

## MOTIVATION

The intuitive definition of an outlier would be “an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism”. [Hawkins, 1980]

An outlying observation, or “outlier,” is one that appears to deviate markedly from other members of the sample in which it occurs. [Grubbs, 1969]

An observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data. [Barnett and Lewis, 1994]

In data mining research many outlier models and variants for improved efficiency have been developed, but each model has strengths and weaknesses. The combination of models for outlier detection is as promising as ensembles are in classification or clustering but did not gain much attention so far.

## RELATED WORK

Feature bagging combines outlier scores learned on different subsets of attributes [Lazarevic and Kumar, 2005]. The problem for such combinations is the comparability of scores that are learned, e.g., in spaces of different dimensionality. Subsequent research studied the problem of score normalization [Gao and Tan, 2006; Nguyen et al., 2010; Kriegel et al., 2011]. Greedy ensemble combines base learners that are as diverse as possible [Schubert et al., 2012].

Methods for inducing diversity in these studies have been using different subspaces, using different parameters, or using different base methods.

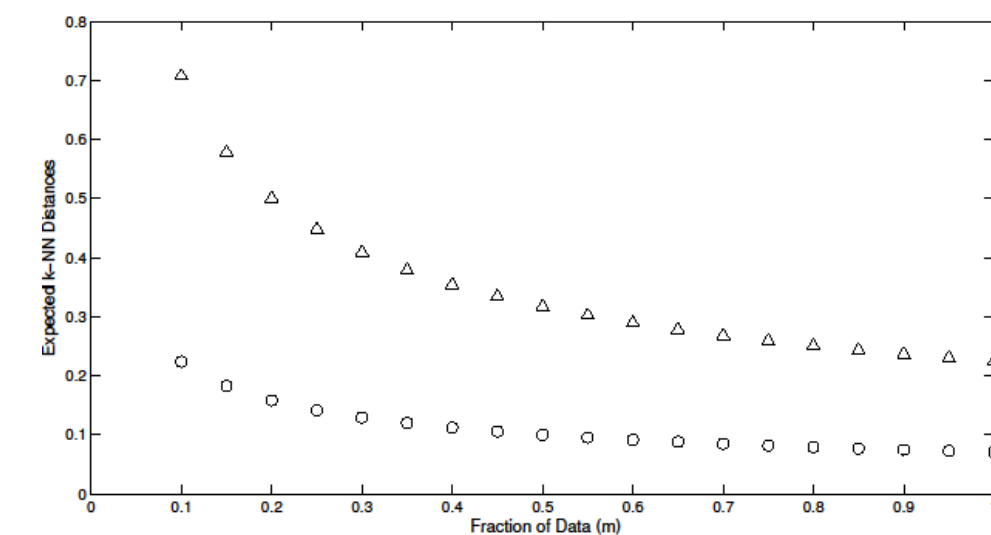
Theoretical insights so far are restricted to an empirical study of the impact of diversity of models [Schubert et al., 2012], and a position paper describing existing algorithmic patterns in two pairs of categories (sequential vs. independent learning of models, data centered vs. model centered ensembles) [Aggarwal, 2012].

One central question remained unanswered: Why should, what has a clear theoretical background in supervised learning, also work in unsupervised outlier detection?

## ENSEMBLES

Outlier detection methods usually rely on density estimates, probably committing some error in the estimate. By averaging outlier scores, we can talk about the expected error and study its impact on the resulting ranking.

As we do not need to preserve the “ideal ranking” (that is, due to the true but unknown underlying probability density distribution, describing the process that generated the data sample) but only to separate outliers from inliers, it turns out that some error can actually be helpful to increase the gap between outliers and inliers.



Expected  $kNN$  distances in volumes of different densities (1000m – circles, 100m – triangles)

