

Winter Institute in Data Science

Ryan T. Moore

2021-01-04

Goals

Skills

Examples

Welcome!

- ▶ Political methodologist in Dept of Government in SPA

- ▶ Political methodologist in Dept of Government in SPA
- ▶ Senior Social Scientist at The Lab @ DC

- ▶ Political methodologist in Dept of Government in SPA
- ▶ Senior Social Scientist at The Lab @ DC
- ▶ Methods Fellow at Office of Evaluation Sciences (US GSA)

Jeff Gill

Jeff Gill

- ▶ Distinguished Professor, Dept of Government and Dept of Math & Stats

Jeff Gill

- ▶ Distinguished Professor, Dept of Government and Dept of Math & Stats
- ▶ Inaugural Fellow of Society for Political Methodology

Jeff Gill

- ▶ Distinguished Professor, Dept of Government and Dept of Math & Stats
- ▶ Inaugural Fellow of Society for Political Methodology
- ▶ NSF, NIH, DOD, ...

Jeff Gill

- ▶ Distinguished Professor, Dept of Government and Dept of Math & Stats
- ▶ Inaugural Fellow of Society for Political Methodology
- ▶ NSF, NIH, DOD, ...
- ▶ Founding Director of AU's Center for Data Science

Data Science

Particular intersection of

- ▶ Statistical practice
- ▶ Computational tools
- ▶ Substantive knowledge

- ▶ Stats: prediction (vs. explanation), algorithms (vs. models)

- ▶ Stats: prediction (vs. explanation), algorithms (vs. models)
- ▶ Computing: addressing problems with data *per se* (size, tidy-ness, un/structure, replicability)

- ▶ Stats: prediction (vs. explanation), algorithms (vs. models)
- ▶ Computing: addressing problems with data *per se* (size, tidy-ness, un/structure, replicability)
- ▶ Substance: social science

Computers!

Data Scientist: The Sexiest Job of the 21st Century

by [Thomas H. Davenport](#) and [D.J. Patil](#)

From the October 2012 Issue

Social Meaning

BLS tracks data science now.

Social Meaning

BLS tracks data science now.

After CA, VA and MD have *next* most data scientists (states only)

Social Meaning

BLS tracks data science now.

After CA, VA and MD have *next* most data scientists (states only)

VA is 5th median salary (\$128,950, after NY, WA, NM, MA)

Social Meaning

BLS tracks data science now.

After CA, VA and MD have *next* most data scientists (states only)

VA is 5th median salary (\$128,950, after NY, WA, NM, MA)

...and stay tuned ...

Goals

Goals of the Course

- ▶ Utilize common computing tools for political data science – applied and scholarly

Goals of the Course

- ▶ Utilize common computing tools for political data science – applied and scholarly
- ▶ Visualize, transform, read, wrangle, tidy, analyze data

Goals of the Course

- ▶ Utilize common computing tools for political data science – applied and scholarly
- ▶ Visualize, transform, read, wrangle, tidy, analyze data
- ▶ Refresh mathematical foundations for modeling

Goals of the Course

- ▶ Utilize common computing tools for political data science – applied and scholarly
- ▶ Visualize, transform, read, wrangle, tidy, analyze data
- ▶ Refresh mathematical foundations for modeling
- ▶ Learn modern scientific communication tools

Goals of the Course

- ▶ Utilize common computing tools for political data science – applied and scholarly
- ▶ Visualize, transform, read, wrangle, tidy, analyze data
- ▶ Refresh mathematical foundations for modeling
- ▶ Learn modern scientific communication tools
- ▶ Learn modern version control

Goals of the Course

- ▶ Utilize common computing tools for political data science – applied and scholarly
- ▶ Visualize, transform, read, wrangle, tidy, analyze data
- ▶ Refresh mathematical foundations for modeling
- ▶ Learn modern scientific communication tools
- ▶ Learn modern version control
- ▶ Gain exposure to machine learning and other modern statistical data science methods and computing tools

Goals of the Course

- ▶ Utilize common computing tools for political data science – applied and scholarly
- ▶ Visualize, transform, read, wrangle, tidy, analyze data
- ▶ Refresh mathematical foundations for modeling
- ▶ Learn modern scientific communication tools
- ▶ Learn modern version control
- ▶ Gain exposure to machine learning and other modern statistical data science methods and computing tools
- ▶ Do original research using data sci methods.
Contribute methods, substance, both.

