Data Mining Basics

Introduction

- **Data is growing at a phenomenal rate**
- **Users expect more sophisticated information**
- **How?**
  - Find hidden information in a database
  - Fit data to a model

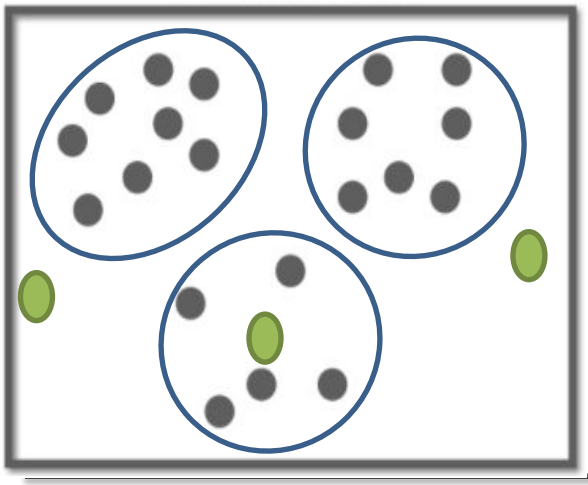
UNCOVER HIDDEN INFORMATION

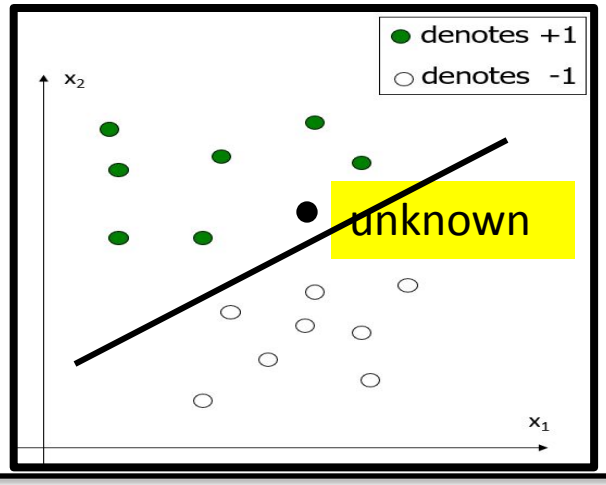
DATA MINING

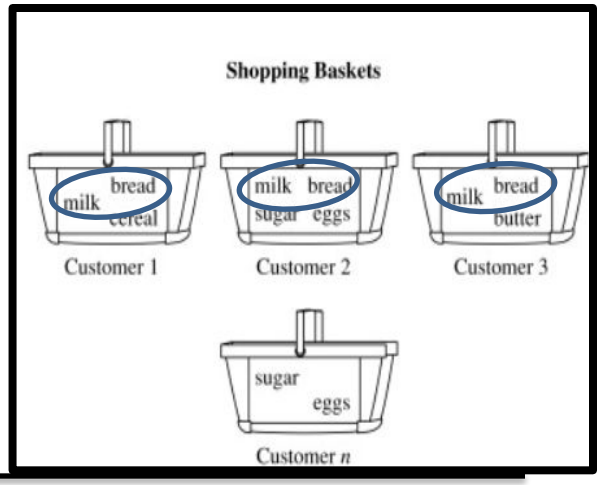
What is Data Mining?

- There are many definitions
- But most mean essentially the following.

Data mining is to discover interesting patterns from the large volumes of data.







What is a Pattern?

- **"A pattern is the opposite of chaos; it is an entity vaguely defined, that could be given a name."**
- **e.g.,**
  - fingerprint image,
  - handwritten word,
  - human face,
  - speech signal,
  - . . .

What is (not) Data Mining?

- **What is not Data Mining?**
  - Look up phone number in phone directory
  - Query a Web search engine for information about “Amazon”

- **What is Data Mining?**
  - Certain names are more prevalent in certain Indian locations
  - Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com,)

