An Overview of Data Mining at NASA

Ashok N. Srivastava, Ph.D.
ashok@email.arc.nasa.gov

Leader: Intelligent Data Understanding Group
NASA Ames Research Center

The Data Mining Team

Group Leader

Ashok N. Srivastava, Ph.D.

Team Members and Collaborators (in color)

Michael Berry, Ph.D.	Dawn McIntosh
Suratna Budalakoti	Rama Nemani, Ph.D.
Pat Castle	Matthew Otey, Ph.D.
Captain Alan Cerino	Nikunj Oza, Ph.D.
Aditi Chattopadhyay, Ph.D.	Loren Rosenthal
Santanu Das	Mark Schwabacher, Ph.D.
Tom Ferryman, Ph.D.	Irv Statler, Ph.D.
Paul Gazis, Ph.D.	Julienne Stroeve, Ph.D.
Dave Iverson	Eugene Turkov
Amy Mai	Michael Way, Ph.D.
Rodney Martin, Ph.D.	Richard Watson
Bryan Matthews	David Wolpert, Ph.D.

Team Members are NASA Employees, Contractors, and Students.

Key Programs

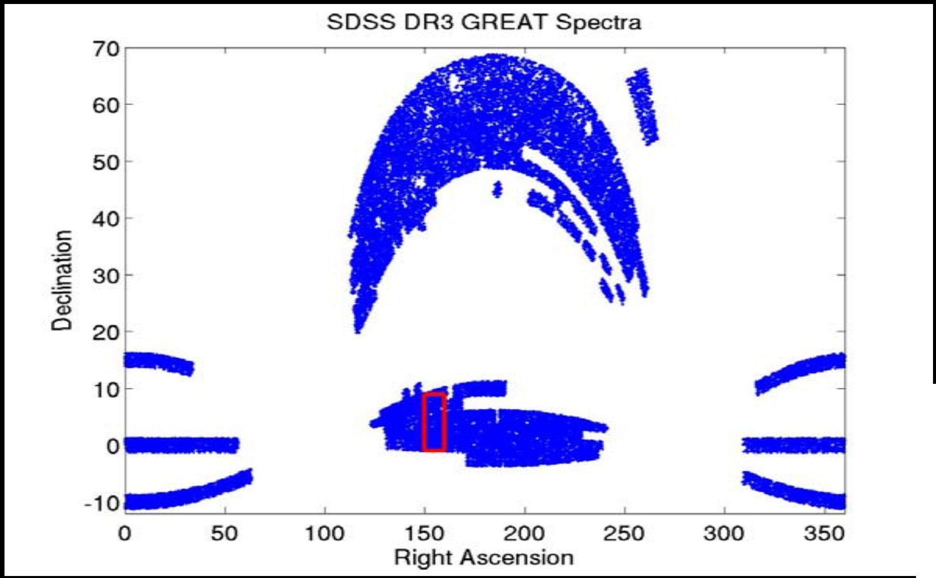
- Aeronautical Research Mission Directorate: Aviation Safety Program
- NASA Engineering and Safety Center
- Exploration Systems Mission Directorate - Exploration Technology Development Program, ISHM Project
- Shuttle Program - Wing Leading Edge Impact Detection
- Science Mission Directorate - AISRP

All schematic diagrams and pictures in this presentation are publicly available on the Internet.

NASA Data Systems

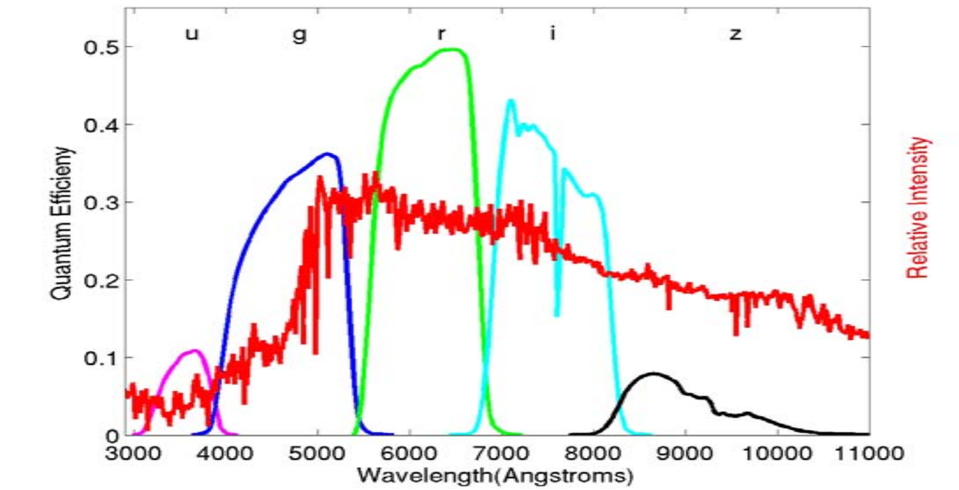
- Earth and Space Science
  - Earth Observing System generates ~21 TB of data per week.
  - Ames simulations generating 1-5 TB per day
- Aeronautical Systems
  - Distributed archive growing at 100K flights per month with 1M flights already.
- Exploration Systems
  - Space Shuttle and International Space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

Characterizing the Large Scale Structure of the Universe



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

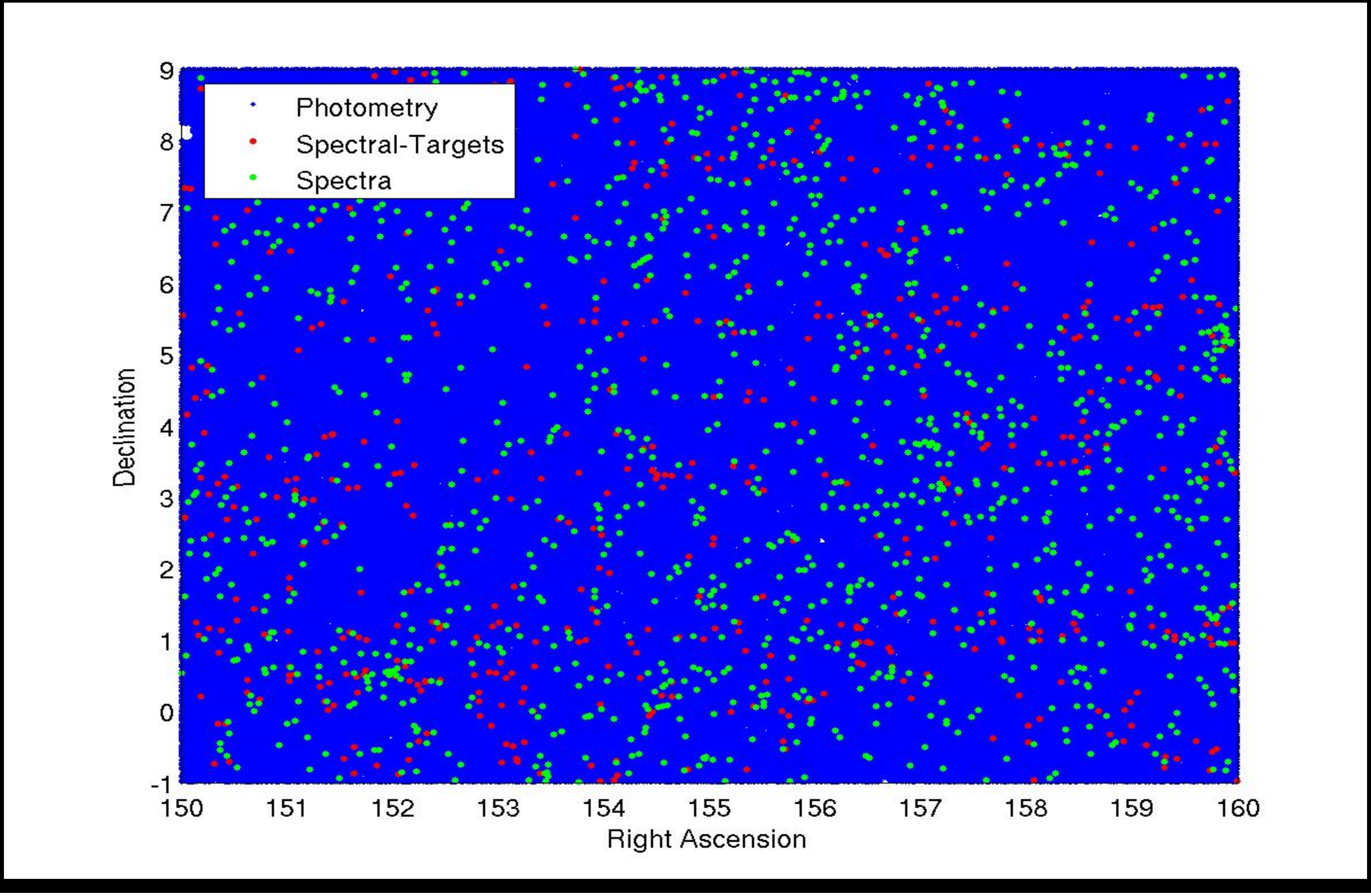
Obtaining a good estimate of their 3-D position in the sky would help determine the filamentary structure of the universe to constrain cosmological models.



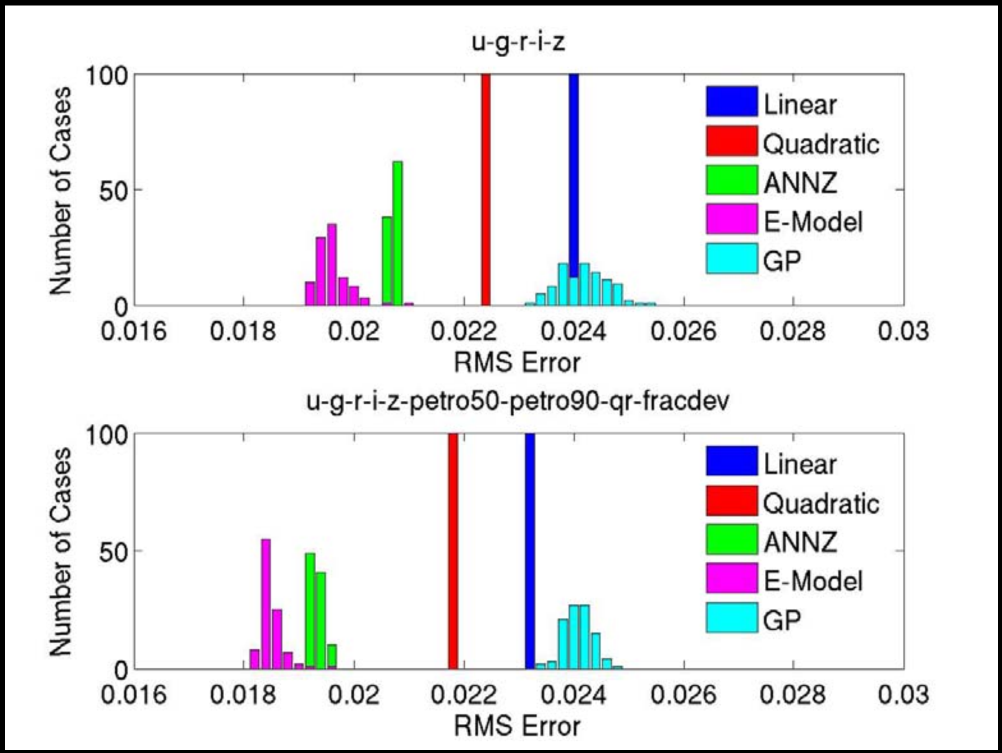
We are building machine learning methods to estimate the redshift of galaxies using broad-band photometry.

