Deep Learning Meets Heterogeneous Computing

Dr. Ren Wu
Distinguished Scientist, IDL, Baidu
wuren@baidu.com

Baidu

Everyday

5b+ queries
500m+ users
100m+ mobile users
100m+ photos

…

As of January 2014, Baidu's market capitalization stood at roughly $55 billion.[8] It is considered one of the most successful internet startups of all time by market capitalization, revenue, growth and cultural impact.[9]

WIKIPEDIA
The Free Encyclopedia

Big Data

Storage	• >2000PB
Processing	• 10-100PB/day
Webpages	• 100b-1000b
Index	• 100b-1000b
Update	• 1b-10b/day
Log	• 100TB~1PB/day

Will Baidu's Data Center Be the World's Largest?

by Mark Hachman | September 6, 2012

Chinese Web giant says that it's spending $1.6 billion on a cloud data center, a massive investment in its infrastructure.

Infrastructure

ARM Servers

- Higher density

GPU服务器

- Much better performance

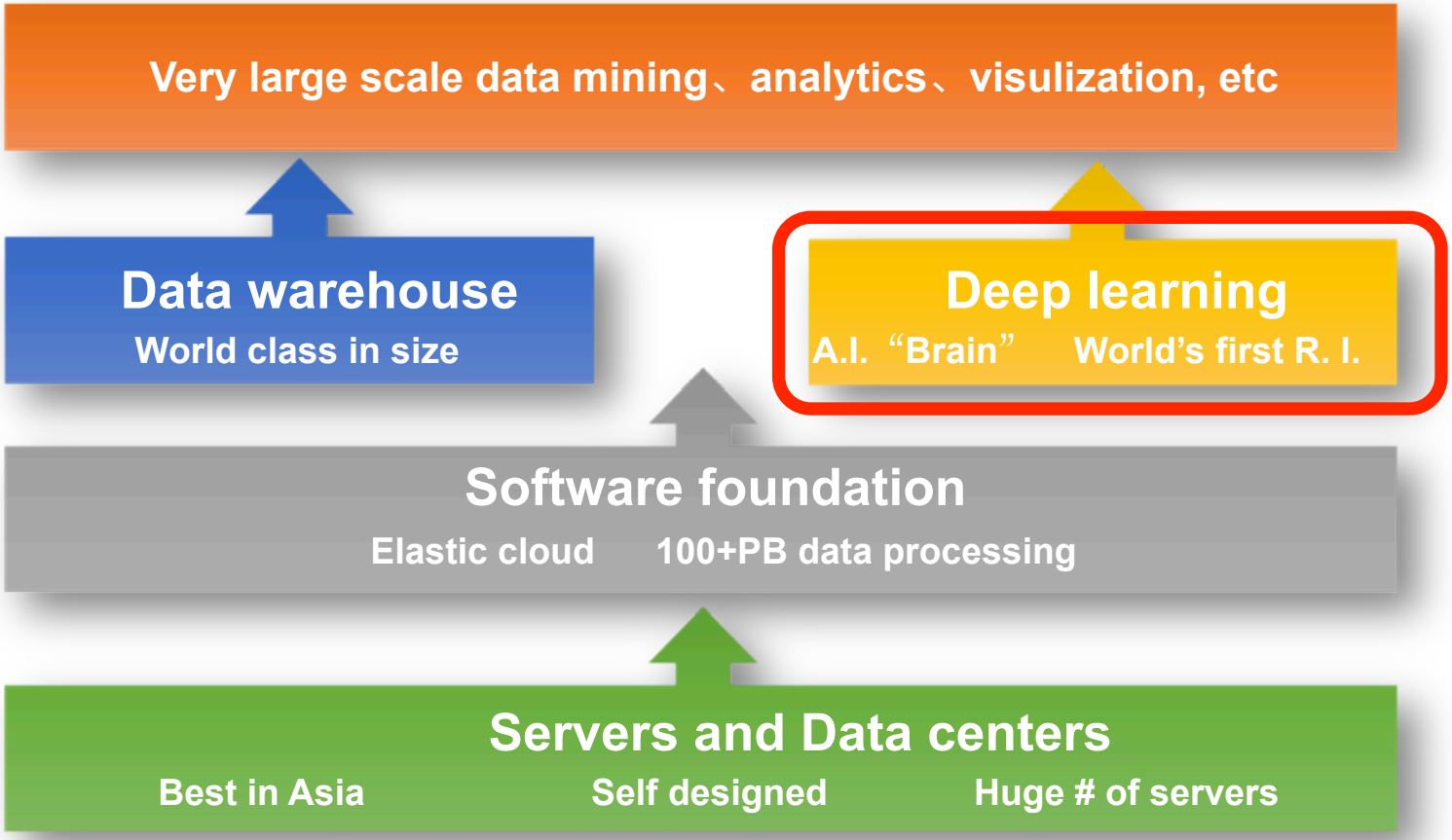
Data center containers

- Faster deployment

Self-design switches

- Much lower cost

Big Data @Baidu



Nine Technology Challenges

On Aug 13, 2012, CEO Robin Li gave a keynote speech at ACM KDD, and proposed nine major technological challenges to the academic research community. The top three are:

