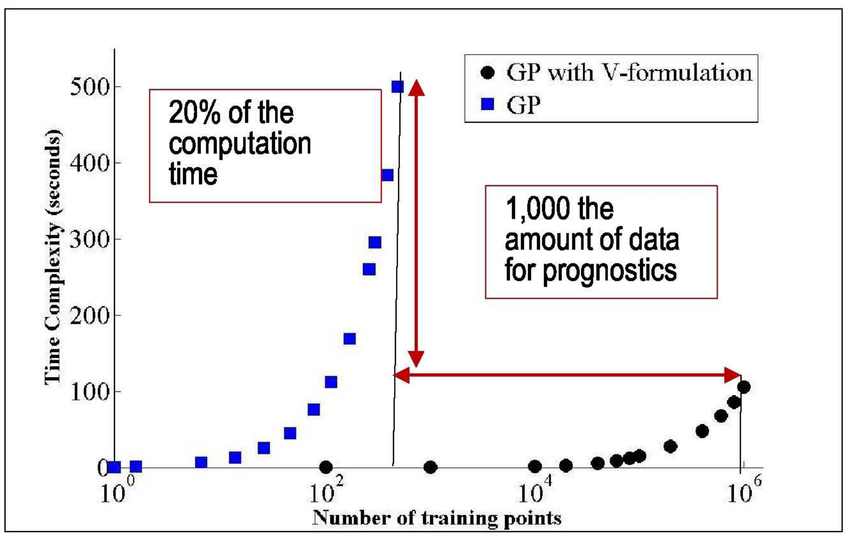
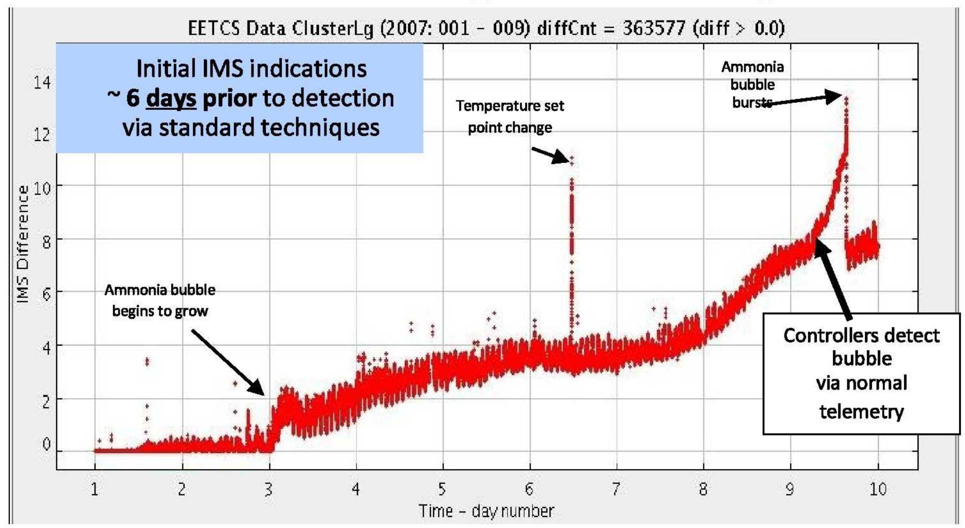
Data Mining at NASA: from Theory to Applications

Ashok N. Srivastava, Ph.D.
Principal Investigator, IVHM Project
Group Lead, Intelligent Data Understanding
ashok.n.srivastava@nasa.gov

Intelligent Data Understanding Group

The IDU group develops novel algorithms to detect, classify, and predict events in large data streams for scientific and engineering systems.





- In early January 2007, ISS Early External Thermal Control System developed an ammonia gas bubble
- Bubble noted by ISS controllers only ~9 hours before it "burst" and dissipated back into liquid

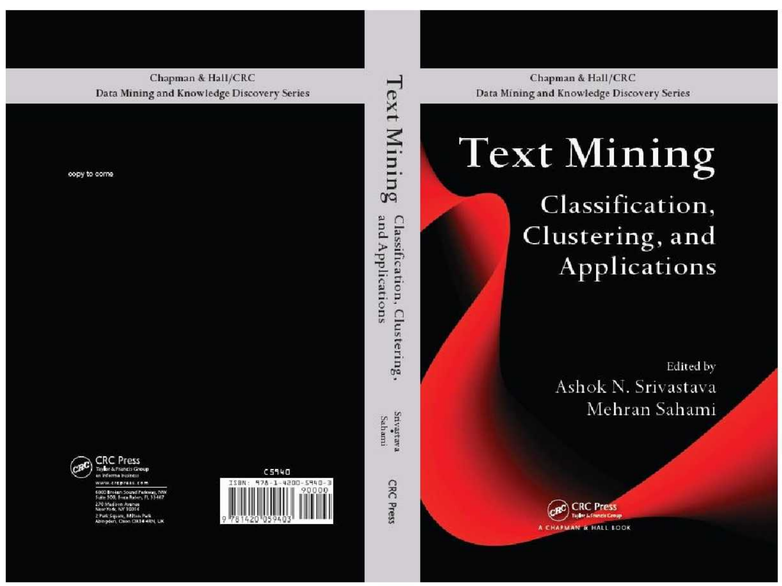
Key areas of research in data mining

Research Topic Areas

- Anomaly Detection
- Prediction Systems
- Text Mining
- Mining Distributed Data Systems and Sensor Networks
- High Performance Time Series Search

Application Areas

- Safety critical systems
- Large scale distributed systems
- Earth Sciences
- Space Sciences
- Systems Health Data from Aeronautical and Space Systems



NASA Data Systems

- Earth and Space Science
  - Earth Observing System generates ~21 TB of data per week.
  - Ames simulations generating 1-5 TB per day
- Aeronautical Systems
  - Distributed archive growing at 100K flights per month with 2M flights already.
- Exploration Systems
  - Space Shuttle and International Space station downlinks about 1.5GB per day.

Developing *Virtual* Sensors

- Virtual Sensors predict the value of one sensor measurement by exploiting the nonlinear correlations between its values and other sensor readings.
- Useful for emulating sensors back in time or estimating the value of one sensor based on other sensor measurements

Z: Sensors measurements
λ: Wavelength or Frequency
u: Position

$$Z(\mathbf{u}, \lambda, t) = [Z_{\mathbf{u}}(\lambda, t)]$$
$$= [Z_{u_1}(\lambda, t), Z_{u_2}(\lambda, t), \ldots, Z_{u_n}(\lambda, t)]^T$$

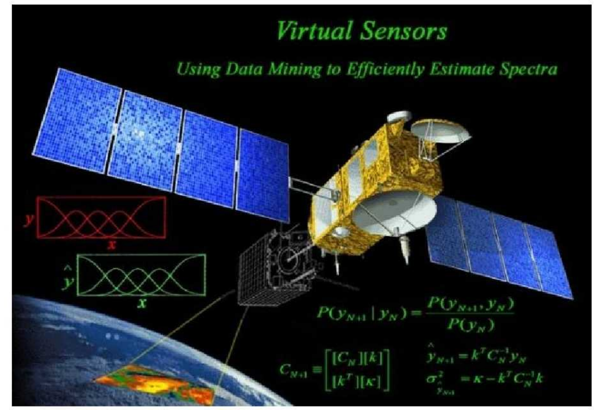
$$\mu(Z(\mathbf{B})) = \int \Gamma(Z(\mathbf{B}))Z(\mathbf{B})d\mathbf{B}$$

Predicted Sensor Measurement

$$\sigma^2(Z(\mathbf{B})) = \int [\Gamma(Z(\mathbf{B})) - \mu(Z(\mathbf{B}))]^2 Z(\mathbf{B})d\mathbf{B}$$

Estimated Uncertainty

Earth and Space Sciences



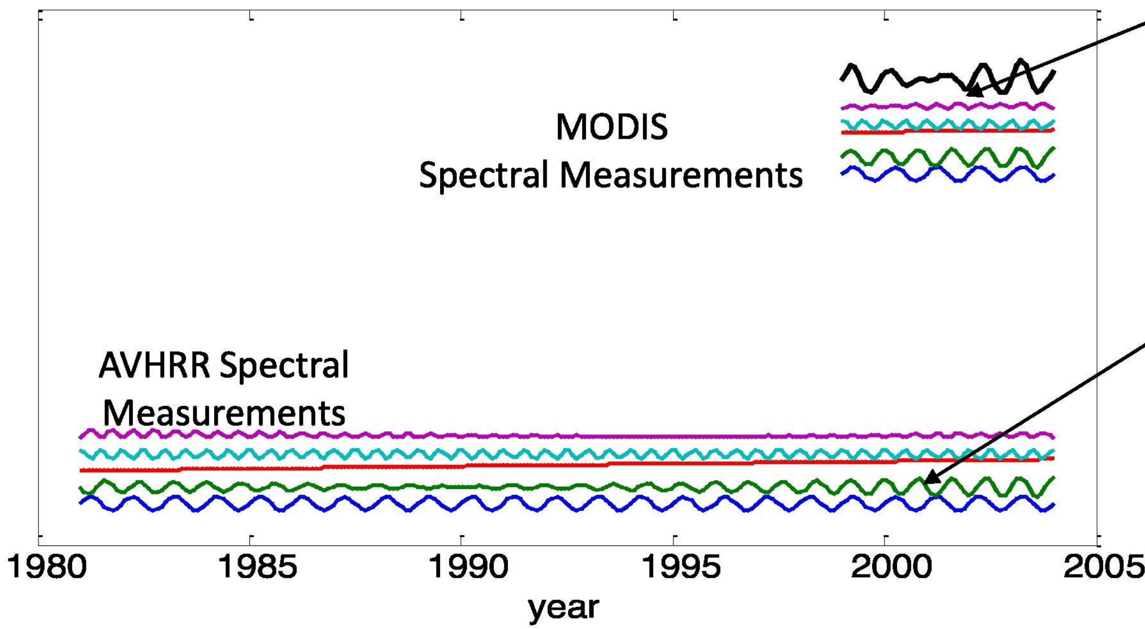
Aeronautics and Space Systems

Virtual Sensors in the Earth Sciences

Collaborators
Ashok N. Srivastava, NASA Ames
Nikunj C. Oza, NASA Ames
Julienne Stroeve, National Snow and Ice Data Center
Ramakrishna Nemani, NASA Ames
Petr Votava, NASA Ames

Has Cloud Cover Changed over Greenland in the past 30 years?

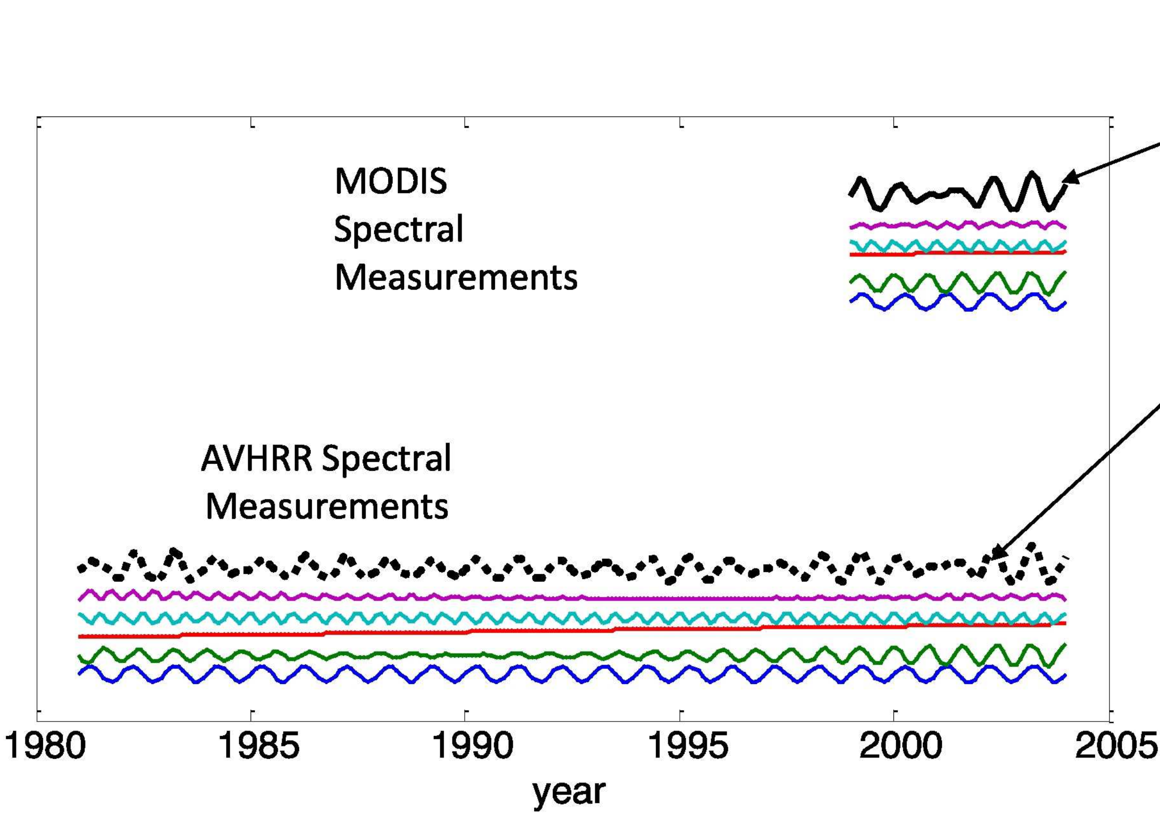
- New sensors on the MODIS system can detect clouds over snow and ice in the 1.6μm band (circa 1999).
- Difficult over snow and ice-covered surfaces because of low contrast in visible and thermal infrared wavelengths.
- Older sensors from the AVHRR system do not detect cloud cover over snow and ice because of poor contrast.

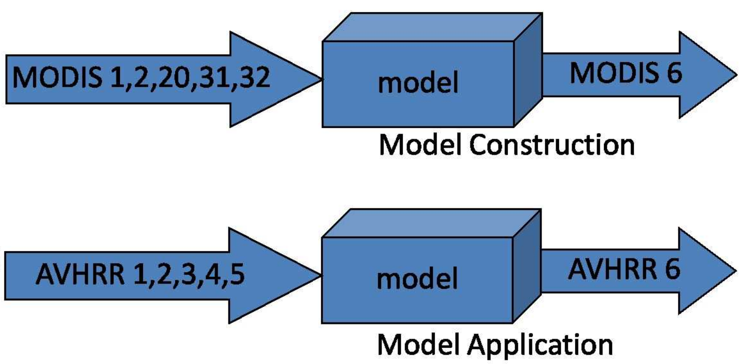


Joint work with Nikunj Oza, Julliene Stroeve, Rama Nemani, Brett Zane-Ulman

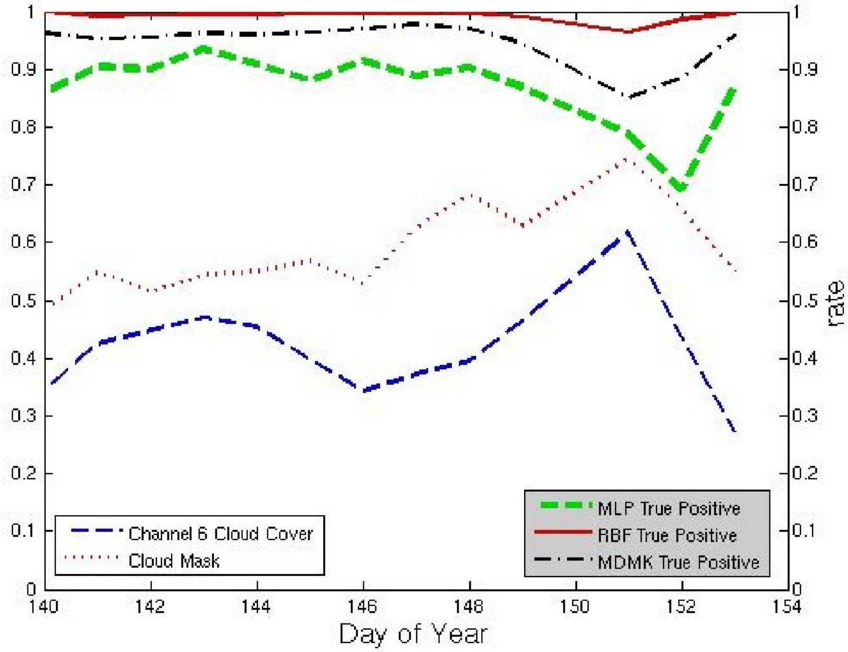
Cloud Detection back in Time

- MODIS 1.6μm has enough contrast for this task.
- However 1.6μm channel not available in AVHRR/2.
- Predict 1.6μm channel using a Virtual Sensor

