

CZ/CE 4032, SC4020

Data Analytics & Mining

Lin Guosheng
School of Computer Science and Engineering
Nanyang Technological University

About me

- Instructor of the 1st half (Week1 – Week6):

A/Prof Lin Guosheng



- Specialized in computer vision
 - Email or MS Teams: gslin@ntu.edu.sg
 - Homepage: <https://guosheng.github.io/>
- Q&A:
 - By emails or MS Teams messages
 - After each lecture or tutorial

- Instructor of the 2nd half (Week7-Week12):

Prof Cong Gao



- Homepage:
<https://personal.ntu.edu.sg/gaocong/>
- Email: gaocong@ntu.edu.sg

Assessment

- Final exam (50%) + Assignments (50%)
 - Closed book exam
- Assignments
 - 2 group-based projects: (25% + 25%)
 - Project 1 will be announced in Week 2
 - Project 2 will be announced in the 2nd half

- Please start to form your group now
 - please edit the online form to create your group.
 - https://docs.google.com/spreadsheets/d/1YMiW326R1CJk_C9DklqqmhtLcQMhNEQ-0nQeKROs20/edit?usp=sharing
 - Each group is limited to 4 members.
 - The first person for each group in the form is the coordinator and contact person.
 - You can form a group of less than 4 members.
- Grouping will be finalized in week 3

Topics (Tentative)

- The 1st half (Week 1 to Week 6)
 - Introduction
 - Clustering – basic methods
 - Link analysis – PageRank
 - Graph neural network
 - Similarity search
 - Clustering – advanced methods
 - Graph community detection
- The 2nd half (Week 7 to Week 12)
 - Association Rule Mining
 - Classification
 - Recommendation Systems
 - Data Processing, Data Cleaning and integration
 - others

- Online materials:

