Data Mining for Aviation Safety: Algorithms and Future Development

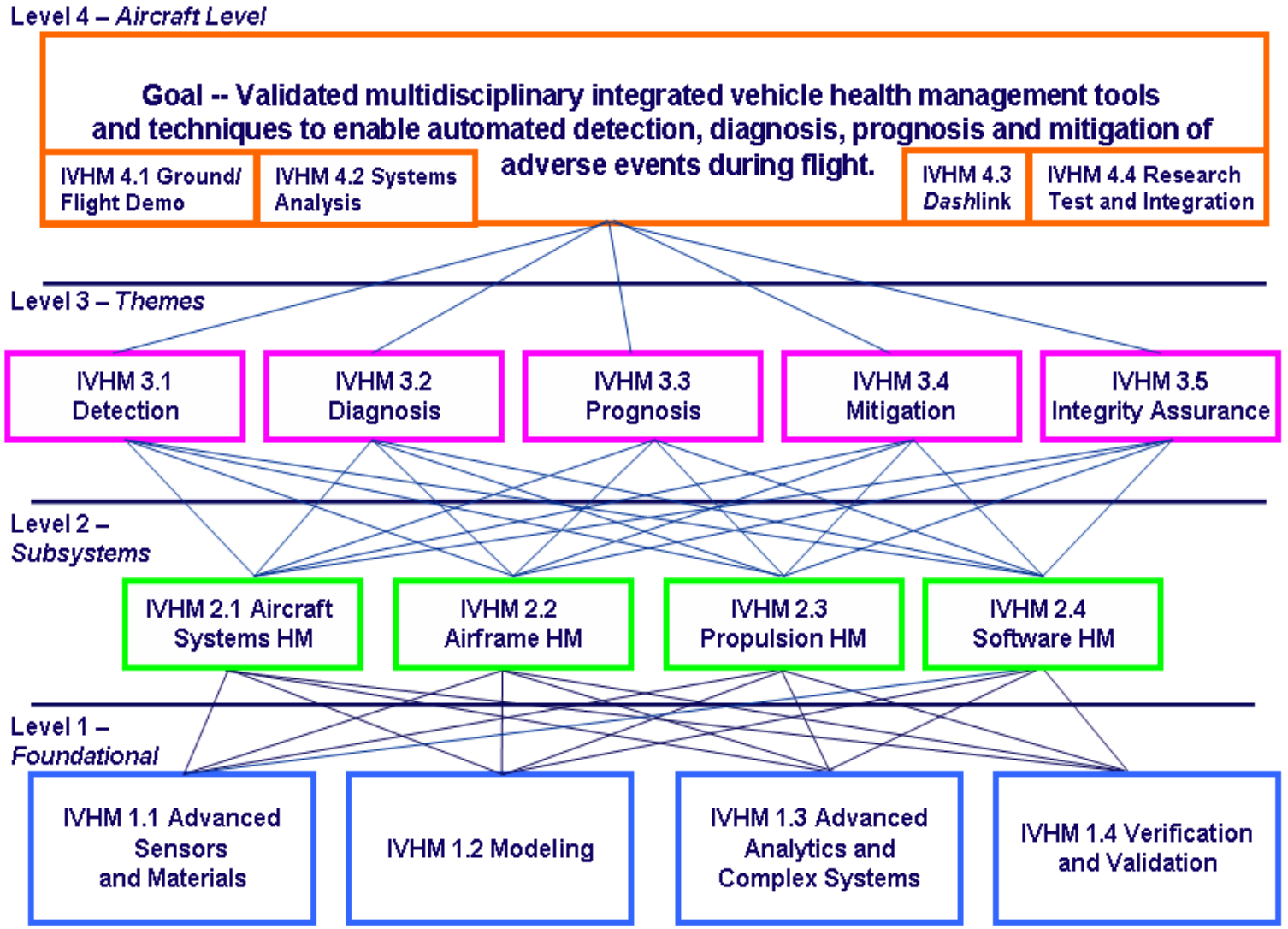
Ashok N. Srivastava, Ph.D.
Principal Investigator, IVHM Project
Group Lead, Intelligent Data Understanding
ashok.n.srivastava@nasa.gov

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

Integrated Vehicle Health Management: An Aviation Safety Project



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

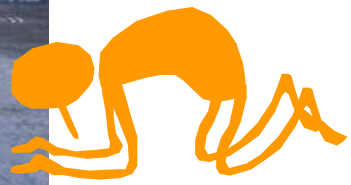
- NASA Aeronautical Research Mission Directorate- IVHM Project
- NASA Engineering and Safety Center
- Exploration Systems Mission Directorate Exploration Technology Development Program, ISHM Project
- Science Mission Directorate

Team Members are NASA Employees, Contractors, and Students.

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.
- Emphasis is on
  - Discovery algorithms: uncovering the unexpected
  - Scalability
  - Fleet-wide or system-wide issues in aeronautics

The Forensic (Historic) Approach to Accident Prevention



VS….