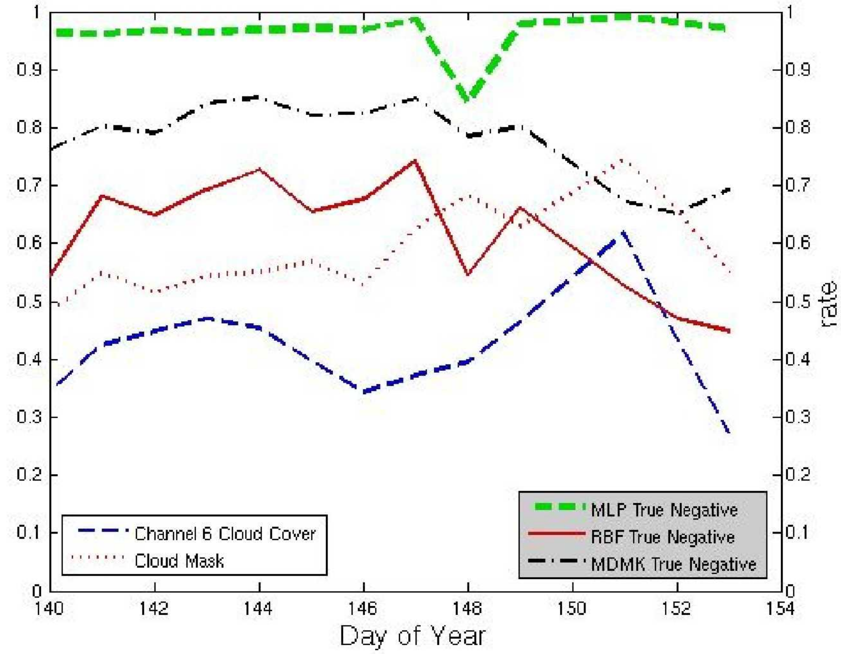




Accuracy Results for Three Models



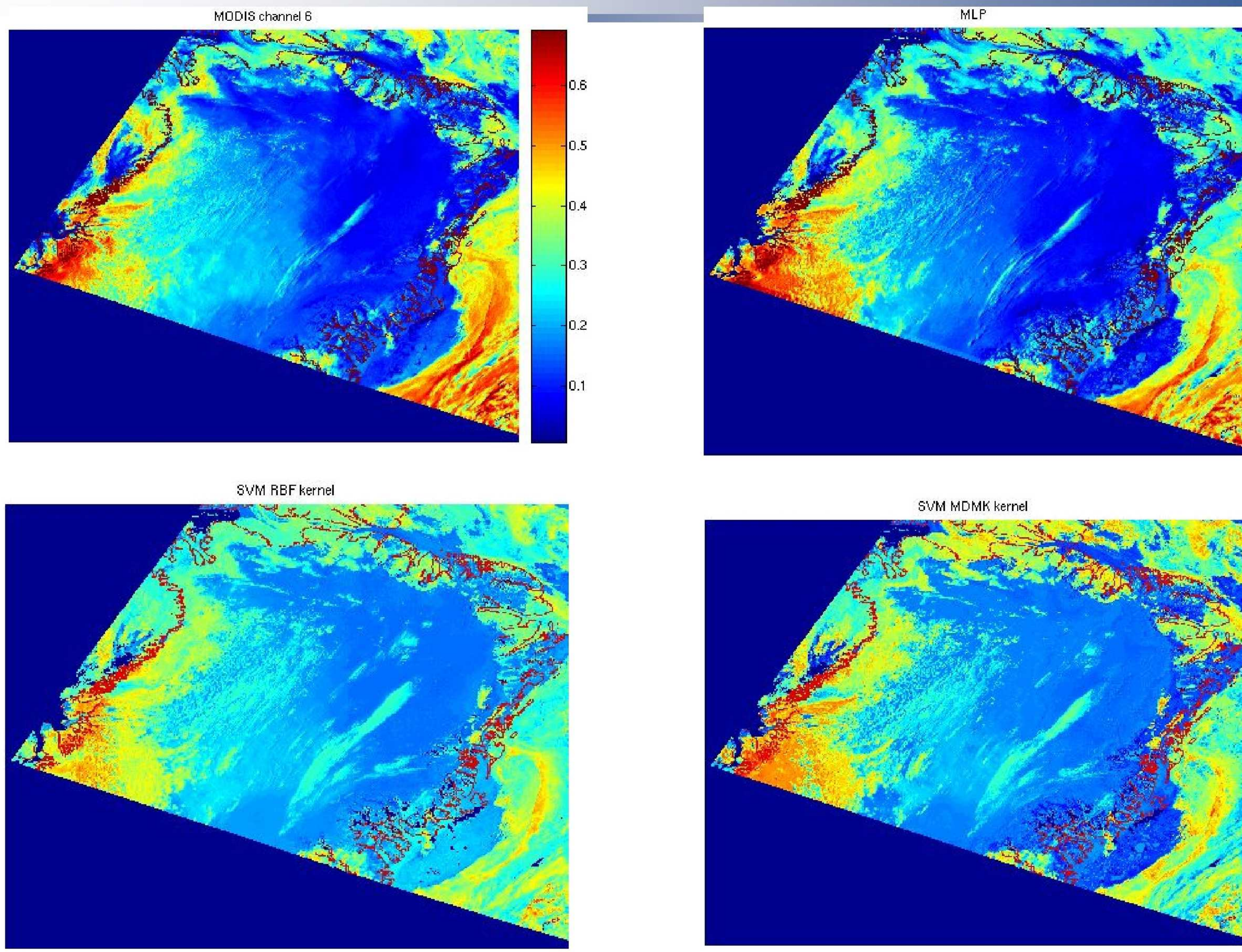
True Positive



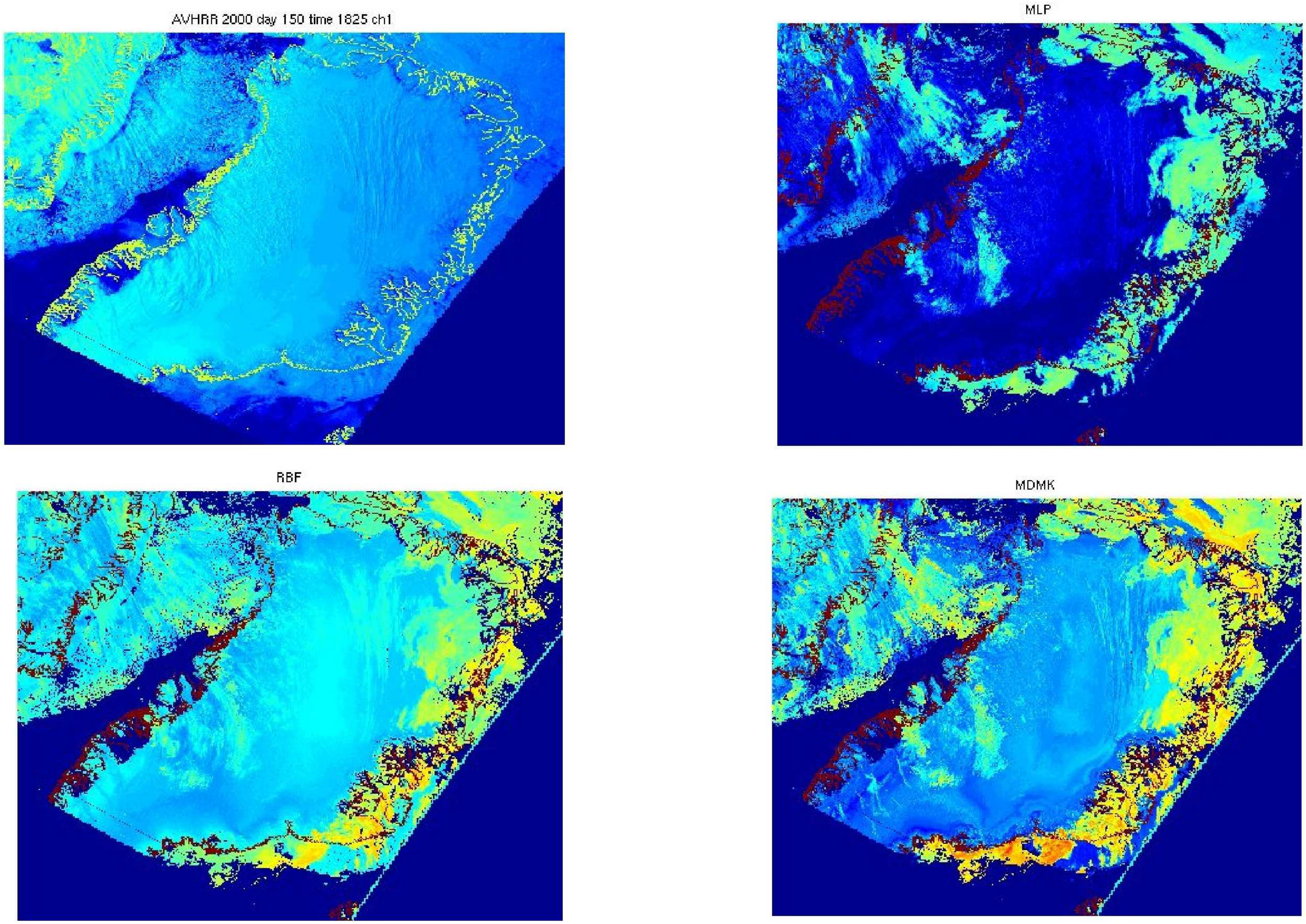
True Negative

- True Positive = number of times channel 6 indicated a cloud and the model predicted cloud
- True Negative = number of times channel 6 indicate no cloud and the model predicted no cloud

Verification of Models on MODIS Data



Application of Models to AVHRR Data



Summary

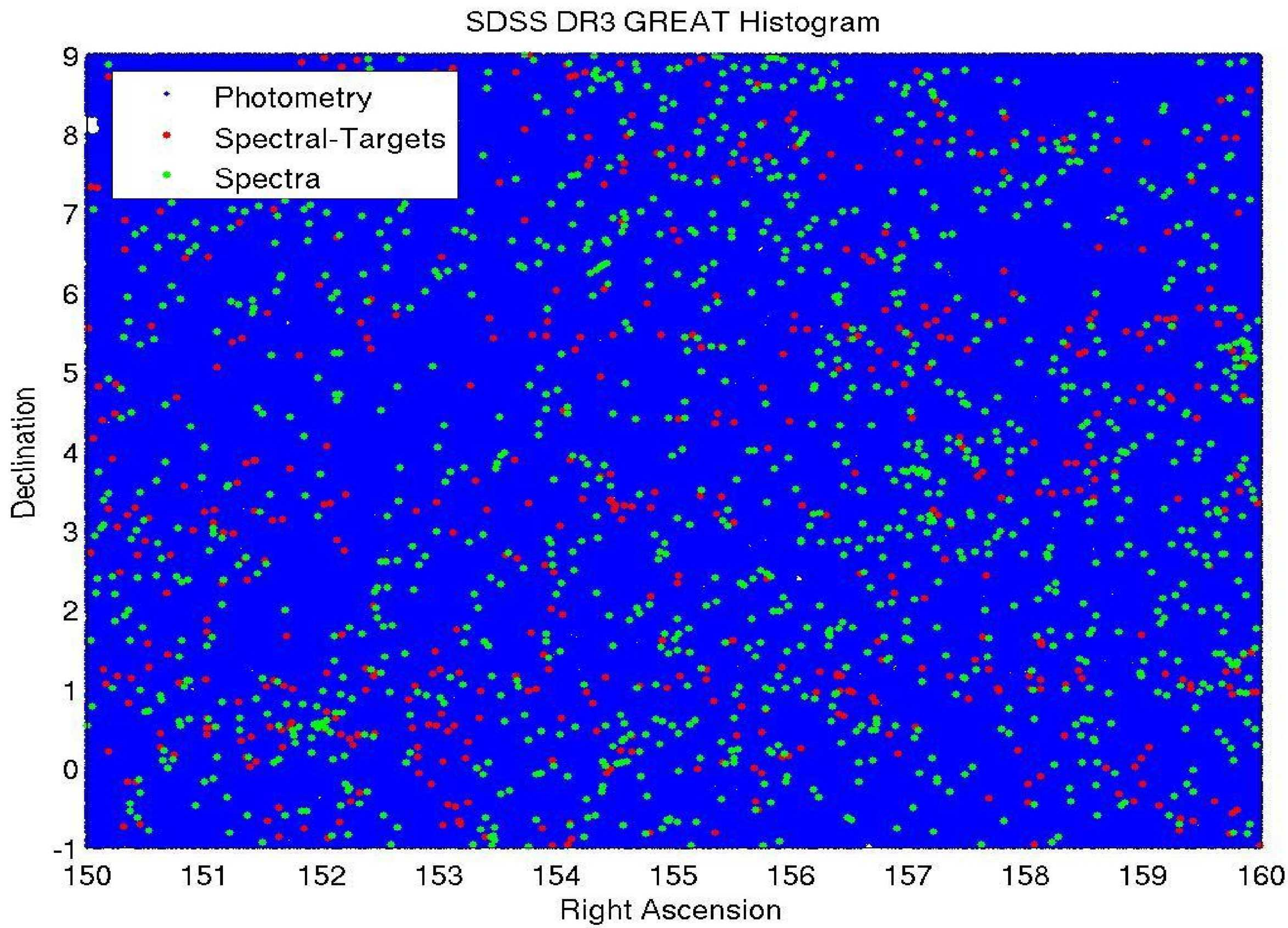
- Application to entire historical record is a significant task because of data quality issues and transitions from one sensor system to another.
- Method applied to emulation of physics models to calculate corrections for surface albedo measurements resulted in an increase in speed by factor of 27 compared to existing methods.
- Potential to deploy Virtual Sensors for generation of a historical cloud mask record.
- Model verification and validation must be done by hand since we have no signal for comparison.

A. N. Srivastava, N. C. Oza, and J. Stroeve, "Virtual Sensors: Using Data Mining Techniques to Efficiently Estimate Remote Sensing Spectra," Special Issue on Advanced Data Analysis, IEEE Transactions on Geoscience and Remote Sensing, March 2005.

Virtual Sensors in Astrophysics

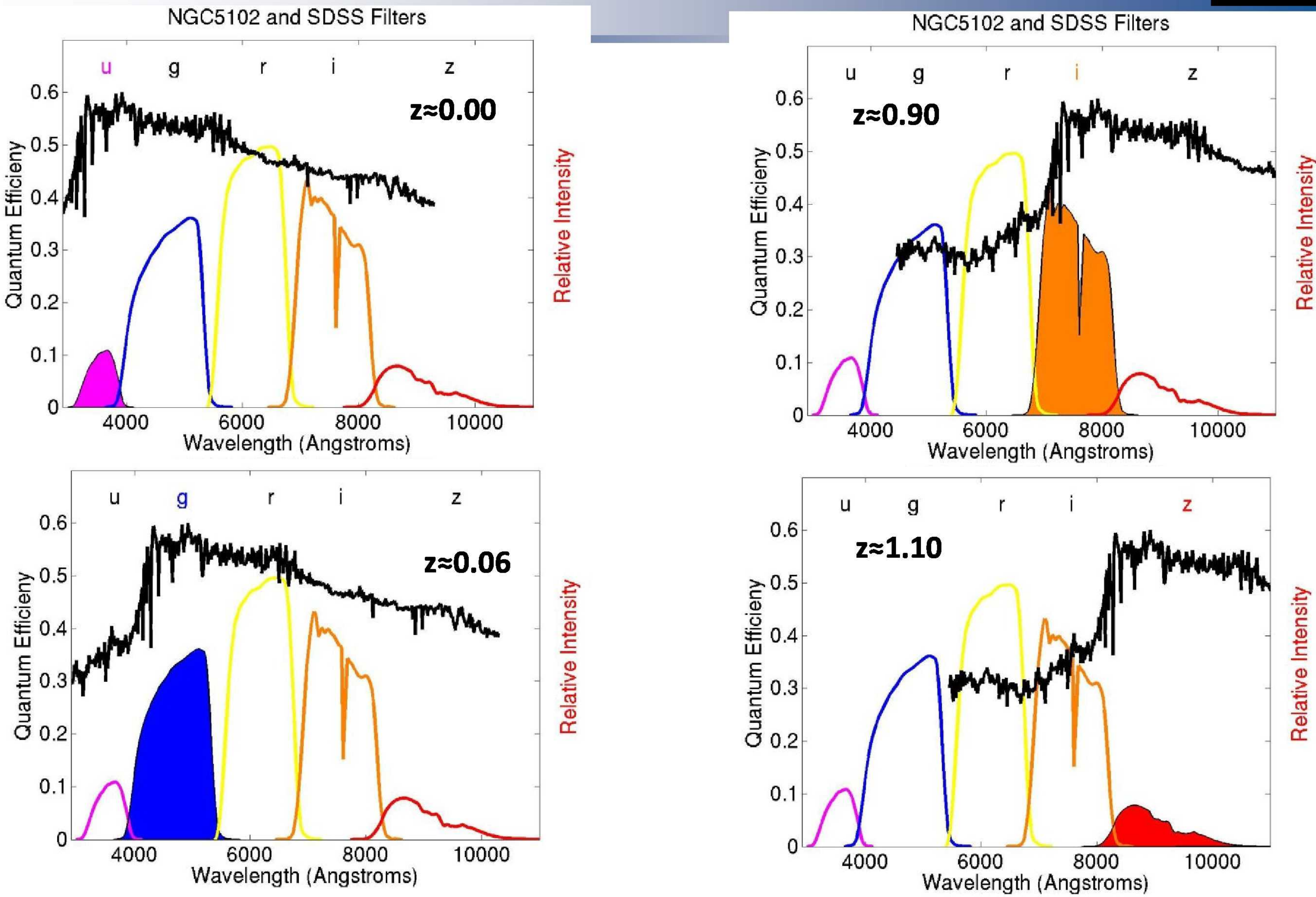
Collaborators
Michael J. Way, NASA Goddard Institute of Space Science
Leslie Foster, San Jose State University
Ashok N. Srivastava, NASA Ames
Paul Gazis, NASA Ames
Jeffery Scargle, NASA Ames

Estimating Photometric Redshifts in the Sloan Digital Sky Survey



Joint work with Michael J. Way, Leslie Foster, Paul Gazis, and Jeffrey Scargle

Photometric Redshifts are Broadband Measurements of Spectra



Gaussian Process Regression

- Can have high accuracy and also measure of uncertainty
- **some *low-rank matrix approximations* work well** but can have numerical problems.

