CPT-S 415

Big Data

Yinghui Wu

EME B45

CPT-S 415 Big Data

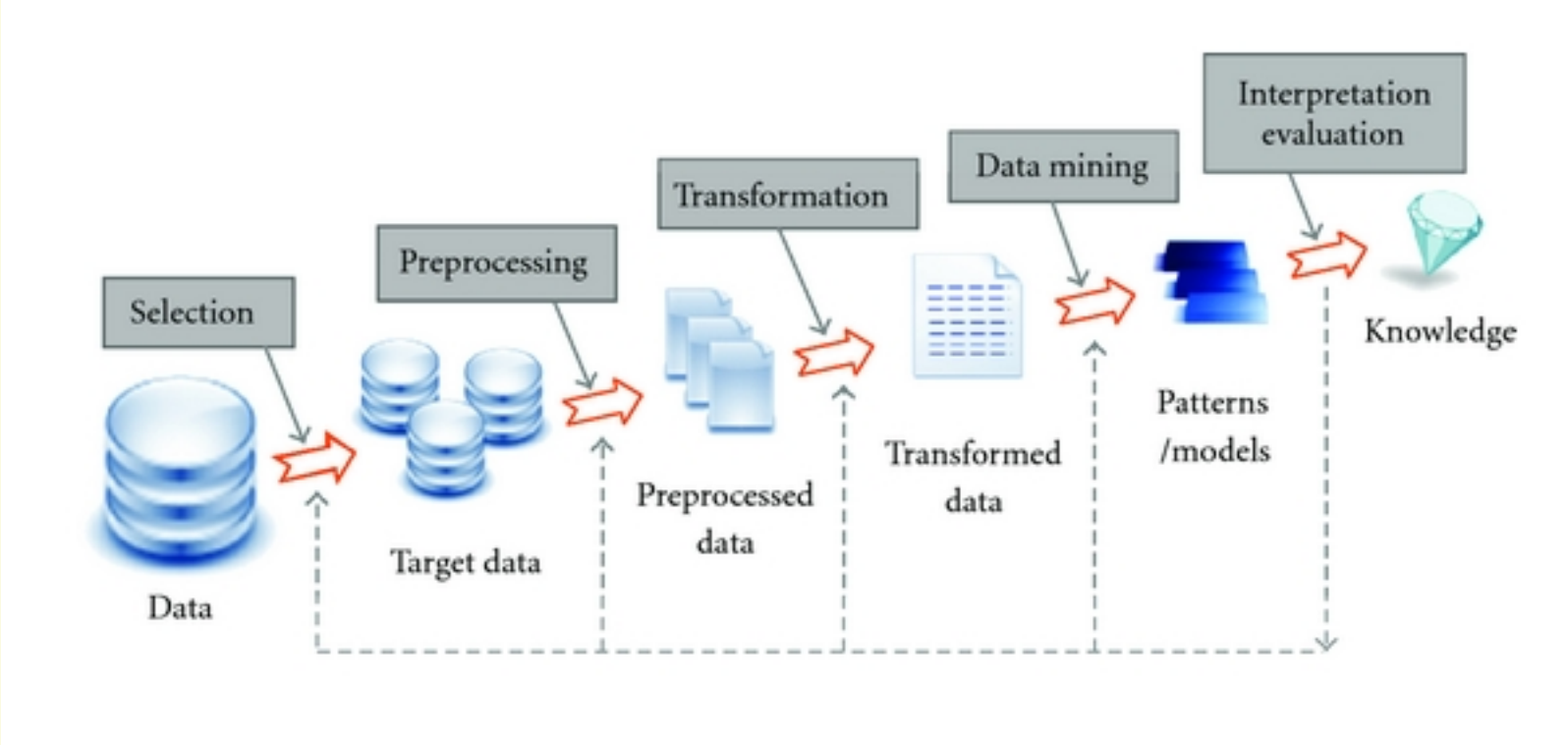
Special topic:

Data Mining & Graph Mining

- ✓ Data mining: from data to knowledge
- ✓ Graph Mining
- ✓ Classification (next week)
- ✓ Clustering (next week)

Data Mining Basics

Data mining


Selection
Preprocessing
Transformation
Data mining
Interpretation evaluation
Data
Target data
Preprocessed data
Transformed data
Patterns /models
Knowledge

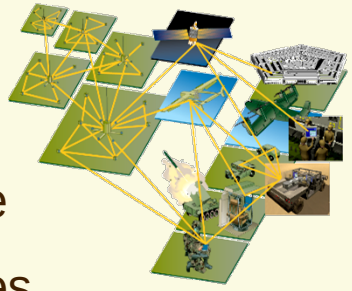
Why Mine Data? Commercial Viewpoint

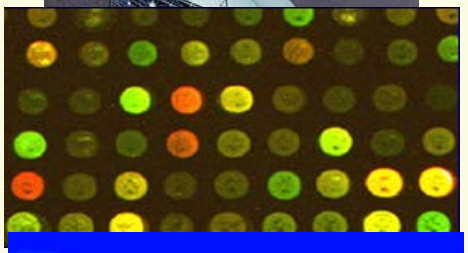
- ✓ Lots of data is being collected and warehoused
  - Web data, e-commerce
  - purchases at department/ grocery stores
  - Bank/Credit Card transactions

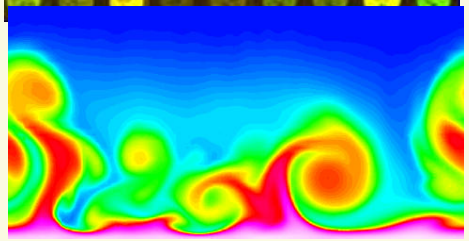
- ✓ Computers have become cheaper and more powerful
- ✓ Competitive Pressure is Strong
  - Provide better, customized services for e.g. Customer Relationship Management)

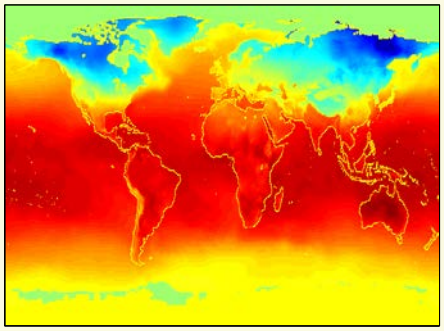
Why Mine Data? Scientific Viewpoint

- ✓ Data collected and stored at enormous speeds (TB/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





