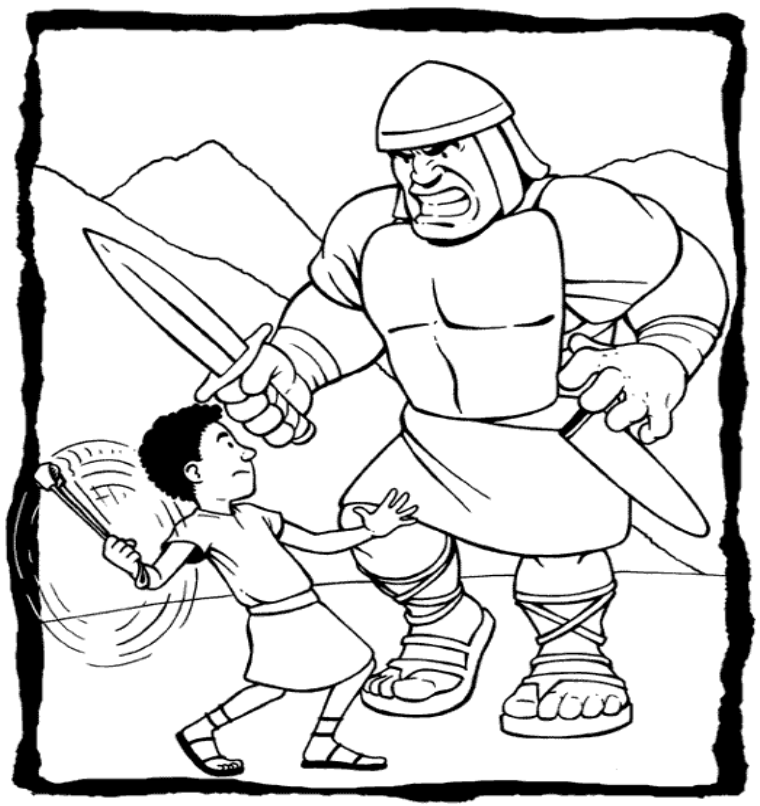
Instance-based Reasoning (Less is More!)

