Hey ML,
what can you do for me?

Javier Pastorino – Ashis Kumer Biswas

Machine Learning Laboratory

Workshop on Machine Learning and Artificial Intelligence in Bioinformatics and Medical Informatics

IEEE AIKE 2020 - Irvine, CA

Virtual

Department of Computer Science and Engineering

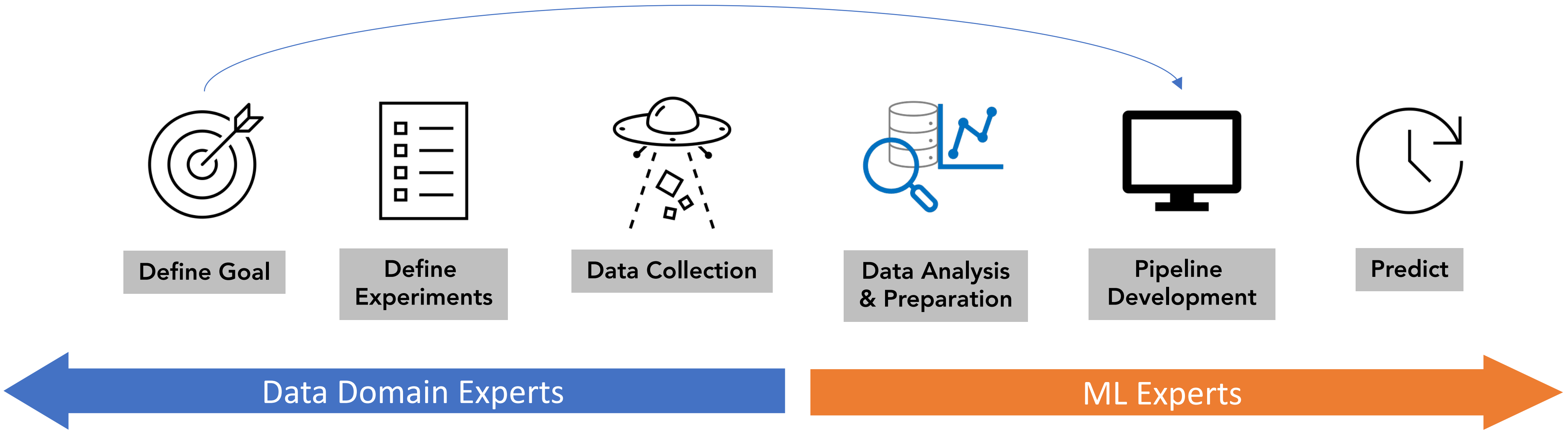
UNIVERSITY OF COLORADO

DENVER | ANSCHUTZ MEDICAL CAMPUS

Agenda

- Motivation
- Problem Description
- Limitation of Existing Methods
- Methodology
- Experiments – Datasets
- Results
- Limitations and Future Work

Motivation



- What other Interesting Problems can be solved with the same data?
- Lack of cross-domain expertise

Problem Description

- Given a raw, unprocessed dataset, identify machine learning tasks that can be predicted using the given dataset without knowledge of the ML algorithms.
- User Friendly Interface

