Data Mining

George Karypis
Department of Computer Science
Digital Technology Center
University of Minnesota, Minneapolis, USA.
http://www.cs.umn.edu/~karypis
karypis@cs.umn.edu

Overview

- Data-mining
  - What is it and why do we care?
- Commercial & Scientific Applications
  - What can it do for me?
- Ongoing Research Activities
  - What is George doing?
- From Research to Technology Transfer
  - How can we help you?

Problem

We have lots of data!

We have little information!

We have no knowledge!

What is Data Mining?

Many Definitions…

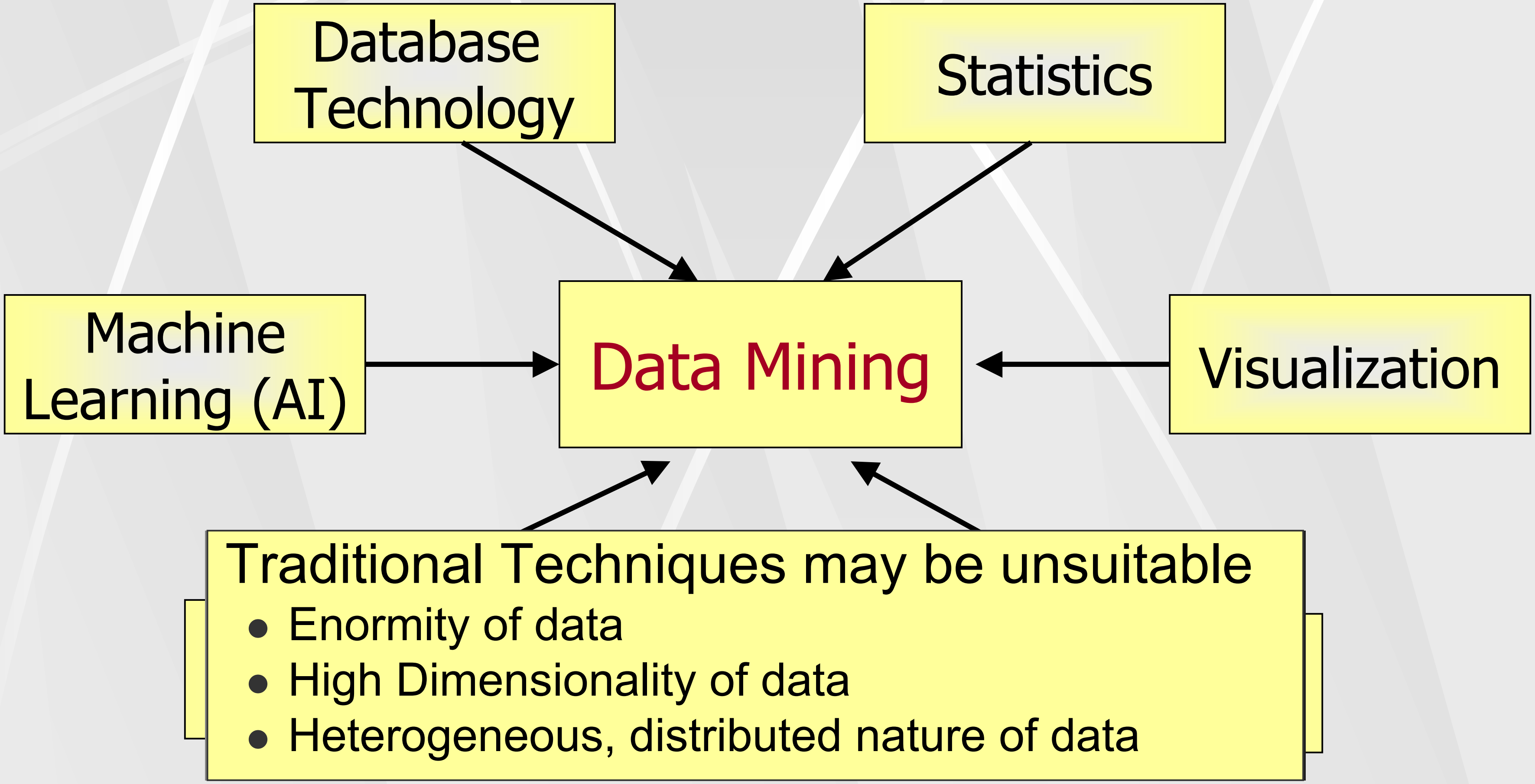
- A short one…

  Search for Valuable Information in Large Volumes of Data.

