

Teaching Precursors to Data Science in Introductory and Second Courses in Statistics

Nicholas Horton, nhorton@amherst.edu

Washington Statistical Society, April 28, 2015

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

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 many of courses (including first and second courses in statistics) often neglect this fact
- This talk 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
- Key foundational ideas:
 - Simulation and computation
 - Multivariate thinking
 - Reproducible analysis

Why are data-related and computational 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?

MAA Committee on Undergraduate Programs in Mathematics (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)

**American Statistical Association
Undergraduate Guidelines Workgroup**

Curriculum Guidelines for Undergraduate Programs in Statistical Science

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

- Carver and Stephens (International Conference on Teaching Statistics 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 (Journal of Statistical Software, 2014) “Tidy data”, <http://www.jstatsoft.org/v59/i10/paper>
- Horton, Baumer, and Wickham (CHANCE, 2015) “Setting the stage for data science”, <http://chance.amstat.org/2015/04/setting-the-stage>

Key idea #1: Why is computing important?

- Setting: Let A , B , and C be independent random variables each distributed uniform in the interval $[0,1]$.
- Question: What is the probability that the roots of the quadratic equation given by $Ax^2 + Bx + C = 0$ are real?
- Source, Rice *Mathematical Statistics and Data Analysis* third edition exercise 3.11 [also in first and second editions]

The analytic solution

The distribution of $Y = B^2$ is given by:

$$f(y) = \begin{cases} \frac{1}{2\sqrt{y}} & \text{if } 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

The distribution of $W = 4AC$ is given by:

$$f(w) = \begin{cases} -\log(w/4)/4 & \text{if } 0 \leq w \leq 4 \\ 0 & \text{otherwise} \end{cases}$$

Since Y and W are independent, the joint distribution is given by:

$$f(y, w) = \begin{cases} \frac{-\log(w/4)}{8\sqrt{y}} & \text{if } 0 \leq y \leq 1 \text{ and } 0 \leq w \leq 4 \\ 0 & \text{otherwise} \end{cases}$$

The analytic solution

The discriminant $B^2 - 4 * AC$ is non-negative when $Y > W$.

$$\begin{aligned} P(Y > W) &= \int_0^1 \int_0^y f(y, w) dw dy \\ &= \int_0^1 \int_0^y \frac{-\log(w/4)}{8\sqrt{y}} dw dy \\ &= \int_0^1 \frac{\sqrt{y}(-\log(y) + 1 + \log(4))}{8} dy \\ &= \frac{5 + \log(64)}{36} \approx 0.254413. \end{aligned}$$

Rice example

- Straightforward to simulate in R (noting that roots will be real only if the discriminant is non-negative):

```
numsim = 1000000
u1 = runif(numsim); u2 = runif(numsim); u3 = runif(numsim)

discrim = u2^2 - 4*u1*u3
realroot = discrim >= 0
table(realroot)/numsim

## realroot
##      FALSE      TRUE
## 0.74554 0.25446
```

Rice example

- Straightforward to simulate in R (noting that roots will be real only if the discriminant is non-negative):

```
numsim = 1000000
u1 = runif(numsim); u2 = runif(numsim); u3 = runif(numsim)

discrim = u2^2 - 4*u1*u3
realroot = discrim >= 0
table(realroot)/numsim

## realroot
##  FALSE    TRUE
## 0.74554 0.25446
```

Rice reports the correct answer as 1/9 (in first two editions and first printing of third!)

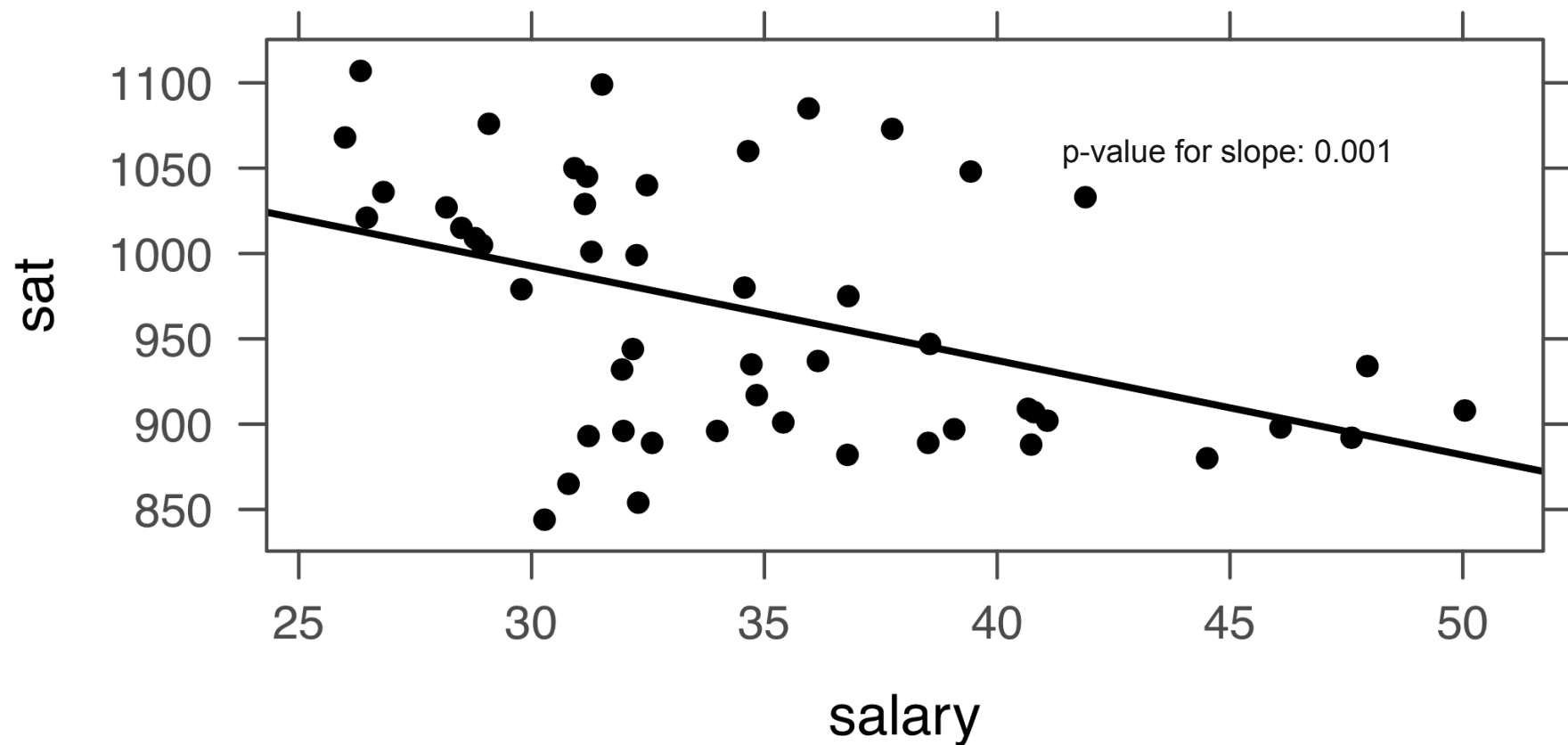
Why is computing important?