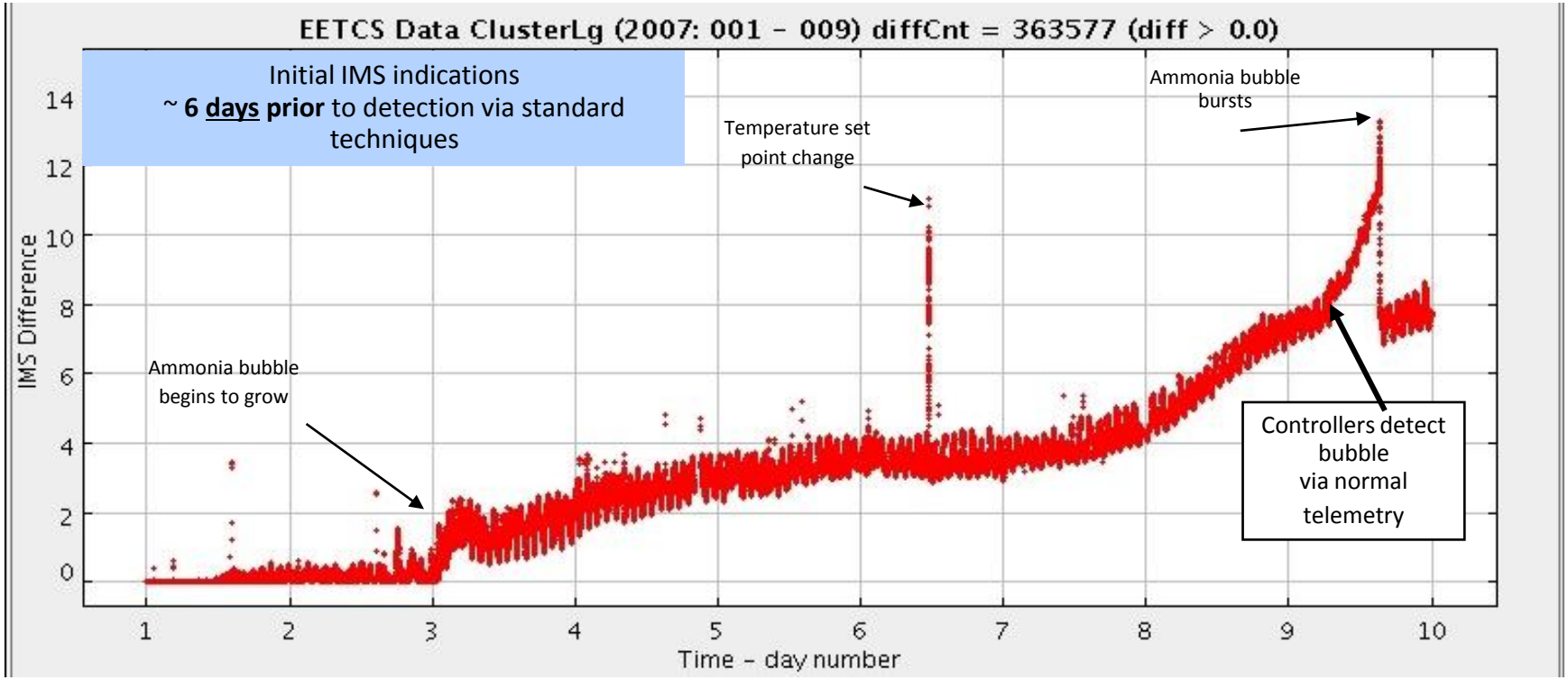
… a More Prognostic Approach

Leads to

Decisions

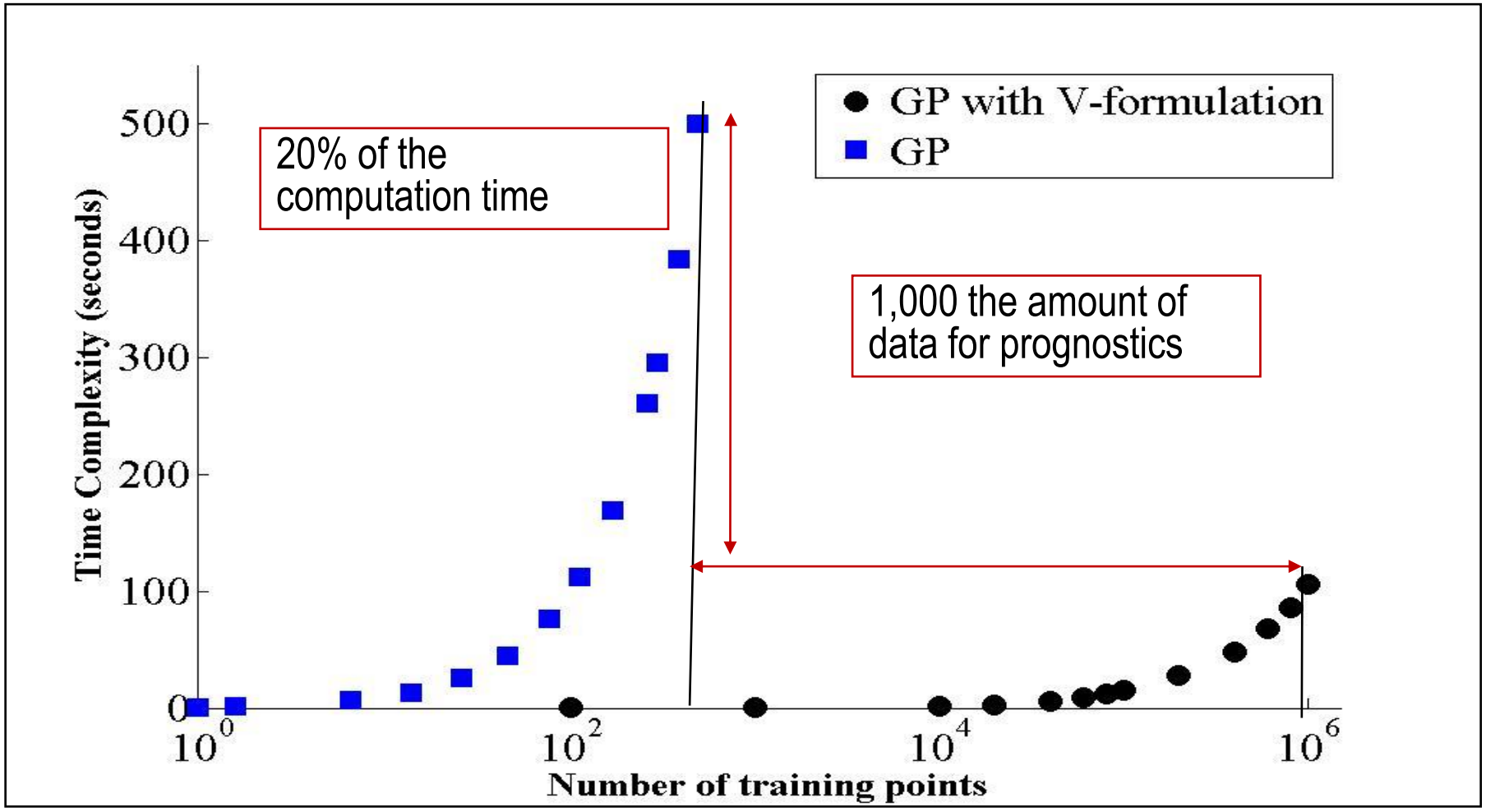
before an Accident or Incident Occurs

Discovery of Anomalies



- In early January 2007, International Space Station 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

Scalability and Accuracy



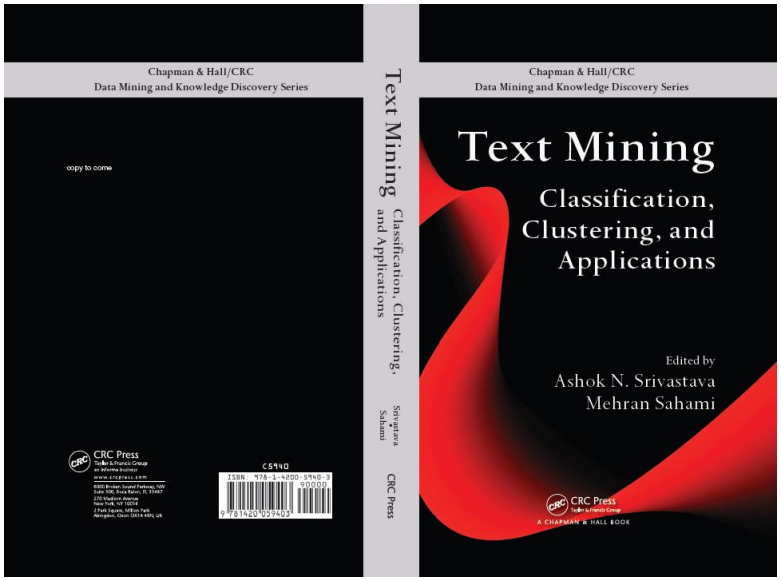
Key areas of research in data mining

Research Topic Areas

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

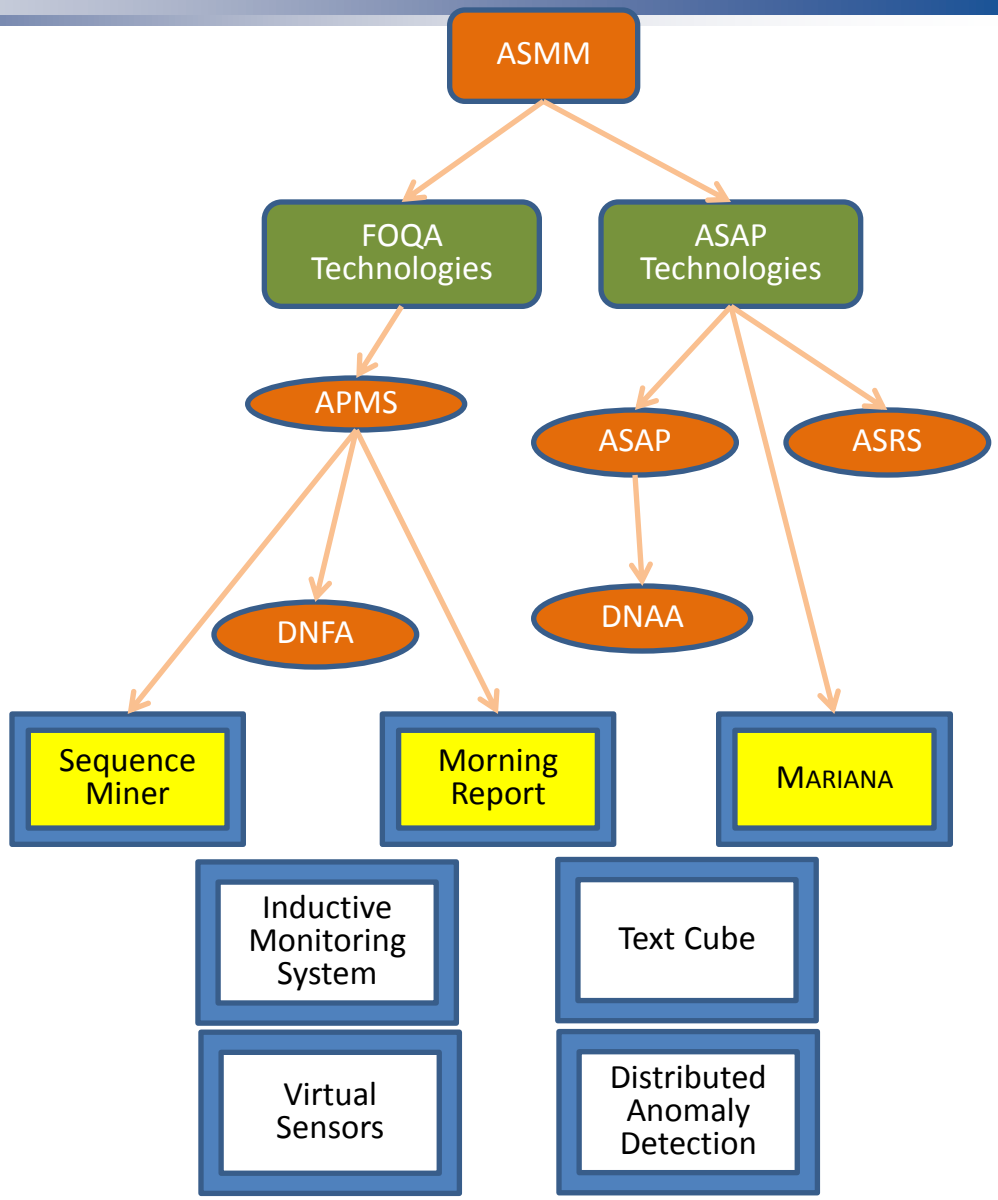
Application Areas

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



Program Background

- Genesis: Aviation System Monitoring & Modeling (ASMM)
- ASMM focus
  - Automated discovery tools
  - Data fusion
  - Finding unexpected events
  - Trends and Causation
- Within confidentiality constraints
- Customers / Stakeholders
  - FAA
  - Air carriers
  - Organized labor
- Milestones
  - FY09: Deployment of data mining tools within ASIAS
  - FY12: Forecasting technology that has the ability to predict at least 3 known anomalies in real or emulated data of large, fleetwide, heterogeneous data sources.
  - Hampered by lack of access to ASIAS data



Algorithms for Distributed Data Mining (DDM) in Large Asynchronous Networks

Kanishka Bhaduri, Ph.D.

