Revised: June 5, 2017

Capstone Project Guidelines

Data Science

Summary

- It's far better to pick a relatively straightforward, 'boring' project that you can deliver on, than to pick a very complex, shiny idea that you'll get stuck in.
- Focus your project around a realistic client and problem.
- Use your mentor as a resource, a sounding board and as a filter.
- Use the student and mentor community for feedback at any time.

TABLE OF CONTENTS

Revised: June 5, 2017

How to pick a good capstone project

When you are working as a data scientist in industry, you're typically required to deliver solutions that are 'good enough' in a limited amount of time. Unlike an academic research project, you may not have the luxury of time to find the best and optimal solution. There's an engineering, product, sales or marketing team waiting on the results of your analysis. It's very important to have a sense of the various tradeoffs between different approaches, and pick one that's well-suited to the problem and resources you have.

How do you pick a course project that reflects this mindset? Here are some general guidelines:

- **Is it a real problem that someone would care about?** Ideally, the result of your project could be something you apply directly at work, or you can add it to your portfolio and brag about it.
- **Is there real data available?** You want to work on a project that has real-world data, not a toy data set or a data set that's used only for academic teaching purposes. The data sets pointed to by the course material typically meet this requirement.
- **Is the data easy enough to acquire and clean?** While you want real-world data, you don't want to spend 100 hours cleaning and wrangling with it. Pick a data set that's relatively clean. As a rule of thumb, if you have to spend more than a week acquiring and cleaning your data, you may want to reconsider.

Basically, to paraphrase Einstein, *keep it as simple as possible, but no simpler* :) Your mentor is there to help you at this stage to decide if your project idea meets these guidelines.

Pick your initial project ideas

Think of up to 3 project ideas that meet the guidelines presented above and that you're excited to work on. You can explore datasets from Quandl, US Government Open Data, UCI Machine Learning Repository, Kaggle competitions or anywhere else you like. There's a great email list called Data is Plural, that lists new and interesting data sets that have been released. Take your mentor's help as needed. Once you have picked them, your goal is to narrow them down to the ONE idea that you'll be working on.

- Write a short blurb for each of your ideas. The blurb should, at high level, describe the problem and the data you'll be using to solve it. At this point, there's no need to talk about specific methods and techniques.

- Post your idea (title and blurb) on the community and solicit feedback from both the mentors and other students.

Pick one idea to work on based on the feedback you get. Discuss the idea with your mentor to ensure that they're on board.

Note: The goal of a project is NOT to do something novel (you're not writing a PhD thesis), it's to demonstrate your competence as a data scientist. It's perfectly acceptable to work on a data set that's been worked on before, and even answer a question that's been answered before, as long as the work is your own.

A word of caution: Kaggle competitions

It's perfectly fine to use a data set from Kaggle in your project. However, many Kaggle competitions are about taking a data set that's already clean and optimized for the specific problem, and tuning a machine learning algorithm for the highest accuracy (or similar metric). While that's an important skill for a data scientist, it's not all. In real-world scenarios, you'll be the person who has to collect, wrangle and clean that data. If a Kaggle competition you're considering falls into that category, here are a couple of ways you could still use the data set:

- Could you use that data set to solve a different problem than the one asked in the competition?
- Could you combine it with other data sets to solve the same problem asked in the competition or a different one?

Basically, we'd like your Capstone project to demonstrate your competence with the entire data science process, not just one aspect of it.

That being said, **your mentor has the final word** on whether a Kaggle competition is appropriate for a Capstone Project or not. Typically, we've found that recruiting competitions sponsored by top companies (e.g. Airbnb) meet the criteria for a Capstone Project.

Another word of caution: Using proprietary data

Many of our students work on Capstone projects that involve proprietary data, from their employer, for example. This is perfectly fine. **We don't require that you share the raw data** with Springboard or your mentor. However, there are a few things you'll need to pay attention to:

1. **Ensure you have the right permissions:** Your mentor is here to guide you through your project. They can only do that effectively if they can look at your code, summarized results, charts etc, even if they don't have access to the data. In

addition, Springboard still requires that you turn in a project report and a slide deck based on your analysis. If your employer (or the people who are providing you the raw data) are not comfortable with (1) or (2) you may need to rethink your project topic. Please check with their legal team to see if you need approval in writing in the form of a legal contract or a Non-Disclosure Agreement (NDA).

2. **Start data collection early:** Even if you have the requisite permissions, please make sure to start the data collection process early, and have a realistic idea of how soon you can actually get the data. Many companies have elaborate processes around data access and extraction (with good reasons around security and privacy), so sometimes, students have become stuck waiting around for their project data to become available.

Project stages

We have broken down the Capstone Project into several stages.

