

Too Much Data

Kelley Rivoire

Almost everyone has seen the unending columns of numbers associated with the movie *The Matrix*. Today, that ceaseless stream of data is reality for researchers in nearly all disciplines. Every day, telemarketers and pollsters survey millions of people. Surely the results aren't analyzed one by one. What about a series of measurements from a device, one every second? How does the owner spot a malfunction? Researchers can't escape the volumes of data accumulating; instead, they must learn to handle and analyze them.

As technology, particularly instrumentation, improves, more and more data can be collected faster and faster, but the management of this data cannot always keep pace. New methods are needed for storage, searching, sort-

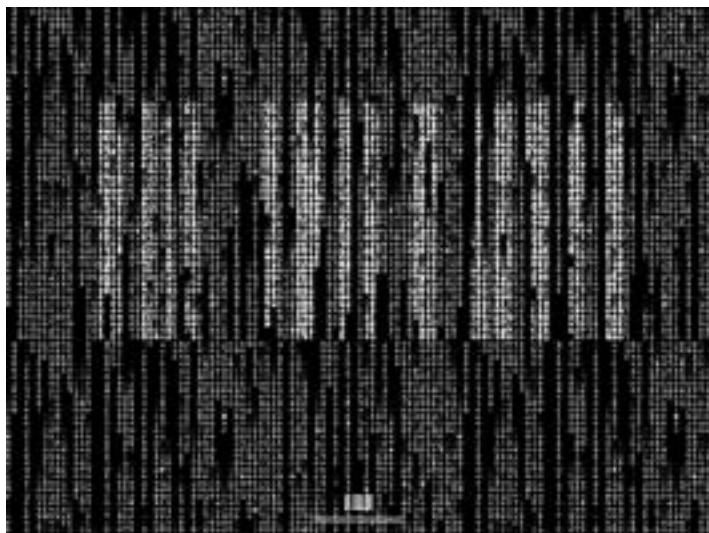
ing, and analyzing. As Caltech Professor of Astronomy George Djorgovski writes, "Raw data, no matter how expensively obtained, are of limited utility without the effective ability to process them quickly and thoroughly, and to refine the essence of scientific knowledge from them." The goal is to find "interesting scientific results" using "statistically sound and objective" techniques that are "automated as much as possible."¹

Usama Fayyad, author of *Advances in Knowledge Discovery and Data Mining*, defines the new field of data mining as "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data."² This problem of finding patterns in large amounts of data unites many scientific disciplines and requires all to face similar dilemmas.

In biology, the recent sequencing of genomes will allow researchers to map genes with their function—if they can cope with thousands of simultaneous measurements. Physicists might soon discover a new, important particle—if they can recognize small anomalies in petabytes of data. Web searchers can access information from people all over the world—if search engines can accurately index and rank Web pages. Examining and understanding data problems in this broad range of fields and the solutions proposed by scientists in each is important; because problems encountered by each technical area are similar, solutions used by one discipline can help researchers in a completely different field.

Web Crawling

Though relatively new, the Internet is inundated with links and is thought by many to be "too unstructured."³ To navigate the Web, a user must sift through vast amounts of data, much of which is irrelevant. In opposition to this hypothesis of an unstructured Web is the idea that, though complicated, the Web is a graph that can be traversed. The successful development of search engines indicates at least a modicum of structure present in the Web. In 1994, the most popular search engine recorded 1,500 queries per day. In 1997, Altavista reported about 20 million queries a day.⁴ Crawling the Web and ranking pages have been the basis of a number of recent computational and mathematical studies, the most prominent of which led to the 1998 creation of the Google search engine by graduate students Larry Page and Sergey Brin at Stanford University.⁵



Google

Prior to Google, many popular search engines such as Yahoo! were run using indices created and maintained by humans. Automated search engines, on the other hand, often led to results listing irrelevant matches. Google sought to use an automated search engine while still maintaining high-quality results. Google's goal was to place the most relevant links at the top of the results list, so that users could sift through fewer irrelevant links.