- Implication: it's hard to get probability problems wrong if you can check them in this manner
- Goal: develop parallel empirical and analytical problem-solving skills (Horton et al TAS, 2004; Horton TAS, 2013; Horton TAS, 2015)
- Still important to be able to get the correct answer (and not just an approximation)
- Helps develop the capacity to solve problems in a creative manner

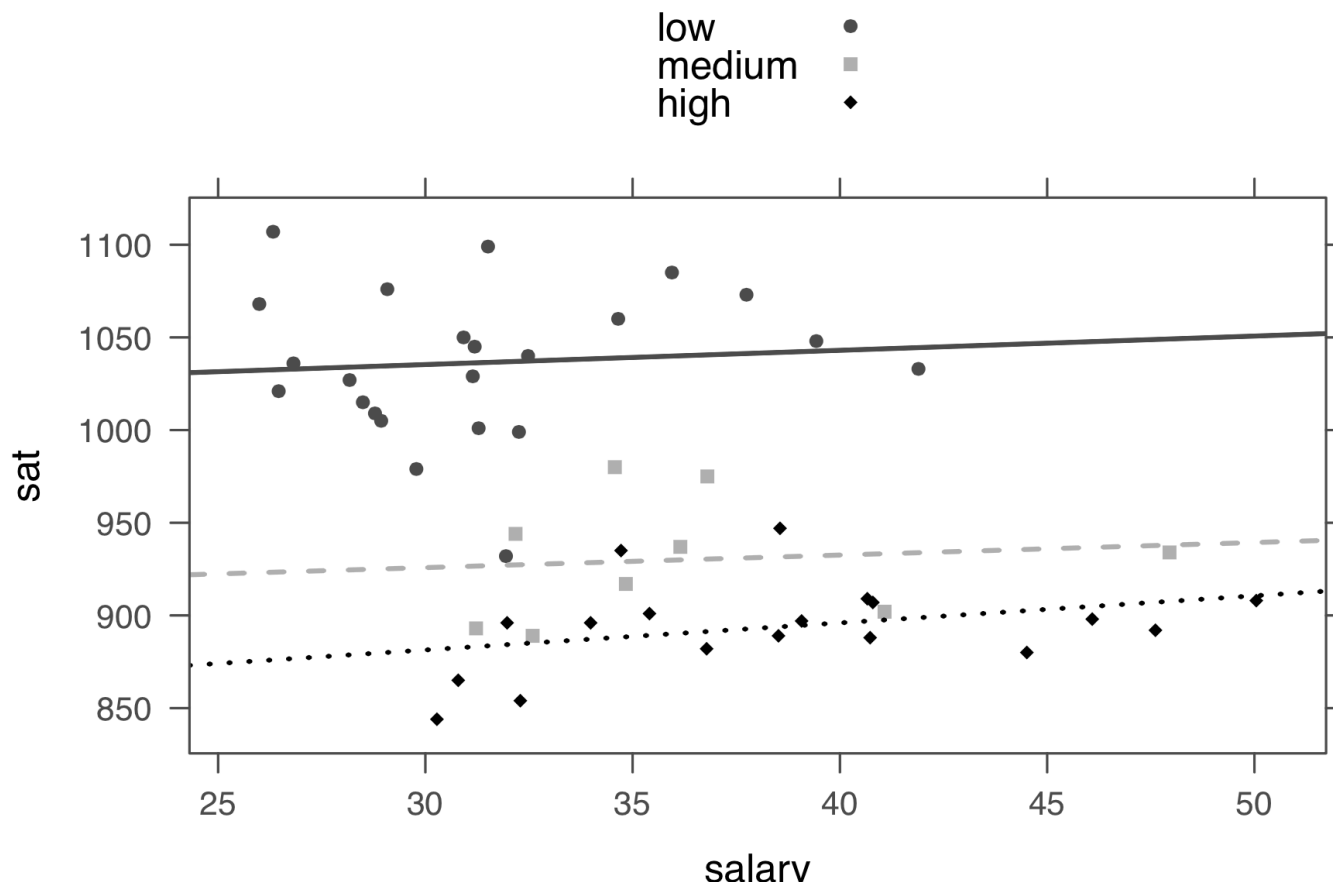
Key idea #2: multivariate thinking

- If data arise from well conducted randomized trial, we can make causal conclusions using two-sample t-test (The pinnacle of statistics? See Cobb “The Introductory Statistics Course: A Ptolemaic Curriculum?” in Technology Innovations in Stat Education 2007 for a dissenting view)
- If not (vast majority of cases) then you are stuck (according to the AP Stat and most intro stat syllabi)
- Net effect: students are paralyzed by what is likely their only stat course (see Meng’s musings and response from Amstat News, 2009)

