

Distributed Data Mining

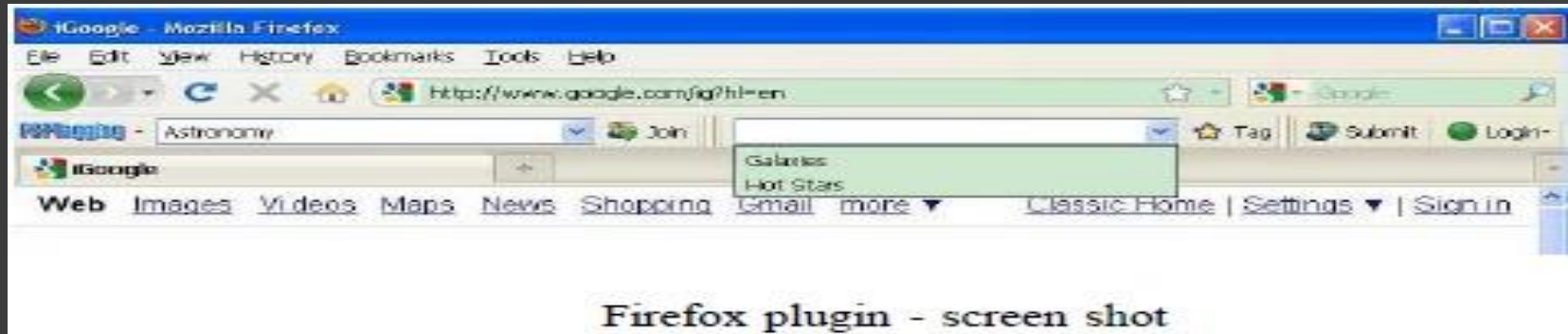
DIADIC Lab

Sandipan Dey

Problems

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

P2P Collaborative Text Classifier



- ◉ 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

$$\operatorname{argmax}_{j \in C} \frac{1}{N} \sum_{i=1}^N P_i(j|x).$$

Randomized Distributed Decision Tree Learning

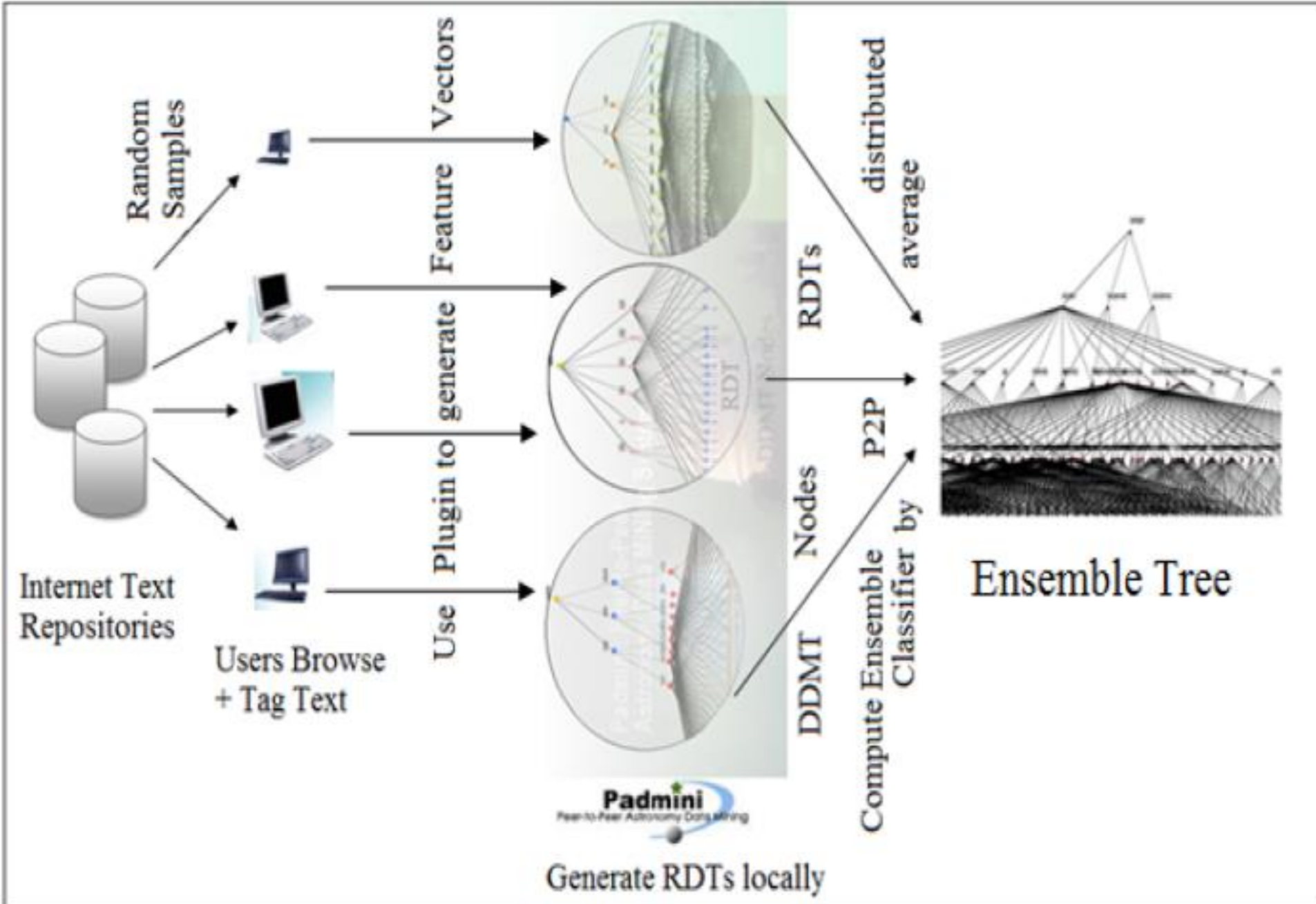
- Sample (WR) from population
- Locally generate RDTs
- Compute leaf class distributions
- Asynchronously (gossip) compute average posterior probability

