Mining Eco-Climate Data

Vipin Kumar
University of Minnesota

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

Research Supported by NSF, NOAA, NASA, Planetary Skin Institute,
Minnesota Futures Program

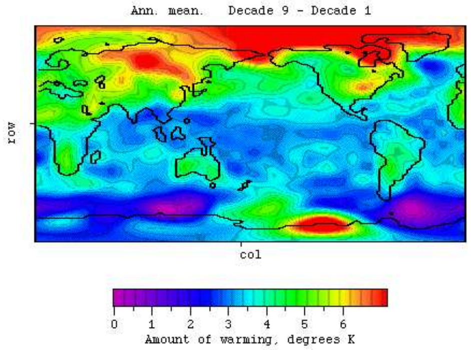
Mining Eco-Climate Data

Science Goal: Understand global scale patterns in biosphere processes

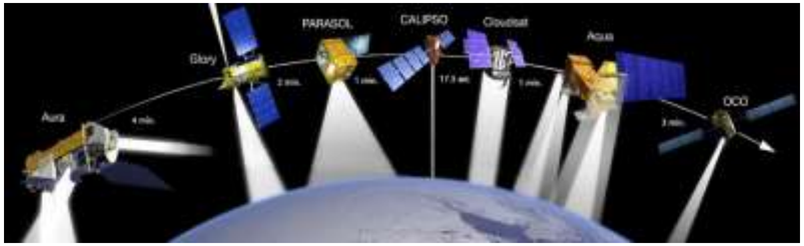
Earth Science Questions:

- 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

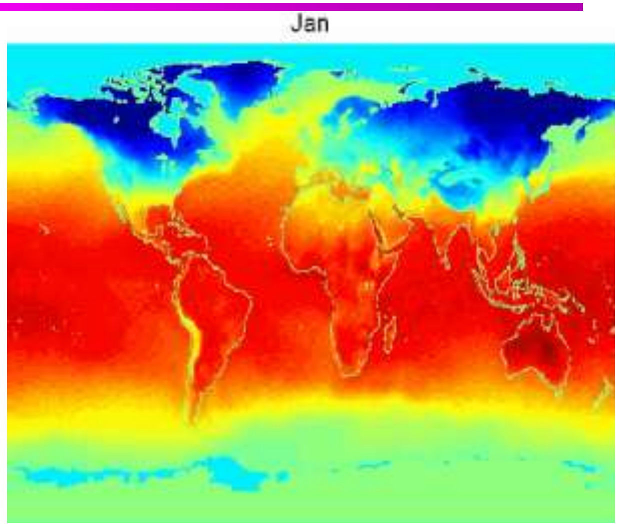


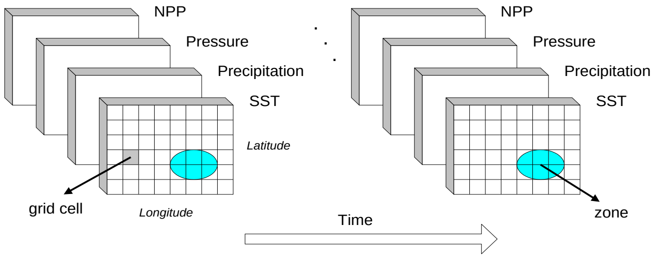
Data guided processes can complement hypothesis guided data analysis to develop predictive insights for use by climate scientists, policy makers and community at large.



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° x 2.5° :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

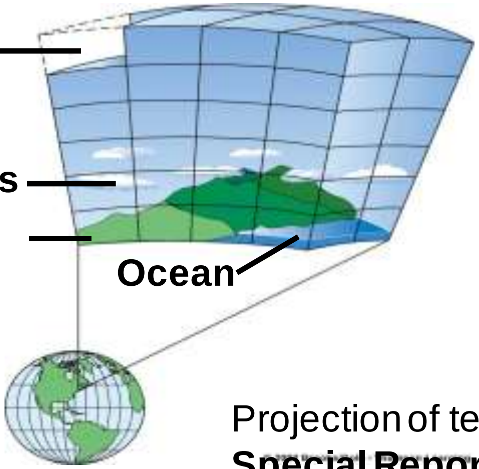
1. Understanding climate change

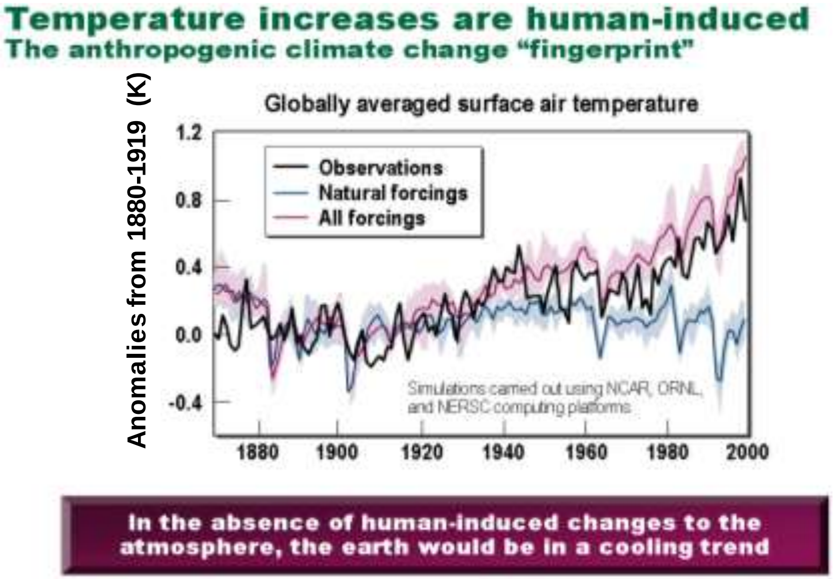
2. Monitoring of global vegetation cover

Understanding Climate Change - Physics based Approach

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

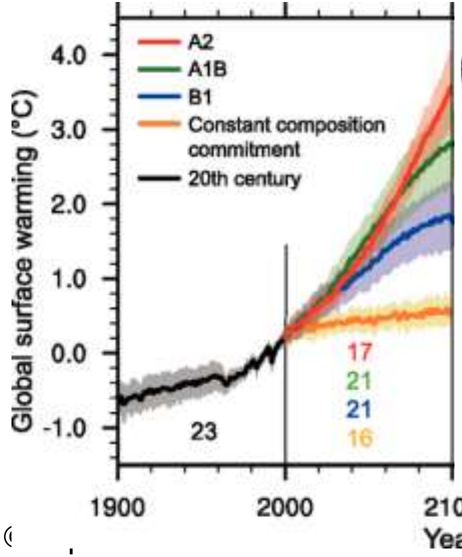
Parameterization and non-linearity of differential equations are sources for uncertainty!




Temperature increases are human-induced
The anthropogenic climate change "fingerprint"

In the absence of human-induced changes to the atmosphere, the earth would be in a cooling trend

Figure Courtesy: ORNL



IPCC (2007)

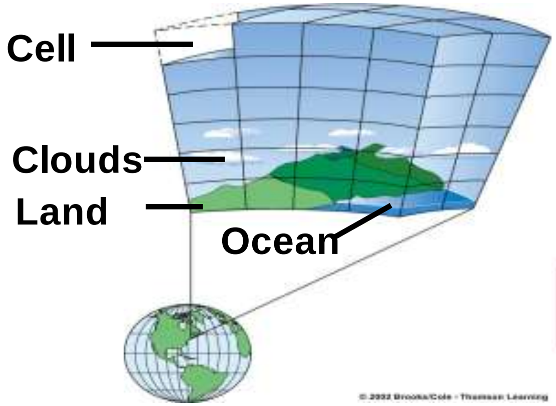
Projection of temperature increase under different **Special Report on Emissions Scenarios** (SRES) by 24 different GCM configurations from 16 research centers used in the **Intergovernmental Panel on Climate Change** (IPCC) 4th Assessment Report.

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

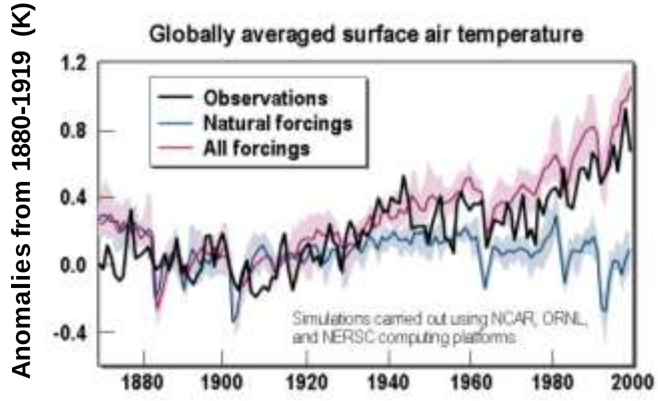
Understanding Climate Change - Physics based Approach

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

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



Temperature increases are human-induced
The anthropogenic climate change "fingerprint"



In the absence of human-induced changes to the atmosphere, the earth would be in a cooling trend

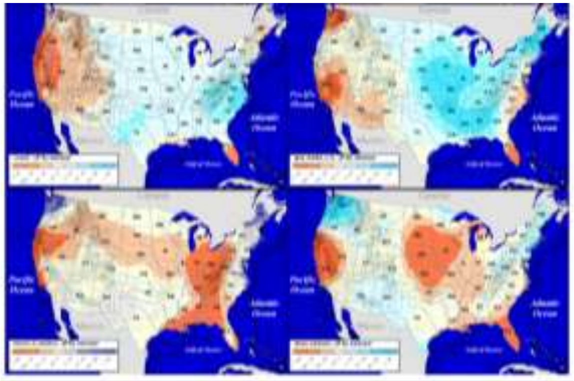
Figure Courtesy: ORNL

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)

