



大数据与城市计算

城市大数据的来源和概念

姚尧 博士，副教授，高级工程师

地理与信息工程学院，地图制图学与地理信息工程

东京大学，空间情报科学研究中心，助教授

阿里巴巴集团，达摩院，访问学者

Email: yaoy@cug.edu.cn

办公地点：未来城校区地信楼522办公室





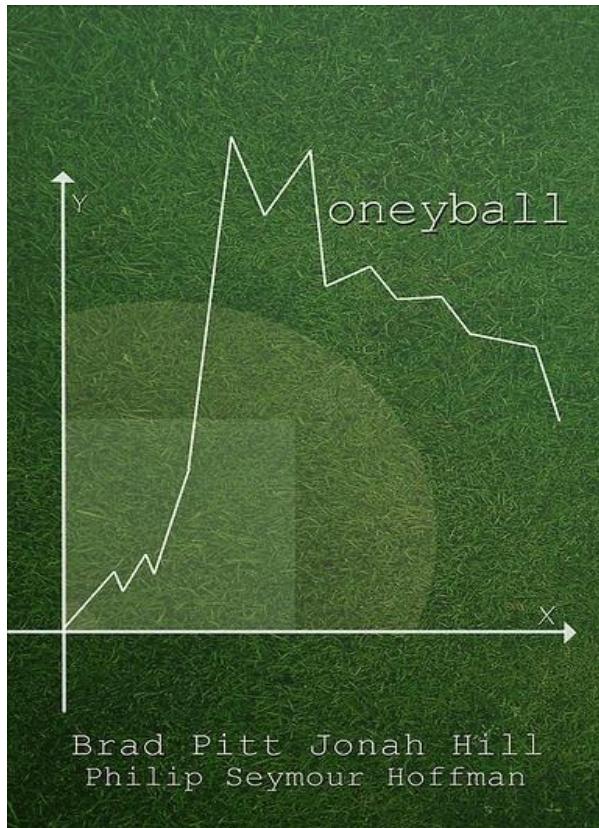
章节内容



- 1 大数据的来源
- 2 大数据的概念
- 3 大数据的影响
- 4 大数据的关键技术
- 5 大数据的计算模式
- 6 大数据产业
- 7 大数据与云计算、物联网的关系

- 1.1 引言
- 1.2 数据的本质是生产资料和资产
- 1.3 数据爆炸式增长为数据资产管理带来挑战
- 1.4 信息科技的进步、数据生产方式的变革以及政府的重视促进大数据时代的到来
- 1.5 城市计算概念
- 1.6 城市大数据的来源
- 1.7 城市大数据的获取

布拉德·皮特主演的《点球成金》是一部美国奥斯卡获奖影片，所讲述的是皮特扮演的棒球队总经理利用计算机数据分析，对球队进行了翻天覆地的改造，让一家不起眼的小球队能够取得巨大的成功。



基于历史数据，利用数据建模定量分析不同球员特点，合理搭配，重新组队；

打破传统思维，通过分析比赛数据，寻找“性价比”最高球员，运用数据取得成功；



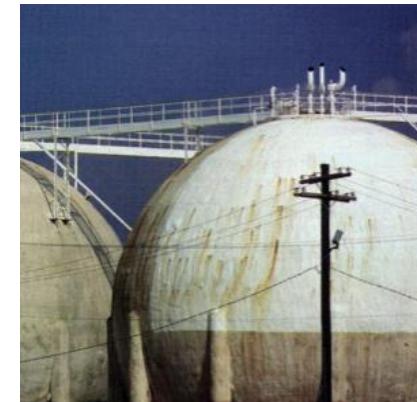
1.2| 数据的本质是生产资料和资产



仅供开采162年



仅供开采45年



仅供开采60年

数据不再是社会生产的“副产品”，而是可被二次乃至多次加工的原料，从中可以探索更大价值，它变成了生产资料。

过去3年数据总量被以往4万年还多

2013年,10分钟的信息总量将达1.8ZB

2010年全球数据总量1.2ZB,年增长50%



不可再生资源VS数据



上传6600张新照片到flickr	13000+个iPhone应用下载	Skype上37万+分钟的语音通话	Twitter上发布98000+新微博
YouTube上上传600+新视频	发出1.68亿+条Email	Facebook上更新69.5万+条新状态	淘宝光棍节10680+个新订单
12306出票1840+张			





尽管“数据是资产”概念已经广为人知，但“如何管理数据资产”仍然缺少成熟理论以及工具手段。

什么是数据资产？

数据资产是企业及组织拥有或控制的，能够带来经济利益的数据资源。

存在什么问题？

- 定义不统一
- 分配不透明
- 数据源不规范
- 治理无力
- 数据不开放
- 加工流程混乱
- 分布杂乱
- 应用低效
- 评估手段缺失
- 处理缓慢
-
- 运营缺失

需求
发现

数据资产管理是企业或组织采取的各种管理活动，用以保证数据资产的安全完整，合理配置和有效利用，从而提高带来的**经济效益**，保障和促进各项事业发展。该领域是大数据时代企业布局竞争的核心，也是目前市场空白。



为什么传统数据管理方式并不适合数据资产管理要求？

传统数据管理方式

外部性管理，依赖管理力度和执行
自律，成难毁易

元数据

数据
稽核

管理
制度

从**范围**来看，非结构化数据、内外部数据混搭、云化处理等都会冲击传统管理模式

挑战1

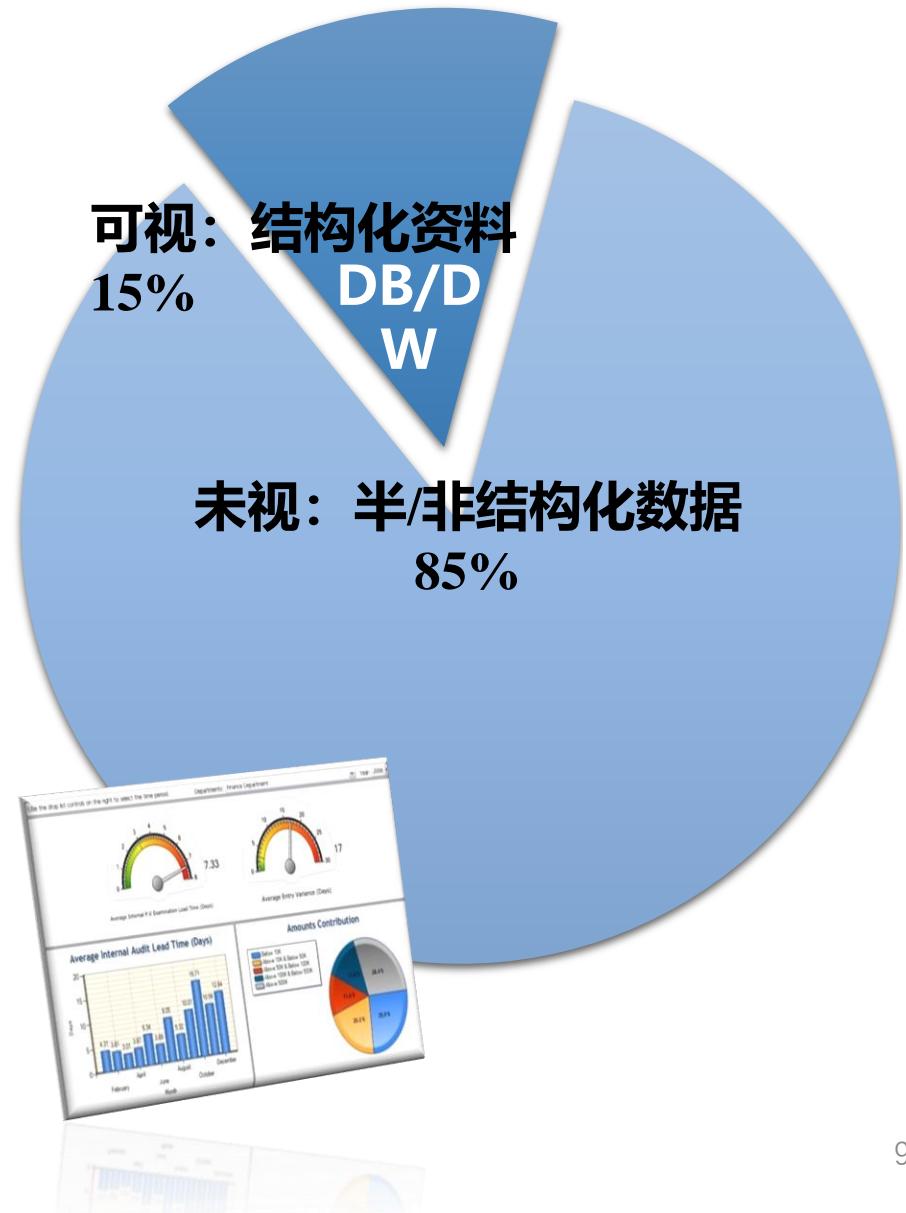
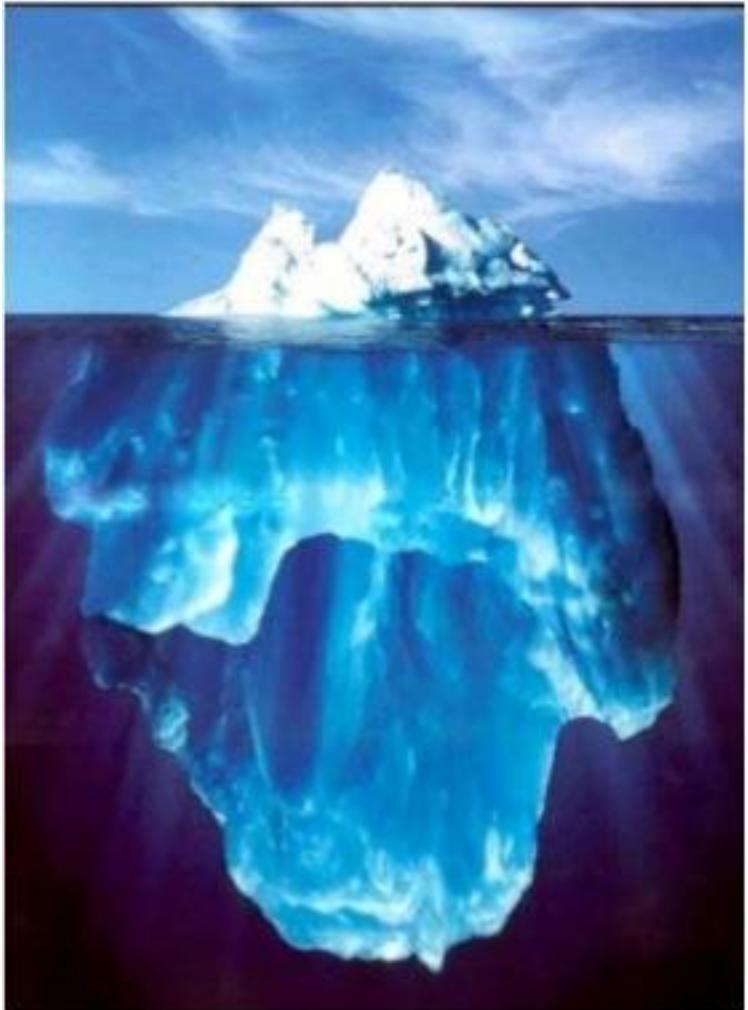
从**形式**来看，数据加工的复杂度和速度要求越来越高，也对传统管理效率提出挑战

