The Anatomy of Data Driven Learning

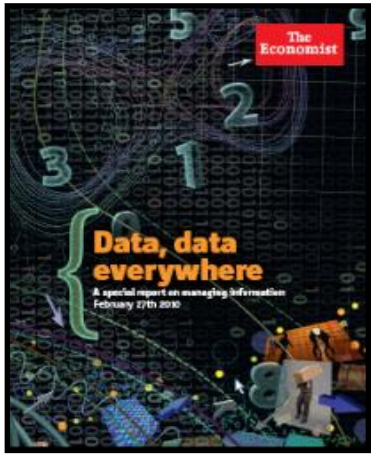
Gui-Rong Xue

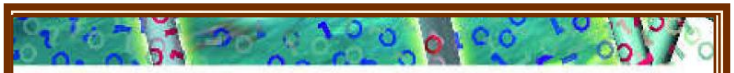
阿里巴巴云计算
Alibaba Cloud Computing

阿里巴巴集团研发院
Alibaba Group Research & Development

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

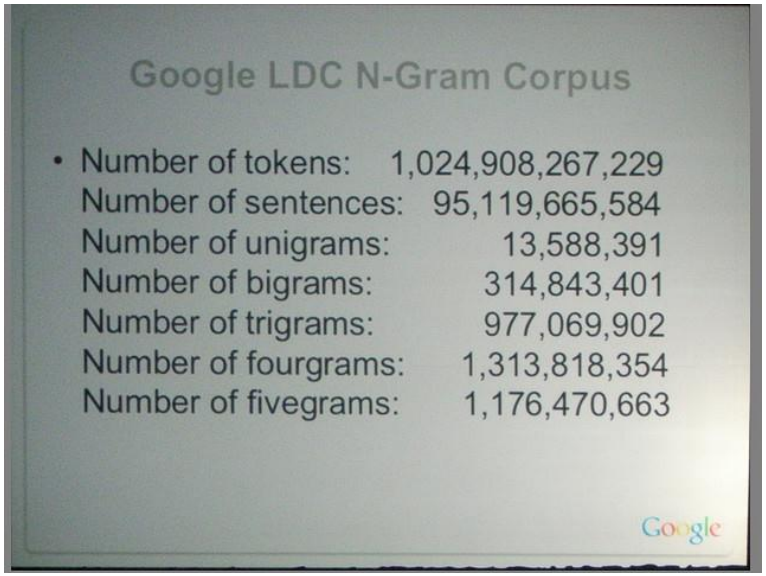
2010《经济学人》“数据、无处不在的数据”专刊

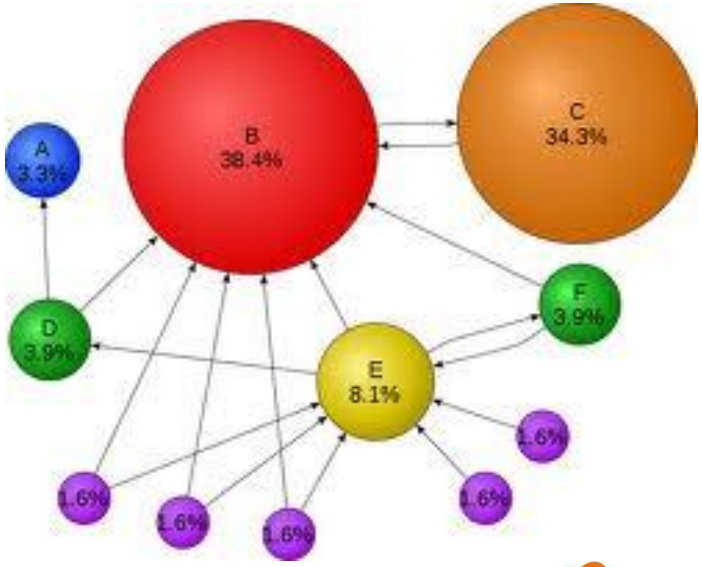
- CERN 每秒产生数据40TB
- Alibaba.com 阿里巴巴 每天10TB交易数据
- facebook 4亿用户，400亿照片

阿里巴巴云计算 Alibaba Cloud Computing

What can big data tell us ?

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





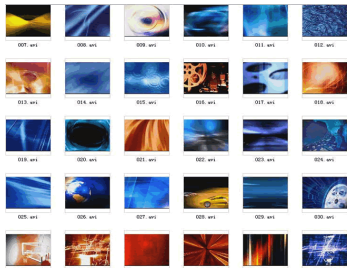
Data Driven Learning

- 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
    - … …

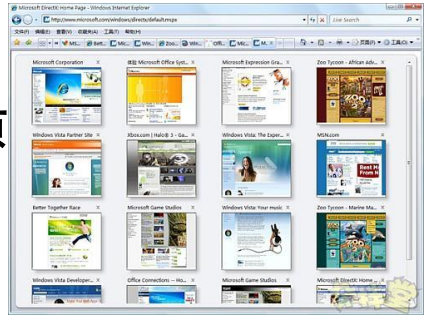
Aliyun.com Open Data Platform

中日英三语语料

1亿图片



10亿网页



开放云服务机群

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

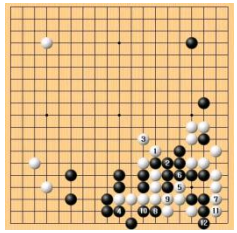
阿里巴巴云计算
Alibaba Cloud Computing

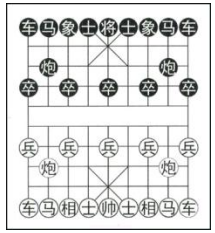
阿里巴巴集团研发院
Alibaba Group Research & Development

