Finding Prototypes* for Nearest Neighbor Algorithms with Application to Forensic Trace Evidence

*to avoid CLIFFs

Fayola Peters

Masters Defense

Committee: Dr. Tim Menzies; Dr. Arun Ross; Dr. Bojan Cukic

West Virginia University LCSEE

November 12, 2010

OUTLINE

Background

CLIFF

CLIFF Experiments

Forensic Application

Conclusion

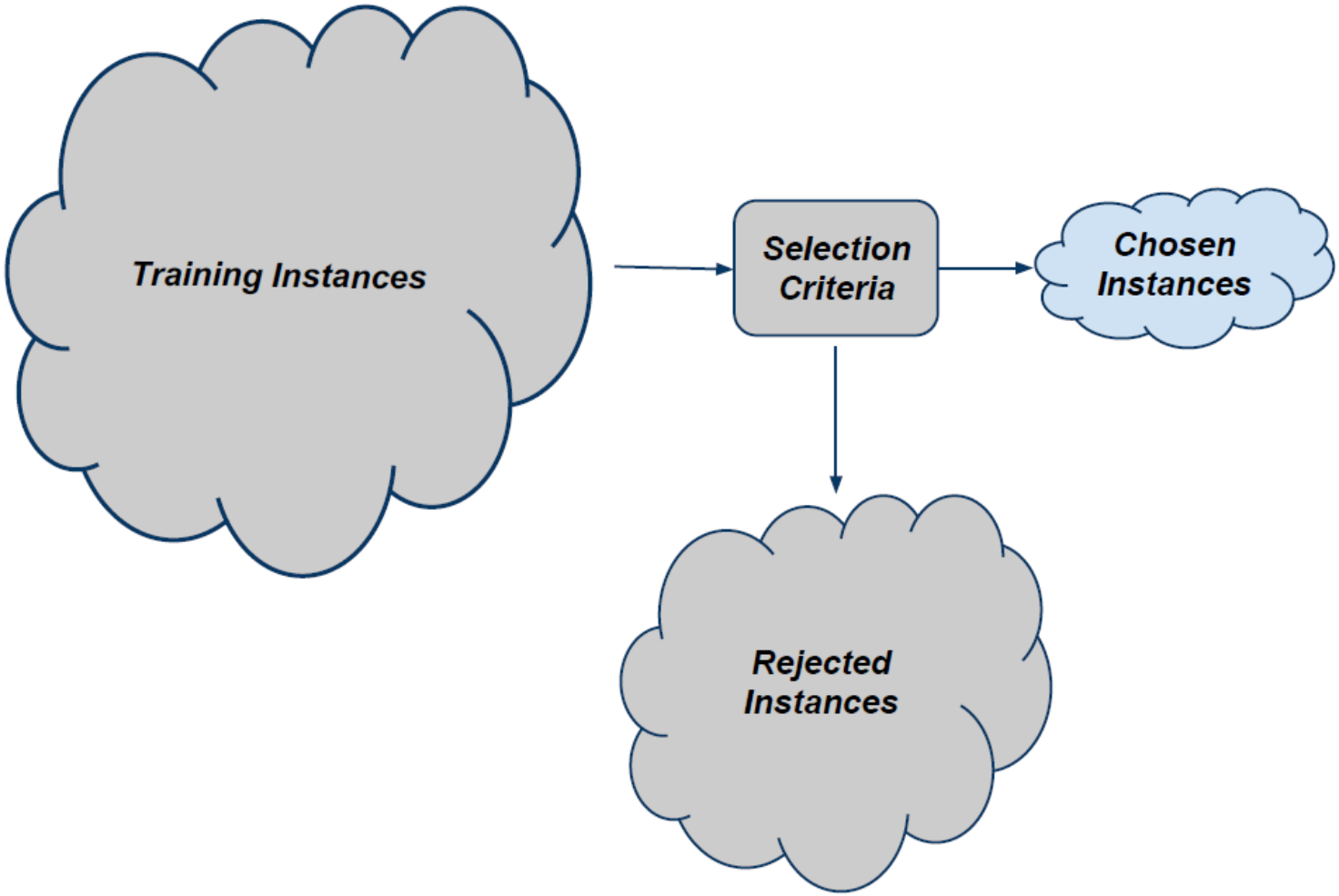
Future Work

K-Nearest Neighbor

- Created in 1967 (Hart'67)

- Issues
  - high computation costs
  - large storage requirement
  - negative effects of outliers
  - overlapping classes
  - low tolerance to noise

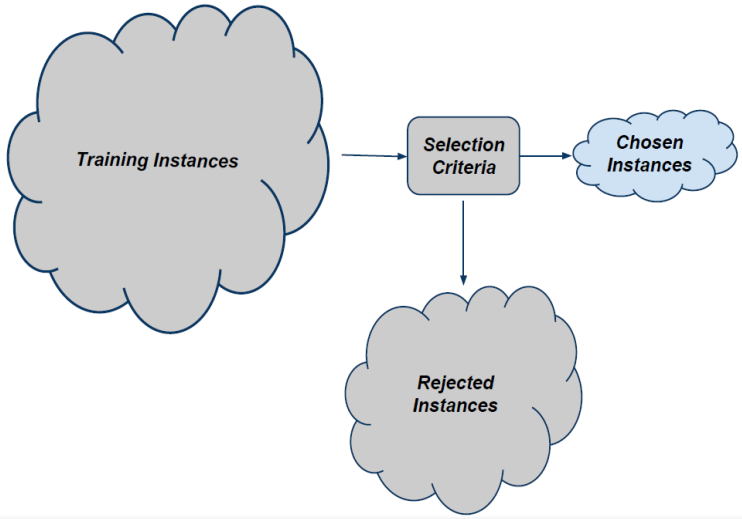
Instance Selection Process



Prototype Learning Schemes

- Instance Selection
  - Condensed Nearest Neighbor (Hart'68)
  - Minimal Consistent Set (Dasarathy'94)
  - Prototype Selection with Clusters (Lopez'10)

- Instance Abstraction
  - Chang & Modified Chang (Chang'74 and Bezdek'98)
  - Learning Vector Quantization (Kohonen'90)



The Time Issue

- Hope for time complexity of $\underline{O(n^2)}$ or less (Wilson'00)
- But…
  - CNN requires that for each new prototype added to the list, consistency must be checked.
  - Chang and Modified Chang algorithms have a consistency issue in that before a new prototype can be created via merging, consistency must be checked.
  - PSC has a three(3) step process :
    1) clustering of the training set
    2) testing whether clusters are homogeneous or heterogeneous, and
    3) finding the border prototypes of heterogeneous clusters.

OUTLINE

Background

CLIFF

CLIFF Experiments

Forensic Application

Conclusion

Future Work

The Idea of CLIFF

- Some ranges of values for attributes can be critical in selecting prototypes for each class

- So we consider using techniques practiced in the field of Feature Subset Selection (FSS) for instance selection

Support Based Bayesian Ranking (SBBR)

- Assume that the target class is divided into one class as best and the other classes as rest
  - This makes it easy to find the attribute values which have a high probability of belonging to the current best class using Bayes theorem.

$$P(best|E) = \frac{like(best|E)}{like(best|E) + like(rest|E)}$$

- To avoid distraction by low frequency evidence, a support term is added.

$$P(best|E) * support(best|E) = \frac{like(best|E)^2}{like(best|E) + like(rest|E)}$$

CLIFF Criteria

- Once ranked, the critical ranges for each attribute are extracted (those with the highest ranks) and used as the criterion for selecting instances from the current best class;
- Each criterion is made up of [attribute, value] pairs;
- Instances are selected using one pair at a time.

CLIFF Time Complexity

- Time complexity for CLIFF can be considered in terms of
  - ranking each value in each attribute, a O(m) operation where m represents attributes
  - finding the criteria for each class, a O(m) + O(k) operation where k represents the class,
  - and selecting instances from each class using the criteria a O(n) operation where n represents the number of instances
- Assuming that n > m > k this process yields a complexity of O(m) + O(m) + O(k) + O(n) which reduces to O(n)

OUTLINE

Background

CLIFF

CLIFF Experiments

Forensic Application

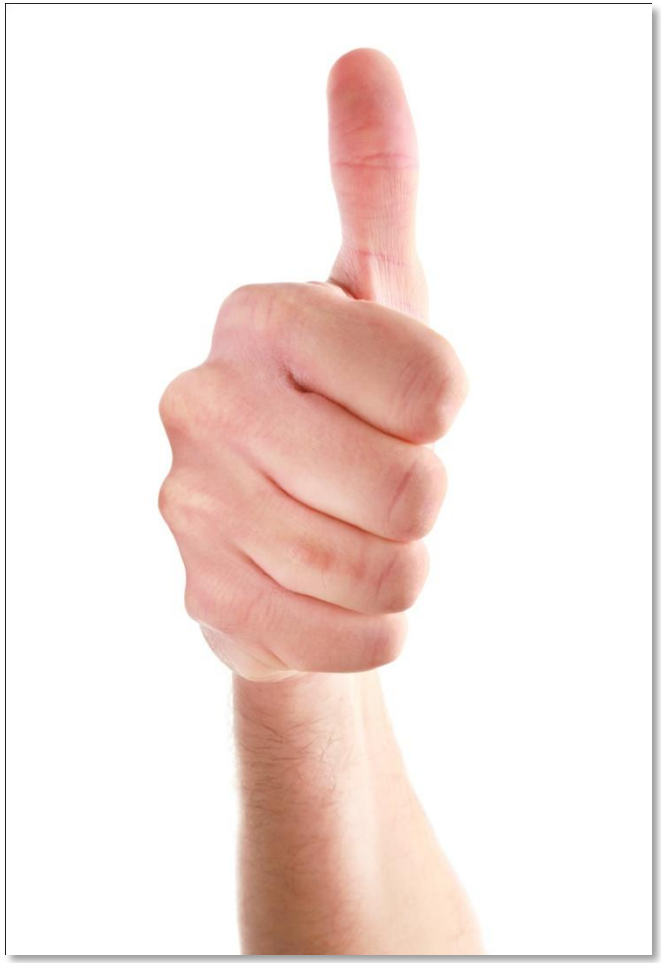
Conclusion

Future Work

Experiments

1. Is CLIFF Viable as a Prototype Learning Scheme?
   - ❖ 5 x 5 cross-validation
   - ❖ K-nearest neighbor classifier; k = 1
2. Does CLIFF handle the presence of noise well?
   - ❖ 10% of target class in training set is swapped randomly with any other class

