LPPS 5740: Data Science for Public Leaders (3 Credit Hours)
Fall 2021: August 24 - December 7, 2021
Thursdays from 2-4:30pm
Final due December 10, 2021

Instructors:
Ben Castleman
Newton and Rita Meyers Associate Professor
in the Economics of Education
Email: castleman@virginia.edu
Office: Ruffner Hall 262
Office Hours: TBD

Brian Heseung Kim
Education Policy PhD student, Batten MPP alum
Adjunct Faculty (Batten)
Email: bhk5fs@virginia.edu
Office: Ruffner Hall 279
Office Hours: TBD

Course Assistant: Emily Arizaga (eea4m@virginia.edu)

Course Description

Data science strategies are now ubiquitous in the private sector and affect our daily lives across a host of contexts: through the entertainment we consume, the money we invest, the ads we are exposed to, and even the people we connect with. But it has only been in the last few years that these same strategies have made their way into the public sector. With the proliferation of data science across academic disciplines and the accelerating availability of large-scale administrative datasets, leaders of public sector organizations are increasingly encountering analyses that leverage data science frameworks to inform and shape policy.

We already have a wide array of salient applications to sectors like health care (Beam & Kohane, 2018), criminal justice (Goel et al., 2016; Kleinberg et al., 2018), education (Lakkaraju et al., 2015; Mduma et al., 2019), and social services (Cuccaro-Alamin et al., 2017). Yet, importantly, there is an equally broad array of circumstances where the application of data science frameworks can calcify historical racial biases (Obermeyer et al., 2019), facilitate reliance on faulty statistics (Mullainathan & Obermeyer, 2017), impede transparency of policy to stakeholders (Cohen et al., 2014), and threaten the privacy of individuals (Eubanks, 2018). To the extent that data science promises great opportunity for the future of policy, it is also accompanied by commensurate risk.

In this course, we will support future public leaders at Batten and the broader University to develop a critical awareness of how data science methods are currently being applied in the public sector and how they will continue to shape policy going forward.

Learning Objectives
Through class participation and completion of the assignments, our intention is that students will be able to:

- Demonstrate an informed understanding of data science methods relevant to public policy—in particular, methods related to policy problem identification; data acquisition; policy targeting; and policy evaluation.
- Interpret, critique, and gauge the relevance of policy research leveraging data science methods
- Think critically about the salient ethical, logistical, and operational concerns for specific applications of data science
- Identify future opportunities where the application of data science methods could inform policy decisions and action
- Facilitate the comprehension of data science methodologies and findings for non-technical audiences and stakeholders

Course Expectations
A Mutual Commitment
We view this course as a mutual commitment to learning. For our part, we make several commitments to our students. We will strive to: provide you with the highest quality course structure possible; facilitate thought-provoking class discussions; involve high-level guests in the class; design assignments that are both challenging and enriching; and support each of you to have a highly-fulfilling and rewarding course experience. In return, we ask several commitments of each of you. We ask that you engage fully and invest wholeheartedly in all aspects of the course, from preparing for class discussions to completing assignments. This is particularly important when we have guest speakers. They will have taken time out of very busy schedules to be with us, and the best way we can signal our appreciation for their time is by having the highest level of preparation for and engagement during their class sessions. We ask also that you reach out for support if you are struggling with any aspect of the class. And we finally ask that you respect, and try your best to learn from, the diverse opinions and perspectives of your classmates – the issues this class will engage with are ones that support many reasonable positions and conclusions. With this mutual commitment in place, we are confident that we will all learn considerably from each other over the course of the semester.

Teaching this class for the first time
This is the first time we have taught this course or more generally a course on data science for public leaders. We want to be transparent about pedagogic decisions we have made with this in mind:

- While we both think deeply about data science applications to public policy and engage actively on data science applications in our own research, neither of us is a professional data scientist. Our value-add will hopefully be in supporting you to build critical awareness of how these methods can be applied in public policy contexts. Perhaps more

often than in a typical class, our initial answer may be "I don't know" to more technical questions we receive. We are committed to promptly finding the answers to your questions, however.
- Our applied examples lean more heavily on data science applications in education and more specifically our own work than we would normally prefer. We chose this approach because we know this work most deeply and so can speak with greater confidence about the data science insights that emerge from these case studies and that are applicable to other policy fields.
- We have worked hard to compile readings, assignments, and visual aids that we think will best serve your learning. Because this is our first time teaching the class, we'll also benefit greatly from your feedback about whether we're striking the right balance between intuition and technical detail, the depth we're giving to different topics, and whether our pacing could be adjusted. We're grateful in advance for the support you'll provide in co-creating a strong learning experience!

