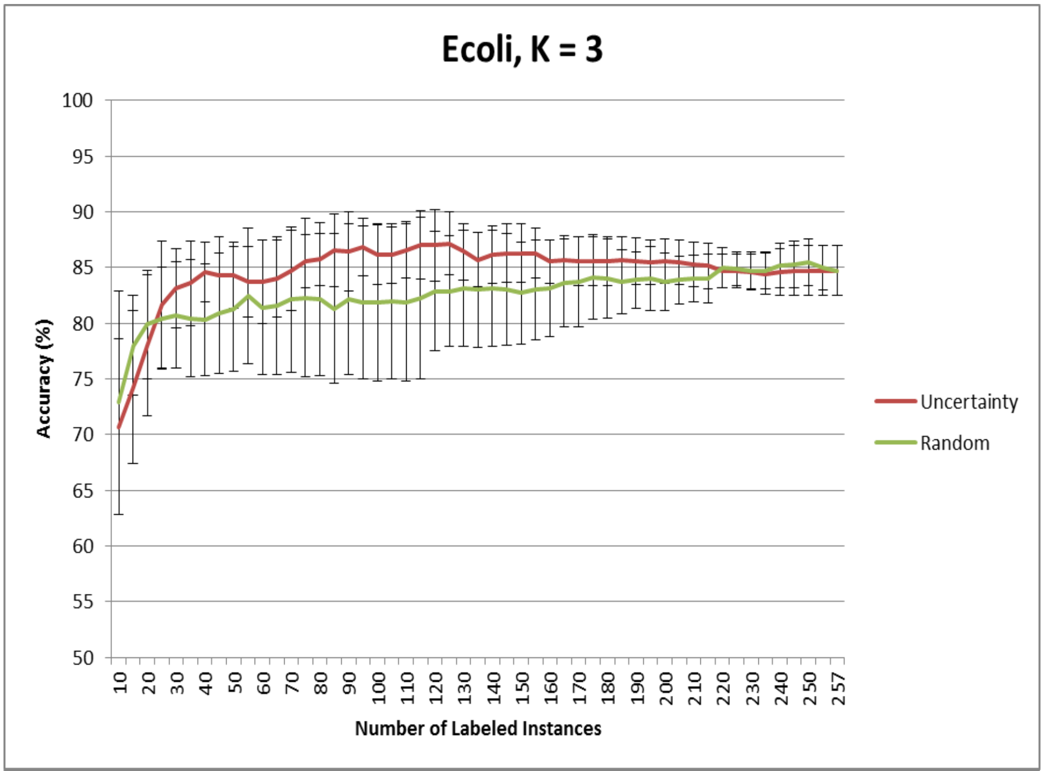
Benjamin Limoges

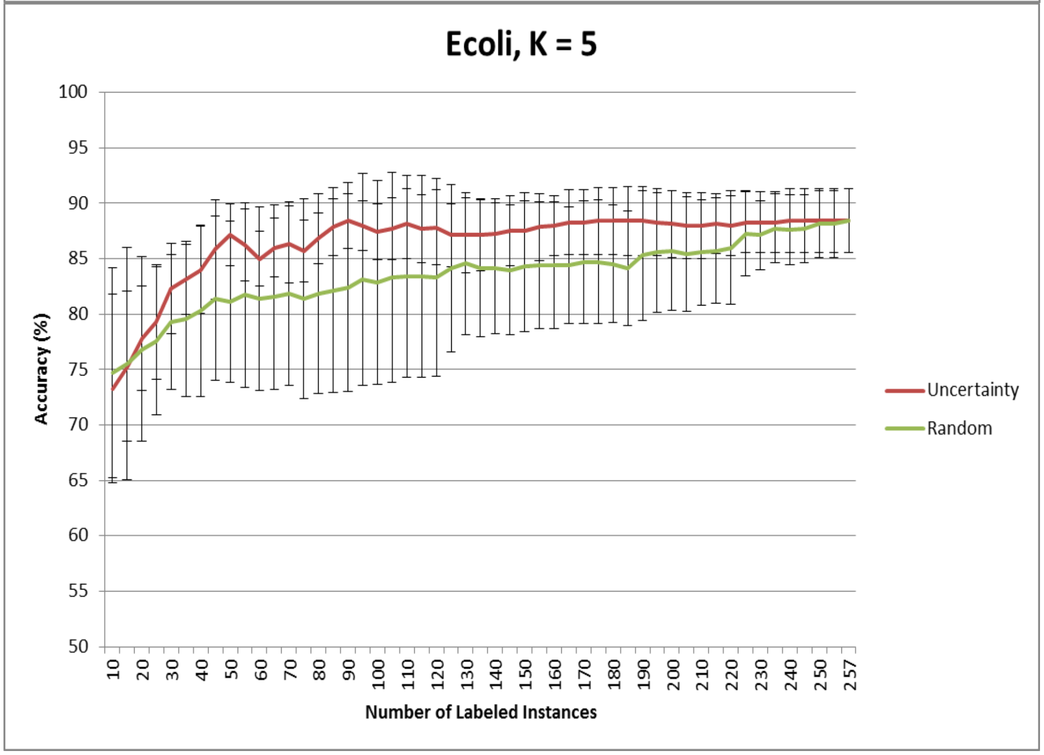
For both datasets, the uncertainty sampling started off slightly worse than random sampling, then uncertainty sampling became much better for a period, followed by tapering off when the two converged. Uncertainty sampling starts off worse most likely by random chance – after two requests (of 5 each) uncertainty sampling has some performance gains at k = 5 for both datasets. This is because the fuzziest points are being removed (and labeled) allowing the classifier to have a sharp increase in accuracy. The random does not cause there to be as sharp of an accuracy gain because it is not getting rid of this "fuzziness."

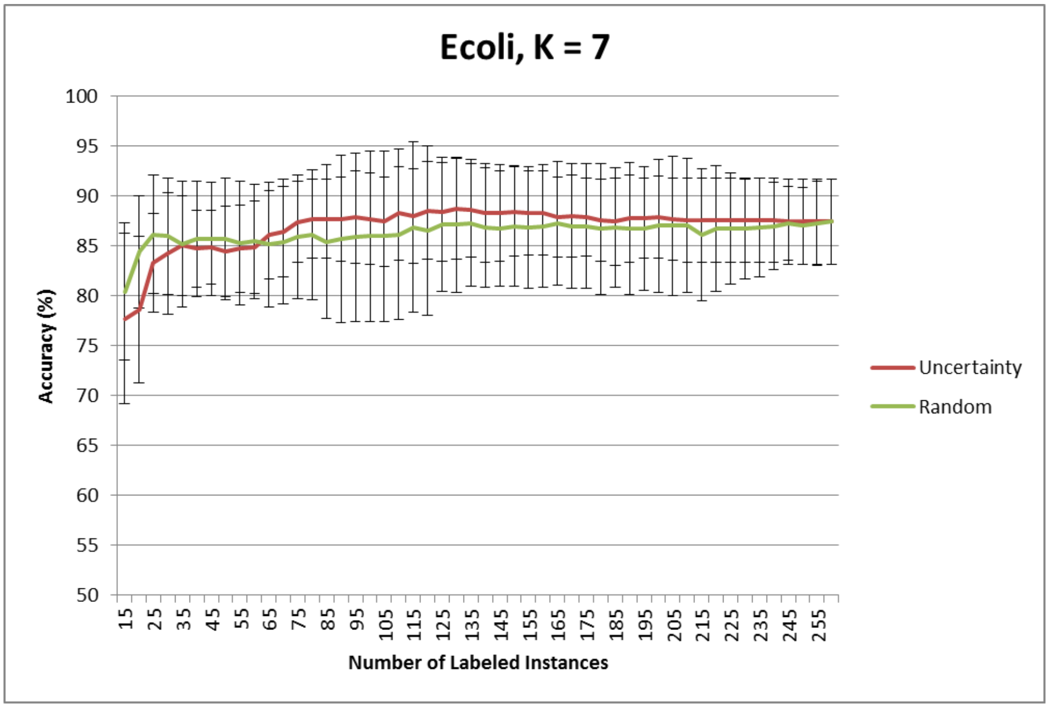
Across different runs of k the two datasets acted differently. For all values of k, the accuracy gain is relatively constant for the Ecoli dataset. There is an initial boost, which tapers off. The random slowly but surely catches up. For the Ionosphere dataset, k=7 seems to get the largest gain in accuracy. The higher value ks are also dramatically improved under uncertainty sampling. I believe that this increase for larger values of k (in the Ionosphere dataset) was because the points are clustered together in several of the feature spaces (this is a conjecture and not verified). By having "too many" neighbors, the learner was needlessly confusing itself. So labeling data defined this boundary really quickly, giving the sharp performance gain.