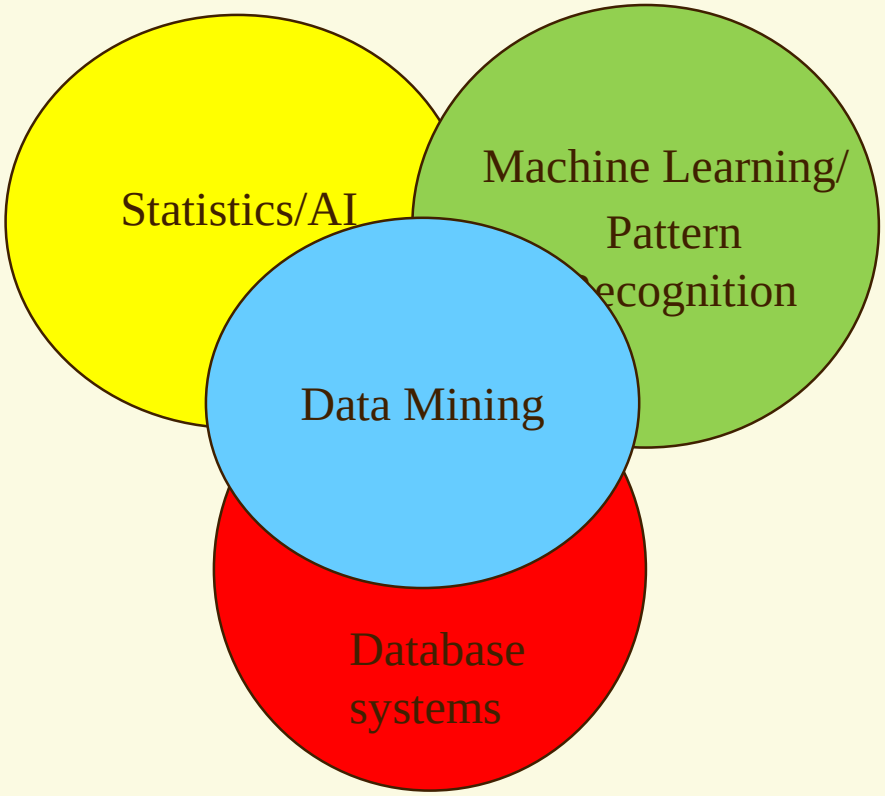




Origins of Data Mining

- ✓ Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems

- ✓ Traditional Techniques may be unsuitable due to
  - Enormity of data
  - High dimensionality of data
  - Heterogeneous, distributed nature of data



Database Processing vs. Data Mining

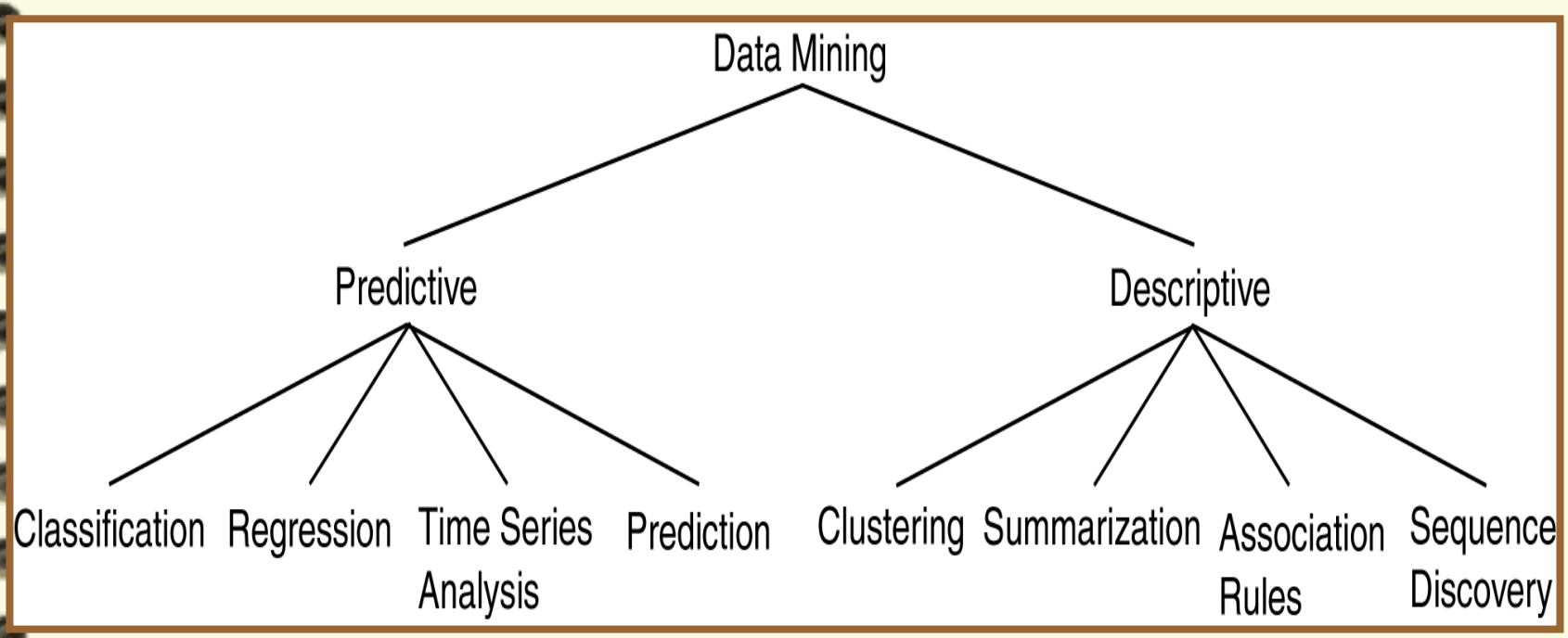
- Query
  - Well defined
  - SQL, SPARQL, Xpath…
  - Find all credit applicants with last name of Smith.
  - Identify customers who have purchased more than $10,000 in the last month.
  - Find all my friends living in Seattle and like French restaurant
- Output
  - Precise
  - Subset of database

- Query
  - Poorly defined
  - No precise query language
  - Find all credit applicants who are poor credit risks. (classification)
  - Identify customers with similar buying habits. (Clustering)
  - Find all my friends who frequently goes to French restaurants if their friends do (association rules)
- Output
  - Fuzzy
  - Not a subset of database

Statistics vs. Data Mining

Feature	Statistics	Data Mining
Type of Problem	Well structured	Unstructured / Semi-structured
Inference Role	Explicit inference plays great role in any analysis	No explicit inference
Objective of the Analysis and Data Collection	First – objective formulation, and then - data collection	Data rarely collected for objective of the analysis/modeling
Size of data set	Data set is small and hopefully homogeneous	Data set is large and data set is heterogeneous
Paradigm/Approach	Theory-based (deductive)	Synergy of theory-based and heuristic-based approaches (inductive)
Type of Analysis	Confirmative	Explorative
Number of variables	Small	Large

Data Mining Models and Tasks

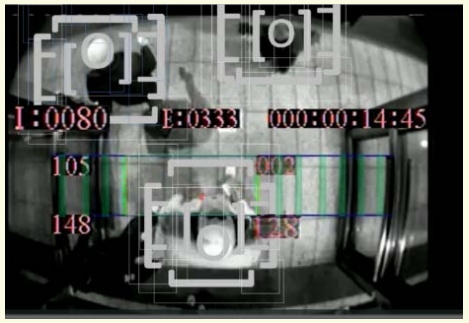


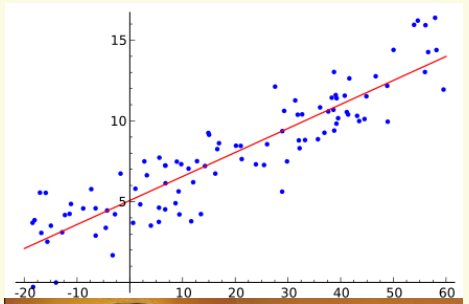
Use variables to predict unknown or future values of other variables.

Find human-interpretable patterns that describe the data.

