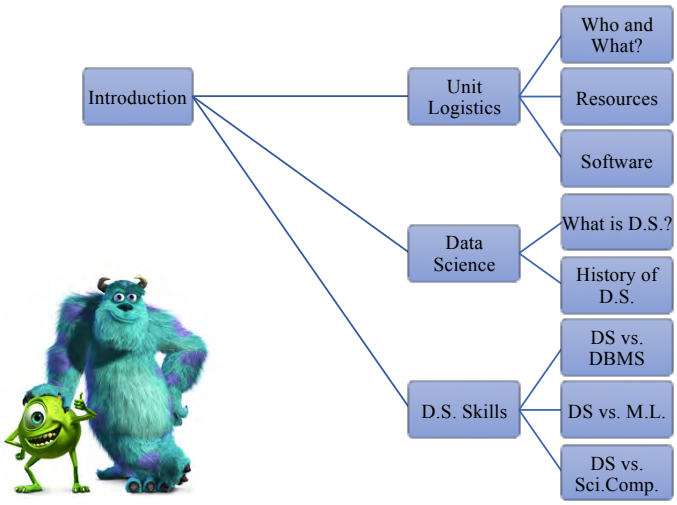
Lecture Notes on

Advanced Data Analytics

Module 01: Data Science

Gang Li
School of Information Technology
Deakin University, VIC 3125, Australia

Road map



Course Logistics

- Course Team
- Unit Resources
  - Textbooks
  - Software
- Unit Learning Outcomes

MDS Team

Unit Chair

Dr. Gang Li

- Email: gang.li@deakin.edu.au
- Phone: 925-17434
- Office: T202.3 (Burwood)
- Other issues by eMail appointment

For all
- On/Off campus issues
- Special Consideration (all students)
- Request for extension (all students)
- Complaints (all students)

Teaching Assistants

- And My Lab's PhD students
  - http://www.tulip.org.au

Instructor

- A/Prof. Gang Li
  - School of Information Technology, Deakin University VIC 3125, Australia
  - gang.li@deakin.edu.au
  - Phone: **+61(3)925-17434**
  - Office: **T207.WS14** (Burwood)

- Faculty HDR Coordinator
- Director of TULIP Lab
  - http://www.tulip.org.au

TULIP Lab, Deakin University (http://www.tulip.org.au)

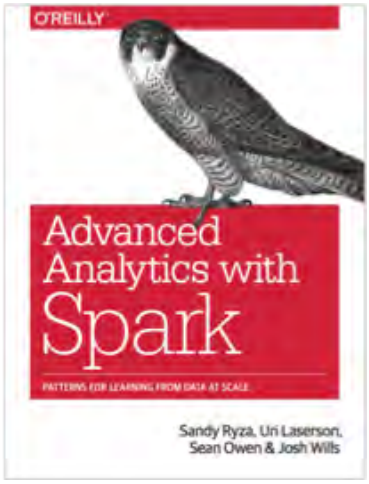
- A/Prof. Gang Li
  - **IEEE Technical Committee**
    - *IEEE Task Force on EDM (Vice chair)*
    - ***Data Mining & Big Data Analytics*** *(Vice Chair 17/18)*
    - *Enterprise Architecture and Engineering*
    - *Enterprise Information Systems*

http://cis.ieee.org/data-mining-tc.html

TULIP Lab
(Team for Universal Learning and Intelligent Processing)

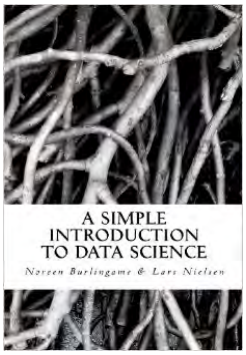
- Research Topics@ TULIP Lab
  - Behaviour Informatics
    - Group Behavior Analysis
    - Abnormal Analysis
  - Abuse Preventive Data Mining
    - Privacy Preservation in Data Mining
    - Information Abuse Prevention
  - Business Intelligence
    - Recommender System
    - Tourism/Hospitality Management

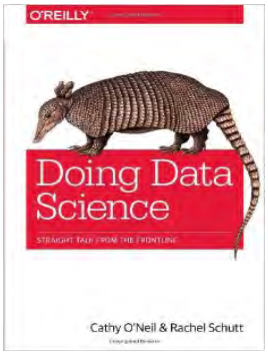
Unit Resources
Recommended Books

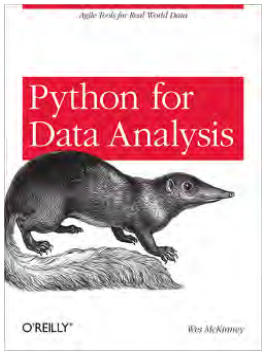


Unit Resources
Recommended Books

- Basic Data Science References

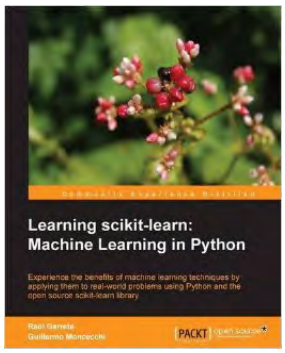


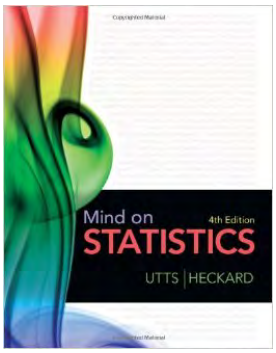


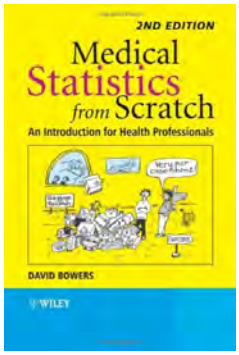


Unit Resources
Recommended Books

- Basic Basic Data Science References

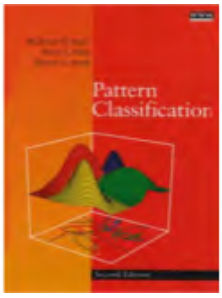


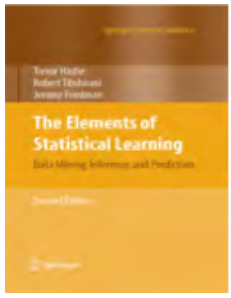




Unit Resources
Recommended Books

- Advanced Data Science References
  - Duda, Hart and Stork, ***Pattern Classification***, Wiley, 2001
  - Trevor Hastie, Robert Tibshirani, Jerome Friedman. *Statistical Learning: Data Mining, Inference, and Prediction*. 2009, Springer





Unit Resources
(Software)

Programming Language

- Python 3.X/2.X
  - Cross platform
  - …

Cloud Platforms

- Spark
  - Installed on clusters of computer, or a single PC with a virtual box image file

Unit Resources (Services)

- For cloud computing, you can use either
  - Amazon CWS or Free/Community accounts at
    - IBM Cloud
    - Databricks Community Version

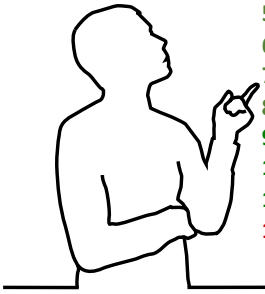
Why *Python*?

- **Relatively easy to learn and use**
  - Simple syntax
  - Interpretive, which makes debugging easier
  - Don't have to worry about managing memory
- **Modern**
  - Supports object-oriented programming,
- **Increasingly popular**
  - Used in majority D.S and C.S units at Deakin
  - Increasing use in industry
  - Large and ever growing set of libraries

Unit Overview

Session Plan

1. R/Python
2. Data Science Overview
3. Data Manipulation: Big Data
4. Data Manipulation: Exploratory Data Analysis
5. Data Analytics: Machine Learning
6. Data Analytics: Supervised Learning
7. Data Analytics: Model Evaluation
8. Data Analytics: Unsupervised Learning
9. Cloud Computing: Programming Spark
10. Cloud Computing: Spark SQL
11. Distributed Data Analytics
12. Deep Learning



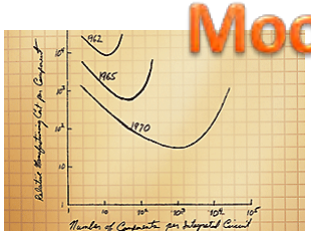
What?

- From Data to Knowledge
- Selected P.R. Applications

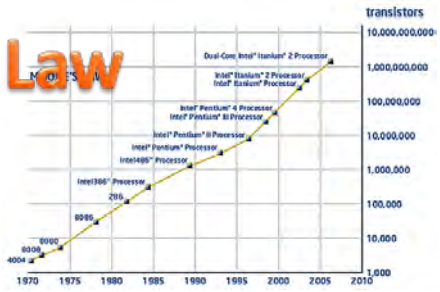
51 Years of Moore's Law

- Hig[illegible]er performance **Computers**
- Hu[illegible]e volume **Storage** media
- Str[illegible]ng **Data Collection** equipment
- Ea[illegible] **Data Generation** equipment

https://en.wikipedia.org/wiki/Moore%27s_law



Moore's Law



Massive Data Generation

- Vatican Square: 2005 vs 2013

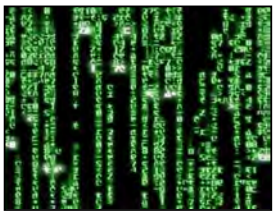
Then End of Science

The quest for knowledge used to begin with grand theories.
Now it begins with massive amounts of data.

Welcome to the Big Data Age.

Data, Information and Knowledge

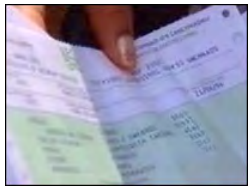
- **Data are** raw facts and figures that on their own have no meaning
  - these can be any alphanumeric characters
    - i.e. text, numbers, symbols
  - For example:
    - Yes, Yes, No, Yes, No, Yes, No, Yes
    - 42, 63, 96, 74, 56, 86
    - 111192, 111234



  - None of the above data sets have any meaning until they are given a CONTEXT and PROCESSED into a usable form

Data, Information and Knowledge

- **Information** is data that has been processed within a context to give it meaning, or data that has been processed into a form that gives it meaning
  - Accurate, relevant, and timely *information* is key to good decision making.



Data, Information and Knowledge

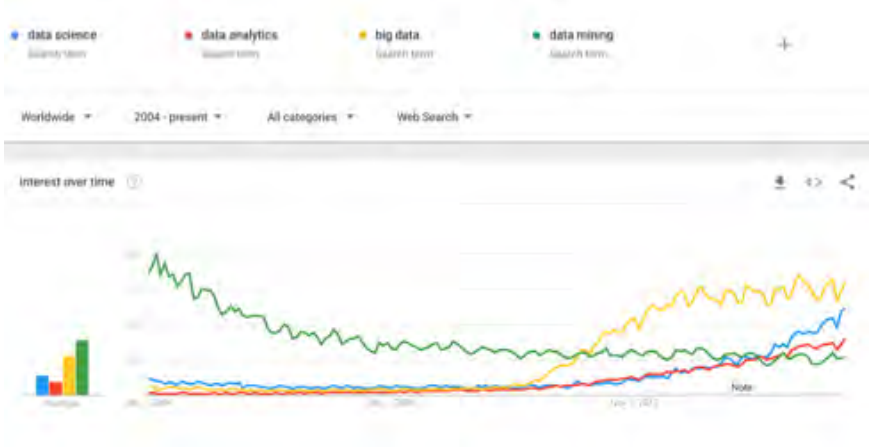
- **Knowledge** is the appropriate collection of information, such that it's intent is to be useful
  - Or put it simply, ***useful information***, for example
    - *Based on previous information, a Marketing Manager could use this information to decide whether or not to raise or lower price y*

http://www.systems-thinking.org/dikw/dikw.htm

Data: The New Currency

- Data is everywhere:
  - Facebook, blogs, emails, cars, sensors, ship, gnome sequence, your Fitbit, and even your body.
- Everyone talks about data these days, and people think that data is the new currency!

Google Trends

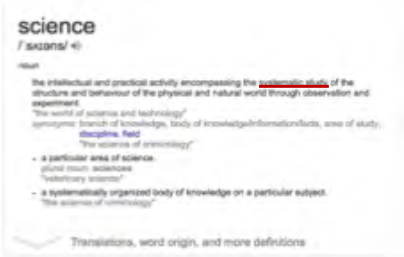


Data Science

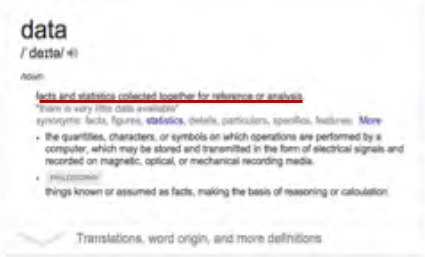
- What is Data Science?
- Data Science History
- Why Now?
- What D.S. work might look like?
- What Skills do you need?

What is Data Science?

- Data Science = Science of Data
  - The intellectual and practical activity encompassing the ***systematic study*** of ***facts and statistics collected together for reference or analysis***.



google define:science



google define:data

What is Data Science?

- However, there is not yet a definition agreed by all.
  - Other examples

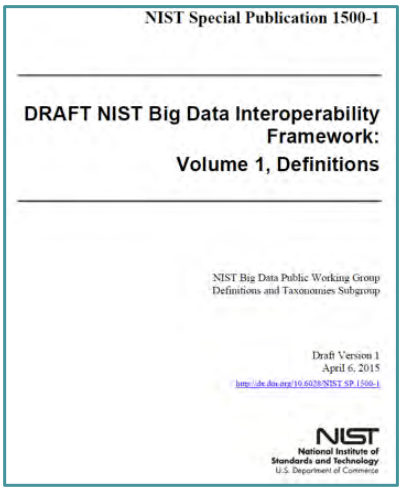
Wikipedia	"Data Science is the extraction of knowledge from large volumes of data that are structured or unstructured"
NIST, 2015	**"Data science is the empirical synthesis of actionable knowledge from raw data through the data lifecycle process"**
Dhar, 2013	"Data science is the study of generalizable knowledge from data"
Peter Naur, 1974	"[data science is] The science of dealing with data, once they have been established, while the relation of the data to what they represent is delegated to other fields and sciences."

What is Data Science?

- Summary
  - Data science is an emerging discipline. It remains a science where new knowledge and tools are still being invented.
    - There is not yet a clear definition agreed by all for the term 'data science'.
    - Different definitions exist from different perspectives (government, business, research, etc.)
  - We adapt NIST's definition:
    - "Data science is the empirical synthesis of actionable knowledge from raw data through the data lifecycle process"
  - **You**, as the future data scientist, will shape the field.

What is Data Science?

- NIST, Big Data Framework, 2015
  - Latest draft on data science and big data analytics.
  - We will use this proposed framework in this Unit.
  - URL: http://dx.doi.org/10.6028/NIST.SP.1500-1


NIST Special Publication 1500-1

DRAFT NIST Big Data Interoperability Framework:
Volume 1, Definitions

NIST Big Data Public Working Group
Definitions and Taxonomies Subgroup

Draft Version 1
April 6, 2015

NIST
National Institute of Standards and Technology
U.S. Department of Commerce

Brief Data Science History

- 1935:
  - R. A. Fisher, "The Design of Experiments"
  - "Correlation does not imply causation"
- 1939:
  - W. E. Demming, "Quality Control"
- 1958:
  - Peter Luhn, "A Business Intelligence System"

Brief Data Science History

- 1962:
  - John W. Tukey:
    - "The Future of Data Analysis"
    - "Exploratory Data Analysis"
- 1974:
  - Peter Naur:
    - "Concise Survey of Computer Methods"
    - "The science of dealing with data, once they have been established, while t f the data to what they represent is delegated to other fields and sciences."
- 1989:
  - Gregory Piatetsky-Shapiro
    - "KDD Workshop"
    - ACM SIGKDD

Brief Data Science History

- 1996:
  - *International Federation of Classification Societies (IFCS)*
    - For the first time, the term "data science" is included in the title of the conference
    - "Data science, classification, and related methods"
- 1997:
  - C. F. Jeff Wu:
    - Statistics → Data Science
- 2001:
  - William S. Cleveland:
    - "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics."

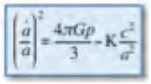
Brief Data Science History

- 2002-2003:
  - Data Science Journal
  - Journal of Data Science
- 2007:
  - Yangyong Zhu and Yun Xiong:
    - "Introduction to Dataology and Data Science," in which they state "Different from natural science and social science, Dataology and Data Science takes data in cyberspace as its research object. It is a new science."

Brief Data Science History

- 2007:
  - Toney Hey, Stewart Tansley, Kristin Tolle, Data-Driven Science, "T… Paradigm"
  - http://research.microsoft.com/en-us/collaboration/fourthparadi…
- 2009:
  - Peter Norvig, "The Unreasonable Effectiveness of Data"
- 2010:
  - Exponential Growth in Data Volume
  - "The Data Deluge"
    - http://www.economist.com/node/15579717

Data Science = 4th Paradigm of Science



(1942-2012)

- **Empirical + Experimental**
  - Thousand years ago
  - Describing natural phenomena
- **Theoretical**
  - Last few hundred years
  - Using models, generalizations
- **Computational**
  - Last few decades
  - Simulating complex Phenomena
- **Data Exploration** (eScience)
  - Unify theory, experiment and simulation
  - Data captured by instruments or generated by simulator
  - Processed by software
  - Information/Knowledge stored in computer
  - Scientist analyzes database/files using data management and statistics

Why Now? Meet the Data Deluge

- Age of Big Data
  - Watch this in real-time:
  - http://onesecond.designly.com/

Source: *Best practice guideline: Big Data*, ADMD
http://www.adma.com.au/assets/Uploads/Downloads/Big-Data-Best-Practice-Guidelines2.pdf

Why Now?

- Meet the Data Deluge
  - "***The average person today processes more data in a single day than a person in the 1500's did in an entire life time***"

Image source: Smolan and Erwitt, *The human face of big data*, 2013.

Why Now?
Meet the Data Deluge

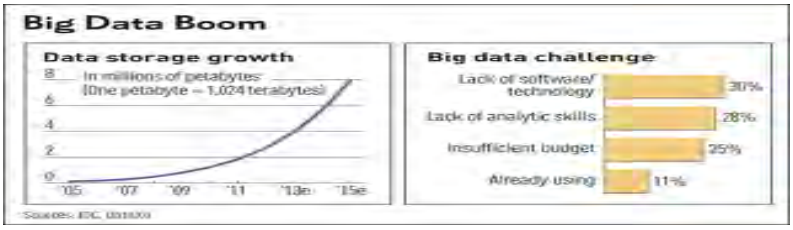
- *"The future belongs to the companies and people that turn data into products"*

  – Mike Loukides
- *"Companies in the top third of their industry in the use of data-driven making were, on average, 5% more productive and 6% more profitable than their competitors."*

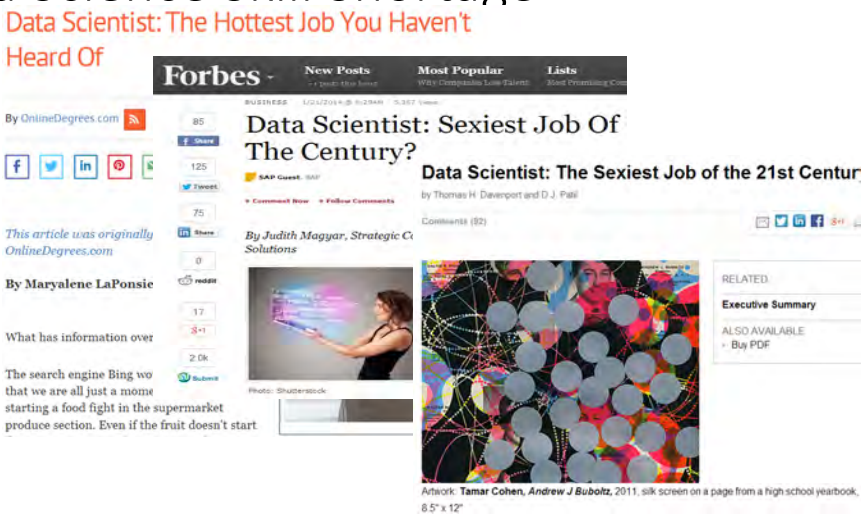
  – Harvard Business Review.

Why Now?
Data Science Skill Shortage

- The Bottleneck is in technology
  - New architecture, algorithms, techniques are needed
- Also in technical skills
  - Experts in using the new technology and dealing with big data

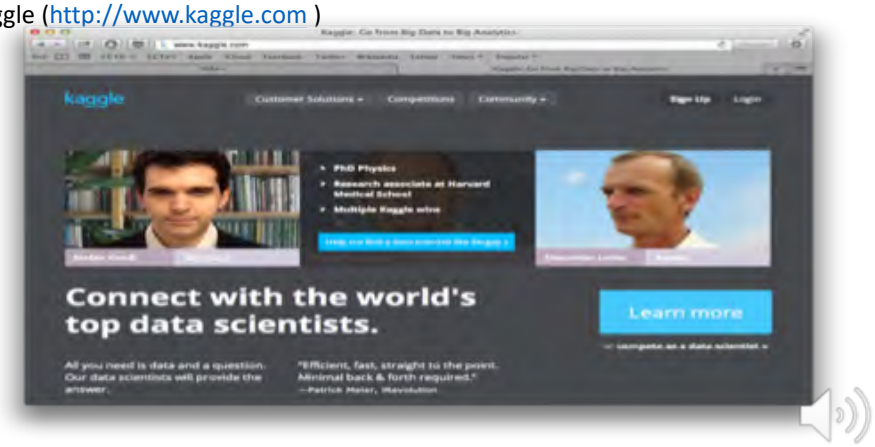


Why Now?
Data Science Skill Shortage



Why Now?

- Global Opportunities to Build Up your Profile
  - Kaggle (http://www.kaggle.com)



Why Now?

- Top Academic Conference Cup

Why Now?