Mission Critical Technology Inc.

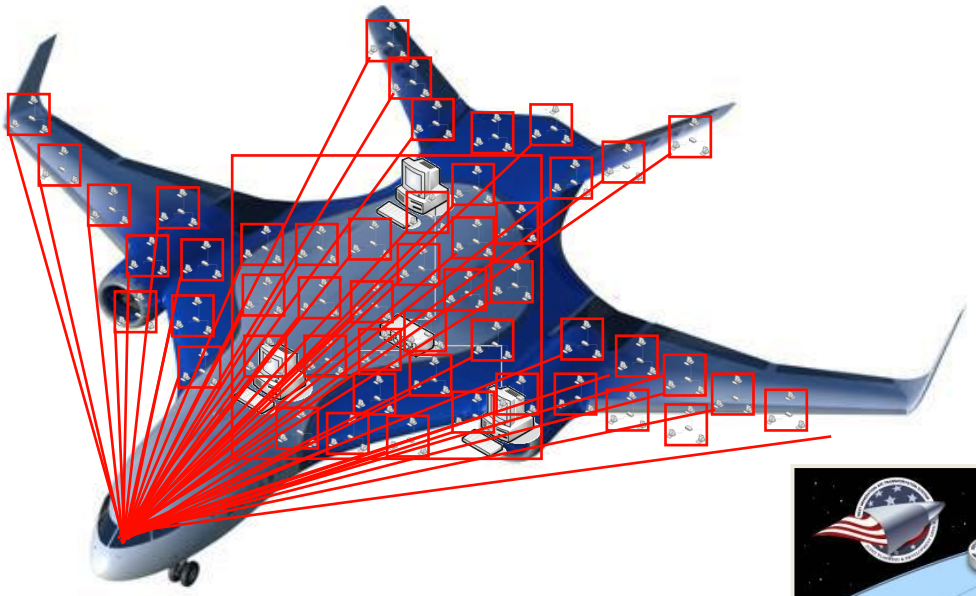
NASA Ames Research Center

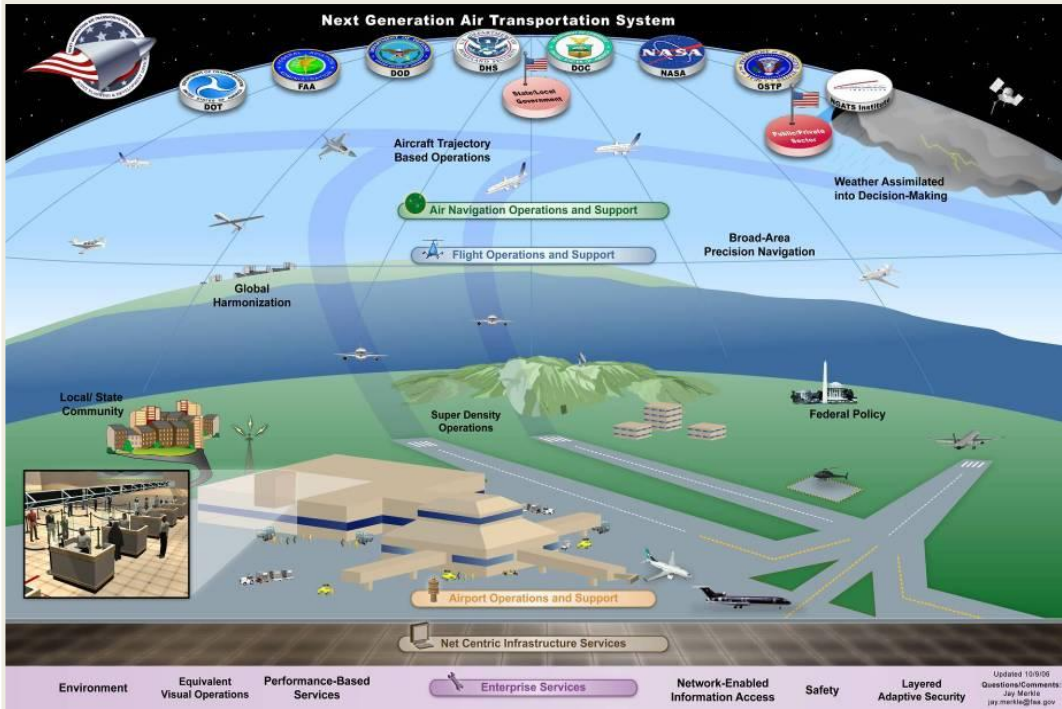
•**Joint work with: R. Wolff, C. Giannella, H. Kargupta, A. N. Srivastava**

Data Mining and Distributed Data Mining

- Data Mining:
  - Discovery of actionable information from large databases with emphasis on:
    - Scalability
    - Communication
    - storage
- Distributed data mining (DDM)
  - Mining data when data and computing resources are distributed

Possible Aviation Related Distributed Systems





Typical Problem Statement

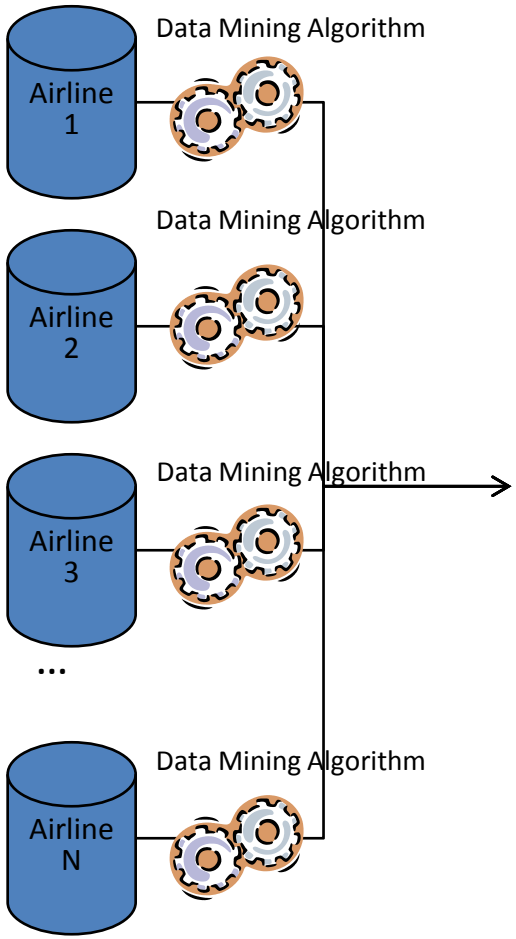
- Consider large network of nodes
  - Each node has local data which change over time
  - Each node exchanges messages
- Develop data mining algorithms for mining global data
- Constraints:
  - low communication overhead
  - no synchronization
  - failure resistance
  - **result *provably* correct with respect to centralized techniques**

Provable Correctness

- If a distributed data mining algorithm is ‘provably correct’ that means that the algorithm will give you the same answer if the data is **centralized** or **distributed**.
- Most attempts at distributed data mining do not obey this principle.
- This verification and validation is **critical** for safety applications.

The way you happen to store data should not change the results of a safety study.

