```
1) Extraction
    For each species in database
        Query points
        Save points into separate CSV w/ unique number(?)
        Counted the points and to lists of species w/ 1,2,3/4,5+ points
2) Chunking
    Divide species with 5+ points into chunks suitable for HPC time limit (=?? species)

3) Loop over 1,2,3&4 ,5+ categories
    1 point → Gentry box (75,000 km2 box centered on point
    2 point → a) bounding box
             b) Gentry box around each point
             Took the larger range area
    3&4 Points → a) bounding box
              b) Gentry box around each point
             Take the larger range area
    5+ points
        Run MaxEnt
            Simplest to just continue to give it the raster and all observation points
            use 100% of points (currently only uses 80%)
            Climate & Spatial
            Increase regularization
            Remove threshhold responses

        if successful
            save Continouous raster surface + metadata w/ threshold criteria
                (for MaxEnt Rawoutput prediction, Results file)
        if failed
            use 3&4 point algorithm

4) Gather outputs
    (LEA / UTM) - what is best?
    From Continuous raster surface + metadata w/ threshold criteria
    Derive
        Raster presence/absence binary
        Shapefile of the range (polygon)
        Range area calculation

Major changes form John’s Scripts
--------------------------------------------
SQL Extraction
queuing & chunking may have changed
MaxEnt options have changed
```

Outputs have been significantly reduced (eliminate excess calculations)

Nice to do

Sampling effort model
Background subsetting
Multiple runs - choose the best

From Cory’s summary doc:

```
\title{Maxent Modeling Ideas for BIEN3}
\author{Cory Merow}
\maketitle

%==================================================================================================
\section{Questions}

\begin{enumerate}
  \item Is overpredicting or underpredicting the range worse for the intended use?
  \item Is the background taken as the entire New World for every species? That could be a real problem if the goal is spatial prediction of current distribution, but solving it will probably be tricky.
  \item What sort of cross validation is currently used? Looks like the code calls it k-fold, but it's just one holdout sample of 20 \%, which isn't used for any sort of model selection.
  \item Does current thresholding of predictions to produce binary maps depend on the output type and its interpretation?
  \item Do the 19 correlated predictors have any undesirable consequences on the predictions, or is the strict interest in spatial distribution and the thresholding of the predictions sufficient motivation to ignore these correlations?
  \item Let's discuss whether we want to toss the duplicates. If we're interpreting output as use-availability and accounting for sampling bias, we maybe shouldn't.
  \item Check whether quantiles (currently 1\%) of outputs give the best thresholded predictions. Past efforts have probably narrowed this down to the best guess to apply to all species. Try to find Steven's criticism of these quantiles.
\end{enumerate}
%==================================================================================================
\section{Easy}

\begin{enumerate}
  \item Increase regularization penalty
  \item Remove threshold features, as these are the most sensitive to sampling idiosyncrasies in geographic space, which can translate to weird jumps in response curves and occurrence probability. Hinge features can probably take care of the legitimate jumps.
  \item Use all data to fit the model, unless we're going to use the holdout data/cross-validation for something (which we should, in the next round of more complicated changes).
%  \item
\end{enumerate}

%================================================================================
==================
\section{Medium}
These options just involve running the existing code multiple times with slight changes in the settings to find the best settings, based on performance under cross validation. We'll have to decide on the best metric for cross-validation, which is probably not AUC.
\begin{enumerate}
  \item Fit models over a range of regularization values and choose the value that's best for each species.
  \item Perform model selection to remove redundant/correlated predictors. Even relatively unimportant ones can lead spikes in suitability if product, hinge and threshold features are used.
%  \item
%  \item
\end{enumerate}

%================================================================================
==================
\section{Hard}

\begin{enumerate}
  \item Model sampling bias based on target group sampling. I suspect the best approach is to build a maxent model for the target group (as if it were one species) and supply this as the sampling bias surface. But I think we'd need to do some tests of 3-4 different approaches to be sure, since no one has ever really addressed this.
  \item Select background specifically for each species, or at least taxonomic group, to avoid spurious effects of comparing used locations to unavailable locations.
  \item Remove spatial autocorrelation. (I'm not totally clear on whether the spatial eigenvectors describe or remove it.)
%  \item
\end{enumerate}

%================================================================================
==================
\section{Someday}
```

```latex
\begin{enumerate}
  \item Map potential spatial distributions, which focus on getting the reasonable environmental responses rather than just the right spatial patterns (with disjoint response curves).
  \item Provide maps of habitat suitability rather than just the binary maps.
  \item Provide uncertainty in predictions over the k-folds used for cross-validation.
  \item Create a metadata file of the settings for each species and allow species specific tuning of model settings (e.g. by BIEN users). For example, an expert might narrow down the 19 bioclim predictors to just a few known to be relevant.
  \item For extremely rare species, consider pooling with other members of a higher taxonomic group to get enough points to build a maxent model.
\end{enumerate}
```