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.
Data guided discovery - A new paradigm

“... data-intensive science [is] ... a new, fourth paradigm for scientific exploration.” - Jim Gray

McKinsey Global Institute

Big data: The next frontier for innovation, competition, and productivity

WIRED MAGAZINE: 16.07

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

By Chris Anderson 06.23.08
Great Opportunities to Solve Society’s Major Problems

- Improving health care and reducing costs
- Finding alternative/green energy sources
- Predicting the impact of climate change
- Reducing hunger and poverty by increasing agriculture production
Sample of Books on Mining Scientific Data
Mining Biomedical Data

- 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

Cost of sequencing has reduced dramatically
Source: www.synthesis.cc

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

4. Biomarkers for Mental Disorders using fMRI Data
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

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

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.

© Vipin Kumar
BPDM—April 27-28, 2012
**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

Enriched with the TNF/NFB signaling pathway which is well-known to be related to lung cancer
P-value: $1.4 \times 10^{-5}$ (6/10 overlap with the pathway)

[Fang et al PSB 2010]

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

(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:
  - GSE7390
  - GSE6532
  - GSE3494
  - GSE1456

- **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)
fMRI Biomarkers for Mental Disorders

Lynall et al. 2010
fMRI Biomarkers for Mental Disorders

Challenges/Opportunities

- Structure in the data
- Interactions between functional connections
- Heterogeneity
- Stability of the functional connections?
Stability of functional connections
Mining Global Eco-Climate Data

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
  - spatial and temporal autocorrelation.
  - Multi-scale/Multi-resolution nature

- Scalability
  - Size of Earth Science data sets can be very large,
  For example, for each time instance,
  - $2.5^\circ \times 2.5^\circ$: 10K locations for the globe
  - 250m x 250m: ~10 billion
  - 50m x 50m: ~250 billion

- High-dimensionality

- Noise and missing values

- Long-range spatial dependence

- Long memory temporal processes

- Nonlinear processes, Non-Stationarity

- Fusing multiple sources of data
Case Studies

1. Understanding climate change

2. Monitoring of global vegetation cover
Understanding Climate Change - Physics based Approach

General Circulation Models: Mathematical models with physical equations based on fluid dynamics

Parameterization and non-linearity of differential equations are sources for uncertainty!

Projection of temperature increase under different Special Report on Emissions Scenarios (SRES) by 24 different GCM configurations from 16 research centers used in the Intergovernmental Panel on Climate Change (IPCC) 4th Assessment Report.

A1B: “integrated world” balance of fuels
A2: “divided world” local fuels
B1: “integrated world” environmentally conscious

© Vipin Kumar
BPDM – April 27-28, 2012
Understanding Climate Change - Physics based Approach

General Circulation Models: Mathematical models with physical equations based on fluid dynamics

Parameterization and non-linearity of differential equations are sources for uncertainty!

Physics-based models are essential but not adequate

- Relatively reliable predictions at global scale for ancillary variables such as temperature
- Least reliable predictions for variables that are crucial for impact assessment such as regional precipitation

“The sad truth of climate science is that the most crucial information is the least reliable”

(Nature, 2010)
NSF Expedition: Understanding Climate Change - *A Data-Driven Approach*

**Project aim:**
A new and transformative data-driven approach that complements physics-based models and improves prediction of the potential impacts of climate change

---

**Transformative Computer Science Research**

- **Predictive Modeling**
  - Enable predictive modeling of typical and extreme behavior from multivariate spatio-temporal data

- **Complex Networks**
  - Enable studying of collective behavior of interacting eco-climate systems

- **High Performance Computing**
  - Enable efficient large-scale spatio-temporal analytics on exascale HPC platforms with complex memory hierarchies

- **Relationship Mining**
  - Enable discovery of complex dependence structures: non-linear associations or long range spatial dependencies

