**NASPAA CHARETTE:** What do we need to create for our students to become data experts for social good?

**Charette date:** October 29th, 2021

**Charrette Chair:** R. Patrick Bixler, Assistant Professor, LBJ School of Public Affairs, RGK Center for Philanthropy and Community Service

**Charrette Support:** Alyssa Studer and Ethan Tenison, RGK Center for Public Affairs

**Charrette participants:**

**Context for this charette:**

Data has transformed modern life and put new pressures on public‐serving organizations. Students in Schools of Public Affairs and Administration are at the forefront of developing skills to lead government and community organizations in this endeavor. This charette was a brainstorming exercise that helped us conceptualize what core competencies would best prepare students to become data specialists and public data scientists. All participant input is available publicly on [Jamboard](https://jamboard.google.com/d/1P12A99ro5p6BImlhqUcyJNGE1wUKglidux_VEz8CYrQ/edit?usp=sharing), and it was used to synthesize the following document. Below are key insights, suggestions, and barriers identified over the course of the charrette.

**Data training vs. Data scientists for Public Affairs students**

From a programmatic perspective, the coursework necessary to ensure data competency for all public affairs students versus creating data scientist for the public good differs. The charette conversation frequently oscillated from one pathway to the other. Data competencies important for all (most?) public affairs students should include:

* dealing with messy data,
* problem identification,
* communicating with stakeholders, and
* data ethics.

Of our self-selected sample of programs, there were many examples of existing courses that cover these topics. However, there was general agreement that training data scientists would require a sequence of coursework that developed advanced skills statistical modeling and programming languages (such as R and Python).

***Core data competencies and data scientist competencies***

|  |  |  |
| --- | --- | --- |
| Competency | Data Specialist | Data Scientist |
| Dealing with Messy Data | ✅ | ✅ |
| Problem Identification | ✅ | ✅ |
| Data Management | ✅ | ✅ |
| Data Ethics | ✅ | ✅ |
| Basic Statistics | ✅ | ✅ |
| Data Storytelling | ✅ | ✅ |
| Data Visualization\* |  | ✅ |
| Econometrics\* |  | ✅ |
| Exploratory Data Analysis\* |  | ✅ |
| Causal Inference\* |  | ✅ |
| Machine Learning\* |  | ✅ |
| Reproducible Reporting\* |  | ✅ |
| Geographic Information Systems \* |  | ✅ |

\*Using R or Python

**R vs. Python for Policy Analysis**

A brief debate among charette participants tracks larger debates in academia. The debate should not impede instruction or academic programming. Above all, data science bilingualism should be promoted to maximize opportunities for students. This is easier than ever because most data science platforms, such as Rstudio and Jupyter, now allow you to use both languages in the same coding environment. Conceptually, these data science platforms are like drills, and R and Python are like drill bits. Students can change the drill bits depending on what their projects require. These languages develop independently from each other and are both awesome in different ways.

Still, the overwhelming response from participants of this charette was that it’s best to choose one language for foundational courses. For example, it can be confusing for students when there are first semester statistical methods courses in both Python and R. These skills are transferable, so it does not make sense to learn about linear regression in both languages.

When deciding which language to choose for foundational courses it is important to consider each language strengths and weakness. The preference for most participants in the charette was R because of its ease of use for non-programmers and powerful reporting tools, which lend themselves to policy settings.

