

大数据机器学习

张潼

2013年9月13日

互联网和大数据



需求：抽取信息；创造用户体验；创造商业价值

怎样去最有效的获取信息？

- 分析大量数据
- 过滤并展示给用户

资源：数据



手段：计算能力
+算法+展现方式

核心技术



大数据管理：infrastructure

大数据分析：machine learning

应用：system integration

机器学习

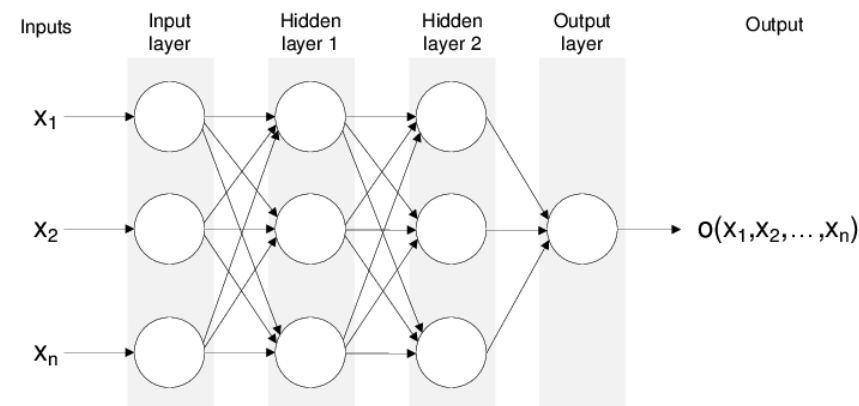
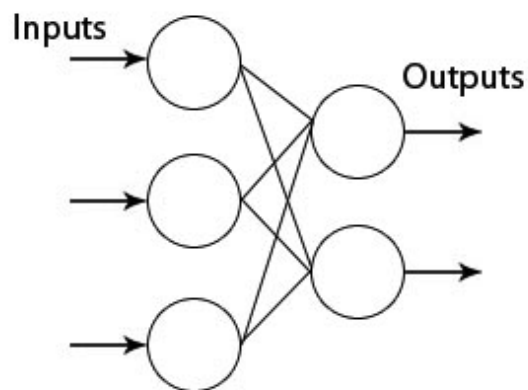
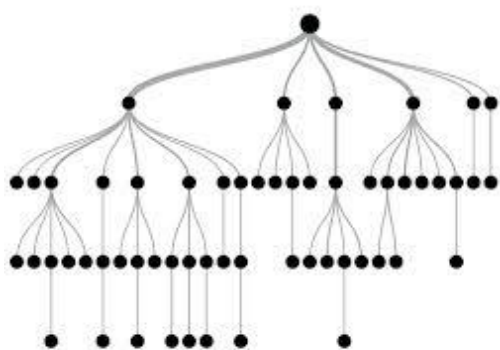
- 目标：让计算机系统更智能
- 方法：大数据+计算能力+复杂模型+高效算法→智能



常用机器学习模型



观察量 \rightarrow 决策的数学模型



主要讨论监督模型

搜索广告



Baidu 百度 [新闻](#) [网页](#) [贴吧](#) [知道](#) [音乐](#) [图片](#) [视频](#) [地图](#) [文库](#) [更多»](#)

北京美食

百度一下

推荐: [用手机随时随地地上百度](#)

[北京美食](#) 首选国内领先的吃喝玩乐信息平台-易吃易乐 [bj.echiele.com](#) 推广链接

[北京美食](#) 首选国内领先的吃喝玩乐信息平台-易吃易乐,每天有上百万网

● [易吃易乐](#) ● [餐饮美食](#) ● [休闲娱乐](#) ● [美容美发](#)

[找北京美食?来DaoDao.com](#) [www.daodao.com](#)

找[北京美食](#)?DaoDao.com为您提供210000条[北京市](#)旅游点评/攻略.

