Teaching precursors to data science introductory and intermediate statistics courses

Nicholas J. Horton

Department of Mathematics and Statistics
Amherst College, Amherst, MA, USA

July 16, 2014

nhorton@amherst.edu

Acknowledgements

- joint work with Ben Baumer (Smith College) and Hadley Wickham (Rice/RStudio)
- supported by NSF grant 0920350 (building a community around modeling, statistics, computation and calculus)
- more information at `http://www.mosaic-web.org`
- examples at `http://www.amherst.edu/~nhorton/icots2014`

Motivation

> *Undoubtedly the greatest challenge and opportunity that confronts today's statisticians is the rise of Big Data: databases on the human genome, the human brain, Internet commerce, or social networks (to name a few) that dwarf in size any databases statisticians encountered in the past.*

(Future of Statistics report (2014), `bit.ly/londonreport`)

Motivation (cont.)

Big Data is a challenge for several reasons:

1. problems of scale
2. different kinds of data
3. additional skills

Motivation (cont.)

- Cobb argued (TISE, 2007) that our courses teach techniques developed by pre-computer-era statisticians as a way to address their lack of computational power
- Do our students see the potential and exciting use of statistics in our classes? (Gould, ISR, 2010)
- How do we prepare them to answer complex questions using richer data?

Undergraduate programs in statistics working group

Draft guidelines suggest specific skill areas:

- Statistical
- Computational
- Data-related
- Mathematical
- Communication

Are we teaching these in our current programs?

Undergraduate programs in statistics working group

Draft guidelines suggest specific skill areas:

- **Statistical**
- **Computational**
- **Data-related**
- Mathematical
- **Communication**

Key "Data Science" topics bolded

Building precursors to data science (and “bigger” data)

How to accomplish this?

- start in the first course
- build on capacities in the second course
- develop more opportunities for students to apply their knowledge in practice (internships, collaborative research, teaching assistants)
- new courses focused on “Data Science” (e.g., Baumer at Smith College)
- “Data Expo” and “Data Fest” opportunities (Gould, *Teaching Statistical Thinking in the Data Deluge*, 2014)
- today’s goal: talk about what can be done in the first and second courses

Data Expo 2009

Ask students: have you ever been stuck in an airport because your flight was delayed or cancelled and wondered if you could have predicted it if you'd had more data? (Wickham, JCGS, 2011)

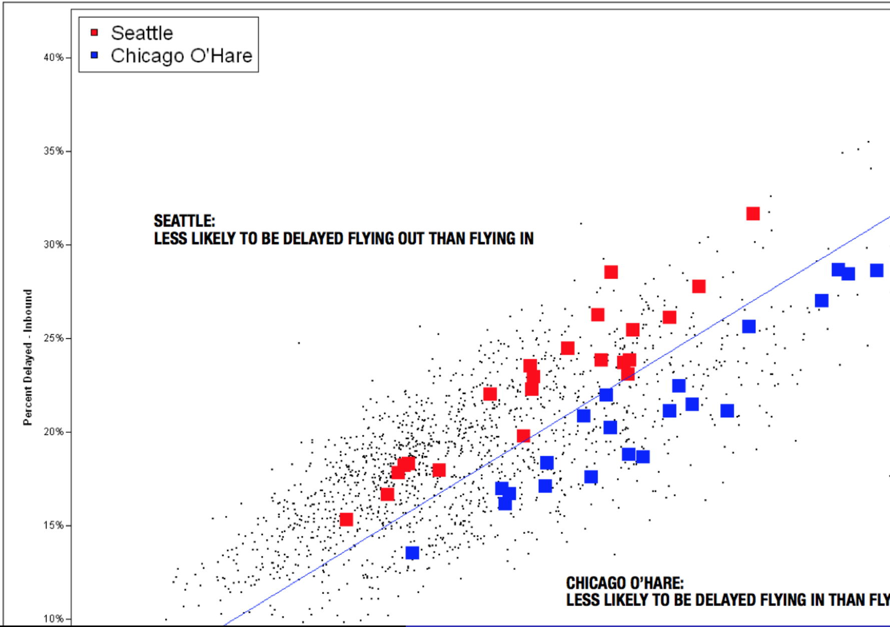
Data Expo 2009

Ask ICOTS attendees: have you ever been stuck in Flagstaff because your flight was delayed or cancelled and wondered if you could have predicted it if you'd had more data?

