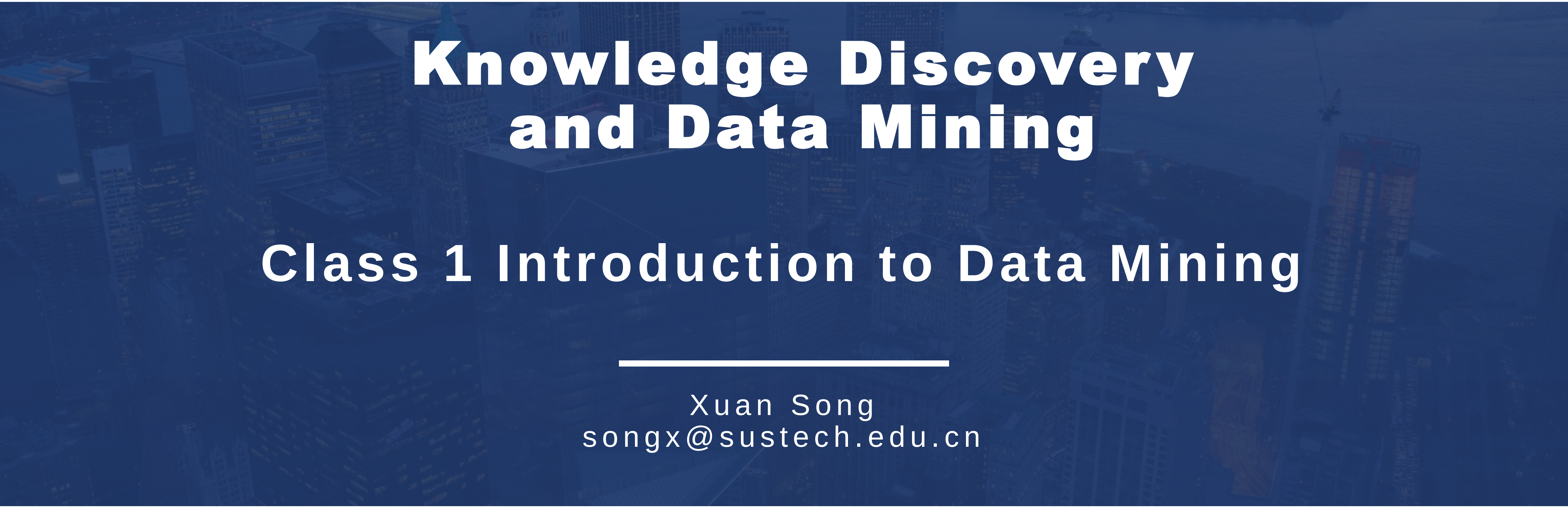


Knowledge Discovery and Data Mining

Class 1 Introduction to Data Mining

Xuan Song
songx@sustech.edu.cn

Contact Information

Instructor: Xuan Song

E-mail: song.x@sustech.edu.cn

Classroom: 荔园6栋403 **Time:** Friday, 10:20-12:10

Laboratory room：荔园6栋402 **Time:** Friday, 14:00-15:50

Office: Innovation Park 10, Room 505

Office Phone: (0755) 8801-5225

Course TA: Yicheng Zhao, Guixu Lin, Yizhuo Wang, Hanchen Liu

Course Wechat Group: Let us establish it now

Grading Policy

	课堂表现 Class Performance		10%		随机随堂测验 Radom in-class quizzes
	平时作业 (Assignments)		40%		平时上机实验 Lab Assignments
	期中考试（前沿文献讨论报告） Mid-Term Test		20%		前沿文献讨论和报告 Paper Reading, Discussion and Presentation
	期末报告（课程项目） Final Presentation		30%		课程项目最终报告、答辩和程序验收 Final Project Presentation

Class Schedule （可能根据学习进度有所调整）

第一周：课程介绍
第二周：数据收集与基础预处理
第三周：监督学习
第四周：无监督学习
第五周：数据降维与模型评判标准
第六周：深度学习与神经网络
第七周：前沿文献阅读、讨论和报告
第八周：前沿文献阅读、讨论和报告

第九周：社交网络
第十周：推荐系统
第十一周：网页数据挖掘
第十二周：城市计算方向的数据挖掘
第十三周：数据挖掘前沿讲座1
第十四周：数据挖掘前沿讲座2
第十五周：课程项目期末答辩
第十六周：课程项目期末答辩

Lab 1: Introduction to Python, Anaconda Jupyter Environment
Lab 2: Introduction to Python Data Crawler
Lab 3: Basic Data Cleaning Skills 1
Lab 4: Basic Data Cleaning Skills 2
Lab 5: Text-based Classification 1
Lab 6: Text-based Classification 2
Lab 7: Text-based Classification 3
Lab 8: Text-based Clustering 1

Lab 9: Text-based Clustering 2
Lab 10: Graphing in Python
Lab 11: Interactive Design
Lab 12: XML Retrieval
Lab 13: Data Visualization on Map
Lab 14: Build a simple Social Network
Lab 15: Final project presentation
Lab 16: Final project presentation

Plagiarism

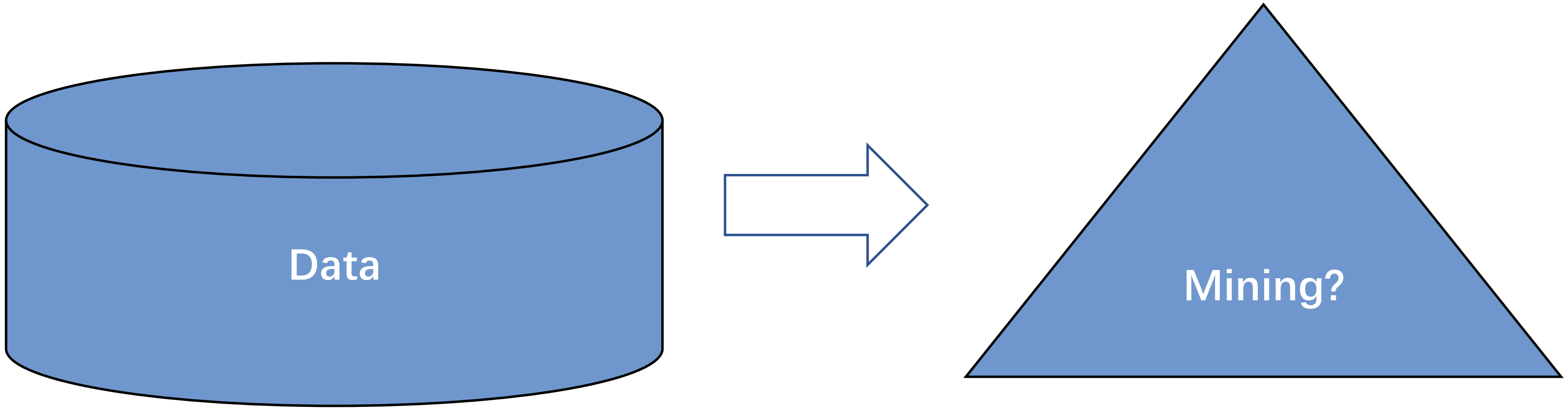
From Spring 2018, the plagiarism policy applied by the Computer Science and Engineering department is the following:

* If an assignment is found to be plagiarized, the first time the score of the assignment will be 0.
•The second time the score of the course will be 0.