- A long one…

  Exploration & Analysis, by Automatic or Semi-Automatic Means, of Large Quantities of Data in order to Discover Meaningful Patterns & Rules.

Origins of Data Mining



Database Technology

Statistics

Machine Learning (AI)

Data Mining

Visualization

Traditional Techniques may be unsuitable

- Enormity of data
- High Dimensionality of data
- Heterogeneous, distributed nature of data

A Brief History of Data Mining Activities

- 1989 IJCAI Workshop on Knowledge Discovery in Databases
  - Knowledge Discovery in Databases (G. Piatetsky-Shapiro and W. Frawley, 1991)
- 1991-1994 Workshops on Knowledge Discovery in Databases
  - Advances in Knowledge Discovery and Data Mining (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1996)
- 1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD'95-98)
  - Journal of Data Mining and Knowledge Discovery (1997)
- 1998 ACM SIGKDD, SIGKDD'1999-2003 conferences, and SIGKDD Explorations
- More conferences on data mining
  - PAKDD, PKDD, SIAM-Data Mining, (IEEE) ICDM, DaWaK, SPIE-DM, etc.

Why Mine Data? Why Now?

- Lots of data is being produced, collected, and warehoused.
- Computing has become affordable.
- Competitive pressures are strong
  - Provide better, customized services for an *edge*.
  - Information is becoming *product* in its own right.
- Data mining has become an integral part of modern CRM.

Data Mining Tasks

- Predictive Tasks
  - Use some variables to predict unknown or future values of other variables.
- Descriptive Tasks
  - Find human-interpretable patterns that describe the data.

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

What Problems Can Data Mining Solve?

- Better/Effective/Personalized/Real-time Marketing
  - Identify the subset of customers that is most likely to buy products from a particular catalog.
  - Create a product catalog that will lead to the highest profit.
  - Identify which of the users that browse my website are most likely to purchase something.
- Fraud/Anomaly Detection
  - Identify a fraudulent credit card transaction given the past transactions of a particular customer.
- Customer Attrition/Churn
  - Identify my customers that are most likely to be lost to a competitor.
  - Identify any actions that I can take in order to retain them.
- Effective Information Filtering/Compression/Navigation
  - Find today’s news articles that I will like to read.
  - Help me find what I’m looking on the web.
  - Please organize my hard-drive (mail folders, bookmarks, contacts, etc).

Ongoing Research Activities

- Customer segmentation
- Marketing campaigns
- Recommender systems
- Document categorization & clustering
- Meta-search engines
- Analysis of web-browsing behavior
- Discovery of complex patterns
- Finding patterns in relational data
- Mining scientific and biological databases

