ECE 594N: Data Mining
Spring 2003

B. S. MANJUNATH
RM 3157 ENGR I
Tel:893-7112
manj@ece.ucsb.edu
http://vision.ece.ucsb.edu/Manjunath

Why data mining?

Our ability to capture and store data far outpaces our ability to process and exploit it
--Fayad and Uturusamy, ACM 2002

- Digital storage capacity has doubled every nine months for the past decade.
  - data tombs –effectively, write-only;
  - data is deposited to merely rest in peace.
- Data rich but information poor
  - Data mining can help discover knowledge

zaiane

Why data mining? Why now?

Necessity is the mother of invention

- Technology is available to help us collect data
  - Bar code, scanners, satellites, cameras, TV news..
- Technology to help us store data
  - Databases, data warehouses,…

Need tools to process and interpret the data

zaiane

What is the need now?

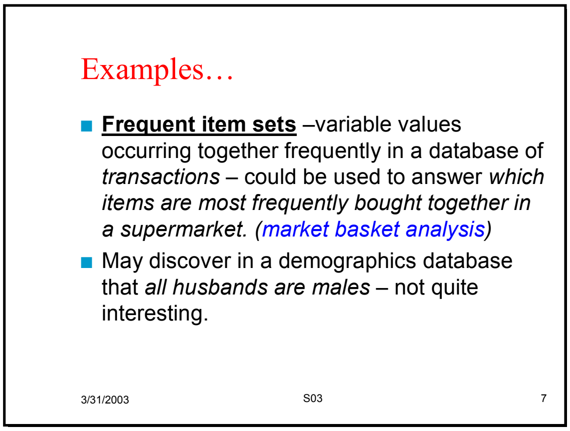
- Extract interesting knowledge --- rules, regularities, patterns, constraints--- from data in large collections.

Statistics vs data mining

- Statistics: data collected using efficient strategies to answer specific questions.
- DM: do not play role in data collection; deals with larger data sets; issues include (a) how to store and access data (b) speed/time needed for analysis; (c) seeking novelty.
- "when a statistician has an idea he or she writes a paper; a computer scientist starts a company"—friedman.

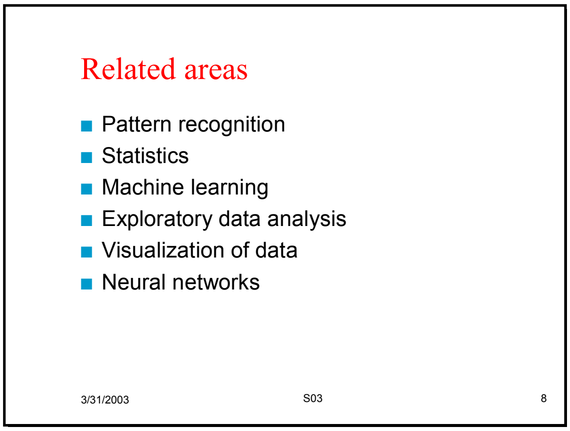
So What is data mining?

- "Analysis of (often large) observational data sets to find unsuspected relationships and to summarize data in novel ways that are both understandable and useful to the data owner" (Hand et. al.)
- Identification of interesting "structure" in data
  - *Structure* →patterns, statistical or predictive *models* of data, or relationships among parts of data
  - *Interesting* is much more difficult to define
- Computer automated exploratory data analysis of large complex data sets.



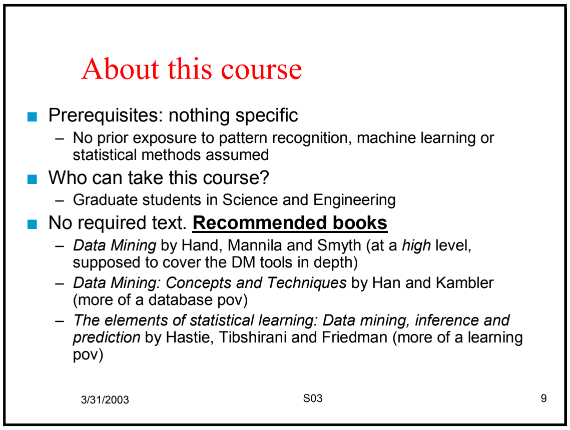
Examples…

- **Frequent item sets** –variable values occurring together frequently in a database of *transactions* – could be used to answer *which items are most frequently bought together in a supermarket. (market basket analysis)*
- May discover in a demographics database that *all husbands are males* – not quite interesting.



