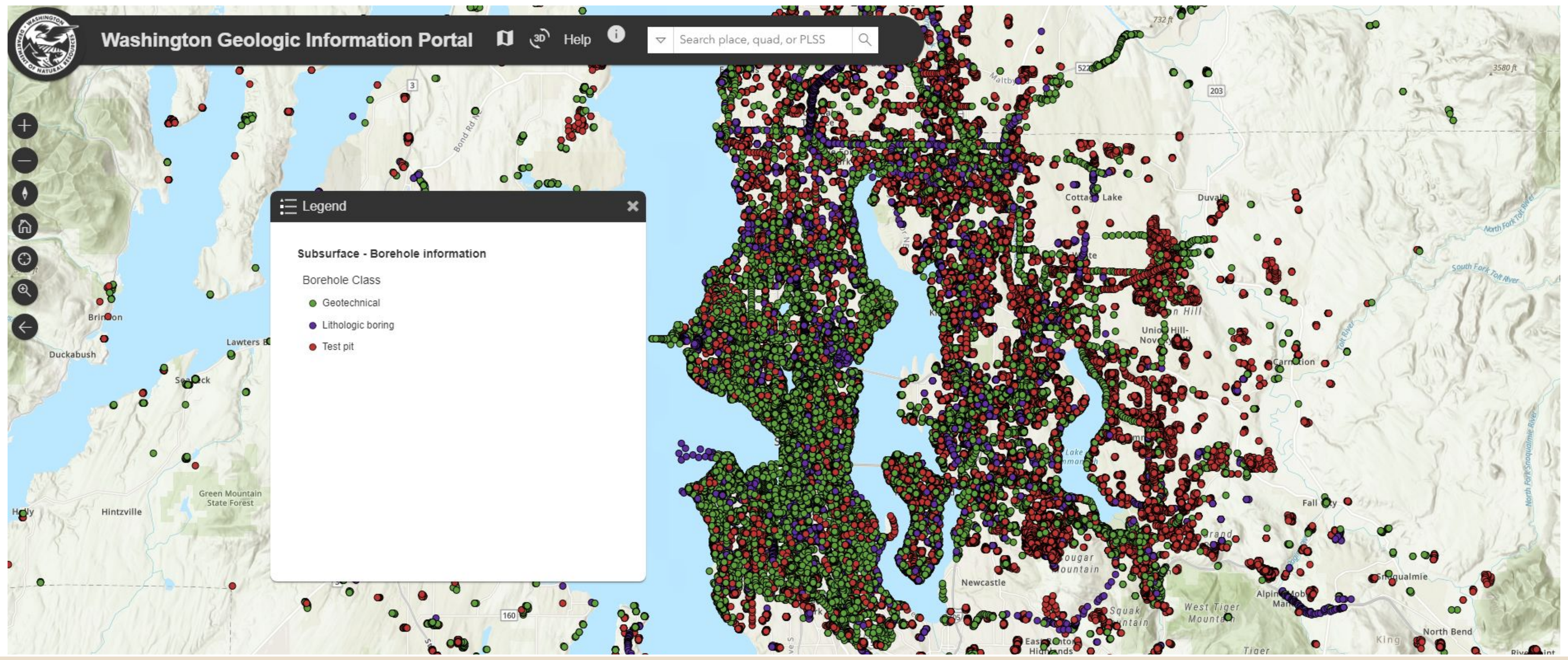
#multi-gpr-soil

Morgan, Mansa, Hailey, Steven, Brendan

#overview:

- Dense, historic soils data have been underutilized
- Gaussian Processes (GPs) can be used to create probabilistic predictions fit to real data
- Soil predictions are foundational to several engineering and environmental applications
- Aim: to predict soil type and thickness across Seattle



#takeaways

- WOW! We learned a lot
- GPs may not be the best model architecture for such dense data
- Multi-output GP models are not suitable for combined classifier and regression models

