Improve Data Mining and Knowledge Discovery through the use of MatLab

Gholam Ali Shaykhian
Information Technology (IT) Directorate
Technical Integration Office (IT-G)
ali.shaykhian@nasa.gov

Dawn (Elliott) Martin
Safety and Mission Assurance Directorate
Technical Management Branch (SA-G1)
dawn.elliott@nasa.gov

Robert Beil
NASA Engineering & Safety Centers (NESC)
Systems Engineering Office(C1-20)
robert.j.beil@nasa.gov

National Aeronautics and Space Administration (NASA)
Kennedy Space Center, FL 32899, United States

Abstract

Data mining is widely used to mine business, engineering, and scientific data. Data mining uses pattern based queries, searches, or other analyses of one or more electronic databases/datasets in order to discover or locate a predictive pattern or anomaly indicative of system failure, criminal or terrorist activity, etc. There are various algorithms, techniques and methods used to mine data; including neural networks, genetic algorithms, decision trees, nearest neighbor method, rule induction association analysis, slice and dice, segmentation, and clustering. These algorithms, techniques and methods used to detect patterns in a dataset, have been used in the development of numerous open source and commercially available products and technology for data mining.

Data mining is best realized when latent information in a large quantity of data stored is discovered. No one technique solves all data mining problems; challenges are to select algorithms or methods appropriate to strengthen data/text mining and trending within given datasets. In recent years, throughout industry, academia and government agencies, thousands of data systems have been designed and tailored to serve specific engineering and business needs. Many of these systems use databases with relational algebra and structured query language to categorize and retrieve data. In these systems, data analyses are limited and require prior explicit knowledge of metadata and database relations; lacking exploratory data mining and discoveries of latent information.

This presentation introduces MatLab® (MATrix LABoratory), an engineering and scientific data analyses tool to perform data mining. MatLab was originally intended to perform purely numerical calculations (a glorified calculator). Now, in addition to having hundreds of mathematical functions, it is a programming language with hundreds built in standard functions and numerous available toolboxes. MatLab's ease of data processing, visualization and its enormous availability of built in functionalities and toolboxes make it suitable to perform numerical computations and simulations as well as a data mining tool. Engineers and scientists can take advantage of the readily available functions/toolboxes to gain wider insight in their perspective data mining experiments.

Improve Data Mining and Knowledge Discovery through the use of MatLab

Sixth International Conference on Dynamic Systems and Applications

May, 25-28, 2011
www.dynamicpublishers.com/icdsa6.htm
Morehouse College, Atlanta, GA, 30314, USA.

Gholam Ali Shaykhian
*Information Technology (IT) DirectorateTechnical Integration Office (IT-G)*ali.shaykhian@nasa.gov

Dawn (Elliott) Martin
Safety and Mission Assurance Directorate
Technical Management Branch (SA-G1)
dawn.elliott@nasa.gov

Robert Beil
NASA Engineering & Safety Centers (NESC)
Systems Engineering Office(C1-20)
robert.j.beil@nasa.gov

National Aeronautics and Space Administration (NASA)Kennedy Space Center, FL 32899, United States

Improve Data Mining and Knowledge Discovery through the use of MatLab

Agenda:

- Abstract
- Data Mining Applications (Medicine, Social Media, Technology Development, and Crime, Terrorism & Security)
- Data Storage Media
- Evolution of Data Mining
- Data Mining Explained
- Analytical techniques/methods and approaches used in data mining
- Latent Dirichlet Allocation (LDA)
- LDA - toolbox
- Closing Remarks
- Suggested Readings

Abstract

Data mining is widely used to mine business, engineering, and scientific data. Data mining uses pattern based queries, searches, or other analyses of one or more electronic databases/datasets in order to discover or locate a predictive pattern or anomaly indicative of system failure, criminal or terrorist activity, etc. There are various algorithms, techniques and methods used to mine data; including neural networks, genetic algorithms, decision trees, nearest neighbor method, rule induction association analysis, slice and dice, segmentation, and clustering. These algorithms, techniques and methods used to detect patterns in a dataset, have been used in the development of numerous open source and commercially available products and technology for data mining.

Data mining is best realized when latent information in a large quantity of data stored is discovered. No one technique solves all data mining problems; challenges are to select algorithms or methods appropriate to strengthen data/text mining and trending within given datasets. In recent years, throughout industry, academia and government agencies, thousands of data systems have been designed and tailored to serve specific engineering and business needs. Many of these systems use databases with relational algebra and structured query language to categorize and retrieve data. In these systems, data analyses are limited and require prior explicit knowledge of metadata and database relations; lacking exploratory data mining and discoveries of latent information.

This presentation introduces MatLab® (MATrix LABoratory), an engineering and scientific data analyses tool to perform data mining. MatLab was originally intended to perform purely numerical calculations (a glorified calculator). Now, in addition to having hundreds of mathematical functions, it is a programming language with hundreds built in standard functions and numerous available toolboxes. MatLab's ease of data processing, visualization and its enormous availability of built in functionalities and toolboxes make it suitable to perform numerical computations and simulations as well as a data mining tool. Engineers and scientists can take advantage of the readily available functions/toolboxes to gain wider insight in their perspective data mining experiments.

Data Mining Applications
Medicine, Social Media, Technology

Medicine

January 21, 2011

http://www.bizjournals.com/

Data mining lifts AIDS research

A regional care provider for individuals with HIV and AIDS is working on creating new revenue streams using thousands of electronic records databases. The organizations used grants to build a new electronic records system that combines data on 6,000 individuals from multiple sources, some of which stretch back 20 years.

The result is Evergreen Community Health Outcomes (ECHO), a data mining tool that will be used to develop de-identified data for researchers and the pharmaceutical industry. Accessible data includes everything from primary health care, syringe exchange programs, health promotion services, HIV testing, mental health counseling and case management and nutrition and housing services.

Data Mining Applications
Medicine, Social Media, Technology

Medicine

January 7, 2011

http://www.reuters.com

US top court to decide state drug data mining law

WASHINGTON, Jan 7 (Reuters) - The U.S. Supreme Court said on Friday that it would decide whether a state law restricting commercial access to information about prescription drug records violated constitutional free-speech rights.

The justices agreed to review a data mining law adopted in 2007 in Vermont that prevented the sale, transmission or use of prescriber-identifiable information for marketing a prescription drug unless the prescribing doctor consented.

Three states have such laws, 25 states considered it

Data Mining Applications
Medicine, Social Media, Technology

Medicine

January 26, 2011

http://7thspace.com

Identification of disease-causing genes using microarray data mining and Gene Ontology

The proposed method addresses the weakness of conventional methods by adding a redundancy reduction stage and utilizing Gene Ontology information.
The empirical results show that our method has improved classification performance in terms of accuracy, sensitivity and specificity. In addition, the study of the molecular function of selected genes strengthened the hypothesis that these genes are involved in the process of cancer growth.

The predictions made in this study can serve as a list of candidates for subsequent wet-lab verification and might help in the search for a cure for cancers.

Data Mining Applications
Medicine, Social Media, Technology

Medicine

http://www.infectioncontroltoday.com

Electronic Surveillance Systems: Data Mining Can Yield Rich Results for Infection Prevention

The use of electronic surveillance systems (ESS) in infection control programs is in its infancy of development and implementation so the ramifications of not using ESSs are still being explored.

With the increasing availability and use of electronic medical records, information technology tools have created opportunities for automation of data collection and the potential to decrease the time spent on conducting manual surveillance.

Rather, data mining can detect new and unexpected patterns and may require additional human resources to analyze and develop interventions.

Social Media

January 29, 2011

http://www.infozine.com

Analyzing Data from Facebook, Twitter, Linkedin, and Other Social Media Sites

In recent weeks, a lot of fuss has been made about data mining, in which popular websites like Facebook and Google sell off information about their users to corporations who are looking to gain information about potential consumers.

The general idea of data mining is simply to give corporations an idea of what people on Facebook and other social networking sites are interested in.

Data Mining Applications
Medicine, Social Media, Technology

Technology Development

January 31, 2011

http://www.theintelligencer.net/

West Virginia University to Help Optimize Natural Gas Production

Reports show the Marcellus Shale natural gas rush sweeping across West Virginia could bring billions of dollars and thousands of new jobs to the state over the next several years.

Now, the state's largest academic institution is looking to help "optimize gas production in the region," as West Virginia University's College of Engineering and Mineral Resources is using data mining (data-intensive science) in an effort to save time and resources during gas development.

Data Mining Applications
Medicine, Social Media, Technology

Crime, Terrorism & Security

January 25, 2011

http://www.popdecay.com/

Department Of Justice Launches Net and Cable Data Retention Dragnet

Deputy Assistant Attorney General Jason Weinstein spoke today before the House Subcommittee on Crime, Terrorism and Homeland Security on the matter of increased data mining by the government in a dragnet-styled effort to thwart ALL crime

Data Mining Applications
Medicine, Social Media, Technology

Crime, Terrorism & Security

January 21, 2011

http://www.zdnet.com.au

Data mining digs up dirt on cheats

This week the NZ Herald reported that some NZ$16 million of benefit fraud was uncovered last year, with 10 social welfare staff getting the sack for ripping off the system.

The data-matching techniques include matching client data with other government agencies like the Inland Revenue, Customs and the Department of Internal Affairs, which handles records associated with dead people.

Unit of Measurements:

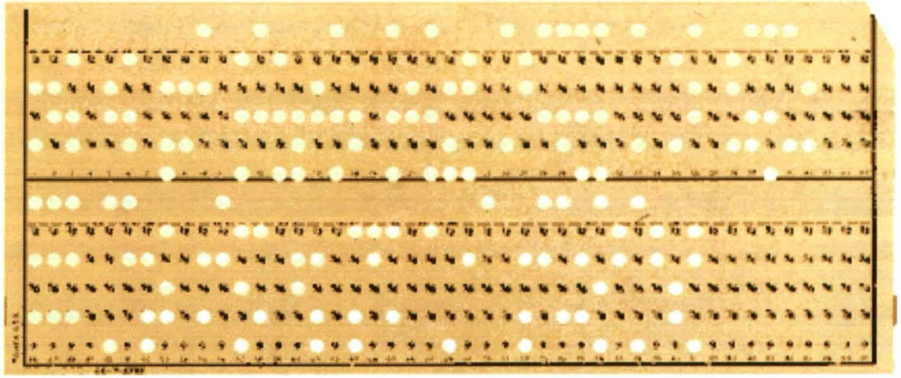
Bits (0,1)	
1 Byte = 8 bits	
1 KB (Kilo Byte)	= 1024 Bytes = 2^10
1 MB (Mega Byte)	= 1024 KB = 2^20
1 GB (Giga Byte)	= 1024 MB = 2^30
1 TB (Tera Byte)	= 1024 GB = 2^40
1 PB (Peta Byte)	= 1024 TB = 2^50
1 EB (Exa Byte)	= 1024 PB = 2^60
1 ZB (Zetta Byte)	= 1024 EB = 2^70
1 YB (Yotta Byte)	= 1024 ZB = 2^80

A side note, the word Google comes from the mathematical term googol, to equal 10^100, a number much larger than the atoms in this universe.

