

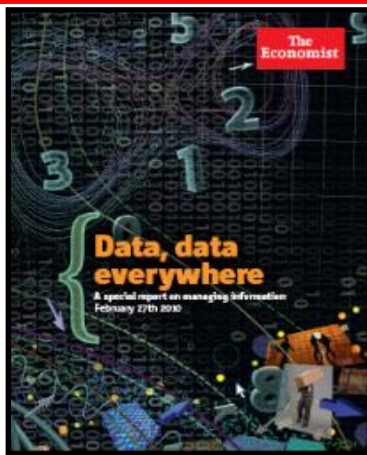
The Anatomy of Data Driven Learning

Gui-Rong Xue



Data, data everywhere

The Economist



Information has gone from scarce to superabundant. That brings huge new benefits, says Kenneth Cukier—but also big headaches

信息从稀缺走向大规模化，
带来便利同时也导致大问题

2010《经济学人》“数据、
无处不在的数据”专刊



每秒产生数据40TB



每天10TB交易数据



4亿用户, 400亿照片

What can big data tell us ?

- Three Examples
 - Machine Translation
 - PageRank
 - N-gram Model

Google LDC N-Gram Corpus	
• Number of tokens:	1,024,908,267,229
Number of sentences:	95,119,665,584
Number of unigrams:	13,588,391
Number of bigrams:	314,843,401
Number of trigrams:	977,069,902
Number of fourgrams:	1,313,818,354
Number of fivegrams:	1,176,470,663

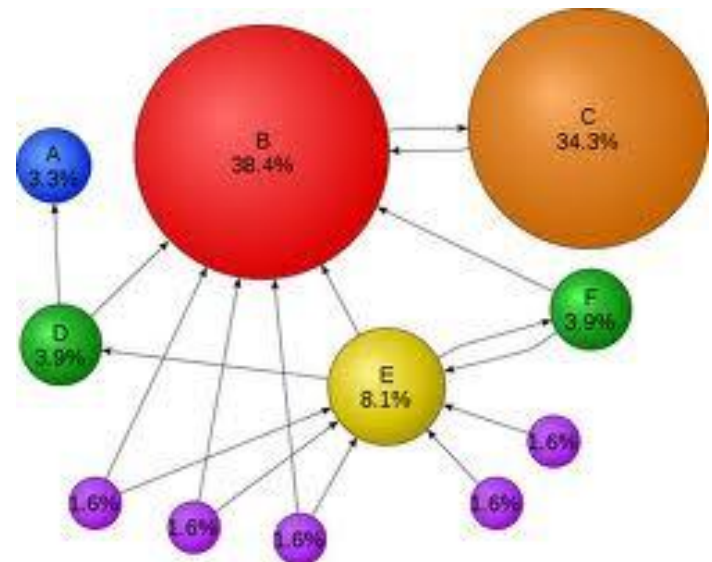
Google 翻译

源语言: 检测到中文 目标语言: 英语 翻译

联合国秘书长潘基文于10月26日开启其亚洲之旅, 期间对中国进行访问。潘基文前往北京、上海和南京等地。此外, 潘基文还出席了2010年上海世博会高峰论坛和世博会的闭幕式。

将中文译成英语

UN Secretary-General Ban Ki-moon opened on October 26 its Asian tour, during a visit to China. Ban Ki-moon to Beijing, Shanghai and Nanjing and other places. In addition, Ban Ki-moon also attended the World Expo 2010 Shanghai World Expo Forum and the closing ceremony.



Data Driven Learning

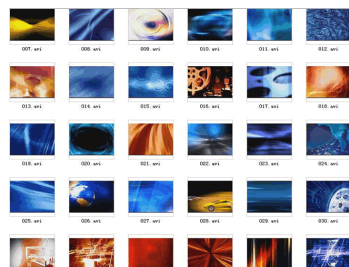
- Facing Issues for Data
 - Large Scale
 - Learning Time Cost
 - Data Sparse
 -
- Solutions
 - Cloud Computing Platform
 - Distributed Computing
 - Distributed Storage
 - GPU
 - Novel Learning Algorithms
 - Matrix Factorization
 - Transfer Learning
 -

Aliyun.com Open Data Platform

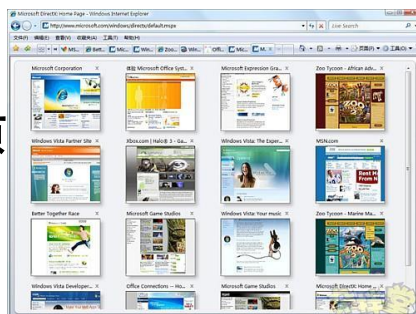
中日英三
语语料



1亿图片



10亿网页



C++
python



开放云服
务机群

计算平台

存储平台

阿里巴巴 云计算
Alibaba Cloud Computing

阿里巴巴 云计算
Alibaba Cloud Computing

A Challenge Issue for Data Driven Learning

- English

- Chinese

Labeled Data	English	Chinese
News	Reuters-21578	?
newsgroups	20 Newsgroups	?
Web pages	Open Directory Project (> 1M)	Very few ODP data (< 20k, ~ 1%)

