

EECS 568: Introduction to Data Mining

Zijun Yao

Assistant Professor, EECS Department

University of Kansas

Class and Office Hour

- **Instructor:** Zijun Yao, Eaton Hall 2048 (Office)
- **Class:** 12:30pm - 1:45pm Mon/Fri, Learned Hall 3151
- **Office hour:** 11:00am - 12:00pm Mon/Fri or by appointment
- **E-mail:** zyao@ku.edu
 - Recommended Subject Line: **EECS568 <Your Full Name> <Brief headline>**
- **Course web:** <https://canvas.ku.edu/>

2021 *This Is What Happens In An Internet Minute*



The Oil of Digital Era

“The world’s most valuable resource is no longer oil, but data.”

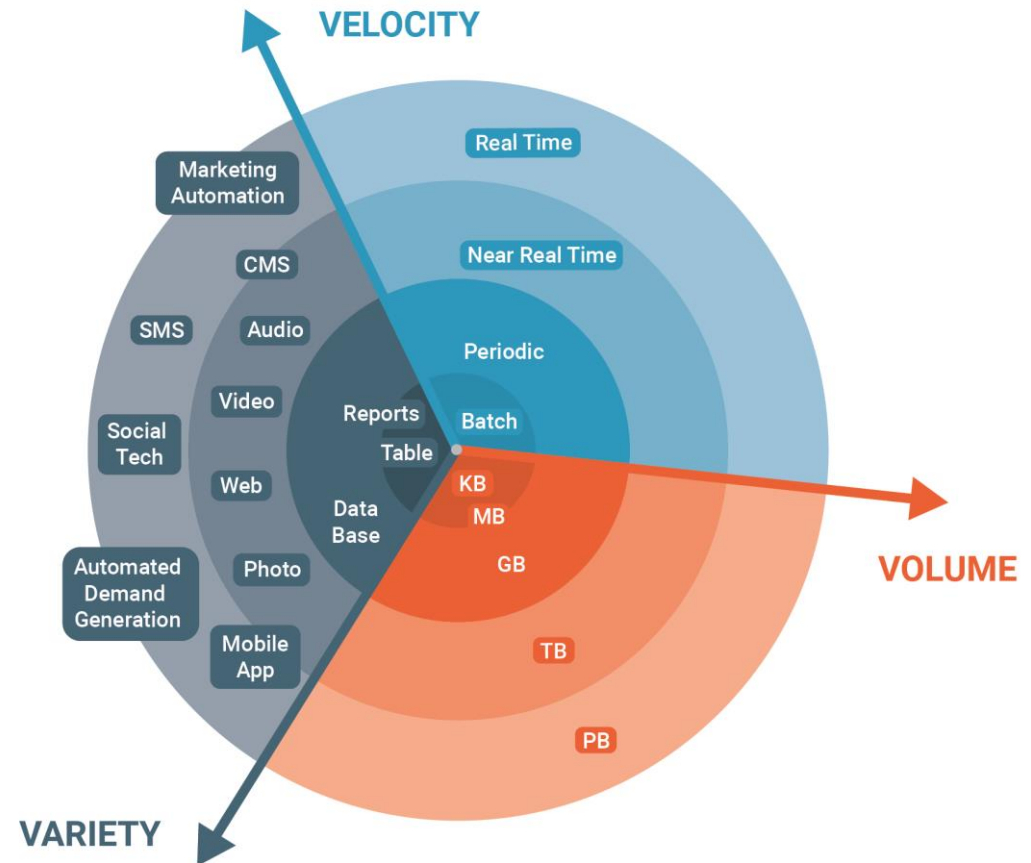
The Economist, May 2017



What is Big Data, exactly?

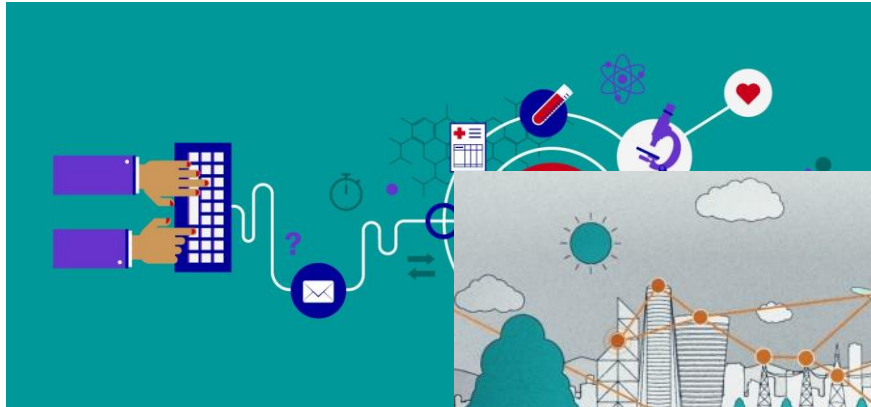
- Data in all forms and sizes is being generated faster than ever before
- **Volume** – terabytes and petabytes data
- **Variety** – structured and unstructured data
- **Velocity** – real-time and streaming data

—• The 3 V's of Big Data



Turning Data into Intelligence

- Use vast amounts of data to provide actionable insights for better and agile decision making.

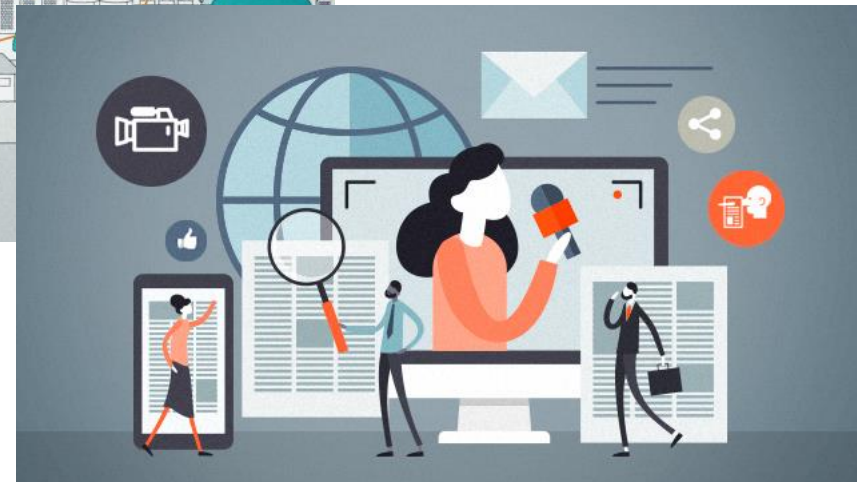


Health analytics



Mobile computing

Media recommendation



Decision Making in Business

- Amazon: which product should be stocked in which warehouse at what time?
- Salesforce: which customers to target, and how to retain them?
- Netflix: what movie should be invested, what movie to recommend?

