Winter Institute in Data Science

Ryan T. Moore

2025-12-15

Intros: Ryan T. Moore & Jeff Gill

Goals

Skills

Examples

Installations and Plan

Welcome!

Intros: Ryan T. Moore & Jeff Gill

- Political methodologist
  (Dept of Government, SPA)

- Political methodologist
  (Dept of Government, SPA)
- Assoc Director, Center for Data Science
  (AU, SPA)

- Political methodologist
  (Dept of Government, SPA)
- Assoc Director, Center for Data Science
  (AU, SPA)
- Senior Social Scientist
  (The Lab @ DC)

- Political methodologist
  (Dept of Government, SPA)
- Assoc Director, Center for Data Science
  (AU, SPA)
- Senior Social Scientist
  (The Lab @ DC)
- (Recently...) Methods Fellow
  (Office of Evaluation Sciences, US GSA)

Jeff Gill, Director, Center for Data Science

Jeff Gill, Director, Center for Data Science

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

Jeff Gill, Director, Center for Data Science

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

Jeff Gill, Director, Center for Data Science

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

Jeff Gill, Director, Center for Data Science

- Distinguished Professor, Dept of Government and Dept of Math & Stats
- Inaugural Fellow of Society for Political Methodology
- NSF, NIH, DOD, ...
- Gosnell Prize for best work in Political Methodology

Jeff Gill, Director, Center for Data Science

- Distinguished Professor, Dept of Government and Dept of Math & Stats
- Inaugural Fellow of Society for Political Methodology
- NSF, NIH, DOD, ...
- Gosnell Prize for best work in Political Methodology
- Career Achievement *and* Excellence in Mentoring Awards (2024, SPM)

Jeff Gill, Director, Center for Data Science

- Distinguished Professor, Dept of Government and Dept of Math & Stats
- Inaugural Fellow of Society for Political Methodology
- NSF, NIH, DOD, ...
- Gosnell Prize for best work in Political Methodology
- Career Achievement *and* Excellence in Mentoring Awards (2024, SPM)
- 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

Harvard Business Review

Data | Data Scientist: The Sexiest Job of the 21st Century

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.

Highest concentrations: DC, CO, NY, UT, NC!

Social Meaning

BLS tracks data science now.

Highest concentrations: DC, CO, NY, UT, NC!

. . . and salaries are high . . .

Social Meaning

BLS tracks data science now.

Highest concentrations: DC, CO, NY, UT, NC!

. . . and salaries are high . . .

Applied quantitative social science increasingly looks like *data science*

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, AWS, Docker, Kubernetes, Code Ocean

- Data analysis
  - R, Python, shell
- Workflow and communication
  - `git`, GitHub, AWS, Docker, Kubernetes, Code Ocean
  - Quarto (/RMarkdown)

- Data analysis
  - R, Python, shell
- Workflow and communication
  - `git`, GitHub, AWS, Docker, Kubernetes, Code Ocean
  - Quarto (/RMarkdown)
  - "projects"

- Data analysis
  - R, Python, shell
- Workflow and communication
  - `git`, GitHub, AWS, Docker, Kubernetes, Code Ocean
  - Quarto (/RMarkdown)
  - "projects"
  - programming practices

- Data analysis
  - R, Python, shell
- Workflow and communication
  - `git`, GitHub, AWS, Docker, Kubernetes, Code Ocean
  - Quarto (/RMarkdown)
  - "projects"
  - programming practices
  - visualization

- Data analysis
  - R, Python, shell
- Workflow and communication
  - `git`, GitHub, AWS, Docker, Kubernetes, Code Ocean
  - Quarto (/RMarkdown)
  - "projects"
  - programming practices
  - visualization
  - cloud and distributed computing

- Data analysis
  - R, Python, shell
- Workflow and communication
  - `git`, GitHub, AWS, Docker, Kubernetes, Code Ocean
  - Quarto (/RMarkdown)
  - "projects"
  - programming practices
  - visualization
  - cloud and distributed computing
- Fundamental statistics

- Data analysis
  - R, Python, shell
- Workflow and communication
  - `git`, GitHub, AWS, Docker, Kubernetes, Code Ocean
  - Quarto (/RMarkdown)
  - "projects"
  - programming practices
  - visualization
  - cloud and distributed computing
- Fundamental statistics
  - descriptive

- Data analysis
  - R, Python, shell
- Workflow and communication
  - `git`, GitHub, AWS, Docker, Kubernetes, Code Ocean
  - Quarto (/RMarkdown)
  - "projects"
  - programming practices
  - visualization
  - cloud and distributed computing
- Fundamental statistics
  - descriptive
  - modeling (linear, GLM, Bayes, and beyond)

- Data analysis
  - R, Python, shell
- Workflow and communication
  - `git`, GitHub, AWS, Docker, Kubernetes, Code Ocean
  - Quarto (/RMarkdown)
  - "projects"
  - programming practices
  - visualization
  - cloud and distributed computing
- Fundamental statistics
  - descriptive
  - modeling (linear, GLM, Bayes, and beyond)
  - inference

