אנאליטיקה של נתונים בזמן
2019-2020

Introduction and Data Mining Refresh

Robert Moskovitch, PhD

Software and Information Systems Engineering

Ben Gurion University

CDAL

- Funding: IBM, Microsoft, Amdocs, MoST, MAFAAT and more.

- Collaborations: Columbia University, Maccabi Healthcare Services, AIIMS/IIT New Delhi, Peking University, UTHealth, UPenn/CHOP and more

- Professional Activities (Organizer/SPC/PC): PLOS ONE – Editor, S/PC: ACM KDD, AAAI, IJCAI, Editorial Board of *Journal of Biomedical Informatics*. AIME 2020 – Co-Chair.

- Students: Roni Mateless, Nofar Sarafian, Guy Danieli, Stav Sapir, Tal Ivshin, Maya Schvatz, Pavel Novitzki, Omer Harel, Amos Zamir, Noa Lemberger, Nevo Itzhak, Ofir Dvir, Guy Shitrit.

Funders and Collaborators

- Funding

IBM

Microsoft

amdocs

PERSON CENTRED SOFTWARE
MOBILE CARE MONITORING

Prime Minister's Office
National Cyber Bureau

חיל הרפואה

משרד המדע,
הטכנולוגיה והחלל
Ministry of Science, Technology & Space

- Collaborators

UCL

Center for Veteran Transition and Integration
Columbia University

Maccabi
The Best Healthcare in Israel

MK&M

AIIMS

Hospital
Assuta
Raising Health Standards

Mount Sinai

UTHealth
The University of Texas
Health Science Center at Houston

Robert Moskovitch, PhD

- Senior Lecturer, Software and Information Systems Engineering, Ben Gurion University
- Adjunct Assistant Professor, Ichan Medical School at Mount Sinai, NYC, USA.
- Head, Complex Data Analytics Lab
- Member, BGU Zlotowsky Neuroscience Center, Brain Medicine Center, @cyber.
- Post Doc, Biomedical Informatics, Columbia University
- R&D Project Manager, Deutsche Telekom Innovation Labs
- Research: Text Retrieval and Categorization, Behavioral (typing and mouse) Verification, Unknown Malware Discovery, Temporal Data Mining (Time Intervals Mining, Classification and Prediction)

Today’s Agenda

- What and Why Temporal Data Analytics?
- Course Goals and TDA Topics
- Course Structure: Assignments and Timeline
- Topics in Brief for the Course Project
- Introduction to Temporal Data Mining
- Prerequisite Methods (without time)
    - Association Rules Mining
    - Classification
    - Clustering

Typical Atemporal Data

- Atemporal data would be a set of values describing an object.

- The description can refer to a moment in time, or a summary of a time period (i.e., an hour, day, year, cet)

- Typically it will be a vector of descriptors described by values: continuous, nominal, and cet.

- However, temporal data based description is much more heterogeneous, and dynamic, which creates a lot of complexity.

Analytics of Data in Time

What? . . . .

Why?... . . i . .l . .. ..

Where? . A AB K IJ LLL O QA

When? . . . : :: .: : :: : : :: . ;: .

Analytics of Data over Time – What?

- Time point values series
  - Fixed frequency (typically, electronic sensors)
    - Different variables may have different frequencies
  - Irregular sampling (typically, manual sampling, or event driven)
- Event series
  - Instantaneous (no duration) events
  - May have different types of events (A, B, C,..)
  - Sampled in fixed frequency
  - Sampled irregularly (manual sampling or event driven)
  - Having duration – time intervals

Today’s Agenda

- What and Why Temporal Data Analytics?
- Course Goals
- Course Structure: Assignments and Timeline
- Topics in Brief for the Course Project
- Introduction to Temporal Data Mining
- Prerequisite Methods (without time)
    - Association Rules Mining
    - Classification
    - Clustering

Analytics of Data in Time - Goals

- Overview of the field of temporal-data-analytics within the field of data analytics (KDD oriented)
- Motivation for Temporal Data Analytics
- Challenges: time, different samplings, irregularity, and cet,.
- Main Methods:
  - Time series analysis: univariate, multivariate, indexing/classification/forecasting ..
  - Sequential data mining
  - Time Intervals Mining
  - Temporal Data based Classification and more.

Today’s Agenda

- What and Why Temporal Data Analytics?
- Course Goals
- Course Structure: Assignments and Timeline
- Topics in Brief for the Course Project
- Introduction to Temporal Data Mining
- Prerequisite Methods (without time)
    - Association Rules Mining
    - Classification
    - Clustering

Analytics of Data in Time – GOAL (not task)

- The main goal is the project
- The course project aims to result in a research reported as an academic paper
  - SIG-KDD style 6-8 pages paper
- Number of Students: 2 depending on the project (1 or 3 is optional)
- Topics: application of TDA method/s on a temporal dataset/s
- Students who have a temporal problem in their thesis are encouraged to work on them as a project – after approval. It can be also sequential, or using ”temporal” methods.
- Otherwise, the students will implement a published paper – after approval.
- And there will be a quiz on the course materials.

An Academic Paper
– we will speak about it more

- Abstract
- Introduction
- Background
- Methods
- Evaluation {Research Questions, Data, Evaluation Plan}
- Results
- Discussion

Today’s Agenda

- What and Why Temporal Data Analytics?
- Course Goals
- Course Structure: Assignments and Timeline
- Topics in Brief for the Course Project
- Introduction to Temporal Data Mining
- Prerequisite Methods (without time)
    - Association Rules Mining
    - Classification
    - Clustering

Analytics of Data in Time – Topics List

- Time Point Series univariate – indexing, match and similarity, search and retrieval, and more
- Multivariate Temporal Data – time series analysis, or heterogeneous variables
- Forecasting
- Clustering
- Classification
- Patterns Discovery
- And more .. we will go through the topics in few more slides ..

Analytics of Data in Time – GOAL (Project)

- Outcomes:
  - Intermediate:
    - A project **proposal** of one page (in English: Introduction(motivation), Methods, Data, Evaluation Goals)
    - A literature **survey** of two pages (in English, including at least 10 refs)
    - A **presentation** of 10 mins (including not more than 7 slides)
      - 3-4 literature survey + 2 Project Methods + 1 Data + 1 Experimental Plan
  - Final:
    - A project report along 6-8 pages in SIG-KDD format
      - – download from ACM SIGKDD 2018 CFP
    - A corresponding presentation of 15 mins

Analytics of Data in Time – a quiz

- The quiz will be based on the contents learnt in class
  - at the end of the semester (or split into two quizzes to make it easier to prepare)

- Will include two hours and contain about 4-6 questions

Course Schedule

Date	Lecture	Project Assignments
October 20, 2021	Lecture	
October 27, 2021	Lecture	
November 03, 2021	Lecture	* Submit Project Proposal (1 page)
November 10, 2021	Lecture	
November 17, 2021	Lecture	
November 24, 2021	Lecture	
December 1, 2021	Lecture	Quiz 1 * Submit Literature Survey
December 08, 2021	Literature Survey PPTs	* Attendance mandatory
December 15, 2019	Literature Survey PPTs	* Attendance mandatory
December 22, 2021	Lecture	
December 29, 2021	Lecture	
January 05, 2022	Projects PPTs	Quiz 2 * Attendance mandatory
January 12, 2022	Projects PPTs (Mandatory)	* Attendance mandatory Submit Project

Analytics of Data in Time – Important Dates

- November 03, 2021 – Projects Proposal in one page (Problem, Methods, Research Questions, Datasets) - better to decide on the first week, to start working.
- December 1, 2021 – Quiz 1 (3 questions)
- December 1, 2021 – submit literature survey (the beginning of the report: Introduction + Background)
  - You can submit also Methods (or more) and have my comments
- December 08,15, 2021 - Literature Survey PPTs, including:
  - 1 slide : Problem/Motivation
  - 3-4 slides : Common Methods ..
  - 1 slide : your project
- January 05, 2022 – Quiz 2 (3 questions)
- January 05, 12, 2022 – Submit Project Reports + Final Project PPTs
  - Shortened literature ppt + Methods + Evaluation (Research Questions + Experimental Plan) + Results + Discussion/Conclusion

Temporal Data Analytics - Grading

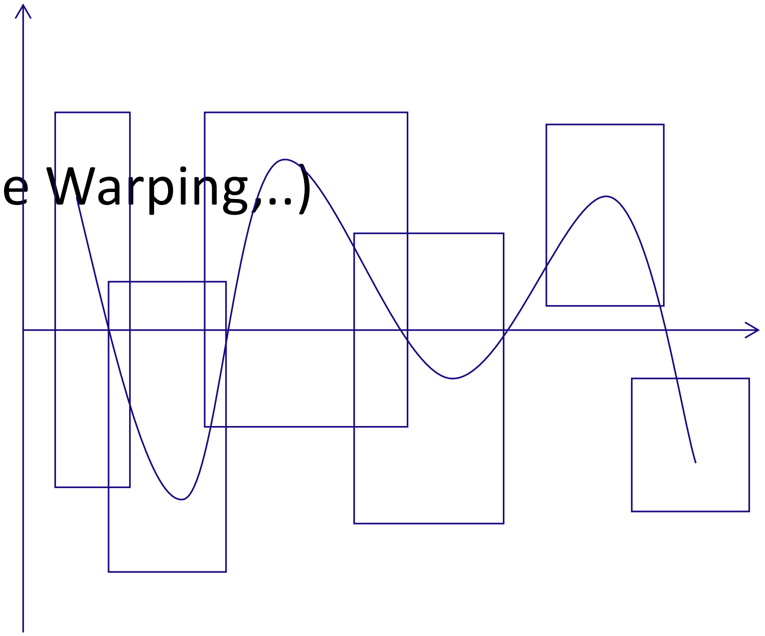
- 15% - Quiz
- 75% - Project
  - 25% - Literature survey + presentation
  - 50% - Final report in academic paper format + ppt
    - 10% originality and innovation
    - 10% complexity
    - 20% writing and presentation
    - 15% soundness and comprehensiveness
- 10% - Impression
  - Attendance – in student presentation classes names will be listed
  - participation in the class, and generally seriousness

