Data Mining: Introduction

Lecture Notes for Chapter 1

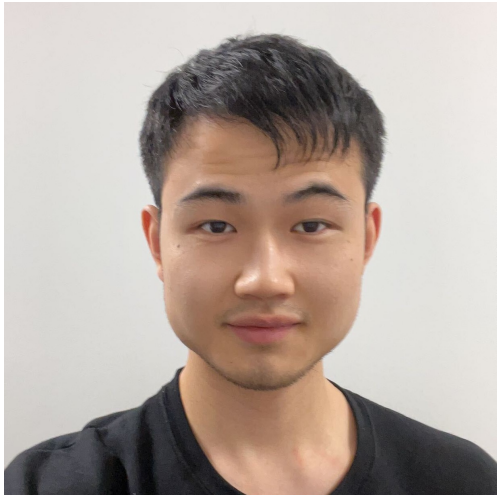
Data Mining

https://ml-graph.github.io/winter-2025/

Yu Wang, Ph.D.
yuwang@uoregon.edu
Assistant Professor
Computer Science
University of Oregon
CS 453/553 – Winter 2025

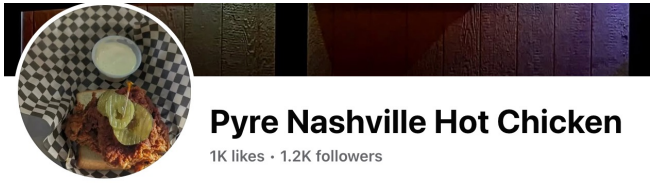
Course Lecture is very heavily based on
"Introduction to Data Mining"
by Tan, Steinbach, Karpatne, Kumar

Self-introduction



Yu (Jack) Wang

You

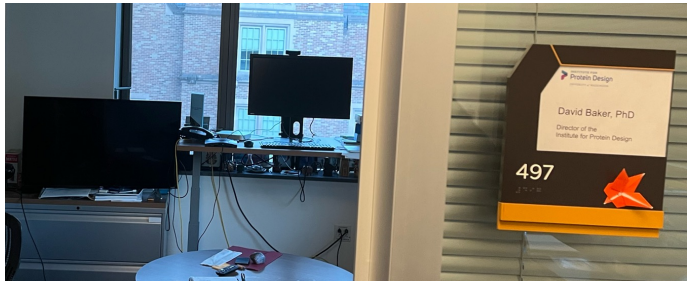


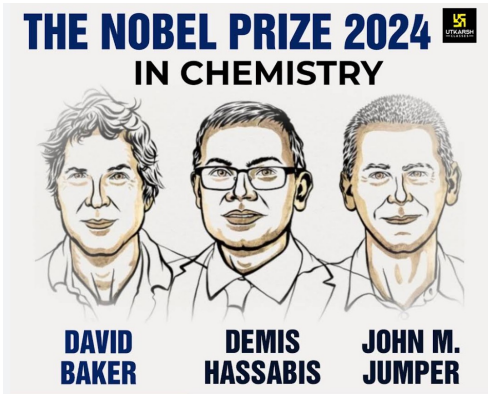
Pyre Nashville Hot Chicken

1K likes · 1.2K followers

Intro

I learned to make Nashville Hot Chicken while living in Nashville. When I moved to Eugene I tried several local places selling "Nashville Hot" but none of it reminded me of Nashville. So I opened Pyre to serve up traditional Nashville Hot Chicken.





Self-introduction

Our lab is actively recruiting!

- **Data Mining and Machine Learning**
- **Graph Machine Learning**
- **Agentic AI**
- **Spatial Temporal Learning**
- **AI/ML Application:**

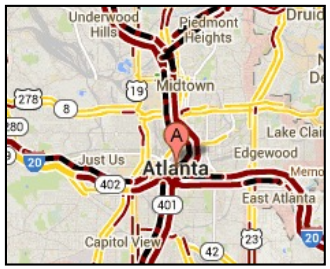
  Information Retrieval/Science/Cyber-security

Large-scale Data is Everywhere!

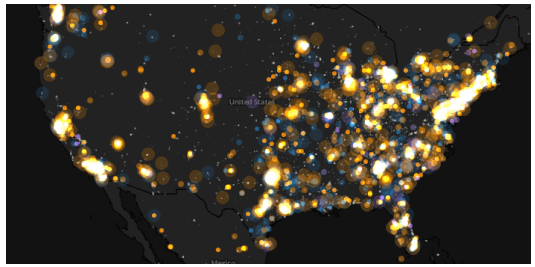
- There has been enormous data growth in both commercial and scientific databases due to advances in data generation and collection technologies
- New mantra
  - Gather whatever data you can whenever and wherever possible.
- Expectations
  - Gathered data will have value either for the purpose collected or for a purpose not envisioned.
- **Model can save the data**
  - **LLM3 – 70B**
  - **Google vs ChatGPT**
  - **RAG**

Cyber Security

E-Commerce

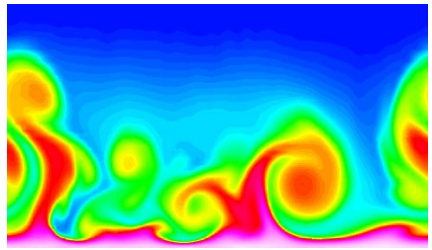


Traffic Patterns



Social Networking: Twitter

Sensor Networks



Computational Simulations

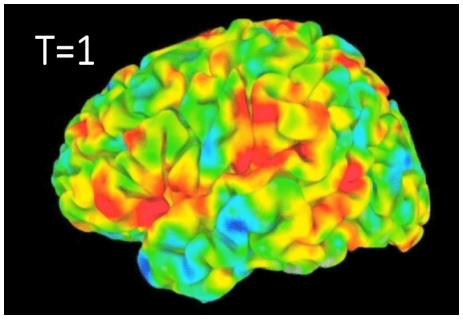
Why Data Mining? Commercial Viewpoint

- Lots of data is being collected and warehoused
  - Web data **1,000 terabytes, 1,000,000,000,000,000= bytes**
    - Google 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)