Provable Correctness of Distributed Data Mining Algorithms



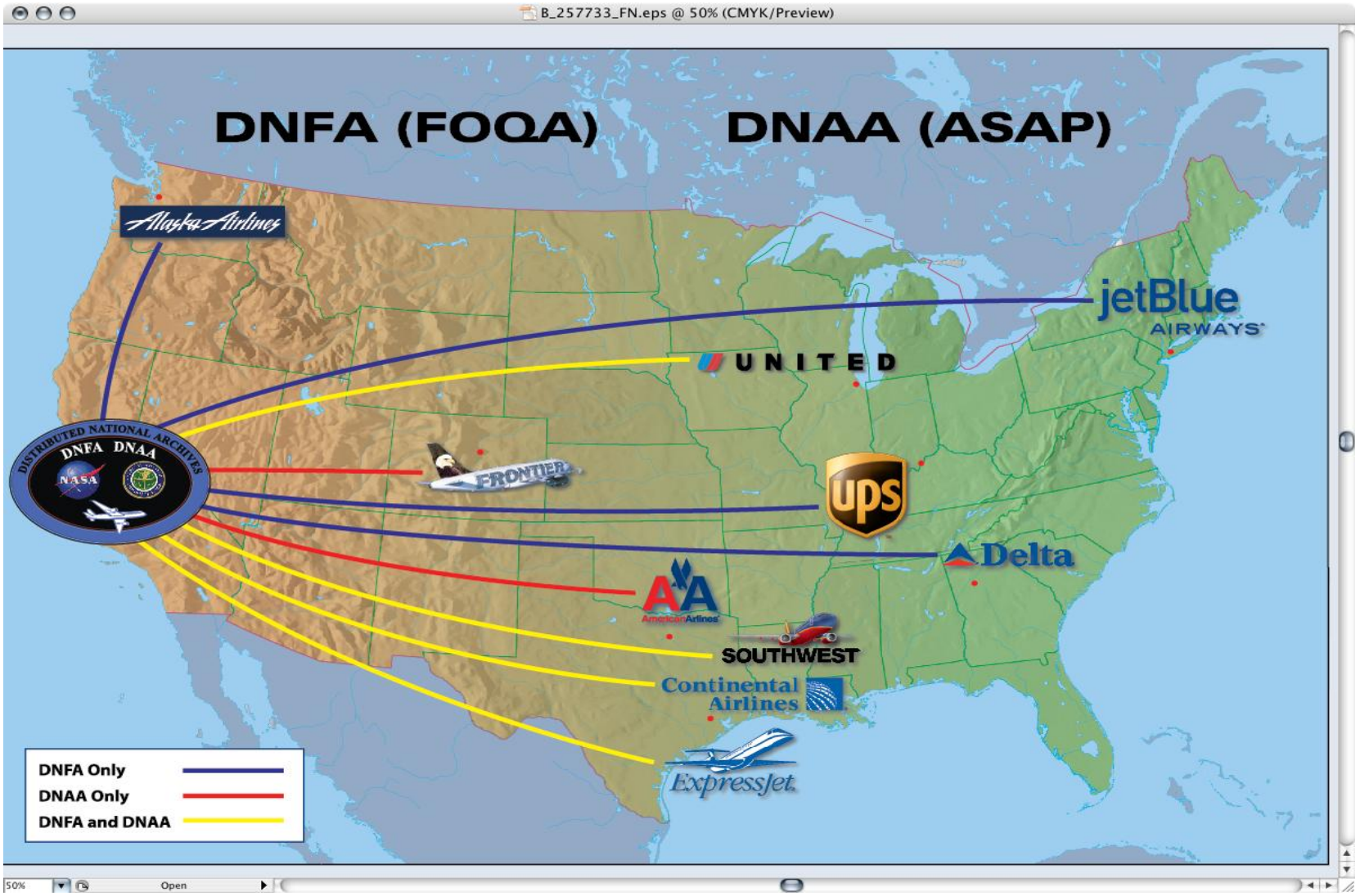
Proprietary Data

Data from real aircraft

Limited publication and distribution rights

- Suppose the output of each data mining algorithm gives the top anomalies in each database.
- The consensus of the output of the algorithms could indicate "**no significant anomalies present**" in the databases.
- If we centralize the data and then analyze it, the algorithm must still say "**no significant anomalies present.**"
- For non-provably correct algorithms, there may be a disagreement.

Distributed National Archives (circa 2007)



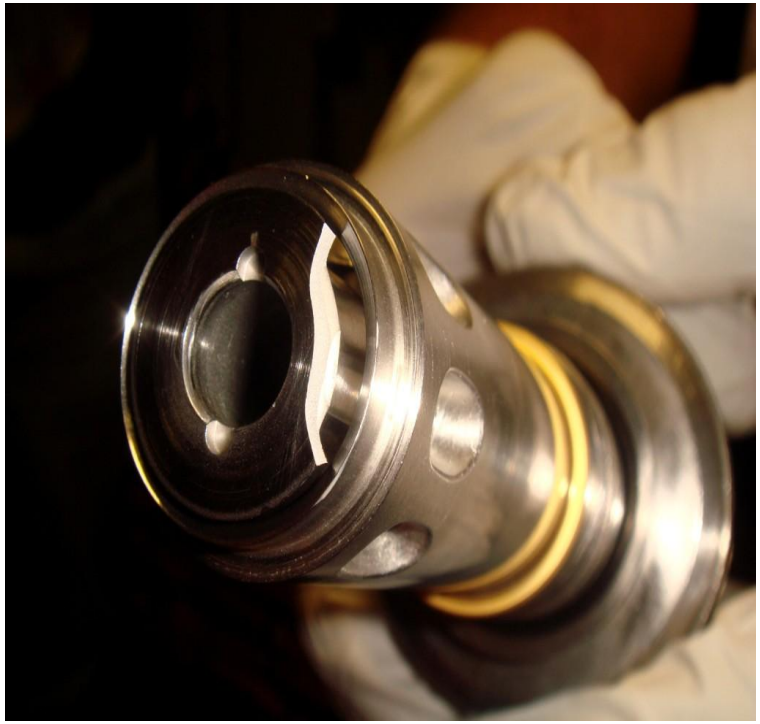
Preliminary Results of Flow Control Valve Data Mining Activity Supporting the Flight Readiness Review for STS-119

Ashok N. Srivastava Ph.D.
Principal Investigator, ARMD-IVHM
Data Mining Group Lead
Dave Iverson, ARC
Bryan Matthews, SGT
February 17, 2009

INTEGRATED VEHICLE HEALTH MANAGEMENT

Overview

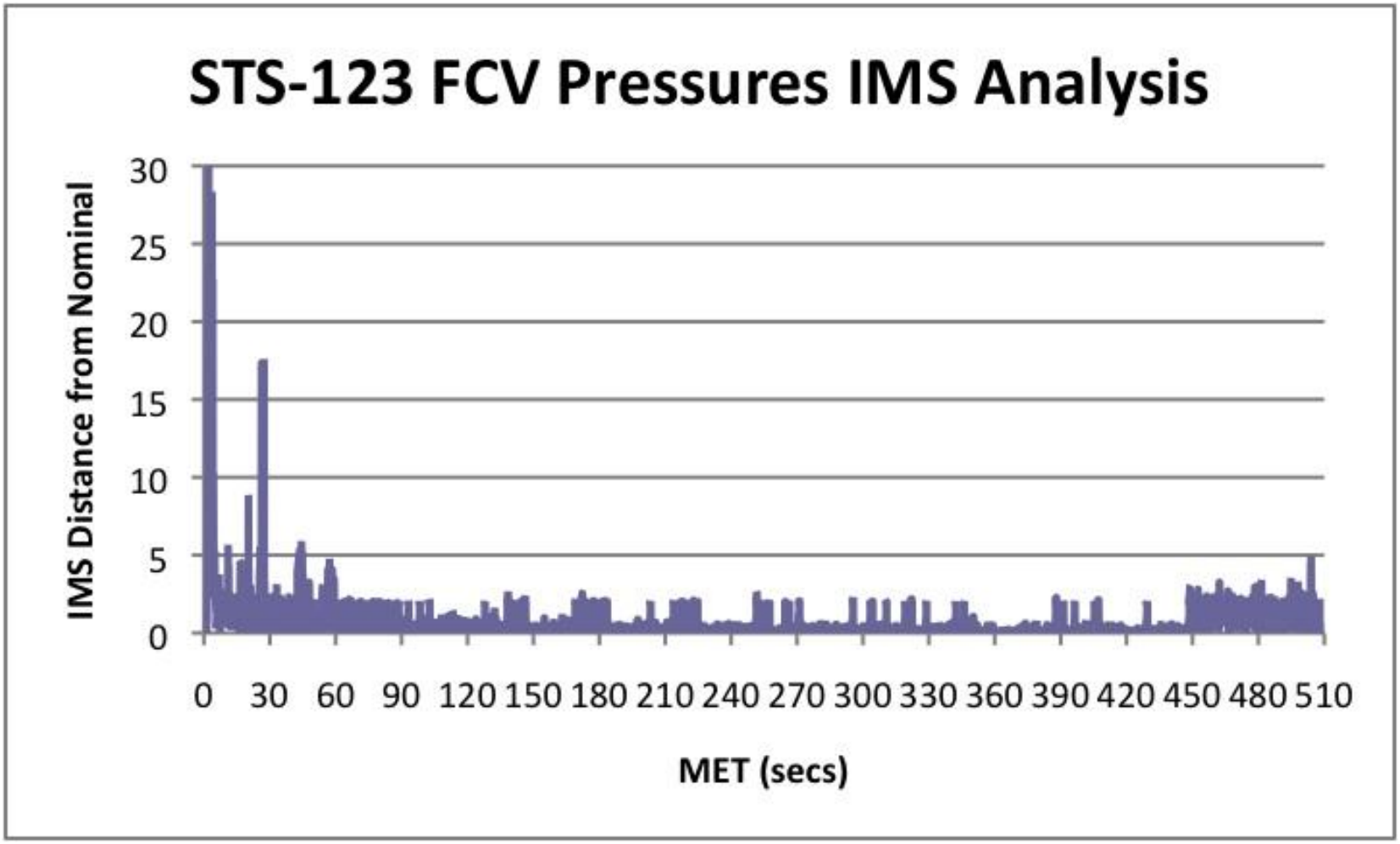
- Ashok received a request to support the Flight Readiness Review for STS-119 which was scheduled for 2/20/09 as the Data Mining Subject Matter Expert.
- Data mining algorithms developed at NASA were applied to these data to determine whether any anomalies can be detected in STS-126 and its predecessor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



