Social Media Summer School 2011 – Antalya, Turkey – June 27, 2011

Image and Video Tagging in the Internet Era

Xian-Sheng Hua
Microsoft, Bing Multimedia Search

About Xian-Sheng Hua

- 2001
  - Ph.D. majored in applied mathematics, Peking University, Beijing, China
- 2001-2010
  - 9.5 years with Microsoft Research Asia, working on multimedia content analysis, search and other related applications
- 2011-now
  - Principal RSDE Lead with Microsoft Bing Multimedia Search
  - Working on turning research into reality

bing

Outline

Session	Time	Topic
0	09:30 – 09:40	Introduction
1	09:40 – 10:20	Learning-Based Tagging
2	10:20 – 11:40	Social Tagging and Tag Processing
		Including a break (10:45 – 11:00)
3	11:40 – 12:10	Data-Driven Web-Scale Tagging
4	12:10 – 12:30	Future Directions/QA

Outline

Session	Time	Topic
0	09:30 – 09:40	Introduction
1	09:40 – 10:20	Learning-Based Tagging
2	10:20 – 11:40	Social Tagging and Tag Processing
		Including a break (10:45 – 11:00)
3	11:40 – 12:10	Data-Driven Web-Scale Tagging
4	12:10 – 12:30	Future Directions/QA

What's Happening

5+ billion (Sep 2010)

- 160 years to view all of them (1s per image)
- **3000+** uploads/minute
- 2% Internet users visit (2009)
- Daily time on site: 4.7 minutes (2009)

400 million (2010)

- 2,000 years to see all of them
- ~20 hours uploaded/minute (09)
- 20% Internet users visit (2009)
- Daily time on site: 23 minutes (2009)
- 2007 bandwidth = entire Internet in 2000
- 3B+ views per day (2010)

60 billion (Dec 2010)

- 1,920 years to view all of them (1s per image)
- ~138M uploads/minute
- 24% Internet users visit (2009)
- Daily time on site: 30 minutes (2009)

Top Five Sites – Alexa Traffic Rank

Characteristics of Internet Multimedia

Huge Amount of Data

Increasing Very Rapidly

Variances Are Very Large

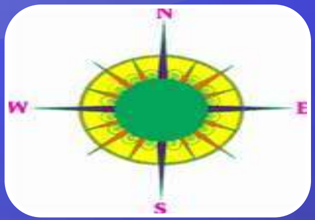
Be consumed Frequently

Affection Highly Involved

Connected to Each Other

Variety of Internet Multimedia Applications


Search



Browsing


Sharing

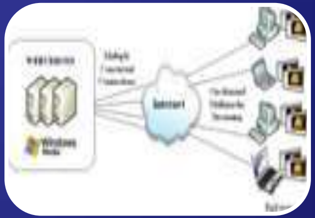

Authoring/Editing


Copy Detection


Recommendation


Tagging


Mining


Streaming


Summarization


Visualization


Advertising


Categorization


Forensics

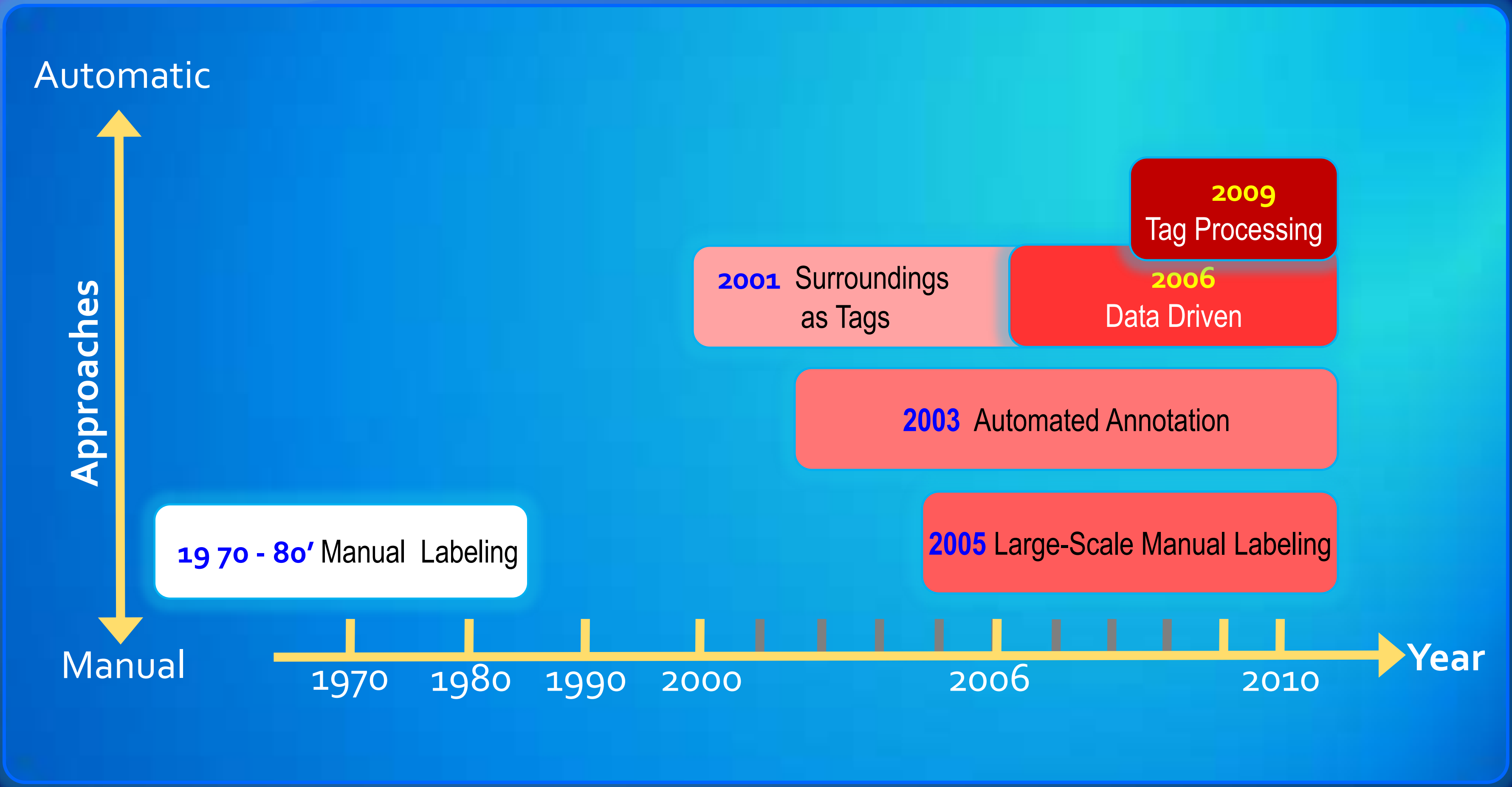

Media on Mobiles

One of the Key Challenge: Search

- Image/Video Tagging is
  - An approach or process of converting visual content into a set of textual words to describe the semantics contained in the image/video to enable content-aware search
  - It is a simplified target for media understanding
  - It is easy to be adopted in a variety of applications

Person, Grass, Tree, Building, Road, Face, …

Evolution of Media Tagging



Different Types of Tags

- Tags learned from examples
  - Video/Image annotation or semantic concept detection
- Owner-input tags
  - For example, tags for Flickr Images
- Tags obtained from labeling platforms
  - ESP game
  - LabelMe
  - reCAPTCHA
  - Amazon Machinery Turk
  - Other labeling/tagging tools
- Tags obtained through search engines
  - Query Association ("Implicit" tagging)

bing

Synonyms

(Automatic) Annotation

≈ Concept Detection

≈ (Automatic) Tagging

≈ (Automatic) Labeling

Tags

≈ Labels

≈ Concepts

bing

Outline

Session	Time	Topic
0	09:30 – 09:40	Introduction
1	09:40 – 10:20	Learning-Based Tagging
2	10:20 – 11:40	Social Tagging and Tag Processing
		Including a break (10:45 – 11:00)
3	11:40 – 12:10	Data-Driven Web-Scale Tagging
4	12:10 – 12:30	Future Directions/QA

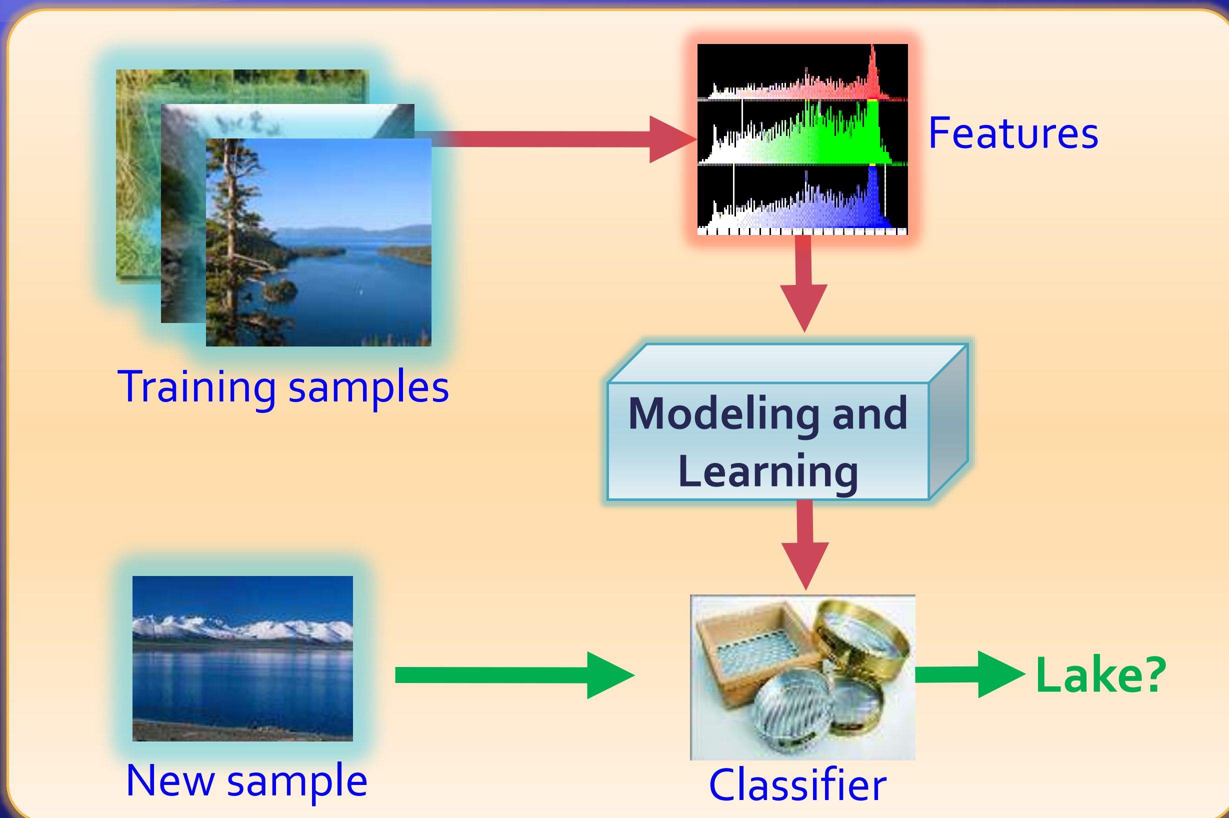
Outline

Session	Time	Topic
0	09:30 – 09:40	Introduction
1	09:40 – 10:20	Learning-Based Tagging
2	10:20 – 11:40	Social Tagging and Tag Processing
		Including a break (10:45 – 11:00)
3	11:40 – 12:10	Data-Driven Web-Scale Tagging
4	12:10 – 12:30	Future Directions/QA

Learning Based Tagging

- Introduction
- Representative approaches
  - Correlative Multi-Labeling Learning
  - Online Active Multi-Label Annotation
- Discussion

Learning-Based Tagging/Annotation



Person

Grass

Tree

Building

Road

Face

Approaches

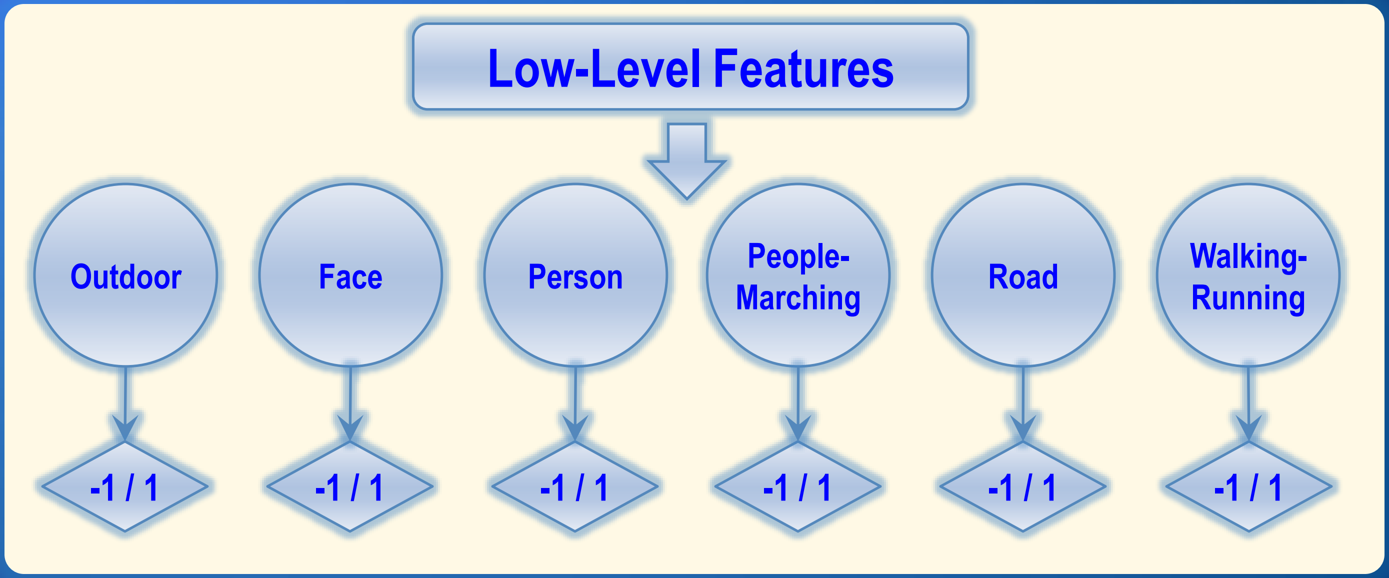
- Different learning approaches
  - Supervised/Semi-Supervised Learning
  - Active Learning
  - Incremental Learning
  - Transfer Learning
  - Single-Label/Multi-Class/Multi-Label Learning
  - Multi-Instance Learning
  - ...
- We choose two exemplar approaches
  - Correlative Multi-Label Learning
  - Online Active Multi-Label Learning

Image/Video Annotation by
Correlative Multi-Label Learning

Credit: Guo-Jun Qi, Xian-Sheng Hua, Yong Rui, et al. Correlative Multi-Label Video Annotation. ACM Multimedia 2007. Augsburg, Germany. (Best Paper Award)

Automated Annotation – 1st Paradigm

- A typical strategy – Individual Concept Detection
  - Annotate multiple concepts separately



To Exploit Label Correlations

√ Person
√ Street
√ Building

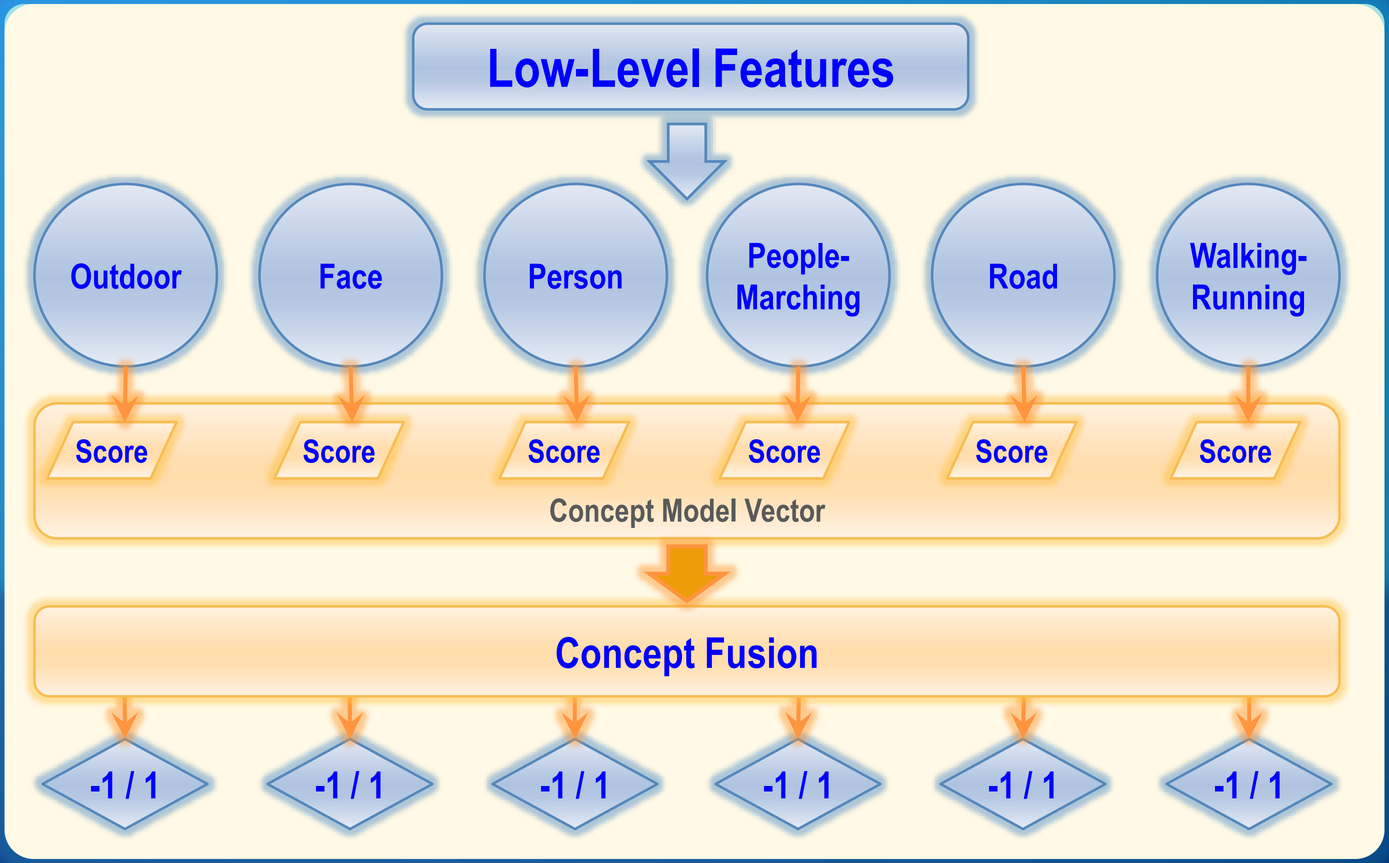
× Beach
× Mountain

√ Crowd
√ Outdoor
√ Walking/Running

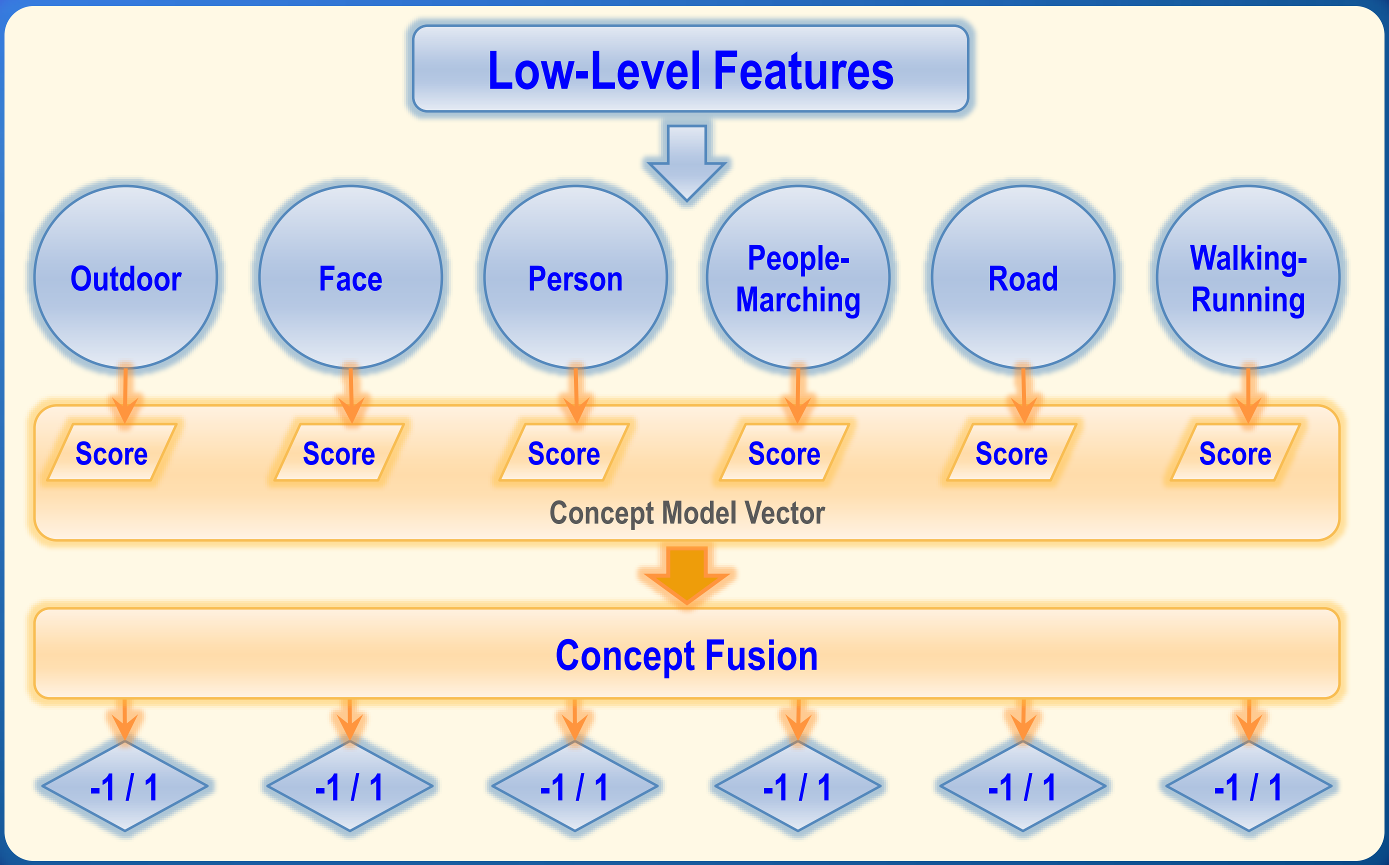
√? Marching

Automated Annotation – 2nd Paradigm

- Another typical strategy – Fusion-Based
  - Context Based Concept fusion (CBCF)

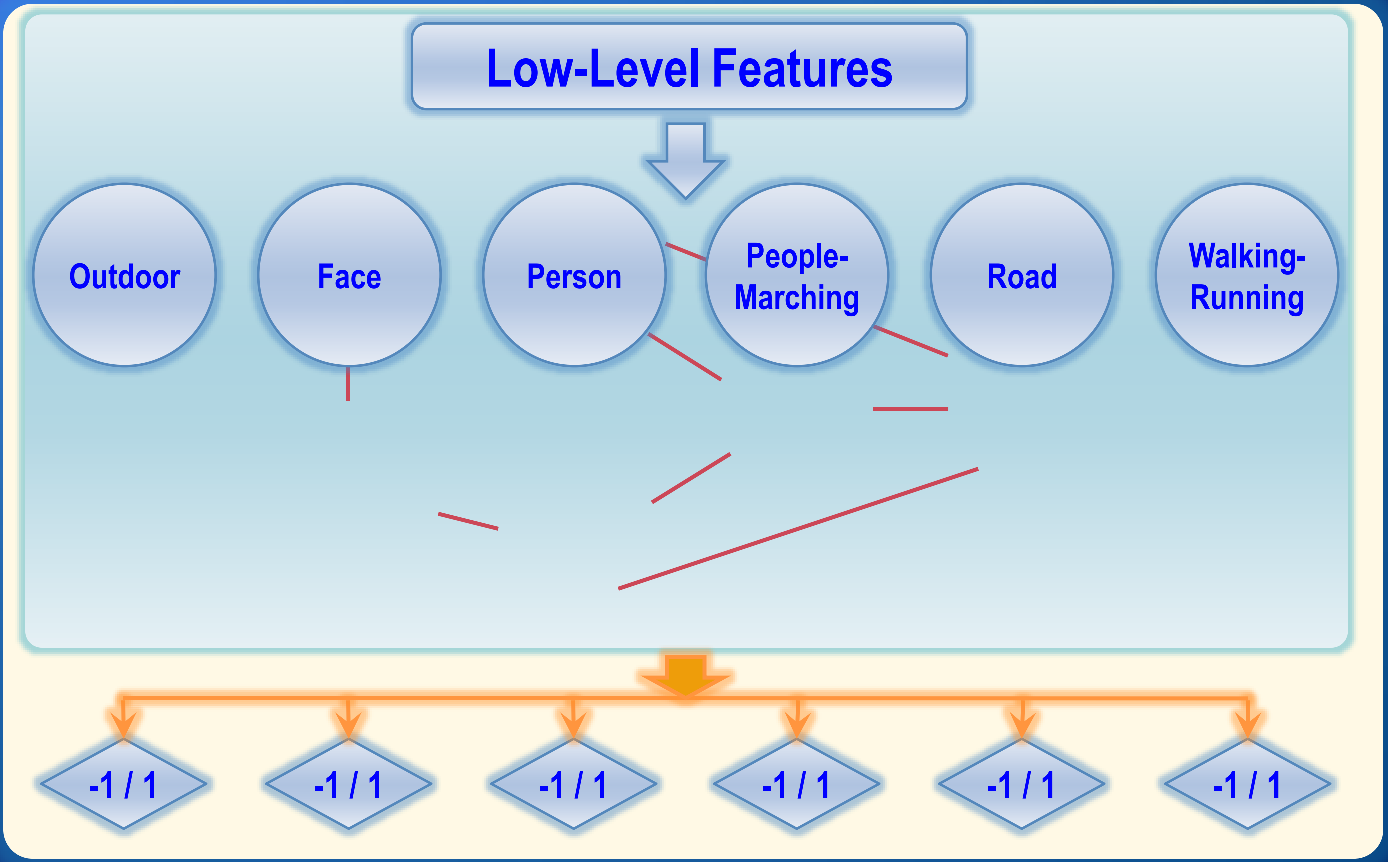


bing


Low-Level Features
Outdoor
Face
Person
People-Marching
Road
Walking-Running
Score
Score
Score
Score
Score
Score
Concept Model Vector
Concept Fusion
-1 / 1
-1 / 1
-1 / 1
-1 / 1
-1 / 1
-1 / 1

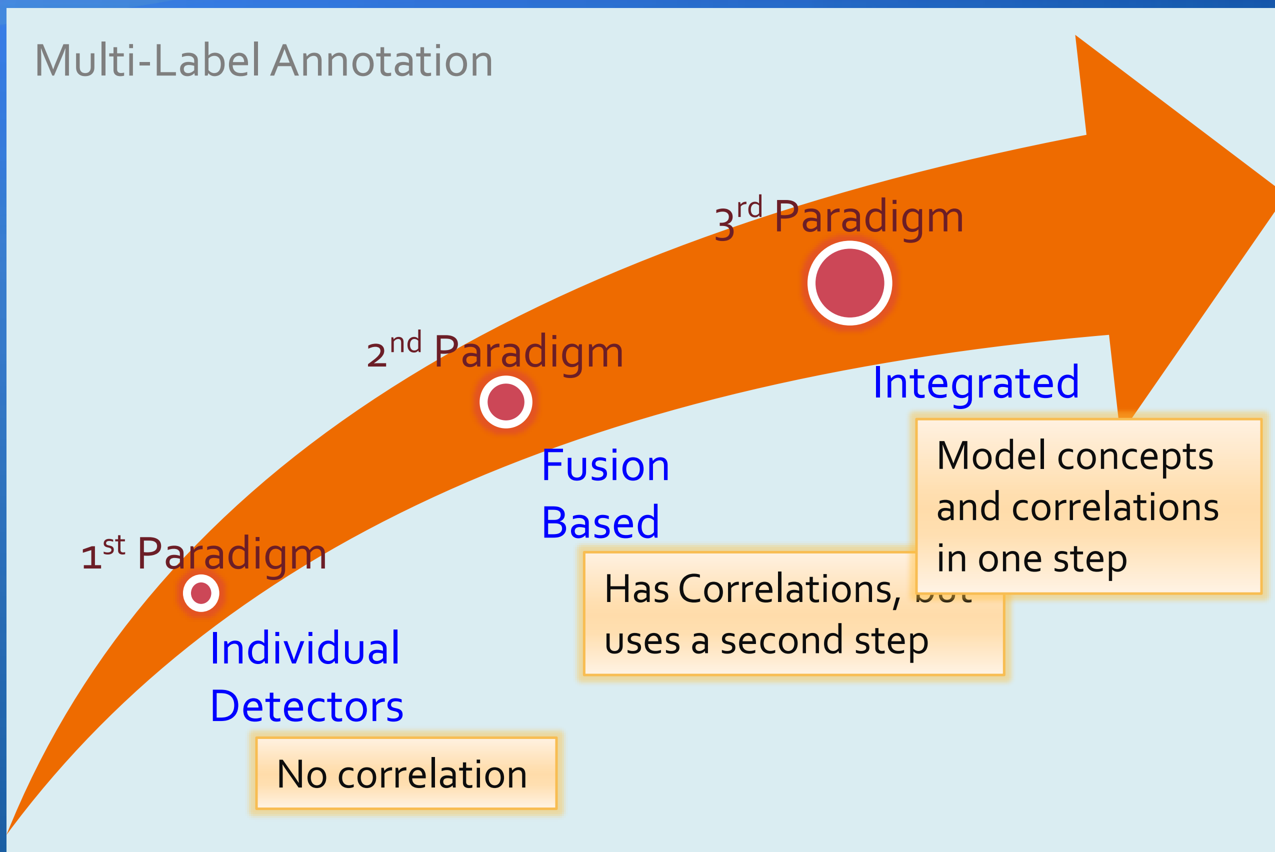
Automated Annotation – 3rd Paradigm

- Our strategy – Integrated Concept Detection
  - Correlative Multi-Label Learning (CML)



bing

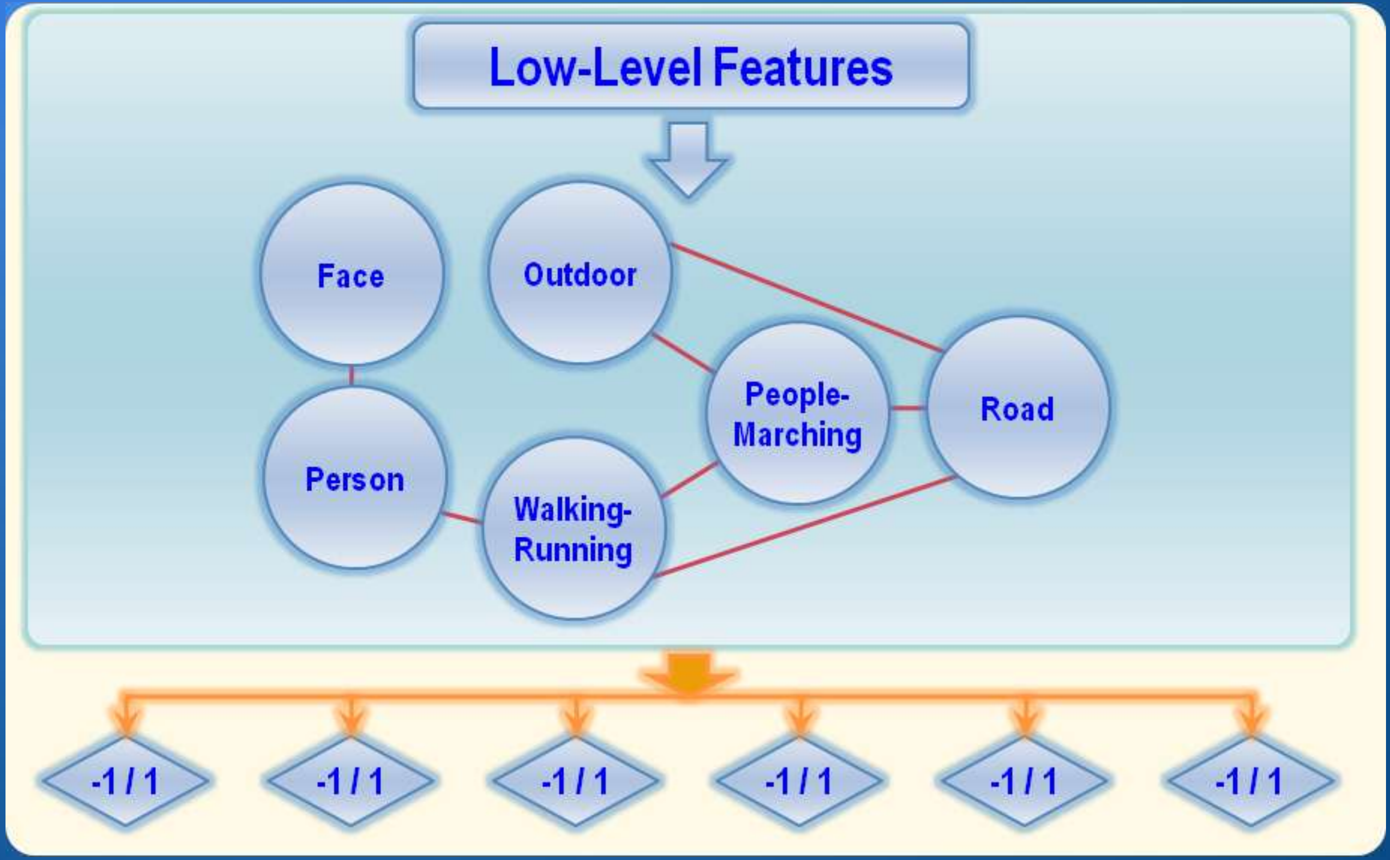
CML Roadmap



Multi-Label Annotation
1st Paradigm
Individual Detectors
No correlation
2nd Paradigm
Fusion Based
Has Correlations,
uses a second step
3rd Paradigm
Integrated
Model concepts and correlations in one step

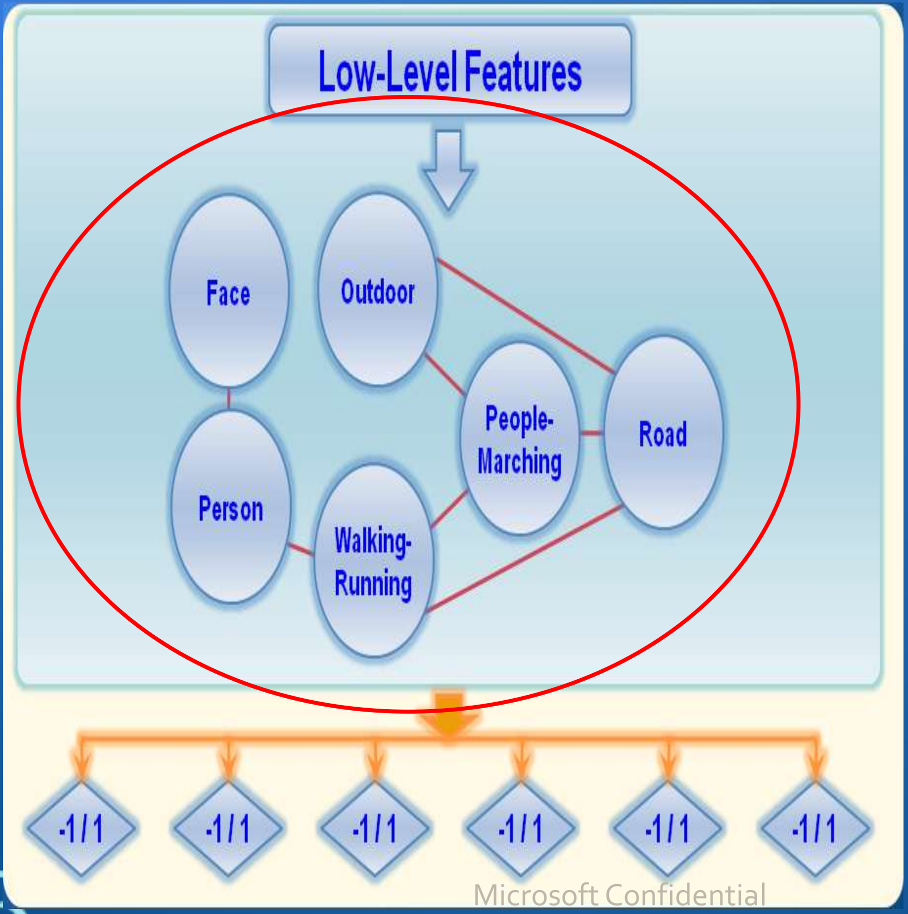
Automated Annotation – 3rd Paradigm

- Our strategy – Integrated Concept Detection
  - Correlative Multi-Label Learning (CML)



How To Model Concept Correlations

- How to model concepts and the correlations among concept in a single step



Our Strategy

Converting correlations into features.