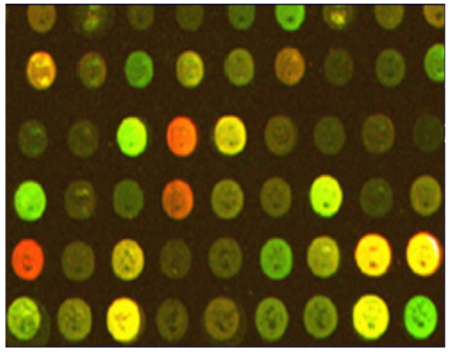
Why Data Mining? Scientific Viewpoint

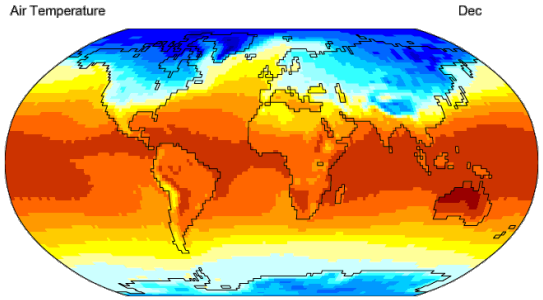
- Data collected and stored at enormous speeds
  - remote sensors on a satellite
    - NASA EOSDIS archives over petabytes of earth science data / year
  - telescopes scanning the skies
    - Sky survey data
  - High-throughput biological data
  - scientific simulations
    - terabytes of data generated in a few hours



fMRI Data from Brain


Sky Survey Data


Gene Expression Data



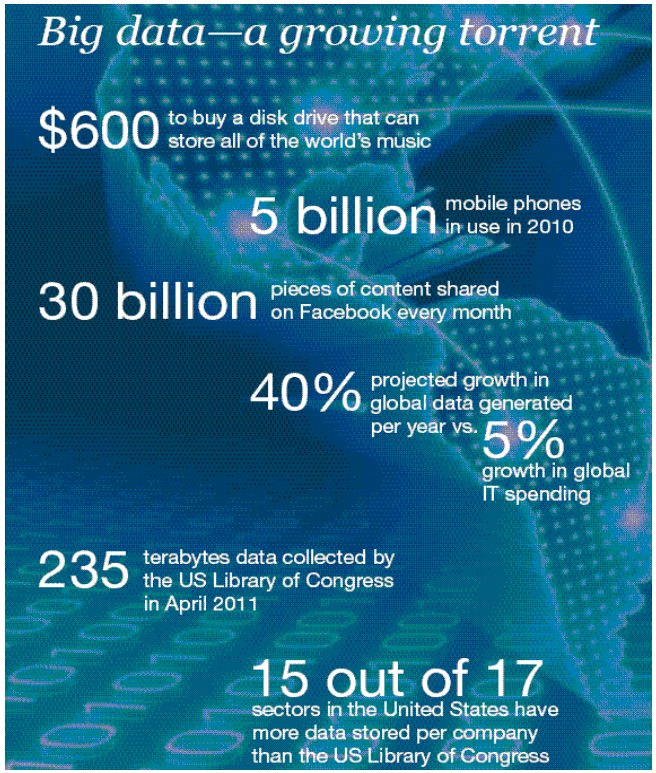
Surface Temperature of Earth

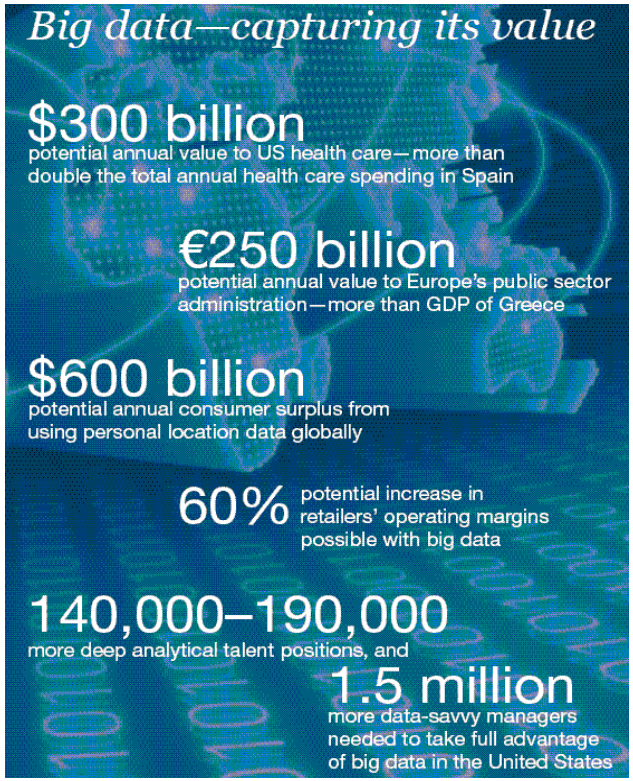
- Data mining helps scientists
  - in automated analysis of massive datasets
  - In hypothesis formation

Great opportunities to improve productivity in all walks of life

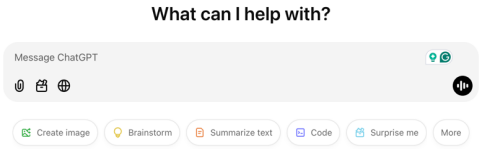
McKinsey Global Institute

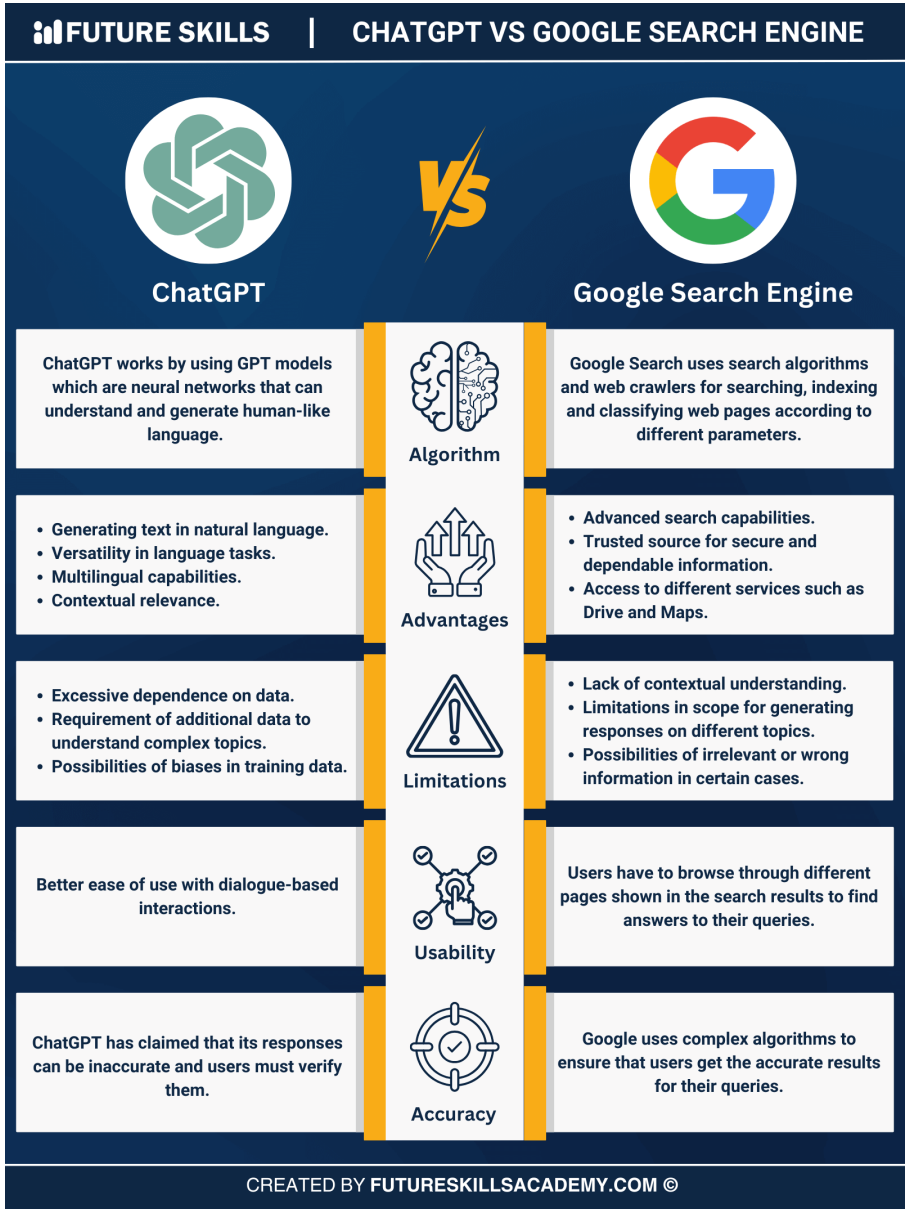
Big data: The next frontier for innovation, competition, and productivity