Basic Data Mining Tasks

- ***Classification*** maps data into predefined groups or classes
  - Supervised learning
  - Pattern recognition
  - Prediction
- ***Regression*** maps a data item to a real valued prediction variable.
- ***Clustering*** groups similar data together into clusters.
  - Unsupervised learning
  - Segmentation
  - Partitioning

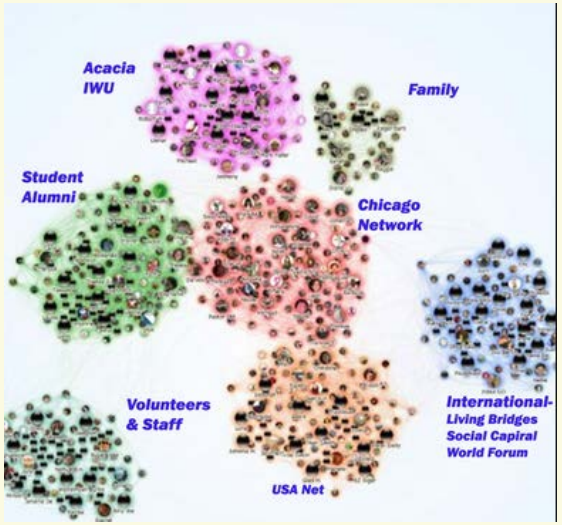




Basic Data Mining Tasks (cont'd)

- ***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.



Classification: Definition

- 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.

Classification: Application 1

- 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 2

- Customer Attrition/Churn:
  - Goal: To predict whether a customer is likely to be lost to a competitor.
  - Approach:
    - Use detailed record of transactions with each of the past and present customers, to find attributes.
      - How often the customer calls, where he calls, what time-of-the day he calls most, his financial status, marital status, etc.
    - Label the customers as loyal or disloyal.
    - Find a model for loyalty.

Classification: Application 3

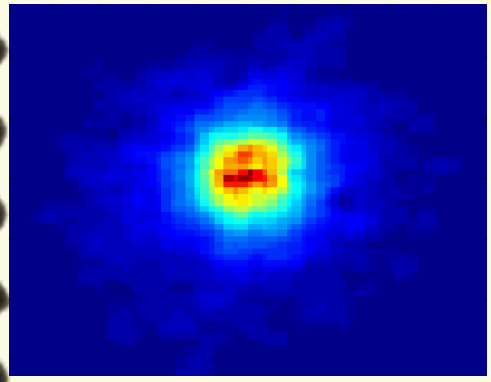
- Fraud Detection
  - Goal: Predict fraudulent cases in credit card transactions.
  - Approach:
    - Use credit card transactions and the information on its account-holder as attributes.
      - When does a customer buy, what does he buy, how often he pays on time, etc
    - Label past transactions 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.

Classification: Application 4

- ✓ Sky Survey Cataloging
  - Goal: To predict class (star or galaxy) of sky objects, especially visually faint ones, based on the telescopic survey images (from Palomar Observatory).
    - 3000 images with 23,040 x 23,040 pixels per image.
  - Approach:
    - Segment the image.
    - Measure image attributes (features) - 40 of them per object.
    - Model the class based on these features.
    - Success Story: Could find 16 new high red-shift quasars, some of the farthest objects that are difficult to find!

Classifying Galaxies

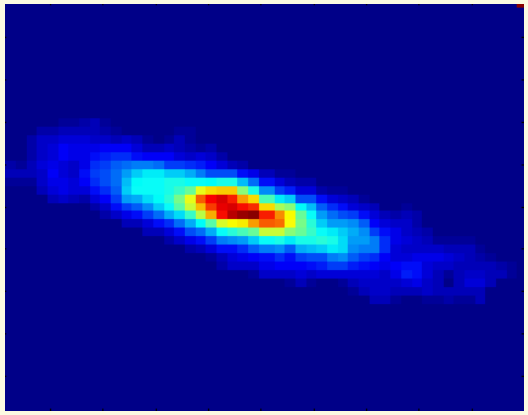
Early



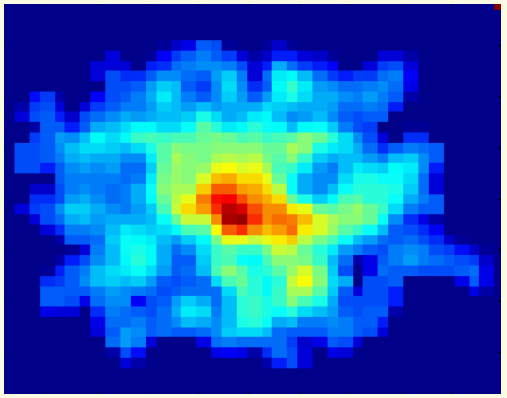
Class:
- Stages of Formation

Attributes:
- Image features,
- Characteristics of light waves received, etc.

Intermediate



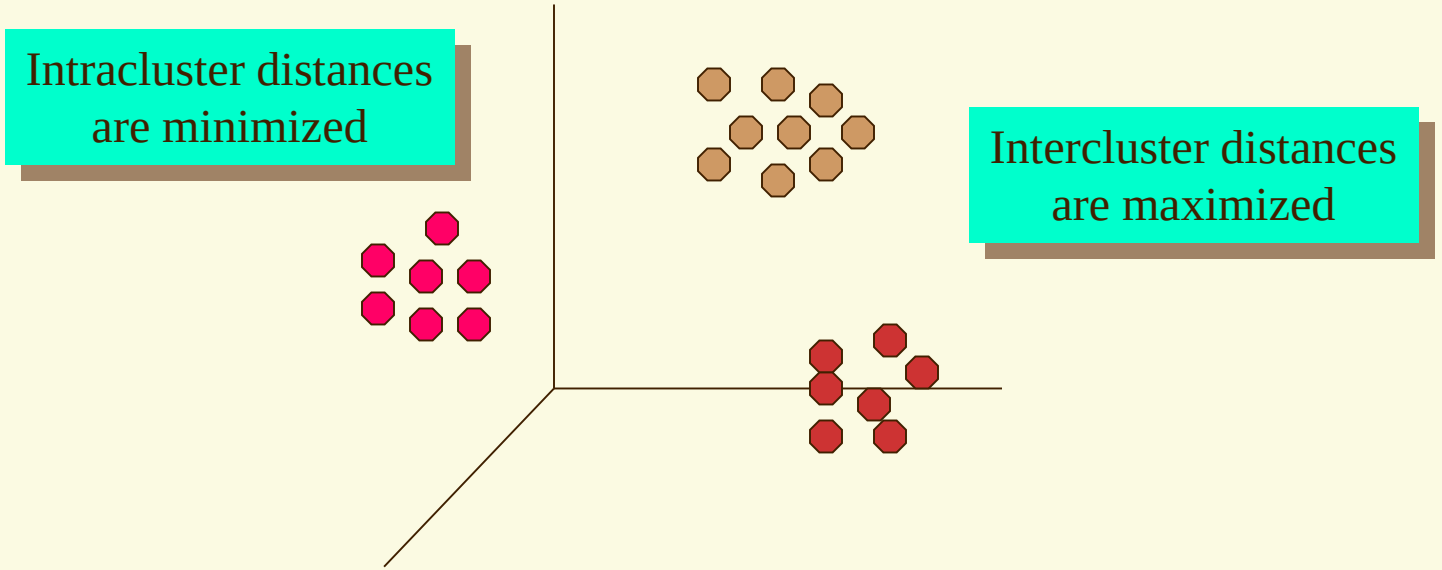
Late



Data Size:
- 72 million stars, 20 million galaxies
- Object Catalog: 9 GB
- Image Database: 150 GB

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 1

- 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 2

- Document Clustering:
  - Goal: To find groups of documents that are similar to each other bas[illegible] in them.
  - Approac[illegible]s in each documer[illegible] the frequenc[illegible]
  - Gain: Inf[illegible]ers to relate a new doc[illegible]uments.

