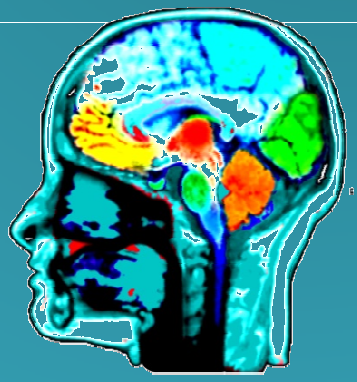
Introduction To Data Mining

Isfahan University of Technology (IUT)

Bahman 1401

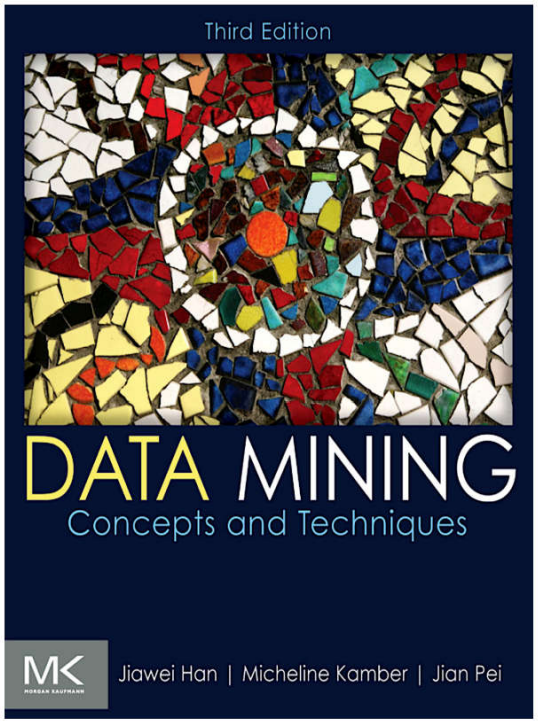


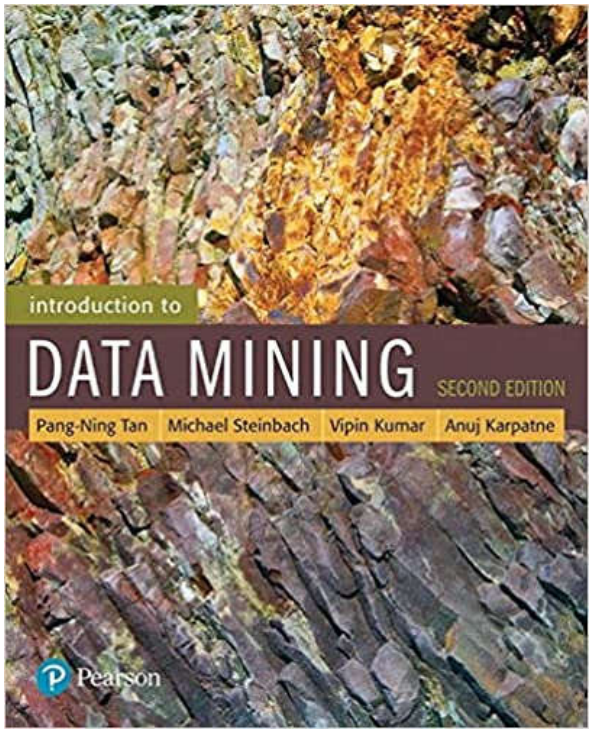
Introduction

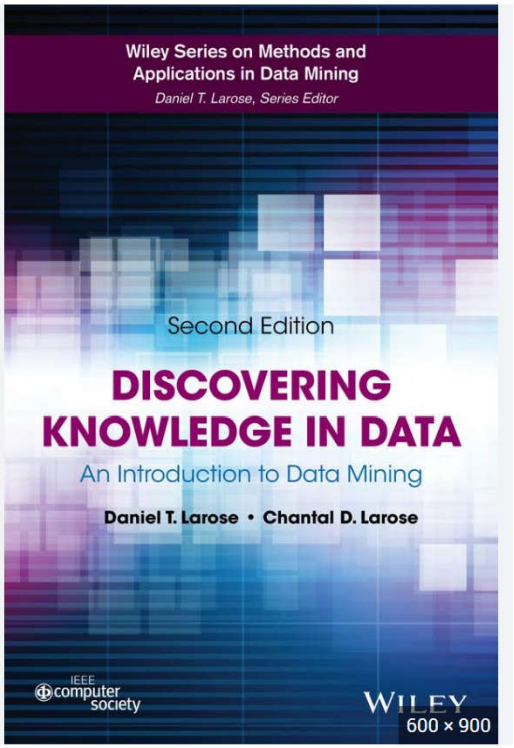
Dr. Hamidreza Hakim
hamid.hakim.u@gmail.com

Content

Refrences


Third Edition
DATA MINING
Concepts and Techniques
Jiawei Han | Micheline Kamber | Jian Pei


introduction to
DATA MINING
SECOND EDITION
Pang-Ning Tan
Michael Steinbach
Vipin Kumar
Anuj Karpatne
Pearson

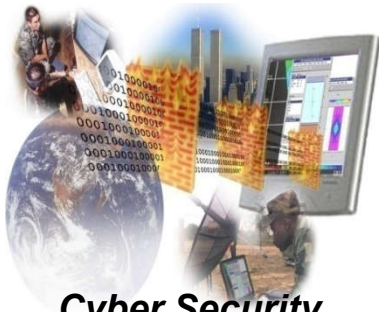

Wiley Series on Methods and Applications in Data Mining
Daniel T. Larose, Series Editor
Second Edition
DISCOVERING KNOWLEDGE IN DATA
An Introduction to Data Mining
Daniel T. Larose • Chantal D. Larose
IEEE computer society
WILEY
600 × 900

Grades

- Project and Presentation: 2 points
- Exercises: 5 points
- Exams and Quizzes: 13 points

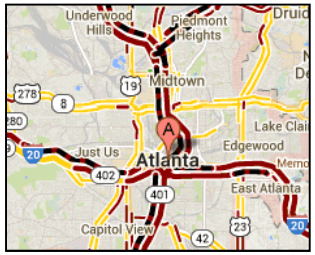
Large-scale Data is Everywhere!

- There has been enormous data growth in both commercial and scientific databases due to advances in data generation and collection technologies
- New mantra
  - Gather whatever data you can whenever and wherever possible.
- Expectations
  - Gathered data will have value either for the purpose collected or for a purpose not envisioned.