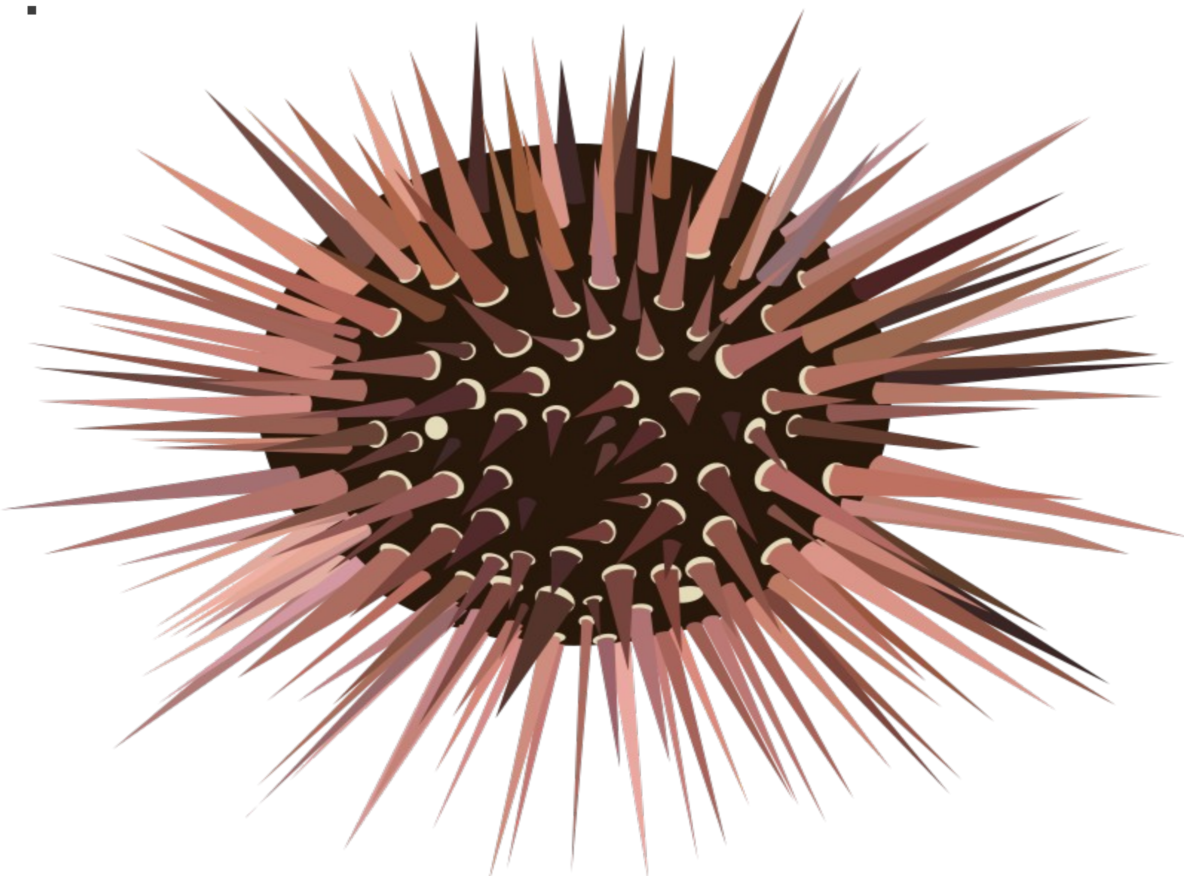


David & Goliath
1 Samuel 17:1–58

Fayola Peters
Lane Department of Computer Science and Electrical Engineering,
West Virginia University

The Problem?

- Only a few instances matter...
- But why?



Outline

- **Previous research = Less is More**
- Why? - The Answer lies in the E(k) matrix
- Now - we exploit instance space

Previous Research = Less is More!

- **TEAK**
- Cross company
- Independent Variable Mutation
- Bias/Variance

TEAK

- Test Essential Assumption Knowledge
- TEAK's design
  - Select a prediction system.
  - Identify the predictor’s essential assumption(s).
  - Recognize when those assumption(s) are violated.
  - Remove those situations.
  - Execute the modified prediction system.
- Conclusion – only few instances matter.

Previous Research = Less is More!