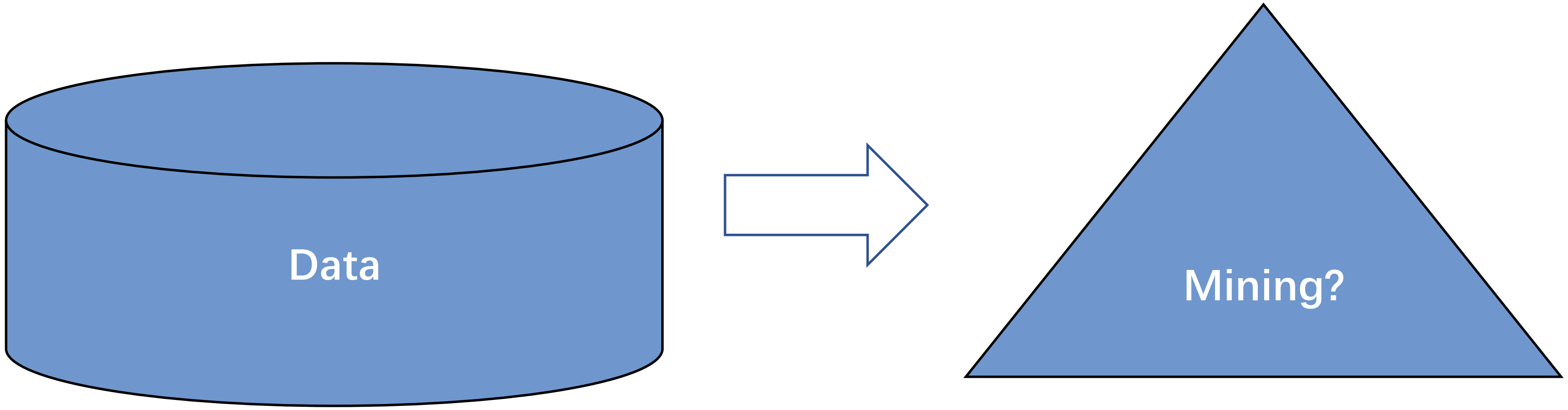
As **it may be difficult** when two assignments are identical or nearly identical who actually wrote it, the policy will apply to BOTH students, unless one confesses having copied without the knowledge of the other.

What is Data Mining?

Why You are Here?



Why You are Here?

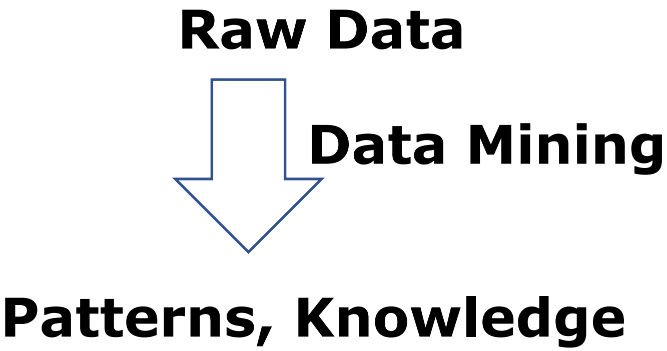


Data

Mining?

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.



Raw Data

Data Mining

Patterns, Knowledge

Data Mining and KDD(Knowledge Discovery in Databases)

- Data Mining and KDD (Knowledge Discovery in Databases)
  - Data mining is an integral part of KDD, which is the overall process of converting raw data into useful information [1].

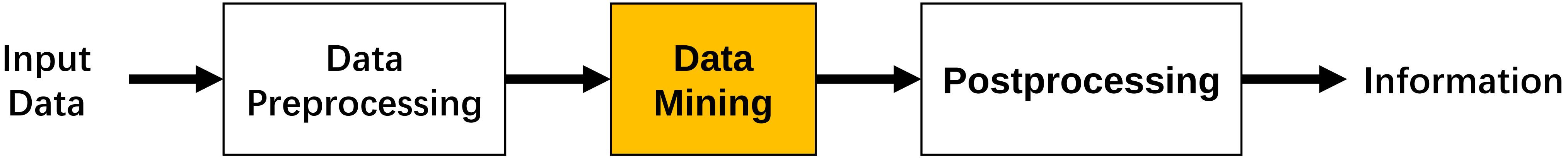


Figure1 The process of knowledge discovery in databases (KDD).

[1] Tan P N, Steinbach M, Kumar V. Introduction to data mining[M]. Pearson Education India, 2016.

Data Mining and KDD(Knowledge Discovery in Databases)

- Data Mining and KDD (Knowledge Discovery in Databases)
  - Data mining is an integral part of KDD, which is the overall process of converting raw data into useful information [1].

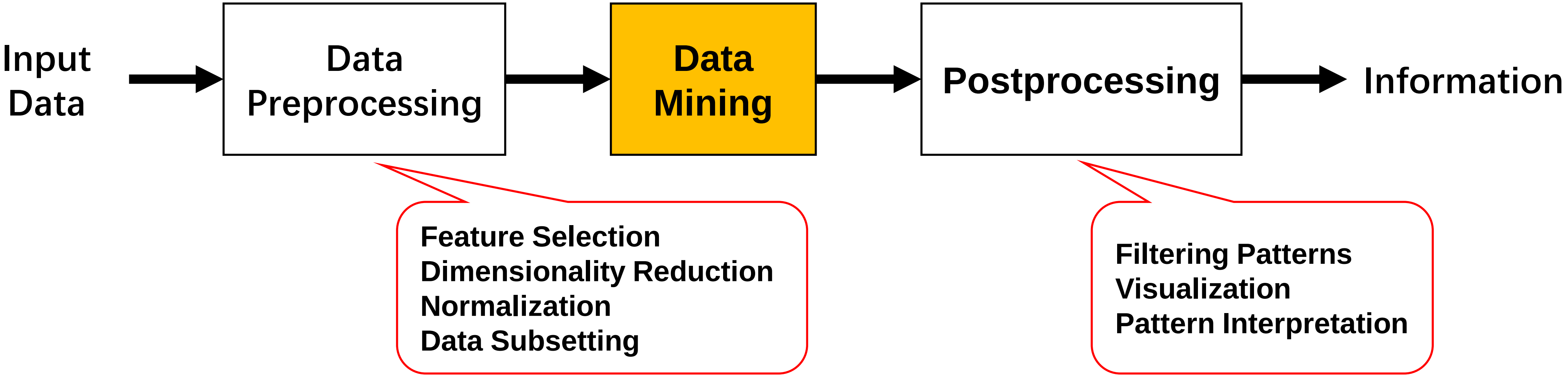


Figure1 The process of knowledge discovery in databases (KDD).

