Predicting Flight Delays *with* Azure ML Studio

DEVELOPER WEEK
AUSTIN 2017

whoami

Paige Bailey
Sr. Cloud Developer Advocate
Machine Learning / AI

Work Experience

- Focus at Microsoft is *machine learning* and *artificial intelligence*.
- Prior to joining Microsoft, was a *data scientist* and *geophysical application developer* in the energy industry for 5 years.
- *GIS Technician* (Esri products) for two years.

Toolkit

- Python *(10 years)*
- R *(4 years)*
- Spark, Kafka, Hive, HBase *(2 years)*

Location: *Austin, TX*

Twitter: *@DynamicWebPaige*

Microsoft | Data Platform

whoami

Paige Bailey
Sr. Cloud Developer Advocate
Machine Learning / AI

@DynamicWebPaige

Cloud Developer Advocates

We write, speak, and dream in code. Our global team is maniacal about making the world amazing for developers of all backgrounds. Connect with us, write code with us, and let's meet up and talk cloud and all things developer!

the situation

Microsoft Interview Schedule: Paige Bailey - Confirmed Interview Schedule- [redacted]-Sr. Cloud Developer Advocate - July 13, 2017 Inbox x

Andrea
to PAIGE.BAILEY, Angela

Jul 12

Hi Paige,

Thank you for your time on the phone today. Below is your schedule for tomorrow – please keep in mind that you could be meeting with 1 more person after Bryan. Once your interview day is complete, the team will let you know and at that point you can be on your way. I'll follow up with you over the phone just as soon as I have all of the feedback gathered.

Feel free to reach out if you have any questions. Otherwise, best of luck to you tomorrow and I'll talk to you soon!