- Top Industry Data Competition

Why Now?

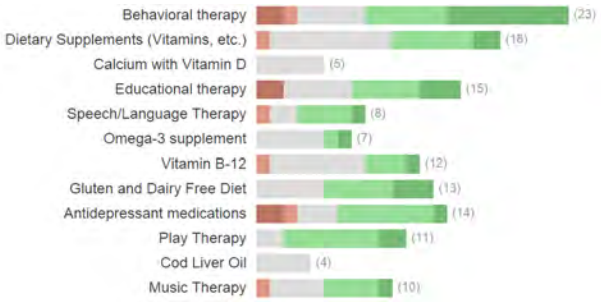
- Summary (Why data science)
  - Data is everywhere and available in huge volume.
  - Tapping and extracting values from data is crucial.
  - The future belong to organizations, government, business, individuals that know how to use data.
  - There is a tremendous shortage of skilled data scientists.
  - Relatively easy to build up your profile through competitions

What D.S. work might look like?

- Case (1)
  - **Collective Intelligence**
    - Use data from the crowd to derive new insights!

"autism"

CureTogether

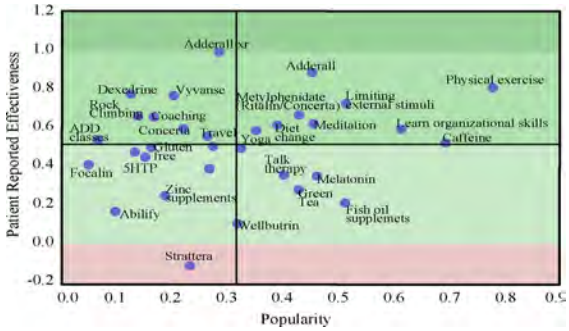


What D.S. work might look like?

- Case (1)
  - **Repurposing** the data
    - Reusing the data to derive new knowledge

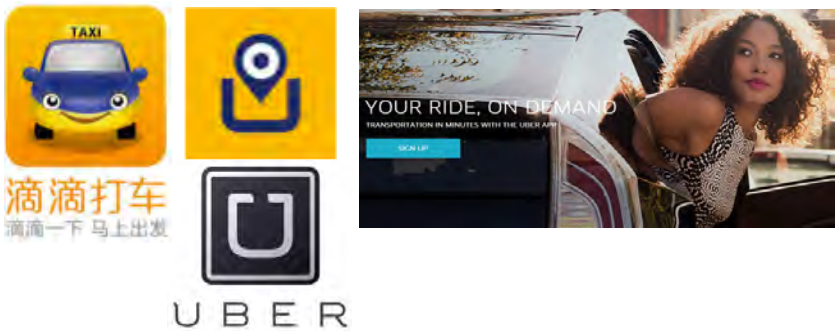
"autism"

CureTogether



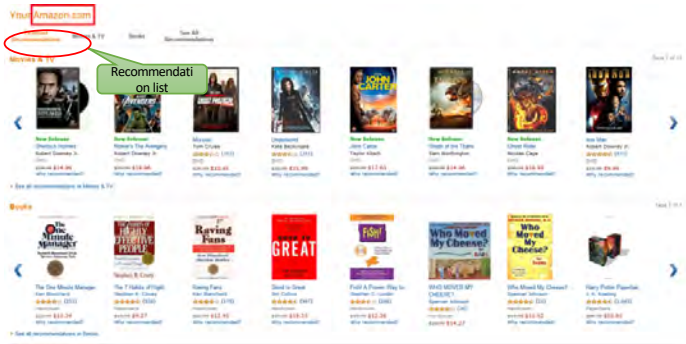
What D.S. work might look like?

- Case (2)
  - Use data to connect **demand** and **supply**



What D.S. work might look like?

- Case (3)
  - Recommender System
    - Use data to understand customer behaviors.



What D.S. work might look like?

- Case (4)
  - Many photo-capturing devices now have built-in global positioning systems (GPS) technology

  - **Geotagged photos**, with embedded **time** and **geographical information**, are shared on social website
    - Flickr (*www.flickr.com*)
    - Panoramio (www.panoramio.com)

What D.S. work might look like?

- Case (4)
  - Task: analyze photos to understand travelers' behavior

What D.S. work might look like?

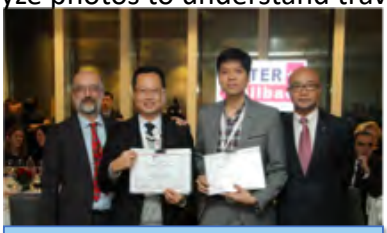
- Case (4)
  - Task: analyze photos to understand travelers' behavior

(A) Asian Tourist (B) Western Tourist

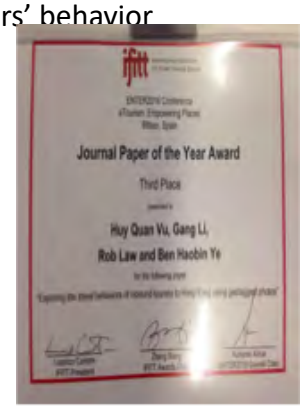
Tourist Traffic Flow in Hong Kong Metropolitan Area.

What D.S. work might look like?

- Case (4)
  - Task: analyze photos to understand travelers' behavior



Huy Quan Vu, Gang Li, Rob Law, Ben Haobin Yip. *Exploring the travel behaviors of inbound tourists to Hong Kong using Geotagged photos*. **Tourism Management**.



What D.S. work might look like?

- Summary:
  - Use data to exact new values, insights and hypothesis.
  - Reuse existing data to derive new knowledge.
  - Use data to understand customers' behaviour.
  - Use data to facilitate the demand market to suppliers.
  - Use data to build recommender system.
  - Use data to build predictive systems.
  - …..
  - DATA = NEW CURRENCY?

What Skills Do You Need?

NIST: Big Data Interoperability Framework, Vol.1 , 2015

Skills for a data scientist

- **Interdisciplinary**:
  - statistics, maths, computer science, programming, engineering, databases, etc…, which can be broadly categorised into three main skills:
  - **Data manipulation**:
    - crawling, cleaning, parsing, formatting, representing, munging, scrapping, ….
  - **Data analytics**:
    - ask the right questions, conjecture hypothesis
    - find dependency, correlation, perform statistical analysis, exploratory data analysis, machine learning, data mining, building predictive models, ….
  - **Communication of results**
    - visualization: charting, graphing, interactive graphics, tools,
    - Present your analysis, results, talk to business partners, etc.
    - build data product, start-up!

Data Science vs Databases

Elements	Data Science	Databases
Data Value	Cheap	Precious
Data Volume	Massive, Big Data	Modest
Priorities	Speed, Scalability, Query Richness, Availability	Consistency, Error Recovery, Auditability
Structured	Semi or Unstructured	Strongly structured
Properties	Eventual consistency CAP theorem (2/3) • Consistency, Availability, Partition Tolerance	Transactions, ACID: Atomicity, Consistency, Isolation and Durability
Realizations	NoSQL: Apache Cassandra, MongoDB, Apache Hbase, etc.	SQL
Now, Past or Future?	Nowcasting, Forecasting • Actionable insight	Querying the past

Data Science vs Machine Learning

Data Science	Machine Learning
Explore many models, build and tune hybrids	Develop new individual models
Understand empirical properties of models	Prove mathematical properties of models
Develop or use tools that can handle massive datasets	Improve/validate on a few, relatively clean, small datasets
Actionable Insights	Theoretical Development

Data Science vs Scientific Computing

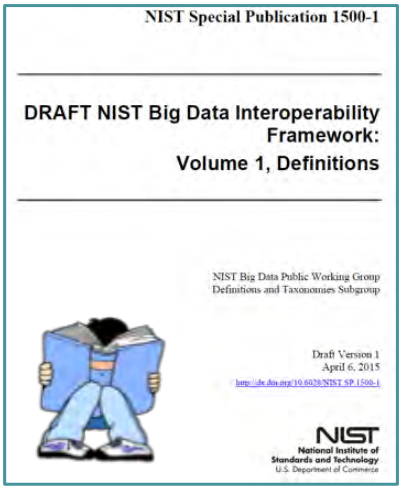
Data Science	Scientific Computing
General Inference engine replaces model	Physics-based models
Structure not related to problem	Problem-structured
Statistical models handle true randomness, and unmodeled complexity	Mostly deterministic, precise
Run on cheaper Computing Clusters (EC2)	Run on Supercomputer or High-end Computing Cluster

This Session's Readings

- A very short history of Data Science
  - https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/#2fade0dc55cf

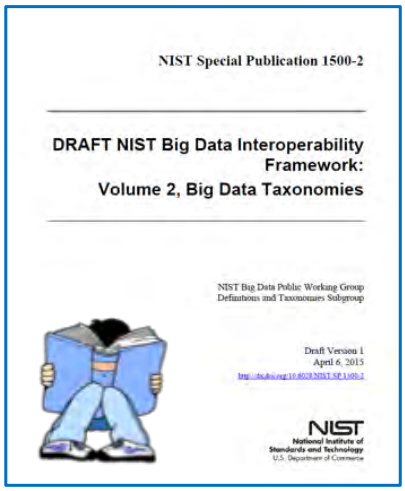
This Session's Readings

- NIST, Big Data Framework, 2015
  - Latest draft on data science and big data analytics.
  - We will use this proposed framework in this Unit.
  - http://dx.doi.org/10.6028/NIST.SP.1500-1



This Session's Readings

- NIST, Big Data Framework, 2015
  - Latest draft on data science and big data analytics.
  - We will use this proposed framework in this Unit.
  - http://dx.doi.org/10.6028/NIST.SP.1500-2



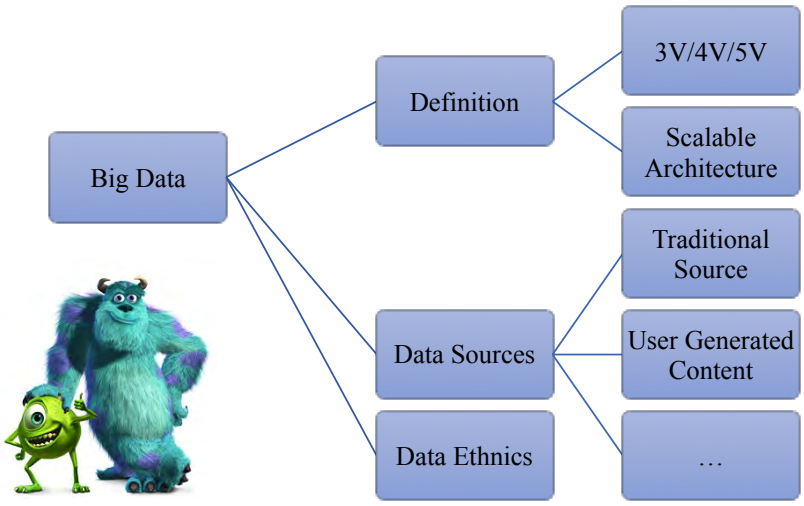
Lecture Notes on

Advanced Data Analytics

Module 01: Big Data

Gang Li
School of Information Technology
Deakin University, VIC 3125, Australia

Road map



Big Data

- Big Data
  - 3V
  - 4V
  - 5V
- Scalable Architecture
  - Scale Out

Meet the Data Deluge

- Age of Big Data
  - "***The average person today processes more data in a single day than a person in the 1500's did in an entire life time***"

Problem with Data Deluge

- Data growing faster than computation speeds
  - Growing data sources
    - Web, mobile, sensor, scientific, etc.
      - Facebook's daily logs: 60TB
      - 1,000 Genomes Projects: 200%B
      - Google Web index: 10+ PB
      - Cost of 1TB of disk: ~ $50
  - Storage getting cheaper
    - Size doubling every 18 months
  - Stalling CPU speeds and storage bottlenecks
    - Time to read 1TB from disk: 3 hours (100MB/S)

Analytics – Computation Ways

- Centralized Model
- Decentralized Model
  - Multiparty computation
  - Centralized Agent based
- Distributed and Parallel
  - Don't Move the data, Move the Computation

Big Data?

- Each person's view is limited to his local region
  - The elephant "feels" like a rope, a hose, a wall, tree etc.
  - Let's assume that
    - the elephant is growing rapidly and its pose changes constantly
    - the blind men also learn from each other while exchanging information on their respective feelings on the elephant.
    - Equivalent to aggregating heterogeneous information
  - Not simple
    - Privacy concerns
    - Network bandwidth
    - Individuals may speak different languages
      - heterogeneous and diverse information sources

Problem with Data Deluge

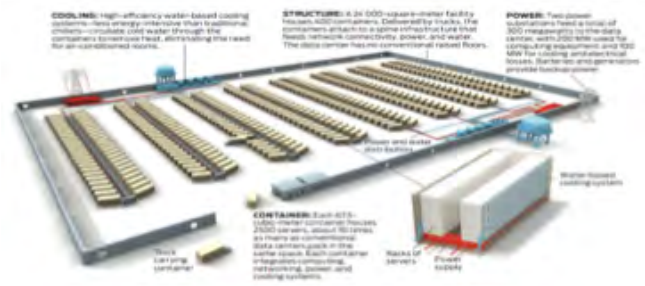
- Tra itional Analysis Tools run on a single machine!
  - single machine can no longer process or even store all the data!
- So tion: distribute data over large clusters

http://http://google-asiapacific.blogspot.com.au/2015/06/inside-our-data-center-in-singapore.html

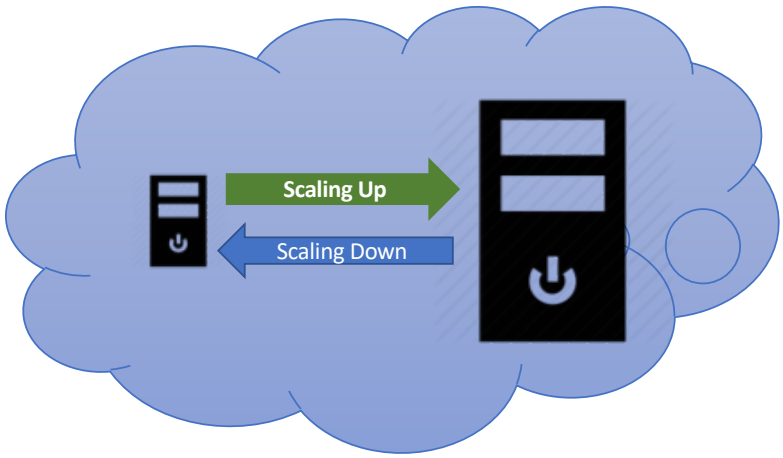
Data Science Enable: Cloud Computing

- Clo d Computing
  - educes computing operating costs
  - nables data science on massive numbers of inexpensive computers (scaling ut)

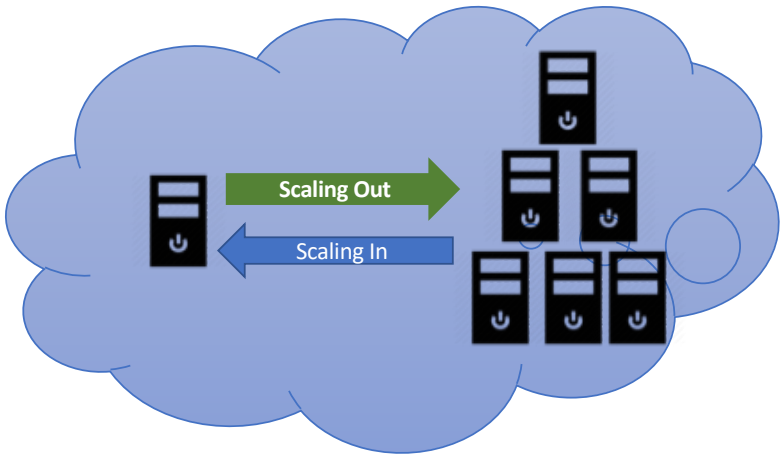
http://spectrum.ieee.org/image/MTk5MTU4



Architectures: Vertical Scaling

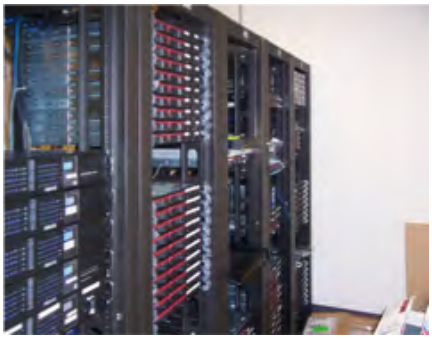


Architectures: Horizontal Scaling



Hardware for Big Data

- Consumer-grade hardware
  - Not "Gold Plated"
- Many desktop-like servers
  - Easy to add capacity
  - Cheaper per CPU/disk
- Complexity in Software



Problems with Cheap Hardware

- Failures, Google's numbers
  - 1-5% hard drivers/year
  - 0.2% DIMMs/year
  - If 1 server fails every 3 years, then with 10K nodes we will see 10 faults/day
  - Even worse: stragglers
- Network speeds versus shared memory
  - Much more latency
  - Network slower than storage

Problems with Cheap Hardware

- Uneven performance
- How do we split work across machines?

Google's Solution

- **Google File System** (GFS)
  - A supporting file system that allows data to be local to computation, and fault tolerant through replication
- **MapReduce**
  - A programming model for processing Big Data
- Publication:
  - Dean, J. and Ghemawat, S. 2004. MapReduce: simplified data processing on large clusters. OSDI'04: Sixth Symposium on Operating System Design and Implementation, San Francisco, CA, December, 2004.
  - Dean, J. and Ghemawat, S. 2008. MapReduce: simplified data processing on large clusters. Communication of ACM 51, 1 (Jan. 2008), 107-113.
  - Ghemawat, S, Gobioff, H. and Leung, S-T. The Google File System. Proceedings of ACM Symposium on Operating Systems Principles. 29-43, 2003

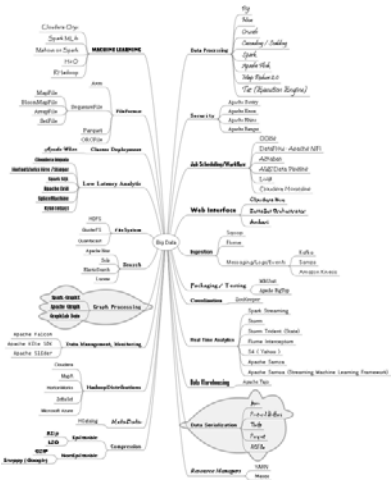
Apache Hadoop

- **Hadoop Distributed File System** (HDFS)
  - GFS is not open source
  - Doug Cutting and Yahoo! Reverse engineered the GFS, and call it HDFS
- **Hadoop MapReduce**

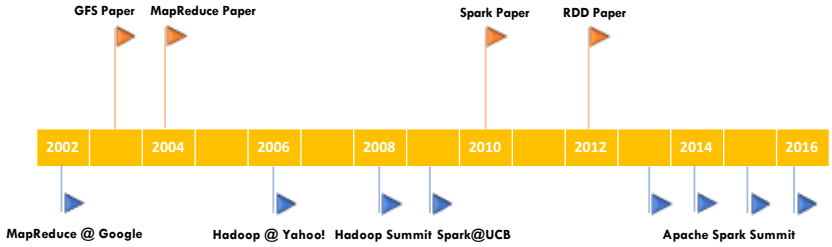
- The software framework that supports HDFS, MapReduce and other related entities is called the ***Project Hadoop***
  - open sourced and distributed by Apache

Hadoop Ecosystem

https://sumitpal.wordpress.com/



History Review

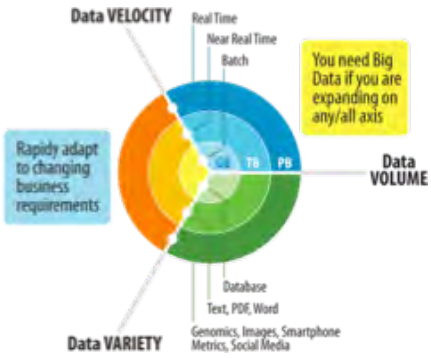


Big Data (3 Vs)

- No consensus but we use the NIST one:

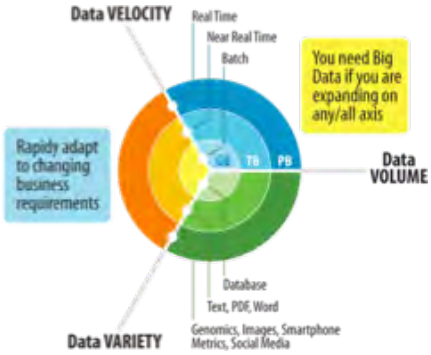
"*Big data consists of extensive **datasets** – primarily in the characteristics of volume, variety, velocity, and/or variability – that require **a scalable architecture** for efficient storage, manipulation, and analysis*"

– [NBDRA,2015]



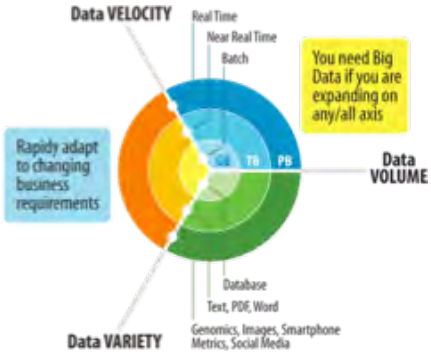
Big Data (3 Vs)

- **Volume**
  - Refer to the size of the dataset: **big**
  - Require new thinking and architecture in storage, query and access.
  - Require new tools and computation methods.
    - Hadoop, MapReduce, etc.
    - NoSQL, Pig, etc…



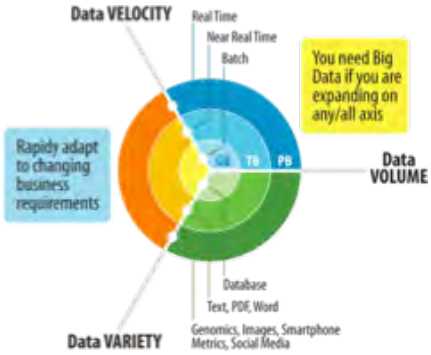
Big Data (3 Vs)

- **Variety**
  - Refer to the diversity of data.
  - Data comes from multiple domains, sources, repositories, or types.