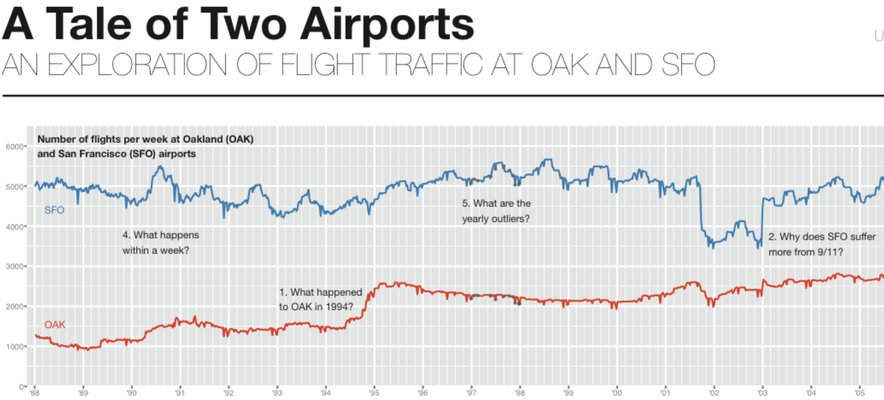
Data Expo 2009

- dataset of flight arrival and departure details for all commercial flights within the USA, from October 1987 to March 2014
- large dataset: more than 150 million records
- aim: provide a graphical summary of important features of the data set
- Expo winners presented at the JSM in 2009; details at `http://stat-computing.org/dataexpo/2009`

Data Expo 2009 winners



Data Expo 2009 winners



A Tale of Two Airports

AN EXPLORATION OF FLIGHT TRAFFIC AT OAK AND SFO

Number of flights per week at Oakland (OAK) and San Francisco (SFO) airports

1. What happened to OAK in 1994?

2. Why does SFO suffer more from 9/11?

4. What happens within a week?

5. What are the yearly outliers?

Find an interesting question

During the month of July in the past few years, what is the distribution of delays for flights leaving Flagstaff?

- what proportion of flights were cancelled?
- what proportion of flights were delayed (15 minutes or more) or cancelled?
- what is the average delay?
- how the average delay relate to time of day?

Accessing the database

Need to utilize a database system (using SQL, structured query language) to easily analyze of this size

```
# establish the connection
require(RMySQL)
con = dbConnect(MySQL(), host="rucker.smith.edu",
  dbname="airlines")
ds = dbGetQuery(con, "SELECT DayofMonth, Month, Year,
  Origin, Dest, UniqueCarrier, TailNum, CRSDepTime ,
  ArrDelay, Cancelled FROM ontime WHERE
    Origin='FLG' AND Year > 2010 AND Month = 7")
```

This returns a data frame which can be analyzed in R
(This is not hard to set up or access.)

What happened on Saturday, July 20, 2014

```
  Day Month Year Origin Dest DepTime ArrDelay Cancelled
1  20     7 2013    FLG  PHX     650       -4         0
2  20     7 2013    FLG  PHX    1030       14         0
3  20     7 2013    FLG  PHX    1200       -8         0
4  20     7 2013    FLG  PHX    1500       -8         0
5  20     7 2013    FLG  PHX    2105      152         0
```