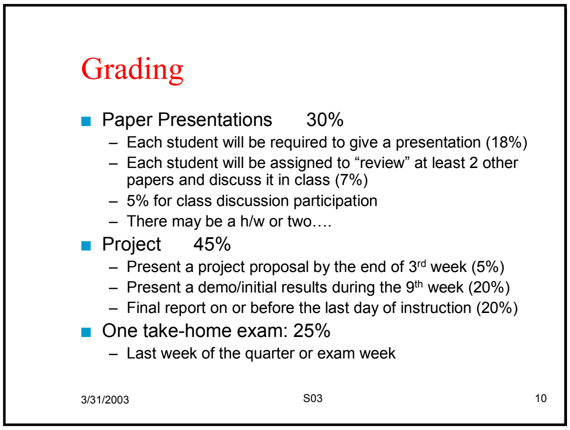
Related areas

- Pattern recognition
- Statistics
- Machine learning
- Exploratory data analysis
- Visualization of data
- Neural networks



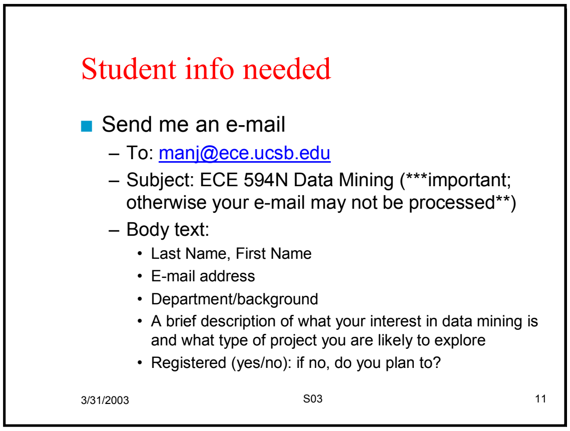
About this course

- Prerequisites: nothing specific
  - No prior exposure to pattern recognition, machine learning or statistical methods assumed
- Who can take this course?
  - Graduate students in Science and Engineering
- No required text. **Recommended books**
  - *Data Mining* by Hand, Mannila and Smyth (at a *high* level, supposed to cover the DM tools in depth)
  - *Data Mining: Concepts and Techniques* by Han and Kambler (more of a database pov)
  - *The elements of statistical learning: Data mining, inference and prediction* by Hastie, Tibshirani and Friedman (more of a learning pov)



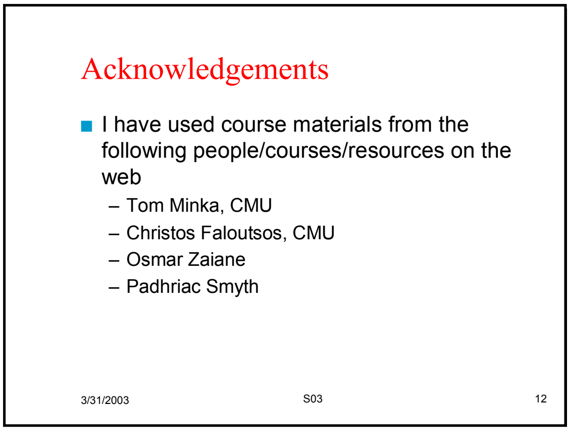
Grading

- Paper Presentations 30%
  - Each student will be required to give a presentation (18%)
  - Each student will be assigned to "review" at least 2 other papers and discuss it in class (7%)
  - 5% for class discussion participation
  - There may be a h/w or two….
- Project 45%
  - Present a project proposal by the end of 3rd week (5%)
  - Present a demo/initial results during the 9th week (20%)
  - Final report on or before the last day of instruction (20%)
- One take-home exam: 25%
  - Last week of the quarter or exam week



Student info needed

- Send me an e-mail
  - To: manj@ece.ucsb.edu
  - Subject: ECE 594N Data Mining (***important; otherwise your e-mail may not be processed**)
  - Body text:
    - Last Name, First Name
    - E-mail address
    - Department/background
    - A brief description of what your interest in data mining is and what type of project you are likely to explore
    - Registered (yes/no): if no, do you plan to?