- TEAK
- **Cross company**
- Independent Variable Mutation
- Bias/Variance

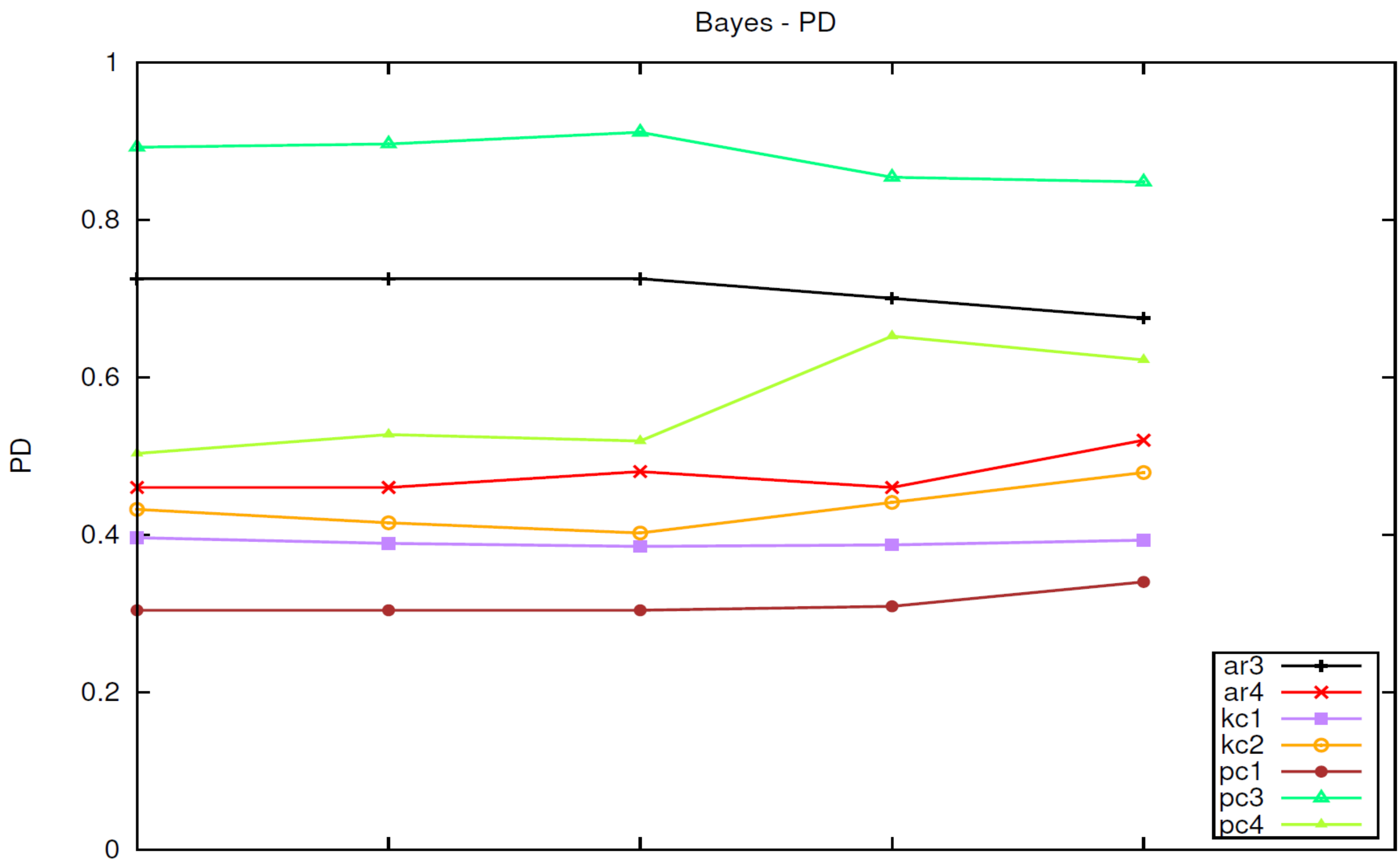
Cross Company

- Acceptable to use cross data sources once a **relevancy filter** is used
- Relevancy filter selects small subset relevant to current test case
