Data Science Education in 2020

Computing, Curricula, and Challenges for the Next 10 Years

Aimee Schwab-McCoy, Catherine Baker, and Rebecca Gasper
Creighton University

AimeeSMcCoy
aimeeschwab-mccoy@creighton.edu

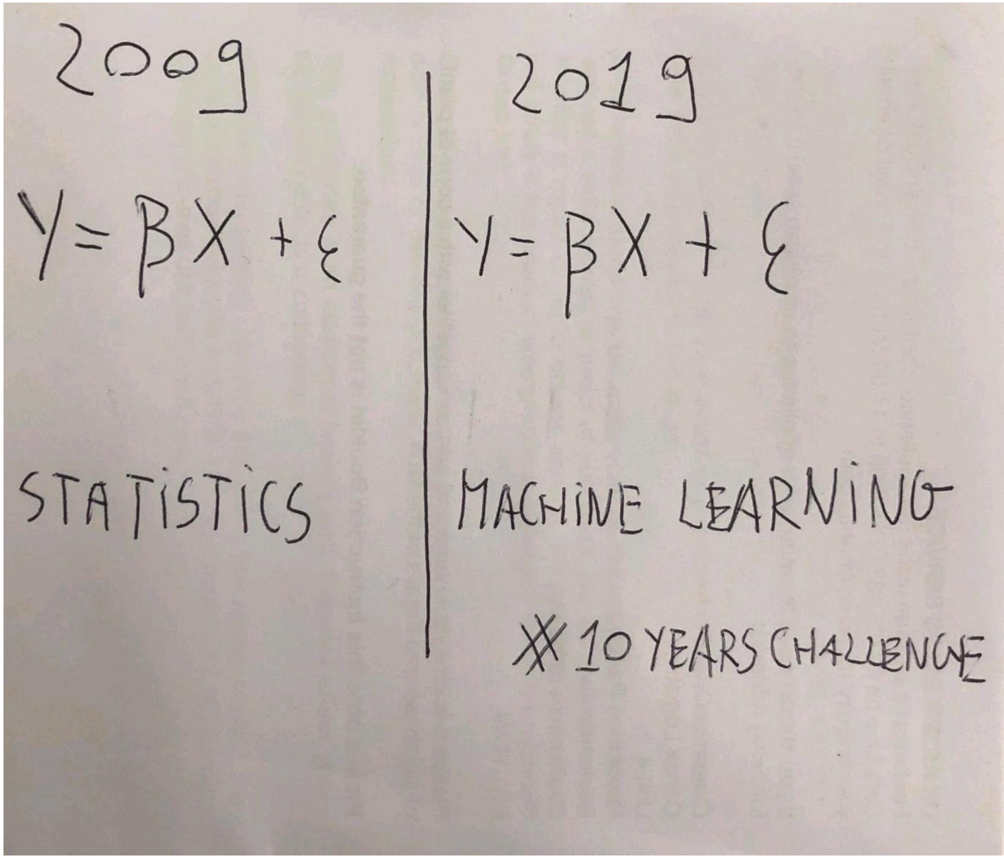
Friday, June 5, 2020

What is *data science*?

> **Data Science**: the science of planning for, acquisition, management, analysis of, and inference from data"
>
> — StatsNSF, 2014

What is data science, **as an academic discipline**?

Some say...



... but it's not so simple.

Existing recommendations

Key Competencies for an Undergraduate Data Science Major

1. Computational and statistical thinking
2. Mathematical foundations
3. Model building and assessment
4. Algorithms and software foundation
5. Data curation
6. Knowledge transference -- communication and responsibility

— Curriculum Guidelines for Undergraduate Programs in Data Science, 2016

Existing recommendations

Data Science Competence Groups

1. Data science analytics (including statistical analysis, machine learning, data mining, business analytics, others)
2. Data science engineering (including software and applications engineering, data warehousing, big data infrastructures and tools)
3. Data management and governance (including data stewardship, curation, and preservation)
4. Research methods and project management
5. Domain knowledge and expertise

— EDISON Data Science Competence Framework, 2018

Existing recommendations

Knowledge Areas for Data Science

1. Analysis and presentation
2. Artificial intelligence
3. Big data systems
4. Computing and computer fundamentals
5. Data acquisition, management, and governance
6. Data mining
7. Data privacy, security, integrity, and analysis for security
8. Machine learning
9. Professionalism
10. Programming, data structures, and algorithms
11. Software development and maintenance

— ACM Task Force on Data Science, 2019

Existing recommendations

Problem: These recommendations are at the *programmatic* level!

- We would hope that **data science majors** have the knowledge or skills listed previously (or some subset depending on the "flavor" of the program).
- But what about the introductory course?

Data Science 101

"Data Science 101" is difficult to define...

- Do we "know it when we see it"?
- Is a "consensus curriculum" on the horizon?

Data Science 101: collecting... data

In Fall 2019, we surveyed faculty who were either currently teaching data science, or planned to in the next 1-2 years.

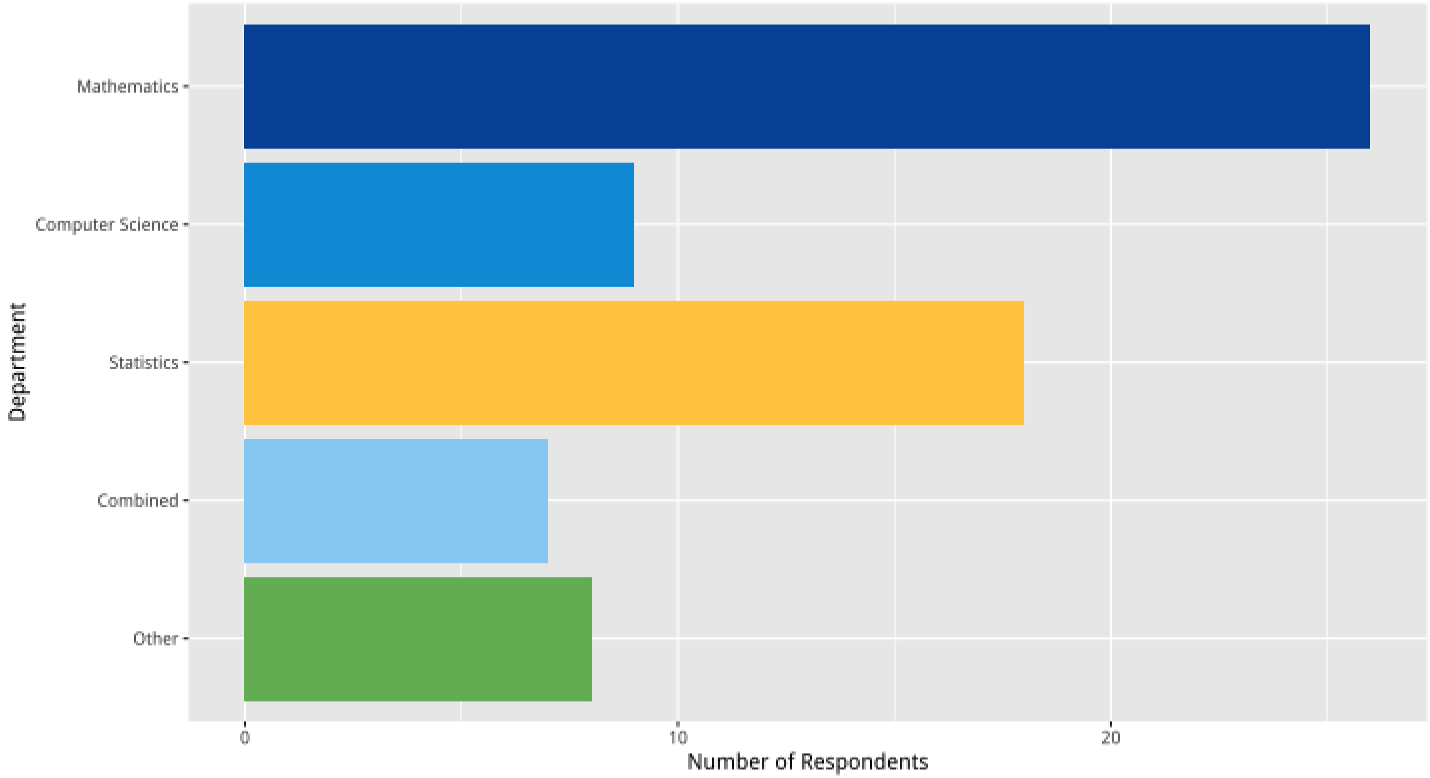
Questions:

1. Experience teaching data science
2. Content areas/topics that should be included in an introductory data science course or program
3. Software/tools used to teach data science
4. Challenges for teaching data science
5. Resources needed to become a better data science instructor

Goal: To establish a snapshot of what data science instruction looks like at the university level *today*, and what is needed to move forward.