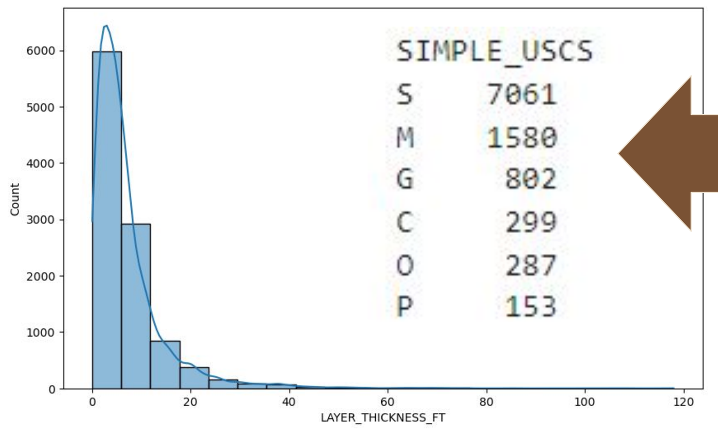
#next_steps

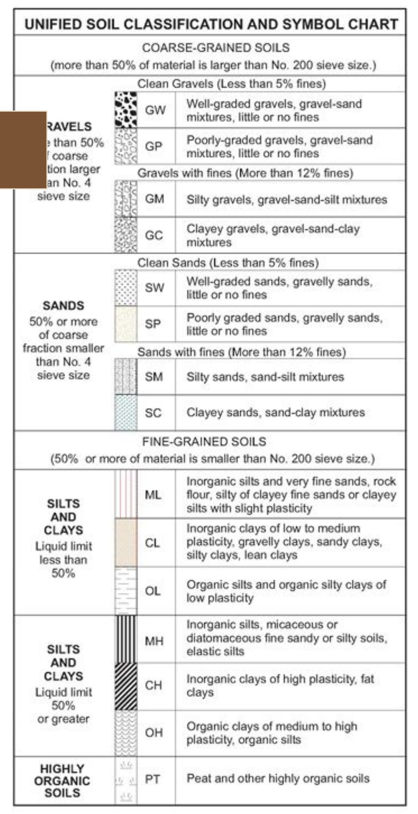
- Try other ML algorithms for comparison (e.g. decision trees, random forest)
- Try to predict thicknesses and soil classes for layers 2, 3, 4+
- Model implementation and visualization for all of Seattle area
- Make a "use case" example to share on the GeoSMART website

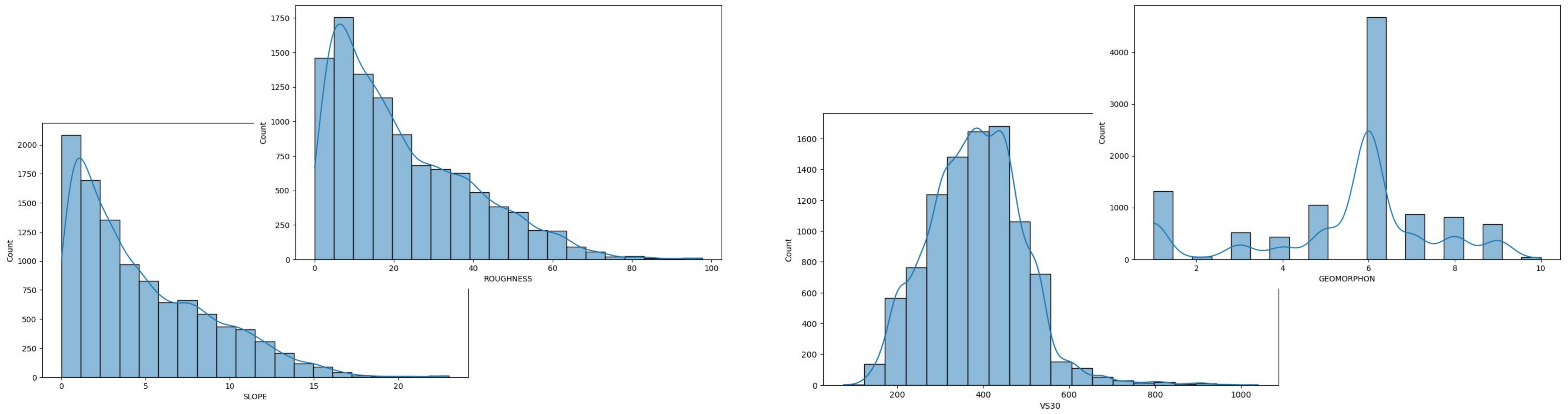
#accomplishments

#data_processing_&_feature_engineering

- Target cleaning and preparation
  - Compute layer thickness (regression)
  - "Simple" USCS (classification)
- Sampling geospatial features
  - Free, publicly available datasets
  - Geospatially continuous = maps
  - Surface proxies of subsurface soil conditions
    - DEM-derived slope, roughness, geomorphon
    - Surface geologic unit
    - Vs30
- Standardization/encoding
  - Regression features: mean-std standardization
  - Regression targets: log standardization
  - Classification features: one-hot encoding
  - Classification targets: label encoding
- Class Imbalance
  - SMOTE=Synthetic Minority Over-sampling Technique







#model_training_&_performance

- scikit learn
  - Classifier
  - Regression
- PyTorch
  - Classifier
    - Loss function: added a class similarity score to the loss function – i.e., predicting some classes are better than others.

$$\mathcal{L} = \lambda\mathcal{L}_{\mathrm{ELBO}} + (1-\lambda)\mathcal{L}_{\mathrm{CUSTOM}}$$

  - Regression
- Cross-Validation & Tuning
  - dask
- Testing & Performance

#classifier_could_use_some_work…_