

# ETC5512: Wild Caught Data

Week 10

## US flights and databases

Lecturer: *Dianne Cook*

Department of Econometrics and Business Statistics

✉ ETC5512.Clayton-x@monash.edu

Image source: [https://commons.wikimedia.org/wiki/File:Another\\_Airplane!\\_\(4676723312\).jpg](https://commons.wikimedia.org/wiki/File:Another_Airplane!_(4676723312).jpg)

# Motivation

American Statistical Association Statistical Graphics and Computing Sections [2009 Data Expo](#) provided all of the commercial flight records for airtravel in the USA from October 1987 to April 2008.

## Questions provided

- ❖ When is the best time of day/day of week/time of year to fly to minimise delays?
- ❖ Do older planes suffer more delays?
- ❖ How does the number of people flying between different locations change over time?
- ❖ How well does weather predict plane delays?
- ❖ Can you detect cascading failures as delays in one airport create delays in others? Are there critical links in the system?

but participants could also decide for themselves what to analyse.

# About the data

- ❖ nearly 120 million records
- ❖ 1.6 gigabytes of space compressed
- ❖ 12 gigabytes when uncompressed

Organisers provided instructions on how to set up an **sqlite database**, and access from R.

# Read about accessing databases from R at

This RStudio site <https://db.rstudio.com/databases/sqlite/> is a good starting place to read about working with a sqlite database.

# The original data source



United States Department of Transportation

Ask-A-Librarian | A-Z Index

Bureau of Transportation Statistics

Topics and Geography      Statistical Products and Data      National Transportation Library      Newsroom

OST-R > BTS

TranStats

Search this site: Go      Advanced Search

On-Time : Reporting Carrier On-Time Performance (1987-present)

Databases    Data Tables    Table Contents

Download Instructions    Filter Geography    Filter Year    Filter Period  
Latest Available Data: February 2020    All    2019    March

Prepped File     % Missing     Documentation     Terms    Download

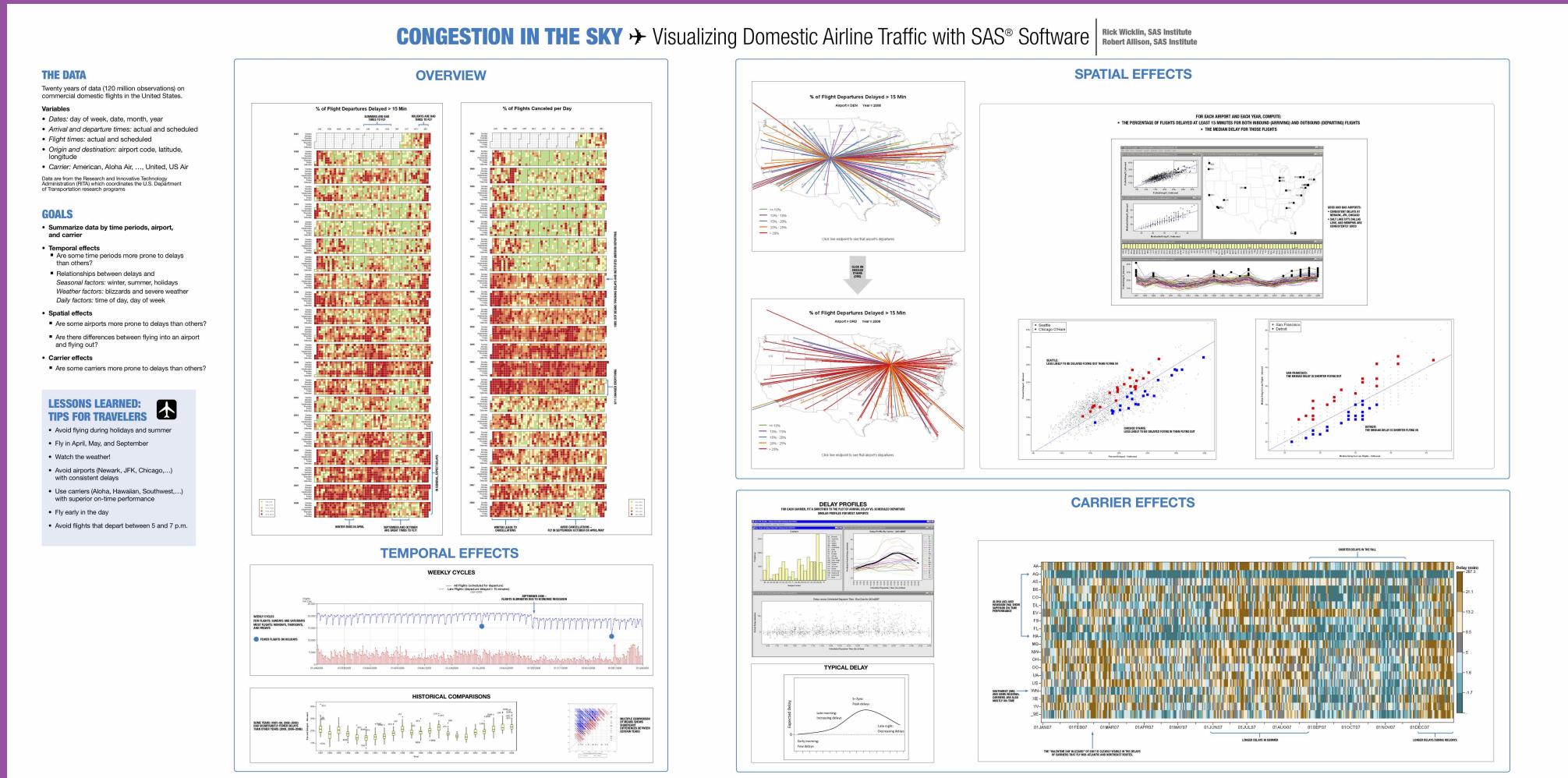
Field Name	Description	Support Table
Time Period		
<input checked="" type="checkbox"/> Year	Year	
<input checked="" type="checkbox"/> Quarter	Quarter (1-4)	<a href="#">Get Lookup Table</a>
<input checked="" type="checkbox"/> Month	Month	<a href="#">Get Lookup Table</a>
<input checked="" type="checkbox"/> DayofMonth	Day of Month	
<input checked="" type="checkbox"/> DayOfWeek	Day of Week	<a href="#">Get Lookup Table</a>
<input checked="" type="checkbox"/> FlightDate	Flight Date (yyyymmdd)	
Airline		
<input checked="" type="checkbox"/> Reporting_Airline	Unique Carrier Code. When the same code has been used by multiple carriers, a numeric suffix is used for earlier uses. For example, PA, PA(1), PA(2). Use this field for analysis across a range of years.	<a href="#">Get Lookup Table</a>
<input checked="" type="checkbox"/> DOT_ID_Reportng_Airline	An identification number assigned by US DOT to identify a unique airline (carrier). A unique airline (carrier) is defined as one holding and reporting under the same DOT certificate regardless of its Code, Name, or holding company/corporation.	<a href="#">Get Lookup Table</a>
<input checked="" type="checkbox"/> IATA_CODE_Reportng_Airline	Code assigned by IATA and commonly used to identify a carrier. As this code may have been assigned to different carriers over time, the code is not always unique. For analysis, use the Unique Carrier Code.	<a href="#">Get Lookup Table</a>
<input checked="" type="checkbox"/> Tail_Number	Tail Number	