Category	Total Articles	Correctly Placed
Financial	555	364
Foreign	341	260
National	273	36
Metro	943	746
Sports	738	573
Entertainment	354	278

Clustering of S&P 500 Stock Data

- ✓ Observe Stock Movements every day.
- ✓ Clustering points: Stock-{UP/DOWN}
- ✓ Similarity Measure: Two points are more similar if the events described by them frequently happen together on the same day. We can use association rules to quantify a similarity measure.

	Discovered Clusters	***Industry Group***
1	Applied-Matl-DOWN,Bay-Network-Down,3-COM-DOWN, Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN, DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN, Micron-Tech-DOWN,Texas-Inst-Down,Tellabs-Inc-Down, Natl-Semiconduct-DOWN,Oracl-DOWN,SGI-DOWN, Sun-DOWN	Technology1-DOWN
2	Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN, ADV-Micro-Device-DOWN,Andrew-Corp-DOWN, Computer-Assoc-DOWN,Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN,Microsoft-DOWN,Scientific-Atl-DOWN	Technology2-DOWN
3	Fannie-Mae-DOWN,Fed-Home-Loan-DOWN, MBNA-Corp-DOWN,Morgan-Stanley-DOWN	Financial-DOWN
4	Baker-Hughes-UP,Dresser-Inds-UP,Halliburton-HLD-UP, Louisiana-Land-UP,Phillips-Petro-UP,Unocal-UP, Schlumberger-UP	Oil-UP

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.

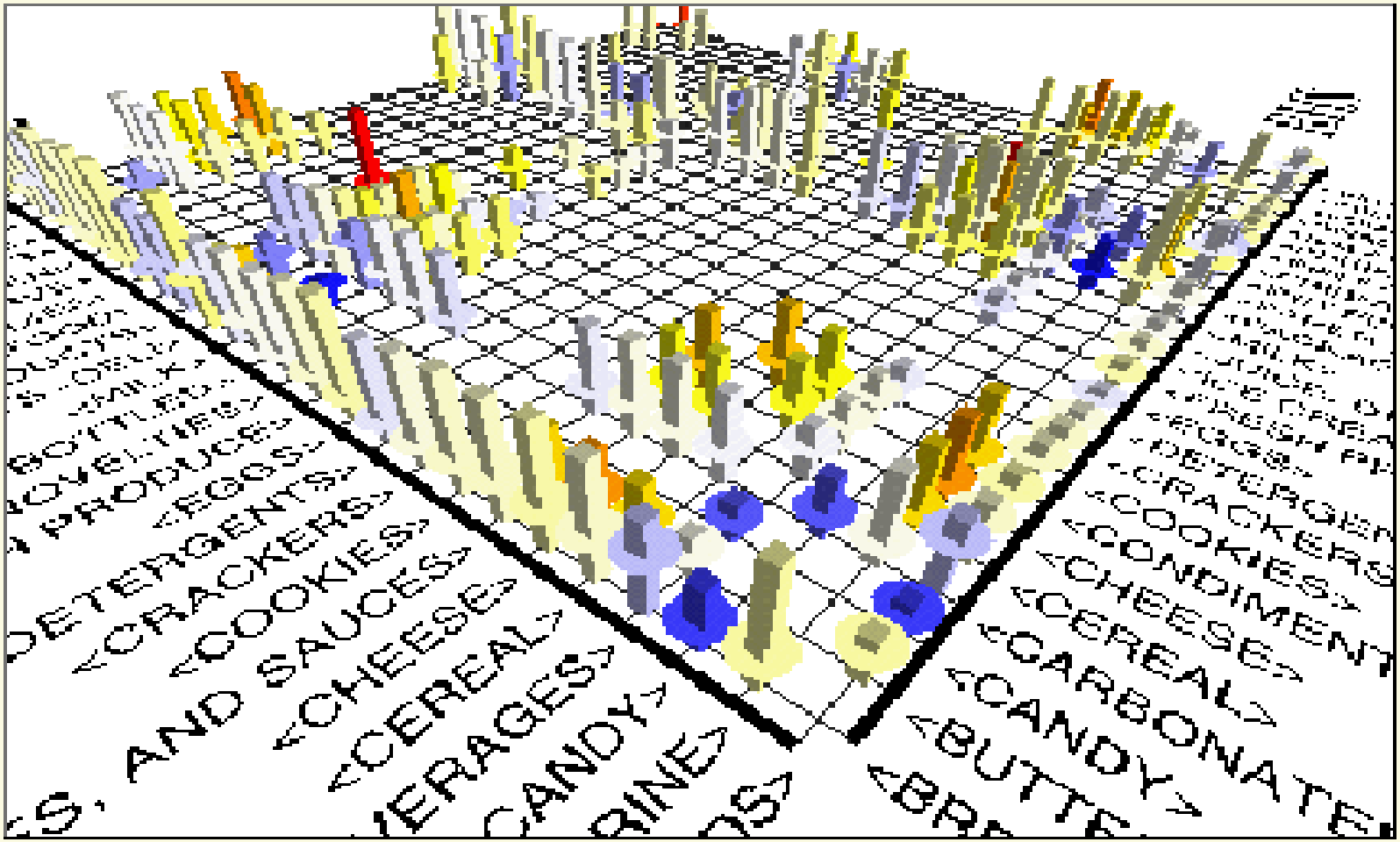
TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

Association Rule Discovery: Application 1



Association Rule Discovery: Application 2

- ✓ 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.
    - So, don't be surprised if you find six-packs stacked next to diapers!

Association Rule Discovery: Application 3

TECH 2/16/2012 @ 11:02AM | 2,689,200 views

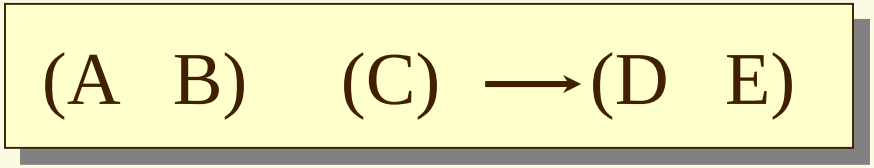
How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

As Pole's computers crawled through the data, he was able to identify about 25 products that, when analyzed together, allowed him to assign each shopper a "pregnancy prediction" score. More important, he could also estimate her due date to within a small window, so Target could send coupons timed to very specific stages of her pregnancy.

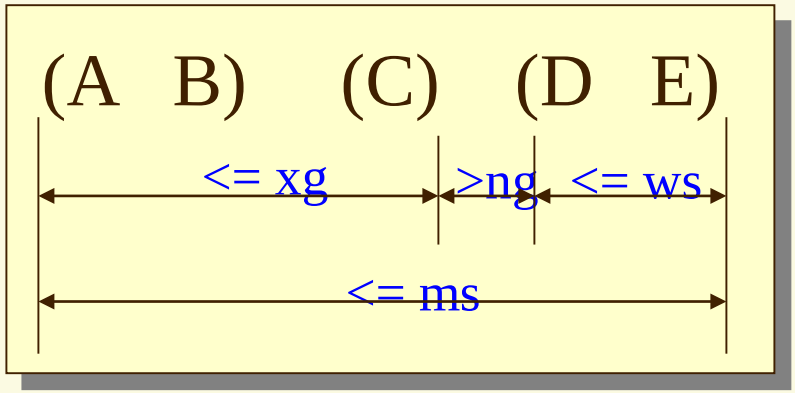
One Target employee I spoke to provided a hypothetical example. Take a fictional Target shopper named Jenny Ward, who is 23, lives in Atlanta and in March bought cocoa-butter lotion, a purse large enough to double as a diaper bag, zinc and magnesium supplements and a bright blue rug. There's, say, an 87 percent chance that she's pregnant and that her delivery date is sometime in late August.

