University of Windsor

03 Introduction to Data Mining

Kalyani Selvarajah
School of Computer Science
University of Windsor

Advanced Database Topics
COMP 8157 01
Fall 2023

Announcement

- Project Proposal
- Assignment 1

Today's Agenda

Introduction to Data Mining

Defining Data Mining

Data Mining Tasks

https://www.33rdsquare.com/best-data-mining-tools/

Reference: Introduction to Data Mining, 2nd Edition and Mining of Massive Datasets

Introductory Questions

Why data is useful?

What do you mean by Data Mining?

What are the step-by-step process of Data Mining?

What are the usages of predictive models? Explain using some applications?

What are the usages of Descriptive methods? Explain using some applications?

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.

The Explosive Growth of Data: from terabytes to petabytes.

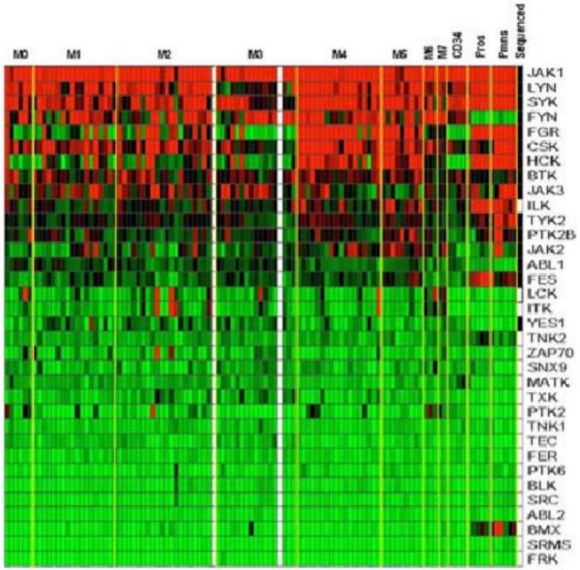
- Major sources of abundant data
  - Business: Web, e-commerce, transactions, stocks, …
  - Science: Remote sensing, bioinformatics, scientific simulation, …
  - Society and everyone: news, digital cameras, YouTube,…
  - Social media: Instagram, Facebook, Twitter,….

Data can help us solve specific problems.

How should these pictures be placed into 3 groups?

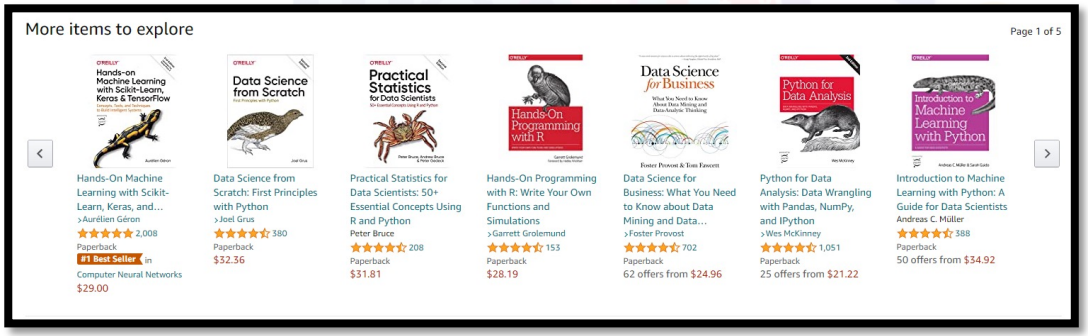
How should these pictures be placed into groups? How many groups should there be?

Data can help us solve specific problems.



Which genes are associated with a disease? How can expression values be used to predict survival?

What items should Amazon display for me?



Data can help us solve specific problems.

Where are the faces in this picture?