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

Each dot (including blue) is a galaxy



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.

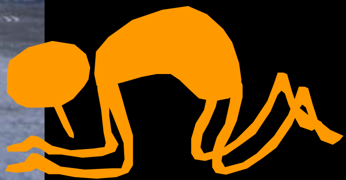
Outline of Talk

Categorizing and detecting anomalies described in safety documents

Detecting anomalies in cockpit switching sequences

Detecting Shuttle wing heating anomalies

The Forensic (Historic) Approach to Accident Prevention



vs….

... a More Prognostic Approach

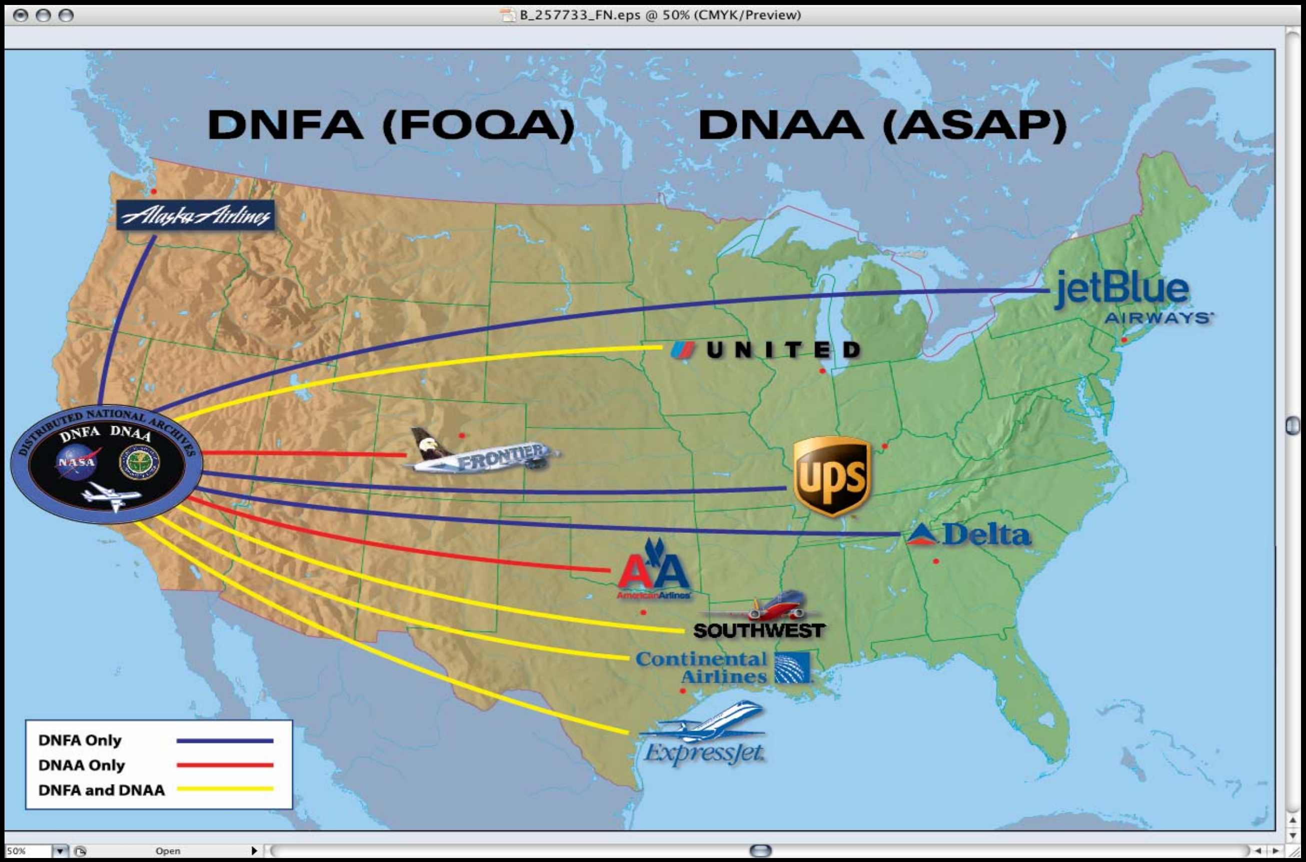
Proactive risk management leads to decisions before an accident occurs

... a More Prognostic Approach

- Identify
  - Monitor and compare with expectations.
  - Uncover potential hazards
- Formulate
  - Consider change
  - Cost-benefit estimate
  - Assess safety risk
- Evaluate
  - Diagnose causation
  - Quantify frequency
  - Assess severity
- Implement
  - Implement locally
  - Evaluate intervention
  - Refine
  - Implement full scale

Proactive risk management leads to decisions before an accident occurs

Distributed National Archives



ASRS Report Excerpt

JUST PRIOR TO TOUCHDOWN, LAX TWR TOLD US TO GO AROUND BECAUSE OF THE ACFT IN FRONT OF US. BOTH THE COPLT AND I, HOWEVER, UNDERSTOOD TWR TO SAY, 'CLRED TO LAND, ACFT ON THE RWY.' SINCE THE ACFT IN FRONT OF US WAS CLR OF THE RWY AND WE BOTH MISUNDERSTOOD TWR'S RADIO CALL AND CONSIDERED IT AN ADVISORY, WE LANDED...

Automatic Categorization of ASRS Reports

ASRS Report Extract

JUST PRIOR TO TOUCHDOWN, LAX TWR TOLD US TO GO AROUND BECAUSE OF THE ACFT IN FRONT OF US. BOTH THE COPLT AND I, HOWEVER, UNDERSTOOD TWR TO SAY, 'CLRED TO LAND, ACFT ON THE RWY.' SINCE THE ACFT IN FRONT OF US WAS CLR OF THE RWY AND WE BOTH MISUNDERSTOOD TWR'S RADIO CALL AND CONSIDERED IT AN ADVISORY, WE LANDED...

?

?

Sample of 60 ASRS Anomaly Categories

Non Adherence to ATC Clearance
Critical Equipment Problem
Runway Incursion
Landing without a Clearance
Air Space Violation
Altitude Deviation Overshoot
Fumes
Altitude Deviation Undershoot
Ground Encounter, Less Severe
...

Automatic Categorization of ASRS Reports

ASRS Report Excerpt

JUST PRIOR TO TOUCHDOWN, LAX TWR TOLD US TO GO AROUND BECAUSE OF THE ACFT IN FRONT OF US. BOTH THE COPLT AND I, HOWEVER, UNDERSTOOD TWR TO SAY, 'CLRED TO LAND, ACFT ON THE RWY.' SINCE THE ACFT IN FRONT OF US WAS CLR OF THE RWY AND WE BOTH MISUNDERSTOOD TWR'S RADIO CALL AND CONSIDERED IT AN ADVISORY, WE LANDED...

Sample of 60 ASRS Anomaly Categories

Non Adherence to ATC Clearance
Critical Equipment Problem
Runway Incursion
Landing without a Clearance
Air Space Violation
Altitude Deviation Overshoot
Fumes
Altitude Deviation Undershoot
Ground Encounter, Less Severe
...

Classification Task

- Automatically map safety reports into Distributed National ASAP Archive (DNAA) anomaly categories.
- New reports entering the DNAA can then be automatically categorized by the classifier.
- Comparison among Natural Language Processing (NLP), statistical methods, and Mariana, which is based on advanced data mining techniques.

Data Mining Approach

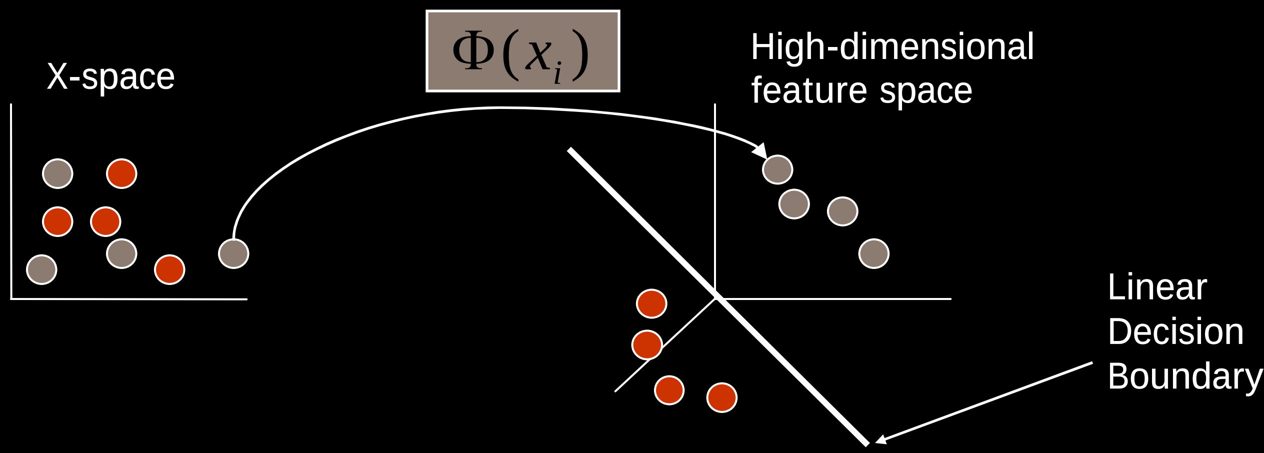
- Convert documents into a vector space representation “Bag of Words” matrix.
- Learn the mappings from documents to categories.
- Typical matrix:
  - 30,000 rows
  - 40,000 dimensions