Ideal Results for CLIFF



- Highest pds
- Lowest pfs
- Smallest size%
- Rank of 1

Results

- CLEAN – without noise
- NOISY – with noise

PD

PLS

Quartile Charts

Clean Dermatology Results

dm	PLS	rank	size%	25%	50%	75%	Q1 median Q3
pd	knn	1	100	89	100	100	
	cliff+knn	2	13	80	93	100	
	cnn+knn	3	27	77	88	100	
	psc+knn	4	10	69	86	96	
	mcs+knn	4	11	73	85	91	
pf	knn	1	100	0	0	2	
	cliff+knn	1	13	0	0	3	
	cnn+knn	1	27	0	0	3	
	psc+knn	1	10	0	0	5	
	mcs+knn	1	11	0	0	5	

0 50 100

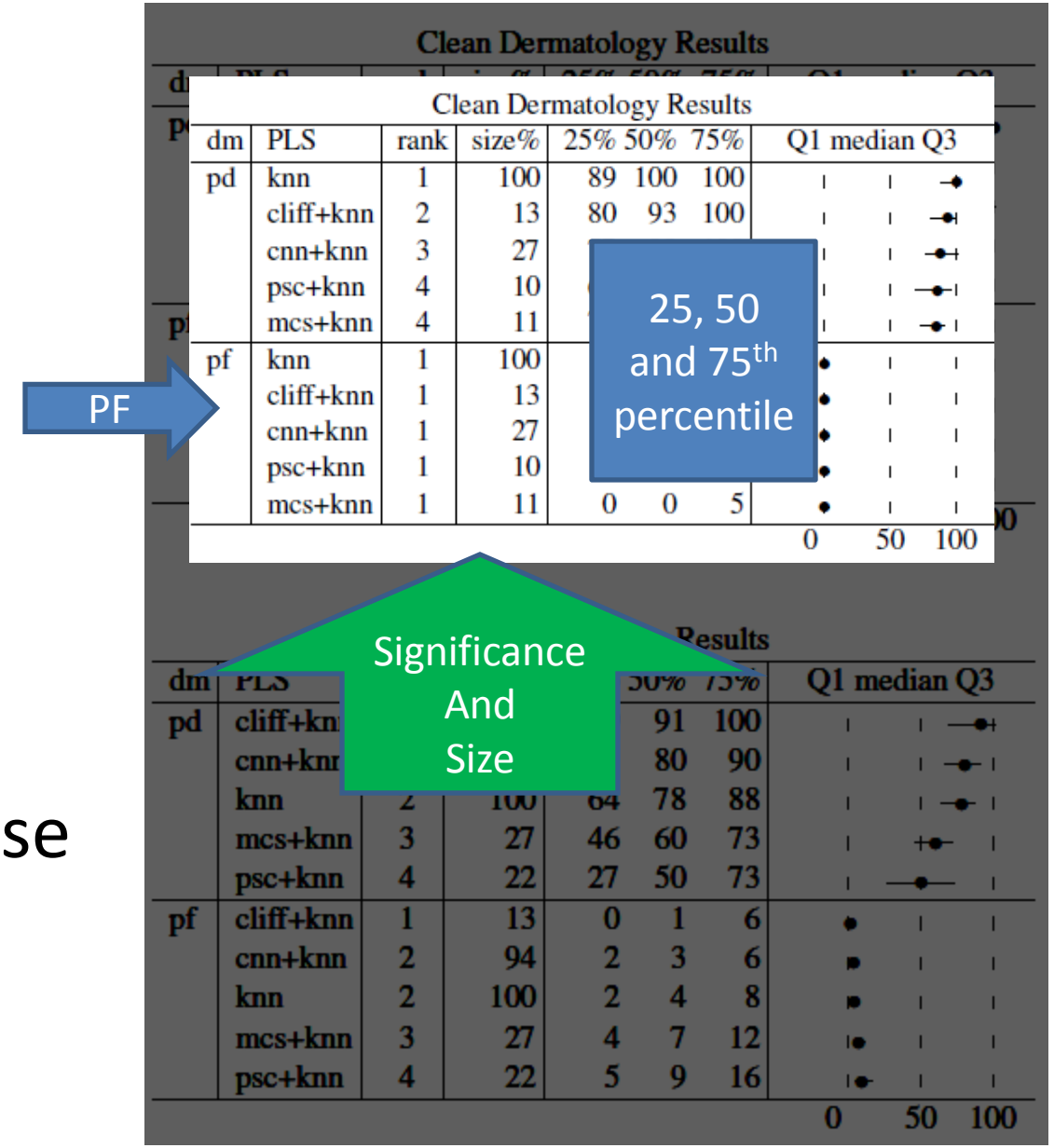
Noisy Dermatology Results

dm	PLS	rank	size%	25%	50%	75%	Q1 median Q3
pd	cliff+	1	13	69	91	100	
	cnn+knn	2	94	67	80	90	
	knn	2	100	64	78	88	
	mcs+knn	3	27	46	60	73	
	psc+knn	4	22	27	50	73	
pf	cliff+knn	1	13	0	1	6	
	cnn+knn	2	94	2	3	6	
	knn	2	100	2	4	8	
	mcs+knn	3	27	4	7	12	
	psc+knn	4	22	5	9	16	

0 50 100

Results

- Clean – without noise
- Noisy – with noise



Dermatology Results

- CLEAN: CLIFF ranks as no.2 for pd with median of 93% and a pd that is indistinguishable.
- NOISY: Size remains the same and median pd and pf show small difference

Clean Dermatology Results

dm	PLS	rank	size%	25%	50%	75%	Q1 median Q3
pd	knn	1	100	89	100	100	
	cliff+knn	2	13	80	93	100	
	cnn+knn	3	27	77	88	100	
	psc+knn	4	10	69	86	96	
	mcs+knn	4	11	73	85	91	
pf	knn	1	100	0	0	2	
	cliff+knn	1	13	0	0	3	
	cnn+knn	1	27	0	0	3	
	psc+knn	1	10	0	0	5	
	mcs+knn	1	11	0	0	5	
							0 50 100

Noisy Dermatology Results

dm	PLS	rank	size%	25%	50%	75%	Q1 median Q3
pd	cliff+knn	1	13	69	91	100	
	cnn+knn	2	94	67	80	90	
	knn	2	100	64	78	88	
	mcs+knn	3	27	46	60	73	
	psc+knn	4	22	27	50	73	
pf	cliff+knn	1	13	0	1	6	
	cnn+knn	2	94	2	3	6	
	knn	2	100	2	4	8	
	mcs+knn	3	27	4	7	12	
	psc+knn	4	22	5	9	16	
							0 50 100

Heart Results

- CLEAN: CLIFF wins for both pd and pf

- NOISY: Size decreases by 2% and median pd and pf show small degradation

Clean Heart (Hungarian) Results

hh	PLS	rank	size%	25%	50%	75%	Q1 median Q3
pd	cliff+knn	1	9	68	82	90	
	knn	1	100	65	75	83	
	cnn+knn	1	65	57	74	85	
	psc+knn	2	14	50	63	75	
	mcs+knn	2	19	53	62	71	
pf	cliff+knn	1	9	10	19	31	
	knn	1	100	16	24	33	
	cnn+knn	1	65	13	25	37	
	mcs+knn	2	19	28	38	46	
	psc+knn	2	14	28	38	48	

0 50 100

Noisy Heart (Hungarian) Results

hh	PLS	rank	size%	25%	50%	75%	Q1 median Q3
pd	cliff+knn	1	7	68	79	89	
	cnn+knn	2	76	60	65	72	
	knn	2	100	58	64	69	
	mcs+knn	2	25	51	59	68	
	psc+knn	3	19	38	53	68	
pf	cliff+knn	1	7	11	21	32	
	knn	2	100	29	35	41	
	cnn+knn	2	76	28	36	41	
	mcs+knn	2	25	28	37	47	
	psc+knn	3	19	28	44	61	

0 50 100

Mammography Results

- CLEAN: CLIFF wins for both pd and pf

Clean Mammography Results

mm	PLS	rank	size%	25%	50%	75%	Q1 median Q3
pd	cliff+knn	1	8	49	62	77	
	cnn+knn	2	57	47	54	61	
	knn	3	100	45	53	57	
	mcs+knn	3	17	48	52	57	
	psc+knn	4	10	44	50	56	
pf	cliff+knn	1	8	21	36	45	
	cnn+knn	2	57	39	46	52	
	knn	3	100	41	46	51	
	mcs+knn	3	17	41	48	52	
	psc+knn	4	10	43	47	55	
							0 50 100

- NOISY: Size increases by 1% and median pd increases by 1% and pf show small degradation

Noisy Mammography Results

mm	PLS	rank	size%	25%	50%	75%	Q1 median Q3
pd	cliff+knn	1	9	51	63	76	
	mcs+knn	2	19	44	51	57	
	psc+knn	2	11	43	50	60	
	cnn+knn	3	61	35	48	57	
	knn	4	100	34	46	53	
pf	cliff+knn	1	9	22	37	52	
	mcs+knn	2	19	39	48	54	
	psc+knn	2	11	41	50	57	
	cnn+knn	3	61	42	51	65	
	knn	4	100	44	54	63	
							0 50 100