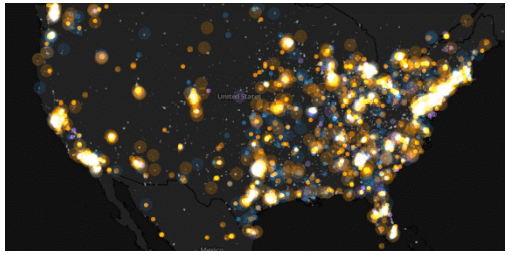


Cyber Security

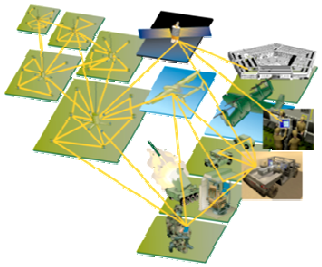
E-Commerce



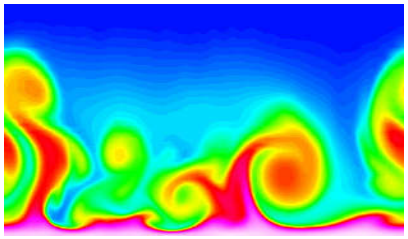
Traffic Patterns



Social Networking: Twitter



Sensor Networks



Computational Simulations

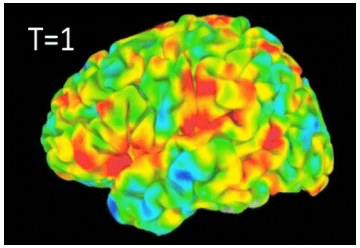
Why Data Mining? Commercial Viewpoint

- Lots of data is being collected and warehoused
  - Web data
    - Google has Peta Bytes of web data
    - Facebook has billions of active users
  - purchases at department/grocery stores, e-commerce
    - Amazon handles millions of visits/day
  - 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)

Why Data Mining? Scientific Viewpoint

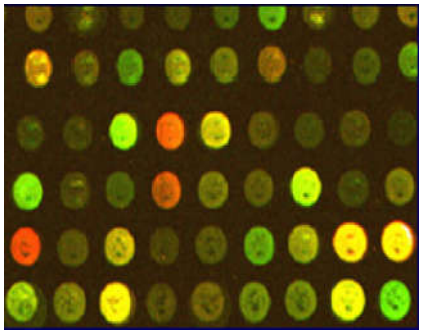
- Data collected and stored at enormous speeds
  - Remote sensors on a satellite
    - NASA EOSDIS archives over petabytes of earth science data/year
  - Telescopes scanning the skies
    - Sky survey data
  - High-throughput biological data
  - Scientific simulations
    - terabytes of data generated in a few hours



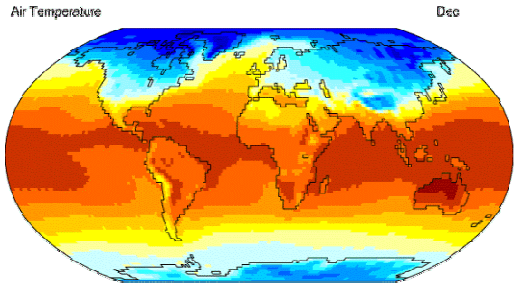
fMRI Data from Brain

پردازش داده های مغزی

Sky Survey Data



Gene Expression Data



Surface Temperature of Earth

بررسی وضعیت اب و هوا

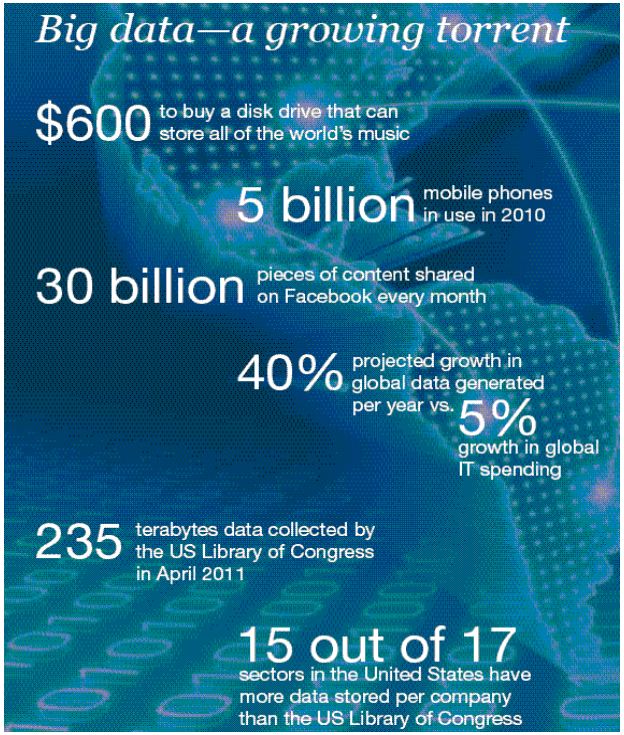
- Data mining helps scientists
  - in automated analysis of massive datasets
  - In hypothesis formation

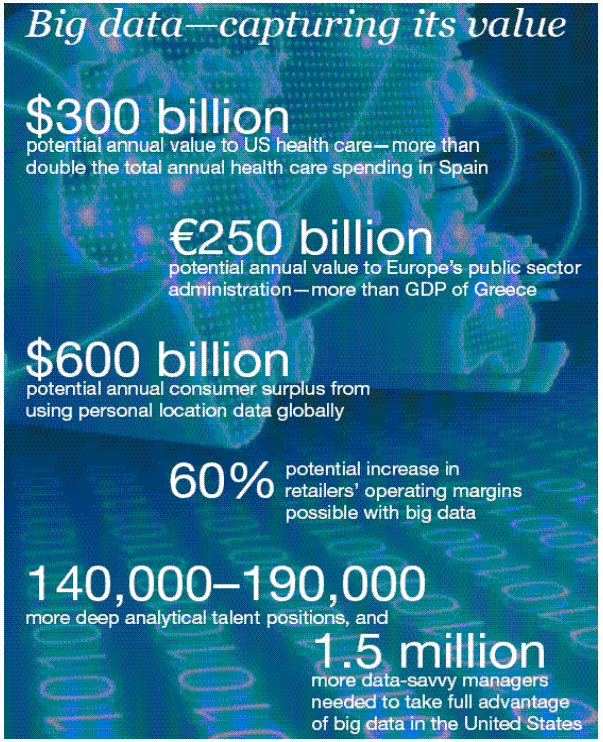
Great opportunities to improve productivity in all walks of life