- CS246: Mining Massive Data Sets

- <https://web.stanford.edu/class/cs246/>

- Many slides used in this course are from CS246

- CS224W: Machine Learning with Graphs

- <http://web.stanford.edu/class/cs224w/>

- Book: Mining of Massive Datasets

- By Jure Leskovec, Anand Rajaraman, Jeff Ullman

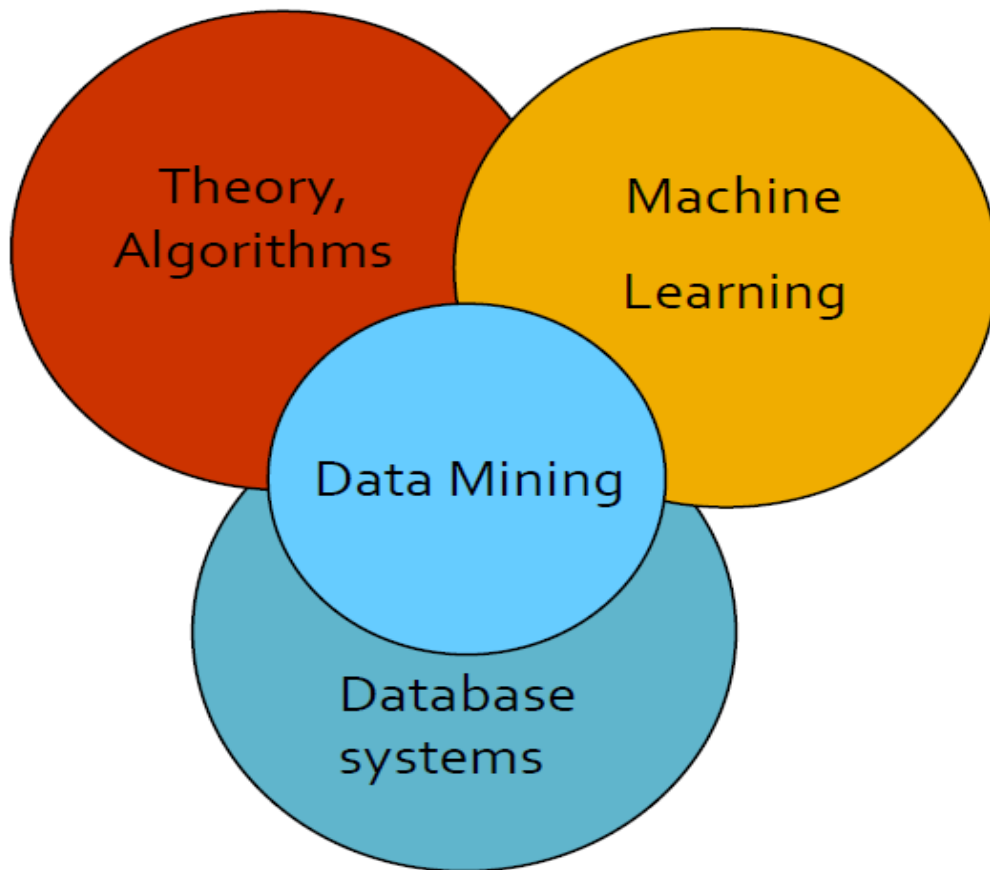
- <http://www.mmids.org/>

- What is data mining
 - Data mining is the process of extracting patterns/knowledge from data.
- Data needs to be
 - Stored (computer systems)
 - Managed (databases)
 - And **ANALYZED (this class)**

Data Mining \approx Knowledge Discovery in Data (KDD)
 \approx Big Data \approx Data Science \approx Machine Learning
Part of Artificial Intelligence

Slides from CS246: Mining Massive Data Sets

Relationship with other subjects:



Also highly related to

1. Computer vision
(image/video data)
2. Natural language process
(text data)
3. Deep learning or
deep neural networks

Application: recommender system

Amazon.com: Bestselling Canon Cameras

[Newsletters](#) | [X](#)

★ [Amazon.com](#) to me

[show details](#) May 30 (9 days ago)

[Reply](#)

amazon.com

More to Explore

Customers who have shown an interest in point-and-shoot cameras might like to see this week's bestselling models.



[Canon PowerShot A495 10.0 MP Digital Camera with 3.3x Optical Zoom and 2.5-Inch LCD \(Blue\)](#)



[Canon PowerShot A3000IS 10 MP Digital Camera with 4x Optical Image Stabilized Zoom and 2.7-Inch LCD](#)



[Canon PowerShot ELPH 300 HS 12 MP CMOS Digital Camera with Full 1080p HD Video \(Black\)](#)




[Canon PowerShot S95 10 MP Digital Camera with 3.8x Wide Angle Optical Image Stabilized Zoom and 3.0-Inch LCD](#)

Amazon.com


<https://research.aimultiple.com/recommendation-system/>

People who are skilled in Microsoft Excel also subscribe to these newsletters

See all


Published weekly

Artificial Intelligence (AI)
Artificial Intelligence & Machine Learning are arguably the most transformative technologies...




Bernard Marr
Internationally best-selling au...

Subscribe


Published weekly

Data Foundation
No organization can thrive without a strong data foundation. Data is the critical asset for...




Jose Almeida
Global Data Management Co...

Subscribe


Published weekly

Mind to Heart
Upgrading to the new human operating system




Rudy de Waele
Keynote Speaker (Virtual & R...

Subscribe


Published weekly

Ludonomics
Ludonomics is a weekly on Allianz markets, macro, sector, and insurance research by its...




Ludovic Subran
Chief Economist at Allianz

Subscribe


Published weekly

Perspectives 4 Active Citizens
Challenged to go beyond tweets, my weekly Standard newspaper...




Irūngū Houghton
Executive Director at Amnest...

Subscribe

Published monthly

Canada Workforce Report
LinkedIn's Workforce Confidence Index shows how Canada's professionals feel now about...



Riva Gold
Editor, Daily News at LinkedIn...