Skills

► Data analysis

- ▶ Data analysis
 - ▶ R, Python, shell

- ▶ Data analysis
 - ▶ R, Python, shell
- ▶ Workflow and communication

- ▶ Data analysis
 - ▶ R, Python, shell
- ▶ Workflow and communication
 - ▶ `git`, GitHub, Docker

- ▶ Data analysis
 - ▶ R, Python, shell
- ▶ Workflow and communication
 - ▶ `git`, GitHub, Docker
 - ▶ RMarkdown

- ▶ Data analysis
 - ▶ R, Python, shell
- ▶ Workflow and communication
 - ▶ `git`, GitHub, Docker
 - ▶ RMarkdown
 - ▶ “projects”

- ▶ Data analysis
 - ▶ R, Python, shell
- ▶ Workflow and communication
 - ▶ `git`, GitHub, Docker
 - ▶ RMarkdown
 - ▶ “projects”
 - ▶ programming practices

- ▶ Data analysis
 - ▶ R, Python, shell
- ▶ Workflow and communication
 - ▶ `git`, GitHub, Docker
 - ▶ RMarkdown
 - ▶ “projects”
 - ▶ programming practices
 - ▶ visualization

- ▶ Data analysis
 - ▶ R, Python, shell
- ▶ Workflow and communication
 - ▶ `git`, GitHub, Docker
 - ▶ RMarkdown
 - ▶ “projects”
 - ▶ programming practices
 - ▶ visualization
 - ▶ cloud and distributed computing

- ▶ Data analysis
 - ▶ R, Python, shell
- ▶ Workflow and communication
 - ▶ `git`, GitHub, Docker
 - ▶ RMarkdown
 - ▶ “projects”
 - ▶ programming practices
 - ▶ visualization
 - ▶ cloud and distributed computing
- ▶ Fundamental statistics

- ▶ Data analysis
 - ▶ R, Python, shell
- ▶ Workflow and communication
 - ▶ git, GitHub, Docker
 - ▶ RMarkdown
 - ▶ “projects”
 - ▶ programming practices
 - ▶ visualization
 - ▶ cloud and distributed computing
- ▶ Fundamental statistics
 - ▶ descriptive

- ▶ Data analysis
 - ▶ R, Python, shell
- ▶ Workflow and communication
 - ▶ git, GitHub, Docker
 - ▶ RMarkdown
 - ▶ “projects”
 - ▶ programming practices
 - ▶ visualization
 - ▶ cloud and distributed computing
- ▶ Fundamental statistics
 - ▶ descriptive
 - ▶ modeling

- ▶ Data analysis
 - ▶ R, Python, shell
- ▶ Workflow and communication
 - ▶ git, GitHub, Docker
 - ▶ RMarkdown
 - ▶ “projects”
 - ▶ programming practices
 - ▶ visualization
 - ▶ cloud and distributed computing
- ▶ Fundamental statistics
 - ▶ descriptive
 - ▶ modeling
 - ▶ inference

- ▶ Data analysis
 - ▶ R, Python, shell
- ▶ Workflow and communication
 - ▶ git, GitHub, Docker
 - ▶ RMarkdown
 - ▶ “projects”
 - ▶ programming practices
 - ▶ visualization
 - ▶ cloud and distributed computing
- ▶ Fundamental statistics
 - ▶ descriptive
 - ▶ modeling
 - ▶ inference
- ▶ Modern statistical computational topics

- ▶ Data analysis
 - ▶ R, Python, shell
- ▶ Workflow and communication
 - ▶ git, GitHub, Docker
 - ▶ RMarkdown
 - ▶ “projects”
 - ▶ programming practices
 - ▶ visualization
 - ▶ cloud and distributed computing
- ▶ Fundamental statistics
 - ▶ descriptive
 - ▶ modeling
 - ▶ inference
- ▶ Modern statistical computational topics
 - ▶ network analysis

- ▶ Data analysis
 - ▶ R, Python, shell
- ▶ Workflow and communication
 - ▶ git, GitHub, Docker
 - ▶ RMarkdown
 - ▶ “projects”
 - ▶ programming practices
 - ▶ visualization
 - ▶ cloud and distributed computing
- ▶ Fundamental statistics
 - ▶ descriptive
 - ▶ modeling
 - ▶ inference
- ▶ Modern statistical computational topics
 - ▶ network analysis
 - ▶ machine learning

- ▶ Data analysis
 - ▶ R, Python, shell
- ▶ Workflow and communication
 - ▶ git, GitHub, Docker
 - ▶ RMarkdown
 - ▶ “projects”
 - ▶ programming practices
 - ▶ visualization
 - ▶ cloud and distributed computing
- ▶ Fundamental statistics
 - ▶ descriptive
 - ▶ modeling
 - ▶ inference
- ▶ Modern statistical computational topics
 - ▶ network analysis
 - ▶ machine learning
 - ▶ clustering