McKinsey Global Institute

Big data: The next frontier for innovation, competition, and productivity

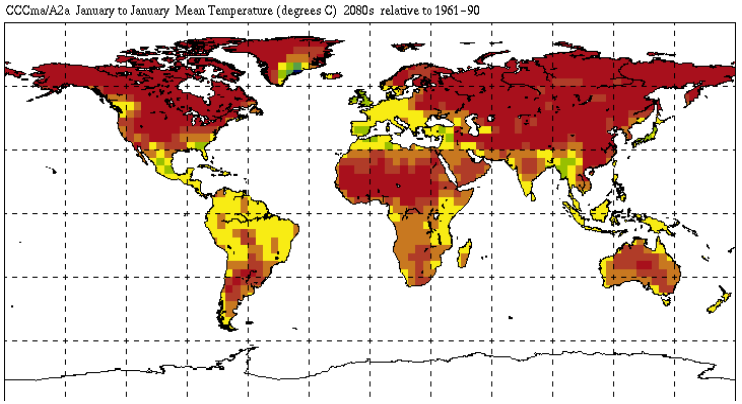
کمک برای بهبود مسائل





Great Opportunities to Solve Society's Major Problems

Improving health care and reducing costs



Predicting the impact of climate change

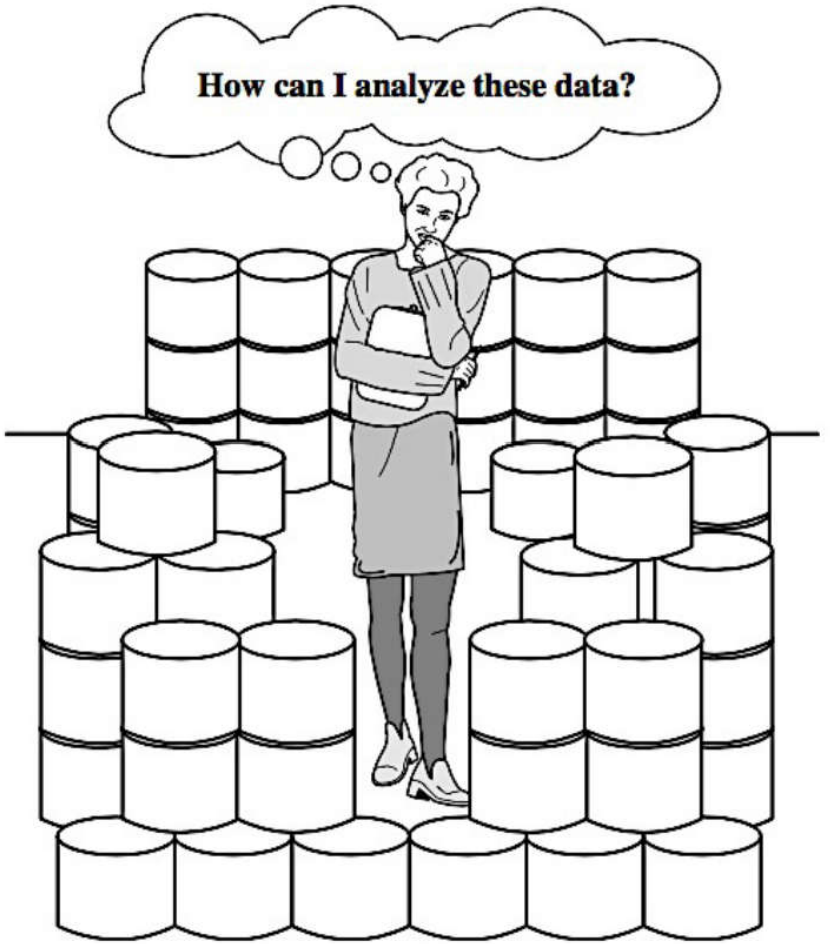
Finding alternative/ green energy sources

Reducing hunger and poverty by increasing agriculture production

We Are Drowning In Data,

But Starving For Knowledge!

ما در انبوهی از داده ها واطلاعات غرق شدیم در حالی که تشنه ی دانش هستیم
داده کاوی :ما یه سری داده داریم که میخایم ازش ارزش افزوده استنباط کنیم
ما قرار نیست دیتا رو پیدا کنیم ما میخایم دانش را پیدا کنیم (دانش های ارزشمند) استخراج میشه



The world is data rich but information poor.

What is Data Mining?

- Many Definitions
  - Non-trivial extraction of implicit, previously unknown and potentially useful information from data
  - Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns.

(Knowledge or Data) Mining!!!

استخراج انالیز و کشف اطلاعات از داده ها به وسیله ی الگوریتم های اتوماتیک و شبه اتوماتیک برای استخراج الگو از داده ها
داده کاوی یک اتفاق نیست، یک فرایند است در پردازش داده ها ما با فرایندی از پردازش روبرو هستیم
مثل فرایند یا چرخه ی تولید نرم افزار

Data mining—searching for knowledge (interesting patterns) in data.

What is Data Mining?

- Data mining,
  - is the process of **uncovering patterns and other valuable information from large data sets**.

- Alternative names

  Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

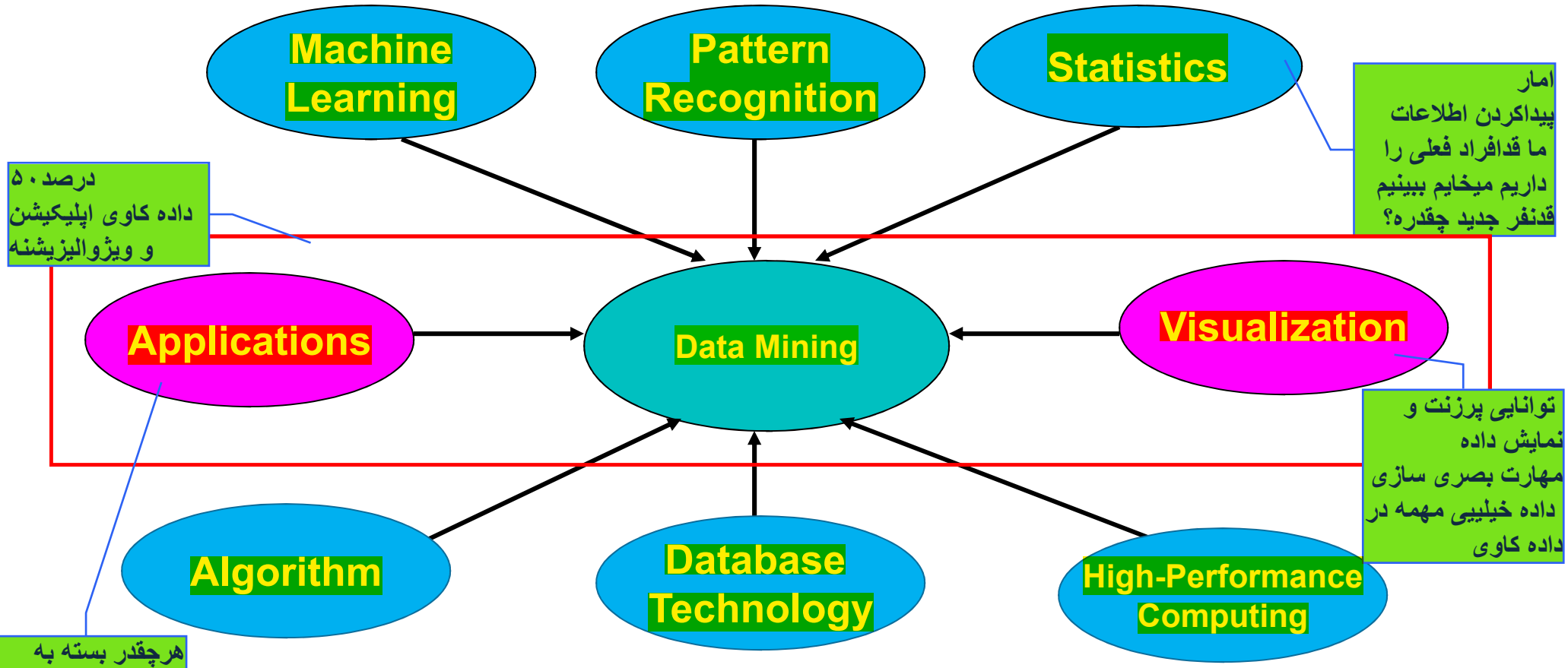
نام های دیگر دیتا ماینینگ

داده کاوی یک ترکیبی از حوزه های مختلف است

Data Mining: Confluence of Multiple Disciplines

رویکردها و جنبه های مهم در داده کاوی

شناسایی الگو



آمار
پیداکردن اطلاعات ما قدافراد فعلی را داریم میخایم ببینیم قدنفر جدید چقدره؟

درصد ۵۰ داده کاوی اپلیکیشن و ویژوالیزیشنه

توانایی پرزنت و نمایش داده
مهارت بصری سازی داده خیلیی مهمه در داده کاوی

هرچقدر بسته به کاربرد دانش مسئله را داشته باشیم بهتر میتونیم حلش کنیم

چه الگویی توی داده ها وجود داره که ما براساس اون الگو یا مدل بیایم راجع به داده هامون قضاوت کنیم درباره ی نفر جدید
الگو میتونه توزیع احتمالاتی باشه

توزیع پردازش، عملیات ضرب ماتریسی

Human role in data mining?