- Removes training instances that create noise in the estimation process
- In theory, this leaves data that adheres to the principal of locality.

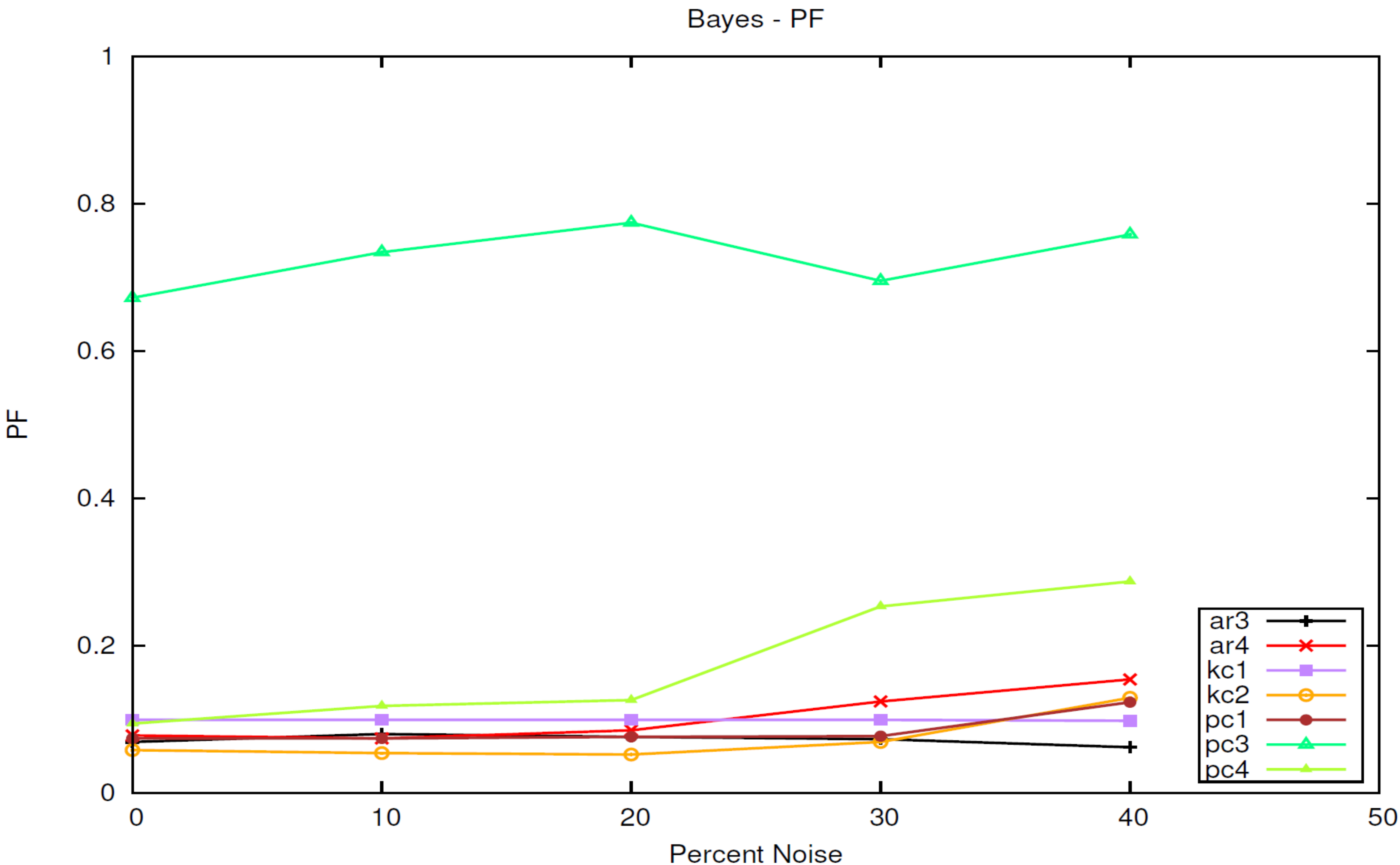
Previous Research = Less is More!

