CS247: ADVANCED DATA MINING

1: Introduction

Instructor: Yizhou Sun

yzsun@cs.ucla.edu

January 6, 2024

Course Information

- Course homepage: https://github.com/yichousun/Winter2024_CS247_AdvDM/
- Class Schedule
  - Slides
  - Assignments
  - Course projects
  - ...
- Piazza: https://piazza.com/ucla/winter2024/cs247

- Prerequisites
  - Required prerequisite courses: CS 145 or CS 146 or equivalent.
  - You are expected to have background knowledge in data structures, algorithms, basic linear algebra, and basic statistics.
  - You are expected to know basic knowledge in data mining and machine learning.
  - You will also need to be familiar with at least one programming language, and have programming experiences.

Meeting Time and Location

- When
  - M&W, 10-11:50pm
- Where
  - BOELTER 4760

Instructor and TA Information

- Instructor: Yizhou Sun
  - Homepage: http://web.cs.ucla.edu/~yzsun/
  - Email: yzsun@cs.ucla.edu
  - Office: 3531F
  - Office hour: Tuesdays 4:45-6:15pm
- TA:
  - Fred Xu (fredxu at cs.ucla.edu), office hours: Monday 4-5PM and Wednesday 4-5PM
  - Yanqiao Zhu (yzhu at cs.ucla.edu), office hours: Monday 8-9PM (Zoom) and Tuesday in person 10-11am
  - TA Office: BH 3256S

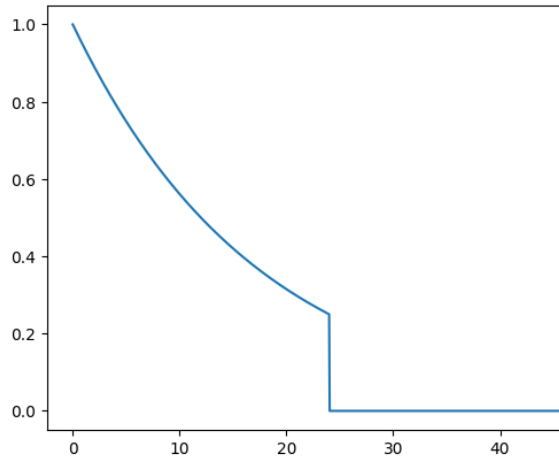
Grading

- Homework: 40%
- Course project: 30%
- Exam:25%
- Participation: 5%

Grading: Homework

- Homework: 40%
  - 6 assignments (highest 5 scores will be picked)
  - Bonus questions might be included, which can be carried over to the final score
  - Deadline: 11:59pm of the indicated due date via *Bruinlearn* system
    - *Late submission policy:* get original score* $1(t <= 24)e^{-(ln(2)/12)*t}$

    if you are t hours late.



  - No copying or sharing of homework!
    - But you can discuss general challenges and ideas with others
    - *Suspicious cases will be reported to The Office of the Dean of Students*

Grading: Course Project

- Course project: 30%
  - Group project (4-5 people for one group)
  - Goal: Solve an open data mining problem, related to text, graph/networks, or recommender systems
  - You are expected to present your project to the class, and submit a project report and your code at the end of the quarter
  - Warning: private peer review will be collected in the end, and free rider will receive a significant lower score

Grading: Exam

- Midterm for in-person sessions: around the Evening of Feb. 22
- Midterm for MSOL: Feb. 10

- Optional Final: March 22, 2024 Friday 11:30am-2:30pm
- Optional Final for MSOL: Saturday, March

- Exam score = max(midterm, final)

Grading: Participation

- Participation: 5%
  - Popup In-class quizzes
    - For Session LEC 1:
      - In Class
    - For Session LEC 80:
      - No
  - Piazza
    - Bonus if you are very active

Q&A

Goals of the Course

- Review and understand fundamentals of basic data mining techniques
- Learn recent data mining techniques on several advanced data types
- Develop skills to apply data mining algorithms to solve real-world applications
- Gain initial experience in conducting research on data mining

Content to be Covered

- Part I: Basics of data mining
- Part II: Text mining
- Part III: Graphs/Network mining
- Part IV: Recommender systems

Part I: Basics of data mining (~2 weeks)

- Basics of Probabilistic Models
- K-means and mixture models
- Neural networks and deep learning

