

大数据技术

汪卫

复旦大学计算机科学技术学院

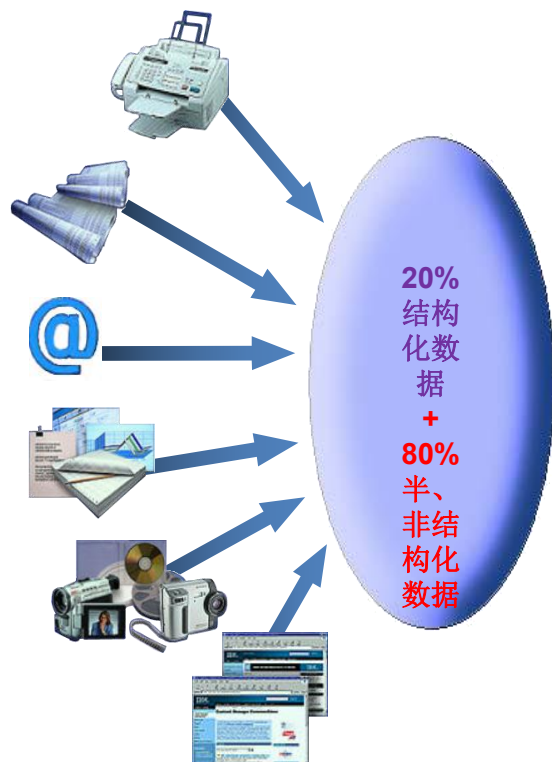
上海数据科学重点实验室

weiwang1@fudan.edu.cn

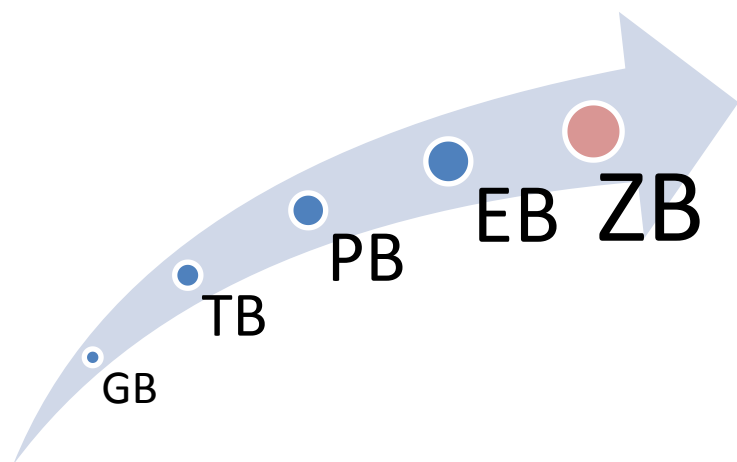
大数据时代的背景

- “大数据”的诞生

— 半个世纪以来，随着计算机技术全面融入社会生活，信息爆炸已经**积累**到了一个开始引发**变革**的程度。



大数据时代数据的爆炸增长



1PB (拍字节) = 2^{50} 字节

1EB (艾字节) = 2^{60} 字节

1ZB (泽字节) = 2^{70} 字节

地球上至今总共的数据量：

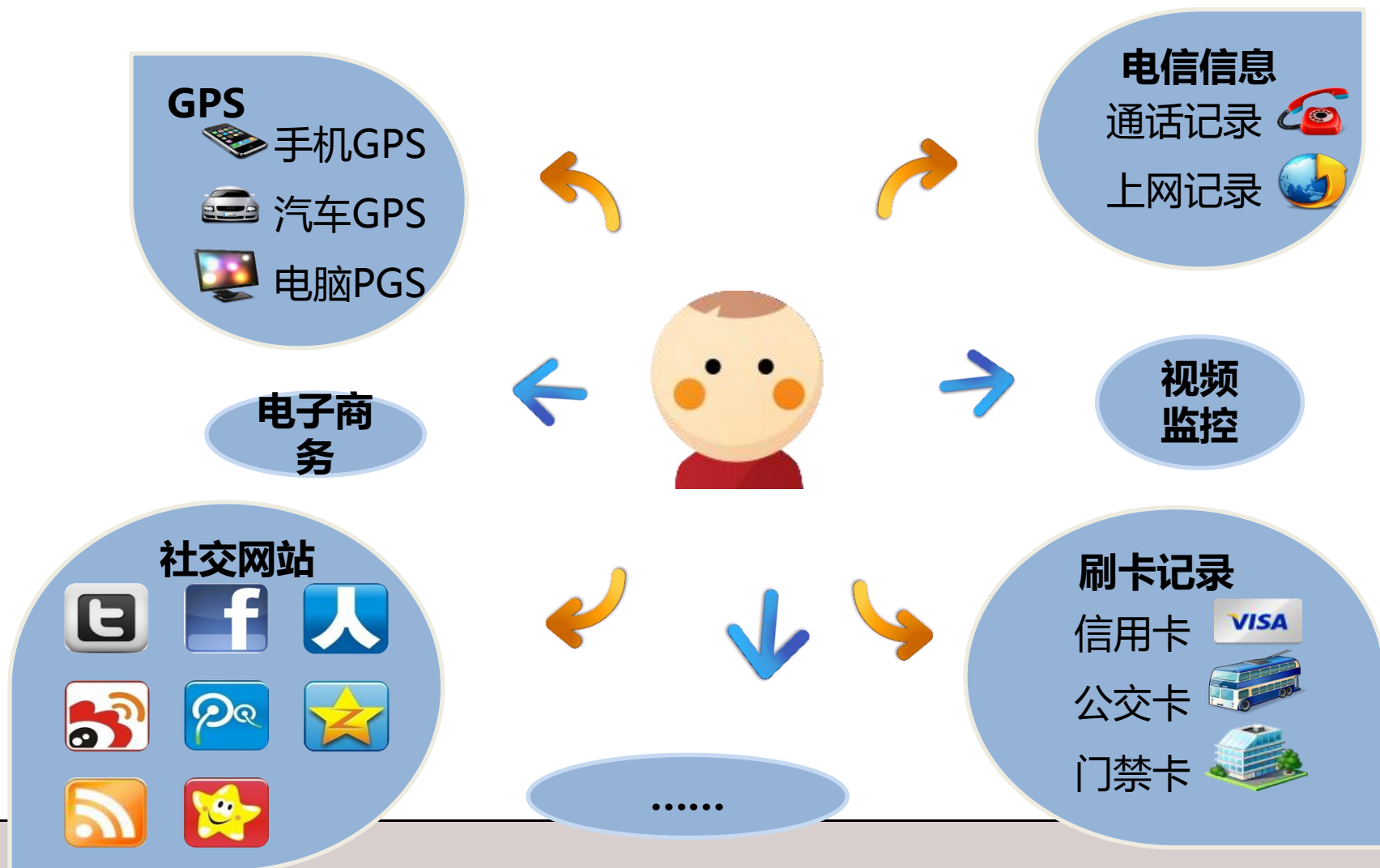
在**2006** 年，个人用户才刚刚迈进**TB**时代，全球一共新产生了约**180EB**的数据；