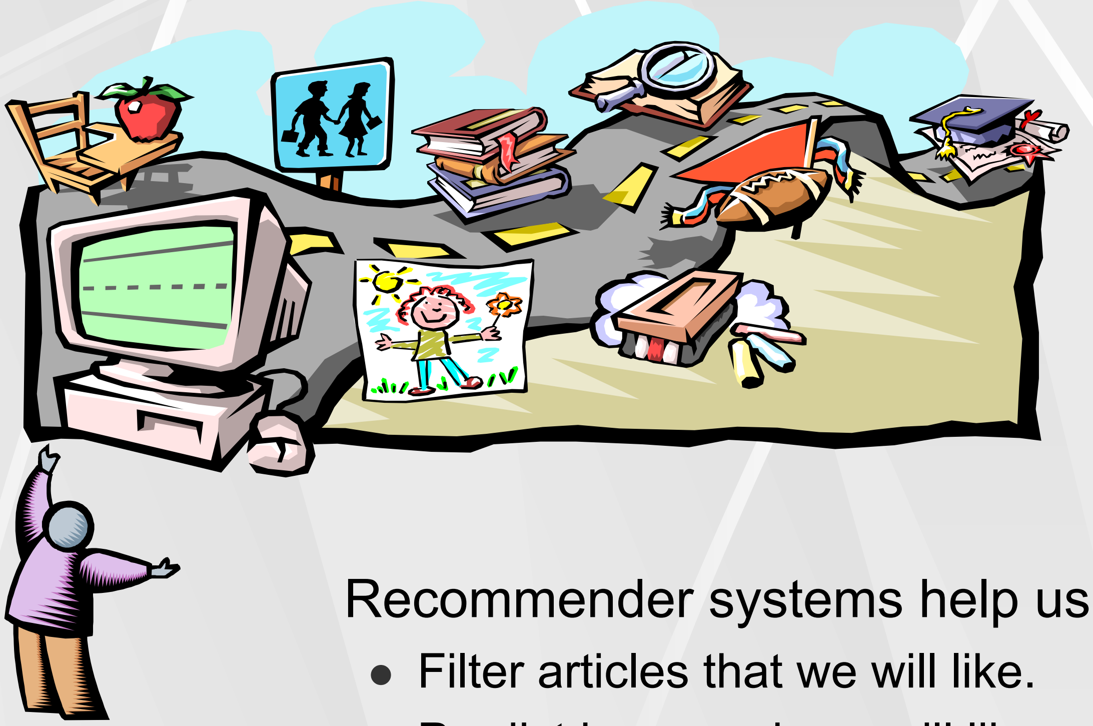
Marketing Campaigns

- Problem:
  - Identify the set of customers that are most likely to buy a set of products.
- Solution:
  - Developed prediction algorithms based on past purchasing and demographic information.

- Problem
  - Create a set of product catalogs and for each catalog identify the subset of customers that it should be mailed to.
- Solution:
  - Developed algorithms to cluster the customers based on predicted future purchases and create the catalogs by analyzing the characteristics of each cluster.

Recommender Systems

- Problem: Information Overload!



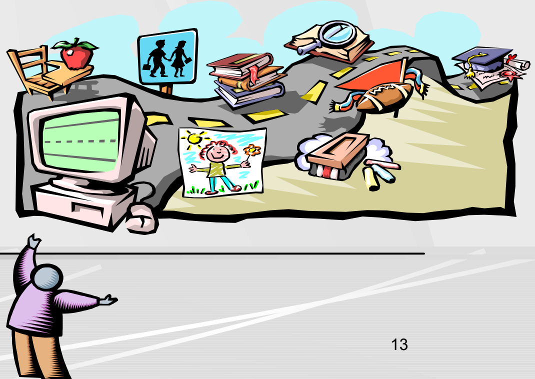
Recommender systems help us identify worthwhile stuff!

- Filter articles that we will like.
- Predict how much we will like a particular book or movie.
- Recommend the top-*N* products that we will most likely buy.

Recommender Systems (cont)

- Developed *scalable* item-based collaborative filtering-based approaches for rating prediction and top-*N* recommendations.

- Collaborative Filtering
  - Key insight—No person is an island!
    - Each individual is a member of a group(s) &
    - The group's collective knowledge can help filter the information!