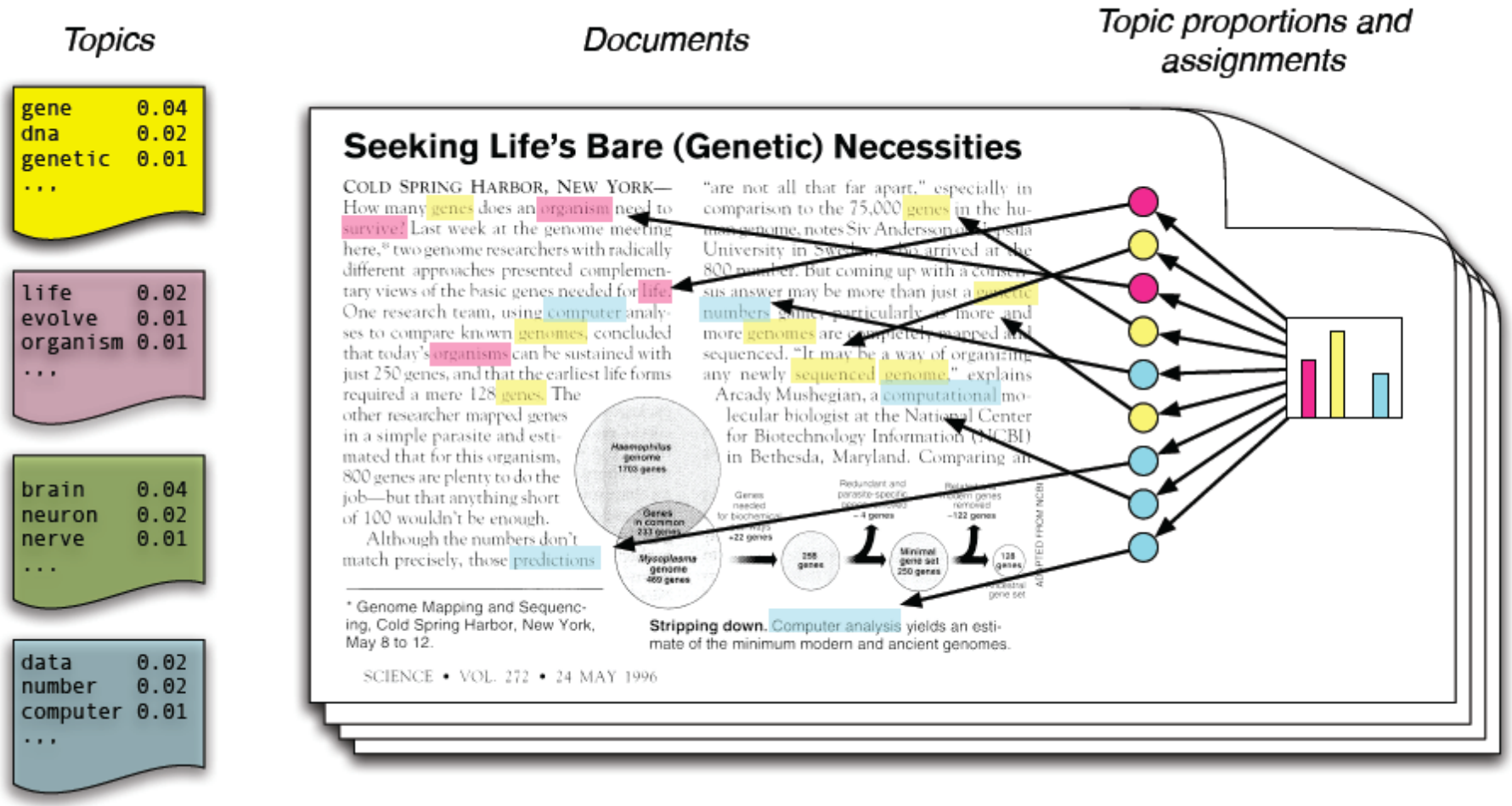
Part II: Text mining (~2 weeks)