Constructing a new feature vector that captures both

- The characteristics of concepts, and
- The correlations among concepts

Correlative Multi-Label Video Annotation

- Notations
  - input pattern $\boldsymbol{x} = (x_1, x_2, \cdots, x_D)^T \in \mathcal{X}$
  - K dimensional concept label $\boldsymbol{y} \in \mathcal{Y} = \{+1, -1\}^K$
  - aims at learning $F(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w}) = \langle \boldsymbol{w}, \theta(\boldsymbol{x}, \boldsymbol{y}) \rangle$
  - new feature vector $\theta(\boldsymbol{x}, \boldsymbol{y})$
  - vector $\boldsymbol{y}^*$ can be predicted by $\boldsymbol{y}^* = \max_{y \in \mathcal{Y}} F(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w})$

bing

Correlative Multi-Label Video Annotation

- Modeling concept and correlations simultaneously

- construct $\theta(\boldsymbol{x}, \boldsymbol{y})$

Type I The elements for *individual* concept modeling:

$$\theta^{l}_{d,p}(\boldsymbol{x}, \boldsymbol{y}) = x_d \cdot \delta\,[\![y_p = l]\!]\,,$$
$$l \in \{+1, -1\}, 1 \le d \le D, 1 \le p \le K$$

-1 1

$(x_1, x_2, \cdots, x_D)$

$(y_1, y_2, \cdots, y_K)$

$\boldsymbol{x} = (0.1,\ 0.2,\ 0.3,\ 0.4,\ 0.5,\ 0.6)$

Five Semantic Concepts : *person, road, beach, car, tree*

$\boldsymbol{y} = (1,\ 1,\ -1,\ -1,\ 1)$

$\theta^{Yes/No}_{feature,\ concept}$

$\theta^{1}_{1,1} = 0.1 \quad \theta^{-1}_{1,1} = 0$

$\theta^{1}_{1,2} = 0.1 \quad \theta^{-1}_{1,2} = 0 \quad \theta^{1}_{1,3} = 0 \quad \theta^{-1}_{1,3} = 0.1$

$\theta^{1}_{2,1} = 0.2 \quad \theta^{-1}_{2,1} = 0 \quad \theta^{1}_{2,2} = 0.2 \quad \theta^{-1}_{2,2} = 0$

Correlative Multi-Label Video Annotation

- Modeling concept and correlations simultaneously

- construct $\theta(\boldsymbol{x}, \boldsymbol{y})$

Type II The elements for concept correlations:

$$\theta_{p,q}^{m,n}(\boldsymbol{x}, \boldsymbol{y}) = \delta [\![y_p = m]\!] \cdot \delta [\![y_q = n]\!]$$
$$m, n \in \{+1, -1\}, 1 \leq p < q \leq K$$

-1 1

$(y_1, y_2, \cdots, y_K)$

$\boldsymbol{x} = (0.1,\ 0.2,\ 0.3,\ 0.4,\ 0.5,\ 0.6)$

Five Semantic Concepts : *person, road, beach, car, tree*

$\boldsymbol{y} = (1,\ 1,\ -1,\ -1,\ 1)$

$\theta^{Y/N,\ Y/N}_{Concept1,\ Concept2}$

$\theta_{1,2}^{1,1} = 1 \quad \theta_{1,2}^{1,-1} = 0 \quad \theta_{1,2}^{-1,1} = 0 \quad \theta_{1,2}^{-1,-1} = 0$

$\theta_{1,3}^{1,1} = 0 \quad \theta_{1,3}^{1,-1} = 1 \quad \theta_{1,3}^{-1,1} = 0 \quad \theta_{1,3}^{-1,-1} = 0$

bing

Correlative Multi-Label Video Annotation

- Modeling concept and correlations

- construct $\theta(\boldsymbol{x}, \boldsymbol{y})$

Type I The elements for *individual* concept modeling:

$$\theta^{l}_{d,p}(\boldsymbol{x}, \boldsymbol{y}) = x_d \cdot \delta\,[\![y_p = l]\!]\,,$$
$$l \in \{+1, -1\}, 1 \le d \le D, 1 \le p \le K$$

Type II The elements for concept correlations:

$$\theta^{m,n}_{p,q}(\boldsymbol{x}, \boldsymbol{y}) = \delta\,[\![y_p = m]\!] \cdot \delta\,[\![y_q = n]\!]$$
$$m, n \in \{+1, -1\}, 1 \le p < q \le K$$

- $\theta(\boldsymbol{x}, \boldsymbol{y})$ is a high-dimensional feature vector $\left(2K(D+K-1)\right)$
- $\theta(\boldsymbol{x}, \boldsymbol{y})$ has very compact kernel representation

$$\langle \theta(\boldsymbol{x}, \boldsymbol{y}), \theta(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) \rangle = \langle \boldsymbol{x}, \tilde{\boldsymbol{x}} \rangle \textstyle\sum_{1 \le k \le K} \delta\,[\![y_k = \tilde{y}_k]\!]$$
$$+ \textstyle\sum_{1 \le p < q \le K} \delta\,[\![y_p = \tilde{y}_p]\!]\, \delta\,[\![y_q = \tilde{y}_q]\!]$$

Correlative Multi-Label Video Annotation

- Learning the classifier

Misclassification Error	$\Delta F_i(\boldsymbol{y}) \triangleq F(\boldsymbol{x}_i, \boldsymbol{y}_i) - F(\boldsymbol{x}_i, \boldsymbol{y})$ $= \langle \boldsymbol{w}, \Delta\theta_i(\boldsymbol{y})\rangle \leq 0, \forall \boldsymbol{y} \neq \boldsymbol{y}_i, \boldsymbol{y} \in \mathcal{Y}$
Loss function	$\ell_h(\boldsymbol{x}_i, \boldsymbol{y}; \boldsymbol{w}) = (1 - \langle \boldsymbol{w}, \Delta\theta_i(\boldsymbol{y})\rangle)_+$
Empirical risk	$\hat{R}_h(\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^n; \boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^n \sum_{\boldsymbol{y}\neq \boldsymbol{y}_i, \boldsymbol{y}\in\mathcal{Y}} \ell_h(\boldsymbol{x}_i, \boldsymbol{y}; \boldsymbol{w})$
Regularization	$\min_{\boldsymbol{w}} \left\{ \hat{R}_h(\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^n; \boldsymbol{w}) + \lambda \cdot \Omega \quad \|\boldsymbol{w}\|^2 \right\}$
Introduce slack variables	$\min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{\lambda}{n} \cdot \sum_{i=1}^n \sum_{\boldsymbol{y}\neq \boldsymbol{y}_i, \boldsymbol{y}\in\mathcal{Y}} \xi_i(\boldsymbol{y})$ $s.t. \langle \boldsymbol{w}, \Delta\theta_i(\boldsymbol{y})\rangle \geq 1 - \xi_i(\boldsymbol{y}), \xi_i(\boldsymbol{y}) \geq 0 \boldsymbol{y} \neq \boldsymbol{y}_i, \boldsymbol{y} \in \mathcal{Y}$
Lagrange dual	$\max_{\alpha} \sum_{i, \boldsymbol{y}\neq \boldsymbol{y}_i} \alpha_i(\boldsymbol{y}) - \frac{1}{2}\sum_{i, \boldsymbol{y}\neq \boldsymbol{y}_i} \sum_{j, \tilde{\boldsymbol{y}}\neq \boldsymbol{y}_j} \langle \Delta\theta_i(\boldsymbol{y}), \Delta\theta_j(\tilde{\boldsymbol{y}})\rangle$ $s.t. 0 \leq \sum_{\boldsymbol{y}\neq \boldsymbol{y}_i, \boldsymbol{y}\in\mathcal{Y}} \alpha_i(\boldsymbol{y}) \leq \frac{\lambda}{n}, \boldsymbol{y} \neq \boldsymbol{y}_i, \boldsymbol{y} \in \mathcal{Y}, 1 \leq i \leq n$
Find solution by SMO	$\boldsymbol{w} = \sum_{1\leq i\leq n, \boldsymbol{y}\in\mathcal{Y}} \alpha_i(\boldsymbol{y})\Delta\theta_i(\boldsymbol{y})$

Correlative Multi-Label Video Annotation

- Connection to Gibbs Random Field

Define a random field

$\wp$ is the set of sites

$\mathcal{N}$ consists of all adjacent sites, that is, this RF is fully connected

$P(y|x,w)$ is a random field

$$\wp = \{i | 1 \le i \le K\}$$

$$\mathcal{N} = \{(p,q) | 1 \le p < q \le K\}$$

Define energy function

$$H(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w}) = -F(\boldsymbol{x},\boldsymbol{y};\boldsymbol{w})$$

Define GRF

$$P(y|x,w) = \frac{1}{Z(x,w)} \exp\{-H(y|x,w)\}$$

Rewrite the classifier

$$F(\boldsymbol{x},\boldsymbol{y};\boldsymbol{w}) = \langle \boldsymbol{w}, \theta(\boldsymbol{x},\boldsymbol{y}) \rangle$$
$$= \sum_{p\in\wp} D_p(y_p;\boldsymbol{x}) + \sum_{(p,q)\in\mathcal{N}} V_{p,q}(y_p,y_q;\boldsymbol{x})$$

$$-H(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w})\}$$

$$D_p(y_p;\boldsymbol{x}) = \sum_{1\le d\le D, l\in\{+1,-1\}} \boldsymbol{w}^l_{d,p}\theta^l_{d,p}(\boldsymbol{x},\boldsymbol{y})$$
$$V_{p,q}(y_p,y_q;\boldsymbol{x}) = \sum_{m,n\in\{+1,-1\}} \boldsymbol{w}^{m,n}_{p,q}\theta^{m,n}_{p,q}(\boldsymbol{x},\boldsymbol{y})$$

Correlative Multi-Label Video Annotation

- Connection to Gibbs Random Field

Define a random field

$\wp$ is the set of sites

$\mathcal{N}$ consists of all adjacent sites, that is, this RF is fully connected

$P(y|x, w)$ is a random field

$$\wp = \{i | 1 \le i \le K\}$$

$$\mathcal{N} = \{(p, q) | 1 \le p < q \le K\}$$

Define energy function

$$H(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}) = -F(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w})$$

Define GRF

$$P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}) = \frac{1}{Z(x, w)} \exp\{-H(y|x, w)\}$$

Rewrite the classifier

$$F(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w}) = \langle \boldsymbol{w}, \theta(\boldsymbol{x}, \boldsymbol{y}) \rangle = \sum_{p \in \wp} D_p(y_p; \boldsymbol{x}) + \sum_{(p,q) \in \mathcal{N}} V_{p,q}(y_p, y_q; \boldsymbol{x})$$

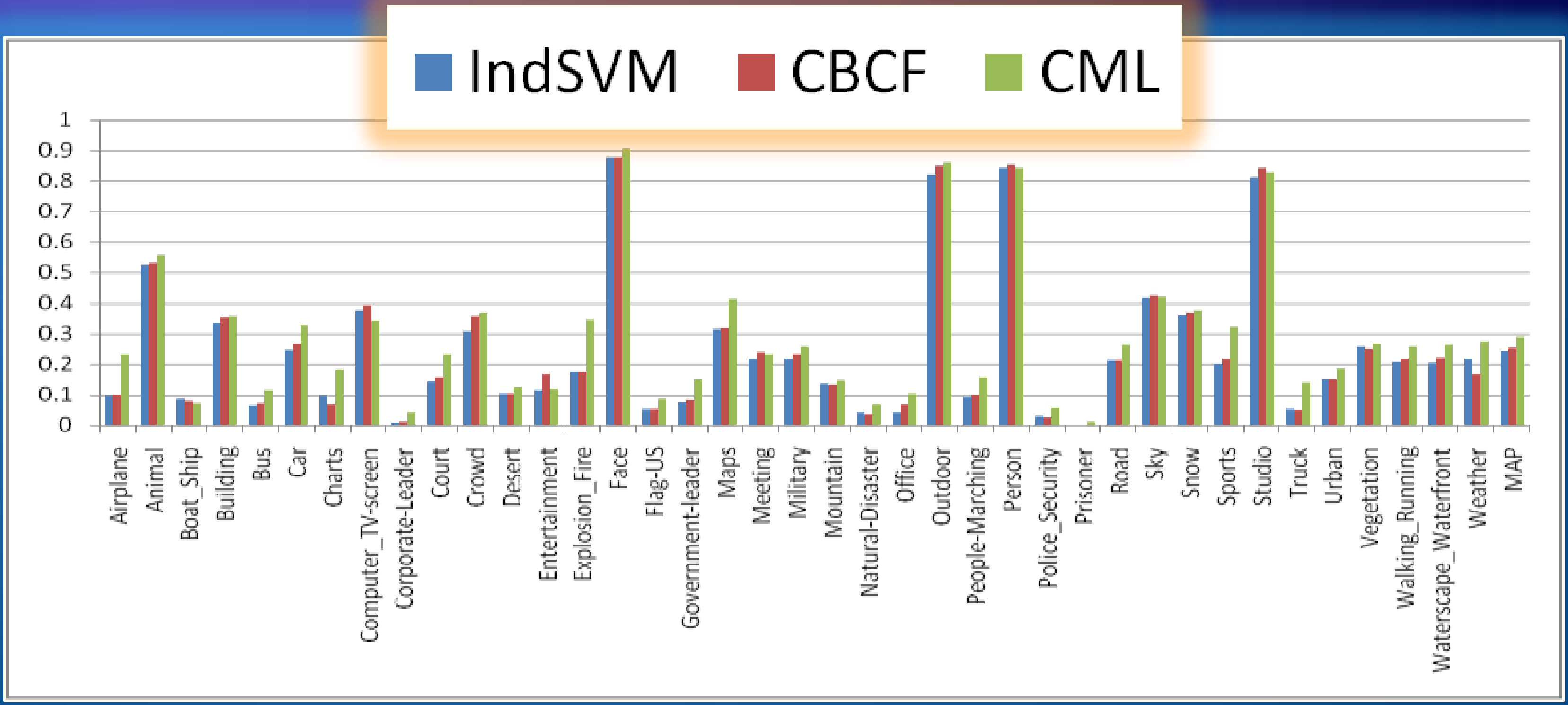
Intuitive explanation of CML

$$P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}) = \frac{1}{Z(\boldsymbol{x}, \boldsymbol{w})} \prod_{p \in \wp} P(y_p|\boldsymbol{x}) \cdot \prod_{(p,q) \in \mathcal{N}} P_{p,q}(y_p, y_q|\boldsymbol{x})$$

$$P(y_p|\boldsymbol{x}) = \exp\{D_p(y_p; \boldsymbol{x})\}$$
$$P_{p,q}(y_p, y_q|\boldsymbol{x}) = \exp\{V_{p,q}(y_p, y_q; \boldsymbol{x})\}$$

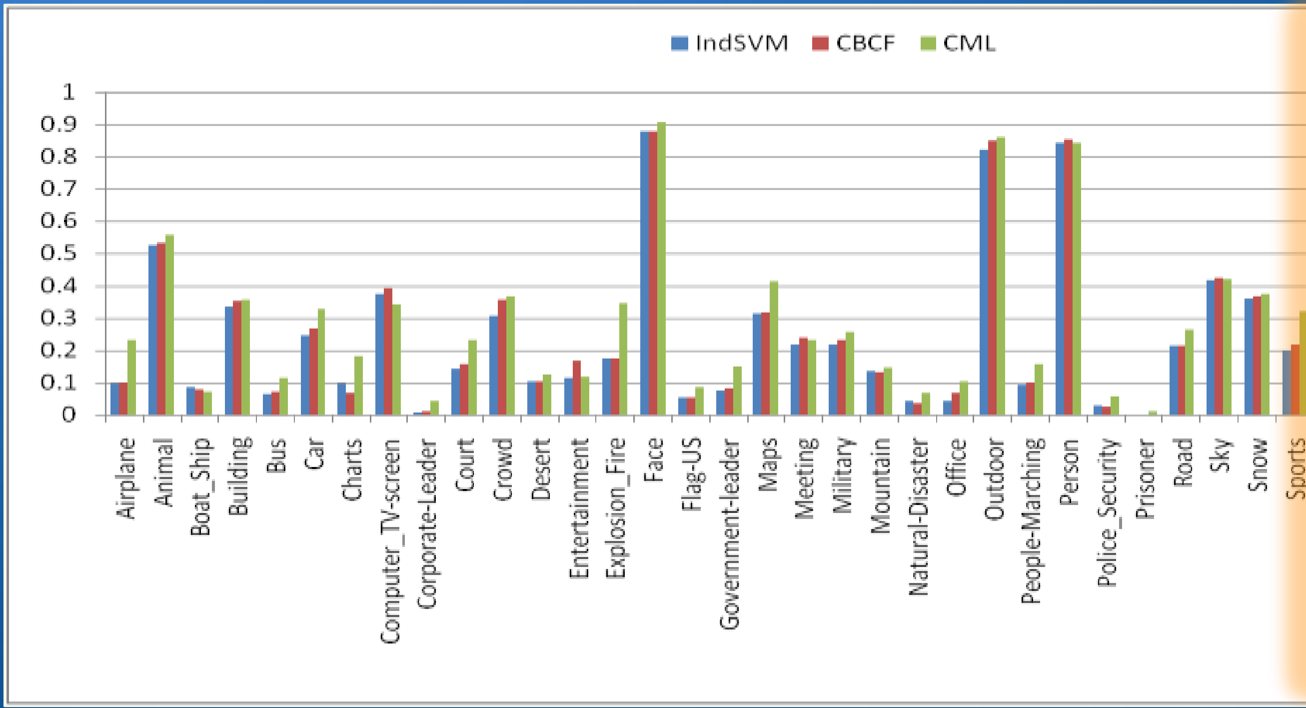
Correlative Multi-Label Video Annotation

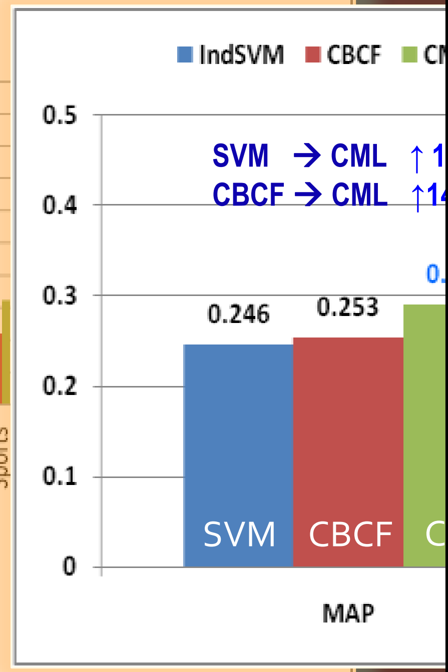
- Experiments
  - TRECVID 2005 dataset (170 hours)
  - 39 concepts (LSCOM-Lite)
  - Training (65%), Validation (16%), Testing (19%)



Correlative Multi-Label Video Annotation

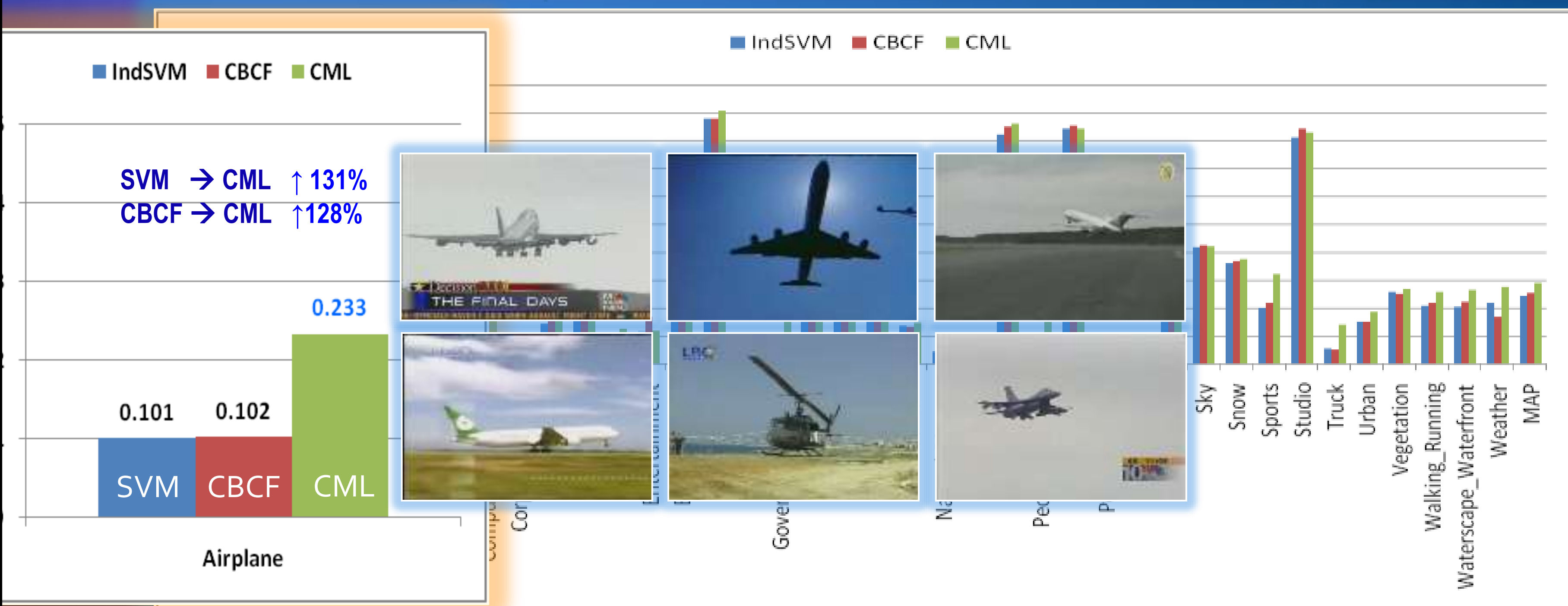
- Experiments
  - TRECVID 2005 dataset (170 hours)
  - 39 concepts (LSCOM-Lite)
  - Training (65%), Validation (16%), Testing (19%)
  - CML (MAP=0.290) improves IndSVM (MAP=0.246) 17% and CBCF (MAP=0.253) 14%





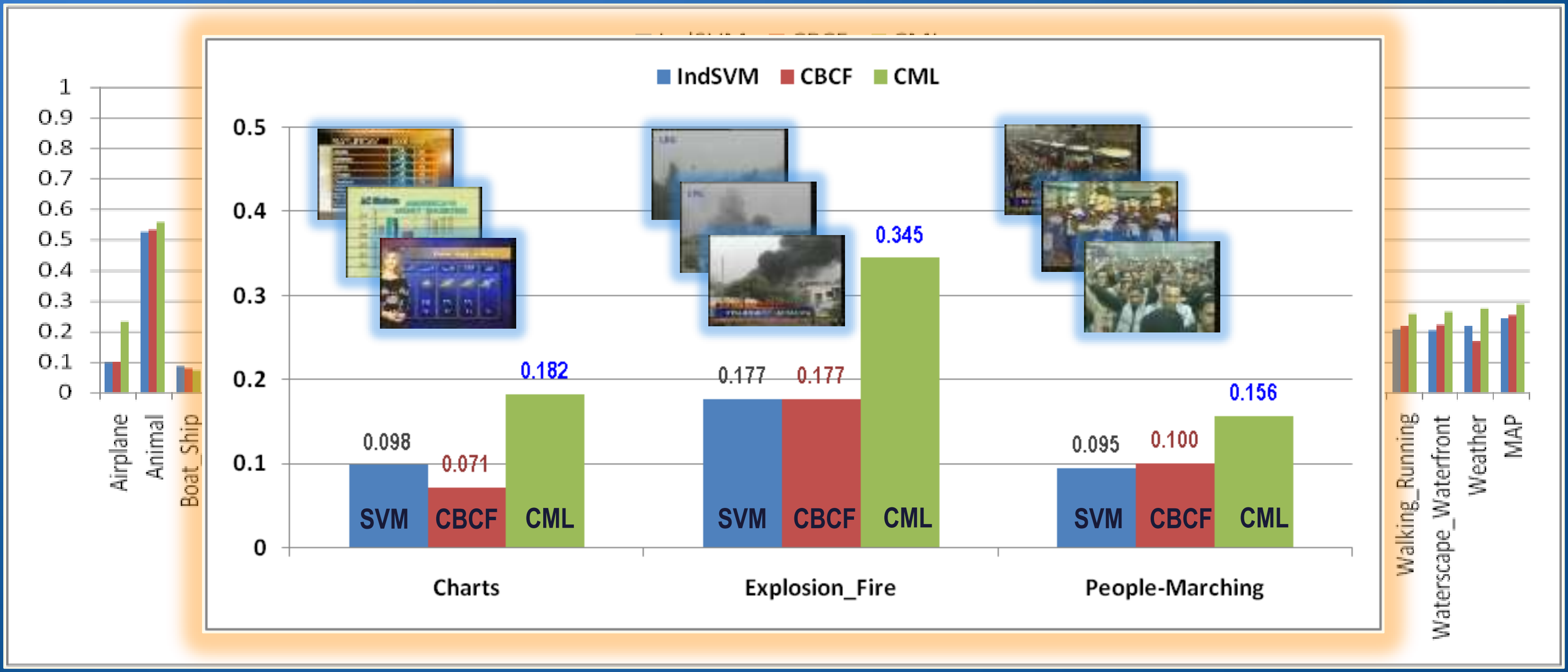
Correlative Multi-Label Video Annotation

- Experiments
  - TRECVID 2005 dataset (170 hours)
  - 39 concepts (LSCOM-Lite)
  - Training (65%), Validation (16%), Testing (19%)
  - CML (MAP=0.290) improves IndSVM (MAP=0.246) 17% and CBCF (MAP=0.253) 14%



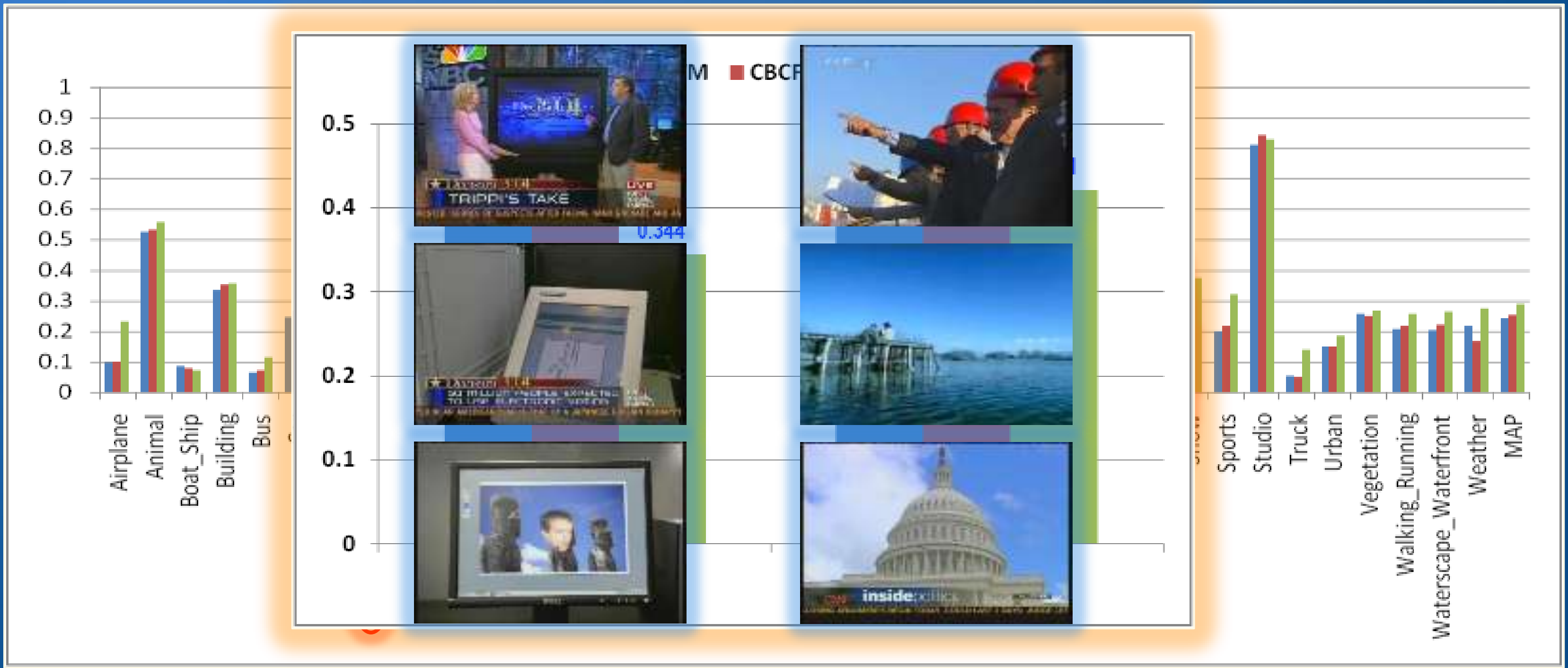
Correlative Multi-Label Video Annotation

- Experiments
  - TRECVID 2005 dataset (170 hours)
  - 39 concepts (LSCOM-Lite)
  - Training (65%), Validation (16%), Testing (19%)
  - CML (MAP=0.290) improves IndSVM (MAP=0.246) 17% and CBCF (MAP=0.253) 14%



Correlative Multi-Label Video Annotation

- Experiments
  - TRECVID 2005 dataset (170 hours)
  - 39 concepts (LSCOM-Lite)
  - Training (65%), Validation (16%), Testing (19%)
  - CML (MAP=0.290) improves IndSVM (MAP=0.246) 17% and CBCF (MAP=0.253) 14%



Image/Video Annotation by
Online Active Multi-Label Learning

Credit: Xian-Sheng Hua, Guo-Jun Qi. Online Multi-Label Active Annotation: Towards Large-Scale Content-Based Video Search. ACM Multimedia 2008. Vancouver, Canada, October 27 - November 1, 2008

bing

Learning Based Tagging

- Challenge - Cannot handle large-scale data/labels
  - Semantic gaps
  - Data/Semantic complexity
  - Computation cost
  - Difficult to scale-up

Where Is the Way Out?

- More training data / More informative training data

Internet Users

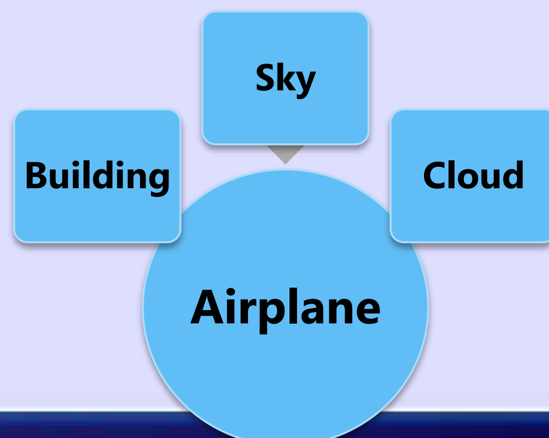
Active Learning

- More efficient training algorithm

Online Learning

- Better statistical model

Correlation Modeling



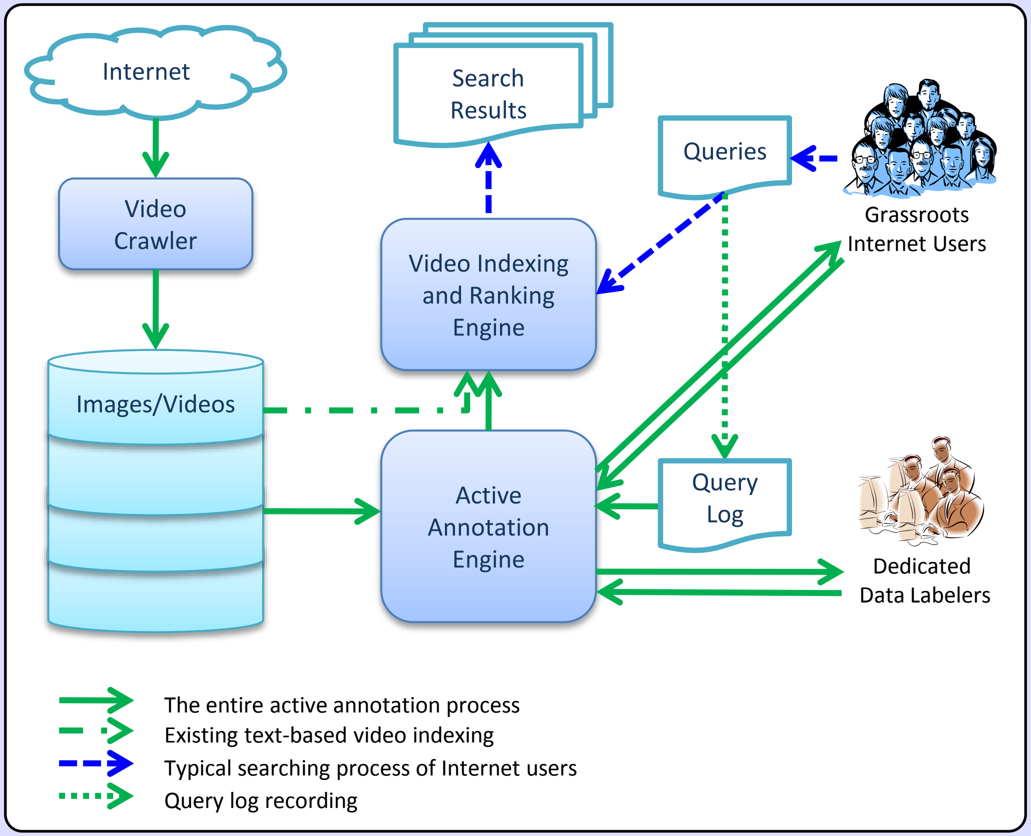
bing

A Promising Direction – Active Annotation

- Basic Idea
  - To bridge the semantic gap by leveraging human factors
  - To model complex semantics by mining correlations
- Basic methods
  - Using backend editors
    - Video indexing refinement by active learning
  - Leveraging grassroots
    - Actively present search results
    - Actively collect grassroots' contribution (**Game / Pay / reCAPTCHA / …**)
  - Modeling correlations
    - Label-Label
    - Label-Instance
    - Instance-Instance (spatial relation)
    - Multi-modality, multi-distance, …

An Attempt

- Online Active Annotation Framework



- Pay for labeling
- Online game
- reCAPTCHA
- Implicit approaches

Online Multi-Label Active Learning

Multi-Label Learning

↓

Multi-Label Active Learning

↓

Online Multi-Label Active Learning

Multi-Label Learning

Active Learning

Online Learning

Online Multi-Label Active Learning

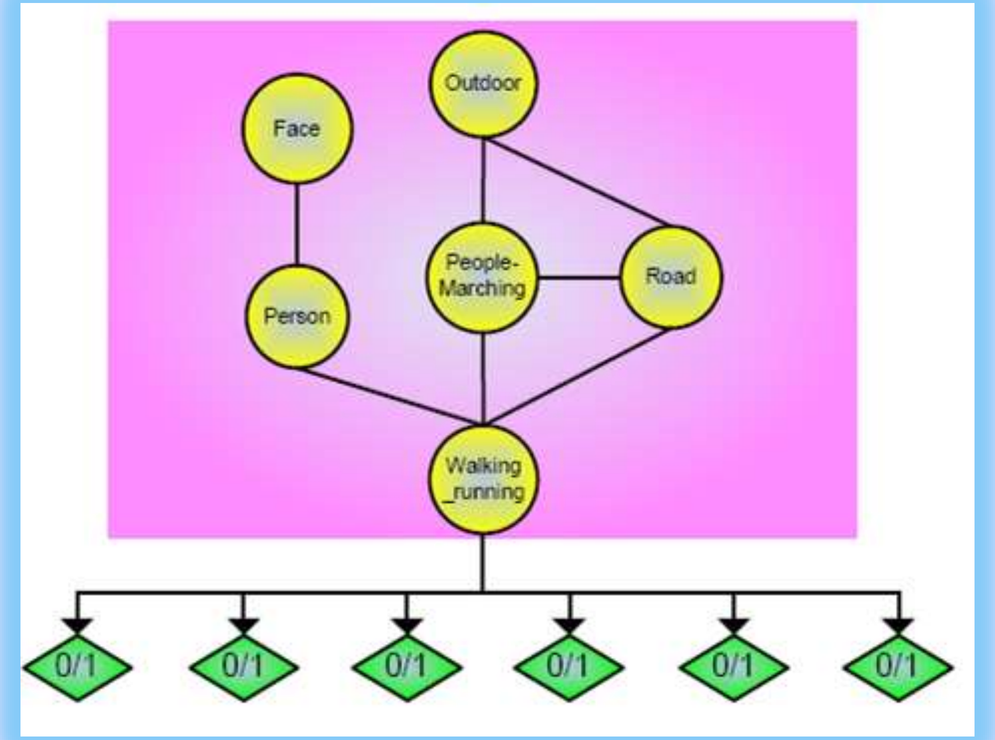
Multi-Label Learning

↓

Multi-Label Active Learning

↓

Online Multi-Label Active Learning



(ACM MM 07 – Best Paper Award)

Online Multi-Label Active Learning

Multi-Label Learning

↓

Multi-Label Active Learning

↓

Online Multi-Label Active Learning

Sky

Water

Mountain

Sands

Scenery

(Multi-Label Multi-Instance Learning - CVPR 08)

Online Multi-Label Active Learning

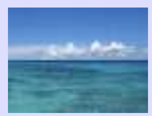
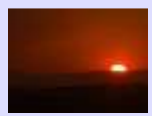
Multi-Label Learning

↓

Multi-Label Active Learning

↓

Online Multi-Label Active Learning

	Outdoor	Water	Sea	People	Crowd	Sky	Cloud
	Y	Y	Y	N	N	Y	Y
	Y	N	N	N	N	Y	N
	Y	N	N	Y	Y	N	N
	N	N	N	Y	N	N	N

(Single-Label Active Learning for Multi-Label Problems)

Online Multi-Label Active Learning

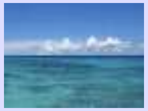
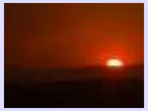
Multi-Label Learning

↓

Multi-Label Active Learning

↓

Online Multi-Label Active Learning

	Outdoor	Water	Sea	People	Crowd	Sky	Cloud
			Y		N	Y	
		N		N		Y	N
	Y	N			Y	N	
	N	N		Y	N		

(Multi-Label Active Learning - CVPR 08)

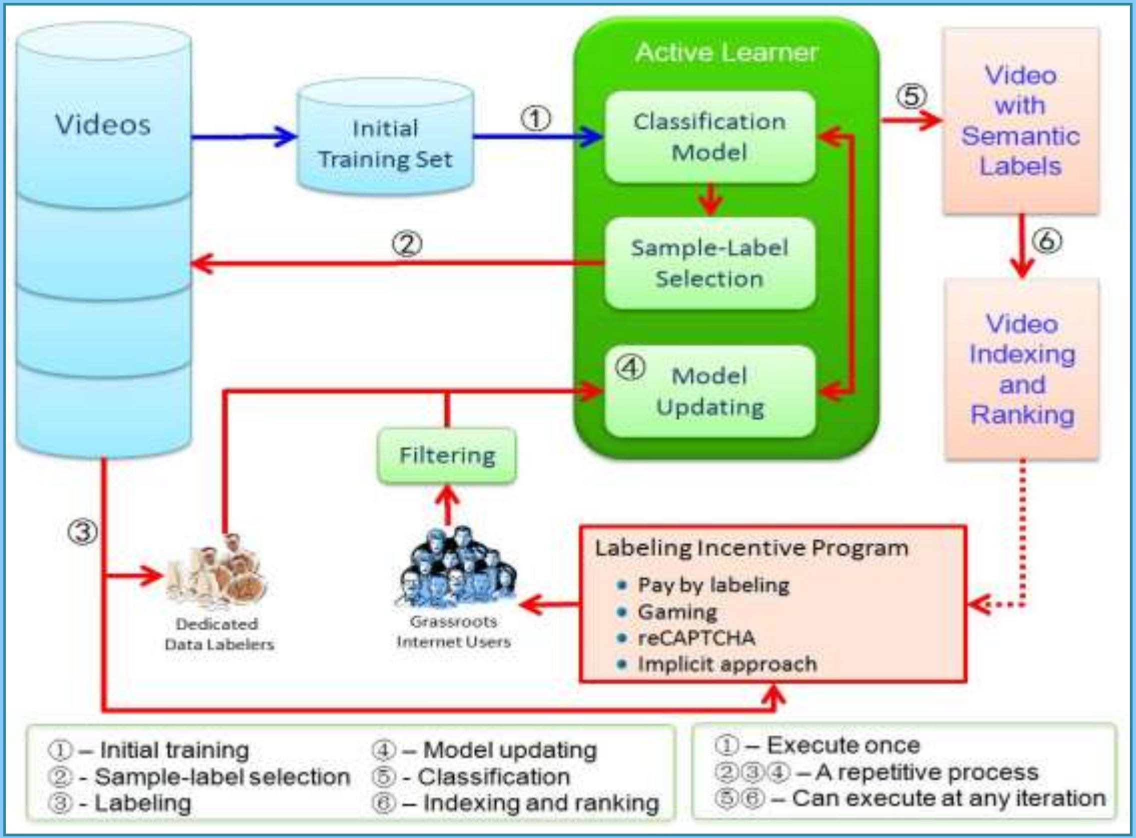
Active Annotation

Online Multi-Label Active Learning

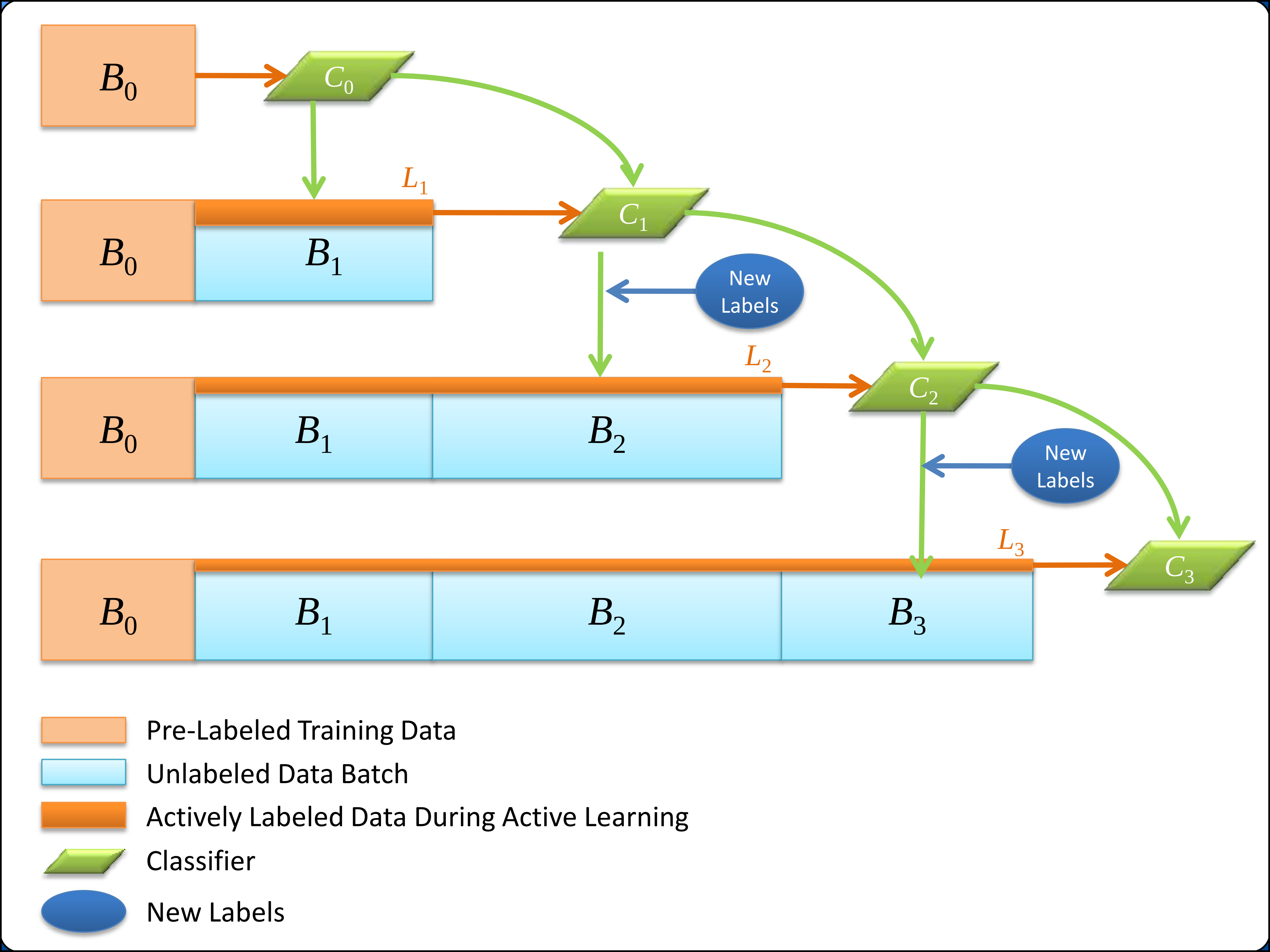
Multi-Label Learning

Multi-Label Active Learning

Online Multi-Label Active Learning



(ACM MM 08)

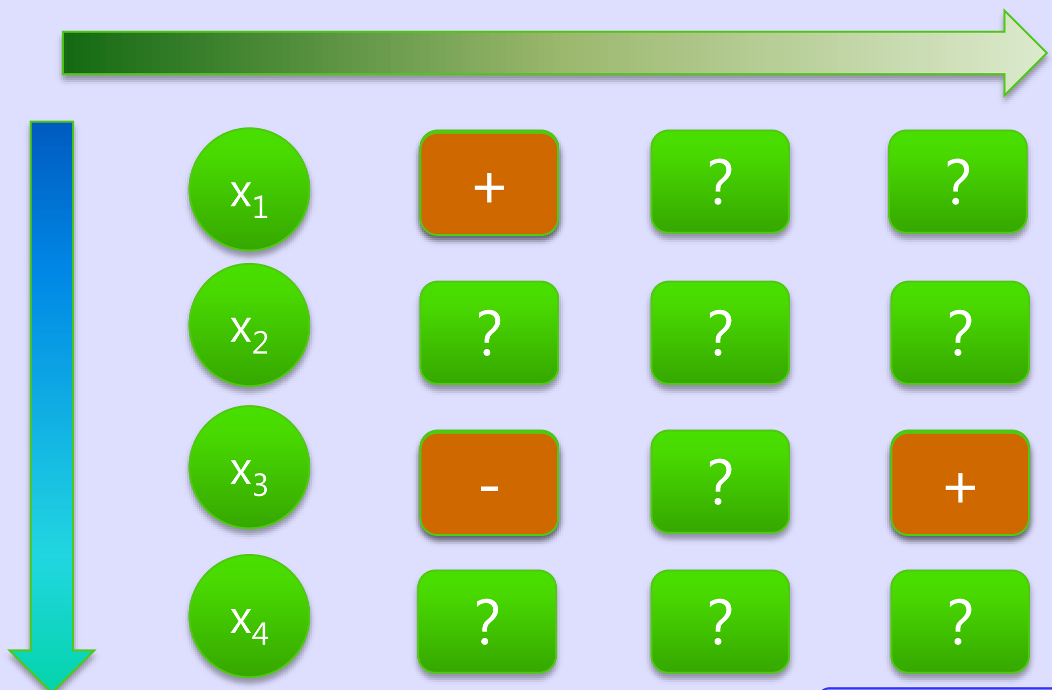


B_0
C_0
L_1
B_0
B_1
C_1
New Labels
L_2
B_0
B_1
B_2
C_2
New Labels
L_3
B_0
B_1
B_2
B_3
C_3
Pre-Labeled Training Data
Unlabeled Data Batch
Actively Labeled Data During Active Learning
Classifier
New Labels

bing

Multi-Label (Two-Dimensional) Active Learning

Label dimension



Two-Dimensional Active Learning

Sample dimension

One Dimensional Active Learning

Sample-label selection criterion

$$(\boldsymbol{x}_s^*, y_s^*) = \mathop{\arg\max}_{\boldsymbol{x}_s \in \mathbf{P}, y_s \in U(\boldsymbol{x}_s)} \left\{ \underbrace{H(y_s \mid y_{L(x_s)}, x_s)}_{\text{Self entropy}} + \underbrace{\sum_{i=1, i \neq s}^{m} MI(y_i; y_s \mid y_{L(x_s)}, x_s)}_{\text{Mutual information between labels}} \right\}$$