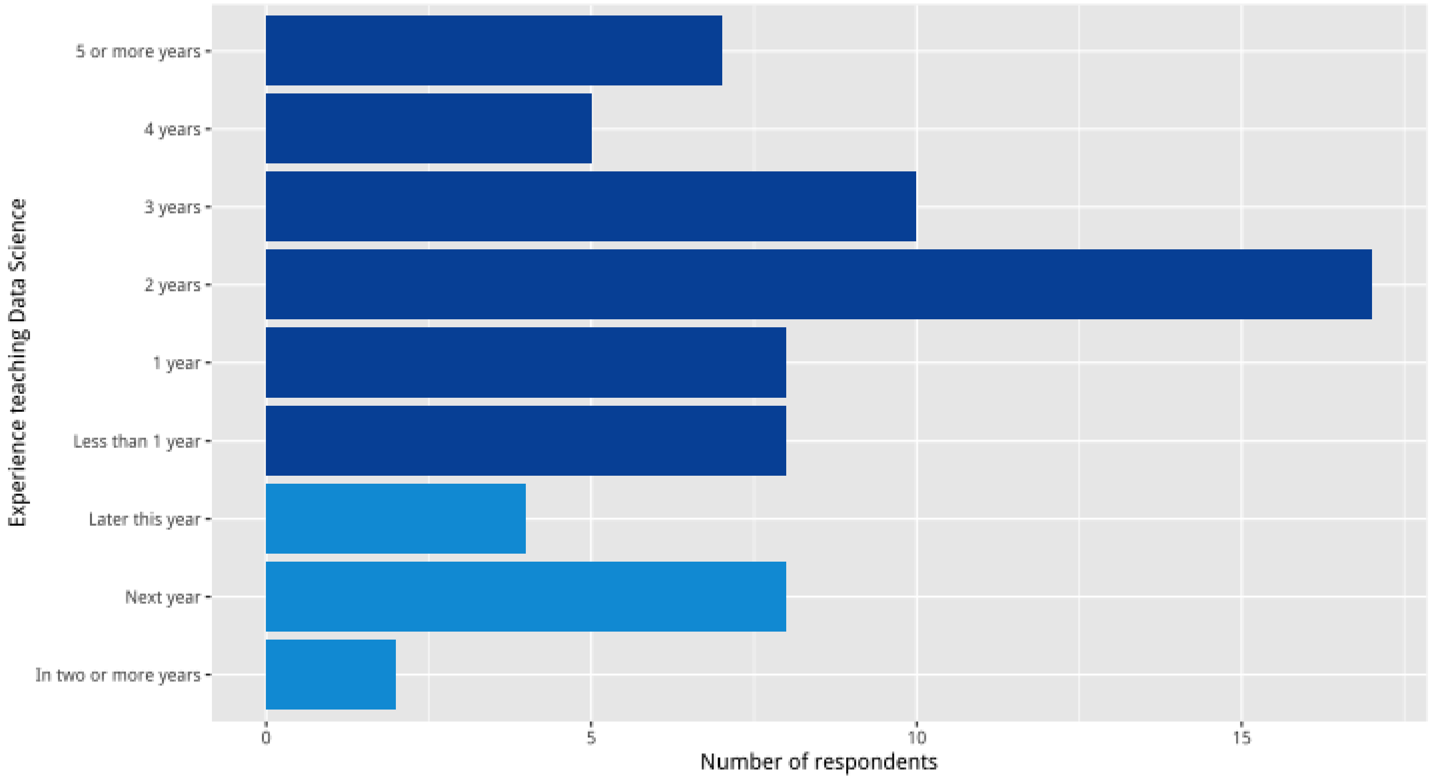
Faculty respondents

In total, 69 faculty members responded to the survey...



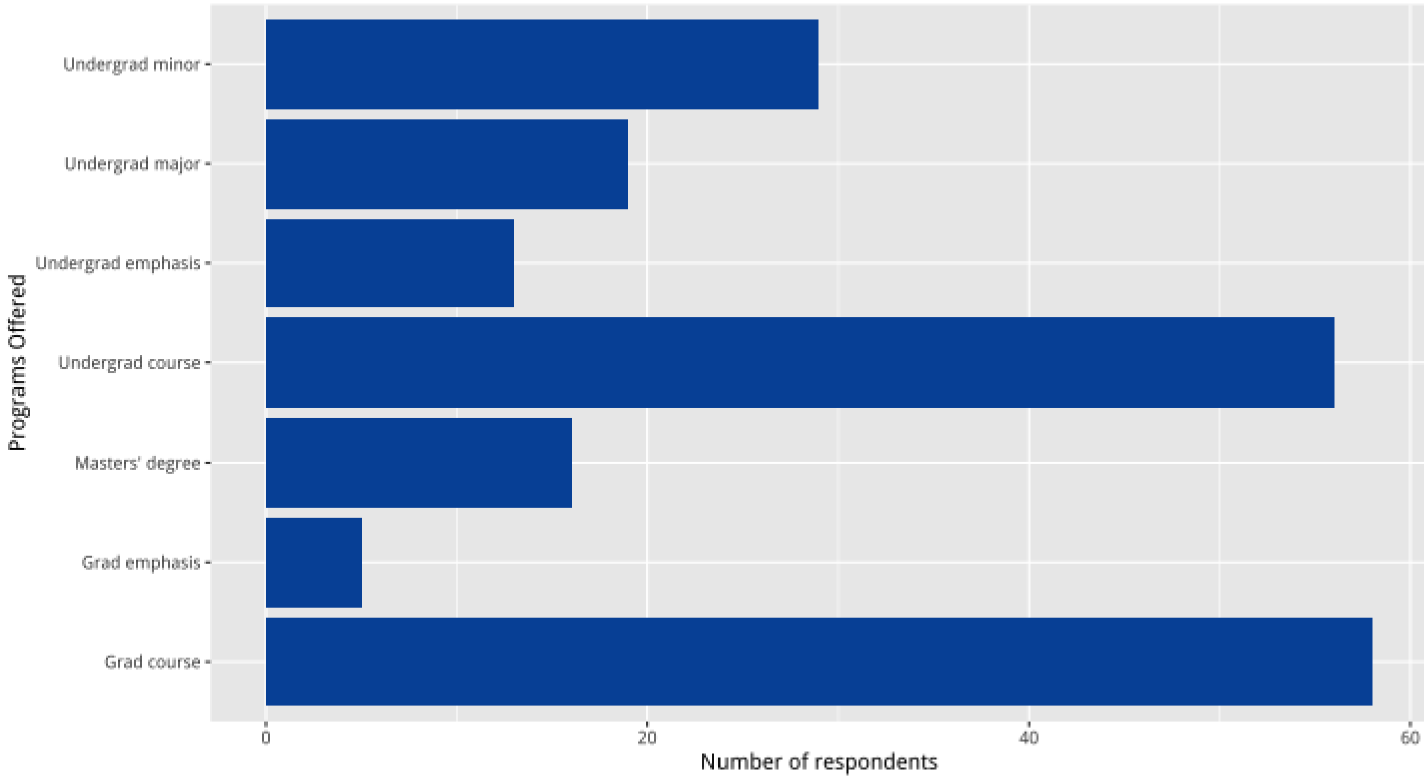
Faculty respondents

… most of whom had taught data science for the first time somewhat recently (or were planning to soon)!



Programs offered

Faculty who responded to our survey taught data science at all levels, but in most cases, the data science curriculum was not yet fully established.



