# CAMIL: Clustering and Assembly with Multiple Instance Learning for Phenotype Prediction

Nathan LaPierre (me), Mohammad Arifur Rahman, and Huzefa Rangwala
George Mason University, Computer Science Department

# Goal: Phenotype Prediction with Metagenomic Data

● Predict disease state ("phenotype") of patient based on metagenome data

● Qin et al. dataset* (2012)
        Type 2 Diabetes
        367 Chinese patients
        Stool samples

* J. Qin et al., "A metagenome-wide association study of gut microbiota in type 2
diabetes," Nature, vol. 490, no. 7418, pp. 55–60, 2012.
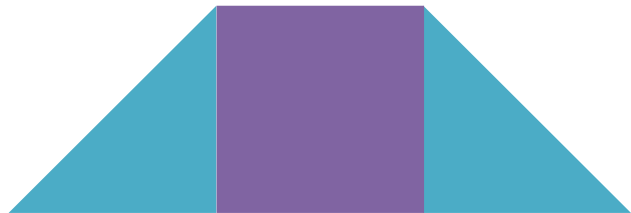
Background: Multiple Instance Learning

- Labeled bags containing unlabeled instances
  - In this context: labeled patients and unlabeled reads

- Instance, Bag, and Embedded Spaces (Amores 2013)*
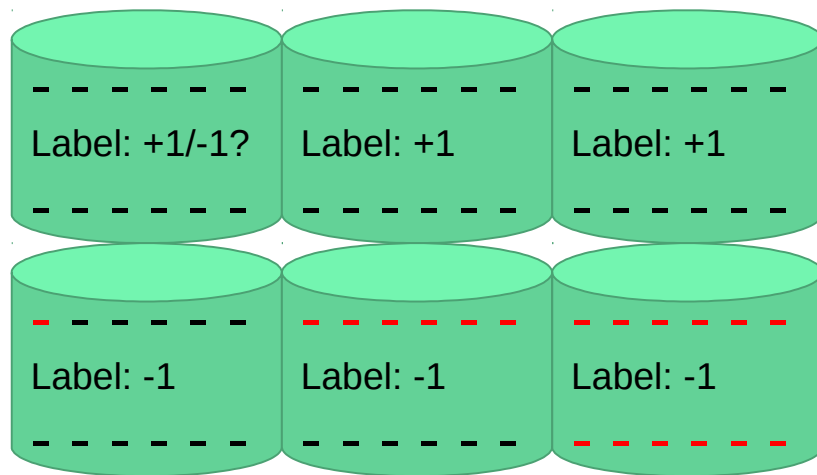
Label: +1/-1?	Label: +1	Label: +1
Label: -1	Label: -1	Label: -1