How does Google rank pages so accurately? The idea is that more relevant pages will have references from many other Web sites, whereas less relevant pages will not be mentioned. How does Google tell the importance of the referring page? It ranks these Web sites also. These ideas are the basis of the PageRank system created by Google to rate each Web page. PageRank counts the number of links pointing to a Web page, normalizing by the total number of links from a given page. The normalization prevents any single Web page's references from contributing too much to a ranking. PageRank is also special because the text that references a link is indexed with the page referred to rather than the referring page. This provides a better description of pages and allows indexing of pages containing images or other data types without text that cannot be otherwise indexed. PageRank also attaches increased weighting to larger, bold words, taking the visual factors of a Web page into account. When a user enters a query, Google counts the weighted hits to create a PageRank. For queries with multiple words, hits occurring near each other receive a greater weighting.

By using these techniques of crawling, indexing, and sorting the Web, Google achieves good precision in a small amount of time in its searches. As time progresses, it will continue to update its methodology to save time and storage space.⁴

Teoma

Though Google is the dominant search engine on the Web today, computer scientists are developing other search engines to improve Google's techniques. *Teoma*, meaning "expert" in Gaelic, was founded by Rutgers University Computer Science Professor Apostolos Gerasoulis in April 2000. One year later, Teoma.com was launched on the Internet, and in September 2001, Ask Jeeves, Inc., began to use the Teoma search engine, making it the third-most used today. Early this year, Teoma released an improved version of its engine, Teoma 2.0, with a set of advanced tools. Teoma hopes to compete with Google and provide an alternative search engine for Web browsers.⁶

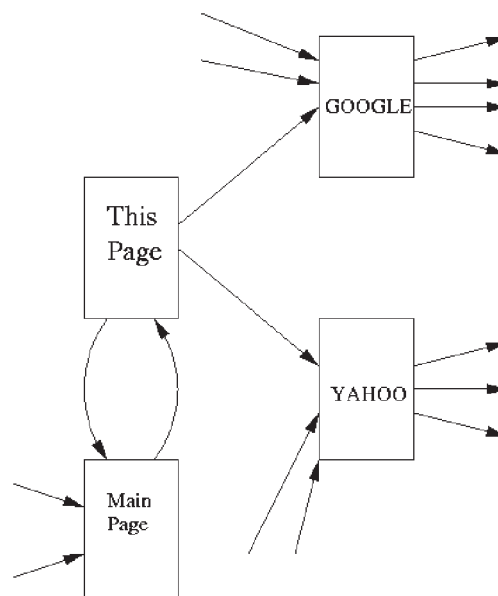
How does Teoma work, and how is it different from Google? Teoma lauds its "community-based" approach

to searching that uses only Web pages in the area of the query. Teoma defines communities as "groups of Web pages that are closely related to the same subject."⁷ In Google, a ranking is developed by counting links from other sites. In Teoma, on the other hand, only sites concerned with the subject of the search string are used in counting links. The idea, as journalist Kieren McCarthy writes, is that "Google asks about a certain expert in the field and then goes and takes a poll from people in the street over which one they think is best. Teoma would

ask all the experts in the field which one of themselves they think is best." By restricting the number of pages used to create the ranking to only pages dealing with the topic of the query, Teoma can often find obscure but relevant sites that Google might have missed.⁸ Teoma calls this idea of counting references only from relevant sites "Subject-Specific PopularitySM."⁷

Teoma also emphasizes a real-time, dynamic search method in contrast to Google's more static system. After the user submits a query, Teoma dynamically looks for communities, more specifically searching for "authorities" in the community. By searching real-time, Teoma can find a community even for new pages.⁹

Features to modify and improve searches also appear in Teoma. After a user enters a query, Teoma not only displays the results but also provides two options to the right of the screen, labeled "refine" and "resources." Refine allows a user to select a specific community of sites developed real-time by Teoma in which to continue the search. This allows a user to further focus his search. The resources option lists sites within a community that contain links to other sites related to the query. This gives the user access to references by authoritative sites on the subject.



So, should users stop using Google and start using Teoma? Maybe not yet. Google contains a much larger index than Teoma, even though Teoma's index has expanded by more than 500 percent to now include more than 500 million links.⁷ Teoma also lacks a cache and advanced searches such as the Boolean search.⁸ By understanding the differences between Teoma's searching techniques and Google's, however, users can decide which will better answer their query.