- TEAK
- Cross company
- **Independent Variable Mutation**
- Bias/Variance

Independent Variable Mutation



Independent Variable Mutation



Previous Research = Less is More!

- TEAC
- Cross company
- Independent Variable Mutation
- **Bias/Variance**

Bias/Variance

- Observations
  - According to theory higher number of smaller test sets, increase the variance and decrease the bias.
  - Extensive study showed that the theory does not hold for effort estimation datasets.
- Conclusion – only few instances matter!

Outline

- Previous research = Less is More
- **Why? - The Answer lies in the E(k) matrix**
- Now - we exploit instance space

Effort Estimation and Active Learning

- Investigation of software effort dataset characteristics
- First application of active learning on software effort estimation
- Active-learning guidance system based on dataset characteristics
  - Reduction in data collection effort

The E(k) Matrix

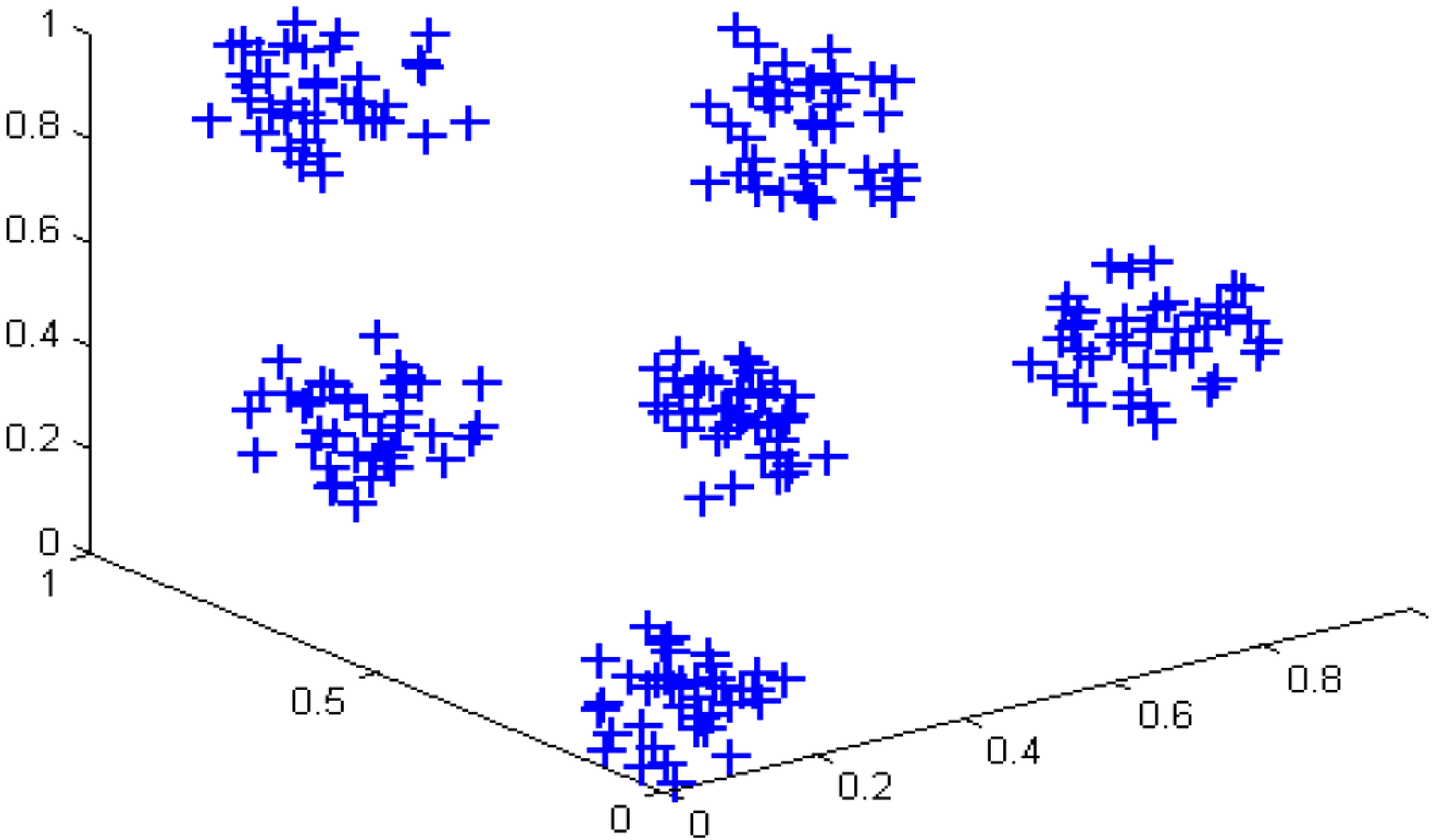
- Everyones k-th nearest matrix
  - The story...

We were interested in the effect of injecting noise to the datasets in the context of ABE.