Training Data:

- X – data matrix of observations – $n \times d$
- y – vector of target data – n $\times$ 1

Testing Data:

- X^* – matrix of new observations – $n^* \times d$

Goal:

- predict y^* corresponding to X^*

- Form covariance matrix K ($n \times n$), cross covariance matrix K^* ($n^* \times n$) and select parameter λ
- predict y^* using

$$\hat{y}^* = K^*(\lambda^2 I + K)^{-1} y$$

- the $n \times n$ matrix $(\lambda^2 I + K)$ is large for large data sets

- Memory: Storing covariance matrix – $O(n^2)$
- Time: Solving linear system – $O(n^3)$
- Numerical stability: accurate calculations.

Standard Least Squares Problems

- Given:

  $n \times m$ matrix A, $\quad n \geq m$

  $n \times 1$ vector y

- Solve $\min ||y - Ax||$
- Normal Equations: $x = (A^T A)^{-1} A^T y$

  potential numerical instabilities

- QR: $A = QR$, $x = R^{-1} Q^T y$

  stable calculation

Computational Challenges

- Subset of Regressors [Wahba, 1990]

$$\widehat{y}^* \cong K_1^*(\lambda^2 K_{11} + K_1^T K_1)^{-1} K_1^T y$$

- Memory: Storing covariance matrix – $O(nm)$
- Time: Solving linear system – $O(nm^2)$
- Numerical stability: ???.

Cures for Numerical Instability: *The V-Method*

Approach

1. Select columns to make K_1 well conditioned
2. Use stable technique for least squares problem such as
   - QR factorization
   - V method
3. Requirement: maintain $O(nm)$ memory use and $O(nm^2)$ efficiency.

Column Selection

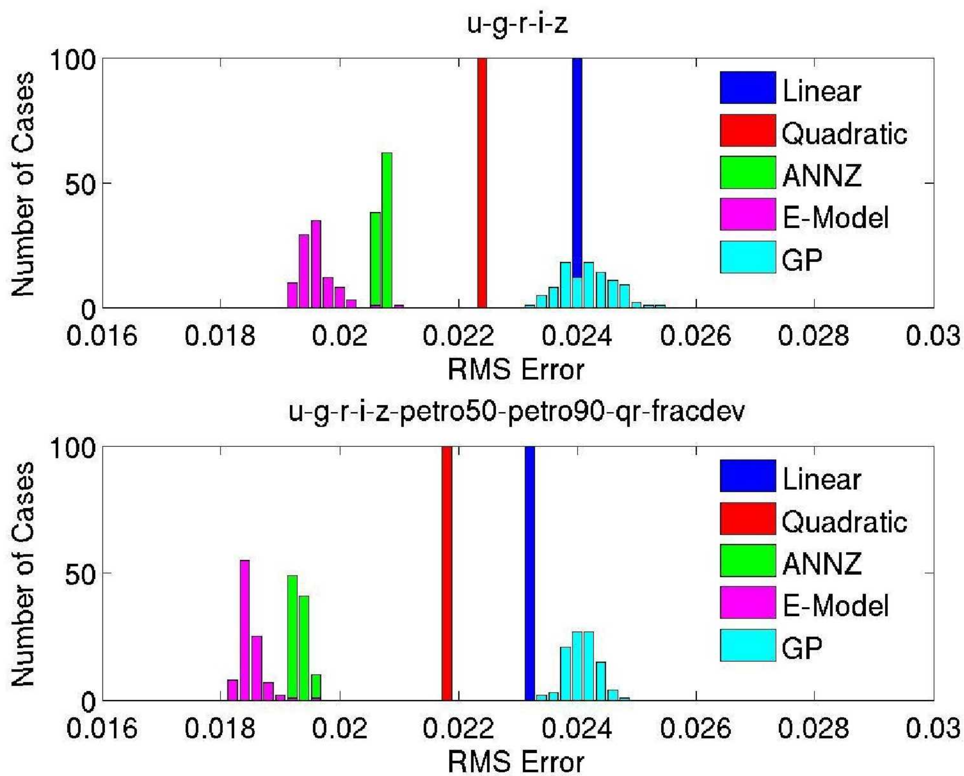
1. Use Cholesky factorization with pivoting to partially factor K
2. selects appropriate columns for K_1
3. K_1 will be well conditioned if $cond(K_1)$ is O(condition of optimal low rank approximation).

The V-Method is the innovation of Leslie Foster and his students at San Jose State University

The V-Method

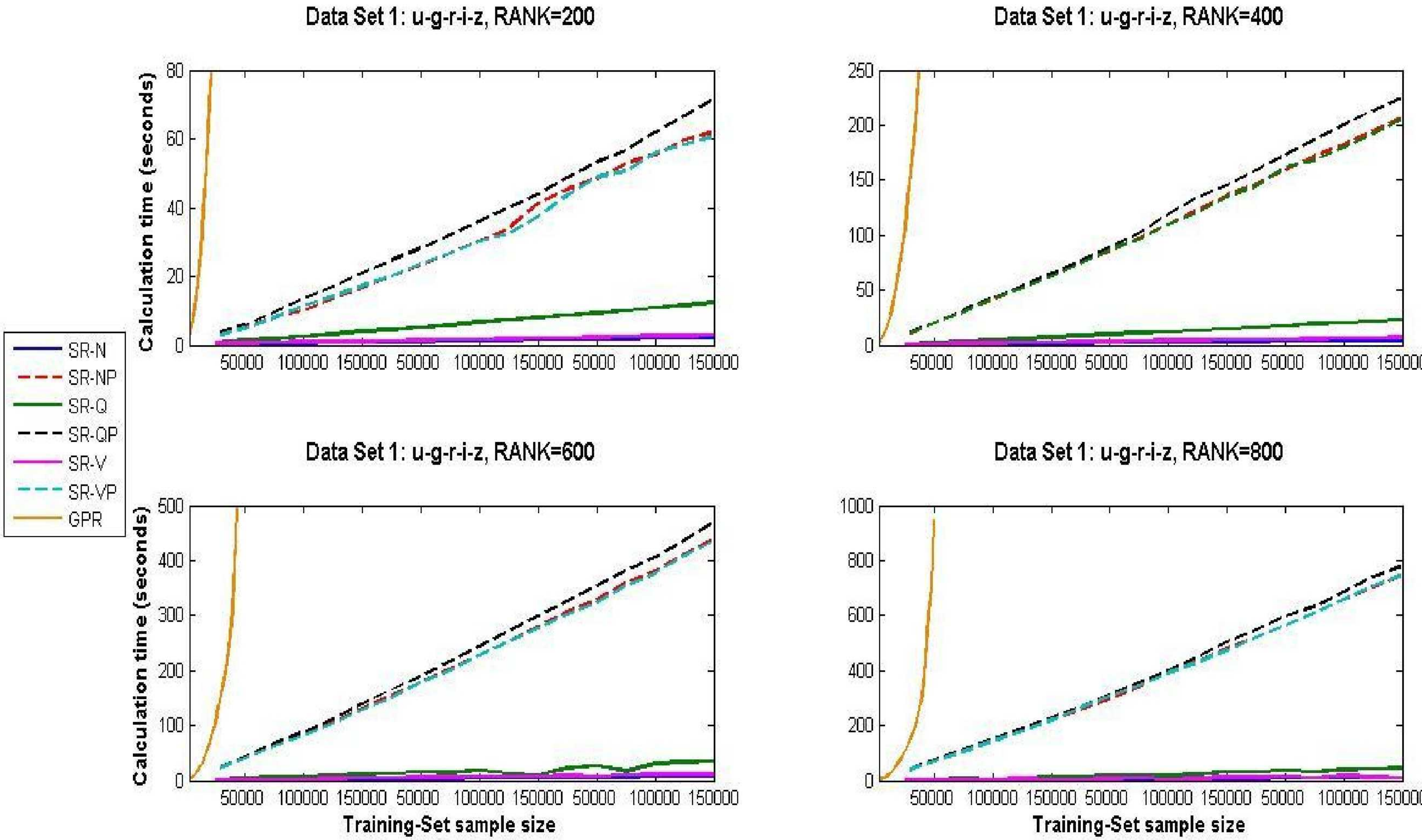
- Factor $K_1 = VV_{11}^T$ where V is $n \times m$ and V_{11} is $m \times m$ lower triangular
- $\widehat{y}^* = K_1^* V_{11}^{-T} (\lambda^2 I + V^T V)^{-1} V^T y$
- V is a rescaling of a well conditioned matrix
- method is numerically stable
- can be faster and need less memory
- related to [Peters and Wilkinson, 1970], [Wahba, 1990]

Prediction Accuracy

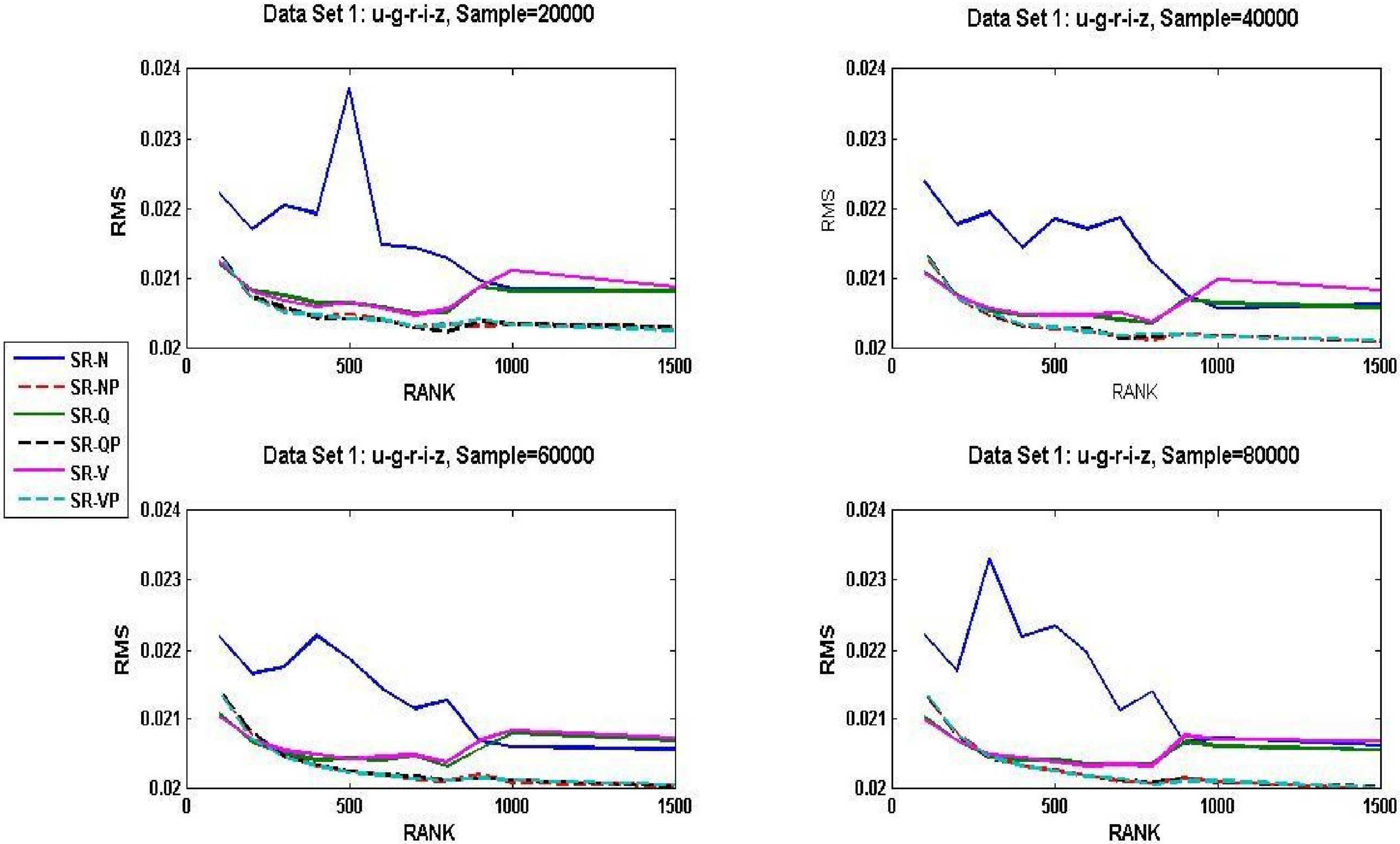


- Our ensemble models produce the best redshift estimates published to date.
- We are developing Gaussian Process Regression methods to scale to 10^6 galaxies and beyond.

Scalability Results



Best Published Results so far*…



* To the best of our knowledge

Results for Redshift Predictions

- The V-Formulation provides an extremely scalable and numerically stable method to compute Gaussian Process Regression for arbitrary kernels.
- With *low-rank matrix inversion approximations* GPs performed better than all other methods.
- Allows us to compute GPs for O(200K) points in a few seconds on a standard desktop PC.

L. Foster, A, A. Waagen, N. Aijaz, M. Hurley, A. Luis, J. Rinsky, C. Satyavolu, M. J. Way, P. Gazis, and A. N. Srivastava, "Stable and Efficient Gaussian Process Calculations," Journal of Machine Learning Research, 10(Apr):857--882, 2009.

Data Mining Supporting the Flight Readiness Review for STS-119

Collaborators

Ashok N. Srivastava, NASA Ames

Dave Iverson, NASA Ames

Bryan Matthews, SGT

Bill Lane, NASA Johnson Space Center

Bob Beil, NASA Kennedy Space Center

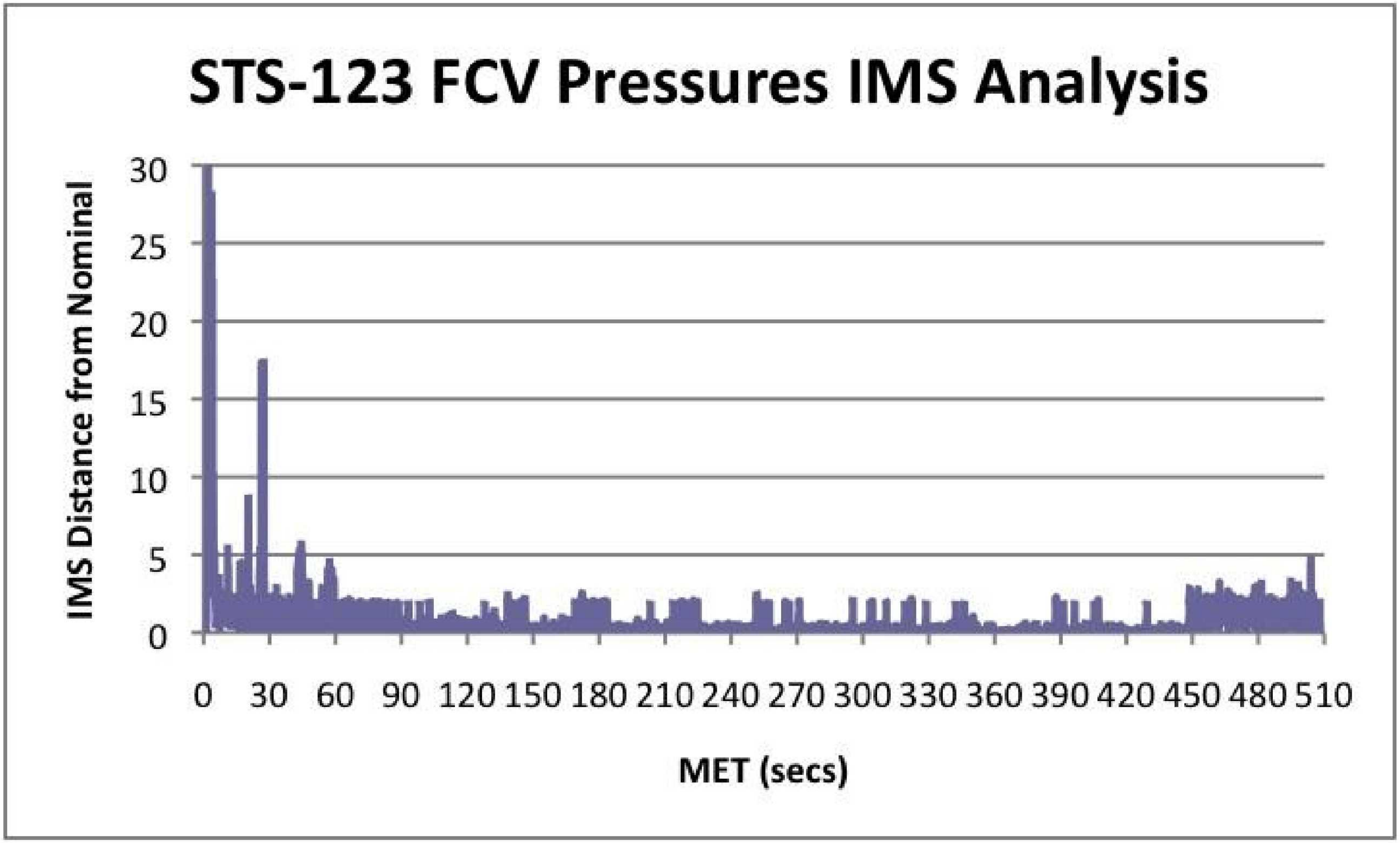
Overview

- Ashok received a request to support the Flight Readiness Review for STS-119 which was scheduled for 2/20/09 as the Data Mining Subject Matter Expert.
- Data mining algorithms developed at NASA were applied to these data to determine whether any anomalies can be detected in STS-126 and its predecessor flight STS-123 for Space Shuttle Endeavor.