在**2011** 年，这个数字达到了**1.8ZB**。

而有市场研究机构预测：

到**2020** 年，整个世界的**数据总量**将会增长**44** 倍，达到**35.2ZB** (**1ZB=10 亿 TB**) ！

信息对人们生活的高度覆盖



大数据的4V特性

Volume

数据量巨大

全球在2010 年正式进入ZB 时代，
IDC预计到2020 年，全球将总共拥有
35ZB 的数据量

Variety

结构化数据、半结构化数据和非结构化数据

如今的数据类型早已不是单一的文本形式，
订单、日志、音频，能力提出了更高的要求

Value

沙里淘金，价值密度低

以视频为例，一部一小时的视频，在连续不间断监控过程中，可能有用的数据仅仅只有一两秒。如何通过强大的机器算法更迅速地
完成数据的价值“提纯”是目前大数据汹涌
背景下亟待解决的难题

Velocity

实时获取需要的信息

大数据区别于传统数据最显著的特征。如今
已是ZB时代，在如此海量的数据面前，处理
数据的效率就是企业的生命

大数据能给我们带来什么？

- 大数据赋予我们洞察未来的能力
 - 马云成功预测2008 年经济危机
 - “2008 年初,阿里巴巴平台上整个买家询盘数急剧下滑, 欧美对中国采购在下滑。海关是卖了货, 出去以后再获得数据; 我们提前半年时间从询盘上推断出世界贸易发生了变化了。”
 - 通常而言, 买家在采购商品前, 会比较多家供应商的产品, 反映到阿里巴巴网站统计数据中, 就是查询点击的数量和购买点击的数量会保持一个相对的数值, 综合各个维度的数据可建立用户行为模型。因为数据样本巨大, 保证用户行为模型的准确性。因此在这个案例中, 询盘数据的下降, 自然导致买盘的下降。



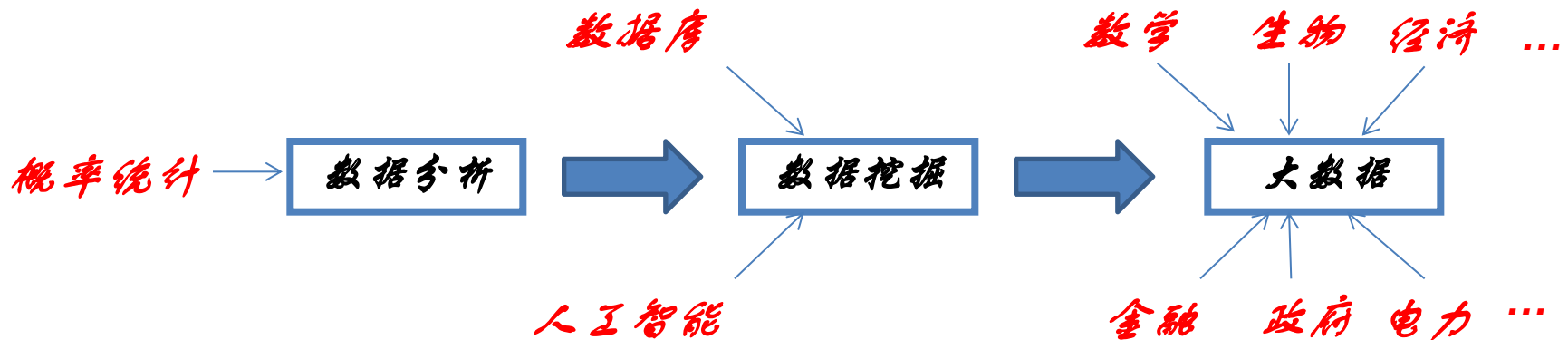
人类从依靠自身判断做决定到依靠数据做决定的转变，也是大数据作出的最大贡献之一。——《大数据时代》