Project proposal

Once you've picked your final capstone project idea, you will write a proposal. The project proposal is a short (1-2 page) document that answers the following questions:

1. What is the problem you want to solve?
2. Who is your client and why do they care about this problem? In other words, what will your client DO or DECIDE based on your analysis that they wouldn't have otherwise?
3. What data are you going to use for this? How will you acquire this data?
4. In brief, outline your approach to solving this problem (knowing that this might change later).
5. What are your deliverables? Typically, this would include code, along with a paper and/or a slide deck.

The proposal will be part of a github repository for your project. All code and further documentation you write will be added to this repository.

Once your mentor has approved your proposal, please share the github repository URL on the community and ask for feedback.

At this point, the project proposal will be considered approved and ready.

Data collection

The first step in your capstone project is to actually collect your data. In some cases, it can be as simple as downloading a data set in a zip file. In other cases, it can require extracting data using a publicly available API or scraping a web site. We urge you to work with your mentor closely to ensure that the data collection process is not too onerous for a capstone project.

Data wrangling

At the end of the material on data wrangling, you'll apply some of the data wrangling techniques you have learned to your capstone data set and create a short document (1-2 pages) in your github describing the data wrangling steps that you undertook to clean your capstone project data set.

- What kind of cleaning steps did you perform?
- How did you deal with missing values, if any?
- Were there outliers, and how did you decide to handle them?

This document will eventually become part of your milestone report.

Exploratory data analysis

After you have obtained the data set for your Capstone project, cleaned and wrangled it into a form that's ready for analysis, you will perform preliminary exploration of the data. This exploratory data analysis (EDA) uses a combination of inferential statistics and data visualization to find interesting trends and identify significant features in the data set. For example:

- Are there variables that are particularly significant in terms of explaining the answer to your project question?
- Are there strong correlations between pairs of independent variables, or between an independent and a dependent variable?

Your findings from this phase will be summarized in another short (1-2 page) document. These findings will not only become part of your milestone report, but will also

Milestone report

You and your mentor will work together to determine at least one milestone for your project. Your milestone will be reached when you produce an early draft of your final project paper. This is a slightly longer (3-5 page) draft that should have the following:

- An introduction to the problem, including a description of the potential client and their motivation
- A deeper dive into the data set:
    - What important fields and information does the data set have?
    - What are its limitations i.e. what are some questions that you cannot answer with this data set?
    - What kind of cleaning and wrangling did you need to do?
- Any preliminary exploration you've performed and your initial findings.
- Based on these findings, what approach are you going to take? How has your approach changed from what you initially proposed, if applicable?

A lot of the work for your milestone should already be done as part of data wrangling and EDA, however, the milestone is an opportunity for you to practice your data storytelling skills. We encourage you and your mentor to plan multiple milestones if possible. This reflects the fact that in industry, you typically have several iterations with your client during a real project.

Add your code and milestone report to the github repository. As before, once your mentor has approved your milestone document, please share the github repository URL on the community and ask the community for feedback.

Final deliverables

The final deliverables for your project, such as your code and report, are part of your portfolio, which means that you'll be sharing them with employers. We also require students to present their project at the data science office hour. This will give you practice talking about your project to potential employers, as well as provide you with a video that demonstrates your technical and communication skills.

When you and your mentor agree on a stopping point for the project, you should have the following deliverables ready *before asking your Student Advisor* to start the completion process.

- **Code** for your project, well-documented on github.
- **Final paper** in your github repository explaining the problem, your approach and your findings. Include ideas for further research, as well as *up to 3 concrete recommendations* for your client on how to use your findings. Please title the file clearly e.g. Capstone_Final_Report.[doc|pdf|...]
- **Slide deck** for your project in your github repository. As a data scientist in a company, you'll be frequently called upon to produce and present slide decks. You

can use any standard presentation tools (Powerpoint, Keynote, Google Slides etc) to create your deck. Please title the file clearly e.g. Capstone_Final_Slides.[ppt|pdf|...]
- **Share your project** in the one of the ways suggested in the following section, namely by presenting in an office hour, creating an online video or writing a blog post.

Guidelines for your code and slide deck are outlined below.

Presenting/Sharing your project

Sharing your project is an important step towards building your brand as a data scientist! We highly encourage you to share your project with the world, and we'd love to support you in that effort. Here are some options you may consider:

- **Present in an Office Hour:** if you are interested, your student advisor will help you with scheduling your presentation as part of the data science office hour.
- **Create an online video of your presentation:** If Office Hours are unavailable or inconvenient, create a screenshare of you presenting your project, put it up online on Youtube or Vimeo, and share the link with your student advisor.
- **Write a blog post:** Blogs are a great way to generate awareness of your work and your brand as a data scientist. Here are Springboard's guidelines for blog posts on our official blog. If you'd like your blog post to be considered for the Springboard blog, please write a draft according to the guidelines and share it with your student advisor. If not, you're welcome to post on your own blog, Medium, LinkedIn or other platforms. Please share the link of your post with your student advisor.