Frequency of term in document

	Term 1	Term 2	Term 3	…
Document 1	0	1	0	4
Document 2	0	3	0	0
…	2	8	1	0

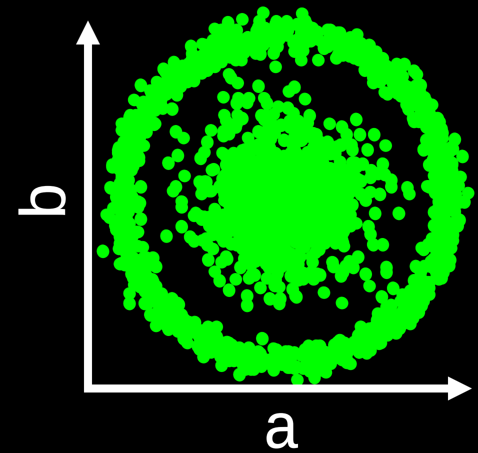
The Support Vector Machine

- Given a set of p-dimensional data $\{x_i\}_{i=1}^{N}$
- Use a possibly infinite dimensional operator $\Phi(x_i)$ to map the data into a feature space.
- Perform linear operations in the feature space.
- Map result back to the original space.
- Can do this operation without explicitly computing $\Phi(x_i)$

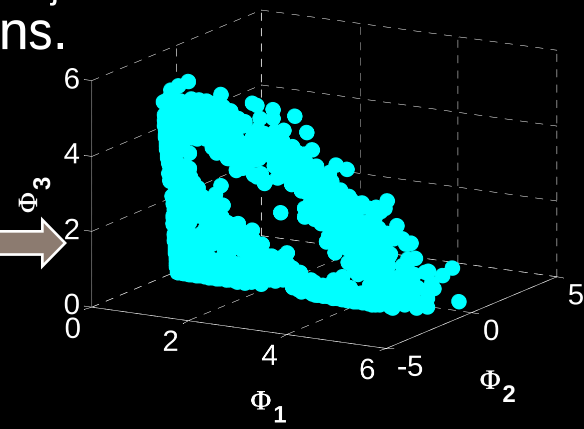


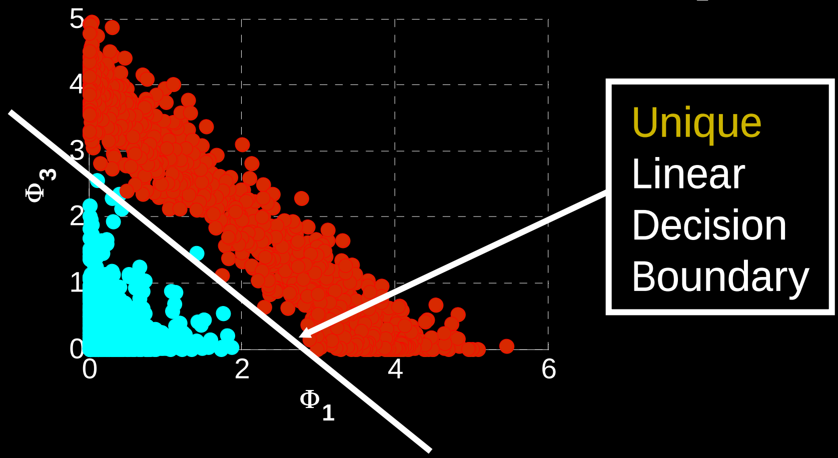
An Example Mapping

Using a kernel function $K(x_i,x_j) = \langle x_i,x_j \rangle^2$, two-dimensional data gets mapped to three dimensions.



$$\Phi(x_i) = \begin{bmatrix} a^2 \\ \sqrt{2}ab \\ b^2 \end{bmatrix}$$





Text Mining with SVMs

- We built 23 instances of a Support Vector Machine, each tuned to classify ASAP documents into DNAA anomaly categories with advanced noise reduction methods.
- We developed Mariana, an advanced Markov Chain Monte Carlo (MCMC) algorithm to find the best SVM hyperparameters.
- Kernel induces an infinite dimensional feature space.

$$K(x_i, x_j) = \Phi^T(x_i)\Phi(x_j) = \exp\left[-\frac{\| x_i - x_j \|^2}{2\sigma^2}\right]$$

Hyperparameters in SVMs

$\min\{ K(x,x_i) + Cw\Sigma_i\varepsilon_i\}$, where ε_i = soft margin, and
where C = error penalty parameter

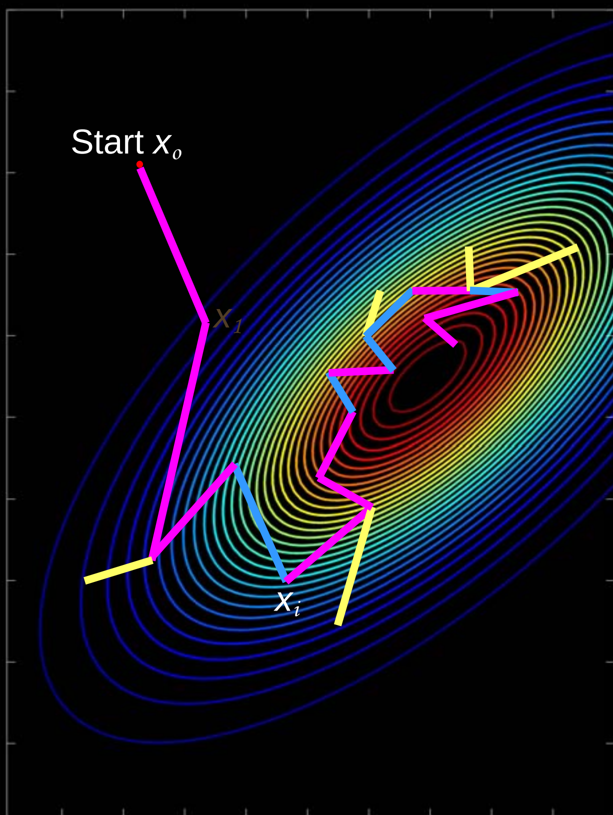
$K(x, x_i) = \exp\{-\gamma||x-x_i||^2\}$, where γ = scale parameter

$$\text{w, class penalty parameter} = \begin{cases} w_1 & y_i = 1 \text{ (in the class)} \\ 1 & y_i = -1 \text{ (out of the class)}, \end{cases}$$

$\boxed{\text{w, C, } \gamma \text{ are model inputs}}$

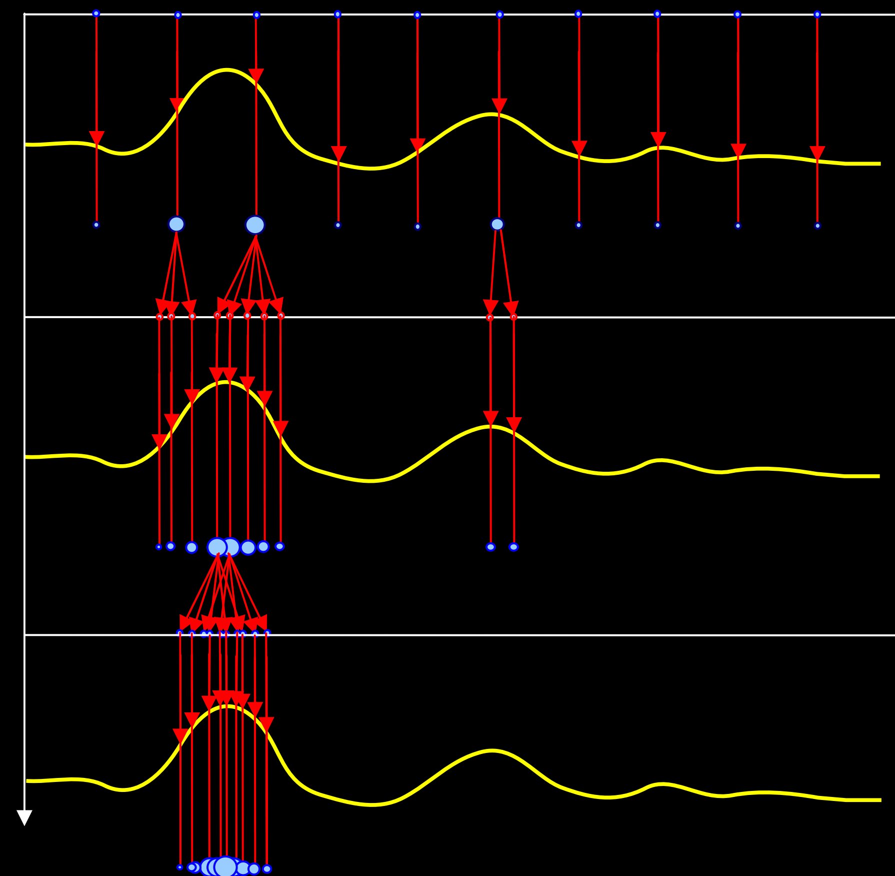
Mariana Statistical Optimization Methods

Current Approach: Simulated Annealing



Accepted step
Accepted, but at lower value
Rejected step

Possible Future Approach: Particle Filter



Less likely to be optimal (maximum) solution
More likely to be optimal, i.e., global max, solution

Natural Language Processing

- NLP extracts and represents concepts in text documents.
- Potentially thousands of hand-crafted rules to extract meaning.
- Example: Identify reports describing "pilot fatigue"
  - Search for: 'fatigue', 'tired', 'last leg of an X day trip', 'sleepy', ...
  - If a document has any of these phrases, tag it as a 'fatigue' document.

