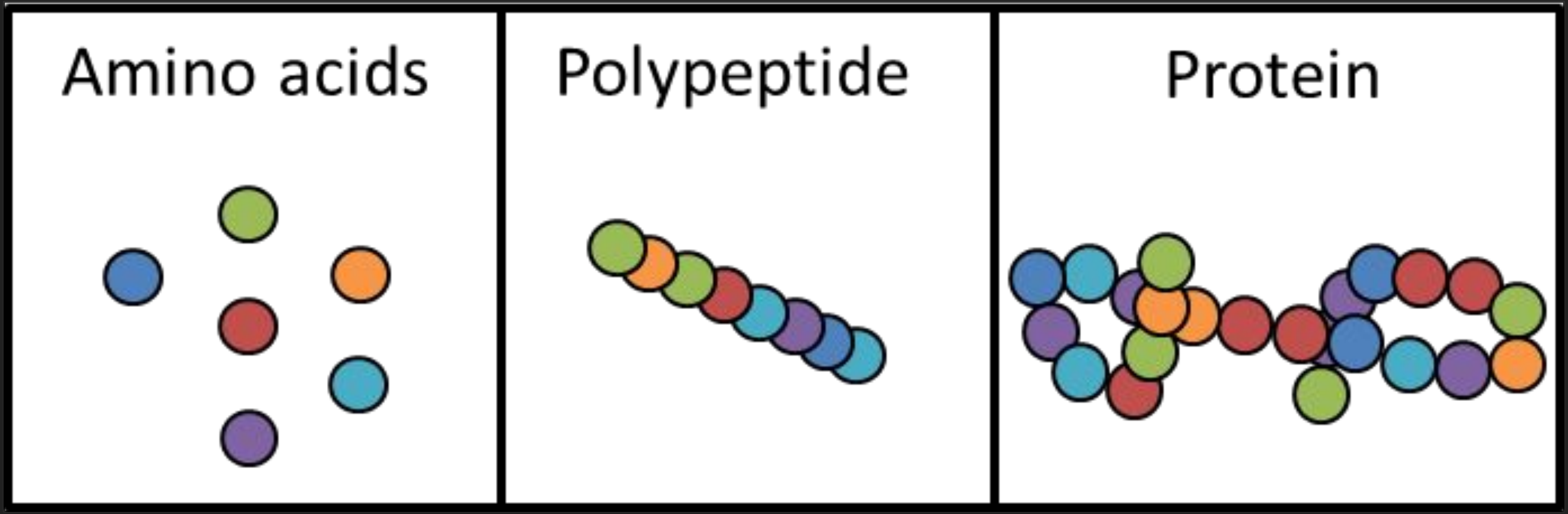
Properly Predicting Protein Localization

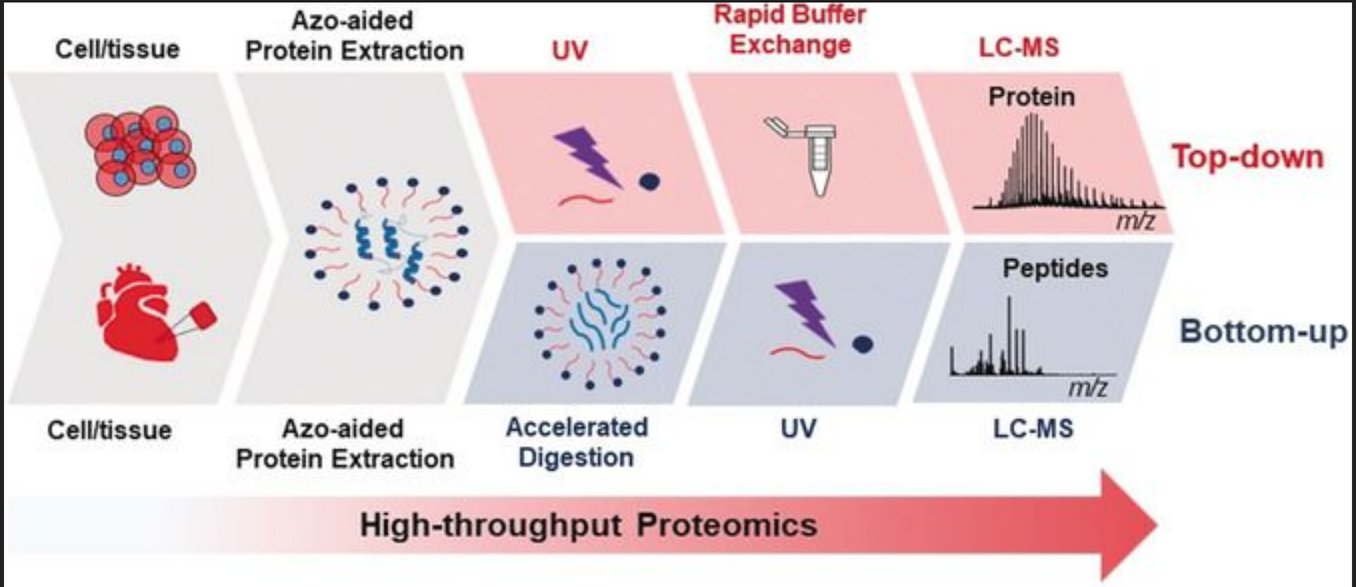
Zach Osking
SpringBoard Capstone 1 Project
Summer 2020

