Recipes for Connector Courses From the Early-Career Board Kitchen

Aleksandrina Goeva¹ Peyton Jones² Sara Stoudt^{3,4} Ana Trisovic⁵

¹Macosko Lab, Stanley Center for Psychiatric Research, Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, Massachusetts, United States of America,

²Harvard College, Cambridge, Massachusetts, United States of America,

³Berkeley Institute for Data Science, University of California Berkeley, Berkeley, California, United States of America,

⁴Statistical & Data Sciences Program, Smith College, Northampton, Massachusetts, United States of America,

⁵Institute for Quantitative Social Science, Harvard University, Cambridge, Massachusetts, United States of America

Published on: Jun 07, 2021

DOI: <https://doi.org/10.1162/99608f92.f7c60746>

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

Whether students begin their data science education through Data 8 or an alternative program, we are especially energized by the concept of connector courses as a way to get experience answering questions in a data-driven domain. We see the potential for these types of courses to provide flexibility in bringing data science to other institutions and increased engagement for students and faculty alike.

For those institutions where interest in data science is growing, but there is no formal structure for it yet, aiming for a handful of connector courses might identify the faculty working on data-intensive research across the curriculum, who may be allies in a larger data science initiative. When first run as pilots, these connector courses would not benefit from the same assumed knowledge of a class like Data 8 right away, but assessing the heterogeneity of background knowledge of students interested in these classes might help define what needs to go in a foundation course moving forward.

Just as the connector course idea broadens the pool of faculty members who may teach a data-related course, these interdisciplinary courses will likely engage a broader pool of students. By showing students all of the different ways data can help us ask and answer questions of interest, data-driven thinking will be solidified as a critical practice within the liberal arts.

Not every program has to offer a major in data science, but the goal may be to make sure every student knows how to approach a problem using data in the context that most interests them. The ideas in “Interleaving Computational and Inferential Thinking: Data Science for Undergraduates at Berkeley” (Adhikari et al., this issue) certainly provide a path forward toward this ideal, and we started to think of the article as providing a recipe for undergraduate data science at UC Berkeley (perhaps cooking is particularly on the brain in these pandemic times). What is data science if not a new fusion cuisine of statistics, computer science, and domain knowledge? We decided to contribute new flavors, cooking up connector courses that we would be excited to take ourselves. Enjoy our tasting menu focused on data science methods and analysis considerations.

Bon appetit!

Data Scientists, Like Chefs, Need Quality Ingredients

Ethical Experimental Design and Survey Sampling

recipe from:

Ingredients:

- 1 sense of right and wrong
- fair trade beverage of choice

Instructions:

1. Consider the number of guests RSVP'd along with their dietary restrictions.
2. Balance the budget with nutritional value to finalize the menu.
3. Ensure the sweetness is "right" and serve.

How do we combat the garbage in, garbage out phenomenon? Just as chefs may ‘pick’ their own ‘ingredients’ and check that they are ethically sourced, data scientists can benefit from getting involved in the data collection process. A connector course could ensure data analysts don’t just treat data as a given; think of a typical experimental design and survey sampling class with a twist.

This class would bring together the principles of data collection and the ethical consequences of that data collection. What questions can we not answer with an experiment? What can go wrong in a survey? What problems can happen downstream as a result of data collection challenges? Students get the chance to evaluate the costs, both literal and social, of their data work alongside traditional tips for sample size and power calculations.

Squeezing Nutrients From Data and Making It Digestible Is Not Always Easy

Extracting Meaning From High Dimensional Data

recipe from:

Ingredients:

- nutrient-rich data
- quality knives dimension reducers

Instructions:

- Wash your raw ingredients.
- Peel, seed, chop, juice, sieve, boil until digestible.
- Assess nutritional value of the final product.

Chef Peyton

As our abilities to store and process large quantities of information have improved in recent decades, fields of all disciplines are leveraging the power of data to solve problems and make discoveries. However, complex data sets can be overwhelming. It can be hard to know where to start, much less discern relevant patterns.

