Mining Scientific Data: Past, Present, and Future

Vipin Kumar
University of Minnesota

kumar@cs.umn.edu
www.cs.umn.edu/~kumar

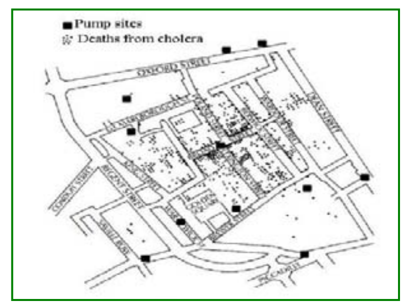
Research Supported by NSF, NOAA, NASA, Planetary Skin Institute, IBM, BICB-UMR, Pfizer, Minnesota Futures Program

Large-scale Data is Everywhere!

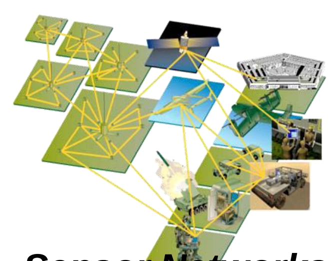
- There has been enormous data growth in both commercial and scientific databases due to advances in data generation and collection technologies
- New mantra
  - Gather whatever data you can whenever and wherever possible.
- Expectations
  - Gathered data will have value either for the purpose collected or for a purpose not envisioned.

Homeland Security

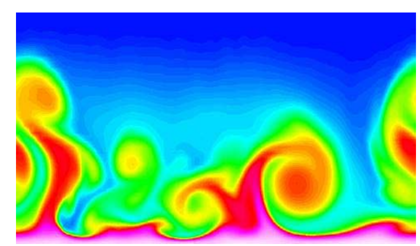
Business Data



Geo-spatial data



Sensor Networks



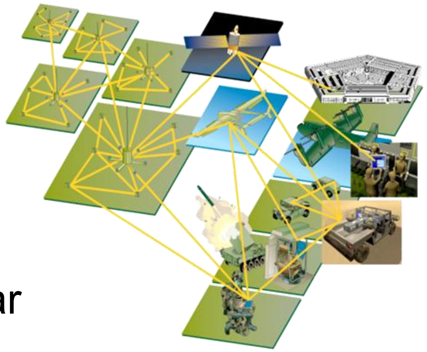
Computational Simulations

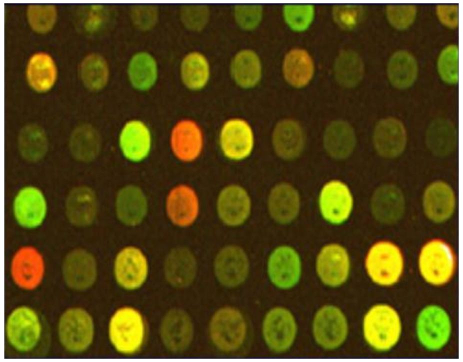
Why Data Mining? Commercial Viewpoint

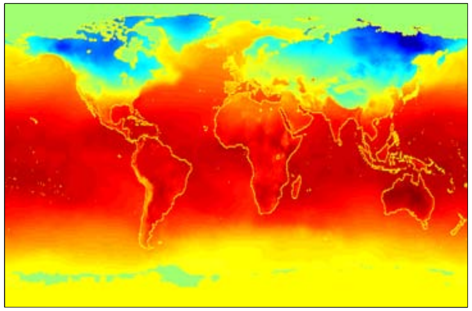
- Lots of data is being collected and warehoused
  - Web data
    - Yahoo has 2PB web data
    - Facebook has 400M active users
  - purchases at department/ grocery stores, e-commerce
    - Amazon records 2M items/day
  - Bank/Credit Card transactions
- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
  - Provide better, customized services for an edge (e.g. in Customer Relationship Management)

Why Data Mining? Scientific Viewpoint

- Data collected and stored at enormous speeds
  - remote sensors on a satellite
    - NASA EOSDIS archives over 1-petabytes of earth science data / year
  - telescopes scanning the skies
    - Sky survey data
  - High-throughput biological data
  - scientific simulations
    - terabytes of data generated in a few hours







- Data mining helps scientists
  - in automated analysis of massive datasets
  - In hypothesis formation

Why Data Mining? Scientific Viewpoint

- Data guided discovery - A new scientific paradigm ?

WIRED MAGAZINE: 16.07

SCIENCE : DISCOVERIES

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

By Chris Anderson 06.23.08

Mining Scientific Data - History

- 1989 : IJCAI Workshop on Knowledge Discovery in Databases
- 1991-1994 : Workshops on Knowledge Discovery in Databases
  - 1995 : First KDD Conference
- 1999-2000 : AHPRC Workshops on Mining Scientific Data
- 2001: SIAM First International Conference in Data Mining
- Past decade has seen a huge growth of interest in mining data in a variety of scientific domains
  - Social Informatics
  - Ecoinformatics
  - Geoinformatics
  - Chemo Informatics
  - Astroinformatics
  - Neuroinformatics
  - Quantum Informatics
  - Health Informatics
  - Evolutionary Informatics
  - Veterinary Informatics
  - Organizational Informatics
  - Pharmacy Informatics

Astronomy

siam® Society for Industrial and Applied Mathematics

SIAM NEWS >

Mining the Sky: Data Analysis Meets Astronomy

April 3, 2002

Chandrika Kamath

Vast amounts of data collected in astronomical surveys is increasingly being analyzed by data mining methods.

- Szalay et al 2001, Burl et al 1998, Kamath et al 2001, Odewahn et al1992

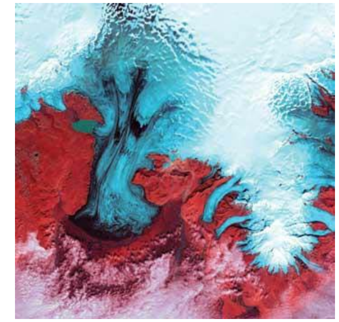
Illustrative Applications

Mining Climate and Eco-system Data

NASA News Archive

July 8, 2003

NASA DATA MINING REVEALS A NEW HISTORY OF NATURAL DISASTERS

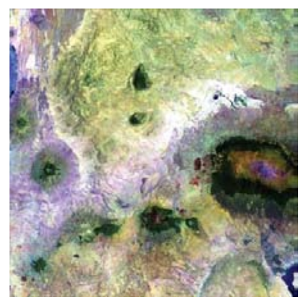


Satellite images and data, such as the one above of the Vatnajökull Glacier in Iceland, assist researchers in tracking teleconnections.

FEATURES | DEPARTMENTS | MARKETPLACE | NEWS UPDATE

ON THE COVER: *Data Mining*

mining what others miss



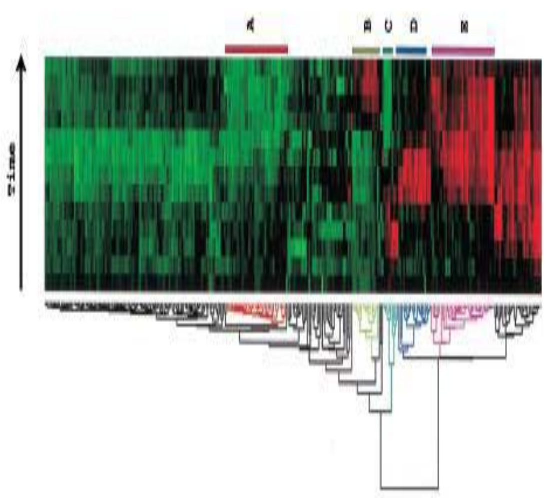
NASA scientists are using satellite data gathered from remote locations (such as Kilimanjaro) to discover changes in the global climate system.

Researchers are using data mining to track the impact of natural disasters, like hurricanes (above), on the global carbon cycle.

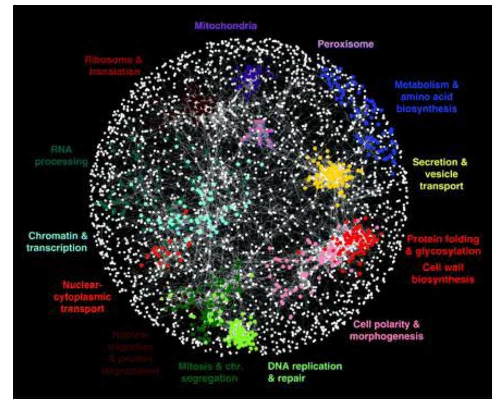
Biological Sciences

Cluster analysis and display of genome-wide expression patterns

MICHAEL B. EISEN*, PAUL T. SPELLMAN*, PATRICK O. BROWN†, AND DAVID BOTSTEIN*‡



Gene expression data



The Genetic Landscape of a Cell
Costanzo et al.
Science 22 January 2010: 425-431
DOI: 10.1126/science.1180823

nature REVIEWS GENETICS

Research Highlight
Bioinformatics: Mining gene expression data
Mark Patterson

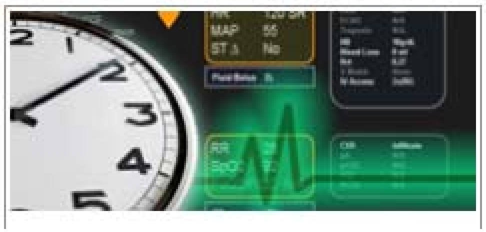
Health Sciences

MAYO CLINIC

Discovery's Edge

Mayo Clinic's Online Research Magazine

Data Mining to Redesign Critical Care Services



When President Barack Obama cites Mayo Clinic as a model healthcare provider, he praises its "smart" practices that offer patients the best possible care at below-normal cost. Mayo's expertise in treating disease is well-known. But the presidential accolades underscore Mayo's pioneer work in an emerging science of healthcare delivery.

Data Mining PhD student Rohit Gupta was selected to present his work on "Colorectal cancer despite colonoscopy" in the clinical science plenary session in DDW 2009, an international conference on gastroenterology recently held in Chicago and attended by more than 15,000 GI professionals.

Illustrative Applications

Social and Political Sciences

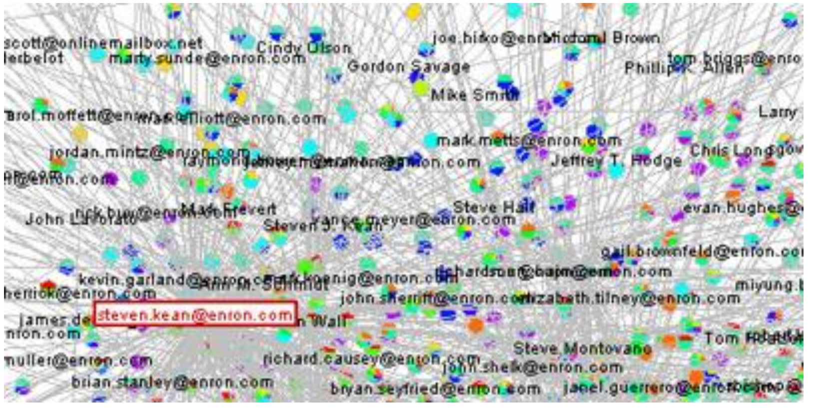
NewScientist Science in Society

Algorithm detects Canadian politicians' spin

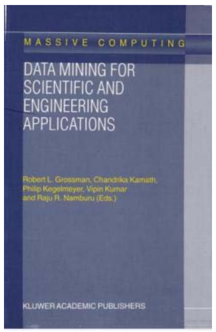
› 16:01 20 January 2006 by Stu Hutson

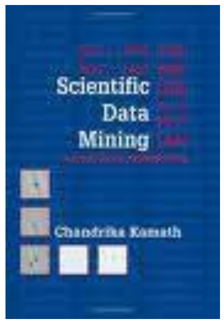
Skillicorn and his team analysed the usage patterns of 88 deception-linked words within the text of recent campaign speeches from the political leaders.

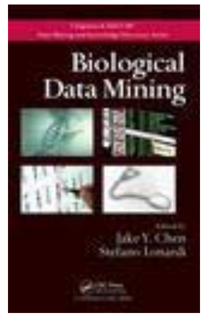
Enron email dataset

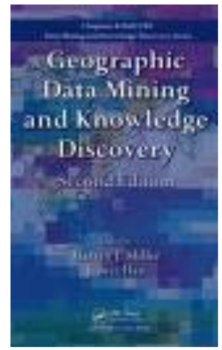


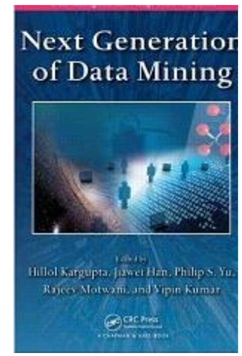
Sample of Books on Mining Scientific Data

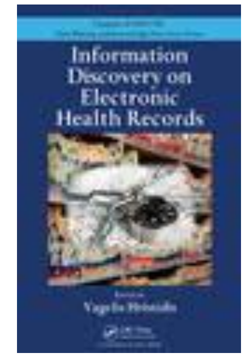


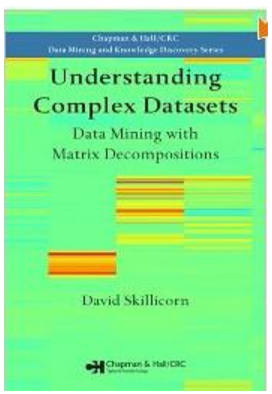


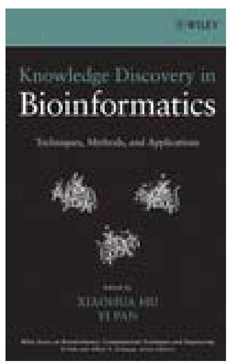


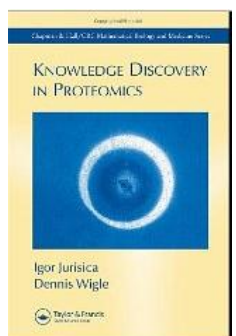


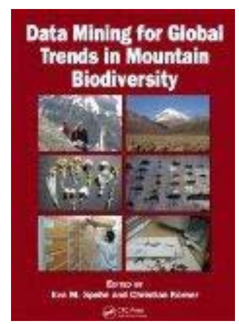


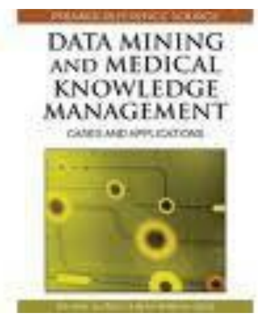








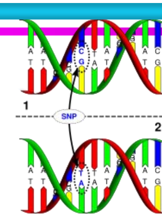


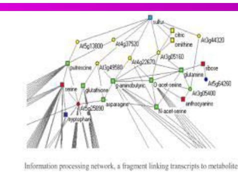


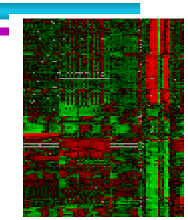
Data Mining for Biomedical Informatics

- Recent technological advances are helping to generate large amounts of clinical and genomic data
  - Biological data sets
    - Gene & protein sequences; Microarray data; Single Nucleotides Polymorphisms (SNPs); Biological networks; Proteomic data; Metabolomics data
  - Electronic Medical Records (EMRs)
    - IBM-Mayo partnership has created a DB of over 6 million patients

