Lecture 1: Overview

Welcome to Spatial Data Mining!

Instructor: Yiqun Xie

Welcome

- Final year for some of you
  - Concepts and techniques in this class may increase your competitiveness for immediate next steps
- First year for some of you
  - May help accelerate your research

COVID-Related Policies

- Updated campus-level policy
  - “Effective Monday, August 29, wearing a mask will not be required while indoors in most situations, including classrooms. As a reminder, masks are a significant defense against the spread of COVID-19 and other respiratory viruses. Therefore, I recommend wearing a KN95 mask while indoors for added protection.”
  - We will follow campus guidelines in case of changes

- If you have symptoms or concerns:
  - HEAL line: https://health.umd.edu/HEAL

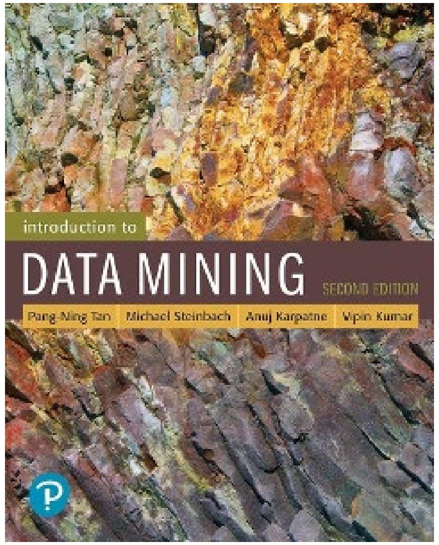
Office Hours

- Right before class
  - 1-2pm, Tu/Th
  - Location: LeFrak hot office: 1111A (or 1111B)
    - I will move to the classroom 10-15 min before the class

- Appointments available

Textbook (Optional Reading)

- Not required for this class
  - Provide additional details for interested students later in the semester
  - Recommendation
    - Introduction to Data Mining, 2nd Edition
    - P. Tan, M. Steinbach, A. Karpatne, and V. Kumar
    - https://www-users.cs.umn.edu/~kumar001/dmbook/index.php



- Lecture notes and slides will be sufficient

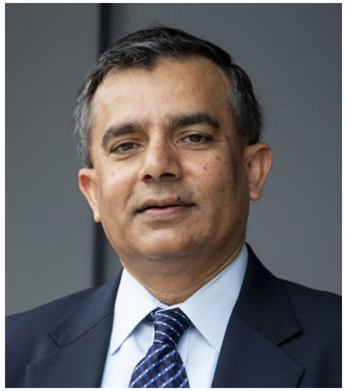
Outline

- Self-introduction
  - Share a few sentences about yourself
  - Background (major), year, experience with data mining
- Syllabus
  - Topics and schedule
  - Grading and policies
- Why data mining?
- Why spatial data mining? What is special?

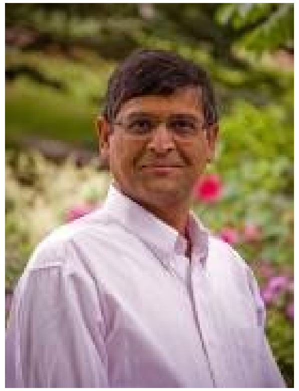
Instructor Self-Introduction

- Name: Yiqun Xie (Yi-cun Sh-yeah)
- Assistant Professor in Geospatial Information Science
- Background: Ph.D. in Computer Science
- Research areas: Spatial AI, Spatial Data Science

PhD Committee



Prof. S. Shekhar
Spatial data mining
Advisor



Prof. V. Kumar
Data mining
(including ST)

Prof. A. Banerjee
Machine learning

Prof. S. Chatterjee
Statistics

Outline

- Self-introduction
  - Share a few sentences about yourself
  - Background (major), year, experience with data mining
- Syllabus
  - Topics and schedule
  - Grading and policies
- Why data mining?
- Why spatial data mining? What is special?

Why We Need to Know about Data Mining?

- Although many of the techniques used to be considered as “high-tech” mastered by a small group of people, they are ubiquitous and fundamental now

- Most large cooperations, institutions, labs, research fields have widely and deeply incorporated the methods

- Skills have high demand in the job market
  - Replacing many traditional approaches (e.g., spam emails)

- Stay competitive and flexible for years to come

Technical Road Map

- Clustering
  - Get familiar with and practice core technical concepts
- Hotspot detection
  - Reinforce key techniques
- Learning and prediction
  - Further advance key techniques and expose to new methods
- Deep learning
  - Higher-level thinking and design

Week #	Week date	Topic	Deliverable*
1	08/29	Overview	
2	09/05	Spatial Data Types and Models	
3	09/12	Early Inception: Database and Spatial Database Classics	
4	09/19	Spatial Statistical Foundations	HW1
5	09/26	Spatial Clustering	
6	10/03	Statistically Robust Clustering: Hotspot Detection I	HW2
7	10/10	Statistically Robust Clustering: Hotspot Detection II	
8	10/17	Midterm Practice & Exam	Midterm Exam
9	10/24	Association Rules, Spatial Co-location, and Outliers	
10	10/31	Prediction Methods I	
11	11/7	Prediction Methods II	HW3
12	11/14	Spatial Prediction Methods III	
13	11/21	Deep Learning for Spatial Data I (Holiday on Thursday)	HW4
14	11/28	Deep Learning for Spatial Data II	
15	12/05	Trends: Trajectory Data Mining, Advanced Learning & Spatial Big Data	HW5 or Proj.

Grading

- Four homeworks (50%)
  - HW1-HW4, each accounts for 12.5% of final grade
- Exams - Midterm (15%)
- Participation (15%)
  - Presentation (groups of one or two)
    - Topic: News/trends related to data mining, or a technique that complements those discussed in class
  - Participate in class
- “Final exam”: (20%)
  - An extra homework for undergraduate students
  - A course project for graduate students (can relate to your own research)

Grading

- Final grades will be curved
  - Separate for undergraduates and graduates
- A mixed view
  - “Curved” here does not mean good grades can only be given to a limited number of students
- Grades are important, but real skills are more critical after college

Policy

- Syllabus

- Best practices
  - You are encouraged to discuss with your peers, but do not share solutions to homework & exam problems
  - Submit your homework early, rather than in the last minute
    - If you submit within 30 minutes past due, valid reasons will be accepted to remove “late submission” penalty; let me know ASAP if that’s the case
    - Submit before the actual due time – do not wait for 30 minutes; Anything after 30-min past due will be strictly late submissions (unless system failure confirmed by system staff)
    - One opportunity allowed to remove a late-submission penalty (HW only)

Important suggestions

- Do not be shy or hesitate to ask (never too late to ask):
  - If you do not know the meaning of a notation (symbol)
  - If you do not know a concept

- Your knowledge builds up if you ask (or search)
- Your confusion builds up if you don’t

- Follow-up after class in office hours or search online
  - Our brains need repetitions
  - Fewer and fewer things you need to search

Acknowledgements and Attributions

- Some slides are modified based on materials from:
  - Prof. Shashi Shekhar
  - Prof. Xun Zhou
  - Prof. Vipin Kumar (open slides)

Outline

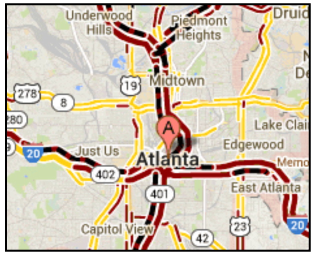
- Self-introduction
  - Share a few sentences about yourself
  - Background (major), year, experience with data mining
- Syllabus
  - Topics and schedule
  - Grading and policies
- Why data mining?
- Why spatial data mining? What is special?