---

- **Science Contributions**
  - Data-guided uncertainty reduction by blending physics models and data analytics
  - A new understanding of the complex nature of the Earth system and mechanisms contributing to adverse consequences of climate change

- **Success Metric**
  - Inclusion of data-driven analysis as a standard part of climate projections and impact assessment (e.g., for IPCC)

---

"... data-intensive science [is] ... a new, fourth paradigm for scientific exploration." - Jim Gray
Some Driving Use Cases: Impact of Climate Change

On Hurricane Frequency, Intensity and Location

- Find non-linear relationships
- Validate w/ hindcasts
- Build hurricane models

On Intensity, Frequency, Duration and Distribution of Extreme Events

- Intensity of heat waves projected from CCSM3.0 climate model using A1F1 forcing for 2045-54 (top panel) and 2090-99 (bottom panel)

Abrupt Climate Change

1930's Dust Bowl

- Affected almost two-thirds of the U.S. Centered over the agriculturally productive Great Plains

Drought initiated by anomalous tropical SSTs (Teleconnections)

Discovering Climate Teleconnections

- Southern Oscillation's impact on land temperature

- Land Impact on Temperature
Monitoring Global Vegetation Cover: Motivation

**Forestry**
- Identify degradation in forest cover due to logging, conversions to cropland or plantations and natural disasters like fires.
- **Applications**: UN REDD+, national monitoring, reporting and verification systems, etc.

**Agriculture**
- Identify changes related to farmland, e.g. conversion to biofuels, changes in cropping patterns and changes in productivity.
- **Applications**: estimating regional food risks and ecological impact of agricultural practices.

**Urbanization**
- Identify scale, extent, timing and location of urbanization.
- **Applications**: policy planning, understanding impact on microclimate, water consumption, etc.
Traditional Approach for Land Cover Change Detection

- Two or more high quality satellite images acquired on different dates are compared for change identification.
- Images differ if a change has occurred between the two dates.

Limitations:
- High quality observations are infrequent in many parts of the world such as the tropics.
- Unable to detect changes outside the image acquisition window.
- Difficult to identify when the change has occurred.
- Parameters such as rate of change, extent, speed, and pattern of growth cannot be derived.
- Requires training data for each specific change of interest making it inherently unsuitable for global analysis.
Alternate Approach: Analyzing Vegetation Time Series

- Time series analysis can be used for
  - Identifying changes in land cover
  - Identifying when the change occurred i.e. the exact date of change

![Images](https://via.placeholder.com/150)

![Vegetation Time Series](https://via.placeholder.com/150)

© Vipin Kumar

BPDM—April 27-28, 2012
Alternate approach: Analyzing Vegetation Time Series

- Daily Remote Sensing observations are available from MODIS aboard AQUA and TERRA satellites.
  - High temporal frequency (daily for multi-spectral data and bi-weekly for the Vegetation index products like EVI, FPAR)

- Time series based approaches can be used for
  - Detection of a greater variety of changes.
  - Identifying when the change occurred
  - Characterization of the type of change eg. abrupt vs gradual
  - Near-real time change identification

- Challenges
  - Poor data quality and high variability
  - Coarse spatial resolution of observations (250 m)
  - Massive data sets: 10 billion locations for the globe

EVI shows density of plant growth on the globe.
EVI time series for a location
Novel Time Series Change Detection Techniques

Existing Time series change detection algorithms do not address unique characteristics of eco-system data like noise, missing values, outliers, high degree of variability (across regions, vegetation types, and time).

**Segmentation based approaches**
- Divide time series into homogenous segments.
- Boundary of segments become the change points.
- Useful for detection land cover conversions like forest to cropland, etc.

**Prediction based approaches**
- Build a prediction model for the location using previous observations.
- Use the deviation of subsequent observations from the predicted value by the model to identify changes/disturbances.
- Useful for detecting deviations from the normal vegetation model.

