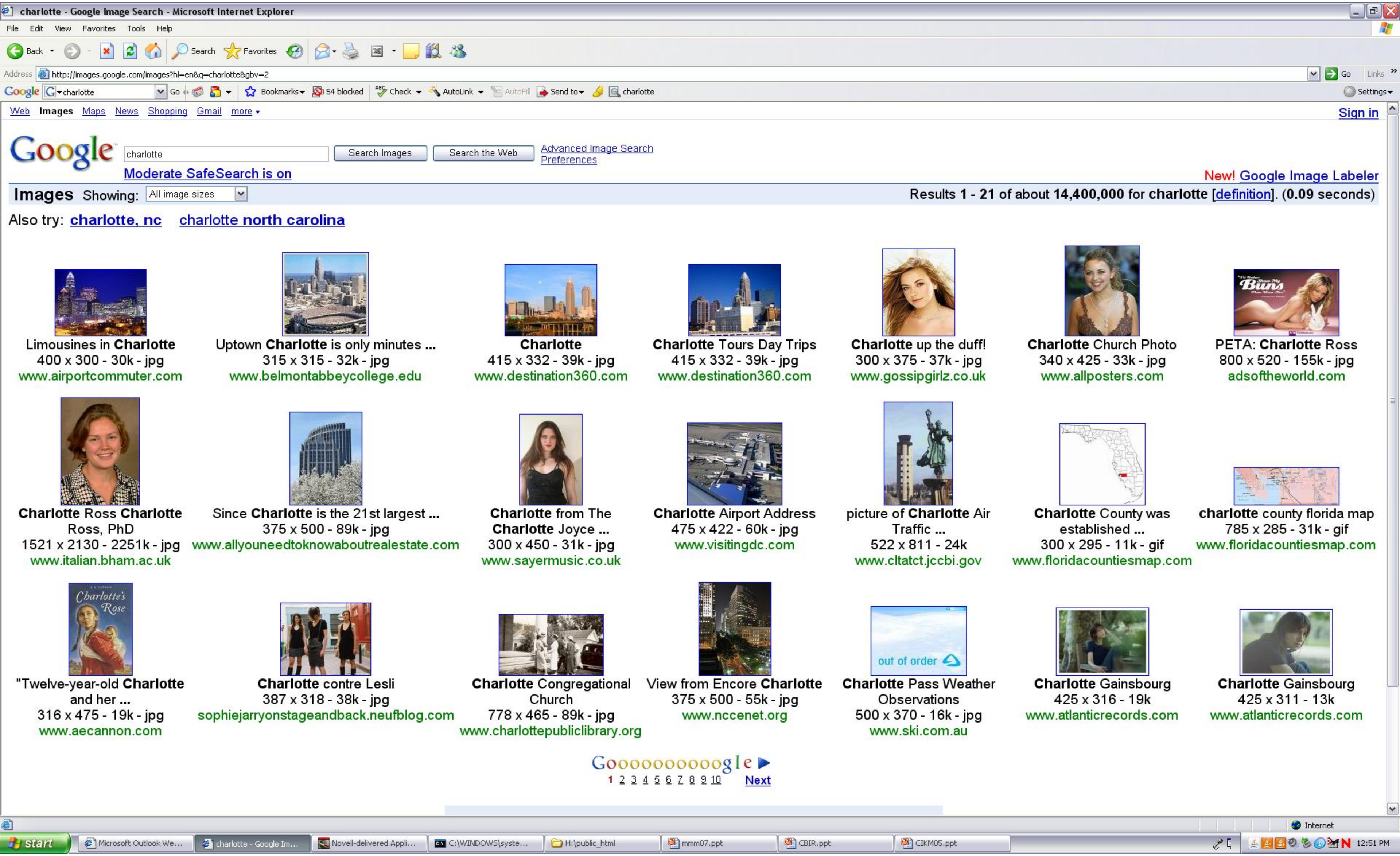
Learning from Large-Scale Online Images
Part I: Junk Image Filtering

Jianping Fan
Department of Computer Science
UNC-Charlotte

Course Website:
http://webpages.uncc.edu/jfan/itcs5152.html

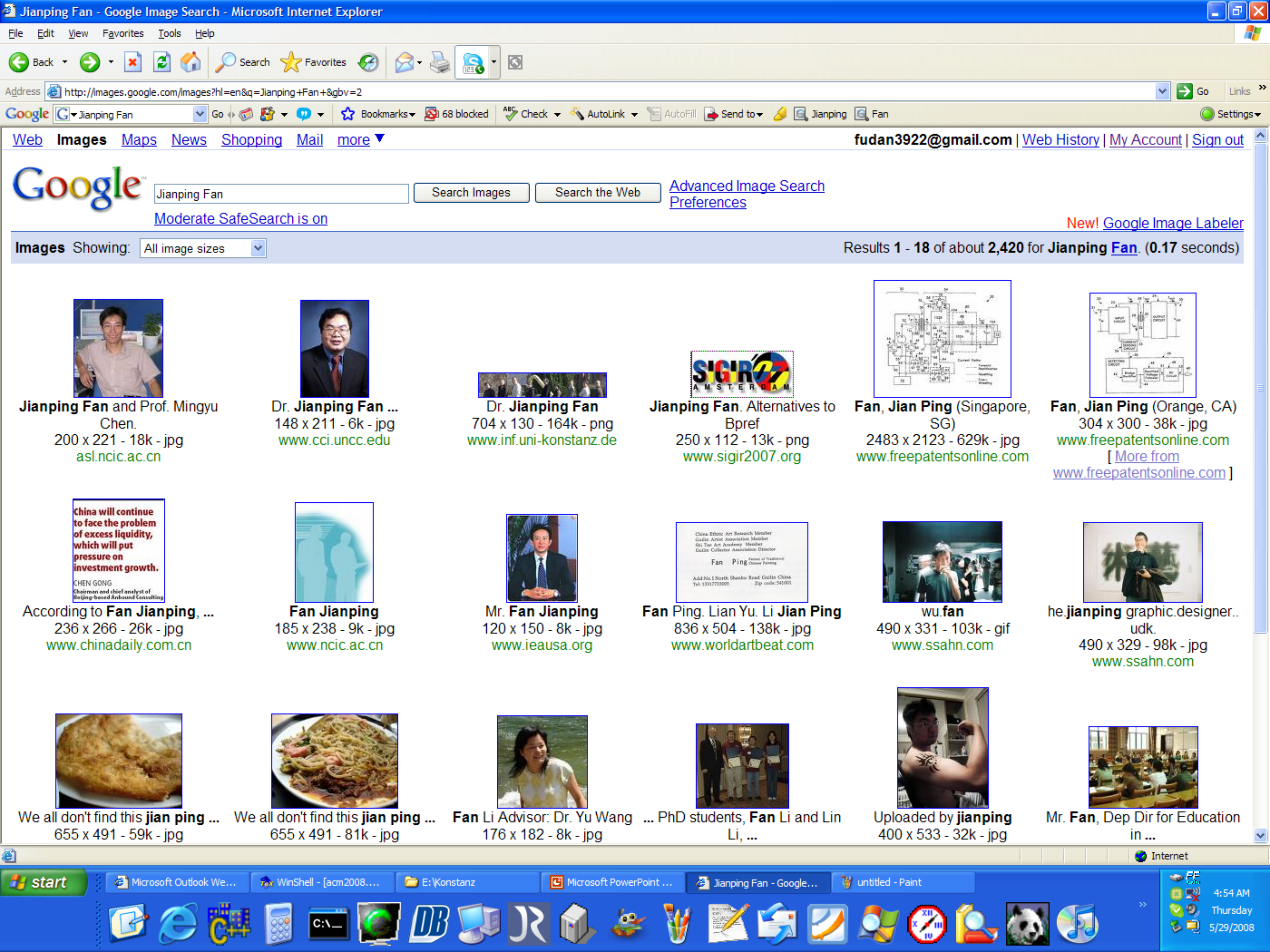
1. Motivation

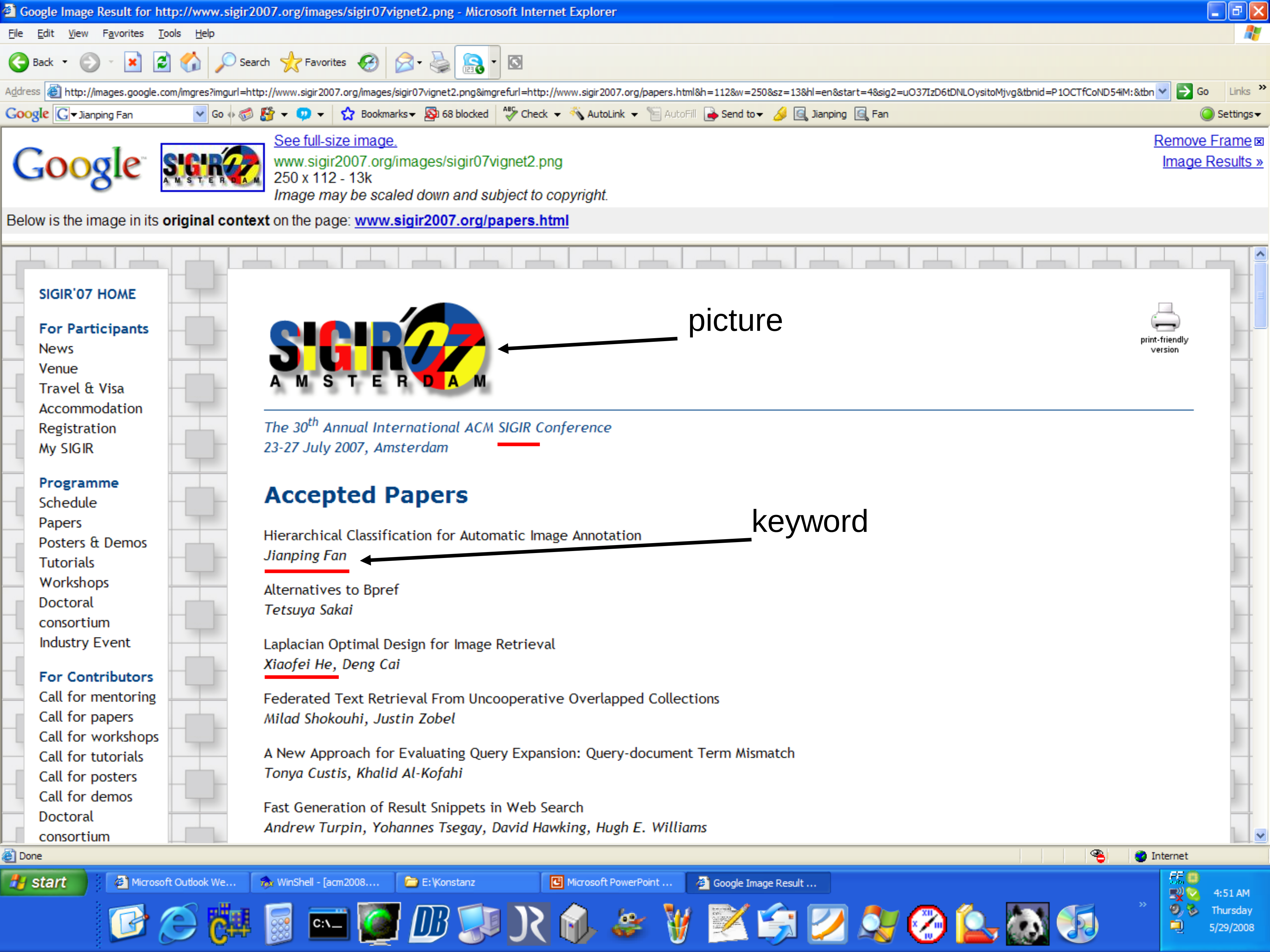


Large amount of junk images may stop users!!

How Google indexes the images?

- Keywords from associated text document
- File names
- URL
- ....

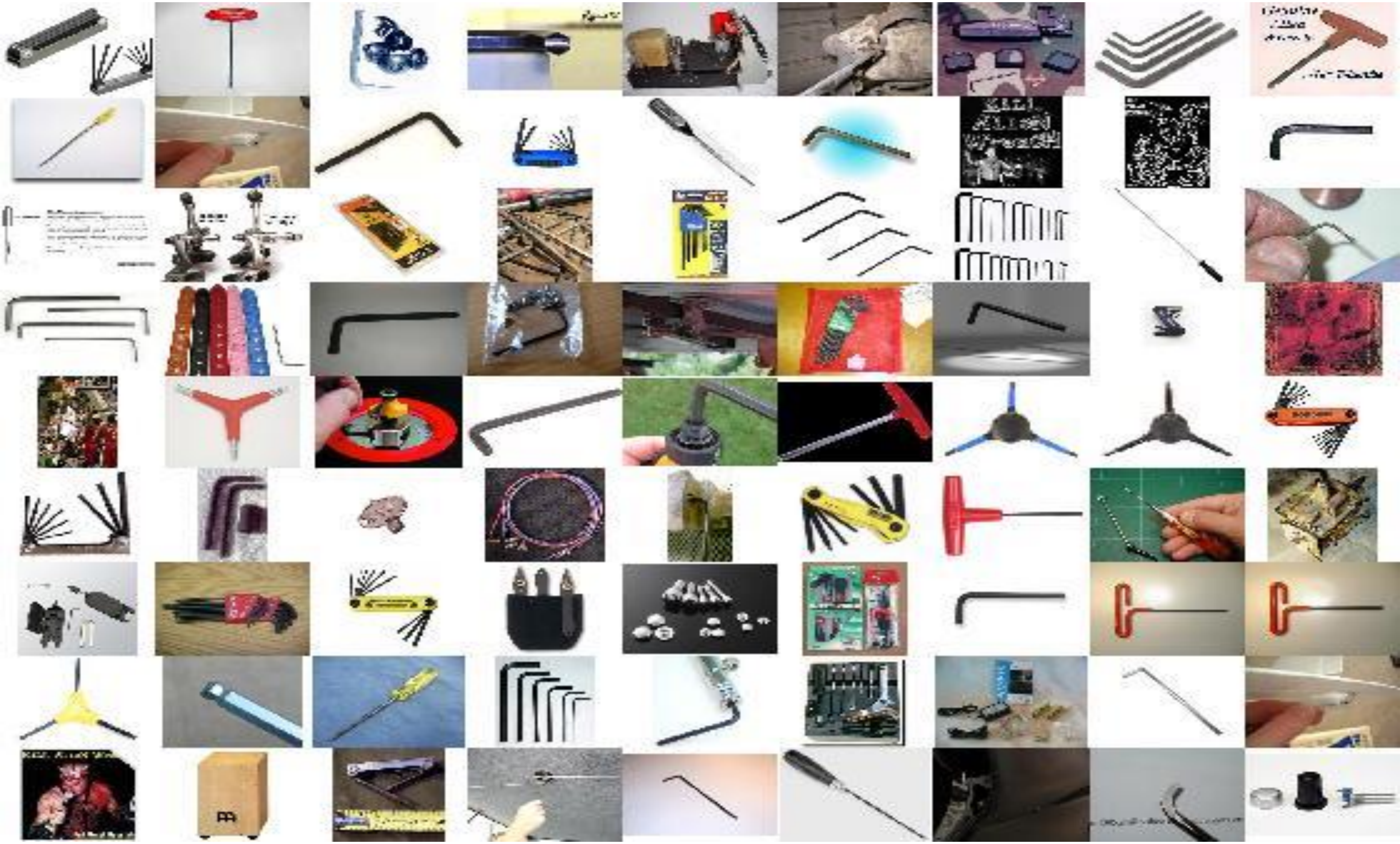

Jianping Fan - Google Image Search - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://images.google.com/images?hl=en&q=Jianping+Fan+&gbv=2
Go Links
Jianping Fan Go Bookmarks 68 blocked Check AutoLink AutoFill Send to Jianping Fan Settings
Web Images Maps News Shopping Mail more
fudan3922@gmail.com | Web History | My Account | Sign out
Jianping Fan
Search Images
Search the Web
Advanced Image Search
Preferences
Moderate SafeSearch is on
New! Google Image Labeler
Images Showing: All image sizes
Results 1 - 18 of about 2,420 for Jianping Fan. (0.17 seconds)
Jianping Fan and Prof. Mingyu Chen.
200 x 221 - 18k - jpg
asl.ncic.ac.cn
Dr. Jianping Fan ...
148 x 211 - 6k - jpg
www.cci.uncc.edu
Dr. Jianping Fan
704 x 130 - 164k - png
www.inf.uni-konstanz.de
SIGIR 07 AMSTERDAM
Jianping Fan. Alternatives to Bpref
250 x 112 - 13k - png
www.sigir2007.org
Fan, Jian Ping (Singapore, SG)
2483 x 2123 - 629k - jpg
www.freepatentsonline.com
Fan, Jian Ping (Orange, CA)
304 x 300 - 38k - jpg
www.freepatentsonline.com
[More from www.freepatentsonline.com]
China will continue to face the problem of excess liquidity, which will put pressure on investment growth.
CHEN GONG
According to Fan Jianping, ...
236 x 266 - 26k - jpg
www.chinadaily.com.cn
Fan Jianping
185 x 238 - 9k - jpg
www.ncic.ac.cn
Mr. Fan Jianping
120 x 150 - 8k - jpg
www.ieausa.org
Fan Ping. Lian Yu. Li Jian Ping
836 x 504 - 138k - jpg
www.worldartbeat.com
wu.fan
490 x 331 - 103k - gif
www.ssahn.com
he.jianping graphic.designer.. udk.
490 x 329 - 98k - jpg
www.ssahn.com
We all don't find this jian ping ...
655 x 491 - 59k - jpg
We all don't find this jian ping ...
655 x 491 - 81k - jpg
Fan Li Advisor: Dr. Yu Wang
176 x 182 - 8k - jpg
... PhD students, Fan Li and Lin Li, ...
Uploaded by jianping
400 x 533 - 32k - jpg
Mr. Fan, Dep Dir for Education in ...
Internet
start
Microsoft Outlook We...
WinShell - [acm2008....
E:\Konstanz
Microsoft PowerPoint ...
Jianping Fan - Google...
untitled - Paint
4:54 AM
Thursday
5/29/2008


Google Image Result for http://www.sigir2007.org/images/sigir07vignet2.png - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://images.google.com/imgres?imgurl=http://www.sigir2007.org/images/sigir07vignet2.png&imgrefurl=http://www.sigir2007.org/papers.html&h=112&w=250&sz=13&hl=en&start=4&sig2=uO37IzD6tDNLOysitoMjvg&tbnid=P1OCTfCoND54iM:&tbn
Google Jianping Fan
See full-size image.
www.sigir2007.org/images/sigir07vignet2.png
250 x 112 - 13k
Image may be scaled down and subject to copyright.
Remove Frame
Image Results »
Below is the image in its original context on the page: www.sigir2007.org/papers.html
SIGIR'07 HOME
For Participants
News
Venue
Travel & Visa
Accommodation
Registration
My SIGIR
Programme
Schedule
Papers
Posters & Demos
Tutorials
Workshops
Doctoral consortium
Industry Event
For Contributors
Call for mentoring
Call for papers
Call for workshops
Call for tutorials
Call for posters
Call for demos
Doctoral consortium
SIGIR'07 AMSTERDAM
picture
print-friendly version
The 30th Annual International ACM SIGIR Conference
23-27 July 2007, Amsterdam
Accepted Papers
keyword
Hierarchical Classification for Automatic Image Annotation
Jianping Fan
Alternatives to Bpref
Tetsuya Sakai
Laplacian Optimal Design for Image Retrieval
Xiaofei He, Deng Cai
Federated Text Retrieval From Uncooperative Overlapped Collections
Milad Shokouhi, Justin Zobel
A New Approach for Evaluating Query Expansion: Query-document Term Mismatch
Tonya Custis, Khalid Al-Kofahi
Fast Generation of Result Snippets in Web Search
Andrew Turpin, Yohannes Tsegay, David Hawking, Hugh E. Williams
Done
Internet
start
4:51 AM
Thursday
5/29/2008

1. Motivation

- **Why Google Images Has so Many Junk Images?**
  - Many keywords are available on the associated text documents, on the other hand, image semantics is interpreted by one or multiple related keywords;
  - Correspondence between the keywords and image semantics is not one-to-one or even there is not exact correspondence between them;
  - Keywords are normally ambiguity.