Proteins



- Large, complex macromolecule
- Perform many cellular functions
- Localization in cell hints at what function a protein may perform

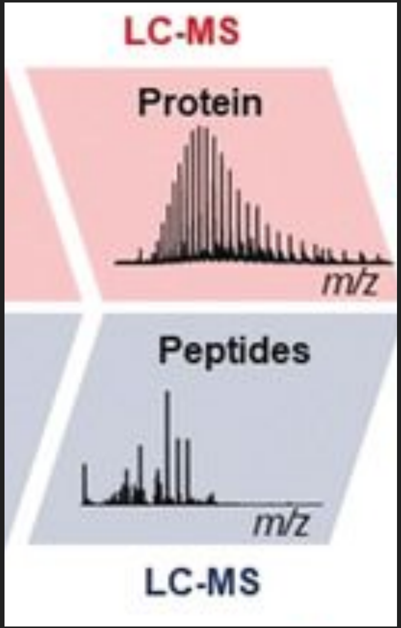
Proteomics


Cell/tissue
Azo-aided Protein Extraction
UV
Rapid Buffer Exchange
LC-MS
Protein
m/z
Top-down
Peptides
m/z
Bottom-up
Cell/tissue
Azo-aided Protein Extraction
Accelerated Digestion
UV
LC-MS
High-throughput Proteomics

Problem

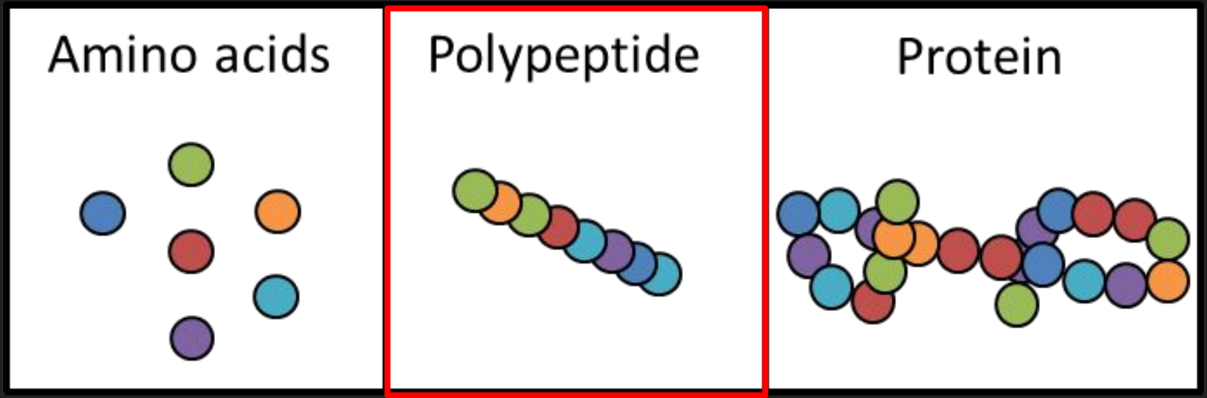
- Output data consists of millions of protein sequences
- If a new sequence appears, how do we evaluate it?
    - Sequence alignment with previously characterized proteins
    - *In vitro* characterization of new sequence

???

