Big Aviation Data Mining for
Robust, Ultra-Efficient Air Transportation

Technical Monitor:
Sarah D’Souza, Systems Analysis Office, NASA Ames Research Center

MIT International Center for Air Transportation

NASA LEARN
Phase 1 Outbrief
16 February 2016
Air Transportation System Challenges

- Air transportation system is very safe, but efficiency & robustness challenges remain.

- Most inefficiencies caused by capacity & demand imbalances at range of spatial & temporal scales.

### Millions of departures / % on time

*BTS Annual*

<table>
<thead>
<tr>
<th>Year</th>
<th>Millions of Departures</th>
<th>% On Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006</td>
<td>4</td>
<td>60%</td>
</tr>
<tr>
<td>2008</td>
<td>4</td>
<td>60%</td>
</tr>
<tr>
<td>2010</td>
<td>4</td>
<td>60%</td>
</tr>
<tr>
<td>2012</td>
<td>4</td>
<td>60%</td>
</tr>
<tr>
<td>2014</td>
<td>4</td>
<td>60%</td>
</tr>
</tbody>
</table>

*NY arrivals*
*NY departures*
*PHL arrivals*
*PHL departures*
*BOS, DC ops*
System Planning

Demographics, economics

Resources, procedures

FAA, Airports

Airlines

Networks, capital, schedules

Air Traffic Control (ATC) Operations

Strategic

Weather forecast

Constraint, capacity forecast

Tactical

Plans

Tactical response & execution

Traffic management

Analytics

NAS structure, resources

Data

Delays, cancellations

Trajectories, resource use

National Airspace System (NAS) ...in a single slide
Space, Time, Data, and Impacts

Goal: Demonstrate Big Data analytic framework for aviation across spatial/temporal scales
## Data Descriptions

<table>
<thead>
<tr>
<th>Data Description</th>
<th>Spatial Extent</th>
<th>Spatial Resolution</th>
<th>Temporal Extent</th>
<th>Temporal Resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Planning</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flight operations</td>
<td>NAS-wide</td>
<td>Airport pair (&gt;300 BTS airports)</td>
<td>2000 - 2014</td>
<td>Annual</td>
</tr>
<tr>
<td><strong>Strategic ATC Operations</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flight delays, cancellations</td>
<td>NAS-wide</td>
<td>Airport pair (&gt;300 BTS airports)</td>
<td>2008 - 2014</td>
<td>Annual, Seasonal, Daily, Hourly</td>
</tr>
<tr>
<td>Traffic Management Initiatives</td>
<td>NAS-wide</td>
<td>N/A</td>
<td>2008 - 2014</td>
<td>Daily</td>
</tr>
<tr>
<td><strong>Tactical ATC Operations</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flight trajectories</td>
<td>Regional (NY, DFW, SFO metro)</td>
<td>~5 miles</td>
<td>2013 - 2015</td>
<td>1 minute</td>
</tr>
<tr>
<td>Weather radar mosaics</td>
<td>Regional (NY, DFW, SFO metro)</td>
<td>1 km</td>
<td>2013 - 2015</td>
<td>2.5 minute</td>
</tr>
<tr>
<td>Convective weather impacts</td>
<td>NY metro</td>
<td>Individual route</td>
<td>2013 - 2015</td>
<td>5 minute</td>
</tr>
<tr>
<td>Terminal wind impacts</td>
<td>NY metro</td>
<td>Individual terminal</td>
<td>2013 - 2015</td>
<td>hourly</td>
</tr>
</tbody>
</table>

*BTS = Bureau of Transportation Statistics*
Anatomy of the Big Data Analysis Framework

Enable insights, applications, solutions

- Identify anomalies, interesting behaviors
- Develop descriptive, predictive models
- Evaluate performance, identify best practices

---

Size of data...

‘Raw’ system data

Aggregate into compact mathematical representation

Derive descriptive metrics

Identify patterns of system behavior

Size of insight...

---

Analytics must be scalable, generalizable, and interpretable

---

Big Data for Aviation - 7
MITLL 2/16/16
• Motivation: Air transportation system challenges and Big Data opportunities

• Technical approach & Selected results:
  – Strategic ATC Operations
  – Tactical ATC Operations
  – Airline Network Planning

• Summary of innovations, Potential impacts and Next step recommendations

• Distribution / Dissemination & Acknowledgements
Space, Time, Data, and Impacts

Scale / scope of impact

NAS+  NAS  Regional  Local

Planning / operational horizon

Minutes  Hours  Days  Years

Strategic ATC Operations

Delay
Cancellation
Traffic management planning
National / regional weather
Route planning, etc.