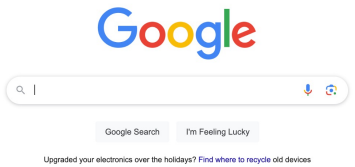




Great opportunities to improve productivity in all walks of life

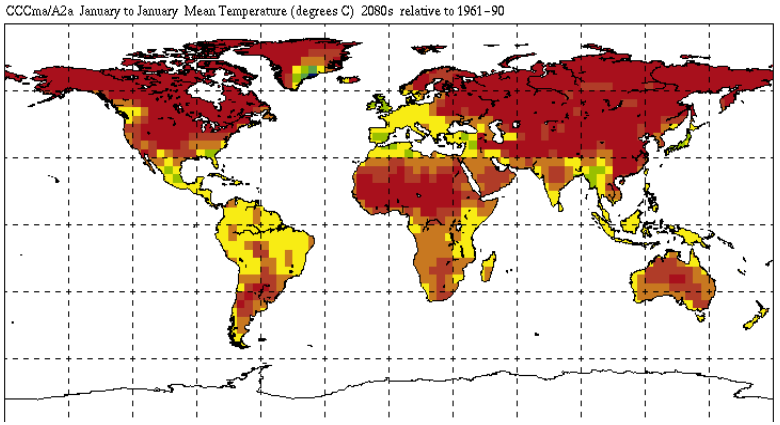






Great Opportunities to Solve Society's Major Problems

Improving health care and reducing costs



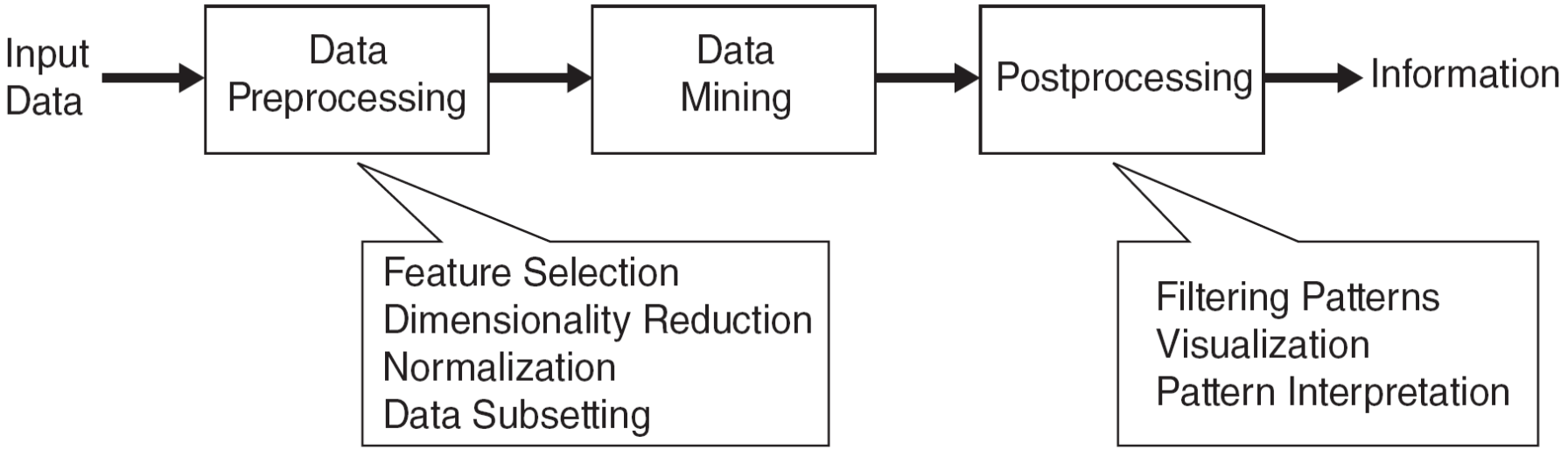
Predicting the impact of climate change

Finding alternative/ green energy sources

Reducing hunger and poverty by increasing agriculture production

What is Data Mining?

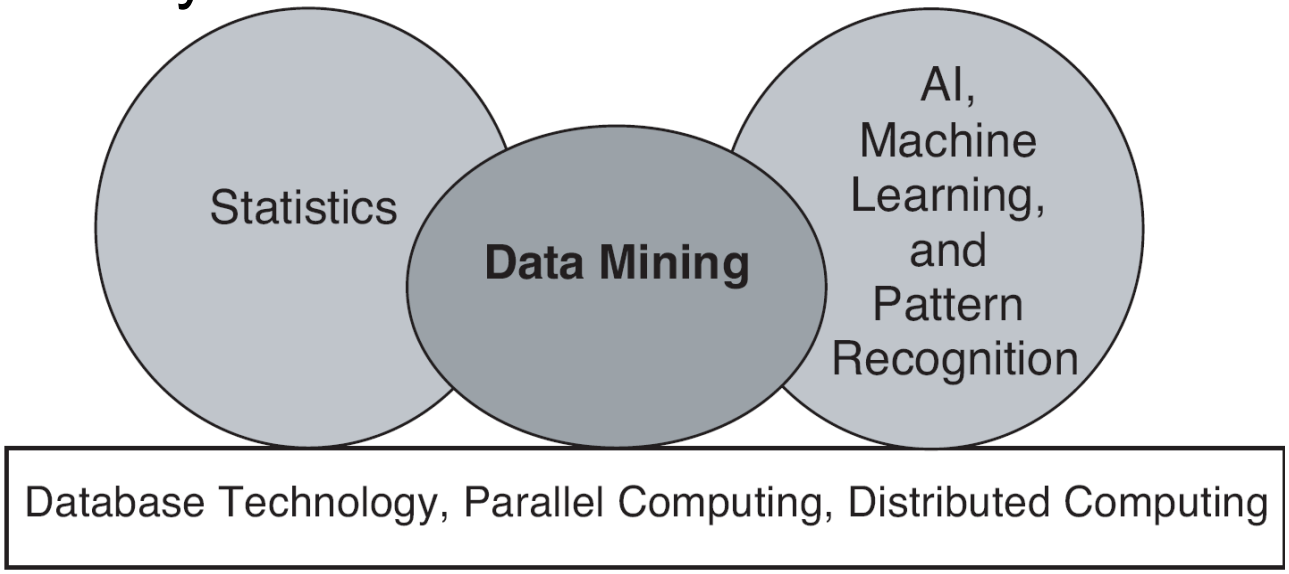
- Many Definitions
  - Non-trivial extraction of implicit, previously unknown and potentially useful information from data
  - Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns



Origins of Data Mining

- Draws ideas from machine learning/AI, pattern recognition, statistics, 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

- Prediction Methods
  - Use some variables to predict unknown or future values of other variables.