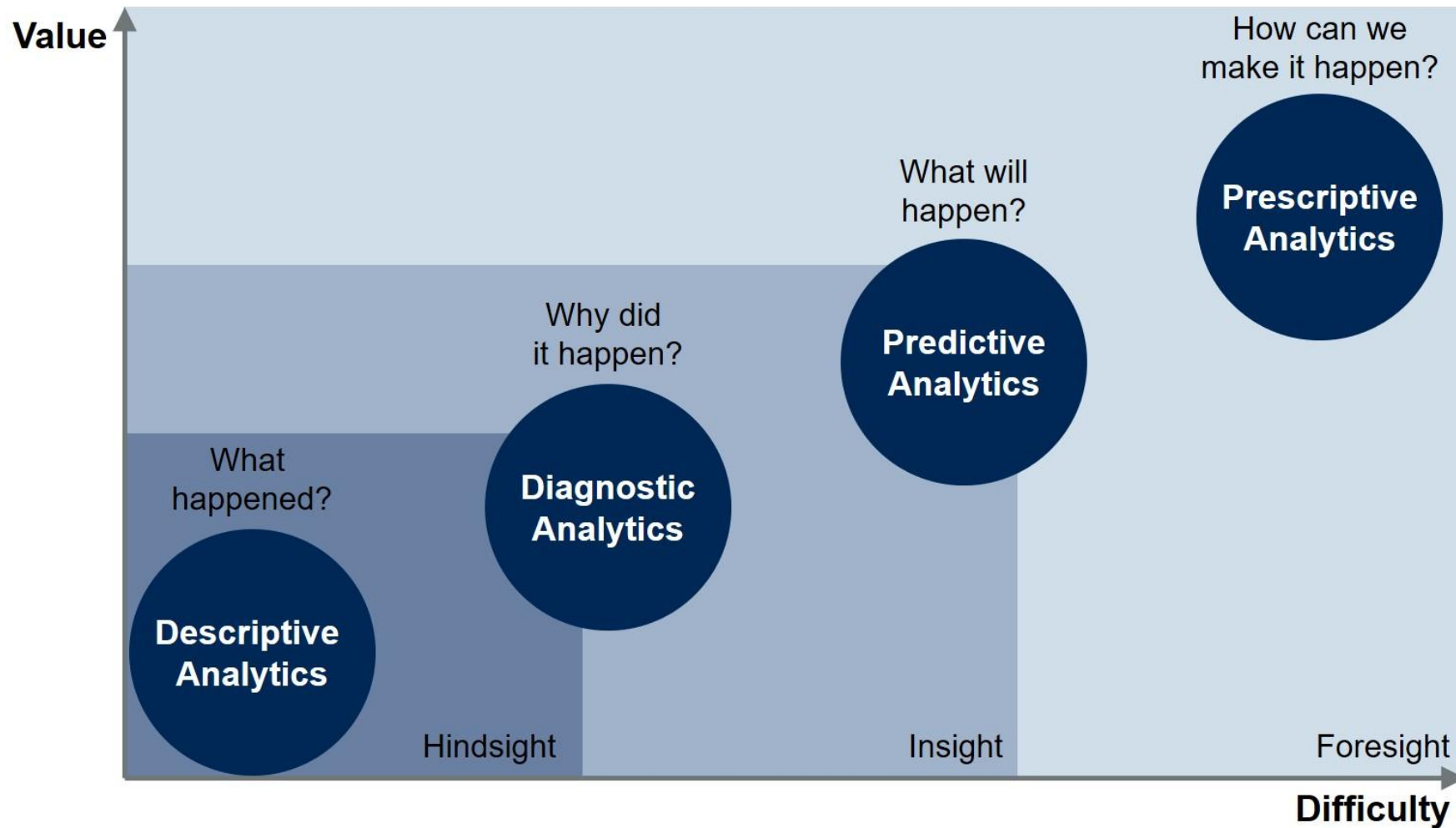


What is Data Mining

- Data mining is the process of automatically discovering **useful information** in **large data repositories**.
- It is a technology that blends traditional data analysis methods with sophisticated **algorithms** for processing **large volumes** of data.
- It has opened up exciting opportunities for exploring and analyzing **new types of data** and for analyzing old types of data in new ways.



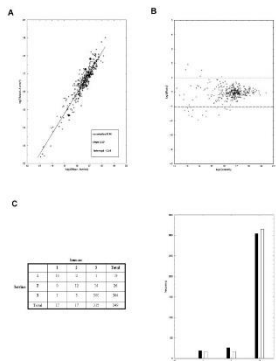
Advanced Analytic Defined



Data Miners are Like Doctors



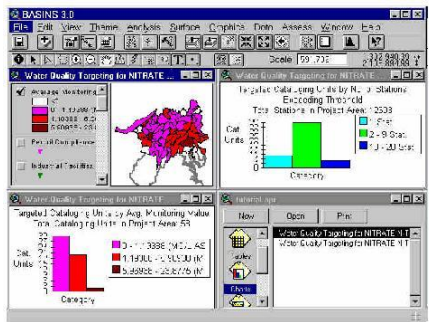
Very Often, No
Standardized Solutions



Data Characteristics



You Symptoms



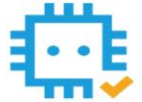
Data Mining Techniques



Medical Devices

Career Future

- Data Scientist: The Sexiest Job of the 21st Century – Harvard Business Review
- Shortage of Data Scientist goes to 250,000 in 2020



The Data Scientist Shortage in 2020

Demand for Data Scientists

Job Listings

37%

Year on Year Growth
in 2019

Job Ranking

#3

Ranking For
Top Jobs in 2020

Salaries

14%

Average Salary
Increase

Hiring

67%

Companies
Expanding the Data
Science Team

Analytics And Data Science

Data Scientist: The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data. by Thomas H. Davenport and DJ Patil

From the Magazine (October 2012)

Analytics And Data Science

Is Data Scientist Still the Sexiest Job of the 21st Century?

by Thomas H. Davenport and DJ Patil

July 15, 2022

<https://quanthub.com/data-scientist-shortage-2020/>

<https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century>

<https://hbr.org/2022/07/is-data-scientist-still-the-sexiest-job-of-the-21st-century>

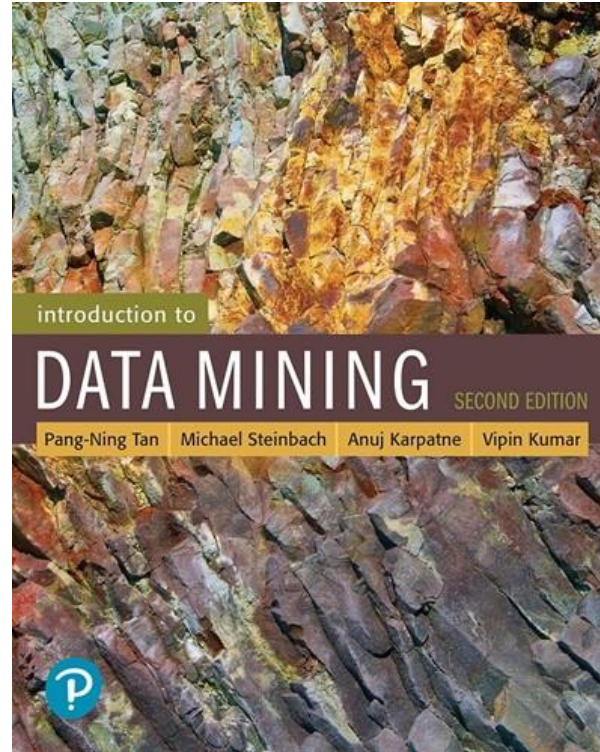
Course Coverage

- Fundamental concepts in data mining
- Computational models and algorithms for various tasks
 - Classification, regression, anomaly detection
- How data mining is used in real-world application
 - Hands-on (eg., computer vision, natural language processing, recommender systems)
 - Analyzing business cases

Course Prerequisites

- **Algorithm skills** - basic algorithms and data structures in python (EECS 168 Programming I or equivalents)
- **Math skills** - linear algebra, probability, and statistics (MATH 526 Applied Mathematical Statistics I or equivalents)
- **Programming** with Python

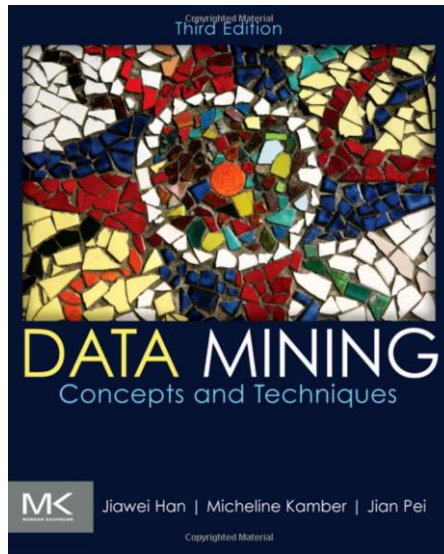
Textbook



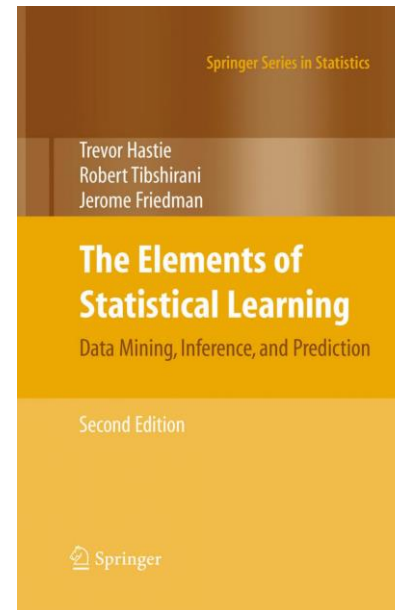
Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar
ISBN: 0133128903, 2018.
(1st Edition also works)