1. OCR in natural images
2. Speech recognition and understanding
3. Content-based image retrieval (visual search)

Deep Learning Since 2006

materials are identical for all configurations. The blue bars in Fig. 1 summarize the measured SHG signals. For excitation of the *LC* resonance in Fig. 1A (horizontal incident polarization), we find an SHG signal that is 500 times above the noise level. As expected for SHG, this signal closely scales with the square of the incident power (Fig. 2A). The polarization of the SHG emission is nearly vertical (Fig. 2B). The small angle with respect to the vertical is due to deviations from perfect mirror symmetry of the SRRs (see electron micrographs in Fig. 1). Small detuning of the *LC* resonance toward smaller wavelength (i.e., to 1.3-μm wavelength) reduces the SHG signal strength from 100% to 20%. For excitation of the Mie resonance with vertical incident polarization in Fig. 1D, we find a small signal just above the noise level. For excitation of the Mie resonance with horizontal incident polarization in Fig. 1C, a small but significant SHG emission is found, which is again po-

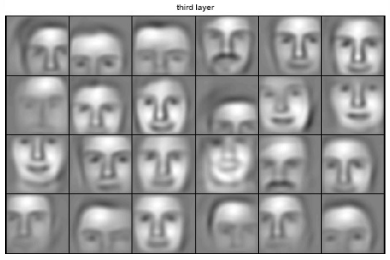
Reducing the Dimensionality of Data with Neural Networks

G. E. Hinton* and R. R. Salakhutdinov

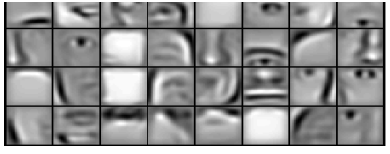
High-dimensional data can be converted to low-dimensional codes by training a multilayer neural network with a small central layer to reconstruct high-dimensional input vectors. Gradient descent can be used for fine-tuning the weights in such "autoencoder" networks, but this works well only if the initial weights are close to a good solution. We describe an effective way of initializing the weights that allows deep autoencoder networks to learn low-dimensional codes that work much better than principal components analysis as a tool to reduce the dimensionality of data.

Dimensionality reduction facilitates the classification, visualization, communication, and storage of high-dimensional data. A simple and widely used method is principal components analysis (PCA), which finds the directions of greatest variance in the data set and represents each data point by its coordinates along each of these directions. We describe a nonlinear generalization of PCA that uses an adaptive, multilayer "encoder" network

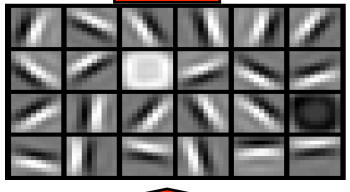
Deep Learning vs. Human Brain



object models

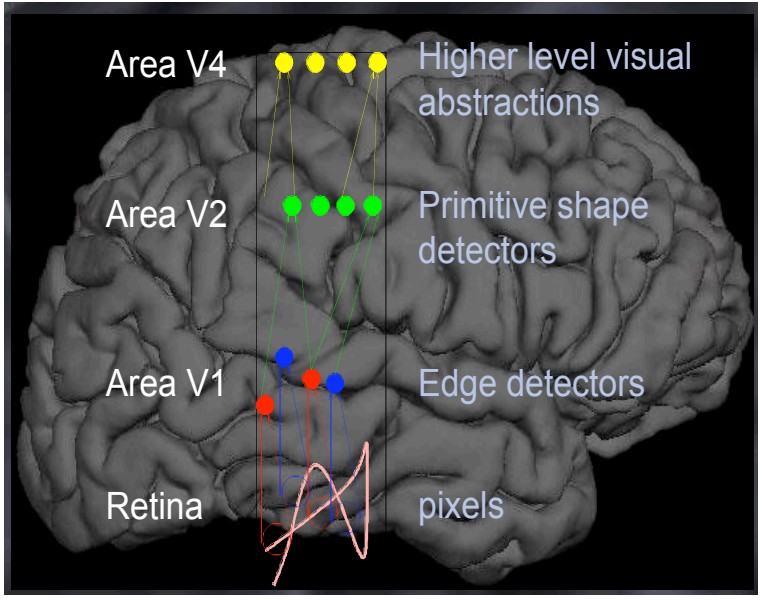


object parts (combination of edges)