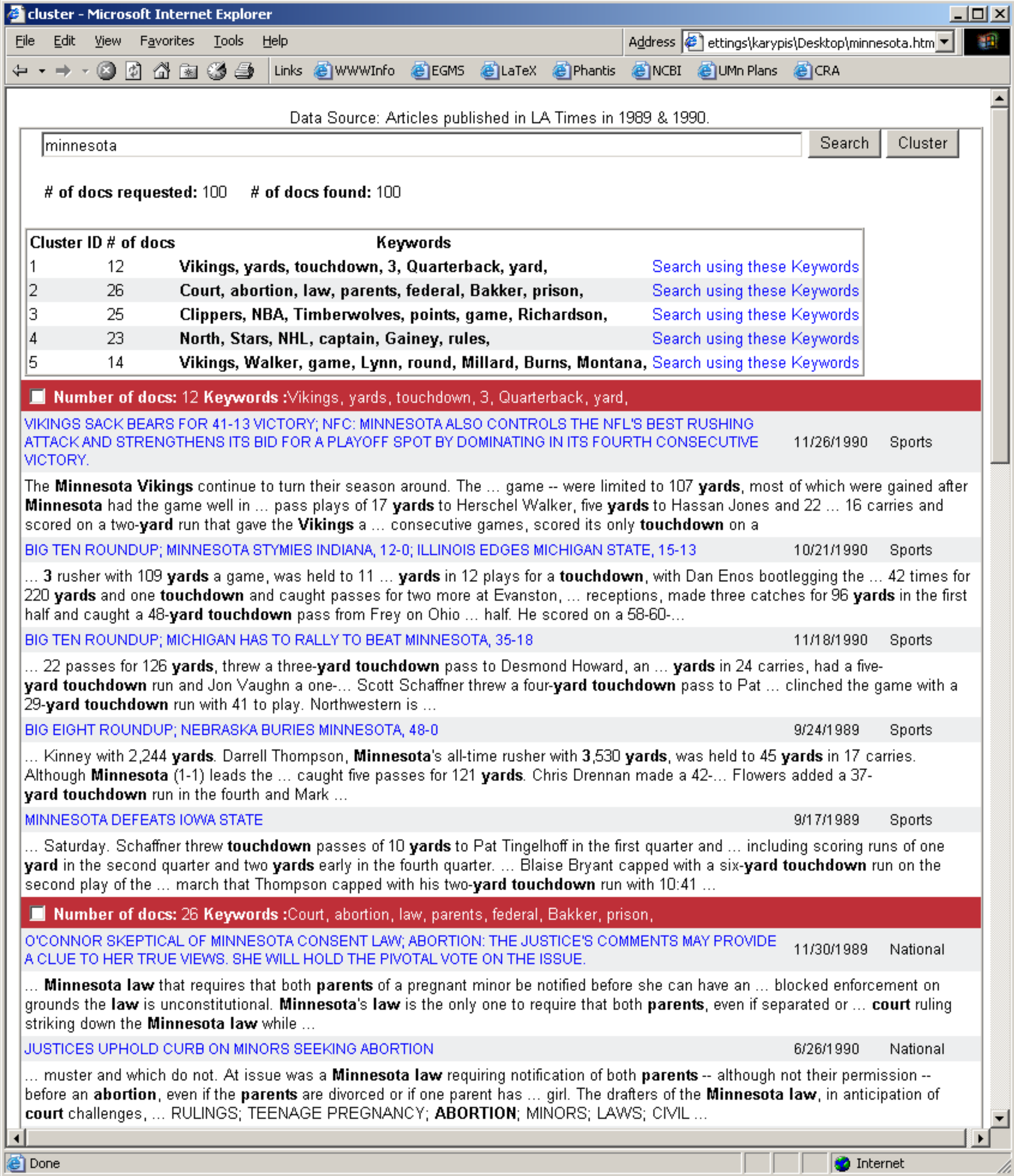
Document Clustering

- Problem:
  - As the amount of textual information increases, there is a need to automatically organize them into meaningful groups and hierarchies, and provide effective summaries.
    - How do you navigate through the 1000+ results that comes back from a Google query?
- Solution:
  - Document Clustering!

Document Clustering (cont)

Developed scalable document clustering algorithms and effective cluster summarization approaches.

