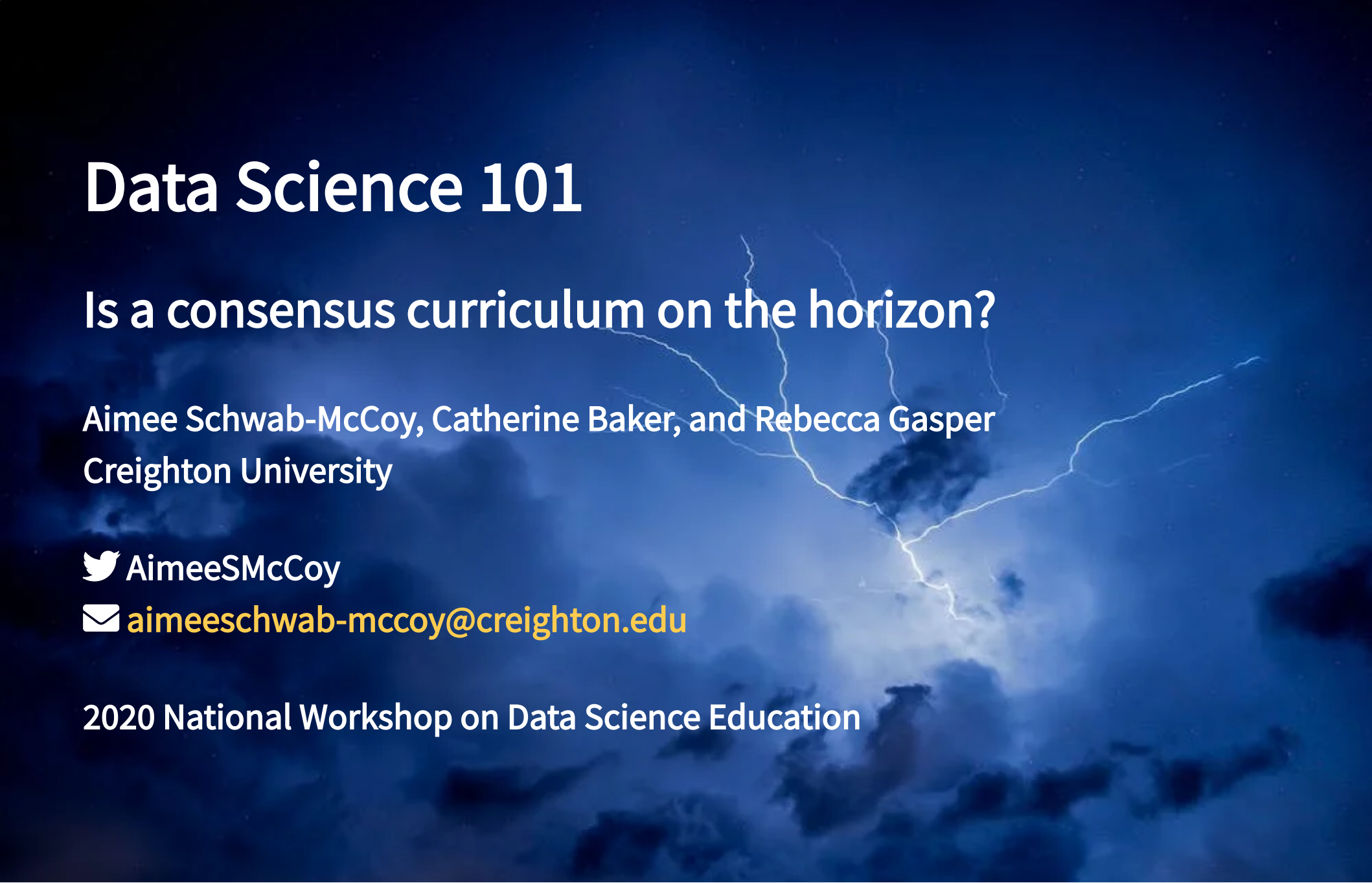

Data Science 101

Is a consensus curriculum on the horizon?

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

AimeeSMcCoy
aimeeschwab-mccoy@creighton.edu

2020 National Workshop on Data Science Education

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.

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

34 final 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

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

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!

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
- Data curation
- Reproducible research
- Regression models
- Data life-cycle

Remember, one "flavor" of introduction to data science probably doesn't suit all!

Is your course:

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