大数据的定义

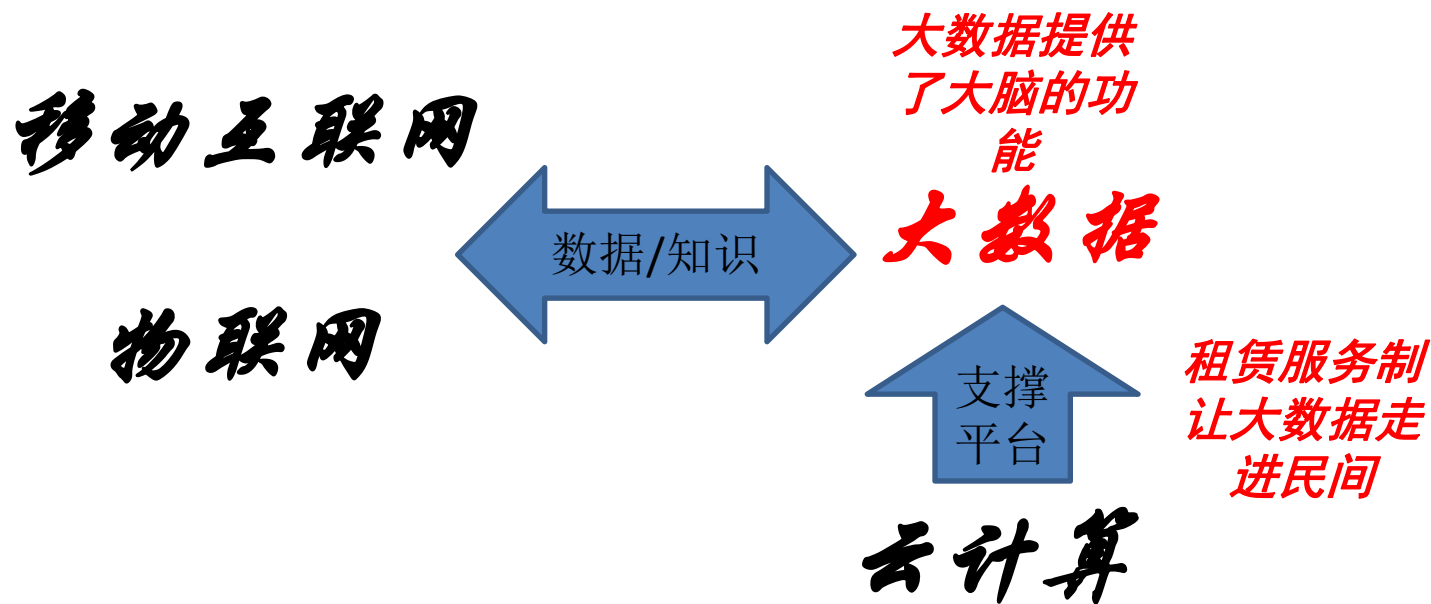
- 大数据(big data), 或称巨量资料, 指的是所涉及的资料量规模巨大到无法通过目前主流软件工具, 在合理时间内达到撷取、管理、处理、并整理成为帮助企业经营决策更积极目的的资讯。
- 大数据指不用随机分析法(抽样调查)这样的捷径, 而采用所有数据的方法
- 大数据是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产
- 大数据的故事
 - 洛杉矶警察局和加利福尼亚大学合作利用大数据预测犯罪的发生。
 - google流感趋势(Google Flu Trends)利用搜索关键词预测禽流感的散布。
 - 统计学家内特·西尔弗(Nate Silver)利用大数据预测2012美国选举结果
 - 麻省理工学院利用手机定位数据和交通数据建立城市规划