Large-scale Data is Everywhere!

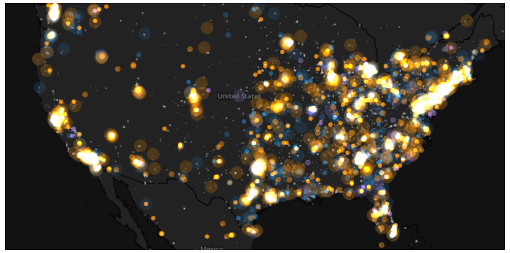
- There has been enormous data growth in both commercial and scientific databases due to advances in data generation and collection technologies
- New mantra
  - Gather whatever data you can whenever and wherever possible.
- Expectations
  - Gathered data will have value either for the purpose collected or for a purpose not envisioned.

Cyber Security

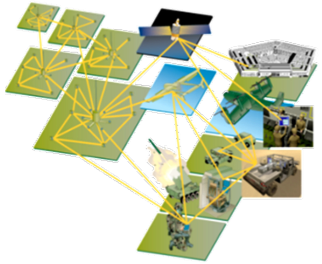
E-Commerce



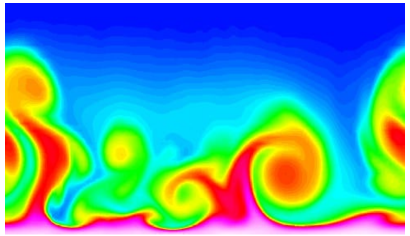
Traffic Patterns



Social Networking: Twitter



Sensor Networks



Computational Simulations

Attribution: https://www-users.cs.umn.edu/~kumar001/dmbook/index.php#item4
Introduction to Data Mining, 2nd Edition   Tan, Steinbach, Karpatne, Kumar

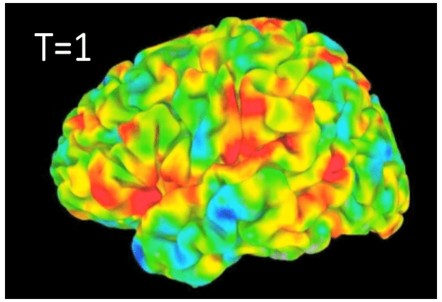
Why Data Mining? Commercial Viewpoint

- Lots of data is being collected and warehoused
  - Web data
    - 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)

Attribution: https://www-users.cs.umn.edu/~kumar001/dmbook/index.php#item4
Introduction to Data Mining, 2nd Edition Tan, Steinbach, Karpatne, Kumar

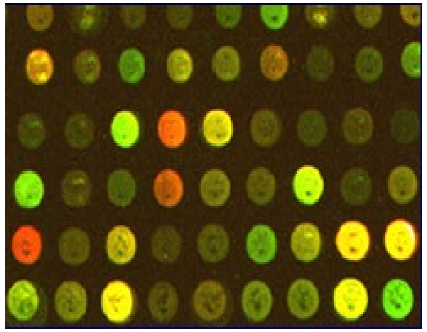
Why Data Mining? Scientific Viewpoint

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

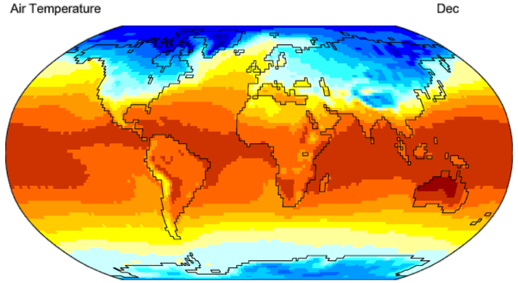


fMRI Data from Brain

Sky Survey Data



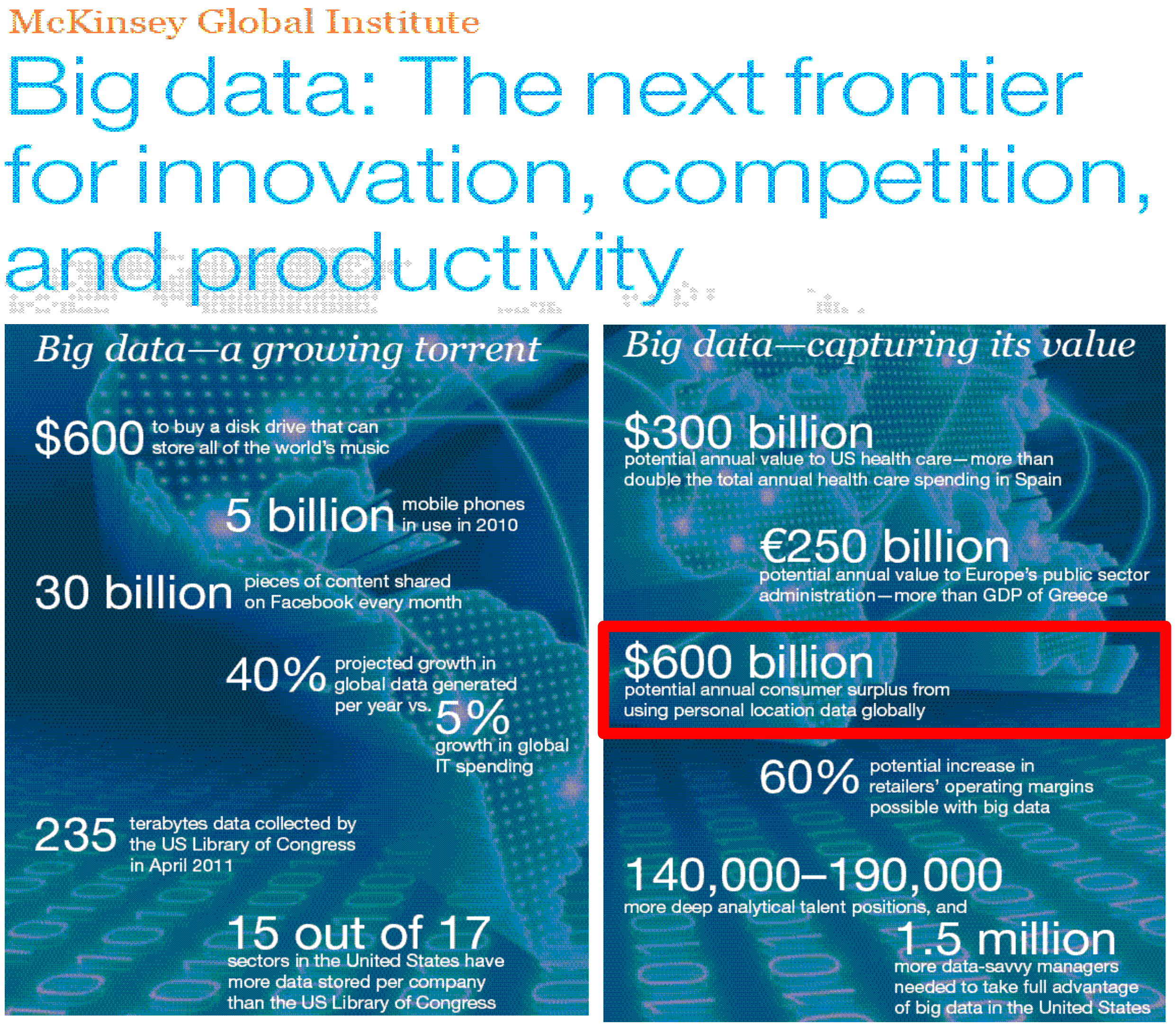
Gene Expression Data



Surface Temperature of Earth

Attribution: https://www-users.cs.umn.edu/~kumar001/dmbook/index.php#item4
Introduction to Data Mining, 2nd Edition Tan, Steinbach, Karpatne, Kumar