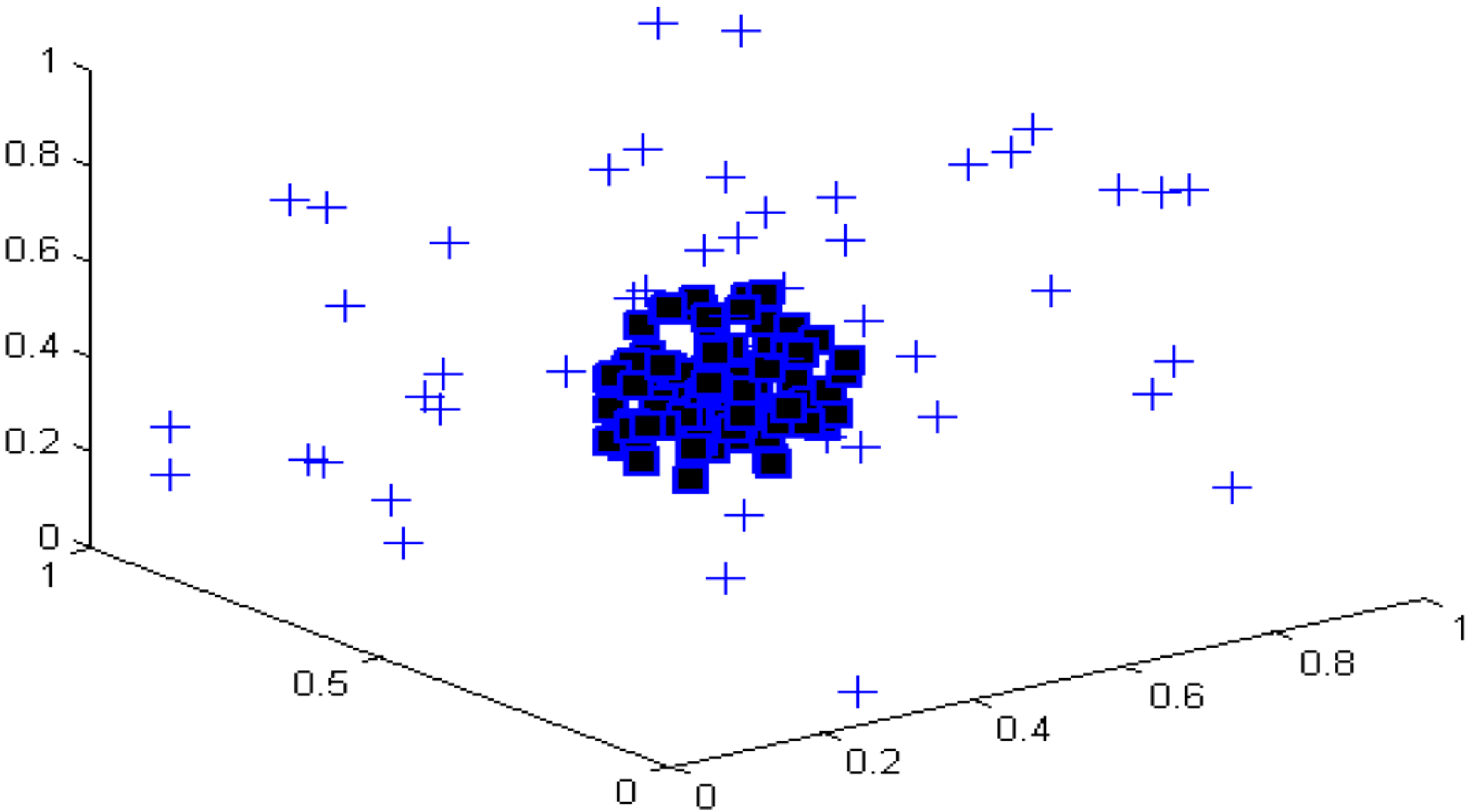
When noise was injected the ABE performances before and after noise injection were statistically the same.

Why? - maybe datasets had a different topology than predicted

Expected Topology



Actual Topology



Outline

- Previous research = Less is More
- Why? - The Answer lies in the E(k) matrix
- **Now - we exploit instance space**

