

E6893 Big Data Analytics:

# Flight Route

## Final Presentation

Guy Farkash



# Motivation

The number of business flights in the US was **488M** in 2016, with average of **1.3M** each day

(By the Global Business Travel Association)

~**6%** of the flights, delayed → **bad weather**

(By the Bureau Of Transportation Statistics)

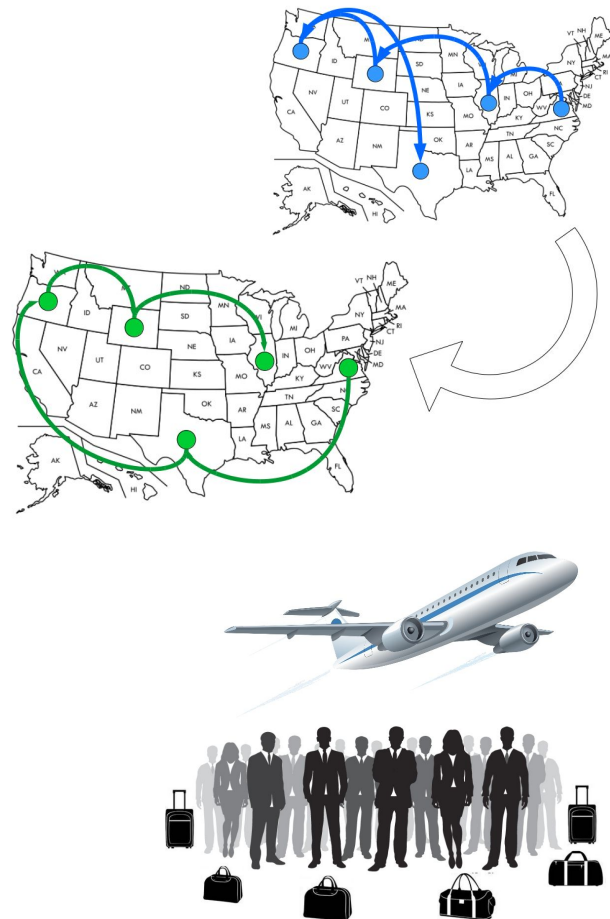
Avoid bad weather flights → **Save time**

Saving **5 min** each trip → Saves **10 months** of work each day

$(1.3M \times 0.06 \times 5) / (60 \times 24 \times 7 \times 4) \sim 9.7$  (for all trips combined)

For **multi-destination trips**

Can choose a cheaper route → **Save money**

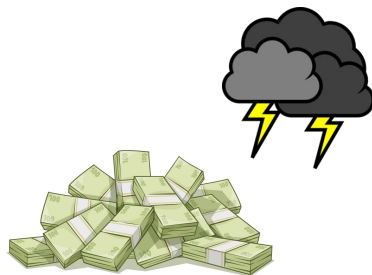


# Goal

## Best flight route:

Minimum predicted weather delay

Break even by choosing minimum cost



## Input:

[Origin City, City 1, City 2, ..., City N]

[City 1 days, City 2 days, ..., City N days]

Starting Date

Return Date

## Output:

Origin City→City 3

[Starting Date, Delay, Fare]

City 3→City 2

[Date 1, Delay, Fare]

.....

City N→City 5

[Date K, Delay, Fare]

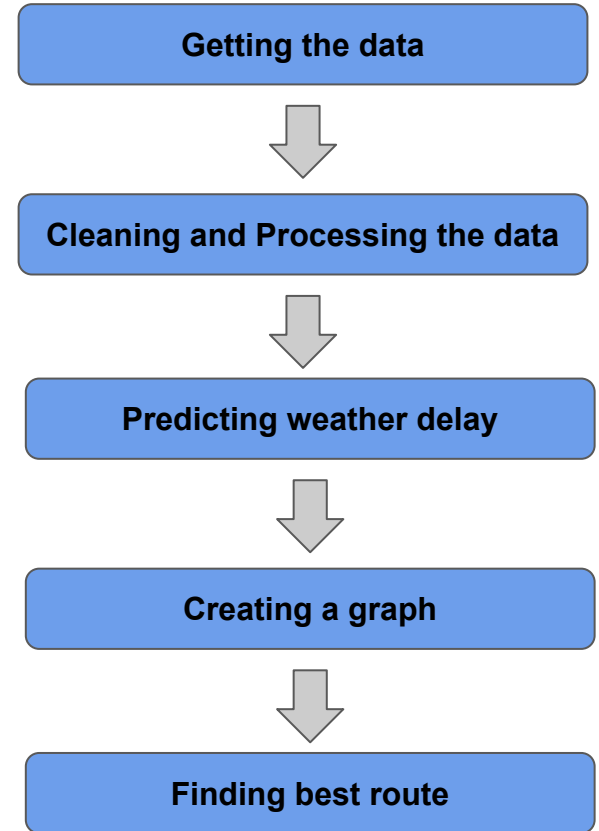
City 5→Origin City

[Return Date, Delay, Fare]