Human Learning

learning

solving





Traditional Machine Learning

Input → Learning → Output → Classifier

Input → Classifier

Classifier → elephant/rhino

Training Data

Test Data

Transfer Learning

Training Data

Auxiliary Data

Test Data

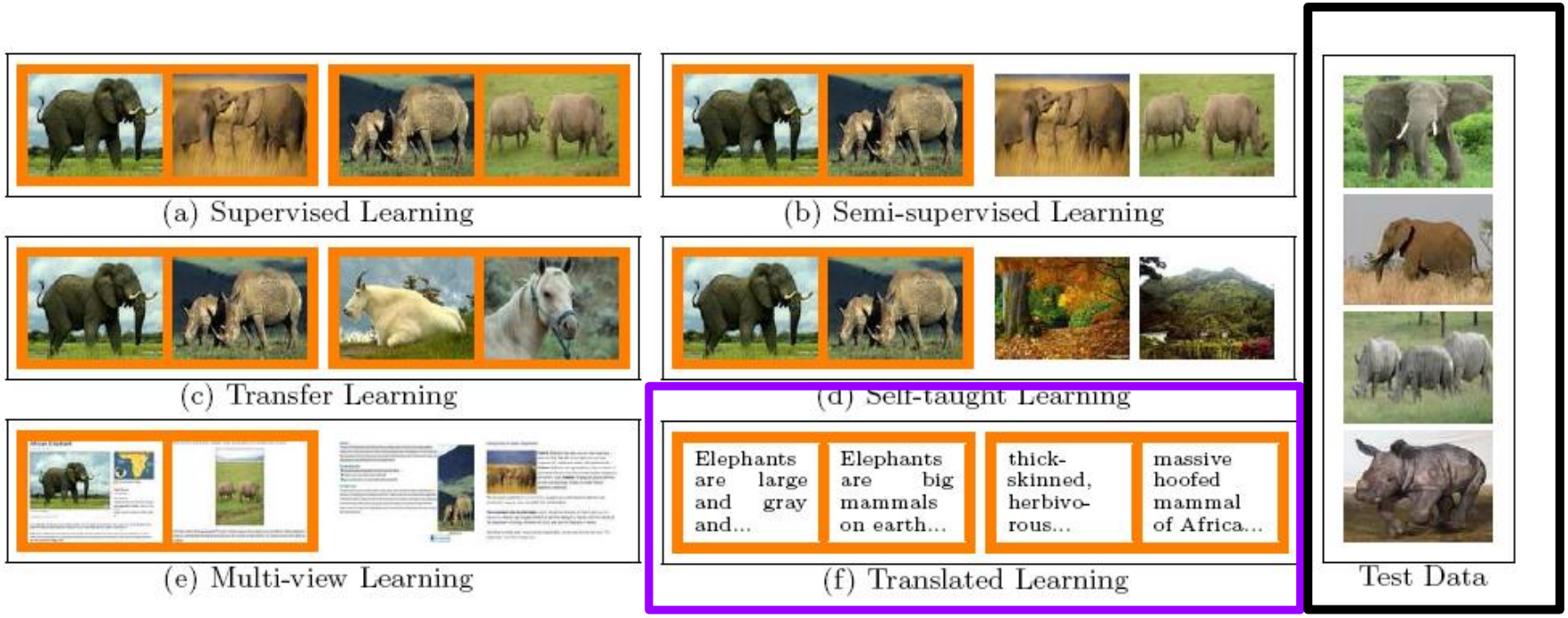
Translated Learning

阿里巴巴云计算
Alibaba Cloud Computing

阿里巴巴集团研发院
Alibaba Group Research & Development

Translated Learning

- Transfer Learning across Different Feature Spaces



Instance-level Translation

- Generally, it is difficult.
  - Easy translation

Hello world! ⇨ 世界你好!