- ✗ You can find the most current data at <https://www.transtats.bts.gov/DataIndex.aspx>
- ✗ Look at the "On-Time Performance 1987-present" table.
- ✗ You can download data a month at a time
- ✗ Its about a month lag in records
- ✗ Explanations of the variables

# Accessing the data

- ❖ Data expo files: the data for the competition is **still available** 
- ❖ Navigating the **BTS web interface**
  - ⌚ What data is available
  - ⌚ How do you download
  - ⌚ Explanations of the records and variables
- ❖ R package NYCflight13: provides a small domesticated data set.  This is a good way to *dip your toes in the water* with the airline data - try this if working with the full data is too scary.

# What did others do? First prize



# Temporal trend

A major component of this data is traffic patterns over time.

# Overview

Its good practice to show a useful view of entire data, to get a rough sense of major patterns.

## Highlights

## Carriers

Are some carriers operating more widely, or more efficiently?

## Spatial pattern

Airports are distributed across the country, explore how the traffic operates relative to this geography

## THE DATA

Twenty years of data (120 million observations) on commercial domestic flights in the United States.

### Variables

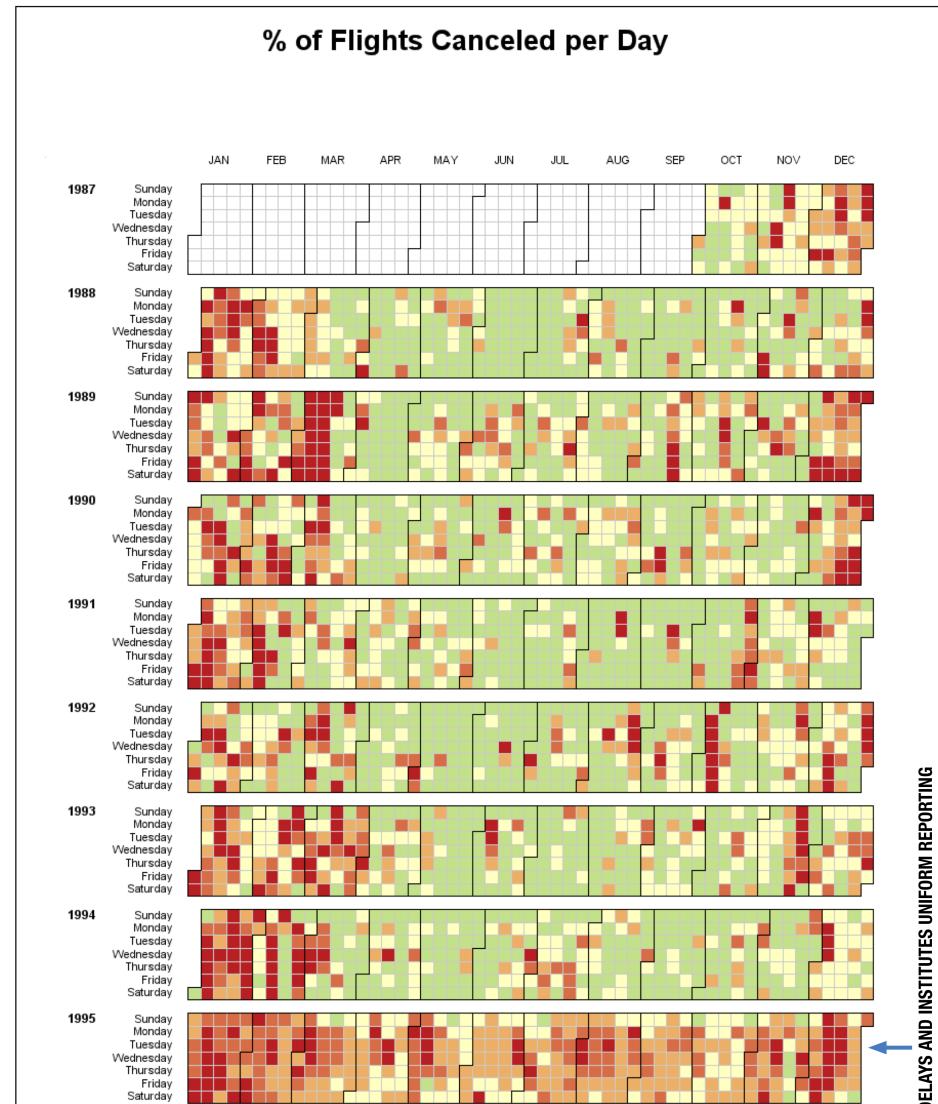
- *Dates*: day of week, date, month, year
- *Arrival and departure times*: actual and scheduled
- *Flight times*: actual and scheduled
- *Origin and destination*: airport code, latitude, longitude
- *Carrier*: American, Aloha Air, ..., United, US Air

Data are from the Research and Innovative Technology Administration (RITA) which coordinates the U.S. Department of Transportation research programs

## GOALS

- **Summarize data by time periods, airport, and carrier**
- **Temporal effects**
  - Are some time periods more prone to delays than others?
  - Relationships between delays and  
*Seasonal factors*: winter, summer, holidays  
*Weather factors*: blizzards and severe weather  
*Daily factors*: time of day, day of week
- **Spatial effects**
  - Are some airports more prone to delays than others?
  - Are there differences between flying into an airport and flying out?
- **Carrier effects**
  - Are some carriers more prone to delays than others?