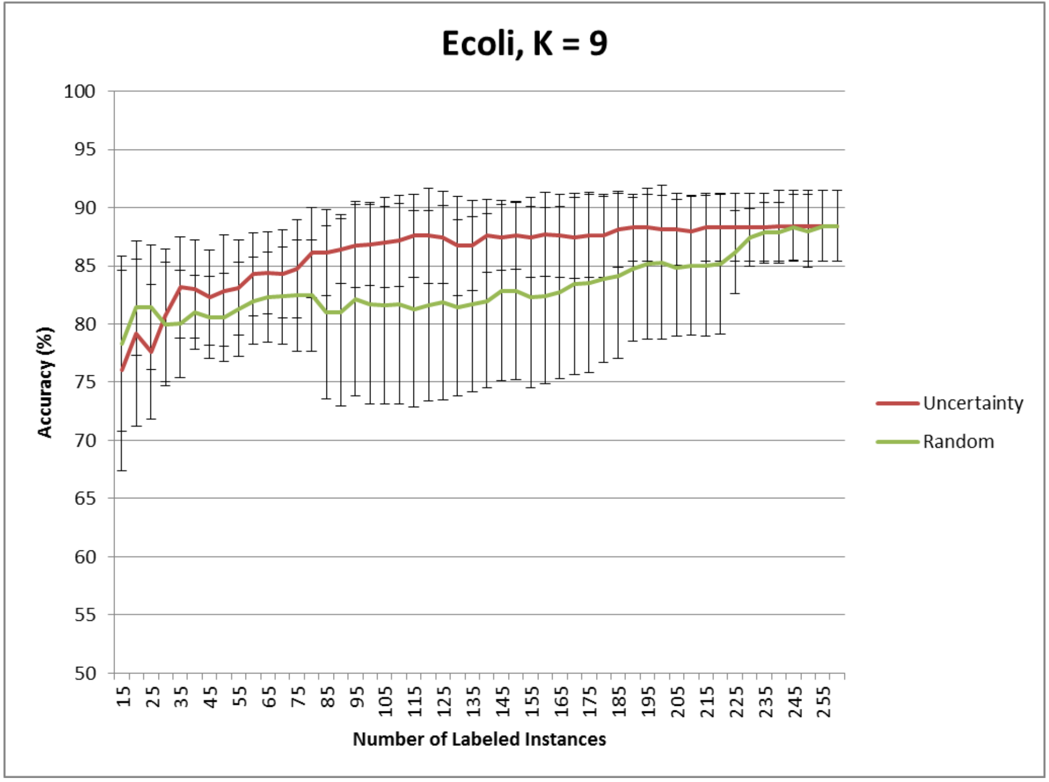
Ionosphere got a larger gain than Ecoli in performance from uncertainty sampling. I have two hypotheses for why this is true. The first is that Ionosphere is a binary class problem, whereas Ecoli is a 5 class problem. This means that with uncertainty sampling, for the Ionosphere dataset, the learner could quickly remove uncertain examples to get a large boost in accuracy. For Ecoli, the multiclass problem made it that the learner had to balance the distinction between 5 classes instead of two. So it may "spend" its 5 additional labels on clearing one boundry without improving performance with the other boundaries. Another issue with the Ecoli dataset was the minority class issue with class 7. This made it hard to pin down – the learner may have been constantly getting this class wrong until most of the datapoints were labeled.