[北京美食-大众点评网](#)

根据合理的商区、地标和[美食](#)商户分类系统,为你提供[北京](#)83892家[美食](#)商户,并通过海量亲身消费者的点评聚合,以各种评分、星级的标准让你选择。

[www.dianping.com/beijing/f...](#) 2013-7-5 - [百度快照](#)

[北京美食攻略_北京美食推荐_美食街,小吃,指南-驴妈妈旅游网](#)

驴妈妈旅游网关于[北京美食](#)攻略,包含更多北京特色美食小吃(美食,餐饮,娱乐),【旅游预订】打折门票,周边酒店,自由行及跟团游信息,就在([www.lvmama.com](#))

[www.lvmama.com/travel/zhongguo_beiji...](#) 2013-6-29 - [百度快照](#)

[北京有什么特色美食?_百度知道](#)

13个回答 - 提问时间: 2011年12月25日

最佳答案: 1.烤鸭:在[北京](#)您要是想吃到便宜实惠的烤鸭,您可以去便宜坊、大鸭梨、安贞烤鸭店。当然您要是想吃最地道的烤鸭那就去和平门的全聚德。 2.涮羊肉:地...

[zhidao.baidu.com/question/3585625...](#) 2013-1-27 - [百度快照](#)

[北京美食_百度百科](#)

[北京美食](#)guide是一款让你随时随地掌握[北京美食](#)信息的手机软件。[北京美食](#)拥有详尽的地图,十多种美食分类。

[基本信息](#) - [软件介绍](#) - [安装指南](#) - [分辨率](#) - [软件截图](#)

[baike.baidu.com/](#) 2013-07-03

[在北京市搜索北京美食_百度地图](#)



A. [辣尚瘾\(人大店\)](#) - (010)82650566

★★★★★ 1229条评论

北京美食

外文名: Guide

版本: V1.8.0

软件大小: 3941KB

来自[百度百科](#)>>

相关食物



[北京小吃](#)



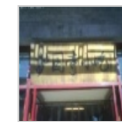
[北京烤鸭](#)



[门钉肉饼](#)



[褡裢火烧](#)



[护国寺小吃](#)

推广链接

[北京美食餐厅预定有折扣](#)

上咕嘟妈咪,方便轻松享优惠[北京美食](#);咕嘟妈咪,不让亲朋排队等.

[www.gudumami.cn](#)

[北京美食,金鼎鱼香渔村欢迎..](#)

金鼎鱼香生态渔村,特色全鱼宴,灶台柴锅水库鱼,柴锅柴鸡,特色烧鸽

[www.myjdyx.com](#)

[北京美食 刷雅酷卡 乐享无限..](#)

找[北京美食](#),精选[北京美食](#)折扣优惠!吃喝玩乐尽在雅酷卡网!

[www.yacool.com](#)

机器学习问题



- 点击率（CTR）预估
- 问题规模：
 - 数据存储和管理：上万台机器
 - 数据量：百亿到千亿级
 - 特征数：百亿到千亿级（稀疏离散值特征）
- 大型线性Logistic Regression模型
- 计算技术：分布式同步CPU并行计算