# Overview



# Think about it 🤔

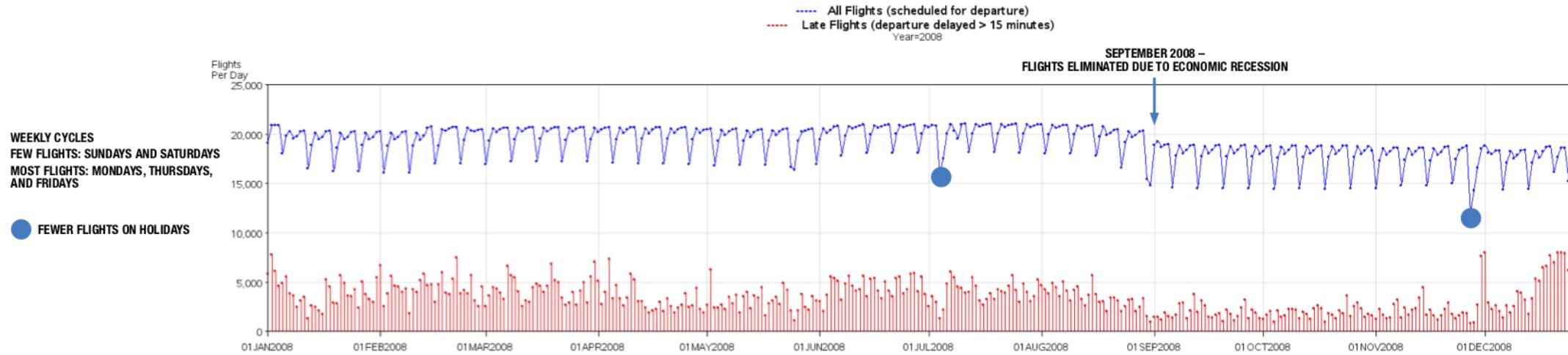
Delay was used in providing an overview.

- ✗ What other aggregates could have been used?
- ✗ Why was delay chosen?

00 : 30

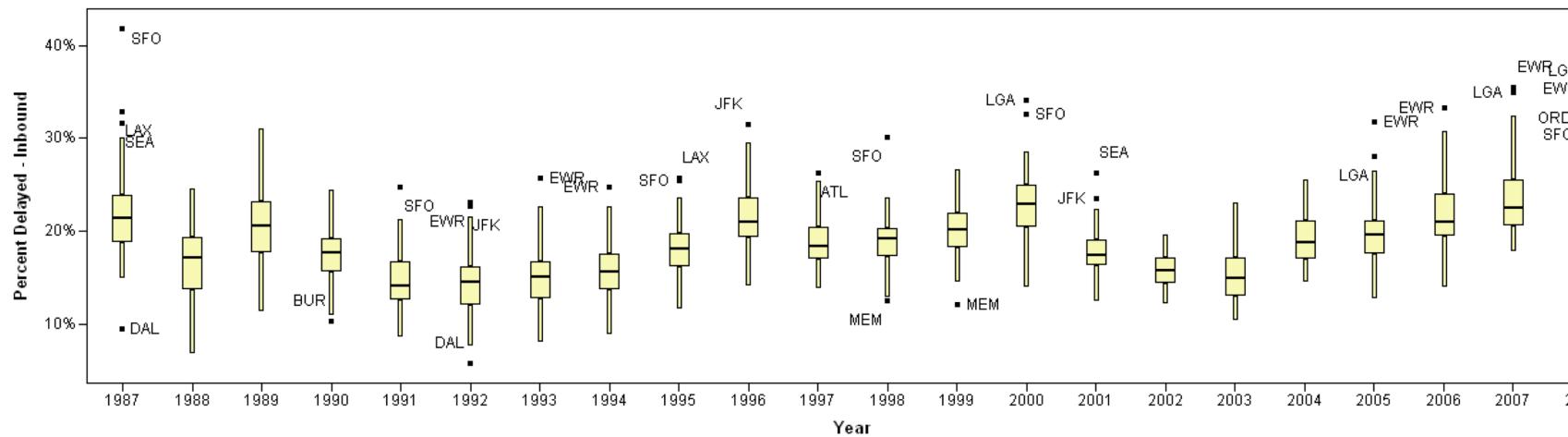
# Temporal trend

## WEEKLY CYCLES

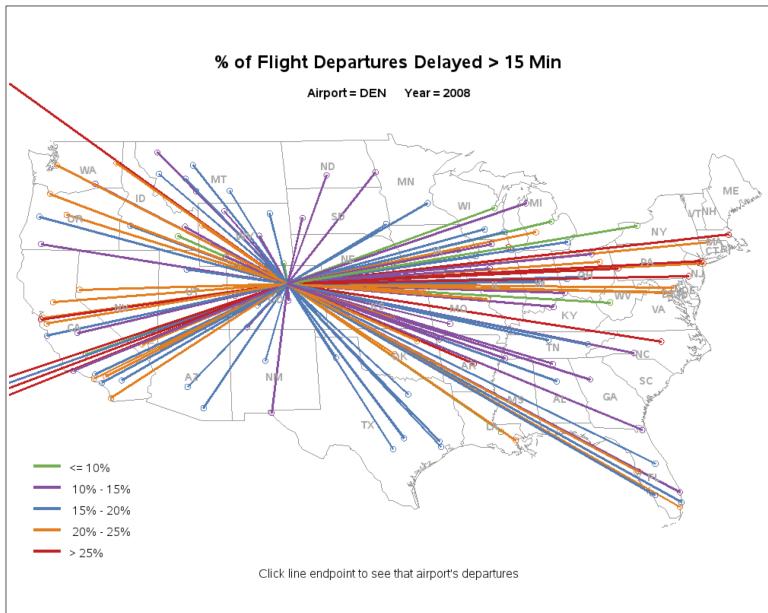


# Temporal trend

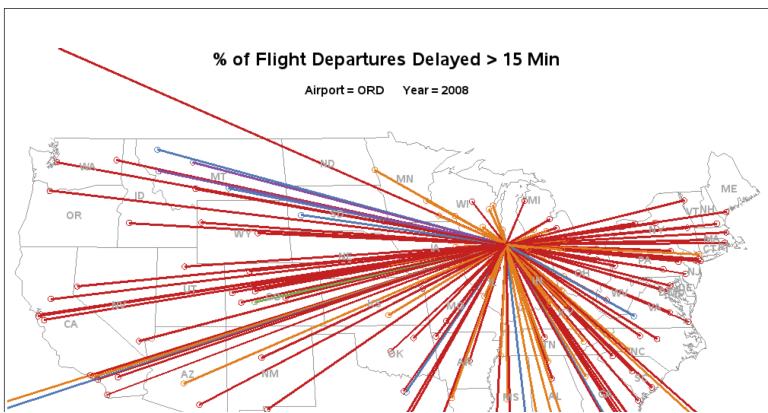
SOME YEARS (1991–94, 2002–2003)  
HAD SIGNIFICANTLY FEWER DELAYS  
THAN OTHER YEARS (2000, 2006–2008)



