Data Mining for Aviation Safety

Nikunj C. Oza, Ph.D.
NASA Ames Research Center
Nikunj.C.Oza@nasa.gov
http://ti.arc.nasa.gov/people/oza

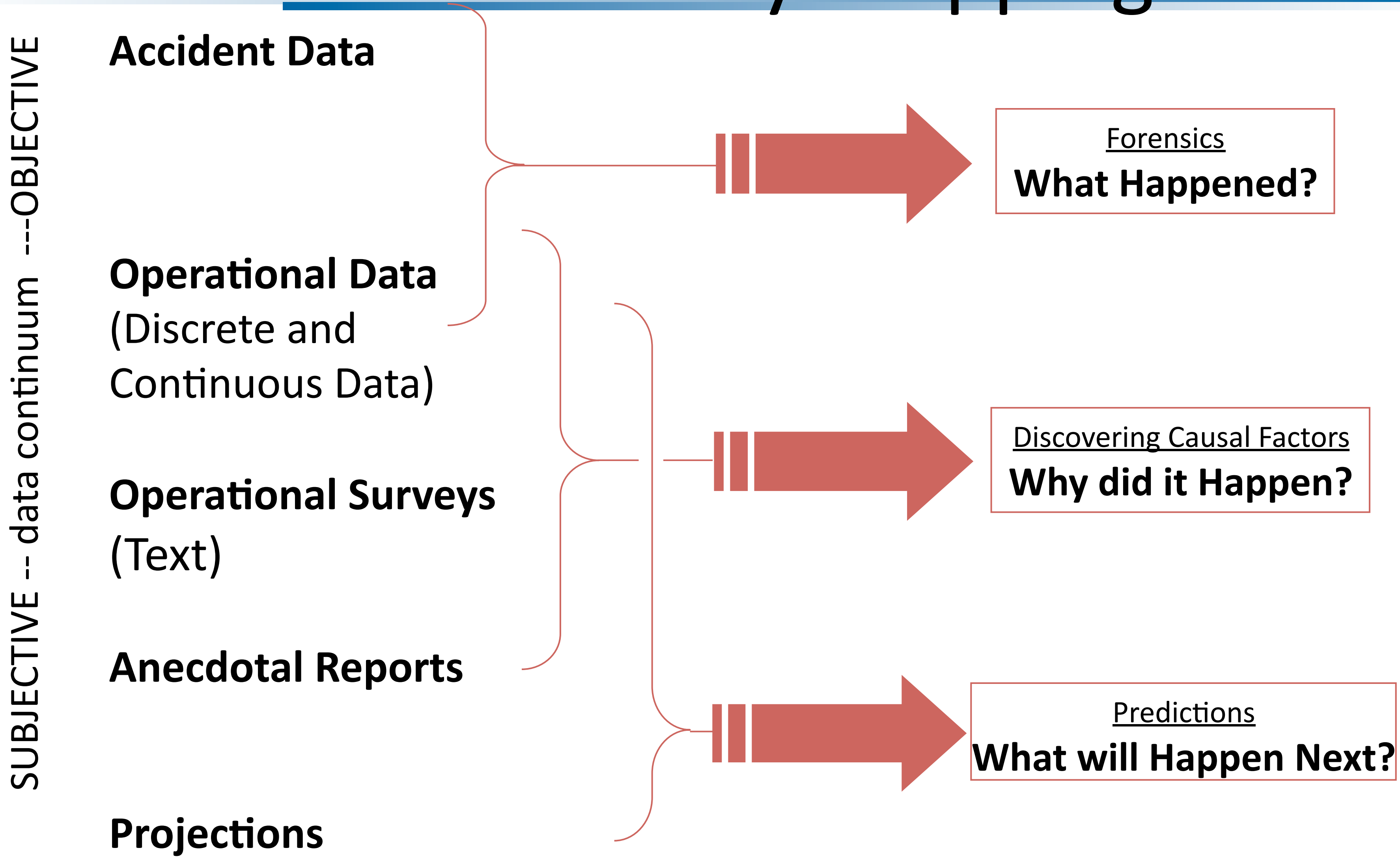
American Statistical Association, San Francisco Bay Area Chapter, October 21, 2010

Outline

- The breadth/depth of the problems
- Combining discrete and continuous sequences: Multiple Kernel Anomaly Detection (MKAD)

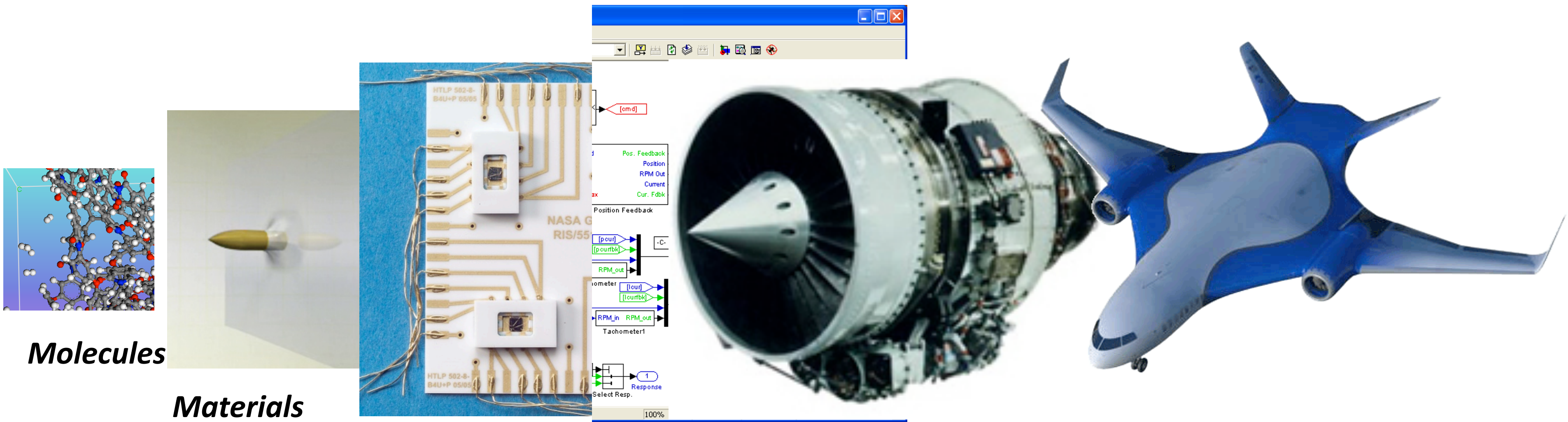
  Derived from anomaly detection methods on discrete and continuous sequences.
- Text Mining: classification, topic modeling
- Ongoing, future work

