Discovering Knowledge in Data

Daniel T. Larose, Ph.D.

Chapter 1
Introduction to Data Mining

Prepared by James Steck and Eric Flores

Discovering Knowledge in Data: An Introduction to Data Mining, Second Edition,
by Daniel Larose and Chantal Larose, John Wiley and Sons, Inc., 2014.

What is Data Mining?

- Data mining is the process of discovering useful patterns and trends in large data sets
  - According to Gartner Group
    - Most American companies with more that 1000 employees have 200TB of data, growing 40% annually
    - Using data mining, retailers could expect to realize an increase in their operating margin of more than 60%
    - Healthcare providers and HMOs (insurance) could achieve $300 million in cost savings annually, through improved efficiency and quality
  - In United States 2012 Presidential Elections (source: MIT Technology Review), Obama campain
    - First identified likely Obama voters using a data mining model, and then made sure that these voters actually got to the polls
    - used a separate data mining model to predict the polling outcomes county-by-county
    - Hamilton, Ohio: the model predicted 56.4% for Obama; actual result was 56.6%, so that the prediction was off by only 0.02%

Discovering Knowledge in Data: An Introduction to Data Mining, Second Edition,
by Daniel Larose and Chantal Larose, John Wiley and Sons, Inc., 2014.

Why Data Mining? *(cont'd)*

- Other examples
  - Bank of America, West Coast customer service call center (source: *CIO Magazine*)
    - 13 million customer calls per month – in the past they all were offered the same products/services
    - Now, with access to customer's individual profile, customer service representatives offer new products or services that may be of greatest interest to an individual customer
  - Supermarkets
    - Each cash-register product scan collected helps to build a profile about the shopping habits of your family, and the other families who are checking out

Discovering Knowledge in Data: An Introduction to Data Mining, Second Edition,
by Daniel Larose and Chantal Larose, John Wiley and Sons, Inc., 2014.

"We are drowning in information but starved for knowledge."
Megatrends, John Naisbitt

Wanted: Data Miners

- We are inundated with data in most fields, but…
- There are not enough trained human analysts available who are skilled to convert the data into knowledge
- According to McKinsey Report
  - "There will be a shortage of talent…"
  - "…particularly of people with deep expertise in statistics and machine learning, and the managers and analysts who know how to operate companies by using insights from big data."
  - Demand for talent to exceed supply "…by 140,000 to 190,000 positions"
  - "… we project a need for 1.5 million additional managers and analysts in the United States"

- Factors
  - Explosive growth in data collection, as in supermarket scanners
  - Storing the data in data warehouses
  - Increased access to data from web navigation and intranets
  - Competitive pressure to increase market share in globalized economy
  - Growth of computing power and storage capacity

Discovering Knowledge in Data: An Introduction to Data Mining, Second Edition, by Daniel Larose and Chantal Larose, John Wiley and Sons, Inc., 2014.

The Need for Human Direction of Data Mining

- Some early data mining definitions described process as "automatic"
- "…this has misled many people into believing data mining is a product that can be bought rather than a discipline that must be mastered." (Berry, Linoff)
- Automation is no substitute for human input
- Data mining is easy to do badly
- Understanding statistical and mathematical model structures of underlying software required
- Humans need to be actively involved in every phase of data mining process
- Task of data mining should be integrated into human process of problem solving

Discovering Knowledge in Data: An Introduction to Data Mining, Second Edition,
by Daniel Larose and Chantal Larose, John Wiley and Sons, Inc., 2014.