Textbooks (optional)

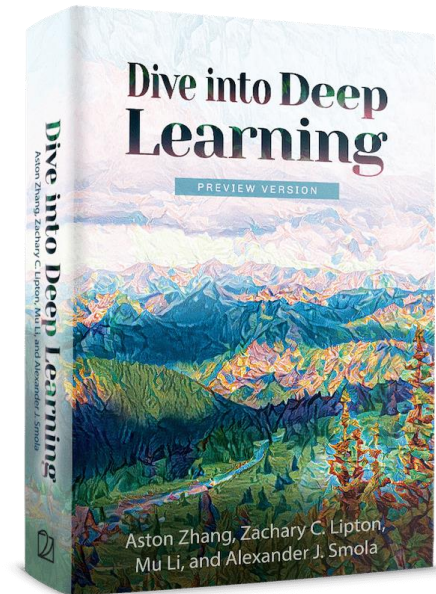
- Use following books as the primary reference



available at [Link](#)



available at
<https://hastie.su.domains/ElemStatLearn/>



available at
<https://d2l.ai/>

Syllabus and Course Schedule

- Course schedule available under [Syllabus](#) section at Canvas (<https://canvas.ku.edu/>)
- Lecture slides and tutorial will be posted under [Course Schedule](#) section at Canvas
- Assignments, exams and other course work will be posted under [Assignments](#) section at Canvas
- You should check [Announcements](#) frequently to remain updated.

Course Components

- Lectures at 12:30am - 1:45am on Monday and Friday
 - Show up and actively engage
- Slides and tutorials
 - Slides and tutorial codes will be posted before lecture
- Readings
 - Read material before the class
- Assignments and exams
 - Submit assignments on time.
- Programming Project
 - Program ML algorithms on real-world data

Grading Policy

1.	Attendance	10%
2.	Assignments	30%
3.	Project	20%
4.	Exam I	20%
5.	Exam II	20%
Total		100%

- **Assignments:** Late submissions will receive penalty
- **Exams:** There will be no make-up exams
- **Course project:** Each student will program certain tasks using Python
- Academic integrity!
- The final grade is based on a curve

Course Expectation

- Develop a strong vocabulary and understanding of data mining terminology
- Communicate with confidence among developers and consultants
- Hands-on experience through exercises using application data
- Exposure to various data mining tools
- Approach business/research problems analytically by identifying opportunities to derive actionable insights from data

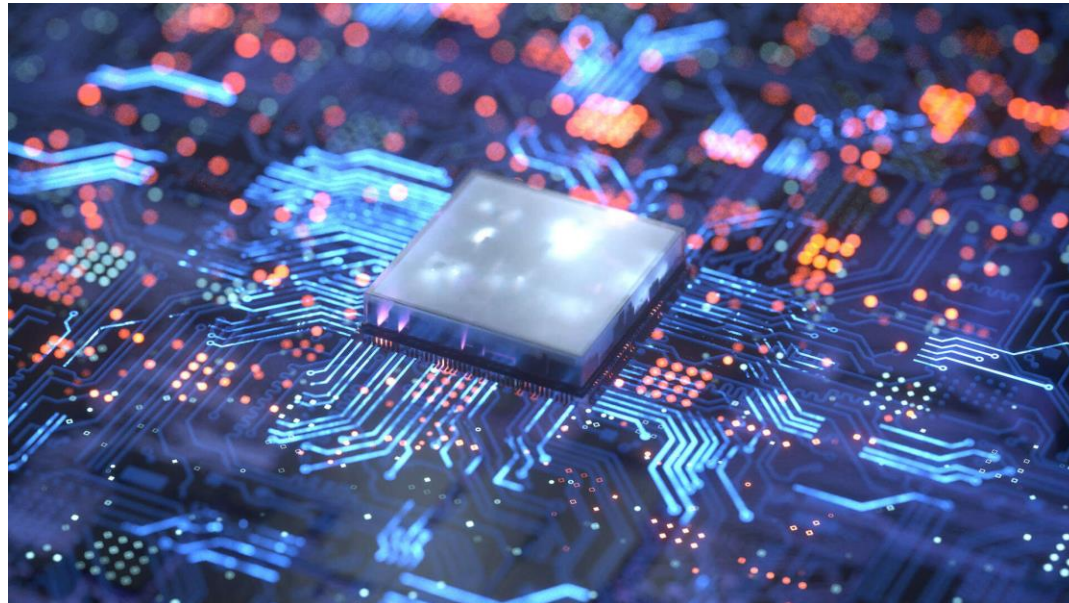
Three Views of Technology - Airplane View

- Airplane View: 30,000 feet above the ground
 - “Above the clouds”, “global view”
 - Technology as a black box, simple



Three Views of Technology - On-the-ground view

- On-the-ground view:
 - Very close view – “bits & bytes”
 - Technology is complex; lots of “messy” details

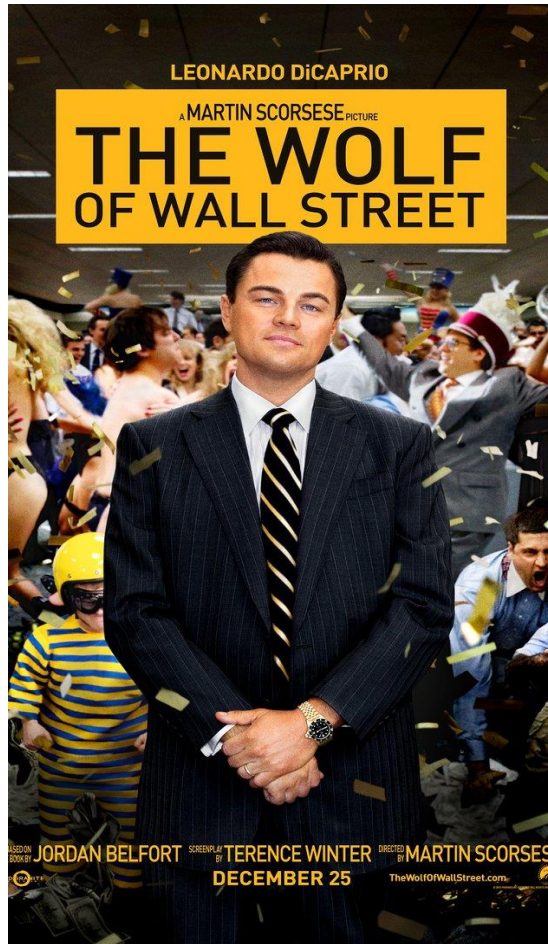


Three Views of Technology - Helicopter View

- Helicopter View:
 - “Below the clouds” view
 - Technology is NOT a black box (will see some “mess”), but no struggle with bits & bytes



Motivation of Data Mining?




Life in 1990's





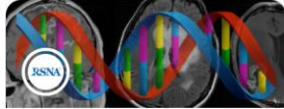













Life in 2010's

Kaggle

- The world's largest data science community
- Datasets are published
- Techniques are shared
- Challenges are posted
- There are huge demand of transforming data into productivity!