挑战2

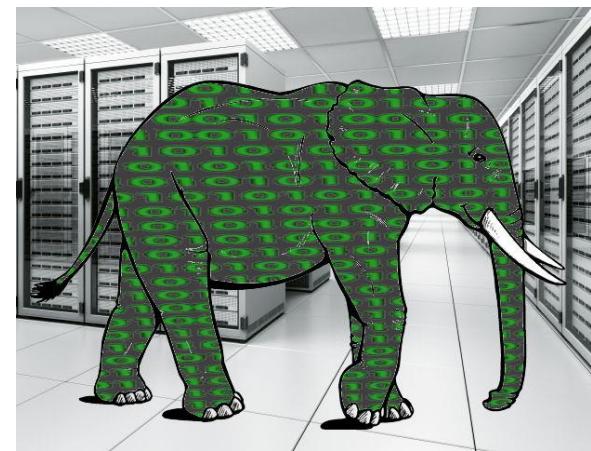
从**内涵**来看，数据的交换、租赁、交易等各种创新模式，也要求新的管理手段

挑战3

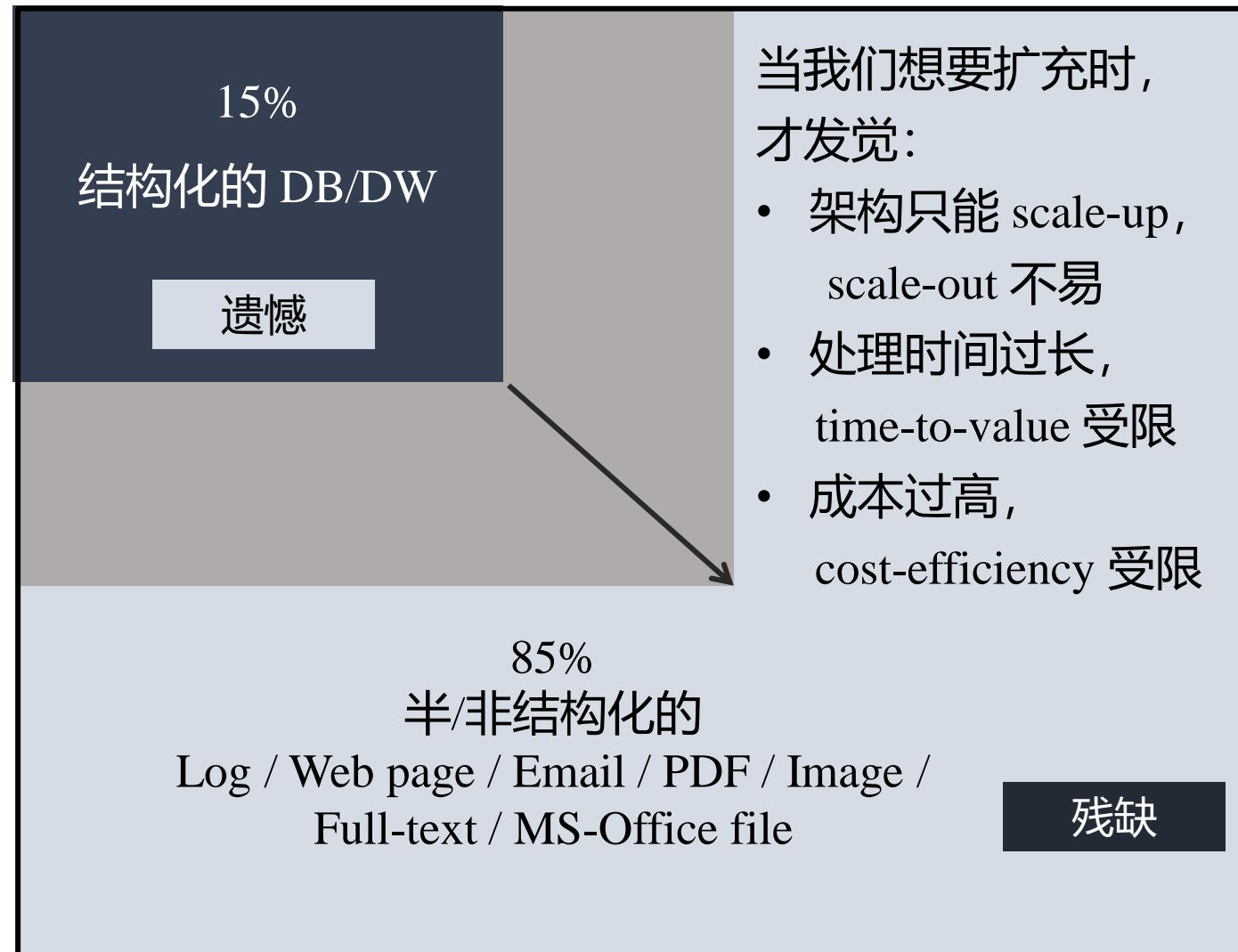
1.3 需要不同“看”数据的方式



1.3 需要更高性价比的数据计算与储存方式



计算更快 存储更省





储存
Storing

每天几百 GB、几 TB 的资料，且持续成长中

计算
Processing

在收数据的同时做必要的前置处理 (preprocessing)
，并区分数据处理的优先等级 (prioritizing)

