Aviation Safety Information Analysis and Sharing

Technology and Tools Symposium

Knowledge Discovery from Aviation Data

ASIAS Data Analysis Insights

Paul Melby Ph.D.

July 27-28, 2009

asias

Outline

- The role of data mining in ASIAS
- Data Sources
- Flight Operations Quality Assurance (FOQA) data
- Aviation Safety Action Program (ASAP) reports
- Vulnerability Discovery
- Status

The Role of Data Mining in ASIAS

- **Data mining plays two important roles within the ASIAS program:**
  - **Data mining of structured data may be one of many analysis approaches used for studies of a known safety issue**
    - **Building predictive models, finding associations between the safety issue and contributing factors, etc.**
  - **Text mining is a critical enabler for using incident and accident reports**
    - **With large numbers of reports, manual review is not possible**
    - **Classification, summarization and information extraction are important for dealing with the large volumes of data**
  - **Data and text mining are two of the primary methods for vulnerability discovery**
    - **Vulnerability discovery: finding previously unknown or underappreciated safety risks**

Overview of Data Sources

Data Sources Supporting the Studies

De-Identified FOQA Data

De-Identified ASAP Data

Aviation Safety Reporting System

Safety Reports

- Runway Incursion
- Surface Incident
- Operational Error / Operational Deviation
- Pilot Deviation
- Vehicle or Pedestrian Deviation
- National Transportation Safety Board
- Accident/Incident Data System
- Service Difficulty Reports

ATC Information

- Traffic Management Reroutes and Delays
- Airport Configuration and Operations
- Sector and Route Structure
- Procedures

Surveillance Data

- En route
- Terminal
- Airport

Other Information

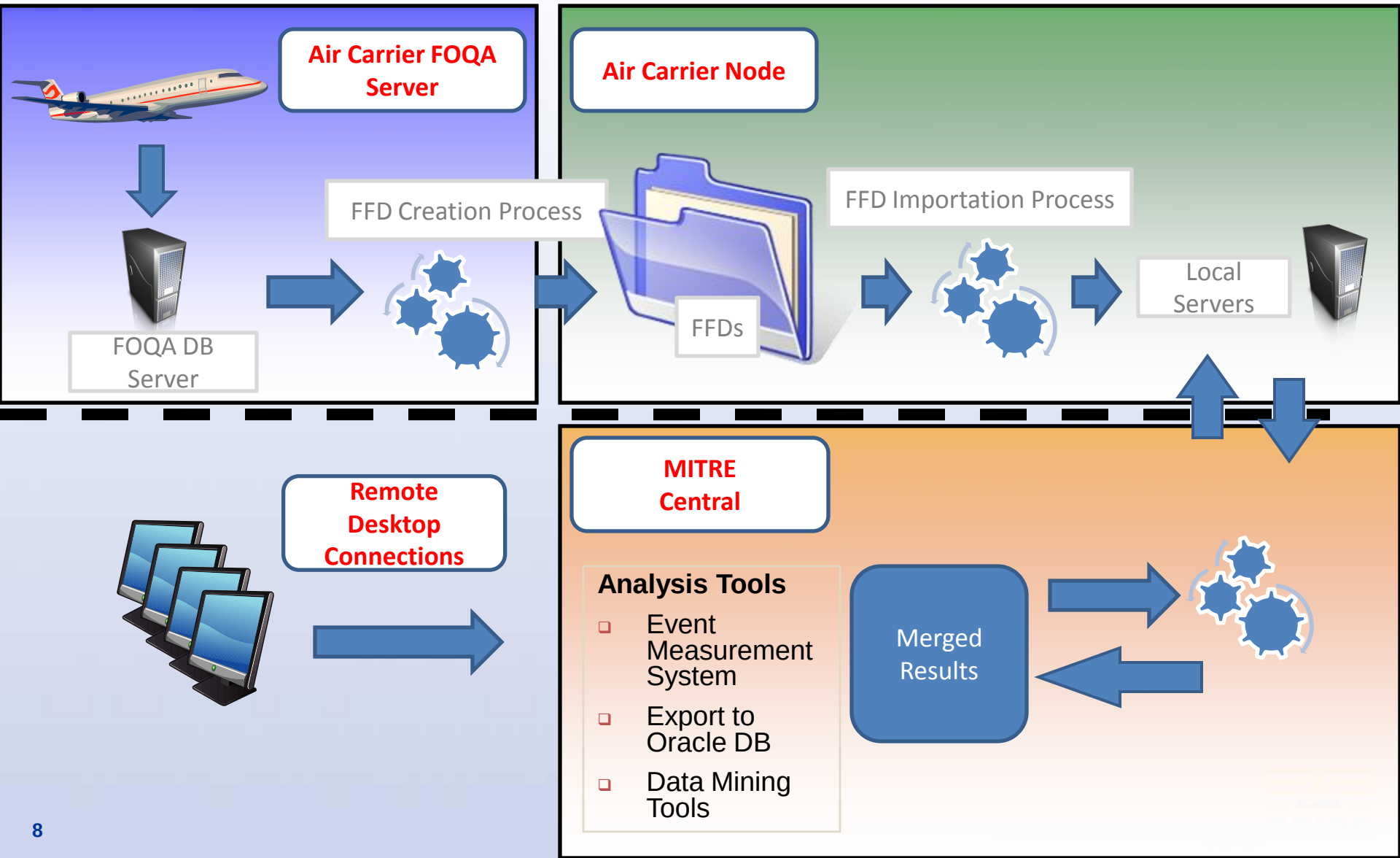
- Bureau of Transportation Statistics
- Weather / Winds
- Manufacturer Data
- Avionics Data
- Worldwide Accident Data