🕒 Active Competitions Hotness ▾ 

 Optiver Realized Volatility Prediction  Apply your data science skills to make fin... Featured Code Competition · 2544 Teams \$100,000 a month to go	 NFL Health & Safety - Helmet Assignment  Segment and label helmets in video foota... Featured Code Competition · 122 Teams \$100,000 2 months to go	 RSNA-MICCAI Brain Tumor Radiogenomic Classification  Predict the status of a genetic biomarker i... Featured Code Competition · 773 Teams \$30,000 2 months to go	 LearnPlatform COVID-19 Impact on Digital Learning  Use digital learning data to analyze the im... Analytics \$20,000 a month to go
 G2Net Gravitational Wave Detection  Find gravitational wave signals from binar... Research 769 Teams \$15,000 a month to go	 chaii - Hindi and Tamil Question Answering  Identify the answer to questions found in ... Research Code Competition · 212 Teams \$10,000 3 months to go	 Lux AI  Gather the most resources and survive th... Featured Simulation Competition · 153 Teams \$10,000 3 months to go	 Google Landmark Recognition 2021  Label famous, and not-so-famous, landm... Research Code Competition · 76 Teams Swag a month to go

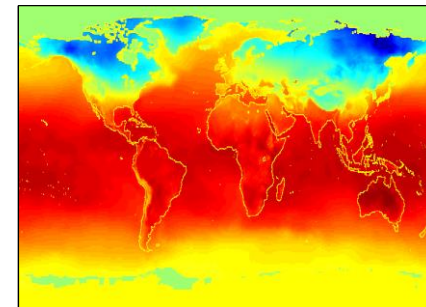
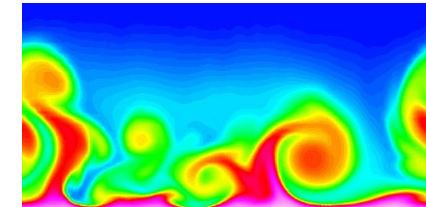
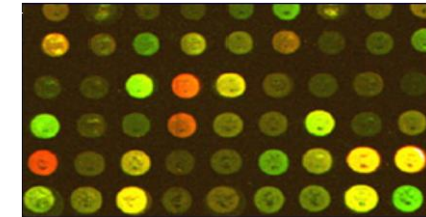
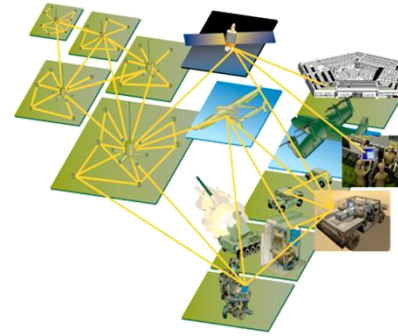
Why Mine Data? Commercial Viewpoint

- Lots of data is being collected and warehoused
 - Web data, e-commerce
 - purchases at department/grocery stores
 - Bank/Credit Card transactions
- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
 - Provide better, customized services customer retention



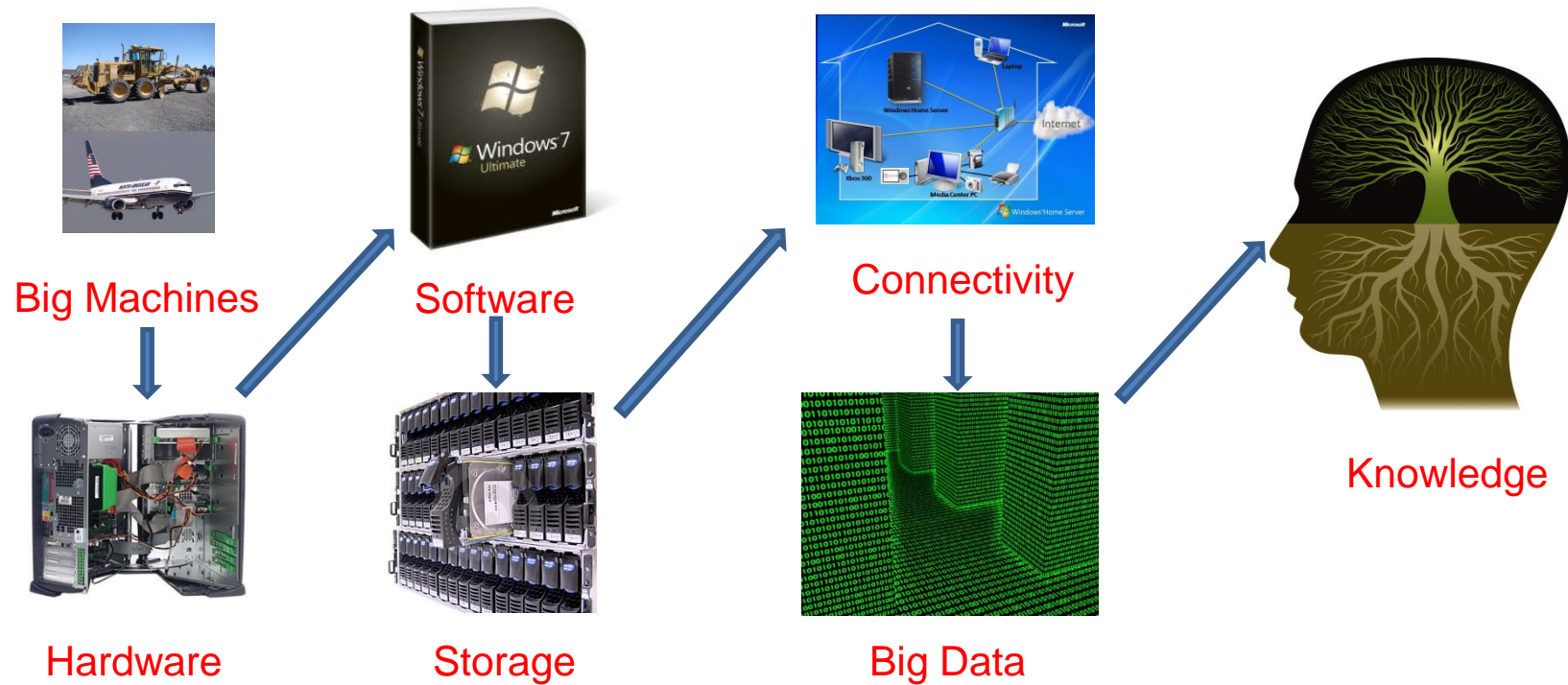
Why Mine Data? Scientific Viewpoint

- Data collected and stored at enormous speeds (GB/hour)
 - remote sensors on a satellite
 - telescopes scanning the skies
 - microarrays generating gene expression data
 - scientific simulations generating terabytes of data
- Traditional techniques infeasible for raw data
- Data mining may help scientists
 - in classifying and segmenting data
 - in Hypothesis Formation



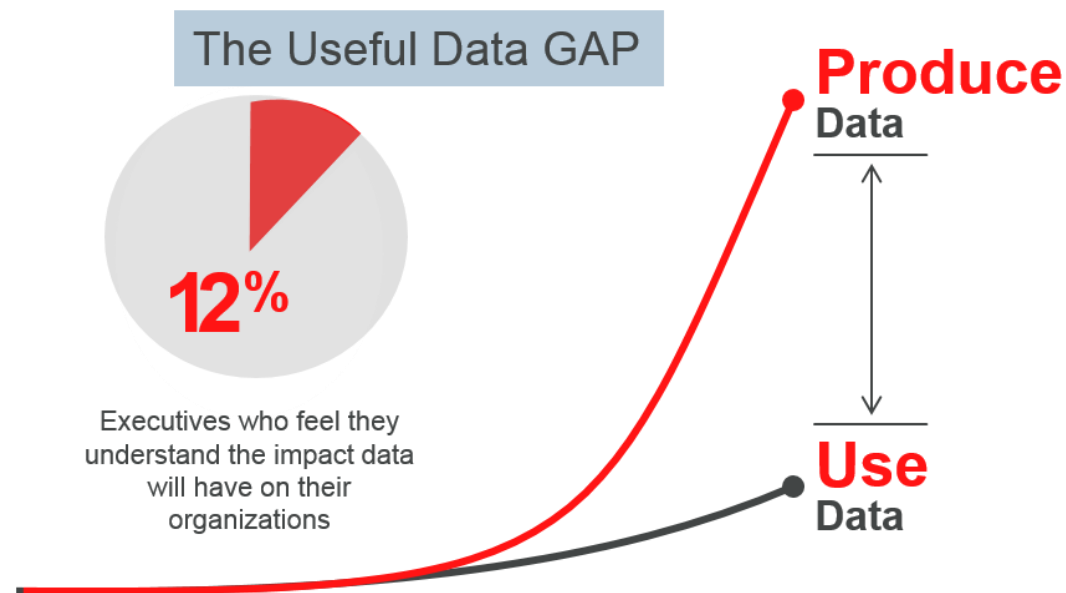
Why mine data: historical view?

