# Rubric

| Section | Deliverable | Description | Due Date | Points |
| --- | --- | --- | --- | --- |
| Proposal | Proposal | Provide a link to the dataset, example images, description of the variation in the dataset (e.g.  categories, size/resolution, etc), description of the intended classification problem (i.e., list of  output categories), estimation of the approximate number of images expected to be in each  category, and a guess of the types of image features that may be useful for this categorization (e.g.edges, histograms, etc) | July 1st | 5 |
| Feature extraction | Report | This part should include code to extract features, illustrations of the features extracted from several  example images, plots showing the amount of variation in the dataset, as well as PCA decomposition and tSNE visualization of features. Be sure to accurately describe and interpret your  methods and results. You must include at least two simple features and at least one complex feature  (see above). Include a detailed explanation of why you chose these particular features for your  dataset and classification problem. |  | 35 |
| Classification | Report | This part should include code to perform classification using at least two methods learned in class,  plots showing the results of classification per category, a discussion of possible reasons why the  classifier might work better for some categories than others, and explanation of the limitations of  the classifier. |  | 35 |
| Generalizability | Report | Your data should be split into train, validation, and test groups before training the classifier, and you  should do a hyperparameter search using parts of the validation set in a way that avoids over-fitting  and maximizes generalizability. Report performance on the test set, and include a discussion of  whether you achieved generalizability, and how your training process might be improved |  | 10 |
| Efficiency vs Accuracy | Report | For the various combinations of three feature vectors and two classifiers, include characterization of  both accuracy and training/inference cost (based on time, assuming equal computational power).  Optimize one solution for accuracy and one for efficiency. Include a discussion of the relative trade-  offs necessary for each of these solutions. |  | 10 |
| Quality of explanation | Report | Overall quality of report, including readability of figures and code comments, quality of analysis,  and discussion of limitations. |  | 5 |

## 

# June 13th-14th (Week 6 of class)

## Agenda

* ~~Go over datasets and select one for the final project: Colombia~~
* ~~Decide how often we want to meet:~~ 
  + weekly Friday morning HKT (7:00 am) <https://berkeley.zoom.us/j/95432733261?pwd=czLShYgUP6Wd3bWnqB41ff8axfRcGH.1>
* ~~Roles & responsibilities: setting up meetings DONE, going to office hours (Juliana), planning the agenda for the meeting (all),~~
* submitting assignments (ask how they’re to be submitted)
* ~~Communication strategies: Slack~~
* Workspace: Collab (Pro if needed) GCP - RESEARCH
* Time constraints of team members:
  + Jacqueline - 2nd week of July in Europe. Last week of July in Japan. No reschedule.
  + Juliana - 1st week of July - Europe. No reschedule.

## Action Items

* ~~Add the dataset to Google sheet (Juliana)~~
* ~~Proposal - July 1st, start a Google doc.~~
* ~~Colab - Jacqueline will set up a folder~~

## Questions

* Which cloud provider - research.

# June 20th-21st (Week 7 of class)

## Agenda

* whether to choose a [subset](https://colab.research.google.com/drive/1fd4bsvs5VJ8wJ8GiqCVaK_Y97wnE-W8i#scrollTo=zbRj24NMQEgT) of classes (3 at the minimum) for the project
  + visually distinguish from each other and
  + contain a large number of images
    - Select top classes in cameras with most images.
    - Augment if necessary
* GCP or Databricks
  + Dataproc makes more sense because of memory limitations, Jupyter notebook on Dataproc

## Action Items

* Camera:class distribution [Juliana Gomez Consuegra](mailto:julianagc@berkeley.edu)
* [Jacqueline Lam](mailto:jacq@berkeley.edu)and [Henry Michael Gardner](mailto:henrygardner@berkeley.edu)will look into GCP
* Next week: 4:00 PT

## Questions

* How should we submit our assignments? Individually or one person per group?
  + Per Albert: “*Historically it has been one submission per group, but please make sure you include who is in your group, so they get credit.*”