Acknowledgements

- I have used course materials from the following people/courses/resources on the web
  - Tom Minka, CMU
  - Christos Faloutsos, CMU
  - Osmar Zaiane
  - Padhriac Smyth

(tentative) course outline

- Introduction (1)
- Data Warehouse (1)
- Data Preprocessing (1)
- Classification Methods (2)
- Clustering Methods (4)
- Pattern finding (2)
- Applications (1-2)
- Multimedia Mining (2)
- Survey of recent research (2)
  - Presentations
- Course project (2)
  - Presentations

achieve a balance between databases and machine learning/pattern recognition related work.

What is unique about data mining?

- Scaling analysis to large databases
  - What can be done with large data sets that can not be loaded and manipulated in main memory?
  - How might we avoid scanning an entire very large database while reliably searching for patterns?
- Scaling to high-dimensional data/models
  - Humans are not quite effective at formulating hypothesis when data sets have large number of variables (hundreds or thousands)
  - Model derived from automated discovery can be used to find lower-dimensional sub-spaces that are better suited for human interpretation

What is unique to data mining?

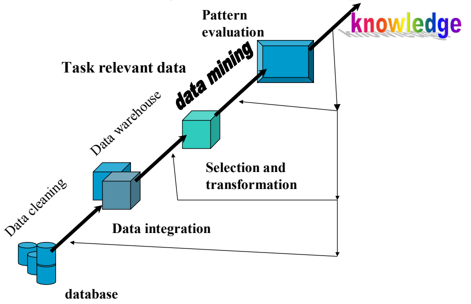
- Automating search
- Finding patterns and models interesting to users
  - Classical methods focus on notions of accuracy (how well the model predicts data) and utility (how to measure the benefit of the derived pattern)
  - New measures? E.g., understandability of a model…..

Data characteristics & issues (PR pov)

- Multivariate data: *N x d* data matrix
  - *N* objects or individuals or transactions
  - Each object has *d* measurements or attributes
  - **x** is a *d*-dimensional vector of measurement, for each of the *N* objects
- *Classification:* learn the mapping from vector **x** to *y*, where *y* is a categorical or scalar target variable.
- *Regression*: *y* takes on real values
- *Clustering*: map **x** into a set of categories
- *Density estimation:* estimate the pdf of **x**.