Limitation of Existing Methods

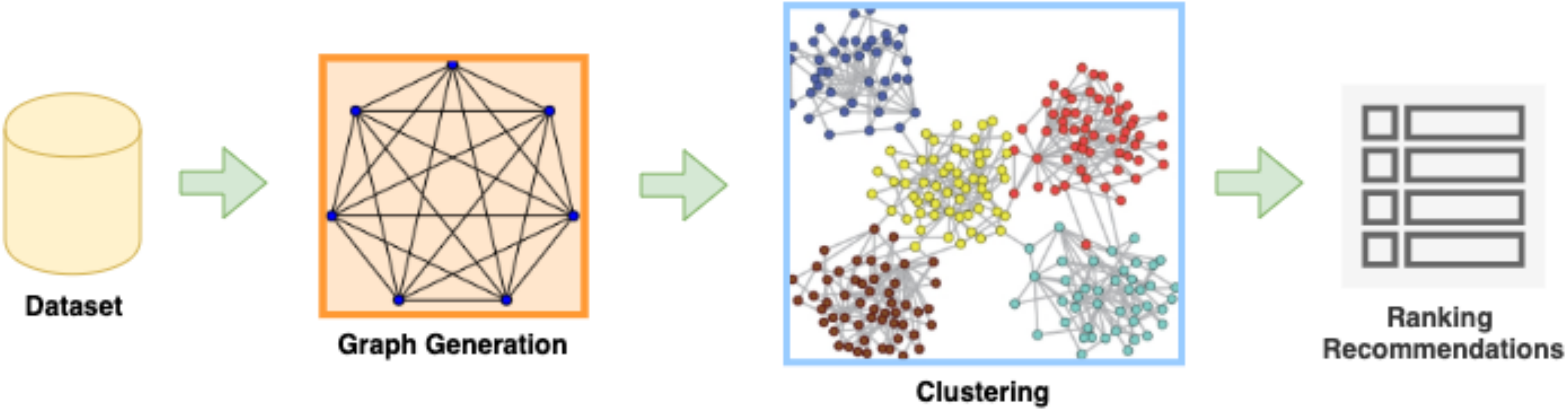
- Manual data analysis needed to assess if a particular task can be predicted by ML.

- Mutual Information [9] and Total Correlation [7]
  - Measure the interaction between two or more features in terms of information sharing.
  - Used for Hierarchical Clustering [6]

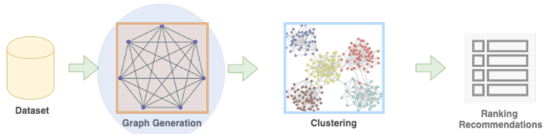
- Straight Forward Empirical Experiments Did Not Predict Target Features Successfully.

[6] T. Ferenci et al., "Using total correlation to discover related clusters of clinical chemistry parameters," in IEEE SISY. Subotica, Serbia: IEEE, 10 2014, pp. 49–54.
[7] S. Watanabe, "Information Theoretical Analysis of Multivariate Correlation," IBM Journal of Research and Development, vol. 4, no. 1, pp. 66–82, 1960.
[9] T. M. Cover et al., Elements of Information Theory. New York, NY: Wiley, 2006.

Methodology

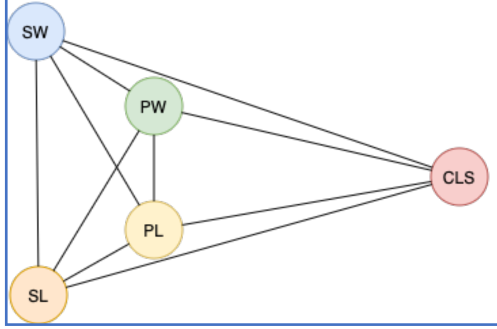


Graph Generation



- Represents the feature dimension as an undirected weighted graph
- **Vertices:** features in the dataset
- **Edges**: distance between features in terms of Mutual Information
  - amount of information a variable X contains about another variable Y
  - $I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \rightarrow [0,\infty)$
  - $w_{i,j} = 1 - I_{norm}(i;j)$

Graph Representation for *Iris* dataset



$$A(K)_{iris} = \begin{pmatrix} 0.0420 & 0.5834 & 0.4026 & 0.5507 & 0.8257 \\ 0.5834 & 0.2030 & 0.5582 & 0.6635 & 0.8985 \\ 0.4026 & 0.5582 & 0.0000 & 0.4618 & 0.7126 \\ 0.5507 & 0.6635 & 0.4618 & 0.1923 & 0.7147 \\ 0.8257 & 0.8985 & 0.7126 & 0.7147 & 0.6851 \end{pmatrix}$$