FOQA Data Details

Flight Operations Quality Assurance (FOQA)

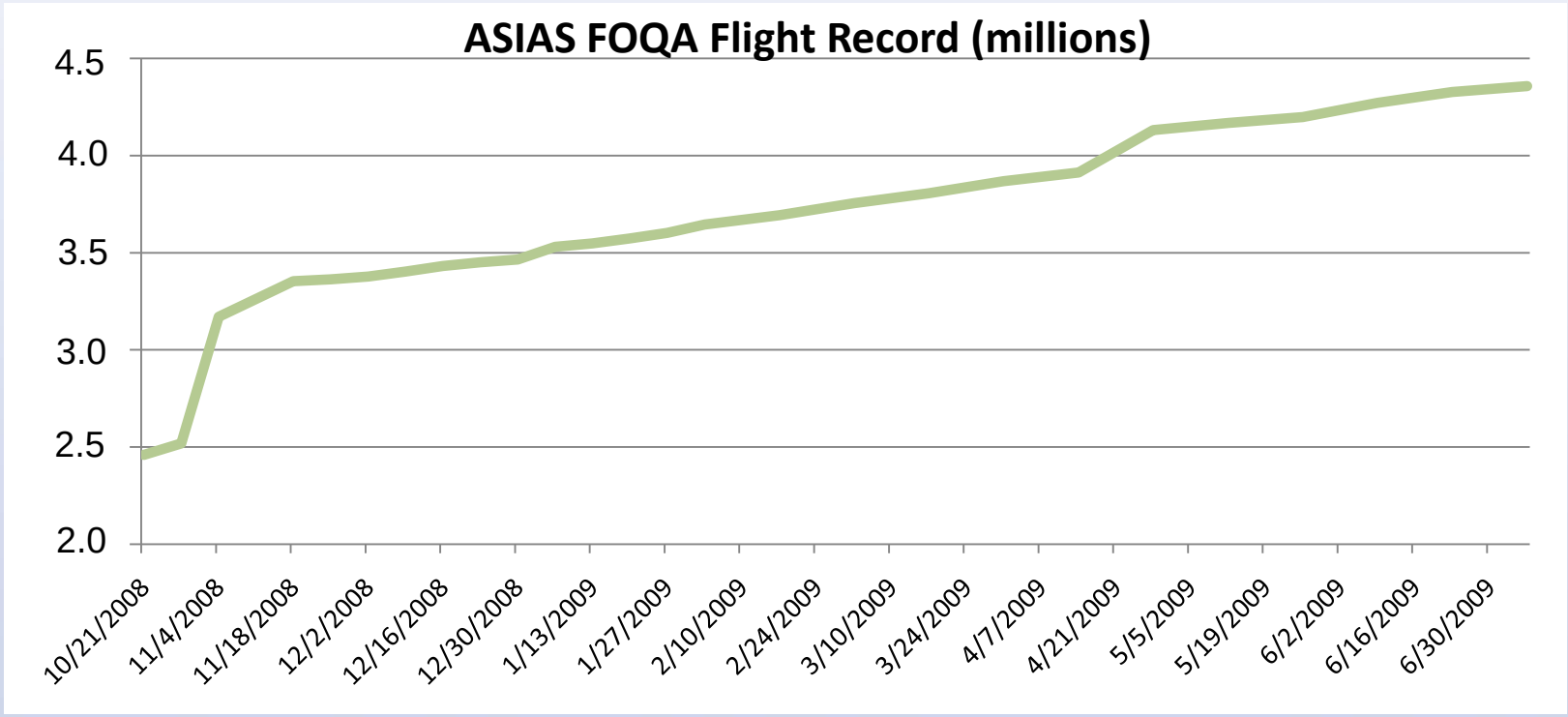
- An FAA approved program for airlines to collect and analyze digital flight data from aircraft
  - Described in FAA Advisory Circular 120-82
- Airlines use FOQA data to monitor fleet for safety events, maintenance, fuel use and other topics of interest
- Airlines with FOQA programs that participate in ASIAS provide MITRE access to a de-identified form of the data for analysis
  - MITRE does not retain local copies of the data