This connector course would aim to help students apply and expand on the foundational data analysis skills learned from Data 8 by making these data sets more approachable through an emphasis on dimensionality reduction and visualization of patterns.

There are many considerations and approaches when reducing the data dimensions. What approaches learned in Data 8 still serve us well? As we start to simplify, how can we make sure we're not over- or under-simplifying? How applicable are particular methods across a variety of domains? Students will learn to wield dimensionality reduction techniques to identify the patterns, connect them to the context, and leave the noise on the chopping block. Visualizations will help reveal the many facets that the data has to offer. Different contexts for motivating examples abound, from linguistics to genomics.

Data Science Is More Like Cooking Than Baking

Choose Your Own Open Data Adventure

recipe from:

Chef Sara

Ingredients:

- 1 publicly available dataset
- handful of questions (to taste)
- Data 8 special sauce

Instructions:

1. Chop your data into bite-sized pieces
2. Choose your favorite cooking method: bakes, deep fry, pan-sear, grill - you name it!

There is no one question to ask of a data set and no one way to answer a question using data. Classic templates exist as a starting point, including those learned in Data 8, but there is room to customize for a given situation or personalize for an individual preference. A connector course centered on open data in general, or even narrowed on a particular open data set like the U.S. Census, American Community Survey, or cities' open data portals, could give students the opportunity to define their own questions of interest and take steps toward answering those questions.

This class would help bridge the gap for students with newfound knowledge and skills from Data 8 who still have logistical questions about how to get started with their own data investigation. Where do I find data? How do I access it once I find it? What questions that I'm interested in can even be answered with this data? Given a 'test kitchen' environment, students can explore the many flavors of data and develop their own analysis tastes.

The Proof Is in the Pudding

Accuracy of Prediction Methods: the What and How	
<p><u>Ingredients:</u></p> <ul style="list-style-type: none"> - prediction tasks (a bunch) - truth (ground) - measuring cups 	<p>recipe from:</p> <p>Chef Aleks</p>
<p><u>Instructions:</u></p> <ol style="list-style-type: none"> 1. Chop data into homogeneous train, validate, and test pieces. 2. Season the method only using the train and validate bits. 3. Experiment with cutting the data in different shapes and see the effect. 	<p><u>Note:</u> if using a new data supplier, make a small test batch before serving!</p>

Or, as the original saying goes: ‘the proof of the pudding is in the eating.’ Just as you have to try out food in order to know if it is good, you have to deploy a prediction method on real-world, unseen-in-training data to know how useful it is. In Data 8, students focus on reasoning with data and are introduced to computation as a tool for inquiry, and to prediction methods as a means of making evidence-based decisions. Since these decisions are often impactful, we need a way of knowing how accurate our predictions would be when deployed in the real world.

Classical statistical wisdom tells us to split the available data for a prediction task into three parts, each representative of the whole data, namely—training, validation, and test sets. But how large, relative to each other, should these three data sets be? Is 80%/10%/10% right? Might 60%/15%/25% be better? What metric should we use to compare the method’s predictions to the ground truth? And how can we check if each subsampled piece is representing the whole data set well? Could it be advantageous to purposefully deplete a subset of the data in the training set, to test how the method performs under ‘stress’ or on unseen scenarios? What about applying our trained, validated, and tested prediction method to new data, in the wild, where we don’t have ground truth to compare the method’s predictions to? What should we do when there is a substantial data set shift, that is, the data the method is trained on is notably different from the data we are hoping to apply it to?

This connector course will go through the dangers of double-dipping (i.e., reporting prediction accuracy on the same data used to train the model), the importance of quantifying data set shift between training/validation/test data and the data that the method will be deployed on in the wild, and strategies for collecting additional ground truth data when the shift is substantial.