## SUBSAMPLING

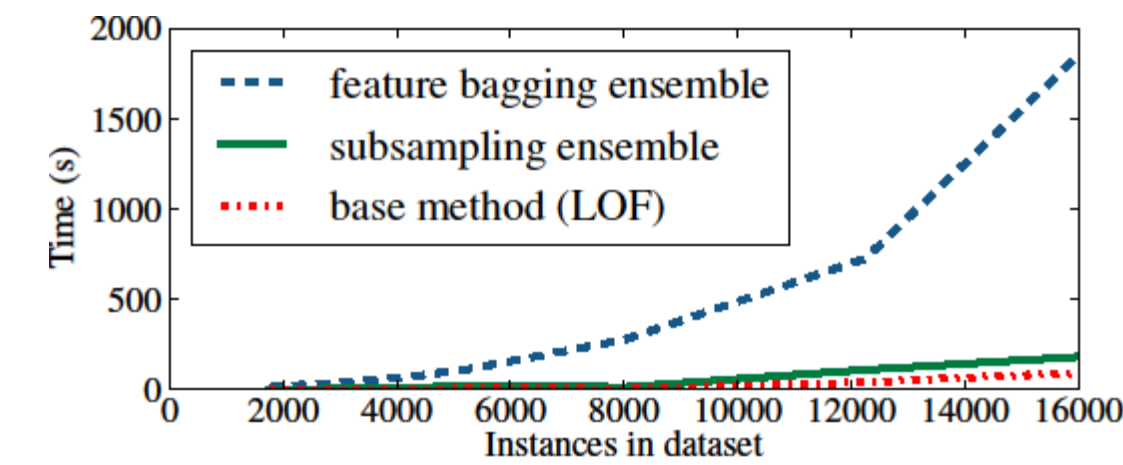
The error induced by subsampling is particularly helpful. While the relative contrast between areas of different densities remains constant, the absolute contrast between low-density and high-density areas is increased.

As an additional benefit, ensembles based on subsampling are efficient: The typical complexity of unsupervised outlier detection methods is  $O(n^2)$ , due to  $kNN$  queries. A common ensemble with  $s$  members would be in  $O(n^2s)$ .

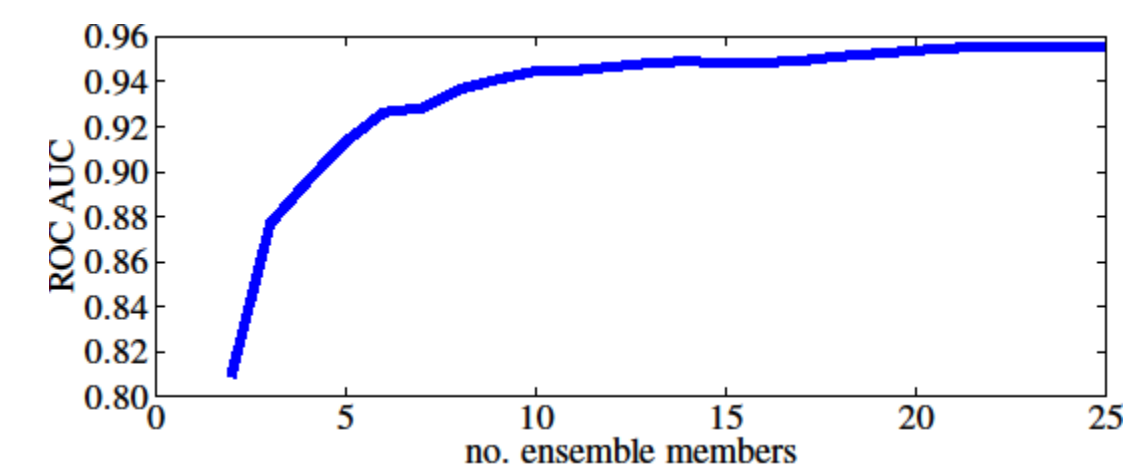
The subsampling ensemble computes the  $kNN$  query for each data object ( $n$ ) only on a subsample of the data set ( $mn$ , with  $0 < m < 1$ ). This is repeated  $s$  times, resulting in a complexity of  $O(n \cdot mn \cdot s)$ .

For example, with a sample rate of 10% and an ensemble size of 10 members, the ensemble would be as efficient as a single base learner, while a standard ensemble would require a runtime of ten times the runtime of the base learner.