* J. Amores, "Multiple instance classification: Review, taxonomy and comparative study," Artificial Intelligence, vol. 201, no. 1, pp. 81–105, 2013.
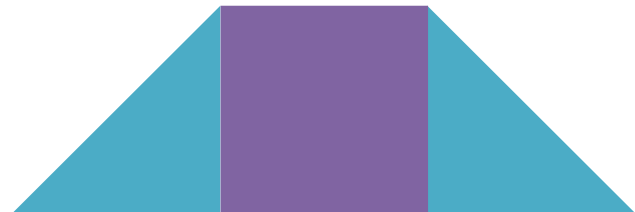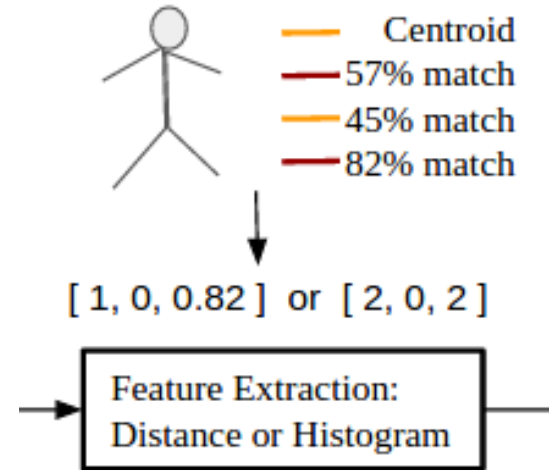
Pipeline
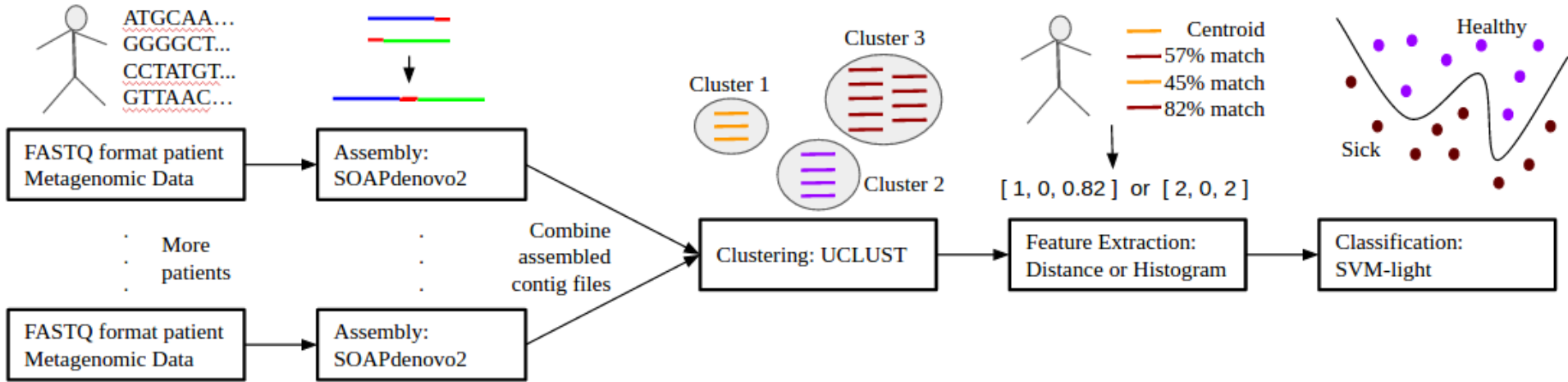
FASTQ Data → Assembly → Clustering → Feature Extraction → SVM Classification
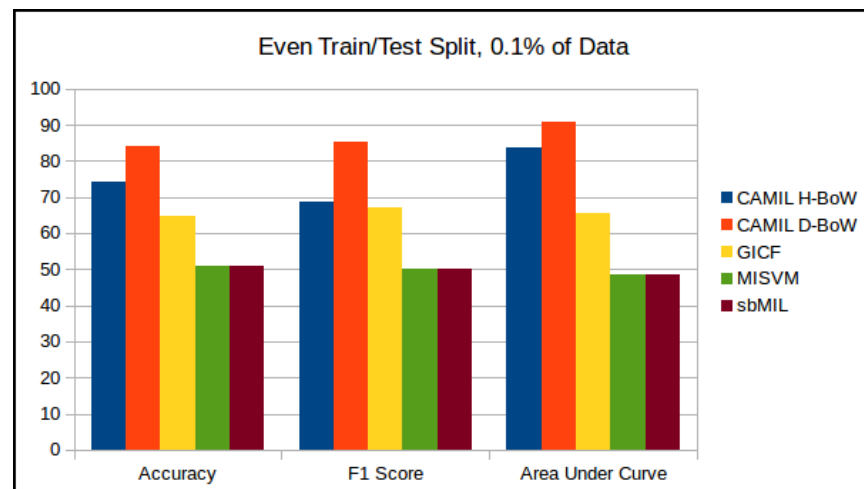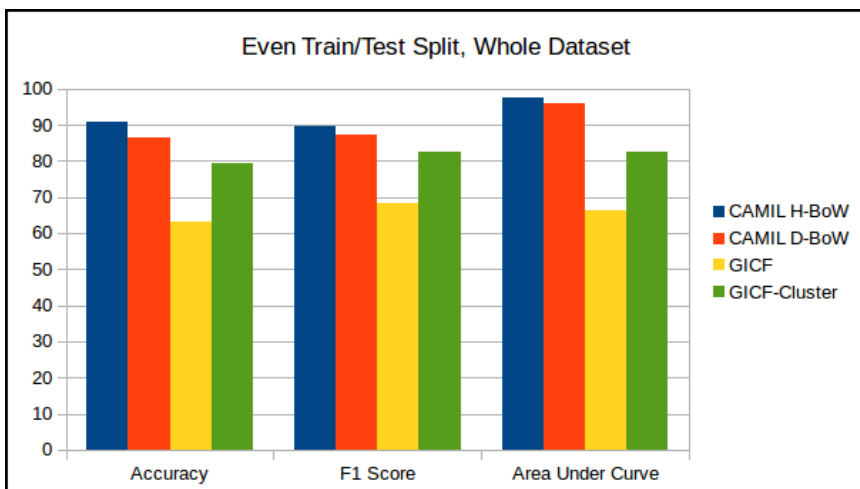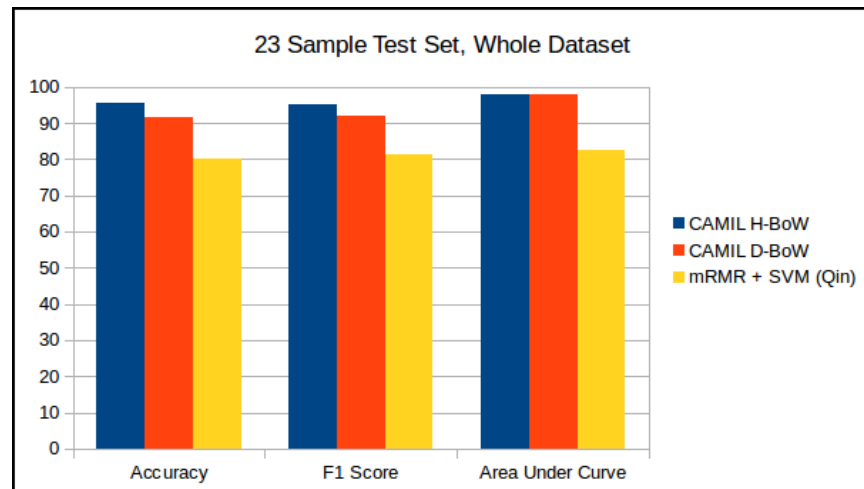
CAMIL Feature Extraction

- MIL: "Bag of Words" (Amores)*

- Histogram Bag of Words (H-BoW)

- Distance Bag of Words (D-BoW)

— Centroid
—57% match
—45% match
—82% match

[1, 0, 0.82] or [2, 0, 2]

Feature Extraction:
Distance or Histogram

* J. Amores, "Multiple instance classification: Review, taxonomy and comparative study," Artificial Intelligence, vol. 201, no. 1, pp. 81–105, 2013.

Results

Conclusion and Future Work

- MIL can be an effective approach towards phenotype prediction

- CAMIL is a general example of this kind of method

- Future: different clustering & assembly algorithms, different MIL-based feature extraction methods, instance labels