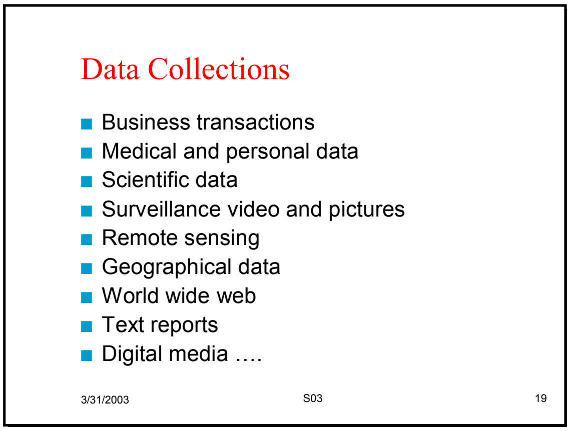
zaiane

Data Mining—a database pov



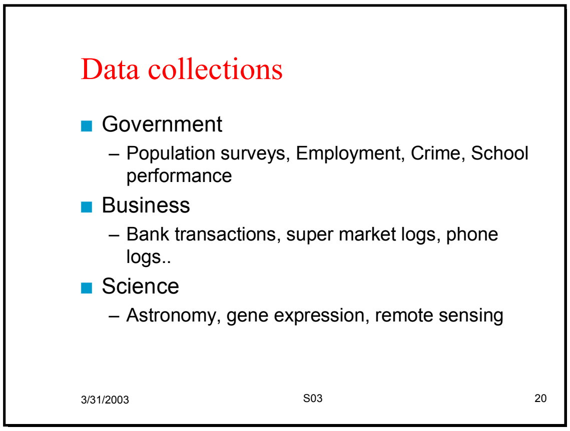
Transactions data

- Both *N* and *d* can be very large in practice;
  - E-commerce web site, N=100,0000 baskets, d=10,000 items.
  - Computing a pairwise correlation matrix needs O(Nd^2) time and O(d^2) memory;
  - Transaction data typically is very sparse, a typical basket may contain 5-10 items.
- *Curse of dimensionality:* amount of data needed for reliable density estimation scales exponentially in *d*.



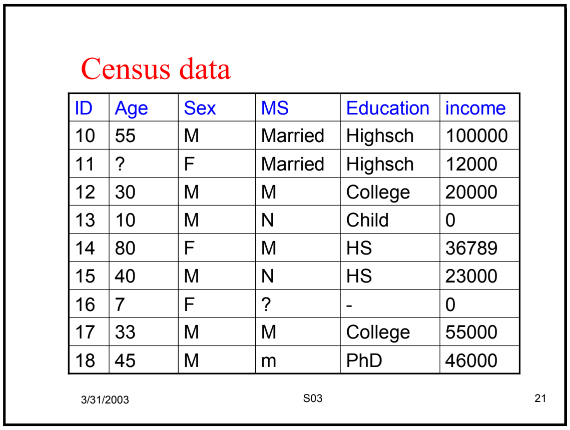
Data Collections

- Business transactions
- Medical and personal data
- Scientific data
- Surveillance video and pictures
- Remote sensing
- Geographical data
- World wide web
- Text reports
- Digital media ….



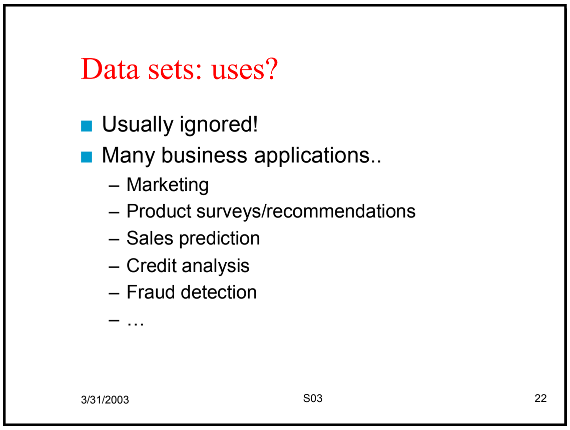
Data collections

- Government
  - Population surveys, Employment, Crime, School performance
- Business
  - Bank transactions, super market logs, phone logs..
- Science
  - Astronomy, gene expression, remote sensing



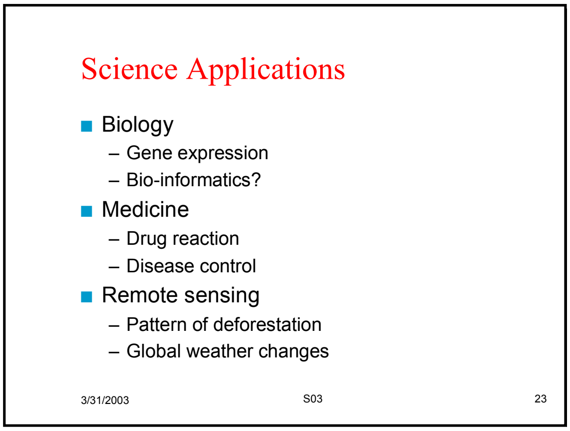
Census data

ID	Age	Sex	MS	Education	income
10	55	M	Married	Highsch	100000
11	?	F	Married	Highsch	12000
12	30	M	M	College	20000
13	10	M	N	Child	0
14	80	F	M	HS	36789
15	40	M	N	HS	23000
16	7	F	?	-	0
17	33	M	M	College	55000
18	45	M	m	PhD	46000