Comparing NLP to Data Mining

NLP

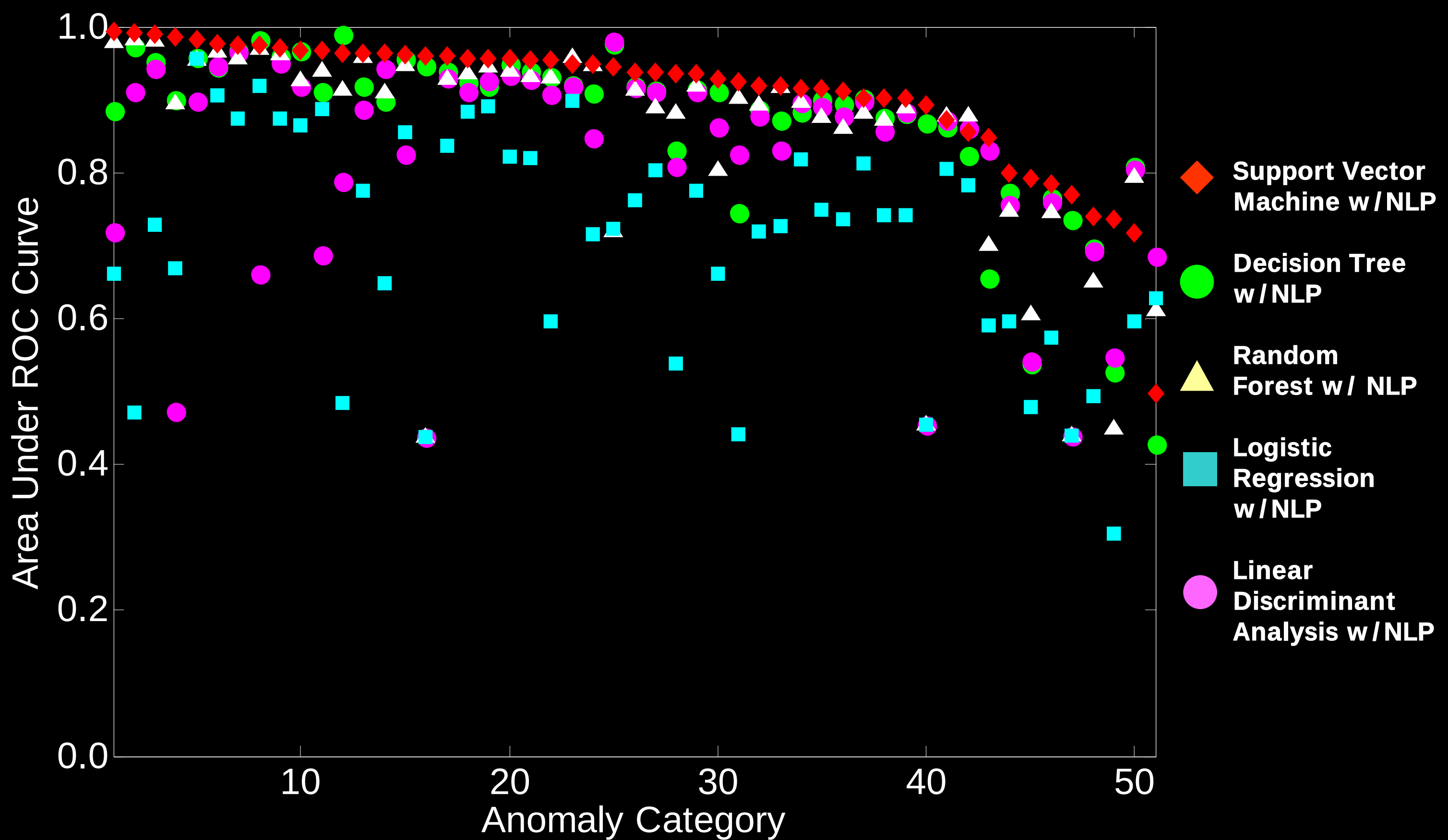
- Very precise representation of concepts.
- Large hand-crafted rule bases.
- Very expensive due manual rule building.

Data Mining

- Very imprecise representation of concepts.
- Word frequencies.
- Inexpensive in terms of manual work.

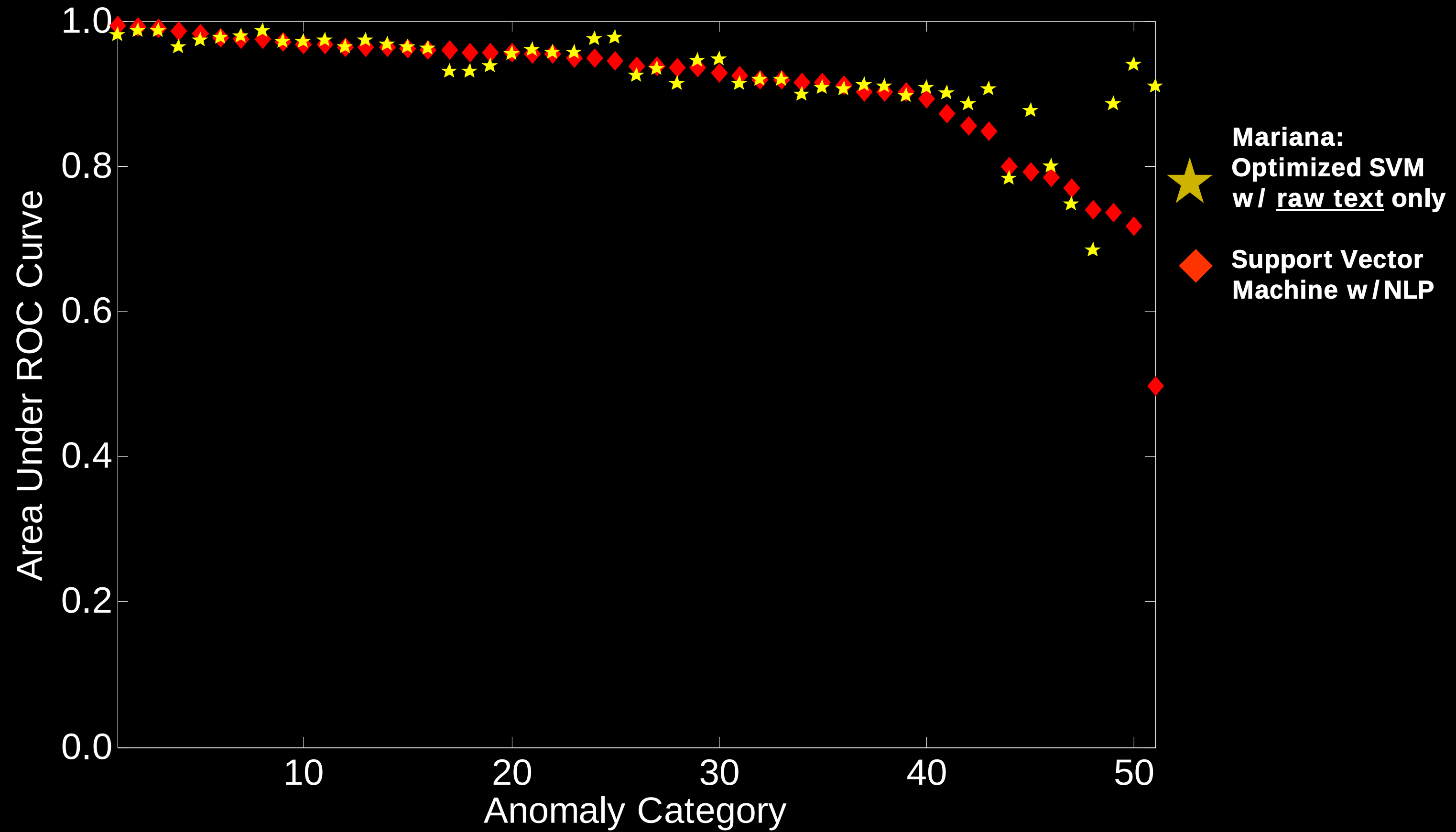
The output of NLP systems can be fed into data mining algorithms to improve accuracy.

Comparing un-optimized SVM and Standard Methods using NLP inputs

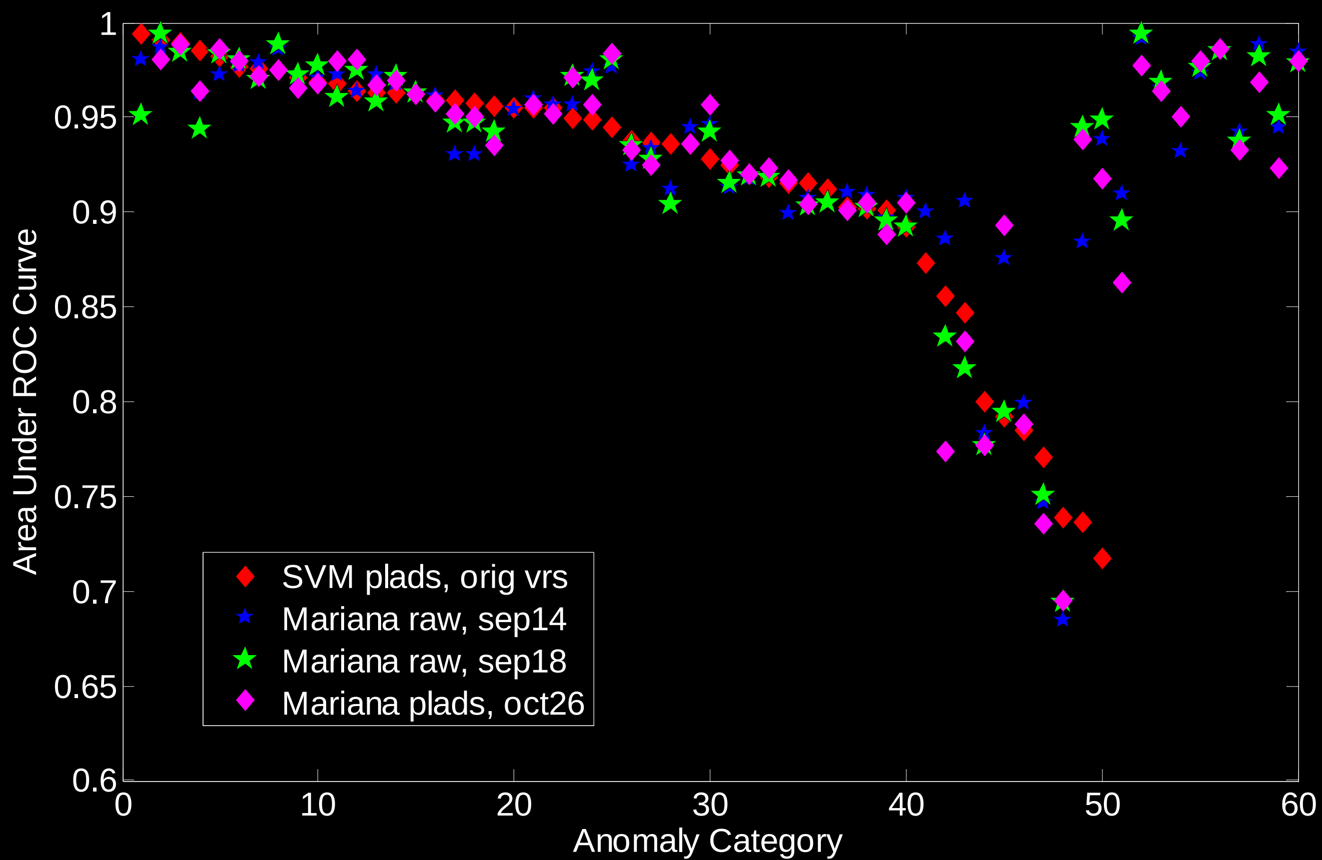


SVM beats other methods 43 out of 51 times.