نقش انسان در پردازش داده ها

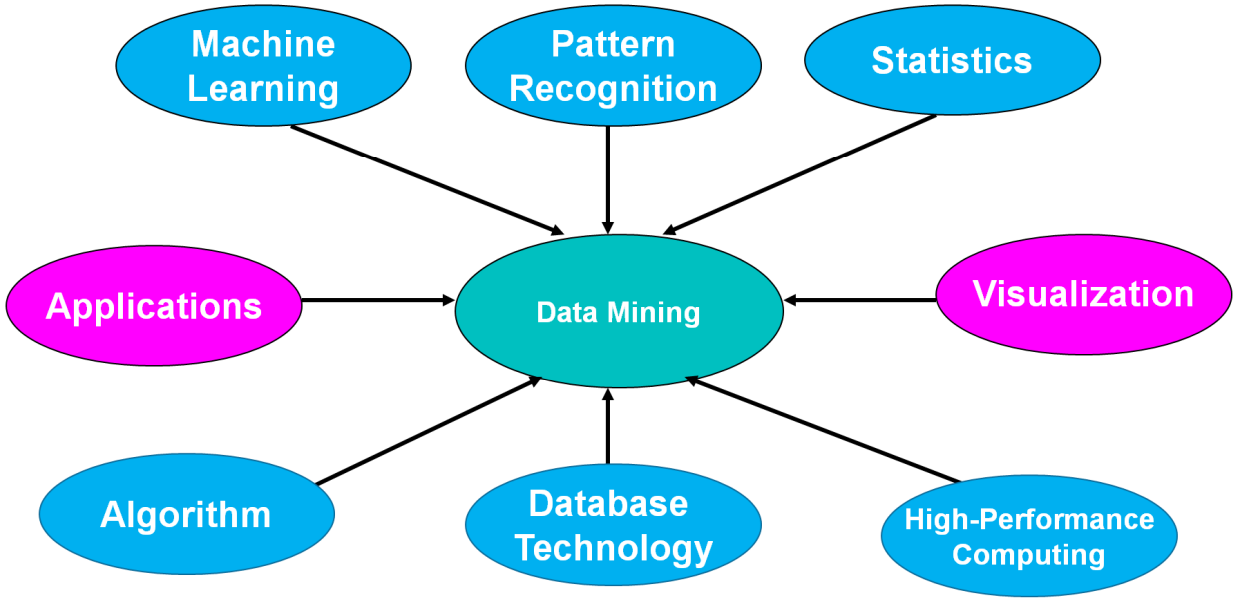
- Berry and Linoff, in their 1997 book gave the following definition for data mining:
  - "Data mining is the process of exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns and rules".

- Three years later, in their Mastering Data Mining book, they mentioned that,
  - "If there is anything we regret, it is the phrase 'by automatic or semiautomatic means' . . . because we feel there has come to be too much focus on the automatic techniques and not enough on the exploration and analysis. This has misled many people into believing that data mining is a product that can be bought rather than a discipline that must be mastered."

Human need to be actively involve in every phase of data mining.

دیتا ماینینگ یک رویه و رویکرد است که ارام ارام حاصل میشه و هزینه بر است و برای این مهارت ها نیاز به انسان داریم

Human role in data mining?



Human need to be actively involve in every phase of data mining.

Question?

Your Name, ID, Major

Q1: What do you think Data Mining is?

Q2: What project have you done so far that you think is most relevant to Data Mining?

Not necessarily research project; can be your course project or any hackathon event you participated in.

Q3: What do you expect to learn from this course?

Multi-Dimensional View of Data Mining

- Data to be mined
  - Database data (extended-relational, object-oriented, heterogeneous), data warehouse, transactional data, stream, spatiotemporal, time-series, sequence, text and web, multi-media, graphs & social and information networks
- Knowledge to be mined (or: Data mining functions)
  - Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, …
  - Descriptive vs. predictive data mining
  - Multiple/integrated functions and mining at multiple levels
- Techniques utilized
  - Data warehousing (OLAP), machine learning, statistics, pattern recognition, visualization, high-performance computing, etc.
- Applications adapted
  - Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.

Data Mining Tasks

- Prediction Methods

متدها برای پیشبینی

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

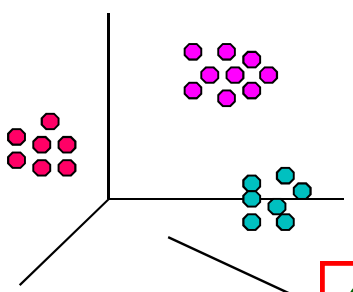
- Description Methods

متدها برای توصیف
یه حجم زیادی از دیتا داریم میخام ببینیم توش چه خبره؟

  - Find human-interpretable patterns that describe the data.

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

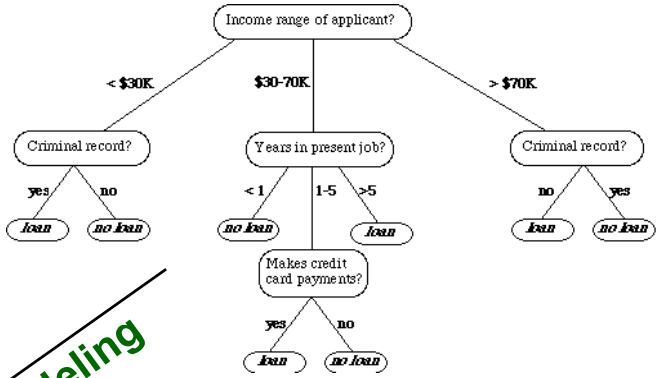
Data Mining Tasks ...