1. Motivation



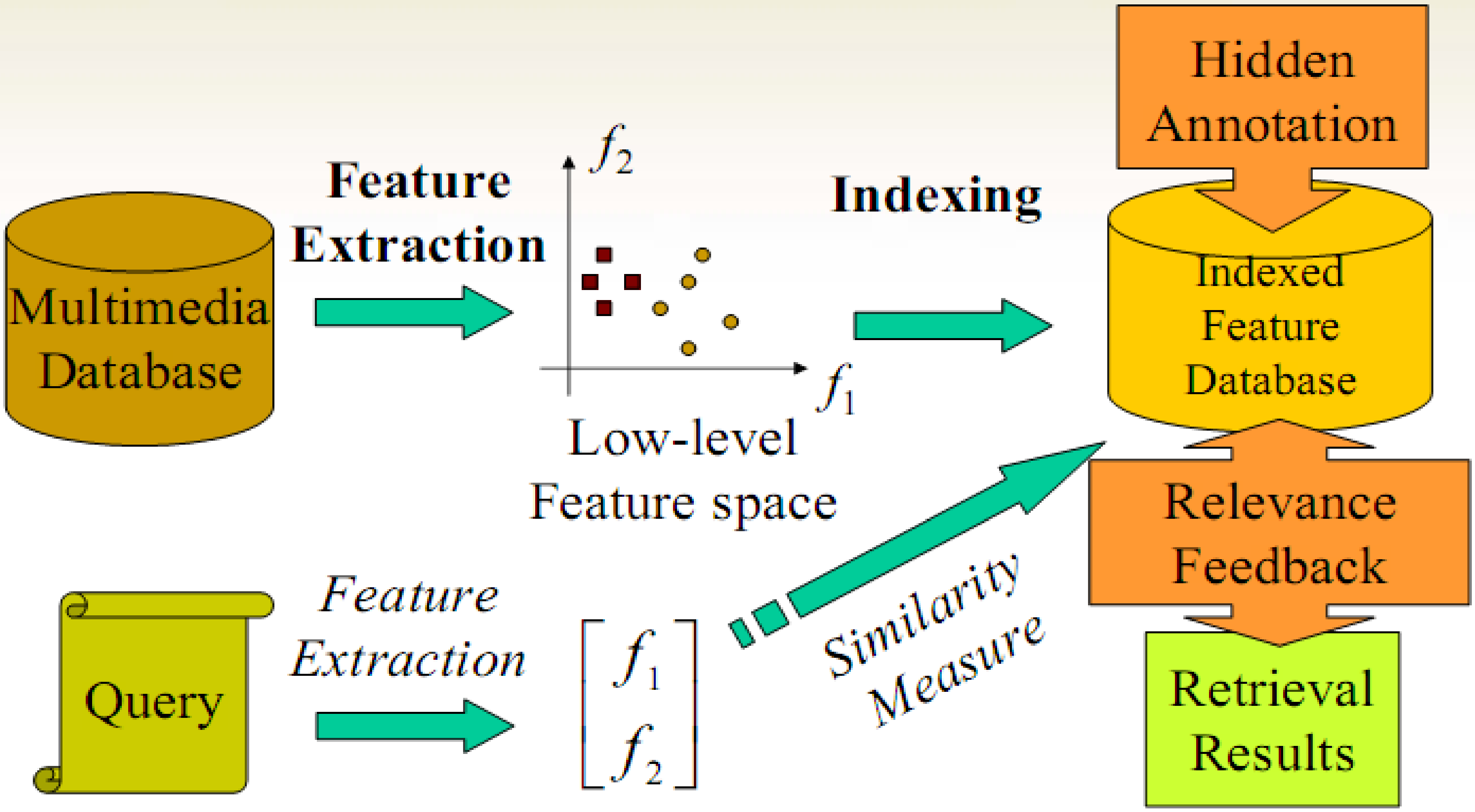
Traditional display cannot characterize image similarity!!

1. Motivation

New Image Search Engines are strongly expected:

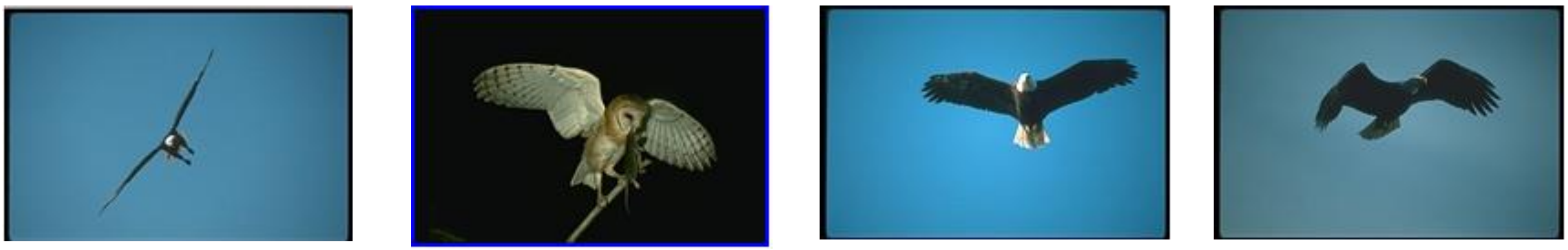
- Filter out the junk images!
- Display the visually-similar images closely!
- Easily for user involvement and capture user’s intention easily!

2. Relevance Feedback

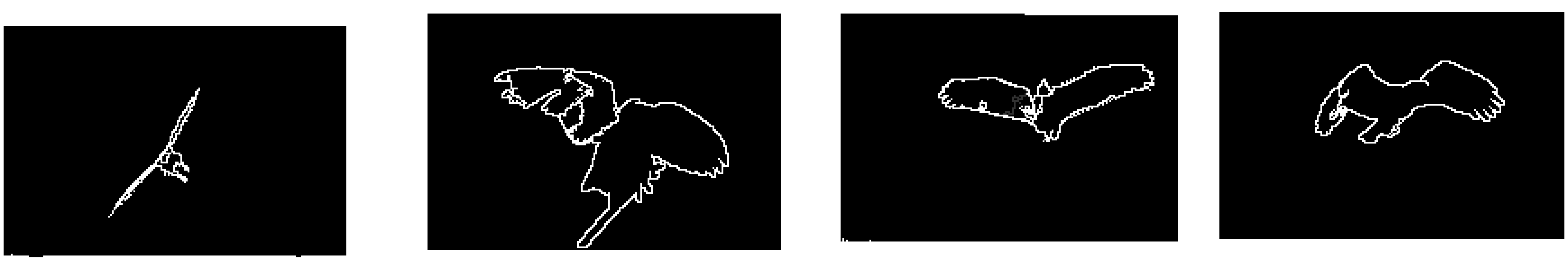


2. Relevance Feedback

a. Taking whole frame as an object



b. Extracting objects or regions from images



2. Relevance Feedback

c. Feature Extraction

Color

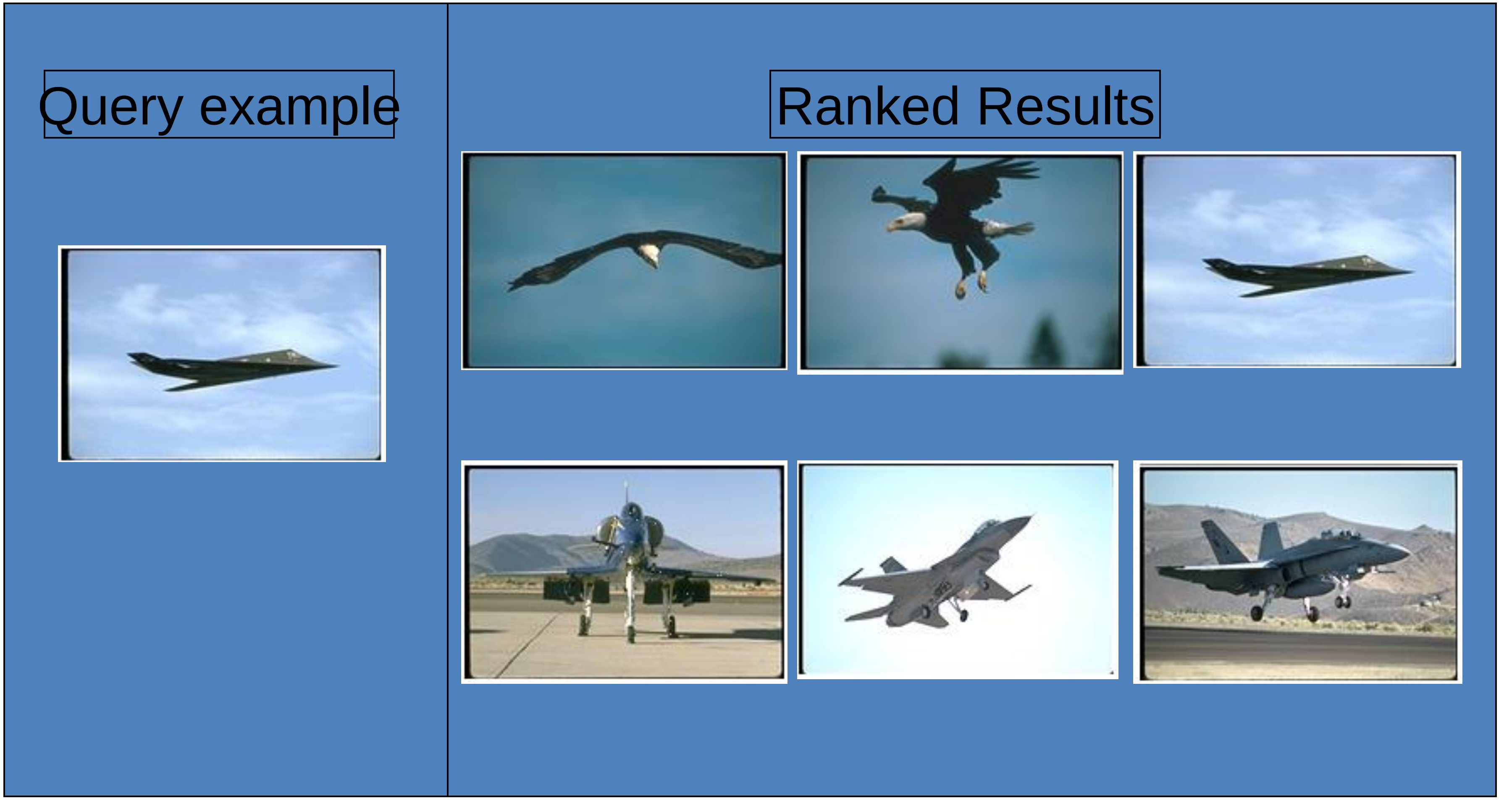
Texture

Shape

2. Relevance Feedback

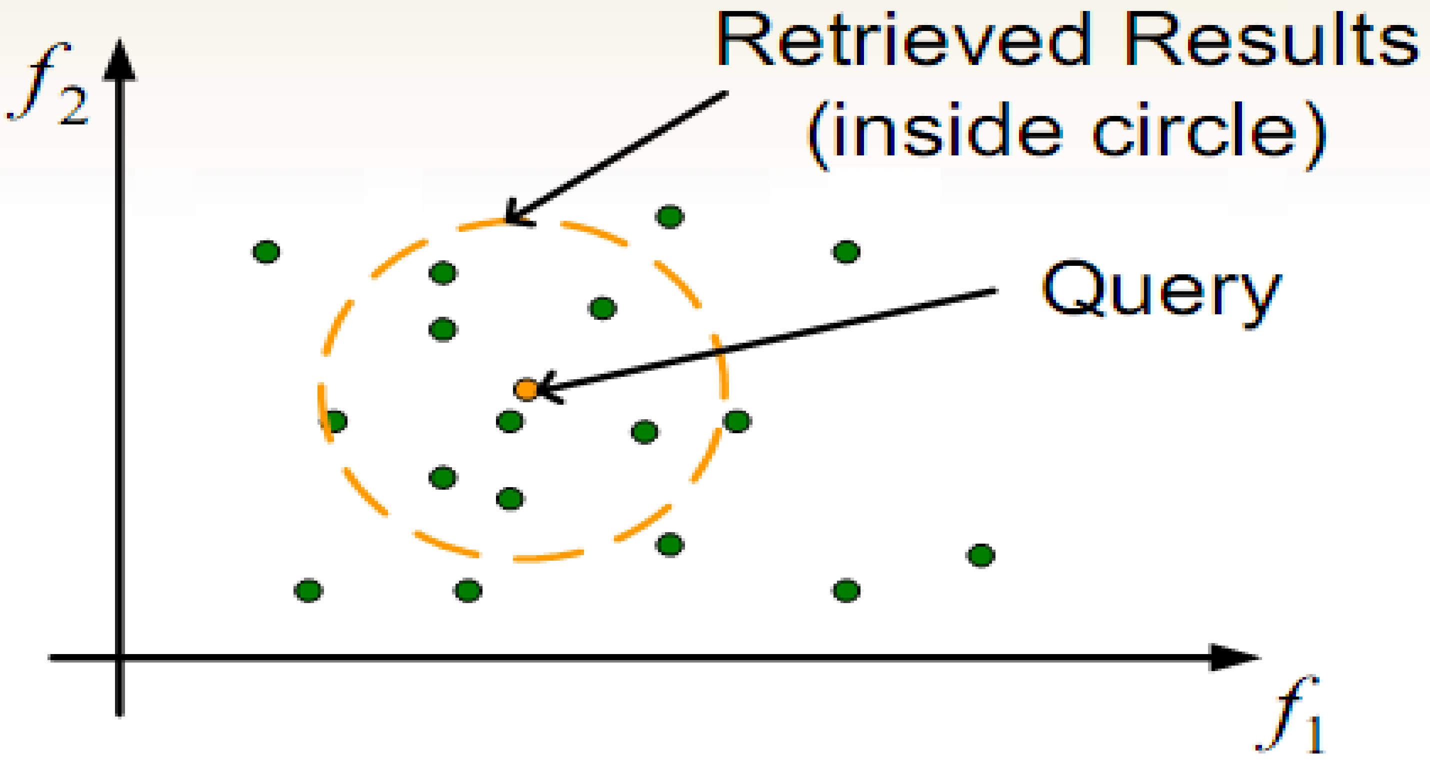
d. Image Classification

2. Relevance Feedback

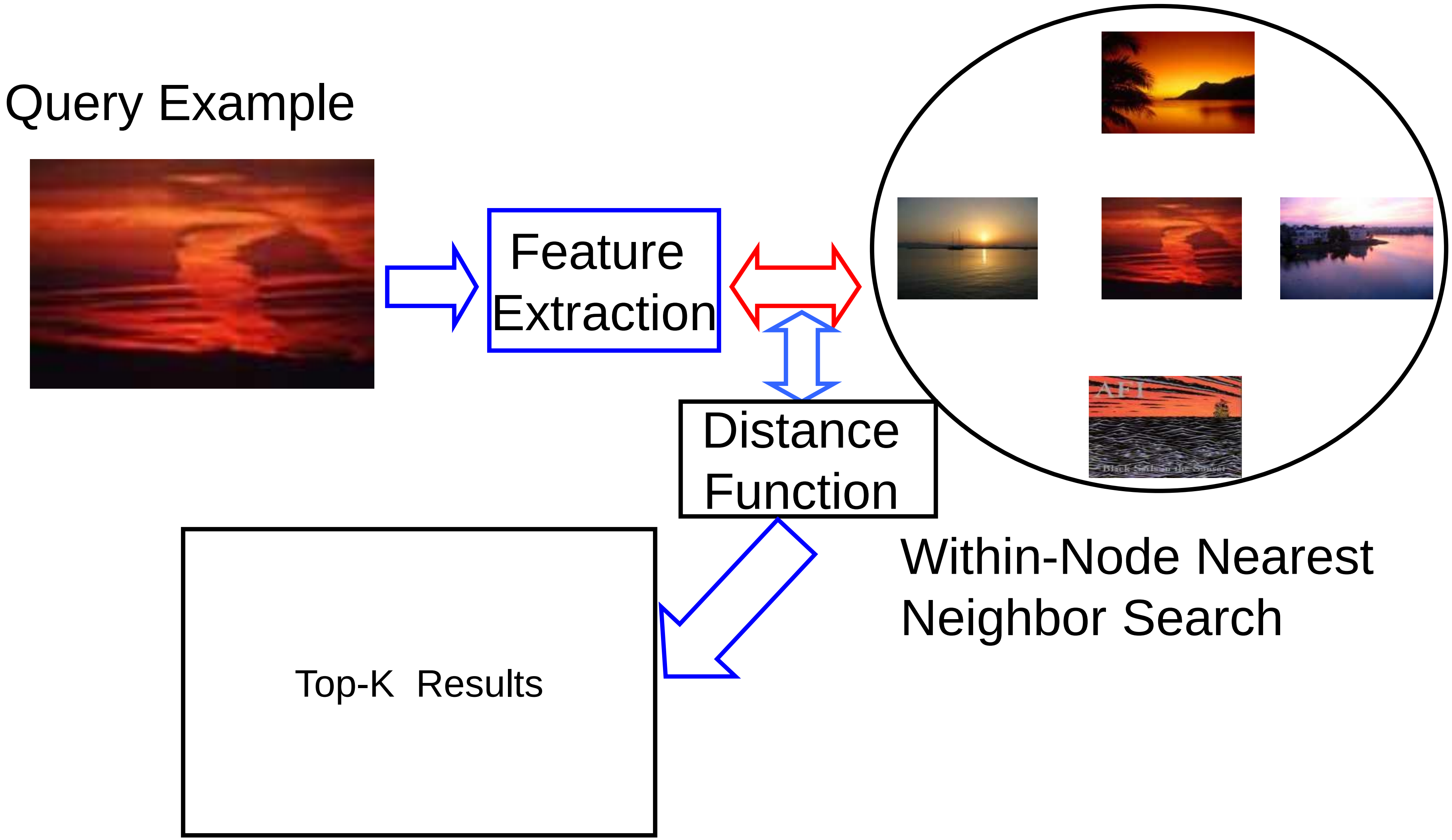


2. Relevance Feedback

An Example



2. Relevance Feedback



2. Relevance Feedback

a. The client send his/her request to the database system;
b. The database system sends him/her some ranked answers;
c. The client can exchange his/her judgment with the system.

