1) Loan Default Modeling [this is my current favorite]

Data come from Lending Club (available on their website) , a peer to peer lending site. The combined data have 1.8 million rows and 151 columns (though some columns have a lot of missing values). The main challenge is what to choose as the target: simply modeling default as a binary variables throws away information, so it is better to model something like expected return to take into account *partial* repayments. This requires choosing a discount rate for future cash flows, and this rate should depend on the risk – which we don't know until *after* the analysis. So the way to proceed seems to be to first choose a reasonable discount rate (maybe based only on FICO-score), then compute the expected return, and then – if I have time – iterate through the process again, each time adjusting the discount rate based on the predicted risk (measured e.g., as the probability of a negative return) from the previous step. I might have to read up on loan default modeling a little bit.
Otherwise, this project seems pretty straightforward.

Pros: large data set; not too much data cleaning necessary; relevant to where I might work (while I don't have any industry preferences, where I live the biggest employer of data scientists by far is a bank)

Cons: Might be harder to tell an engaging story since the subject matter is a little dry.

2) Bank Telemarketing analysis: Which contacted customers will subscribe to term deposit?

Data come from UCI repository. 45k rows and 18 columns , which is okay but not as good as the Lending Club data.
The dependent variable is dichotomous, so this is the classification task. It seems even more straightforward than idea 1).

Pros: Like idea 1), it is relevant to where I might work and doesn't seem to require too much data cleaning; I have a slight preference for classification tasks, because most of my experience so far is with regression.

Cons: Both less observations and less attributes then data for idea 1), so it will require less complex models; presumably this data set has been analyzed much more often than the Lending Club data, because it is available on the UCI repository.

3) Car Safety (this is my favorite for capstone 2, because it is more involved)

Analysis of driver death rates to determine the influence of the car model. The Insurance Institute for Highway Safety(IIHS) already does this analysis, but only does an okay job (e.g., a lot of relevant information is not used). As a result, the confidence intervals are so high that the results only give a very rough picture of the safety of specific car models.

The initial reason I consider this project is because it seems ideal for a Bayesian hierarchical model analogous to what I'm currently working on (estimating the sale price of hotels clustered in different groups, e.g., Metro area). Similarly, cars are clustered into make, model, generation, etc. If – like the IIHS – we estimate a separate intercept for each generation of each model, we are overfitting the data, especially for groups with a relatively small number of observations. It is better to assume that, say, the safety scores of different generations of a particular model come from the same distribution, and are thus more likely to be similar to each other than to the safety scores of a different model. Hierarchical models take this into account by employing adaptive regularization (shrinking the coefficient estimates towards the group mean, with the amount of shrinkage depending on the number of observations we have for the particular group).

In addition, I also want to try out different machine learning methods to model this, and then do a comparison which performs more accurately. Since this is more of an inference problem than a prediction problem, it is not the typical machine learning problem. But even though more complex machine learning models will not give an easily interpretable coefficient estimate for each car model, it should still be possible to simulate predictions for new data, which can be used to isolate the effect of the variable of interest.

The data come from the National Highway Traffic Safety Administration (and also includes the VIN number, which can be used to get additional vehicle data through an API). It is also available on Google Bigtable.

Pros: The topic is interesting to me, and it seems to be a good project for telling an interesting story.

Cons: More of an inference than a prediction problem, so not ideal for machine learning; data seems to need more cleaning, so better suited for capstone 2.