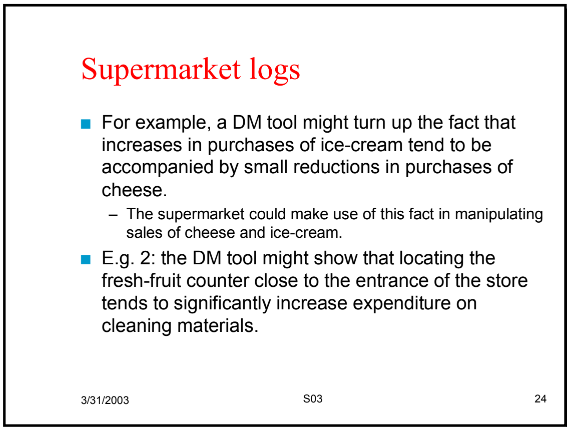
Data sets: uses?

- Usually ignored!
- Many business applications..
  - Marketing
  - Product surveys/recommendations
  - Sales prediction
  - Credit analysis
  - Fraud detection
  - …



Science Applications

- Biology
  - Gene expression
  - Bio-informatics?
- Medicine
  - Drug reaction
  - Disease control
- Remote sensing
  - Pattern of deforestation
  - Global weather changes



Supermarket logs

- For example, a DM tool might turn up the fact that increases in purchases of ice-cream tend to be accompanied by small reductions in purchases of cheese.
  - The supermarket could make use of this fact in manipulating sales of cheese and ice-cream.
- E.g. 2: the DM tool might show that locating the fresh-fruit counter close to the entrance of the store tends to significantly increase expenditure on cleaning materials.

Problems with data

- Most data too complex for conventional tools
  - Too fragmented (phone calls)
  - Too complex to model directly (images, cells,..)
  - Too much spurious phenomena (register logs)
  - Too much variation even within a given domain/data

Problems with machine learning

- Too specific (blind?)
- Require pre-defined goals; modeling strategy
- Precise, automated answer to specific queries
  - May overlook other crucial aspects of data

Data Mining

Utilization of statistics/machine learning methods within an exploratory framework

Emphasizes:

- Visualization
- Exploratory data analysis
- Non-parametric methods
- Serendipity

Commercial software: what do they have?

- Functionality
  - Decision trees
  - Association Rules
  - Nearest neighbor methods
  - Clustering
  - Feature extraction
  - Visualization
- Methods
  - Neural networks, Bayesian methids, Genetic algorithms, SOM, Fuzzy systems,….

E.g. VisualMine

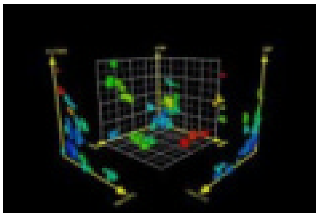
- “**Visual Data Mining**
  The discovery of information hidden in data is achieved through visual metaphors that, according to the principles stated by cognitive sciences, increase the intuitive comprehension of complex phenomena”

Commercial software

- IBM Intelligent miner
- Mine set (SGI)—no longer available
- DB miner
- DigiMine

visualmine

Customer segmentation: red clusters depict most profitable customers



Associations (examples from IBM)

- Given a database of transactions, where each transaction consists of a set of items, discover all associations such that the presence of one set of items in a transaction implies the presence of another set of items.
  - *"30% of people who buy diapers also buy beer."*

Classification

- Given examples of objects belonging to different groups, develop profile of each group in terms of attributes of the objects. This profile is then used to predict the group of a new object.
  - *"Buyers of expensive sport cars are typically young urban professionals whereas luxury sedans are bought by elderly wealthy persons."*
- Application: A bank wants to assess the credit-worthiness of its customers. By analyzing the loan-history records with the classification technique, the bank gets a precise profile of high, medium, and low-risk customers.