Clustering

Data

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes



Predictive Modeling

Association Rules

Anomaly Detection

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 Analysis: Applications

تحلیل سبد بازار
- قوانین برای ارتقای فروش، مدیریت قفسه و مدیریت موجودی استفاده می شود

- Market-basket analysis
  - Rules are used for sales promotion, shelf management, and inventory management

حوزه ی مخابرات

تشخیص دزدگیر مخابراتی
- از قوانین برای یافتن ترکیبی از آلارم هایی که اغلب در یک دوره زمانی با هم رخ می دهند استفاده می شود

- Telecommunication alarm diagnosis
  - Rules are used to find combination of alarms that occur together frequently in the same time period

حوزه ی پزشکی

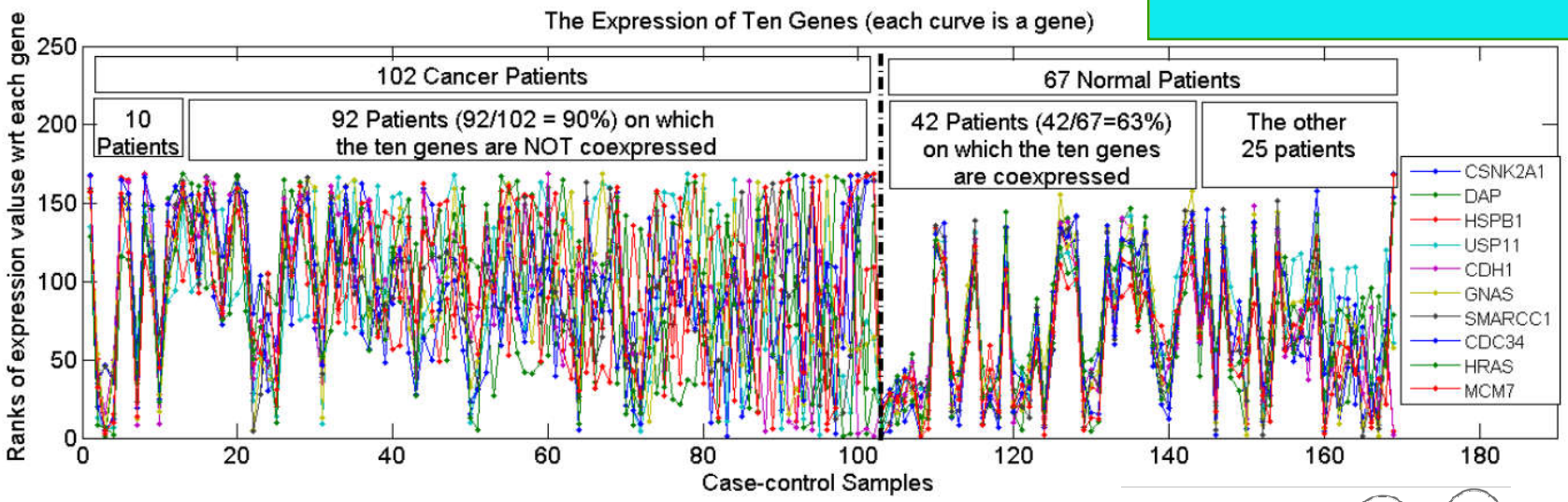
انفورماتیک پزشکی
- قوانین برای یافتن ترکیبی از علائم بیمار و نتایج آزمایش مرتبط با بیماری های خاص استفاده می شود

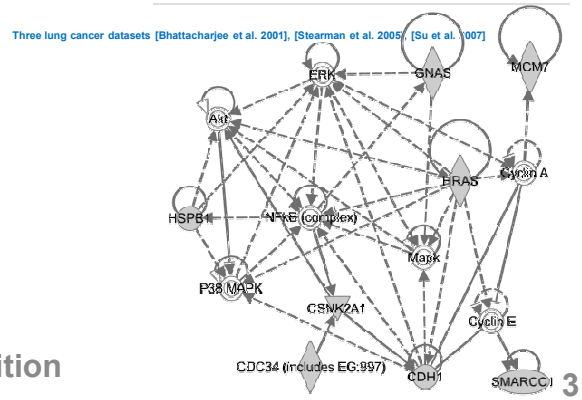
- Medical Informatics
  - Rules are used to find combination of patient symptoms and test results associated with certain diseases

Association Analysis: Applications

- An Example Subspace Differential Coexpression Pattern from lung cancer dataset

نمونه ای از الگوی هم بیان دیفرانسیل
زیرفضایی از مجموعه داده سرطان ریه





Enriched with the TNF/NFB signaling pathway which is well-known to be related to lung cancer

P-value: $1.4*10^{-5}$ (6/10 overlap with the pathway)

[Fang et al PSB 2010]

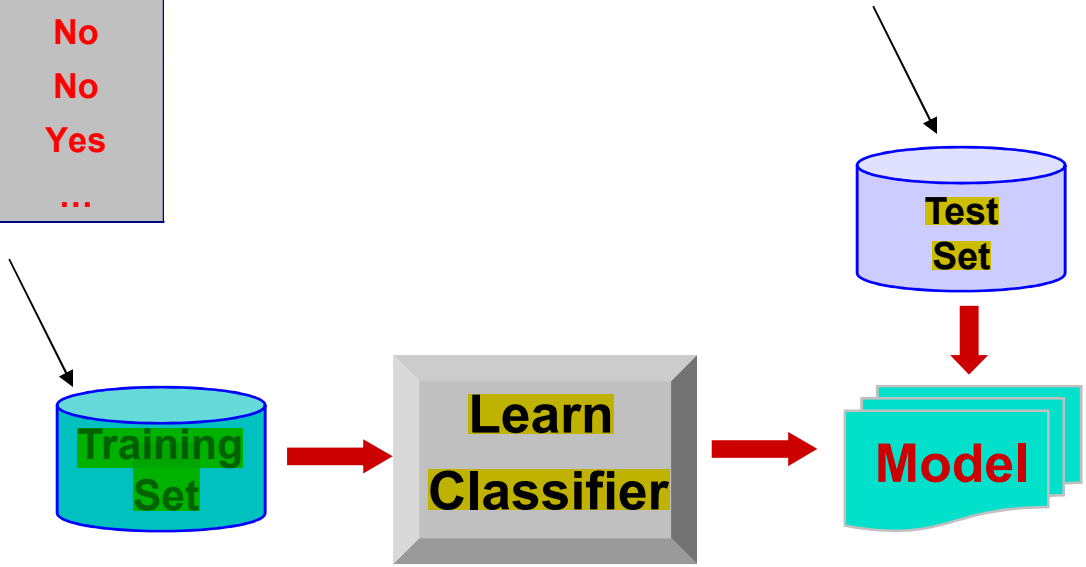
Classification Example

predicting credit worthiness

categorical categorical quantitative class

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Undergrad	7	?
2	No	Graduate	3	?
3	Yes	High School	2	?
…	…	…	…	…



