Essence: Pervasive & Distributed Intelligence

BIG DATA THINNING: KNOWLEDGE DISCOVERY FROM RELEVANT DATA

MR NAJI SHEHAB

SUPERVISED BY: DR CHRISTOS ANAGNOSTOPOULOS

@ SAWB- F121; 21 MAR 2019

THE PROBLEM

What is the problem?	• Building Predictive Models Over Big Data
Who has this problem?	• Data Scientists - Data Analysts – IOT Applications – Edge Centric Services
Why should this problem be solved?	• Big Data Is Big – It Can Be Thinned
The Solution ?	• Exploit & Learn Only The Most Interested & Important Subspaces

HYPOTHESES

HYPOTHESIS 1

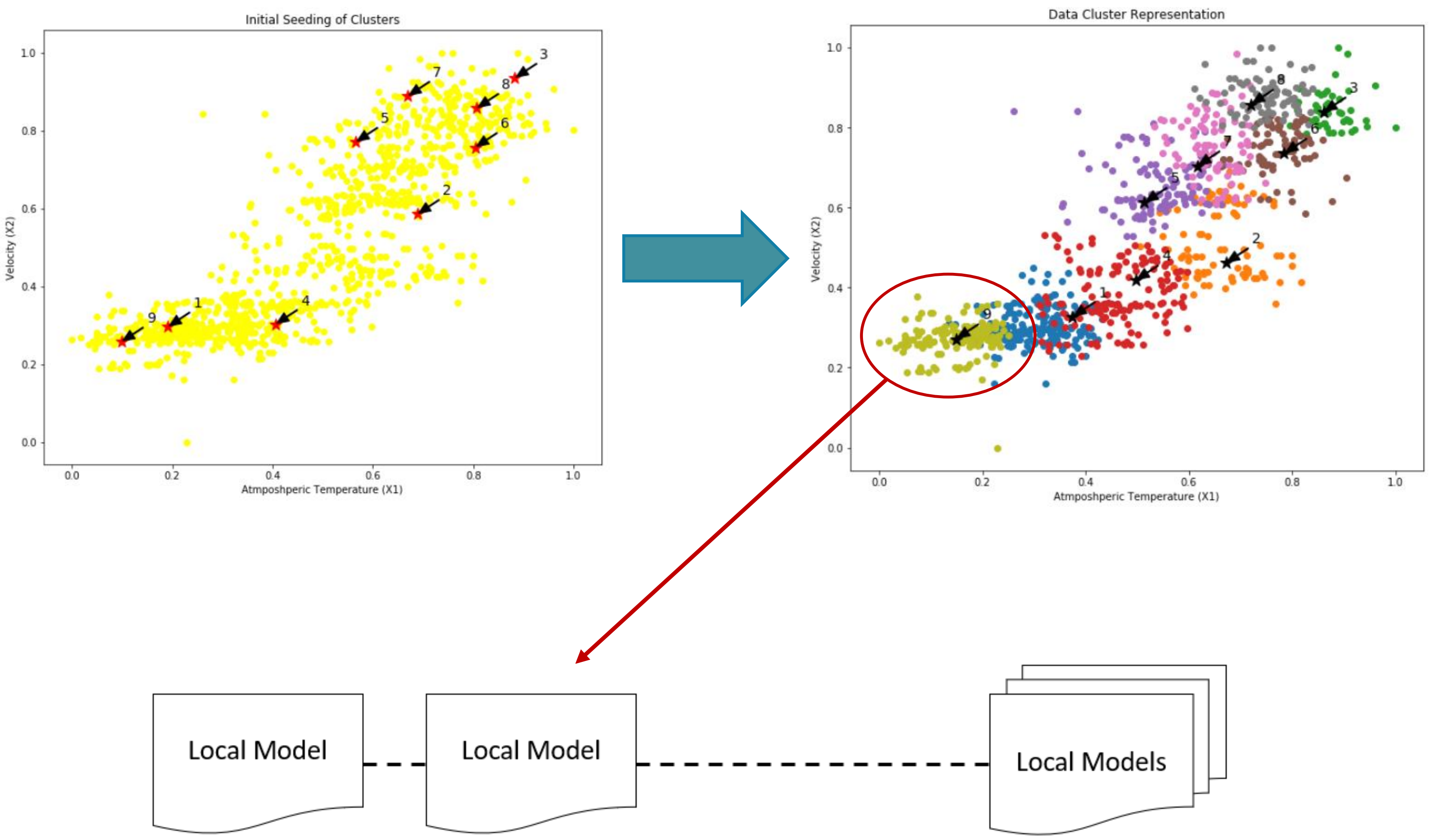
"Upon quantizing a large-scale dataset, each respective cluster can be associated with an individual predictive local model. Given this, the local models should predict more accurately than its counterpart global model over the entire data space."