- Text

- ODP: 4,616,309
labeled web pages

- Image

- Caltech256: 30,607
labeled images

Our Solution

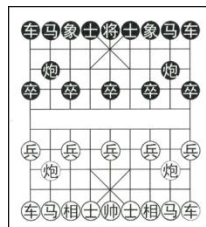
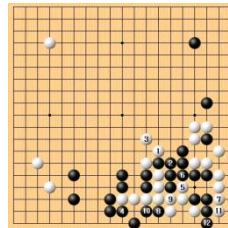
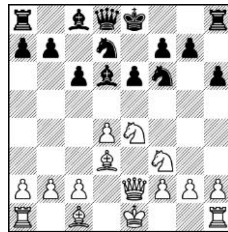
- Translated Learning
 - A variation for transfer learning

Transfer Learning



Human Learning

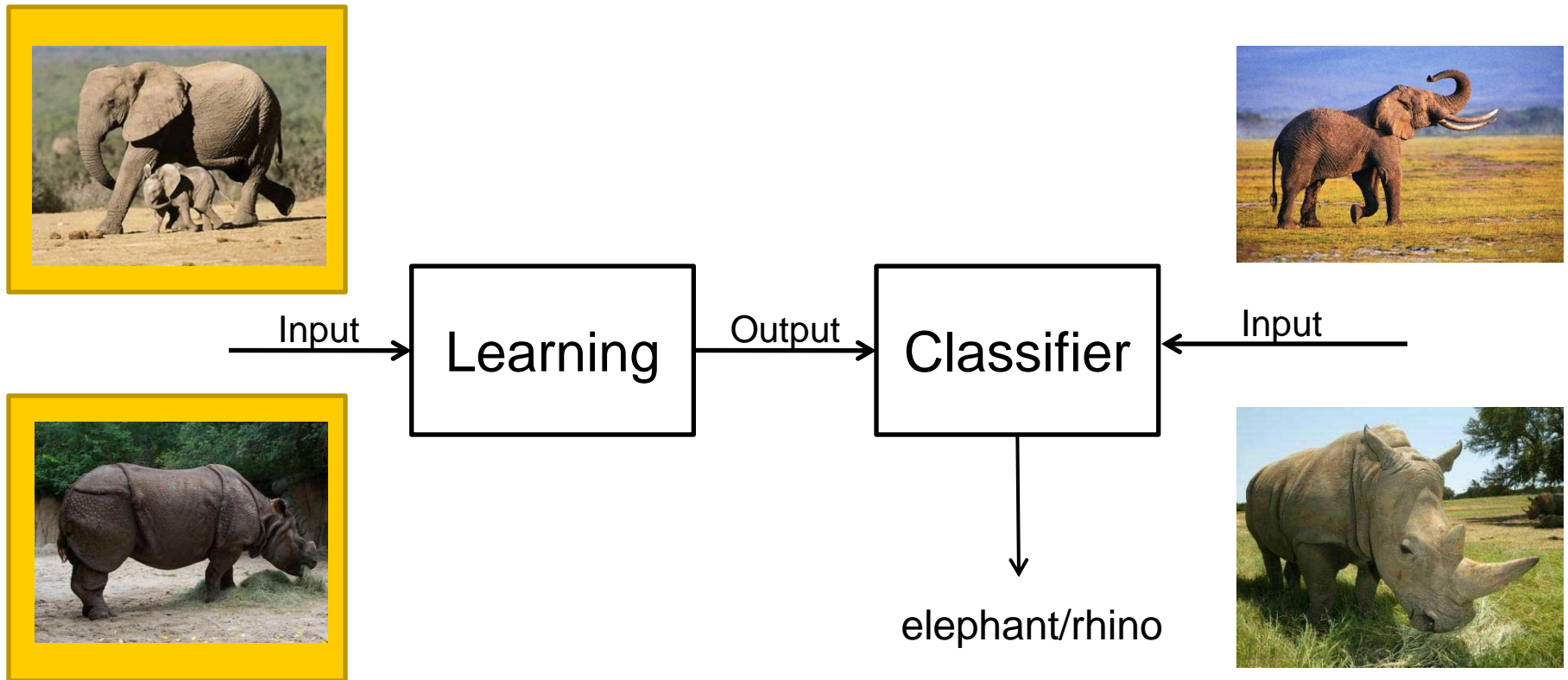
learning



solving



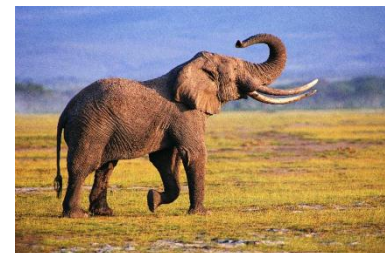
Traditional Machine Learning



Training Data