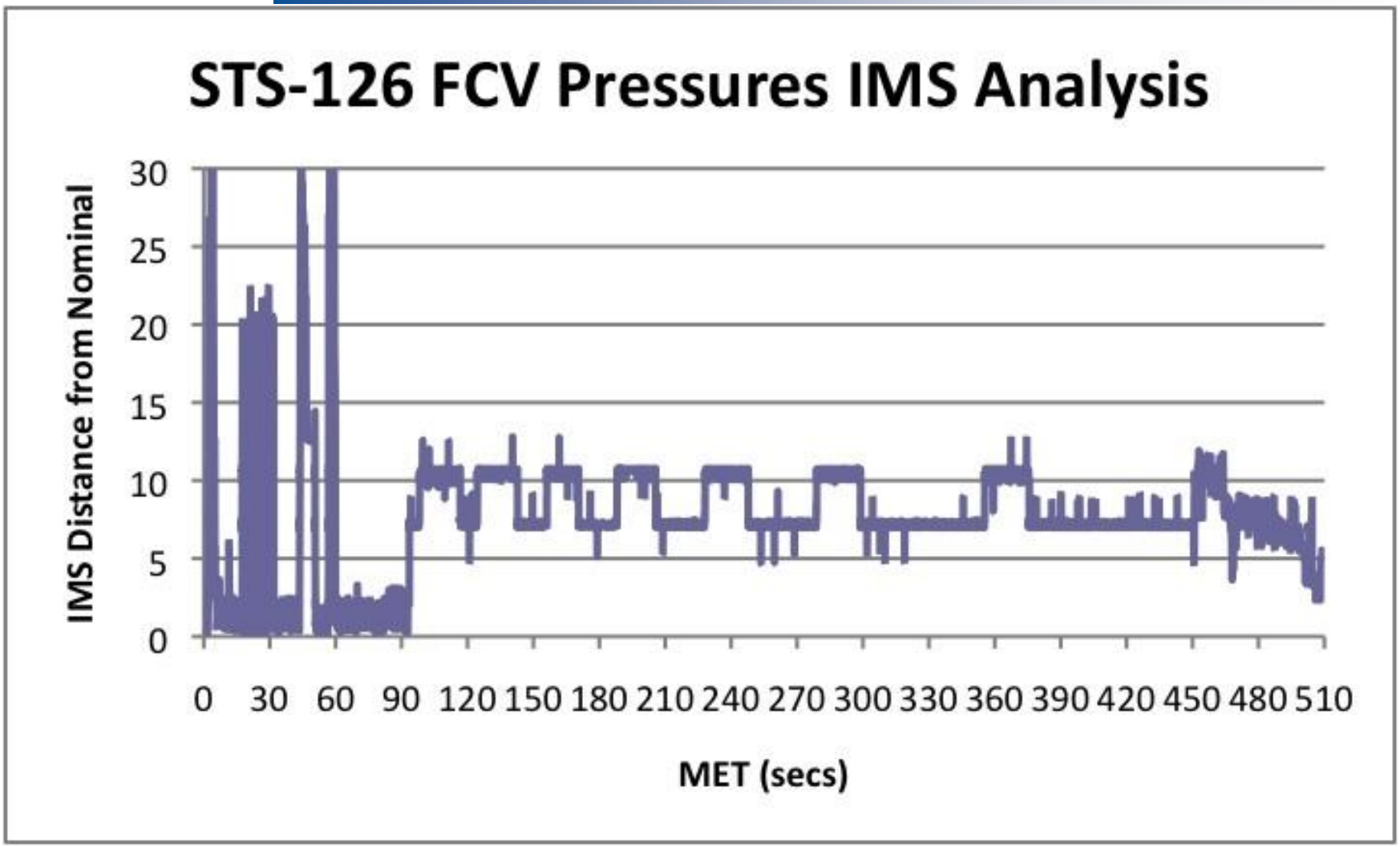
Algorithms and Data

- 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.

