Booz | Allen | Hamilton

Bayesian Networks for Departure Delay Prediction

NASA Ames Research Center

Airline Operations Workshop

Alex Cosmas

Chief Scientist

Booz Allen Hamilton

Booz Allen Hamilton and Client proprietary and business confidential

SE2020

TASK ORDER NO. 67, TORP 1543

In support of:

FAA NextGen Advanced Concepts and Technology Development Group

Agenda

+ Project Overview
+ Bayesian Networks
+ SMDP Model Development
+ Questions

Research Overview

- Most existing models that are employed in practice (for instance by the FAA) use simulation techniques, which are based on:
  - Regression / Stochastic / Behavioral Models
  - “Causal Patterns” that are based on theoretical knowledge
  - Iterative, manual, and time-consuming calibration processes

- Several academic studies propose the use of Bayesian modeling techniques for predicting flight delays

- BBNs represent a paradigm shift as they:
  - Have a structure that is machine-learned from data and does not require assumptions about “causal” patterns
  - Can produce estimates even in situations with sparse or limited data
  - Can be used well in advance of the actual flight, as they can predict based on only partial evidence

SMDP represents a paradigm shift in solving the problem of predicting departure time

Booz | Allen | Hamilton

GOAL: Develop a probabilistic model using machine learning algorithms and data mining techniques to improve departure time predictions for real-time TFM in the NAS

Statistical Methods for Departure Prediction (SMDP)

PHASE 1 (2013-2014):

- Developed a Proof of Concept for Boston Logan Intl Airport.
- Used machine learning techniques and 52M flight records to predict departure delays utilizing 47 different variables.

PHASE 2 (2015-2016):

- Update the BOS Model with additional data sets: TFMS and CCFP.
- Develop individual models for the Core 30 Airports.

PHASE 3 (TBD):

- Identify use cases and carry out field tests.
- Develop and test multiple BBN model network.
- Operationalize tool with incoming data feed (e.g. SWIM data) and real-time capabilities.

Agenda

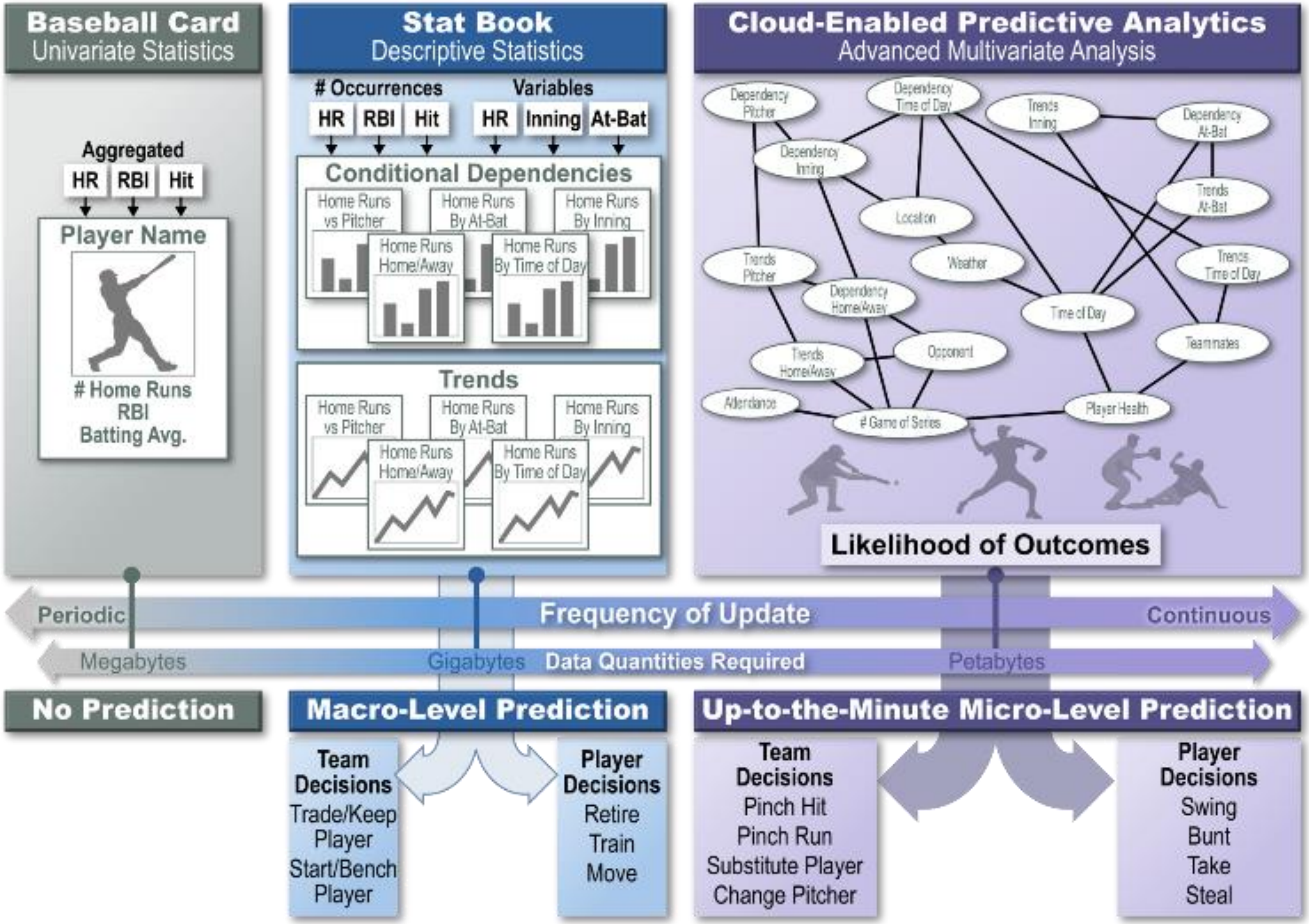
+ Project Overview
+ $\boxed{\text{Bayesian Networks}}$
+ SMDP Model Development
+ Questions