# June 27th-28th (Week 8 of class)

## Agenda

* [Jacqueline Lam](mailto:jacq@berkeley.edu) and [Henry Michael Gardner](mailto:henrygardner@berkeley.edu): discuss GCP advancements
  + Collab / Collab Pro if necessary.
* ~~Peruse proposal to see if what Juliana initially wrote down makes sense (and feel free to change whatever you see fit, I apologize again for missing this week’s meeting)~~

## Action Items

* ~~Include additional features that would make sense for our species~~
* ~~Finalize the proposal~~

## Questions - Answered in Rachel’s OH

* How to change from infrared to grayscale?
  + See percentage of infrared vs RGB
  + Just use the red channel from daytime images, so they would have a similar color profile.
  + Feature extraction pipeline: first detect color or infrared image, and then if color, extract these, else….separate classifier for each type. Each line is one image, each column is one feature, have to be the same for all of them. They won’t be the same. Multuheaded classifier.
* Different cameras - identifying cameras vs identifying species
  + Within vs across group variance: make sure you have multiple cameras for multiple classes.
* Can we change classes along the way (worst case scenario)
  + Yes, we can, we’re sure to.
* Should we add EDA to our proposal?
  + Not mandatory.

Additional notes:

* Find published papers and look at what other people have done, as long as you cite their data
* If we have multiple shots in a sequence of the same species, make sure we group them together either in train set or test set, to avoid data leakage.

# July 4th-5th (Week 9 of class)

## Agenda

* Review Henry’s 2 minute summary, split parts

## Action Items (For two weeks!!!)

* Anticipated problems:
  + Animal is asleep in front of camera - remove background?
  + Animal is too close to camera - if cosine similarity is too high, keep just one, for independent
  + Animal walking slowly in front of camera - same as above
  + Data leakage from similar images
* Downsample dataset
  + Take just the red channel
* Preprocess (change bottom pixels to black)
* Split dataset
* Each select a feature
* Read paper from data

## Questions

# July 11th-12th (Week 10 of class)

## Agenda

* Split tasks
  + Downsampling [Henry Michael Gardner](mailto:henrygardner@berkeley.edu) will finish look at Juliana’s [notebook](https://colab.research.google.com/drive/1st-PFF8sgeQO-omC7Izq8bHGi5b1j1hC?usp=drive_link) and add to it. ETA: Friday evening (HK Saturday)
  + Preprocessing (Saturday / HK Sunday)
    - Red channel only [Jacqueline Lam](mailto:jacq@berkeley.edu)
    - Change bottom pixels to black [Juliana Gomez Consuegra](mailto:julianagc@berkeley.edu)
    - Splitting images into train/test, making sure all frames from a specific species/time are in just one of the groups, to avoid data leakage [Jacqueline Lam](mailto:jacq@berkeley.edu)[Man Eunice Ngai](mailto:eunicen@berkeley.edu)[Juliana Gomez Consuegra](mailto:julianagc@berkeley.edu)[Henry Michael Gardner](mailto:henrygardner@berkeley.edu)(whoever’s done can tackle this: Json file has frame number)
    - Saving image after processing [Man Eunice Ngai](mailto:eunicen@berkeley.edu)
  + Feature extraction (Sunday / HK Monday)
  + Reading papers to find more features
    - This [paper](https://drive.google.com/file/d/1Eve4fQ1voNjYUjSGI8t5SG3XFUUlss-l/view?usp=drive_link) has feature extraction for some of our species, and uses a technique similar to HOG, which extracts textures, so that might me worth looking into.
  + Assignment 6: do it together at some point?? Yes, discuss over Slack

## Action Items

* Would everyone be okay with creating their individual folders inside our drive, so that we can see what others are working on to not repeat work?