HYPOTHESIS 2

"Upon observing user queries analytics, we can learn and make use of the principles of the RPCL and Learning Automata methodologies in order to extract the distributions and associations of the queries over the large-scale data space. Given this and Hypothesis 1; we should be able to create respective local models over the past issued query spaces providing high quality analytics."

HYPOTHESIS 1

FORMING LOCAL MODELS



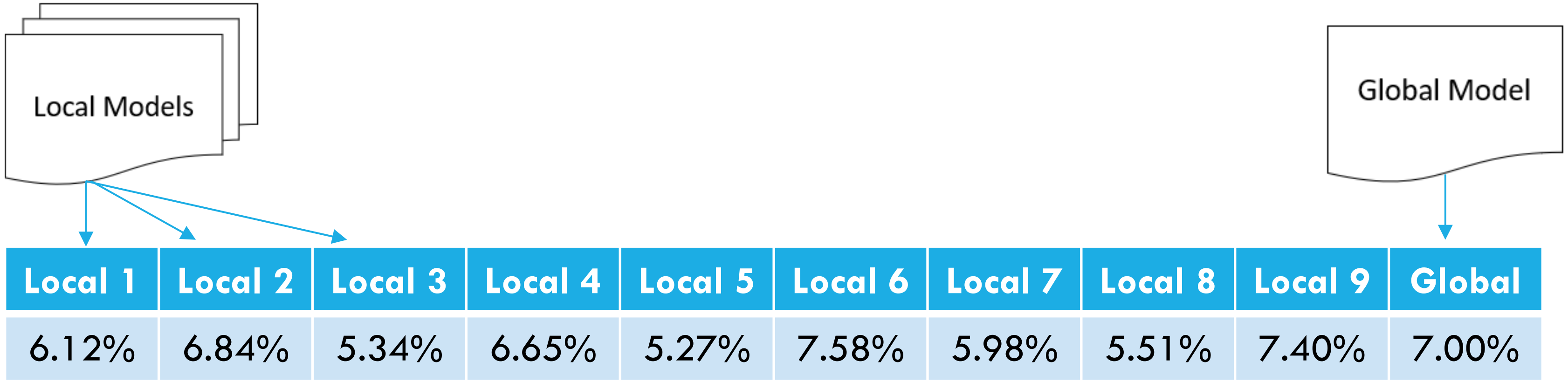
Local Model

Local Model

Local Models

HYPOTHESIS 1

RMSE: LOCAL VS GLOBAL

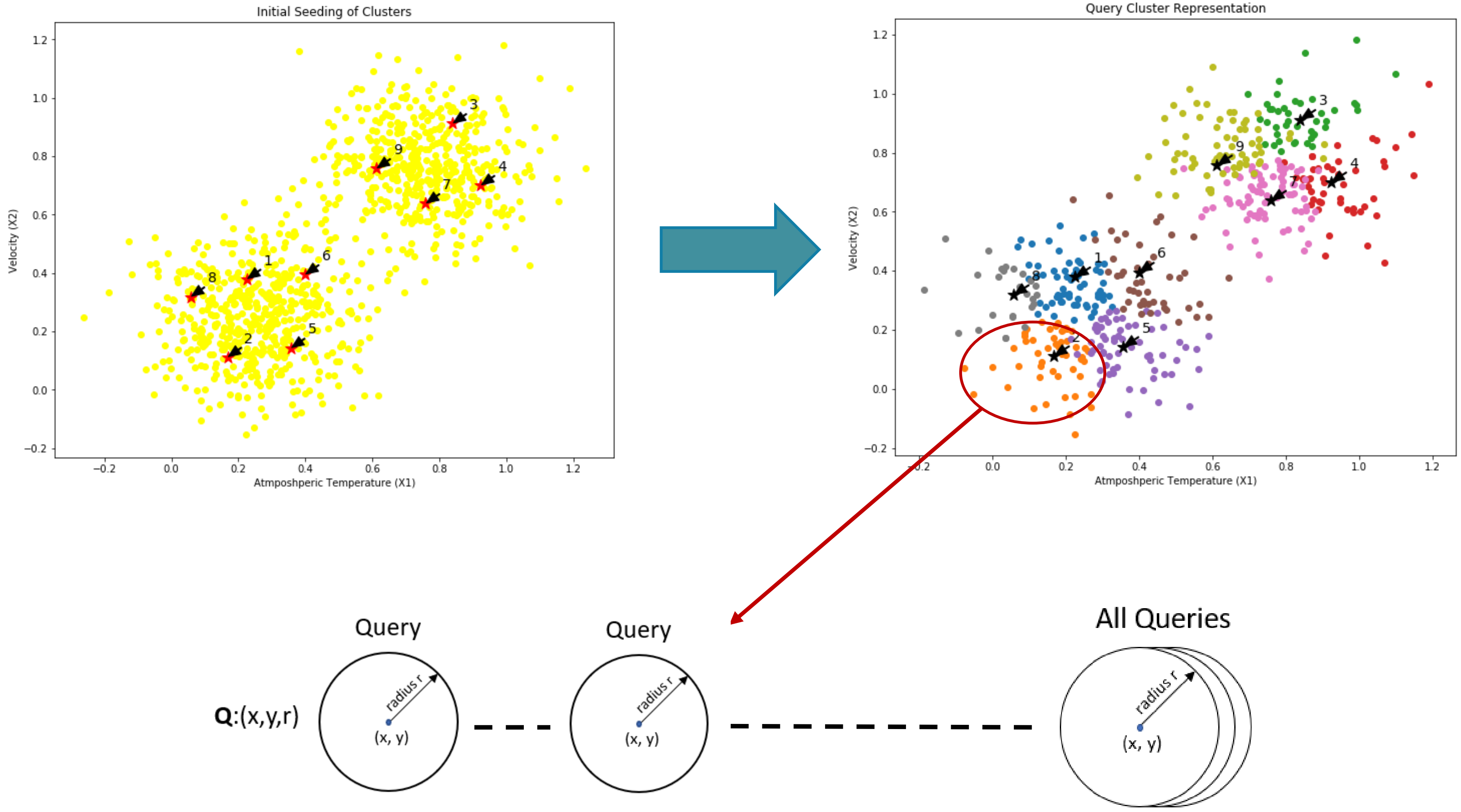


Local 1	Local 2	Local 3	Local 4	Local 5	Local 6	Local 7	Local 8	Local 9	Global
6.12%	6.84%	5.34%	6.65%	5.27%	7.58%	5.98%	5.51%	7.40%	7.00%

Each local model produces on average a **10%** increase in predictive performance with the use of less than **15%** of the entire dataset

