text data
- WWW data
- Time-series data, temporal data, sequence data
- Structure data, graphs, social networks and multi-linked data

Definitions

- **Source data:** Information from any source in any format.
- **Analytical file:** A set of information items from (possibly) multiple sources; that information is composed into one row of information about some entity (e.g., a customer).
- **Record** : One row in the analytical file.
- **Attribute:** An item of data that describes the record in some way.

- **Variable:** An attribute installed into a column (field) of the entity record.
- **Target variable:** A variable in the entity record to be predicted by the model.
- **Predictor variable:** A variable in the entity record that is a candidate for inclusion in the model as a predictor of the target variable.
- **Numeric variable:** A variable with only numbers in it; it is treated as a number.
- **Categorical variable:** A variable with any character in it; the character may be a number, but it is treated as text.
- **Dummy variable:** A variable created for each member of the list of possible contents of a categorical variable (e.g., “red,” “green,” “blue”)
- **Surrogate variable:** A variable that has an effect on the target variable very similar to that of another variable in the record

Data Understating: Basic Issues

- How do I find the data I need for modeling?—Data Acquisition
- How do I integrate data I find in multiple disparate data sources?—Data Integration
- What do the data look like?—Data Description
- How clean is the data set?—Data Assessment

Data Preparation: Basic Issues

- How do I clean up the data?—Data Cleansing
- How do I express data variables?—Data Transformation
- How do I handle missing values?—Data Imputation
- Are all cases treated the same?—Data Weighting and Balancing
- What do I do about outliers and other unwanted data?—Data Filtering
- How do I handle temporal (time-series) data?—Data Abstraction
- Can I reduce the amount of data to use?—Data Reduction
- Records?—Data Sampling
- Variables?—Dimensionality Reduction
- Values? —Data Discretization
- Can I create some new variables?—Data Derivation

Data Understanding

1. Data Acquisition
2. Data Extraction
3. Data Description
4. Data Assessment
5. Data Profiling
6. Data Cleansing
7. Data Transformation
8. Data Imputation
9. Data Weighting and Balancing
10. Data Filtering and Smoothing
11. Data Abstraction
12. Data Reduction
13. Data Sampling
14. Data Discretization
15. Data Derivation

1. Data Acquisition

- Gaining access to data is not as easy
- Companies have portions of the data you need stored in different data “silos.”
- The separate data stores may exist in different departments, spreadsheets, miscellaneous databases, printed documents, and handwritten notes.
- The initial challenge is to identify where the data are and how you can get this information.

Common modes of access to business data

- **Query-based data :** SQL-99
- **High-level query languages:** Elaborations of SQL optimized for data mining include Modeling Query Language (MQL; Imielinski and Virmani, 1999) and Data Mining Query Language (DMQL; Han et al., 1996). It's attractive, but the high-level languages are not in standard use
- **Low-Level and ODBC Database Connections:**

NCR Teradata in Warehouse Miner: can access data directly and create descriptive statistical reports and some analytical modeling operations (e.g., Logistic Regression).

Some other data mining tools picked up on that concept to provide in-database access to data via ODBC or other proprietary low-level interfaces (SAS-Enterprise Miner, SPSS Clementine, and STATISTICA). This approach yields several benefits:

- Removes time and space constraints in moving large volumes of data.
- Helps keep management and provisioning of data centralized.
- Reduces unnecessary proliferation of data.
- Facilitates better data governance to satisfy compliance concerns.

2. Data Extraction

- Now you have your data in some form (flat-file format).
- How do you put all the pieces together?
- The challenge before you now is to create a combined data structure suitable for input to the data mining tool.

File #1: Name & Address

Name	Address	City	State	Zipcode
John Brown	1234 E St.	Chicago	IL	60610
Jean Blois	300 Day St.	Houston	TX	77091
Neal Smith	325 Clay St.	Portland	OR	97201

File #2: Product

Name	Address	Product	Sales Date
John Brown	1234 E. St.	Mower	1/3/2007
John Brown	1234 E. St.	Rake	4/16/2006
Neal Smith	325 Clay St.	Shovel	8/23/2005
Jean Blois	300 Day St.	Hoe	9/28/2007

- Create a separate output records for each product sold to John Brown.