Aviation Safety Mapping



From Irving Statler, Aviation Safety Monitoring and Modeling Project

Data Arises from Molecular to Global Scales

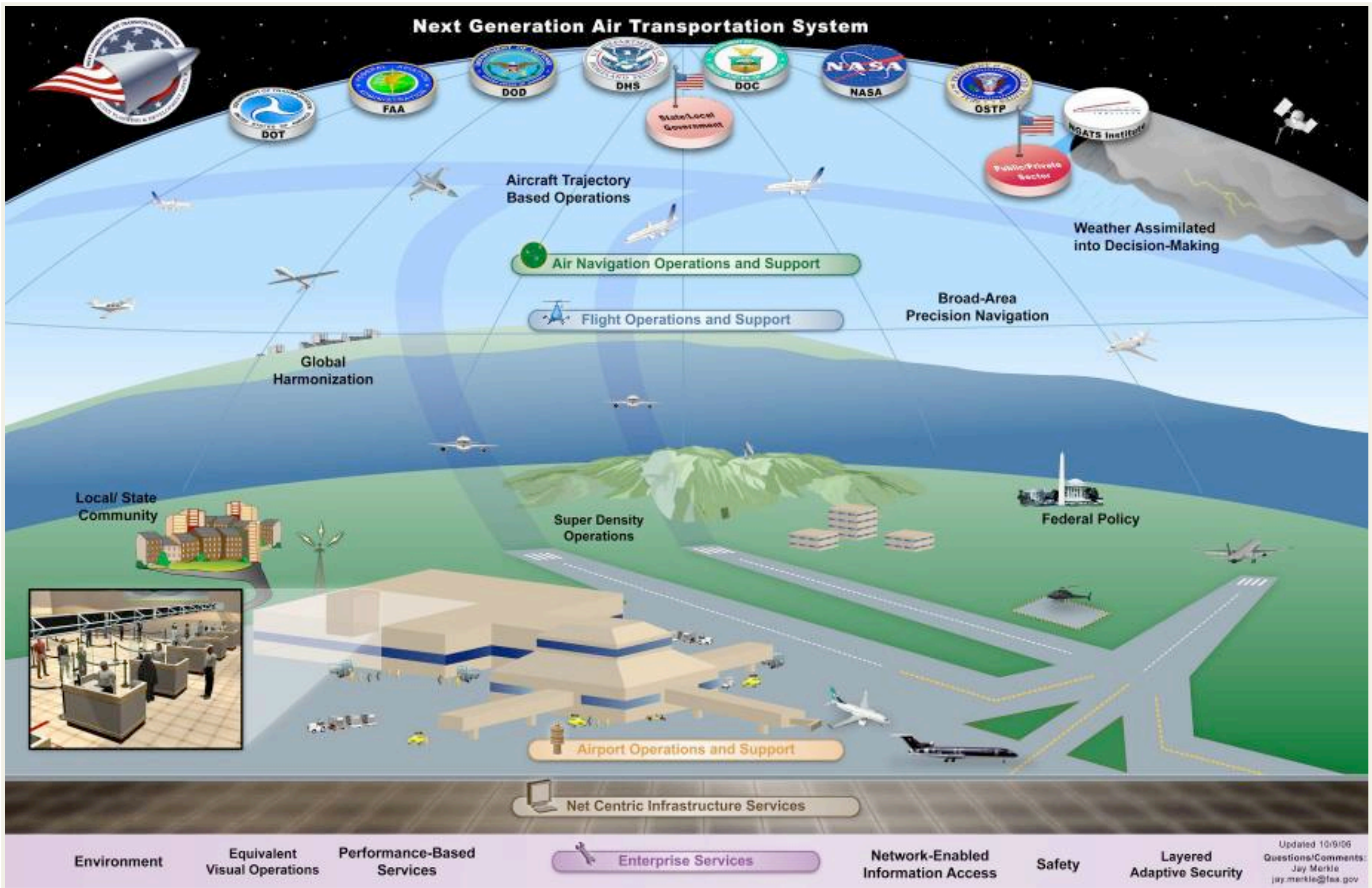


Molecules ***Materials*** ***Sensors*** ***Software*** ***Engines*** ***Aircraft***

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}

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 Mining

- IVHM/SSAT project goals require leveraging substantial data.
  - Aircraft-produced: Sensor data, flight-related data (e.g., origin, destination), covering many flights over many years.
  - Other: Safety reports, trajectories
- Transform data into useful knowledge
  - Tools for Detection, Diagnosis, Prognosis, Mitigation
  - Levels ranging from flight-level to national air space level

Outline

- The breadth/depth of the problems
- Combining discrete and continuous sequences: Multiple Kernel Anomaly Detection (MKAD)

  Derived from anomaly detection methods on discrete and continuous sequences.
- Text Mining: classification, topic modeling
- Ongoing, future work

Anomalies in Discrete Sequences

- Need to model the behavior of discrete sensors and switches 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.

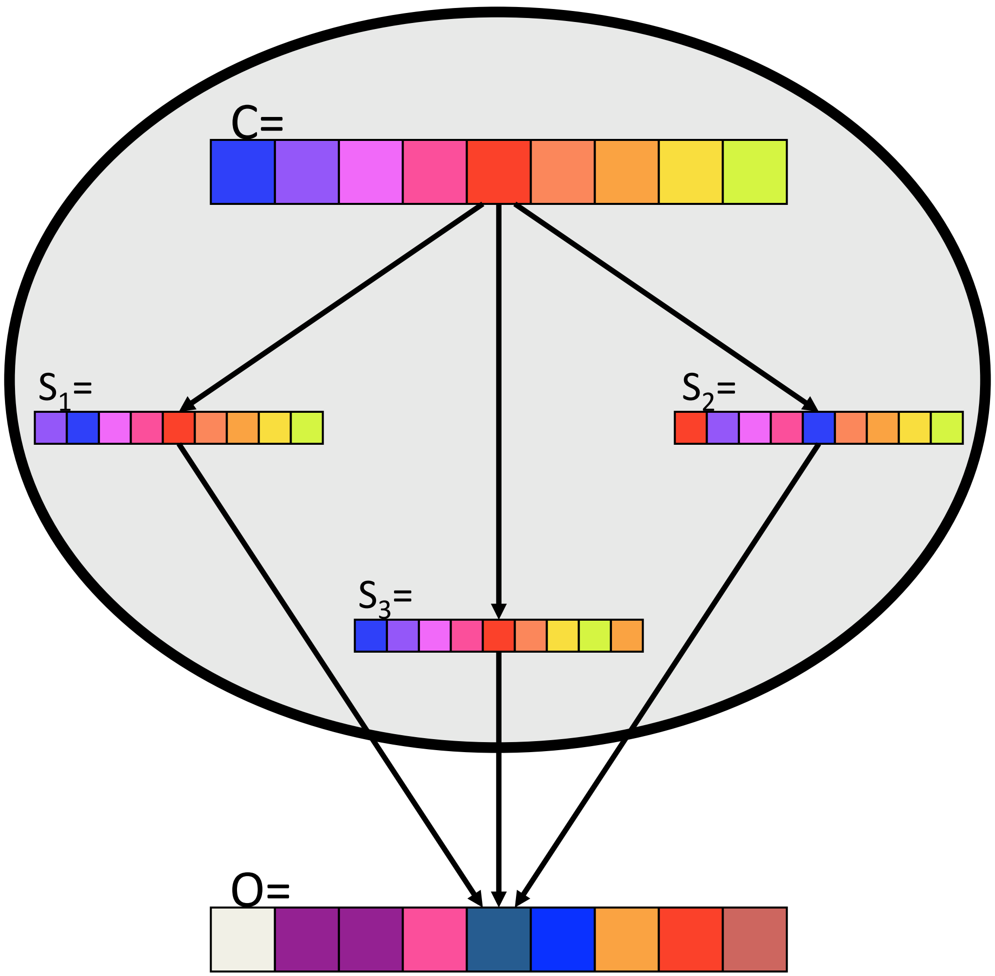
Solution

- We developed sequenceMiner:

  Each flight is analyzed as a sequence of events, accounting for

  - order in which switches change values
  - frequency of occurrence of switches
- Two Tasks:
  - Given a group of flights, find flights that are anomalous.
  - Given an anomalous flight, describe the anomalies and the degree of anomalousness.
- Method based on techniques used in bioinformatics.

Bayesian Model



- Sequences S_i in a cluster is dependent on a prototype C
- The outlier O is dependent on the S_i
- Therefore:

$$P(O|C) = \sum_i P(O|S_i) \times P(S_i|C)$$

- This forms the basis of our objective function *F* that allows us to describe why sequences are anomalous.