管理
Managing

如何有效的避免因硬件毁坏所导致的资料损毁

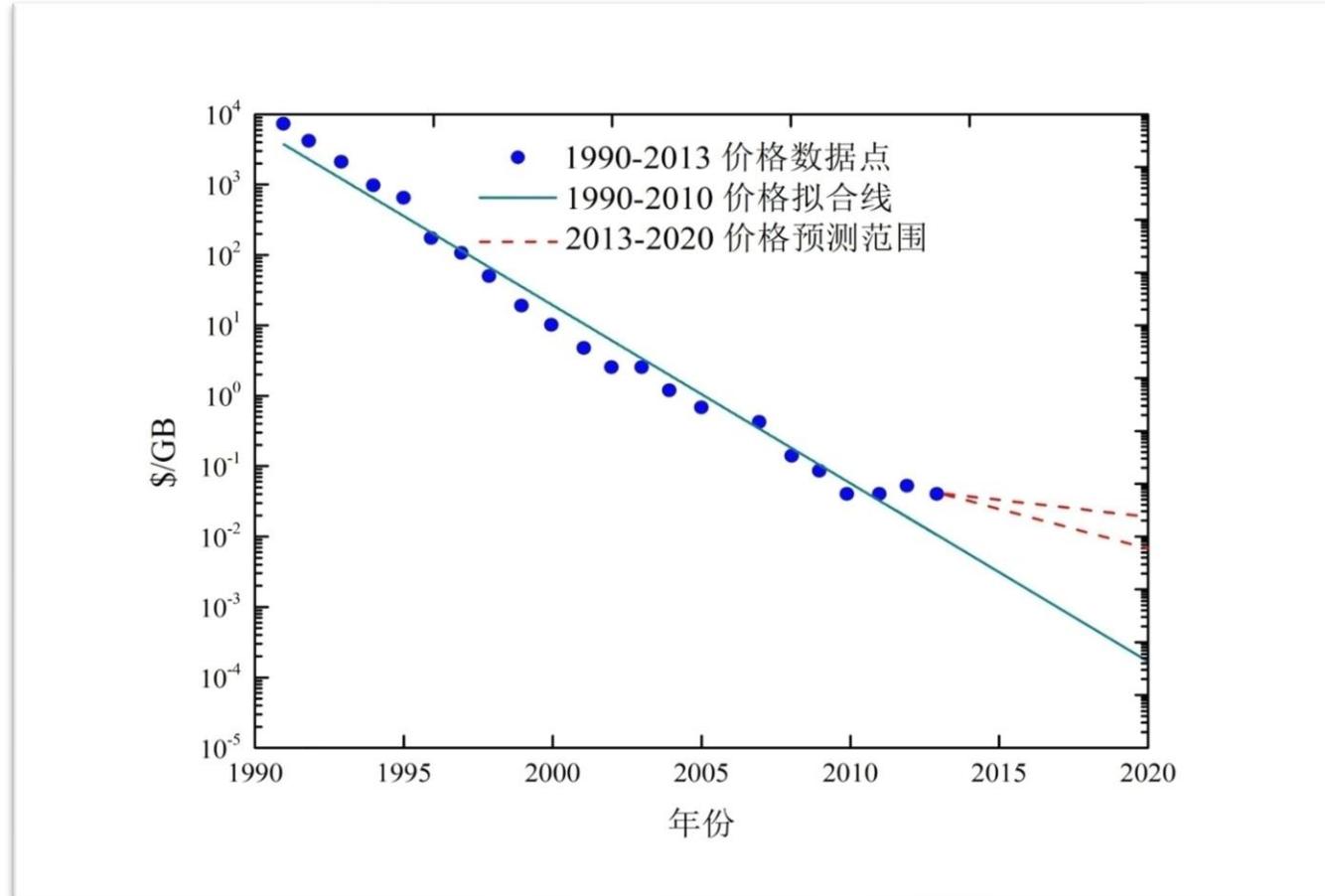
分析
Analyzing

如何从中挖掘出所关注事件的 pattern 或 behavior

1.4 | 信息科技的进步



1. 存储设备容量不断增加



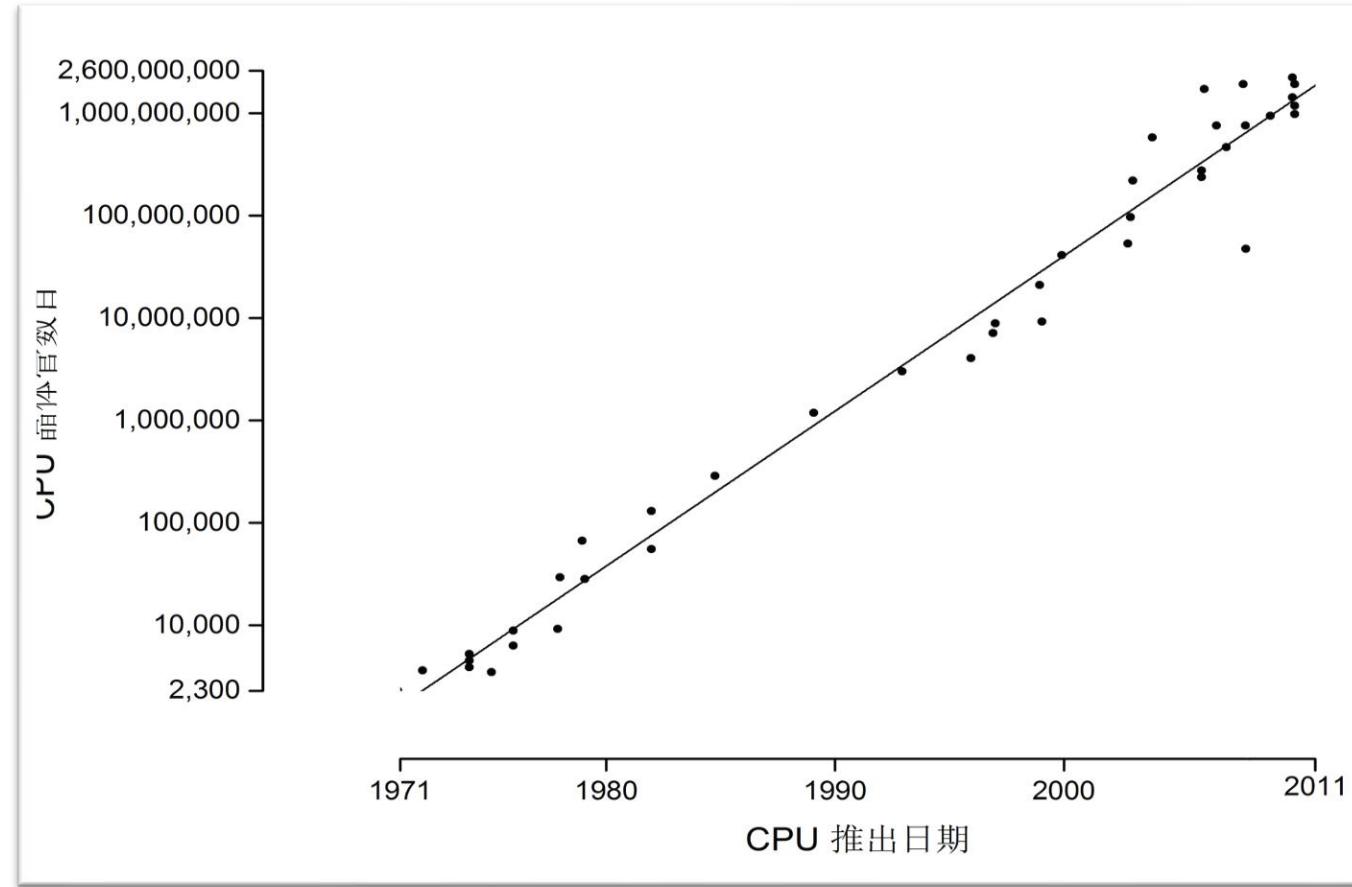
存储价格随时间变化情况

1.4 | 信息科技的进步



来自斯威本科技大学 (Swinburne University of Technology) 的研究团队，在2013年6月29日刊出的《自然通讯 (Nature Communications)》杂志的文章中，描述了一种全新的数据存储方式，可将1PB (1024TB) 的数据存储到一张仅DVD大小的聚合物碟片上。

