Data Mining: Introduction and a Health Care Application

Prem Swaroop (pswaroop@rhsmith.umd.edu); Dr Bruce Golden (bgolden@rhsmith.umd.edu)
Robert H Smith School of Business, University of Maryland, College Park

Case: Medicare and Hospitalization Costs

Medicare has announced that it will no longer reimburse hospitals for errors, and for nosocomial infections. To reduce the occurrence of resistant infections, one 1000-bed hospital wants to implement a protocol to prevent them. High-risk patients admitted for elective surgery will be identified. They will be admitted to the hospital 24 hours prior to surgery (the usual protocol has them admitted after surgery) and placed on IV vancomycin. The antibiotic will be continued until discharge.

Risk is not uniform and is based on a combination of patient demographics, diagnoses, and procedures. Further, the cost to treat infection is not uniform across patients. A superficial wound generally will not add to the overall length of stay. A deep skin wound requires ten days to three weeks of IV antibiotics (the date of discharge will vary). An infection in the bone requires six weeks of antibiotics, and carries the additional risk to the patient of limb amputation or death. A resistant infection in the lungs is life-threatening and the patient will be moved to ICU where the daily costs to treat increase substantially.

Case: Breast Cancer Detection

- *Breast cancer* is a disease in which malignant (cancer) cells form in the tissues of the breast. Breast cancer is the second leading cause of cancer deaths in women today (after lung cancer) and is the most common cancer among women, except for skin cancers. About 1.3 million women are expected to be diagnosed annually with breast cancer worldwide, and about 465,000 will die from the disease. In the United States alone, in 2007 an estimated 240,510 women were expected to be diagnosed with breast cancer, and 40,460 women are expected to have died from breast cancer.
- *Screening* is looking for cancer in asymptomatic people – i.e., before a person has any symptoms of the disease. Cancer screening can help find cancer at an early stage. When abnormal tissue or cancer is found early, it is often easier to treat. By the time symptoms appear, cancer may have begun to spread. The good news is that breast cancer death rates have been dropping steadily since 1990, both because of earlier detection via screening and better treatments.

Steps for confirming cancer:

1. Screening mammography
2. Diagnostic mammography (5-10% of screened women)
3. Biopsy (3-10 cases per 1000 screened women)

Undetected cancers at screening stage

- Multiple reasons
- Double reading implemented at many sites
- Increased detection by 4-15%; expensive

Computer Aided Detection

- High rate of detection
- Improvements would help tremendously in:
  - Sensitivity – (Pred) True Positive/Act. Positive
  - Specificity – (Pred) False Positive/Act. Negative

4 stage system of CAD

1. Candidate generation
2. Feature extraction
3. Classification (detection)
4. Visual presentation

Data Mining: Introduction

- Motivation: Why mine data
- Definition: What is data mining
- Functionality: Key data mining tasks
- Classification: Multi-dimensional view
- Techniques and Example Applications
- Challenges
- Software Tools
- References

Motivation: Why mine data – commercial viewpoint

- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - purchases at department/grocery stores
  - Bank/Credit Card transactions
- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
  - Provide better, customized services for an *edge* (e.g. in Customer Relationship Management)

Motivation: Why mine data – scientific viewpoint

- Data collected and stored at enormous speeds (GB/hour)
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating gene expression data
  - scientific simulations generating terabytes of data
- Traditional techniques infeasible for raw data
- Data mining may help scientists
  - in classifying and segmenting data
  - in Hypothesis Formation

Motivation: Why mine data – Summary

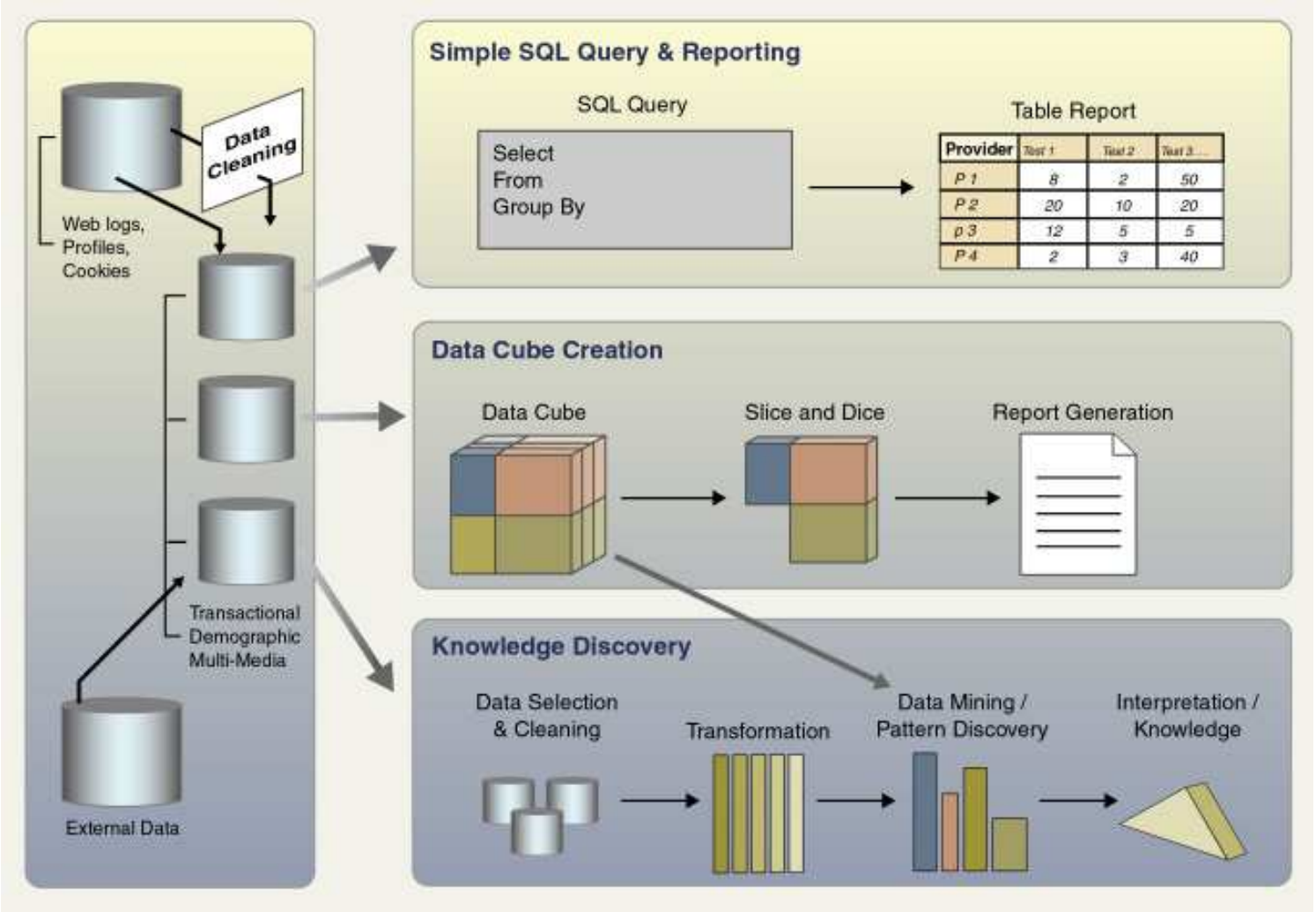
- There is often information “hidden” in the data that is not readily evident
- Human analysts may take weeks to discover useful information
- Much of the data is never analyzed at all
- We are drowning in data, but starving for knowledge!
- “Necessity is the mother of invention”—Data mining—