Sequential Patterns

- Given a database of transactions over a period of time, find inter-transaction patterns such that the presence of a set of items is followed by another set of items.
  - *"10% of people with diabetes develop a treatable loss in eyesight."*
- Application: A direct mailer wants to maximize cross-selling opportunities. By applying the *Associations and Sequential Patterns* technique to historical order data, the direct mailer can find out what articles sell together and what articles are bought in a sequence over time.
  - The mailer uses this information to decide on placements of articles in the catalog and for deciding what flyers to attach with a bill.

Similar Time Sequences

- *Given a database of time sequences, find sequences similar to a given one, or find all occurrences of similar sequences.*
  - *"The closing net asset value of the Harbor International mutual fund has been similar to that of Ivy International and Scudder Global Fund."*
- Application: A retailer wants to optimize purchasing and store-keeping. By applying the Similar Time Sequences technique, the retailer can find groups of products that have similar forecasted seasonal sales for next year and use this information for combining purchases and inventory replenishment.

More examples

- An auto insurer wants to study lapsing and retention among their customers. By applying the *Sequential Patterns technique*, the insurer can understand what events lead to lapses.
- A medical insurer is interested in detecting insurance fraud. By applying the *associations technique*, the insurer can determine if there is a ring of providers indulging in ping-ponging of patients between them.

Examples (Time, dec 2002)

- Software developed by Autonomy, based in Cambridge, England, connected BAE's research databases and alerted civilian aircraft engineers to the fact that the wing-construction problem they were working on was also being addressed by the company's military division. Ending this duplication helped the company save millions of dollars.

Examples (Time dec 2002)

- The data-mining algorithms of ClearForest, based in New York City, are at work within both Israeli security agencies and NASDAQ. Israel uses them to drill for hidden connections among suspected terrorists: say, a pattern of phone calls shortly before each of several suicide bombings. NASDAQ uses the same software to detect block trades of stock quietly placed just before the release of company news — including sales by relatives of ImClone's founder, Sam Waksal, who this fall pleaded guilty to insider-trading charges, and his friend Martha Stewart, who remains under investigation (and has denied any wrongdoing).

Is DM an intellectual discipline?

--J. H. Friedman

- So far ----- No, not yet!
  - DM packages implement well known methods from machine learning, pattern recognition,…
  - Emphasize GUI/visualization
  - No regard to performance (black box?)
  - Goal is to get to market quickly
- However, in the future, yes!
  - “every time the amount of data increases by a factor of ten, we should totally rethink how we analyze it”

Requirements/Challenges revisited

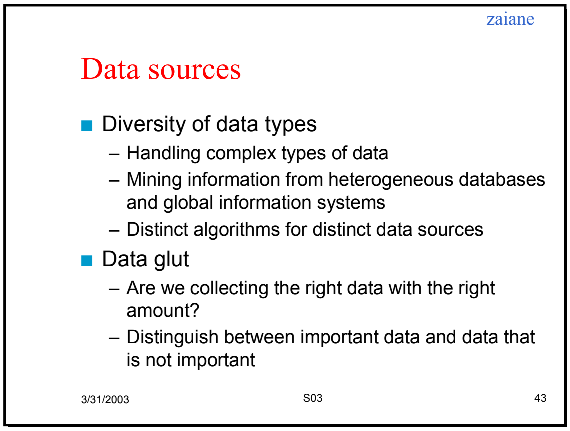
- Security and social issues
- User interface issues
- Mining methodology
- Performance issues
- Data sources issues

Security and social issues

- Societal impact
  - Private and sensitive data is gathered and mined without individual’s knowledge and/or consent.
  - New, implicit knowledge disclosed (confidentiality, integrity)
  - Appropriate use and distribution of discovered knowledge (sharing)
- Regulations
  - Need for privacy and DM policies