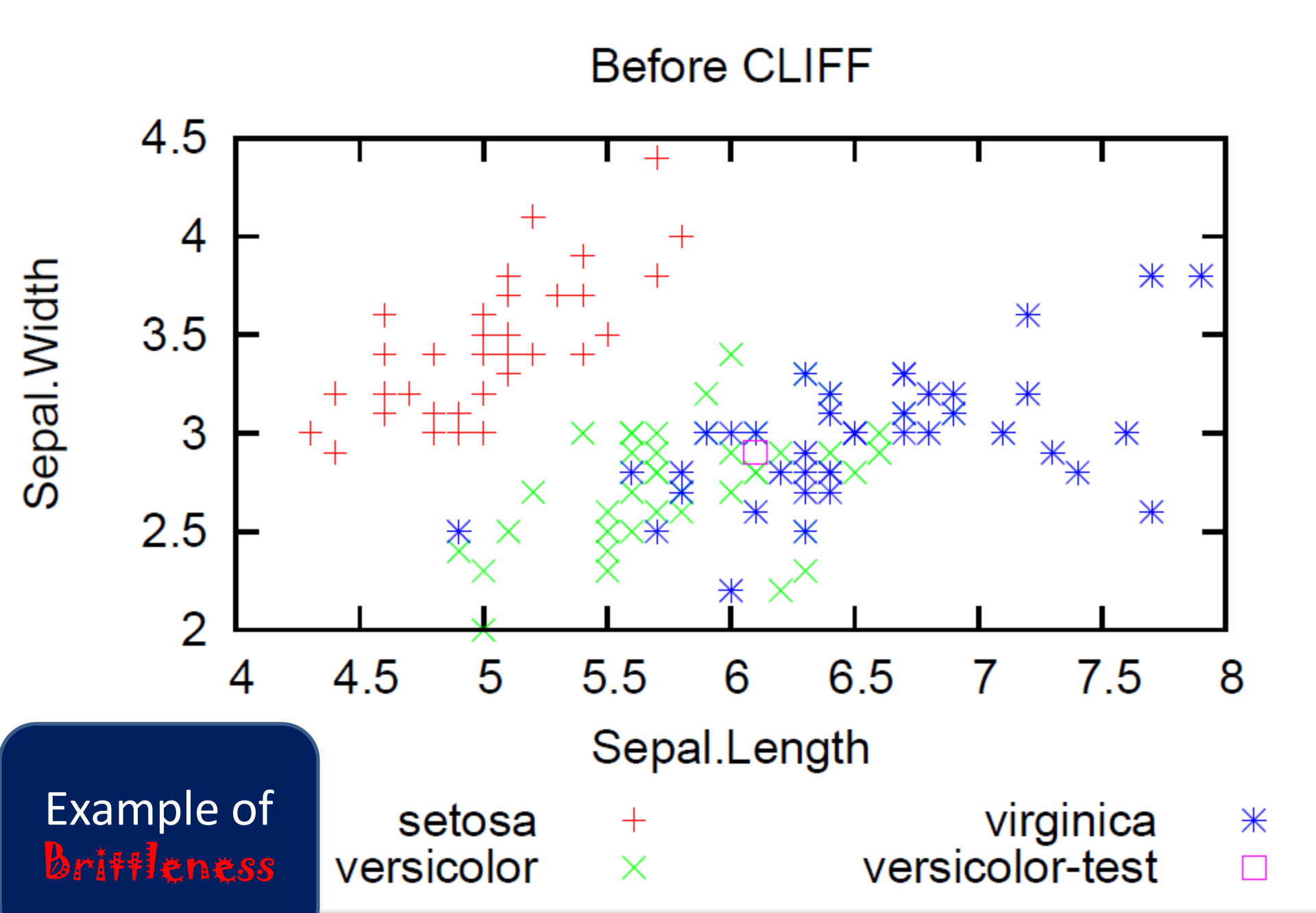
Summary of Results

- CLIFF has competitive pds showing similar or better results for Heart and Mammography respectively.
- CLIFF has the lowest pfs most of the time
- Noise does not increase the size% substantially

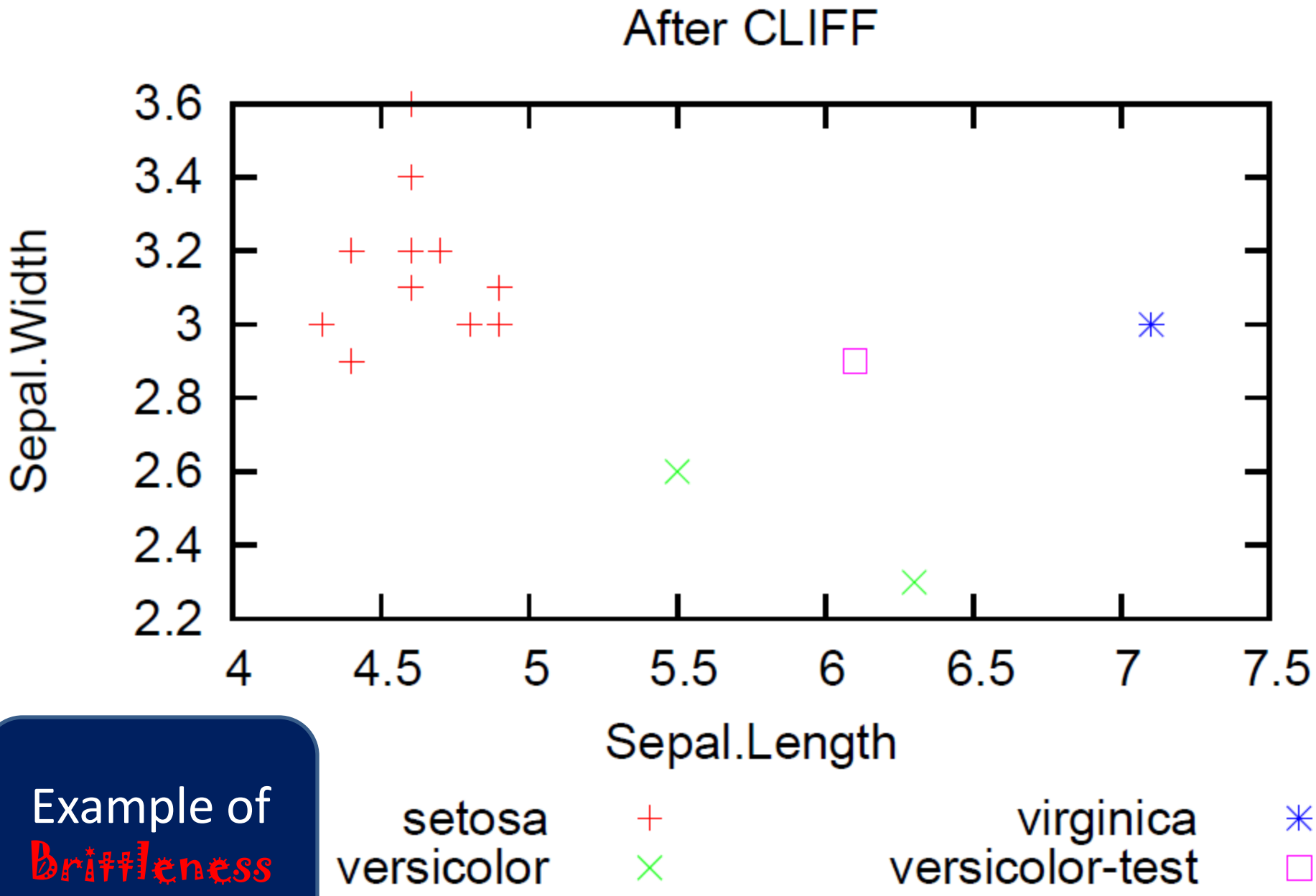
Experiment 3

Can CLIFF reduce Brittleness

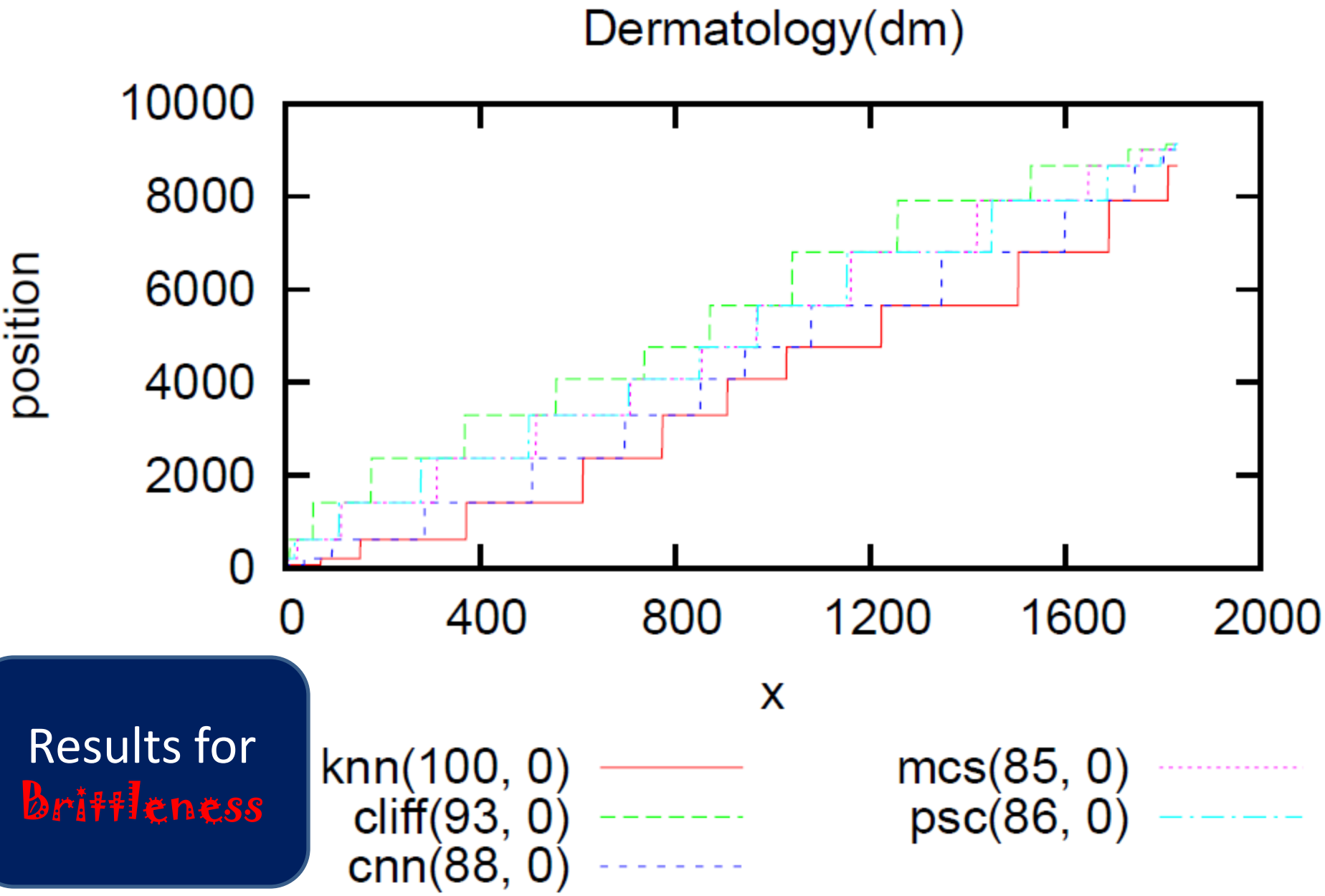
Brittleness is a measure of whether a solution (predicted target class) comes from a region of similar solutions or from a region of dissimilar solutions. Or, looking at this another way, how far would a test instance have to move before a different target class is predicted.