2. CPU处理能力大幅提升



CPU晶体管数目随时间变化情况

1.4 | 数据产生方式的变革



3. 网络带宽不断增加

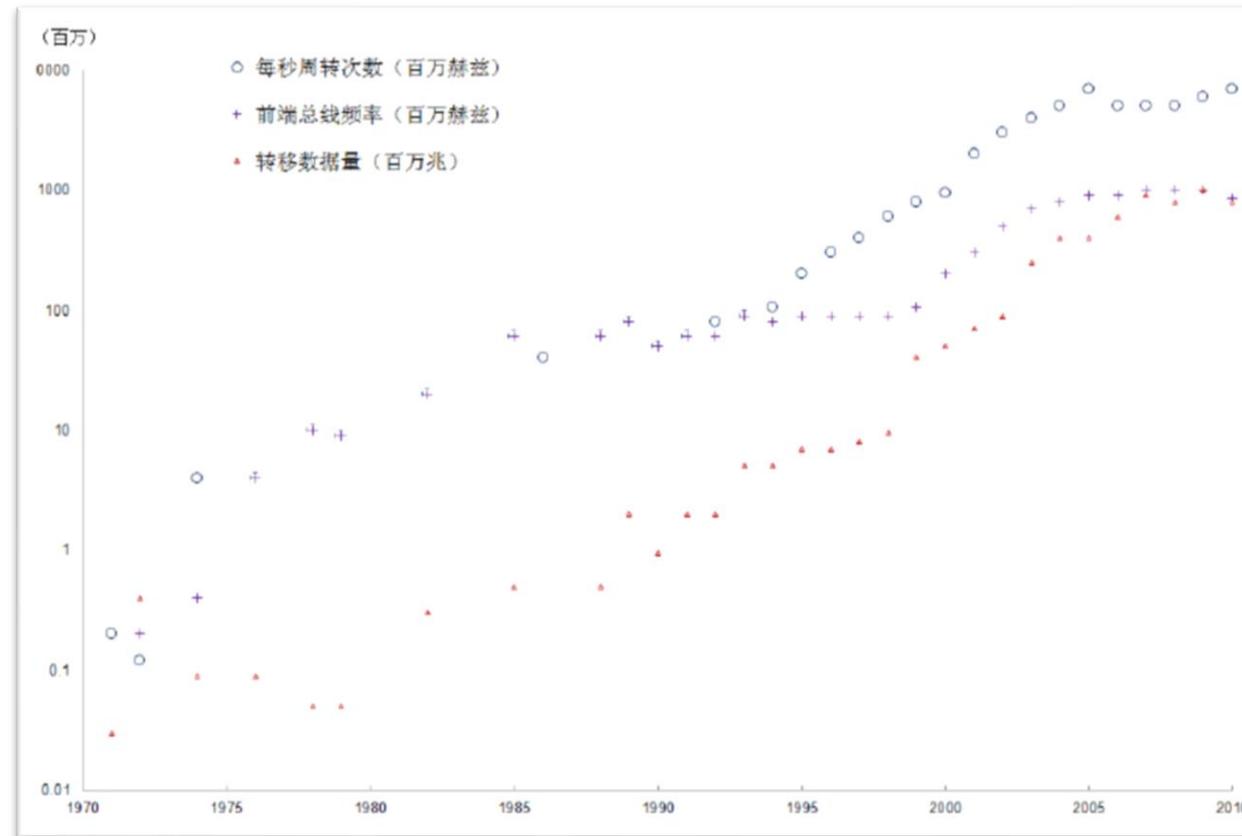


图1-4 网络带宽随时间变化情况

1.4 | 数据产生方式的变革



运营式系统阶段

用户原创内容阶段

感知式系统阶段

数据库的出现使得数据管理的复杂度大大降低，数据往往伴随着一定的运营活动而产生并记录在数据库中，数据的产生方式是被动的

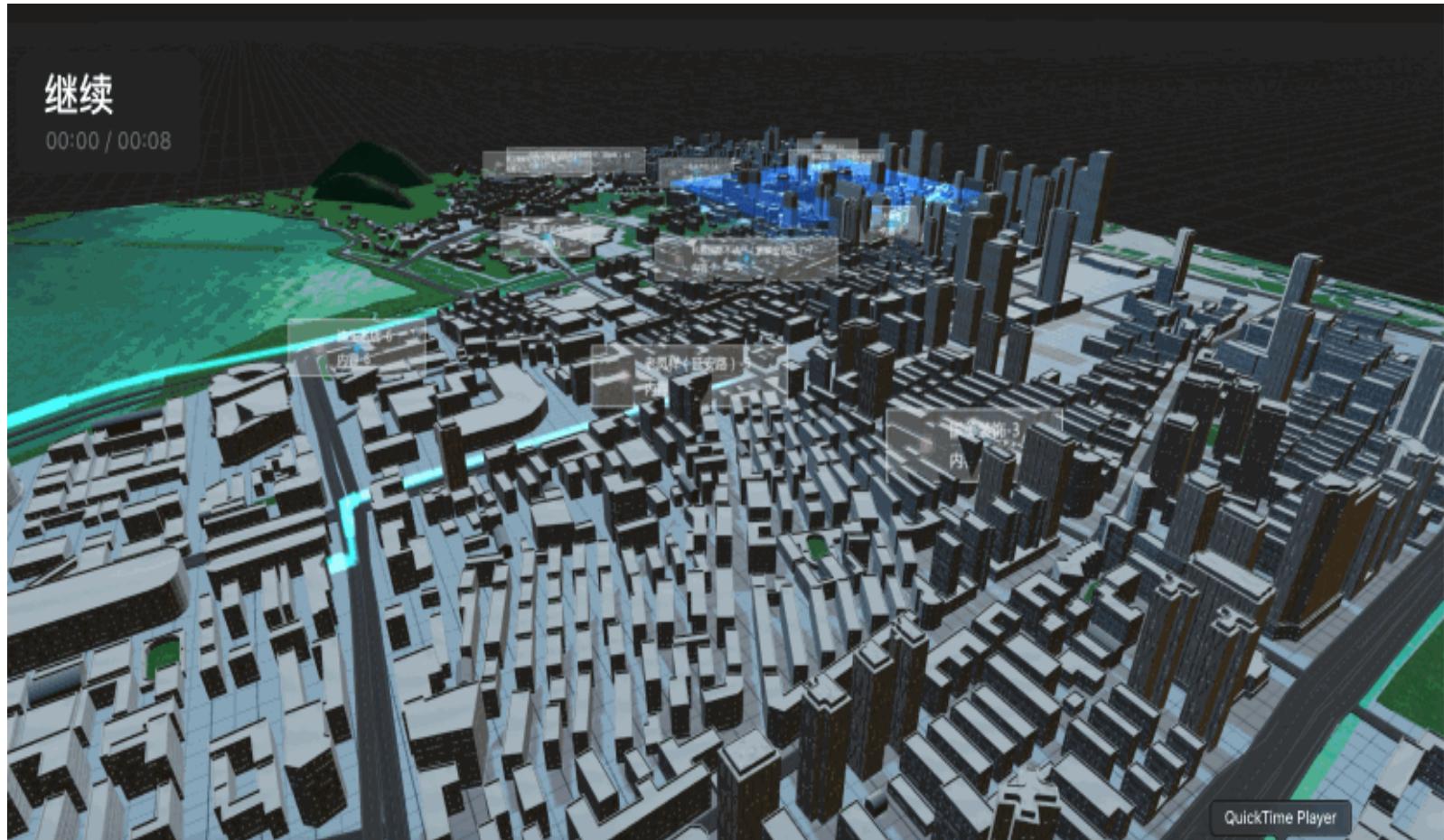
- 数据爆发产生于 Web 2.0 时代，而 Web 2.0 的最重要标志就是用户原创内容
- 智能手机等移动设备加速内容产生
- 数据产生方式是主动的

- 感知式系统的广泛使用
- 人类社会数据量第三次大的飞跃最终导致了大数据的产生

中央政府对大数据的重视程度



习近平	政府管理不仅要讲究策略，还要讲究手段，比如大数据技术的应用，2014年3月8日“大数据”首次写入政府工作报告
奥巴马	“将投入巨资拉动与大数据相关的产业” “数据为“未来的石油”，是美国综合国力的一部分，是与陆权、海权、空权同等重要的“国家核心资产”。
李克强	加快推进全国中小企业征信系统建设,通过大数据等技术优化中小企业征信资质。
李克强	经济数据和目标的进一步调整，中小企业将面临更大的压力，互联网金融除了解决便利性问题外，更重要的是如何围绕特有的大数据资源展开对实体经济的服务
汪洋	数据为王，财政工作离不开大数据

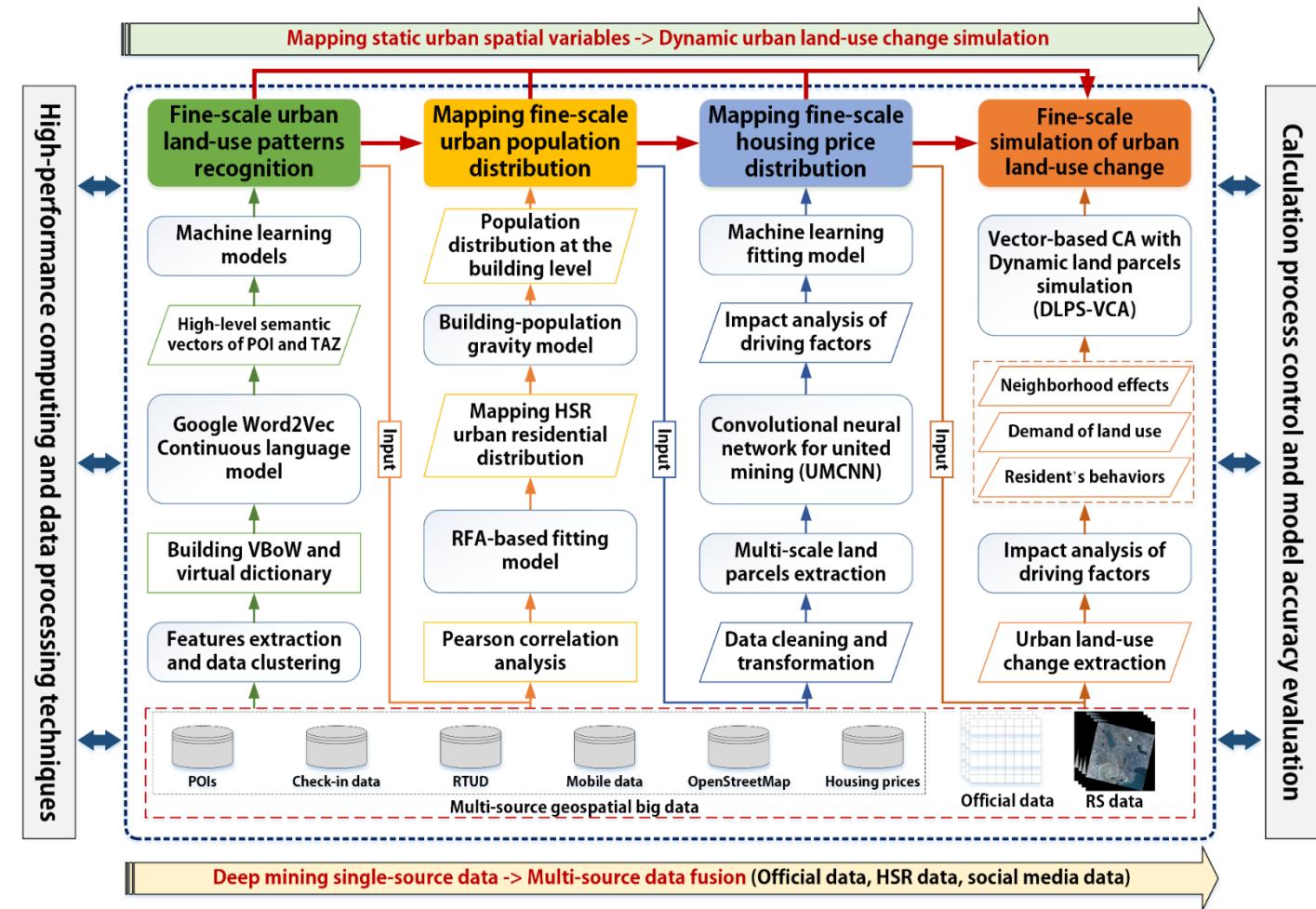
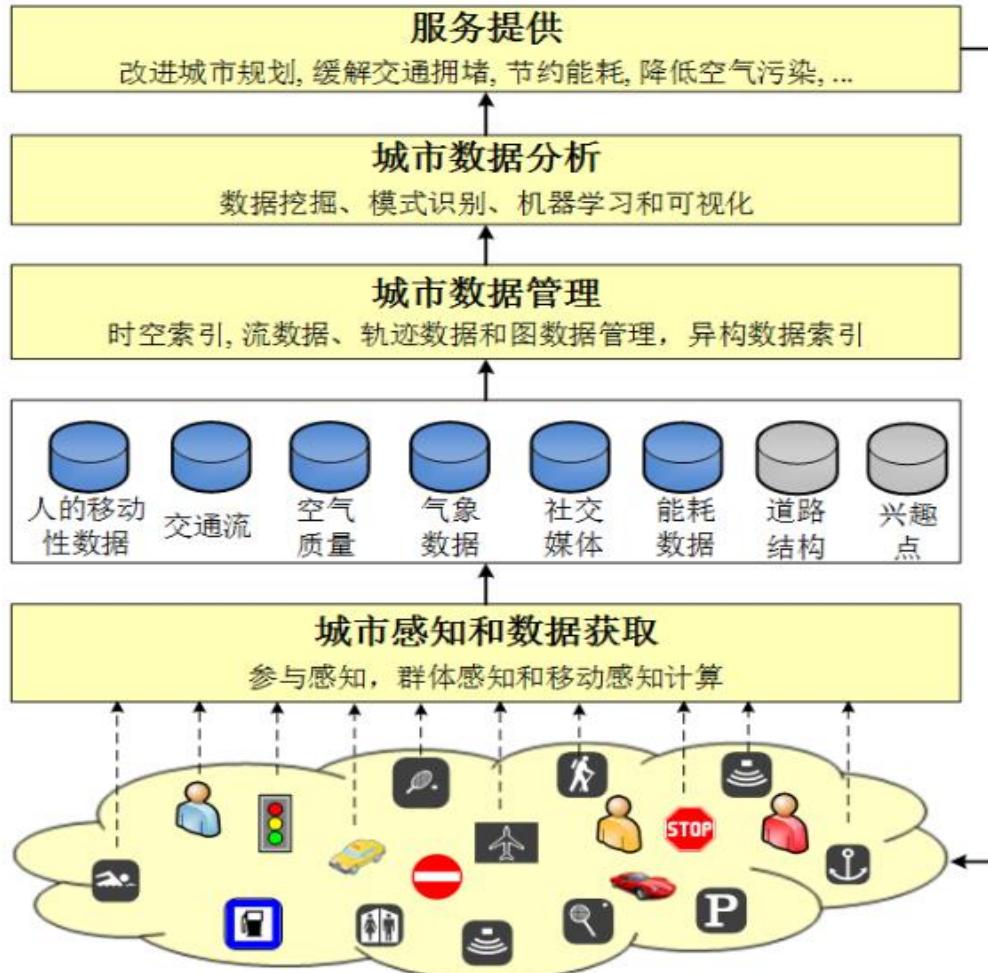


城市计算是一个交叉学科，是计算机科学以城市为背景，跟城市规划、交通、能源、环境、社会学和经济等学科融合的新兴领域。

城市计算是一个通过不断获取、整合和分析城市中多源异构的大数据来解决城市所面临的挑战的过程。

(郑宇, 2015)