Disagreement between IPCC models



Regional hydrology exhibits large variations among major IPCC model projections

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

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

Transformative Computer Science Research

Predictive Modeling

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

Relationship Mining

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

Complex Networks

Enable studying of collective behavior of interacting eco-climate systems

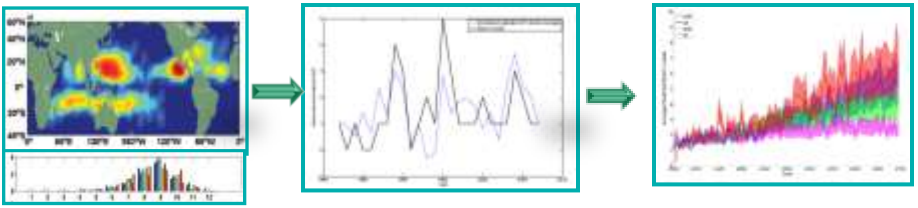
High Performance Computing

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

- *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)

Some Driving Use Cases : Impact of Climate Change

On Hurricane Frequency, Intensity and Location



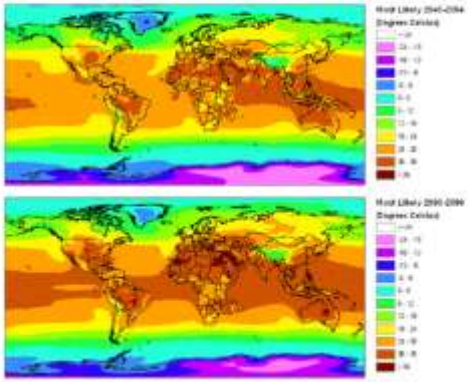
Find non-linear relationships

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)

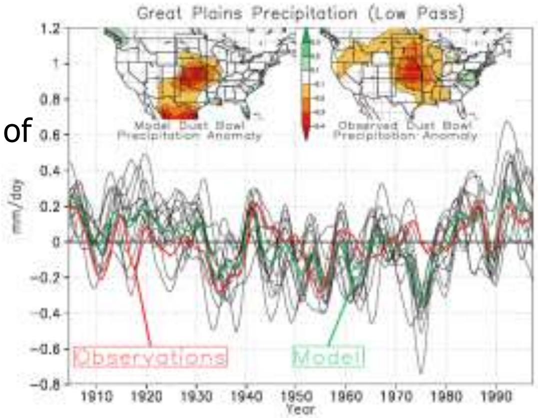


Abrupt Climate Change

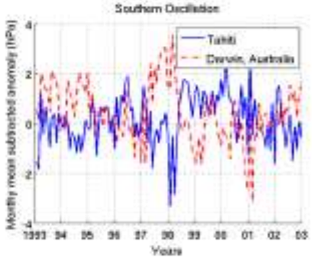
1930's Dust Bowl

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

Drought initiated by anomalous tropical SSTs (Teleconnections)

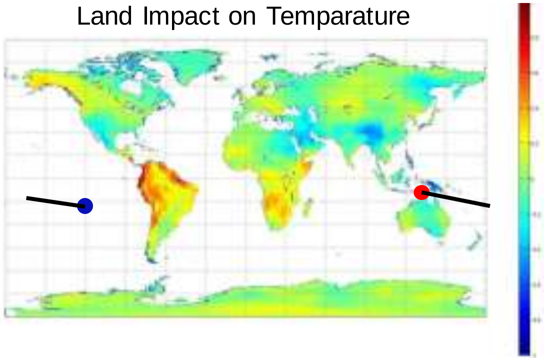


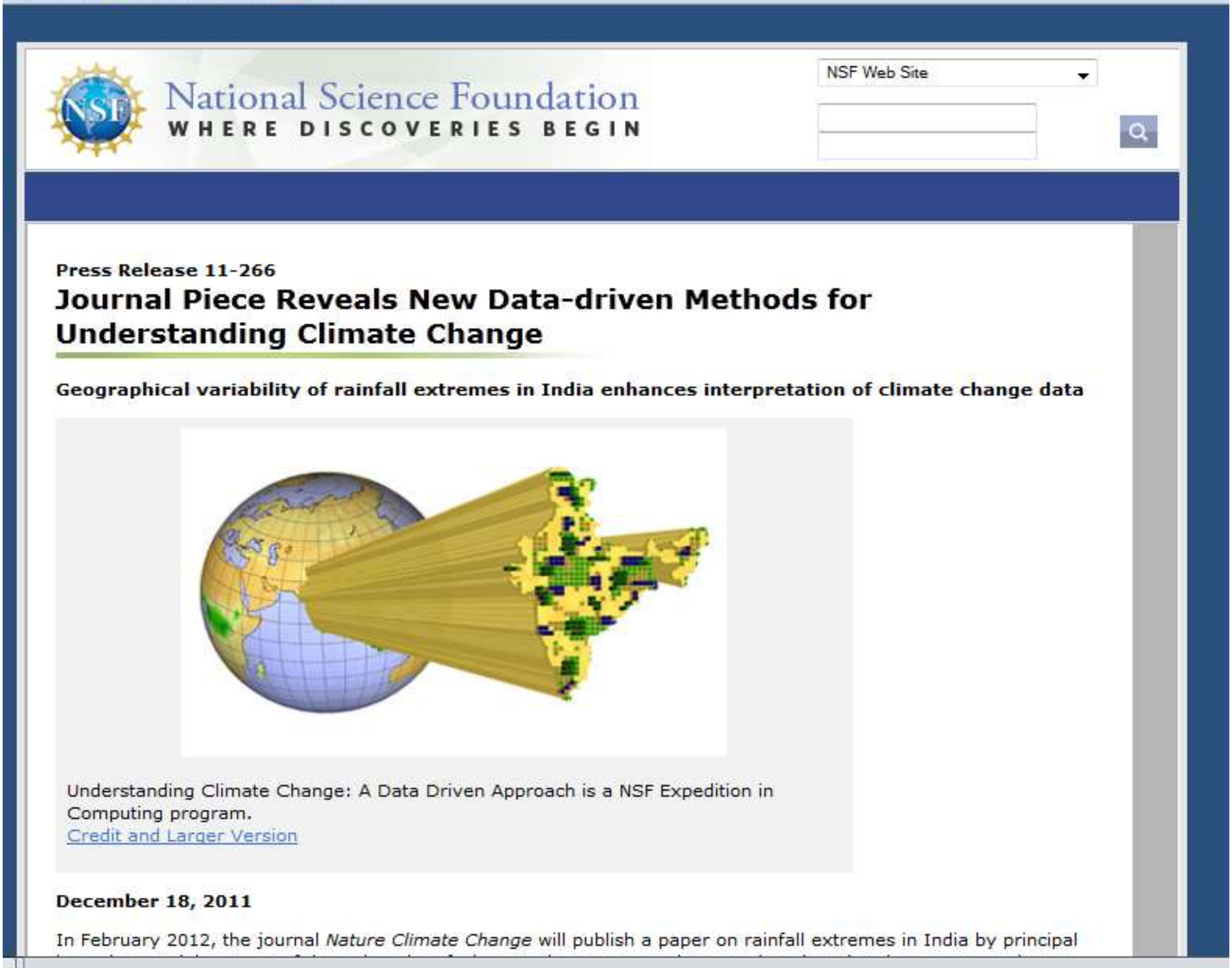
Discovering Climate Teleconnections



Southern Oscillation's impact on land temperature

Land Impact on Temparature





National Science Foundation
WHERE DISCOVERIES BEGIN

Press Release 11-266

Journal Piece Reveals New Data-driven Methods for Understanding Climate Change

Geographical variability of rainfall extremes in India enhances interpretation of climate change data

Understanding Climate Change: A Data Driven Approach is a NSF Expedition in Computing program.
Credit and Larger Version

December 18, 2011

In February 2012, the journal *Nature Climate Change* will publish a paper on rainfall extremes in India by principal