Comparison of Mariana with Raw Text and SVM with Raw Text + NLP

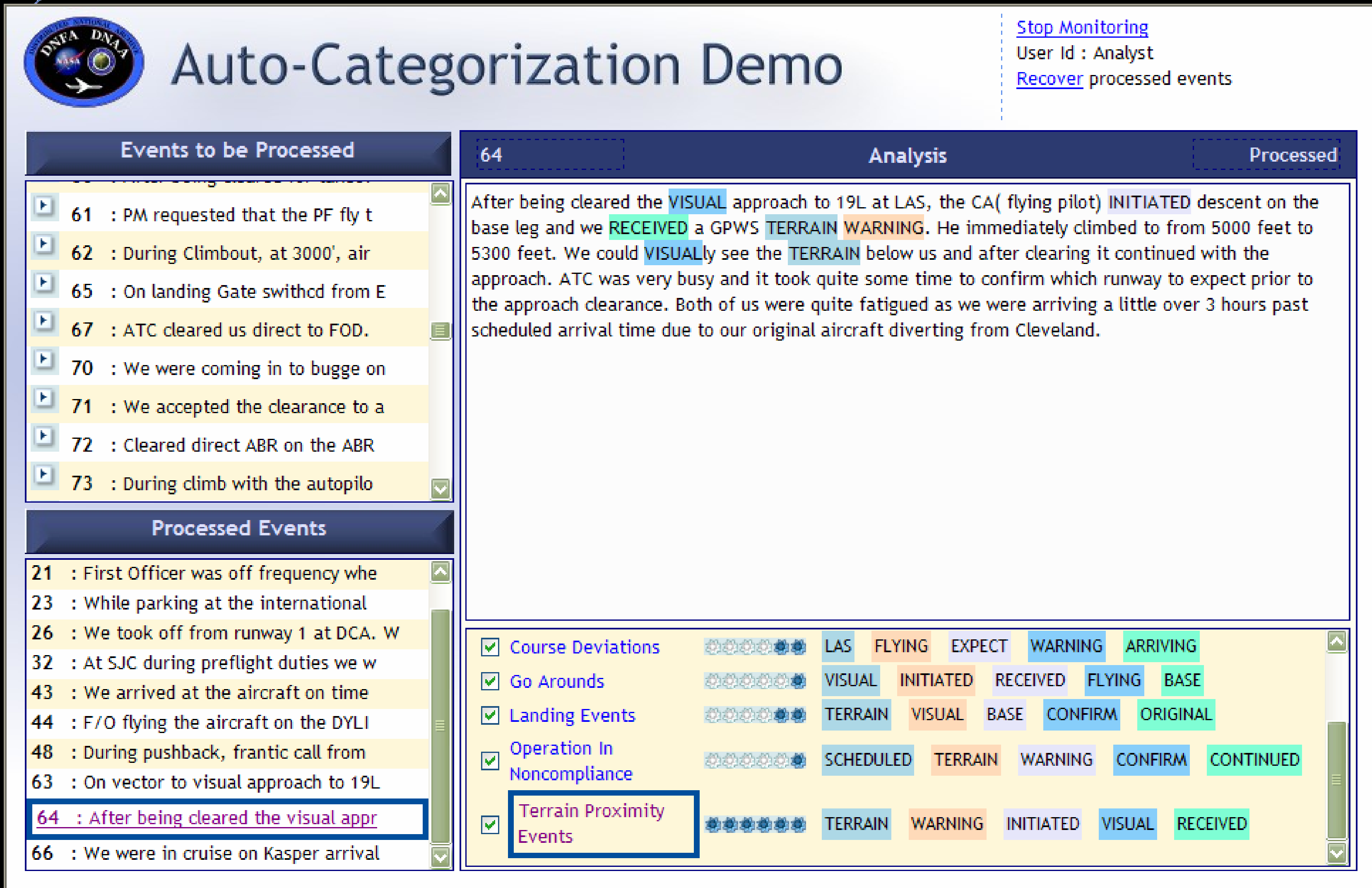


Comparison of NLP and Mariana



Mariana's performance with raw text matches the performance with text and NLP.

Mariana to be deployed at Air Carriers



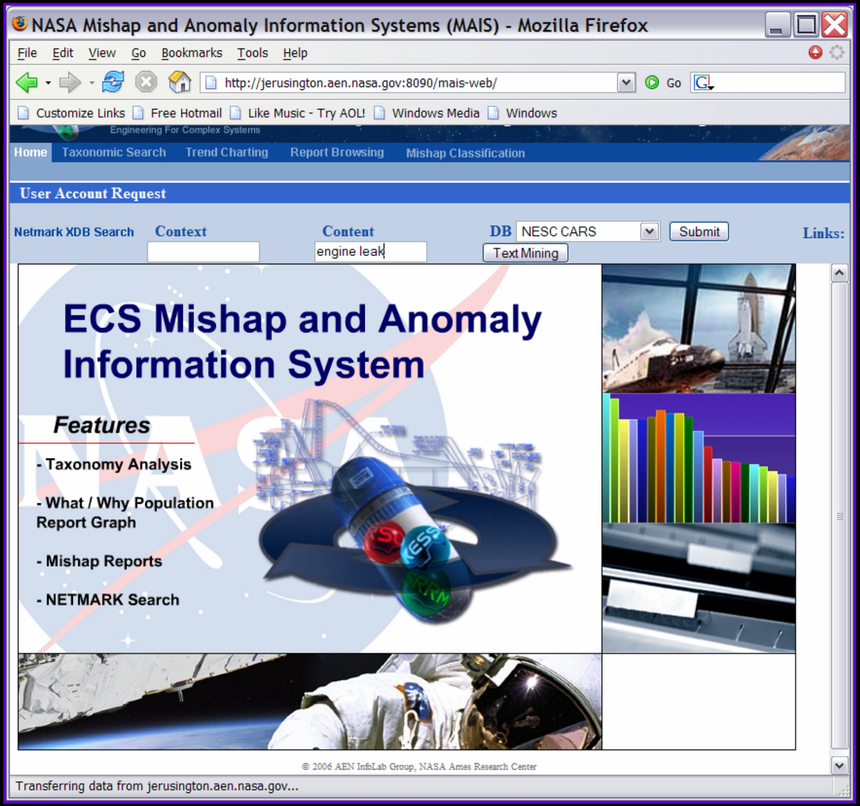
Our Innovations

- Mariana searches for the best SVM hyperparameters using Markov Chain Monte Carlo techniques.
- Mariana performs as well as or better than the SVM built using NLP techniques without the overhead.
- Our methods for term selection and noise reduction reduce false positive rates by as much as 30%.

Searching for Recurring Anomalies

Enabling discovery of anomalous trends in complex aerospace systems

Research sponsored by: NASA Engineering and Safety Center

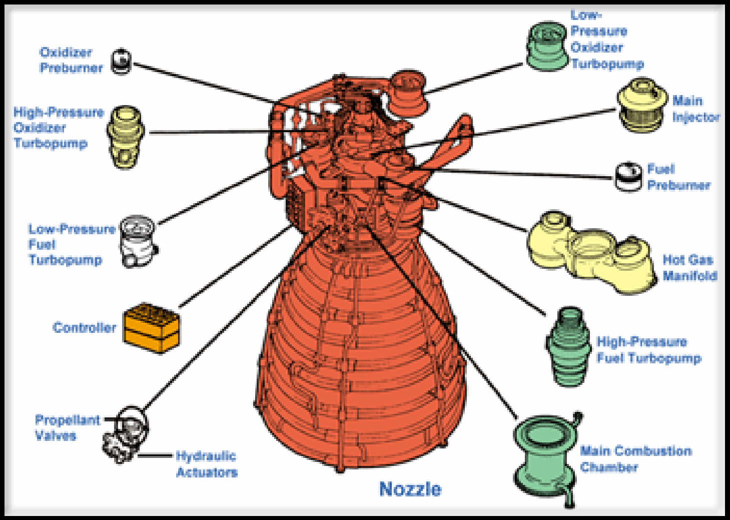


Searching for Recurring Anomalies

- These reports *do not* have an anomaly category associated with them.
- Potentially several hundred thousand reports.
- Some systems have been around for decades.
- Enables analysis of trends of anomalies (trending).
- Can't be addressed using standard clustering techniques.
- Our systems use content-based similarity as well as statistical similarity.

NESC Definition of Recurring Anomalies

- Recurrent failures described in text reports.
- Problems that cross traditional system boundaries.
- Problems that have been accepted by repeated waivers.
- Discrepant conditions repeatedly accepted by routine analysis.
- Events with unknown causes.

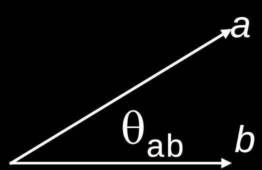


Detecting Recurring Anomalies

1. *Calculate cosine similarity between all document vectors.*

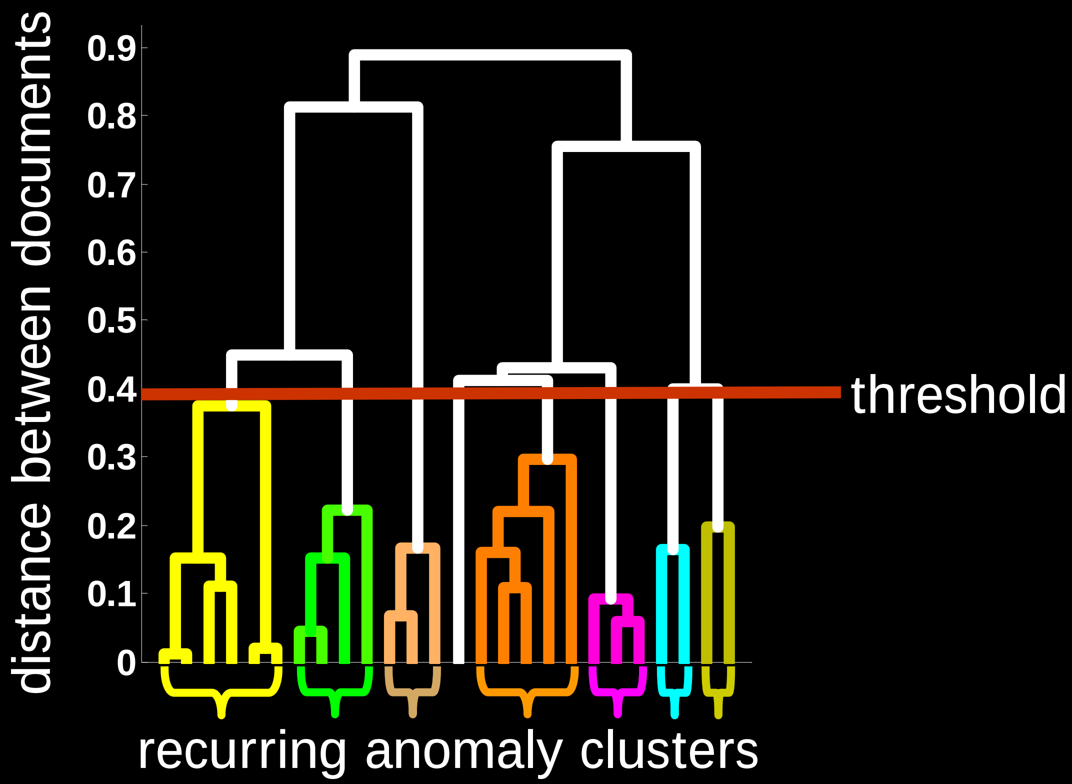
	Term 1	Term 2	Term 3	Term 4	...
Doc a	3	2	1	5	...
Doc b	0	1	4	1	...
...	...	...	...	...	...