Project Topics List

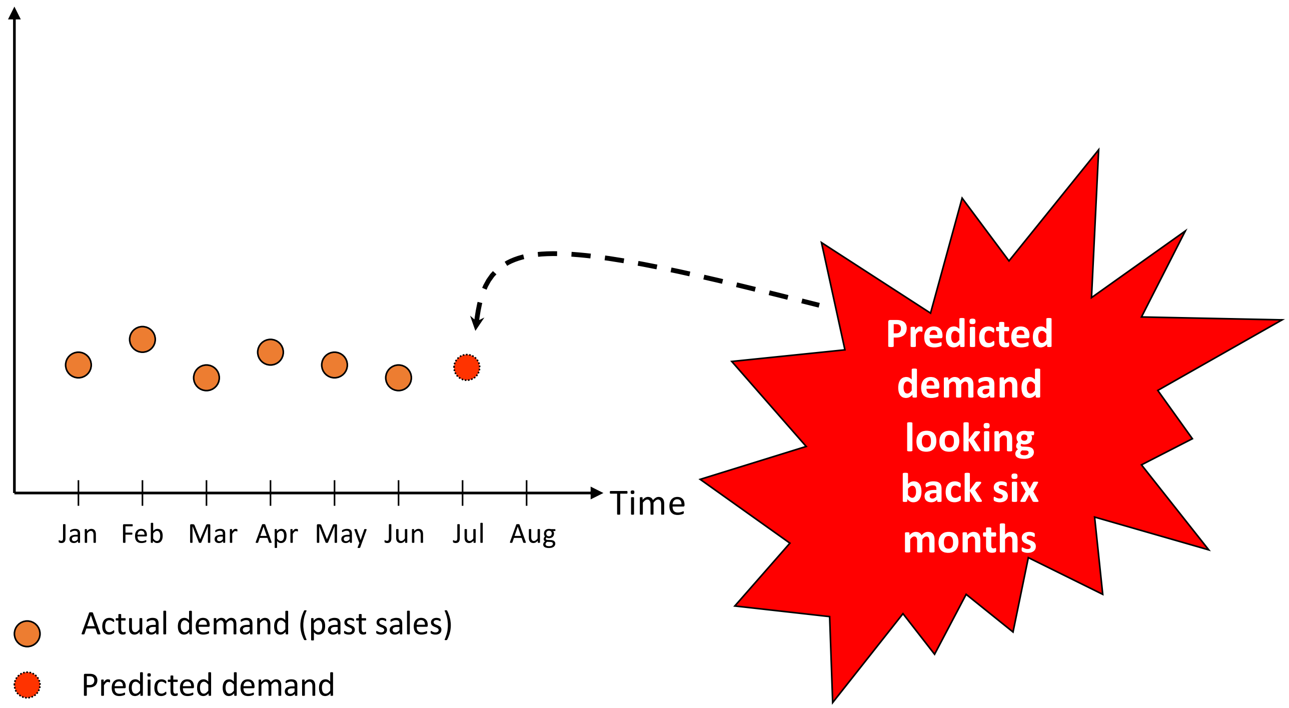
- Time Point Series univariate – indexing, match and similarity, search and retrieval, and more
- Multivariate Temporal Data – time series analysis, or heterogeneous variables
- Forecasting
- Clustering
- Classification
- Patterns Discovery
- And more .. we will go through the topics in few slides

Univariate Time Series Indexing and Matching

- Indexing and retrieval
- Using raw time series values
- Similarity functions (Euclidean, Dynamic Time Warping,..)
- Using discretization (PAA, SAX, TD4C,..)



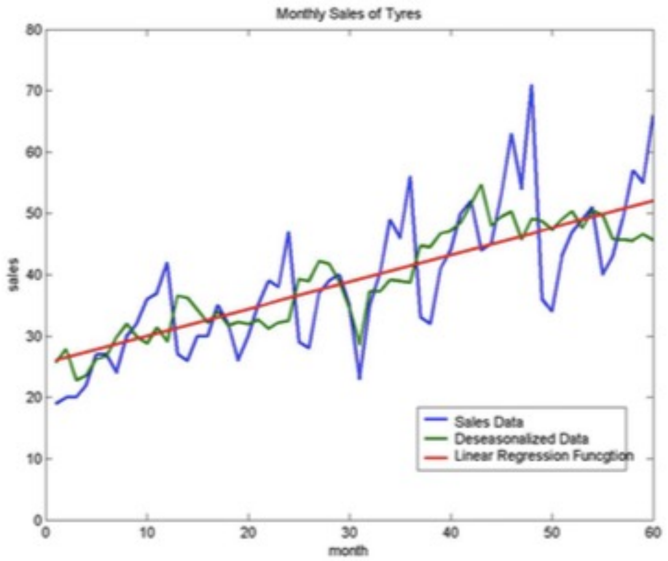
Forecasting



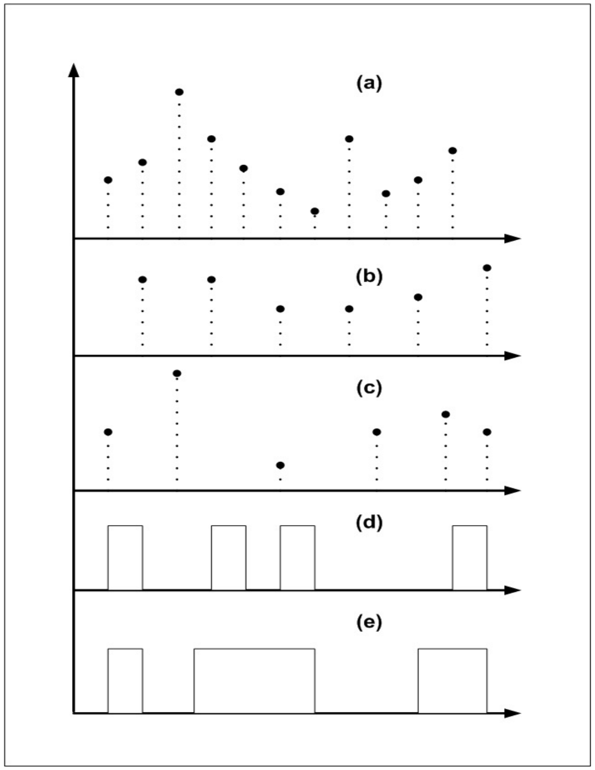
Predicted demand looking back six months

Forecasting

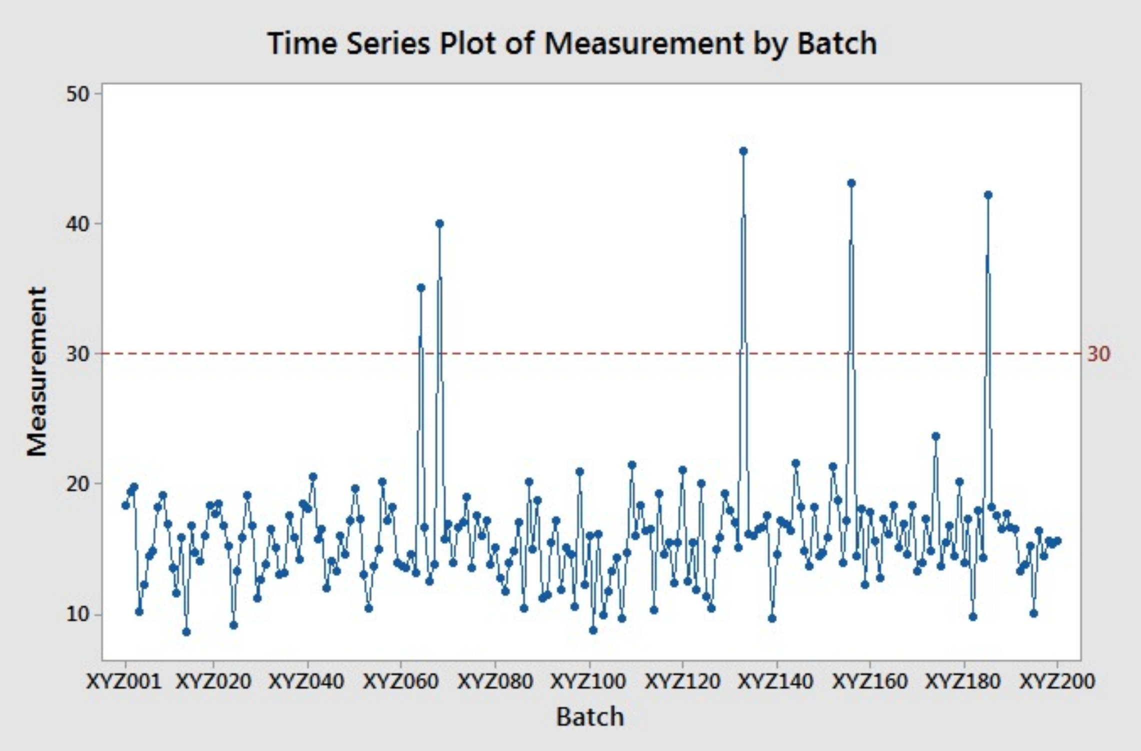
- Forecasting:
  - Autoregressive Models
  - Autoregressive Moving Average Models
    - moving average
    - Weighted moving average
    - Exponential moving average
  - Multivariate Forecasting with Hidden Variables
    - ARMA
    - ARIMA