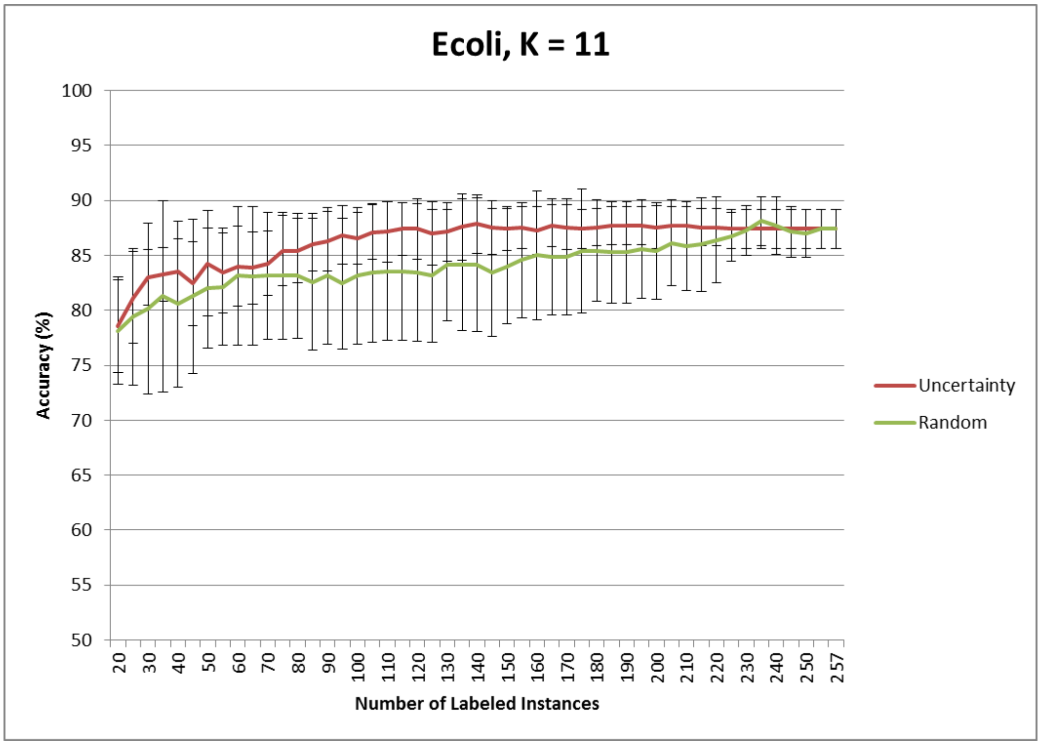
The convergence of random sampling and uncertainty sampling as more labels were added is a because the pool of unlabeled instances was becoming increasingly small. If there were an infinite pool of unlabeled instances to draw from, then the convergence would be much slower; by the end of the finite pool of unlabeled instances the two methods were drawing extremely similar sets of points to be labeled. In the asymptotic case with infinite data, the two would converge simply because the boundaries would become clearer and clearer, since the random draws will eventually include the "key" points that define the boundary between the classes – uncertainty sampling gets there quicker by defining the boundary immediately, whereas random sampling fills in the entire space.