Data Storage Media

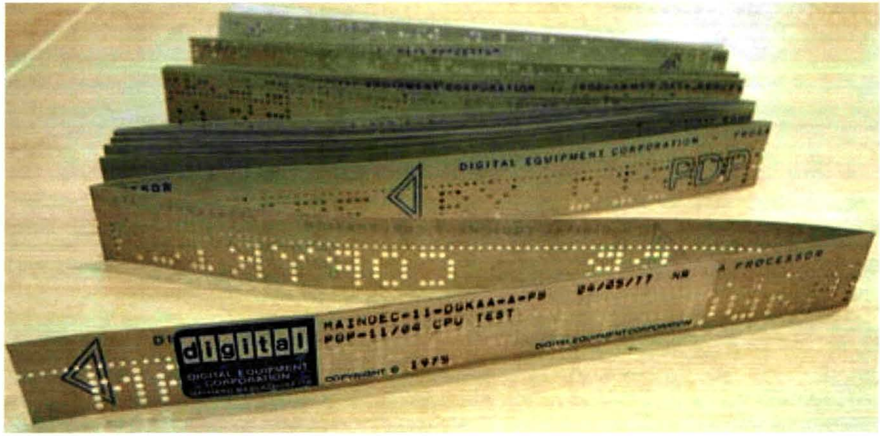
History -Data Storage

Punch Cards



Punch Tape

Each row on the tape represents one Character



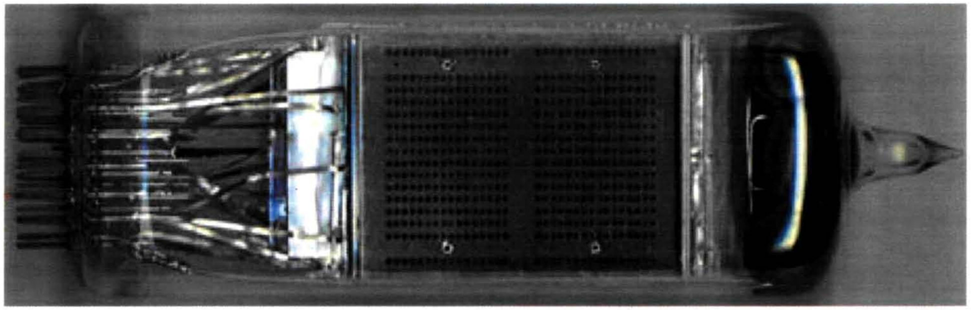
Data Storage Media

History -Data Storage

Selectron Tubes

Largest Selectron Tubes (10 inches)

Could store 4096 bits



Magnetic Tape

1 Magnetic Tape = 10,000 Punch Cards

Data Storage Media

History -Data Storage

Compact Cassette

1 DVD = 4500 Compact Cassette

It takes 281 days to restore the data

Magnetic Drum

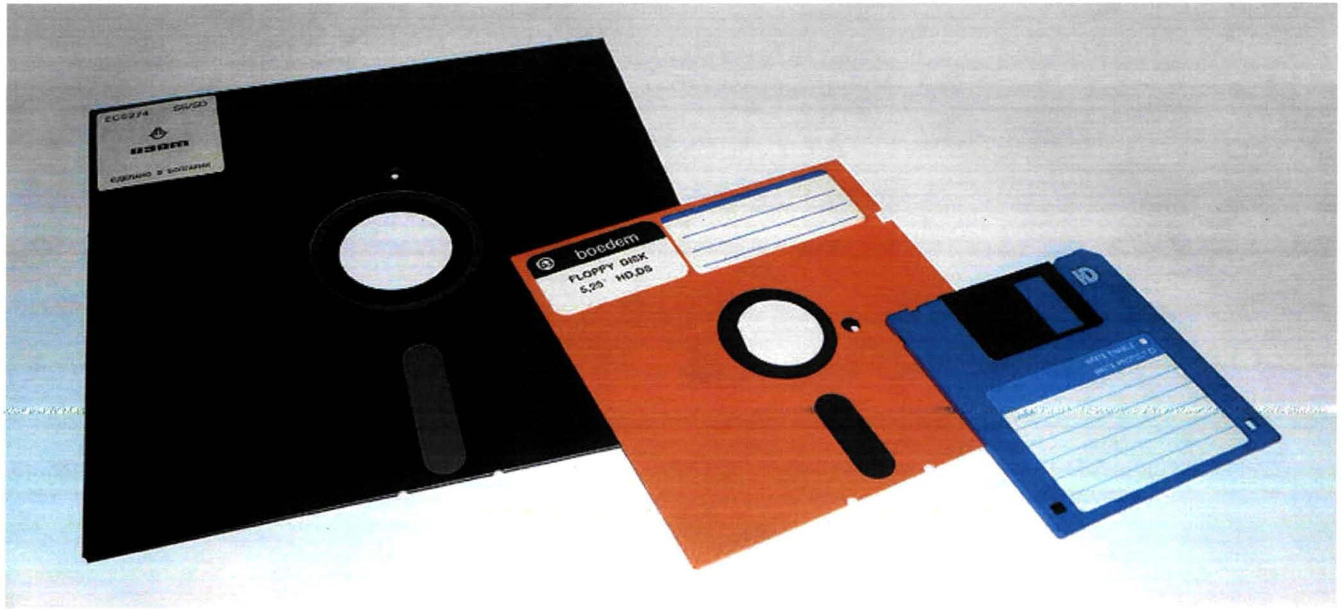
16 inch long (12,500 RPM)

Storage -10,000 Characters

Data Storage Media

History -Data Storage

Floppy Disk



1971 - 8 inch = 80KB

1976 - 5.25 inch = 110/160/180/360 KB, 1.2MB

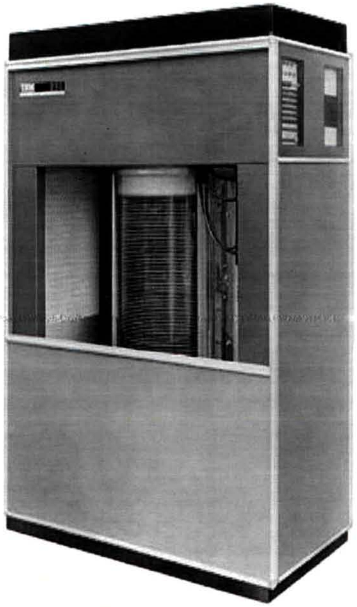
1987 - 3.5 inch = 720KB, 1.4MB

http://en.wikipedia.org/wiki/Floppy_disk

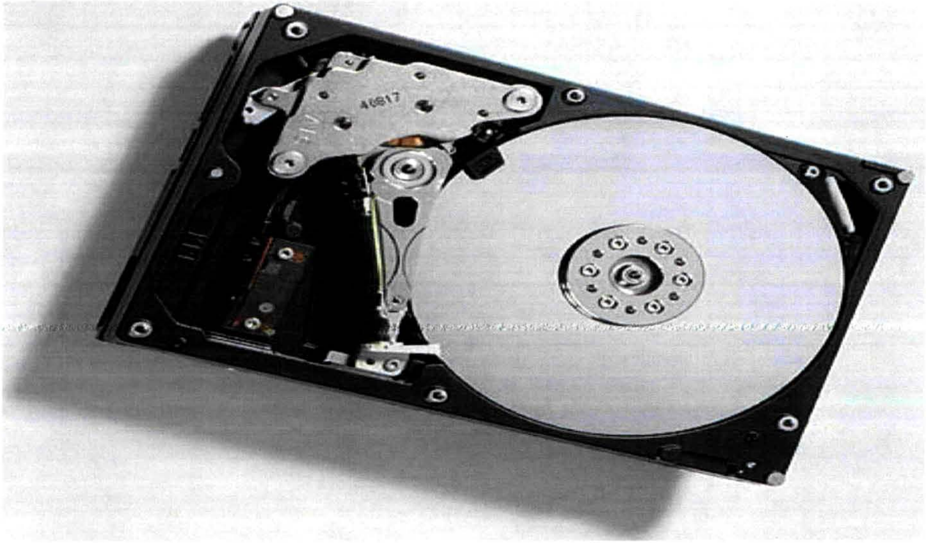
Data Storage Media

History -Data Storage

Hard Drive



1956 - IBM 305 RAMAC
50 24-inch magnetic disks
Leased for $3,200 per month
4.4 MB



Under Constant Development
A 500 GB Hard Drive sells for less than $200 and
120,000 times more storage than the first Hard Drive

Data Storage Media

History -Data Storage

Laser Disk

Laser Disk was invented in 1958
Become available on the Market in 1978

History -Data Storage

Compact Disk

Compact Disk was developed in 1979
A CD can store 700MB of data

History -Data Storage

DVD

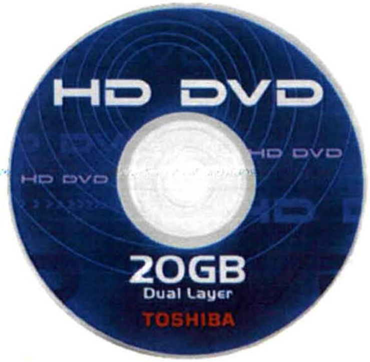
DVD is a CD that uses different kind of laser technology
A dual layer DVD can store 8.5GB of data

Data Storage Media

History -Data Storage

Blu-Ray & HD DVD

Supersede the DVD format



Blu-Ray
Single layer capacity – 25 GB
Dual layer capacity – 50 GB

HD DVD
Single layer capacity – 20 GB
Dual layer capacity – 45 GB

Data Storage Media

History -Data Storage

The Future is here!

Holographic Versatile Disc (HVD)

HVD stores 3.9 Tetrabyte of data
20 Blu-Ray Disk = 1 HDV
Holographic drives are projected to initially cost around US$15,000
A single disc around US$120–180 (although prices are expected to fall steadily)

Data Storage Media

History -Data Storage

The Future is here!