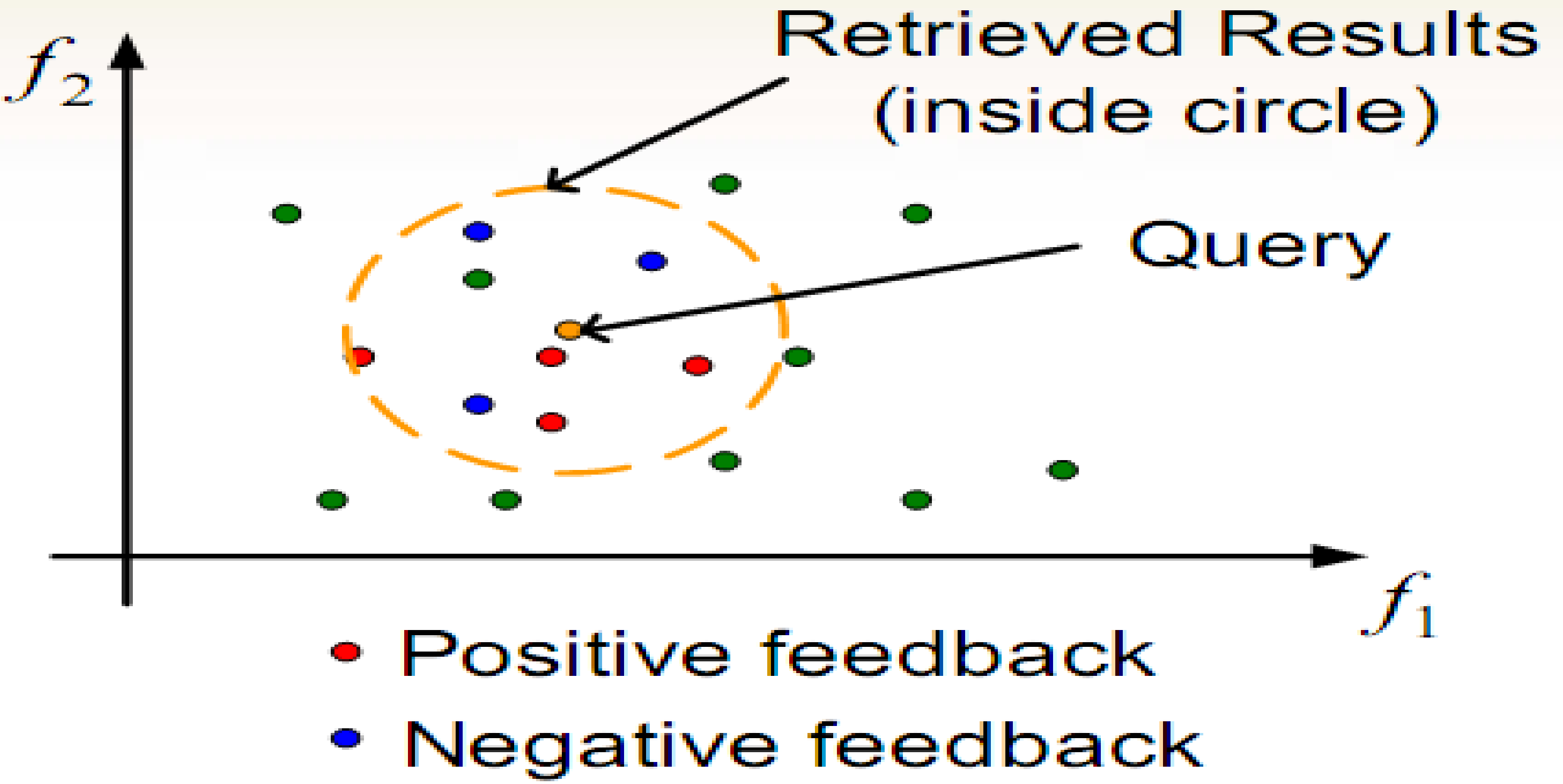


relevant? no no



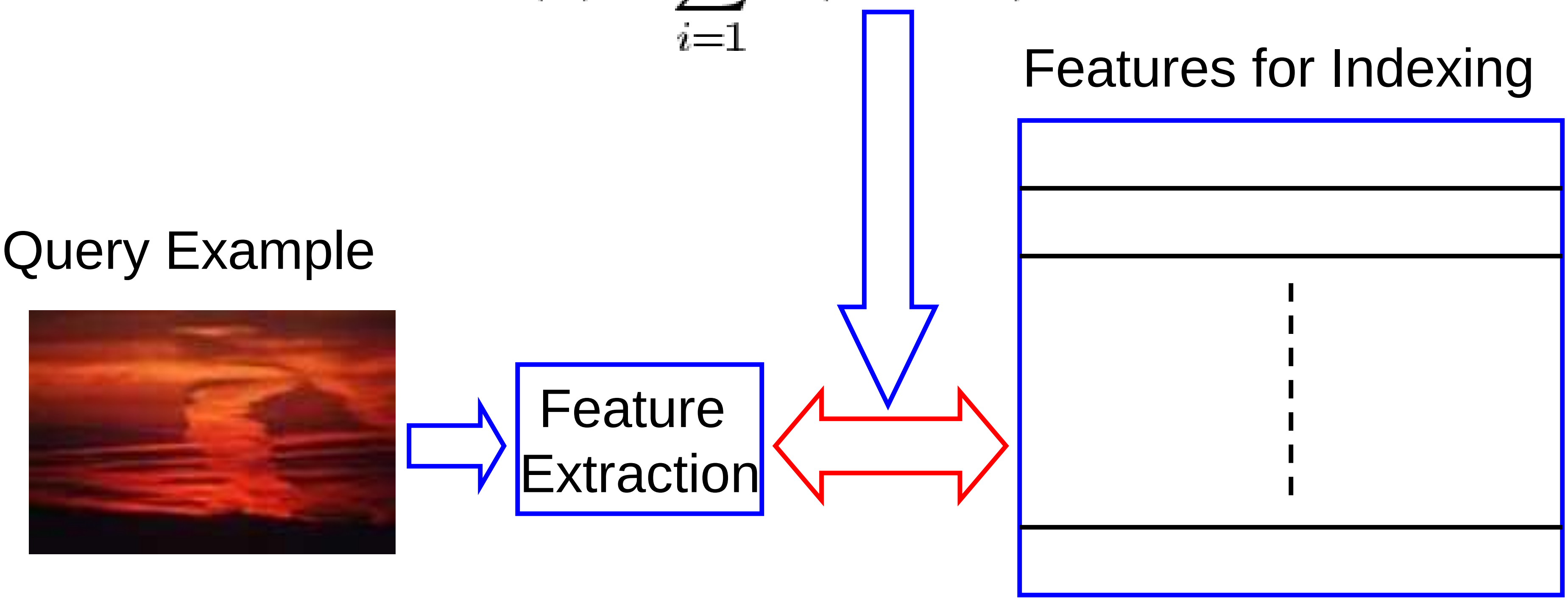
2. Relevance Feedback

User Feedback

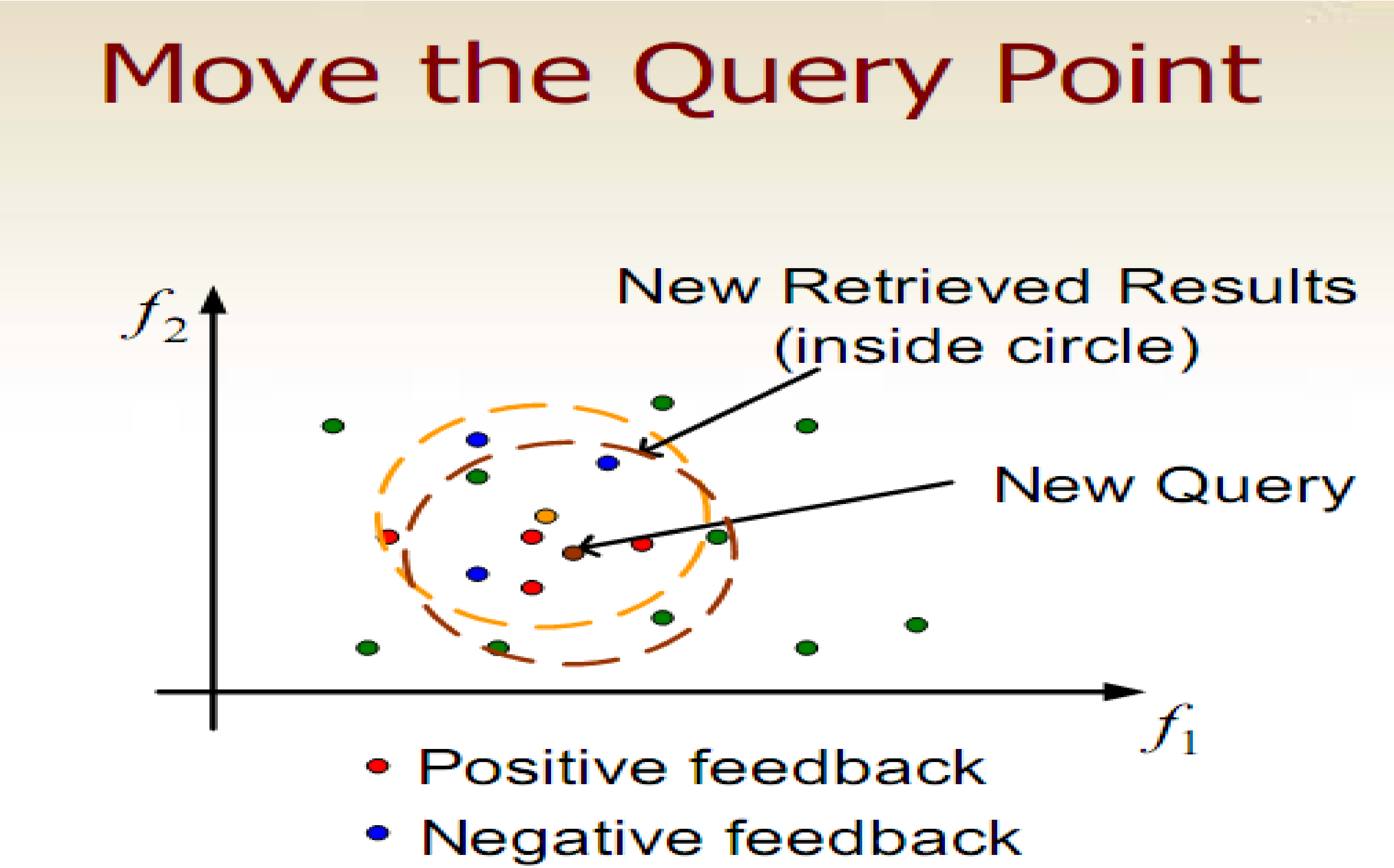


- Distance Weighting Approach:

$$\phi = \log(L) = \sum_{i=1}^{I} \alpha_i (A_i - R_i)$$

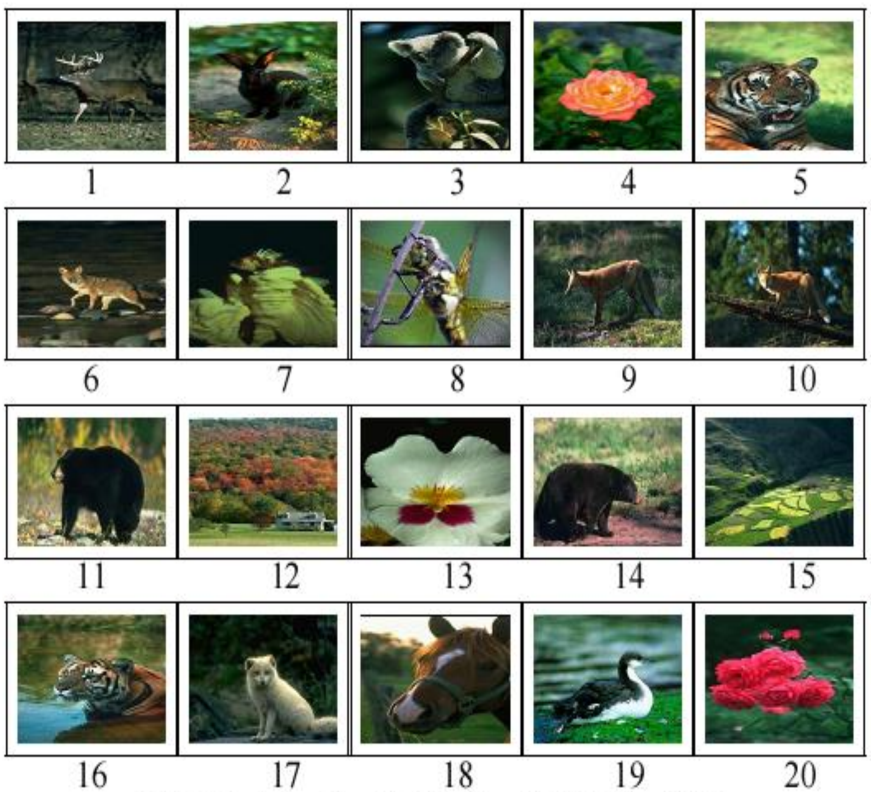


2. Relevance Feedback

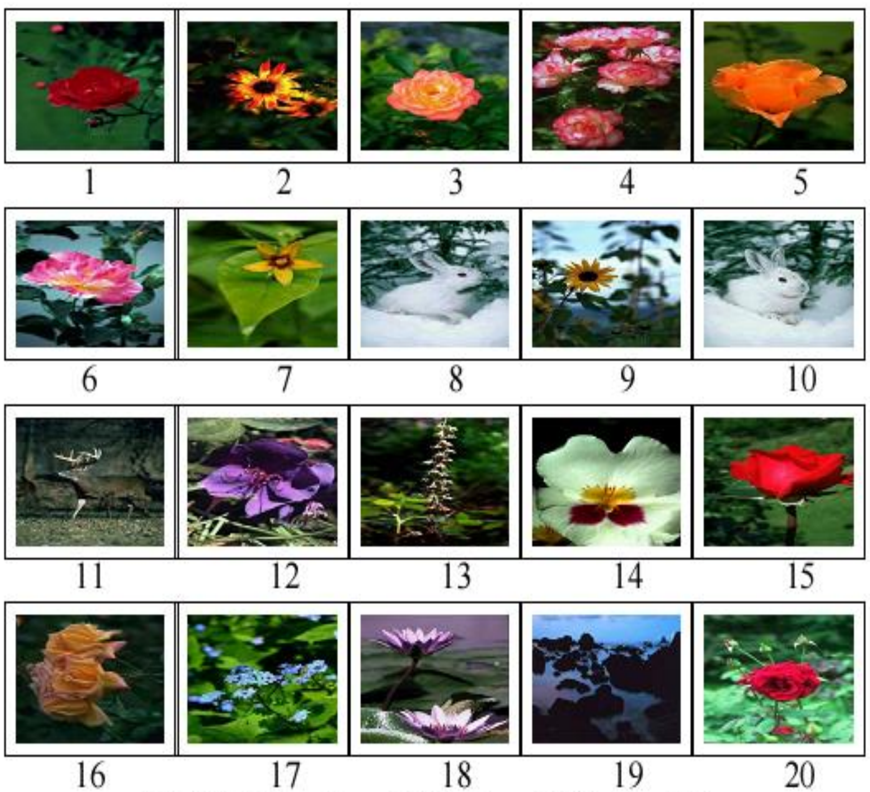


- Effectiveness of Feature Weighting

(a) Query image: Flower and vegetation



(b) Weights: S = 0.33, C = 0.33, T = 0.33.



(c) Weights: S = 0.05, C = 0.05, T = 0.9.

2. Relevance Feedback

- Query Updating

$$\overline{Q}^{(k+1)} = \overline{Q}^{(k)} + \delta \sum_{i=1}^{N_{rel}} \overline{X}_i^{(k)} - \varepsilon \sum_{j=1}^{M_{notrel}} \overline{Y}_j^{(k)}$$

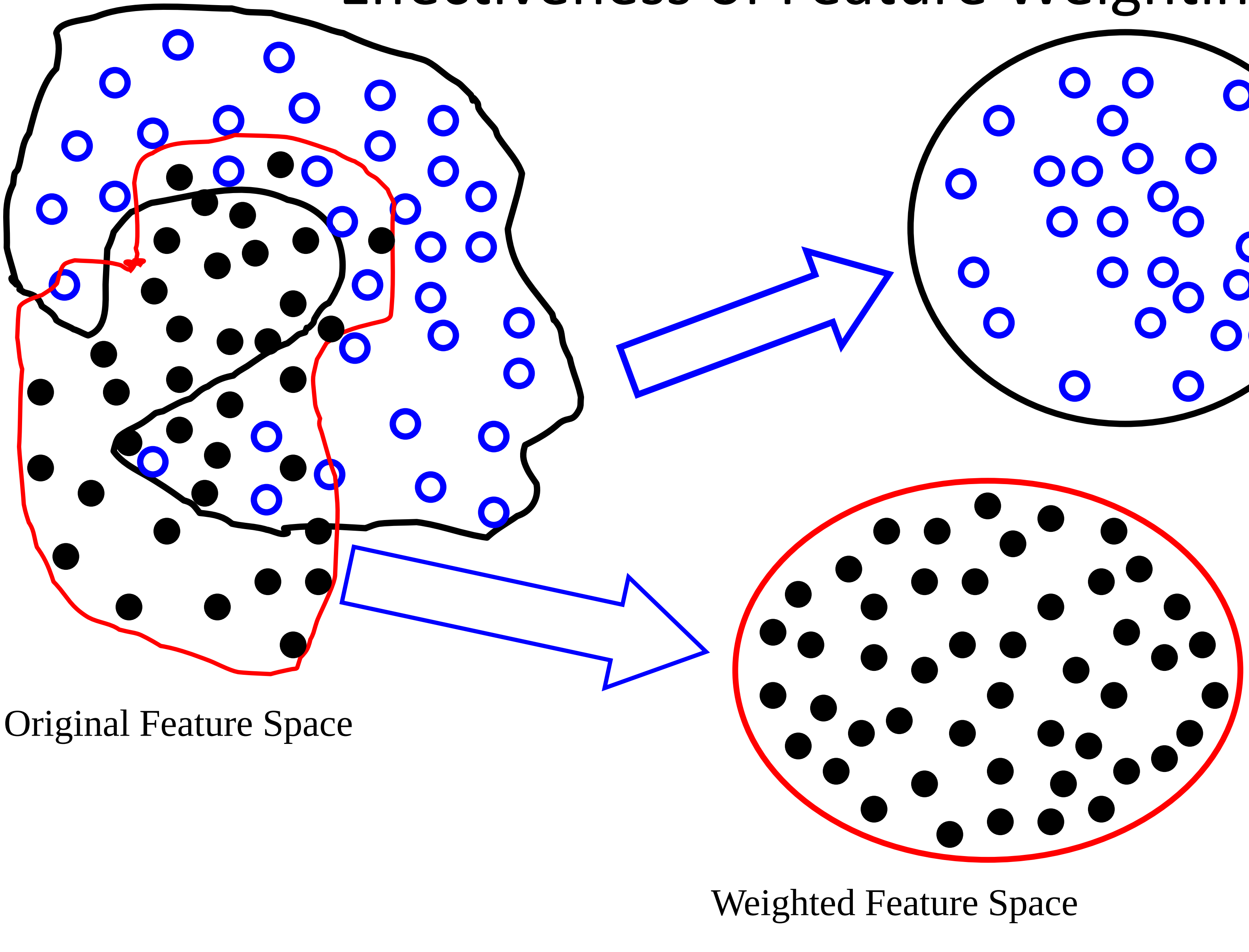
New Query Vector

Previous Query Vector

Vectors for Positive Images

Vectors for Negatives

- Effectiveness of Feature Weighting:



- Problem for Feature Weighting Approach

a. Cost-Sensitive: It is very expensive to update the feature weights on real time!

b. Semantic Gap: The distance functions may not be able to characterize the underlying image similarity effectively!

c. Visualization: The underlying image display tools may separate similar images in different places, it is hard for users to evaluate the visual similarity (relevance) between the images!