Online Multi-Label Learner

- Preserve existing knowledge
- Comply with the new coming examples

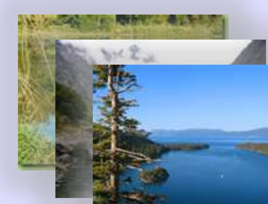
Old Multi-Labe Model

Preserve existing knowledge

New Multi-Labe Model

Comply with the new examples

New coming examples



Online Multi-Label Learner

Minimize KLD between old model and new one

$$\hat{P}^{\tau+1}(\boldsymbol{y}\,|\,\boldsymbol{x}) = \arg\min_{P^{\tau+1}} \left\langle D_{KL}(P^{\tau+1}(\boldsymbol{y}\,|\,\boldsymbol{x}) \,\|\, p^{\tau}(\boldsymbol{y}\,|\,\boldsymbol{x})) \right\rangle_{\tilde{P}}$$

Comply with multi-label constraints

$$s.t. \quad \left\langle y_i \right\rangle_{P^{\tau+1}} = \left\langle y_i \right\rangle_{\tilde{P}} + \eta_i, 1 \le i \le m$$

$$\left\langle y_i y_j \right\rangle_{P^{\tau+1}} = \left\langle y_i y_j \right\rangle_{\tilde{P}} + \theta_{ij}, 1 \le i < j \le m$$

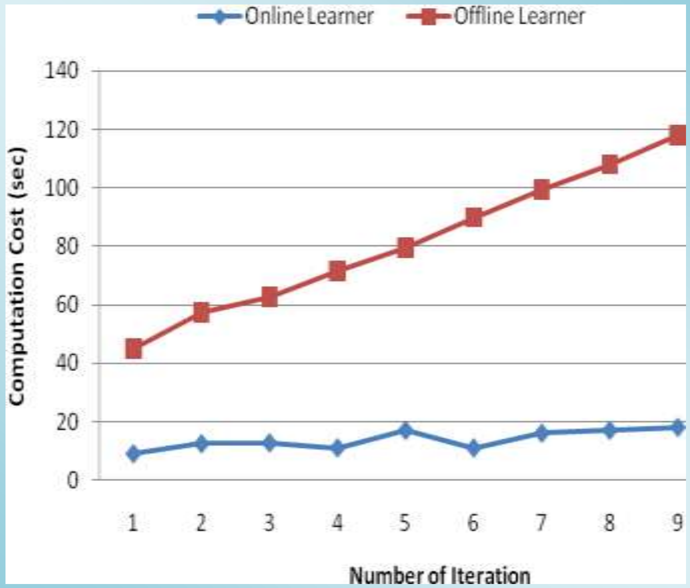
$$\left\langle y_i x_l \right\rangle_{P^{\tau+1}} = \left\langle y_i x_l \right\rangle_{\tilde{P}} + \varphi_{il}, 1 \le i \le m, 1 \le l \le d$$

$$\sum_{y} P^{\tau+1}(\boldsymbol{y}\,|\,\boldsymbol{x}) = 1$$

$$\sum_{i} \frac{\eta_i^2}{2\sigma_\eta^2 / n} + \sum_{i<j} \frac{\theta_{ij}^2}{2\sigma_\theta^2 / n} + \sum_{i,l} \frac{\phi_{il}^2}{2\sigma_\phi^2 / n} \le C$$

Experiments

- Online vs. Offline
  - On multi-label scene dataset: 2407 images, 6 labels
  - Performance is very close (F1 score differences are less than 0.001)

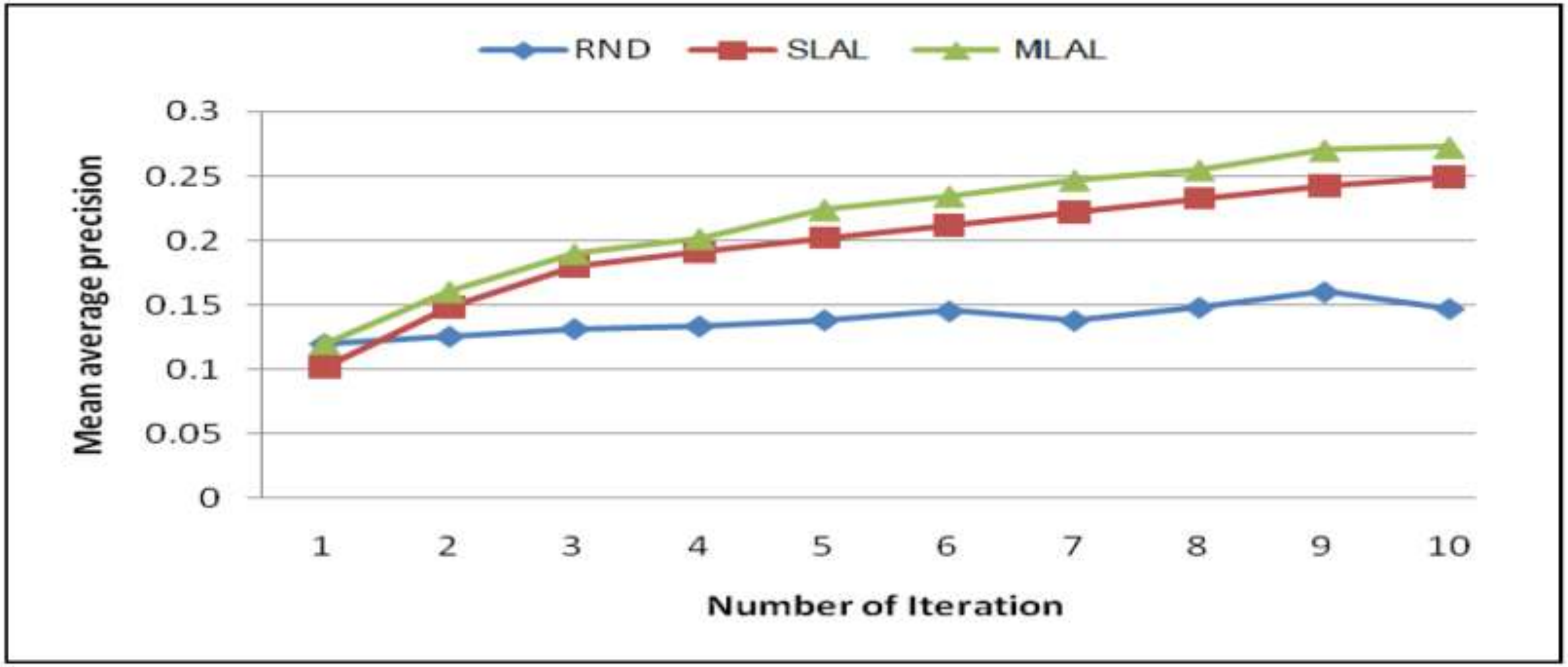


TRECVID Data

Iteration Number	Online Learner	Offline Learner
1	**0.12038**	0.11327
2	0.16071	**0.16159**
3	**0.19009**	0.18918
4	0.20179	**0.20894**
5	**0.22422**	0.22364
6	**0.23469**	0.22852
7	0.24708	**0.24795**
8	**0.25502**	0.25370
9	**0.27041**	0.26346
10	**0.27259**	0.26314

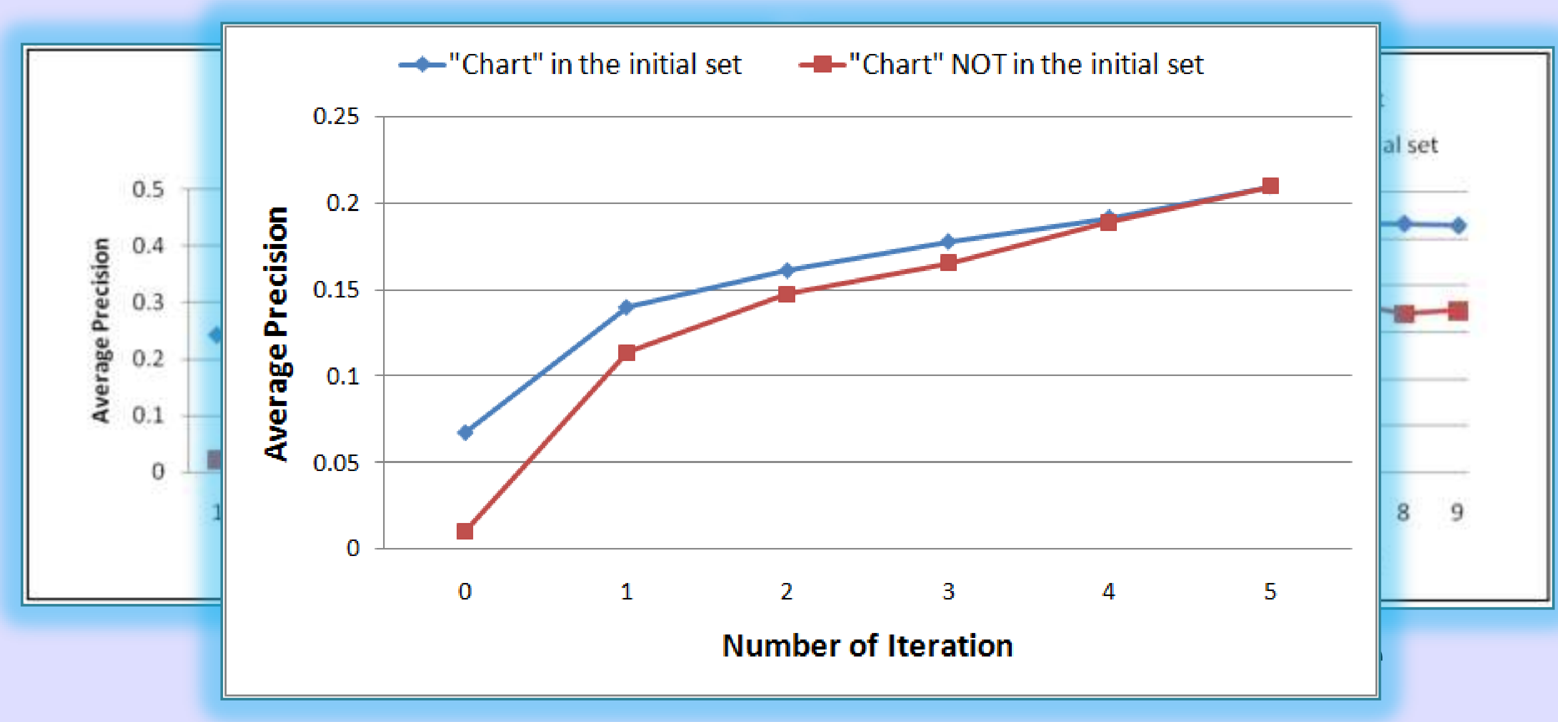
Experiments

- Single-Label vs. Multi-Label Active Learning
  - On TRECVID dataset: 2006 Dev set, 61901 shots, 39 concepts
  - Initial labeling: 10000, each step 39000 sample-label pairs



Experiments

- Adding new labels
  - On TRECVID dataset: 2006 Dev set, 61901 shots, 39 concepts
  - Initial labeling: 10000, each step 39000 sample-label pairs



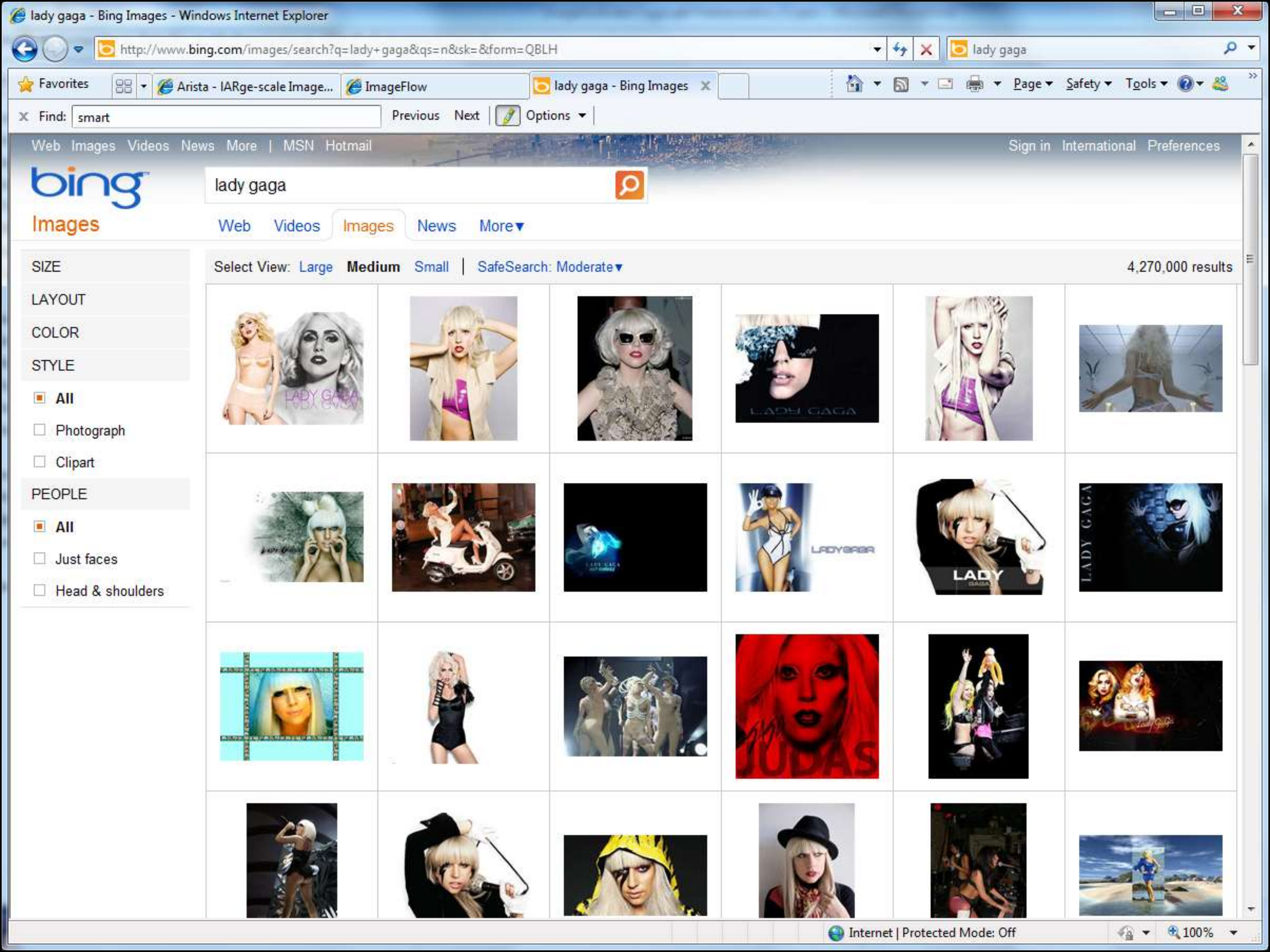
Active Annotation

- Key Challenge
  - Handling large scale content-based multimedia search
- Key Idea
  - Leveraging dedicated data labelers and large amount of grassroots Internet users to enable scalable multimedia semantic annotation
- Key Vehicle
  - Online multi-label active learning with incentive programs

Active Annotation is not only applicable for multimedia search, but may also be applicable for ranking, text search and other large-scale classification problems.

How Learning-Based Tagging is Used

- Demo
  - Content-based filters in Bing image search


lady gaga - Bing Images - Windows Internet Explorer
http://www.bing.com/images/search?q=lady+gaga&qs=n&sk=&form=QBLH
lady gaga
Favorites
Arista - lARge-scale Image...
ImageFlow
lady gaga - Bing Images
Page
Safety
Tools
Find: smart
Previous
Next
Options
Web Images Videos News More | MSN Hotmail
Sign in International Preferences
bing
Images
lady gaga
Web Videos Images News More
SIZE
LAYOUT
COLOR
STYLE
All
Photograph
Clipart
PEOPLE
All
Just faces
Head & shoulders
Select View: Large Medium Small | SafeSearch: Moderate
4,270,000 results
LADY GAGA
LADY
JUDAS
LADY GAGA
Internet | Protected Mode: Off
100%

Learning-Based Tagging

- Discussion
  - Variances are very high in real world data
  - Increasing training data helps, but still …
  - Exploit context/correlations helps
  - Though active learning and online learning helps, it is still not easy to scale up
  - Model-only approaches look like not the best solution for multimedia tagging

bing

Test

- What’re the limitations of learning-based tagging? Why?
- What can be used to improve learning-based tagging?

bing

Outline

Session	Time	Topic
0	09:30 – 09:40	Introduction
1	09:40 – 10:20	Learning-Based Tagging
2	10:20 – 11:40	Social Tagging and Tag Processing
		Including a break (10:45 – 11:00)
3	11:40 – 12:10	Data-Driven Web-Scale Tagging
4	12:10 – 12:30	Future Directions/QA

Outline

Session	Time	Topic
0	09:30 – 09:40	Introduction
1	09:40 – 10:20	Learning-Based Tagging
2	10:20 – 11:40	Social Tagging and Tag Processing
		Including a break (10:45 – 11:00)
3	11:40 – 12:10	Data-Driven Web-Scale Tagging
4	12:10 – 12:30	Future Directions/QA

Social Tagging and Tag Processing

- Introduction
- Tag Processing
  - Tag Ranking
  - Tag Recommendation
- Tag Analysis – Flickr Distance
- Discussion

bing

Social Media and the Associated Tags -
Towards Large-Scale Content-Based Multimedia Search

flickr

www, www2009, madrid, spain
w3c, Don Quixote, Don, Quixote
cervantes, Sancho, …

www2009, w3c, futuro, future, workshop, congreso
palacio, municipal, Madrid, consortium, consorcio
20, aniversario, España, Spain, Vinton, …

YouTube

bing

Social tags are good, but they are

- Noisy
- Ambiguous
- Incomplete
- No relevance information

Two directions to improve tag quality

- During tagging – Tag Recommendation
- After tagging – Tag Refinement/Ranking

Tag Ranking

Credit: Dong Liu, Xian-Sheng Hua, Linjun Yang, Meng Wang, Hong-Jiang Zhang. Tag Ranking. WWW 2009. Madrid, Spain.

The most relevant tag is NOT at the top position in

Social tags for online images are better than automatic annotation in terms of both scalability and accuracy.

This phenomenon is widespread on social media websites such as Flickr.

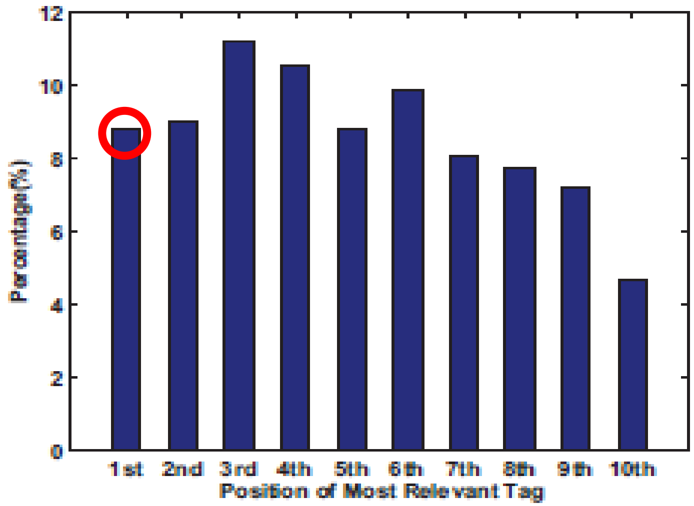
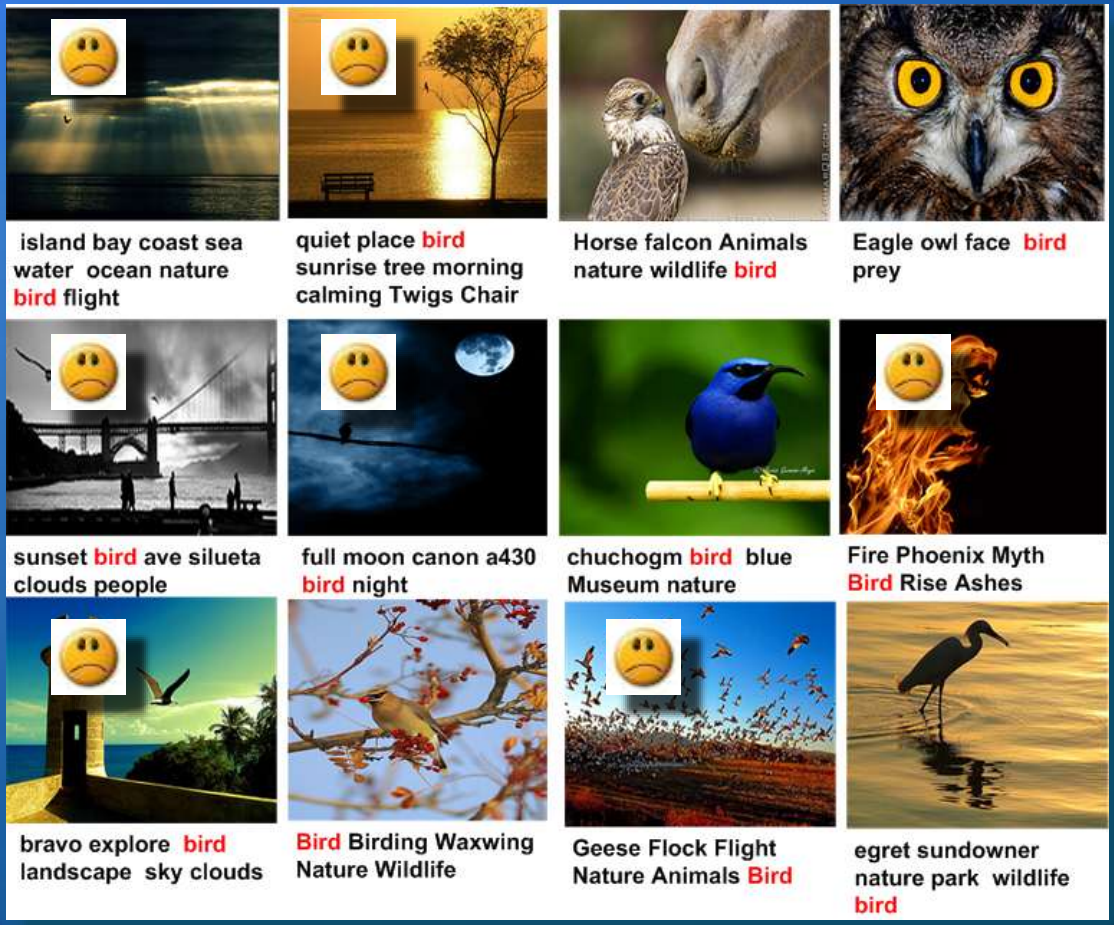


Figure 2: Percentage of images that have their most relevant tag at the n-th position in the associated tag list, where $n = 1, 2, 3 \ldots, 10$.

Only less than 10% images have their most relevant tag at the top position in their tag list.

This has significantly limited the performance of tag-based image search and other applications.

For example, when we search for “bird” on Flickr.



What we are going to do:

Rank the tags according to their relevance to the image.

bing

But how can we make it?

Automatically.

Image Ranking v.s. Tag Ranking

- Image Ranking
  - Order images according to the relevance (of the images) to the query term
- Tag Ranking
  - Order tags according to the relevance (of the tags) to the image

Image *Re*ranking v.s. Tag Ranking

- Image Reranking
  - Initial image ranking list → Improved ranking list
- Tag Ranking
  - Initial tag list (no order)→ Ranked tag list

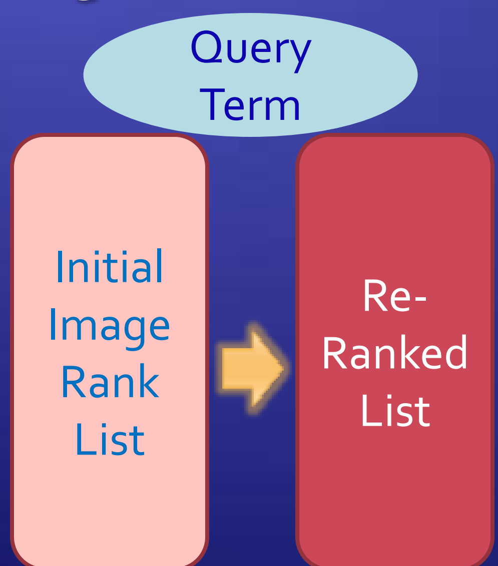
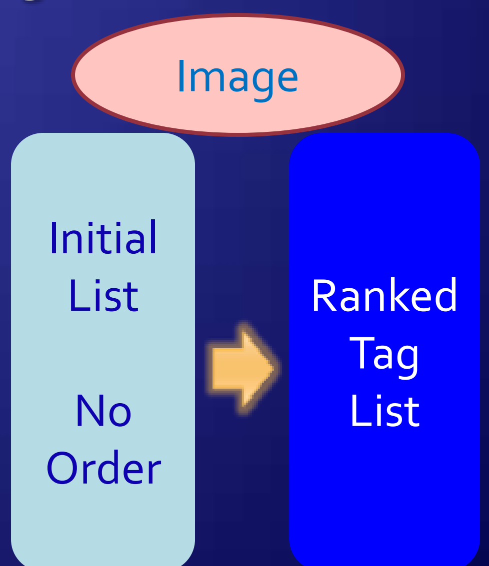


Image ReRanking



Tag Ranking

Can we borrow some idea from image reranking?

bing

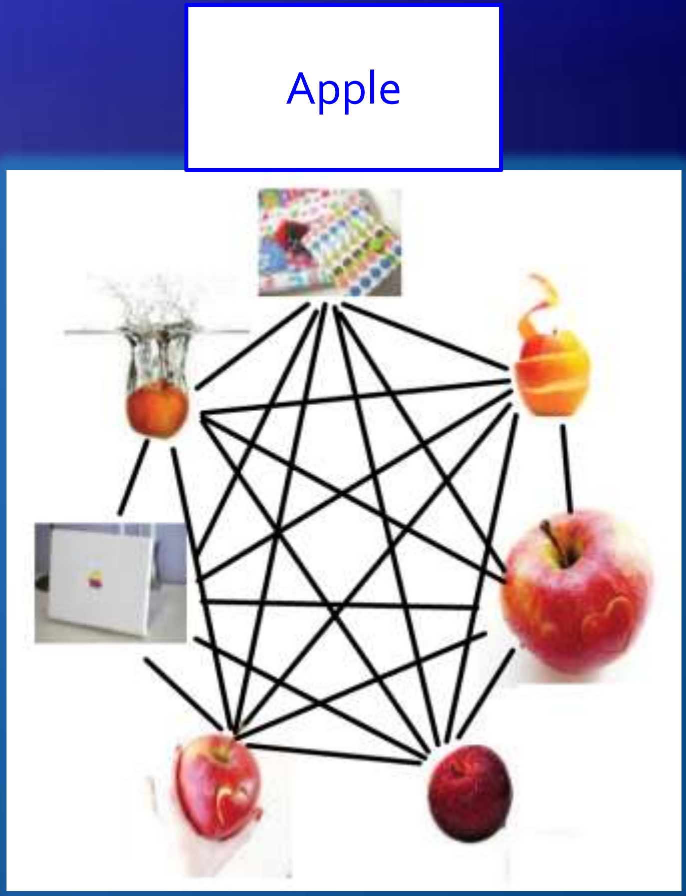
Basic Assumptions

- Image Reranking
  - Large image clusters should be promoted
  - Visually similar images should be ranked closely

bing

Typical Image ReRanking – Random Walk

- Graph construction
  - Images as nodes
  - Rank or ranking score of an image as the value of the node
  - Visual similarities of images are the edges
  - Transition probability between two nodes
- Graph Iteration
  - To refine the relevance scores step by step
  - With the help of the scores of the visually similar images

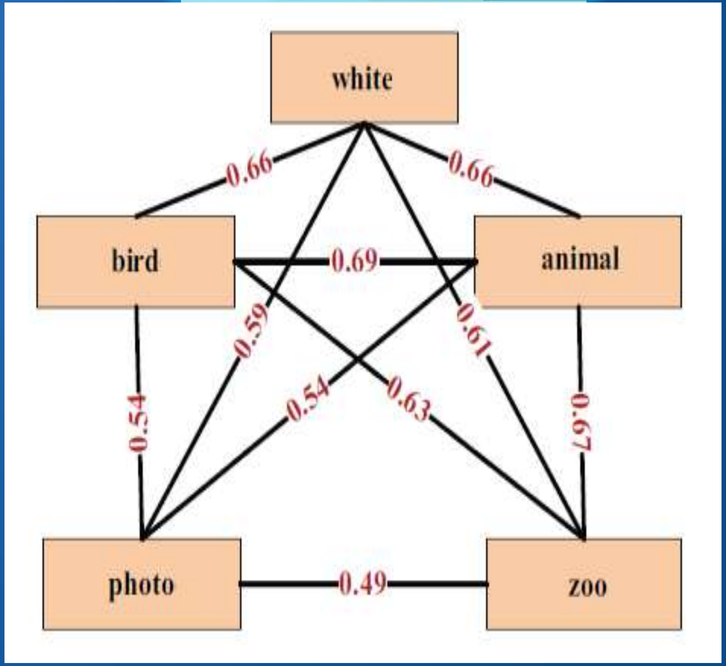


Basic Assumptions

- Image Reranking
  - Large image clusters should be promoted
  - Visually similar images should be ranked closely
- Tag Ranking
  - Large tag clusters should be promoted
  - Semantically close tags should be ranked closely

Tag Ranking (for each image)

- Graph construction
  - Tags as nodes
  - Rank of a tag as the value of the node
  - Semantic similarities of tags are the edges
  - Transition probability between two nodes



- Graph Iteration
  - To refine the relevance scores step by step
  - With the help of the scores of the semantically close tags

The problem is:

How can we calculate the similarity or distance of two tags?

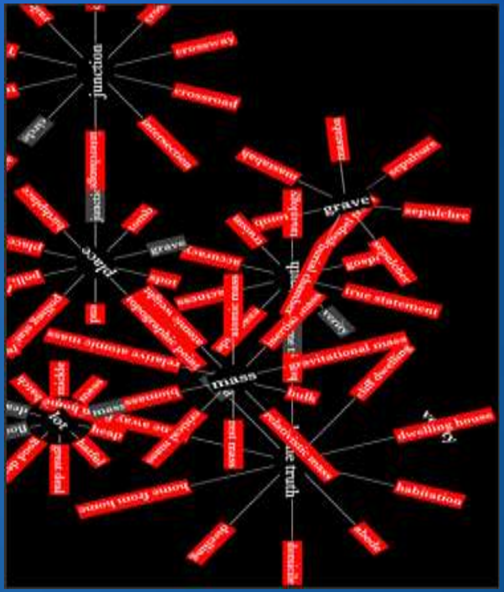
What We Can Use

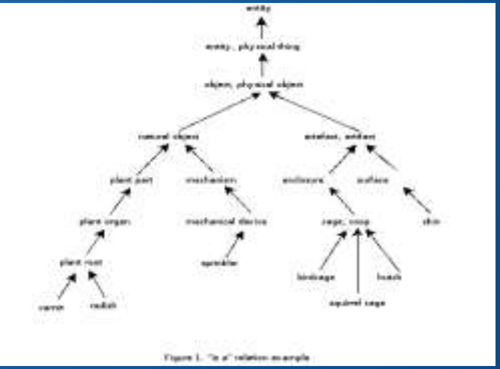
- WordNet distance
- Google distance

WordNet Distance

PRINCETON UNIVERSITY

- WordNet
  - 150,000 words
- WordNet Distance
  - Quite a few methods to get it in WordNet
  - Basic idea is to measure the length of the path between two words
- Pros and Cons
  - **Pros**: Built by human experts, so close to human perception
  - **Cons**: Coverage is limited and difficult to extend





Google Distance

UNIVERSITY OF AMSTERDAM

- Normalized Google Distance (NGD)
  - Reflects the concurrency of two words in Web documents
  - Defined as

$$NGD(x,y) = \frac{max(\log f(x), \log f(y)) - \log f(x,y)}{\log N - min(\log f(x), \log f(y))}$$

- Pros and Cons
  - **Pros**: Easy to get and huge coverage
  - **Cons**: Only reflects concurrency in textual documents. Not really concept distance (semantic relationship)

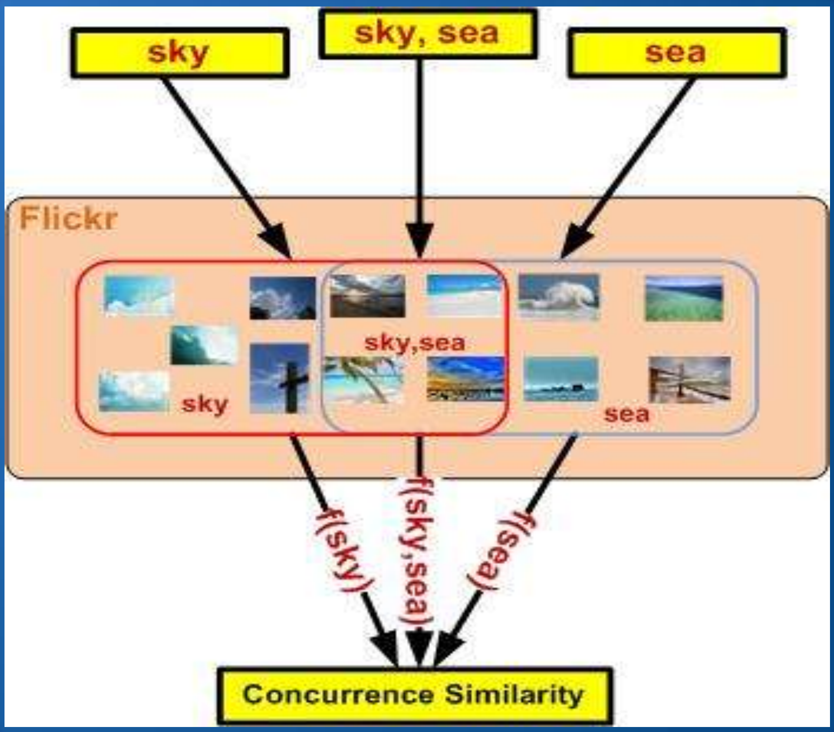
bing

What We Can Use

- WordNet distance
- Google distance
- Tag Concurrence Distance
- Tag2Image Distance

Tag Concurrence Distance

- Image Tag Concurrence Distance
  - Reflects the frequency of two tags occur in the same images
  - Based on the same idea of NGD but regards image tags as document



- Pros and Cons
  - Pros: Images are taken into account
  - Cons: Tags are not complete and noisy so visual concurrency is not well reflected. In addition, the distance is image independent

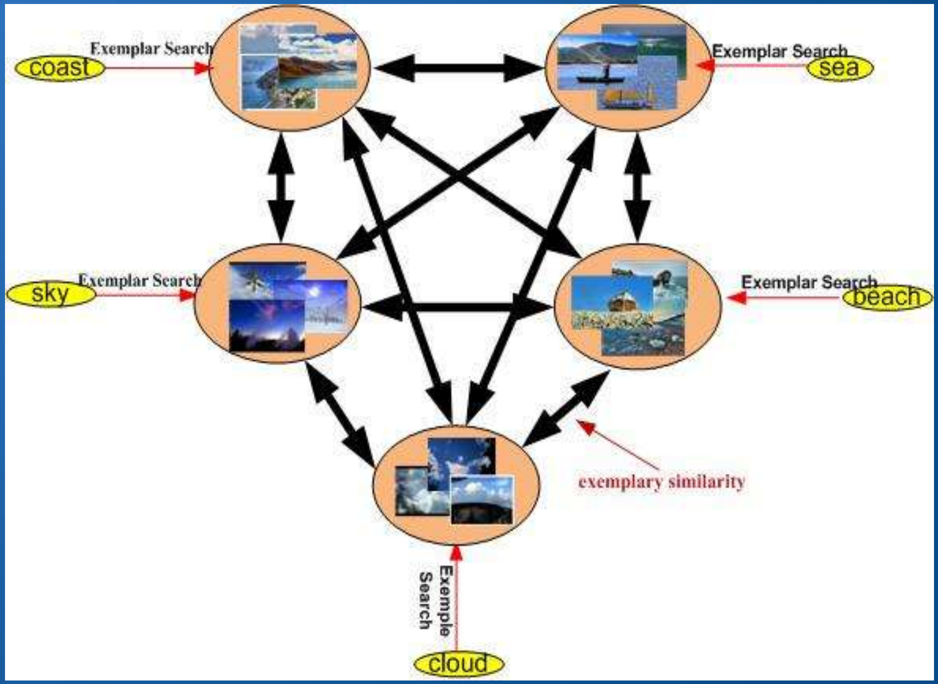
bing

dog , grass, tree, leaf

tree , grass, dog, leaf

Tag2Image Distance

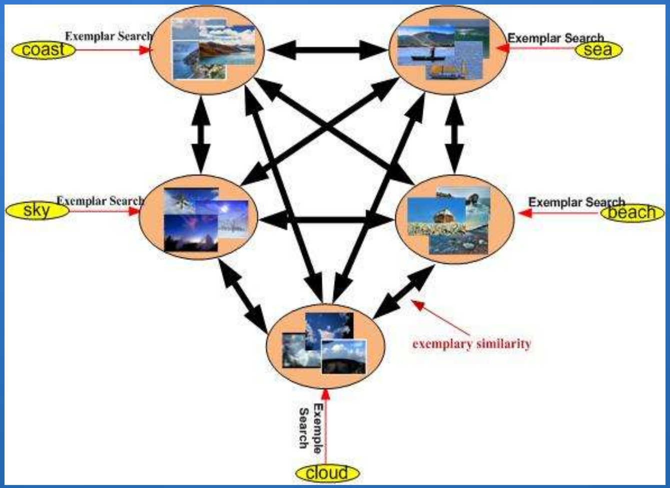
- Tag2Image
  - Find images with a particular tag
  - Keep those close to the target image (finding N-neighborhood)
  - Named as "Tag2Image Set"
- Tag2Image Distance
  - Distance between the corresponding tag2image sets of the two tags
- Pros and Cons
  - **Pros**: Images are taken into account and the distance is image dependent
  - **Cons**: Finding neighbors may be expensive

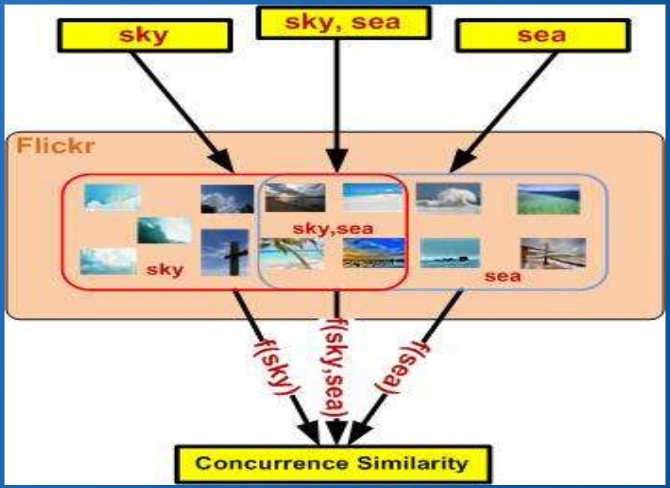


bing

Random Walk Based Tag Ranking

- Tag Graph Construction
  - Tag2Image similarity & Concurrence similarity



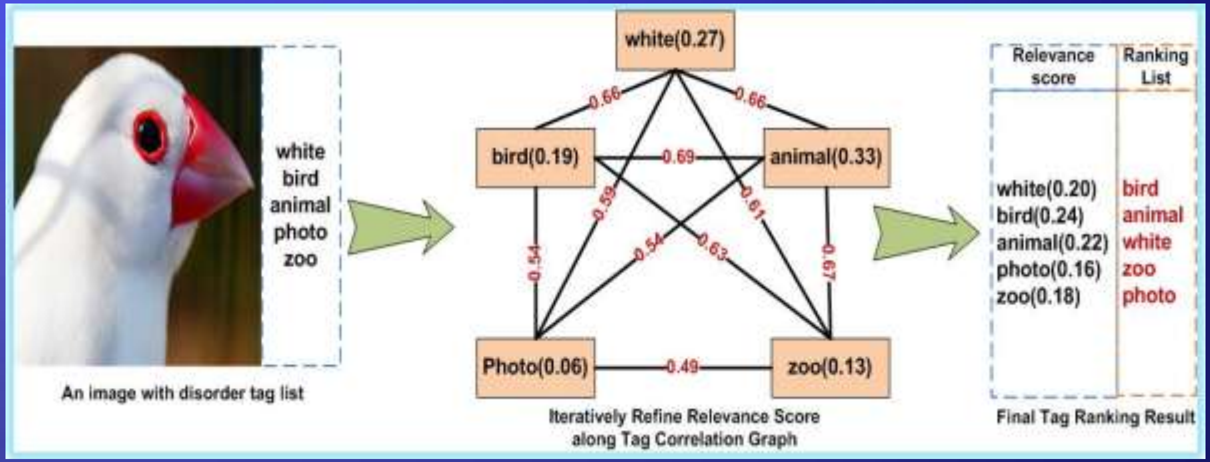


  - Combination $s_{ij} = s(t_i, t_j) = \lambda \cdot \varphi_e(t_i, t_j) + (1 - \lambda) \cdot \varphi_c(t_i, t_j)$

Visual similarity   Concurrence similarity

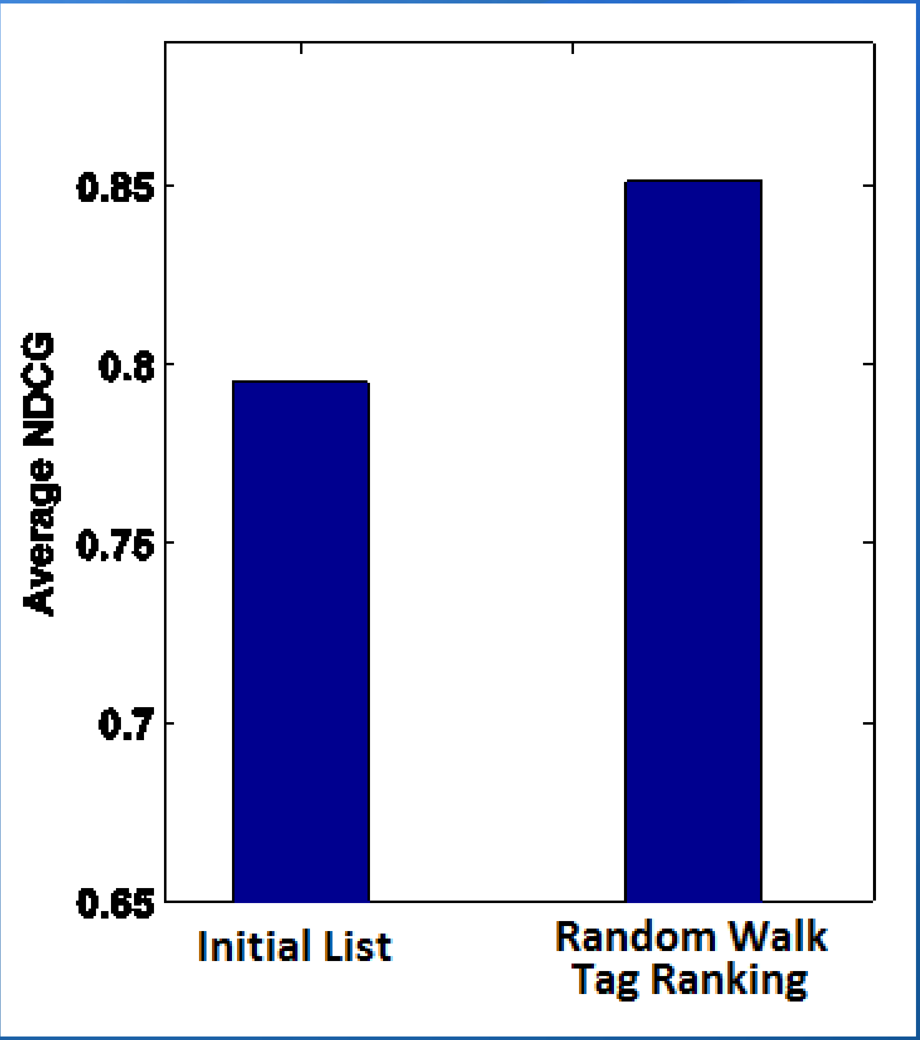
Random Walk Based Tag Ranking

- Random walk over tag graph



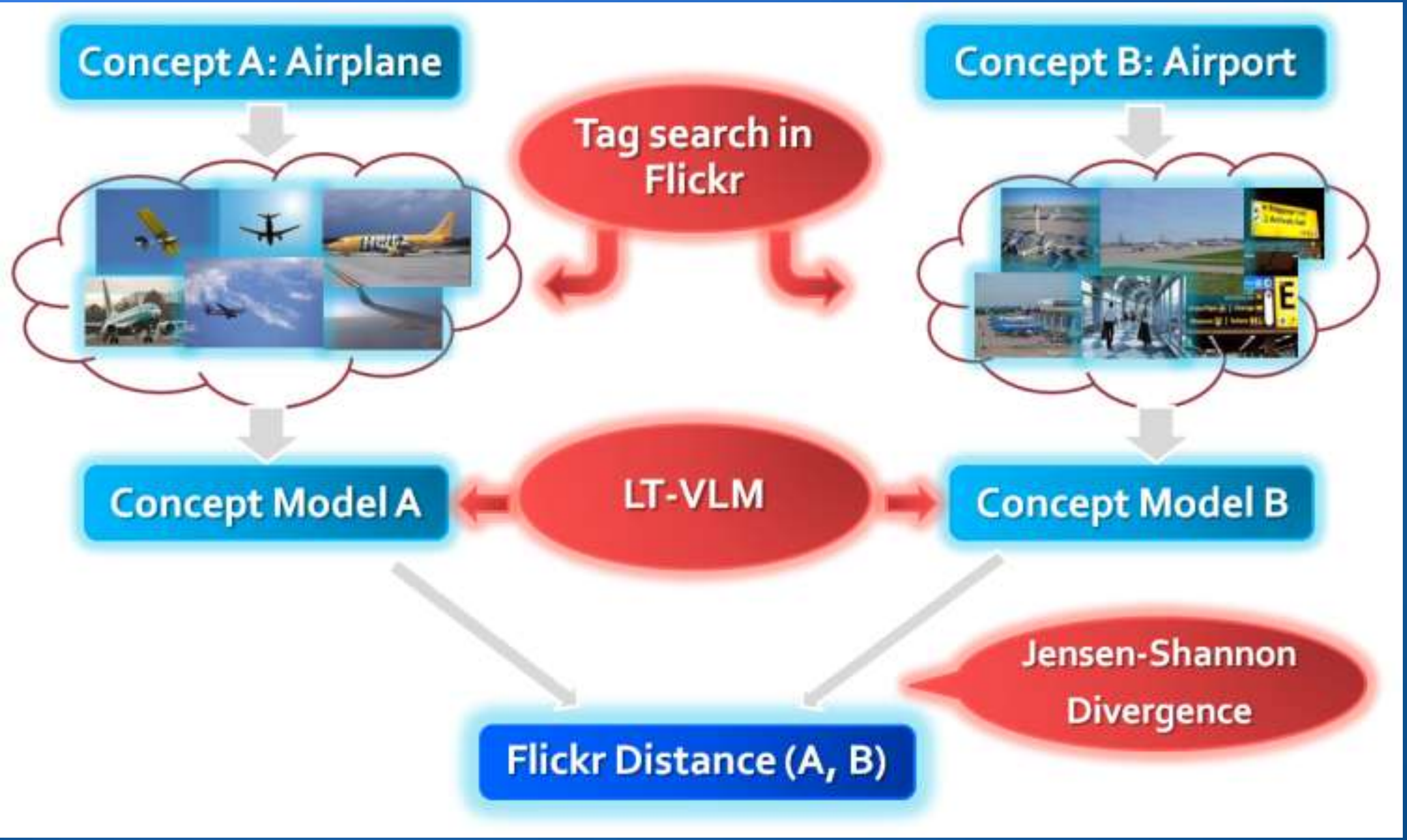
$$\mathbf{r}_k = \alpha \mathbf{P} \mathbf{r}_{k-1} + (1-\alpha)\mathbf{v}$$

- Transition matrix P denotes the row- normalized matrix of similarity matrix S.
- r is the vector of relevance score for each tag of the image.
- v is the vector of relevance score obtained by initial probabilistic tag relevance estimation.
- Alpha is the weighting parameter.