## Questions

# July 18th-19th (Week 11 of class)

## Agenda

* Questions:
  + ~~Frames? (Group\_Images\_by\_sequenceID notebook)~~
* Split tasks
  + Jacqueline/Henry: lower resolution - Friday
  + More features: Eunice/Juliana
  + Extract mask of animal from empty image (subtract image without animal from image with animal)
    - Might be because of intensity of the light
      * ~~Compare luminance~~
      * Compare green
  + ~~Eunice: Split into train/validation/test 70:10:20~~ 
    - Preprocess only train set
  + ~~Jacqueline: look at curve and HOG feature - Saturday~~

## Action Items

1. Keep only frame 0-(lowest number of frames when more than one frame), exclude classes with < 1000 [Man Eunice Ngai](mailto:eunicen@berkeley.edu)
2. Remove duplicates (where 2 animals are in the same frame)[Man Eunice Ngai](mailto:eunicen@berkeley.edu)
3. Downsampled all classes to 1000 images per class[Man Eunice Ngai](mailto:eunicen@berkeley.edu)
4. Split our data by sequence id and common name into train\_val\_test (70:10:20)[Man Eunice Ngai](mailto:eunicen@berkeley.edu)
5. Preprocess black metadata for downsampled [Jacqueline Lam](mailto:jacq@berkeley.edu)
6. Preprocess red channel for downsampled [Jacqueline Lam](mailto:jacq@berkeley.edu)
7. Preprocessing on train set only for now:
   1. Lower resolution [Henry Michael Gardner](mailto:henrygardner@berkeley.edu)
   2. Remove background with blank image as a basis [Juliana Gomez Consuegra](mailto:julianagc@berkeley.edu)
8. Features:
   1. HOG [Jacqueline Lam](mailto:jacq@berkeley.edu)
   2. Curve [Jacqueline Lam](mailto:jacq@berkeley.edu)
   3. [Henry Michael Gardner](mailto:henrygardner@berkeley.edu) [Jacqueline Lam](mailto:jacq@berkeley.edu) [Juliana Gomez Consuegra](mailto:julianagc@berkeley.edu)[Man Eunice Ngai](mailto:eunicen@berkeley.edu)ideas for new features?
   4. Complex feature from other models
9. PCA

## Timeline

| **Step** | **Date** |
| --- | --- |
| 1 | Friday 9:00 PM PST |
| 2 | Friday 9:00 PM PST |
| 3 | Friday 9:00 PM PST |
| 4 | Friday 9:00 PM PST |
| 5 | Friday 9:00 AM PST notebook ready to run |
| 6 | Friday 9:00 AM PST notebook ready to run |
| 7 | 1. [Henry Michael Gardner](mailto:henrygardner@berkeley.edu)what time will you have this?   B. Saturday 9:00 AM PST (on sample images) |
| 8 | 1. Saturday 6:00 AM PST (on sample images) 2. Saturday 6:00 AM PST (on sample images) 3. Saturday 6:00 AM PST   Run notebooks on features with the whole dataset for the rest of **Saturday PST**. |
| 9 | Sunday morning: meet |

