CSC550Z: Data Mining for Business Intelligence

Week 1: Introduction to Data Mining

Tuan Tran, Ph.D.

Assistant Professor
College of Information and Computer Technology
Sullivan University
Email: ttran@sullivan.edu

October 27, 2016

Objectives

- Understand data mining concepts
- Understand steps in a data mining project
- Identify types of variables and understand data pre-processing
- Know how to normalize data
- Understand the issues of overfitting and how to fix it
- Steps in constructing data mining model using XLMiner
- Know how to read and intepret output

Outline

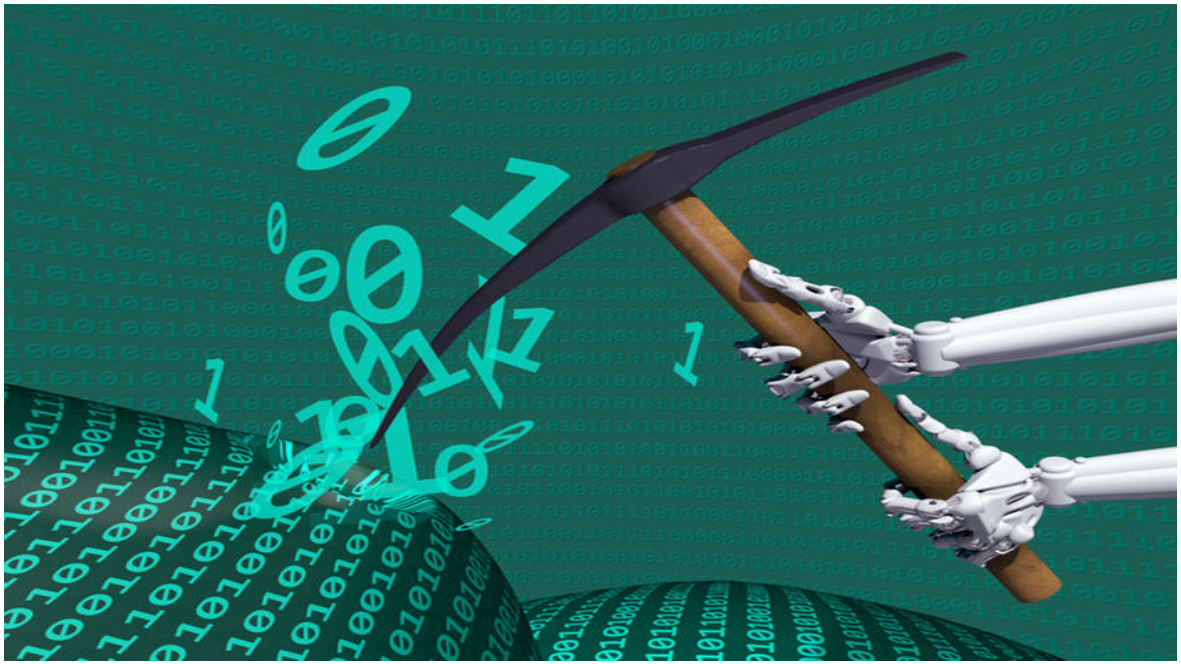
Why Mine Data?

- **Massive data is collected and warehoused**
  - Web data, e-commerce
  - Purchase at grocery stores
  - Bank/credit card transactions
- **Computing system is cheaper and more powerful**
- **Competitive pressure is strong**
  - Provide better customized services

What is Data Mining?

There are many definitions

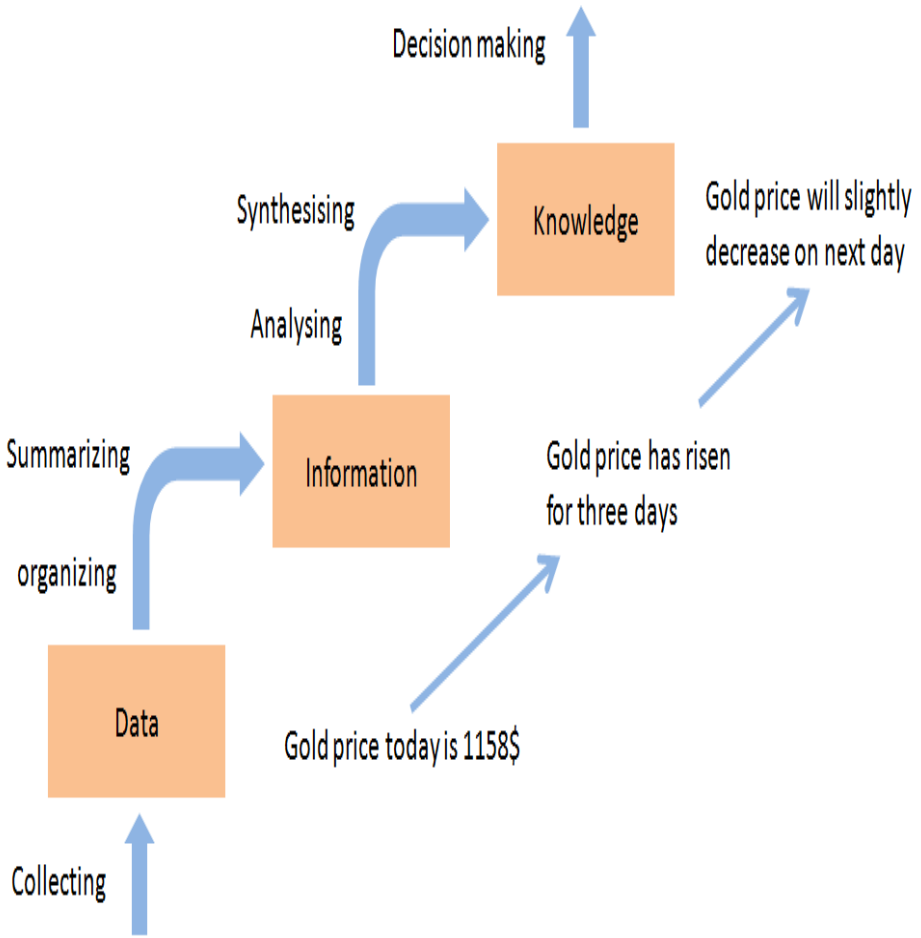
- **Non-trivial extraction** of implicit, previously unknown and potentially useful information from data
- **Exploration** & **analysis**, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns



Core Ideas in Data Mining

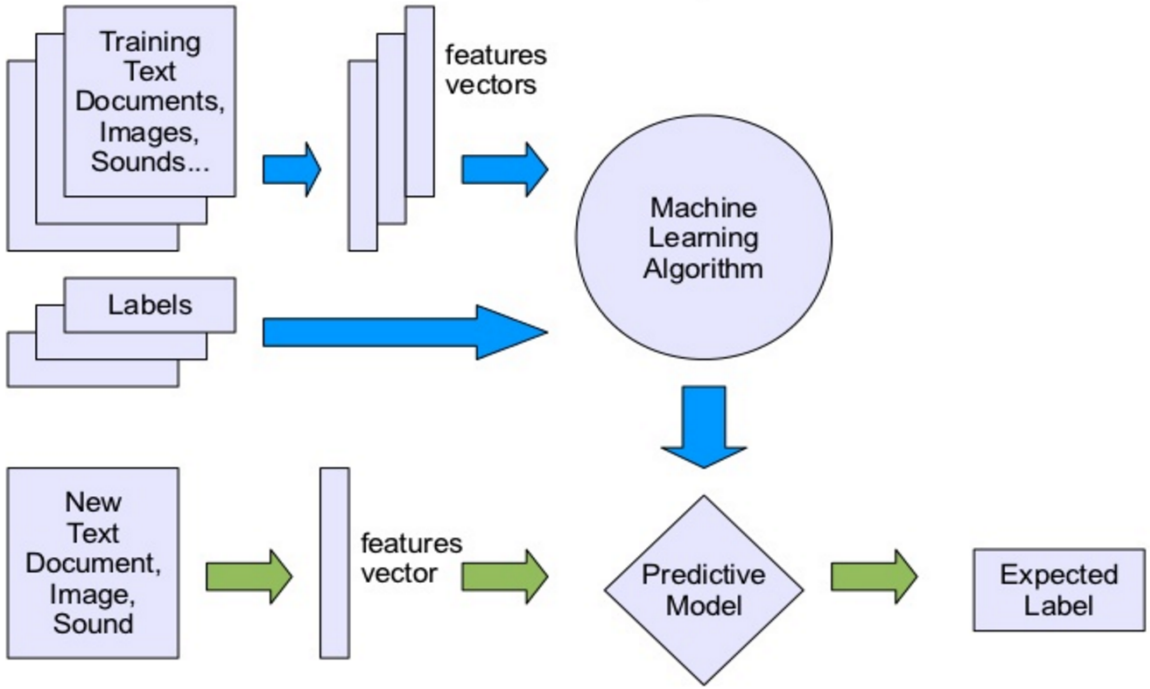
Consisting of several aspects

- **Classification**: Basic form of data analysis (purchase/no purchase)
- **Prediction**: Predict value of a numerical variable (amount of purchase)
- **Association Rules**: What goes with what
- **Data Reduction**: Transform into simpler data
- **Data Exploration**: Review and examine data
- **Visualization**: Graphical analysis



Supervised Learning

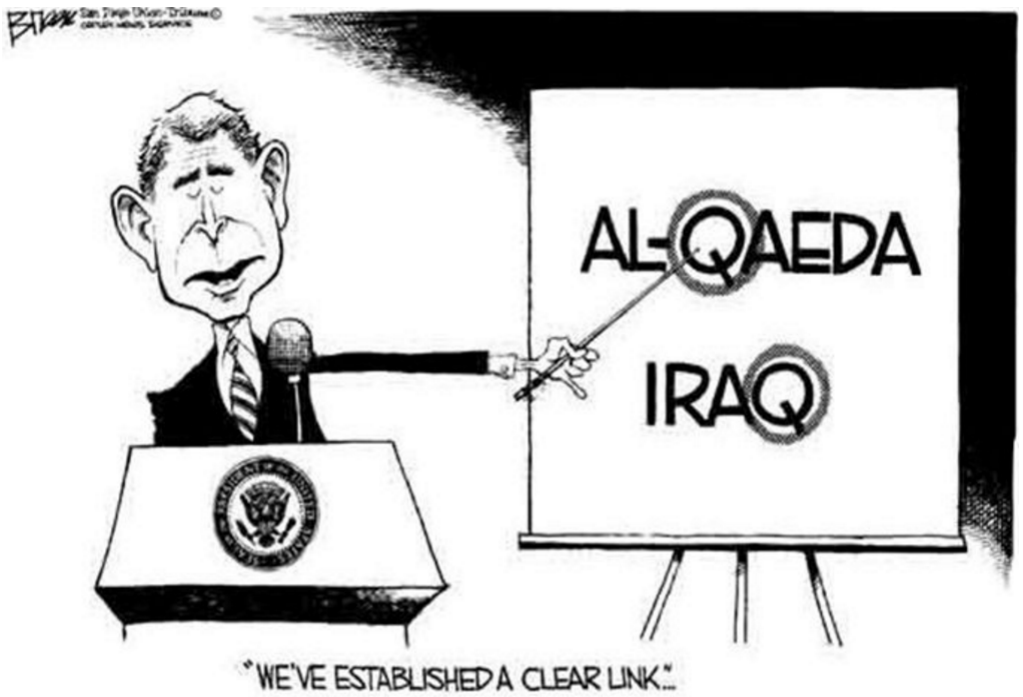
- **Goal**: Predict a single "target" or "outcome" variable
  - **Training data**: target value is known
  - **New data**: score to new data where (target) value is not known
- **Methods**: Classification and Prediction
  - **Example**: Classify emails (spam/no-spam), predict house value, etc.



Source: www.astroml.org

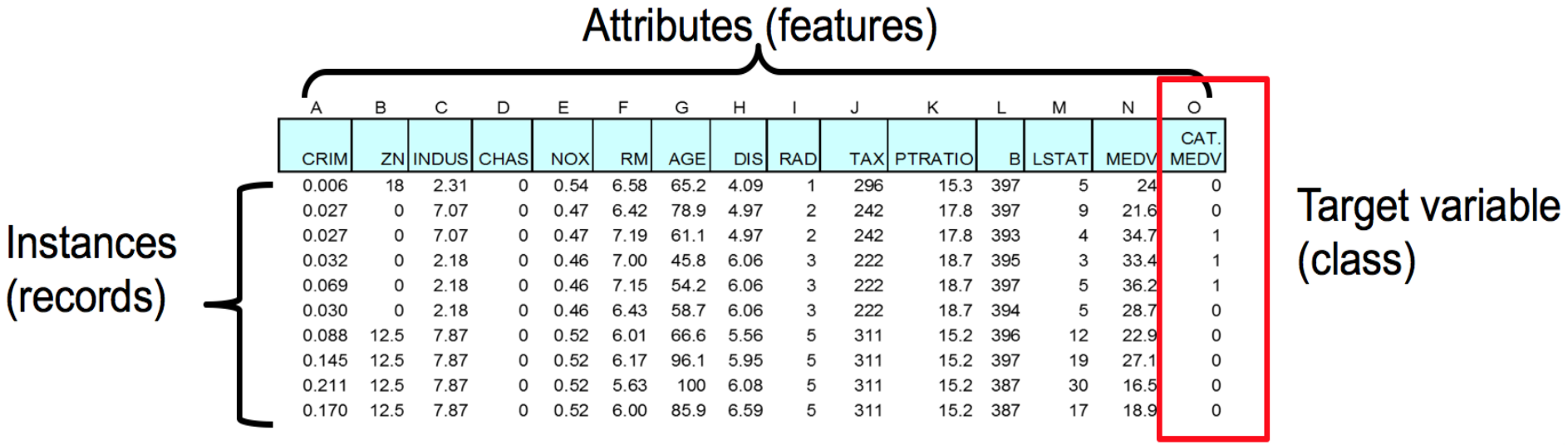
Unsupervised Learning

- **Goal**: Segment data into meaningful segments; detect patterns
  - Draw inferences from datasets consisting of input data without labeled responses
  - There is no target (outcome) variable to predict or classify
- **Methods**: Association rules, data reduction & exploration, visualization
  - **Example**: shopping basket, clustering, etc.