Proportion delayed ($>$ 15 min) or cancelled

Among $n = 553$ flights in July 2011, July 2012, and July 2013

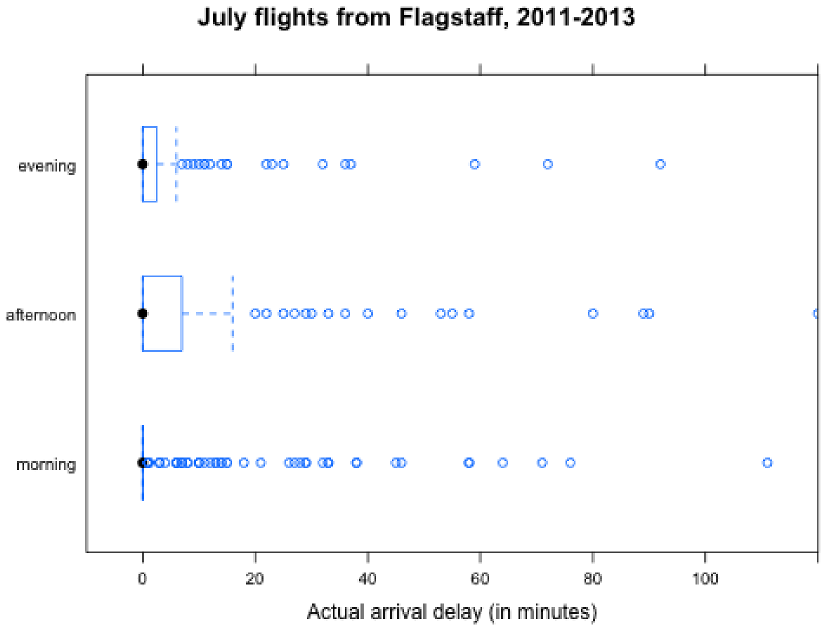
```
> tally(~ Cancelled, format="percent", data=ds)
    0      1
96.93  3.07

> # delayed if ArrDelay > 15
> tally(~ delayorcancel, format="percent", data=ds)
  no  yes
86.3 13.7
```

Distribution of Arrival Delay by Time of Day

```
> favstats(ActDelay ~ TimeOfDay, data=ds)
    group min Q1 med  Q3 max  mean   sd   n miss
  morning   0  0   0 0.0 250  5.95 22.9 274    5
afternoon   0  0   0 6.5 300 15.01 44.5 166   11
  evening   0  0   0 2.5 152  7.24 21.1  95    2
```

Distribution of Arrival Delay by Time of Day



How to introduce? (first course)

Garfield et al: Model Eliciting Activity http://serc.carleton.edu/sp/library/mea/examples/example5.html

- how would you determine if one airline was more reliable than another?
- give students a small sample from the airlines dataset for one city pair for two airlines
- Is there a difference in the reliability as measured by arrival time delays for these two regional airlines out of Chicago? Or are both airlines pretty much the same in terms of their arrival time delays?
- original MEA requires no technology

How to introduce? (first course)

- use R Markdown as a mechanism to simplify access to code (see Baumer et al, *TISE* 2014, "R Markdown: Integrating A Reproducible Analysis Tool into Introductory Statistics")
- provide scaffolding for extensions
- prepare datasets for students to answer specific questions of their own
- let them explore the performance of their "rules" on samples (or the whole population of flights)
- visualize larger datasets (and start thinking about data cleaning and consistency checking)
- database system is hidden to them

How to introduce? (second course)

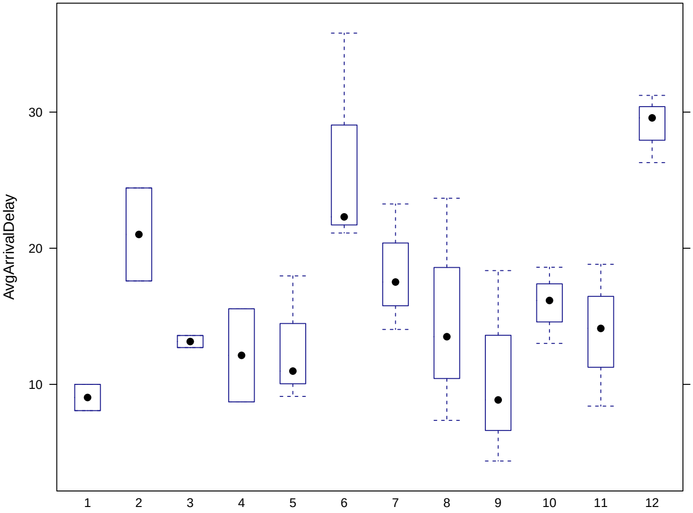
Start thinking about specific learning outcomes for data management and computation

- introduce a framework for the fundamentals of data management
- Hadley Wickham's 6 key verbs for "Tidy Data": arrange, filter, mutate, select, group by
- introduce students to database systems
- scaffold using R Markdown (to allow reproducibility and minimize need to start from scratch)
- focus on telling a story using data (as always, in the context of answering a statistical question)