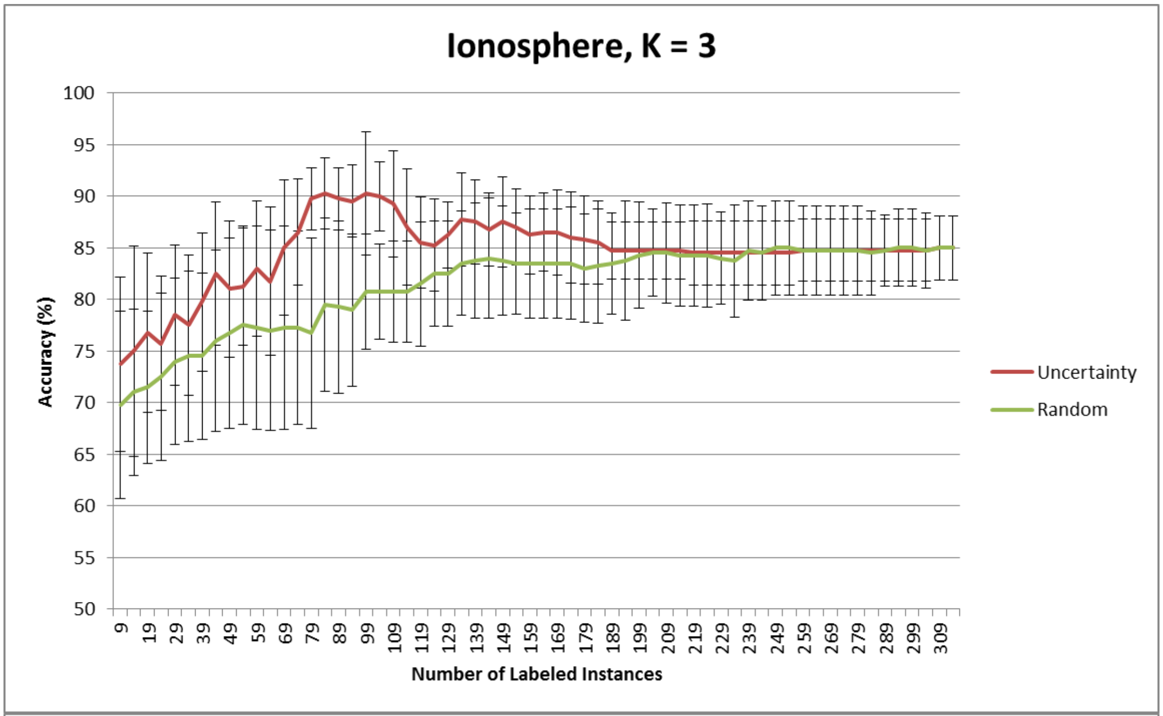

Ecoli, K = 3
100
95
90
85
80
75
70
65
60
55
50
Accuracy (%)
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
200
210
220
230
240
250
257
Number of Labeled Instances
Uncertainty
Random