Is this spam?

```
hi backpackers,

i saw that close to my hotel there is a pub with
bowling (it's on market between 9th and 10th avenue).
are you up to it?  i think it is about 20 years i
haven't played...  if you like the idea what about
8.30 there?

otherwise any suggestion welcome.  i can survive
another 20 years without bowling.
```


POPULAR SCIENCE
THE FUTURE NOW
THE CONTROL CENTERS
Using Data to Feed the World, Solve Cold Cases, Battle Malware, Predict Our Fate
OFFICER ALGORITHM
Can a Crime Be Prevented Before It Begins?
NEW WAYS OF SEEING
A Gallery of Extraordinary Infographics
PLUS
Juan Enriquez Reprograms Life
James Gleick Unsplits the Bit
AND
Lawrence Weschler Questions the Cloud
SPECIAL ISSUE
DATA IS POWER
HOW INFORMATION IS DRIVING THE FUTURE

Data Mining

What is Data Mining?

To extract the knowledge data needs to be

- Stored (System)
- Managed (Database) and
- Analyzed (Data mining)

Data Mining ≈ Big Data ≈ Predictive Analytics ≈ Data Science

What is Data Mining?

Many Definitions:

Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data.

Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns

Alternative names:

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

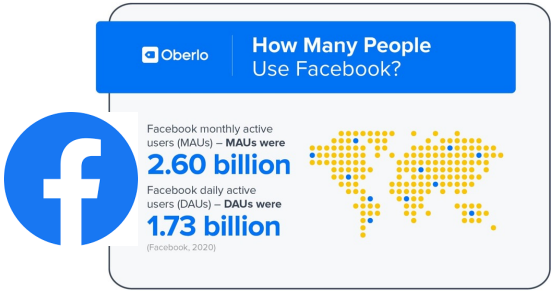
What is Data Mining?

Given lots of data

Discover patterns and models that are:

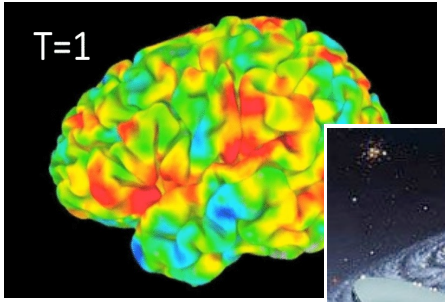
- Valid: hold on new data with some certainty
- Useful: should be possible to act on the item
- Unexpected: non-obvious to the system
- Understandable: humans should be able to interpret the pattern

Why Data Mining? Commercial Viewpoint



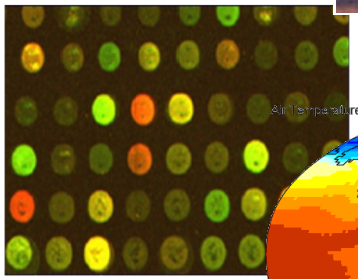
- Lots of data is being collected and warehoused
  - Web data
    - Yahoo has Peta Bytes of web data
    - Facebook has billions of active users
  - purchases at department/grocery stores, e-commerce
    - Amazon handles millions of visits/day
  - Bank/Credit Card transactions
- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
  - Provide better, customized services for an edge (e.g., in Customer Relationship Management) example: Spam Filter

Why Data Mining? Scientific Viewpoint

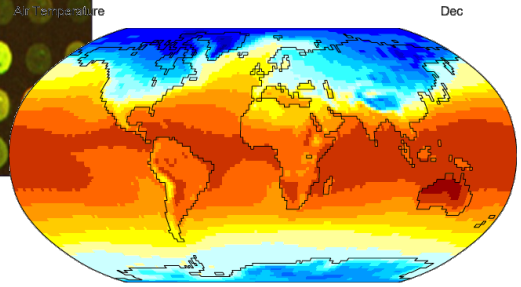


fMRI Data from Brain

Sky Survey Data



Gene Expression Data



Surface Temperature of Earth

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

Why Data Mining? Various Application

RETAIL/MARKETING

- ✓ Identifying buying patterns of customers
- ✓ Finding associations among customer demographic characteristics
- ✓ Predicting response to mailing campaigns
- ✓ Market basket analysis

BANKING

- ✓ Detecting patterns of fraudulent credit card use
- ✓ Identifying loyal customers
- ✓ Predicting customers likely to change their credit card affiliation
- ✓ Determining credit card spending by customer groups

INSURANCE

- ✓ Claims analysis
- ✓ Predicting which customers will buy new policies

MEDICINE

- ✓ Characterizing patient behavior to predict surgery visits
- ✓ Identifying successful medical therapies for different illnesses

Is everything “data mining”?