Average NDCG
0.85
0.8
0.75
0.7
0.65
Initial List
Random Walk
Tag Ranking

A Better Measure: Flickr Distance

- Lei Wu, Xian-Sheng Hua, et al. Flick Distance. *ACM Multimedia 2008* (**ACMMM 2008**). Vancouver, Canada, October 2008. (Best Paper Award Candidate)



bing

Is It Enough?

RECALL

Basic Assumptions

- Image Reranking
  - Large image clusters should be promoted
  - Visual similar images should be ranked closely
  - Initial ranks need to be kept as much as possible

Typically got from text-based ranking

- Tag Ranking
  - Large tag clusters should be promoted
  - Semantically close tags should be ranked closely
  - We don't have initial rank

How can we get it?

bing

Initial Relevance Estimation

- A possible estimation

$$s(t,x) = p(t|x)$$

- A better estimation (normalized by frequency)

$$s(t,x) = p(t|x)/p(t)$$

- After some calculation based on Bayesian Rule

$$s(t,x) = \frac{p(x|t)p(t)}{p(x)p(t)} = \frac{p(x|t)}{p(x)}$$

- It is about a particular image x, so $p(x)$ is a constant, therefore

$$s(t,x) \doteq p(x|t)$$

- What is it now?
  - Density of image x in the image space with tag t

bing

Initial Rank Estimation

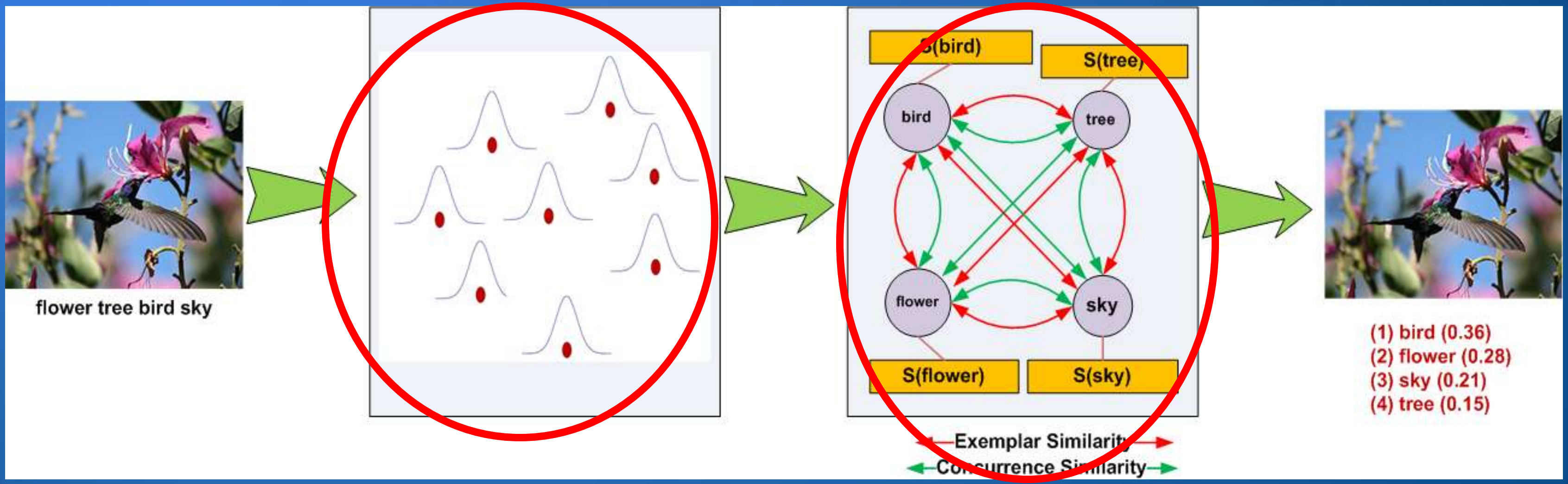
- Can be estimated by Kernel Density Estimation

$$s(t_i, x) = p(x|t_i) = \frac{1}{|X_i|} \sum_{x_k \in X_i} K_\sigma(x - x_k)$$

- An intuitive explanation
  - For image x, X_i can be regarded as x's friends with tag t_i
  - The sum of the similarities estimated based on Gaussian kernel can be regarded as the soft voting from the friends
  - So the initial relevance is actually estimated based on "collective intelligence" from its friend images

In Summary: Tag Ranking

- Two-step strategy



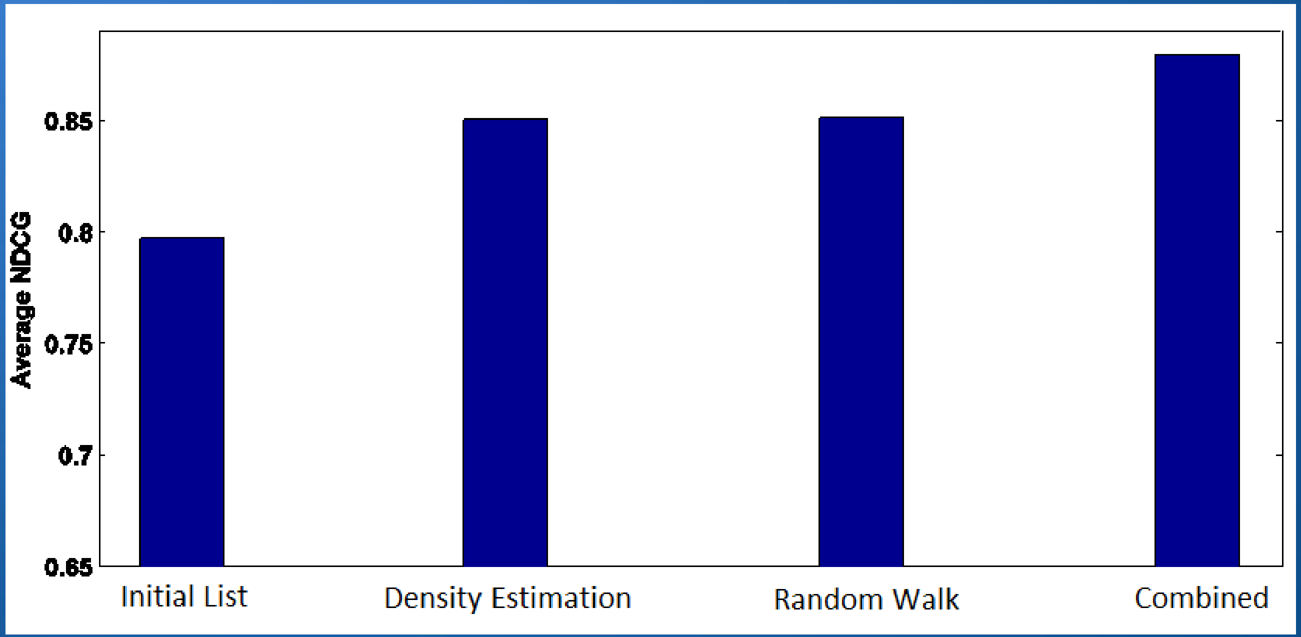
Probabilistic Tag Relevance Estimation

Random Walk Refinement

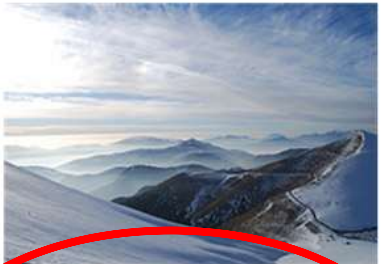
bing

Performance Evaluation

- In term of average NDCG
  - 50,000 Flickr images (to mine distance and estimate density)
  - 13,330 unique tags
  - 10,000 test images (each was labeled by 5 persons with five levels of relevance)



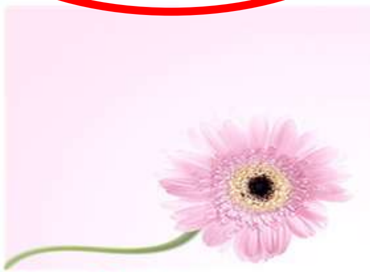
bing



Original Tag List:
blue winter sky white mountain snow photography gold nikon paradise view top greece drama
Ranked Tag List:
mountain sky white snow winter blue nikon photography view paradise gold greece top drama

Original Tag List:
ocean city summer brazil praia beach water architecture fantastic warm paradise desert great playa best resort arena
Ranked Tag List:
beach water ocean summer architecture fantastic paradise great resort playa city brazil best desert praia arena warm

Original Tag List:
blue pakistan portrait green bird nature yellow gold powershot karachi
Ranked Tag List:
bird nature blue green yellow portrait gold powershot pakistan karachi



Original Tag List:
pink light white flower green nature yellow spring flora gerbera
Ranked Tag List:
flower white pink nature light green yellow spring flora gerbera

Original Tag List:
sun sunlight animal cat kitten kitty gata gatto
Ranked Tag List:
cat kitty kitten animal sunlight sun gata gatto

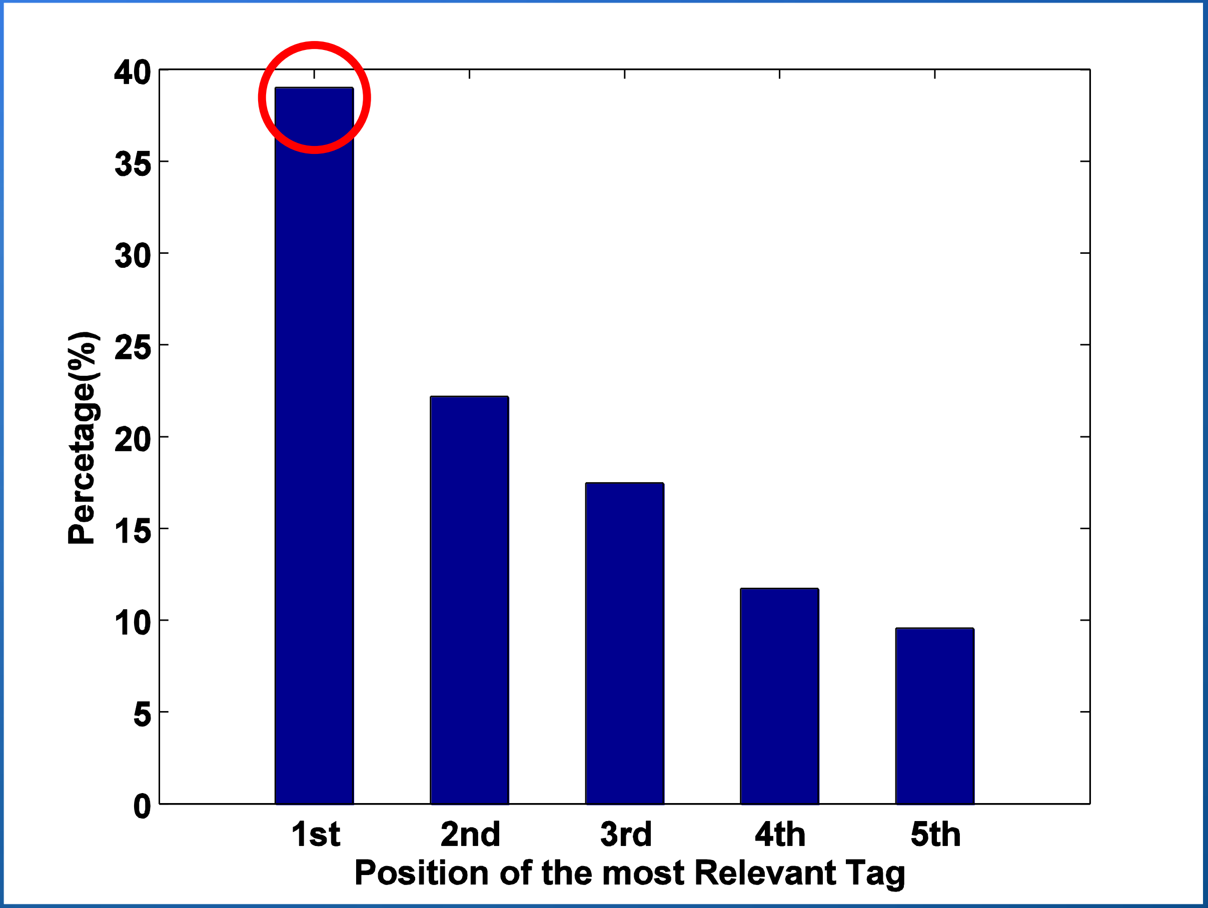
Original Tag List:
family wedding friends sunset red sea love beach silhouette nikon flickr day colours maldives
Ranked Tag List:
sunset sea red beach nikon silhouette maldives love colours flickr friends family day wedding

Original Tag List:
park morning mist holland tree bird water fog duck baum
Ranked Tag List:
tree water bird fog park mist morning duck holland baum

Original Tag List:
ocean travel blue sea water philippines adventure
Ranked Tag List:
sea water ocean blue travel philippines adventure

Original Tag List:
ferrari concept car auto automobile
Ranked Tag List:
automobile car auto ferrari concept

After tag ranking, almost 40% images have their most relevant tag appear at the top position in their tag list.



Application 1: Tag-based search

- Use tag position as relevance measure

$$r(x_i) = -\tau_i + 1/n_i$$

- Ranking result for query "water"

bing

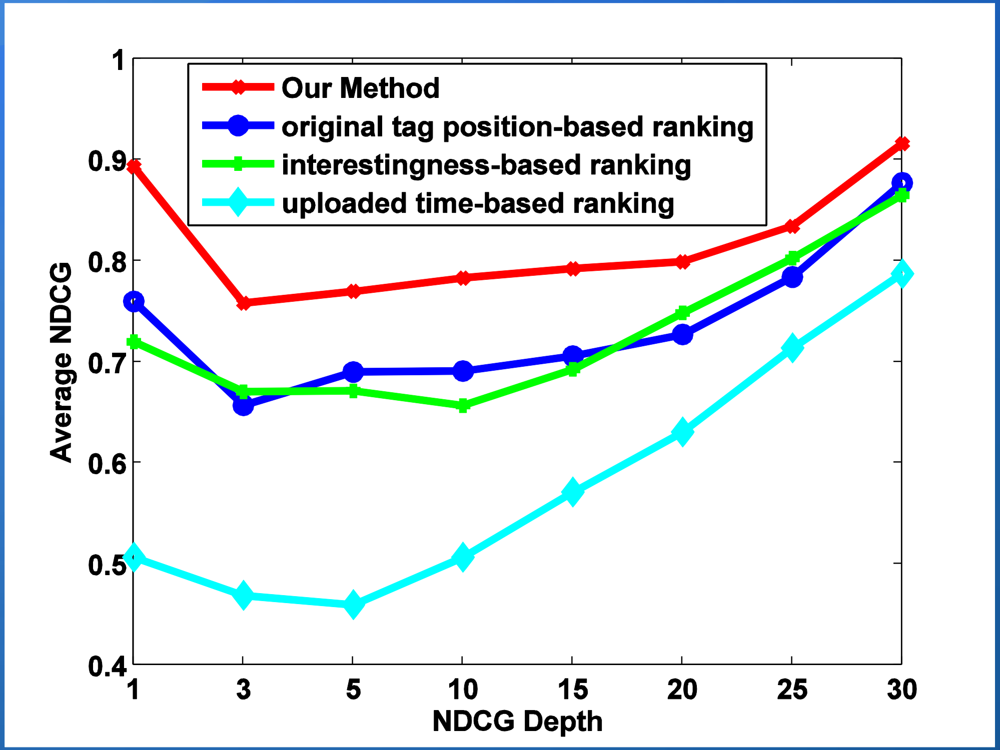
Application 1: Tag-based search

- Use tag position as relevance measure

$$r(x_i) = -\tau_i + 1/n_i$$

- Ranking result for query "bird"

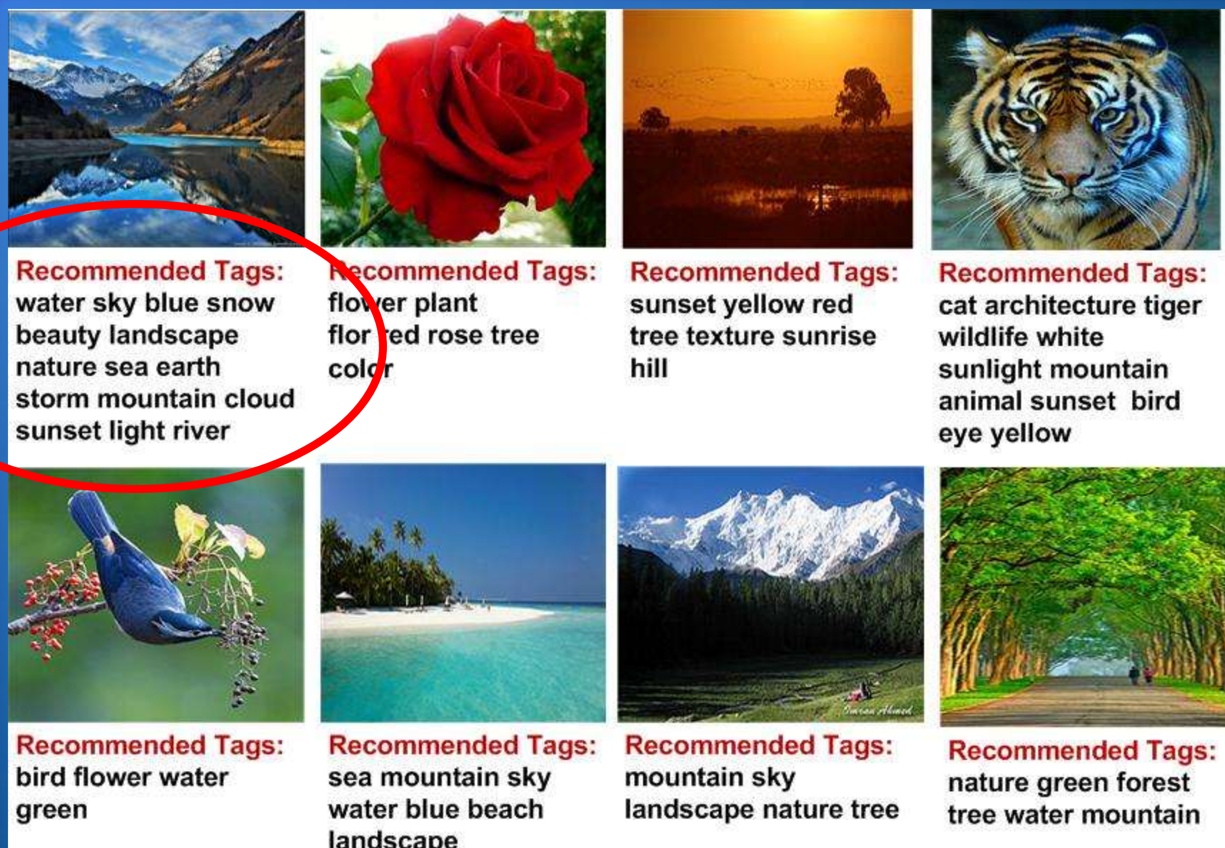
Performance of Tag-Based Search



Our tag position-based ranking strategy outperforms all other image ranking strategies on Flickr

Application 2: Auto Tagging

- Use top tags of similar images as tags for a new uploaded image



Performance of Auto Tagging

	Prec@1	Prec@5	Prec@all
Original(Baseline)	0.5858	0.4980	0.4980
Recommendation	0.7255	0.5799	0.5772
Improvement(%)	23.9	16.5	15.9

Using top tags after tag ranking to perform auto tagging

even outperforms human being

bing

Application 3: Group recommendation

- Use the top tags of an image as query keywords to search for its potentially suitable groups.

Performance of Group Recommendation

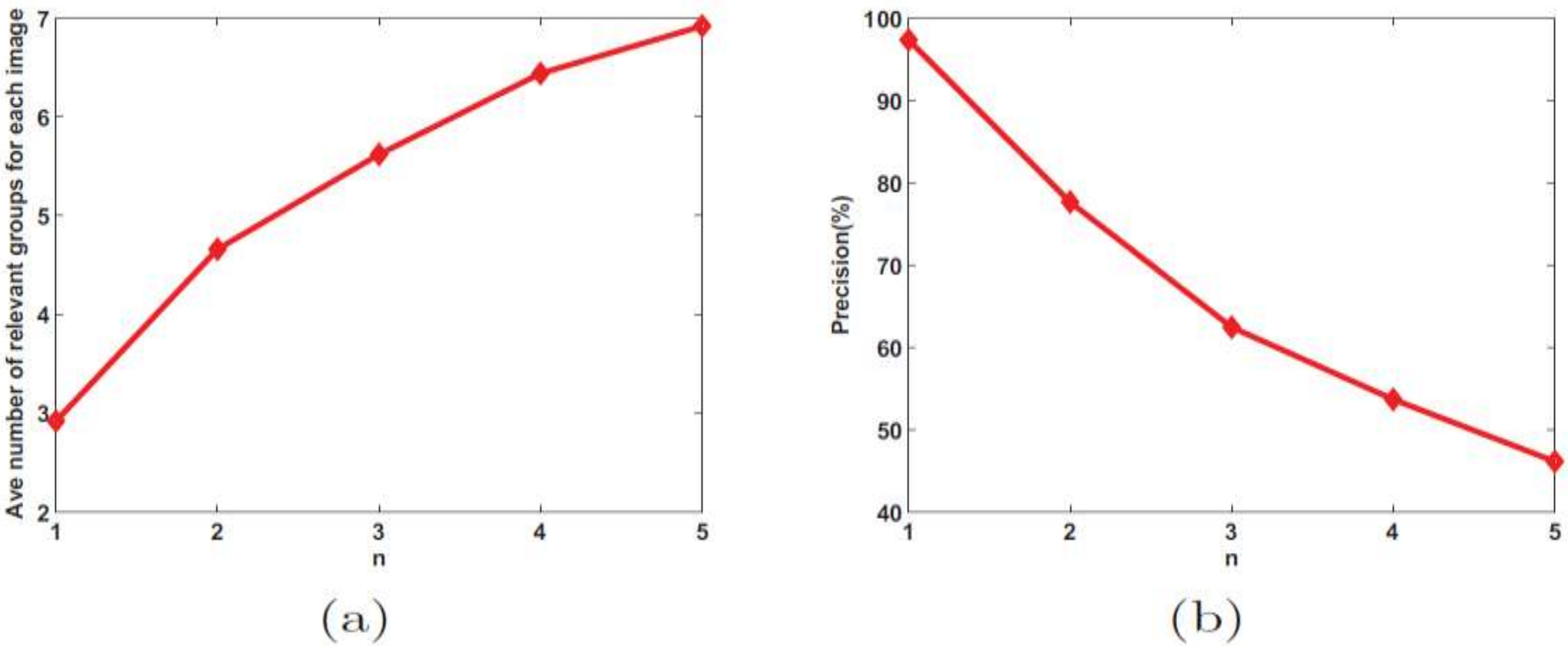


Figure 18: Performance of group recommendation with different n. (a) illustrates the average numbers of relevant recommended groups and (b) illustrates the recommendation precisions.

Tag ranking based group recommendation can help users better share their media content

Conclusion

- Initial tags are orderless in term of the relevance which limits the performance of tag-based search and other applications based on tags
- We propose a tag ranking strategy to solve this problem:
  - Density estimation to obtain initial rank scores
  - Refined by random walk based on image-dependant tag graph
- Tag ranking benefits a series of tag-based applications on social media websites

bing

Outline

Session	Time	Topic
0	09:30 – 09:40	Introduction
1	09:40 – 10:20	Learning-Based Tagging
2	10:20 – 11:40	Social Tagging and Tag Processing
		Including a break (10:45 – 11:00)
3	11:40 – 12:10	Data-Driven Web-Scale Tagging
4	12:10 – 12:30	Future Directions/QA

Learning To Tag

Credit: Lei Wu, Xian-Sheng Hua, et al. Learning To Tag. WWW 2009. Madrid, Spain.

Why Need Tag Recommendation?

Four Issues: Tags are

1. **Ambiguous**
2. **Incomplete**
3. **Noisy**
4. **No relevance**

Possible Tags:

Apple
Fruit
Red
Corporation
Logo
Products

Ambiguous

Incomplete

Tag Distribution on Flickr

Tags are noisy

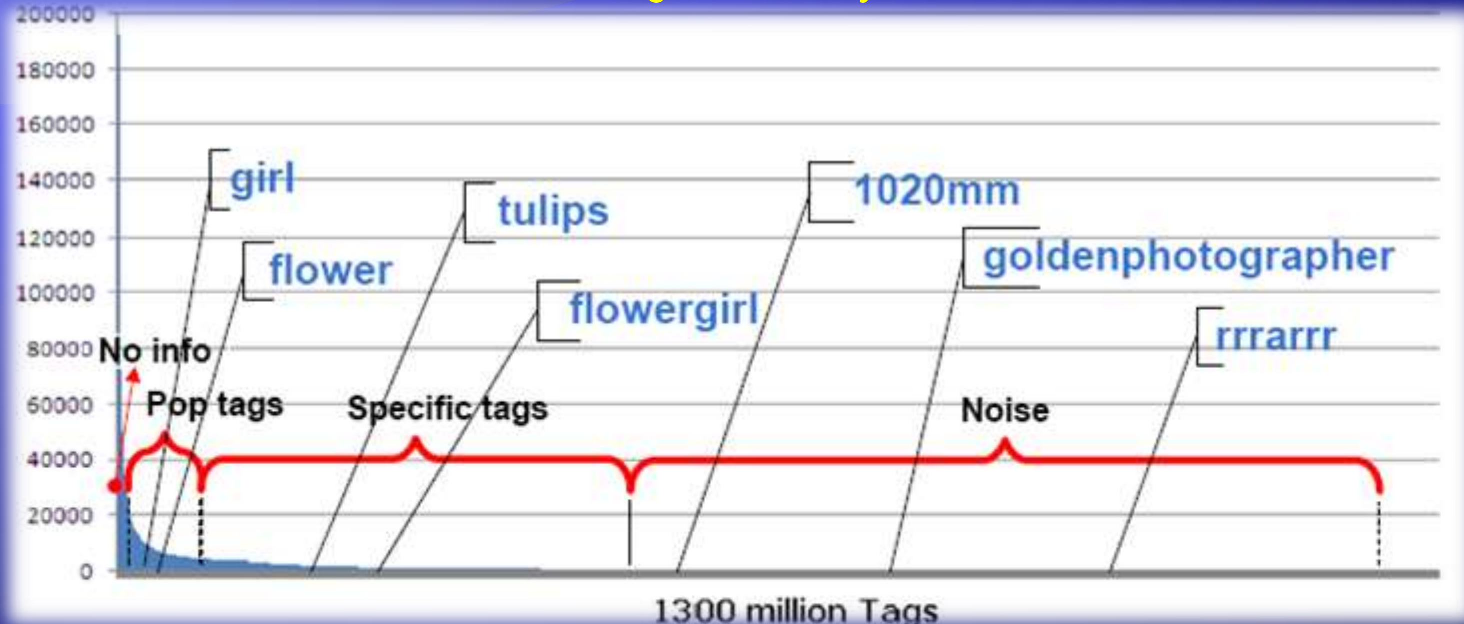


Figure 1: Tag distribution over a collection of 640 million images from Flickr.com. There are totally 1,300 million tags. Around 1% of the tags appearing more than 20,000 times, which contain little information. Around 5.82% of the tags have appeared more than 5,000 in the collection, which are considered as popular tags. 33.21% of the tags appears more than 50 and less than 5,000 times, which are defined as specific tags. 60% of the tags have appeared less than 50 times

Tag Recommendation

- What is Tag Recommendation
  - Given one or more initial tags for an image
  - Provide a list of possible tags automatically (ordered by relevance scores)
- Advantages of Tag Recommendation
  - Enable fast tagging
  - Enable high-quality tagging
    - Higher Relevance/Accuracy
    - Wider Coverage
    - Less noises

Straightforward Method

- Tag concurrence based recommendation

$$R^{a}_{tag}(t_i, t_j) = \frac{|t_i \bigcap t_j|}{|t_i|}$$

- Drawback:
  - Cannot deal with synonym
    - i.e. "table tennis" == "ping-pong"
  - Cannot deal with polysemy
    - i.e. "apple" fruit != "apple" logo
  - Cannot deal with meronymy
    - i.e., car vs. wheel
  - Target image is not taken into account (image independent)
    - Different images with the same initial tags will get the same recommended tag list

bing

Learning to Tag

- Three Features to Tag Correlations
  - Tag Concurrence
    - The same as previous work
  - Tag Content Correlation
    - To solve the first three issues (synonym, polysemy & meronymy)
  - Image-Conditioned Tag Correlation
    - To solve the fourth and second issue (image independent, & polysemy)

bing

Tag Content Correlation

Based on "Flickr Distance" or "PicNet Distance" – ACM MM 09 Best Paper Candidate

Can handle concurrency, synonym, polysemy & meronymy

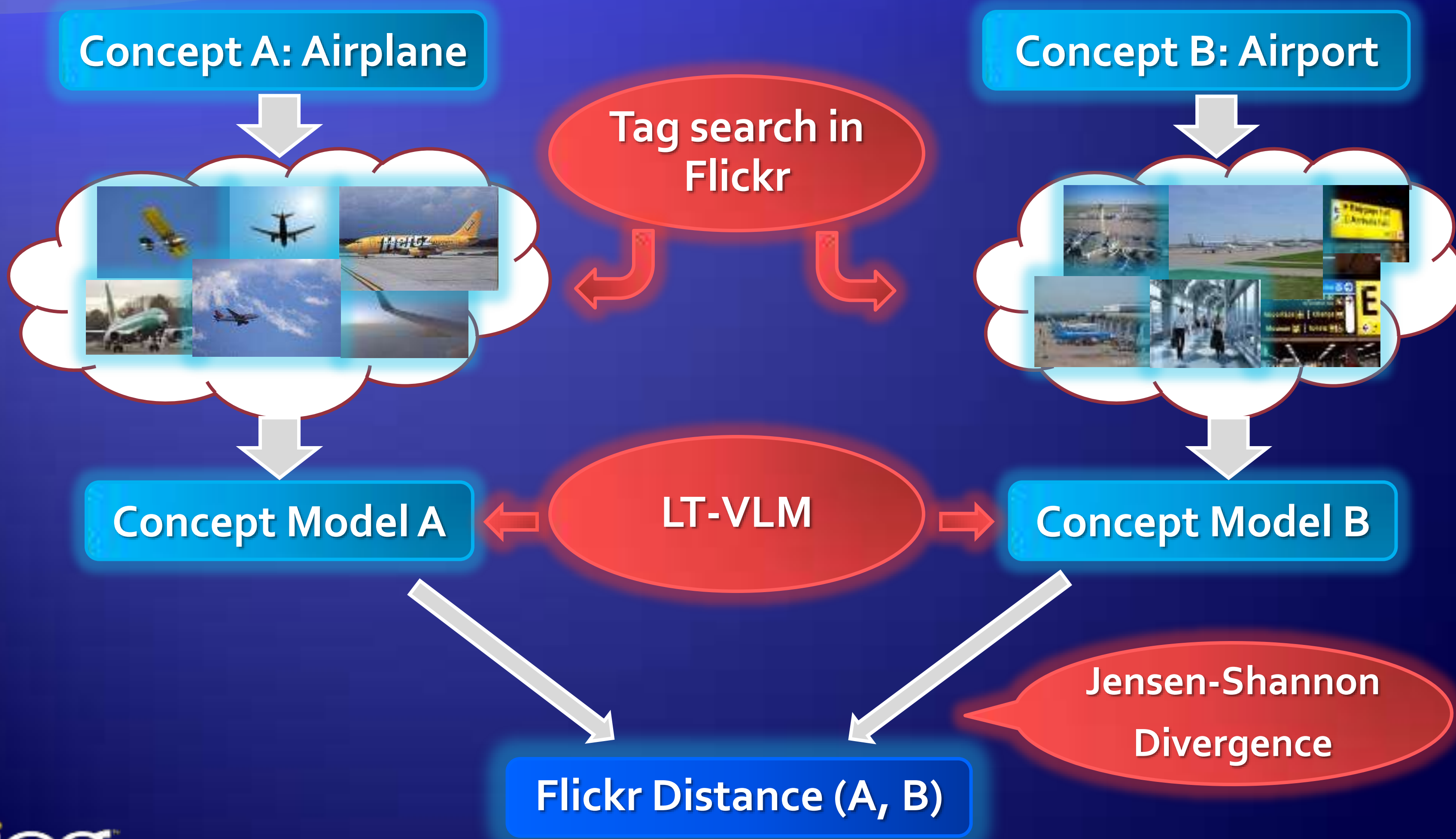


Image-Conditioned Tag Correlation

Based on Visual Language Model (VLM)

Visual Words

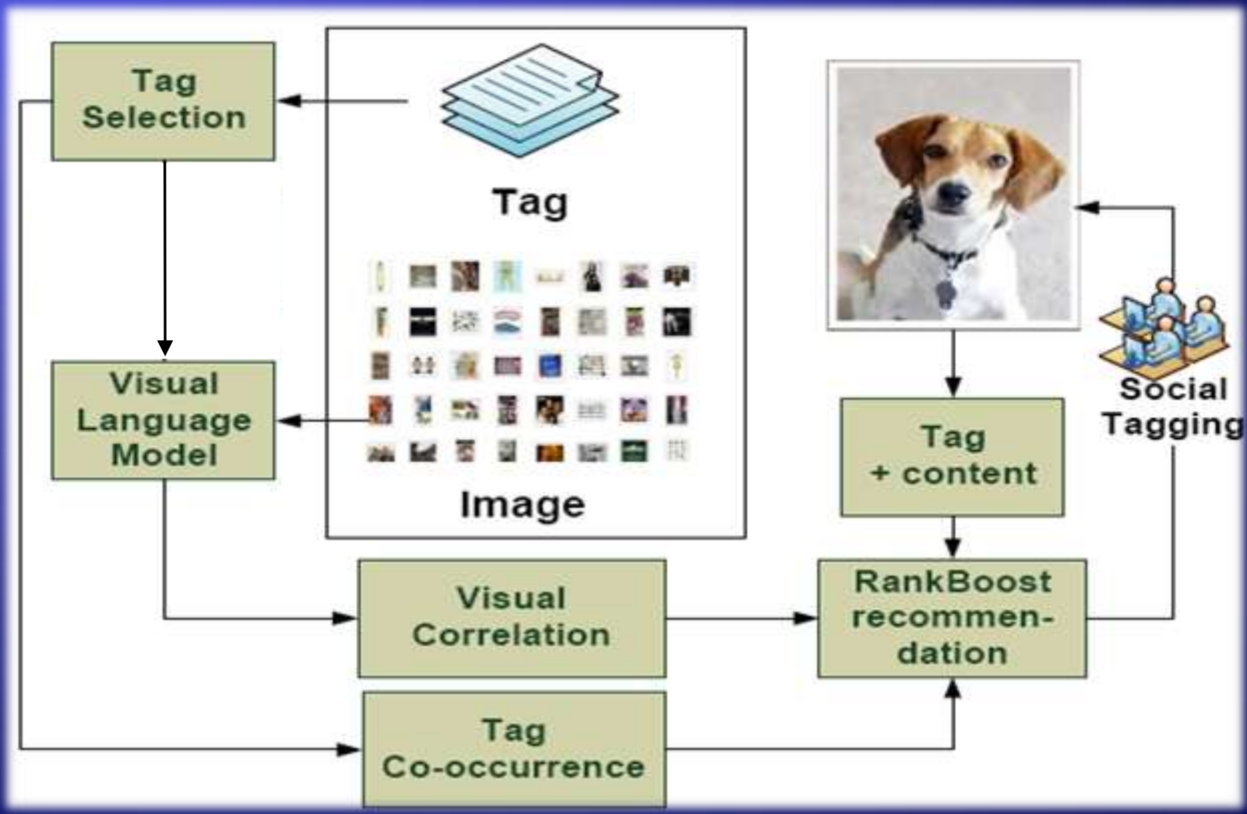
VLM of Tag 1 | VLM of Tag 2 | …… | VLM of Tag N

Projection 1 | Projection 2 | …… | Projection N

New Features for Tags (Image Dependant)

Learning To Tag

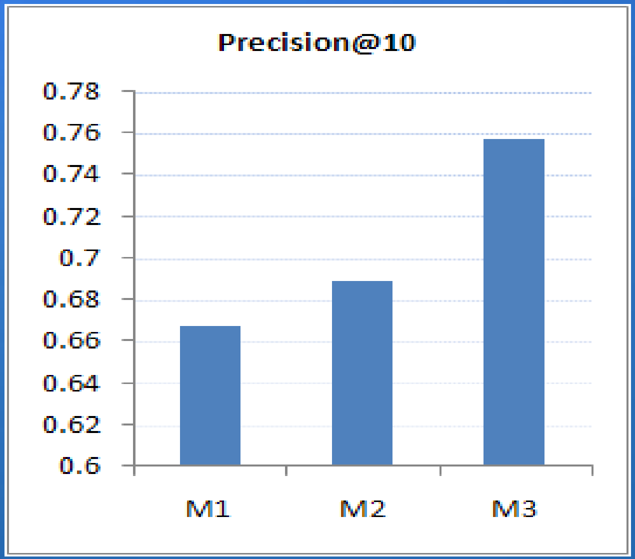
- Our Approach

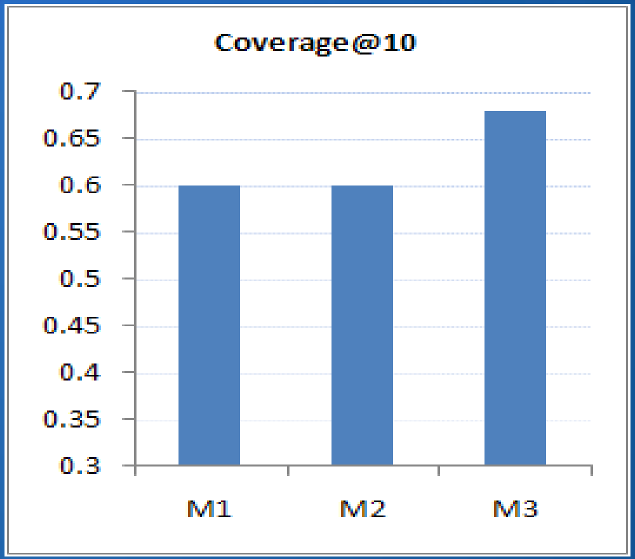


Results

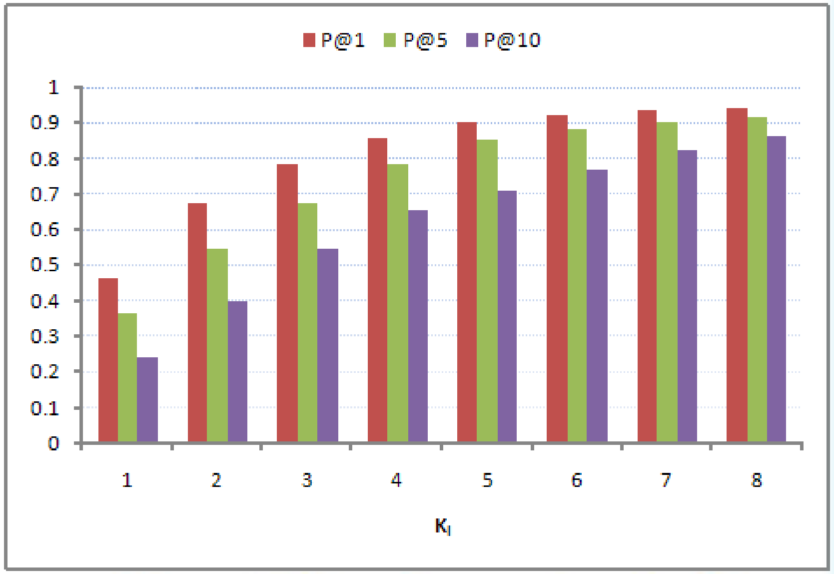
			
Init Tags	Cruise party boat purple spectrum	Travel sea seaweed water colors	Travel family sea sun beach
Tag Concurrence Only	friends fun birthday art girls summer Florida winter snow flower	vacation Asia trip holiday nature city cannon tree Europe building	vacation holiday Europe nature city water trip building Asia light
Tag Concurrence & Tag Content	friends girls music fun night love art holiday vacation trip	vacation holiday trip Asia Europe nature city fun music friends	vacation trip Asia holiday water Europe nature tree friends sun
Tag Concurrence & Tag Content & Image-Cond. Correlation	friends dance fun girls night music love men happy laugh	trip ocean sky island nature landscape blue umbrella red men	vacation fun water kids ocean sky holiday sand wave blue

Results

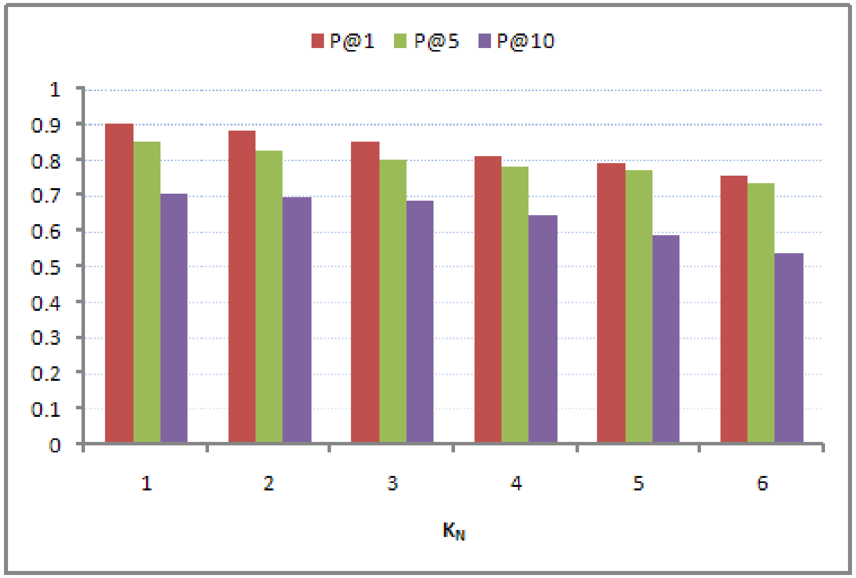

Precision@10
0.78
0.76
0.74
0.72
0.7
0.68
0.66
0.64
0.62
0.6
M1
M2
M3


Coverage@10
0.7
0.65
0.6
0.55
0.5
0.45
0.4
0.35
0.3
M1
M2
M3

Results



precision@N with K_I initial tags



precision@N with K_N irrelevant tags

Flickr Distance

Credit: Lei Wu, Xian-Sheng Hua, Nenghai Yu, Wei-Ying Ma, and Shipeng Li. Flickr Distance. ACM Multimedia 2008. (Best Paper Candidate).