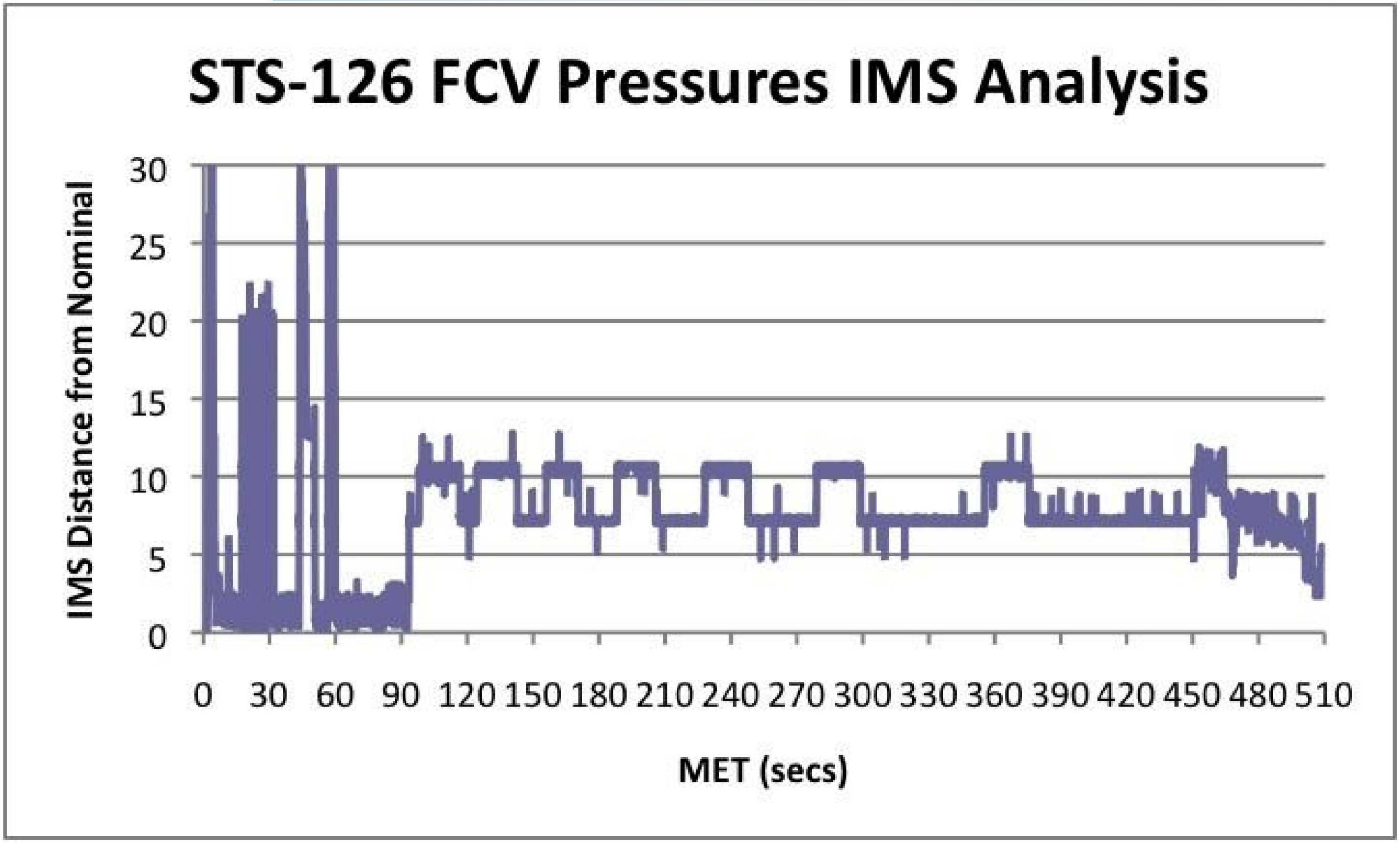
Algorithms and Data

- IMS (Inductive Monitoring System): a data point is anomalous if it is far away from clusters of nominal points.
- Orca: a data point is anomalous if it is far away from its nearest neighbors.
- Virtual Sensor: a data point is anomalous if the actual value is far away from the predicted value.
- Data: 13 pressure, temperature, and control variables related to the Flow Control Valve subsystem.

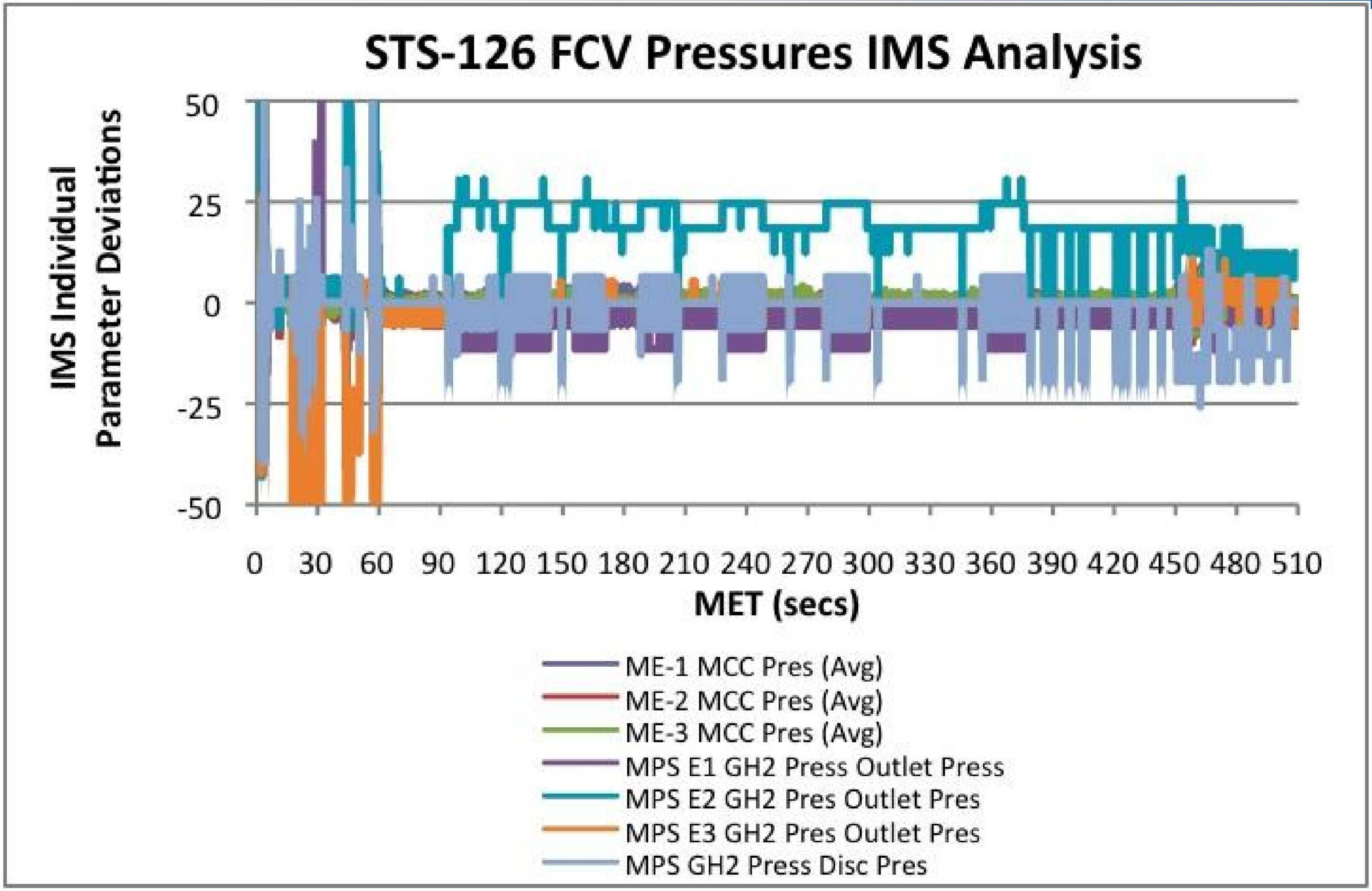
IMS Anomaly Score



IMS Anomaly Score



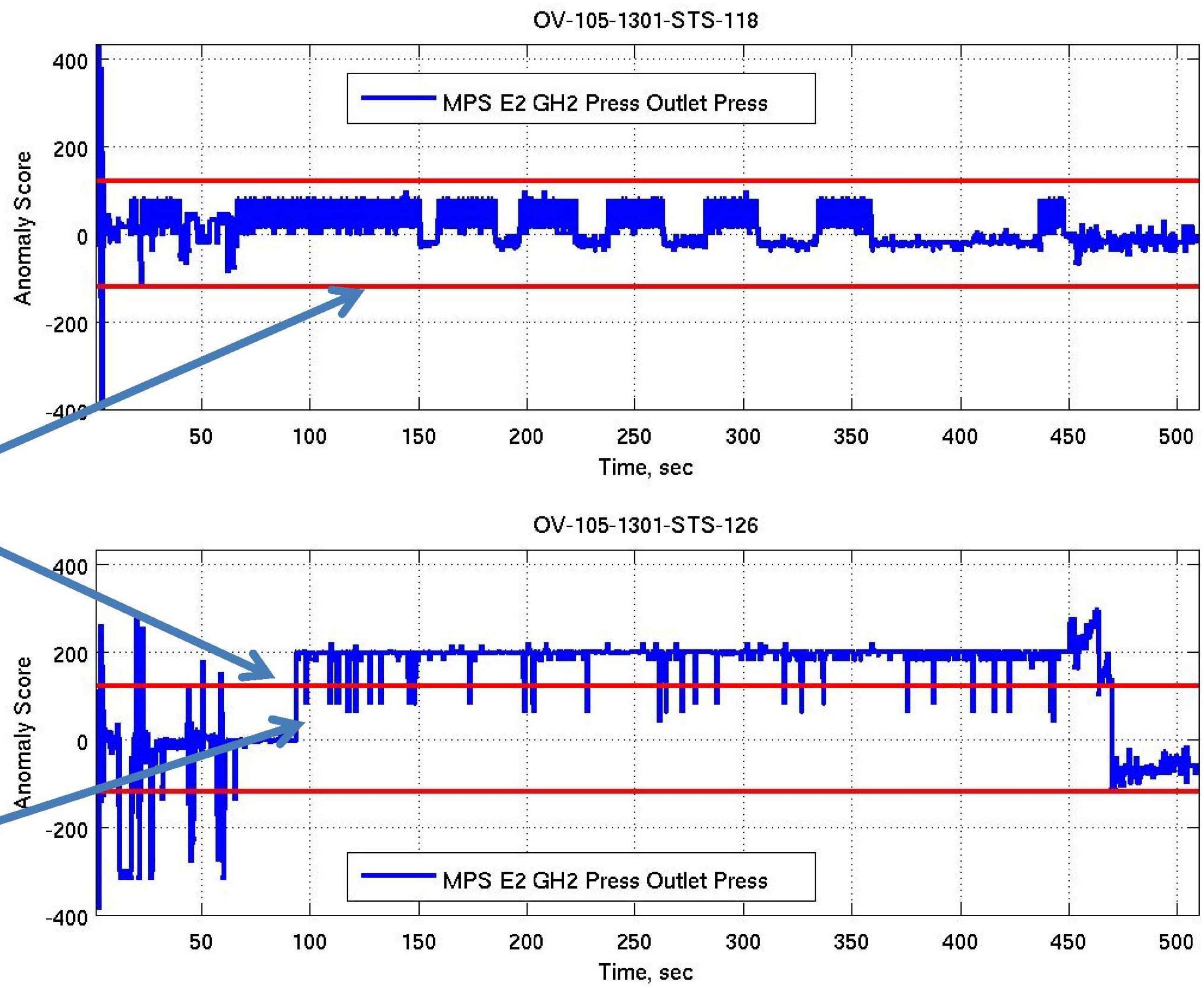
IMS Anomaly Score



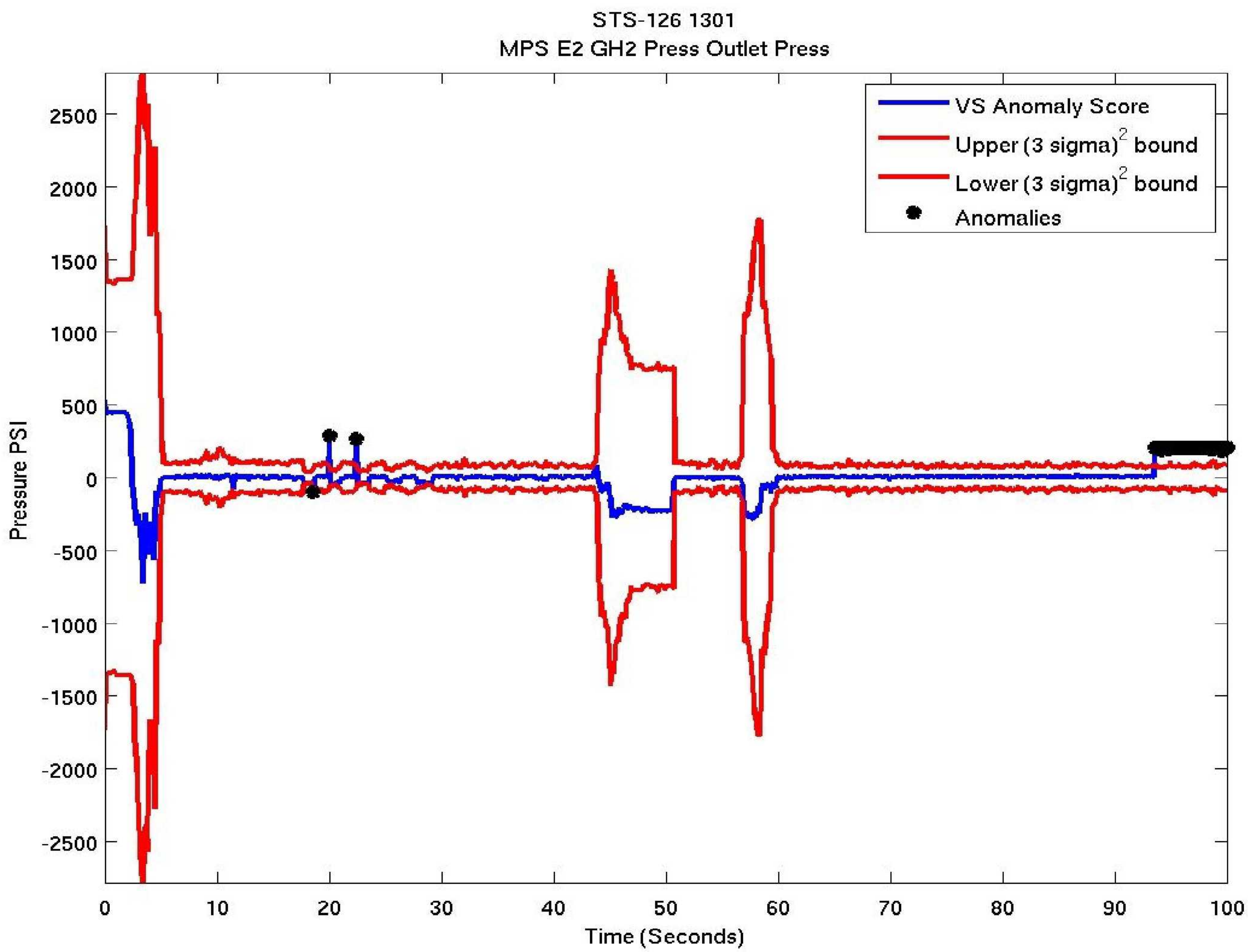
Virtual Sensor: STS-118 and STS-126

• Redlines correspond to 3-sigma nominal error rate on STS-118.