Example: state average SAT scores and teacher salaries



Example: state average SAT scores and teacher salaries



Multivariate thinking

AP Statistics Vocabulary

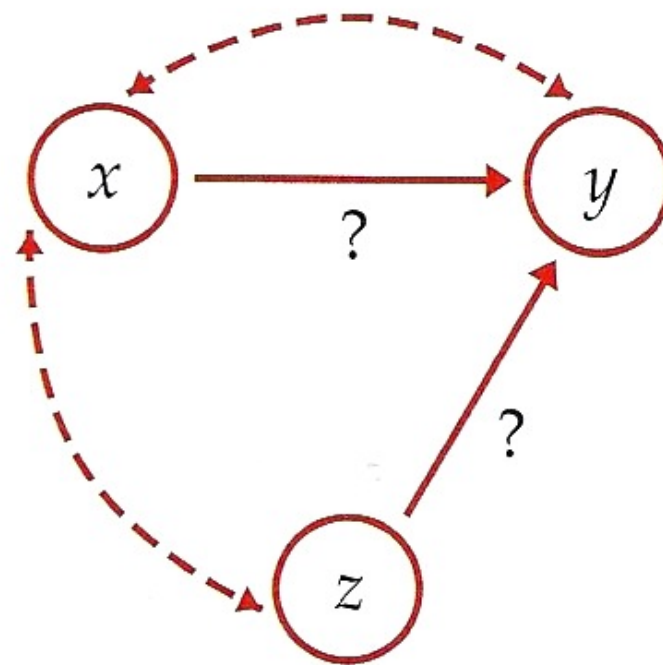


☒ Both Sides

confounding

when the levels of one factor are associated with the levels of another factor so their effects cannot be separated

Multivariate thinking



Confounding

Idea #3: Growing importance of data related skills

- George Lee of Goldman Sachs estimates that 90% of the world's data have been created in the last two years
- AAAS 2015 annual meeting theme: “Science and technology are being transformed by new ways to collect and use information. Progress in all fields is increasingly driven by the ability to organize, visualize, and analyze data.”

Idea #3: Growing importance of data related skills

- Working with data requires extensive computing skills far beyond those we have traditionally taught
- 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

Bad jokes about data scientists

- Question: What is a data scientist?

Bad jokes about data scientists

- Question: What is a data scientist?
A statistician that works in California.

Bad jokes about data scientists

- Question: What is a data scientist?
A statistician that works in California.
- Question: What is a data scientist?

Bad jokes about data scientists

- Question: What is a data scientist?
A statistician that works in California.
- Question: What is a data scientist?
A statistician that is useful.

Bad jokes about data scientists

- Question: What is a data scientist?
A statistician that works in California.
- Question: What is a data scientist?
A statistician that is useful.
- Question: What is a statistician?

Bad jokes about data scientists

- Question: What is a data scientist?
A statistician that works in California.
- Question: What is a data scientist?
A statistician that is useful.
- Question: What is a statistician?
A mathematician that can't program.

Threats

- ACM White Paper on Data Science
www.cra.org/cac/files/docs/init/bigdatawhitepaper.pdf
- The promise of data-driven decision-making is now being recognized broadly, and there is growing enthusiasm for the notion of ``Big Data." (first line)

Threats

- ACM White Paper on Data Science
www.cra.org/cac/files/docs/init/bigdatawhitepaper.pdf
- The promise of data-driven decision-making is now being recognized broadly, and there is growing enthusiasm for the notion of ``Big Data." (first line)
- Methods for querying and mining Big Data are fundamentally different from traditional statistical analysis on small samples. (first mention of statistics, page 7)

Growth of data science (separate from statistics)

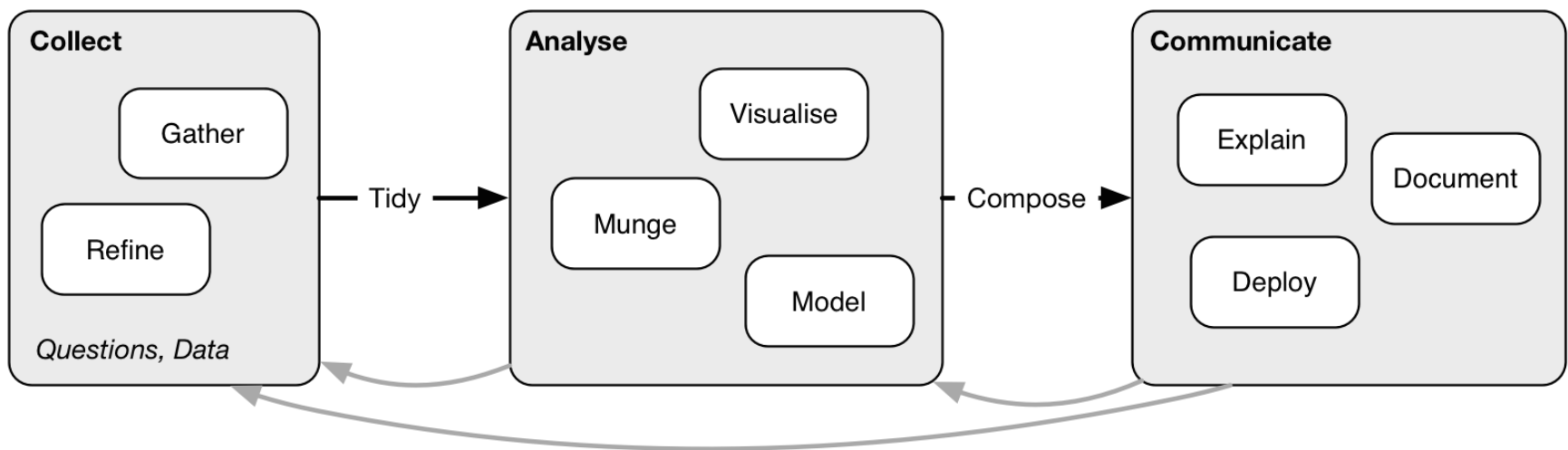
- Obama video intro to Big Data conference keynote:
<https://www.youtube.com/watch?v=vbb-AjiXyh0>
- “understanding and innovating with data has the potential to change how we do almost anything for the better”



President Barack Obama's Big Data Keynote -- Strata + Hadoop World 2015

A view of the data science process (Wickham)