城市计算的基本框架





➤ 城市感知（计算）

- 城市计算就是要用城市中的大数据来解决城市本身所面临的挑战，通过对多种异构数据的整合、分析和挖掘，来提取知识和智能，并用这些智能来创造“人 - 环境 - 城市”三赢的结果。
- 如何利用城市中现有的资源（如手机、传感器、车辆和人等），在不干扰人们生活的前提下，不断地自动感知城市的韵律，是一个重要的研究课题。

➤ 海量异构数据的管理

- 城市产生的数据属性差别很大，如何管理和整合大规模的异构数据将是一个新的挑战。
- 尤其是在一个应用中使用多种数据时，只有提前建立起不同数据之间的关联，才能让后面的分析和挖掘过程变得高效、可行。

➤ 异构数据的协同计算

- 如何从不同的数据源中获取相互增强的知识是一个新的课题。
- 在保证知识提取深度的同时，如何提高对大数据的分析效率，从而满足城市计算中众多实时性要求较高的应用（如空气质量预测、异常事件监测等），也是一个难题。
- 数据维度的增加也容易导致数据稀疏性问题，如何应对大数据的数据稀疏性问题，也很重要。

➤ 虚实结合的混合式系统

- 城市计算常常催生混合系统，比如云加端模式，即信息产生在物理世界，通过终端设备被收集到云端（虚拟世界）分析和处理，最后云再将提取的知识作为服务提供给物理世界的终端用户。



(李清泉, 2017)

- 个体时空行为数据涌现
 - 大规模 (百万、千万、亿级别)
 - 高质量 (精细的时空分辨率)
- GIS分析方法和工具提出要求
 - 大数据 (架构、平台)
 - 实时 (存储、查询、分析)
 - 从自然、地理现象到以人中心的研究
- 研究范式: 微观 → 宏观 (Bottom → Top)
- 三大价值: 科学价值、社会价值、商业价值

1.6 | 城市大数据的来源



➤ 手机信令

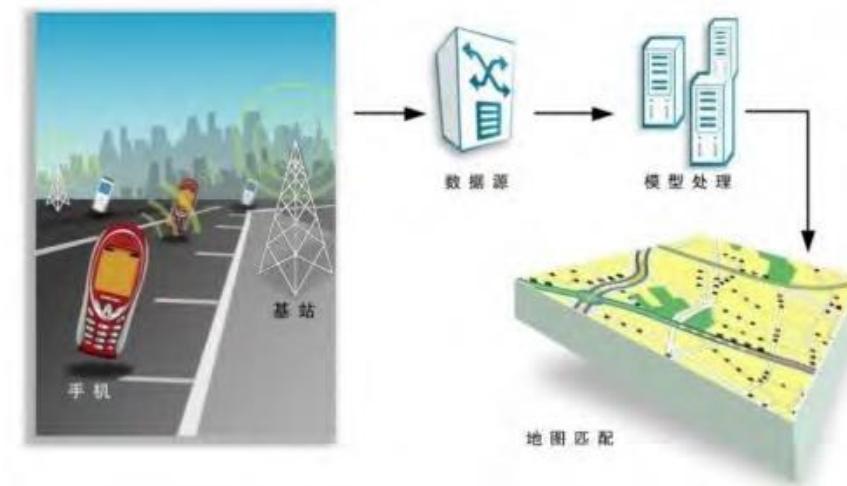
某运营商提供的数据说明

➤ 信令数据：2G，3G信令统计数据。
内容为每个基站每小时出现的信令数量



➤ 用户数据：2G，3G信令统计数据。
内容为每个基站每小时统计的用户数量

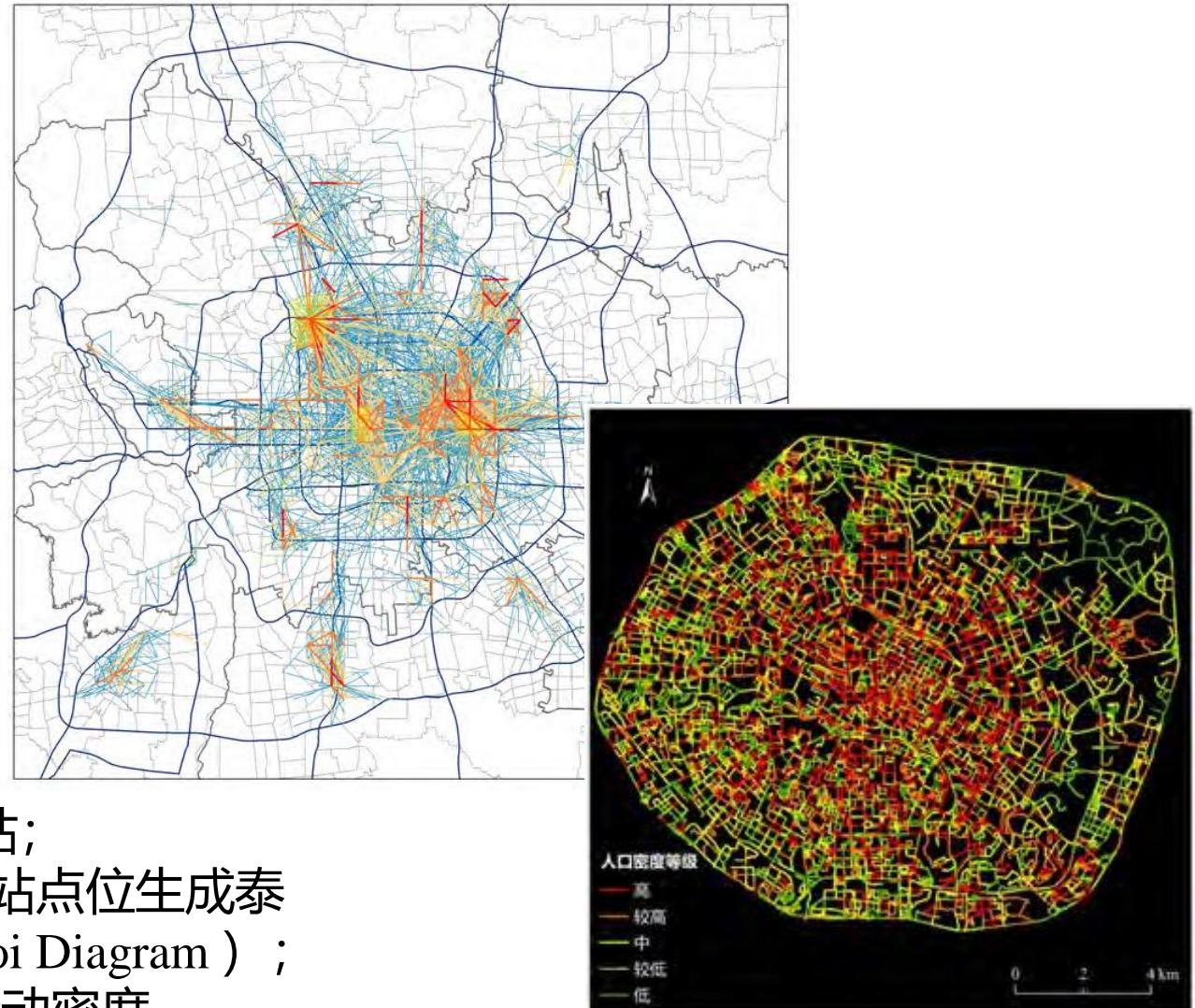
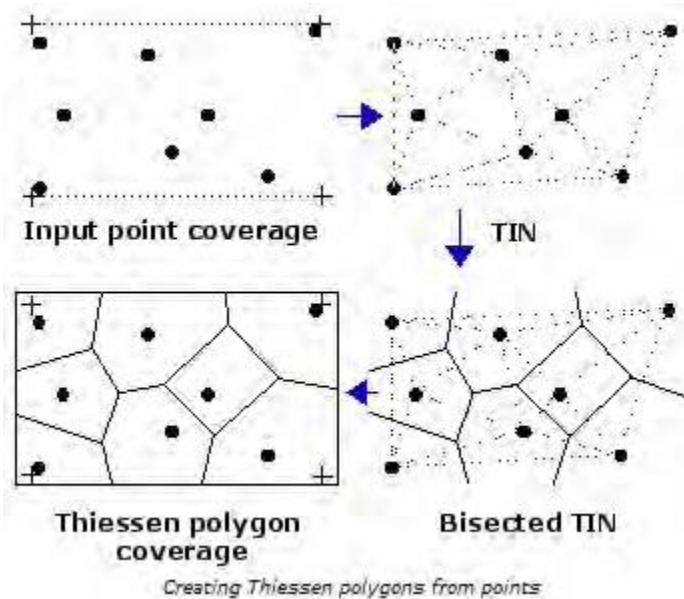
➤ 轨迹数据：2G，3G信令统计数据。
内容为上午用户的移动OD（基站之间移动
的用户数量）



1.6 | 城市大数据的来源



➤ 手机信令



手机信令信息空间匹配至相应小区基站；

- 合并位置完全相同的基站，基于基站点位生成泰森（Thiessen）多边形（亦为Voronoi Diagram）；
- 计算每个Thiessen多边形内的人类活动密度。

1.6 | 城市大数据的来源



▶ 公共交通刷卡记录

- 大量城市的公共交通系统采用智能卡作为交通收费手段 (AFC)
- 公共交通刷卡记录是AFC的副产品
- 空间分辨率为站点 metro station/bus stop, 时间分辨率精确到秒
- 轨道交通、分段计价公交、一票制公交
- 辅助数据：公交车GPS、居民家庭出行调查数据、线路、站点、交通分析小区 (TAZ)

