Big Data Analytics for Aeronautics

NASA SMARTNAS 2.4 NRA
Interim Results

ATAC Corporation
Metron Inc.
University of California Berkeley
January 30, 2019

ATAC
Aviation Analysis Experts

Contents

- Overview Team Progress
  - NRA Objectives
  - Year 3 Work Plan & Activities
  - Data Set Utilized & Methodologies
- Integration with NASA Systems
  - Sherlock ATM Data Warehouse
  - ATM-X Testbed
- Anomaly Detection Updates
  - Updated indicators
  - NSB Scoring Results
- Analysis of Data
  - Update of voice data analysis
  - Building a prognostic model for go-arounds
- Next Steps

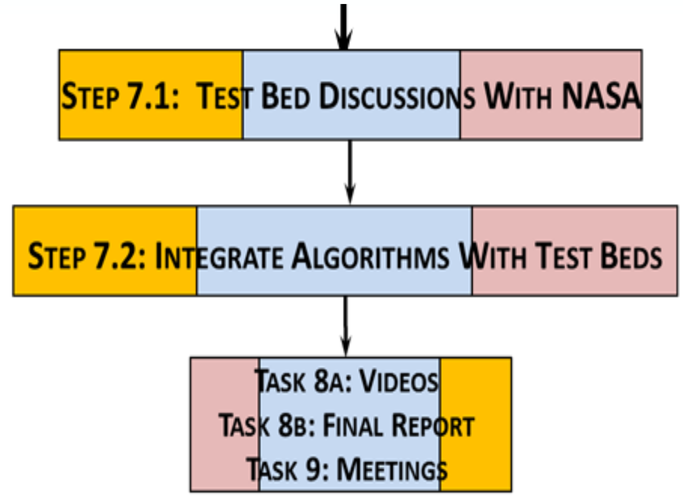
NRA Objectives

- Develop and apply data mining algorithms that identify degraded states of the NAS and their precursors
  - Identify sequences of states that lead from precursor to degraded states with higher than normal probability
  - Accommodate supervised learning through human feedback
  - Indicate operationally significant incidents
- Develop data mining algorithms to aid in the development of metrics associated with safety and efficiency of the NAS
- Year 2 - Add capability of data mining algorithms to be updated daily
- Year 3 - Deploy algorithms to the SMARTNAS testbed or other NASA Platforms

Year 3 Work Plan Overview

Year 3

Tasks 7, 8,9



- Develop approach for ATM-X testbed integration through discussions with Testbed personnel (already started).
- Continue iterative anomaly detection development
  - Incorporate energy features into anomaly detection
  - Add metrics derived from automated voice processing to features
- Continue to develop approaches for prognostic modeling (go arounds)
- Continue to develop continuous processing moving towards real-time model updates

Year 3 Work Activities to date

- Finalize additional safety-based indicators to augment the current set
  - Overtake situations
  - High-Energy approaches
- Finalize voice metrics to include in anomaly detection
- Continued data preparation for training data sets
- Development of go-around causal factor analysis to lead to predictive model for go-arounds
- Initial design for integration with NASA systems
  - Sherlock ATM Data Warehouse
  - ATM-X Testbed

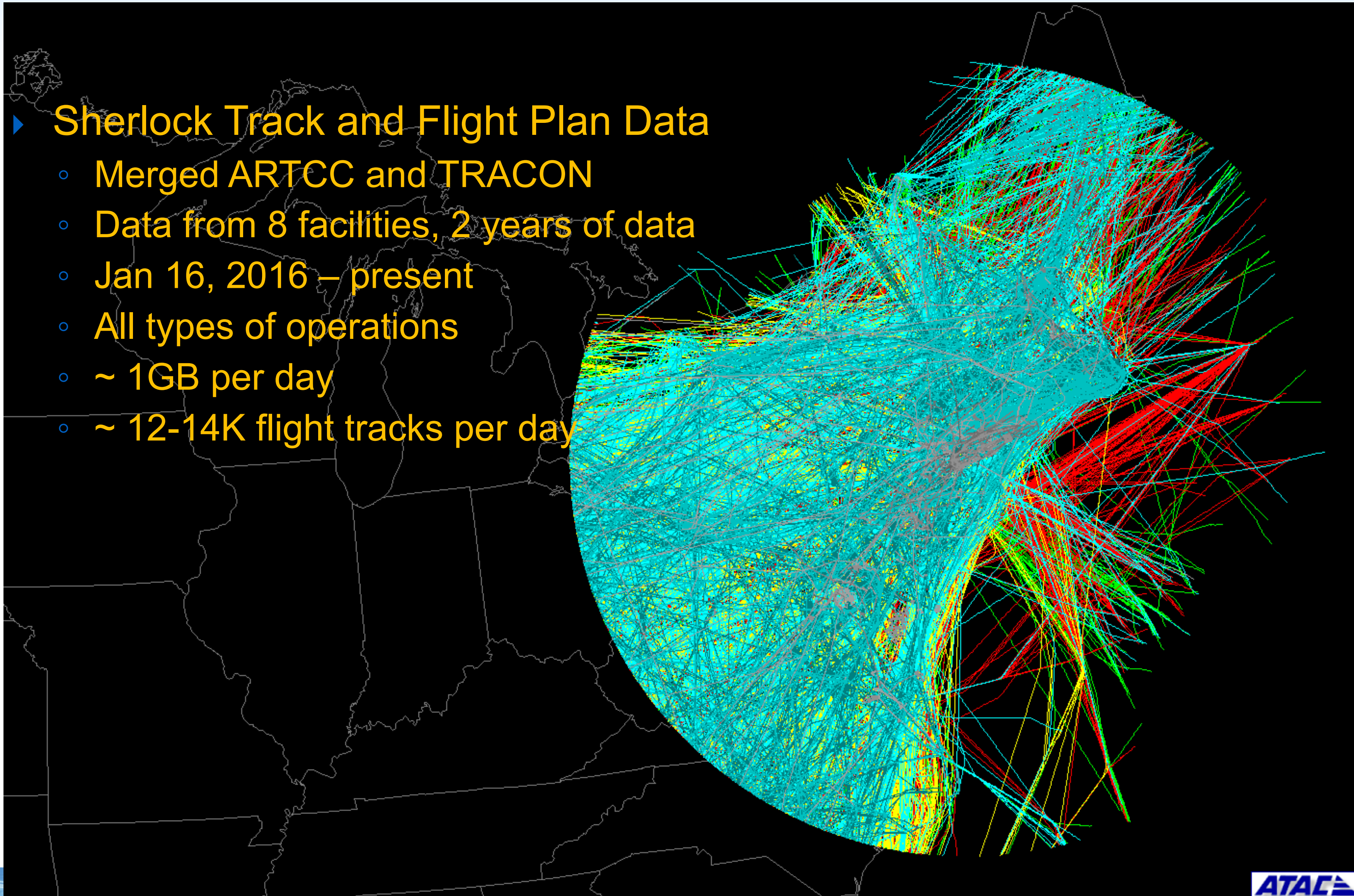
Data Sets Utilized & Methodologies

Additional Data Sets Selected/Prepared

- Sherlock ATM Data Warehouse Track and Flight Plan Data for NY Area
  - Merged 8 ATC facilities – N90, ZNY, ZOB, ZID, ZDC, ZBW, ZTL, ZAU
- Processing expanded to Jan 2016 – present ~ 3+ years of operational data.
- Performance Data from Sherlock Reports
  - Turn to Final (measures that characterize the final approach)
- ATC Voice Data
  - Downloading Voice Recordings from liveatc.net, starting from 2/13/17
    - Focus on JFK tower, final, and approach
      - KJFK tower (3 frequencies)
      - KJFK final (1)
      - KJFK CAMRN approach (4)
      - KJFK ROBER approach (2)

Data Sets…

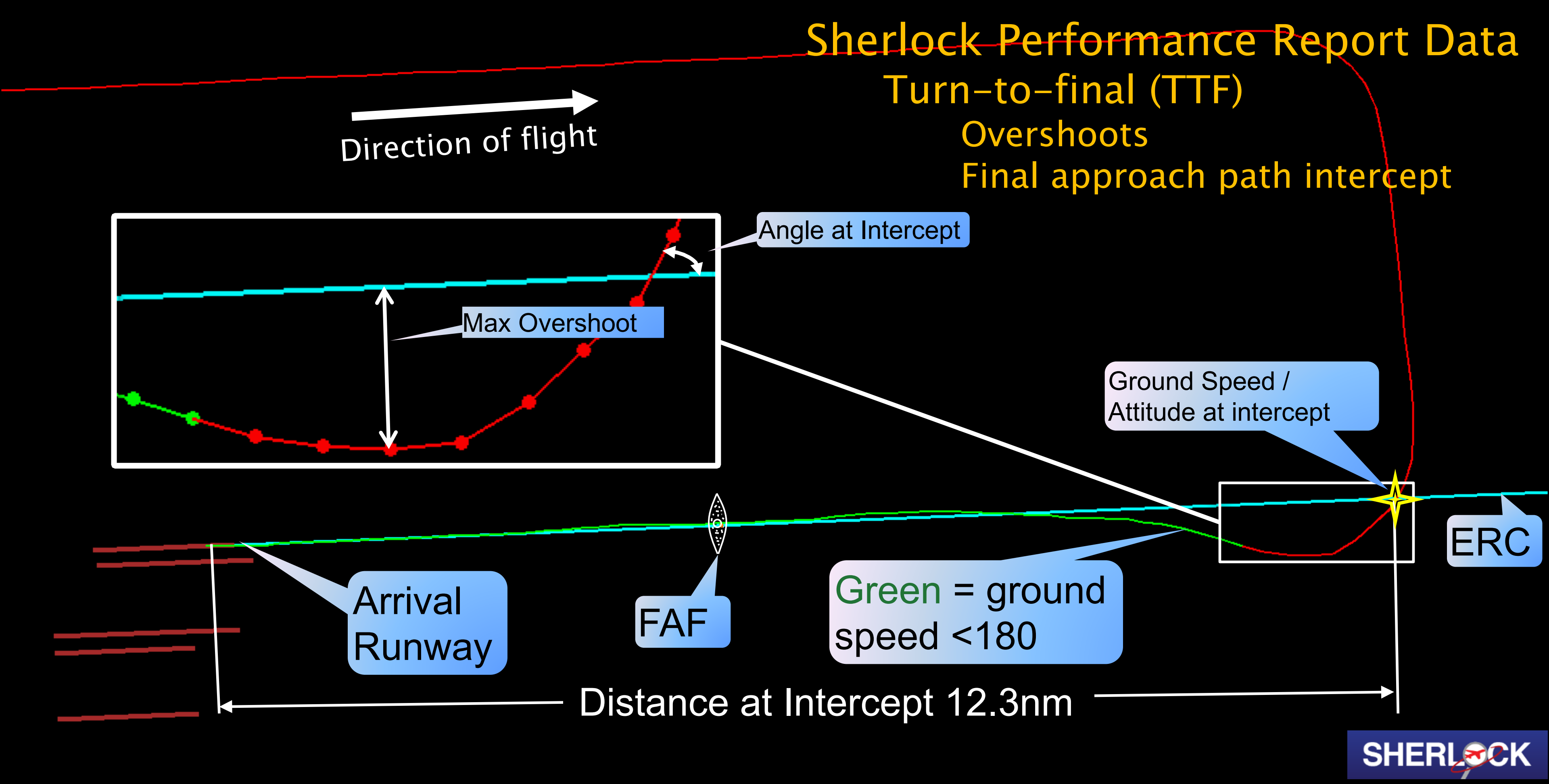
- Sherlock Track and Flight Plan Data
  - Merged ARTCC and TRACON
  - Data from 8 facilities, 2 years of data
  - Jan 16, 2016 – present
  - All types of operations
  - ~ 1GB per day
  - ~ 12-14K flight tracks per day