  - Hard translation

⇨ ? in text

Model-level Translation

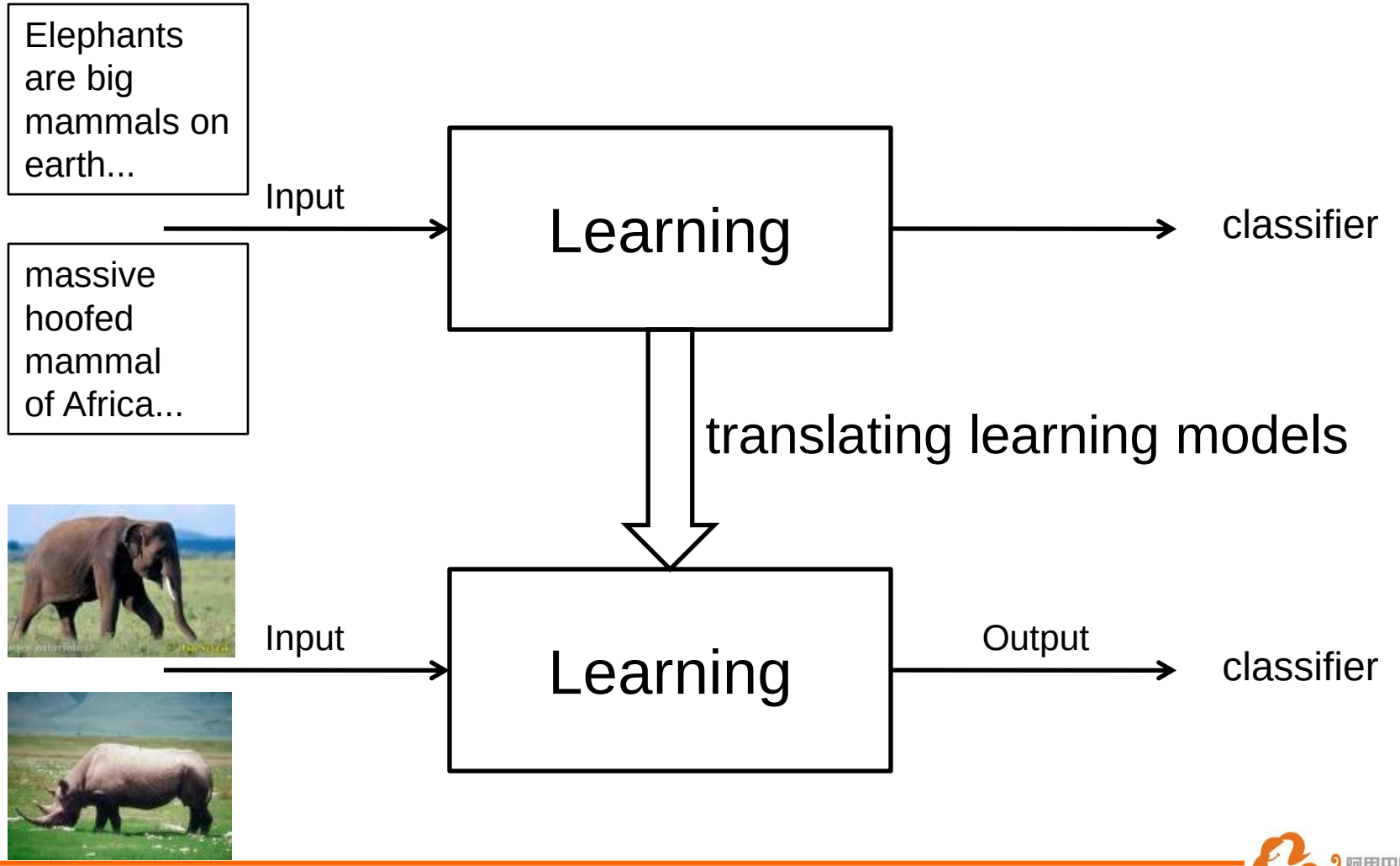
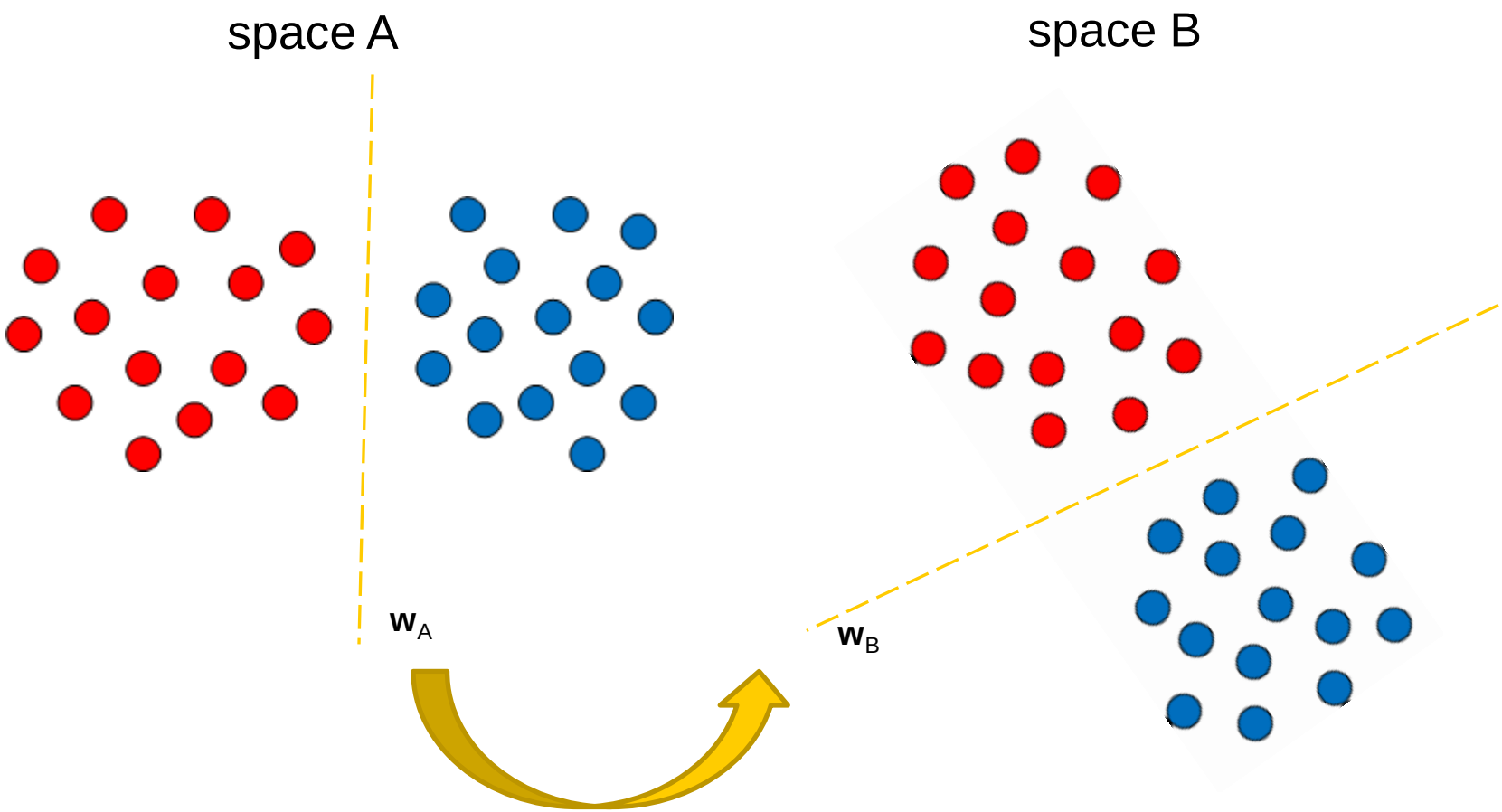