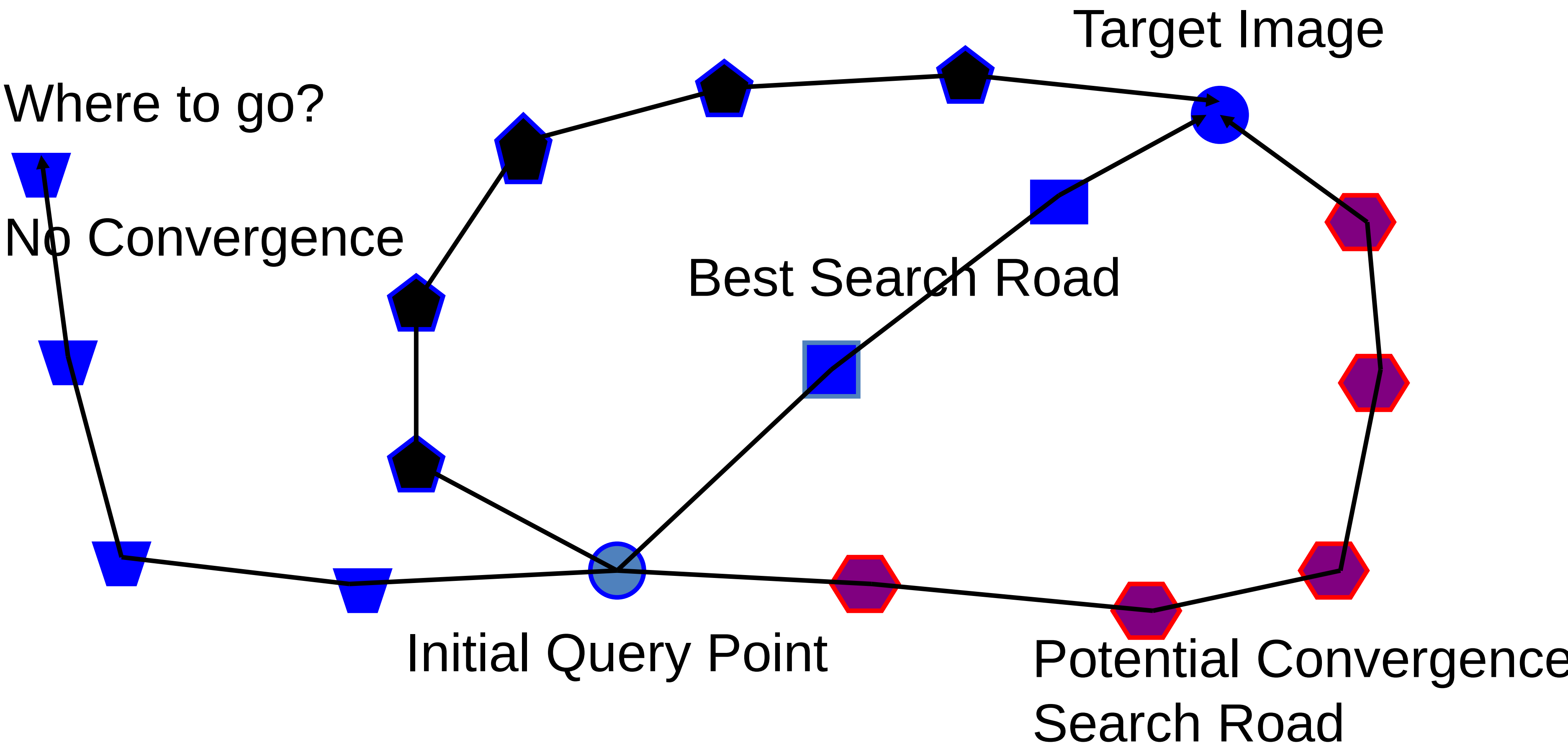
2. Relevance Feedback

- Two More Issues for Feature Weighting

a. Informative Sample Generation: what we should return to users, so that they can make good decision on relevance *vs.* irrelevance?

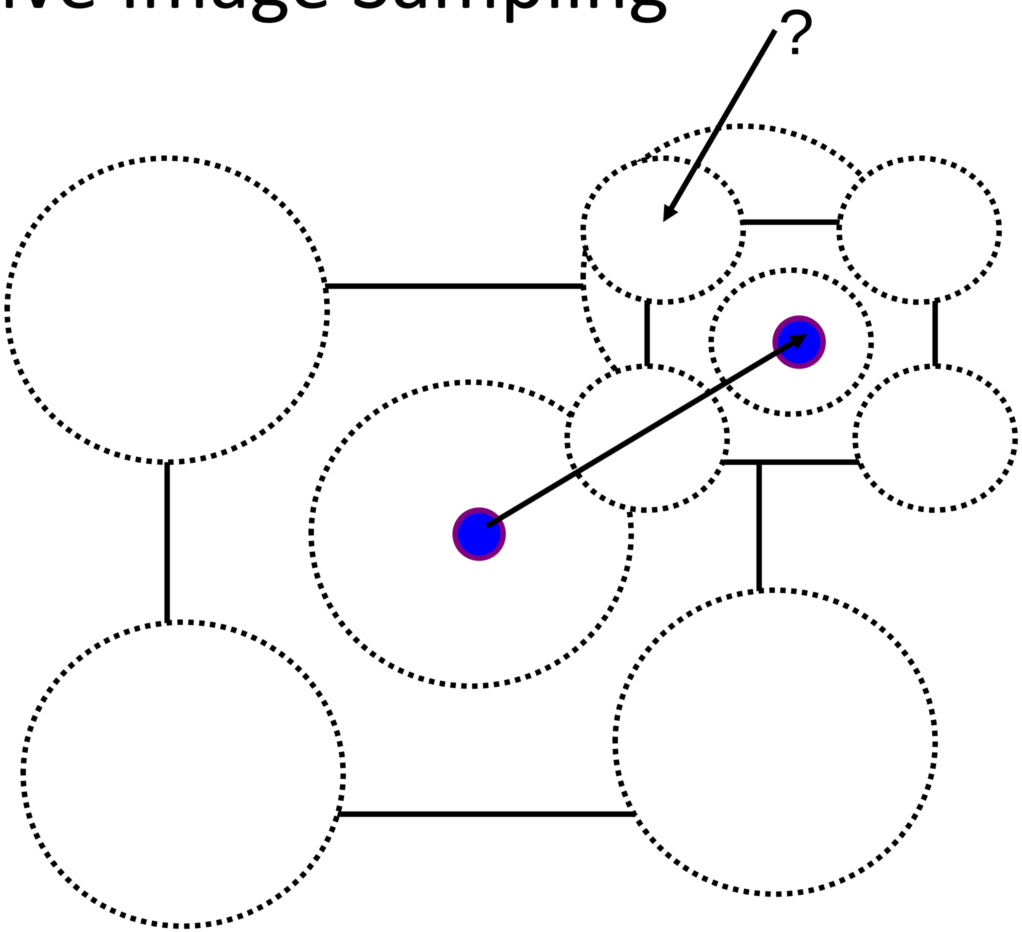
b. Query Movement Control: Through weighting the features, it is able for us to control the importance between the features for image similarity characterization. However, for image retrieval application, we also need to control the query point to move to target in the best way!

- Query Point Movement Control



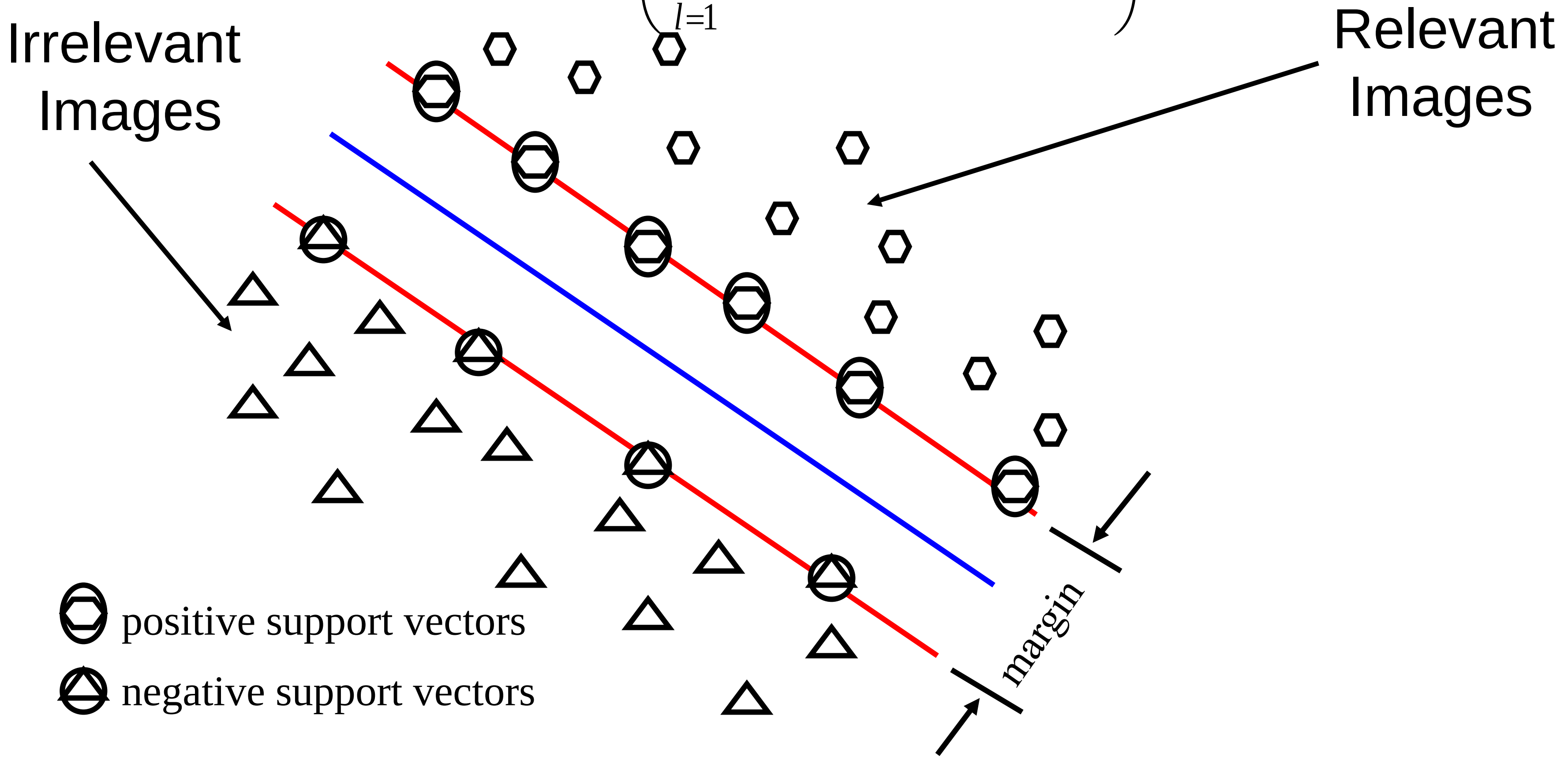
2. Relevance Feedback

- Informative Image Sampling



- **Classification-Based Approach**

$$f(X) = \mathrm{sign}\left(\sum_{l=1}^{M} \beta_l Y_l \hat{\mathrm{K}}(X_l, X) + b \right)$$



2. Relevance Feedback

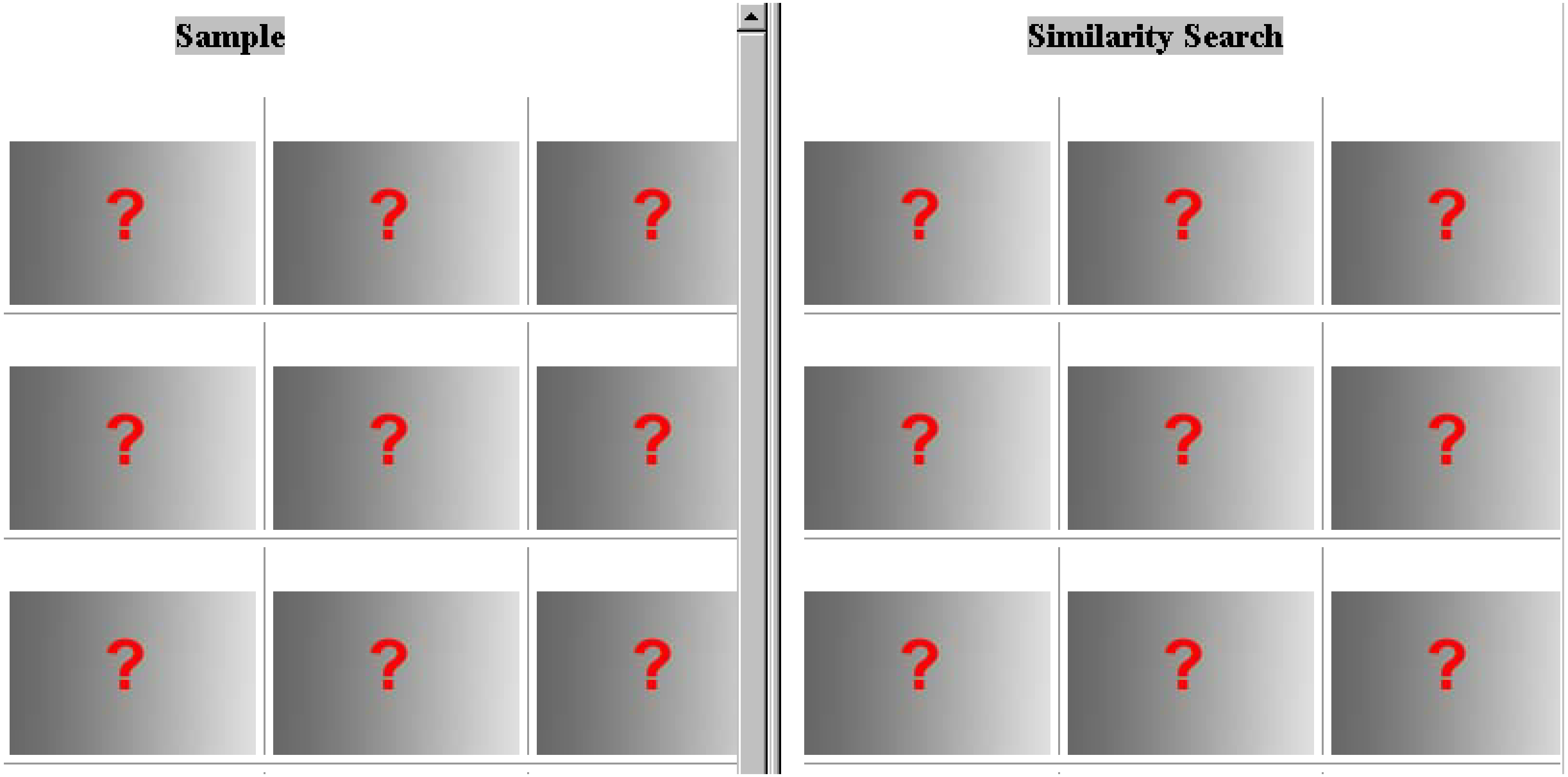
- Increment SVM Updating

$$f_i(x_i) = \begin{cases} x_i & x_i < u_i \\ a \cdot (x_i - u_i) + u_i & x_i \in [u_i, v_i] \\ x_i + (a-1) \cdot (v_i - u_i) & x_i > v_i \end{cases}$$

Piece-wise updating of kernel function
by using users selected images!

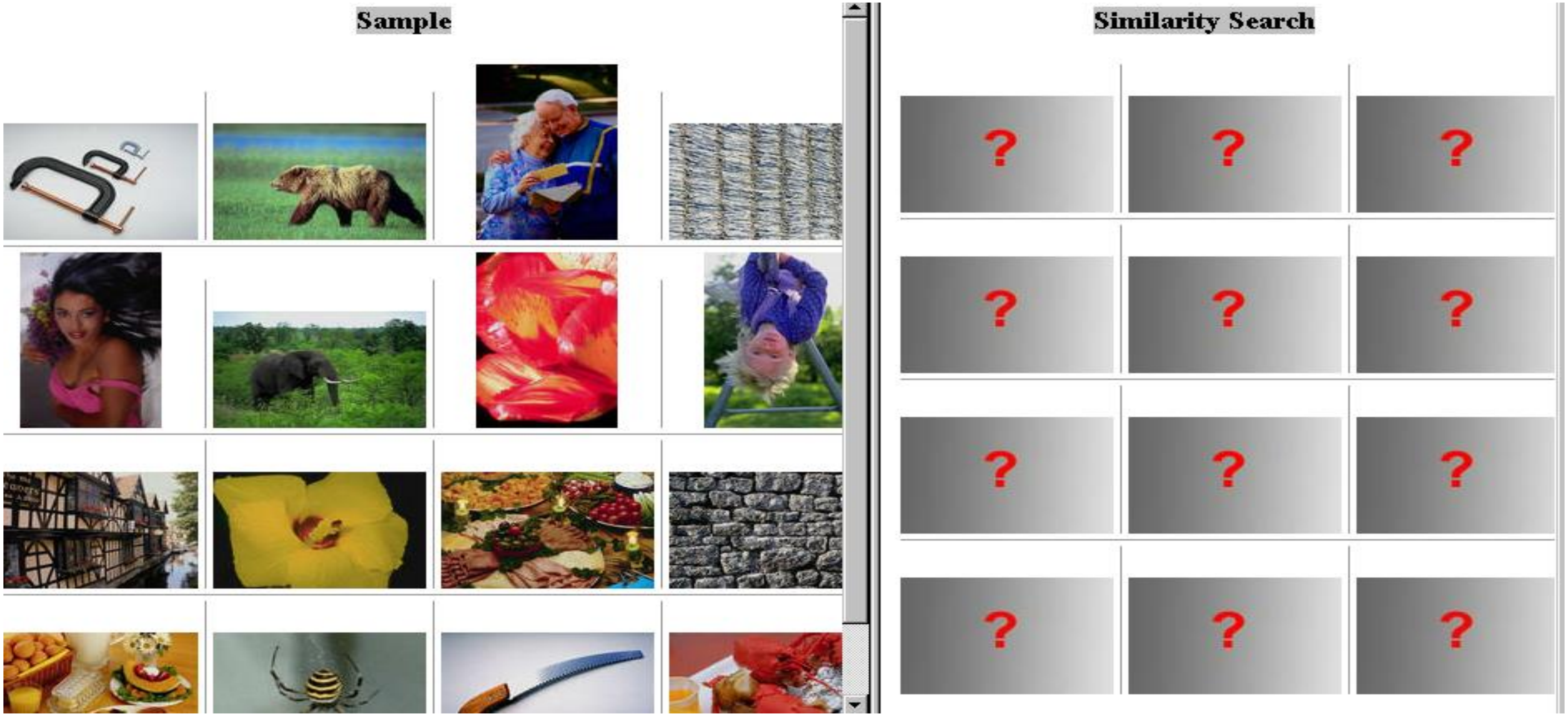
2. Relevance Feedback

a. How to initialize the query?