Missing and Extra Switches

- $P(O|S_i)$ is proportional to the normalized length of the longest common subsequence of O and S_i
  - Discovery of missing and extra symbols is done by aligning the sequences
- Missing and extra symbols are those whose addition to or deletion from an outlier sequence O would make O more normal with respect to the objective function *F*

Alignment

Cluster Members	A	B	_	_	E	_
	A	B	C	_	E	F
	A	B	C	_	_	F
Outlier Sequence	A	_	C	D	E	F

Missing a "B"

Has extra "D"

SequenceMiner

Normalized Longest Common Subsequence (NLCS)

$$\frac{L\left(h\left(s_i, s_j\right)\right)}{\sqrt{L\left(s_i\right) \times L\left(s_j\right)}}$$

where the functions $h(.)$ and $L(.)$ calculate the longest common subsequence and length of the sequences.

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
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

	seconds to touchdown												
	309	309	309	309	309	309	308	308	308	302	288	275	258
on	▲	▲				▲			▲	▲	▲		▲
	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
off			▼	▼	▼		▼	▼				▼	
	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.

MKAD for Fleet wide analysis

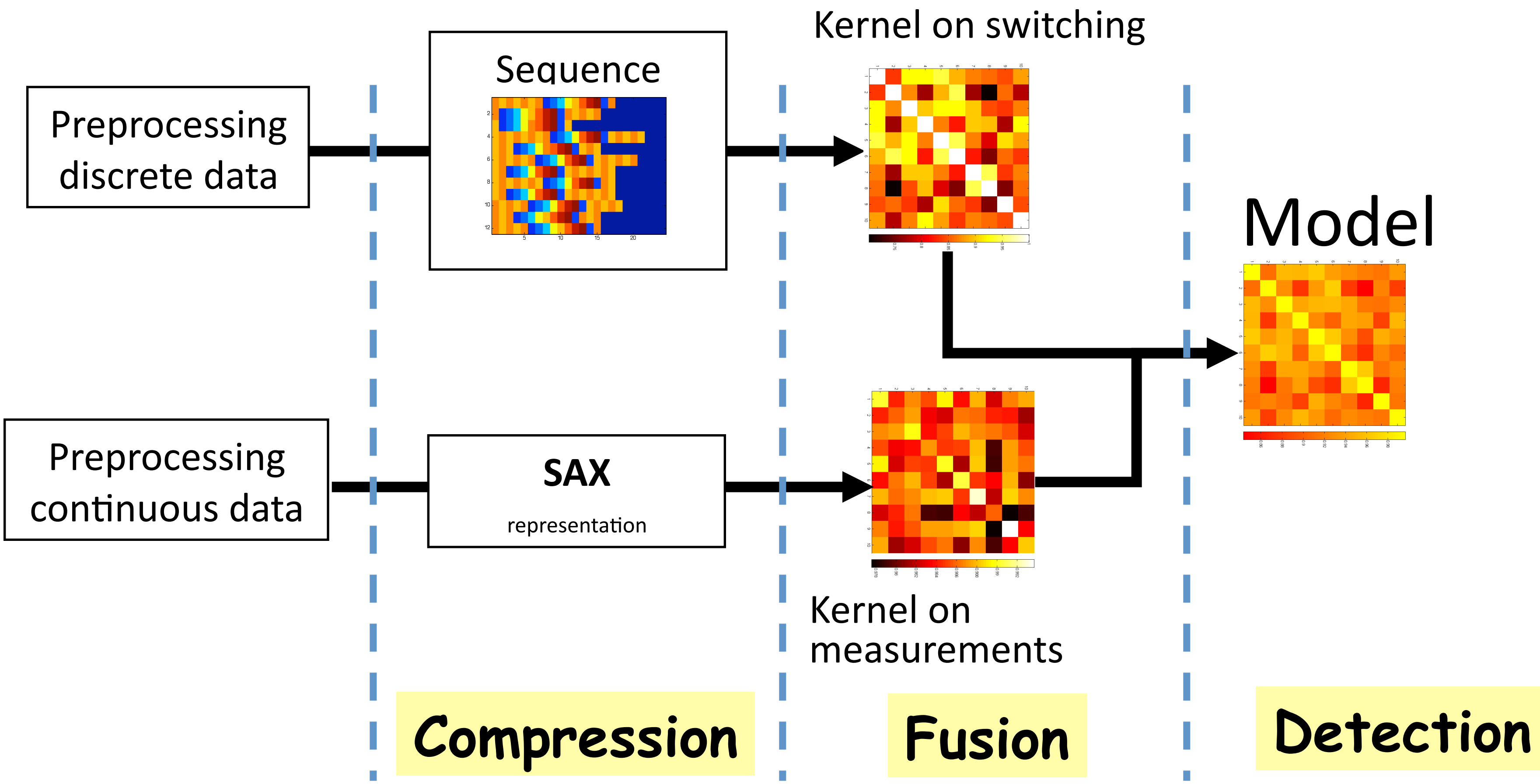
Flight Data Monitoring

Sequences D and continuous data streams C interactions

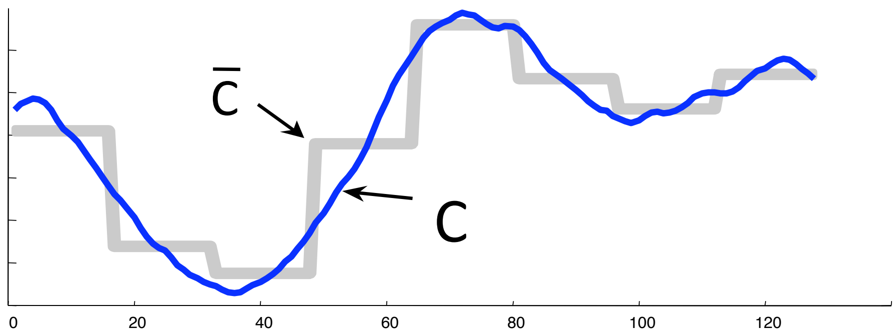
How to integrate all information in a **concise** and **intuitive** manner?

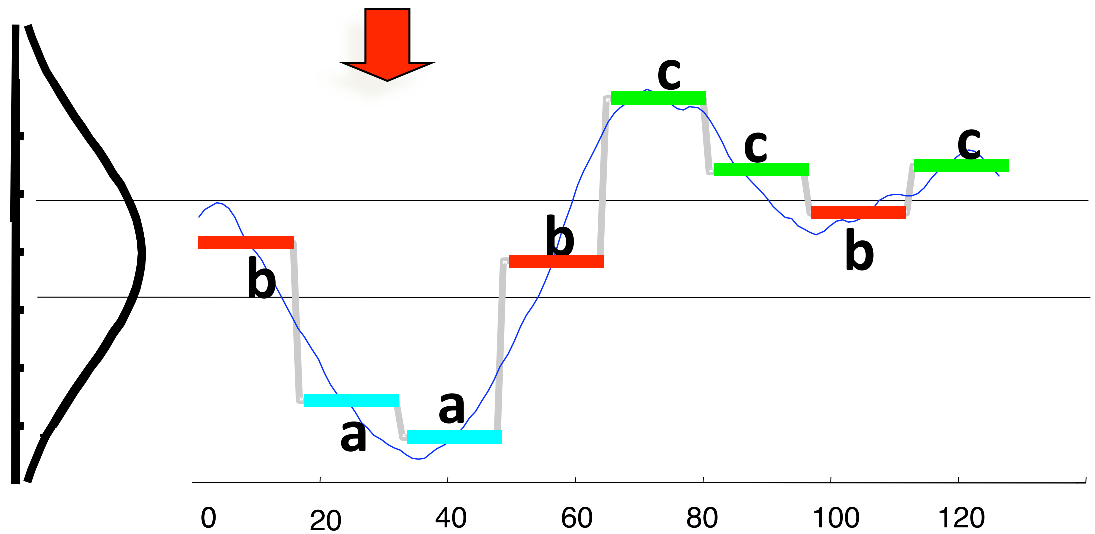
Compression,
Feature extraction,
Fusion,
Anomaly detection