大数据分析技术的形成基础

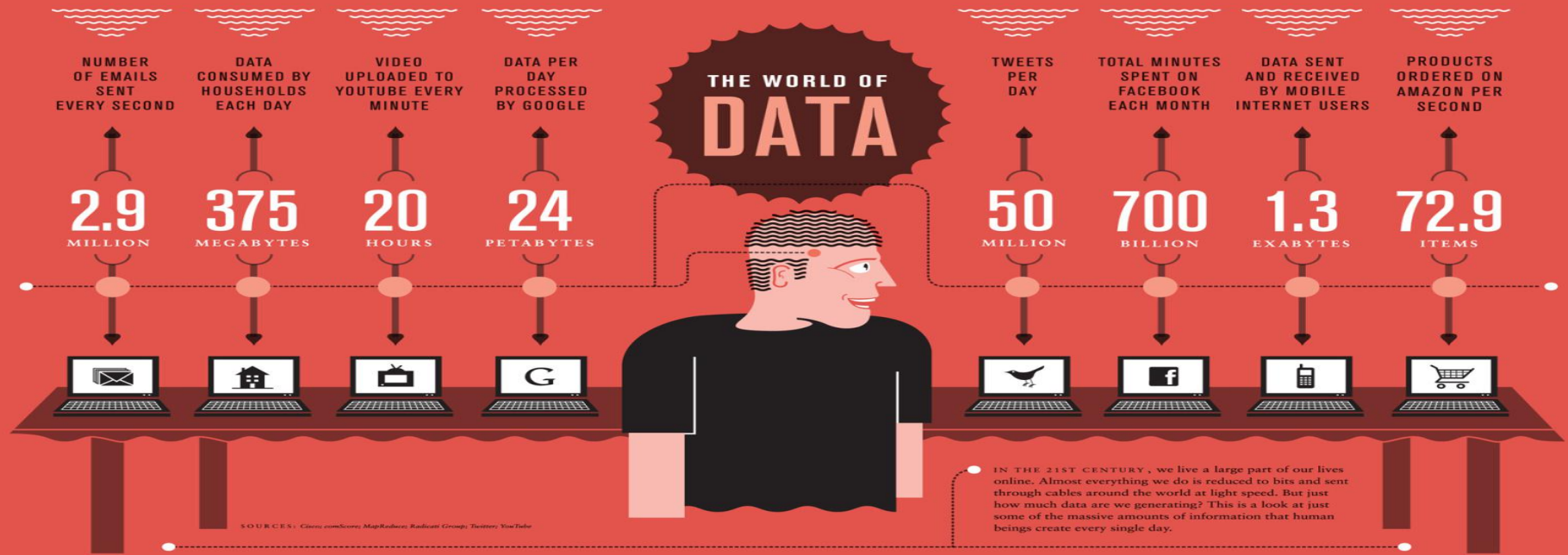
**从数据到知识到智能是人们
长期以来追求的目标！**



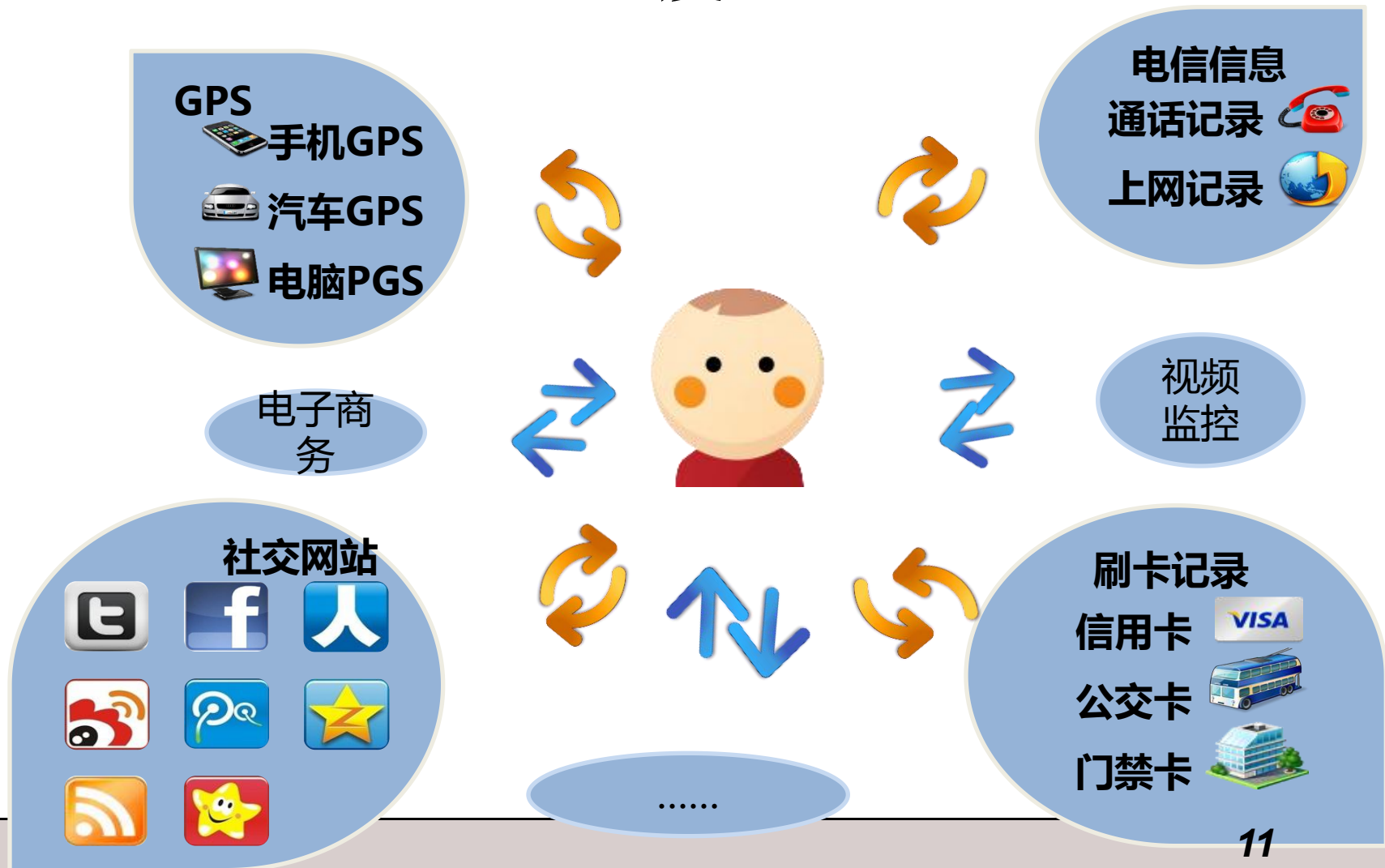
大数据的形成基础



大数据之大



大数据之全-信息对人们生活的高度覆盖



大数据思维

关联



采样与全数据

- 小数据时代
 - 随机采样利用最小的数据获得更多的信息
 - 分析的精确性同采样的随机性相关
- 大数据时代
 - 随着收集数据的便捷性和数据处理速度的加快，我们可以分析更多的数据，甚至是关于某些现象的所有数据
- 相比小数据，大数据更能反映事物的多样性
 - 在社交网络分析中人们发现
 - 社区中删除连接关系比较多的人，会降低社区的效率但是不会破坏社区的结构
 - 若删除同社区外的很多人有连接的节点会使整个社团破碎成很多小块

大数据要容忍不准确

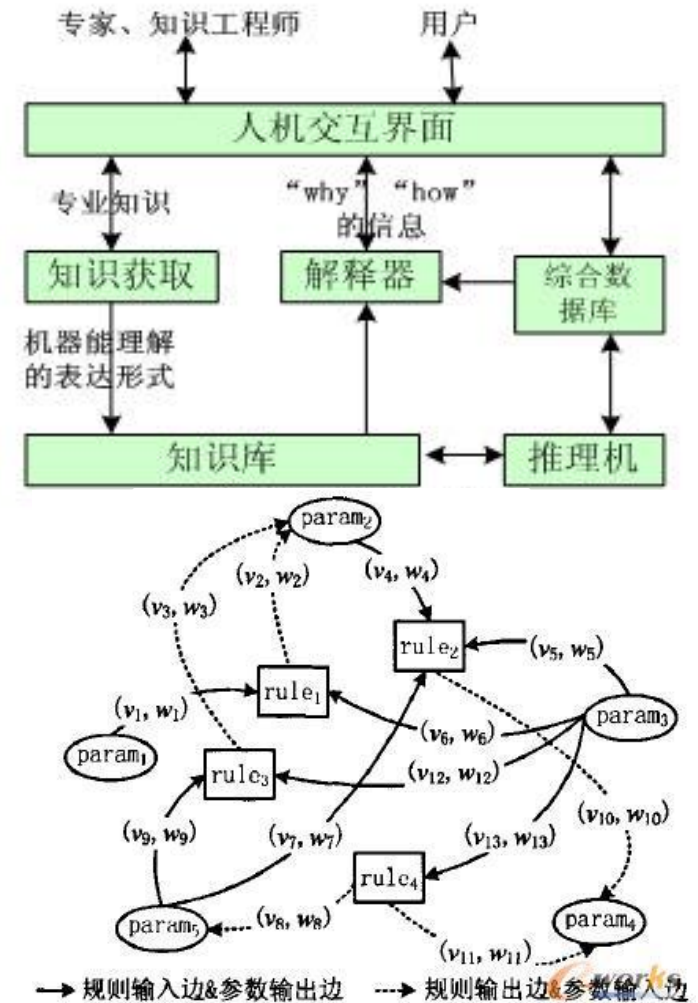
- 数据的大幅度增加将导致错误数据的引入
 - 用一个温度计统计葡萄园的温度，只要保证这个温度计工作正常就可以了
 - 如果用100个温度计，则很难保证所有的温度计都运行正常
 - 我们获取的大量非结构化数据（95%）精确性难以保证
- 但是大数据能够更快地反映大的趋势
 - 美国劳工统计局雇佣大量的人员收集全美90个城市的商品价格
 - 麻省理工学院的经济学家利用互联网上收集的50万中商品的价格于2008年9月预测到通货紧缩的趋势，而官方在11月才正式发布