Holographic Versatile Disc (HVD)

HVD stores 3.9 Tetrabyte of data
20 Blu-Ray Disk = 1 HDV
Holographic drives are projected to initially cost around US$15,000
A single disc around US$120–180 (although prices are expected to fall steadily)

What We know?

- Data – Data are stored in one or more tables, matrices or have relations.

- Information (actionable) – The patterns, associations, or relationships among *data* can provide *information*.

- Knowledge – Information can be converted into *knowledge* depicting historical patterns and future trends.

Evolutions of Data Mining

Classical Statistics

- Statistics are the foundation of most technologies on which data mining is built.
- Concepts such as regression analysis, standard distribution, standard deviation, standard variance, discriminant analysis, cluster analysis, and confidence intervals, all of which are used to study data and data relationships.

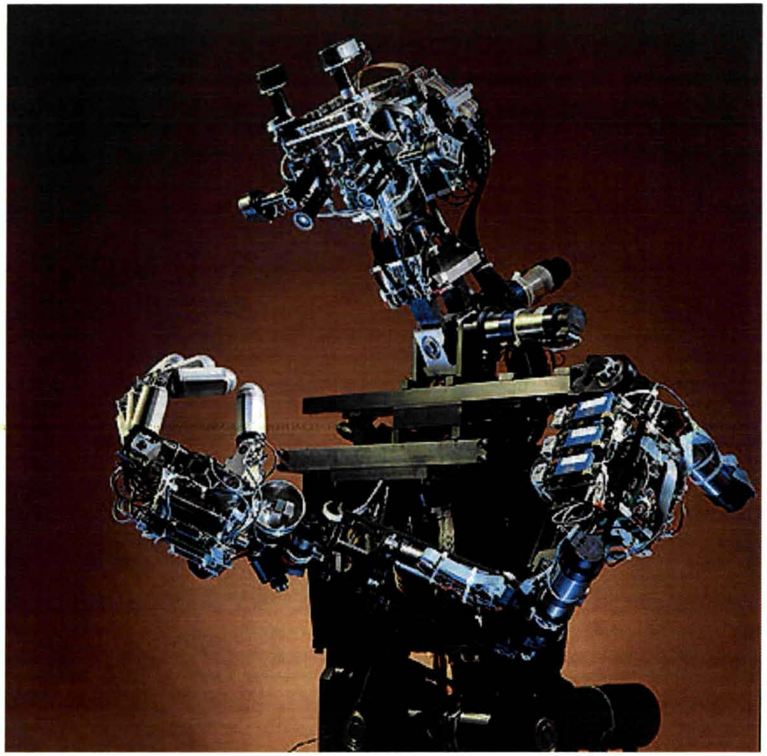
$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\bar{x} - x_i)^2 \qquad \mu = \bar{x} \pm \frac{ts}{\sqrt{N}} \qquad F = \frac{s_1^2}{s_2^2}$$

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{s}\sqrt{\frac{n_1 n_2}{n_1 + n_2}} \qquad s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}$$

Evolutions of Data Mining

Artificial Intelligence

- Artificial Intelligence (AI) is built upon heuristics as opposed to statistics, attempts to apply human-thought-like processing to statistical problems.



Evolutions of Data Mining

Machine Learning

- Machine learning, a union of statistics and AI, it could be considered an evolution of AI, because it blends AI heuristics with advanced statistical analysis.

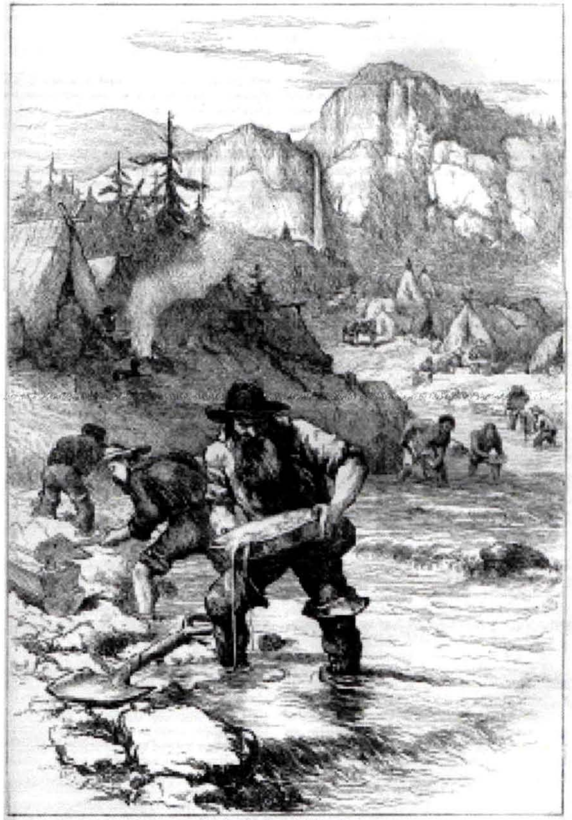
- Machine learning attempts to let computer programs learn about the data they study, such that programs make different decisions based on the qualities of the studied data.

Evolutions of Data Mining

Data Mining (Statistics + AI + Machine Learning)

- Data mining is best described as the union of historical and recent developments in statistics, AI, and machine learning.

- Data mining is finding increasing acceptance in science and business areas which need to analyze large amounts of data to obtain useful knowledge



GOLD-WASHING IN CALIFORNIA.

Data Mining Explained

Definition

"Data mining is the process of discovering meaningful new correlations, patterns, and trends by sifting through large amounts of data stored in repositories and by using pattern recognition technologies as well as statistical and mathematical techniques." (M. J. Berry and G. Linoff)

"Data mining is the process of extracting hidden patterns from large amounts of data." (P. Lyman and H. Varian)

Data Mining Explained

- Data mining is a new discipline; it involves intelligent technical steps to search the data using mining algorithms to output patterns and relationships - Data patterns and relationships are used for interpretation/evaluation in knowledge discovery

- Data mining is the process of analyzing data from different perspectives and summarizing it into useful information (knowledge discovery)

- Data mining is the science of extracting useful information from large data sets or databases to discover patterns and trends that go beyond simple analysis; involves using sophisticated mathematical algorithms to segment the data and evaluate the probability of future events.

Data Mining Explained

- Data mining is the non-trivial discovery of useful patterns, trends, and anomalies in large data sets.

- Data types included in data mining are numeric, text, images, symbolic, and their combinations.

- Data mining tools are software implementations of algorithms generally based on mathematics, statistics, artificial intelligence, machine learning, probability theory, and decision theory.

Data Mining Explained

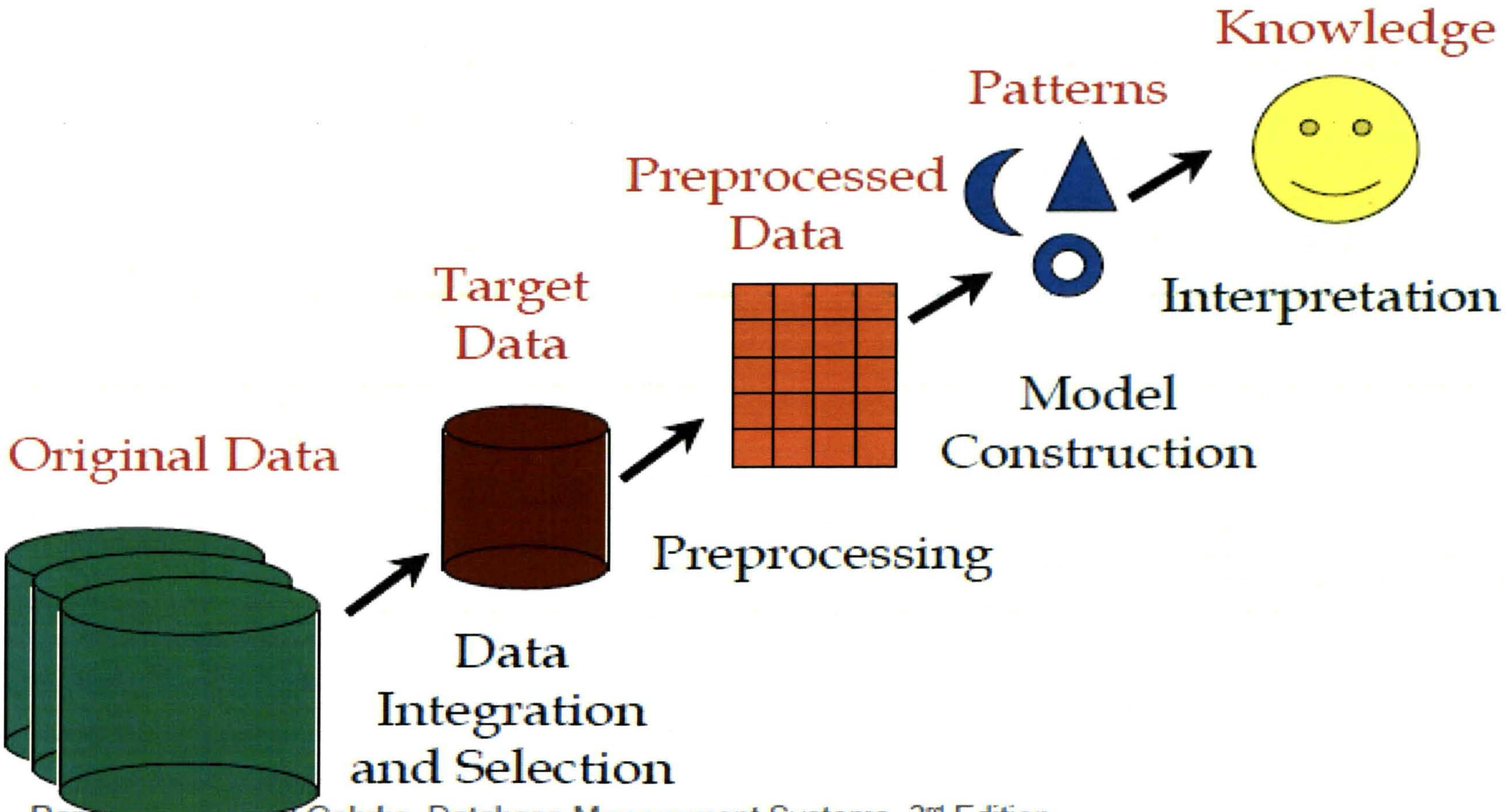
- Data mining is a practice by which we sift through large quantities of data, through exploration and analysis, in order to discover meaningful patterns, associations, or relationships among the data

- Data mining facilitates discovery and prediction – the purpose of data mining is to transform data into actionable information in a wide range of disciplines including science, engineering, marketing, fraud detection, etc. to **discover** explicit characteristics of data and to **predict** future events.