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

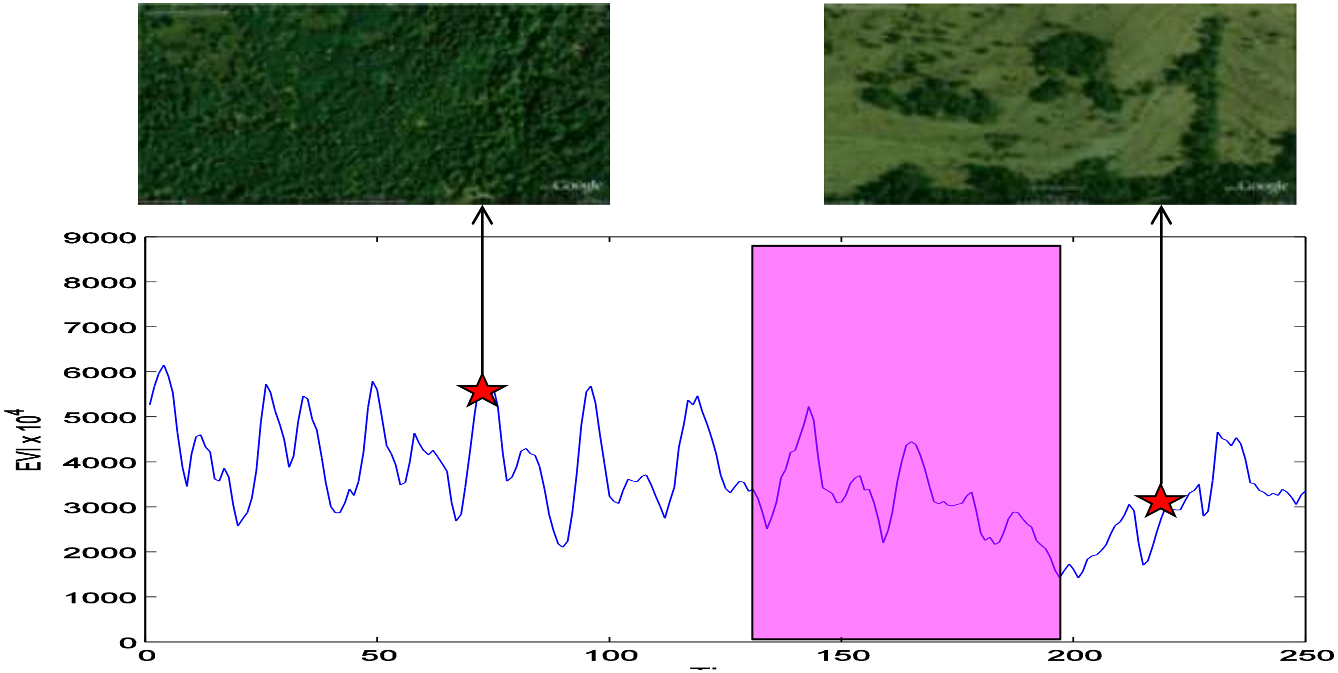
- Two or more high quality satellite images acquired on different dates are compared for change identification.
- Images differ if a change has occurred between the two dates.

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

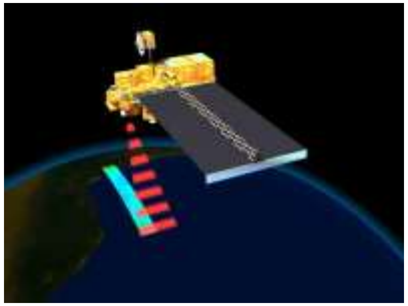
- Time series analysis can be used for
  - Identifying changes in land cover
  - Identifying when the change occurred i.e. the exact date of change

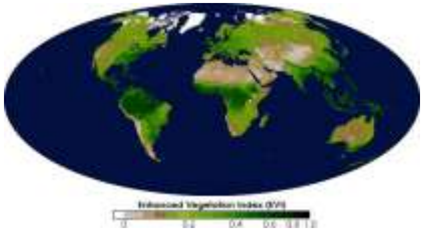


Images

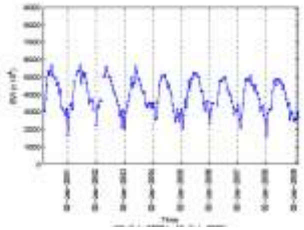
Vegetation Time Series

Alternate approach: Analyzing Vegetation Time Series



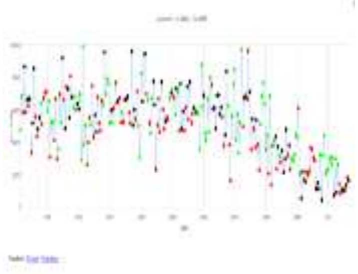


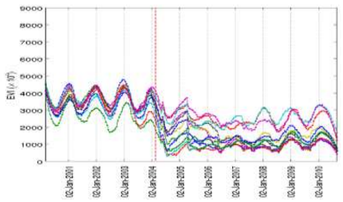
EVI shows density of plant growth on the globe.

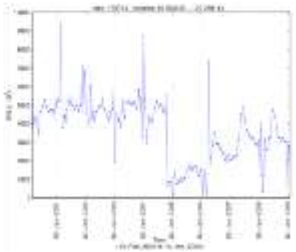


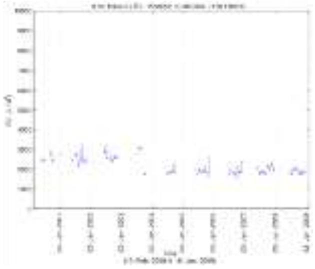
EVI time series for a location

- 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



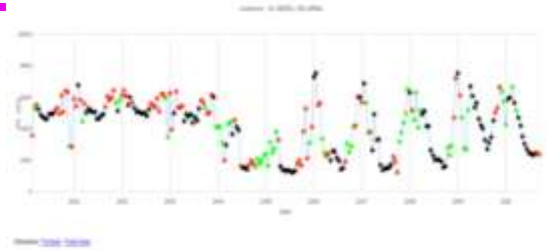






Novel Time Series Change Detection Techniques

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).



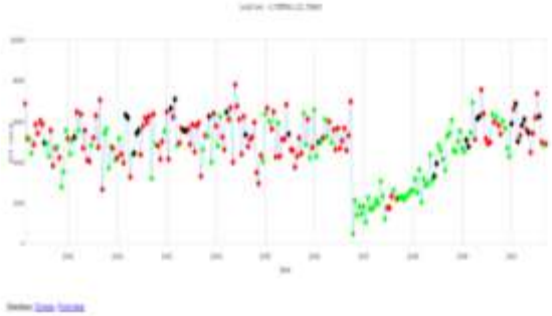
EVI time series for a 250 m by 250 m of land in Iowa, USA that changed from fallow land to agriculture land.

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

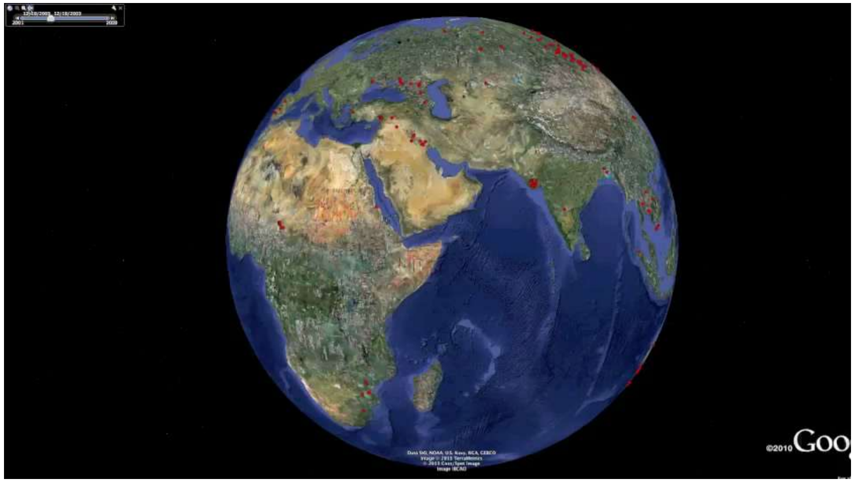
- Build a prediction model for the location using previous observations.
- Use the deviation of subsequent observations from the predicted value by the model to identify changes/disturbances.
- Useful for detecting deviations from the normal vegetation model.



FPAR time series for a forest fire location in California, USA.

- S. Boriah, V. Kumar, M. Steinbach, et al., *Land cover change detection: a case study*, *KDD* 2008.
- V. Mithal, S. Boriah, A. Garg, M. Steinbach, V. Kumar et al., *Monitoring global forest cover using data mining. ACM* Transactions on Intelligent Systems and Technology, 2011 (In Press)

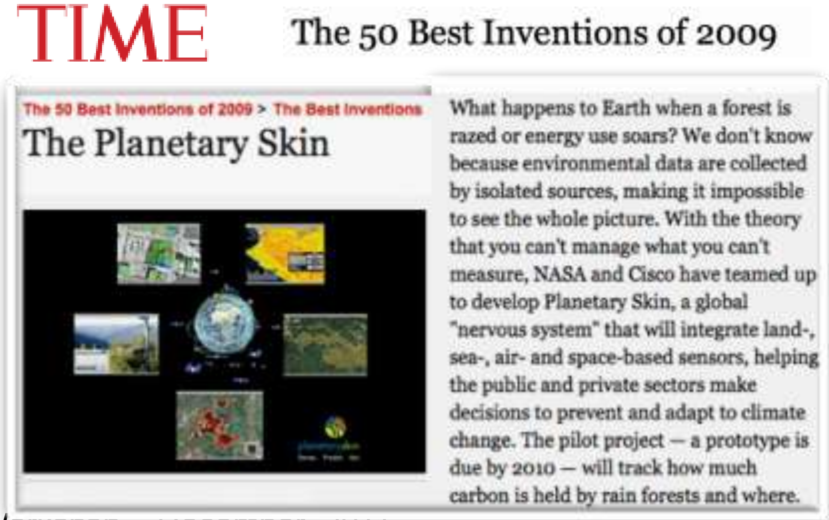
Automated Land change Evaluation, Reporting and Tracking System (ALERT)