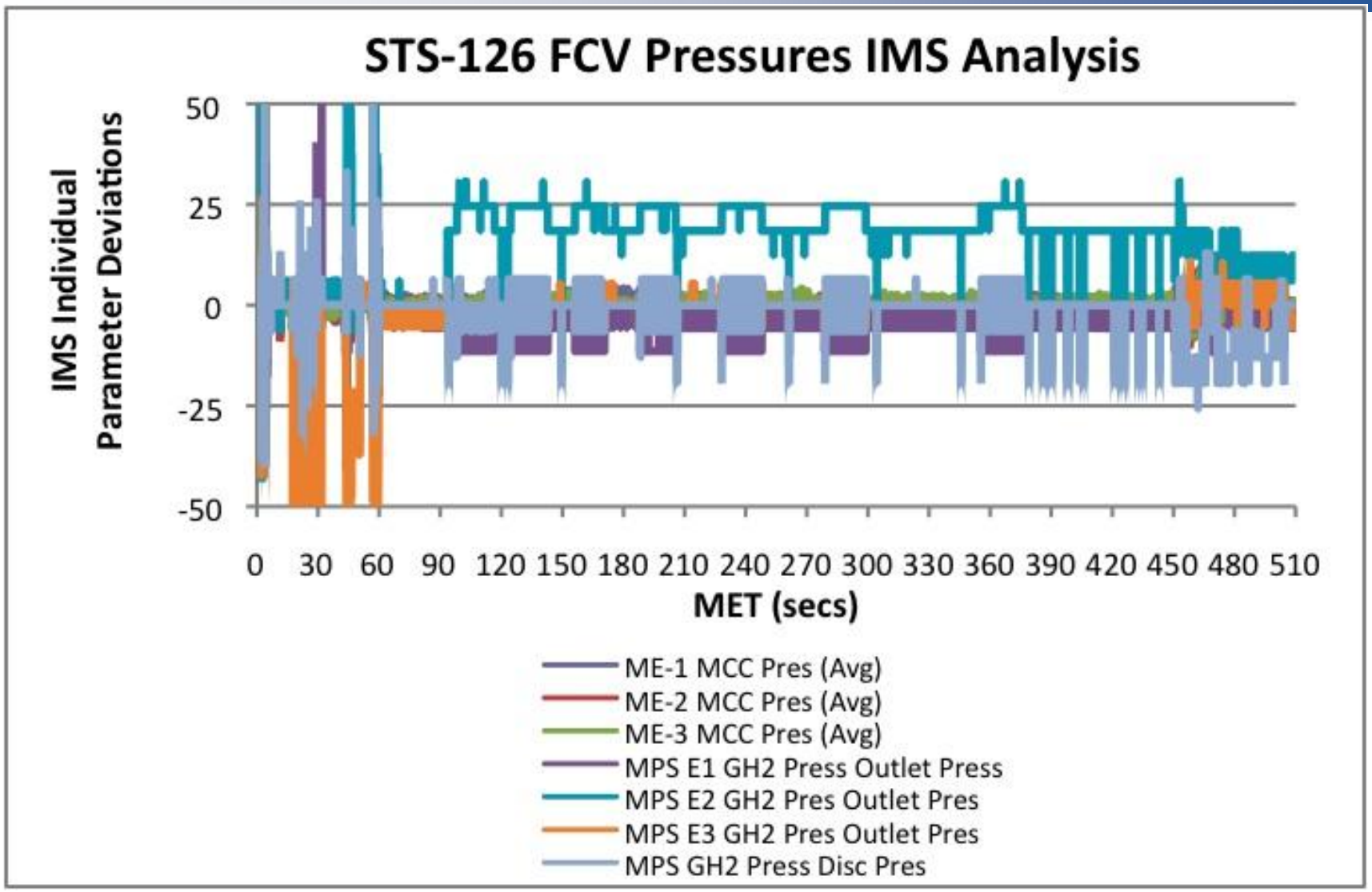
IMS Anomaly Score



IMS Anomaly Score



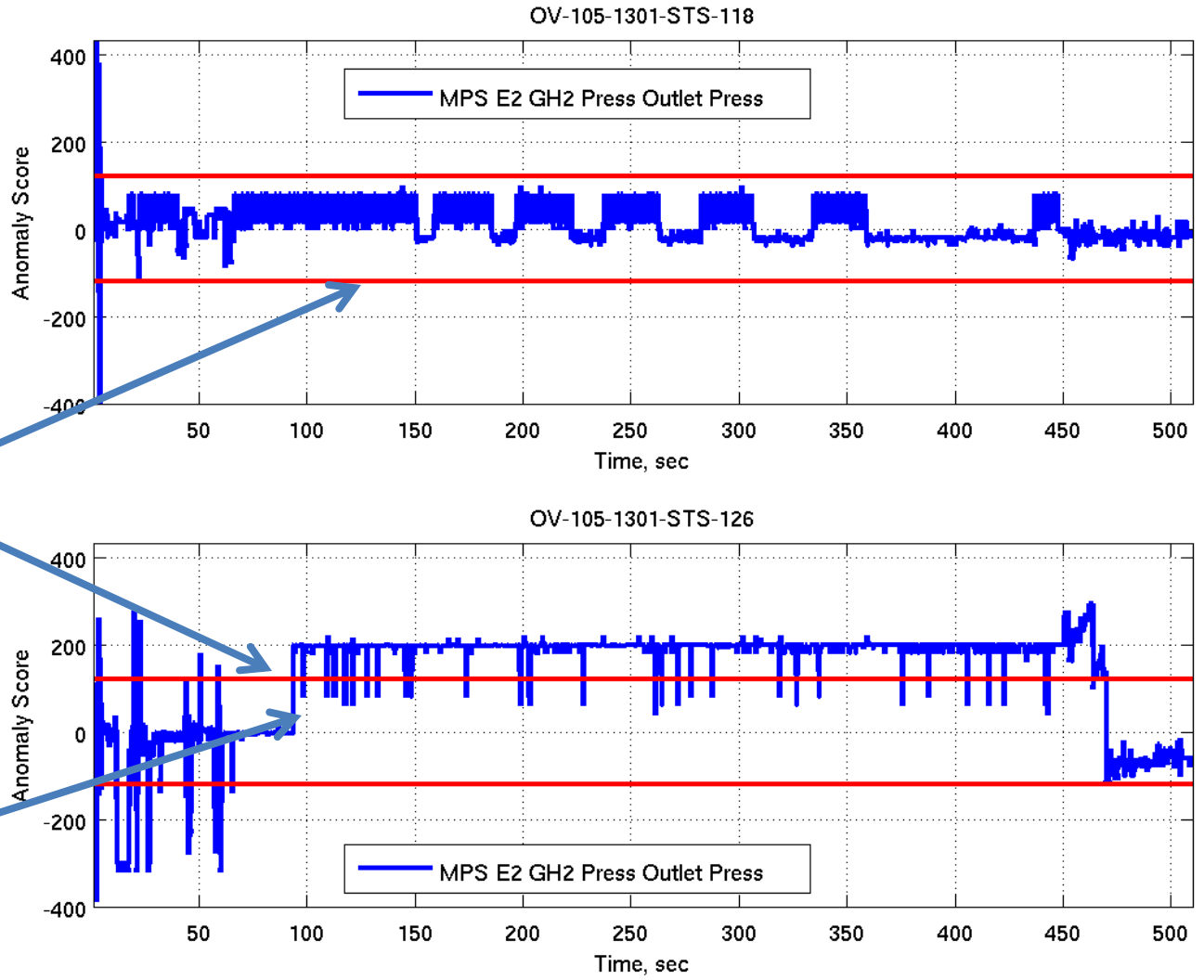
IMS Anomaly Score



Virtual Sensor: STS-118 and STS-126

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

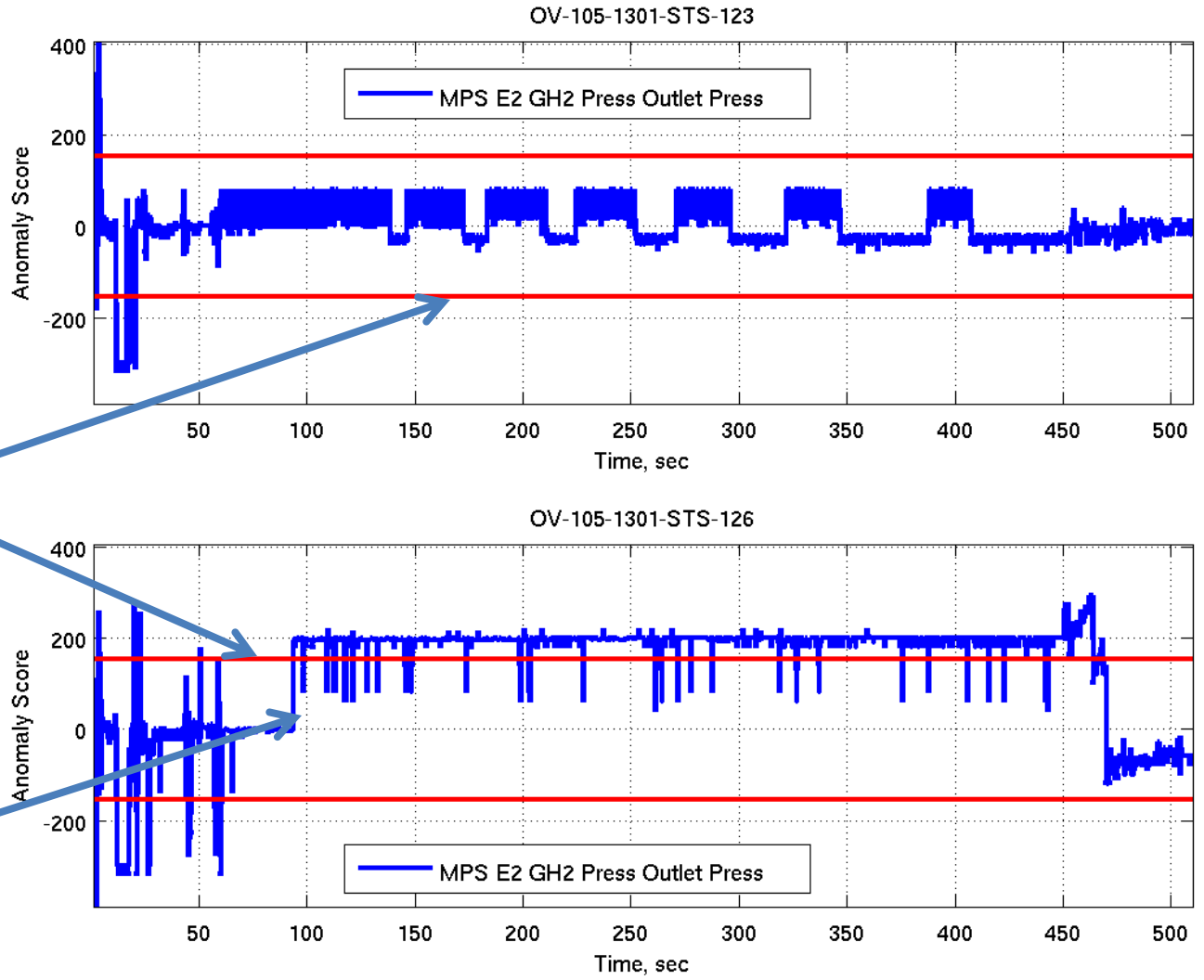
- STS-126 shows anomalous behavior after 93.6 seconds.