Tactical ATC Operations

Airline Network Planning
FAA System Planning

Big Data for Aviation - 9
MITLL 2/16/16
NAS-Wide Operational Network

At a glance...

Airport Connections

Links colored by delay

Normalized departure delay

(airport pair)

Normalized departure delay

BOS

ATL

SFO

LAX

MIA

DFW

ORD

NYC

WAS
Strategic ATC Operations: Analyzing the NAS-Wide Network

**Adjacency matrix**

**Demand-weighted adjacency matrix**

**Delay, cancellation weighted adjacency matrix**

**Hub, authority metrics:**
Asymmetrical propagation of delay, cancellation

**Application:**
Propagation of weighting metric (delay, cancellation, etc.)

**KEY:**
- Airport
- Flight connection

**Eigencentrality:**
- Airport connectivity
- Airport throughput

**Application:**
- Network structure
- Network capacity

**Goal:** Characterize and model NAS-wide network dynamics and performance

**Approach:** Apply novel adjacency matrix weightings and metrics to define NAS-wide states that characterize propagation of disruptions
Delay State Identification: Methodology

Flight delays, cancellations (2008-2014)

Aggregate (daily, hourly) weighted connectivity matrices (delay, cancellation)

Calculate Hub, Authority scores for major airports

Cluster into propagation patterns

Daily Delay / Cancellation States
Post-event performance evaluation

Hourly Cancellation States
Dynamic delay propagation for predictive modeling

Framework key:
- Raw data
- Aggregate
- Metrics
- Patterns
- Insights
Delay Distribution by Daily Delay State
Selected (5 of 12) *Persistent* Delay States (2008-2014)

**Daily Delay States provide insights into the scale and propagation of delay**
Total delay is similar (but propagation is not) in single-airport dominated states.
Total delay in NAS-wide states tends to the extremes.
Hourly Delay States
Capturing Dynamics of Delay Propagation

- Hourly Delay States capture delay *propagation structure, magnitude, and trends*
  - Local delays build and spread
  - Propagation is widest as delays peak and begin decrease

- Observed Hourly Delay State *transition probabilities, and dwell times* can be calculated
Day Delay Cancelled

July 26, 2012

26808 hours
554

Avg: 2008-2014
13054 hours
295

Normalized departure delay (airport pair)
Network Dynamics Case Study
26 July, 2012

NY Ground Delay Program (GDP) to reduce demand as thunderstorms impact local operations

NY GDP continues & delays persist and propagate as weather dissipates and major traffic corridors clear
Network Dynamics Case Study
26 July, 2012

Delays rapidly increase storms bisect the NAS (but coastal corridor remains clear)

Delay growth and propagation appear to be driven by weather-related airspace constraints and control decisions with long time constants

Delay State dwell times, transition probabilities provide insight into NAS system response times
Strategic ATC Operations: Next Steps

Delay Propagation Modeling
*Markov Jump Linear System*

\[ \bar{x}(t + 1) = \Gamma_m(t)\bar{x}(t) \]

\[ \pi_{ij}(t) = \text{pr}[m(t + 1) = j|m(t) = i] \]

- \( \bar{x}(t) \): Vector of airport delays at time \( t \)
- \( m(t) \): Delay state at time \( t \)
- \( \Gamma_m(t) \): Delay-state dependent system matrix
- \( \pi_{ij} \): Probability of transition from delay state \( i \) to state \( j \)

Delay / demand prediction modeling
Control strategy assessment
Space, Time, Data, and Impacts

- **Flight trajectories**
- Local / regional
- high resolution
- weather
- Tactical decisions
- Etc.

Tactical ATC Operations

Strategic ATC Operations

Airline Network Planning
FAA System Planning

Planning / operational horizon

- Minutes
- Hours
- Days
- Years

Scale / scope of impact

NAS+
NAS
Regional
Local
Tactical ATC Operations
NY Metro Focus

NY Metro Arrival Trajectories

Fair weather operations

Convective weather operations

Key:
LGA
EWR
JFK

Goal: Develop a generalizable method to characterize tactical use of terminal and transition airspace to guide airspace design and support operational best practices

Approach: Identify patterns of arrival / departure resource use through trajectory analysis and link them to constraints and outcomes

‘arrival (departure) resource’ = routinely used arrival (departure) path
Tactical ATC Operations: Methodology

**Resource Identification**
- Cluster trajectories using DBSCAN

**Resource Use**
- Assign trajectories to resources using Random Forest & identify non-conforming trajectories

**Operational Patterns**
- Cluster Resource Use Vectors to identify patterns of hourly use

**Framework key:**
- Raw data → Aggregate → Metrics → Patterns → Insights

- 13 day training set
- 57 day weather impact dataset
- 1000 day pattern dataset (2013-2015)