•STS-126 shows anomalous behavior after 93.6 seconds.



Virtual Sensors with Adaptive Thresholds

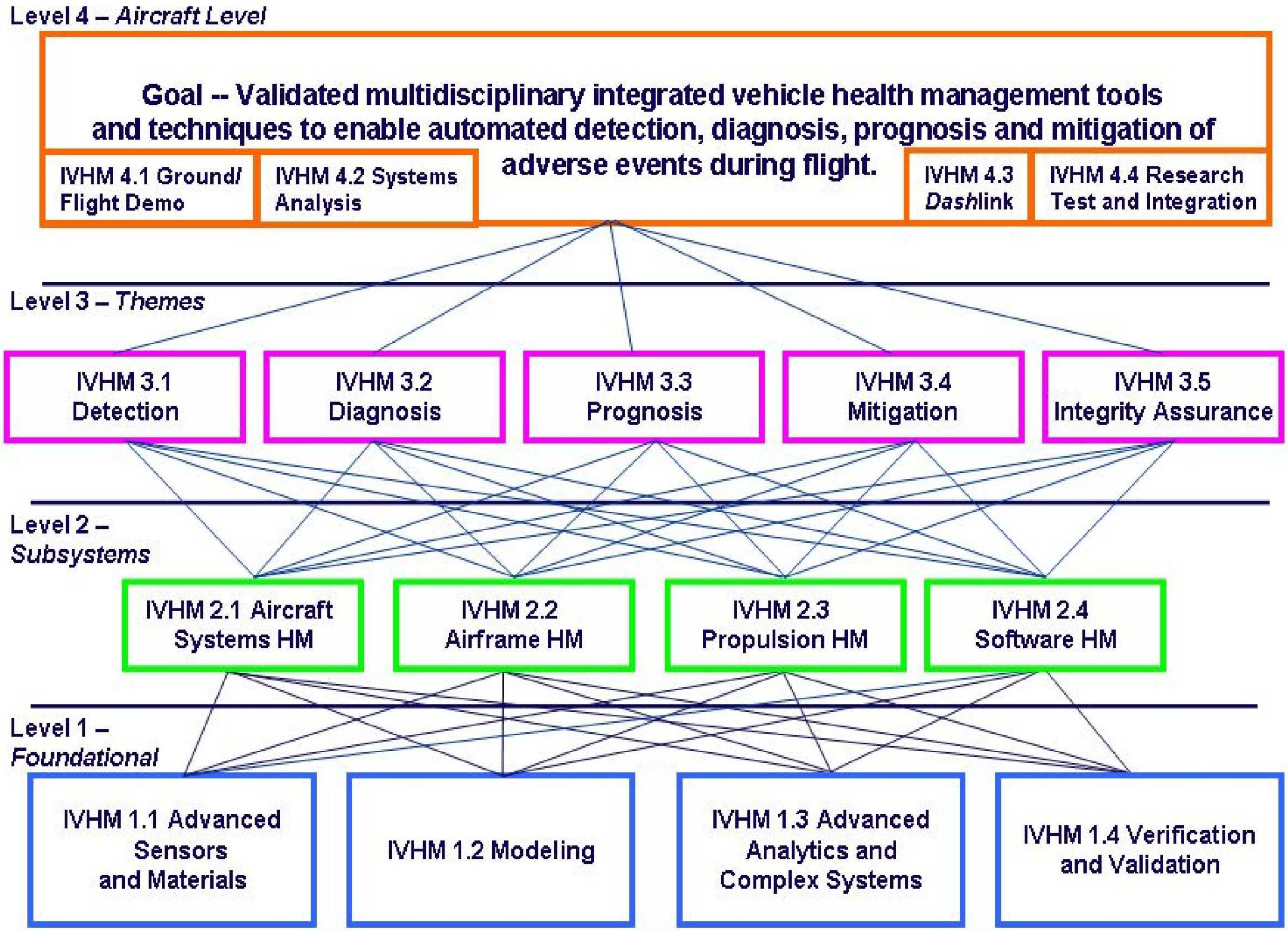


A. N. Srivastava, B. Matthews, D. Iverson, B. Beil, and B. Lane, "Multidimensional Anomaly Detection on the Space Shuttle Main Propulsion System: A Case Study," submitted to IEEE Transactions on Systems, Man, and Cybernetics, Part C, 2009.

The Role of Data Mining in Aviation Safety

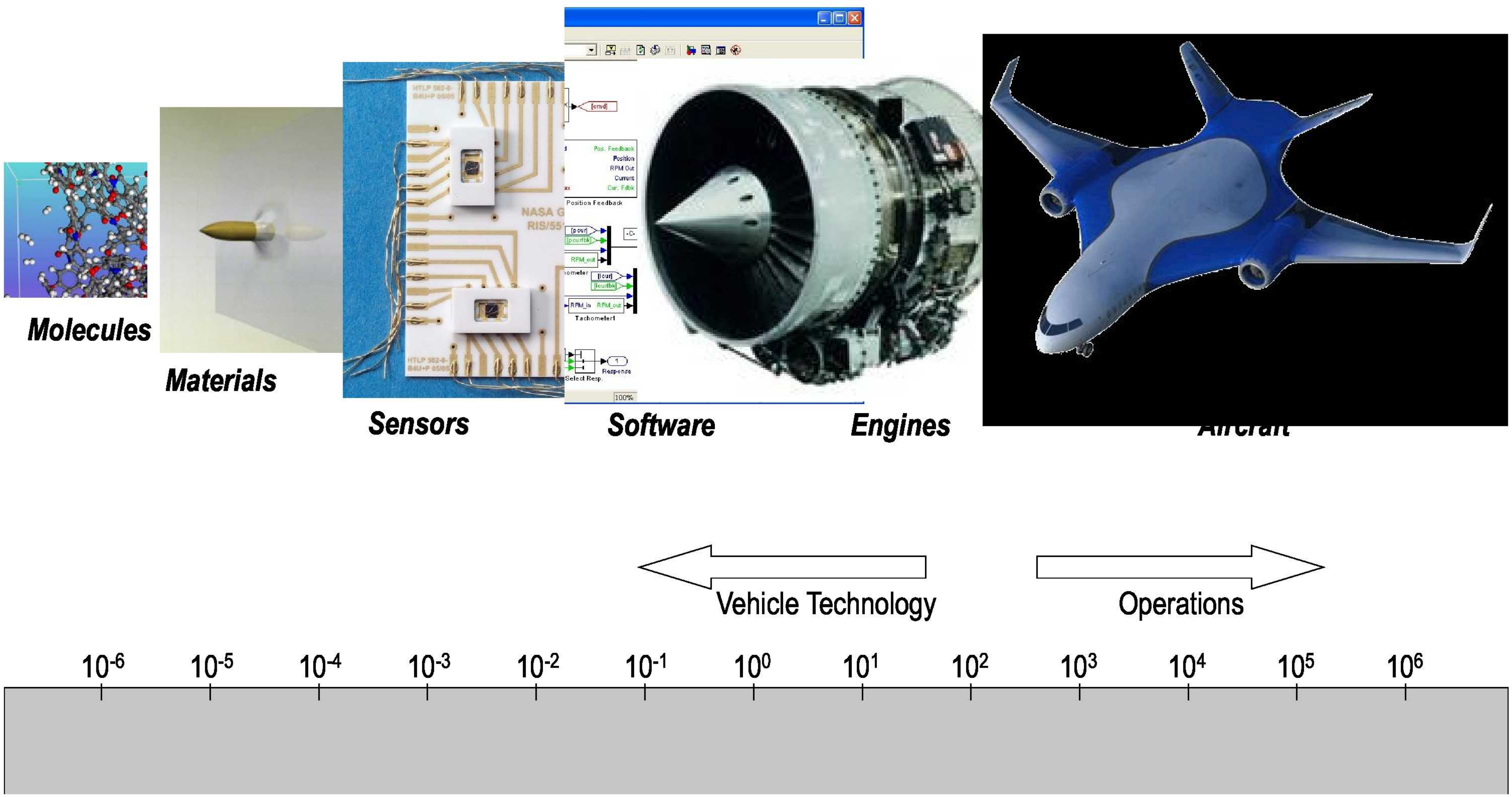
Ashok N. Srivastava, Principal Investigator
Claudia Meyer, Project Manager
Robert Mah, Project Scientist

Integrated Vehicle Health Management: An Aviation Safety Project

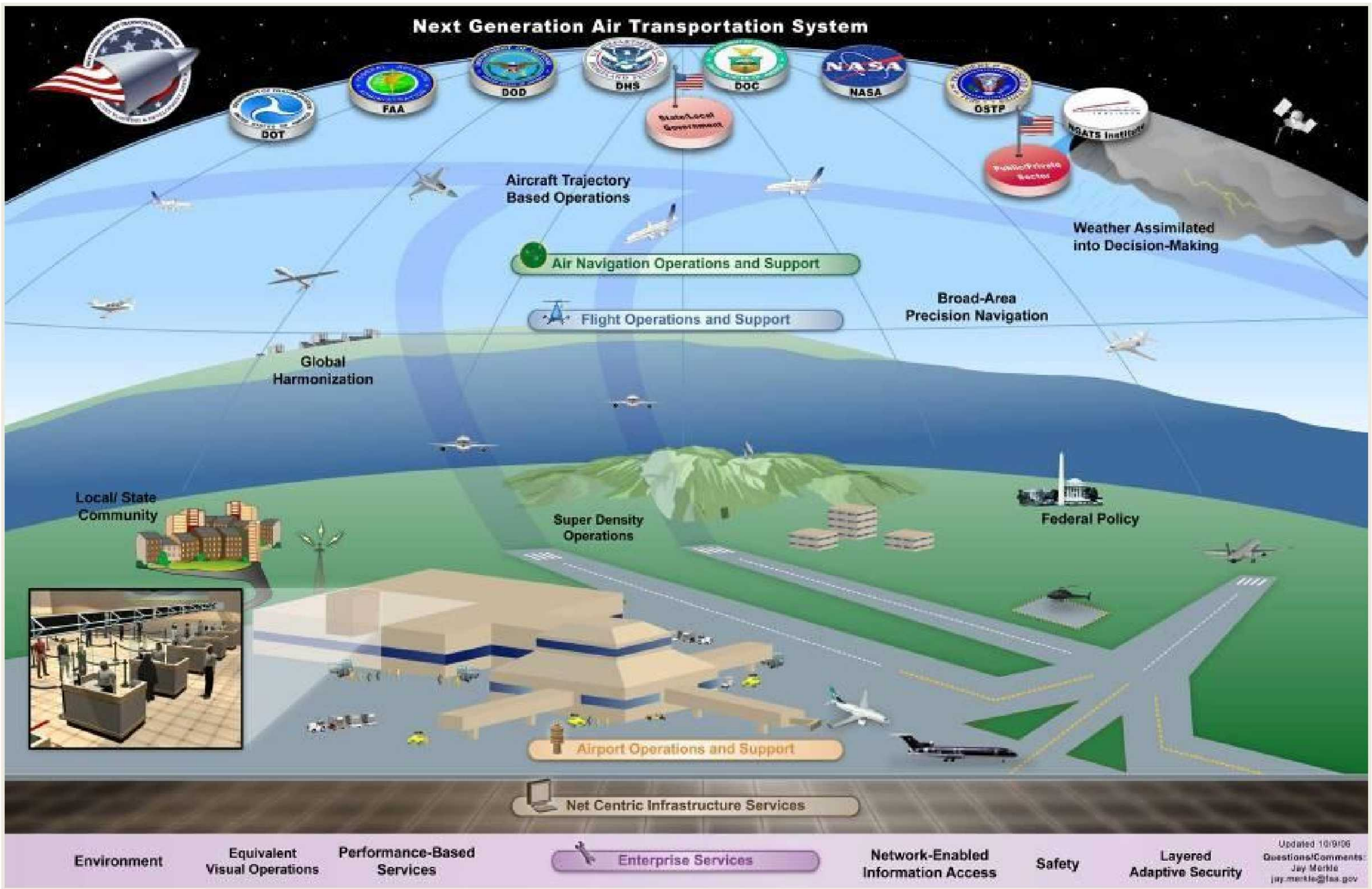


Some Partners of the IVHM Project

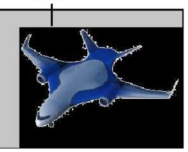
IVHM Covers a broad range of technology



Data Mining in Support of Global Operations



10^{-6} 10^{-5} 10^{-4} 10^{-3} 10^{-2} 10^{-1} 10^{0} 10^{1} 10^{2} 10^{3} 10^{4} 10^{5} 10^{6}



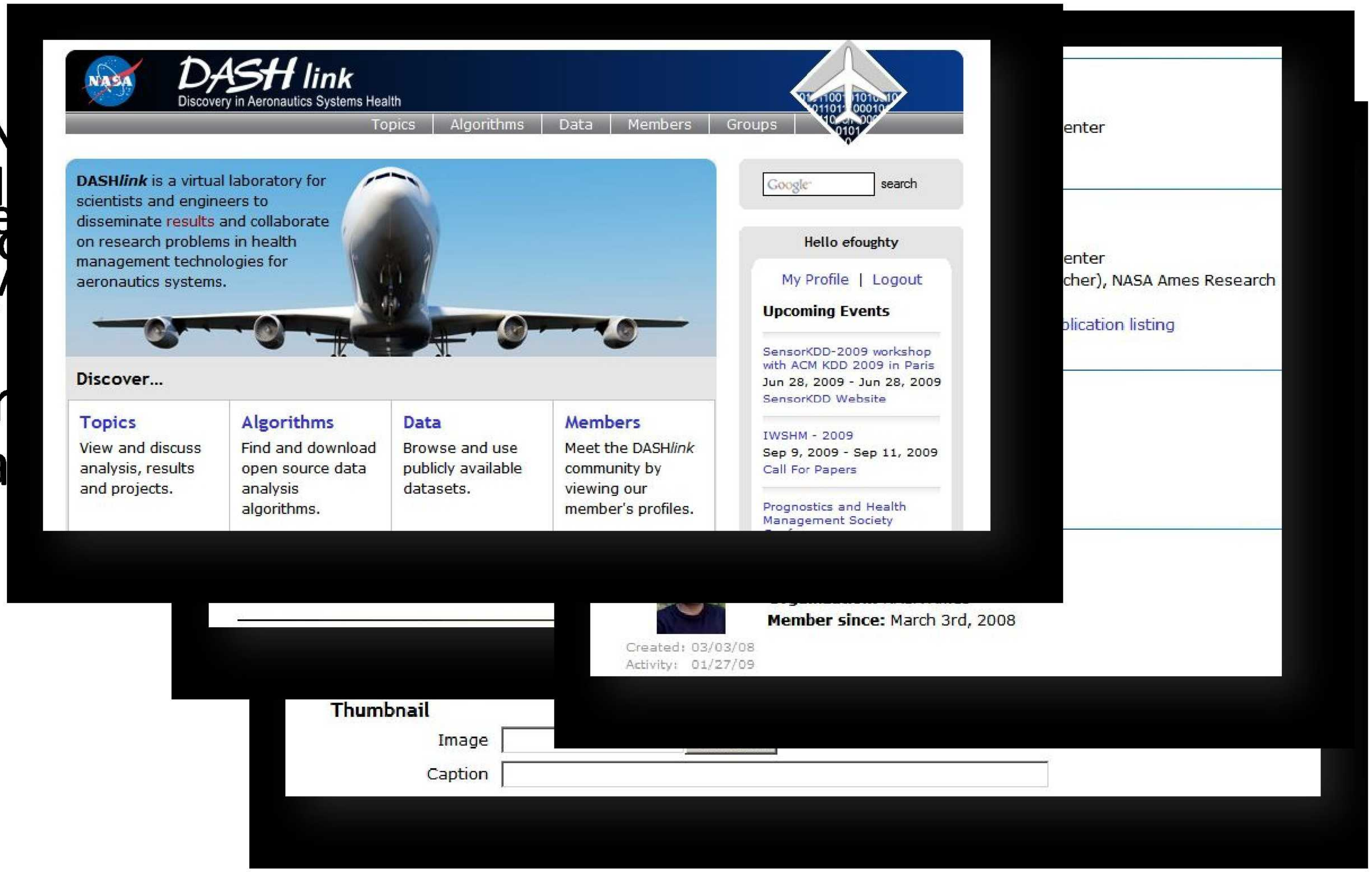
DASHlink.arc.nasa.gov

DASHlink harnesses the power of web 2.0 to further Systems Health and Data Mining research

Download N… papers, and… (including s… in this talk).