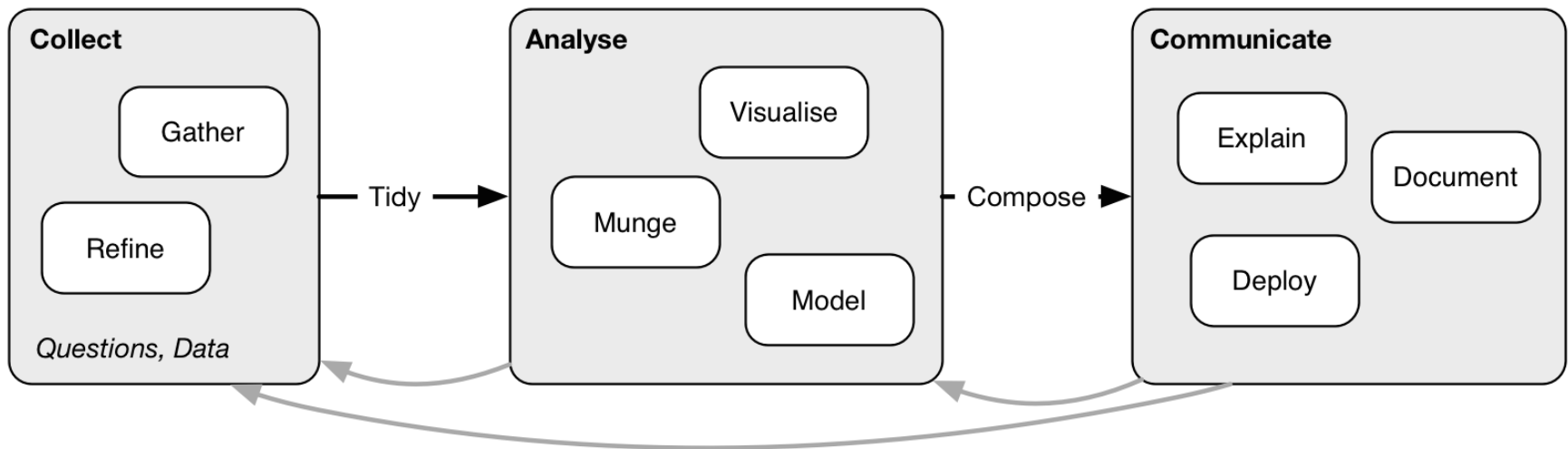
The data science process





A view of the statistical analysis process

The data science process



Key skills

- Effective statisticians at any level display an integrated combination of skills that are built upon statistical theory, statistical application, data management, 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

Important data-related topics

- 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

How to make this happen? 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: see resources)
 - nycflights13 package in R (n=336,776 flights)

Key verbs for data management/wrangling (dplyr)

- Minimal set of powerful idioms: “Less Volume, More Creativity”
- Select variables (or columns)
- Subset observations (or rows)
- Add new variables (or columns)
- Reduce to a single row (aggregate)
- Merge datasets

Key verbs for data management/wrangling (dplyr)

Verb	Meaning
<code>select()</code>	Select variables (or columns)
<code>filter()</code>	Subset observations (or rows)
<code>mutate()</code>	Add new variables (or columns)
<code>summarise()</code>	Reduce to a single row
<code>group_by()</code>	Aggregate
<code>left_join()</code>	Merge two data objects
<code>distinct()</code>	Remove duplicate entries

mosaic package vignette

Enough R for Intro Stats

Numerical Summaries

These functions have a formula interface to match plotting.

```
favstats()    # mosaic  
tally()       # mosaic  
mean()        # mosaic augmented  
median()      # mosaic augmented  
sd()          # mosaic augmented  
var()         # mosaic augmented
```

Randomization/Simulation

```
rflip()       # mosaic  
do()          # mosaic  
sample()      # mosaic augmented  
resample()    # with replacement  
shuffle()     # mosaic  
rbinom()  
rnorm()       # etc, if needed
```


Data Wrangling with dplyr and tidyr

Cheat Sheet



Syntax - Helpful conventions for wrangling

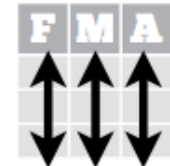
`dplyr::tbl_df(iris)`

Converts data to tbl class. tbl's are easier to examine than data frames. R displays only the data that fits onscreen:

```
Source: local data frame [150 x 5]
  Sepal.Length Sepal.Width Petal.Length
1           5.1           3.5           1.4
2           4.9           3.0           1.4
3           4.7           3.2           1.3
4           4.6           3.1           1.5
5           5.0           3.6           1.4
...           ...           ...
Variables not shown: Petal.Width (dbl),
Species (fctr)
```

Tidy

In a tidy
data set:

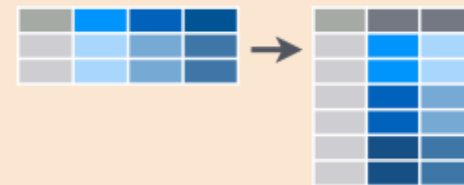


&

Each **variable** is saved
in its own **column**

E
S

Reshape



`tidyr::gather(cases, "year", "n", 2:4)`

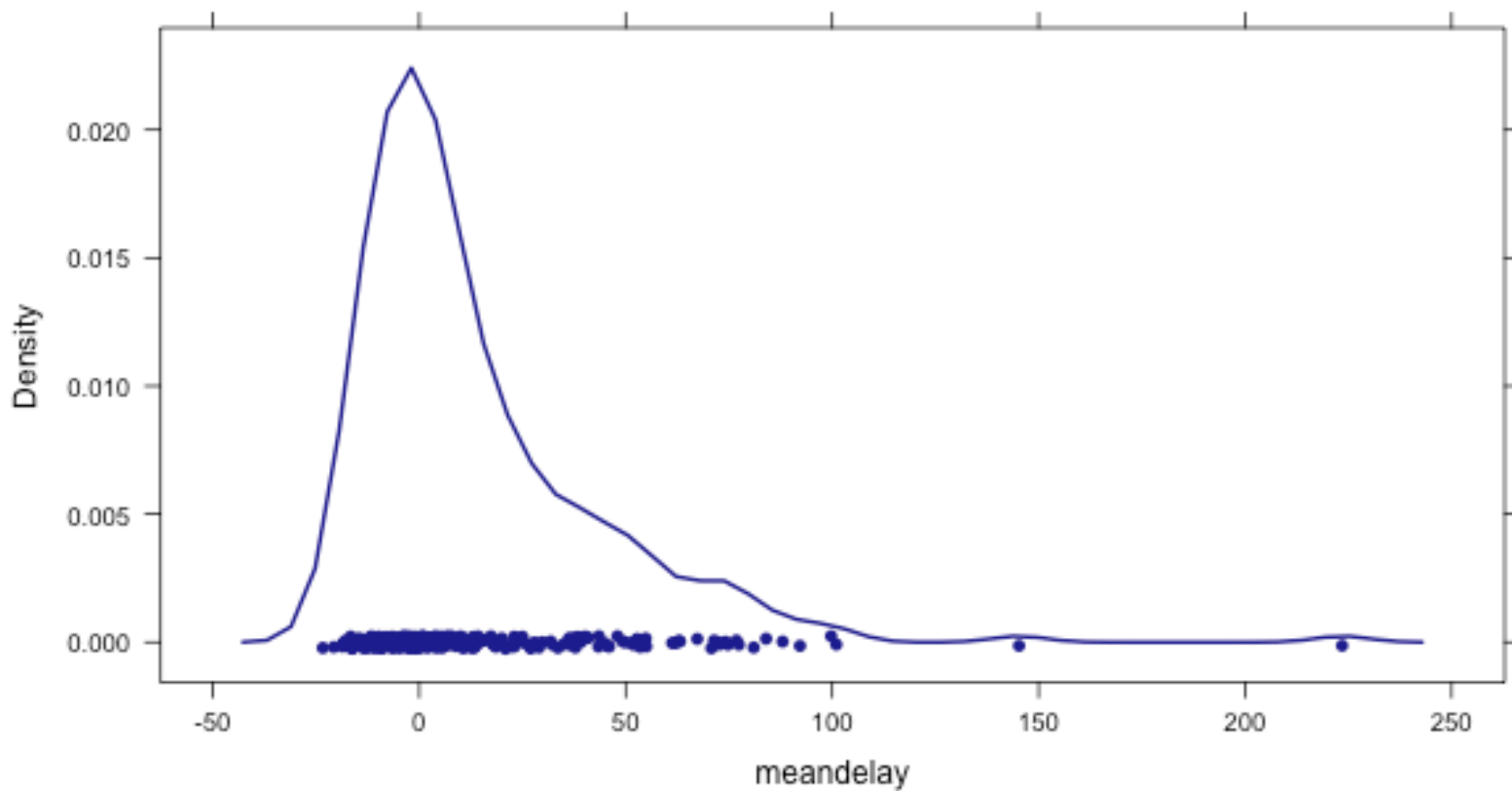
Gather columns into rows.



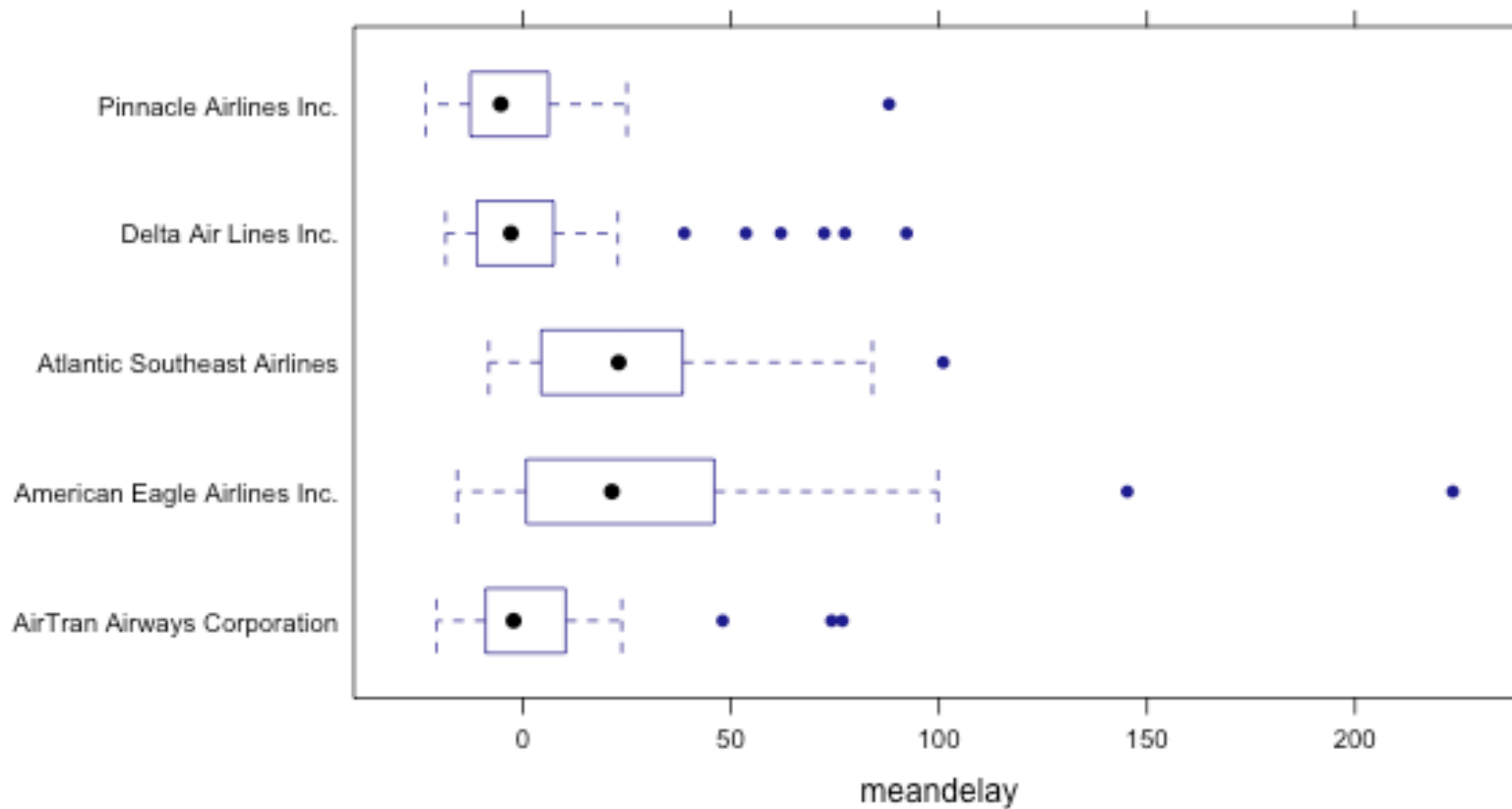
`tidyr::separate(storms, date, c("y", "m", "d"))`

Separate one column into several.