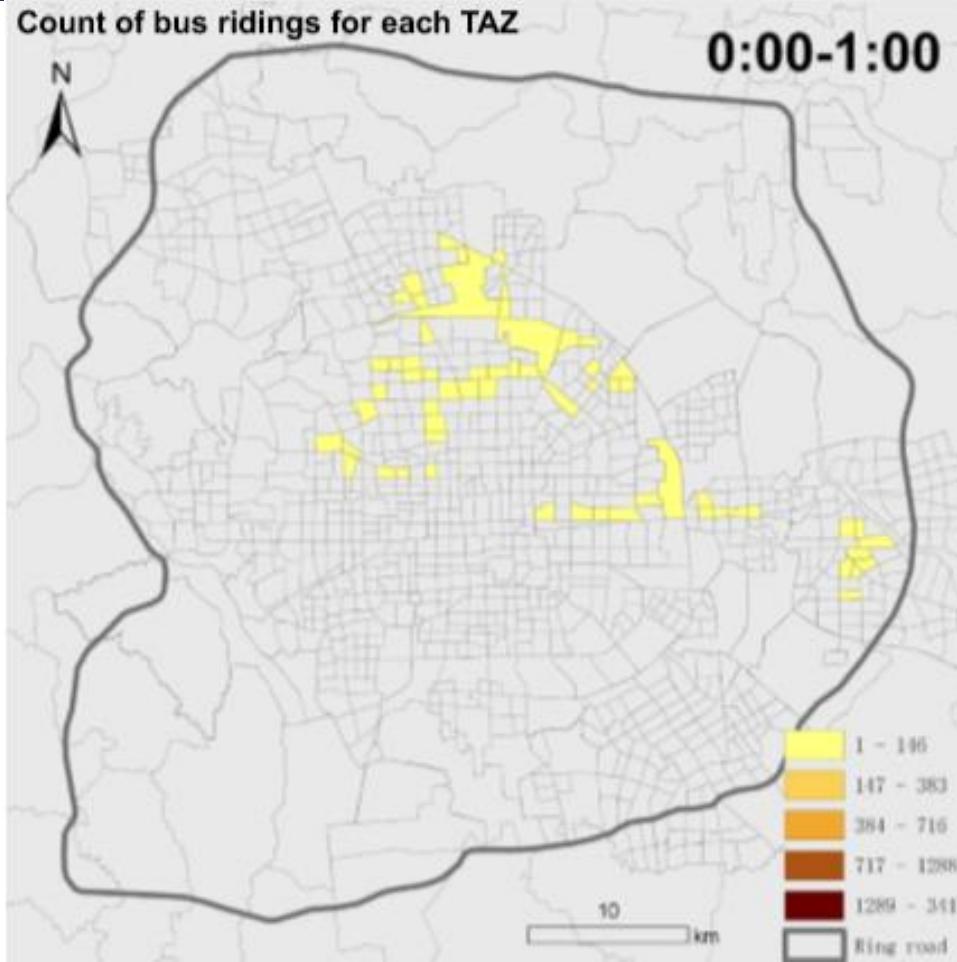
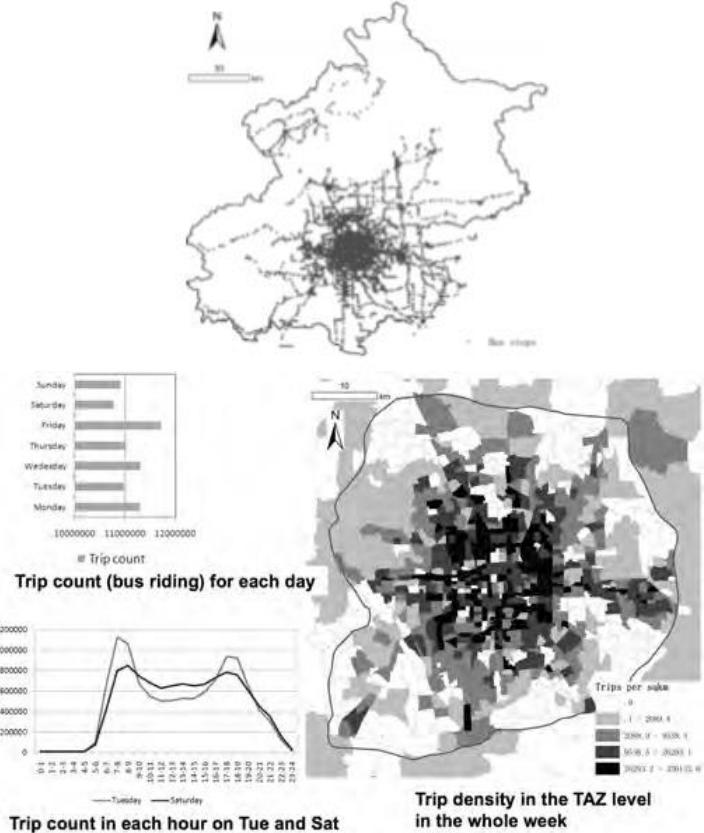
Variable	Exemplified Values
Card ID	“10007510038259911”, “10007510150830716”
Card Type	1, 2, 3, 4
Line ID	602, 40, 102
Line Type	0, 1
Driver ID	11032, 332
Vehicle ID	111223, 89763
Departure Date	2008-04-08
Departure Time	“06-22-30”, “11-12-09”
Departure Stop	11, 5, 14
Arrival Time	“09-52-05”, “19-07-20”
Arrival Stop	3, 14, 9



1.6 | 城市大数据的来源



➤ 公共交通刷卡记录

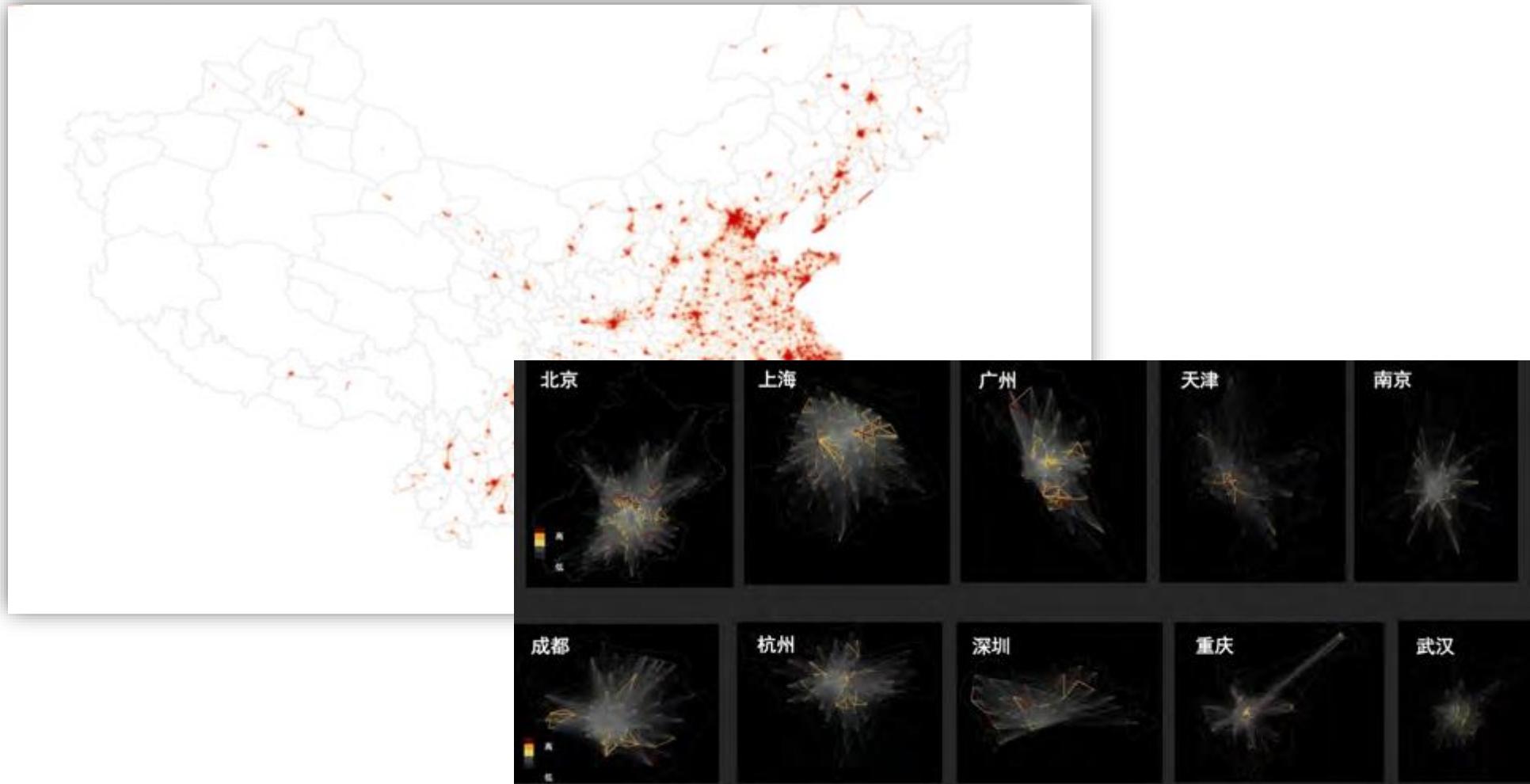


- 潜在应用领域：职住平衡、城市贫困、极端出行、乘客画像、线路调整、规划实施评价、群体出行、学生出行、灰色人群、城市功能识别（北京研究：上千万持卡人连续一周近亿次出行）
- 龙瀛, 孙立君, 陶遂. 基于公共交通智能卡数据的城市研究综述. 城市规划学刊, 2015, 3:70-77.