Accuracy of the Ensemble

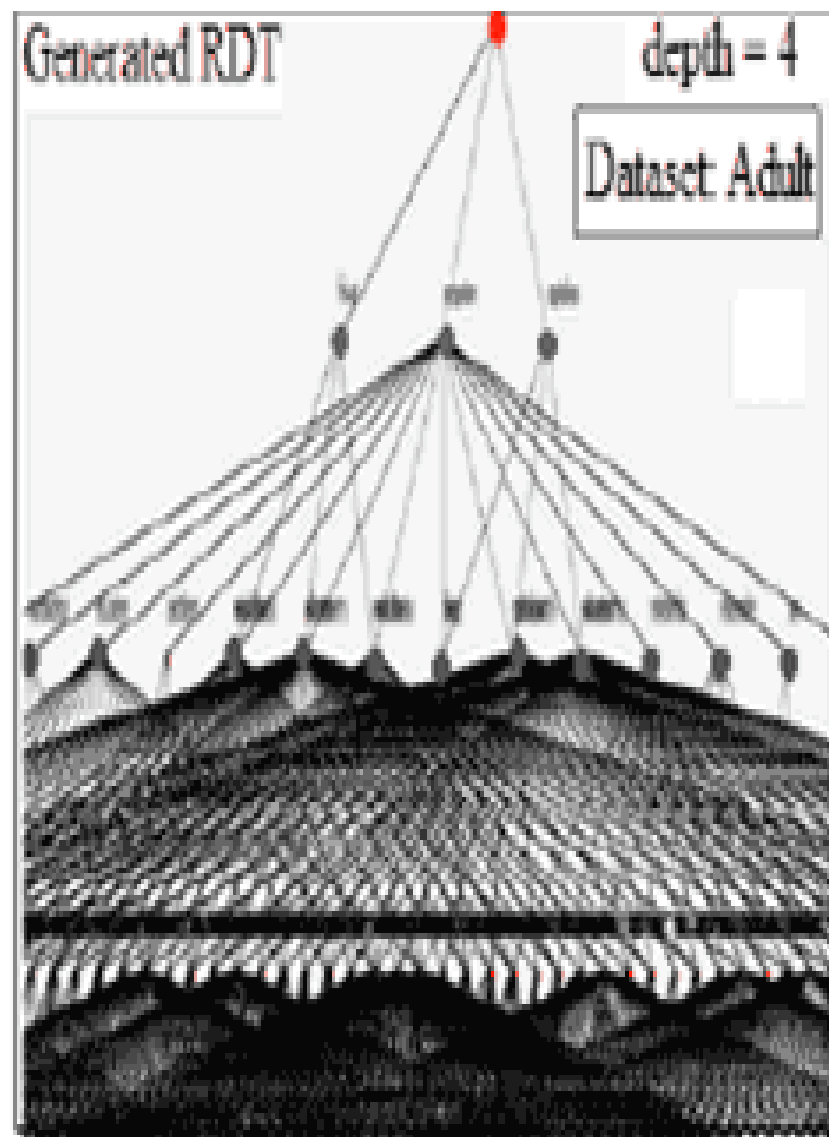
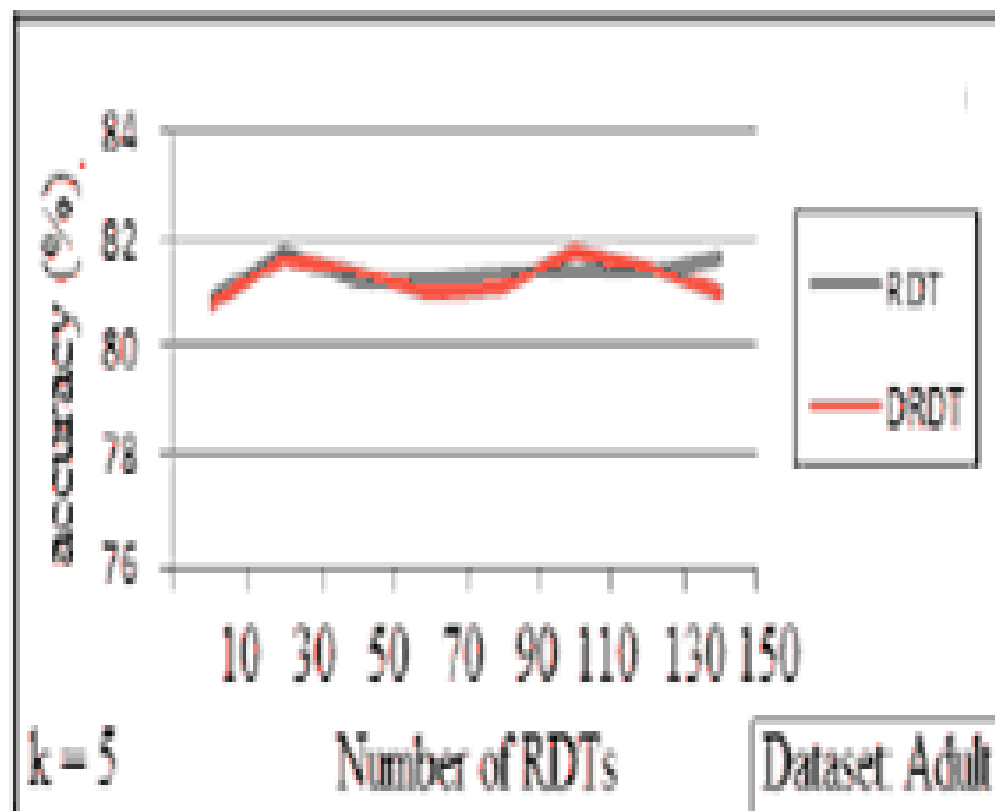
- Ensemble acc. = $1 - \exp\left(-\frac{1}{2p} \frac{(Np - \lfloor \frac{N}{2} \rfloor + 1)^2}{N}\right)$
N = #RDTs, p = accuracy of an RDT