Airline delays: carrier average flight delays to DAY in January (2010-2014)



Airline delays: flights to DAY in January



Airline delays: use of 5 idioms

```
delays <- ontime %>%  
  select(Origin, Dest, Year, Month, DayOfMonth,  
         UniqueCarrier, ArrDelay) %>%  
  filter(Dest == 'DAY' & Month == 1 & Year > 2009) %>%  
  group_by(Year, Month, DayOfMonth, UniqueCarrier) %>%  
  summarise(meandelay = mean(ArrDelay), count = n())  
merged <- left_join(delays, airlines)
```

See the data wrangling cheatsheet at:

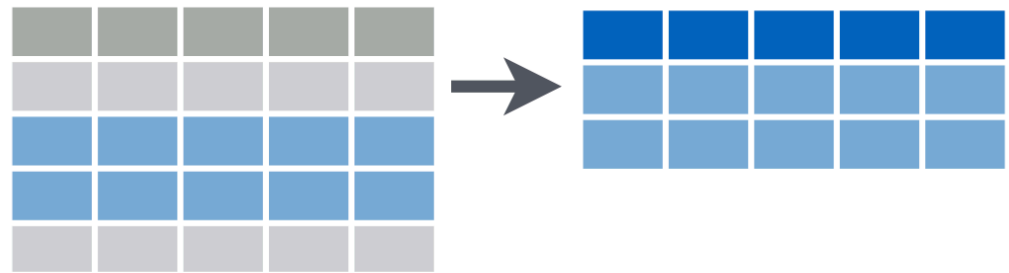
<http://www.rstudio.com/resources/cheatsheets/>

Airline delays: use of 5 idioms

```
delays <- ontime %>%  
  select(Origin, Dest, Year, Month, DayOfMonth,  
         UniqueCarrier, ArrDelay) %>%  
  filter(Dest == 'DAY' & Month == 1 & Year > 2009) %>%  
  group_by(Year, Month, DayOfMonth, UniqueCarrier) %>%  
  summarise(meandelay = mean(ArrDelay), count = n())  
merged <- left_join(delays, airlines)
```

Shiny app at: <http://rstudio.calvin.edu:3838/rpruim/dataOps/>

When do students learn these idioms?



```
dplyr::filter(iris, Sepal.Length > 7)
```

Extract rows that meet logical criteria.

<http://www.rstudio.com/resources/cheatsheets/>

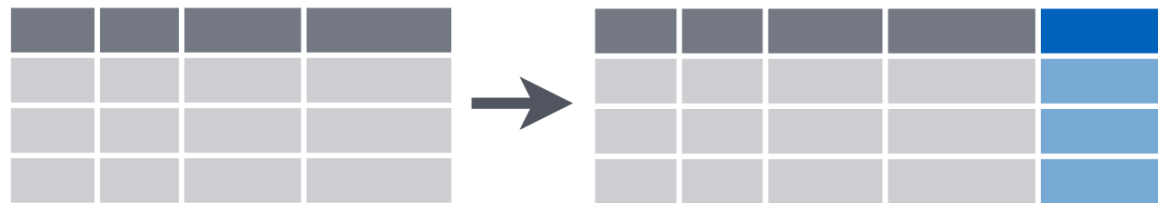
When do students learn these idioms?



```
dplyr::select(iris, Sepal.Width, Petal.Length, Species)
```

<http://www.rstudio.com/resources/cheatsheets/>

When do students learn these idioms?



```
dplyr::mutate(iris, sepal = Sepal.Length + Sepal.Width)
```

Compute and append one or more new columns.

<http://www.rstudio.com/resources/cheatsheets/>

When do students learn these idioms?

Group Data

dplyr::group_by(iris, Species)

Group data into rows with the same value of Species.

dplyr::ungroup(iris)

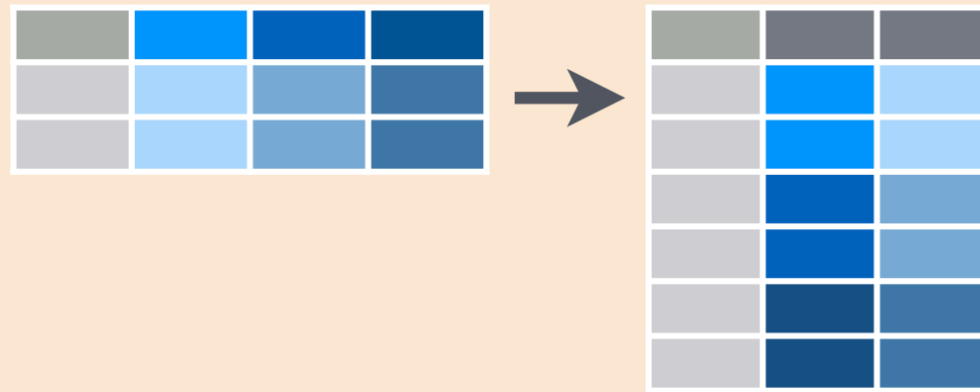
Remove grouping information from data frame.

iris %>% group_by(Species) %>% summarise(...)

Compute separate summary row for each group.



When do students learn these idioms?