# Spatial

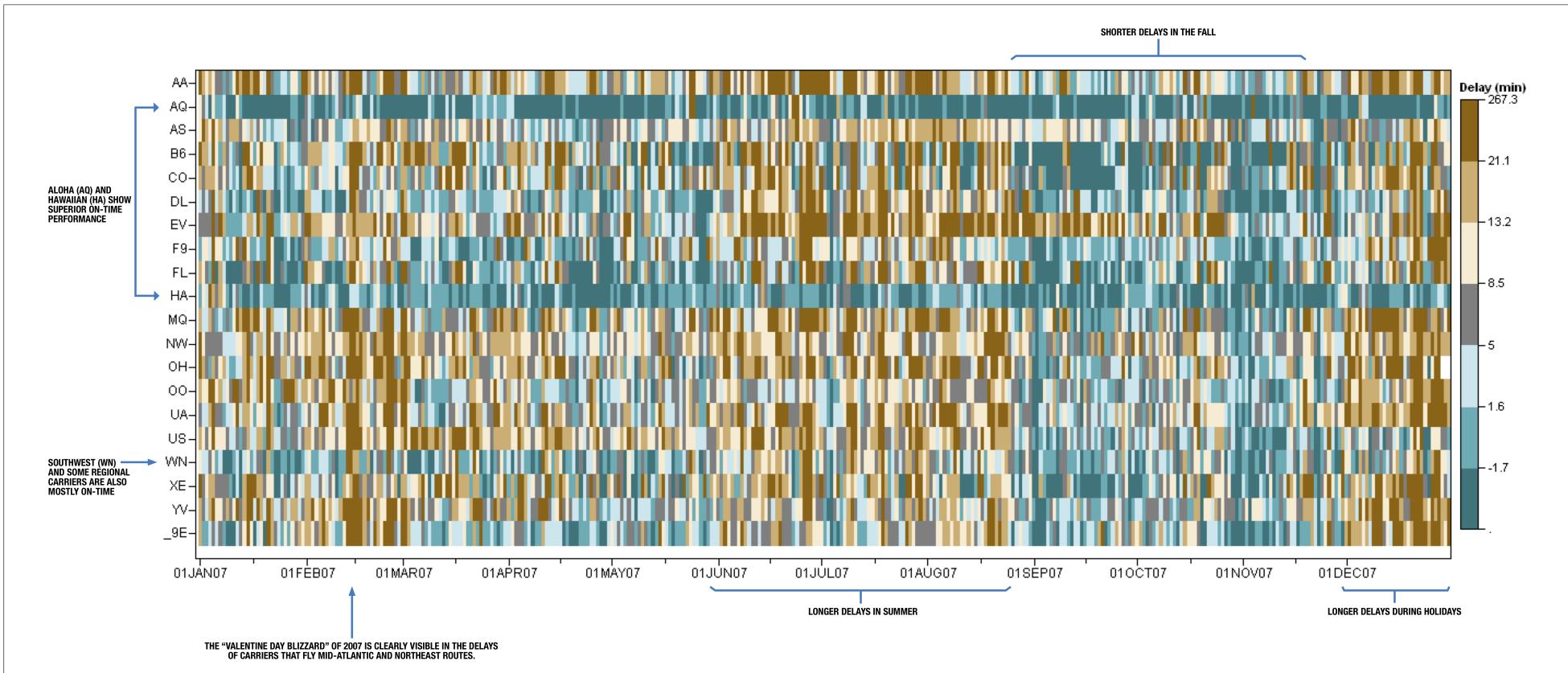


CLICK ON  
CHICAGO  
O'HARE  
(ORD)



# Carrier

## CARRIER EFFECTS



## LESSONS LEARNED: TIPS FOR TRAVELERS



- Avoid flying during holidays and summer
- Fly in April, May, and September
- Watch the weather!
- Avoid airports (Newark, JFK, Chicago,...) with consistent delays
- Use carriers (Aloha, Hawaiian, Southwest,...) with superior on-time performance
- Fly early in the day
- Avoid flights that depart between 5 and 7 p.m.

# What did others do? Second prize



**Delayed, Cancelled, On Time, Boarding, ..... Flying in the USA**

Heike Hofmann, Di Cook, Chris Kielion, Barret Schloerke, Jon Hobbs, Adam Loy, Lawrence Mosley, David Rockoff, Yuanyuan Sun, Danielle Wrolstad, Tengfei Yin  
Department of Statistics, Iowa State University

**Data**

The data are provided by Research and Innovative Technology Administration (RITA) and Bureau of Transportation Statistics (BTS). Arrival and departure details for 123 million commercial flights throughout the United States are recorded between October 1987 and December 2008, representing 29 commercial airlines and 3,376 airports. About 2.3 million flights were cancelled, 25 million flights were at least 15 minutes late.

**Additional Sources**

Additional BTS Data:  
 \* Monthly fuel cost and consumption data by carrier  
 \* Fleet information by carrier  
 Hourly weather details for each airport from Weather Underground at <http://www.wunderground.com>

**FLIGHTS OF '07**  
 The maps above and below show all flights from 2007. Southwest Airlines (WN) operates without a hub system, the other airlines' hubs are prominent. Small carriers tend to operate locally.

**Flight volume is on the increase - dramatically so since 2000. Structural shifts (below) in the flight load for airports lead to minimal average delays (above, right) in '02 and '03. Delays have been on the increase since.**

**Volume of Major Airports: Atlanta, Chicago, Seattle, St. Louis**

**Large changes in airports' daily flight volume are triggered by different events, including strikes, FAA order, and seasonal fluctuations,**

**Delays**

ARRIVAL & DEPARTURE  
 Most flights have 0 delay, with fewer and fewer flights having increasing delays. A secondary peak occurs at 24 hours suggesting a limit of 24 hours delay is used by some carriers. Some data is likely incorrect, e.g. flights arriving 24 hours early.

BY SCHEDULED DEPARTURE TIME & YEAR  
 Delays increase as day progresses. In 2001 they show an overall decrease, likely to be structural change, maybe FAA policy. Delays deteriorated again after 2003.

BY DAY OF WEEK  
 Best days to travel and avoid delays are Saturdays, and Tuesdays or Wednesdays. Fridays are bad for delays.

BY AIRPORT  
 EWR (Newark) is the worst. ORD (Chicago O'Hare) is not good, but also has high volume. DFW (Dallas-Fort Worth) is relatively good - high traffic but relatively small delay. Weather plays a huge role in delays - any kind of precipitation, high winds, or reduced visibility increases delays (scatterplots above).