Sequential Pattern Discovery: Definition

- 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.

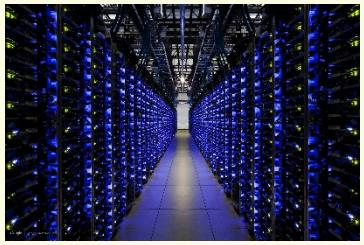


Sequential Pattern Discovery: Examples

- 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)

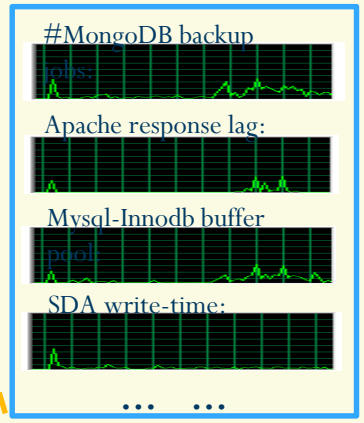
Example: Massive Monitoring Sequences Mining

Data center



Monitoring data

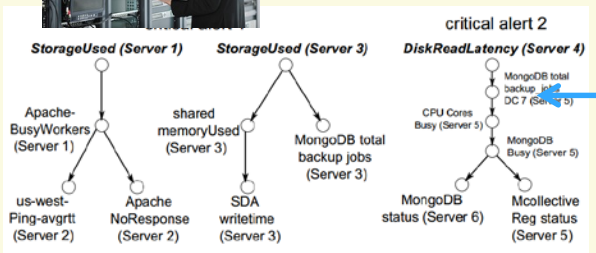
@Server-A

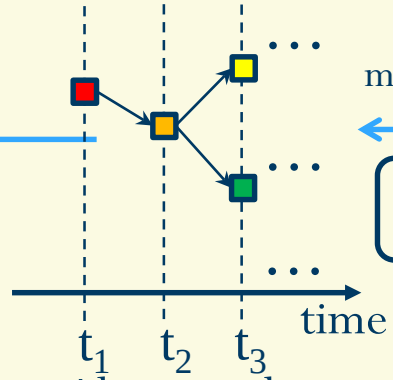


Alert @server-A

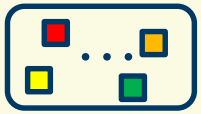
01:20am: #MongoDB backup jobs ≥ 30
01:30am: Memory usage ≥ 90%
01:31am: Apache response lag ≥ 2 seconds
01:43am: SDA write-time ≥ 10 times slower than average performance
…
09:32pm: #MySQL full join ≥ 10
09:47pm: CPU usage ≥ 85%
09:48pm: HTTP-80 no response
10:04pm: Storage used ≥ 90%
…

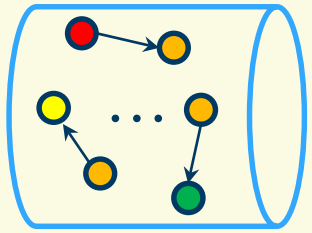
120-server data center can generate monitoring data 40GB/day





Online maintenance





Dependency rules

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.

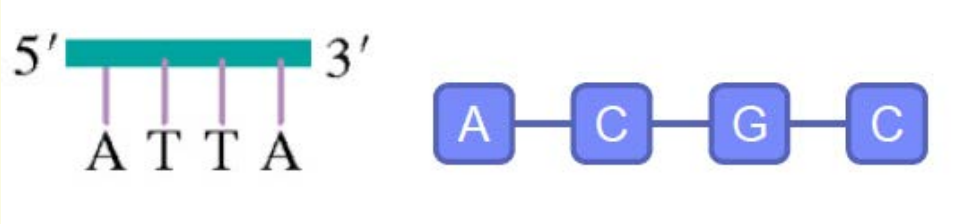
Challenges of Data Mining

- ✓ Scalability
- ✓ Dimensionality
- ✓ Complex and Heterogeneous Data
- ✓ Data Quality
- ✓ Data Ownership and Distribution
- ✓ Privacy Preservation
- ✓ Streaming Data