语音识别

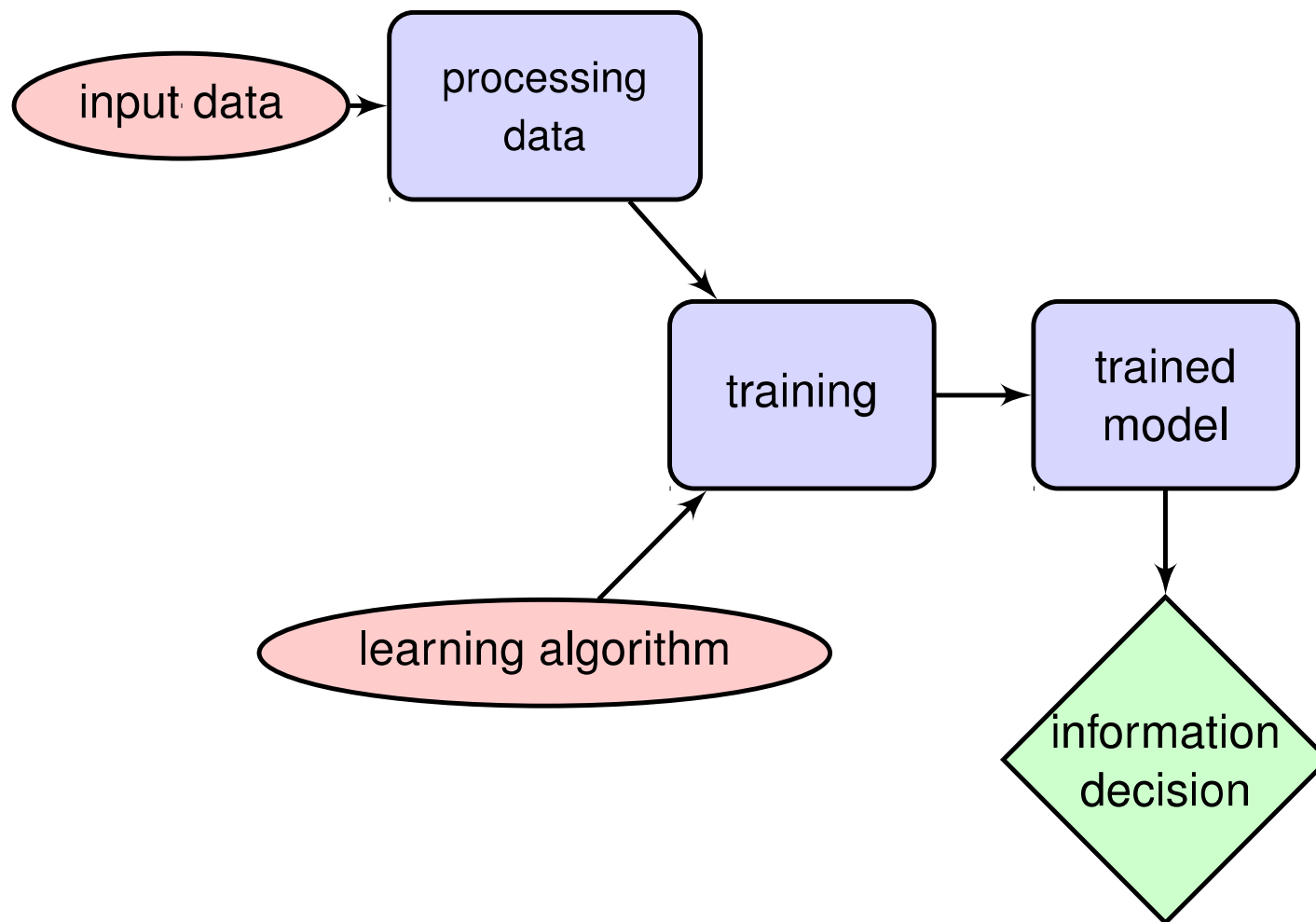


机器学习问题



- 把声学信号变成文字：多分类问题
- 问题规模
 - 万小时级语料
 - 百亿级训练数据
 - 上万类别；几百维特征（稠密连续值特征）
- 深度神经网络模型
- 计算技术：分布式异步GPU计算

机器学习流程



大规模机器学习



- 基础架构

分布式数据存储，管理，和分析

分布式CPU/GPU计算平台

- 算法

模型和特征提取

数据抽样

大型机器学习模型优化

数据管理

- Hadoop: 大数据存储 de facto standard
分布式文件系统 (HDFS)

Map-Reduce



可用于机器学习特征提取

计算性能



CPU: complex tasks
Large memory (128G)
Few cores (8)
Peak 100+Gflops

适合：

稀疏离散特征
树模型



GPU: simple tasks extreme
parallel
Small memory (5G)
Many cores (2K cores)
Peak 3Tflops