Test Data

Transfer Learning



Training Data

Auxiliary Data

Test Data

Translated Learning



Translated Learning

- Transfer Learning across Different Feature Spaces



(a) Supervised Learning



(b) Semi-supervised Learning



(c) Transfer Learning



(d) Self-taught Learning



(e) Multi-view Learning



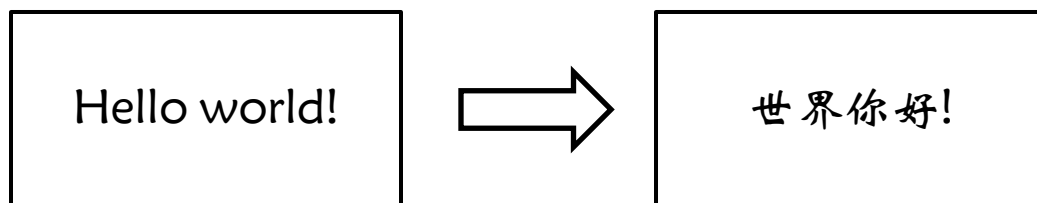
(f) Translated Learning



Test Data

Instance-level Translation

- Generally, it is difficult.
 - Easy translation



- Hard translation



Model-level Translation

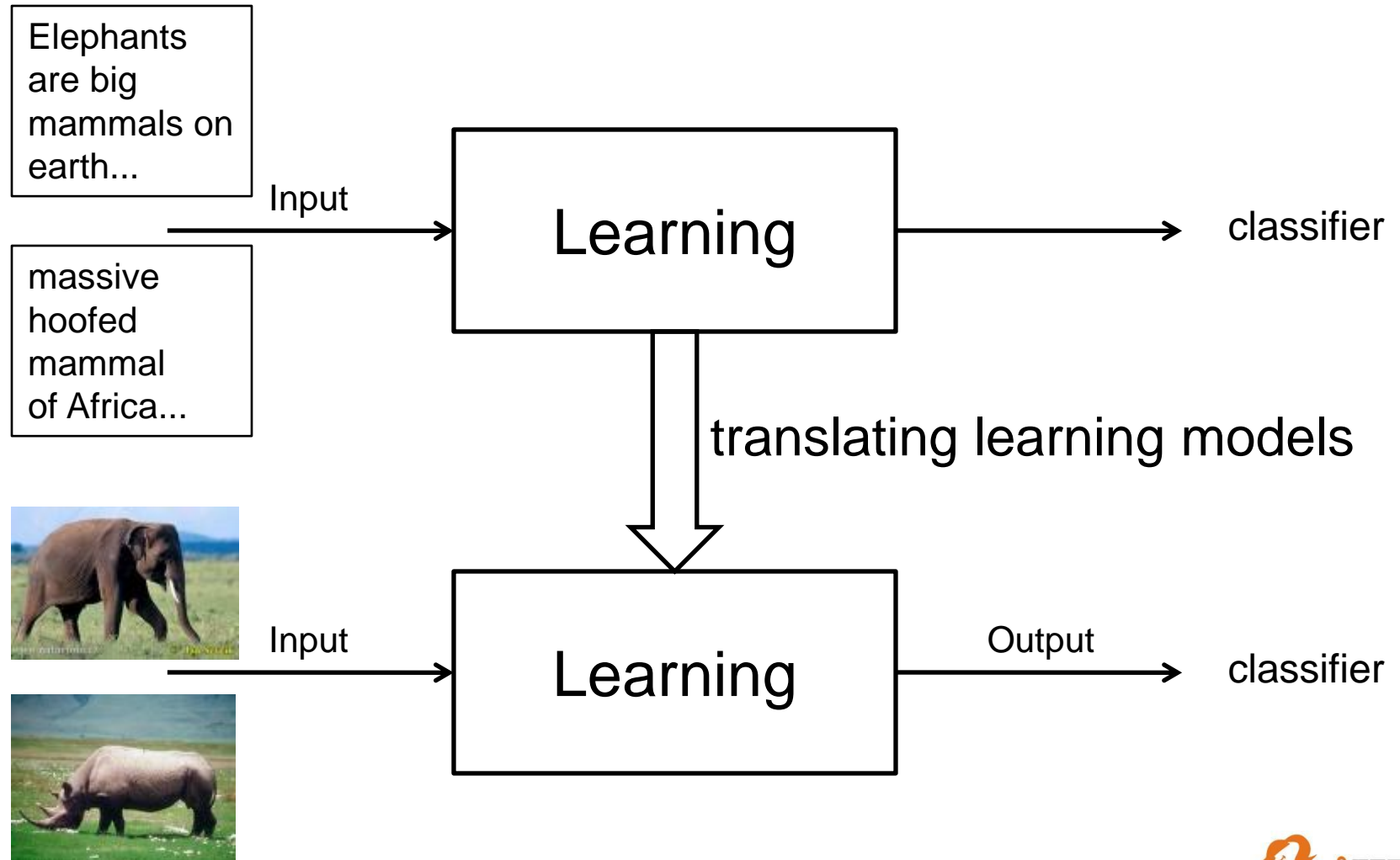
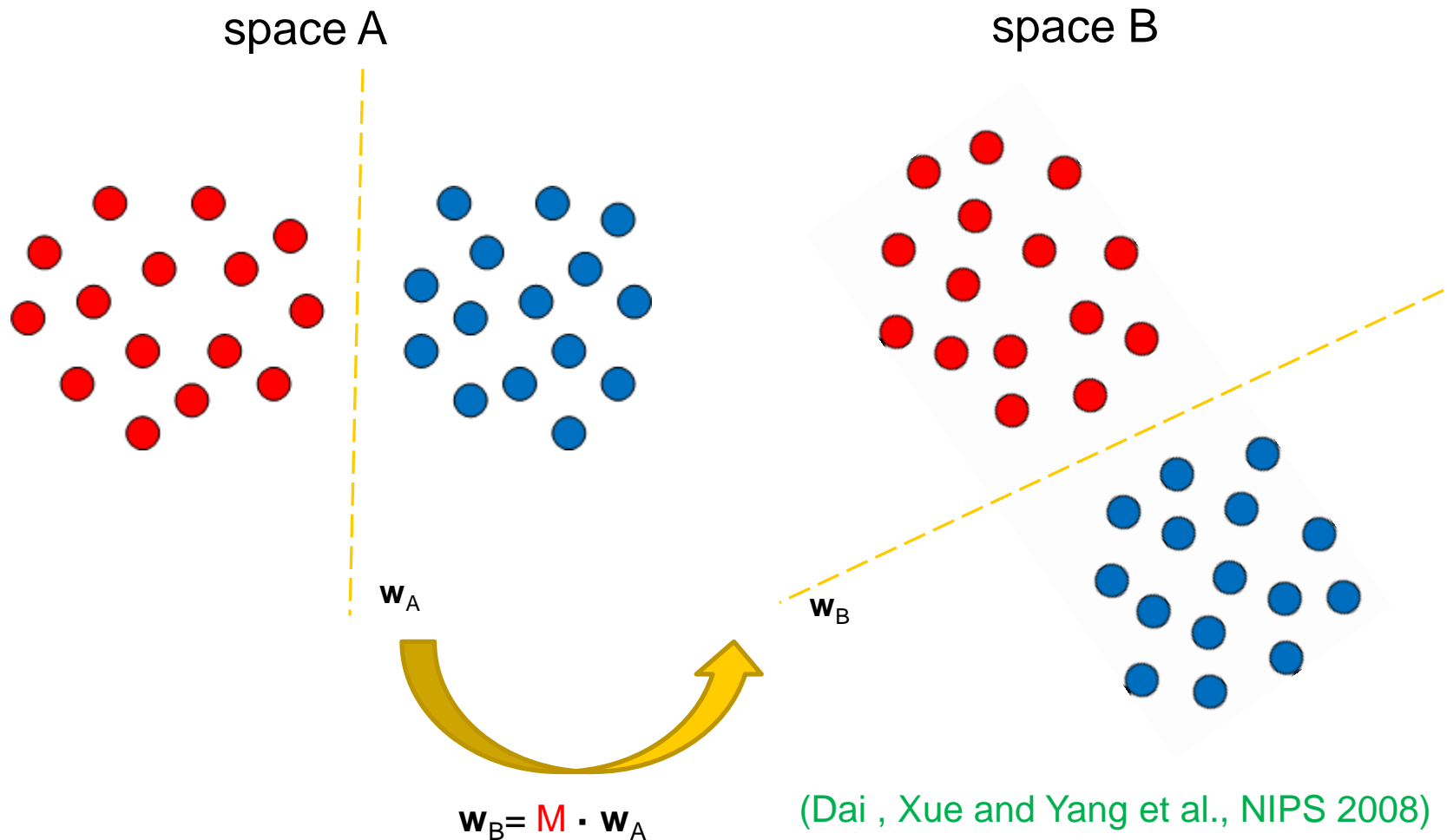


Illustration in Vector Space Model



(Dai, Xue and Yang et al., NIPS 2008)

(Yang, Chen and Xue et al, ACL 2009)

(Chen, Xue et al, APWEB 2010)

Chen, Xue, Yang et al. AAAI 2010)

Text Classification Model

quadruped ornithischian dinosaur of four long bony spikes on a flexible tail and two rows of upright triangular bony **plates** running along the back...