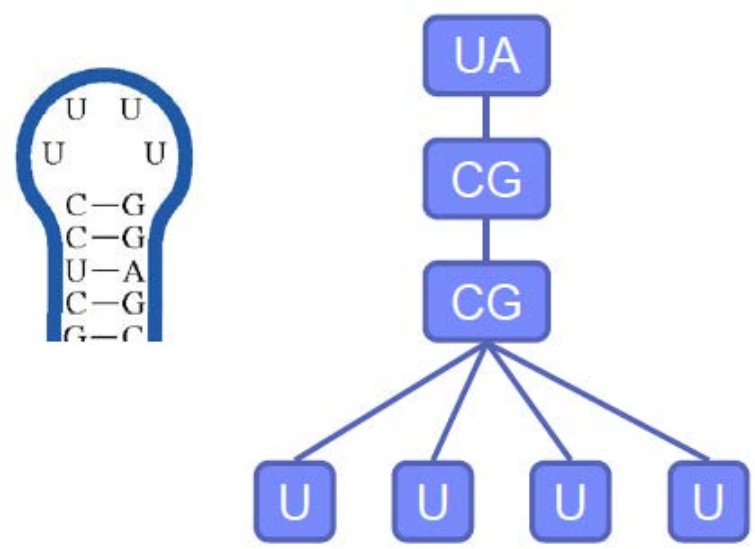
Graph Mining

Graph Data Mining

- ✓ DNA sequence

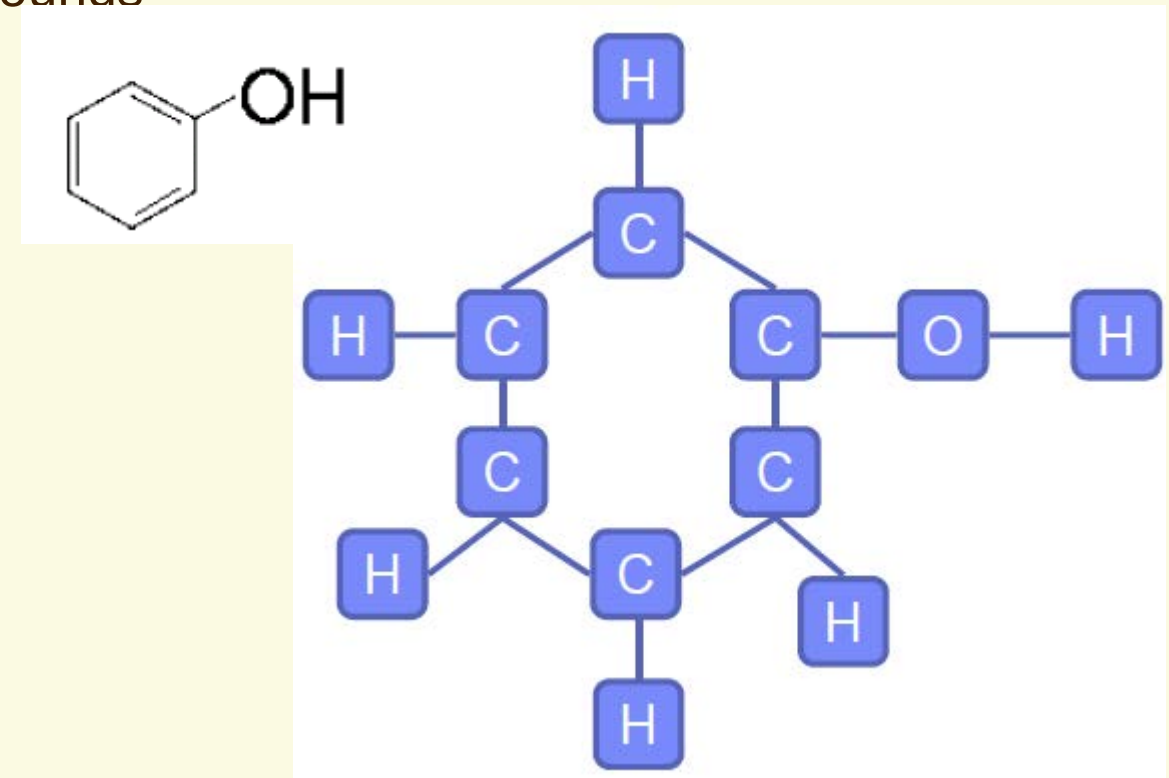


- ✓ RNA

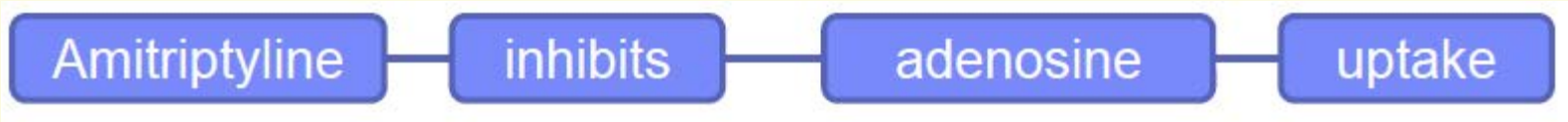


Graph Data Mining

- Compounds



- Texts



Graph Mining

- ✓ Graph Pattern Mining
  - Mining Frequent Subgraph Patterns
  - Graph Indexing
  - Graph Similarity Search
- ✓ Graph Classification
  - Graph pattern-based approach
  - Machine Learning approaches
- ✓ Graph Clustering
  - Link-density-based approach

Graph Pattern Mining

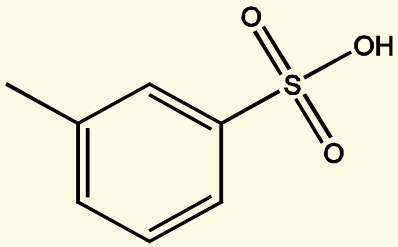
Graph Pattern Mining

- *Frequent* subgraphs
  - A (sub)graph is ***frequent*** if its *support* (occurrence frequency) in a given dataset is no less than a *minimum support* threshold
- Support of a graph g is defined as the percentage of graphs in G which have g as subgraph

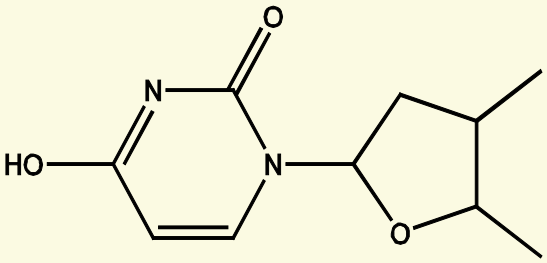
- Applications of graph pattern mining
  - Mining biochemical structures
  - Program control flow analysis
  - Mining XML structures or Web communities
  - Building blocks for graph classification, clustering, compression, comparison, and correlation analysis

Example: Frequent Subgraphs

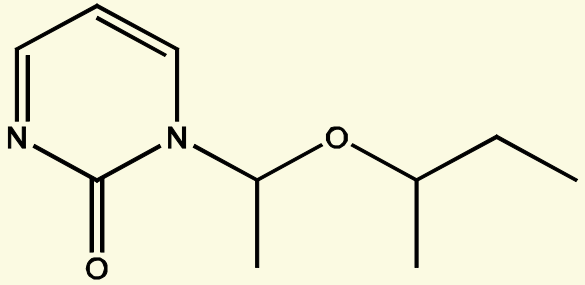
GRAPH DATASET



(A)



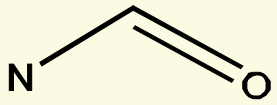
(B)



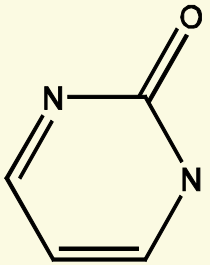
(C)

FREQUENT PATTERNS (MIN SUPPORT IS 2)

(1)

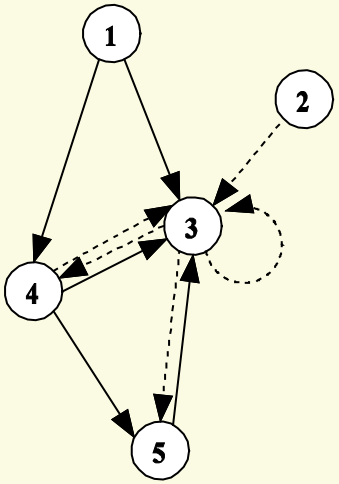


(2)

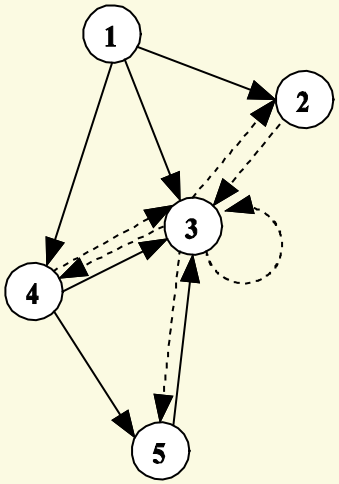


Example

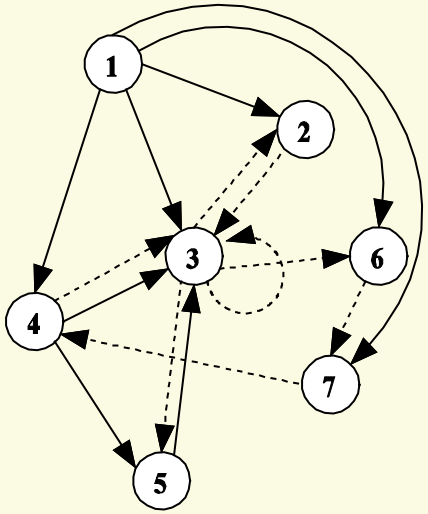
GRAPH DATASET



(1)



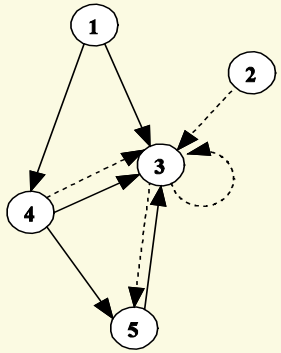
(2)



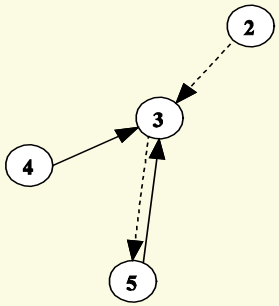
(3)