Illustration in Vector Space Model



$\mathbf{w}_B = M \cdot \mathbf{w}_A$

(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…

• • •

Text Labeled Data

Training

stegosaurus

forelimbs

small brain

plates

carnivore

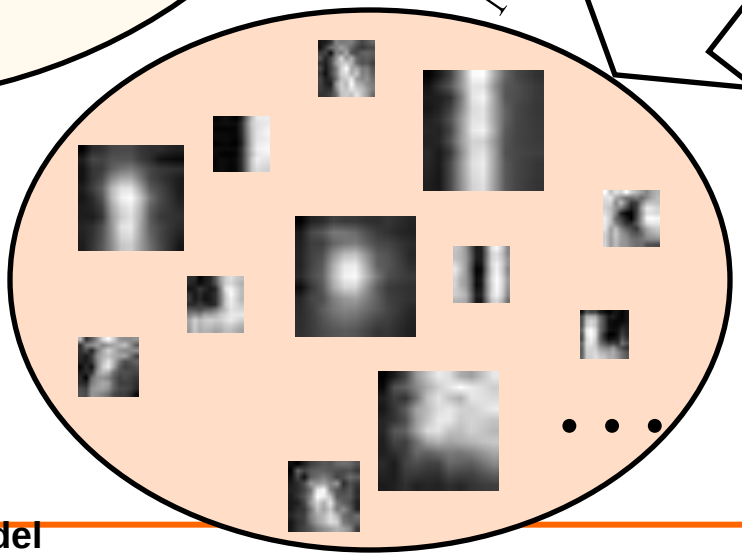
back

bony spikes

tyrannosaurus

• • •

Feature Mapping



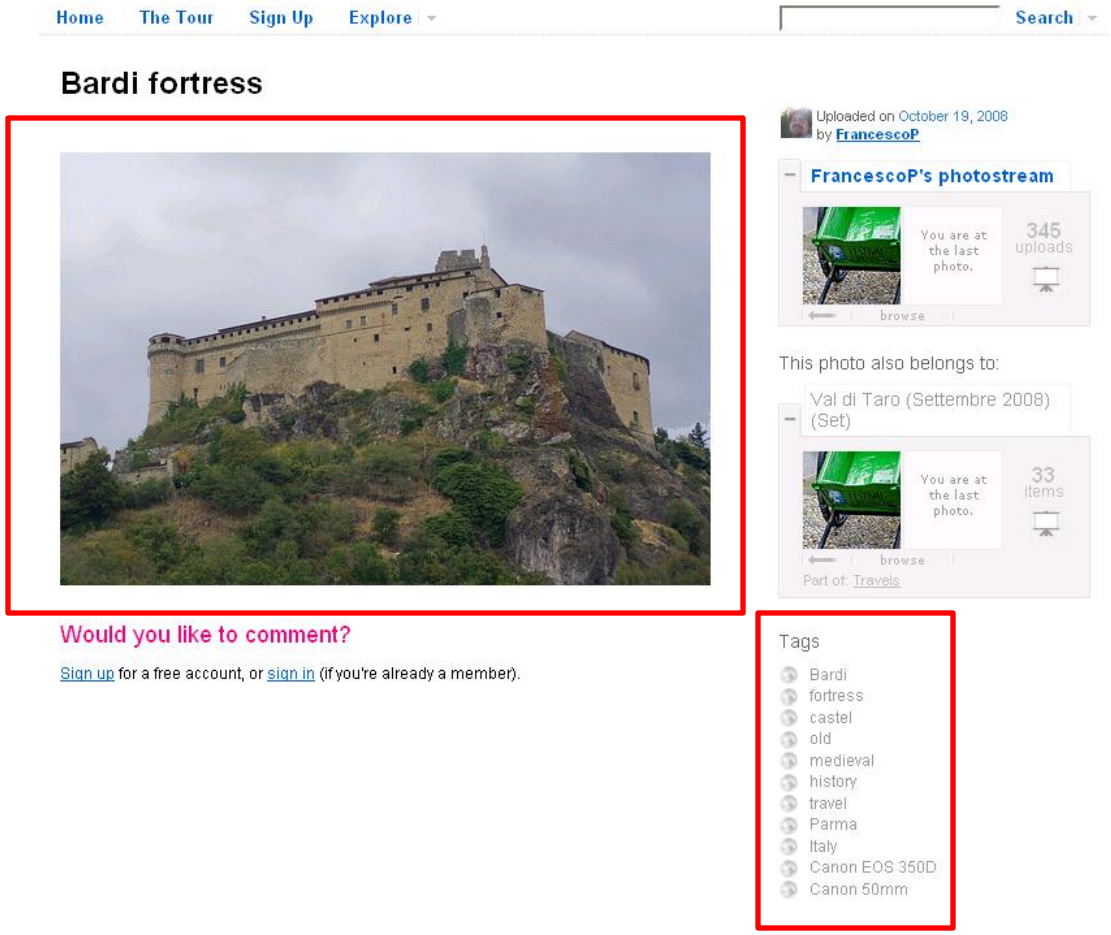
• • •

Image Classification Model

(Dai, Xue, Yang, NIPS 2008)