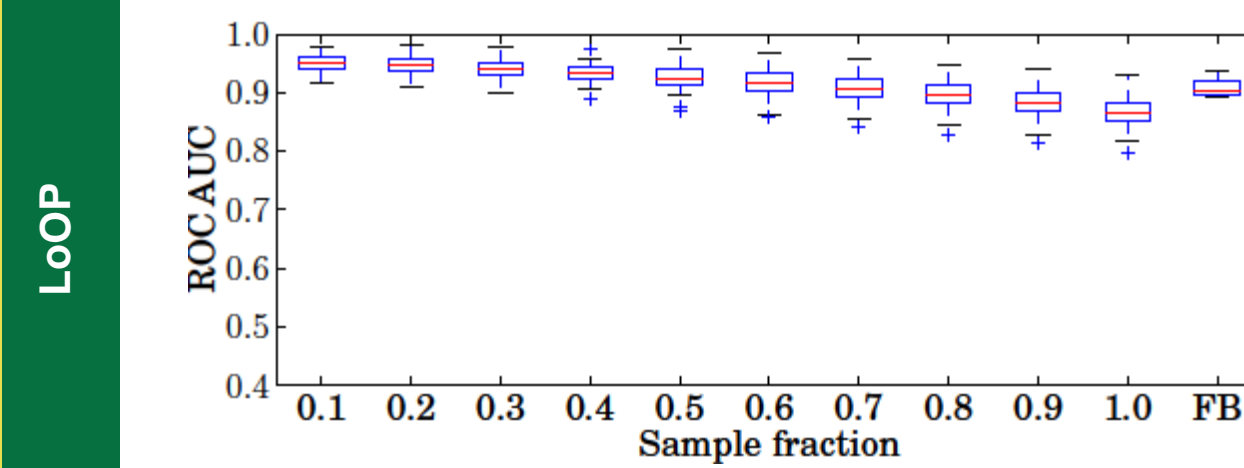
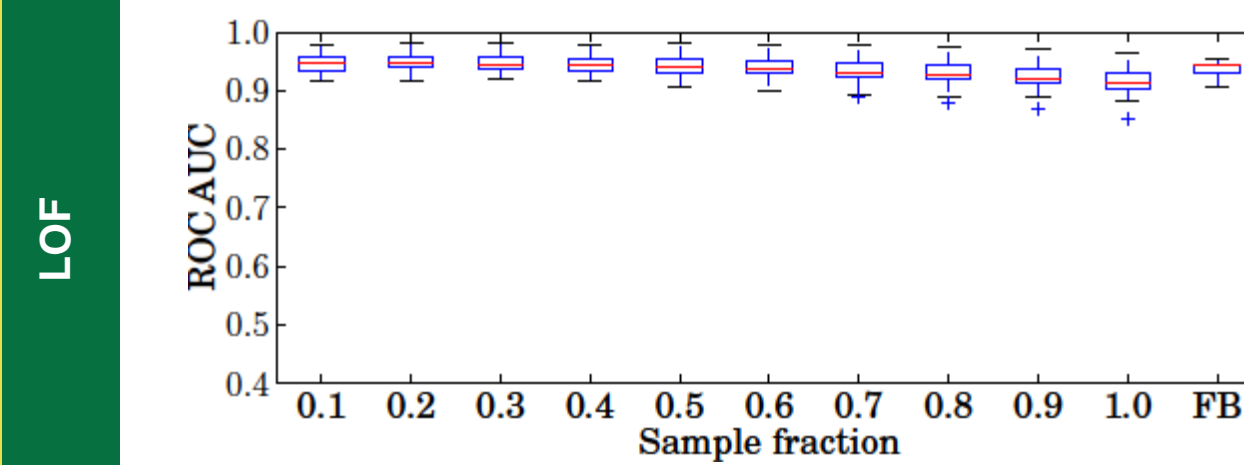
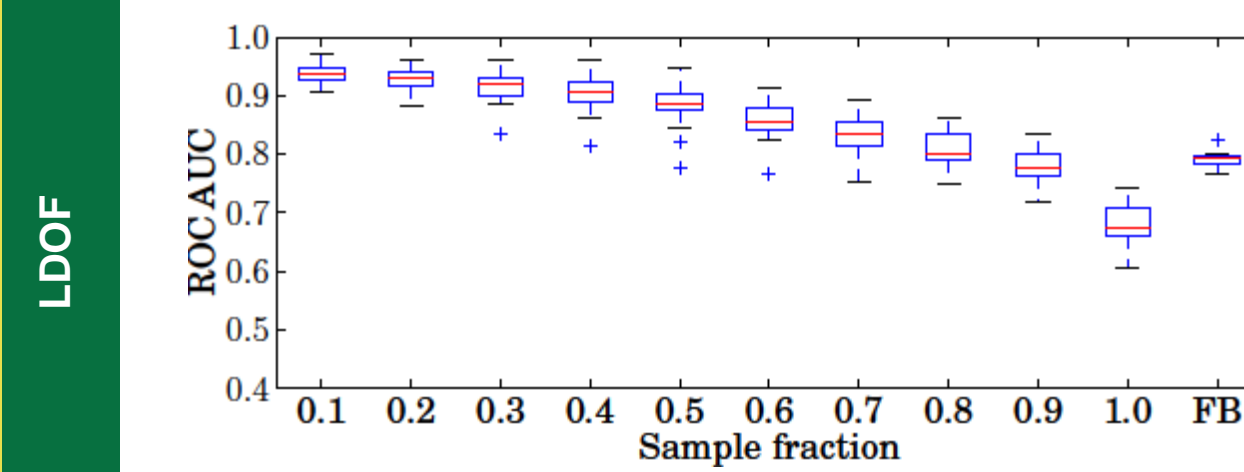
## EVALUATION