2. Relevance Feedback

b. Send the query to system

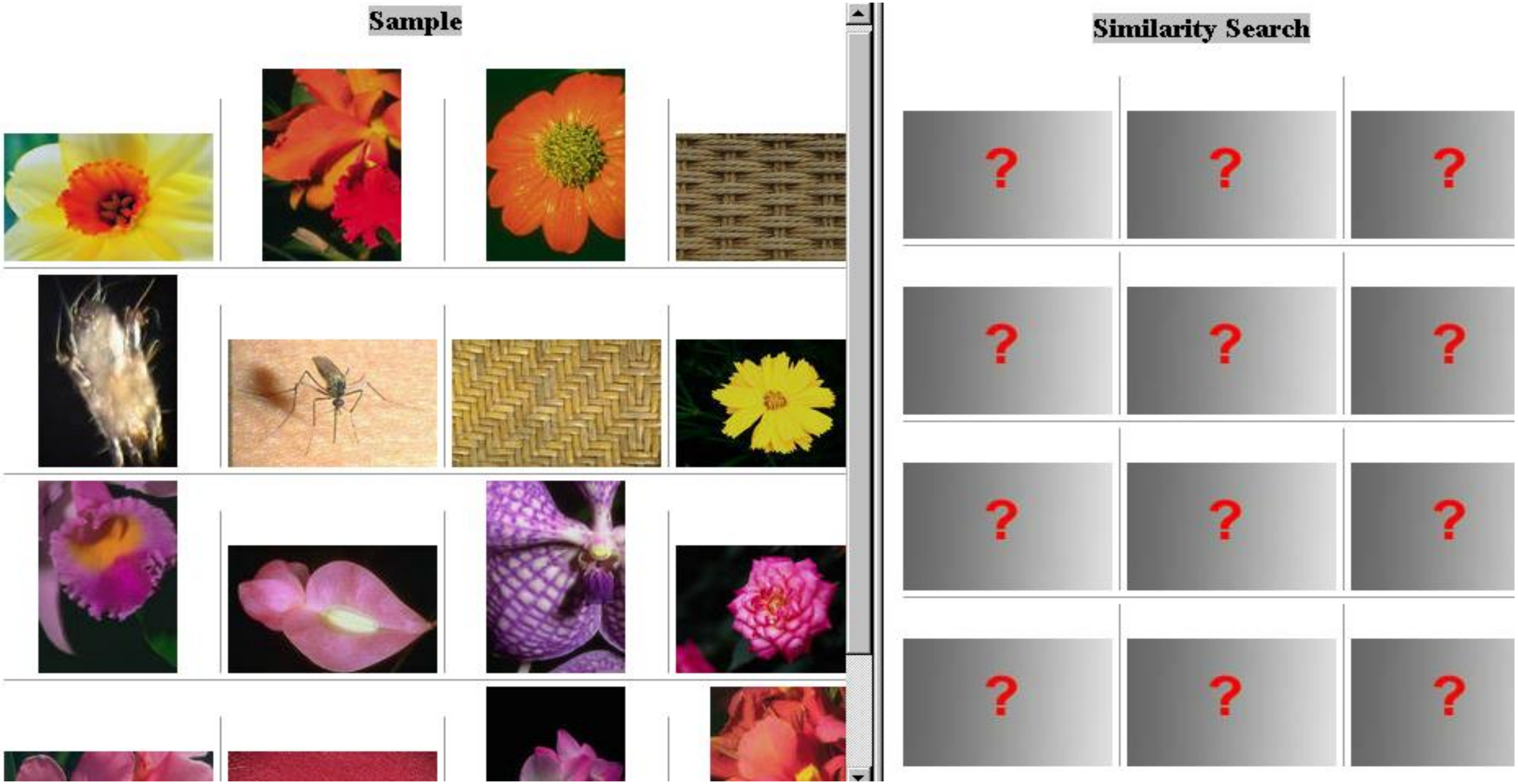


2. Relevance Feedback

c. Client mark the relevant examples

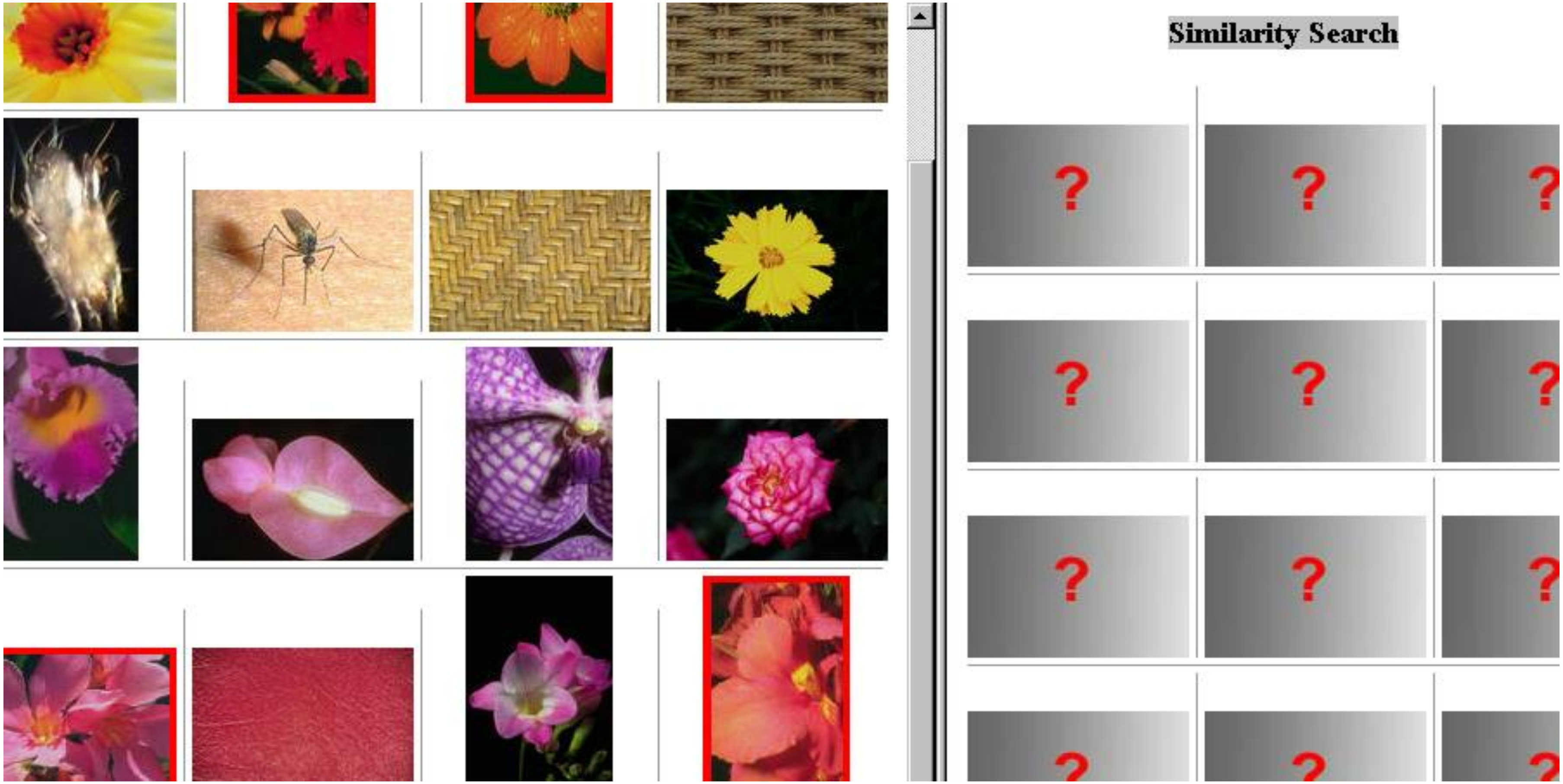
2. Relevance Feedback

d. System Evaluation according to client feedback



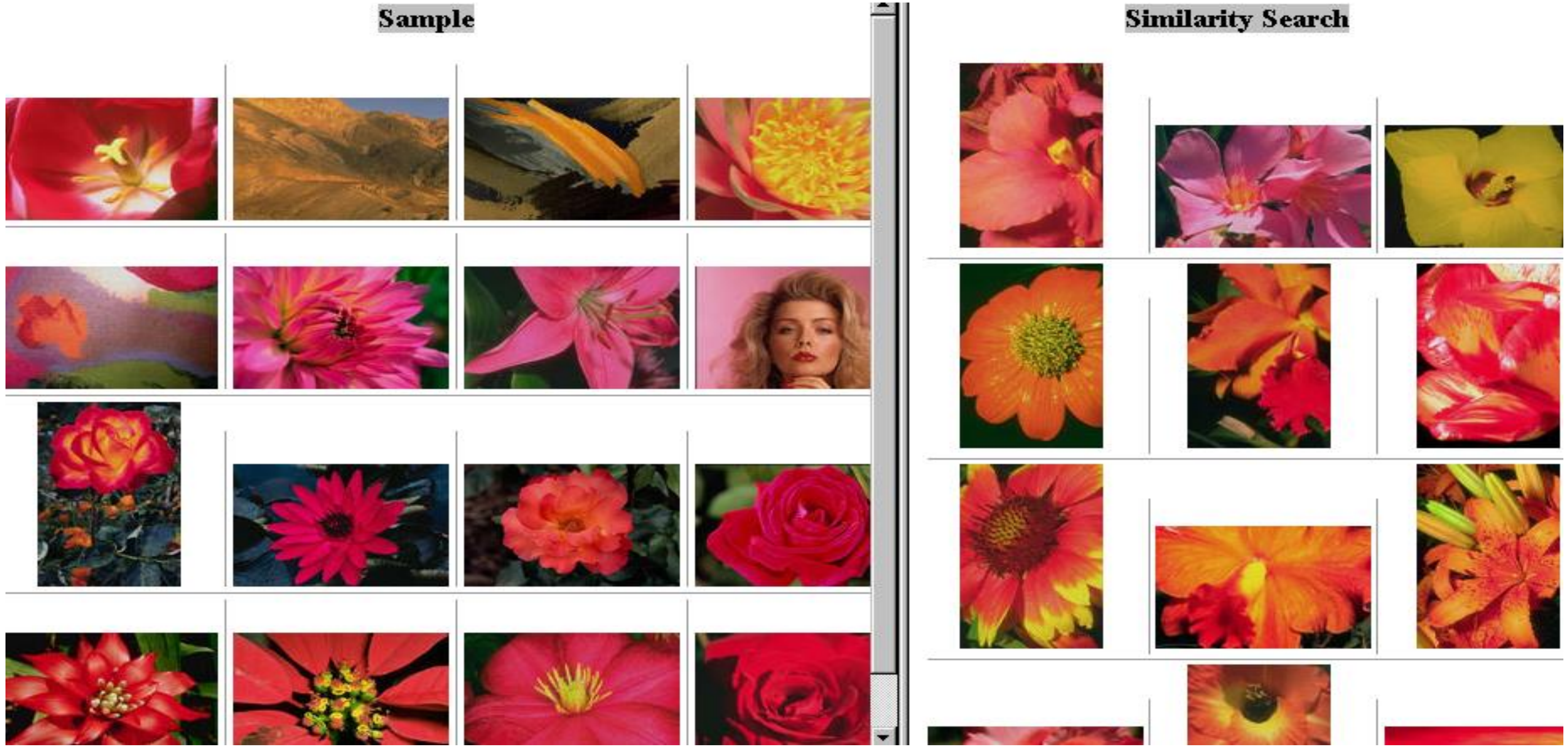
2. Relevance Feedback

e. Second client feedback



2. Relevance Feedback

f. Second System Evaluation



2. Relevance Feedback

- Problems for Classification-Based Approach

a. Convergence: It is very important to guarantee the algorithm for kernel updating is converged!

b. Cost Reduction: It is very important to reduce the cost for kernel updating!

3. Our Proposed Solutions

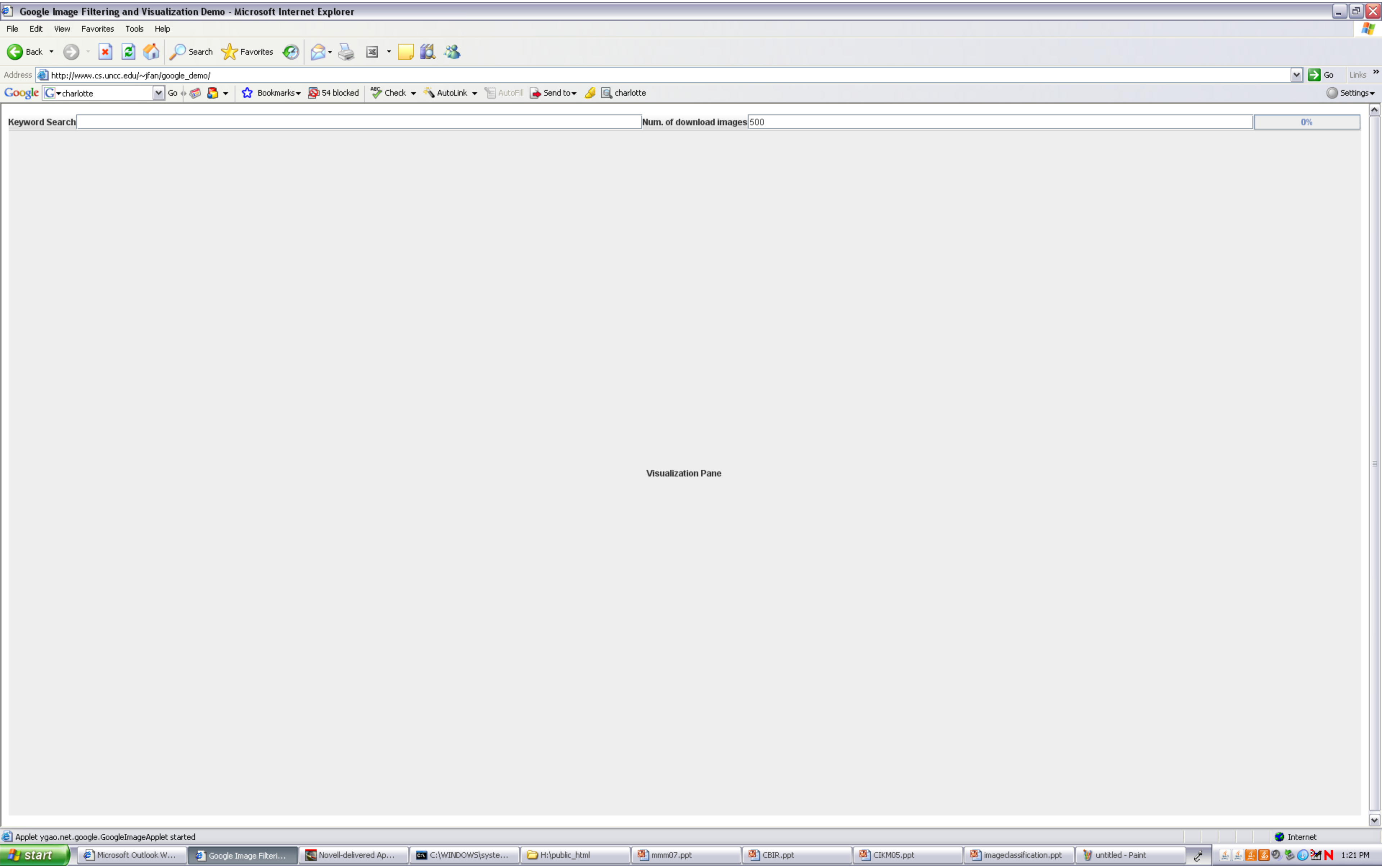
- Kernel-Based Clustering of Google Search Results
- Similarity-Based Image Projection and Visualization
- Intention capturing and Kernel Selection for Junk Image Filtering

Relevance is user-dependent!

3. Our Proposed Solutions

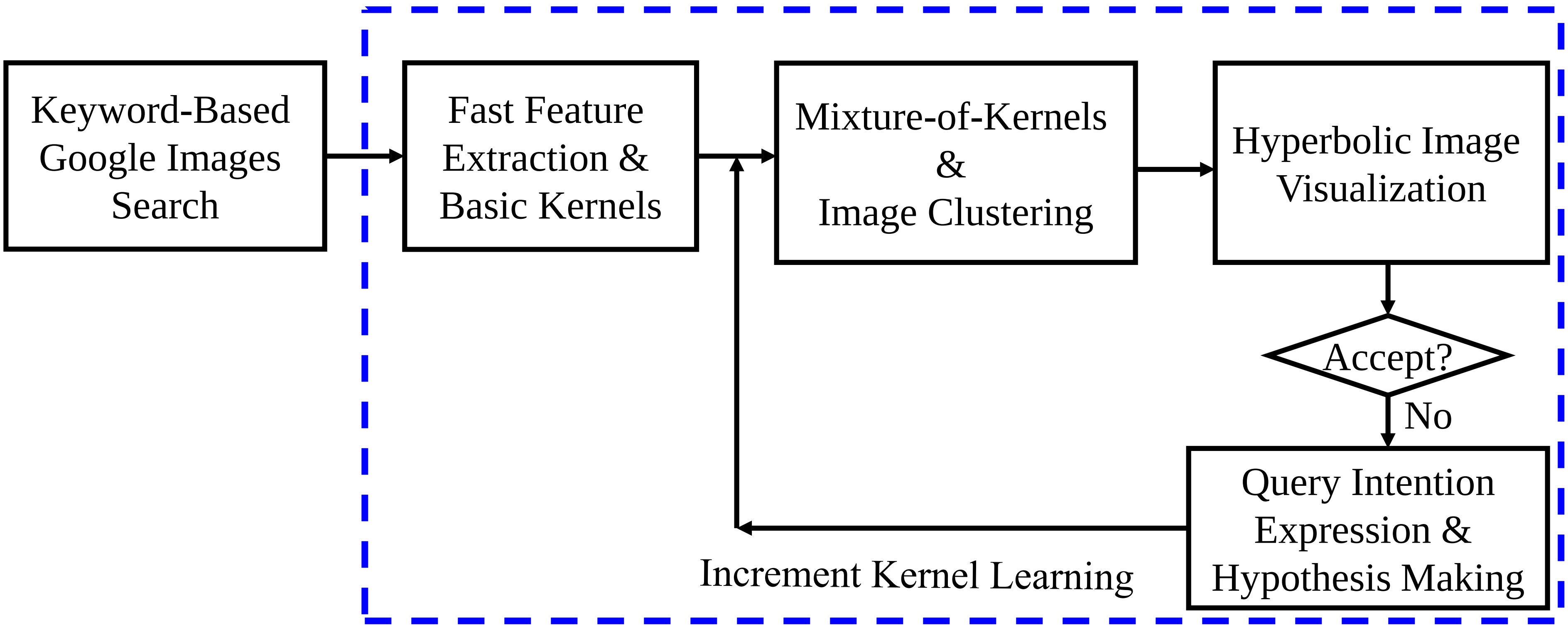
- **Requirements for such new search engine**:
  - Fast algorithm for feature extraction;
  - Multiple kernels for diverse image similarity characterization;
  - Implicit query intention capturing and real-time kernel updating

Keyword-Based Image Search



Users are allowed to submit keyword-based queries!

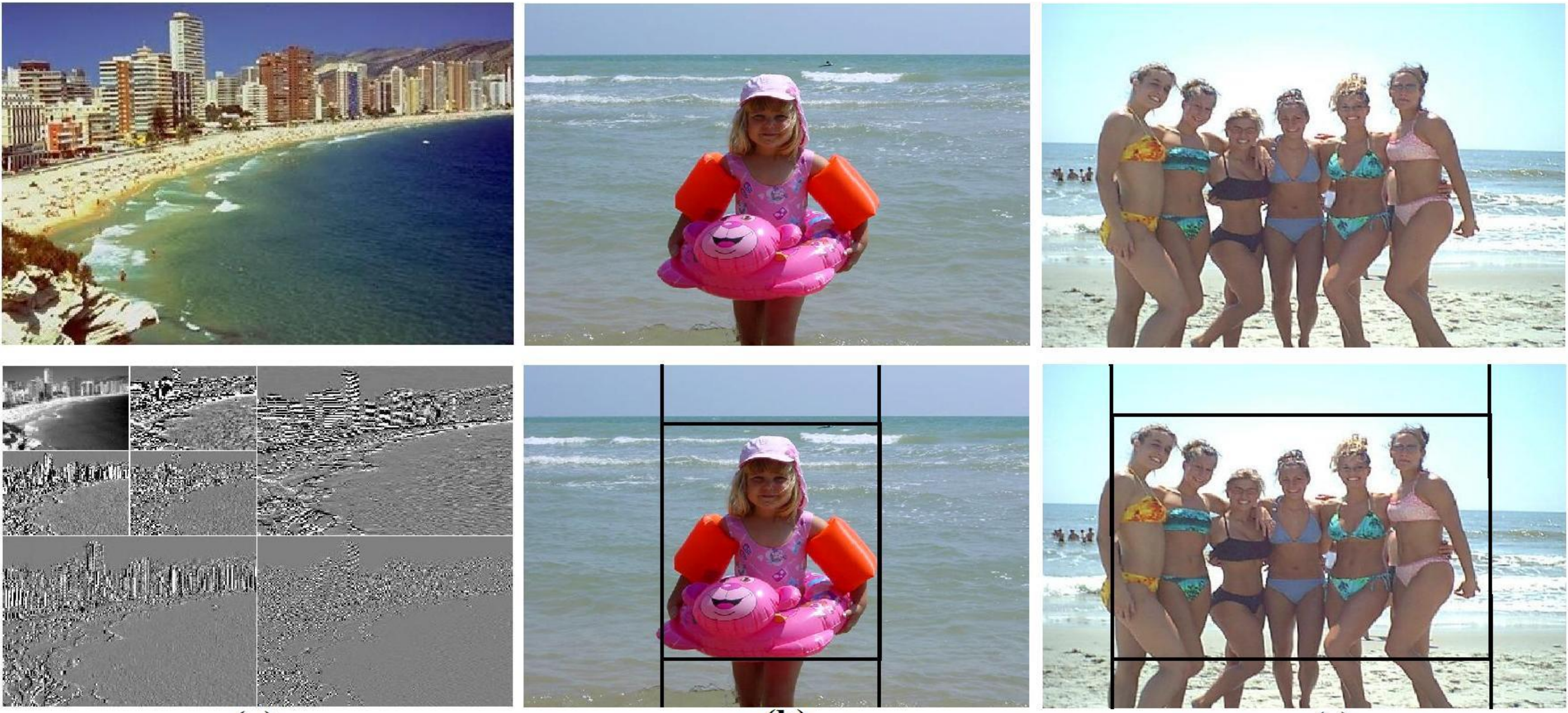
3. Our Proposed Solutions



Through incremental learning, we can consider multiple competing hypotheses for the same task!