Learning Feature Mappings

- Cooccurrence data



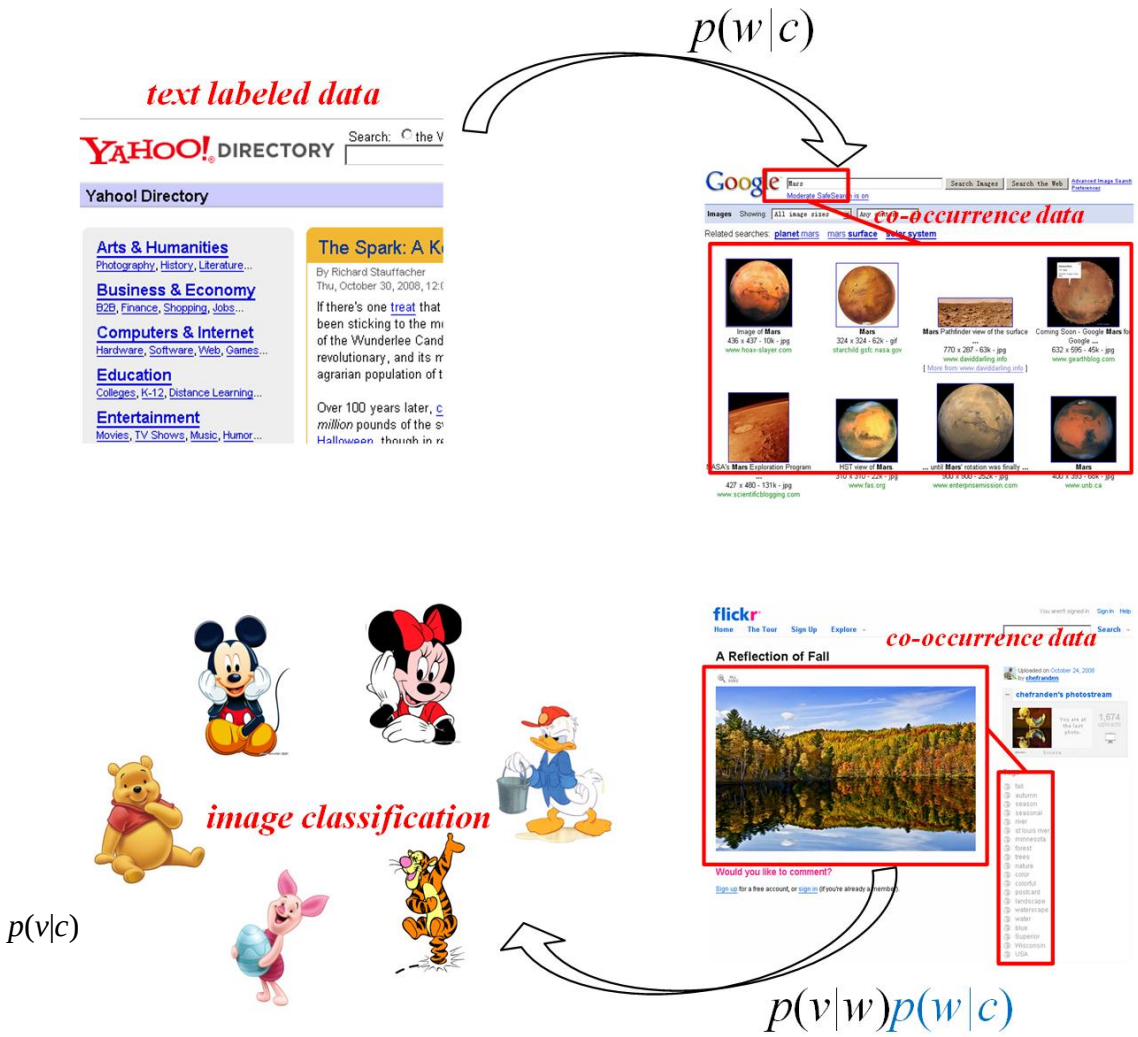
17

Learning Feature Mappings

- Search engine