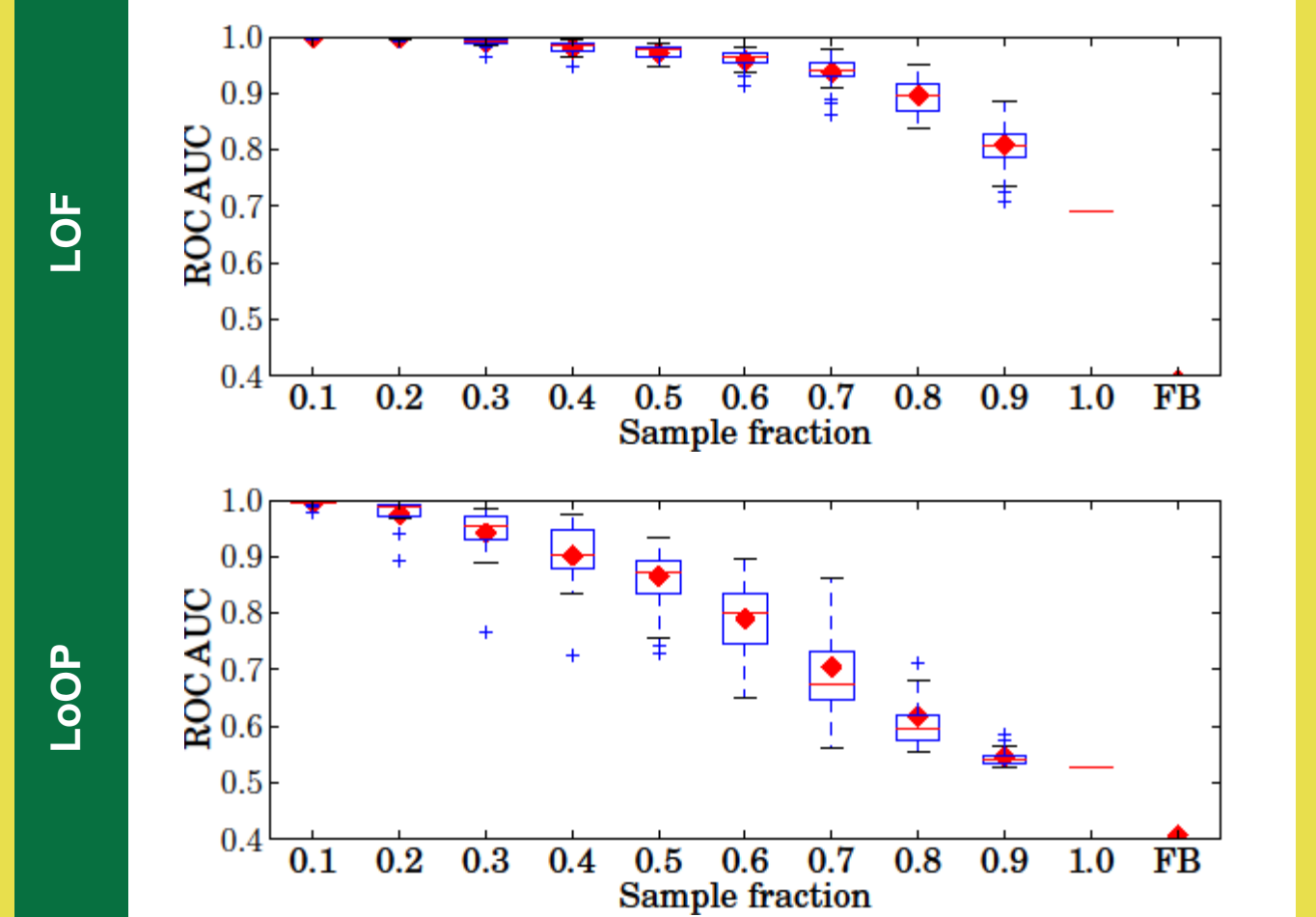
Scalability



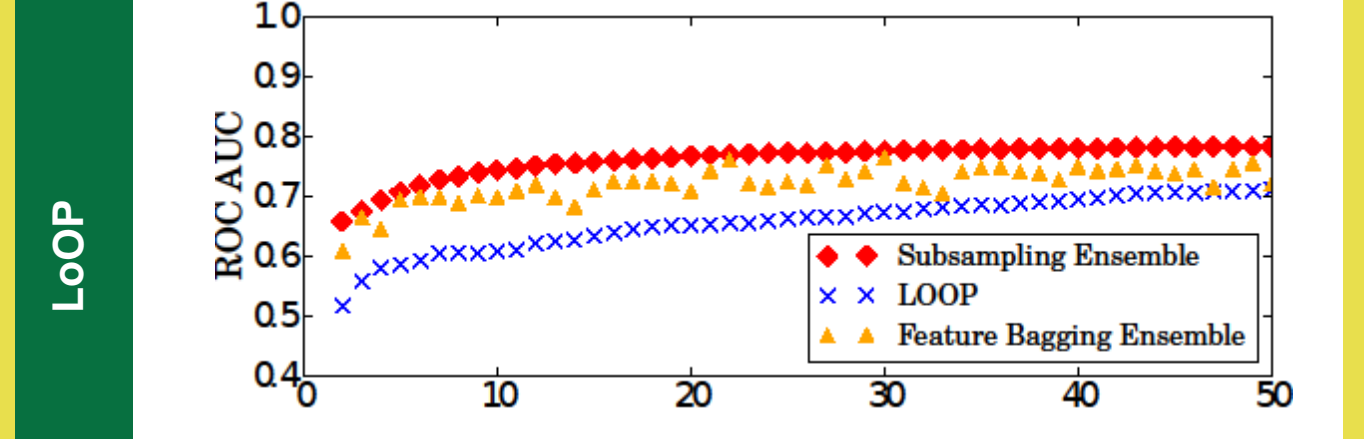