- Looking up individual records using database management system.
  Look up phone number in phone directory

- Finding particular web pages via query to an internet search engine.
  Query a Web search engine for information about “Amazon”

The goal of Data Mining is **the extraction of patterns and knowledge** from large amounts of data, not the extraction (mining) of data itself

The process of knowledge discovery in databases (KDD)

involves transforming raw data into an understandable format.

display the knowledge and transform it into a form understandable to the end user

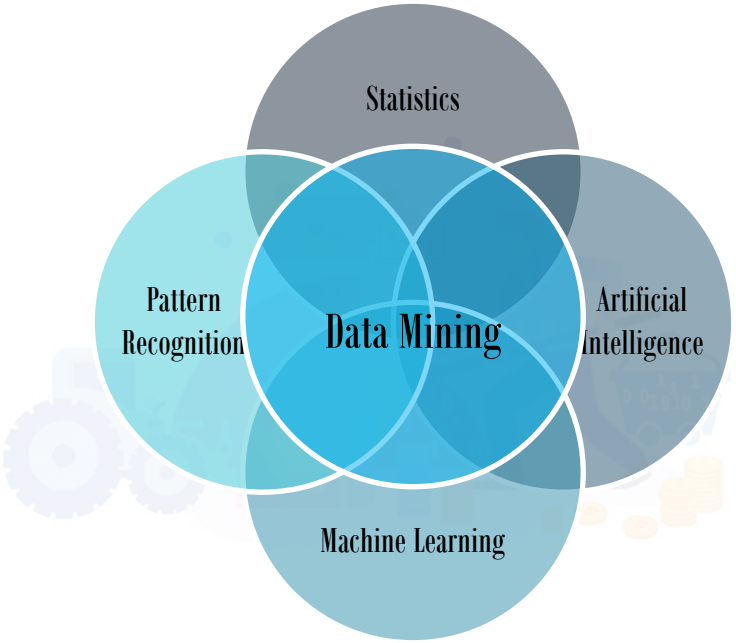
Data integration
Normalization
Feature selection
Dimension reduction

Pattern discovery
Association & correlation
Classification
Clustering
Outlier analysis

Pattern evaluation
Pattern selection
Pattern interpretation
Pattern visualization

The Origins of Data Mining

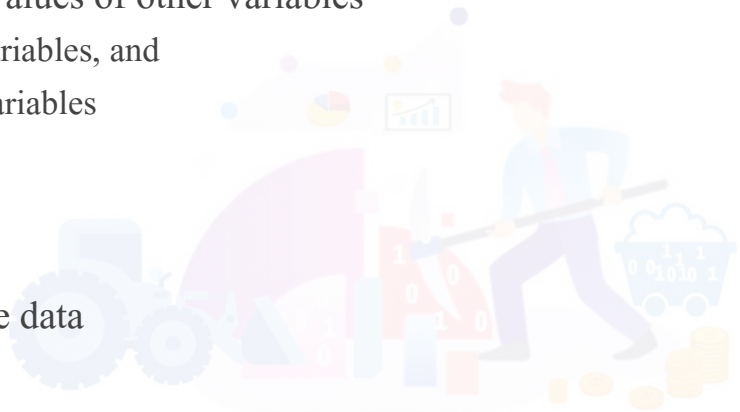
- Draws ideas from machine learning/AI, pattern recognition, statistics, visualization and database systems
- Traditional techniques may be unsuitable due to data that is
  - ✓ Large-scale
  - ✓ High dimensional
  - ✓ Heterogeneous
  - ✓ Complex
  - ✓ Distributed
- A key component of the emerging field of data science and data-driven discovery