Multivariate Heterogeneous Temporal Data



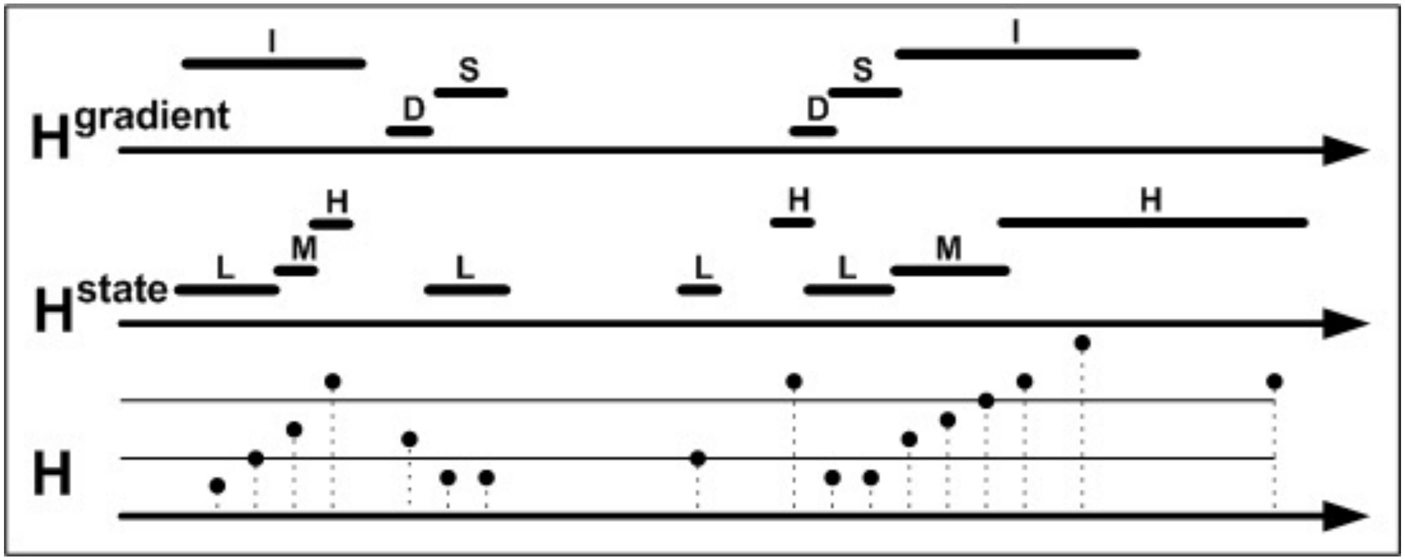
Outliers and Anomalies



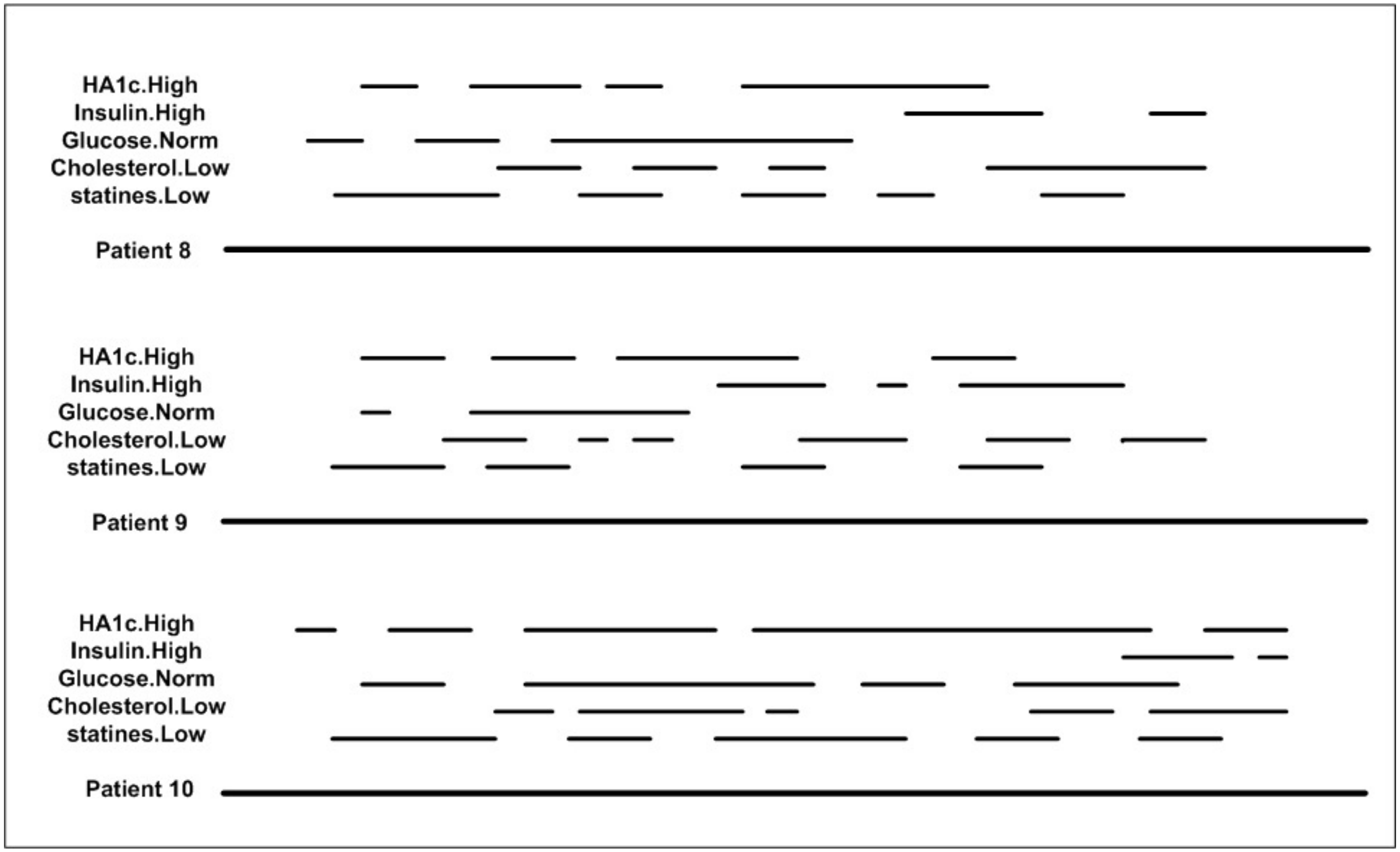
Clustering of Multivariate Temporal Data

- Through clustering of univariate or multivariate time series we can determine common types of ”temporal behavior”
- Clustering via traditional ”static” methods
- Similarity temporal functions
- Clustering via frequent temporal patterns, especially useful for multivariate clustering through frequent temporal patterns:
  - Sequential mining
  - Time intervals mining
  - Markov Chains

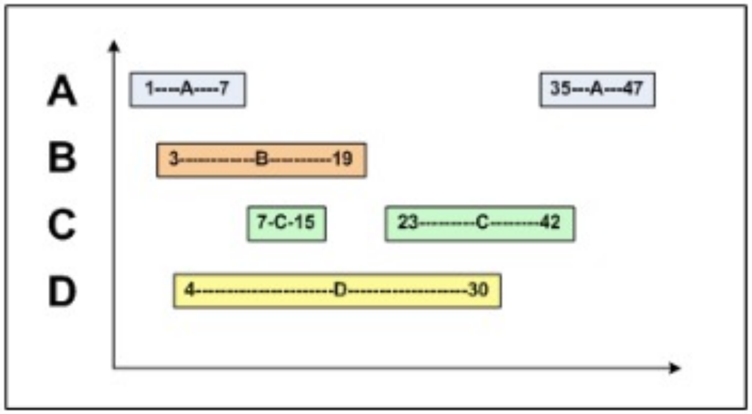
From Time Points to Time Intervals Series



Time Intervals Related Patterns Discovery – an illustration



Time Intervals Related Pattern - **TIRP**


A
B
C
D
1---A----7
35---A---47
3-----------B----------19
7-C-15
23---------C---------42
4--------------------D-----------------30

A TIRP is a conjunction of pairwise temporal relations

{A_1 o B, A_1 o D, A_1 m C_1, A_1 b C_2, A_1 b A_2, B o D, B c C_1, B b C_2, B b A, C_1 b C_2, C_1 b A, C_2 o A}

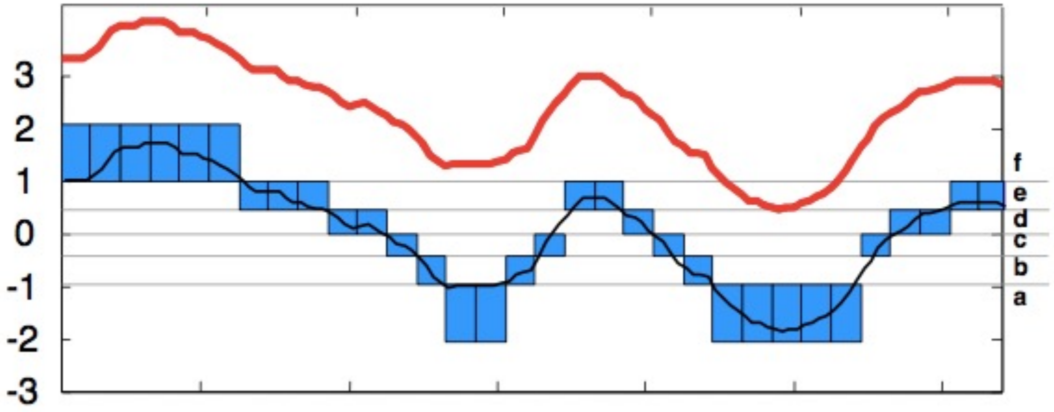
A k-sized TIRP includes k(k-1)/2 = (k^2-k)/2 temporal relations

SAX [Lin et al, 2003]

