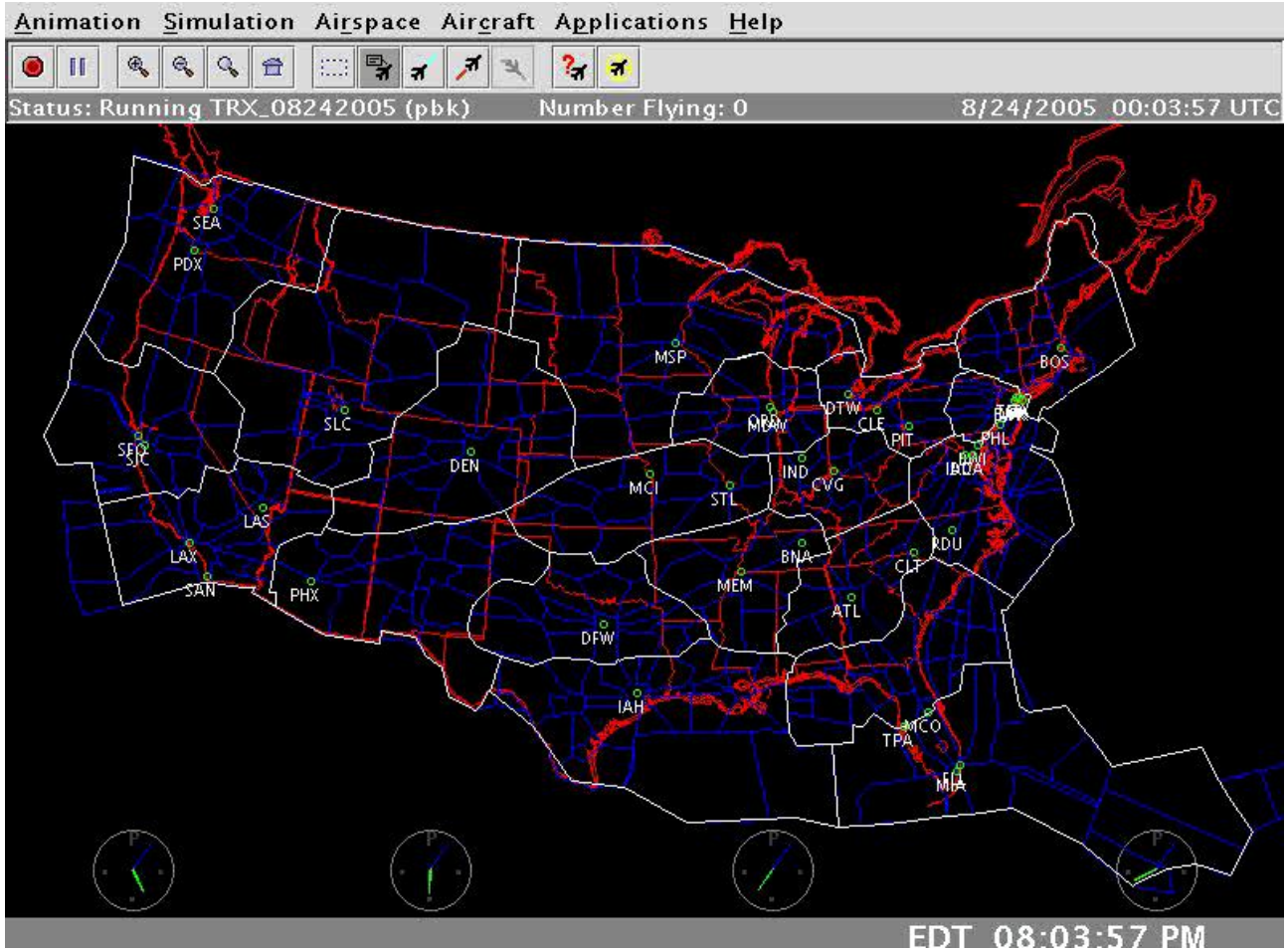
Discovering Precursors to Aviation Safety Incidents: from Massive Data to Actionable Information

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

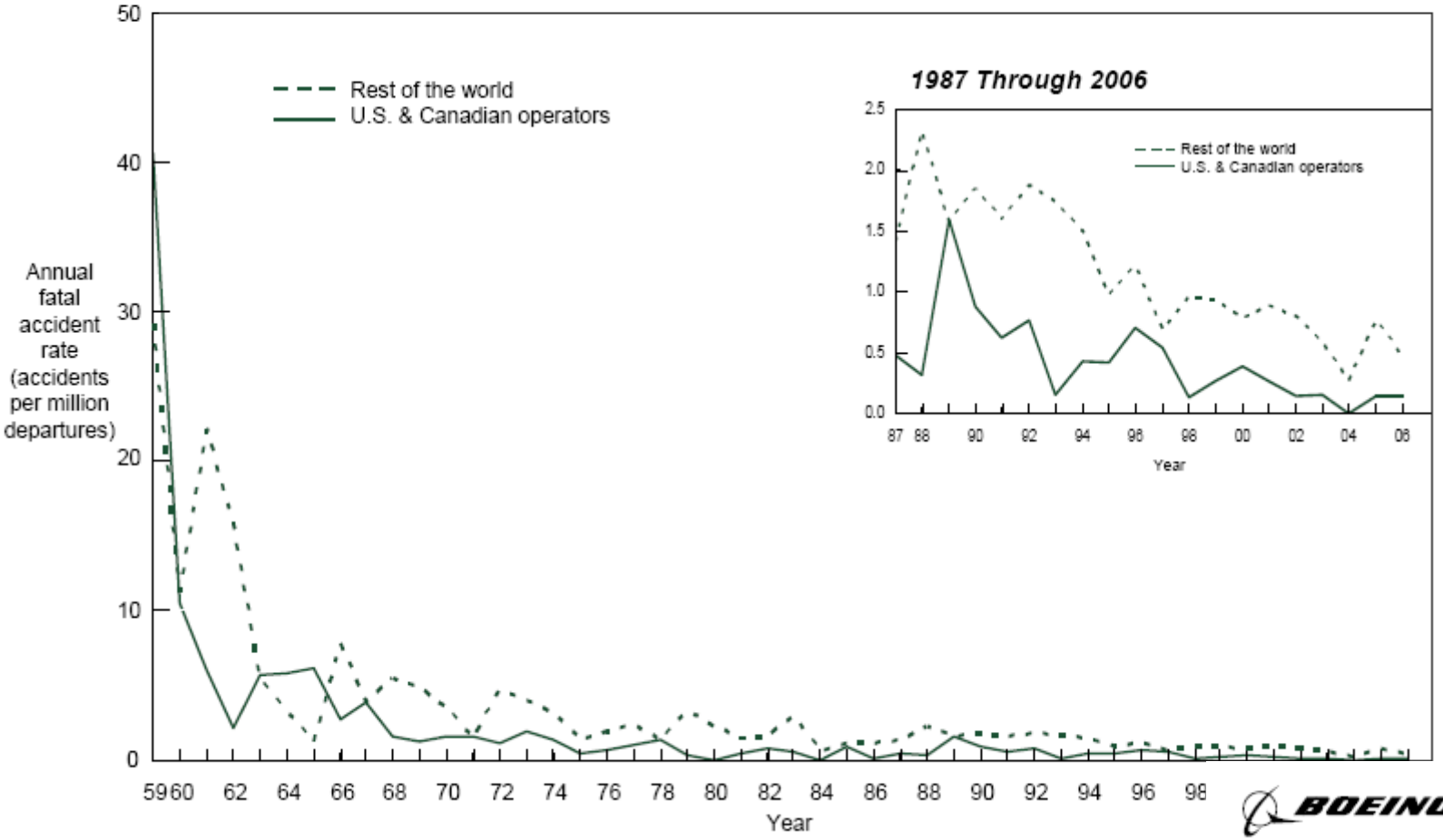
A Day in the Life of the National Airspace