Turn to Final Overview – measures used as features for anomaly detection

Turn To Final

Event Date/Time(UTC	ACID	Runway	Aircraft Type	Airport	MaxOverShoot(ft)	Dist@Int	Angle@Intercept	Speed@Int	Total
11/03/2009 00:23:22	ASQ5529	26R	CRJ2	ATL	3482	12.32	60	213	1



Methodologies Employed

- Iterative Development, Analysis, Review

2. Prepare Data Sets

1. SME Consultation/Review

3. Run Anomaly Detection

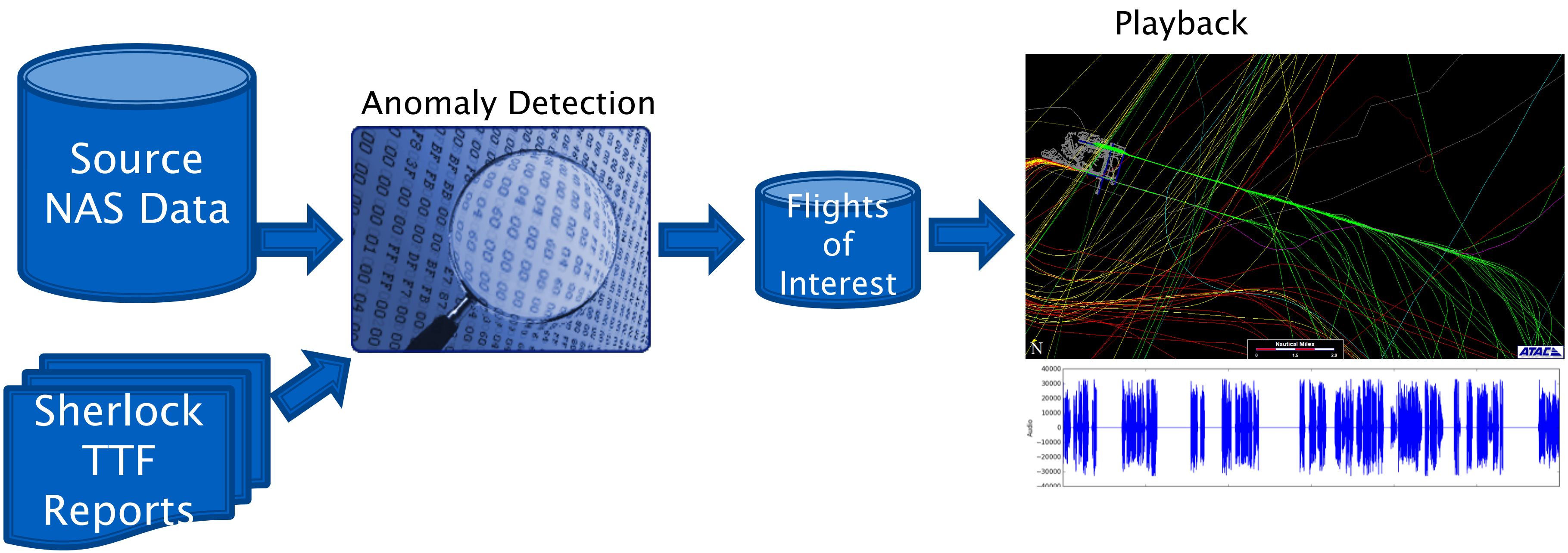
5. Analyze Results

4. Add Supplemental Data

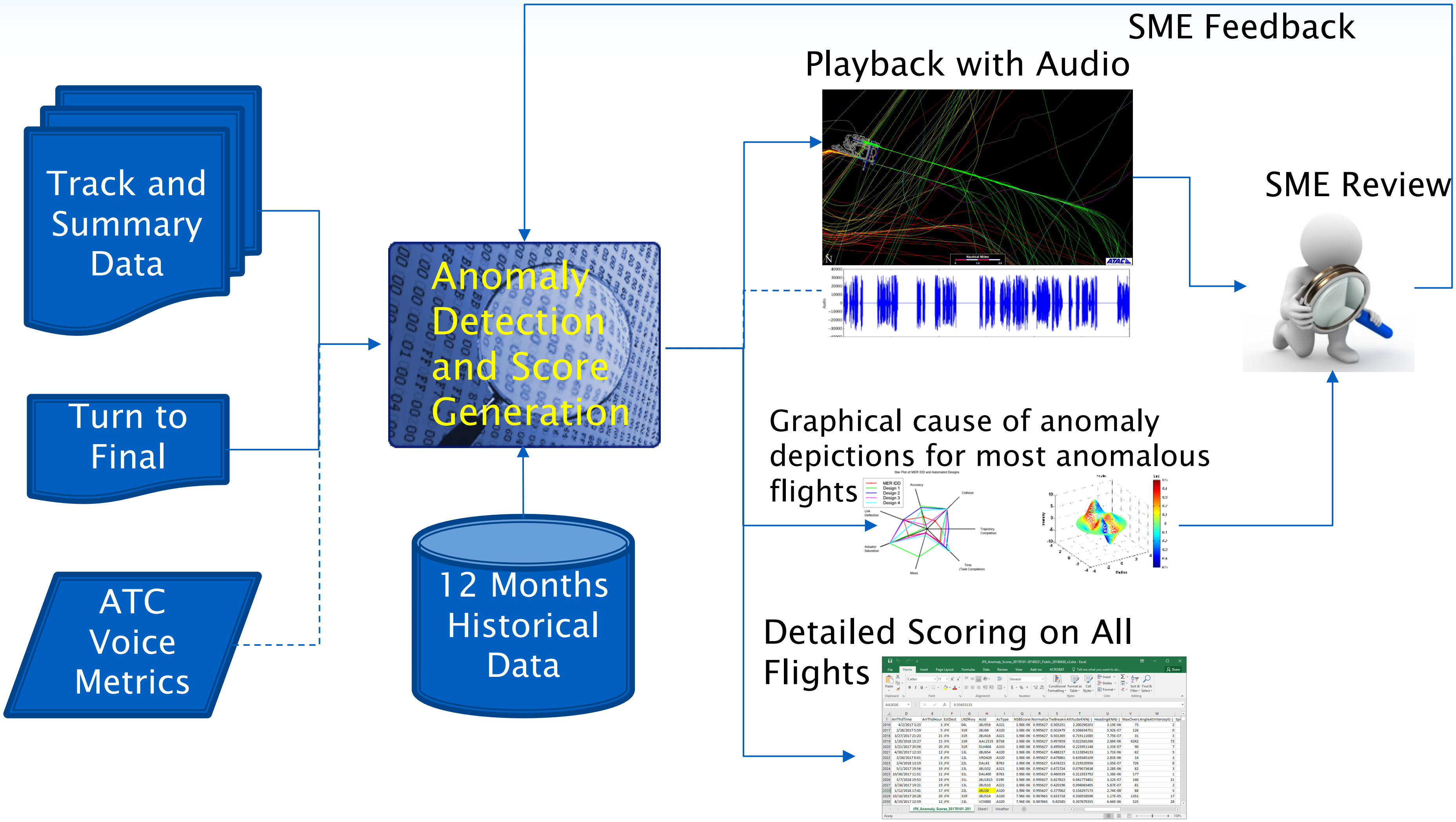
~Quarterly Frequency

SME Review Tool

- Automatically makes videos of top “X” anomalous flights
- Merges and syncs voice recording (when available)
- Allows for quick SME review
- Facilitates supervised learning



Overnight Update in Development System (current)