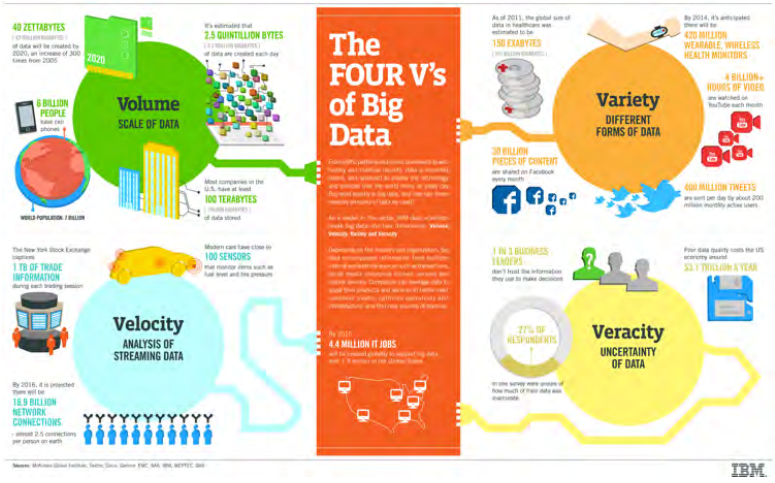


Big Data (3 Vs)

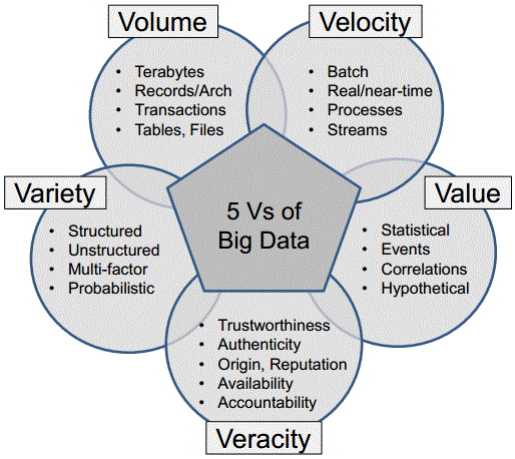
- **Velocity**
  - Refer to the rate of data growth.
  - Most of data nowadays is streaming.
    - e.g.: Facebook messages delivered to you in real-time



Big Data (4 Vs)



Big Data (5 Vs)

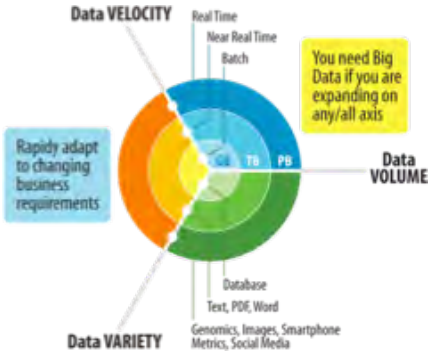


Big Data (3 Vs)

- No consensus but we use the NIST one:

"*Big data consists of extensive* ***datasets*** *– primarily in the characteristics of volume, variety, velocity, and/or variability – that require* ***a scalable architecture*** *for efficient storage, manipulation, and analysis*"

– [NBDRA,2015]

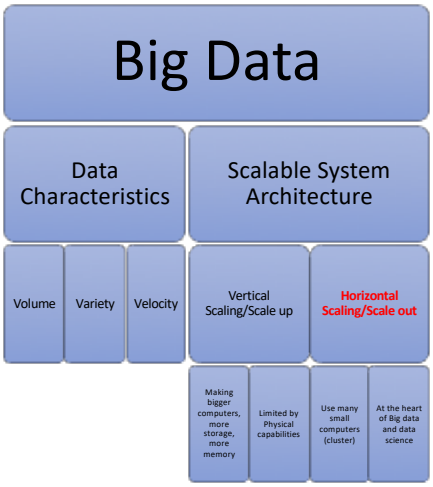


Big Data (3 Vs)

- No consensus but we use the NIST one:

"*Big data consists of extensive **datasets** – primarily in the characteristics of volume, variety, velocity, and/or variability – that require **a scalable architecture** for efficient storage, manipulation, and analysis*"

– [NBDRA,2015]

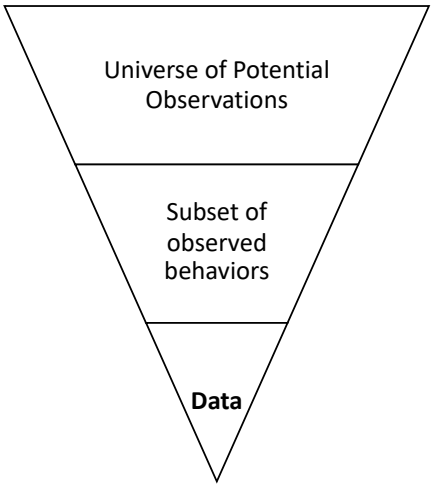


Where Does Data Come From?

- Traditional Sources
- User Generated Content
- Environmental Data
- Sensor Data
- Machine Log Data
- Government Data
- Enterprise Data

Where Does Data Come From?

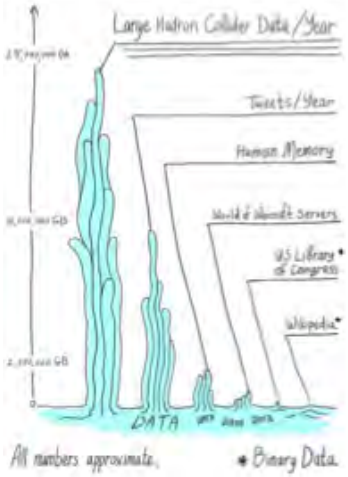
- **Traditional Data Sources**
  - The interaction of the researcher with the environment leads to data
  - Conditioned by the experimental design or measurement instrument, we have ***the universe of potential observations***
  - Only ***a sub-set of the potential observations*** will be observed
  - The observed behaviors have to be ***encoded*** in terms of some set of rules



Where Does Data Come From?

- **Traditional Data Sources (1)**
  - Health and Scientific Computing

http://www.symmetrymagazine.org/article/august-2012/particle-physics-tames-big-data



Where Does Data Come From?

- **Tra**[**ditional Data Sources (2)**](http://www.symmetrymagazine.org/article/august-2012/particle-physics-tames-big-data)
  - Media, News
  - Archive of scanned historical documents, scanned documents, books, medical records, etc.

Where Does Data Come From?

- **User Generated Content**
  - Social networking:
    - Facebook, Twitter, Reddit, …
  - Blogosphere:
    - WordPress, DataSift, Blogspot, …
  - Wikis:
    - Wikipedia, PmWiki, …
  - Web behaviours:
    - send trend, google search, network traffic

Where Does Data Come From?

- **Environmental/astronomical/economic:**
  - Weather data, Square Kilometre Array (SKA), Satellite imagery (MODIS, Landsat, nearmap, worldbank, bloomberg)

Where Does Data Come From?

- **Sensor data**
  - Medical devices, fitness sensors, car sensors, road cameras, satellites, cell towers, buildings, etc. …

Where Does Data Come From?

- **Machine Log Data**
  - Event logs, network traffic logs, business process logs, mobile location, …

Where Does Data Come From?

- **Government Data Sources**
  - http://data.gov.au/
  - http://data.gov/
  - Innovate with Open Government Data.
    - https://www.govhack.org/

Where Does Data Come From?

- **Enterprises Data Sources**
  - http://kaggle.com/
  - http://tianchi.aliyun.com/
  - Using Data Science to find top Data Scientists

Where Does Data Come From?

Crowdsourcing +

Terrain Model +

Sensor Data

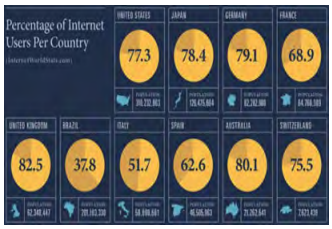
➔

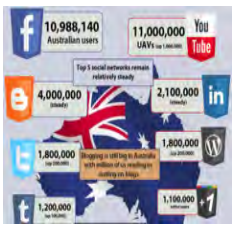
Data Ethics and Privacy

- Data Ethical Issues
- Data Privacy Issues

The Popularity of Online Social Networks

- The popularity of OSNs has greatly increased in recent years
  - A large amount of Internet users also use OSN services,
  - Australia has up to 11 million Facebook users.





The Popularity of Online Social Networks

- People share news, interests and ideas in OSNs,
  - **though** these platforms also spread email malware, rumours, gossips and malicious links, and also leak our privacy.

Data Ethical Issues

- Ethical issues such as:
  - When is it appropriate to use electronic medical records?
  - Is it okay to use demographics (sex, gender, religion) for loan application?
  - Data may reveal locations, when is it okay to use?
  - Is your data product or experiment cause distress, effect mental health?
- Questions to ask:
  - Who has the right to access the data?
  - What can the data be used for?
  - How can we preserve the privacy of individuals?

Data Ethical Issues

- Case (1): Mood Manipulation


Everything We Know About Facebook's Secret Mood Manipulation Experiment

It was probably legal. But was it ethical?

What did the paper find?

- The study found that by manipulating the news feeds displayed to 689,003 Facebook users, it could affect their content posted to Facebook.
- More negative news feeds lead to more negative status messages, as more positive news feeds led to positive statuses.

Data Ethical Issues

- Case (1): Mood Manipulation
  - It is probably legal. But was it ethical?


Facebook's Mood Manipulation Experiment Might Have Been Illegal

Two University of Maryland law professors allege that the social network's experiments—and OkCupid's—count as "research," and thus violate state statute.

Data Ethical Issues

http://www.informationweek.com/big-data/big-data-analytics/data-scientists-want-big-data-ethics-standards/d/d-id/1316258

InformationWeek CONNECTING THE BUSINESS TECHNOLOGY COMMUNITY

Home News & Commentary Authors Video Reports White Papers Events University

STRATEGIC CIO | IoT | DEVOPS | SOFTWARE | SECURITY | CLOUD | MOBILE

BIG DATA // BIG DATA ANALYTICS

NEWS
9/17/2014
09:05 AM

Data Scientists Want Big Data Ethics Standards

Nearly half of data scientists surveyed last month say Facebook's controversial "mood manipulation study" was unethical, and many support ethics guidelines for big data research.

Is O.S.N. a Secure Place to Show off?

Did Facebook photo trigger murder?

Aubrey Whelan, *INQUIRER STAFF WRITER*

Last week, Tony Harris, a 50-year-old electronics repairman, uploaded a photo to his Facebook page - something different from his standard fare of videos and memes and images of his three children.

This one showed his wife, Amber Crane, grinning up at the camera, clutching thick stacks of cash in both hands. More money sat piled in her lap.

"I misplaced $60,000.00. I hope my wife didn't go shopping with it," the caption joked.

"Stop playing," a friend wrote back.

Precisely a week later about 11:30 at night, police said, three young men walked into his home on the 1100 block of South Ruby Street in Kingsessing, where he has lived for decades. Where he always kept his

Application of Geo Photos

- There are many geotagged photos available. However :
  - *Data is noisy or misleading.*
  - *Photos are taken in transit rather than at the attractions.*
  - *Photo is static media, whereas, travel behavior is dynamic.*

How to mine these data for travel behaviour analysis?

Tourist Movement Analysis

(A) Asian Tourist

(B) Western Tourist

Tourist Traffic Flow in Hong Kong Metropolitan Area.

Tourist Movement Analysis

Huy Quan Vu, Gang Li, Rob Law, Ben Haobin Yip. *Exploring the travel behaviors of inbound tourists to Hong Kong using Geotagged photos*. **Tourism Management**.

Concerns on Data Privacy

- Privacy Breach by Released Photos
- Privacy Breach by Released Datasets

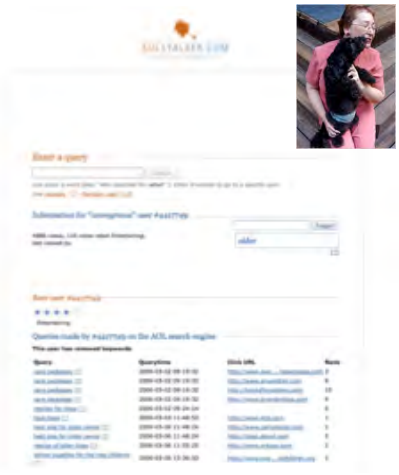
I Know Where Your Cat Lives

- This project randomly selected one million images that include the word "**cat**" acr s public photo sites, which plots the location coordinates from each photo aga st a map to show where each cat lives. ---- July 22, 2014 ***Time***

http://time.com/3019671/i-know-where-your-cat-lives-website-data-visualization/

AOL Dataset Debacle

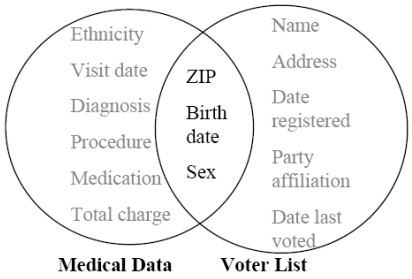
- AOL search data leak (2006):
  - 36 million search terms of 650,000 users
    - http://search-id.com
    - http://www.aolstalker.com
- I Know who you are
  - **Click history** can uniquely identify a person
  - Find all log entries for AOL user 4417749
  - Multiple queries for businesses and services in *Lilburn, GA* (population 11K)
    - queries for *Jarrett Arnold*
    - Lilburn has 14 people with the last name Arnold
    - NYT contacts them, finds out User 4417749 is *Thelma Arnold*



Breach of medical record

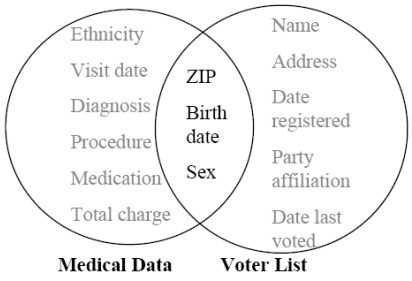
- 87% of Americans can be uniquely identified by
  - {***zip code, gender, date of birth***}
  - *Latanya Sweeney* re-identify the medical record of an ex-governor of Massachusetts.

[***International Journal on Uncertainty, Fuzziness and Knowledge-based Systems***, 2002]



Breach of medical record

- **Linkage Attacks**
  - Using "innocuous" data in one dataset to identify a record in a different dataset containing both innocuous and sensitive data
  - At the heart of the voluminous research on hiding small cell counts in tabular data



One of the 1st Concerns in Statistics

- Dalenius, T., "***Towards a methodology for statistical disclosure control,***" Statistisktidskrift, 5, 429–444, 1977.


Towards a methodology for statistical disclosure control

by Tore Dalenius*

Special Issues on Privacy

- **IEEE Spectrum**, Aug. 2014
  - On the Internet, nobody knows you are a dog (1993).
  - Interested parties not only know you are a dog, but also know the colour of your fur (2014)

Special Issues on Privacy

- **Communication of ACM** Sept. 2014
  - Federal law governing student privacy and the release of student records suggests that anonymizing student data can hardly protects student privacy.


practice

Quality social science research and the privacy of human subjects require trust.

BY JON P. DAVIES, JUSTIN REICH, JIM WALDO, ELISE M. YOUNG, JONATHAN WHITTINGHILL, ANDREW DEAN HO, DANIEL THOMAS SEATON, AND ISAAC CHUANG

Privacy, Anonymity, and Big Data in the Social Sciences

Special Issues on Privacy

- **Science**, Jan. 2015
  - Data pour out of us and our devices every second of every day, and people no longer control their personal privacy.

Challenge: Tradeoff P&U

- ***Privacy***
  - bounded by the privacy budget
- ***Utility***
  - diverse measurements according to different application requirements
    - Recommendation system: similarity covariance
    - Classification: misclassified rate
- ***Privacy vs. Utility***
  - Both mechanisms sacrifice utility to gain privacy
  - ***Tradeoff***: To get the maximal utility in a fixed

Challenge: Tradeoff P&U

- **Big Data is a double edged sword**
  - provides reliable resources for data mining
  - Brings potential threats to enterprise
- **Privacy Preserving in Data Releasing**
  - T. Zhu, G. Li, W. Zhou, P.S. Yu. *Differential Privacy and Its Applications: Survey*, **IEEE Transactions on Data and Knowledge Engineering**, 2017
  - T. Zhu, G. Li, W. Zhou, P.S. Yu. *Differential Privacy and Its Applications*, Springer 2017
  - **Coupled Differential Privacy**
    - T. Zhu, P. Xiong, G. Li, W. Zhou. *Correlated Differential Privacy: Hiding Information in Non-IID Dataset*. **IEEE Transactions on Information Forensics & Security**, 2015, 10(2), 229-242

Pri acy and Law

- EU ourt Orders Google to Respect User's "**Right to be Forgotte** (26 05/2014)
- Go gle has faced a range of lawsuits from American citizens and pri acy groups over a range of their products. (23/07/2014)

http://www.smh.com.au/digital-life/digital-life-news/google-to-face-privacy-lawsuit-in-the-us-20140723-zvwop.html

http://www.ipbrief.net/2014http://sorbenchmark.org//05/26/eu-court-orders-google-to-respect-users-right-to-be-forgotten/

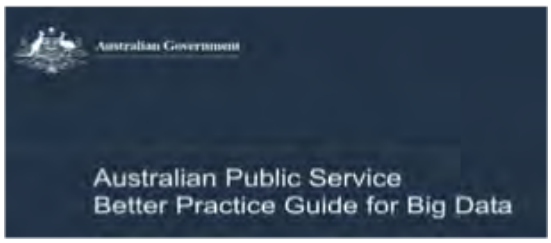
Data Ethics and Privacy Standards

- **Australia's Privacy Act** 1988
- **Australia Privacy Principles** (APP) become effective from March, 2014.

Australian Government
Office of the Australian Information Commissioner

Data Ethics and Privacy Standards

- Australian Government
  - Australian Public Service Better Practice Guide for Big Data
    - http://www.finance.gov.au/sites/default/files/APS-Better-Practice-Guide-for-Big-Data.pdf



Data Ethics and Privacy Standards

- ADMA Guideline
  - Transparency is the key to trust. Use your privacy policy and collection notices to develop consumer trust in your data collection practice"
  - Businesses should comply with legal requirements in the collection and use of data, as per the APPs."
  - Businesses should assess, beyond the legal requirements, whether their use of data will be within customers' expectations."
  - Businesses should use data and analytics in a responsible manner, and review their practices so that they ensure they are delivering benefits to consumers, not just the business."
  - What is a responsible use of data and analytics will be determined by the circumstances, and the specific risks that any particular data use creates. "
  - Data security should be assessed on the basis of the kinds of information collected and used, and the relative risks associated with that."
  - Businesses should consider the vulnerabilities of particular market segments, such as children, in their use of data and analytics."
  - Be aware that if you breach privacy principles or expectations, you will invite further gov

http://www.adma.com.au/assets/Uploads/Downloads/Big-Data-Best-Practice-Guideline-v2.pdf

ADMA Association for data-driven marketing & advertising

Data Ethics and Privacy

- Issues arise in practice.
  - Be aware of privacy and ethical issues in using data.
- Questions to ask:
  - Who has the right to access the data?
  - What can the data be used for?
  - What measure and actions needed, e.g., do you need an ethics approval, etc.
  - How can we preserve the privacy of individuals?