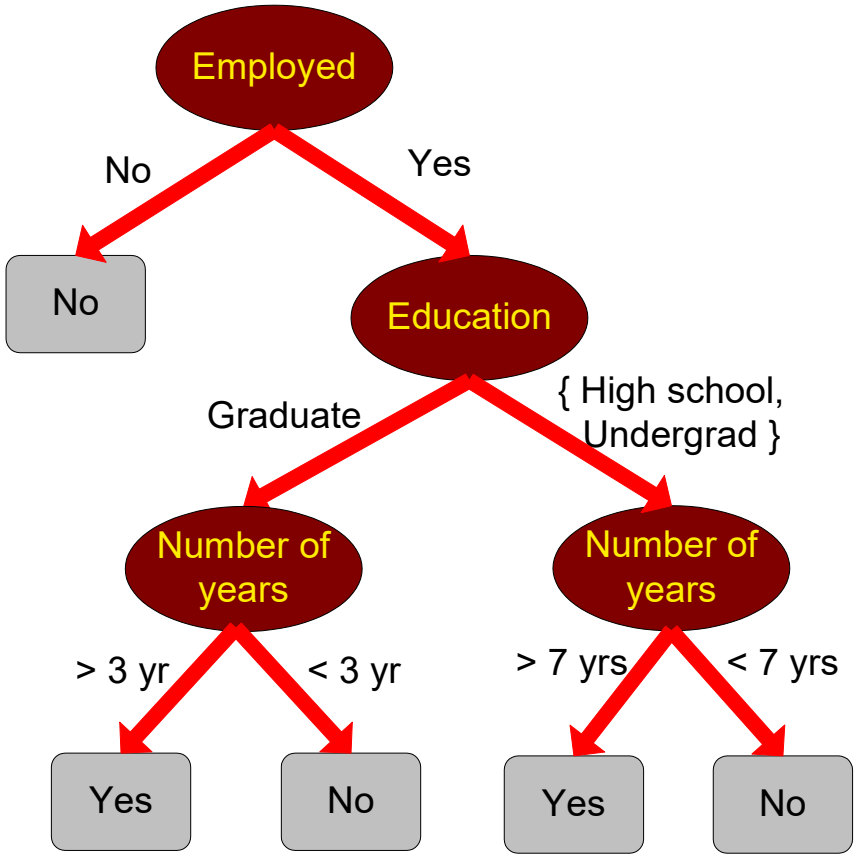
Predictive Modeling: Classification

- Find a model for class attribute as a function of the values of other attributes

Class

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

Model for predicting credit worthiness



Classification: Application 1

Fraud Detection

- **Goal:** Predict fraudulent cases in credit card transactions.

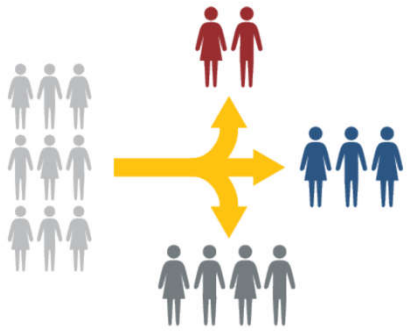
- **Approach:**
  - Use credit card transactions and the information on its account-holder as attributes.
    - When does a customer buy, what does he buy, how often he pays on time, etc
  - Label past transactions as fraud or fair transactions. This forms the class attribute.
  - Learn a model for the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

Classification: Application 2

Churn prediction for telephone customers

- **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.

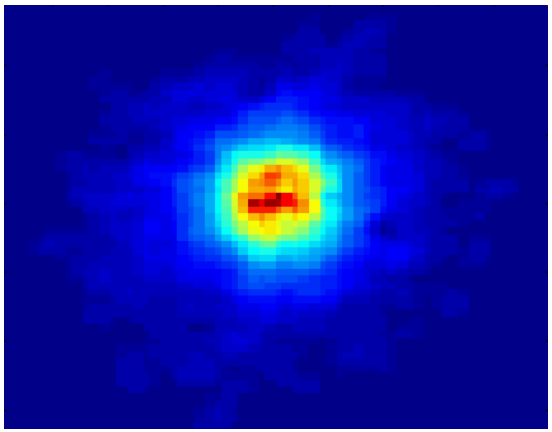


From [Berry & Linoff] Data Mining Techniques, 1997

Classifying Galaxies

Courtesy: http://aps.umn.edu

Early



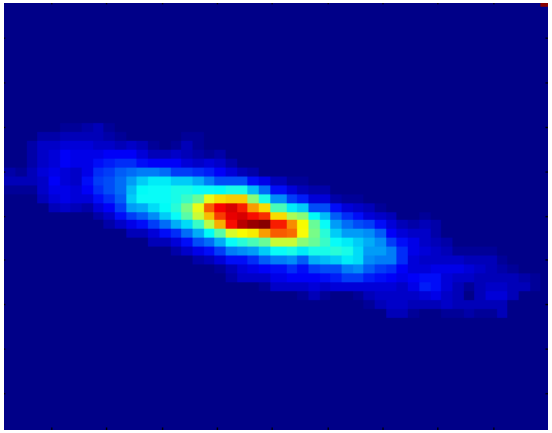
Class:

- **Stages of Formation**

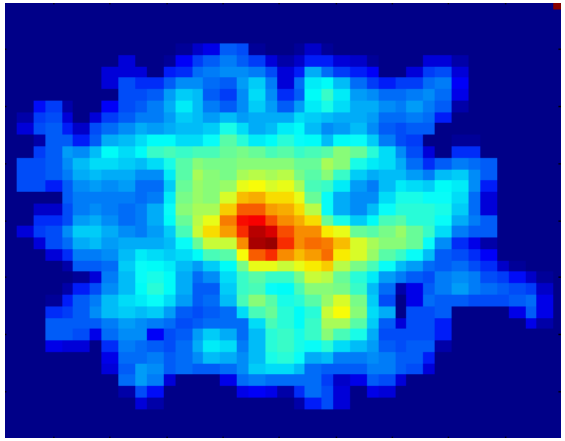
Attributes:

- **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**

Classification: Application 3

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.

فهرست نویسی بررسی آسمان
– هدف: پیشبینی کلاس (ستاره یا کهکشان) اجرام آسمان، به ویژه آنهایی که از نظر بصری کمنور هستند، بر اساس تصاویر بررسی تلسکوپی (از رصدخانه پالومار)

- **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!

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

Examples of Classification Task

- Classifying land covers (water bodies, urban areas, forests, etc.) using satellite data
- Categorizing news stories as finance, weather, entertainment, sports, etc
- Identifying intruders in the cyberspace
- Predicting tumor cells as benign or malignant
- Classifying secondary structures of protein as alpha-helix, beta-sheet, or random coil

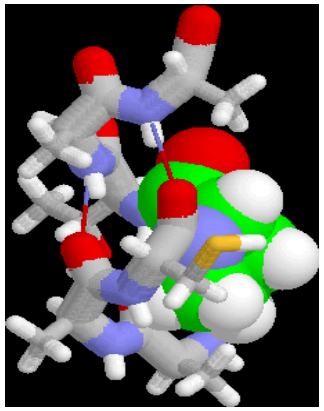
طبقه بندی پوشش های زمین (آب، مناطق شهری، جنگل ها و غیره) با استفاده از داده های ماهواره ای

طبقه بندی اخبار به عنوان امور مالی، آب و هوا، سرگرمی، ورزش و غیره

شناسایی مزاحمان در فضای مجازی

پیش بینی سلول های تومور به عنوان خوش خیم یا بدخیم

طبقه بندی ساختارهای ثانویه پروتئین به عنوان آلفا مارپیچ، بتاسیت یا سیم پیچ تصادفی



Regression

مسائل پیش بینی و کار با متغیرهای تصادفی پیوسته

قیمت یک خانه
شاخص الودگی هوا
گزارش زمان

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

پیش بینی میزان فروش محصول جدید بر اساس هزینه های تبلیغاتی.
- پیش بینی سرعت باد به عنوان تابعی از دما، رطوبت، فشار هوا و غیره.
- پیش بینی سری زمانی شاخص های بورس

پدیده هایی که بعد زمانی دارند

سری زمانی قیمت یک کالا
اگه بخایم پیشبینی داشته باشیم درباره ی اینده ی اون کالا مساله از جنس رگرشن است