Data Mining Tasks

- Predictive methods
  - Use some variables to predict unknown or future values of other variables
    - classification, which is used for discrete target variables, and
    - regression, which is used for continuous target variables
    - Example: Recommender systems

- Descriptive methods
  - Find human-interpretable patterns that describe the data
    - Example: Clustering



Predictive Methods: Classification

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

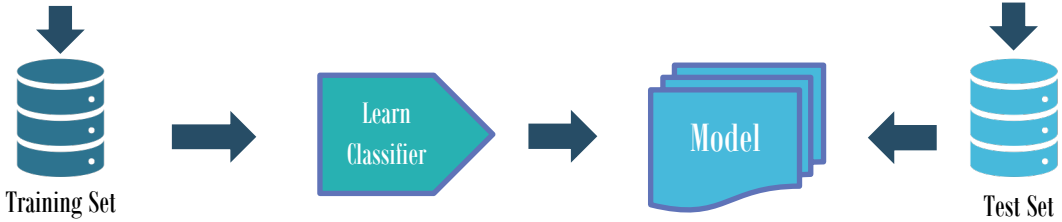
Classification Example

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

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

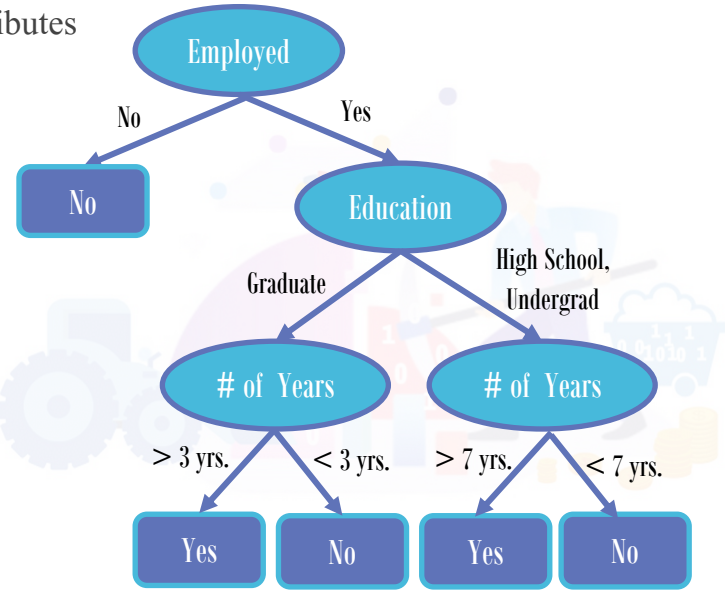
Tid	Employed	Level of Education	Years at present address	Credit Worthy
1	Yes	Undergrad	7	?
2	No	Graduate	3	?
3	Yes	High School	2	?
	...	..	..	..



Classification Example

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

Tid	Employed	Level of Education	Years at present address	Credit Worthy
	Categorical	Categorical	Quantitative	Class
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

Spam Filter

Input: an email

Output: spam/ham

Setup:

Get a large collection of example emails, each labeled “spam” or “ham”

Note: someone has to hand label all this data!

Want to learn to predict labels of new, future emails

Features: The attributes used to make the ham / spam decision

Words: FREE!

Text Patterns: $dd, CAPS

Non-text: SenderInContacts, WidelyBroadcast

…

Dear Sir.

First, I must solicit your confidence in this transaction, this is by virture of its nature as being utterly confidencial and top secret. ...

TO BE REMOVED FROM FUTURE MAILINGS, SIMPLY REPLY TO THIS MESSAGE AND PUT "REMOVE" IN THE SUBJECT.

99 MILLION EMAIL ADDRESSES
FOR ONLY $99

Ok, Iknow this is blatantly OT but I'm beginning to go insane. Had an old Dell Dimension XPS sitting in the corner and decided to put it to use, I know it was working pre being stuck in the corner, but when I plugged it in, hit the power nothing happened.