[1] Tan P N, Steinbach M, Kumar V. Introduction to data mining[M]. Pearson Education India, 2016.

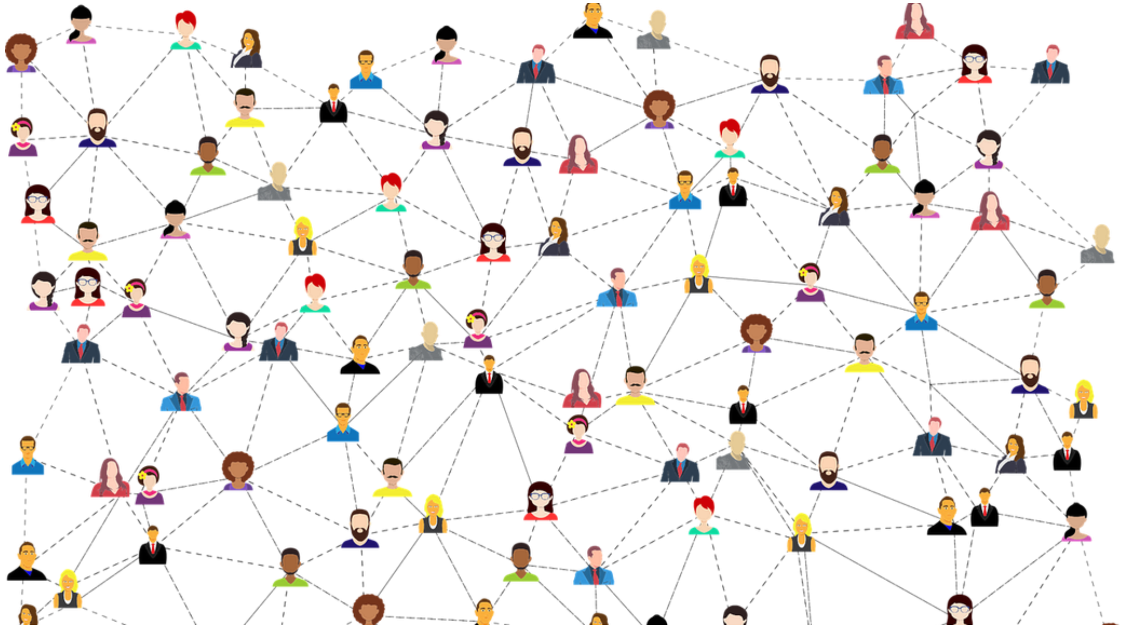
Why Mine Data?

Why Mine Data?

Before answering this question, we must know the fact that there has been enormous data growth in both commercial and scientific databases due to advances in data generation and collection technologies.

Bigger and Bigger Volumes of Data

- Social Network



[1] https://www.thebalancesmb.com/define-social-networks-1794438
[2] https://www.euroscientist.com/imagine-a-social-network-like-facebook-with-no-facebook/

Bigger and Bigger Volumes of Data

- Sensor Network

[1] https://www.youtube.com/watch?app=desktop&v=W1aMmCZ25fw&ab_channel=ToshibaNewsandHighlights
[2] https://www.channelfutures.com/iot/microsoft-pumps-5-billion-into-iot-eyeing-a-more-intelligent-edge-2

Bigger and Bigger Volumes of Data

- Traffic patterns

A Day in the Life of Uber San Francisco [1]

NYC Subway Traffic [2]

[1] https://eng.uber.com/data-visualization-intelligence/
[2] http://piratefsh.github.io/mta-maps/public/