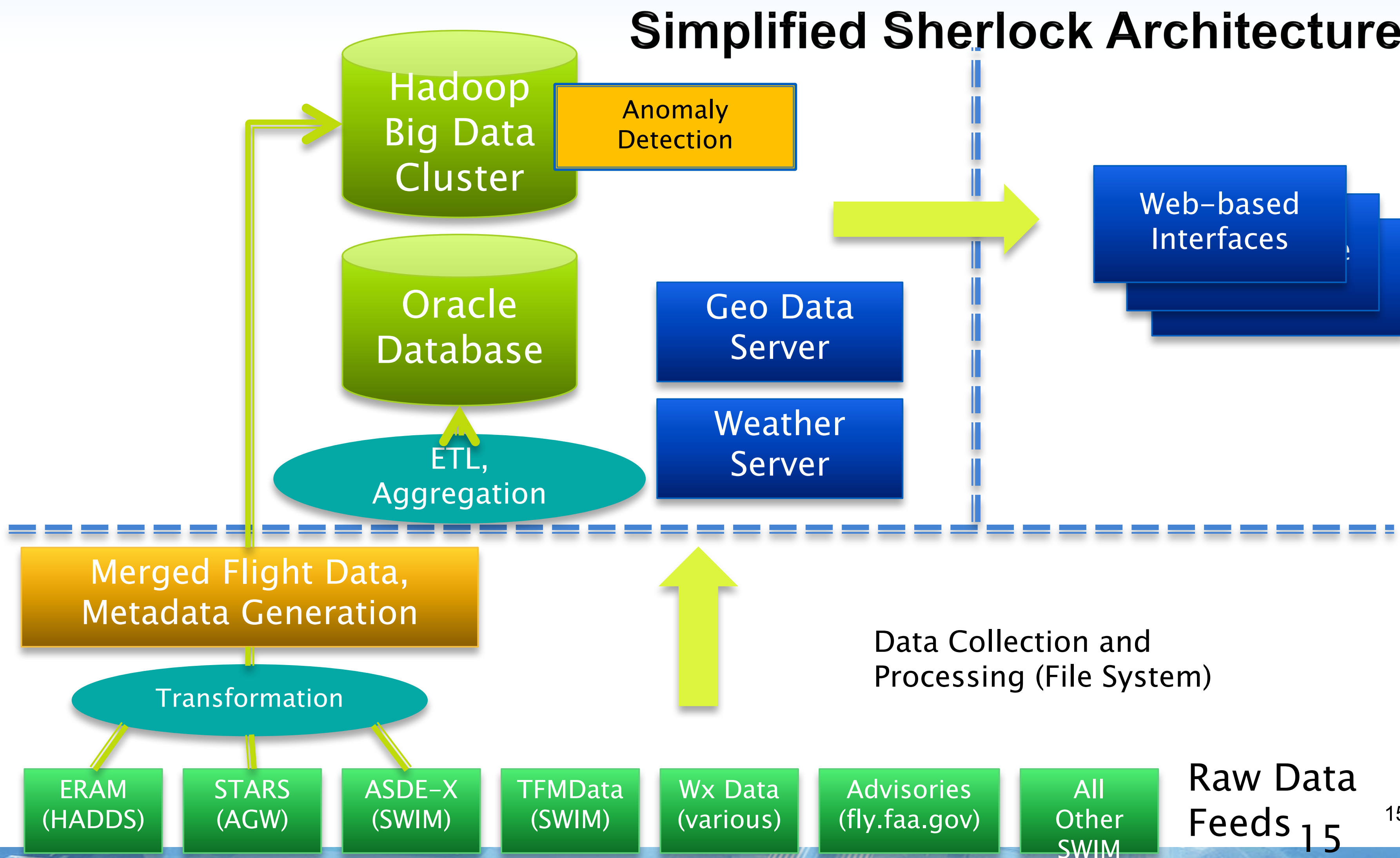
To be implemented in NASA systems…

Integration with NASA Systems

Integration Overview

- Integration with NASA systems includes 2 phases:
  1. Phase I – Migrate anomaly detection processing to Sherlock ATM Data Warehouse Big Data computing cluster
  2. Phase II – Integrate with ATM-X testbed by producing an Anomaly Detection Service
- Advantages to this approach:
  - Sherlock provides access to the data (IFF/RD/ and TTF)
  - Leverages Sherlock existing Big Data computing assets
  - Integration is internal inside NASA programs (no need for SAA or other external access mechanism)

Phase I: Migration of Anomaly Detection to Sherlock



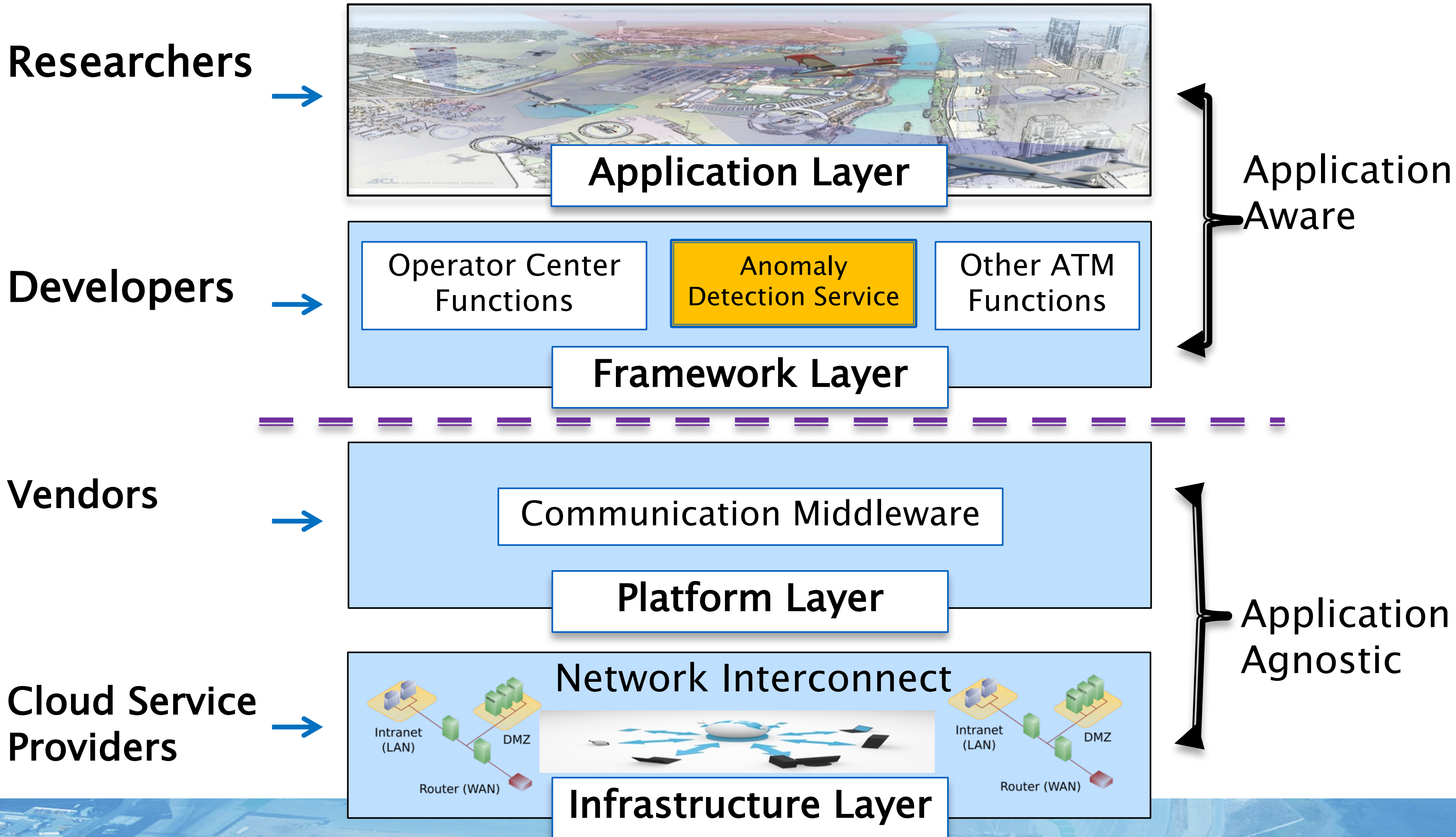
METRON SCIENTIFIC SOLUTIONS ATAC Berkeley UNIVERSITY OF CALIFORNIA

Migration of Anomaly Detection to Sherlock

Sherlock Big Data System

- SuperMicro Engineered System
- Cloudera Hadoop stack
- 42U rack
- Total of 480 CPU Cores, 1752 TB Storage
- 1 Management Node
- 3 Name Nodes (Dual 6 Core, 256 GB RAM each)
- 36 Data Nodes (Dual 6 Core, 128 GB RAM each)

Phase II: Integration with ATM-X Testbed Architecture



Implementation Schedule

- Phase I – 1st Quarter 2019
- Phase II – 2nd Quarter 2019
- Government shutdowns could affect the overall schedule