- Data mining offers potential solution for analysis of this large-scale biomedical data
  - Novel associations between genotypes and phenotypes
  - Biomarker discovery for complex diseases
  - Prediction of the functions of anonymous genes
  - Personalized Medicine – Automated analysis of patients history for customized treatment

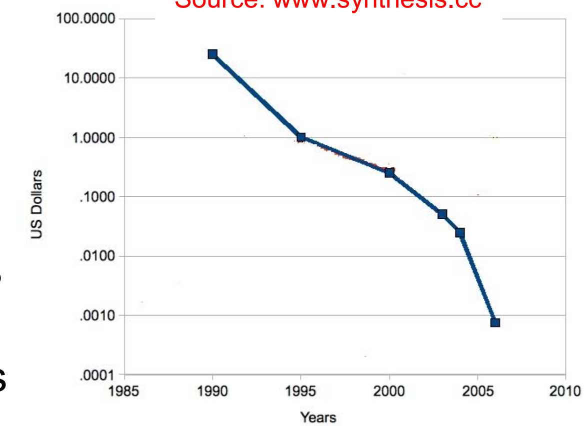

SNP


PPI Data

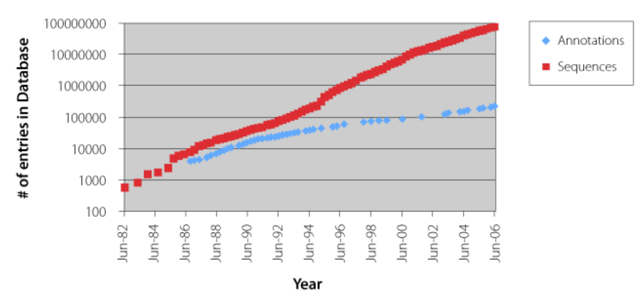

Gene Expression Data

Cost of sequencing has reduced dramatically

Source: www.synthesis.cc



Growth of sequences and annotations since 1982



Increasing gap between genome sequences and functional annotations [Meyers August 2006]

Challenges in Analyzing Biomedical Data

- High dimensionality in the number of attributes (genes, SNPs) and relatively low sample size
  - Difficult to find statistically significant results
    - e.g., associations between gene(s) and disease phenotype
- Heterogeneous data
  - Structured and unstructured data elements, different types of data attributes
    - e.g, gene expression data, networks and pathways, lab tests and pathology reports
- Data is noisy, error-prone and has missing values
  - Difficult to discover true structure due to poor data quality
- Different biological data types provide complimentary but limited information
  - Need to develop approaches that integrates multiple data sets

Case studies

1. Discovering novel associations among SNPs and disease phenotypes
   - Addressing issue of high dimensionality

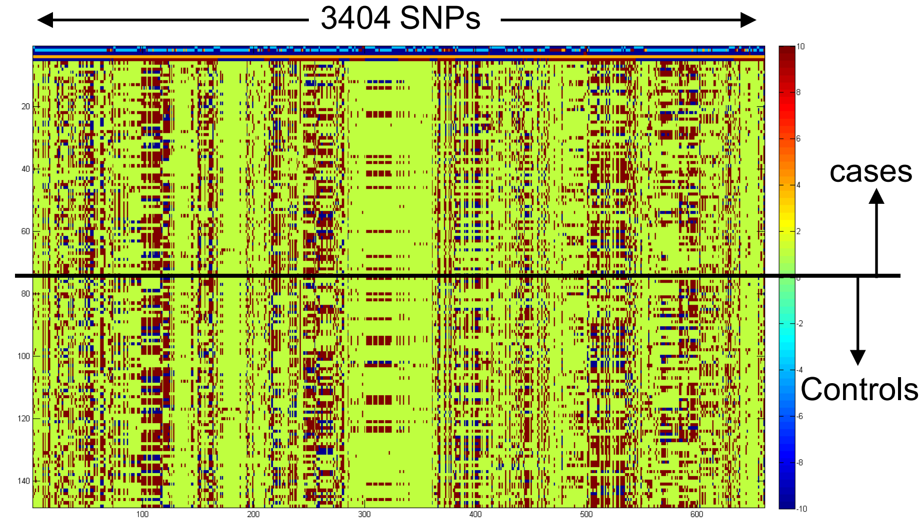
2. Subspace differential co-expression analysis for discovering disease subtypes
   - Addressing the issue of high dimensionality and genetic heterogeneity

3. Error-tolerant pattern mining based biomarker discovery for breast cancer metastasis
   - Addressing issue of data noise

1. Modeling functional inter-relationship among gene annotations for improving protein function prediction
   - Addressing complex functional annotation structure of biological entities

Case Study 1:
Discovering SNP Biomarkers

- Given a SNP data set of Myeloma patients, find a combination of SNPs that best predicts survival.



  - 3404 SNPs selected from various regions of the chromosome
  - 70 cases (Patients survived shorter than 1 year)
  - 73 Controls (Patients survived longer than 3 years)

Complexity of the Problem:

•Large number of SNPs (over a million in GWA studies) and small sample size

•Complex interaction among genes may be responsible for the phenotype

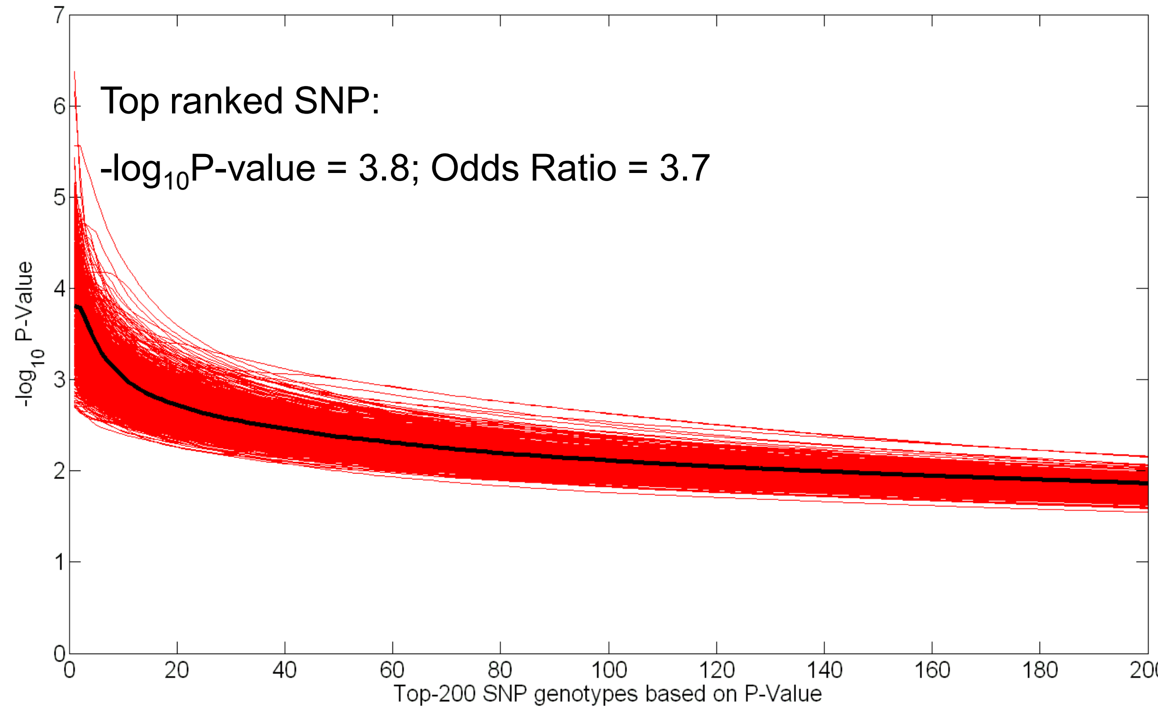
•Genetic heterogeneity among individuals sharing the same phenotype (due to environmental exposure, food habits, etc) adds more variability

•Complex phenotype definition (eg. survival)

Case Study 1:

Issues with Traditional Methods

- Each SNP is tested and ranked individually
- Individual SNP associations with true phenotype are not distinguishable from random permutation of phenotype



Van Ness et al 2009

A comprehensive review of genetic association studies.
by: Joel N. Hirschhorn, Kirk Lohmueller, Edward Byrne, Kurt Hirschhorn

Genetics in medicine, Vol. 4, No. 2. (r 2002), pp. 45-61.

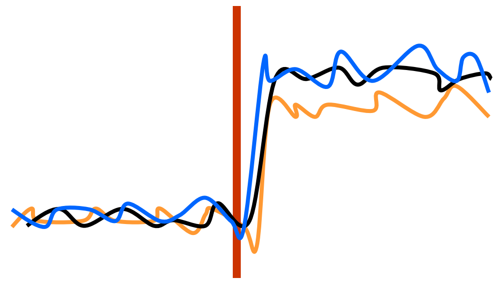
However, most reported associations are not robust: of the 166 putative associations which have been studied three or more times, only 6 have been consistently replicated.

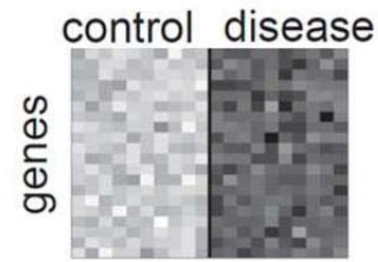
Case Study 2:

Discovering Multi-Gene Biomarkers

- Differential Expression (DE)
  - Traditional analysis targets changes of expression level

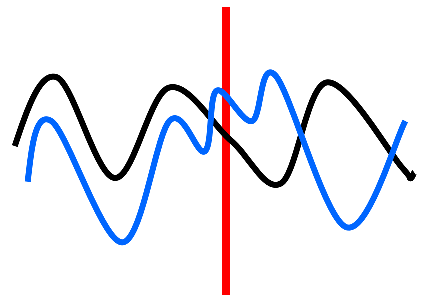
[Silva et al., 1995], [Li, 2002], [Kostka & Spang, 2005], [Rosemary et al., 2008], [Cho et al. 2009] etc.

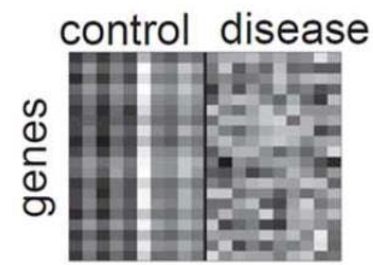




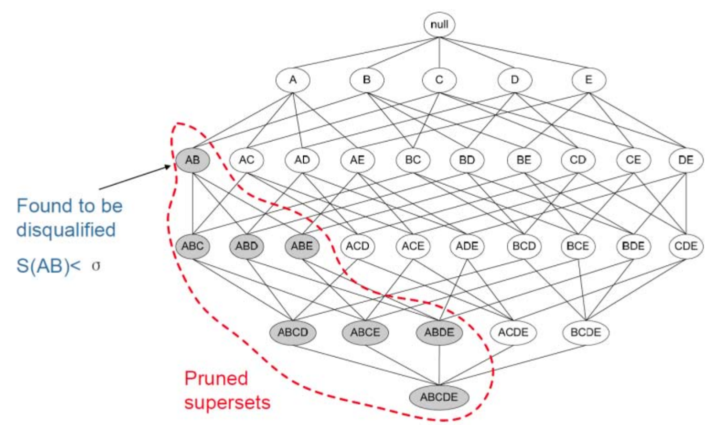
- Differential Coexpression (DC)
  - Changes of the coherence of gene expression

[Eisen et al. 1999] [Golub et al., 1999], [Pan 2002], [Cui and Churchill, 2003] etc.





- Combinatorial Search
- Genetic Heterogeneity
  - calls for subspace analysis

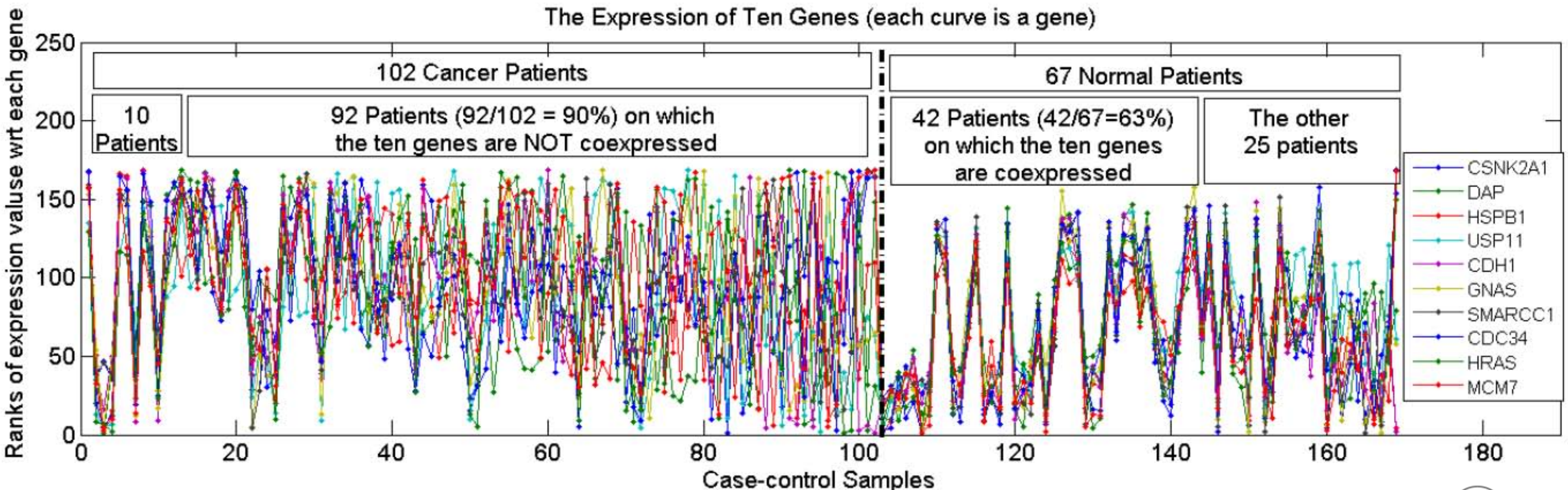


Case Study 2:

Discovering Multi-Gene Biomarkers

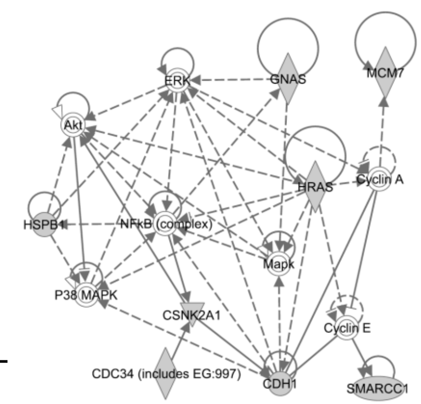
- An Example Subspace Differential Coexpression Pattern from lung cancer dataset

Three lung cancer datasets [Bhattacharjee et al. 2001], [Stearman et al. 2005], [Su et al. 2007]



Enriched with the TNF/NFB signaling pathway which is well-known to be related to lung cancer

P-value: $1.4*10^{-5}$ (6/10 overlap with the pathway)