Subscribe

LinkedIn

<https://research.aimultiple.com/recommendation-system/>

NETFLIX

Browse

Personalize

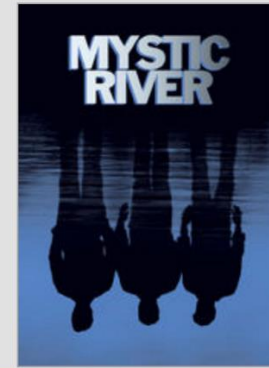
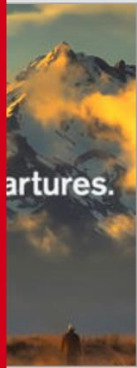
DVDs

Search

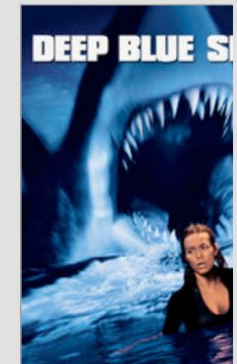
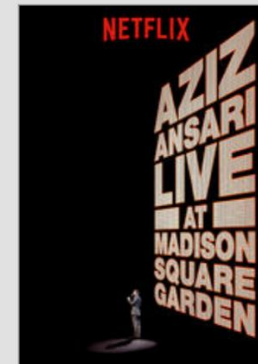
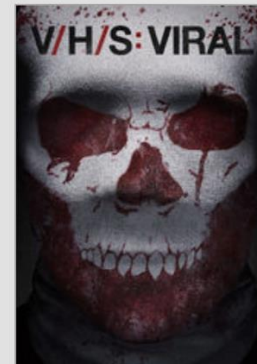
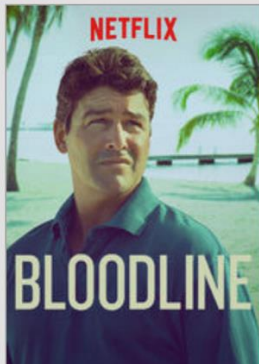