Multimedia Information Retrieval

bing

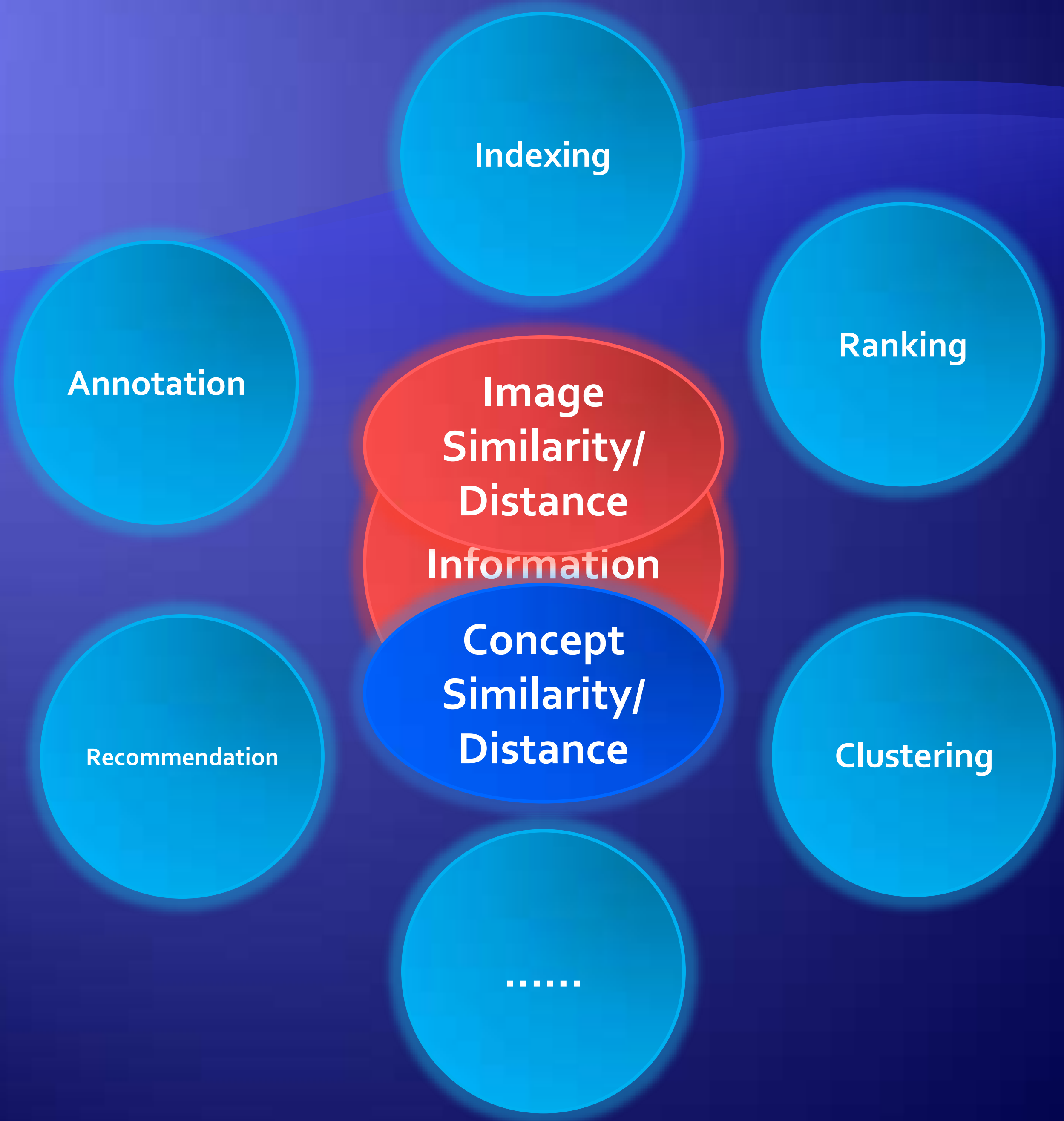

Indexing
Annotation
Ranking
Image Similarity/ Distance
Information
Concept Similarity/ Distance
Recommendation
Clustering
……

Image Similarity/Distance

Distance

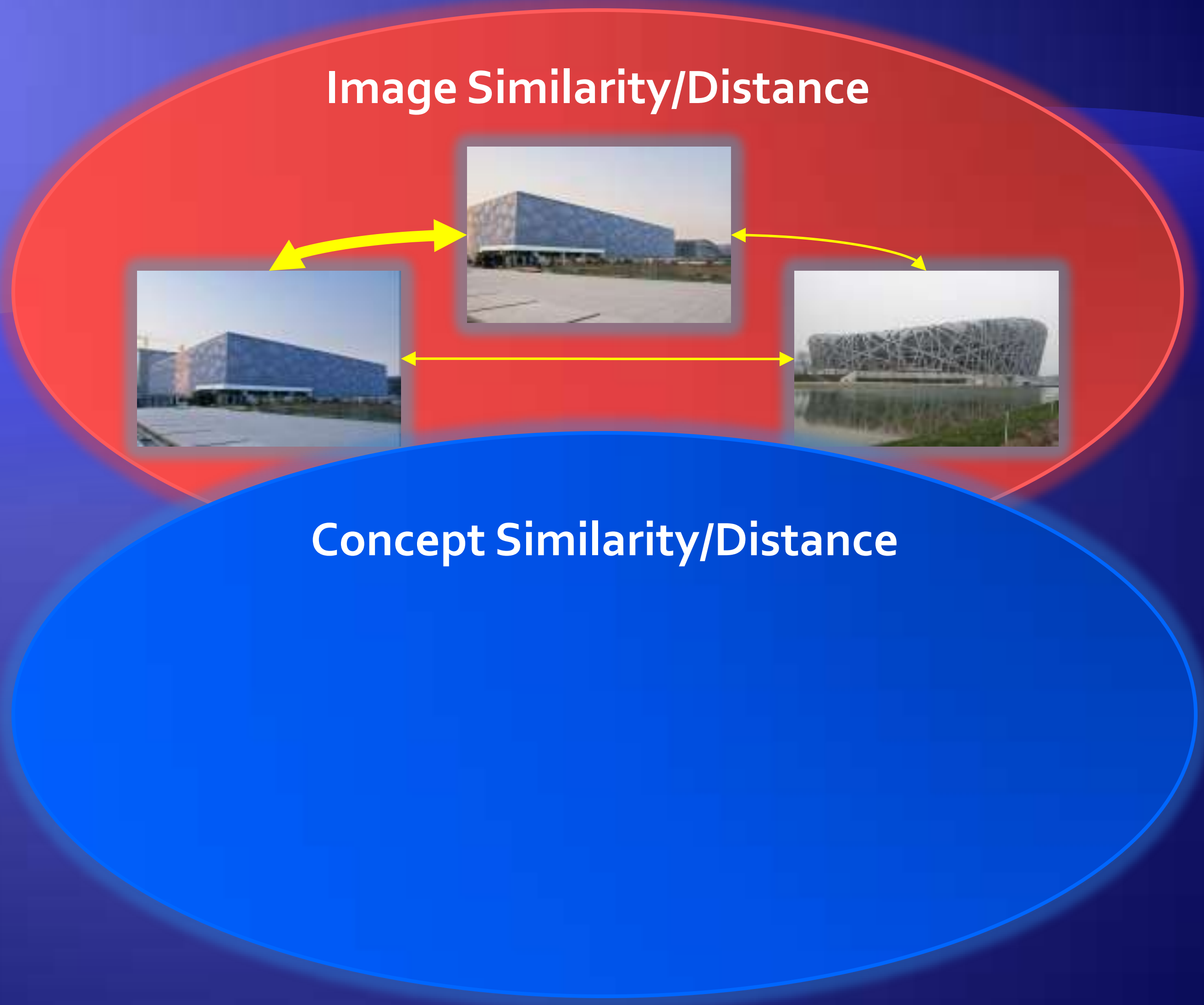
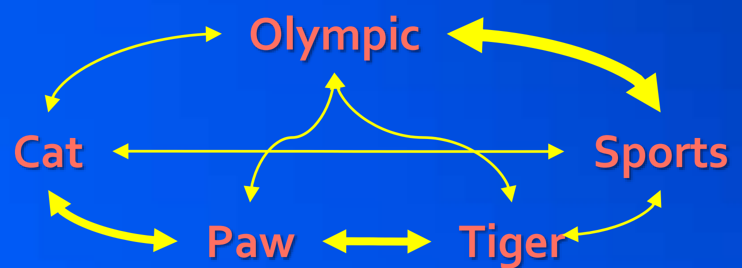


Image Similarity/Distance
Concept Similarity/Distance

Image Similarity/Distance

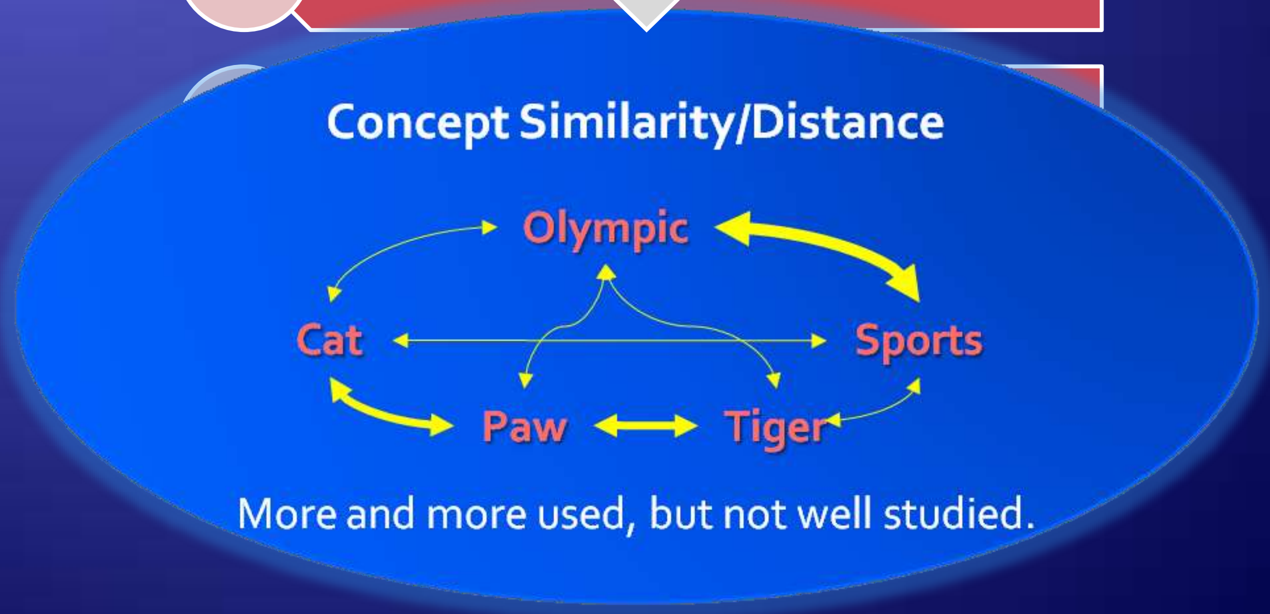
Numerous efforts have been made.

Concept Similarity/Distance



More and more used, but not well studied.

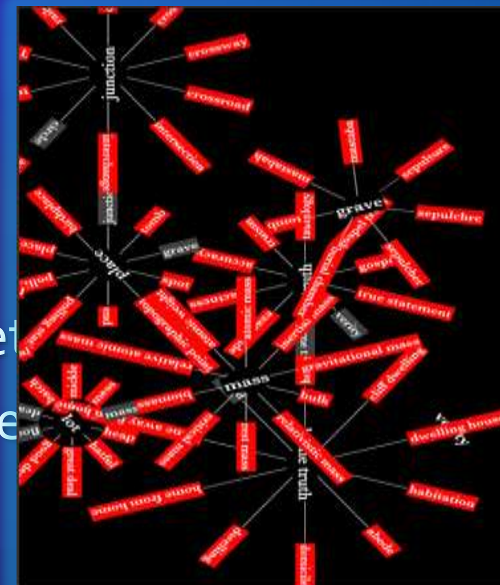
WordNet Distance

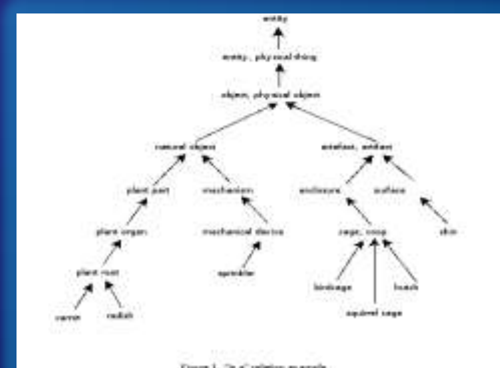


WordNet Distance

PRINCETON UNIVERSITY

- WordNet
  - 150,000 words
- WordNet Distance
  - Quite a few methods to get it in WordNet
  - Basic idea is to measure the length of the path between two words
- Pros and Cons
  - **Pros**: Built by human experts, so close to human perception
  - **Cons**: Coverage is limited and difficult to extend





Google Distance

UNIVERSITY OF AMSTERDAM

- Normalized Google Distance (NGD)
  - Reflects the concurrency of two words in Web documents
  - Defined as

$$NGD(x,y) = \frac{max(\log f(x), \log f(y)) - \log f(x,y)}{\log N - min(\log f(x), \log f(y))}$$

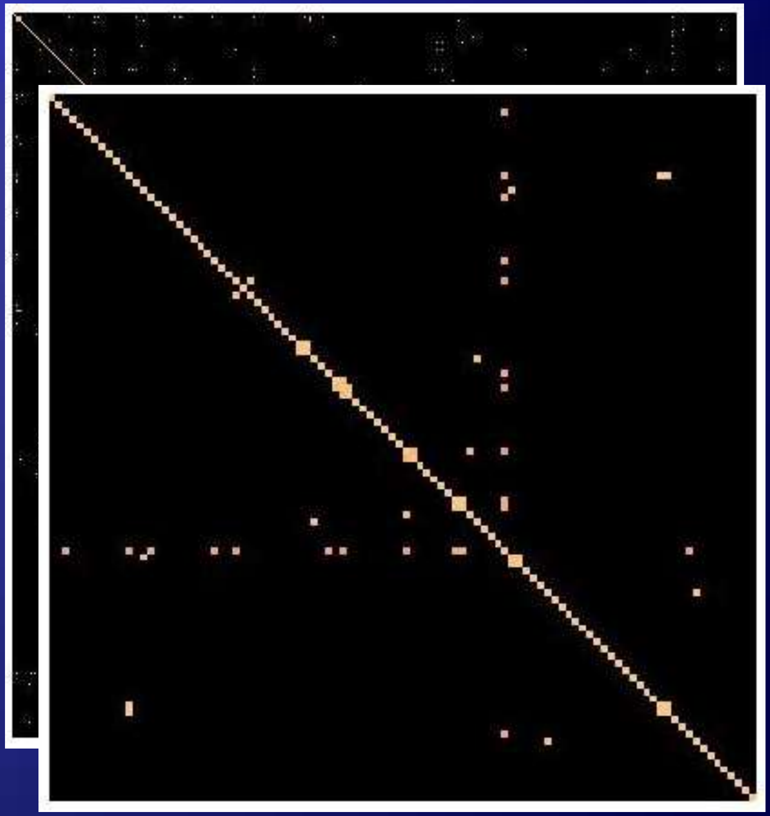
- Pros and Cons
  - **Pros**: Easy to get and huge coverage
  - **Cons**: Only reflects concurrency in textual documents. Not really concept distance (semantic relationship)

bing

Concept Pairs	Google Distance
Airplane – Dog	0.2562
Football – Soccer	0.1905
Horse – Donkey	0.2147
Airplane – Airport	0.3094
Car – Wheel	0.3146

Tag Concurrence Distance

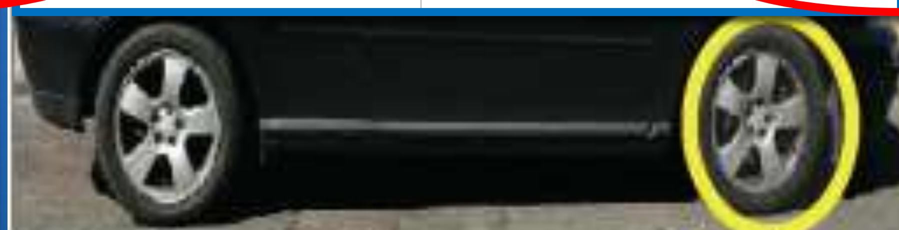
- Image Tag Concurrence Distance (Qi, Hua, et al. ACMMM07)
  - Reflects the frequency of two tags occur in the same images
  - Based on the same idea of NGD
  - Mostly is sparse (> 95% are zero in the similarity matrix)



similarity matrix: 50 tags

- Pros and Cons
  - **Pros**: Images are taken into account
  - **Cons**: a) Tags are sparse so visual concurrency is not well reflected
    b) Training data is difficult to get

Concept Pairs	Google Distance	ence Distance
Airplane – Dog	0.2562	532
Football – Soccer	0.1905	739
Horse – Donkey	0.2147	513
Airplane – Airport	0.3094	833
Car – Wheel	0.3146	617



Different Concept Relationships

table tennis — ping-pong

Synonymy

different words but the same meaning

horse — donkey

Visually Similar

similar things or things of same type

car — wheel

Meronymy

part and the whole

airplane — airport

Concurrency

exist at the same scene/place

Concept Distance

Image tag concurrence distance implicitly uses image information, but tags are too sparse

Mine from image tags

Google distance's coverage is very high, but it is for text domain

Mine from text documents

WordNet distance is good, but coverage is too low

Mine from ontology

Can we mine concept distance from image content?

Some Facts

- Semantic concept distance is based on human's cognition
- 80% of human cognition comes from visual information
- There are around 5 billion photos on Flickr (by 2010)
- In average each Flickr image has around 10 tags

To mine concept distance from a large tagged image collection based on image content

bear, fur, grass, tree

polar bear, water, sea

polar bear, fighting, usa

bing

Overview of Flickr Distance

Concept A: Airplane

Concept Model A

Concept B: Airport

Concept Model B

Flickr Distance (A, B)

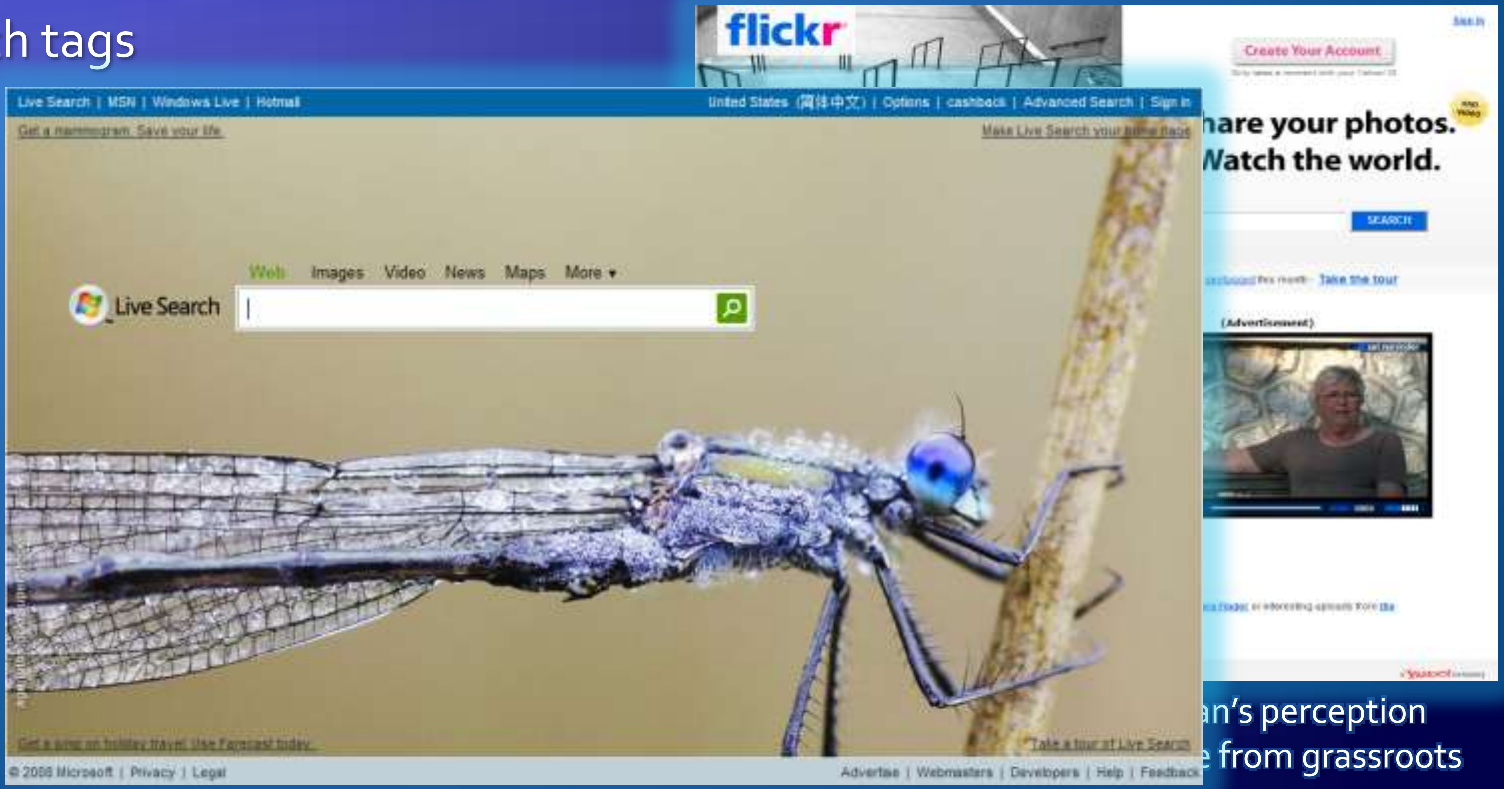
Concept Pairs	Google Distance	Tag Concurrent Distance	istance
Airplane – Dog	0.2562	0.8532	51
Football – Soccer	0.1905	0.1739	15
Horse – Donkey	0.2147	0.4513	31
Airplane – Airport	0.3094	0.1833	76
Car – Wheel	0.3146	0.9617	08

Flickr Distance is able to cover the four different semantic relationships

Synonymy, Visually Similar, Meronymy, and Concurrency

What We Need

- R1: A Good Image Collection
  - Large
  - High coverage, especially on real-life world
  - With tags



an's perception
from grassroots

bing

What We Need

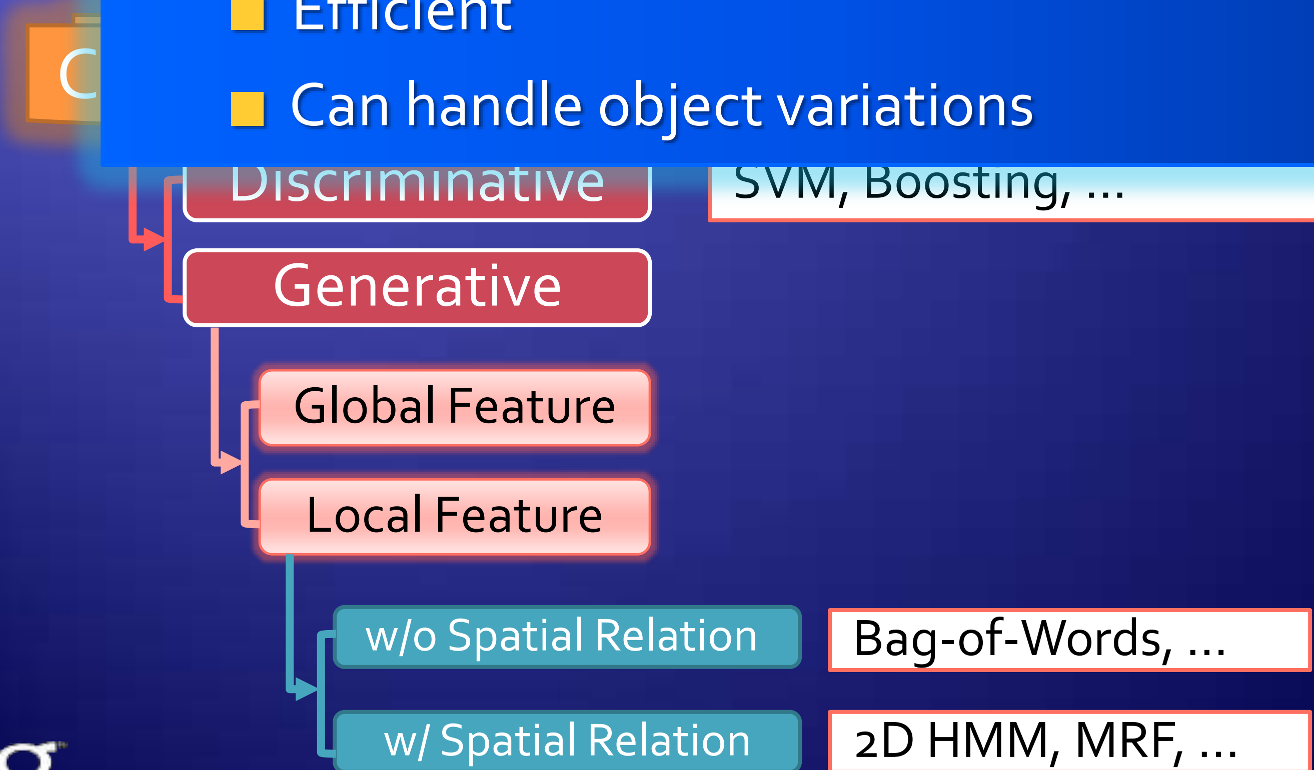
- R2: A Good Concept Representation or Model
  - Based on image content
  - Can cover wider concept relationships
  - Can handle large-concept set

Concept Models

- Discriminative — SVM, Boosting, ...
- Generative
  - Global Feature
  - Local Feature
    - w/o Spatial Relation — Bag-of-Words (pLSA, LDA), ...
    - w/ Spatial Relation — 2D HMM, MRF, ...

What We Need

- **VLM – Visual Language Model**
  - Spatial-relation sensitive
  - Efficient
  - Can handle object variations



Statistical Language Model

I am talking about statistical language model.

Unigram Model	$P(w_x \mid w_1 w_2 \cdots w_n) = P(w_x)$
Bigram Model	$P(w_x \mid w_1 w_2 \cdots w_n) = p(w_x \mid w_{x-1})$
Trigram Model	$P(w_x \mid w_1 w_2 \cdots w_n) = P(w_x \mid w_{x-1} w_{x-2})$

Visual Language Model (VLM)

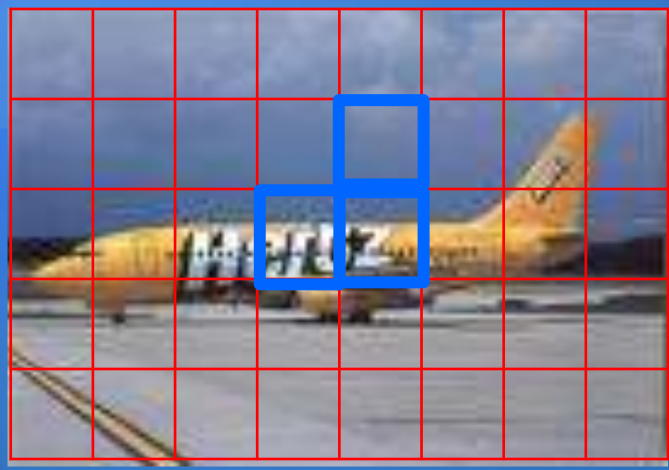


Image → Patch

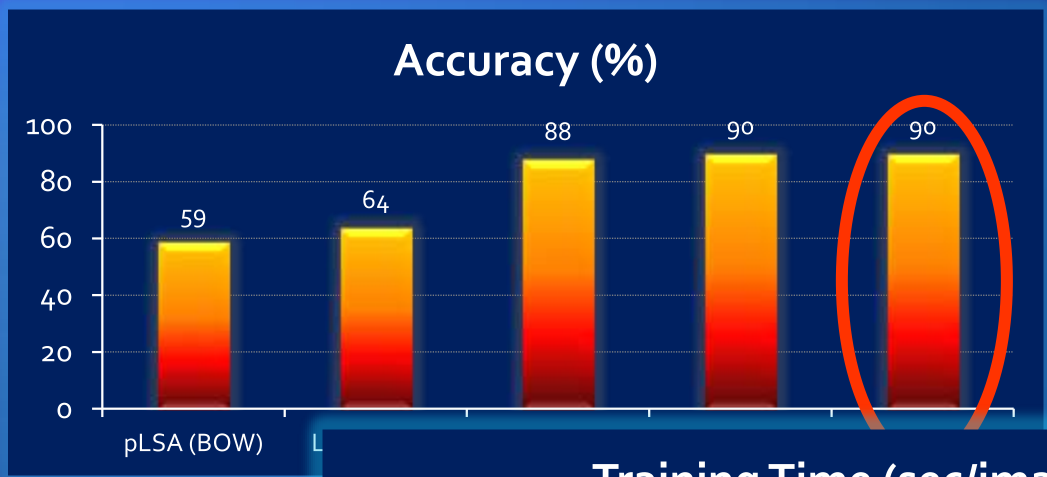
Visual Word Generation

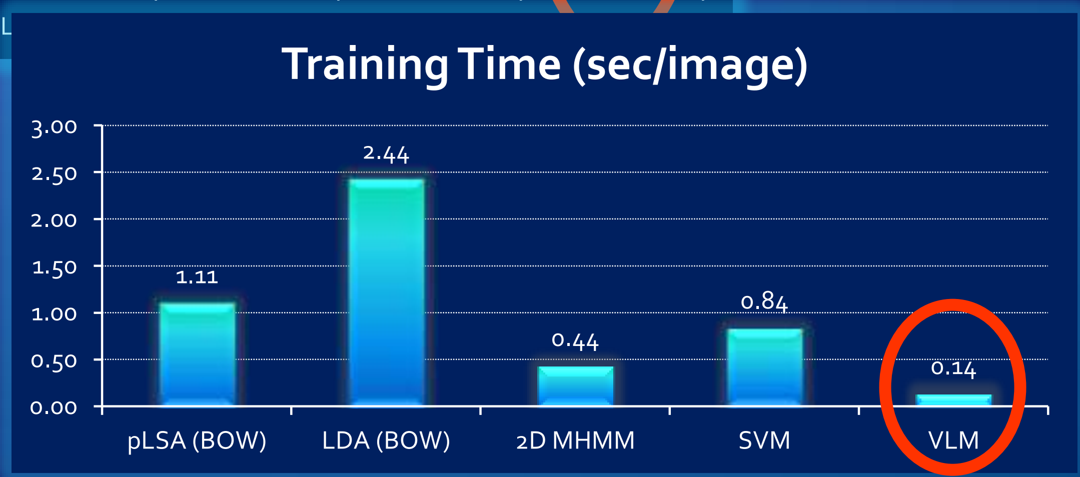
Unigram Model	$P\left(w_{xy} \middle| w_{11}w_{12}\cdots w_{mn}\right) = P\left(w_{xy}\right)$
Bigram Model	$P\left(w_{xy} \middle| w_{11}w_{12}\cdots w_{mn}\right) = P\left(w_{xy} \middle| w_{x-1,y}\right)$
Trigram Model	$P\left(w_{xy} \middle| w_{11}w_{12}\cdots w_{mn}\right) = p\left(w_{xy} \middle| w_{x-1,y}\, w_{x,y-1}\right)$

Trigram VLM is estimated by directly counting from sufficient samples of each category.
To avoid the bias in the sampling, back-off smoothing method is adopted.

Performance of VLM

- Comparison on Image Categorization
  - Caltech 8 categories / 5097 images





Latent-Topic VLM (1)

- Why Latent-Topic

- Latent-Topic VLM
  - Visual variations of concept are taken as latent topics

$$P\left(w_{xy} \middle| w_{x-1,y} w_{x,y-1}, d_j^C\right) = \sum_{k=1}^{K} P\left(w_{xy} \middle| w_{x-1,y} w_{x,y-1}, z_k^C\right) P\left(z_k^C \middle| d_j^C\right)$$

C : *A concept*

d_j^C : *the j^{th} image in concept C*

z_k^C : *the k^{th} latent topic of concept C*

Latent-Topic VLM (2)

Latent-Topic VLM Training

- Solved by EM algorithm,
- The objective function is to maximize the joint distribution of concept and its visual word arrangement A_w

E step:

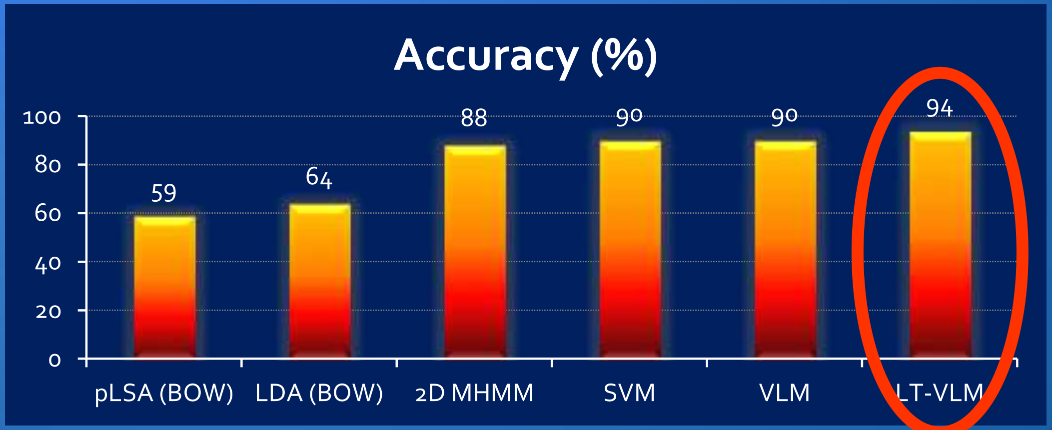
Estimate the posteriors of the hidden topics

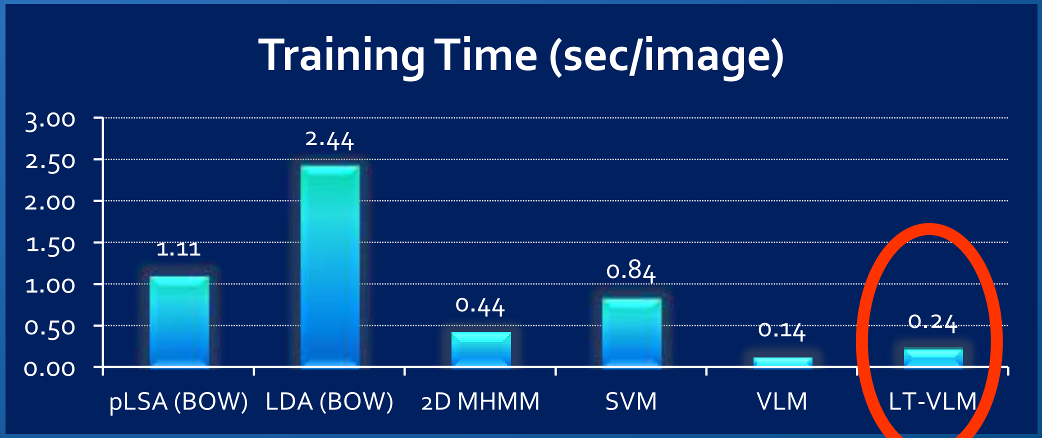
M step:

Maximize the likelihood of visual arrangement

Performance of LT-VLM

- Comparison on Image Categorization
  - Caltech 8 categories / 5097 images





Flickr Distance

- Kullback – Leibler (KL) divergence
  - Good, but not symmetric

$$D_{KL}(P_{Z_i^{c_1}} | P_{Z^{c_2}}) = \sum_l P_{Z_i^{c_1}}(l) \log \frac{P_{Z_i^{c_1}}(l)}{P_{Z_j^{c_2}}(l)}$$

topic distance

- Jensen –Shannon (JS) divergence
  - Better, as it is symmetric
  - And, square root of JS divergence is a metric, so is Flickr Distance

$$D_{JS}(P_{Z_i^{c_1}} | P_{Z^{c_2}}) = \frac{1}{2} D_{KL}(P_{Z_i^{c_1}} | M) + \frac{1}{2} D_{KL}(P_{Z_j^{c_2}} | M)$$