- Moved from things we could touch and see; Fell to things that are conceptual.



Why Do We Need Data Mining ?

- Leverage organization's data assets
 - Only a small portion (typically - 5%-10%) of the collected data is ever analyzed
 - Data that may never be analyzed continues to be collected, at a great expense, out of fear that something which may prove important in the future is missing
 - Growth rates of data precludes traditional “manually intensive” approaches



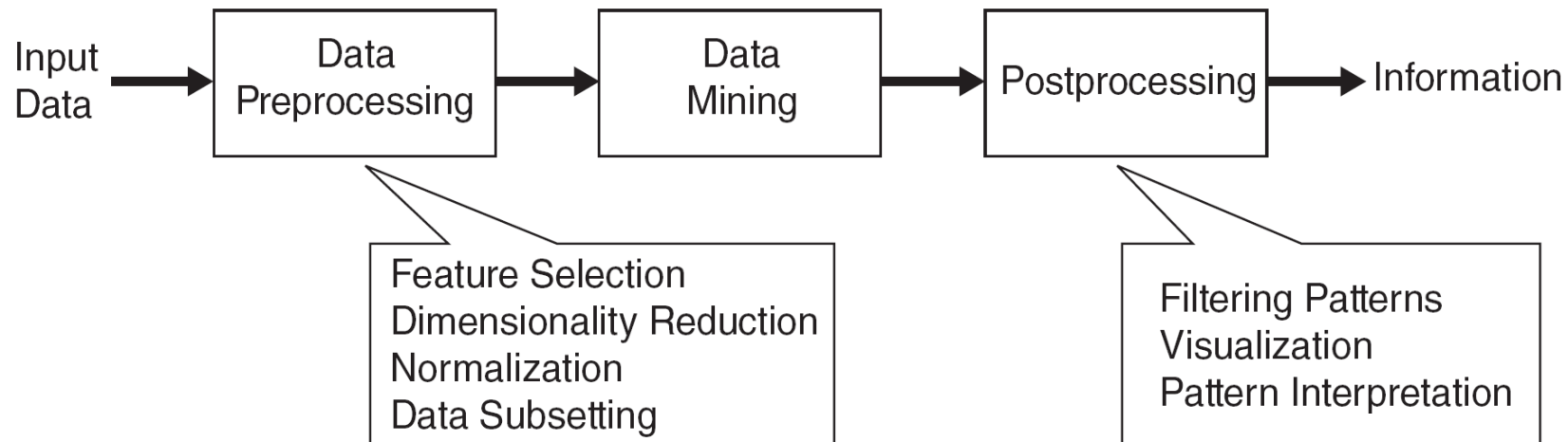
Analytics Provides a Solid Foundation for Career or Research in Many Fields

A word cloud featuring various industries and sectors, each in a different color and tilted at an angle. The words are: Life Sciences (yellow), Education (purple), Financial Services (green), Automotive (blue), Energy (red), Insurance (dark blue), Transportation & Logistics (black), Information Technology (black), Entertainment (grey), Consulting (light blue), Telecommunications (red), Government (dark grey), Food Processing (pink), Electronics (grey), Manufacturing (pink), Retail Trade (green), and Banking (orange).

The best thing about analytics is that it lets you play in everyone's sandbox

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



What is (not) Data Mining?

What is not Data Mining? (facts)

- Look up phone number in phone directory
- Check the dictionary for the meaning of a word

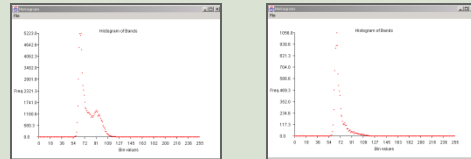
Rely on traditional computer science techniques and obvious features of the data to create index structures for efficiently organizing and retrieving information

What is Data Mining? (Intelligence)

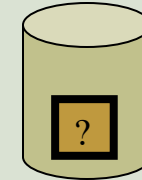
- Supermarkets place beer and diapers together to boost sales
- Mobile services create personalized advertising and retain customers who are looking to change the vendors