Antsy Ann

Top Picks for Antsy Ann



Popular on Netflix



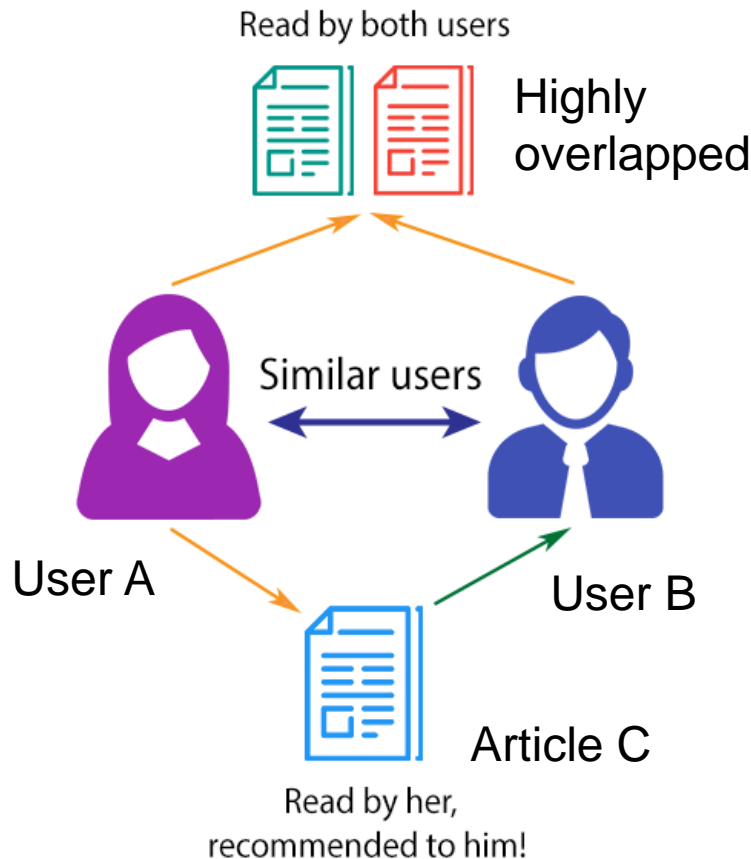
<https://rpubs.com/dhnanjay/286571>

Netflix Prize (2006): recommendation competition

<https://www.thrillist.com/entertainment/nation/the-netflix-prize>

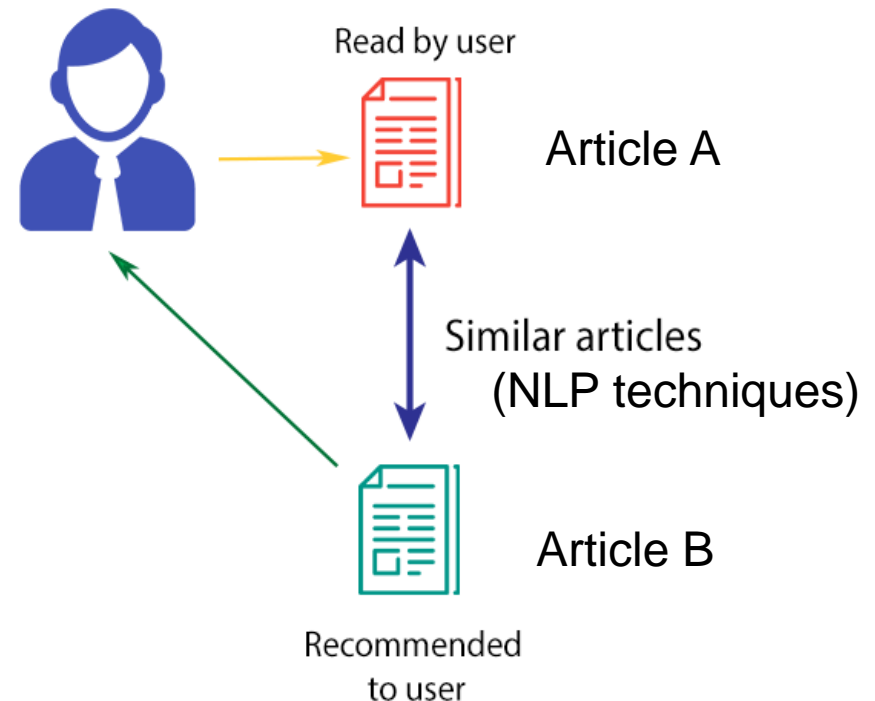
Recommender system

COLLABORATIVE FILTERING



Focus on the similarity between users

CONTENT-BASED FILTERING



Focus on the similarity between items

A simple example for collaborative filtering.