  Automated analysis of massive data sets

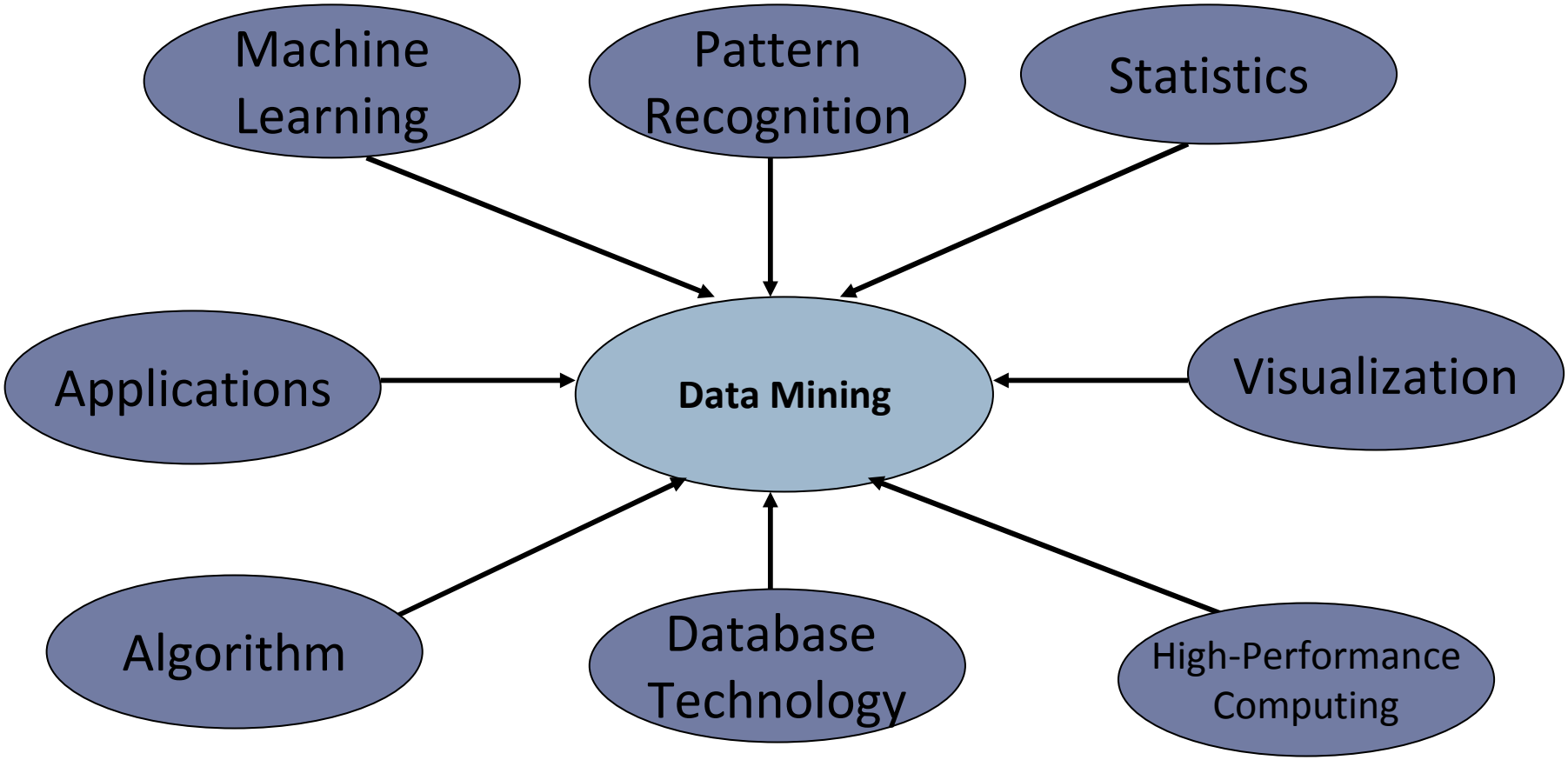
Definition: What is data mining

- Many Definitions
- Knowledge discovery from data
- Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
- Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns
- Alternative names:
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