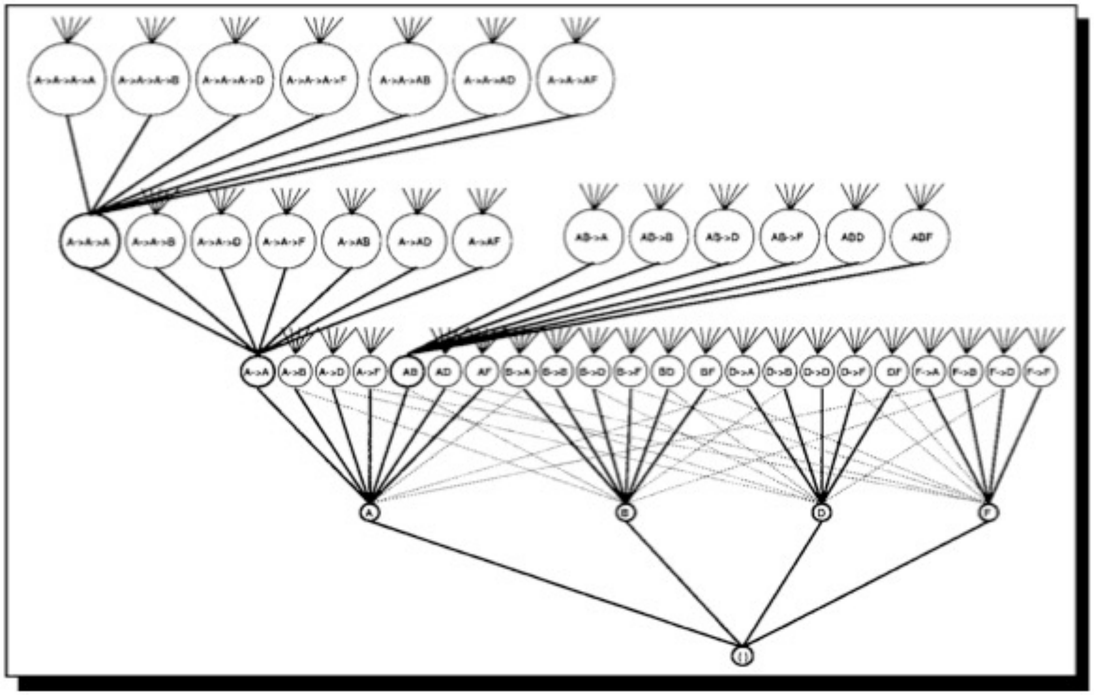
- The states' cutoffs are defined by the normal distribution of the values: mean and standard-deviations

- SAX provides a tradeoff between efficiency and approximation accuracy.

- Can be:
  - Symbolic time series
  - Symbolic time intervals



Sequential Mining



Classification of Temporal Data

- Classification of univariate/multivariate
- Related issues - Time windows, imputation (e.g., mean values)
- Discretization (unsupervised/supervised)
- Features:
    - Markov Models as features
    - Shapelets
    - Frequent Sequences (patterns)
    - Furrier-Transforms as features
    - More ..

Frequent Temporal Patterns Discovery

- Sequential Mining
- Time Intervals Mining
- Their use for Temporal Knowledge Discovery
- Their use for Classification
- Their use for Clustering (Each pattern is a culster)
- Metrics for discovery
- Interestingness measures
- Visualization

Other topics

- Other temporal data mining research topics proposed by the students are possible too, after approval.
- Students are encouraged to work on topics from their thesis (msc or phd)

Data Mining

Since taking Machine Learning or Data Mining courses is not a prerequisite, we will do a brief overview of Data Mining ..

Why (Temporal) Data Mining?

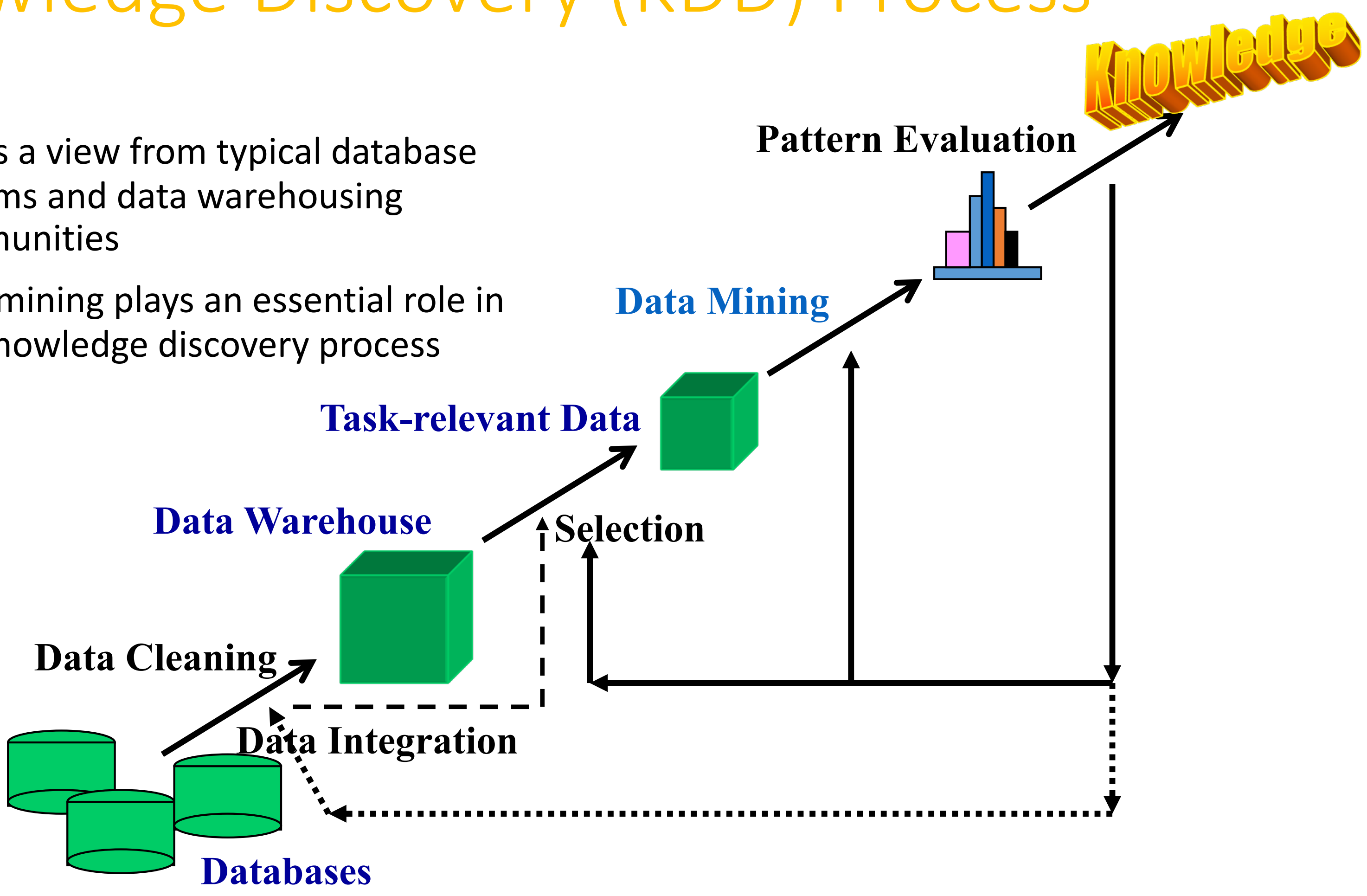
- The Explosive Growth of Data: from terabytes to petabytes
  - Data collection and data availability
    - Automated data collection tools, database systems, Web, computerized society, IoT
  - Major sources of abundant data
    - Business: Web, e-commerce, transactions, biomedical informatics, stocks, …
    - Science: Remote sensing, bioinformatics, scientific simulation, …
    - Society and everyone: news, digital cameras, YouTube
- We are drowning in data, but starving for knowledge!
- “Necessity is the mother of invention”—Data mining—Automated analysis of massive data set

What is Data Mining?

- Data mining (knowledge discovery from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- Watch out: Is everything "data mining"?
  - Simple search and query processing
  - expert systems

Knowledge Discovery (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



Example: Medical Data Mining

- Health care & medical data mining – often adopted such a view in statistics and machine learning
  - Typically, in medical data longitudinal analysis is intrinsic and crucial
- Preprocessing of the data (including feature extraction and dimension reduction)
- Classification or/and clustering processes
- Post-processing for presentation

Data Mining: Association and Correlation Analysis

- Frequent patterns (or frequent itemsets, for example)
  - What items are frequently purchased together in your Walmart?

- Association, correlation vs. causality
  - A typical association rule
    - Diaper → Beer [0.5%, 75%] (support, confidence)
  - Are strongly associated items also strongly correlated?

- How to mine such patterns and rules efficiently in large datasets?

- How to use such patterns for classification, clustering, and other applications?

Data Mining: Classification

- Classification and label prediction
  - Construct models (functions) based on some training examples
  - Describe and distinguish classes or concepts for future prediction
    - E.g., classify countries based on (climate), or classify cars based on (gas mileage)
- Typical methods
  - Decision trees, naïve Bayesian classification, support vector machines, neural networks (deep learning), rule-based classification, pattern-based classification, logistic regression, …,..
- Typical applications:
  - Credit card fraud detection, direct marketing, diseases, web-pages, …
  - Temporal – Outcomes Prediction and ”forecasting”, Diagnose, Reason ..

Data Mining: Cluster Analysis

- Unsupervised learning (i.e., Class label is unknown)
- Group data to form new categories (i.e., clusters), e.g., cluster patients to find disease progress patterns
- Principle: Maximizing intra-class similarity & minimizing interclass similarity
- Many methods and applications
  - In the temporal context, it is especially relevant to temporal patterns discovery that are – clusters of temporal behavior
  - Clustering stocks longitudinally
  - Clustering users behaviors on the internet

Data Mining: Outlier Analysis

- Outlier analysis
  - Outlier: A data object that does not comply with the general behavior of the data
  - Noise or exception? — One person's garbage could be another person's treasure
  - Methods: by product of clustering or regression analysis, …
  - Useful in fraud detection, rare events analysis

- Outliers may be looked for error measurements, but another perspective is anomaly detection
- Anomaly detection can be a deviation from the a typical (temporal) behavior ..