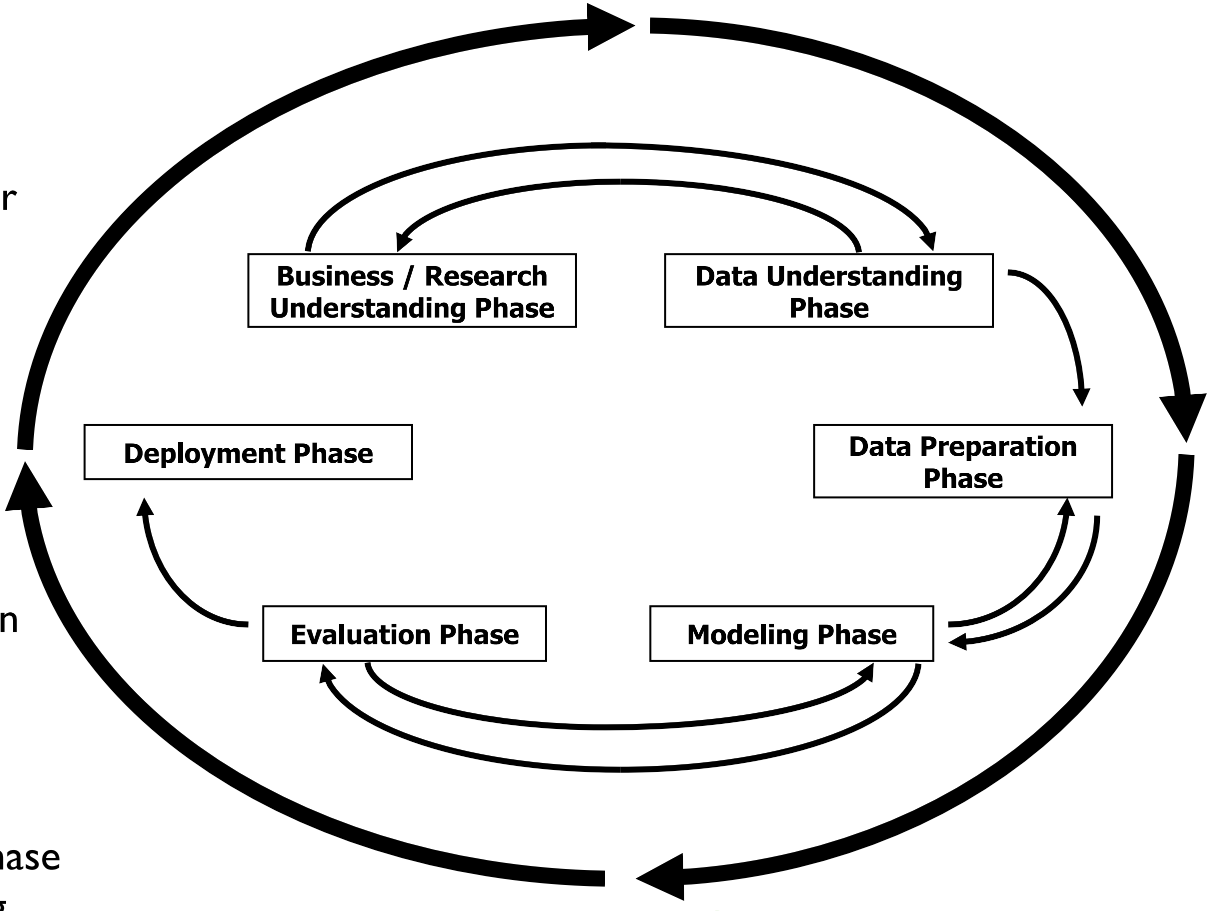
Cross Industry Standard Process: CRISP-DM

- Cross-Industry Standard Process for Data Mining (CRISP-DM) developed in 1996
  - Fits data mining into the general problem-solving strategy of business/research unit
  - Contributors include DaimlerChrysler, SPSS, and NCR
  - Industry, tool and application neutral
  - Non-proprietary and freely available
  - Data mining projects follow iterative, adaptive life cycle consisting of 6 phases
  - Phase sequences are <u>adaptive</u> to your specific application

Discovering Knowledge in Data: An Introduction to Data Mining, Second Edition,
by Daniel Larose and Chantal Larose, John Wiley and Sons, Inc., 2014.