Supervised Learning: Classification

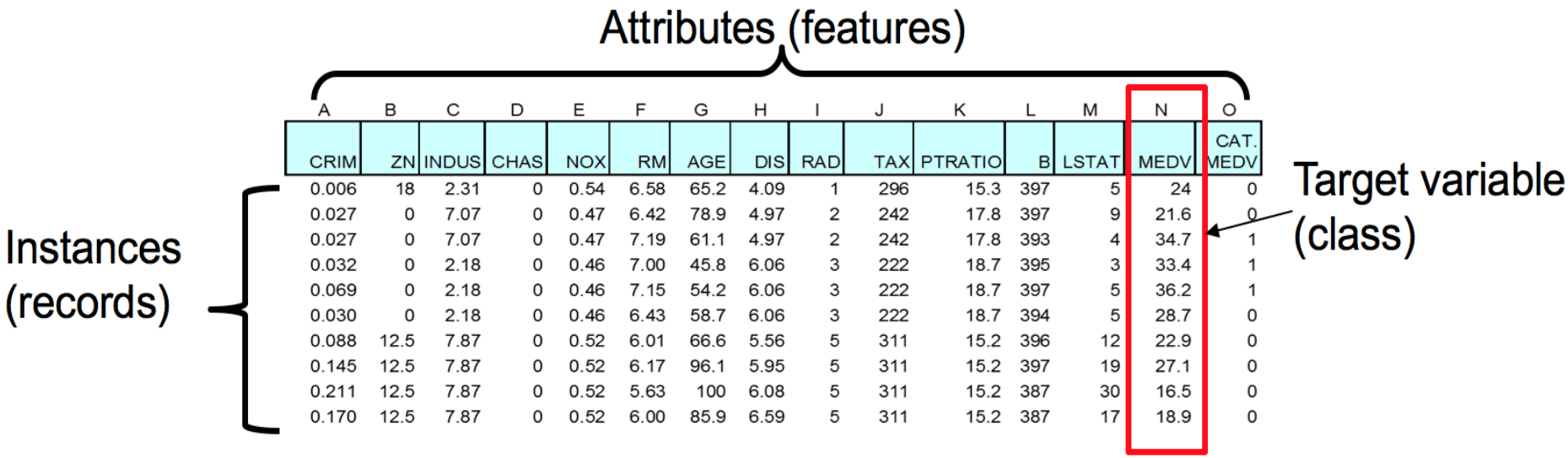
- **Goal**: Predict categorical target (outcome) variable
- **Examples**: Purchase/no purchase, fraud/no fraud, creditworthy/not creditworthy
  - Each row is a case/instance/record (customer, tax return, applicant)
  - Each column is a variable (attribute, feature)
- **Target types**: Target variable is often binary (yes/no)



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV	CAT. MEDV
0.006	18	2.31	0	0.54	6.58	65.2	4.09	1	296	15.3	397	5	24	0
0.027	0	7.07	0	0.47	6.42	78.9	4.97	2	242	17.8	397	9	21.6	0
0.027	0	7.07	0	0.47	7.19	61.1	4.97	2	242	17.8	393	4	34.7	1
0.032	0	2.18	0	0.46	7.00	45.8	6.06	3	222	18.7	395	3	33.4	1
0.069	0	2.18	0	0.46	7.15	54.2	6.06	3	222	18.7	397	5	36.2	1
0.030	0	2.18	0	0.46	6.43	58.7	6.06	3	222	18.7	394	5	28.7	0
0.088	12.5	7.87	0	0.52	6.01	66.6	5.56	5	311	15.2	396	12	22.9	0
0.145	12.5	7.87	0	0.52	6.17	96.1	5.95	5	311	15.2	397	19	27.1	0
0.211	12.5	7.87	0	0.52	5.63	100	6.08	5	311	15.2	387	30	16.5	0
0.170	12.5	7.87	0	0.52	6.00	85.9	6.59	5	311	15.2	387	17	18.9	0

Supervised Learning: Prediction

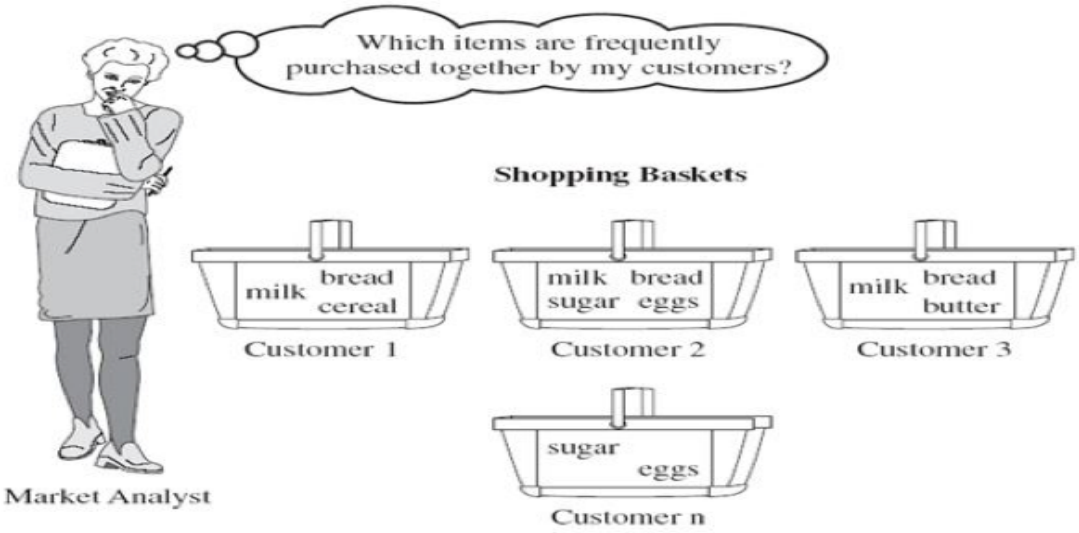
- **Goal**: Predict numerical target (outcome) variable
- **Examples**: sales, revenue, performance
  - Each row is a case/instance/record (customer, tax return, applicant)
  - Each column is a variable (attribute, feature)
- Classification and prediction together constitute "predictive analytics"



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV	CAT. MEDV
0.006	18	2.31	0	0.54	6.58	65.2	4.09	1	296	15.3	397	5	24	0
0.027	0	7.07	0	0.47	6.42	78.9	4.97	2	242	17.8	397	9	21.6	0
0.027	0	7.07	0	0.47	7.19	61.1	4.97	2	242	17.8	393	4	34.7	1
0.032	0	2.18	0	0.46	7.00	45.8	6.06	3	222	18.7	395	3	33.4	1
0.069	0	2.18	0	0.46	7.15	54.2	6.06	3	222	18.7	397	5	36.2	1
0.030	0	2.18	0	0.46	6.43	58.7	6.06	3	222	18.7	394	5	28.7	0
0.088	12.5	7.87	0	0.52	6.01	66.6	5.56	5	311	15.2	396	12	22.9	0
0.145	12.5	7.87	0	0.52	6.17	96.1	5.95	5	311	15.2	397	19	27.1	0
0.211	12.5	7.87	0	0.52	5.63	100	6.08	5	311	15.2	387	30	16.5	0
0.170	12.5	7.87	0	0.52	6.00	85.9	6.59	5	311	15.2	387	17	18.9	0

Unsupervised Learning: Association Rules

- **Goal**: Produce rules that define "what goes with what"
- **Examples**: "If X was purchased, Y was also purchased"
  - Rows are transactions
  - Each column is a variable (attribute, feature)
- Used in recommender systems - "Our records show you bought X, you may also like Y"
- Also called "affinity analysis"



Unsupervised Learning: Association Rules

- How Can We Qualtify It?

Dataset

Transactions	Items
12345	ABC
12346	AC
12347	AD
12348	BEF

Support

Itemset	Support
A	75%
B	50%
C	50%
A,C	50%

- Parameters: support (50%) and confidence (e.g, 50%)
- **Rules**: $A \rightarrow C$
  - support (A,C) = 50%
  - confidence ($A \rightarrow C$) = support(A,C)/support(A)=66,6%

Unsupervised Learning: Data Reduction

- **Goal**: Distillation of complex/large data into simpler/smaller data
- Reducing the number of variables/columns (e.g., principal components)
- Reducing the number of records/rows (e.g., clustering)