**Fuel Efficiency**

FUEL USE  
 Consumption versus distance flown. Three groups of carriers are operating in 2000 - American Airlines is moving into the lead. Southwest is moving out of the middle group, closing up with AA and has overtaken AA on distance flown and efficiency 2007-8.

ARRIVAL & DEPARTURE  
 Delays increase as day progresses. In 2001 they show an overall decrease, likely to be structural change, maybe FAA policy. Delays deteriorated again after 2003.

**Crosswinds**

FUEL EFFICIENCY  
 Smaller carriers more efficient, probably due to use of smaller planes. Larger carriers American is one of the least efficient, Southwest is most efficient. Hawaiian Airlines very inefficient.

**Ghosts of Flights**

CAN WE SEE WHAT IS NOT THERE?  
 Planes have, for reasons such as maintenance, weather, or schedule fly empty between airports as so-called **Ghosts**. By tracking individual planes, we reveal their paths, including situations, where a plane lands in a different airport than where it takes off later, i.e. a ghost.

Example: US Airways Aircraft N-881 - Ghostflight from PIT to RIC (222 miles)

Year	Month	Day	Dest.	ArrTime	Origin	Dest. Diverted	
1995	3	8	1182	1256	PIT	CVG	0
1995	3	8	1311	NA	CVG	PIT	1
1995	3	8	1913	2050	RIC	PIT	0
1995	3	8	2134	2300	PIT	MSY	0

Ghost Flight Totals: over 1 million flights since 1995, with an average distance between airports of 1000 miles, corresponding to about 1.5 million gallons of fuel. Since 2001 the number of ghost flights is cut, with major airlines also decreasing the fuel consumption. Smaller carriers are facing increasing costs from ghost flights.

**Believe it or not??**

Racing Balloons  
 Three of American Airlines' registered vehicles in the database are RAVEN hot air balloons. Based on the data, they cruise at an impressive average speed of 430 miles an hour. Fasten those hats!

I'm sorry, Sir, but your flight left 12h early!  
 According to the data, this was said to all passengers of 247 flights. Another 165 flights left at least 2h early.

"Within-City Hoppers" - if you're in a real time crunch  
 A total of 232,809 flights cover a distance of less than 50 miles. The shortest commercial flights occur between the New York airports La Guardia (LGA) and John F. Kennedy (JFK). The distance is 11 miles, which according to google.maps can be covered in 18 min by car, and according to data, takes 14 min of air time.

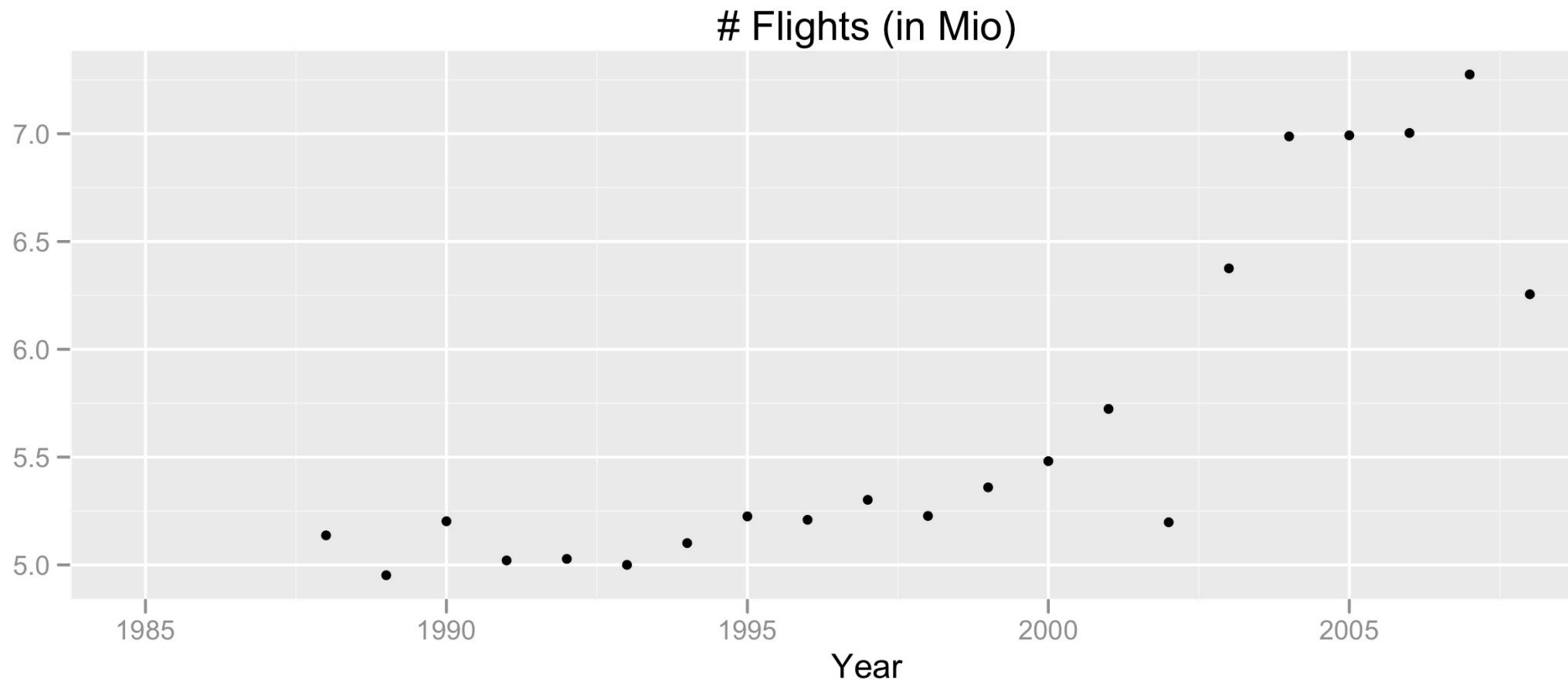
**Tools**

MySQL database on fast server (thanks to Ted Peterson)  
 R and packages  
 - ggplot2 (Hadley Wickham)  
 - DBI, RMySQL (David James, Jeff Horner)  
 Google Earth  
 supported by NSF grant # 0706949