Find and inte… Data Mining/… researchers.

Easily shar… own resea…



Organization of IVHM

Principal Investigator: Ashok Srivastava
Project Scientist: Robert Mah
Project Manager: Claudia Meyer

•Project Operations
Manager: Jeff Rybak

•NRA
Manager: Lilly Spirkovska

ARC, DFRC, GRC, LaRC Center POCs

ARC APM: Steve Jacklin
DFRC POC: Mark Dickerson
GRC APM: Bob Kerczewski
LaRC APM: Sharon Graves

•Level 4

- **•Multidisciplinary Ground/ Flight Demos** — •Leads: PI, PS, PM
- **•Systems Analysis for Health Management** — Lead: Mary Reveley
- **•Research Test and Integration** — Lead: Robert Mah
- **•DASHlink** — Lead: Elizabeth Foughty

•Level 3 Associate Principal Investigators

- **•Detection** — API: John Lekki
- **•Diagnosis** — API: Rick Ross
- **•Prognosis** — API: Kai Goebel
- **•Mitigation** — API: Eric Cooper
- **•Integrity Assurance** — API: Eric Cooper

•Level 2

- **Aircraft Systems**
- **Airframe**
- **Propulsion Systems**

***Traditional** Aircraft Subsystems – well represented in Levels 1,3 and 4*

- **Software** — **Lead: Paul Miner**

***Newly Recognized** Aircraft Subsystem*

•Level 1 Lead Researchers

- **•Advanced Sensors and Materials** — Lead: Tim Bencic
- **•Modeling** — Lead: Kevin Wheeler
- **•Advanced Analytics and Complex Systems** — Lead: Nikunj Oza
- **•Verification and Validation** — Lead: Steve Jacklin

The Data Mining Team

Group Members

Kanishka Bhaduri, Ph.D.
Santanu Das, Ph.D.
Elizabeth Foughty
Dave Iverson
Rodney Martin, Ph.D.
Bryan Matthews
Nikunj Oza, Ph.D.
Mark Schwabacher, Ph.D.
John Stutz
David Wolpert, Ph.D.

Funding Sources

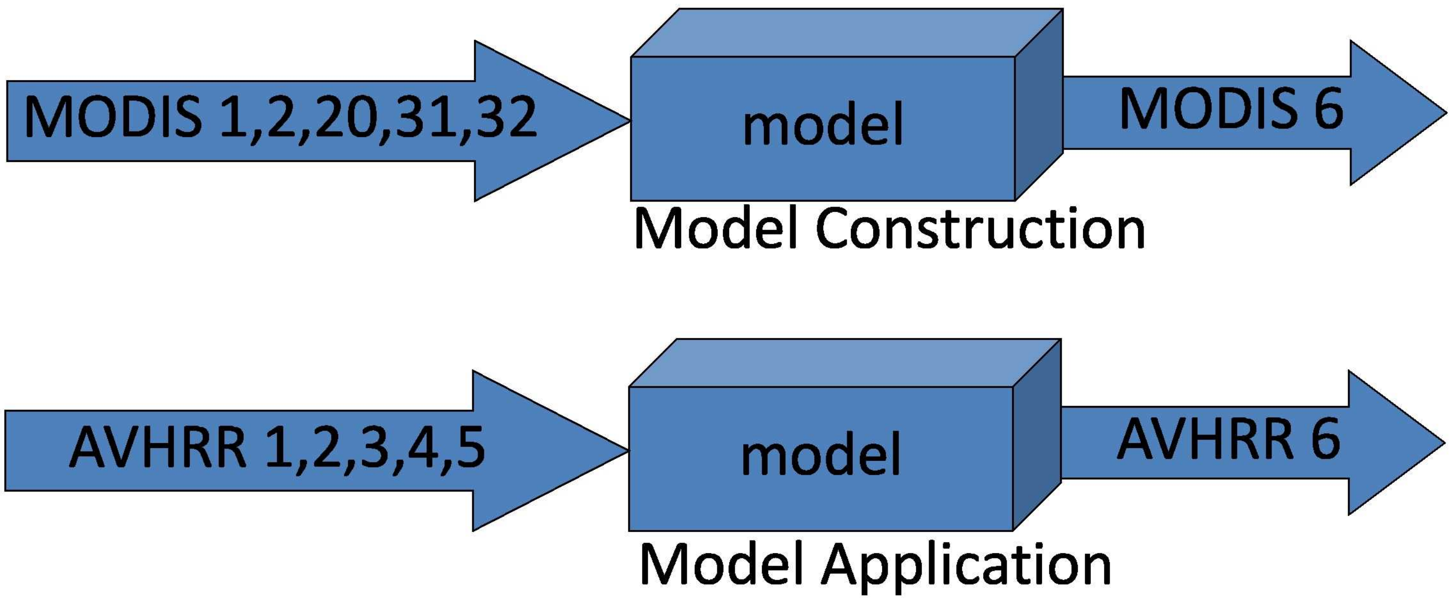
- NASA Aeronautics Research Mission Directorate- IVHM Project
- NASA Engineering and Safety Center
- Exploration Systems Mission Directorate Exploration Technology Development Program, ISHM Project
- Science Mission Directorate

Team Members are NASA Employees, Contractors, and Students.

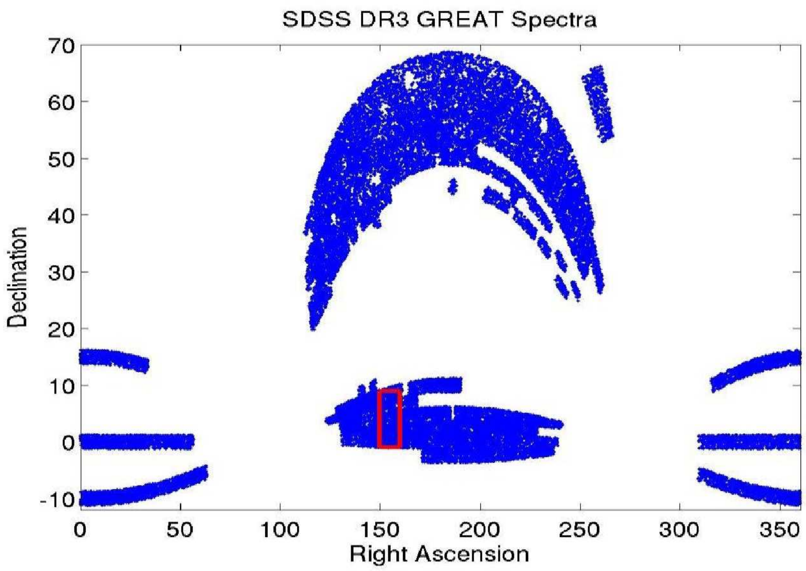
APPENDIX

Virtual Sensors Approach

- Given MODIS channels 1, 2, 20, 31, 32 correspond to five AVHRR/2 channels
- Develop a model for MODIS channel 6 (1.6mm) as a function of these channels
- Use function to construct estimate of 1.6mm channel for AVHRR/2



Characterizing the Large Scale Structure of the Universe

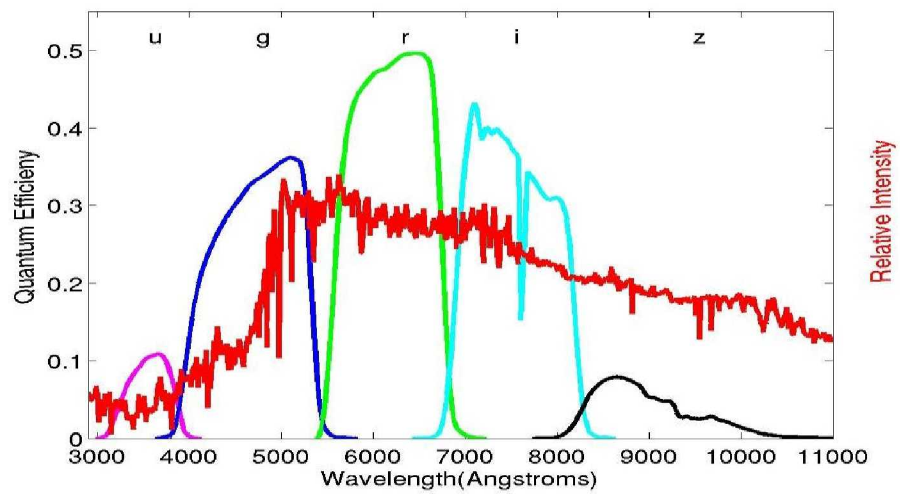


There are between 125 and 500 billion galaxies in the universe.

Obtaining a good estimate of their 3-D position in the sky would help determine the filamentary structure of the universe to constrain cosmological models.

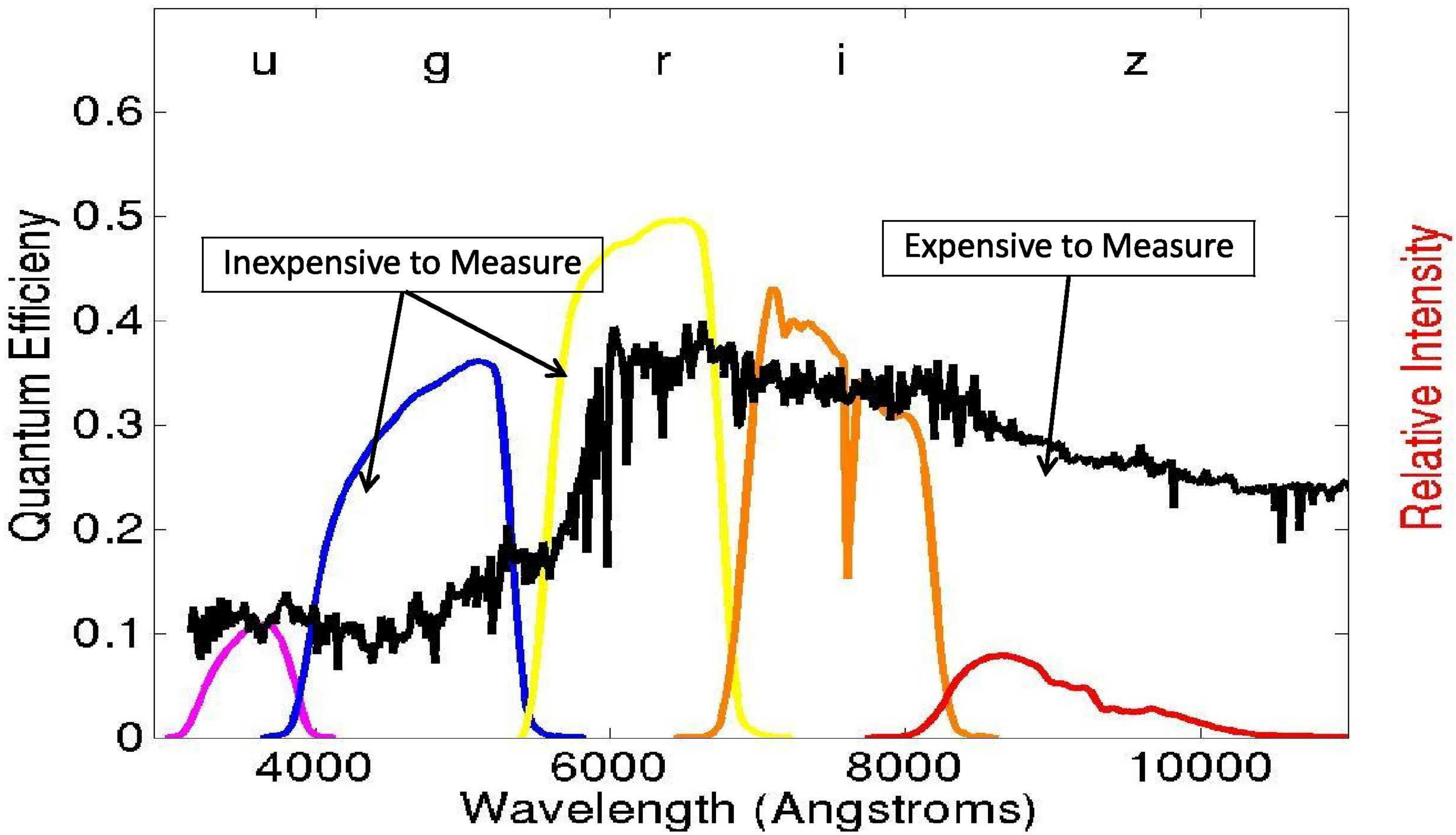
We are building machine learning methods to estimate the redshift of galaxies using broad-band photometry.

If these estimates are of high enough accuracy, it would enable a better understanding of how the universe evolved after the Big Bang.



What are Photometric Redshifts?

Photometric Redshifts: A **rough** estimate of the redshift of a galaxy without having to measure a spectrum.



The Empirical Approach to Redshift Estimation

Training sample consists of galaxies with

- known spectroscopic redshift
- a comparable range of **magnitudes (**u g r i z**)** to our photometric survey objects

Galaxy Photometric Redshift Prediction History

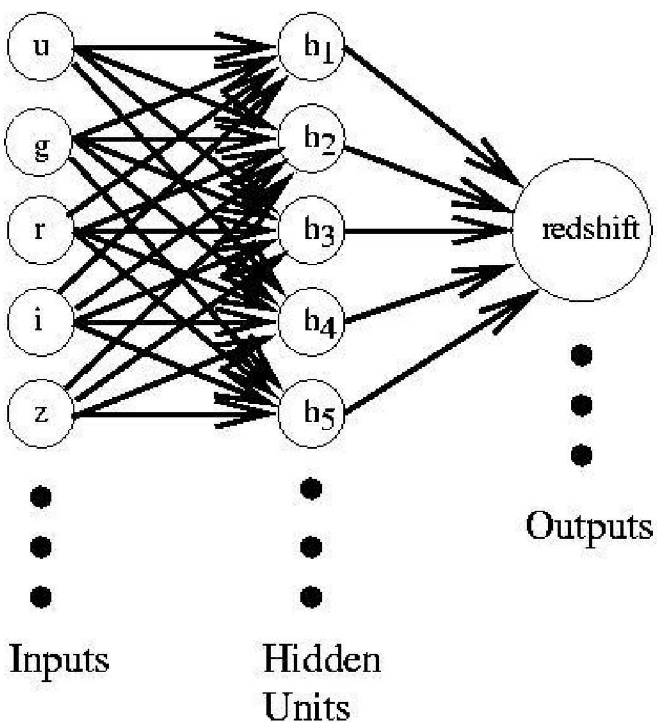
- Linear Regression was first tried in the 1960s
- Quadratic & Cubic Regression (1970s)
- Polynomial Regression (1980s)
- Neural Networks (1990s)
- Kd Trees & Bayesian Classification Approaches (1990s)
- Support Vector Machines & GP Regression (2000s)

Kernels Incorporate Prior Knowledge

Gaussian Process Regression

A large # of hidden units in a Neural Network

Gaussian Process Regression (Neal 1996).