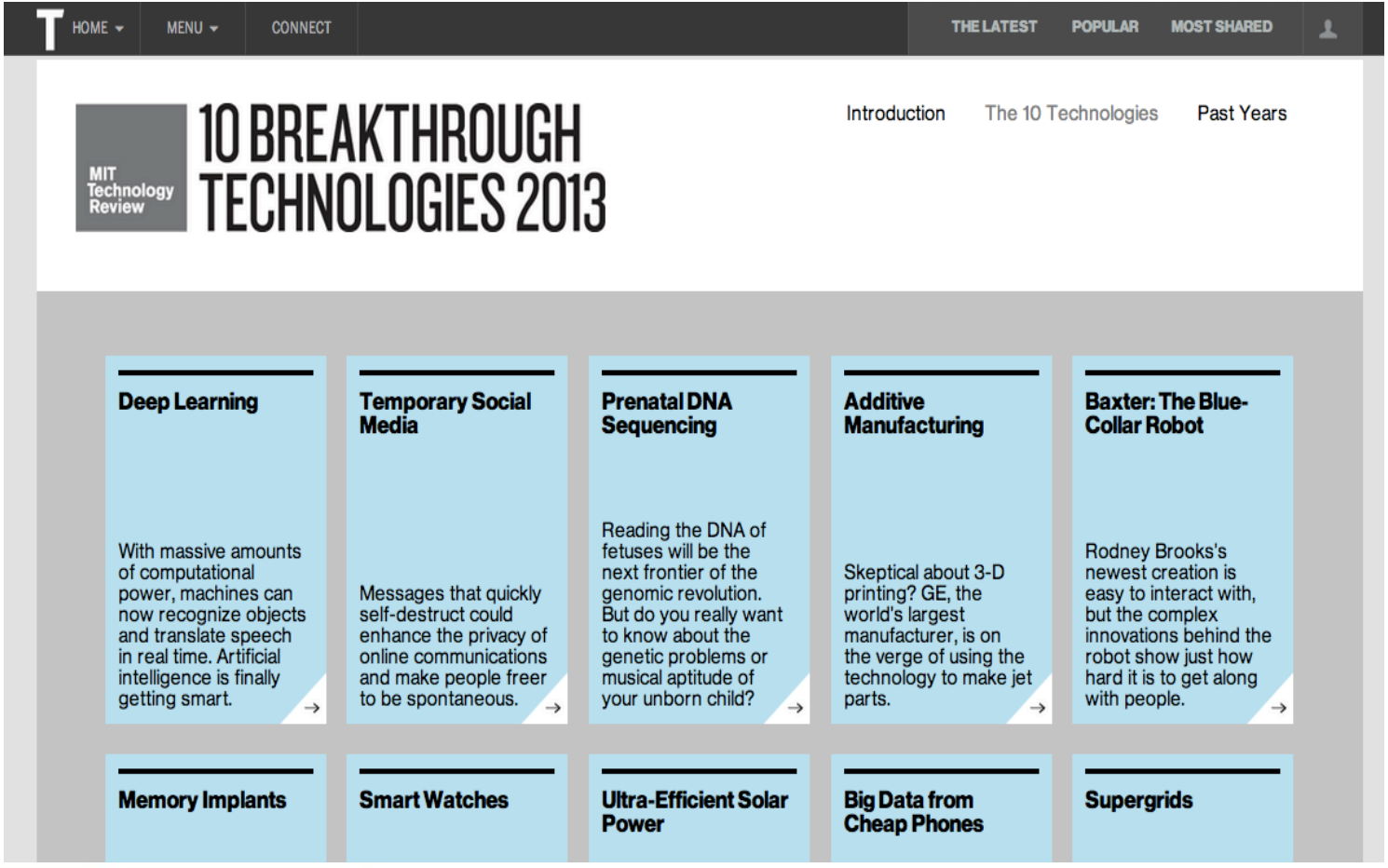
edges

pixels



Slide credit: Andrew Ng

Top breakthrough technology 2013



MIT Technology Review, April 23rd, 2013

Baidu IDL

- Announced its first research arm in Jan. 2013
- Institute of Deep Learning (IDL)
- The focus is Artificial Intelligence
- Two locations: Beijing and Silicon Valley

Progress of Deep Learning at Baidu

- Big improvement on speech & image recognition (2013)
  - Speech: error rate reduced by 25%
  - OCR: error rate reduced by 30%
  - Face: LFW benchmark, 94% correct

- DNN CTR for search ads was launched on May 20th 2013, serving billions of search queries everyday – substantial improvement

Baidu – Visual Search

Baidu 识图

http://stu.baidu.com

Visual Search: Faces

Visual Search Example

Baidu 识图 | 本地上传 | 请粘贴图片网址或拖拽图片至此

670x753

您的图片有可能是：

邓丽君　成龙　陈晓旭

为您找到：

邓丽君

› 来自百度百科

邓丽君（Teresa Teng，1953年1月29日—1995年5月8日），是一位在亚洲地区和全球华人社会极具影响力的台湾歌唱家，亦是20世纪后半叶最富盛名的日本歌坛巨星之一。她的歌曲在华人社会广泛的知名度和经久不衰的传唱度为其赢得了"十亿个掌声"的美誉，被日本艺能界尊为"亚洲歌唱女王"。其生前演艺

› 人物概况　› 早年经历　› 演艺经历　› 个人生活　› 主要作品

相似图片

› 全部相似图片

相关网页

› 更多来源

揭秘: 邓丽君 ,梅艳芳,陈晓旭的

她的歌声甜蜜依旧 她的笑容甜美动人 她的情路一波三折 她的柔情似水东流 ...