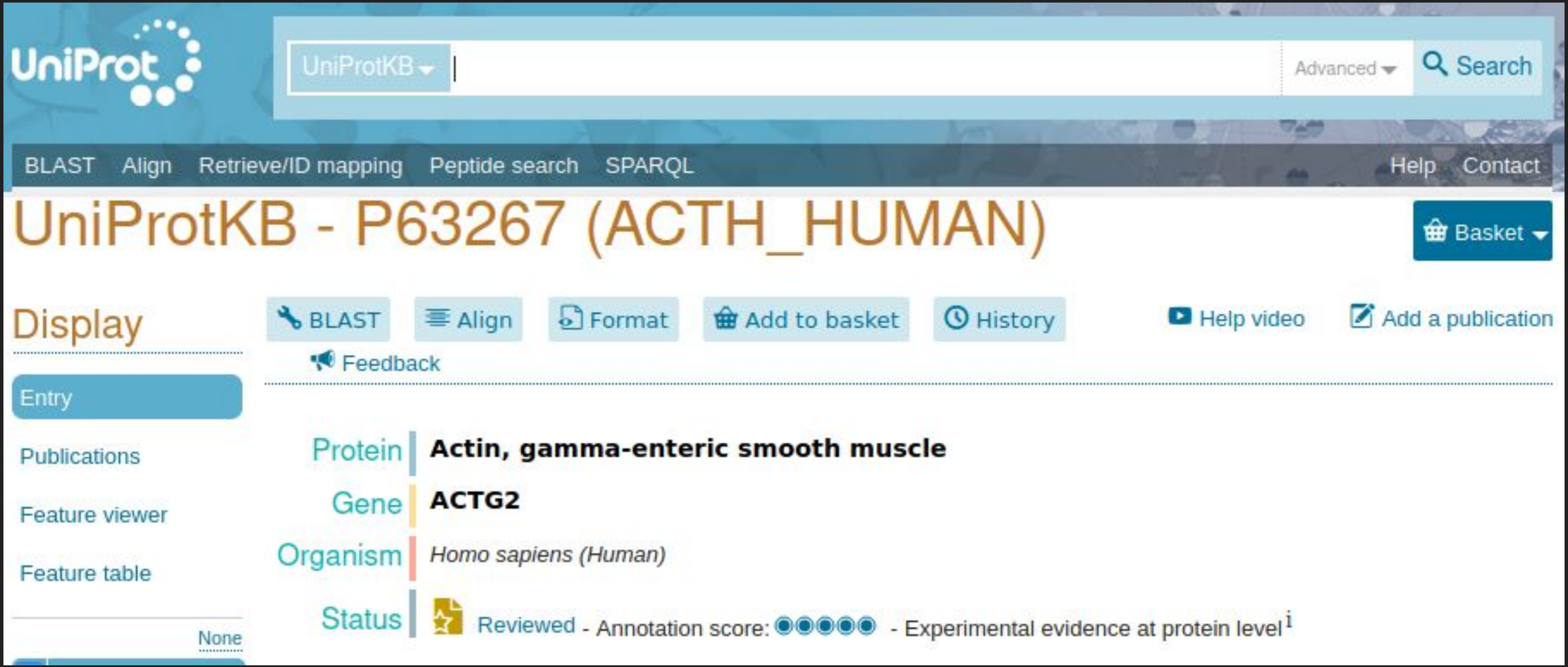


Solution - Machine Learning!

- Generate as many relevant features as possible using only the primary amino acid sequence
- Apply ML and NLP techniques to predict the localization of an uncharacterized protein within a cell



Dataset - Uniprot Database



UniProtKB - P63267 (ACTH_HUMAN)

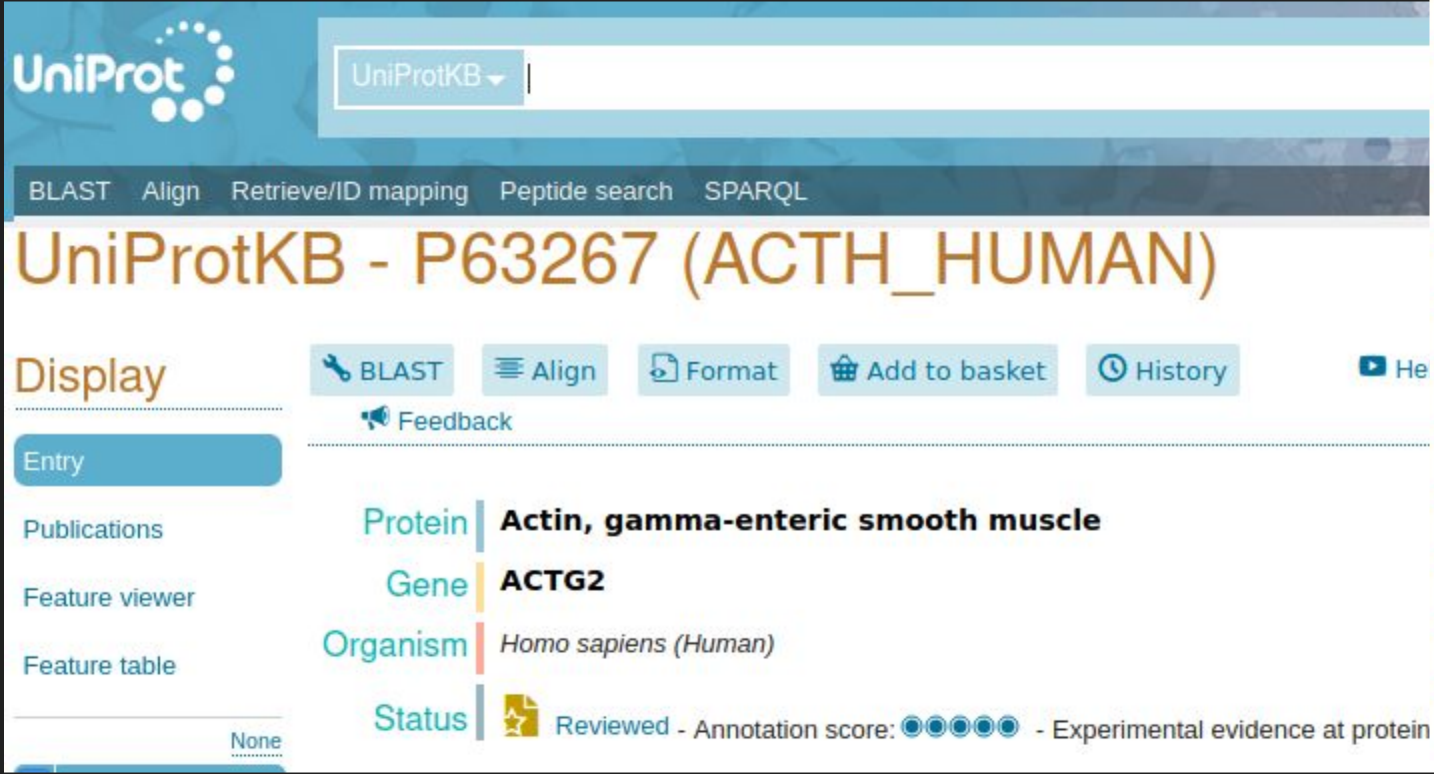
Protein | **Actin, gamma-enteric smooth muscle**