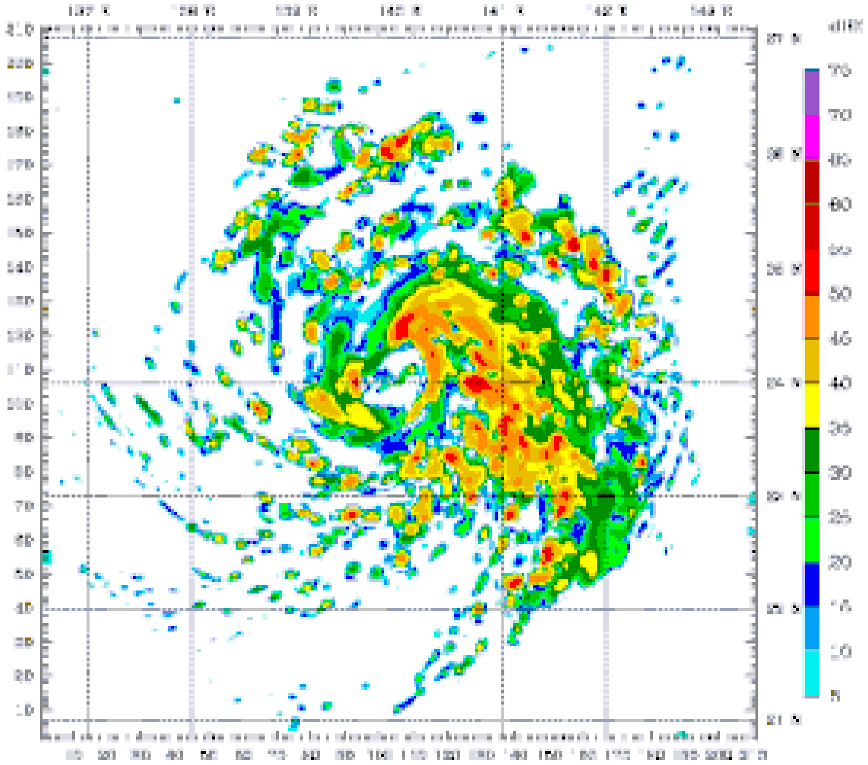
Bigger and Bigger Volumes of Data

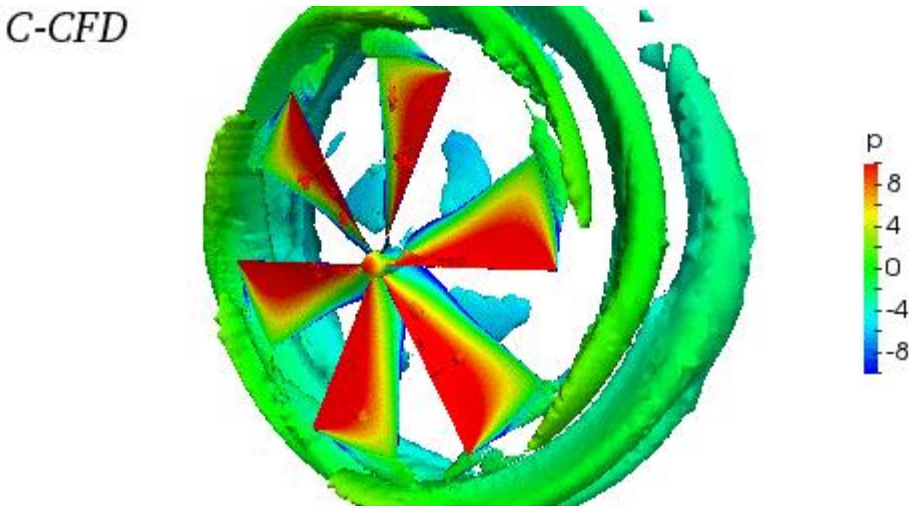
- E-Commerce

[1] https://co-well.vn/en/tech-blog/why-is-ecommerce-important-for-your-business/
[2] https://bulbandkey.com/blog/business/what-is-the-impact-of-e-commerce-on-society/

Bigger and Bigger Volumes of Data

- Computational Simulation





[1] https://en.wikipedia.org/wiki/Computer_simulation
[2] https://centerforcfd.com/

Why Data Mining? Commercial Viewpoint

- Lots of data is being collected and warehoused.
- Computers have become cheaper and more powerful.
- Competitive Pressure is Strong.
  - Provide better, customized services for an edge (e.g. in Customer Relationship Management).

Data → **Knowledge/Decision/Understanding/Profit**

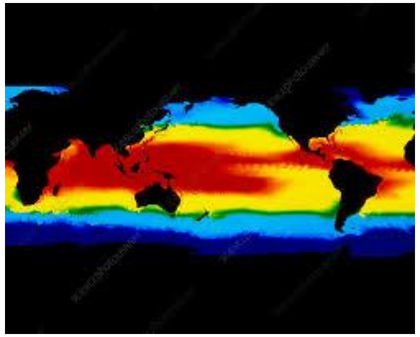
Why Data Mining? Scientific Viewpoint

- Data collected and stored at a very high speed.

fMRI Data

Sky Survey Data

Gene Expression Data



Computational Simulation Data

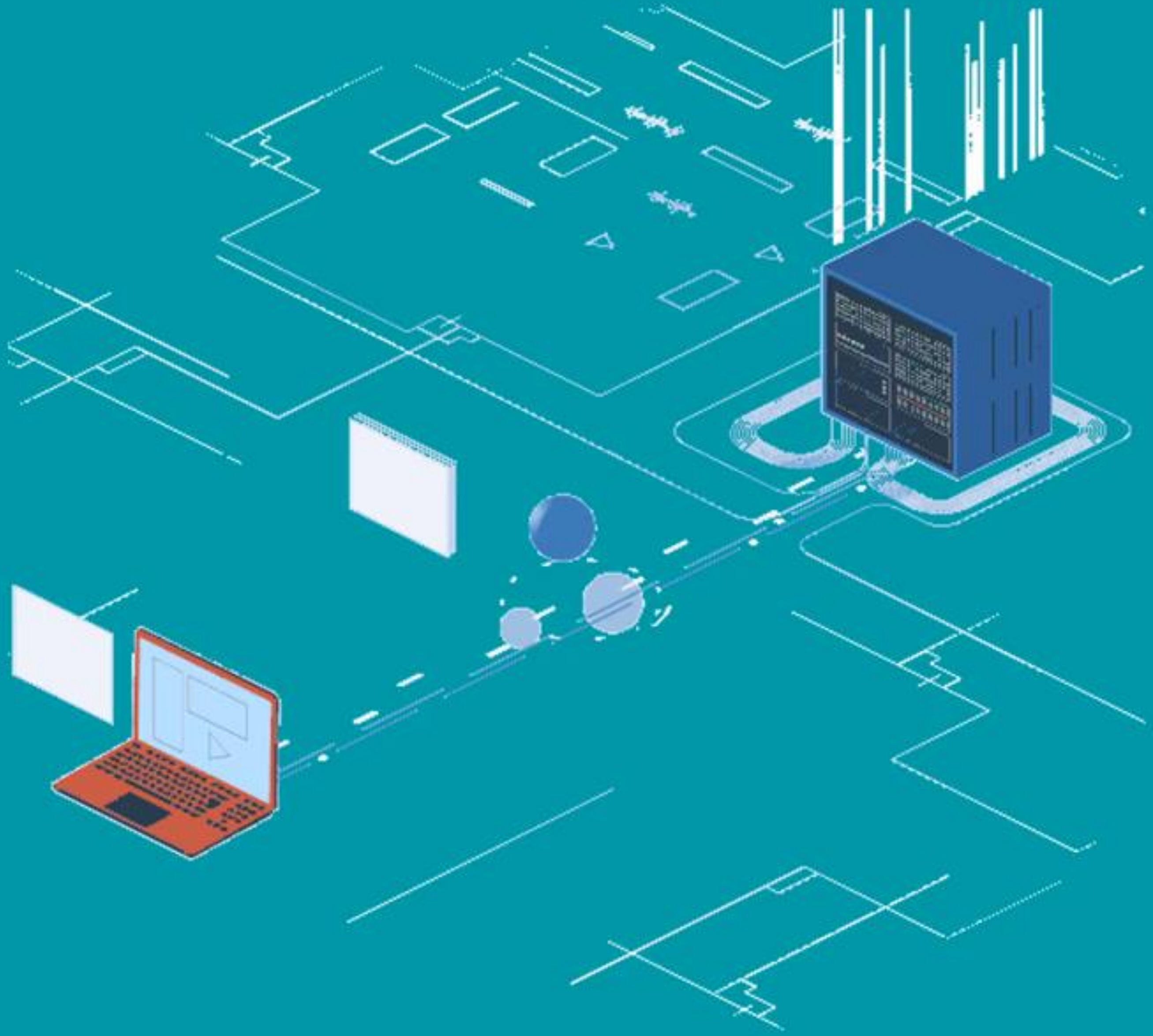
- Data mining helps scientists
  - in automated analysis of massive datasets
  - In hypothesis formation

https://www.spectrumnews.org/news/toolbox/online-tool-can-mix-match-gene-expression-data/
https://www.sciencephoto.com/media/165997/view/computer-simulation-of-global-sea-surface-temp-