- Topic Models
- Word Embedding
- Transformers

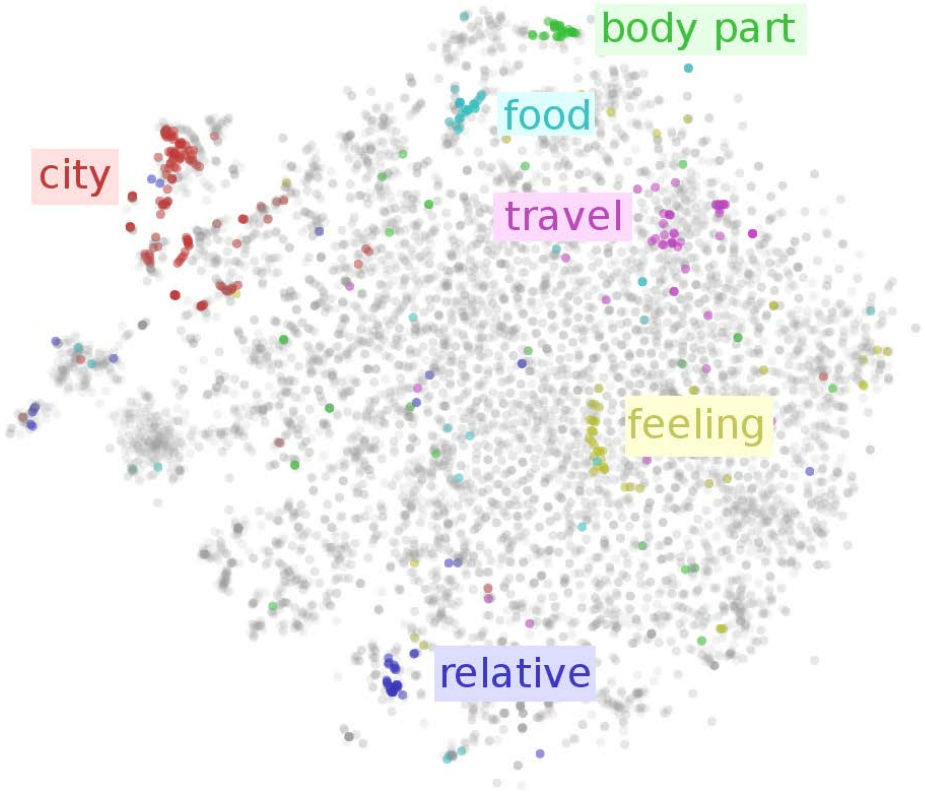
Text Data

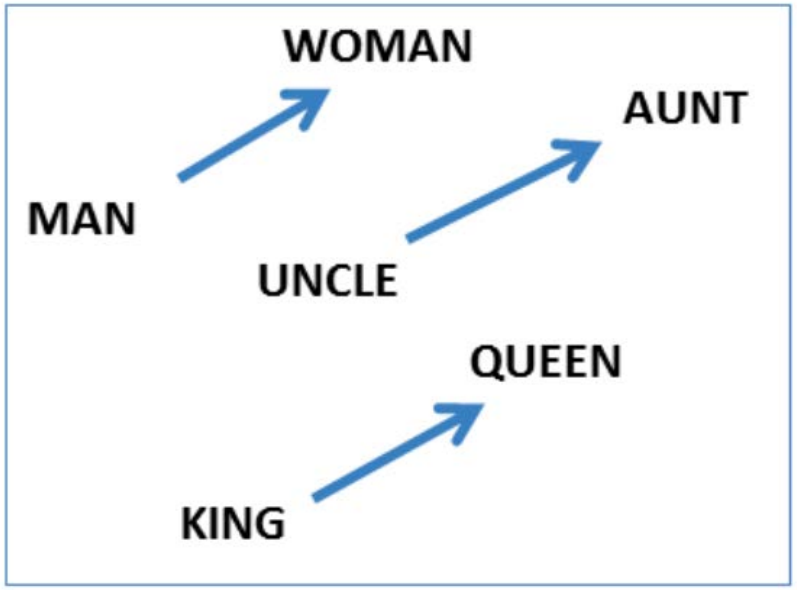
- “Text mining, also referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' in text mining usually refers to some combination of relevance, novelty, and interestingness. Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling (i.e., learning relations between named entities).” –from wiki