Exploiting Instance Space

- E(k) and guidance system
- CLIFF

E(k) Matrix and Guidance System

- Simple example
-

Project	KLOC	Effort
P_1	20	3
P_2	10	4
P_3	40	7

E(k) Matrix and Guidance System

- Step 1: Build distance matrix
-

	P_1	P_2	P_3
P_1	0	0.34	0.66
P_2	0.34	0	1
P_3	0.66	1	0

E(k) Matrix and Guidance System

- Step 2: Create E(k) Matrix
-

	P_1	P_2	P_3
P_1	na	1	2
P_2	1	na	2
P_3	1	2	na

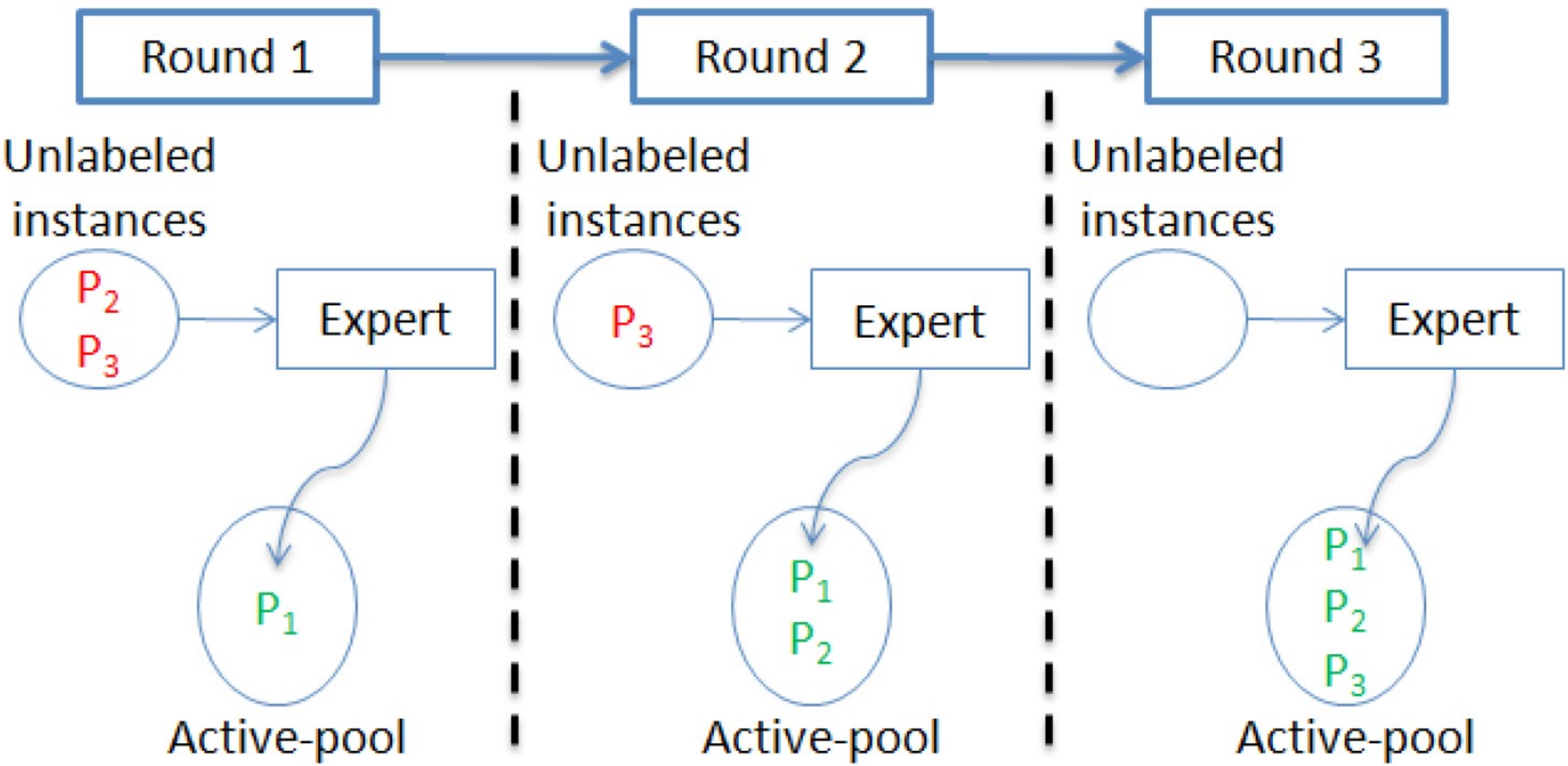
E(k) Matrix and Guidance System

- Step 3: Calculate Popularity Index

		P_1	P_2	P_3
	P_1	0	1	0
	P_2	1	0	0
$+$	P_3	1	0	0
	$Popularity:$	2	1	0

E(k) Matrix and Guidance System

- Visualization of Process



CLIFF – Immunizer Against Noise

- Simple example
-

#	forecast	temp	humidty	windy	play
1.	sunny	hot	high	FALSE	no
2.	sunny	hot	high	TRUE	no
3.	overcast	hot	high	FALSE	yes
4.	rainy	mild	high	FALSE	yes
5.	rainy	cool	normal	FALSE	yes
6.	rainy	cool	normal	TRUE	no
7.	overcast	cool	normal	TRUE	yes
8.	sunny	mild	high	FALSE	no
9.	sunny	cool	normal	FALSE	yes
10.	rainy	mild	normal	FALSE	yes
11.	sunny	mild	normal	TRUE	yes
12.	overcast	mild	high	TRUE	yes
13.	overcast	hot	normal	FALSE	yes
14.	rainy	mild	high	TRUE	no