United 1157 UAL1157 / UA1157

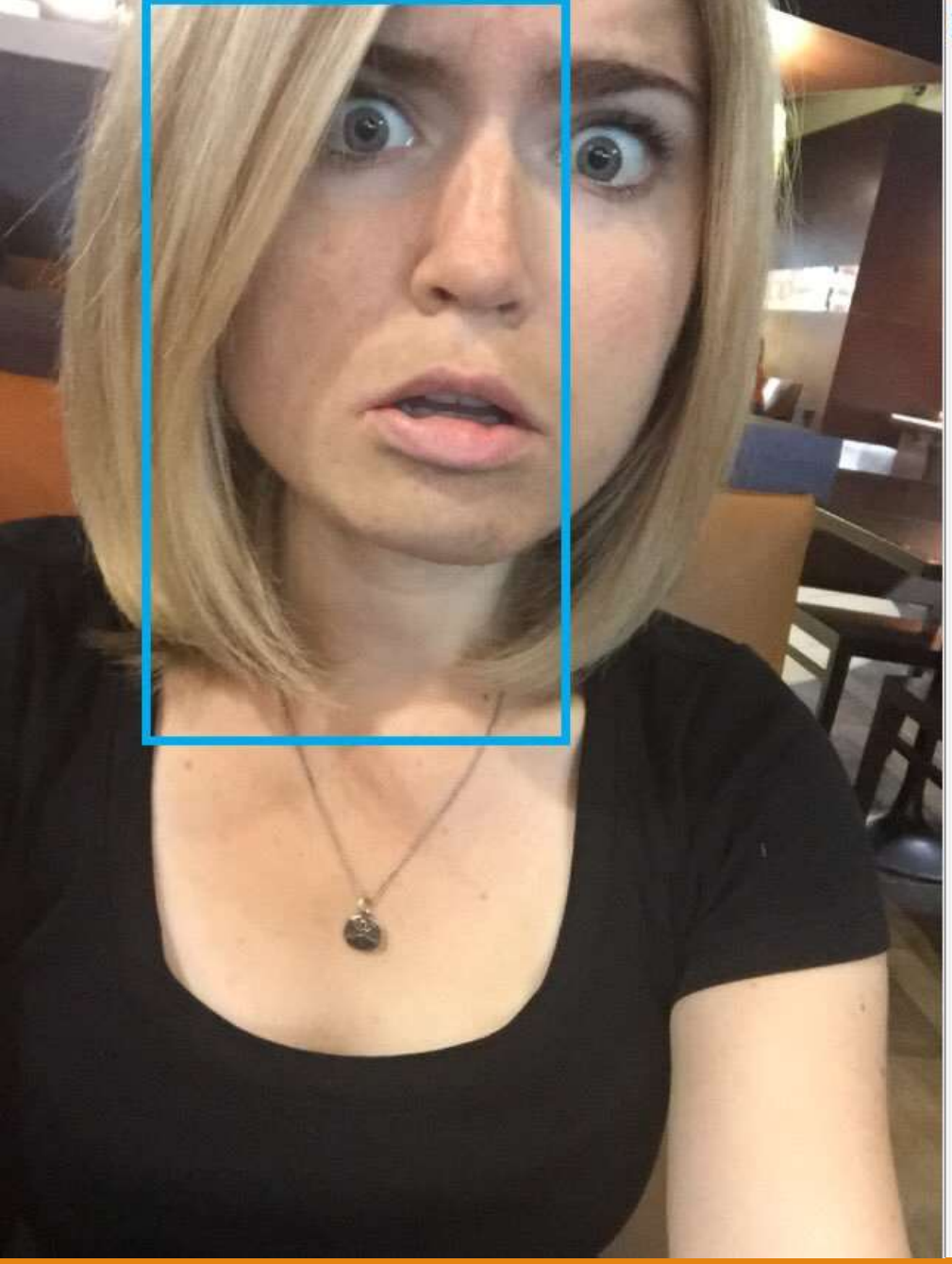
EXPECTED TO DEPART IN OVER 20 HOURS

Where is my plane now?

IAH	SEA
HOUSTON, TX	**SEATTLE, WA**
departing from **GATE E1**	arriving at **GATE A13**
Houston Bush Int'ctl - **IAH**	Seattle-Tacoma Intl - **SEA**
WEDNESDAY 12-JUL-2017	THURSDAY 13-JUL-2017
09:25PM CST (on time)	(on time) ***12:10AM PST***