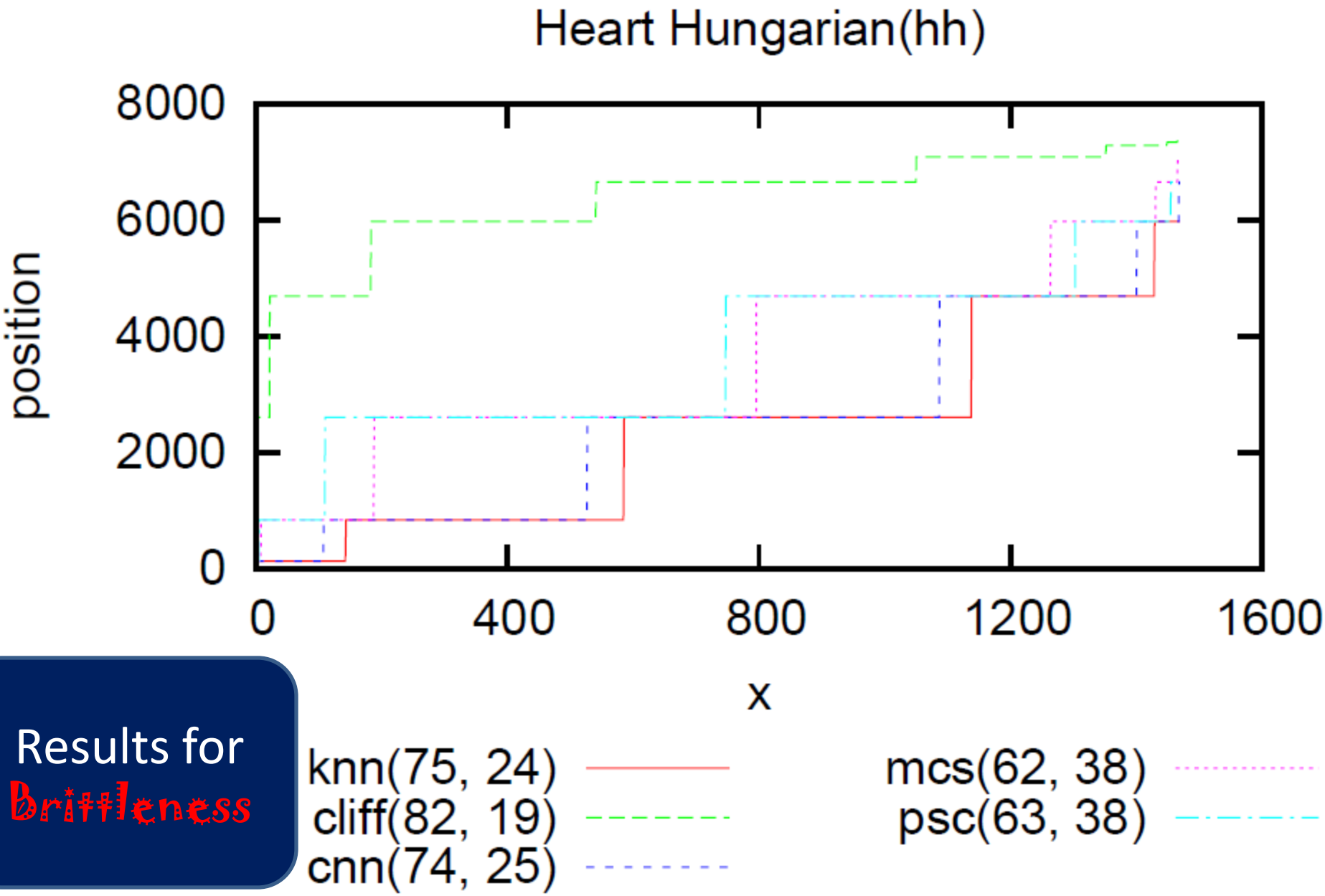
Example of **Brittleness**



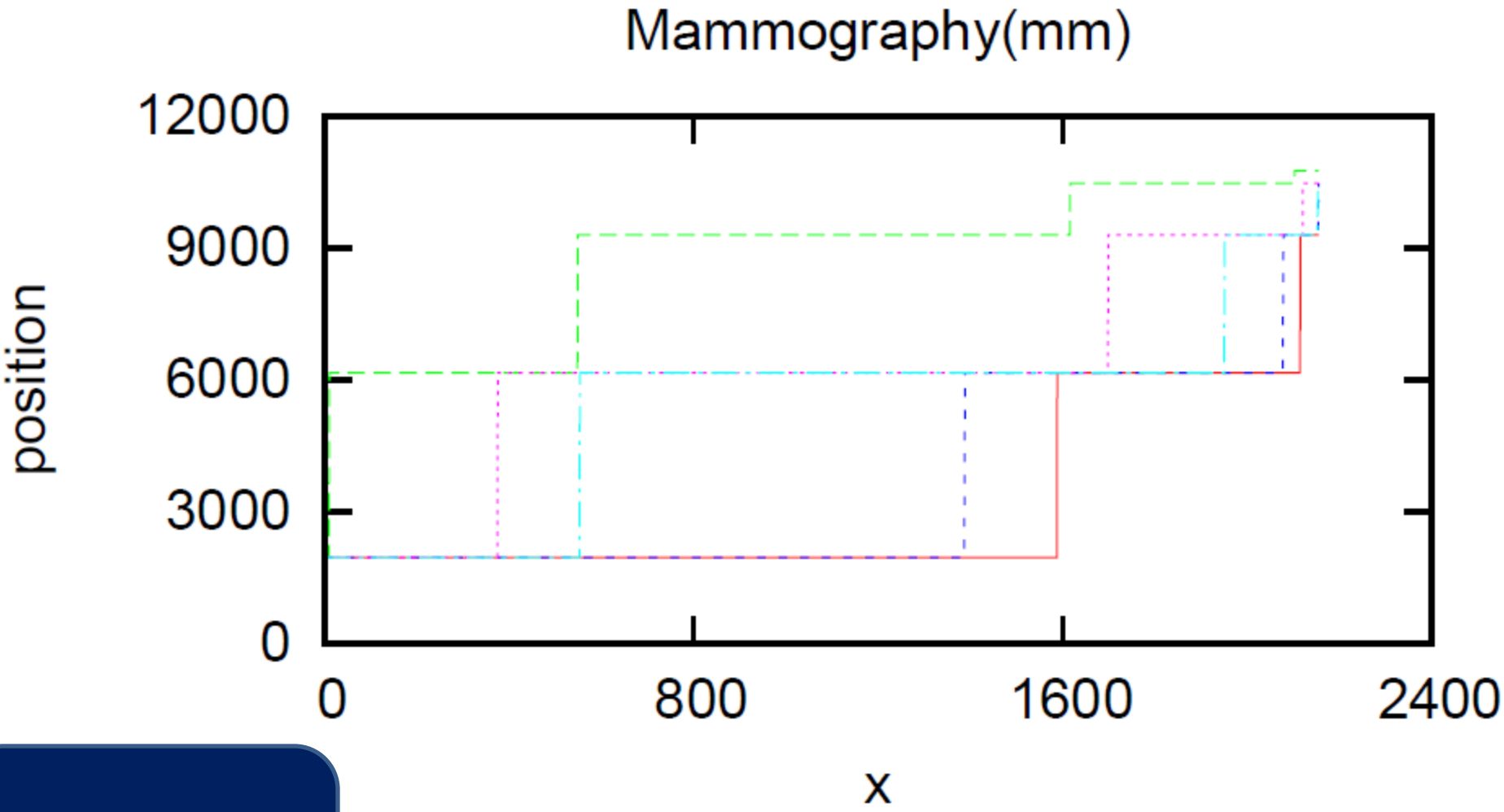
Example of Brittleness



Results for Brittleness



Results for Brittleness



Results for Brittleness

OUTLINE

Background

CLIFF

CLIFF Experiments

Forensic Application

Conclusion

Future Work

Motivation

Brittleness

With the exception of nuclear DNA analysis, ...no forensic method has been rigorously shown to have the capacity to consistently, and with a high degree of certainty, demonstrate a connection between evidence and a specific individual or source. [36]

(NAS'09)