- 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



TIME

The 50 Best Inventions of 2009

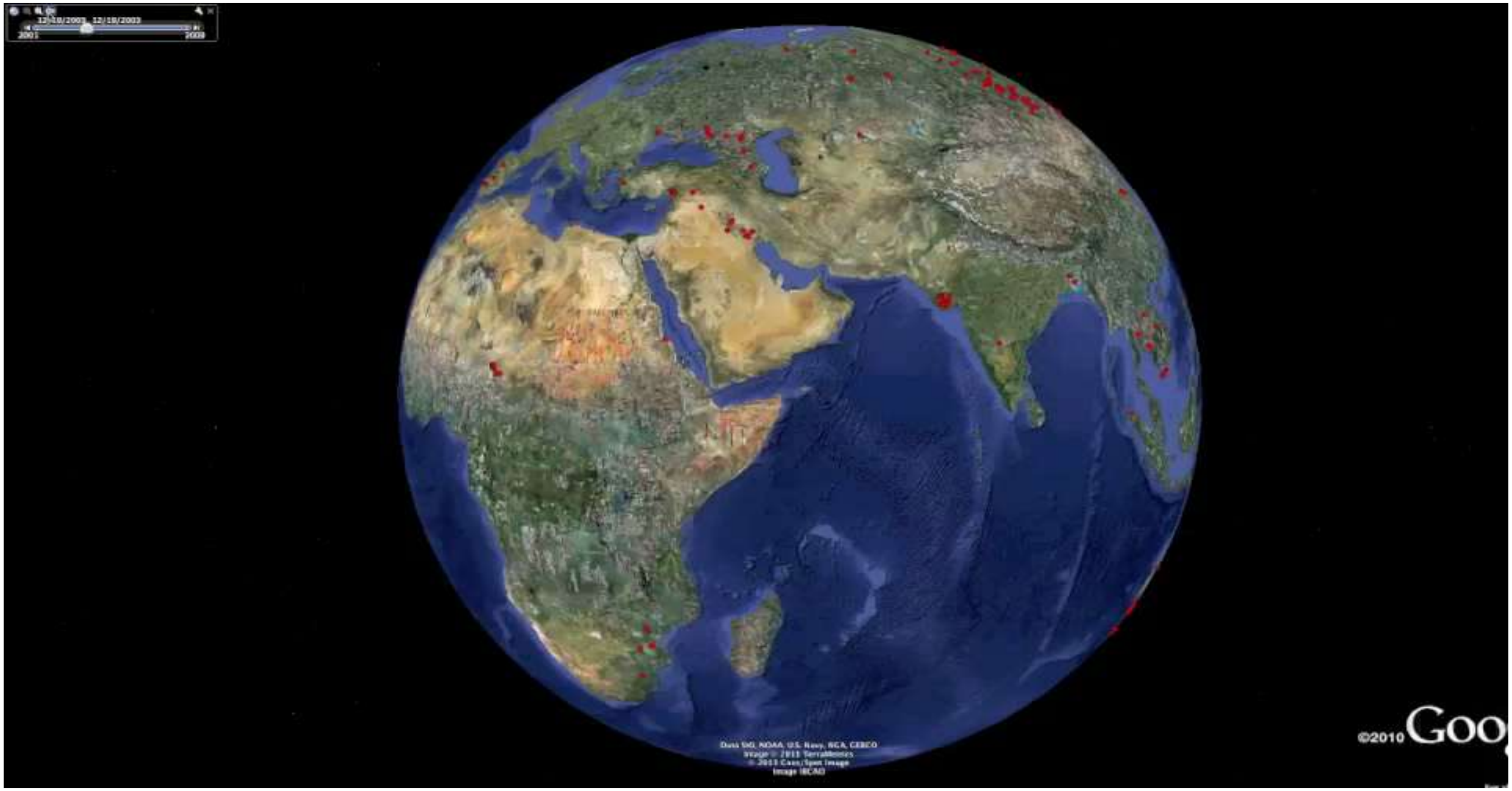
The 50 Best Inventions of 2009 > The Best Inventions

The Planetary Skin

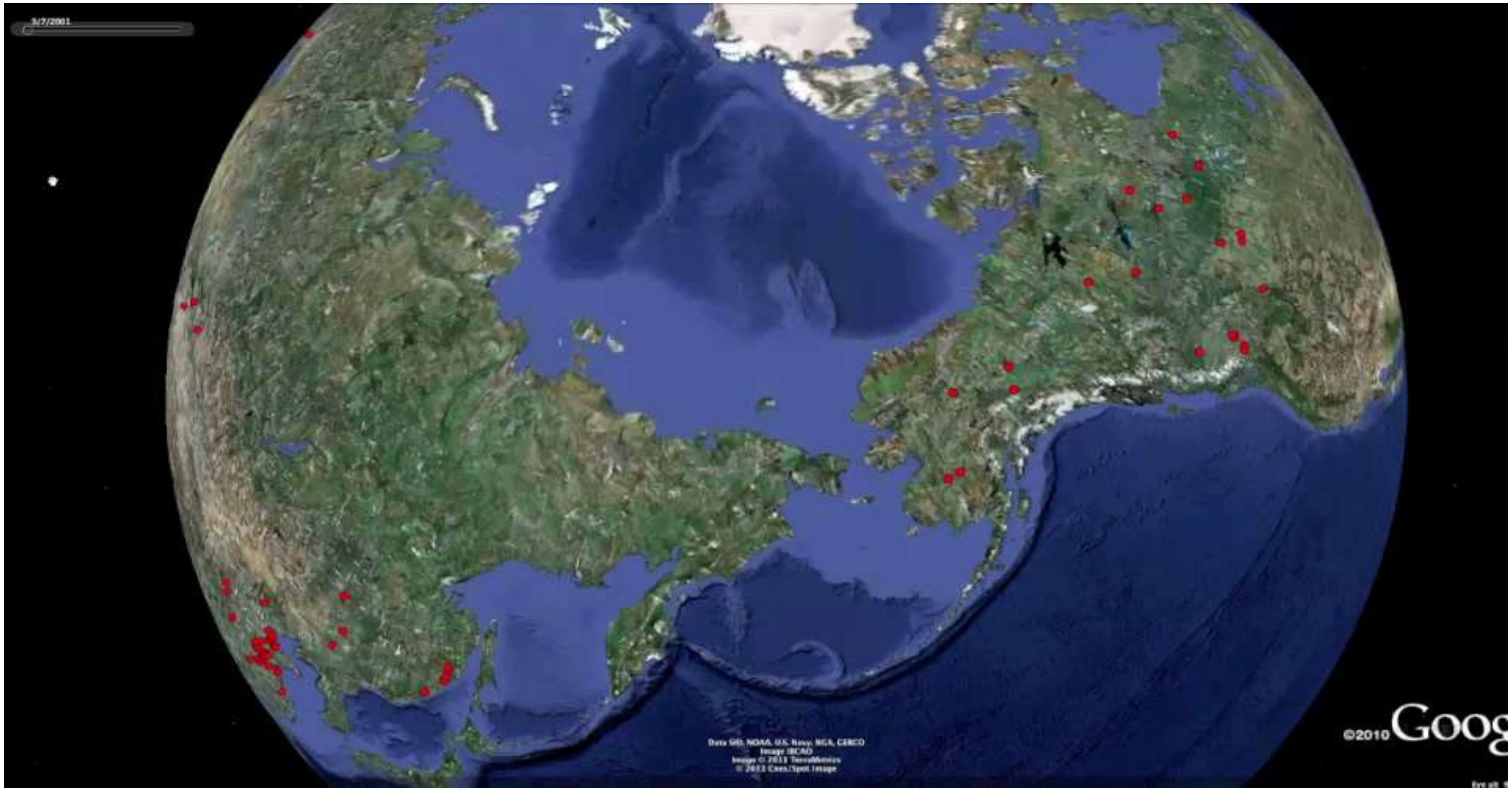
What happens to Earth when a forest is razed or energy use soars? We don't know because environmental data are collected by isolated sources, making it impossible to see the whole picture. With the theory that you can't manage what you can't measure, NASA and Cisco have teamed up to develop Planetary 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.

Case Study 1:

Monitoring Global Forest Cover

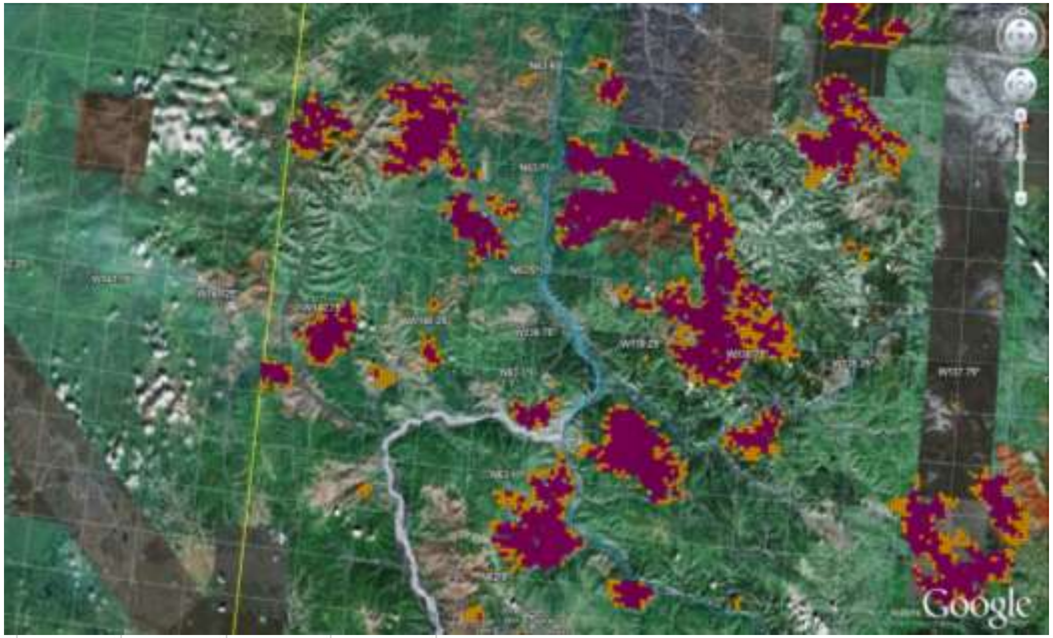


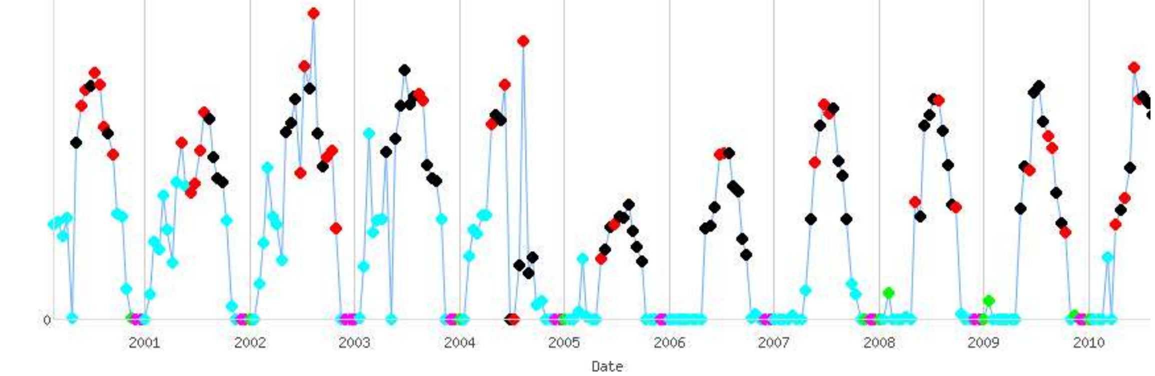
Fires in Northern Latitude (Canada/Russia) 2001-2009


5/7/2001
Data SIO, NOAA, U.S. Navy, NGA, GEBCO
Image IBCAO
Image © 2011 TerraMetrics
© 2011 Cnes/Spot Image
©2010 Goog

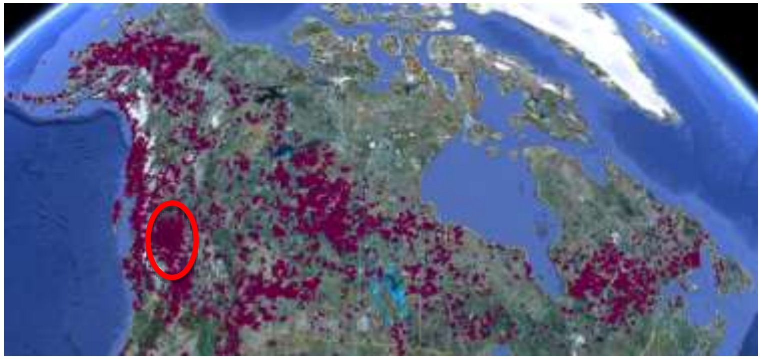
Forest Fires in Canada

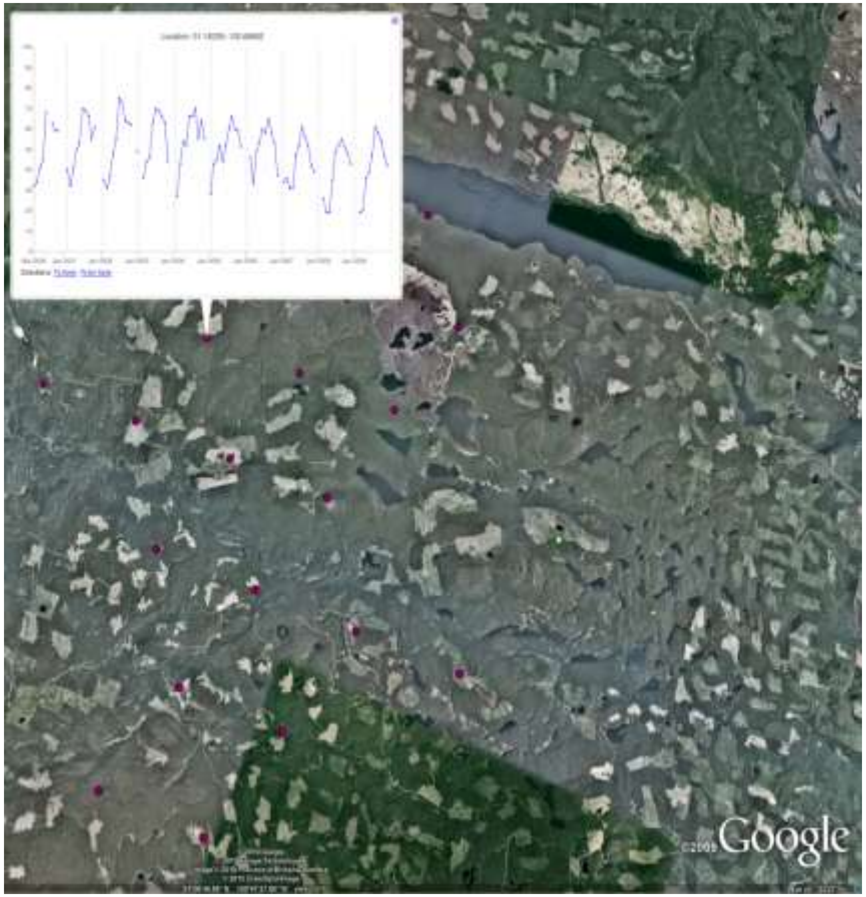
Massive Fires in Canada have converted the forests into source of carbon in the atmosphere.





Logging in Canada



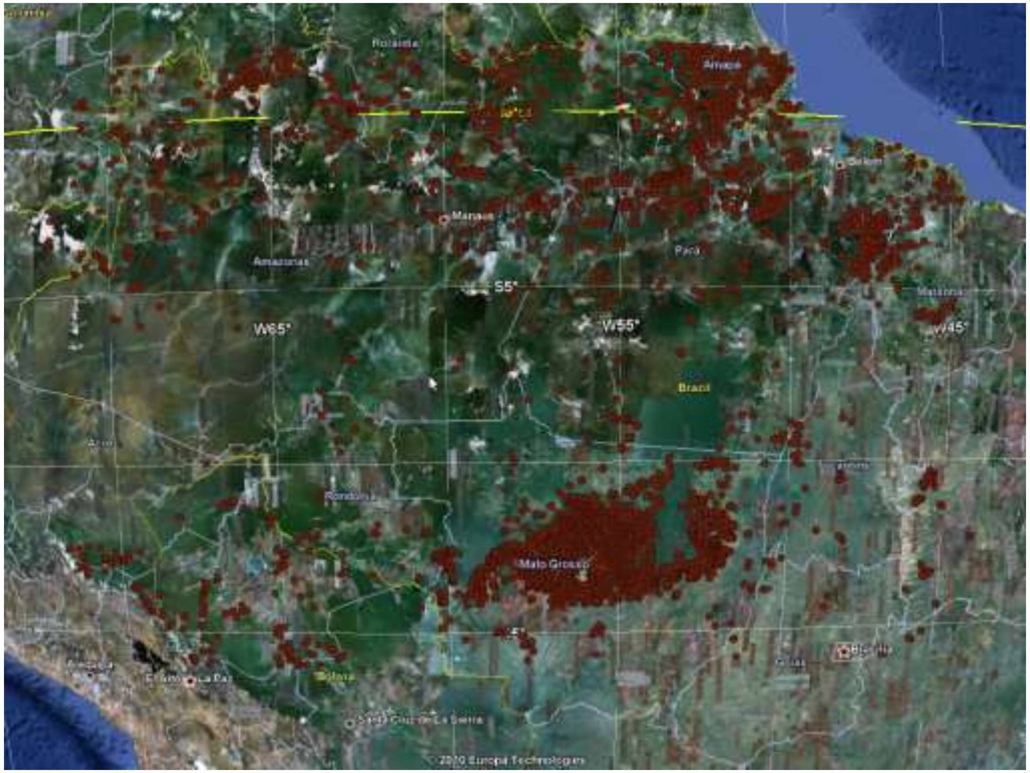


- Logging has produced clear cut areas in British Columbia, which can be identified as regular, generally rectangular shapes.

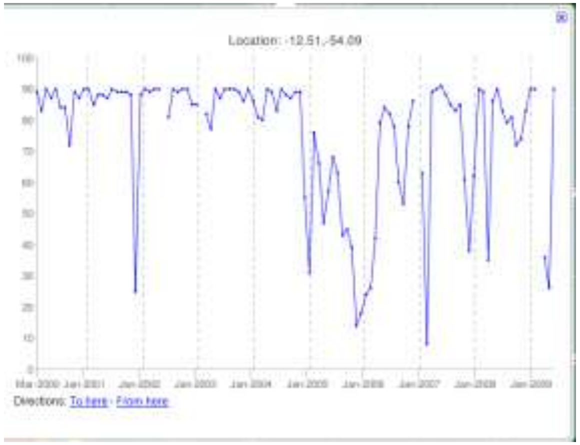
- The highly reflective clear cut areas stand out in marked contrast to the dark green forested areas.