相关性是大数据核心概念

- 相关关系是量化两个数据值之间的数理关系，如果一个数据值增加另一个也增加
- 建立数据之间相关性
 - 亚马逊发掘两个产品之间的关联
 - 沃尔玛发现飓风来临时手电和蛋挞的销量会同时增加
- 相关性分析有助于我们分析一个现象，但是还要揭示其内部的运作机制，大数据时代让我们有更多的数据和更好的分析工具发现关联关系

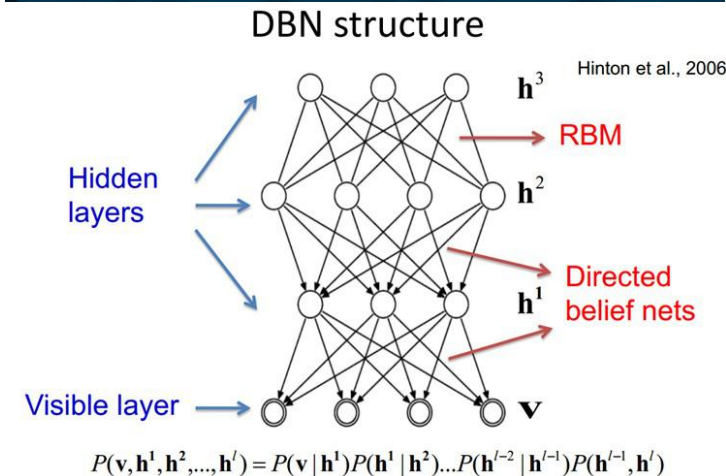
大数据与人工智能

- 以规则推理为基础的人工智能在上世纪末达到顶点
 - 专家系统
 - 辅助决策系统
- 规则的复杂性和基础事实的缺乏制约了传统人工智能技术的发展
- 神经网络、贝叶斯等技术为新一代人工智能技术的发展奠定了基础

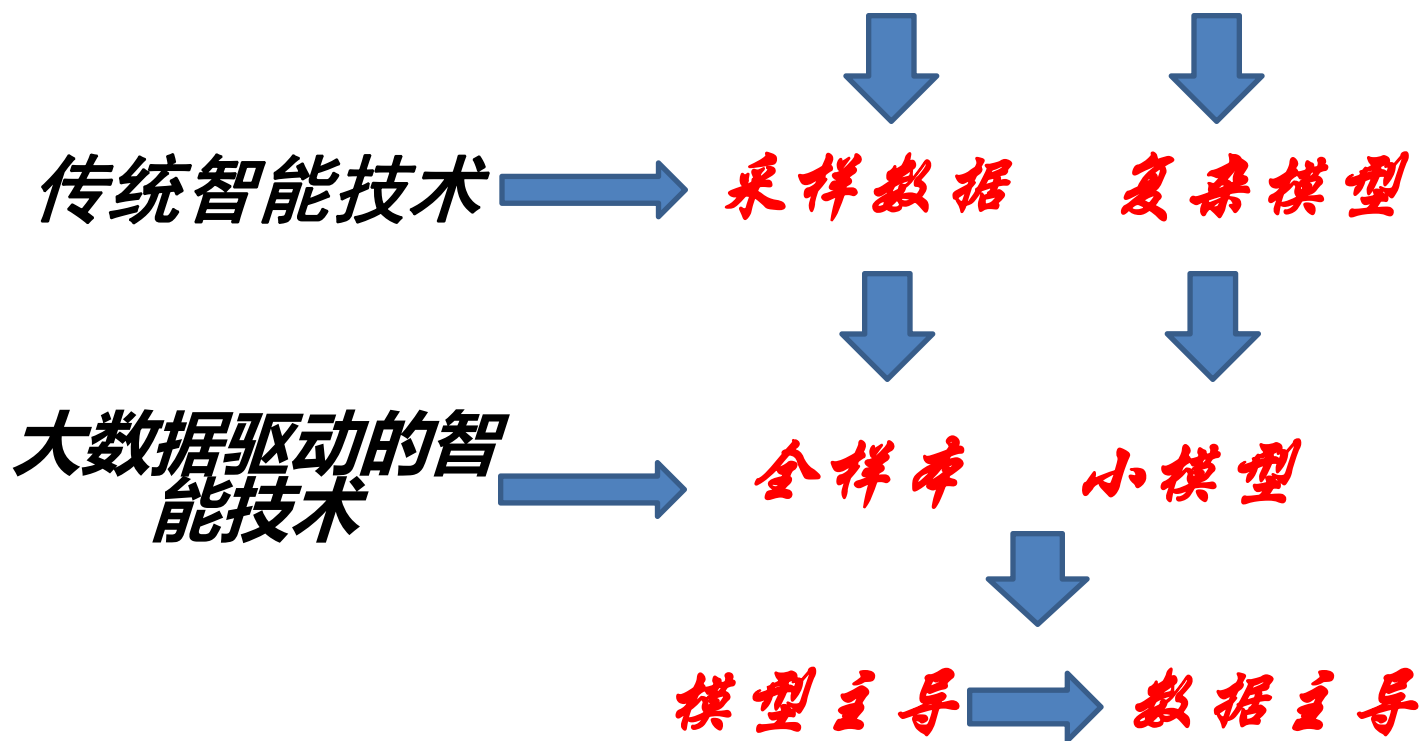


大数据与人工智能

- 机器学习、深度学习、知识图谱技术的发展形成了新一代人工智能技术的技术基础
 - 面向海量不同层次和类型的数据的信息抽取与整理使得面向通用或领域的知识图谱（知识库）系统成为可能
 - 海量的标注或未标注数据使得深度学习的效果得到了保证
- 以Waston、Alpha Go、图像识别等系统成为新一代人工智能技术典型代表