Classification : Application 2

Digit Recognition

Input: images / pixel grids

Output: a digit 0-9

Setup:

- Get a large collection of example images, each labeled with a digit
- Note: someone has to hand label all this data!
- Want to learn to predict labels of new, future digit images

Features: The attributes used to make the digit decision

- Pixels: (6,8)=ON
- Shape Patterns: NumComponents, AspectRatio, NumLoops
- …
- Features are increasingly induced rather than crafted


0


1


2

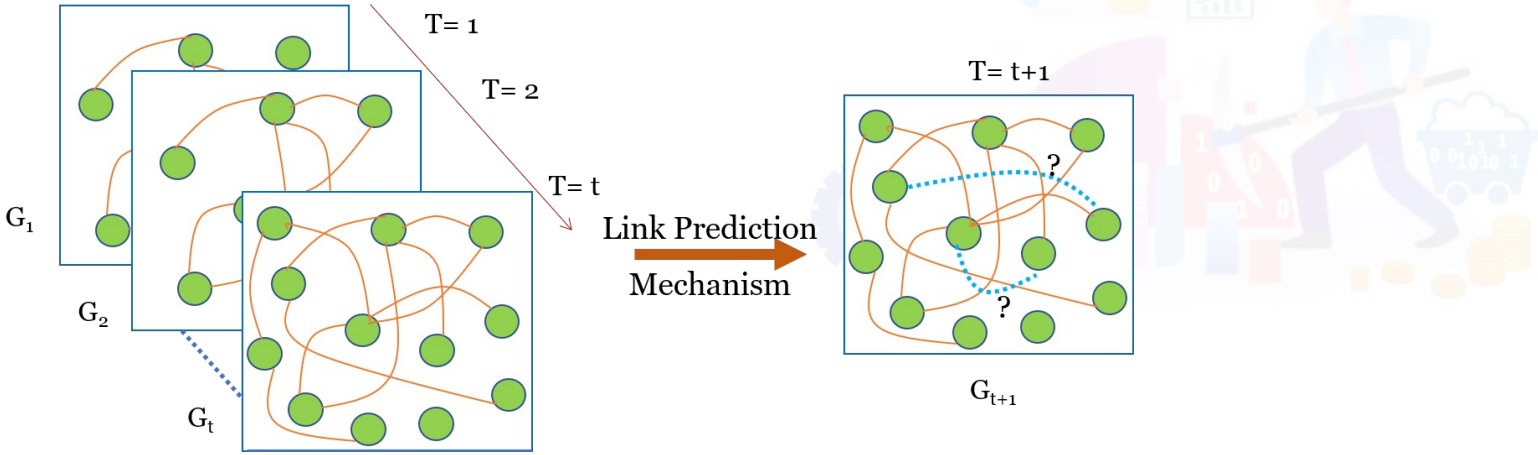

1


??

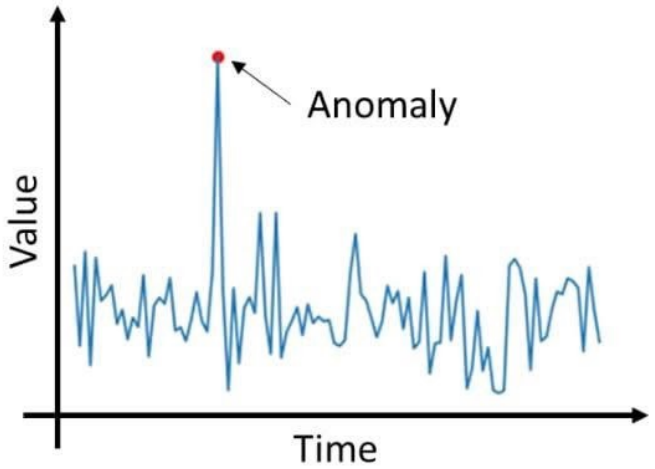
Predictive Modeling: My research work

Real-world networks always evolve over time with the occurrence and vanishing of nodes and links.