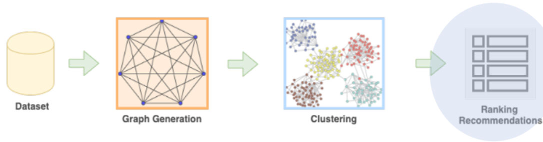
Methodology

Clustering

- **Groups Communities of Features that share the same amount of information (MI)**
- **Spectral Clustering**
- **Optimization of Number of Clusters**
  - Min #Clusters = 2; Max #Clusters=d
  - Grid search
  - **Silhouette Metric** to measure each configuration's performance:
    - Quantifies the **quality** of a set of clusters
    - Compares tightness and separations, inter/intra-cluster

Methodology

Ranking



- **Tentative Problem/Task**:
  - each computed cluster is considered a tentative predictive task

- **Candidate Task Definition:**
  - $\{\boldsymbol{x_1}, \dots, \boldsymbol{x_i}\} \rightarrow \{\boldsymbol{y_1}, \dots, \boldsymbol{y_j}\}$
  - Target: $\{y_1, \dots, y_j\}$
  - Input: $\{x_1, \dots, x_i\}$ *(the remainder features)*

- **Ranking Score**:
  - **Conditional Entropy** $\{\boldsymbol{x_1}, \dots, \boldsymbol{x_i}\} \rightarrow \{\boldsymbol{y_1}, \dots, \boldsymbol{y_j}\}$
    - Provides the amount of information input X provides to predict target Y
    - $H(Y|X) = -\sum_x \sum_y p(x,y) \log p(y|x)$

Methodology

Evaluation Strategy

- **top-$\boldsymbol{k}$ Precision and Recall**
  - ratio of ground truth retrieved in the ranking items
  - ratio of raking items that belong to ground truth

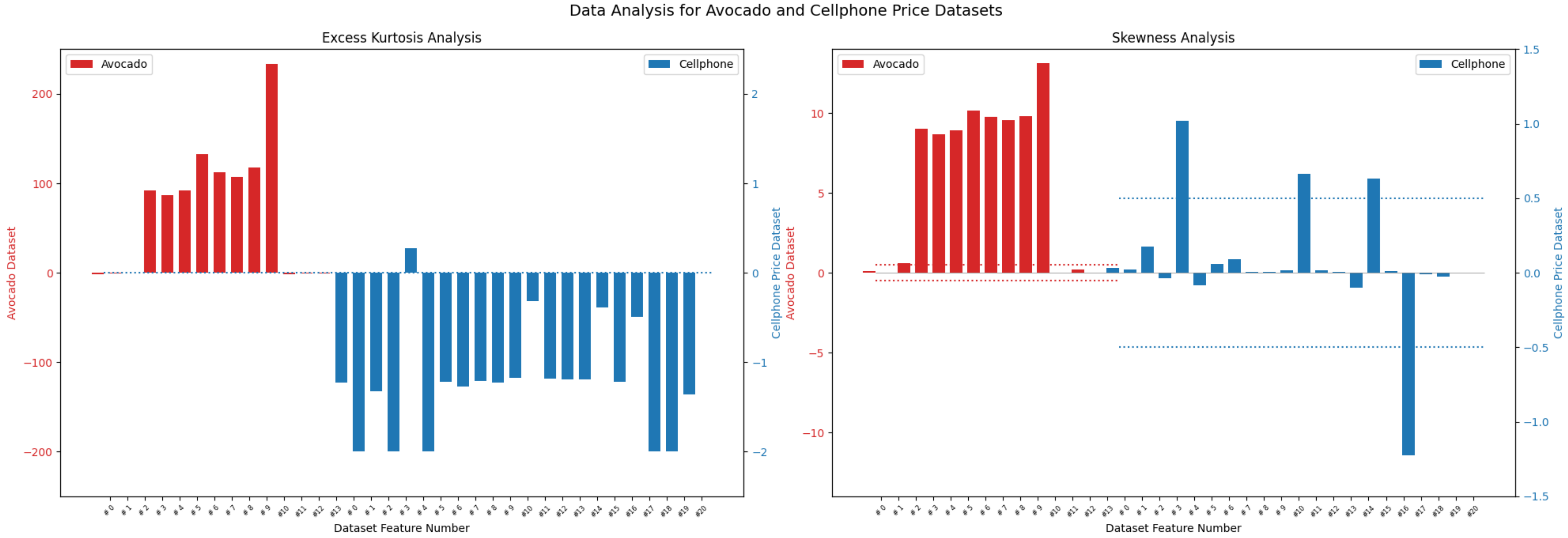
- **Mean Reciprocal Rank @$\boldsymbol{k}$**
  - Measures how well the target is placed in the ranking
  - $MMR(target) = \frac{1}{|k|}\sum_{i=1}^{|k|}\frac{1}{rank_i}$