Name	Address	City	State	Zipcode	Product1	Product2
John Brown	1234 E. St.	Chicago	IL	60610	Mower	Rake
Neal Smith	325 Clay St.	Portland	OR	97201	Shovel	
Jean Blois	300 Day St.	Houston	TX	77091	Hoe	

If you want to include fields like Product as predictors in the model. In this case, you must create separate fields for each record and copy the relevant data into them. The second process is called flattening or denormalizing the database.

Customer Analytical Record

- All relevant data for each customer are listed in the same record.

3. Data Description

- Descriptive statistical metrics for individual variables (univariate analysis)

Mean, Standard deviation Minimum, Maximum, Standard deviation, Frequency tables, Histograms

- Assessment of relationships between pairs of variables (bivariate analysis)

- Visual/graphical techniques for viewing more complex relationships between variables.

4. Data Assessment

- Before fixing any problems in the data set, find them and decide how to handle them.
- Some problems will become evident during data description operations.
- Data auditing is similar to auditing in accounting, and includes two operations: **data profiling and the analysis of the impact of poor-quality data.**

5. Data Profiling

Look at the data distributions of each variable, and note the following:

- The central tendency of data in the variable
- Any potential outliers
- The number of and distribution of blanks across all the cases
- Any suspicious data, like miscodes, training data, system test data, or just plain garbage

Your findings should be presented in the form of a report and listed as a milestone in the project plan

6. Data Cleansing

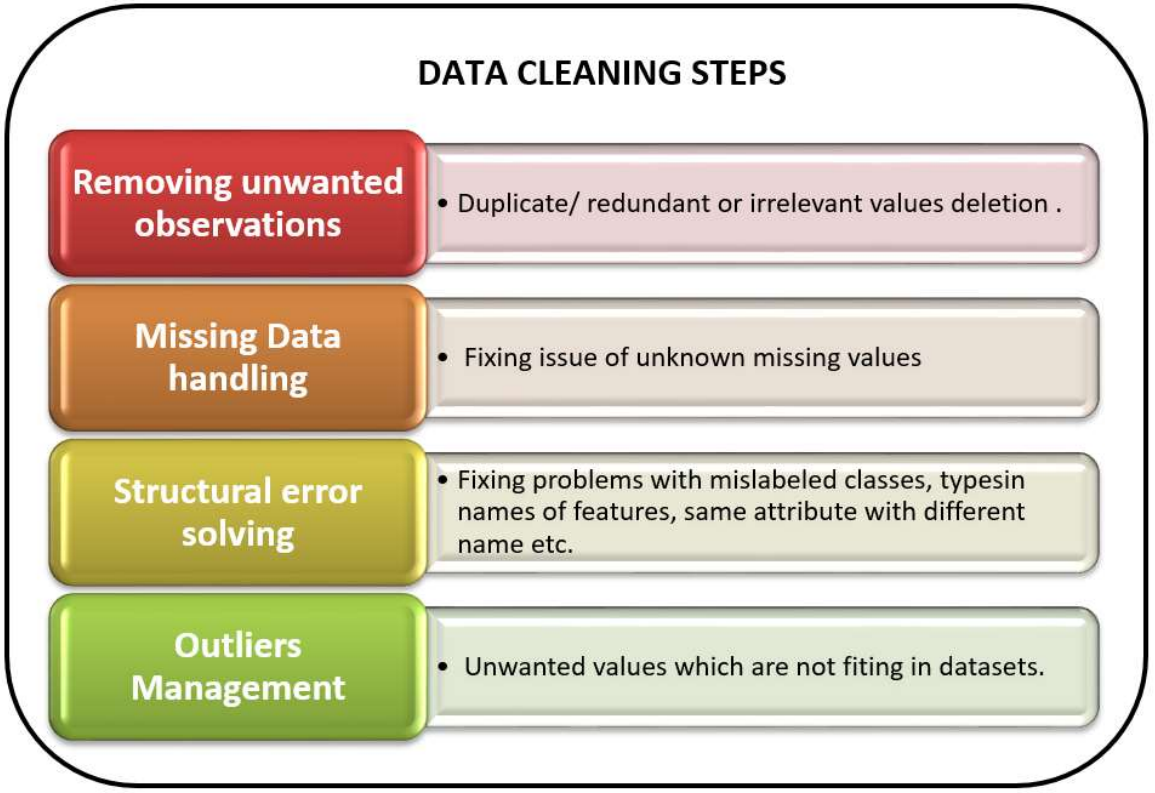
Data cleansing includes operations that correct bad data, filter some bad data out of the data set, and filter out data that are too detailed for use in your model