Definition: What is data mining – Computational Knowledge Discovery

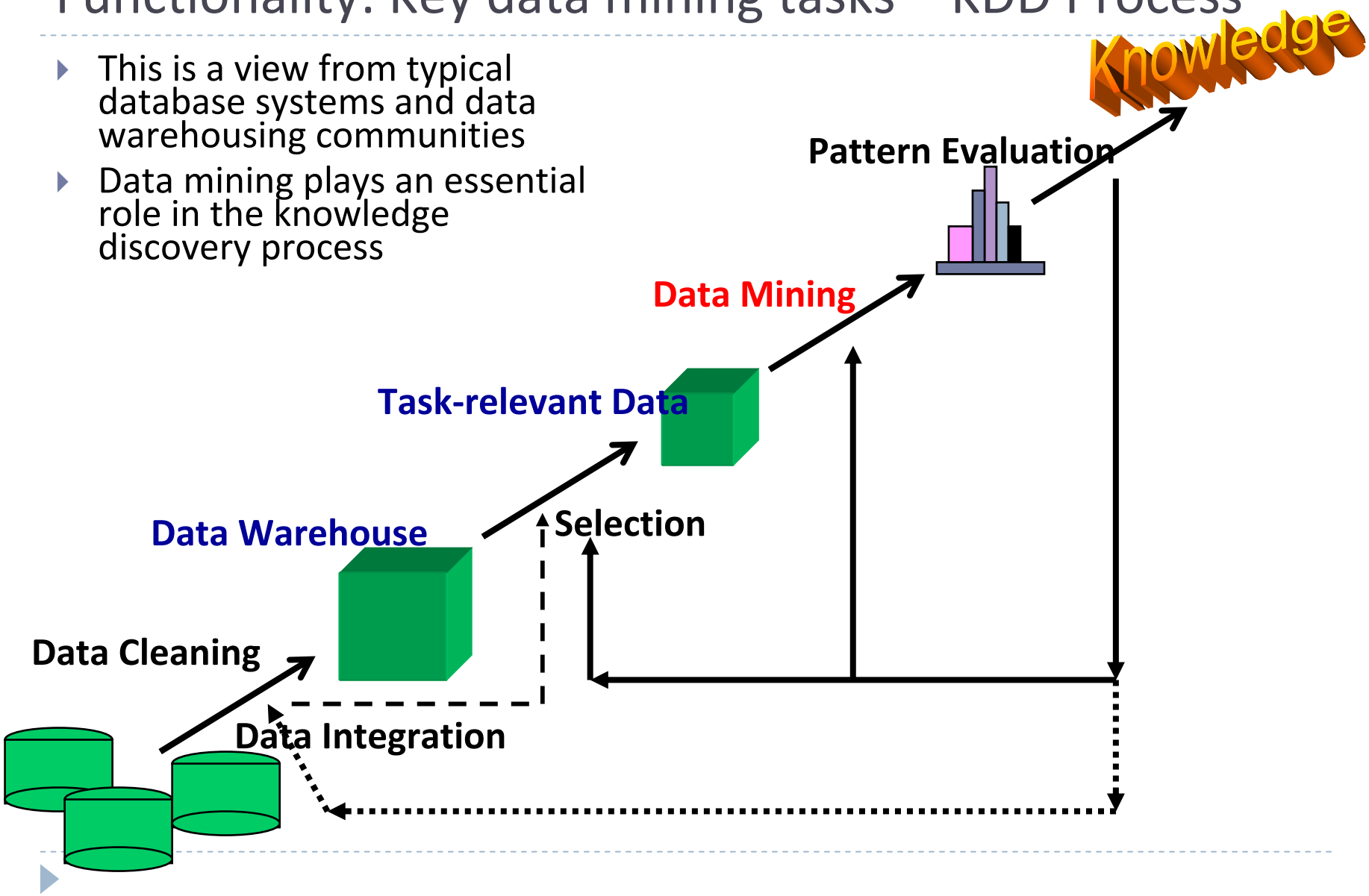


Definition: What is data mining – Origins

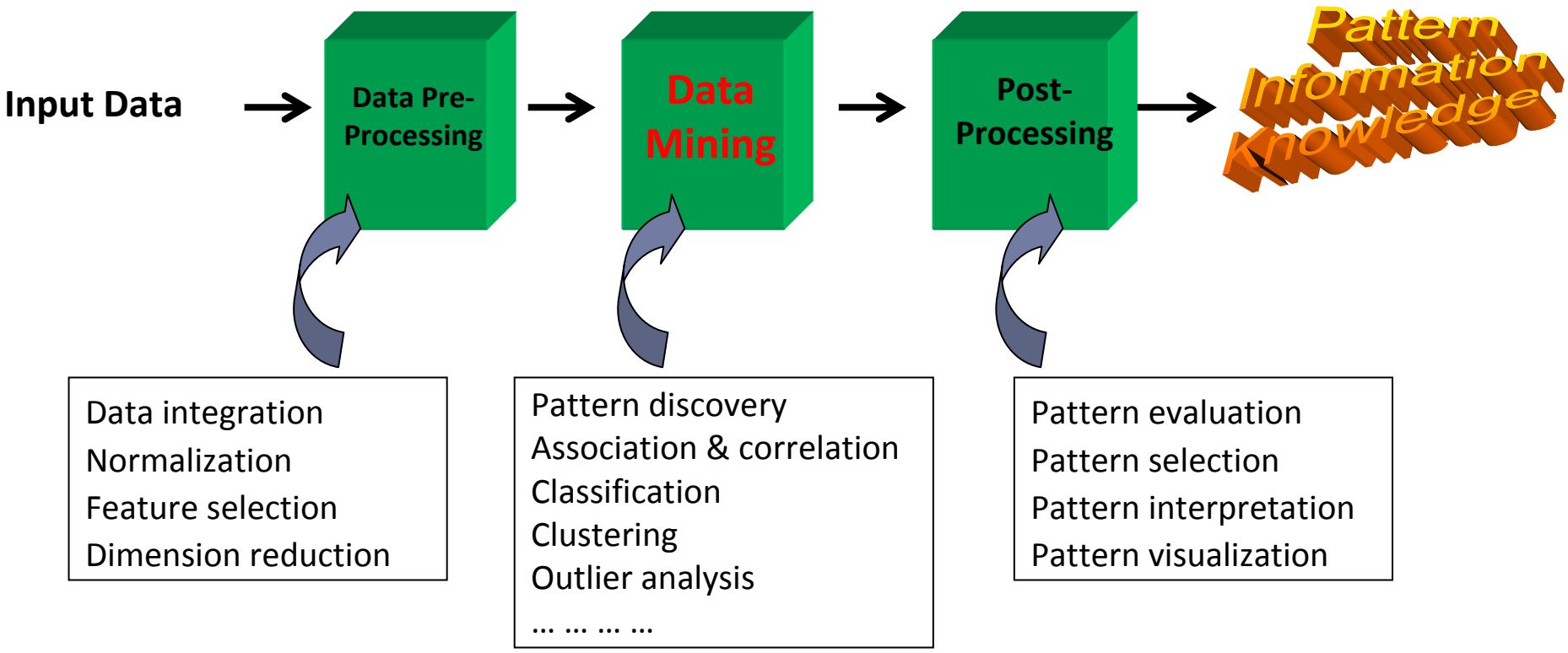


Functionality: Key data mining tasks – KDD Process

- This is a view from typical database systems and data warehousing communities
- Data mining plays an essential role in the knowledge discovery process