BBNs have historically been a tool for the researcher; their potential is extraordinary as a tool for the business

+ 90% of the world's data was created in the past 2 years
+ That metric is expected to hold true in another 2 years
+ Data Miners produce Snapple cap facts
+ Data Scientists produce insights - they require the intellectual curiosity to ask "why" and "so what"?

Real fact #855: Animals that lay eggs do not have belly buttons

Moneyball 2.0: The data revolution is enabling real-time predictive analysis

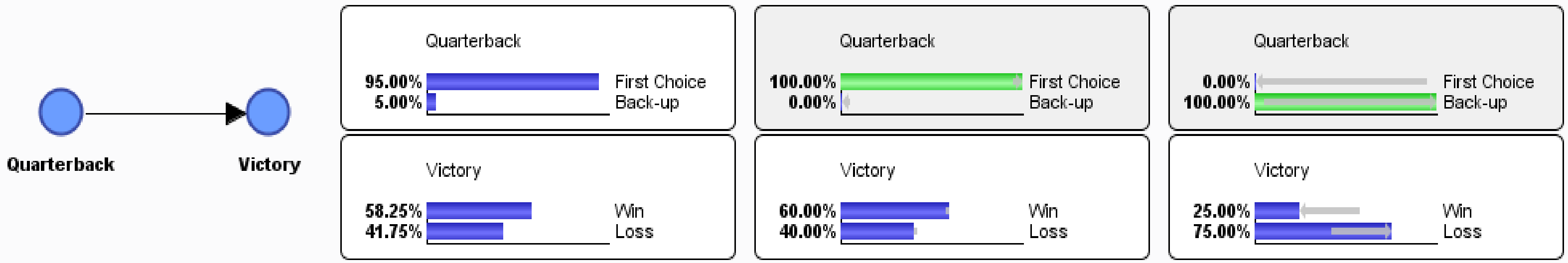


BBN Overview and Use Cases

What are Bayesian Belief Networks?

- A BBN is a graphical model representing the conditional relationships between variables
  - All variables (continuous or discrete) are modeled in terms of probability distributions
  - Relationships between the variables are modeled in terms of the conditional probability tables

Illustration of a Simple BBN



- In the illustrative BBN, variables Quarterback and Victory have two states each and the corresponding probabilities. For example, there's 95% chance of first choice QB opening the game.
- When the status of the playing QB is "known", the distribution function for Quarterback changes, and the effect is propagated through the arc influencing the distribution of the Victory variable

BBN Overview and Use Cases

BBNs offer significant advantages over traditional models:

Traditional Models

Form of the model has to be selected by the modeler

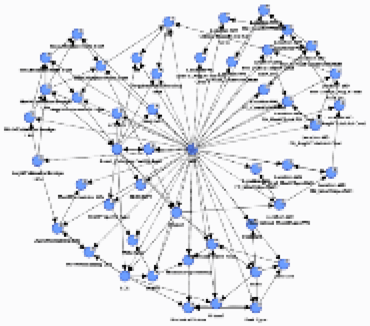
$$Y = \alpha + \sum_{i=1}^{N} \beta_i X_i$$

Only **parameters, not the model form** can be identified based on the data

- For example, linear regression, joint probability analysis, etc.
- One size fits all solution
- Observations with missing data are thrown away
- Variables with non-numerical values such as *"Color of car"* cannot be modeled

BBN Models

	QB: First Choice	QB: Back-up
Win	60%	25%
Loss	40%	75%



Parameters

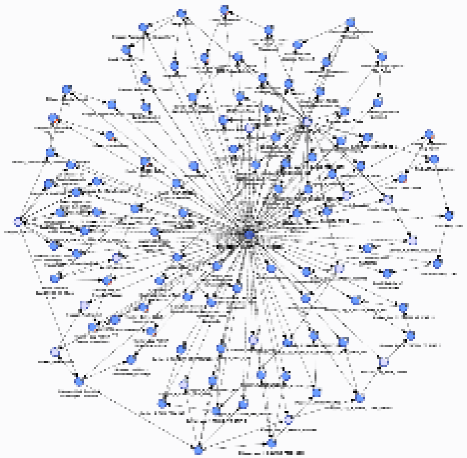
Form

Both learned from the data

- Optimized network structure (form) learned from the data
- Missing values are *inferred*[1] during the machine learning process
- Discretization of variables allows for non-numerical variables to be modeled