Creative Learning

Data Mining

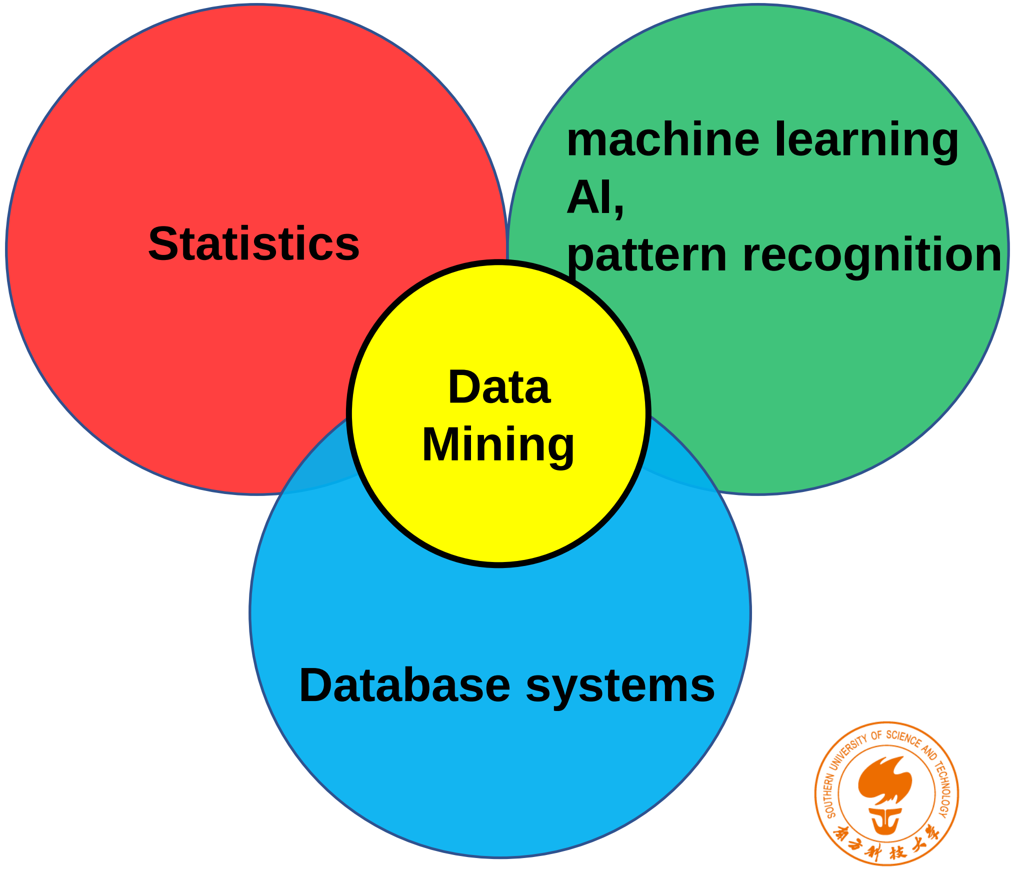
By: Mariam Rashid



Data Mining Tasks

Origins of Data Mining

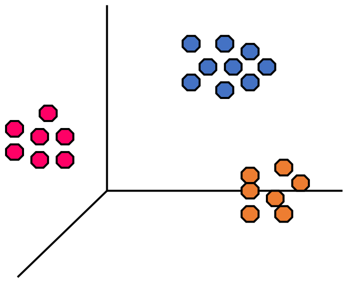
- Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems.
- Traditional techniques may be unsuitable due to:
  - Enormity of data
  - High dimensionality of data
  - Heterogeneous, distributed nature of data



Data Mining Tasks

- Prediction Tasks
  - Use some variables to predict unknown or future values of other variables.
- Description Tasks
  - Find human-interpretable patterns that describe the data.
  - The objective is to derive patterns (correlations, trends, clusters, trajectories, and anomalies) that summarize the underlying relationships in data.

Data Mining Tasks…



Clustering

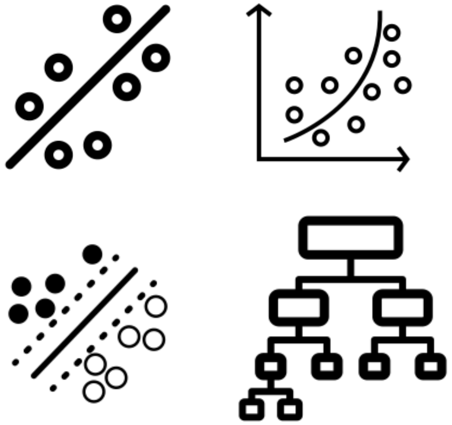
[Descriptive]

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

[Predictive]

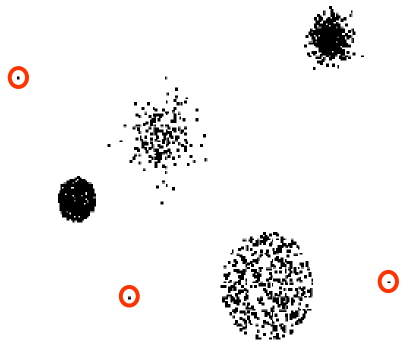


Association Rules

[Descriptive]

Anomaly Detection

[Descriptive]



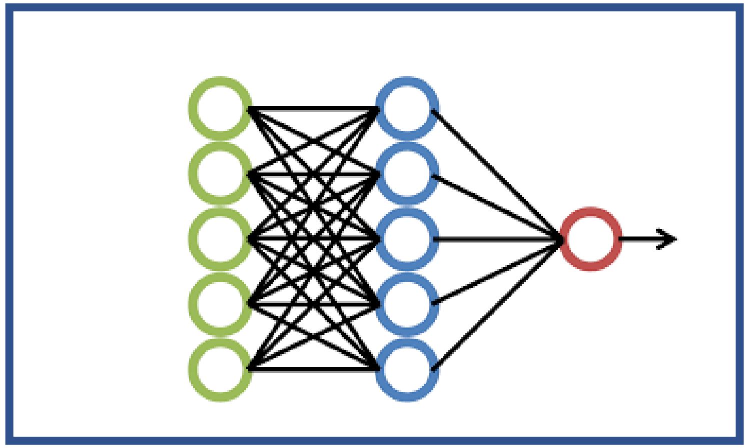
Predictive Modeling: Classification

- 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.