Data Mining: Confluence of Multiple Disciplines

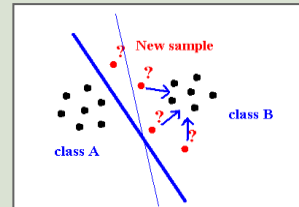
Statistics



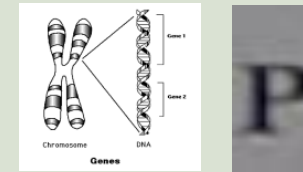
Database Techniques



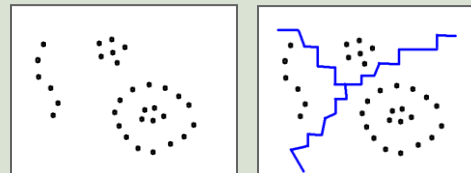
Machine Learning



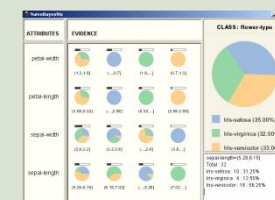
Optimization Techniques



Pattern Recognition



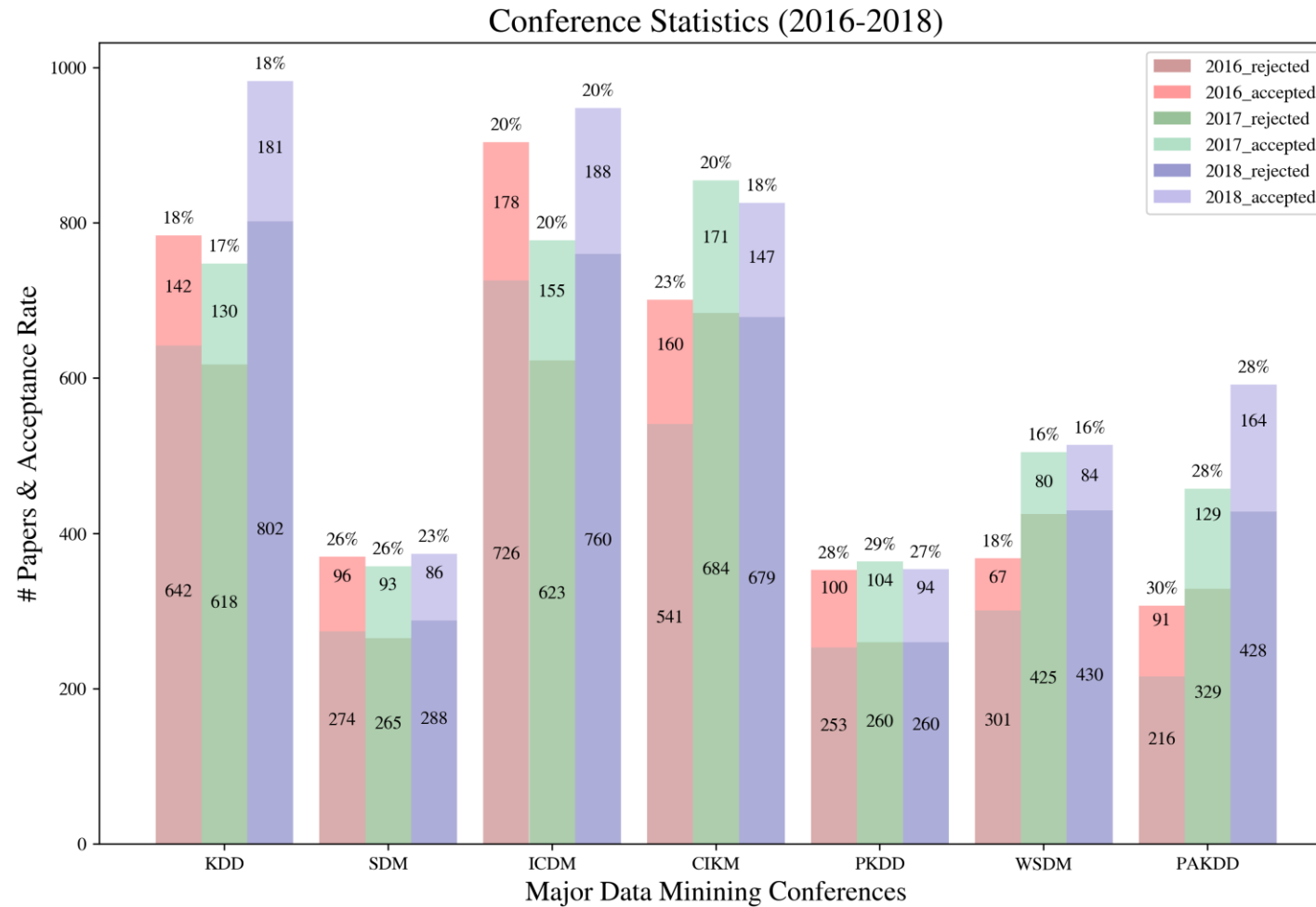
Visualization



Main Forums in Data Mining

- Conferences:
 - The birth of data mining/KDD: 1989 IJCAI Workshop on Knowledge Discovery in Databases
 - 1991-1994 Workshops on Knowledge Discovery in Databases
 - 1995 – date: International Conferences on Knowledge Discovery and Data Mining ([KDD](#))
 - 2001 – date: [IEEE ICDM](#) and [SIAM-DM \(SDM\)](#)
 - Other conferences, incl. PAKDD (since 1997) & PKDD (since 1997) & WSDM (since 2008)
- Journals:
 - [Data Mining and Knowledge Discovery](#) (DMKD, since 1997)
 - Knowledge and Information Systems (KAIS, since 1999)
 - [IEEE Trans. on Knowledge and Data Engineering \(TKDE\)](#)
 - Many others, incl. TPAMI, TKDD, ML, MLR, VLDBJ ...

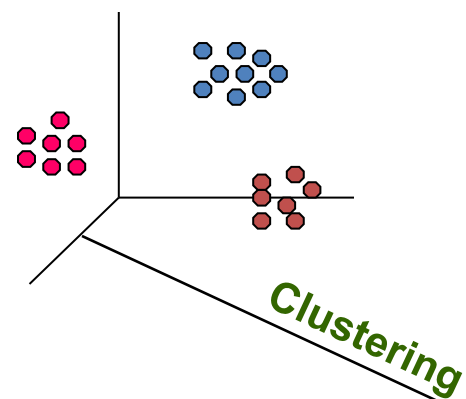
Data Mining Conferences



Data Mining Tasks

- Predictive Tasks
 - Predict the value of a particular attribute based on the values of other attributes.
 - The attribute to be predicted is commonly known as the **target** or **dependent variable**, while the attributes used for making the prediction are known as the **explanatory** or **independent variables**.
- Descriptive Tasks
 - Derive patterns (correlations, trends, clusters, trajectories, and anomalies) that summarize the underlying relationships in data.

Data Mining Tasks ...



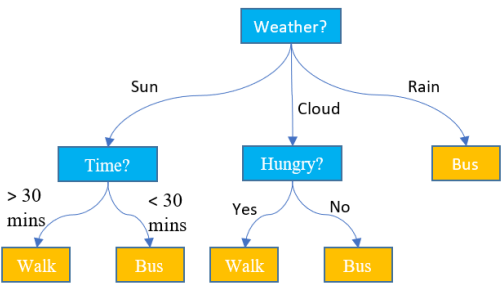
Data

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

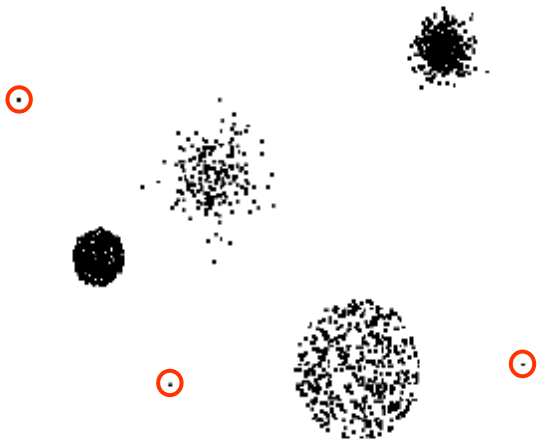
Association Analysis



Predictive Modeling



Anomaly Detection

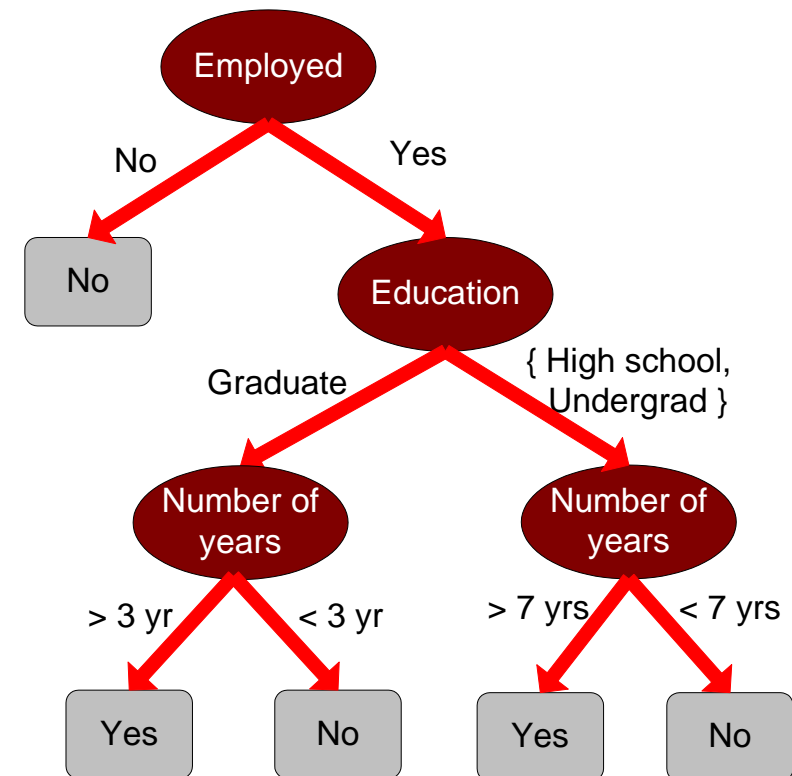


Predictive Modeling: Classification

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

Features				Class
<i>Tid</i>	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