Pre-requisites
Students should have completed the Batten Research Methods and Data Analysis first-year course sequence or its equivalent:

- **Data Science**:
    - STAT 6021 Linear Models for Data Science
    - DS 6001 Practice and Application of Data Science
- **Education:**
    - EDLF 5330 Quantitative Methods and Data Analysis I
    - EDLF 7420 Quantitative Methods II: General Linear Models
- **Stats:**
    - STAT 6120 Linear Models
    - STAT 6190 Introduction to Mathematical Statistics

Interested students are welcome to contact either instructor to gauge whether their prior coursework provides a sufficient foundation for the course. We intend to introduce fundamental concepts of data science and machine learning at a level that builds on the econometric training Batten students receive and that builds strong intuitive understanding. The goal of the course is not to equip students with the technical skills they need to directly apply data science methods themselves, but rather to equip students to be critical consumers of these strategies.

Laptop/Tablet Policy
The success of this course depends on students being fully engaged in course discussions. In our experience—both as students and now as professors—laptops and tablets can create substantial distractions. Even when we had the best intentions as students to stay focused in class and participate in discussions, we inevitably found ourselves online or reading through emails rather than actively listening to our peers and adding our own insights. New rigorous research supports that students learn less in courses that allow laptops and tablets than in courses that do not.

For this reason, we do not allow students to use laptops, tablets, or phones on a regular basis during class, except for students who have an accommodation to use technology to maximize their learning and retention. If you have such an accommodation, please reach out to Ben or Brian to discuss this. We will provide advance notice if there are class sessions when it would be helpful to bring a laptop or tablet with you.

To reassure you, the class assessments will ask you to synthesize high-level concepts from class and from the readings—we will not ask you to provide detailed facts that you might otherwise want to take notes on during class. If you would still like to take notes during class on paper/pencil, you're of course more than welcome to do so, and we can provide notebooks and pencils to anyone for whom buying one would create a financial hardship.

Course Texts

All readings are described in the course outline below. We have posted articles on Collab unless we provide a web link below. There is no textbook for the course.

Course Outline

Dates	**Topics**	**Readings**
	Unit 1: Course Introduction & Foundation-Building	
Class 1 Aug. 30	**Introduction to Key Concepts and Frameworks for Data Science and Public Policy** **Guest Speakers:** Batten MPP graduates in policy	● Grimmer, J. (2014). We are all social scientists now: How big data, machine learning, and causal inference work together. PS: *Political Science and Politics*. ● Engler, A. (2020). What all policy analysts need to know about data science. *Brookings.* ● Davenport, T.H. & Patil, D.J. (2012). Data scientist: The sexiest job of the 21st century. *Harvard Business Review.*
Class 2 Sept. 6	**The Policy Analysis Process Mapped to Data Science**	● Arnold, K., Fleming, S., DeAnda, M., Castleman, B. & Lynk Wartman, K. (2009). The summer flood: The invisible gap among low-income students. *The NEA Higher Education Journal.* ● Castleman, B. & Page, L. (2013). A trickle or a torment? Understanding the extent of summer "melt" among college intending high school graduates. *Social Science Quarterly.* ● Castleman, B. & Page, L. (2015). Summer nudging: Can personalized text messages and peer mentor outreach increase college going among low-income high school graduates? *Journal of Economic*

		Behavior & Organization. **(Read intro only)** ● Bird, K., Castleman, B., Denning, J., Goodman, J., Lamberton, C., Ochs Rosinger, K. (2021). Nudging at scale: Experimental evidence from FAFSA completion campaigns. *Journal of Economic Behavior & Organization.* **(Read intro only)** ● Kleinburg, J., Ludwig, J. & Mullainathan, S. (2016). A guide to solving social problems with machine learning. *Harvard Business Review.*
Class 3 Sept. 13	**Best Practices in Applied Data Science** **Guest Speaker:** TBD	● Barshay, J. & Aslanian, S. (2019). Colleges are using big data to track students in an effort to boost graduation rates, but it comes at a cost. *The Hechinger Report.*
Unit 2: Data Acquisition		
Class 4 Sept. 20	**Data Science & Data Acquisition: Web Scraping 101**	***Unit 1 reflection due by the start of class*** ● Tiecke, T. & Gros, A. (2016). Connecting the world with better maps. *Facebook Engineering.* ● Tiecke et. al. (2017). Mapping the world population one building at a time. *Cornell University.* ● Anglin, K.L. (2019). Gather-narrow-extract: A framework for studying local policy variation using web-scraping and natural language processing. *Journal of Research on Educational Effectiveness 12(4).* ● Chouldechova, A. & Roth, A. (2018). The frontiers of fairness in machine learning. *Cornell University.*
Class 5 Sept. 27	**Harnessing Insights from Audio-Visual Data: Machine Vision and Hearing** **Guest Speaker:** TBD	● 3Brown1Blue. (2017, October 5). *But what is a neural network? Deep learning, chapter 1.* [**Video**]. Youtube. ● St-Aubin, P., Saunier, N. & Miranda-Moreno, L. (2015). Large-scale automated proactive road safety analysis using video data. *Transportation Research Part C: Emerging Technologies*, 58 (Part B), p.363-379. ● Adukia et. al. (2021). What we teach about race and gender: Representation in images and text of children's books. *University of Chicago, Becker Friedman Institute for Economics.*