CLIFF – Immunizer Against Noise

- Step 1: Get Criteria

{{forecast, rainy} {temp, mild} {humidity, high} {windy, FALSE}}

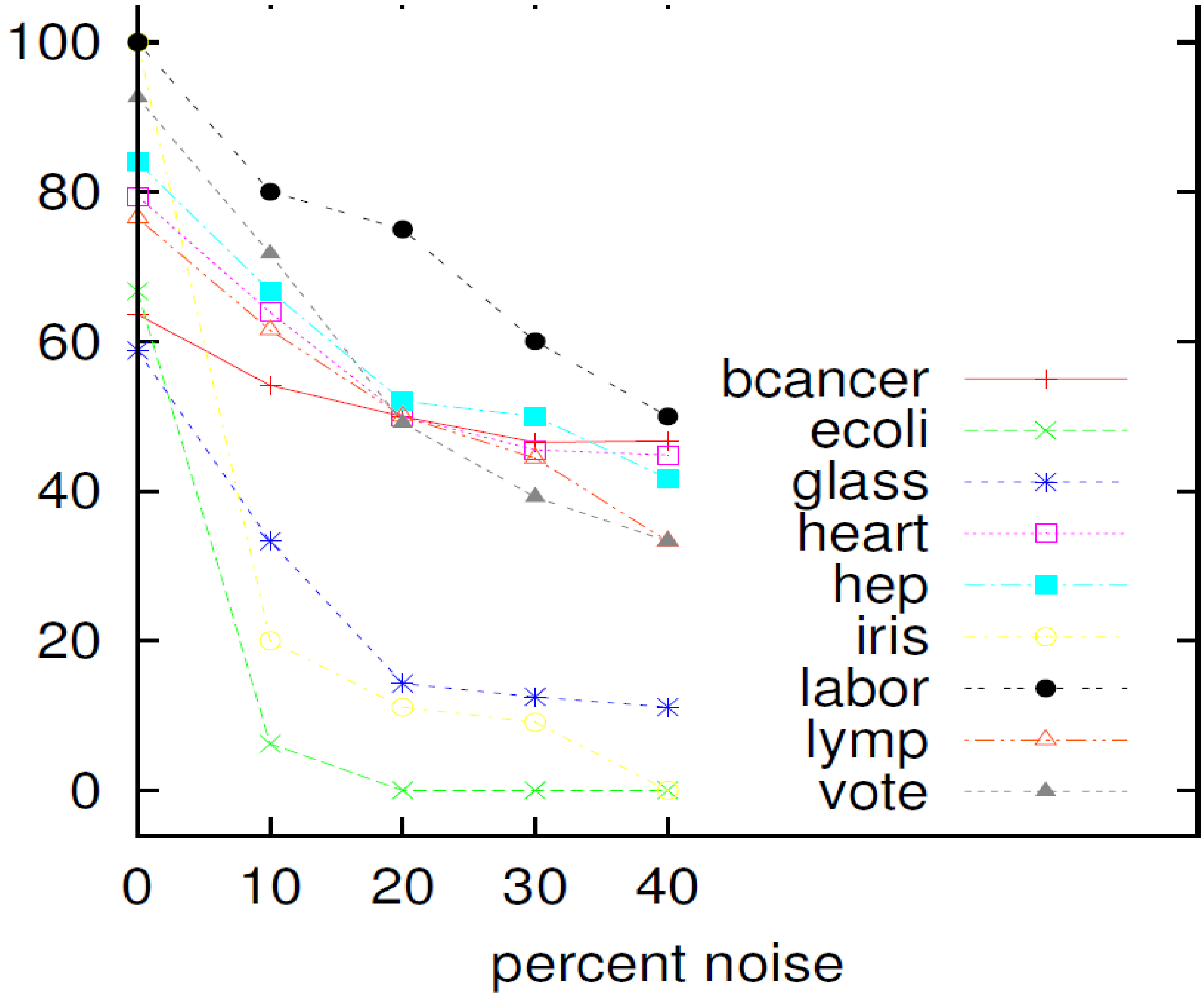
- Step 2: Apply Criteria

4.	rainy	mild	high	FALSE	yes
5.	rainy	cool	normal	FALSE	yes
10.	rainy	mild	normal	FALSE	yes

4.	rainy	mild	high	FALSE	yes
10.	rainy	mild	normal	FALSE	yes

CLIFF vs KNN

- KNN pd



CLIFF vs KNN

- CLIFF pd