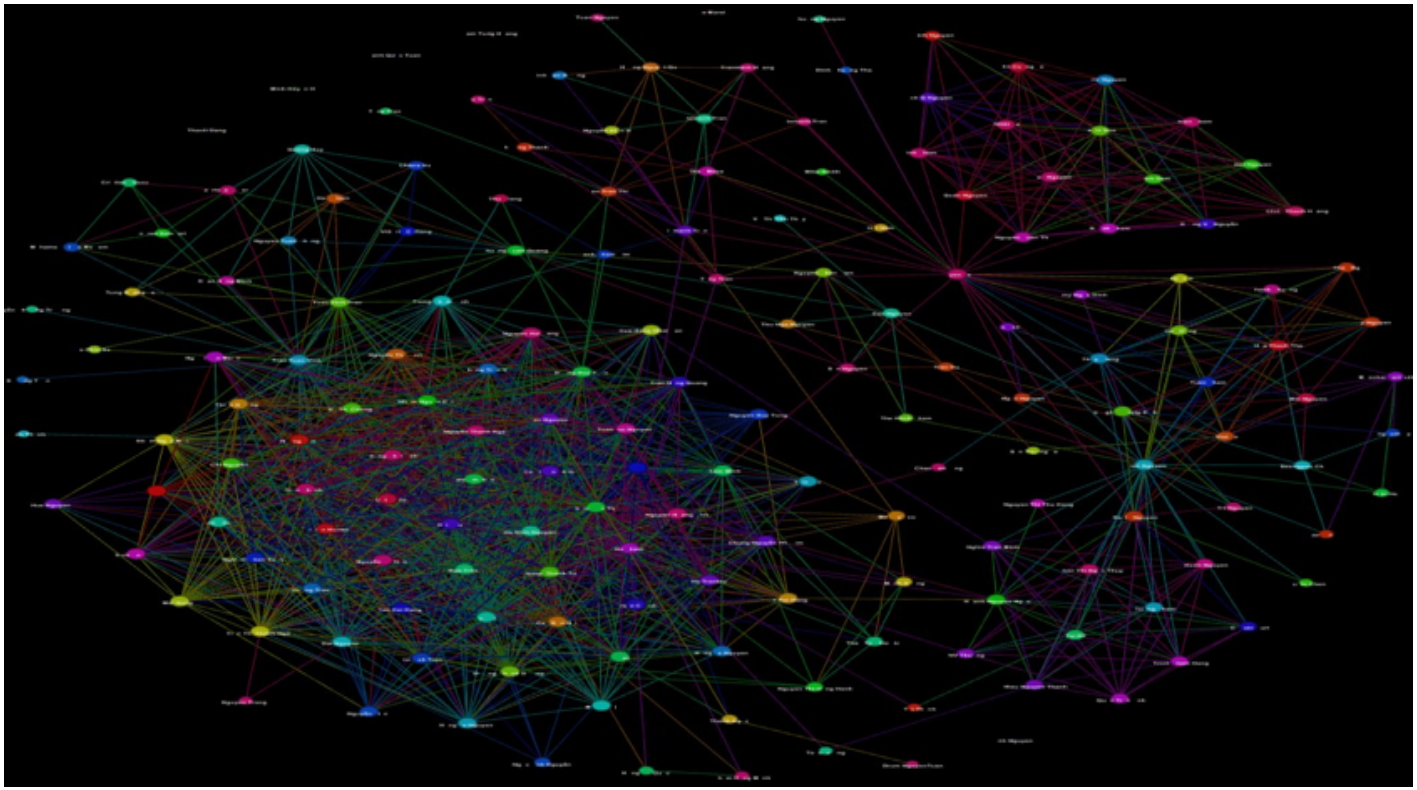
Attributes (features)

Instances (records)

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV	CAT. MEDV
0.006	18	2.31	0	0.54	6.58	65.2	4.09	1	296	15.3	397	5	24	0
0.027	0	7.07	0	0.47	6.42	78.9	4.97	2	242	17.8	397	9	21.6	0
0.027	0	7.07	0	0.47	7.19	61.1	4.97	2	242	17.8	393	4	34.7	1
0.032	0	2.18	0	0.46	7.00	45.8	6.06	3	222	18.7	395	3	33.4	1
0.069	0	2.18	0	0.46	7.15	54.2	6.06	3	222	18.7	397	5	36.2	1
0.030	0	2.18	0	0.46	6.43	58.7	6.06	3	222	18.7	394	5	28.7	0
0.088	12.5	7.87	0	0.52	6.01	66.6	5.56	5	311	15.2	396	12	22.9	0
0.145	12.5	7.87	0	0.52	6.17	96.1	5.95	5	311	15.2	397	19	27.1	0
0.211	12.5	7.87	0	0.52	5.63	100	6.08	5	311	15.2	387	30	16.5	0
0.170	12.5	7.87	0	0.52	6.00	85.9	6.59	5	311	15.2	387	17	18.9	0

Unsupervised Learning: Data Visualization

- **Goal**: Graphs and plots of data for visualization
  - Histograms, boxplots, bar charts, scatterplots
- Especially useful to examine relationships between pairs of variables
  - Graphical network of my social network links



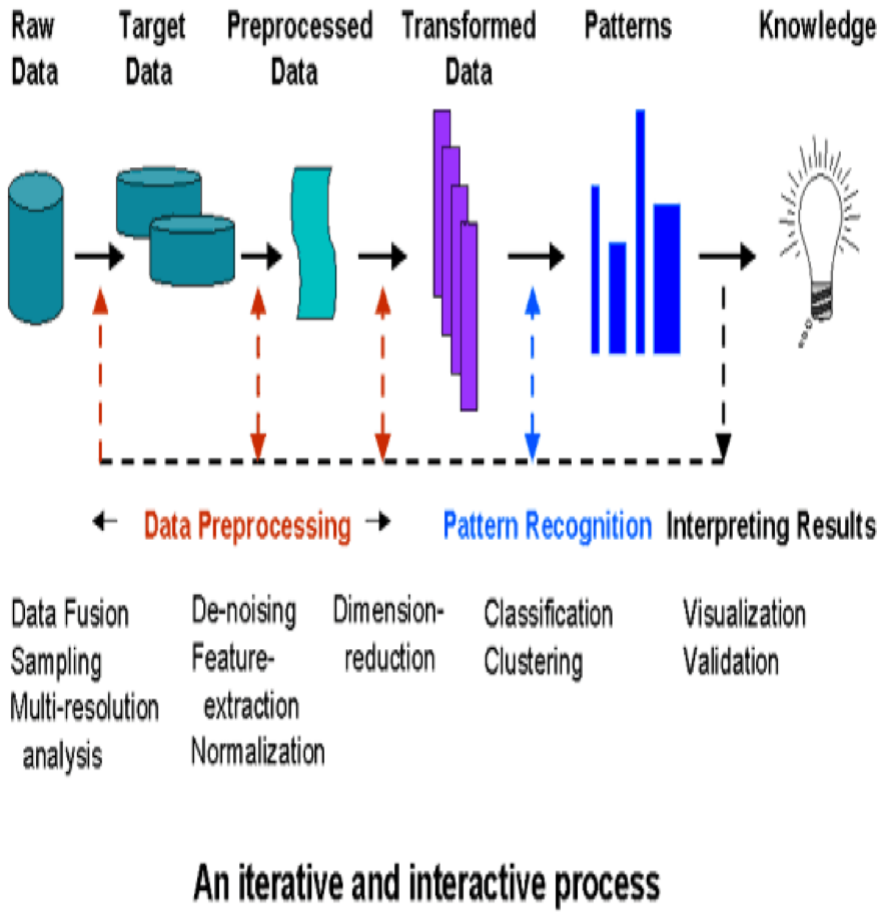
Unsupervised Learning: Data Exploration

- Data sets are typically large, complex & messy (dirty)
  - Need to review the data to help refine the task
- Use techniques of Reduction and Visualization