Can be formulated as a missing value prediction problem:

	Movie A	Movie B	Movie C	Movie D
User 1	5	4	1	1
User 2	2	3	2	4
User 3	2	4	?	1
User 4	2	3	1	?

Fill in missing value (User 4, Movie D) in the table

A simple example for collaborative filtering

	Movie A	Movie B	Movie C	Movie D
User 1	5	4	1	1
User 2	2	3	2	4
User 3	2	4	?	1
User 4	2	3	1	?

Predict the missing value (User 4, Movie D) in the table

Collaborative filtering:

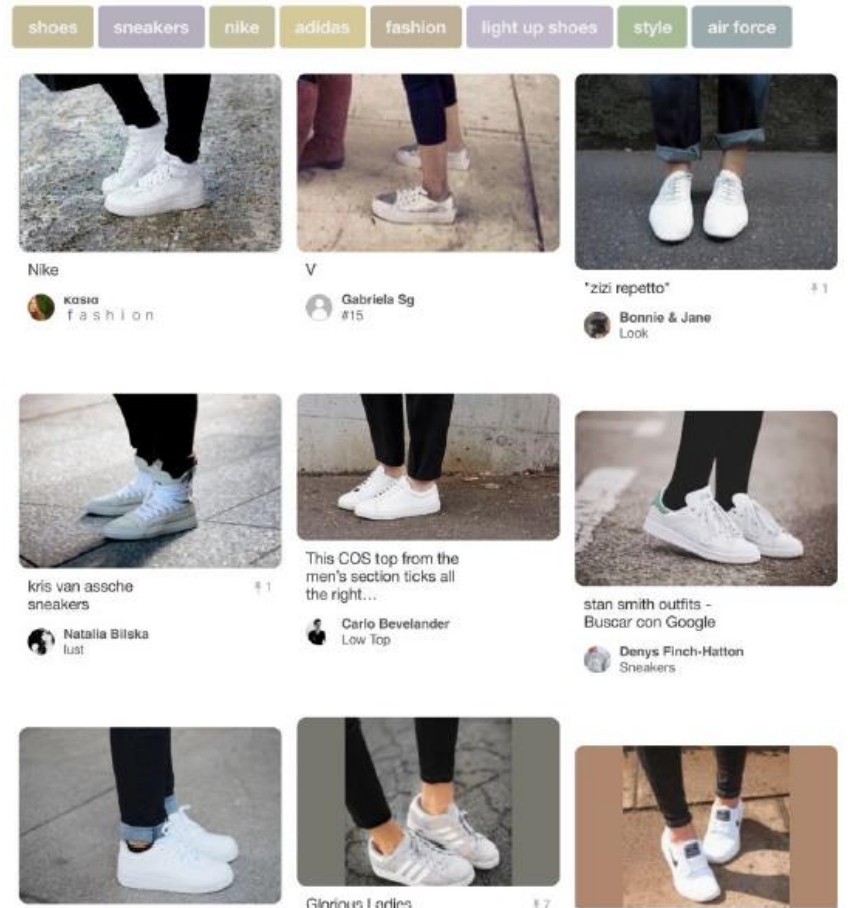
User 4 is similar to user 2, so we can predict 4 for (User 4, Movie D)