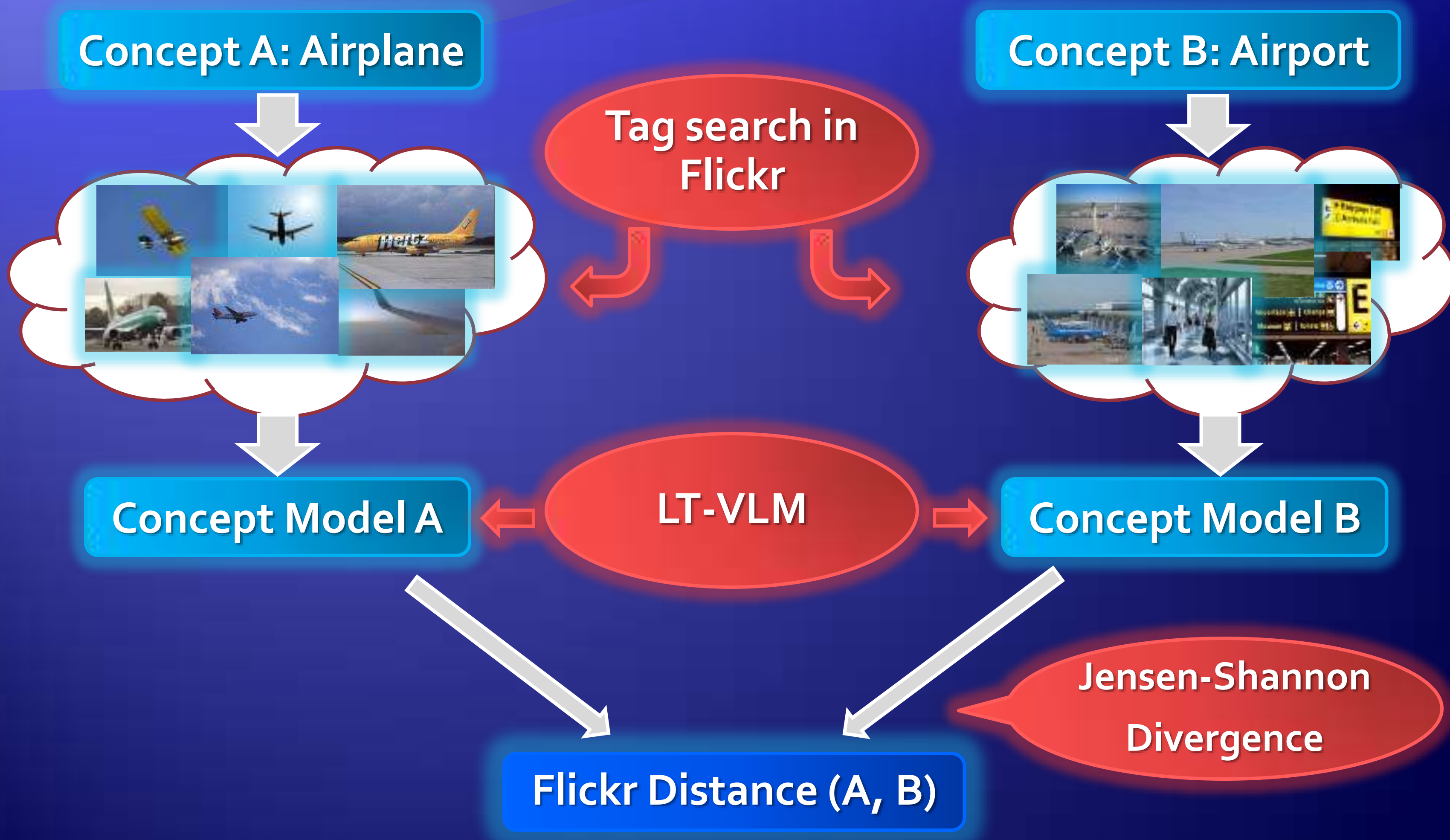
$$M = \left(P_{Z_i^{c_1}} + P_{Z_i^{c_1}}\right)/2$$

topic distance

$$D_{Flickr}(C_1, C_2) = \sqrt{\sum_{i=1}^{K} \sum_{j=1}^{K} P(z_i^{C_1} | C_1) P(z_j^{C_2} | C_2) D_{JS}(P_{z_i^{c1}} | P_{z_j^{c2}})}$$

concept distance

Procedure of Flickr Distance



Experiments

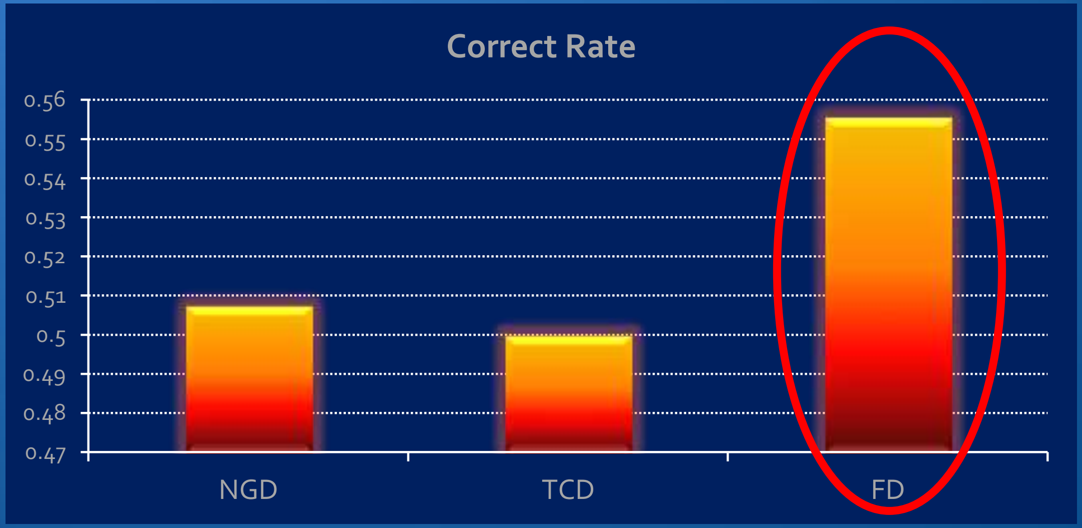
- Evaluation
  - Objective evaluation
  - Subjective evaluation
- Applications
  - Concept clustering
  - Image annotation
  - Tag recommendation

Experiments - Configurations

- Images
  - 6,400,000 from Flickr
- Concepts
  - 130,000,000 different tags
  - 10,000,000 filtered tags
  - 1,000 randomly-selected tags
- Comparison
  - Normalized Google Distance (NGD)
  - Tag Concurrence Distance (TCD)
  - Flickr Distance (FD)

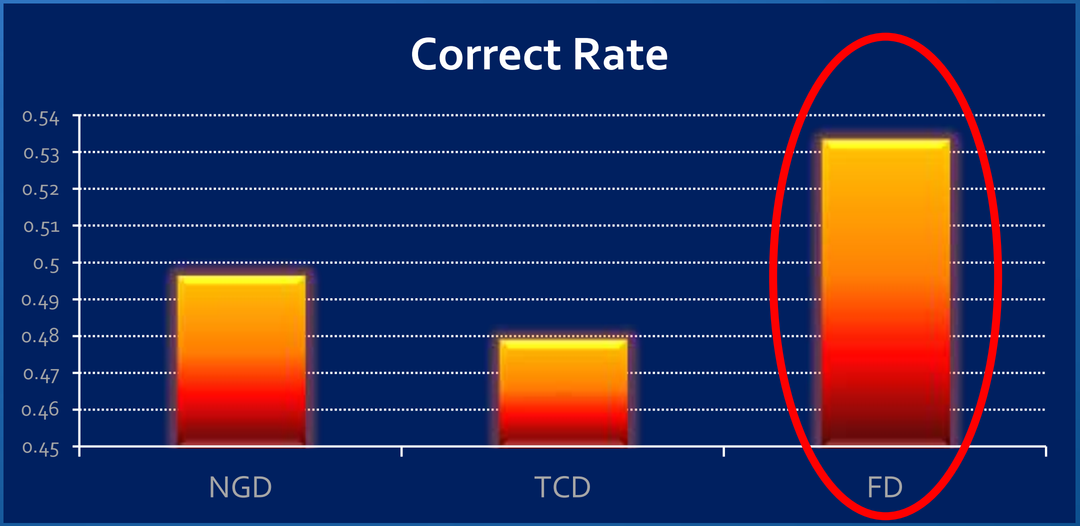
Eva1: Subjective Evaluation

- Ground-Truth
  - **12 persons are asked to score semantic correlation of each concept pair**
  - Average scores are taken as ground-truth
- Evaluate Accuracy of "Relative Distance Pairs"
  - Step 1: Find all distance pairs D(a,b) and D(c,d)
  - Step 2: Check whether the order of D(a,b) and D(c,d) is consistent with ground-truth



Eva2: Objective Evaluation

- Ground-Truth
  - **WordNet Distance**
  - Only 497 concepts (overlap of WordNet and the 1000 concepts)
- Evaluate Accuracy of "Relative Distance Pairs"
  - Step 1: Find all distance pairs D(a,b) and D(c,d)
  - Step 2: Check whether the order of D(a,b) and D(c,d) is consistent with ground-truth



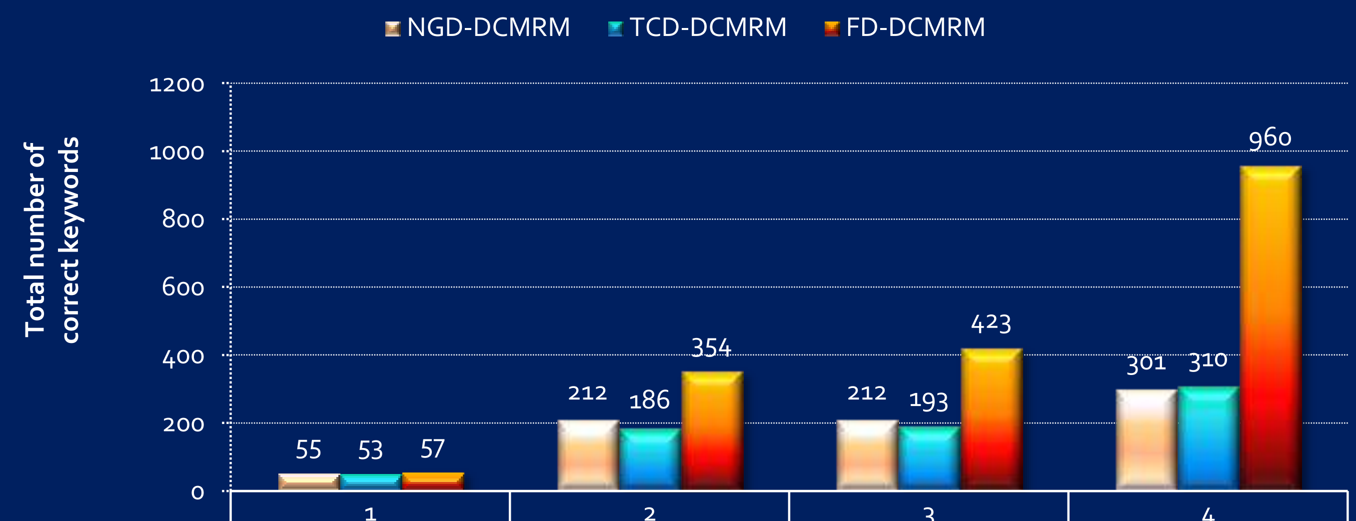
App1: Concept Clustering

- Concept Clustering
  - 23 concepts;
  - 3 groups – (1) outer space, (2) animal and (3) sports

Normalized Google Distance			Tag Concurrence Distance			Flickr Distance		
Group1	Group2	Group3	Group 1	Group2	Group3	Group1	Group2	Group3
bears	**bowling**	baseball	moon	baseball	basketball	moon	bears	baseball
horses	dolphin	basketball	space	**donkey**	**bears**	Saturn	dolphin	basketball
moon	donkey	football	Venus	softball	bowling	space	donkey	football
space	**Saturn**	golf	**whale**	**wolf**	**dolphin**	Venus	**golf**	**snake**
	sharks	soccer			football		horses	soccer
	snake	tennis			golf		sharks	bowling
	softball	volleyball			**horses**		spiders	softball
	spiders				**Saturn**		**tennis**	volleyball
	turtle				**sharks**		whale	
	Venus				soccer		wolf	
	whale				**spiders**			
	wolf				tennis			
					turtle			
					volleyball			

App2: Image Annotation

- Based on an approach using concept relation
  - Dual Cross-Media Relevance Model (DCMRM, J. Liu et al. ACMMM 2007)
  - On 79 concepts / 79,000 images

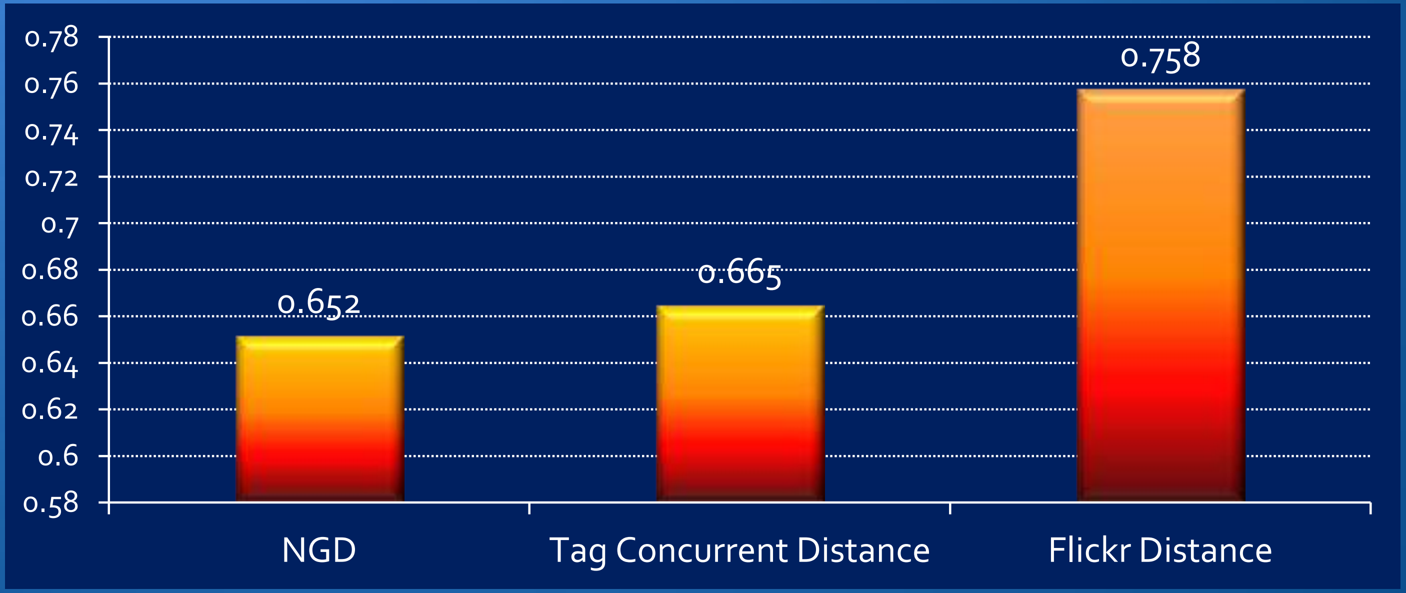


	1	2	3	4
NGD-DCMRM	55	212	212	301
TCD-DCMRM	53	186	193	310
FD-DCMRM	57	354	423	960

The number of correctly annotated keywords at the first N words

App3: Tag Recommendation

- To Improve Tagging Quality
  - Eliminating tag incompletion, noises, and ambiguity
  - 500 images / 10 recommended tags per image



Precision @ 10

bing

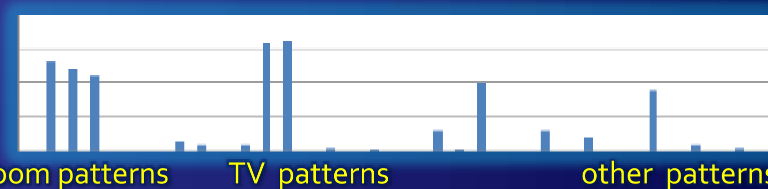
Discussion

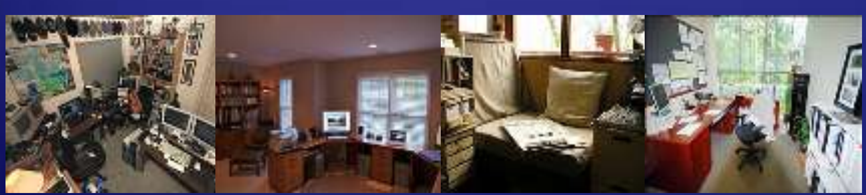
- Why VLM divergence can estimate concept distance?
- Why FD works well even tags are not complete?

Computer

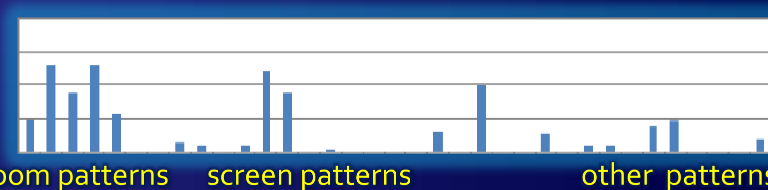
VLM: distribution of trigrams

TV





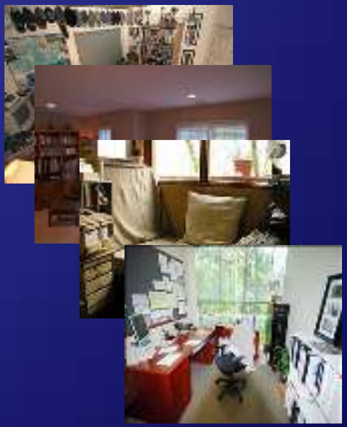
Office



If we find similar patterns in the images associated with different concepts,

the corresponding concept relationships can be discovered.

Computer



Office

Flickr Distance

- A novel approach to discover semantic relationships from image content
  - based on real-life images from the Web
  - based on collective intelligence from grassroots
- A distance more consistent with human's perception
- A measurement more effective in many applications

bing

Social Tagging and Tag Processing

- Discussion
  - User-input tags are better than surrounding text
  - Tags are not perfect
  - Quality of tags can be enhanced, whether during tagging or after tagging through mining the correlations among tags, among images and between tags and images
  - Implicit tagging is extensively used in search engines, but not easy to be studied in academia
  - Tag has limitations – still far from a complete description

Test

- Why user-input tags are not perfect? What are the major problems for those tags?
- Why we are able to improve the quality of user-input-tags without any additional manual labeling?

Outline

Session	Time	Topic
0	09:30 – 09:40	Introduction
1	09:40 – 10:20	Learning-Based Tagging
2	10:20 – 11:40	Social Tagging and Tag Processing
		Including a break (10:45 – 11:00)
3	11:40 – 12:10	Data-Driven Web-Scale Tagging
4	12:10 – 12:30	Future Directions/QA

Outline

Session	Time	Topic
0	09:30 – 09:40	Introduction
1	09:40 – 10:20	Learning-Based Tagging
2	10:20 – 11:40	Social Tagging and Tag Processing
		Including a break (10:45 – 11:00)
3	11:40 – 12:10	Data-Driven Web-Scale Tagging
4	12:10 – 12:30	Future Directions/QA

bing

Data Driven Web-Scale Tagging

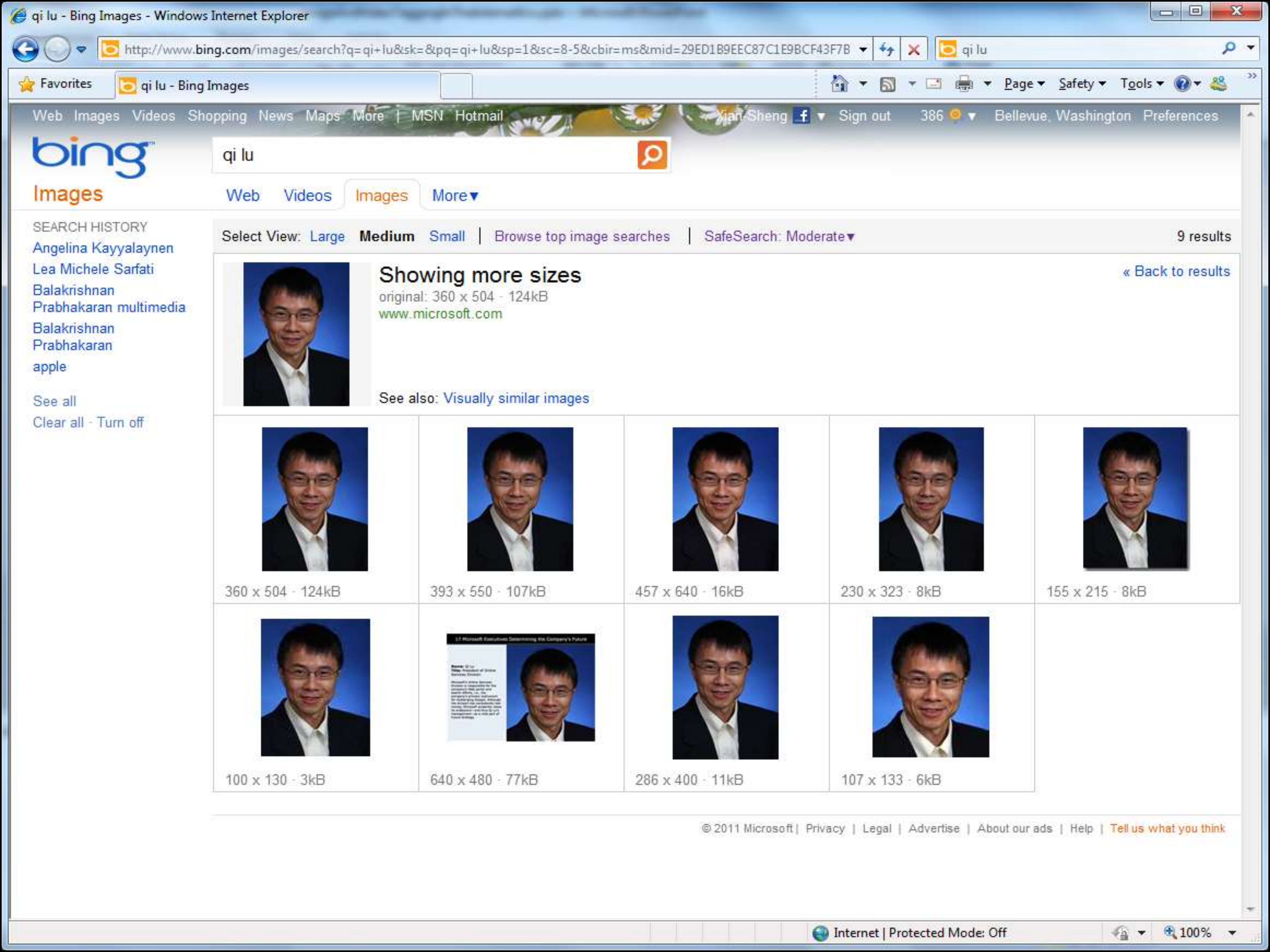
- CBIR-Based Tagging
- Building Web-Scale Image Graph
- Discussion

bing

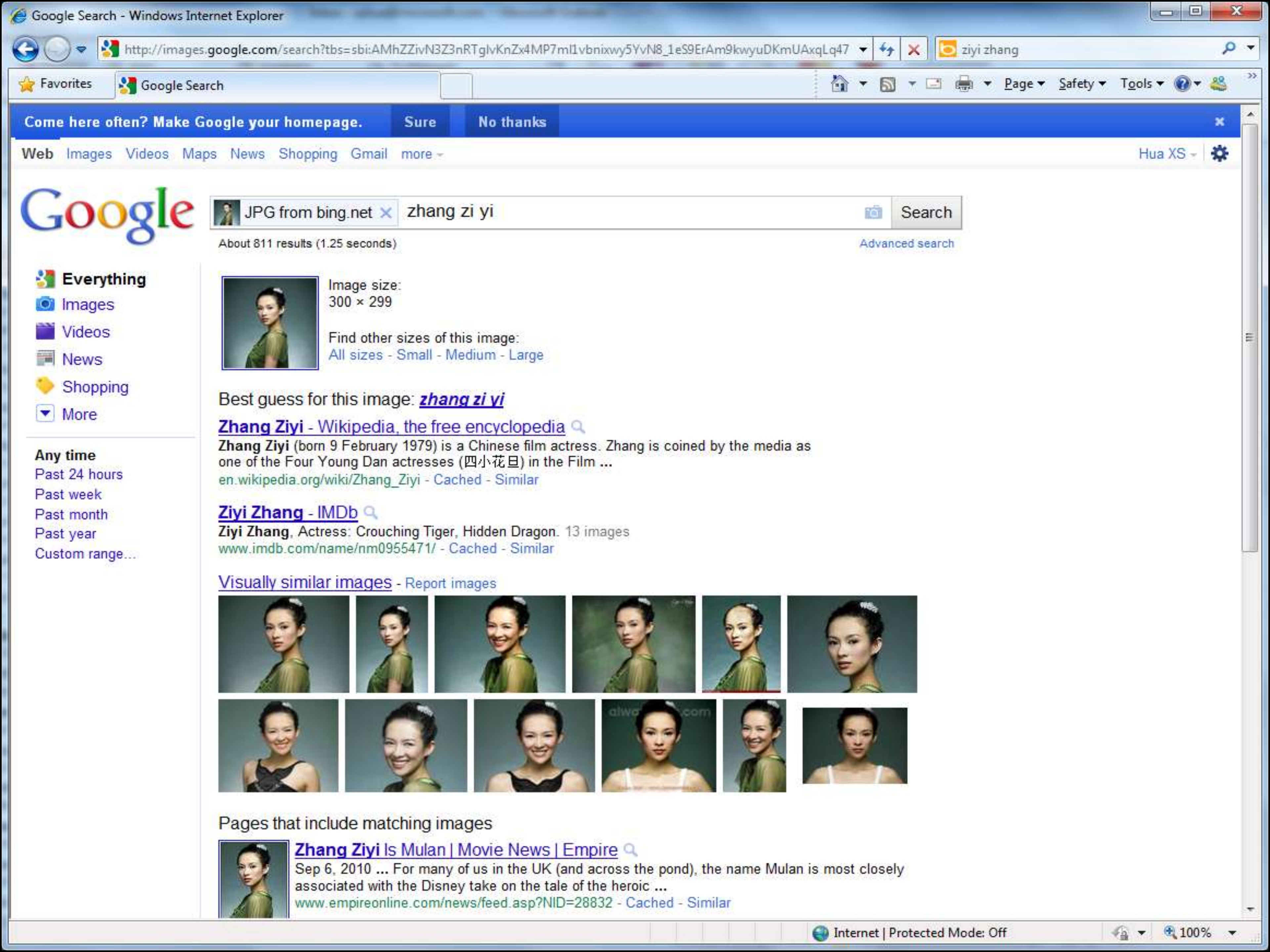
Introduction

- Some facts
  - A huge number of images available on the Internet
    - A portion of them has textual descriptions
    - A small portion of them has tags
  - Many images have multiple copies/versions
    - Images are reused in different places
    - Some with textual description or tags, while some not

bing



qi lu - Bing Images - Windows Internet Explorer
http://www.bing.com/images/search?q=qi+lu&sk=&pq=qi+lu&sp=1&sc=8-5&cbir=ms&mid=29ED1B9EEC87C1E9BCF43F7B
qi lu
Favorites
qi lu - Bing Images
Page
Safety
Tools
Web Images Videos Shopping News Maps More | MSN Hotmail
Xian-Sheng
Sign out
386
Bellevue, Washington
Preferences
bing
Images
qi lu
Web
Videos
Images
More
SEARCH HISTORY
Angelina Kayyalaynen
Lea Michele Sarfati
Balakrishnan Prabhakaran multimedia
Balakrishnan Prabhakaran
apple
See all
Clear all · Turn off
Select View: Large Medium Small | Browse top image searches | SafeSearch: Moderate
9 results
Showing more sizes
original: 360 x 504 · 124kB
www.microsoft.com
« Back to results
See also: Visually similar images
360 x 504 · 124kB
393 x 550 · 107kB
457 x 640 · 16kB
230 x 323 · 8kB
155 x 215 · 8kB
100 x 130 · 3kB
640 x 480 · 77kB
286 x 400 · 11kB
107 x 133 · 6kB
© 2011 Microsoft | Privacy | Legal | Advertise | About our ads | Help | Tell us what you think
Internet | Protected Mode: Off
100%


Google Search - Windows Internet Explorer
http://images.google.com/search?tbs=sbi:AMhZZivN3Z3nRTglvKnZx4MP7ml1vbnixwy5YvN8_1eS9ErAm9kwyuDKmUAxqLq47
ziyi zhang
Favorites
Google Search
Page
Safety
Tools
Come here often? Make Google your homepage.
Sure
No thanks
Web Images Videos Maps News Shopping Gmail more
Hua XS
Google
JPG from bing.net
zhang zi yi
Search
About 811 results (1.25 seconds)
Advanced search
Everything
Images
Videos
News
Shopping
More
Any time
Past 24 hours
Past week
Past month
Past year
Custom range...
Image size:
300 × 299
Find other sizes of this image:
All sizes - Small - Medium - Large
Best guess for this image: zhang zi yi
Zhang Ziyi - Wikipedia, the free encyclopedia
Zhang Ziyi (born 9 February 1979) is a Chinese film actress. Zhang is coined by the media as one of the Four Young Dan actresses (四小花旦) in the Film ...
en.wikipedia.org/wiki/Zhang_Ziyi - Cached - Similar
Ziyi Zhang - IMDb
Ziyi Zhang, Actress: Crouching Tiger, Hidden Dragon. 13 images
www.imdb.com/name/nm0955471/ - Cached - Similar
Visually similar images - Report images
Pages that include matching images
Zhang Ziyi Is Mulan | Movie News | Empire
Sep 6, 2010 ... For many of us in the UK (and across the pond), the name Mulan is most closely associated with the Disney take on the tale of the heroic ...
www.empireonline.com/news/feed.asp?NID=28832 - Cached - Similar
Internet | Protected Mode: Off
100%

How Many Duplicates on the Web?

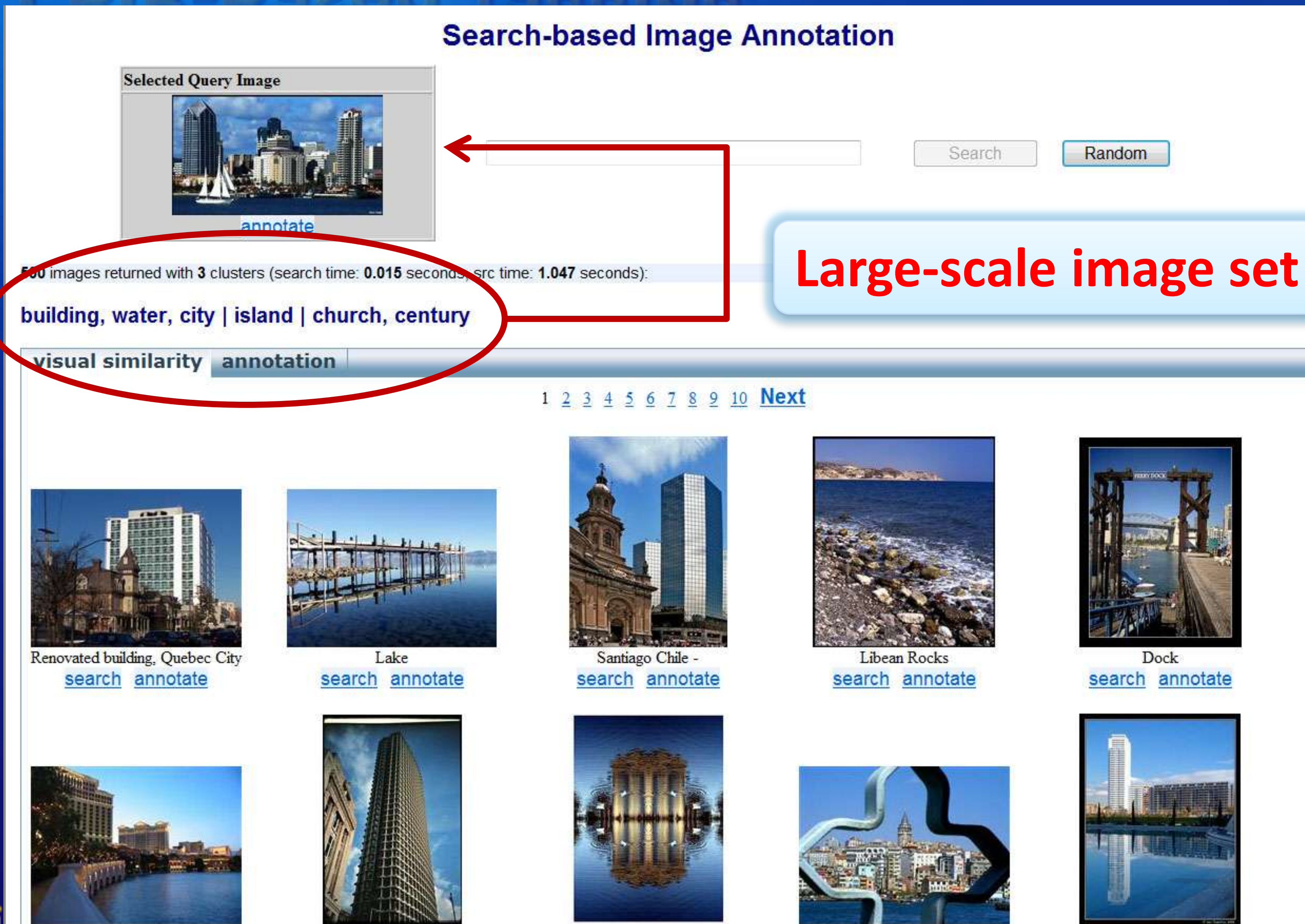
- 8.1% web images has no less than 10 duplicates
- 28.5% images shown in top 40 search results has no less than 10 duplicates

bing

CBIR-Based Tagging

Credit: Xin-Jiang Wang, Lei Zhang, et al. CVPR 2006.

CBIR-Based Tagging



Large-scale image set

Near-dup search based Tagging on 2 Billion Images – Basic Idea

The Value of Duplicates

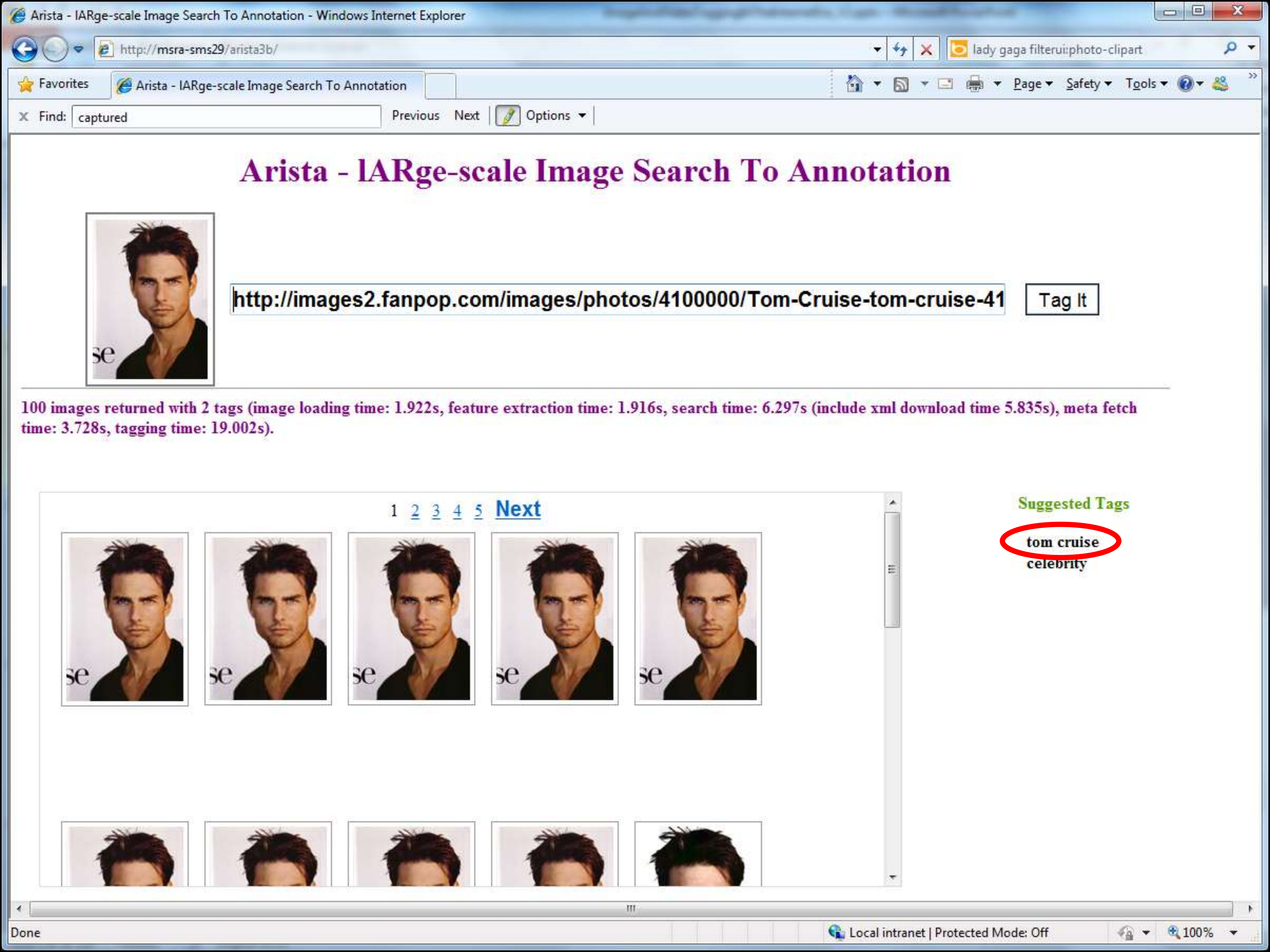
- Duplicate search is a well-defined problem
- Frequent terms/phrases indicate semantics

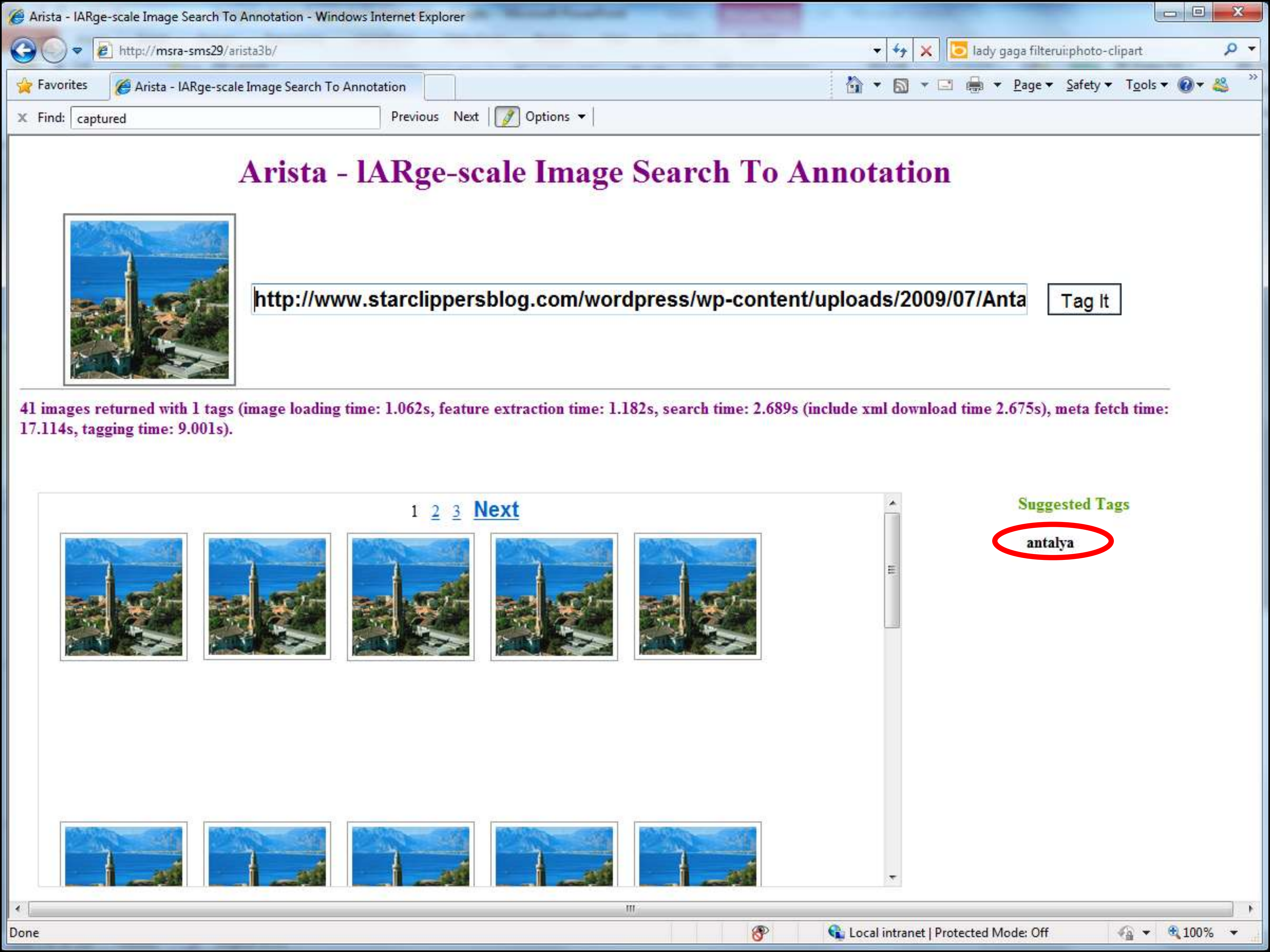
 prison break sarah callies sara tancredi looking (339 dups)	**sarah wayne callies** picture thread bild-quelle edit by annika beitraege in einen...	 aeon concept phone mobile phone cell phone touch screen nokia phone mobile nokia (1888 dups)	**nokia aeon** was presented by **nokia** on their website in the research development...
	prison break is paging dr. sara. if you are one of the many **prison break** fans...		**nokia aeon concept phone** (no ratings yet) sexy is the word to describe it **nokia** is ...
	prison break - dr **sara tancredi** is not dead you knew that, right?dr **sara tancredi** ...		**nokia aeon** - future **mobile phone**
	dr. **sara** comes back to **prison break**?		**nokia aeon concept phone nokia** has unveiled its latest concept unbelievable ...
 costa rica golden toad climate amphibian (18 dups)	this is a picture of male **golden toads** congregating for breeding...	 sydney opera house australia (19 dups)	enjoying the wet season in **australia sydney**...
	is there a relationship between climate variability & amphibian declines? **golden toad**		150975_**sydney_opera_house** next ...
	male **golden toads** at a breeding pool in indigenous to monteverde **costa rica**...		07/12, 1. tag in **sydney** > **opera house** ...
	amphibian declines in the cloud forests of **costa rica** ...		kirsty and trudy drink wine **sydney opera house** ...