# The Process

1. **Getting the data**  
Flight reports with weather delay for years 2012 – 2017  
Flight fares for 2017
2. **Cleaning and Processing the data**  
Fixing cities names  
Creating cities IDs  
Creating the database
3. **Predicting weather delay**  
Calculating the delay for each flight in each date
4. **Creating a graph**  
Building the vertices and nodes files
5. **Finding best route**  
Using the graph database



# Getting The Data

On-Time Performance database  
from Bureau of Transportation Statistics

Years 2012 - 2017      **3.2GB** of data

Consumer Airfare report  
from US department of transportation

Flights average fare for every quarter of 2017

## Delays Data

	Field Name	Description
1	Year	Year
2	Month	Month
3	DayOfMonth	Day
4	OriginCityName	Origin City
5	DestCityName	Destination City
6	WeatherDelay	Weather Delay, in Minutes.

## Fares Data

	Field Name	Description
1	city1	Origin city
2	city2	Destination city
3	fare	Average fare of all flights and airlines

# Cleaning and Processing The Data

Removing records with no delay information

Renaming city names to match for all records

Removing \$ signs from the fare column

Removing unnecessary columns

Adding unique city ID number

Adding fare

Combining all the the files to one database

## Processed Data

	Field Name	Description
1	YEAR	The year → [2012, 2014, 2015, 2016, 2017]
2	MONTH	The month → integer number [1, 2, ..., 11, 12]
3	DAY_OF_MONTH	The day of month → integer number [1, 2, ..., 30, 31]
4	ORIGIN_CITY_ID	Origin city name → string
5	ORIGIN_CITY_NAME	Origin city ID → unique integer number
6	DEST_CITY_ID	Destination city name → string
7	DEST_CITY_NAME	Destination city ID → unique integer number
8	WEATHER_DELAY	Weather Delay, in Minutes → integer number
9	FARE	Fare in Dollars → floating point number

# Predicting Weather Delay



Several delay records for each flight:

YEAR	MONTH	DAY	ORIGIN CITY NAME	DEST CITY NAME	WEATHER DELAY
2012	1	22	New York, NY	Boston, MA	6
2014	1	22	New York, NY	Boston, MA	25
2014	1	22	New York, NY	Boston, MA	75
2016	1	22	New York, NY	Boston, MA	59

Number of records can be 1 - 45

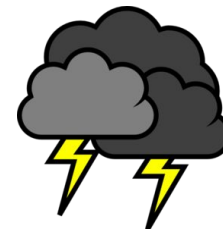
Using ML predictions:

1. Naive Bayes
2. Random Forest
3. Multilayer Perceptron
4. One-vs-Rest
5. Decision Tree

**The information is too scarce to train a good model**

**All the ML models produced accuracy < 30%**

# Predicting Weather Delay



Creating a predictor of the average delay with confident value

Combining records for each flight to calculate average delay:

MONTH	DAY	ORIGIN CITY NAME	DEST CITY NAME	AVG WEATHER DELAY	RECORDS
1	22	New York, NY	Boston, MA	41.25	4

Adding records from +-5 days of current date: (increase the confidence in the result)

MONTH	DAY	ORIGIN CITY NAME	DEST CITY NAME	AVG WEATHER DELAY	RECORDS
1	22	New York, NY	Boston, MA	40.13	31

Assuming that the weather is similar in a 10 days window



# Predicting Weather Delay



## Examples:

### Low Confidence:

MONTH	DAY	ORIGIN CITY NAME	DEST CITY NAME	AVG WEATHER DELAY	RECORDS
1	5	New York, NY	Dallas, TX	16	1

### High Confidence:

MONTH	DAY	ORIGIN CITY NAME	DEST CITY NAME	AVG WEATHER DELAY	RECORDS
1	6	New York, NY	Los Angeles, CA	63	79

# Testing The Predictions

**P** = Predicted Delay

**A** = Actual Delay

“False Alarm” prediction  $P > A$

“Miss Detect” prediction  $P < A$

Predictor error [min] 
$$Avg\ Err = \frac{\sum A - P}{N}$$

**A Good predictor**

Low “Miss Detect” and “False Alarm” error rate

“Miss Detect” error rate < “False Alarm” error rate

Splitting the data

**Training Data** (2012 - 2016)

**Testing Data** (2017)

Using the training data to build the database

Using the testing data to test the predictor

**“Miss Detect”** average error = **~9.5 [min]**

**“False Alarm”** average error = **~20 [min]**

Total average delay is 46 [min]

