Teaching precursors to data science in introductory and second courses in statistics

Nicholas Horton, nhorton@amherst.edu

CAUSE Teaching and Learning webinar, February 24, 2015

Resources available at http://www.amherst.edu/~nhorton/precursors

Teaching precursors to data science in introductory and second courses in statistics

CAUSE Teaching and Learning webinar: Nicholas Horton

While we wait, please check out the related resources, papers, and materials at:

http://www.amherst.edu/~nhorton/precursors

Questions for the speaker can be submitted at any time

Acknowledgements

- Joint work with Ben Baumer, Hadley Wickham, Danny Kaplan, and Randy Pruim
- Funded by NSF grant 0920350 (Phase II: Building a Community around Modeling, Statistics, Computation, and Calculus)
- More information at http://mosaic-web.org

Plan and outline

- Statistics students need to develop the capacity to make sense of the staggering amount of information collected in our increasingly data-centered world
- Data science is an important part of modern statistics, but our introductory and second statistics courses often neglect this fact
- This webinar discusses ways to provide a practical foundation for students to learn to “compute with data” as defined by Nolan and Temple Lang (2010), as well as develop “data habits of mind” (Finzer, 2013)

Plan and outline

- By introducing students to commonplace tools for data management, visualization, and reproducible analysis in data science and applying these to real-world scenarios, we prepare them to think statistically in the era of big data
- See resources page or https://www.causeweb.org/ecots/ecots14/31/ for background on R markdown (all examples in this talk are available using this technology)

Why are data-related skills important?

- McKinsey & Company report stated that “by 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions”.
- A large number of those workers will be at the bachelors level (and the vast majority will not major in statistics)
- How do we ensure that they have the appropriate training to be successful?

CUPM 2015

Cognitive Recommendation 3:

- Students should learn to use technological tools. Mathematical sciences major programs should teach students to use technology effectively
- Use of technology should occur with increasing sophistication throughout a major curriculum.

Content Recommendation 3: Mathematical sciences major programs should include concepts and methods from data analysis, computing, and mathematical modeling.

ASA's undergraduate guidelines (endorsed 2014)

