

Data Science



and the Knowledge Discovery Adventure

Bryan Matthews (NASA/KBR Inc.)

June 28th 2022



Overview of the Knowledge Discovery Process



- Initial fact gathering (interview domain experts, review reports, articles, state-of-the-art)
- Identify the problem type (prediction, classification, statistical analysis, etc.)
- Survey supporting data sources
- Understand the data (numerical, categorical, text, sampling rate, data quality issues, etc.)
- Select relevant features and sources
- Acquire the data (set up agreements with the data stewards, APIs to download, etc.)
- Merge data sources (temporal, spatial, common key, other ontologies...)
- Feature Engineering (non linear domain knowledge or physics-based relationships)
- Build data processing pipeline (may need to tap into data stream, develop parallel processing algorithm, federated learning etc.)
- Build model and test (tune hyper-parameters, cross validation.)
- Analyze/Validate results (do the results make sense. Does it answer the original question).
- Deploy/Publish (Monitor and assess benefits)

- Interview domain experts
 - Learn how the domain currently operates.
 - Is there automation, what role do humans have?
- Review reports
 - Are their logs, complaints/errors, text reports that support the need to address the deficiency in the system?
- Review State-of-the-art research
 - Has anyone addressed this or part of this problem and how?
 - Have similar problems been solved in other domains?
- Is the data annotated or labeled?
 - One of the most significant Data Science challenges.

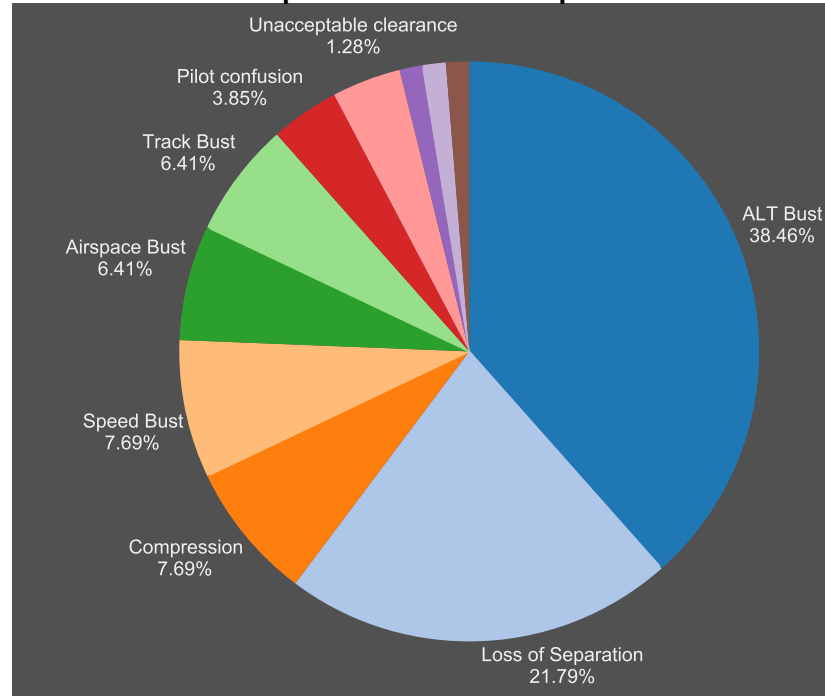
- Obtaining Labels:
 - Supervised approaches provide superior performance over unsupervised
 - However, require extensive labels.
 - SME time is a limited commodity
 - How to intelligently select examples for SME labels that maximize ROI?



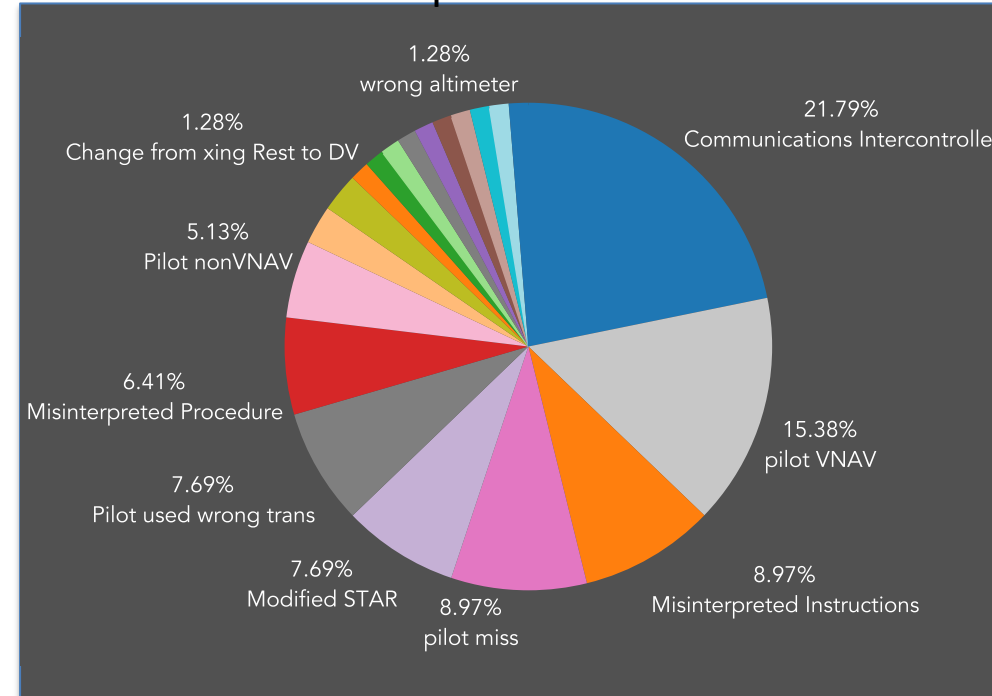
Initial fact gathering: Case Study



Reported Consequences



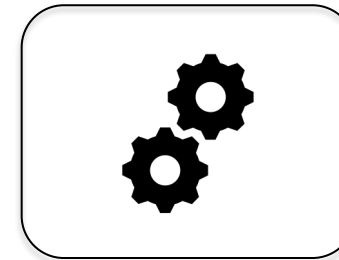
Reported Causes



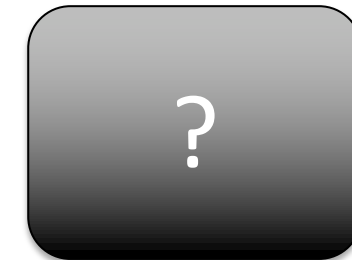
- Survey user reports :
 - Controllers report in free-text format safety issues in Aviation Safety Reporting System.
 - Highlight causes and consequences to safety issues.
 - Use this to identify adverse events and areas of non-adherence
- Identify deficiency in the system
 - Current state-of-the-art does not measure full adherence to procedures
 - Procedural/Environmental/Human factors have not been explored to explain factors of non-adherence.

- What are the shortcomings that need to be addressed
 - What is well understood and what is not?
 - What do we need to learn?
 - How can data mining/machine learning help achieve this?
- What sort of problem is being addressed?
 - Supervised/Unsupervised/Semi-Supervised
 - Prediction
 - Event/category or scalar value
 - Classification
 - Binary/multi-class
 - Statistical Analysis
 - Identifying trends
 - Summary statistics
 - Choosing White/Black Box Method