4h 45m total flight time

NOT YOUR FLIGHT? UAL1157 flight schedule

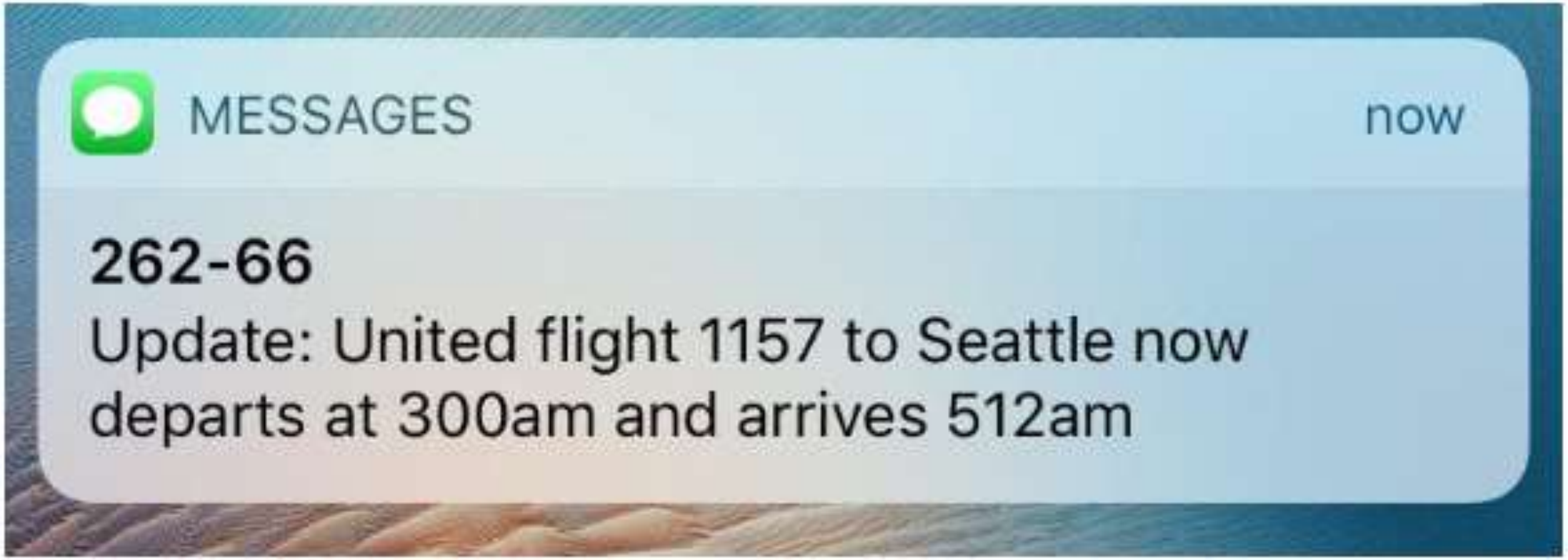


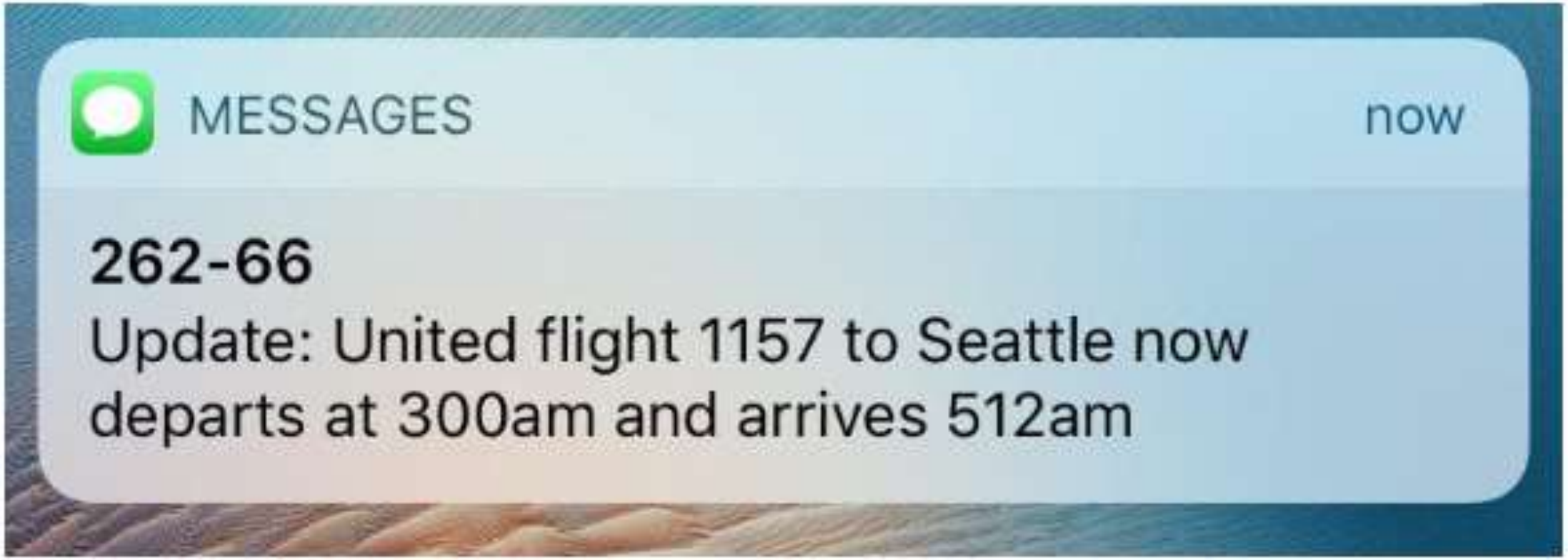
```
Detection result:
1 faces detected

JSON:
[
  {
    "FaceRect
      "Top":
      "Left": 200,
      "Width": 576,
      "Height": 565
    },
    "Scores": {
      "Anger": 0.00478247926,
      "Contempt": 0.001889224,
      "Disgust": 0.003453348,
      "Fear": 0.405526668,
      "Happiness": 6.23536062E-06,
      "Neutral": 0.107686535,
      "Sadness": 0.107204422,
      "Surprise": 0.3694511
    }
  }
```