Text Data – Topic Modeling



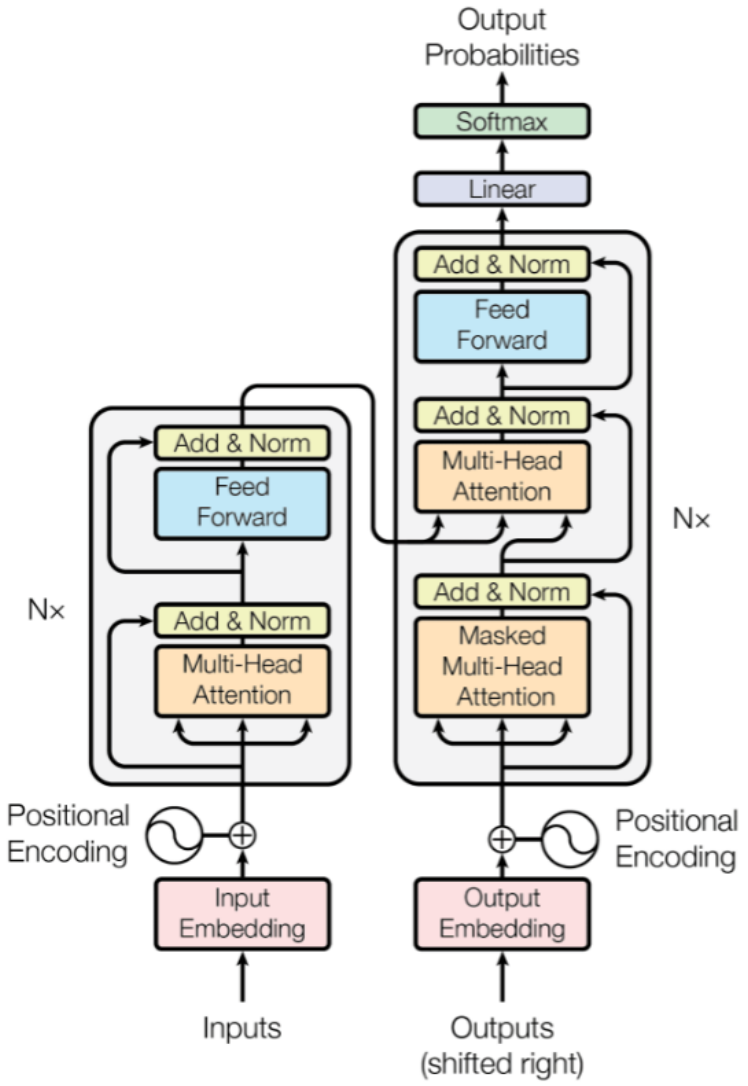
Text Data – Word Embedding





king - man + woman = queen