Time and Ordering: Sequential Pattern, Trend and Evolution Analysis

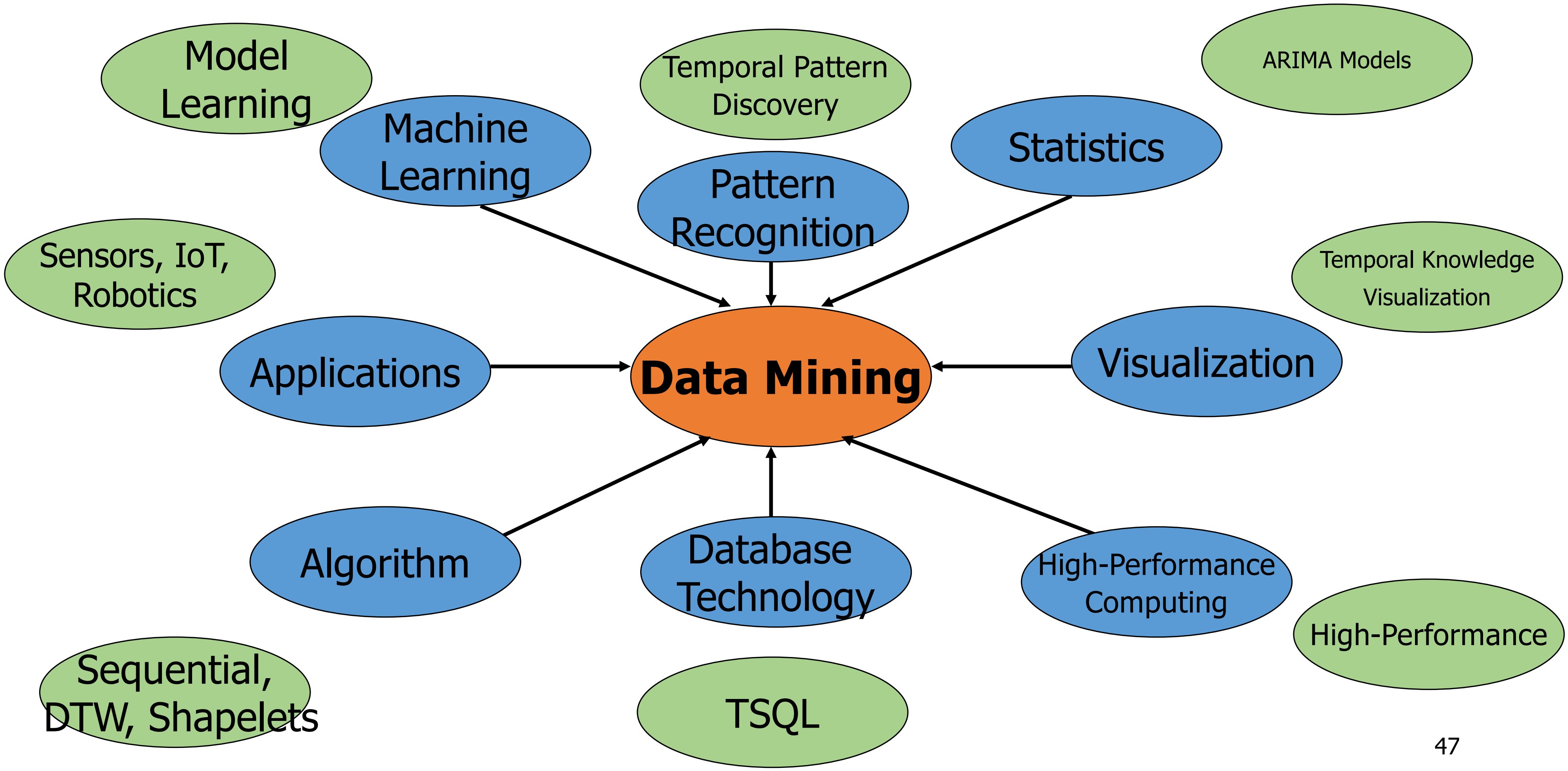
- Sequence, trend and evolution analysis
  - Trend, time-series, and deviation analysis: e.g., regression and value prediction
  - Sequential pattern mining
    - e.g., first smart phone, then buy smart watch
  - Periodicity analysis
  - Similarity-based analysis

- Mining data streams
  - Ordered, time-varying, potentially infinite, data streams
  - Piecewise Aggregate Approximation, SAX,..

Evaluation of Knowledge

- Are all mined knowledge interesting?
  - One can discover tremendous amount of "patterns" and knowledge
  - Some may fit only certain dimension space (time, location, …)
- We want meaningful, ideally significant, actionable knowledge
- Evaluation of mined knowledge → directly mine only interesting knowledge?
  - Descriptive vs. Predictive
  - Typicality vs. novelty
  - Accuracy
  - …

Data Mining: Confluence of Multiple Disciplines



Applications of Data Mining and temporal

- Web page analysis: from web page classification, clustering to PageRank
  - Temporal – Click Stream Analysis, and Sequential Pages Analysis

- Recommender systems
  - Temporal - sequence based recommendations (People who bought A, bought B after 2 months)

- Basket data analysis to targeted marketing
  - Temporal – sequences of basket purchases

- Biological and medical data analysis: classification, cluster analysis, biological sequence analysis, biological network analysis
  - Temporal – Electronic Health Records Analysis

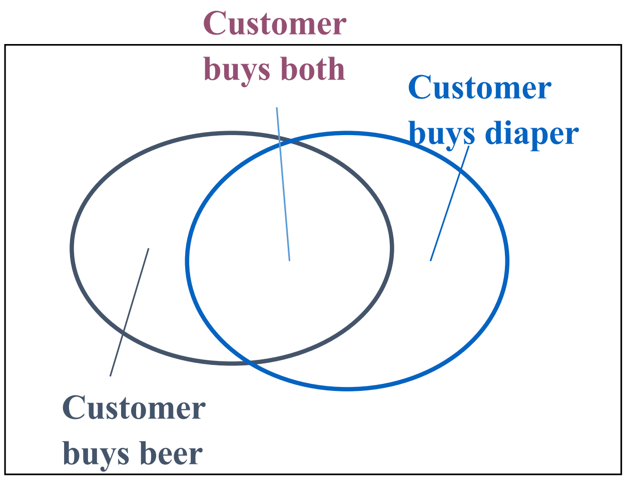
Temporal Data Mining Venues and Journals

- Conferences:
- ACM SIGKDD Int. Conf. on Knowledge Discovery in Databases and Data Mining (KDD)
- SIAM Data Mining Conf. (SDM)
- IEEE Int. Conf. on Data Mining (ICDM)
- European Conf. on Machine Learning and Principles and practices of Knowledge Discovery and Data Mining (ECML-PKDD)
- Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD)
- Other related conferences
  - DB conferences: ACM SIGMOD, VLDB, ICDE, …
  - Web and IR conferences: WWW, SIGIR, ..
  - ICML, AAAI, IJCAI, ..
- Journals: Data Mining and Knowledge Discovery (DAMI or DMKD), IEEE Trans. On Knowledge and Data Engineering (TKDE), KDD Explorations, ACM Transactions on KDD (TKDD), Knowledge and Information Systems (KAIS)

Overview on Tools in Data Mining/Machine Learning

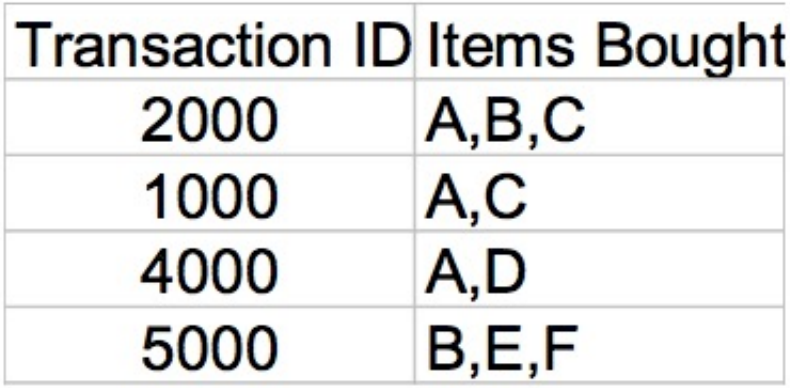
- Time Series Analysis
  - Forecasting, Auto Regression, ARIMA
- Pattern Mining
  - Association Rules Mining, Sequential Mining, Time Intervals Mining
- Clustering
  - K-Means, Hierarchical Clustering
- Classification
  - Decision Trees, Random Forests, Naïve Bayes, Deep Learning

Association Metrics: Support and Confidence



Find all the rules *X & Y Z* with minimum confidence and support

- support, *s*, probability that a transaction contains {X & Y Z}
- confidence, *c*, conditional probability that a transaction having {X & Y} also contains *Z*



Transaction ID	Items Bought
2000	A,B,C
1000	A,C
4000	A,D
5000	B,E,F

Let minimum support 50%, and minimum confidence 50%, we have

- *A C* (50%, 66.6%)
- *C A* (50%, 100%)

Applications Examples

- Market Basket Analysis
  - *Maintenance Agreement* (What the store should do to boost Maintenance Agreement sales?)
  - *Home Electronics* (What other products should the store stocks up on if the store has a sale on Home Electronics?)
  - Drug Drug Interactions -> Adverse Events Reactions (Conditions or AERs?)
  - Attached mailing in direct marketing
  - AMAZON

Association Rules Mining – Problem Statement

- $\mathcal{I} = \{i_1, i_2, \ldots, i_m\}$: a set of literals, called items
- Transaction $\mathcal{T}$: a set of items s.t. $\mathcal{T}$ □ $\mathcal{I}$
- Database $\mathcal{D}$: a set of transactions
- A transaction contains X, a set of items in $\mathcal{I}$, if X□$\mathcal{T}$
- An association rule is an implication of the form X Y,