Predicting the occurrence of future links in a given dynamic network is a challenging process.



Predictive Methods: Deviation/Anomaly/Change Detection



Detect significant deviations from normal behavior:

deviation detection is often a source of true discovery, because it identifies outliers, which express deviation from some previously known expectation and norm.

Applications:

Credit Card Fraud Detection

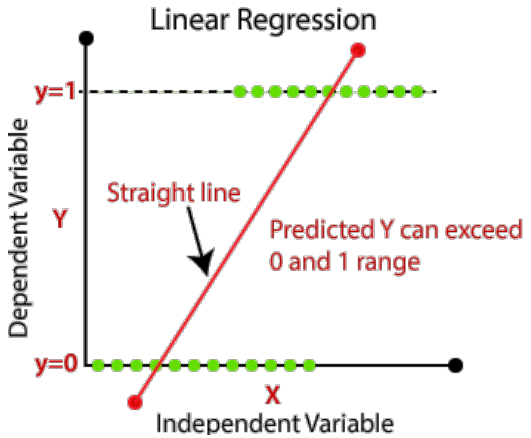
Network Intrusion Detection

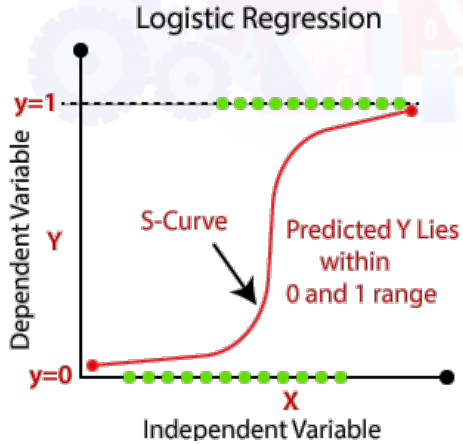
Identify anomalous behavior from sensor networks for monitoring and surveillance.

Detecting changes in the global forest cover.

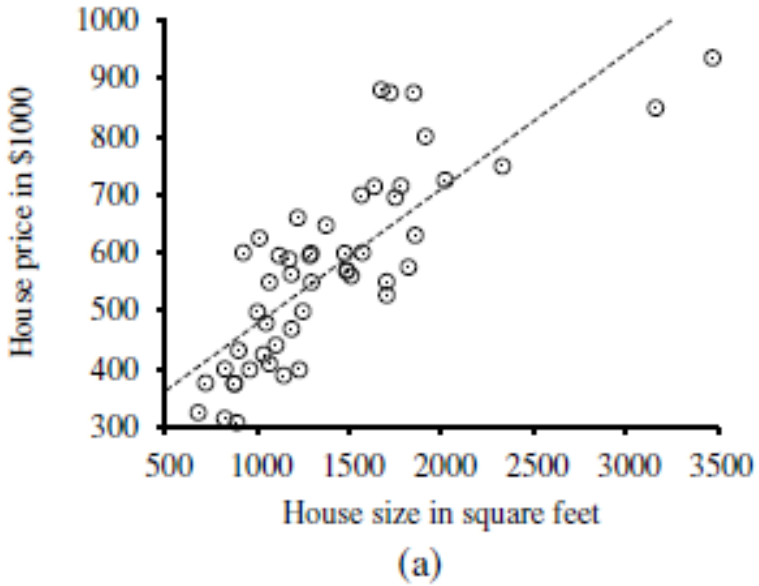
Predictive Methods: Regression

- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Extensively 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.