Automated Land change Evaluation, Reporting and Tracking System (ALERT)

- 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
Case Study 1:

Monitoring Global Forest Cover
Fires in Northern Latitude (Canada/ Russia) 2001-2009
Massive Fires in Canada have converted the forests into source of carbon in the atmosphere.
Logging in Canada

• Logging has produced clear cut areas in British Columbia, which can be identified as regular, generally rectangular shapes.

• The highly reflective clear cut areas stand out in marked contrast to the dark green forested areas.

(Source: NASA)
Deforestation in the 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 Deforestation Animation 2001-2009
Deforestation in the Amazon Rainforest: Comparison with PRODES

The blue polygons are deforestation changes marked by PRODES. Yellow dots are events detected by our algorithm.

PRODES is a system for monitoring deforestation in Brazilian Amazon.
Deforestation in the Amazon Rainforest: Comparison with PRODES

PRODES is a system for monitoring deforestation in Brazilian Amazon.

The blue polygons are deforestation changes marked by PRODES.
Yellow dots are events detected by our algorithm.
Gold Mine in Protected Forest, Tanzania
Reforestation near Guangting Reservoir, China

• These reforestation events are around Guangting Reservoir, a reservoir around 100 miles away from Beijing.

• Around 20 years ago, Guanting Reservoir used to play an important role of serving water for people in Beijing and Zhangjiakou.

• The environment around the reservoir got polluted after years, due to lack of protection.

• It is located very close to Beijing and plays an important role, therefore the government began to give a comprehensive treatment for this area.

• Part of the treatment is planting trees around Guangting Reservoir which started in 2003 and is still going on.

News Articles:

http://news.china.com.cn/rollnews/2010-06/04/content_2514320.htm
Detecting other land cover changes

Shrinking of Lake Chad, Nigeria

Damage to vegetation by hurricane Katrina

Flooding along Ob River, Russia

Farm abandonment in Zimbabwe during political conflict between 2004 and 2008.
ALERT Platform
Impact on REDD+

“The [Peru] government needs to spend more than $100m a year on high-resolution satellite pictures of its billions of trees. But … a computing facility developed by the Planetary Skin Institute (PSI) … might help cut that budget.”

“ALERTS, which was launched at Cancún, uses … data-mining algorithms developed at the University of Minnesota and a lot of computing power … to spot places where land use has changed.”

- The Economist 12/16/2010
Monitoring Forest Cover Change: Challenges Ahead

Designing robust change detection algorithms

Characterization of land cover changes

Multi-resolution analysis (250m vs 1km vs 4km)
  - Different kinds of changes are visible at different scales

Multivariate analysis
  - Detecting some types of changes (e.g. crop rotations) will require additional variables.

Data quality improvement
  - Preprocessing of data using spatio-temporal noise removal and smoothing techniques can increase performance of change detection.

Incremental update and Real-time detection

Spatial event identification

Spatial-Temporal Querying

Applications in variety of domains:
  - Climate, agriculture, energy
  - Economics, health care, network traffic
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, Gang Fang, Karsten Steinhaeuser, Gowtham Atluri, Varun Mithal, Vanja Paunic, Sanjoy Dey, Sean Landman, Wen Wang, Marc Dunham, Matt Kappel, Ivan Brugere, Anuj Karpatne, Xi Chen, Ayush Singhal, Sanyam Mehta

Biomedical Informatics:
Brian Van Ness, Bill Oetting, Gary L. Nelsestuen, Christine Wendt, Piet C. de Groen, Michael Wilson, Rui Kuang, Chad Myers, Angus McDonald III, Kelvin Lim

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).


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


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


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.


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, 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 npre.2008.2184.1)

Rohit Gupta, Gang Fang, Blayne Field, Michael Steinbach and Vipin Kumar, Quantitative Evaluation of Approximate Frequent Pattern Mining Algorithms,

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


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, 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


S Nijssen, T Guns, L De Raedt, Correlated itemset mining in ROC space: a constraint programming approach, KDD 2009