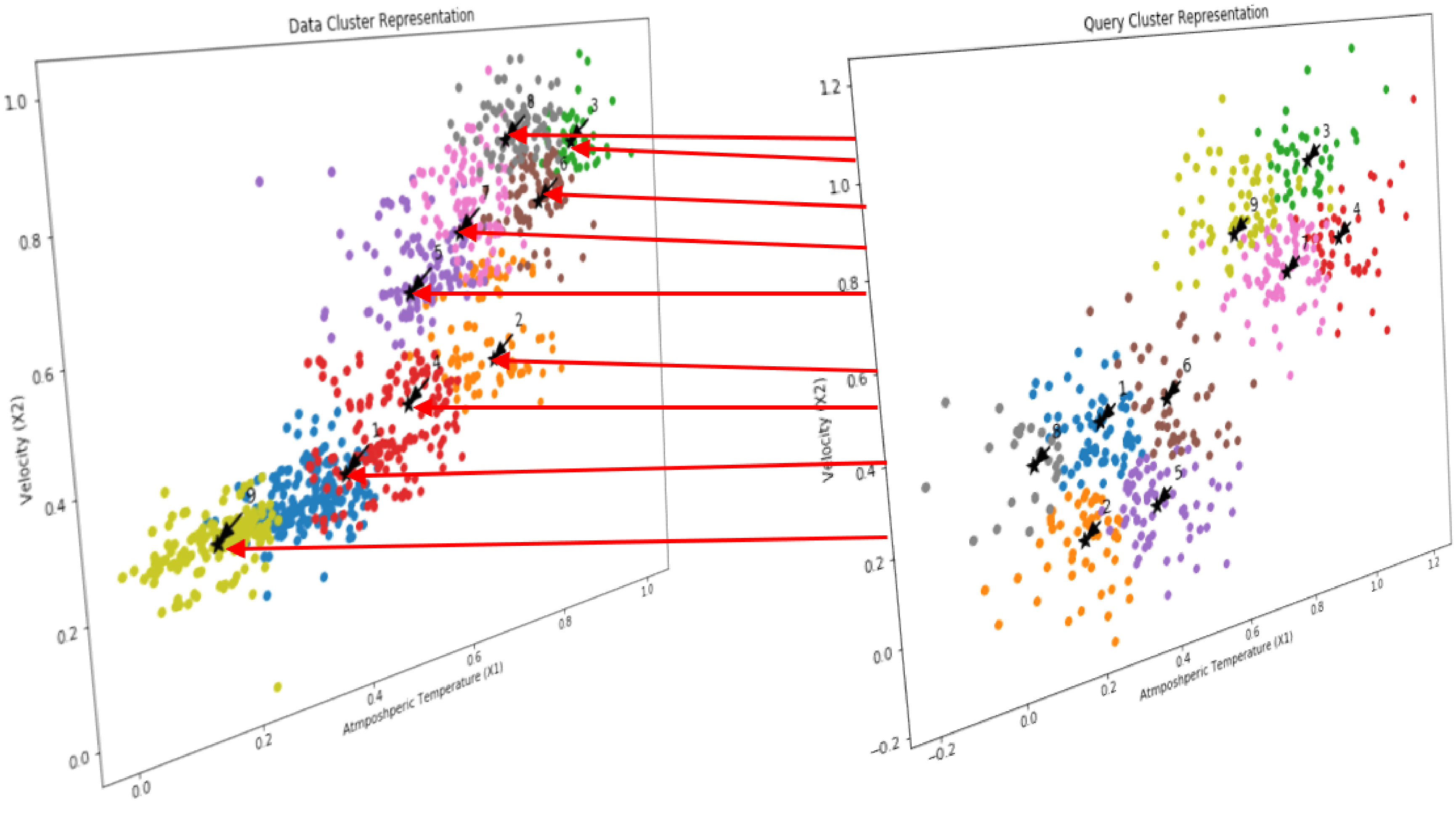
HYPOTHESIS 2

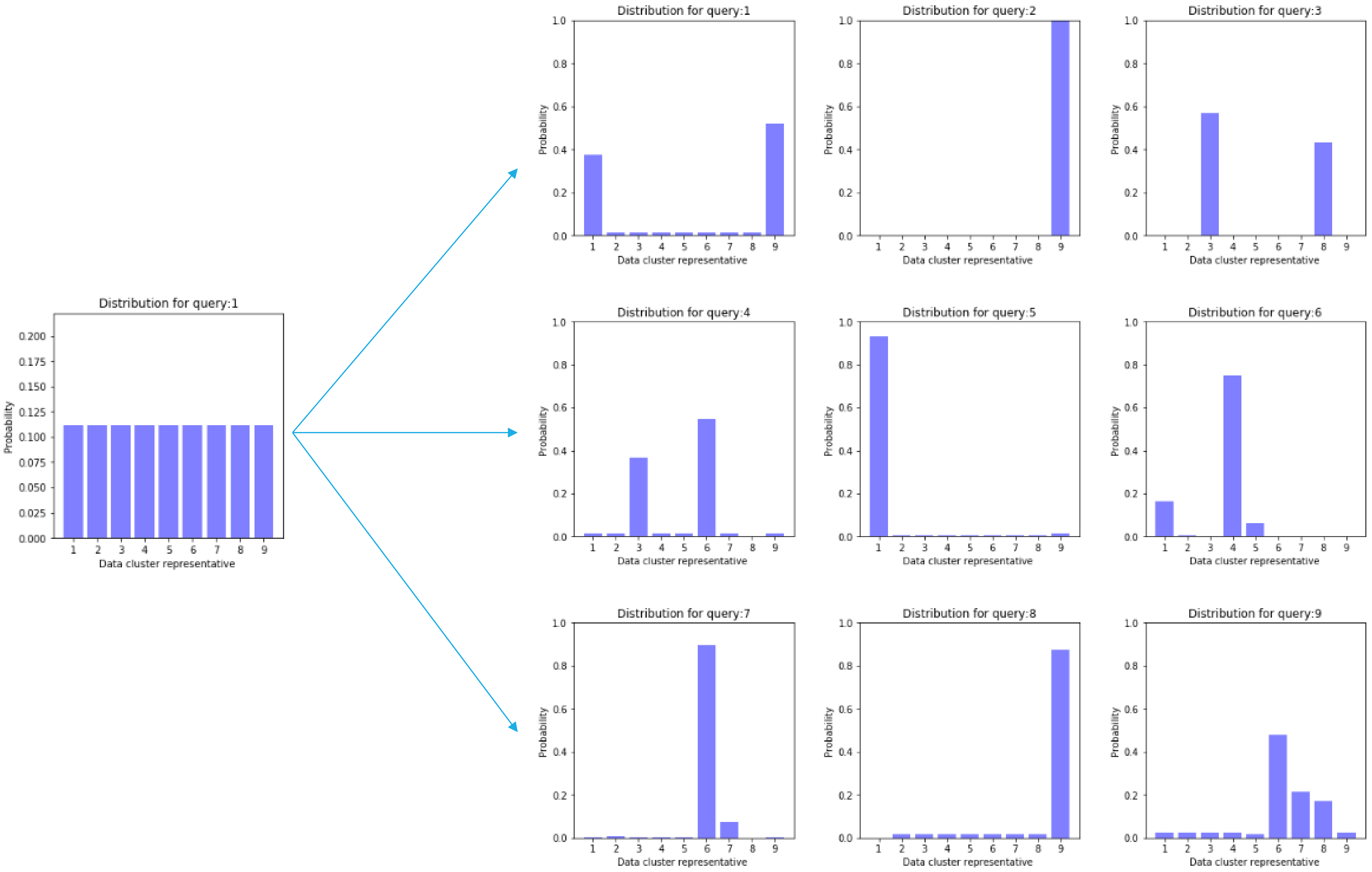
ANALYSING USER QUERIES



HYPOTHESIS 2

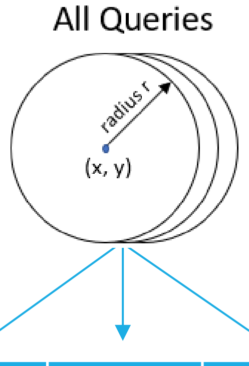
MAPPING USER QUERIES ONTO DATA

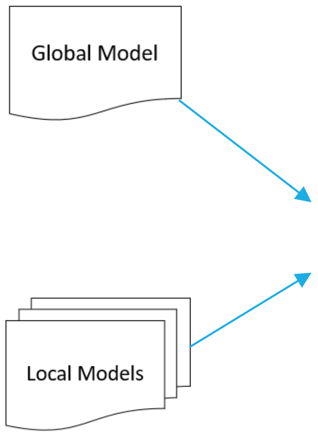



Distribution for query:1
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
Probability
1 2 3 4 5 6 7 8 9
Data cluster representative
Distribution for query:1
Distribution for query:2
Distribution for query:3
Distribution for query:4
Distribution for query:5
Distribution for query:6
Distribution for query:7
Distribution for query:8
Distribution for query:9
1.0
0.8
0.6
0.4
0.2
0.0
Probability
1 2 3 4 5 6 7 8 9
Data cluster representative

HYPOTHESIS 2

RMSE: LOCAL VS GLOBAL





Query 1	Query 2	Query 3	Query 4	Query 5	Query 6	Query 7	Query 8	Query 9
6.8%	8.9%	8.2%	7.8%	5.3%	7.0%	5.0%	8.1%	6.7%
6.8%	8.2%	6.3%	6.1%	5.4%	6.3%	5.6%	9.9%	6.2%

Set of local models hold on average a **4%** increase in performance with the use of less than **20%** of the entire data

CONCLUSION

- We were able to exploit important subspaces from query analytics
- Map each significant subspace into small predictive local models that:
  - Yielded higher predictive accuracy.
  - Made use of less data.
- Thus thinning the data space, and making use of less training data instances.