Proposal Review Checklist

# Define Business Goal, Explore Data, Define Problem & Metric

* What exactly is the business problem to be solved?
* Is the data science solution formulated appropriately to solve this business problem?
* What business entity does an instance/example correspond to?
* Is the problem a supervised or unsupervised problem?
* If supervised, is a *target* variable defined? Think about the values it can take.
* Are the attributes defined precisely? Think about the values they can take.
* For supervised problems: will modeling this target variable improve the stated business problem? An important subproblem? If the latter, is the rest of the business problem addressed?
* Does framing the problem in terms of expected value help to structure the subtasks that need to be solved?
* If unsupervised, is there an “exploratory data analysis” path well defined?
* Is there a plan for domain-knowledge validation?
  + Will domain experts or stakeholders want to vet the model before deployment?
  + Will the model be in a form they can understand?
* Against what baselines will the results be compared?
  + Why do these make sense in the context of the actual problem to be solved?
  + Is there a plan to evaluate the baseline methods objectively as well?

# Implement Model

## Data Preparation

* Will it be practical to get values for attributes and create feature vectors, and put them into a single table?
* If not, is an alternative data format defined clearly and precisely? Is this considered in the later stages of the project?
* How exactly will the values for the target variable be acquired? Are there any costs involved? If so, are the costs considered in the proposal?
* Are the data being drawn from the similar population to which the model will be applied? If there are discrepancies, are the selection biases noted clearly? Is there a plan for how to compensate for them?

## Modeling

* Is the choice of model appropriate for the choice of target variable?
* Does the model/modeling technique meet the other requirements of the task? (Generalization performance, comprehensibility, speed of learning, speed of application, amount of data required, type of data, missing values)
* Is the choice of modeling technique compatible with prior knowledge of problem (e.g., is a linear model being proposed for a nonlinear problem)?
* Should various models be tried and compared (in evaluation)?
* For clustering, is there a similarity metric defined? Does it make sense for the business problem?

## Evaluation

* Is the evaluation setup and metric appropriate for the business task?
  + Are business costs and benefits considered?
  + For classification, how is a classification threshold chosen?
  + Are probability estimates used directly?
  + Is ranking more appropriate (e.g., for a fixed budget)?
  + For regression, how will you evaluate the quality of numeric predictions? Why is this the right way in the context of the problem?
* Does the evaluation use holdout data? Cross-validation is one technique.
* For clustering, how will the clustering be understood?

## Deployment

* Will deployment as planned actually (best) address the stated business problem?
* If the project expense must be justified to stakeholders, what is the plan to measure the final (deployed) business impact?

# Take Actions, Monitor Performance