18

Translated Learning Algorithm



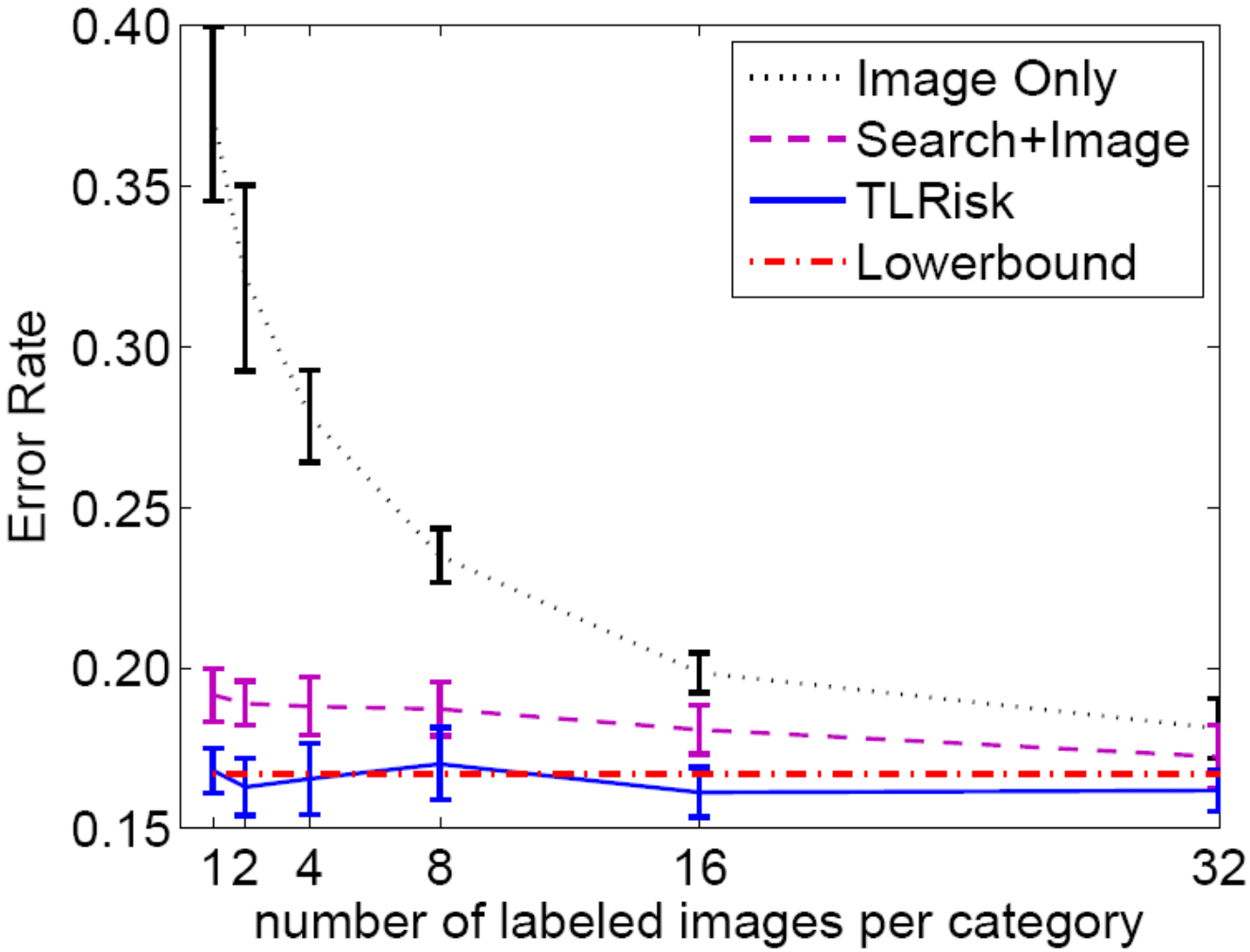
19

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*]