[1] Missing values of a variable are inferred from the known values for the same variable from other "similar" observations

BBN Overview and Use Cases



Risk Visualization and Prediction

<u>Goal:</u>
Predict likely locations of serious incidents arising from the transport of hazardous material

<u>Challenge:</u>
Historical incident data spread across disparate sources had many missing values

<u>Data:</u>
113 variables
225,000 rows
Source Data Size: 3 GB

Asset Management

<u>Goal:</u>
Estimate the reliability of thousands of assets

<u>Challenge:</u>
Sparse information on individual assets

<u>Data:</u>
49 variables
83,000 rows
Source Data Size: 10 GB

<u>Operationalized BBNs</u>:

- General Electric (failure detection based on sensor data)
- Intel Corporation (processor fault diagnosis)
- Proctor and Gamble (market research and consumer loyalty)
- SABRE Online Reservation System (bug detection)
- Ministry of Defense, UK (TRACS, military vehicle location software)
- Philips Consumer Electronics (testing process quality and software product quality)
- Inrix Traffic (predicting road traffic flows)
- Microsoft Office Assistant (enabling proactive tips based on user usage)
- Reasoning Under Uncertainty, Monash University (missing person search and rescue)
- National Institute of Water and Atmospheric Research, New Zealand (forest resources management)

BBN Application – Solving The Monty Hall Problem

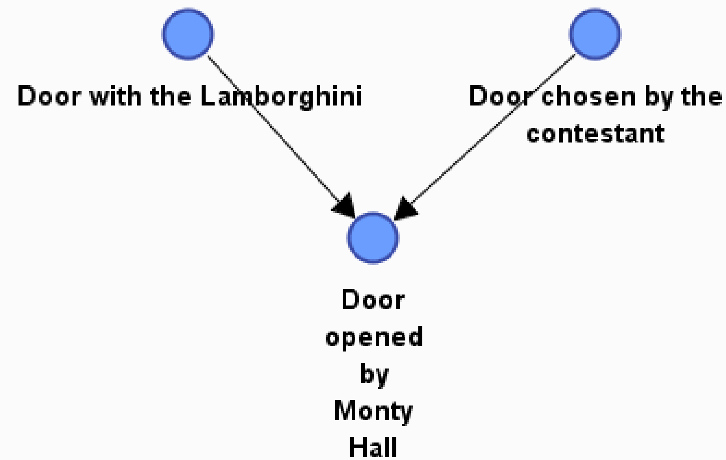
$$p(A\,|\,B) = \frac{p(A,B)}{p(B)} = \frac{p(B\,|\,A)p(A)}{p(B)}$$

$$p(A_i\,|\,E) = \frac{p(E\,|\,A_i)p(A_i)}{p(E)} = \frac{p(E\,|\,A_i)p(A_i)}{\sum_i p(E\,|\,A_i)p(A_i)}$$

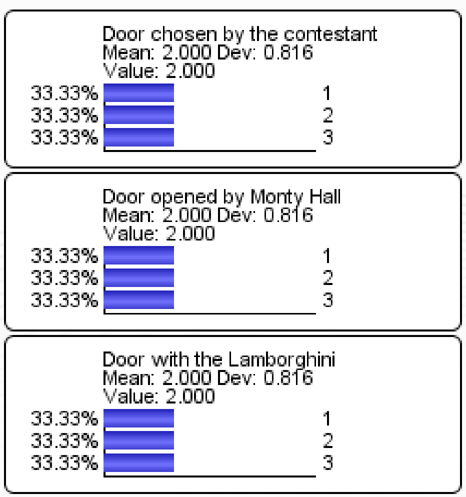
When problem first appeared in *Parade*, approximately 10,000 readers, including 1,000 PhDs, wrote claiming the solution was wrong.

BBN Application – Solving The Monty Hall Problem

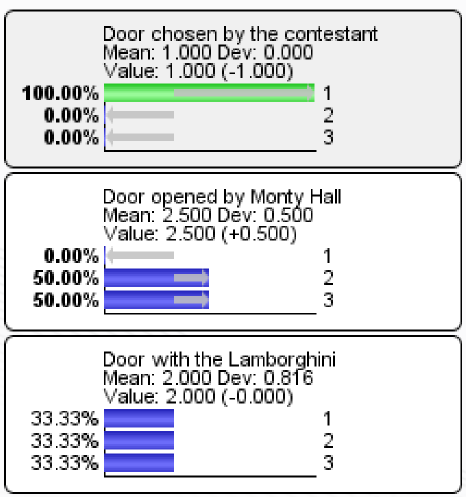
- The BBN model for the Monty Hall Problem has three nodes
  - Door with the car
  - Door chosen by the contestant
  - Door opened by Monty Hall
- Since Monty Hall knows the door with the car and the contestant's choice, that variable is affected by other two variables as signified by the arc



