DSC250: Advanced Data Mining

Overview

Zhiting Hu

Lecture 1, Jan 7, 2025

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

Logistics

- Class webpage: http://zhiting.ucsd.edu/teaching/dsc250winter2025

Logistics

Instructor: Zhiting Hu

Email: zhh019@ucsd.edu

Office hours: Thursday 1-2pm

Location: HDSI 442

TA: Yi Gu

Email: yig025@ucsd.edu

Office hours: Mon 2-3pm

Location: TBA

- Discussion forum: Piazza
- Homework & writeup submission: Gradescope

Logistics: grading

- 2 Homework assignments (30% of grade)
- Paper presentation (20%)
- Course project (46%)
- Participation (4%)

Logistics: grading

- 2 Homework assignments (30% of grade)
  - Theory exercises
  - 3 total late days without penalty
- Paper presentation (20%)
- Course project (46%)
- Participation (4%)

Logistics: grading

- 2 Homework assignments (30% of grade)
- Paper presentation (20%)
  - Each **student or pair** will give an oral presentation on a research paper
    - 10 mins = 8 mins presentation + 2 mins QA (*tentative*)
      - Timing -- hard time constraint: if you run over the expected time limit (8min), there will be no QA session for your presentation, and thus no credits for the QA component
    - **Critical thinking:** discuss both strengths and limitations of the paper
    - Sign up in a google sheet (TBA)
    - Design quiz questions for audience
  - **Peer grading:** other students will rate and give feedback (5% of grade)
  - Starting later part of the quarter, after the class size is stabilized
- Course project (46%)
- Participation (4%)

Depending on #enrollments

Logistics: grading

- 2 Homework assignments (30% of grade)
- Paper presentation (20%)
- Course project (46%)
  - 3 or 4-member **team** to be formed and sign up in a google sheet (TBA)
  - Designed to be as similar as possible to researching and writing a **conference-style paper**:
    - Due to tight timeline, fine to use synthetic/toy data for proof-of-concept experiments + explanation of theory/intuition of why your approach is likely to work
  - **Proposal** : 2 pages excluding references (10%) -- due in 2 or 3 weeks (TBA)
    - Overview of project idea, literature review, potential datasets and evaluation, milestones
  - **Midway Report** : 4-5 pages excluding references (20%)
  - **Presentation** : oral presentation, 15-20mins (20%)
    - Peer grading (5%)
  - **Final Report** : 6-8 pages excluding references (50%)

Logistics: grading

- 2 Homework assignments (30% of grade)
- Paper presentation (20%)
- Course project (46%)
- Participation (4%)
  - Submission of quiz answers and feedback on paper/project presentations
  - Contribution to discussion on Piazza
  - Completion of final course evaluation
  - Any constructive suggestions

Logistics: grading

- 2 Homework assignments (30% of grade)
- Paper presentation (20%)
- Course project (46%)
- Participation (4%)

Logistics: grading

- 2 Homework assignments (30% of grade)
  - Theory exercises, implementation exercises
  - 3 total late days without penalty
- Paper presentation (20%)
- Course project (46%)
- Participation (4%)

Logistics: grading

- 2 Homework assignments (30% of grade)
- Paper presentation (20%)
  - Each student will give an oral presentation on a research paper
  - 10 mins = 8 mins presentation + 2 mins QA *(tentative)*
  - Discuss both strengths and limitations of the paper
  - Sign up in a google sheet (TBA)
  - Starting 2nd half of the quarter
- Course project (46%)
- Participation (4%)

Logistics: grading

- 2 Homework assignments (30% of grade)
- Paper presentation (20%)
- Course project (46%)
  - 3 or 4-member team to be formed and sign up in a google sheet (TBA)
  - Designed to be as similar as possible to researching and writing a conference-style paper:
    - Due to tight timeline, fine to use synthetic/toy data for proof-of-concept experiments + explanation of theory/intuition of why your approach is likely to work
  - **Proposal** : 2 pages excluding references (10%) -- Due in 2 or 3 weeks (TBA)
    - Overview of project idea, literature review, potential datasets and evaluation, milestones
  - **Midway Report** : 4-5 pages excluding references (20%)
  - **Presentation** : oral presentation, 15-20mins (20%)
  - **Final Report** : 6-8 pages excluding references (50%)

Logistics: grading

- 2 Homework assignments (30% of grade)
- Paper presentation (20%)
- Course project (46%)
- Participation (4%)
  - Contribution to discussion on Piazza
  - Complete mid-quarter evaluation
  - Any constructive suggestions

Data Mining

Why Data Mining

- The Explosive Growth of Data: from terabytes to petabytes

Why Data Mining

- The Explosive Growth of Data: from terabytes to petabytes
  - Data collection and data availability
    - Automated data collection tools, database systems, Web, computerized society

> - Facebook: one billion images uploaded per day
> - 300 hours of video are uploaded to YouTube every minute

Why Data Mining

- The Explosive Growth of Data: from terabytes to petabytes
  - Data collection and data availability
    - Automated data collection tools, database systems, Web, computerized society
  - Major sources of abundant data
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: Remote sensing, bioinformatics, scientific simulation, …
    - Society and everyone: news, digital cameras, YouTube

Why Data Mining

- The Explosive Growth of Data: from terabytes to petabytes
  - Data collection and data availability
    - Automated data collection tools, database systems, Web, computerized society
  - Major sources of abundant data
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: Remote sensing, bioinformatics, scientific simulation, …
    - Society and everyone: news, digital cameras, YouTube
- We are drowning in data, but starving for knowledge!
- Data Mining: Automated analysis of massive datasets

Evolution of Sciences

- Before 1600, **empirical science**
- 1600-1950s, **theoretical science**
  - Each discipline has grown a *theoretical* component. Theoretical models often motivate experiments and generalize our understanding.
- 1950s-1990s, **computational science**
  - Over the last 50 years, most disciplines have grown a third, *computational* branch
    - e.g. empirical/theoretical/computational ecology, or physics, or linguistics.
  - Computational Science traditionally meant simulation. It grew out of our inability to find closed-form solutions for complex mathematical models.
- 1990-now, **data science**
  - **Mountains of data from several converging trends:**
    - The flood of data from new scientific instruments and simulations
    - The ability to economically store and manage petabytes of data online
    - The Internet and computing Grid that makes all these archives universally accessible

Jim Gray and Alex Szalay, *The World Wide Telescope: An Archetype for Online Science*, Comm. ACM, 45(11): 50-54, Nov. 2002

Han, Kamber, and Pei, *Data Mining: Concepts and Techniques* 3rd edition

What is Data Mining

- Data mining (knowledge discovery from data; KDD)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data