# Analysis overview

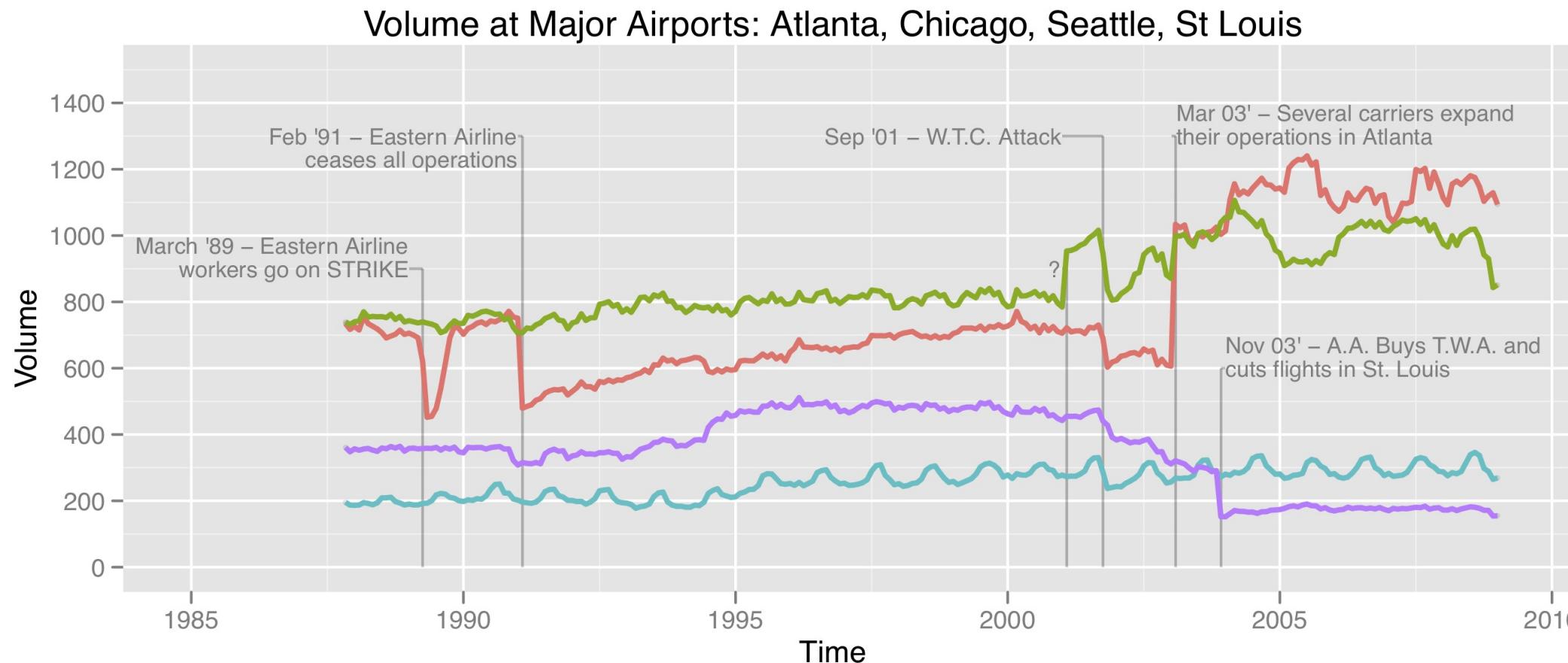
- ❖ Traffic patterns over time
- ❖ Delays
- ❖ Ghost flights
- ❖ Mapping traffic

# Traffic patterns over time



Number of flights in millions per year: steadily increasing volume until 2001, with a big drop in 2002. Volume recovered in 2003, and flattens 2004-7, with another drop in 2008. What happened in 2001? What was happening in 2008?