**“Miss Detect”** accuracy = ~79%

**“False Alarm”** accuracy = ~56%

# Creating a Graph

## Origin Vertices File Contains:

Origin cities name  
Origin cities ID

## Destination Vertices File Contains:

Destination cities name  
Destination cities ID

## Nodes File Contains:

Year, Month, Day  
Origin cities name  
Origin cities ID  
Destination cities name  
Destination cities ID  
Predicted weather delay  
Number of records  
Fares

## Loading the graph to GrapheneDB



## Origin Vertices

	Field Name	Description
1	ORIGIN_CITY_NAME	Origin city name → string
2	ORIGIN_CITY_ID	Origin city ID → unique integer number

## Destination Vertices

	Field Name	Description
1	DEST_CITY_NAME	Destination city name → string
2	DEST_CITY_ID	Destination city ID → unique integer number

## Nodes

	Field Name	Description
1	YEAR	The year → [2012, 2014, 2015, 2016, 2017]
2	MONTH	The month → integer number [1, 2, ..., 11, 12]
3	DAY_OF_MONTH	The day of month → integer number [1, 2, ..., 30, 31]
4	ORIGIN_CITY_ID	Origin city name → string
5	ORIGIN_CITY_NAME	Origin city ID → unique integer number
6	DEST_CITY_ID	Destination city name → string
7	DEST_CITY_NAME	Destination city ID → unique integer number
8	AVG_WEATHER_DELAY	Weather Delay, in Minutes → integer number
8	DELAY_RECORDS	Number of records → integer number
9	FARE	Fare in Dollars → floating point number

# Finding The Best Route



## Best Route Algorithm:

### 1) Find all routes possibilities

All cities - dates combinations

### 2) Get delay and fare data for each route

### 3) Choose the best route with:

- a) Less delay
- b) High confidence (large records number)
- c) Less fare

**Start**...Date1...Date2...Date3.....**End**  
**Start**.....Date1...Date2.....Date3...**End**  
**Start**...Date1.....Date2...Date3.....**End**  
**Start**.....Date1....Date2.....Date3.....**End**

Start	Date 1	Date 2	Date 3	End
Ori	City 1	City 2	City 3	Ori
Ori	City 1	City 3	City 2	Ori
Ori	City 2	City 1	City 3	Ori
Ori	City 2	City 3	City 1	Ori
Ori	City 3	City 1	City 2	Ori
Ori	City 3	City 2	City 1	Ori