**Daily Resource Use Matrices**
- Post-event analysis of operational dynamics

**Hourly Resource Use Vectors**
- Real time operational dynamics

**Hourly Resource Use Patterns**
- Predictive modeling
Resource Identification

‘Emergence’ of EWR Arrival Resources

13 days of arrivals...
...23 clusters...
...23 cluster centroids = Arrival Resources

- Cluster algorithm parameterization involves tradeoffs between compactness, separability, and dissimilarity of clusters
- Resulting clusters captured ~92% of all trajectories
Random Forest trajectory classification assigns individual trajectories to resources and identifies non-conforming trajectories.

Non-conforming trajectories take many forms:
- Dynamically alter flow structure
- Workload consequences for Air Traffic Control?
Non-conformance and Weather

- Trajectories assigned for dataset of 56 days including weather impacted (convection or adverse winds / ceiling / visibility) and fair weather days
- Significant increase in non-conforming trajectories during weather impacted days
NY Metro Operational Dynamics
A Tale of Two Days… (EWR Arrivals)

Resource Use Matrix

Full day summary

Period of convective impacts

October 8, 2014: Fair weather

July 14, 2015: Convective impacts
Hourly Resource Use Patterns (RUP)

RUP 1: Departure

Average Hourly Number of Flights

RUP 2: JFK, EWR Arrival

Average Hourly Number of Flights

EWR Arrivals

EWR Departures

Big Data for Aviation - 27
MITLL 2/16/16
Hourly Resource Use Patterns (RUP)

RUP 3: Arrival / Low Throughput

Average Hourly Number of Flights

RUP 4: Very Low Throughput

Average Hourly Number of Flights
Hourly Resource Use Patterns (RUP)

RUP 5: High Demand / High Throughput

Average Hourly Number of Flights

RUP 6: High Demand / High Non-conformance

Average Hourly Number of Flights

EWR Arrivals

EWR Departures

JFK ARR

JFK DEP

EWR ARR

EWR DEP

LGA ARR

LGA DEP

NC

NUMBER OF FLIGHTS

0 5 10 15 20 25 30 35
Occurrence of Resource Use Patterns
By Hour

**Clear Weather**
(427 days)

- High demand, high throughput
- Departure
- Arrival / Low throughput
- JFK / EWR arrival
- Very low throughput
- High demand / High non-conformance

**Convection / Rain**
(523 days)

High non-conforming (High Throughput) RUP observed more (less) frequently on days with measurable convection / rain impacts
Tactical ATC Operations: Next Steps

Resource Use Matrices

Weather impact / constraint

Clustering to identify days with similar constraints, resource use

Constraint-normalized performance assessment

Case day identification / scenario generation

Daily Aggregations

Hourly Aggregations

Resource Use Patterns

Weather impact / constraint

Correlation of Resource Use Patterns with constraints, demand

Constrained capacity modeling and prediction for decision support

Development of best practices
Space, Time, Data, and Impacts

Planning / operational horizon

Minutes  Hours  Days  Years

Scale / scope of impact

Local  Regional  NAS  NAS+

Strategic ATC Operations

Tactical ATC Operations

Airline Network Planning
FAA System Planning

Schedule
Economic
Demographic
Climatology, Etc.
Air Carrier Competition: Methodology

2000 - 2014

Extract all city pairs

Identify top 40 routes
Calculate # of flights, # of airlines on each

Define use, competition network structures

Annual Route Use,
Competition Networks
Inputs to Strategic Operations analyses
Basis for predictive models to guide capital investment

Number of City Pairs

Number of Flights on route (x 10^4)
Top 40 Routes
By number of operations

Annual number of departures

2006 2007 2008
2009 2010 2011
2012 2013 2014

$\times 10^4$
Competition on Top 40 Routes

Number of airline operators

Number of flight operators

2006 2007 2008

2009 2010 2011

2012 2013 2014

Big Data for Aviation - 35
MITLL 2/16/16
Air Carrier Competition: Next Steps

Network operations

Market competition

Characterize operational and competitive network structure as weighted connectivity matrices

Effect of structure on annual NAS performance measured by delay, cancellation

Correlation to observed frequency of Delay, Cancellation States
Outline

• Motivation: Air transportation system challenges and Big Data opportunities

• Technical approach & Selected results:
  – Strategic ATC Operations
  – Tactical ATC Operations
  – Airline Network Planning

