Thinking with Data in the Second Course

Nicholas J. Horton

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

August 4, 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/jsm2014

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.)

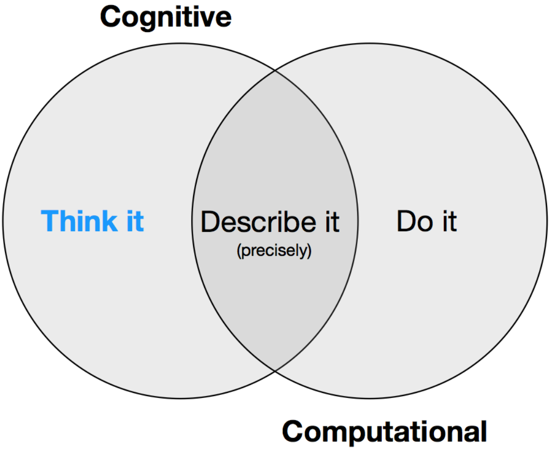
> *Data science is the study of the generalizable extraction of knowledge from data, yet the key word is science. It incorporates varying elements and builds on techniques and theories from many fields, including signal processing, mathematics, probability models, machine learning, statistical learning, computer programming, data engineering, pattern recognition and learning, visualization, uncertainty modeling, data warehousing, and high performance computing with the goal of extracting meaning from data and creating data products.*

Wikipedia, `https://en.wikipedia.org/wiki/Data_science`, accessed July 31, 2014

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?
- These are necessary precursors to move towards bigger data

Computational overview (Wickham)



Undergraduate programs in statistics working group

Draft guidelines suggest specific skill areas:

www.amstat.org/education/curriculumguidelines.cfm

- Statistical Methods and Theory
- Computational/Data-related
- Mathematical
- Statistical Practice

Are we teaching these in our current programs?

Undergraduate programs in statistics working group

Draft guidelines suggest specific skill areas:

`www.amstat.org/education/curriculumguidelines.cfm`

- **Statistical Methods and Theory**
- **Computational/Data-related**
- Mathematical
- **Statistical Practice**

Key “Data Science” topics bolded

Building precursors to data science (and "bigger" data)

How to accomplish this?

- start in the first course (using approach outlined by Pruim)
- build on capacities in the second course
- develop more opportunities for students to apply their knowledge in practice (internships, collaborative research, teaching assistants: see Legler's talk)
- new courses focused on "Data Science" (e.g., Baumer at Smith College, see related Wednesday 10:30am session)
- "Data Expo" and "Data Fest" opportunities (Gould, *Teaching Statistical Thinking in the Data Deluge*, 2014 and session on Wednesday at 2:00pm)
- today's goal: what can be done in the second course?

Possible models for the second course

- Intermediate Statistics/Applied Regression
- Data Science/Statistical Computing
- “Foundations of Statistics” (formerly Mathematical Statistics)
- Data and Computing Fundamentals (1 credit course)

Intermediate Statistics/Applied Regression

- often taught from the Sleuth or the STAT2 text
- usually provides predigested datasets
- range of statistical topics
- projects provide opportunity to build capacities
- could add new data-related learning outcomes early, then reinforce using projects

Data Science/Statistical Computing

- see `http://www.stat.berkeley.edu/~statcur` or Baumer's course at Smith College
- explicit focus on computing
- grounded in answering a statistical question
- projects provide even more opportunity to build capacities
- typically a new course

“Foundations of Statistics” (formerly Math Stats)

- still sometimes the first course beyond intro!
- still often reflects curricular choices of Hogg and Craig
- relatively rare to include computing or real data (but see Nolan and Speed’s *Stat Labs*)
- lots of opportunities to reformulate (see Horton “I hear I forget” *TAS* 2013 paper) to include more varied data

Data and Computing Fundamentals (1 credit course)

Week 1: Introduction, data files, documentation markup, and elementary data visualization

Week 2: Relational database operations, intermediate data visualization

Week 3: More data operations, map visualization

Week 4: Basic models, fitting, and summaries

Week 5: Clustering

Week 6: Dimension reduction

Week 7: Putting it all together

`www.macalester.edu/hhmi/curricularinnovation/data`

What to include?

