# CS506 Projects: Key Components and Expectations

## Spring 2018

A successful CS506 project must have the following components:

1. **Formulating one or more questions**: Before you dive into complex analyses, it’s important to lay out your goals. What are you trying to find out? What would constitute a successful outcome for your project?
   1. By the time the project is complete, you should be able to specify a major question that connects all you results, as well as several specific questions that contributed to answering the big question
2. **Obtaining data**: Part of the project consists in obtaining and/or combining several interesting data sources to help you answer your questions.
   1. Getting one or more relevant data sets (could involve crawling, using relevant API’s, obtaining via partners)
   2. Cleaning and preparing data (could involve matching records to merge data sets, marking missing values, or otherwise getting the data into analyzable form)
3. **Data Analysis**
   1. Exploratory analysis
      1. Get to know your data: visualization, basic descriptive statistics
   2. Main data analysis: Based on the questions you aim to answer, you should:
      1. Identify and run relevant algorithms
         1. At least one should be from class, though more advanced algorithms are encouraged
      2. Compute diagnostic quantities to help with interpretation, depending on the type of analysis (these could include confidence intervals, p-values, and cross-validation error estimates)
      3. Present experimental results corresponding to each of the questions
4. **Interpreting your results**
   1. How do the results help you answer the question you started with?
      1. Are they meaningful?
      2. Are they sufficient to answer your questions?
         1. Your results only provide incomplete answers. Regardless, you should be able to explain whether and how the results solve the problem you started with
   2. Visualize results
5. **Formulation of revised or new questions**, and follow-up analysis (see “revised final report” under Milestones, below): Based on your first analysis, you should perform a revised analysis, involving
   1. New questions
      1. Did you get a meaningful answer to your original questions?
      2. Is there a better, modified version of your initial question that now makes more sense?
      3. What new questions does your initial analysis raise?
   2. If possible, new data
      1. What other data sources could you use to confirm your analysis?
      2. Are there new/other data sources that could help?
6. **Final conclusions**
   1. How have your efforts contributed to answering the original question? Be nuanced.
   2. If time allowed, what further analyses would you perform? On what other data sources?

The following items are not expected for all projects, but are encouraged:

* Using data that were scraped, or derived and integrated from multiple sources.
* Using large data sources ( > 5GB)
* Providing exceptional insightful or sophisticated visualizations (maps, for example)
* Implementing or using an approach from a research paper or advanced textbook. It is important though that the team knows how the approach works both, in theory and practice.

# Milestones

Once you know your team and whether or not you’ll be working with a partner, you’ll need to make a plan for the project and then carry it out. We’ve asked for documents from you along the way from you to help keep you on track and to provide you with feedback.

Exact due dates for the milestones will be maintained here:

<https://tinyurl.com/cs506-spring2018>

**Proposal** (mid-late February)

* Topic, and questions to be addressed (½ page)
  + The description of the general topic can be short -- a paragraph should suffice. If you feel you need to provide extra information or context, it is ok (but not necessary!) to add a separate appendix with relevant background.
  + The questions should be reasonably clear. It is good to have several questions, some more complicated or ambitious than others. It is good to have at least one fairly straightforward question (e.g., “what factors are correlated with X?”) to get you started.
  + For Spark!-partnered project, at least one of the specific questions should be formulated by you (the students). We encourage you to discuss with and get feedback from the partners, but please be considerate of their time.
* Ideas for specific data sets (¼ to ½ page)
  + A brief description of the initial data sets you intend to use.
  + If possible, please provide details on the size of the data sets (number of records, types of features, and how they are accessible.
* See this discussion guide for questions you should be asking as a team as you write your proposal.

**Interim report** (mid-late March): 1-2 pages

* By this point you should have
  + at least one clearly defined question and hypothesis you plan to investigate
  + At least one data source chosen and retrieved
  + Initial exploratory analysis (histograms, etc)

**Final report** (early April): 5 pages

* This report should explain how you have achieved key components 1 through 4 (that is, your first analysis and interpretation should be complete).
* The report should also detail the follow-up analyses you plan to do as well as any additional data sources you plan to use.
* The report should contain the following sections
  + Problem Formulation
  + Data Analysis
  + Methodologies
  + Experimental Results
  + Conclusions so far
  + Intended follow-up
* In each section, explain the reasoning behind your choices (why a particular data set was chosen, how different problems were addressed, why specific algorithms were chosen e.t.c.)

**Revised final report** (late April), 10%

* This report should address all key project components, and follow the structure of the final report

**Poster** (last week of class)

* The poster should be structured similarly to the report (Abstract, Data Collection/Analysis, Methods, Experimental Results, Conclusions). Part of the poster grade will be for the short presentation you give at the poster session.

**Poster session**

* At the poster session, each team will present their poster briefly to the instructors. You will have a chance to walk around and see others posters, and hear about what they did.

Some example of posters and reports from previous semesters are available here:

* <https://drive.google.com/drive/folders/1NT1FaHA1RUX-XszyfQ1RveHzUFb28U2V?usp=sharing>