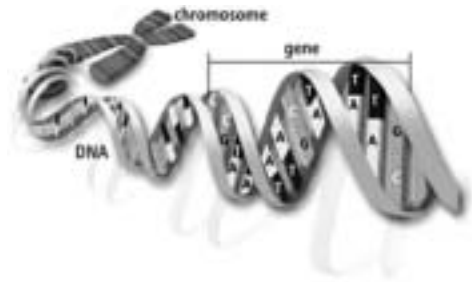
Biology

DNA Microarrays

A key recent development in biology is the sequencing of genomes of different species. A genome consists of all the DNA in an organism, including genes that contain information for making proteins that determine the appearance and functionality of an organism. Even a simple organism such as a bacterium has 1.8 million bases with one gene per thousand bases. The human genome consists of roughly 3 billion DNA base pairs with about 30,000 genes: one gene for every hundred thousand bases. This April, the International Human Genome Sequence Consortium announced the completion of the Human Genome Project, which sequenced each of the DNA base letters in humans with an estimated accuracy of less than one error out of every 10,000 genes sequenced. Now that biologists know the gene sequence of humans and other organisms, they can try to determine gene expressions and coexpressions, that is, traits to which the genes map. The implications of their results will be enormous: Understanding the genes linked with disease could lead to improved understanding and treatment of ailments such as heart disease, diabetes, cancer, and deafness.¹⁰ Already, more than 1,400 genes have been linked to specific diseases.¹¹

How do biologists search for genes linked with specific expressions out of the millions of base pairs in the genome, especially since many sections of the genome appear to have no function?¹² One of the most popular methods to map genes to their expressions is cDNA microarray analysis, which allows simultaneous measurements of several thousands of genes.¹³

To perform a microarray analysis, a biologist first extracts mRNA from a sample. Because mRNA in organisms is later transcribed into proteins, by measuring mRNA, the biologist can indirectly determine protein level. The mRNA is then changed into cDNA, a synthesized, single-strand form of DNA.¹⁴ This cDNA is labeled with fluorescent dyes that bind to a slide with DNA sequences from known genes. The activity level of each gene can be determined by looking at the fluorescence intensity and locations of fluorescence. By simultaneously examining mRNA levels in thousands of genes, scientists can find relationships between subsets of the genes, such as in feedback loops.¹⁵



The staggering number of concurrent measurements, however, causes problems.^{14, 16} As Sylvia Spengler, an expert in biotechnology databases with the Center for Bioinformatics and Computational Genomics at Lawrence National Laboratories writes, "Clearly, it is shortsighted to gather large amounts of data that cannot be analyzed in a timely manner."¹⁷ For experiments to be useful, they must be designed for analyses, and specific analytical methods should be determined before measurements are made.

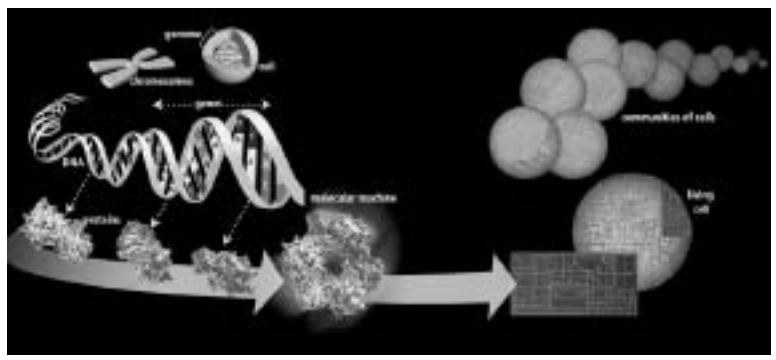
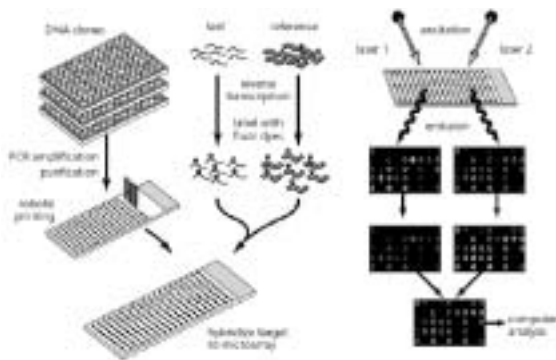
Experimental Design