- framework for data wrangling
- more complex data formats and technologies
- reliable workflow and reproducible analyses
- precursors to bigger data
- grounded in answering a statistical question of some substance

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)

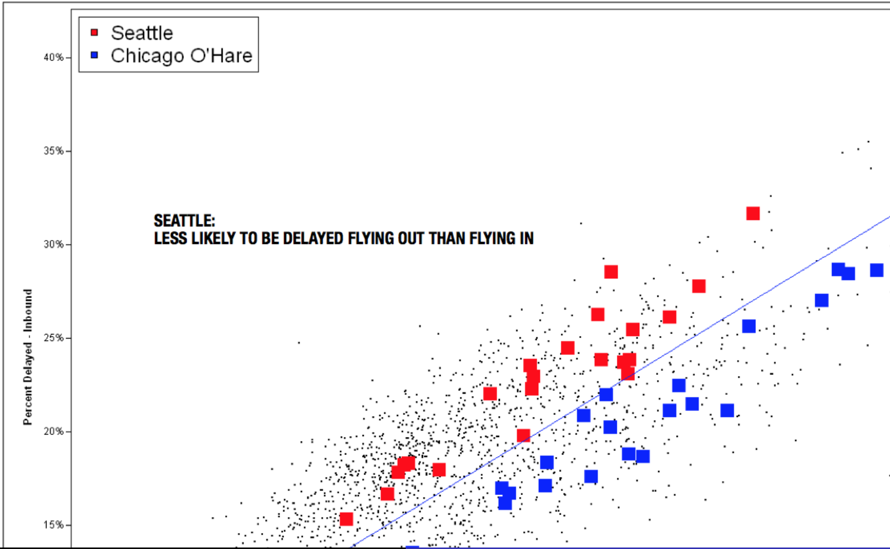
Data Expo 2009

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

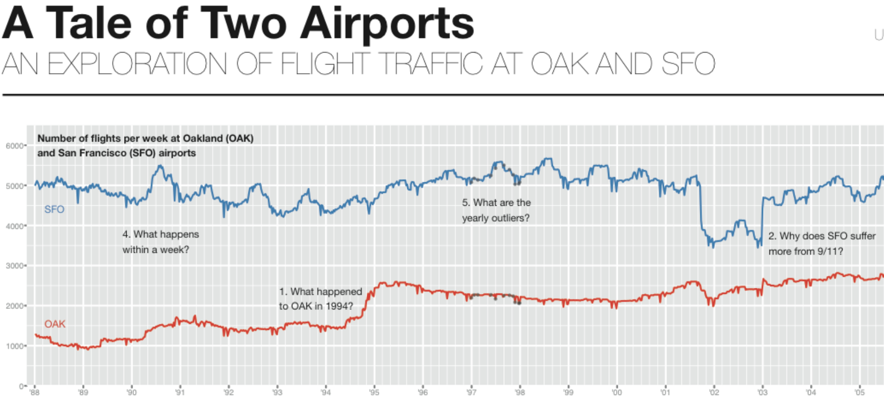
Data Expo 2009

- 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 August in the past few years, what is the distribution of delays for flights leaving Boston's Logan Airport?

- 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

```
require(dplyr)
my_db = src_mysql(host="rucker.smith.edu",
  user="mth292", password="XX", dbname="airlines")
my_tbl = tbl(my_db, "ontime")
bosFlights = my_tbl %>%
  filter(Origin=="BOS" & Year > 2010 & Month==8) %>%
  select(DayofMonth, Month, Year, Origin, Dest,
    UniqueCarrier, TailNum, CRSDepTime, ArrDelay,
    Cancelled)
```

This returns a data frame which can be analyzed in R (using piping syntax from dplyr)

(Brief) 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
- free course: `class.stanford.edu/courses/DB/2014/SelfPaced/about`

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` and `dplyr` packages
7. establish a connection and start to make selections using functions in `dplyr`

What happened on Thursday, August 8th, 2013?