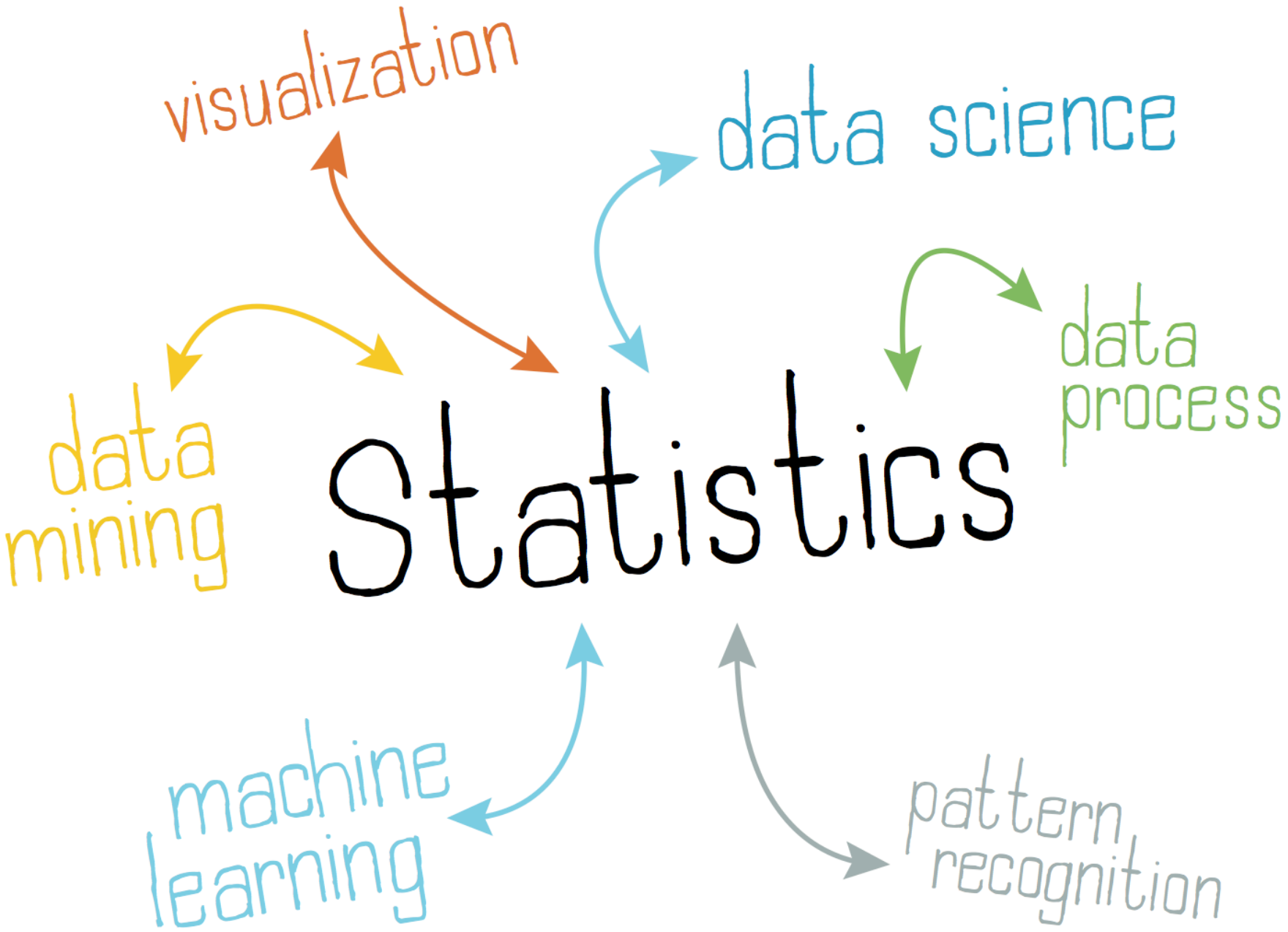
The American Statistical Association endorses the value of undergraduate programs in statistics as a reflection of the increasing importance of the discipline. We expect statistics programs to provide sufficient background in the following core skill areas: statistical methods and theory, data manipulation, computation, mathematical foundations, and statistical practice. Statistics programs should be flexible enough to prepare bachelor's graduates to either be functioning statisticians or go on to graduate school.

Other calls

- Finzer (TISE, 2013) “The data science education dilemma”, http://escholarship.org/uc/item/7gv0q9dc
- Carver and Stephens (ICOTS, 2014) “It is time to include data management in introductory statistics, http://iase-web.org/icots/9/proceedings/pdfs/ICOTS9_C134_CARVER.pdf
- Wickham (JSS, 2014) “Tidy data”, http://www.jstatsoft.org/v59/i10/paper