Experiments – Datasets

	Dataset	# Features	# Data Samples
UCI	**Abalone**	9	4,177
	Auto MPG	8	406
	Bike Sharing	14	731
	Iris	5	150
	Naval Propulsion Plants	18	11,934
	Superconductivty	81	21,263
	Wine Quality	12	4,898
	Yacht Hydrodynamics	7	308
Kaggle	**Avocado**	14	18,250
	CellPrice	21	2,001

Experiments – Datasets

Distribution Analysis



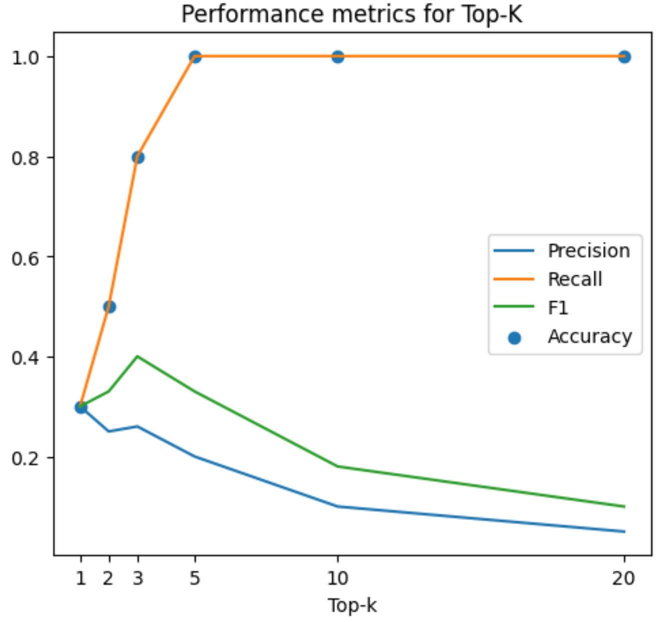
Data is not normal distributed: Shapiro-Wilk Test with *p-value* < 0.05