Search to Browse…


cluster - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address ettings\karypis\Desktop\minnesota.htm
Links WWWInfo EGMS LaTeX Phantis NCBI UMn Plans CRA
Search
Cluster
Done
Internet

Data Source: Articles published in LA Times in 1989 & 1990.

minnesota

# of docs requested: 100 **# of docs found:** 100

Cluster ID	# of docs	Keywords	
1	12	**Vikings, yards, touchdown, 3, Quarterback, yard,**	Search using these Keywords
2	26	**Court, abortion, law, parents, federal, Bakker, prison,**	Search using these Keywords
3	25	**Clippers, NBA, Timberwolves, points, game, Richardson,**	Search using these Keywords
4	23	**North, Stars, NHL, captain, Gainey, rules,**	Search using these Keywords
5	14	**Vikings, Walker, game, Lynn, round, Millard, Burns, Montana,**	Search using these Keywords

☐ **Number of docs:** 12 **Keywords :**Vikings, yards, touchdown, 3, Quarterback, yard,

VIKINGS SACK BEARS FOR 41-13 VICTORY; NFC: MINNESOTA ALSO CONTROLS THE NFL'S BEST RUSHING ATTACK AND STRENGTHENS ITS BID FOR A PLAYOFF SPOT BY DOMINATING IN ITS FOURTH CONSECUTIVE VICTORY. 11/26/1990 Sports

The **Minnesota Vikings** continue to turn their season around. The ... game -- were limited to 107 **yards**, most of which were gained after **Minnesota** had the game well in ... pass plays of 17 **yards** to Herschel Walker, five **yards** to Hassan Jones and 22 ... 16 carries and scored on a two-**yard** run that gave the **Vikings** a ... consecutive games, scored its only **touchdown** on a

BIG TEN ROUNDUP; MINNESOTA STYMIES INDIANA, 12-0; ILLINOIS EDGES MICHIGAN STATE, 15-13 10/21/1990 Sports

... **3** rusher with 109 **yards** a game, was held to 11 ... **yards** in 12 plays for a **touchdown**, with Dan Enos bootlegging the ... 42 times for 220 **yards** and one **touchdown** and caught passes for two more at Evanston, ... receptions, made three catches for 96 **yards** in the first half and caught a 48-**yard touchdown** pass from Frey on Ohio ... half. He scored on a 58-60-...

BIG TEN ROUNDUP; MICHIGAN HAS TO RALLY TO BEAT MINNESOTA, 35-18 11/18/1990 Sports

... 22 passes for 126 **yards**, threw a three-**yard touchdown** pass to Desmond Howard, an ... **yards** in 24 carries, had a five-**yard touchdown** run and Jon Vaughn a one-... Scott Schaffner threw a four-**yard touchdown** pass to Pat ... clinched the game with a 29-**yard touchdown** run with 41 to play. Northwestern is ...

BIG EIGHT ROUNDUP; NEBRASKA BURIES MINNESOTA, 48-0 9/24/1989 Sports

... Kinney with 2,244 **yards**. Darrell Thompson, **Minnesota**'s all-time rusher with **3**,530 **yards**, was held to 45 **yards** in 17 carries. Although **Minnesota** (1-1) leads the ... caught five passes for 121 **yards**. Chris Drennan made a 42-... Flowers added a 37-**yard touchdown** run in the fourth and Mark ...

MINNESOTA DEFEATS IOWA STATE 9/17/1989 Sports

... Saturday. Schaffner threw **touchdown** passes of 10 **yards** to Pat Tingelhoff in the first quarter and ... including scoring runs of one **yard** in the second quarter and two **yards** early in the fourth quarter. ... Blaise Bryant capped with a six-**yard touchdown** run on the second play of the ... march that Thompson capped with his two-**yard touchdown** run with 10:41 ...