Database Processing vs. Data Mining Processing

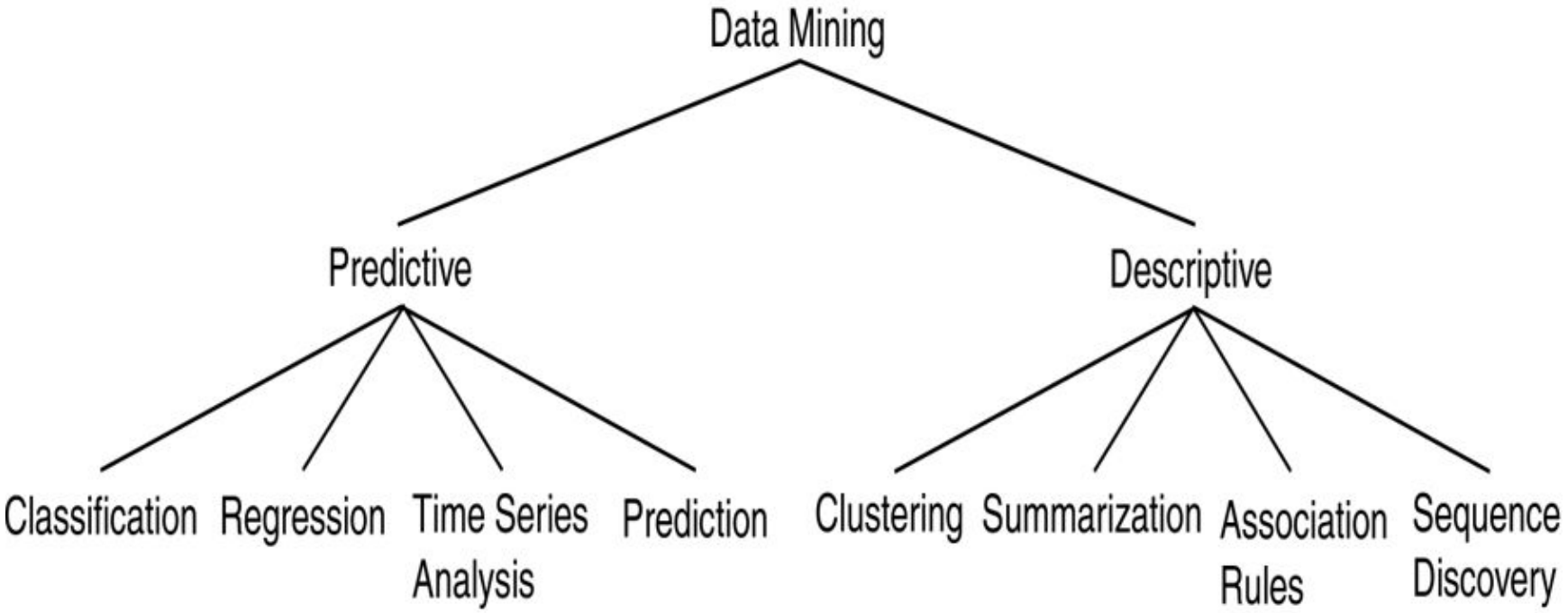
- **Query**
  - Well defined
  - SQL
- **Data**
  - Operational data
- **Output**
  - Precise
  - Subset of database

- **Query**
  - Poorly defined
  - No precise query language
- **Data**
  - Not operational data
- **Output**
  - Fuzzy
  - Not a subset of database

Query Examples

- **Database**
  - Find all credit applicants with last name of Smith.
  - Identify customers who have purchased more than $10,000 in the last month.
  - Find all customers who have purchased milk
- **Data Mining**
  - Find all credit applicants who are poor credit risks. (classification)
  - Identify customers with similar buying habits. (Clustering)
  - Find all items which are frequently purchased with milk. (association rules)

Data Mining Models and Tasks



Data Mining Tasks

- **Two main tasks**
  - Predictive data mining
    - Produces new, nontrivial information based on available data.
    - Predicts future values, or unknown values
  - Descriptive mining
    - Produces the model of the system described by the given set.