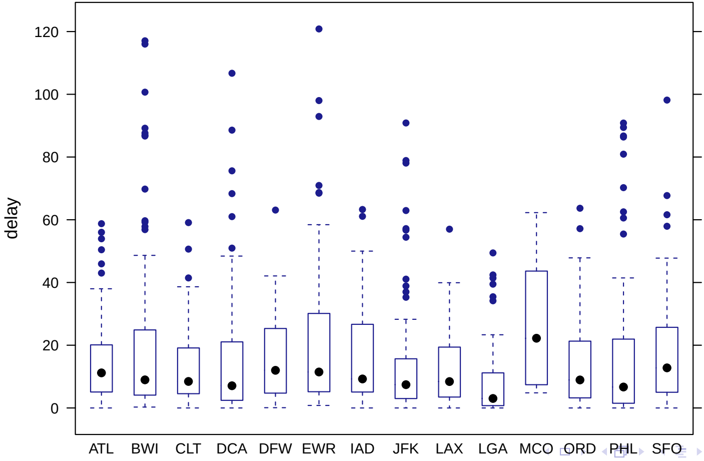
```
> dim(oneday)
[1] 354   8
> head(oneday)
  DayofMonth Month Year Origin Dest CRSDepTime ArrDelay
1          8     8 2013    BOS  JFK        600      -14
2          8     8 2013    BOS  MEM       1635       34
3          8     8 2013    BOS  CVG        630       -6
4          8     8 2013    BOS  CVG       1040      -12
5          8     8 2013    BOS  ORF       1905      -17
6          8     8 2013    BOS  ORF        840      -24
```

Proportion delayed (> 15 min) or cancelled

Among $n = 29,442$ flights in August 2011, August 2012, and August 2013

```
> favstats(DelayOrCancel ~ TimeOfDay, data=bosFlights)
     .group min Q1 median Q3 max  mean    sd     n missing
1   morning   0  0      0  0   1 0.163 0.369 14130       0
2 afternoon   0  0      0  1   1 0.285 0.451 10004       0
3   evening   0  0      0  1   1 0.376 0.485  5308       0
```

Distribution of Average Daily Arrival Delay for Common Destinations



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 (pick their own airports? city pairs? season?)
- 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)

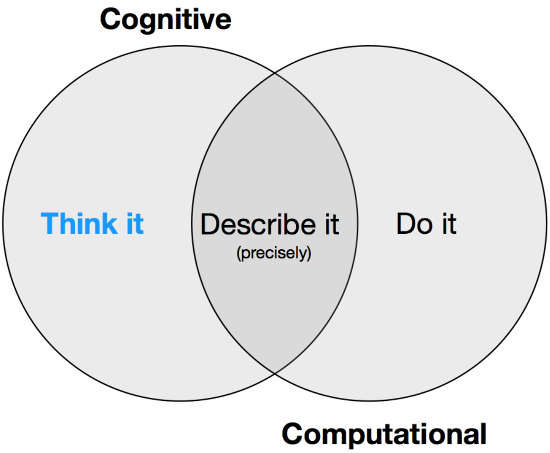
Start thinking about specific learning outcomes for data management and computation

- introduce a framework for the fundamentals of data management (see Finzer's "Data Habits of Mind" paper in *TISE*, 2013)
- basic data manipulation (Wickham's "Tidy Data")
- 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)

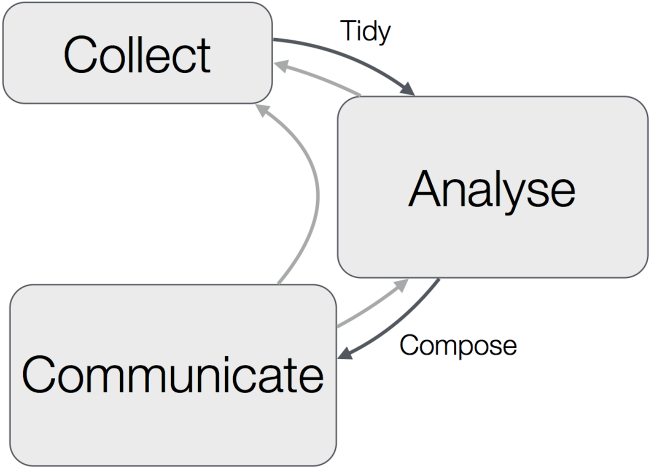
How to introduce? (second course)