$$\cos\theta_{ab} = \frac{<a,b>}{\|a\| \cdot \|b\|}$$



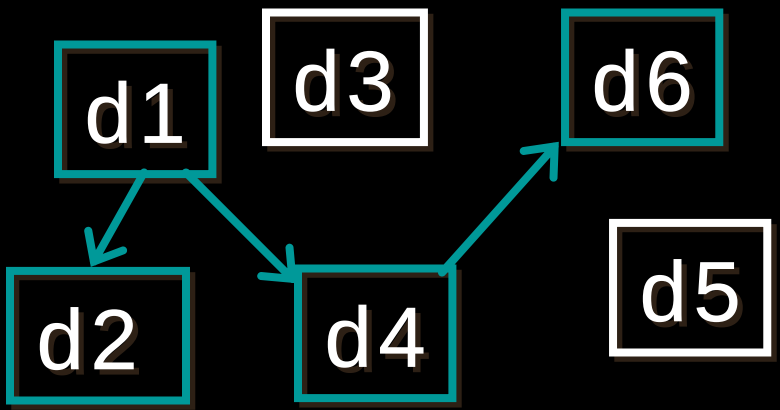
Detecting Recurring Anomalies

2. Apply agglomerative clustering.



Detecting Recurring Anomalies

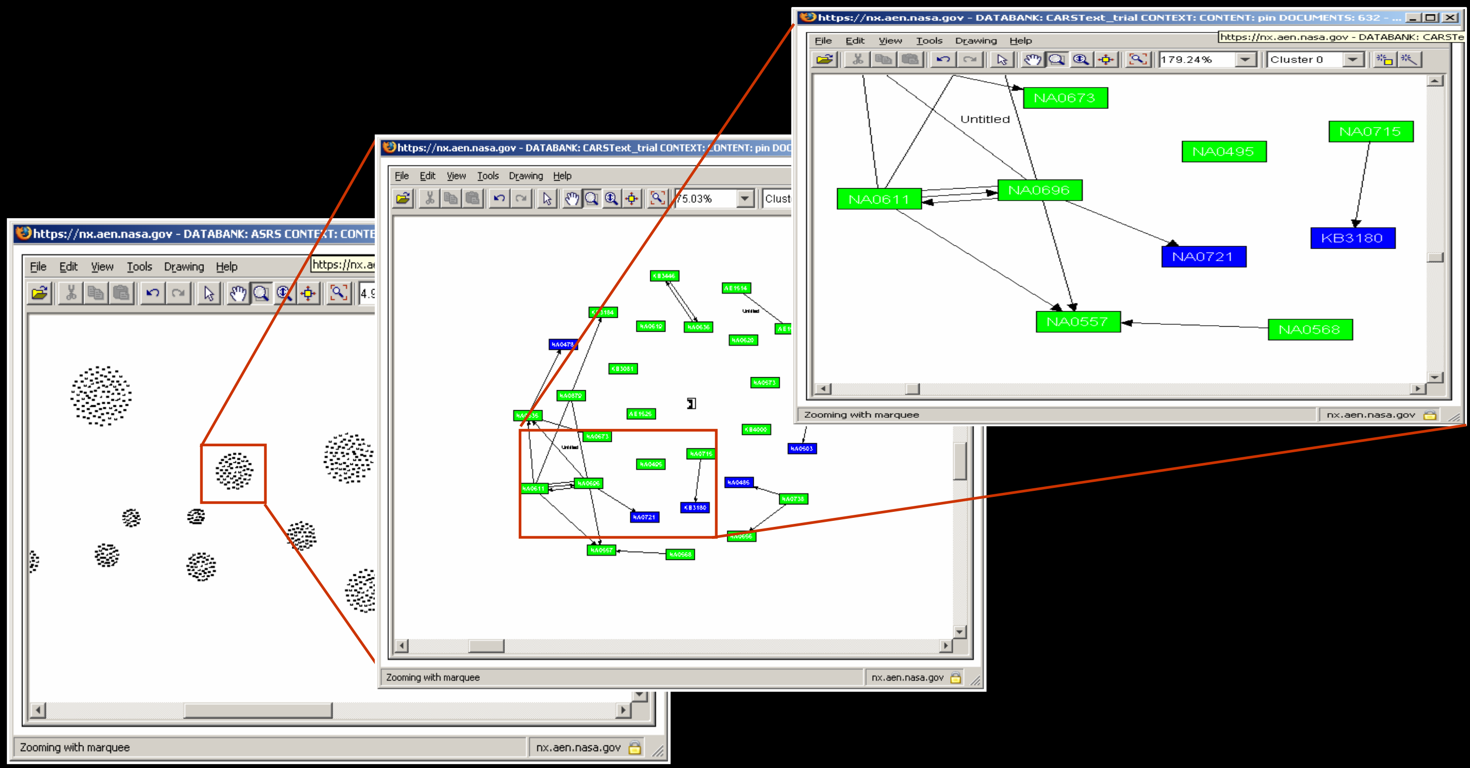
3. Identify referenced documents.



If d1 refers to d2 and d4, and d4 refers to d6, then d1, d2, d4, & d6 are considered a <u>recurring anomaly</u>.

Detecting Recurring Anomalies

4. Identify & visualize possible recurring anomalies.



Testing the Recurring Anomaly Detection System (ReADS)

- Experts reviewed a subset of the Shuttle Orbiter Corrective Action Records (CARs) to identify recurring anomalies.
- We extracted 333 reports to test the performance of our system called REcurring Anomaly Detection System (ReADS).
- Of those 333 reports, the experts identified 20 recurring anomalies and ReADS identified 39 recurring anomalies.

Performance of ReADs

On a subset (333) of the Shuttle Orbiter Records:

58% of the records were eliminated as non-recurring anomalies (RAs) by ReADS.

12 exact matches between RAs discovered by experts and RAs discovered by ReADS.

6 previously unidentified RAs discovered by ReADS which were confirmed by experts.

1 record was identified by experts as being part of an RA and was missed by ReADS.

5% of the expert RAs were separated by ReADS into more than one RA.

8% of the ReADS RAs combined two expert RAs into a single RA.

Our Innovations

- Enable analysis of anomaly trends using a combination of content and statistical search methods.
- ReADS is a novel tool designed especially for identifying recurring anomalies across multiple databases.
- Development of robust platform to analyze and visualize recurring anomalies.

Detecting Anomalies in Cockpit Switch Sequences

Enabling discovery of anomalous switching events

Research sponsored by: NASA ARMD

Background

- sequenceMiner analyzes large repositories of discrete sequences and identifies operationally significant anomalies.
- Learns the typically observed switching patterns directly from discrete data streams.
- This method outperforms others in terms of speed, comprehensibility, and stability, and does not require knowledge of Standard Operating Procedures.

Example Sequence Anomaly Detection Problem

Typically Observed Switching Patterns

A B C D A D D A G F Q ...

Example Observed Switching Sequence

A B G F Q C D A D D A ...

Problems: (1) Discover Typically Observed Switching patterns given thousands of flights.

(2) Discover outlying sequences.

Outline of Approach

- sequenceMiner discovers typically observed switching patterns using Multiple Sequence Alignment.
  - Normalized Longest Common Subsequence as a similarity measure
  - Optimized for speed. Analyzes 7400 flights in 6 minutes.
- sequenceMiner discovers:
  - Switches absent in an expected sequence position.
  - Switches inserted in an unexpected sequence position.
  - Switches that are out of order from what is expected.
- sequenceMiner describes why flights are called anomalous and provides a degree of anomalousness.

Multiple Sequence Alignment (MSA)

- Used in bioinformatics to compare DNA sequences of organisms descended from a common ancestor.
- Can identify mutation inside a sequence by comparing it to other sequences.
- In the context of flights, these mutations are the points where a flight deviated from the norm.

Incorporating Operational Information

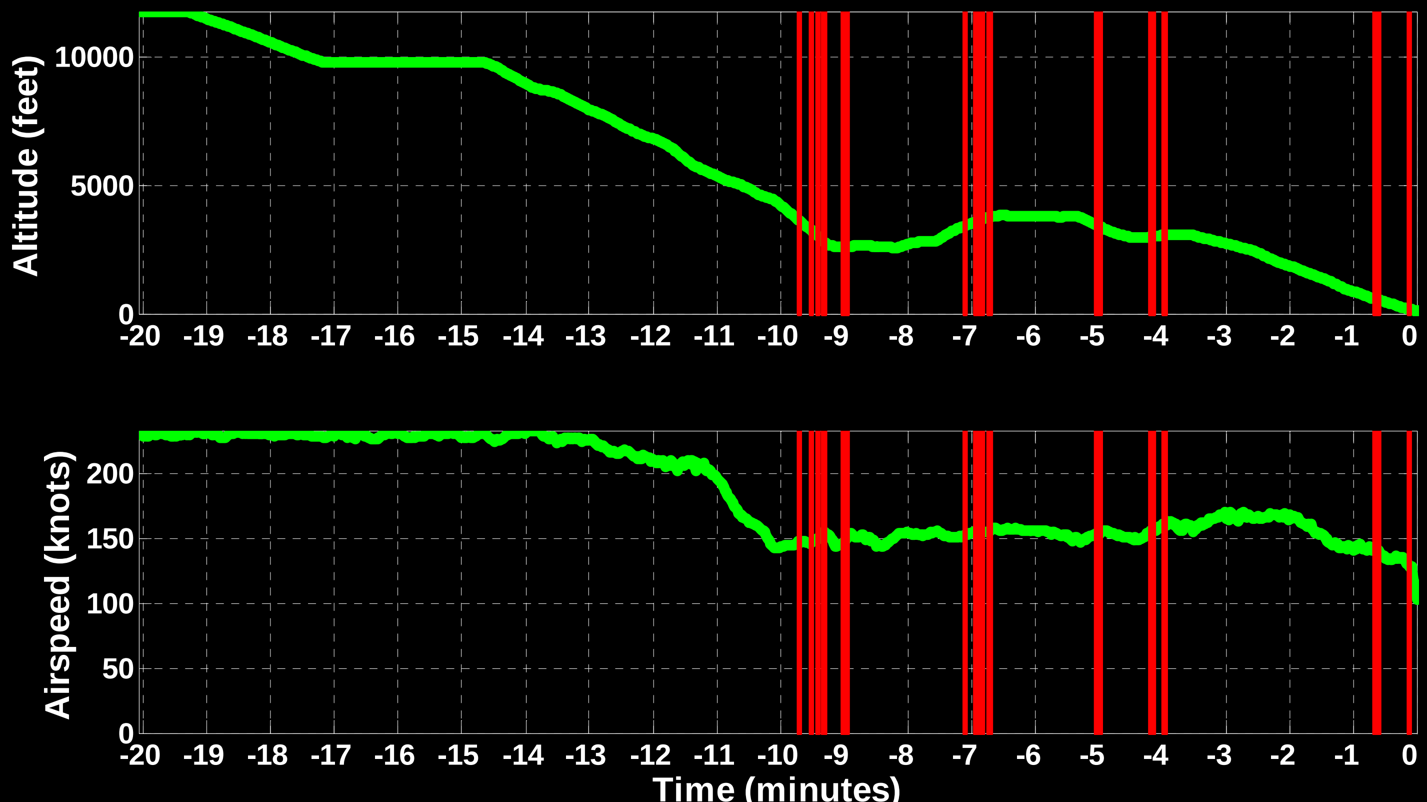
- Weighting of Switches
  - Measures its importance to flight.
  - Used during clustering and anomaly detection.
  - Sequences are identified that have more highly weighted switches out of sequence, instead of simply the number of switches out of sequence.