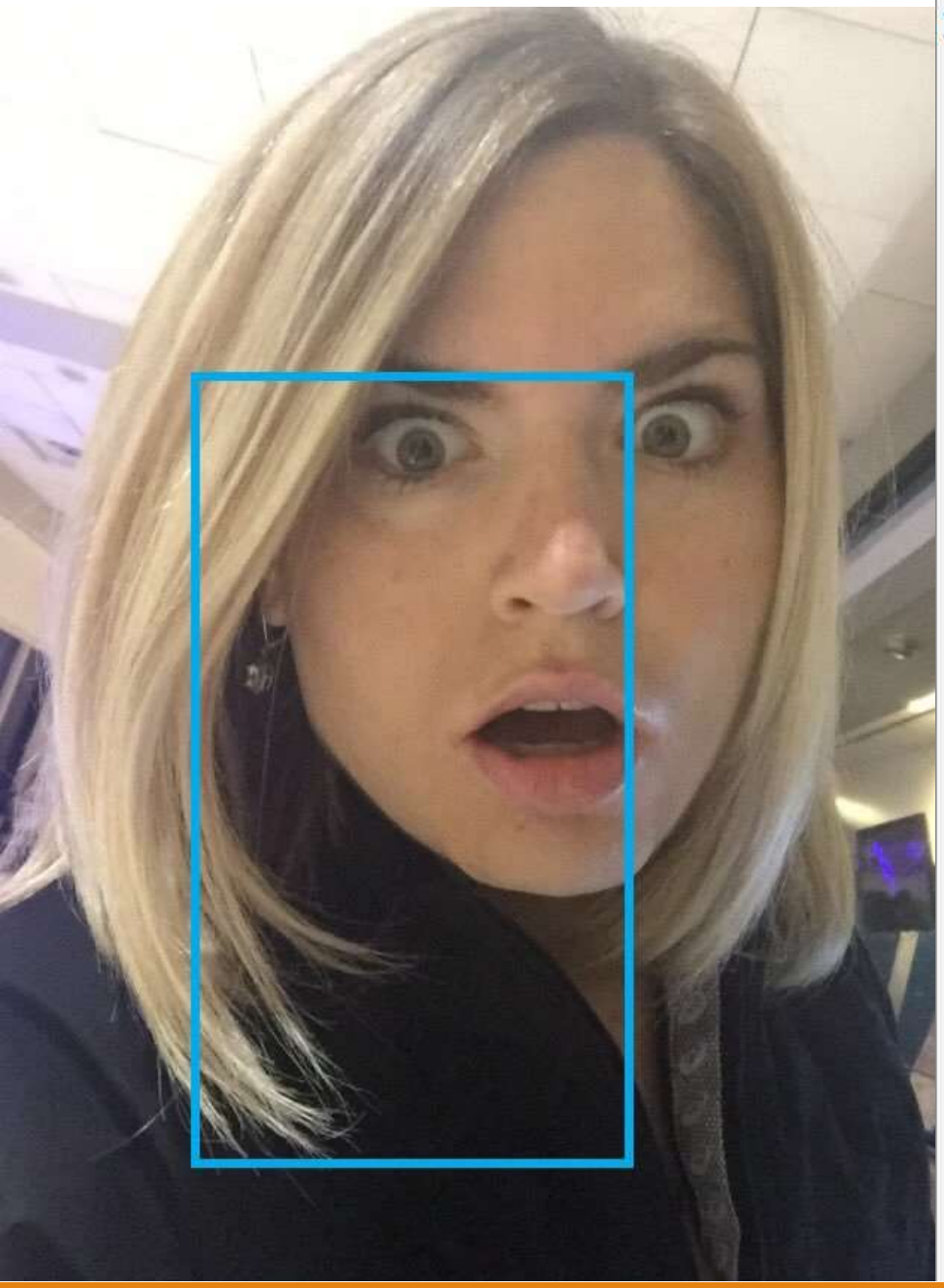
@DynamicWebPaige
@DynamicWebPaige

tfw you realize your flight will not arrive until 2:30am


MESSAGES
now
262-66
Update: United flight 1157 to Seattle now departs at 300am and arrives 512am



...interview starts at 7:30am.



```
Detection result:
1 faces detected

JSON:
[
  {
    "FaceRect
      "Top":
      "Left": 279,
      "Width": 591,
      "Height": 591
    },
    "Scores": {
      "Anger": 0.007640295,
      "Contempt": 0.000296741229,
      "Disgust": 0.001998572,
      "Fear": 0.0390212275,
      "Happiness": 6.261075E-05,
      "Neutral": 0.008304399,
      "Sadness": 0.000473582826,
      "Surprise": 0.942202568
    }
```

@DynamicWebPaige
@DynamicWebPaige

...actually, make that 5:12am. @microsoft, I think I broke your algorithm.