# Traffic patterns at selected airports

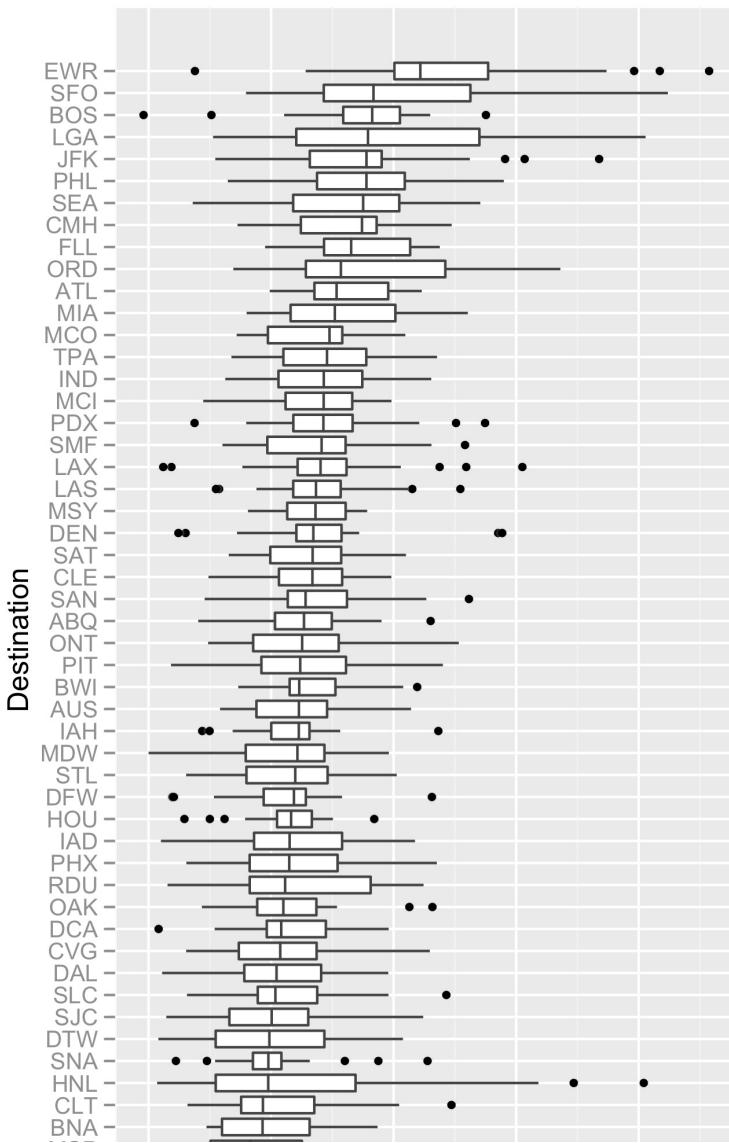


# Delays

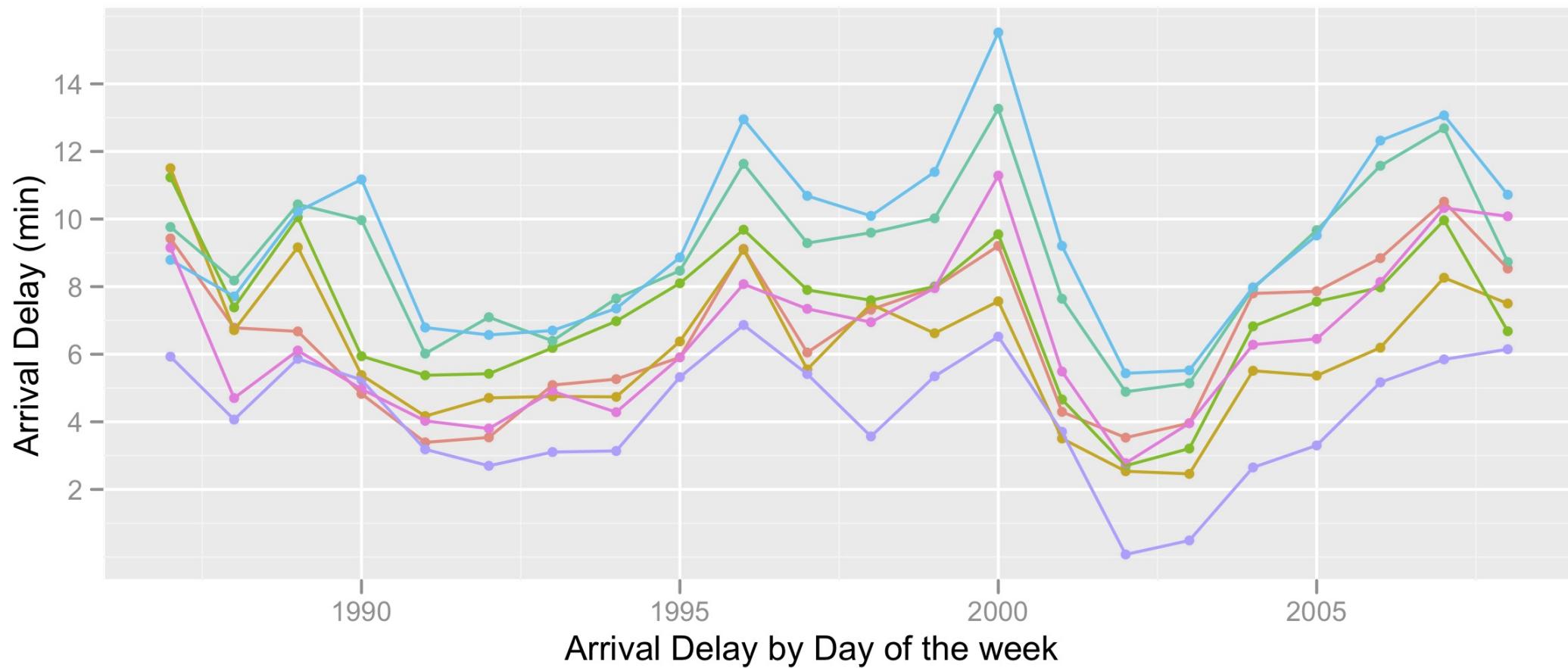
# Delays, by year

# Delays, by carrier

# Delays, by airport



# Delays, by day



# Fuel use by carrier

# Fuel efficiency

# Ghost flights

## CAN WE SEE WHAT IS NOT THERE?

Planes have, for reasons such as maintenance, weather, or schedule empty between airports as so-called *Ghosts*. By tracking individual we reveal their paths, including situations, where a plane lands in a ent airport than where it takes off later, i.e. a ghost:

Example: US Airways Aircraft N-881 - Ghostflight from PIT to RIC (222 miles)

Year	Month	Day	DepTime	ArrTime	Origin	Dest	Diverted
1995	3	8	1102	1256	PIT	CVG	0
1995	3	8	1311	NA	CVG	PIT	1
1995	3	8	1913	2050	RIC	PIT	0
1995	3	8	2134	2300	PIT	MSY	0

# Ghost flights, wasted fuel

# What tools were used and why

A subset of the analysis materials including data and code can be downloaded from the [paper site](#)

- ✗ sqlite database: Inspired by the guidelines provided by the organisers we created a mysql database, on a central server that all team members could access with a password. Each person accessed the data through R.
- ✗ R packages: RMySQL, DBI, ggplot2

# A brief introduction to working with databases

Working from these notes

<https://db.rstudio.com/databases/sqlite/>

Why should I use a database?

- ✗ The data is too large to load into memory, ie work directly with it in R
- ✗ Database can make more efficient calculations
- ✗ Only load the data needed for specific analysis tasks

# Connecting to an existing database

The packages `DBI`, `RMySQL`, `RPostgreSQL`, `RSQLite`, `bigrquery`, `odbc` enable connection to many different types of databases. The package `dbplyr` enables tidy style access to the databases.

# Set up connection

Download the [supplementary material](#) for Hofmann et al (2012) and you will find:

```
library(RMySQL)
m <- dbDriver("MySQL")
co <- dbConnect(m, user="2009Expo",
                password="R R0cks",
                port=3306,
                dbname="data_expo_2009",
                host="headnode.stat.iastate.edu")
```

✋ Stop, read the Securing Credentials documentation

# SQL

```
dtme <- dbGetQuery(co, "select Year,  
avg(CRSArrTime), avg(ArrTime),  
count(*) as count,avg(ArrDelay) from  
ontime group by Year, (CRSArrTime div 100)")
```

```
qplot(`avg(CRSArrTime)` , `avg(ArrDelay)` ,  
      geom="point" ,  
      data=subset(dtme, Year > 1998) ,  
      xlab="Scheduled Arrival" ,  
      ylab="Average Arrival Delay (in mins)") +  
      facet_wrap(facets=~Year, ncol=5) +  
      geom_hline(yintercept=c(0,15))
```

# Set up connection, using SQLite

```
# Set up connection  
library(DBI)  
library(RSQLite)  
con <- dbConnect(RSQLite::SQLite(), ":flights:")
```

This creates the link between R and the database.

# Suppose we want to set up a database

One month of air traffic data is quite manageable in an R session. We can use this to get started.

To populate our SQLite database with airlines data, you need

```
copy_to(con, d, "flights",
  temporary = FALSE,
  indexes = list(
    c("FlightDate",
      "Reporting_Airline",
      "Tail_Number",
      "Origin",
      "Dest"
    )
  ))
```

Setting up the indexes makes it faster to process data on the database.