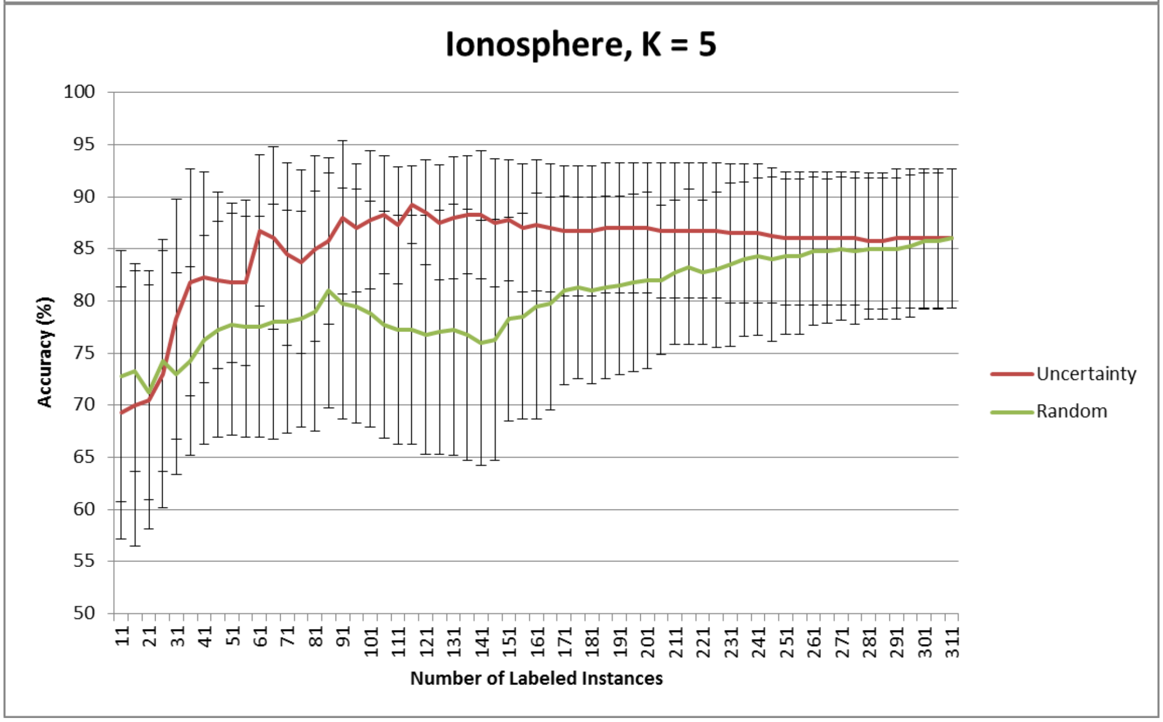

Ecoli, K = 5
100
95
90
85
80
75
70
65
60
55
50
Accuracy (%)
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
200
210
220
230
240
250
257
Number of Labeled Instances
Uncertainty
Random

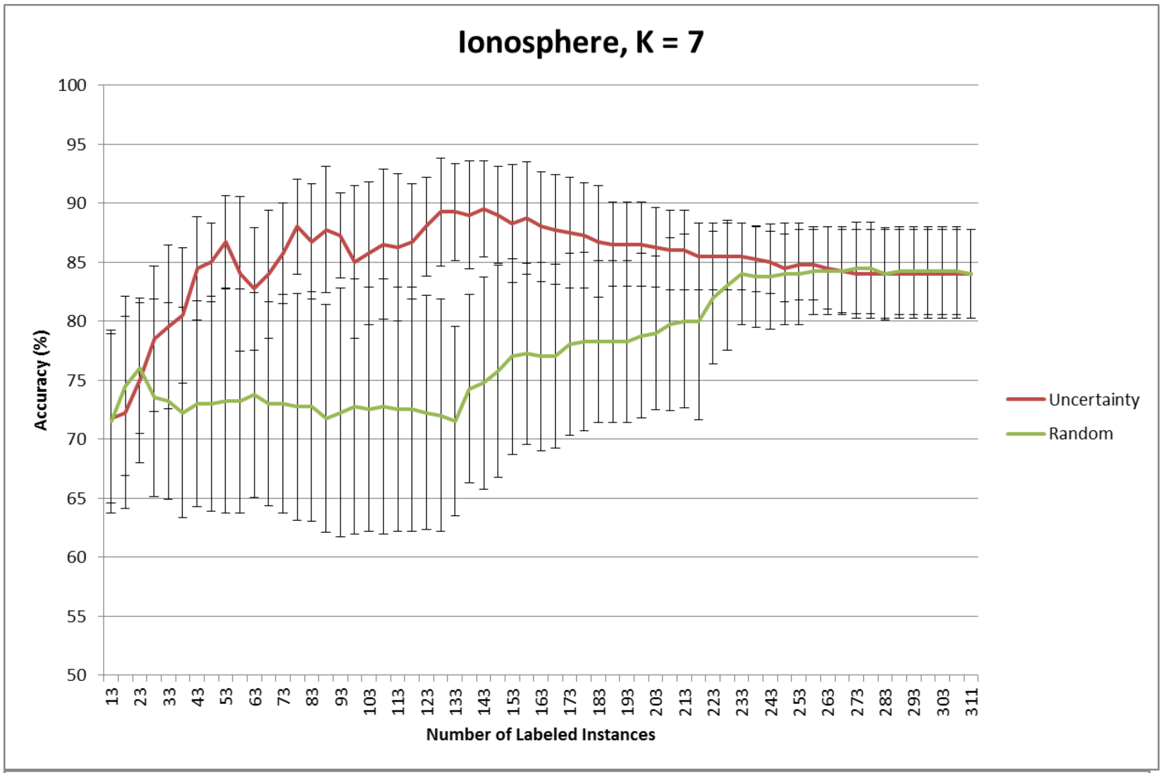

Ecoli, K = 7
Accuracy (%)
100
95
90
85
80
75
70
65
60
55
50
15
25
35
45
55
65
75
85
95
105
115
125
135
145
155
165
175
185
195
205
215
225
235
245
255
Number of Labeled Instances
Uncertainty
Random