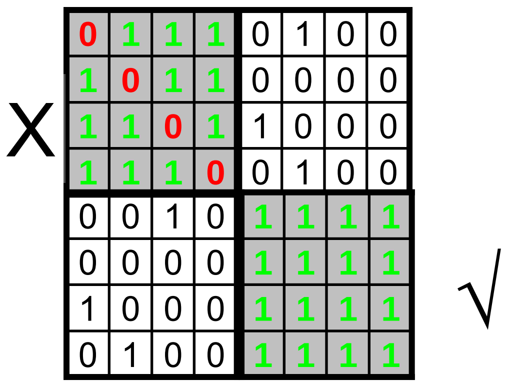


[Fang et al PSB 2010]

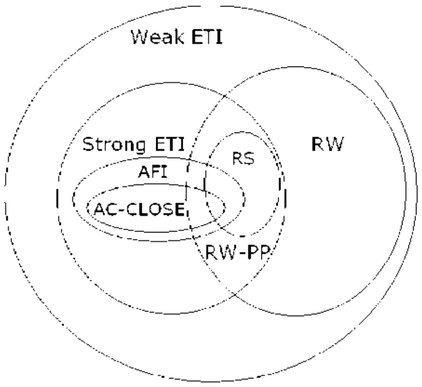
Case Study 3:

Biomarker discovery using error-tolerant patterns

- Association pattern mining is a potential approach to discover multiple markers, however,
  - Too many spurious patterns at low support level
  - True patterns cannot be found at desired level of support as they are fragmented due to random noise



- Possible solution: Error-tolerant patterns
  - These patterns differ in the way errors/noise in the data are tolerated
  - [Yang et al 2001]; [Pei et al 2001]; [Seppanen et al 2004]; [Liu et al 2006]; [Cheng et al 2006]; [Gupta et al., KDD 2008]; [Poernomo et al 2009]

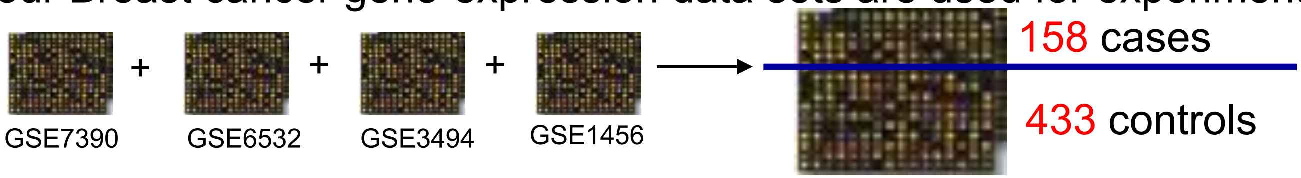


(See Gupta et al KDD 2008 for a survey)

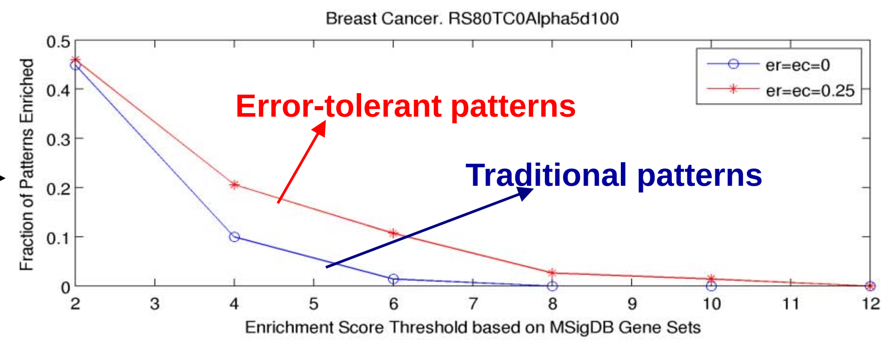
Case Study 3:

Error-tolerant vs. traditional Association patterns

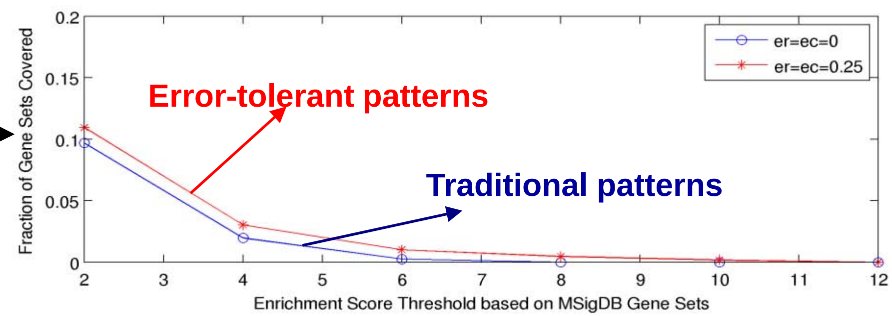
- Four Breast cancer gene-expression data sets are used for experiments:



- Cases: patients with metastasis within 5 years of follow-up;
- Controls: patients with no metastasis within 8 years of follow-up
- Discriminative Error-tolerant and traditional association patterns case/control are discovered and evaluated by enrichment analysis using MSigDB gene sets (Gupta et al 2010)
- Greater fraction of error-tolerant patterns enrich at least one gene set (higher precision)



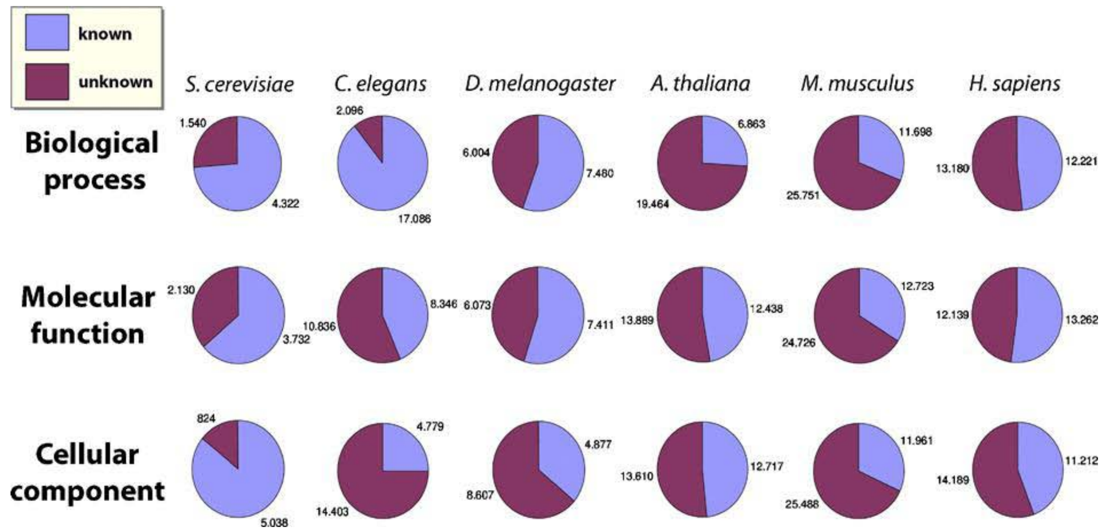
- Greater fraction of gene sets are enriched by at least one error-tolerant pattern (higher recall)



Case Study 4:

Protein Function Prediction

	F_1	F_2	…	F_m	L_1
P_1					✓
P_2					
…					
P_n					✓

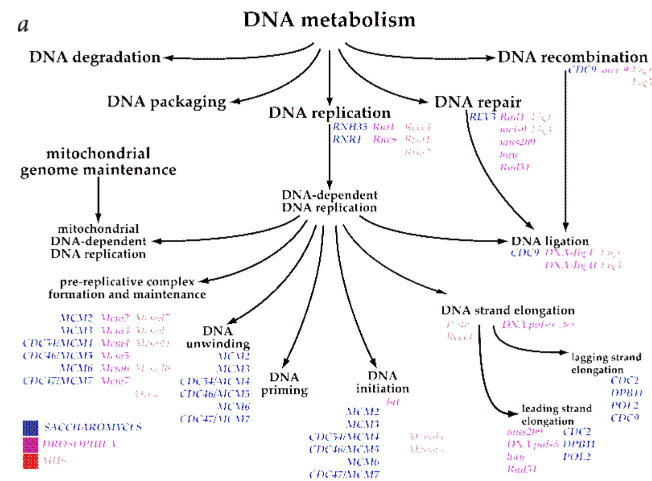


- Protein function prediction one of the most important problems in computational biology.
  - Classification is one of the standard approaches for this problem
    - Pandey *et al.* (2006), “Computational Approaches for Protein Function Prediction: A Survey”, TR 06-028, Dept. of Comp. Sc. & Engg. UMN
    - To be published as a book in the Wiley Bioinformatics series.

Case Study 4:

Protein Function Prediction

	F_1	F_2	…	F_m	L_1	L_2	…	L_k
P_1					✓		✓	
P_2						✓		✓
…								
P_n					✓	✓		✓

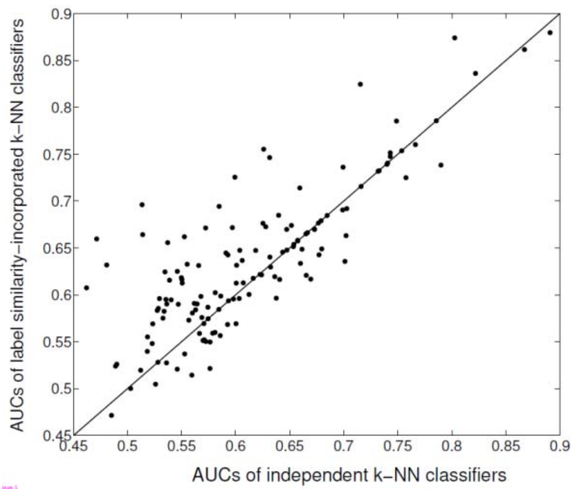


- Inherently a multi-label classification problem
  - Each protein can perform multiple functions.
  - Most labels are infrequent (rare classes)
- Labels (functions) are inter-related in terms of parent-child as well as distant (e.g., sibling) relationships.
  - Inter-relationships captured by Gene Ontology (Ashburner *et al.*, 2000)

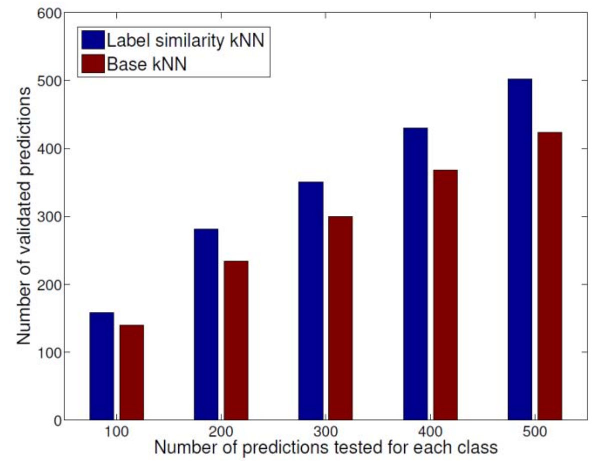
Case Study 4:
Recent Approaches and Future Directions

- k-NN-based approach for incorporating inter-relationships
  - Pandey *et al.*, BMC Bioinformatics, 2009, Tao *et al.*, Bioinformatics, 2007
- SVM+BN approach for enforcing parent-child relationships
  - Barutcuoglu *et al.*, Bioinformatics, 2006

Sample results from Pandey *et al.* (2009) on an yeast gene expression data set.



AUC comparison shows that classification performance is improved, particularly for small (rare) classes.



Classifiers trained on data until Feb, 2008 and tested on annotations added to GO between Feb-Sep, 2008 shows that incorporation enables higher recovery of true annotations.

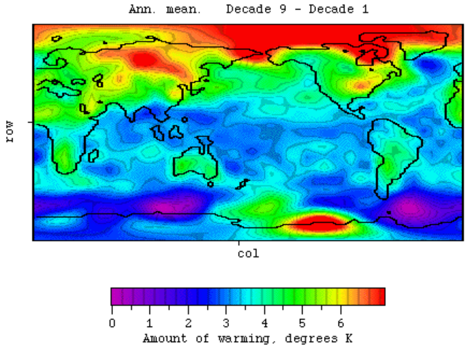
- Possible future directions:
  - Incorporation of label correlations into SVM
  - Design of new measures for capturing label correlation
  - Large-scale incorporation, e.g., all inter-relationships between all classes in GO

Discovery of Climate Patterns from Global Data Sets

Science Goal: Understand global scale patterns in biosphere processes

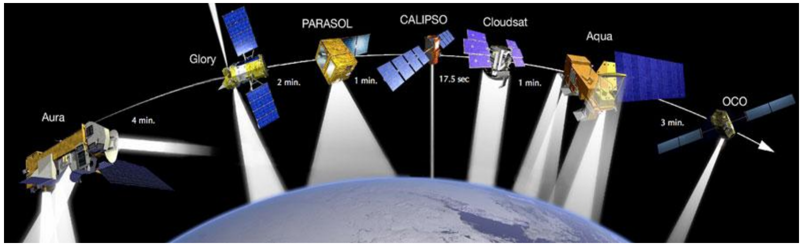
Earth Science Questions:

- When and where do ecosystem disturbances occur?
- What is the scale and location of human-induced land cover change and its impact?
- How are ocean, atmosphere and land processes coupled?



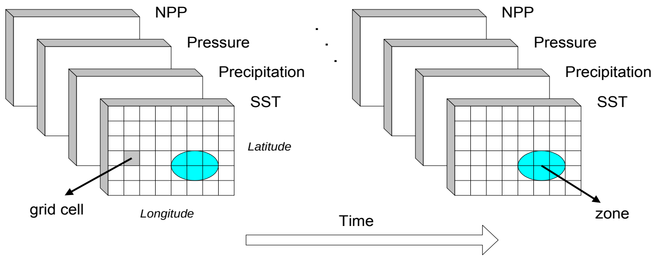
- Data sets need to answer the questions above are becoming available
  - Remote Sensing data from satellites and weather radars
  - Data from in-situ sensors and sensor networks
  - Output from climate and earth system models
  - Geographic Information Systems

Data guided processes can complement hypothesis guided data analysis to develop predictive insights for use by climate scientists, policy makers and community at large.

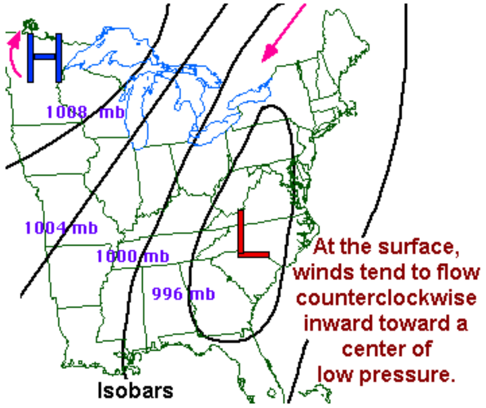


Data Mining Challenges

- **Spatio-temporal nature of data**
  - Traditional data mining techniques do not take advantage of spatial and temporal autocorrelation.