(Source: NASA)

Deforestation in the Amazon Rainforest



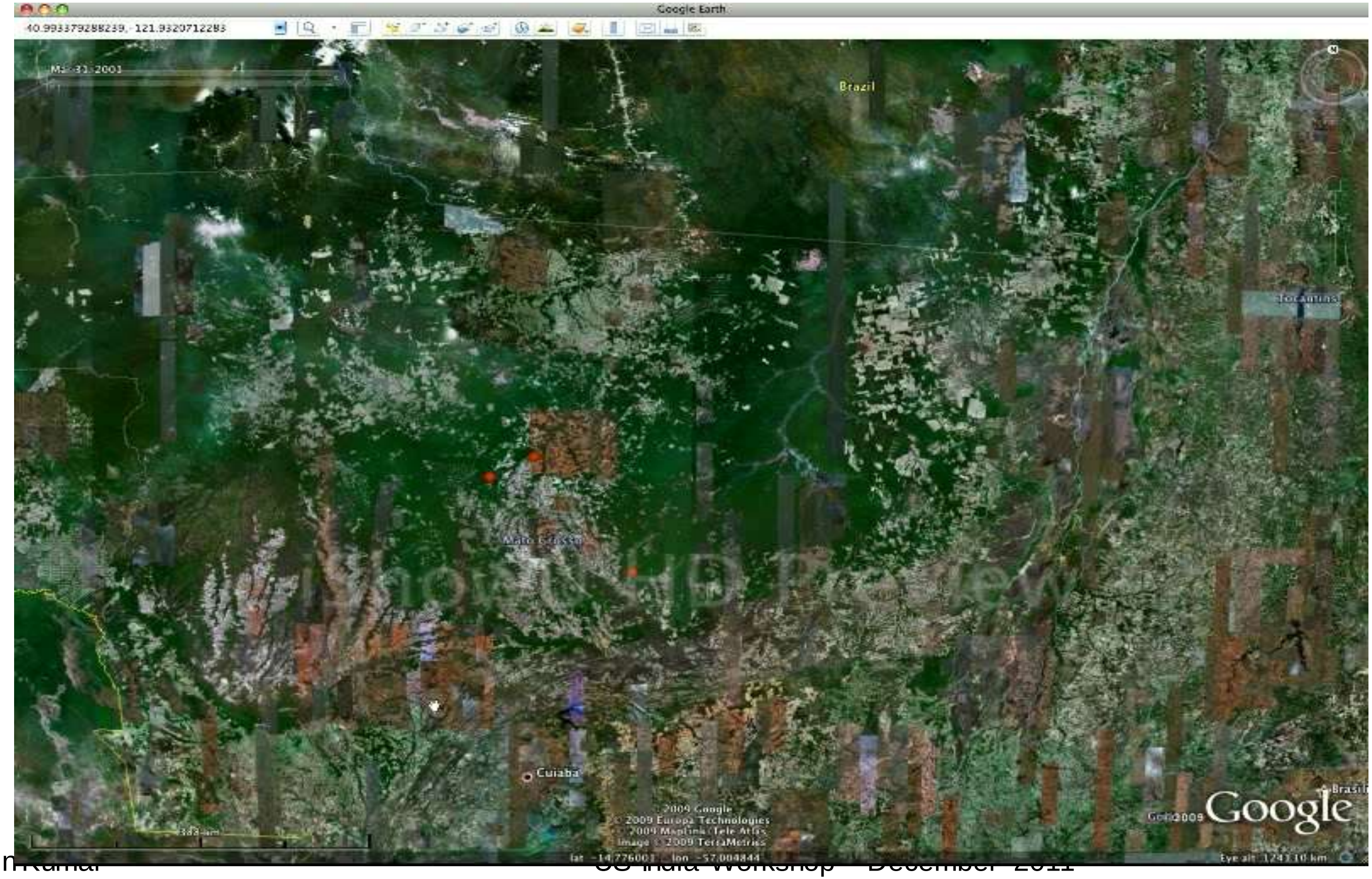
Brazil Accounts for almost 50% of all humid tropical forest clearing, nearly 4 times that of the next highest country, which accounts for 12.8% of the total.



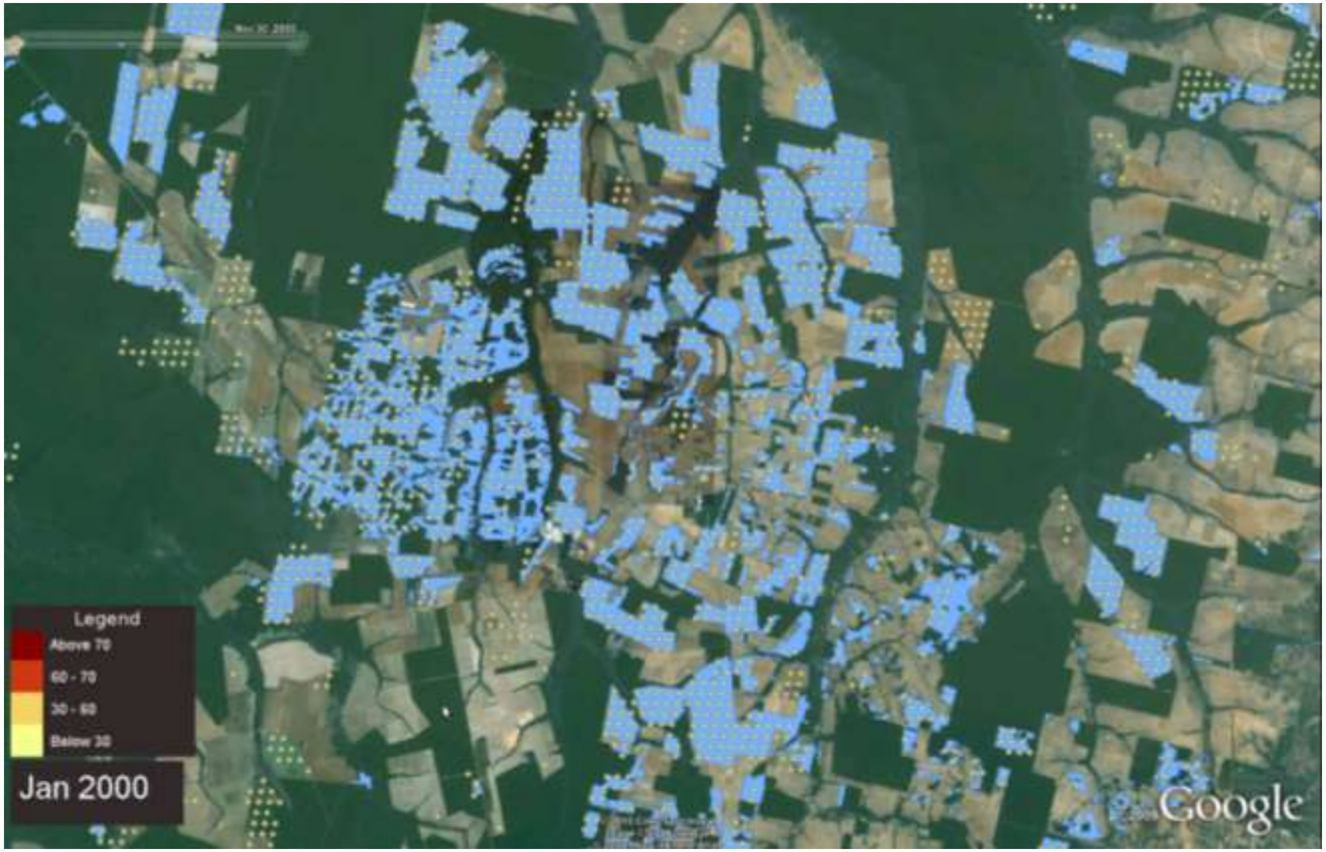
© Vipin Kumar

US India Workshop – December 2011

Amazon Deforestation Animation 2001-2009



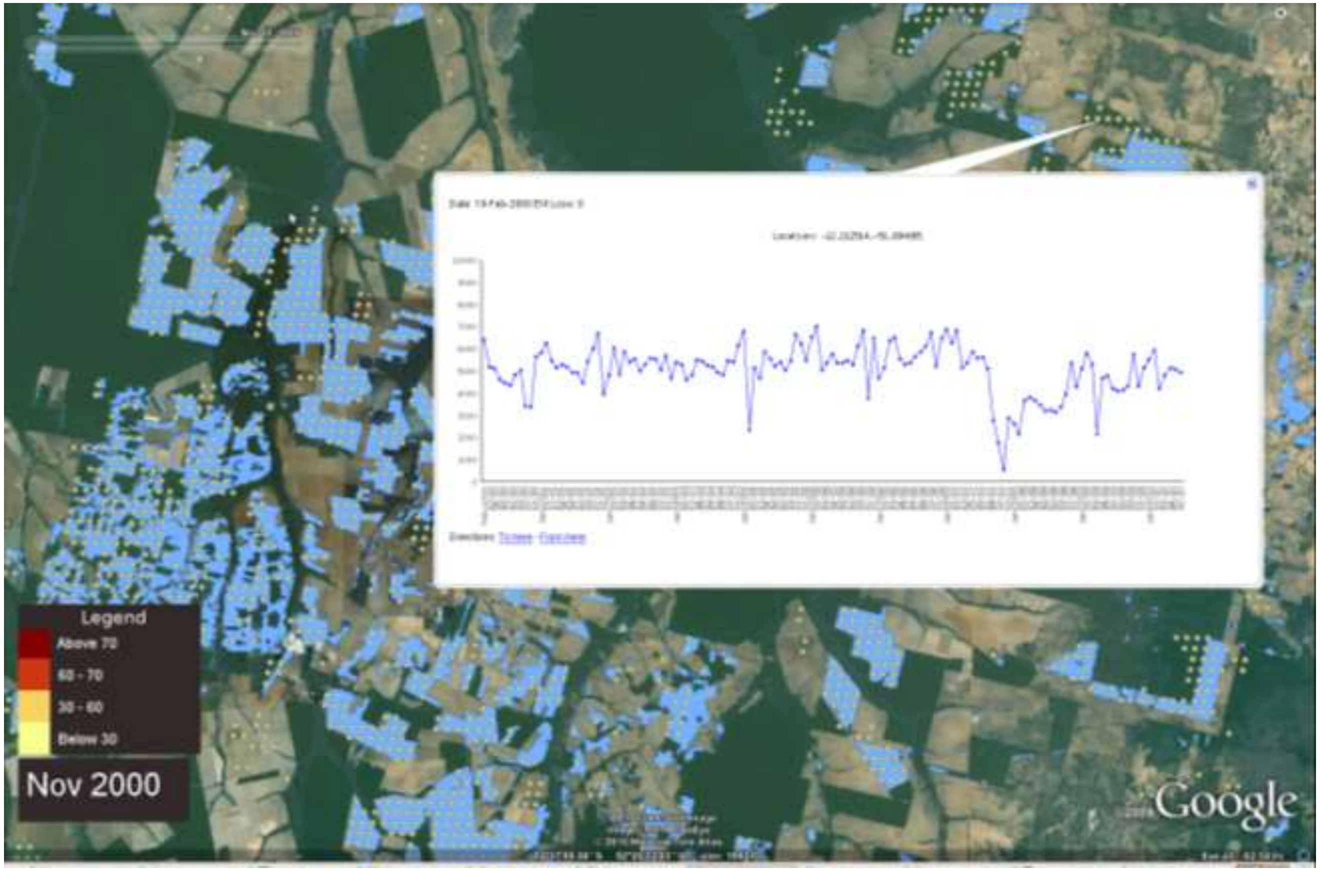
Deforestation in the Amazon Rainforest: Comparison with PRODES