Gene | **ACTG2**

Organism | *Homo sapiens (Human)*

Status | Reviewed - Annotation score: ●●●●● - Experimental evidence at protein level i

Dataset - Uniprot Database



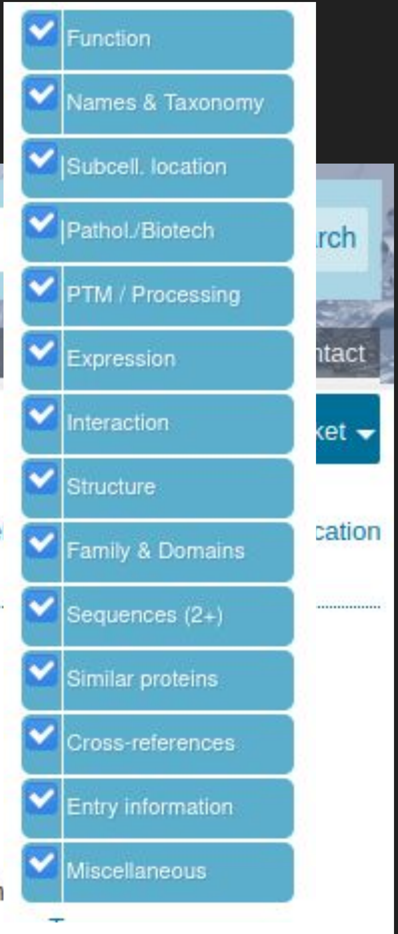
UniProtKB - P63267 (ACTH_HUMAN)

Protein | **Actin, gamma-enteric smooth muscle**

Gene | **ACTG2**

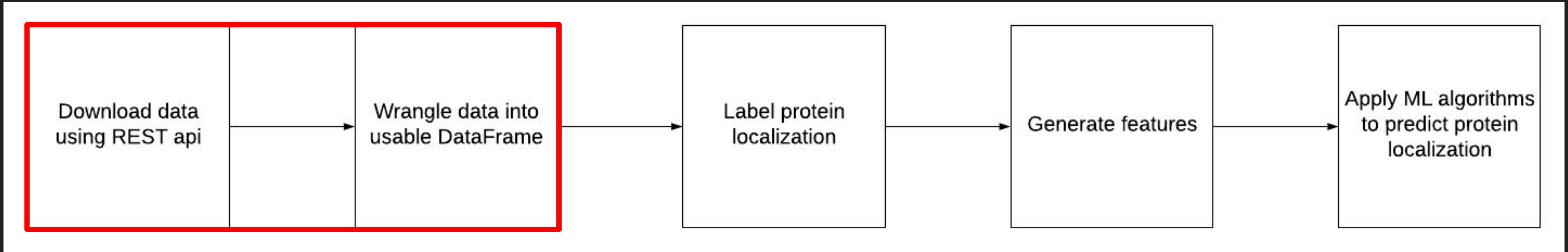
Organism | *Homo sapiens (Human)*

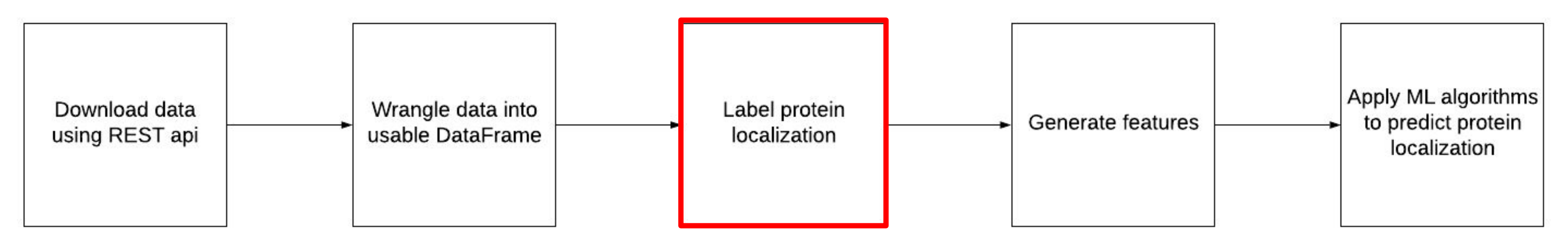
Status | Reviewed - Annotation score: ●●●●● - Experimental evidence at protein