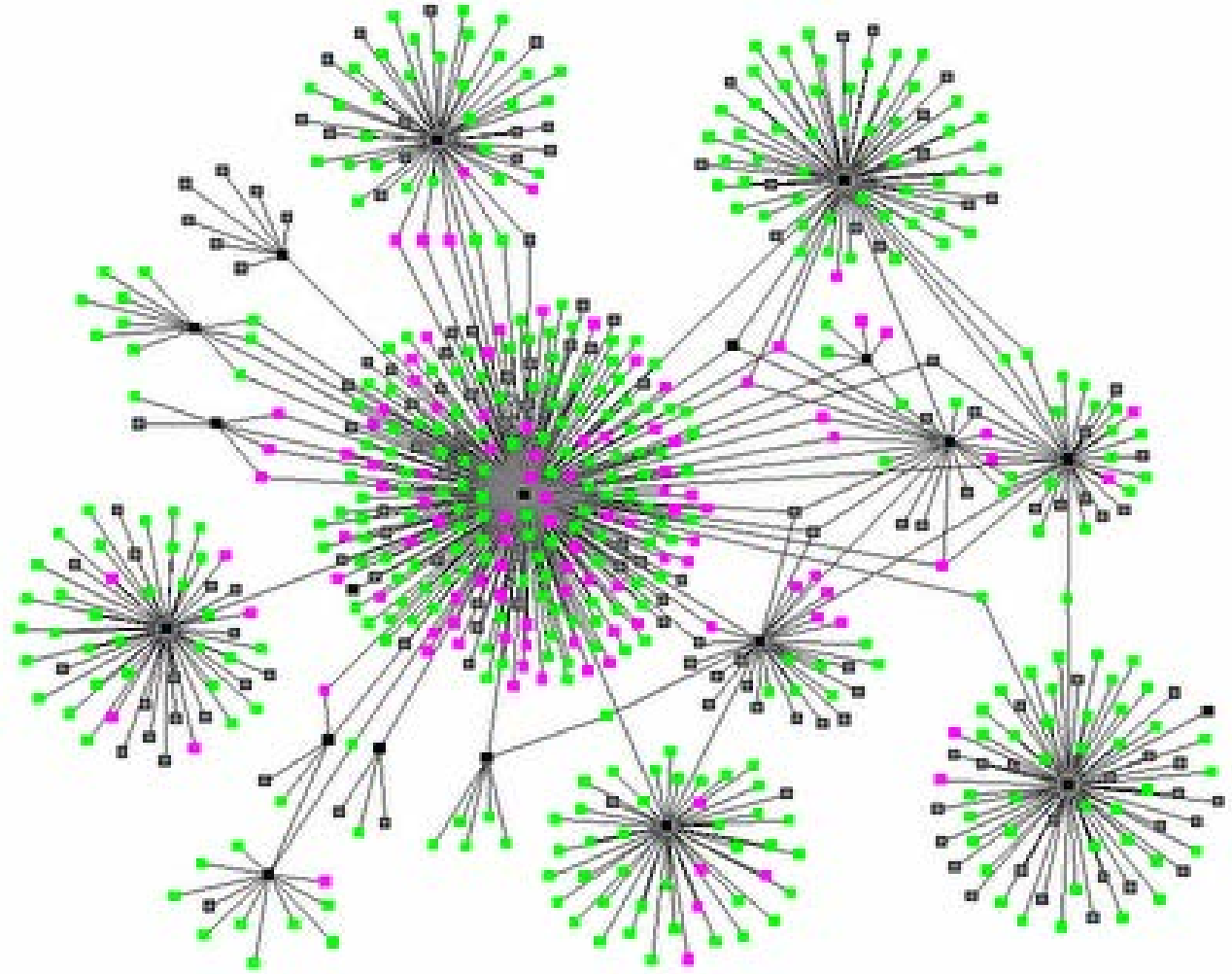
Text Data - Transformers



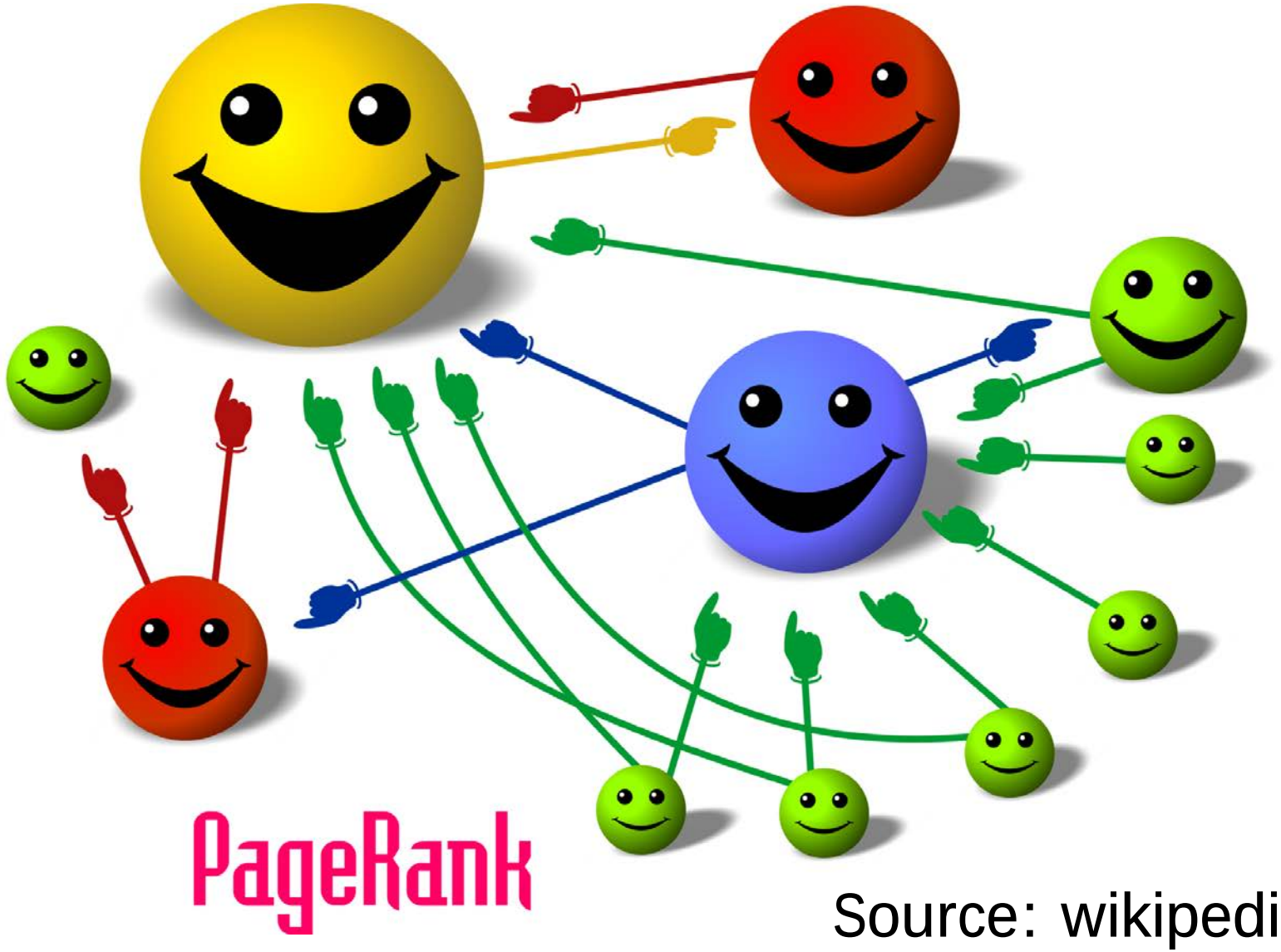
Part III: Graphs/Network mining (~2.5 weeks)