Brittleness and Published Models

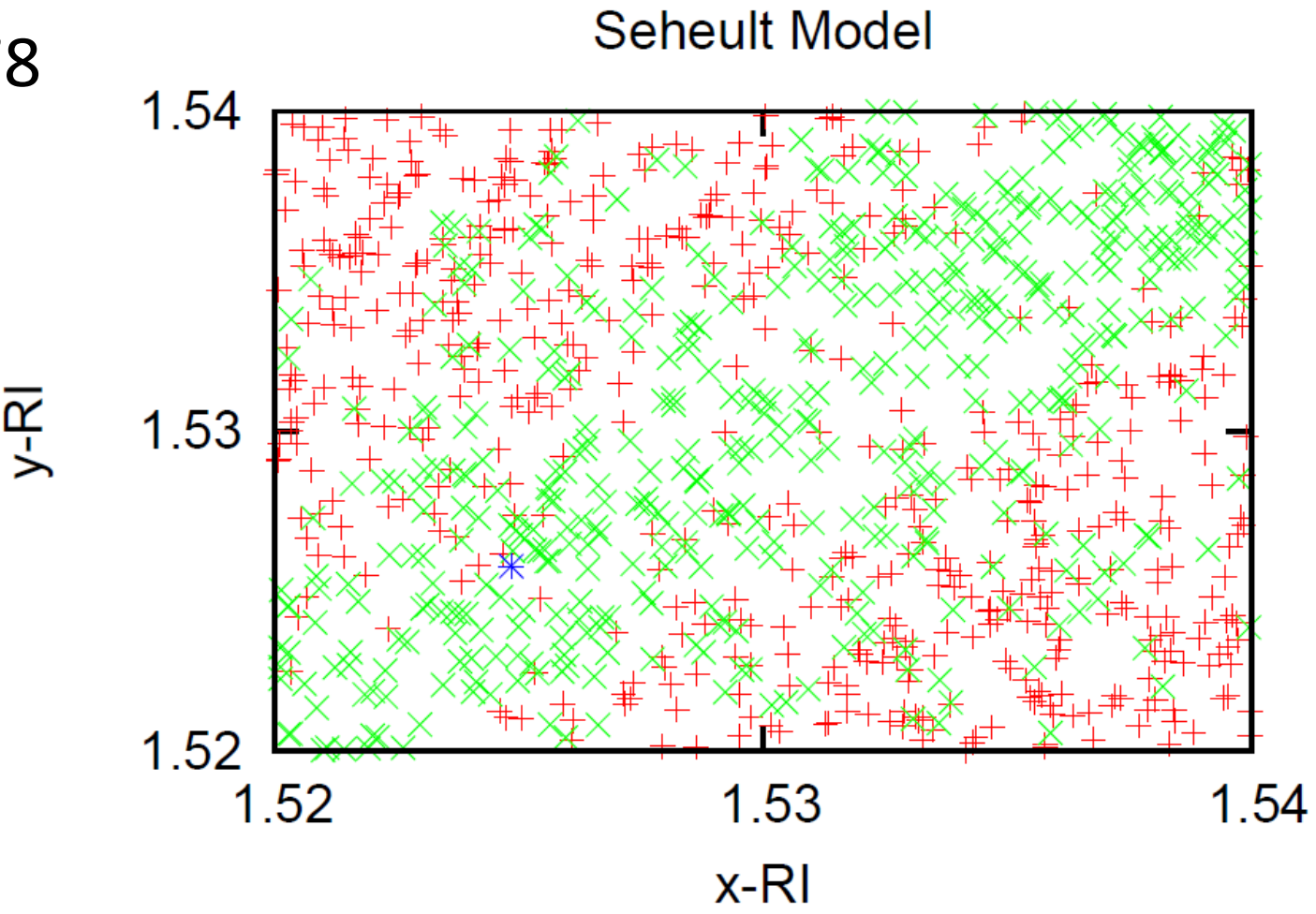
- Seheult’78

$$\frac{1+\lambda^2}{\lambda(2+\lambda^2)^{1/2}} \qquad -\frac{1}{2(1+\lambda^2)}\cdot(u^2-v^2)$$

- Grove’80
- Evett’95
- Walsh’96

Brittleness and Published Models

- Seheult'78
- Grove'80
- Evett'95
- Walsh'96



Brittleness and Published Models

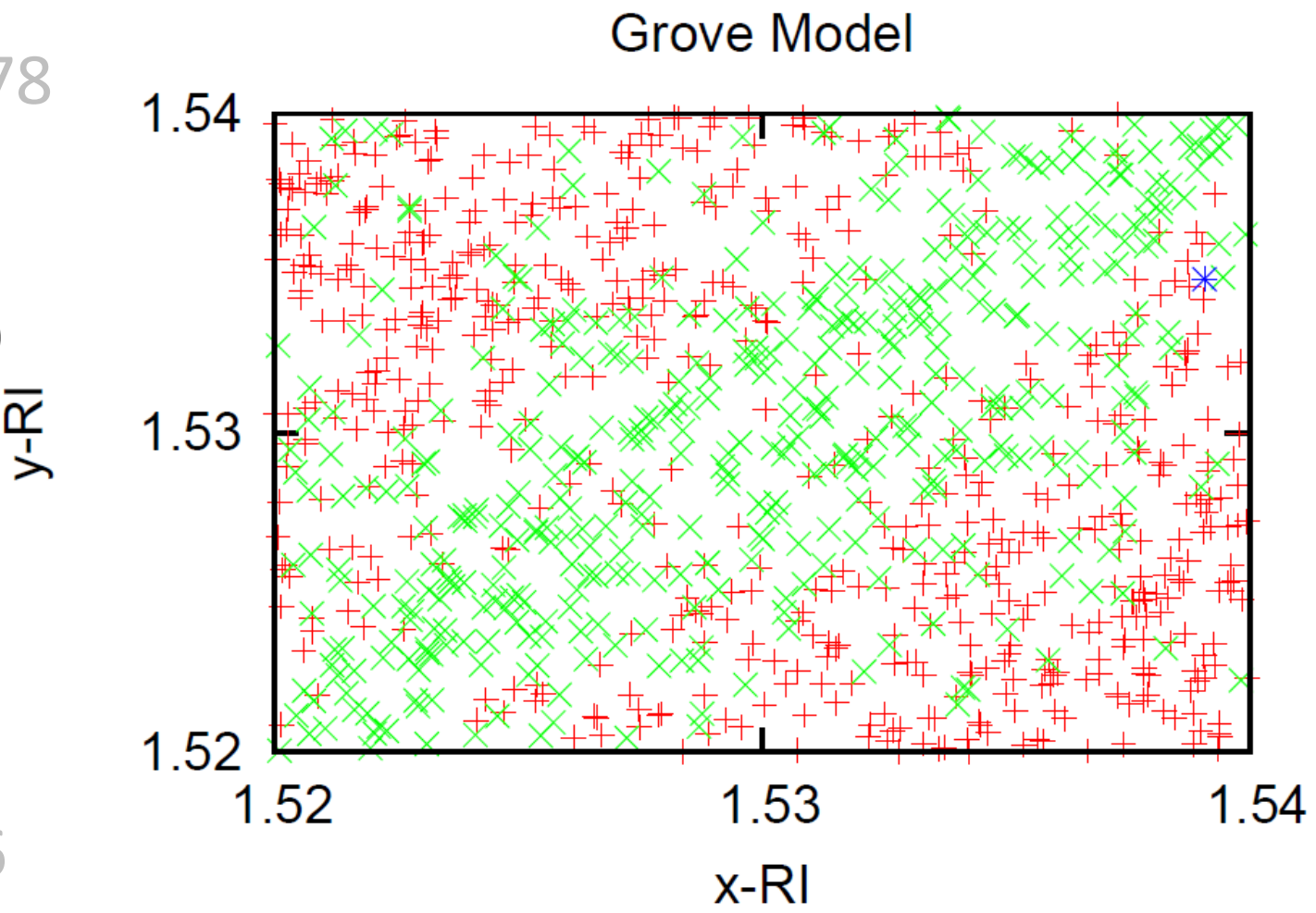
- Seheult’78
- Grove’80

$$\frac{\tau}{\sigma} \cdot e^{\left\{ \frac{-(X-Y)^2}{4\sigma^2} + \frac{(Y-\mu^2)}{2\tau^2} \right\}}$$