- Generally, data mining activity is an afterthought activity, we collect data for "primary" reason, we then want to find unsuspected relationships among these data; mining analysis concerns finding values of interest hidden from database owners.

Data Mining Explained

“The nontrivial extraction of implicit, previously unknown, and potentially useful information from data.” by Frawley et al. (1992)



Ramakrishnan and Gehrke. Database Management Systems, 3rd Edition.

Data Mining Explained

Common usage of data mining (Business transactions, Scientific/Engineering data, Web, text, images, voice, video) :

- Science and engineering
- Surveillance
- Pattern mining
- Subject-based data mining
- Games
- Business

Analytical techniques/methods and approaches used in data mining

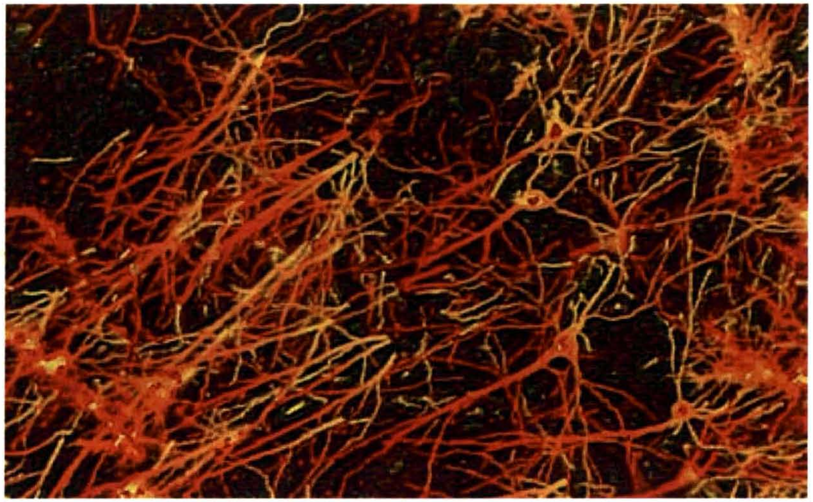
Data mining models use different levels of analytical methodologies including:

- Neural networks
- Genetic algorithms
- Decision trees
- Nearest neighbor method (linear programming)
- Rule induction
- Data visualization
- Association analysis
- Slice and Dice
- Segmentation
- Clustering

Analytical techniques/methods and approaches used in data mining

Neural networks

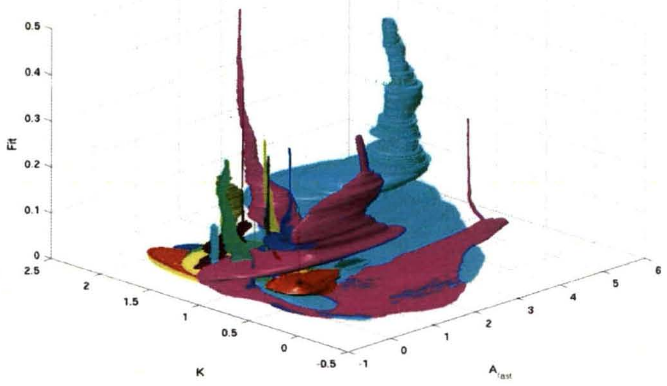
Neural networks are probably the most common data mining technique. Neural networks learn from a training set, generalizing patterns inside it for classification and prediction. Neural networks are also interesting because in their most common incarnation, they detect patterns in data in a matter analogous to human thinking (simulation of a biological brain).



Analytical techniques/methods and approaches used in data mining

Genetic Algorithms

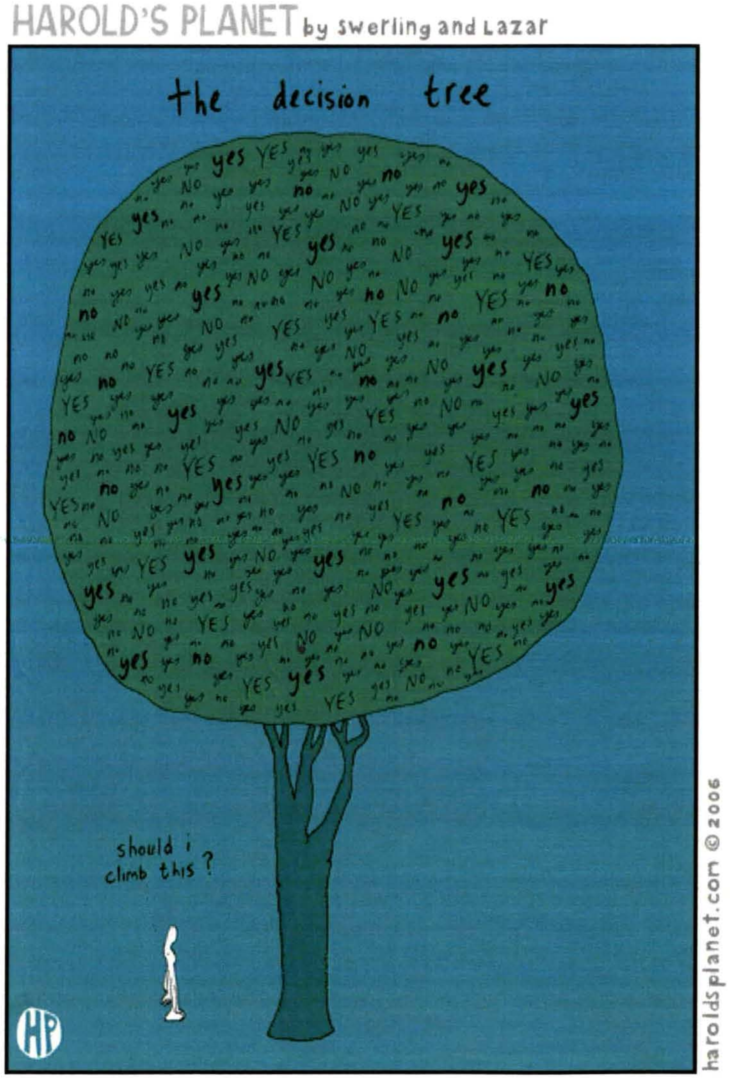
The Genetic Algorithm (GA), inspired by Darwin's theory of evolution and employed to solve optimization problems uses an evolutionary process. It is a search algorithm based on mechanics of natural selection and natural genetics. The GA uses the selection, crossover, and mutation operators to evolve successive generations of solutions. As the generations evolve, only the most predictive survive, until the functions converge on an optimal solution.



Analytical techniques/methods and approaches used in data mining

Decision trees

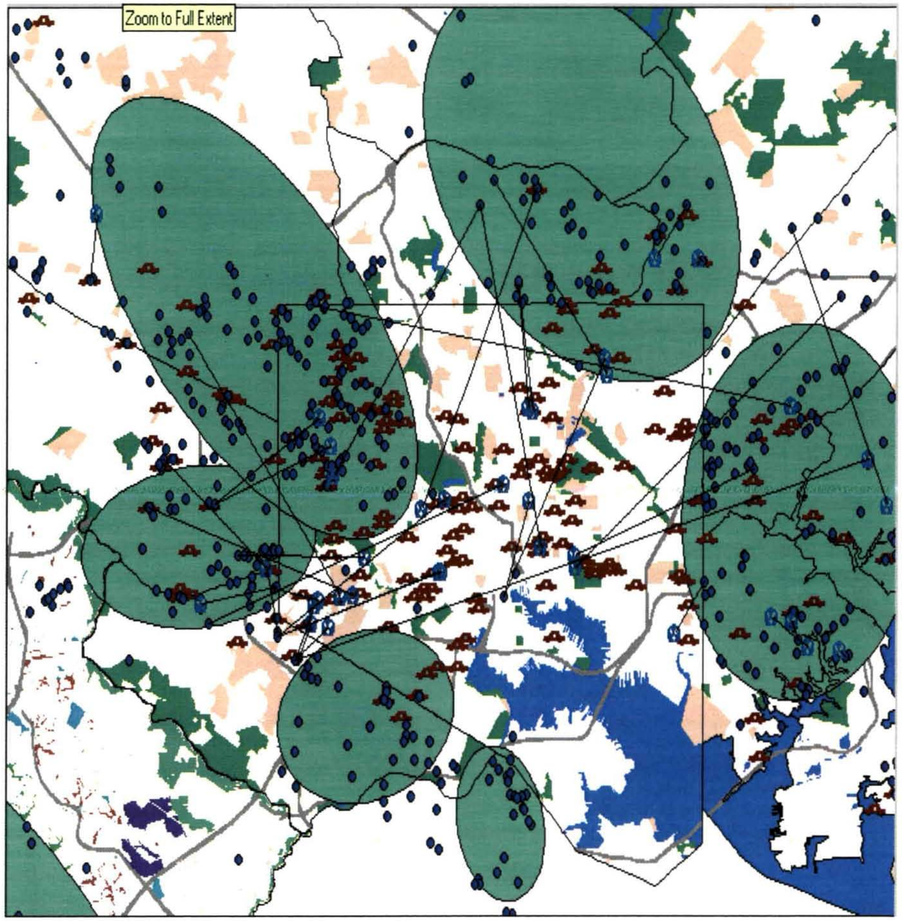
Tree-shaped structures that represent sets of decisions for classification of a dataset. Decision trees provide a set of rules to apply to a new dataset to predict which records will have a given outcome. Decision trees are used for directed data mining; they divide the records into disjoint subsets, each of which is described by a simple rule on one or more fields.