- **Scalability**
  - Size of Earth Science data sets can be very large, especially for data such as high-resolution vegetation.
  - For example, for each time instance,
    - 2.5° x 2.5° :10K locations for the globe
    - 250m x 250m: ~10 billion
    - 50m x 50m : ~250 billion



- **High-dimensionality**
  - Long time series are common in Earth Science
- **Noise and missing values, Nonlinear processes**
- **Multi-Scale nature, Long range dependency**
- **Non-Stationarity**

Case Studies

1. Monitoring of global ecosystem
2. Discovering teleconnections among climate variables
3. Predicting the impacts of climate change

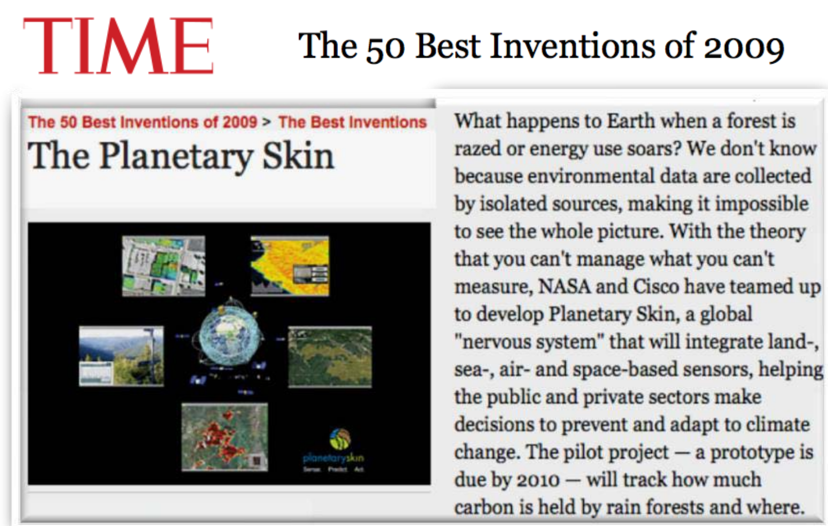
Case Study 1:

Monitoring of global ecosystem

- Planetary Information System for assessment of ecosystem disturbances
  - Forest fires
  - Droughts
  - Floods
  - Logging/deforestation
  - Conversion to agriculture

- This system will help
  - quantify the carbon impact of these changes
  - Understand the relationship to global climate variability and human activity

- Provide **ubiquitous web-based access** to changes occurring across the globe, creating public awareness



TIME The 50 Best Inventions of 2009

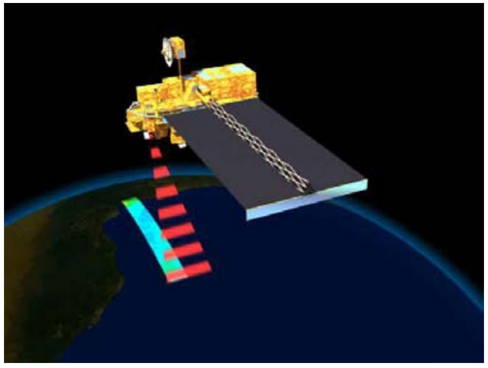
The 50 Best Inventions of 2009 > The Best Inventions

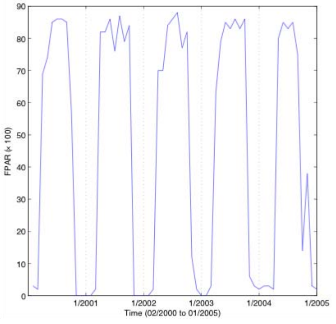
The Planetary Skin

What happens to Earth when a forest is razed or energy use soars? We don't know because environmental data are collected by isolated sources, making it impossible to see the whole picture. With the theory that you can't manage what you can't measure, NASA and Cisco have teamed up to develop Planetary Skin, a global "nervous system" that will integrate land-, sea-, air- and space-based sensors, helping the public and private sectors make decisions to prevent and adapt to climate change. The pilot project — a prototype is due by 2010 — will track how much carbon is held by rain forests and where.

Novel Algorithms for Monitoring Global Eco-system

- State of the art algorithm for land cover change detection do not scale
- Existing Time series change detection algorithms are not suitable for eco-system data
- New algorithms build a non-parametric model of different segments of the time series and use them to capture the degree of change





MODIS captures high quality vegetation index data from Feb 2000 to present at various spatial resolutions.

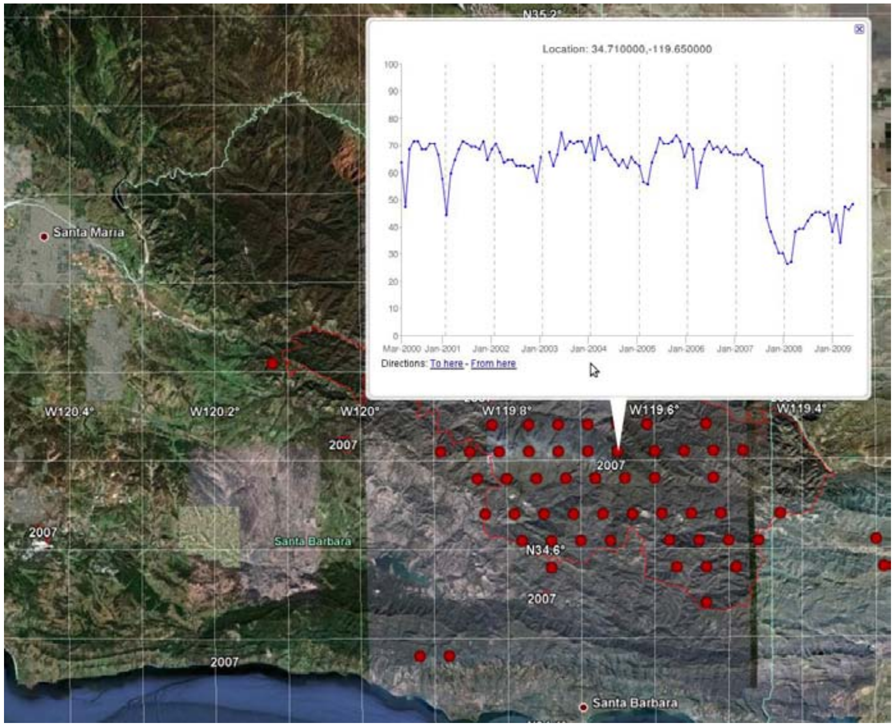
Challenges:
Noise, missing values, outliers, high degree of variability (across regions, vegetation types, and time)

S. Boriah, V. Kumar, M. Steinbach, et al., *Land cover change detection: a case study*, *KDD* 2008.

Case Study 1:

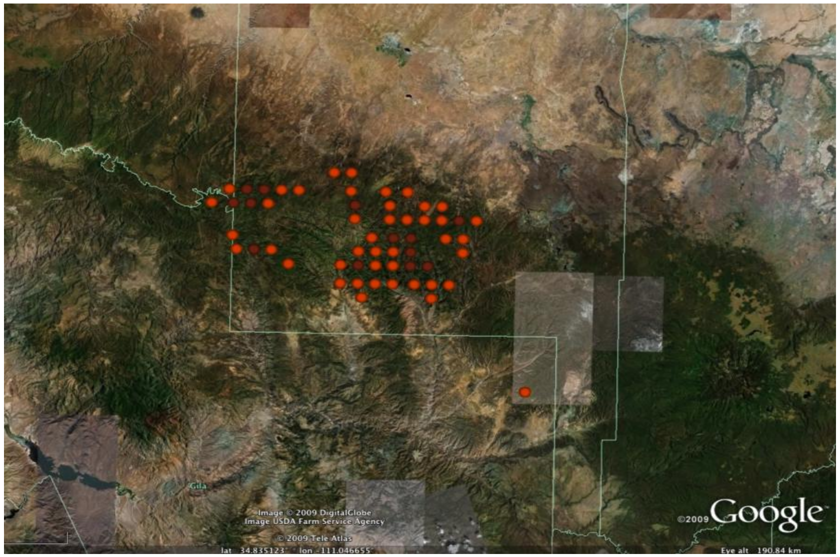
Monitoring Global Forest Cover

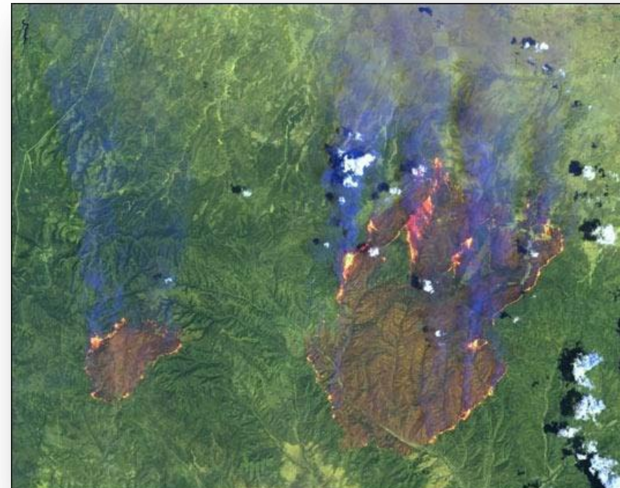
California Fires: 2007 Santa Barbara Fire

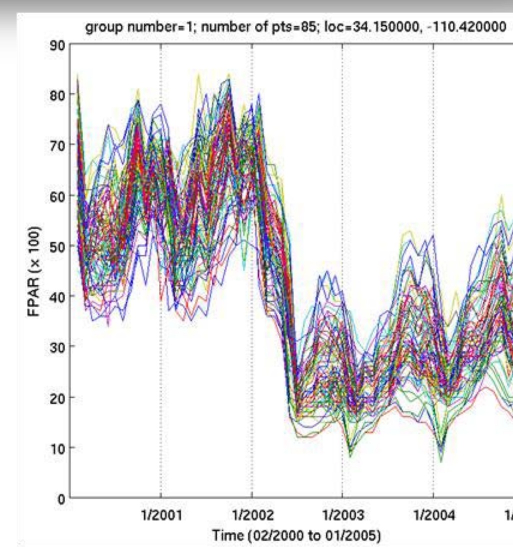


- Fire detected is the well documented Zaca Fire. It began burning about 15 miles northeast of Buellton, California. The fire started on July 4, 2007 and by August 31, it had burned over 240,207 acres (972.083 km^2), making it California's second largest fire and Santa Barbara‟s county largest fire.
- The fire was human induced and started as a result of sparks from a grinding machine on private property which was being used to repair a water pipe. The fire cost $118.3 million to fight and involved 21 fire crews.

Arizona





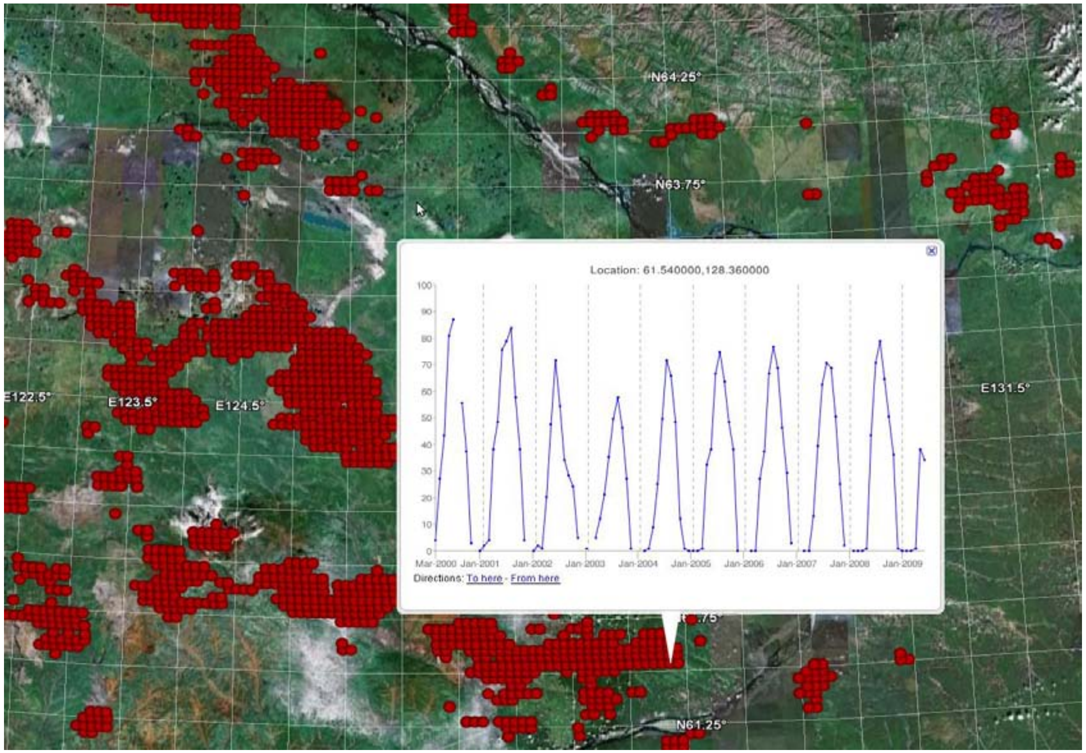


Two huge forest fires have become one giant inferno sweeping across the American state of Arizona.

http://news.bbc.co.uk/cbbcnews/hi/world/newsid_2061000/2061402.stm

June 2002

Large Outbreak of Fires near Yakutsk, Russia



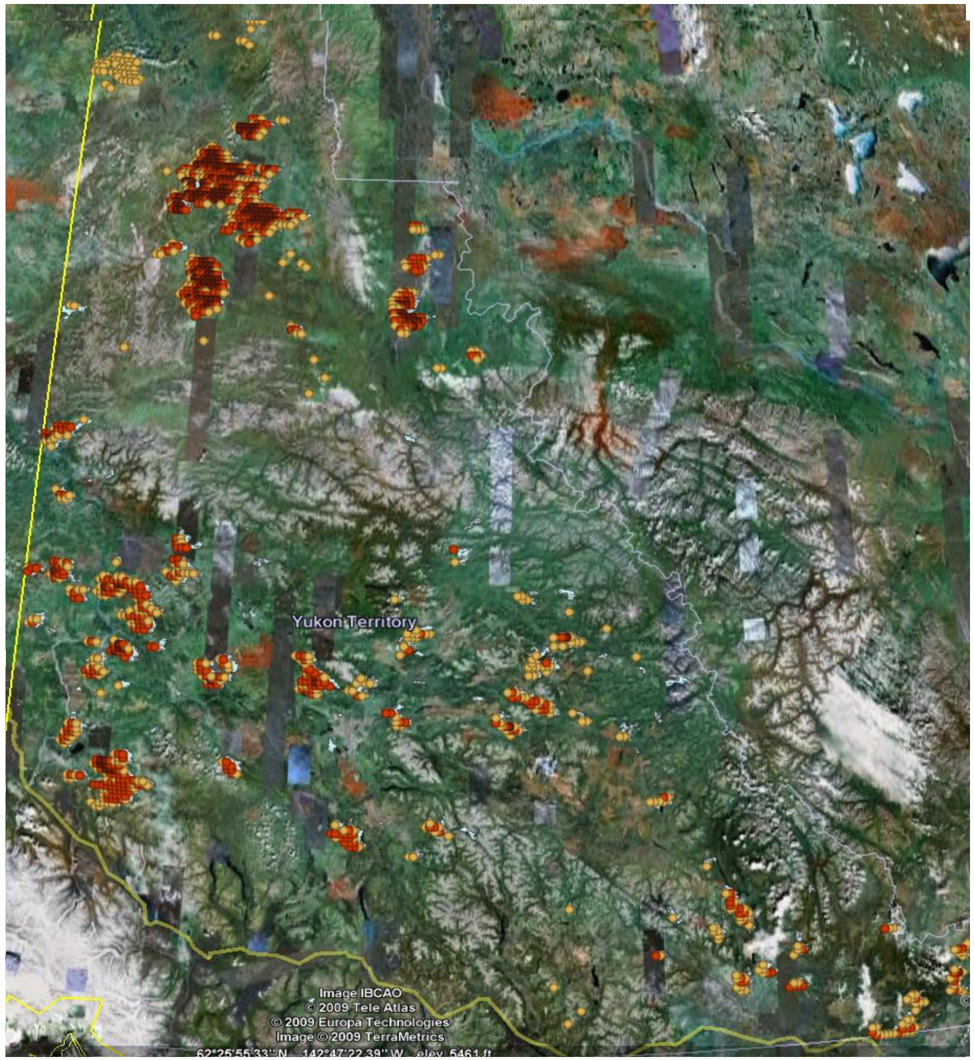
During the summer months in the Northern Hemisphere, many fires are ignited in the boreal forests of Canada and Russia by lightning striking the surface