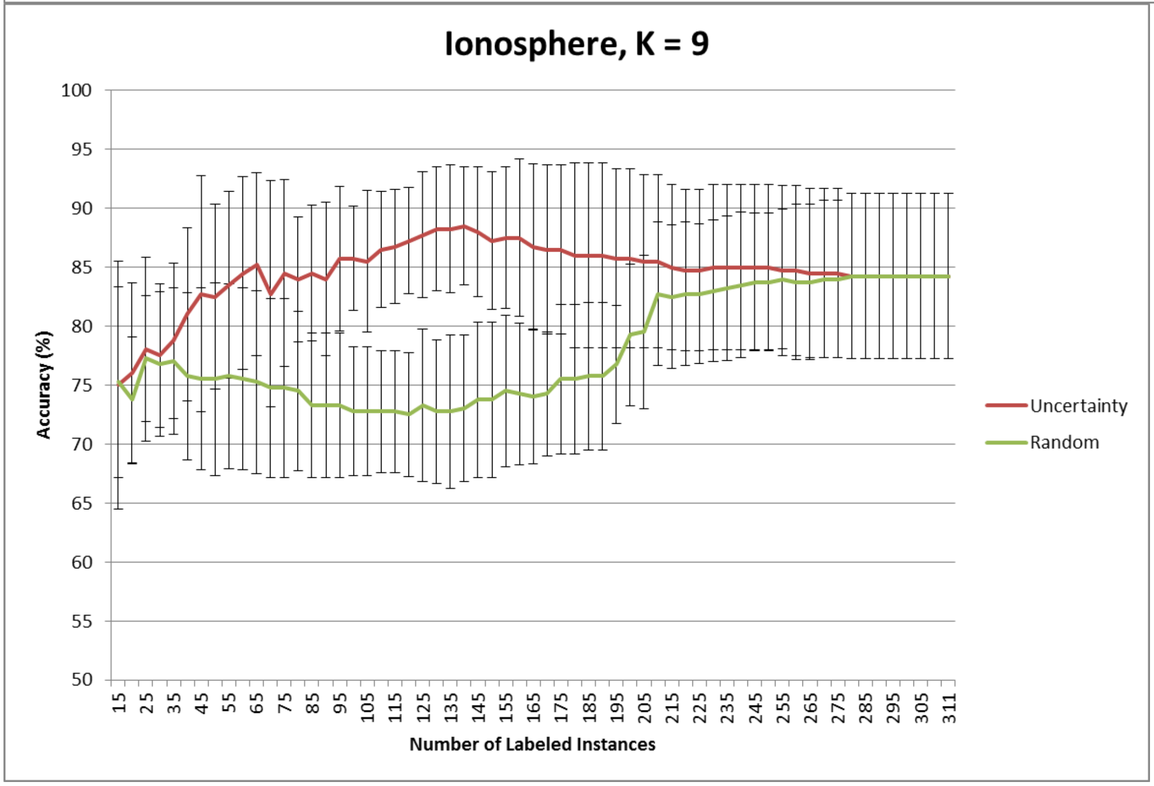

Ecoli, K = 9
Accuracy (%)
100
95
90
85
80
75
70
65
60
55
50
15
25
35
45
55
65
75
85
95
105
115
125
135
145
155
165
175
185
195
205
215
225
235
245
255
Number of Labeled Instances
Uncertainty
Random