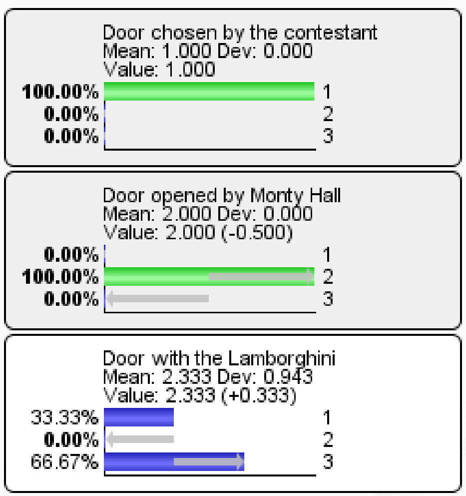
At the start of the show, all three variables have equal chance of being door 1, 2 or 3



After contestant's selection, Monty Hall may choose any of the other two doors



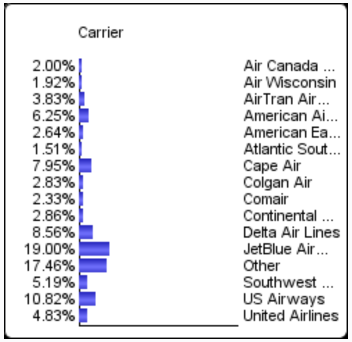
Monty Hall's selection of Door 2 pushes its 1/3 probability to Door 3

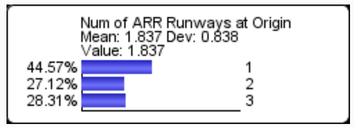


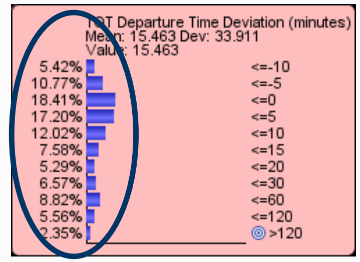
Booz | Allen | Hamilton

Application of the Model for Delay Prediction

- Variables are modeled in terms of probability distribution functions

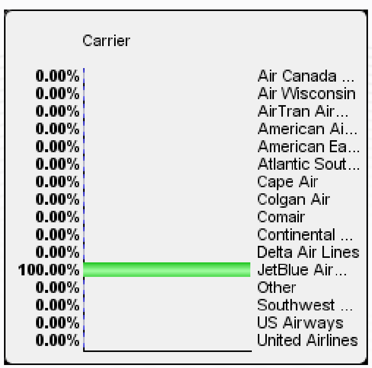


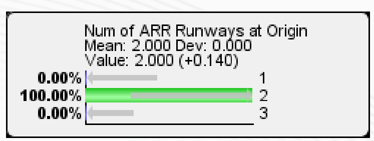


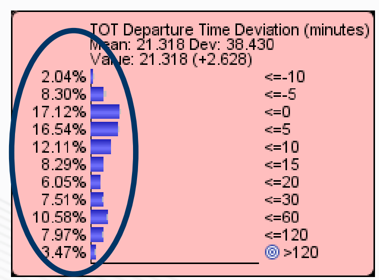


No evidence applied

- The probabilistic relationships between the variables are represented in terms of the conditional probability tables defined by the arcs in the network
- Any change in the information about the variables propagates that information through these arcs of the connected model network and alter the distribution of other variables