Steps in Data Mining

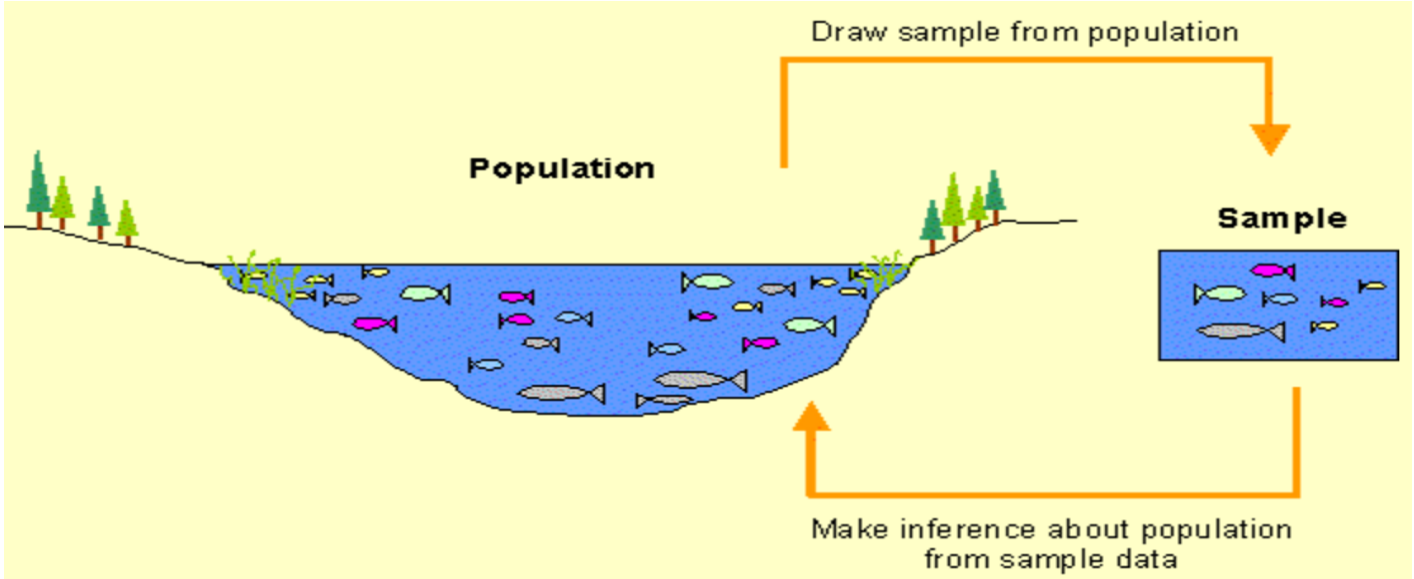
Consisting of several steps

1. **Understanding**: Define purpose
2. **Data Collection**: Obtain data (random sampling)
3. **Cleaning**: Explore, pre-process data
4. **Data Reduction**: Reduce the data; if supervised DM, partition it
5. **Identify Task**: Specify task (classification, clustering, etc.)
6. **Technique Selection**: Choose the techniques (CART, NN, etc.)
7. **Parameter Tunning**: Iterative and "tuning"
8. **Evaluation**: Assess & compare models
9. **Deployment**: Deploy best model



Obtaining Data: Sampling

- Why Sampling?
  - Massive dataset with correlation instances
  - Software limitation, e.g., XLMiner, e.g., limits the "training" partition to 10,000 records
- How to Sample?
  - Use build-in function of software
- What Is the Size of Sample?
  - Data mining typically deals with huge datasets
  - Produce statistically-valid results



Source: http://labs.geog.uvic.ca/

Rare event oversampling

- Often the event of interest is rare
  - **Examples**: response to mailing, fraud in taxes, ...
- Sampling may yield too few "interesting" cases to effectively train a model
- A popular solution: **oversample** the rare cases to obtain a more balanced training set
  - Later, need to adjust results for the oversampling

Types of Variables

- Determine the types of pre-processing needed, and algorithms used
- Main distinction: Categorical vs. numeric
- Numeric
  - Continuous
  - Integer
- Categorical
  - Ordered (low, medium, high) (ordinal)
  - Unordered (male, female) (nominal)

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV	CAT. MEDV
0.006	18	2.31	0	0.54	6.58	65.2	4.09	1	296	15.3	397	5	24	0
0.027	0	7.07	0	0.47	6.42	78.9	4.97	2	242	17.8	397	9	21.6	0
0.027	0	7.07	0	0.47	7.19	61.1	4.97	2	242	17.8	393	4	34.7	1
0.032	0	2.18	0	0.46	7.00	45.8	6.06	3	222	18.7	395	3	33.4	1
0.069	0	2.18	0	0.46	7.15	54.2	6.06	3	222	18.7	397	5	36.2	1
0.030	0	2.18	0	0.46	6.43	58.7	6.06	3	222	18.7	394	5	28.7	0
0.088	12.5	7.87	0	0.52	6.01	66.6	5.56	5	311	15.2	396	12	22.9	0
0.145	12.5	7.87	0	0.52	6.17	96.1	5.95	5	311	15.2	397	19	27.1	0
0.211	12.5	7.87	0	0.52	5.63	100	6.08	5	311	15.2	387	30	16.5	0
0.170	12.5	7.87	0	0.52	6.00	85.9	6.59	5	311	15.2	387	17	18.9	0

Variable handling

- Numeric
  - Most algorithms in XLMiner can handle numeric data
  - May occasionally need to "bin" into categories
- Categorical
  - Naive Bayes can use as-is
  - In most other algorithms, must create binary dummies (number of dummies = number of categories - 1)

Variable Handling

- An outlier is an observation that is "extreme", being distant from the rest of the data (definition of "distant" is deliberately vague)
- Outliers can have disproportionate influence on models (a problem if it is spurious)
- An important step in data pre-processing is detecting outliers
- Once detected, domain knowledge is required to determine if it is an error, or truly extreme
- In some contexts, finding outliers is the purpose of the DM exercise (airport security screening). This is called "anomaly detection"

Handling Missing Data

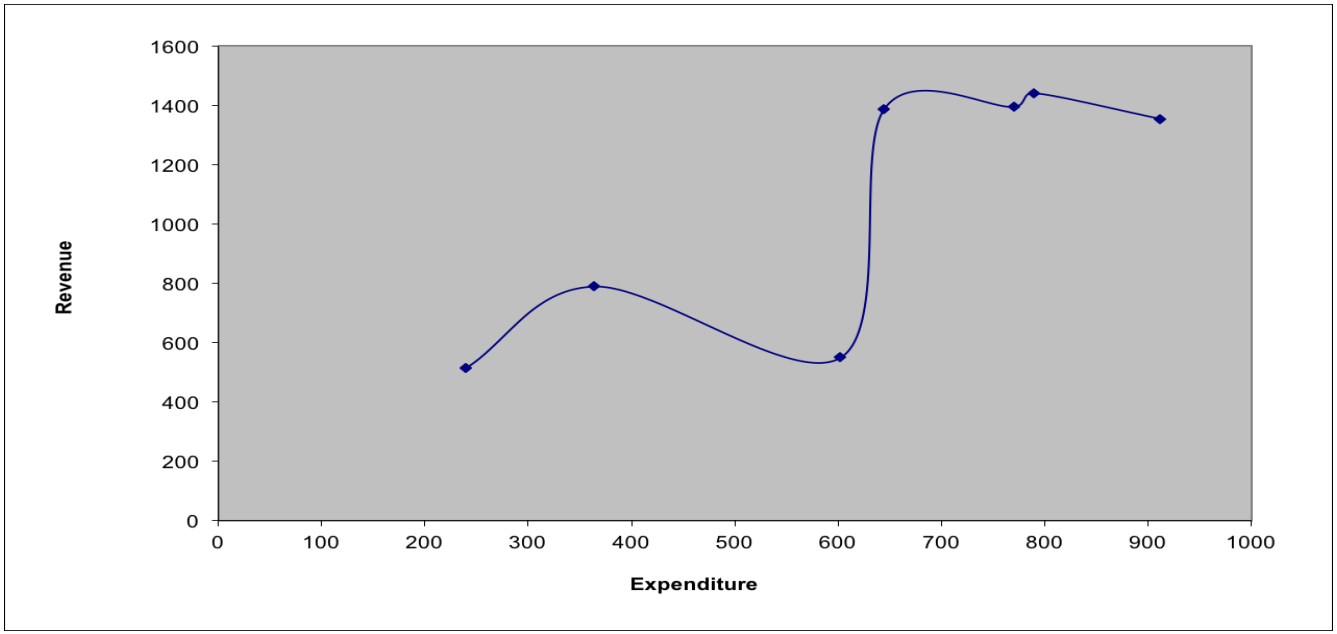
- Most algorithms will not process records with missing values. Default is to drop those records.
- Solution 1: Omission
  - If a small number of records have missing values, can omit them
  - If many records are missing values on a small set of variables, can drop those variables (or use proxies)
  - If many records have missing values, omission is not practical
- Solution 2: Imputation
  - Replace missing values with reasonable substitutes
  - Lets you keep the record and use the rest of its (non-missing) information