Un-personalized prediction (prior):

Use the average score as the prediction: $(1+4+1)/3 = 2$

Application: information retrieval

Visually similar results

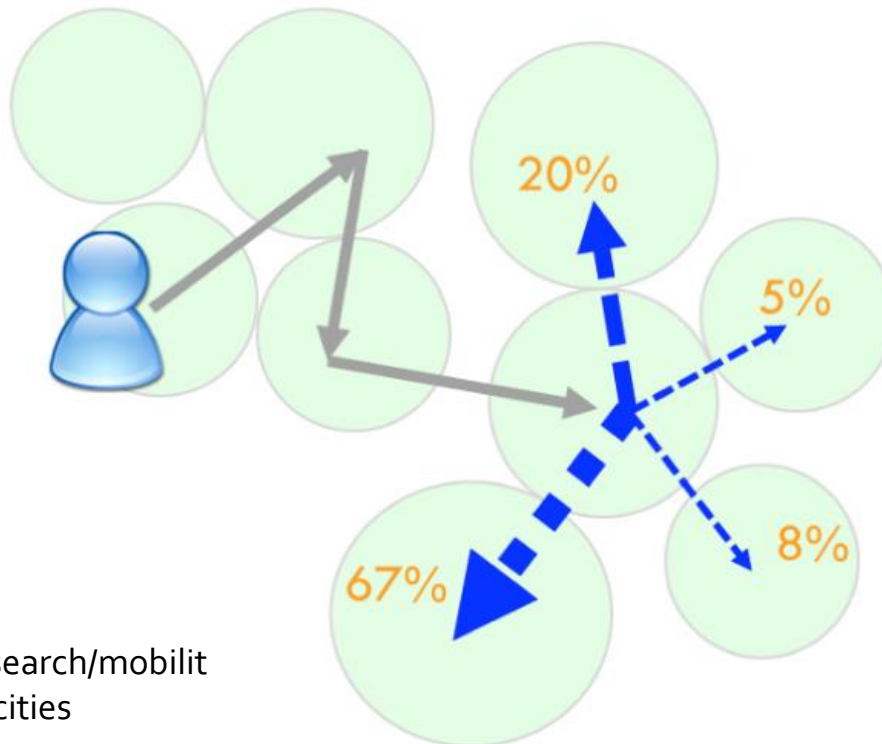




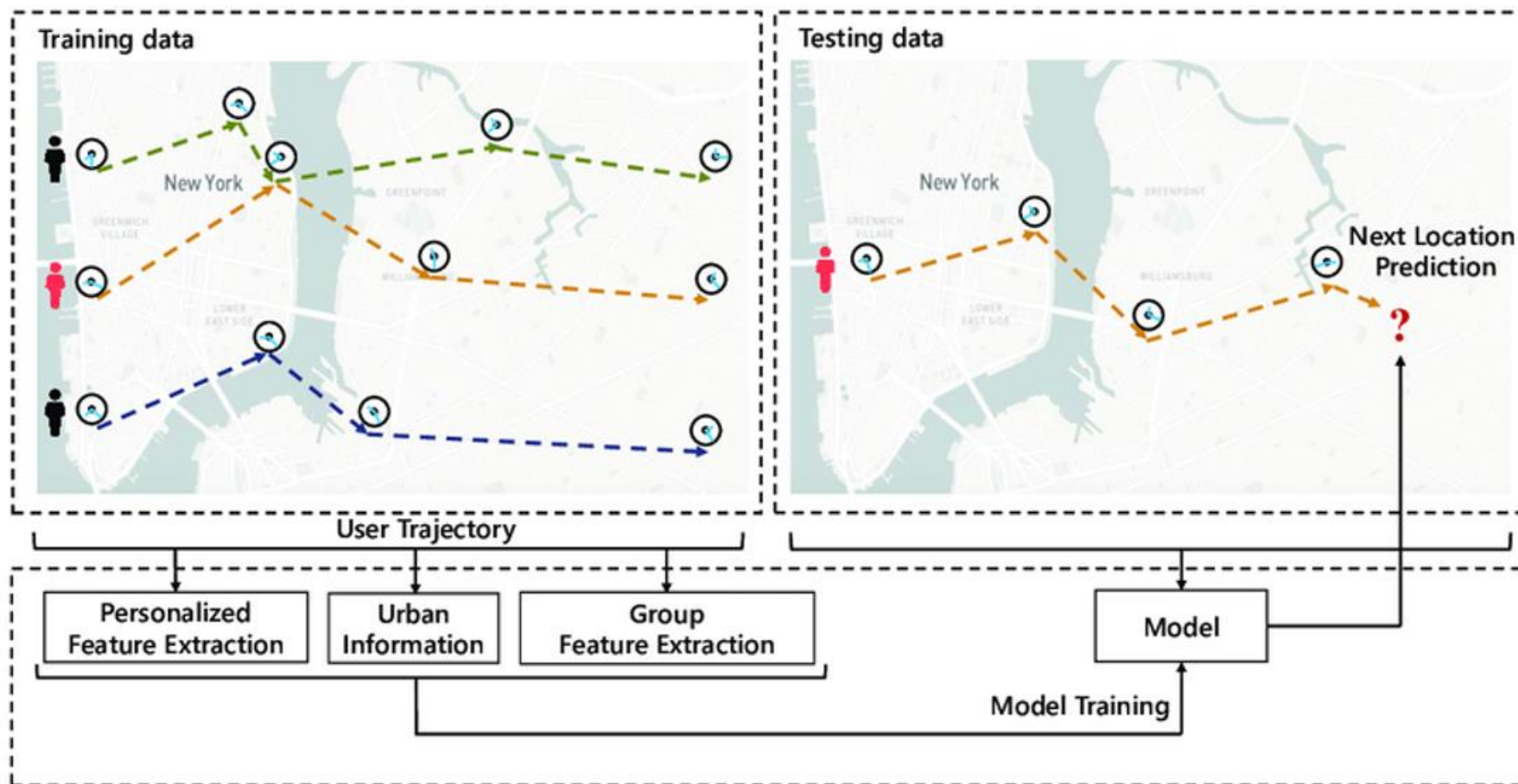
Source: FRSFN: A semantic fusion network for practical fashion retrieval

Application

- Urban computing, smart city
 - Next location prediction/ future position prediction
 - Spatial-temporal trajectory data (e.g., GPS data + time)
 - Predict traffic congestion, improve traffic management



Source:
<https://kdd.isti.cnr.it/research/mobility-data-mining-science-cities>



Source: PG2Net: Personalized and Group Preferences Guided Network for Next Place Prediction

Application

- Urban computing, smart city

- Urban planning

- Choose the address for a hospital/bus stop?

- Social computing

Analyse human behavior based on social data

- Examples of social data: social media, blogs, social networks
 - Social media platforms: Twitter, Facebook, etc.