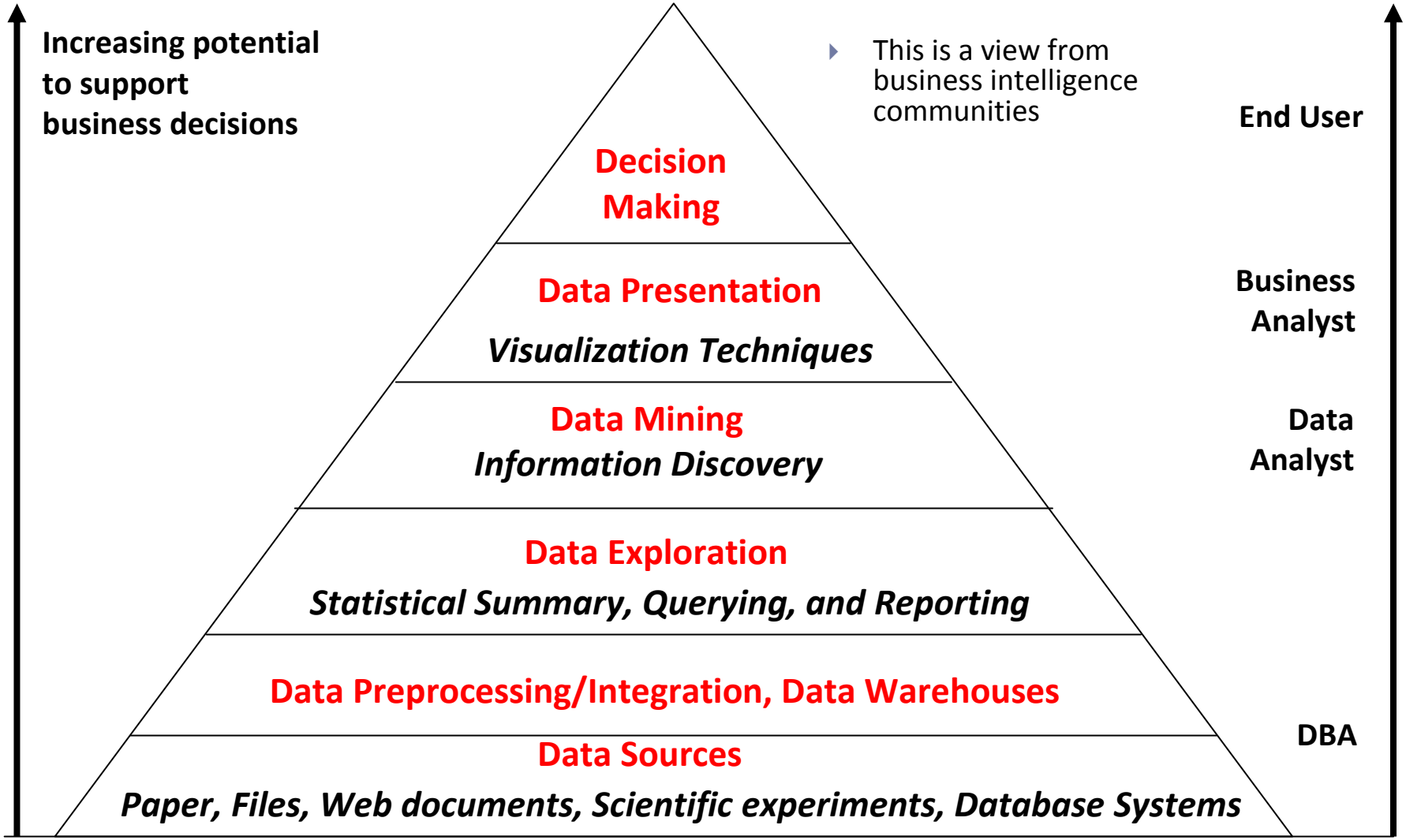


Functionality: Key data mining tasks – KDD Process



- This is a view from typical machine learning and statistics communities

Functionality: Key data mining tasks – KDD Process



Classification: Multi-dimensional view

- General functionality
  - Descriptive data mining
  - Predictive data mining
- Different views lead to different classifications
  - Data view: Kinds of data to be mined
  - Knowledge view: Kinds of knowledge to be discovered
  - Method view: Kinds of techniques utilized
  - Application view: Kinds of applications adapted

Classification: Multi-dimensional view

- **Data to be mined**
  - Relational, data warehouse, transactional, stream, object-oriented/relational, active, spatial, time-series, text, multi-media, heterogeneous, legacy, WWW
- **Knowledge to be mined**
  - Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, etc.
  - Multiple/integrated functions and mining at multiple levels
- **Techniques utilized**
  - Database-oriented, data warehouse (OLAP), machine learning, statistics, visualization, etc.
- **Applications adapted**
  - Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.

Techniques and Example Applications

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation Detection [Predictive]
- Structure and Network Analysis

Techniques – Classification

- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

Example Application – Classification

- Direct Marketing
  - Goal: Reduce cost of mailing by *targeting* a set of consumers likely to buy a new cell-phone product.
  - Approach:
    - Use the data for a similar product introduced before.
    - We know which customers decided to buy and which decided otherwise. This *{buy, don't buy}* decision forms the *class attribute*.
    - Collect various demographic, lifestyle, and company-interaction related information about all such customers.
      - Type of business, where they stay, how much they earn, etc.
    - Use this information as input attributes to learn a classifier model.

Techniques – Clustering

- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- Similarity Measures:
  - Euclidean Distance if attributes are continuous.
  - Other Problem-specific Measures.

Example Application – Clustering

- Market Segmentation:
  - Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
  - Approach:
    - Collect different attributes of customers based on their geographical and lifestyle related information.
    - Find clusters of similar customers.
    - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Technique & Example Application – Association Rules

- Given a set of records each of which contain some number of items from a given collection;
  - Produce dependency rules which will predict occurrence of an item based on occurrences of other items.
- Marketing and Sales Promotion:
  - Let the rule discovered be

    {Bagels, … } --> {Potato Chips}
  - Potato Chips as consequent => Can be used to determine what should be done to boost its sales.
  - Bagels in the antecedent => Can be used to see which products would be affected if the store discontinues selling bagels.
  - Bagels in antecedent *and* Potato chips in consequent => Can be used to see what products should be sold with Bagels to promote sale of Potato chips!

Technique & Example Application – Sequential Patterns

- Given is a set of *objects*, with each object associated with its own *timeline of events*, find rules that predict strong sequential dependencies among different events.
- Rules are formed by first discovering patterns. Event occurrences in the patterns are governed by timing constraints.
- In telecommunications alarm logs,
  - (Inverter_Problem Excessive_Line_Current)
    (Rectifier_Alarm) --> (Fire_Alarm)
- In point-of-sale transaction sequences,
  - Computer Bookstore:
    (Intro_To_Visual_C) (C++_Primer) -->
    (Perl_for_dummies,Tcl_Tk)
  - Athletic Apparel Store:
    (Shoes) (Racket, Racketball) --> (Sports_Jacket)

Technique & Example Application – Regression

- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Greatly studied in statistics, neural network fields.
- Examples:
  - Predicting sales amounts of new product based on advertising expenditure.
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Time series prediction of stock market indices.

Technique & Example Application – Deviation

- Detect significant deviations from normal behavior
- Applications:
  - Credit Card Fraud Detection
  - Network Intrusion Detection
  - Auto insurance: ring of collisions
  - Money laundering: suspicious monetary transactions
  - Medical insurance
    - Professional patients, ring of doctors, and ring of references
    - Unnecessary or correlated screening tests
  - Telecommunications: phone-call fraud
    - Phone call model: destination of the call, duration, time of day or week. Analyze patterns that deviate from an expected norm
  - Retail industry
    - Analysts estimate that 38% of retail shrink is due to dishonest employees
  - Anti-terrorism

Technique – Structure and Network Analysis

- Graph mining
  - Finding frequent sub graphs (e.g., chemical compounds), trees (XML), substructures (web fragments)
- Information network analysis
  - Social networks: actors (objects, nodes) and relationships (edges)
    - e.g., author networks in CS, terrorist networks
  - Multiple heterogeneous networks
    - A person could be multiple information networks: friends, family, classmates, …
  - Links carry a lot of semantic information: Link mining
- Web mining
  - Web is a big information network: from PageRank to Google
  - Analysis of Web information networks
    - Web community discovery, opinion mining, usage mining, …

Challenges in Data Mining

- Efficiency and scalability of data mining algorithms
- Parallel, distributed, stream, and incremental mining methods
- Handling high-dimensionality
- Handling noise, uncertainty, and incompleteness of data
- Incorporation of constraints, expert knowledge, and background knowledge in data mining
- Pattern evaluation and knowledge integration
- Mining diverse and heterogeneous kinds of data: e.g., bioinformatics, Web, software/system engineering, information networks
- Application-oriented and domain-specific data mining
- Invisible data mining (embedded in other functional modules)
- Protection of security, integrity, and privacy in data mining

Data Mining Software Tools

- Top ten most used packages as per KDD Nuggets Survey (May 2007)
  1. SPSS/ SPSS Clementine
  2. Salford Systems CART/MARS/TreeNet/RF
  3. Yale (now Rapid Miner) *(open source)*
  4. SAS / SAS Enterprise Miner
  5. Angoss Knowledge Studio / Knowledge Seeker
  6. KXEN
  7. Weka *(open source)*
  8. R *(open source)*
  9. Microsoft SQL Server
  10. MATLAB
      - Source: the-data-mine.com; kdnuggets.com

References

- Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, 2nd edition, Morgan Kaufmann, 2006
- Pang-Ning Tan, Michael Steinbach and Vipin Kumar, Introduction to Data Mining, Pearson Addison Wesley, 2005
- Ian Witten and Eibe Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd Edition, Morgan Kaufmann, 2005

Cost Model Formulation for Treatment of Patients at Risk of Nosocomial (Hospital-Acquired) Infections using Data Mining

Prem Swaroop

Overview

- Problem Description
- Cost Model
  - Motivation and Definitions
- Formulation
- Parameter Estimation

Problem Description

Medicare has announced that it will no longer reimburse hospitals for errors, and for nosocomial infections. To reduce the occurrence of resistant infections, one 1000-bed hospital wants to implement a protocol to prevent them. High-risk patients admitted for elective surgery will be identified. They will be admitted to the hospital 24 hours prior to surgery (the usual protocol has them admitted after surgery) and placed on IV vancomycin. The antibiotic will be continued until discharge.