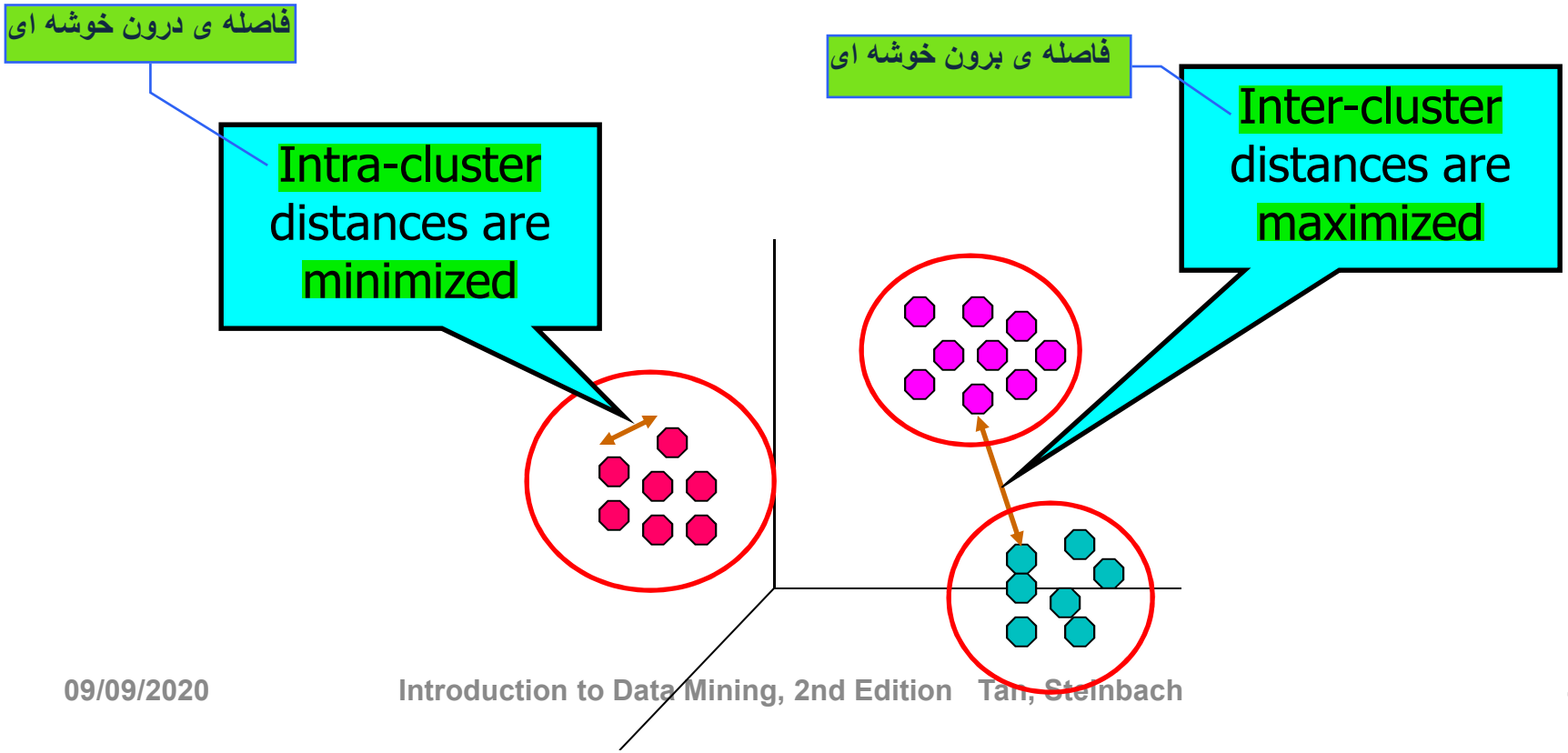
دسته بندی یک سری نمونه

مسائل خوشه بندی

Clustering Task

یک شاخصی از فاصله میسازه و تلاش میکنه فاصله ی اشیایی که توی یک خوشه هستند از هم کم باشه و فاصله ی دو شی در دوخوشه ی متفاوت زیاد باشه

Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups.



Clustering: Application 1

تقسیم بندی بازار:
- هدف: تقسیم بازار به زیرمجموعه های متمایز از مشتریان که در آن هر زیرمجموعه ای ممکن است به عنوان هدف بازار انتخاب شود تا با یک آمیخته بازاریابی مجزا به آن دست یابید.

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.

رویکرد: ویژگی های مختلف مشتریان را بر اساس اطلاعات جغرافیایی و سبک زندگی آنها جمع آوری کنید.
خوشه هایی از مشتریان مشابه را بیابید.
کیفیت خوشه بندی را با مشاهده الگوهای خرید مشتریان در همان خوشه در مقابل مشتریان خوشه های مختلف اندازه گیری کنید.

Clustering: Application 2

خوشه بندی اسناد:
- هدف: یافتن گروه هایی از اسناد مشابه یکدیگر بر اساس اصطلاحات مهم موجود در آنها.
– رویکرد: برای شناسایی اصطلاحات رایج در هر سند. یک معیار تشابه را بر اساس فراوانی اصطلاحات مختلف تشکیل دهید. از آن برای خوشه بندی استفاده کنید

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.

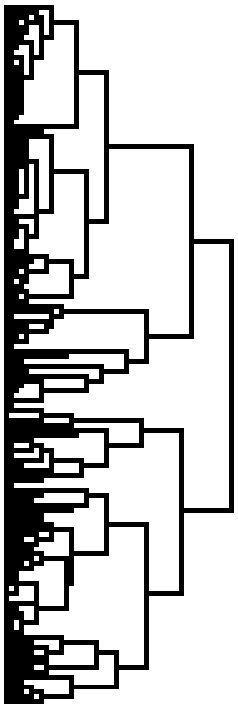
Enron email dataset

Applications of Cluster Analysis

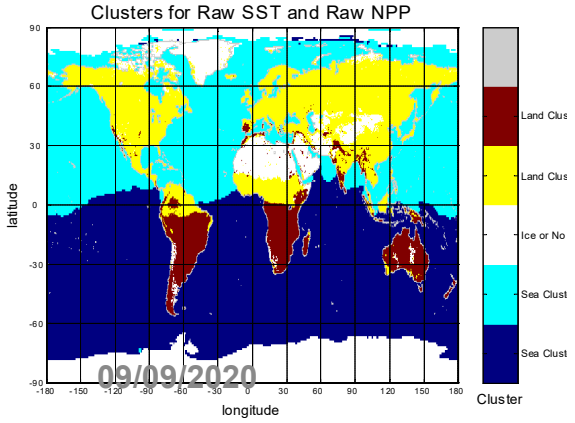
- **Understanding** نمادهای بورس
  - Custom profiling for targeted marketing
  - Group related documents for browsing
  - Group genes and proteins that have similar functionality
  - Group stocks with similar price fluctuations
- **Summarization**
  - Reduce the size of large data sets