☐ **Number of docs:** 26 **Keywords :**Court, abortion, law, parents, federal, Bakker, prison,

O'CONNOR SKEPTICAL OF MINNESOTA CONSENT LAW; ABORTION: THE JUSTICE'S COMMENTS MAY PROVIDE A CLUE TO HER TRUE VIEWS. SHE WILL HOLD THE PIVOTAL VOTE ON THE ISSUE. 11/30/1989 National

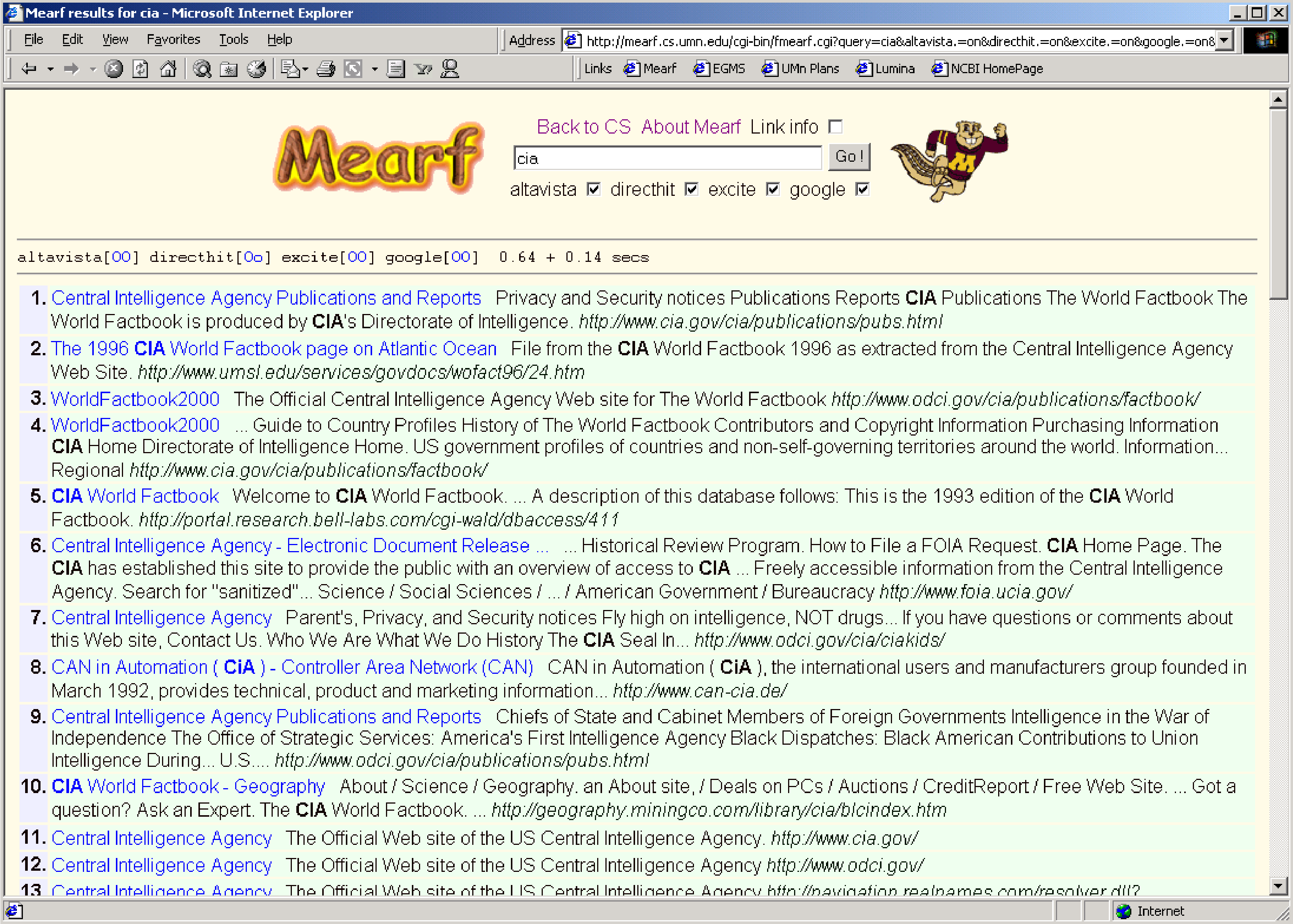
... **Minnesota law** that requires that both **parents** of a pregnant minor be notified before she can have an ... blocked enforcement on grounds the **law** is unconstitutional. **Minnesota**'s **law** is the only one to require that both **parents**, even if separated or ... **court** ruling striking down the **Minnesota law** while ...