Johann Carl Friedrich Gauss (1777–1855), painted by [Christian Albrecht Jensen] (wikipedia)

Large Scale Gaussian Processes

With our SDSS (DR3) Main Galaxy spectroscopic sample (180,000 galaxies) the matrix size is 180,000 x 180,000

- Need a supercomputer with a LOT of ram and cpu time?
- One can take a random sample of ~1000 galaxies & invert that while bootstrapping n times from full sample
- **However, some *low-rank matrix approximations* work well** such as Cholesky Decomposition, Subset of Regressors but can have numerical problems.
- Solution: V-method (Cholesky decomposition with pivoting)

The V-Method is the innovation of Leslie Foster and his students at San Jose State University

Numerical Instability in Subset of Regressors Method

- In SR formula consider special case $\lambda = 0$
- $\widehat{y}^* = K_1^*(K_1^T K_1)^{-1} K_1^T y$
- Exactly normal equations solution to the least squares prediction problem: min $||y - K_1 x||$ and $\widehat{y}^* = K_1^* x$
- Note: can be easily extended for $\lambda \neq 0$
- Potential numerical instability

Low Rank Approximations

$$K = \begin{array}{cc} & \begin{array}{cc} m & n-m \end{array} \\ \begin{array}{c} m \\ n-m \end{array} & \begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{pmatrix} \end{array} = \begin{array}{cc} & \begin{array}{cc} m & n-m \end{array} \\ n & \begin{pmatrix} K_1 & K_2 \end{pmatrix} \end{array}$$

$$K^* = \begin{array}{cc} & \begin{array}{cc} m & n-m \end{array} \\ n^* & \begin{pmatrix} K_1^* & K_2^* \end{pmatrix} \end{array}$$

$$K \cong \widehat{K} \equiv K_1 K_{11}^{-1} K_1^T$$

$$K^* \cong \widehat{K}^* \equiv K_1^* K_{11}^{-1} K_1^T$$

Results from Other Authors

Method Name	σ_{rms}	Dataset[1]	Inputs[2]	Source
CWW	0.0666	SDSS-EDR	ugriz	Csabai et al. (2003)
Bruzual-Charlot	0.0552	SDSS-EDR	ugriz	Csabai et al. (2003)
ClassX	0.0340	SDSS-DR2	ugriz	Suchkov et al. (2005)
Polynomial	0.0318	SDSS-EDR	ugriz	Csabai et al. (2003)
Support Vector Machine	0.0270	SDSS-DR2	ugriz	Wadadekar (2005)
Kd-tree	0.0254	SDSS-EDR	ugriz	Csabai et al. (2003)
Support Vector Machine	0.0230	SDSS-DR2	ugriz+r50+r90	Wadadekar (2005)
Artificial Neural Network	0.0229	SDSS-DR1	ugriz	Collister & Lahav (2004)

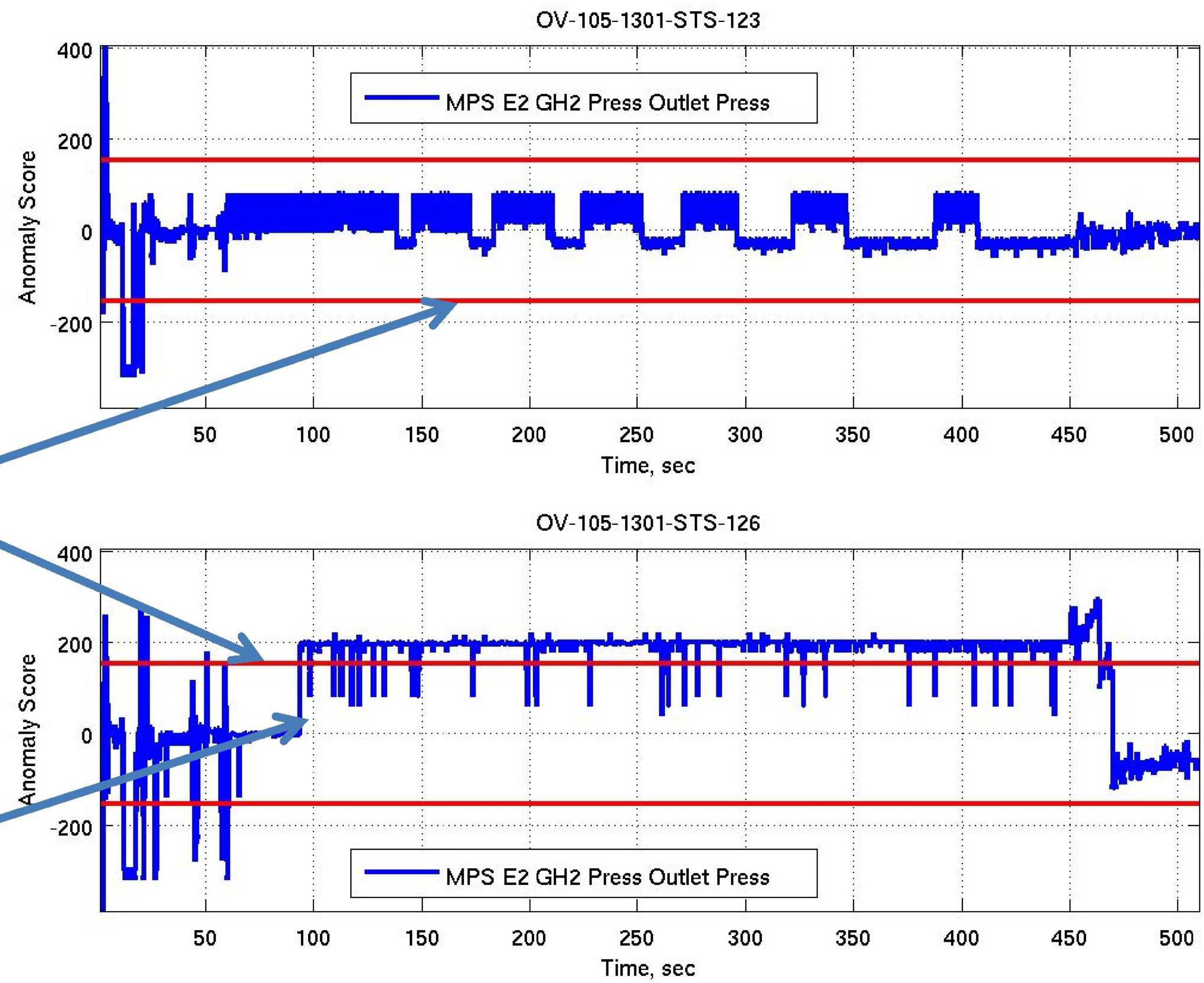
Summary of Our Results

Results: SDSS (DR3) Main Galaxy Sample

- Paper I: Compared linear, quadratic, Neural Networks and GPs on the SDSS
- With ONLY 1000 samples GPs performed well compared to the other methods
- Paper II: With *low-rank matrix inversion approximations* GPs performed better than all other methods

Virtual Sensor: STS-123 and STS-126

- Redlines correspond to 3-sigma nominal error rate on STS-123.
- STS-126 shows anomalous behavior after 93.6 seconds.



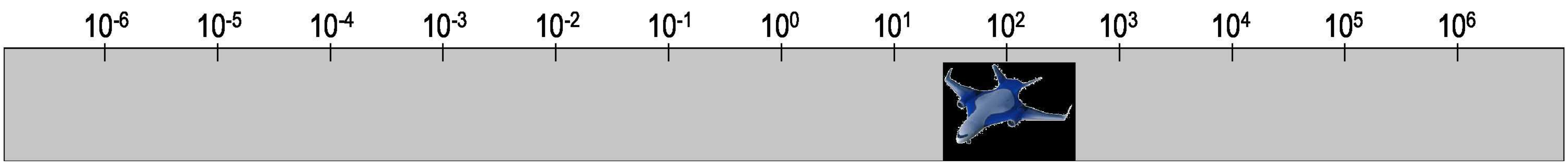
Summary of Research Needs in Aviation Safety

- Aircraft aging and durability
  - Full fundamental knowledge about legacy aircraft
  - Start on knowledge about likely emerging materials and structures
- On-board system failures and faults – airframe, propulsion, aircraft systems (physical and software)
  - Early prediction, detection and diagnosis
  - Prognosis
  - Mitigation
- Monitoring for problems before they become accidents
  - Vehicle issues
  - Airspace issues
- Loss-of-control
  - Understanding aircraft dynamics of current and future vehicles in damaged and upset conditions
  - Control systems robust to the unanticipated and anticipated
  - Aircraft guidance for emergency operation
- Flight in hazardous conditions
  - Modeling and sensing airframe and engine icing and icing conditions
  - Sensing and portraying environmental hazards
- New operations
  - Design of robust collaborative work environments
  - Design of effective, robust human-automation systems
  - Information management and portrayal for effective decision making

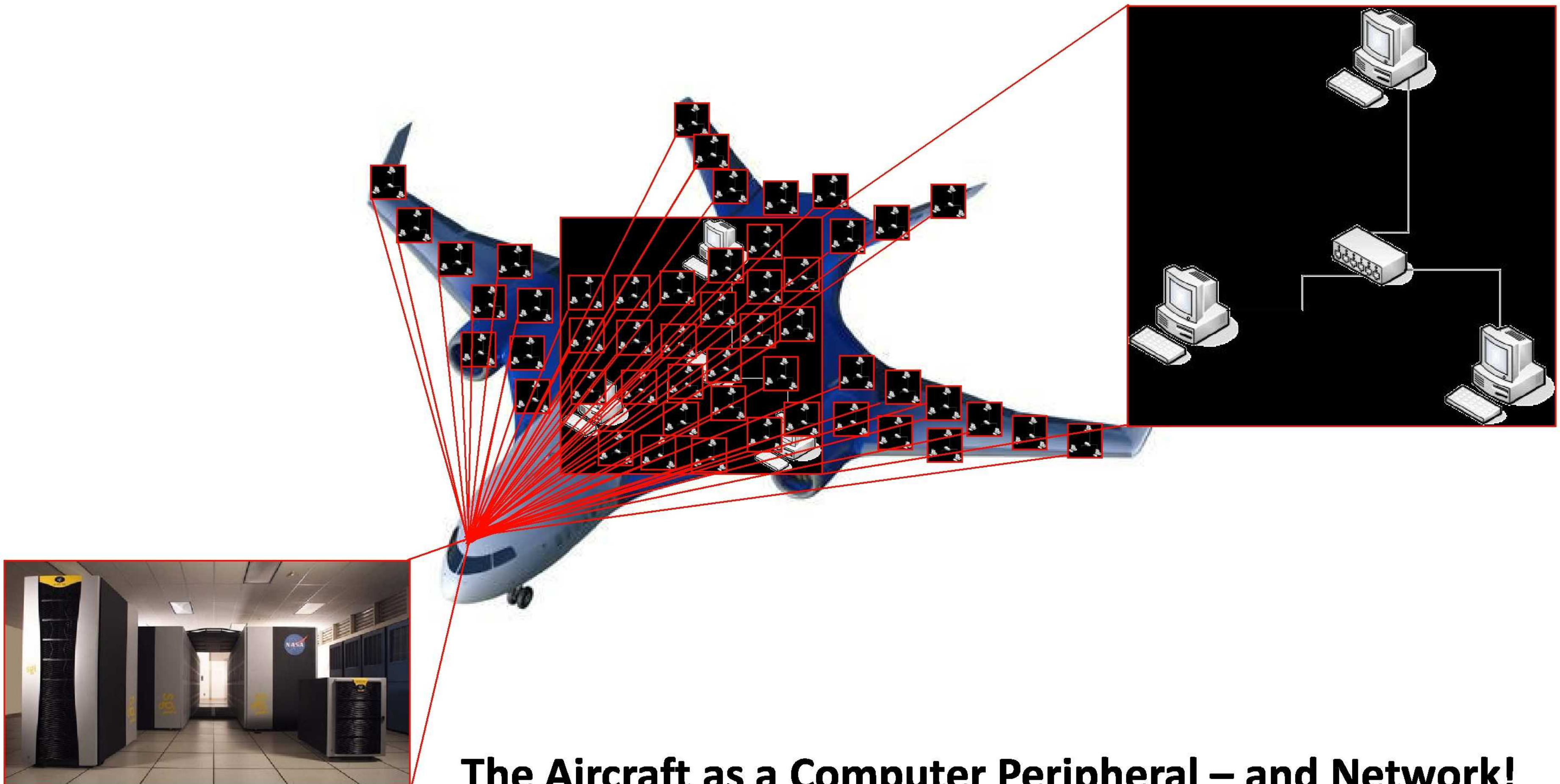
Integrated Vehicle Health Management

The Powers of Aviation Safety – 10^{-6} - 10^{6}

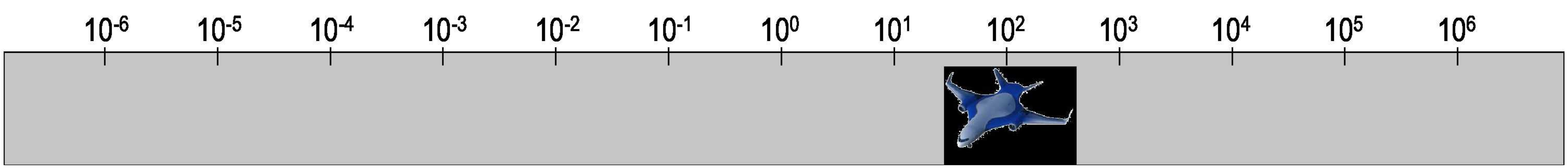
- There is no one 'silver bullet' – we must look at all contributors to safety
- Consider the space we must consider:
  - Safety at the smallest level
  - Safety spanning the nation (and the world!)
- Let us consider these different sizes, expressed as 'Powers of Ten'



10^2 – The Aircraft



The Aircraft as a Computer Peripheral – and Network!



Organization of IVHM

Principal Investigator: Ashok Srivastava
Project Scientist: Robert Mah
Project Manager: Claudia Meyer

- **Project Operations**
  Manager: Jeff Rybak
- **NRA**
  Manager: Lilly Spirkovska

ARC, DFRC, GRC, LaRC Center POCs

ARC APM: Steve Jacklin
DFRC POC: Mark Dickerson
GRC APM: Bob Kerczewski
LaRC APM: Sharon Graves

- ***Level 4***

- **Multidisciplinary Ground/ Flight Demos**
  - Leads: PI, PS, PM
- **Systems Analysis for Health Management**
  Lead: Mary Reveley
- **Research Test and Integration**
  Lead: Robert Mah
- **DASHlink**
  Lead: Elizabeth Foughty

- ***Level 3 Associate Principal Investigators***

- **Detection**
  API: John Lekki
- **Diagnosis**
  API: Rick Ross
- **Prognosis**
  API: Kai Goebel
- **Mitigation**
  API: Eric Cooper
- **Integrity Assurance**
  API: Eric Cooper

- ***Level 2***

Aircraft Systems | **Airframe** | **Propulsion Systems**

***Traditional** Aircraft Subsystems – well represented in Levels 1,3 and 4*

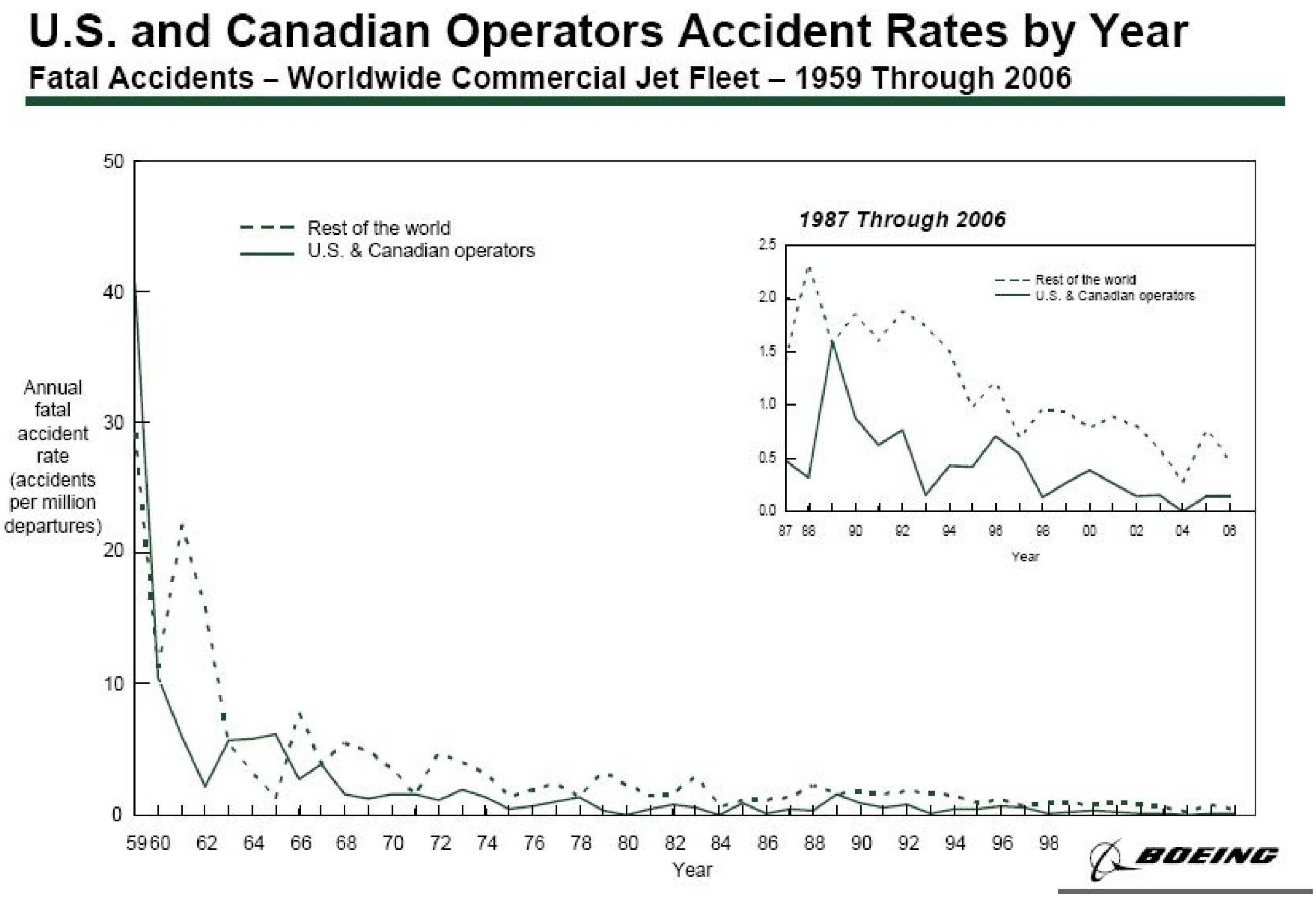
Software
Lead: Paul Miner

***Newly Recognized** Aircraft Subsystem*

- ***Level 1 Lead Researchers***

- **Advanced Sensors and Materials**
  Lead: Tim Bencic
- **Modeling**
  Lead: Kevin Wheeler
- **Advanced Analytics and Complex Systems**
  Lead: Nikunj Oza
- **Verification and Validation**
  Lead: Steve Jacklin

Recent Safety Advances



References

L. Connell, "Incident Reporting: The nasa aviation safety reporting system" ,*GSE Today,* pp. 66-68, 1999.

T.K. Landauer, D. Laham, and P. Foltz, "Learning human-like knowledge by singular value decomposition: A progress report," in *Advances in Neural Information Processing Systems,* M. I. Jordan, M. J. Kearnes, and S. A. Solla, Eds., vol. 10. The MIT Press, 1998. [online]. Available: cite-seer.ist.ppsu.edu/landauer/98learning.html.