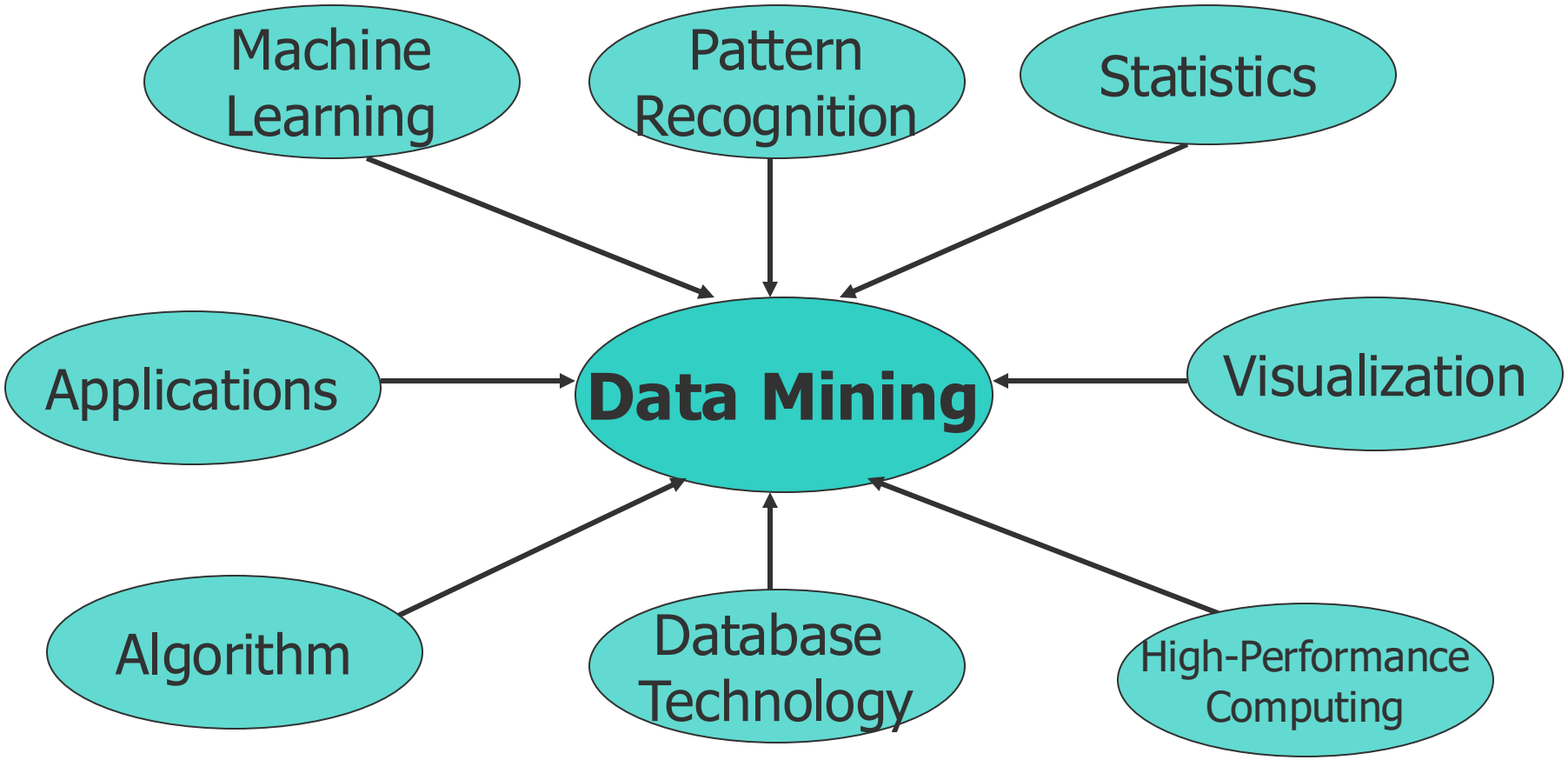
What is Data Mining

- Data mining (knowledge discovery from data; KDD)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
- 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

Question: *What's the difference between Data Mining vs Machine Learning?*

Confluence of Multiple Disciplines

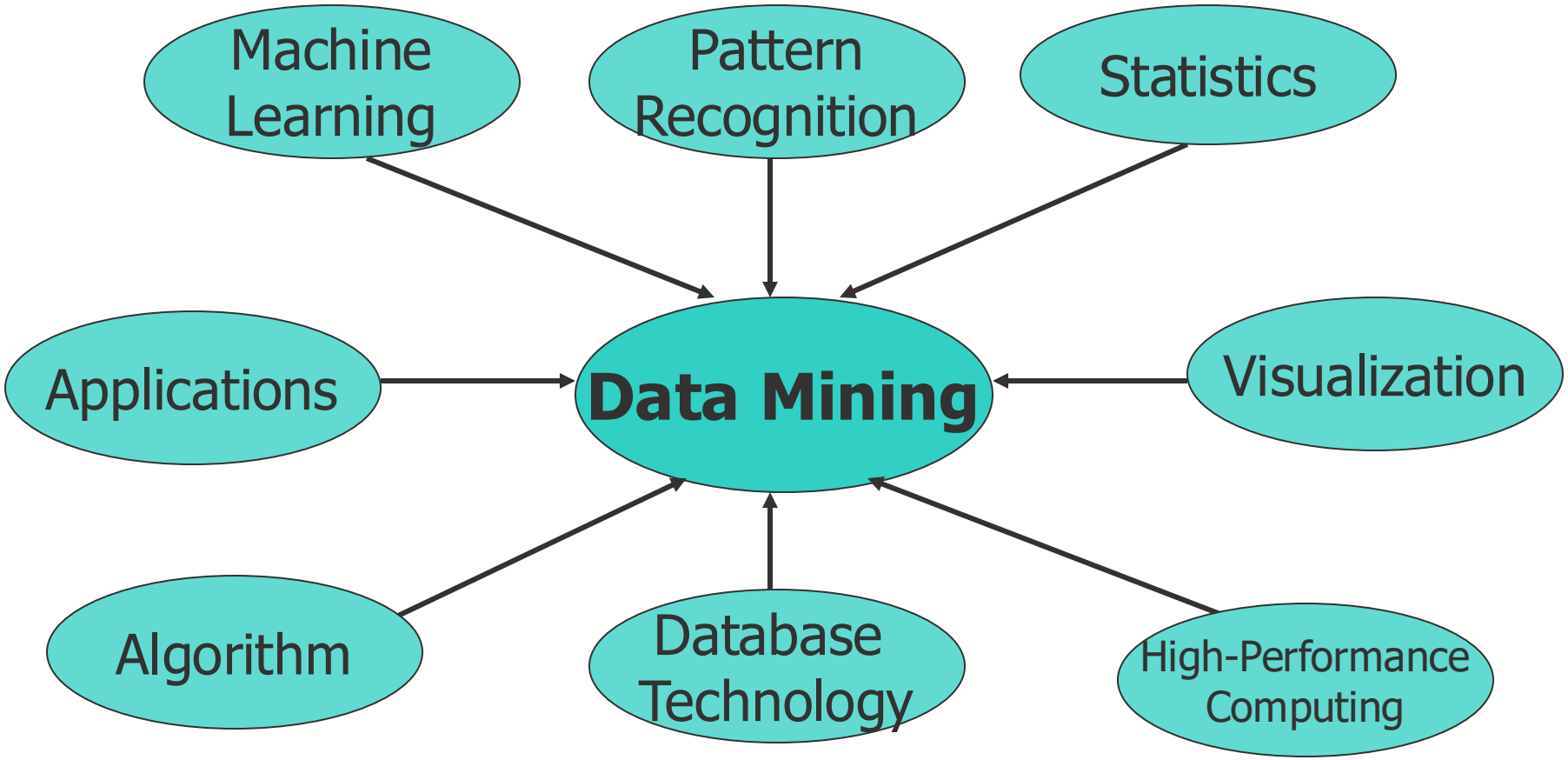


What is Data Mining

The age of pursuing "AGI": the boundary between these fields is getting blurred…

E.g., ***Question:*** *Is training LLMs part of Data Mining?*

Confluence of Multiple Disciplines

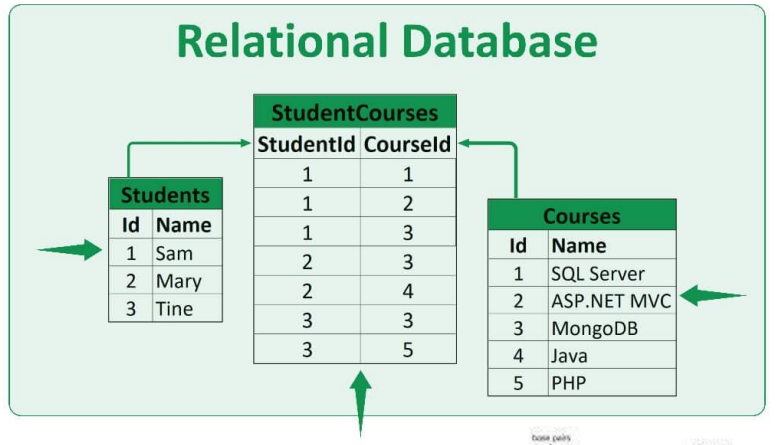


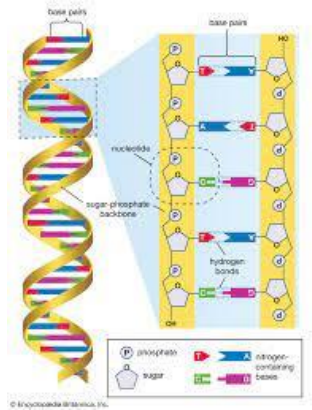
Different Dimensions of Data Mining

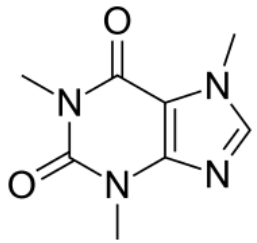
- **Data to be mined**
  - Database data (extended-relational, object-oriented, heterogeneous, legacy), data warehouse, transactional data, stream, spatiotemporal, time-series, sequence, text and web, multi-media, graphs & social and information networks
- **Knowledge to be mined (or: Data mining functions)**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