Cross Industry Standard Process: CRISP-DM *(cont'd)*

- Iterative CRIP-DM process shown in outer circle
- Most significant dependencies between phases are shown
- Next phase depends on results from preceding phase
- Returning to earlier phase possible before moving forward



CRISP-DM Lifecycle

Discovering Knowledge in Data: An Introduction to Data Mining, Second Edition, by Daniel Larose and Chantal Larose, John Wiley and Sons, Inc., 2014.

Cross Industry Standard Process: CRISP-DM *(cont'd)*

- (1) Business/Research Understanding Phase
  - Define project requirements and objectives
  - Translate objectives into data mining problem definition
  - Prepare preliminary strategy to meet objectives
- (2) Data Understanding Phase
  - Collect data
  - Perform exploratory data analysis (EDA)
  - Assess data quality
  - Optionally, select interesting subsets
- (3) Data Preparation Phase
  - Prepares data for modeling in subsequent phases
  - Select cases and variables appropriate for analysis
  - Cleanse and prepare data so it is ready for modeling tools
  - Perform transformation of certain variables, if needed

Discovering Knowledge in Data: An Introduction to Data Mining, Second Edition,
by Daniel Larose and Chantal Larose, John Wiley and Sons, Inc., 2014.

Cross Industry Standard Process: CRISP-DM *(cont'd)*

- (4) Modeling Phase
  - Select and apply one or more modeling techniques
  - Calibrate model settings to optimize results
  - If necessary, additional data preparation may be required for supporting a particular technique
- (5) Evaluation Phase
  - Evaluate one or more models for effectiveness
  - Determine whether defined objectives achieved
  - Establish whether some important facet of the problem has not been sufficiently accounted for
  - Make decision regarding data mining results before deploying to field

Cross Industry Standard Process: CRISP-DM *(cont'd)*

- (6) Deployment Phase
  - Make use of models created
  - Simple deployment example: generate a report
  - Complex deployment example: implement parallel data mining effort in another department
  - In businesses, customer often carries out deployment based on your model

Discovering Knowledge in Data: An Introduction to Data Mining, Second Edition,
by Daniel Larose and Chantal Larose, John Wiley and Sons, Inc., 2014.

Fallacies of Data Mining

- Four Fallacies of Data Mining (Louie, Nautilus Systems, Inc.)

	Fallacy	Reality
1	• Set of tools can be turned loose on data repositories • Finds answers to all business problems	• No automatic data mining tools solve problems • Rather, data mining is a process (CRISP-DM) • Integrates into overall business objectives
2	• Data mining process is autonomous • Requires little oversight	• Requires significant intervention during every phase • After model deployment, new models require updates • Continuous evaluation of quality measures by analysts
3	• Data mining quickly pays for itself	• Return rates vary • Depending on startup, personnel, data preparation costs, etc.
4	• Data mining software easy to use	• Ease of use varies across projects • Analysts must combine subject matter knowledge with specific problem domain

Discovering Knowledge in Data: An Introduction to Data Mining, Second Edition,
by Daniel Larose and Chantal Larose, John Wiley and Sons, Inc., 2014.

Fallacies of Data Mining (cont'd)

- Other Fallacies of Data Mining (Larose)

	Fallacy	Reality
5	• Data mining identifies causes of business problems	• Knowledge discovery process uncovers patterns of behavior • Humans interpret results and identify causes
6	• Data mining automatically cleans data in databases	• Data mining often uses data from out-dated systems • Data possibly not examined or used in years • Organizations starting data mining efforts confronted with huge data preprocessing task
7	• Data mining always provides positive results.	• There is no guarantee of positive results • But used properly, data mining can provide actionable and highly profitable results.

Discovering Knowledge in Data: An Introduction to Data Mining, Second Edition, by Daniel Larose and Chantal Larose, John Wiley and Sons, Inc., 2014.