Great opportunities to improve productivity in all walks of life



McKinsey Global Institute

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

Big data—a growing torrent

$600 to buy a disk drive that can store all of the world's music

5 billion mobile phones in use in 2010

30 billion pieces of content shared on Facebook every month

40% projected growth in global data generated per year vs. 5% growth in global IT spending

235 terabytes data collected by the US Library of Congress in April 2011

15 out of 17 sectors in the United States have more data stored per company than the US Library of Congress

Big data—capturing its value

$300 billion potential annual value to US health care—more than double the total annual health care spending in Spain

€250 billion potential annual value to Europe's public sector administration—more than GDP of Greece

$600 billion potential annual consumer surplus from using personal location data globally

60% potential increase in retailers' operating margins possible with big data

140,000–190,000 more deep analytical talent positions, and

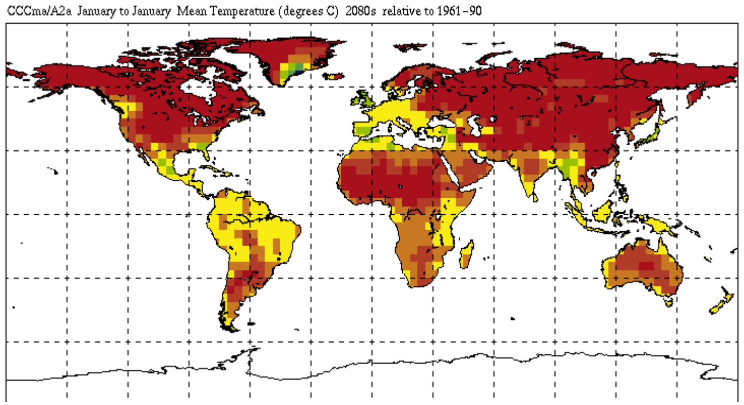
1.5 million more data-savvy managers needed to take full advantage of big data in the United States

"...services enabled by personal-location data can allow consumers to capture **$600 billion**..."

Attribution: https://www-users.cs.umn.edu/~kumar001/dmbook/index.php#item4
Introduction to Data Mining, 2nd Edition Tan, Steinbach, Karpatne, Kumar

Great Opportunities to Solve Society’s Major Problems

Improving health care and reducing costs


CCCma/A2a January to January Mean Temperature (degrees C) 2080s relative to 1961–90

Predicting the impact of climate change

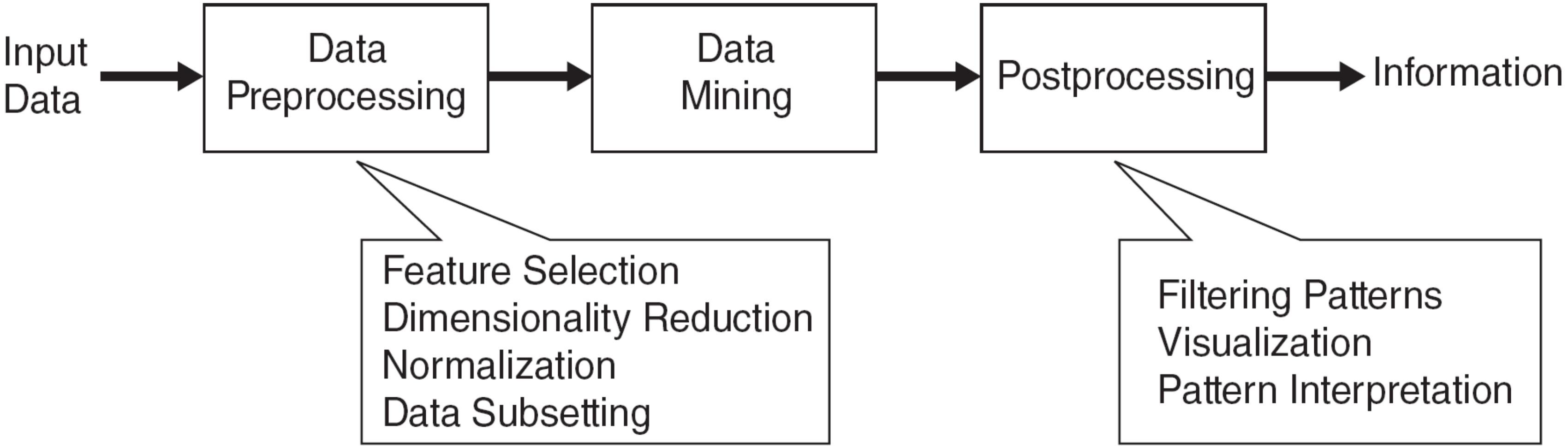
Finding alternative/ green energy sources

Reducing hunger and poverty by increasing agriculture production

Attribution: https://www-users.cs.umn.edu/~kumar001/dmbook/index.php#item4
Introduction to Data Mining, 2nd Edition Tan, Steinbach, Karpatne, Kumar

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

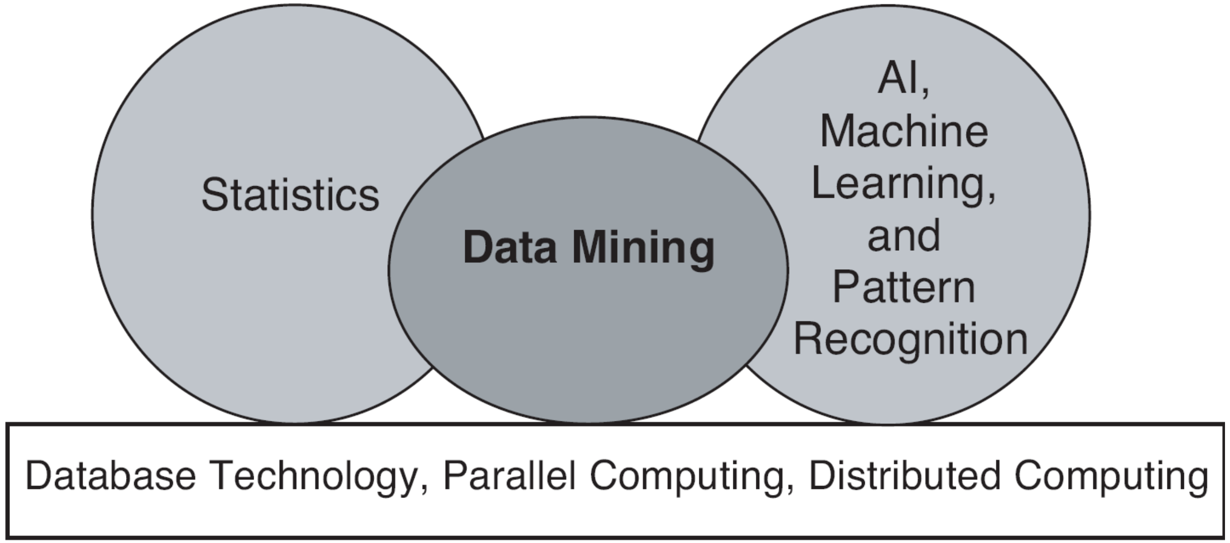


Attribution: https://www-users.cs.umn.edu/~kumar001/dmbook/index.php#item4
Introduction to Data Mining, 2nd Edition Tan, Steinbach, Karpatne, Kumar