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

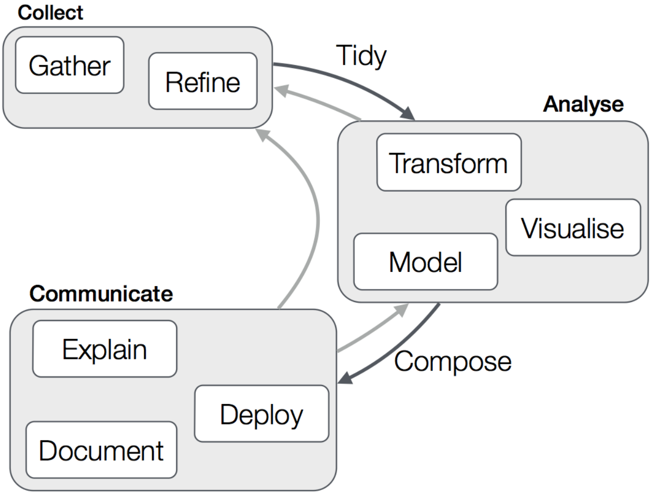
Computational overview (Wickham)



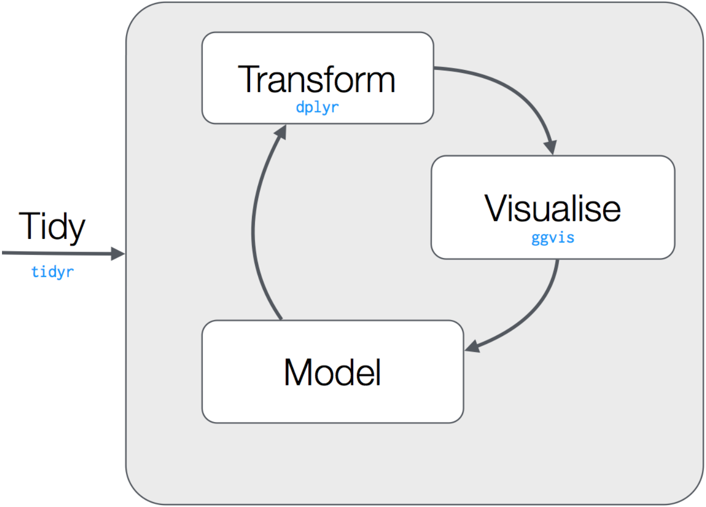
Data science cycle (Wickham)



Data science cycle (Wickham)



Statistical data analysis cycle (Wickham)



Key verbs (Wickham)

filter:	keep rows matching criteria
select:	pick columns by name
arrange:	reorder rows
mutate:	add new variables
summarise:	reduce variables to values
group by:	collapse groups

Other key idioms (Wickham)

join: variety of merges

gather: gather columns (to reshape)

spread: inverse of gather

Small set of powerful tools (“less volume, more creativity”)

Other related and useful resources

- http://blog.rstudio.org/2014/08/01/the-r-markdown-cheat-sheet/
- http://blog.rstudio.org/2014/07/31/httr-0-4/
- http://blog.rstudio.org/2014/07/23/new-data-packages/
- http://blog.rstudio.org/2014/07/16/rstudio-presents-essential-tools-for-data-science-with
- http://blog.rstudio.org/2014/06/23/introducing-ggvis/

Undergraduate programs in statistics working group

Draft guidelines suggest specific skill areas: `www.amstat.org/education/curriculumguidelines.cfm`

- **Statistical Methods and Theory**
- **Computational/Data-related**
- Mathematical
- **Statistical Practice**

Need to establish a framework and process to introduce and reinforce these skills ("less volume, more creativity")

Closing thoughts

- MEA's bring big ideas into the classroom ("excitement of statistics")
- we need to revise our second courses to ensure that students develop more sophisticated data related skills (ability to "think with data" as described by Diane Lambert of Google)
- Wickham's `tidyr` and `dplyr` plus `ggvis` (different talk) facilitate moving in this direction
- markdown helps simplify the use of more sophisticated code (and can be introduced early in introductory statistics)
- allows students to tackle more interesting questions

Thinking with Data in the Second Course

Nicholas J. Horton

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

August 4, 2014

nhorton@amherst.edu

Examples and slides at
`http://www.amherst.edu/~nhorton/jsm2014`