Distributed Data Mining

DIADIC Lab

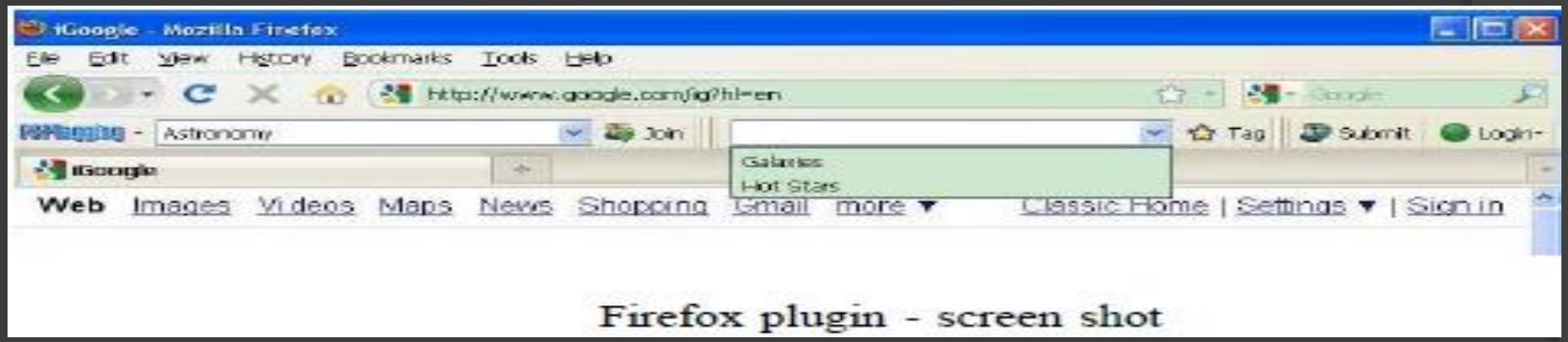
Sandipan Dey

UMBC
AN HONORS UNIVERSITY IN MARYLAND

Problems

- Randomized construction of distributed decision tree (collaborative text classifier)
- Outlier Detection
  - Parallel (Hadoop based)
  - P2P (DDMT / erlang)
- PADMini System

P2P Collaborative Text Classifier



Firefox plugin - screen shot

- Tag Text through plugin
- Generate feature vector
- Learn Classifier
- Centralized – privacy issue

Randomized Distributed Decision Tree Learning

- Greedy decision / synchronization overhead
- Random construction of trees (Wei Fan), ensemble learning
- Simple averaging $\underset{j \in C}{argmax} \frac{1}{N} \sum_{i=1}^{N} P_i(j|x).$