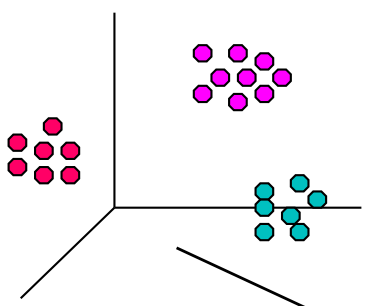
  Rent Prediction

- Description Methods
  - Find human-interpretable patterns that describe the data.

  The larger the apartment, the higher the price

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

Data Mining Tasks ...

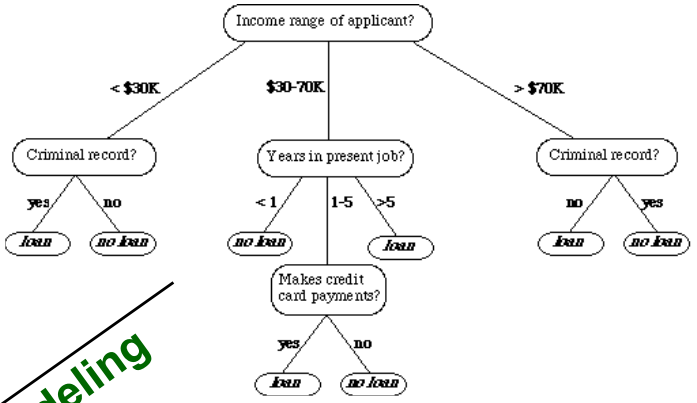


Clustering

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

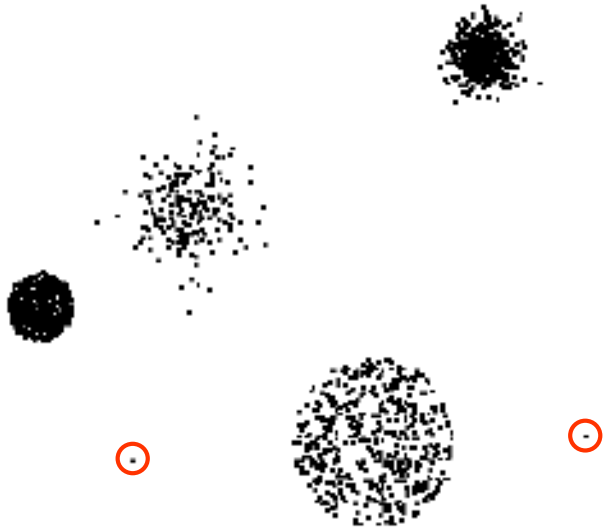
Predictive Modeling



Association Rules

Milk

Anomaly Detection



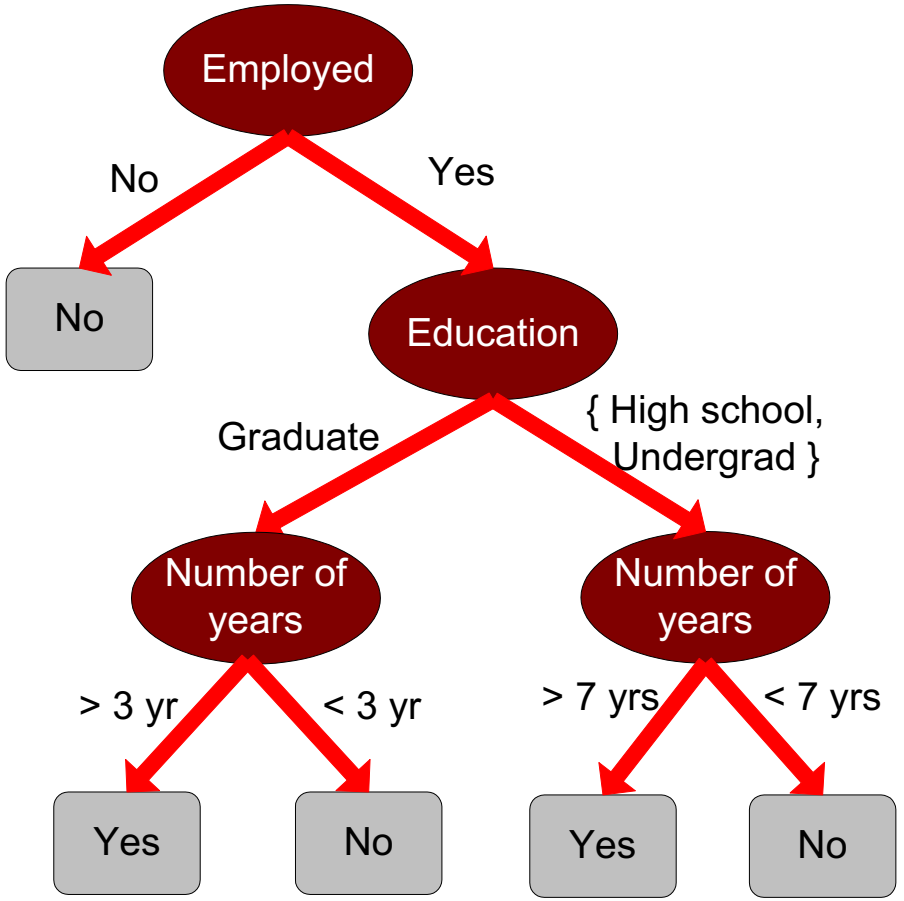
Predictive Modeling: Classification

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

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
…	…	…	…	…

Model for predicting credit worthiness

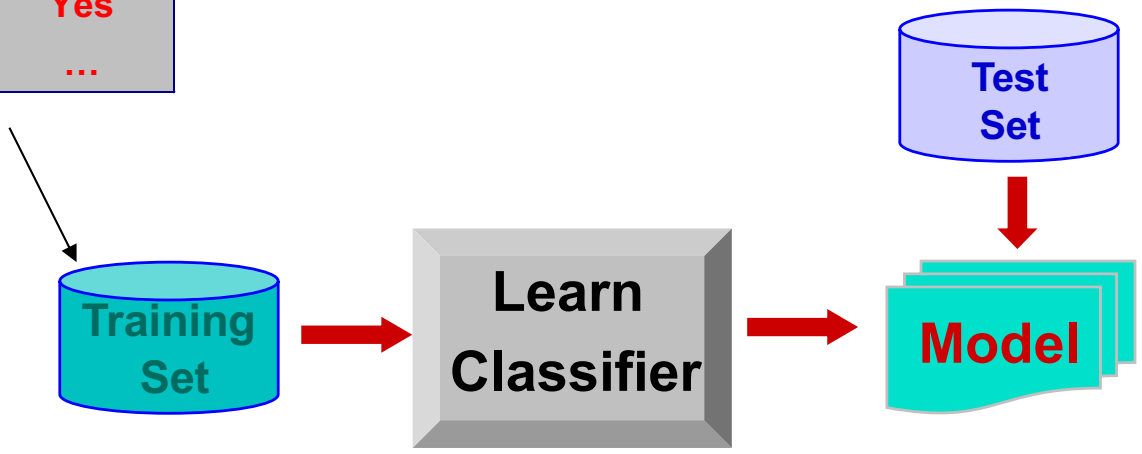


Classification Example

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	?
…	…	…	…	…

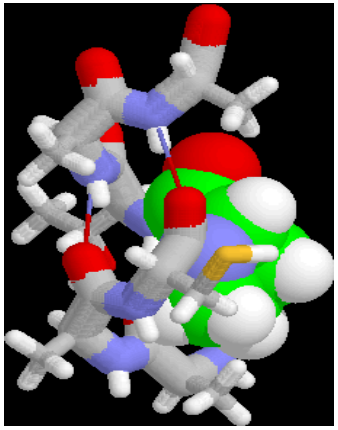


Examples of Classification Task

- Classifying credit card transactions as legitimate or fraudulent

- Classifying land covers (water bodies, urban areas, forests, etc.) using satellite data