flight delay predictions

are a hugeeeeeeeee thing

Package 'nycflights13'

January 27, 2017

Title Flights that Departed NYC in 2013
Version 0.2.2
Description Airline on-time data for all flights departing NYC in 2013. Also includes useful 'metadata' on airlines, airports, weather, and plane
License CC0
LazyData true
Depends R (>= 2.10)
Imports tibble
Suggests dplyr
URL http://github.com/hadley/nycflights13
BugReports https://github.com/hadley/nycflights13/issues
RoxygenNote 5.0.1.9000

kaggle

coursera

Piping

You can use magrittr pipes to write cleaner syntax. Using the same example from above, you can write a much cleaner version like this:

```
c4 <- flights %>%
  filter(month == 5, day == 17, carrier %in% c('UA', 'WN', 'AA', 'DL')) %>%
  select(carrier, dep_delay, air_time, distance) %>%
  arrange(carrier) %>%
  mutate(air_time_hours = air_time / 60)
```

Grouping

The `group_by` function corresponds to the `GROUP BY` statement in SQL.

```
c4 %>%
  group_by(carrier) %>%
```

Google Scholar

flight delay prediction

Articles

About 260,000 results (0.08 sec)

Any time
Since 2017
Since 2016
Since 2013
Custom range...

Sort by relevance
Sort by date

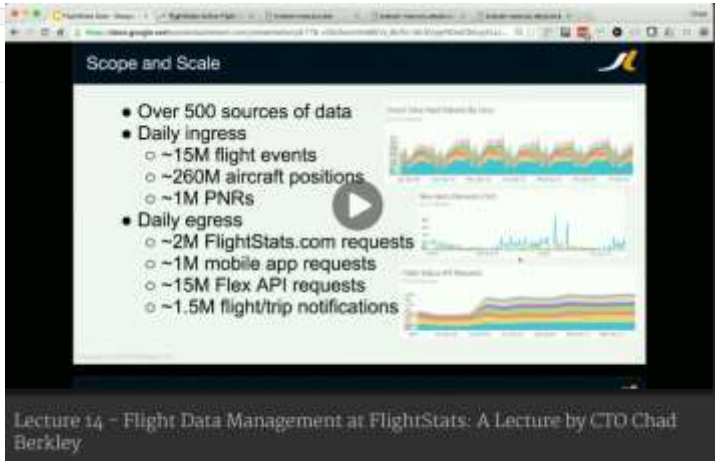
include patents
include citations

Create alert

[PDF] **Flight Delay Prediction**
V Martinez - 2012 - e-collection.library.ethz.ch
Abstract Flight delays are quite frequent (19% of the US domestic flights arrive more than 15 minutes late), and are a major source of frustration and cost for the passengers. As we will see, some flights are more frequently delayed than others, and there is an interest in
☆ 99 Related articles All 2 versions »

RIA-based visualization platform of **flight delay** intelligent **prediction**
R Yao, W Jiandong, D Jianli - ... , Communication, Control, and ..., 2009 - ieeexplore.ieee.org
Abstract: In order to provide a **flight delay prediction** tool based on software system for airports and airlines, a visualization platform of **flight delay** intelligent **prediction** is designed and implemented. The platform consists of airport data acquisition front-end computer,
☆ 99 Cited by 7 Related articles All 2 versions

Flight turnaround time analysis and **delay prediction** based on Bayesian network
W Cao, X Lin - Computer Engineering and Design, 2011 - en.cnki.com.cn
From the standpoint of **flight delay** propagation **prediction**, multi-factors that influenced **flight** turnaround time are analyzed, and a Bayesian network model is established, which could clearly reflected the influence of various factors on the downstream **flight** tur-naround time. A
☆ 99 Cited by 6 Related articles All 2 versions »