MKAD Framework



SAX Representation





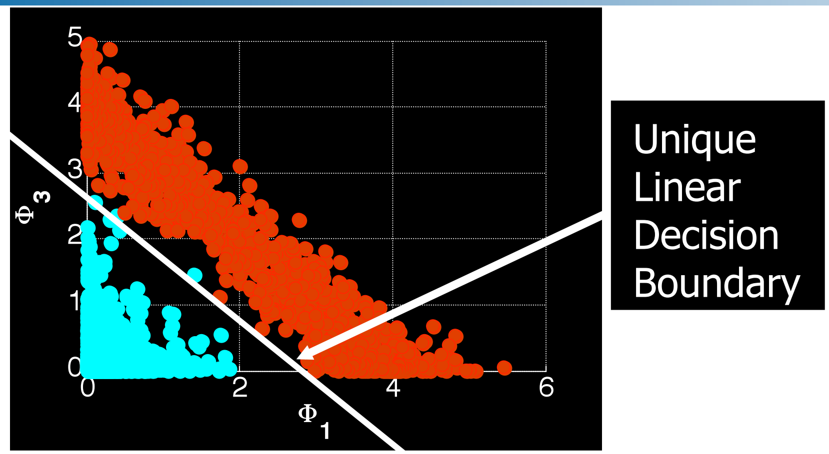
First convert the time series to Piecewise Aggregate Approximation (PAA) representation, then convert to symbols

It takes linear time

baabccbc

Slide courtesy of Eamonn Keogh and Jessica Lin

Optimization problem



$$\text{minimize} \quad Q = \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j \left(\beta K_d(x_i, x_j) + (1-\beta) K_c(x_i, x_j)\right)$$

$$\text{subject to} \quad 0 \le \alpha_i \le \frac{1}{\ell\nu}, \quad \nu \in [0,1], \quad \sum_i \alpha_i = 1$$

Discrete kernel

Continuous kernel

In the objective function, each entry of the discrete kernel and the continuous kernel represents the score obtained using longest common subsequence (LCS) of discrete signals and SAXified continuous signals, respectively.

Pairwise Similarity Measure

Kernel on discrete : Normalized Longest Common Subsequence (NLCS)

$$K_d(f_i, f_j) = \frac{L(h(s_i, s_j))}{\sqrt{L(s_i) \times L(s_j)}}$$

Kernel on continuous : Inversely proportional to distance between SAX representations of sequences

General Case, Multiple Kernels

One class SVMs training algorithms require solving the quadratic problem

Dual form

$$\min Q = \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j\left(\boxed{\sum_{\lambda}\beta_\lambda K_\lambda\left(x_i,x_j\right)}\right)$$

Subject to:

$$\sum_i \alpha_i = 1$$

Linear equality constraint

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

Control parameter

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

Bounds on design variables

Also: $$\sum_{\lambda}\beta_\lambda = 1$$

α : Lagrange multipliers of the primal QP problem

Anomaly scores

Decision boundary is determined only by margin and non-margin support vectors obtained by solving the QP problem

$$h(\alpha, \beta, f_z, \rho) = \sum_i \alpha_i \left(\sum_\lambda \beta_\lambda K_{i,z}^\lambda \right) - \rho$$

Datapoints with $\alpha_k > 0$ will be the support vectors

Indicator

Sign of *h: if negative – outlier*
if positive - normal

Magnitude of *h: degree of normality/anomalousness*

Synthetic Experiment

Simulation data

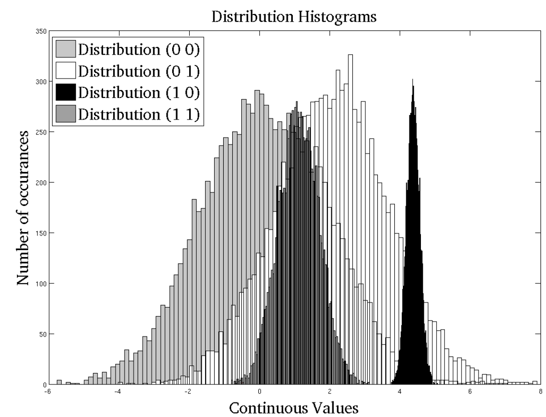
Type 1 – (Missing event) Flaps were not extended to normal full deployment at landing.

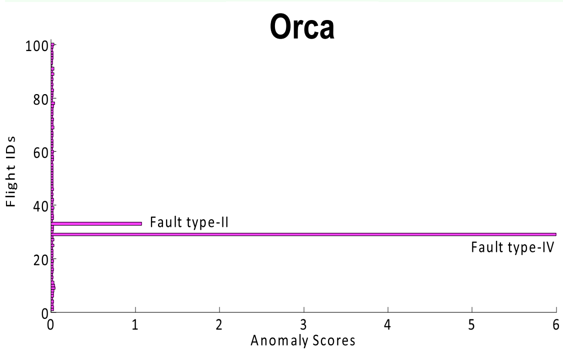
Type 2 - (Extra event) Landing gear was retracted after being deployed on final approach.

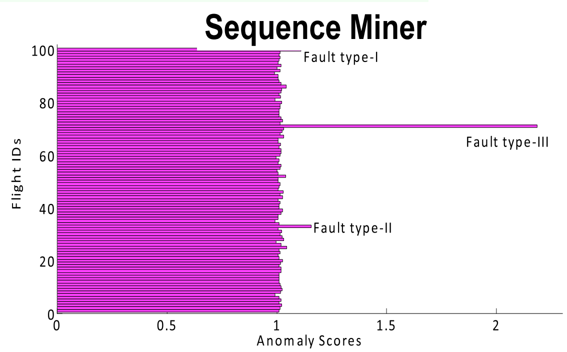
Type 3 – (Out of order event) Gear deployed before initial flaps below flaps limit.

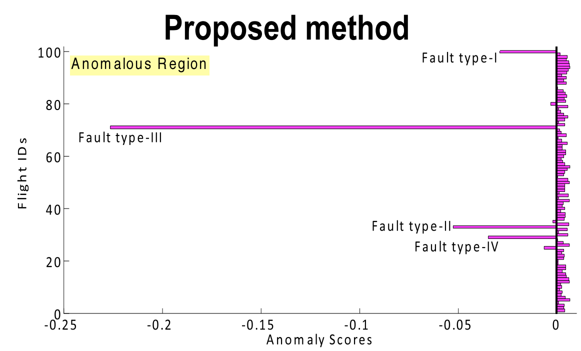
Type 4 – (Continuous anomaly) High bank angles or rate of descent below 1,000 ft.

Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
3	0	0	0	0	1	0	1	0	1	0
4	0	0	0	0	1	0	1	0	1	0
5	0	0	0	0	1	0	0	0	1	0
6	1	1	1	0	1	1	0	1	0	1
7	0	1	1	0	1	1	0	1	0	1
8	0	1	1	0	1	1	0	1	0	1
9	0	1	1	0	1	1	0	1	0	1