JUSTICES UPHOLD CURB ON MINORS SEEKING ABORTION 6/26/1990 National

... muster and which do not. At issue was a **Minnesota law** requiring notification of both **parents** -- although not their permission -- before an **abortion**, even if the **parents** are divorced or if one parent has ... girl. The drafters of the **Minnesota law**, in anticipation of **court** challenges, ... RULINGS; TEENAGE PREGNANCY; **ABORTION**; MINORS; LAWS; CIVIL ...

Meta-Search Engines

- P
- S



Mearf results for cia - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://mearf.cs.umn.edu/cgi-bin/fmearf.cgi?query=cia&altavista.=on&directhit.=on&excite.=on&google.=on&
Links Mearf EGMS UMn Plans Lumina NCBI HomePage
Mearf
Back to CS About Mearf Link info □
Go !
Internet

cia

altavista ☑ directhit ☑ excite ☑ google ☑

altavista[00] directhit[0o] excite[00] google[00] 0.64 + 0.14 secs

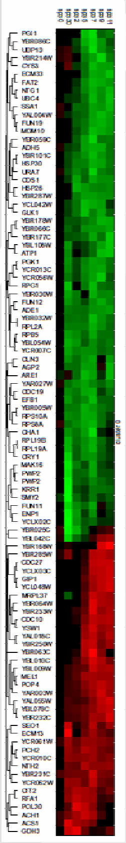
1. Central Intelligence Agency Publications and Reports Privacy and Security notices Publications Reports **CIA** Publications The World Factbook The World Factbook is produced by **CIA**'s Directorate of Intelligence. *http://www.cia.gov/cia/publications/pubs.html*
2. The 1996 **CIA** World Factbook page on Atlantic Ocean File from the **CIA** World Factbook 1996 as extracted from the Central Intelligence Agency Web Site. *http://www.umsl.edu/services/govdocs/wofact96/24.htm*
3. WorldFactbook2000 The Official Central Intelligence Agency Web site for The World Factbook *http://www.odci.gov/cia/publications/factbook/*
4. WorldFactbook2000 ... Guide to Country Profiles History of The World Factbook Contributors and Copyright Information Purchasing Information **CIA** Home Directorate of Intelligence Home. US government profiles of countries and non-self-governing territories around the world. Information... Regional *http://www.cia.gov/cia/publications/factbook/*
5. **CIA** World Factbook Welcome to **CIA** World Factbook. ... A description of this database follows: This is the 1993 edition of the **CIA** World Factbook. *http://portal.research.bell-labs.com/cgi-wald/dbaccess/411*
6. Central Intelligence Agency - Electronic Document Release ... ... Historical Review Program. How to File a FOIA Request. **CIA** Home Page. The **CIA** has established this site to provide the public with an overview of access to **CIA** ... Freely accessible information from the Central Intelligence Agency. Search for "sanitized"... Science / Social Sciences / ... / American Government / Bureaucracy *http://www.foia.ucia.gov/*
7. Central Intelligence Agency Parent's, Privacy, and Security notices Fly high on intelligence, NOT drugs... If you have questions or comments about this Web site, Contact Us. Who We Are What We Do History The **CIA** Seal In... *http://www.odci.gov/cia/ciakids/*
8. CAN in Automation (**CiA**) - Controller Area Network (CAN) CAN in Automation (**CiA**), the international users and manufacturers group founded in March 1992, provides technical, product and marketing information... *http://www.can-cia.de/*
9. Central Intelligence Agency Publications and Reports Chiefs of State and Cabinet Members of Foreign Governments Intelligence in the War of Independence The Office of Strategic Services: America's First Intelligence Agency Black Dispatches: Black American Contributions to Union Intelligence During... U.S.... *http://www.odci.gov/cia/publications/pubs.html*
10. **CIA** World Factbook - Geography About / Science / Geography. an About site, / Deals on PCs / Auctions / CreditReport / Free Web Site. ... Got a question? Ask an Expert. The **CIA** World Factbook. ... *http://geography.miningco.com/library/cia/blcindex.htm*
11. Central Intelligence Agency The Official Web site of the US Central Intelligence Agency. *http://www.cia.gov/*
12. Central Intelligence Agency The Official Web site of the US Central Intelligence Agency *http://www.odci.gov/*
13. Central Intelligence Agency The Official Web site of the US Central Intelligence Agency *http://navigation.realnames.com/resolver.dll?*