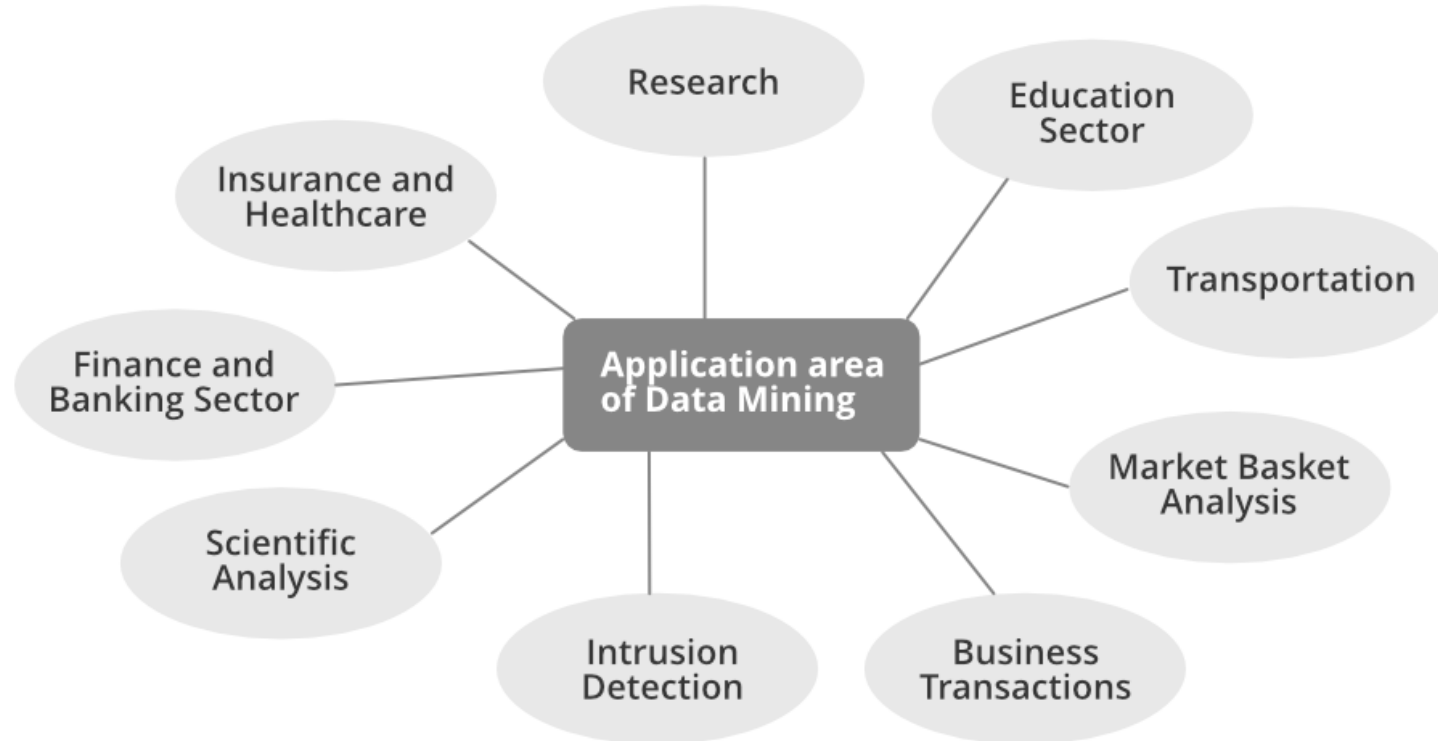
- Social Media Analysis

- Understand user behavior and activities

- Social Network Analysis

- Community detection

Many other applications



<https://www.geeksforgeeks.org/applications-of-data-mining/>

Conferences and Journals

Where you can find various research topics and research papers on data mining

- KDD 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](#))
 - Int. Conf. on Web Search and Data Mining ([WSDM](#))
- Other related conferences
 - DB conferences: ACM SIGMOD, VLDB, ICDE, EDBT, ICDT, ...
 - Web and IR conferences: WWW, SIGIR, WSDM
 - ML conferences: ICML, NeurIPS
 - CV conferences: CVPR, ICCV
 - NLP conferences: ACL
- Journals
 - Data Mining and Knowledge Discovery (DAMI or DMKD)
 - IEEE Trans. On Knowledge and Data Eng. (TKDE)
 - KDD Explorations
 - ACM Trans. on KDD