Data Science 101: the three Cs

Computing

- Languages and software used

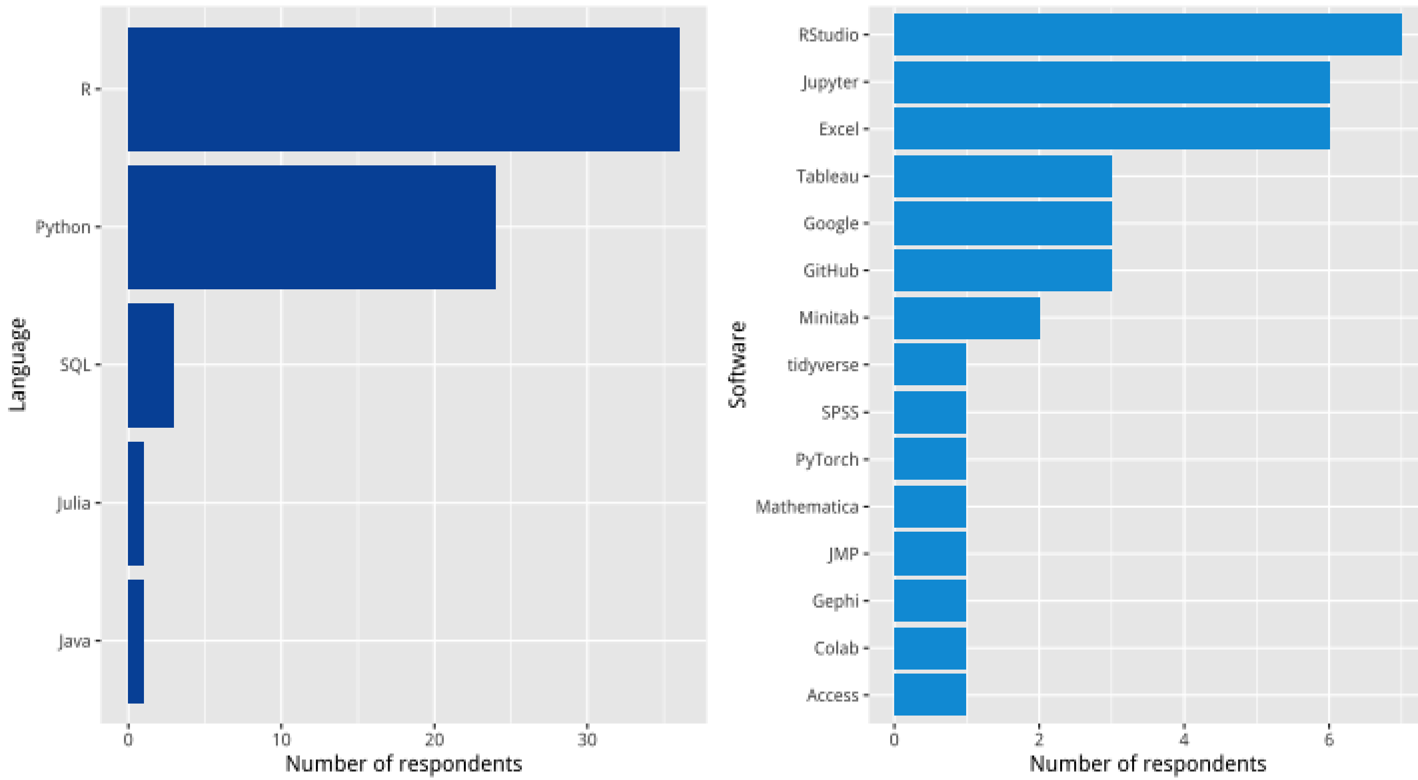
Curriculum

- Topic areas or knowledge competencies specific to "Data Science 101"
- Topic areas or knowledge competencies covered *elsewhere* in the data science curriculum
- Pre-requisites for taking data science

Challenges (and Resources)

- What challenges do new (or seasoned) data science instructors struggle with the most?
- What resources have instructors used (or would instructors need) to meet these challenges?

Language and software



Curriculum: topic areas

Faculty were asked to rate a list of provided knowledge or topic areas as:

- Included in the introductory course
- Included elsewhere in the curriculum
- Not covered in the curriculum
- Unknown

Curriculum: topic areas

Knowledge or topic areas were constructed based on three resources:

- EDISON Data Science Framework [(1)]
- Curriculum Guidelines for Undergraduate Programs in Data Science ("Park City Report") [(2)]
- ACM Task Force on Data Science Education Draft Report [(3)]

(1) Demchenko, Y., Belloum, A., & Wiktorski, T. (2017). Edison Data Science Framework: Part 1. Data Science Competence Framework. http://edison-project.eu/data-science- competence-framework-cf-ds
(2) De Veaux, R. D., Agarwal, M., Averett, M., Baumer, B. S., Bray, A., Bressoud, T. C., ... Ye, P. (2017). Curriculum Guidelines for Undergraduate Programs in Data Science. Annual Reviews of Statistics and Its Application, 4, 15–30. https://doi.org/10.1146/annurev-statistics-060116- 053930
(3) Danyluk, A., College, W., Leidig, P., Buck, S., Cassel, L., McGettrick, A., ... Wang, H. (2019). ACM Data Science Task Force Draft Report. http://dstf.acm.org/DSReportInitialFull.pdf

In total, we included 34 topics, spread across eight content areas...

Computing

- Big data infrastructures: high-performance networks, infrastructure and services management and operations (**Comp_BigData**)
- Big data technologies: batch and parallel processing, streaming systems (**Comp_BigData2**)
- Cloud computing (**Comp_Cloud**)
- Systems engineering and software engineering principles (**Comp_SysEng**)

Data management

- Data acquisition through web scraping and/or API calls (**Data_API**)
- Data architecture, data types, and data formats (**Data_Arch**)
- Data cleaning and wrangling (i.e. `tidyverse` operations) (**Data_Cleaning**)
- Data governance: managing the availability, usability, integrity and security of data (**Data_Gov**)
- Data curation and data quality (**Data_Quality**)
- Relational and non-relational databases; SQL, NoSQL, data warehouse solutions... (**Data_SQL**)
- Data storage systems, archive services, digital libraries (**Data_Storage**)

In total, we included 34 topics, spread across eight content areas (continued)...

Information technology

- Applications software requirements and design (IT_Applications)
- Information systems and collaborative systems (IT_InfoSys)
- Systems and data security: data anonymization, federated access control (IT_Systems)

Mathematics

- Linear algebra: matrix manipulation, eigenvalues, singularity (Math_Linear)
- Graph analysis: paths, social networks (Math_Graph)
- Optimization and numerical algorithms (EM) (Math_Optimization)
- Markov chain Monte Carlo and other simulation methods (Math_MCMC)

In total, we included 34 topics, spread across eight content areas (continued)...

Machine learning

- Machine learning (supervised): decision trees, Naive Bayes classification, neural networks, support vector machines (SVM), ensemble methods (ML_Supervised)
- Machine learning (reinforced): Q-learning, TD- learning, genetic algorithms (ML_Reinforced)
- Machine learning (unsupervised): clustering algorithms, principal components analysis (PCA), singular value decomposition (SVD), independent components analysis (ICA) (ML_Unsupervised)
- Text mining: natural language processing, topic modeling, text visualization (i.e. word clouds, frequency plots) (ML_TextMining)

Statistics

- Traditional statistical inference: t/z-based hypothesis tests, confidence intervals (Stat_Inference)
- Simulation- based inference: bootstrapping, randomization tests (Stat_SBI)
- Regression models: simple linear regression, least squares regression, logistic regression (Stat_Regression)
- Research methods: research cycle, hypothesis definition and testing (Stat_ResearchMethods)
- Experimental design, modeling, and planning (Stat_ExpDesign)

In total, we included 34 topics, spread across eight content areas (continued)...

Professional practice

- Data ethics and responsible data use (Prof_Ethics)
- Data life-cycle and data collection, data quality evaluation (Prof_Lifecycle)
- Data protection, backup, and privacy (Prof_Protect)
- Reproducible research: computational and empirical reproducibility, statistical reproducibility (Prof_Reproducible)

Visualization

- Exploratory data analysis: multivariate plots and graphs, summary statistics (Viz_EDA)
- Customizing data visualizations: color and composition, accessibility, "grammar of graphics" (Viz_Custom)
- Advanced data visualizations: creating animated or dynamic visualizations, maps (Viz_Advanced)

Most popular topics in "DS 101"

Topic	Intro Course	Program	Covered	Not Covered	Unknown	Total
Exploratory data analysis	57	5	62	2	2	66
Data cleaning/wrangling	52	7	59	6	0	65
Ethics in data science	44	15	59	3	4	66
Data curation	36	15	51	7	7	65
Reproducible research	36	18	54	5	7	66
Regression models	36	27	63	1	2	66
Data lifecycle	35	11	46	12	8	66
Data architecture	28	28	56	9	2	67
Text mining	28	23	51	15	2	68
Research methods	28	33	61	3	3	67
Customizing data visualizations	28	25	53	8	6	67
Supervised learning	27	28	55	9	4	68

Most common topics in *curriculum*

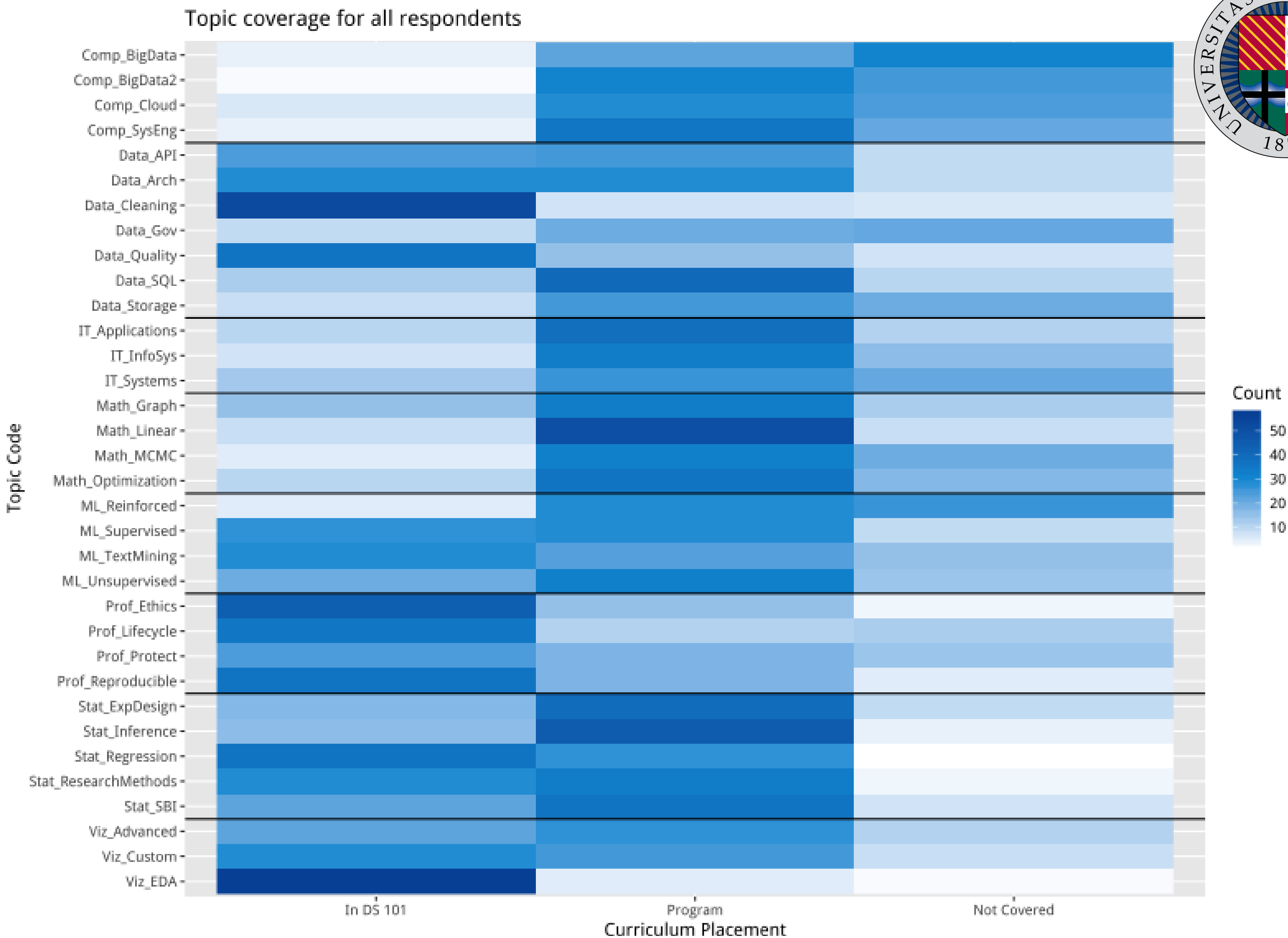
Topic	Intro Course	Program	Covered	Not Covered	Unknown	Total
Regression models	36	27	63	1	2	66
Exploratory data analysis	57	5	62	2	2	66
Traditional inference	16	45	61	4	3	68
Research methods	28	33	61	3	3	67
Data cleaning/wrangling	52	7	59	6	0	65
Ethics in data science	44	15	59	3	4	66
Linear algebra	8	50	58	8	2	68
Simulation-based inference	22	36	58	7	3	68
Data architecture	28	28	56	9	2	67
Experimental design	17	39	56	9	3	68
Supervised learning	27	28	55	9	4	68
Reproducible research	36	18	54	5	7	66