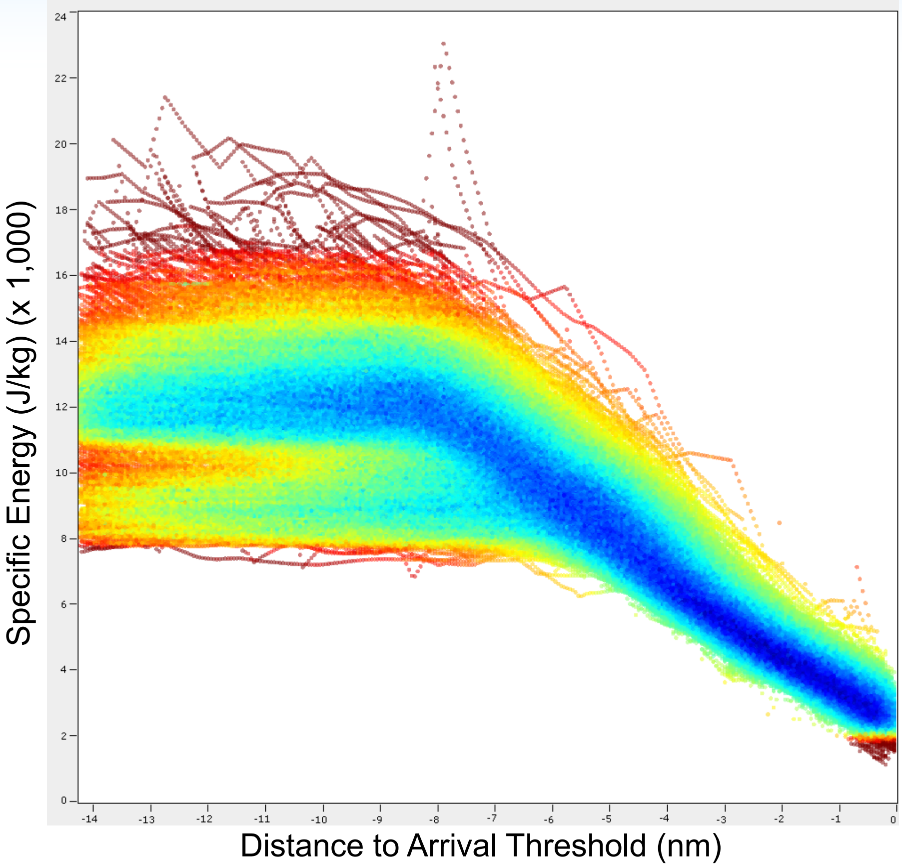
Anomaly Detection Updates

Anomaly Detection Overview

- Compute nine anomaly indicators:
  (those in bold developed under NASA Phase 2 SBIR)
  - **Heading Trajectory k-Nearest Neighbor**
  - **Altitude Trajectory k-Nearest Neighbor**
  - Angle and Speed at Intercept
  - Maximum Overshoot
  - Glide Path Angle at Intercept (Altitude divided by Dist. at Intercept)
  - **Final Approach Positions (unusual locations 1-5nm before runway)**
  - **Overtake Potential (one aircraft closing in on another near runway)**
  - Aircraft Energy (unusually high or low specific energy on approach)
- Normalcy Score Broker (NSB) combines indicators into single anomaly score to identify flights that are outliers in one or more indicators

Aircraft Energy Anomalies

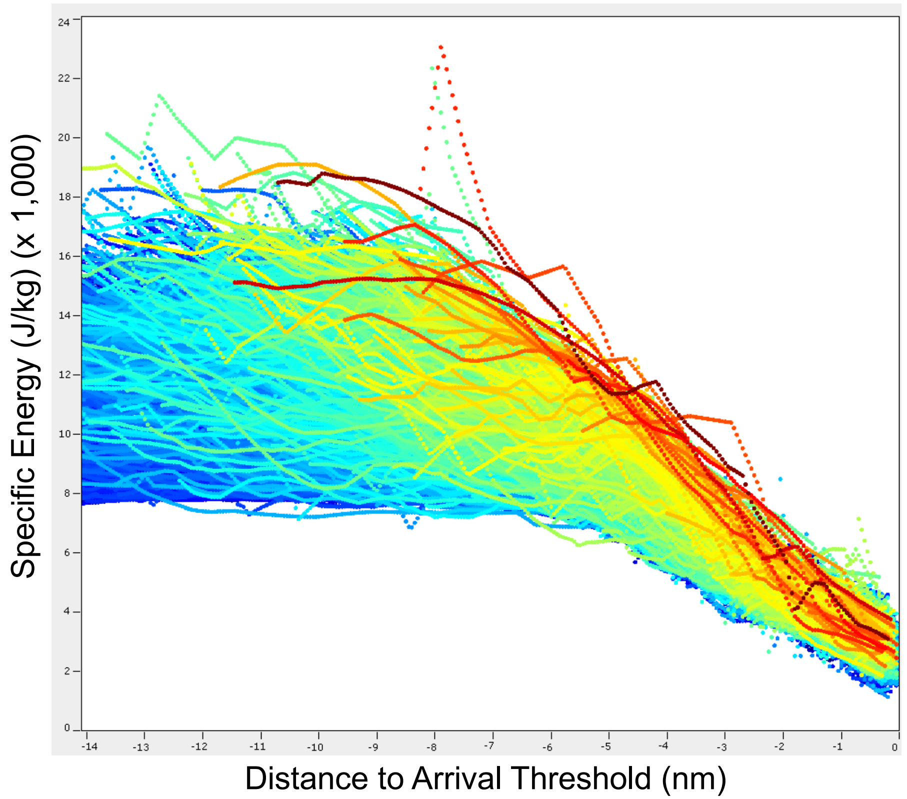
- Identifies flights with unusual specific energy on approach
  - Too high & fast or low & slow
  - Specific energy $\frac{1}{2}v^2 + gh$
    - For velocity v and altitude h
- Measured over approach's final ~15 nm
  - Sample points every 0.05 nm along typical approach path
  - Velocities and positions smoothed using improved Kalman filter
- Energy paths have multiple clusters (see figure, right)
  - Different approaches & runways



JFK 31R energy points individually colored by normalcy over 2018

Aircraft Energy Anomalies

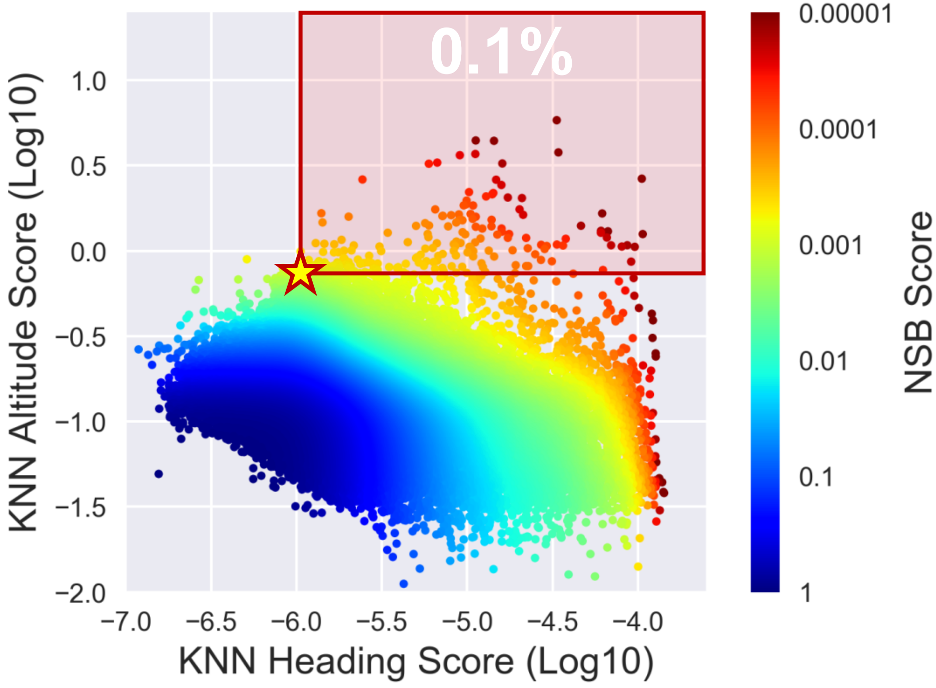
- Energy tracks compared to find anomalies
- Energies normalized to z-scores at each sampled distance
  - Enables comparison of scores across distances with different variances
- Use k-Nearest Neighbor (k-NN) to identify anomalous energy tracks
  - Compare tracks with L1 norm
  - Use exponential weighted average over k=0.5% nearest neighbor distances



2018 JFK 31R flight energy tracks colored by Aircraft Energy indicator

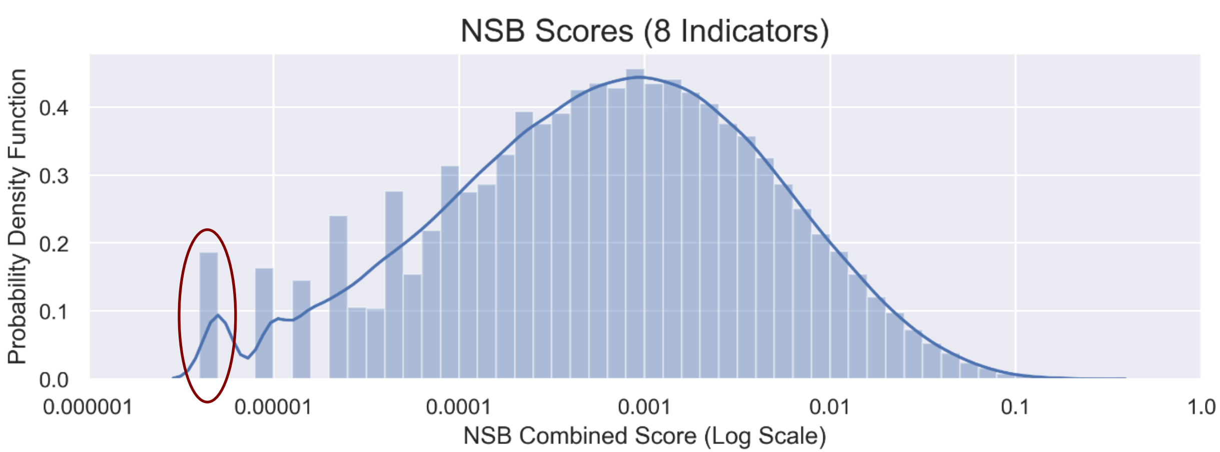
Combined Anomaly Scores

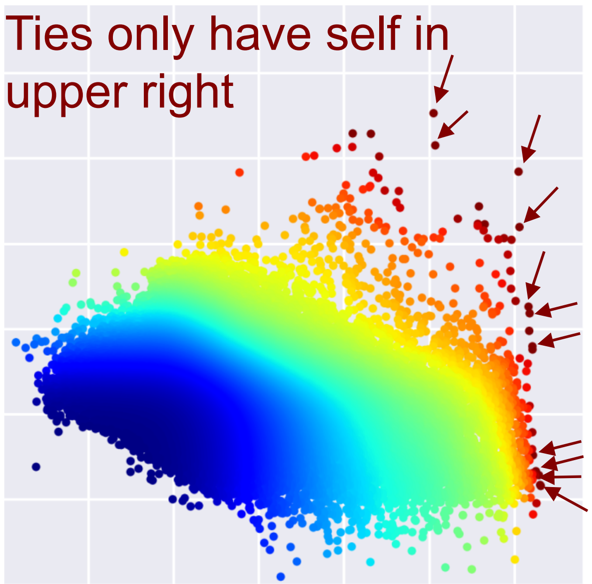
- Normalcy Score Broker (NSB) combines multiple anomaly indicators into single score
- Combined score is proportion of flights at least as anomalous in every indicator
  - Joint CDF measures mass of distribution in upper right
- Ex: Starred flight's score is proportion of flights in red rectangle (including self)
  - Only 0.1% of flights have both indicator scores at least as anomalous as the starred flight



