IDENTIFYING CORE COMPETENCIES IN INTRODUCTORY DATA SCIENCE FOR A DIVERSE STUDENT AUDIENCE: A NATIONAL STUDY

Jennifer Broatch (Arizona State Univ.)
Matthew Beckman (Penn State)

OVERVIEW

This project explores the current national landscape of introductory data science education, identifying the **core competencies** emphasized in existing courses and examining how students engage with and develop data literacy across **diverse learning environments**. The study focuses on institutions offering introductory data science courses for a general audience, where course structures, student backgrounds, and learning objectives vary widely.

To provide a **comprehensive picture** of data science education for non-majors, the study integrates:

Syllabi Review: Content analysis of previously collected syllabi to identify common coverage of topics

Introductory Data Science Topic Instructor Survey (MASDER -SDSattitudes.com): Collect detailed course information, including topic coverage and teacher demographics

Instructor interviews: Explore deeper insights into learning experiences, instructional methods, and perceptions of essential competencies

PHASE II: Student Survey of Motivational Attitudes toward Data Science (S-SOMADS) : Assess what students' attitudes toward data science are, and how attitudes relate to student and instructor characteristics and classroom environment

HOW TO GET INVOLVED

If you teach or have taught an Introductory Data Science course, please consider lending us your expertise.

There are varying levels of **PAID** involvement.

- Data Science Instructor Survey
- Follow-up interview,
- Student survey administration (Phase II).

Total Compensation level: $100 (Can be prorated)

Any level of participation is welcome.

SCAN ME

Mid-Career Advancement

Research funded by an NSF Mid-Career Advancement Grant MCA: Cultivating a Leader in Statistics and Data Science Education through Strategic Partnerships - If you are a mid-career faculty member, ask me about this funding opportunity.

EMERGING THEMES

Topic	Description	Keywords
Programming Languages & Software (Basics)	Focuses on the programming languages (R, Python) and software tools (RStudio, Jupyter, SQL) used to manage, analyze, and visualize data. Students learn to write code, automate analysis, and use key packages like pandas, numpy, and sklearn.	R, Python, pandas, numpy, sklearn, RStudio, Jupyter, SQL
Data Visualization	Covers principles and tools for creating effective data visualizations using libraries like ggplot. Emphasizes design, interpretation, and storytelling with visual data.	visualization, graph, plot, ggplot, heatmap, histogram, scatterplot, boxplot, barplot, facet
Data Wrangling	Involves cleaning, transforming, reshaping, and managing datasets using techniques such as filtering, joining, and tidying data. Ensures that raw data is converted into a format suitable for analysis.	wrangling, cleaning, tidy, reshape, filter, mutate, join, pivot, data cleaning, data transformation, data manipulation
Statistical Foundations	Provides the theoretical grounding in statistics and probability necessary for data analysis, including descriptive statistics, hypothesis testing, regression, distributions, confidence intervals, and ANOVA.	probability, descriptive statistics, distribution, hypothesis, inference, regression, ANOVA, p-value, confidence interval
Machine Learning	Introduces the concepts of building predictive models using algorithms for classification, clustering, and dimensionality reduction. Covers supervised and unsupervised learning techniques, including k-NN, decision trees, random forest, and PCA.	classification, clustering, prediction, k-NN, decision tree, random forest, PCA, unsupervised learning, supervised learning
Ethics and Communication	Explores ethical considerations in data science including fairness, bias, privacy, and the social impact of data-driven decisions. Also focuses on effective communication, storytelling, and visualization to convey findings to varied audiences.	ethics, bias, misrepresentation, communication, storytelling, data ethics, reproducibility, responsibility, social justice
Data Science Workflow	Emphasizes the end-to-end process of data science: defining problems, exploratory data analysis (EDA), feature engineering, building and evaluating models, and communicating results.	data science process, pipeline, EDA, exploratory, project, analysis, workflow, feature engineering, model evaluation, cross-validation
Advanced or Extended Topics	Covers computational tools such as APIs, web-scraping, text mining, version control (GitHub), and interactive dashboards and graphics (Shiny, Tableau) to automate, document, and deploy data science work. Includes advanced/extended areas of deep learning, cloud computing, and geospatial analysis.	APIs, web-scraping, GitHub, version control, Shiny, Tableau, deep learning, neural network, GPU, high throughput, cloud computing, text mining, sentiment analysis, geospatial, GIS, interactive graphics

Initial Results

- Machine learning and ethics and communication topics were not emphasized across most courses.
- Topics covered in courses entitled" Introductory Data Science" varied.
- Common topics include basic programming in R/Python, data visualization, statistical foundations, and data science workflow.

Next Steps

- Further specify skills and topics addressed, detailing both the depth and breadth of their coverage.
- Identify common patterns in topic sequencing and examine the pedagogical rationale that informs these choices.
- Assess how topic coverage varies in relation to key factors such as departmental affiliation and prerequisite requirements.

TOPIC COVERAGE

Course Title	Programming Languages & Software (Basics)	Data Visualization	Statistical Foundations	Machine Learning	Ethics and Communication	Data Science Workflow	Advanced or Extended Topics
Introduction to Data Science	1	1	1	0	0	1	1
Introduction to Data Science	1	1	1	1	1	1	1
Introduction to Data Science	1	1	1	1	0	1	1
Introduction to Data Science	1	1	1	1	0	1	1
Introduction to Data Science	1	1	1	1	1	1	1
Introduction to Data Science	1	1	1	1	1	1	1
Introduction to Data Science	1	1	1	1	1	1	1
Introduction to Data Science	1	1	1	0	0	1	1
Introduction to Data Science	1	1	1	1	1	1	1
Introduction to Data Science	1	1	1	1	1	1	1
Introduction to Data Science	1	1	1	0	0	1	0
Introduction to Data Science	1	1	1	1	0	1	0
Introduction to Data Science	1	1	0	0	0	1	1
Computing across the Disciplines	1	1	1	0	0	1	0
Data Analysis	1	1	1	1	1	1	1
Data Science	1	1	1	0	1	1	1
DS and Predictive Analytics (Health)	1	1	1	1	0	1	1
Data Science for Social Impact	0	0	0	0	1	0	1
Data Science I	1	1	1	1	0	1	0
Elements of Data Science	1	1	1	1	1	1	1
Foundations of Data Science	1	1	1	1	1	1	1
Foundations of Data Science	1	1	1	1	1	1	0
Information Exposition (CS)	1	1	1	0	1	1	0
Introduction to Data Analysis	1	1	1	1	1	1	1
Introduction to R	1	1	1	0	0	1	1
Introduction to Stat Methods	0	1	1	1	1	1	0
Mathematical Computing	1	0	1	0	0	0	0
Principles and Tech of Data Science	1	1	1	1	1	1	1
Visualization and Communication in DS	1	1	1	0	1	1	1

Connect with us!

This material is based upon work supported by the National Science Foundation under Grant No. DUE-2423026-. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

jennifer.broatch@asu.edu