Key skills

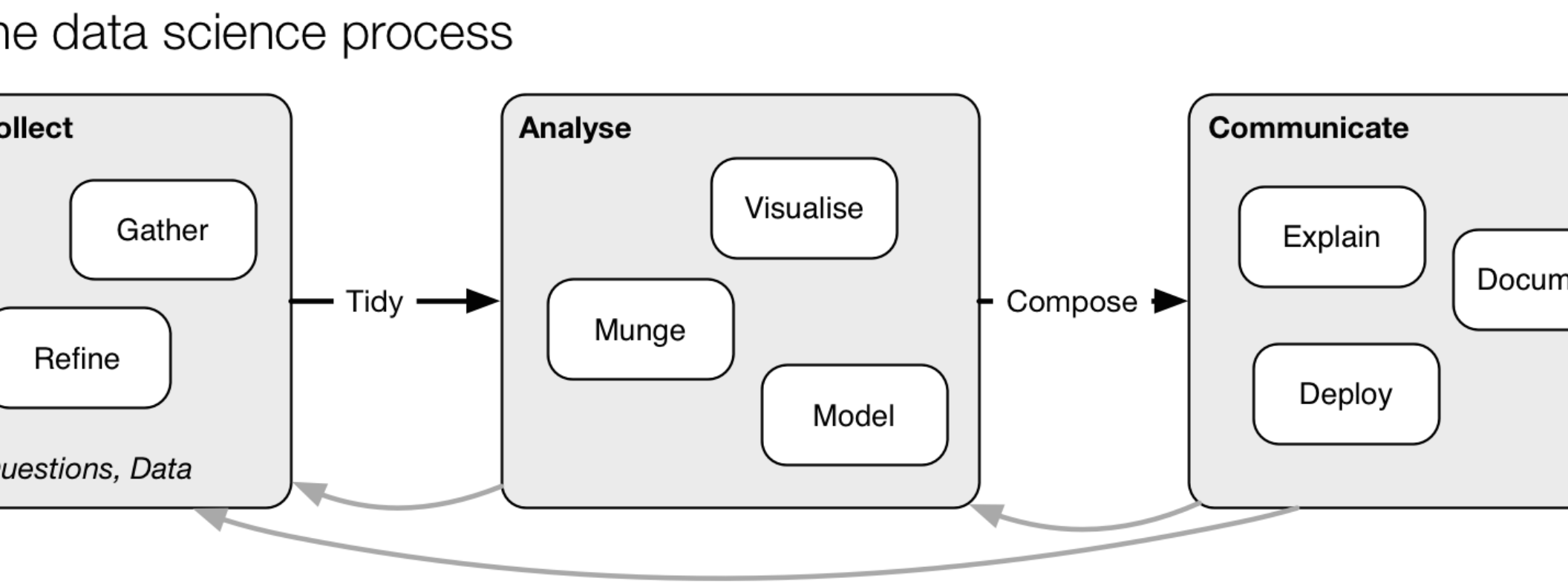
- Effective statisticians at any level display an integrated combination of skills that are built upon statistical theory, statistical application, data manipulation, computation, and communication
- Students need scaffolded exposure to develop connections between statistical concepts and theory and their application to statistical practice
- Students need to be able to “think with data” to solve problems


visualization
data science
data
process
data
mining
Statistics
machine
learning
pattern
recognition

Key changes: importance of data science

- Working with data requires extensive computing skills far beyond those described in the previous guidelines
- Students need facility with professional statistical analysis software, the ability to access and “wrangle” data in various ways, and the ability to utilize algorithmic problem-solving
- Students need to be able to be fluent in higher-level languages and be facile with database systems

A view of the data science process (Wickham)



Data-related topics (undergrad guidelines)

- Use of one or more professional statistical software environments
- Data analysis skills undertaken in a well-documented and reproducible manner
- Students should be able to manage and manipulate data, including joining data from different sources and restructuring data into a form suitable for analysis