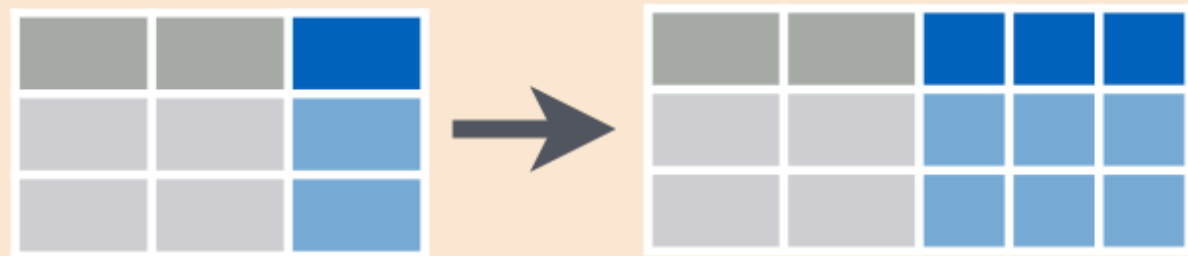


```
tidyr::gather(cases, "year", "n", 2:4)
```

Gather columns into rows.

<http://www.rstudio.com/resources/cheatsheets/>

When do students learn these idioms?



tidyr::separate(storms, date, c("y",
Separate one column into several.

When do students learn these idioms?

a		b	
x1	x2	x1	x3
A	1	A	T
B	2	B	F
C	3	D	T

+

=

Mutating Joins

x1	x2	x3
A	1	T
B	2	F
C	3	NA

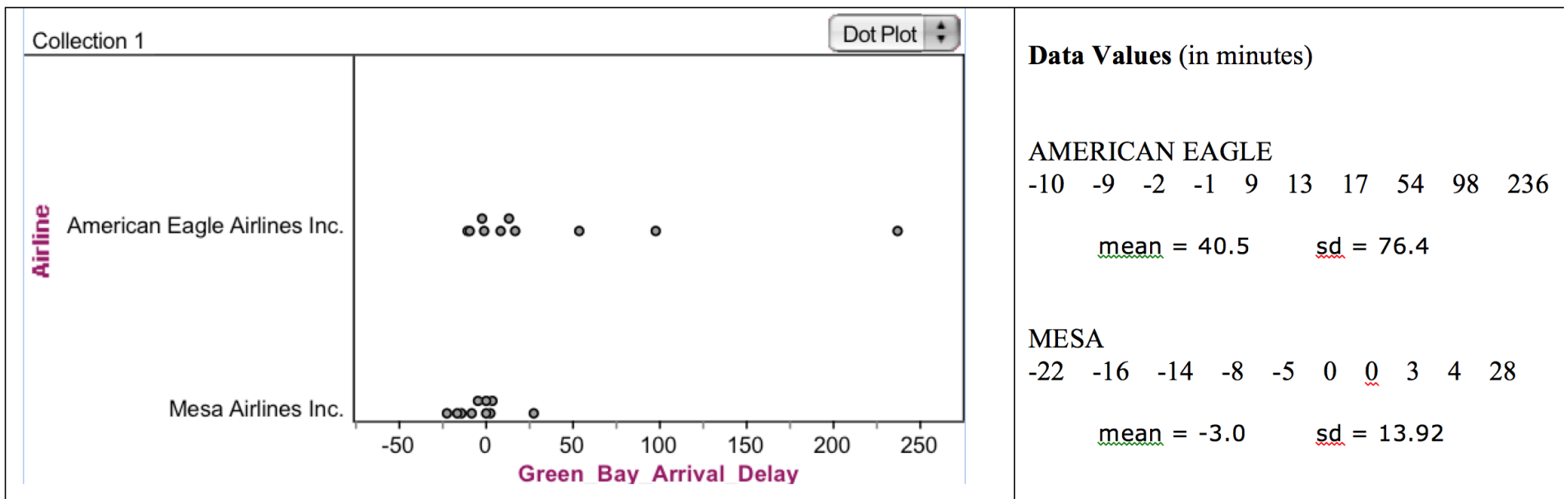
dplyr::left_join(a, b, by = "x1")

Join matching rows from b to a.

Airline delays: starting in intro stats

- Model Eliciting Activity: CATALST
<http://serc.carleton.edu/sp/library/mea/examples/example5.html> 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?
- If there are differences, are these differences consistent from city to city?

Airline delays: starting in intro stats



Airline delays MEA

- Pick five summary statistics to characterize the difference between the ontime performance of the airlines
- Interpret these (in the context of the problem)
- Use two or more of these statistics to come up with a rule to determine if one airline is more reliable than another
- Deliverable: short report (no jargon!) to the editor of *Chicago Magazine*
- Later in the semester: apply their rule to the full dataset

Airline delays: extensions

- 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!

Use of a database

- Fairly large dataset (100MB csv file per year compressed, 650MB per year uncompressed)
- SQLite database approximately 30GB in size (4 tables)
- Straightforward to access via R remotely or locally
- Facilitated for students using markdown (see resources)
- This is a skill that students need in the workforce
- Other multi-GB examples: NYC taxis, CitiBikes, Lahmann baseball, PGA every stroke, ...

R Markdown

- R Markdown is a format for writing reproducible, dynamic reports with R. Use it to embed R code and results into slideshows, pdfs, html documents, Word files and more
- See Baumer et al (TISE, 2014) for assessment of using this in intro stats (at Duke, Smith, and Amherst College)

<https://escholarship.org/uc/item/90b2f5xh>

R Markdown Cheat Sheet

learn more at rmarkdown.rstudio.com

rmarkdown 0.2.50 Updated: 8/14



embed R code

i. **Open** - Open a file that uses the .Rmd extension



2. Open File Start by saving a text file with the extension .Rmd, or open an RStudio Rmd template

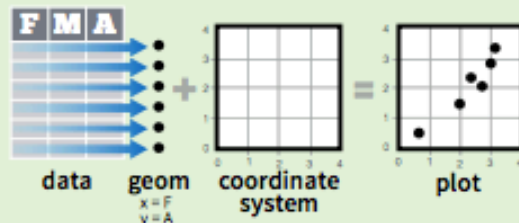
Data Visualization with ggplot2

Cheat Sheet



Basics

ggplot2 is based on the **grammar of graphics**, the idea that you can build every graph from the same few components: a **data** set, a set of **geoms**—visual marks that represent data points, and a **coordinate system**.