Class

Model for predicting credit worthiness

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
…	…	…	…	…

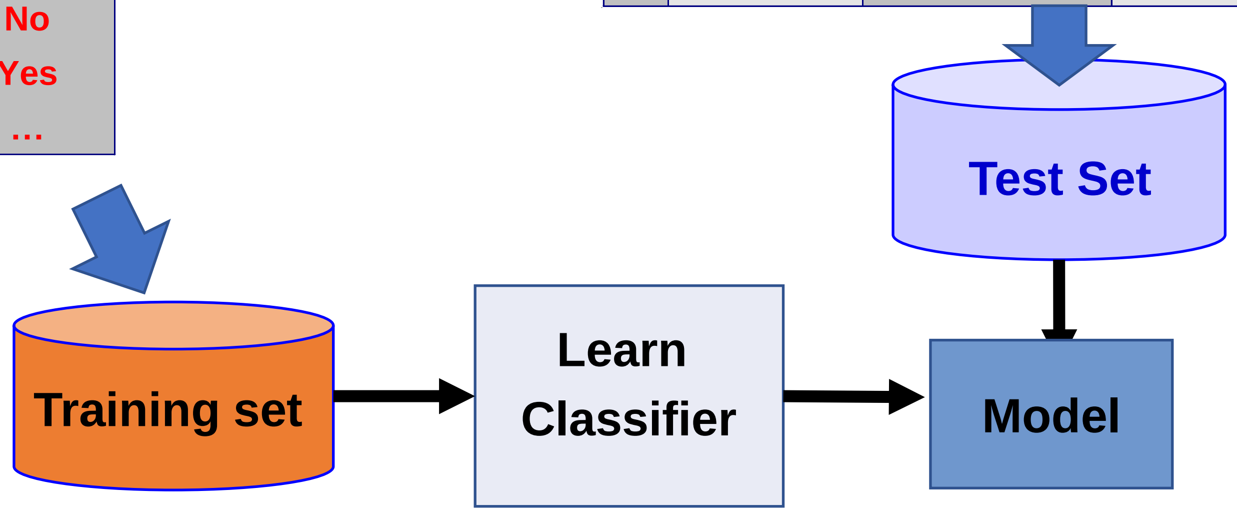


Predictive Modeling: Classification

- 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.

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	?
…	…	…	…	…

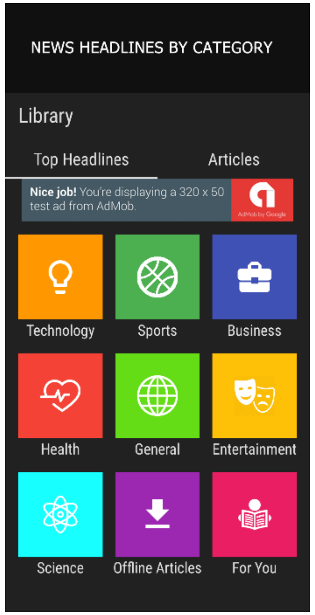


Classification: Application

- Classifying credit card transactions as legitimate or fraudulent

Classification: Application

- Categorizing news stories as finance, weather, entertainment, sports, etc.



NEWS HEADLINES BY CATEGORY

Library

Top Headlines | Articles

Nice job! You're displaying a 320 x 50 test ad from AdMob.

https://in.pinterest.com/pin/631770653951937625/
https://honestreporting.com/news-literacy-the-eight-categories-of-media-bias/

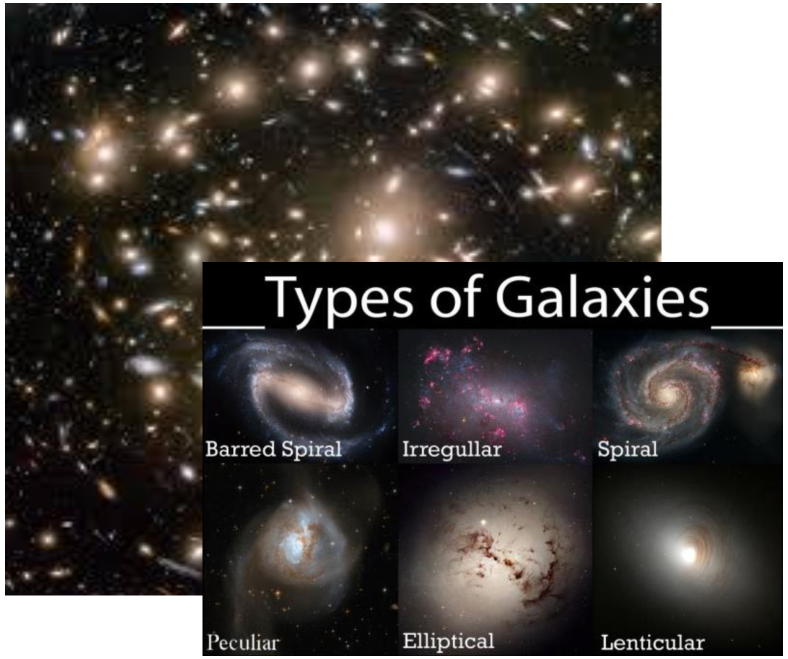
Classification: Application

- Classifying emails as normal or spam

Blanzieri, E. & A. Bryl. “A survey of learning-based techniques of email spam filtering”
Artificial Intelligence Review
March 2008, Vol. 29, Issue 1, pp 63–92

