Data Science for Two Year Colleges

Rob Gould

Roxy Peck

Brad Thompson

A Few Notes Before We Get Started…

- Please use the mute button on the bottom left side of your screen to mute your microphone.
- Questions? Please enter them into the “Chat” window using the Chat button at the bottom of your screen. The questions will be monitored and used as the basis for our Q&A time at the end of the webinar.
- This webinar is being recorded and will be posted on the project website: https://www.amstat.org/ASA/Education/Two-Year-College-Data-Science-Summit.aspx

Two-Year College Data Science Summit May 2018

- Project goal: Develop guidelines and recommendations for data science programs at two-year colleges.
- Summit funded by NSF with additional support from ASA and Booz Allen Hamilton.
- Over 70 participants from two-year colleges, four-year colleges, industry and government.
- Summit speakers and group discussions were designed to inform writing of guidelines and recommendations.

Three Types of Potential Data Science Programs at Two-Year Colleges

- Associate degree to workforce
  - Goal is to produce “middle-skills” data practitioners. Graduates would typically work as part of a data science team with duties that might include data acquisition, data cleaning, data base management, combining data from different sources, producing simple data visualizations, and basic data analysis.
- Associates degree for transfer to a four-year data science program
  - Many challenges in designing these types of programs. There is no consensus among four-year programs on curriculum content or on the mix and balance of foundations in mathematics, statistics, computer science. This makes articulation difficult
- Certificate programs
  - Diverse in content and focus, usually tailored to local and regional workforce needs. Examples include data management, cyber security, business analytics.

Two-Year College Data Science Summit - May 2018

Key Questions Raised:

- Which type of program is most feasible?
- What specific jobs will graduates be prepared for?
- What level of mathematics should be required?
- What would a data science capstone course in a two-year program consist of?
- In what academic department should these programs reside?
- What specific technologies/methods should be in the curriculum?

Important Issues Discussed:

- Will industry accept associate level degrees or certificates in this field?
- How critical will partnerships with local industry be for program and student success?

Post-Summit

- Writing Teams
  - Three writing teams, each made up of 3 – 4 people, were charged with synthesizing the discussions that occurred at the summit into recommendations and guidelines. There was one team for each of the three program types.
- Rob Gould served as overall editor, and merged the work of the three writing teams into a draft report.
- Summit attendees were provided with an opportunity to review the report and provide feedback.
- Report was revised based on input from the writing teams and summit attendees.

It is recommended by the TYCDSS committee that two-year colleges developing new programs in data science should:

Note: Each recommendation is explored more in-depth within the 2YCDSS Report

Recommendation 1

Create courses that provide students with a modern and compelling introduction to statistics.

Include exploratory data analysis, the use of simulations, randomization-based inference, and an introduction to confounding and causal inference.

Recommendation 2

Ensure that students have ample opportunities to engage with realistic problems using real data so that they see statistics as an important investigative process useful for problem solving and decision-making.

Recommendation 3

Explore ways of reducing mathematics as a barrier to studying data science while addressing the needs of the target student populations and ensuring appropriate mathematical foundations.

Consider a "math for data science" sequence that emphasizes applications and modeling.

Recommendation 4

Design courses so that students solve problems that require both algorithmic and statistical thinking.

Recommendation 5

Expose students to technology tools for reproducibility, collaboration, database query, data acquisition, data curation, and data storage.

Require students to develop fluency in at least one programming language used in data science and encourage learning a second language.

Recommendation 6

Infuse ethical issues and approaches throughout the curriculum in any program of data science.

Recommendation 7

Foster active learning and use real data in realistic contexts and for realistic purposes.

Programs should consider portfolios as summative and formative assessment tools that both improve and evaluate student learning.

Program Learning Outcome Categories

- Computational Foundations
- Computational Thinking
- Statistical Foundations
- Statistical Thinking
- Statistical Modeling
- Data Management and Curation
- Mathematical Foundations
- Productivity Foundations

Statistical thinking vs. Outcomes

Foundations

- Determine if conclusions are appropriate based on study design.
- Produce and interpret data visualizations
- Produce and interpret numerical summaries
- Produce and interpret confidence intervals
- Formulate statistical claims in terms of hypotheses; carry out and interpret tests.
- Investigate and explore multivariate relationships

Thinking

- Recognize questions and problems that can be investigated with data.
- Identify data appropriate to answer statistical questions.
- Explain the role of randomization and random sampling
- Identify sources of variability including sampling variability when drawing conclusions.