- Ignore order of switches within a one-minute time interval.
  - This step reduces alarms by around 30%.

Data and Methodology

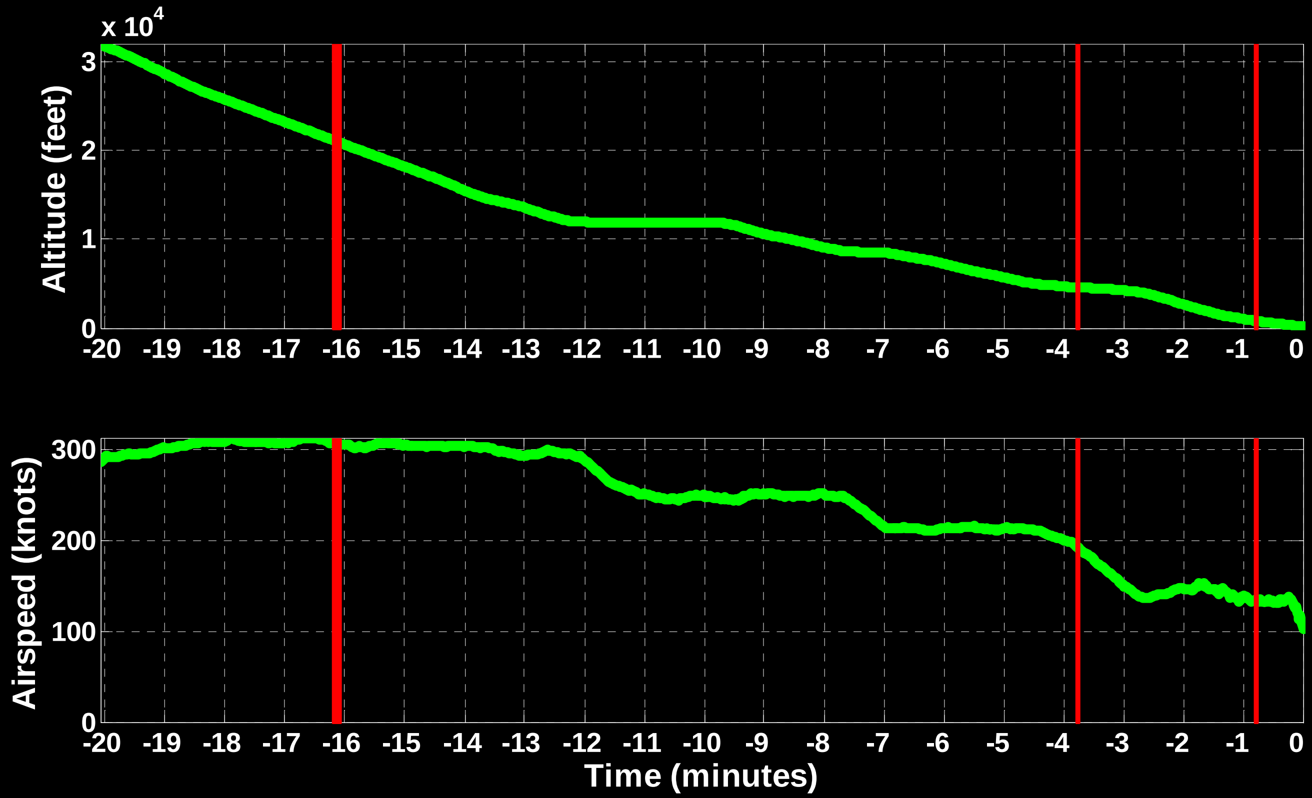
- Initial Dataset
  - 7400 flights from a single fleet and airline.
  - Recordings of 1038 primary and secondary binary switches.
  - 111 primary switches were selected from a subset of 2225 flights.
  - Landing phase to a specific destination airport.
- The 13 most anomalous flights identified by sequenceMiner were analyzed by a 747 pilot who was our expert.
  - 5 were judged to be bad data.
  - 3 were judged to be normal.
  - 5 were judged to be operationally significant anomalies.

sequenceMiner Discovered Anomalous Presses of the Igniter Switch (Red Bars)



Expert: “Pilot switched igniter on and off at atypical times. Possible engine malfunction.”

sequenceMiner Discovered Anomalous Engagement of the Autopilot (Red Bars)



Expert: "Auto-pilot used too many times. Possible case of mode confusion."

sequenceMiner Discovered Anomalous Usage of Speed Brakes (Red Bars)

Expert: “Overuse of speed brakes. Possibly a high energy approach.”

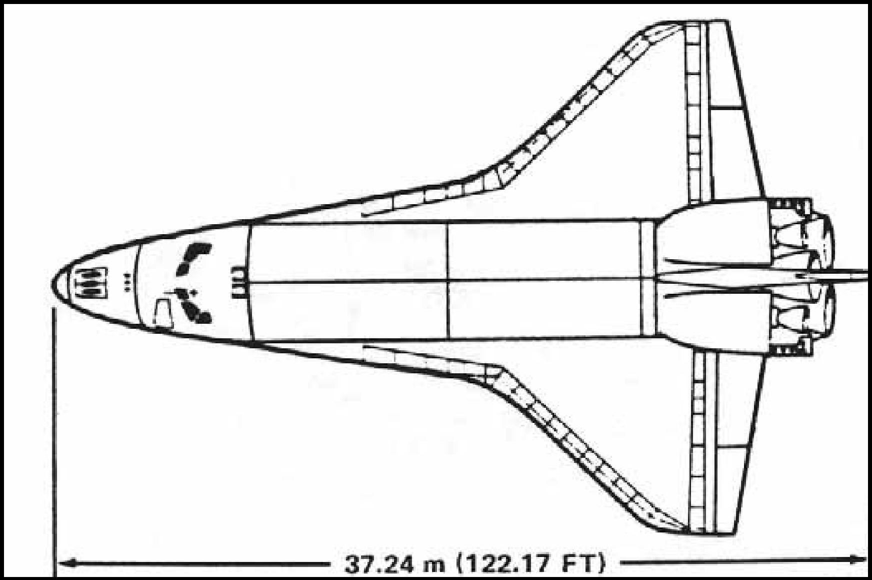
Our Innovations

- sequenceMiner is a fast and reliable system to learn typically observed switching patterns from large volumes of discrete data.
- This system outperforms other algorithms in terms of speed and reliability.
- Discovers operationally significant events such as mode-confusion and high-energy approaches.

Detecting Anomalies in Shuttle Systems

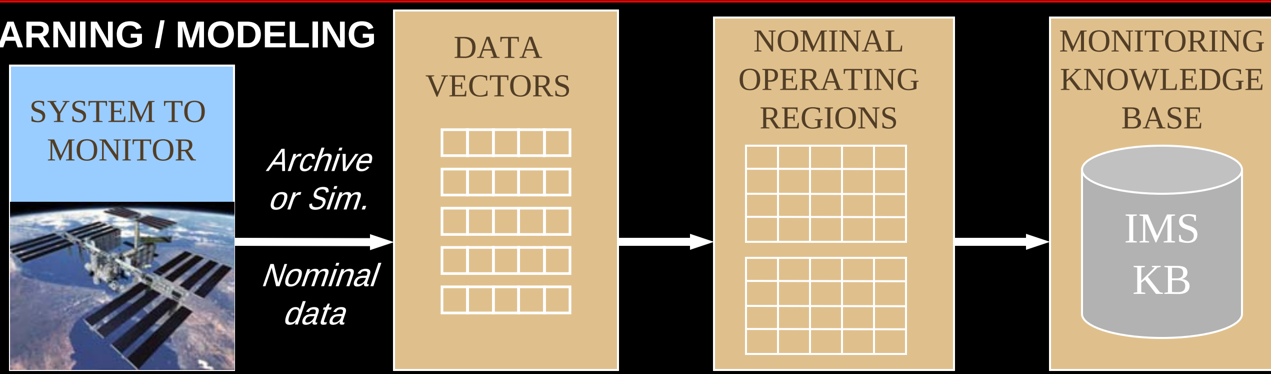
Enabling discovery of anomalies in continuous data streams

Research sponsored by:
NASA ESMD ETDP - ISHM Program



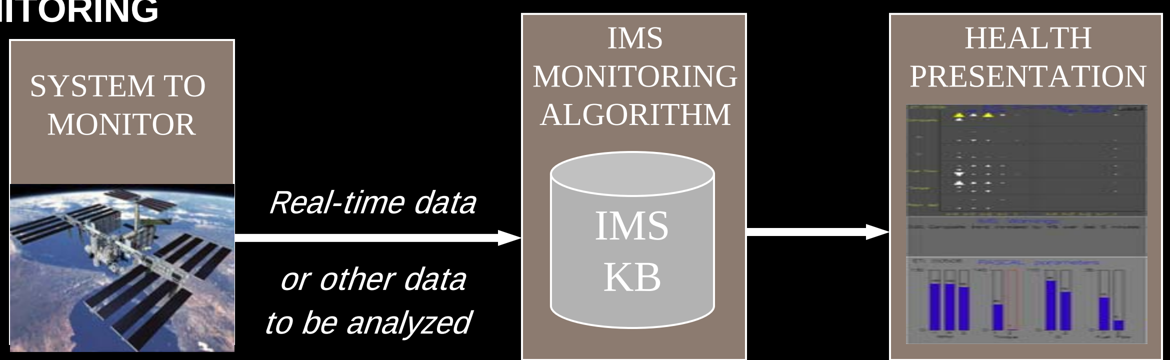
Inductive Monitoring System

LEARNING / MODELING



IMS learns nominal system behavior from archived or simulated system data, automatically builds a “model” of nominal operations, and stores it in a knowledge base.

MONITORING



IMS real-time monitor & display informs users of degree of deviation from nominal performance. Trend analysis can detect conditions that may indicate incipient failure or required system maintenance.