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



Statistics

Data Mining

AI,
Machine
Learning,
and
Pattern
Recognition

Database Technology, Parallel Computing, Distributed Computing

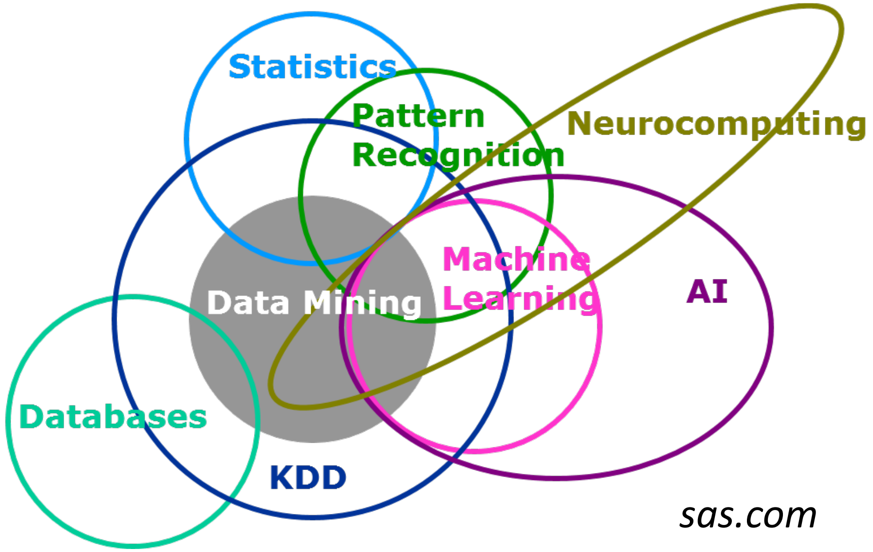
- A key component of the emerging field of data science and data-driven discovery

Attribution: https://www-users.cs.umn.edu/~kumar001/dmbook/index.php#item4
Introduction to Data Mining, 2nd Edition Tan, Steinbach, Karpatne, Kumar

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



sas.com

- A key component of the emerging field of data science and data-driven discovery

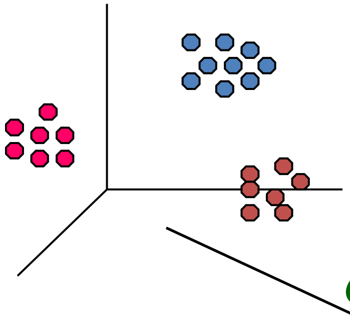
Data Mining Tasks

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

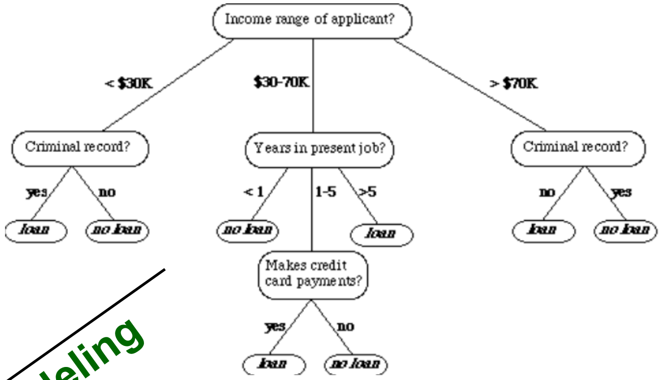
- Description Methods
  - Find human-interpretable patterns that describe the data.

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

Data Mining Tasks …



Clustering



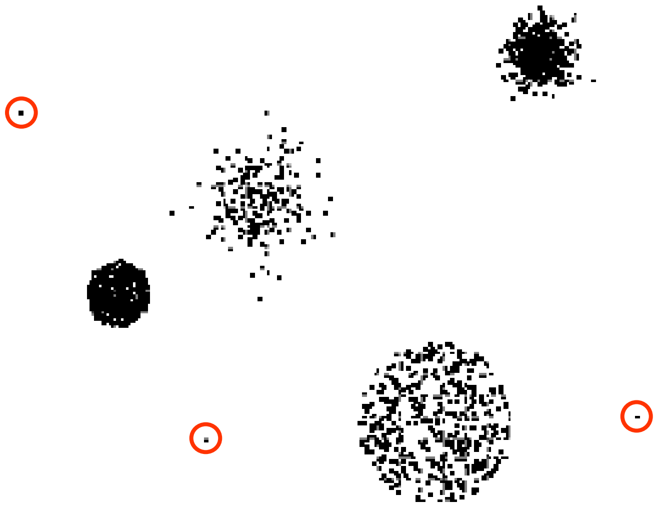
Predictive Modeling

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

Association Rules

Anomaly Detection



Attribution: https://www-users.cs.umn.edu/~kumar001/dmbook/index.php#item4 Introduction to Data Mining, 2nd Edition   Tan, Steinbach, Karpatne, Kumar