- Evett’95
- Walsh’96

Brittleness and Published Models

- Seheult'78
- Grove'80
- Evett'95
- Walsh'96



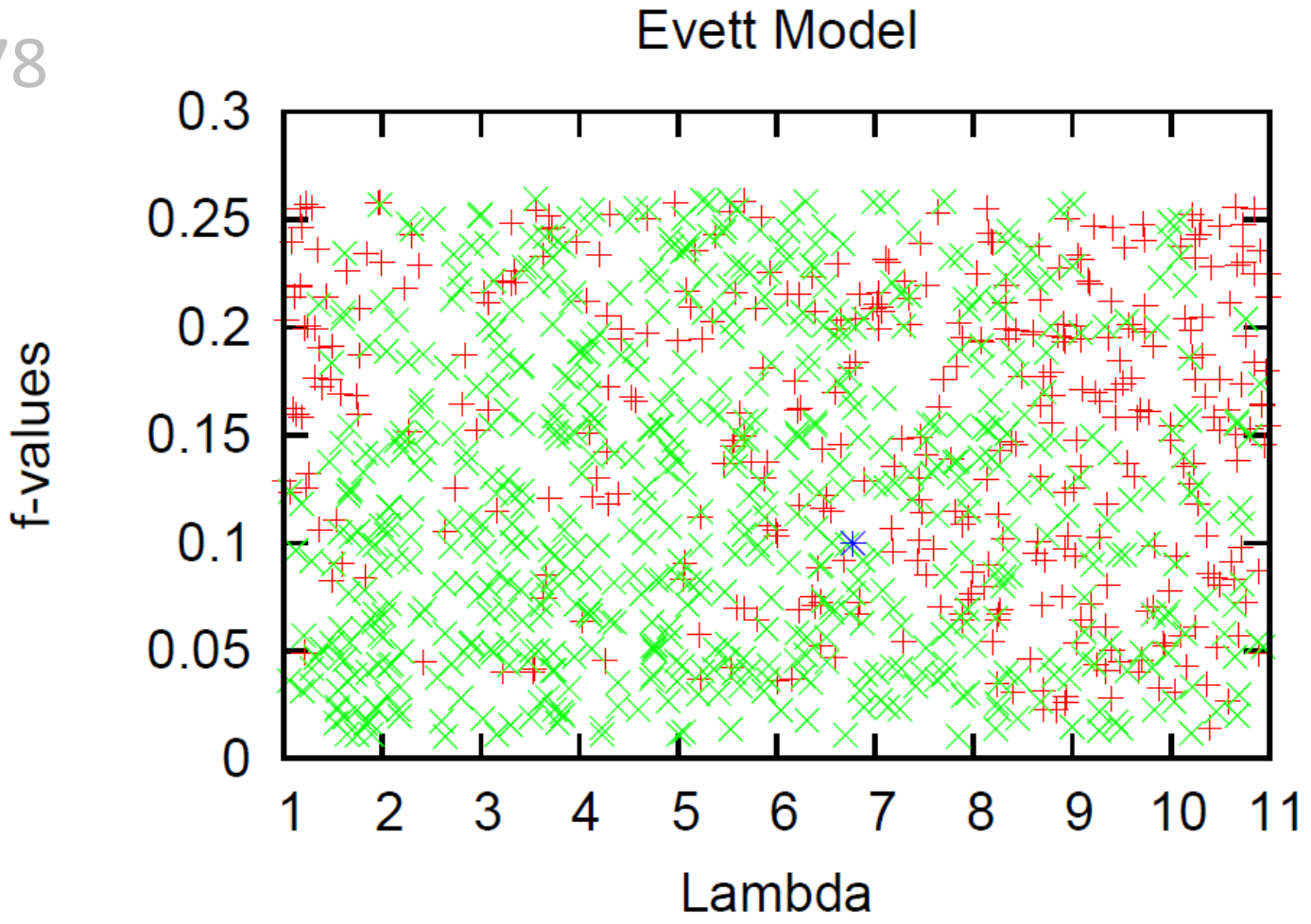
Brittleness and Published Models

- Seheult'78
- Grove'80
- Evett'95
- Walsh'96

$$LR = \frac{P_0 T_n}{P_1 S_n f} + T_0$$

Brittleness and Published Models

- Seheult'78
- Grove'80
- Evett'95
- Walsh'96



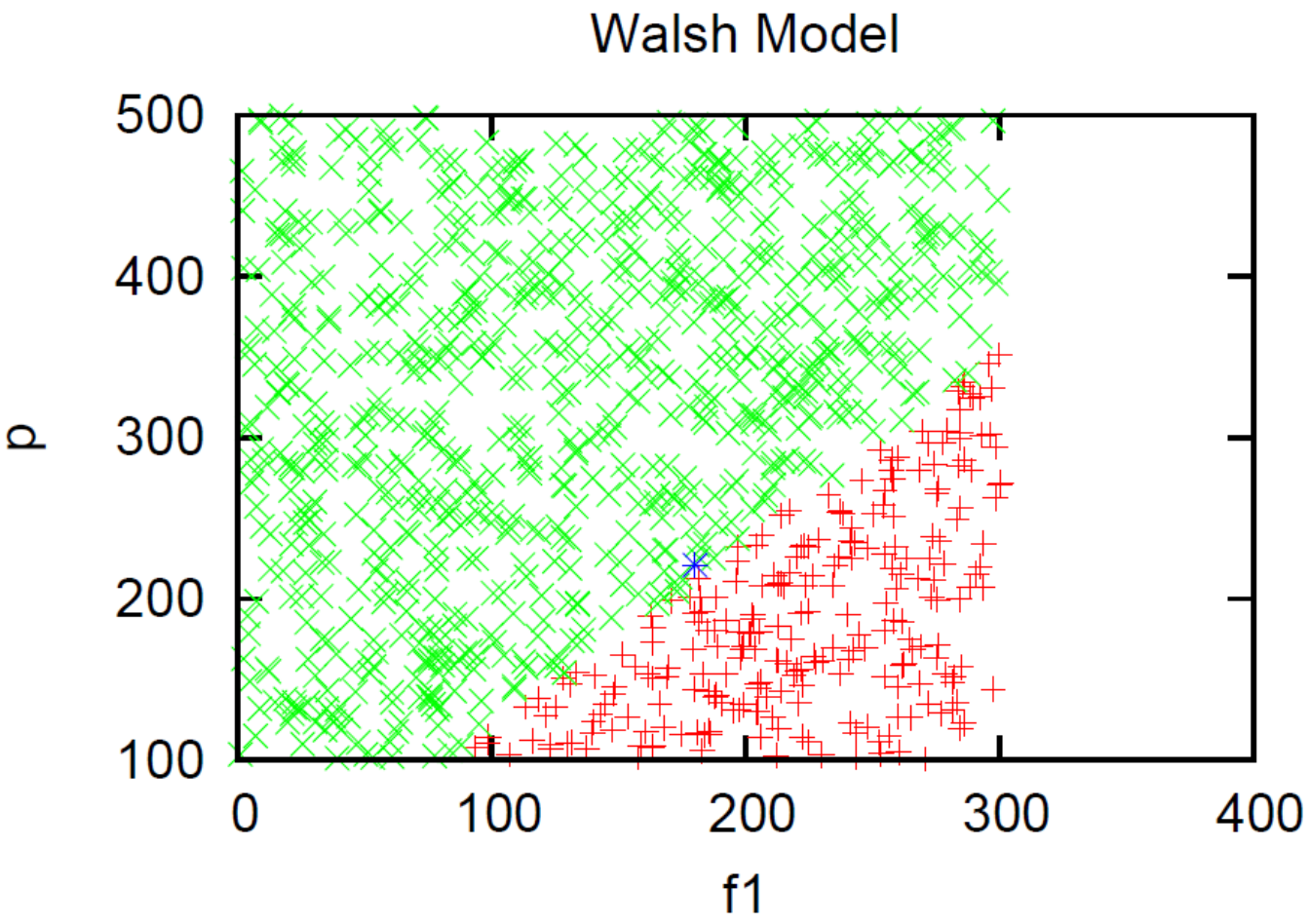
Brittleness and Published Models

- Seheult'78
- Grove'80
- Evett'95
- Walsh'96

$$\frac{T_L P_0 p(X, Y | S_y, S_x)}{P_1 S_L f_1}$$

Brittleness and Published Models

- Seheult'78
- Grove'80
- Evett'95
- Walsh'96



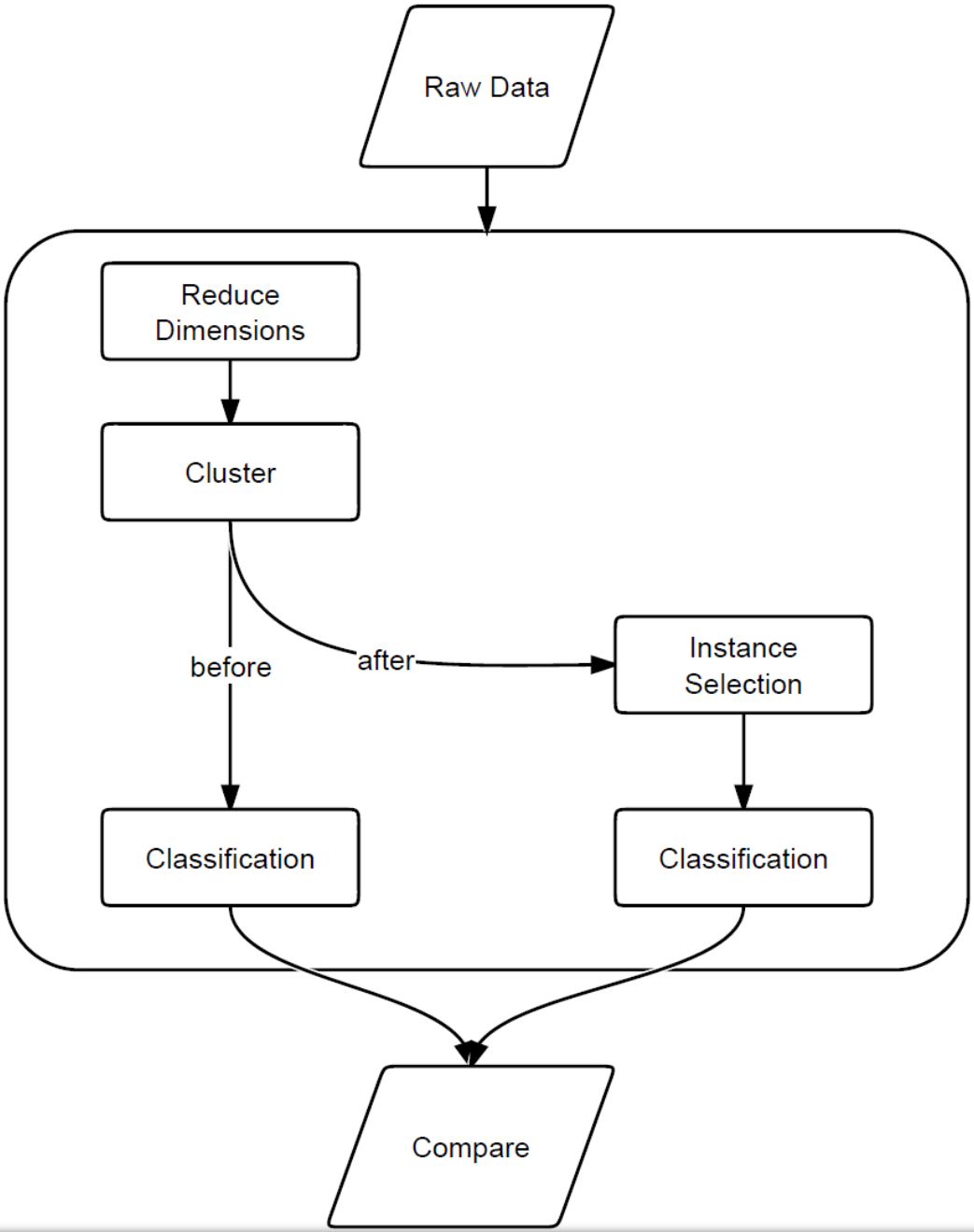
What Can We Do To Reduce Brittleness In Forensic Models?