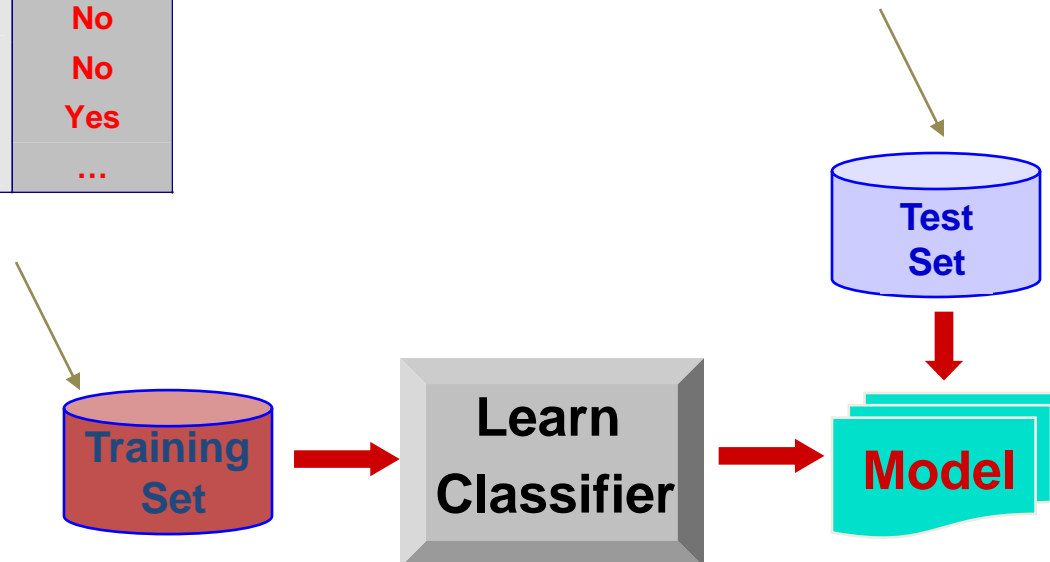


Classification Example

categorical *categorical* *quantitative* *class*

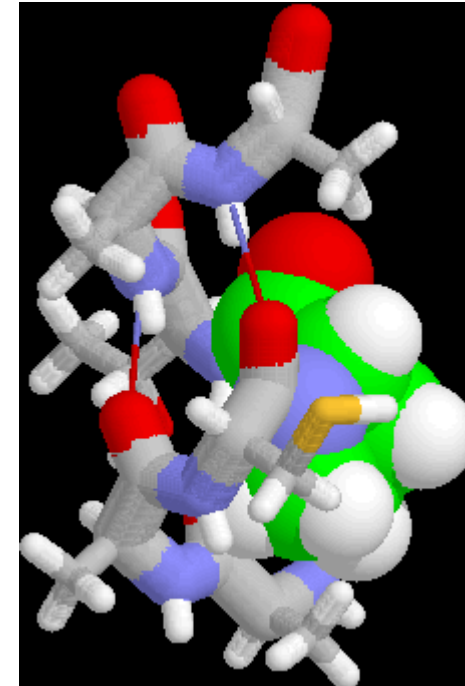
<i>Tid</i>	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
...

<i>Tid</i>	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

- Predicting tumor cells as benign or malignant
- Classifying credit card transactions as legitimate or fraudulent
- Categorizing news stories as finance, weather, entertainment, sports, etc
- Classifying secondary structures of protein as alpha-helix, beta-sheet, or random coil
- Identifying intruders in the cyberspace



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.

Classification: Application 2

- Churn prediction for telephone customers
 - **Goal:** To predict whether a customer is likely to be lost to a competitor.
 - **Approach:**
 - Use detailed record of transactions with each of the past and present customers, to find attributes.
 - How often the customer calls, where he calls, what time-of-the day he calls most, his financial status, marital status, etc.
 - Label the customers as loyal or disloyal.
 - Find a model for loyalty.

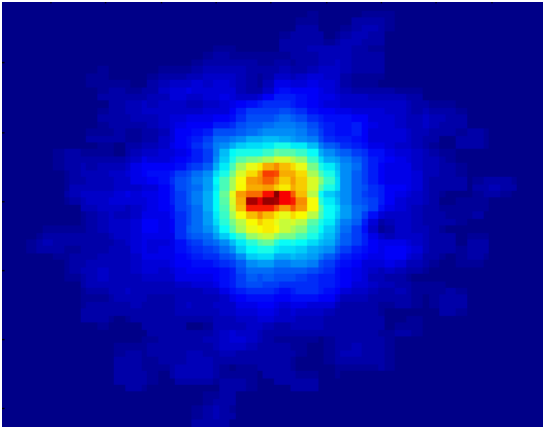
From [Berry & Linoff] Data Mining Techniques, 1997

Classification: Application 3

- Sky Survey Cataloging
 - **Goal:** To predict class (star or galaxy) of sky objects, especially visually faint ones, based on the telescopic survey images (from Palomar Observatory).
 - 3000 images with 23,040 x 23,040 pixels per image.
 - **Approach:**
 - Segment the image.
 - Measure image attributes (features) - 40 of them per object.
 - Model the class based on these features.
 - Success Story: Could find 16 new high red-shift quasars, some of the farthest objects that are difficult to find!

Classifying Galaxies

Early



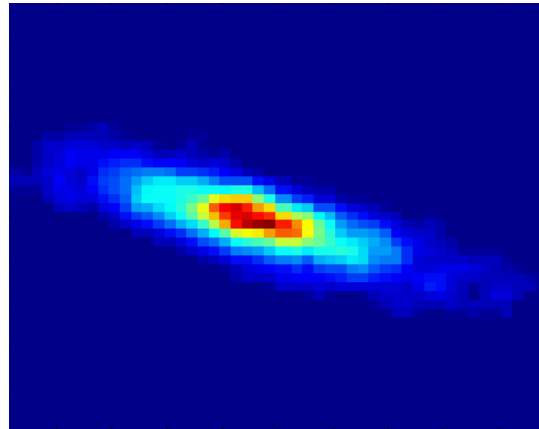
Class:

- Stages of Formation

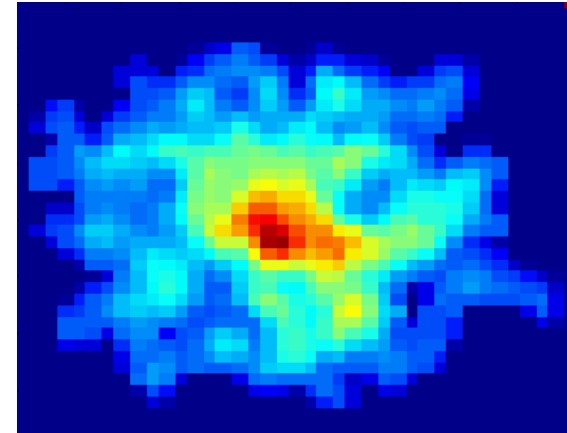
Attributes:

- 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

<http://aps.umn.edu>

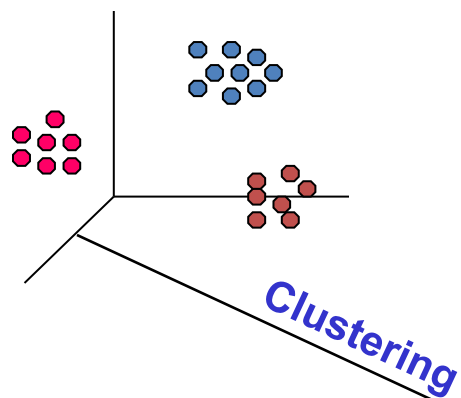
Classification Techniques

- Base Classifiers
 - Decision Tree based Methods
 - Rule-based Methods
 - Nearest-neighbor
 - Neural Networks
 - Naïve Bayes and Bayesian Belief Networks
 - Support Vector Machines
- Ensemble Classifiers
 - Boosting, Bagging, Random Forests

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

Data Mining Tasks ...