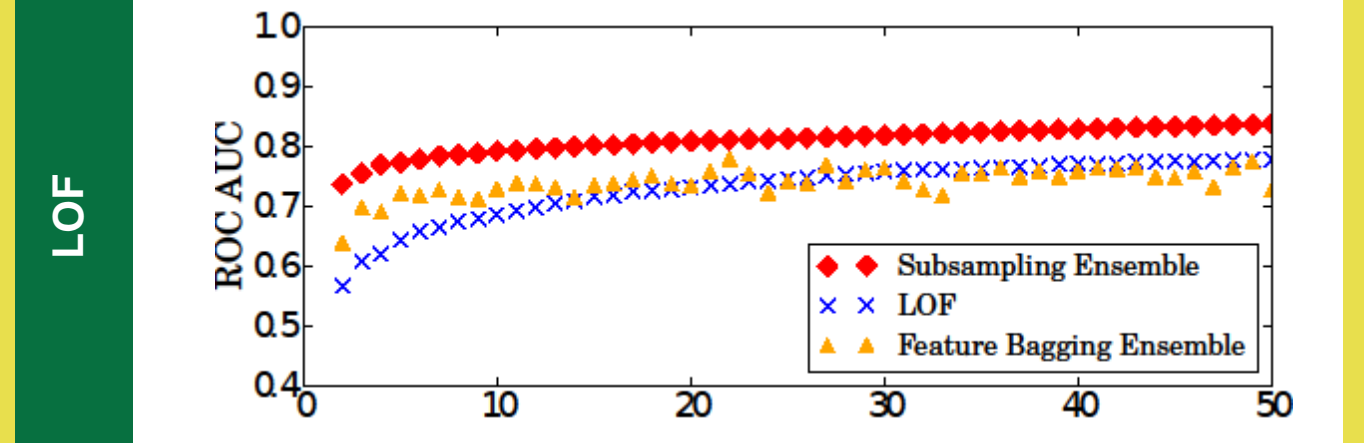
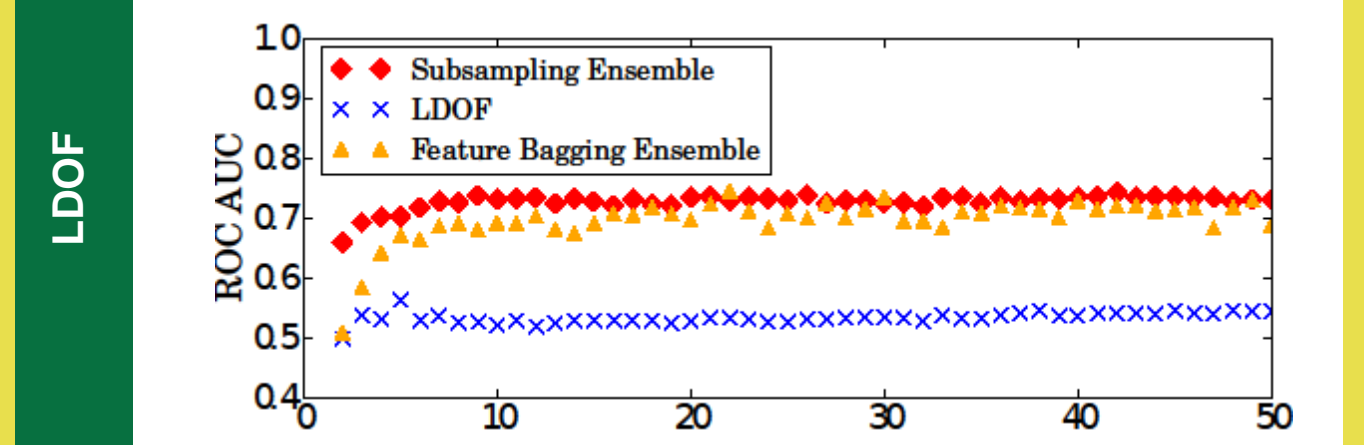
Quality with increasing ensemble size