Analytical techniques/methods and approaches used in data mining

Nearest neighbor method

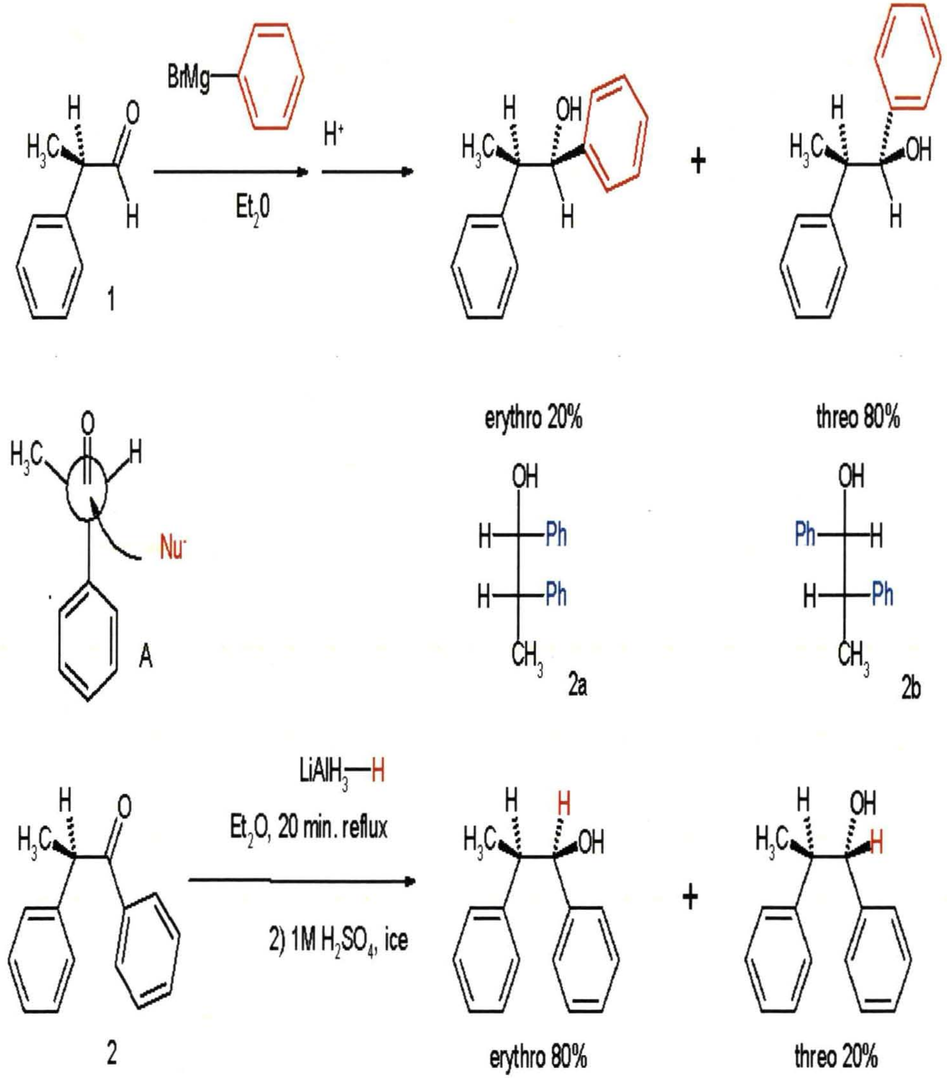
A technique that uses the ratio of expected and observed mean value of the nearest neighbor distances to determine if a data set is clustered; it classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k 1).



Analytical techniques/methods and approaches used in data mining

Rule induction

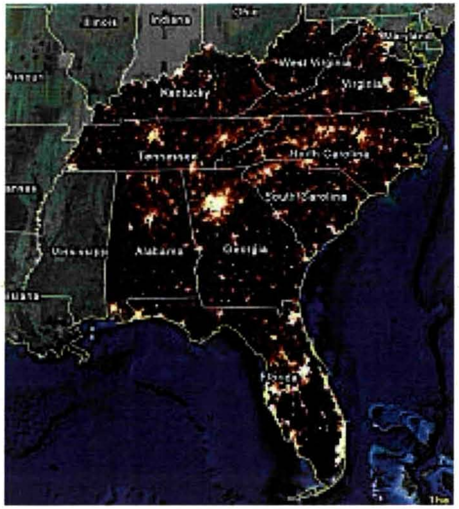
Rule induction is an area of machine learning in which formal rules are extracted from a set of observations. The rules extracted may be useful if based on statistical significance.



Improve Data Mining and Knowledge Discovery through the use of MatLab

Data visualization

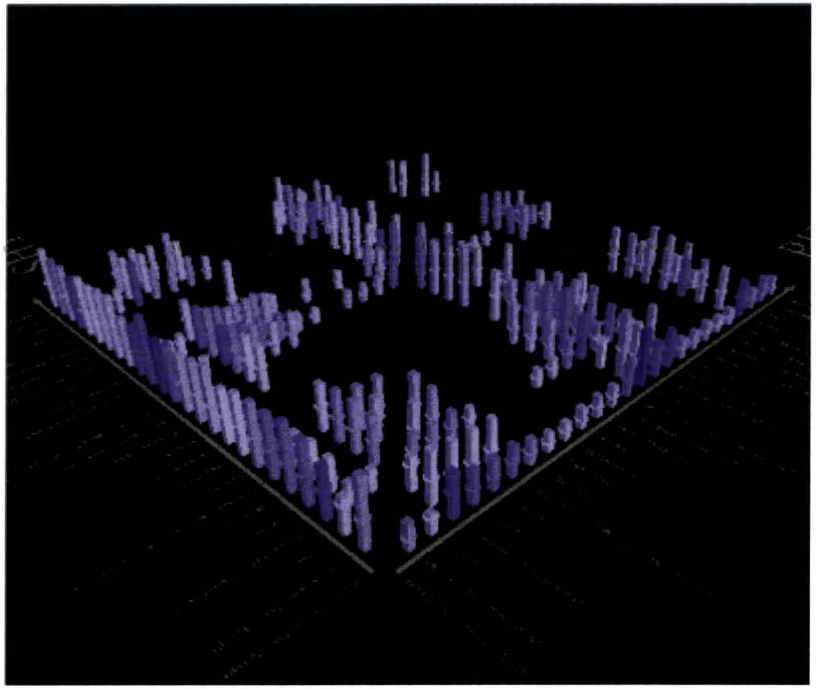
Data visualization is a technique for creating visual interpretation of complex relationships in multidimensional data. Its focus is on human information discourse (interaction) within massive, dynamically changing information spaces.



Analytical techniques/methods and approaches used in data mining

Association Analysis

Association analysis is a method for discovering latent relations among variables in large databases. Association analysis seeks strong rules among data in the database that have different measures of latency; for example, in Amazon.com's Web site, a customer that bought a C++ book may be offered to buy a C++ book and a UML book combo- Based on prior discovery of people buying both books.



Analytical techniques/methods and approaches used in data mining

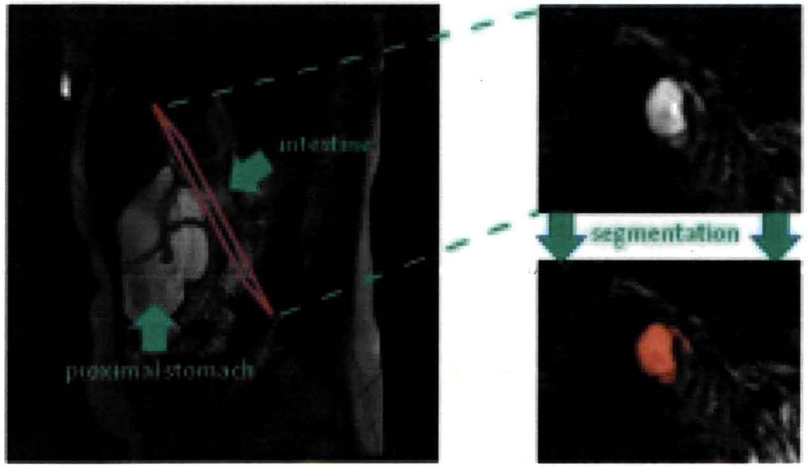
Slice & Dice

Slice and dice business intelligence tools break a body of information down into smaller parts through a systematic reduction of a body of data. Slice and dice method provides the presentation of information in a variety of different and useful ways.

Analytical techniques/methods and approaches used in data mining

Segmentation Algorithms

Segmentation algorithm groups data into segments according to a specific property; typically used to identify characteristics of specific aspects of a research question. In segmentation, the value of data mining is to tell us which data about our research question is relevant and which we could ignore.



Analytical techniques/methods and approaches used in data mining

Clustering Algorithms

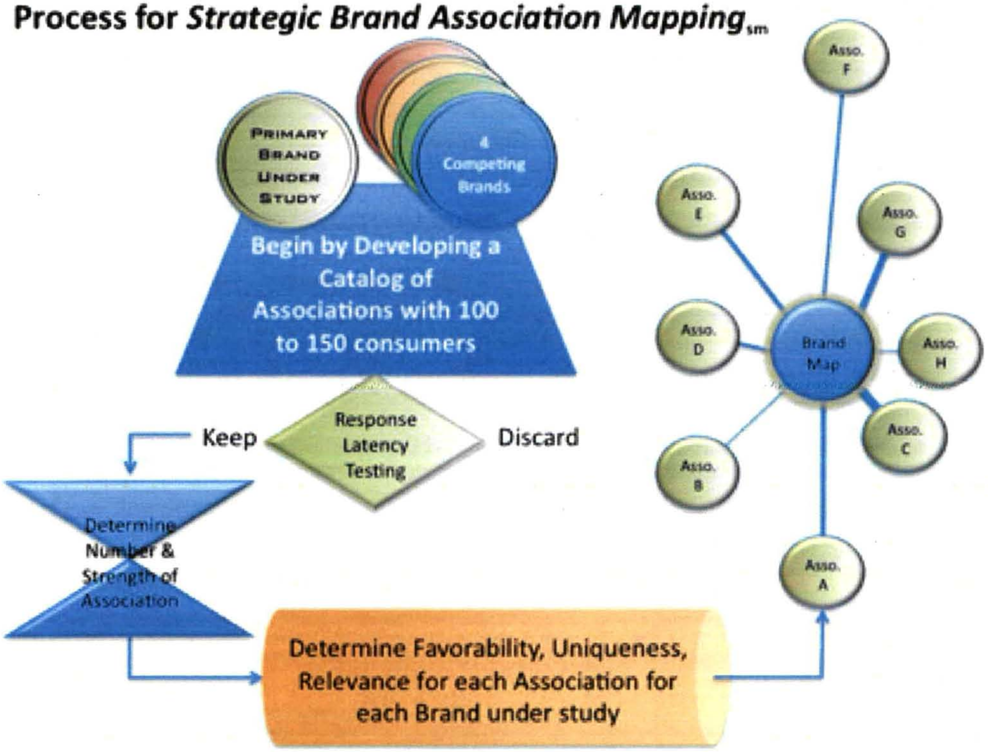
Clustering is a method which aims to partition *n* objects into *k* clusters in which each object belongs to the cluster with the nearest mean. The clustering method involves finding data records that are similar to each other; then clump the self-similar records in clusters (group into clusters simply on the basis of similarity)

Clusters- Group data by logical relationships or preferences; for example, failure data can be grouped according to failure types; “burned fuse” –This information can be mined to identify hardware segments that experience burned fuse.