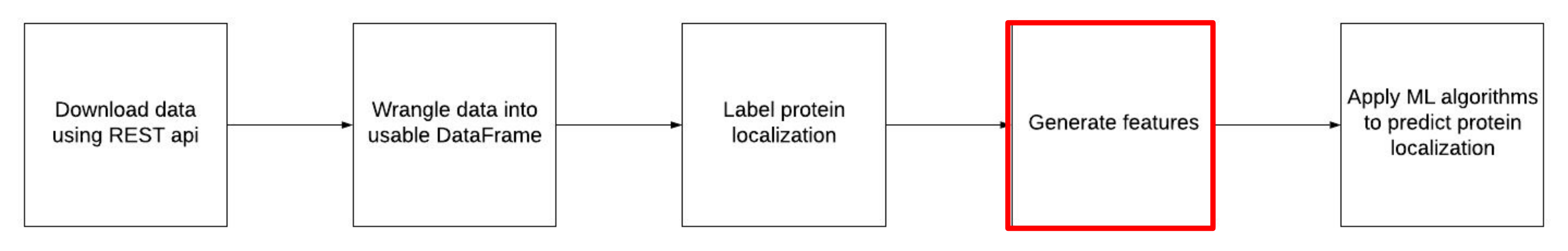
- ☑ Function
- ☑ Names & Taxonomy
- ☑ Subcell. location
- ☑ Pathol./Biotech
- ☑ PTM / Processing
- ☑ Expression
- ☑ Interaction
- ☑ Structure
- ☑ Family & Domains
- ☑ Sequences (2+)
- ☑ Similar proteins
- ☑ Cross-references
- ☑ Entry information
- ☑ Miscellaneous

Generalized Workflow


Download data using REST api
Wrangle data into usable DataFrame
Label protein localization
Generate features
Apply ML algorithms to predict protein localization


Download data using REST api
Wrangle data into usable DataFrame
Label protein localization
Generate features
Apply ML algorithms to predict protein localization

- Dozens of features contained in DataFrame related to protein localization
- Manually created lists of keywords related to certain localizations



Download data using REST api

Wrangle data into usable DataFrame

Label protein localization

Generate features

Apply ML algorithms to predict protein localization

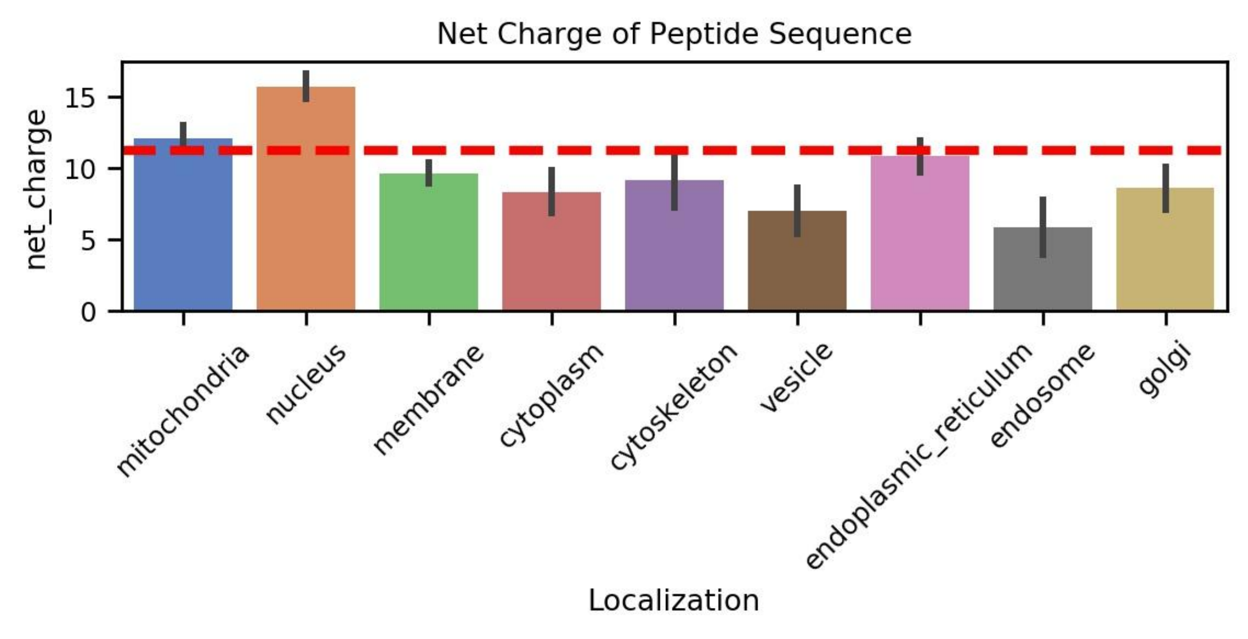
- Features generated based on basic properties of amino acid sequence
- Natural language processing techniques used on primary amino acid sequence