For bonus points, actually present or send your report/slide deck to your designated client and let us know what kind of response you received from them. Completely optional, but our previous students have found this very rewarding!

Remember that both your paper and slides should be targeted to the client that you picked in your proposal.

What makes a good presentation?

A good project presentation lasts about 20 minutes, with 10 minutes for Q&A. Here are some suggested guidelines for the structure of the presentation:

- A clear explanation of the problem, the client and the motivation for this project in (1-2 slides)
- The formulation of the project as a data science problem, and a description of the data set (1-2 slides)
- Data wrangling steps (1-2 slides)
- Exploratory analysis and interesting findings therein (3-4 slides)

- In-depth analysis (e.g. machine learning) at a high level e.g. which method did you choose and why? (1-2 slides)
- Results of your analysis (1-2 slides)
- Recommendations for your client based on your results (1-2 slides)
- Practical considerations and suggestions for improvement (1 slide)

Please look at the sample presentations provided at the end of this document for inspiration.

Guidelines for a good portfolio

Your portfolio consists of all of your data science projects, including the code and documentation, usually in your github account. Typically, a data scientist who looks at your portfolio wants to see evidence of both your technical and your communication skills. It is your responsibility to ensure that your portfolio is clear and easy to navigate. Here are a few tips that'll help your portfolio stand out:

- Every project should ideally be in a separate repository that is clearly named.
- For each project:
    - Make sure you have a README page that provides an executive summary of the project i.e. summarizes the problem, approach and final results.
    - The README should also include a list of the important files that the reader should look at. The files themselves should be as clearly named and organized as possible.
    - Clean up your code and document it, so that your approach and methodology are clear to any technical reader. You do not need to document every line of your code, but have comments or text explaining important decision points and why you chose them.
    - Include any other documents that you have created e.g. a report, slide deck etc in the same repository as the code. Make sure the README points them out to the reader.
- Ensure that your portfolio is not cluttered with "junk" i.e. repositories that are incomplete, irrelevant or undocumented.

Overall, try to put yourself in the shoes of an experienced data scientist who has a limited amount of time to look at your portfolio. How can you ensure that you make it easy for them to get a good idea of your skills and abilities? Your mentor and the community should also be able to provide good feedback on your portfolio, so please use them as resources.

Project Evaluation

For Springboard to consider your workshop complete, and issue a certificate of completion, your mentor needs to approve your final project submission per the rubric described below. In case the project is not approved, please discuss the feedback from your mentor and resubmit in case improvements are necessary for the approval. Your Student Advisor will not be able to process your workshop completion until your project is approved by your mentor!

Capstone Project Rubric for Data Science

We use the following rubric for evaluating final Capstone Projects. Please take a good look at it and make sure you discuss with your mentor to agree on success criteria.

View the Capstone Project Rubric here

The first tab is the rubric itself, and the next two tabs are sample projects that have been graded using the rubric. The rubric consists of several evaluation criteria, each graded on a 3-point scale: *Below Expectations*, *Meets Expectations*, or *Exceeds Expectations*. Your project is considered Complete if you get a *Meets Expectations* or *Exceeds Expectations* on **ALL** of the criteria.

Because your mentor decides your grade, it's vital that you work with your mentor during the entire project to understand and agree on what the bar is for each criterion, and incorporate their feedback during the intermediate stages.

For now, your mentor will be using a copy of this sheet to grade your project and returning the graded sheet by email along with your assessment.

Appendix: Sample Projects

Here are some sample projects by our students. We'll add more exemplary projects to this section periodically.

Foundations of Data Science

- Robert Chen: "Eat, Rate, Love" – A Proposal for Modifying Yelp's Rating Systems
    - Report
    - Presentation
    - Github
- Stan Siranovich: Predicting Protein Tertiary Structure
    - Report

    - Presentation
    - Github
- Arti Annaswamy: Predicting Opening Weekend Movie Returns
    - Report
    - Presentation
    - Github

Data Science Intensive

- Charles Franzen: Assessing risk for US work visas
    - Report
    - Presentation
    - Github
- Sayan de Sarkar: Predicting customer subscriptions to bank term deposits
    - Report
    - Presentation
    - Github
- Naresh Vempala: Neighborhood crime in the city of Toronto
    - Report
    - Github
- Ravikiran Durbha: Predicting SQL execution
    - Report
    - Presentation
    - Github