# GUI

## A convenient user interface makes it easy to use

Origin city = **Boston**

Destination 1 = **New York**

Destination 2 = **Las Vegas**

Destination 3 = **San Francisco**

Flight Route

Search

Origin City	Start Month	Start Day	Destination Cities	End Month	End Day	Days Per City
Aberdeen, SD	1	1	Lake Charles, LA	1	1	5
Abilene, TX	2	2	Lansing, MI	2	2	5
Adak Island, AK	3	3	Laramie, WY	3	3	5
Aguadilla, PR	4	4	Laredo, TX	4	4	
Akron, OH	5	5	Las Vegas, NV	5	5	
Albany, GA	6	6	Latrobe, PA	6	6	
Albany, NY	7	7	Lawton/Fort Sill, OK	7	7	
Albuquerque, NM	8	8	Lewisburg, WV	8	8	
Alexandria, LA	9	9	Lewiston, ID	9	9	
Allentown/Bethlehem/Easton, PA	10	10	Lexington, KY	10	10	
Alpena, MI	11	11	Lihue, HI	11	11	
Amarillo, TX	12	12	Lincoln, NE	12	12	
Anchorage, AK		13	Little Rock, AR		13	
Appleton, WI		14	Long Beach, CA		14	
Arcata/Eureka, CA		15	Longview, TX		15	
Asheville, NC		16	Los Angeles, CA		16	
Aspen, CO		17	Louisville, KY		17	
Atlanta, GA		18	Lubbock, TX		18	
Atlantic City, NJ		19	Madison, WI		19	
Augusta, GA		20	Mammoth Lakes, CA		20	
Austin, TX		21	Manchester, NH		21	
Bakersfield, CA		22	Manhattan/Ft. Riley, KS		22	
Baltimore, MD		23	Marquette, MI		23	
Bangor, ME		24	Martha's Vineyard, MA		24	
Barrow, AK		25	Medford, OR		25	
Baton Rouge, LA		26	Melbourne, FL		26	
Beaumont/Port Arthur, TX		27	Memphis, TN		27	
Bellingham, WA		28	Meridian, MS		28	
Bemidji, MN		29	Miami, FL		29	
Bend/Redmond, OR		30	Midland/Odessa, TX		30	
Bethel, AK		31	Milwaukee, WI		31	
Billings, MT			Minneapolis, MN			
Binghamton, NY			Minot, ND			
Birmingham, AL			Mission/McAllen/Edinburg, TX			
Bismarck/Mandan, ND			Missoula, MT			
Bloomington/Normal, IL			Moab, UT			
Boise, ID			Mobile, AL			
Boston, MA			Modesto, CA			
Bozeman, MT			Moline, IL			
Brainerd, MN			Monroe, LA			
Branson, MO			Monterey, CA			
Bristol/Johnson City/Kingsport, TN			Montgomery, AL			
Brownsville, TX			Montrose/Delta, CO			
Brunswick, GA			Mosinee, WI			
Buffalo, NY			Muskegon, MI			
Burbank, CA			Myrtle Beach, SC			
Burlington, VT			Nantucket, MA			
Butte, MT			Nashville, TN			
Carlsbad, CA			New Bern/Morehead/Beaufort, NC			
Casper, WY			New Orleans, LA			
Cedar City, UT			New York, NY			
Cedar Rapids/Iowa City, IA			Newark, NJ			
Champaign/Urbana, IL			Newburgh/Poughkeepsie, NY			
Charleston, SC			Newport News/Williamsburg, VA			
Charleston/Dunbar, WV			Nome, AK			


The best route has avg delay of 109 minutes:

Boston, MA ----> San Francisco, CA @ 1/23  
The avg delay is 36 (21.0 reports)  
The fare is 324.72\$

San Francisco, CA ----> Las Vegas, NV @ 2/4  
The avg delay is 9 (12.0 reports)  
The fare is 133.6\$

Las Vegas, NV ----> New York, NY @ 2/9  
The avg delay is 28 (5.0 reports)  
The fare is 278.83\$

New York, NY ----> Boston, MA @ 2/14  
The avg delay is 36 (82.0 reports)  
The fare is 148.51\$



# Example Results

**Input:**      Origin city = **Boston**      Destination 1 = **New York**      (5 days)      Starting date: **Jan. 23th**  
   Destination 2 = **Las Vegas**      (5 days)      Ending date: **Feb. 14th**  
   Destination 3 = **San Francisco**      (5 days)

## Output:

The best route has avg  
**delay of 109** minutes:

**Boston, MA ---> San Francisco, CA @ 1/23**

The avg delay is 36 (21.0 reports)

The fare is 324.72 \$

**San Francisco, CA ---> Las Vegas, NV @ 2/4**

The avg delay is 9 (12.0 reports)

The fare is 133.67 \$

**Las Vegas, NV ---> New York, NY @ 2/9**

The avg delay is 28 (5.0 reports)

The fare is 278.83 \$

**New York, NY ---> Boston, MA @ 2/14**

The avg delay is 36 (82.0 reports)

The fare is 148.51 \$

# Example Results

**Best Route Predicted Delay:**     **109** minutes

**Boston, MA ---> San Francisco, CA @ 1/23**

The avg delay is 36 (21.0 reports)

**San Francisco, CA ---> Las Vegas, NV @ 2/4**

The avg delay is 9 (12.0 reports)

**Las Vegas, NV ---> New York, NY @ 2/9**

The avg delay is 28 (5.0 reports)

**New York, NY ---> Boston, MA @ 2/14**

The avg delay is 36 (82.0 reports)

**Real World Delay:**     **99** minutes

**Boston, MA ---> San Francisco, CA @ 1/23**

The delay was 23

**San Francisco, CA ---> Las Vegas, NV @ 2/4**

The delay was 0

**Las Vegas, NV ---> New York, NY @ 2/9**

The delay was 0

**New York, NY ---> Boston, MA @ 2/14**

The avg delay is 76

**Other Route Predicted Delay:**     **208** minutes

**Boston, MA ---> New York, NY ---> Las Vegas, NV ---> San Francisco, CA ---> Boston, MA (different order)**

# Thank You