Hint!

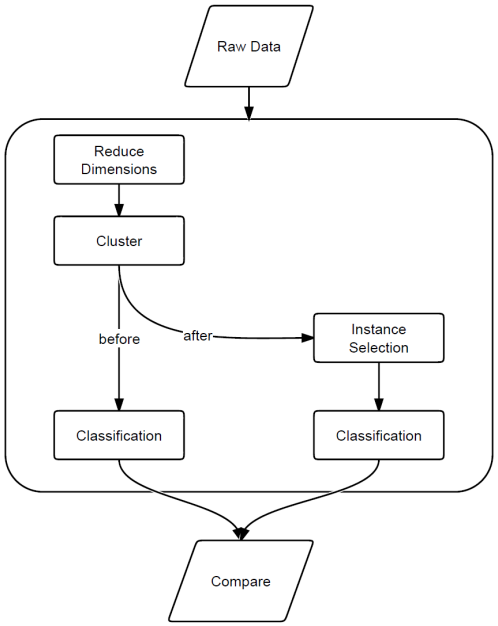
Move away from statistical models.

Introducing

The CLIFF Avoidance Model (CAM)



Data Set



Name	Attributes	Instances	Classes
Clear Coat Paint	1151	185	37

Experiment 1: CAM as a Forensic Model

- FastMap is used to create 4 dimensions
- Kmeans is used to create 4 data sets with 3, 5, 10 and 20 clusters respectively
- CAM is benchmarked against 1-nearest neighbor classifier

Results

- For all data sets except the one with 3 clusters, pds and pfs measures are ideal for both CAM and 1NN

- For data set with 3 clusters there is a small degradation in the CAM pd result – 91% vs 1NN at 100%

3	PLS	rank	size%	25%	50%	75%	Q1 median Q3
pd	1nn	1	100	90	100	100	
	cam	1	25	63	91	100	
pf	1nn	1	100	0	0	4	
	cam	2	25	0	4	16	
							0 50 100

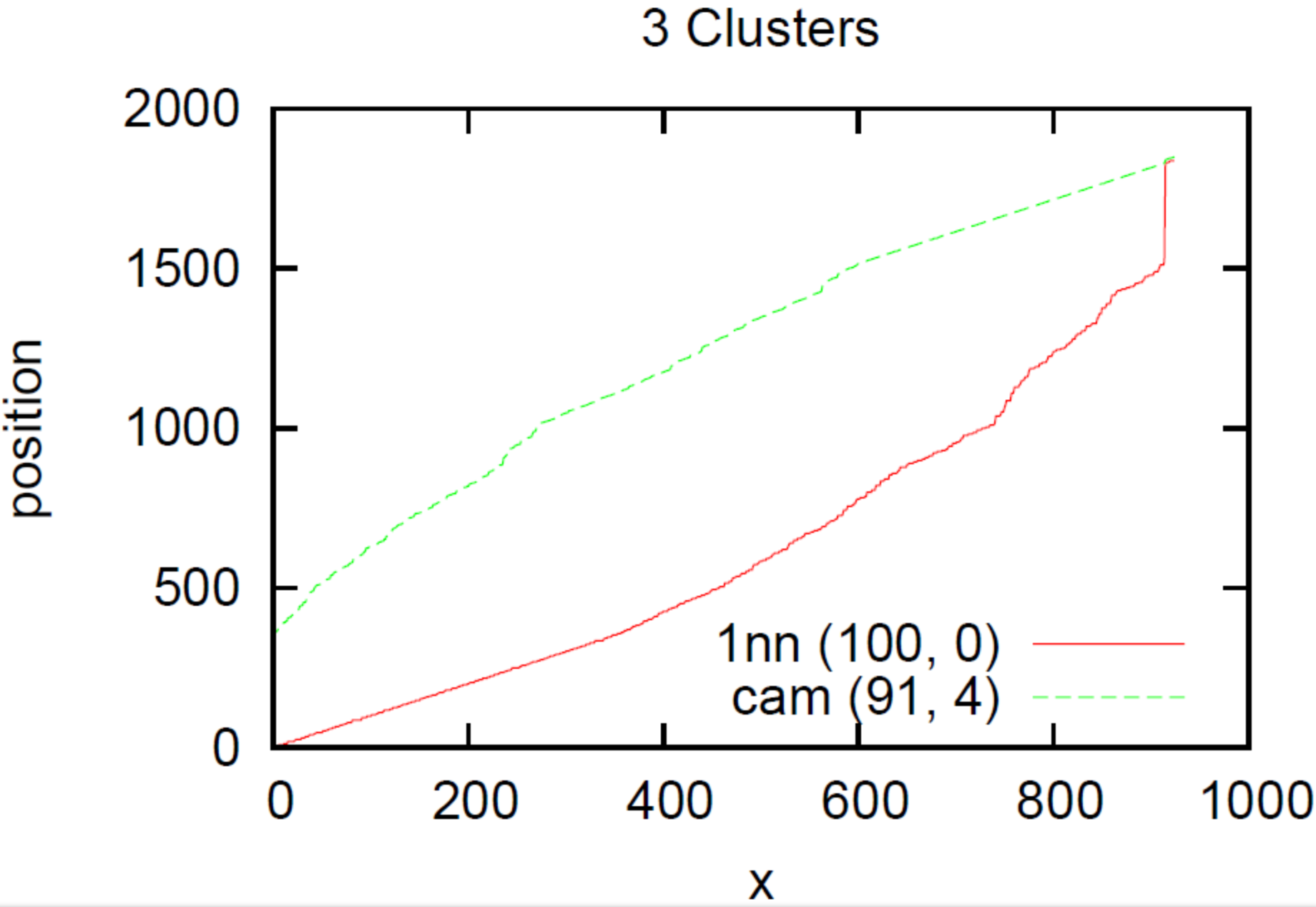
5	PLS	rank	size%	25%	50%	75%	Q1 median Q3
pd	1nn	1	100	91	100	100	
	cam	1	30	88	100	100	
pf	1nn	1	100	0	0	3	
	cam	1	30	0	0	3	
							0 50 100

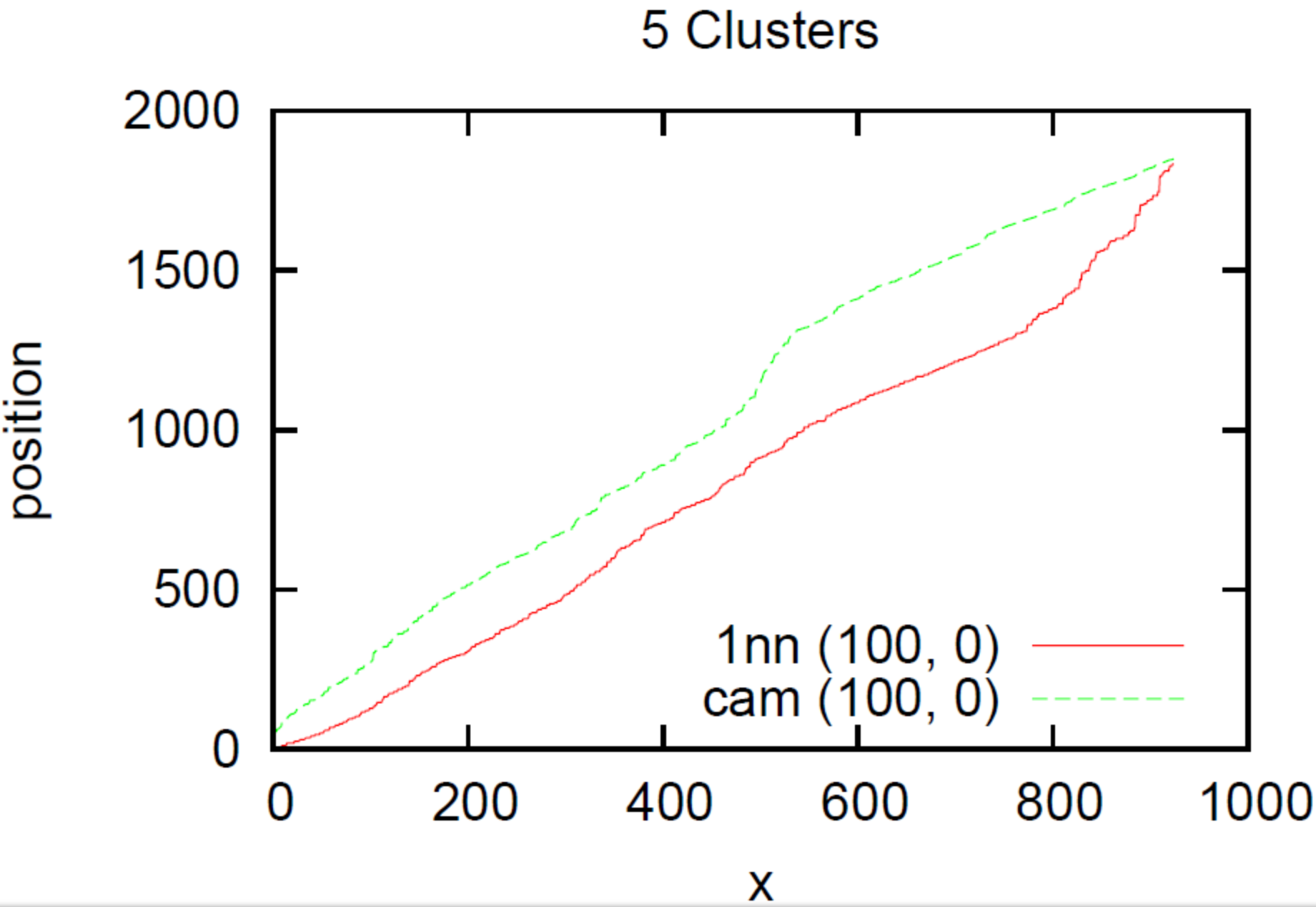
10	PLS	rank	size%	25%	50%	75%	Q1 median Q3
pd	1nn	1	100	92	100	100	
	cam	1	36	67	100	100	
pf	1nn	1	100	0	0	0	
	cam	1	36	0	0	3	
							0 50 100