Class 6 Oct. 4	**Harnessing Insights from Text Data: Introduction to Text Mining and Natural Language Processing** ***Note:*** *In-class activity will require technology; please bring your laptop, tablet, or other device to class.*	● Fesler et. al. (2019). Text as data methods for education research. *Journal of Research on Educational Effectiveness,* 12(4), p.707-727. ● Quinn et. al. (2009). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science.*
Oct. 11	*UVA Reading Day*	
Class 7 Oct. 18	**Mid-Semester Check-in & Project Workshop Session**	*No readings for this class.*
Unit 3: Targeting		
Class 8 Oct. 25	**Why (and how) Do We Target Policies and Programs?**	***Unit 2 reflection due by the start of of class*** ● Baker, M. & Fink, S. (2020). At the top of the COVID-19 Curve, how do hospitals decide who gets treatment? *The New York TImes.* ● Weissman, J. (2020). What the Shake Shack debacle tells us about the flawed small-business bailout. *Slate.* ● Moving to opportunity. *National Bureau of Economic Research.* ● Sullivan, Z., Castleman, B.L. & Bettinger, E. (2019). College advising at a national scale: Experimental evidence from the CollegePoint initiative. *EdWorkingPapers.*
Class 9 Nov. 1	**The Promise, Overpromise, and Peril of Prediction in Public Policy** **Guest Speaker:** TBD	● Bringing transparency to predictive analytics in higher education (URL to be included) ● Ekowo, M. & Palmer, I. (2016). The promise and peril of predictive analytics in higher education. *New America, Education Policy.*
Class 10 Nov. 8	**Predictive Modeling in Other Policy Contexts** ***Note:*** *In-class activity will require technology; please bring your laptop, tablet, or other device to class.*	**All students read:** ● Amatriain, X. & Basilico, J. (2012). Netflix recommendations: Beyond the 5 stars (Part 1). *The Netflix Tech Blog.* ● Amatriain, X. & Basilico, J. (2012). Netflix recommendations: Beyond the 5 stars (Part 2). *The Netflix Tech Blog.*

		Before class we will assign small groups to one of the following: ● Stevenson, M. & Doleac, J. (2019). Algorithm risk assessment in the hands of humans. *Social Science Research Network.* ● Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science.* ● Seck, H. (2019). SOCOM is using AI to search for the ideal MARSOC raider. *Military.com.* ● Llorente, A., Garcia-Herranz, M.Cebrian, M. & Moro, E. (2015). Social media fingerprints of unemployment. *PLoS ONE, 10(5).*
Class 11 Nov. 15	**Recommender Systems in Public Policy** **Guest speaker:** TBD	● Tugend, A. (2010). Too many choices: A problem that can paralyze. *The New York Times.* ● Personalized pathways to successful community college transfer: Leveraging machine learning strategies to customize transfer guidance and support. *Nudge[4] Solutions Lab.*
Class 12 Nov. 22	**Recommender Systems Continued**	● Kim, B., Bird, K. & Castleman, B. (2021). Crossing the finish line but losing the race? Socioeconomic inequalities in the labor market trajectories of community college graduates. *Nudge4 Solutions Lab.*
Class 13 Nov. 29	**Causal Forests** **Group presentations**	***Unit 3 reflection due by the start of class*** *No readings for this class.*
Class 14 Dec. 6	**Group presentations & course wrap-up**	***Group projects due by the start of class*** *No readings for this class.*
Dec. 10	***Final project due on Canvas***	

Assessments

1. **Class Participation (25 percent)**. The learning experience in the class will be strongest with active engagement and perspective-sharing from every student. We will consider several inputs into the class participation grade:
    a. Class attendance. If you cannot be in class, please reach out to us ahead of time to let us know.
    b. Active participation in and contribution to class discussions.

c. Active engagement with guest speakers. This includes both preparation (e.g., formulating and submitting questions ahead of time) and engaging in any participatory activities and discussion they elicit during their lectures.
d. Weekly asynchronous responses to reflection questions we post alongside course readings. The purpose of these asynchronous responses is to provide the opportunity for us to learn from and engage with students who may be more reticent to engage as actively in class.