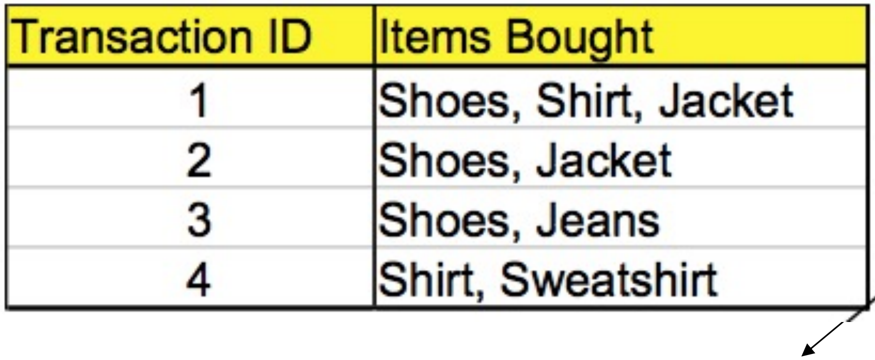
  where X,Y□$\mathcal{I}$
- The rule X Y has support s in the transaction set $\mathcal{D}$ if s% of transactions in $\mathcal{D}$ contain X and Y
- The rule X Y holds in the transaction set $\mathcal{D}$ with confidence c if c% of transactions in $\mathcal{D}$ that contain X also contain Y [sup(X,Y)/sup(X)]
- Find all rules that have support and confidence greater than user-specified min support and min confidence

Problem Decomposition

1. Find all sets of items that have minimum support (frequent itemsets)

2. Use the frequent itemsets to generate the desired rules

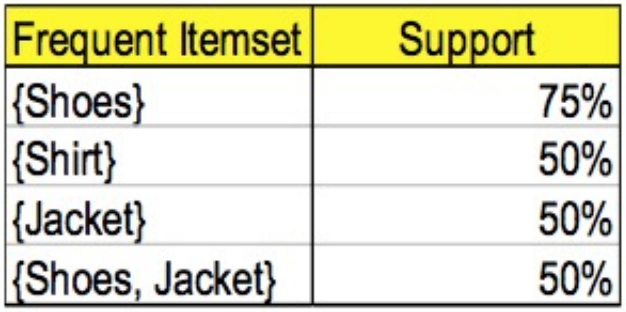
Problem Decomposition – Example



Transaction ID	Items Bought
1	Shoes, Shirt, Jacket
2	Shoes, Jacket
3	Shoes, Jeans
4	Shirt, Sweatshirt

For min support = 50% = 2 trans,

and min confidence = 50%



Frequent Itemset	Support
{Shoes}	75%
{Shirt}	50%
{Jacket}	50%
{Shoes, Jacket}	50%

For the rule Shoes ⍰ Jacket

- Support = Sup({Shoes,Jacket)}=50%
- Confidence = $\frac{50\%}{75\%}$ = 66.6%

{Jacket , *Shoes} has 50% support and 100% confidence*

Discovering Rules

- Naïve Algorithm

```
for each frequent itemset $l$ do
 for each subset $c$ of $l$ do
   if (support($l$) / support($l - c$) >= minconf) then
     output the rule ($l - c$)  $c$,
       with confidence = support($l$) / support ($l - c$ )
        and support = support($l$)
```

Mining Frequent Itemsets: the Key Step

- Find the *frequent itemsets*: the sets of items above minimum support
  - A subset of a frequent itemset must also be a frequent itemset
    - i.e., if {*AB*} is a frequent itemset, both {*A*} and {*B*} should be a frequent itemset
  - Iteratively find frequent itemsets with cardinality from 1 to *k (k*-itemset*)*
- Use the frequent itemsets to generate association rules.

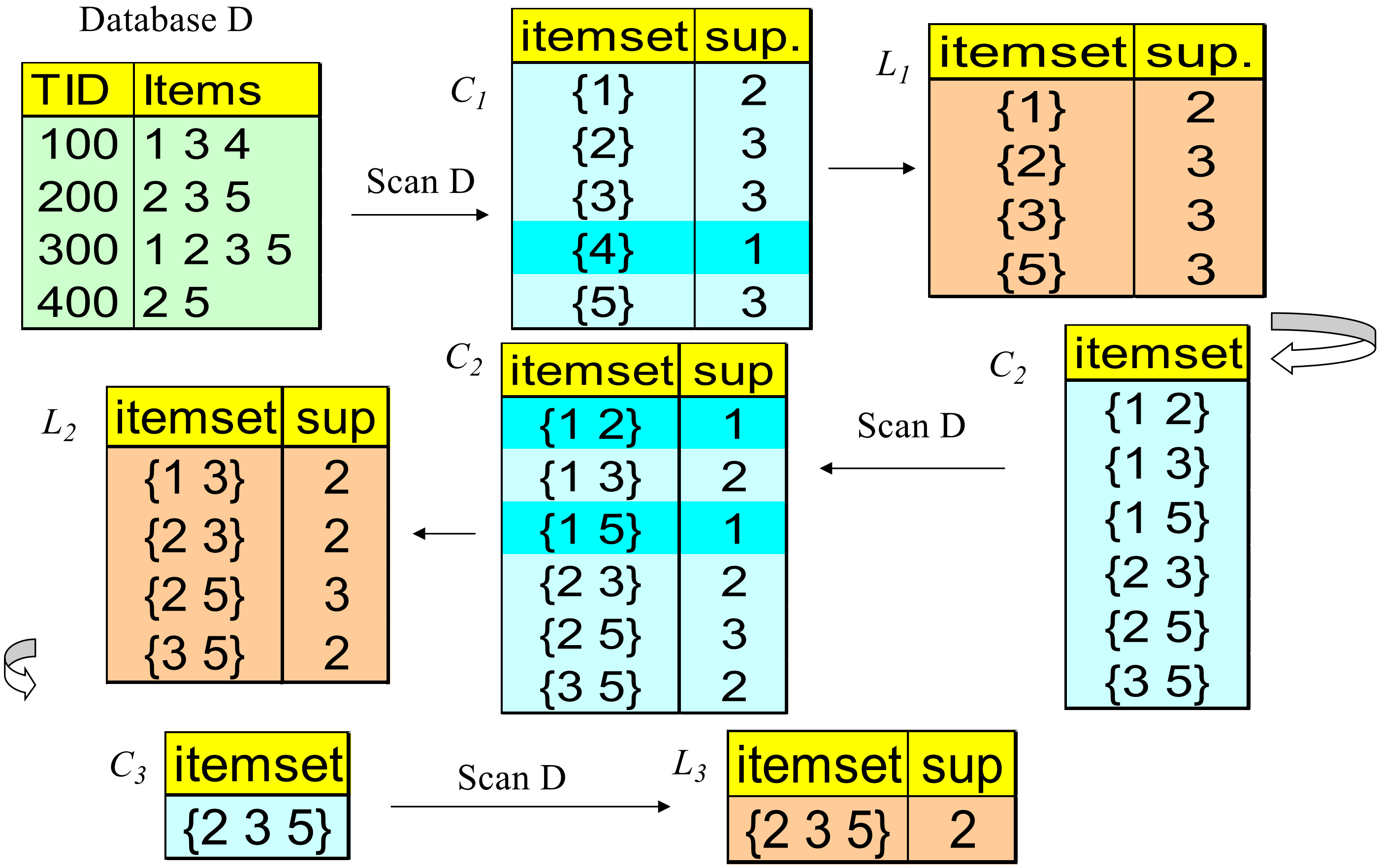
The Apriori Algorithm

- L_k: Set of frequent itemsets of size k (those with min support)
- C_k: Set of candidate itemset of size k (potentially frequent itemsets)

```
$L_1$ = {frequent items};
for ($k$ = 1; $L_k$ != K ; $k$++) do begin
    $C_{k+1}$ = candidates generated from $L_k$;
   for each transaction $t$ in database do
          increment the count of all candidates in $C_{k+1}$
        that are contained in $t$
   $L_{k+1}$ = candidates in $C_{k+1}$ with min_support
   end
return $L_k$;
```