What Tasks Can Data Mining Accomplish?

- Six common data mining tasks
  - Description
  - Estimation
  - Prediction
  - Classification
  - Clustering
  - Association

Discovering Knowledge in Data: An Introduction to Data Mining, Second Edition,
by Daniel Larose and Chantal Larose, John Wiley and Sons, Inc., 2014.

What Tasks Can Data Mining Accomplish? *(cont'd)*

1. Description
   - Describes patterns or trends in data
     - For example, polls may uncover patterns suggesting those presidents laid-off less likely to support the incumbent. President.
     - Descriptions of patterns, often suggest possible explanations
     - For example, those laid-off now are less financially secure; therefore, prefer alternate candidate
   - Data mining models should be transparent
     - That is, results should be interpretable by humans
     - Some data mining methods more transparent than others
     - For example, Decision Trees (transparent) > Neural Networks (non-transparent)
   - High-quality description accomplished using Exploratory Data Analysis (EDA)
     - Graphical method of exploring patterns and trends in data

Discovering Knowledge in Data: An Introduction to Data Mining, Second Edition,
by Daniel Larose and Chantal Larose, John Wiley and Sons, Inc., 2014.

What Tasks Can Data Mining Accomplish? *(cont'd)*

2. Estimation (1/3)

- Similar to Classification task, except target variable is numeric
- Models built from complete data records
  - Records include values for each predictor field and numeric target variable in training set
- For new observations, estimate the target variable

- Example: Estimate a patient's systolic blood pressure, based on patient's age, gender, body-mass index, and sodium levels
  - a) Use training data to develop model that estimates systolic blood pressure based on predictor variables
  - b) Apply model to new cases, to obtain estimated systolic blood pressure

Discovering Knowledge in Data: An Introduction to Data Mining, Second Edition, by Daniel Larose and Chantal Larose, John Wiley and Sons, Inc., 2014.

What Tasks Can Data Mining Accomplish? *(cont'd)*

2. Estimation (2/3) – Further examples
   - Estimate amount of money, family of four will spend on back-to-school shopping
   - Estimate percentage decrease in rotary movement sustained to NFL player with knee injury
   - Estimate number of points basketball player scores when double-teamed in playoffs
   - Estimate GPA of graduate student, based on student's undergraduate GPA

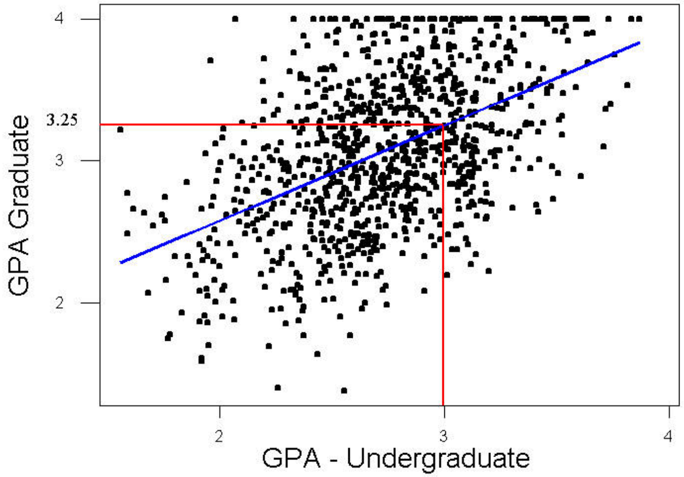
<u>Statistical Analysis</u> uses several estimation methods: point estimation, confidence interval estimation, linear regression and correlation, and multiple regression

Discovering Knowledge in Data: An Introduction to Data Mining, Second Edition,
by Daniel Larose and Chantal Larose, John Wiley and Sons, Inc., 2014.