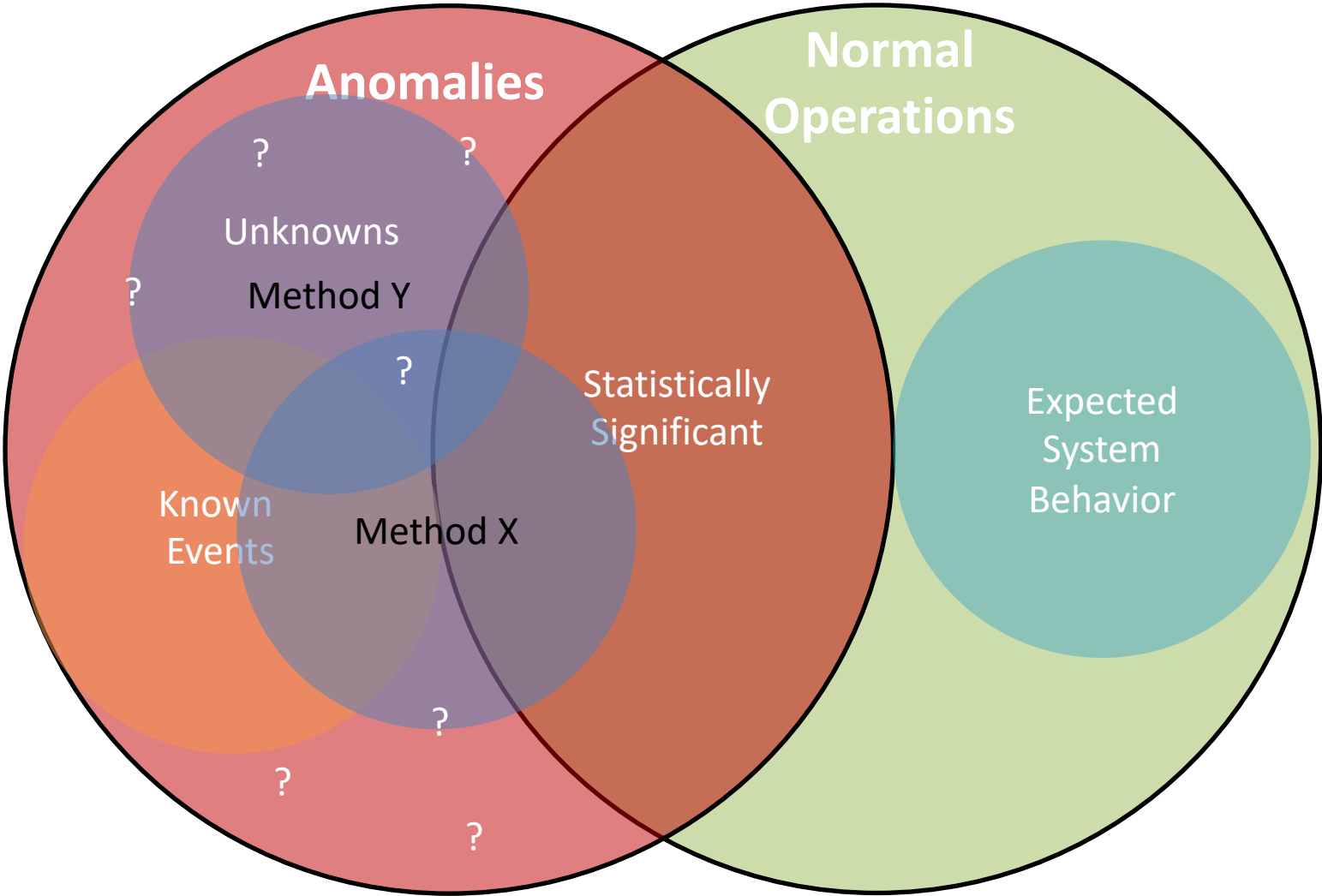
Model: A



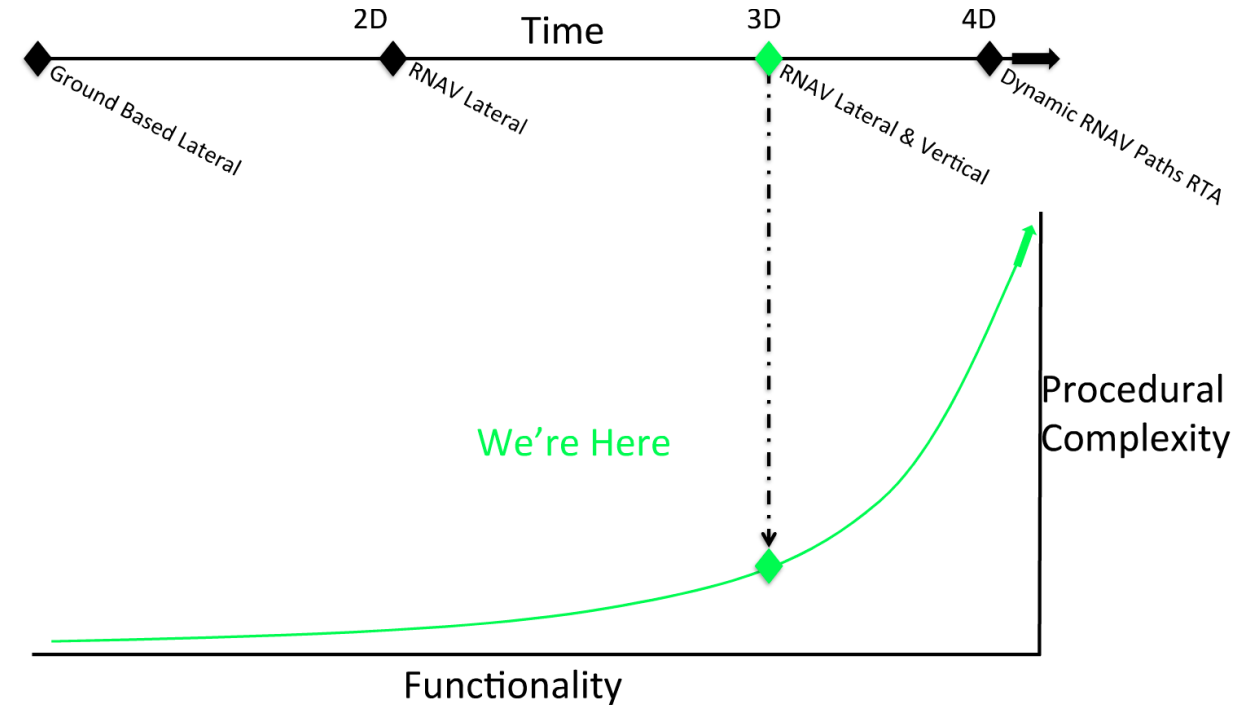
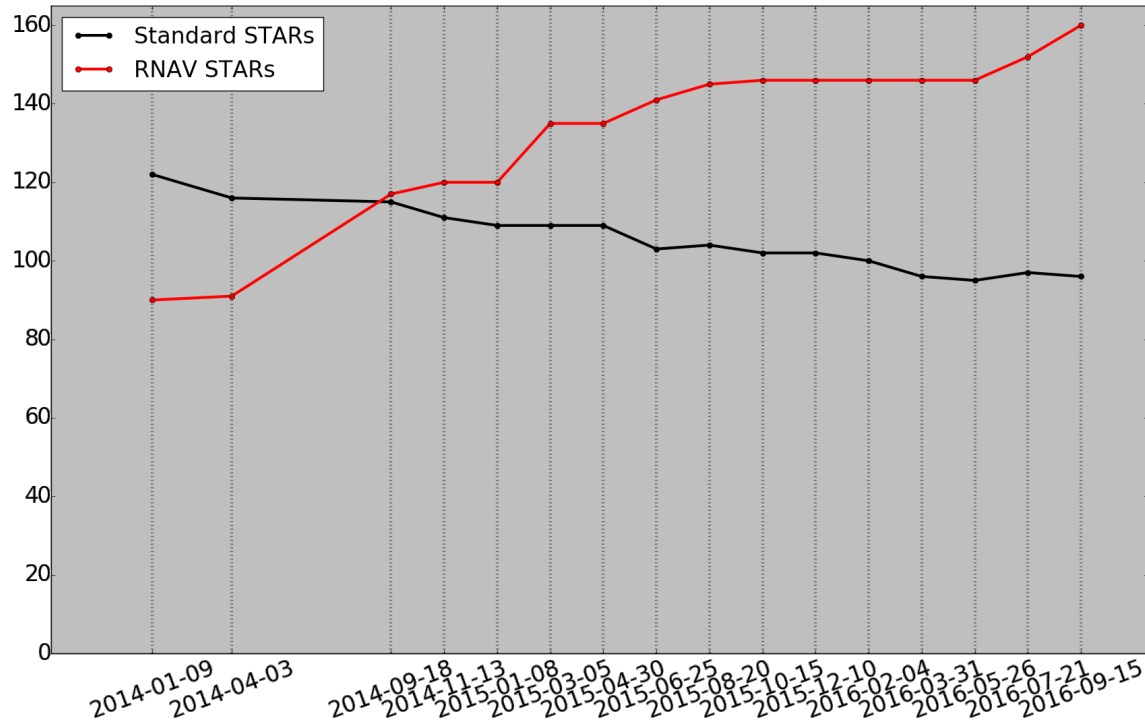
Model: B



Understanding Operational Significant Vs Statistically Significant



Identify Problem: Case Study



- What is the impact of the increase in more complex operations?
 - How often are these utilized? (Statistical assessment)
 - Does this help or hinder efficiency/safety? (Statistical assessment)
 - If not fully utilized what factors affect usability? (Predictability of event type)
 - White box algorithm needed for interpretability
 - Classification problem to identify adverse events (non-adherence)

- What sources of data are available?
 - Traditional data sources numerical, text
 - Non-traditional data sources (social media/crowd sourcing)
 - What is the reliability of these data sources
- What is the scale of data (Mb/Gb/Tb/Pb/...)
- Who owns the data?
 - Publicly available
 - Does an agreement need to be set up to acquire (programmatic/legal)?
 - Is the data sensitive PII, SBU, Export Controlled, Classified etc.
 - What layer of protection does it need?
 - Encryption
 - Secure server
 - Limited user access

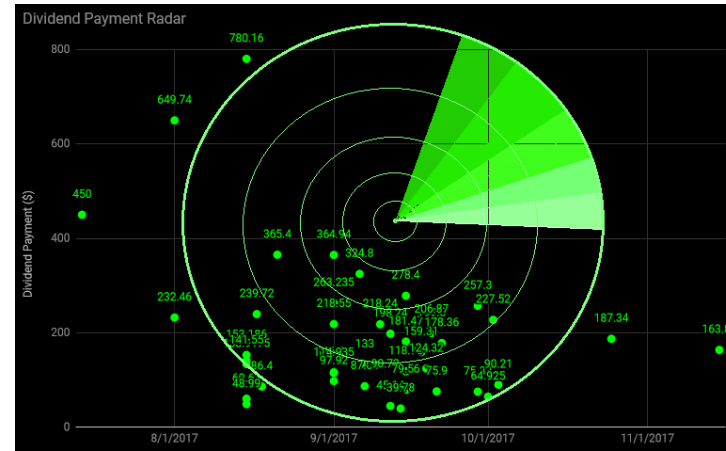
- Data Quantity and Quality:
 - Small Data: Unusual behavior occurrence rates are rare.
 - Big Data: How to drill down and identify operationally significant events.
 - Sensor/Corrupted Data:
 - Anomaly detection is sensitive to these issues.
 - How is Reliability/Robustness impacted?