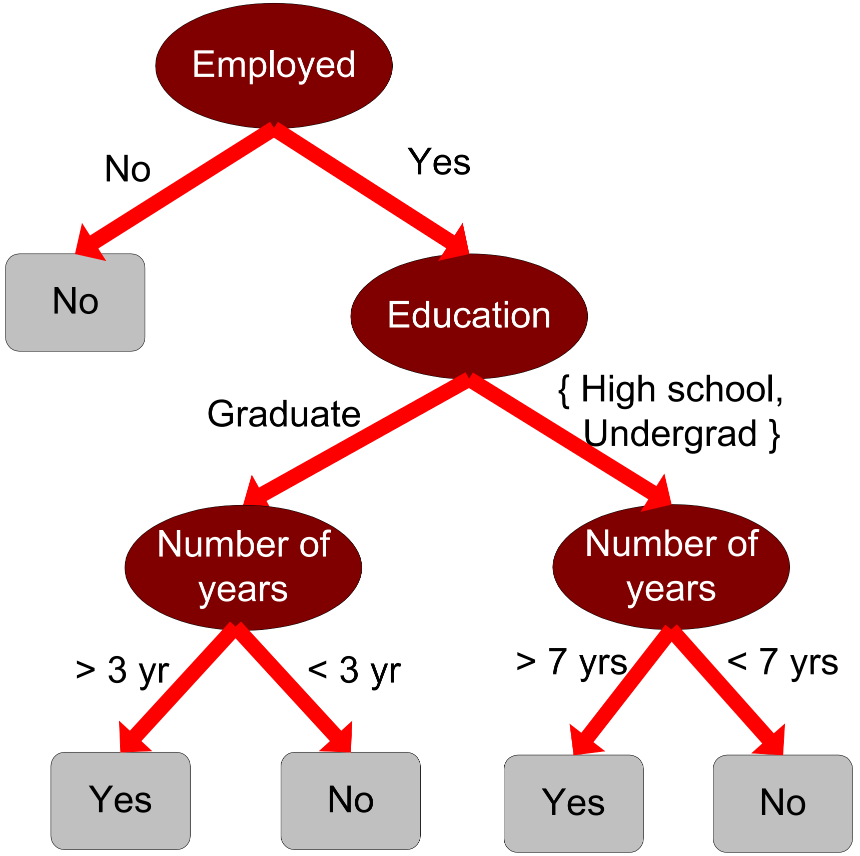
Predictive Modeling: Classification

- 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



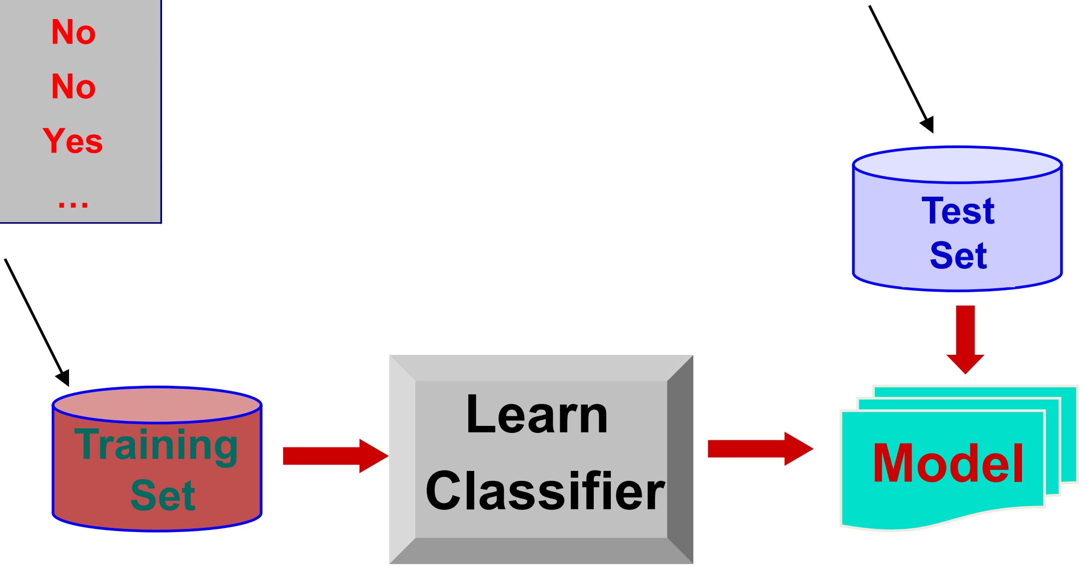
Attribution: https://www-users.cs.umn.edu/~kumar001/dmbook/index.php#item4
Introduction to Data Mining, 2nd Edition   Tan, Steinbach, Karpatne, Kumar

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	?
…	…	…	…	…



Attribution: https://www-users.cs.umn.edu/~kumar001/dmbook/index.php#item4
Introduction to Data Mining, 2nd Edition Tan, Steinbach, Karpatne, Kumar

Examples of Classification Task

- Classifying credit card transactions as legitimate or fraudulent

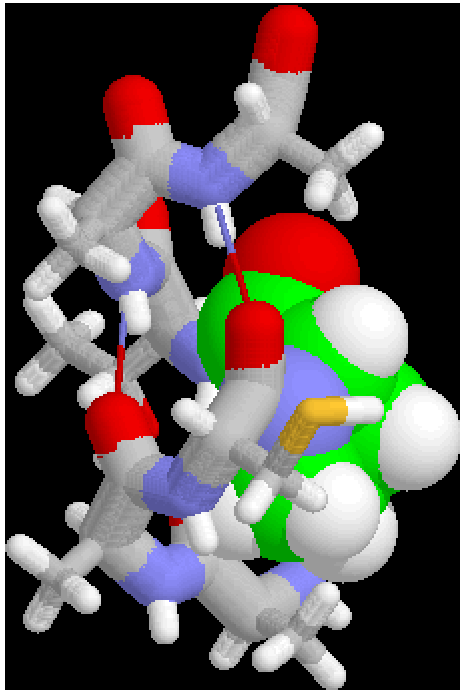
- Classifying land covers (water bodies, urban areas, forests, etc.) using satellite data

- 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



Attribution: https://www-users.cs.umn.edu/~kumar001/dmbook/index.php#item4
Introduction to Data Mining, 2nd Edition   Tan, Steinbach, Karpatne, Kumar

Classifying fraudulent behaviors

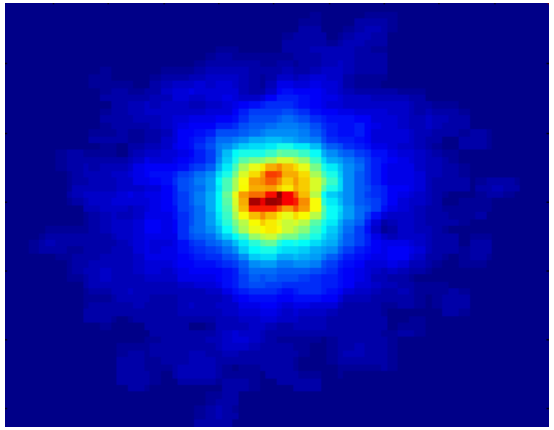
- Fraud Detection
  - **Goal:** Predict fraudulent cases in credit card transactions.
  - **Approach:**
    - Use credit card transactions and the information on its account-holder as attributes.
      - When does a customer buy, what does he buy, how often he pays on time, etc
    - Label past transactions 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.

Attribution: https://www-users.cs.umn.edu/~kumar001/dmbook/index.php#item4
Introduction to Data Mining, 2nd Edition   Tan, Steinbach, Karpatne, Kumar

Classifying Galaxies

Courtesy: http://aps.umn.edu

Early



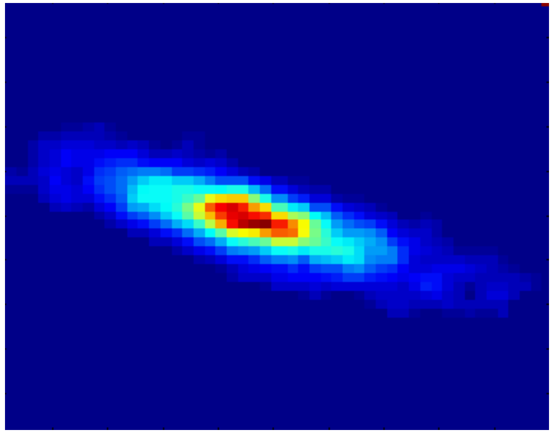
Class:

- **Stages of Formation**

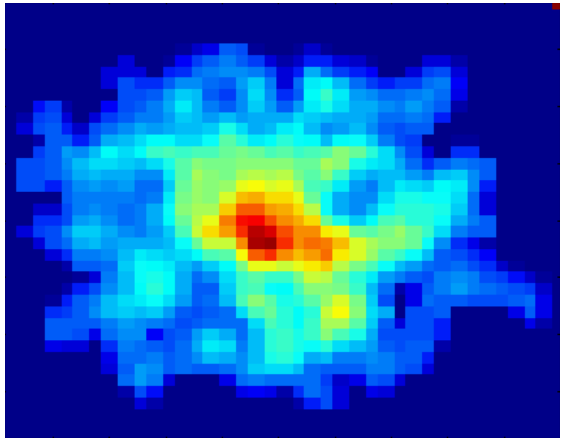
Attributes:

- **Image features,**
- **Characteristics of light waves received, etc.**

Intermediate



Late



Data Size:

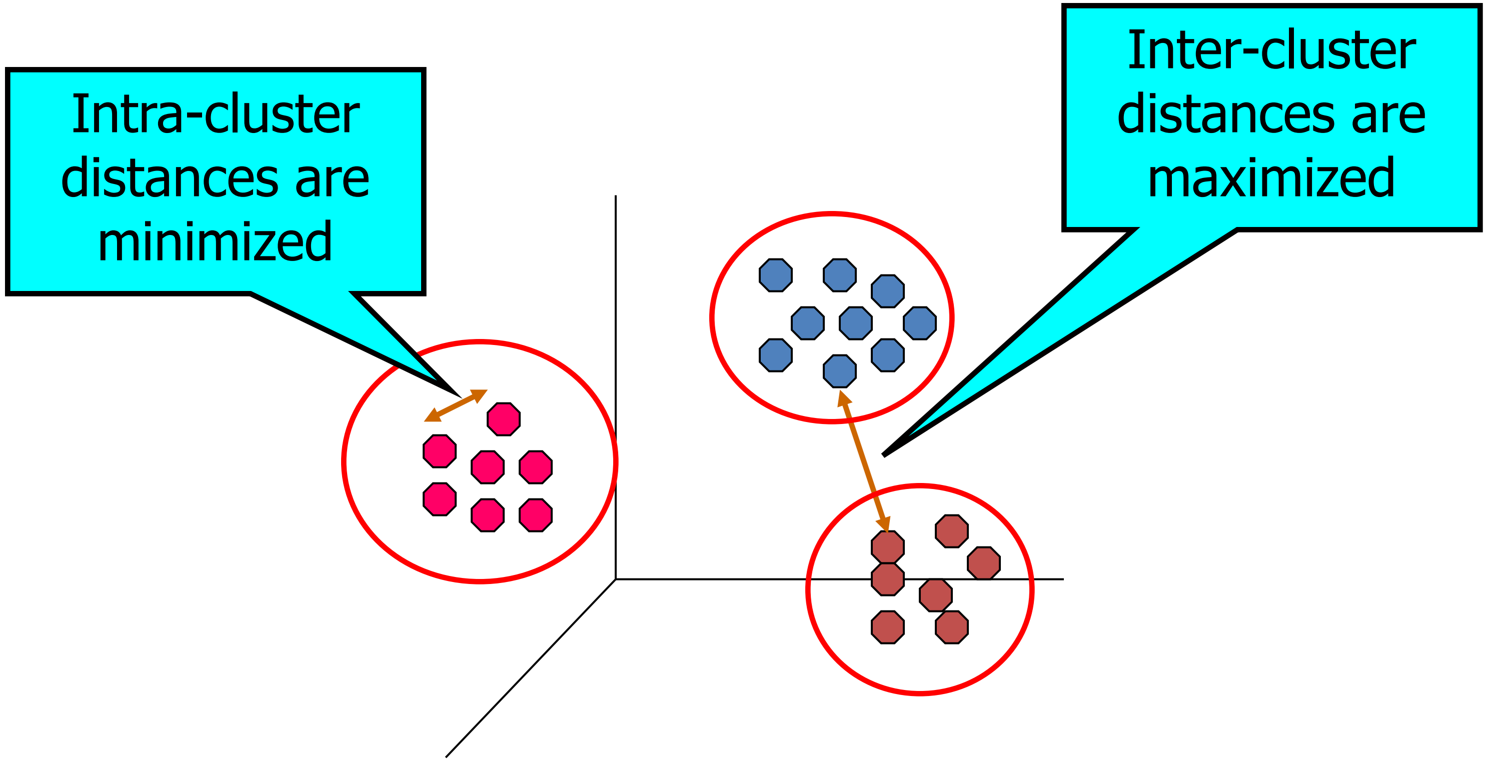
- **72 million stars, 20 million galaxies**
- **Object Catalog: 9 GB**
- **Image Database: 150 GB**

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



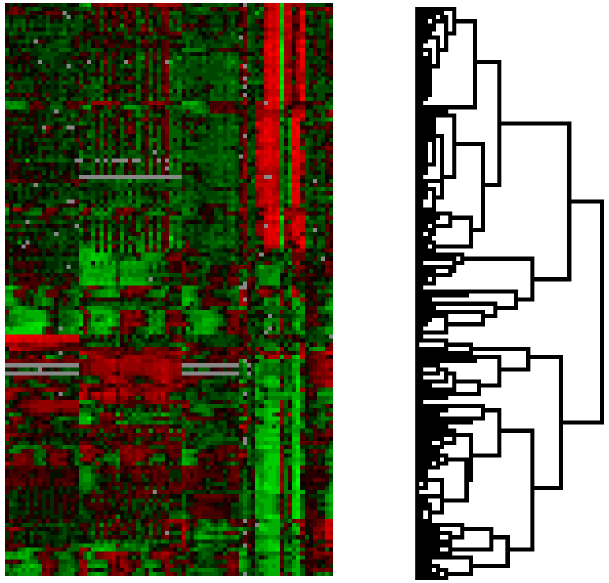
Intra-cluster distances are minimized

Inter-cluster distances are maximized

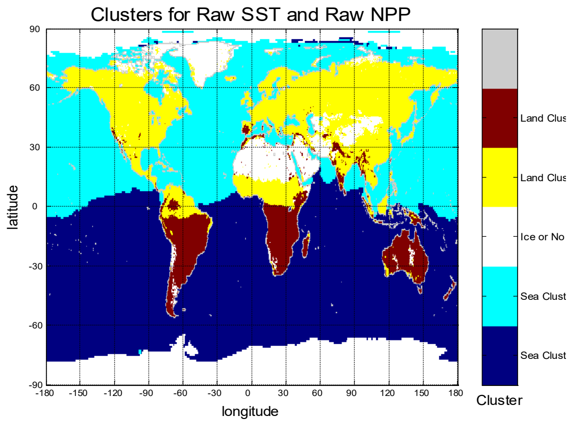
Attribution: https://www-users.cs.umn.edu/~kumar001/dmbook/index.php#item4
Introduction to Data Mining, 2nd Edition Tan, Steinbach, Karpatne, Kumar