Normalizing (Standardizing) Data

- Used in some techniques when variables with the largest scales would dominate and skew results
- Puts all variables on same scale
- Normalizing function: subtract mean and divide by standard deviation (used in XLMiner)
  - Mean of n samples x_i: $\mu = \sum_{i=1}^{n} x_i / n$
  - Standard deviation $\sigma = \sqrt{\sum_{i=1}^{n} (x_i - \mu)^2 / n}$
  - Standardized samples: $x_i = (x_i - \mu)/\sigma$
- Alternative function: scale to 0-1 by subtracting minimum and dividing by the range
  - Range $r = x_{max} - x_{min}$
  - Normalized samples $x_i = (x_i - x_{min})/r$
  - Useful when the data contain dummies and numeric

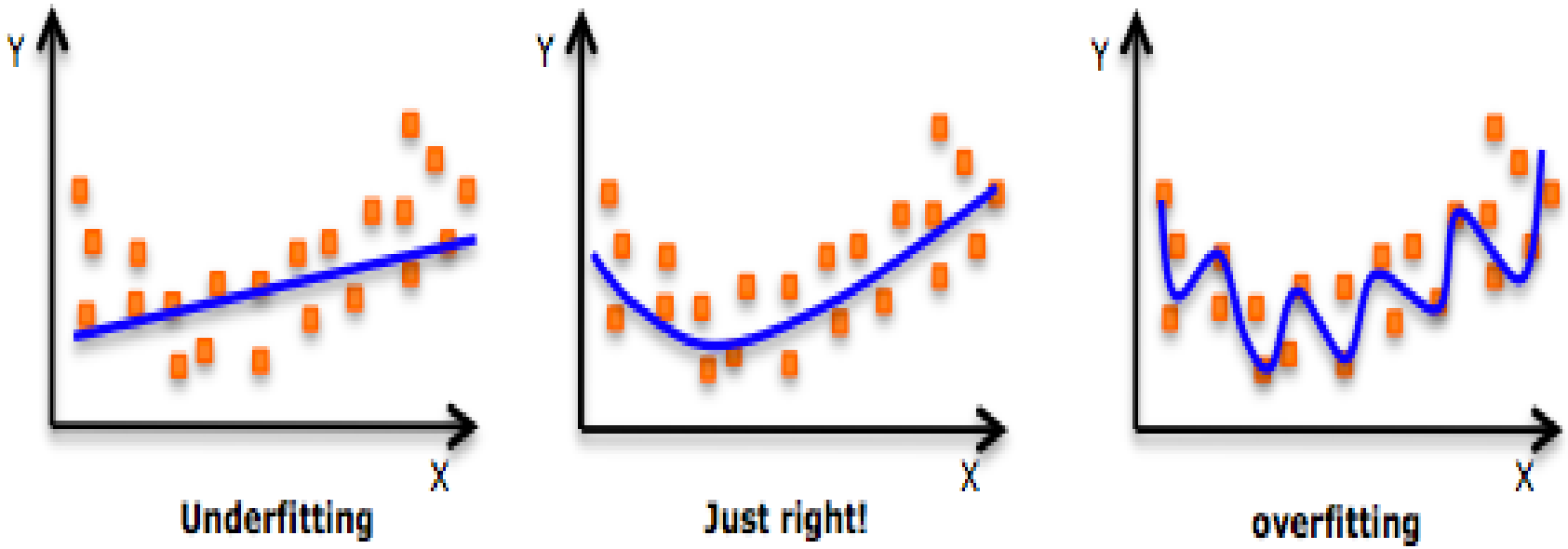
The Problem of Overfitting

- Statistical models can produce highly complex explanations of relationships between variables
- The "fit" may be excellent
- When used with new data, models of great complexity do not do so well
- 100% fit - not useful for new data



The Problem of Overfitting (cont.)

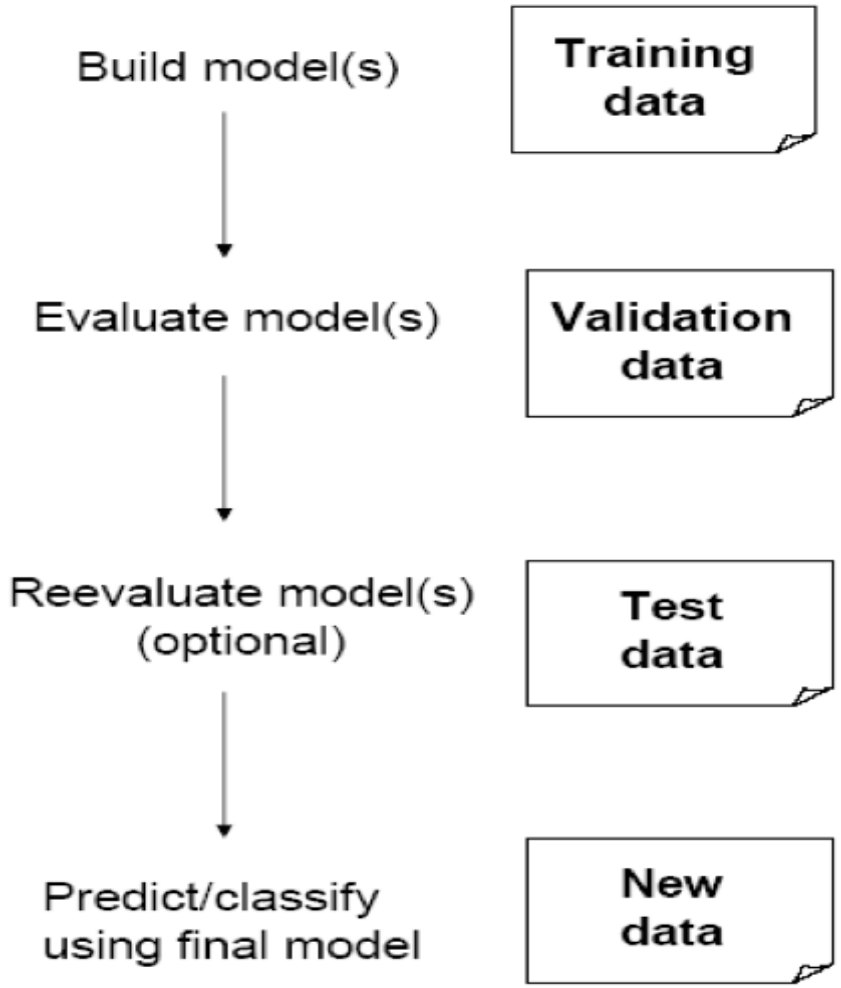
- Causes:
  - Too many predictors
  - A model with too many parameters
  - Trying many different models
- Consequence: Deployed model will not work as well as expected with completely new data.