A four-legged herbivore from the Mid-Jurassic to the Late Cretaceous time. Its two rows of **bony** plates and tail **spikes** probably provided it much protection against large predators like Tyrannosaurus rex...

Stegosaurus was up to 26-30 feet long, about 9 feet tall, and weighed about 6,800 pounds. Its **small brain** was only the size of a walnut. Its skull was long, pointed, and narrow; it had a toothless beak and small cheek teeth...

Training

stegosaurus

forelimbs

small brain

carnivore

plates

back

bony spikes

tyrannosaurus

...

Feature Mapping

Text Labeled Data

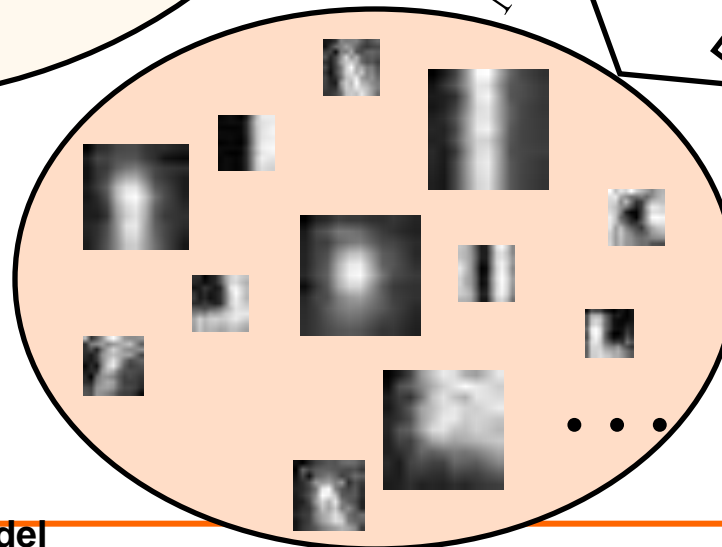


Image Classification Model


(Dai, Xue, Yang, NIPS 2008)

Learning Feature Mappings

- Cooccurrence data

Home The Tour Sign Up Explore Search

Bardi fortress



Would you like to comment?
[Sign up](#) for a free account, or [sign in](#) (if you're already a member).

Uploaded on October 19, 2008
by [FrancescoP](#)

FrancescoP's photostream

You are at the last photo. 345 uploads

browse

This photo also belongs to:

Val di Taro (Settembre 2008) (Set)

You are at the last photo. 33 items

browse

Part of: [Travels](#)

