# CPSC 5305: Introduction to Data Science

# The Data Science Project

The project is a useful and inevitable part of any broad-topic Data Science class, and most specialized data science courses too. For this class the project has two main goals

- Demonstrate your ability to execute on all phases of the data science methodology, from defining a problem precisely to implementing a solution
- Gain more in-depth experience in some aspect of data science concepts and technologies. This might be in the area of systems or algorithms, it might be applying some some concept from the class in more depth, or doing something completely new.

In other words, the project is open-ended and is intended to let you explore or deepen your understanding of some area of data science.

Here are some example final projects

- From the General Assembly Intensive class in Data Science: https://gallery.generalassemb.ly/DS
- From a UC Berkeley Intro to Data Science course: https://bcourses.berkeley.edu/courses/1377158/pages/project-suggestions

## Project Teams

You will form teams of three for the project. Teams of two will be OK only if the number of people in the class is not divisible by three. The course schedule will have deadlines as to when project groups must be formed.

## Selection of Topic

Your topic selection will be an important decision obviously, and you probably want to approach it in one of two ways

1. There is some data problem you are aware of (i.e. it has arisen at work) and you want to explore it further. In that case, since you're starting with an ill-defined problem, you will spend more effort on the earlier stages of the methodolgy, and less on details of the technical solution
2. There is some known defined data problem or data set that you think is interesting. You will pursue a solution to the problem, and find challenges in the technical piece of the solution rather than in defining the problem in the first place

No matter which route you go, your project must do two things

1. Embody the data science methodology from beginning to end
2. Be challenging in some aspect, which you will determine. Probably it will be one of these things

a. *The problem is ill-defined* so the earlier stages of the process – defining the problem, finding adequate data, initial proof of concept – are more complicated and risky.
b. *The algorithm is not obvious* – you have to experiment with different algorithms and parameters to improve performance.
c. *Scale is an issue* – your data is of a size where running the whole data set on a regular laptop is infeasible, so you will have to use some other technology, or reduce the number of data cases, or reduce the number of variables, in the data.
d. *Uncharted territory* – you are experimenting either with some algorithm, or some processing or storage technology, or some problem that was not addressed in depth in the course, so you will have to do some additional research.

Most important is for you to enjoy the experience! This is not supposed to be a high-pressure class or a high-pressure project, so just settle on something interesting, do a beautiful job, and don't worry. If you find yourselves worrying about or not understanding the project, let me know! We can spend class-time discussion on the topic, and I'm always happy to talk to you in person about it.

## Deliverables

- **Problem Statement** – this will be a written document, and specifics will be sketched out below, but more detail will be provided. Exact due date will be on the course schedule, but probably the 4th week of the course
- **Progress Report** – again a written report, due probably around the 7th week of the course
- **Final Report and Presentation** – the final presentation will be at the final exam time; the final report will be a notebook combining results and writeup, due at the time of the presentation

## Problem Statement

This deliverable is to make sure you have a well-defined problem, and have given some thought to what data you will need. You formulate the question to be answered, and identify data sources relevant to the problem. You may or may not have made any decisions yet about what methods you will use. You outline next steps, and call out risks to successful completion. The whole document should be approximately 2 pages, and should contain the following sections

1. Header: project title
2. Team members
3. Problem statement. Be sure to state it as a "business problem" without going into the solution. What question(s) are you answering, and why are they important?
4. Data sources. What data sources are you planning to use, and why are they sufficient to answer the question(s). Will there be a large amount of data preprocessing and cleaning required before the data sets are useful?[1] It's OK not to have finalized the data sources, but if you are still looking, you need to outline a plan for how and when you will have your data sources available.

[1] For example, I advised a project where the student wanted to process race results for a series of motorcycle races. The race results were published (only) as PDF files, and getting from PDF to a usable data set was a major undertaking, one that we didn't realize until we actually tried to do it. Catch these challenges early!

5. Algorithms / solution technologies. What algorithms and technologies do you think will be required to process this data in order to answer your questions. It's fine to be speculative here. For example, I would not expect you to have settled on which machine learning algorithm you will use!
6. Risks. What are the unknown or semi-known factors that might threaten your project in terms of time or resources? Remember, "risk" really just means "unknown" in this context
7. Challenge: what is the aspect of the project that will be your "differential challenge" as explained above
8. Citations: cite any related projects you came across during your initial investigation

### *Use of External Work*

In putting together your problem statement it is expected that you will do research on the web and choose a topic that has been solved before – or something close to it. It is fine to do that, but (a) be sure to cite any related work you have consulted, and (b) for the problem statement and your throughout your project, it is not OK to use somebody else's work (e.g. Python code or notebook). In deciding how much you can legitimately draw from the work of others, remember that you can draw inspiration from other work, but you cannot copy a solution. For example, you are fine reading a report from a related project, both to get ideas, and even to understand the author's approach to solving the problem. But once you have done so, put the external work away, don't look at again, do your own solution, and be sure to cite the external work.

## Progress Report

- This report will be delivered roughly the 7th week.
- At this point you should have access to your data sources and have done some EDA, so you are more confident that you can get to good answer(s) to your project question(s)
    - Likewise if there is a systems aspect to your problem – for example if you were using Spark to handle large scale – you should at this point have settled on your technology platform and are convined that it is sufficient for you to solve the problem at scale
- This report will build off your problem statement. In fact, it will be your edited problem statement with a section at the end documenting the current state of your project and what needs to be done before the final presentation

### *Format for the Progress Report*

- Begin by editing your Problem Statement to reflect the true state of your project right now.
- The "Algorithms and Solution Techniques" section should be better developed now. I expect at this point you have a well-defined predictive model and have run at least some algorithm against the model. Describe your independent variables (X) and your dependent variable (y) in detail. Describe any preliminary results (e.g. accuracy or error from preliminary runs).
- Update the "Risks" and "Challenges" section.
    - The update to the "Risks" section is especially important, because now Risks represents any uncertainties or impediments to having a good solution and presentation done by the presentation day
- Add any Citations to the document as per the discussion in the Problem Statement requirements

- Directly before the Citations section, add a short section (half to single page is fine) "Plan for Completion"

Your "Plan for Completion" section should

- Briefly reiterate where you are now (we have our data set, we have run a preliminary model, etc.) Be honest about this! The point is not to penalize you, just to give us all a realistic assessment of what needs to happen in the next three weeks
- List the steps you need to take between now and the final presentation, with the estimated number of person/hours required for each. Obviously the last one will be "prepare presentation"
- List any outstanding issues or concerns remaining, that were not covered in the Risks section. It is fine if there are no outstanding issues that you did not already cover – if so, omit this section

## Final Report and Presentation

- The presentation
    - Plan on approximately 10 minutes for your presentation
    - Assume your audience is non-technical, so emphasize the problem motivation and an intuitive explanation of your solution. It's fine and desirable to include technical results, just don't make it entirely technical and leave out the motivation
    - Your presentation should follow the data science methodology, taking us from problem statement to solutions
        - It should also include some retrospective material on what you learned from the project, what you would do different next time, etc.
- Your report will be a document (e.g. Word). It can closely mirror your presentation – it too takes us from problem statement to results, but in the notebook you tell the story primarily through numbers and graphics
- There are lots of good examples of project presentations through the links above, and also here: https://bcourses.berkeley.edu/courses/1377158/pages/final-projects