适合：

稠密连续值
深层神经网络计算

数学问题

- 大型机器学习训练优化问题：

$$\min_w \frac{1}{n} \sum_{i=1}^n f_i(w)$$

- 分布式多机并行训练

问题的分配方式



分配数据到多机

- 每个机器有所有模型参数
- 每个机器也不同数据

分配特征到多机

- 每个机器有所有数据的一些特征
- 每个机器有不同参数

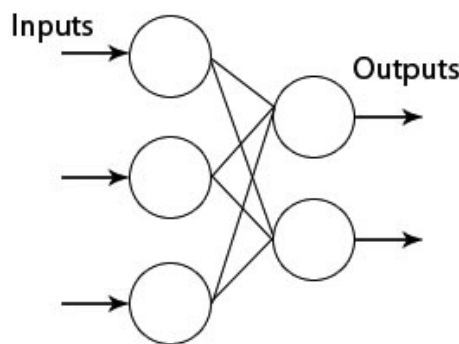
分配数据和特征到多机

- 每个机器有一些特征和一些参数

大型线性模型

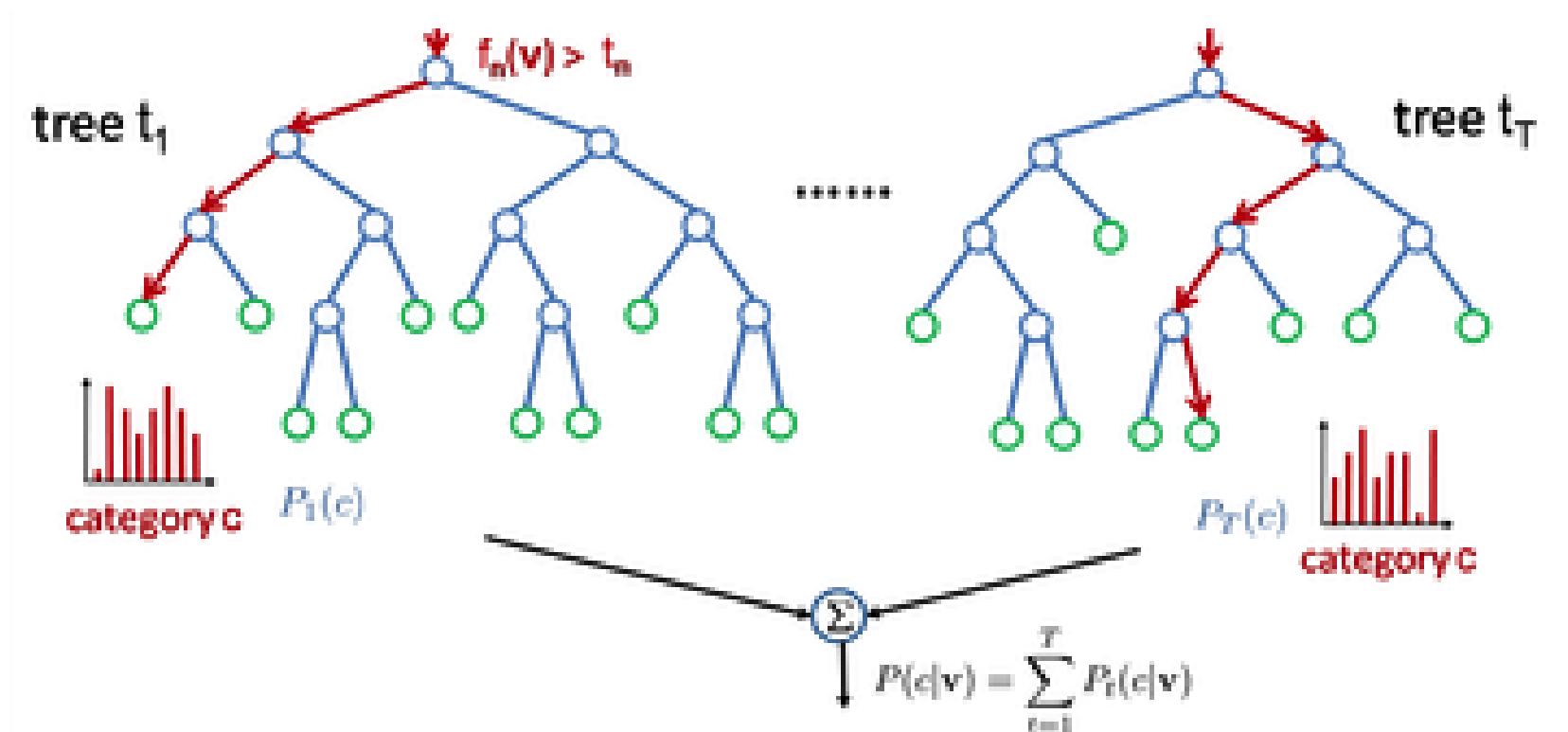
$$\min_w \frac{1}{n} \sum_{i=1}^n f_i(w)$$