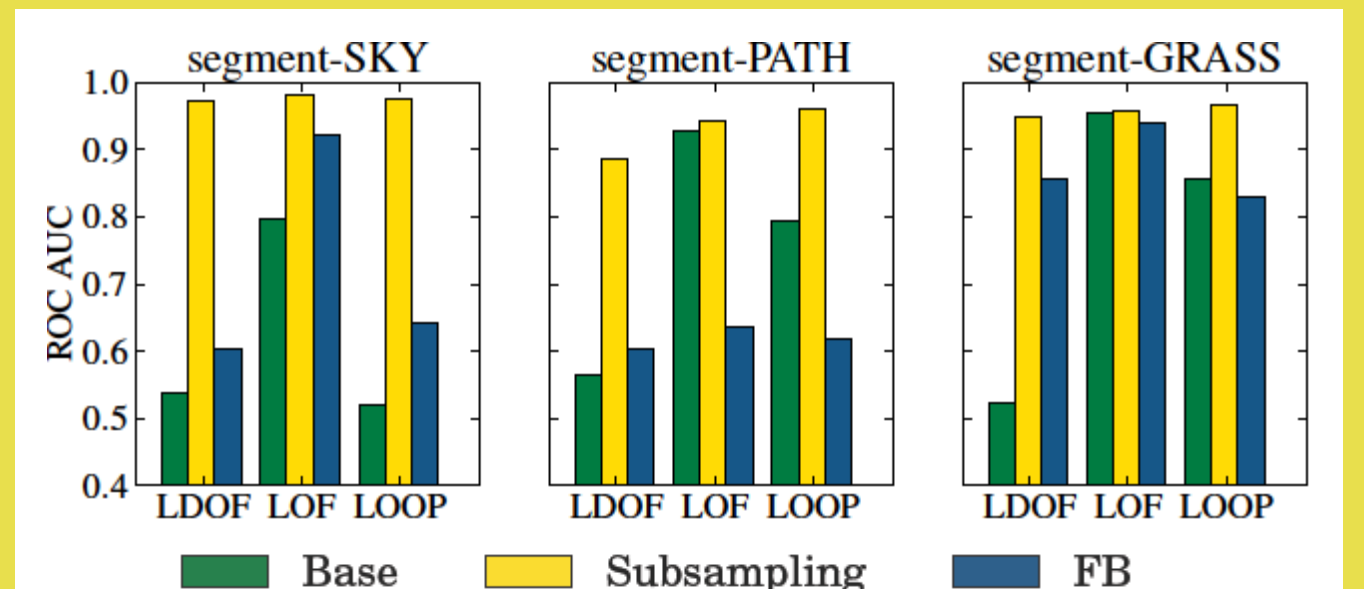
ROC AUC with varying sample sizes, distribution over 30 synthetic datasets



ROC AUC with varying sample sizes on dataset satimage-2



ROC AUC with varying  $k$  on dataset waveform



ROC AUC for all methods,  $k=20$ , variants of SEGMENT data