Many random variables enter into the measuring process. Noise, for instance, can occur at any stage in the process. Within a single slide, temperature and illumination changes, or even dust, can affect measurements. From slide to slide, other variables such as the use of different dyes can vary fluorescence readings. This alone can significantly affect data.^{16, 17} A recent simulation model of cDNA microarrays required more than 20 parameters, each with a probability distribution, to represent factors such as spot geometry and background noise.¹³

Even without noise, statistical difficulties are inherent in the design of a microarray experiment. A good statistical experiment tests several variables with many repetitions. Due to the high cost of gene chips, generally about \$1,000 each, and the scarcity of rare gene samples, repeated measurements typically do not occur. This means that although many conclusions seem significant, they actually occur only by random variation. Commonly, scientists assign the chance to be 5 percent that a variable will appear to be significant when it truly is not (i.e., a "false positive"). This factor practically vanishes in most datasets. When testing 10,000 genes at once, however, 5 percent suddenly becomes 500 "significant" results generated by chance alone. Typically, statisticians perform tests to reduce this kind of false positive; however, many of these "multiple comparison" tests are not designed for the extremely large number of comparisons that microarray analysis presents and are therefore not necessarily applicable.¹⁶ New methods for reducing the number of false positive results are being developed, such as that by Westfall and Young specifically formulated for microarray analysis but applicable to other fields of research as well.¹⁸

Data Analysis

Regardless of these false positives, how do scientists even begin to find significant results? Statisticians and



data analysts are being challenged by this problem. Data is first standardized by a process called normalization to try to reduce noise in the measurements. Then data analysts use pattern recognition methods to look for significant relationships in the data.

One method to find correlations is to use a permutation test. A permutation is a rearrangement of elements in a group, in this case, a random arrangement of the measured microarray data. The permutation test is used to compare the level of significance in the original data to the levels of significance in sets of randomly permuted data: data mixed up in a random order. This random data can then serve as a comparison to the original, nonpermuted data. If the permuted, random data shows a similar level of statistical significance to the original data, then the trends in the original data are reproducible by chance alone and are therefore insignificant. If the original data shows a much stronger statistical significance than the random data, then the statistician can conclude that a true significant pattern exists. To visualize these potential significance patterns in both the original and the permuted data, topological terrain maps can be employed using recently developed software.¹⁹ By using the permutation test, a statistician has a higher level of certainty that his results are truly “significant” than if he looked only at the original data.

Cluster analysis, another technique to study patterns in data, has also been used in analyzing microarray data. Cluster analysis compares data by creating a vector for each gene and each experiment, creating a data matrix. Genes are then “clustered” into groups, and the distances between their expression vectors are measured.¹⁷ Many variations of clustering exist. In hierarchical clustering, clusters are connected in a tree-like progression. In k-means clustering, a specific number of clusters is specified, and data is classified into clusters to minimize distances within a cluster as compared to the distances between clusters. So-called “supervised” methods such as support vector machines (SVM) use labeled data to train a machine to distinguish observational data as members or nonmembers of a group. Used in correlation with these methods are data reduction techniques, such as principal component analysis (PCA), which reduce the dimensionality of data and are useful to determine the number of groups necessary for a cluster analysis such as k-means.²⁰