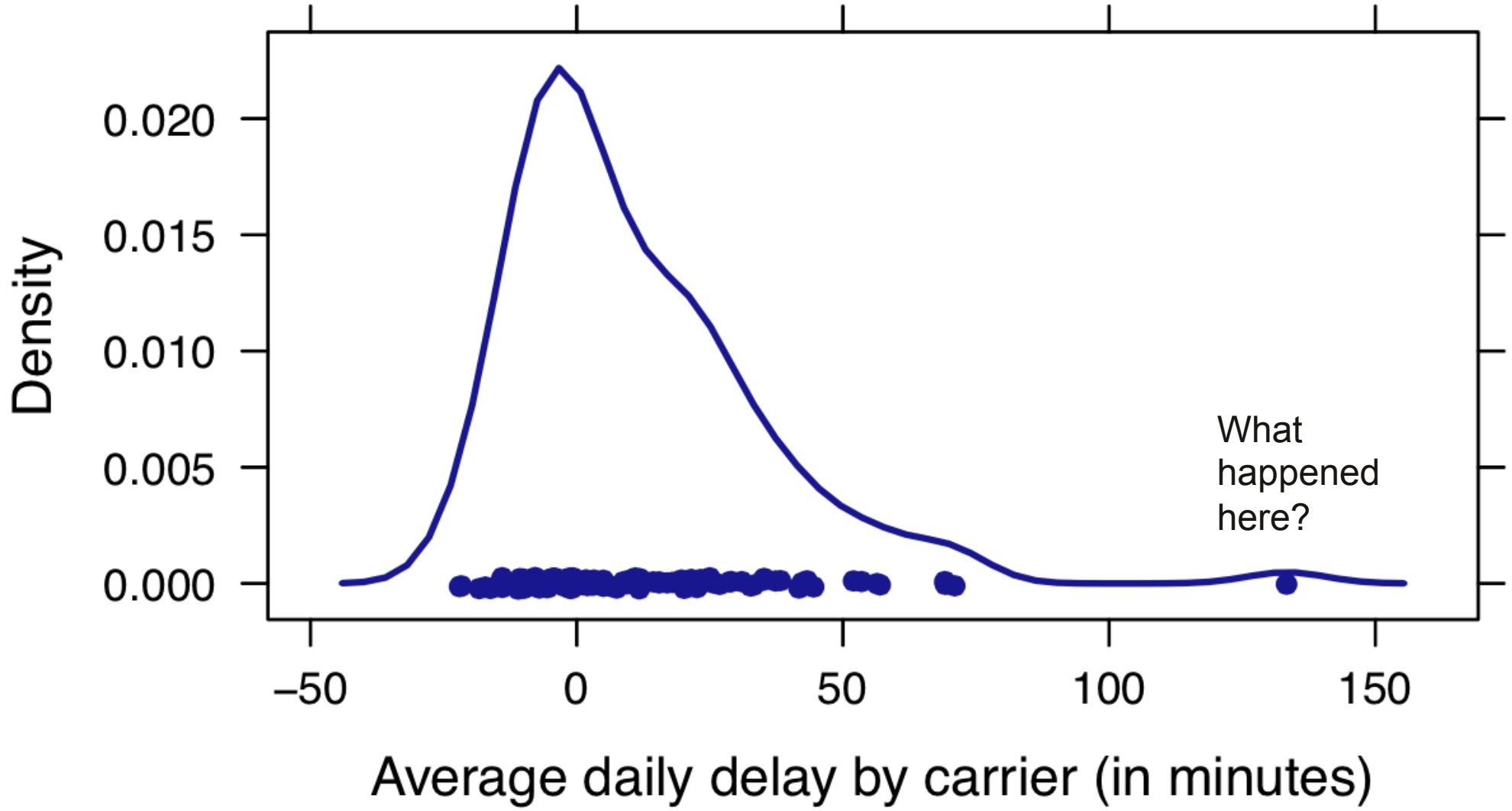
Case study: flight delays

- Ask students: have you ever been stuck in an airport because your flight was delayed or cancelled and wondered if you could have predicted the delay if you'd had more data?
- Enter the airline delays dataset:
  - More than 180 million flights since 1987 (needs database: different webinar, but see resources)
  - nycflights13 package in R (n=336,776 flights)

Key verbs for data management/wrangling (dplyr)

rb	Meaning
ect()	Select variables (or columns)
r()	Subset observations (or rows)
tate()	Add new variables (or columns)
nmarise()	Reduce to a single row
up_by()	Aggregate
_join()	Merge two data objects
tinct()	Remove duplicate entries