1/2/2013 2:28:11 PM

http://ent.rednet.cn/c/2009/04/10/1741997_2.htm

"聆听永恒的经典"纪念 邓丽君 20

旷世美女 邓丽君 的终身遗憾

邓丽君 成龙恋情昙花一现 盘点

芝兰之室(a) - 邓丽君 的绝版私

人脸搜索

› 更多人脸搜索

Visually similar images

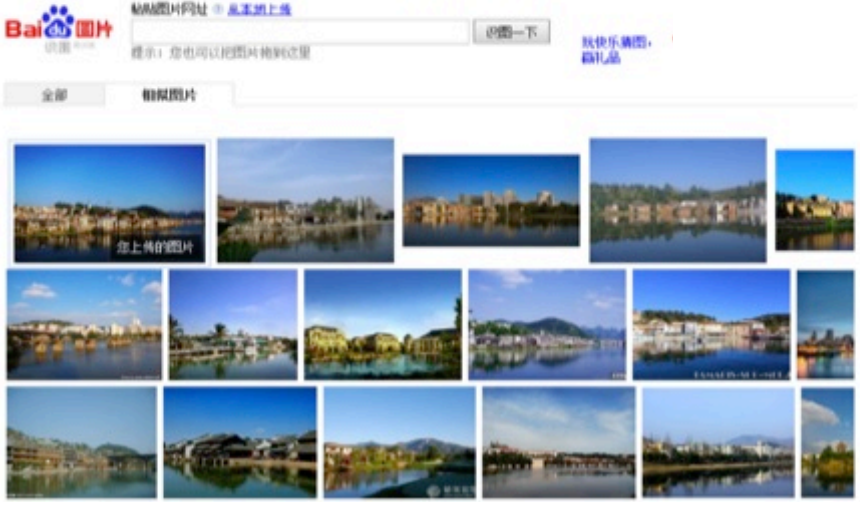
The competition

Baidu

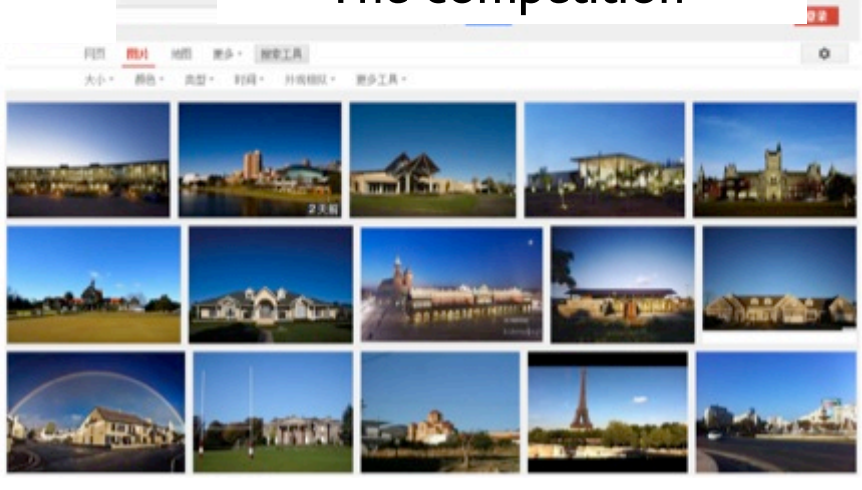
Another Example

Image uploaded

Baidu

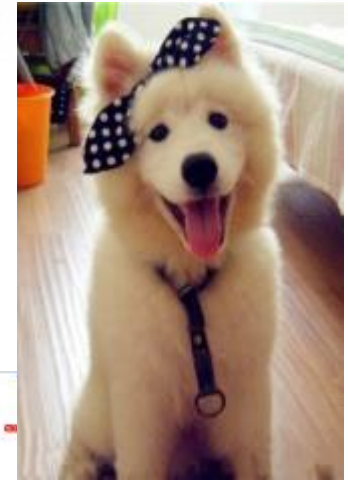


The competition



Visually Similar Images - Comparison