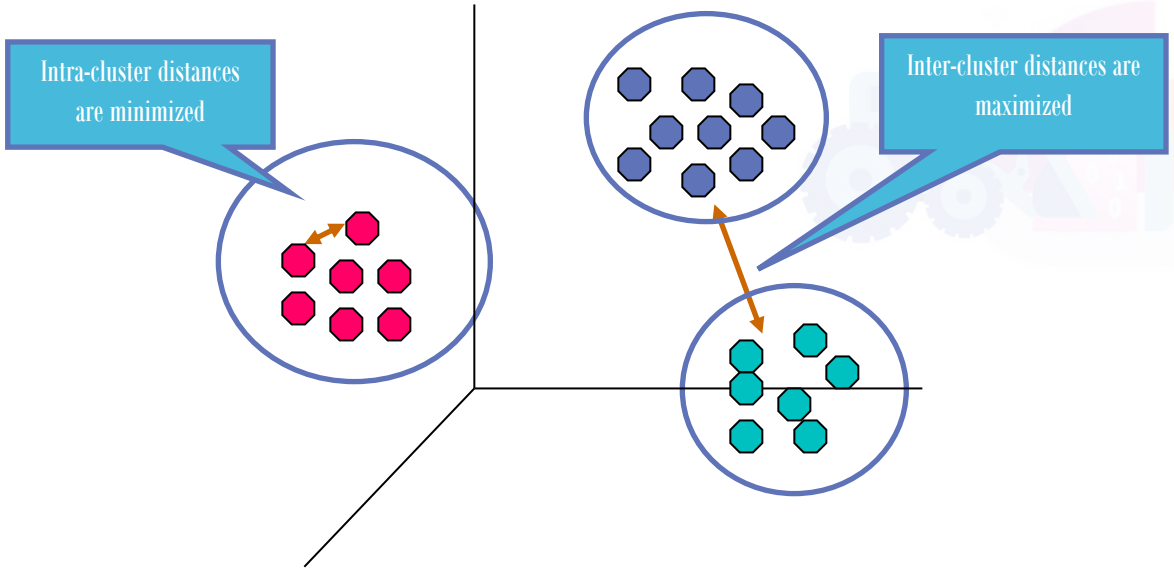
Regression: Application 1



- Data points of price versus floor space of houses for sale in Berkeley, CA, in July 2009.
- along with the linear function hypothesis that minimizes squared error loss: $y = 0.232x + 246$

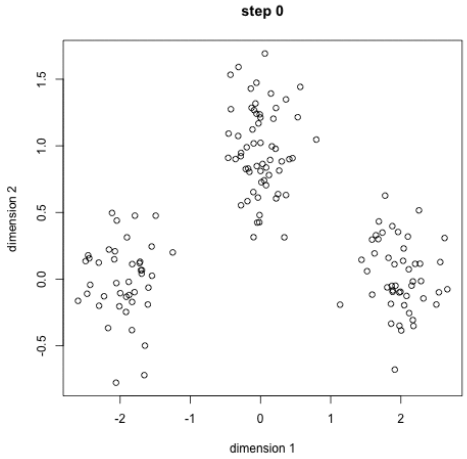
Descriptive methods: Clustering

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.