智能 = 数据 + 推理



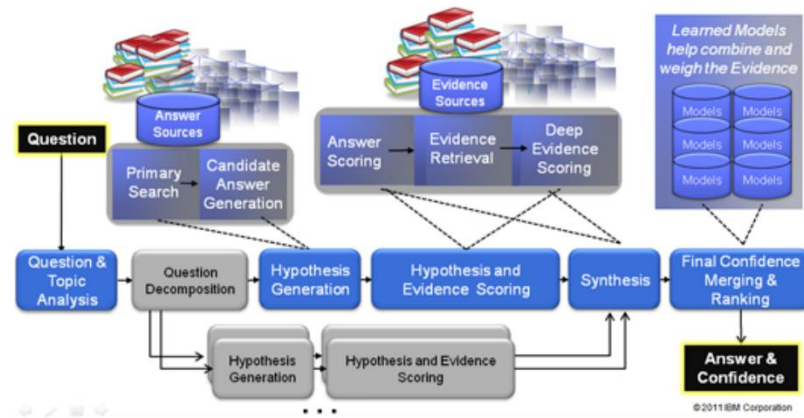
海量数据隐含着更准确的事实

- 2001年, 一个基于事实的简短问答研究, 如提问: Who shot Abraham Lincoln? 在很大的数据集时, 只要使用简单的模式匹配方法, 找到在 “shot Abraham Lincoln” 前面的部分即可快速得到准确答案: John Wilkes Booth
- 2007, Brants et al. 描述了一个基于2万亿个单词训练数据集的语言模型, 比较了当时最先进的Kneser-Ney smoothing 算法与他们称之为 “stupid backoff “ (愚蠢退避) 的简单算法, 最后发现, 后者在小数据集时效果不佳, 但在大数据集时, 该算法最终居然产生了更好的语言模型!

大数据集上的简单算法能比小数据集上的复杂算法产生更好的结果!

基于大数据的分析与融合

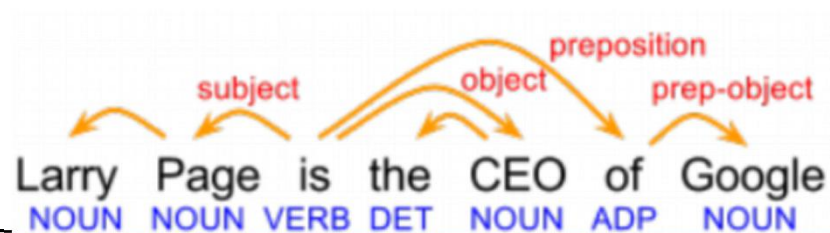
- “沃森”
 - 2011年2月17日由IBM和美国德克萨斯大学联合研制的超级电脑“沃森”(Watson)今日在美国最受欢迎的智力竞猜电视节目《危险边缘》中击败该节目历史上两位最成功的选手肯-詹宁斯和布拉德-鲁特，成为《危险边缘》节目新的王者
 - 基础
 - 存储了大量图书、新闻和电影剧本资料、辞海、文选和《世界图书百科全书》(World Book Encyclopedia)等数百万份资料。每当读完问题的提示后，“沃森”就在不到三秒钟的时间里在长达2亿页的漫漫资料里展开搜索



基于大数据的分析与融合

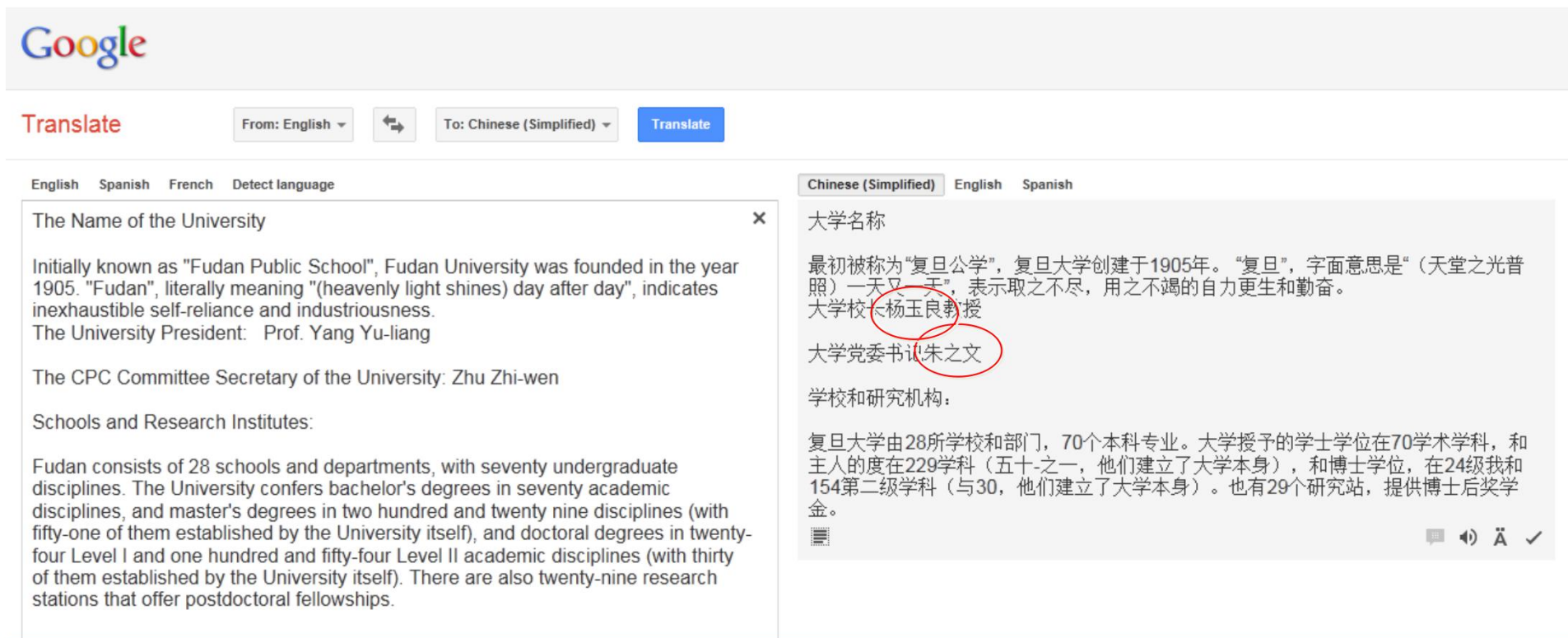
- 机器翻译系统

- 大数据+机器学习+语言学
- Google、微软等均提供了基于统计的机器翻译系统
- 统计机器翻译系统的基础
 - 超过1百万个单词的双语文本语料库
 - 超过10亿个单词的语种语料库
 - 从两个语料库训练的统计模型



基于大数据的分析与融合

- 机器翻译系统



The screenshot displays the Google Translate interface. The source text on the left is in English, and the target text on the right is in Chinese (Simplified). The interface includes a search bar, a 'Translate' button, and a language selection dropdown. The source text is titled 'The Name of the University' and describes the history and structure of Fudan University. The target text is titled '大学名称' and provides the Chinese translation of the same information. Red circles highlight specific names in the Chinese text: '杨玉良' (Yang Yuliang) and '朱之文' (Zhu Zhiwen).