The Apriori Algorithm — Example

Min support =50% = 2 trans



Database D

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

Scan D

C_1

itemset	sup.
{1}	2
{2}	3
{3}	3
{4}	1
{5}	3

L_1

itemset	sup.
{1}	2
{2}	3
{3}	3
{5}	3

C_2

itemset
{1 2}
{1 3}
{1 5}
{2 3}
{2 5}
{3 5}

Scan D

C_2

itemset	sup
{1 2}	1
{1 3}	2
{1 5}	1
{2 3}	2
{2 5}	3
{3 5}	2

L_2

itemset	sup
{1 3}	2
{2 3}	2
{2 5}	3
{3 5}	2

C_3

itemset
{2 3 5}

Scan D

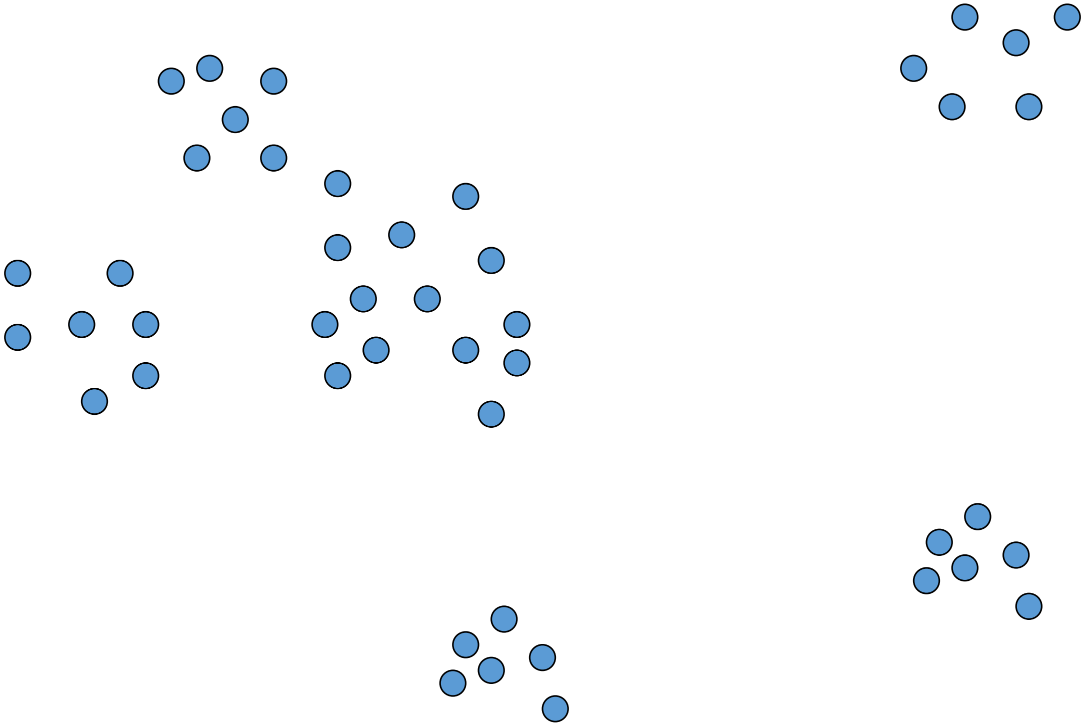
L_3

itemset	sup
{2 3 5}	2

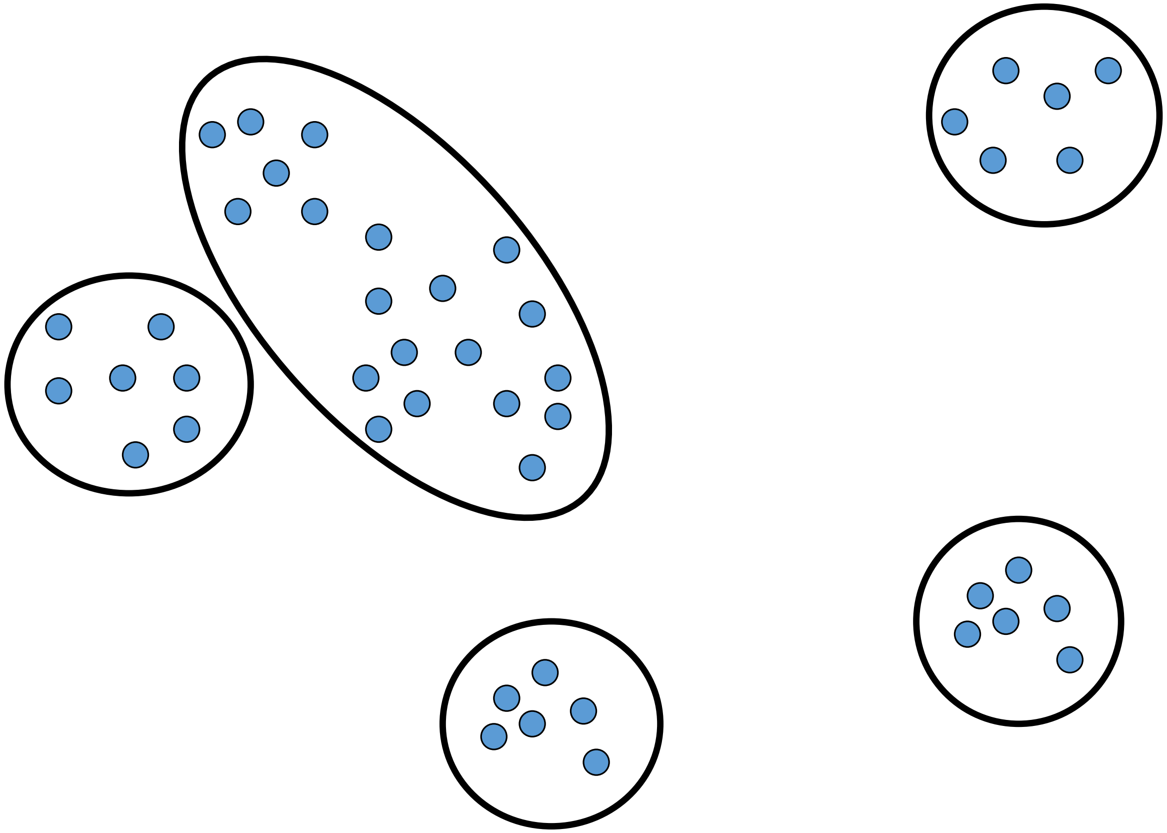
Methods to Improve Apriori's Efficiency

- Transaction reduction: A transaction that does not contain any frequent k-itemset is useless in subsequent scans
- Partitioning: Any itemset that is potentially frequent in DB must be frequent in at least one of the partitions of DB
- Sampling: mining on a subset of given data
- Dynamic itemset counting: add new candidate itemsets only when all of their subsets are estimated to be frequent (apriori-all)

Clustering



Clustering



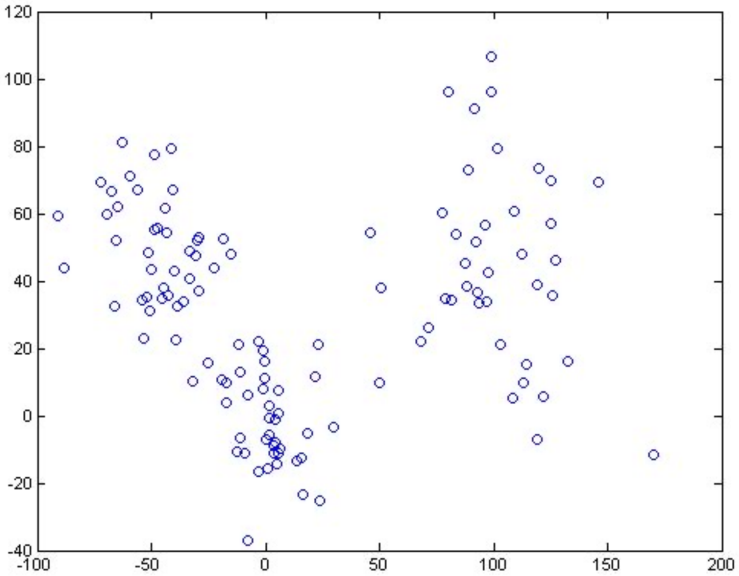
Clustering considerations

- What does it mean for objects to be similar?
- What algorithm and approach do we take?
  - Top-down: k-means
  - Bottom-up: hierarchical agglomerative clustering
- Do we need a hierarchical arrangement of clusters?
- How many clusters?
- Can we label or name the clusters?
- How do we make it efficient and scalable?

K-means Clustering

- Choose a number of clusters *k*
- Initialize cluster centers c_1,… c_k
  - Could pick *k* data points and set cluster centers to these points
  - Or could randomly assign points to clusters and take means of clusters
- For each data point, compute the cluster center it is closest to (using some distance measure) and assign the data point to this cluster
- Re-compute cluster centers (mean of data points in cluster)
- Stop when there are no new re-assignments

K-means Clustering (cont.)



How many clusters do you think there are in this data? How might it have been generated?

K-means Clustering Demo

$k = 2$

K-means Clustering Issues

- Random initialization means that you may get different FINAL clusters each time

- Data points are assigned to only one cluster (hard assignment)

- You have to pick the number of clusters…

Determining the “correct” number of clusters

- We’d like to have a measure of cluster quality *Q* and then try different values of *k* until we get an optimal value for *Q*
- But, since clustering is an unsupervised learning method, we can’t really expect to find a “correct” measure *Q*…
- So, once again there are different choices of *Q* and our decision will depend on what dissimilarity measure we’re using and what types of clusters we want

Cluster Quality Measures

- a measure that emphasizes cluster tightness or homogeneity:

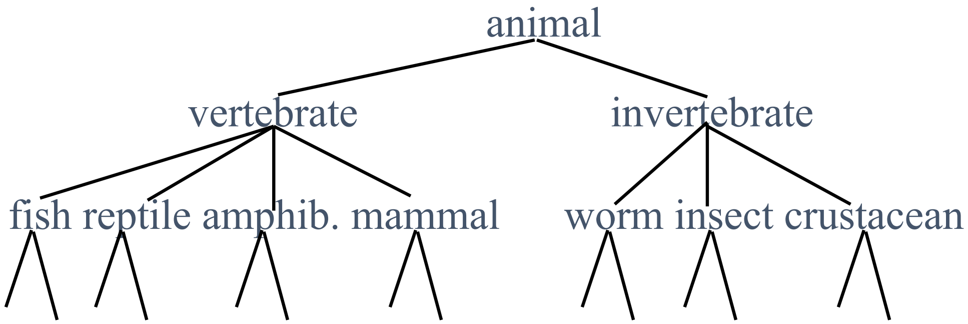
$$Q = \sum_{i=1}^{k} \frac{1}{|C_i|} \sum_{\boldsymbol{x} \in C_i} d(\boldsymbol{x}, \mu_i)$$

- $|C_i|$ is the number of data points in cluster i

- Q will be small if (on average) the data points in each cluster are close

Hierarchical Clustering

- Build a tree-based hierarchical taxonomy (*dendrogram*) from a set of documents.