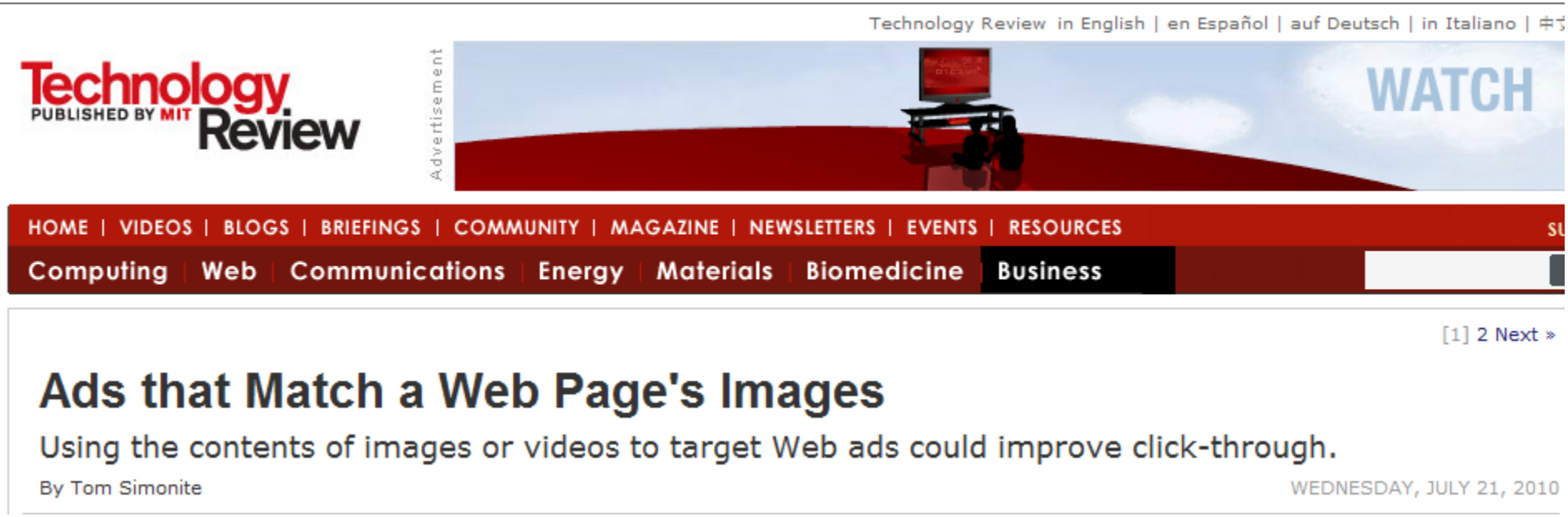
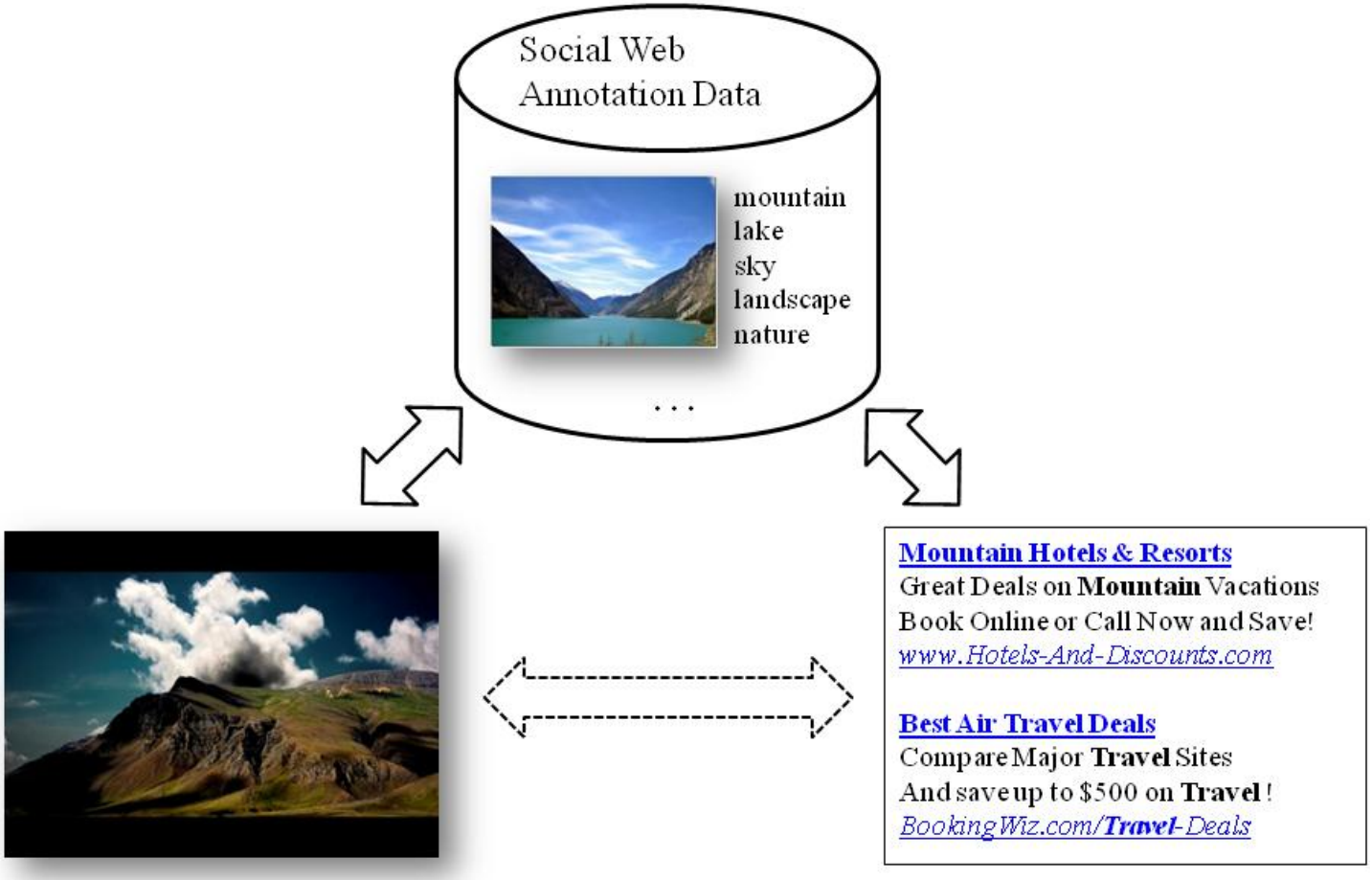