خلاصه سازی حجم زیادی از اطلاعات

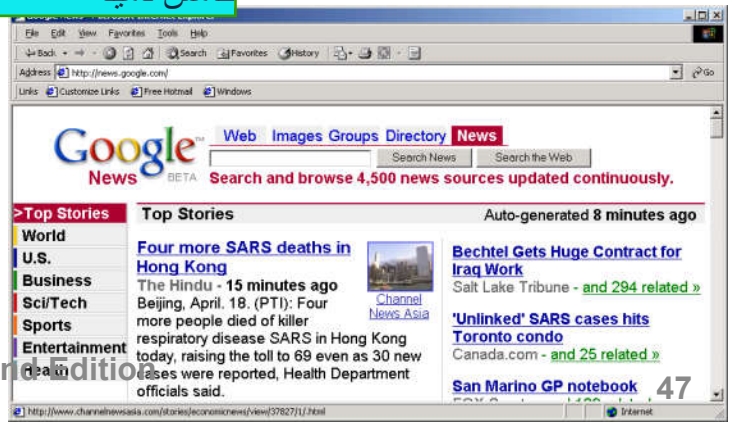
درک کردن
- پروفایل سفارشی برای بازاریابی هدفمند
– گروه بندی اسناد مرتبط برای مرور
– گروه بندی ژن ها و پروتئین هایی که عملکرد مشابهی دارند
– گروه بندی سهام با نوسانات قیمتی مشابه
خلاصه سازی
- اندازه مجموعه داده های بزرگ را کاهش دهید



hael Eisen



Use of K-means to partition Sea Surface Temperature (SST) and Net Primary Production (NPP) into clusters that reflect the Northern and Southern Hemispheres.



Deviation/Anomaly/Change Detection

در یک مساله ی دیتاماینینگ ما یه سری رکوردهای داده ای داریم که هررکوردی یه سری اطلاعات به ما میده و قراره به کمک این رکوردها ما یه مساله ای رو حل کنیم مثل تایید هویت یک ادم برای اختصاص وام ما سن و تحصیلات و شغل و اطلاعات دیگرش را داشتیم و ستون اخرمون این بود ایا طرف معتبر بوده یا نه پس ما یه چارچوب برای دیتاستمون داریم

Detect significant deviations from normal behavior

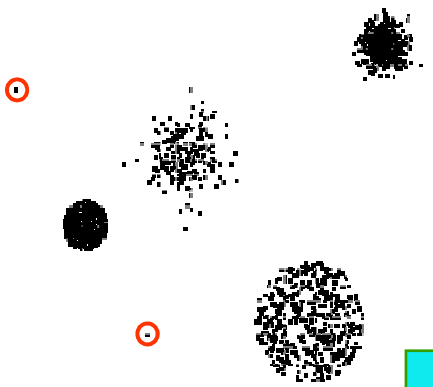
Applications:

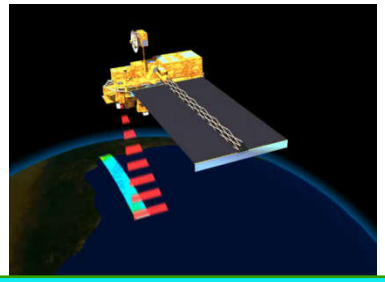
- Credit Card Fraud Detection
- Network Intrusion Detection

تشخیص نفوذ در شبکه ها

- Identify anomalous behavior from sensor networks for monitoring and surveillance.
- Detecting changes in the global forest cover.

نواحی غیرعادی در تصاویر ماهواره ای





انحرافات قابل توجه از رفتار عادی را تشخیص دهید
برنامه های کاربردی:
- تشخیص تقلب در کارت اعتباری
– تشخیص نفوذ شبکه
- شناسایی رفتار غیرعادی از شبکه های حسگر برای نظارت .
- تشخیص تغییرات در پوشش جنگلی جهانی

Major Issues in Data Mining (1)

- Mining Methodology
  - Mining various and new kinds of knowledge(New Question)
  - Mining knowledge in multi-dimensional space (Traffic[#+Speed])
  - Data mining: An interdisciplinary effort (bug mining)
  - Boosting the power of discovery in a networked environment (hybrid)
  - Handling noise, uncertainty, and incompleteness of data
  - Pattern evaluation and pattern- or constraint-guided mining (ADHD types details)
- User Interaction
  - Interactive mining(Search Engine(Similar))
  - Incorporation of background knowledge(Multi Judge)
  - Presentation and visualization of data mining results (Better Presentations)

روش شناسی کاوش کردن
- استخراج انواع مختلف و جدید دانش (سوال جدید)
- دانش ماینینگ در فضای چند بعدی
– داده کاوی: یک تلاش بین رشته ای (باگ کاوی)
- تقویت قدرت کشف در محیط شبکه ای (هیبرید)
- مدیریت نویز، عدم قطعیت و ناقص بودن داده ها
- ارزیابی الگو مبتنی بر محدودیت
تعامل با کاربر
- استخراج تعاملی (موتور جستجو (مشابه))
- ترکیب دانش پیشینه (چند داور)
- ارائه و بصری سازی نتایج داده کاوی

Major Issues in Data Mining (2)

- Efficiency and Scalability

  (running time of a data mining algorithm must be predictable)

  - Efficiency and scalability of data mining algorithms
  - Parallel, distributed, stream, and incremental mining methods(|.|)
- Diversity of data types
  - Handling complex types of data (Time Series)
  - Mining dynamic, networked, and global data repositories(Social Network)
- Data mining and society
  - Social impacts of data mining (The social Dilem
  - Privacy-preserving data mining(Fanavard)
  - Invisible data mining(Search Engine)

کارایی و مقیاس پذیری
(زمان اجرای یک الگوریتم داده کاوی باید قابل پیش بینی باشد)
– کارایی و مقیاس پذیری الگوریتم های داده کاوی
– روش های استخراج موازی، توزیعی، جریانی و افزایشی
تنوع انواع داده ها
- مدیریت انواع پیچیده داده ها (سری های زمانی)
- استخراج مخازن داده پویا، شبکه ای و جهانی (شبکه اجتماعی)
داده کاوی و جامعه
- اثرات اجتماعی داده کاوی (معضل اجتماعی!)
– داده کاوی با حفظ حریم خصوصی (فناورد)
– داده کاوی نامرئی (موتور جستجو)

Where to Find References? DBLP, CiteSeer, Google

Data mining and KDD (SIGKDD)

- Conferences: ACM-SIGKDD, IEEE-ICDM, SIAM-DM, PKDD, PAKDD, etc.
- Journal: Data Mining and Knowledge Discovery, KDD Explorations, ACM TKDD¨

Database systems (SIGMOD)

- Conferences: ACM-SIGMOD, ACM-PODS, VLDB, IEEE-ICDE, EDBT, ICDT, DASFAA
- Journals: IEEE-TKDE, ACM-TODS/TOIS, JIIS, J. ACM, VLDB J., Info. Sys., etc. ¨

AI & Machine Learning

- Conferences: Machine learning (ML), AAAI, IJCAI, COLT (Learning Theory), CVPR, NIPS, etc.
- Journals: Machine Learning, Artificial Intelligence, Knowledge and Information Systems, IEEE-PAMI, etc.

Where to Find References? DBLP, CiteSeer, Google

Web and IR

Conferences: SIGIR, WWW, CIKM, etc.

Journals: WWW: Internet and Web Information Systems

..

Statistics

Conferences: Joint Stat. Meeting, etc.

Journals: Annals of statistics, etc.

..

Visualization

Conference proceedings: CHI, ACM-SIGGraph, etc.

Journals: IEEE Trans. visualization and computer graphics, etc.

Exercise2

- One data mining case example from Kaggle in the following format

Case	Feature #	Train Sample #	Test Sample #

The social Dilemma