Image courtesy Jacques Descloitres, MODIS Land Rapid Response Team

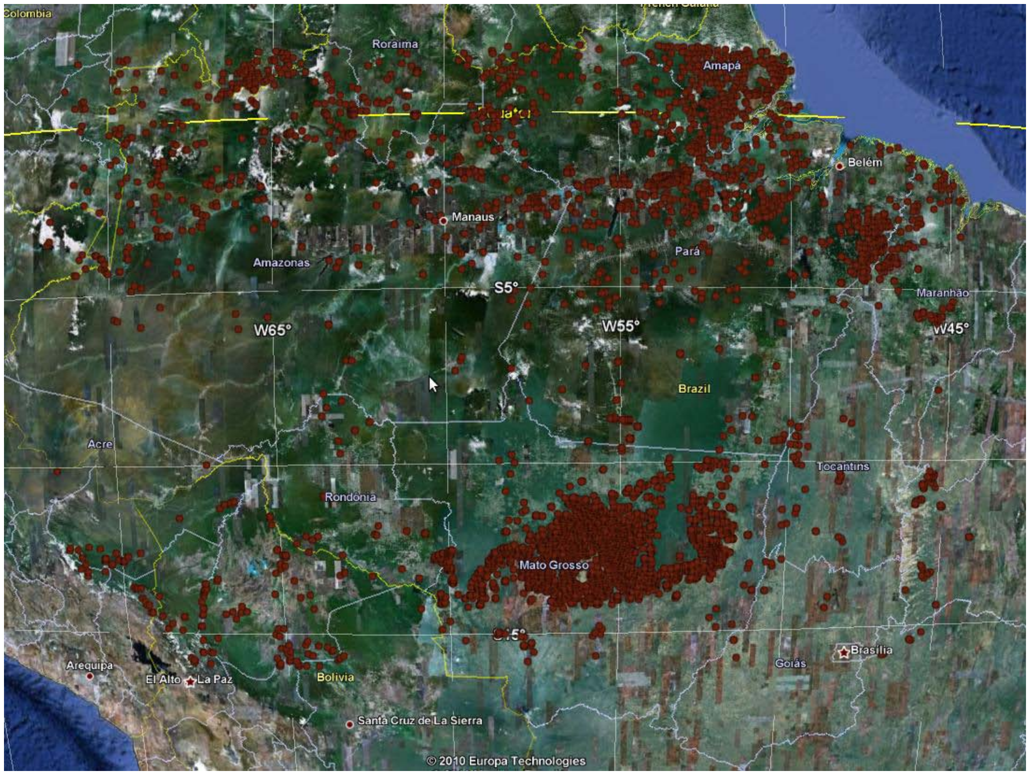
2:40 UTC

4:15 UTC

Canada: Fires in Yukon Province

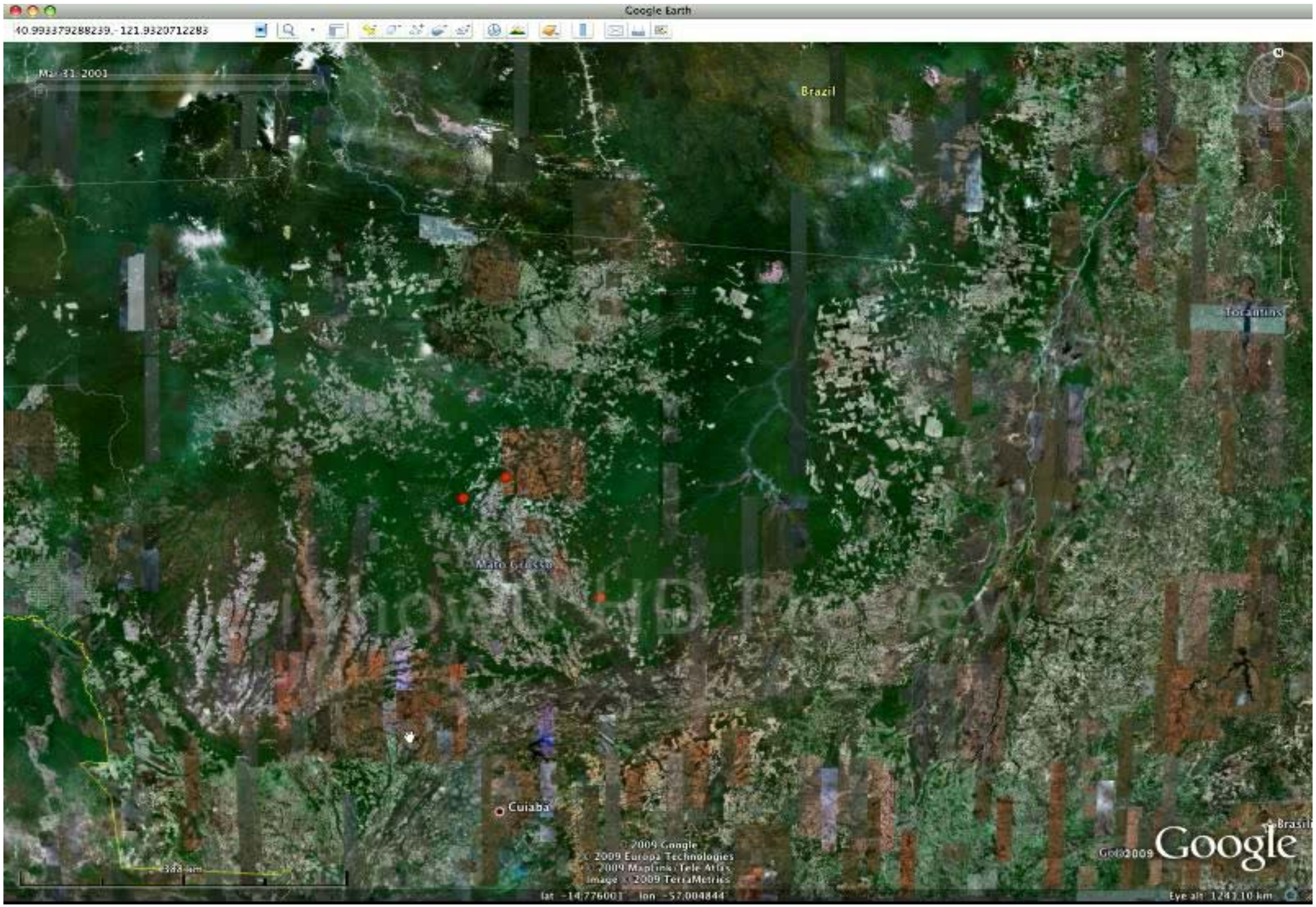


Amazon Rainforest

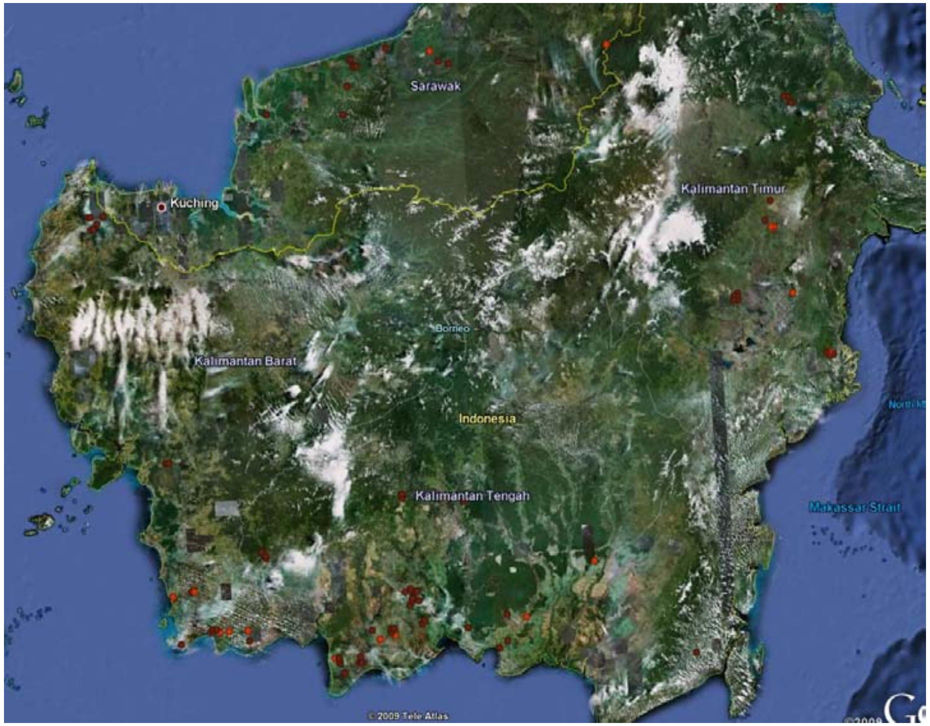


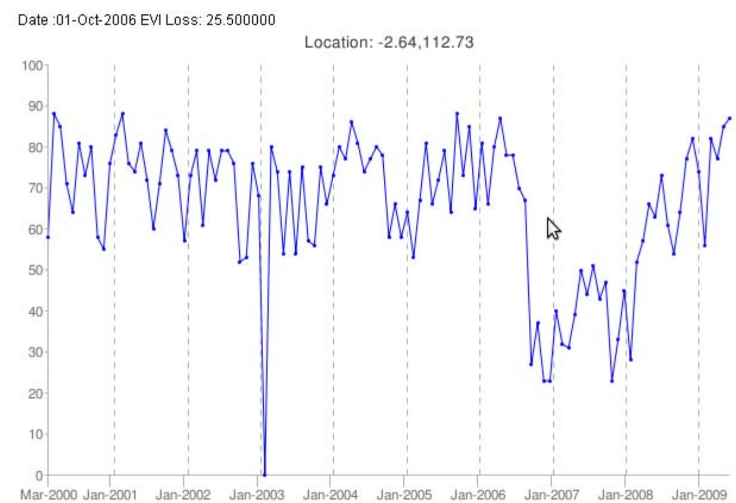
Brazil Accounts for almost 50% of all humid tropical forest clearing, nearly 4 times that of the next highest country, which accounts for 12.8% of the total.

Amazon Animation



Indonesia



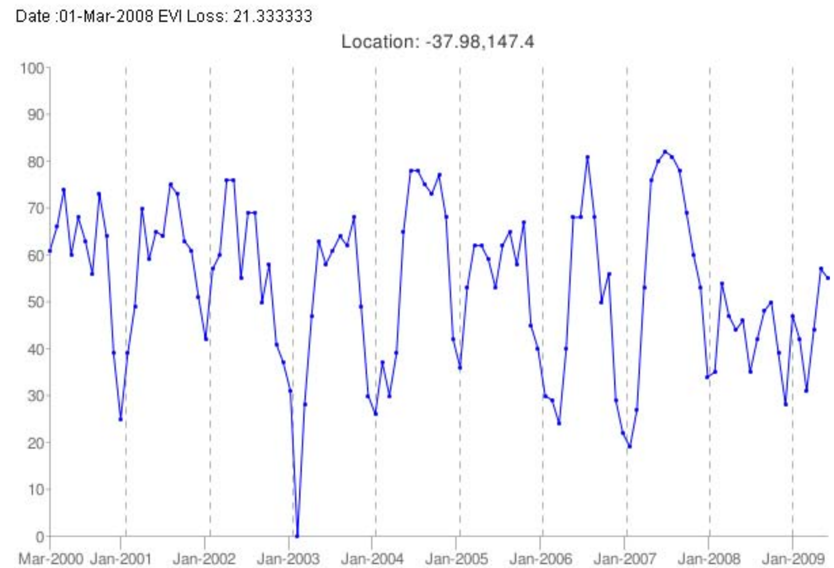


newsVOA.com

11 September 2006

Forest Fires Sweep Indonesia Borneo and Sumatra.
Officials in Indonesia say illegal burning to clear land has caused rampant wildfires across Borneo and Sumatra ... eight million hectares have gone up in smoke over the last month, and fires are still burning out of control on the island of Borneo.

Victoria (Australia)



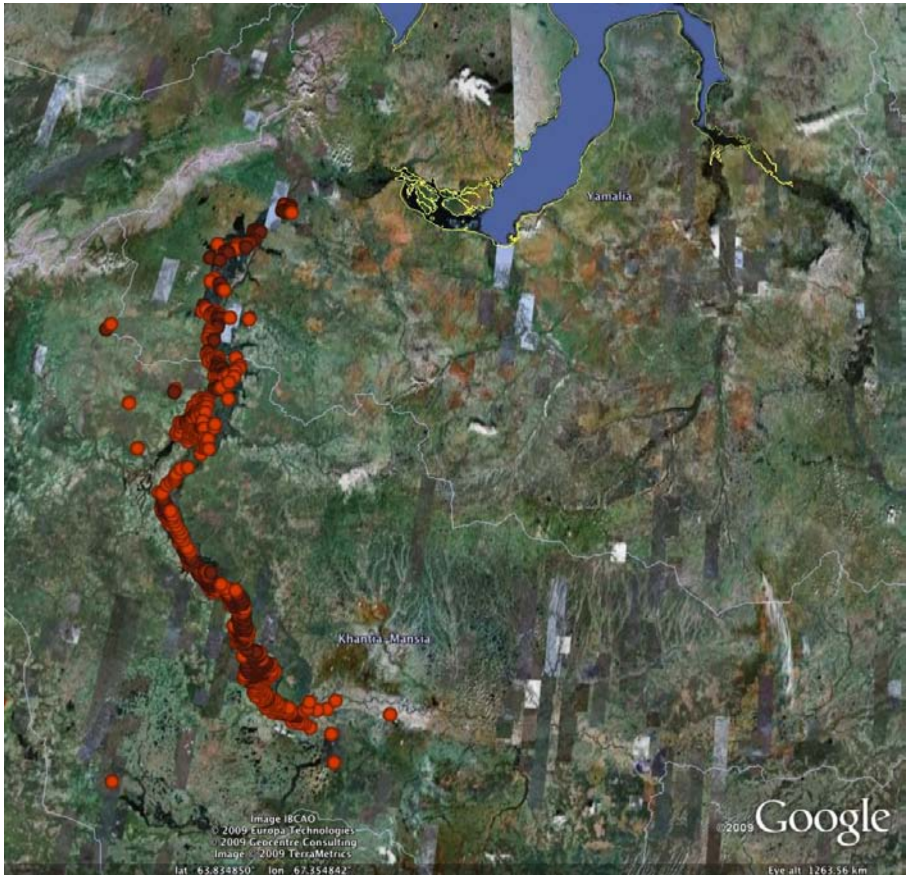
Drought in southern Australia declared 'worst on record'

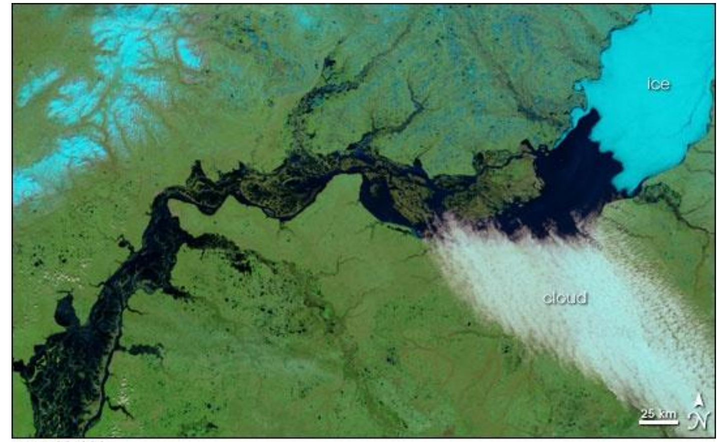
October 10, 2008

David Jones, the head of climate analysis at the Bureau of Meteorology, said the drought affecting south-west Western Australia, south-east South Australia, Victoria and northern Tasmania "is now very severe and without historical precedent".