From: English **To:** Chinese (Simplified) **Translate**

English Spanish French Detect language

The Name of the University

Initially known as "Fudan Public School", Fudan University was founded in the year 1905. "Fudan", literally meaning "(heavenly light shines) day after day", indicates inexhaustible self-reliance and industriousness.
The University President: Prof. Yang Yu-liang

The CPC Committee Secretary of the University: Zhu Zhi-wen

Schools and Research Institutes:

Fudan consists of 28 schools and departments, with seventy undergraduate disciplines. The University confers bachelor's degrees in seventy academic disciplines, and master's degrees in two hundred and twenty nine disciplines (with fifty-one of them established by the University itself), and doctoral degrees in twenty-four Level I and one hundred and fifty-four Level II academic disciplines (with thirty of them established by the University itself). There are also twenty-nine research stations that offer postdoctoral fellowships.

Chinese (Simplified) English Spanish

大学名称

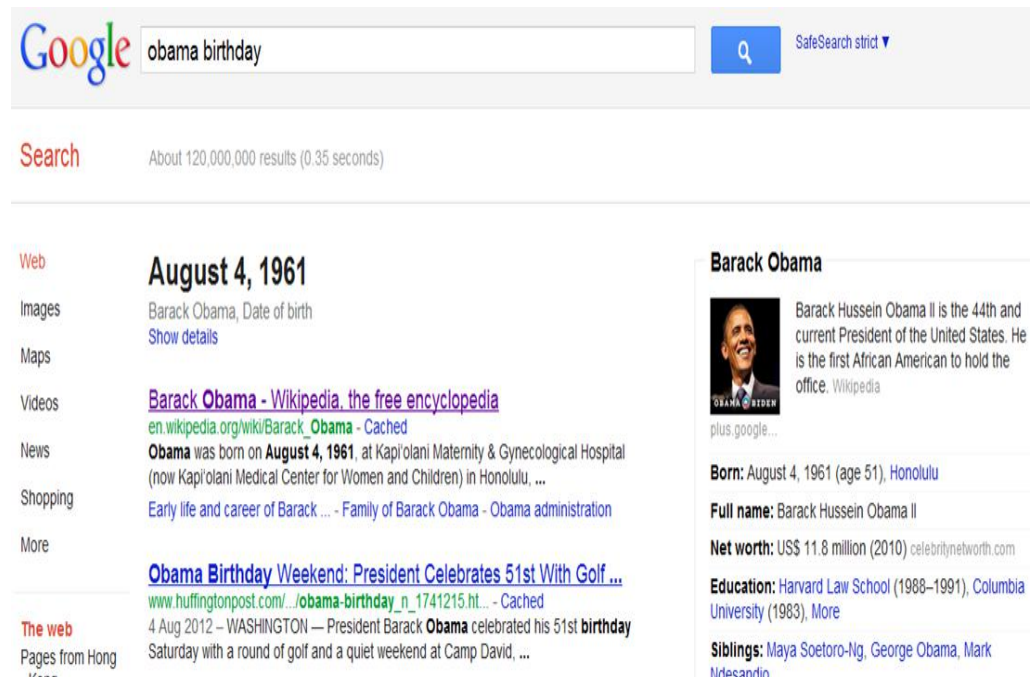
最初被称为“复旦公学”，复旦大学创建于1905年。“复旦”，字面意思是“（天堂之光普照）一天又一天”，表示取之不尽，用之不竭的自力更生和勤奋。
大学校长杨玉良教授
大学党委书记朱之文
学校和研究机构：

复旦大学由28所学校和部门，70个本科专业。大学授予的学士学位在70学术学科，和主人的度在229学科（五十之一，他们建立了大学本身），和博士学位，在24级我和154第二级学科（与30，他们建立了大学本身）。也有29个研究站，提供博士后奖学金。

知识图谱

Google于2012年5月份提出的从多种数据源抽取并构建的一种用于实现**语义**搜索的知识库

- 从包括CIA Factbook、Freebase以及Wiki等数据源，抽取知识，表达了近**5亿**实体，近**35 Billion**的事实信息



Google search for "obama birthday" results.

Search About 120,000,000 results (0.35 seconds)

Web

August 4, 1961
 Barack Obama, Date of birth
[Show details](#)

[Barack Obama - Wikipedia, the free encyclopedia](#)
en.wikipedia.org/wiki/Barack_Obama - Cached

Obama was born on **August 4, 1961**, at Kapi'olani Maternity & Gynecological Hospital (now Kapi'olani Medical Center for Women and Children) in Honolulu, ...

[Early life and career of Barack ...](#) - [Family of Barack Obama](#) - [Obama administration](#)

Obama Birthday Weekend: President Celebrates 51st With Golf ...
www.huffingtonpost.com/.../obama-birthday_n_1741215.html - Cached

4 Aug 2012 - WASHINGTON — President Barack **Obama** celebrated his 51st birthday Saturday with a round of golf and a quiet weekend at Camp David, ...