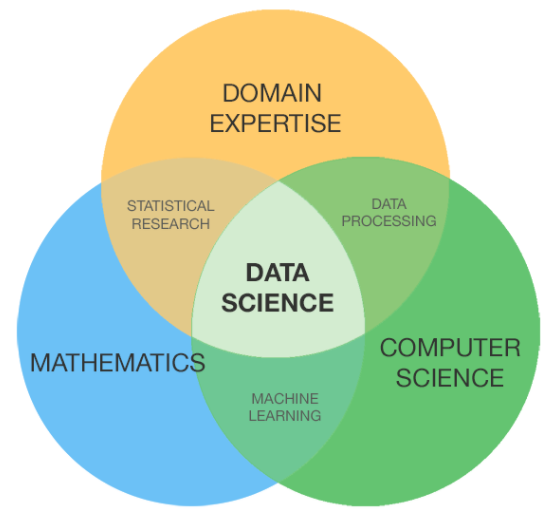
**R Strengths:**

* R has become the world’s largest repository of statistical knowledge with reference implementations for thousands, if not tens of thousands, of algorithms that have been vetted by experts.
* R has a very low barrier to entry for doing exploratory analysis, and converting that work into a great report, data visualization, or dashboard.
* It’s easier to learn when students have zero coding background.
* R has a great community of supportive data scientists from diverse backgrounds. For example, [R-Ladies](https://rladies.org/about-us/) is a global organization dedicated to promoting gender diversity in the R Community.
* R users tend to come from a much more diverse range of domain expertise (ecology, economics, psychology, bioinformatics, policy analysis, etc.)

**Python Strengths:**

* Python is a great general programming language, with many libraries dedicated to data science.
* It’s easier to learn for students who already know programming
* Python is the go-to language in commercial settings that include computer scientists, who generally already know Python.
* For some organizations, Python is easier to deploy, integrate and scale than R, because Python tooling already exists within the organization.

**How can schools of public policy better prepare students to apply data science for the social and public good?**

** [[1]](#footnote-1)**

Public Policy schools focus on domain knowledge, and access to real data, is both a challenge and barrier. Programs that focus solely on mathematics and computer science fail to develop student skills in critical thinking and communication. Data scientists in the policy sector must be able to identify problems, evaluate metrics, and communicate their findings to non-technical audiences and stakeholders. These skills are already taught in Public Affairs programs in their core curriculum and can be altered to integrate more data literacy.

**Service and experiential learning**

The role and importance of applied experiences were emphasized by charette participants. Programs can promote experiential learning by funding student internships/fellowships and by hosting hack-a-thons or micro-internships. Experiential learning is also a creative opportunity to incorporate an interdisciplinary approach by involving researchers and experts from other colleges and universities. Challenges around involving community organizations such as organizing courses for both content and service, and forming, managing, and cultivating client relationships were noted.

**What are potential barriers to implementing this new course/series?**

Overall, the biggest barrier we gathered from the charrette was the distinction between larger schools and smaller schools and the ability to integrate data competencies, much less a data scientist sequence, into curriculum. To address this challenge, ideas were generated around consortium models with shared expertise and possibly coursework. Other ideas and examples were shared related to data/programming upskilling for existing faculty.

One of the biggest barriers for students getting involved is the initial learning curve, which can be a significant barrier for students that have never programmed before. One institution offered a data bootcamp. Taking place before classes start, a boot camp could provide that extra push to help students gain confidence with data science tools (R/[tidyverse](https://www.tidyverse.org/); Python).

**Example Course Series for Data Science**

1. Data Science Bootcamp in R (Taken before their regular courses begin)
   * Basics of R and Rstudio
   * Students learn basic data exploration, visualization, and data cleaning using the Tidyverse
2. Public Management (and Data Management)
   * Focus on the use of real messy data
   * Problem Identification
   * Learn how agencies use large-scale data to inform policy design, increase stakeholder engagement, and improve service delivery
3. Research Methods
   * With integration into a programming language
   * Real world examples using census data (tidycensus package)
   * Including data ethics
4. Data Visualization and Data Storytelling
   * Experiential learning with real-world clients
   * Focus on reproducible reporting
5. Geographic Information Systems
   * Focus on community level data
   * GIS dashboards using Tableau or R Shiny
6. Machine Learning
   * Goals: Identify machine learning techniques to solve real world problems
   * Using *An Introduction to Statistical Learning: with Applications in R*
7. Advanced Research Methods (electives)
   * Conducted in either R or Python
   * Casual Inference
   * Network Analysis
   * Data Life Cycle
   * Econometrics
   * Program Evaluation
   * Bayesian Statistics
8. Capstone Project

**Example Introduction to Data Science for Policy Analysis Course**

The initial framing of the charette was to think about how to integrate data science into a core public management course. We had limited success keeping our conversation focused to that scope. Some participants didn’t believe a six-course sequence was enough to train data scientist. There was, however, overwhelming consensus on the need for and importance of exposing public affairs students in data core competencies and training data scientists for the public good.

1. https://medium.com/@balajiciet/how-am-i-evolving-as-a-data-science-practitioner-5d105f38525f [↑](#footnote-ref-1)