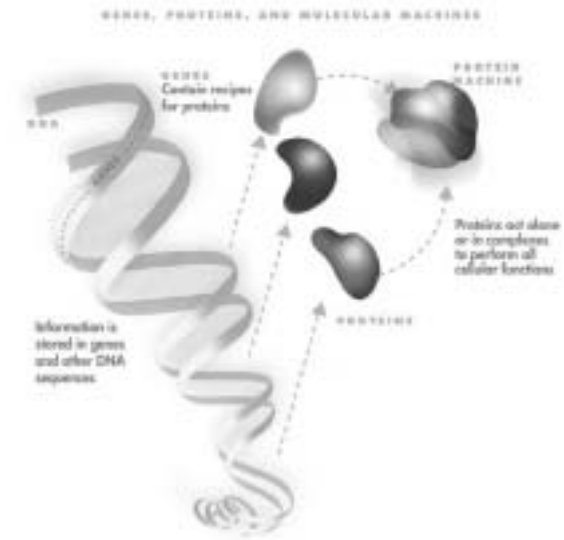
Future of Microarray Analysis

Microarray analysis clearly presents a number of difficulties in both experimental design and analysis stages. Analysis of the data is tricky because, as John Quackenbush with the Institute for Genomic Research writes, “The methods that are used to analyze the data can have a profound influence on the interpretation of the results.”²⁰ New guidelines stressing minimum information about a microarray experiment (MIAME) have been created by the Microarray Gene Expression Data Society. These guide researchers in the areas of repeated measurements and design. Journals are also mandating stricter statistical analyses of data, eliminating “sloppy statistics” that lead to “faulty conclusions” and mandating data to be submitted with papers.¹⁶ Members of the field recognize the difficulties and are working to minimize these problems so that microarray analysis can revolutionize biology, leading to a better understanding of genetic diseases.

Physics

Particle Physics

Particle physics, the study of basic atomic and subatomic elements of matter, uses high-energy particle accelerators to collide and detect particles at high speeds. This results in large volumes of recorded data. The problems associated with sharing these vast datasets resulted in the creation of the World Wide Web in 1990 by Tim





Berners-Lee, a scientist at CERN, the European Laboratory for Particle Physics, to allow physicists all over the world to share data. Even the World Wide Web, though, is not sufficient to handle the amounts of data physicists will soon collect.²¹

Researchers at CERN, located in Geneva, Switzerland, are currently building the Large Hadron Collider (LHC). The LHC will search for the Higgs particle, a currently undetected particle thought to give other particles their masses. In the LHC, collisions will occur at energies of up to 14 TeV, the highest ever; collision numbers could be as high as a billion; the annual volume of data collected will be five petabytes (10 to the 15th power). Though scientists predict 800 million collisions will occur every second in the LHC, Higgs particles are likely to be seen in only 100.²² This amounts to finding a needle in a haystack bigger than any before. Data storage, data accessibility, and data analysis are all problems. Several international initiatives have since been created to handle these problems. The Grid Physics Network addresses the IT problems present using Petascale Virtual Data Grids. The project works on creating systems of software to allow users in locations across the world to analyze data. The project aims to make a data toolkit to aid with this complicated analysis.²³ Another project dealing with similar problems is GIOD.

GIOD

A collaborative project between Caltech, CERN, and the Hewlett Packard Corporation, Globally Interconnected Object Databases (GIOD) deals with information technology and data issues of the LHC, creating an object database with reconstruction, analysis, and visualization tools. The database is complicated by a need to make it accessible to an international group of physicists. The systems created must be tested for scalability with simulated high-energy physics data to ensure correct handling of the many megabyte events.

Since the data in the GIOD project is so complex, an object-oriented system is used to describe them. Object-oriented programming languages such as C++ and Java are focused around creating a system in which data can be represented as “objects” according to their attributes and behavior. Examples of objects can be particle tracks, intersection points of tracks, and particle assignments.²⁴ Commercial object-oriented database management systems are used in GIOD to help sort and move through data quickly. These objects can keep track of the complicated relationships present in the data and help physicists extract data of interest in searches.²⁵