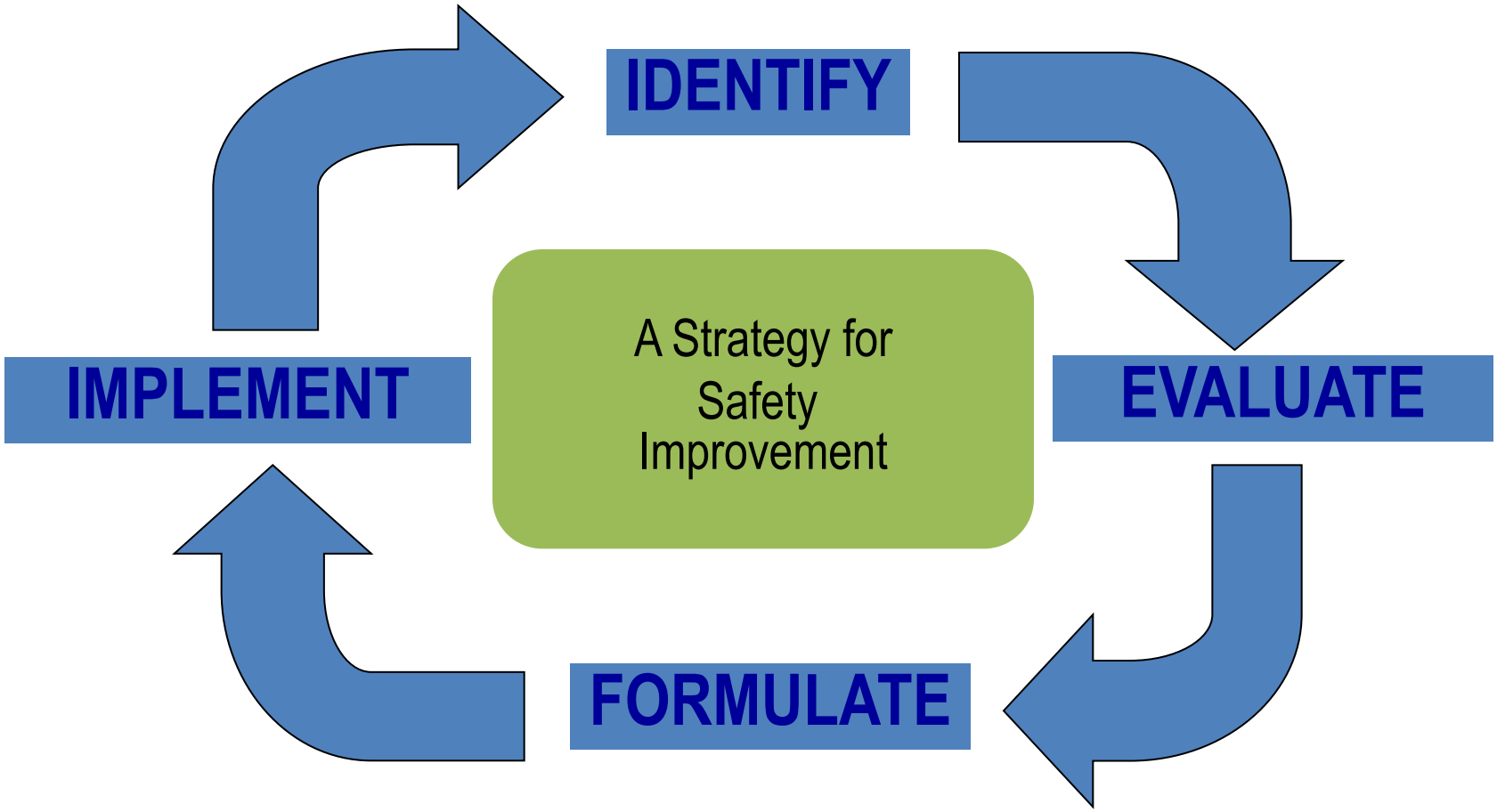
Recent Safety Advances

U.S. and Canadian Operators Accident Rates by Year

Fatal Accidents – Worldwide Commercial Jet Fleet – 1959 Through 2006

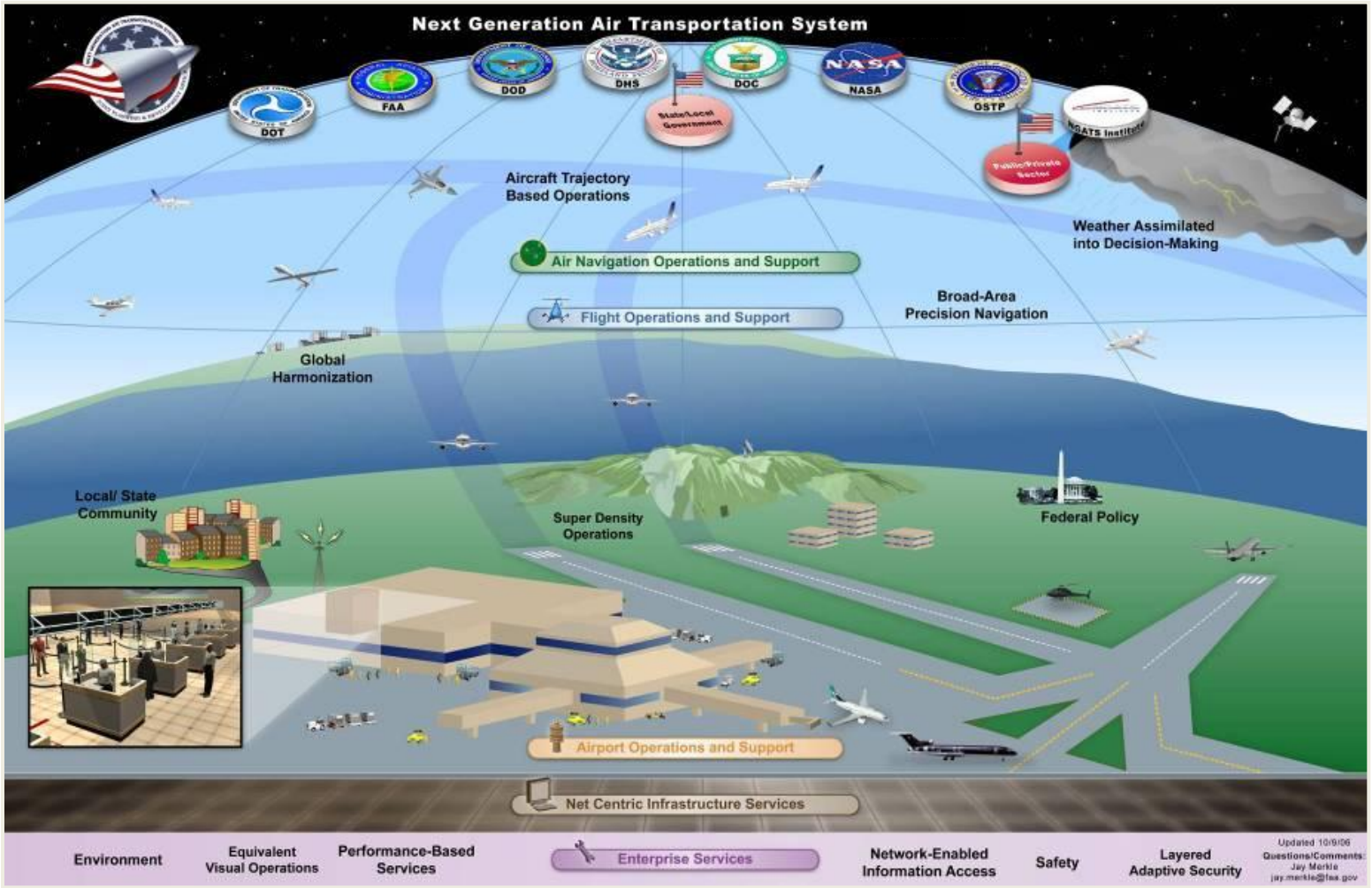


Proactively Managing Risk



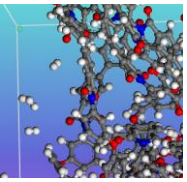
Reduce the accident rate by identifying and responding to precursor events *before* accidents occur.

Data Mining in Support of Global Operations



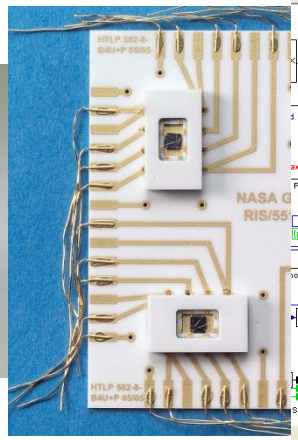
10^{-6} 10^{-5} 10^{-4} 10^{-3} 10^{-2} 10^{-1} 10^{0} 10^{1} 10^{2} 10^{3} 10^{4} 10^{5} 10^{6}