Clustering

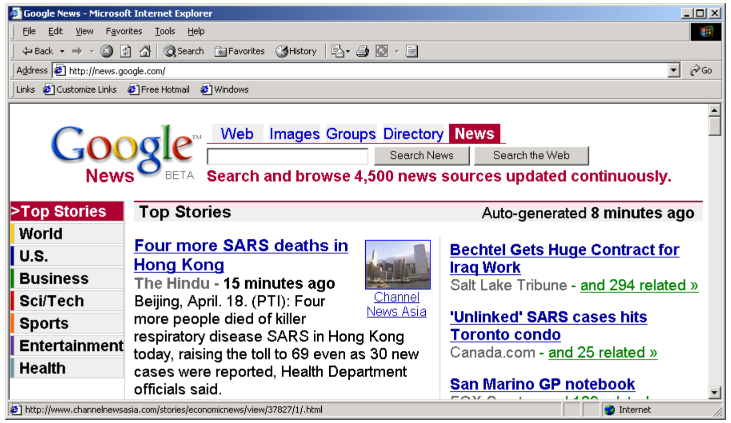
- **Understanding**
  - Custom profiling for targeted marketing
  - Group related documents for browsing
  - Group genes and proteins that have similar functionality
  - Group stocks with similar price fluctuations
- **Summarization**
  - Reduce the size of large data sets



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.



Search and browse 4,500 news sources updated continuously.

Top Stories Auto-generated 8 minutes ago

Four more SARS deaths in Hong Kong

The Hindu - 15 minutes ago

Beijing, April. 18. (PTI): Four more people died of killer respiratory disease SARS in Hong Kong today, raising the toll to 69 even as 30 new cases were reported, Health Department officials said.

Channel News Asia

Bechtel Gets Huge Contract for Iraq Work

Salt Lake Tribune - and 294 related »

'Unlinked' SARS cases hits Toronto condo

Canada.com - and 25 related »

San Marino GP notebook

Attribution: https://www-users.cs.umn.edu/~kumar001/dmbook/index.php#item4

Introduction to Data Mining, 2nd Edition Tan, Steinbach, Karpatne, Kumar

Example

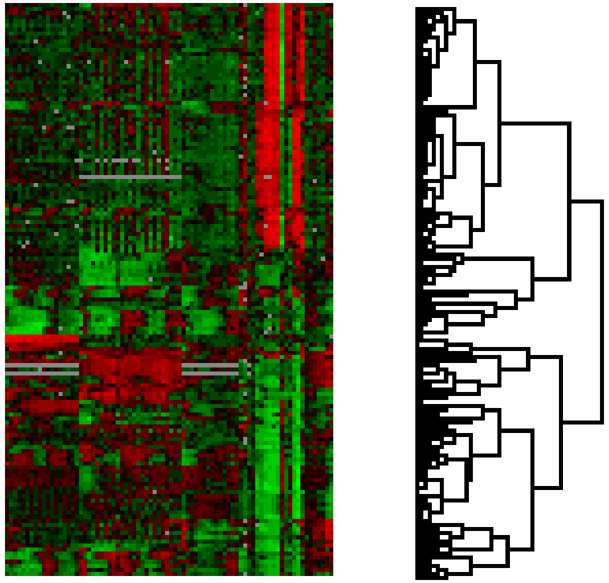
- ABC News
  - https://www.youtube.com/watch?v=f2Kji24833Y

Clustering: Application 1

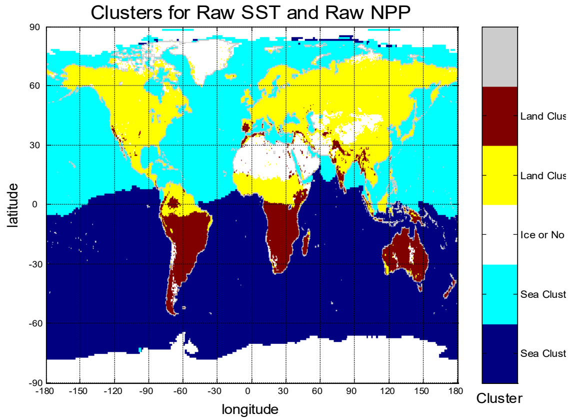
- Market Segmentation:
  - **Goal:** subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
  - **Approach:**
    - Collect different attributes of customers based on their geographical and lifestyle related information.
    - Find clusters of similar customers.
    - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Clustering

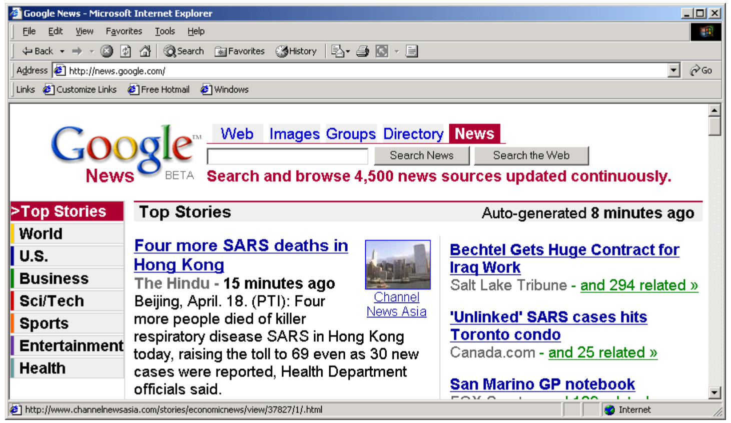
- **Understanding**
  - Custom profiling for targeted marketing
  - Group related documents for browsing
  - Group genes and proteins that have similar functionality
  - Group stocks with similar price fluctuations
- **Summarization**
  - Reduce the size of large data sets



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.



Search and browse 4,500 news sources updated continuously.

>Top Stories
World
U.S.
Business
Sci/Tech
Sports
Entertainment
Health

Top Stories — Auto-generated 8 minutes ago

Four more SARS deaths in Hong Kong
The Hindu - 15 minutes ago
Beijing, April. 18. (PTI): Four more people died of killer respiratory disease SARS in Hong Kong today, raising the toll to 69 even as 30 new cases were reported, Health Department officials said.

Channel News Asia

Bechtel Gets Huge Contract for Iraq Work
Salt Lake Tribune - and 294 related »

'Unlinked' SARS cases hits Toronto condo
Canada.com - and 25 related »

San Marino GP notebook

http://www.channelnewsasia.com/stories/economicnews/view/37827/1/.html

Internet

Attribution: https://www-users.cs.umn.edu/~kumar001/dmbook/index.php#item4
Introduction to Data Mining, 2nd Edition   Tan, Steinbach, Karpatne, Kumar

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}

Attribution: https://www-users.cs.umn.edu/~kumar001/dmbook/index.php#item4
Introduction to Data Mining, 2nd Edition Tan, Steinbach, Karpatne, Kumar