➤ 出租车/网约车轨迹



➤ 出租车/网约车轨迹



➤ 共享单车骑行轨迹



- 具有GPS模块的共享单车（如所有的摩拜和部分OFO）骑行过程中产生的定位轨迹数据、开关锁记录等
- <https://mobike.com>、<http://www.ofo.so>
- 以及其他共享单车公司推测的辅助数据如用户居住地、就业地、使用习惯、年龄阶段等，以及故障报告记录
- 2017年共享单车与城市发展白皮书：<https://zhuanlan.zhihu.com/p/26443639>

➤ 银联消费



- 银联智惠：<https://www.unionpaysmart.com>
- 上海主要商圈消费及客流数据解读（城市数据团与银联智惠研究院共同发布）：
<https://mp.weixin.qq.com/s/boc2mchWPO7jGhRuV0Tn2g>

➤ 智慧足迹



- 中国联通的智慧足迹：<http://www.smartsteps.com>
- 将手机信令加工成时空标签，反映人的空间活动，从而研其踪而知其人。通过匿名、聚合、外推的大数据能力，帮助政府精准服务、精确决策、精细分析，帮助企业挖掘潜客、选址营销、业务创新

➤ 百度慧眼



Baidu 地图 | 慧眼

首页 产品介绍 成功案例 行业报告 联系我们

百度慧眼

商业地理大数据服务专家

勾勒顾客画像，展现顾客轨迹；竞品分析对比，客流来源去向，助您广拉新客，精准营销，对目标区域进行全面位置评估。

地理数据查询

- 城市基础数据
- 商圈分布数据
- 实地全景
- 商业设施
- 品牌分布
- 交通分布

人口数据分析

- 居住人口
- 工作人口
- 人群画像
- 分时段客流分析
- 实时客流热力
- OD通勤分析

位置信息管理

- 点数据管理
- 面数据管理
- 个性化地图
- 渗透率分析

<http://huiyan.baidu.com>

➤ 阿里数据



<https://dt.alibaba.com>

➤ 腾讯大数据



<http://bigdata.qq.com>

➤ TalkingData



- 是中国最大的独立第三方移动数据服务平台，覆盖全国，收集大量移动终端的定位数据
(TD托管Android平台的大量APP)
- <http://www.talkingdata.com>

1.6 | 城市大数据的来源



➤ 公共设施（市政）大数据（水、电和天然气等）

The screenshot shows two windows side-by-side. The left window is titled '故障抢修工作台' (Fault Repair Workstation) from '江苏电力 - Microsoft Internet Explorer'. It displays a map of a city area with numerous red markers indicating specific locations. The map includes street names like '小王巷', '小板巷', '张府园', and '七家湾'. A legend on the left shows icons for water, electricity, and gas. A scale bar at the bottom indicates 100米 (100 meters) and 500 英尺 (500 feet). The right window shows a list of documents under '高级应用' (Advanced Application), with file names and sizes listed:

文件名	大小
通知, 最好明天下班	52.7K
发包人要求"这个模	616.0K
现场测试自动化技术	41.4K
	345.4K
上午12点以前发给我)	7.3M
	140.3K
意见, 进行了报奖书	17.6M
前将下月重点工作计	3.0K
本周四下班填妥附表	41.1K
	12.7M
请审阅!	96.6K
	1.3M
	453.1K
	570.5K
	55.7K

图片来源：东南大学杨俊宴团队

公用设施的智能计量表能够实时记录住宅、办公和商业场所的资源/能源消耗，所产生的数据除了支持收费外，还可以用于研究使用者的行为，评价建筑物的节能效率等工作。

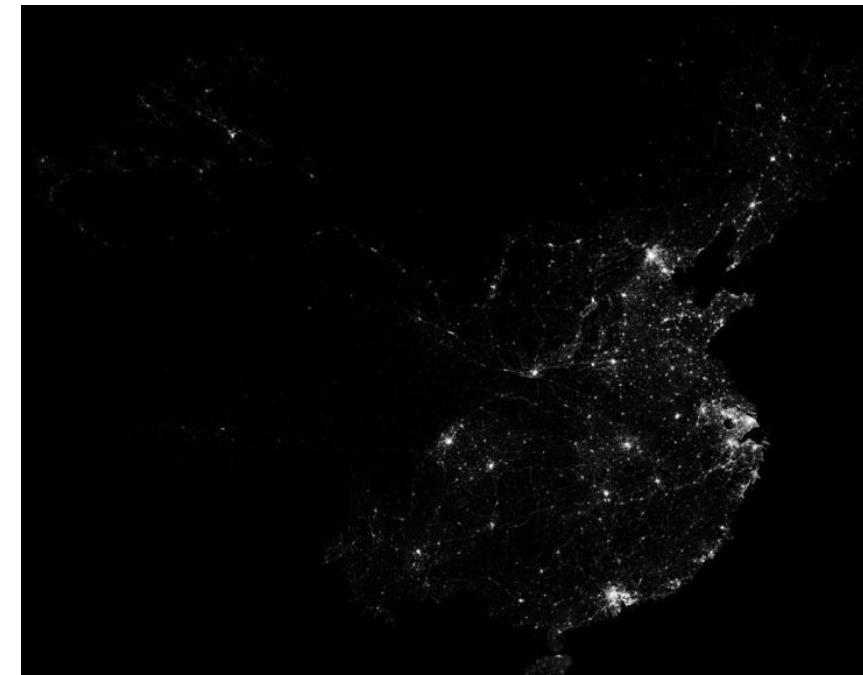
1.6 | 城市大数据的来源



➤ 兴趣点

兴趣点（英语：point of interest，通常缩写成POI）是电子地图上的某个地标、景点，用以标示出该地所代表的政府部门、各行各业之商业机构（加油站、百货公司、超市、餐厅、酒店、便利商店、医院等）、旅游景点（公园、公共厕所等）、古迹名胜、交通设施（各式车站、停车场、超速照相机、速限标示）等处所。

多个互联网公司均提供兴趣点获取的API（Application Programming Interface，应用程序编程接口）



➤ 大众点评

- 大众点评是中国领先的本地生活信息及交易平台，也是全球最早建立的独立第三方消费点评网站。大众点评不仅为用户提供商户信息、消费点评及消费优惠等信息服务，同时亦提供团购、餐厅预订、外卖及电子会员卡等O2O（OnlineToOffline）交易服务。
- <http://www.dianping.com>



➤ 美团



美团

搜索商家或地点

美团外卖 猫眼电影 美团酒店 民宿 / 公寓 商家入驻 美团公益

全部分类

- 美食
- 外卖
- 酒店 HOT
- 榛果民宿
- 猫眼电影
- 机票 / 火车票
- 休闲娱乐 / KTV
- 生活服务
- 丽人 / 美发 / 医学美容
- 结婚 / 婚纱摄影 / 婚宴
- 亲子 / 儿童乐园 / 幼教
- 运动健身 / 健身中心
- 家装 / 建材 / 家居
- 学习培训 / 音乐培训
- 医疗健康 / 宠物 / 爱车

住酒店 品质出游 特价酒店

住酒店 秒天天特价 享超值优惠

涨姿势 快来干掉无趣

我是商家 我想合作

Hi! 你好

注册 立即登录

美团APP手机版

1元起 吃喝玩乐

➤ www.meituan.com 饿了么 (<https://www.ele.me>) 则主要侧重餐饮外卖

➤ 百度搜索结果：地名共现（地名+地名）

利用百度新闻搜索sina.com.cn中任意两个中国省级行政区(一共31个，除了港澳台)名称共现的页面数目。之所以做这样的限定，是为了避免搜到过多的垃圾页面，并且新闻中的地名往往更可靠。这样得到465个共现页面数值，我们认为它反映了相应省份的联系强度



地名共现可简单归纳为以下四种情况：

- (1) 整体部分关系，如“海淀是北京的人口第二多的区”；
- (2) 空间相近或者相邻，如“山西位于陕西东边”；
- (3) 空间交互，如“从四川到广东的人口迁徙”；
- (4) 属于同一类别，如“广东和江苏都是中国经济发达的省份”。

- 基于地名共现网页数的相关性研究，
http://blog.sina.com.cn/s/blog_17288fcf30102xk5i.html

1.6 | 城市大数据的来源



➤ 12306

中国铁路12306
12306 CHINA RAILWAY

搜索车票/餐饮/常旅客/相关规章

我的12306 | 登录 注册

首页 车票 ▾ 团购服务 ▾ 会员服务 ▾ 站车服务 ▾ 商旅服务 ▾ 出行指南 ▾ 信息查询 ▾

车票 常用查询 订餐

单程 往返 接续换乘 退改签

出发地：简拼/全拼/汉字

到达地：简拼/全拼/汉字

出发日期：2020-01-14

学生 高铁/动车

查询

江西
风/景/独/好
中国铁路旅游

最新发布 联系客服 APP下载 旅客出行温馨提示

- 两个城市间的火车班次、余票等信息，可以用于支持研究城市网络（交通联系）
- www.12306.cn

1.6 | 城市大数据的来源



➤ 携程/去哪儿

The screenshot shows the homepage of Ctrip.com. At the top, there's a navigation bar with links for '让旅行更幸福' (Travel Happier), 'Language', '您好, 请登录' (Hello, Please Log In), '免费注册' (Free Registration), '消息' (Messages), '我的携程' (My Ctrip), '我的订单' (My Orders), '客服中心' (Customer Service), and a phone icon. Below the navigation is the Ctrip logo and a search bar with placeholder text '搜索旅行地/酒店/旅游/景点门票/交通'. To the right of the search bar are buttons for '境内: 95010' (Domestic: 95010) and '(或) 400-830-6666'. The main menu below the search bar includes categories like 首页 (Home), 酒店 (Hotels), 旅游 (Travel), 跟团游 (Group Tours), 自由行 (Independent Travel), 机票 (Air Tickets), 火车 (Trains), 汽车·船 (Cars·Boats), 用车 (Car Services), 门票 (Tickets), 攻略 (Travel Guide), 全球购 (Global Purchase), 礼品卡 (Gift Cards), 商旅 (Business Travel), 邮轮 (Cruises), 目的地 (Destinations), 金融 (Finance), and 更多 (More). A prominent red promotional banner on the right side of the page says '过全世界的年' (Celebrate the New Year Around the World) and '瓜分¥100,000,000新春福利' (Share ¥100,000,000 Spring Festival Welfare). On the left, there's a sidebar with links for 酒店 (Hotels), 机票 (Air Tickets), 旅游 (Travel), 跟团游 (Group Tours), 自由行 (Independent Travel), 火车 (Trains), and 用车 (Car Services). The bottom of the page features a 'HOT' section with links for '热门' (Hot), '特价旅游' (Special Price Travel), '出境游' (Overseas Travel), '境内游' (Domestic Travel), '周边游' (Local Travel), '邮轮' (Cruises), '门票' (Tickets), '当地玩乐' (Local Entertainment), '主题游' (Theme Travel), and '高端游' (High-end Travel). A '武汉出发' (Departing from Wuhan) button is also visible.

