Common Local Directory Structure

The big idea is that a uniform folder structure can be hardcoded into dataset logic to simplify any ETL processes, as it removes the variance in directory structure for each dataset. All operations can be performed by just passing in the root folder. There is need to conserve the original dataset structure, since after the schema transformation step, it will be downloaded from personal remote storage anyway.

- $HOME/datasets/
  - dataset_name/
    - downloads
    - archives
      - .tar / .zip
    - metadata
      - metadata.csv
    - imagefolder
      - images
      - masks
      - labels.csv
    - hdf5
      - dataset_name.h5 (encode directory structure, instructions on how to open and load data from the file inside the .h5 file itself)
    - litdata
      - train
      - val
      - test
- $HOME/experiments/project_name/
  - experiment_name/
    - dataset.csv
    - metrics.csv (special case for test run)
    - learning_curve.png
    - hparams.yaml
    - wandb/
    - checkpoints/
      - step=1_epoch=1/
        - confusion_matrix.npy
        - predictions.csv
        - step=1_epoch=1.ckpt (will need to modify model checkpoint)
        - inference/
        - attribution/
          - best_k_prediction/
          - worst_k_predictions/
      - test/ ...

Common S3 Object Structure

- s3://dataset_name is root bucket, rest is the same as local
- s3://project_name is root bucket, rest is the same as local
- s3://imagenet
  - metadata

    - metadata.csv
  - archives
    - imagenet.zip
  - hdf5
    - imagenet.h5
- s3://imagenet_classification
  - alexnet_pretrained
  - resent18_from_scratch