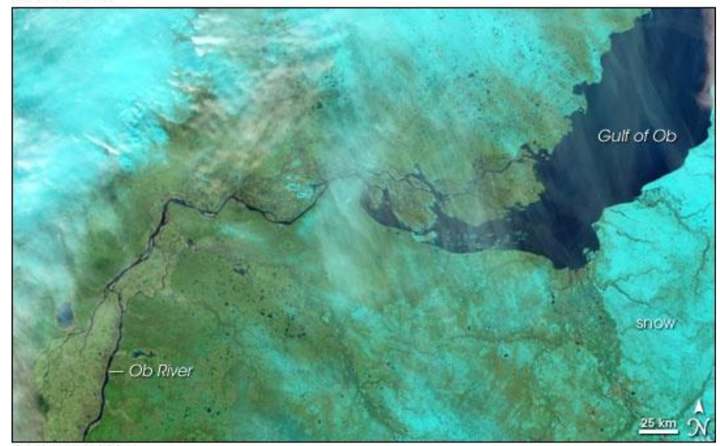
Source: climateprogress.org

Flooding along Ob River, Russia





June 20, 2007

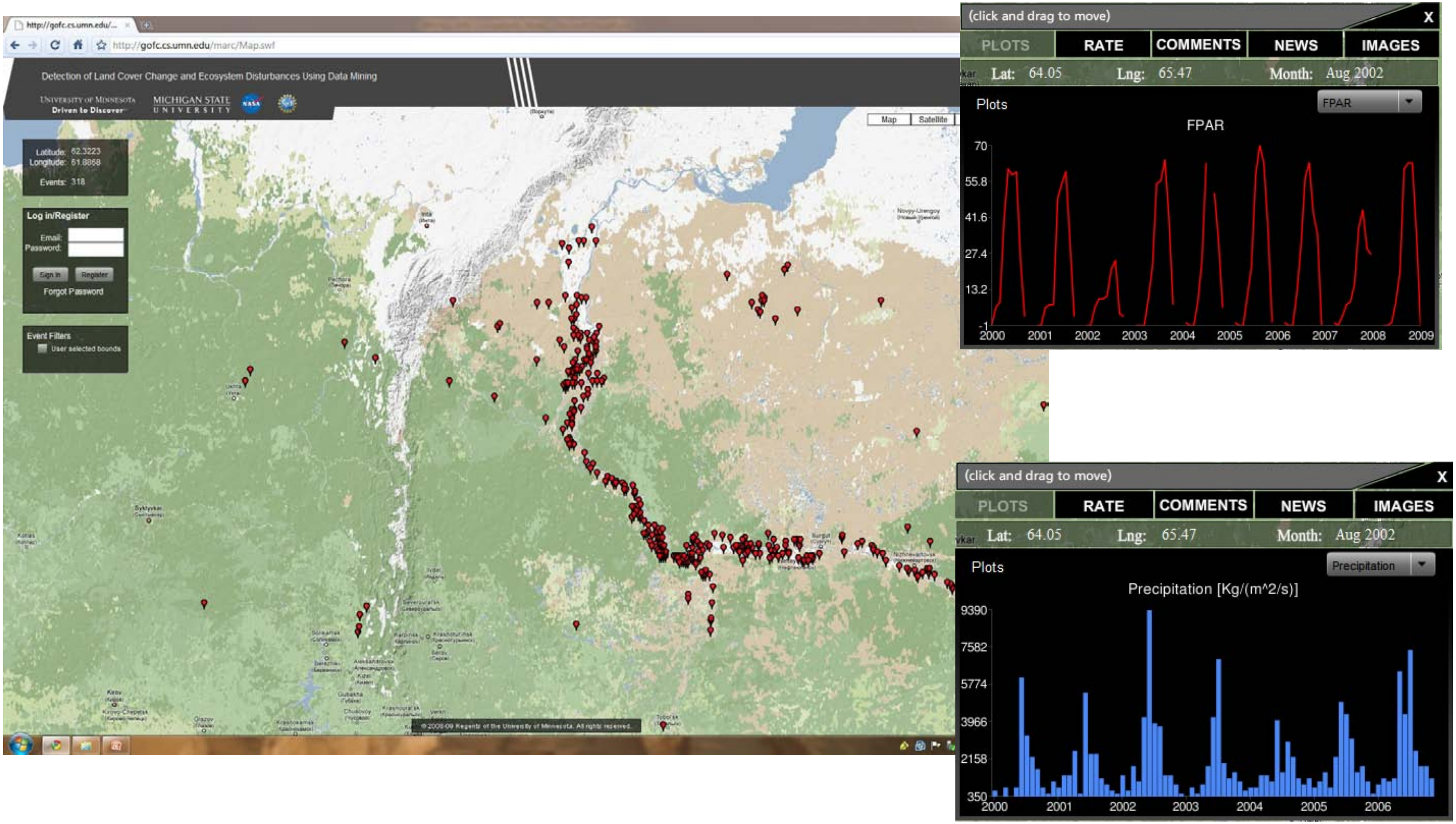


October 5, 2006

The river flows north and is blocked by ice (top right), which causes flooding. Under normal circumstances the river flows into the Gulf of Ob.

Source: NASA Earth Observatory

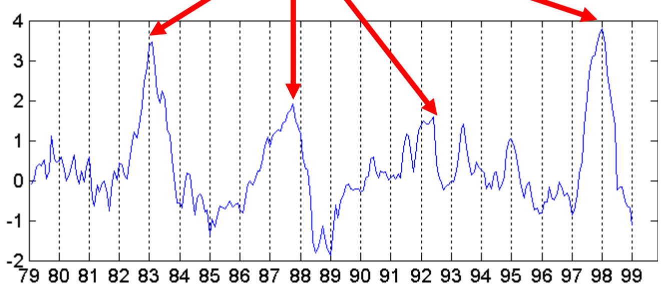
Web 2.0 interface for planetary information system



Case Study 2:
Discovering teleconnections:
Relationship among ocean/atmosphere and the land

- Climate indices capture teleconnections (in both space and time)
  - The simultaneous variation in climate and related processes over widely separated points on the Earth

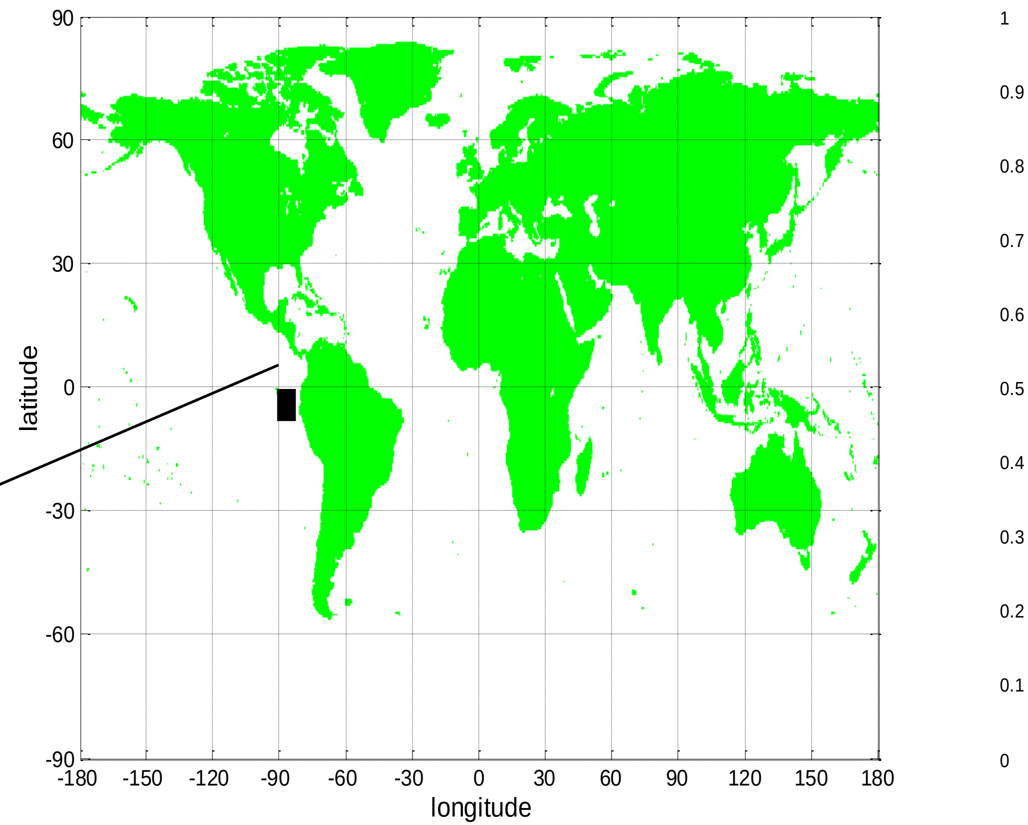
El Nino Events



Nino 1+2 Index

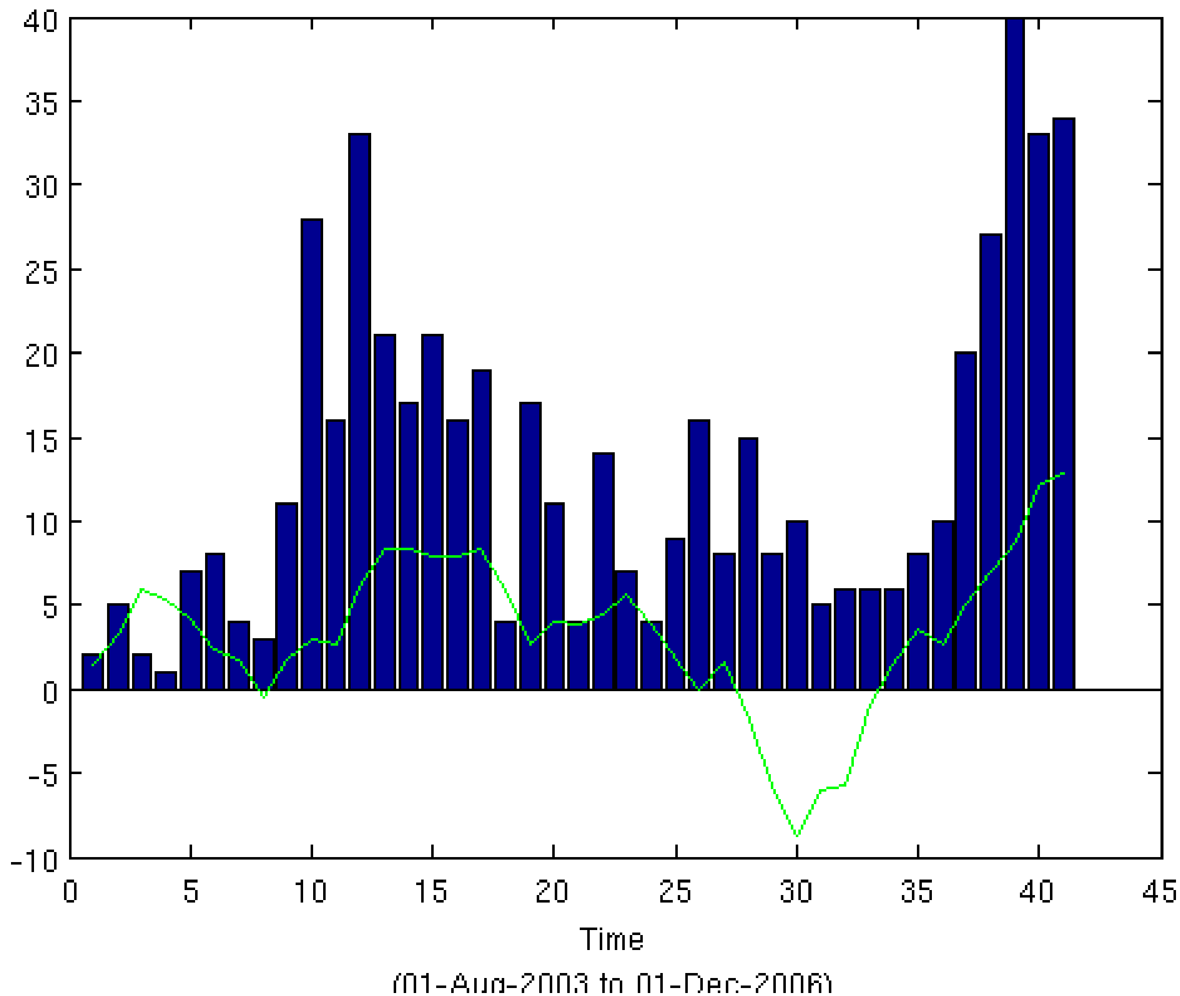
Sea surface temperature anomalies in the region bounded by 80° W-90° W and 0° - 10° S

Correlation Between ANOM 1+2 and Land Temp (>0.2)



Effects: Drought in Australia, warmer winter in North America, flooding in coastal Peru, increased rainfall in East Africa

Relationship between El Nino and Fires in Indonesia

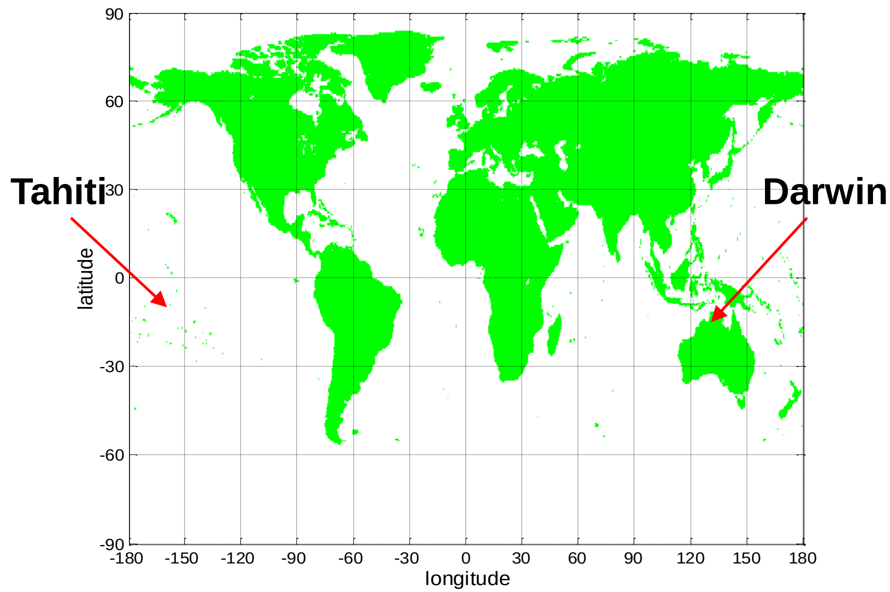


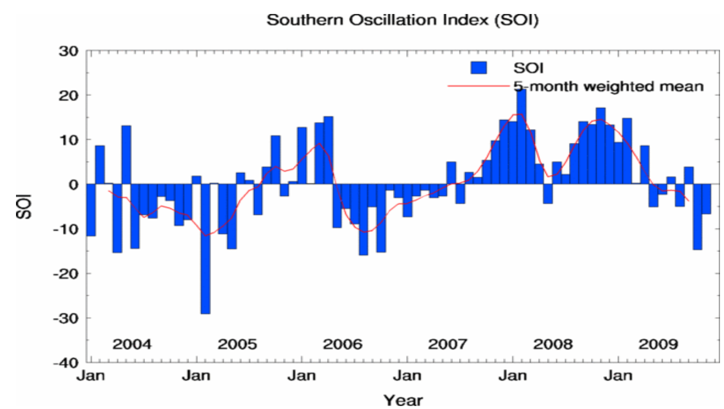
A Pressure Based El Niño Index: SOI

- The Southern Oscillation Index (SOI) is also associated with El Niño.