ASIAS FOQA Data/Information Flow



ASIAS Participants

Participating Airlines and Fleets															
A300	**A320**	**A330**	**B717**	**B727**	**B737**	**B747**	**B757**	**B767**	**B777**	**DC8**	**DC9**	**E145**	**E190**	**MD11**	**MD80**
AAL	FFT	NWA	TRS	UPS	AAL	NWA	AAL	AAL	AAL	UPS	NWA	BTA	JBU	UPS	AAL
UPS	JBU	USA			ASA	UAL	COA	COA	COA				USA		DAL
	NWA				COA	UPS	DAL	DAL	DAL						
	UAL				DAL		NWA	UAL	UAL						
	USA				SWA		UAL	UPS							
					TRS		UPS	USA							
					UAL		USA								
					USA										



Architecture

- **Over 4 million flight records**
- **Data resides at airline nodes, maintained on private network**
- **There are multiple systems available for analyzing FOQA data within ASIAS:**
  - **Commercial FOQA Analysis tools used by airlines**
    - **Contains FFD data plus many derived and calculated parameters**
    - **Proprietary query interface**
  - **Distributed Oracle Database**
    - **Contains FFD data, derived data from COTS tools, analysis results**
    - **Allows for *ad hoc* queries**

FFD Data

- **Key Features:**
  - Data recorded during entire flight, from gate to gate
  - Measurements 64 times per second up to one measurement every 64 seconds
  - Over 400 parameters measured (varies by airframe):
    - Continuous parameters
      - » Latitude, Longitude, Altitude, Airspeed, Bank Angle, Engine Power, etc
    - Discrete parameters
      - » Flap handle position, autopilot on/off, landing gear selected up/down, etc.
    - Demographic data (only 1 value per flight)
      - » Arrival/Departure airports, make/model, fleet, de-identified tail number, etc.

- **Database:**
  - FFD data is loaded into Oracle database at each airline node
    - Data is sampled at 1 second intervals
  - Data is accessed through a distributed query

FOQA Analysis System

- **Derived and corrected parameters**
- **Over 1000 operationally relevant measurements on every flight**
  - **New measurements can be specified and calculated by the system**
- **Measurements of safety events:**
  - **Many built in safety events, such as stall warnings, ground proximity warnings**
  - **The growing collection of events from ASIAS studies**

Source: Austin Digital, Inc

FOQA Data Complexity Challenges

- High Dimensional (over 400 parameters)
- Sequential
- Combination of continuous and discrete parameters
- Complexity of real operations
  - Phases of flight
  - Variations in duration of flight (30 minute cruise versus 6+ hours)
  - Variations in airport layout and procedure design
  - Different routes, impacted by weather and traffic
  - Differences in aircraft types
- Lack of context
  - No information on traffic, limited, possibly erroneous weather information