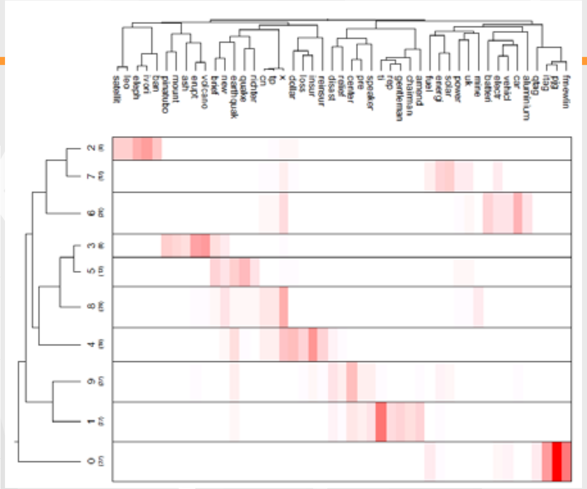
From Research to Technology Transfer

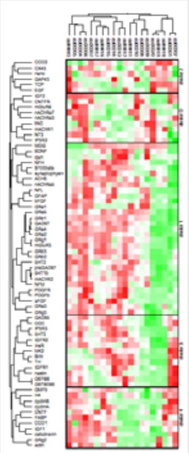
- We put a considerable effort to provide industrial-strength implementations of the various algorithms that we are developing.
- Currently available software tools:
  - CLUTO for clustering
    - http://www.cs.umn.edu/~cluto
  - PAFI for frequent pattern discovery
    - http://www.cs.umn.edu/~pafi
  - SUGGEST for top-*N* recommendation
    - http://www.cs.umn.edu/~karypis/suggest
- These tools are used extensively by various academic, government, and commercial entities.

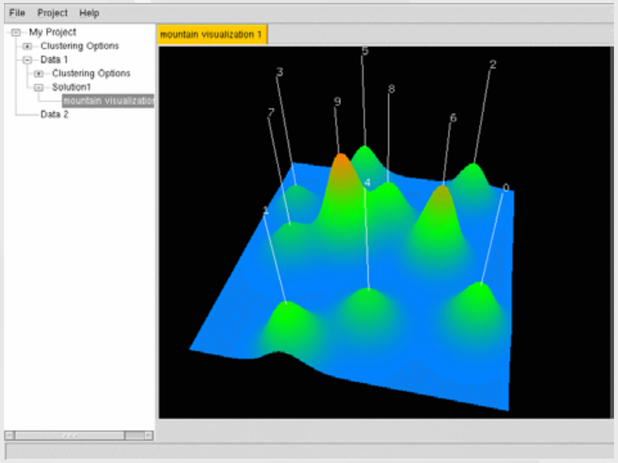
CLUTO

- Clustering Algorithms
  - High-performance & High-quality partitional clustering
  - High-quality agglomerative clustering
  - High-quality graph-partitioning-based clustering
  - Hybrid partitional & agglomerative algorithms for building trees for very large datasets.
- Cluster Analysis Tools
  - Cluster signature identification
  - Cluster organization identification
- Visualization Tools
  - Hierarchical Trees
  - High-dimensional datasets
  - Cluster relations
- Interfaces
  - Stand-alone programs
  - Library with a fully published API
- Available on Windows, Sun, and Linux









http://www.cs.umn.edu/~cluto

Thank you

George Karypis
Department of Computer Science
Digital Technology Center
University of Minnesota, Minneapolis, USA.
http://www.cs.umn.edu/~karypis
karypis@cs.umn.edu