- Data analysis
  - R, Python, shell
- Workflow and communication
  - `git`, GitHub, AWS, Docker, Kubernetes, Code Ocean
  - Quarto (/RMarkdown)
  - "projects"
  - programming practices
  - visualization
  - cloud and distributed computing
- Fundamental statistics
  - descriptive
  - modeling (linear, GLM, Bayes, and beyond)
  - inference
- Modern statistical computational topics

- Data analysis
  - R, Python, shell
- Workflow and communication
  - `git`, GitHub, AWS, Docker, Kubernetes, Code Ocean
  - Quarto (/RMarkdown)
  - "projects"
  - programming practices
  - visualization
  - cloud and distributed computing
- Fundamental statistics
  - descriptive
  - modeling (linear, GLM, Bayes, and beyond)
  - inference
- Modern statistical computational topics
  - network analysis

- Data analysis
  - R, Python, shell
- Workflow and communication
  - `git`, GitHub, AWS, Docker, Kubernetes, Code Ocean
  - Quarto (/RMarkdown)
  - "projects"
  - programming practices
  - visualization
  - cloud and distributed computing
- Fundamental statistics
  - descriptive
  - modeling (linear, GLM, Bayes, and beyond)
  - inference
- Modern statistical computational topics
  - network analysis
  - machine learning

- Data analysis
  - R, Python, shell
- Workflow and communication
  - `git`, GitHub, AWS, Docker, Kubernetes, Code Ocean
  - Quarto (/RMarkdown)
  - "projects"
  - programming practices
  - visualization
  - cloud and distributed computing
- Fundamental statistics
  - descriptive
  - modeling (linear, GLM, Bayes, and beyond)
  - inference
- Modern statistical computational topics
  - network analysis
  - machine learning
  - clustering

- Data analysis
  - R, Python, shell
- Workflow and communication
  - `git`, GitHub, AWS, Docker, Kubernetes, Code Ocean
  - Quarto (/RMarkdown)
  - "projects"
  - programming practices
  - visualization
  - cloud and distributed computing
- Fundamental statistics
  - descriptive
  - modeling (linear, GLM, Bayes, and beyond)
  - inference
- Modern statistical computational topics
  - network analysis
  - machine learning
  - clustering
  - neural nets

- Data analysis
  - R, Python, shell
- Workflow and communication
  - `git`, GitHub, AWS, Docker, Kubernetes, Code Ocean
  - Quarto (/RMarkdown)
  - "projects"
  - programming practices
  - visualization
  - cloud and distributed computing
- Fundamental statistics
  - descriptive
  - modeling (linear, GLM, Bayes, and beyond)
  - inference
- Modern statistical computational topics
  - network analysis
  - machine learning
  - clustering
  - neural nets
  - text as data, NLP

- Data analysis
  - R, Python, shell
- Workflow and communication
  - `git`, GitHub, AWS, Docker, Kubernetes, Code Ocean
  - Quarto (/RMarkdown)
  - "projects"
  - programming practices
  - visualization
  - cloud and distributed computing
- Fundamental statistics
  - descriptive
  - modeling (linear, GLM, Bayes, and beyond)
  - inference
- Modern statistical computational topics
  - network analysis
  - machine learning
  - clustering
  - neural nets
  - text as data, NLP
  - modeling, Bayes, AI

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, (hhsize > 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, (hhsize > 1) & (primary2004 == 0))
```

```
## # A tibble: 161,275 x 6
##    sex    yearofbirth primary2004 messages   primary200
##    <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
##  9 female        1968           0 Control             0
## 10 male          1967           0 Control             0
## # ... with 161,265 more rows
```

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")
```