Airline delays: flights to MSP in January, 2013

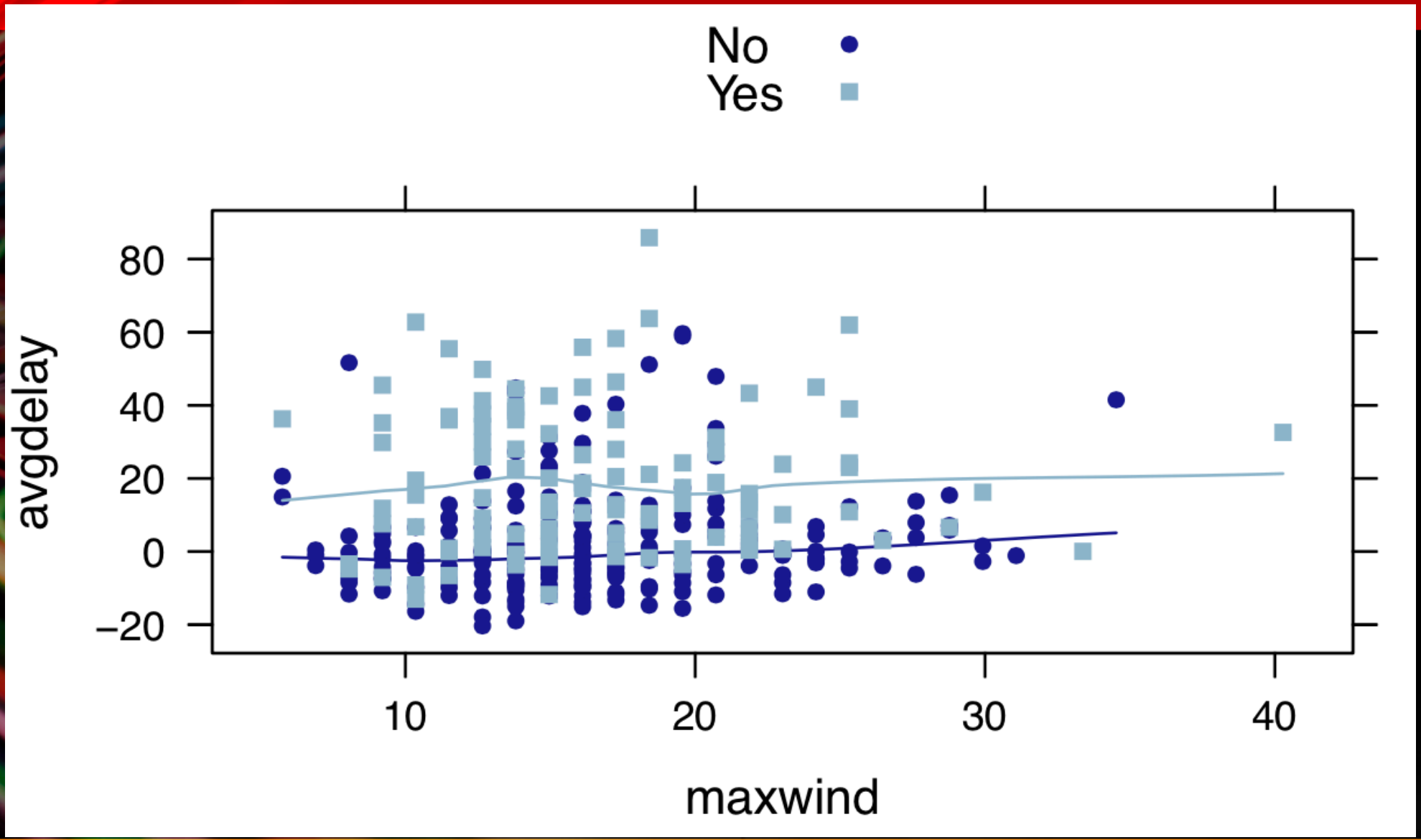


Airline delays: use of 5 idioms

```
delays <- flights %>%
  select(origin, dest, year, month, day, carrier, arr_delay) %>%
  filter(dest == 'MSP' & month == 1) %>%
  group_by(year, month, day, carrier) %>%
  summarise(meandelay = mean(arr_delay), count = n())
merged <- left_join(delays, airlines, by=c("carrier" = "carrier"
```

- See the data wrangling cheatsheet at: http://www.rstudio.com/resources/cheatsheets/

Airline delays: is it the wind, or the precipitation?



Airline delays: nycflights13

- Lots of other interesting questions
  - Tracing individual airplanes
  - Assessing the impact of weather events on flight delays
  - Cascading delays through hub airports
  - Time of day, day of week, and holiday effects
  - Smoothing and visualization important
  - Minimal need for p-values and models

To be relevant in the era of data science

- Need to be able to think creatively about data
- Need to be able to “tidy” data
- Need facility with data sets of varying sizes
- Need experience wrestling with large, messy, complex, challenging datasets
- Need an ethos of reproducibility
- Introduce small but powerful set of tools

Caveats

- Can be implemented in many systems (e.g. JMP, see Carver paper)
- Requires additional learning outcomes
- Requires faculty development
- Need more examples and sample activities

Key recommendations: if not now, when?

- Students need to be able to “think with data” (Lambert)
- They need multiple opportunities to analyze messy data using modern statistical practices
- Key theoretical concepts (design and confounding!) need to be integrated with theory, practice, and computation
- If not included as a part of our first and second courses in statistics, the vast majority of our students will not see these topics
- If we don’t teach these things, others will!

Teaching precursors to data science in introductory and second courses in statistics

Nicholas Horton, nhorton@amherst.edu

CAUSE Teaching and Learning webinar, February 24, 2015

Resources available at http://www.amherst.edu/~nhorton/precursors