2. **Unit reflections (20 percent).** After each unit we will ask you to complete brief reflections (500 words) that convey:
   a. Key insights and frameworks that you take away from the unit.
   b. Remaining questions or areas of uncertainty you have. We will make a point of addressing these topics in future classes or with you individually.
   c. Potential applications of the data science methods to a policy area of interest for you. We will ask you to identify this policy area at the start of the semester and to consider applications for this policy area throughout the term, largely as preparation for the individual final project (see #4 below).

   These are due at the beginning of the following class as noted in the schedule above.

3. **Group presentation on current application of data science in public policy (20 percent).** In small groups (3-4), students will find and closely analyze a recent paper or report that applies data science methods to public policy and prepare a 15-20 minute presentation that synthesizes its details for the class. The presentation should be supported by a deck with a maximum of 10 slides and cover:
   a. The policy question or policy application that the report focuses on
   b. Primary data used in the application and source(s) of that data. Be sure to describe any particular strengths or weaknesses of these data (reliability, generalizability, etc.)
   c. How the data science methods were applied in that policy context. What was done, and what were the justifications for the various design decisions made?
   d. A translational and intuitive explanation of the methods the report uses and of the report's main conclusions
   e. Cautionary notes or concerns about how those methods were applied
   f. Opportunities to apply additional data science methods that the report or paper did not apply
   g. Broader insights and implications for either the policy or the methods applied

   Benchmarks and timeline for the group presentation:
   - **Grading Rubric:** Group Presentation
   - End of class 3 (September 9): Identify group members (3-4 max) and the likely policy focus for your presentation (e.g., health, finance, education, etc.).
   - End of class 5 (September 23): Select the article or report your group's presentation will focus on. Please email your article to both of us and we will either give the OK to proceed or set up a meeting to discuss any reservations we have about the selection and brainstorm potential alternatives.

- Beginning of class 7 (October 7): Read your article (pending instructors' approval) and generate initial reactions, impressions, concerns, etc. We will randomly select a few groups to describe the policy area their presentation is focused on, the article they have identified, and any initial insights or questions that have emerged from your review of the article. This intended to be an informal, workshop-oriented presentation.
- End of class 10 (October 28): Send completed draft slides to Ben and Brian. We will provide you feedback within a week.
- Beginning of classes 13 and 14 (November 18 and December 2): Groups present in a randomly selected order (15-20 mins each) with time for additional audience questions after.

Additional expectations for the group presentation
In addition to assessing each group's presentation against the rubric, we will ask each student to submit a confidential report to us on completion describing:
- The specific contributions they individually made to the group project
- Whether they view these contributions as equitable to, more, or less than other group members' contributions
- Whether there were group dynamics that meaningfully affected the final product, e.g. one group member carrying the lion's share of the work, or a group member whose contributions were very limited

If all group members describe relatively equitable contributions to the final product, the group grade and the individual grade will likely be the same for this assignment. Based on the feedback we get from each group, however, we may adjust the individual grades we ultimately assign for the group project. For instance, if an individual's self-reflection and the description of other group members indicates s/he carried most of the work, we may increase the student's grade relative to the overall group grade. Conversely, if an individual's self-reflection and the description of other group members indicates their contributions were limited, we may reduce the student's grade relative to the overall group grade.

4. **Final Project: Proposal for Data Science Applications to a Specific Policy Topic (35 percent).** Drawing on the methods we have studied throughout the semester, each student will complete a final proposal (2,000 words) for a state agency leader on how they could incorporate data science methods into the agency's work. The proposal will follow the format of the policy analysis process we have used as a guidepost throughout the semester—in other words, the proposal will identify:
   a. A problem the agency faces or may face in the future.
   b. Prior attempts by the agency (or other agencies) to address the problem and a summary of evidence on the efficacy of these attempts
   c. How the agency could use data acquisition methods to support investigation of this question, alongside concrete explanations for how these approaches would improve upon the current "business-as-usual"

d. How the agency could use prediction methods to identify which constituents are most affected by this problem and/or recommendation methods to identify resources or pathways that could improve constituents' outcomes, alongside concrete explanations for how these approaches would improve upon the current "business-as-usual"
e. Limitations and cautions for responsible application of data science methods to this problem—and to the agency's work more broadly. For example, how might these approaches be unsuccessful? How might these approaches result in inequitable outcomes for constituents of the agency? What are the implications of these approaches for transparency in agency activities?