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

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



P2P Collaborative Text Classification DI



Experimental Results for DRDT and comparison with RDT

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*

$$\hat{X}^i = X_i \cdot \hat{V}_g^k \cdot \hat{V}_g^{kT}$$

$$s_j^i = \frac{\|X_j^i - \hat{X}_j^i\|_2}{\max_j \|X_j^i - \hat{X}_j^i\|_2}$$

Meta-data

data

HDFS

E_g^k

data with outlier scores

Read (Ra, Dec)
coordinates

Map

Map

Map

Map

Normalize Data
Compute local
Covariance Matrices

Query V.O.
fetch data



V.O.

Write data to HDFS

Reduce

Combine

Compute global Covariance Matrix
& top-k global Eigen Vectors

Start Job

Read
Data &
top-k
global
E Vectors

Map

Map

Map

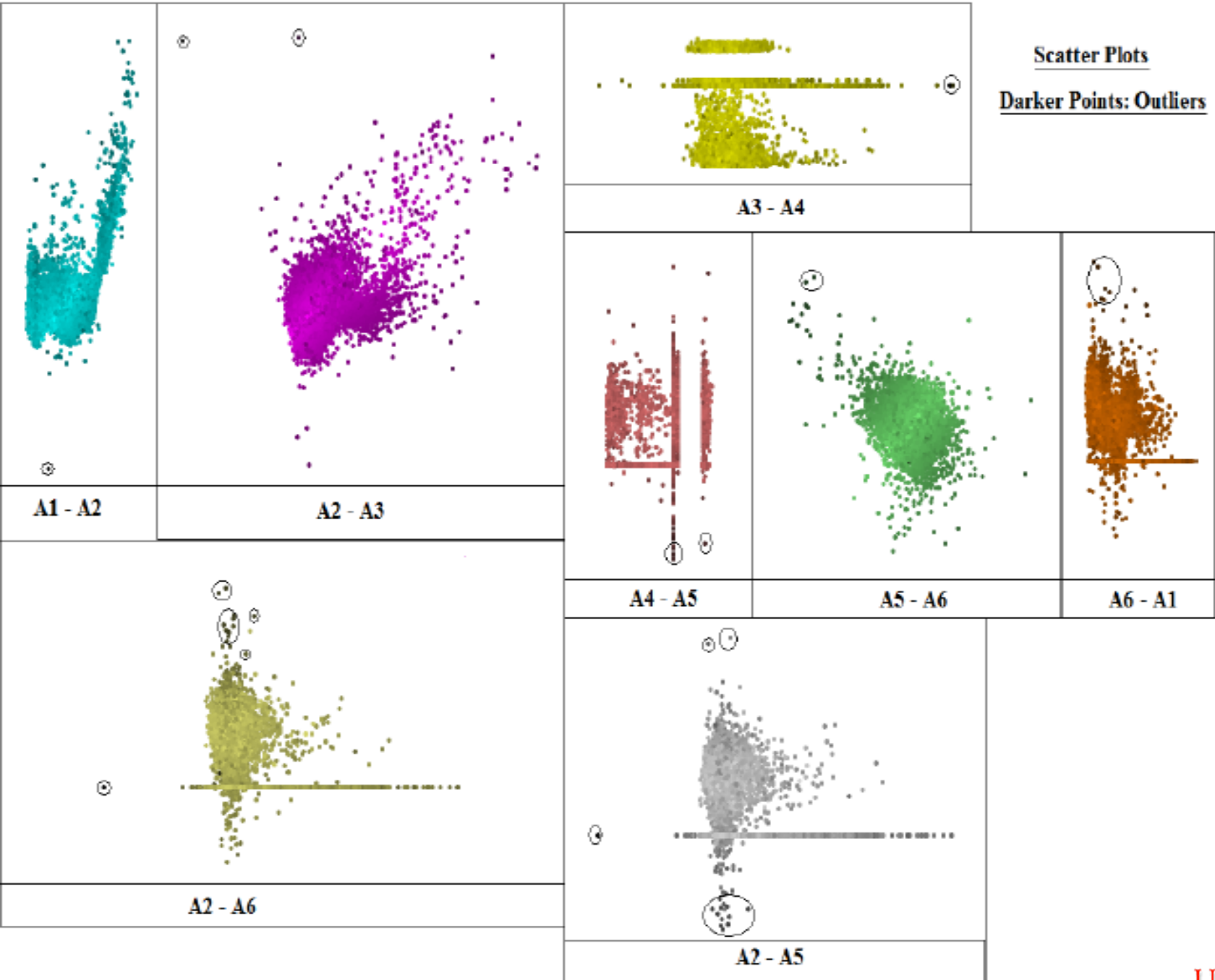
Map

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



Step 1: Choose Mode

☐ Stars ☐ Galaxies

Step 2: Provide Coordinates

Either enter a single region in the input boxes (option 2A) or choose region(s) directly on the map (option 2B).

Option 2A: Single Input

Enter co-ordinates below and click plot to preview

Ra:
hrmin:sec

Dec:
deg:min:sec

Ra:
arcmin

Plot

Option 2B: Select Multiple Regions on Map

Directly draw one or more circles by choosing a center and a point on the circumference.



Powered by Google

Ra: 14h 11m 25.0s
Dec: 69° 34' 11.8"

Unselect

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