STS-107 Launch Analysis

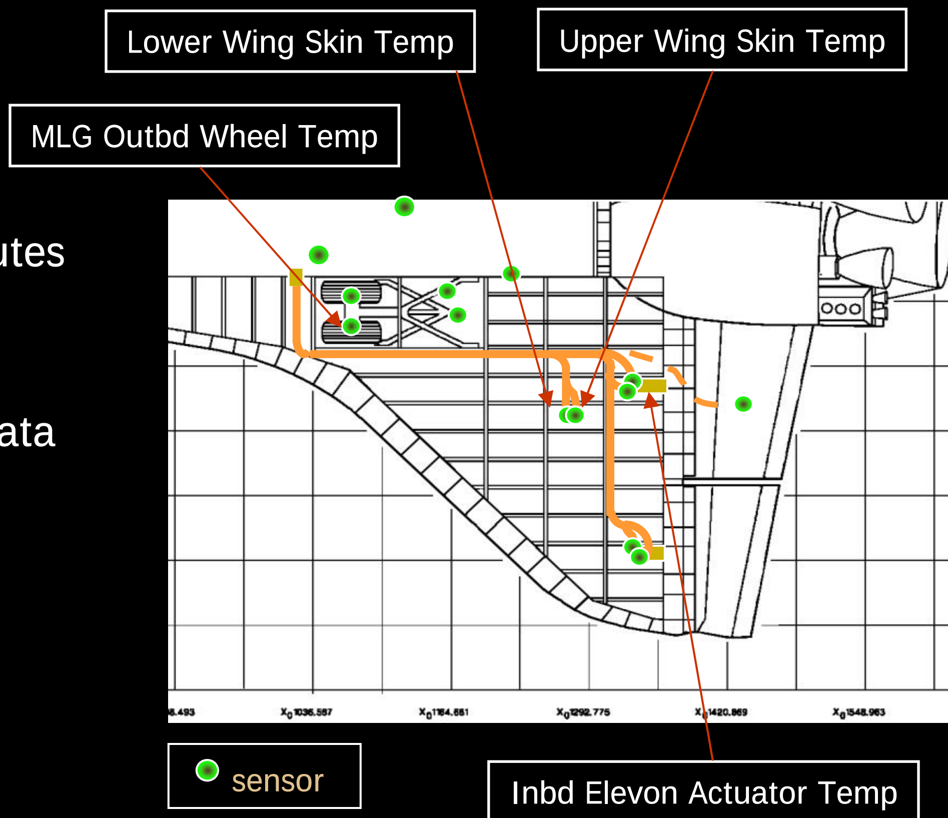
- The IMS method can help identify subtle but meaningful changes in system behavior.
- A comparison of STS-107 ascent telemetry data to data from previous Columbia flights indicates that there may have been enough information to detect a wing-heating anomaly.

STS-107 Ascent - IMS Analysis

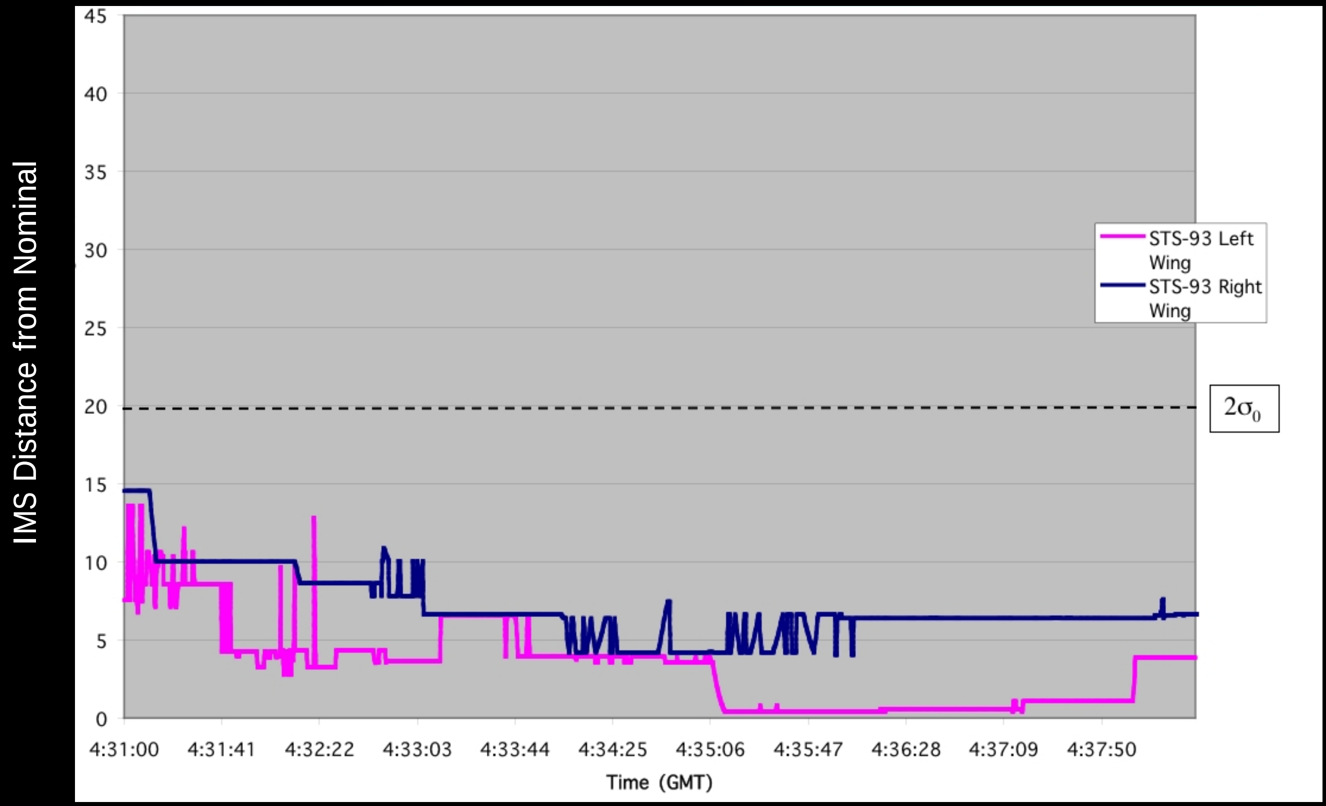
- Data vectors formed from 4 temperature sensors inside the wing
- Data covered first 8 minutes of each flight (Launch to Main Engine Cut Off)
- Trained on telemetered data from 10 previous Columbia flights

Normalization:

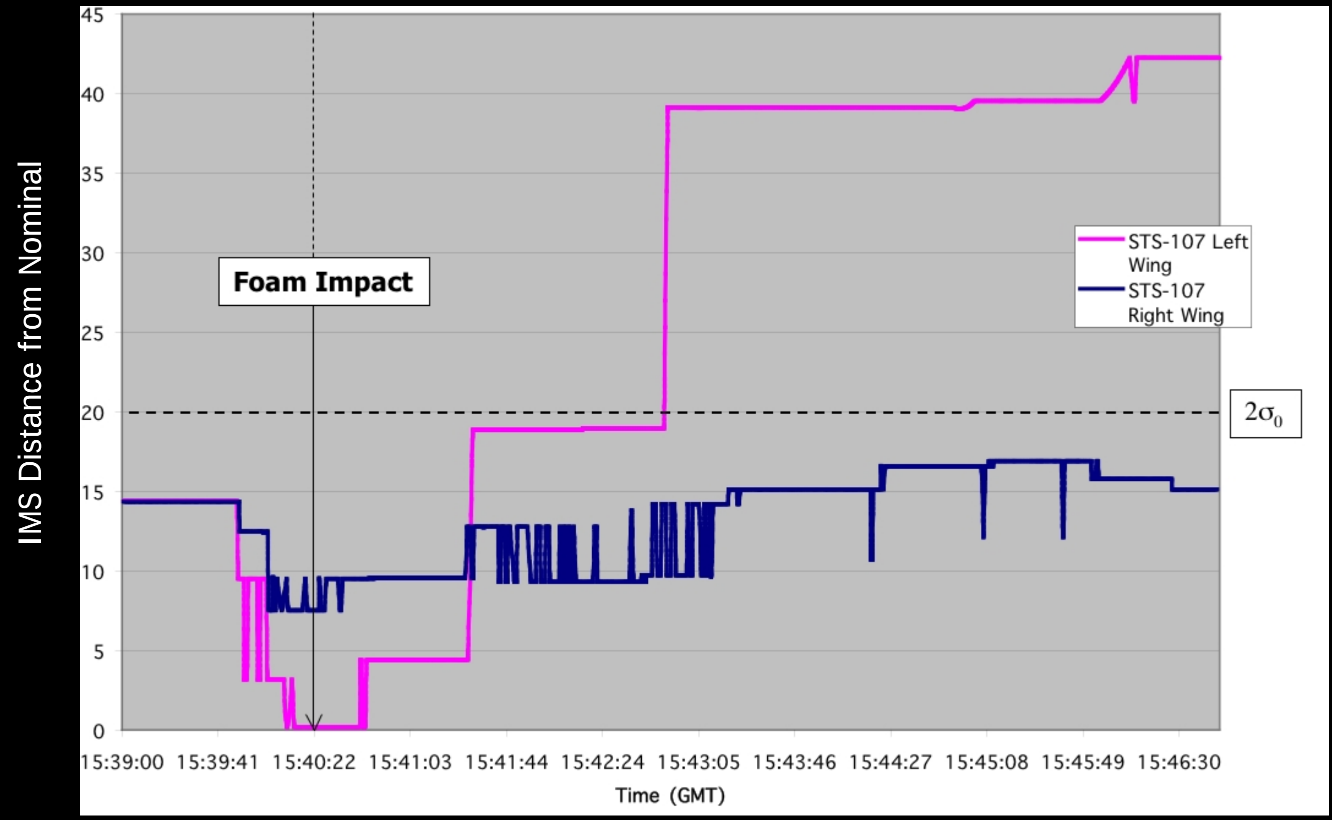
- Data expressed as value relative to a reference sensor



STS-93 Launch IMS Analysis



STS-107 Launch IMS Analysis



Our Innovations

- The IMS system automatically learns a model for nominal behavior to detect system anomalies.
- Orca provides a flexible platform to detect anomalies in massive data sets.
- IMS is used to detect wing impacts in support of STS-121 and STS-115.
- IMS will be deployed on Console at Mission Operations Directorate, JSC.

Conclusions

- Demonstrated transparent mining of discrete, continuous, and textual information to uncover safety anomalies.
- Enabling automated analysis of the Distributed National ASAP and FOQA Archives.
- The methods we discuss provide a comprehensive capability to monitor, detect, and analyze system anomalies.

Future Directions

- Advanced methods to analyze heterogeneous data sets.
- Prognostic and diagnostic methods for aircraft and space systems.
- Potential new book on text mining (Srivastava and Sahami): A collaboration between NASA and Google.
- SIAM Text Mining Competition: Classification of ASRS reports, sponsored by NASA .
- Data Mining in Science, Aeronautics, and Exploration Systems Conference 2007

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.