Data to be mined

- Database-oriented data sets and applications
  - Relational database, data warehouse, transactional database
- Advanced data sets and advanced applications
  - Text databases
  - Structure data, graphs, social networks and multi-linked data
  - Time-series data, temporal data, sequence data (incl. bio-sequences)
  - Data streams and sensor data
  - Heterogeneous databases and legacy databases
  - Spatial data and spatiotemporal data
  - Multimedia database
  - The World-Wide Web







Knowledge to be mined (i.e., data mining functions)

(Ex-1) Association and Correlation Analysis

- Frequent patterns (or frequent itemsets)
  - What items are frequently purchased together in your Walmart?
- Association, correlation vs. causality
  - A typical association rule
    - Diaper → Beer [0.5%, 75%] (support, confidence)
- How to mine such patterns and rules efficiently in large datasets?
- How to use such patterns for classification, clustering, and other applications?

Knowledge to be mined (i.e., data mining functions)

(Ex-2) Classification

- Classification and label prediction
  - Construct models (functions) based on some training examples
  - Describe and distinguish classes or concepts for future prediction
    - E.g., classify countries based on (climate), or classify cars based on (gas mileage)
  - Predict some unknown class labels
- Typical methods
  - Decision trees, naïve Bayesian classification, support vector machines, neural networks, rule-based classification, pattern-based classification, logistic regression, …
- Typical applications:
  - Credit card fraud detection, direct marketing, classifying stars, diseases, web-pages, …

Knowledge to be mined (i.e., data mining functions)

(Ex-3) Cluster Analysis

- Unsupervised learning (i.e., Class label is unknown)
- Group data to form new categories (i.e., clusters), e.g., cluster houses to find distribution patterns
- Principle: Maximizing intra-class similarity & minimizing interclass similarity
- Many methods and applications

Knowledge to be mined (i.e., data mining functions)

(Ex-4) Outlier Analysis

- Outlier: A data object that does not comply with the general behavior of the data
- Noise or exception? — One person's garbage could be another person's treasure
- Methods: by product of clustering or regression analysis, …
- Useful in fraud detection, rare events analysis

Knowledge to be mined (i.e., data mining functions)

(Ex-5) Time and Ordering: Sequential Pattern, Trend and Evolution Analysis

- Sequence, trend and evolution analysis
  - Trend, time-series, and deviation analysis: e.g., regression and value prediction
  - Sequential pattern mining
    - e.g., first buy digital camera, then buy large SD memory cards
  - Periodicity analysis
  - Motifs and biological sequence analysis
    - Approximate and consecutive motifs
  - Similarity-based analysis
- Mining data streams
  - Ordered, time-varying, potentially infinite, data streams

Knowledge to be mined (i.e., data mining functions)

(Ex-6) Structure and Network Analysis

- Graph mining
  - Finding frequent subgraphs (e.g., chemical compounds), trees (XML), substructures (web fragments)
- Information network analysis
  - Social networks: actors (objects, nodes) and relationships (edges)
    - e.g., author networks in CS, terrorist networks
  - Multiple heterogeneous networks
    - A person could be multiple information networks: friends, family, classmates, …
  - Links carry a lot of semantic information: Link mining
- Web mining
  - Web is a big information network: from PageRank to Google
  - Analysis of Web information networks
    - Web community discovery, opinion mining, usage mining, …