Benchmarks and timeline for the final proposal

- **Grading Rubric**: Final Project
- End of class 5 (September 23): Select policy focus and potential agency collaborator (just hypothetical, we do not expect you to engage with agencies)
- Beginning of class 7 (October 7): Come to class with your current best thinking about what policy focus, potential agency collaborator, and problem statement (an issue within your policy area that could potentially benefit from applications of data science) will be for the final proposal. Be ready to share with peers to get feedback.
- End of class 10 (October 28): Send Brian and Ben a draft problem statement, relevant to your selected (hypothetical) agency collaborator. This draft problem statement should include background research/policy context (no more than 1-2 paragraphs) to help us understand the **current** state of the policy problem.
- End of class 13 (November 18): Develop a bulleted outline of how you think data science methods could be applied to this problem statement (addressing points C and D) above) alongside potential issues (addressing point E above). Send to Ben and Brian for feedback.
- December 10: Final proposals due.

Assignment grading and regrade policy:
We (Ben and Brian) will divvy up and alternate grading of all submitted work so that you all have each of us review approximately half of your work. Any assignment that one of us assigns a grade lower than B+ will be reviewed by the other. If you would like to dispute the grade you receive on an assignment, you may write to Ben a maximum 100-word explanation of why you think you deserve a different grade, and Ben will review your assignment. He will review these requests and assign the grade we think the assignment merits—note that this could result in a grader that is higher *or* lower than what the initial grader provided.

University Email Policy
Students are expected to activate and then check their official UVA email addresses on a frequent and consistent basis to remain informed of University communications, as certain communications may be time sensitive. Students who fail to check their email on a regular basis are responsible for any resulting consequences.

Because the course uses Canvas as a course website, we also expect that you are checking the Canvas website multiple times a week for any course notifications.

University of Virginia Honor System

All work should be pledged in the spirit of the Honor System of the University of Virginia. The instructor will indicate which assignments and activities are to be done individually and which permit collaboration. The following pledge should be written out at the end of all quizzes, examinations, individual assignments and papers: "I pledge that I have neither given nor received help on this examination (quiz, assignment, etc.)". The pledge must be signed by the student. For more information please visit http://www.virginia.edu/honor/.

Special Needs

It is the policy of the University of Virginia to accommodate students with disabilities in accordance with federal and state laws. Any student with a disability who needs accommodation (e.g., in arrangements for seating, extended time for examinations, or note-taking, etc.), should contact the Learning Needs and Evaluation Center (LNEC) and provide them with appropriate medical or psychological documentation of his/her condition. Once accommodations are approved, it is the student's responsibility to follow up with the instructor about logistics and implementation of accommodations.

If students have difficulty accessing any part of the course materials or activities for this class, they should contact the instructor immediately. Accommodations for test taking should be arranged at least 14 business days in advance of the date of the test(s). Students with disabilities are encouraged to contact the LNEC*:* 434-243-5180/Voice, 434-465-6579/Video Phone, 434-243-5188/Fax. For more information, visit the U.Va. Special Needs website at http://www.virginia.edu/studenthealth/lnec.html.

On Well-Being

If you are feeling overwhelmed, stressed, isolated, or otherwise unwell, there are many resources at UVa ready and wanting to support you. The Student Health Center offers Counseling and Psychological Services (CAPS) for its students, and they are an incredible resource to be aware of as a student here. They offer free consultations and a number of free sessions as appropriate. Mental and emotional well-being is a wide spectrum, and any reason is a good reason to seek support if you feel you need it (grad school is *hard*). Call 434-243-5150 (or 434-972-7004 for after hours and weekend crisis assistance) or visit their website to get started and schedule an appointment. If you prefer to speak anonymously and confidentially over the phone, call Madison House's HELP Line at any hour of any day: 434-295-8255.

If you or someone you know is struggling with gender, sexual, or domestic violence, there are many specialized community and University of Virginia resources available. The Office of the Dean of Students, Sexual Assault Resource Agency (SARA), Shelter for Help in Emergency (SHE), and UVA Women's Center are all fantastic and eager to help.