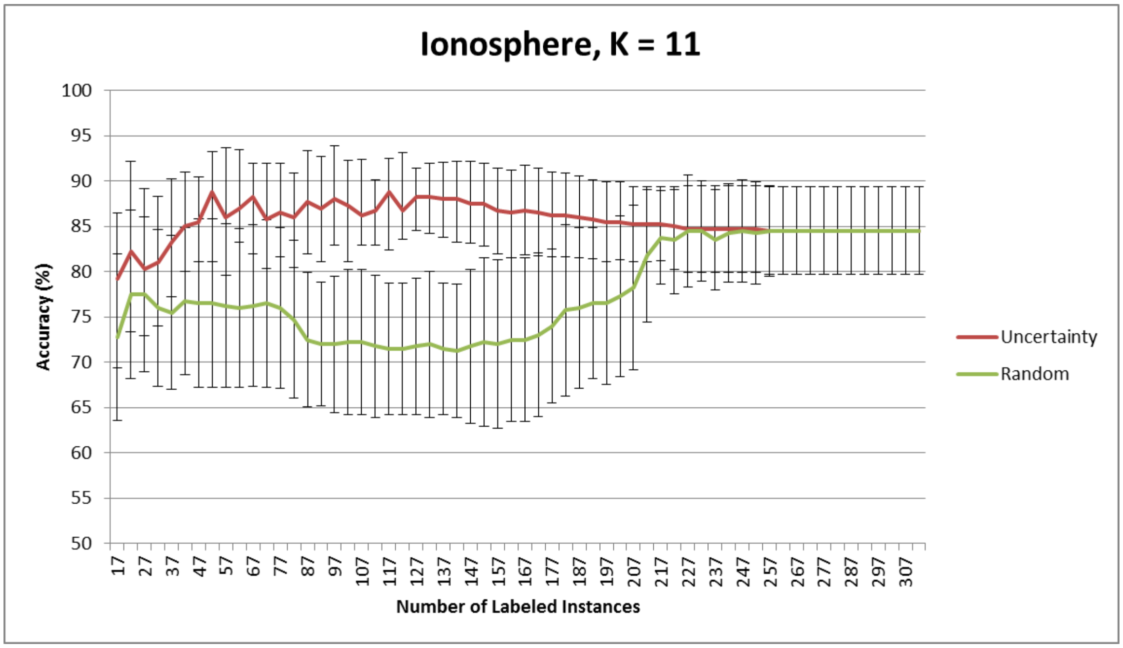

Ecoli, K = 11
Accuracy (%)
100
95
90
85
80
75
70
65
60
55
50
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
200
210
220
230
240
250
257
Number of Labeled Instances
Uncertainty
Random


Ionosphere, K = 3
Accuracy (%)
100
95
90
85
80
75
70
65
60
55
50
9
19
29
39
49
59
69
79
89
99
109
119
129
139
149
159
169
179
189
199
209
219
229
239
249
259
269
279
289
299
309
Number of Labeled Instances
Uncertainty
Random


Ionosphere, K = 5
Accuracy (%)
100
95
90
85
80
75
70
65
60
55
50
11
21
31
41
51
61
71
81
91
101
111
121
131
141
151
161
171
181
191
201
211
221
231
241
251
261
271
281
291
301
311
Number of Labeled Instances
Uncertainty
Random


Ionosphere, K = 7
100
95
90
85
80
75
70
65
60
55
50
Accuracy (%)
13
23
33
43
53
63
73
83
93
103
113
123
133
143
153
163
173
183
193
203
213
223
233
243
253
263
273
283
293
303
311
Number of Labeled Instances
Uncertainty
Random


Ionosphere, K = 9
100
95
90
85
80
75
70
65
60
55
50
Accuracy (%)
15
25
35
45
55
65
75
85
95
105
115
125
135
145
155
165
175
185
195
205
215
225
235
245
255
265
275
285
295
305
311
Number of Labeled Instances
Uncertainty
Random


Ionosphere, K = 11
Accuracy (%)
100
95
90
85
80
75
70
65
60
55
50
17
27
37
47
57
67
77
87
97
107
117
127
137
147
157
167
177
187
197
207
217
227
237
247
257
267
277
287
297
307
Number of Labeled Instances
Uncertainty
Random