This Session's Readings

- A Very Short History of Big Data:
  - https://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/#3d64ff6865a1
- Have a look at Four Milestone Papers (Difficult to read)
  - *The Google File System*. Ghemawat, S, Gobioff, H. and Leung, S-T. Proceedings of ACM Symposium on Operating Systems Principles. 29-43, 2003
  - *MapReduce: Simplified Data Processing on Large Clusters* Jeffrey Dean and Sanjay Ghemawat, 2004
  - *Spark: Cluster Computing with Working Sets* Matei Zaharia, Mosharaf Chowdhury, Michael J. Franklin, Scott Shenker, Ion Stoica USENIX HotCloud (2010)
  - *Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing,* Matei Zaharia, Mosharaf Chowdhury, Tathagata Das, Ankur Dave, Justin Ma, Murphy McCauley, Michael J. Franklin, Scott Shenker, Ion Stoica, NSDI (2012)

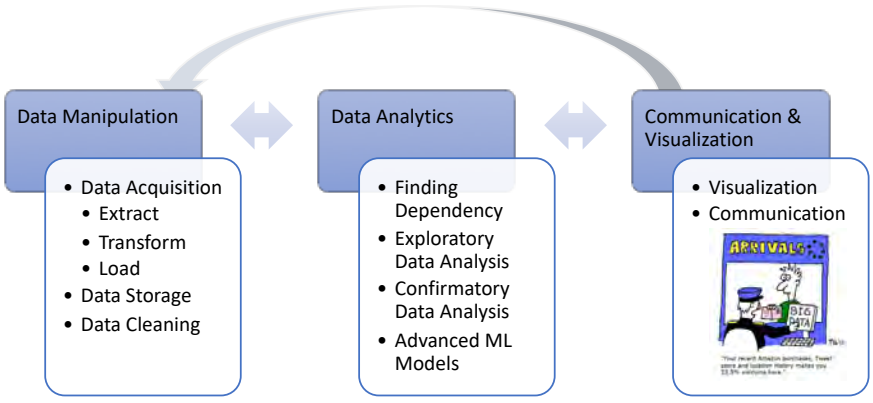
Lecture Notes on

Advanced Data Analytics

Module 01: Data Mining Case Study

Gang Li
School of Information Technology
Deakin University, VIC 3125, Australia

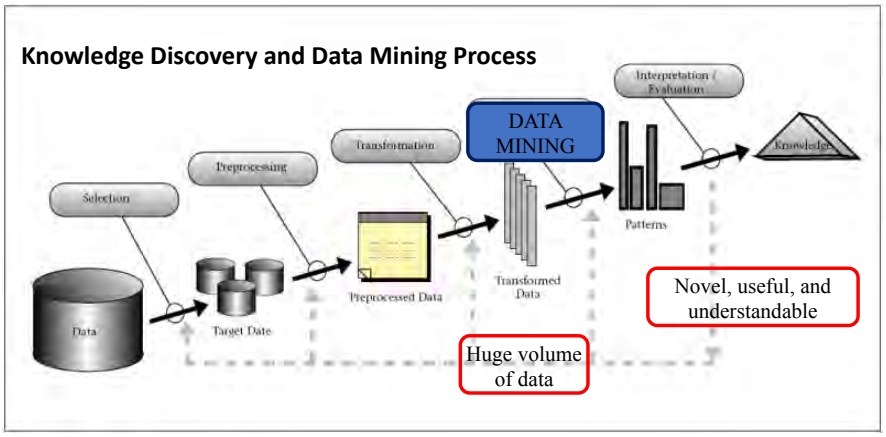
Data Science Process



Data Mining Process

varied from U. Fayad, et al. 's definition of KDD at KDD96

- Data mining is the ***non-trivial*** process of identifying ***valid***, ***novel***, ***potentially useful***, and ***ultimately understandable*** patterns from ***huge volume of data***



Data Mining Process

varied from U. Fayad, et al. 's definition of KDD at KDD96

- Data mining is the ***non-trivial*** process of identifying ***valid***, ***novel***, ***potentially useful***, and ***ultimately understandable*** patterns from ***huge volume of data***

- Why ***huge volume of data***?
  - Analyzing small volume of data does not require data mining
- Why ***non-trivial***?
  - Data mining is not so easy as SQL queries
- Why ***valid***?
  - Incorrect patterns are valueless
- Why ***novel***?
  - Investment on known-knowledge is wasteful
- Why ***potentially useful***?
  - Patterns will be used in decision making for future affairs
- Why ***ultimately understandable***?
  - Patterns will be presented to decision makers

Self-Organizing Maps

- Market Segmentation
- Clustering
- SOFM
- Biological Inspiration of SOM
- Learning Algorithm

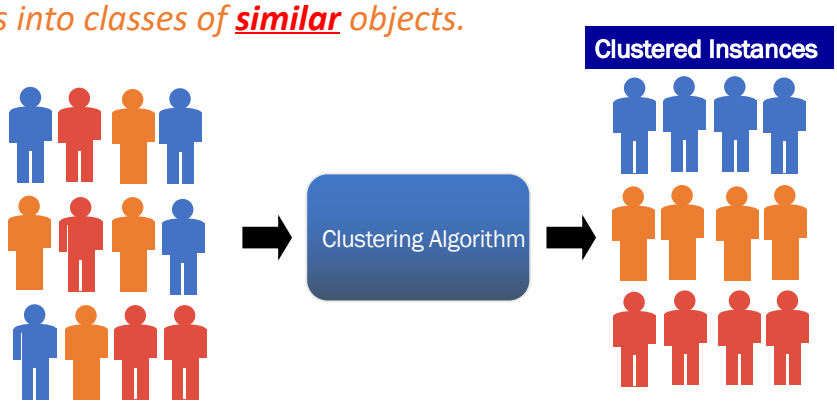
Market Segmentation

- **Market Segmentation** refers to the process of forming groups of people, whereby the groups are homogeneous in terms of demand elasticity and are accessible via marketing Strategies
  - identifying the appropriate segments for ***target marketing***,
  - gaining competitive advantage through ***product differentiation***, or at least
  - enabling business to ***target customers*** more effectively

Kotler, P. (1980). Marketing management: analysis, planning and control, P.H.

Clustering

- Clustering is *the process of grouping a set of physical or abstract objects into classes of **similar** objects.*



Major Clustering Approaches

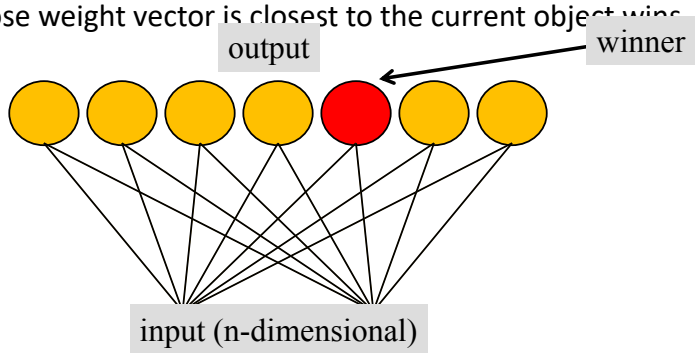
- **Partitioning** algorithms:
  - Construct various partitions and then evaluate them by some criterion, e.g. K-Means, etc.
- **Hierarchy** algorithms:
  - Create a hierarchical decomposition of the set of data (or objects) using some criterion, e.g. DIANA, etc.
- **Density-based**:
  - based on connectivity and density functions, e.g. EM, etc.
- **Grid-based**:
  - based on a multiple-level granularity structure, e.g., STING, etc.
- **Model-based**:
  - A model is hypothesized for each of the clusters and the idea is to find the best fit of that model to each other, eg, SOM

Self-Organizing Maps (SOMs)

- SOM (SOFM)
  - a.k.a as Self-Organizing Feature Map (SOFM), Kohonen networks
  - It captures essential features of dimensionality reduction in Brain
    - The brain maps the external multidimensional representation of the world (including its spatial relations) into a similar 1 or 2 - dimensional internal representation.
    - That is, the brain processes the external signals in a topology-preserving way
  - So, if we are to have a hope of mimicking the way the brain learns, our system should be able to do the same thing.

Self-Organizing Maps (SOMs)

- SOM (SOFM)
  - Clustering is performed by having several units competing for the current object
  - The unit whose weight vector is closest to the current object wins



Self-Organizing Maps (SOMs)

- Competitive learning
  - Determine the winner (the neuron of which the weight vector has the smallest distance to the input vector)
  - Move the weight vector **w** of the winning neuron towards the input **i**

Before learning *After learning*

Self-Organizing Maps (SOMs)

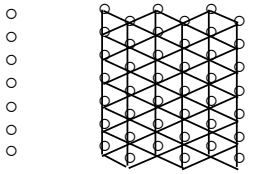
- Kohonen's Idea
  - Impose a topological order onto the competitive neurons (e.g., rectangular map)
  - Let neighbours of the winner share the "prize" (The "postcode lottery" principle.)
  - After learning, neurons with similar weights tend to cluster on the map

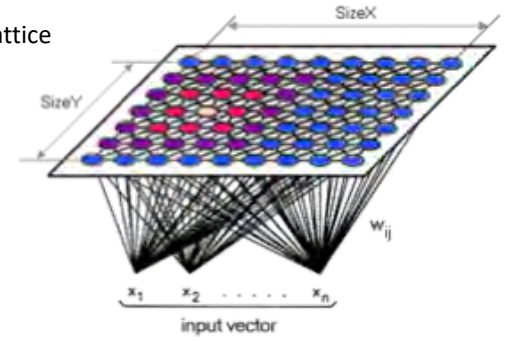
Biological Inspiration of SOM

- Brain is organized such a way that different sensory data is represented by topologically ordered computational maps
  - tactile, visual, acoustic sensory input are mapped onto areas of cerebral cortex in topologically ordered manner
  - building block of information processing infrastructure of nervous system

Topological order

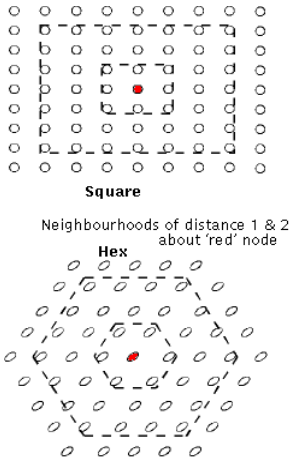
- SOM contains two layers:
  - Input Layer
  - Kohonen Layer
    - neurons placed at the nodes of lattice
    - one or two dimension





Topological order

- neighborhoods
- Square
  - winner (red)
  - Nearest neighbors
- Hexagonal
  - Winner (red)
  - Nearest neighbors



Learning a SOM Model

- Transform incoming input pattern into the Kohonen layer (discrete map of 1-D or 2-D)
  - adaptively topologically ordered fashion
    - Topology-preserving transformation
  - input pattern is represented as a localized region or spot of activities in the network
- General Procedure
  - After initialization, iterate over three essential processes
    - competition
    - cooperation
    - synaptic adaptation

Competitive Process

- Find best match of input vector with synaptic weight

  $x = [x_1, x_2, \ldots, x_3]^T$

  $w_j = [w_{j1}, w_{j2}, \ldots, w_{jm}]^T, j = 1, 2, 3, l$

- Best matching, winning neuron

  $i(x) = \arg\min ||x - w_j||, j = 1,2,3,..,l$

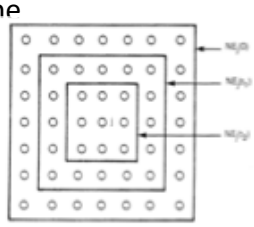
- Determine the location where the topological neighborhood of excited neurons is to be centered
- continuous input space is mapped onto discrete output space of neuron by competitive process

Cooperative Process

- For a winning neuron, the neurons in its immediate neighborhood excite more than those farther away
- topological neighborhood decay smoothly with lateral distance
  - Symmetric about maximum point defined by $d_{ij} = 0$
  - Monotonically decreasing to zero for $d_{ij} \rightarrow \infty$
  - Neighborhood function: Gaussian case
- Size of neighborhood shrinks with time

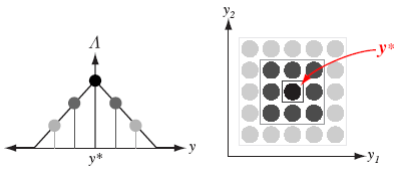
$$h_{j,i(x)} = \exp\left(-\frac{d_{j,i}^2}{2\sigma^2}\right)$$

$$\sigma(n) = \sigma_0 \exp\left(-\frac{n}{\tau_1}\right), \quad n = 0,1,2,3$$



Cooperative Process

- Typical window function



Adaptive process

- Synaptic weight vector is changed in relation with input vector

$$w_j(n+1) = w_j(n) + \eta(n)\, h_{j,i(x)}(n)\, (x - w_j(n))$$

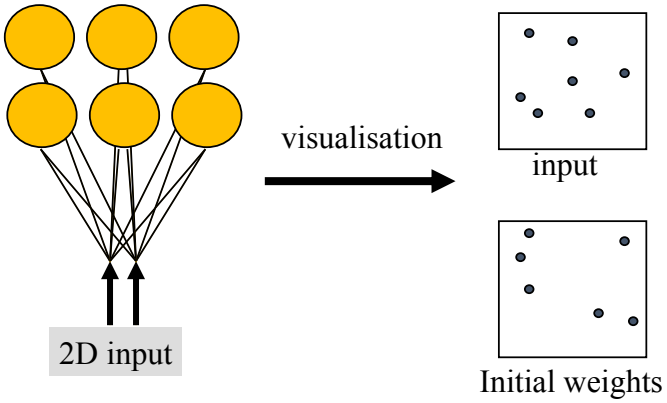
- applied to all neurons inside the neighborhood of winning neuron i
  - effect of moving weight w_j toward input vector x
- upon repeated presentation of the training data, weight tend to follow the distribution
- Learning rate $\eta(n)$: may decay with time

SOFM algorithm

1. initialize w's by random number
2. For input $\underline{x}(n)$, find nearest cell
   $i(\underline{x}) = \text{argmin}_j \| \underline{x}(n) - \underline{w}_j(n) \|$
3. update weights of neighbors
   $\underline{w}_j(n+1) = \underline{w}_j(n) + \eta(n)\, h_{j,i(x)}(n) [\, \underline{x}(n) - \underline{w}_j(n) \,]$
4. reduce neighbors and η
5. Go to 2

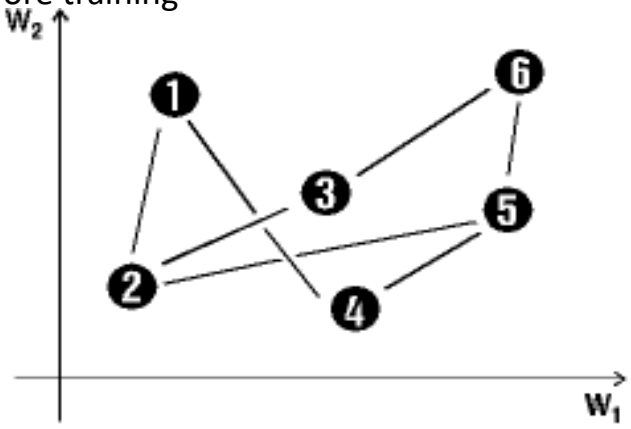
A simple example

- A topological map of 2 x 3 neurons and two inputs



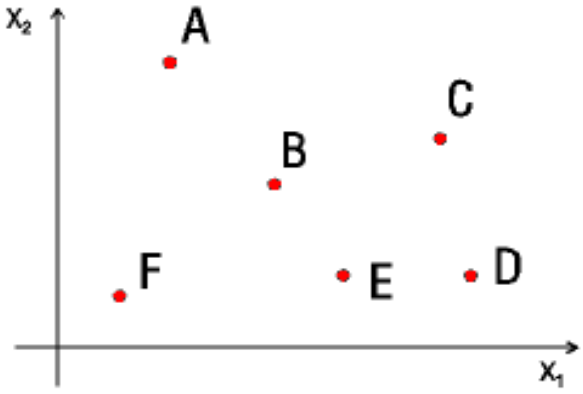
A simple example

- Weights before training



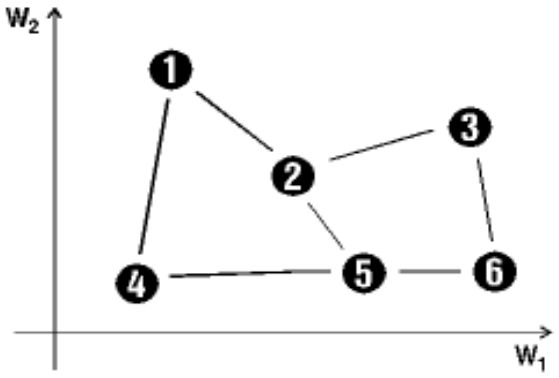
A simple example

- Input patterns (note the 2D distribution)

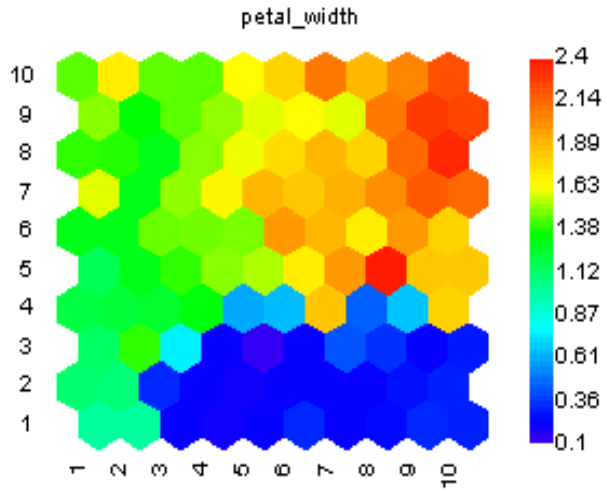


A simple example

- Weights after training



SOM Component Plane

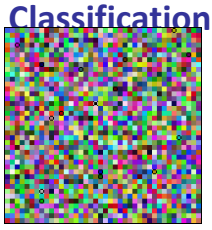


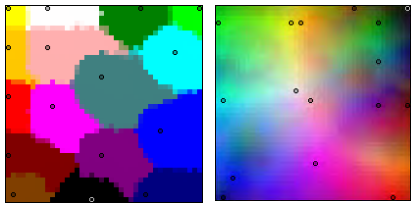
SOFM Applications

- http://www-ti.informatik.uni-tuebingen.de/~goeppert/KohonenApp/KohonenApp.html
- http://davis.wpi.edu/~matt/courses/soms/applet.html

SOFM Applications

Colour Classification





Car Clustering

Case Study: Travelers' Profile

- Background
- The Model
- The Result

Motivation

- **Significant contributions** of tourism receipts to a economy
- However, **few** attempts to ***understand the profile and behavioral patterns*** of international travelers
  - Hong Kong Tourism Board [HKTB] only show the number and percentages of travelers from different major source markets
- Expected to know
  - Who are they?
  - What they like?

Background: Data Set

- Traveler Expenditure Data Set
  - 1,282 visitors were interviewed at the HKIA
    - 303 of them named themselves as ***business travelers***
  - The following information was collected for visitors
    - Basic Variables
      - *Demographic Information*,
      - *Trip Information*, and
      - *Expenditure Information*
    - Associate Variables
      - *Motivation Information*,
      - *Activities Information*,
      - *Satisfactory Information*

Basic Variables

Demographic Information		Trip Information		Expenditure Information	
Variable	Description	Variable	Description	Variable	Description
LANGUAGE	Which language does the questionnaire use?	DESTINAT	Flight destination	EXPENSE	Total Expenses
COUNTRY	Country of Residence	RETURNHO	Does the respondent return home?	EXPENCUR	Total expenses (currency)
	Province in	TOTALLOS	Total length of stay (whole trip)	EXPENHKD	Expenses in HK in HKD
WEUROPE	Western European Countries	HKLOS	Length of Stay in HK?	TOTALEXR	Recoded total expenses in HKD
GENDER	Gender	FIRSTVHK	First Visit to HK?		
	Age	TRIPNO	Number of Trips made to HK including the current one		
EDUCATIO	Highest education level attained	ONLYDEST	Is HK the only destination you will visit during this trip?		
INCOME	Annual household income	MAINDEST	HK as the main destination?		
SIMILARI	How similar is HK's culture to your home culture?	MAINPURP	Main purpose for visiting HK?		
INTERN_T	International travel experience	TRAVELMO	Mode of Travel		
		TTPARTY	Total travel party in your group		
		TOTALLOR	Recoded total length of stay (whole trip)		
		HKLOSR	Recoded length of stay in HK		
		TTPARTYR	Recoded total travel party in your group		

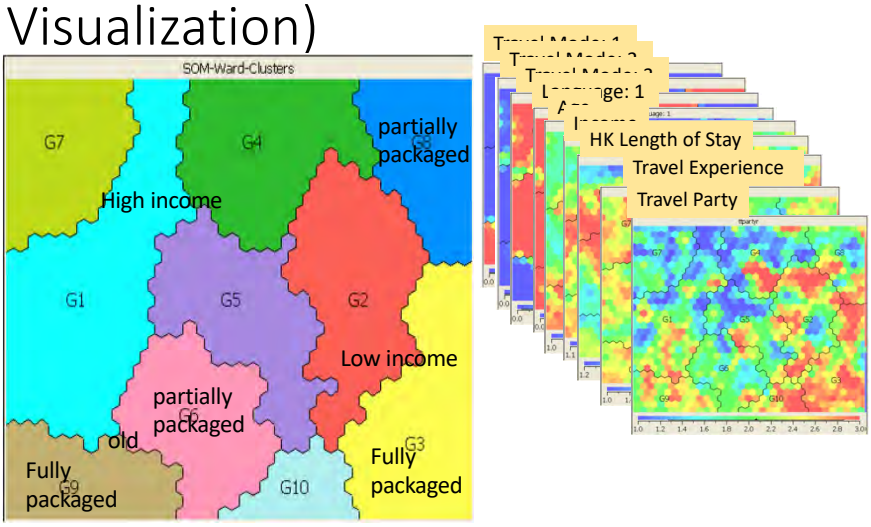
Associate Variables

Motivation Information		Activities Information		Satisfaction Indicator	
Variable	Description	Variable	Description	Variable	Description
motivat1	spend time with family, friends and relatives	act1	sightseeing	attractr	Attractiveness of (recode)
motivat2	meet different people	act2	shopping	satisfar	Satisfaction of (recode)
motivat3	rest and relax	act3	eating different foods	serquar	Overall service quality in (recode)
motivat4	get away from daily routine	act4	visiting nightclubs	vfmoner	Value for money (recode)
motivat5	discover new places and/or things	act5	visiting museums	returnr	Likelihood of return to (recode)
motivat6	increase my knowledge	act6	ecotourism		
motivat7	do business	act7	visiting religious sites		
motivat8	a convenient stopover before or after visiting	act8	visiting theme parks		
		act9	going to beaches		
		act10	playing sports		
		act11	riding public transport		
		act12	visiting		
		act13	attending family events		
		act14	family gathering		
		act15	attending seminars		
		act16	visiting festivals		
		act17	cross border tourism		

Variable Selection for SOM

- We intended to use SOM to
  - Identify the inherent segments of International Travellers
  - Identify How many Segments?
  - Find out how similar one segment is to the other ones
- Variable Selection
  - Use *Basic Variables*, so the segments are based on the similarity on the demographic, trip and expenditure information
  - After Removing Redundant Variables, we have 15 variables
    - *TTPARTYR*, *TRAVELMO*, *FIRSTVHK*, *GENDER*, *AGE*, *INCOME*, *COUNTRY*, *EDUCATIO*, *TOTALLOR*, *LANGUAGE*, *SIMILARI*, *INTERN_T*, *RETURNHO*, *ONLYDEST*, and *HKLOSR*

SOM Output (for Visualization)

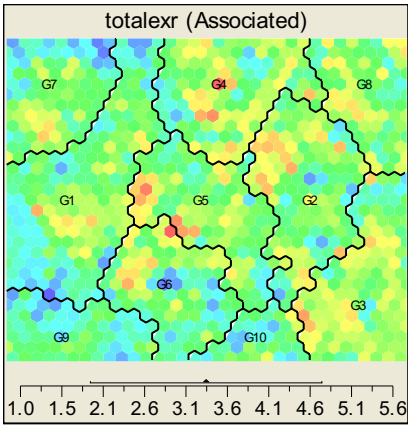


Segment Profile

Segment	Frequency	Features
G1	15.68%	English speaking, HK not as the only Destination, Independent Travelers, High income, Returning Home after HK
G2	9.67%	Chinese speaking, Independent Travelers, First Time visitor
G3	15.52%	Chinese speaking, Full Packaged Visitors, Returning Home
G4	10.76%	Chinese speaking, Not First Time Visitor, Independent Travelers
G5	7.41%	English speaking, Staying Longest in HK, Independent Travelers, HK as the only destination
G6	7.57%	English speaking, partially-packaged visitors
G7	9.28%	Not Returning Home, Travel alone, English speaking, High Education, Independent Travelers, High income
G8	9.44%	Chinese speaking, partially-packaged visitors
G9	11.00%	English speaking, Full Packaged Visitors, Senior Age, High income, Stay overseas longer (2-3 weeks)
G10	3.67%	Not Returning Home, Travel with 3 or more

Who Spend More?