This Course

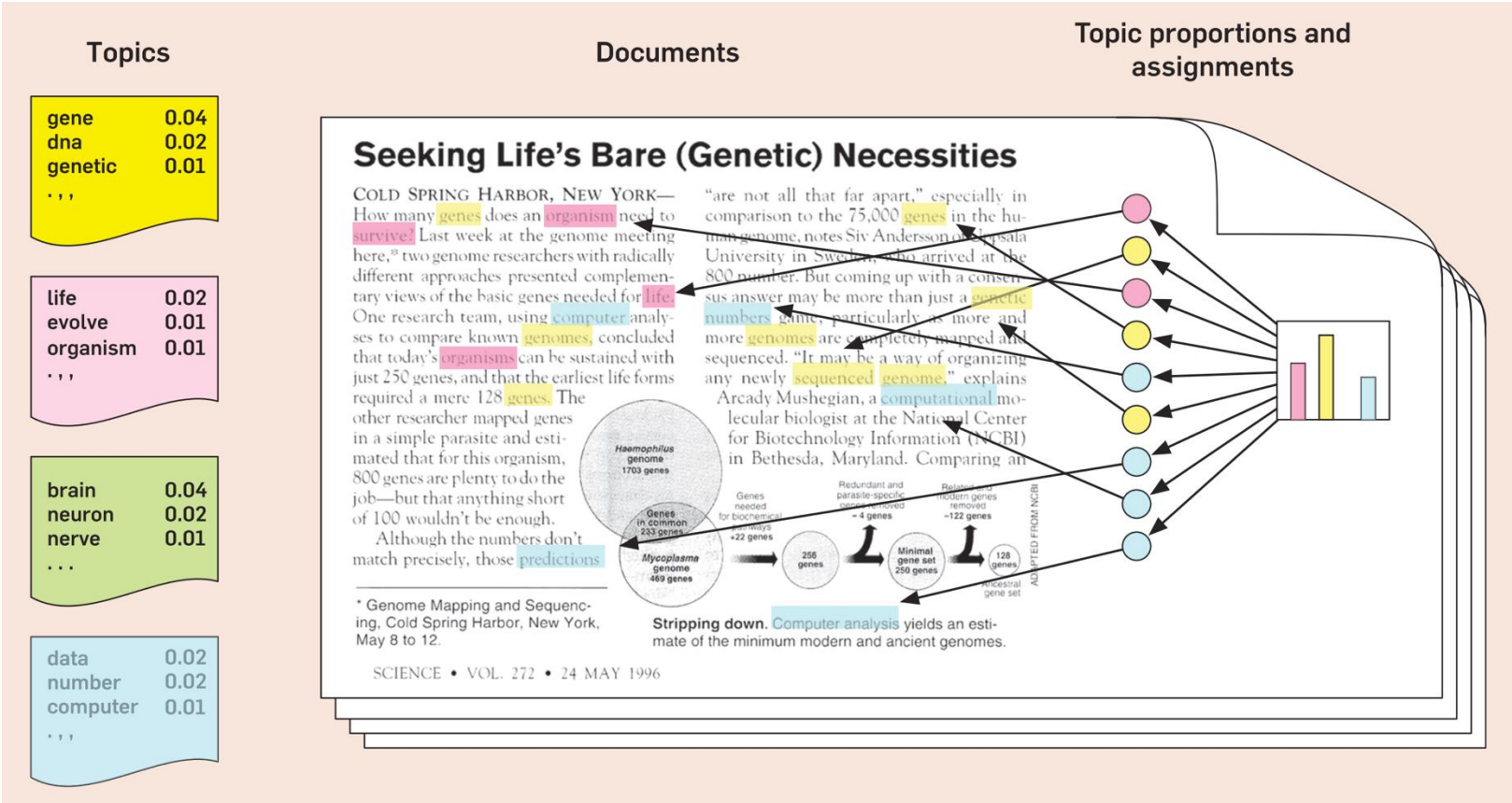
1) Text (multi-modal) mining

2) Graph/network mining

3) Recommender systems

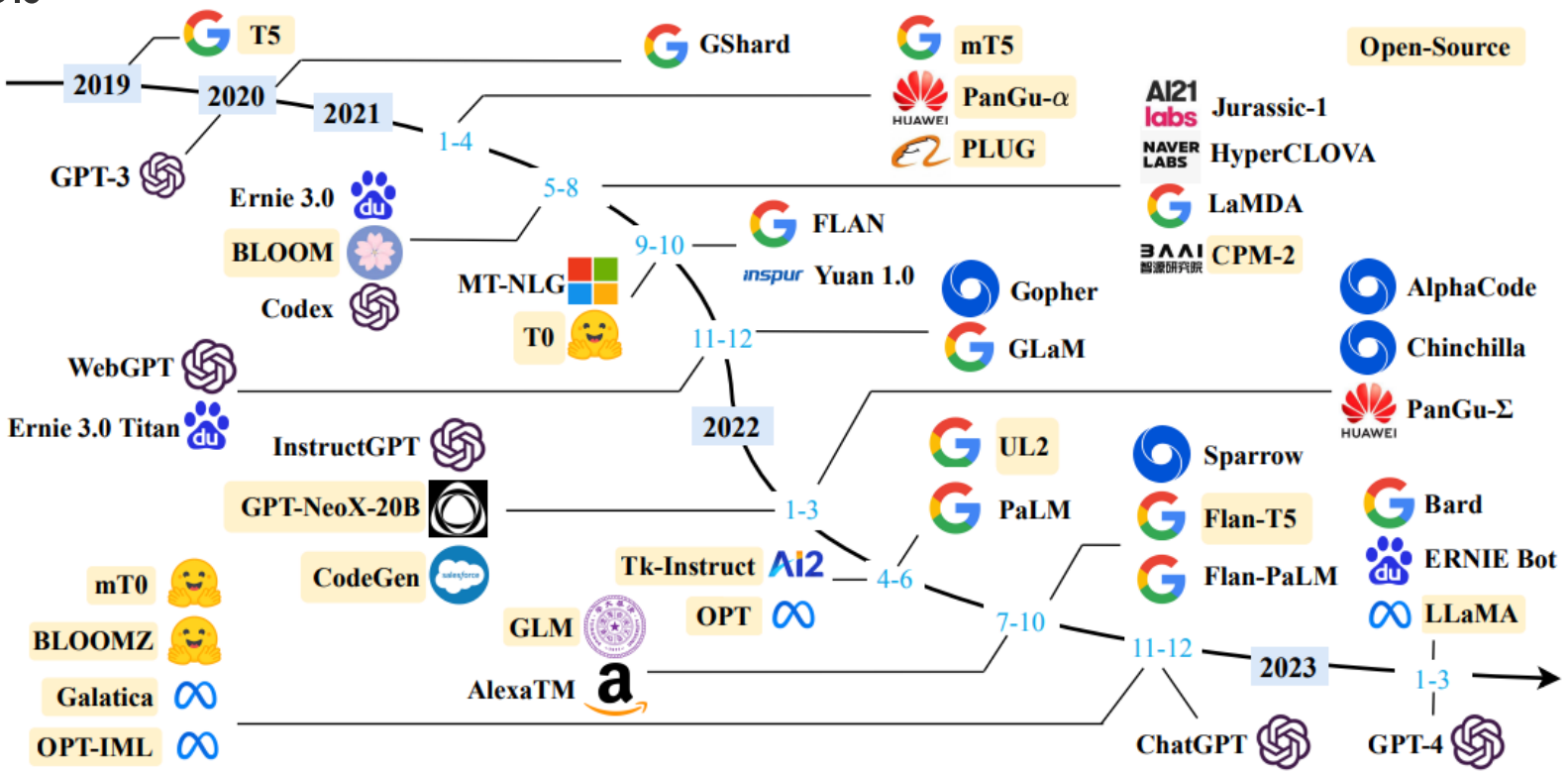
This Course

1) Text (multi-modal) mining
   - Topic models
     - LDA, Expectation Maximization, variational inference



This Course

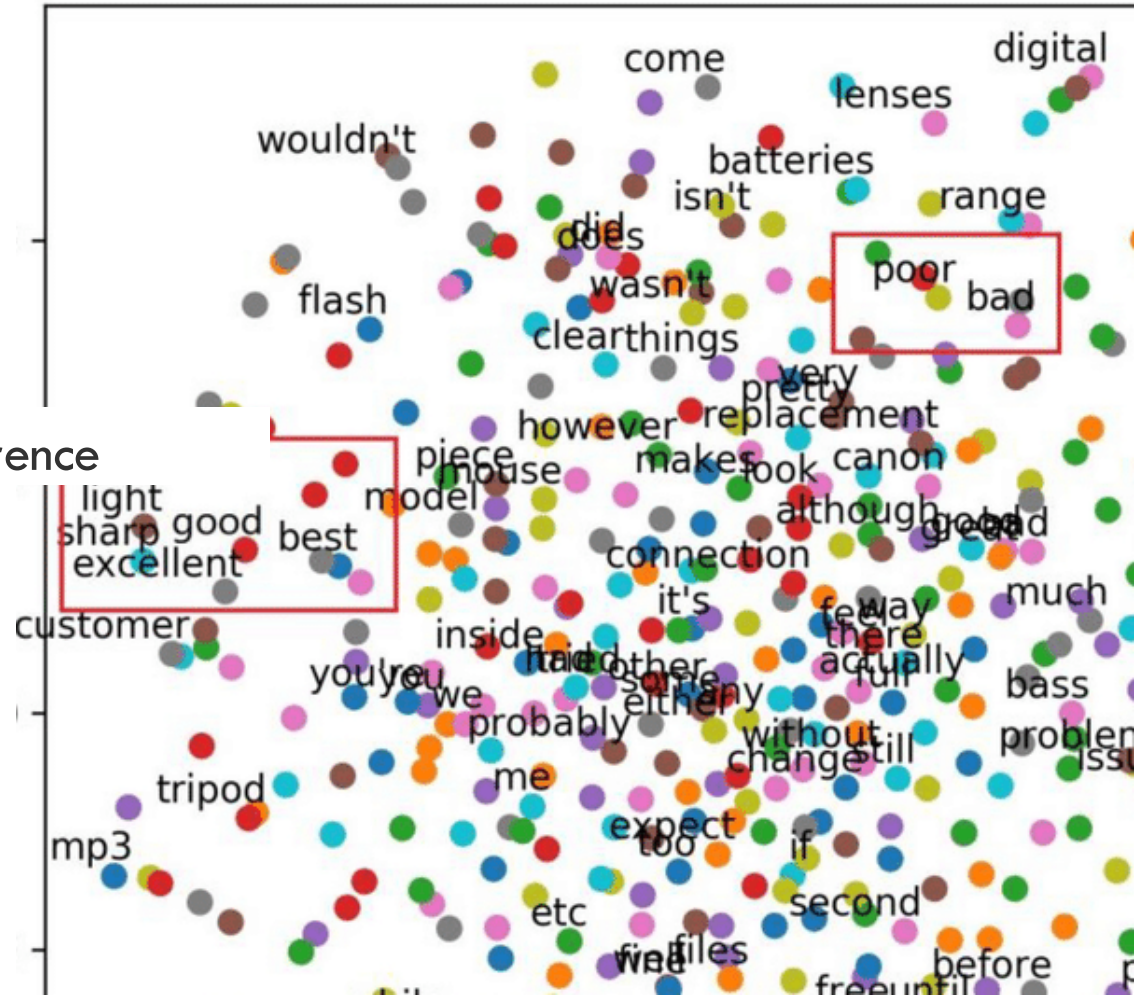
1) Text (multi-modal) mining
   - Topic models
     - LDA, Expectation Maximization, variational inference
   - Language models



https://medium.com/@thedatabeast/top-free-courses-on-large-language-models-abf2722d15c5