Classification: Application

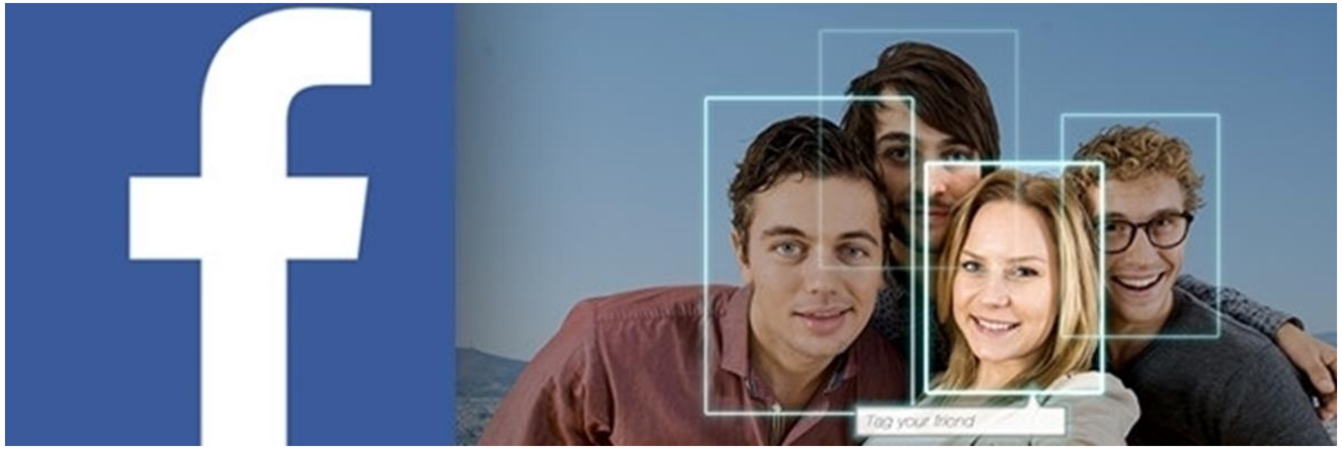
- Classifying galaxies in the universe

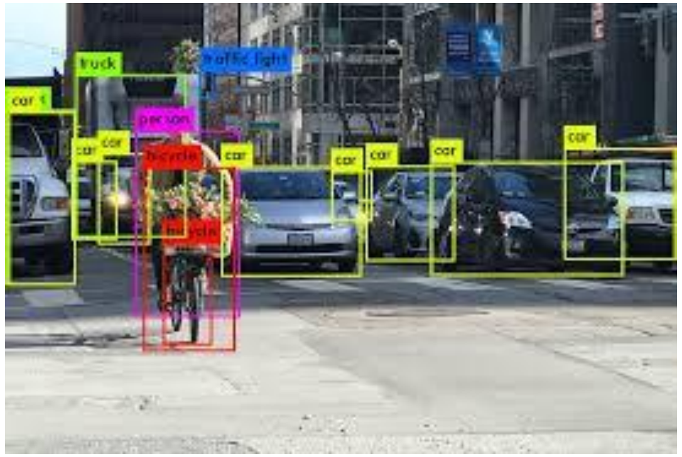


Fowler, L., Schawinski, K., & Brandt, B.-E.
Galaxy Classification using Machine Learning.
Paper presented at the American Astronomical
Society Meeting Abstracts. 2017

Classification: Application

- Image and video processing

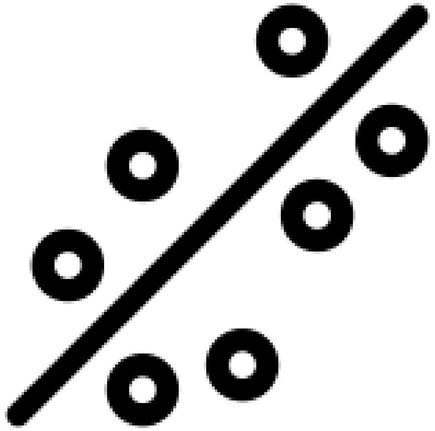


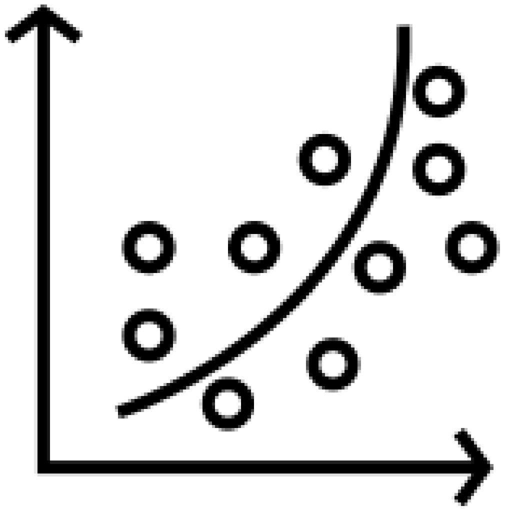


https://www.classaction.org/blog/facebook-sued-over-face-recognition-feature

Predictive Modeling: Regression

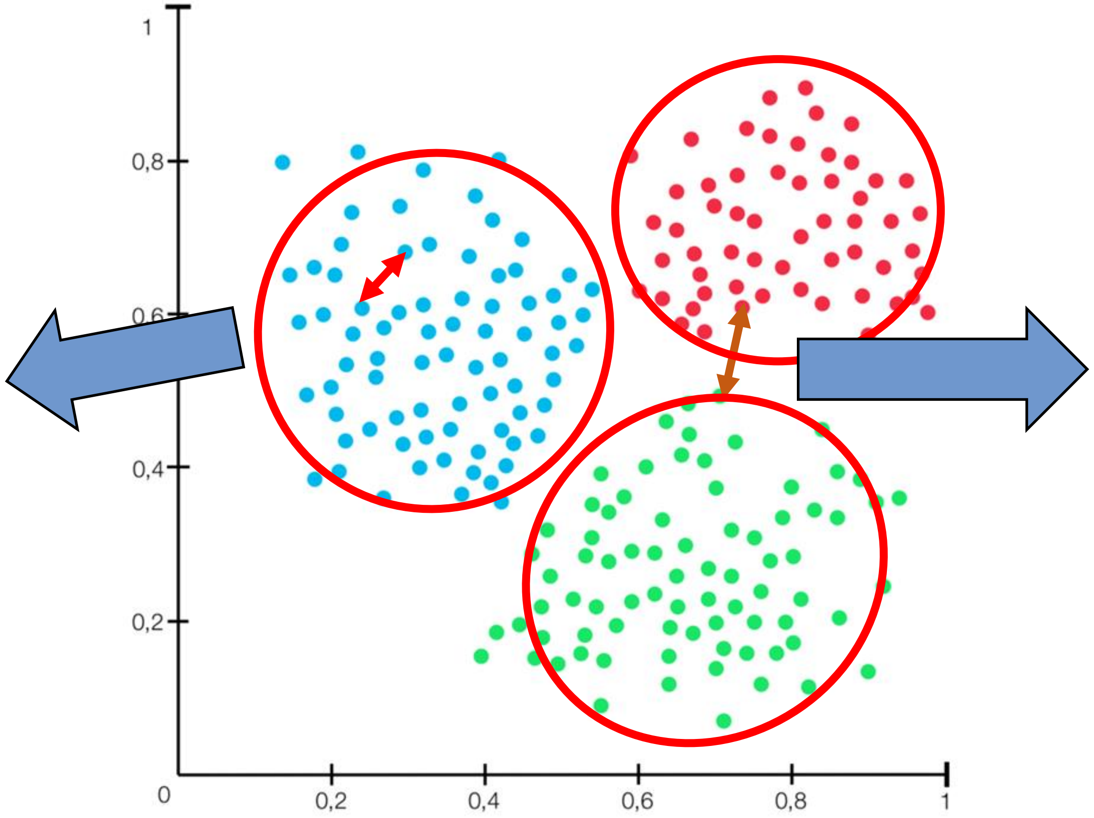
- Predict a value of a given **continuous** valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.