FOQA Data Quality Challenges

- **Variation between aircraft types**
  - Different aircraft types have different sensors and may measure parameters in slightly different way or not at all
  - Differences between manufacturers are very large
    - Some differences due to equipment on aircraft
    - Some differences due to differing designs
- **Variation between FOQA programs**
  - Different FOQA vendors provide different data quality processes
  - Variations in the translation to FFD files
  - Multiple flights in a single FFD file
- **Bad sensors**
- **Inaccurate measurements**
  - e.g., Lat/Lon measurements can be significantly off for aircraft without GPS

FOQA Analysis: The Vision

- **FOQA analysis tools, methods and architecture that make it easy for the analyst to:**
  - **Find all relevant flights on a topic of interest and summarize what they contain**
  - **Group flights by operationally significant groups: airport, runway, waypoint, procedure, etc.**
  - **Understand links between safety events and/or contributing factors**
  - **Find the flights that exhibit the highest risk events**
  - **Find flights and groups of flights that discuss previously uncharacterized events or contributing factors**
  - **Allow for interactive exploration of the flights to find very rare safety issues of interest**

FOQA Tools and Current Status

- **The following tools are available and currently being applied or evaluated:**
  - **MITRE Developed Tools (internal R&D project)**
  - **NASA Developed Tools (presentation by Srivastava)**
  - **Commercial FOQA analysis Tools:**
    - **EMS system from Austin Digital Inc.**
    - **Automated Ground System (AGS) from SAGEM**
  - **Oracle**
    - **11g Enterprise Edition with Data Mining and OLAP**
  - **PASW Modeler and PASW Statistics (formerly SPSS Clementine and SPSS Statistics)**
    - **COTS data mining and statistics software from SPSS Inc**
  - **Other COTS products:**
    - **Tableau, Matlab, JMP, many others**
- **Most work so far has been analyzing known safety risks and setting up the infrastructure to enable data mining**
  - **Development of safety benchmarks**
  - **Development of Oracle database**
  - **Export of EMS measurements to Oracle database**
  - **Visualization capabilities**

Aviation Safety Action Program (ASAP) Reports

Aviation Safety Action Program (ASAP)

- **An FAA approved program for airlines to collect voluntary incident reports from different groups of personnel:**
  - **Flight Ops (pilot), Maintenance, Dispatcher, Flight Attendant, Ramp/Load Planner**
- **Airlines use ASAP data to monitor safety concerns**
  - **Most reported incidents would be unknown to the airline without an ASAP report**
- **Airlines with ASAP programs that participate in ASIAS provide MITRE access to a de-identified form of the data for analysis**
  - **MITRE does not retain local copies of the data**

ASAP Overview: Collection, Review and Archiving of Incident Data

ASAP Process

1. Pilot Report Submission

10 - 400 reports per month per Airline

→

2. Reviewed for acceptance by ASAP ERC committee

Corrective actions determined

→

3. Reports closed and archived for further analyses

de-identified, Analyzed

→

4. Reports made available to ASIAS

asias

ASIAS ASAP Event Report Count as of June 2009

asias

ASIAS ASAP Taxonomy

- **Airline taxonomies mapped to common taxonomy:**
  - **Demographic:** location, origin and destination airports, phase of flight of event, aircraft type, etc.
  - **Event Type:** the type of incident is being reported such as a runway incursion or altitude deviation
    - 29 Primary event types, 246 secondary types
  - **Internal Factors:** contributing factors originating within the cockpit, such as improper use of automation or flight crew coordination
    - 14 Primary internal factors, 88 secondary
  - **External Factors:** contributing factors originating outside the cockpit, such as turbulence or equipment malfunction
    - 13 Primary External Factors, 82 secondary
  - **Narrative:** includes flight crew's description of the incident, a summary of the event, and flight crew's suggestions of preventative measures that would have avoided the event