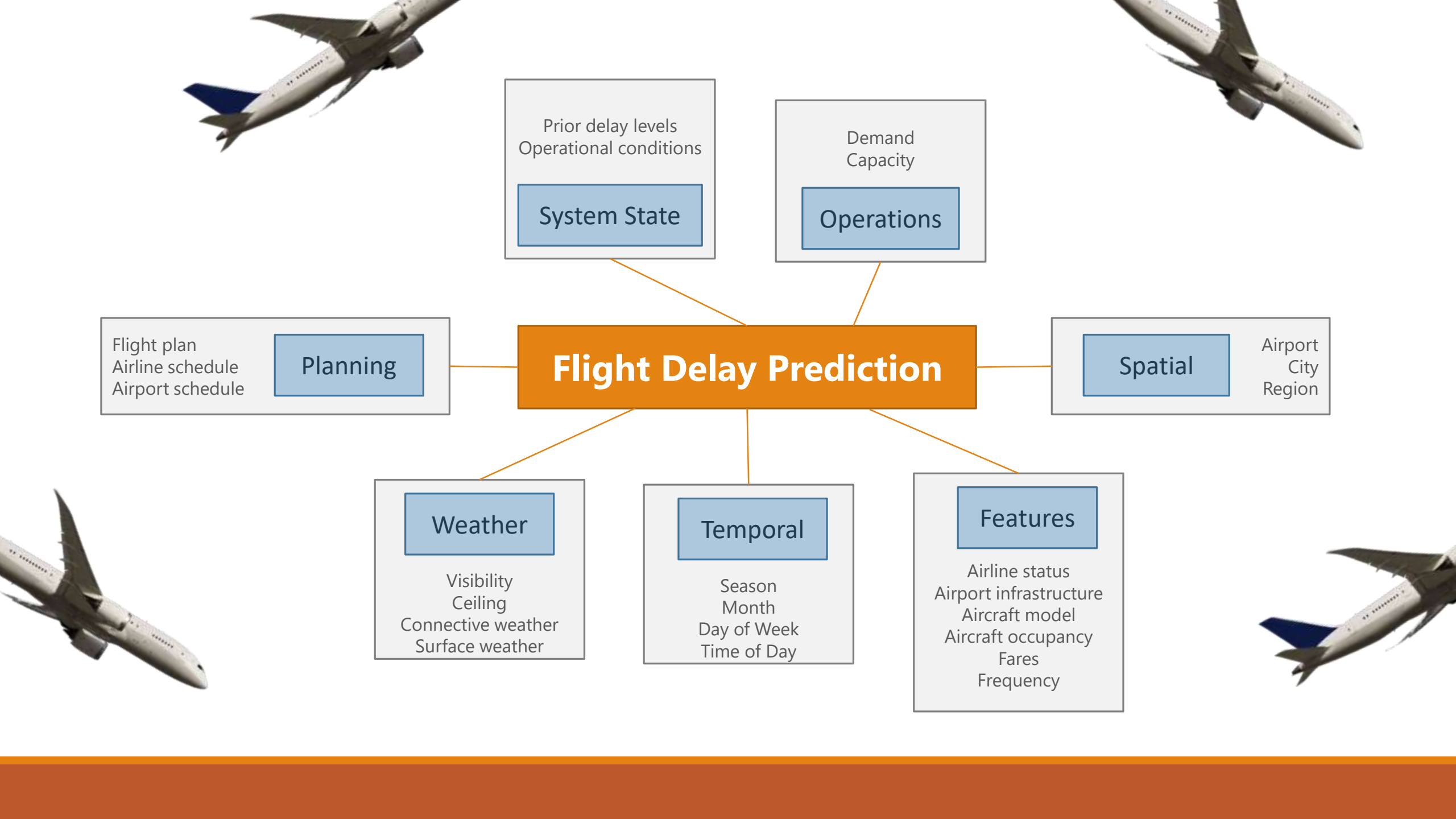
DataCamp

freely available data sources

Region	Ensemble	Airline	Airport
Asia	2	1	1
Brazil	2	0	0
Europe	7	2	7
United States	11	7	16

(full list of sources for the above datasets in Appendix)

United States Bureau of Transportation Statistics


Prior delay levels
Operational conditions
System State
Demand
Capacity
Operations
Flight plan
Airline schedule
Airport schedule
Planning
Flight Delay Prediction
Spatial
Airport
City
Region
Weather
Visibility
Ceiling
Connective weather
Surface weather
Temporal
Season
Month
Day of Week
Time of Day
Features
Airline status
Airport infrastructure
Aircraft model
Aircraft occupancy
Fares
Frequency

Methods

In general, there are five approaches to predicting flight delays.