Data Arises from Molecular to Global Scales

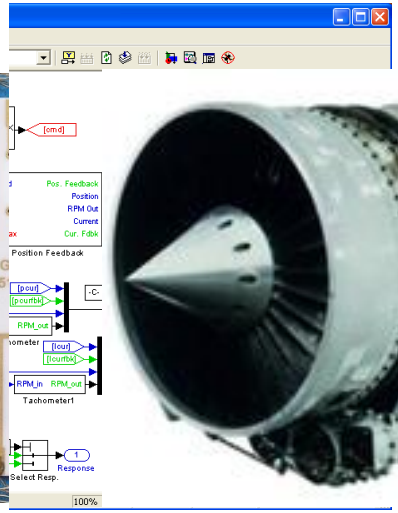


Molecules

Materials



Sensors



Software

Engines

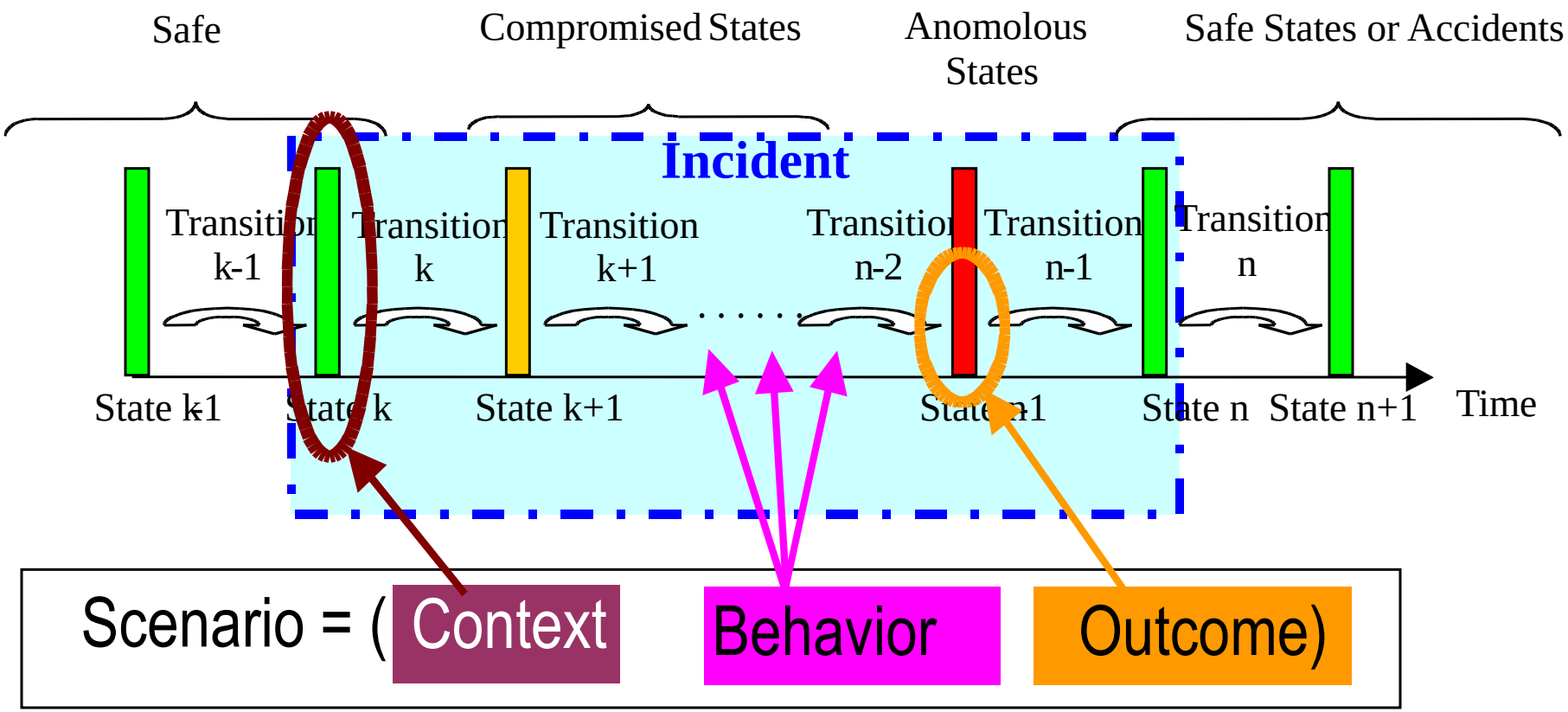
Aircraft

Vehicle Technology

Operations

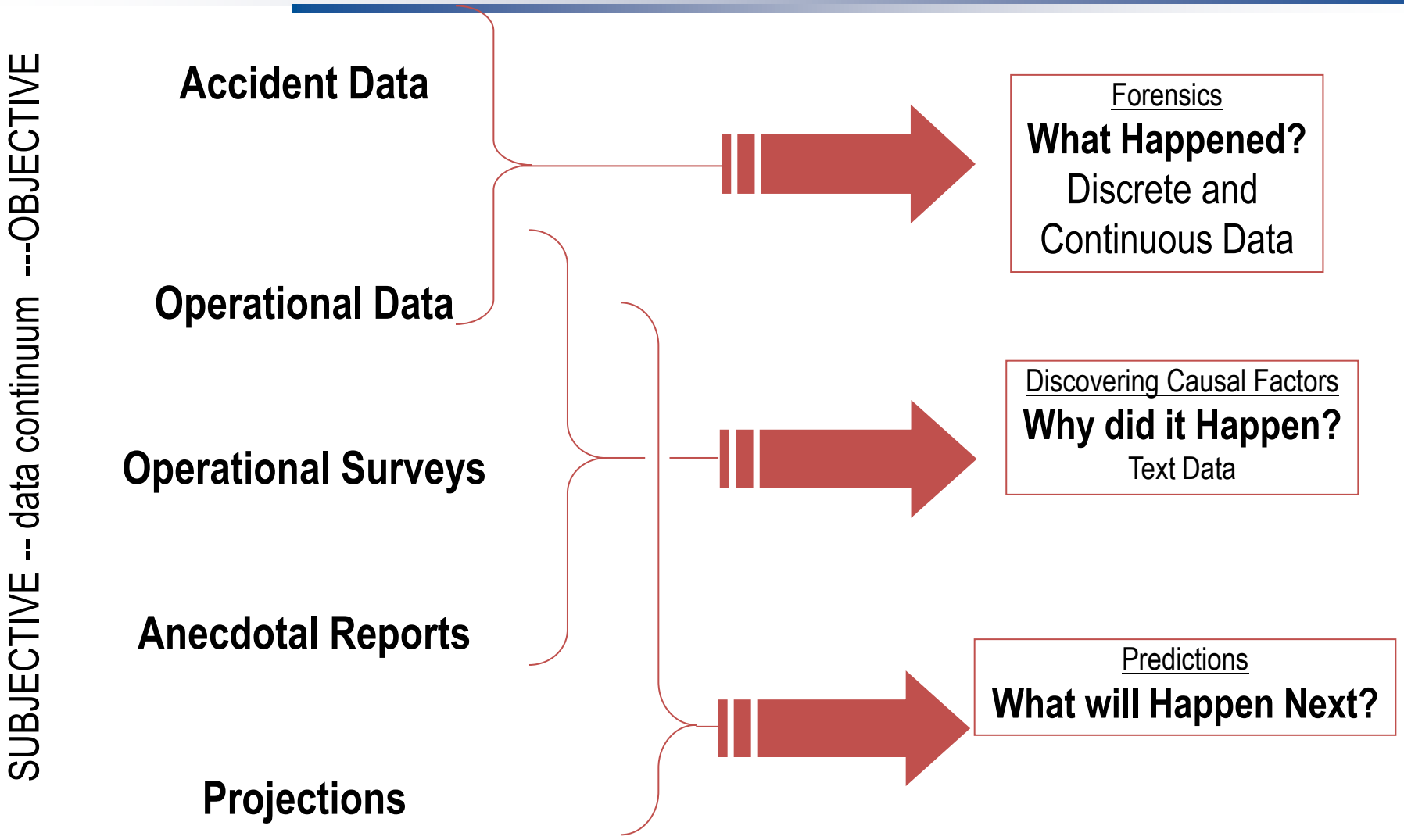
10^{-6} 10^{-5} 10^{-4} 10^{-3} 10^{-2} 10^{-1} 10^{0} 10^{1} 10^{2} 10^{3} 10^{4} 10^{5} 10^{6}