- Spectral clustering
- Label propagation
- Graph/Network shallow embedding
- Knowledge graph embedding
- Graph Neural Networks

Graph / Network

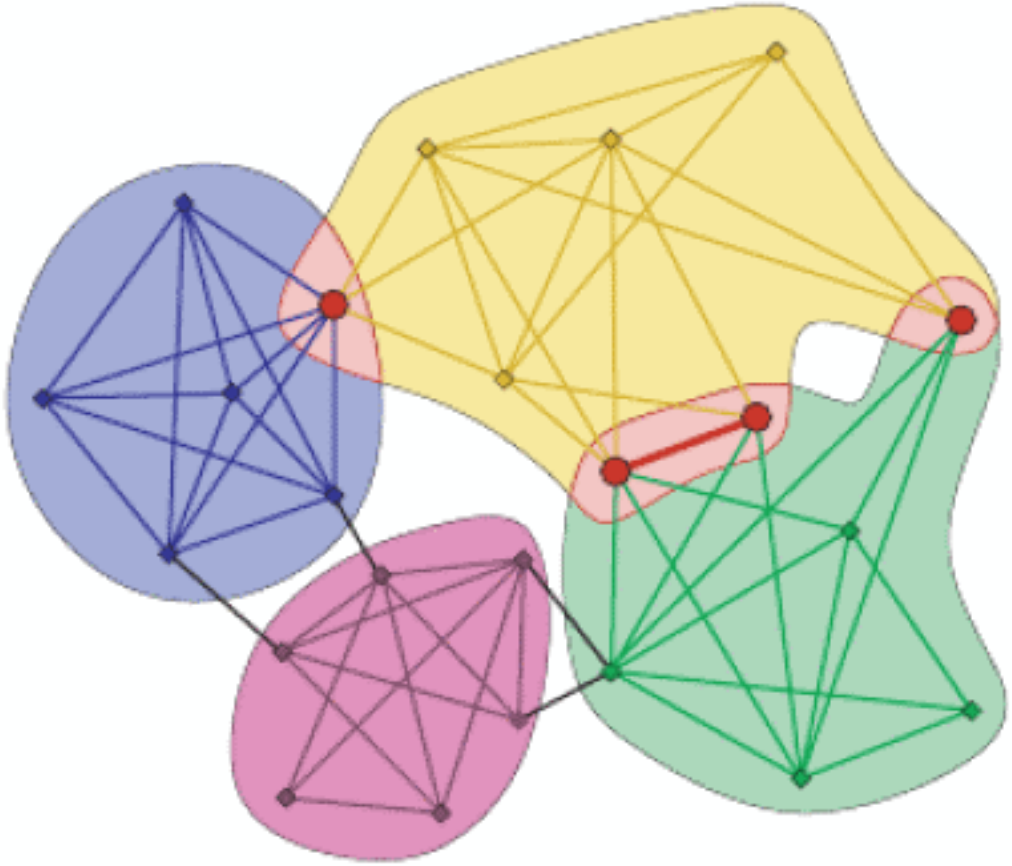


PageRank

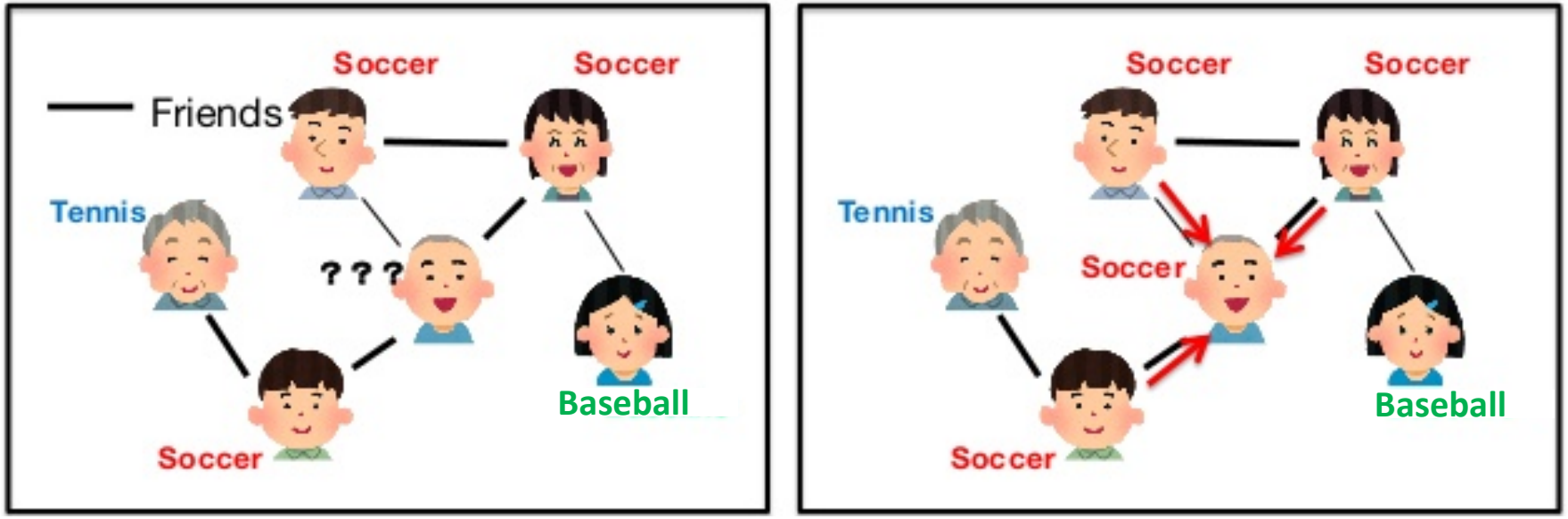


Source: wikipedia

Community Detection



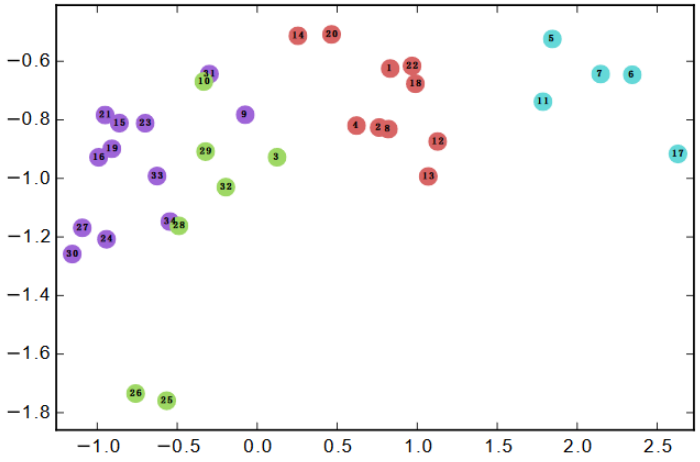
Label Propagation



Source: https://www.slideshare.net/yamaguchiyuto/when-does-label-propagation-fail-a-view-from-a-network-generative-model