$$f_i(w) = \ln(1 + e^{-w^\top x_i y_i})$$



多机CPU分布式计算：特征和样本发给多机

树模型

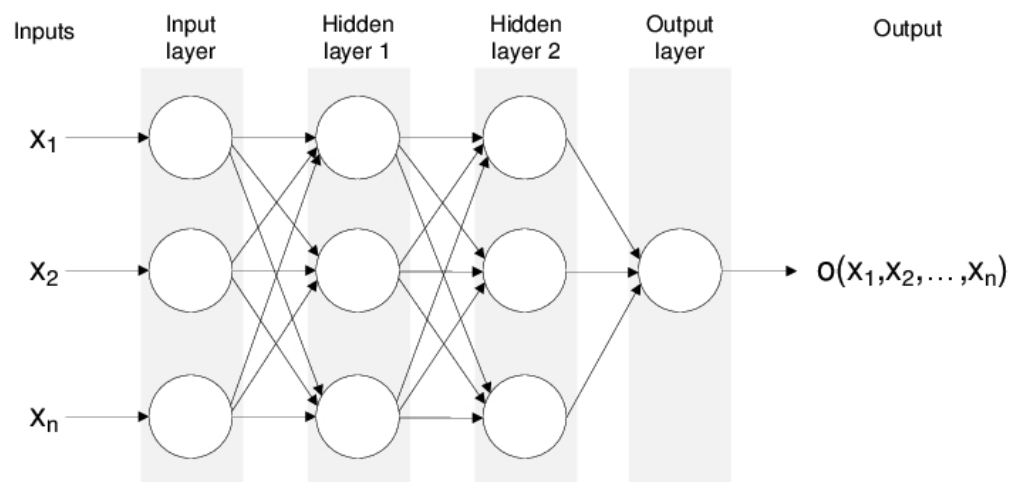


多机CPU分布式计算

把特征发到不同机器

深度神经网络

$$\min_w \frac{1}{n} \sum_{i=1}^n f_i(w)$$

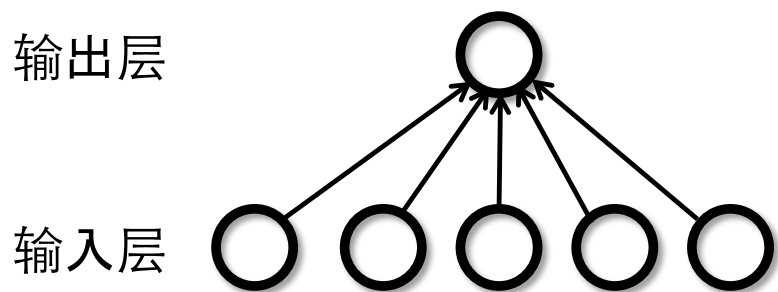


多机GPU/CPU分布式计算