This Course

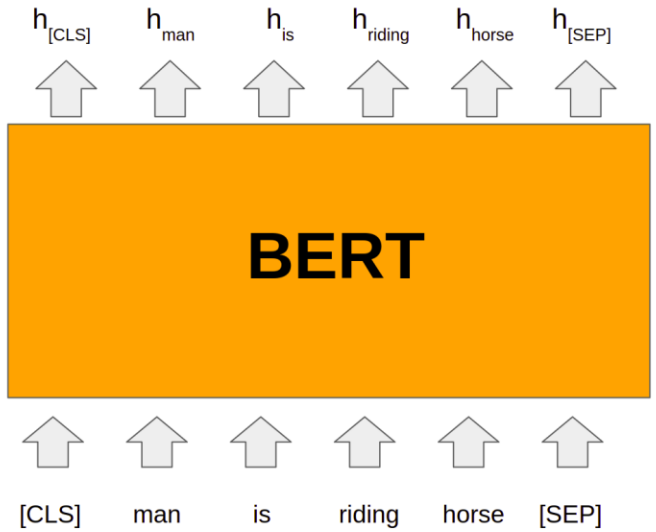
1) Text (multi-modal) mining
   - Topic models
     - LDA, Expectation Maximization, variational inference
   - Language models
   - Text representation learning (embedding)



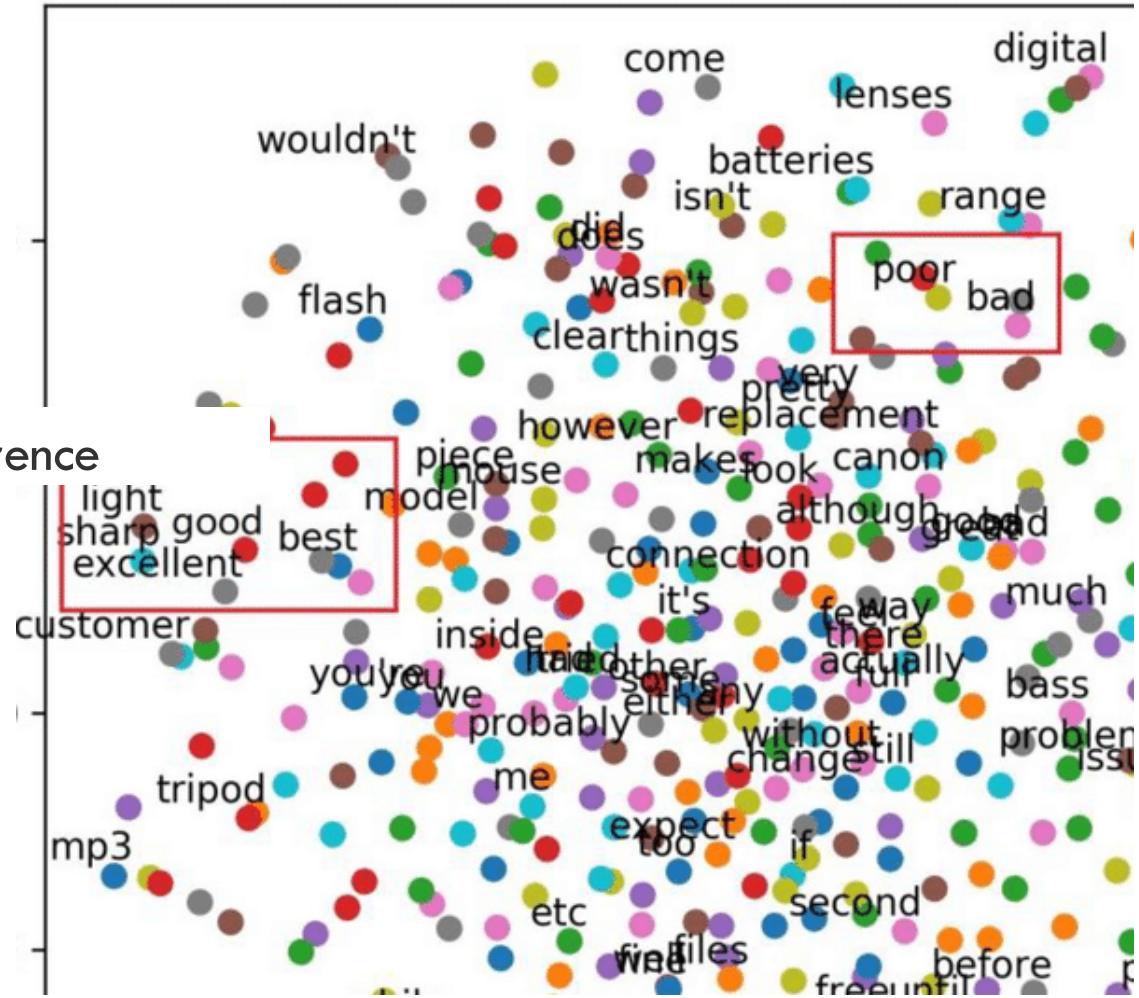
Word embedding

This Course

1) Text (multi-modal) mining
   - Topic models
     - LDA, Expectation Maximization, variational inference
   - Language models
   - Text representation learning (embedding)