- Categorizing news stories as finance, weather, entertainment, sports, etc
- Identifying intruders in the cyberspace
- Predicting tumor cells as benign or malignant



- Classifying secondary structures of protein as alpha-helix, beta-sheet, or random coil

Classification: Application 1

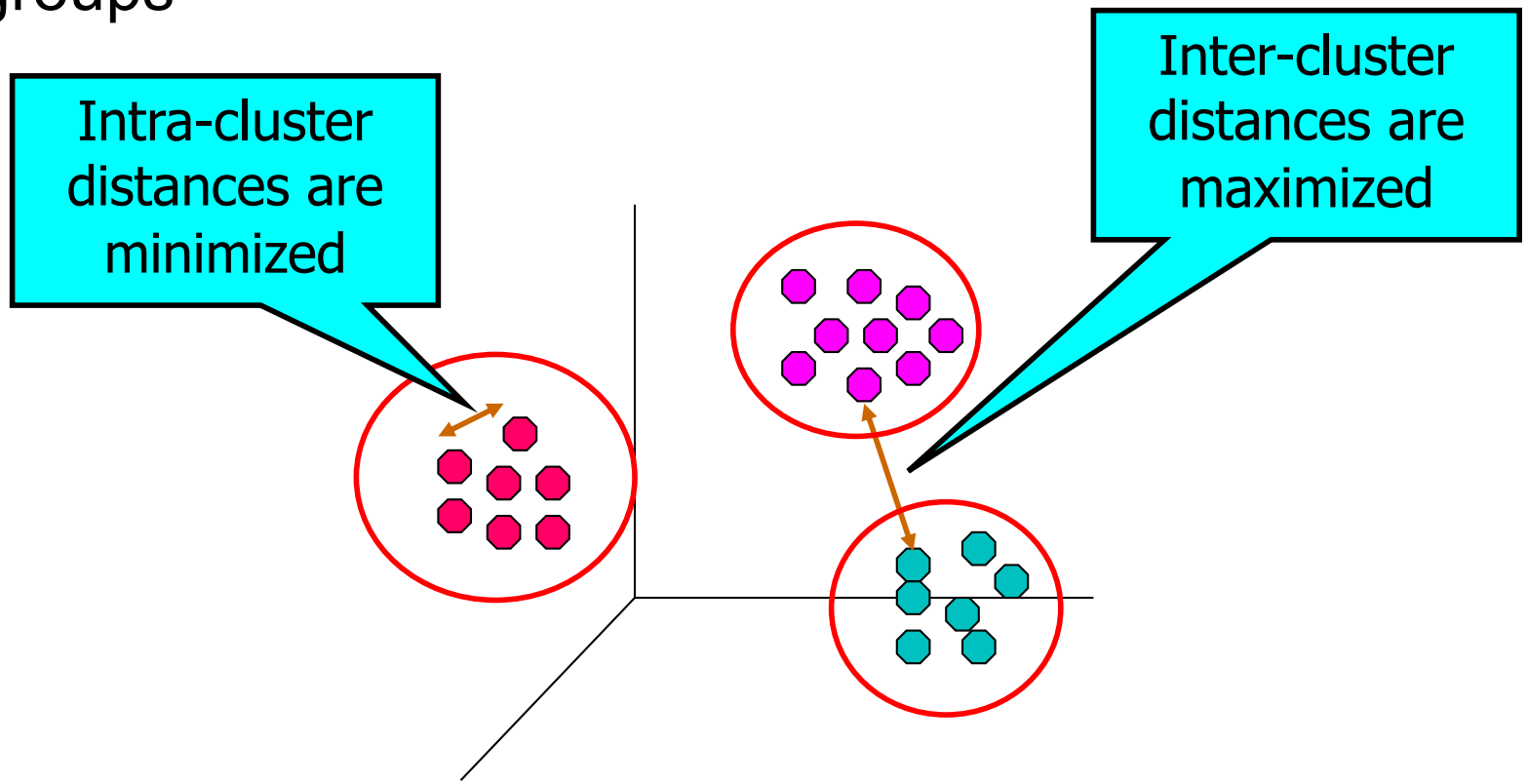
- Fraud Detection
  - **Goal:** Predict fraudulent cases in credit card transactions.
  - **Approach:**
    - Use credit card transactions and the information on its account-holder as attributes.
      - When does a customer buy, what does he buy, how often he pays on time, etc
    - Label past transactions as fraud or fair transactions. This forms the class attribute.
    - Learn a model for the class of the transactions.
    - Use this model to detect fraud by observing credit card transactions on an account.

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 advetising expenditure.
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Time series prediction of stock market indices.