Analytical techniques/methods and approaches used in data mining

Associations

Data can be mined to identify associations. For example, identifying “burned fuse” due to lightening – at the time of the failure report, the weather condition was unsuspected.



Process for *Strategic Brand Association Mapping*sm

Primary Brand Under Study
4 Competing Brands
Begin by Developing a Catalog of Associations with 100 to 150 consumers
Keep
Response Latency Testing
Discard
Determine Number & Strength of Association
Determine Favorability, Uniqueness, Relevance for each Association for each Brand under study
Asso. F
Asso. E
Asso. G
Asso. D
Brand Map
Asso. H
Asso. C
Asso. B
Asso. A

Analytical techniques/methods and approaches used in data mining

Sequential Patterns

Data patterns and relationships are used for interpretation/evaluation in knowledge discovery. Data is mined to anticipate behavior patterns and trends. For example, most "burned fuses", are discovered in facilities without adequate lightening protection systems.

Analytical techniques/methods and approaches used in data mining

- No one technique solves all data mining problems. Familiarity with a variety of techniques is necessary to provide the best approach to solving data mining problems.

- The real challenge is to select the data mining method/ approach that can best discover the latent information

- To select the best data mining technique/model requires deep understanding of the semantic of each specific problem

- Common problems in selecting a model - Sometimes, models do not work very well. Two common causes are underfitting and overfitting the data.

- Underfitting occurs when the resulting model fails to match patterns of interest in the data.
- Overfitting can also occur when the predicted field is redundant.

Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA)

In statistics, Latent Dirichlet Allocation (LDA) is a generative model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics. LDA is an example of a topic model and was first presented as a graphical model for topic discovery by David Blei, Andrew Ng, and Michael Jordan in 2002.

Latent Dirichlet Allocation (LDA)

Topics in LDA

In LDA, each document may be viewed as a mixture of various topics. This is similar to probabilistic latent semantic analysis (pLSA), except that in LDA the topic distribution is assumed to have a Dirichlet prior. In practice, this results in more reasonable mixtures of topics in a document. It has been noted, however, that the pLSA model is equivalent to the LDA model under a uniform Dirichlet prior distribution.

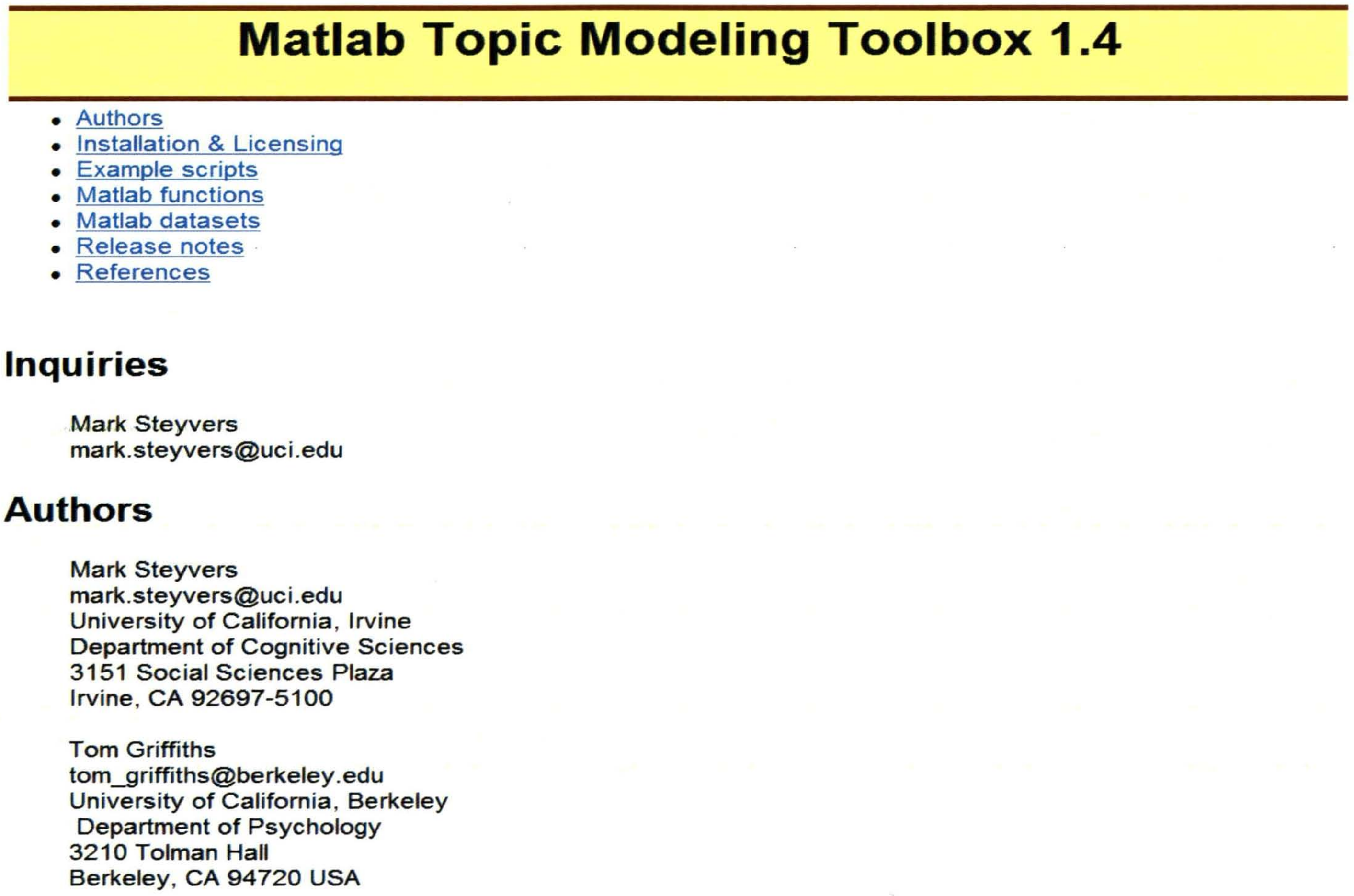
For example, an LDA model might have topics that can be classified as CAT and DOG. However, the classification is arbitrary because the topic that encompasses these words cannot be named. Furthermore, a topic has probabilities of generating various words, such as milk, meow, and kitten, which can be classified and interpreted by the viewer as "CAT". Naturally, cat itself will have high probability given this topic. The DOG topic likewise has probabilities of generating each word: puppy, bark, and bone might have high probability. Words without special relevance, such as the (see function word), will have roughly even probability between classes (or can be placed into a separate category).

A document is given the topics. This is a standard bag of words model assumption, and makes the individual words exchangeable.

Source: WikiPedia (http://en.wikipedia.org/wiki/Latent_Dirichlet_allocation)

Latent Dirichlet Allocation: LDA – MatLab Toolbox

http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm



Matlab Topic Modeling Toolbox 1.4

- Authors
- Installation & Licensing
- Example scripts
- Matlab functions
- Matlab datasets
- Release notes
- References

Inquiries

Mark Steyvers
mark.steyvers@uci.edu

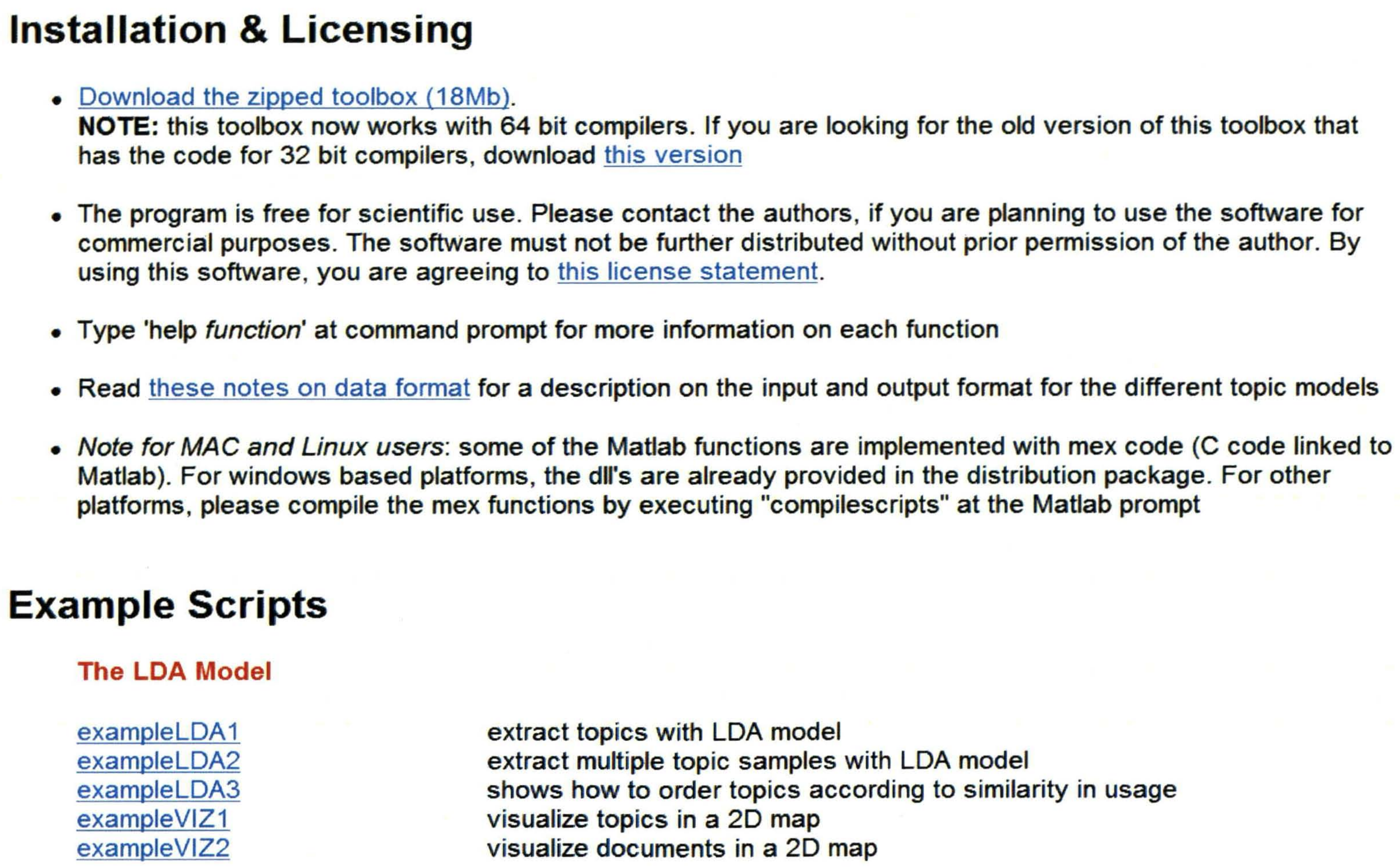
Authors

Mark Steyvers
mark.steyvers@uci.edu
University of California, Irvine
Department of Cognitive Sciences
3151 Social Sciences Plaza
Irvine, CA 92697-5100