The Anatomy of an Aviation Safety Incident



From Irving Statler, Aviation Safety Monitoring and Modeling Project

Where are Precursors Found?



From Irving Statler, Aviation Safety Monitoring and Modeling Project

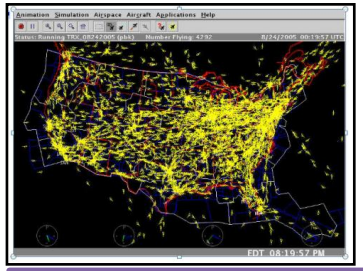
Aviation Safety Report Excerpt

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...

- Can answers **why** an incident happened
- Over 100K reports
- Multiple authors (multi lingual writers)
- Inconsistent use of abbreviations and punctuation

Mining Heterogeneous Data is the Key

- Primary Source: aircraft
- Can answer **what** happened in an aircraft during an Aviation Safety Incident



- Primary Source: Radar data
- Can answer what happened in the National Airspace during Aviation Safety Incident

Discrete

Continuous

$K(x_i, x_j)$

Textual

Networks

- Primary Source: Humans
- Can answer **why** an Aviation Safety Incident happened

Sample Text Report

JUST PRIOR TO TOUCHDOWN, LAX TWR TOLD US TO GO AROUND BECAUSE OF THE ACFT IN FRONT OF US.
...

Massive Data: Archives growing at 100K flights per month.

Mercer Kernel Based Approaches

for Discrete, Continuous, Textual, and other Sources

Multiple Kernel Anomaly Detection (KDD 2010)

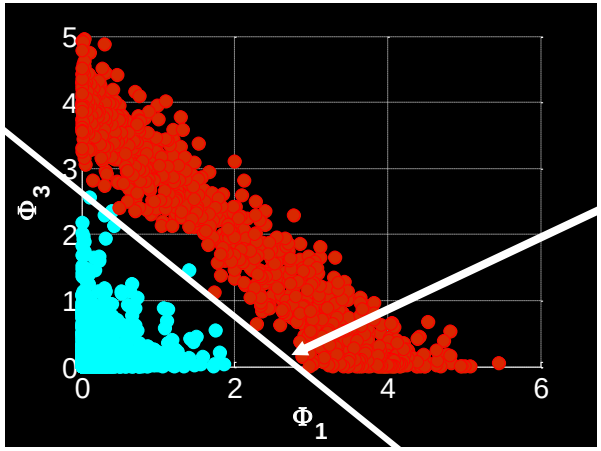
$$\min_{Q} = \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j\left(\sum_{\lambda}\beta_\lambda K^{\lambda}_{i,j}\right)$$

Subject to:

$$\sum_i \alpha_i = 1$$

$$\nu \in [0,1],$$

$$0 \le \alpha_i \le \frac{1}{l\nu}, \forall i$$



Unique Linear Decision Boundary

Multiple Kernel Learning for Heterogeneous Anomaly Detection: Algorithm and Aviation Safety Case Study

Santanu Das
UARC, UCSC
NASA Ames Research Center
Moffett Field, CA 94035
Santanu.Das-1@nasa.gov

Bryan L. Matthews
SGT Inc.
NASA Ames Research Center
Moffett Field, CA 94035
Bryan.L.Matthews@nasa.gov

Ashok N. Srivastava
NASA Ames Research Center
Moffett Field, CA 94035
Ashok.Srivastava@nasa.gov

Nikunj C. Oza
NASA Ames Research Center
Moffett Field, CA 94035
Nikunj.C.Oza@nasa.gov

ABSTRACT

The world-wide aviation system is one of the most complex dynamical systems ever developed and is generating data at an extremely rapid rate. Most modern commercial aircraft record several hundred flight parameters including information from the guidance, navigation, and control systems, the

General Terms

Algorithms, Human Factors, Performance, Reliability

Keywords

Aeronautics, Anomaly Detection, Prediction, Prognostics

Talk at 5 pm today, Independence A

Gaussian Process Regression

for Discrete and Continuous Sources (typically for small problems, see below)

- Training data
  - X data matrix of observations – n x d
  - y vector of target data – n x 1
- Test data
  - X* matrix of new observations – n* x d
- Covariance function
  - $K_{ij} = k(x_i, x_j), K^*_{ij} = k(x^*_i, x_j)$
- Goal
  - Predict y* corresponding to X*

- Model building
  - Train hyperparameters on a sample of X
  - Compute covariance matrix K (n x n)
- Prediction
  - Compute cross covariance matrix K* (n* x n)
  - Compute mean prediction on y* using

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

  - Compute variance of prediction using

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

Algorithm Analysis

- Storage Complexity: Storing covariance matrix $O(n^2)$
- Time Complexity: Computing matrix inversion $O(n^3)$