Image uploaded



Baidu

The competition

CBIR – The Competition

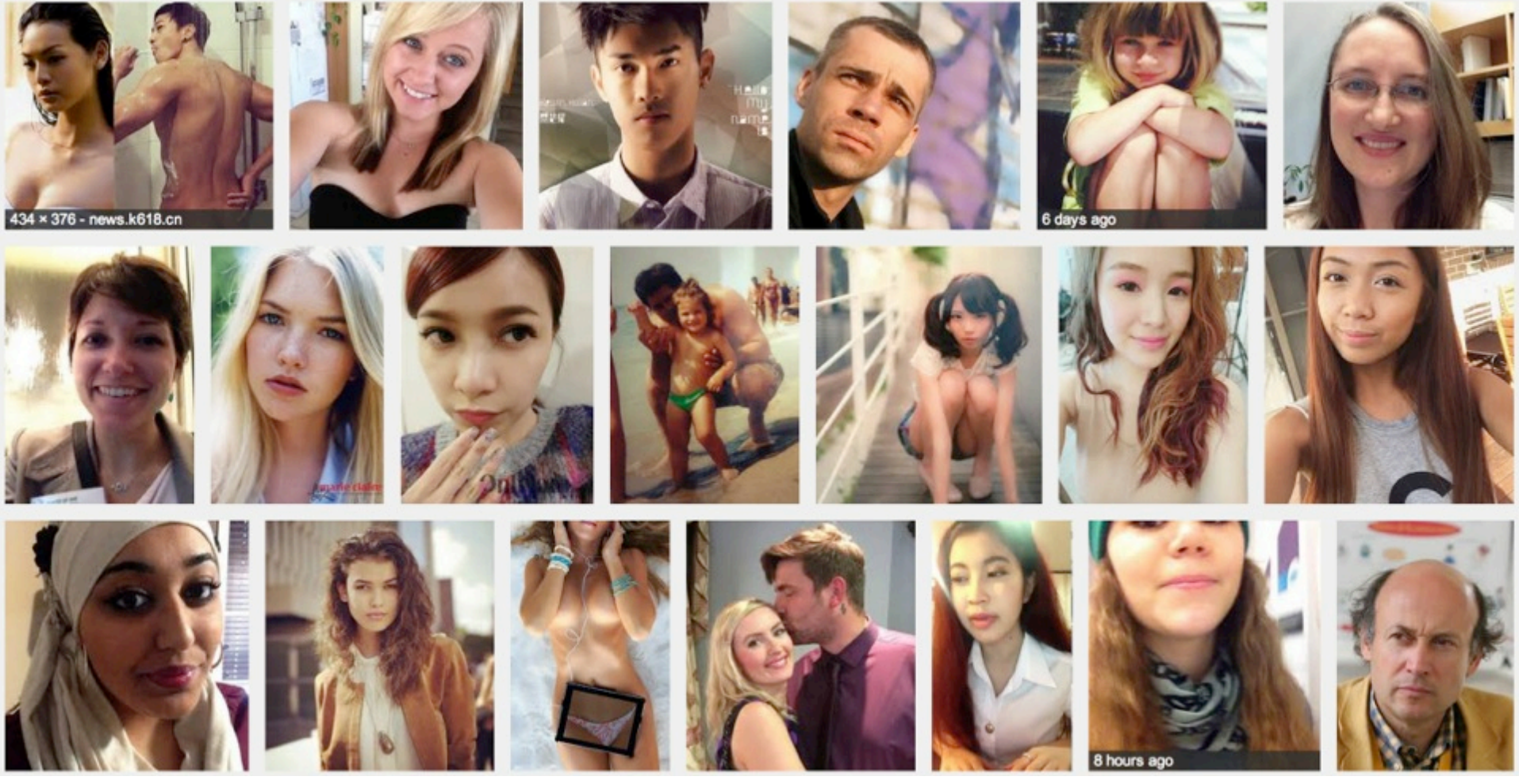


Image Recognition - Flowers

原图尺寸:750x502

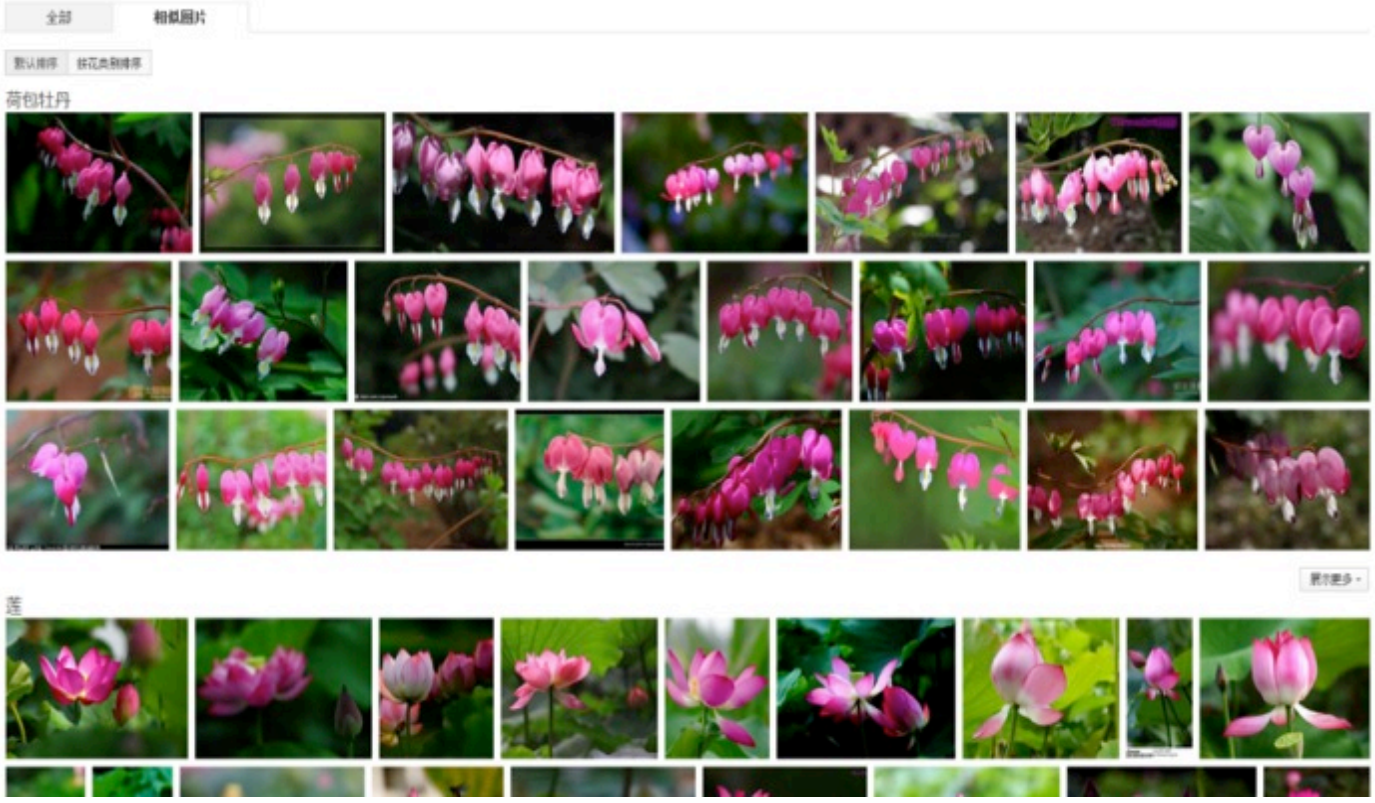
我猜测这种花是：

荷包牡丹

99%的可能性

荷包牡丹原产中国、西
茎。叶对生，有长柄，
萼片两枚，早落。花瓣
荷...

植物简介 - 科属分类 -



百度魔图：PK大咖

Peak uploadin

)S APP #1 for 3 weeks

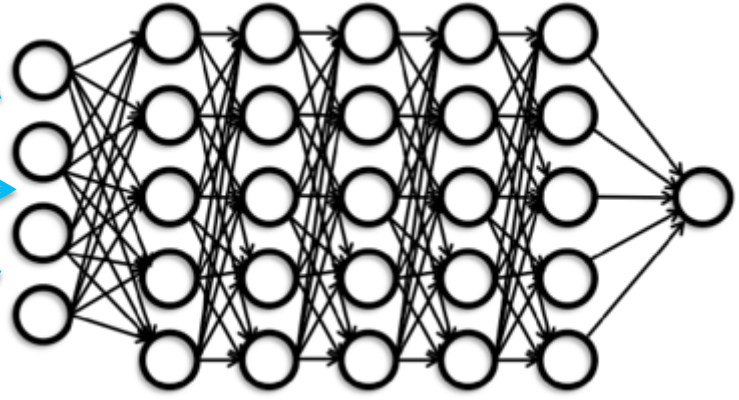
Deep Learning

DNN for Speech
10k hours of voice data
10b training samples
Months on a GPU cluster

Voice, Text

Image

User



Typical scale of training data

Datasets

- Image recognition: 100 millions
- OCR: 100 millions
- Speech: 10 billions
- CTR: 100 billions

Training time:

Weeks to Months
on GPU clusters

Big data + Deep learning + HPC = Success

Projected training data to grow 10x each year

Mobile Applications of DNN

“手机百度 随时知道”