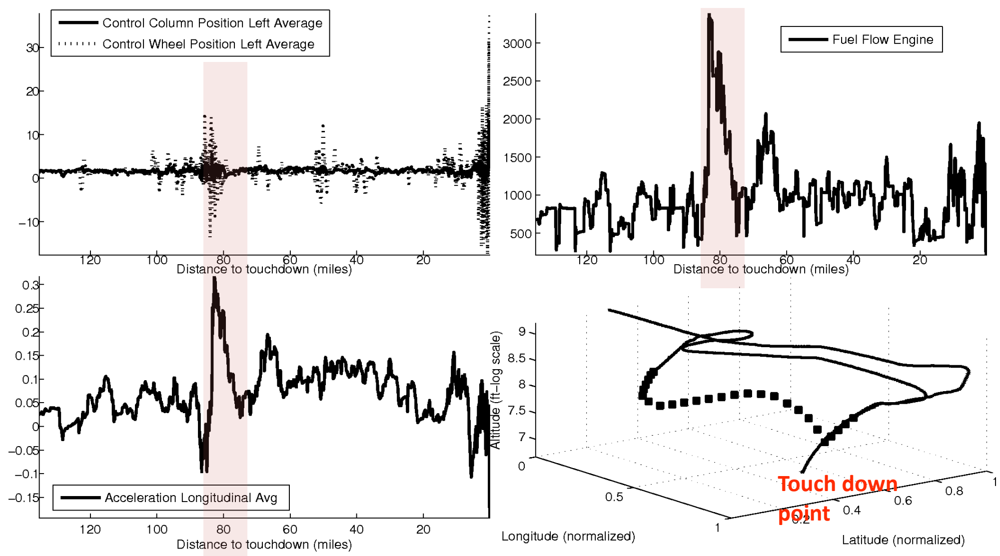








Case study: FOQA anomaly detection



Normal Extra Missing

- Flaps1 On
- Landing_Gear_Sel_Dwn On
- Flaps2 On
- FlapsFull On
- Ground_Spoilers_Deployed On
- Landing_Gear_Sel_Dwn Off
- FlapsFull Off
- Flaps2 Off
- Flaps1 Off
- Flaps1 On
- Landing_Gear_Sel_Dwn On
- Flaps2 On
- FlapsFull On

- The tradition methods cannot detect and monitor these anomalous activities that may have occurred simultaneously and are heterogeneous in nature.

MKAD Summary

Performs

.... anomaly detection on multivariate mixed attributes where discrete may influences the system dynamics which is reflected on the continuous data streams.

Highlights

.. High detection rate on most operationally significant anomalies in fleet wide analysis on large datasets

.. Discover some "unknown unknowns"

Application

1. Support flights safety experts
2. Schedule maintenance

Outline

- Combining discrete and continuous sequences: Multiple Kernel Anomaly Detection (MKAD)
  - Derived from anomaly detection methods on discrete and continuous sequences.
- Text Mining: classification, topic modeling
- Ongoing, future work

Aviation Safety Text Reports

- Pilots are encouraged to report incidents, concerns, or unsafe conditions. Two repositories:
  - The NASA-FAA Aviation Safety Reporting System (ASRS).
  - Each individual airline has an Aviation Safety Action Program (ASAP).
- Reports can be analyzed to improve aviation safety.
- Reports need to be correctly and consistently categorized by event type to determine dangerous situations and track trends of incidents and events.
- Event types are not enough to capture all anomalies. Topic identification needed to identify new event types, combinations of event types.

Manually Classifying Reports

- Review and analysis of reports is labor intensive.
  - In ASRS, reviewers have classified about 135,000 ASRS reports of 715,000 total reports submitted into 60 overlapping anomaly categories.
  - The length of the reports range from 0.5-4 pages long.
  - ASRS reports accrue at about 3,000 per month.
  - Classifications not always consistent due to multiple reviewers, changing experiences.
- Historical reports must be reread when new categories are added or changed.
- Current systems rely heavily on human memories for historical perspectives.

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...

Note: Industry specific vocabulary and abbreviations.

Automatic Categorization of ASRS Reports

A single report can be in multiple categories.