## Questions

# July 25th-26th (Week 12 of class)

## Agenda

* Blue images / zeros update [Jacqueline Lam](mailto:jacq@berkeley.edu)
* CNN - where do we stand on that? Do we have to rerun feature vectors on it? (I mention this because we have hog2 and sobel2, but not cnn2) [Jacqueline Lam](mailto:jacq@berkeley.edu)
* Remove crappy images?
  + First run without removing extra images, if accuracy bad, resplit it.
* ~~SIFT + BOVW update~~
  + ~~I think we’re ready to include the feature vectors with tf-idf. I’ll pass them through PCA tonight and add them to our feature vector matrix.~~
* ~~Image differencing update~~ [~~Henry Michael Gardner~~](mailto:henrygardner@berkeley.edu)
  + Final week!
* Which classifiers do we want to use? (the rubric states “two methods learned in class”)
  + LDA [Man Eunice Ngai](mailto:eunicen@berkeley.edu)
  + SVM [juliana.consuegra77@gmail.com](mailto:juliana.consuegra77@gmail.com)
  + Simple perceptron [Man Eunice Ngai](mailto:eunicen@berkeley.edu)[juliana.consuegra77@gmail.com](mailto:juliana.consuegra77@gmail.com)[Henry Michael Gardner](mailto:henrygardner@berkeley.edu)[Jacqueline Lam](mailto:jacq@berkeley.edu)
* We split our data into training, validation and test, but the rubric says “*For full credit, implement a hyperparameter search using subsets of the validation data*.”
  + Split test / validation [Man Eunice Ngai](mailto:eunicen@berkeley.edu)
* For our feature extraction, we’re missing “*plots showing the amount of variation in the dataset*”
  + SIFT [juliana.consuegra77@gmail.com](mailto:juliana.consuegra77@gmail.com)
  + HOG [Jacqueline Lam](mailto:jacq@berkeley.edu)
  + Sobel [Jacqueline Lam](mailto:jacq@berkeley.edu)
* Presentation
* Report [Henry Michael Gardner](mailto:henrygardner@berkeley.edu)

## Action Items (with timeline)

* Rerun features to correct for cv blue and missing zeros
* Rerun PCA with sift and add to feature vector matrix
* Variation in the dataset - feature extraction.
  + HOG average histogram
  + Maybe also plot luminance, day vs night images, etc?
* Hyperparameter search
* Classification
  + LDA
  + SVM
* Presentation
* Report
* Organize Google Drive so that Rachel can understand it. Possible folders based on the rubric:
  + Proposal
  + Feature Extraction
  + Classification

## Questions

* Juliana asked on BCourses: “*I've been trying to apply SIFT with Visual Bag of Words to our animal classification problem, and have experimented with different numbers of clusters, but a lot of the background is still classified as descriptors. The resolution of the images has been significantly lowered (to 10% of the original size), and we have performed histogram equalization (CLAHE). My question is: what else can I do to improve this feature vector? (i.e. ignore all descriptors in the background).*”
  + Rachel’s answer: “*There is really no way to "ignore" the background unless you have a high quality mask to remove it. You could conceivably create that by subtracting your "blank" image from the animal image, and then setting a difference threshold for the subtraction image (i.e. find the highest density cluster of high-difference values and set a bounding box around that cluster). However, I would encourage you to try using your BOVW even with the background, because it might still work! The idea behind BOW generally is that you would remove words that are very common (think "a", "the", "of" etc) and make a dictionary focused on the medium or low frequency words. I would expect that the visual characteristics of the background are pretty consistent across images, and would therefore hopefully not be as important in your final dictionary. You might need to change some of your dictionary formation settings to make this happen.*”

# August 1st-2nd (Week 13 of class)

## Agenda

* ~~Updates:~~ 
  + ~~Image differencing~~
  + ~~LDA~~
  + ~~SVM~~
  + ~~NN~~
* ~~Standardize code for training/validation/test times to report along all models~~
* ~~Standardize confusion matrix, so that we show % per class instead of raw numbers, since classes have different sizes~~

## Action Items

* ~~Check labels for all images in our training set~~[~~Juliana Gomez Consuegra~~](mailto:julianagc@berkeley.edu)
* Re-run LDA on individual feature vectors?
  + If time allows
* Re-run SVM on individual feature vectors [Juliana Gomez Consuegra](mailto:julianagc@berkeley.edu)
  + ~~HOG 0.76~~
  + Sobel
  + ~~SIFT 0.53~~
  + CNN
* Test sets
  + LDA
  + SVM
  + NN
* ~~Image differencing~~
* Add to slide deck (each one add to their part)
* Read report and add to it if necessary
* Practice presentation
  + Add to speaker notes.