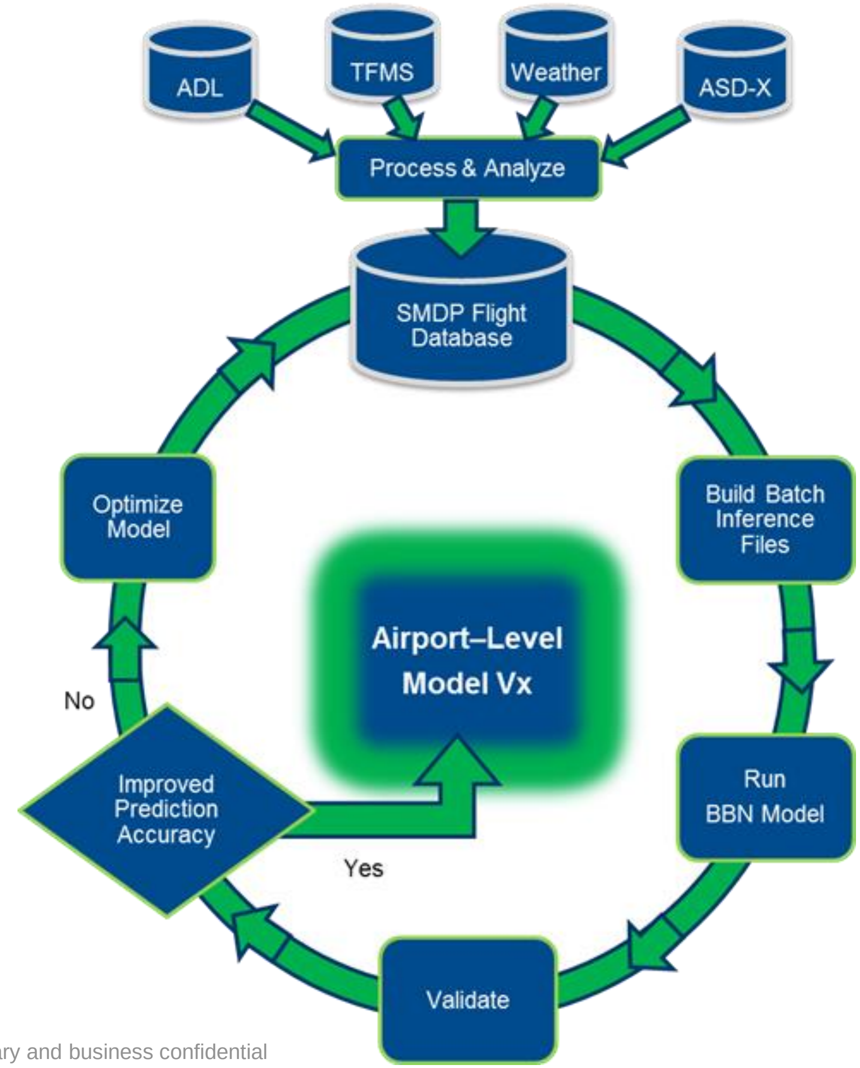


Evidence applied on Carrier and Number of runways

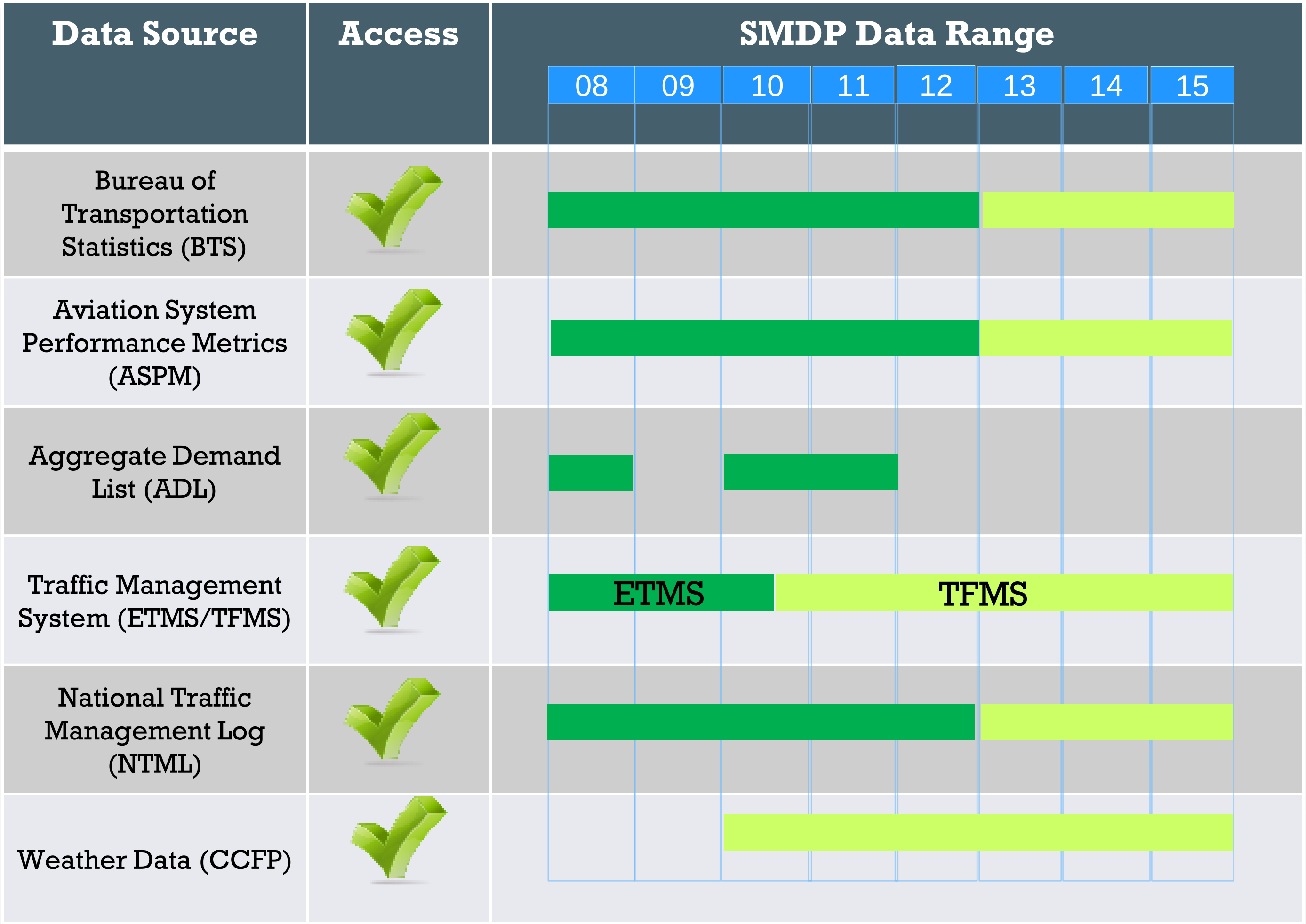
Agenda

+ Project Overview
+ Bayesian Networks
+ $\boxed{\text{SMDP Model Development}}$
+ Questions