Tags

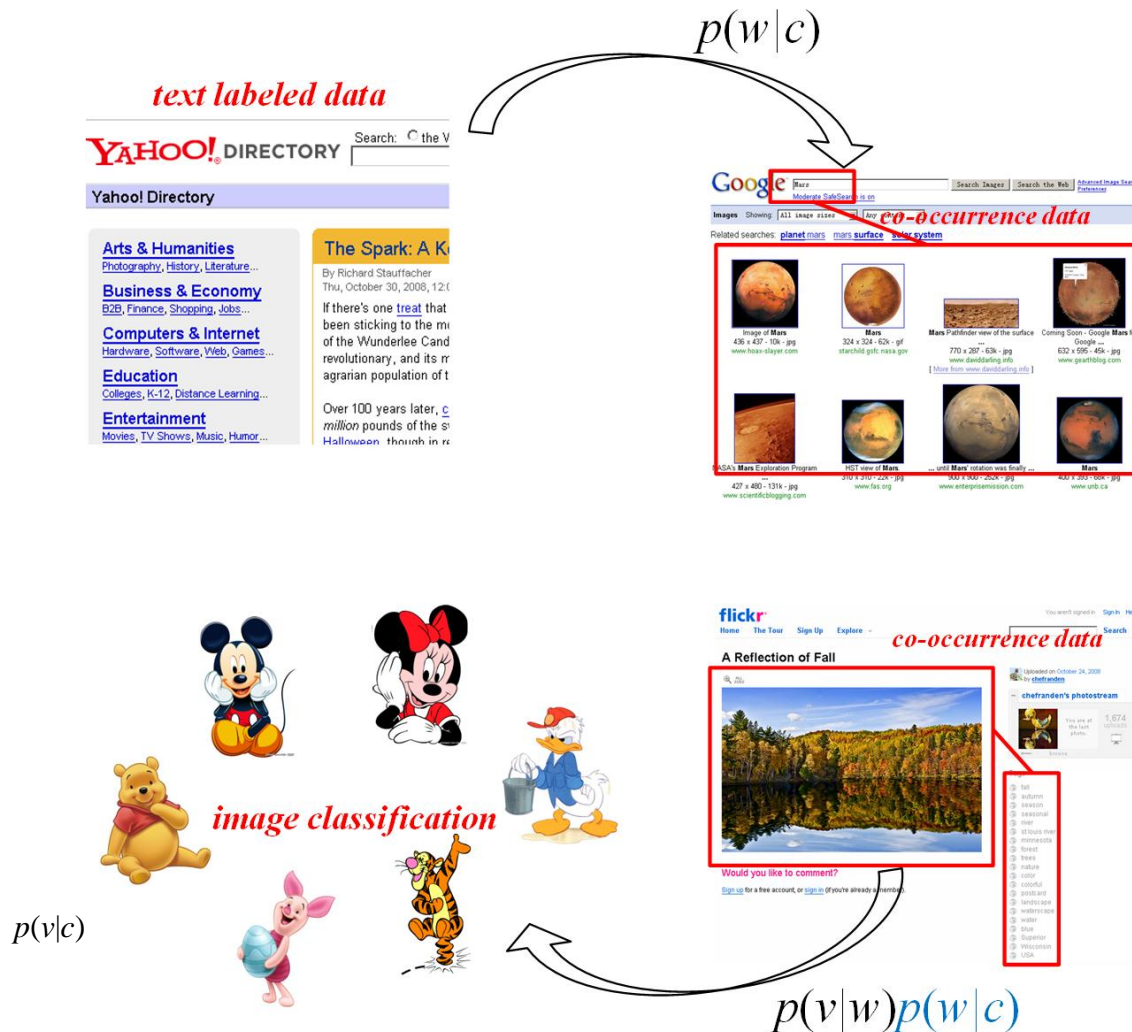
- Bardi
- fortress
- castel
- old
- medieval
- history
- travel
- Parma
- Italy
- Canon EOS 350D
- Canon 50mm