- Defined as the normalized pressure differences between Tahiti and Darwin Australia.

- Both temperature and pressure based indices capture the same El Niño climate phenomenon.

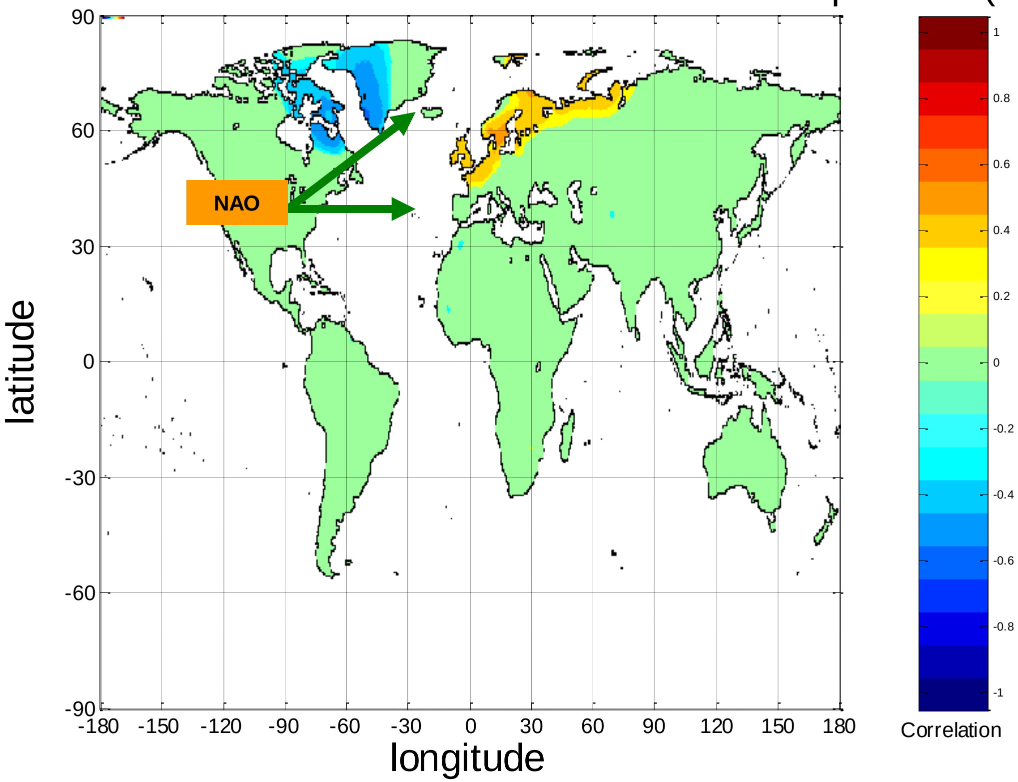


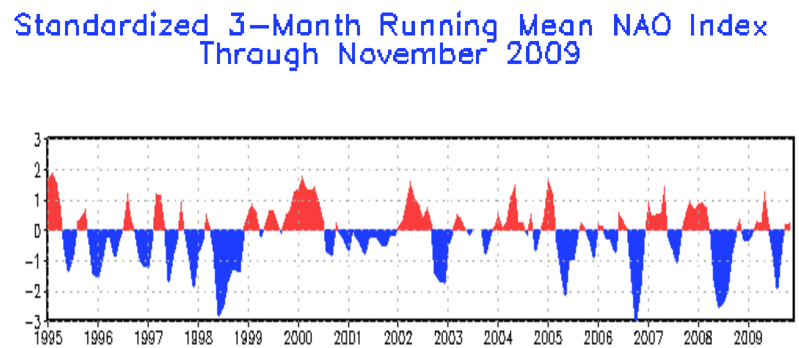


NAO (North Atlantic Oscillation)

- NAO computed as the normalized difference between SLP at a pair of land stations in the Arctic and the subtropical Atlantic regions of the North Atlantic Ocean

Correlation Between NAO and Land Temperature (>0.3)





List of Well Known Climate Indices

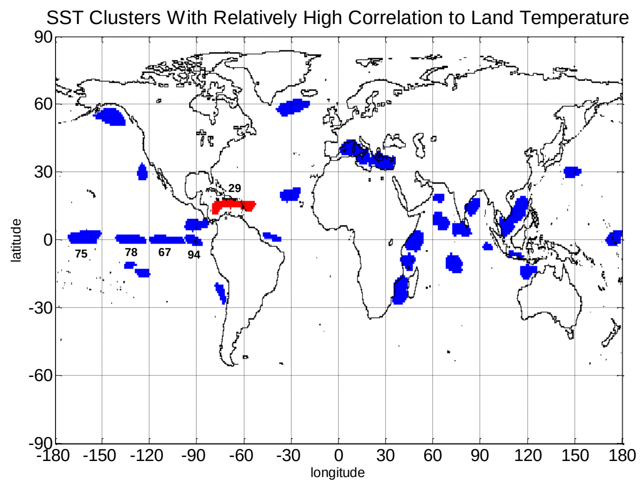
Index	Description
SOI	**Southern Oscillation Index:** Measures the SLP anomalies between Darwin and Tahiti
NAO	**North Atlantic Oscillation:** Normalized SLP differences between Ponta Delgada, Azores and Stykkisholmur, Iceland
AO	**Arctic Oscillation:** Defined as the _first principal component of SLP poleward of 20° N
PDO	**Pacific Decadel Oscillation:** Derived as the leading principal component of monthly SST anomalies in the North Pacific Ocean, poleward of 20° N
QBO	**Quasi-Biennial Oscillation Index:** Measures the regular variation of zonal (i.e. east-west) strato-spheric winds above the equator
CTI	**Cold Tongue Index:** Captures SST variations in the cold tongue region of the equatorial Pacific Ocean (6° N-6° S, 180° -90° W)
WP	**Western Pacific:** Represents a low-frequency temporal function of the „zonal dipole' SLP spatial pattern involving the Kamchatka Peninsula, southeastern Asia and far western tropical and subtropical North Pacific
NINO1+2	Sea surface temperature anomalies in the region bounded by 80° W-90° W and 0° -10° S
NINO3	Sea surface temperature anomalies in the region bounded by 90° W-150° W and 5° S-5° N
NINO3.4	Sea surface temperature anomalies in the region bounded by 120° W-170° W and 5° S-5° N
NINO4	Sea surface temperature anomalies in the region bounded by 150° W-160° W and 5° S-5° N

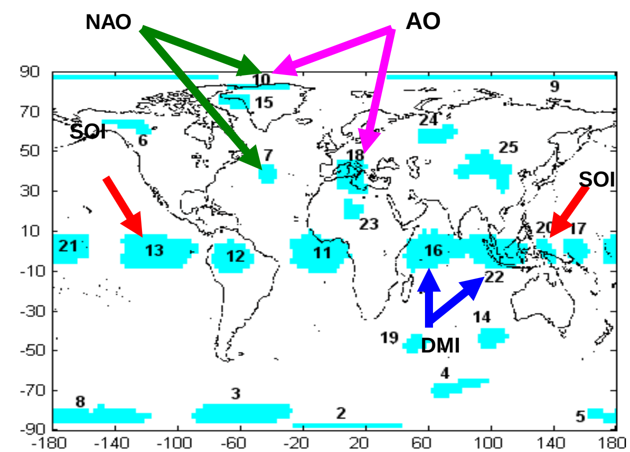
Discovered primarily by human observation

Discovery of Climate Indices Using Clustering

- Clustering provides an alternative approach for finding candidate indices.
- Clusters are found using the Shared Nearest Neighbor (SNN) method that eliminates “noise” points and tends to find homogeneous regions of “uniform density”.
- Clusters are filtered to eliminate those with low impact on land points

Cluster	Nino Index	Correlation
94	NINO 1+2	0.9225
67	NINO 3	0.9462
78	NINO 3.4	0.9196
75	NINO 4	0.9165



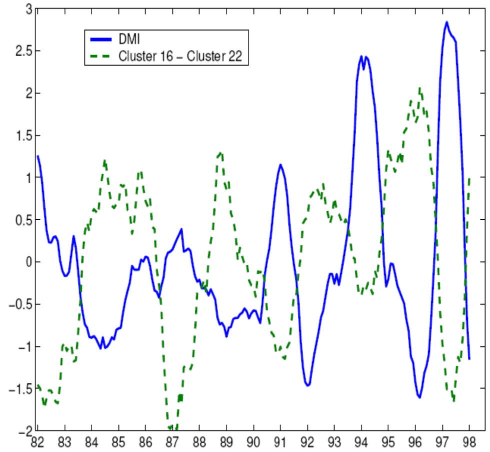


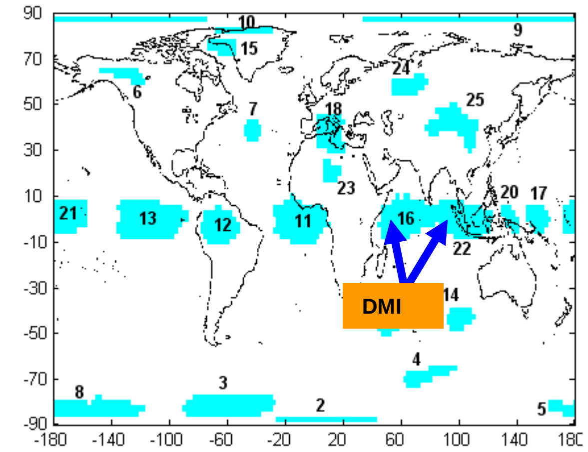
M. Steinbach, P. Tan, V. Kumar, C. Potter and S. Klooster. Discovery of Climate Indices Using Clustering, *Proceedings of KDD 2003*.

Automated Discovery of Climate Indices: Opportunities and Challenges

Opportunities:

- Discover new relationships that are difficult to find manually
- Example:
  - **DMI** is a temperature based index which is an indicator of weak mansoon over Indian subcontinent and heavy rainfall over east Africa.
  - **Clustering** finds a pressure based surrogate

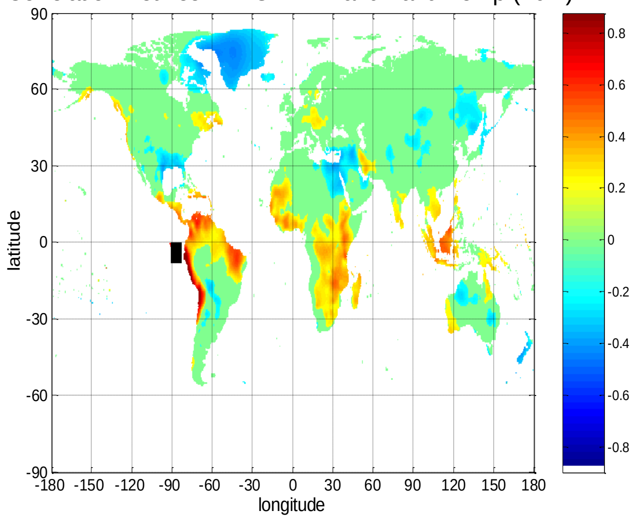




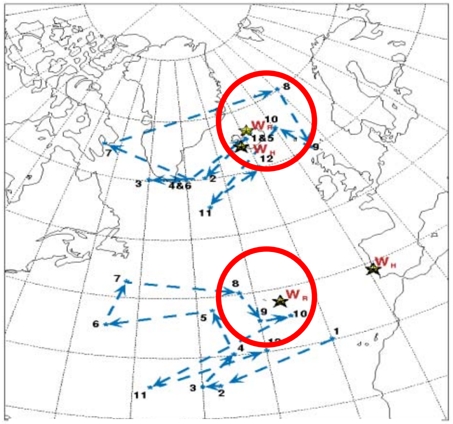
Challenges:

- Nonlinear, dynamic relationships
- Long term spatial and temporal dependence
- Spatio temporal auto-correlation
- Multi-scale multi-resolution
- Distinguishing spurious relationships from real

Correlation Between ANOM 1+2 and Land Temp (>0.2)



Phenomenon underlies NAO is dynamic



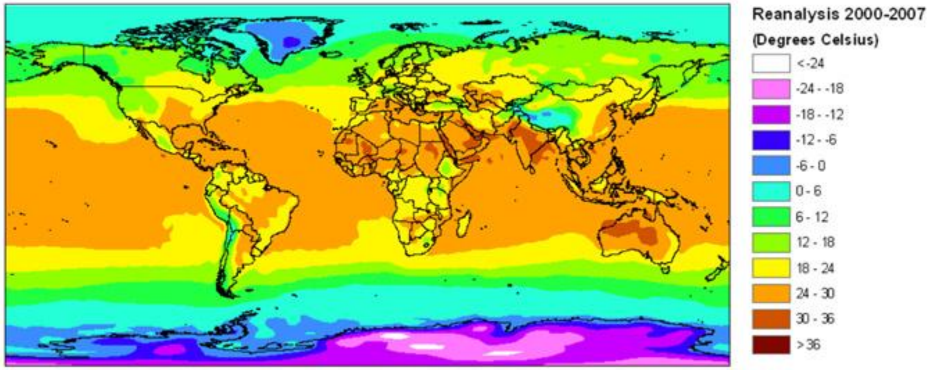
Source: Portis et al, Seasonality of the NAO, AGU Chapman Conference, 2000.