SMDP machine learning is based on an iterative process that tests thousands of alternative model structures



The datasets acquired in Phase I (Dark Green) have been integrated with additional years and types of data in Phase II (Light Green)



Phase I

Phase II

SMDP BBN Model Data Elements

ASPM

Departure YYYYMM, Departure Day, Departure Hour, Departure QTR, Arrival YYYYMM, Arrival Day, Arrival Hour, Arrival QTR, OFF YYYYMM, OFF Day, OFF Hour, OFF QTR, ON YYYYMM, ON Day, ON Hour, ON QTR, FAACARRIER, Flight Number, Tail Number, ETMS EQPT, Departure Airport, Arrival Airport, Flight Type, OAG ACID,USER CLASS, Scheduled OUT Time, Flight Plan Departure Time, Actual OUT Time, Nominal Taxi Out, Actual Taxi Out, Scheduled OFF Time, Flight Plan OFF Time, ETMSOFF Time, EDCTOFF Time, Actual OFF Time, DZ Time, GAP DZ, Delay Scheduled OUT, Delay Flight Plan OUT, Delay EDCT, Delay Taxi Out, Delay Scheduled OFF, Delay Flight Plan OFF, Flight Plan ETE, Actual AIR, DZ2AZ, Delay AIR, AZ Time, GAP AZ, EDCT ON Time, Actual ON Time, EDCT ARR Difference, Nominal Taxi In, Actual Taxi In, Scheduled BLOCK, Actual BLOCK, Scheduled IN Time, Flight Plan IN Time, Actual IN Time, Delay Taxi In, Delay BLOCK, Delay Scheduled ARR, Delay Flight Plan ARR, ASQP Reported Carrier Delay, ASQP Reported Weather Delay, ASQP Reported NAS Delay, ASQP Reported Security Delay, ASQP Reported Late Arrival Delay, OPSNET Delay Cause, Departure Wind, Departure Ceiling, Departure Visibility, Departure Nearby TS, Departure Weather, Arrival Wind, Arrival Ceiling, Arrival Visibility, Arrival Nearby TS, Arrival Weather*

TFMS/ETMS

Flight Plan, Flight Plan Amendments, Flight Routing, Track Data, Flight Time data, Cancelations, Controlled Flights, Airline data, Sector loadings, Aircraft data*

SMDP BBN Model

BTS

Year, Month, Day Of Month, Day Of Week, Flight Date, Unique Carrier, Carrier, Tail Number, Flight Number, Origin, Dest, Crs Departure Time, Departure Time, Departure Delay, Departure Delay New, Departure Delay 15,Departure Delay Group, Taxi Out, Wheels Off, Wheels On, Taxi In, Crs Arrival Time, Arrival Time, Arrival Delay, Arrival Delay New, Arrival Del15, Arrival Delay Group, Cancelled, Cancellation Code, Diverted, Crs Elapsed Time, Actual Elapsed Time, Air Time, Distance, Distance Group, Carrier Delay, Weather Delay, NAS Delay, Security Delay, Late Aircraft Delay, First Departure Time, Total Additional Gate Time, Longest Additional Gate Time*

ADL

Flight airline updates, Flight ETD updates, Flight departure fix, Flight departure fix time, Flight arrival fix, Flight arrival fix time . Flight transition fix, Flight control times

Weather

Terminal Area Forecast (TAF), Collaborative Convective Forecast Product (CCFP), Corridor Integrated Weather System (CIWS), Weather Impacted Traffic Index (WITI)*