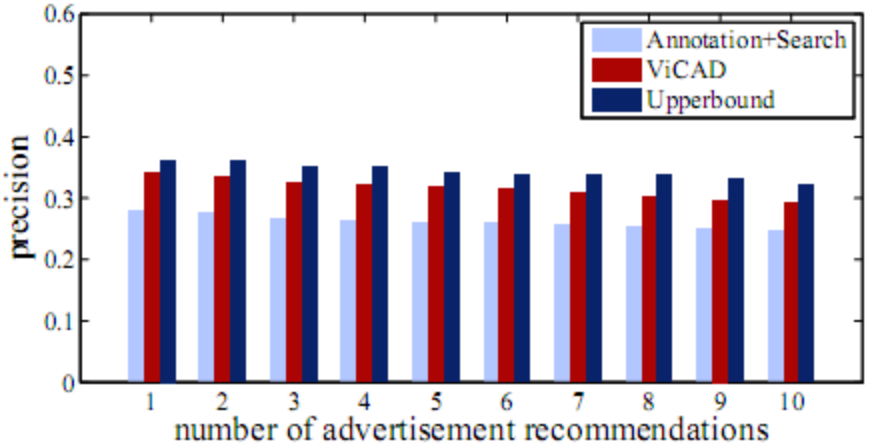
Figure illustration of Visual Contextual Advertising



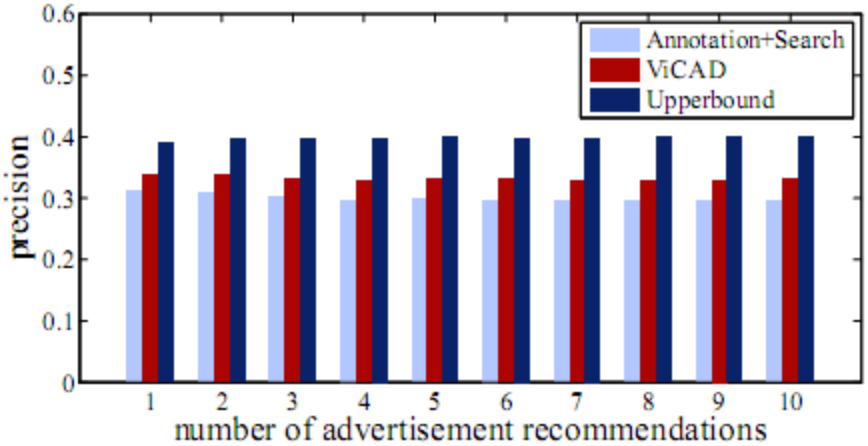
Experimental Results

- 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!

阿里巴巴云计算
Alibaba Cloud Computing

阿里巴巴集团研发院
Alibaba Group Research & Development