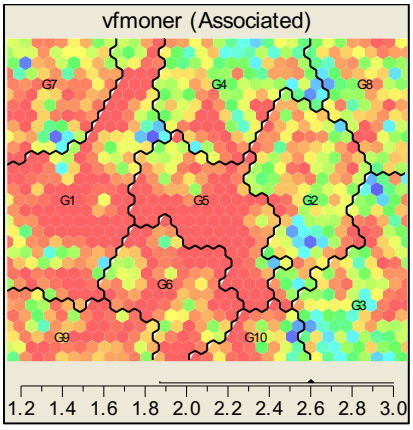
- Superimpose TOTALEXR over the map



Segment		*MAINPURP*	*TOTALEXR*	*VFMONER*	*RETURNR*
G1	Mean	1.522	3.040	2.770	2.637
	std	.794	1.281	.591	.737
G2	Mean	1.556	3.549	2.532	2.766
	std	1.171	1.494	.737	.626
G3	Mean	1.276	**3.716**	*2.372*	2.698
	std	.931	**1.314**	*.848*	.703
G4	Mean	**1.899**	3.654	2.384	**2.775**
	std	**1.083**	1.405	.840	**.616**
G5	Mean	1.811	3.500	**2.863**	2.600
	std	1.075	1.365	**.452**	.751
G6	Mean	1.680	3.260	2.825	2.722
	std	1.123	1.430	.540	.657
G7	Mean	1.798	2.940	2.619	2.597
	std	.869	1.434	.715	.751
G8	Mean	1.669	3.639	2.529	2.744
	std	1.028	1.394	.731	.639
G9	Mean	*1.043*	*2.779*	2.738	*2.298*
	std	*.356*	*1.474*	.605	*.924*
G10	Mean	1.255	2.826	2.617	2.702
	std	.765	1.354	.709	.720
Total	Mean	1.54	2.826	2.617	2.702
	std	.971	1.426	.718	.732

Value for Money?

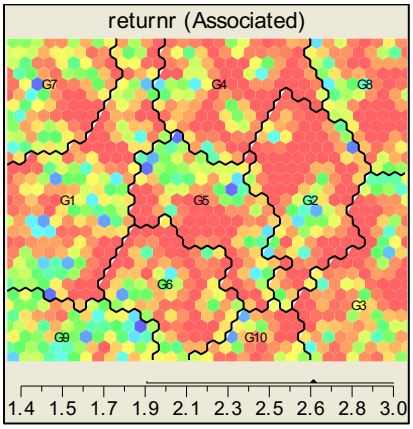
- Superimpose VFMONER over the map



Segment		*MAINPURP*	*TOTALEXR*	*VFMONER*	*RETURNR*
G1	Mean	1.522	3.040	2.770	2.637
	std	.794	1.281	.591	.737
G2	Mean	1.556	3.549	2.532	2.766
	std	1.171	1.494	.737	.626
G3	Mean	1.276	**3.716**	*2.372*	2.698
	std	.931	**1.314**	*.848*	.703
G4	Mean	**1.899**	3.654	2.384	**2.775**
	std	**1.083**	1.405	.840	**.616**
G5	Mean	1.811	3.500	**2.863**	2.600
	std	1.075	1.365	**.452**	.751
G6	Mean	1.680	3.260	2.825	2.722
	std	1.123	1.430	.540	.657
G7	Mean	1.798	2.940	2.619	2.597
	std	.869	1.434	.715	.751
G8	Mean	1.669	3.639	2.529	2.744
	std	1.028	1.394	.731	.639
G9	Mean	*1.043*	*2.779*	2.738	*2.298*
	std	*.356*	*1.474*	.605	*.924*
G10	Mean	1.255	2.826	2.617	2.702
	std	.765	1.354	.709	.720
Total	Mean	1.54	2.826	2.617	2.702
	std	.971	1.426	.718	.732

Likelihood of Return

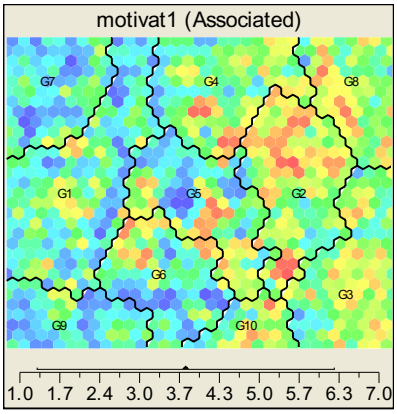
- Superimpose RETURNR over the map



Segment		*MAINPURP*	*TOTALEXR*	*VFMONER*	*RETURNR*
G1	Mean	1.522	3.040	2.770	2.637
	std	.794	1.281	.591	.737
G2	Mean	1.556	3.549	2.532	2.766
	std	1.171	1.494	.737	.626
G3	Mean	1.276	**3.716**	*2.372*	2.698
	std	.931	**1.314**	*.848*	.703
G4	Mean	**1.899**	3.654	2.384	**2.775**
	std	**1.083**	1.405	.840	**.616**
G5	Mean	1.811	3.500	**2.863**	2.600
	std	1.075	1.365	**.452**	.751
G6	Mean	1.680	3.260	2.825	2.722
	std	1.123	1.430	.540	.657
G7	Mean	1.798	2.940	2.619	2.597
	std	.869	1.434	.715	.751
G8	Mean	1.669	3.639	2.529	2.744
	std	1.028	1.394	.731	.639
G9	Mean	*1.043*	*2.779*	2.738	*2.298*
	std	*.356*	*1.474*	.605	*.924*
G10	Mean	1.255	2.826	2.617	2.702
	std	.765	1.354	.709	.720
Total	Mean	1.54	2.826	2.617	2.702
	std	.971	1.426	.718	.732

Motivation of Travellers (M1: Spend with Family)

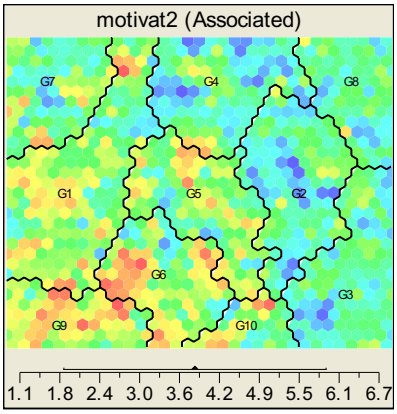
- Superimpose Motivation1 over the map



Segment		*MOTIVAT1*	*MOTIVAT2*	*MOTIVAT3*	*MOTIVAT4*	*MOTIVAT5*	*MOTIVAT6*	*MOTIVAT7*	*MOTIVAT8*
G1	Mean	3.120	4.307	4.335	4.327	5.455	5.447	3.208	3.528
	std	2.420	2.005	2.092	2.254	1.997	1.768	2.605	2.437
G2	Mean	**4.960**	3.113	5.669	4.540	5.105	5.218	2.098	2.721
	std	**2.210**	1.777	1.733	2.038	1.941	1.769	1.856	2.141
G3	Mean	4.460	3.226	**5.734**	4.613	5.236	5.414	1.843	2.444
	std	2.259	1.887	**1.444**	1.948	1.792	1.640	1.522	1.963
G4	Mean	4.109	*3.103*	4.551	4.170	*4.185*	*4.191*	4.239	3.125
	std	2.449	*1.956*	2.234	2.163	*2.078*	*2.006*	2.665	2.327
G5	Mean	3.674	4.263	4.042	4.442	5.337	5.221	3.558	*2.074*
	std	2.595	1.964	2.212	2.196	2.097	2.043	2.665	*1.758*
G6	Mean	3.381	4.845	4.742	**5.072**	6.010	5.917	2.887	3.608
	std	2.539	1.944	2.103	**2.048**	1.539	1.554	2.474	2.321
G7	Mean	2.748	3.822	*3.992*	*4.026*	4.658	4.847	**4.286**	**4.254**
	std	2.233	2.082	*2.098*	*2.246*	2.182	2.078	**2.662**	**2.461**
G8	Mean	4.667	3.342	5.250	4.538	4.933	5.025	3.562	2.892
	std	2.201	1.845	1.898	1.974	1.922	1.840	2.565	2.169
G9	Mean	*2.652*	**4.914**	4.241	4.763	**6.504**	**6.243**	*1.400*	4.079
	std	*2.214*	**1.984**	2.056	2.045	**.976**	**1.174**	*1.065*	2.390
G10	Mean	4.362	4.383	5.340	4.660	5.830	5.702	2.340	3.468
	std	2.335	1.929	1.536	1.868	1.340	1.502	1.821	2.339
Total	Mean	3.78	3.87	4.80	4.49	5.29	**5.31**	*2.91*	3.20
	std	2.462	2.047	2.055	2.105	1.947	**1.834**	*2.441*	2.329

Motivation of Travellers (M2: Meet Different People)

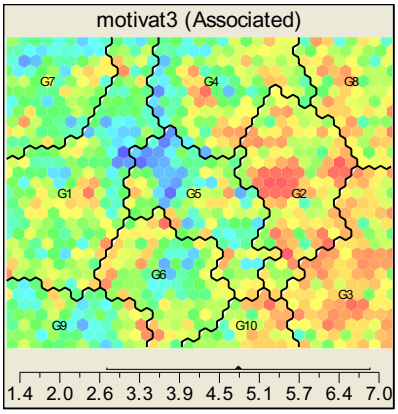
- Superimpose Motivation2 over the map



Segment		MOTIVAT1	MOTIVAT2	MOTIVAT3	MOTIVAT4	MOTIVAT5	MOTIVAT6	MOTIVAT7	MOTIVAT8
G1	Mean	3.120	4.307	4.335	4.327	5.455	5.447	3.208	3.528
	std	2.420	2.005	2.092	2.254	1.997	1.768	2.605	2.437
G2	Mean	**4.960**	3.113	5.669	4.540	5.105	5.218	2.098	2.721
	std	**2.210**	1.777	1.733	2.038	1.941	1.769	1.856	2.141
G3	Mean	4.460	3.226	**5.734**	4.613	5.236	5.414	1.843	2.444
	std	2.259	1.887	**1.444**	1.948	1.792	1.640	1.522	1.963
G4	Mean	4.109	*3.103*	4.551	4.170	*4.185*	*4.191*	4.239	3.125
	std	2.449	*1.956*	2.234	2.163	*2.078*	*2.006*	2.665	2.327
G5	Mean	3.674	4.263	4.042	4.442	5.337	5.221	3.558	*2.074*
	std	2.595	1.964	2.212	2.196	2.097	2.043	2.665	*1.758*
G6	Mean	3.381	4.845	4.742	**5.072**	6.010	5.917	2.887	3.608
	std	2.539	1.944	2.103	**2.048**	1.539	1.554	2.474	2.321
G7	Mean	2.748	3.822	*3.992*	*4.026*	4.658	4.847	**4.286**	**4.254**
	std	2.233	2.082	*2.098*	*2.246*	2.182	2.078	**2.662**	**2.461**
G8	Mean	4.667	3.342	5.250	4.538	4.933	5.025	3.562	2.892
	std	2.201	1.845	1.898	1.974	1.922	1.840	2.565	2.169
G9	Mean	*2.652*	**4.914**	4.241	4.763	**6.504**	**6.243**	*1.400*	4.079
	std	*2.214*	**1.984**	2.056	2.045	**.976**	**1.174**	*1.065*	2.390
G10	Mean	4.362	4.383	5.340	4.660	5.830	5.702	2.340	3.468
	std	2.335	1.929	1.536	1.868	1.340	1.502	1.821	2.339
Total	Mean	3.78	3.87	4.80	4.49	5.29	**5.31**	*2.91*	3.20
	std	2.462	2.047	2.055	2.105	1.947	**1.834**	*2.441*	2.329

Motivation of Travellers (M3: Spend with Family)

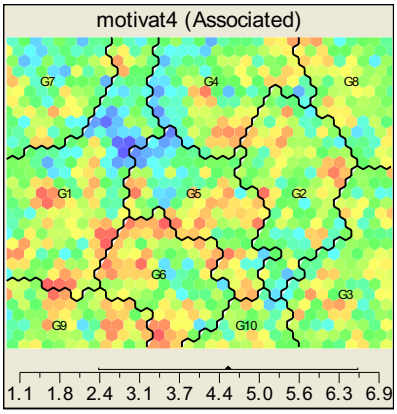
- Superimpose Motivation3 over the map



Segment		MOTIVAT1	MOTIVAT2	MOTIVAT3	MOTIVAT4	MOTIVAT5	MOTIVAT6	MOTIVAT7	MOTIVAT8
G1	Mean	3.120	4.307	4.335	4.327	5.455	5.447	3.208	3.528
	std	2.420	2.005	2.092	2.254	1.997	1.768	2.605	2.437
G2	Mean	**4.960**	3.113	5.669	4.540	5.105	5.218	2.098	2.721
	std	**2.210**	1.777	1.733	2.038	1.941	1.769	1.856	2.141
G3	Mean	4.460	3.226	**5.734**	4.613	5.236	5.414	1.843	2.444
	std	2.259	1.887	**1.444**	1.948	1.792	1.640	1.522	1.963
G4	Mean	4.109	*3.103*	4.551	4.170	*4.185*	*4.191*	4.239	3.125
	std	2.449	*1.956*	2.234	2.163	*2.078*	*2.006*	2.665	2.327
G5	Mean	3.674	4.263	4.042	4.442	5.337	5.221	3.558	*2.074*
	std	2.595	1.964	2.212	2.196	2.097	2.043	2.665	*1.758*
G6	Mean	3.381	4.845	4.742	**5.072**	6.010	5.917	2.887	3.608
	std	2.539	1.944	2.103	**2.048**	1.539	1.554	2.474	2.321
G7	Mean	2.748	3.822	*3.992*	*4.026*	4.658	4.847	**4.286**	**4.254**
	std	2.233	2.082	*2.098*	*2.246*	2.182	2.078	**2.662**	**2.461**
G8	Mean	4.667	3.342	5.250	4.538	4.933	5.025	3.562	2.892
	std	2.201	1.845	1.898	1.974	1.922	1.840	2.565	2.169
G9	Mean	*2.652*	**4.914**	4.241	4.763	**6.504**	**6.243**	*1.400*	4.079
	std	*2.214*	**1.984**	2.056	2.045	**.976**	**1.174**	*1.065*	2.390
G10	Mean	4.362	4.383	5.340	4.660	5.830	5.702	2.340	3.468
	std	2.335	1.929	1.536	1.868	1.340	1.502	1.821	2.339
Total	Mean	3.78	3.87	4.80	4.49	5.29	**5.31**	*2.91*	3.20
	std	2.462	2.047	2.055	2.105	1.947	**1.834**	*2.441*	2.329

Motivation of Travellers (M4: Get Away from Routine)

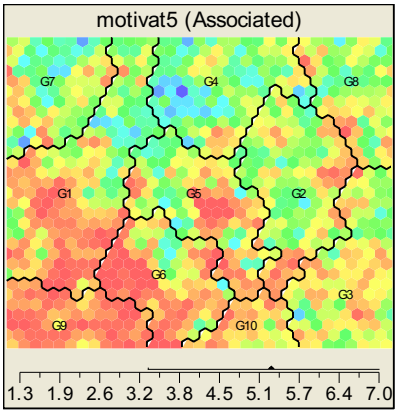
- Superimpose Motivation4 over the map



Segment		MOTIVAT1	MOTIVAT2	MOTIVAT3	MOTIVAT4	MOTIVAT5	MOTIVAT6	MOTIVAT7	MOTIVAT8
G1	Mean	3.120	4.307	4.335	4.327	5.455	5.447	3.208	3.528
	std	2.420	2.005	2.092	2.254	1.997	1.768	2.605	2.437
G2	Mean	**4.960**	3.113	5.669	4.540	5.105	5.218	2.098	2.721
	std	**2.210**	1.777	1.733	2.038	1.941	1.769	1.856	2.141
G3	Mean	4.460	3.226	**5.734**	4.613	5.236	5.414	1.843	2.444
	std	2.259	1.887	**1.444**	1.948	1.792	1.640	1.522	1.963
G4	Mean	4.109	*3.103*	4.551	4.170	*4.185*	*4.191*	4.239	3.125
	std	2.449	*1.956*	2.234	2.163	*2.078*	*2.006*	2.665	2.327
G5	Mean	3.674	4.263	4.042	4.442	5.337	5.221	3.558	*2.074*
	std	2.595	1.964	2.212	2.196	2.097	2.043	2.665	*1.758*
G6	Mean	3.381	4.845	4.742	**5.072**	6.010	5.917	2.887	3.608
	std	2.539	1.944	2.103	**2.048**	1.539	1.554	2.474	2.321
G7	Mean	2.748	3.822	*3.992*	*4.026*	4.658	4.847	**4.286**	**4.254**
	std	2.233	2.082	*2.098*	*2.246*	2.182	2.078	**2.662**	**2.461**
G8	Mean	4.667	3.342	5.250	4.538	4.933	5.025	3.562	2.892
	std	2.201	1.845	1.898	1.974	1.922	1.840	2.565	2.169
G9	Mean	*2.652*	**4.914**	4.241	4.763	**6.504**	**6.243**	*1.400*	4.079
	std	*2.214*	**1.984**	2.056	2.045	**.976**	**1.174**	*1.065*	2.390
G10	Mean	4.362	4.383	5.340	4.660	5.830	5.702	2.340	3.468
	std	2.335	1.929	1.536	1.868	1.340	1.502	1.821	2.339
Total	Mean	3.78	3.87	4.80	4.49	5.29	**5.31**	*2.91*	3.20
	std	2.462	2.047	2.055	2.105	1.947	**1.834**	*2.441*	2.329

Motivation of Travellers (M5: Discovery New Things)

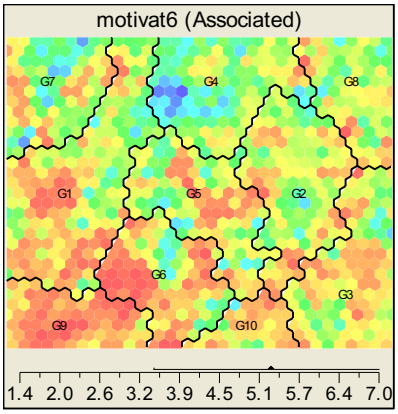
- Superimpose Motivation5 over the map



Segment		MOTIVAT1	MOTIVAT2	MOTIVAT3	MOTIVAT4	MOTIVAT5	MOTIVAT6	MOTIVAT7	MOTIVAT8
G1	Mean	3.120	4.307	4.335	4.327	5.455	5.447	3.208	3.528
	std	2.420	2.005	2.092	2.254	1.997	1.768	2.605	2.437
G2	Mean	**4.960**	3.113	5.669	4.540	5.105	5.218	2.098	2.721
	std	**2.210**	1.777	1.733	2.038	1.941	1.769	1.856	2.141
G3	Mean	4.460	3.226	**5.734**	4.613	5.236	5.414	1.843	2.444
	std	2.259	1.887	**1.444**	1.948	1.792	1.640	1.522	1.963
G4	Mean	4.109	*3.103*	4.551	4.170	*4.185*	*4.191*	4.239	3.125
	std	2.449	*1.956*	2.234	2.163	*2.078*	*2.006*	2.665	2.327
G5	Mean	3.674	4.263	4.042	4.442	5.337	5.221	3.558	*2.074*
	std	2.595	1.964	2.212	2.196	2.097	2.043	2.665	*1.758*
G6	Mean	3.381	4.845	4.742	**5.072**	6.010	5.917	2.887	3.608
	std	2.539	1.944	2.103	**2.048**	1.539	1.554	2.474	2.321
G7	Mean	2.748	3.822	*3.992*	*4.026*	4.658	4.847	**4.286**	**4.254**
	std	2.233	2.082	*2.098*	*2.246*	2.182	2.078	**2.662**	**2.461**
G8	Mean	4.667	3.342	5.250	4.538	4.933	5.025	3.562	2.892
	std	2.201	1.845	1.898	1.974	1.922	1.840	2.565	2.169
G9	Mean	*2.652*	**4.914**	4.241	4.763	**6.504**	**6.243**	*1.400*	4.079
	std	*2.214*	**1.984**	2.056	2.045	**.976**	**1.174**	*1.065*	2.390
G10	Mean	4.362	4.383	5.340	4.660	5.830	5.702	2.340	3.468
	std	2.335	1.929	1.536	1.868	1.340	1.502	1.821	2.339
Total	Mean	3.78	3.87	4.80	4.49	5.29	**5.31**	*2.91*	3.20
	std	2.462	2.047	2.055	2.105	1.947	**1.834**	*2.441*	2.329

Motivation of Travellers (M6: Increase Knowledge)