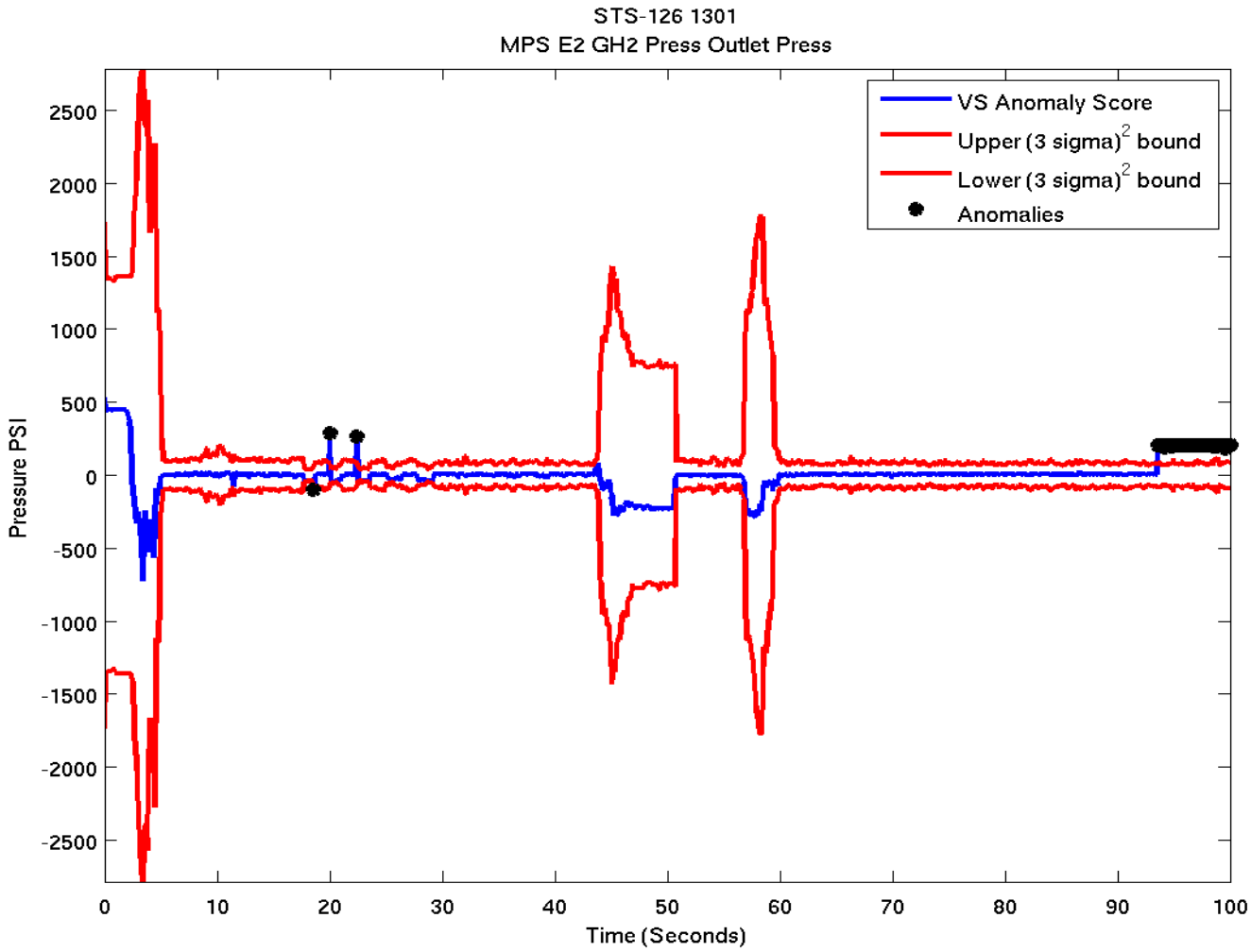
Virtual Sensor: STS-123 and STS-126

- Redlines correspond to 3-sigma nominal error rate on STS-123.

- STS-126 shows anomalous behavior after 93.6 seconds.



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.

SequenceMiner

Identifying Anomalous Flights from Discrete Data

Problem Specification

- Develop an approach to model the behavior of discrete sensors in an aircraft during flight.
- Focus is on primary sensors that record pilot actions.
- The aim is to discover atypical behavior that has possible operational significance.

Solution

We developed sequenceMiner,

Each flight is analyzed as a sequence of events, taking into account the order in which switches change values as well as the frequency of occurrence of switches.

Two Tasks:

Given a group of flights, find flights that are anomalous compared to the rest of the flights in the group.

Given a flight known to be anomalous, describe the anomalies in the flight and the degree to which they are anomalous.

Incorporating Operational Information

Switch Weight:

A weight is attached to each switch.

- A measure of its importance to flight based on a discussion with SMEs.

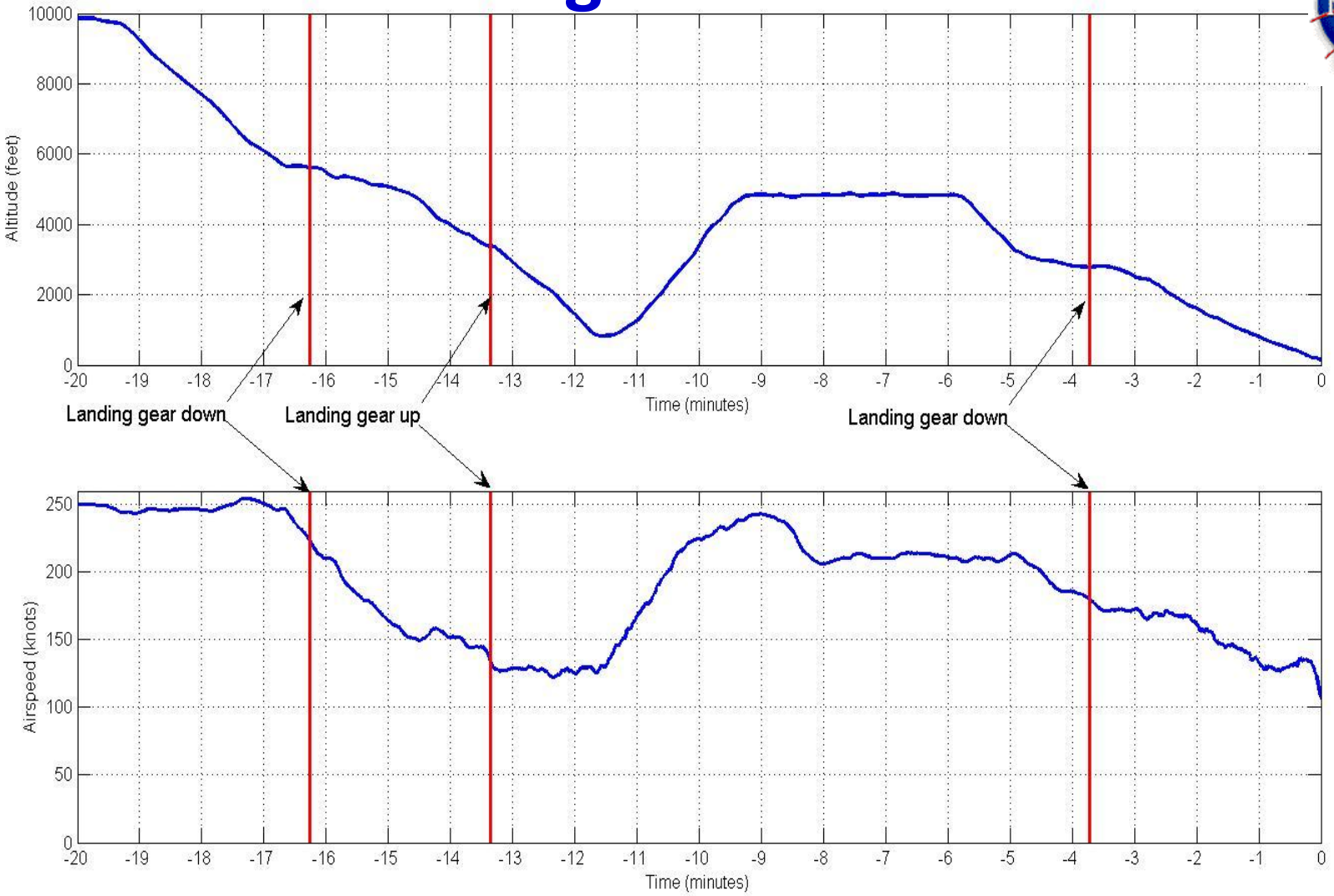
In the measure of similarity, the weights of the switches as well as the number of switches out of sequence are considered.

Ignore deviations with a small time gap:

Suppose the algorithm finds that a switch was not pressed when it was expected.

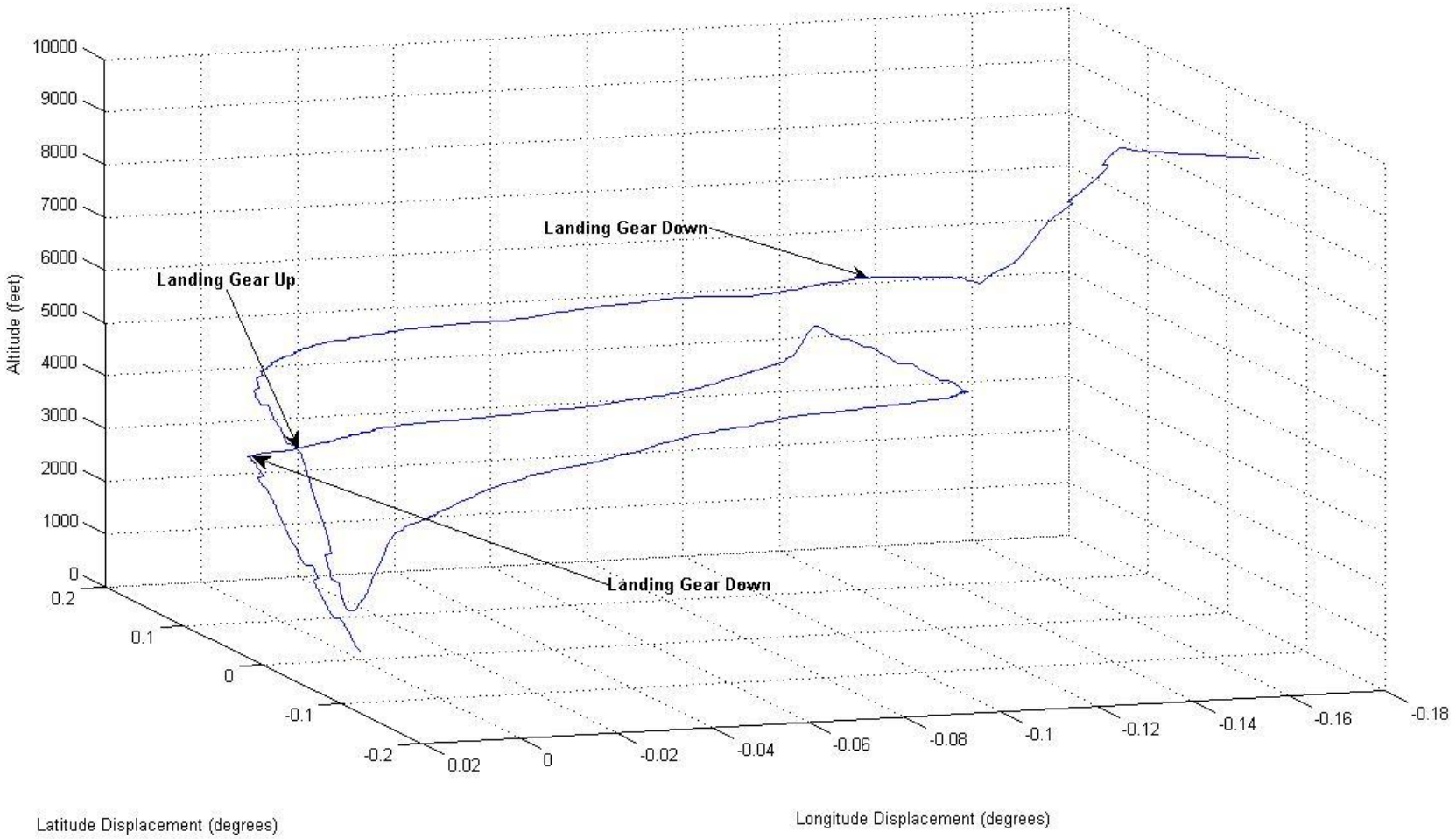
- It searches to see if the switch was pressed within one minute of the expected time.
- If the switch was pressed within a minute of the time point, it ignores the alarm.