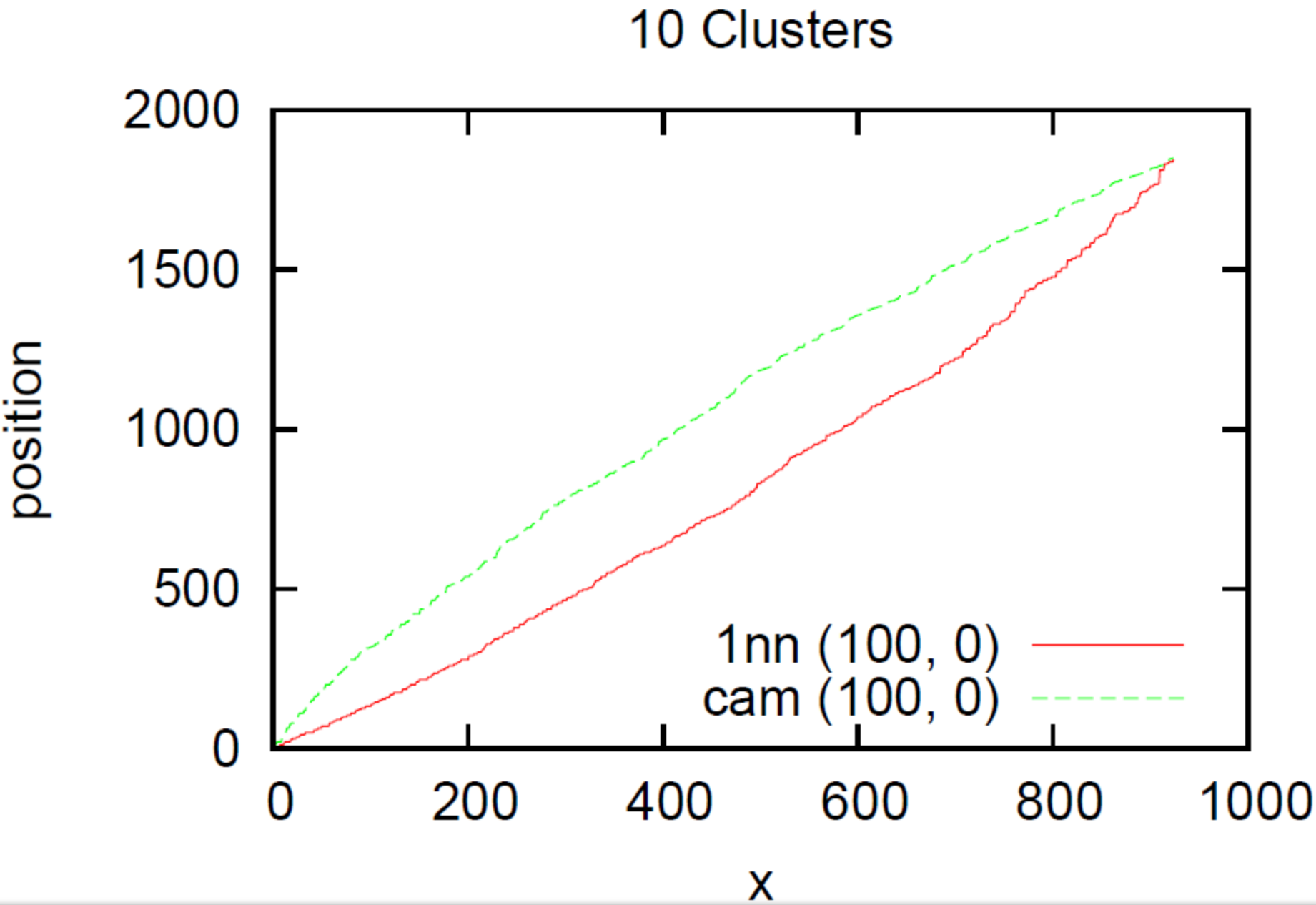
20	PLS	rank	size%	25%	50%	75%	Q1 median Q3
pd	1nn	1	100	33	100	100	
	cam	1	53	75	100	100	
pf	1nn	1	100	0	0	0	
	cam	1	53	0	0	3	
							0 50 100

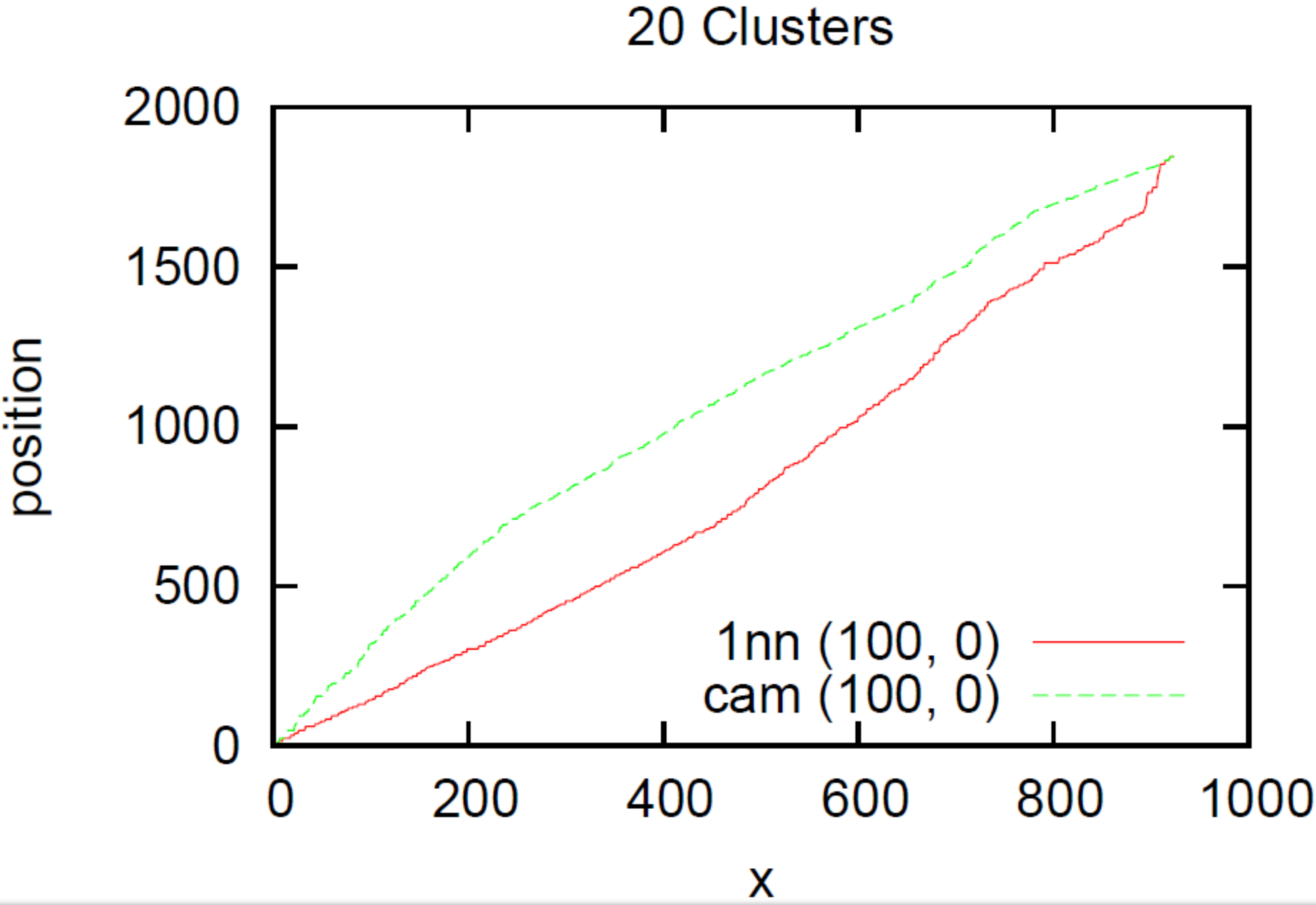
Experiment 2: Does CAM Reduce Brittleness?

- CLIFF is used to create prototypes
  - Using the classified test set
  - Distances of nearest unlike neighbor found for CAM and 1NN
  - These are ranked according to position in population and plotted


3 Clusters
2000
1500
1000
500
0
position
0
200
400
600
800
1000
x
1nn (100, 0)
cam (91, 4)


5 Clusters
2000
1500
1000
500
0
position
0
200
400
600
800
1000
x
1nn (100, 0)
cam (100, 0)


10 Clusters
2000
1500
1000
500
0
position
0
200
400
600
800
1000
x
1nn (100, 0)
cam (100, 0)


20 Clusters
2000
1500
1000
500
0
position
0
200
400
600
800
1000
x
1nn (100, 0)
cam (100, 0)

OUTLINE

Background

CLIFF

CLIFF Experiments

Forensic Application

Conclusion

Future Work

Conclusions

- We showed that CLIFF
  - has a time complexity of O(n);
  - reduces training sets to a range of 9 to 15%;
  - has pd and p f results which compares favorably with 1NN and other PLS in several standard data sets;
  - does not significantly increase the number of instances selected in the presence of noise as compared with other PLS;
  - reduces brittleness substantially in most data sets used.

OUTLINE

Background

CLIFF

CLIFF Experiments

Forensic Application

Conclusion

Future Work

Future Work

- Using CLIFF with Other Classifiers
- Using CLIFF to Optimized Feature Subset Selection
- Comparing CAM to Other Forensic Models and Forensic Data Sets

Questions?

References