• Summary of innovations, Potential impacts and Next step recommendations

• Distribution / Dissemination & Acknowledgements
• Developed Big Data analysis framework using novel metrics & analytics to provide new insight across a range of fundamental scales in air transport:

<table>
<thead>
<tr>
<th>Aggregate</th>
<th>Metrics</th>
<th>Patterns</th>
<th>Insights</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tactcal ATC Operations</td>
<td>• Terminal area trajectory clustering under range of operating conditions</td>
<td>• Assignment of trajectories to standard resources</td>
<td>• Identification of small number of key resource use patterns</td>
</tr>
<tr>
<td>Strategic ATC Operations</td>
<td>• Airport-pair delay and cancellation weighted directional connectivity matrices</td>
<td>• NAS network hub and authority scores at range of temporal scales</td>
<td>• Identification of small number of key NAS-wide delay and cancellation states</td>
</tr>
<tr>
<td>Airline/FAA Planning</td>
<td>• Airline network definitions across decades</td>
<td>• Top route and competition evolutions over decades</td>
<td>• Identification of dominant scheduled routes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Competition dynamics</td>
</tr>
</tbody>
</table>

• Insights provide foundation for performance evaluation and predictive models
Phase 1 Innovation & Impact Summary

Phase 1

Data Layer
- Flight trajectories
- Flight delay
- Weather
- Cancellations
- Schedules

Analytics Layer
- Tactical ATC Operations Analysis
- Strategic ATC Operations Analysis
- Airline/FAA Planning Analysis

Application Layer
- Diagnostic system characterization
- Baseline, anomaly, scenario identification

Impact & Tech Transfer Layer
- NASA: technical interchange meetings
- Other: Publications

Phase 2

- Traffic Management Initiatives
- Emerging data types (FAA SWIM, other?)
- Database structure & technology

- Refinements across areas
- Extensions where appropriate

- Predictive modeling
- Control action analysis
- Tool building (visualization & analysis)

- NASA: tools for integration into existing programs
- FAA / Industry: performance analysis
Current & Potential Future Connections to NASA Efforts

• **Tactical Operations / 4D-TBO**: end-to-end modeling of TBO-based traffic management (illustrated)

• **Strategic, Tactical Operations / SMART-NAS Testbed**: real-time analytics and visualization tools
  - Simulation modules
  - Review of archives to identify case studies and define scenarios

• **All / Sherlock Data Warehouse**: information models for analytic products

4D-TBO = 4 Dimensional-Trajectory Based Operations
SMART-NAS = Shadow Mode Assessment Using Realistic Technologies for the National Airspace System

System Planning

Air Traffic Control (ATC) Operations

Strategic

Tactical

Analytics

Structural inefficiencies
Capital needs projection

Performance-driven best practices
(post-event analysis)
Operational decision support
(real-time predictive models)
Outline

• Motivation: Air transportation system challenges and Big Data opportunities

• Technical approach & Selected results:
  – Strategic ATC Operations
  – Tactical ATC Operations
  – Airline Network Planning

• Summary of innovations, Potential impacts and Next step recommendations

• Distribution / Dissemination & Acknowledgements
• Papers
  – “Multi-Scale Data Mining for Air Transportation System Diagnostics”, accepted to *16th AIAA Aviation Technology, Integration, and Operations Conference*, 13-17 June 2016, Washington DC.
  – “Clusters and Communities in Air Traffic Delay Networks”, accepted to *2016 IEEE American Control Conference*, 6-8 July 2016, Boston, MA.

• Presentations
  – “Big Aviation Data Mining for Robust, Ultra-Efficient Air Transportation”, Kick-off Meeting & Overview for NASA ARC Aviation Systems Division researchers, NASA Ames Research Center, 4 April 2015.

• Other
  – Numerous telcons with NASA researchers to discuss potential mutual value from collaboration (including SMART-NAS, 4D-TBO, Sherlock data warehouse programs)
Acknowledgments

• Many thanks to the following:

  – **NARI** for supporting the project and promoting collaboration

  – **Sarah D’Souza and Michael Bloem**, NASA ARC for providing excellent technical oversight and helping connect us to relevant NASA researchers

  – **NASA ARC program researchers** for their invaluable technical discussions, feedback on our approach and identification of relevant problem areas
    * 4D-TBO (Paul Lee, Heather Arneson, Tony Evans, …)
    * SMART-NAS (John Robinson, Kee Palopo, Gano Chatterji, …)
    * Sherlock data warehouse team (Michelle Eshow, Rich Keller, Ron Reisman, …)
    * William Chan (Branch Chief)
    * Sandy Lozito (Division Chief)
Thank you!