Source: http://pingax.com

Partitioning the Data

- **Problem**: How well will our model perform with new data?
- **Solution**: Separate data into two parts
  - Training partition to develop the model
  - Validation partition to implement the model and evaluate its performance on "new" data
- Addresses the issue of overfitting



Test Partition

- When a model is developed on training data, it can overfit the training data (hence need to assess on validation)
- Assessing multiple models on same validation data can overfit validation data
- Some methods use the validation data to choose a parameter. This too can lead to overfitting the validation data
- **Solution**: final selected model is applied to a test partition to give unbiased estimate of its performance on new data

Example: MLR for Predicting Boston Housing (XLMiner)

- This example shows data mining steps for predicting house values in Boston areas
  - Use BostonHousing.xls dataset
  - Construct a multiple linear regression (MLR) model to predict house values based on house features
  - Intepret output of the model
- Sample of the dataset

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV	CAT. MEDV
0.006	18	2.31	0	0.54	6.58	65.2	4.09	1	296	15.3	397	5	24	0
0.027	0	7.07	0	0.47	6.42	78.9	4.97	2	242	17.8	397	9	21.6	0
0.027	0	7.07	0	0.47	7.19	61.1	4.97	2	242	17.8	393	4	34.7	1
0.032	0	2.18	0	0.46	7.00	45.8	6.06	3	222	18.7	395	3	33.4	1
0.069	0	2.18	0	0.46	7.15	54.2	6.06	3	222	18.7	397	5	36.2	1
0.030	0	2.18	0	0.46	6.43	58.7	6.06	3	222	18.7	394	5	28.7	0
0.088	12.5	7.87	0	0.52	6.01	66.6	5.56	5	311	15.2	396	12	22.9	0
0.145	12.5	7.87	0	0.52	6.17	96.1	5.95	5	311	15.2	397	19	27.1	0
0.211	12.5	7.87	0	0.52	5.63	100	6.08	5	311	15.2	387	30	16.5	0
0.170	12.5	7.87	0	0.52	6.00	85.9	6.59	5	311	15.2	387	17	18.9	0

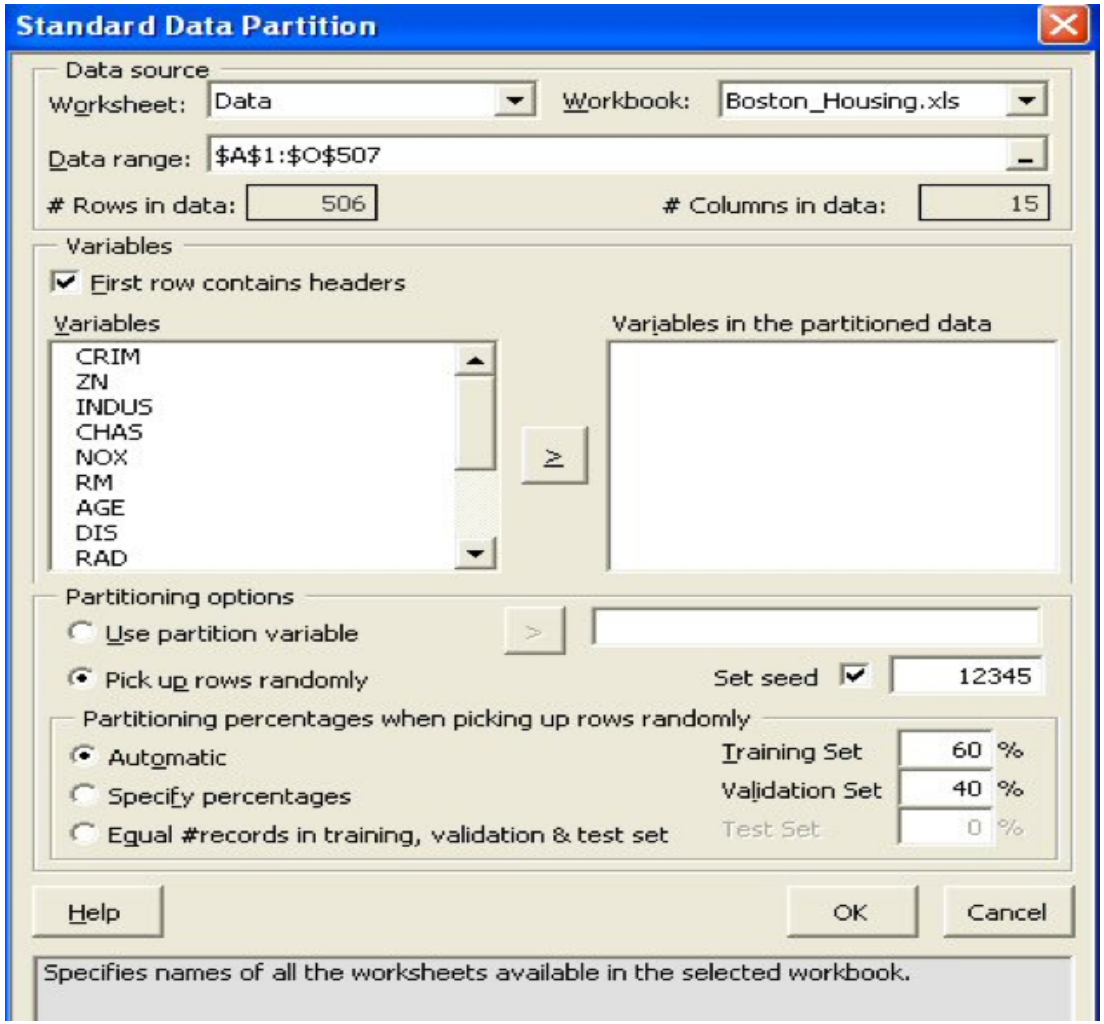
Example: MLR for Predicting Boston Housing (XLMiner) (cont.)

Description of variables

CRIM	per capita crime rate by town
ZN	proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS	proportion of non-retail business acres per town.
CHAS	Charles River dummy variable (1 if tract bounds river; 0 otherwise)
NOX	nitric oxides concentration (parts per 10 million)
RM	average number of rooms per dwelling
AGE	proportion of owner-occupied units built prior to 1940
DIS	weighted distances to five Boston employment centres
RAD	index of accessibility to radial highways
TAX	full-value property-tax rate per $10,000
PTRATIO	pupil-teacher ratio by town
B	1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
LSTAT	% lower status of the population
MEDV	Median value of owner-occupied homes in $1000

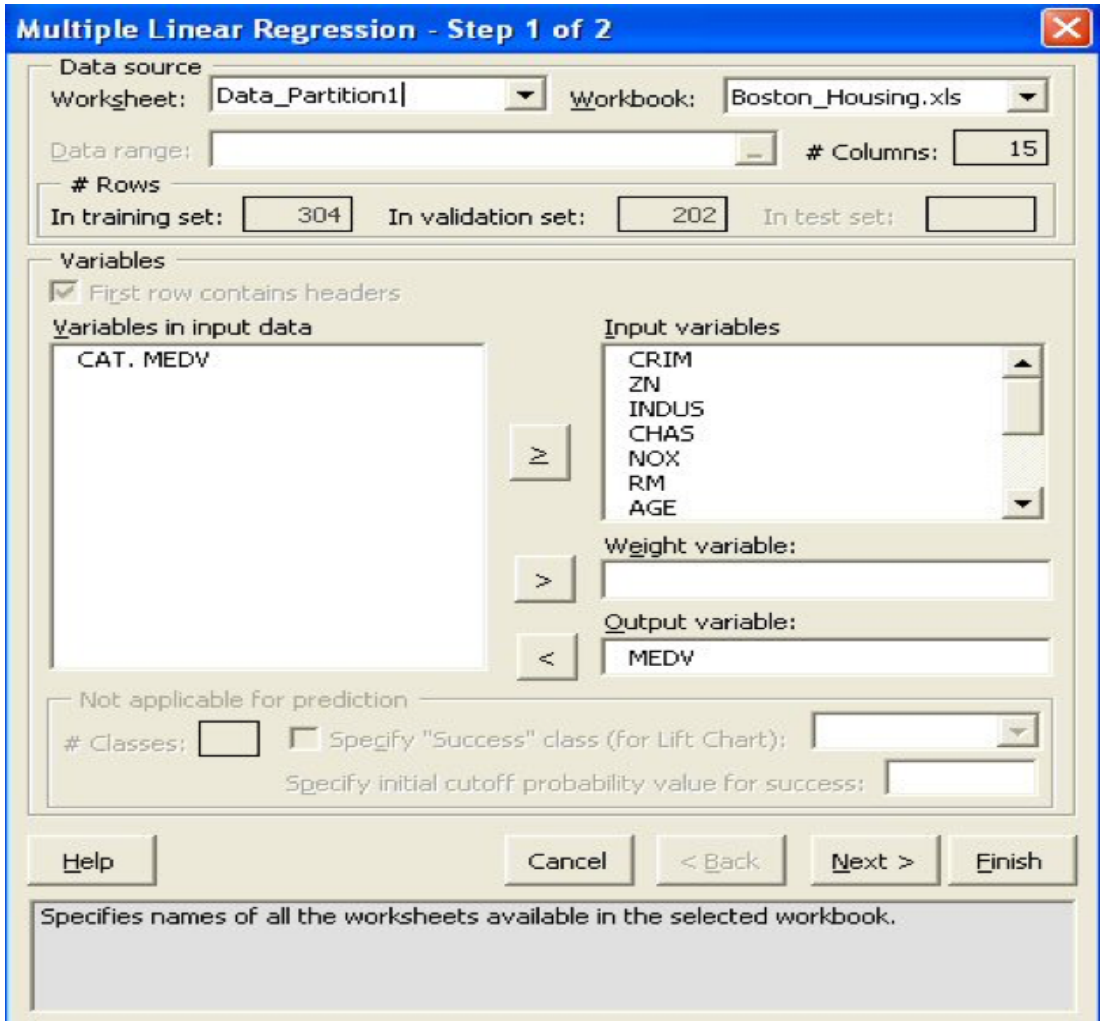
Example: MLR for Predicting Boston Housing (XLMiner) (cont.)