Network Embedding

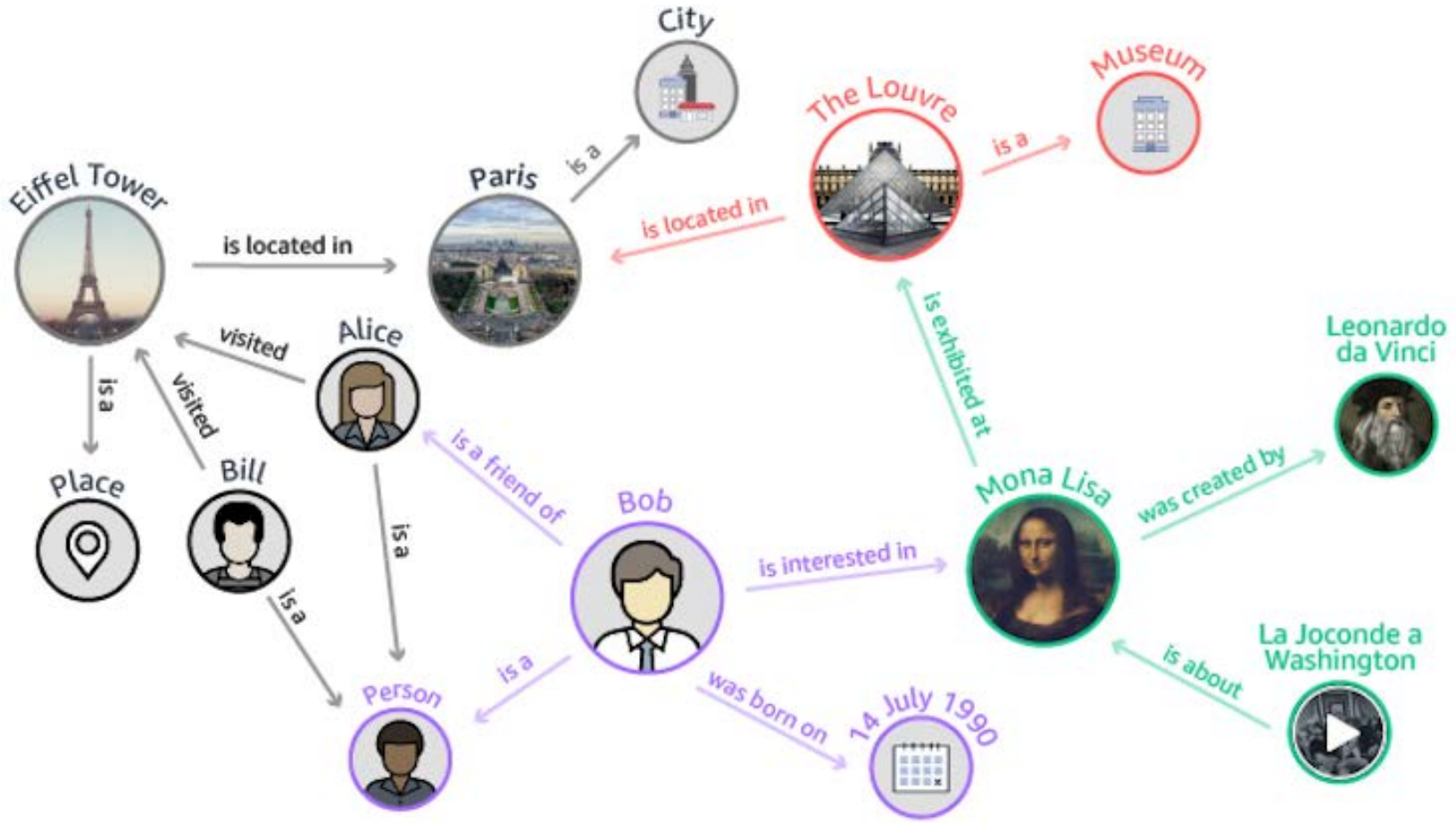
(a) Input: karate network



(b) Output: representations

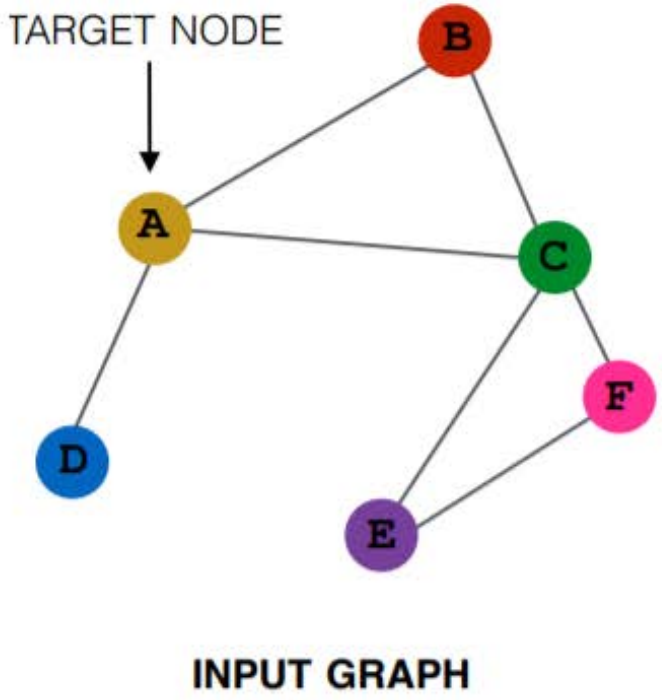
Source: DeepWalk

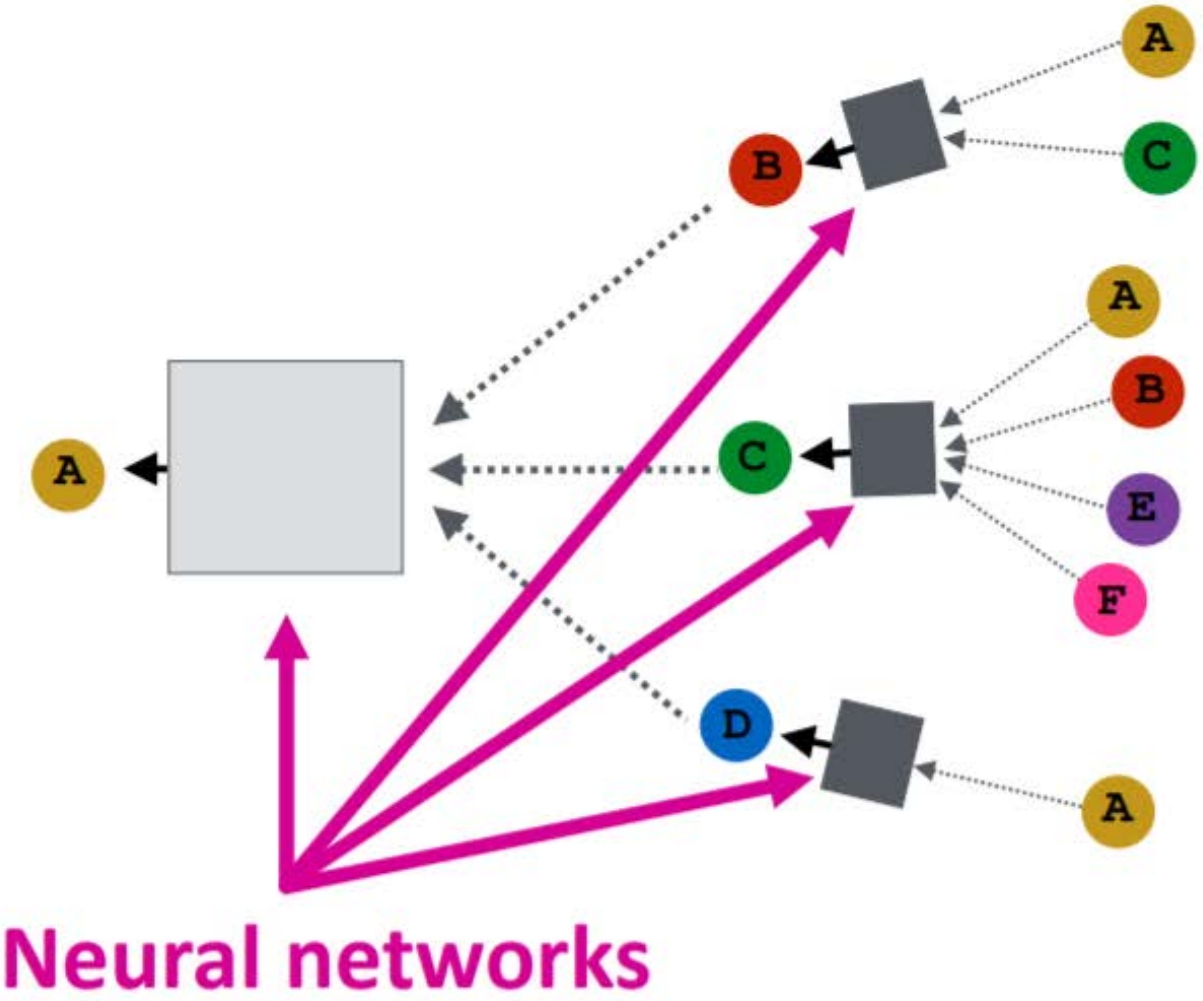
Knowledge Graph



https://geomarketing.com/amazons-neptune-database-will-expand-the-knowledge-graph-heres-how/amazon-neptune-knowledge-graph