Credit: <https://xkcd.com/1838/>

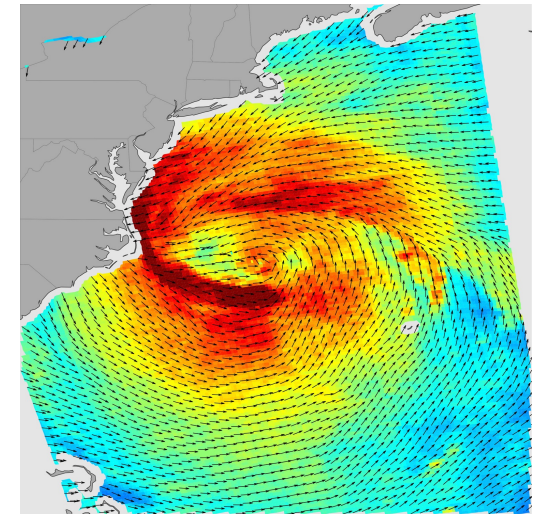
Survey Supporting Data Sources: Case Study

- How do we measure the flight paths?

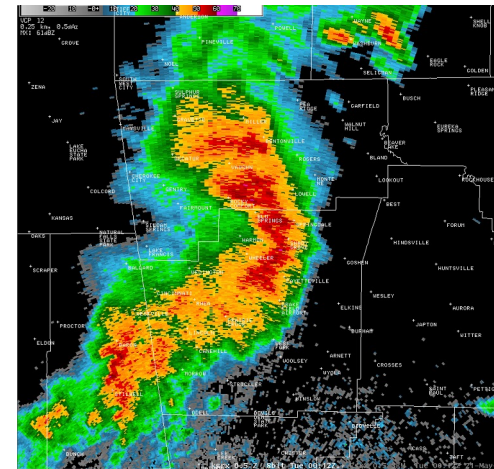


Flight Radar Track Points

- What impact do tail winds have on the operations?

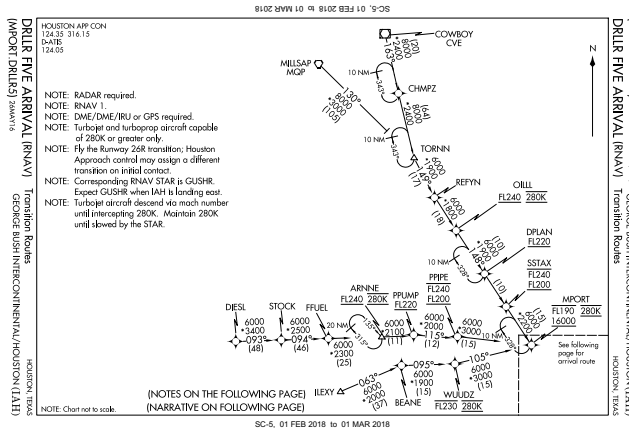


Wind Vectors



Convective Weather

- What type of weather impacts the operations?



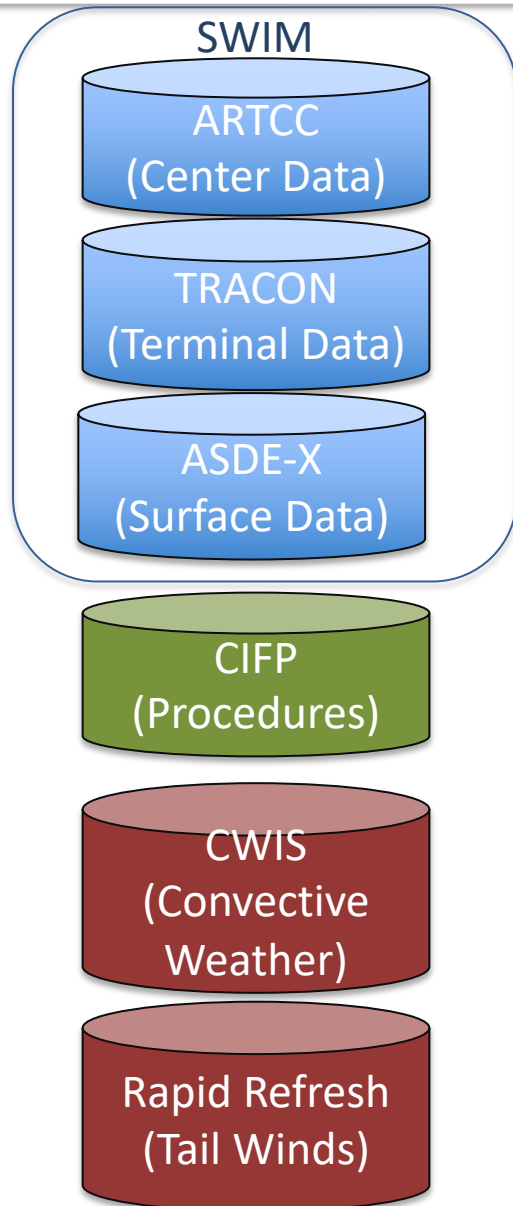
Procedural Data

Data Features are Like Ingredients to an Algorithmic Recipe

- Domain experts can help identify relevant features
- Identify sources that are reliable
 - Key features may be unreliable and unusable unless handled properly
- Identify redundant features
 - Feature pruning



Selecting Relevant Features and Sources: Case Study

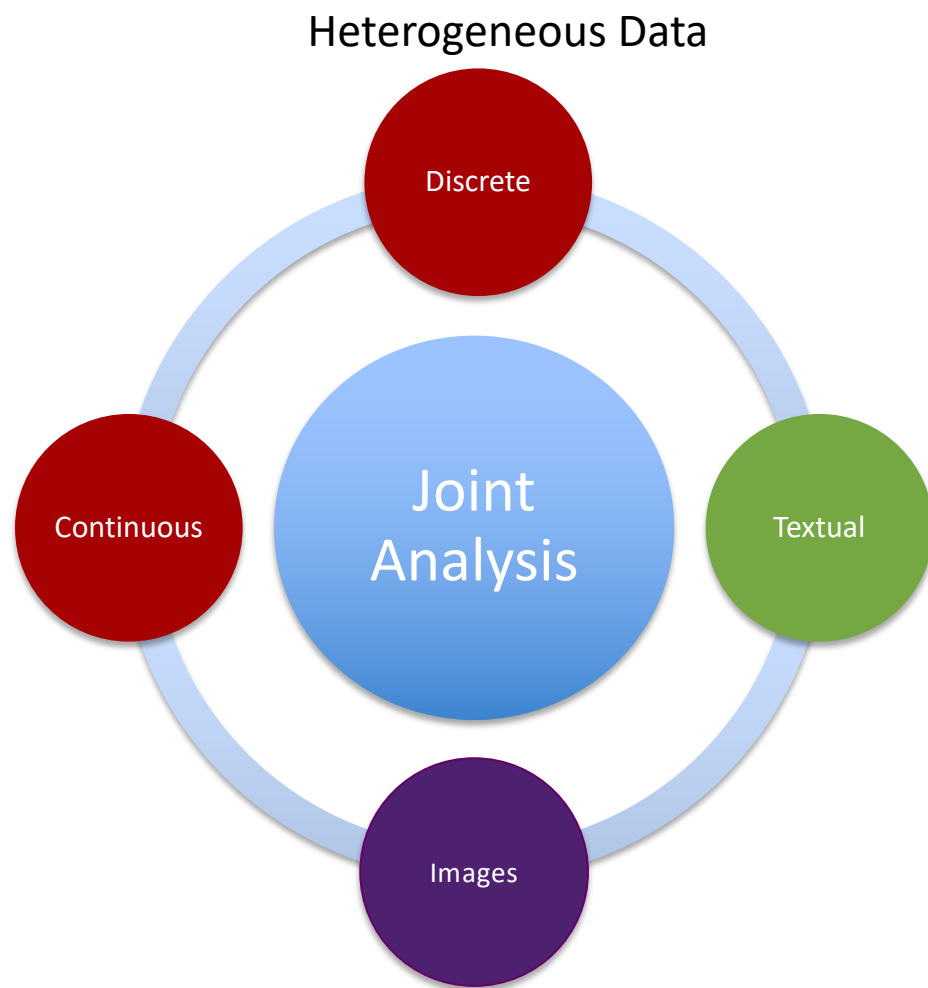


- What records the flight track positions?
 - FAA SWIM network provides access to historical radar track data feeds (lat/lon/alt/speed) from air traffic control facilities.
- How are the procedures defined?
 - Coded Instrument Flight Procedures defines routes with waypoint coordinates with altitude/speed restrictions.
- Is there data that indicates severe weather?
 - MIT Lincoln Labs produces the Corridor Integrated Weather System (CWIS), which defines severe convective weather.
- Can we characterize the tail winds?
 - Rapid Refresh (NOAA product) defines a grid estimate of winds aloft.