3. Our Proposed Solutions

Fast Feature Extraction



(a) (b) (c)

3. Our Proposed Solutions

Image Representation & Similarity

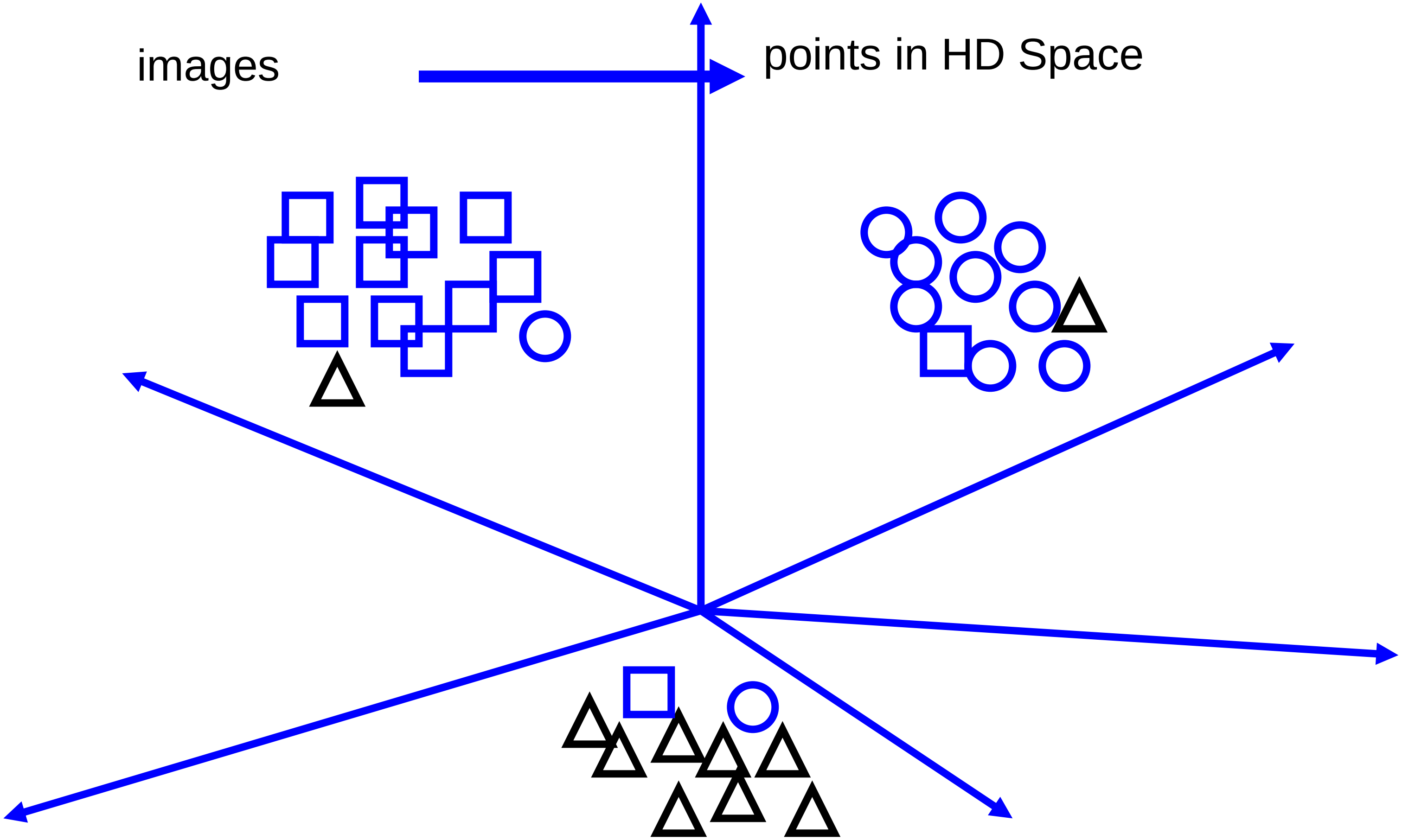
a. Color histogram for whole image

b. 10 color histograms for different patterns

c. Wavelet transformation

Time-Constrainted Image Analysis!

3. Our Proposed Solutions



They are invisible for human eye!

3. Our Proposed Solutions

Basic kernels for image similarity characterization:

- Color Histogram Kernel

$$\kappa(u,v) = e^{-\chi^2(u,v)/\sigma_c} \qquad \chi^2(u,v) = \frac{1}{2}\sum \frac{(u_i - v_i)^2}{u_i + v_i}$$

- Wavelet Filter Bank Kernel

$$\kappa_w(u,v) = \prod_{i=1}^{m} e^{-\chi^2(u_i,v_i)/\sigma_w}$$

- Sub-Image Color Histogram Kernel

$$\kappa_I(u,v) = e^{-D(u.v)/\sigma_I}$$

5. Image Similarity Characterization

Mixture-of-kernels for diverse similarity characterization:

$$K(u,v) = \sum_{i=1}^{\tau} \alpha_i \kappa_i (u,v) \qquad \sum_{i=1}^{\tau} \alpha_i = 1$$

(a) It could be expensive for learning a good combination!

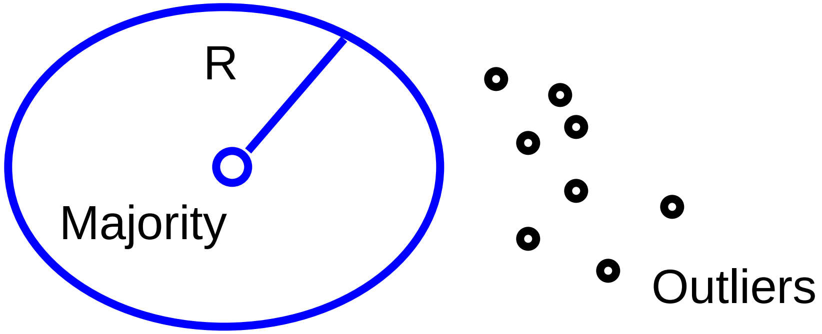
(b) The similarity between the images depends on the given kernel function!

3. Our Proposed Solutions

Hypothesis Making & Initial Analysis

$$\forall_{j=1}^{N}: \quad \|\phi(x_j) - \mu^{\phi}\|^2 \leq R^2$$

$$\mu^{\varphi} = \sum_{j=1}^{N} \phi(x_j)$$



$$min\left\{R^2 + \frac{C}{N}\sum_{j=1}^{N}\xi_j\right\}$$ subject to: $$\forall_{j=1}^{N}: \quad \|\phi(x_j) - \mu^{\phi}\|^2 \leq R^2 + \xi_j, \qquad \xi_j \geq 0$$

Decision function:

$$f(x) = R^2 - \sum_{i,j}^{N} \alpha_i \alpha_j \kappa(x_i, x_j) + 2\sum_{j}^{N} \alpha_j \kappa(x_j, x) - \kappa(x, x)$$

3. Our Proposed Solutions

Similarity-Preserving Image Projection

$$y_i = A^T \phi(x_i) \qquad \Xi_{ij} = \phi(x_i)^T \phi(x_j) = \kappa(x_i, x_j) = \sum_{l=1}^{3} \beta_l K_l(x_i, x_j)$$

$$A_{optimal} = \underset{A}{argmin} \left\{ \sum_{i,j}^{N} (y_i - y_j)^2 \Xi_{ij} \right\} = \underset{A}{argmin} \left\{ \sum_{i,j}^{N} (A^T \phi(x_i) - A^T \phi(x_j))^2 \Xi_{ij} \right\}$$

$$= \underset{A}{argmin} \; A^T \phi(X) \Delta \phi^T(X) A$$

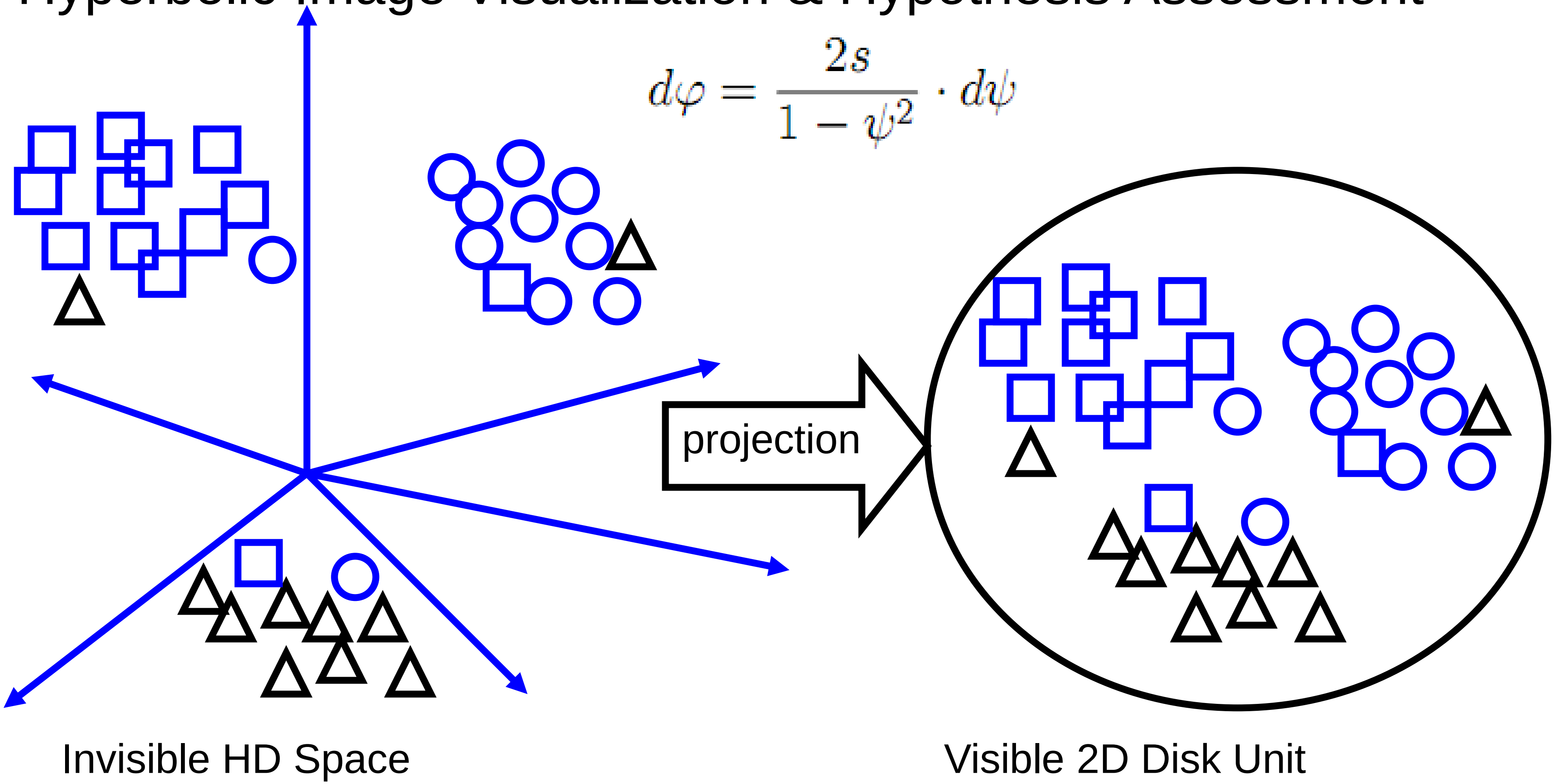
$$\Delta = D - \Xi, \qquad D_{ii} = \sum_{j}^{N} \Xi_{ij}$$

Transform large amount of images (represented by high-dimensional visual features) into their similarity contexts for enabling better visualization!