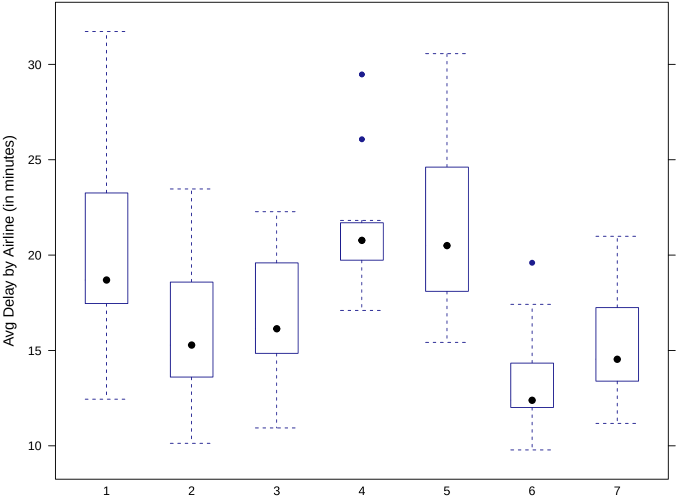
How to introduce? (second course)

- answer other questions
- merge other tables (e.g., information about airports, individual planes)
- visualize large datasets
- communicate results

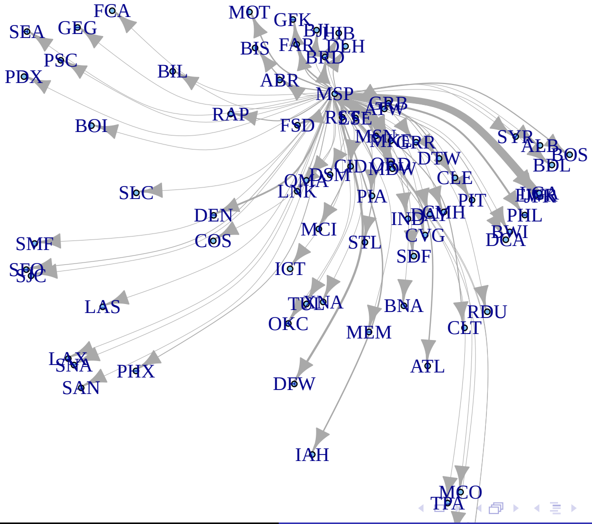
Which month is it best to travel (airline averages/BDL)?



Which day is it best to travel (airline averages from BDL)?



Maps and visualization



Undergraduate programs in statistics working group

Draft guidelines suggest specific skill areas:

- **Statistical**
- **Computational**
- **Data-related**
- Mathematical
- **Communication**

Key "Data Science" topics bolded

Closing thoughts

- MEA's bring big ideas into the classroom ("excitement of statistics")
- we need to think more about teaching data related skills (ability to "think with data" as described by Diane Lambert of Google)
- markdown helps simplify the use of more sophisticated code (and can be introduced early in introductory statistics)
- this pair of technologies helps to allow instructors (and in later classes, students) to tackle more interesting questions

Teaching precursors to data science introductory and intermediate statistics courses

Nicholas J. Horton

Department of Mathematics and Statistics
Amherst College, Amherst, MA, USA

July 16, 2014

nhorton@amherst.edu

Examples at at `http://www.amherst.edu/~nhorton/icots2014`

Background on databases and SQL

- no technology needed for initial MEA
- modest investment can allow use of a rich dataset
- instructors need some background on databases and SQL
- relational databases (invented in 1970)
- like electronic filing cabinets to organize masses of data (terabytes)
- fast and efficient
- useful reference: *Learning MySQL*, O'Reilly 2007

Creating the airline delays database (approx. 1 hour for SQLite)

1. download and install SQLite from `sqlite.org`
2. download the data (1.6gb compressed, 12gb uncompressed)
3. create a table with fields that match the csv files
4. load the data with the `.import` directive
5. add indices (to speed up access to the data, takes some time)
6. install and load the `RSQLite` package
7. establish a connection (using `dbConnect()`)
8. start to make selections (which will be returned as data frames) using the `dbGetQuery()` function