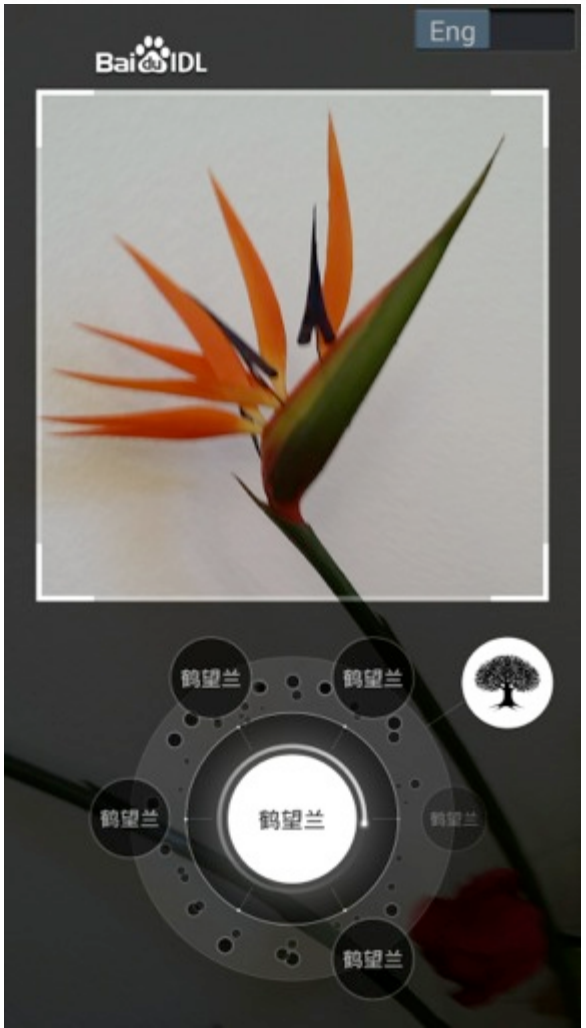
DNN – Anywhere, Anytime

- DNN-based image recognition on mobile device
- No connectivity needed
- Real time, directly works on video stream
- Everything is done within the device
- What you point is what you get

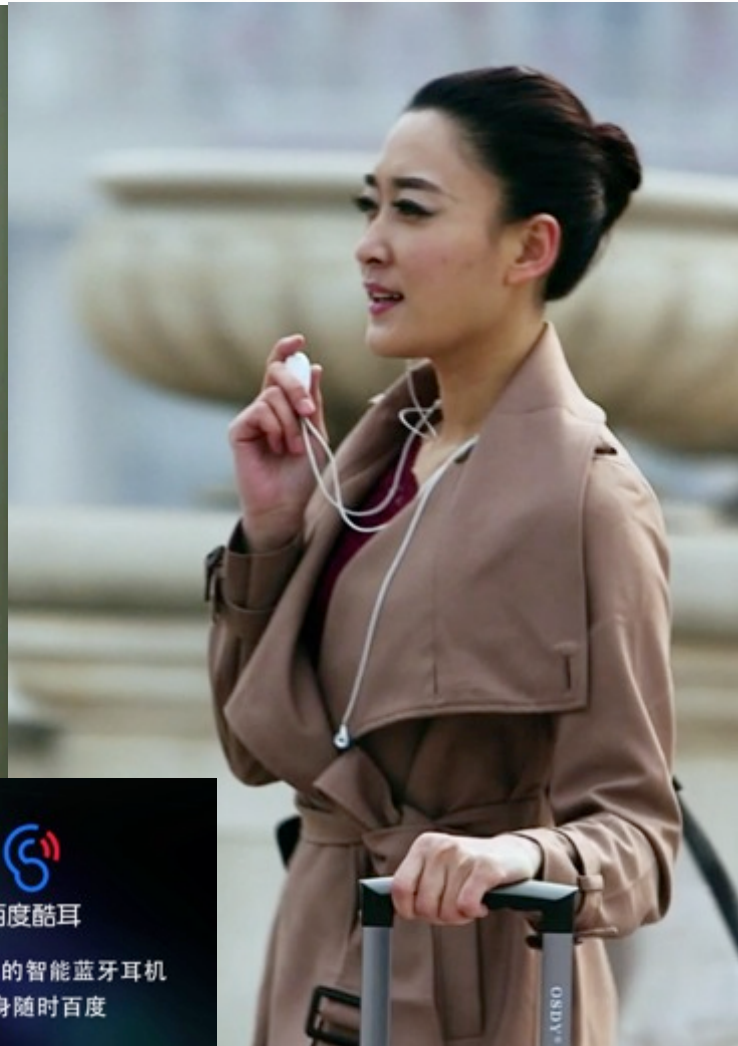
- OpenCL based, highly optimized
- Large deep neural network models
- Thousands of objects, flowers, dogs, and bags etc
- Unleashed the full potential of the device hardware

- World's first in-place mobile DNN app?
- And the best!