NTML

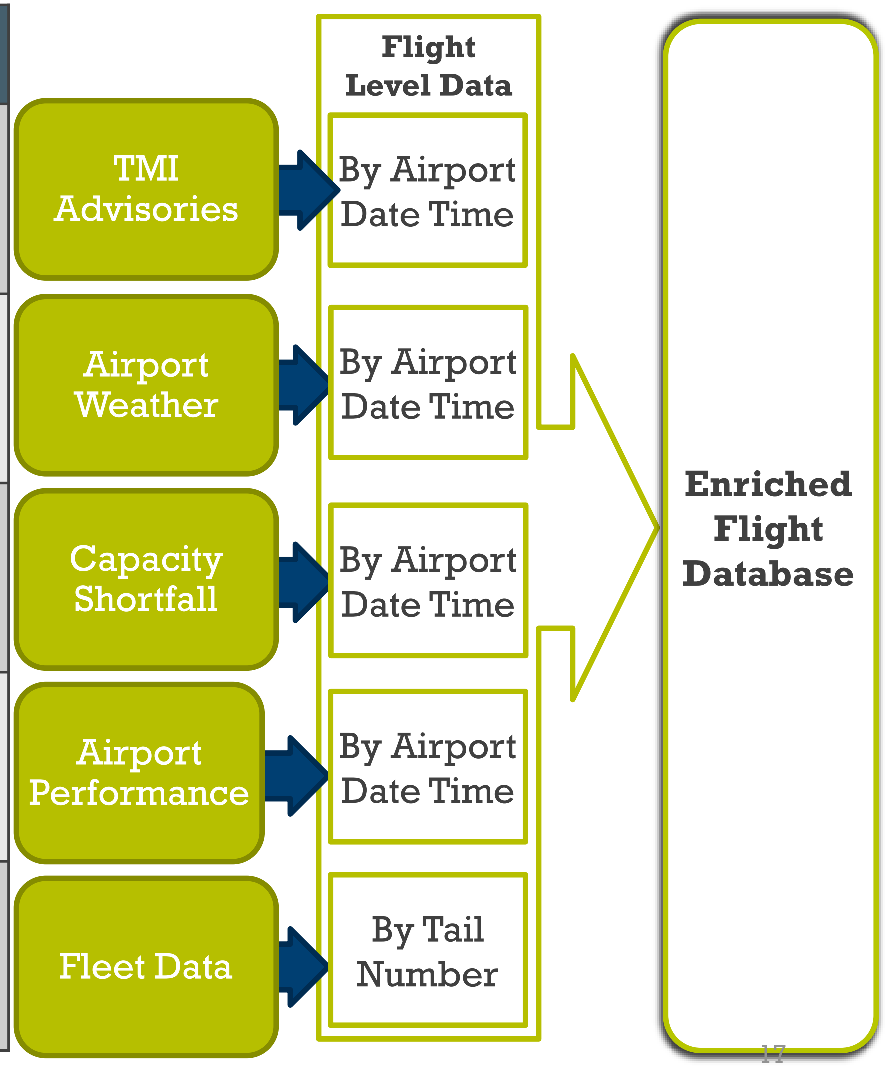
Ground Delay Program, GDP Revision, Ground Stop, Ground Stop Revision, Compression, Airspace Flow Program, Airspace Flow Program Revision, Miles In Trail*

- Data is available and processed (green)
- Partial data and not processed (yellow)
- Data is not available or not processed (red)

Booz | Allen | Hamilton

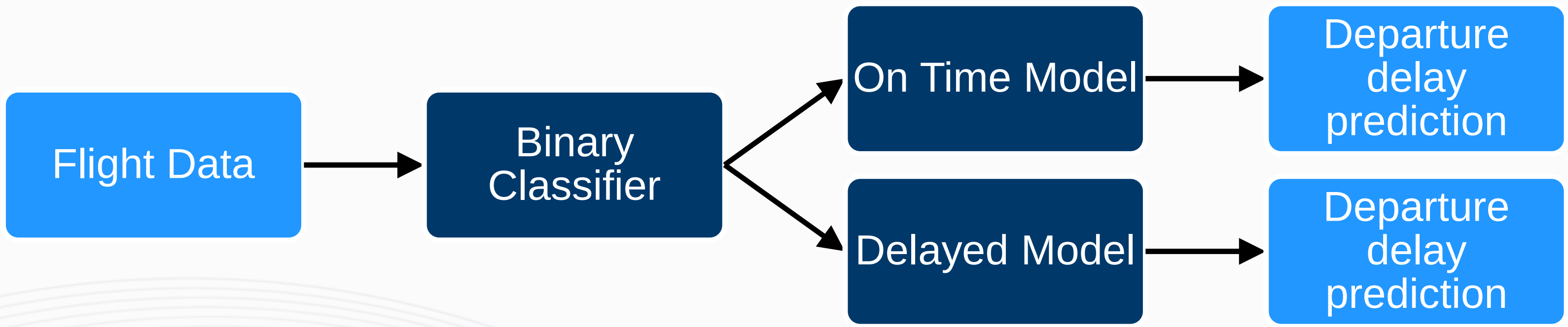
The current Boston model was developed in Phase I by learning information from data integration process outlined below; the model is being updated to include the information provided by the new data sources

Source	Hypothesis	Goal
NTML Data Warehouse	Cause of delay, length of delay are important indicators	Correlate advisory, and TMI program information to flight level data
ASPM Terminal Weather	Weather, and % on-time information are indicators for delay	Correlate weather information to flight level data
ASPM Quarterly Data	Capacity shortfall directly impacts delays	Correlate airport capacity shortfalls to flight level data
ASPM ASQP	There are airport specific delay indicators	Correlate airport on-time characteristics to flight level data
JP Fleets	Age, equipment details correlate well with mechanical issue related delay	Add aircraft specific details to flight level data