Most common topics in *curriculum*

Topic	Intro Course	Program	Covered	Not Covered	Unknown	Total
Regression models	36	27	63	1	2	66
Exploratory data analysis	57	5	62	2	2	66
Traditional inference	16	45	61	4	3	68
Research methods	28	33	61	3	3	67
Data cleaning/wrangling	52	7	59	6	0	65
Ethics in data science	44	15	59	3	4	66
Linear algebra	8	50	58	8	2	68
Simulation-based inference	22	36	58	7	3	68
Data architecture	28	28	56	9	2	67
Experimental design	17	39	56	9	3	68
Supervised learning	27	28	55	9	4	68
Reproducible research	36	18	54	5	7	66

Most common topics *beyond intro*

Topic	Intro Course	Program	Covered	Not Covered	Unknown	Total
Traditional inference	16	45	61	4	3	68
Research methods	28	33	61	3	3	67
Linear algebra	8	50	58	8	2	68
Simulation-based inference	22	36	58	7	3	68
Data architecture	28	28	56	9	2	67
Experimental design	17	39	56	9	3	68
Supervised learning	27	28	55	9	4	68
Customizing data visualizations	28	25	53	8	6	67
Databases (SQL)	12	40	52	10	4	66
Unsupervised learning	20	32	52	14	2	68
Text mining	28	23	51	15	2	68



Topic coverage for all respondents
Comp_BigData
Comp_BigData2
Comp_Cloud
Comp_SysEng
Data_API
Data_Arch
Data_Cleaning
Data_Gov
Data_Quality
Data_SQL
Data_Storage
IT_Applications
IT_InfoSys
IT_Systems
Math_Graph
Math_Linear
Math_MCMC
Math_Optimization
ML_Reinforced
ML_Supervised
ML_TextMining
ML_Unsupervised
Prof_Ethics
Prof_Lifecycle
Prof_Protect
Prof_Reproducible
Stat_ExpDesign
Stat_Inference
Stat_Regression
Stat_ResearchMethods
Stat_SBI
Viz_Advanced
Viz_Custom
Viz_EDA
Topic Code
In DS 101
Program
Not Covered
Curriculum Placement
Count
50
40
30
20
10

Pre-requisites

Are certain topics omitted because they were covered *before* introductory data science?

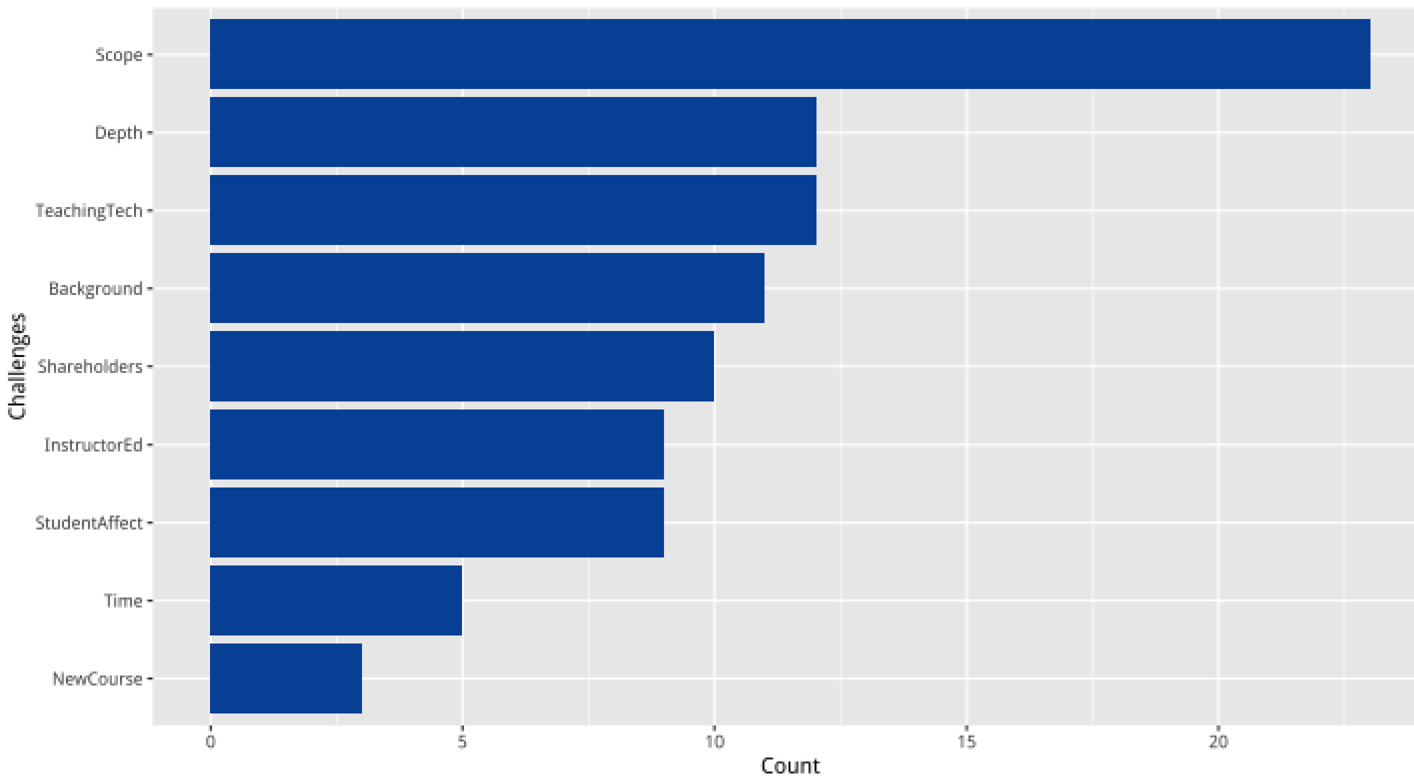
- 28% of instructors (19/69) taught courses with a **computing** pre-requisite
- 25% of instructors (17/69) taught courses with a **statistics** pre-requisite
- 13% of instructors (9/69) taught courses with **both** statistics and computing required