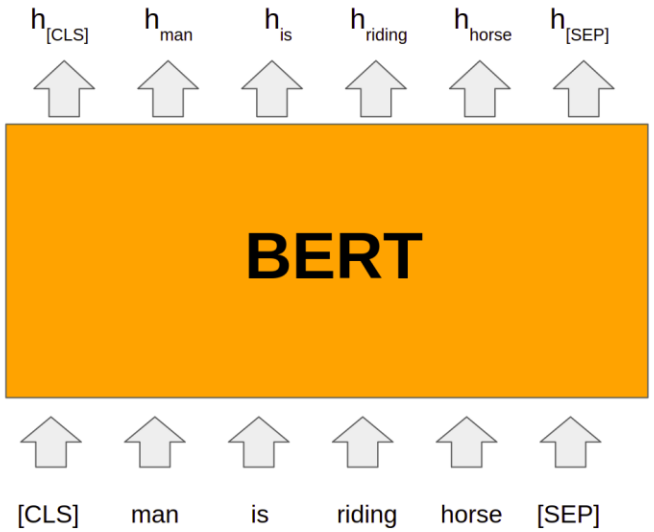
Contextualized embedding



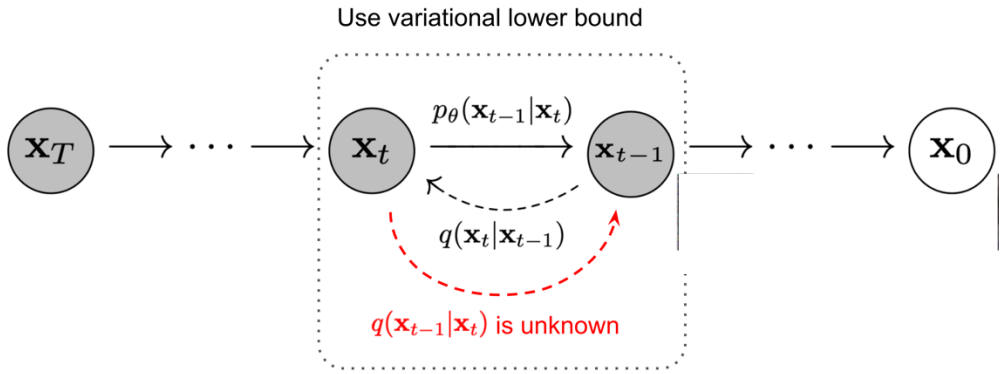
Word embedding

This Course

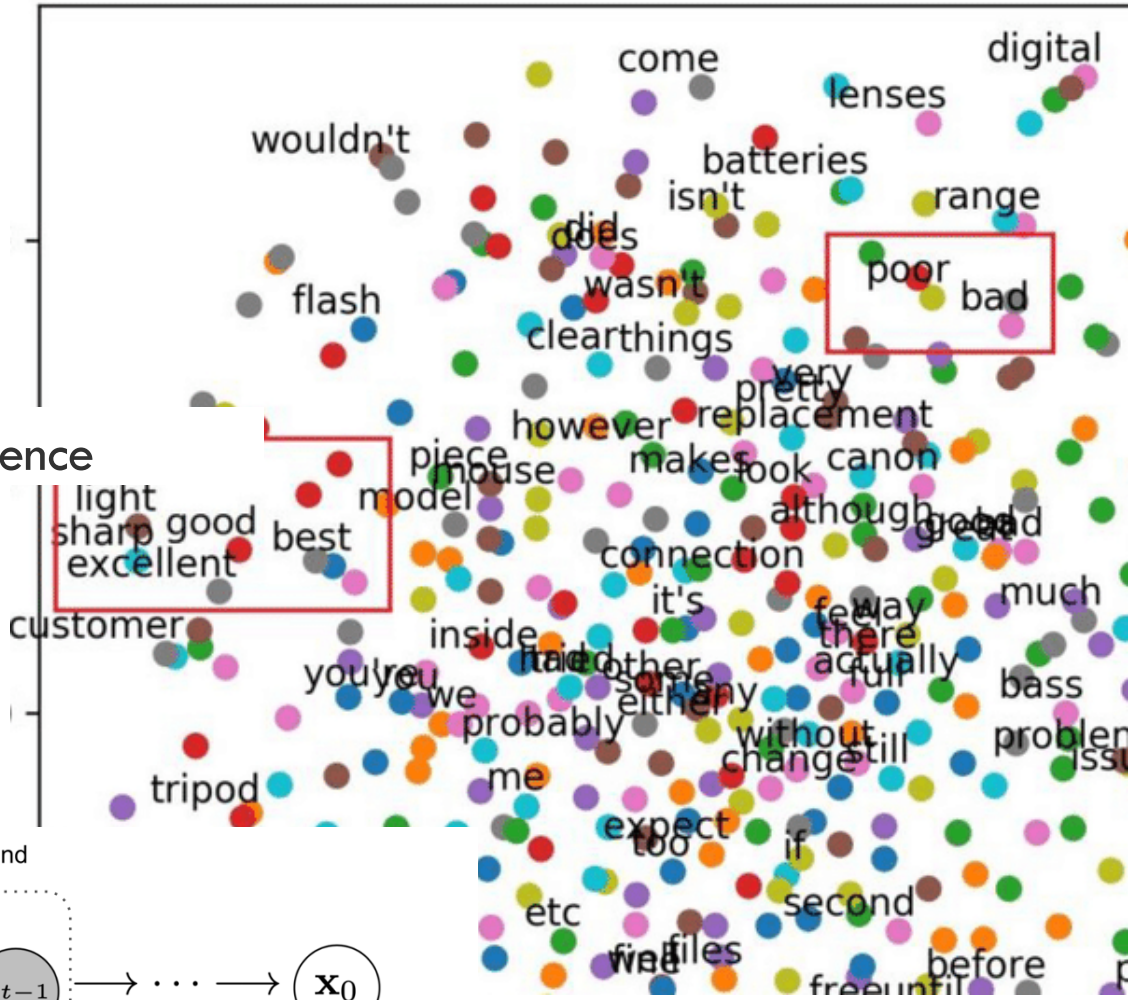
1) Text (multi-modal) mining
   - Topic models
     - LDA, Expectation Maximization, variational inference
   - Language models
   - Text representation learning (embedding)



Contextualized embedding



More advanced methods: e.g., diffusion model

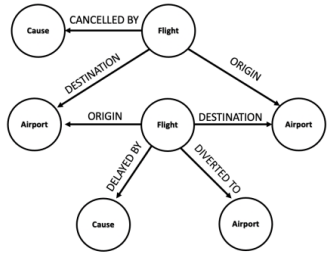


Word embedding

This Course

2) Graph/network mining

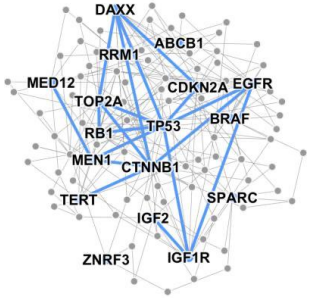
Graphs are a general language for describing and analyzing entities with relations/interactions



Event Graphs

Image credit: SalientNetworks

Computer Networks



Disease Pathways

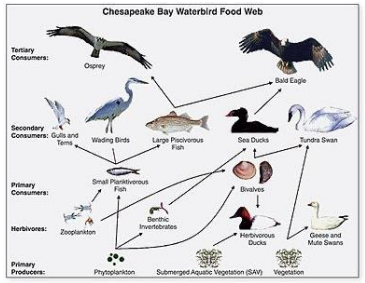


Image credit: Wikipedia

Food Webs

Image credit: Pinterest

Particle Networks

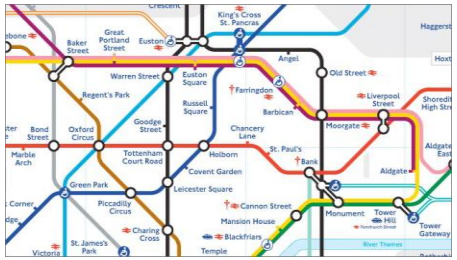


Image credit: visitlondon.com

Underground Networks

[Jure Leskovec, Stanford CS224W: Machine Learning with Graphs]

This Course

2) Graph/network mining

Graphs are a general language for describing and analyzing entities with relations/interactions

Image credit: Medium

Social Networks

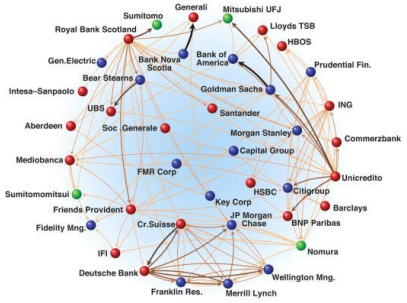


Image credit: Science

Economic Networks

Image credit: Lumen Learning

Communication Networks

Citation Networks

Image credit: Missoula Current News

Internet

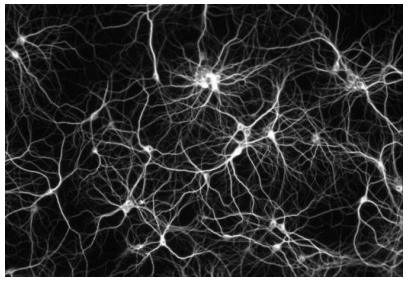


Image credit: The Conversation

Networks of Neurons

[Jure Leskovec, Stanford CS224W: Machine Learning with Graphs]

This Course

2) Graph/network mining

Graphs are a general language for describing and analyzing entities with relations/interactions