从浅层到深度学习



浅层网络：
人工特征抽取
学习线性组合

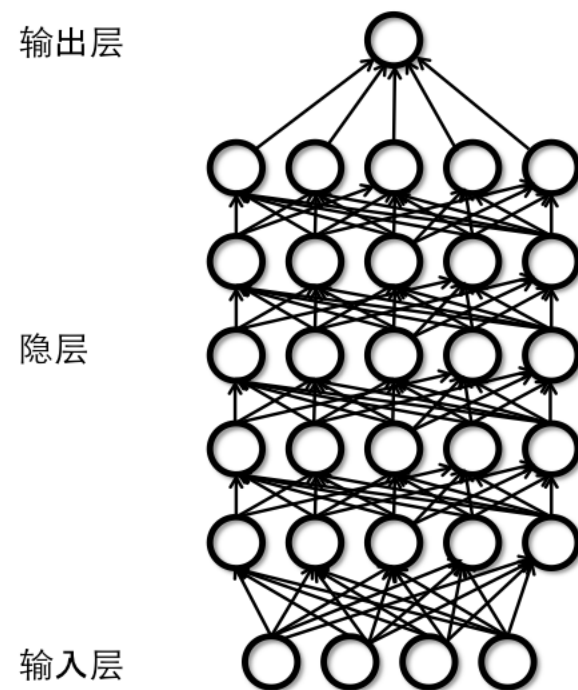


不含隐层的浅层学习模型

深层网络：

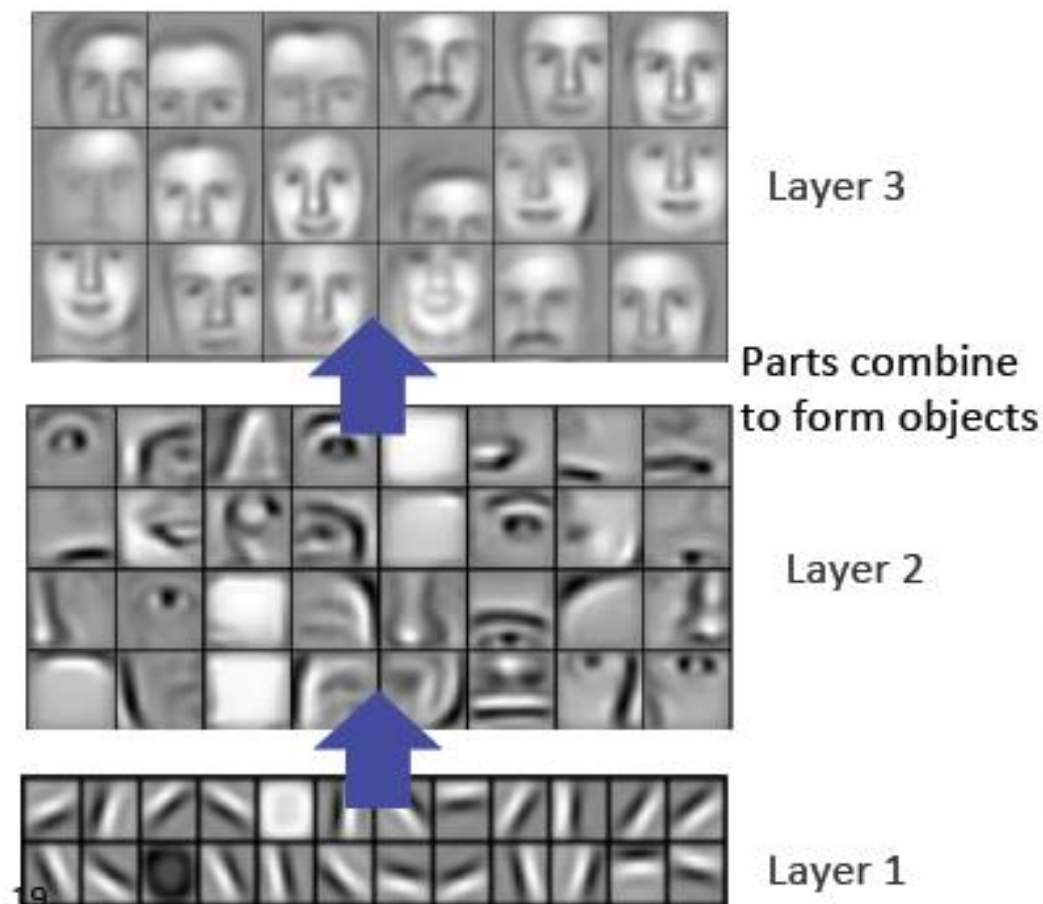
从原始特征出发

自动学习高级特征组合

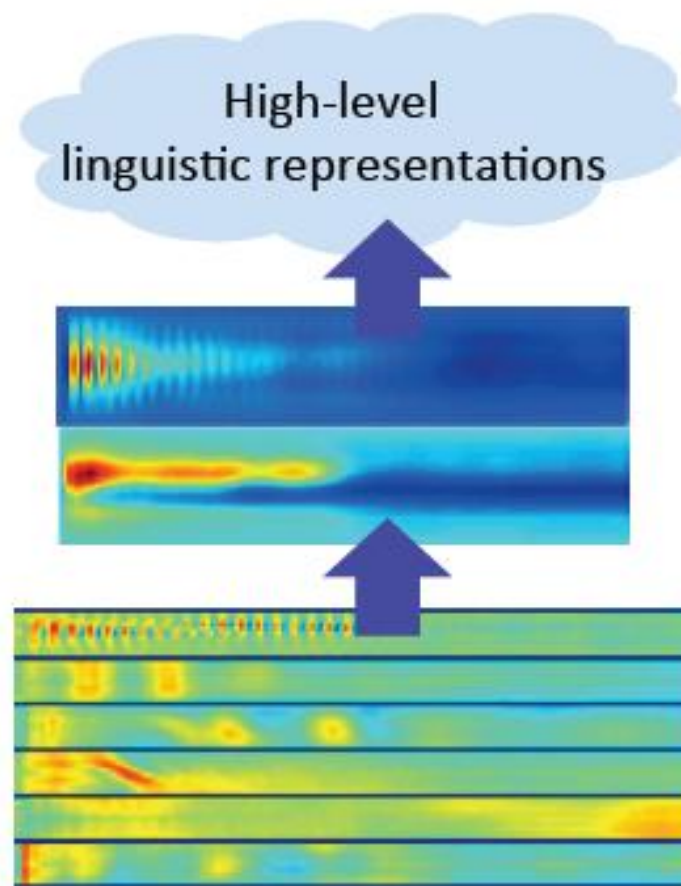


含多个隐层的深度学习模型

高级特征



Image



Speech