Basic Data Mining Tasks

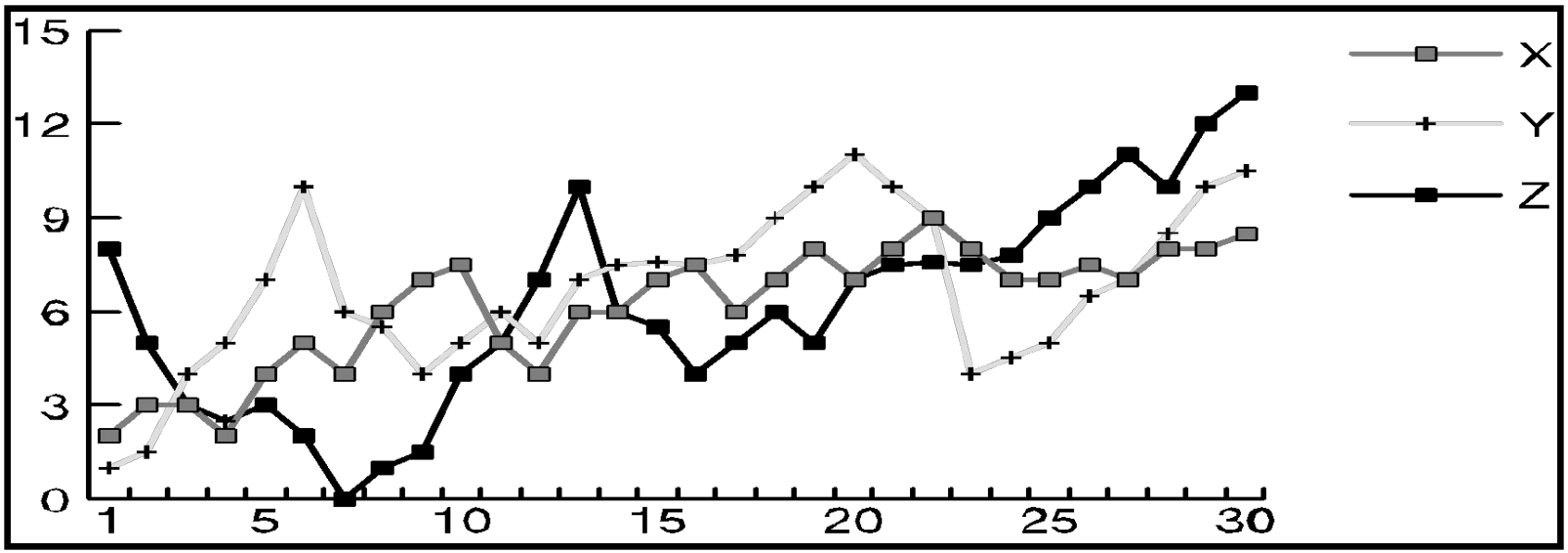
- **Classification maps data into predefined groups or classes**
  - Supervised learning
- **Regression is used to map a data item to a real valued prediction variable.**
- **Clustering groups similar data together into clusters.**
  - Unsupervised learning

Basic Data Mining Tasks

- **Summarization maps data into subsets with associated simple descriptions.**
  - Characterization
  - Generalization
- **Link Analysis uncovers relationships among data.**
  - Affinity Analysis
  - Association Rules
  - Sequential Analysis determines sequential patterns.

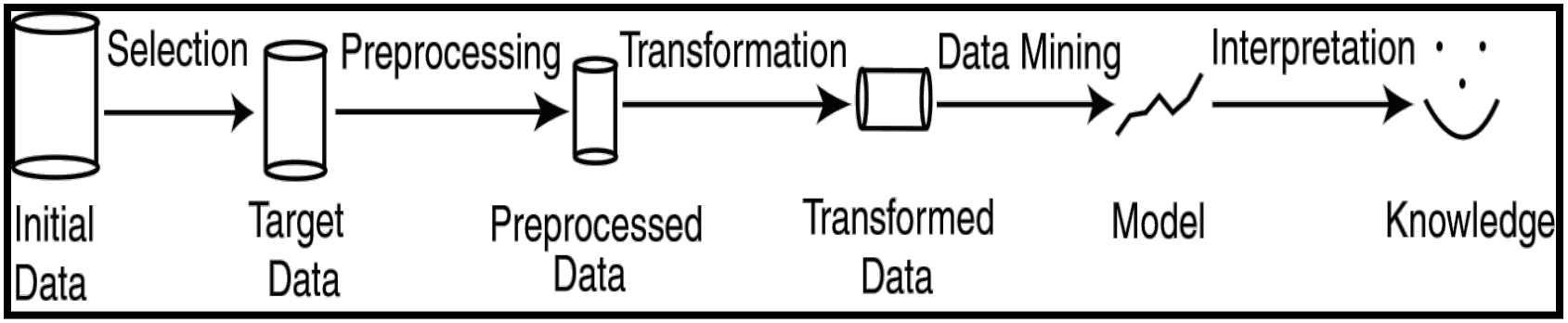
Ex: Time Series Analysis

- **Example: Stock Market**
  - Predict future values
  - Determine similar patterns over time
  - Classify behavior