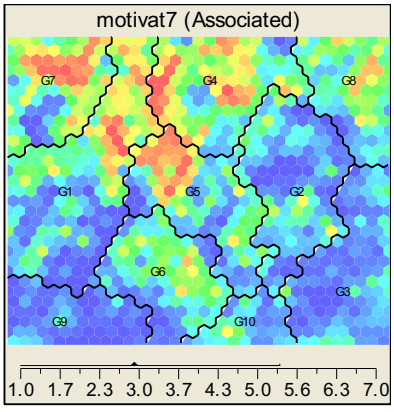
- Superimpose Motivation6 over the map



Segment		MOTIVAT1	MOTIVAT2	MOTIVAT3	MOTIVAT4	MOTIVAT5	MOTIVAT6	MOTIVAT7	MOTIVAT8
G1	Mean	3.120	4.307	4.335	4.327	5.455	5.447	3.208	3.528
	std	2.420	2.005	2.092	2.254	1.997	1.768	2.605	2.437
G2	Mean	**4.960**	3.113	5.669	4.540	5.105	5.218	2.098	2.721
	std	**2.210**	1.777	1.733	2.038	1.941	1.769	1.856	2.141
G3	Mean	4.460	3.226	**5.734**	4.613	5.236	5.414	1.843	2.444
	std	2.259	1.887	**1.444**	1.948	1.792	1.640	1.522	1.963
G4	Mean	4.109	*3.103*	4.551	4.170	*4.185*	*4.191*	4.239	3.125
	std	2.449	*1.956*	2.234	2.163	*2.078*	*2.006*	2.665	2.327
G5	Mean	3.674	4.263	4.042	4.442	5.337	5.221	3.558	*2.074*
	std	2.595	1.964	2.212	2.196	2.097	2.043	2.665	*1.758*
G6	Mean	3.381	4.845	4.742	**5.072**	6.010	5.917	2.887	3.608
	std	2.539	1.944	2.103	**2.048**	1.539	1.554	2.474	2.321
G7	Mean	2.748	3.822	*3.992*	*4.026*	4.658	4.847	**4.286**	**4.254**
	std	2.233	2.082	*2.098*	*2.246*	2.182	2.078	**2.662**	**2.461**
G8	Mean	4.667	3.342	5.250	4.538	4.933	5.025	3.562	2.892
	std	2.201	1.845	1.898	1.974	1.922	1.840	2.565	2.169
G9	Mean	*2.652*	**4.914**	4.241	4.763	**6.504**	**6.243**	*1.400*	4.079
	std	*2.214*	**1.984**	2.056	2.045	**.976**	**1.174**	*1.065*	2.390
G10	Mean	4.362	4.383	5.340	4.660	5.830	5.702	2.340	3.468
	std	2.335	1.929	1.536	1.868	1.340	1.502	1.821	2.339
Total	Mean	3.78	3.87	4.80	4.49	5.29	**5.31**	*2.91*	3.20
	std	2.462	2.047	2.055	2.105	1.947	**1.834**	*2.441*	2.329

Motivation of Travellers (M7: do Business)

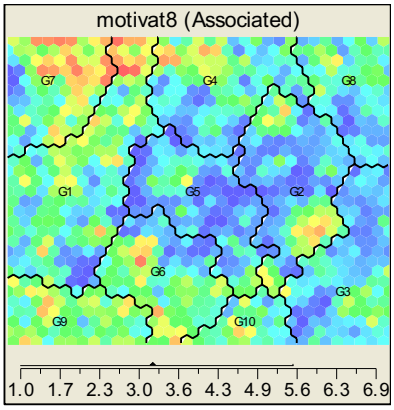
- Superimpose Motivation7 over the map



Segment		MOTIVAT1	MOTIVAT2	MOTIVAT3	MOTIVAT4	MOTIVAT5	MOTIVAT6	MOTIVAT7	MOTIVAT8
G1	Mean	3.120	4.307	4.335	4.327	5.455	5.447	3.208	3.528
	std	2.420	2.005	2.092	2.254	1.997	1.768	2.605	2.437
G2	Mean	**4.960**	3.113	5.669	4.540	5.105	5.218	2.098	2.721
	std	**2.210**	1.777	1.733	2.038	1.941	1.769	1.856	2.141
G3	Mean	4.460	3.226	**5.734**	4.613	5.236	5.414	1.843	2.444
	std	2.259	1.887	**1.444**	1.948	1.792	1.640	1.522	1.963
G4	Mean	4.109	*3.103*	4.551	4.170	*4.185*	*4.191*	4.239	3.125
	std	2.449	*1.956*	2.234	2.163	*2.078*	*2.006*	2.665	2.327
G5	Mean	3.674	4.263	4.042	4.442	5.337	5.221	3.558	*2.074*
	std	2.595	1.964	2.212	2.196	2.097	2.043	2.665	*1.758*
G6	Mean	3.381	4.845	4.742	**5.072**	6.010	5.917	2.887	3.608
	std	2.539	1.944	2.103	**2.048**	1.539	1.554	2.474	2.321
G7	Mean	2.748	3.822	*3.992*	*4.026*	4.658	4.847	**4.286**	**4.254**
	std	2.233	2.082	*2.098*	*2.246*	2.182	2.078	**2.662**	**2.461**
G8	Mean	4.667	3.342	5.250	4.538	4.933	5.025	3.562	2.892
	std	2.201	1.845	1.898	1.974	1.922	1.840	2.565	2.169
G9	Mean	*2.652*	**4.914**	4.241	4.763	**6.504**	**6.243**	*1.400*	4.079
	std	*2.214*	**1.984**	2.056	2.045	**.976**	**1.174**	*1.065*	2.390
G10	Mean	4.362	4.383	5.340	4.660	5.830	5.702	2.340	3.468
	std	2.335	1.929	1.536	1.868	1.340	1.502	1.821	2.339
Total	Mean	3.78	3.87	4.80	4.49	5.29	**5.31**	*2.91*	3.20
	std	2.462	2.047	2.055	2.105	1.947	**1.834**	*2.441*	2.329

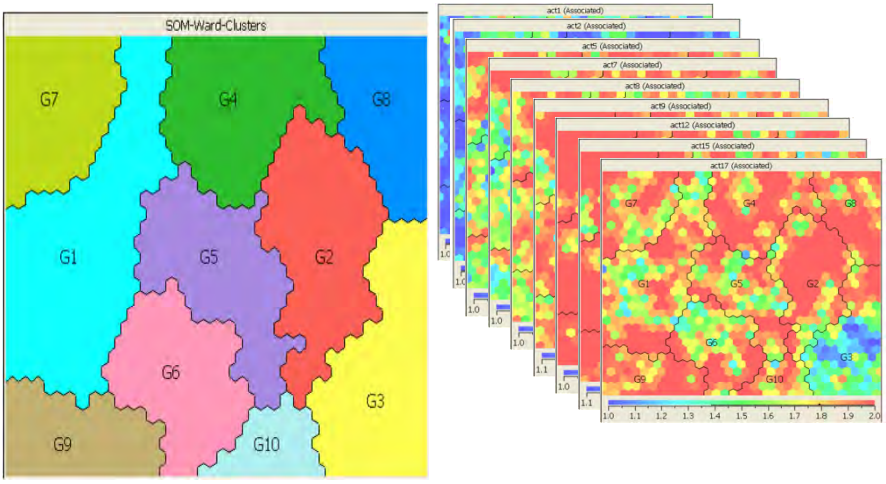
Motivation of Travellers (M8: Stopover for other city)

- Superimpose Motivation8 over the map



Segment		MOTIVAT1	MOTIVAT2	MOTIVAT3	MOTIVAT4	MOTIVAT5	MOTIVAT6	MOTIVAT7	MOTIVAT8
G1	Mean	3.120	4.307	4.335	4.327	5.455	5.447	3.208	3.528
	std	2.420	2.005	2.092	2.254	1.997	1.768	2.605	2.437
G2	Mean	**4.960**	3.113	5.669	4.540	5.105	5.218	2.098	2.721
	std	**2.210**	1.777	1.733	2.038	1.941	1.769	1.856	2.141
G3	Mean	4.460	3.226	**5.734**	4.613	5.236	5.414	1.843	2.444
	std	2.259	1.887	**1.444**	1.948	1.792	1.640	1.522	1.963
G4	Mean	4.109	*3.103*	4.551	4.170	*4.185*	*4.191*	4.239	3.125
	std	2.449	*1.956*	2.234	2.163	*2.078*	*2.006*	2.665	2.327
G5	Mean	3.674	4.263	4.042	4.442	5.337	5.221	3.558	*2.074*
	std	2.595	1.964	2.212	2.196	2.097	2.043	2.665	*1.758*
G6	Mean	3.381	4.845	4.742	**5.072**	6.010	5.917	2.887	3.608
	std	2.539	1.944	2.103	**2.048**	1.539	1.554	2.474	2.321
G7	Mean	2.748	3.822	*3.992*	*4.026*	4.658	4.847	**4.286**	**4.254**
	std	2.233	2.082	*2.098*	*2.246*	2.182	2.078	**2.662**	**2.461**
G8	Mean	4.667	3.342	5.250	4.538	4.933	5.025	3.562	2.892
	std	2.201	1.845	1.898	1.974	1.922	1.840	2.565	2.169
G9	Mean	*2.652*	**4.914**	4.241	4.763	**6.504**	**6.243**	*1.400*	4.079
	std	*2.214*	**1.984**	2.056	2.045	**.976**	**1.174**	*1.065*	2.390
G10	Mean	4.362	4.383	5.340	4.660	5.830	5.702	2.340	3.468
	std	2.335	1.929	1.536	1.868	1.340	1.502	1.821	2.339
Total	Mean	3.78	3.87	4.80	4.49	5.29	**5.31**	*2.91*	3.20
	std	2.462	2.047	2.055	2.105	1.947	**1.834**	*2.441*	2.329

Activity Pattern Analysis



Activity Pattern Profile

Segment	Features
G1	English speaking, HK not as the only Destination, Independent Travelers, High income, Returning Home after HK
G2	Chinese speaking, Independent Travelers, First Time visitor
G3	Chinese speaking, Full Packaged Visitors, Returning Home
G4	Chinese speaking, Not First Time Visitor, Independent Travelers
G5	English speaking, Staying Longest in HK, Independent Travelers, HK as the only destination
G6	English speaking, partially-packaged visitors
G7	Not Returning Home, Travel alone, English speaking, High Education, Independent Travelers, High income
G8	Chinese speaking, partially-packaged visitors
G9	English speaking, Full Packaged Visitors, Senior Age, High income, Stay overseas longer (2-3 weeks)
G10	Not Returning Home, Travel with 3 or more

- **Visitors in G1**
  - were most interested in ACT11 (riding public transportation),
  - but least interested in ACT13 (attending family events).
- **Visitors in G2**
  - were most interested in ACT12 (visiting Disneyland).
- **Visitors in G3**
  - were most interested in ACT2 (shopping), ACT7 (visiting churches), ACT8 (visiting theme parks), ACT9 (going to beaches),
  - but least interested in ACT3 (food) and ACT14 (family gathering).

Activity Pattern Profile

Segment	Features
G1	English speaking, HK not as the only Destination, Independent Travelers, High income, Returning Home after HK
G2	Chinese speaking, Independent Travelers, First Time visitor
G3	Chinese speaking, Full Packaged Visitors, Returning Home
G4	Chinese speaking, Not First Time Visitor, Independent Travelers
G5	English speaking, Staying Longest in HK, Independent Travelers, HK as the only destination
G6	English speaking, partially-packaged visitors
G7	Not Returning Home, Travel alone, English speaking, High Education, Independent Travelers, High income
G8	Chinese speaking, partially-packaged visitors
G9	English speaking, Full Packaged Visitors, Senior Age, High income, Stay overseas longer (2-3 weeks)
G10	Not Returning Home, Travel with 3 or more

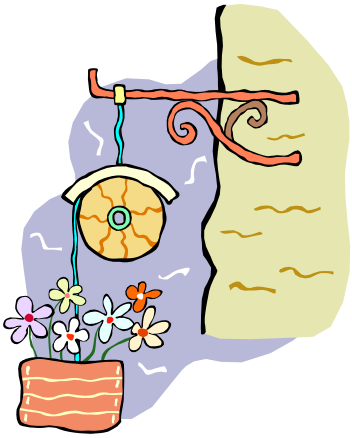
- **Visitors in G4**
  - most interested in ACT15 (attending training), but least interested in ACT1 (sightseeing), ACT5 (visiting museum) and ACT7 (visiting churches).
- **Visitors in G5**
  - most interested in ACT3 (food), ACT4 (visiting nightclub), ACT10 (sports) and ACT14 (family gathering).
- **Visitors in G6**
  - least interested in *ACT6* (eco-tourism).
- **Visitors in G7**
  - least interested in ACT8 (visiting theme park) and ACT9 (going to beach).

Activity Pattern Profile

Segment	Features
G1	English speaking, HK not as the only Destination, Independent Travelers, High income, Returning Home after HK
G2	Chinese speaking, Independent Travelers, First Time visitor
G3	Chinese speaking, Full Packaged Visitors, Returning Home
G4	Chinese speaking, Not First Time Visitor, Independent Travelers
G5	English speaking, Staying Longest in HK, Independent Travelers, HK as the only destination
G6	English speaking, partially-packaged visitors
G7	Not Returning Home, Travel alone, English speaking, High Education, Independent Travelers, High income
G8	Chinese speaking, partially-packaged visitors
G9	English speaking, Full Packaged Visitors, Senior Age, High income, Stay overseas longer (2-3 weeks)
G10	Not Returning Home, Travel with 3 or more

- **Visitors in G8**
  - most interested in *ACT13* (attending family events).
- **Visitors in G9**
  - most interested in *ACT5* (visiting museum) and *ACT16* (visiting festival),
  - but least interested in *ACT10* (sports), *ACT12* (visiting Disneyland), *ACT15* (attending training) and *ACT17* (cross-border tourism).
- **Visitors in G10**
  - most interested in *ACT1* (sightseeing) and *ACT6* (eco-tourism),
  - but least interested in *ACT4* (visiting nightclub), *ACT11* (riding public transportation) and *ACT16* (visiting festival).

Market Basket Analysis



- Market Basket Analysis

Market Baskets Analysis

- People go shopping everyday, So ***Transaction data*** are accumulated day by day, month by month and year by year.
  - And eventually a huge amount of data are collected and become available
- Then the market analyser want to know **what is the customers' *consumption preference*.**

Market Basket Analysis

- Analyze tables of transactions
- Which items are frequently purchased together by customers?

Customer	Basket
C_1	Chips, Salsa, Cookies, Crackers, Coke, Beer
C_2	Lettuce, Spinach, Oranges, Celery, Apples, Grapes
C_3	Chips, Salsa, Frozen Pizza, Frozen Cake
C_4	Lettuce, Spinach, Milk, Butter

Market Baskets Analysis

- Did customers buy chips also buy Salsa?

Chips ⇒ Salsa

- Did customers prefer buying Lettuce together with Spinach?

Lettuce ⇒ Spinach

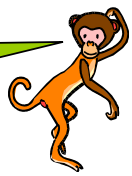
(Positive) Associations Rules

- Association Rules express relationships between items
  - e.g.

cereal, milk ⇒ fruit

  - "*Peoples who bought cereal and milk also bought fruit*"

Stores might want to offer specials on milk and cereal to get people to buy more fruit.



Measuring Interesting (Positive) Rules

- Rules are mined based on two metrics

$$\sup port(A \rightarrow B) = \frac{\#\_of\_Trans\_containing\_both\_A\_and\_B}{\#\_of\_Trans}$$

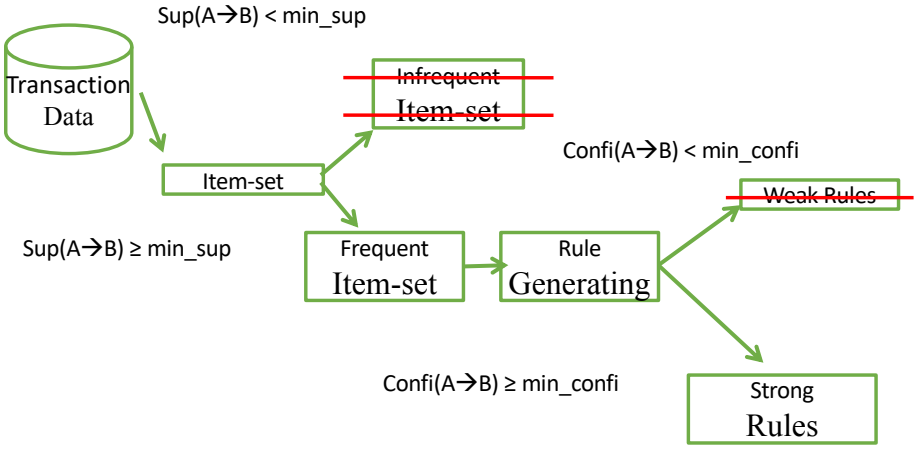
$$confidence(A \rightarrow B) = \frac{\#\_of\_Trans\_containing\_both\_A\_and\_B}{\#\_of\_Trans\_containing\_A}$$

Market Basket Analysis

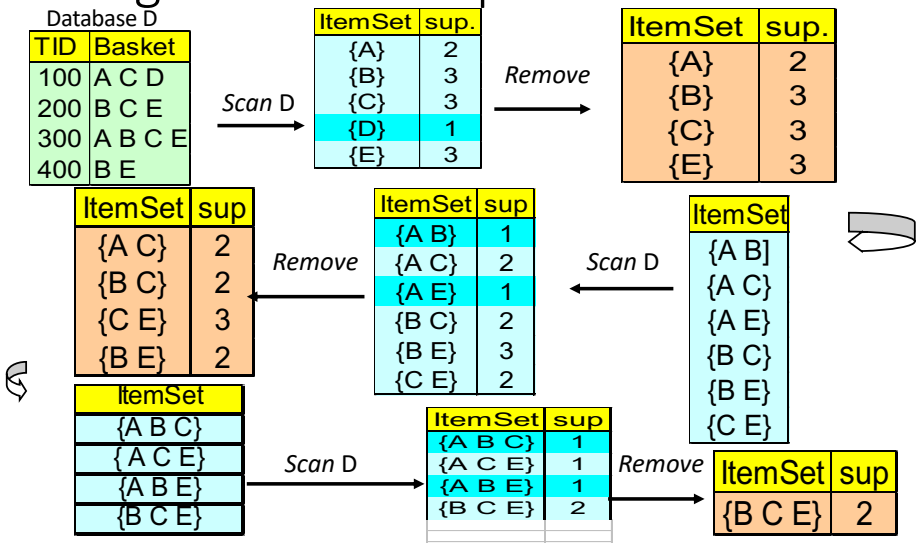
Transaction	Item Sets
T_1	Chips, Salsa, Cookies, Crackers, Coke, Beer
T_2	Lettuce, Spinach, Oranges, Celery, Apples, Grapes
T_3	Chips, Salsa, Frozen Pizza, Frozen Cake
T_4	Lettuce, Spinach, Milk, Butter

- What is ***support(Chips=>Salsa)***?
  - 2/4 = 0.5
- What is ***confidence(Chips=>Salsa)***?
  - 2/2 = 1

Association Rule Mining



Apriori Algorithm: Example



Apriori Algorithm: Example

Database D

TID	Trans.
100	A C D
200	B C E
300	A B C E
400	B E

Possible Association Rules that can be derived from {*B C E*}:

B, C --> E	*support= 2/4*	*confidence = 2/2 = 1.0*
B, E --> C	*support = 2/4*	*confidence = 2/3 = 0.6*
C, E --> B	*support = 2/4*	*confidence = 2/2 = 1.0*
B --> C, E	*support = 2/4*	*confidence = 2/3 = 0.6*
C --> B, E	*support = 2/4*	*confidence = 2/3 = 0.6*
E --> B, C	*support = 2/4*	*confidence = 2/3 = 0.6*

Limitations of (Positive) Association Rules

- **Problem #1**
  - All possible rules A=>B are generated, User can not specify the target value of B that they are interested.
- For example :

Transaction	Item Sets
T_1	Chips, Salsa, Cookies, Crackers, Coke, Beer
T_2	Lettuce, Spinach, Oranges, Celery, Apples, Grapes
T_3	Chips, Salsa, Frozen Pizza, Frozen Cake
T_4	Lettuce, Spinach, Milk, Butter

- {Chips, Salsa} => {Cookies}
- {Chips, Salsa} => {Frozen Pizza}
- {Crackers, Coke}=> {Beer}
- .....

Waste computing time and effort to extract The un-interested rule

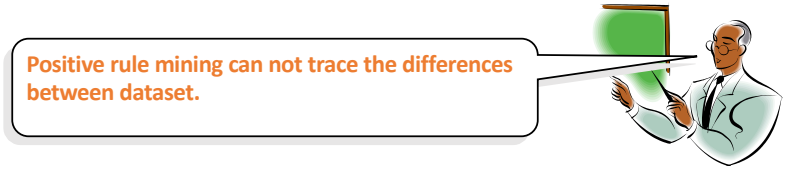
Limitations of (Positive) Association Rules

- **Problem #2** (in tourism research)
  - (Positive) Association Rule mining can find rule such as **Male => Macau**
  - But never find the rules such as **Female=> ~ Macau**

- Consider the following case
  - **Female** is frequent item, **Macau** is also a frequent item
  - **Female** U **Macau** is infrequent
    - Though the support(**Female** *U* **~ Macau**) can be high.

Limitations of (Positive) Association Rules

- **Problem #3**
  - But (Positive) Association Rule mining can only work on single data set. A strong rule in this data set may not be strong in other data sets.
- Consider the following case
  - **Male => Macau** is a strong rule in 2009 dataset, but not a strong rule in 2010 dataset.
  - **High Income =>~ Oversea** is a strong rule in 2009 dataset, but weak in 2010 dataset.