NSB Score Ties

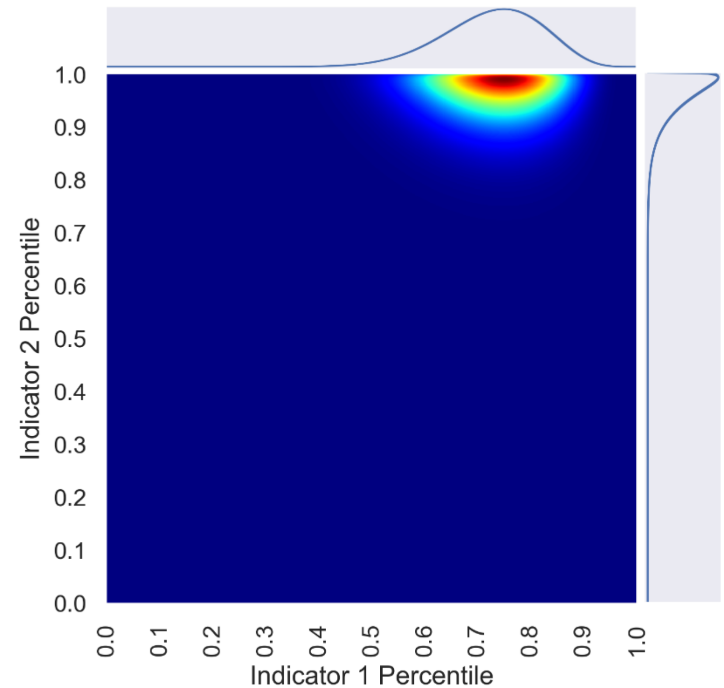
- Normalcy Score Broker (NSB) can result in many ties for the most anomalous combined score
  - More indicators (higher dimensions) generally leads to more ties
  - Negatively correlated indicators lead to more ties
- Some nearby flights of interest fall in the rankings





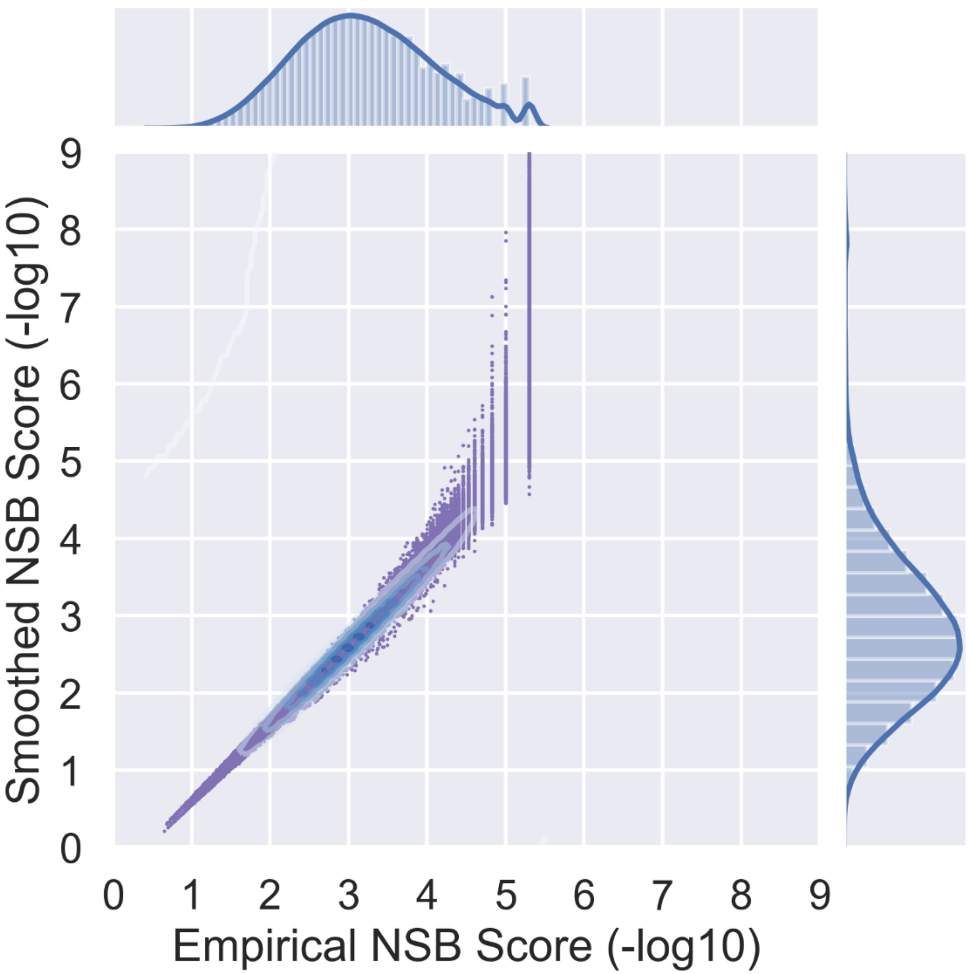
Smoothed NSB Scores

- Break ties and elevate nearby flights by kernel-smoothing the "mass" of each flight
  - First, convert each indicator into a percentile value (does not change ordering and therefore NSB score remains)
  - Then, replace the point-mass of each flight with a multivariate beta distribution
- Example (at right):
  - A flight with indicator percentiles 0.75, 0.99
  - Multivariate beta distribution smooths flight's mass over region $[0, 1]^2$
  - *Example uses exaggerated smoothing bandwidth for improved visualization*
- Smoothed NSB score computes total mass in upper-right of the flight's indicator percentiles



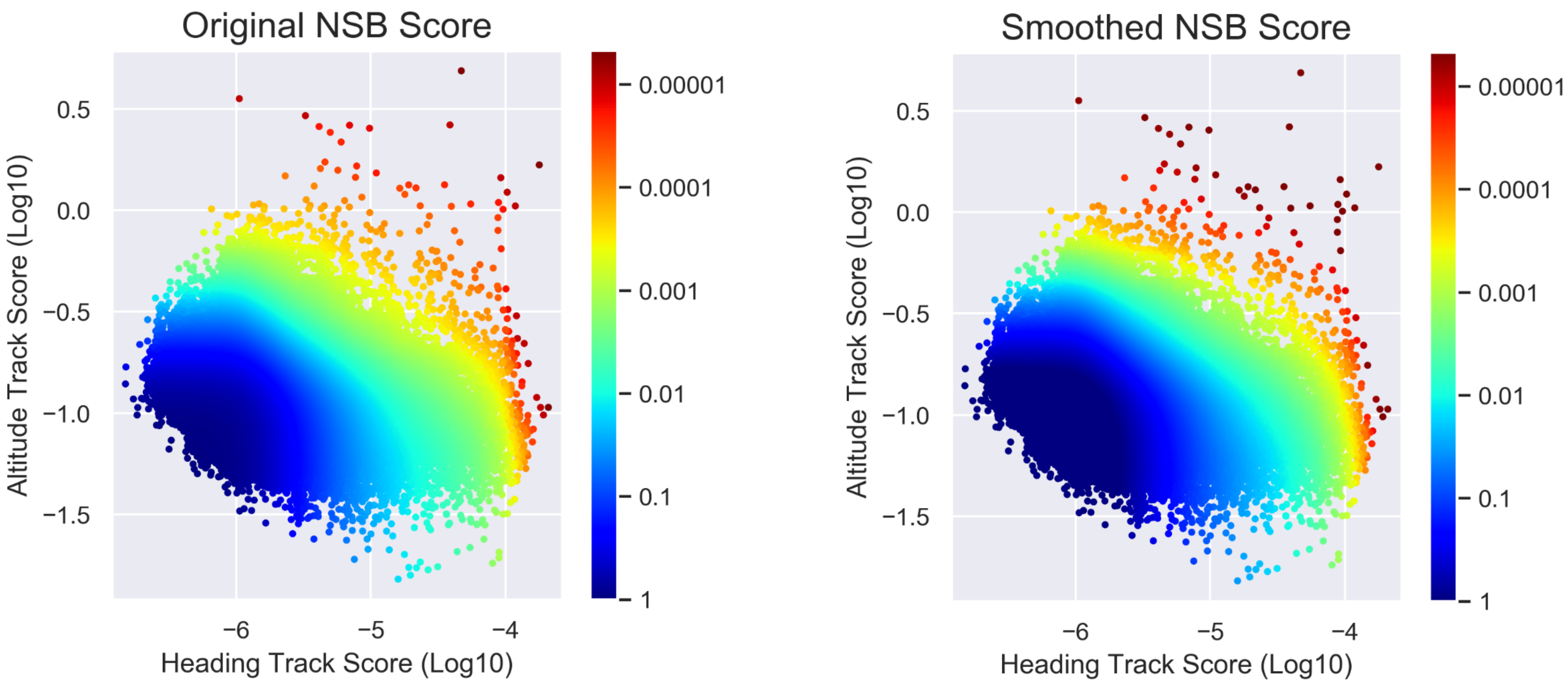
Smoothed NSB Score Results: 8 Indicators