DNN – Anywhere, Anytime



百度酷耳



DNNs Everywhere

Supercomputers

Datacenters
(cloud)

Tablets, smartphones

Wearable devices
IoTs

DNNs Everywhere

Supercomputers

1000s GPUs

Datacenters

100k-1m servers

Tablets, smartphones

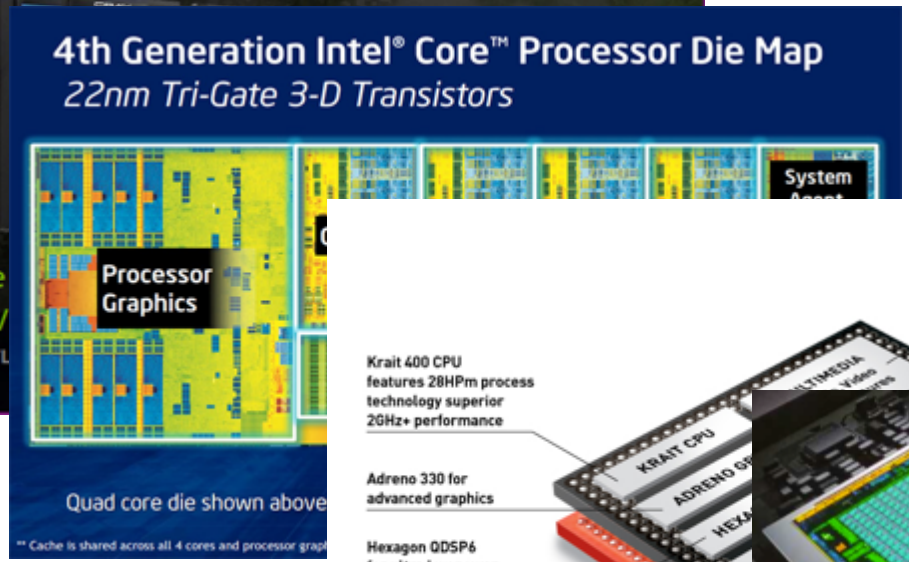
700m (in China)

Wearable devices
IoTs

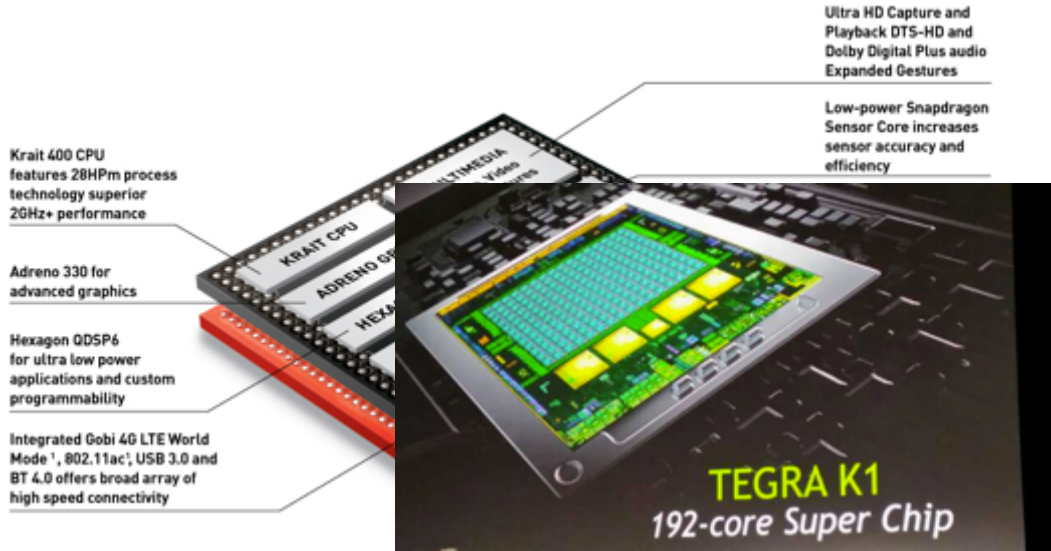
Billions?

Supercomputer used for training
Trained DNNs then deployed to data centers (cloud), smartphones, and even wearables and IoTs

Heterogeneous Computing



Supercomputers
Data centers (cloud)
Smart phones
Wearable devices!



Big data +
Deep learning +
~~HPC~~ HC = Success

OpenCL-based Open ECO-SYSTEM

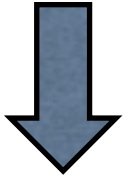
- Diverse industry participation, from cell phones to supercomputers
  - Processor vendors, system OEMs, middleware vendors, application developers.
- OpenCL is the industry standard embraced by many companies.

* Courtesy of Simon McIntosh-Smith and Tom Deakin

Summary

Big data + Deep learning + High performance computing =

Intelligence



Big data + Deep learning + Heterogeneous computing =

Success

Baidu USA

And we are hiring

- Heterogeneous Computing experts
- Parallel algorithm and performance experts
- CUDA/OpenCL Experts
- FPGA experts
- Andrios/IOS experts
- Data scientist
- Infrastructure Engineer

…

usdc-jobs@baidu.com
wuren@baidu.com

http://usa.baidu.com/

Thank you!

Dr. Ren Wu
wuren@baidu.com
@韧在百度