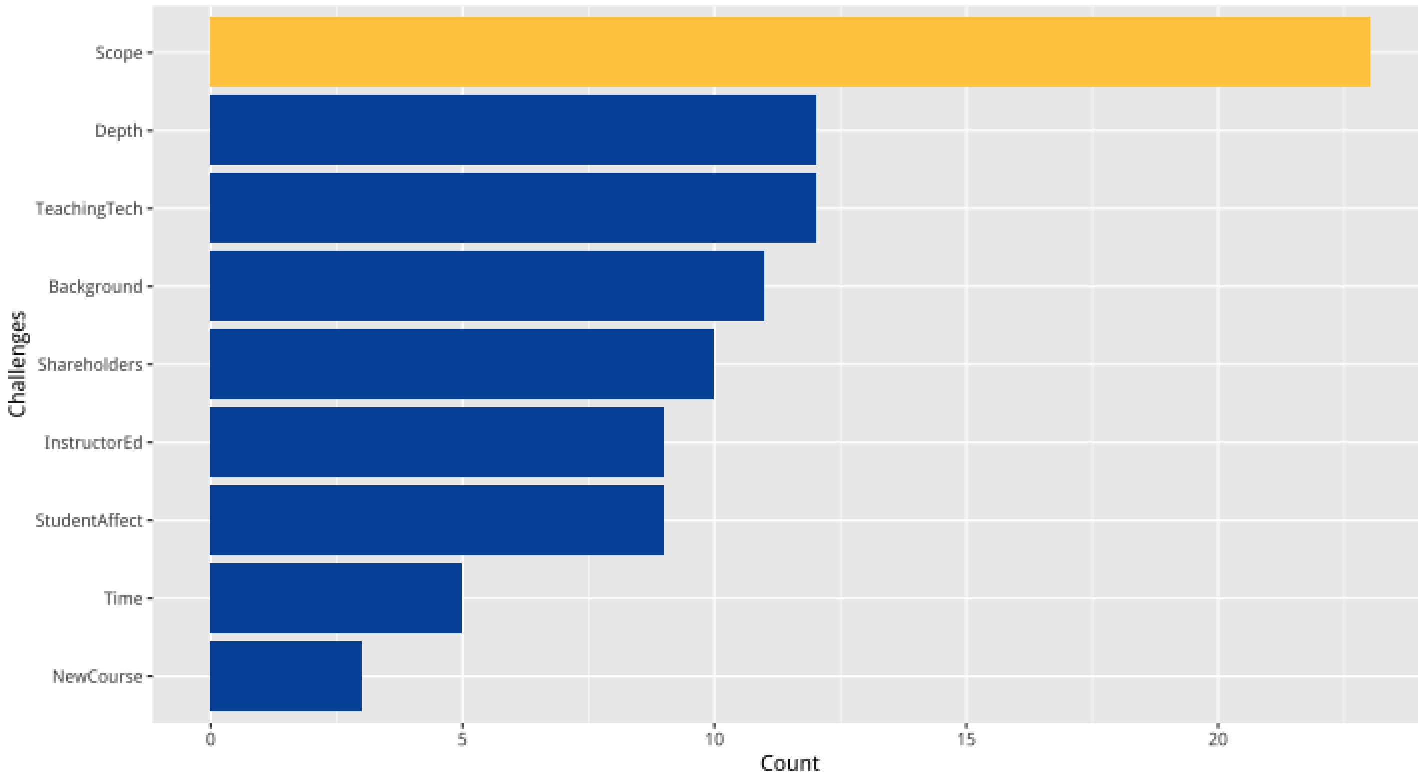
For many students, introductory data science is their first exposure to computing, coding, statistics, or data analysis!

Challenges

Faculty provided short responses for the biggest challenges they faced teaching data science.

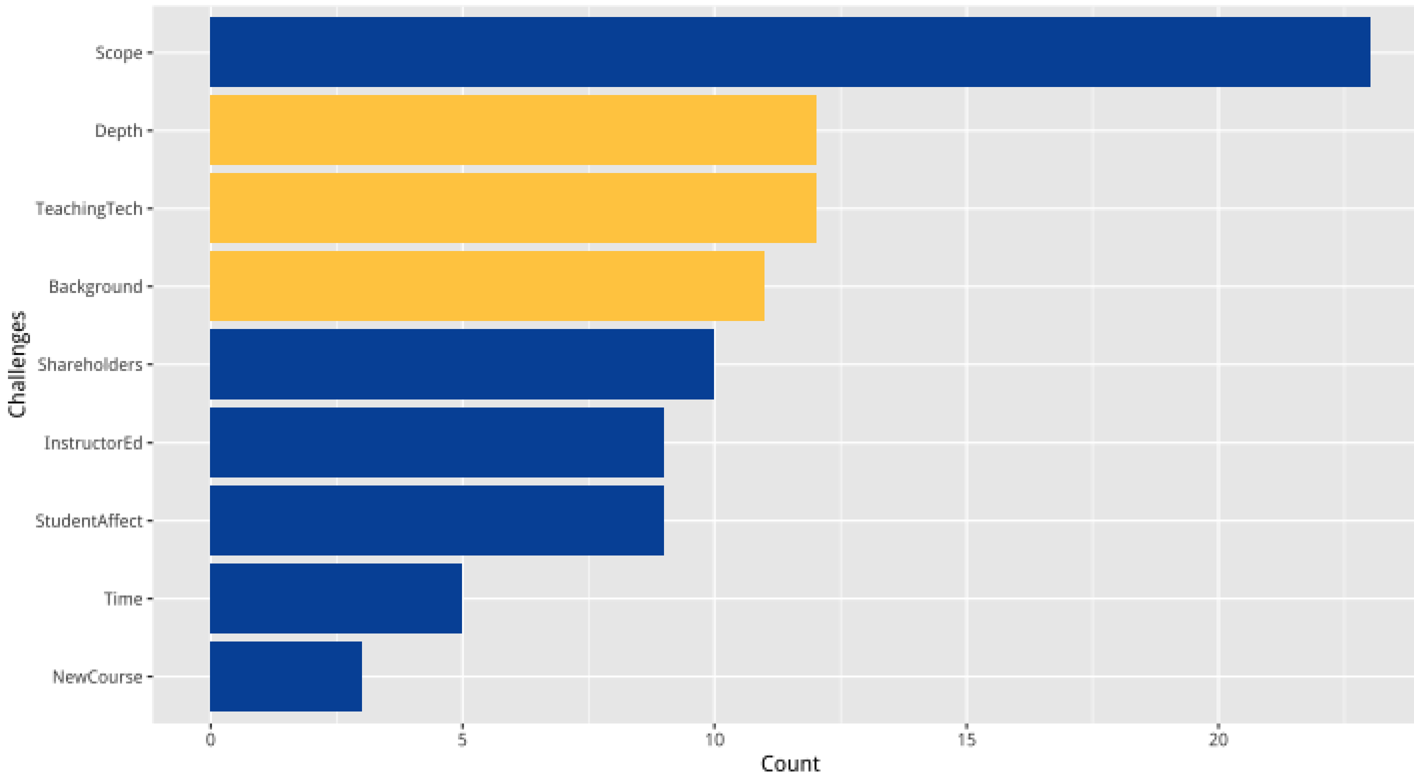
- Using two-step axial coding resulted in the categories below:





Scope represented the “breadth” of the course, including topic standardization, course pacing, and building unified themes throughout the introductory data science course.

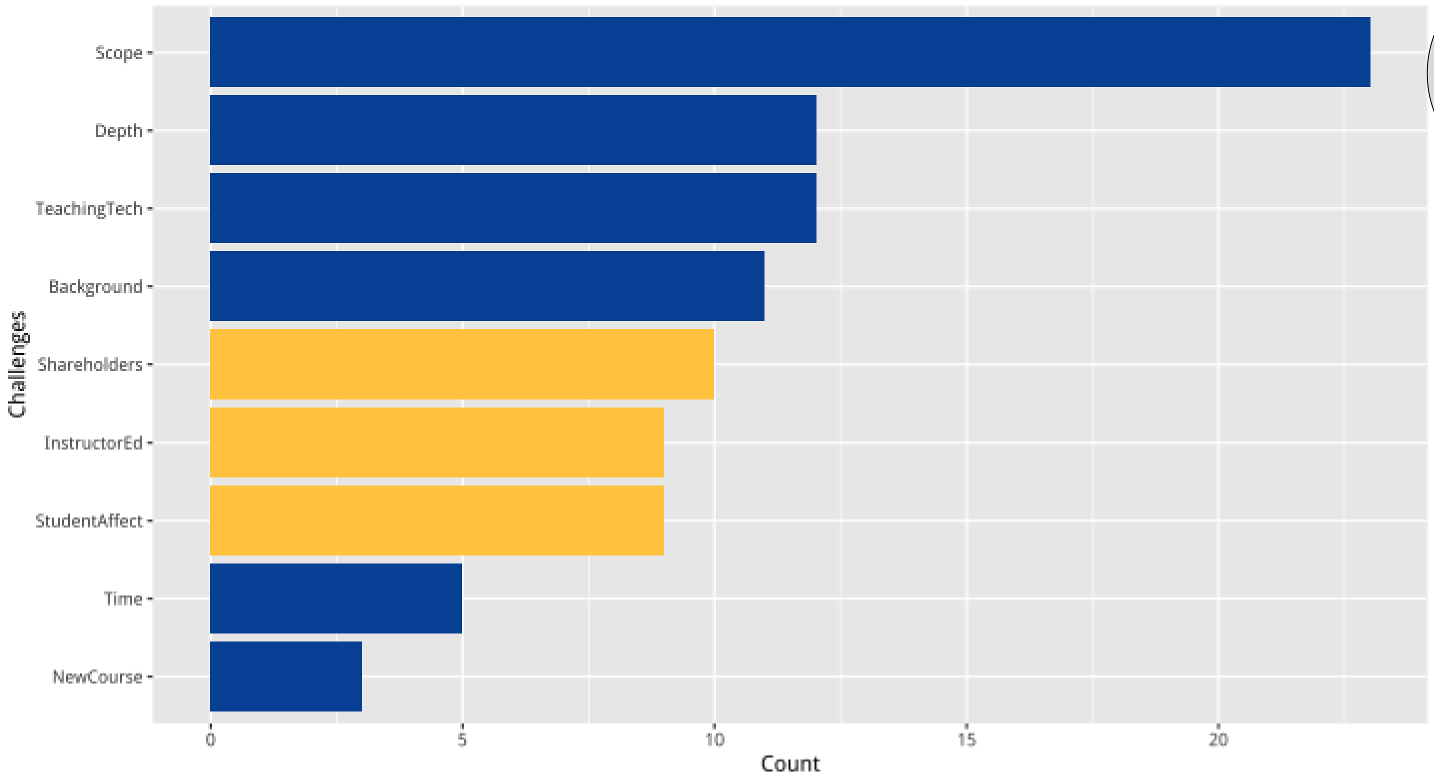
- Difficulty “[n]arrowing down the vast amount of information into one semester introductory course"
- Differentiating introductory data science courses from introductory statistics course
- Lack of “consensus curriculum”: “Obtaining a clear definition of DS and what is required and what is recommended”



Depth: Are we covering topics in sufficient detail to stimulate critical thinking about data science?