(Lee, Largman, Pham & Ng, NIPS 2009)
(Lee, Grosse, Ranganath & Ng, ICML 2009)

深度学习成功条件

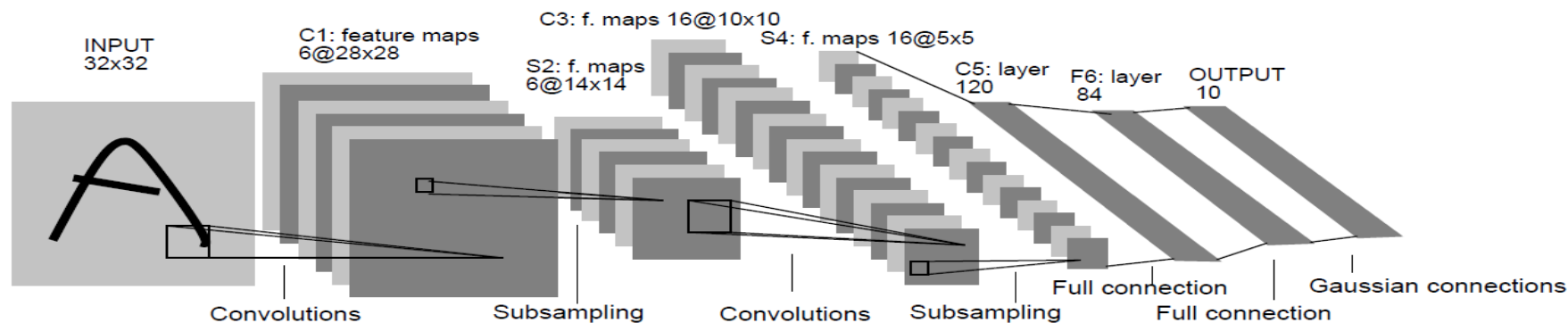


- 2010-今：在工业界取得巨大成功
 - 复杂模型
- 大数据：100x
- 大规模计算能力：1000x
- 大数据+计算能力+复杂模型+高效算法

- 图像： 数千万训练样本
- 语音： 数百亿训练样本
- 广告： 数千亿训练样本
- ...