Barack Obama

 Barack Hussein Obama II is the 44th and current President of the United States. He is the first African American to hold the office. [Wikipedia](#)

[plus.google...](#)

Born: August 4, 1961 (age 51), Honolulu

Full name: Barack Hussein Obama II

Net worth: US\$ 11.8 million (2010) [celebritynetworth.com](#)

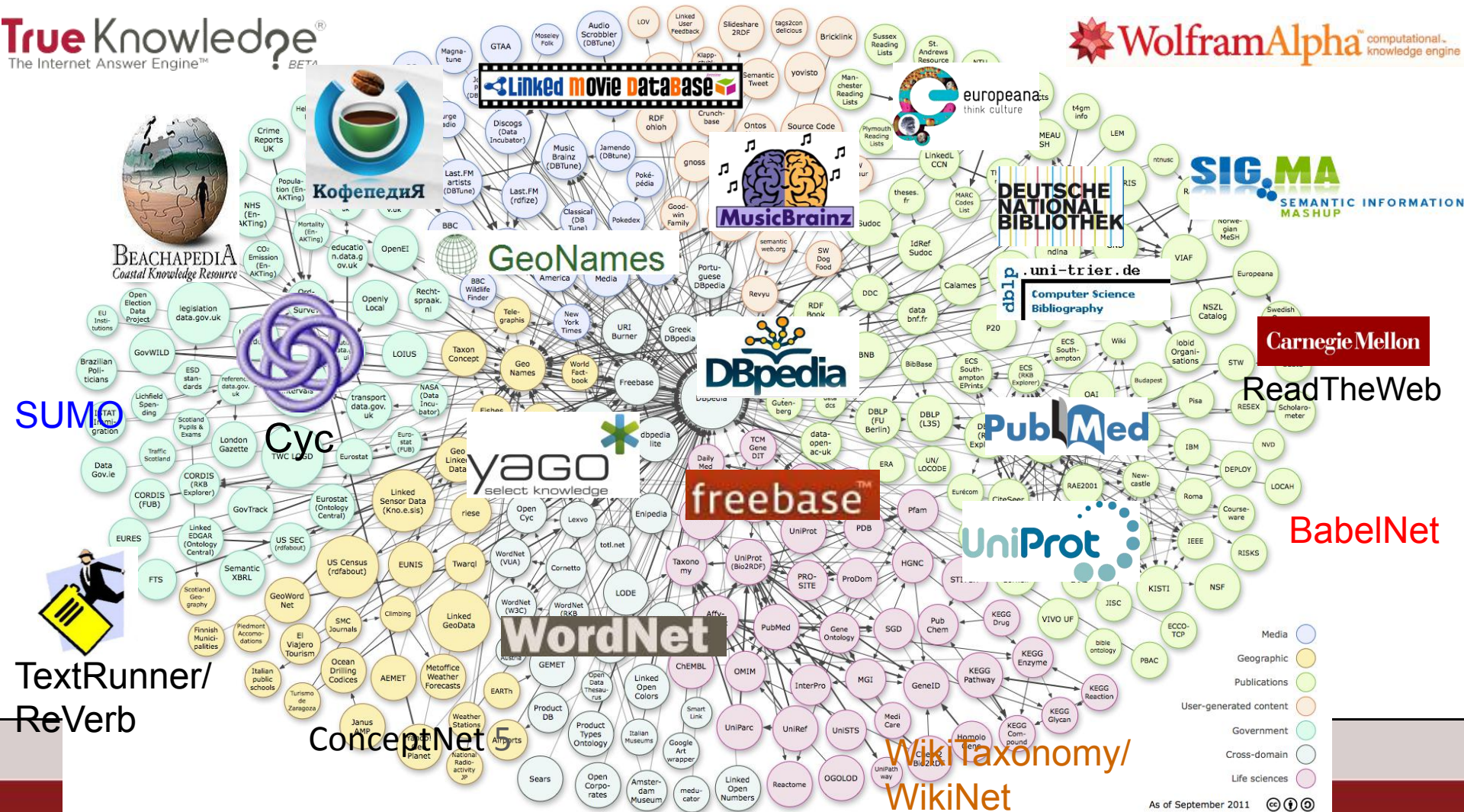
Education: Harvard Law School (1988–1991), Columbia University (1983), [More](#)

Siblings: [Maya Soetoro-Ng](#), [George Obama](#), [Mark Niles](#), [Cassandra](#)



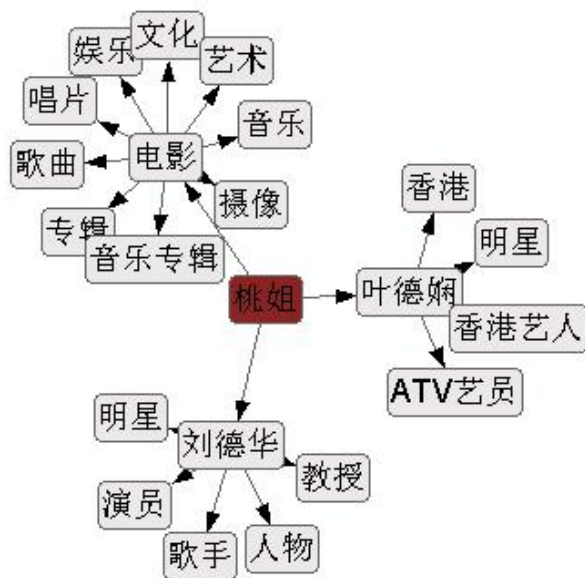
网络空间的知识库

250 Bio. SPO triples (RDF) and growing



复旦中文知识图谱平台

- 2000万实体，5000万关系
- 部分领域准确率，比如军事，准确率达到95%
- 链接: <http://gdm.fudan.edu.cn/Soso/index.jsp>

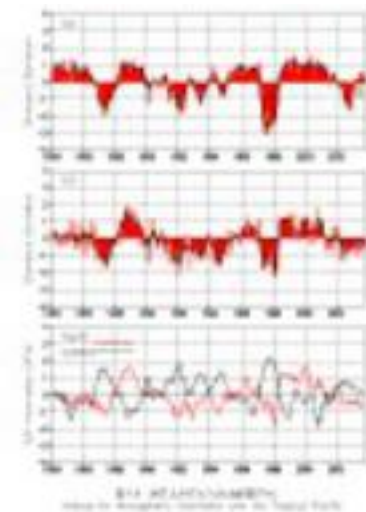


语义网 分词

搜索

例: 刘德华, 华仔, 刘德华 代表作品

物联网大数据管理



物联网技术提高了智能感知能力

谢谢！