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.

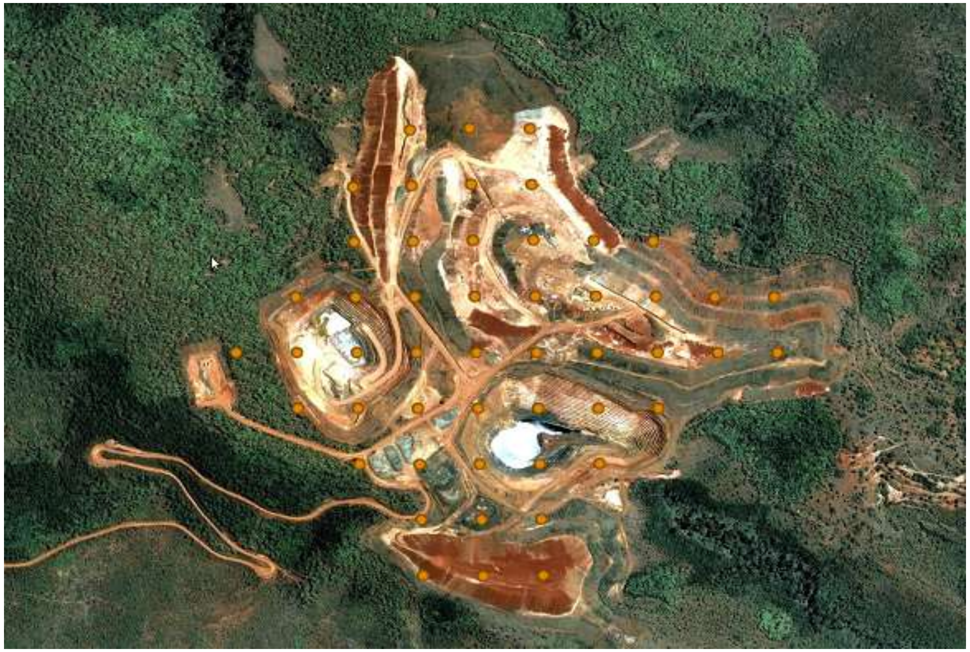
Deforestation in the Amazon Rainforest: Comparison with PRODES

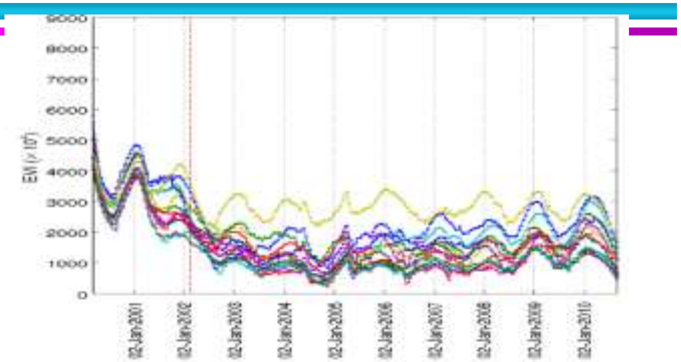


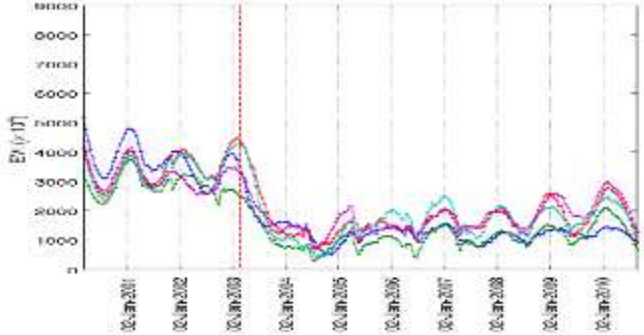
PRODES is a system for monitoring deforestation in Brazilian Amazon.

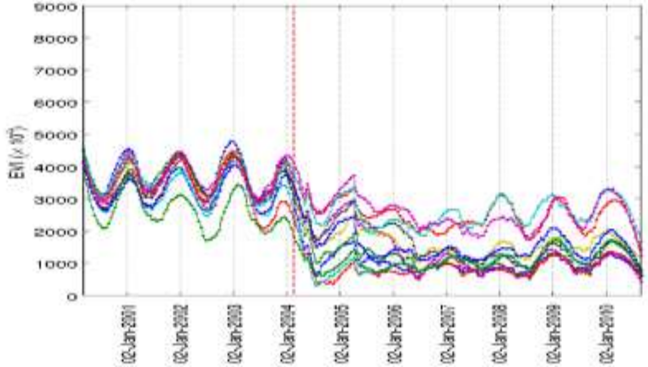
The blue polygons are deforestation changes marked by PRODES.
Yellow dots are events detected by our algorithm.

Gold Mine in Protected Forest , Tanzania

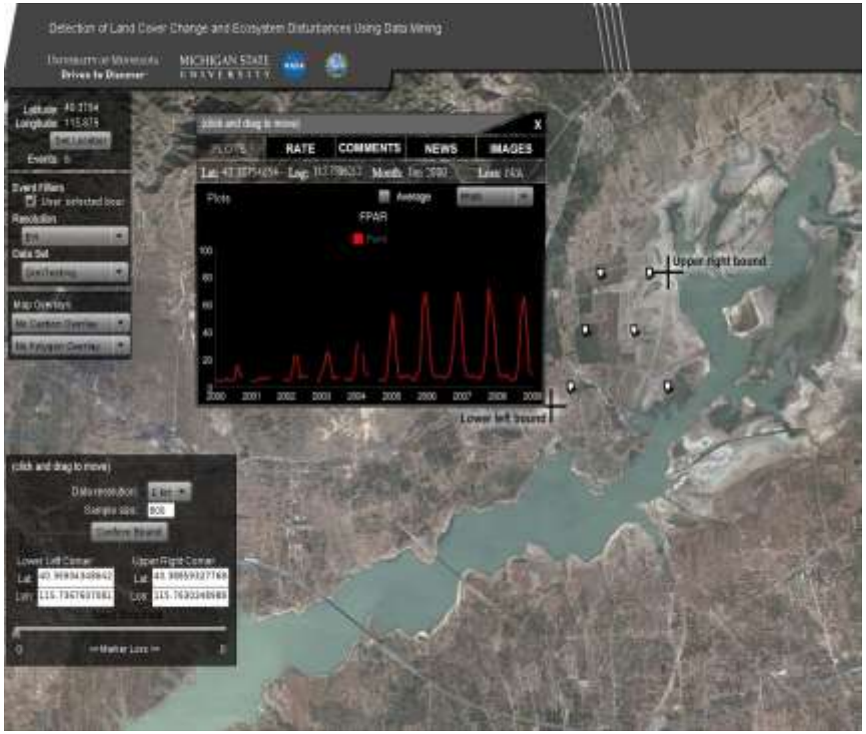








Reforestation near Guangting Reservoir, China



•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.

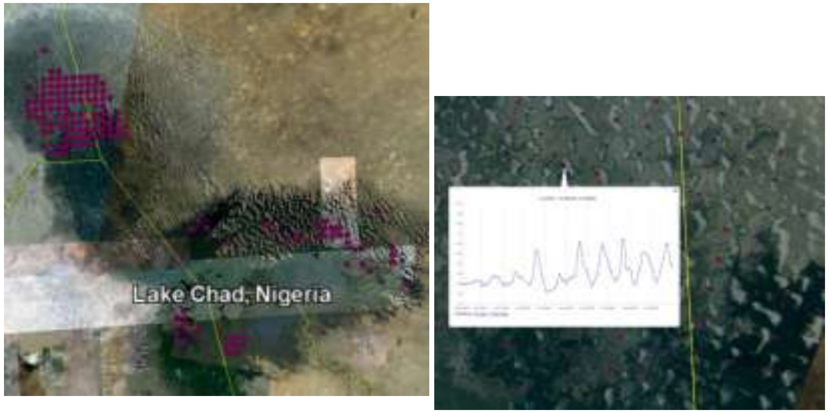
•It is located very close to Beijing and plays an important role, therefore the government began to give a comprehensive treatment for this area.