Feature Creation with Natural Language Processing Techniques

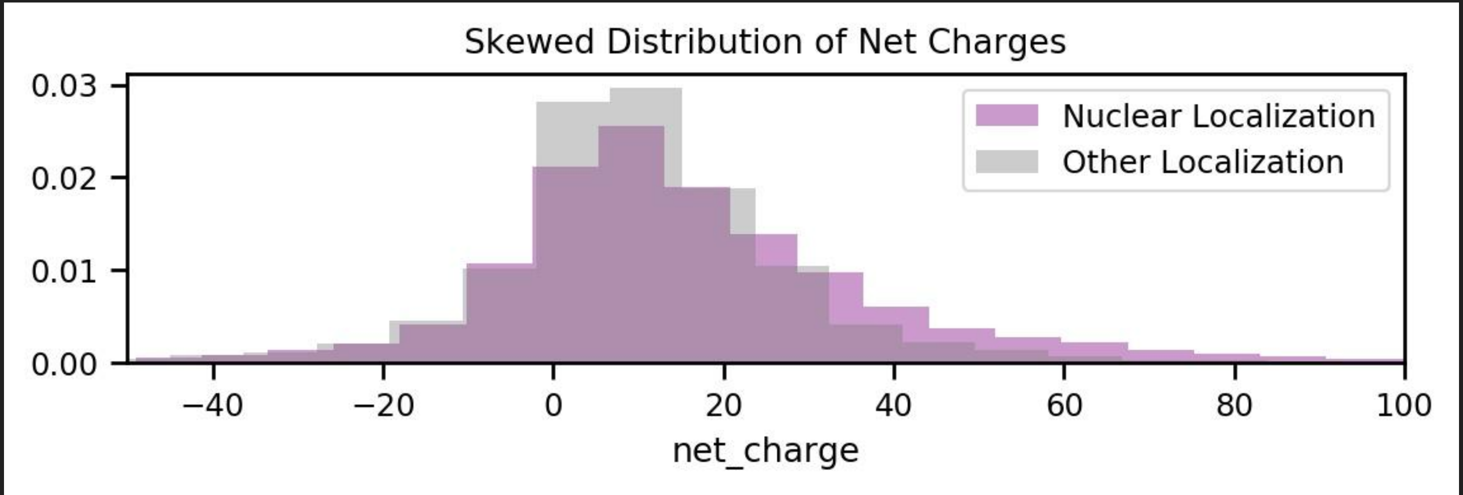
- AA sequence = “sentence” describing protein
- Proteome = corpus
- Natural Language Toolkit (NLTK) used to convert AA sequence information into vectors (bag-of-words)
- Word2Vec used to (hopefully) identify contextual patterns in AA sequences

- **Final DataFrame dimensions: 28,056 rows (proteins) x 151 features**

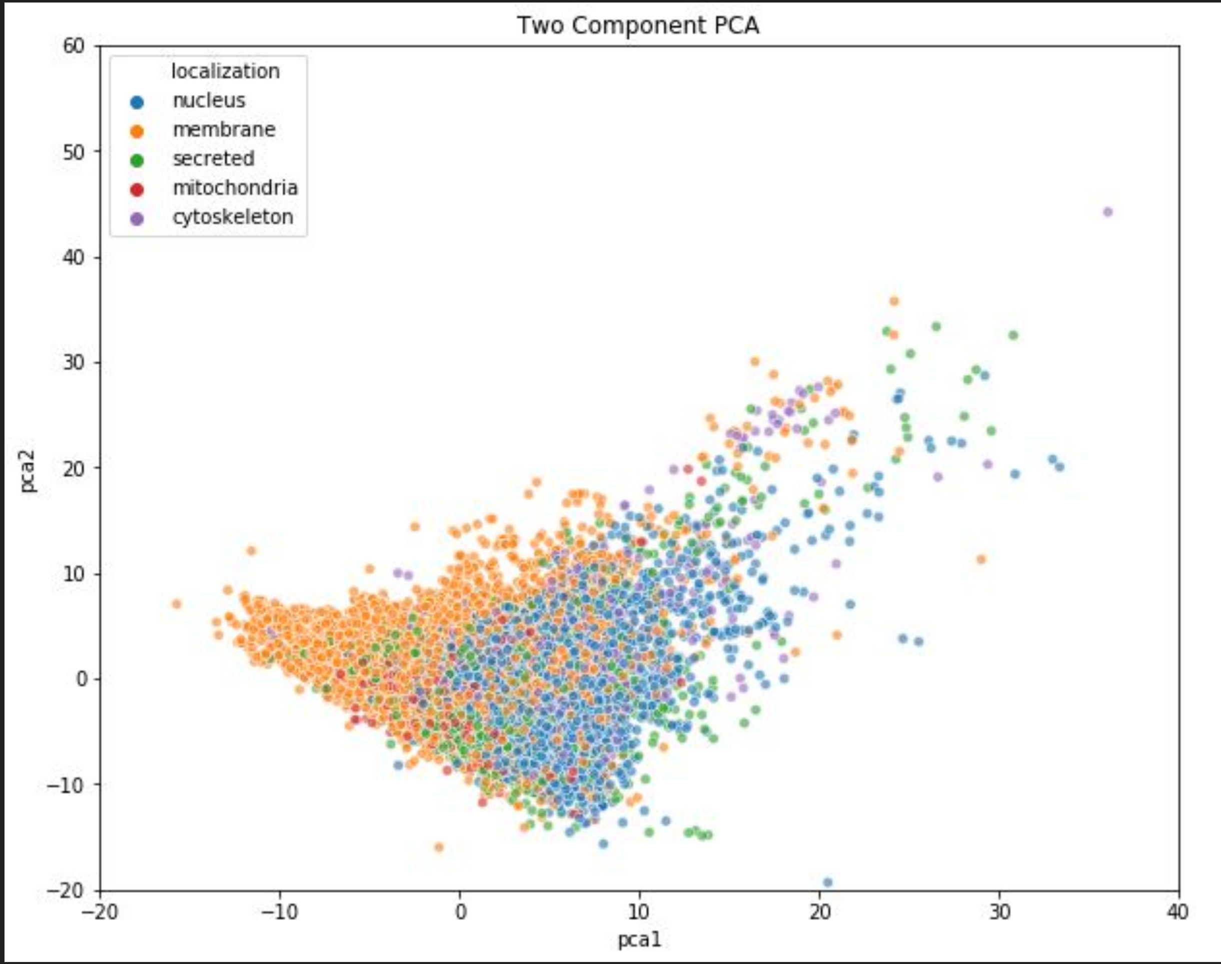
EDA - Relationship Between Net Charge and Protein Localization