Analysis

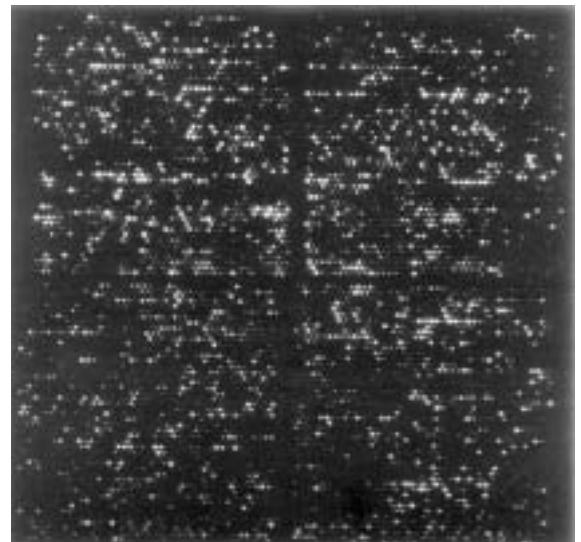
Analysis of particle physics results involves significant data mining. One common method is the use of an artificial neural network (ANN).²⁶ A neural network, also called a neural net, is an algorithm that uses a computer to recognize patterns in a set of training data. This creates a set of criteria by which to predict the classification of other samples. ANNs and other statistical software modified for use by physicists have been created and made available.²⁷ The Stanford Linear Accelerator Center (SLAC), for example, maintains a collection of software for everything from data acquisition to data analysis of high-energy physics experiments.²⁸ By using ANNs and these other highly complicated tools, physicists can search for anomalies or patterns, giving them the possibility of finding a Higgs particle among millions of collisions.

Astronomy and Astrophysics

In order for astronomical experiments to be effective, they must draw data from locations as far as light years away with complicated instruments that make enormous numbers of measurements. Astronomical experiments such as the Laser Interferometer Gravitational-wave Observatory to observe gravitational waves of pulsars and supernovae as well as Sloan Digital Sky Galaxy to create a database of features in the sky, also face problems in data collection and analysis.²³

World-Wide Telescope

A project called the World-Wide Telescope or Virtual Observatory will collect astronomical data and allow access by Internet users. Vast quantities of astronomical data are collected annually from locations such as the Hubble Space Telescope, the Chandra X-Ray Observatory, and the Digitized Palomar Observatory Sky Survey. Because astronomical objects like galaxy clusters can behave differently in different wavebands, observations across the entire sky are necessary. Each year, with instrumentation improvements, more data can be collected. Even a single spectral band from the sky can hold



as much as a few terabytes of information. Each scientist cannot individually keep this data, so astronomers need a digital repository to allow access to the data from numerous locations. Astronomical data is further complicated by the need to calibrate data to the instrument on which it was measured, meaning that data must undergo corrections before it is useful, and the individual groups who measured the data must perform these calculations.

The World-Wide Telescope project will allow access to data from the entire sky for all historical experiments, so that international groups of scientists can analyze data. It is also hoped that this project will help regulate the way in which data is measured, as consistency in terminology, units, representations, technology, and data structures will be required of such a database.²⁹

Data-Mining in the Sky

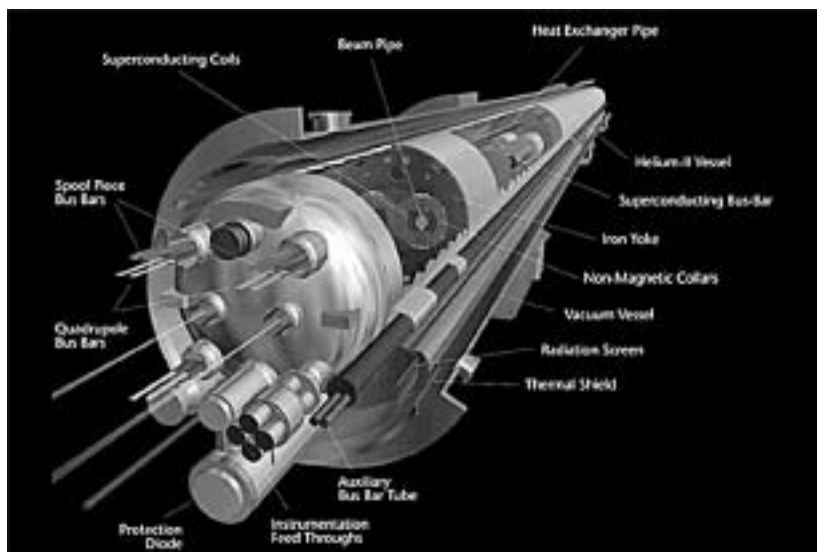
Even with a digital repository of data, searching for specific objects such as quasars or galaxy clusters can be difficult in these terabytes of pixels. Statistically sound methods from machine learning and artificial intelligence must be utilized for these vast searches. Clustering techniques, similar to those in microarray analysis, are used to group objects that belong together. To do this, first the number of clusters must be estimated, usually assuming what statisticians call a normal, Gaussian distribution: the bell curve. One way to determine the number of clusters is to use a Bayesian technique that uses probabilities of previous events, called prior probabilities, to more accurately predict outcomes. To look for patterns and classify data, other methods using these Bayesian prior probabilities techniques can be used. Bayesian inference methods predict the most likely fit for data. Further refinement can be made by adding a penalty term to reflect the bias inherent in the model due to the reliance on prior probabilities.³⁰