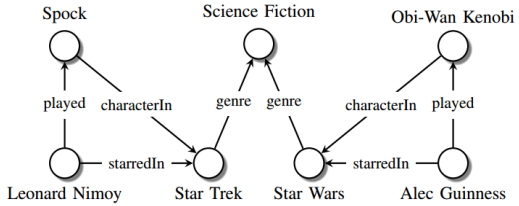


Image credit: Maximilian Nickel et al

Knowledge Graphs

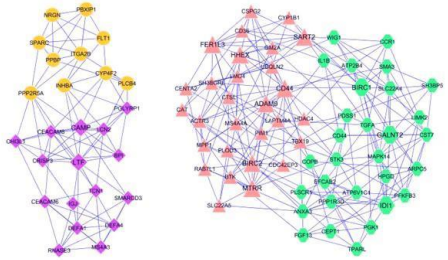


Image credit: ese.wustl.edu

Regulatory Networks

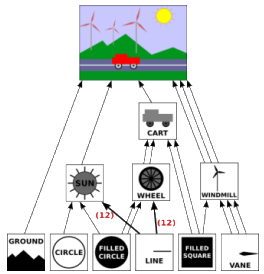


Image credit: math.hws.edu

Scene Graphs

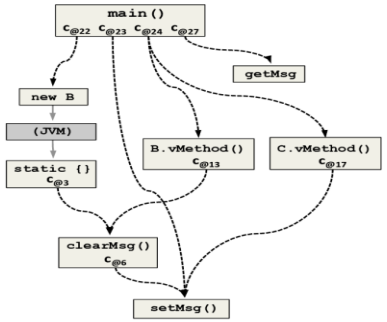


Image credit: ResearchGate

Code Graphs

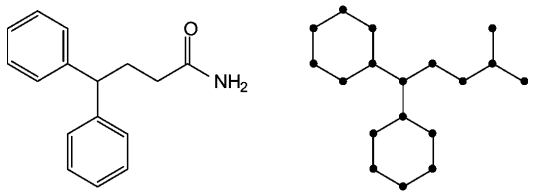


Image credit: MDPI

Molecules

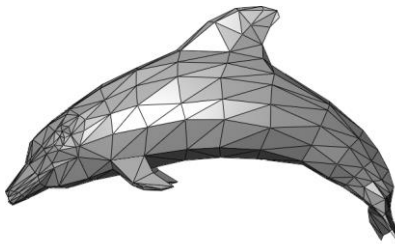


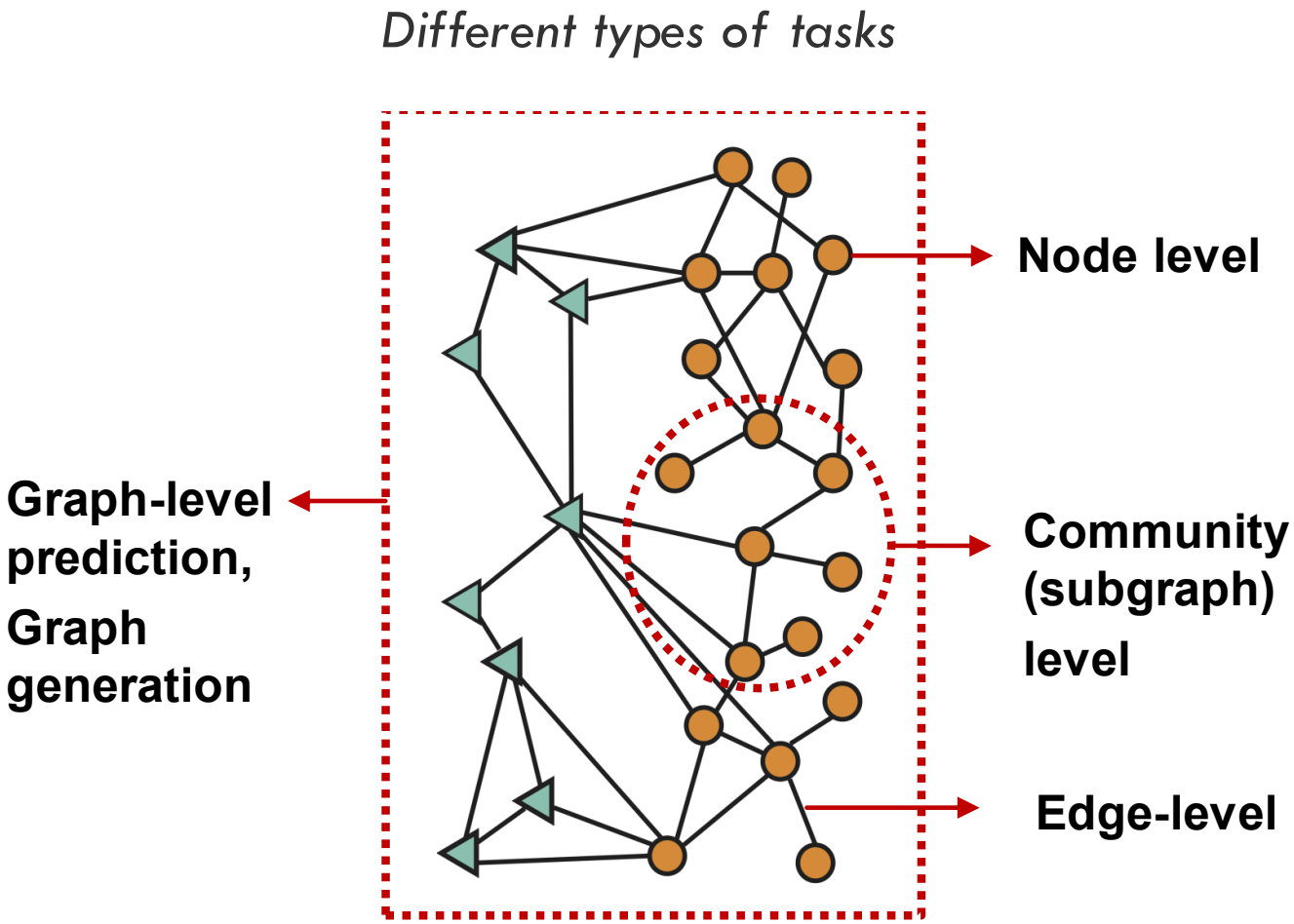
Image credit: Wikipedia

3D Shapes

[Jure Leskovec, Stanford CS224W: Machine Learning with Graphs]

This Course

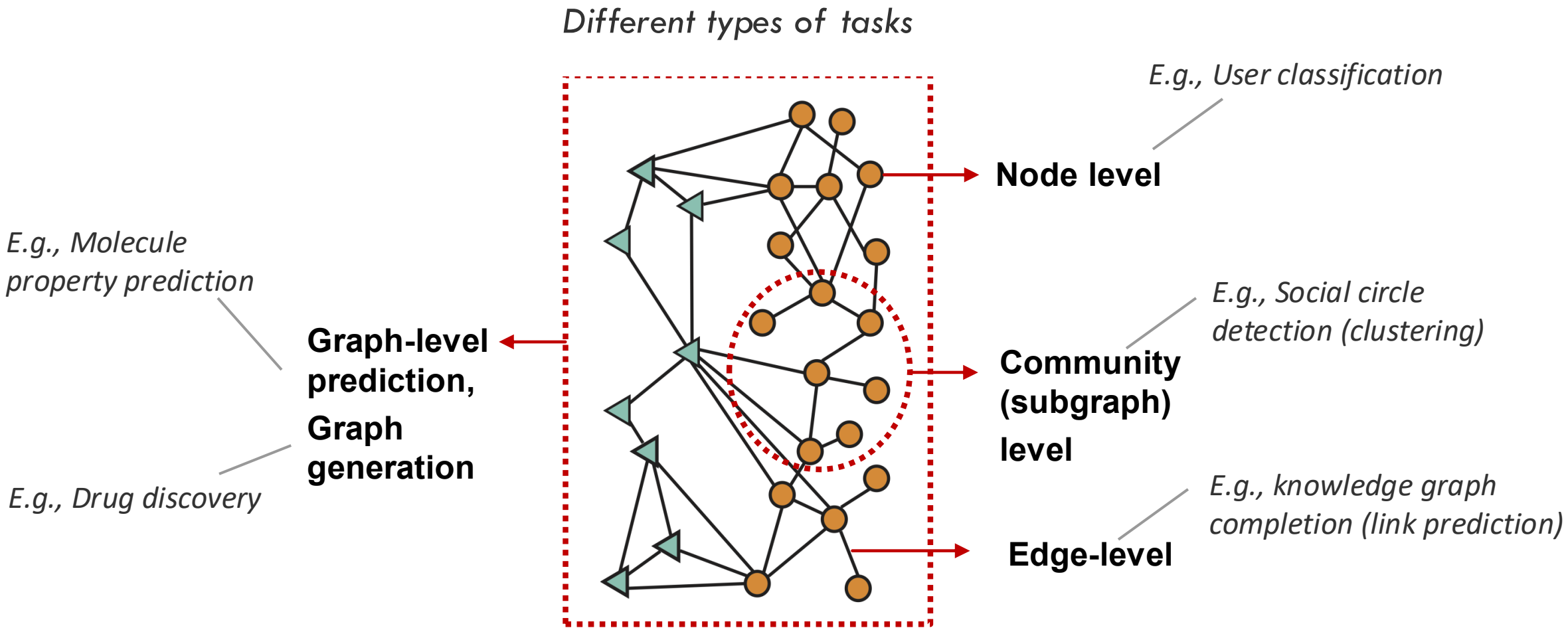
2) Graph/network mining



[Jure Leskovec, Stanford CS224W: Machine Learning with Graphs]

This Course

2) Graph/network mining



[Jure Leskovec, Stanford CS224W: Machine Learning with Graphs]

This Course

2) Graph/network mining
   - Node embedding
   - Graph neural networks
   - Knowledge graphs and reasoning
   - …