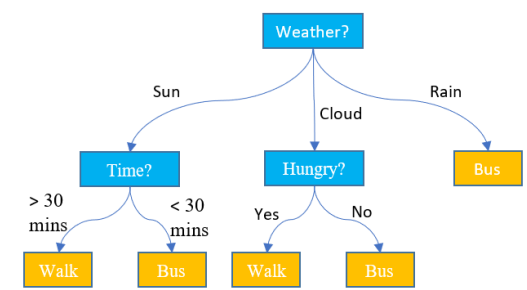
Data

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

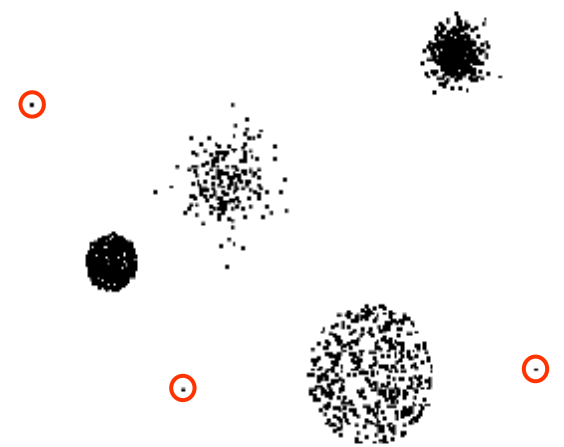
Association Analysis



Predictive Modeling

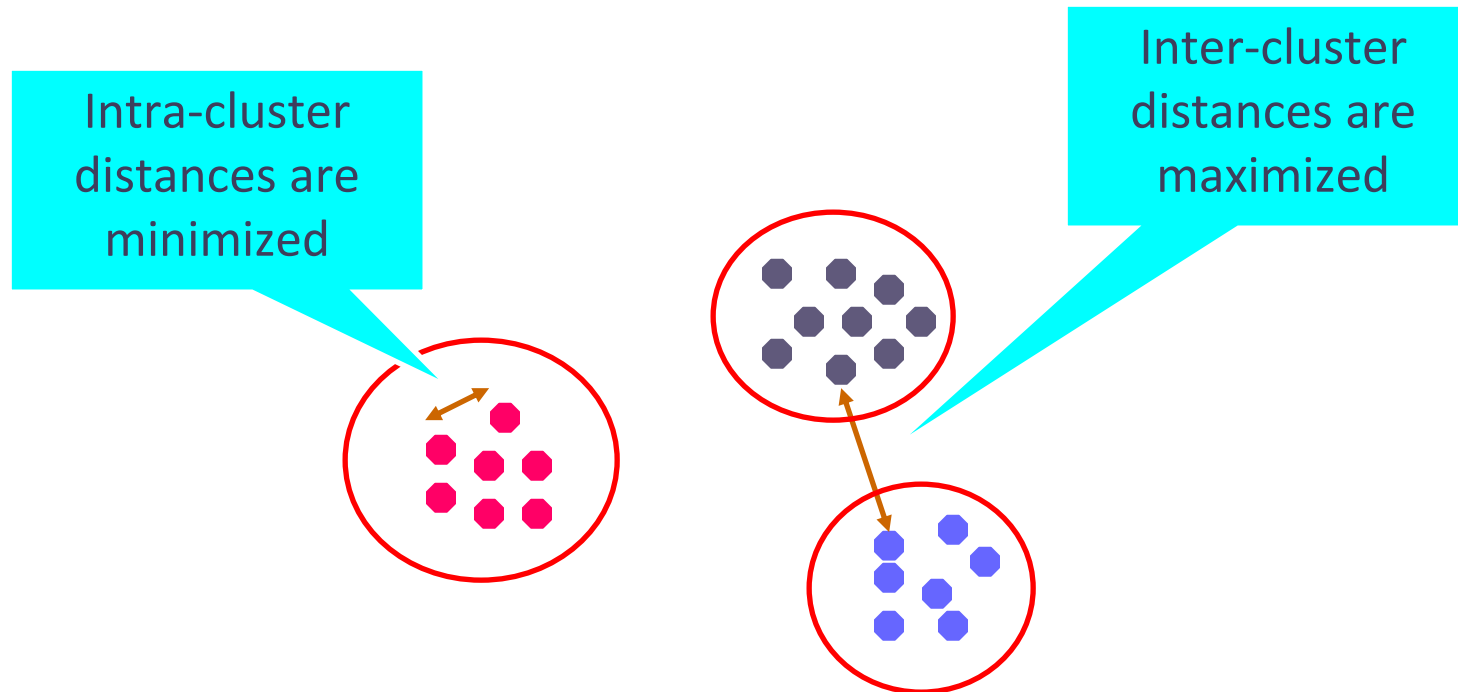


Anomaly Detection



Clustering

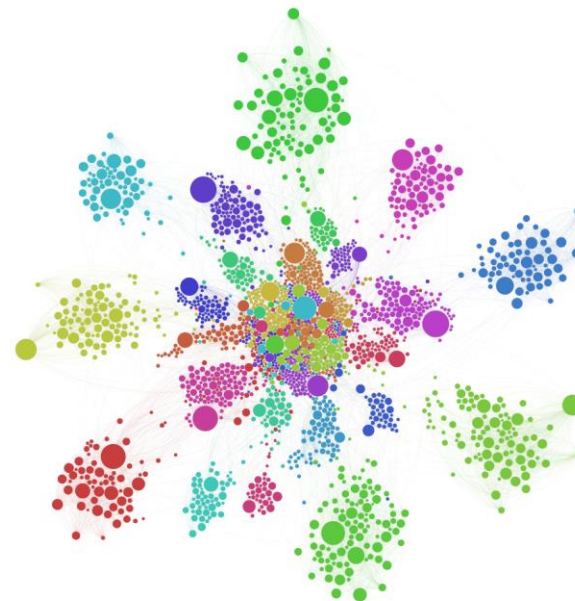
- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



Applications of Cluster Analysis

- Understanding
 - 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
 - Visualizing High Dimensional Clusters

	<i>Discovered Clusters</i>	<i>Industry Group</i>
1	Applied-Matl-DOWN,Bay-Network-DOWN,3-COM-DOWN,Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN,DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN,Micron-Tech-DOWN,Texas-Inst-Down,Tellabs-Inc-Down,Natl-Semiconduct-DOWN,Oracl-DOWN,SGI-DOWN,Sun-DOWN	Technology1-DOWN
2	Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN,ADV-Micro-Device-DOWN,Andrew-Corp-DOWN,Computer-Assoc-DOWN,Circuit-City-DOWN,Compaq-DOWN,EMC-Corp-DOWN,Gen-Inst-DOWN,Motorola-DOWN,Microsoft-DOWN,Scientific-Atl-DOWN	Technology2-DOWN
3	Fannie-Mae-DOWN,Fed-Home-Loan-DOWN,MBNA-Corp-DOWN,Morgan-Stanley-DOWN	Financial-DOWN
4	Baker-Hughes-UP,Dresser-Inds-UP,Halliburton-HLD-UP,Louisiana-Land-UP,Phillips-Petro-UP,Unocal-UP,Schlumberger-UP	Oil-UP



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: 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.

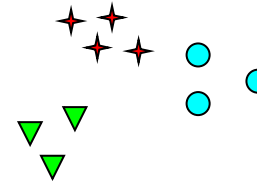
What is not Cluster Analysis?

- Simple segmentation
 - Dividing students into different registration groups alphabetically, by last name
- Results of a query
 - Groupings are a result of an external specification
 - Clustering is a grouping of objects based on the data
- Supervised classification
 - Have class label information
- Association Analysis
 - Local vs. global connections

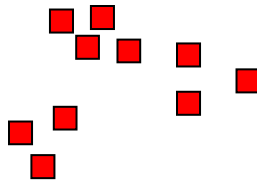
Notion of a Cluster can be Ambiguous



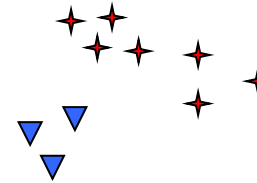
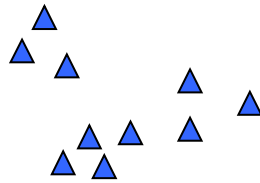
How many clusters?



Six Clusters



Two Clusters



Four Clusters