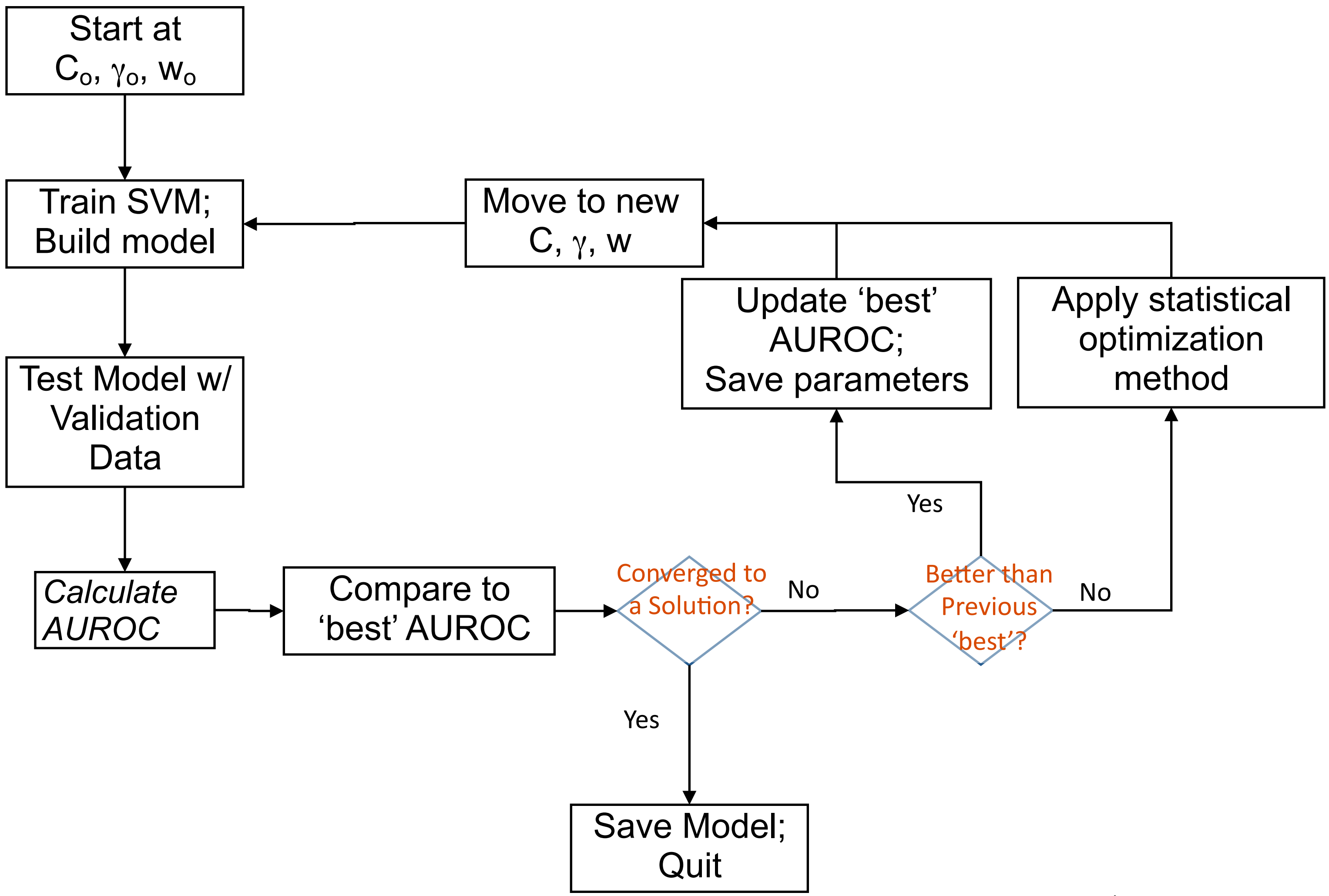
ASRS Report Excerpt

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

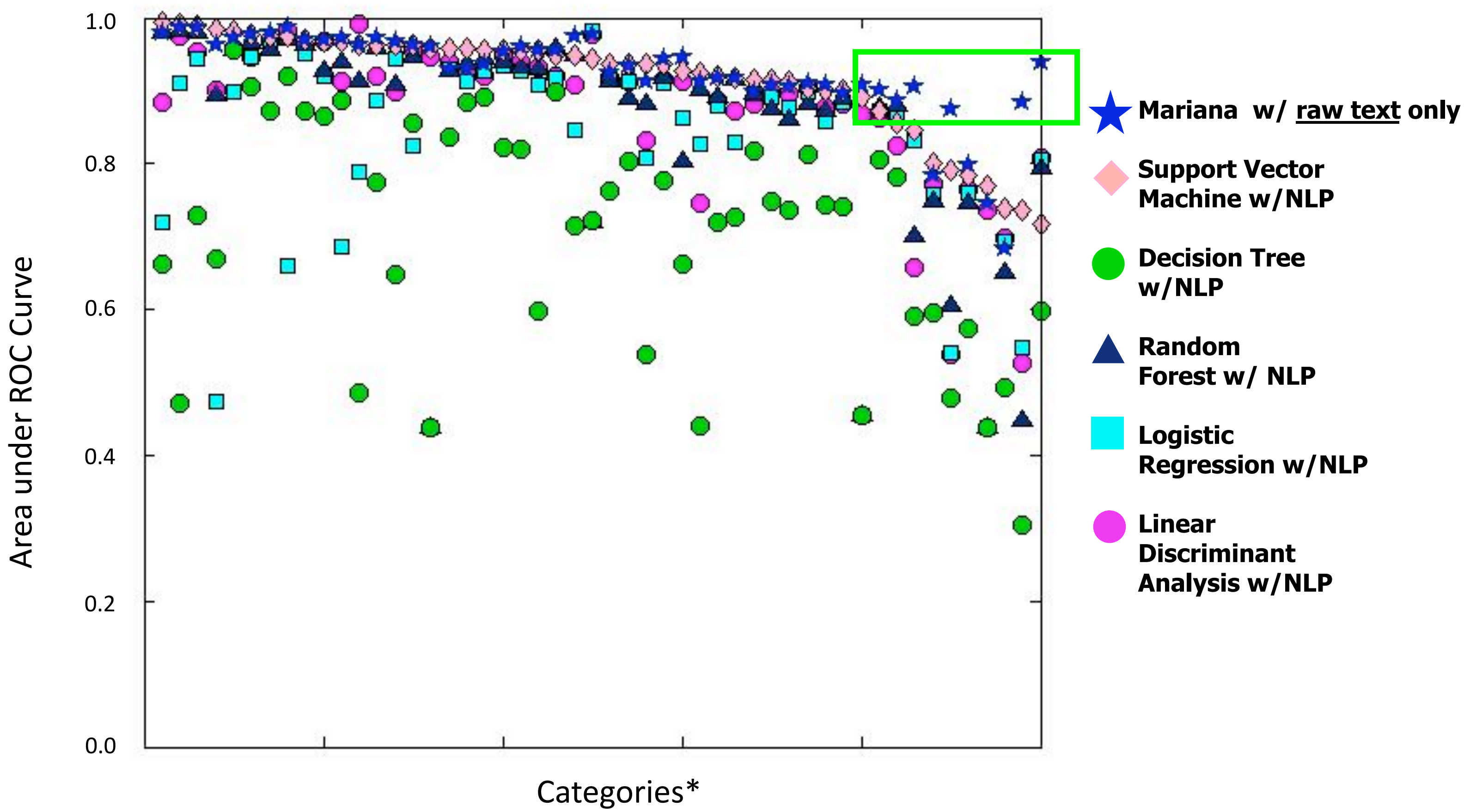
Non Adherence to ATC Clearance
Critical Equipment Problem
Runway Incursion
Landing without a Clearance
Air Space Violation
Altitude Deviation Overshoot
Fumes
Altitude Deviation Undershoot
Ground Encounter, Less Severe
...

Mariana Statistical Model Optimization Flowchart



AUROC - area under receiver operator curve

Mariana vs. Methods using Natural Language Processing



The Mariana algorithm can give substantially better results over other methods (green box), even when processing raw text.

Outline

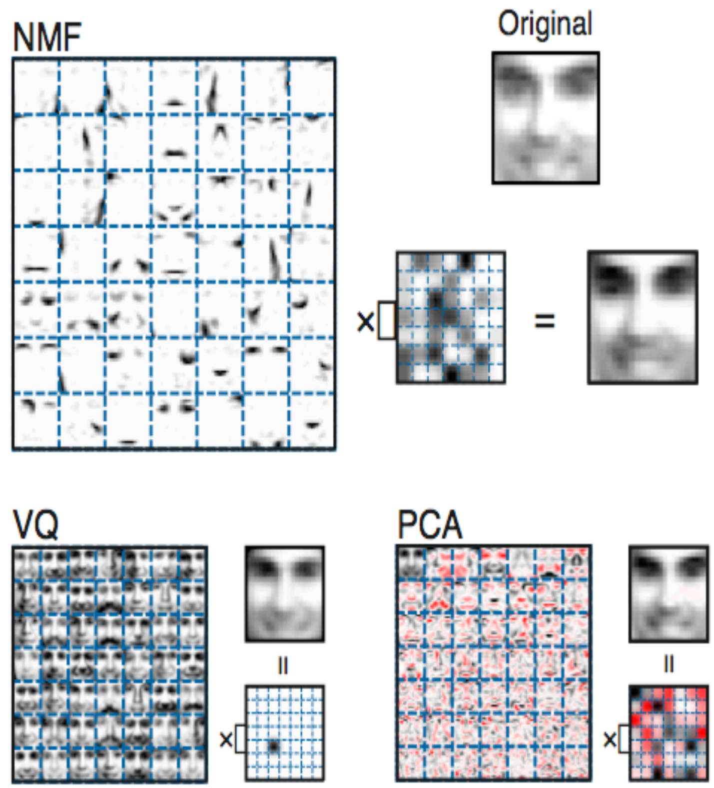
- Combining discrete and continuous sequences: Multiple Kernel Anomaly Detection (MKAD)
  - Derived from anomaly detection methods on discrete and continuous sequences.
- Text Mining: classification, topic modeling
- Ongoing, future work

Subspace Approximation

- Goals
  - System often assumed explainable by simple model plus noise. Construction of simple approximation may reduce noise.
  - Derive small, key set of features to explain system behavior.
  - Lower storage requirements.
- Issue
  - Derived features more intuitive if they are positive, reflecting presence of key components that add up to characterize the item of interest.

Non-negative Matrix Factorization

- NMF finds factors representing parts of faces (e.g., nose, mouth). Easy to interpret.
- Vector Quantization and PCA find holistic representations (face +/- other stuff). Difficult to interpret.



Non-negative Matrix Factorization

Problem: Given a nonnegative matrix $A \in R^{mxn}$ and a positive integer $k < \min\{m,n\}$ find nonnegative matrices to minimize $W \in R^{mxk}, H \in R^{kxn}$

$$f(W,H) = \frac{1}{2}\|A - WH\|_F^2$$

WH is a nonnegative matrix factorization. Columns of W represent k basis vectors captured from n examples having m feature values each. H gives factor-example weights. k is problem-specific and chosen by user.

NMF for classification

- NMF factors can be used for clustering, classification, regression.
- Classification of ASAP reports into one or more contributing factors.