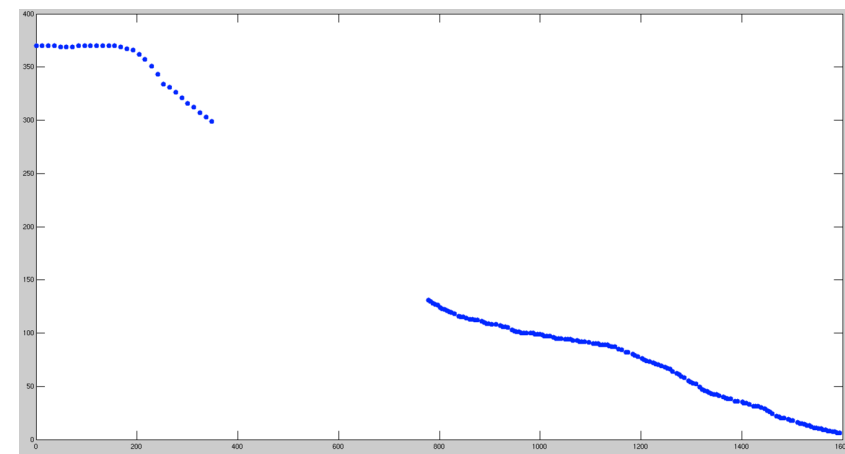
- Review initial sample data
 - Identify known data quality issues (veracity) and mitigation approaches
 - Gaps in the data
 - Sampling rate inconsistencies
 - Corrupted fields
 - If possible resolve issues at source
 - Assess features:
 - Numerical (continuous/binary)
 - Categorical
 - Text



Understand the data: Case Study



Missing/Gaps in Data



Corrupted Data



- Initial data dump
 - Shipping/hand delivery external hard drive
 - File transfer over network
 - Data security encryption keys etc.
- Ongoing acquisition
 - API needed to tap into streaming data
 - Automatic SFTP transfer
 - Scripts to verify and ensure data are transferred successfully



- Identify ways to merge the data (Data Ontology):
 - Temporal
 - Spatial
 - Individual's ID
 - Common key, other.
- How to synchronize the data
 - High/low sample rates
 - Sample hold
 - Interpolate
- Across physical network boundaries
 - Sensitive data vs public
 - VPN/Firewall network technical issues

Unix time	Latitude	Longitude	Flight ID
1461976876	40.63463	-73.78158	UAL123
1461978073	40.63472	-73.78175	UAL123
1461979526	40.63482	-73.78196	UAL123
1461979533	40.63491	-73.78216	UAL123
...	...	...	..

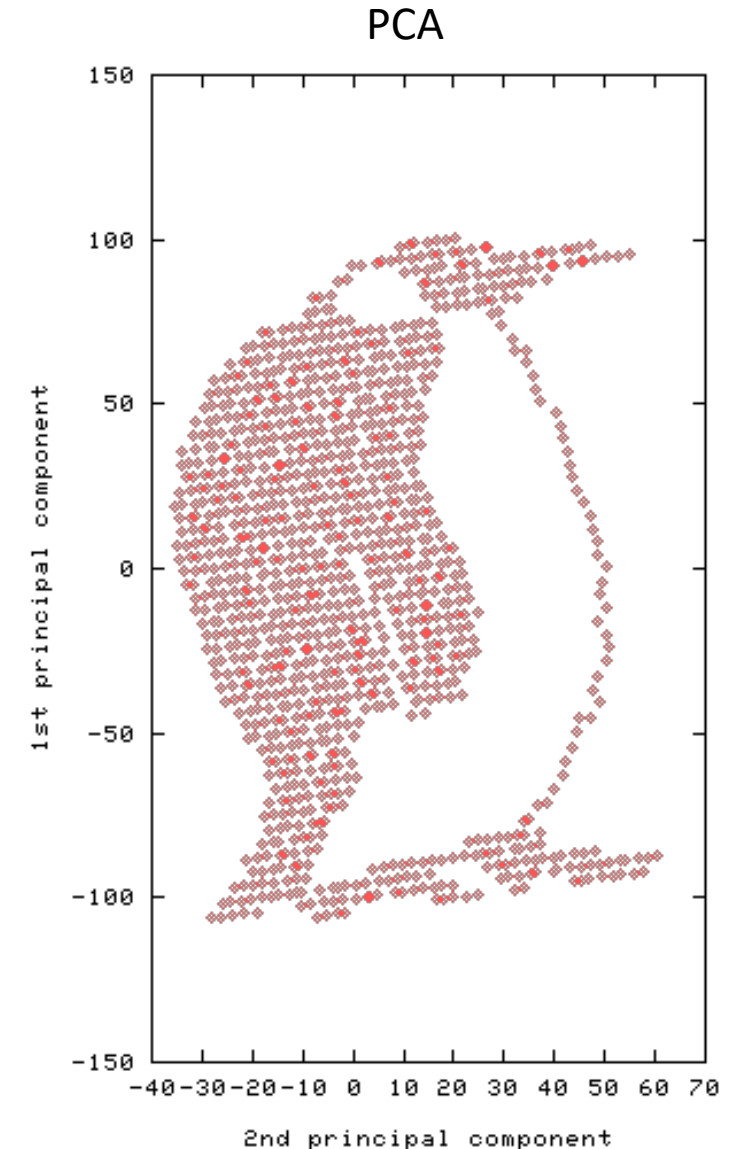
CWAM DP % 0 10 20 30 40 50 60 70 80 90 100

Jul 14, 2006 18:30

0 min @ 32000 ft

[illegible][illegible]

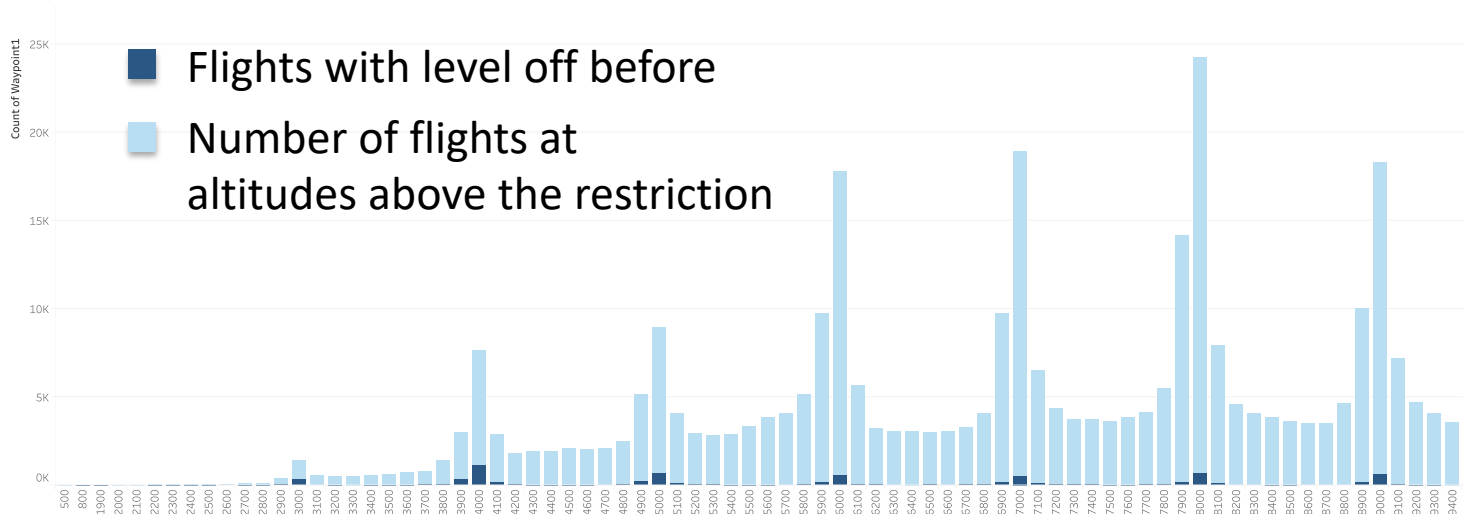
- Non-linear relationships
 - Physics based models to derive relevant parameters
 - Frequency based features
 - Compute deviations from distributions (statistical features)
- Dimensionality reduction
 - Principle Component Analysis (1901)
 - Independent Component Analysis (1980's)
 - Non-negative Matrix Factorization (1990's)
 - T-distributed Stochastic Neighbor Embedding (2008)
- Signal processing techniques
 - Isolate signal
 - Low-pass/high-pass/band-pass filters
- Filter with meta data