## Questions

### From office hours:

* For the various combinations of three feature vectors and two classifiers, include characterization of both accuracy and training/inference cost (based on time, assuming equal computational power).
* Optimize one solution for accuracy and one for efficiency. Include a discussion of the relative trade-offs necessary for each of these solutions.
  + **Accuracy vs efficiency:** You can get the best accuracy when you Put everything in there. HIghest accuracy, but also look at other models with lower accuracy, and see a trade-off in terms of training time. If there’s a huge gap, CAUTION (best model is miles apart from the second model), make sure you’re doing good faith in second best model. i.e. NN vs canny filter.
  + **Report training time:** how long it took. At least report training time, possibly also inference time. But definitely not just inference time. Training time: time to train some batch number, or the whole dataset. Main thing: collect for all different models, to report apples-to-apples.
* Other questions from OH from other people:
  + Austen
    - Complex vector: feature vector derived from a network, or visual bag of words.
    - Neural network done on feature vectors and not on the raw images. So there’s no end-to-end (pass an image and get a classification).

### To discuss with the group:

* After k-fold cross validation, we perform inference with our folds.
  + With our best performing model:
    - K-fold train and validation report accuracy
    - Or report train accuracy, validation and test accuracy with our modeo, either one is fine.
    - So, how to report validation accuracy with k-folds?

# August 5th-9th (Week 14 of class)

## Agenda

### From Henry’s notes section:

* ~~NN confusion matrix to percentages?~~ [~~Jacqueline Lam~~](mailto:jacq@berkeley.edu)
* ~~any other visuals?~~
  + ~~Update tsne~~ [~~Juliana Gomez Consuegra~~](mailto:julianagc@berkeley.edu)
  + ~~Update nn to %~~ [~~Jacqueline Lam~~](mailto:jacq@berkeley.edu)
  + ~~Update edge detection~~ [~~Jacqueline Lam~~](mailto:jacq@berkeley.edu)
* ~~Change data to actual source instead of link~~
* ~~F1-score and other metrics~~

### Juliana’s comments:

* ~~Modify our report based on our best-performing model (NN?) Nope~~
* ~~Decide how we want to frame using sobel or canny~~
* ~~Select which results to show on the report:~~
  + ~~suggestion: PCA for each of the features/classifiers plus full feature matrix, in order of complexity i.e:~~

| Classifier | Features | Training Accuracy | Test Accuracy | Training Time |
| --- | --- | --- | --- | --- |
| LDA | Sobel PCA |  |  |  |
| HOG PCA |  |  |  |
| SIFT PCA |  |  |  |
| CNN PCA |  |  |  |
| Sobel-HOG-SIFT-CNN PCA |  |  |  |
| SVM | Sobel PCA |  |  |  |
| HOG PCA |  |  |  |
| SIFT PCA |  |  |  |
| CNN PCA |  |  |  |
| Sobel-HOG-SIFT-CNN PCA |  |  |  |
| NN | Sobel PCA |  |  |  |
| HOG PCA |  |  |  |
| SIFT PCA |  |  |  |
| CNN PCA |  |  |  |
| Sobel-HOG-SIFT-CNN PCA |  |  |  |

* ~~Select which notebooks we’re going to leave on Google Drive~~
  + ~~all the ones mentioned in the table above (plus notebooks 1-4).~~
  + [~~Juliana Gomez Consuegra~~](mailto:julianagc@berkeley.edu)~~move SVM~~
* Make sure all the visualizations we include are referenced in the document. [Juliana Gomez Consuegra](mailto:julianagc@berkeley.edu)

### Missing from the rubric:

* + Readability of the code
  + ~~Inference time for all models~~
    - ~~Add for LDA~~ [~~Henry Michael Gardner~~](mailto:henrygardner@berkeley.edu) ~~and NN (full feature matrix)~~ [~~Jacqueline Lam~~](mailto:jacq@berkeley.edu)

## Action Items

* ~~LDA results include in the report - notebooks with PCA vectors.~~ [~~Henry Michael Gardner~~](mailto:henrygardner@berkeley.edu)

## 