Technology: What is "best" to use, and to what extent?

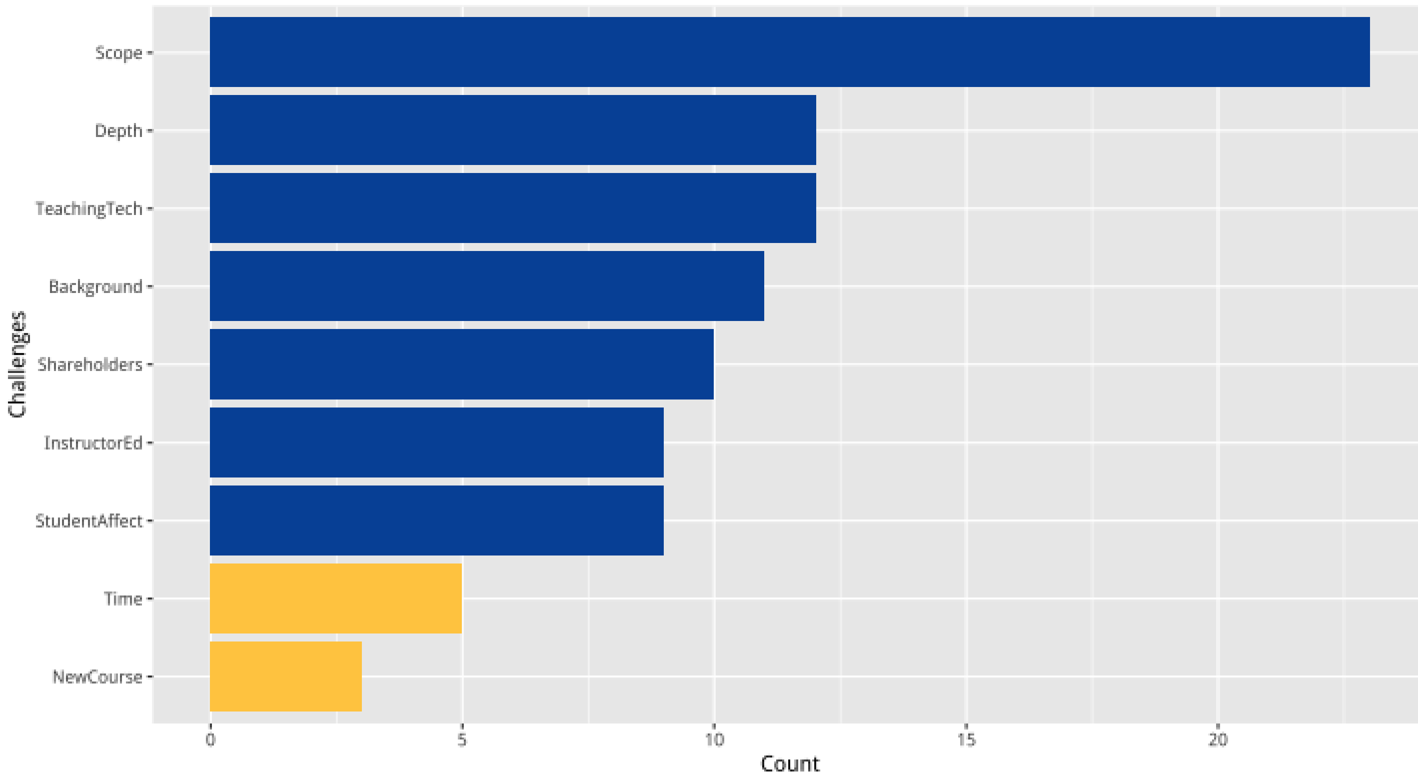
Background: Are students adequately prepared to take data science?



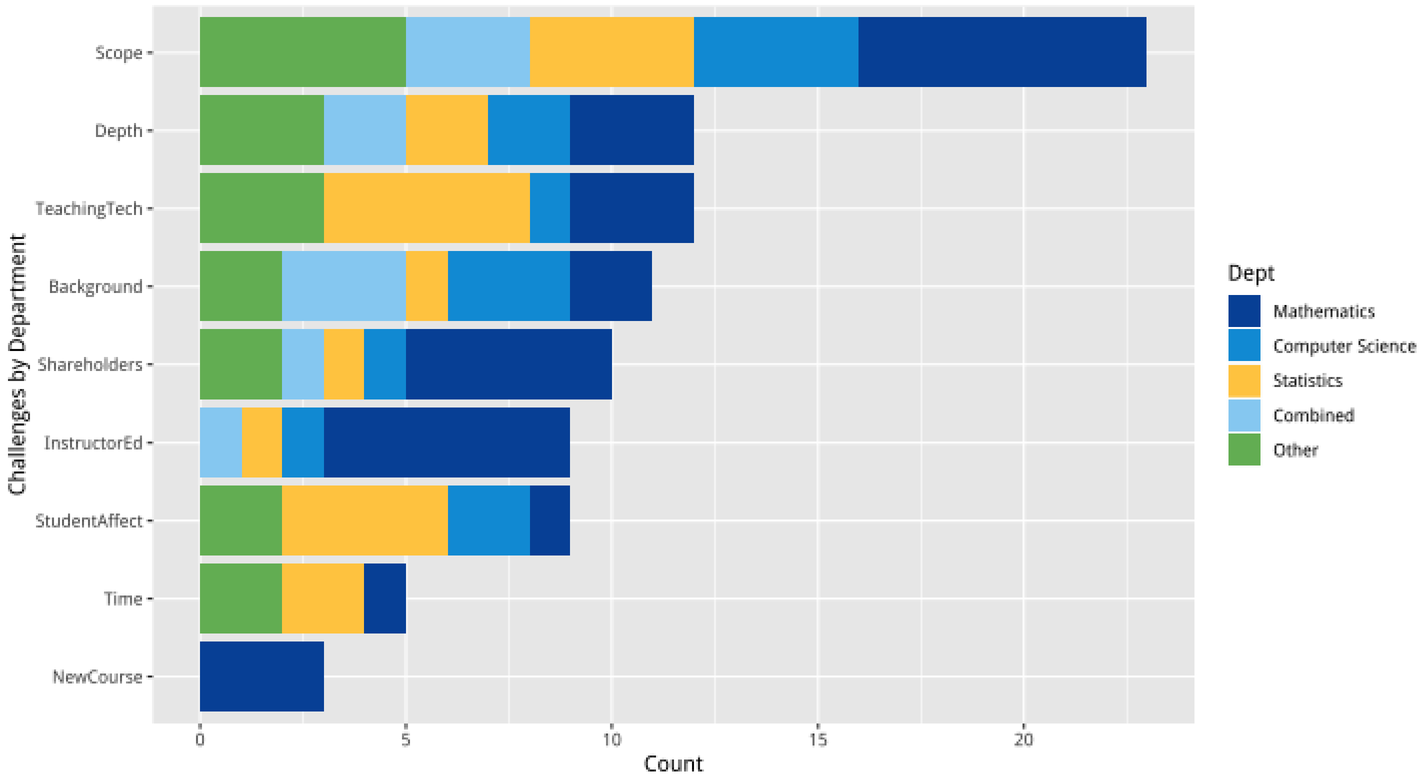
Shareholders: Are our graduates prepared? Are our administrators supportive?

InstructorEd: Are *instructors* adequately prepared to *teach* data science?

StudentAffect: Do students *care* about data science?