Derive New Features: Case Study

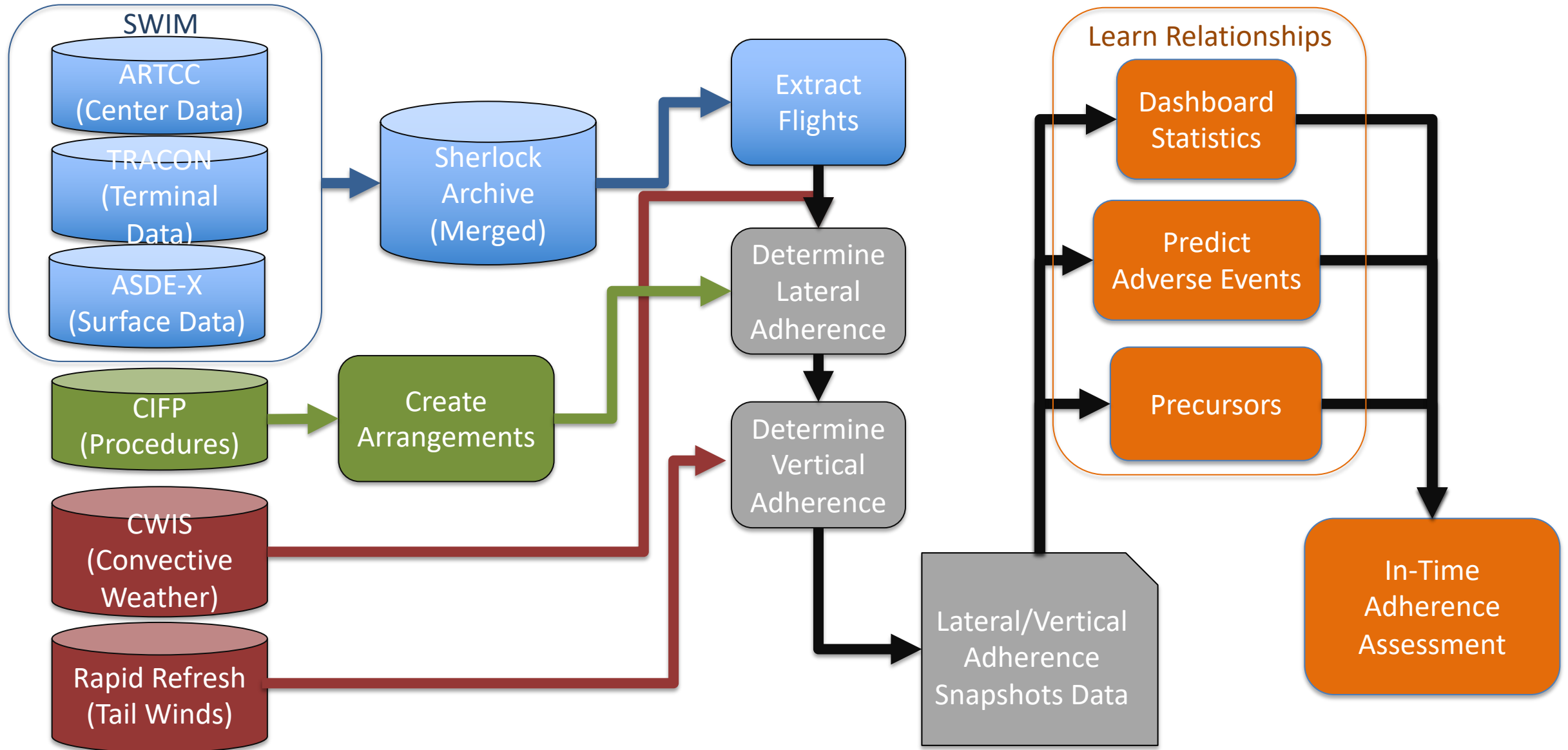


- New Features
 - Time Conflicts at Merging Waypoints
 - Level off before waypoint (indicate intentional deviations)
 - Miles in Trail
- Filtering:
 - Used Runway Landing to our advantage.
- Statistical Features:
 - Computed deviations from mean for altitudes and wind speeds.

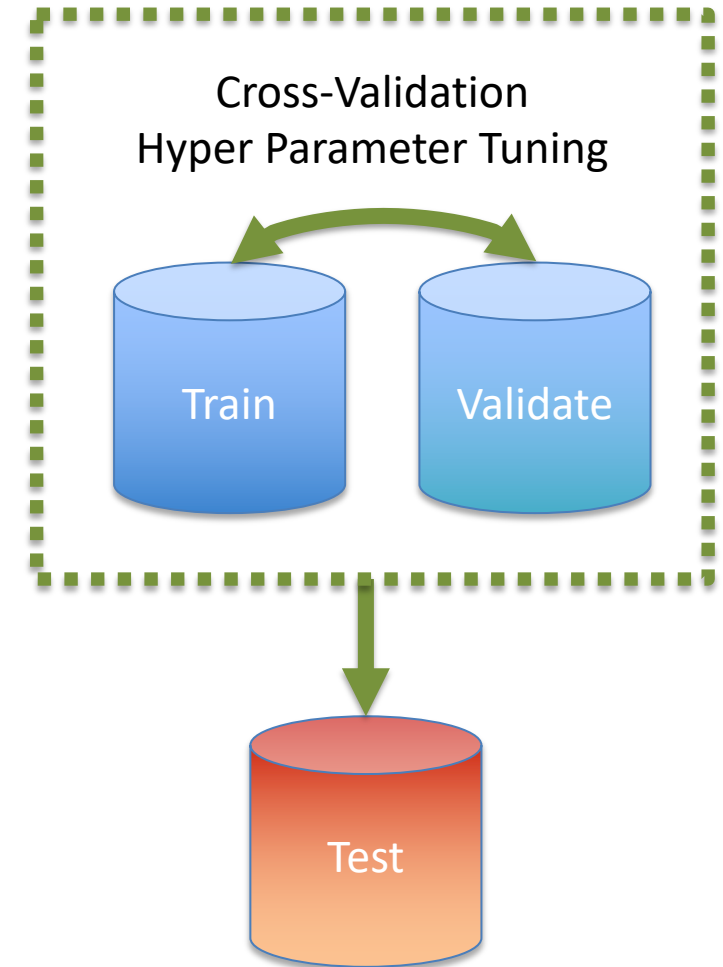


- Determine input file formats
- Address error handling
- Identify System Requirements
 - Disk space
 - Computing resources
 - Memory
 - Network bandwidth
 - Determine if parallel computing is necessary
 - Is Map/Reduce framework necessary or is batch processing or single thread sufficient?
- Choose appropriate programming language
 - Scala
 - Python
 - Java
 - C/C++
- Develop and Test Code
 - Verify & Validate Output (Make sure the processed data maps to the original problem)

Build Data Processing Pipeline: Case Study



- Based on problem definition select algorithm
 - Black box (typically non-linear methods)
 - Good for performance
 - Poor for interpretability
 - White box
 - Good for interpretability
 - May not give best performance
- Train model
 - Determine what hyper parameters generalize well
 - Grid search
 - Cross validation
- Test model
 - If data is stationary a fixed model will work
 - If non-stationary (in temporal applications) updates to the model may be needed to account for shifts in data.
- Interpretability
 - Summary statistics
 - Parameter significance (weights)
 - Thresholds (decision trees)
 - Probabilistic models with confidence bounds