Randomized Distributed Decision Tree Learning

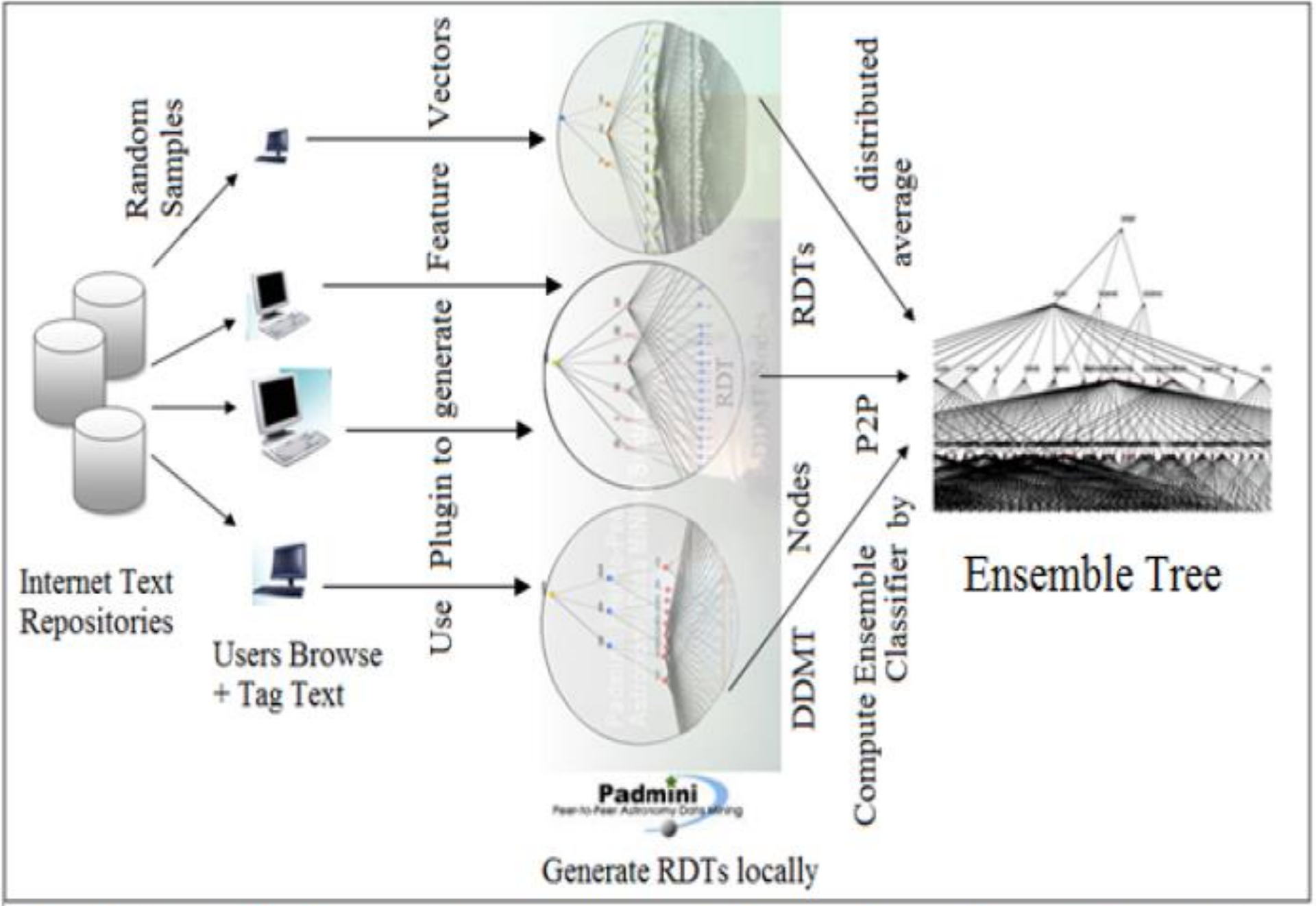
- Sample (WR) from population
- Locally generate RDTs
- Compute leaf class distributions
- Asynchronously (gossip) compute average posterior probability

Accuracy of the Ensemble

- Ensemble acc.= $1 - \exp\left(-\frac{1}{2p}\frac{\left(Np - \lfloor \frac{N}{2} \rfloor + 1\right)^2}{N}\right)$
  N = #RDTs, p = accuracy of an RDT

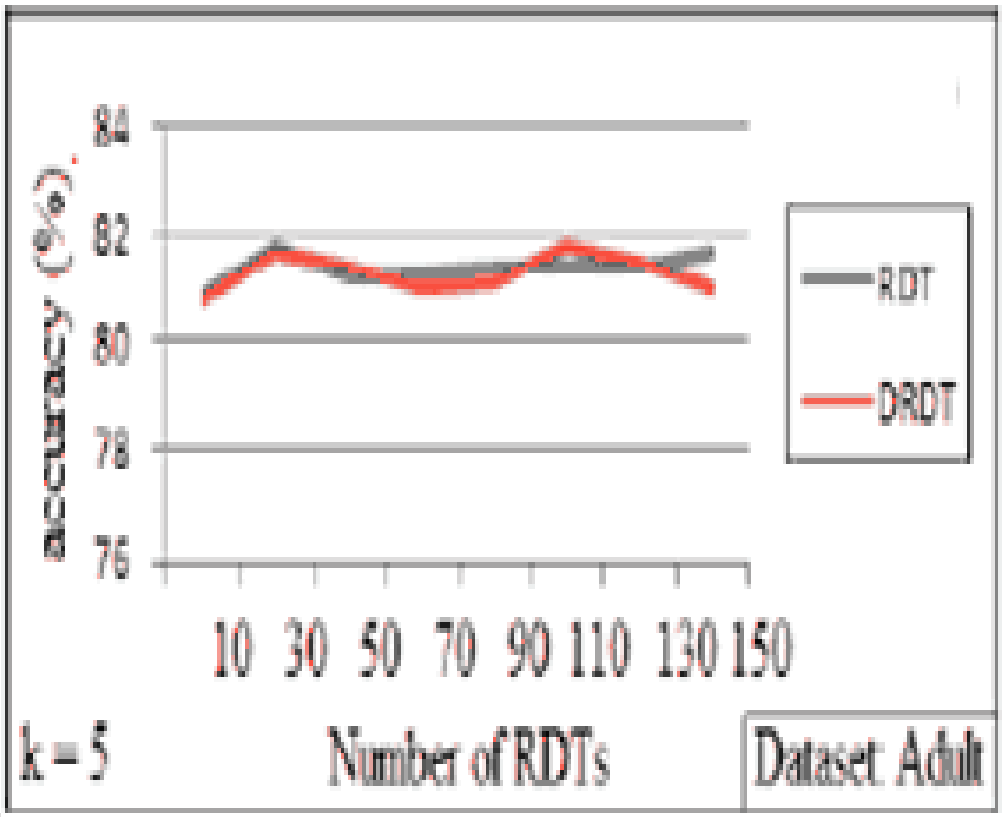
- Generalization Error of ensemble Classifier (Leo Breiman, Random Forest)