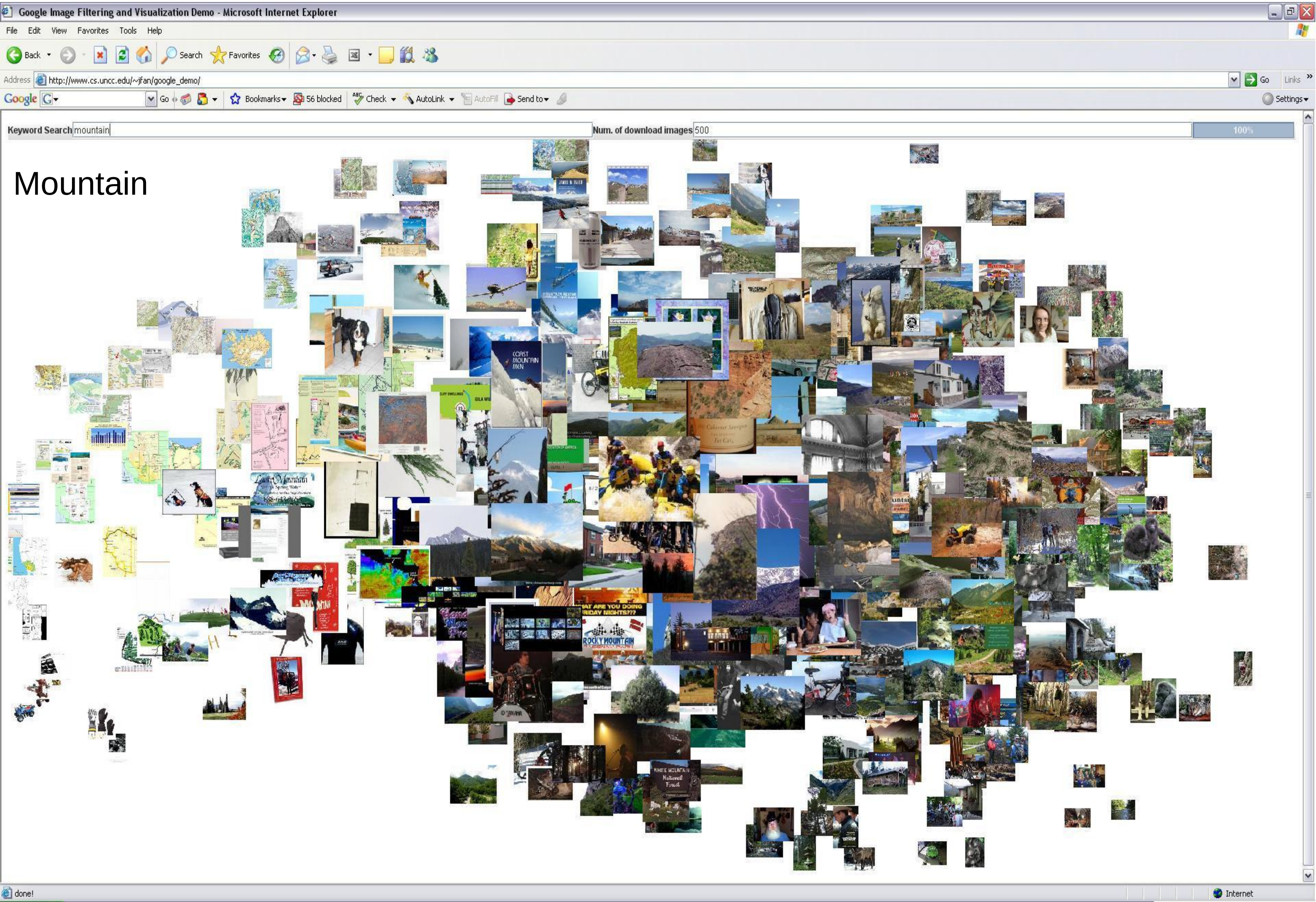
3. Our Proposed Solutions

Hyperbolic Image Visualization & Hypothesis Assessment

$$d\varphi = \frac{2s}{1-\psi^2} \cdot d\psi$$



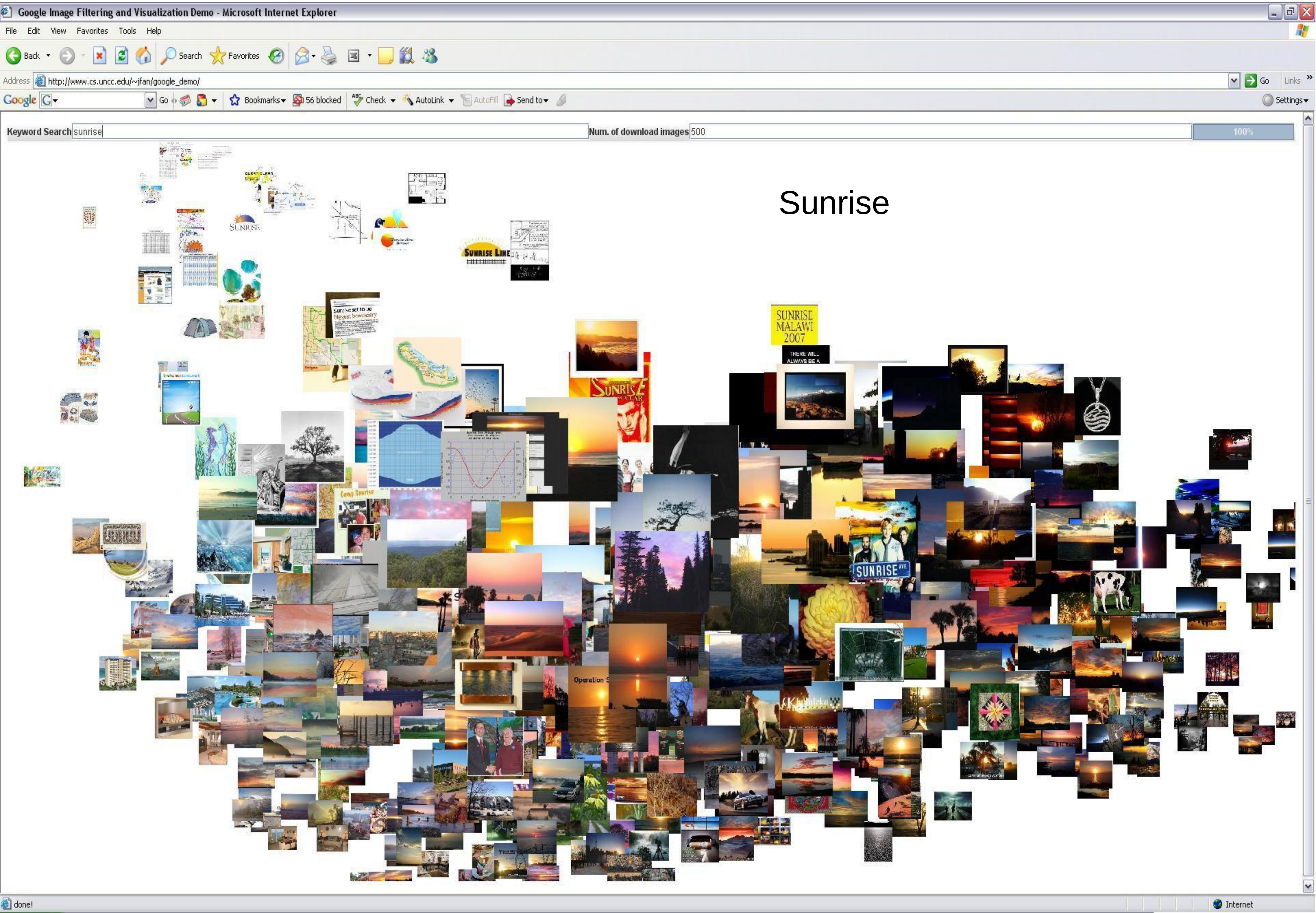
3. Our Proposed Solutions



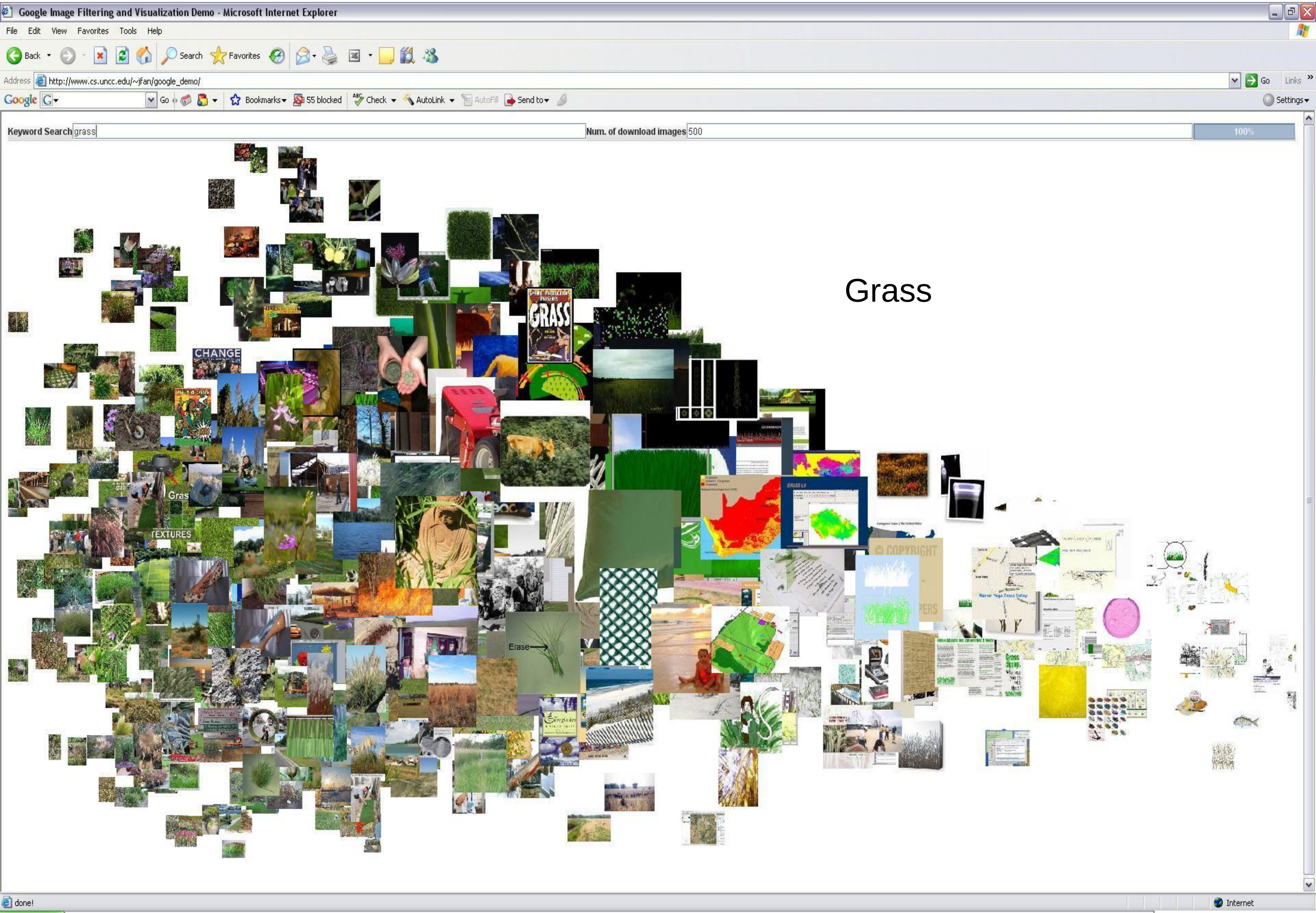
3. Our Proposed Solutions

Ocean

3. Our Proposed Solutions



3. Our Proposed Solutions



3. Our Proposed Solutions

User-System Interaction for Making New Hypothesis

Hypothesis-Driven Image Re-Clustering

3. Our Proposed Solutions

Hypothesis-Driven Data Analysis:

a. Updating decision function: margin between relevant images and irrelevant images!

b. Updating the combination of feature subsets!

c. Updating image projection optimization criteria to obtain more accurate projection!

d. Updating image representation!

3. Our Proposed Solutions

Incremental Learning: Update decision function

$$\min\left\{\frac{1}{2}\left\|W - W_0\right\|^2 + \alpha\sum_{l=1}^{m}[1 - Y_l(W^T \bullet X_l + b)]\right\}$$

- Dual Problem

$$\min\left\{\frac{1}{2}\sum_{l=1}^{m}\sum_{h=1}^{m}\alpha_l\alpha_h Y_l Y_h X_l^T X_h - \sum_{l=1}^{m}\alpha_l(1 - Y_l W_0^T X_l)\right\}$$

Subject to:

$$\forall_{l=1}^{m} : 0 \le \alpha_l \le C, \sum_{l=1}^{m}\alpha_l Y_l = 0$$

3. Our Proposed Solutions

Incremental Learning: Update decision function

a. Old decision function

$$f_p(x) = W_0^T \phi(x) + b \qquad W_0 = \sum_{i=1}^{L} \alpha_i^* y_i \phi(x_i)$$

b. New decision function with user's feedbacks

$$W = W_0 + \sum_{l=1}^{m} \alpha_l^* y_l \phi(x_l) = \sum_{i=1}^{L} \alpha_i^* y_i \phi(x_i) + \sum_{l=1}^{m} \alpha_l^* y_l \phi(x_l)$$

$$f_c(x) = W^T \phi(x) + b = \sum_{i=1}^{L} \alpha_i^* y_i \kappa(x, x_i) + \sum_{l=1}^{m} \alpha_l^* y_l \kappa(x, x_l) + b$$

3. Our Proposed Solutions

Incremental Learning: Update Feature Weights

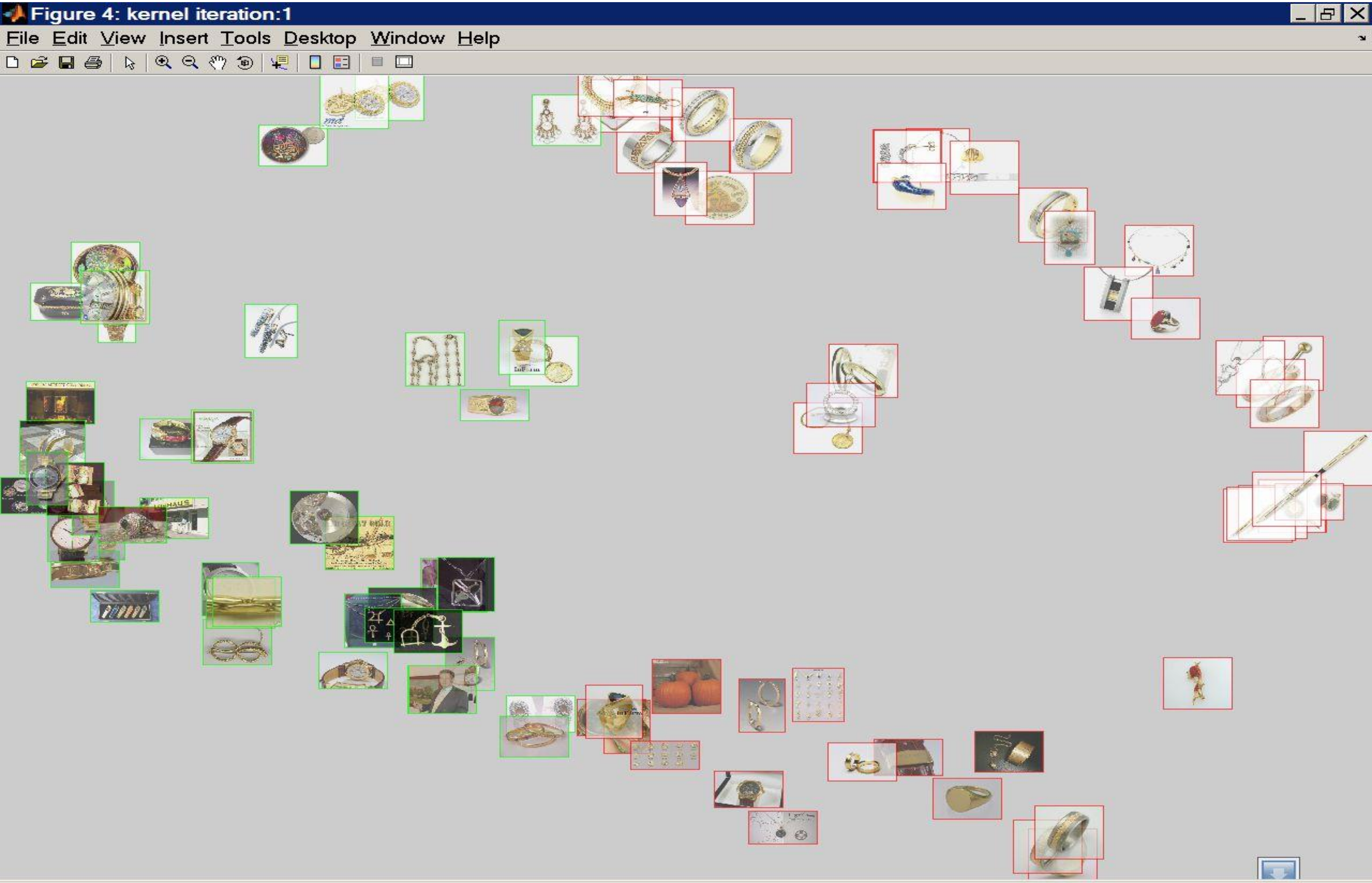
$$J(\beta) = \frac{1}{2}\sum_{l=1}^{m}\sum_{h=1}^{m}\alpha_l^*\alpha_h^* y_l y_h \sum_{i=1}^{3}\beta_i K_i(x_l, x_h) - \sum_{l=1}^{m}\alpha_l^*\left(1 - y_l\sum_{i=1}^{L}\alpha_i^* y_i \sum_{i=1}^{3}\beta_i K_i(x_i, x_l)\right)$$

$$\forall_{i=1}^{3}: \quad \frac{\partial J(\beta)}{\partial \beta_i} = \frac{1}{2}\sum_{l=1}^{m}\sum_{h=1}^{m}\alpha_l^*\alpha_h^* y_l y_h K_i(x_l, x_h) + \sum_{l=1}^{m}\sum_{i=1}^{L}\alpha_l^*\alpha_i^* y_l y_i K_i(x_i, x_l)$$

$$\forall_{i=1}^{3}: \quad \beta_i^{t+1} = \beta_i^t + \gamma_t\left[\frac{1}{2}\sum_{l=1}^{m}\sum_{h=1}^{m}\alpha_l^*\alpha_h^* y_l y_h K_i(x_l, x_h) + \sum_{l=1}^{m}\sum_{j=1}^{L}\alpha_l^*\alpha_i^* y_l y_j K_i(x_j, x_l)\right]$$

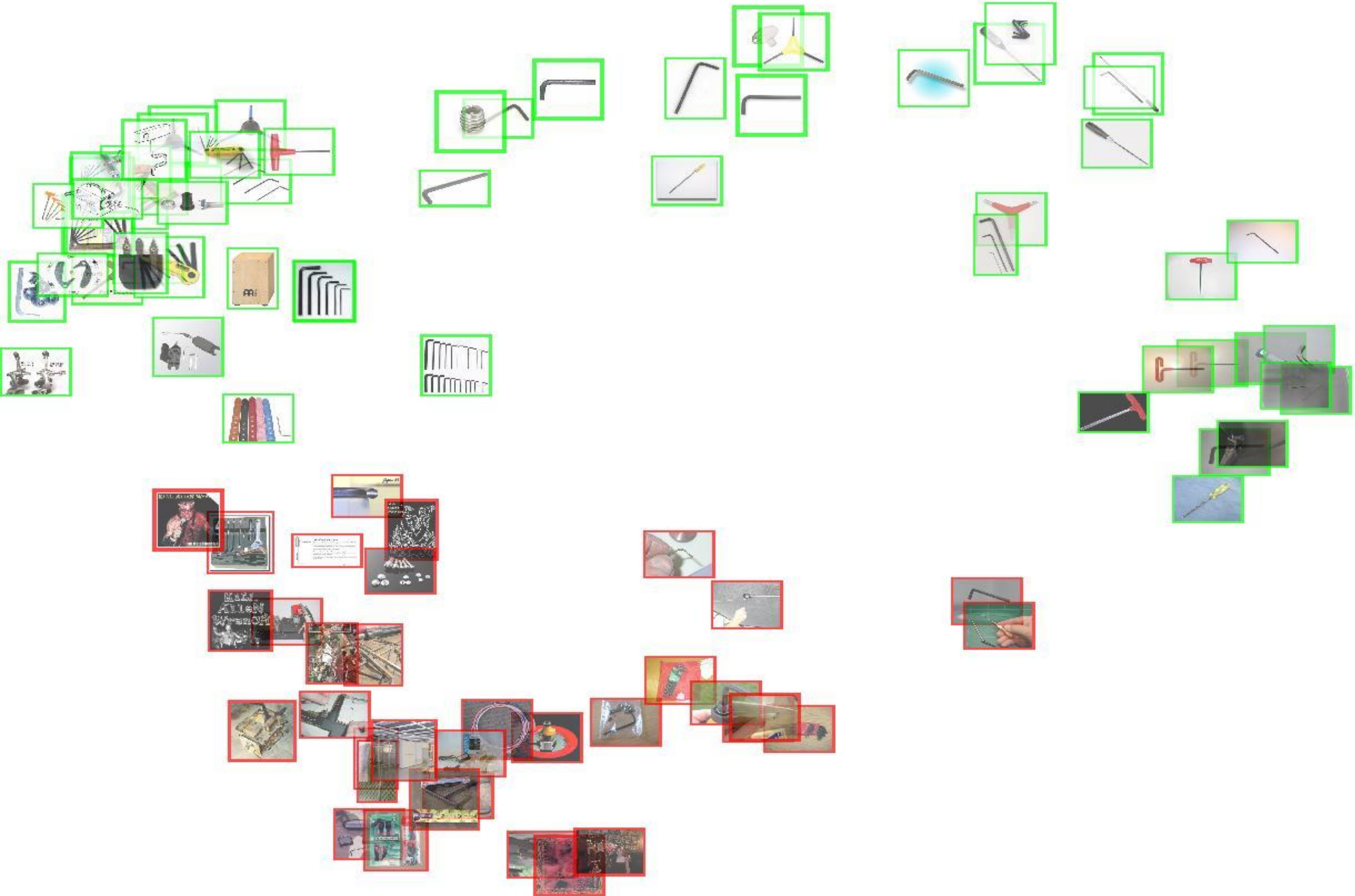
3. Our Proposed Solutions

Make the decision function to be visible!



3. Our Proposed Solutions

Enlarge the margin between two classes!



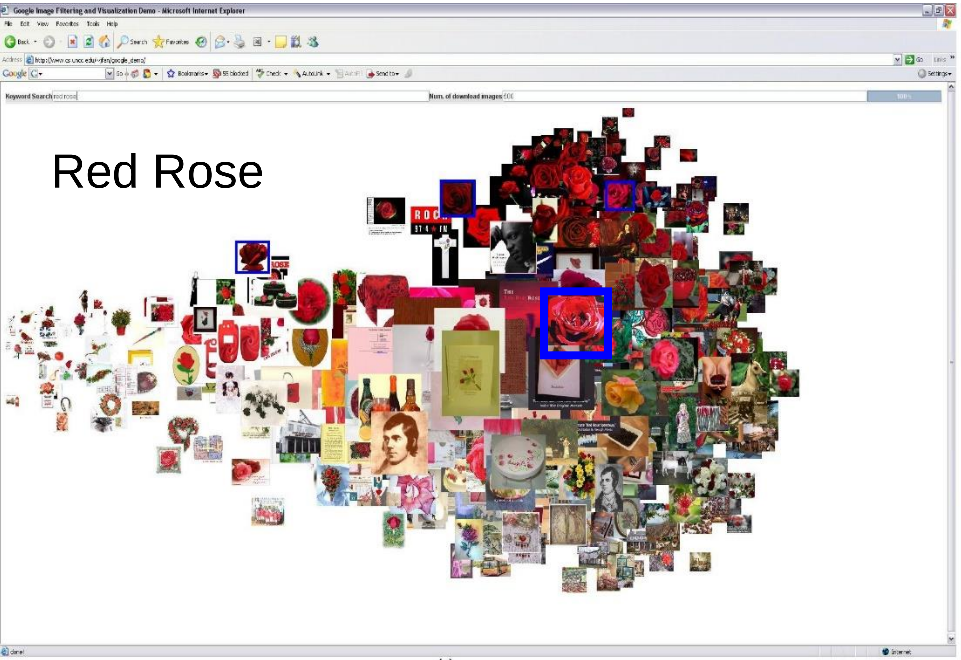
3. Our Proposed Solutions

Enlarge the margin between two classes!

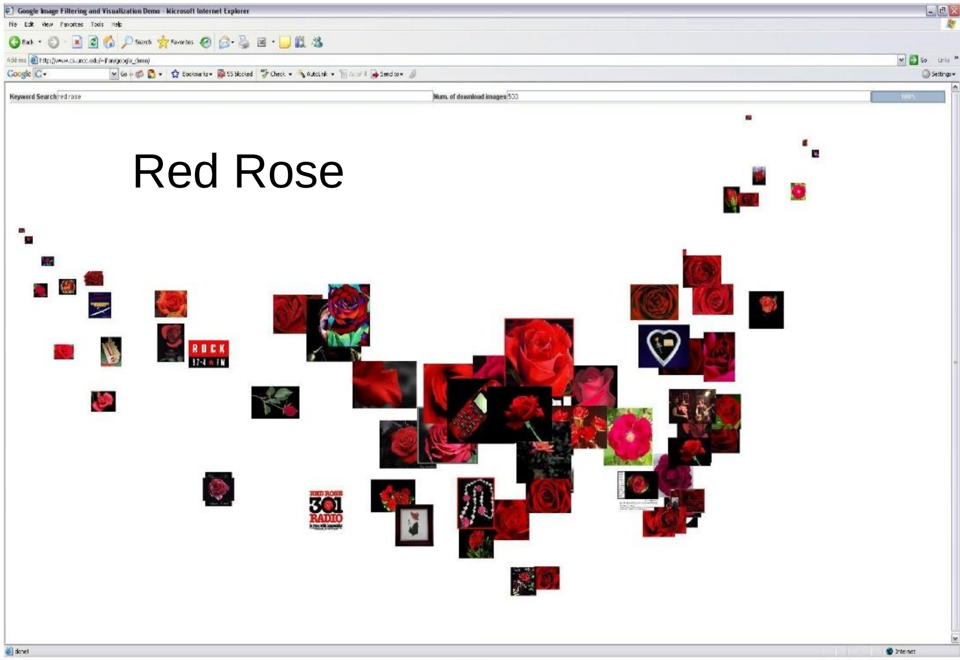
3. Our Proposed Solutions

Larger margin has good generalization property!