Other Examples

Problems of large data sets and data mining occur in many other situations. Detecting anomalies is one general application. Companies can collect information to try to detect fraud, such as in credit card scams. Governments can collect information about millions of individuals to search for patterns that might identify a terrorist. Large databases can also be used to analyze consumer spending patterns, satellite and meteorological data, process control, and even the likelihood of risk for insurance and actuarial purposes.^{2, 17, 27, 31, 32, 33, 34}

Commonalities Across Disciplines

Across all of these disciplines, many of the same problems with these large datasets appear. First, data must be measured and investigated. Often this involves using dimensionality-reduction techniques such as principal component analysis, multidimensional scaling, and new methods such as independent component analysis used in functional magnetic resonance imaging. Data can then be explored using vector clustering techniques.



By assuming items in the same clusters share features, researchers can create and test hypotheses. Models are then derived using learning algorithms. Two kinds of models are typically employed: generative models that capture causal relationships and Bayesian networks that use previous outcomes to predict future results. A model of connected points, called a graph, can also be used, as in the case of the World Wide Web. Next, predictions made by the models can be tested. For this, large amounts of labeled data must be used. Finally, the model is revised, and data is again collected to test the corrected model.³⁵

These steps of collecting and analyzing large amounts of data allow scientists to examine gene expression, particles such as the Higgs boson, astronomical data, and the World Wide Web. Specific analytical techniques are applicable to all these data mining situations, such as dimensionality reduction methods, clustering techniques, and classification algorithms. The techniques used in microarray analysis to find significant trends among thousands of concurrent measurements can be used by other fields searching for significant patterns in their datasets. The databases and software being created by particle physics give others ideas about how to circulate their data among an international group. The ideas of modeling the World Wide Web as a graph can be used for other problems involving complicated graphs with nonobvious connections between elements. A solution in one field is often broadly applicable to many other problems.

Skills for Undergraduates to Handle Large Datasets

With the vast amounts of data aggregated today, what subjects should undergraduates study to prepare themselves for this influx? Data-mining itself represents the boundary of many fields, among them database management, artificial intelligence, machine learning, pattern recognition, and data visualization.

These areas fall under the main subject lines of computer science, engineering, and statistics. Though statistics is generally concerned with analysis of data, statisticians tend to want proof in more rigorous ways than data-mining allows. As a result, statisticians have been left out of the process, and other disciplines have taken over.^{2, 35, 36} Truly, however, to best understand the data, members of all these fields should contribute. The general steps used to analyze large datasets of observation, data reduction, data clustering, hypothesis generation, modeling, and testing of predictions involve all of these subjects.³³ Statistics must be utilized for data reduction techniques like principle component analysis and hypothesis

testing. Computer science disciplines such as artificial intelligence and machine learning are used for data clustering, hypothesis generation, and modeling. Engineering and statistical considerations are frequently taken into account in setting up an experiment prior to data collection. All these disciplines are important for complex datasets, and students would be well advised to be familiar with each. Furthermore, whether biologists or meteorologists, research groups should employ diverse team members such as statisticians and computer scientists to properly handle the data and correctly interpret results. In this way, a world growing increasingly complex every day can be analyzed reliably and accurately. ■