$$f(W_T, H_T) = \frac{1}{2}\|A_T - W_T H_T\|_F^2$$

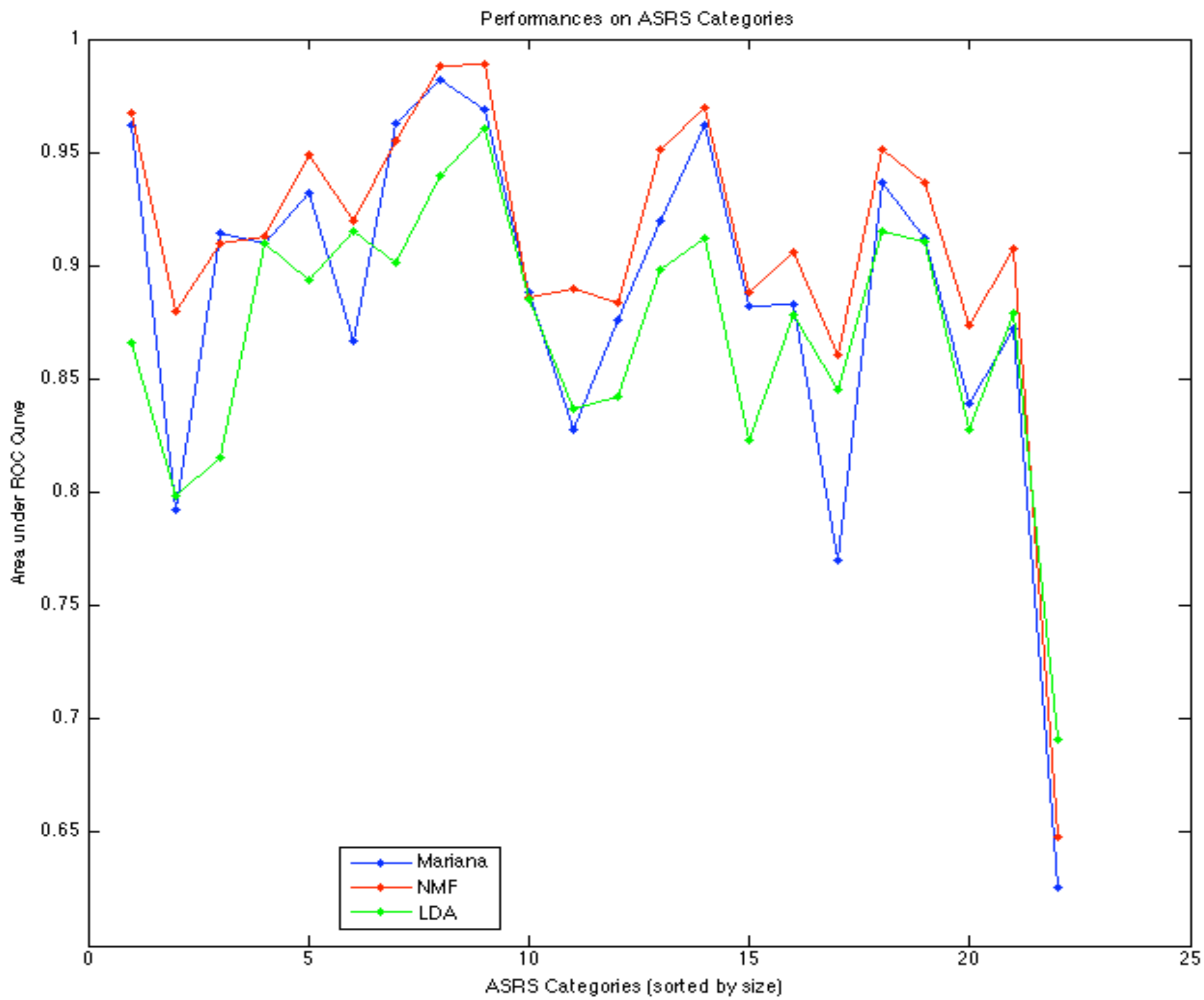
$$f(H_E) = \frac{1}{2}\|A_E - W_T H_E\|_F^2$$

$$C_E \leftarrow H_E^T \backslash (H_T C_T)$$

Sample (ASRS) Basis Vectors

	Basis Vector 1			Basis Vector 2		
Run	1	2	3	1	2	3
	FUEL	FUEL	FUEL	INSTALL	INSTALL	INSTALL
	TANK	TANK	TANK	INSPECT	INSPECT	REMOVE
	POUND	POUND	POUND	REMOVE	REMOVE	REPLACE
	GALLON	GALLON	GALLON	REPLACE	MECHANIC	ENGINEER
	GAUGE	GAUGE	GAUGE	MECHANIC	REPLACE	MANUAL
	PUMP	PUMP	PUMP	FOUND	PART	INSPECT
	FUELTANK	BURN	BURN	WORK	MANUAL	WORK
	BURN	FUELTANK	FUELTANK	MANUAL	WORK	SHIFT
	FUELER	FUELER	FUELER	REPAIR	REPAIR	FOUND
	FUELQUANTITY	FUELQUANTITY	FUELQUANTITY	PART	FOUND	ASSEMBLE
	CENTER	CENTER	CENTER	ENGINEER	SIGN	TECHNICIAN
	MAINTANK	DISPATCH	FUELGAUGE	TEST	ENGINEER	REPORT
	FUELGAUGE	FUELGAUGE	MAINTANK	CHECK	NUMBER	PANEL
	IMBAL	MAINTANK	IMBAL	SHIFT	SHIFT	REPAIR
	REFUEL	IMBAL	REFUEL	SIGN	MAINTAIN	JOB
	CROSSFEED	REFUEL	PLAN	ASSEMBLE	TEST	XYZ
	QUANTITY	QUANTITY	CALCULATE	MAINTAIN	ASSEMBLE	BOLT
	BALANCE	PLAN	CROSSFEED	SERVE	AIRCRAFT	CARD
	CALCULATE	CROSSFEED	BALANCE	CARD	XYZ	LEAK
	EMPTY	CALCULATE	EMPTY	TECHNICIAN	TECHNICIAN	JOBCARD

ASRS Classification Results



Outline

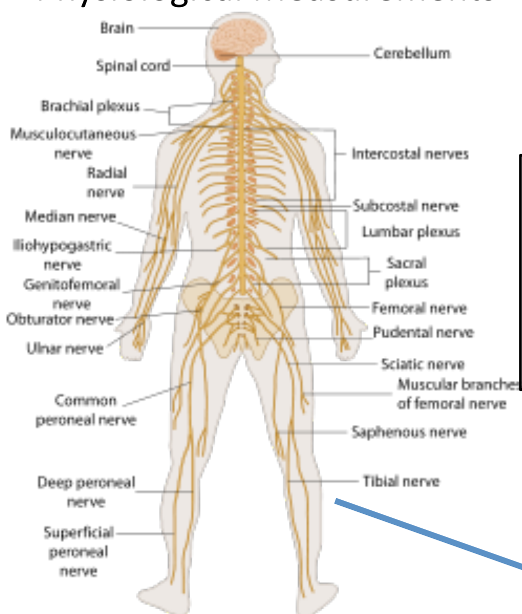
- Combining discrete and continuous sequences: Multiple Kernel Anomaly Detection (MKAD)
  - Derived from anomaly detection methods on discrete and continuous sequences.
- Text Mining: classification, topic modeling
- Ongoing, future work

Ongoing, Future Work

- Anomaly detection over discrete, continuous, and text (utilize when available, don't penalize when not).
- Anomaly cause/precursor identification.
- Prediction over multiple scales: within flight, across flights, across fleets over years.

How Does Human Performance Affect Aviation Safety?

Physiological Measurements*



easyJet.com

Luton, UK

To EasyJet
Near real-time decision support

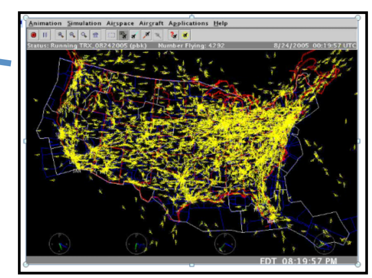
Understanding pilot fatigue

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

Sample Text Report

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

NASA Data Mining Lab (Mountain View, CA)

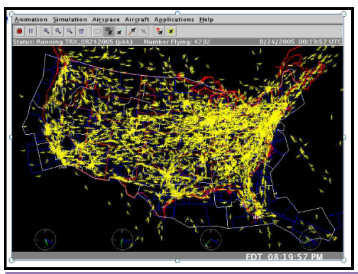


Multiple Kernel Anomaly Detection
(KDD 2010)