Association Analysis: Applications

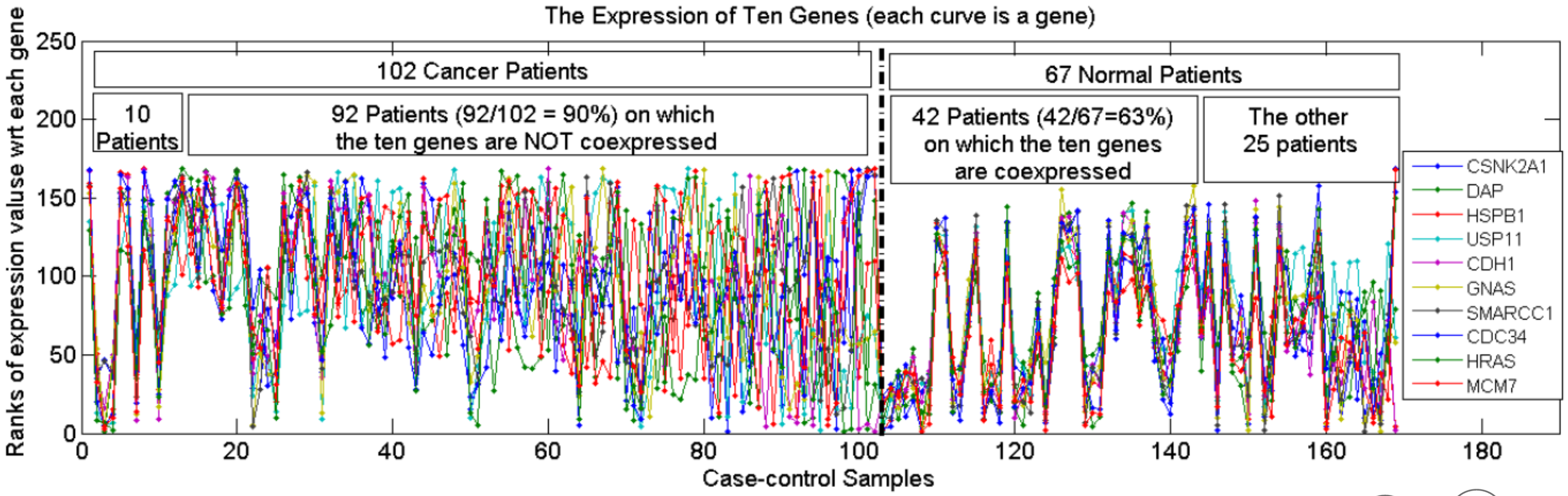
- Market-basket analysis
  - Rules are used for sales promotion, shelf management, and inventory management
- Telecommunication alarm diagnosis
  - Rules are used to find combination of alarms that occur together frequently in the same time period
- Medical Informatics
  - Rules are used to find combination of patient symptoms and test results associated with certain diseases

Attribution: https://www-users.cs.umn.edu/~kumar001/dmbook/index.php#item4
Introduction to Data Mining, 2nd Edition   Tan, Steinbach, Karpatne, Kumar

Association Analysis: Applications

- An Example Subspace Differential Coexpression Pattern from lung cancer dataset

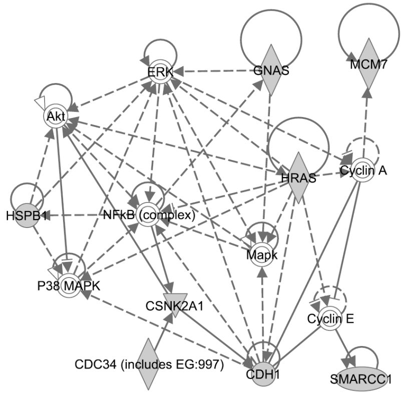
Three lung cancer datasets [Bhattacharjee et al. 2001], [Stearman et al. 2005], [Su et al. 2007]



Enriched with the TNF/NFB signaling pathway which is well-known to be related to lung cancer

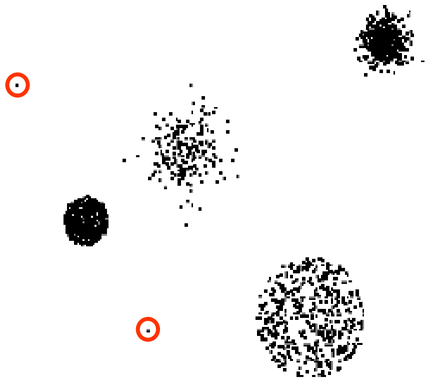
P-value: 1.4*10^{-5} (6/10 overlap with the pathway)

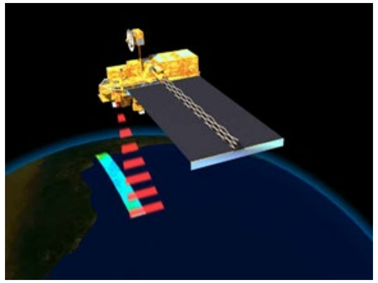
[Fang et al PSB 2010]

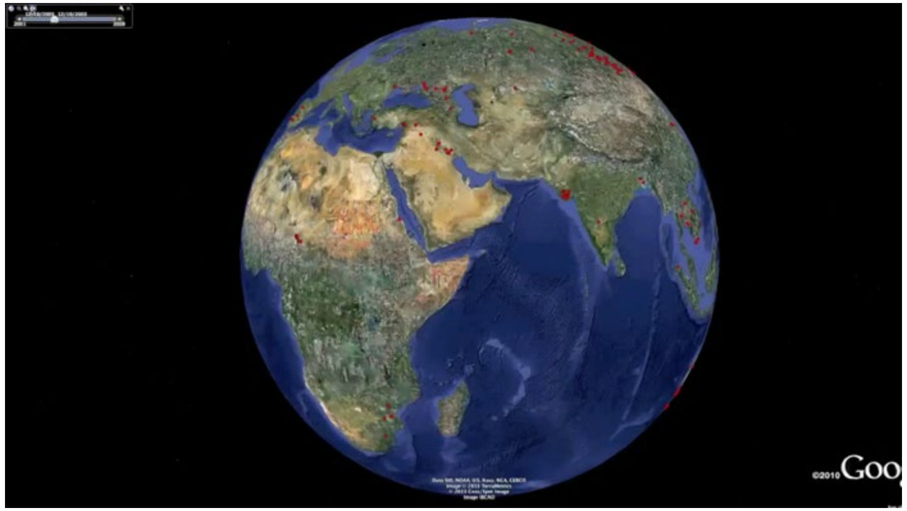


Deviation/Anomaly/Change Detection

- 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.







Attribution: https://www-users.cs.umn.edu/~kumar001/dmbook/index.php#item4
Introduction to Data Mining, 2nd Edition Tan, Steinbach, Karpatne, Kumar

Quiz

- Which of the followings are NOT considered as a data mining task?
  A. Finding clusters of customers
  B. Searching for words in an editor
  C. Predicting fraudulent activities
  D. Finding cities with population greater than a million
  E. Finding associations between events (e.g., gene expressions and diseases)
  F. Finding intersections of polygons
  G. Calculating derivatives of a complex function

Quiz

- Which of the followings are NOT considered as a data mining task?
  A. Finding clusters of customers
  B. Searching for words in an editor
  C. Predicting fraudulent activities
  D. Finding cities with population greater than a million
  E. Finding associations between events (e.g., gene expressions and diseases)
  F. Finding intersections of polygons
  G. Calculating derivatives of a complex function

Non-trivial extraction of **implicit**, previously **unknown** and potentially **useful** information from **data**

Outline

- Self-introduction
  - Share a few sentences about yourself
  - Background (major), year, experience with data mining
- Syllabus
  - Topics and schedule
  - Grading and policies
- Why data mining?
- Why spatial data mining? What is special?