- Smoothed scores more accurately reflect the underlying joint probability distribution
- Ties in anomaly tail are eliminated
- Flights previously tied for second place are promoted
    - Receive scores similar to “nearby” flights



Smoothed NSB Score Results: 2 Indicator Example

- “Nearby” flights receive more similar scores (subtle)

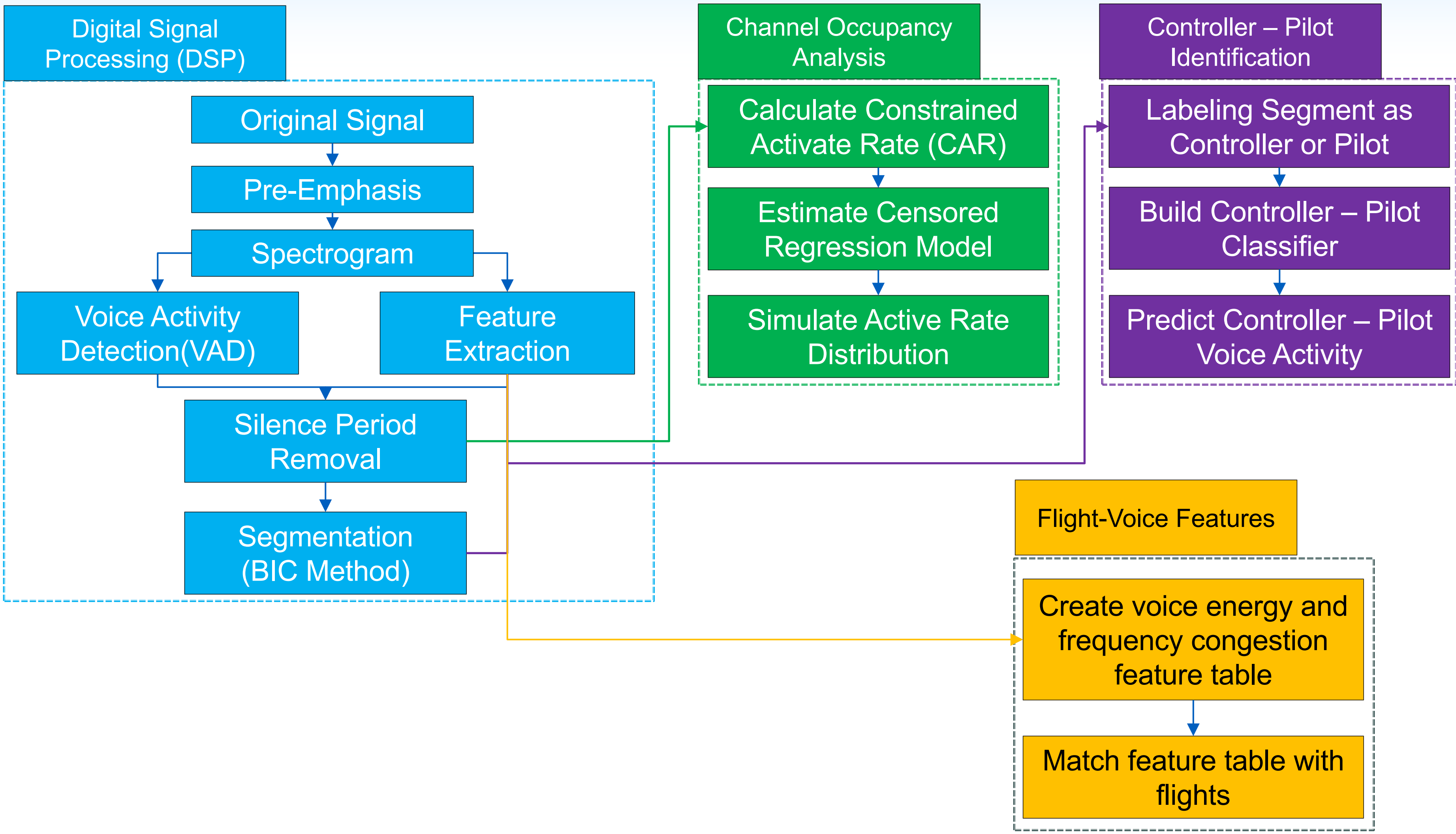


Analysis of Unstructured Data (ATC Voice)

Background & Objectives

- ATC voice data from LiveATC.com records the message exchange between the pilots and the controllers
- Incorporate ATC voice metrics as additional anomaly detection indicators, and explore the correlation between voice features and flight traffic
- Initial trial of speech transcription has poor performance due to lack of training dataset (corpus)
- Instead, spectrum analysis algorithm was applied to extract representative features from the ATC audio data

Methodology – Framework



01

Digital Signal Processing

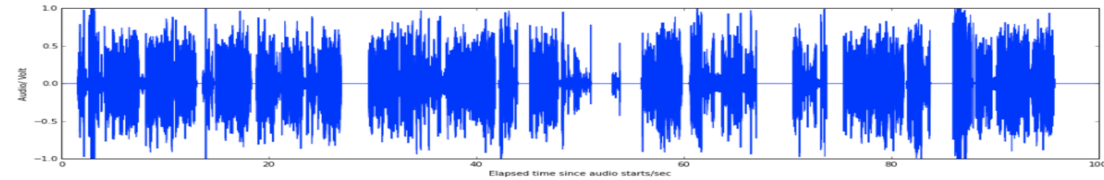
METRON SCIENTIFIC SOLUTIONS

ATAC

Berkeley UNIVERSITY OF CALIFORNIA

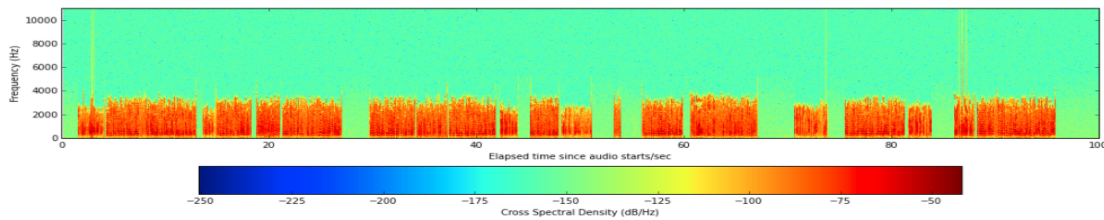
Digital Signal Processing Overview

Pre-Emphasis



Original signal – time domain samples from ATC tower audio

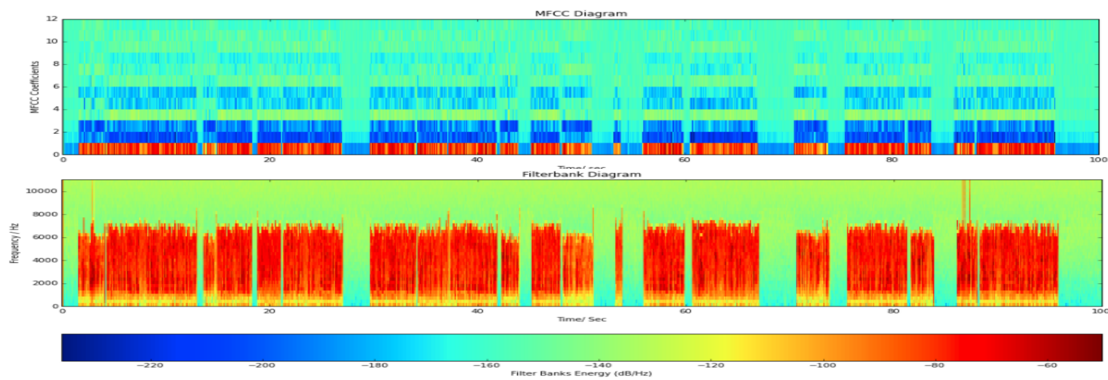
Spectrogram (STFT)



Spectrogram – converting signals into (frequency, time, energy) tuples.

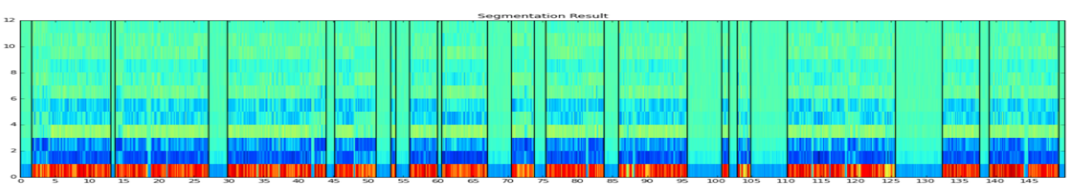
Voice Activity Detection (thresholding)

Feature Extraction (MFCC & Filters)



Feature map – Each **frame** is a vector of features for a short time period (e.g., 20 ms)

Segmentation (BIC method)



