Characteristics of Datasets:

1. Number of objects
2. Distribution (i.e., shape) of
3. Instance distribution: Object width/length
4. Number of instances
5. Instance probabilities

Schemes:

1. Baseline 1: Sequential with pruning
2. Baseline 2: Parallel Naïve partition, no pruning (i.e., pruning 1.1), no labeling (i.e., labeling 1.2)
3. Pruning 1.1 (scheme 1)
4. Labeling 1.2 (scheme 2)
5. Pruning + Labeling (scheme 3 – put it all together)

Experiment 1:

* What to compare? baseline 1 and scheme 5
* Purpose: Overall performance.
* Compare Baseline 1 with scheme 3 to evaluate the speedup

Experiment 2:

* What to compare? Baseline 2, scheme 3, scheme 4, and scheme 5.
* Purpose: compare the pruning and labeling techniques

Experiment 3:

* Purpose: impact of datasets on our algorithm.
* Datasets: uniform, anti-correlated, correlated.

Experiment 4:

* Purpose: data scalability
* Datasets: increase the number of dimensions
* Datasets: increase the number of objects
* Datasets: increase the number of instance

Experiment 5:

* Purpose: algorithm’s scalability
* Hadoop cluster: increase the number of data nodes