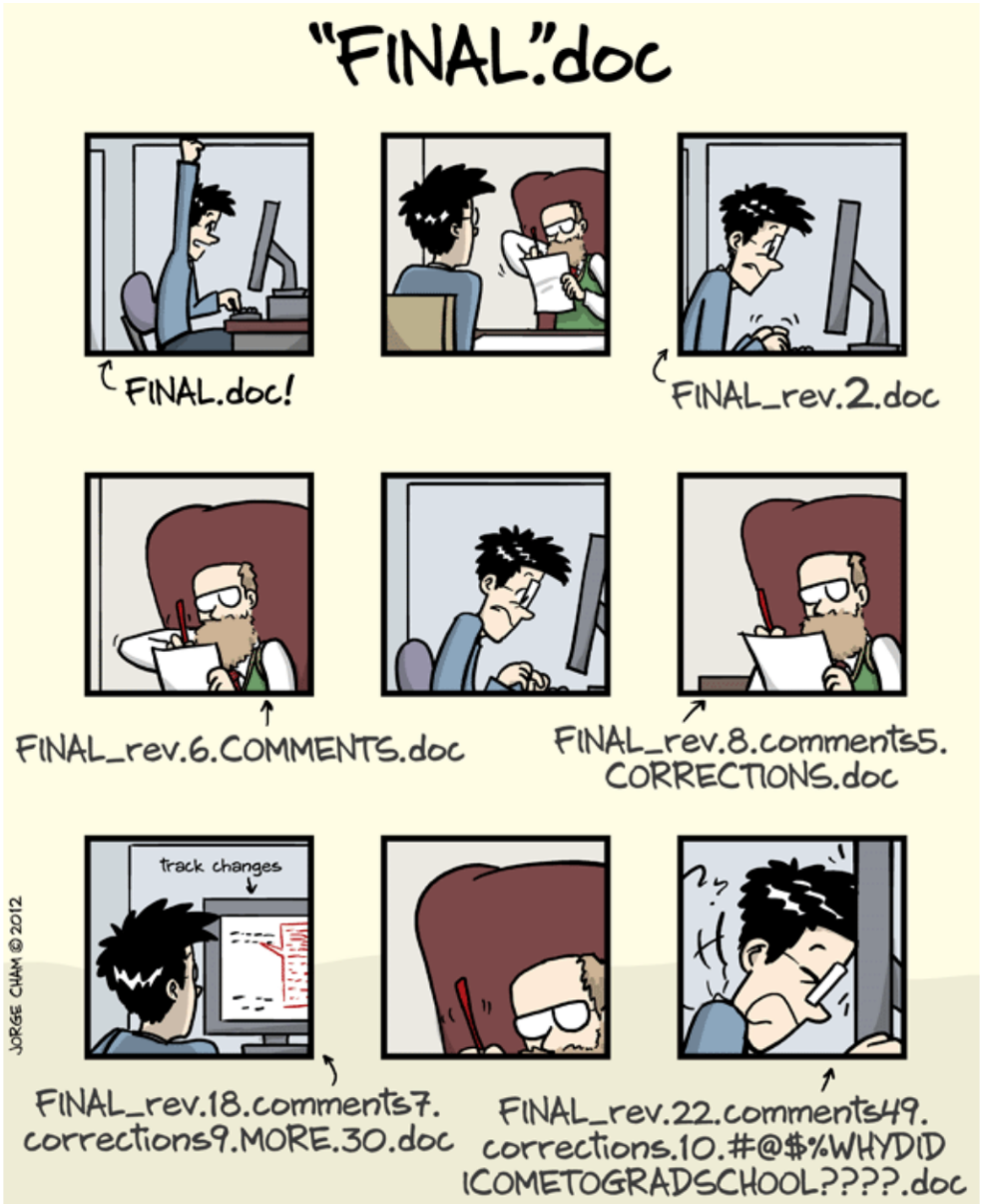
```
## [1] "2000-01-15"
```

What is a data science task?

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

What is a data science task?

I 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.

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 crashes?

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 crashes?

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 crashes?

- How do we compare models/prediction strategies?

Course GitHub page:

https://github.com/ryantmoore/winter-inst-2026

(syllabus tour)

Installations and Plan

Plan

Today

- Now → 10:00: Installations
- 10:00-12:00: R/tidyverse
- 13:00-14:00: LaTeX, Quarto, RMarkdown
- 14:30-16:00: Data wrangling, EDA
- 16:30-17:00: Final projects

Plan

Today

- Now → 10:00: Installations
- 10:00-12:00: R/tidyverse
- 13:00-14:00: LaTeX, Quarto, RMarkdown
- 14:30-16:00: Data wrangling, EDA
- 16:30-17:00: Final projects

Tomorrow: Math, stats, programming practices

Plan

Today

- Now → 10:00: Installations
- 10:00-12:00: R/tidyverse
- 13:00-14:00: LaTeX, Quarto, RMarkdown
- 14:30-16:00: Data wrangling, EDA
- 16:30-17:00: Final projects

Tomorrow: Math, stats, programming practices

January 5-9: Guest speakers + RTM on variety of data science methods

Plan

Today

- Now → 10:00: Installations
- 10:00-12:00: R/tidyverse
- 13:00-14:00: LaTeX, Quarto, RMarkdown
- 14:30-16:00: Data wrangling, EDA
- 16:30-17:00: Final projects

Tomorrow: Math, stats, programming practices

January 5-9: Guest speakers + RTM on variety of data science methods

January 10: Final presentations, team work, project completion 15:00

Assignment	Weight	Due date
Final presentation	20%	10 January
Final project	50%	10 January
Final peer evaluation	10%	10 January
Participation	10%	daily
Attendance	10%	daily

Table 1: Course Assessment Summary

Installations

- R (4.5.2):
  https://cran.r-project.org
- RStudio (Desktop):
  https://posit.co/download/rstudio-desktop/
- Python (includes IDLE):
  https://www.python.org/downloads/
- (Positron?):
  https://positron.posit.co/install.html
- Anaconda: (can skip registration)
  https://www.anaconda.com/download/
- Tour RStudio!