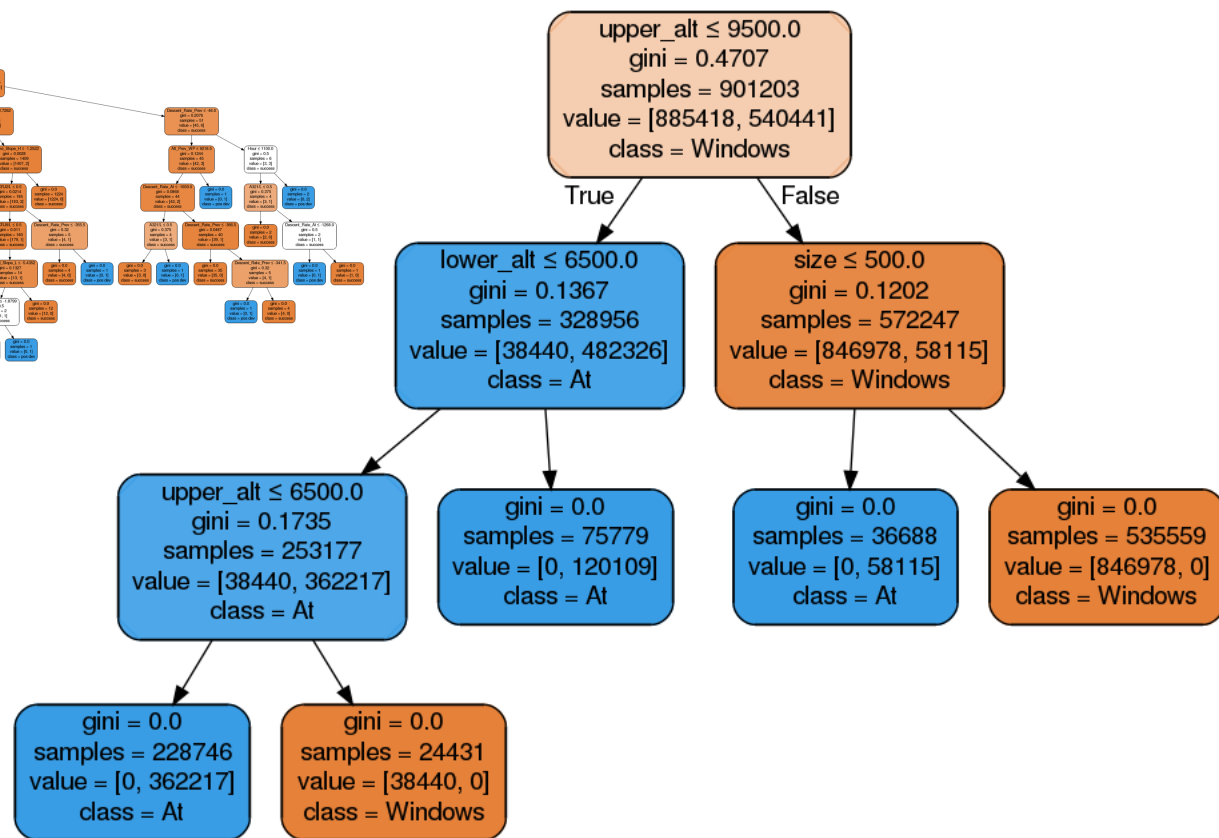
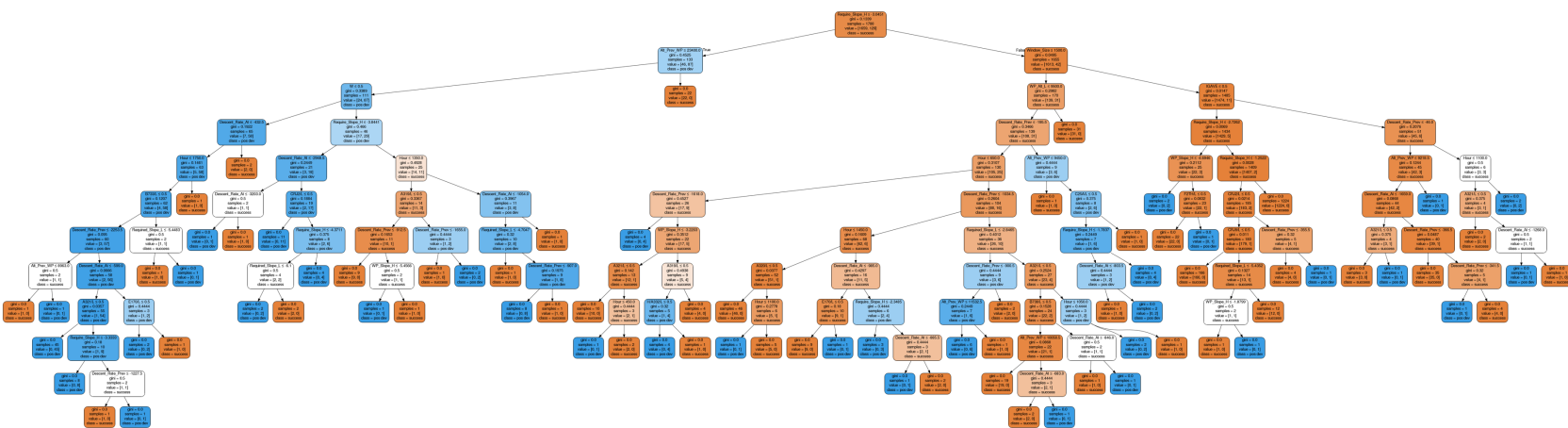
Build Model and Test: Case Study



Complex

Vs

Interpretable



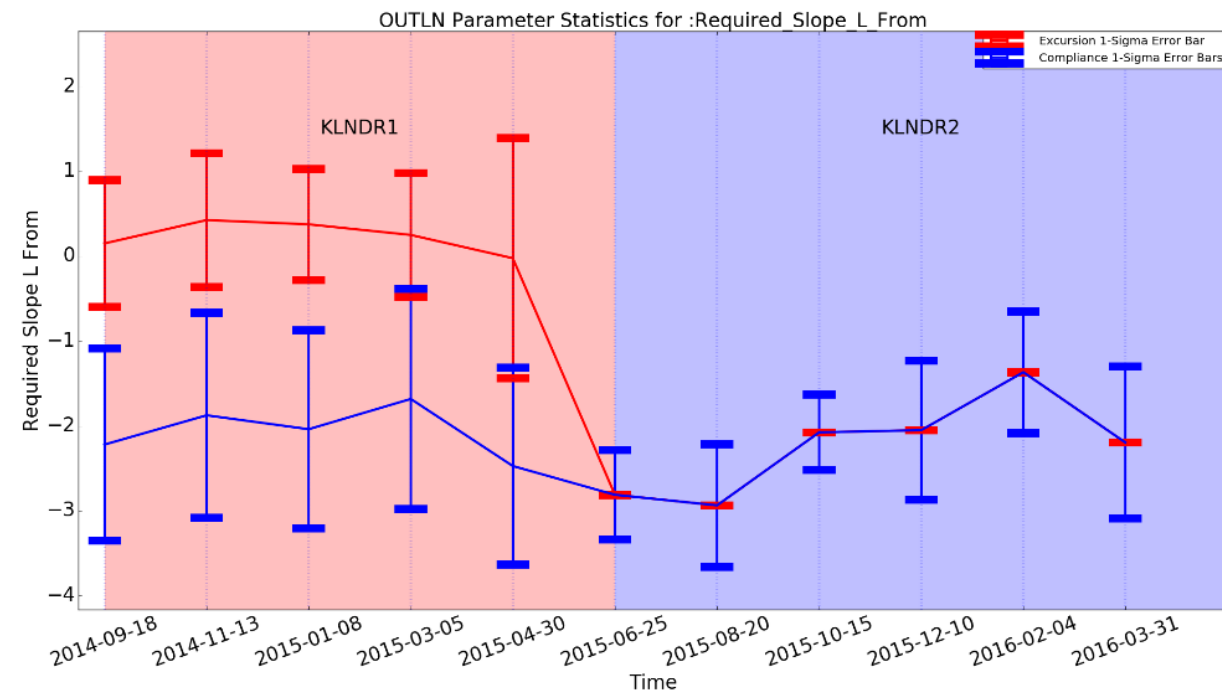
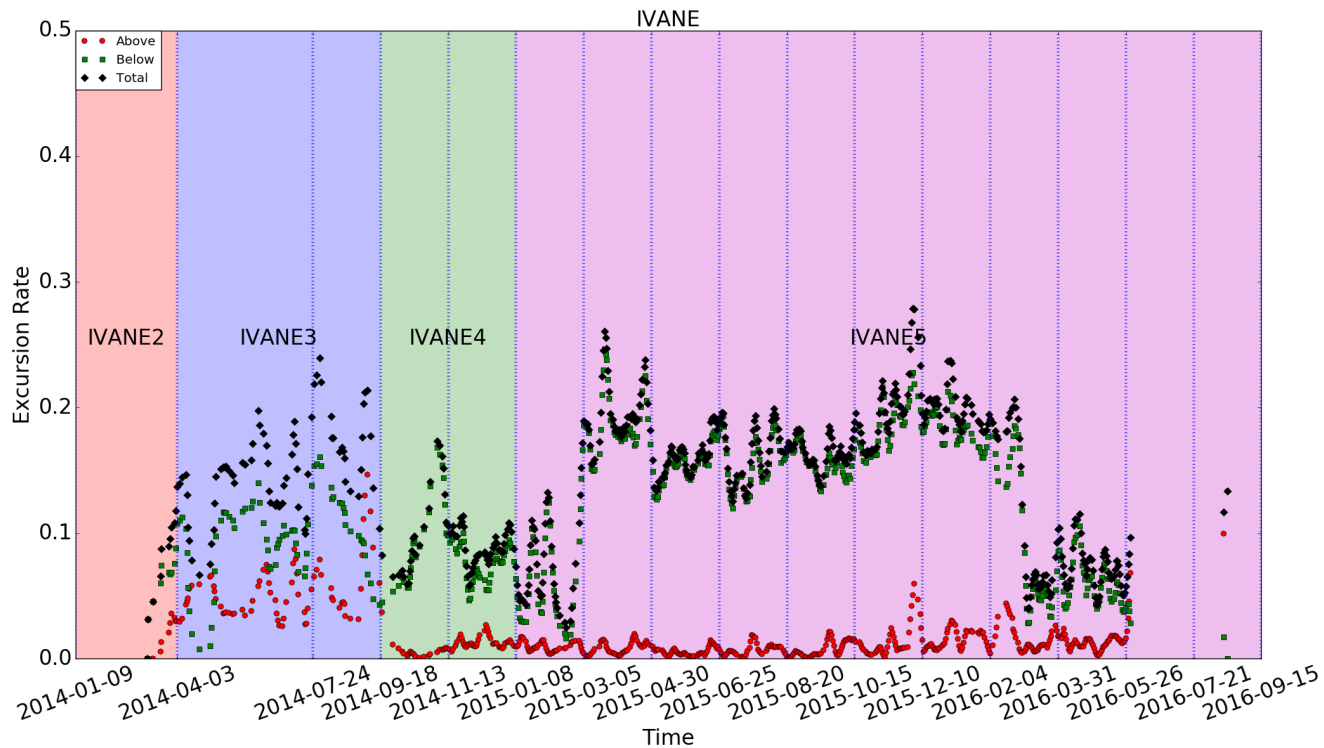
- Explainable AI:
 - Trade off between superior performance of complex models and model understanding.
 - Low model transparency affects trust and adoption within the industry.
 - Tying the model decisions back to the original input space (both temporal and across features).
 - Model robustness and how stable the model is with respect to perturbations in the data.