Learning Feature Mappings

- Search engine



Translated Learning Algorithm

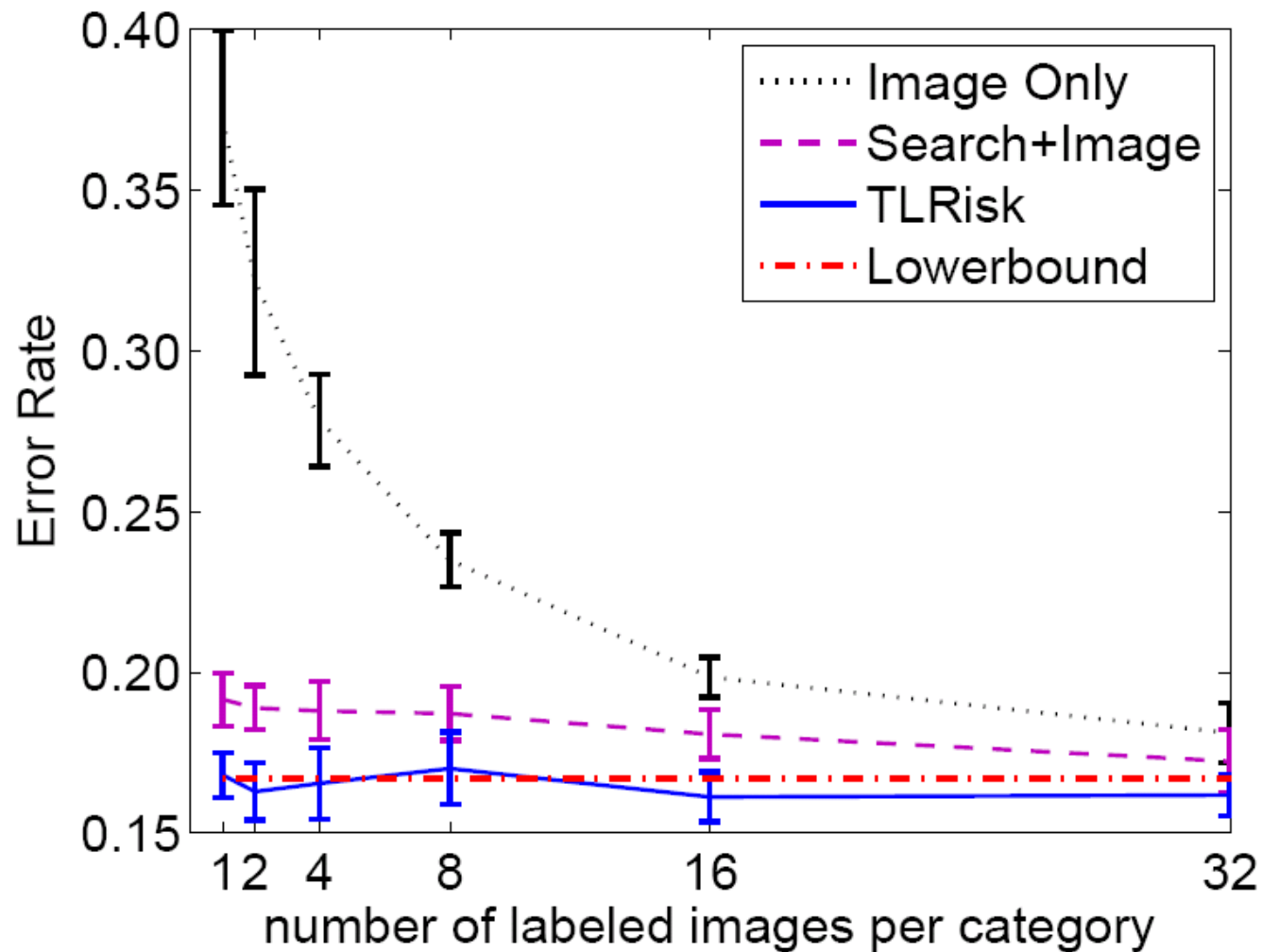


Experiments on Classification

- Documents: Open Directory Project (ODP)
- Images: Caltech-256
- Feature Mapping: Google Image Search

DATA SET	DATA SIZE				DATA SET	DATA SIZE			
	DOCUMENTS		IMAGES			DOCUMENTS		IMAGES	
	+	−	+	−		+	−	+	−
horse vs coin	1610	1345	270	123	dog vs canoe	1084	1047	102	103
kayak vs bear	1045	885	102	101	greyhound vs cd	380	362	94	102
electric-guitar vs snake	335	326	122	112	stained-glass vs microwave	331	267	99	107
cake vs binoculars	265	320	104	216	rainbow vs sheet-music	261	256	102	84
laptop vs sword	210	203	128	102	tomato vs llama	175	172	102	119
bonsai vs comet	166	164	122	120	frog vs saddle	150	148	115	110

Experimental Results



Application (Advertising)

- Visual Contextual Advertising
 - [Chen et al. AAAI 2010]
- Image to Text Ads
 - [News from MIT *Technology Review*]



[1] 2 Next »