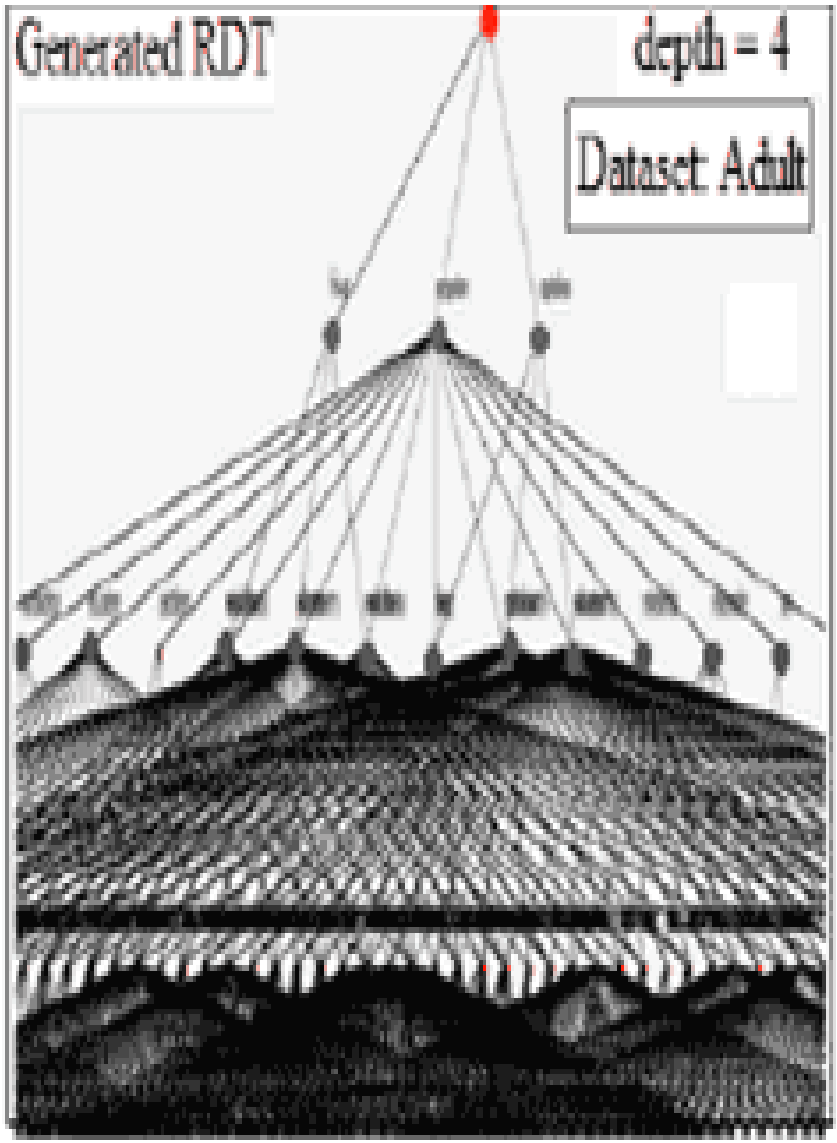
$$PE^* \leq \bar{\rho}\frac{1 - s^2}{s^2}$$