Tom Griffiths
tom_griffiths@berkeley.edu
University of California, Berkeley
Department of Psychology
3210 Tolman Hall
Berkeley, CA 94720 USA

Latent Dirichlet Allocation: LDA – MatLab Toolbox

http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm



Installation & Licensing

- Download the zipped toolbox (18Mb).
  NOTE: this toolbox now works with 64 bit compilers. If you are looking for the old version of this toolbox that has the code for 32 bit compilers, download this version
- The program is free for scientific use. Please contact the authors, if you are planning to use the software for commercial purposes. The software must not be further distributed without prior permission of the author. By using this software, you are agreeing to this license statement.
- Type 'help *function*' at command prompt for more information on each function
- Read these notes on data format for a description on the input and output format for the different topic models
- *Note for MAC and Linux users*: some of the Matlab functions are implemented with mex code (C code linked to Matlab). For windows based platforms, the dll's are already provided in the distribution package. For other platforms, please compile the mex functions by executing "compilescripts" at the Matlab prompt

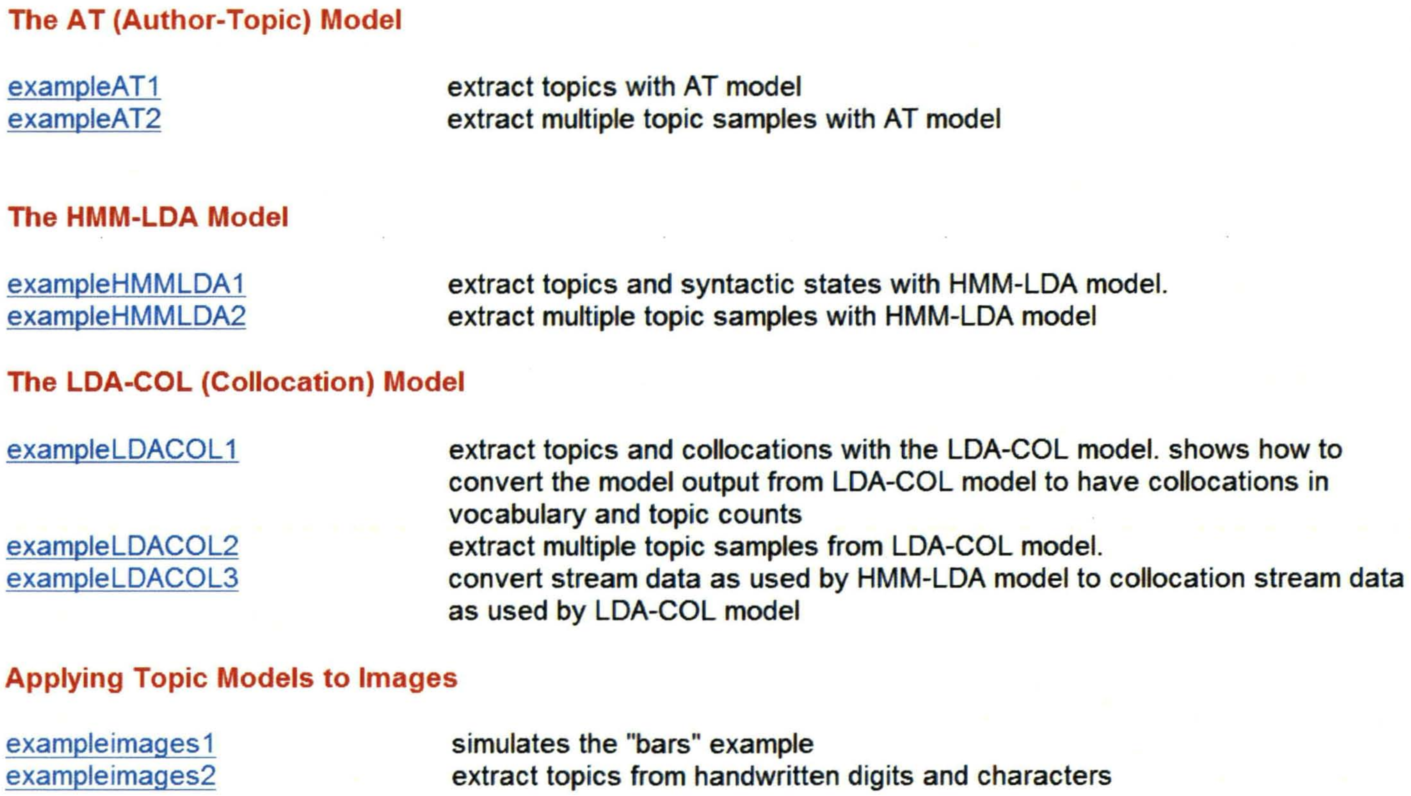
Example Scripts

The LDA Model

exampleLDA1	extract topics with LDA model
exampleLDA2	extract multiple topic samples with LDA model
exampleLDA3	shows how to order topics according to similarity in usage
exampleVIZ1	visualize topics in a 2D map
exampleVIZ2	visualize documents in a 2D map

Latent Dirichlet Allocation: LDA – MatLab Toolbox

http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm



The AT (Author-Topic) Model

exampleAT1	extract topics with AT model
exampleAT2	extract multiple topic samples with AT model

The HMM-LDA Model

exampleHMMLDA1	extract topics and syntactic states with HMM-LDA model.
exampleHMMLDA2	extract multiple topic samples with HMM-LDA model

The LDA-COL (Collocation) Model

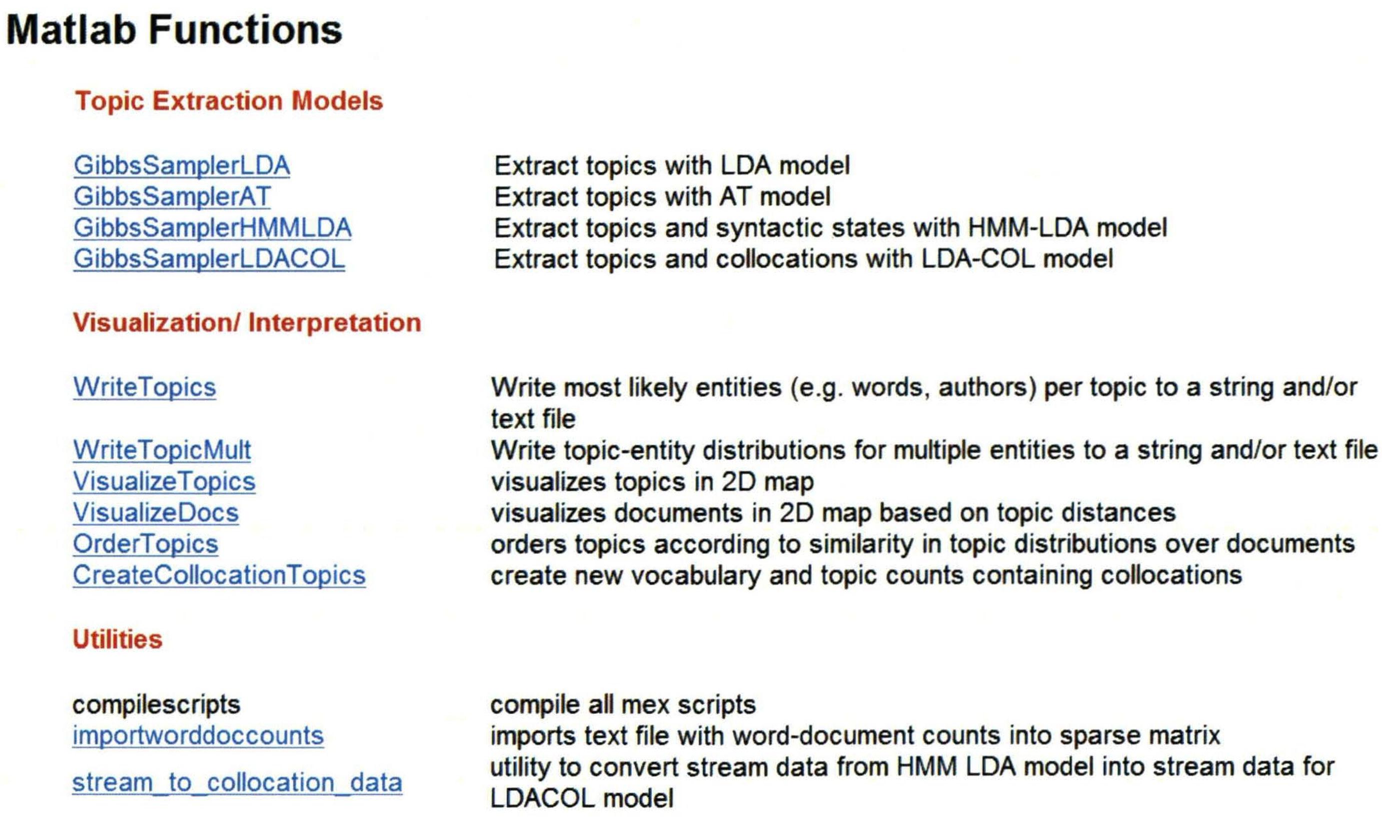
exampleLDACOL1	extract topics and collocations with the LDA-COL model. shows how to convert the model output from LDA-COL model to have collocations in vocabulary and topic counts
exampleLDACOL2	extract multiple topic samples from LDA-COL model.
exampleLDACOL3	convert stream data as used by HMM-LDA model to collocation stream data as used by LDA-COL model

Applying Topic Models to Images

exampleimages1	simulates the "bars" example
exampleimages2	extract topics from handwritten digits and characters

Latent Dirichlet Allocation: LDA – MatLab Toolbox

http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm



Matlab Functions

Topic Extraction Models

GibbsSamplerLDA	Extract topics with LDA model
GibbsSamplerAT	Extract topics with AT model
GibbsSamplerHMMLDA	Extract topics and syntactic states with HMM-LDA model
GibbsSamplerLDACOL	Extract topics and collocations with LDA-COL model