What Tasks Can Data Mining Accomplish? *(cont'd)*

2. Estimation (3/3) – continued

- Figure 1.2 shows scatter plot of graduate GPA against undergraduate GPA (1000 students)
- Linear regression finds line (blue) best approximating relationship between two variables



- Regression line estimates student's graduate GPA based on their undergraduate GPA, resulting in the following model:

$$\hat{y} = 1.24 + 0.67x$$

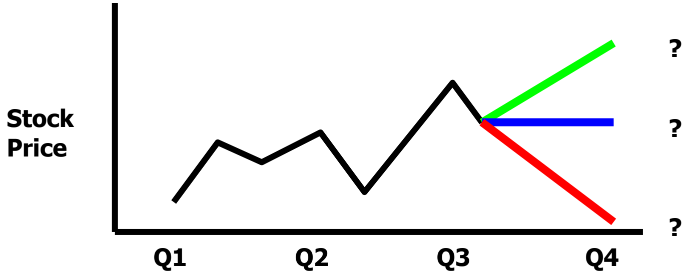
- For example, suppose student's undergraduate GPA = 3.0
- According to estimation model, estimated student's graduate GPA = 1.24 + 0.67(3.0) = 3.25
- Point ($x = 3.0$, $\hat{y} = 3.25$) lies on regression line

Discovering Knowledge in Data: An Introduction to Data Mining, Second Edition, by Daniel Larose and Chantal Larose, John Wiley and Sons, Inc., 2014.

What Tasks Can Data Mining Accomplish? *(cont'd)*

3. Prediction
   - Similar to classification and estimation, except results lie in the future
   - Methods used for classification and prediction applicable to prediction
     - Includes point estimation, confidence interval estimation, linear regression and correlation, multiple regression, k-nearest neighbor, decision trees and neural networks

- Example prediction tasks in business and research:



- Predict price of stock 3 months into future, based on past performance
- Predict percentage increase in traffic deaths next year, if speed limit increased
- Predicting the winner of this fall' s World Series, based on a comparison of the team statistics
- Predict whether molecule in newly discovered drug leads to profitable pharmaceutical drug

Discovering Knowledge in Data: An Introduction to Data Mining, Second Edition, by Daniel Larose and Chantal Larose, John Wiley and Sons, Inc., 2014.

What Tasks Can Data Mining Accomplish? *(cont'd)*

3. Classification (1/5)

- Similar to Estimation task, except target variable is <u>categorical</u>
- Models built from complete data records
  - Records include values for each predictor field and <u>categorical</u> target variable in training set (rather than numeric)
- For <u>new observations</u>, estimate the target variable

- Example: Classify the Income Bracket of an individual as Low, Middle or High based their Age, Gender and Occupation
  a) Use training data to develop model that classifies Income Bracket based on predictor variables
  b) Apply model to cases not currently in the database, to obtain estimated Income Bracket classification

Discovering Knowledge in Data: An Introduction to Data Mining, Second Edition, by Daniel Larose and Chantal Larose, John Wiley and Sons, Inc., 2014.

What Tasks Can Data Mining Accomplish? *(cont'd)*

3. Classification (2/5) – Example in detail

- Using the training data set, the algorithm would:
  - Examine the data set containing both the predictor variables and the (already classified) target variable, *income bracket*
  - Algorithm (software) "learns about" which combinations of variables are associated with which income brackets (for example, Older females -> High Income)
- Then, when looking at new records with no income information, the algorithm would:
  - Based on the classification in the training set, would assign classifications to the new records (for example, 63-year-old female professor -> high)

Subject	Age	Gender	Occupation	Income Bracket
001	47	F	Software Engineer	High
002	28	M	Marketing Consultant	Middle
003	35	M	Unemployed	Low
…	…	…	…	…

Discovering Knowledge in Data: An Introduction to Data Mining, Second Edition, by Daniel Larose and Chantal Larose, John Wiley and Sons, Inc., 2014.