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

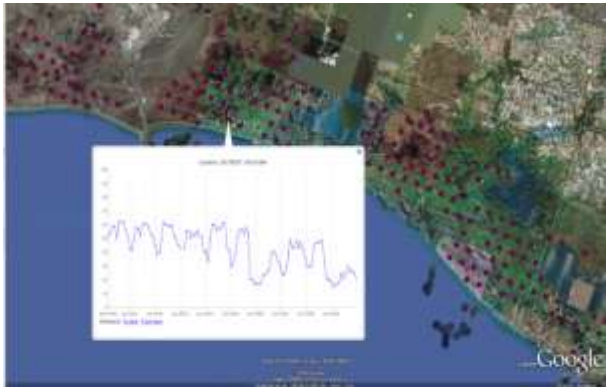
News Articles:

- http://www.yzhbw.net/news/shownews-22_510.aspx
- http://news.china.com.cn/rollnews/2010-06/04/content_2514320.htm

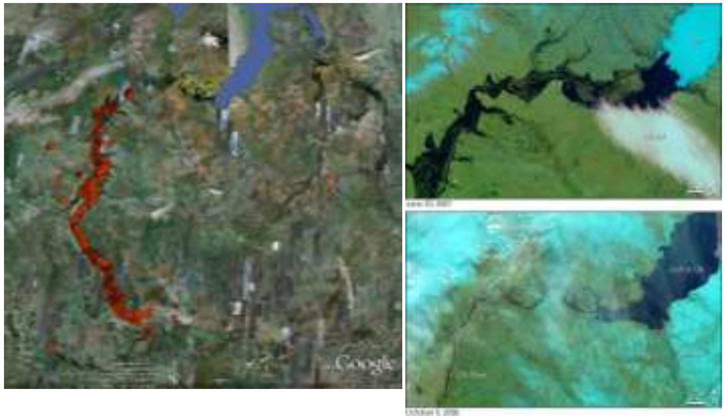
Detecting other land cover changes



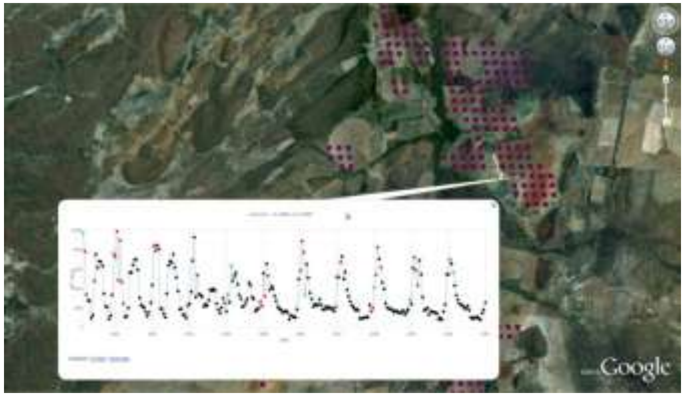
Shrinking of Lake Chad, Nigeria



Damage to vegetation by hurricane Katrina



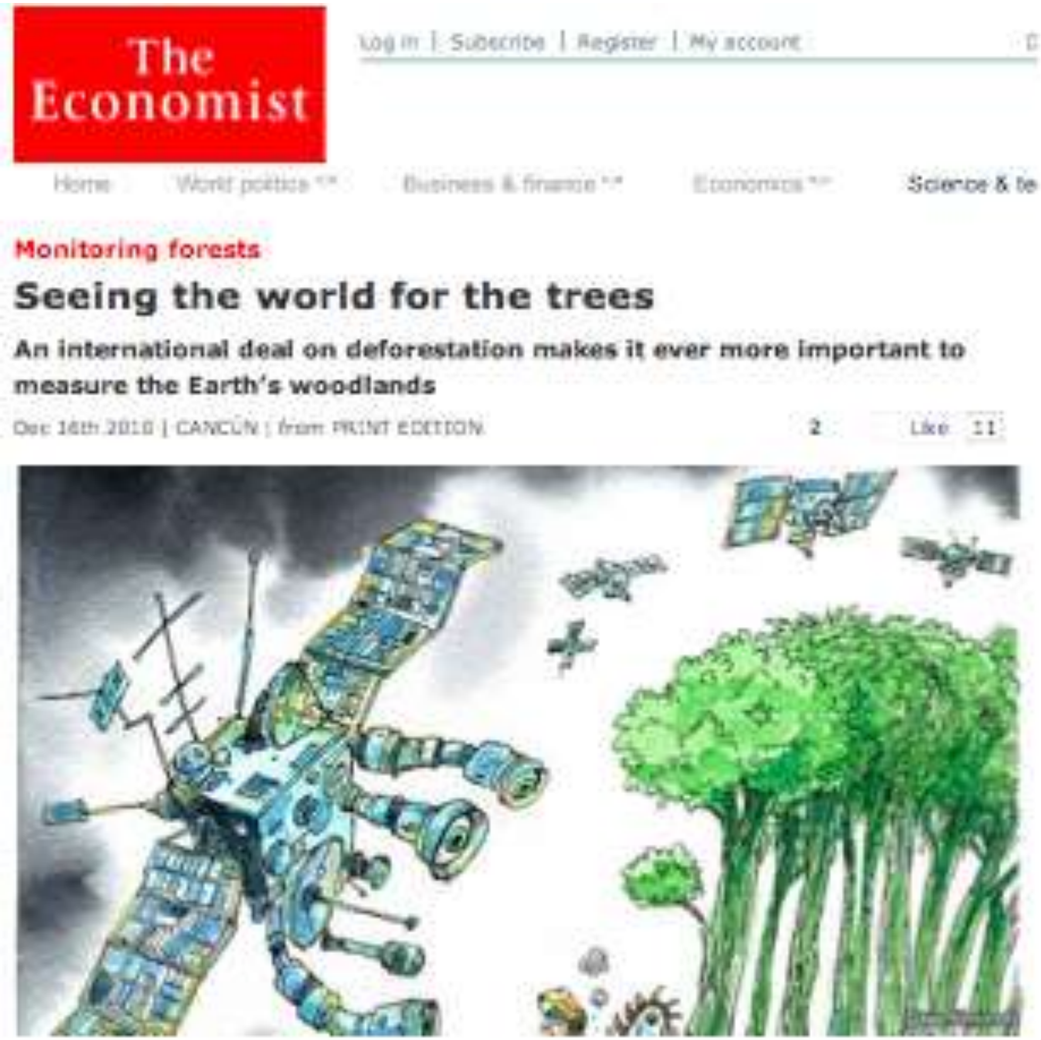
Flooding along Ob River, Russia



Farm abandonment in Zimbabwe during political conflict between 2004 and 2008.

ALERT Platform

Impact on REDD+



The Economist

Log in | Subscribe | Register | My account

Home   World politics   Business & finance   Economics   Science & te

Monitoring forests

Seeing the world for the trees

An international deal on deforestation makes it ever more important to measure the Earth's woodlands

Dec 16th 2010 | CANCÚN | from PRINT EDITION   2   Like 11

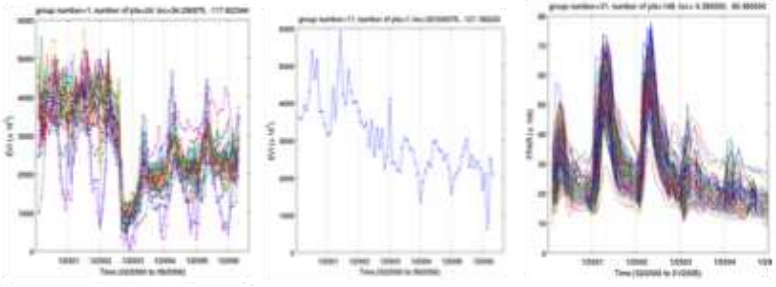
"The [Peru] government needs to spend more than $100m a year on high-resolution satellite pictures of its billions of trees. But … a computing facility developed by the Planetary Skin Institute (PSI) … might help cut that budget."

"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."

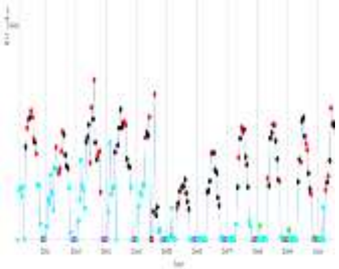
- The Economist 12/16/2010

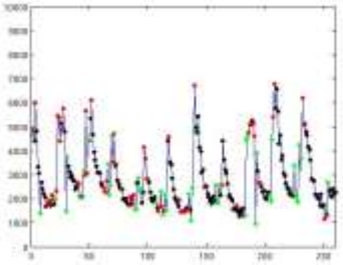
Monitoring Forest Cover Change: Challenges Ahead

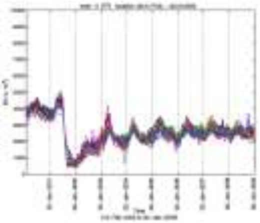
- Designing robust change detection algorithms
- Characterization of land cover changes
- Multi-resolution analysis (250m vs 1km vs 4km)
  - Different kinds of changes are visible at different scales
- Multivariate analysis
  - Detecting some types of changes (e.g. crop rotations) will require additional 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



(a) Fire (b) Logging (c) Drought







Summary

- Data driven discovery methods hold great promise for advancement in a variety of scientific disciplines
- Challenges arise due to the complex nature of eco-climate data sets
    - 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

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).
- 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