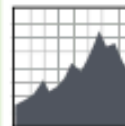


Geoms - Use a geom to represent data

One Variable

Continuous

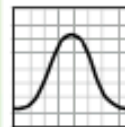
```
a <- ggplot(mpg, aes(hwy))
```



a + geom_area(stat = "bin")

x, y, alpha, color, fill, linetype, size

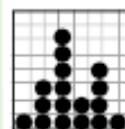
b + geom_area(aes(y = ..density..), stat = "bin")



a + geom_density(kernel = "gaussian")

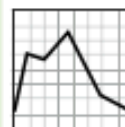
x, y, alpha, color, fill, linetype, size, weight

b + geom_density(aes(y = ..count..))



a + geom_dotplot()

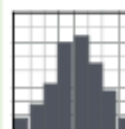
x, y, alpha, color, fill



a + geom_freqpoly()

x, y, alpha, color, linetype, size

b + geom_freqpoly(aes(y = ..density..))



a + geom_histogram(binwidth = 5)

x, y, alpha, color, fill, linetype, size, weight

b + geom_histogram(aes(v = ..density..))

Shiny Cheat Sheet

learn more at shiny.rstudio.com

Shiny 0.10.0 Updated: 6/14



2. server.R A set of instructions that build the R components of your app. To write server.R:

Some other thoughts about data science

- There is not a widely held definition of data science
- Data science, while not well defined, is different from statistics
- Data science seems to refer to a set of skills and techniques that include statistics, data mining, computer science (munging, coding, visualizing, etc.), domain expertise, and communication
- Data science is already more recognizable than statistics

To be relevant in the era of data science

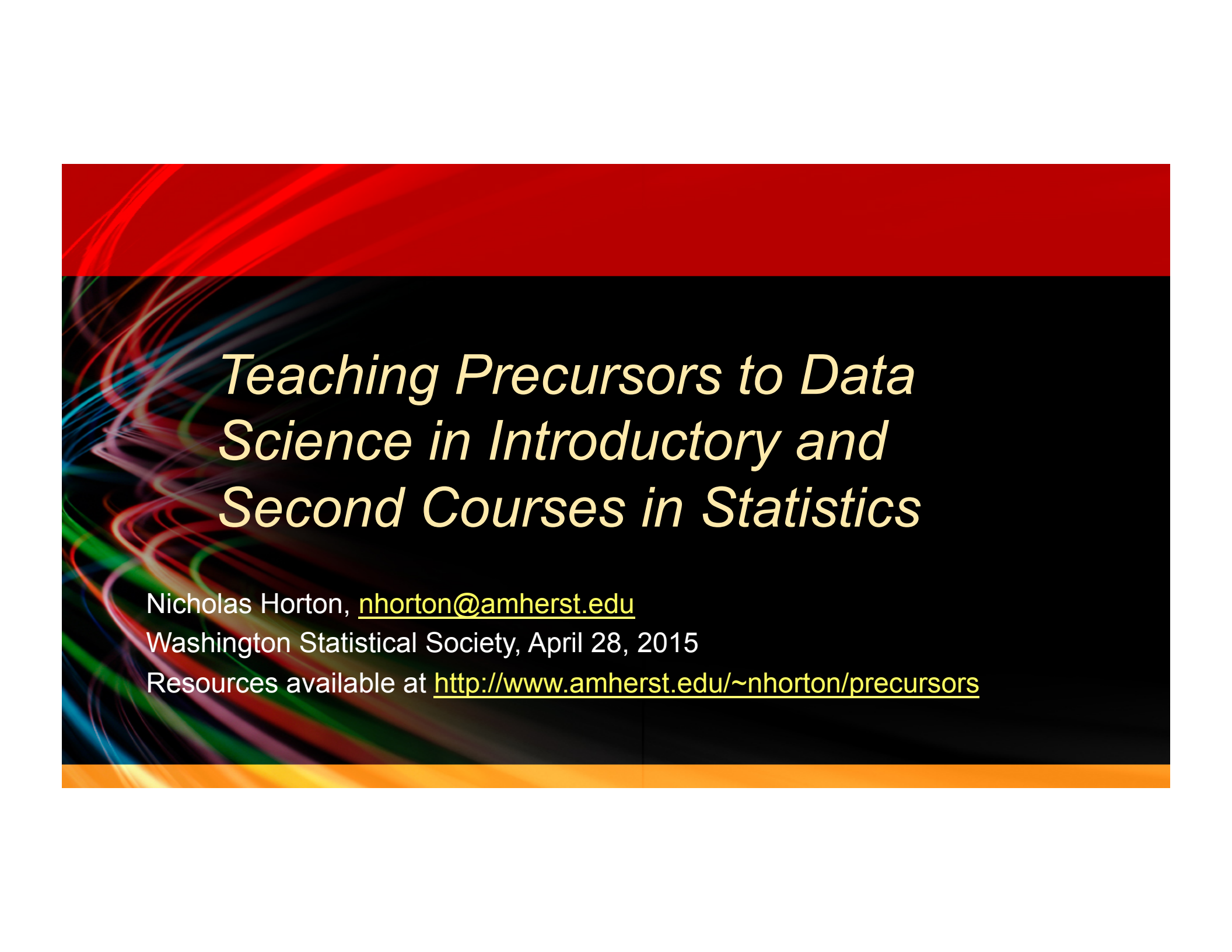
- Need to be able to think creatively about data
- Need to be able to “tidy” data
- Need facility with data sets of varying sizes
- Need experience wrestling with large, messy, complex, challenging datasets
- Need an ethos of reproducibility
- Need an understanding of ethics
- How to make it happen: introduce a small but powerful set of tools

Caveats

- This approach can be implemented in many systems (e.g. SAS or JMP, see Carver paper from ICOTS)
- I couldn't do this in R without R Markdown
- Requires additional learning outcomes...
- Requires faculty development...
- Need more examples and sample activities
- Need to determine what to do less of... (null hypothesis testing?)

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!
- Students with these skills will get jobs



Teaching Precursors to Data Science in Introductory and Second Courses in Statistics

Nicholas Horton, nhorton@amherst.edu

Washington Statistical Society, April 28, 2015

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