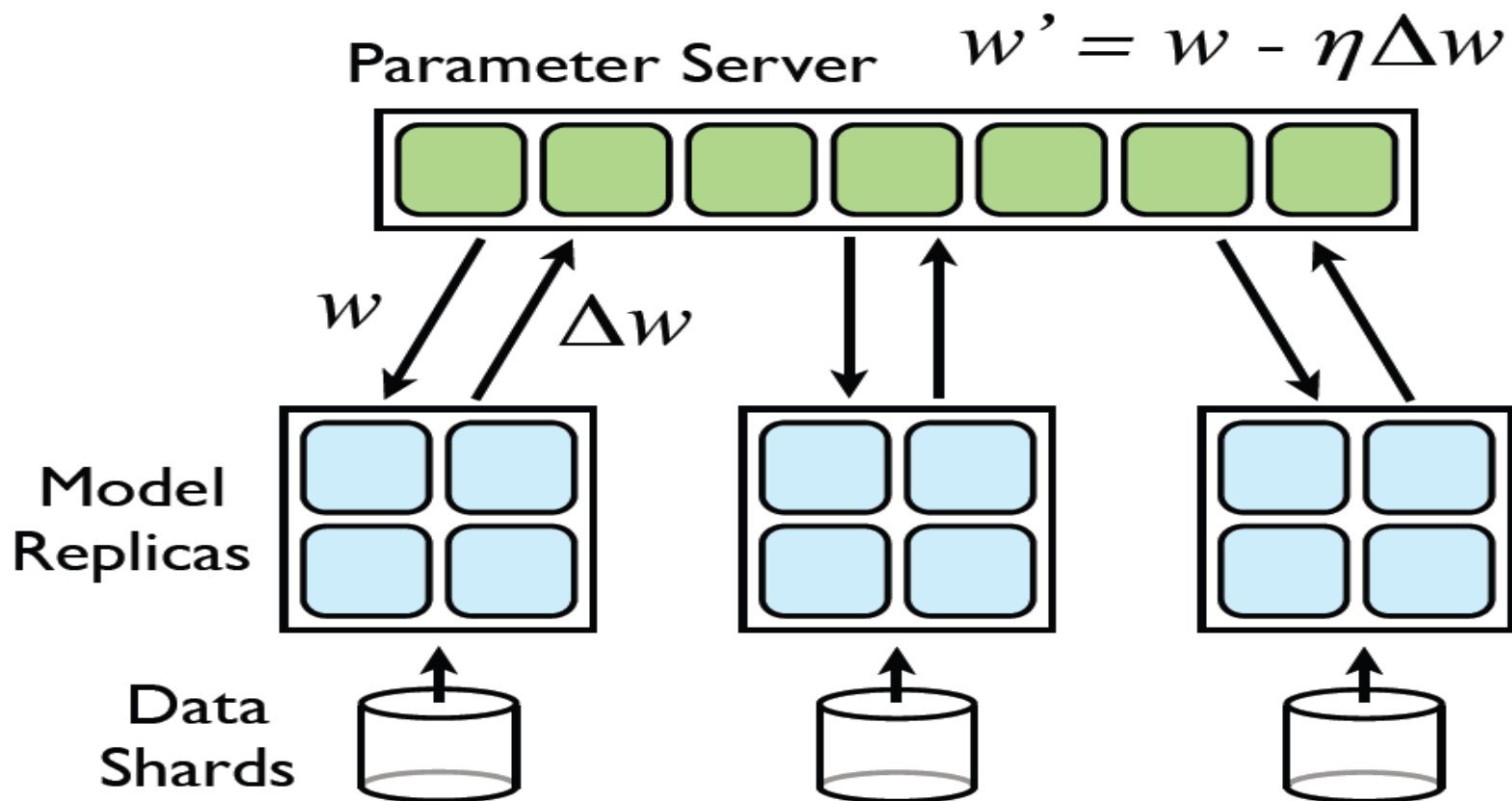
训练数据每年成倍增长 ...

深度学习模型：图像



计算资源和算法

- 几十台GPU并行计算
- 分布式算法



大数据研究



问题

用什么数据解决什么问题

数据

数据融合，结构；噪声过滤和纠偏；数据抽样方法；数据降维

模型

图模型，树模型，深度神经网络；

和问题相关的模型结构和高级特征的学习

分布式计算

同步和异步；鲁棒性；大模型；理论分析

大数据和深度学习的意义



- 目标：计算机智能和人工智能
- 手段：
 - 大数据
 - 复杂模型
 - 计算能力和算法
 - 系统集成
- 深度学习：最接近人脑的复杂模型
 - 目前向人工智能走得最近的方法