Partitioning the data



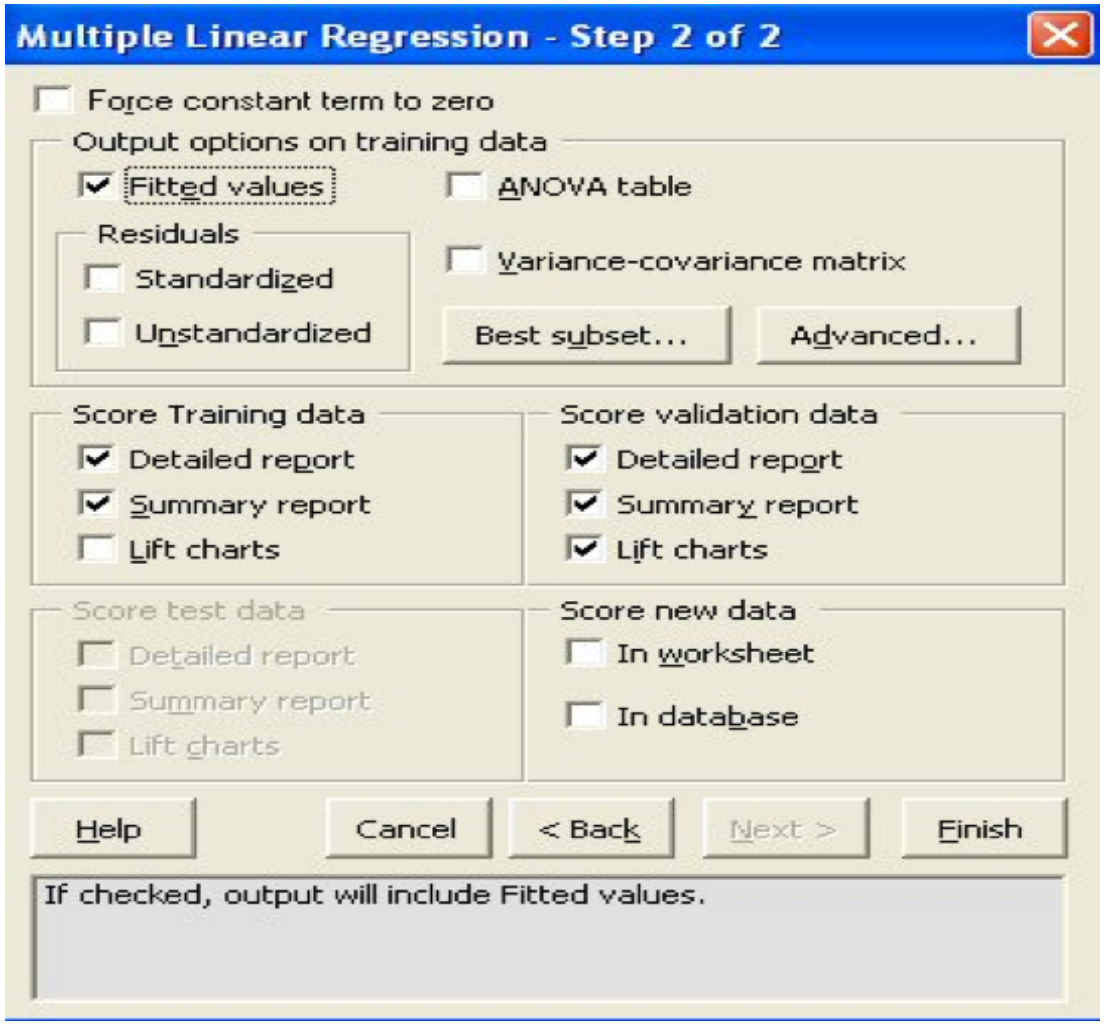
Example: MLR for Predicting Boston Housing (XLMiner) (cont.)

Using XLMiner for Multiple Linear Regression



Example: MLR for Predicting Boston Housing (XLMiner) (cont.)

Specifying Output



Example: MLR for Predicting Boston Housing (XLMiner) (cont.)

Prediction of Training Data

Row Id.	Predicted Value	Actual Value	Residual
1	30.24690555	24	-6.246905549
4	28.61652272	33.4	4.783477282
5	27.76434086	36.2	8.435659135
6	25.6204032	28.7	3.079596801
9	11.54583087	16.5	4.954169128
10	19.13566187	18.9	-0.235661871
12	21.95655773	18.9	-3.05655773
17	20.80054199	23.1	2.299458015
18	16.94685562	17.5	0.553144385

Example: MLR for Predicting Boston Housing (XLMiner) (cont.)

Prediction of Validation Data

Row Id.	Predicted Value	Actual Value	Residual
2	25.03555247	21.6	-3.435552468
3	30.1845219	34.7	4.515478101
7	23.39322259	22.9	-0.493222593
8	19.58824389	27.1	7.511756109
11	18.83048747	15	-3.830487466
13	21.20113865	21.7	0.498861352
14	19.81376359	20.4	0.586236414
15	19.42217211	18.2	-1.222172107
16	19.63108414	19.9	0.268915856

Example: MLR for Predicting Boston Housing (XLMiner) (cont.)

Summary of errors

- Error = actual - predicted
- RMS = Root-mean-squared error (Square root of average squared error)
- In example, sizes of training and validation sets differ, so only RMS Error and Average Error are comparable

Training Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
6977.106	4.790720883	3.11245E-07

Validation Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
4251.582211	4.587748542	-0.011138034

Using Excel and XLMiner for Data Mining

- Excel is limited in data capacity
- However, the training and validation of DM models can be handled within the modest limits of Excel and XLMiner
- Models can then be used to score larger databases
- XLMiner has functions for interacting with various databases (taking samples from a database, and scoring a database from a developed model)

Summary

- Data Mining consists of supervised methods (Classification & Prediction) and unsupervised methods (Association Rules, Data Reduction, Data Exploration & Visualization)
- Before algorithms can be applied, data must be characterized and pre-processed
- To evaluate performance and to avoid overfitting, data partitioning is used
- Data mining methods are usually applied to a sample from a large database, and then the best model is used to score the entire database