1: makepat
2: esc
3: addstr
4: getccl
5: dodash
6: in_set_2
7: stclose

FREQUENT PATTERNS (MIN SUPPORT IS 2)



(1)



(2)

Graph Mining Algorithms

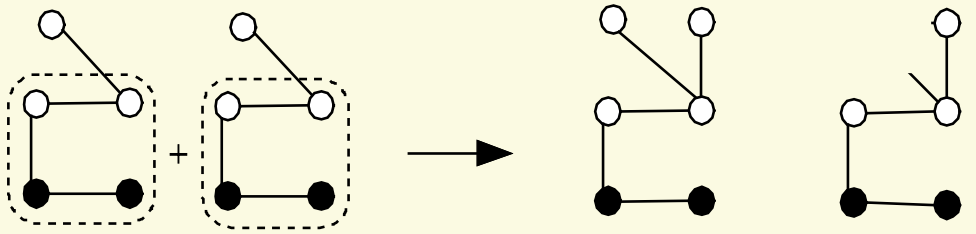
- ✓ Incomplete beam search – Greedy (Subdue)
- ✓ Inductive logic programming (WARMR)
- ✓ Graph theory-based approaches
  - Apriori-based approach
  - Pattern-growth approach

Properties of Graph Mining Algorithms

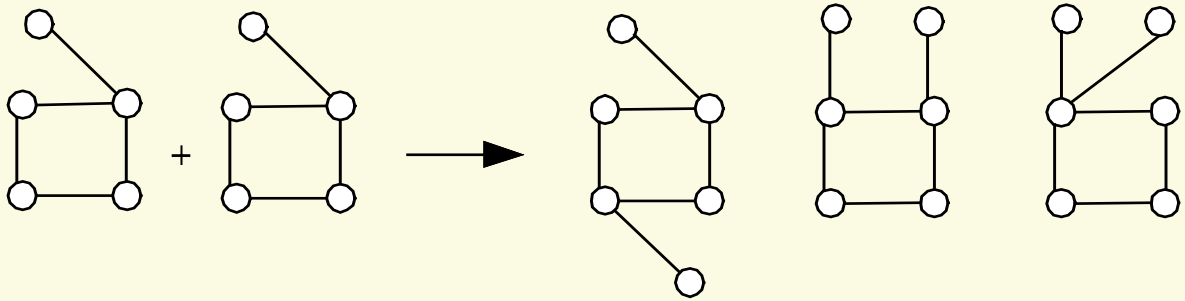
- ✓ Search order
  - breadth vs. depth
- ✓ Generation of candidate subgraphs
  - apriori vs. pattern growth
- ✓ Elimination of duplicate subgraphs
  - passive vs. active
- ✓ Support calculation
  - embedding store or not
- ✓ Discover order of patterns
  - path → tree → graph