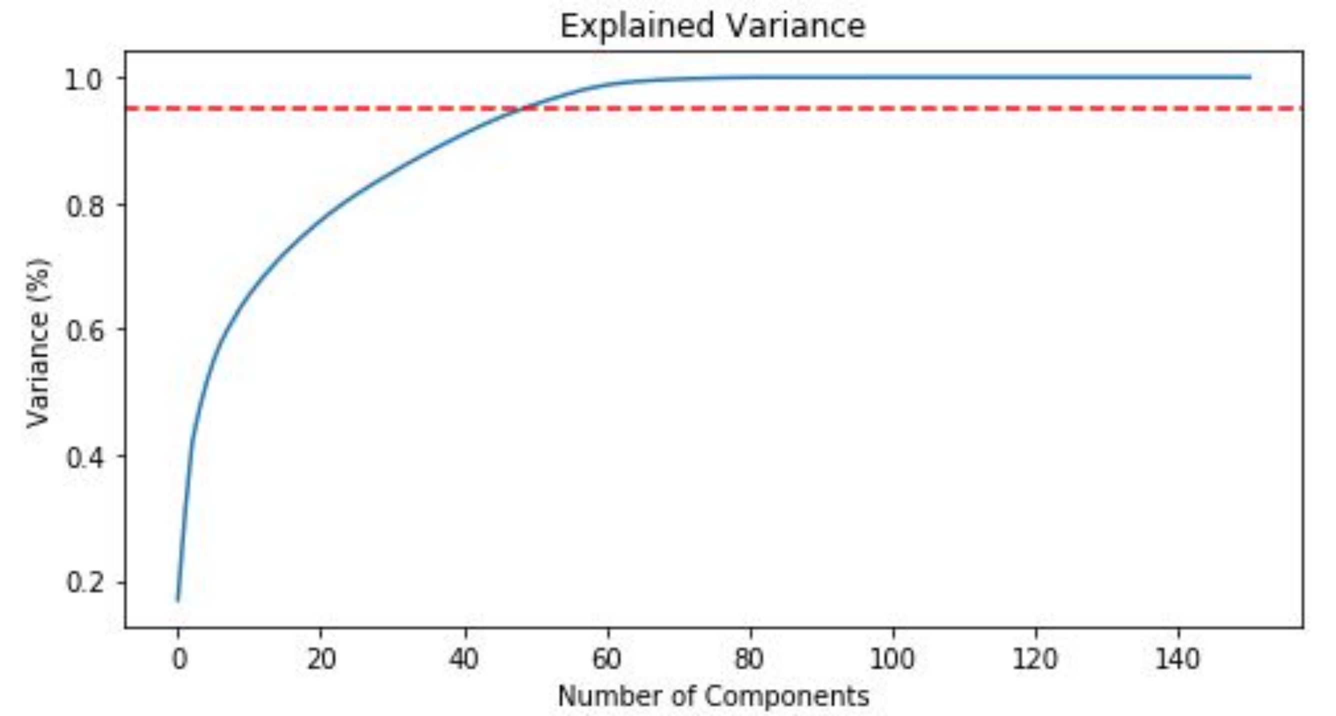
EDA - Relationship Between Net Charge and Protein Localization



Principal Component Analysis



Principal Component Analysis



Choosing ML Algorithms

Algorithm	Categorical Data	Multiple Categories	Small Sample Size	Unbalanced Categories
Logistic Regression	✓	−	✓	−
Support Vector Machine	✓	✓	−	−
Random Forest Classifier	✓	✓	✓	✓
Multi-Layer Perceptron	✓	✓	−	✓
Gradient Boosting Classifier	✓	✓	−	−