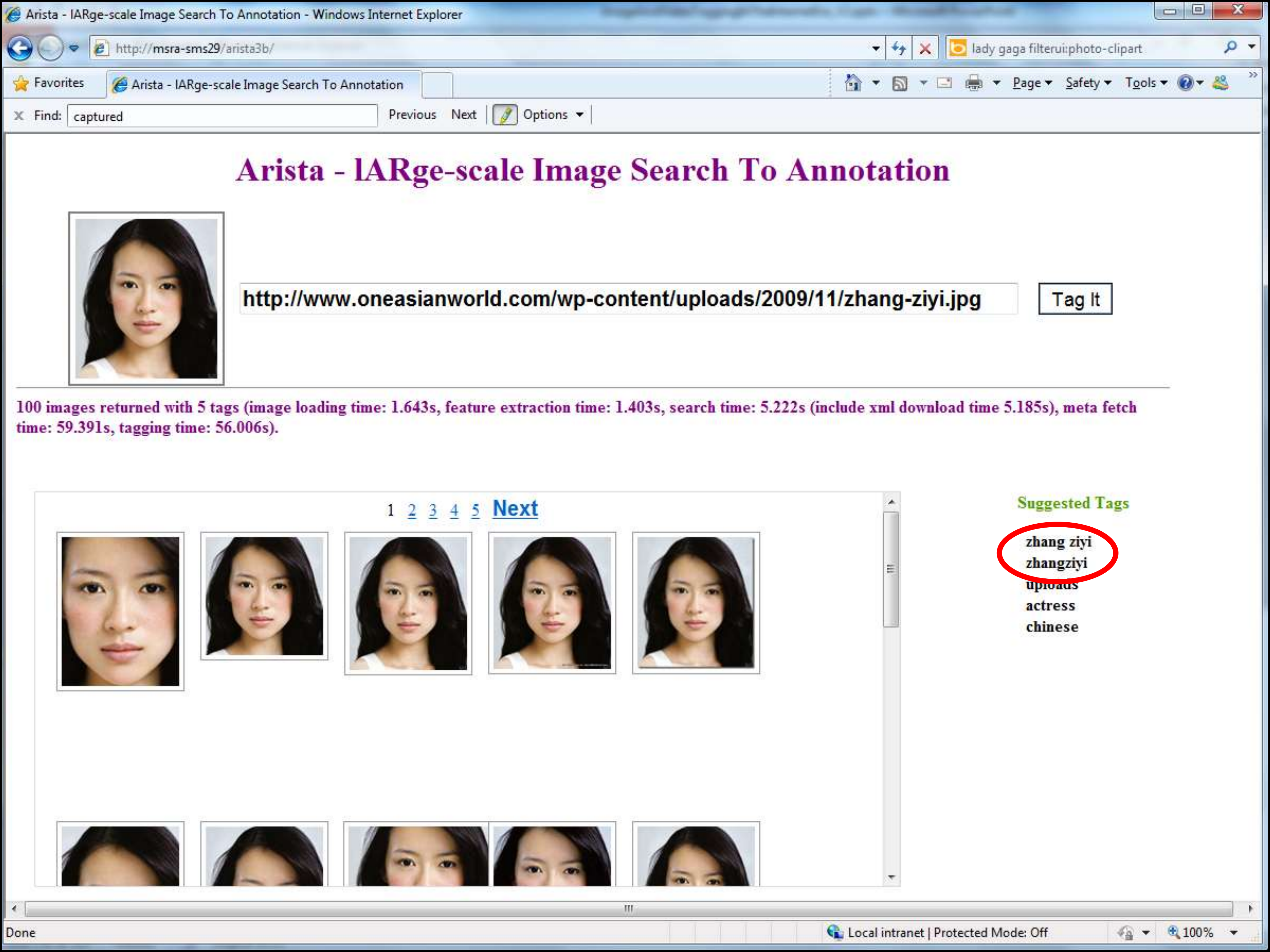
- Dup Images favor the concepts closer to Web users’ interest
  - Celebrity, product, Landmark, Cartoon, Painting, …
- However, not well for personal images
  - When there is no duplicate, the system will fail.
  - Similarity-based search is still needed, but very challenging.

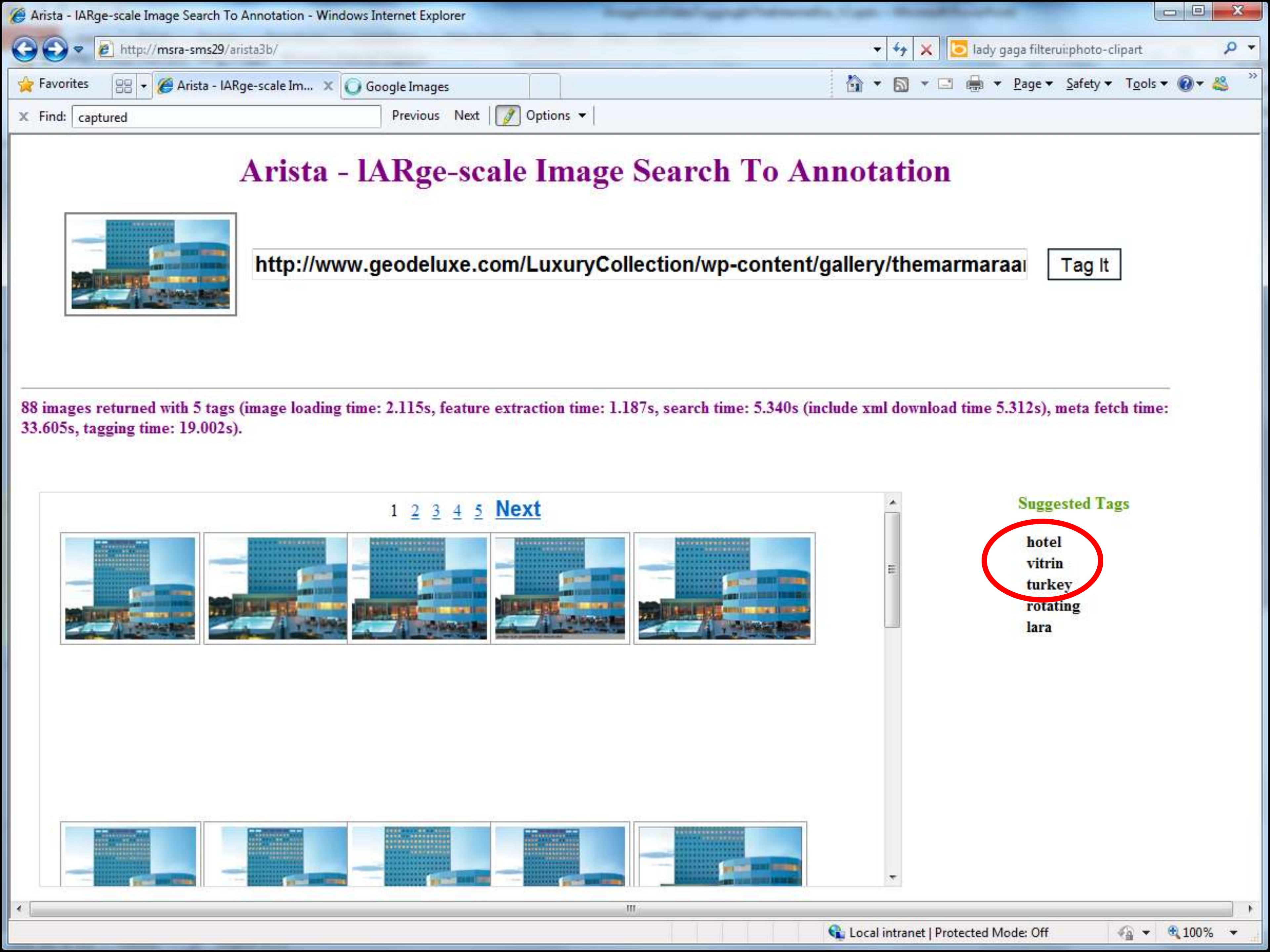
DEMOS

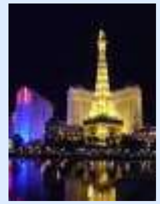
- Demo Link 1
  - Submit URL of an image and see the tags
  - Online tagging: generally needs about 5 sec to get the results (large time cost on xml downloading)
- Demo Link 2
  - Shows the improved precision with a Bayesian model


Arista - IARge-scale Image Search To Annotation - Windows Internet Explorer
http://msra-sms29/arista3b/
lady gaga filterui:photo-clipart
Favorites
Arista - IARge-scale Image Search To Annotation
Page
Safety
Tools
Find: captured
Previous
Next
Options
Arista - IARge-scale Image Search To Annotation
http://images2.fanpop.com/images/photos/4100000/Tom-Cruise-tom-cruise-41
Tag It
100 images returned with 2 tags (image loading time: 1.922s, feature extraction time: 1.916s, search time: 6.297s (include xml download time 5.835s), meta fetch time: 3.728s, tagging time: 19.002s).
1 2 3 4 5 Next
Suggested Tags
tom cruise
celebrity
Done
Local intranet | Protected Mode: Off
100%


Arista - IARge-scale Image Search To Annotation - Windows Internet Explorer
http://msra-sms29/arista3b/
lady gaga filterui:photo-clipart
Favorites
Arista - IARge-scale Image Search To Annotation
Page Safety Tools
Find: captured
Previous Next Options
Arista - IARge-scale Image Search To Annotation
http://www.starclippersblog.com/wordpress/wp-content/uploads/2009/07/Anta
Tag It
41 images returned with 1 tags (image loading time: 1.062s, feature extraction time: 1.182s, search time: 2.689s (include xml download time 2.675s), meta fetch time: 17.114s, tagging time: 9.001s).
1 2 3 Next
Suggested Tags
antalya
Done
Local intranet | Protected Mode: Off
100%


Arista - IARge-scale Image Search To Annotation - Windows Internet Explorer
http://msra-sms29/arista3b/
lady gaga filterui:photo-clipart
Favorites
Arista - IARge-scale Image Search To Annotation
Page
Safety
Tools
Find: captured
Previous
Next
Options
Arista - IARge-scale Image Search To Annotation
http://www.oneasianworld.com/wp-content/uploads/2009/11/zhang-ziyi.jpg
Tag It
100 images returned with 5 tags (image loading time: 1.643s, feature extraction time: 1.403s, search time: 5.222s (include xml download time 5.185s), meta fetch time: 59.391s, tagging time: 56.006s).
1 2 3 4 5 Next
Suggested Tags
zhang ziyi
zhangziyi
uploads
actress
chinese
Done
Local intranet | Protected Mode: Off
100%


Arista - IARge-scale Image Search To Annotation - Windows Internet Explorer
http://msra-sms29/arista3b/
lady gaga filterui:photo-clipart
Favorites
Arista - IARge-scale Im...
Google Images
Page
Safety
Tools
Find: captured
Previous
Next
Options
Arista - IARge-scale Image Search To Annotation
http://www.geodeluxe.com/LuxuryCollection/wp-content/gallery/themarmaraa
Tag It
88 images returned with 5 tags (image loading time: 2.115s, feature extraction time: 1.187s, search time: 5.340s (include xml download time 5.312s), meta fetch time: 33.605s, tagging time: 19.002s).
1 2 3 4 5 Next
Suggested Tags
hotel
vitrin
turkey
rotating
lara
Local intranet | Protected Mode: Off
100%

	mercedes benz; swarovski crystal		Logo; mercedes benz; mercedes van; mercedes logo		chocolate, Red, Favorites
	Las vegas		Vegas; las vegas		sacre coeur; Paris; location vacances
	paris hilton; hollywood gossip;		barack obama; presidential candidate		bill gates
	frida kahlo; hope,tree,art; masters painter		van gogh; oil painting; drinkers, vangogh		van gogh; night café; oil paintings
	Happy birthday dog balloons; Glitter		Simpsons movie		travel inn; premier inn; Accommodation; city centre; basildon hotel
	pearl harbor josh hartnett		timber wolf		Monkey

Some Numbers

- Precision
  - 57.9% as reported in CVPR'10
  - improved to 92% (on 500 randomly selected images, Bayesian model)
- However, can hardly scale-up
  - Needs ~1 year to process all Bing images

Towards Efficient Near-dup Search

- Local feature-based near-dup detection (Bing MM)
  - Low efficiency: Needs ~1yr to process all Bing images
- MD5-based exact-dup detection (Bing MM)
  - Low coverage: 2.07% of #dup>=3
- Hashcode-based near-dup detection (Lei Zhang) *
  - 28-bits hashing: ~4 times of MD5 method, precision 97.1%
  - 40-bits hashing: ~1.3 times of MD5 method, precision 99.8%

* Rough estimation, numbers may be affected by two problems:

- different images having the same image key
- the population of non-photo images

bing

Hash Code Based Near-dup Detection

Bin Wang, et.al. *Large-scale duplicate detection for web image search*, In ICME 2006.

- Input Image
- Divide image to blocks
- Extract mean gray values
- Concatenate values
- Project to PCA space
- Quantization

Summary - Dup search methods

	T2S2	md5	Hash Code
Feature	Sift-like Local features	Md5 on first 3M pixels	8x8 grid→ Hash code
Coverage	High	Low	High
Efficiency	Low	High	High
Similarity Search	No	No	No
Service	Available	Available	Under Research

bing

Building Web-Scale Image Graph

Credit: Jingdong Wang, Xian-Sheng Hua and Shipeng Li

Techniques - Scalable k-NN graph construction

- Organize images with *hybrid neighborhood graph* that is constructed using visual and textual features
  - 3M images 3 hours, 80% accuracy
- Fast hybrid similarity search with *neighborhood propagation*
  - Below 50ms per image
- Advantages
  - Easy insertion and deletion
  - O(1) time

Key technologies

- **Multiple random divide-and-conquer**

- **Neighborhood propagation**

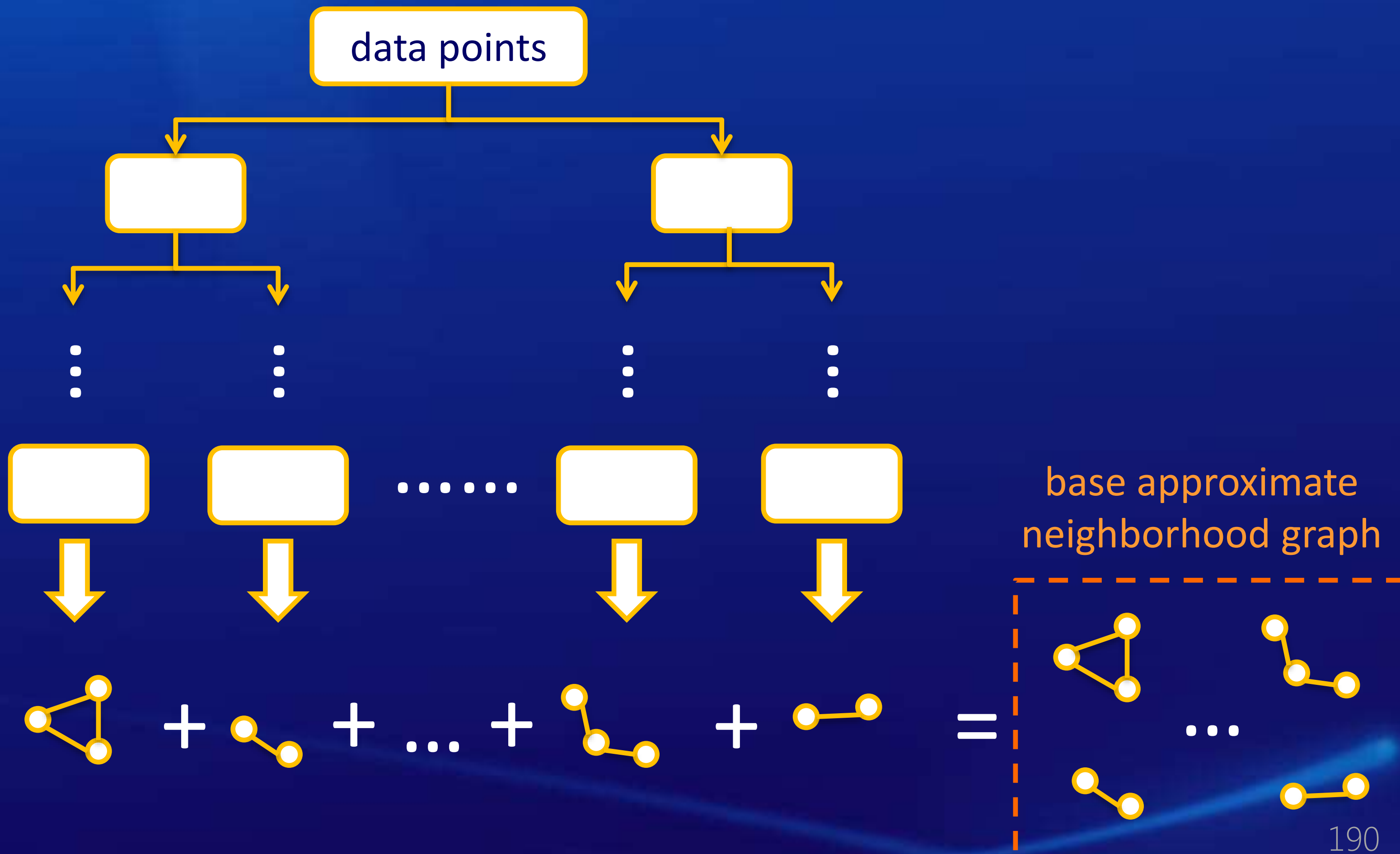
Random divide-and-conquer

- Hierarchically binary partition
  - Divide the data points into isolated subsets
  - Build the neighborhood graph for each subset
  - Achieve a base approximate neighborhood graph

- Partition hyperplane
  - Randomized principal directions – reduce the diameter of each subset efficient
  - Fast computed using the Lanczos algorithm

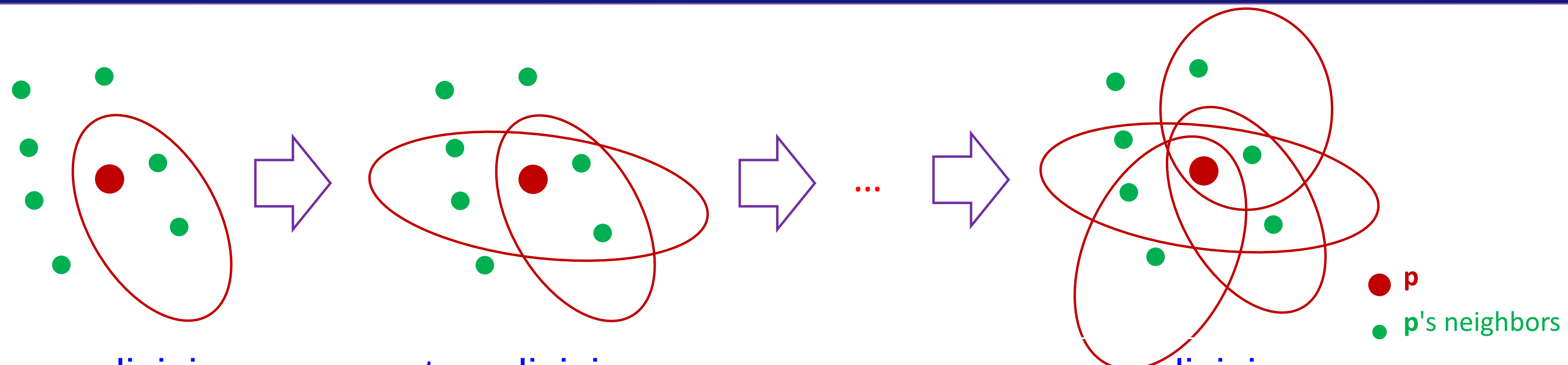
Illustration of a random division

A base approximate neighborhood graph by random division



Multiple random divide-and-conquer

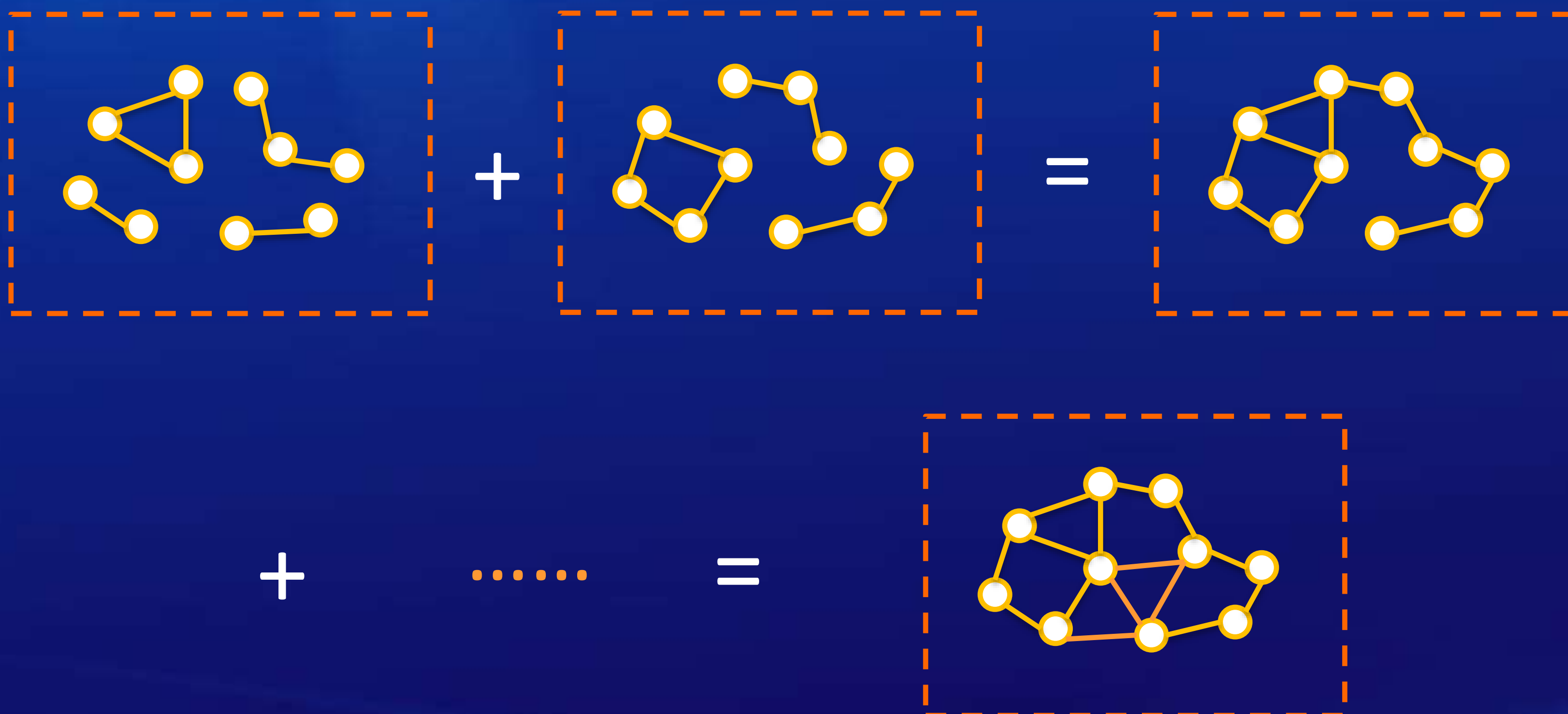
- Assemble base approximate neighborhood graphs from multiple random divisions
- Consider a point p
  - For each division, some neighbors of p lie in the same subset with p
  - With more division, more neighbors are covered



one division two divisions more divisions

Illustration of multiple divisions

Assemble multiple base neighborhood graphs

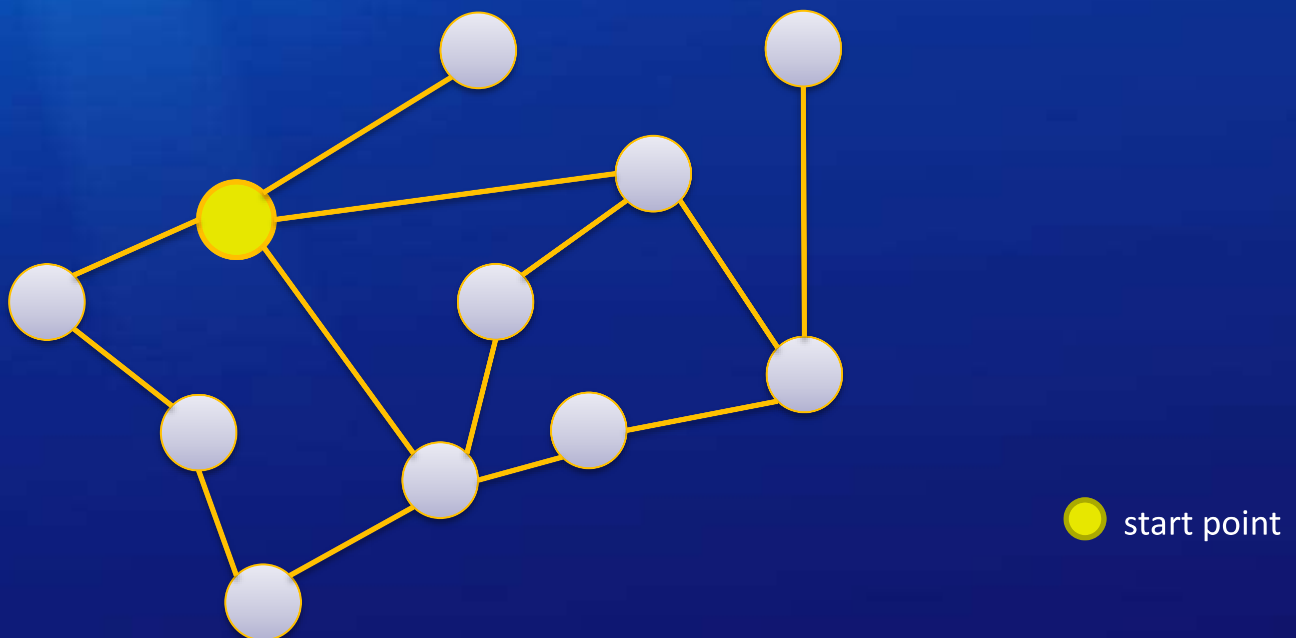


Neighborhood propagation

- Neighborhood propagation
  - Fast neighborhood refinement
  - For each point, expand its neighborhoods in a best-first manner so that the propagation path is towards its true nearest neighbors

- Hybrid neighborhood propagation
  - Computing the distance from both visual and textual information
  - The distance is positive infinity if two images don't share any common tag

Illustration of neighborhood propagation



priority queue

Illustration of neighborhood propagation

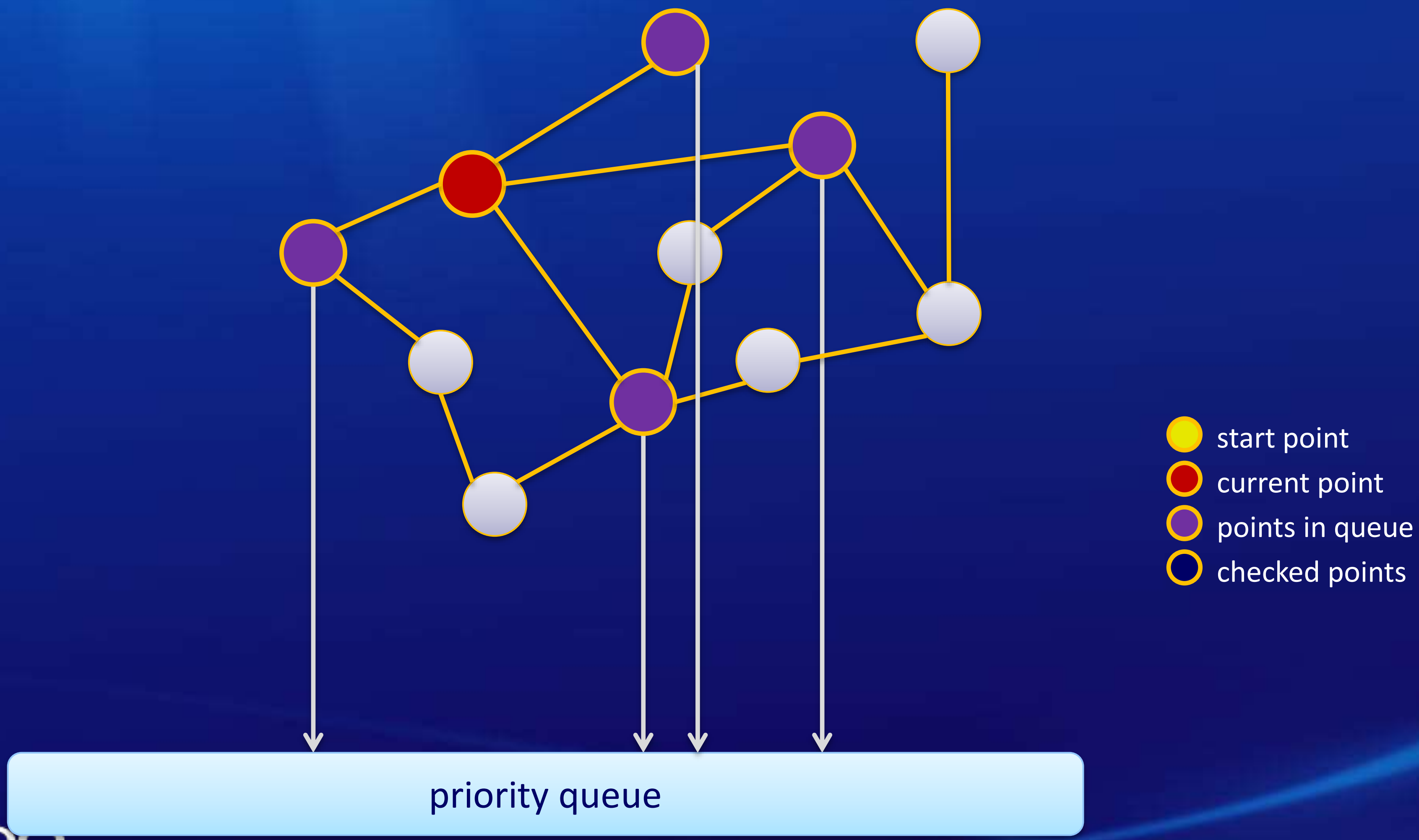


Illustration of neighborhood propagation

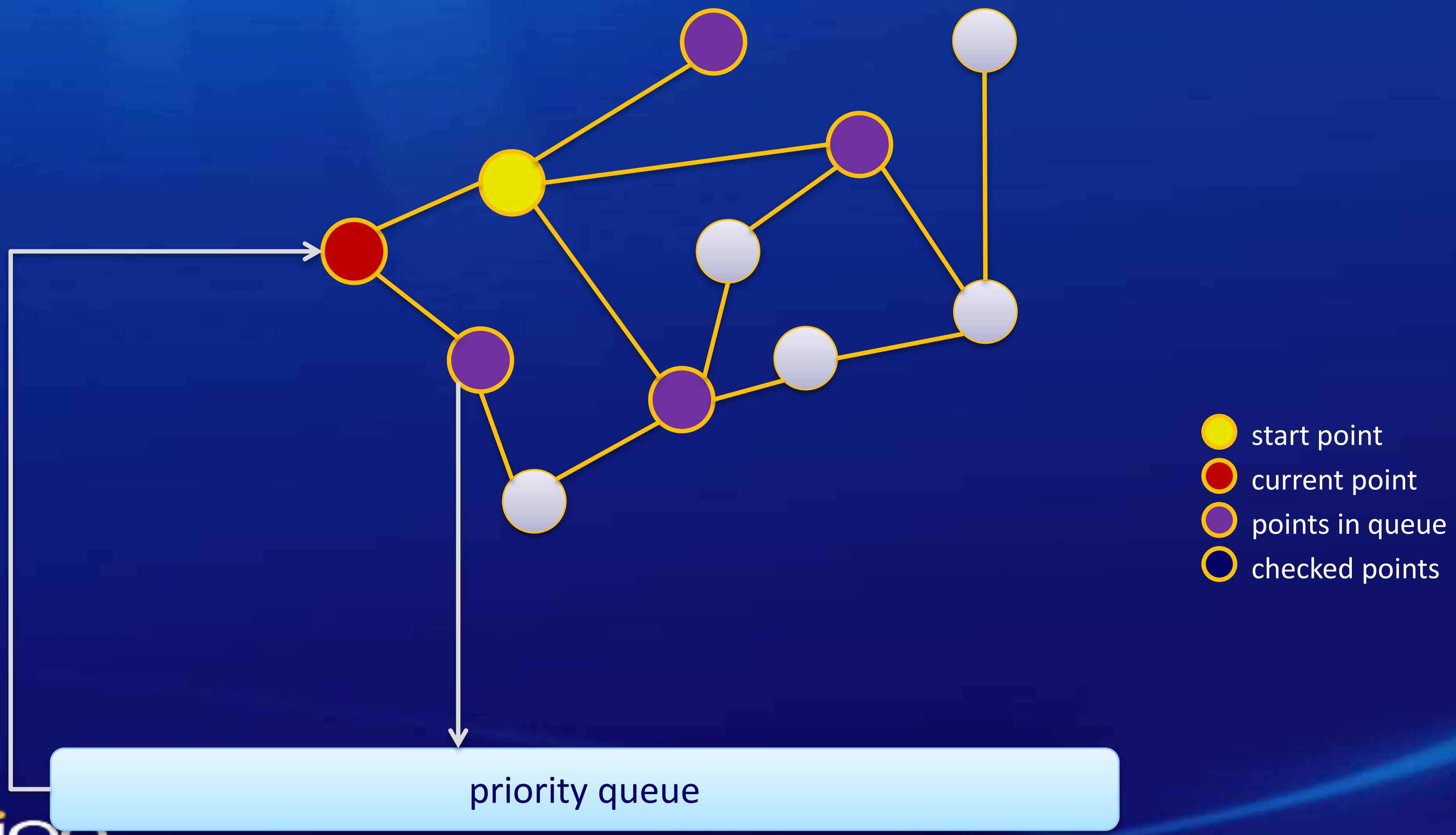


Illustration of neighborhood propagation

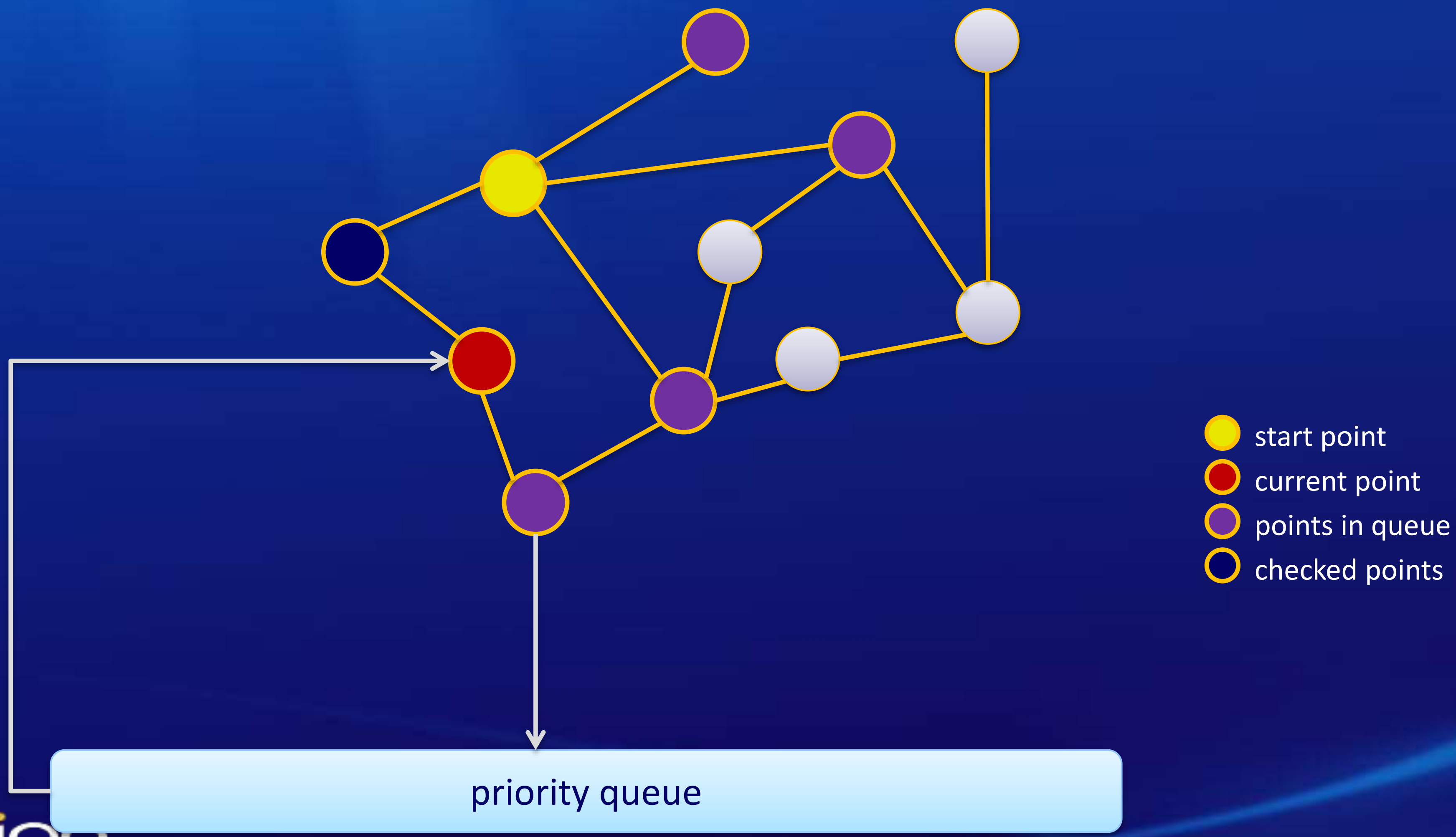


Illustration of neighborhood propagation

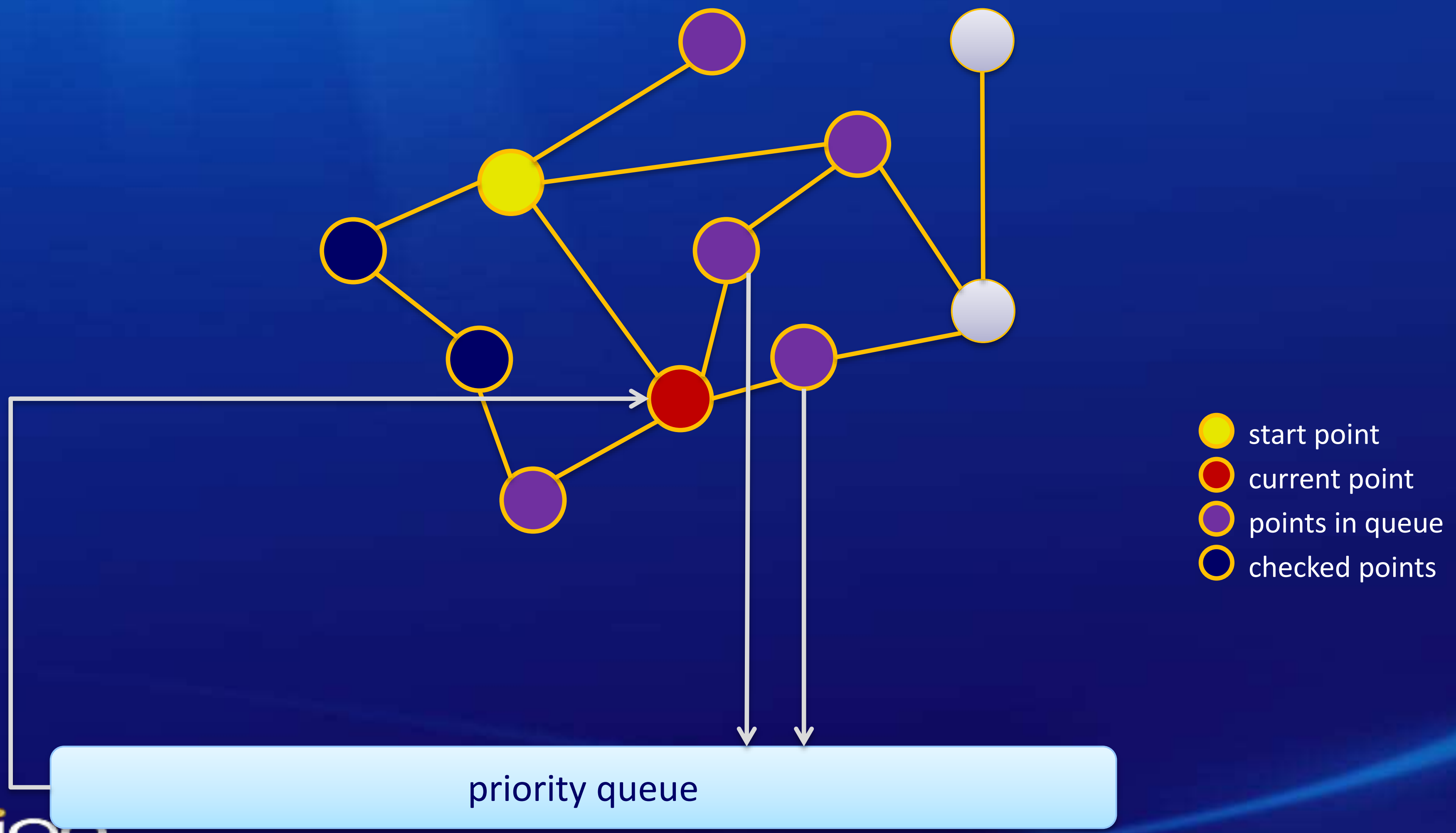
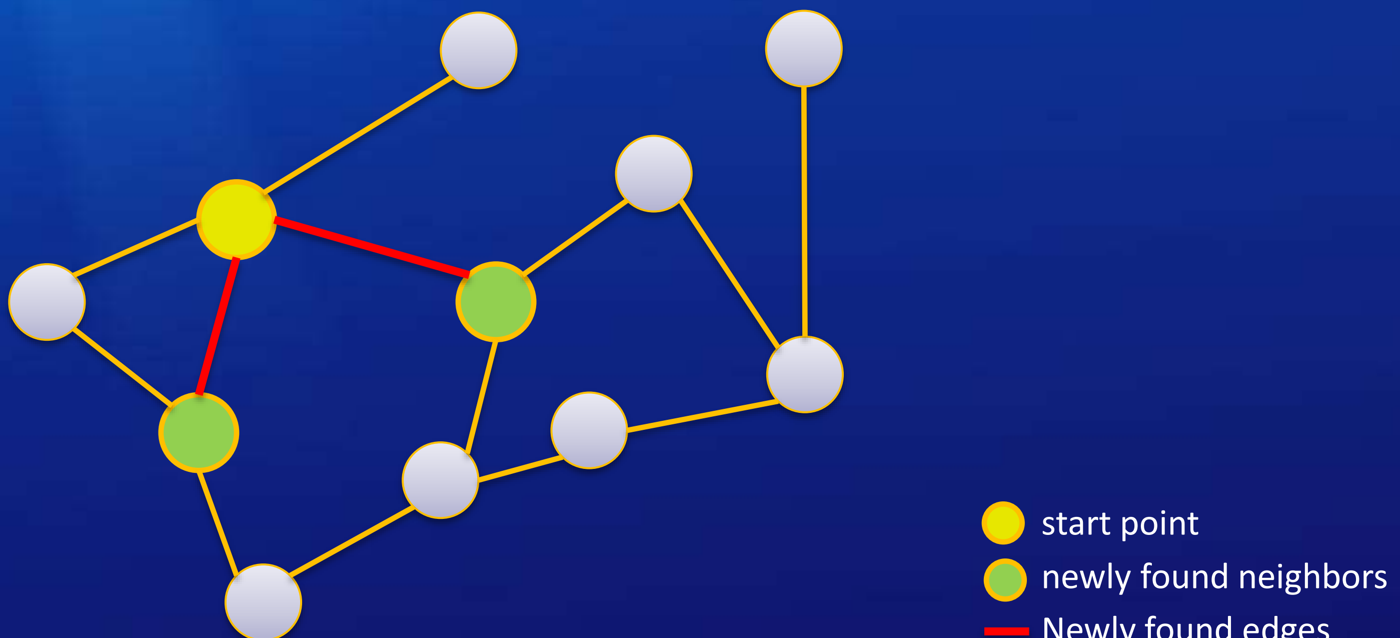