References

1. L. Getoor. "Link Mining: A New Data Mining Challenge." *SIGKDD Explorations*. July 2003; 5(1): 84–89.
2. J. Friedman. "Data Mining and Statistics: What's the Connection?" *Keynote Address of 29th Symposium on the Interface*, Houston, TX, May 1997.
3. O. Etzioni. "The World Wide Web: quagmire or gold mine?" *Communications of ACM* 1996; 39(11): 65–68.
4. S. Brin, L. Page. "The Anatomy of a Large-Scale Hypertextual Web Search Engine." *Proceedings of the Seventh International World Wide Web Conference*, Brisbane, Australia, 1998, Elsevier.
5. www.infosatellite.com/news/2002/04/p030402 teoma.html
6. sp.teoma.com/docs/teoma/about/developmentteam history. html
7. sp.teoma.com/docs/teoma/about/searchwithauthority. html
8. www.theregister.co.uk/content/6/20614.html
9. searchenginewatch.com/searchday/article.php/ 2159601
10. www.ornl.gov/TechResources/Human_Genome/
11. www.genome.gov
12. I. Wickelgren. "Spinning Junk into Gold." *Science* 2003; 300(5626): 1646–1649.
13. Y. Balagurunathan, E.R. Dougherty, Y. Chen, M.L. Bittner, J.M. Trent. "Simulation of cDNA microarrays via a parametrized random signal model." *Journal of Biomedical Optics* Jul 2002; 7(3): 507–523.
14. www.axon.com/mr_ Glossary.html
15. D. Page, M. Graven. "Biological Applications of Multi-Relational Data Mining." *SIGKDD Explorations* July 2003; 5(1): 69–79.
16. C. Tilstone. "Vital Statistics." *Nature* 2003; 424(6949): 610–612.
17. S. Spengler. "Bioinformatics in the Information Age." *Science* 2000; 287(5456): 1221–1223.
18. P. Westfall, S. Young. *Resampling-based Multiple Testing*. John Wiley and Sons: New York (1993).
19. S.K. Kim, J. Lund, M. Kiraly, K. Duke, M. Jiang, J.M. Stuart, A. Eizinger, B.N. Wylie, G.S. Davidson. "A Gene Expression Map for *Caenorhabditis elegans*." *Science* 2001; 293(5537): 2087–2092.
20. J. Quackenbush. "Computational Analysis of Microarray Data." *Nature Reviews Genetics* July 2001; 2: 418–427.
21. public.web.cern.ch/public/about/achievements/www/ www.html
22. J. Bunn, K. Holtman, H. Newman. "Object Database Scalability for Scientific Workloads." Technical Report.
23. www.griphyn.org/projinfo/summary.php
24. http://nscp.upenn.edu/sc95/cdf.html
25. http://pcbunn.cithec.caltech.edu/projdesc/project_descriptions.htm
26. www.eu-crossgrid.org/ physics.htm
27. www.hyperdictionary.com
28. http://heplibw3.slac.stanford.edu/FIND/FHMAIN.html
29. A. Szalay, J. Gray. "The World-Wide Telescope." *Science* 2001; 293(5537): 2037–2040.
30. S.G. Djorgovski, R.R. Calvarlho, S.C. Odewahn, R.R. Gal, J. Roden, P. Storlorz, A. Gray. "Data-Mining a Large Digital Sky Survey: From the Challenges to the Scientific Results." *Applications of Digital Image Processing XX* 1997, ed. A. Tescher, *Proc. S.P.I.E.* 3164: 98–109.
31. M.N. Garofalakis, R. Rastogi, S. Seshadri, K. Shim. "Data Mining and the Web: Past, Present, and Future." *Workshop on Web Information and Data Management*. 1999: 43–47.
32. J. Srivastava, R. Cooley, M. Deshpande, P. Tan. "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data." *SIGKDD Explorations* Jan 2000; 1(2): 12–23.
33. P. Domingos. "Prospects and Challenges for Multi-Relational Data Mining." *SIGKDD Explorations* July 2003; 5(1): 80–83.
34. B. Thuraisingham. "Data Mining, National Security, Privacy, and Civil Liberties." *SIGKDD Explorations* Jan 2003; 4(2): 1–5.
35. E. Mjolsness, D. DeCoste. "Machine Learning for Science: State of the Art and Future Prospects." *Science* 2001; 293(5537): 2051–2055.
36. D. Hand. "Statistics and Data Mining: Intersecting Disciplines." *SIGKDD Explorations* June 1999; 1(1): 16–19.