ASAP Data Characteristics

- **Reports are de-identified by airline**
  - Pilot names, flight id and tail number removed
  - Date set to 1st of the month
- **All reports are retained**
- **No processing of narrative except for de-identification**
- **Merged report: reports from individual flight crew members are combined into a single event report**
- **Categorization:**
  - Multiple categories: each report can list multiple event types and contributing factors
  - Unbalanced: most categories are true for only 1-5% of the reports and are frequently true much for less than 1% of the reports

ASAP Data Quality Challenges

- **The quality of ASAP reports vary significantly due to several factors:**
  - **Categorization by flight crew and ASAP Analyst:**
    - **Depending on the data collection process and review process at an airline, the quality of the structured fields can be very low**
  - **Self-reported data contains errors and bias**
  - **Fields like phase of flight may have fuzzy definitions or boundaries and events may occur across multiple phases**
  - **Airline to ASIAS ASAP Taxonomy mapping:**
    - **Imperfect mapping**
    - **Missing/extra categories**
  - **The narratives are full of jargon, abbreviations, misspellings and other grammatical errors**

Text Analysis: The Vision

- **Text analysis tools, methods and architecture that make it easy for the analyst to:**
  - **Find all relevant reports on a topic of interest and summarize what they contain**
  - **Group data by operationally significant groups: airport, runway, waypoint, procedure, etc.**
  - **Understand links between event types and/or contributing factors**
  - **Find the reports that describe the highest risk events**
  - **Find reports and groups of reports that discuss events or factors that are not in the current taxonomy**
  - **Allow for interactive exploration of the reports to find very rare safety issues of interest**

Text Mining Challenges Areas

- **Search/Classification:**
  - In ASAP, the structured fields for event types and contributing factors are not well populated
  - Searches based only on structured fields will miss a lot of reports
  - Keyword searches result in large numbers of false positives
- **Review Process:**
  - How do we capture all of the relevant information from a SME that reviews a report?
  - How can we minimize the time needed for SME review?
- **Summarization:**
  - Given thousands of reports of a particular event type, what's the fastest way to summarize them?
  - What concepts are linked together that we didn't realize were linked?
- **Information Extraction:**
  - What airports, waypoints, procedures, etc. are mentioned in the data set?

Text Analysis Tools and Current Status

- **A number of text analysis tools are being developed and/or available:**
  - **ASIAS Report Reviewer**
    - SME review and classification tool
  - **ASIAS Regular Expression Generator**
    - Iterative search method that generates regular expressions which can be imported into the ASIAS report reviewer or used as features for other text mining methods
  - **NASA's Mariana:**
    - Optimized Support Vector Machine (SVM) classifier
  - **PASW Text Analytics**
    - COTS text mining tool from SPSS Software (formerly Text Mining for SPSS Clementine)
  - **Other MITRE tools:**
    - Aviation Safety Workbench: includes MITRE patented similarity matching algorithm
    - MITRE developed annotation and information extraction tools (presentation by Doran on Tuesday)
    - Latent Dirichlet Allocation based SVM classifier (presentation by Booker on Tuesday)
- **The initial focus has been to streamline the process of creating validated sample data**
  - This is the key bottleneck in the development of classifiers
  - Poorly classified data is one of the key bottlenecks in performing other analysis, such as tracking reporting rates and linking events and contributing factors

Vulnerability Discovery

What is Vulnerability Discovery?

- **Some examples of vulnerability discovery:**
  - **Discovering previously unknown or underappreciated links between types of safety events, contributing factors**
  - **Raising awareness of little known event types or contributing factors**
  - **Discovering new contributing factors to known event types**
  - **Discovering new safety event types**

Approaches to Vulnerability Discovery

- **Analyst/Subject Matter Expert (SME) Driven (Manual)**
  - Focus on hot spots or negative trends
  - Input and feedback from stakeholders

- **Data Driven (Automated) : focus of this symposium**
  - Statistics
    - Correlations between contributing factors and event types
    - Characterization of typical
  - Data Mining
    - Characterization of typical
    - Pattern identification: typical and atypical
    - Predictive and classification rules, event associations
  - Text Mining
    - Looking for correlations between contributing factors and event types in narrative data
    - Looking for reports that would be categorized as “high risk”