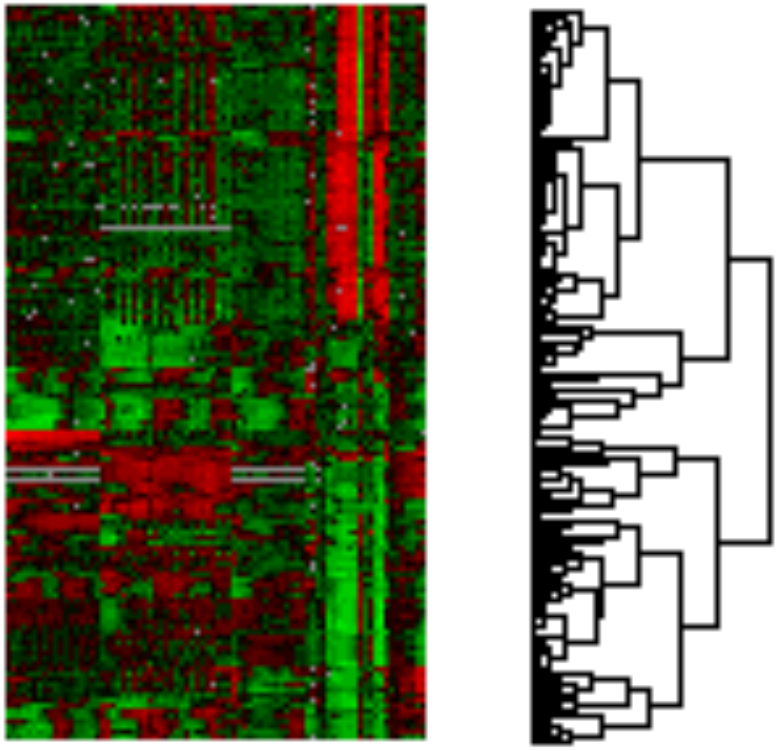
Clustering

- 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

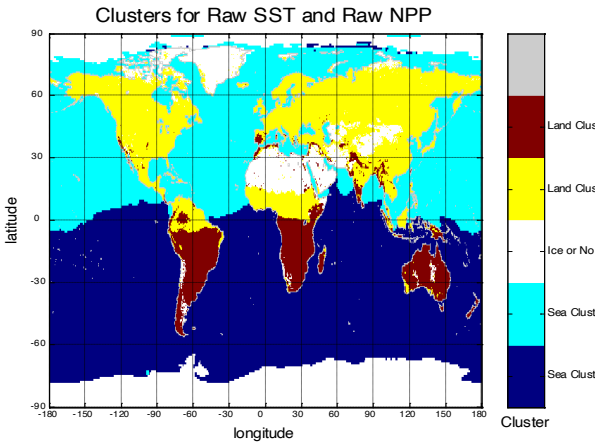


Applications of Cluster Analysis

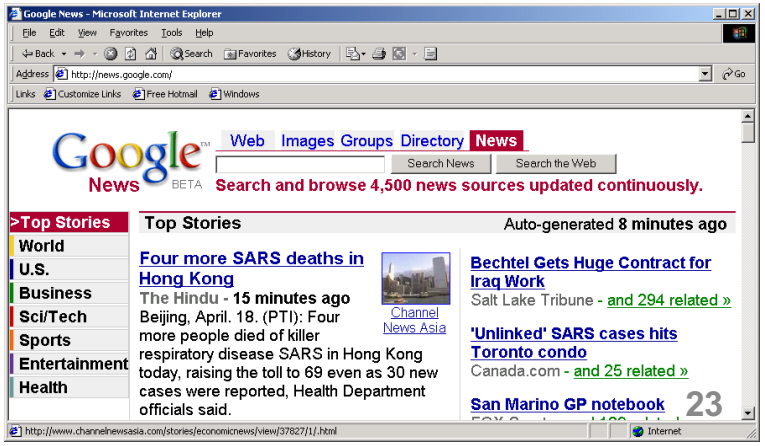
- **Understanding**
  - Custom profiling for targeted marketing
  - Group related documents for browsing
  - Group genes and proteins that have similar functionality
  - Group stocks with similar price fluctuations
- **Summarization**
  - Reduce the size of large data sets



Courtesy: Michael Eisen



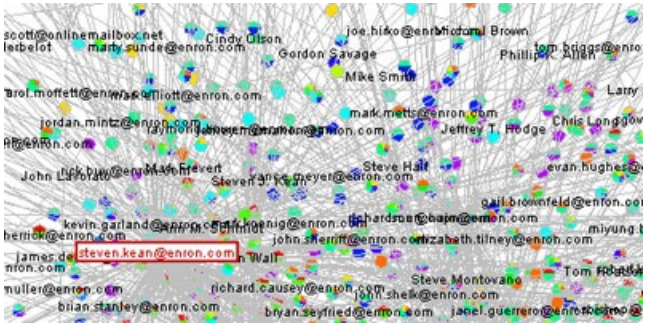
Use of K-means to partition Sea Surface Temperature (SST) and Net Primary Production (NPP) into clusters that reflect the Northern and Southern Hemispheres.



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.

Enron email dataset



Association Rule Discovery: Definition

- 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}

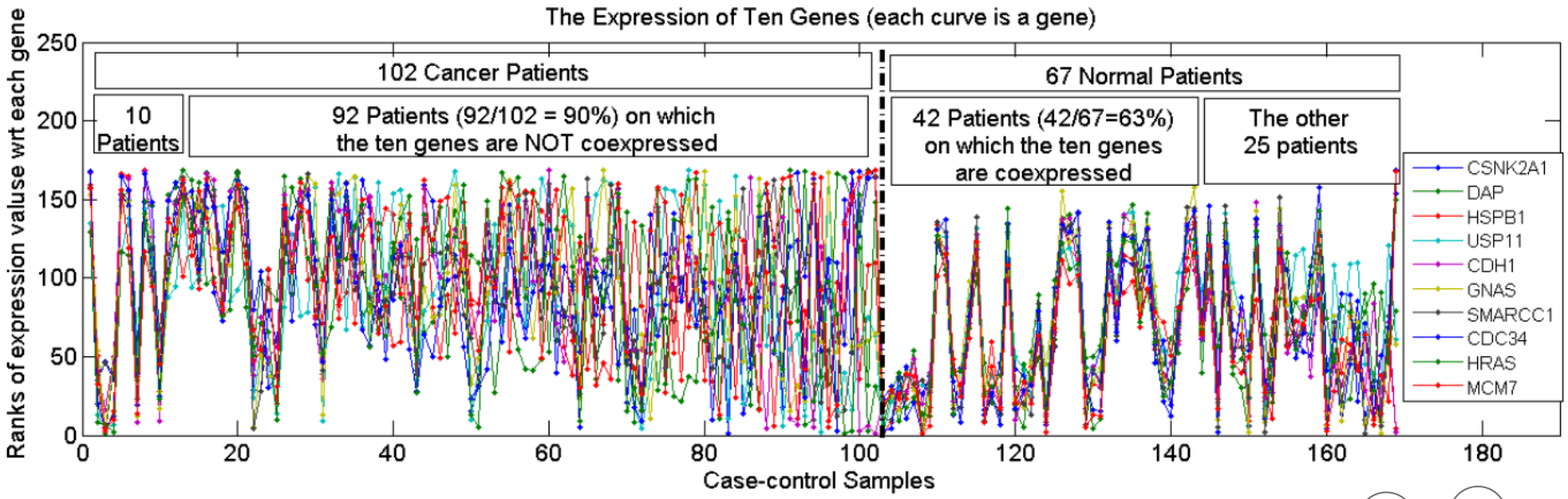
Association Analysis: Applications

- 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

Association Analysis: Applications

- An Example Subspace Differential Coexpression Pattern from lung cancer dataset

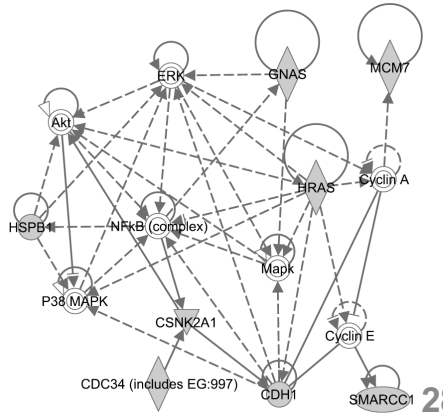
Three lung cancer datasets [Bhattacharjee et a 2001], [Stearman et al. 2005], [Su et al. 2007]