Apriori-Based, Breadth-First Search

- Methodology: breadth-search, joining two graphs

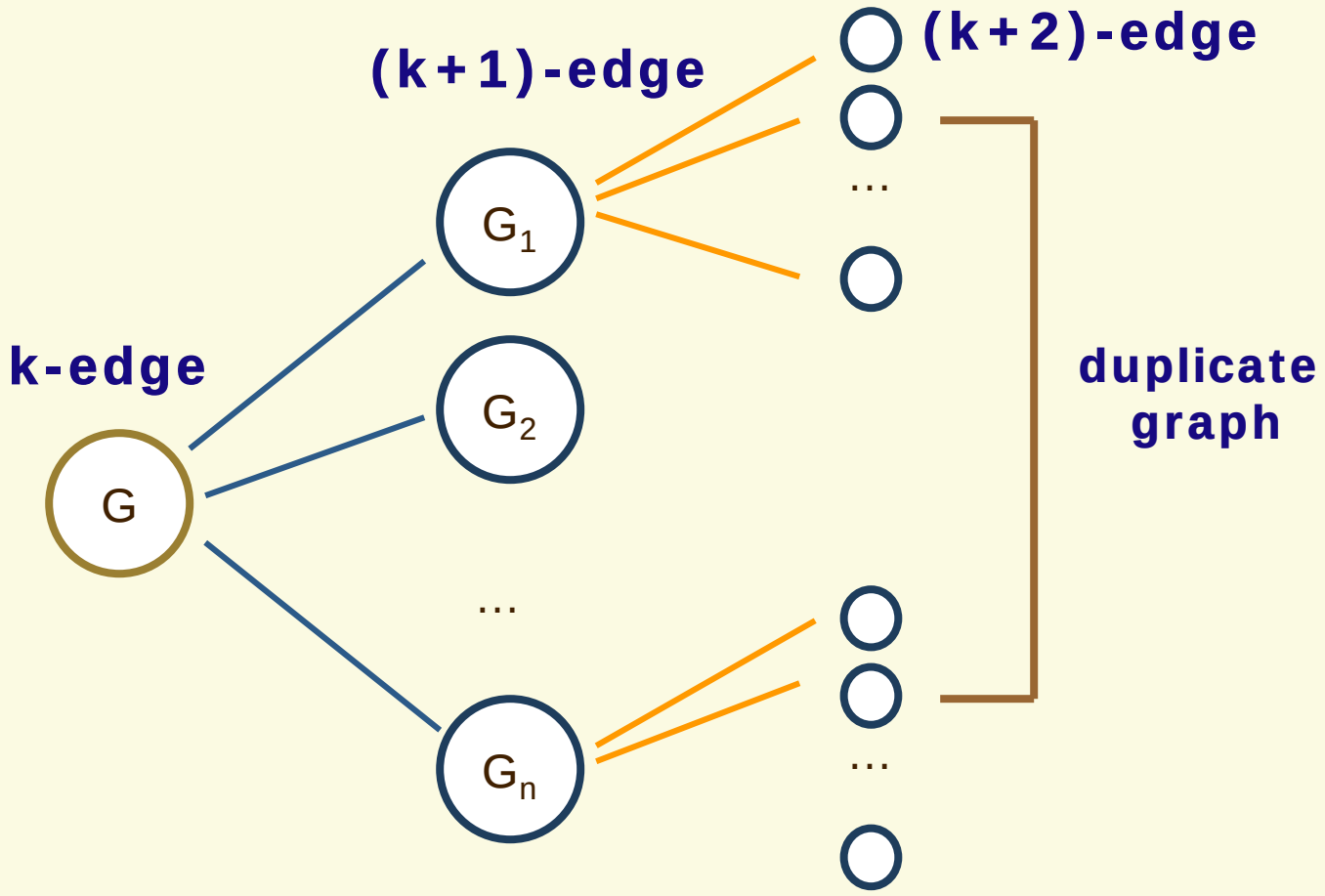


- AGM (Inokuchi, et al.)
  - generates new graphs with one more node



- FSG (Kuramochi and Karypis)
  - generates new graphs with one more edge

Pattern Growth Method



Graph Pattern Explosion Problem

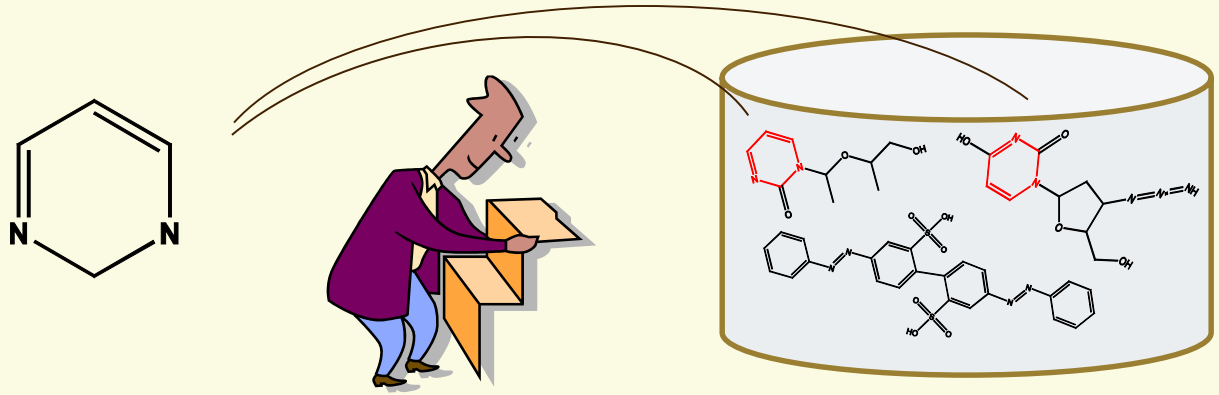
- If a graph is frequent, all of its subgraphs are frequent
  - **the Apriori property**
- An **n**-edge frequent graph may have 2^n subgraphs
- Among **422** chemical compounds which are confirmed to be active in an AIDS antiviral screen dataset,
  - there are **1,000,000** frequent graph patterns if the minimum support is 5%

Closed Frequent Graphs

- A frequent graph G is closed
  - if there exists no supergraph of G that carries the same support as G
- If some of G's subgraphs have the same support
  - it is unnecessary to output these subgraphs
  - nonclosed graphs
- Lossless compression
  - Still ensures that the mining result is complete

Graph Search

- Querying graph databases:
  - Given a graph database and a query graph, find all the graphs containing this query graph



query graph **graph database**

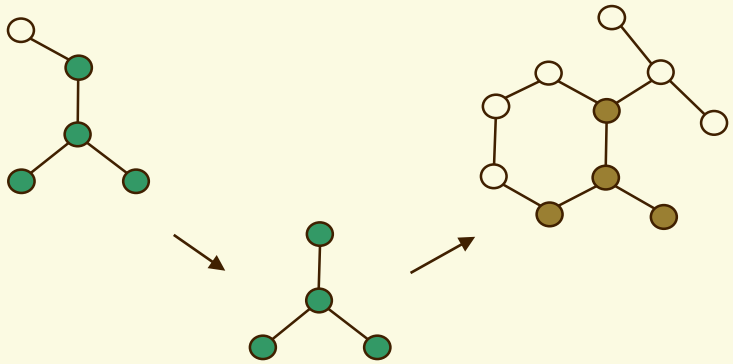
Scalability Issue

- Naïve solution
  - Sequential scan (Disk I/O)
  - Subgraph isomorphism test (NP-complete)
- Problem: Scalability is a big issue
- An indexing mechanism is needed

Indexing Strategy

Query graph (Q)

Graph (G)



Substructure

If graph G contains query graph Q, G should contain any substructure of Q

Remarks

- Index substructures of a query graph to prune graphs that do not contain these substructures

Indexing Framework

- Two steps in processing graph queries

Step 1. Index Construction

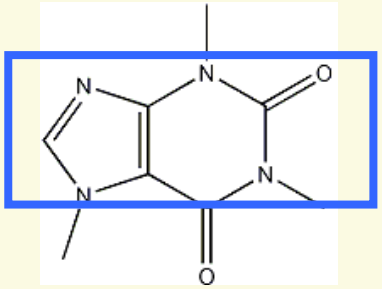
- Enumerate structures in the graph database, build an inverted index between structures and graphs

Step 2. Query Processing

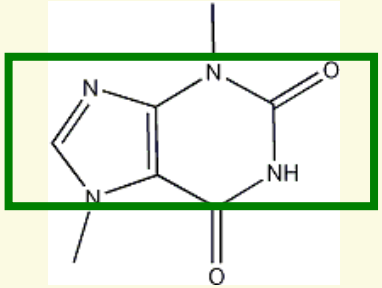
- Enumerate structures in the query graph
- Calculate the candidate graphs containing these structures
- Prune the false positive answers by performing subgraph isomorphism test

Structure Similarity Search

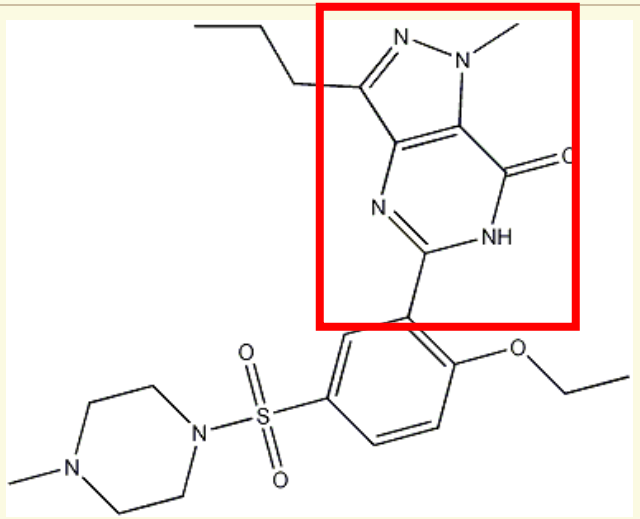
- **CHEMICAL COMPOUNDS**



(a) caffeine

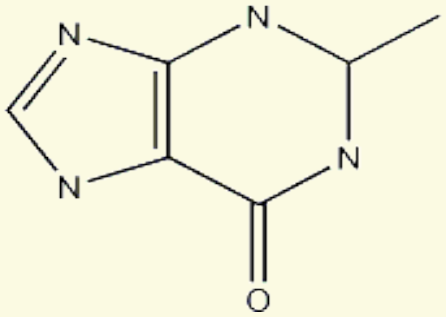


(b) diurobromine



(c) sildenafil

- **QUERY GRAPH**



Substructure Similarity Measure

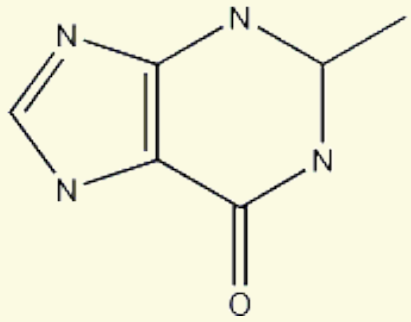
- **Feature-based similarity measure**
  - Each graph is represented as a feature vector

$$X = \{x_1, x_2, \ldots, x_n\}$$

  - Similarity is defined by the distance of their corresponding vectors
  - Advantages
    - Easy to index
    - Fast
    - Rough measure

Some “Straightforward” Methods

- Method1: Directly compute the similarity between the graphs in the DB and the query graph
  - Sequential scan
  - Subgraph similarity computation



- Method 2: Form a set of subgraph queries from the original query graph and use the exact subgraph search
  - Costly: If we allow 3 edges to be missed in a 20-edge query graph, it may generate 1,140 subgraphs

Index: Precise vs. Approximate Search

- ✓ Precise Search
  - Use frequent patterns as indexing features
  - Select features in the **database space** based on their selectivity
  - Build the index
- ✓ Approximate Search
  - Hard to build indices covering similar subgraphs
    - explosive number of subgraphs in databases
  - Idea: (1) keep the index structure

    (2) select features in the **query space**