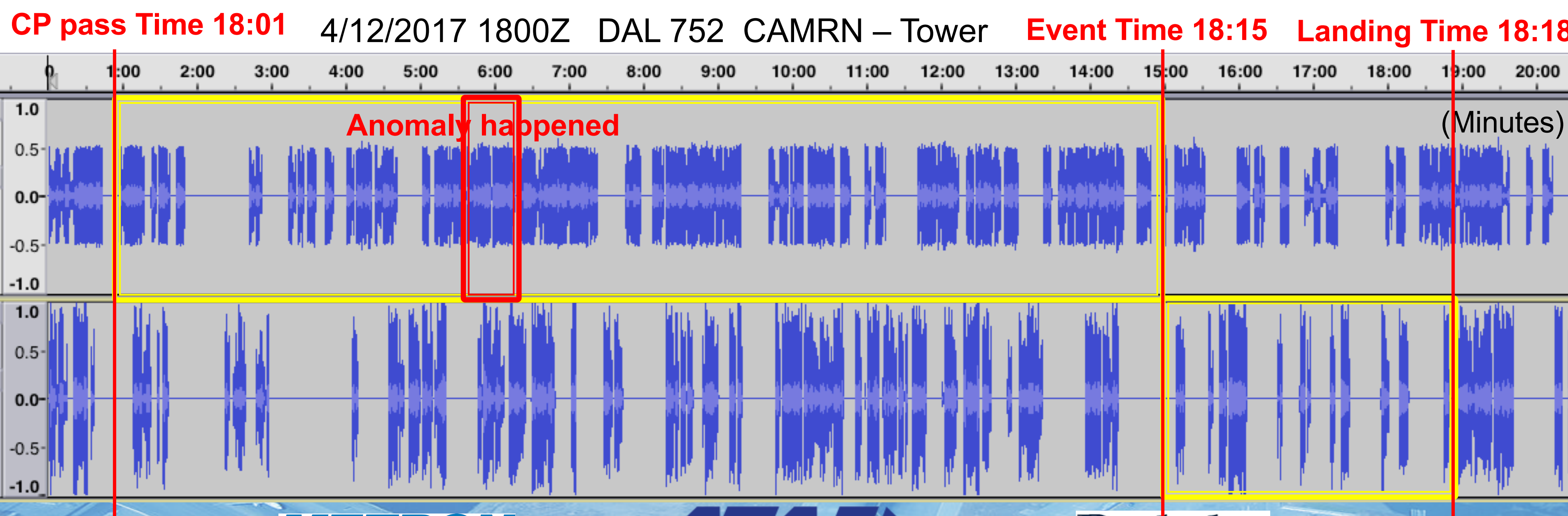
Segmentation – Each **segment** contains only one speaker

02

Flight-Voice Features

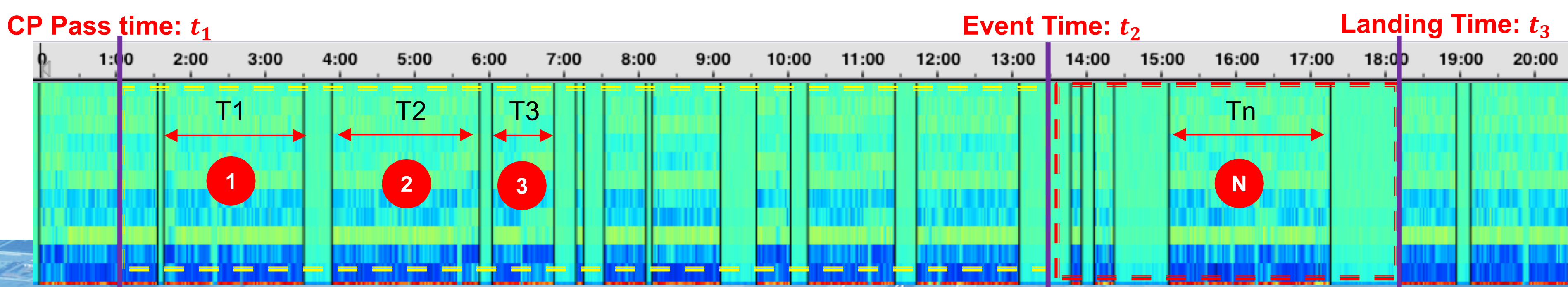
Flight-Voice Feature Analysis

- Three key timestamps identified for each flight operation:
  - Corner post passing time
  - Event time: time to pass intercept
  - Landing time: time to land
- Extract flight-level features from voice data for every flight:
  - TRACON channel: from CP time to event time.
  - Tower channel: from event time to landing time.
- Case study for one specific anomalous flight



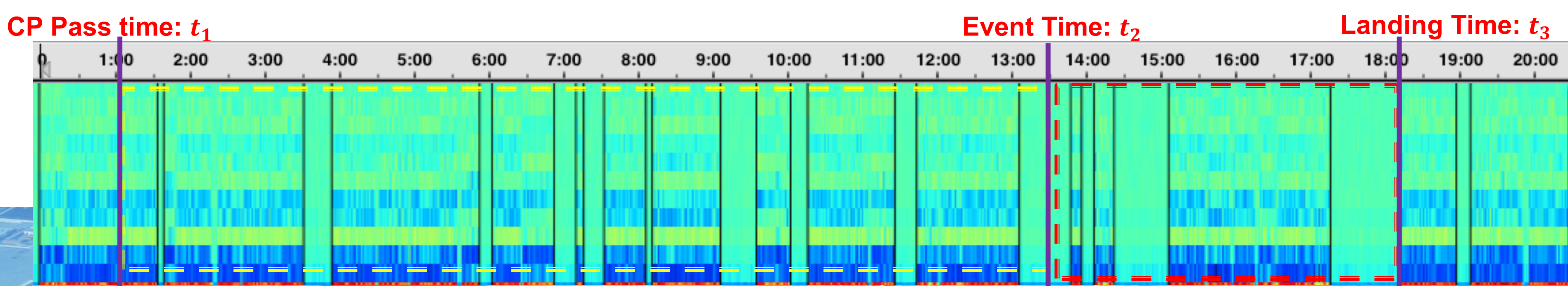
Flight-Voice Feature Analysis

- Approach
  - The total number of events per unit time within a flight time window, λ
  - The average duration (μ) of voice activities (events) within a flight time window
- Calculation
  - N_{tracon} = number of voice communications in time interval $[t_1, t_2]$.
  - N_{twr} = number of voice communications in time interval $[t_2, t_3]$.
  - $\lambda_{tracon} = \frac{N_{tracon}}{t_2 - t_1}; \lambda_{twr} = \frac{N_{twr}}{t_3 - t_2}$
  - $\mu_{tracon} = \frac{\sum_i^{N_{tracon}} T_i}{N_{tracon}}; \mu_{twr} = \frac{\sum_i^{N_{twr}} T_i}{N_{twr}}$



Flight-Voice Feature Analysis

- Calculation
  - Summarize voice energy statistics every second.
    - Max, avg, 75q, 90q of energy statistics for every second (~25 frames).
    - Each voice tape will have a feature matrix with dimension (1800, 4).
  - Map every flight's time windows $[t_1, t_2]$ and $[t_2, t_3]$ to feature matrix. Compute:
    - Average audio energy within the time window.
    - Max, min, 25q, 50q, 75q, 90q, avg of the within-second-avg.
    - Max, min, 25q, 50q, 75q, 90q, avg of the within-second-max.
    - Max, min, 25q, 50q, 75q, 90q, avg of the within-second-75q.
    - Max, min, 25q, 50q, 75q, 90q, avg of the within-second-90q.



03

Pilot-Controller Identification

Pilot-Controller Identification

- Labeling
  - Use the segmentation results (small λ) to aid us listening to audios.
  - For each segment, assign a label as either pilot (1) or controller (2). All non-speech segments will be assigned as 0.
  - For each labeled segment, assign its label to all frames in the segment.

<table>
<tr><td>Segments</td><td colspan="10">Segment 1: controller (2)</td><td colspan="8">Segment 2: pilot (1)</td><td colspan="4">Silence (0)</td><td colspan="9">Segment 4: controller (2)</td></tr>
<tr><td>Frames</td><td>2</td><td>2</td><td>2</td><td>2</td><td>2</td><td>2</td><td>2</td><td>2</td><td>2</td><td>2</td><td>1</td><td>1</td><td>1</td><td>1</td><td>1</td><td>1</td><td>1</td><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td><td>2</td><td>2</td><td>2</td><td>2</td><td>2</td><td>2</td><td>2</td><td>2</td><td>2</td></tr>
<tr><td></td><td colspan="10">All frames belonging to segment 1 will be labeled as controller</td><td colspan="21"></td></tr>
</table>

- Build Classifier
  - Training
    - Build a classifier to predict the label for each frame, using 123 dimensional features (filter bank and FOS and SOS).
  - Testing
    - Predict the label for each frame of the audio clip(s).
    - Apply segmentation algorithm to audio clip(s).
    - For each segment, the final label will be the majority of the frames' label.

<table>
<tr><td colspan="12">Segment 1</td><td colspan="8">Segment 2</td><td colspan="6">Silence (by VAD)</td><td colspan="9">Segment 4</td></tr>
<tr><td>2</td><td>2</td><td>1</td><td>1</td><td>2</td><td>2</td><td>1</td><td>1</td><td>2</td><td>2</td><td>2</td><td>2</td><td>2</td><td>2</td><td>1</td><td>1</td><td>1</td><td>1</td><td>1</td><td>2</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td><td>2</td><td>2</td><td>2</td><td>2</td><td>1</td><td>2</td><td>2</td><td>2</td></tr>
<tr><td colspan="12">2: controller</td><td colspan="8">1: pilot</td><td colspan="6">0: vacant</td><td colspan="9">2: controller</td></tr>
</table>

Pilot-Controller Identification

- Manually label 3 audio clips, each of which covers a 30-minute ATC tower communication.
- Select two labeled audio clip (4/28/2017 1830 Z & 4/28/2017 1800 Z) as training set and one (5/26/2017 2030 Z) as testing set.