Build model and test: Case Study



Behavior Can Change Over Time

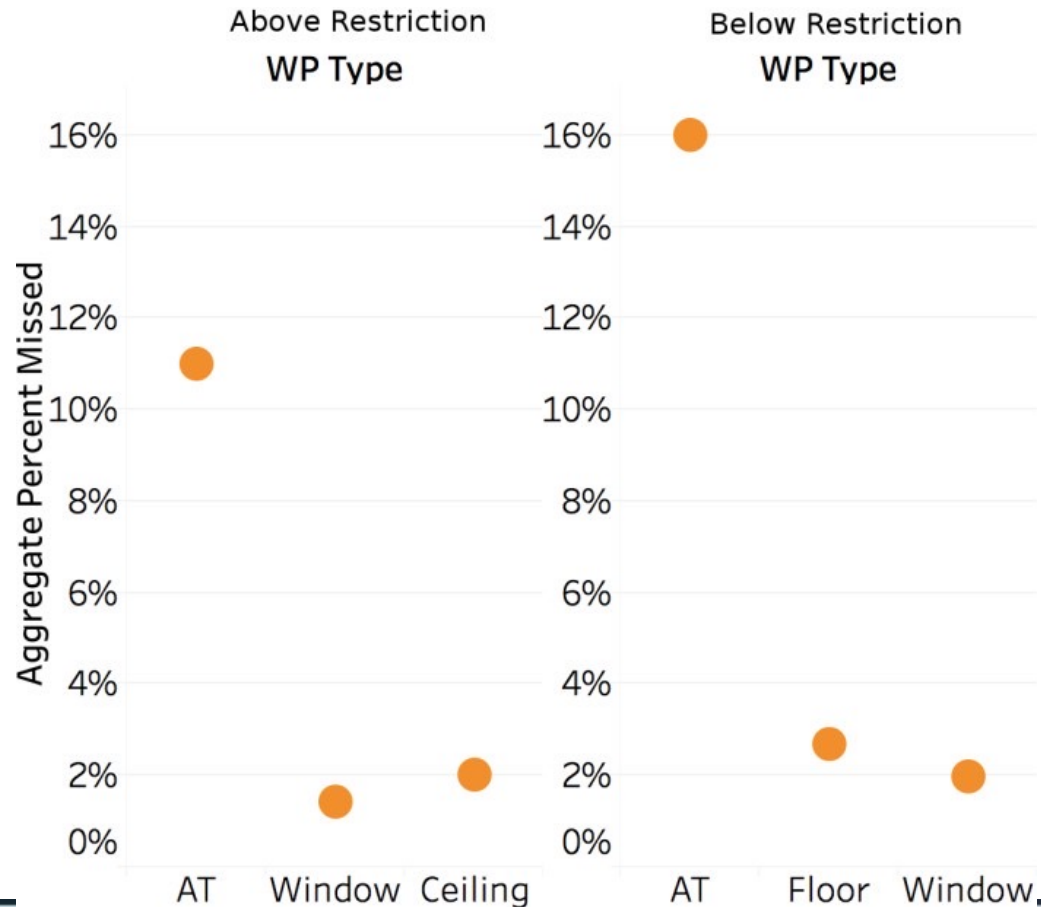


- Do the results make sense?
 - Is the performance what you expected? (100% can be fishy)
 - Review and Validate with Subject Matter Expert
 - In white box algorithms do highly weighted features match what the domain says should be important?
- Does it answer the original question?
 - Are results significant?
 - Can the results leveraged to improve the system?
 - Is there added value over the previous state-of-the-art?
 - Saves (Time/\$\$)
 - Improves safety/lowers risks
 - Improves user experience

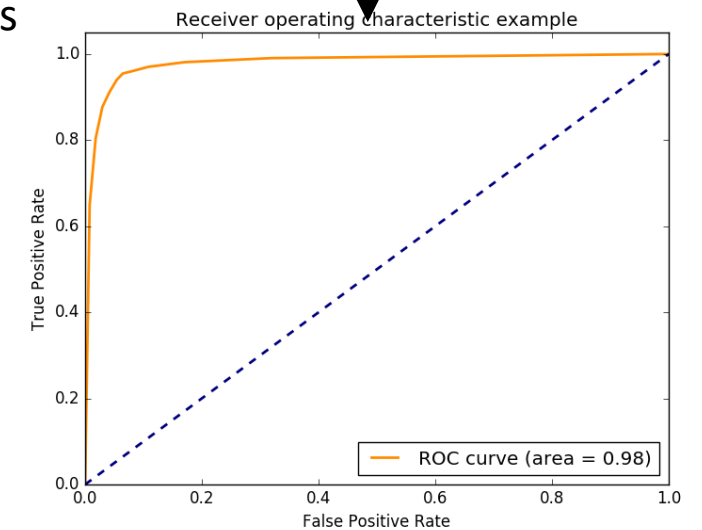
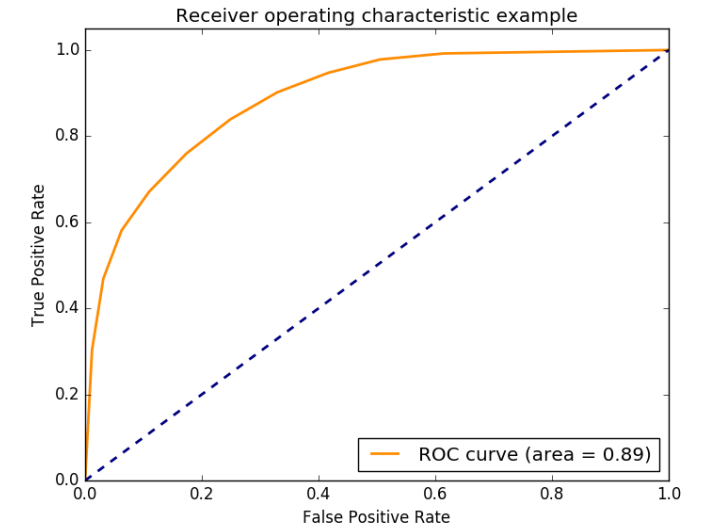
Analyze Results: Case Study