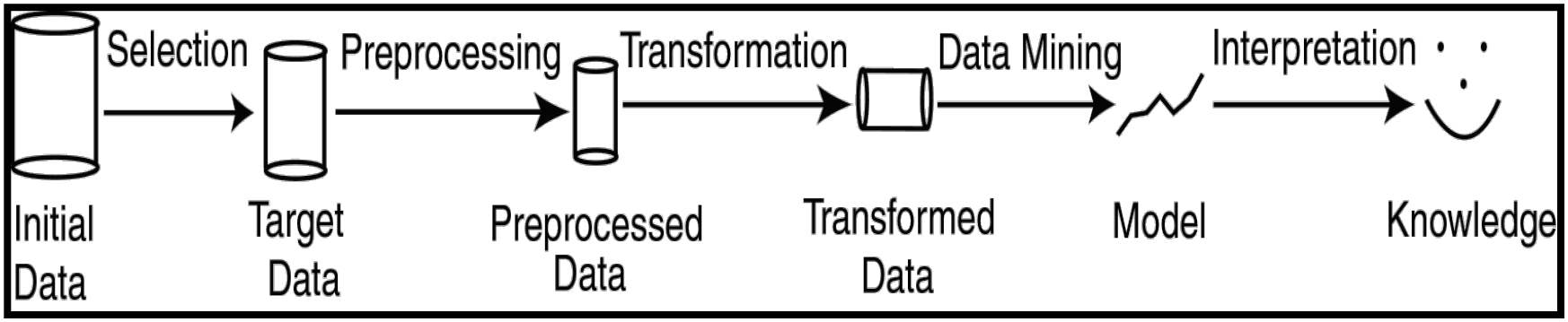


Data Mining vs. KDD

- ***Knowledge Discovery in Databases (KDD):*** **process of finding useful information and patterns in data.**
- ***Data Mining:*** **Use of algorithms to extract the information and patterns derived by the KDD process.**



KDD Process



- ***Selection:*** **Obtain data from various sources.**
- ***Preprocessing:*** **Cleanse data.**
- ***Transformation:*** **Convert to common format. Transform to new format.**
- ***Data Mining:*** **Obtain desired results.**
- ***Interpretation/Evaluation:*** **Present results to user in meaningful manner.**

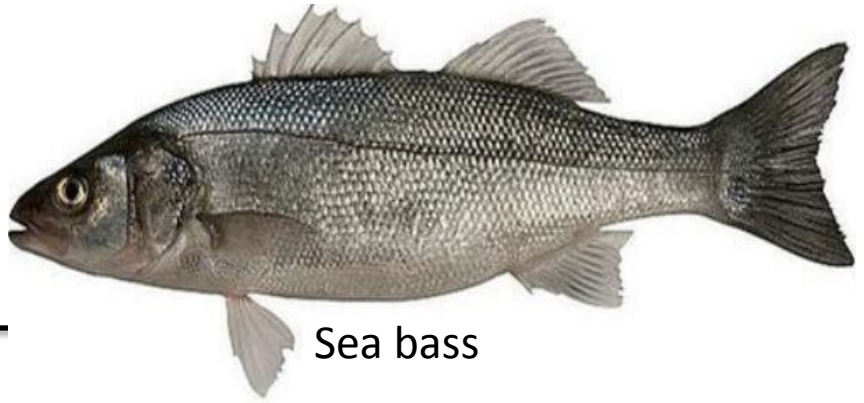
Classification

- **Given a collection of records (training set)**
  - Each record contains a set of attributes and have an associated class label.
- **Find a model for class label 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.

Classification - Example

- **A fish-packing plant wants to automate the process of sorting incoming fish according to species**
- **As a pilot project, it is decided to try to separate sea bass from salmon using optical sensing**
- **Features (to distinguish):**
  - Length
  - Lightness
  - Width
  - Position of mouth

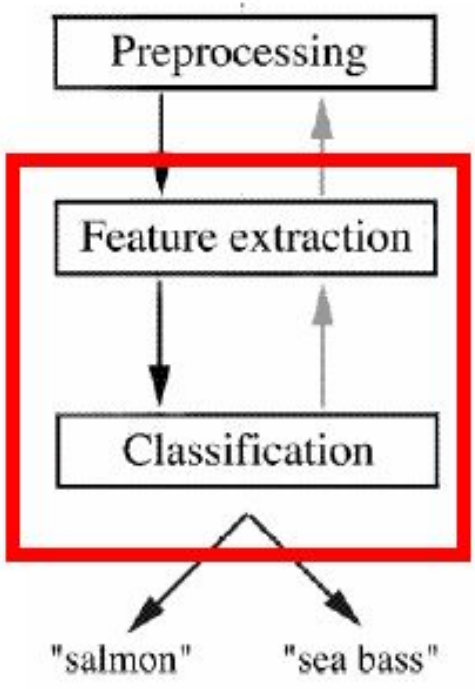
Salmon



Sea bass

Classification - Example

- **Preprocessing:**
  - Images of different fishes are isolated from one another and from background;
- **Feature extraction:**
  - The information of a single fish is then sent to a feature extractor, that measure certain “features” or “properties”;
- **Classification:**
  - The values of these features are passed to a classifier that evaluates the evidence presented, and build a model to discriminate between the two species