Classifier	Frame-wise accuracy	Segment-wise accuracy	Pros	Cons
Logistic regression	75.0%	75%	Easy to train	Loss of temporal relations Hard to update
Linear SVM	75.3%	74%		
BiRNN	87.3%	78%	Easy to update with new data Capable of transfer learning (e.g., speech to text)	Hard to train

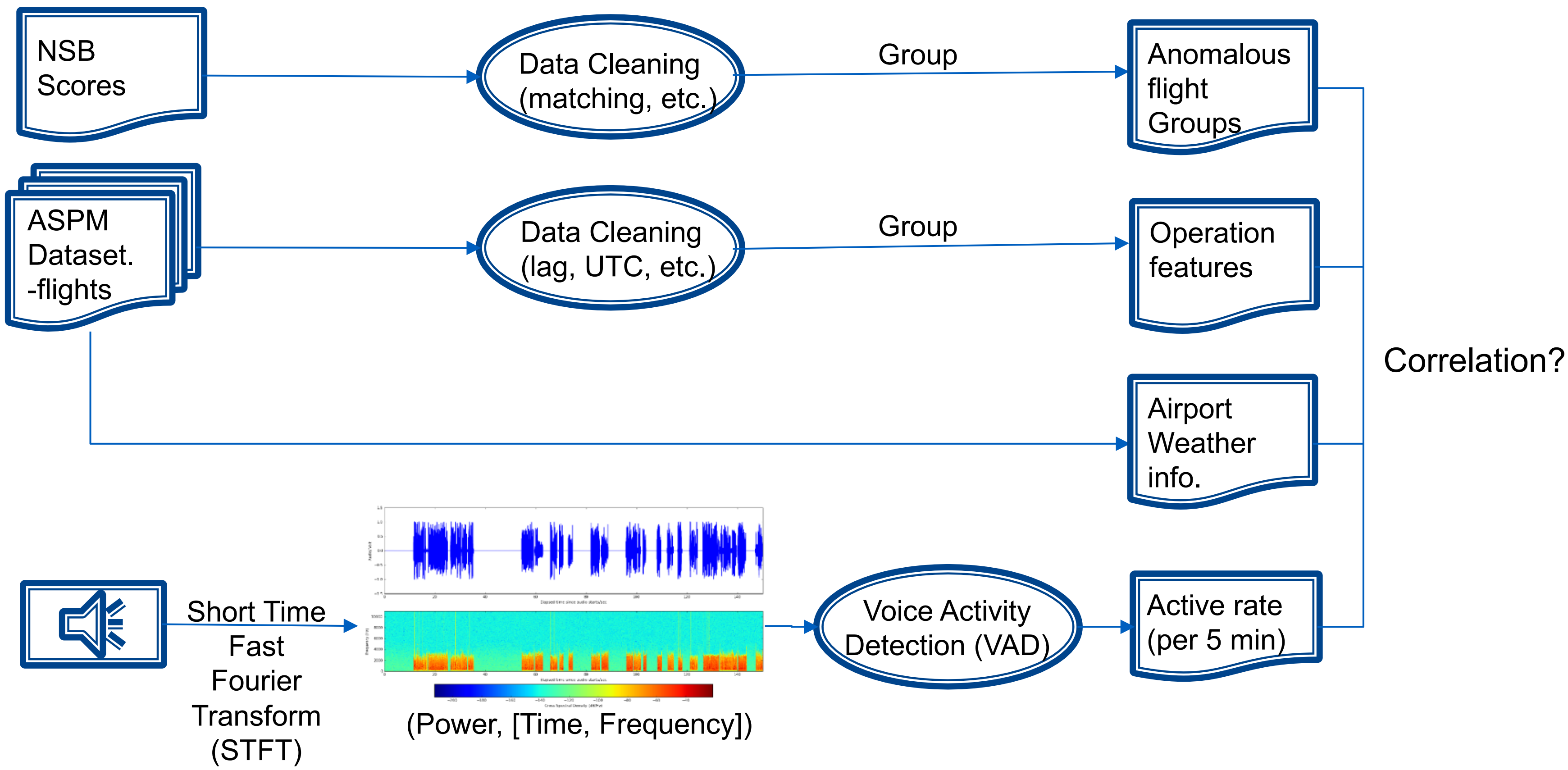
- Further experiments are required to validate our results – coincidently, there is a woman controller in both the training audio clips (two on 4/24/2017) and testing audio (one on 5/26/2017).

04

Channel Occupancy Analysis

Channel Occupancy Analysis

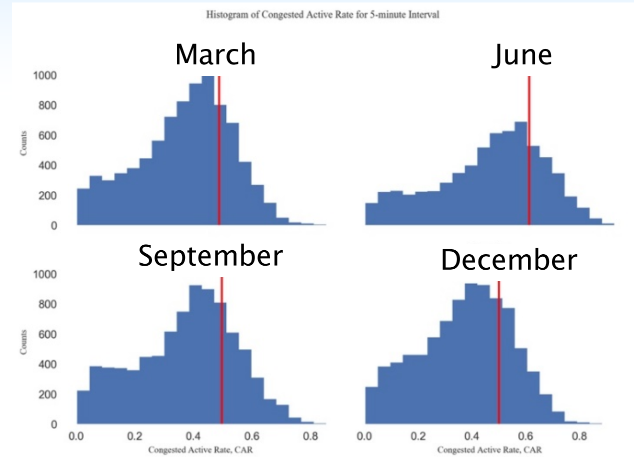
Data matching for each 5-minute time period:



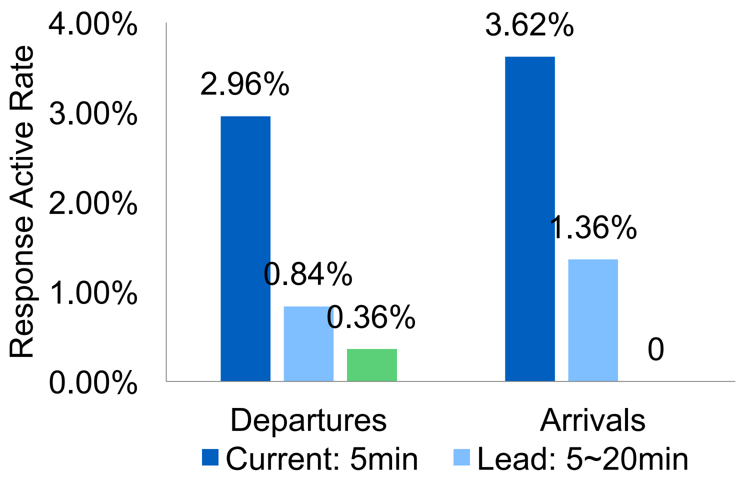
METRON SCIENTIFIC SOLUTIONS
ATAC
Berkeley UNIVERSITY OF CALIFORNIA

Channel Occupancy Analysis

- (Constrained) Active Rate: the percentage of time a communication channel is utilized within time interval
- Result
  - Right censored threshold limit for ATC voice communication is 60.69%
  - Arrivals have stronger impact on the active rate and the leading effect dissipates over time
  - Higher visibility decreases active rate
  - Positive daytime effect
  - Stronger winds lead to more voice activities. Tailwind speed has the strongest impact
  - Flights with high NSB scores require more communication
  - Runway configuration fixed effect increase the active rate as the runway utilization decreases



Incremental Effect of Active Rate with one flight operation adding in different period



Analysis of Go-arounds

Go-around Analysis

- Deeper look into special anomaly events, such as go-arounds
- Study period: 2018/04/01 – 2018/09/30 (JFK), with 445 go-arounds and 101,932 non go-around flights
- Predict Go-Arounds based on features selected from PCA dedicating to analyze both quantitative and qualitative variables (Pagès 2004)
- Estimate logistic regression model
  - Dependent variable: whether a flight is a go-around
  - Independent variables: principal components formed by features
- Varimax Rotation is done for interpreting the effects of each components
- Quantify the contributions of causal factors

Go-around Analysis

- Intercept with final approach features
  - *DIST_AT_INT, ANGLE_AT_INT, INT_RUNWAY_DIST, INT_TYPE = Int Outside Gate* have positive impact on go-around probability
  - *FinalApproachCylinder(-), GlideslopeAtIntercept(-), INT_TYPE=Int Inside FAF* have negative impact on go-around probability
  - *ALT_DIFF_AT_INT, MAX_VERT_FT, MAX_HORIZ_FT* have extremely small positive impact on go-around probability (coef. ≈ 0)

Go-around Analysis

- Separation Feature
  - Incremental effect of go-arounds with 1nm adding in different segments

(nautical mile)

0	Overtake(+)_1	1	Overtake(+)_2	2.5	Overtake(+)_3	5	Overtake(+)_4	8	Overtake(+)_5
	−4.82		−0.94		−0.07		−0.03		−0.00

  - The difference between theoretical (required) separation and real separation increases the probability of go-arounds
    - Theoretical separation: FAA Wake Separation Standards based on weight class pair
    - Real separation: for each aircraft leading-trailing pair, resample and interpolate the time series of positions (latitude, longitude, altitude), then get the minimum separation between two trajectory segments

Go-around Analysis

- Visibility Feature
  - Incremental effect of go-arounds with 1nm adding in different segments

(statute mile)

0	VISIBLE_1	3	VISIBLE_2	5	VISIBLE_3	10
	−0.21		−0.11		−0.10	

  - Go-arounds less likely under visual conditions
- Weight Class

Variable	Coef.
WC_LEAD=F	-
WC_LEAD=H	0.43
WC_LEAD=L	-
WC_LEAD=N	1.08
WC_LEAD=S	-1.01

Variable	Coef.
WC_TRA=F	-0.46
WC_TRA=H	0.62
WC_TRA=L	-0.29
WC_TRA=N	-
WC_TRA=S	-2.95

Go-around Analysis

- Winds
  - Strong tailwind increases the probability of go-arounds
- Agglomeration Effect
  - The number of go-arounds in the 30-minute window, surrounding the landing time of aircraft, has strong impact in increasing go-arounds
  - The time interval between the final approach start time and the closest go-around time, in minutes, weakly decreases the probability of go-arounds
  - The number of aircrafts intending to arrive for the 15-minute period has positive impact on go-around probability

Next Steps

Next Steps

- Complete Development of Anomaly Detection System (Version 1.0)
  - Additional SME involvement through review of energy and voice metric features
  - Finalize voice metrics to include in anomaly detection
  - SME review of high emery feature outliers
  - Develop initial go-around prediction model
- Implement Phase I – Migrate anomaly detection to Sherlock
  - Create one year training set for anomaly detection model
  - Deploy anomaly detection software to Sherlock Big Data System
  - Configure data flows for overnight update
  - V&V of data
- Prepare for Phase II – Integrate with ATM-X Testbed
  - Meetings with ATM-X testbed personnel
  - Determine best design for testbed plug in adapter and Webservice
  - Configure testbed connection
  - V&V of data