T. Joachims, "A Probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization," in *Proceedings of ICML-97, 14th International Conference on Machine Learning,* D. H. Fisher Ed. Nashville, US: Morgan Kaufman Publishers, San Francisco, US, 1997, pp. 143-151.

I.T. Jolliffe, *Principle Components Analysis.* New York: Springer Verlag, 1986.

M.I. Jordan and R.A. Jacobs, "Hierarchical mixtures of experts and the EM algorithm, Tech. Rep. AIM-1440, 1993. [online]. Available: citeseer.ist.psu.edu/article/jordan94hierarchical.html.

J.W. Sammon, "A nonlinear mapping for data structure analysis," *IEEE Transactions on Computers,* Vol. C-18, pp. 401-409, 1969.

A. Ng. M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," 2001. [Online]. Available: citeseer.ist.psu.edu/ng01spectral.html.

C. Linde and R. Wales, "Work process issues in nasa's problem reporting and corrective action (praca) database," NASA Ames Research Center, Human Factors Division, Tech. Rep., 2001. [Online]. Available: human-factors.arc.nasa.gov/april01-workshop/2pg.linde3.doc.

References

References for slides on IMS

D. Dvorak and B. Kuipers. "Model-Based Monitoring of Dynamic Systems", *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence (IJCAI-89)*, Morgan Kaufman, Los Altos, CA., 1989.

R. Reiter. "A Theory of Diagnosis from First Principles", *Artificial Intelligence*, 32(1):57-96, Elsevier Science, 1987.

P.S. Bradley, O.L. Mangasarian, and W.N. Street. "Clustering via Concave Minimization", *Advances in Neural Information Processing Systems 9*, M.C. Mozer, M.I. Jordon, and T. Petsche(Eds.), pp 368-374, MIT Press, 1997.

P.S. Bradley and U. M. Fayyad. "Refining initial points for K-means clustering", in *Proceedings of the International Conference on Machine Learning (ICML-98),* pp 91--99, July 1998.

M. Ester, H-P Kreigel, J. Sander, and X. Xu. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", *Proceedings of the 2nd ACM SIGKDD*, pp 226-231, Portland, OR, 1996.

W.C. Hamscher. "ACP: Reason maintenance and inference control for constraint propagation over intervals", *Proceedings of the 9th National Conference on Artificial Intelligence*, pp 506-511, Anaheim, CA, July, 1991.

J.M Kleinberg. "Two Algorithms for Nearest-Neighbor Search in High Dimensions", *Proceedings of the 29th Annual ACM Symposium on Theory of Computing*, pp 599-608, El Paso, TX, May, 1997.

H.W. Gehman, et al., "Columbia Accident Investigation Board Report", U.S. Government Printing Office, Washington, D.C., August 2003.

References

References for slides on sequenceMiner

L. Kaufman and P.J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*,John Wiley and Sons, Inc., New York (1990).

T. Cormen, C. Leiserson, R. Rivest and C. Stein, *Introduction to algorithms*, The MIT Press; 2nd edition.

James W. Hunt and Thomas G. Szymanski, *A Fast Algorithm for computing Longest Common Subsequences*. Communications of the ACM, Volume 20, Issue 5 (May 1977),Pages: 350 - 353.

D. S. Hirschberg, *Algorithms for the Longest Common Subsequence Problem*, Journal of the ACM, Volume 24, Issue 4 (October 1977),Pages: 664 - 675.

D. S. Hirschberg, *A Linear Space Algorithm for computing Maximal Common Subsequences*, Communications of the ACM, Volume 18, Issue 6 (June 1975),Pages: 341 - 343.

L. Bergroth, H. Hakonen and T. Raita, *A Survey of Longest Common Subsequence Algorithms*, Proceedings of the Seventh International Symposium on String Processing Information Retrieval(SPIRE), 2000.

K. Sequeira and M. Zaki, *ADMIT: Anomaly based Data Mining for Intrusions*, Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(SIGKDD), 2002.

Scott Coull, Joel Branch and Boleslaw Szymanski, *Intrusion Detection: A Bioinformatics Approach*, Proceedings of the 19th Annual Computer Security Applications Conference(ACSAC), 2003.

A. Banerjee and J. Ghosh, *Clickstream Clustering using Weighted Longest Common Subsequence*, Proceedings of the 1st SIAM International Conference on Data Mining (SDM): Workshop on WebMining, 2001

T. Lane and C. Brodley, *Temporal sequence learning and data reduction for anomaly detection,* ACM Transactions on Information and System Security (TISSEC), Volume 2, Issue 3 (August 1999), Pages: 295 - 331.

A. N. Srivastava, *Discovering System Health Anomalies using Data Mining Techniques*, Proceedings of the 2005 Joint Army Navy NASA Airforce Conference on Propulsion, 2005.

References

References for slides on Orca

C.C. Aggarwal and P.S. Yu. Outlier detection for high dimensional data. In Proceedings of the ACM SIGMOD International Conference on Management of Data, 2001

F. Angiulli and C. Pizzuti. Past outlier detection in high dimensional spaces. In Proceedings of the Sixth European Conference on the Principle of Data Mining and Knowledge Discovery, pages 15-26, 2002

V. Barnett and T. Lewis. Outliers in Statistical Data. John Wiley & Sons, 1994

J.L. Bentley. Multidimensional binary search trees used for associative searching. Communications of the ACM, 18(9): 509-517, 1975

S. Berchtold, D. Keim, and H.-P. Kreigel. The X-tree: an index structure for high-dimensional data. In Proceedings of the 22nd International Conference on Very Large Databases, pages 28-39, 1996

G. Bisson, Learning in FOL with a similarity measure. In Proceedings of the Tenth National Conference on Artificial Intelligence, pages 82-87, 1992.

R.J. Bolton and D.J. Hand. Statistical fraud detection: A review (with discussion). Statistical Science, 17(3): 235-255,2002

M.M. Breunig, H. Kriegel, R.T. Ng. and j. Sander. LOF: Identifying density-based local outliers. In Proceedings of the ACM SIGMOD International Conference on Management of Data, 2000

W. Emde and D. Wettschereck. Relational instance-based learning. In Proceedings of the thirteenth International Conference on Machine Learning, 1996

E. Eskin, A. Arnold, M. Prerau. L. Portnoy, and S. Stolfo. A Geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data. In Data mining for Security Applications, 2002.

IDU Publications

Barrientos, F., Foughty, E., Matthews, B., and McIntosh, D. Bringing Web 2.0 to Government Research: A Case Study. Computer Human Interaction, 2009.

Bay, S.D., and Schwabacher, M. Mining Distance-Based Outliers in Near Linear Time with Randomization and a Simple Pruning Rule. KDD, 2003.

Bay, S.D., and Schwabacher, M. Near Linear Time Detection of Distance-Based Outliers and Applications to Security. SIAM Data Mining Conference, Workshop on Data Mining for Counter Terrorism and Security, San Francisco, CA, 2003.

Bhaduri, K., Kargupta, H. An Efficient Local Algorithm for Distributed Multivariate Regression in Peer-to-Peer Networks. SIAM International Conference on Data Mining, Atlanta, Georgia. pp. 153-164. 2008. (Best of SDM'08).

Bhaduri, K., and Srivastava, A. N. A Local Scalable Distributed Expectation Maximization Algorithm for Large Peer-to-Peer Networks. 15th ACM SIGKDD Conference On Knowledge Discovery and Data Mining.

Bhaduri, K., Stafanski, M., and Srivastava, A.N. Privacy Preserving Outlier Detection through Random Nonlinear Data Distortion. IEEE Transactions on Knowledge and Data Engineering.

Bieniawski. S., Kroo, I., and Wolpert, D.H. Flight Control with Distributed Effectors, AIAA Paper 2005-6074, Proceedings of the 2005 AIAA Guidance, Navigation, and Control Conference, San Francisco, CA, August 15-18, 2005.

Budalakoti, S., Srivastava, A. N. , and Otey, M. Anomaly Detection and Diagnosis Algorithms for Discrete Symbol Sequences with Applications to Airline Safety, IEEE Transactions on Systems, Man, and Cybernetics, Part C, 2008.

Coelho, C. K., Das, S., and Chattopadhyay, A. A Hierarchical Classification Scheme for Computationally Efficient Damage Classification, Journal of Aerospace Engineering, 2008

Christensen, D. and Das, S. Highly Scalable Matching Pursuit Signal Decomposition Algorithm. International Workshop on Structural Health Monitoring.

Das, S., Chattopadhyay, A., Srivastava, A.N. Classifying Induced Damage in Composite Plates using One Class Support Vector Machines, AIAA Journal, February 12, 2009

Figueroa, F., Aguilar, R., Schwabacher, M., Schmalzel, J., and Morris, J. Integrated System Health Management (ISHM) for Test Stand and J-2X Engine: Core Implementation. AIAA/ASME/SAE/ASEE Joint Propulsion Conference, 2008.

IDU Publications

Foster, L., Waagen, A.,. Aijaz, N., Hurley, M., Luis, A., Rinsky, J., Satyavolu, C., Way, M. J., Gazis, P., and Srivastava, A. N. Stable and Efficient Gaussian Process Calculations, Journal of Machine Learning Research, accepted for publication, Jan 14, 2009.

Kurklu, E., Morris, R., and Oza, N. Learning Points of Interest for Observation Flight Planning Optimization: A Preliminary Report. In Workshop on AI Planning and Learning, International Conference on Automated Planning and Scheduling (ICAPS), September 2007.

Lawson, J and Wolpert, D.H. Adaptive Programming of Unconventional Nano-Architectures, Journal of Computational and Theoretical Nanoscience, 3, 272-279, 2006.

Iverson, D.L. Martin, R. Schwabacher, M, Spirkovska, L. Taylor, W. Mackey, R. & Castle. J.P.. General purpose data-driven system monitoring for space operations, Proceedings of the AIAA Infotech@Aerospace Conference, Seattle, Washington, April 2009.

Martin, R. Investigation of Optimal Alarm System Performance for Anomaly Detection. In National Science Foundation Symposium on Next Generation of Data Mining and Cyber-Enabled Discovery for Innovation, Baltimore, MD, October 2007.

Martin, R., Unsupervised Anomaly Detection and Diagnosis for Liquid Rocket Engine Propulsion. In Proceedings of the IEEE Aerospace Conference, Big Sky, MT, March 2007.

Martin, R. An Investigation of State-Space Model Fidelity for SSME Data. In Proceedings of the International Conference on Prognostics and Health Management. IEEE, October 2008.

Martin, R. Schwabacher, M. Oza, N., Srivastava, A.N. Comparison of Unsupervised Anomaly Detection Methods for Systems Health Management Using Space Shuttle Main Engine Data. In Proceedings of the 54th Joint Army-Navy- NASA-Air Force Propulsion Meeting, Denver, CO, May 2007.

Martin, R. Approximations of Optimal Alarm Systems for Anomaly Detection. IEEE Transactions on Information Theory, 2007.

Oza, N. and Tumer, K. Key Real-World Applications of Classifier Ensembles. Information Fusion, Special Issue on Applications of Ensemble Methods, 9(1):4–20, 2008.

IDU Publications

Oza, N. Online Bagging and Boosting. In International Conference on Systems, Man, and Cybernetics, Special Session on Ensemble Methods for Extreme Environments, pp. 2340–2345, Institute for Electrical and Electronics Engineers, New Jersey, October 2005.

Oza, N., Srivastava, A.N., Strove, J. Improvements in Virtual Sensors: Using Spatial Information to Estimate Remote Sensing Spectra. In Proceedings of the International Geoscience and Remote Sensing Symposium, Institute for Electrical and Electronics Engineers, New Jersey, July 2005.

Oza, N. AveBoost2: Boosting for Noisy Data In Fifth International Workshop on Multiple Classifier Systems, pp. 31–40, Springer-Verlag, Cagliari, Italy, June 2004.

Oza, N. Boosting with Averaged Weight Vectors. In Fourth International Workshop on Multiple Classifier Systems, pp. 15–24, Springer-Verlag, Guildford, UK, June 2003.

Oza, N., Tumer, I., Tumer, K., and Huff, E. Classification of Aircraft Maneuvers for Fault Detection. In Proceedings of the Fourth International Workshop on Multiple Classifier Systems, pp. 375–384, Surrey, UK, June 2003.

Rajnarayan, D. and Wolpert, D.H., Exploiting Parametric Learning to Improve Black-Box Optimization, Proceedings of ECCS 2007, J. Jost et al. (Ed.)

Rajnarayan, D., Wolpert, D.H., Kroo, I. Optimization Under Uncertainty Using Probability Collectives, Proc. 11 AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, Portsmouth, VA, AIAA-2006-7033, 2006.

Schwabacher, M., Aguilar, R., and Figueroa, F. Using Decision Trees to Detect and Isolate Simulated Leaks in the J-2X Rocket Engine. JANNAF 6th Modeling and Simulation Subcommittee / 4th Liquid Propulsion Subcommittee / 3rd Spacecraft Propulsion Subcommittee Joint Meeting, 2008. ITAR Restricted. To appear.

Schwabacher, M., Aguilar, R., and Figueroa, F. Using Decision Trees to Detect and Isolate Simulated Leaks in the J-2X Rocket Engine. IEEE Aerospace Conference, 2009. To appear.

Schwabacher, M., and Waterman, R. Pre-Launch Diagnostics for Launch Vehicles. IEEE Aerospace Conference, 2008.

Schwabacher, M., and Goebel, K. A Survey of Artificial Intelligence for Prognostics. AAAI Fall Symposium, 2007.

Schwabacher, M., Oza, N., and Matthews, B. Unsupervised Anomaly Detection for Liquid-Fueled Rocket Propulsion Health Monitoring. AIAA Infotech@Aerospace Conference, 2007.

Schwabacher, M. Machine Learning for Rocket Propulsion Health Monitoring. SAE World Aerospace Congress, 2005.

IDU Publications

Schwabacher, M. A Survey of Data-Driven Prognostics. AIAA Infotech@Aerospace Conference, 2005.

Schwabacher, M., Oza, N., and Matthews, B. Unsupervised Anomaly Detection for Liquid-Fueled Rocket Propulsion Health Monitoring. Journal of Aerospace Computing, Information, and Communication (to appear).

Srivastava, A. N. and Das, S. Detection and Prognostics for Low-Dimensional Systems, IEEE Transactions on Systems Man and Cybernetics-C, 2008, August 29, 2008.

Srivastava, A.N., Oza, N., and Stroeve, J. Virtual Sensors: Using Data Mining to Efficiently Estimate Spectra. IEEE Transactions on Geosciences and Remote Sensing, Special Issue on Advances in Techniques for Analysis of Remotely Sensed Data, 43(3):590–600, 2005.

Srivastava, A.N., Stroeve, J., and Oza, N. Using Kernel Methods to Detect Clouds, Snow, Ice and other Geophysical Processes. In Transactions of the American Geophysical Union, Fall Meeting Supplement, pp. C12A08–65, June 2003.

Tumer, K., and Oza, N. Input Decimated Ensembles. Pattern Analysis and Applications, 6(1):65–77, 2003.

Wolff, R., Bhaduri, K., Kargupta, H. A Generic Local Algorithm for Mining Data Streams in Large Distributed Systems. IEEE Transactions on Knowledge and Data Engineering (accepted in press). 2008.

Wolpert, D.H. and Kulkarni, N., Game-theoretic Management of Interacting Adaptive Systems, Proc. 2008 NASA/ESA Conference on Adaptive Hardware and Systems, in press.

Wolpert, D.H., and Lee, C.F. An adaptive Metropolis-Hastings scheme: sampling and optimization, Europhysics Letters, 76, 353-359, 2006.

Wolpert, D.H., Strauss, C.E.M., Rajnarayan, D. Advances in Distributed Optimization using Probability Collectives, Advances in Complex Systems, 9, 2006.