Risk is not uniform and is based on a combination of patient demographics, diagnoses, and procedures. Further, the cost to treat infection is not uniform across patients. A superficial wound generally will not add to the overall length of stay. A deep skin wound requires ten days to three weeks of IV antibiotics (the date of discharge will vary). An infection in the bone requires six weeks of antibiotics, and carries the additional risk to the patient of limb amputation or death. A resistant infection in the lungs is life-threatening and the patient will be moved to ICU where the daily costs to treat increase substantially.

Problem Description

Objective:

Given two years of patient data, including whether or not the patient contracted an infection during a surgical procedure, and the cost to treat that infection, determine an optimal strategy for choosing patients from the test group to minimize the total cost of medication.

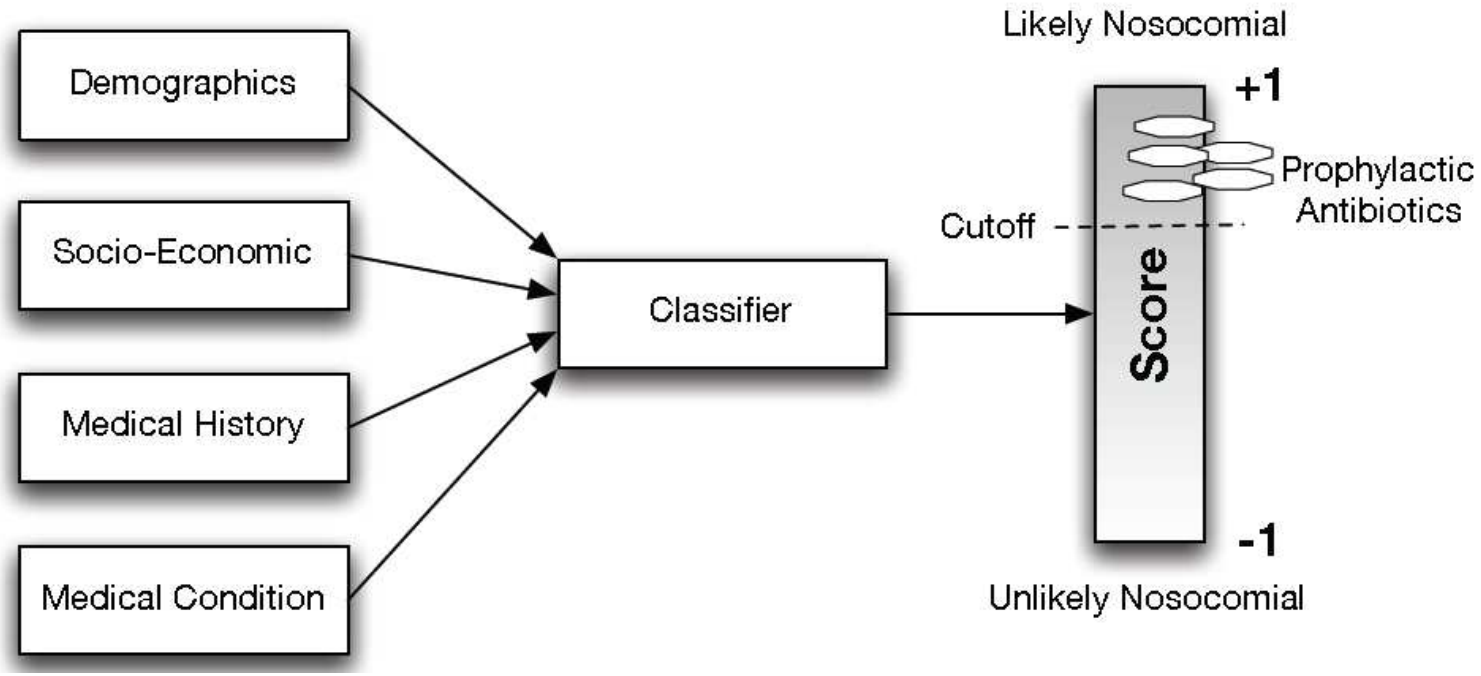
Part 1 (classification): For an unseen patient dataset, determine probability of each patient to be diagnosed with MRSA.
Part 2 (policy): Develop and justify a realistic cost model (including cost of prophylactic treatment, cost of MRSA treatment, and probabilities from your predictive model), and use it to maximize the total cost savings of the proposed strategy.

Problem Description – Summary

- Proposed Strategy
  - Patients to be administered preventive antibiotics 24 hours prior to start of treatment
- Realistic Cost Model
  - Predicted probabilities
  - Costs of treatment
- Policy Design
  - Maximize Total Cost Savings

Classification

Classification



4 datasets

Cleaning:

- Merge
- Remove redundant fields
- Handle noise

Preprocessing:

- Association Rules to identify most frequently occurring diseases in history and current diagnosis together with nosocomial infection

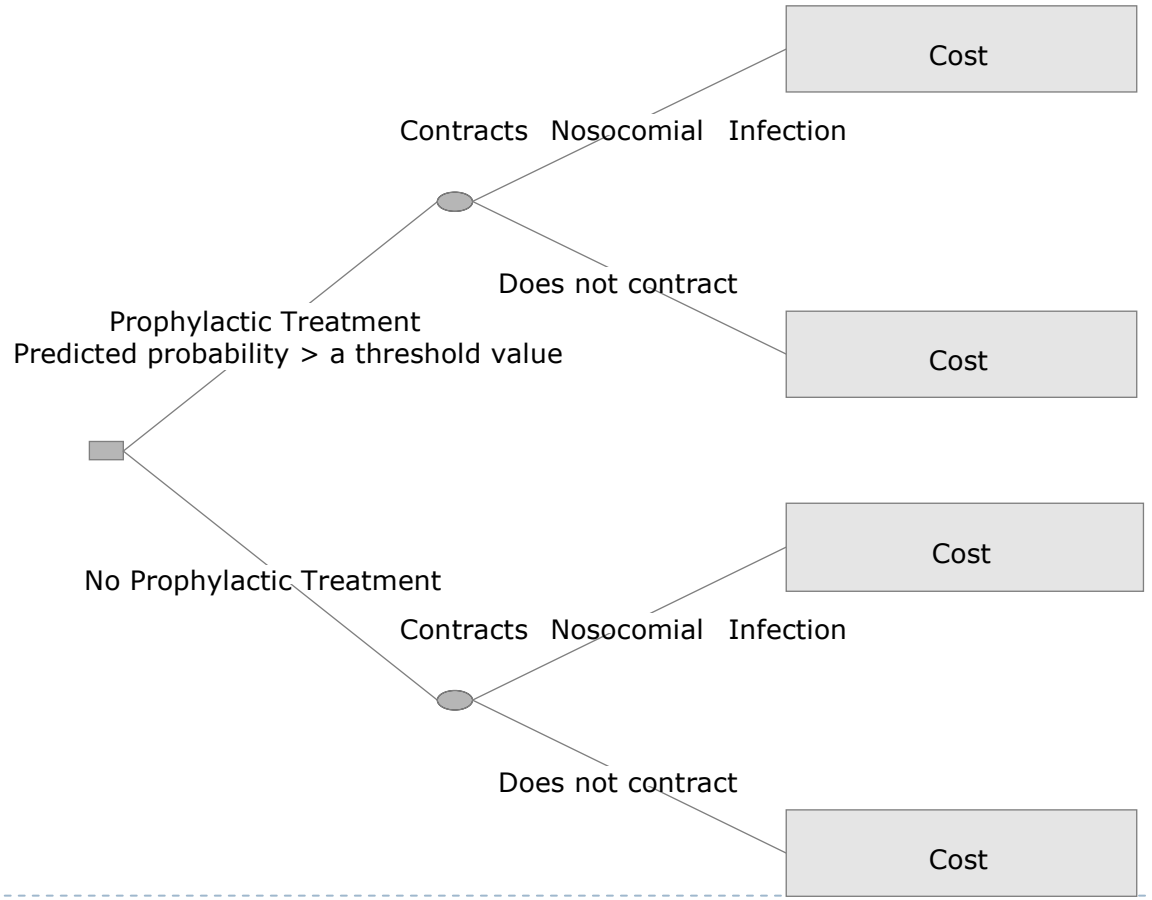
Classification:

- Run multiple algorithms
- Evaluate prediction results
- Select best predicting ones
- *Support Vector Machines*

Policy

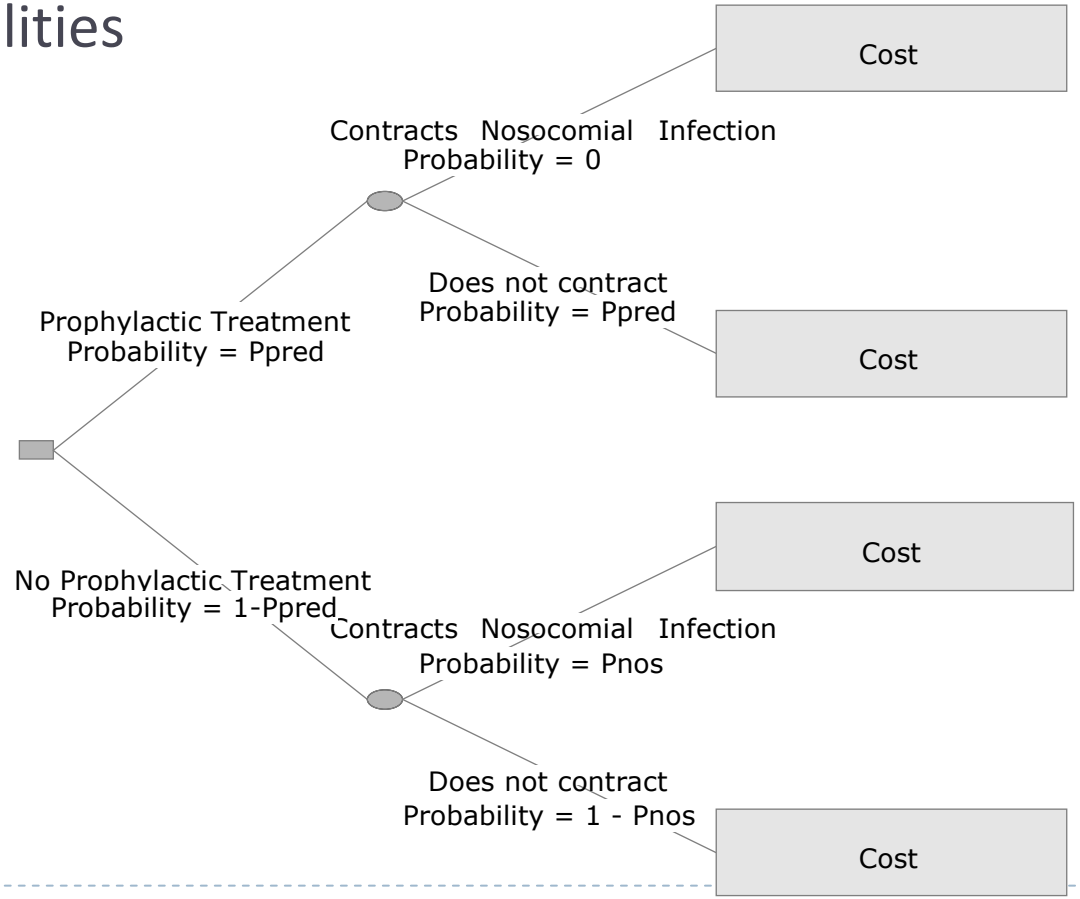
Cost Model

- Motivation



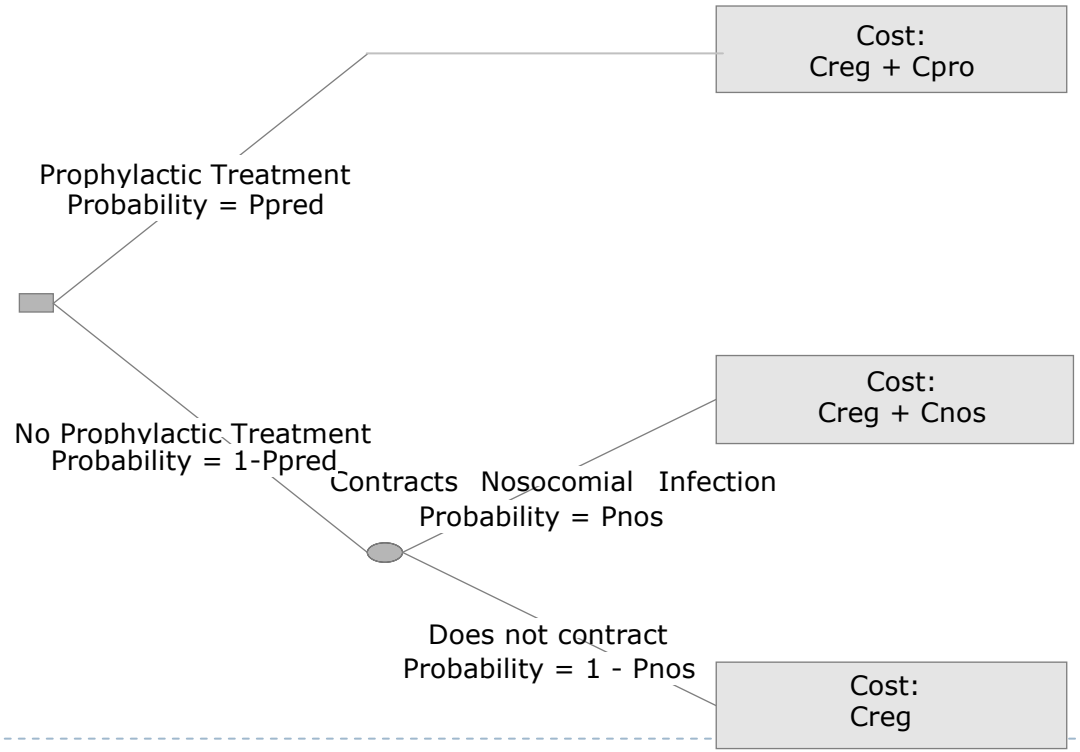
Cost Model

- Definitions
  - Probabilities



Cost Model

- Definitions
  - Total Costs

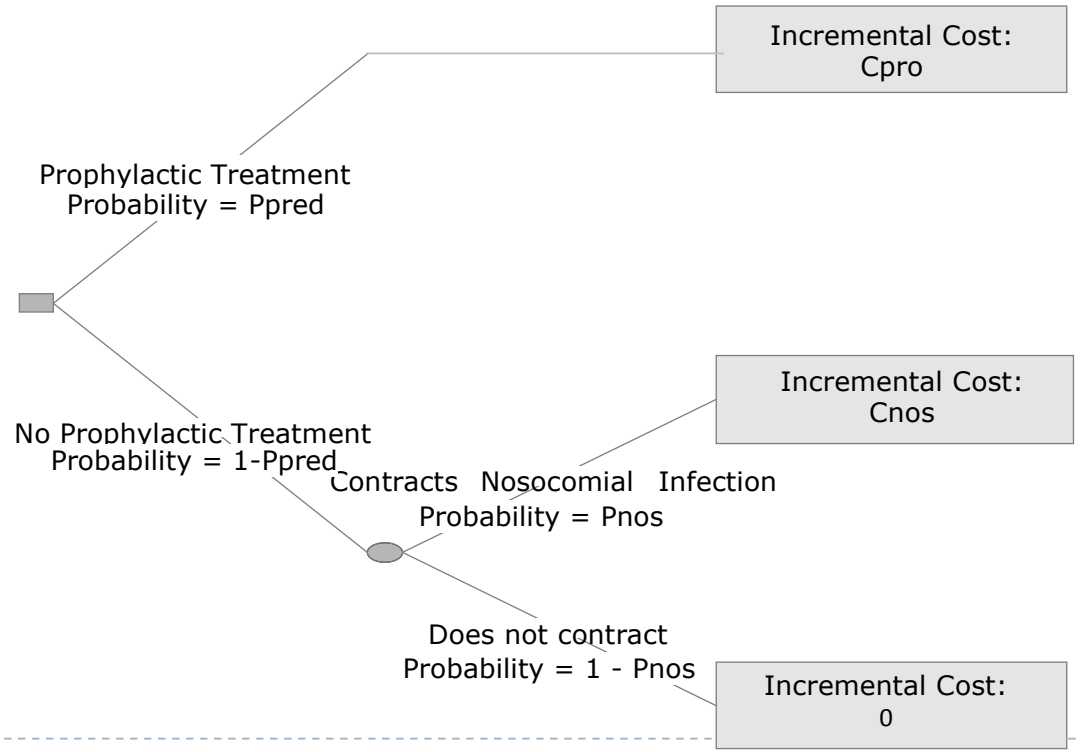


Cost Model

- Definitions
  - Regular treatment cost, Creg:
    - Hospitalization cost, Chos:
      - Facilities cost, Cfc
      - Physician charges, Cmd
    - Medication cost, Crx
  - Prophylactic treatment cost, Cpro:
    - Antibiotic cost, Canti times
    - Length of Stay, LOS + 1
  - Treatment cost of nosocomial infection, Cnos:
    - Similar components as of Creg
    - Random in nature

Formulation

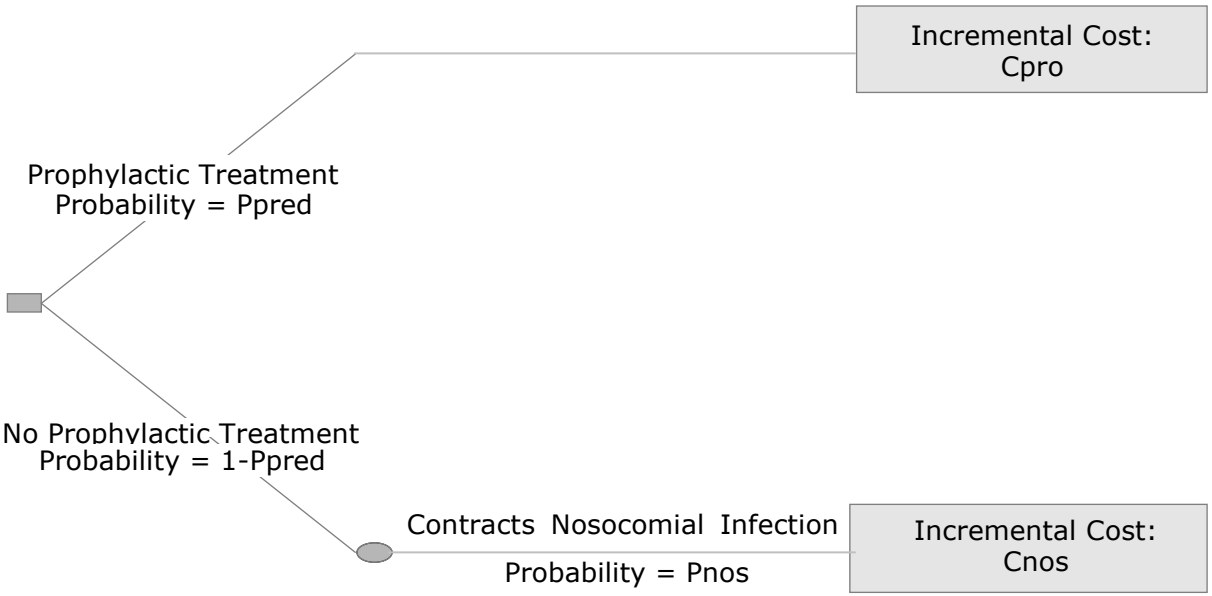
- Objective
  - minimize expected incremental cost



Formulation

- Minimize expected incremental cost =

$$\sum_{i = 1..N \text{ patients}} \{(Ppred[i] \times Cpro[i]) + ((1 - Ppred[i]) \times Pnos[i] \times Cnos[i])\}$$



Formulation

- Minimize expected incremental cost =

$$\sum_{i = 1..N \text{ patients}} \{(Ppred[i] \times Cpro[i]) + ((1 - Ppred[i]) \times Pnos[i] \times Cnos[i])\}$$

subject to

∀ patients i = 1 .. N:

Ppred[i] = f(patient i's risk rating for contracting nosocomial pneumonia)

Cpro[i] = (LOS[i]+1) x Canti

Cnos[i] = ø(severity of complications due to nosocomial pneumonia)