Mathematical Foundations

- Calculus; Matrices and basic linear algebra; Basic probability.
- Math for Data Science:
  Emphasize modeling, connection to real-world problems, linear algebra.
- See "Mathematical Foundations I and II" from Park City Math Institute Curriculum Guideline for UG Programs in DS
  https://www.amstat.org/asa/files/pdfs/EDU-DataScienceGuidelines.pdf

"Productivity"

- Communication
- Collaboration
- "Habits of mind"

Ethics

- Embed throughout. A course devoted exclusively to ethics may not be necessary if topics are explicitly included in classes, but is probably a good idea.
- Throughout, students should gain understanding of how privacy and confidentiality may be compromised, how societal inequities may be propagated, and the importance of consent in data collection.

Achieving Outcomes

- "Exposure" to an outcome: students have seen the content, are aware of its existence, and can describe what it means.
- "Working Knowledge" means students can apply this outcome to routine problems, but may need some assistance in novel situations.
- "Mastery" means students have a fairly deep understanding that allows them to apply a concept to a new situation or transfer to a new context.

Example Data Management and Curation Outcome Levels

Outcome	**To Work**	**Transfer**	**Certificate**
DMC.A. Demonstrate ability to acquire and represent data in diverse formats and structures, such as databases, web pages, JSON, etc.	working knowledge/mastery	exposure	working knowledge/mastery
DMC.B. Apply exploratory data analysis to identify problems in the data.	working knowledge	exposure	exposure/working knowledge
DMC.C. Clean data and prepare data for analysis and identify problems that might arise from assumptions made during this process.	working knowledge	working knowledg e	exposure/working knowledge
DMC.D. Identify problems associated with missing data.	exposure	exposure	exposure/working knowledge
DMC.E. Combine multiple data sources to address a given statistical goal.	working knowledge	exposure	working knowledge
DMC.F. Manage databases.	working knowledge/mastery	exposure	working knowledge/mastery
DMC.G. Explain issues related to data privacy and security.	working knowledge	exposure	working knowledge/mastery
DMC.H. Construct, maintain, and share files in a version control system.	working knowledge	exposure	working knowledge/mastery

Achieving Mastery

- Multiple exposures in different concepts.
- Capstone project or extensive project work with realistic problems based on real data.

Sample Curriculum for a Direct-to-Work program

	Semester 1	**Semester 2**	**Semester 3**	**Semester 4**
Course 1	Intro to Data Science	Database Design and Programming	Math for Data Science	Applied Predictive Modeling
Course 2	Intro to Programming	Data Visualization	Intro to Predictive Modeling	Data Analytics
Course 3	Intro to Statistics	Applied Data Programming	Data Structures	Capstone

Other questions and concerns raised during the TYCDSS

- Is a data science program feasible at a particular two-year institution?

- What are the minimum degree requirements for data science job postings in the area served by the college?

- Is an associate degree in data science creating a new role within the industry?...”Entry-level data scientist”? “Data Practitioner”?

- Aside from what type of program, in what department will the program operate?
- Will there be qualified faculty available to develop and teach these new courses?
- As data science is an emerging field, how will curriculum be continually evaluated and redesigned to meet changing needs of industry?

Continuing the Conversation…

Keep in Touch

- To join the Google group discussion forum, please email Steve Pierson at **spierson@amstat.org**

- https://groups.google.com/forum/#!forum/2ydatascience

Acknowledgements: Steering Committee

- Rob Gould and Roxy Peck, chairs; Beth Hawthorn, Nicholas Horton, Randy Kochevar, Brian Kotz, Donna LaLonde, Steve Pierson, Mary Rudis, Brad Thompson, Heikki Topi

- Writing Team: Rob Gould and Roxy Peck; Julie Hanson, Nicholas Horton, Brian Kotz, Kathy Kubo, Joyce Malyn-Smith, Mary Rudis, Brad Thompson, Mark Daniel Ward, Rebecca Wong.

- Summit Participants: Please see full list at https://www.amstat.org/asa/files/pdfs/2018-2YCDSS-Participants.pdf