Illustration of neighborhood propagation - final



bing

Quantitative evaluation

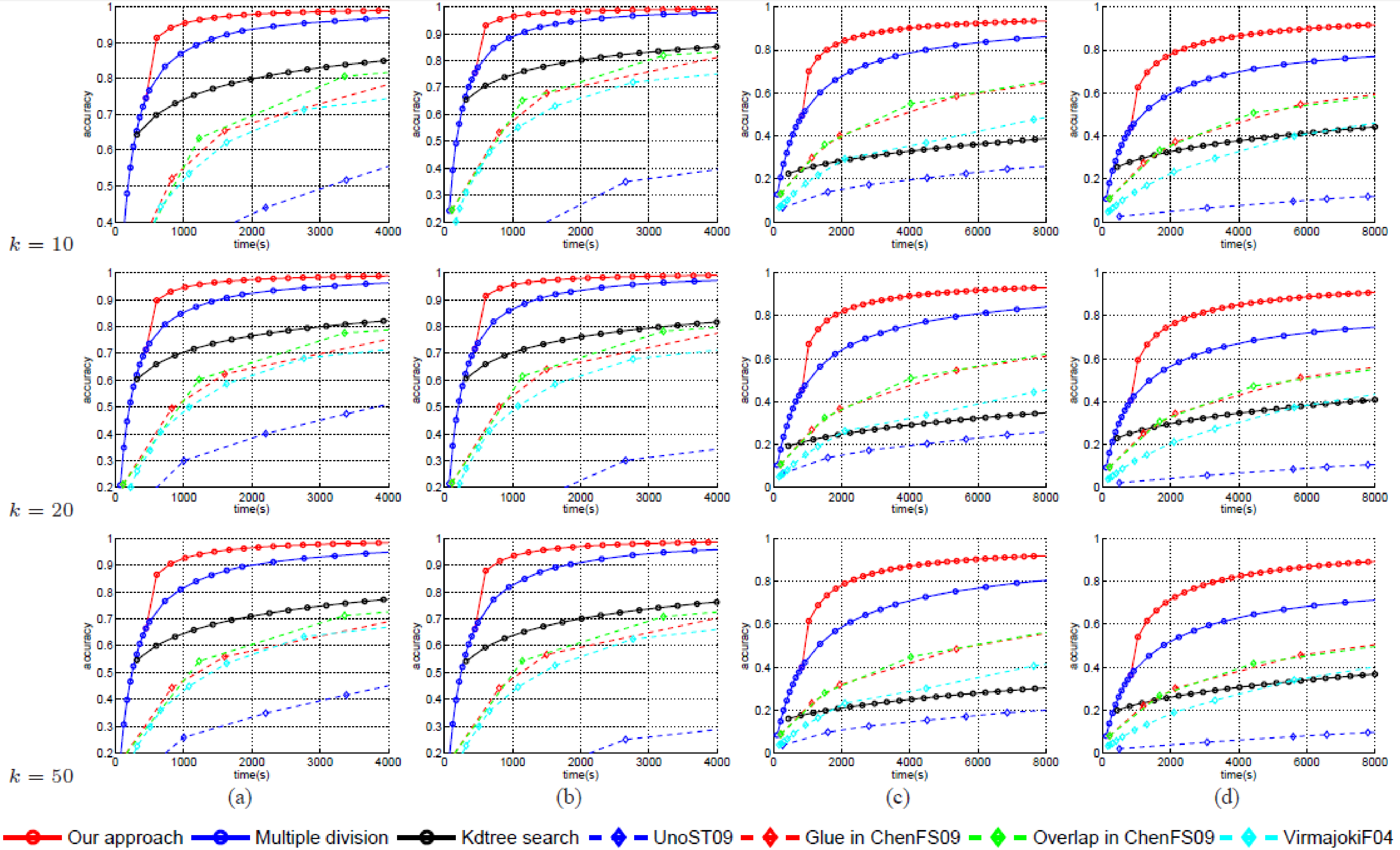
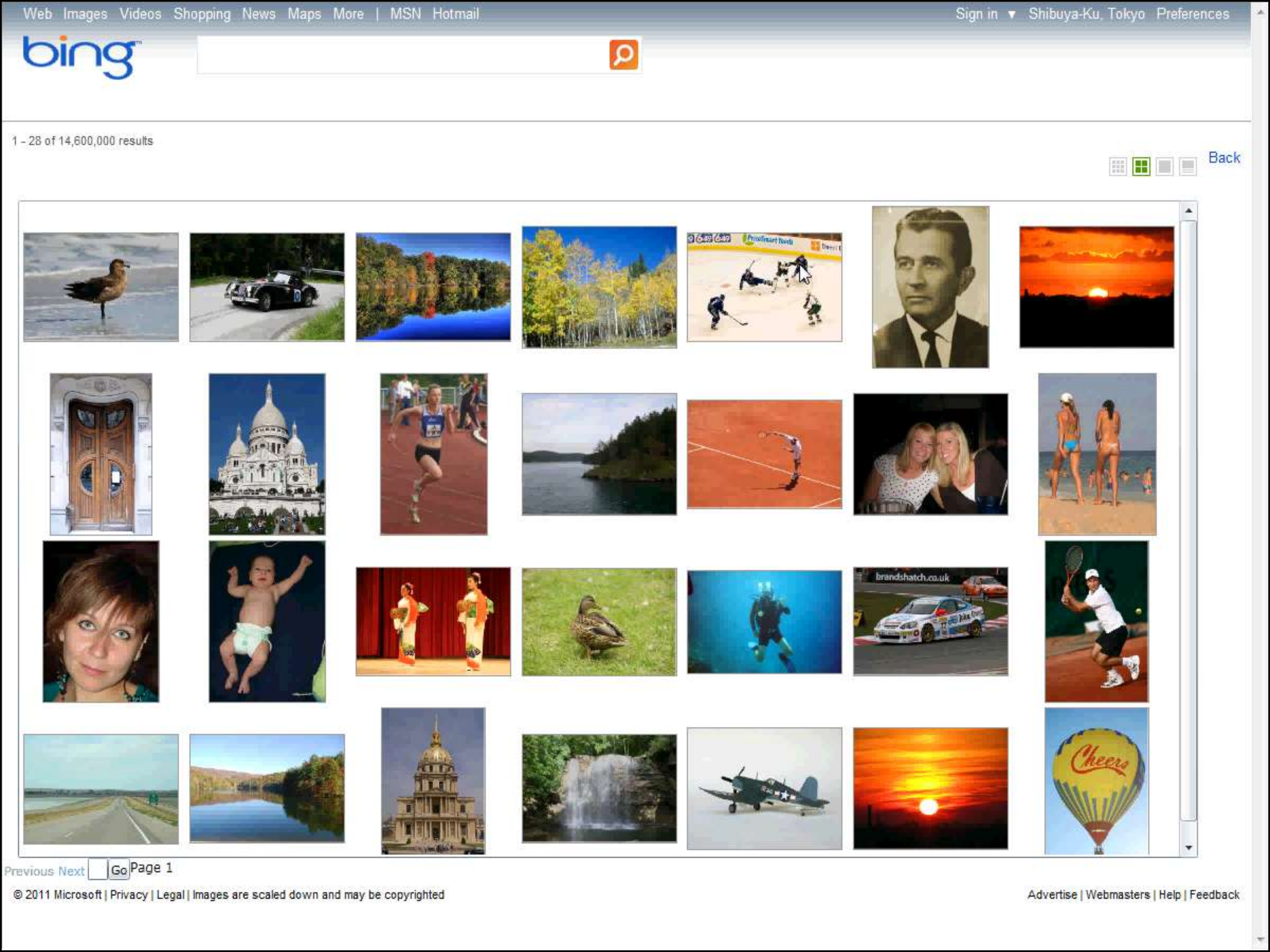


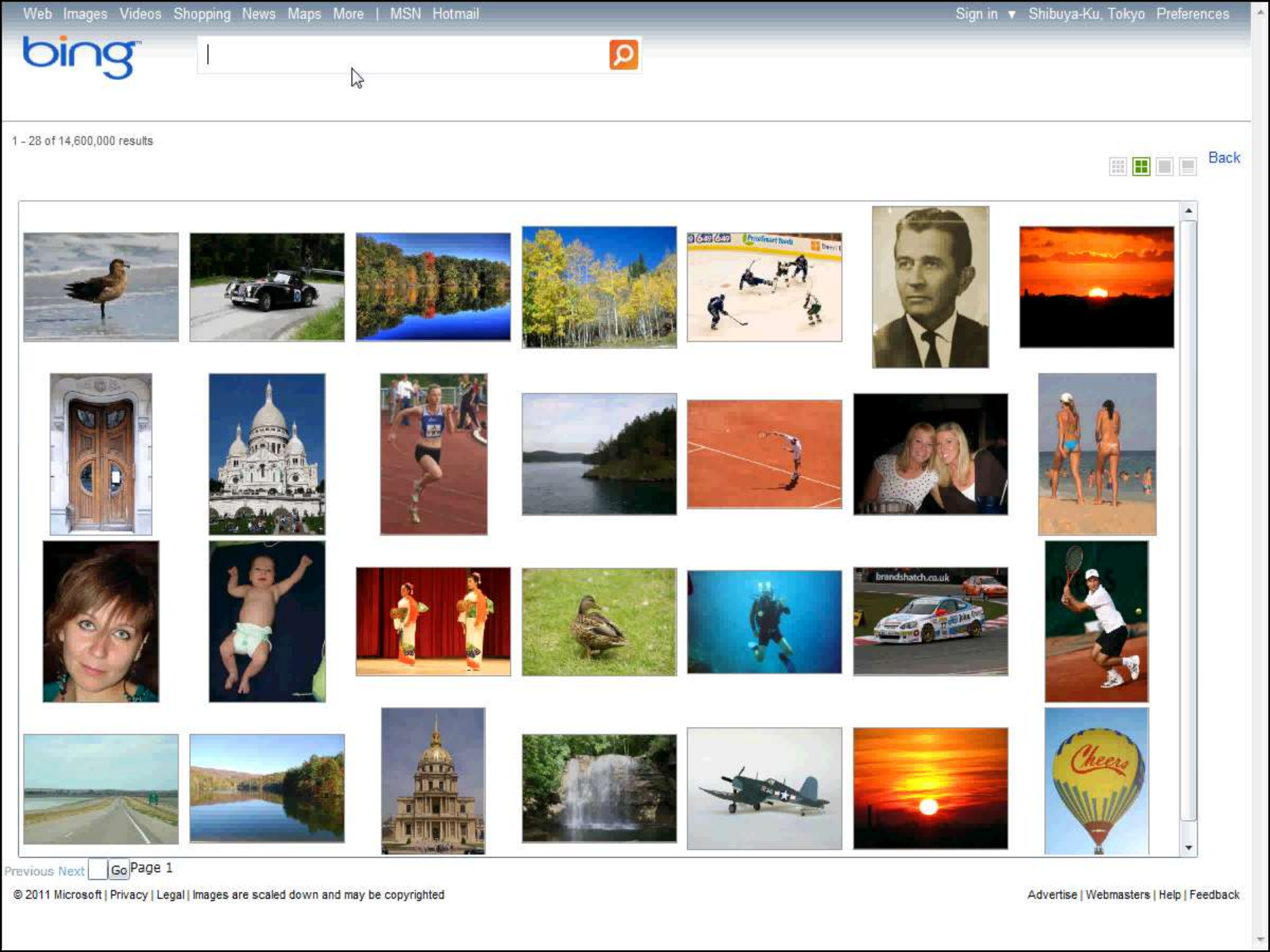
Figure 5. Performance comparison on (a) Caltech 101, (b) Recognition Benchmark, (c) Imagenet, (d) TinyImage

Demo

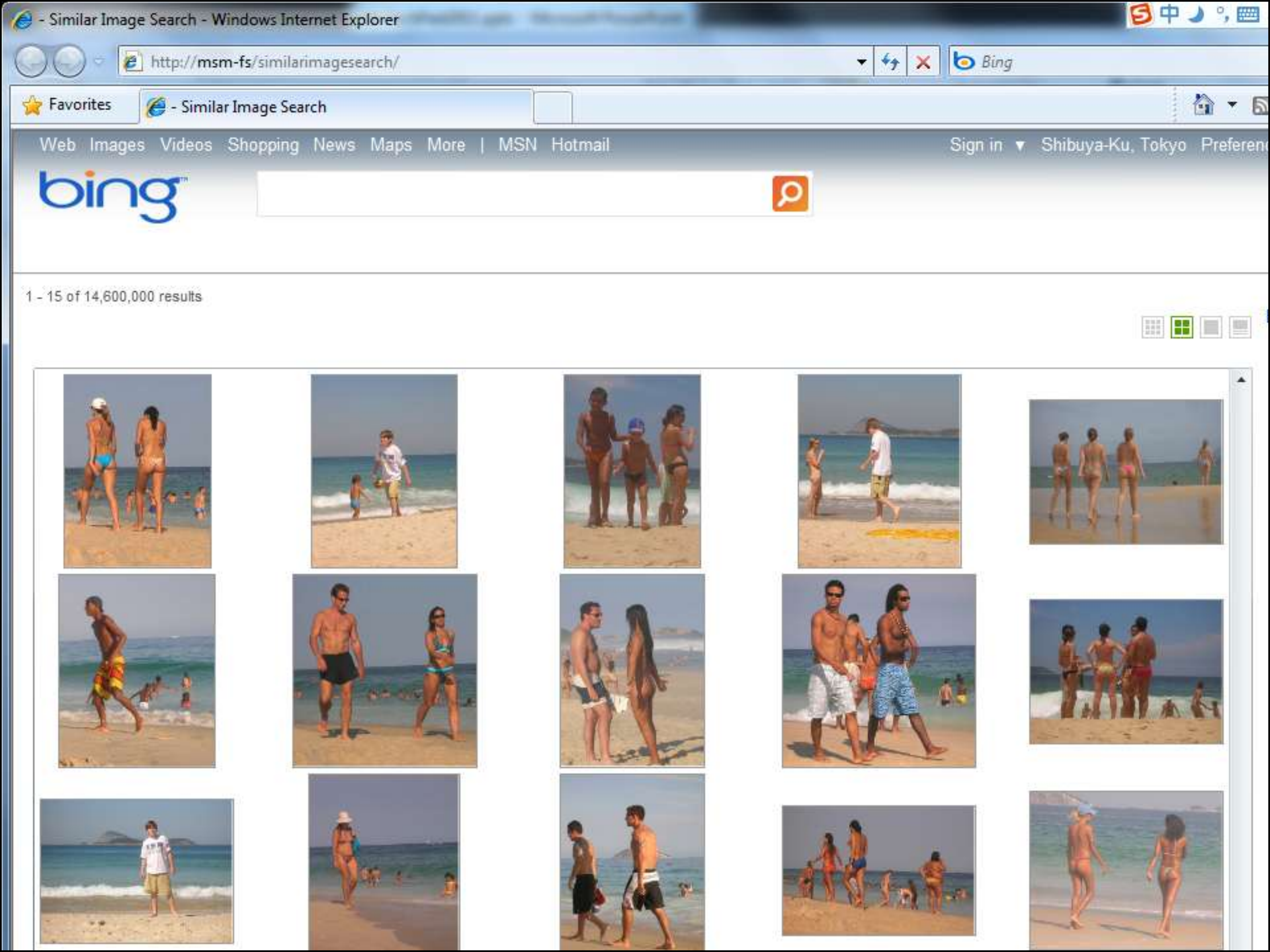
- Data set
  - 3M flickr images, storage ~400G
  - ~5000 tags
  - ~20 tags each image

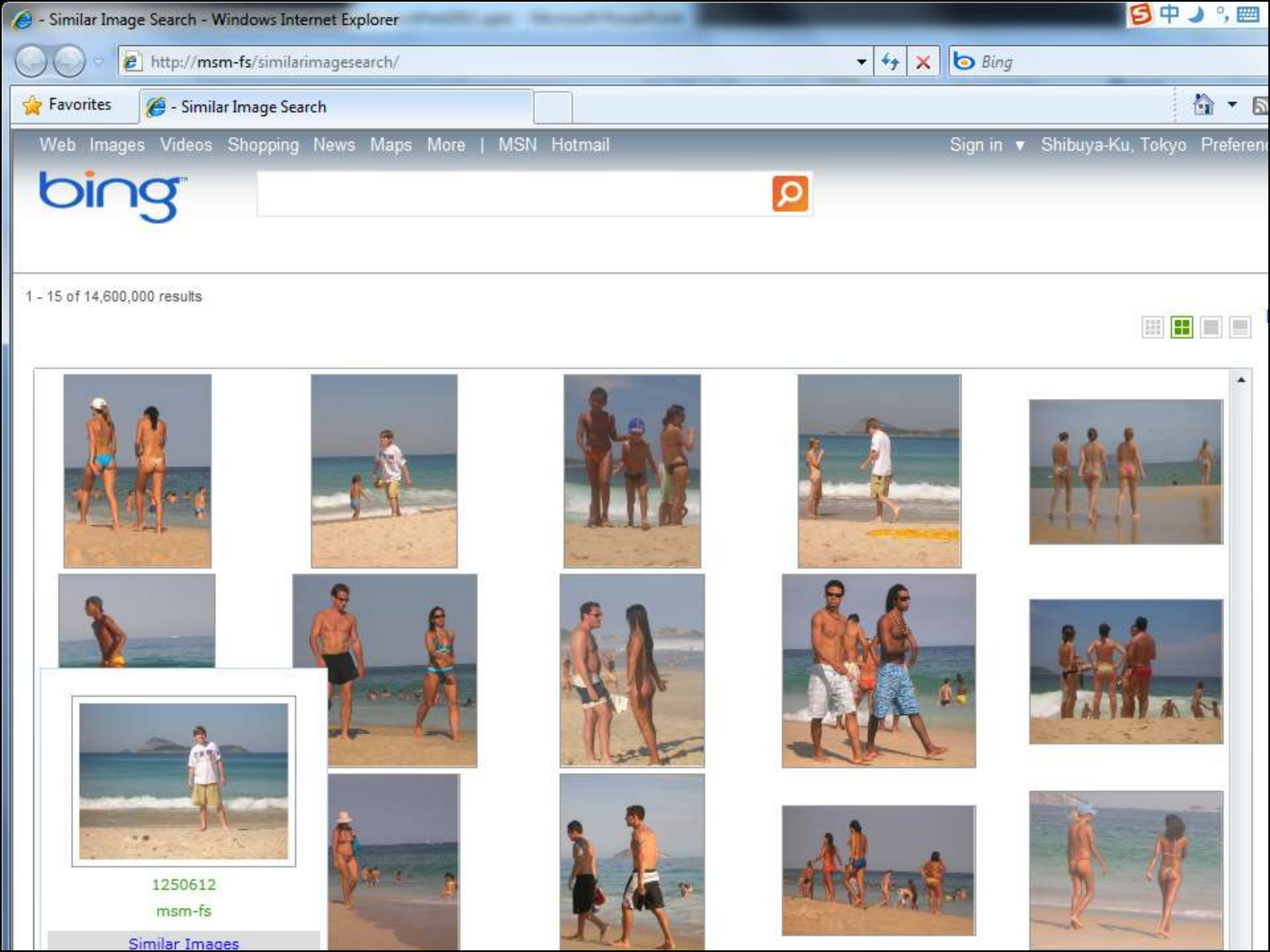
- Storage
  - Index structure ~500M (can be reduced to ~300M)
  - GIST ~1G
  - Color ~500M

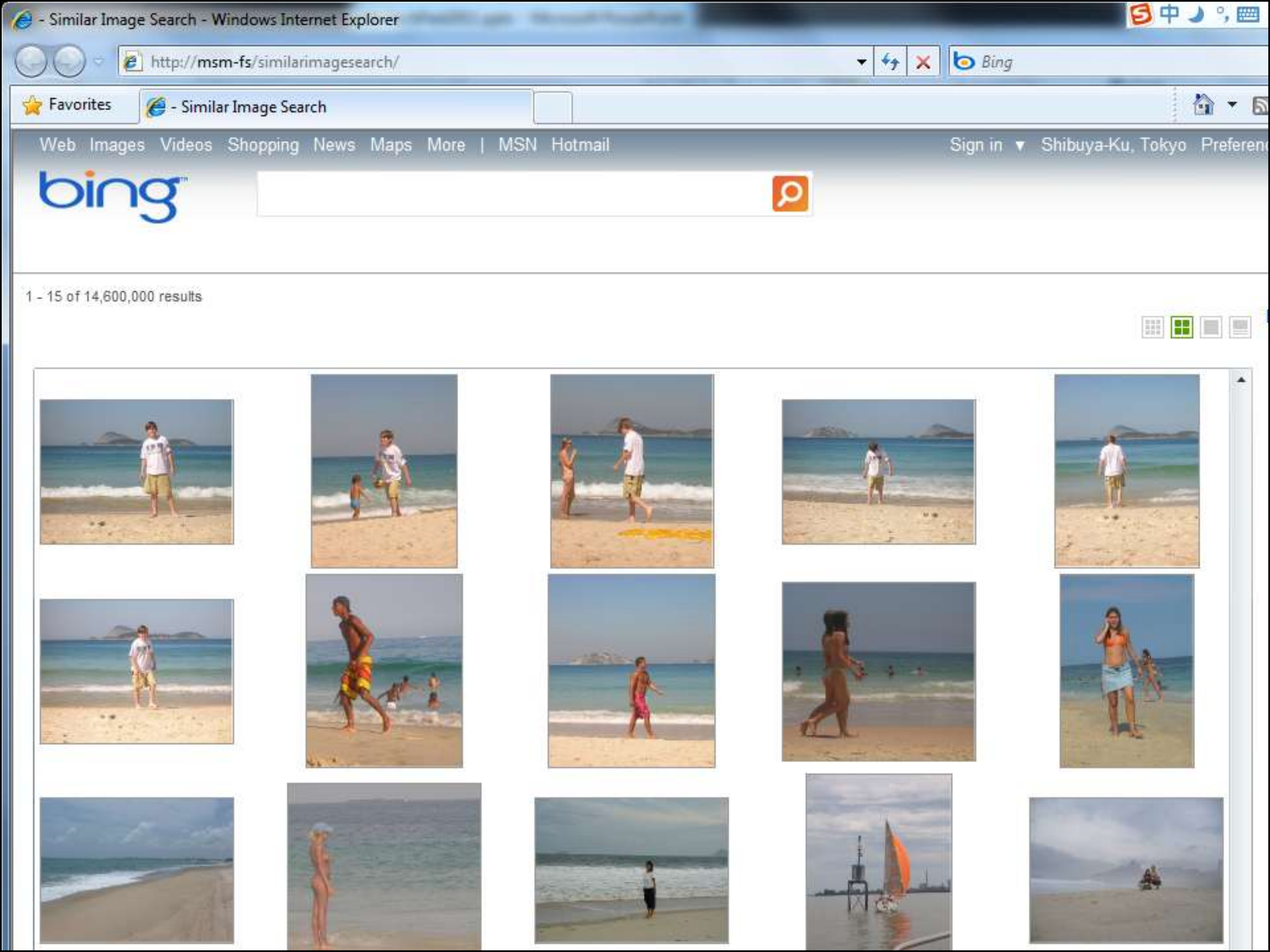

Web Images Videos Shopping News Maps More | MSN Hotmail
Sign in ▼ Shibuya-Ku, Tokyo Preferences
bing
1 - 28 of 14,600,000 results
Back
brandshatch.co.uk
Cheers
Previous Next Go Page 1
© 2011 Microsoft | Privacy | Legal | Images are scaled down and may be copyrighted
Advertise | Webmasters | Help | Feedback


Web Images Videos Shopping News Maps More | MSN Hotmail
Sign in ▾ Shibuya-Ku, Tokyo Preferences
bing
1 - 28 of 14,600,000 results
Back
brandshatch.co.uk
Cheers
Previous Next Go Page 1
© 2011 Microsoft | Privacy | Legal | Images are scaled down and may be copyrighted
Advertise | Webmasters | Help | Feedback

Real Demo


- Similar Image Search - Windows Internet Explorer
http://msm-fs/similarimagesearch/
Bing
Favorites
- Similar Image Search
Web Images Videos Shopping News Maps More | MSN Hotmail
Sign in
Shibuya-Ku, Tokyo
bing
1 - 15 of 14,600,000 results


- Similar Image Search - Windows Internet Explorer
http://msm-fs/similarimagesearch/
Bing
Favorites
- Similar Image Search
Web Images Videos Shopping News Maps More | MSN Hotmail
Sign in ▼ Shibuya-Ku, Tokyo
bing
1 - 15 of 14,600,000 results
1250612
msm-fs
Similar Images


- Similar Image Search - Windows Internet Explorer
http://msm-fs/similarimagesearch/
Bing
Favorites
- Similar Image Search
Web Images Videos Shopping News Maps More | MSN Hotmail
Sign in
Shibuya-Ku, Tokyo
bing
1 - 15 of 14,600,000 results

Not Finished Yet

CBIR-Based Tagging

+

Large-Scale Similarity Graph

bing

Test

- How duplicates help tagging?
- Why building and using large-scale similarity graph are difficult? How to do it?

bing

Outline

Session	Time	Topic
0	09:30 – 09:40	Introduction
1	09:40 – 10:20	Learning-Based Tagging
2	10:20 – 11:40	Social Tagging and Tag Processing
		Including a break (10:45 – 11:00)
3	11:40 – 12:10	Data-Driven Web-Scale Tagging
4	12:10 – 12:30	Future Directions/QA

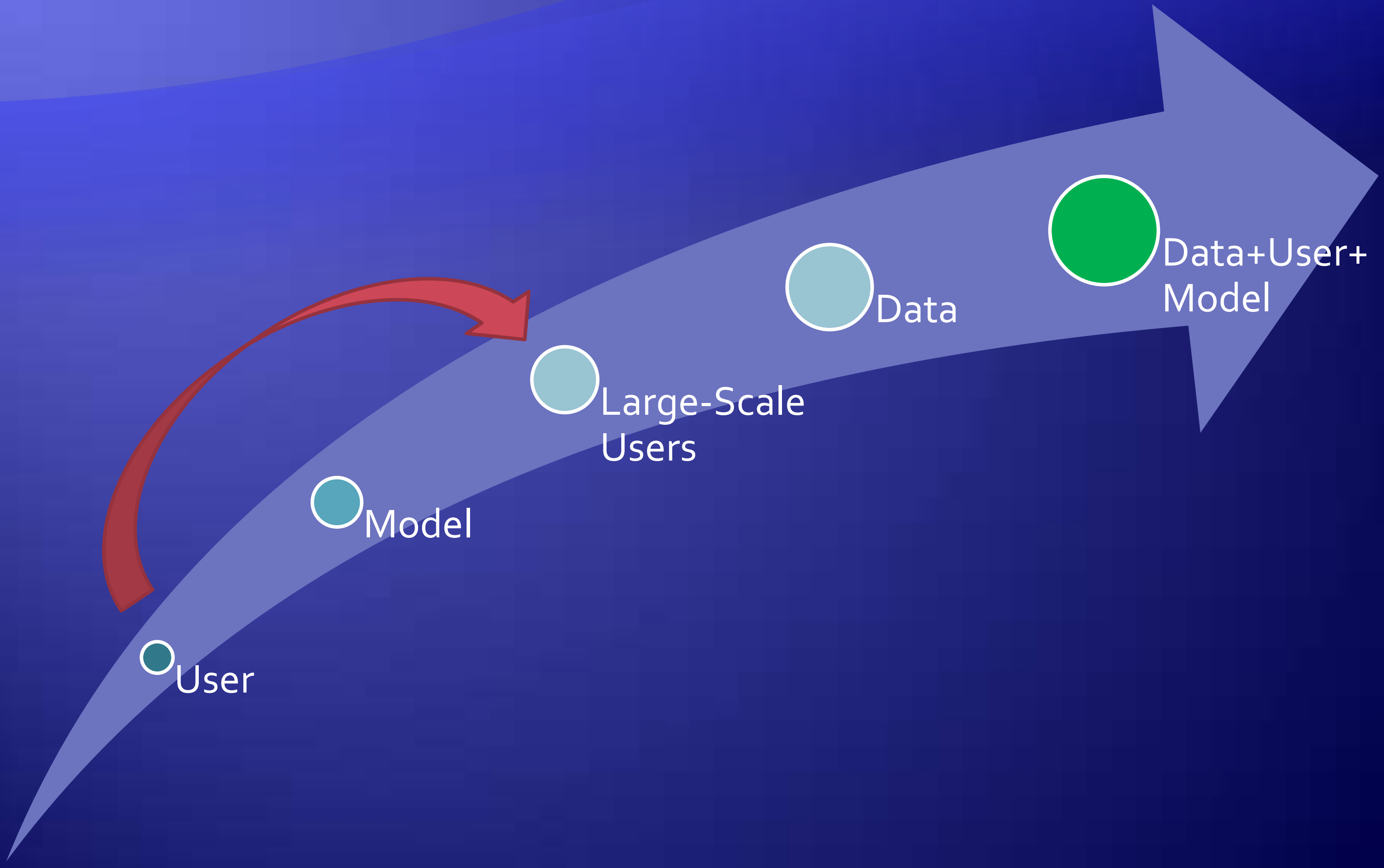
bing

Outline

Session	Time	Topic
0	09:30 – 09:40	Introduction
1	09:40 – 10:20	Learning-Based Tagging
2	10:20 – 11:40	Social Tagging and Tag Processing
		Including a break (10:45 – 11:00)
3	11:40 – 12:10	Data-Driven Web-Scale Tagging
4	12:10 – 12:30	Future Directions/QA

bing

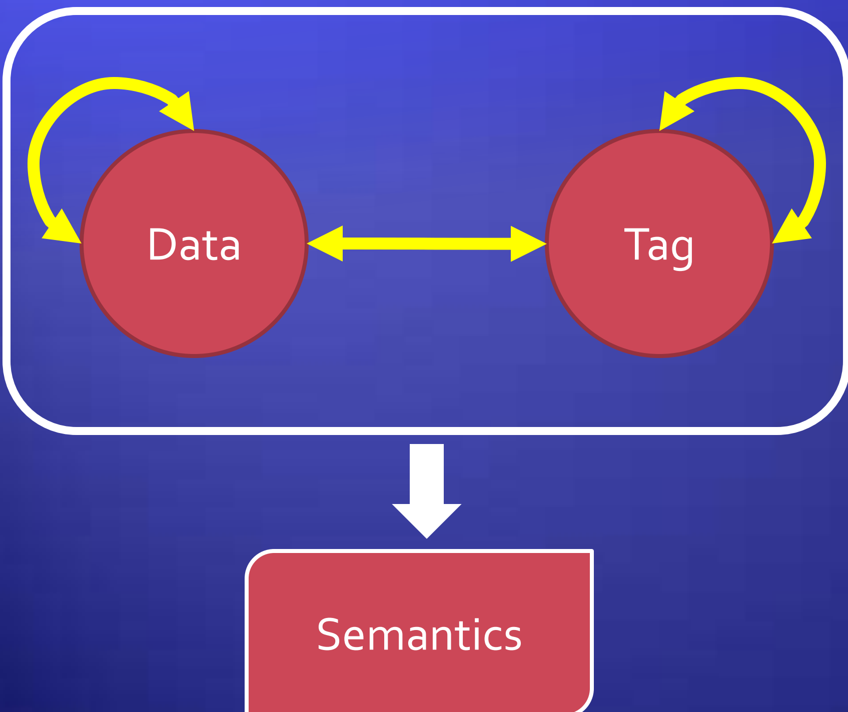
Evolution of Extracting Semantics



Idea Solution

- Data + User + Model

Rethinking Bridging Semantic Gap



Basic Assumptions

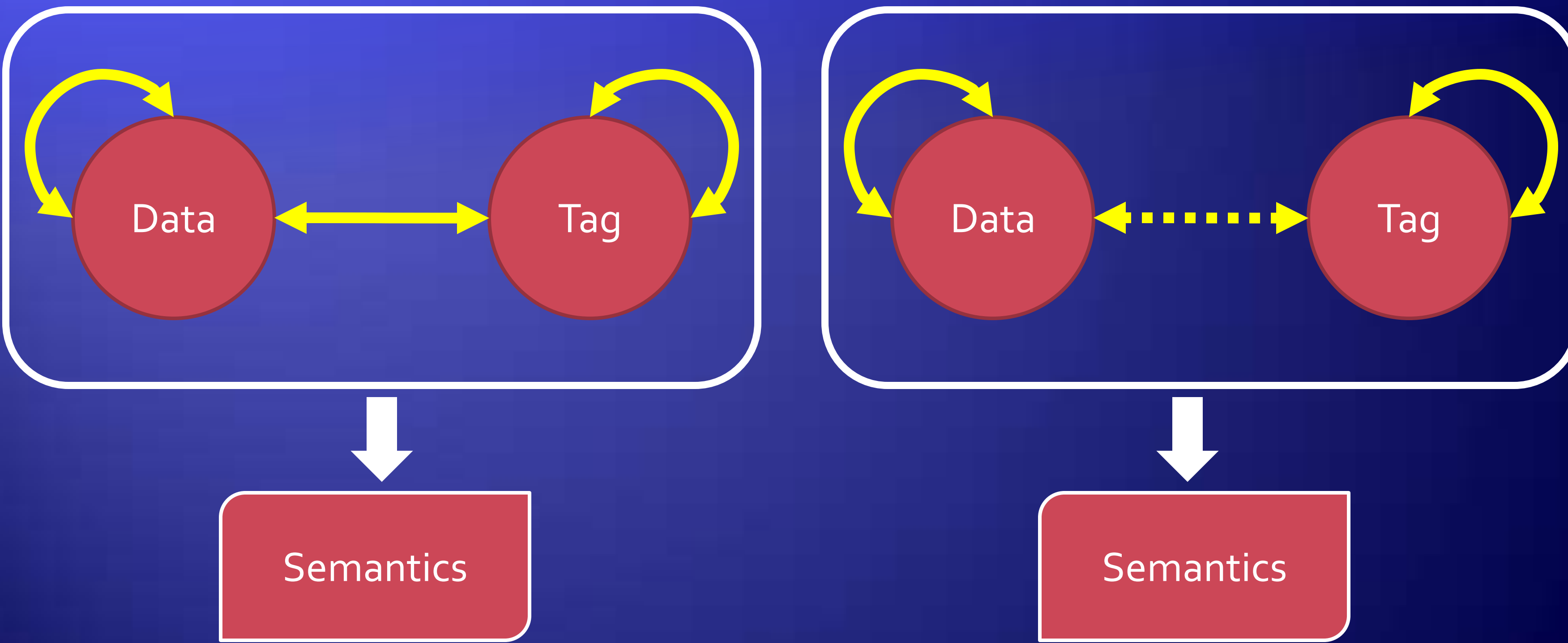
of all automatic and semi-automatic approaches

(model driven / data driven approaches)

1. Images have correlations
2. Tags have correlations
3. Tags have correlation with image content ?

bing

Rethinking Bridging Semantic Gap



For "Model-able" Tags

For "Un-model-able" Tags

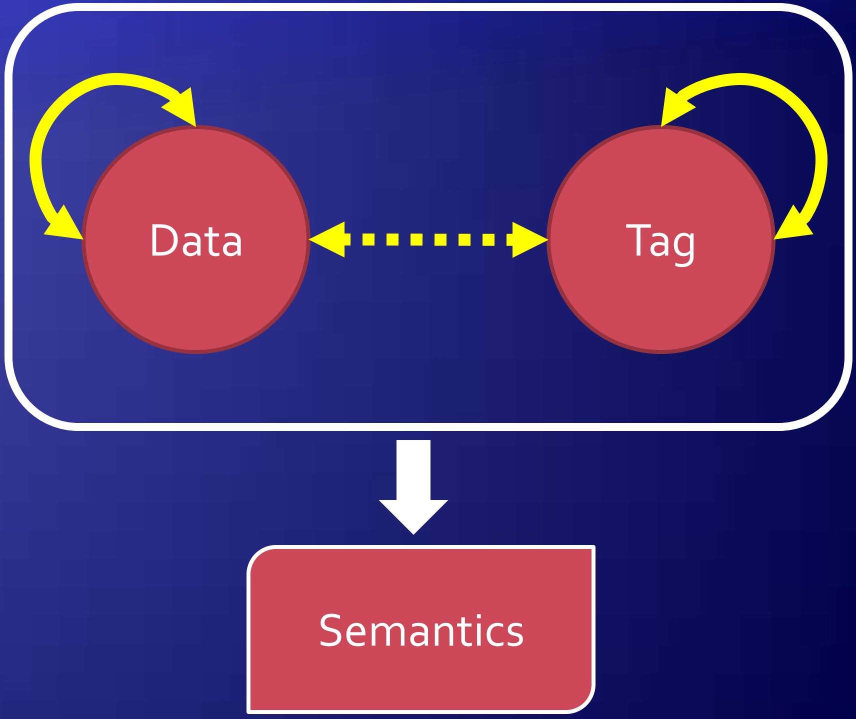
Rethinking Bridging Semantic Gap

We need

- Large amounts of tagged data
- Large amounts of users to do labeling
- **"CrowdSourcing"**

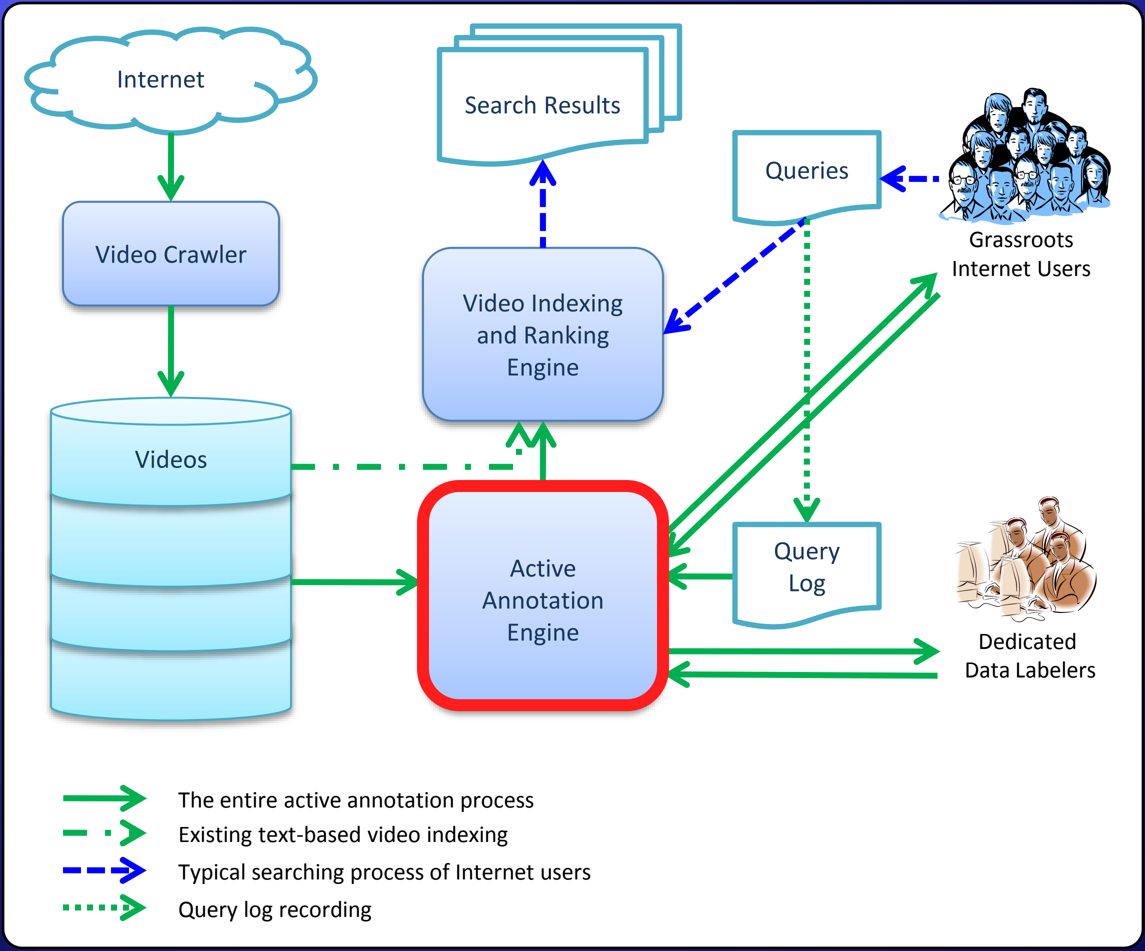
Can be regarded as a combination of

- Manual labeling
- Model based annotation
- Data driven tag processing



A Preliminary Attempt

- Online Active Annotation Framework [Qi, Hua et al, PAMI09, ACM MM 08]



- Pay for labeling (MTurk)
- Online game (ESP, …)
- reCAPTCHA
- Implicit approaches

56 million online game users in China

Increasing faster than Internet users

0.25 million online game developer

2008 revenue: 2.7 billion USD

BUT,

1/3 are non-workers and

others' monthly income is 150 ~300USD

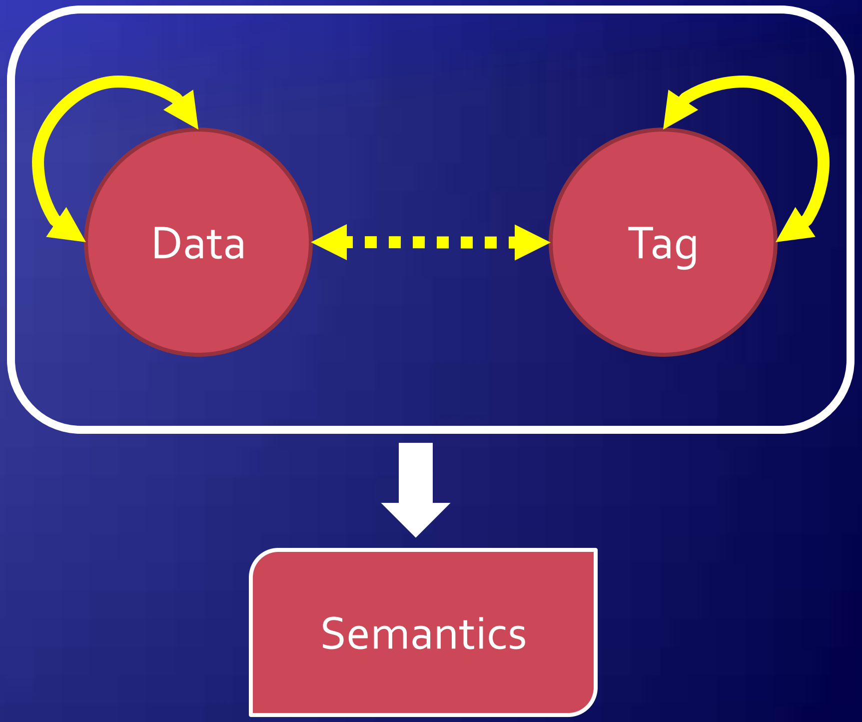
bing

Rethinking Bridging Semantic Gap

Research Challenges:

- Large-scale content-based indexing
- Large-scale active learning
- Large-scale online learning

- Labeling psychology and incentive
- Labeling quality estimation/evaluation
- Labeling Interface
- Anti-spam/cheating in labeling
- ......



bing

Future Directions

- More proactively collect data
- More aggressively leverage users
- Continuing using models and data mining techniques

bing

Thank You