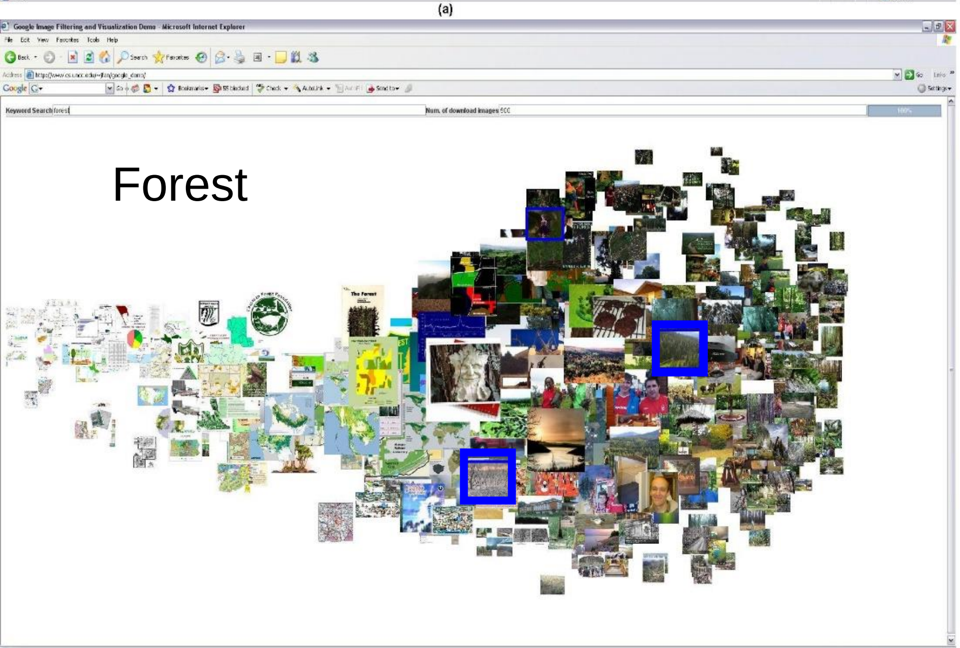
3. Our Proposed Solutions



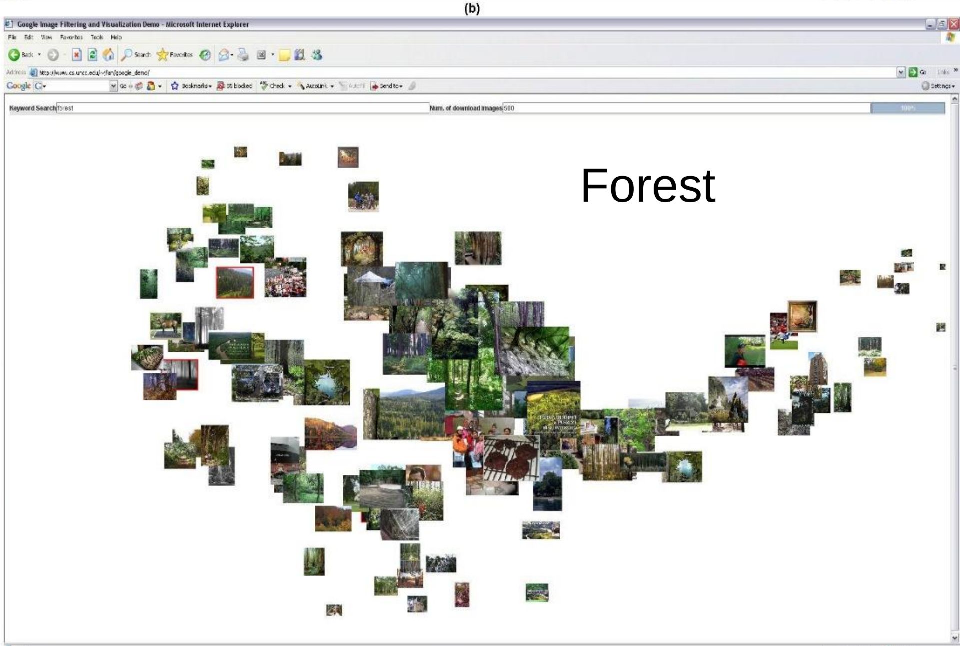
(a)



(b)

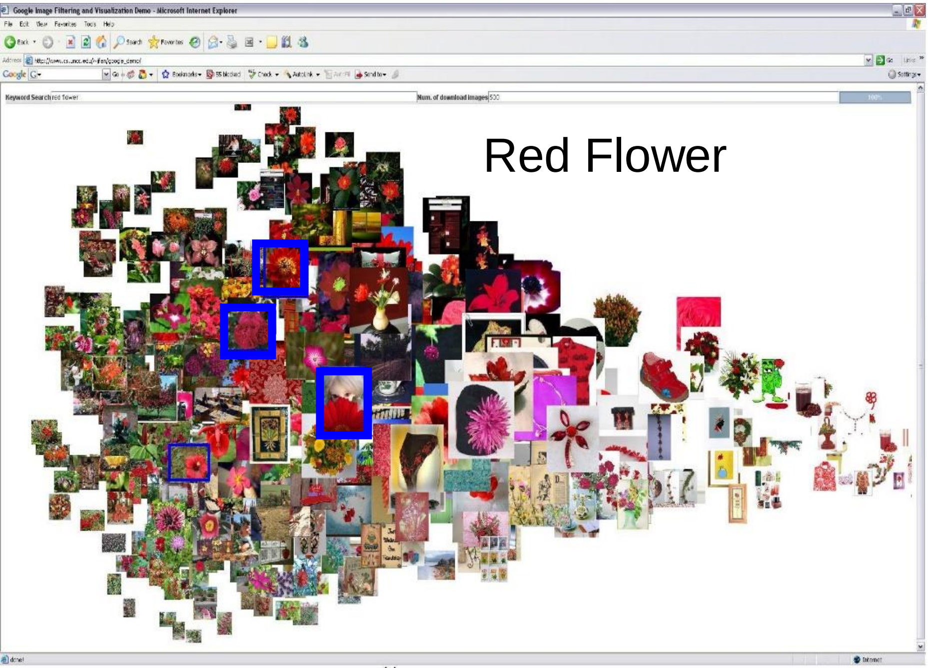


(a)

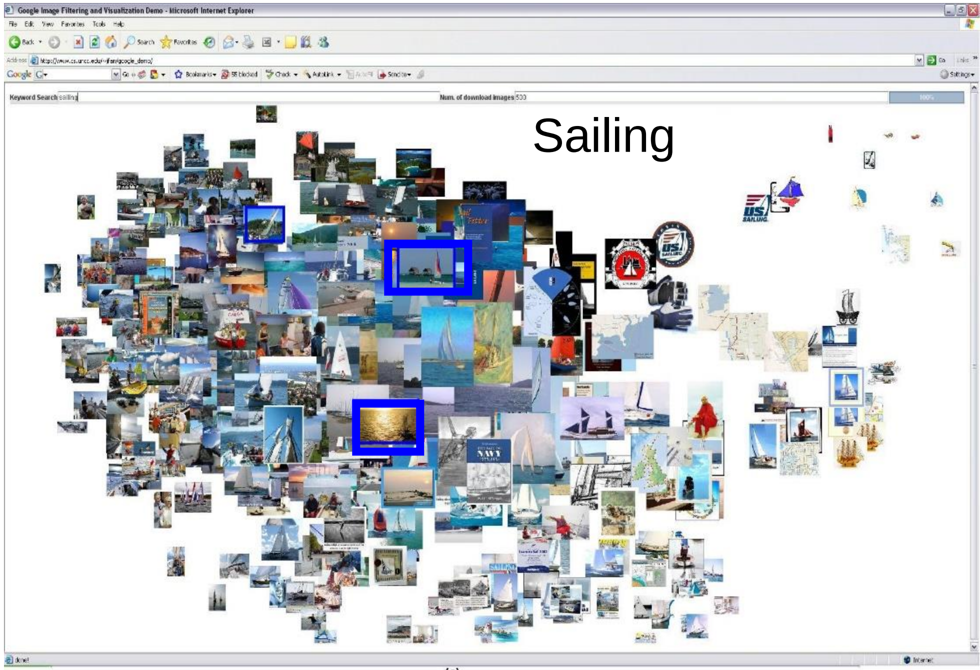


(b)

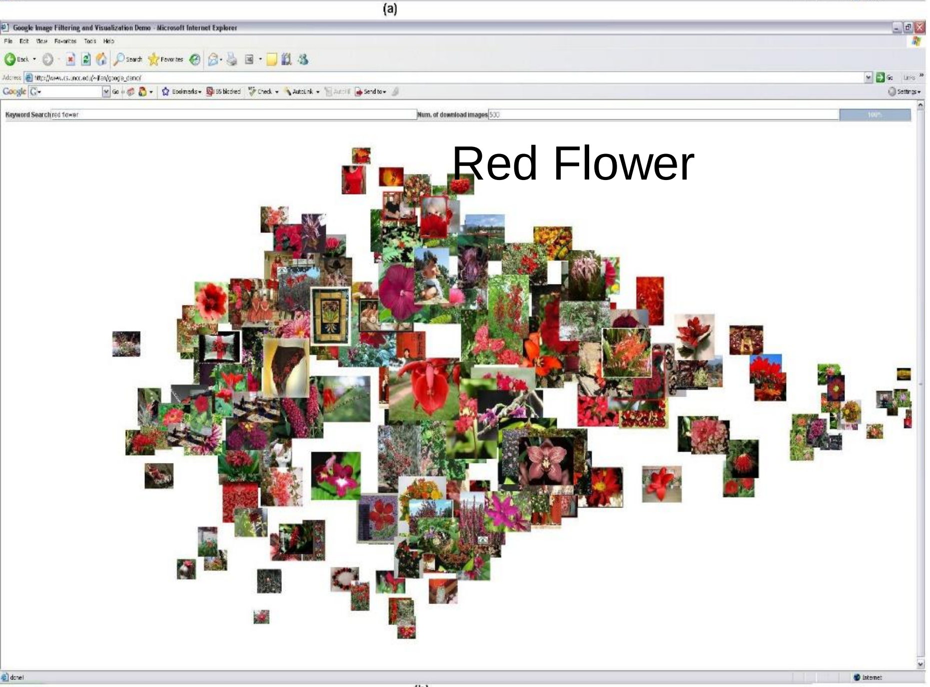
3. Our Proposed Solutions



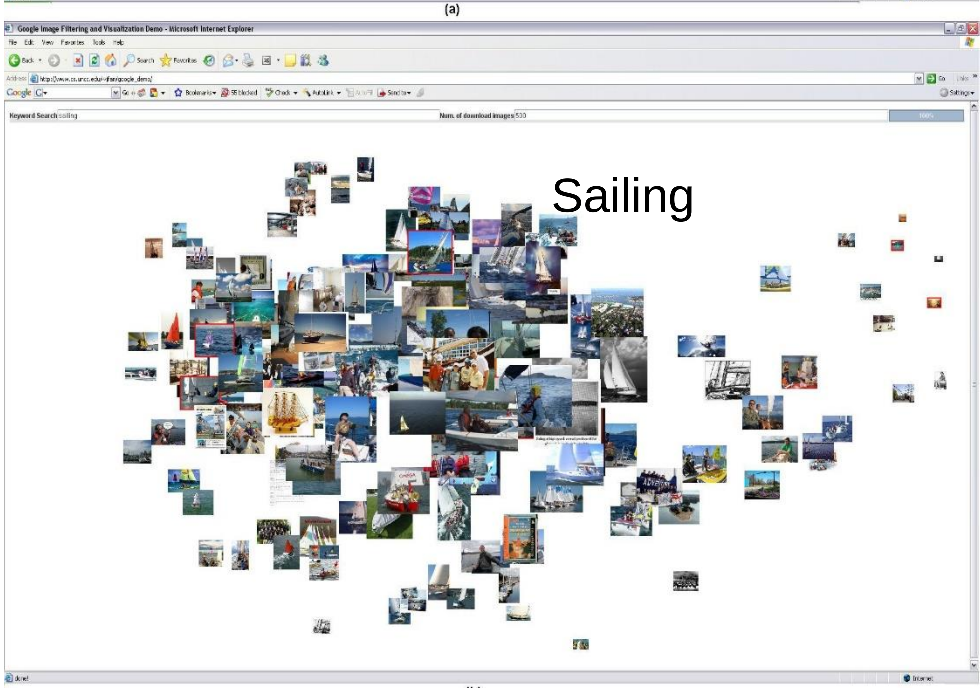
(a)



(a)



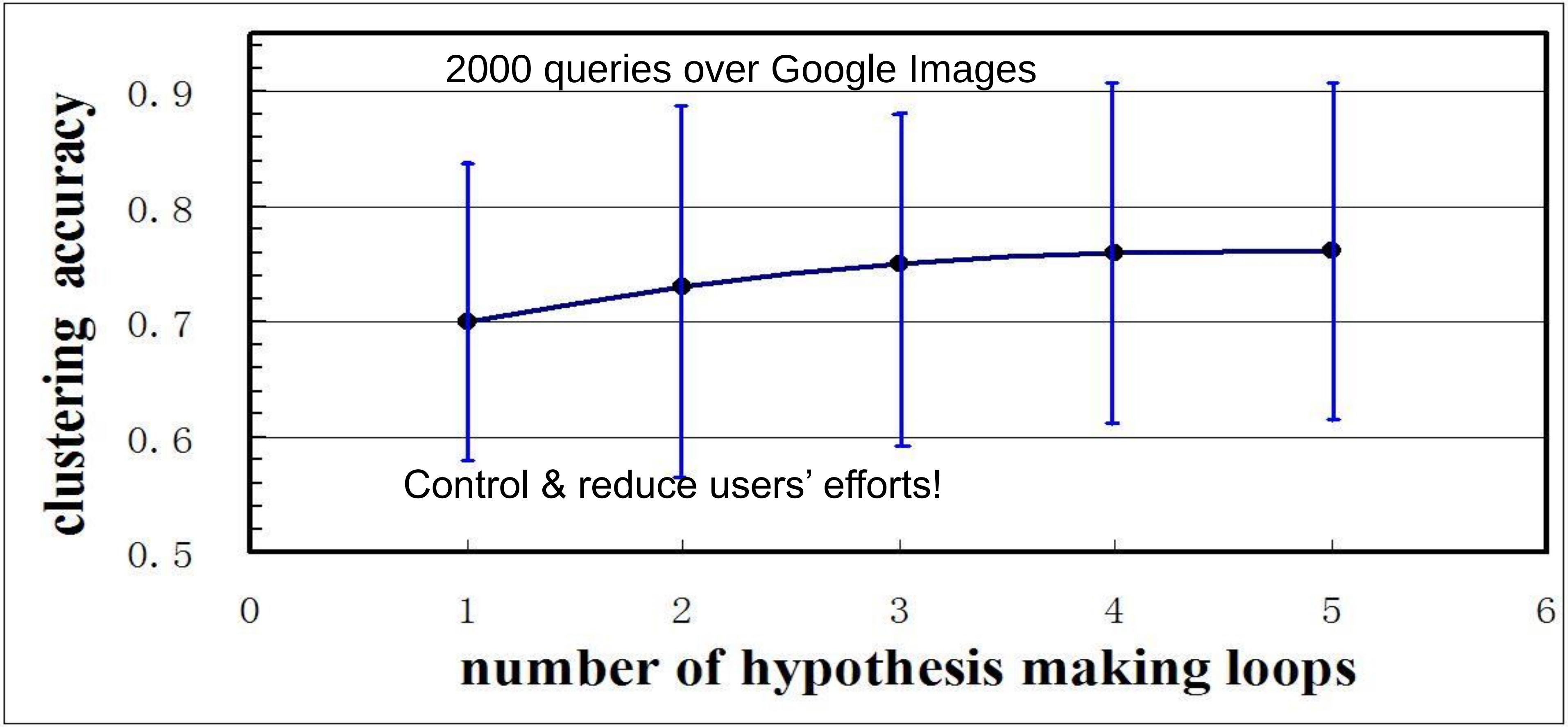
(b)



(b)

3. Our Proposed Solutions

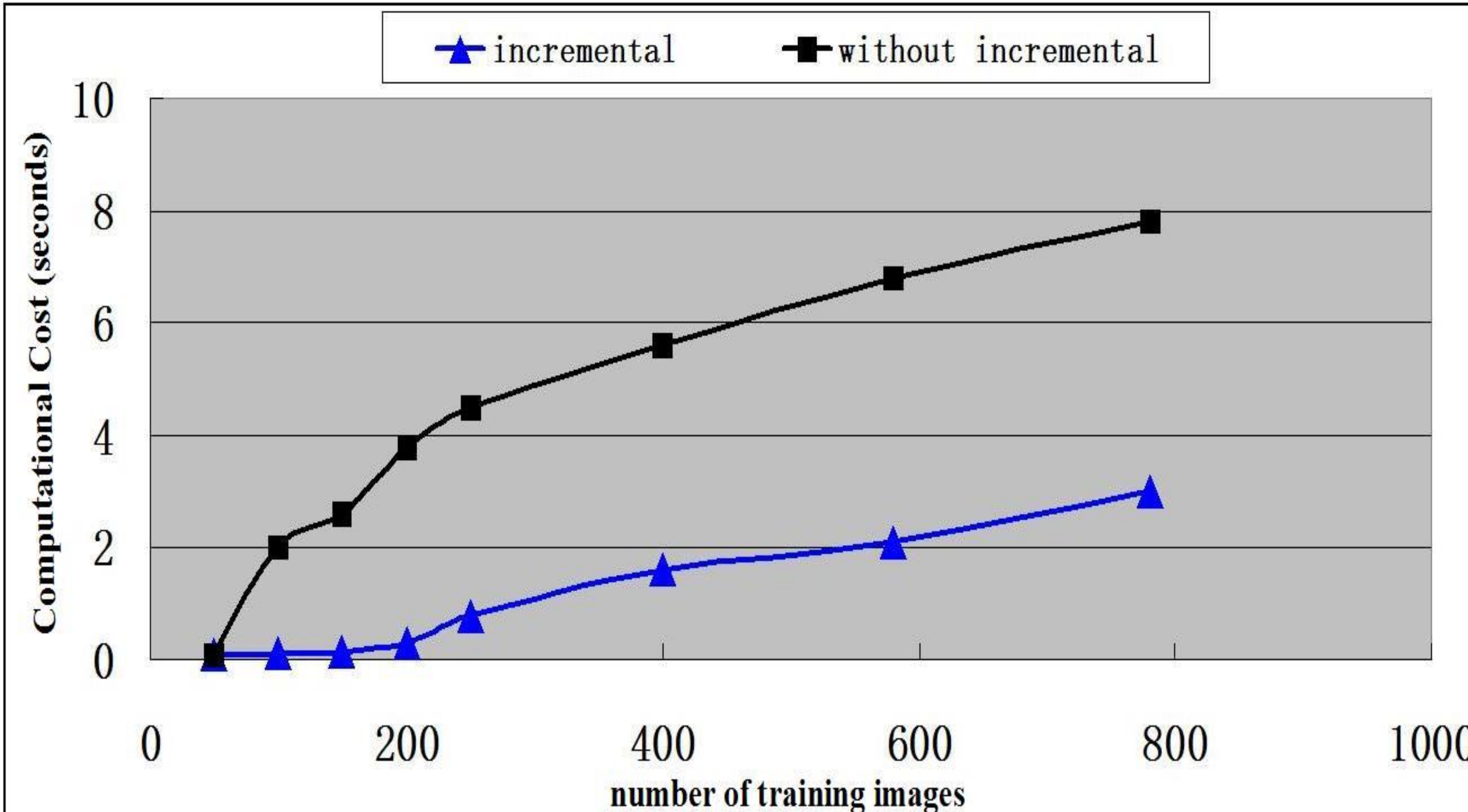
Convergence for Incremental Learning



http://www.cs.uncc.edu/~jfan/google_demo/

3. Our Proposed Solutions

Incremental Learning is critical for Visual Analytics



Conclusions

In this presentation, we have introduced a novel approach to filter out the junk images from Google search results!

Multiple techniques are integrated for achieving this goal:

(a) More effective query intention capturing via hyperbolic image visualization;

(b) Kernel-based image clustering;

(a) Iterative kernel updating.