Enriched with the TNF/NFB signaling pathway

which is well-known to be related to lung cancer

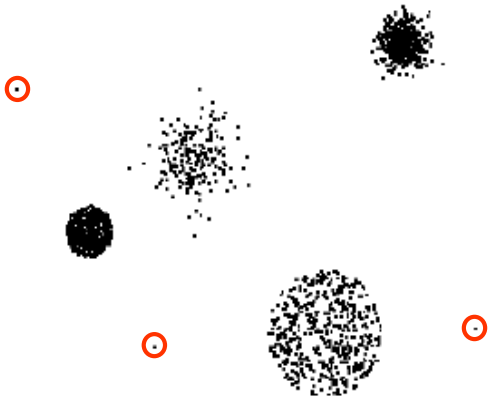
P-value: $1.4*10^{-5}$ (6/10 overlap with the pathway)

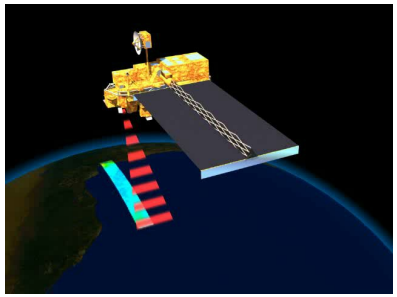


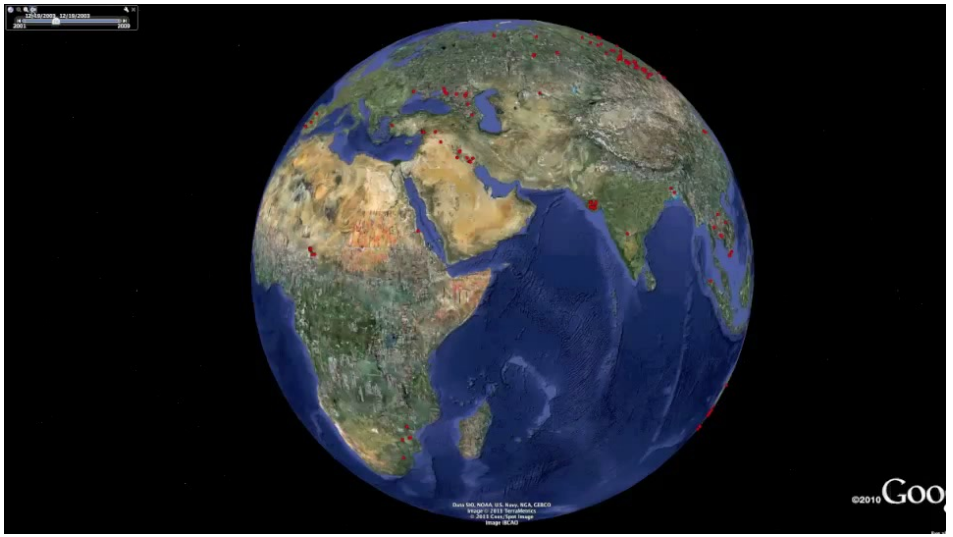
[Fang et al PSB 2010]

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.







Motivating Challenges

- Scalability
- High Dimensionality
- Heterogeneous and Complex Data
- Data Ownership and Distribution
- Non-traditional Analysis

Data Mining in Modern Era

- LLM is like a human agent
- Neural-Symbolic Harmonization
- If you want to choose the career in DS/MLE, become very familiar/deep in your Statistics/ML rather than tool-user/leetcode-lover
- Or become a domain-expert (symbolic knowledge to the extreme case) and control LLM to help you do domain problem

Data Mining in Modern Era

OpenAI cofounder Ilya Sutskever says the way AI is built is about to change

Ilya Sutskever. Photo by JACK GUEZ/AFP via Getty Images

/ “We’ve achieved peak data and there’ll be no more,” OpenAI’s former chief scientist told a crowd of AI researchers.

By **Kylie Robison**, a senior AI reporter working with The Verge's policy and tech teams. She previously worked at Fortune Magazine and Business Insider.