- ▶ Data analysis
 - ▶ R, Python, shell
- ▶ Workflow and communication
 - ▶ git, GitHub, Docker
 - ▶ RMarkdown
 - ▶ “projects”
 - ▶ programming practices
 - ▶ visualization
 - ▶ cloud and distributed computing
- ▶ Fundamental statistics
 - ▶ descriptive
 - ▶ modeling
 - ▶ inference
- ▶ Modern statistical computational topics
 - ▶ network analysis
 - ▶ machine learning
 - ▶ clustering
 - ▶ neural nets

- ▶ Data analysis
 - ▶ R, Python, shell
- ▶ Workflow and communication
 - ▶ git, GitHub, Docker
 - ▶ RMarkdown
 - ▶ “projects”
 - ▶ programming practices
 - ▶ visualization
 - ▶ cloud and distributed computing
- ▶ Fundamental statistics
 - ▶ descriptive
 - ▶ modeling
 - ▶ inference
- ▶ Modern statistical computational topics
 - ▶ network analysis
 - ▶ machine learning
 - ▶ clustering
 - ▶ neural nets
 - ▶ text as data

Examples

What is a data science task?

“Keep only non-voters who might be subject to interference”

What is a data science task?

“Keep only non-voters who might be subject to interference”

```
social <- read_csv("http://j.mp/2Et71U0")  
filter(social, (hhsiz > 1) & (primary2004 == 0))
```

What is a data science task?

“Keep only non-voters who might be subject to interference”

```
social <- read_csv("http://j.mp/2Et71U0")
filter(social, (hysize > 1) & (primary2004 == 0))
```

```
social <- read_csv("http://j.mp/2Et71U0")
filter(social, (hysize > 1) & (primary2004 == 0))
```

A tibble: 161,275 x 6

##	sex	yearofbirth	primary2004	messages	primary2004
##	<chr>	<dbl>	<dbl>	<chr>	<dbl>
##	1 male	1941	0	Civic Duty	0
##	2 female	1947	0	Civic Duty	0
##	3 male	1951	0	Hawthorne	1
##	4 female	1950	0	Hawthorne	1
##	5 female	1982	0	Hawthorne	1
##	6 male	1981	0	Control	0
##	7 female	1959	0	Control	1
##	8 male	1956	0	Control	1

What is a data science task?

“I need to read these dates from Spanish \rightsquigarrow standard format”

What is a data science task?

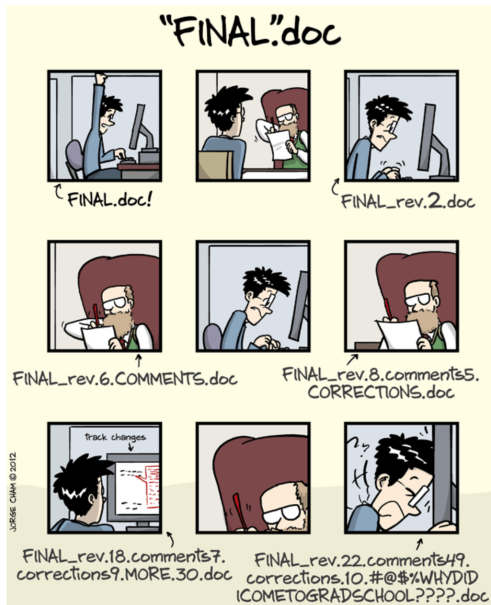
“I need to read these dates from Spanish \rightsquigarrow standard format”

```
parse_date("15 enero 2000",  
           locale = locale("es"),  
           format = "%d %B %Y")
```

```
## [1] "2000-01-15"
```


What is a data science task?

I collaborate, but FinalLAST draft.v.2.doc (1) is painful.



What is a data science task?

I need to collaborate, but FinalFinalLAST draft.v.2.doc
(1) isn't working for me anymore.

```
git add paper.tex  
git commit paper.tex  
git push
```

What is a data science question?

- ▶ Can we predict which registrants are most likely to reply to which email appeals?

What is a data science question?

- ▶ Can we predict which registrants are most likely to reply to which email appeals?
- ▶ What characteristics of rodent complaints actually lead to successful abatement?

What is a data science question?

- ▶ Can we predict which registrants are most likely to reply to which email appeals?
- ▶ What characteristics of rodent complaints actually lead to successful abatement?
- ▶ How can we fairly estimate probability defendant will appear?

What is a data science question?

- ▶ Can we predict which registrants are most likely to reply to which email appeals?
- ▶ What characteristics of rodent complaints actually lead to successful abatement?
- ▶ How can we fairly estimate probability defendant will appear?
- ▶ Are intersections with new patterns less prone to traffic accidents?

Course GitHub page:

<https://github.com/ryantmoore/winter-inst-2021>

(syllabus tour)

Installations

- ▶ R:
<https://cran.r-project.org>
- ▶ RStudio (Desktop):
<https://rstudio.com/products/rstudio/>
- ▶ Anaconda:
<https://www.anaconda.com/distribution/>