Case Study 3:

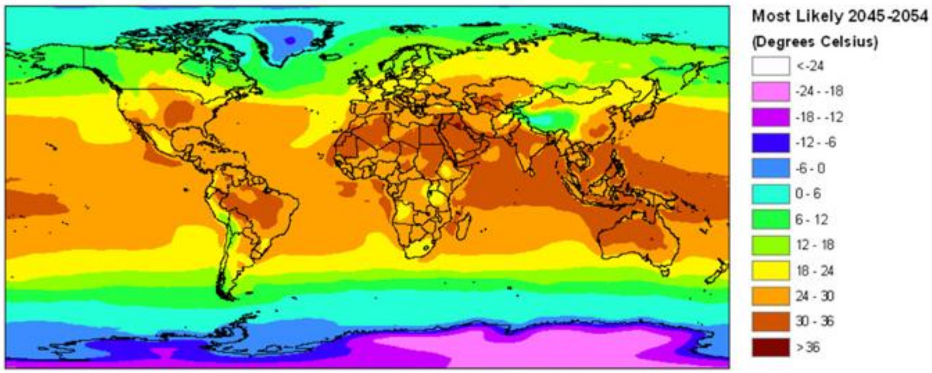
Planning for Climate Change and Extreme Events

- One of the predicted impacts of climate change is an increase in climate extremes:
  - Droughts, fires, cyclones, severe storms, heat waves
- Many of these impacts cannot be predicted using physics based models
- A possible pproach:
  - Extract climate indices and features for extreme events from past observations.
  - Develop predictive capabilities for extreme events using these features
  - Generate climate forecasts using climate indices and Global Circulation Models (GCMs)

Higher trends but larger uncertainty and geographic variability in 21st century temperature and heat waves, Ganguly, Steinhaeuser, et. al, PNAS 2009

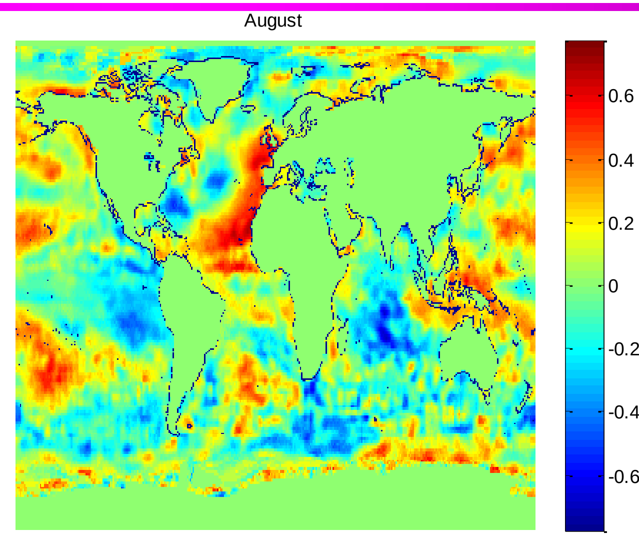


Heat wave intensity from reanalysis data for 2000–2007



2050 heat wave projections from the A1FI climate scenario

Predicting Tropical Storm Counts from Climate Model Projections for SST



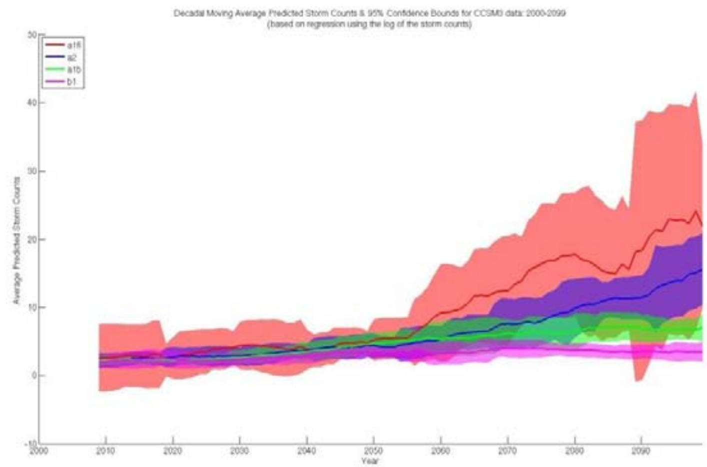
Correlation between Sea Surface Temperature and the number of tropical cyclones off the western coast of Africa from 1982 to 2007.

Approach

- Build a regression model that relates August SST values off the western coast with the August tropical storm counts.
- Use predicted SST from climate scenarios produced by Global Climate Models (GCMs) to compute projected cyclones.

Challenges

- multi-scale nature,
- nonlinearity,
- long range spatial and temporal relationships



10 year moving average of predictions based on linear regression, and SST from four IPCC climate scenarios.

(Joint work with Ganguly and Semazzi).

Summary

- Data driven discovery methods hold great promise for advancement in a variety of scientific disciplines
- Challenges arise due to the complex nature of scientific data sets
  - Climate:
    - Significant amounts of missing values, especially in the tropics
    - Multi-scale/Multi-resolution nature, Variability
    - Spatio-temporal autocorrelation
    - Long-range spatial dependence
    - Long memory temporal processes (teleconnections)
    - Nonlinear processes, Non-Stationarity
    - Fusing multiple sources of data
  - Bioinformatics:
    - High dimensionality
    - Heterogeneous nature
    - Noise, missing values
    - Integration of heterogeneous data

Team Members and Collaborators

Michael Steinbach, Shyam Boriah, Gaurav Pandey, Rohit Gupta, Gang Fang, Gowtham Atluri, Varun Mithal, Ashish Garg, Vanja Paunic, Sanjoy Dey, Deepthi Cheboli, Marc Dunham, Divya Alla, Matt Kappel, Ivan Brugere, Vikrant Krishna

Bioinformatics:

Brian Van Ness, Bill Oetting, Gary L. Nelsestuen, Christine Wendt, Piet C. de Groen, Michael Wilson, Rui Kuang, Chad Myers

Climate and Eco-system:

Sudipto Banerjee, Chris Potter, Fred Semazzi, Steve Klooster, Auroop Ganguly, Pang-Ning Tan, Joe Knight, Arindam Banerjee

Project websites
Bioinformatics: www.cs.umn.edu/~kumar/dmbio
Climate and Eco-system: www.cs.umn.edu/~kumar/nasa-umn

References

- Gaurav Pandey, Chad L. Myers and Vipin Kumar, Incorporating Functional Inter-relationships into Protein Function Prediction Algorithms, BMC Bioinformatics, 10:142, 2009 (Highly Accessed).
- Brian Van Ness, Christine Ramos, Majda Haznadar, Antje Hoering, Jeff Haessler, John Crowley, Susanna Jacobus, Martin Oken, Vincent Rajkumar, Philip Greipp, Bart Barlogie, Brian Durie, Michael Katz, Gowtham Atluri, Gang Fang, Rohit Gupta, Michael Steinbach, Vipin Kumar, Richard Mushlin, David Johnson and Gareth Morgan, Genomic Variation in Myeloma: Design, content and initial application of the Bank On A Cure SNP Panel to detect associations with progression free survival, BMC Medicine, Volume 6, pp 26, 2008.
- TaeHyun Hwang, Hugues Sicotte, Ze Tian, Baolin Wu, Dennis Wigle, Jean-Pierre Kocher, Vipin Kumar and Rui Kuang, Robust and Efficient Identification of Biomarkers by Classifying Features on Graphs, Bioinformatics, Volume 24, no. 18, pages 2023-2029, 2008
- Rohit Gupta, Smita Agrawal, Navneet Rao, Ze Tian, Rui Kuang, Vipin Kumar, Integrative Biomarker Discovery for Breast Cancer Metastasis from Gene Expression and Protein Interaction Data Using Error-tolerant Pattern Mining, Proceedings of the International Conference on Bioinformatics and Computational Biology (BICoB), March 2010 (Also published as CS Technical Report).
- Gang Fang, Rui Kuang, Gaurav Pandey, Michael Steinbach, Chad L. Myers and Vipin Kumar, Subspace Differential Coexpression Analysis: Problem Definition and A General Approach, Proceedings of the 15th Pacific Symposium on Biocomputing (PSB), 15:145-156, 2010. (software and codes)
- Gang Fang, Gaurav Pandey, Manish Gupta, Michael Steinbach, and Vipin Kumar, Mining Low-support discriminative patterns from Dense and High-dimensional Data, TR09-011, CS@UMN, 2009
- Rohit Gupta, Navneet Rao, Vipin Kumar, "A Novel Error-Tolerant Frequent Itemset Model for Binary and Real-Valued Data", CS Technical Report 09-026, University of Minnesota.
- Gaurav Pandey, Gowtham Atluri, Michael Steinbach, Chad L. Myers and Vipin Kumar, An Association Analysis Approach to Biclustering, Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD) 2009.
- Gaurav Pandey, Gowtham Atluri, Gang Fang, Rohit Gupta, Michael Steinbach and Vipin Kumar, Association Analysis Techniques for Analyzing Complex Biological Data Sets, Proceedings of the IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS), in press, 2009.
- Gowtham Atluri, Rohit Gupta, Gang Fang, Gaurav Pandey, Michael Steinbach and Vipin Kumar, Association Analysis Techniques for Bioinformatics Problems, Proceedings of the 1st International Conference on Bioinformatics and Computational Biology (BICoB), pp 1-13, 2009 (Invited paper).
- Rohit Gupta, Michael Steinbach, Karla Ballman, Vipin Kumar, Petrus C. de Groen, "Colorectal Cancer Despite Colonoscopy: Critical Is the Endoscopist, Not the Withdrawal Time", [Abstract] Gastroenterology, Volume 136, Issue 5, Supplement 1, May 2009, Pages A-55. (Selected for presentation in clinical science plenary session in DDW 2009) [Recipient of Student Abstract Prize]
- Rohit Gupta, Michael Steinbach, Karla Ballman, Vipin Kumar, Petrus C. de Groen, "Colorectal Cancer Despite Colonoscopy: Estimated Size of the Truly Missed Lesions". [Abstract] Gastroenterology, Volume 136, Issue 5, Supplement 1, May 2009, Pages A-764. (Presented in DDW 2009)
- Rohit Gupta, Brian N. Brownlow, Robert A. Domnick, Gavin Harewood, Michael Steinbach, Vipin Kumar, Piet C. de Groen, Colon Cancer Not Prevented By Colonoscopy, American College of Gastroenterology (ACG) Annual Meeting, 2008 (Recipient of the 2008 ACG Olympus Award and the 2008 ACG Presidential Award)
- Gaurav Pandey, Lakshmi Naarayanan Ramakrishnan, Michael Steinbach and Vipin Kumar, Systematic Evaluation of Scaling Methods for Gene Expression Data, Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp 376-381, 2008.
- Gaurav Pandey, Gowtham Atluri, Michael Steinbach and Vipin Kumar, Association Analysis Techniques for Discovering Functional Modules from Microarray Data , Proceedings of the ISMB satellite meeting on Automated Function Prediction 2008 (Also published as Nature Precedings 10.1038/npre.2008.2184.1)
- Rohit Gupta, Gang Fang, Blayne Field, Michael Steinbach and Vipin Kumar, Quantitative Evaluation of Approximate Frequent Pattern Mining Algorithms,

References (Cont..)

- Gaurav Pandey, Michael Steinbach, Rohit Gupta, Tushar Garg and Vipin Kumar, Association Analysis-based Transformations for Protein Interaction Networks: A Function Prediction Case Study, Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), pp 540-549, 2007 (Also selected for a Highlight talk at ISMB 2008).
- Gaurav Pandey and Vipin Kumar, Incorporating Functional Inter-relationships into Algorithms for Protein Function Prediction, Proceedings of the ISMB satellite meeting on Automated Function Prediction 2007
- Rohit Gupta, Tushar Garg, Gaurav Pandey, Michael Steinbach and Vipin Kumar, Comparative Study of Various Genomic Data Sets for Protein Function Prediction and Enhancements Using Association Analysis, Proceedings of the Workshop on Data Mining for Biomedical Informatics, held in conjunction with SIAM International Conference on Data Mining, 2007
- Hui Xiong, X. He, Chris Ding, Ya Zhang, Vipin Kumar and Stephen R. Holbrook, Identification of Functional Modules in Protein Complexes via Hyperclique Pattern Discovery, pp 221-232, Proc. of the Pacific Symposium on Biocomputing, 2005
- Benjamin Mayer, Huzefa Rangwala, Rohit Gupta, Jaideep Srivastava, George Karypis, Vipin Kumar and Piet de Groen, Feature Mining for Prediction of Degree of Liver Fibrosis, Proc. Annual Symposium of American Medical Informatics Association (AMIA), 2005
- Gowtham Atluri, Gaurav Pandey, Jeremy Bellay, Chad Myers and Vipin Kumar, Two-Dimensional Association Analysis For Finding Constant Value Biclusters In Real-Valued Data, Technical Report 09-020, July 2009, Department of Computer Science, University of Minnesota
- Gaurav Pandey, Gowtham Atluri, Michael Steinbach and Vipin Kumar, Association Analysis for Real-valued Data: Definitions and Application to Microarray Data, Technical Report 08-007, March 2008, Department of Computer Science, University of Minnesota
- Gaurav Pandey, Lakshmi Naarayanan Ramakrishnan, Michael Steinbach, Vipin Kumar, Systematic Evaluation of Scaling Methods for Gene Expression Data, Technical Report 07-015, August 2007, Department of Computer Science, University of Minnesota
- Gaurav Pandey, Vipin Kumar and Michael Steinbach, Computational Approaches for Protein Function Prediction: A Survey, Technical Report 06-028, October 2006, Department of Computer Science, University of Minnesota
- G. Dong and J. Li. Efficient mining of emerging paterns: Discovering trends and differences. In Proceedings of the 2001 ACM SIGKDD international conference on knowledge discovery in databases, pages 43–52, 1999
- S. Bay and M. Pazzani. Detecting group differences: Mining contrast sets. Data Mining and Knowledge Discovery, 5(3):213–246, 2001.
- H. Cheng, X. Yan, J. Han, and C.-W. Hsu. Discriminative frequent pattern analysis for effective classification. In Proceedings of International Conference on Data Engineering, pages 716–725, 2007.
- H. Cheng, X. Yan, J. Han, and P. Yu. Direct discriminative pattern mining for effective classification. In Proceedings of International Conference on Data Engineering, pages 169–178, 2008.
- J. Li, G. Liu, and L. Wong. Mining statistically important equivalence classes and delta-discriminative emerging patterns. In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 430–439. 2007.
- W. Fan, K. Zhang, H. Cheng, J. Gao, X. Yan, J. Han, P. S. Yu, and O. Verscheure. Direct mining of discriminative and essential graphical and itemset features via model-based search tree. In Proceeding of the ACM SIGKDD international conference on knowledge discovery in databases, pages 230–238, 2008.
- S Nijssen, T Guns, L De Raedt, Correlated itemset mining in ROC space: a constraint programming approach, KDD 2009
- PK Novak, N Lavrac, GI Webb, Supervised Descriptive Rule Discovery: A Unifying Survey of Contrast Set, Emerging Pattern and Subgroup Mining The Journal of Machine Learning, 2009