P2P Collaborative Text Classification DI

UMBC
AN HONORS UNIVERSITY IN MARYLAND



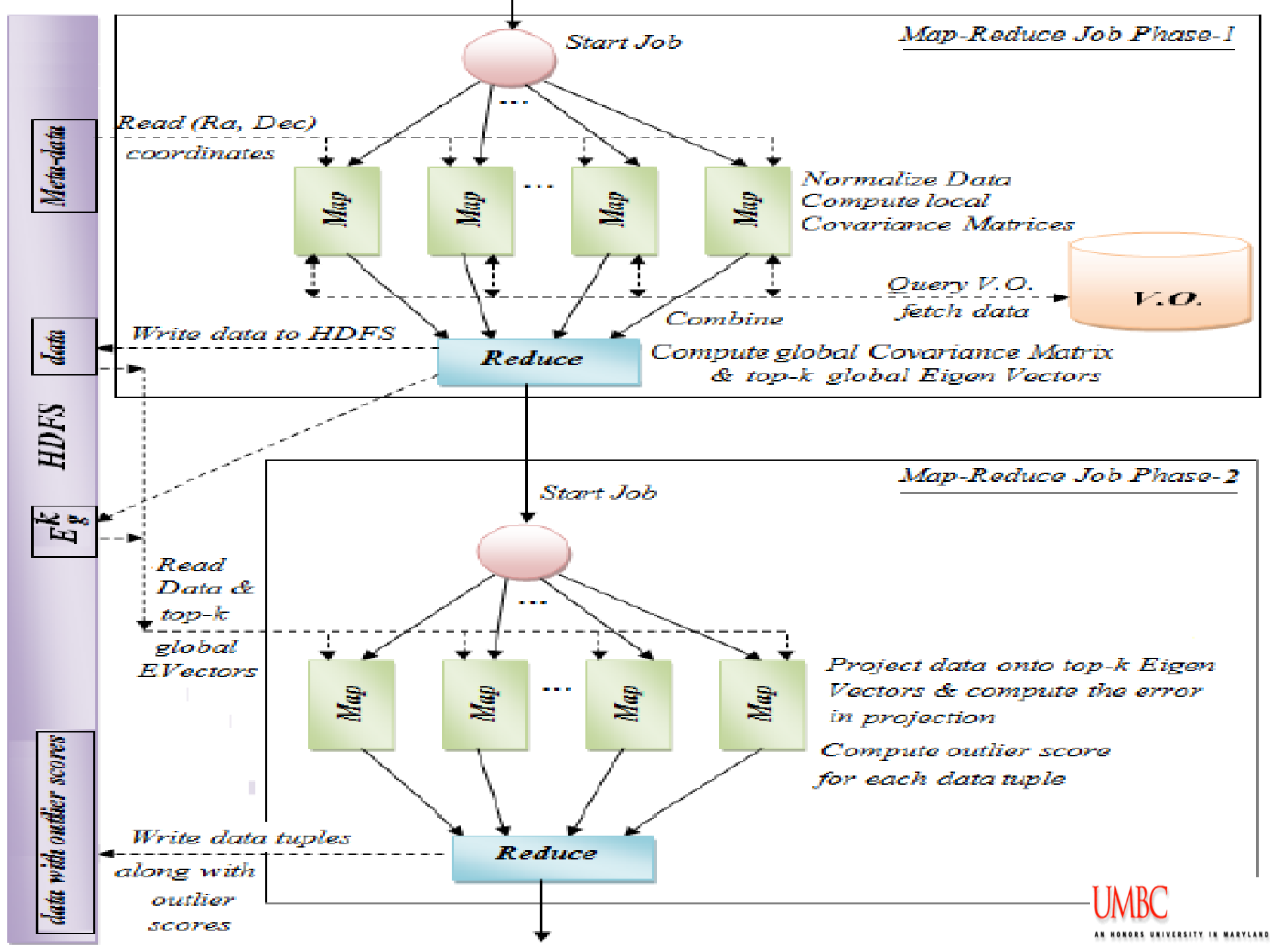


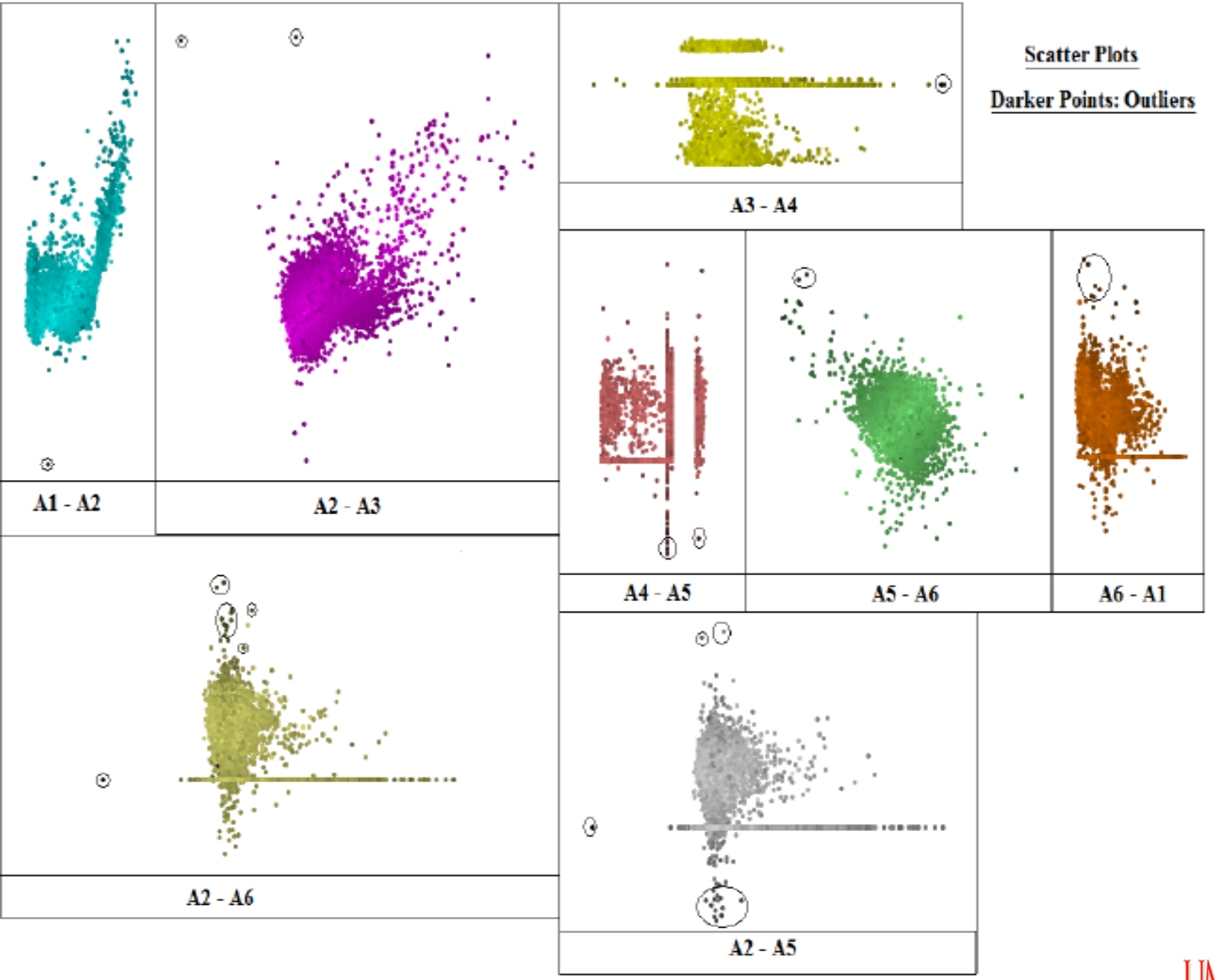
Experimental Results for DRDT and comparison with RDT

UMBC
AN HONORS UNIVERSITY IN MARYLAND

Distributed Outlier Detection in PADMini (using Hadoop)

- *Couple of Map-Reduce Phases*
- *Use Eigen Analysis to find outliers* $C_g = V_g \Lambda_g V_g^T$
- *Use additive decomposability of covariance matrix* $C_g = \frac{\sum_{i=1}^{N} m_i C_i}{\sum_{i=1}^{N} m_i}$
- *Compute outlier score for each tuple* $\hat{X}^i = X_i . \hat{V}_g^k . \hat{V}_g^{kT}$ $s_j^i = \frac{||X_j^i - \hat{X}_j^i||_2}{\max_j ||X_j^i - \hat{X}_j^i||_2}$