Machine Learning
- Clustering
- Recommendation Systems
- Prediction

Operational Research
- Simulation
- Queueing Models

Network Representation
- Graph approaches
- Probability networks
- Distributions

Probabilistic Models
- Conditional probability
- Survival model

Statistical Analysis
- Econometric models
- Tests
- Correlation analysis

Methods

In general, there are five approaches to predicting flight delays.

Machine Learning

- Clustering
- Recommendation Systems
- Prediction

Operational Research

- Simulation
- Queueing Models

Network Representation

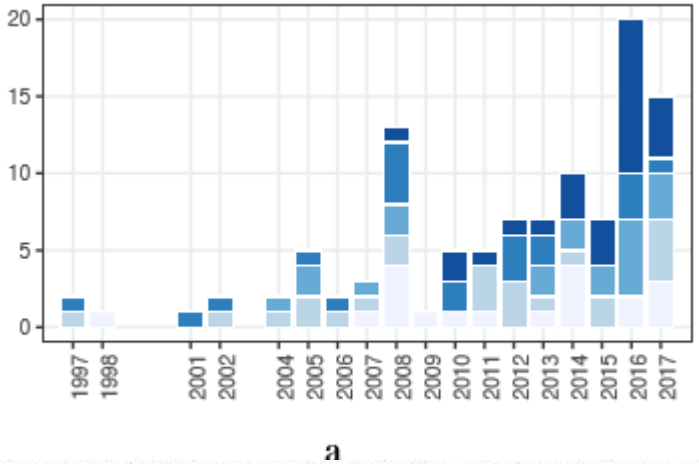
- Graph approaches
- Probability networks
- Distributions

Probabilistic Models

- Conditional probability
- Survival model

Statistical Analysis

- Econometric models
- Tests
- Correlation analysis



a

Figure 5: (a) Publication in years according to main methods: Statistical Analysis, Probabilistic Models, Network Representation, Operational Research, Machine Learning;

Team Data Science Process

For more information on the Team Data Science Process, be sure to check out its detailed documentation:

https://azure.microsoft.com/en-us/documentation/learning-paths/data-science-process/

Step 1: Business Understanding

- Identify your Scenario

Step 2: Data Acquisition and Understanding

- Load data into storage environments
- Prepare data
- Explore data
- Sample data

Step 3: Modeling

- Engineer features
- Train the Model
- Evaluate and Tune the Model

Step 4: Deployment

- Publish the model as a Web service
- Consume a model in Excel
- Consume a model programmatically

it’s dangerous to go alone, take this

Houston Flight (delayed) – UA 1157 on July 12	
IAH Airport Code	12266
SEA-TAC Airport Code	14747
Austin Airport Code	10423
Departure / Arrival	9:25pm / 12:19pm

Comparison Flight (on time) – ASA461 on July 12	
Departure / Arrival	7:30pm / 10:18pm

Upcoming Flight – ASA671 on November 15	
Departure / Arrival	7:40pm / 10:08pm

[Algorithm Cheat Sheet]

LIVE DEMO

The most terrifying experience of them all.

improve performance

- Increase the size of the dataset
What if you had three years of data, instead of just one month? Or data from November?

- Additional feature engineering
What if you included additional columns in the data set?

- Data quality
What if some of the values are skewed, and delayed flights are marked on-time?

- Additional data sets
Ex: weather, geopolitical events, natural disasters

- Algorithms
Perhaps we should try a different algorithm!

- Hyperparameter tuning
Changing the analysis parameters for the algorithm can sometimes improve performance.

Thank you!

Appendix

Literature Review
Demoed Products

A Review on Flight Delay Prediction (Sternberg, Soares, Carvalho, Ogasawara). https://arxiv.org/pdf/1703.06118.pdf. 2017.

Bureau of Transportation Statistics: Flight Data. https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time (*usually a couple of months late on refresh*)

Azure Machine Learning Studio: http://studio.azureml.net

Azure Notebooks: http://notebooks.azure.com

Azure Machine Learning Studio Documentation: https://docs.microsoft.com/en-us/azure/machine-learning/studio/what-is-machine-learning

Azure Machine Learning Modules: https://msdn.microsoft.com/library/azure/6d9e2516-1343-4859-a3dc-9673ccec9edc/

Azure Machine Learning Workbench: https://docs.microsoft.com/en-us/azure/machine-learning/preview/quickstart-installation

Cortana Intelligence Gallery: https://gallery.cortanaintelligence.com/