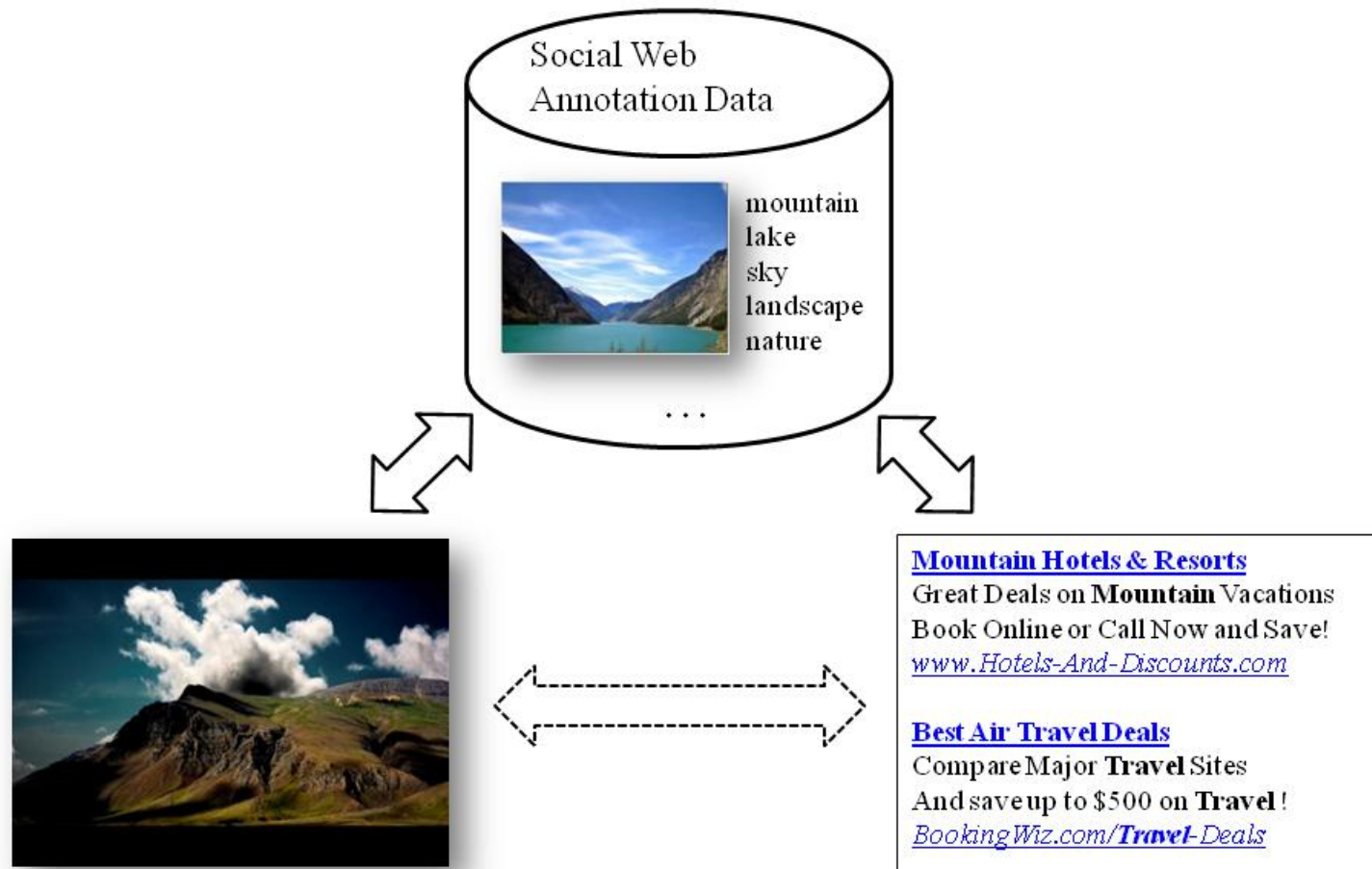
Ads that Match a Web Page's Images

Using the contents of images or videos to target Web ads could improve click-through.

By Tom Simonite

WEDNESDAY, JULY 21, 2010

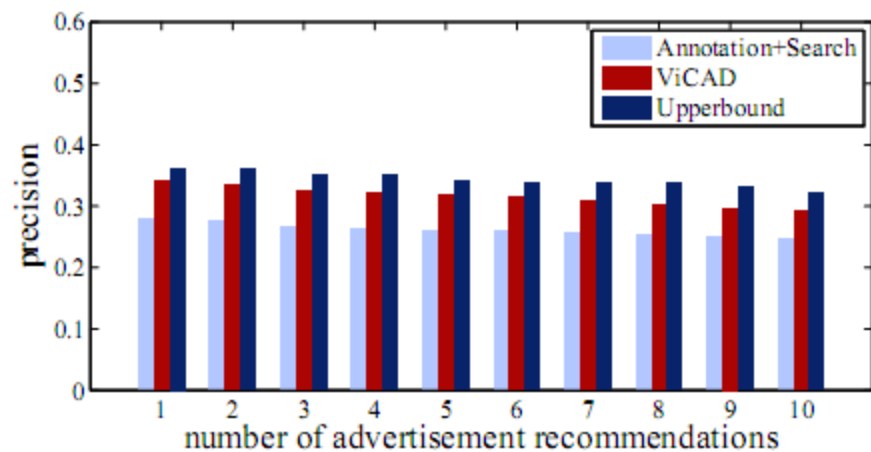
Figure illustration of Visual Contextual Advertising



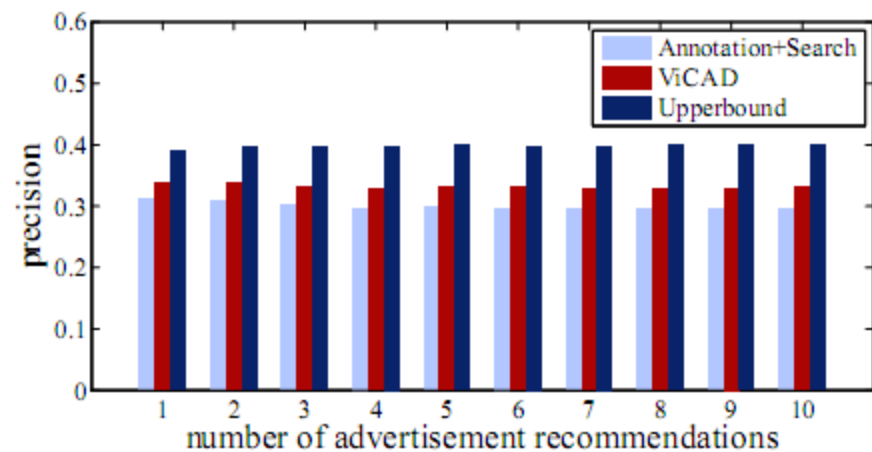
Experimental Results

- Co-occurrence data from Flickr.
- Test Image from Flickr and Fifteen scene data set
- Advertisement are crawled from MSN search engine with queries chosen from AOL query log.

Experimental Results



(a) Flickr image set



(b) Fifteen scene data set

Q / A ?

Thanks!