Visualization/ Interpretation

WriteTopics	Write most likely entities (e.g. words, authors) per topic to a string and/or text file
WriteTopicMult	Write topic-entity distributions for multiple entities to a string and/or text file
VisualizeTopics	visualizes topics in 2D map
VisualizeDocs	visualizes documents in 2D map based on topic distances
OrderTopics	orders topics according to similarity in topic distributions over documents
CreateCollocationTopics	create new vocabulary and topic counts containing collocations

Utilities

compilescripts	compile all mex scripts
importworddoccounts	imports text file with word-document counts into sparse matrix
stream_to_collocation_data	utility to convert stream data from HMM LDA model into stream data for LDACOL model

Latent Dirichlet Allocation: LDA – MatLab Toolbox

Matlab Datasets

Psych Review Abstracts (bag of words)

bagofwords_psychreview	document word counts
words_psychreview	vocabulary

Psych Review Abstracts (word stream)

psychreviewstream	successive word and document indices

Psych Review Abstracts (collocation word stream)

psychreviewcollocation	successive word and document indices with function words removed

NIPS proceedings papers (bag of words)

bagofwords_nips	document word counts
words_nips	vocabulary
titles_nips	titles of papers
authors_nips	names of authors
authordoc_nips	document author counts

NIPS proceedings papers (word stream)

nips_stream	successive word and document indices (*note*: the document indices in this dataset do not align with the bag-of-words dataset for nips)

NIPS proceedings papers (collocation stream)

nipscollocation	successive word and document indices with function words removed

Image Data

binaryalphabet	a set of handwritten digits and characters. See exampleimages2 for an application of topic models to this data

Closing Remarks

Data mining is the exploration and analysis of large quantities of data in order to discover ***valid***, ***novel***, potentially ***useful***, and ultimately ***understandable*** patterns in data.

Valid: The patterns hold in general.

Novel: We did not know the pattern beforehand.

Useful: We can devise actions from the patterns.

Understandable: We can interpret and comprehend the patterns.

Closing Remarks

Thank you!

Questions

Suggested Readings

References

LDA MODEL

Steyvers, M. & Griffiths, T. (2007). Probabilistic topic models. In T. Landauer, D McNamara, S. Dennis, and W. Kintsch (eds), *Latent Semantic Analysis: A Road to Meaning.* Laurence Erlbaum

Griffiths, T.L., Steyvers, M., & Tenenbaum, J.B.T. (2007). Topics in Semantic Representation. *Psychological Review*, 114(2), 211-244.

Griffiths, T., & Steyvers, M. (2004). Finding Scientific Topics**.** *Proceedings of the National Academy of Sciences, 101 (suppl. 1), 5228-5235.*

D. Blei, A. Ng, and M. Jordan (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993-1022

AT (AUTHOR-TOPIC) MODEL

Steyvers, M., Smyth, P., Rosen-Zvi, M., & Griffiths, T. (2004). Probabilistic Author-Topic Models for Information Discovery. *The Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* Seattle, Washington.

Rosen-Zvi, M., Griffiths T., Steyvers, M., & Smyth, P. (2004). The Author-Topic Model for Authors and Documents. *In 20th Conference on Uncertainty in Artificial Intelligence. Banff, Canada*

M. Rosen-Zvi, T. Griffiths, P. Smyth, M. Steyvers (submitted). Learning author-topic models from text corpora.

HMM-LDA MODEL

Griffiths, T.L., & Steyvers, M., Blei, D.M., & Tenenbaum, J.B. (2004). Integrating Topics and Syntax**.** In: *Advances in Neural Information Processing Systems, 17.*

LDA-COL MODEL

Griffiths, T.L., Steyvers, M., & Tenenbaum, J.B.T. (2007). Topics in Semantic Representation. *Psychological Review*, 114(2), 211-244. **See pages 234-236.**

Suggested Readings

1. A. Acquisti, S. Gritzalis, C. Lambrinoudakis and S. Vimercati, Digital Privacy: Theory, Technologies, and Practices, Auerbach Publications, 2008, ISBN:9781420052176
2. B. Ville, Decision Trees for Business Intelligence and Data Mining: Using SAS Enterprise Miner, SAS Publishing, 2006, ISBN:9781590475676
3. B. Larson, Delivering Business Intelligence with Microsoft SQL Server 2008, McGraw-Hill/Osborne, 2009, ISBN:9780071549448
4. D G. Schwartz, Encyclopedia of Knowledge Management, IGI Publishing, 2006, ISBN:9781591405733
5. D. T. Larose, Discovering Knowledge in Data: An Introduction to Data Mining, John Wiley & Sons, 2005, ISBN:9780471666578
6. D. Taniar, Research and Trends in Data Mining Technologies and Applications, IGI Publishing, 2007, ISBN:9781599042718
7. D. Taniar, Data Mining and Knowledge Discovery Technologies, IGI Publishing, 2008, ISBN:9781599049601
8. D. Hand, H. Mannila and P. Smyth, Principles of Data Mining, The MIT Press, 2001, ISBN:9780262082907
9. D. Laha and P. Mandal, Handbook of Computational Intelligence in Manufacturing and Production Management, IGI Publishing, 2008, ISBN:9781599045825
10. D. Harts, Microsoft Office 2007 Business Intelligence: Reporting, Analysis, and Measurement from the Desktop, McGraw-Hill/Osborne, 2008, ISBN:9780071494243
11. D. Zhang and J. Tsai, Advances in Machine Learning Applications in Software Engineering, IGI Publishing, 2007, ISBN:9781591409410

Suggested Readings

12. E. Veerman, T. Lachev, D. Sarka and J. Loria, MCTS Self-Paced Training Kit (Exam 70-445): Microsoft SQL Server 2005 Business Intelligence: Implementation and Maintenance, Microsoft Press, 2008, ISBN:9780735623415
13. F. Masseglia, P. Poncelet and M. Teisseire, Successes and New Directions in Data Mining, IGI Publishing, 2008, ISBN:9781599046457
14. G. Felici and C. Vercellis, Mathematical Methods for Knowledge Discovery and Data Mining, IGI Publishing, 2008, ISBN:9781599045283
15. G. S. Linoff and M. J. Berry, Mining the Web: Transforming Customer Data into Customer Value, John Wiley & Sons, 2001, ISBN:9780471416098
16. H. O. Nigro, S. E. Císaro and D. H. Xodo, Data Mining with Ontologies: Implementations, Findings, and Frameworks, IGI Publishing, 2008, ISBN:9781599046181
17. H. Hsu, Advanced Data Mining Technologies in Bioinformatics, IGI Publishing, 2006, ISBN:9781591408635
18. J. Wang, Encyclopedia of Data Warehousing and Mining, Volume II, I-Z , Idea Group Publishing, 2006, ISBN:9781591405573

19. J. Wang, Encyclopedia of Data Warehousing and Mining, Volume I, A-H, IGI Publishing, 2006, ISBN:9781591405573
20. H. Nemati, Information Security and Ethics: Concepts, Methodologies, Tools, and Applications, IGI Publishing, 2008, ISBN:9781599049373
21. K. E. Voges and N. L. Pope, Business Applications and Computational Intelligence, IGI Publishing, 2006, ISBN:9781591407027
22. L. Lobel, A. Brust and S. Forte, Programming Microsoft SQL Server 2008, Microsoft Press, 2009, ISBN:9780735625990

Suggested Readings

23. L. C. Rivero, J. H. Doorn and V. E. Ferraggine, Encyclopedia of Database Technologies and Applications, IGI Publishing, 2006, ISBN:9781591405603
24. L. Langit, K. S. Goff, D. Mauri, S. Malik and J. Welch, Smart Business Intelligence Solutions with Microsoft SQL Server 2008, Microsoft Press, 2009, ISBN:9780735625808
25. M. T. Jones, Artificial Intelligence: A Systems Approach, Infinity Science Press, 2008, ISBN:9780977858231
26. M. S. Hodges, Computers: Systems, Terms and Acronyms, 17th Edition, SemCo, 2007, ISBN:9780979443206
27. M. Dodgson, D. Gann and A. Salter, Think, Play, Do: Technology, Innovation, and Organization, Oxford University Press, 2005, ISBN:9780199268092
28. M. Khosrow-Pour, Encyclopedia of Information Science and Technology, Volume II, Idea Group Publishing, 2005, ISBN:9781591405535
29. M. Edesess, The Big Investment Lie: What Your Financial Advisor Doesn't Want You to Know, Berrett-Koehler Publishers, 2007, ISBN:9781576754078
30. O. Maimon and L. Rokach, Data Mining and Knowledge Discovery Handbook, Springer, 2005, ISBN:9780387244358
31. P. Janus, Pro PerformancePoint Server 2007: Building Business Intelligence Solutions, Apress, 2008 , ISBN:9781590599617
32. R. Matignon, Data Mining Using SAS Enterprise Miner, John Wiley & Sons, 2007, ISBN:9780470149010
33. R. S. Busch, Healthcare Fraud: Auditing and Detection Guide, John Wiley & Sons, 2008, ISBN:9780470127100

Suggested Readings

34. R. Feldman and J. Sanger, The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data, Cambridge University Press, 2007, ISBN:9780521836579

35. S. J. Ovaska, Computationally Intelligent Hybrid Systems: The Fusion of Soft Computing and Hard Computing, John Wiley & Sons, 2005, ISBN:9780471476689

36. S. Ananiadou and J. McNaught, Text Mining for Biology and Biomedicine, Artech House, 2006, ISBN:9781580539845

37. T. C. Redman, Data Driven: Profiting from Your Most Important Business Asset, Harvard Business Press, 2008 , ISBN:9781422119129

38. V. Sugumaran, Intelligent Information Technologies: Concepts, Methodologies, Tools, and Applications, IGI Publishing, 2008, ISBN:9781599049410

39. W. W. Eckerson, Performance Dashboards: Measuring, Monitoring, and Managing Your Business, John Wiley & Sons, 2006, ISBN:9780471724179

40. W. Hsu, M. L. Lee and J. Wang, Temporal and Spatio-Temporal Data Mining, IGI Publishing, 2008, ISBN:9781599043876

41. X. and I. Davidson, Knowledge Discovery and Data Mining: Challenges and Realities, IGI Publishing, 2007, ISBN:9781599042527

42. Z. Ma, Intelligent Databases: Technologies and Applications, IGI Publishing, 2007, ISBN:9781599041209