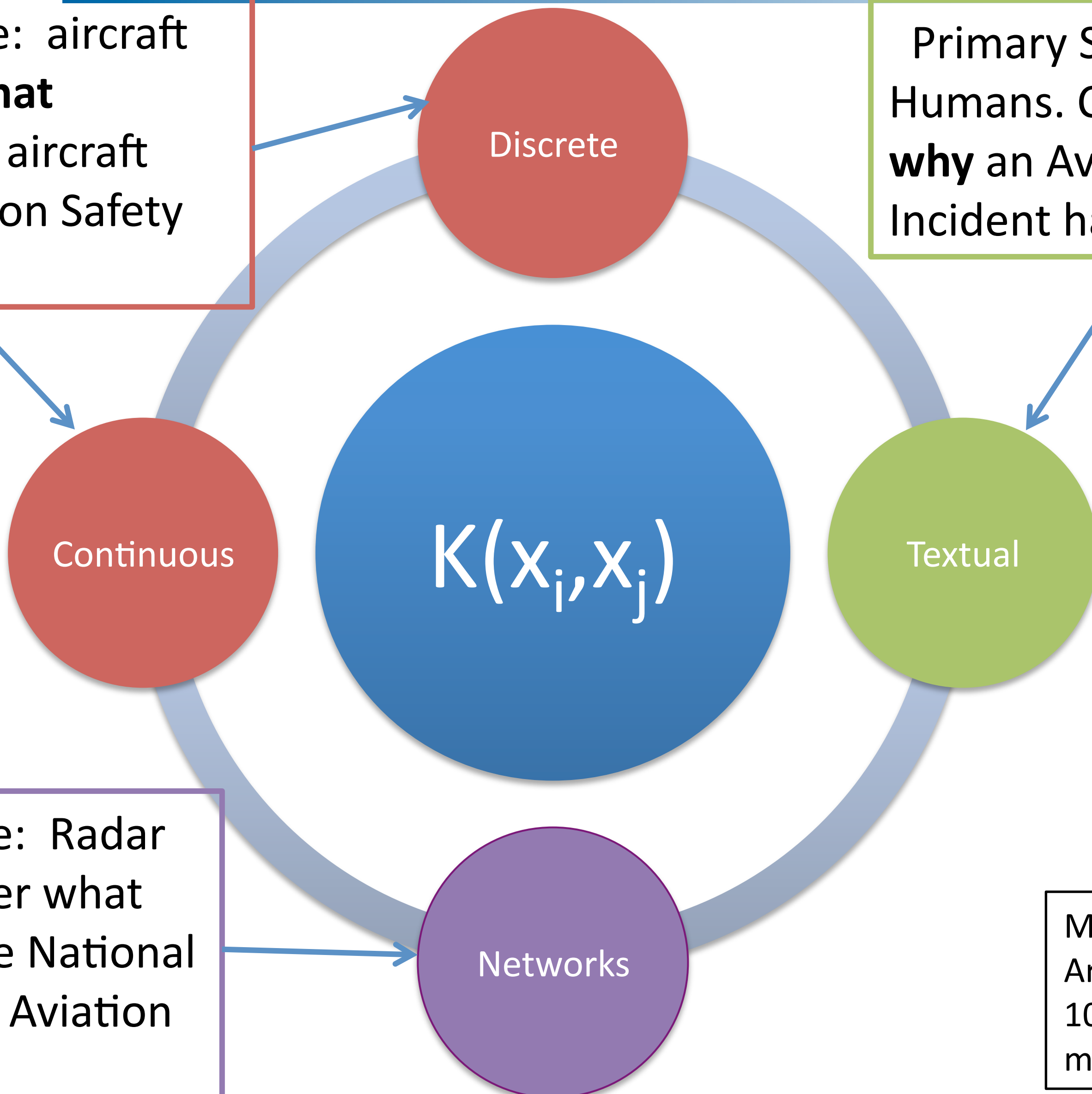
* Diagram for notional purposes only

Mining Heterogeneous Data is the Key

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



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



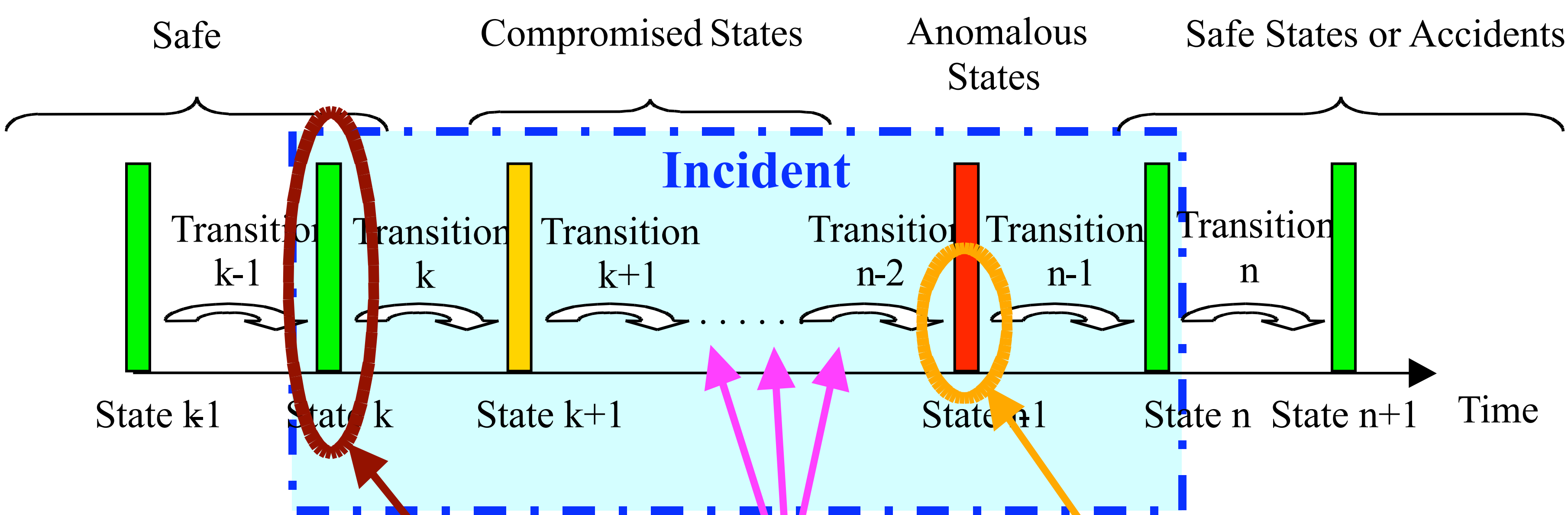
Primary Source: Humans. Can answer **why** an Aviation Safety Incident happened

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.

The Anatomy of an Aviation Safety Incident



Scenario = (Context Behavior Outcome)

From Irving Statler, Aviation Safety Monitoring and Modeling Project

Join DASHlink!

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**.

Conference on Intelligent Data Understanding - 2010

Call for Participation

- Conference focused on theory and applications of data mining and machine learning to Earth Science, Space Science, Engineering Systems
- Location: Computer History Museum, Mountain View, CA
- Date: October 5-6, 2010
- Registration: Free ✔

- Steering Committee
  - Ashok Srivastava (chair)
  - Stephen Boyd
  - Jiawei Han
  - Eamonn Keogh
  - Vipin Kumar
  - Zoran Obradovic
  - Nikunj Oza
  - Raghu Ramakrishnan
  - Ramasamy Uthurusamy
  - Ramasubbu Venkatesh
  - Xindong Wu

Program Chairs: Nitesh Chawla and Philip Yu

References

- S. Budalakoti, A. N. Srivastava, and M. E. Otey, Anomaly detection and diagnosis algorithms for discrete symbol sequences with applications to airline safety, *in IEEE SMC Part C, 39(1), 2008*.
- Santanu Das, Kanishka Bhaduri, Nikunj C. Oza, and Ashok N. Srivastava, nu-Anomica: A Fast Support Vector Based Novelty Detection Technique, *in ICDM 2009*.
- Santanu Das, Bryan L. Matthews, Ashok N. Srivastava, and Nikunj C. Oza, Multiple kernel learning for heterogeneous anomaly detection: algorithm and aviation safety case study, *in KDD 2010*.
- Nikunj C. Oza, J. Patrick Castle, and John Stutz, Classification of Aeronautics System Health and Safety Documents, *in IEEE SMC Part C, 39(6), 2009*.