Rationale for Split Model Approach

- Our analysis indicates that the intrinsic causal drivers of delay differ for narrowly delayed and severely delayed flights
- <u>On-Time flights</u>:
  - Flights with 15 minutes of delay or less (different thresholds were tested)
  - Usually exhibit regular day's operations and follow historically average trends
- <u>Delayed flights</u>:
  - Flights delayed more than 15 minutes (different thresholds were tested)
  - Usually subjected to few abnormal factors, such as extreme weather or network delays

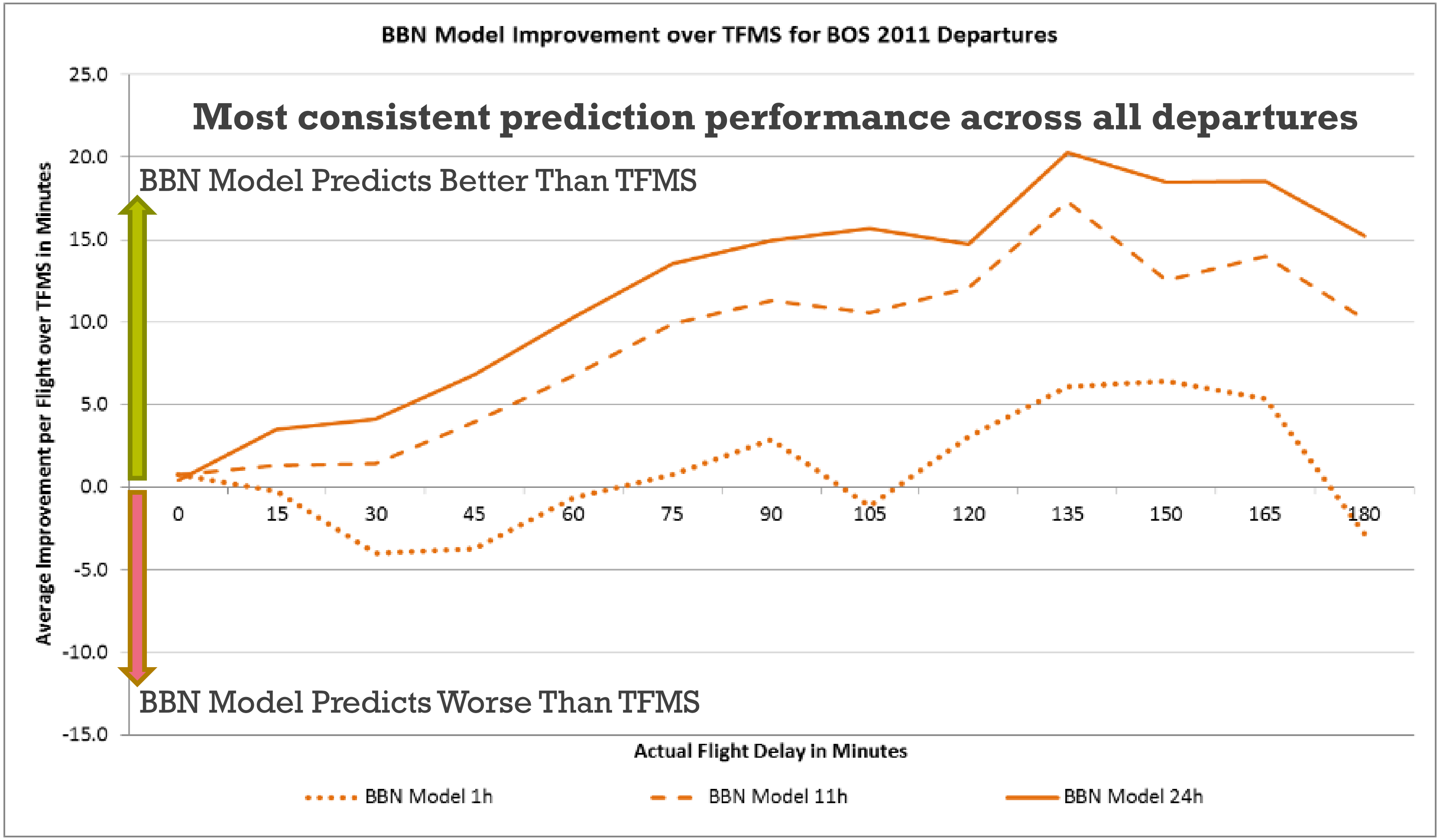


- A two-tiered split model approach where a separate model is trained on two subsets of flight data, each belonging to one type of flights defined above, is adopted
  - Step 1: A binary classifier is used to classify a future flight as On-Time or Delayed flight
  - Step 2: Based on the classification generated in Step 1, the departure delay of the concerned future flight is derived from the appropriate model

BayesiaLab was used to build a BBN model that provided the best representation of the ingested integrated data with minimal complexity and broader generalization ability

- The model takes as input the file with flight-level data for BOS in 2011 (**47 variables**)
- The **target variable** in the model is named ‘TOT Departure Time Deviation”
- The variables were imported into **BayesiaLab** (the modelling software) and discretized
- **4 machine learning algorithms** were tested to develop the final model.

The BBN model produces departure time predictions that consistently outperform the TFMS predictions



Questions