Flight 1147



Captain Cirino's comments: "Landing gear goes up and down more than once. Go-around. Unable to determine pilot or ATC-related. Low descent after bringing landing gear up. Needs to be investigated (with flap information applied)."

Flight 1147 Flight Path



Change-in-Runway Study

There is evidence that a change in landing runway during approach is a causal factor in computer errors.

Can sequenceMiner find evidence of mode confusion?

Flight Management System (FMS) Switch Activations on Approach with a CIR Event

			seconds to touchdown					
	639	639	457	457	457	305	305	266
on	▲						▲	▲
off		▼	▼	▼	▼	▼		
	AP_Heading_Select_Mode	AP_LNAV_Mode	AP_Engaged_L	AP_Engaged_R	Autothrottle_Engaged	AP_Heading_Select_Mode	AP_Localizer_Engaged	AP_Glide_Slope_Engage

SME opinion: No evidence of mode confusion.

Switch Activations during a Change to a Parallel Runway

(NOTE: Approximately same time interval as previous slide)

	seconds to touchdown																							
	600	600	365	365	364	364	364	359	359	359	358	358	351	309	309	309	309	309	309	308	308	308	302	288	275	258
on	▲		▲					▲	▲				▲	▲	▲				▲			▲	▲	▲		▲
											✔	✔		✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
off		▼		▼	▼	▼	▼			▼	▼	▼				▼	▼	▼		▼	▼				▼	
	AP_Heading_Select_Mode	AP_LNAV_Mode	AP_Localizer_Engaged	Autothrottle_Engaged	AP_Engaged_L	AP_Engaged_R	AP_Heading_Select_Mode	AP_Engaged_L	AP_Engaged_R	Flight_Director_On_R	AP_Engaged_L	AP_Engaged_R	AP_Glide_Slope_Engage	AP_Engaged_L	AP_Engaged_R	AP_Glide_Slope_Engage	AP_Localizer_Engaged	Flight_Director_On_L	Flight_Director_On_R	AP_Engaged_L	AP_Engaged_R	Flight_Director_On_L	AP_Localizer_Engaged	AP_Glide_Slope_Engage	Flight_Director_On_R	AP_Approach_Mode

✔ = out of sequence switch activation detected by seqeunceMiner

SME opinion: Possible evidence of mode confusion.

Switch Activations during a Change to a Crossing Runway

	seconds to touchdown																
	376	376	245	244	244	172	172	172	172	171	171	171	171	169	167	167	167	167
on	▲					▲	▲					▲	▲		▲	▲		
			✔			✔	✔			✔	✔	✔	✔	✔	✔	✔	✔	✔
off		▼	▼	▼	▼			▼	▼	▼	▼			▼			▼	▼
	AP_Heading_Select_Mode	AP_LNAV_Mode	Autothrottle_Engaged	AP_Engaged_L	AP_Engaged_R	AP_Engaged_L	AP_Engaged_R	AP_Heading_Select_Mode	Flight_Director_On_R	AP_Engaged_L	AP_Engaged_R	AP_Heading_Select_Mode	Flight_Director_On_R	Flight_Director_On_R	AP_Engaged_L	AP_Engaged_R	AP_Heading_Select_Mode	Flight_Director_On_L

SME opinion: Possible evidence of mode confusion.

sequenceMiner Conclusions

A flight's anomalous behavior can be characterized by it's missing and extra switches, which sequenceMiner is able to describe.

We applied sequenceMiner to the CIR study and identified instances of mode confusion occurring during a CIR.

An article describing the sequenceMiner algorithms in detail, along with associated experimental results has been accepted for publication.

We have started the process to release sequenceMiner as Open Source software.

Some Partners of the IVHM Project

References

Primary References:

A. N. Srivastava, "Learning Kernels with Mixture Densities," in preparation for IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005.

A. N. Srivastava, "Mixture Density Mercer Kernels: A Method to Learn Kernels Directly from Data, Proceedings of the 2004 SIAM Data Mining Conference, Orlando FL.

A. N. Srivastava and N. Oza, "Knowledge Driven Image Mining with Mixture Density Mercer Kernels," European Space Agency Special Publication #553, Proceedings of the European Image Information Mining Coordination Group, Madrid, Spain 2004.

A. N. Srivastava and B. Zane-Ulman, "Discovering Hidden Anomalies in Text Reports Regarding Complex Space Systems", IEEE Aerospace Conference, Big Sky, MT, 2005.

A. N. Srivastava, "Discovering Anomalies in Sequences with Applications to System Health," Proceedings of the 2005 Joint Army Navy NASA Air Force Interagency Conference on Propulsion, Charleston SC, 2005.

A. N. Srivastava, R. Akella, et. al., "Enabling the Discovery of Recurring Anomalies in Aerospace System Problem Reports using High-Dimensional Clustering Techniques," accepted for publication in the 2006 Proceedings of the IEEE Aerospace Conference.

M. J. Way and A. N. Srivastava, "Novel Methods for Predicting Photometric Redshifts from Broadband Photometry using Virtual Sensors." Astrophysical Journal, 647:102-115, 2006.

S. Budalakoti, A. N. Srivastava, R. Akella, "Discovering Atypical Flights in Sequences of Discrete Flight Parameters," accepted for publication in the 2006 Proceedings of the IEEE Aerospace Conference.

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

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

D. Iverson, "Inductive System Health Monitoring," Published in the Proceedings of The 2004 International Conference on Artificial Intelligence (IC-AI'04), CSREA Press, Las Vegas, NV, June 2004.

References

B. Amidan, and T. Ferryman, "Atypical Event and Typical Pattern Detection within Complex Systems," IEEE Aerospace Conference, 2005.

L. Atlas and G. Bloor, *An evolvable tri-reasoner ivhm system,* ISIS Vanderbilt Website (1999).

A. Banerjee, I. Dhillon, J. Ghosh, and S. Sra, *Generative model-based clustering of directional data,* 2003.

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

I.T. Joliffe, *Principle component analysis,* Springer, 2002.

Eamonn J. Keogh, Selina Chu, David Hart, and Michael J. Pazzani, *An online algorithm for segmenting time series,* ICDM, 2001, pp. 289-296.

T. Lane, "Machine Learning Techniques for the computer security domain of anomaly detection" , Ph.D. Thesis, CERIAS TR 2000-12, Purdue University, August 2000.

M. Last, Y. Klein, and A. Kandel, *Knowledge discovery in time series databases,* 2001.

R.T. Ng. and Jiawei Han, "CLARANS: a method for clustering objects for spatial data mining", IEEE Transactions on Knowledge and Data Engineering, Volume 14, Issue 5 (Sept/Oct 2002), Pages: 1003-1016.

L. R. Rabiner, *A Tutorial on hidden markov models and selected applications in speech recognition,* Proceedings of the IEEE 77 (1989), no. 2, 257-286.

K. R. Pattipati J. Ying, T. Kirubarajan and A. Patterson-Hine, *A hidden markov model-based algorithm for online fault diagnostic with partial and imperfect tests,* IEEE Transactions on SMC: Part C 30 (2000), no. 4, 463-473.

D.B. Skillicorn, *Clusters within clusters: Svd and counterterrorism,* SIAM Workshop on Counterterrorism (2003).

References

L. Connel, "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.