This Course

3) Recommender systems

Example recommender systems

Facebook–"People You May Know"

Netflix–"Other Movies You May Enjoy"

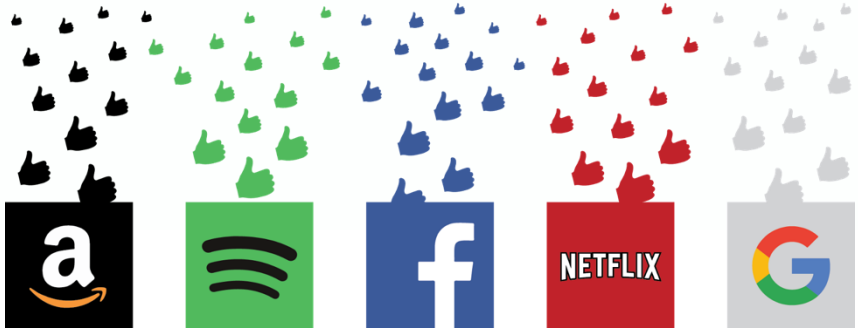
LinkedIn–"Jobs You May Be Interested In"

Amazon–"Customer who bought this item also bought …"

YouTube–"Recommended Videos"

Google–"Search results adjusted"

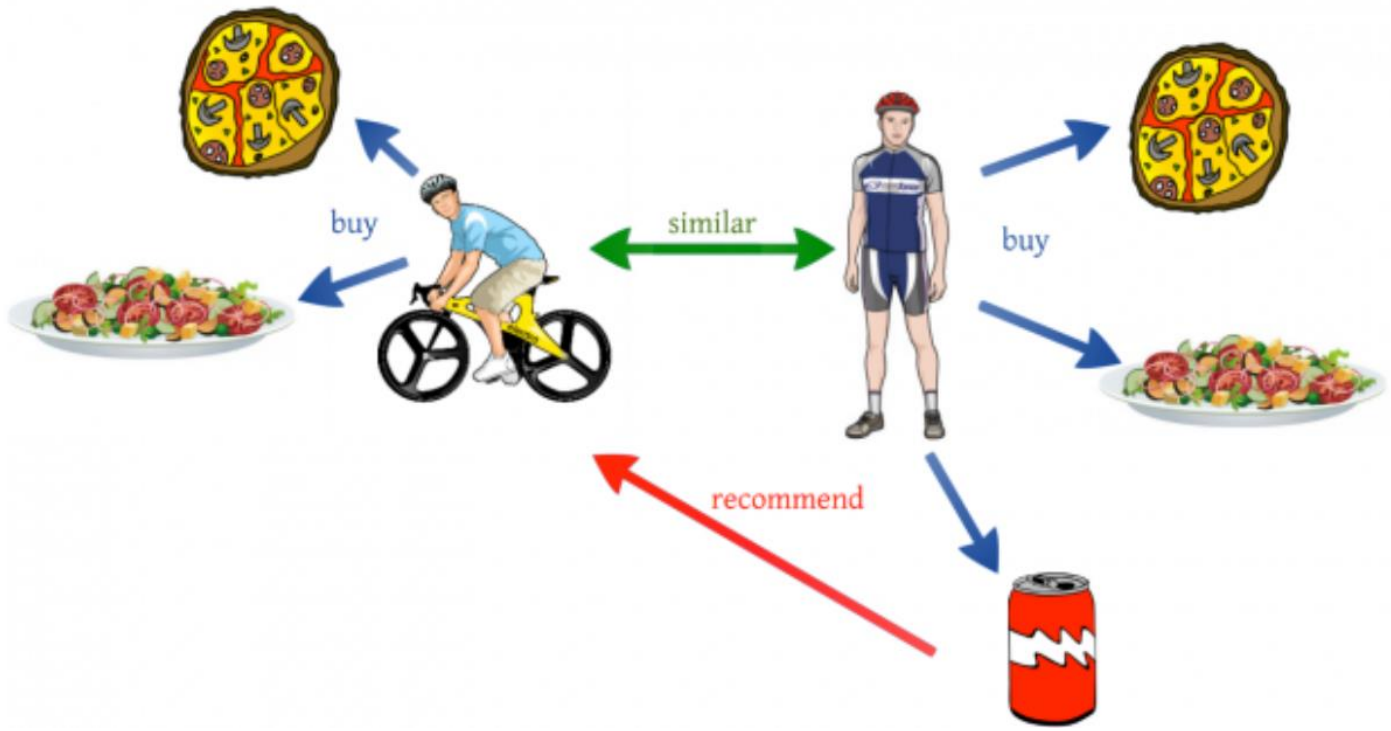
Pinterest–"Recommended Images"



[Crossing Minds, https://www.slideshare.net/CrossingMinds/recommendation-system-explained]

This Course

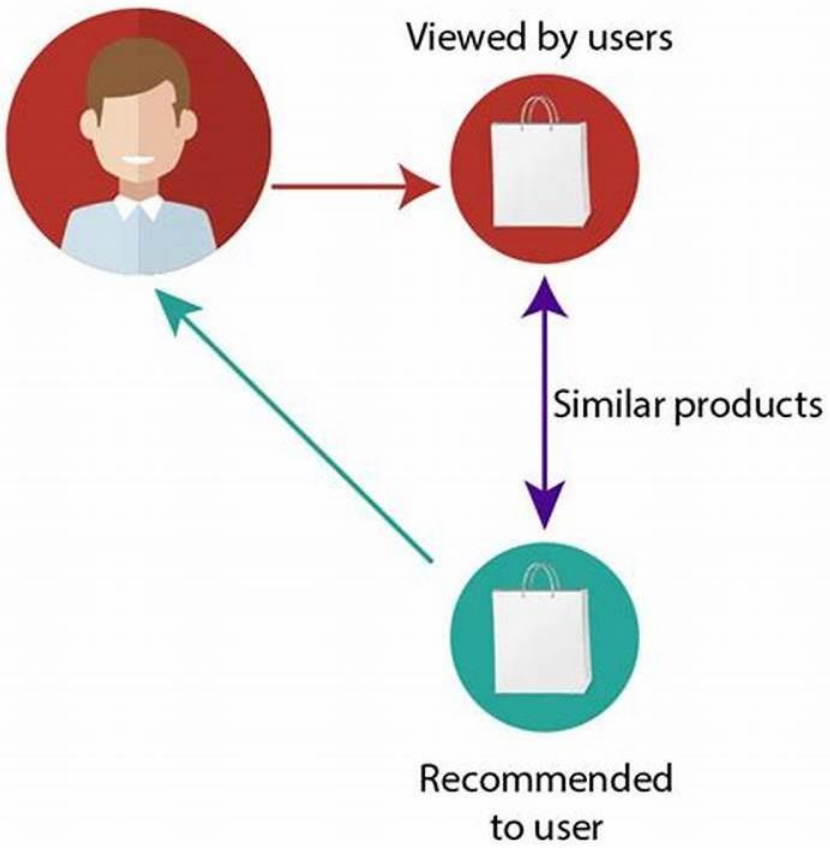
3) Recommender systems
   - Collaborative filtering
     - Matrix factorization, deep learning methods, …



https://d4datascience.com/category/predictive-analytics/

This Course

3) Recommender systems
   - Collaborative filtering
     - Matrix factorization, deep learning methods, …
   - Content-based recommendation
     - Object similarity measure



https://d4datascience.com/category/predictive-analytics/

This Course

3) Recommender systems
   - Collaborative filtering
     - Matrix factorization
   - Content-based recommendation
     - Object similarity measure
   - Graph neural networks for recommendation

https://d4datascience.com/category/predictive-analytics/

This Course: Summary

1) Text (multi-modal) mining
   - Topic models
     - LDA, Expectation Maximization, variational inference
   - Language models
   - Text representation learning (embedding)
2) Graph/network mining
   - Node embedding
   - Graph neural networks
   - Knowledge graphs and reasoning
3) Recommender systems
   - Collaborative filtering
     - Matrix factorization
   - Content-based recommendation
     - Object similarity measure
   - Graph neural networks for recommendation

Questions?