Time and **NewCourse**: The usual anxieties of being a professor



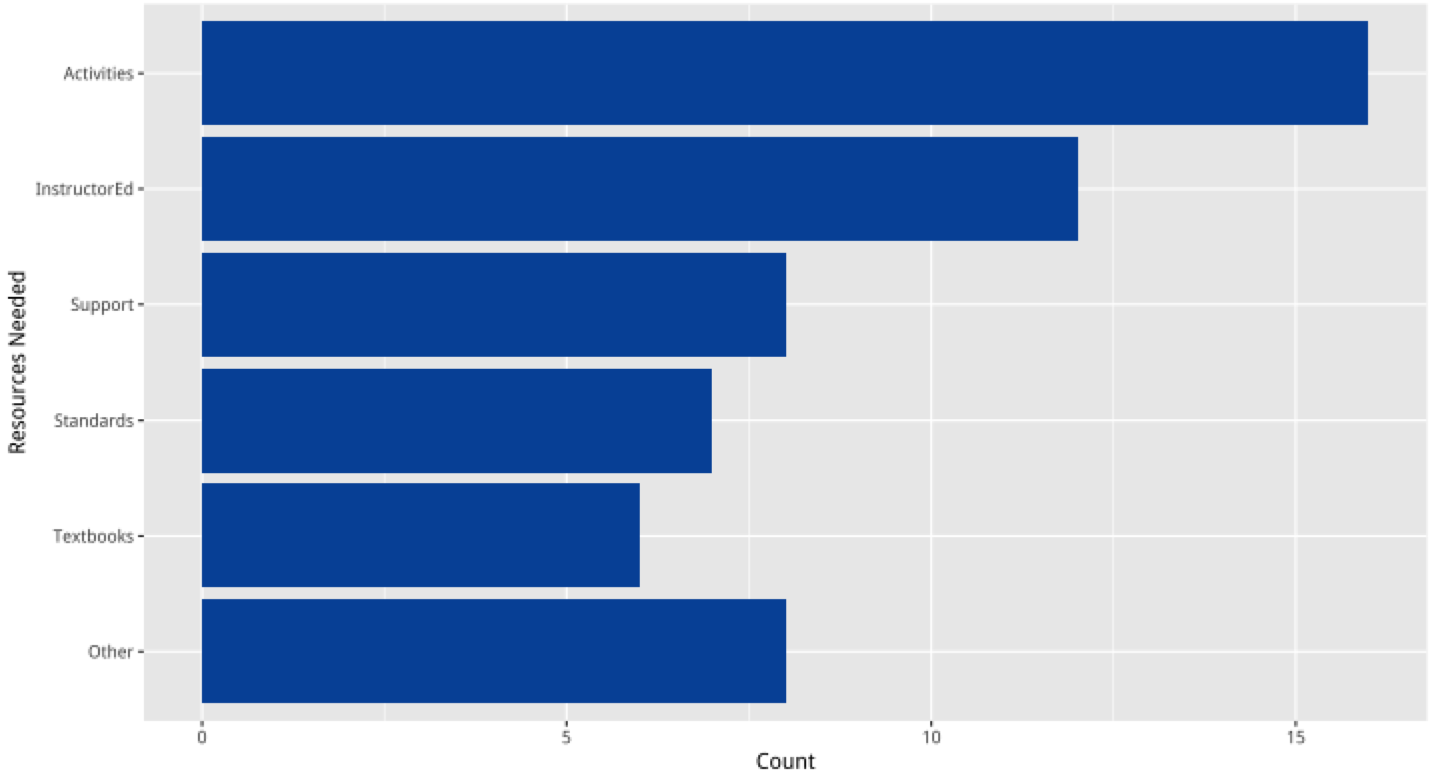
Many of these challenges applied across the board, regardless of instructor background and department.

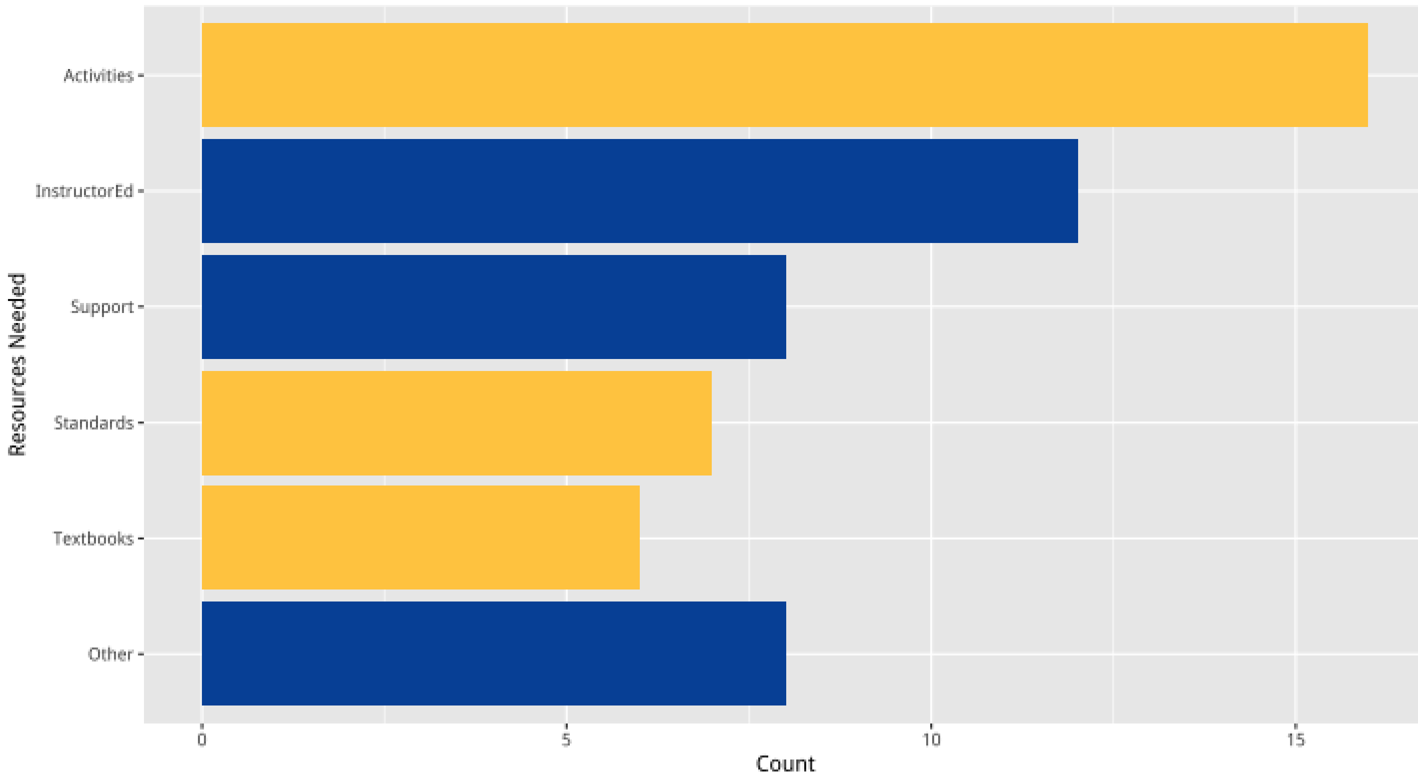
- Possible exception: - *Mathematics* faculty were more interested in continuing professional development and instructor education before teaching data science.

Resources needed

So how do we address these challenges?

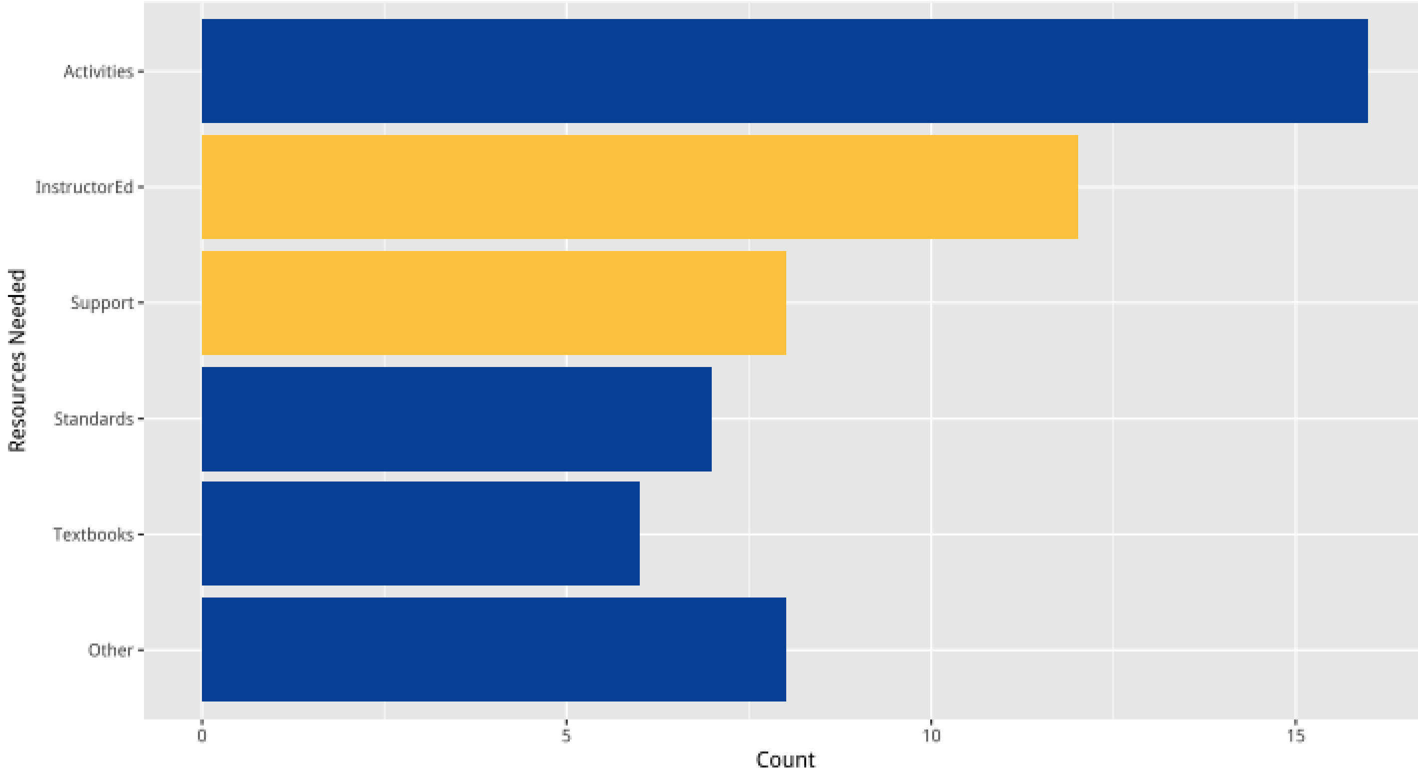
- Using the same two-step axial coding procedure:



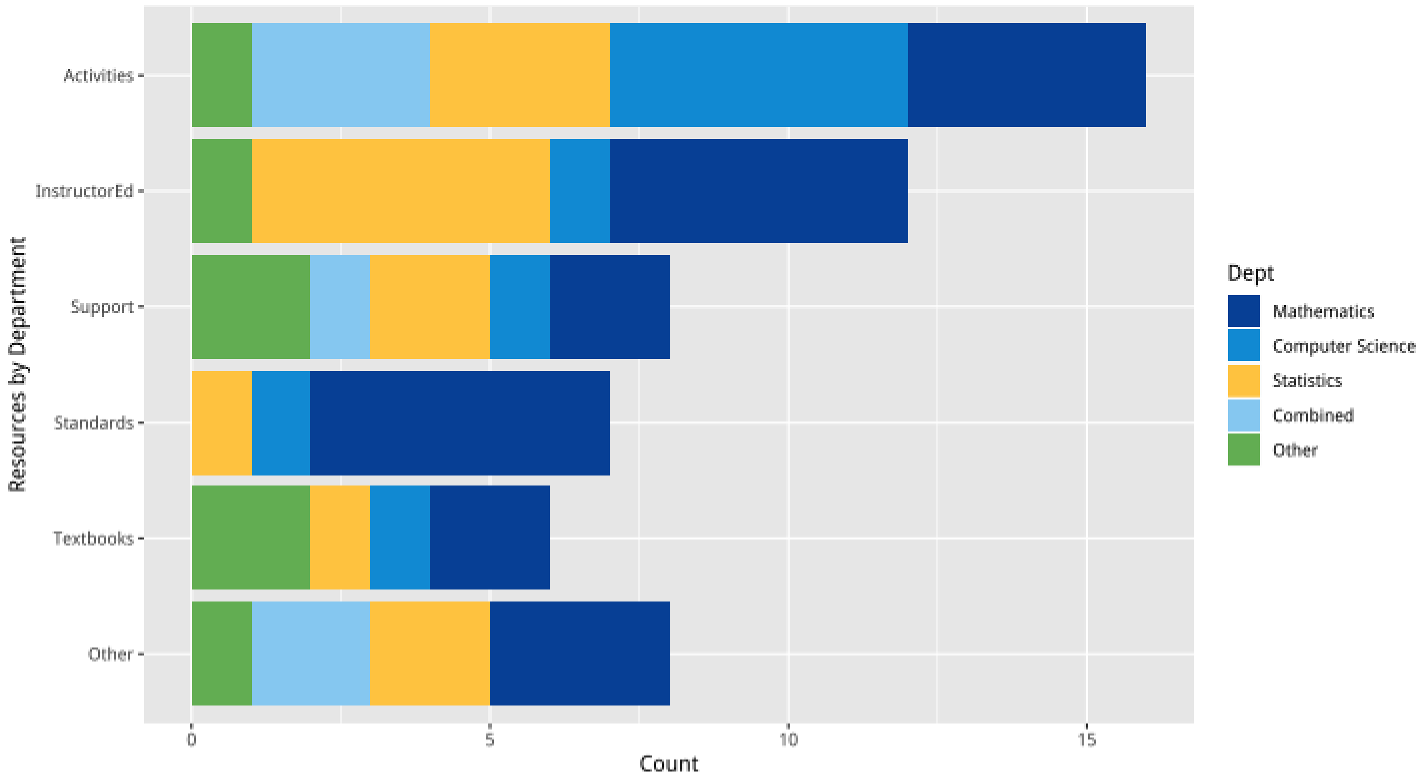


Many of the resources requested were direct solutions to the questions of course scope and depth:

- **Activities** suitable for classroom adoption
- A set of **standards** for course topics and content
- **Textbook** resources



Other instructors wrote about the need for additional institutional **support** or professional support through **instructor education**.



There were some differences in the requested resources by discipline:

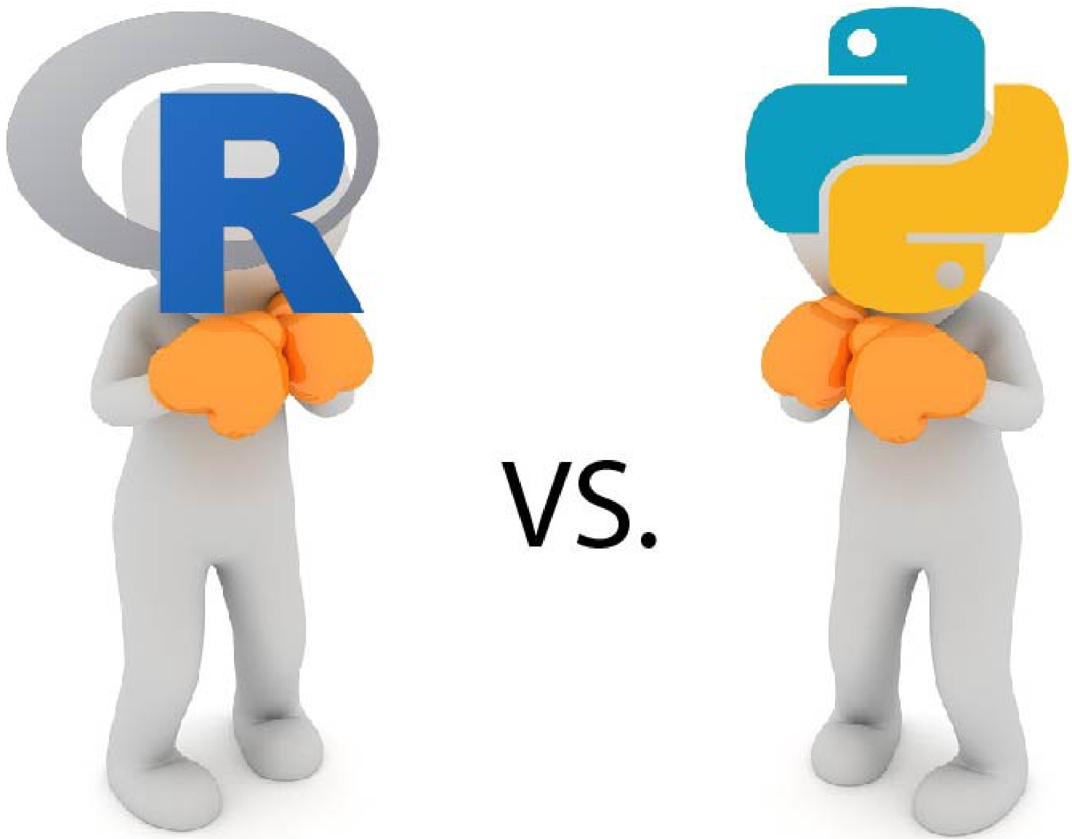
- *Mathematics* instructors were more interested in a set of standards for a data science course.
- *Statistics* and *Mathematics* instructors wrote more often about instructor education through professional development workshops.
- *Computer science* instructors were primarily interested in activities and labs to incorporate into their own courses.

What have we learned?

Are we at a "consensus" as to what should be taught in introductory data science?

We're getting there! Over 50% of all respondents included these topics in their introductory course:

- Exploratory data analysis and data visualization
- Data cleaning and wrangling
- Data science ethics
- Date curation
- Reproducible research
- Regression models
- Data life-cycle



We are not getting into a language debate (this is the wrong talk for that), but we're glad to see an abundance of coding in the classroom!

Like ice cream, one flavor of introduction to data science doesn't suit all!

- Computationally-infused statistics?
- Statistically-infused computation?
- Somewhere in-between?

Where do our classes fall? Where will they fall 10 years from now?

Some final thoughts

There are definitely challenges for instructors that need to be addressed!

- Building a community of practice for **data science education** and continuing to support great open-source material is one way to help data science meet the growing demand.

Lingering questions:

1. *How* is coding being taught in the data science classroom, and does that differ from introductory computing/statistics?

2. Instructional and assessment modes for data science: active learning, inquiry-based learning, lecture-style, etc?

3. Growth and trajectory of data science courses/programs?

Thank you!

Shameless plug: We are recruiting volunteers to participate in short interviews (~30-90 minutes) about teaching data science!

Interested participants should contact Aimee Schwab-McCoy:

- AimeeSMcCoy
- aimeeschwab-mccoy@creighton.edu