Targeted Negative Rules Mining

- The Problem Defined
- Targeted Negative Rules
- The Algorithm

Targeted Association Rules

- **Target Rules**:
  - Let $\mathbf{B_t}$ be the target item, our method focuses on the mining of targeted rules, including
    - Positive Rules with the form **A** => $\mathbf{B_t}$
    - Negative Rules with the form **A** => ~ $\mathbf{B_t}$

Male, High Income => Travel to Overseas

Female, Low Income => ~Travel to Overseas

Being Tourism Manager, I am only interested in Tourists who travel to Overseas

Negative Association Rules

- **Negative Association Rules**:
  - a rule that contains a negation of an item.
    - **A=>~B, ~A=>B, ~A=>~B**

- Examining both item-set having support above and below *min_support*.

Female, Low Income → ~Travel to Overseas

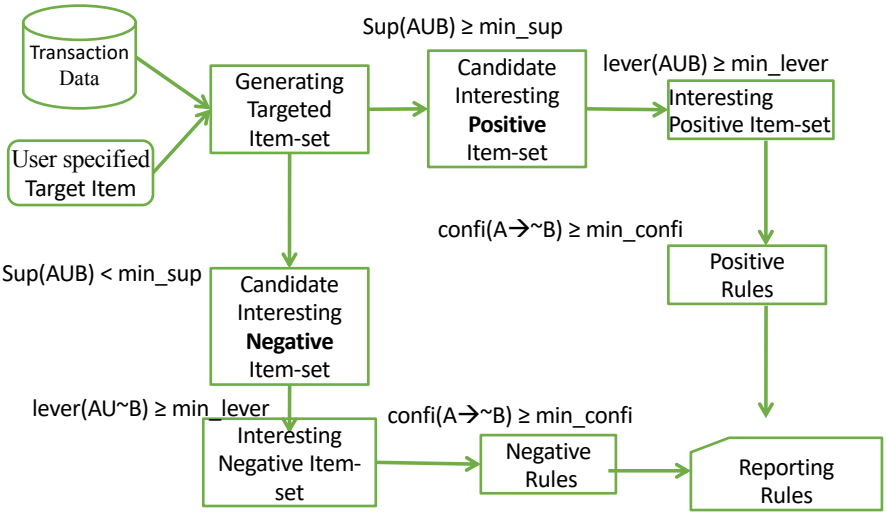
Male, High Income → ~Travel to Mainland China

Measuring Interesting Negative Rules

- The mining negative Rules are mined based on three User Specified Parameters
  - ***support***
    - how frequently an association rule appear in transactions
  - ***confidence***
    - how frequently the left hand side of a rule implies the right hand side
  - ***interestingness (Leverage)***
    - how interesting an item-set is

$$Leverage(A,B) = | Supp(AUB) - Supp(A)Supp(B) |$$
$$Leverage(A,\sim B) = | Supp(AU \sim B) - Supp(A)Supp(\sim B) |$$

Mining Pos/Negative Rules



Rules from Tourism Data.

- We apply this algorithm to Hong Kong outbound tourism data set in 2007 (2008 and 2009 data set is used for validation of results)
  - Parameters: min_supp = 0.1, min_inter = 0.01, min_confi = 0.5

Association Rules			Confidence	Validation	Rule (ID)
webExp0, education1	→	¬ML,OS	0.5899	0.6789	R_1
webExp0, education1	→	¬OS	0.5983	0.5151	R_2
webExp0, education1, guandong0	→	¬ML,OS	0.8886	0.9569	R_3
webExp0, education1, guandong0	→	¬OS	0.6193	0.6015	R_4
webExp0, travelExp1	→	¬ML,OS	0.8794	0.9250	R_5
webExp0, travelExp1	→	¬OS	0.6983	0.7689	R_6
guangdong0, travelExp1	→	¬ML,OS	0.9272	0.9105	R_7
guangdong0, travelExp1	→	¬OS	0.6993	0.7437	R_8
travelExp1	→	¬ML,OS	0.8328	0.7938	R_{09}
travelExp1					
webExp0, mac					
webExp0, mac					94
income1, trave					43
income2, travelExp1	→	NoTravel	0.6004	0.7291	

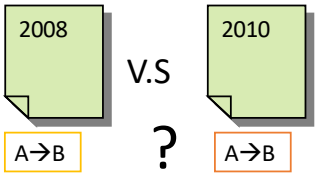
Yeah, Being Tourism Manager, I 'd like to know about these.

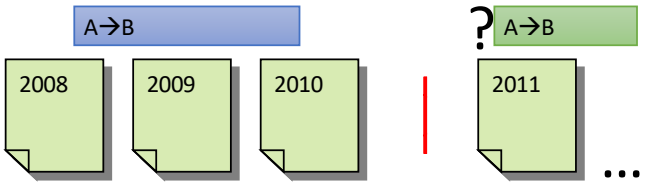
Contrast Targeted Rules Mining

- Contrast Targeted Rules
- The Algorithm

Contrast Targeted Rules Mining

- Contrast Targeted Rules Algorithm (CTR):
  - Detect the changes of rules strength between differences year data set.
  - Tracing the trend for predicting future potential.



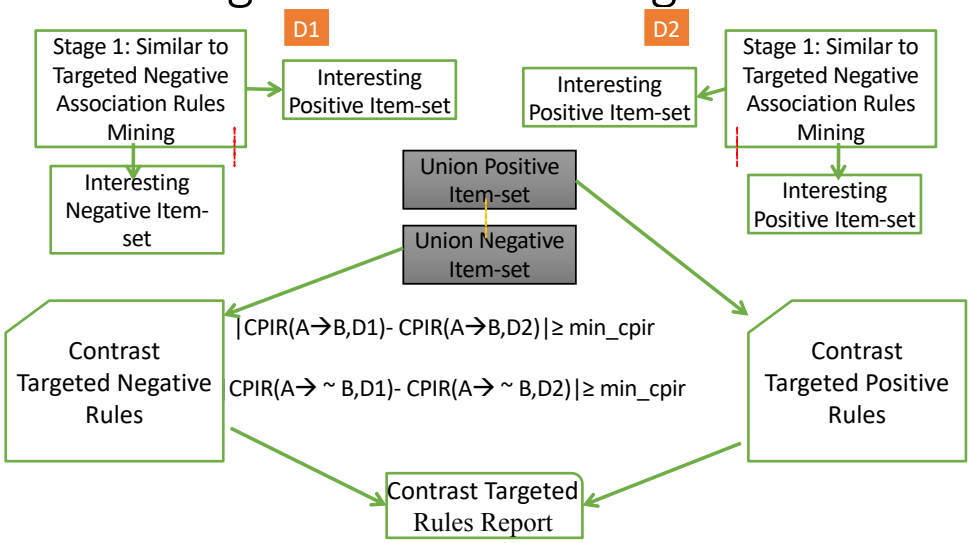


Measuring Interesting Negative Rules

- The mining negative Rules are mined based on three User Specified Parameters
  - ***support***
    - how frequently an association rule appear in transactions
  - ***interestingness (Leverage)***
  - ***Conditional Probability Increment Ratio (CPIR)***
    - Reflect the dependency between item set (positive or negative) – better than *confidence* value when show these relation.

$$CPIR(A \to B, Di) = \frac{p_{Di}(B|A) - p_{Di}(B)}{1 - p_{Di}(B)} \quad CPIR(A \to \neg B, Di) = \frac{p_{Di}(B|A) - p_{Di}(\neg B)}{1 - p_{Di}(\neg B)}$$

Contrast Targeted Rules Mining



Contrast Targeted Rules Mining

- We apply this algorithm to Hong Kong outbound tourism data set in 2005 and 2009
  - Parameters: ms = 0.05, mi = 0.01, mc = 0.05

Association Rules			CPIR 2005	CPIR 2009	Difference	Rate	Rule(ID)
mainland	→	*¬meetpeople*	0.6488	0.2426	**-0.4062**	167%	R_{ddtm1}
mainland, ¬importance	→	*¬meetpeople*	0.8244	0.1219	**-0.7025**	**576%**	R_{ddtm2}
¬mainland	→	*¬discovery*	0.0610	0.3456	0.2845	**466%**	R_{ddtm3}
¬mainland, ¬overseas	→	*¬discovery*	0.0495	0.3660	**0.3165**	**640%**	R_{ddtm4}
mainland	→	*knowledge*	0.5131	0.2028	**-0.3103**	153%	R_{ddtm5}
mainland, ¬importance	→	*knowledge*	0.4537	0.0934	**-0.3603**	**386%**	R_{ddtm6}

Contrast Targeted Rules Mining

Association Rules			CPIR 2005	CPIR 2006	CPIR 2007	CPIR 2008	CPIR 2009	Difference	Rate	Rule(ID)
¬futuretrip	→	*¬macau*	0.4545	0.3832	0.3448	0.2328	0.2098	-0.2447	117%	R_{tdtd1}
discovery, knowledge	→	*¬overseas*	0.5437	0.5165	0.4647	0.4023	0.3299	-0.2137	65%	R_{tdtd2}
discovery, knowledge	→	*mainland*	0.7558	0.7753	0.2554	0.2322	0.1946	**-0.5612**	**74%**	R_{tdtd3}

- (Rule 1) Chance for people with future trip motivation to go to Macau is decreasing
- People with motivation of *discovery* and getting new k*nowledge* are also more likely to take overseas trip (Rule 2) and less likely to visit Mainland China(Rule 3)

Yeah, That is great, I can predict what will happen in 2010 and 2011

Lecture Notes on

Advanced Data Analytics

Module 07: Text Mining

Gang Li
School of Information Technology
Deakin University, VIC 3125, Australia

Text Analysis

- Text Analysis
- Inverted Index

Need for Text Analysis

- Approximately 90% of the World's data is held in unstructured formats
  - Web pages
  - Emails
  - Technical documents
  - Books
  - Digital libraries
  - Customer complaint letters
- Growing rapidly in size and importance
- Various needs for information
  - Search for documents that fall in ***a given topic***
  - Search for ***specific*** information
  - Search an ***answer*** to a question
  - Search for information in ***a specific language***

Text Analysis

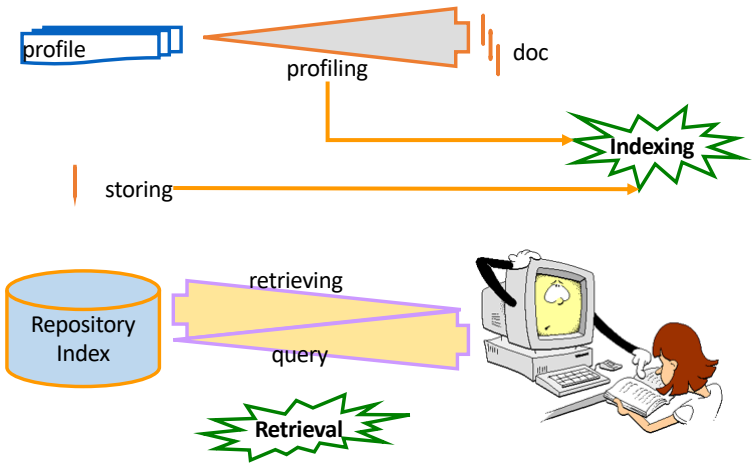
- Categories of Text Analysis
  - Information retrieval
  - Information extraction
  - Named-entity recognition (NER)
  - Question answering (QA)
  - Machine translation (ML)
  - Foreign language reading and writing
  - Speech recognition
  - Text proofing
  - Optical character recognition (OCR)

Definition of IR

- Kowalski (1997)
  - "An ***Information Retrieval system*** is a system that is capable of ***storage***, ***retrieval***, and ***maintenance*** of ***information***. Information in this context can be composed of ***text*** (including ***numeric*** and ***date*** data), ***images***, ***audio***, ***video***, and other ***multimedia*** objects."

IR deals with ***unstructured*** data.

Two Main Stages



Two Main Stages

- **Indexing Process**
  - Involves ***pre-processing*** and ***storing*** of information in a repository
  - It usually involves ***5*** pre-processing stages
    - ***Lexical Analysis***
      - Determines the words of the document (Tokenization, etc.)
    - ***Stop-word Elimination***
      - Filter Out words that occur in most of the documents (like, “the”, “of”)
    - ***Stemming***
      - Replace all variants of a word with a single stem of the word
    - ***Index-Term Selection***
      - Select a set of terms for indexing documents
    - ***Thesauri***
      - Standardize the index terms that were selected

Two Main Stages

- **Retrieval Process**
  - At retrieval time, the ***query*** is ***indexed***. Then index files (postings) are fetched and analysed so as to ***return most relevant documents***
  - Basic object is a ***document***

Inverted Index

- What is Inverted Index?
- Constructing an Inverted Index

Need for Inverted Index

- An obvious option in searching for a basic query is to scan the text sequentially
  - ***sequential search***, or ***online search***
  - It is appropriate when
    - the text is ***small***
      - e.g., a few megabytes
    - the text collection is very ***volatile***
      - e.g., undergoes changes frequently
    - index space overhead cannot be afforded
- However, most real world text collection is:
  - ***Large***,
  - ***Semi-Static***

Need for Inverted Index

- Most successful techniques used ***Inverted Index*** (a.k.a ***Inverted File***) to build data structures over the text to speed up the search

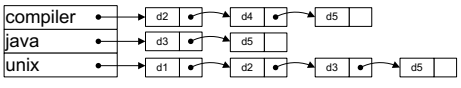
docs	t1	t2	t3
D1	1	0	1
D2	1	0	0
D3	0	1	1
D4	1	0	0
D5	1	1	1
D6	1	1	0
D7	0	1	0
D8	0	1	0
D9	0	0	1
D10	0	1	1

Invert documents into a big index (organized by terms vs. documents)

Terms	D1	D2	D3	D4	D5	D6	D7	…
t1	1	1	0	1	1	1	0	
t2	0	0	1	0	1	1	1	
t3	1	0	1	0	1	0	0	

Inverted Index

- ***Inverted Index*** consists of ***an ordered list of indexing terms***, each indexing term is associated with some document identification numbers
  - **Vocabulary**:
    - the set of all distinct words in the text
  - **Occurrences**:
    - lists containing information for each word of the vocabulary. At minimum this will be a list of documents that the word appears in.



Constructing Inverted Index

- **Step 1:**
  - Documents are parsed to extract tokens.
  - These are saved with the Document ID.

Doc 1

Now is the time for all good men to come to the aid of their country

Doc 2

It was a dark and stormy night in the country manor. The time was past midnight

Term	Doc #
now	1
is	1
the	1
time	1
for	1
all	1
good	1
men	1
to	1
come	1
to	1
the	1
aid	1
of	1
their	1
country	1
it	2
was	2
a	2
dark	2
and	2
stormy	2
night	2
in	2
the	2
country	2
manor	2
the	2
time	2
was	2
past	2
midnight	2

Constructing Inverted Index

- **Step 2:**
  - After all documents have been parsed the inverted file is sorted alphabetically.

Term	Doc #
now	1
is	1
the	1
time	1
for	1
all	1
good	1
men	1
to	1
come	1
to	1
the	1
aid	1
of	1
their	1
country	1
it	2
was	2
a	2
dark	2
and	2
stormy	2
night	2
in	2
the	2
country	2
manor	2
the	2
time	2
was	2
past	2
midnight	2

Term	Doc #
a	2
aid	1
all	1
and	2
come	1
country	1
country	2
dark	2
for	1
good	1
in	2
is	1
it	2
manor	2
men	1
midnight	2
night	2
now	1
of	1
past	2
stormy	2
the	1
the	1
the	2
the	2
their	1
time	1
time	2
to	1
to	1
was	2
was	2

Constructing Inverted Index

- **Step 3:**
  - Multiple term entries for a single document are merged.
  - Within-document term frequency information is compiled.

Term	Doc #
a	2
aid	1
all	1
and	2
come	1
country	1
country	2
dark	2
for	1
good	1
in	2
is	1
it	2
manor	2
men	1
midnight	2
night	2
now	1
of	1
past	2
stormy	2
the	1
the	1
the	2
the	2
their	1
time	1
time	2
to	1
to	1
was	2
was	2

→

Term	Doc #	Freq
a	2	1
aid	1	1
all	1	1
and	2	1
come	1	1
country	1	1
country	2	1
dark	2	1
for	1	1
good	1	1
in	2	1
is	1	1
it	2	1
manor	2	1
men	1	1
midnight	2	1
night	2	1
now	1	1
of	1	1
past	2	1
stormy	2	1
the	1	2
the	2	2
their	1	1
time	1	1
time	2	1
to	1	2
was	2	2

Constructing Inverted Index

- **Step 4:**
  - Then the file can be split into
    - A *Dictionary file*
    - A *Postings file*

Term	Doc #	Freq
a	2	1
aid	1	1
all	1	1
and	2	1
come	1	1
country	1	1
country	2	1
dark	2	1
for	1	1
good	1	1
in	2	1
is	1	1
it	2	1
manor	2	1
men	1	1
midnight	2	1
night	2	1
now	1	1
of	1	1
past	2	1
stormy	2	1
the	1	2
the	2	2
their	1	1
time	1	1
time	2	1
to	1	2
was	2	2

→

Constructing Inverted Index

Term	Doc #	Freq
a	2	1
aid	1	1
all	1	1
and	2	1
come	1	1
country	1	1
country	2	1
dark	2	1
for	1	1
good	1	1
in	2	1
is	1	1
it	2	1
manor	2	1
men	1	1
midnight	2	1
night	2	1
now	1	1
of	1	1
past	2	1
stormy	2	1
the	1	2
the	2	2
their	1	1
time	1	1
time	2	1
to	1	2
was	2	2

Dictionary File

Term	N docs	Tot Freq
a	1	1
aid	1	1
all	1	1
and	1	1
come	1	1
country	2	2
dark	1	1
for	1	1
good	1	1
in	1	1
is	1	1
it	1	1
manor	1	1
men	1	1
midnight	1	1
night	1	1
now	1	1
of	1	1
past	1	1
stormy	1	1
the	2	4
their	1	1
time	2	2
to	1	2
was	1	2

Postings File

Doc #	Freq
2	1
1	1
1	1
2	1
1	1
1	1
2	1
2	1
1	1
1	1
2	1
1	1
2	1
2	1
1	1
2	1
2	1
1	1
1	1
2	1
2	1
1	2
2	2
1	1
1	1
2	1
1	2
2	2

What is in "Dictionary File"?

- Fields in the ***Dictionary File***
  - ***Index Term***
  - ***Number of Entries*** in ***Posting Files***:
    - it appeared in how many documents
  - ***Pointer*** to ***Postings File***
  - Additional Fields Including
    - Total Frequency of Term
    - ***IDF*** Value of Term
- Normally the size of dictionary file is small enough to be fitted into memory

What is in "Postings"?

- Fields in ***Postings Files***
  - The ***sorted*** ***Document ID Number***
    - Boolean Model
  - Document ID Number with ***Term Frequency*** (TF)
    - Vector Model
  - Document ID Number with ***Position*** in the Document
    - Proximity Operators, etc.

- For Boolean Model, posting file is around 10% size of the document
- For Vector Model, posting file is around 20% size of the document

Statistical Properties of Text

- Word Frequency
- Zipf's Law
- Heap's Law

Two Questions?

- How is the ***frequency of different words*** distributed?
  - ***Zipf's Law***
- How fast does ***vocabulary size*** grow with ***the size of a corpus (collection of documents)***?
  - ***Heaps' Law***

Word Frequency

- A few words are ***very common***.
  - 2 most frequent words ("the", "of") can account for about 10% of word occurrences.
  - ***Stopwords***
- Most words are ***very rare***.
  - Half the words in a corpus appear only once, called *hapax legomena* (Greek for "read only once")
- Called a "***heavy/long tailed***" distribution, since most of the probability mass is in the "tail"

Frequent Word	Number of Occurrences	Percentage of Total
the	7,398,934	5.9
of	3,893,790	3.1
to	3,364,653	2.7
and	3,320,687	2.6
in	2,311,785	1.8
is	1,559,147	1.2
for	1,313,561	1.0
The	1,144,860	0.9
that	1,066,503	0.8
said	1,027,713	0.8

Frequencies from 336,310 documents in the 1GB TREC Volume 3 Corpus
125,720,891 total word occurrences; 508,209 unique words

Zipf's Law

- Zipf (1949) "discovered" that
  - The frequency of ***the r-th most frequent*** term is $1/r^{\vartheta}$ times that of ***the most frequent*** term
    - E.g., the most frequent term occurred K times
    - the 2nd most frequent term occurred $K \times 1/2^{\vartheta}$ times
    - the 3rd most frequent term occurred $K \times 1/3^{\vartheta}$ times
  - In the most simple formulation, $\vartheta = 1$

$$f \propto \frac{1}{r^{\theta}}$$

Zipf and Term Weighting

- Luhn (1958) suggested that both ***extremely common*** and ***extremely uncommon words*** were **not very useful** for indexing.