- Do Results Make sense?
 - Tighter restrictions result in less adherence

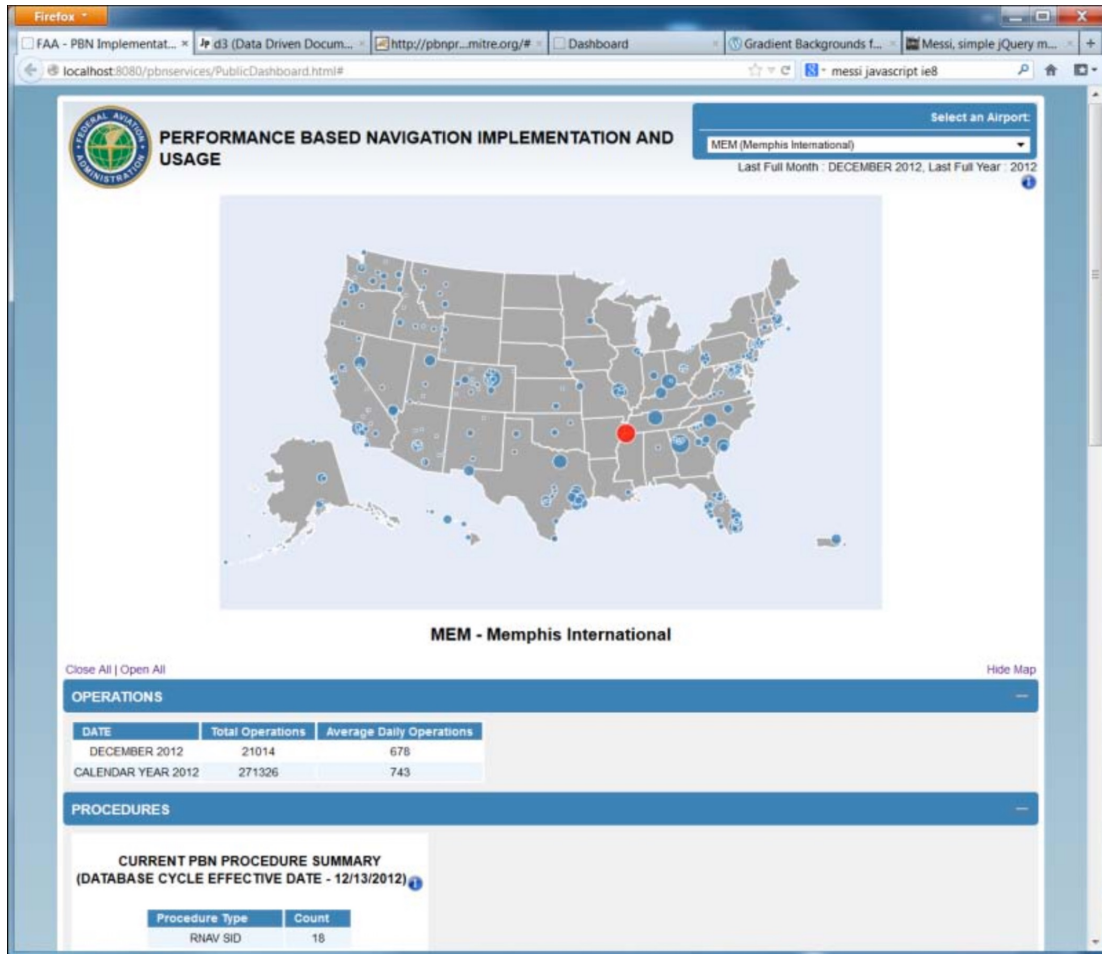


- Is there improved performance?
 - False Positive/
True Positive Rates



- Document process so results can be reproduced
- Deploy
 - Review prototype
 - Are there areas that need improvement
 - Error handling
 - Port to more efficient coding language
 - Identify dedicated machinery to run production environment
 - Write user guide
 - Write code to work out of the box on sample data for demo purposes and package handoff.
- Publish
 - Identify appropriate venue
 - Cite foundational work
 - Distill findings so that there is a clear message describing the new approach's added value
 - When appropriate release code and data for transparency and reproducibility

Deploy/Publish: Case Study



Performance Based Navigation Dashboard



T. Becher, "Performance-based navigation analysis and reporting," *2013 Integrated Communications, Navigation and Surveillance Conference (ICNS)*, Herndon, VA, 2013, pp. 1-17. doi: 10.1109/ICNSurv.2013.6548637

URL: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6548637&isnumber=6548506>



Objective Assessment Method for RNAV STAR Adherence

Michael Stewart
San José State University Research Foundation
NASA Ames Research Center
Moffett Field USA
Michael.J.Stewart@nasa.gov

Bryan Matthews
Stinger Ghaffarian Technologies Inc.
NASA Ames Research Center
Moffett Field USA
Bryan.L.Matthews@nasa.gov

Abstract—Flight crews and air traffic controllers have reported many safety concerns regarding area navigation standard terminal arrival routes (RNAV STARs). Specifically, optimized profile descents (OPDs). However, our information sources to quantify these issues are limited to subjective reporting and time-consuming case-by-case investigations. This work is a preliminary study into the objective performance of instrument procedures and provides a framework to track procedural concepts and assess design specifications. We created a tool and analysis methods for gauging aircraft adherence as it relates to RNAV STARs. This information is vital for comprehensive understanding of how our air traffic behaves. In this study, we mined the performance of 24 major US airports over the preceding three years. Overlaying 4D radar track data onto RNAV STAR routes provided a comparison between aircraft flight paths and the waypoint positions and altitude restrictions. NASA Ames Supercomputing resources were utilized to perform the data mining and processing. We assessed STARs by lateral transition path (*full-lateral*), vertical restrictions (*full-lateral/full-vertical*), and skipped waypoints (*skips*). In addition, we graphed frequencies of aircraft altitudes relative to the altitude restrictions. *Full-lateral* adherence was always greater than *Full-lateral/full-vertical*, as it is a subset, but the difference between the rates was not consistent. *Full-lateral/full-vertical* adherence medians of the 2016 procedures ranged from 0% in KDEN (Denver) to 21% in KMEM (Memphis). Waypoint skips ranged from 0% to nearly 100% for specific waypoints. Altitude restrictions were sometimes missed by systematic amounts in 1,000 ft. increments from the restriction, creating multi-modal distributions. Other times, altitude misses looked to be more normally distributed around the restriction. This tool may aid in providing acceptability metrics as well as risk assessment information.

Keywords—data mining; RNAV STAR; procedures; adherence; waypoint

I. INTRODUCTION

Area navigation (RNAV) is a cornerstone of the Federal Aviation Administration's (FAA) plan for future instrument procedures. Between 2009 and 2016, 264 RNAV standard terminal arrival routes (STARs) were implemented, and they are still continuously increasing [1]. This rapid implementation has incurred unforeseen side effects. Major and regional airlines

have both voiced safety concerns about deviations at a NASA workshop meeting in November 2015. Boston's STARs were redesigned in 2016 after numerous operational problems, and Atlanta was forced to stop using the vertical and speed profiles of their newly designed RNAV STARs [2],[3]. Our interest in RNAV procedures is pointed specifically towards RNAV OPDs and stems from their increased functionality (e.g., vertical profiles and speed control) and its resulting complexity. To fully understand the operational effects of this paradigm change, either from the flight deck or air traffic control perspective, we need a source of objective data describing operational performance of RNAV procedures. That is, a method for comprehensively capturing and describing the adherence trends of instrument procedures. We introduce the idea that procedures are important entities in their own right and should be treated as the unit of analysis. They are not just collections of waypoints, and it is performance along a path through a STAR that should be analyzed. By monitoring the flightpaths of aircraft, we can understand how STARs are functioning in the airspace. Currently no system exists that can monitor the adherence of aircraft utilizing RNAV procedures in a comprehensive way.

RNAV STARs are often designed to enhance efficiency and regulate throughput. This expands upon the previous intentions of instrument procedures: terrain avoidance and standardized routing. This expansion of functionality increases complexity, particularly, the addition of the speed restricted vertical profile for the precision it requires, and the addition of transition routes to control specific runway assignments. Vertical profiles add variables pertaining to aircraft performance such as drag, idle, and auto-flight variability; transitions add rigidity by limiting lateral path flexibility.

A. Purpose

In order to maintain or reduce the current level of risk we need to understand the implications of existing and new procedures and our system of controlling air traffic. A possible risk with deviating from a STAR is loss of separation leading to a midair collision. Complying with a STAR keeps the airplane on a known trajectory so it can be protected. Issues such as energy management, loss of control, controlled flight into terrain, wake turbulence, and weather are also possible areas of concern for RNAV STAR adherence. We do not

- Process may vary from application to application, but generally the data understanding and acquisition accounts for >80% of the effort.
- Understanding the data and domain (strengths and weaknesses) is key to successful machine learning.
- As in all V&V processes it is important to continually be checking that each step maps back to the original problem statement.
 - If you do change direction, make sure all previous steps map to the new problem.

Questions?