Data Scientists Often Find Themselves in Pressure Cooker Environments

Data Science Principles to Uphold Under Pressure

recipe from:

Ingredients:

- $\frac{1}{2}$ data you requested (the rest is on backorder)
- $\frac{1}{3}$ time you originally budgeted
- double the ~~steaks~~ stakes

Instructions:

1. Cut original recipe in half.
2. Accept that the outcome will be a little rare.
3. Use a meat thermometer to make sure no one will get salmonella.

Chef Sara

Data science often happens under constraints. When tackling a problem we assess our resources including time, money, and data availability. We might need to adjust our goals with these limitations in mind. A case studies-based connector course could give students a taste of doing high-stakes data analysis under a time crunch to prepare them for fast-paced data science environments.

This class would be structured around case studies from a variety of data-related fields. The common theme would be some kind of constraint that had to be accommodated in the analysis (think the modern equivalent of the O-ring debacle, where extrapolation occurred under some political pressure). What choices lead to ‘good enough’ results? What aspects of an analysis should not be sacrificed at any cost? What would our Data 8 professors do in a pinch? Students would learn apprentice style as they worked through examples of others’ decision-making processes.

Recipes Provide Good Models for Data Science Documentation

Automation, Preservation, and Dissemination in Data Science

recipe from:

Ingredients:

- all used files from soup to nuts, especially data and code
- helper appliances (optional)

Instructions:

1. Consider using ~~plastic~~ containers for data and code.
2. Share your data, workflow, and results, but preserve them too!

Chef Ana

Many research results are taken with an extra grain of salt if they could not be independently reproduced. Some of the reasons for the observed lack of reproducibility are unavailable data, obsolete code, or an overall lack of transparency about what happened behind the scenes. A connector course that would help students understand the benefits of computational reproducibility and help them achieve it in practice would produce well-equipped data scientists conducting reliable and reusable research.

This connector course would address the following three questions: (1) What are the ethics behind data and code sharing? It is important to understand that data sharing is not always possible, but also that in some cases, sensitive data can be shared using novel tools. (2) How do we achieve analysis automation and reproducibility? There is a surge of existing infrastructures like data and software repositories, tools like containers and workflow engines, and functionalities within a native programming language to help the researchers develop portable and reusable code. (3) How do we comprehensively document resources and enable their reuse? Adequate repository organization and documentation are critical for reuse, and they should be emphasized in the classroom. Lastly, open research resources need to be accompanied by permissive licenses, and the students should learn how to choose and use them. As a result, the students will become proficient curators of their work, competently sharing their work for easy review or reuse.

Results Do Not Stand Alone; Their Presentation Matters

The Craft of Written and Visual Arguments

recipe from:

Ingredients:

- charcuterie board of others' work (fresh)
- dash of rhetoric and perception theory

Instructions:

1. Group work by flavor profile and reduce each until thick (with meaning).
2. Use reductions as glaze for your own argument.

Chef Sara

Just as photos and menu descriptions play a large role in convincing diners to order a dish, figures and reports help convince others that what we did is defensible, useful in context, and contributes to the field of data science. As a complement to the current Writing Data Stories connector course, another communication-style course could focus on argument-making over narrative-building. Another possible distinction is to have a connector course focused on communication to broader audiences and one focused on more formal communication to our peers.

How do we distill a data-driven argument from a body of work and hone that argument in both written and visual form? Students can learn from others' approaches by reading data-driven arguments and evaluating data visualizations, determining what is and is not convincing. They can build their own set of best practices for writing and data visualization, compare them to classical theories of rhetoric and perception, and implement them to strengthen their own data-driven arguments.

Disclosure Statement

Aleksandrina Goeva, Peyton Jones, Sara Stoudt, and Ana Trisovic have no financial or non-financial disclosures to share for this article.

©2021 Aleksandrina Goeva, Peyton Jones, Sara Stoudt, and Ana Trisovic. This article is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](#), except where otherwise indicated with respect to particular material included in the article.