## Alternative approach using DBI database functions

```
dbWriteTable(con, "flights", d)  
dbListTables(con)
```

and check it

```
flights_db <- tbl(con, "flights")  
flights_db
```

```
feb1 <- flights_db  %>%
  filter(DayofMonth==1) %>%
  select(DayofMonth, Origin, Dest) %>%
  collect()
feb1
```

or with SQL

```
dbListFields(con, "flights")
res <- dbSendQuery(con, "SELECT * FROM
                           flights WHERE DayofMonth=1")
firstday <- dbFetch(res)
```

# Add a table on airport details

Information about airport location and details is found in a different table at the BTS site:

[https://www.transtats.bts.gov/Fields.asp?Table\\_ID=288](https://www.transtats.bts.gov/Fields.asp?Table_ID=288) . We will download this and add to our database to use for plotting flights on a map.

```
airports <- read_csv("data/402312038_T_MASTER_CORD.csv") %>%  
  select(-X29)  
copy_to(con, airports, "airports",  
       temporary = FALSE  
     )  
dbListTables(con)
```

# Its easy to forget what variables are in the table

You can check this with

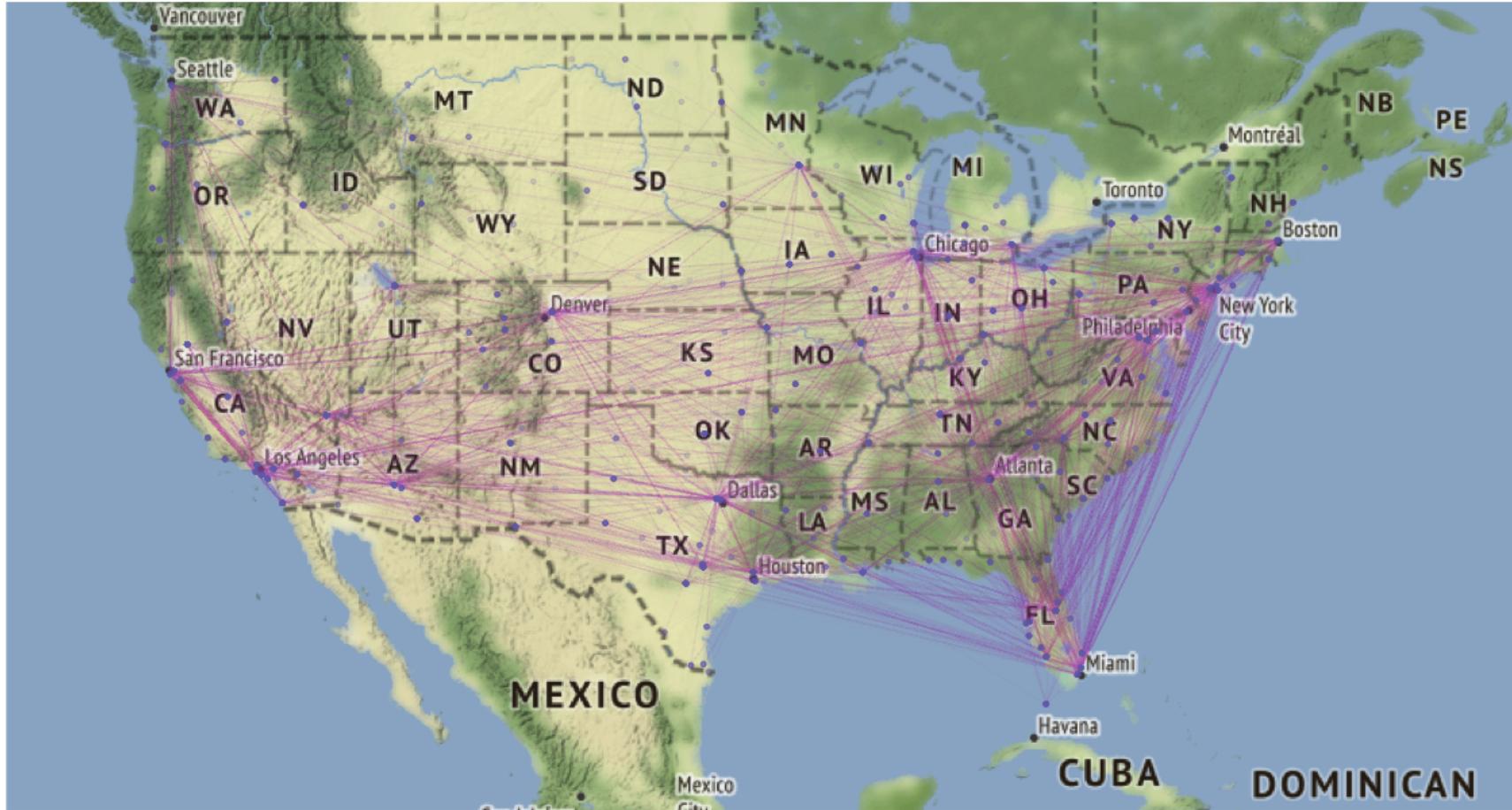
```
dbListFields(con, "airports")
```

# Make a map of flights for Feb 1

```
airport_locations <-tbl(con, "airports") %>%  
  filter(AIRPORT_IS_LATEST == 1, AIRPORT_COUNTRY_CODE_ISO == "US") %>%  
  select(AIRPORT, DISPLAY_AIRPORT_NAME, LONGITUDE, LATITUDE) %>%  
  collect()  
  
feb1_flights <- feb1 %>%  
  left_join(airport_locations, by=c("Origin" = "AIRPORT")) %>%  
  rename(Origin_lon = LONGITUDE, Origin_lat = LATITUDE,  
         Origin_name = DISPLAY_AIRPORT_NAME) %>%  
  left_join(airport_locations, by=c("Dest" = "AIRPORT")) %>%  
  rename(Dest_lon = LONGITUDE, Dest_lat = LATITUDE,  
         Dest_name = DISPLAY_AIRPORT_NAME)
```

```
library(ggmap)
usa_bbox <- c(-130, # min long
              20, # min lat
              -60, # max long
              50) # max lat
usa_map <- get_map(location = usa_bbox, source = "osm")
ggmap(usa_map)
```

```
library(ggthemes)
ggmap(usa_map) + geom_segment(data=feb1_flights,
                               aes(x=origin_lon,
                                   xend=Dest_lon,
                                   y=origin_lat,
                                   yend=Dest_lat),
                               colour="#9651A0", alpha=0.01) +
geom_point(data=feb1_flights, aes(x=origin_lon, Origin_lat),
           colour="#746FB2", alpha=0.1, size=1) +
theme_map()
```



# Animating flights for one day



Data Visualization and Statistical Graphics in Big Data Analysis:  
Video 1 from Annual Reviews on Vimeo.

Code is here

# That's it!



This work is licensed under a [Creative Commons Attribution-ShareAlike 4.0 International License](#).

Lecturer: Dianne Cook

Department of Econometrics and Business Statistics

✉ ETC5512.Clayton-x@monash.edu