What Tasks Can Data Mining Accomplish? *(cont'd)*

3. Classification (3/5) – Further examples in Business and Research

- Determining whether a particular credit card transaction is fraudulent
- Placing a new student into a particular track with regard to special needs
- Assessing whether a mortgage application is a good or bad credit risk
- Diagnosing whether a particular disease is present
- Determining whether a will was written by the actual deceased, or fraudulently by someone else,
- Identifying whether or not certain financial or personal behavior indicates a possible terrorist threat

Discovering Knowledge in Data: An Introduction to Data Mining, Second Edition, by Daniel Larose and Chantal Larose, John Wiley and Sons, Inc., 2014.

What Tasks Can Data Mining Accomplish? *(cont'd)*

3. Classification (4/5) – The drug prescription example

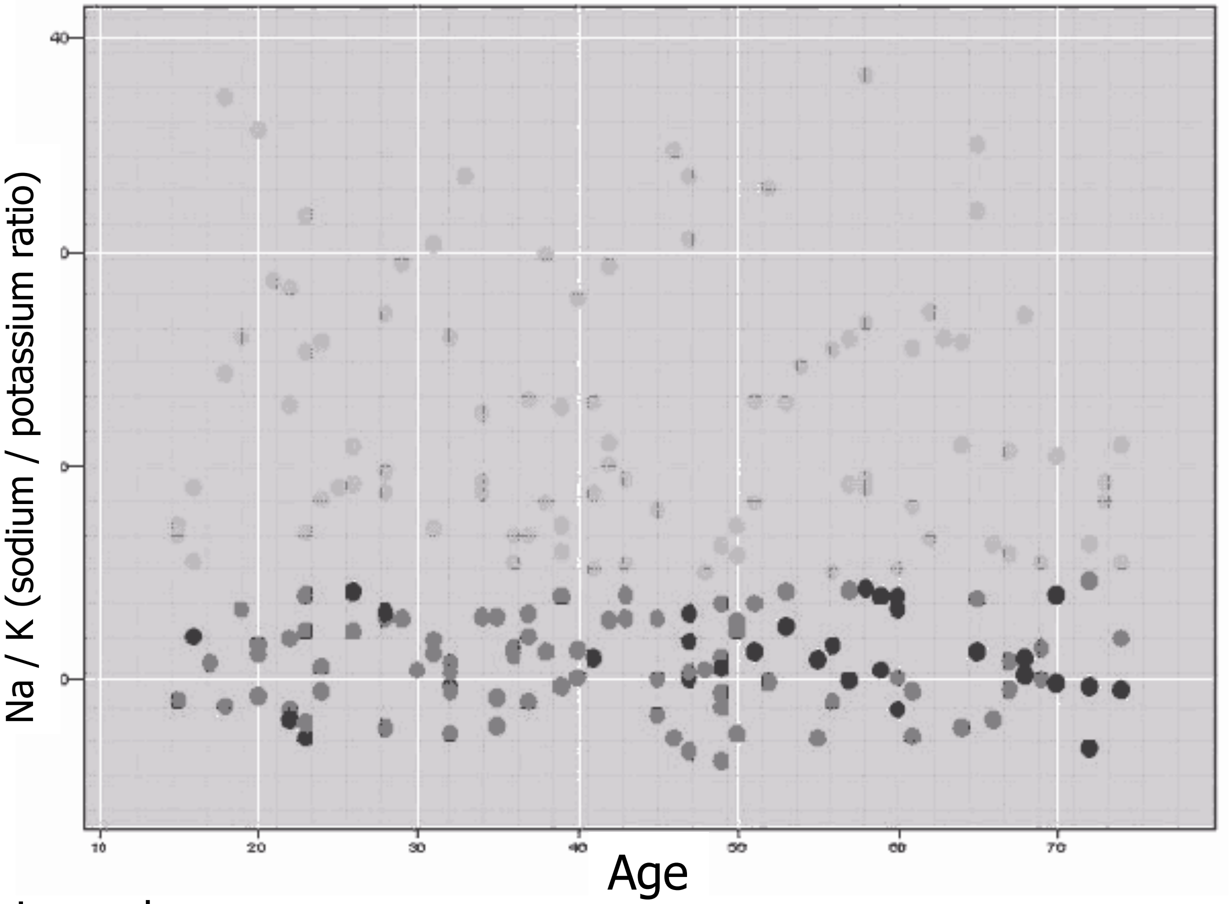
- Interested in classifying the type of drug a patient should be prescribed, based on age of the patient, and the patient's sodium / potassium ratio
- Next slide presents scatter plot of 200 patients with their sodium/potassium ratios against age, and the particular drug prescribed by the shade of the points
- What drug should be prescribed for:

- Young patient with high Na/K ratio?
  - Young patients with high Na/K are in the upper left region
  - Past patients in this region got Drug Y
  - The recommended classification for such patients is Drug Y

- Older patient with low Na/K ratio?
  - Lower right region
  - Past patients in this region got either dark gray (Drugs B or C) or medium gray (Drugs A or X).
  - Definitive classification not possible without further information

Discovering Knowledge in Data: An Introduction to Data Mining, Second Edition, by Daniel Larose and Chantal Larose, John Wiley and Sons, Inc., 2014.

Figure 1.3 - Which Drug Should Be Prescribed for Which Type of Patient?



Dot Legend:
Light gray – Drug Y
Medium gray – Drugs A or X
Dark gray – Drugs B or C

Discovering Knowledge in Data: An Introduction to Data Mining, Second Edition, by Daniel Larose and Chantal Larose, John Wiley and Sons, Inc., 2014.

What Tasks Can Data Mining Accomplish? *(cont'd)*

4. Classification (5/5) – Handling many predictors

- Classification tasks with 2 or 3 predictors
  - Can be analyzed using charts and plots like the drug example above
- Many datasets have multiple predictors
  - This requires common data mining methods for classification like:
    - k-nearest neighbor (Chapter 7)
    - decision trees (Chapter 8)
    - neural networks (Chapter 9).

Discovering Knowledge in Data: An Introduction to Data Mining, Second Edition,
by Daniel Larose and Chantal Larose, John Wiley and Sons, Inc., 2014.

What Tasks Can Data Mining Accomplish? *(cont'd)*

5. Clustering (1/2)

   - Refers to grouping records into classes of similar objects
   - Cluster – a collection of records similar to one another, and dissimilar to records in other clusters
   - Clustering algorithm seeks to segment data set into homogeneous subgroups
   - Target variable <u>not specified</u>
     - Clustering does not try to classify/estimate/predict target variable

   - For example, Claritas, Inc. PRIZM software clusters demographic profiles for different geographic areas according to zip code
     - It describes every American zip code area in terms of distinct lifestyle types (see next slide for example)

Discovering Knowledge in Data: An Introduction to Data Mining, Second Edition,
by Daniel Larose and Chantal Larose, John Wiley and Sons, Inc., 2014.