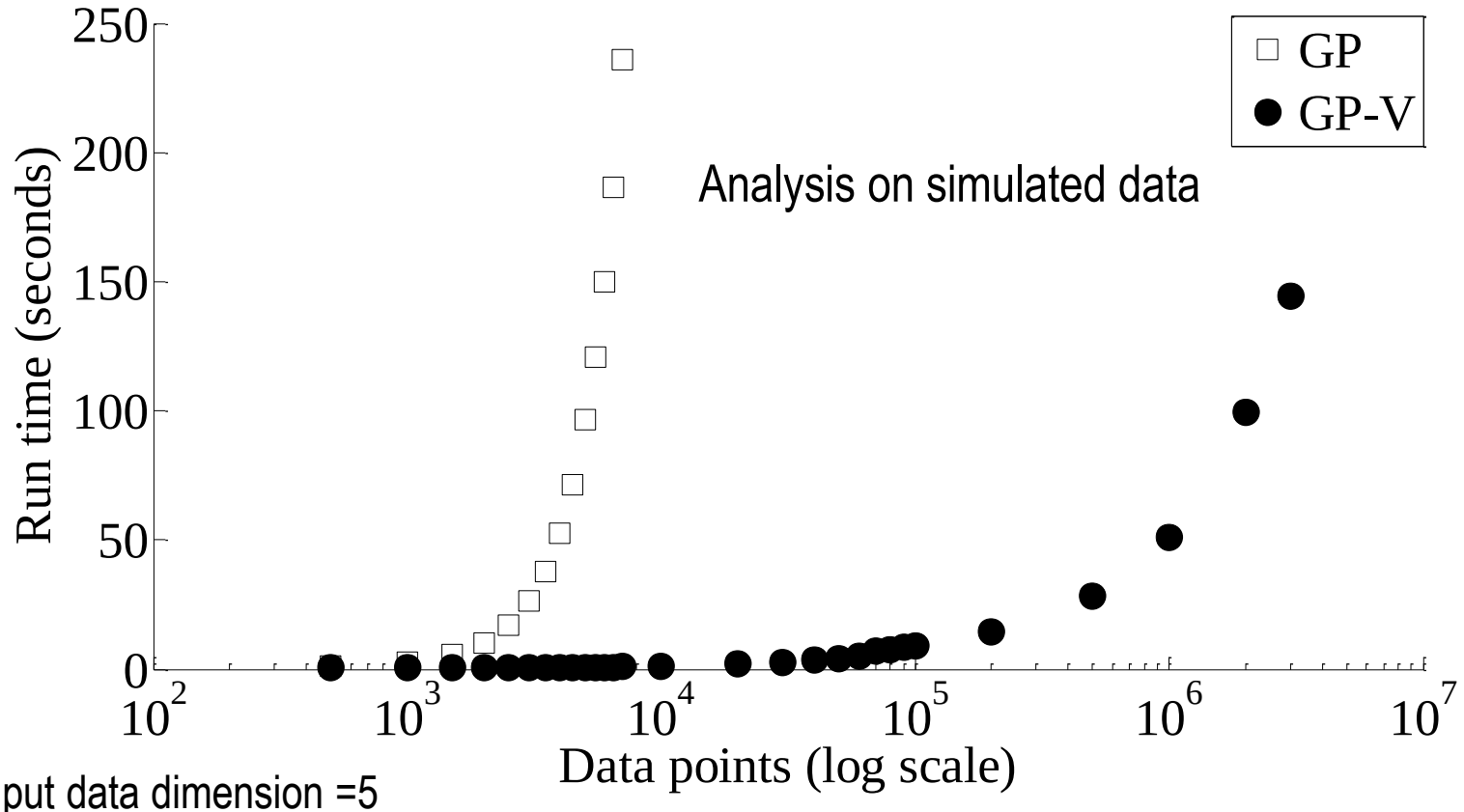
Mixtures of Gaussian Processes

Volker Tresp
Siemens AG, Corporate Technology, Department of Neural Computation
Otto-Hahn-Ring 6, 81730 München, Germany
Volker.Tresp@mchp.siemens.de

Table 1: The table shows results using artificial and real data sets of size $N = 100$ using $M = 10$ GPR models. The data set ART is generated by adding Gaussian noise with a standard deviation of 0.2 to a map defined by 5 normalized Gaussian bumps. *numin* is

V. Tresp. *Mixtures of Gaussian processes*. In Todd K. Leen, Thomas G. Dietterich, and Volker Tresp, editors, *Advances in Neural Information Processing Systems 13, pages 654–660*. MIT Press, 2001.

GP–V: Scaling to 3 million points



- Input data dimension =5
- Number of sample points = 3 million. **New method under review at ICDM 2010 for distributed implementation for massive data sets** (K. Das and A. N. Srivastava).
- Run time = time to build the model + time to evaluate 500 test points
- Hyper parameters trained on 100 sample points
- Accuracy does not degrade with approximation

L. Foster, 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.

How do we get the Word Out?

DASHlink

disseminate. collaborate. innovate.
https://dashlink.arc.nasa.gov/

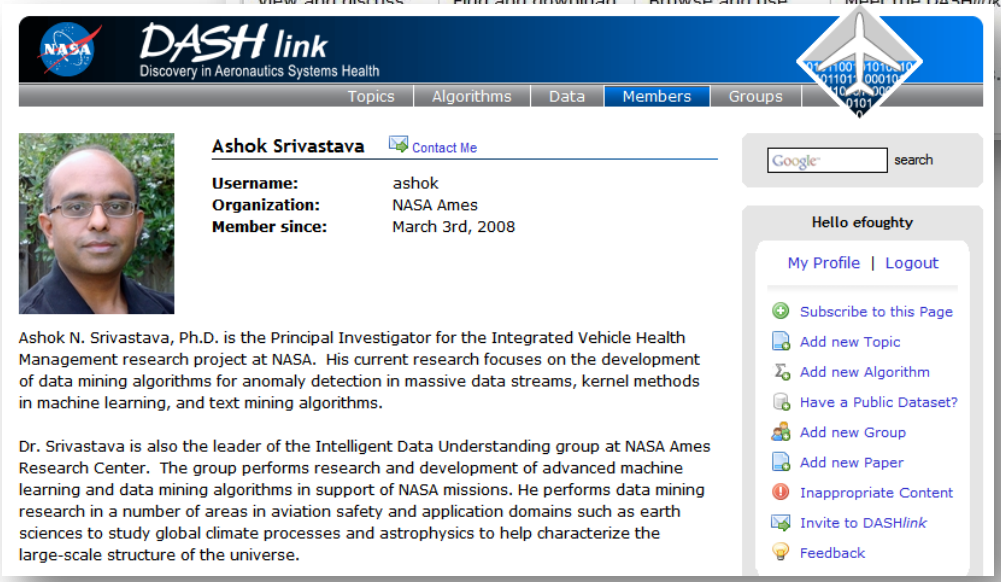
DASHlink is a collaborative website designed to promote:

- Sustainability
- Reproducibility
- Dissemination
- Community building

Users can create profiles

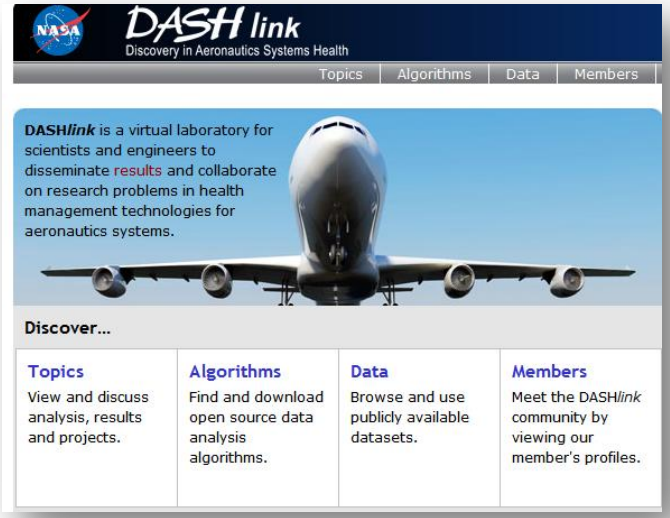
- Share papers, upload and download opensource algorithms
- Find NASA data sets.

Coming Soon… **DASHlink 2.0**.



Real World Impact: Flight Operations at Southwest Airlines

- NASA has **open-sourced** many of its key data mining algorithms for analysis of data from flight data recorders through **DASHlink**, our Web 2.0 portal for the world.
- **Southwest Airlines** obtained copies of sequenceMiner and Orca, two advanced anomaly detection techniques.
- Early results indicate that **operationally significant** events have been discovered by these algorithms that would not be triggered by their existing methods.

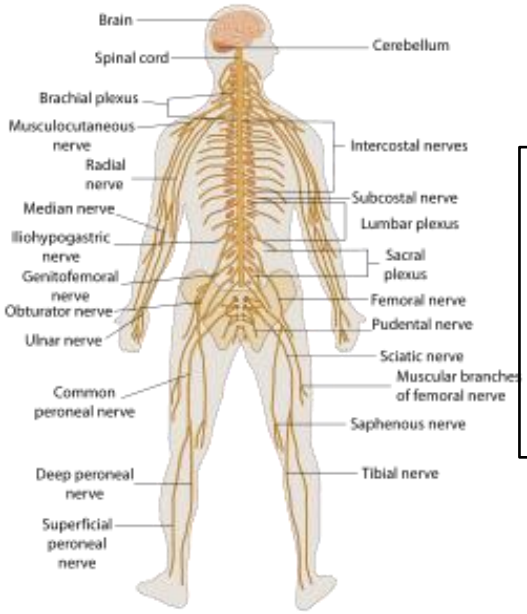


dashlink.arc.nasa.gov

- Southwest Airlines plans to incorporate algorithms into **daily operations**.

Mining Human Performance and Flight Data

Physiological Measurements



easyJet.com

Luton, UK

To EasyJet
- Near real-time decision support
- Understanding fatigue with an objective approach

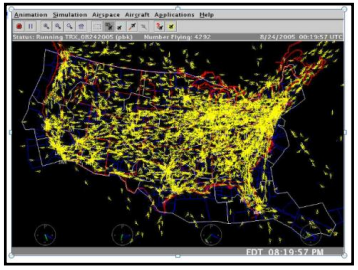
To NASA Data Mining Team
- Daily data
- 300 GB flights per month
- Physiology, text, cockpit, engines and flight parameters, flight path, network nformation.

NASA Data Mining Lab (Mountain View, CA)

Sample Text Report

JUST PRIOR TO TOUCHDOWN, LAX TWR TOLD US TO GO AROUND BECAUSE OF THE ACFT IN FRONT OF US. …

Key question: How does fatigue effect aviation safety?



Data Mining in Scientific Domains

- Aeronautical Systems
  - 100K flights per month
- Earth and Space Science
  - Earth Observing System generates ~21 TB of data per week.
  - Ames simulations generating 1-5 TB per day
- Exploration Systems
  - Space Shuttle and International Space station downlinks about 1.5GB per day.
- Kepler Planetary Discovery Data Systems
  - 100K solar systems every 30 minutes

Researchers world-wide can work with us in numerous domains through our new collaborative portal