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



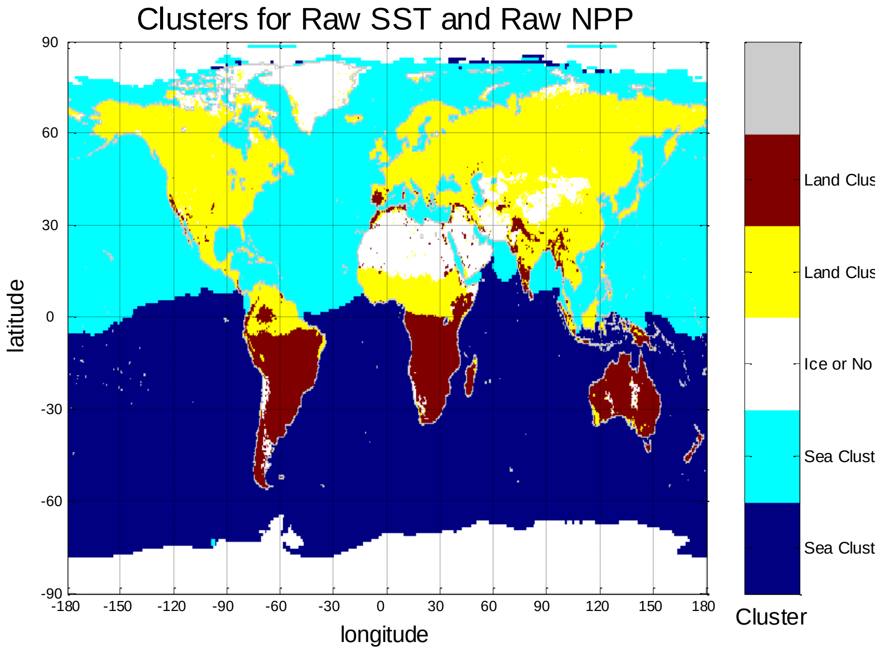
Intra-cluster distances are minimized

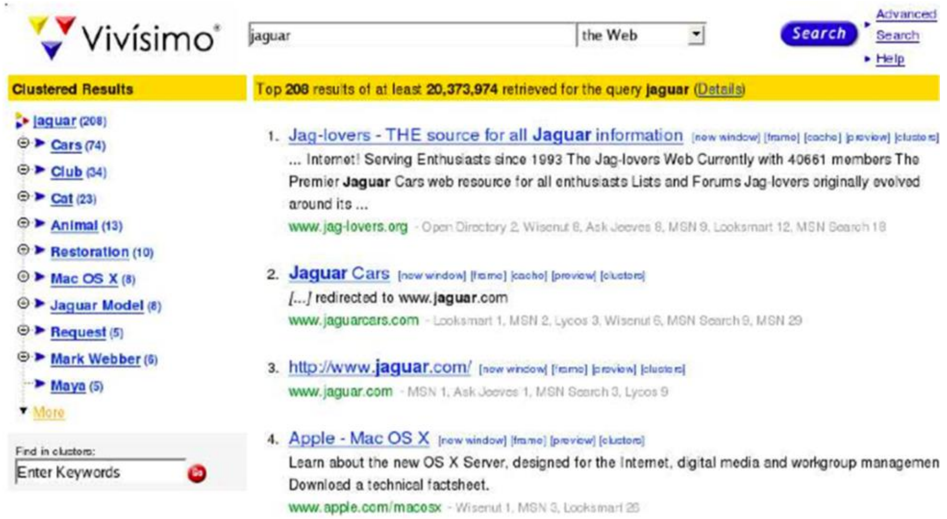
Inter-cluster distances are maximized

Applications of Cluster Analysis

- Data summarization, compression, and reduction
- Collaborative filtering, recommender systems, or customer segmentation
  - Find like-minded users or similar products

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





Clustered Results

- jaguar (208)
  - Cars (74)
  - Club (34)
  - Cat (23)
  - Animal (13)
  - Restoration (10)
  - Mac OS X (8)
  - Jaguar Model (8)
  - Request (5)
  - Mark Webber (6)
  - Maya (5)
  - More

Top 208 results of at least 20,373,974 retrieved for the query **jaguar** (Details)

1. Jag-lovers - THE source for all Jaguar information [new window] [frame] [cache] [preview] [clusters]
   ... Internet! Serving Enthusiasts since 1993 The Jag-lovers Web Currently with 40661 members The Premier **Jaguar** Cars web resource for all enthusiasts Lists and Forums Jag-lovers originally evolved around its ...
   www.jag-lovers.org - Open Directory 2, Wisenut 8, Ask Jeeves 8, MSN 9, Looksmart 12, MSN Search 18

2. Jaguar Cars [new window] [frame] [cache] [preview] [clusters]
   [...] redirected to www.**jaguar**.com
   www.jaguarcars.com - Looksmart 1, MSN 2, Lycos 3, Wisenut 6, MSN Search 9, MSN 29

3. http://www.jaguar.com/ [new window] [frame] [preview] [clusters]
   www.jaguar.com - MSN 1, Ask Jeeves 1, MSN Search 3, Lycos 9

4. Apple - Mac OS X [new window] [frame] [preview] [clusters]
   Learn about the new OS X Server, designed for the Internet, digital media and workgroup managemen
   Download a technical factsheet.
   www.apple.com/macosx - Wisenut 1, MSN 3, Looksmart 26

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.
  - Association rule discovery can be applied to find items that are frequently bought together by customers.

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

- 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.

Deviation/Anomaly/Change Detection

- Detect significant deviations from normal behavior
- 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.

Other Data Mining Resources

Data Mining Resource: Books

- "Data Mining: Practical Machine Learning Tools and Techniques (4th Edition)" I.H. Witten, E. Frank, M. Hall, C. Pal. Morgan Kaufmann Publishers. 2017.
- Introduction to Data Mining (2nd edition) P.-N. Tan, M. Steinbach, A. Karpatne, V. Kumar. Pearson, 2018.
- "Data Mining: Concepts and Techniques (3rd Edition)". J. Han and M. Kamber. Morgan Kaufmann Publishers. 2012.
- "Advances in Knowledge Discovery and Data Mining". Eds.: Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy. The MIT Press, 1995.
- …

Data Mining Resource: Journals

- Data Mining and Knowledge Discovery Journal
- ACM SIGKDD Explorations Newsletter
- TKDE: IEEE Transactions in Knowledge and Data Engineering
- TODS: ACM Transactions on Database Systems
- JACM: Journal of ACM
- Data and Knowledge Engineering
- JIIS: Intl. Journal of Intelligent Information Systems
- …

Data Mining Resource: Conferences

- KDD: ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining
- ICDM: IEEE International Conference on Data Mining,
- SIAM International Conference on Data Mining
- PKDD: European Conference on Principles and Practice of Knowledge Discovery in Databases
- PAKDD Pacific-Asia Conference on Knowledge Discovery and Data Mining
- DaWak: Intl. Conference on Data Warehousing and Knowledge Discovery

Other related Conferences:

- ICML: Intl. Conf. On Machine Learning
- IDEAL: Intl. Conf. On Intelligent Data Engineering and Automated Learning
- IJCAI: International Joint Conference on Artificial Intelligence
- AAAI: American Association for Artificial Intelligence Conference
- SIGMOD/PODS: ACM Intl. Conference on Data Management
- ICDE: International Conference on Data Engineering
- VLDB: International Conference on Very Large Data Bases

SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY
南方科技大学

End of Class 1