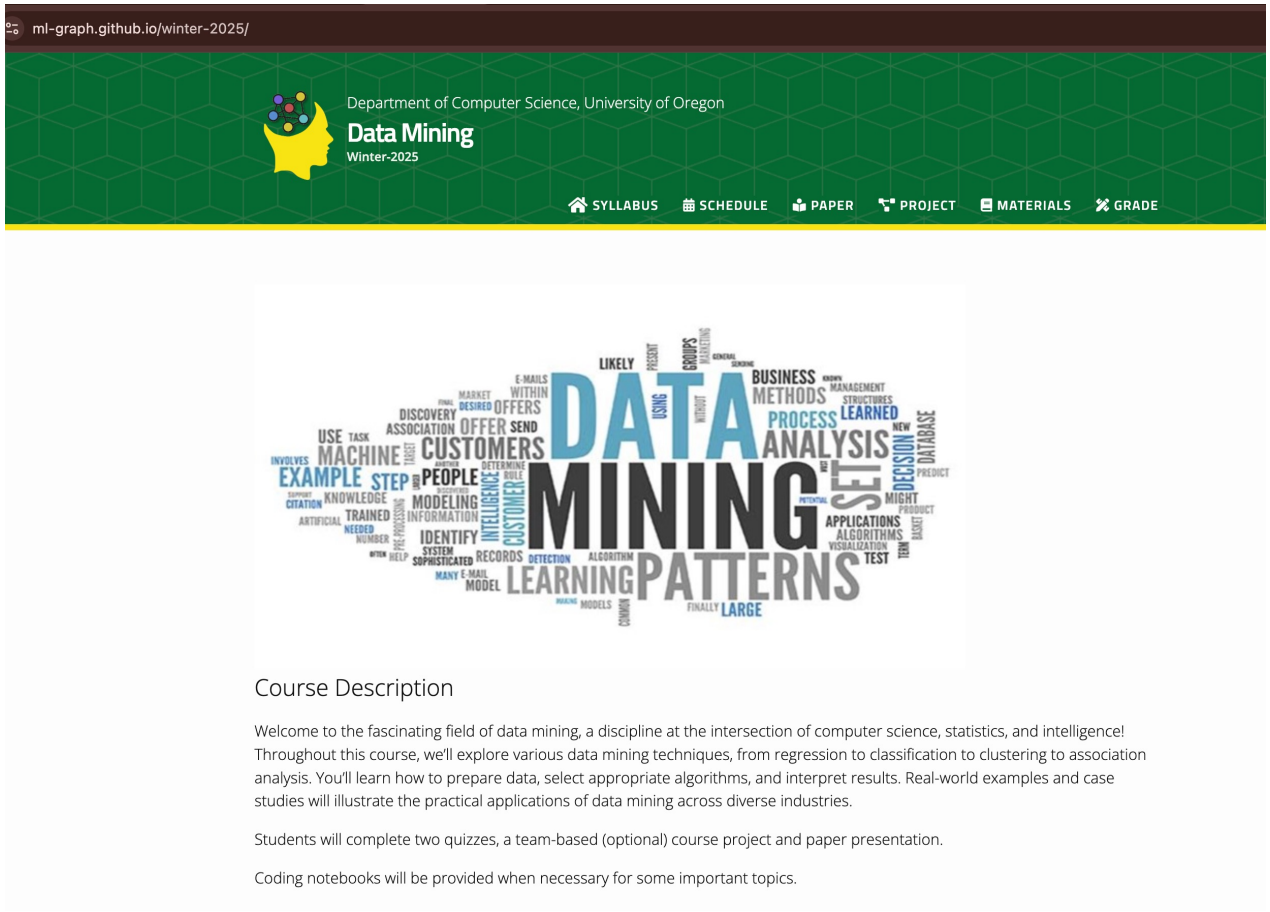
Dec 13, 2024, 4:34 PM PST

13 Comments (13 New)

Any Question?

1. "Judge a man by his questions rather than by his answers."
   – Voltaire
2. "If I had an hour to solve a problem, I'd spend 55 minutes thinking about the problem and 5 minutes thinking about solutions."
   – Albert Einstein
3. "The art and science of asking questions is the source of all knowledge."
   – Thomas Berger
4. "Asking the right questions takes as much skill as giving the right answers."
   – Robert Half
5. "The wise man doesn't give the right answers, he poses the right questions."
   – Claude Lévi-Strauss
6. "Great questions make great companies."
   – Peter Drucker

Course Logistics - Overview



https://ml-graph.github.io/winter-2025/

Course Logistics - Overview

Goals:

- **Broad overview of Data Mining**
- **Basic Data Mining Knowledge/Algorithms, Data Processing Tool**
- **Very classic Data Mining Code**
- **Master real-world GML/DM applications**

Requirements:

- **Little Knowledge in ML**
- **Basic linear algebra, probability and statistics, and calculus**
- **Programming – Python**
- **Coding - PyTorch**

Course Logistics - Overview

Times:

- **Classes: Monday/Wednesday 10:00-11:20 am, 166 LA**
- **Office hours: Friday 4:00-5:00 pm PST, other time by appointment**
- **Zoom: https://uoregon.zoom.us/j/4052006678**

Components:

Course Assessment and Grading Scale

Category	CS-453 (%)	CS-553 (%)
Quzz 1	20%	15%
Quzz 2	20%	15%
Project	40%	45%
Participation	5%	5%
Paper Presentation	15%	20%
Overleaf Bonus	5%	5%

Course Logistics - Overview

Quiz:

- **Test the basic knowledge**
- **Do not be Afraid**
- **As long as you understand the content, you will be good**

Participation:

- **Expected to be on-site**
- **But allow virtual attend upon request**
- **But still I do not have any right to force you to attend onsite ☺**

Course Logistics - Overview

Project:

- **Test the basic knowledge**
- **Do not be Afraid**
- **As long as you understand the content, you will be good**

Participation:

- **Expected to be on-site**
- **But allow virtual attend upon request**
- **But still I do not have any right to force you to attend onsite ☺**

Course Logistics - Overview

Paper Presentation (Why we need?):

1. Introduction and Background – What is the general impact and background of the topic?

2. Motivation and Problem – What is the core research problem and why do we study it?

3. Related Work and Challenges – How did previous works on this problem and what are some challenges?

4. Proposed Solutions/Methods and Rationale – What are the proposed methods/techniques and why propose them? What specific reasons that solving this problem would require these proposed methods/techniques

5. Experimental Setting, Results and Analysis – What experiments are designed to verify the proposed method? How are results being discussed and analyzed? Are there any interesting findings?

6. Conclusion and Future Work

Course Logistics - Overview

Project will Release Soon!

Course Logistics - Overview

EVENT	DATE	DESCRIPTION	COURSE MATERIAL
Lecture	01/06/2025 Monday	Overview Syllabus	**Course Materials:** ◦ Slides
Lecture	01/08/2025 Wednesday	Introduction	**Course Materials:** ◦ Slides
Paper Presentation	01/12/2025 04:30 Sunday	Topic of Paper Release.	
Assignment	01/12/2025 Sunday	**Project released!**	[Project]
Lecture	01/13/2025 Monday	Understanding Data	**Course Materials:** ◦ Slides
Lecture	01/15/2025 Wednesday	Understanding of Data	**Course Materials:** ◦ Slides
Martin Luther King, Jr holiday	01/20/2025 04:30 Monday	Enjoy :)	
Lecture	01/22/2025 Wednesday	Basics of Classification	**Course Materials:** ◦ Slides
Due	**01/24/2025 23:59 Friday**	**Project Proposal Due**	
Lecture	01/27/2025 Monday	Overfitting	**Course Materials:** ◦ Slides
Lecture	01/29/2025 Wednesday	Decision Trees	**Course Materials:** ◦ Slides
Lecture	02/03/2025 Monday	Artificial Neural Networks 1	**Course Materials:** ◦ Slides
Exam	**02/03/2025 16:00 Monday**	**Quzz 1**	**Topics:** ◦ TBD
Lecture	02/05/2025 Wednesday	Artificial Neural Networks 2	**Course Materials:** ◦ Slides
Lecture	02/10/2025 Monday	Rule-based Classifier	**Course Materials:** ◦ Slides
Lecture	02/12/2025 Wednesday	Nearest Neighbor Classifiers	**Course Materials:** ◦ Slides
Lecture	02/17/2025 Monday	Cluster Analysis 1	**Course Materials:** ◦ Slides
Lecture	02/19/2025 Wednesday	Cluster Analysis 2	**Course Materials:** ◦ Slides
Lecture	02/24/2025 Monday	Naive Bayes Classifier 1	**Course Materials:** ◦ Slides
Lecture	02/26/2025 Wednesday	Naive Bayes Classifier 2	**Course Materials:** ◦ Slides
Lecture	03/03/2025 Monday	Support Vector Machine	**Course Materials:** ◦ Slides
Exam	**03/03/2025 16:00 Monday**	**Quzz 2**	**Topics:** ◦ TBD
Lecture	03/10/2025 Monday	Ensemble Methods	**Course Materials:** ◦ Slides