Ppred[i], Pnos[i] =~ [0 .. 1], real

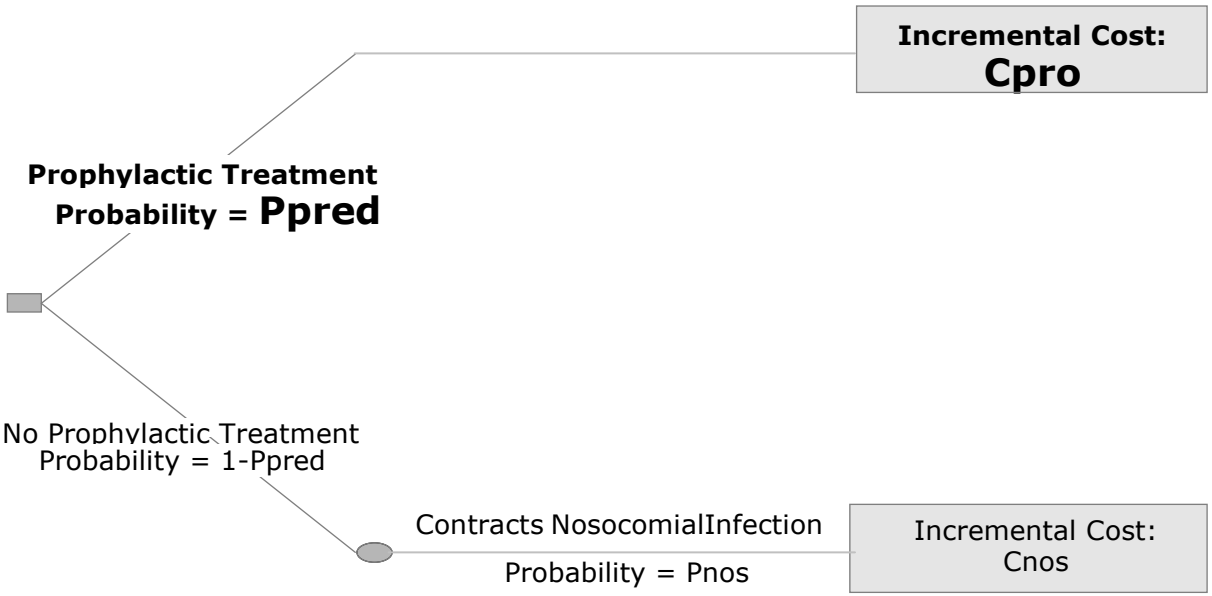
Canti, Cpro[i], Cnos[i] > 0, real

LOS[i] >= 0, integer

Parameter Estimation

- Minimize expected incremental cost =

$$\sum_{i = 1..N \text{ patients}} \{(Ppred[i] \times Cpro[i]) + ((1 - Ppred[i]) \times Pnos[i] \times Cnos[i])\}$$



Parameter Estimation

- Ppred: probability of the patient contracting nosocomial infection
  - AFTER physician's diagnosis, AND recommendation that patient be admitted
  - Use data mining classifier
  - Ppred is confidence score of the classifier
  - For +ve predictions
  - Possible improvements:
    - Adjust Ppred for classifier's performance on independent test set or using n-fold cross-validation
    - Instead of taking all +ve predictions, use a threshold prediction to select patients, truncate Ppred at threshold

Parameter Estimation

- Cpro: cost of prophylactic treatment
- Cpro = (LOS + 1) x Canti
- LOS: length of stay
  - Physician's estimate of LOS upon diagnosis
  - Possible improvement:
    - Incorporate physicians' estimate errors over long-term

Parameter Estimation

- Canti: cost of antibiotics per day
  - Antibiotics to treat “regular” pneumonia
  - Nosocomial pneumonia infected patients would have had additional treatments

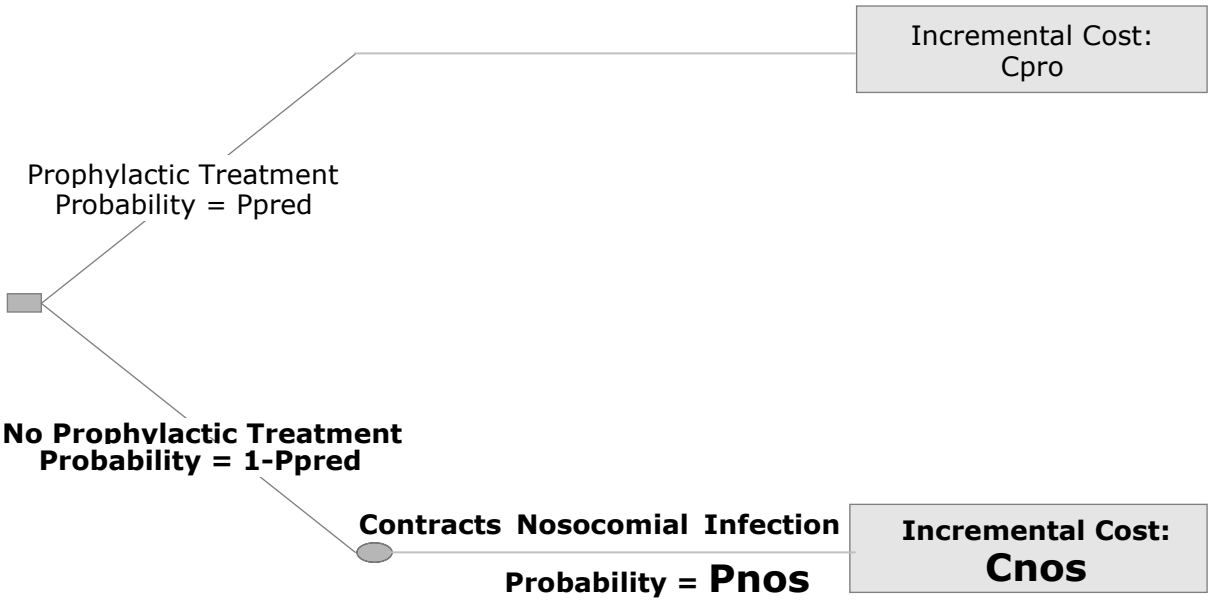
1. from hospital dataset, identify patients who had pneumonia, and separate them from those with nosocomial pneumonia.
2. from the medications dataset, identify the set of common drugs across the regular pneumonia patients, validate this against an external information set (eg, www.drugs.com). Compute total cost for these drugs per patient in the given year.
3. in the hospital dataset, for those regular pneumonia patients whose NUMNIGHX has been recorded, compute the cost of drugs per day by dividing the total cost of the stated drugs by NUMNIGHX.
4. take mean of valid costs per day to be Canti.

- Improvements possible:
  - Use domain knowledge of medical experts

Parameter Estimation

- Minimize expected incremental cost =

$$\sum_{i = 1..N \text{ patients}} \{(Ppred[i] \times Cpro[i]) + ((1 - Ppred[i]) \times Pnos[i] \times Cnos[i])\}$$



Parameter Estimation

- Pnos: probability of a low-risk patient contracting nosocomial pneumonia
  - Classifier's false negative predictions
  - Use evaluation of classifier's performance on independent test set or n-fold cross validation
  - Possible improvements:
    - Keep tuning the classifier!

Parameter Estimation

- Cnos: treatment cost of nosocomial pneumonia
  - Has components: hospitalization and medication
  - Random in nature
    1. from hospital dataset, identify patients who had pneumonia, and separate them from those with nosocomial pneumonia.
    2. from the medications dataset, identify the set of common drugs across the nosocomial pneumonia patients, validate this against an external information set (eg, www.drugs.com). Compute total cost for these drugs per patient in the given year.
    3. in the hospital dataset, note total hospitalization costs for the nosocomial pneumonia patients, and subtract from these individually mean hospitalization costs for regular patients of the respective primary conditions.
    4. add medication and hospitalization costs for each nosocomial pneumonia patient to arrive at the distribution of Cnos; use its mean as Cnos estimate for population.
  - Possible improvements:
    - Use regression or another data mining approach to predict severity and therefore cost of treatment for given profile of a patient