Classification - Example

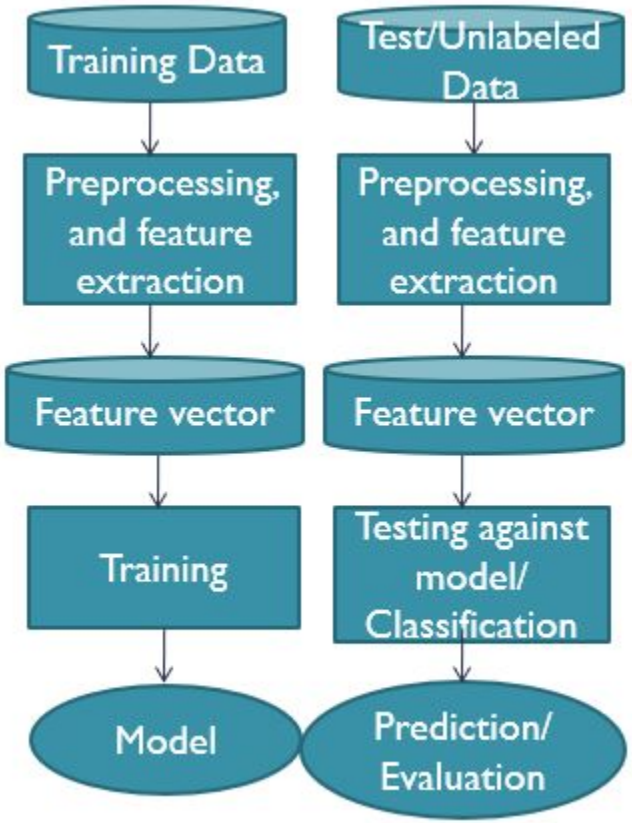
- **Domain knowledge:**
  - A sea bass is generally longer than a salmon
- **Related feature: (or attribute)**
  - Length
- **Training the classifier:**
  - Some examples are provided to the classifier in this form: <fish_length, fish_name>
  - These examples are called training examples
  - The classifier learns itself from the training examples, how to distinguish Salmon from Bass based on the fish_length

Classification - Example

- **Classification model (hypothesis):**
  - The classifier generates a model from the training data to classify future examples (test examples)
  - An example of the model is a rule like this:
  - If Length >= l* then sea bass otherwise salmon
  - Here the value of l* determined by the classifier
- **Testing the model**
  - Once we get a model out of the classifier, we may use the classifier to test future examples
  - The test data is provided in the form <fish_length>
  - The classifier outputs <fish_type> by checking fish_length against the model

Classification - Example

- **So the overall classification process goes like this □**



Classification: Application I

- **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.

Classification: Application II

- **Fraud Detection**
  - Goal: Predict fraudulent cases in credit card transactions.
  - Approach:
    - Use credit card transactions and the information on its account-holder as attributes.
    - When does a customer buy, what does he buy, how often he pays on time, etc
    - Label past transactions as fraud or fair transactions. This forms the class attribute.
    - Learn a model for the class of the transactions.
    - Use this model to detect fraud by observing credit card transactions on an account.

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.

Clustering: Application I

- **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.

Clustering: Application II

- **Document Clustering:**
  - Goal: To find groups of documents that are similar to each other based on the important terms appearing in them.
  - Approach: To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.
  - Gain: Information Retrieval can utilize the clusters to relate a new document or search term to clustered documents.

Association Rule Discovery: Definition

- **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.

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

Association Rule Discovery: Application I

- **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!

Association Rule Discovery: Application II

- **Supermarket shelf management.**
  - Goal: To identify items that are bought together by sufficiently many customers.
  - Approach: Process the point-of-sale data collected with barcode scanners to find dependencies among items.
  - A classic rule --
    - If a customer buys diaper and milk, then he is very likely to buy beer:

$$Diapers \rightarrow Beer,\ support = 20\%,\ confidence = 85\%$$

Acknowledgements

- **Lecture slides modified from**
  - Overview of Data Mining, Mehedy Masud
  - DATA MINING Introductory and Advanced Topics Part I, Margaret H. Dunham
  - Data Mining Basics, Arun K Pujari
  - An Introduction to Data Mining, Prof. S. Sudarshan