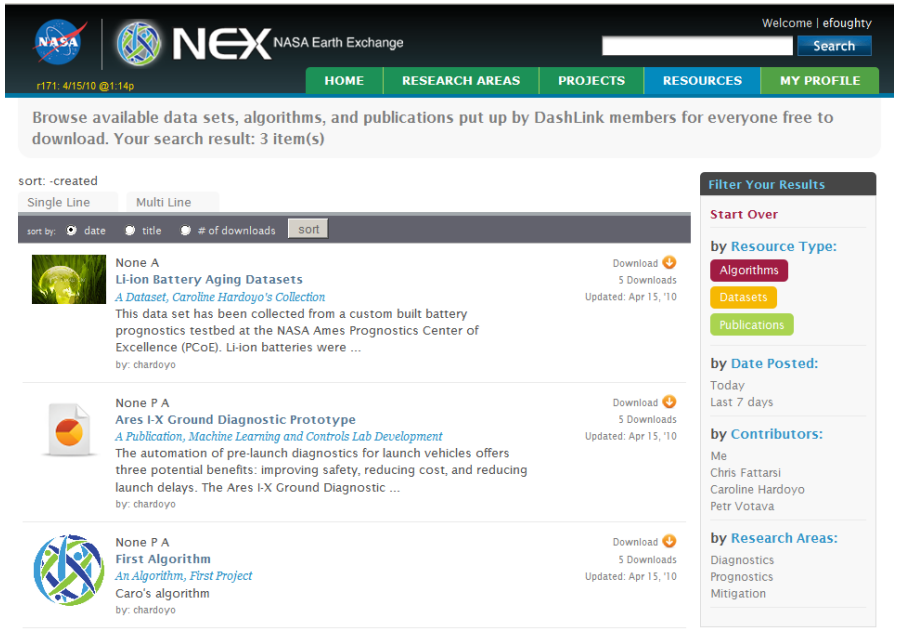
Introducing NASA Earth Exchange (NEX)
Collaborative Earth Science

NEX provides a platform for researchers to share information and data derived from NASA's Earth observing satellites. Its goal is to encourage exploration and collaboration to create new ways to understand and improve life here on Earth.

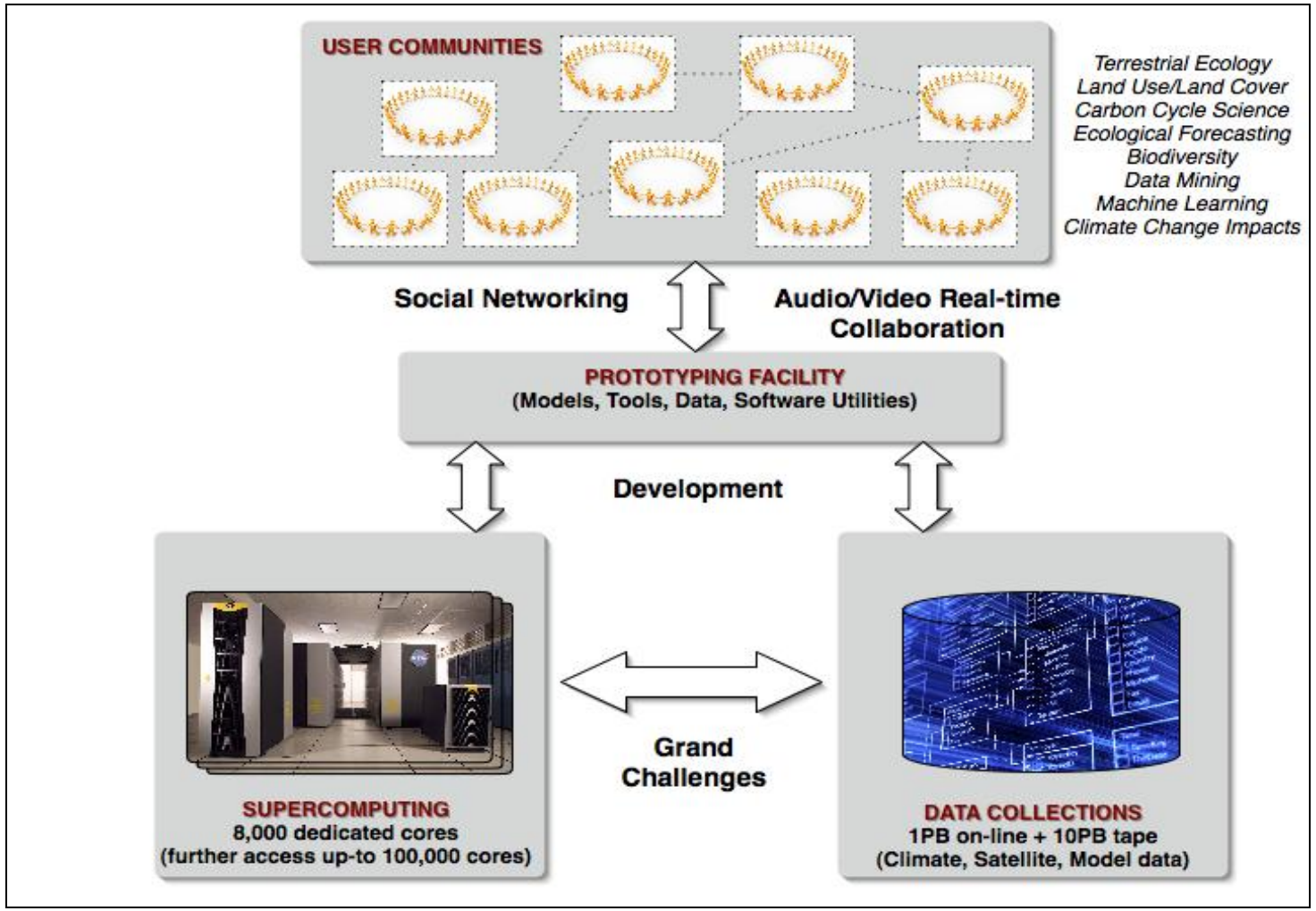
Public Release: August 2010

NEX website and DASHlink 2.0 Capabilities

- Collaboration with NASA on Projects (and initiating the process for further collaboration with NEX researchers on **NASA supercomputing** facilities).
- Share Resources: Papers, algorithms, data sets etc.
- Find and contact other researchers in your field via member profiles.
- Learn about other cutting edge work at NASA and beyond.
- Cross reference work on DASHlink and NEX.



NASA Earth Exchange Components



The Data Mining Team

Group Members

Ashok N. Srivastava
Kanishka Bhaduri
Kamalika Das
Santanu Das
Elizabeth Foughty
Dave Iverson
Rodney Martin
Bryan Matthews
Dawn McIntosh
Nikunj Oza
Mark Schwabacher
John Stutz
David Wolpert
+ 7 summer students

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.