Distributed Outlier Detection

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

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

- *Combining local eigenvectors to form global eigenvectors*
- *How to find weights?*

Chrome Extension P2P Commenting

ebay Kodak EASYSHARE C190 ...

http://cgi.ebay.com/Kodak-EASYSHARE-C190-Digital-Camera-/230504525112?...

Customize Links Suggested Sites Web Slice Gallery Paper List The Stand...

ebay® Welcome! Sign in or register.

CATEGORIES FASHION MOTORS DAILY DEALS CLASSIFIEDS

Back to previous page | Listed in category: Cameras & Photo > Digital Cameras

Kodak EASYSHARE C190

12 MEGAPixels

FREE shipping

Item condition: **Manufacturer**

Quantity: 1 More

Price: **US \$58.99**

Shipping: **FREE shipping** Service See more Estimated delivery

Padmini
Peer-to-Peer Astronomy Data Mining

Write your comments

- ☐ Great size (not too small so it is easy to push buttons and hold) but small enough to put in any pocket.
- ☐ This camera is great for first time users.
- ☐ The cameras designed for children were expensive and did not appear to take decent pictures.
- ☒ The picture is perfect, the Zoom is very dependable and it take the picture in all angle.
- ☐ Camera works great for taking pictures of grandson because the picture is...

This camera is|

To Save, enter the code as shown

47 7 6 7 5 4

Remove

Save this se
See other ite

Thank You