Python:

Pandas &
NumPy

R package:

Plyr, tidyr



7. Data Transformation

- **Numerical Variables:** With parametric statistical modeling algorithm, you should transform any variables forming exponential (nonlinear) curves. Otherwise, estimation errors caused by the violation of the assumption of linearity could invalidate predictions made by the model.
- **Standardization** means to transform all numerical values to a common range.

$$z = (\text{value} - \text{mean})/\text{standard deviation}$$

Parametric machine learning algorithm	Nonparametric machine learning algorithm
Logistic Regression, LDA, Perceptron, Naive Bayes Simple Neural Networks	k-Nearest Neighbors, Decision Trees like CART and C4.5, SVM
Simpler: These methods are easier to understand and interpret results. **Speed**: Parametric models are very fast to learn from data. **Less Data**: They do not require as much training data and can work well even if the fit to the data is not perfect.	**Flexibility**: Capable of fitting a large number of functional forms. **Power**: No assumptions (or weak assumptions) about the underlying function. **Performance**: Can result in higher performance models for prediction.
Constrained: By choosing a functional form these methods are highly constrained to the specified form. **Limited Complexity**: The methods are more suited to simpler problems. **Poor Fit**: In practice the methods are unlikely to match the underlying mapping function.	**More data**: Require a lot more training data to estimate the mapping function. **Slower**: A lot slower to train as they often have far more parameters to train. **Overfitting**: More of a risk to overfit the training data and it is harder to explain why specific predictions are made.

Categorical Variable

- **Dummy variables** transform categorical (discrete) data into numerical data.

Coding of Dummy Variables for the Variable Color

Case	Color	Color-Red	Color-Blue	Color-Yellow	Color-Green
1	Red	1	0	0	0
2	Blue	0	1	0	0
3	Yellow	0	0	1	0
4	Green	0	0	0	1
5	Blue	0	1	0	0

Algorithms that depend on calculations of covariance (e.g., regression) or that require other numerical operations (e.g., most neural nets) must operate on numbers.

Some statistical packages recode categorical data with a set of sequential numbers automatically and treat them numerically.

8. Data Imputation

- Process of replacing **missing data** with intuitive data
- Assign data to the blank based on some reasonable heuristic (a rule or set of rules)
- Selection of the proper technique for handling missing values depends on making the right assumption about the pattern of “missingness” in the data set.
- If there is a strong pattern among the missing values of a variable (e.g., caused by a broken sensor), the variable should be eliminated from the model.

9. Data Weighting and Balancing

Parametric statistical algorithms measure **how far various derived metrics** (e.g., means, standard deviations, etc.) are from critical values defined by the characteristics of the .data distribution.

For example, if a value in a data set is beyond 1.96 standard deviation units from the mean (= the z-value), it is beyond the value where it could be a part of the other data 95% of the time. This limit is called the 95% Confidence Level (or 95% CL).

- Simple Linear regression learn things about the data by using all cases to calculate the metrics (e.g., mean and standard deviation), and compare all data values in relation to those metrics and standard tables of probability to decide if a relationship exists.

Weight

- **Example:** a data value input from sensor-A may be twice as accurate as data from sensor-B. In this case, we can apply a weight of 2 for all values input by sensor-A and a weight of 1 for values input from sensor-B.
- **Example: Weight in ML :** Machine learning (ML) algorithms learn in a very different way. Instead of going through all of the cases to calculate summary metrics, machine learning algorithms learn case by case.

- **We will study about it in Neural network**

10. Data Filtering and Smoothing

- Data filtering refers to eliminating rows (cases) to remove unnecessary information
- It reduces the noise below the level that can confuse the analysis
- **Removal of Outliers:** Removing outliers is good but some outliers are of primary interest to the modelling of credit risk, fraud, and other rare events.

Outlier Detection and Removal

Descriptive statistics: Minimum and maximum, Histogram Boxplot, Percentiles

Python: Isolation Forest, Minimum Covariance Determinant, Local Outlier Factor, One-Class SVM

R:Hampel filter, Grubbs's test, Dixon's test, Rosner's test

Homework

- **SQL Tutorial**
- **https://www.w3schools.com/sql/**
- **Database Queries With R**

https://db.rstudio.com/getting-started/database-queries/

- **SQL queries in Python**

https://towardsdatascience.com/sql-queries-in-python-51ef85b92c1e