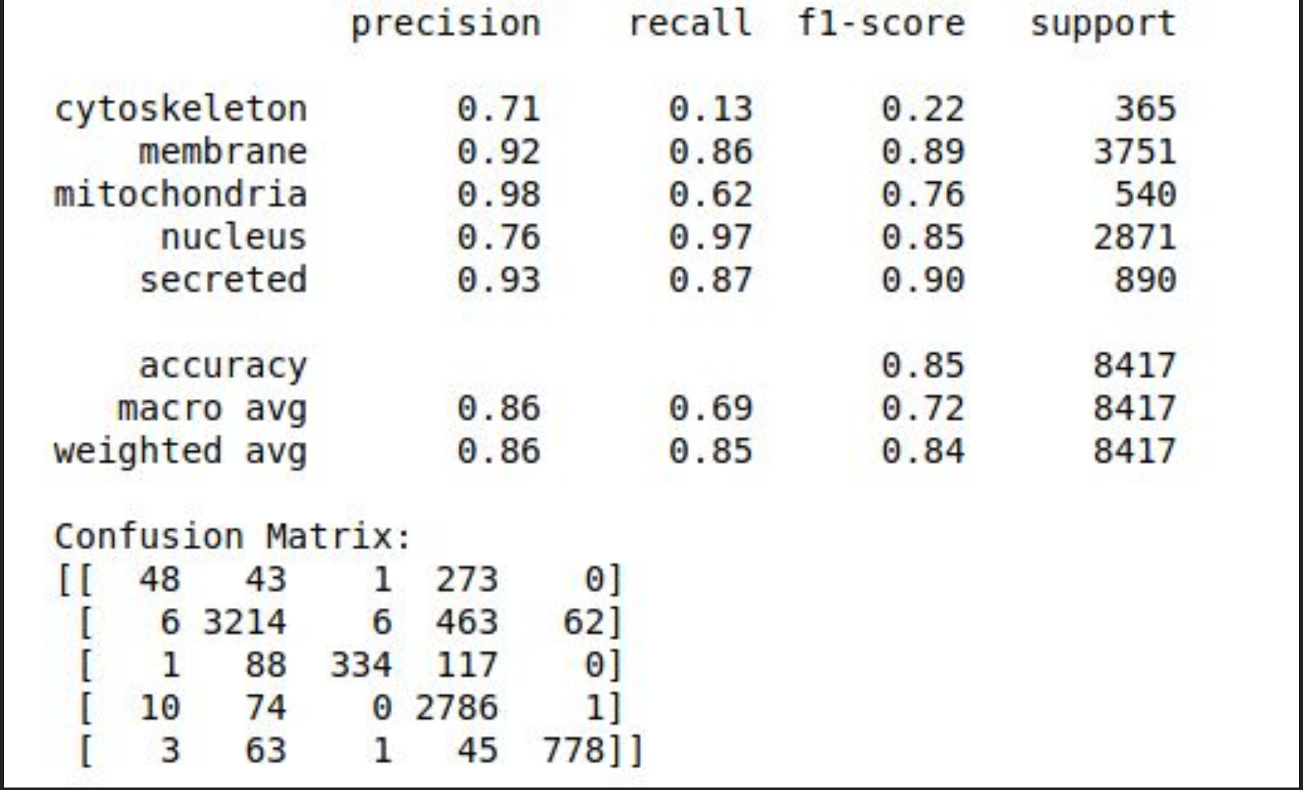
Preprocessing Data

Feature Set	Preprocessing
X	None
X_scaled	SciKitLearn StandardScaler
X_up	Upsampling of underrepresented categories
X_up_scaled	Upsampling of underrepresented categories, SciKitLearn StandardScaler

Results of Initial Algorithm Screening

Algorithm	Feature Set	Average f1-score	Weighted Average f1-score
Logistic Regression	X_scaled	0.70	0.82
	X_scaled_up	0.66	0.76
Support Vector Machine	X_scaled	0.72	0.84
	X_scaled_up	0.76	0.84
Random Forest Classifier	X	0.79	0.87
	X_up	0.81	0.87
Multi-Layer Perceptron	X_scaled	0.79	0.87
	X_scaled_up	0.78	0.85
Gradient Boosting	X_scaled_up	0.75	0.84

Example Classification Report and Confusion Matrix



```
              precision    recall  f1-score   support

cytoskeleton       0.71      0.13      0.22       365
    membrane       0.92      0.86      0.89      3751
mitochondria       0.98      0.62      0.76       540
     nucleus       0.76      0.97      0.85      2871
    secreted       0.93      0.87      0.90       890

    accuracy                           0.85      8417
   macro avg       0.86      0.69      0.72      8417
weighted avg       0.86      0.85      0.84      8417

Confusion Matrix:
[[  48   43    1  273    0]
 [   6 3214    6  463   62]
 [   1   88  334  117    0]
 [  10   74    0 2786    1]
 [   3   63    1   45  778]]
```

Cross-Validation of Top 3 Algorithms

- Support Vector Machine, Random Forest Classifier, and Multi-Layer Perceptron chosen as candidate algorithms
- GridSearchCV used for hyperparameter tuning and cross-validation

Algorithm	**Average f1-score**	**Change in f1-score**
Support Vector Machine	0.81	+0.05
Random Forest Classifier	0.83	+0.03
Multi-Layer Perceptron	0.82	+0.03

Conclusions

- Accuracy (f1-score) tops out around 90% with these data
- Three classes routinely perform well (f1-score ~0.90):
  - Membrane
  - Nucleus
  - Secreted
- Mitochondrially localized proteins perform moderately well (f1-score ~0.75-0.85)
- Cytoskeletally localized proteins perform poorly (best f1-score ~0.70)

Future Directions

1. More data
2. ML stacking
3. Refinement of feature set
4. Using different classes or redefining how classes are labelled

Questions?