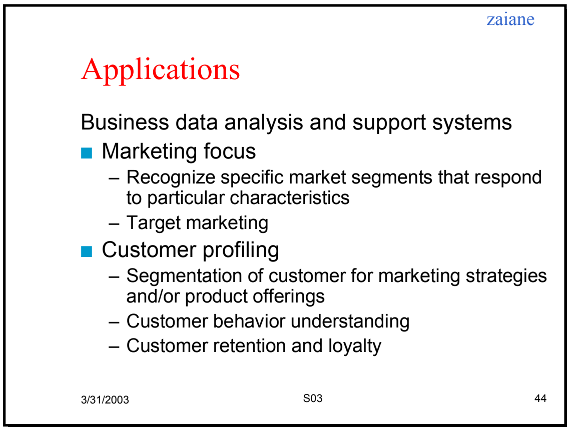
Performance issues

- Efficiency and scalability of DM algorithms
  - Linear algorithms are needed; certainly no exponential schemes
  - Sampling
- Parallel and distributed methods
  - Incremental mining
  - Can we divide and conquer



Data sources

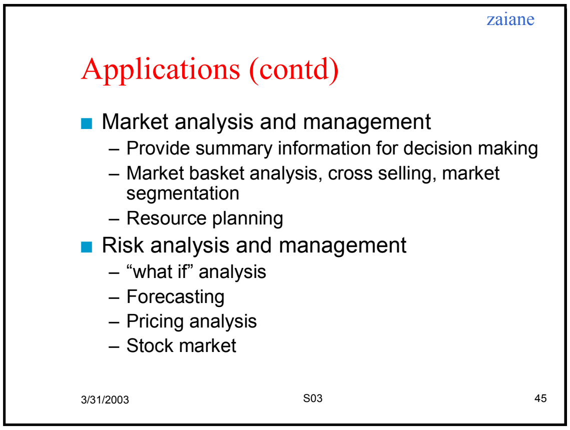
- Diversity of data types
  - Handling complex types of data
  - Mining information from heterogeneous databases and global information systems
  - Distinct algorithms for distinct data sources
- Data glut
  - Are we collecting the right data with the right amount?
  - Distinguish between important data and data that is not important



Applications

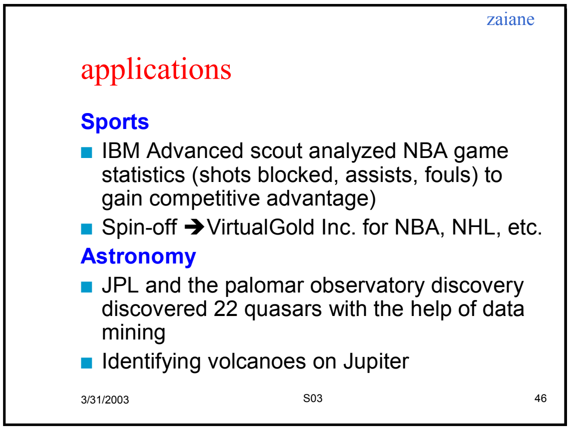
Business data analysis and support systems

- Marketing focus
  - Recognize specific market segments that respond to particular characteristics
  - Target marketing
- Customer profiling
  - Segmentation of customer for marketing strategies and/or product offerings
  - Customer behavior understanding
  - Customer retention and loyalty



Applications (contd)

- Market analysis and management
  - Provide summary information for decision making
  - Market basket analysis, cross selling, market segmentation
  - Resource planning
- Risk analysis and management
  - “what if” analysis
  - Forecasting
  - Pricing analysis
  - Stock market



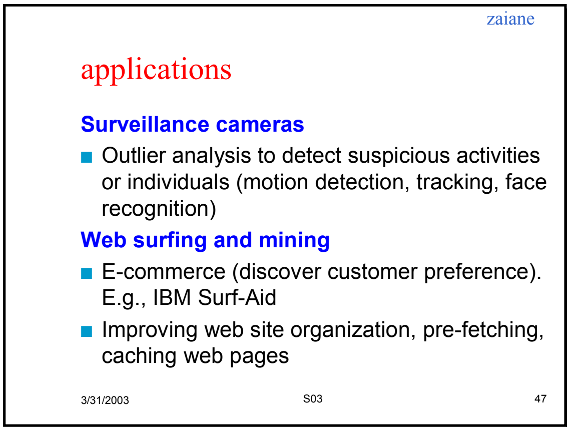
applications

Sports

- IBM Advanced scout analyzed NBA game statistics (shots blocked, assists, fouls) to gain competitive advantage)
- Spin-off → VirtualGold Inc. for NBA, NHL, etc.