Nielsen Claritas' *PRIZM* segmentation system

01 Upper Crust	02 Blue Blood Estates	03 Movers and Shakers
04 Young Digerati	05 Country Squires	06 Winner's Circle
07 Money and Brains	08 Executive Suites	09 Big Fish, Small Pond
10 Second City Elite	11 God's Country	12 Brite Lites, Little City
13 Upward Bound	14 New Empty Nests	15 Pools and Patios
16 Bohemian Mix	17 Beltway Boomers	18 Kids and Cul-de-sacs
19 Home Sweet Home	20 Fast-Track Families	21 Gray Power
22 Young Influentials	23 Greenbelt Sports	24 Up-and-Comers
25 Country Casuals	26 The Cosmopolitans	27 Middleburg Managers
28 Traditional Times	29American Dreams	30 Suburban Sprawl
31 Urban Achievers	32 New Homesteaders	33 Big Sky Families
34 White Picket Fences	35 Boomtown Singles	36 Blue-Chip Blues
37 Mayberry-ville	38 Simple Pleasures	39 Domestic Duos
40 Close-in Couples	41 Sunset City Blues	42 Red, White and Blues
43 Heartlanders	44 New Beginnings	45 Blue Highways
46 Old Glories	47 City Startups	48 Young and Rustic
49 American Classics	50 Kid Country, USA	51 Shotguns and Pickups
52 Suburban Pioneers	53 Mobility Blues	54 Multi-Culti Mosaic
55Golden Ponds	56 Crossroads Villagers	57 Old Milltowns
58 Back Country Folks	59 Urban Elders	60 Park Bench Seniors
61 City Roots	62 Hometown Retired	63 Family Thrifts
64 Bedrock America	65 Big City Blues	66 Low-Rise Living

Table 1.2 The 66 clusters used by the *PRIZM* segmentation system.

- Clusters for zip code 90210, Beverly Hills, California are:
  - #01: Upper Crust Estates
  - #03: Movers and Shakers
  - #04:Young Digerati
  - #07: Money and Brains
  - #16: Bohemian Mix
- The description for Cluster # 01: Upper Crust
  - The nation's most exclusive address
  - the wealthiest lifestyle in America
  - Haven for empty-nesting couples between the ages of 45 and 64
  - highest concentration of residents with:
    - over $100,000/year
    - Most opulent standard of living.

Discovering Knowledge in Data: An Introduction to Data Mining, Second Edition,
by Daniel Larose and Chantal Larose, John Wiley and Sons, Inc., 2014.

What Tasks Can Data Mining Accomplish? *(cont'd)*

5. Clustering (2/2) - Clustering Tasks in Business and Research:
   - Target marketing niche product for small business that does not have large marketing budget
   - For accounting purposes, to segmentize financial behavior into benign and suspicious categories
   - Use as dimensionality-reduction tool for data set having several hundred inputs
   - For gene expression clustering, where very large quantities of genes may exhibit similar behavior
   - As preliminary step in data mining
     - Resulting clusters used as input to different technique downstream, such as neural networks

See more about hierarchical and k-means clustering in Chapter 10, and Kohonen networks in Chapter 11

What Tasks Can Data Mining Accomplish? *(cont'd)*

6. Association (1/2)
   - Find out which attributes "go together"
   - Commonly used for Market Basket Analysis (aka Affinity Association)
   - Quantify relationships between two or more attributes in the form of rules as:

     IF *antecedent* THEN *consequent*

   - Rules measured using support and confidence:
     - Support = P(antecedent), confidence = P(consequent|antecedent)
   - Example: A particular supermarket might find that:
     - Thursday night 200 of 1,000 customers bought diapers, and of those buying diapers, 50 purchased beer
     - Association Rule: "IF buy diapers, THEN buy beer"
     - Support = 200/1,000 = 5%, and confidence = 50/200 = 25%

Discovering Knowledge in Data: An Introduction to Data Mining, Second Edition, by Daniel Larose and Chantal Larose, John Wiley and Sons, Inc., 2014.

What Tasks Can Data Mining Accomplish? (cont'd)

6. Association (2/2) - Association Tasks in Business and Research:
   - Investigating the proportion of subscribers to your company's cell phone plan that respond positively to an offer of a service upgrade,
   - Examining the proportion of children whose parents read to them who are themselves good readers,
   - Predicting degradation in telecommunications networks,
   - Finding out which items in a supermarket are purchased together, and which items are never purchased together,
   - Determining the proportion of cases in which a new drug will exhibit dangerous side effects.

The *a priori* and the *GRI* association rule algorithms are visited in Chapter 12

Discovering Knowledge in Data: An Introduction to Data Mining, Second Edition, by Daniel Larose and Chantal Larose, John Wiley and Sons, Inc., 2014.