Meta-data
data
HDFS
E_v^k
data with outlier scores
Map-Reduce Job Phase-1
Start Job
Read (Ra, Dec) coordinates
Map
Map
Map
Map
Normalize Data
Compute local
Covariance Matrices
Query V.O.
fetch data
V.O.
Combine
Write data to HDFS
Reduce
Compute global Covariance Matrix
& top-k global Eigen Vectors
Map-Reduce Job Phase-2
Start Job
Read
Data &
top-k
global
EVectors
Map
Map
Map
Map
Project data onto top-k Eigen
Vectors & compute the error
in projection
Compute outlier score
for each data tuple
Write data tuples
along with
outlier
scores
Reduce



Scatter plots with different 2-attributes combinations and color coded displa

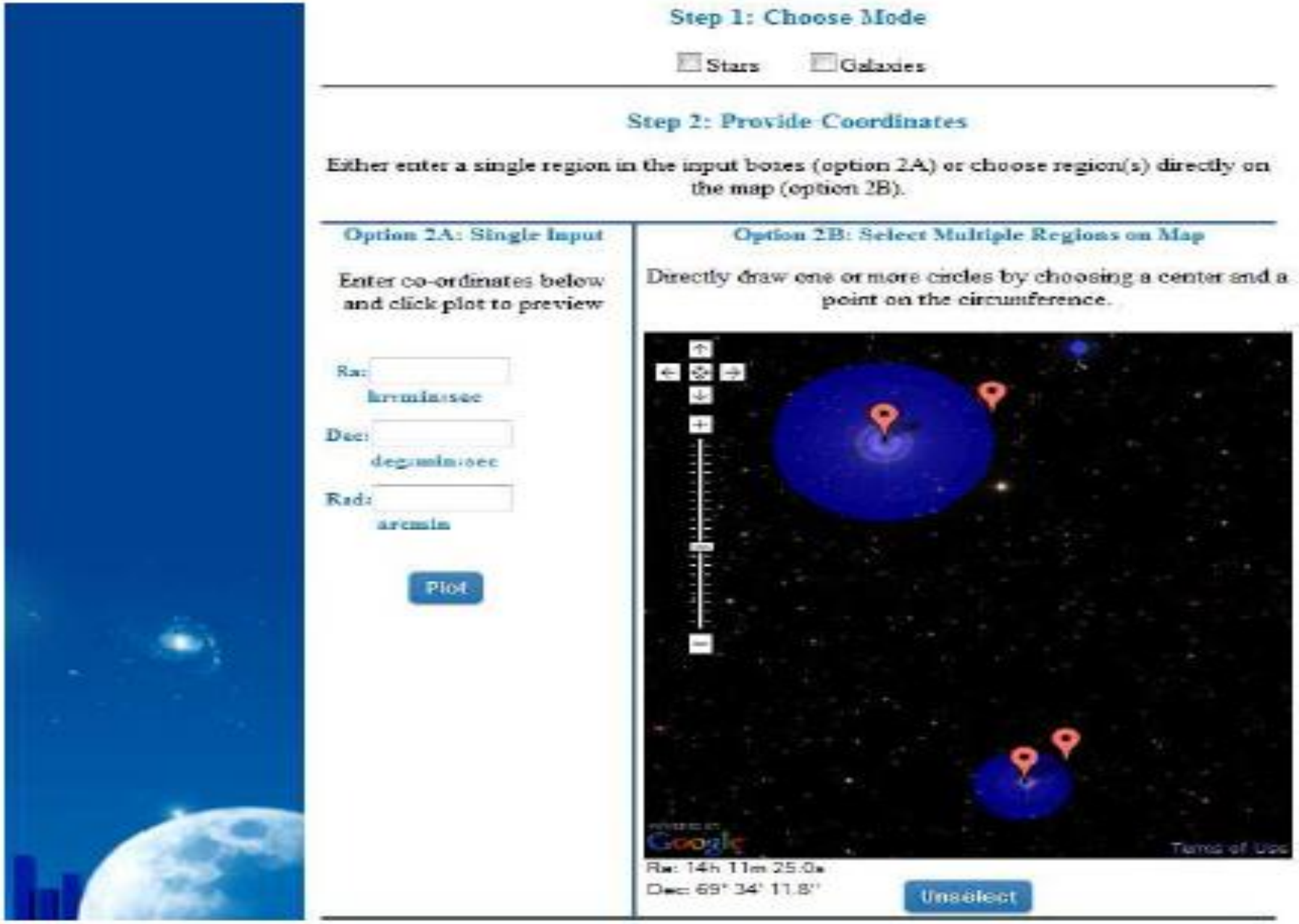


Fig. 3. Google Maps interface for selecting regions of the sky

Distributed Outlier Detection

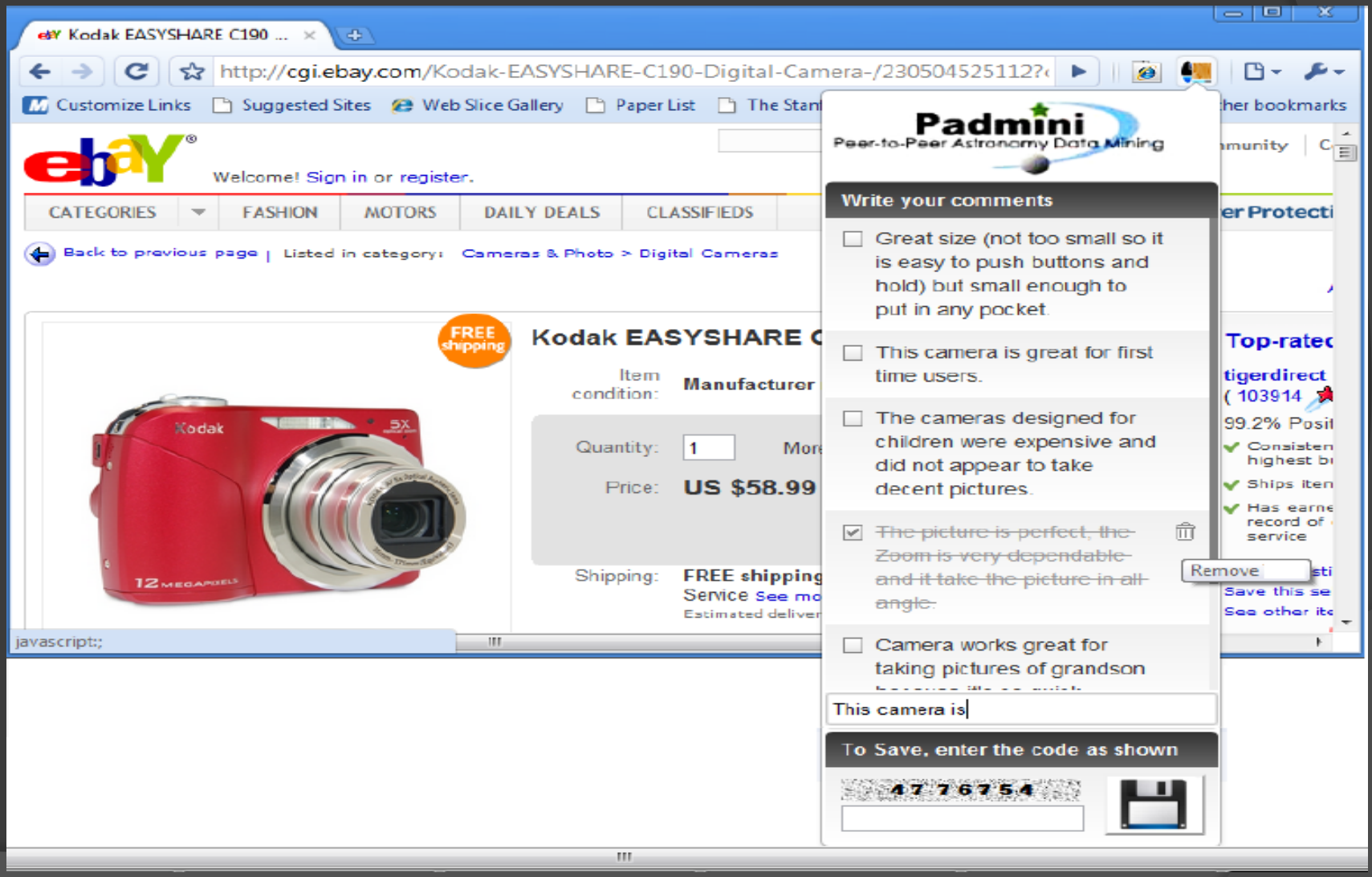
- ***Using Entropy Minimization / Spectral Decomposition – define Outlierness***

$$\min_{O \subseteq D} H(D - O) \quad s.t.\ |O| = k.$$

- ***Combining local eigenvectors to form global eigenvectors***

- ***How to find weights?***

Chrome Extension P2P Commenting



Thank You