Astronomy

- JPL and the palomar observatory discovery discovered 22 quasars with the help of data mining
- Identifying volcanoes on Jupiter



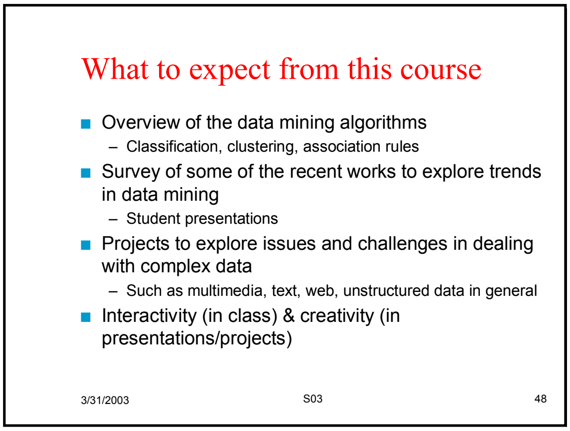
applications

Surveillance cameras

- Outlier analysis to detect suspicious activities or individuals (motion detection, tracking, face recognition)

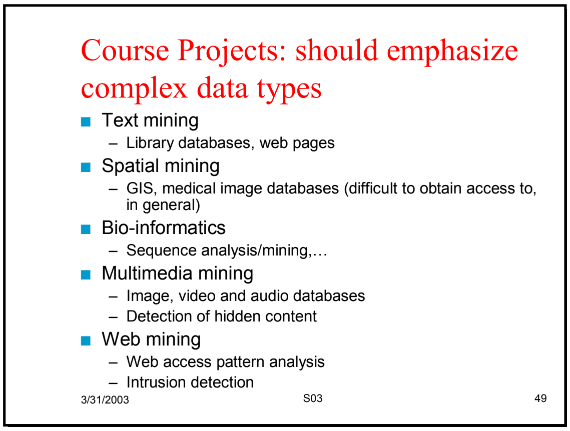
Web surfing and mining

- E-commerce (discover customer preference). E.g., IBM Surf-Aid
- Improving web site organization, pre-fetching, caching web pages



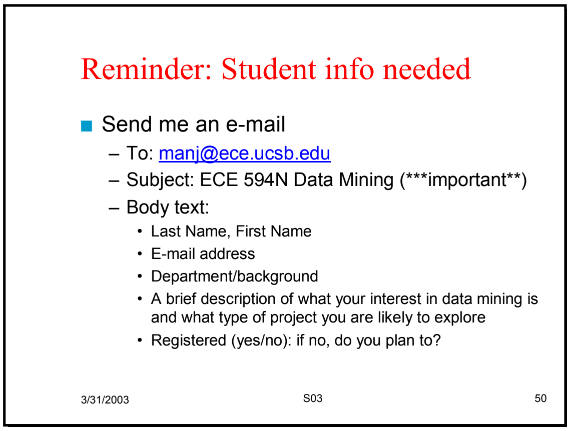
What to expect from this course

- Overview of the data mining algorithms
  - Classification, clustering, association rules
- Survey of some of the recent works to explore trends in data mining
  - Student presentations
- Projects to explore issues and challenges in dealing with complex data
  - Such as multimedia, text, web, unstructured data in general
- Interactivity (in class) & creativity (in presentations/projects)



Course Projects: should emphasize complex data types

- Text mining
  - Library databases, web pages
- Spatial mining
  - GIS, medical image databases (difficult to obtain access to, in general)
- Bio-informatics
  - Sequence analysis/mining,…
- Multimedia mining
  - Image, video and audio databases
  - Detection of hidden content
- Web mining
  - Web access pattern analysis
  - Intrusion detection



Reminder: Student info needed

- Send me an e-mail
  - To: manj@ece.ucsb.edu
  - Subject: ECE 594N Data Mining (***important**)
  - Body text:
    - Last Name, First Name
    - E-mail address
    - Department/background
    - A brief description of what your interest in data mining is and what type of project you are likely to explore
    - Registered (yes/no): if no, do you plan to?