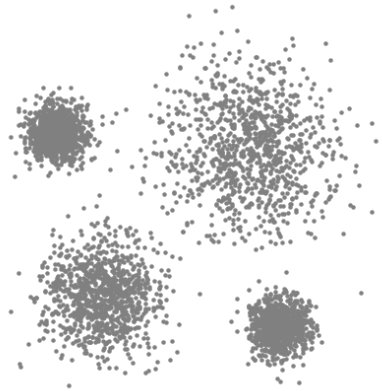


Source: https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68

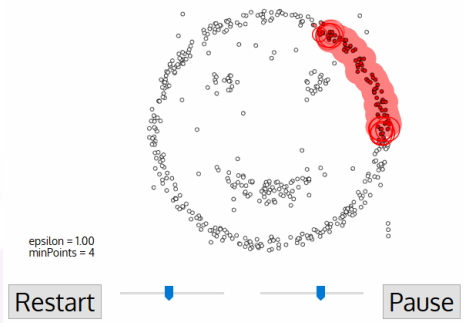
Clustering Algorithms: Example



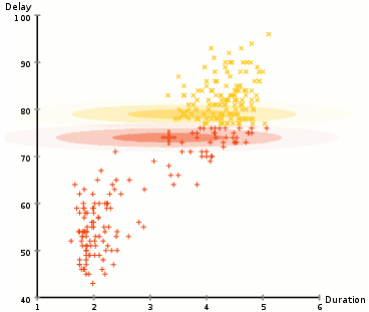
K-Means Clustering



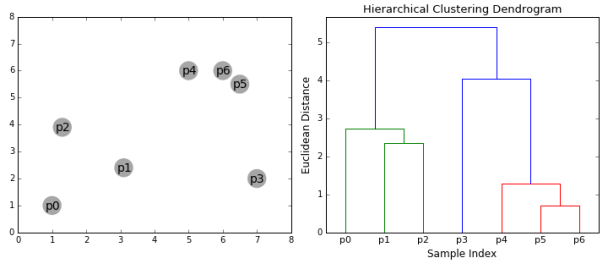
Mean-Shift Clustering



Density-Based Spatial Clustering of Applications with Noise (DBSCAN)



Expectation–Maximization (EM) Clustering using Gaussian Mixture Models (GMM)



Agglomerative Hierarchical Clustering

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.

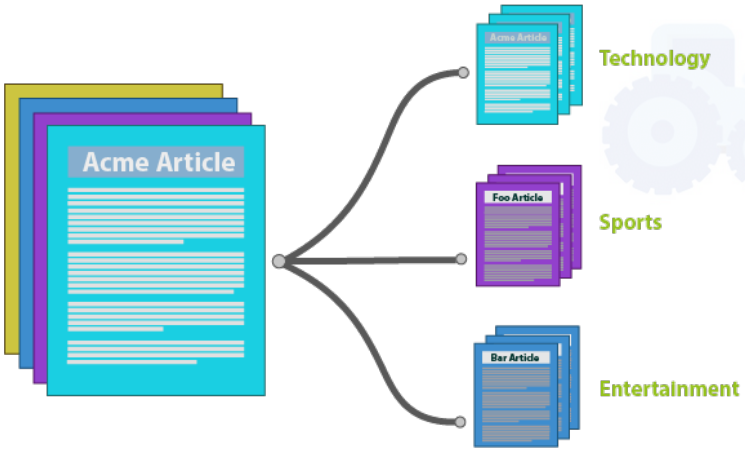
Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Clustering: Application 2

Document Clustering:

Goal: To find groups of documents that are similar to each other 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.



Descriptive methods: Association Rule Discovery

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: Applications

Market-basket analysis:

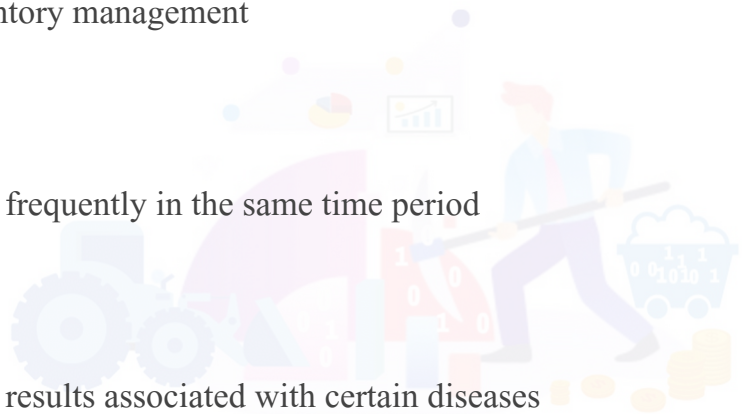
Rules are used for sales promotion, shelf management, and inventory management

Telecommunication alarm diagnosis:

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

Medical Informatics:

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



Summary

We discussed how data is spread out there.

We defined the meaning of data mining.

We then discussed the process of knowledge discovery in databases (KDD)

We discussed two different methods of datamining and the application of those methods.