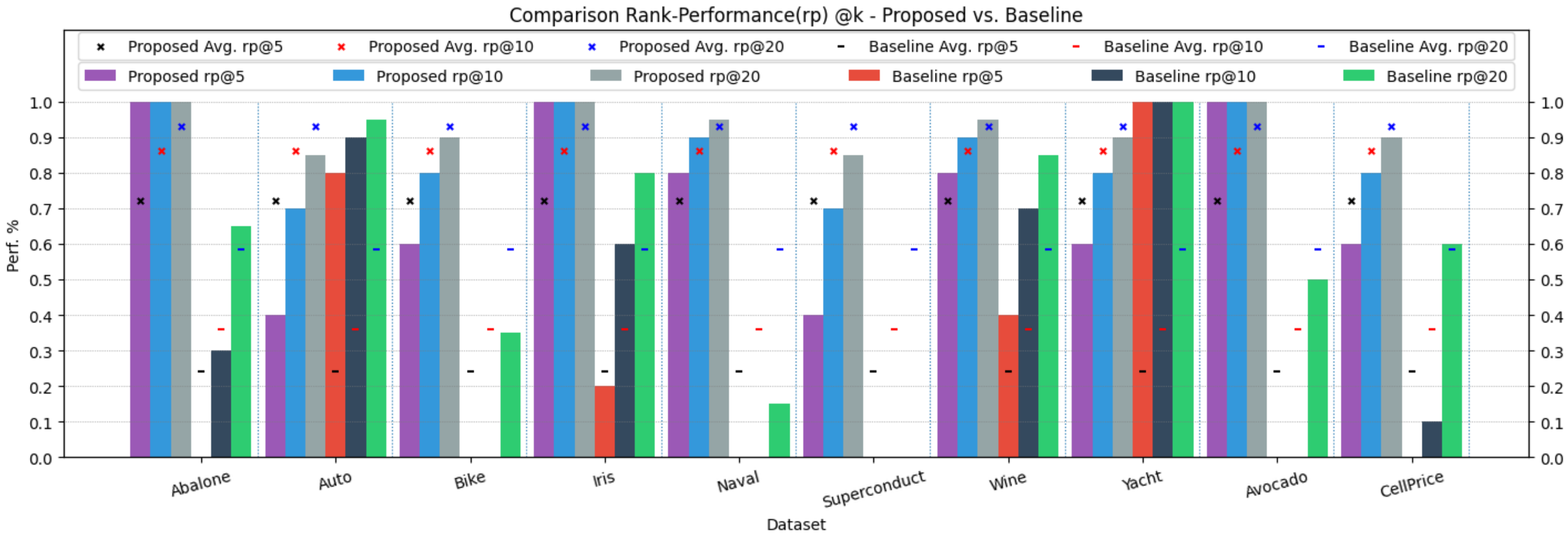
Results – Demo / Output Sample

```
+----------------------------------------------------------------+
|                           Main Menu                            |
+----------------------------------------------------------------+
|                       Available Datasets                       |
+----------------------------------------------------------------+
|  1 - abalone                                                   |
|  2 - auto-mpg                                                  |
|  3 - avocado                                                   |
|  4 - bike-sharing                                              |
|  5 - cellphone_price                                           |
|  6 - insurance car                                             |
|  7 - insurance health                                          |
|  8 - iris                                                      |
|  9 - naval                                                     |
| 10 - superconduct                                              |
| 11 - winequality-white                                         |
| 12 - yacht_hydrodynamics                                       |
+----------------------------------------------------------------+
████████████████████ Choose file to process [1-12]▷ 8
iris Dataset File Selected

+----------------------------------------------------------------+
|                Processing File: ../data/iris.csv               |
+----------------------------------------------------------------+
██████ Number of Features:  5                ██████
██████ Number of Examples: 149               ██████
██████ Optimized Number of Clusters:  5      ██████

+----------------------------------------------------------------+
|        Machine Learning Problem Recommendations for iris.csv   |
+----------------------------------------------------------------+
|With a 80.5% confidence, we recommend that ['target'] can be predicted by a machine learning task.
|With a  7.1% confidence, we recommend that ['sepal_width'] can be predicted by a machine learning task.
|With a  6.7% confidence, we recommend that ['petal_width'] can be predicted by a machine learning task.
|With a  3.1% confidence, we recommend that ['sepal_length'] can be predicted by a machine learning task.
|With a  2.5% confidence, we recommend that ['petal_length'] can be predicted by a machine learning task.
+----------------------------------------------------------------+
```

Demo recording available at:
https://github.com/jpastorino/heyml

Results – Performance





Limitations and Future Work

- **Limitations**
  - Information Graph Generation $O(d^2)$

- **Future Work**
  - Reach out to Data Domain Experts to validate other tasks.
  - Work to optimize the use computational resources.
  - Experiment with larger, complex datasets to improve the algorithm scalability.

Thank you!

Questions?