Welcome to the Session on Data Mining and Search

- Dianne O'Leary
- Chidanand Apte
- Raúl Valdés-Pérez
- Jon Kleinberg
- Laszlo Lovasz

What is data mining / search?

Considerable variation in jargon:

- data mining
- information extraction
- knowledge discovery
- information retrieval
- ...

A Working Definition:

Data mining and search refers to a set of techniques used to automatically find useful information.

A “classic” example:

Project Dendral, Stanford University, 1970’s:
From chemists’ interpretation of many spectra from mass spectrometers and nuclear resonance spectrometers, researchers tried to extract *rules* so that a computer program could interpret spectra automatically.

Today:
There are similar efforts for *automated* rule extraction.

A “legendary” example: Beer and Diapers

A 1992 marketing study: NCR found a high correlation between sales of beer and sales of diapers, between 5 and 7 pm, at Osco Drugs.

Today: There are efforts to discover unexpected correlations across multiple, mixed-media data sources.

Everyday example:

Today:
We might use our favorite search engine to locate sites with information on safety records of trucks.

Future:
We might use our personal assistant to compile all available information on safety records of 1999 Ford trucks.

The field already has considerable mathematical sophistication:

- 1996 workshop at IMA (Minnesota), on the Mathematics of Information Coding, Extraction, and Distribution.

- Considerable use of statistical techniques and optimization technology.

An example of the potential of mathematics:

Problem: text summarization

Best published algorithm: relies on sophisticated natural language processing techniques to extract features and choose summary sentences.

Recent research: A simple linear algebra approach (pivoted QR decomposition) outperforms the best published algorithm.

The Challenge:

We need to extract information from data with

- •Non-uniform encodings (html, postscript, pdf, etc.)
- •Non-uniform formats (text, image, audio, etc.)
- •Non-uniform terminology (conflicting jargon, multiple languages, etc.)
- •Non-uniform indexing conventions.

The Potential:

Progress in knowledge discovery could revolutionize the way we structure

- scientific inquiry.
- commerce.
- world finance.
- medicine.
- food production.

Our Purpose today….

To uncover research opportunities by completing this sentence:

It would cause a revolution in data mining

and search if we could solve the problem of ….