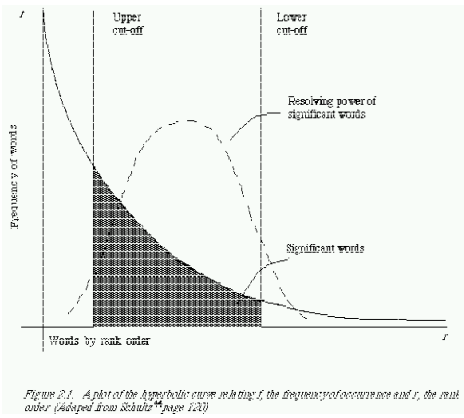


Figure 2.1. A plot of the hyperbolic curve relating f, the frequency of occurrence and r, the rank order (Adapted from Schultz[4] page 120)

Predict Occurrence Frequency

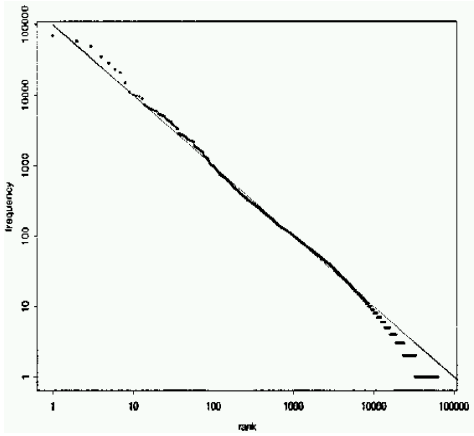
- From ***Zipf's*** Law, it implies that
  - In a document of length ***n*** with a vocabulary of ***V*** terms, the ***i*-th *most frequent term*** appears $n/(i^{\vartheta} H_V(\vartheta))$ times, where $H_V(\vartheta)$ is the harmonic number of order ϑ of ***V***, defined as

$$H_V(\theta) = \sum_{j=1}^{V} \frac{1}{j^{\theta}}$$

Real Data & Zipf's Law

- A law of the form $y = kx^c$ is called a ***power law***.
  - On a ***log-log*** plot, power laws give a straight line with slope *c*.
  - **$\log(y) = \log(kx^c) = \log k + c \log(x)$**
- Zipf is quite accurate except for very high and low rank.

$$f \propto \frac{1}{r^{\theta}} = 1 \times r^{-\theta}$$

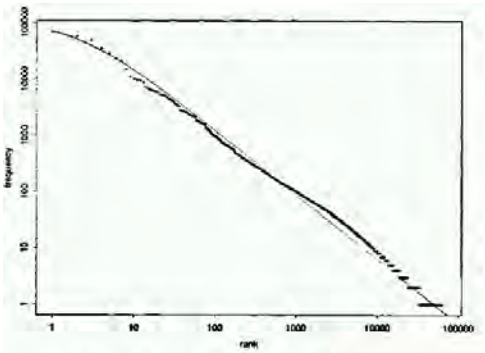


Mandelbrot Correction

- In 1954, ***Mandelbrot*** proposed a correction of ***Zipf's*** Law and gave a better fit

$$f = P(r+\rho)^{-B}$$

for constants P, B, ρ



Mandelbrot's function on Brown corpus

Summary of Zipf's Law

Good News	Bad News
☺Stopwords will account for a large fraction of text so eliminating them greatly reduces inverted-index storage costs ☺Li (1992) shows that just random typing of letters including a space will generate "words" with a Zipfian distribution. http://linkage.rockefeller.edu/wli/zipf/	☹For most words, gathering sufficient data for meaningful statistical analysis (e.g. for correlation analysis for query expansion) is difficult since they are extremely rare.

Vocabulary Growth

- How does ***the size of the overall vocabulary*** (number of unique words) grow with ***the size of the corpus***?
  - This determines how the size of the inverted index will scale with the size of the corpus.
  - Vocabulary not really upper-bounded due to proper names, typos, etc.

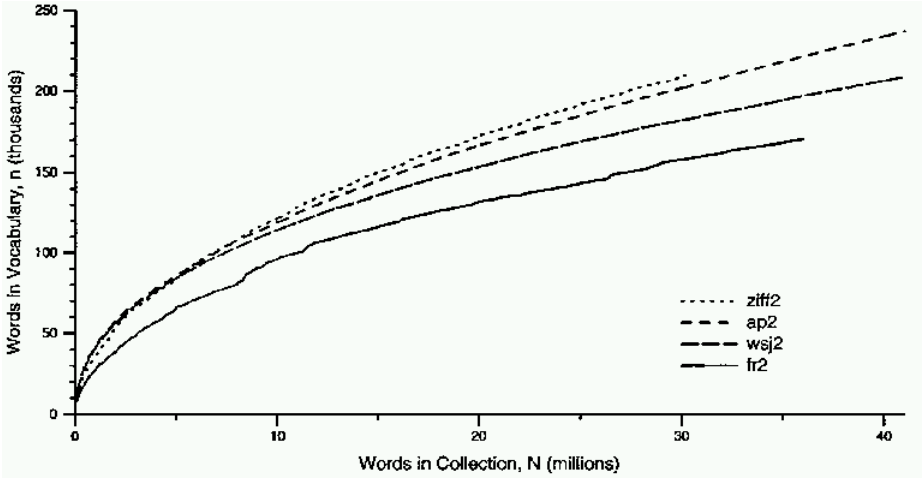
Heaps' Law

- If ***V*** is the size of the vocabulary and the ***n*** is the length of the corpus in words:

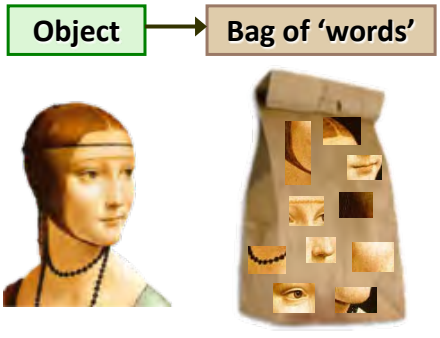
$$V = Kn^{\beta} \quad \text{with constants } K,\ 0 < \beta < 1$$

- Typical constants:
  - $K \approx 10\text{–}100$
  - $\beta \approx 0.4\text{–}0.6$ (approx. square-root)

Heaps' Law Data



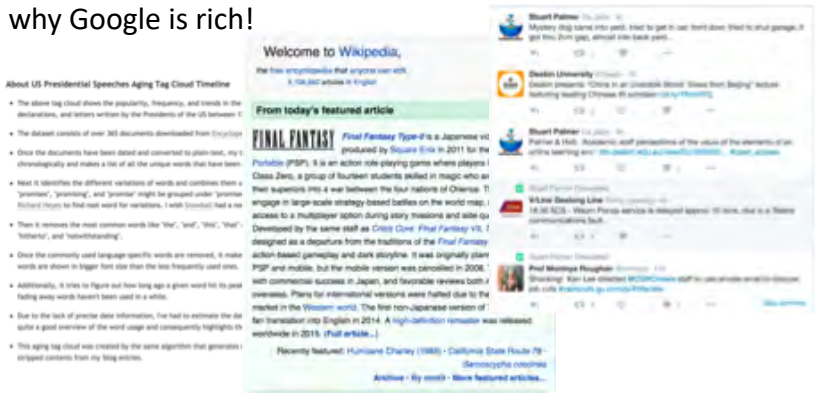
Bag of Words Model



- Unstructured Data Presentation
- Bag of Words model
  - text

Unstructured Data (Text)

- **Text** stores most of human knowledge (80%)
  - that's why Google is rich!



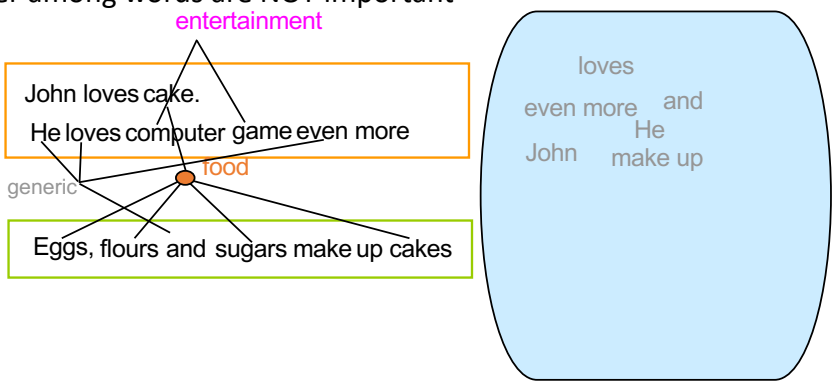
Unstructured Data Representation

- Is there a universal approach for **data representation**?
  - How to deal with heterogeneity from unstructured data:
    - images, texts, videos, signals… !
  - How to approach and represent the data for an analysis task?

- Bag-of-word representation!

Bag-of-Words (Text)

- Collect "important" (unique) words into a bag
  - Order among words are NOT important



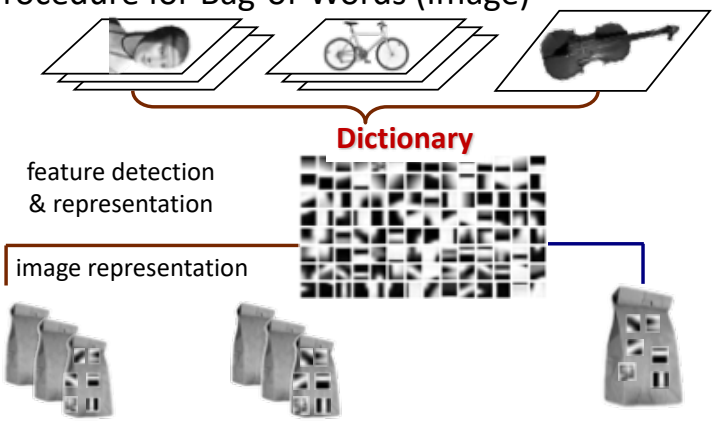
Bag-of-Words (Text)

- Or er-less document representation:
  - tart with a corpus of documents.
  - lse bag-of-word representation based on **Dictionary**.
  - urn the corpus into a matrix of numerical values.
  - ows are terms, columns are documents or vice versa.

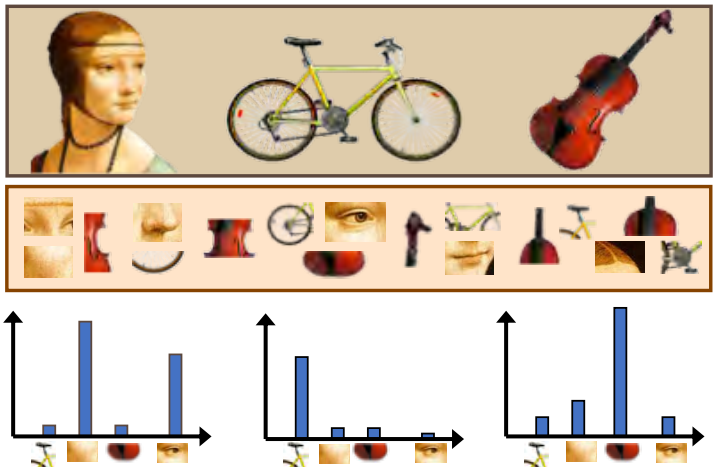
http://chir.ag/phernalia/preztags/

Bag-of-Words (Image)

- General Procedure for Bag-of-Words (Image)



Bag-of-Words (Image)



Bag-of-Words and Vector Model

- Summary
  - **Bag-of-Words** Representation
    - First, take the data repository, extract features, and build up a "dictionary" – a list of common features
    - Given a new record, extract features and build a histogram – for each feature, find the closest "word" in the "dictionary"
  - **Vector Model**
    - Each record is represented as a vector of numerical values.
    - Computation can NOW be computed using the vector representation.
    - It is the fundamental technique in information retrieval system, or other tasks such as clustering

Vector Model

- Need for Vector Model
- Vector Model
  - Basic Concepts
  - Similarity
- Vector Model Example

Review: Inverted Index

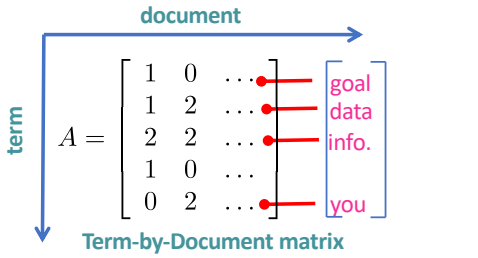
vocabulary	Doc1	Doc2
goal	1	0
data	1	2
information	2	2
insight	1	0
you	0	2

Document 1

The goal is to turn data into information, and information into insight"
Carly Fiorina

Document 2

"You can have data without information, but you cannot have information without data."
Daniel Keys Moran



Vector Model

- Each document is now represented as a vector.
- We can use distance between vectors to compute their similarity.
- We can then make quantitative statement about which document is more similar to each other!
- This is the core of information retrieval!

Document 1

The goal is to turn data into information, and information into insight"
Carly Fiorina

Document 2

"You can have data without information, but you cannot have information without data."
Daniel Keys Moran

$x_1 = [1\ 1\ 2\ 1\ 0]^T$ $\quad x_2 = [0\ 2\ 2\ 0\ 2]^T$

$$\cos\theta = \frac{\sum_i w_{iq} w_{id}}{\sqrt{\sum_i w_{iq}^2 \sum_i w_{id}^2}}$$

Euclidean distance:

$$\sqrt{(1-0)^2 + (1-2)^2 + (2-2)^2 + (1-0)^2 + (0-2)^2}$$
$$= \sqrt{0+1+0+1+4} = \sqrt{6} \approx 2.45$$

Need for Vector Model

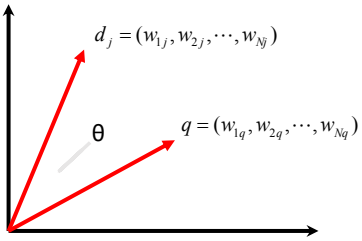
- Users need
  - ***partial*** matching of documents
  - return top ***K*** related documents
- Problems with Boolean Models
  - Use of Binary weights is too limiting
    - just "***Present***" or "***Absent***", no information on how frequent

Vector Model

- Documents are represented as ***vectors*** in a ***N***-dimensional space
  - ***N*** is the number of terms in the corpus
- ***Query is treated as a document***
  - also a vector
- Relevance is measured by ***similarity***
  - A document is relevant to the query if its vector is similar to the query's vector

Vector Model

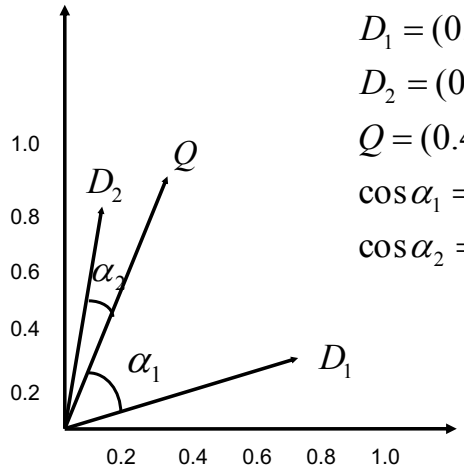
- Relevance is measured by similarity
  - One possible measure of similarity is the ***cosine*** of the angle between a document vector and the query vector.
  - A document is retrieved even if it matches the query terms only partially



$$sim(q,d) = \cos\theta = \frac{\sum_i w_{iq} w_{id}}{\sqrt{\sum_i w_{iq}^2 \sum_i w_{id}^2}}$$

Vector Model

- Example



$$D_1 = (0.8, 0.3)$$
$$D_2 = (0.2, 0.7)$$
$$Q = (0.4, 0.8)$$
$$\cos\alpha_1 = 0.7328$$
$$\cos\alpha_2 = 0.9829$$

Vector Model

- How to determine the vector?
  - **TF-IDF**
    - TF (***Term Frequency***):
      - reflects the ***intra-document*** contents (***similarity***)
      - the frequency with which term k_i appears in document d_j
    - IDF (***Inverse-Document Frequency***)
      - reflect the ***inter-document*** separation (***dissimilarity***)
      - it is defined on the term k_i
  - A good term-weighting schemes is given by
    - w_{ij} = *TF(i,j) * IDF(i)*

Vector Model

- How to determine the weights?
  - Let
    - ***N*** be the **total number** of documents in the collection
    - n_i be the number of documents which **contain the term** k_i
    - *freq(i,j)*: raw frequency of term k_i within document d_j
  - the TF is computed as
    - *TF(i,j) = freq(i,j)*
  - the IDF is computed as
    - *IDF(i)* = $log_2(N/ n_i)$
    - = $-log_2 (ni/N)$

- the ***log*** is used to make the values of TF and IDF comparable
- It can also interpreted as the amount of information associated with term k_i

Vector Model: Example

	nova	galaxy	heat	H'wood	film	role	diet
Query	1	1	1				
D1	5	2			20	20	20
D2				2	1	5	
D3			3	4	1		

- IDF(nova) = log (3/1) = 1.58
- IDF(galaxy) = log(3/1) = 1.58
- IDF(heat) = log(3/1) = 1.58
- IDF(H'Wood) = log(3/2) = 0.58
- IDF(role) = log (3/2) = 0.58
- IDF(film) = log(3/3) = 0
- IDF(diet) = log(3/1) = 1.58

$IDF(i) = log2(N/ n_i)$

Vector Model: Example

	nova	galaxy	heat	H'wood	film	role	diet
Query	1	1	1				
D1	5	2			20	20	20
D2				2	1	5	
D3			3	4	1		

- Terms: <nova, galaxy, heat, h'wood, film, role, diet>
- IDF weight vector: <1.58, 1.58, 1.58, 0.58, 0, 0.58, 1.58>
- TF weights for Query: <1,1,1,0,0,0,0>
- TF-IDF weights for documents:
  - D1: <7.9, 3.16, 0, 0, 0, 11.60, 31.60>
  - D2: <0,0,0,1.16, 0, 2.9, 0>
  - D3: <0,0,4.74, 2.32, 0, 0, 0>

$w_{ij} = TF(i,j) * IDF(i)$

Vector Model: Example

- TF weights for Query: <1,1,1,0,0,0,0>
- TF-IDF weights for documents:
  - D1: <7.9, 3.16, 0, 0, 0, 11.60, 31.60>
  - D2: <0,0,0,1.16, 0, 2.9, 0>
  - D3: <0,0,4.74, 2.32, 0, 0, 0>

$$\cos\theta = \frac{\sum_i w_{iq} w_{id}}{\sqrt{\sum_i w_{iq}^2 \sum_i w_{id}^2}}$$

$$sim(Q, D1) = \frac{(1\times 7.9) + (1\times 3.16) + (1\times 0)}{\sqrt{(1^2+1^2+1^2)} \times \sqrt{(7.9^2+3.16^2+11.6^2+31.6^2)}} = 0.1839$$

$$sim(Q, D2) = \frac{0}{\sqrt{(1^2+1^2+1^2)} \times \sqrt{(1.16^2+2.9^2)}} = 0$$

$$sim(Q, D3) = \frac{1\times 4.74}{\sqrt{(1^2+1^2+1^2)} \times \sqrt{(4.74^2+2.32^2)}} = 0.5186$$

Vector Model

- **TF-IDF** is mainly used for ***the documents vectors***
- **TF** is usually used for ***the query vector***
  - Sometimes TF-IDF weights are also used for the query vector, a better suggestion is:

$$w_{iq} = \frac{0.5 + 0.5 \times freq(i,q)}{\max freq(i,q)} \times \log(N / n_i)$$

Vector Model

- In addition to cosine measure of similarity, other measures have also been proposed
- The ***DICE*** coefficient is defined as follows:

$$sim(q,d) = \frac{2 \times \sum_i w_{iq} w_{id}}{\sqrt{\sum_i w_{iq}^2} + \sqrt{\sum_i w_{id}^2}}$$

Summary of Vector Model

Strength	**Weakness**
☺Cosine ranking sorts documents according to degree of similarity ☺term-weighting improves quality of the answer set ☺partial matching allowed ☺TF-IDF is a good strategy ☺Simple, fast and elegant	☹Assumes independence of index terms, although not clear that this is bad or not

Text Mining

- Documents Classify
- Documents Cluster
- Find patterns or trends across documents

Text Mining

- At this point the Text mining process merges with the traditional Data Mining process.
- Classic Data Mining techniques are used on the structured database that resulted from the previous stages.
- This is a purely application-dependent stage.

Text Mining

- Determining the topic of an article or a book
  - Deciding if an email is spam or not
  - Determining who wrote a text
  - Determining the meaning of a word in a particular context
- Classification
  - Types
    - Open-class classification - set of labels is not defined in advance
    - Multi-class classification - each instance may be assigned multiple labels
    - Sequence classification - a list of inputs are jointly classified.
  - What do we have to do?
    - choose a particular class label for a given input
    - identify particular features of language data that are salient for classifying it
    - construct models of language that can be used to perform language processing tasks automatically
    - learn about text/language from these models

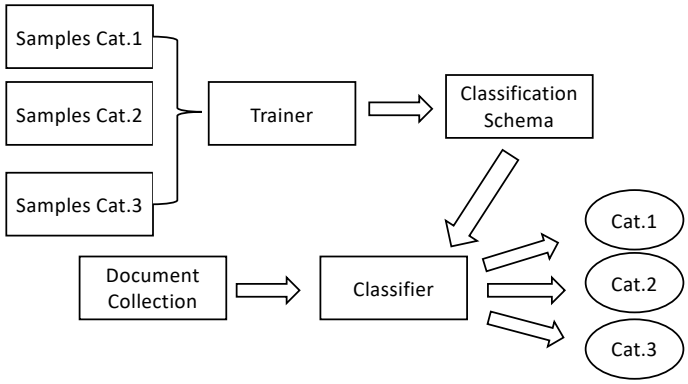
Text Mining

- Sentiment classification
  - Use corpora where documents have been labeled with categories
    - build classifiers that will automatically tag new documents with appropriate category labels
    - Use the Movie Review Corpus, which categorizes reviews as positive or negative to construct a list of documents
  - Human experts classify a set of documents
    - training data set
  - Induce a classification model

	Terms					Class
	Oil	Iraq	build	France	…	Interesting/Not interesting
Document1	0.01	0.05	0.03	0		Interesting
Document2	0	0.05	0	0.01		Not interesting
…	…	…	…	…		

Document Classification

- Classification Schema



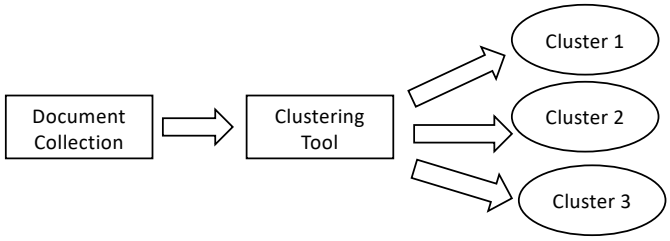
Document Clustering

- Finding Groups of Similar Documents
  - Partitioning Methods: k-means
  - Hierarchical Methods: Agglomerative or Divisive

	Terms					Class
	Oil	Iraq	build	France	…	?
Document1	0.01	0.05	0.03	0		?
Document2	0	0.05	0	0.01		?
…	…	…	…	…		

Document Clustering

- Clustering Schema



Questions?