The Role of Risk

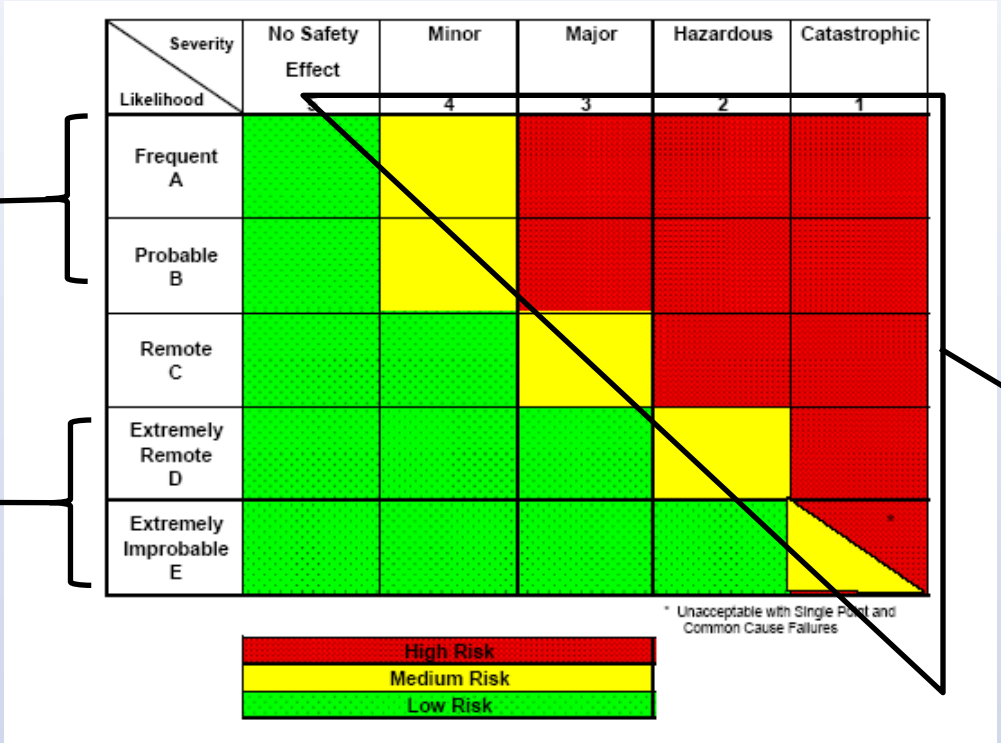
Finding anomalies is not enough – the goal is to find high risk events

- **Risk combines severity and likelihood**

These would be characterized as “frequent” or “normal” behavior, which is not always safe

Anomaly detection focuses here, but not all anomalies are unsafe or high risk

Risk Assessment Matrix



These high risk factors are the most important for safety

Source:
FAA SMS Manual v1.1

Technical Challenges

- **Heterogeneous databases: FOQA, ASAP, ASRS, SDR, NOP, Advisory, METAR, Airline Flight Ops, many others.**
- **High volume, dimensionality of data**
- **Data Quality concerns**
- **ASIAS isolated computer network**
- **De-identification requirements**
- **Complexity of flight operations**

Long term Vision

- **Tools and automation of:**
  - **Data fusion**
  - **Anomaly detection based on established patterns**
  - **Data characteristics trend tracking**
  - **Risk assessment**
  - **Visualization of data**

Summary

- **Large, challenging and important problem area**
- **Lots of rich and complex data sources**
  - **Applications for data mining, text mining, visualization, data fusion**
- **Big issues to address:**
  - **Data quality, dimensionality, sequence based, data fusion**
  - **Assessment of risk to find ‘interesting’ results**
  - **Implementation, deployment and integration of tools**

Approved for Public Release; Distribution Unlimited. Case Number: 09-09-2907

The contents of this material reflect the views of the author and/or the Director of the Center for Advanced Aviation Systems Development. Neither the Federal Aviation Administration nor the Department of Transportation makes any warranty or guarantee, or promise, expressed or implied, concerning the content or accuracy of the views expressed herein