Graph Neural Networks





Part IV: Recommender systems (~2 weeks)

- Collaborative filtering, matrix factorization, Bayesian personalized ranking
- Factorization machine and deep collaborative filtering
- Recommendation from graph perspective

Recommender systems

You may also like

Jack & Jones
JAMIE - Polo shirt - orange
£21.00
Free delivery & returns

ALTERNATIVE PRODUCTS

Beko Washing Machine
Code: WMB81431LW
£269.99

Zanussi Washing Machine
Code: ZWH6130P
£269.99

Blomberg Washing Machine
Code: WNF6221
£299.99

You may also like

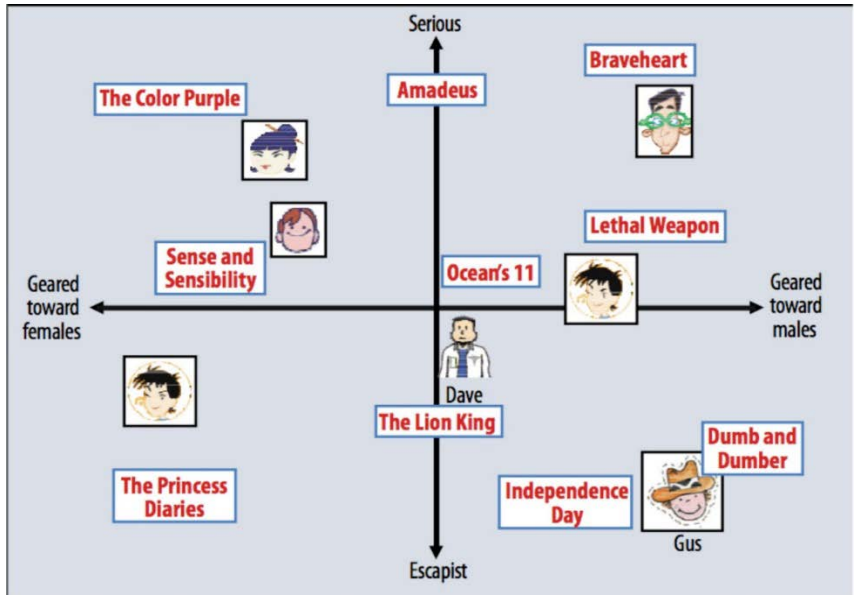
★★★★☆ (109)

★★★★½ (53)

★★★½☆ (33)

A Matrix Representation View

Users	Movie1	Movie2	Movie3	Movie4	Movie5	Movie6	...
User1	?	?	4	?	1	?	...
User2	2	5	2	?	?	2	...
User3	?	?	5	3	2	4	...
User4	1	?	?	4	?	?	...
User5	2	3	?	?	?	?	...
...	...	...	...	...	...	...	...



Summary: Methods to Learn

- Part I: Basics of data mining (~2 weeks)
  - Naïve Bayes, Logistic regression, K-Means, mixture models, ANN
- Part II: Text mining (~ 2 weeks)
  - Topic Models: pLSA, LDA
  - Word embedding
  - Transformers
- Part III: Graphs/Network mining (~2.5 weeks)
  - spectral clustering, graph embedding, label propagation, knowledge graph embedding, GNNs
- Part IV: Recommender systems (~ 2 weeks)
  - Collaborative filtering, matrix factorization, Bayesian personalized ranking
  - Factorization machine, neural collaborative filtering
  - Recommendation as link prediction

Where to Find References? DBLP, CiteSeer, Google

- Data mining and KDD (SIGKDD: CDROM)
  - Conferences: ACM-SIGKDD, IEEE-ICDM, SIAM-DM, PKDD, PAKDD, etc.
  - Journal: Data Mining and Knowledge Discovery, KDD Explorations, ACM TKDD
- Database systems (SIGMOD: ACM SIGMOD Anthology—CD ROM)
  - 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: AAAI, IJCAI, COLT (Learning Theory), CVPR, ICML, NIPS, ICLR etc.
  - Journals: Machine Learning, Artificial Intelligence, Knowledge and Information Systems, IEEE-PAMI, etc.
- NLP
  - ACL, EMNLP, NAACL, etc.
- 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.

Textbook

- Recommended: Jiawei Han, Micheline Kamber, and Jian Pei. Data Mining: Concepts and Techniques, 3rd edition, Morgan Kaufmann, 2011
- References
  - "Data Mining: The Textbook" by Charu Aggarwal (http://www.charuaggarwal.net/Data-Mining.htm)
  - “Neural Networks and Deep Learning” (https://www.springer.com/us/book/9783319944623)
  - "Data Mining" by Pang-Ning Tan, Michael Steinbach, and Vipin Kumar (http://www-users.cs.umn.edu/~kumar/dmbook/index.php)
  - "Machine Learning" by Tom Mitchell (http://www.cs.cmu.edu/~tom/mlbook.html)
  - "Introduction to Machine Learning" by Ethem ALPAYDIN (http://www.cmpe.boun.edu.tr/~ethem/i2ml/)
  - "Pattern Classification" by Richard O. Duda, Peter E. Hart, David G. Stork (http://www.wiley.com/WileyCDA/WileyTitle/productCd-0471056693.html)
  - "The Elements of Statistical Learning: Data Mining, Inference, and Prediction" by Trevor Hastie, Robert Tibshirani, and Jerome Friedman (http://www-stat.stanford.edu/~tibs/ElemStatLearn/)
  - "Pattern Recognition and Machine Learning" by Christopher M. Bishop (http://research.microsoft.com/en-us/um/people/cmbishop/prml/)

Q&A