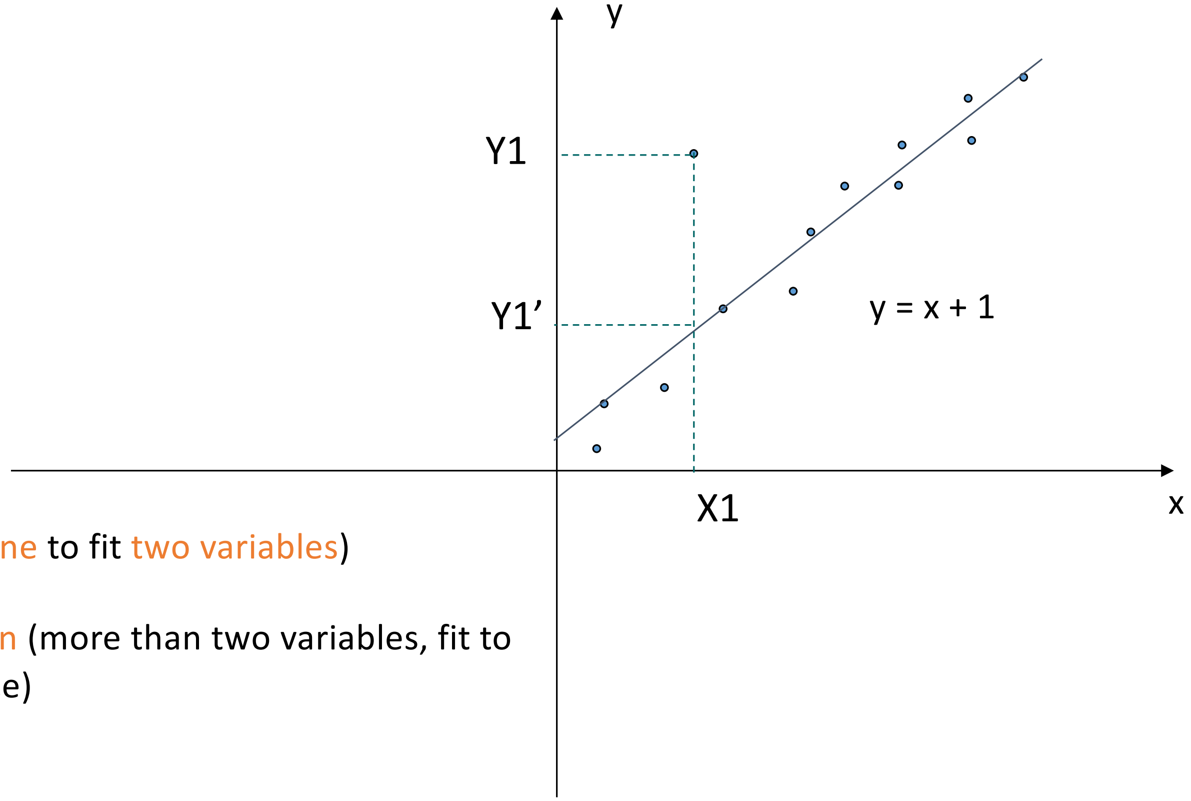


How could you do this with k-means?

Hierarchical Clustering algorithms

- **Agglomerative (bottom-up):**
  - Start with each document being a single cluster.
  - Eventually all documents belong to the same cluster.
- **Divisive (top-down):**
  - Start with all documents belong to the same cluster.
  - Eventually each node forms a cluster on its own.
  - Could be a recursive application of k-means like algorithms
- Does not require the number of clusters ***k*** in advance
- Needs a termination/readout condition

Regression



- Linear regression (best line to fit two variables)

- Multiple linear regression (more than two variables, fit to a multidimensional surface)

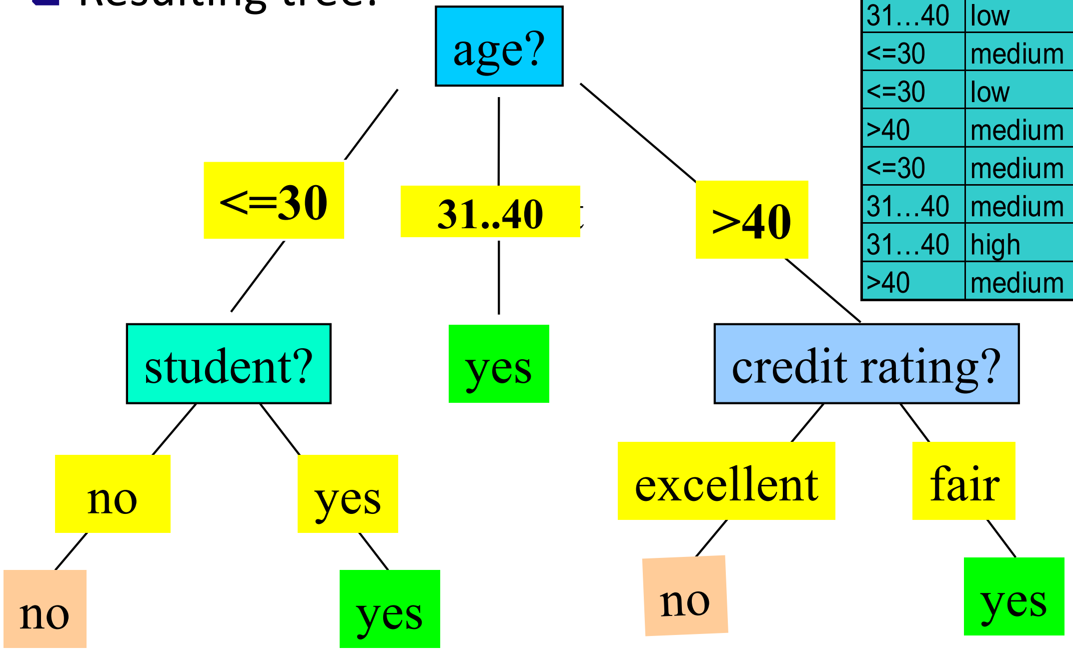
Prediction: Classification vs. Numeric

- Classification
  - predicts categorical class labels (discrete or nominal)
  - classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data
- Numeric Prediction
  - models continuous-valued functions, i.e., predicts unknown or missing values (imputation)
- Typical applications
  - Credit/loan approval, Medical diagnosis, Fraud detection, Web page categorization: which category it is, or recommendation
  - Temporal – incorporates multivariate temporal behavior

Decision Tree Induction: An Example

- Training data set: Buys_computer
- The data set follows an example of Quinlan's ID3 (Playing Tennis)
- Resulting tree:

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no



Bayesian Classification: Why?

- A statistical classifier: performs *probabilistic prediction*, *i.e.*, predicts class membership probabilities

- Foundation: Based on Bayes' Theorem.

- Incremental: Each training example can incrementally increase/decrease the probability that a hypothesis is correct — prior knowledge can be combined with observed data

Prediction Based on Bayes' Theorem

- Given training data **X**, *posteriori probability of a hypothesis* H*,* P(H|**X**)*,* follows the Bayes' theorem

$$P(H\,|\,\mathbf{X}) = \frac{P(\mathbf{X}|H)P(H)}{P(\mathbf{X})} = P(\mathbf{X}|H) \times P(H) / P(\mathbf{X})$$

- Informally, this can be viewed as

  posteriori = likelihood x prior/evidence

- Practical difficulty: It requires initial knowledge of many probabilities, involving significant computational cost

Model Evaluation and Selection

- Evaluation metrics: How can we measure accuracy? Other metrics to consider?
- Use **validation test set** of class-labeled tuples instead of training set when assessing accuracy

- Methods for estimating a classifier's accuracy:
  - Cross-validation

- Comparing classifiers:
  - Confidence intervals
  - Cost-benefit analysis and ROC Curves

Classifier Evaluation: Confusion Matrix

Confusion Matrix:

Actual class\Predicted class	C_1	¬ C_1
C_1	**True Positives (TP)**	**False Negatives (FN)**
¬ C_1	**False Positives (FP)**	**True Negatives (TN)**

Example of Confusion Matrix:

Actual class\Predicted class	buy_computer = yes	buy_computer = no	Total
buy_computer = yes	**6954**	**46**	7000
buy_computer = no	**412**	**2588**	3000
Total	7366	2634	10000

Classifier Evaluation : Accuracy, Error Rate, Sensitivity and Specificity

A\P	C	¬C	
C	**TP**	**FN**	**P**
¬C	**FP**	**TN**	**N**
	P’	**N’**	**All**

- **Classifier Accuracy,** or recognition rate: percentage of test set tuples that are correctly classified

  Accuracy = (TP + TN)/All

- **Error rate:** *1 – accuracy*, or

  Error rate = (FP + FN)/All

- **Class Imbalance Problem**:
  - One class may be *rare*, e.g. fraud, or HIV-positive
  - Significant *majority of the negative class* and minority of the positive class

Classifier Evaluation: Precision and Recall, and F-measures

- **Precision**: exactness – what % of tuples that the classifier labeled as positive are actually positive

$$precision = \frac{TP}{TP + FP}$$

- **Recall:** completeness – what % of positive tuples did the classifier label as positive?

$$recall = \frac{TP}{TP + FN}$$

- Perfect score is 1.0
- Inverse relationship between precision & recall
- ***F* measure (*F_1* or *F*-score)**: harmonic mean of precision and recall,

$$F = \frac{2 \times precision \times recall}{precision + recall}$$

- ***$F_{ß}$*:** weighted measure of precision and recall
  - assigns ß times as much weight to recall as to precision

$$F_\beta = \frac{(1 + \beta^2) \times precision \times recall}{\beta^2 \times precision + recall}$$

Evaluating Classifier Accuracy: Holdout & Cross-Validation Methods

- **Holdout method**
  - Given data is randomly partitioned into two independent sets
    - Training set (e.g., 2/3) for model construction
    - Test set (e.g., 1/3) for accuracy estimation
  - Random sampling: a variation of holdout
    - Repeat holdout k times, accuracy = avg. of the accuracies obtained
- **Cross-validation** (*k*-fold, where k = 10 is most popular)
  - Randomly partition the data into *k mutually exclusive* subsets, each approximately equal size
  - At *i*-th iteration, use D_i as test set and others as training set
  - Leave-one-out: *k* folds where *k* = # of tuples, for small sized data
  - ***Stratified cross-validation***: folds are stratified so that class dist. in each fold is approx. the same as that in the initial data

More in DM and ML

- There are more topics related to Data Mining and Machine Learning
  - Feature Selection, Dimensionality reduction
  - Missing values and imputation
  - and more ..

Recommended References

- J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 3rd ed., 2011
- T. M. Mitchell, Machine Learning, McGraw Hill, 1997
- I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 2nd ed. 2005
- T. Mitsa, Temporal Data Mining, Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, 2010.

Complex Data Analytics Lab

robertmo@bgu.ac.il

Building 96, room 310