- www.ctrip.com, www.qunar.com
- 机票和汽车票同样也可以体现研究城市网络（交通联系）

➤ 淘宝

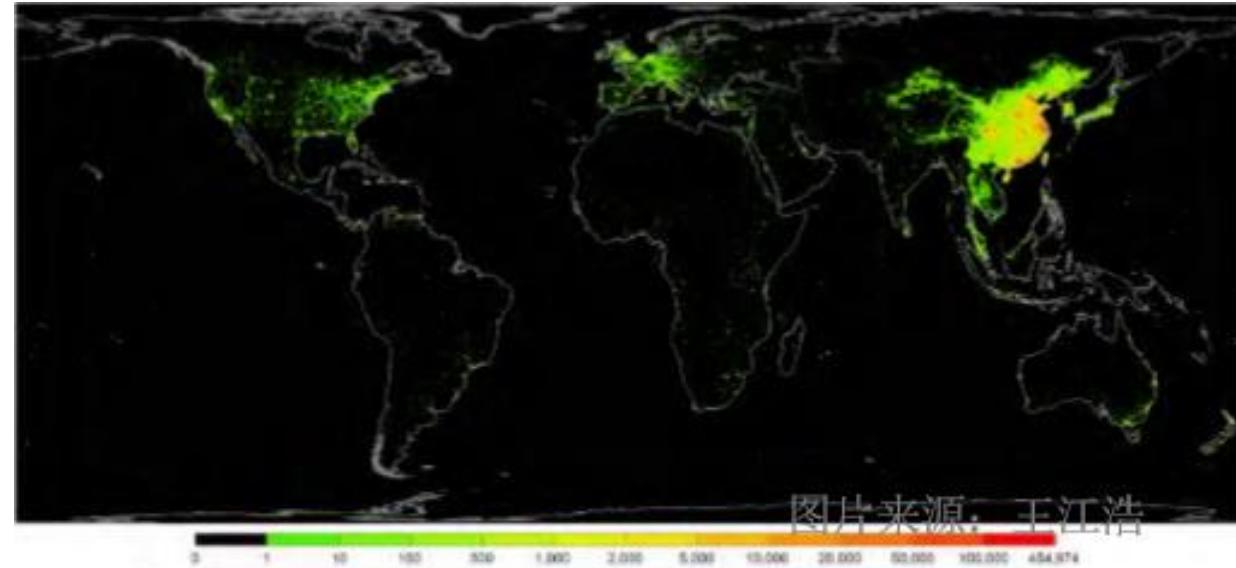


www.taobao.com

1.6 | 城市大数据的来源

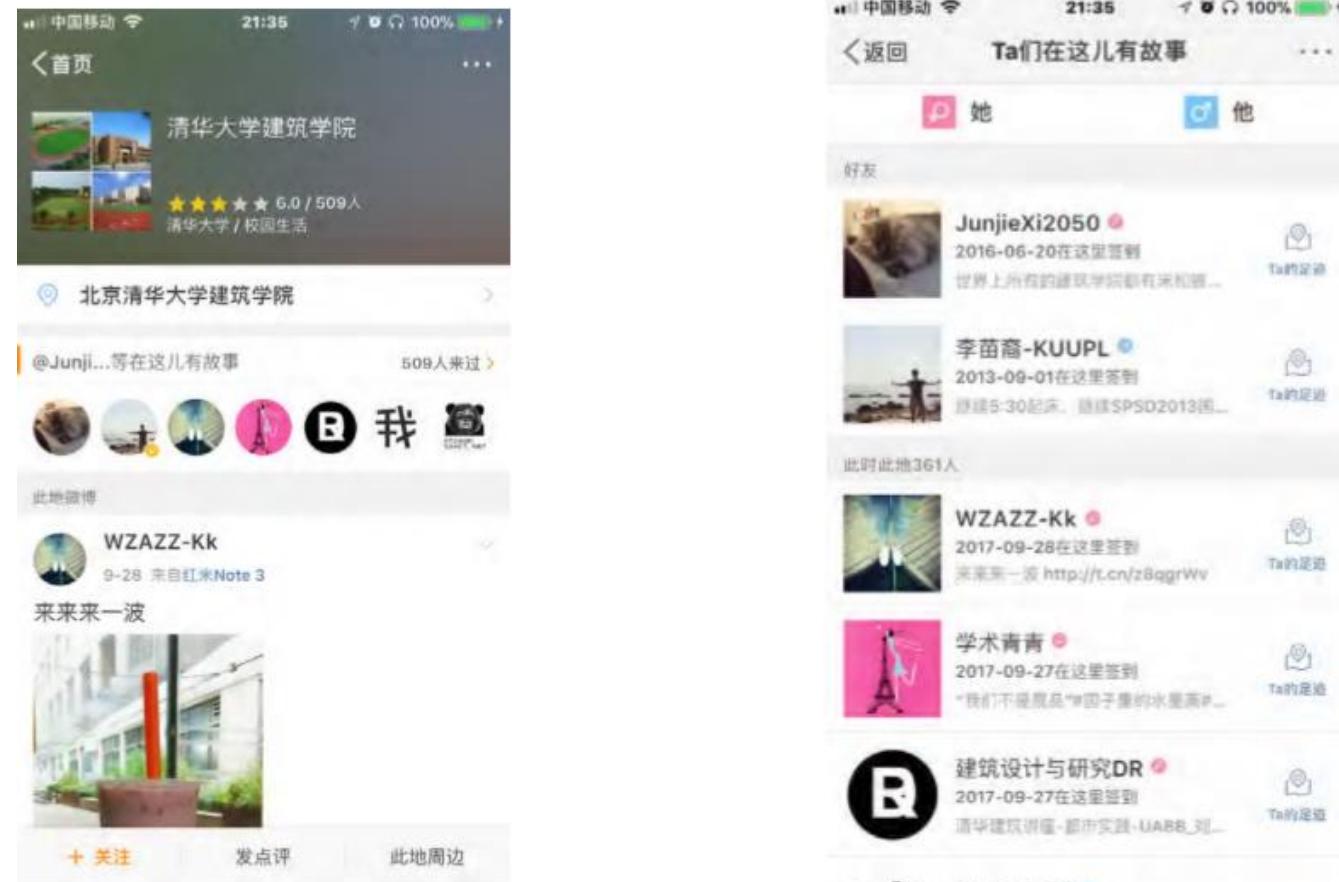


➤ (位置) 微博



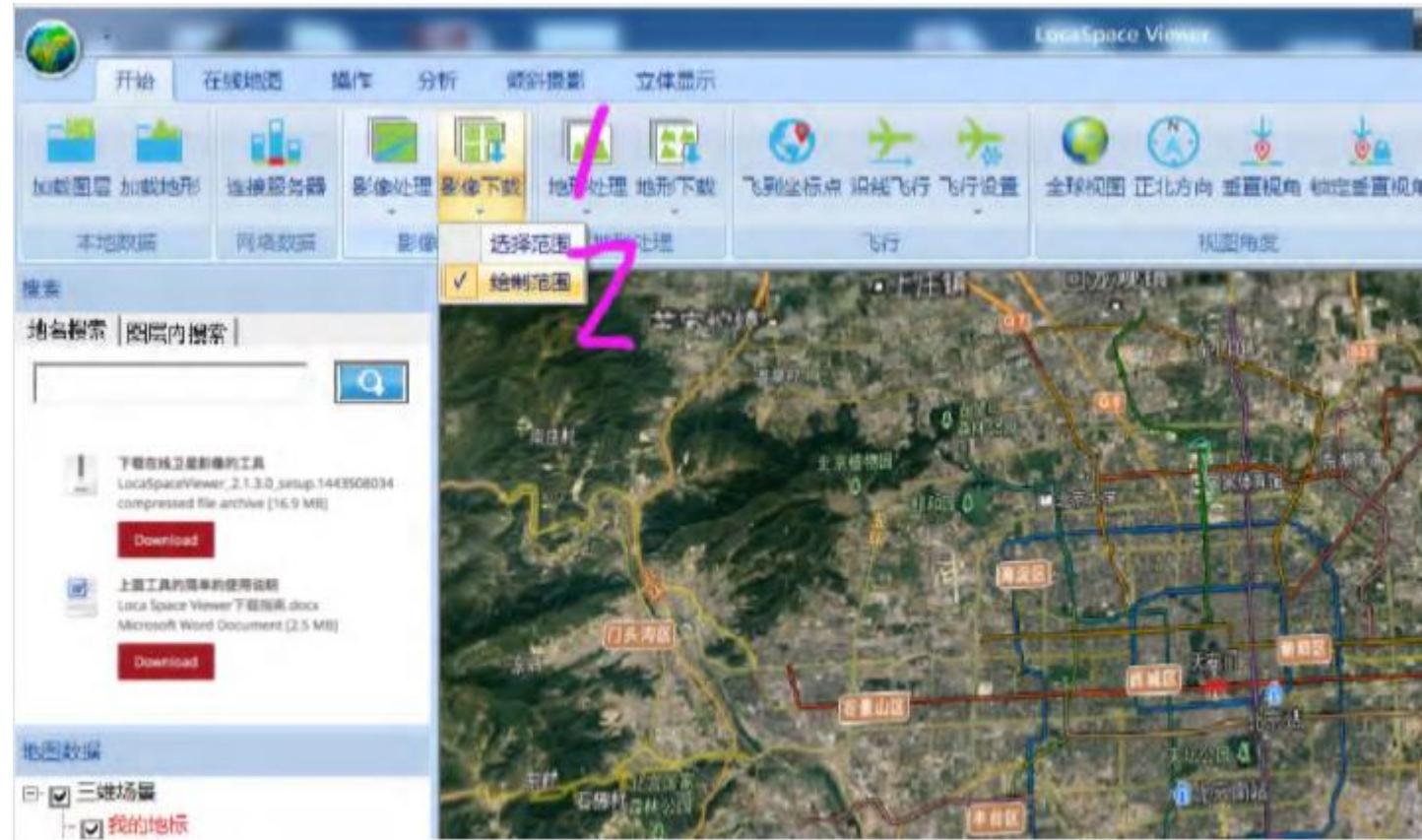
- www.weibo.com
- 约1%微博具有位置信息，微博抓取难度日益提高
- 微博数据处理，涉及空间分析、文本分析、图片分析等
- 新浪微博创立于2009年

➤ 微博签到



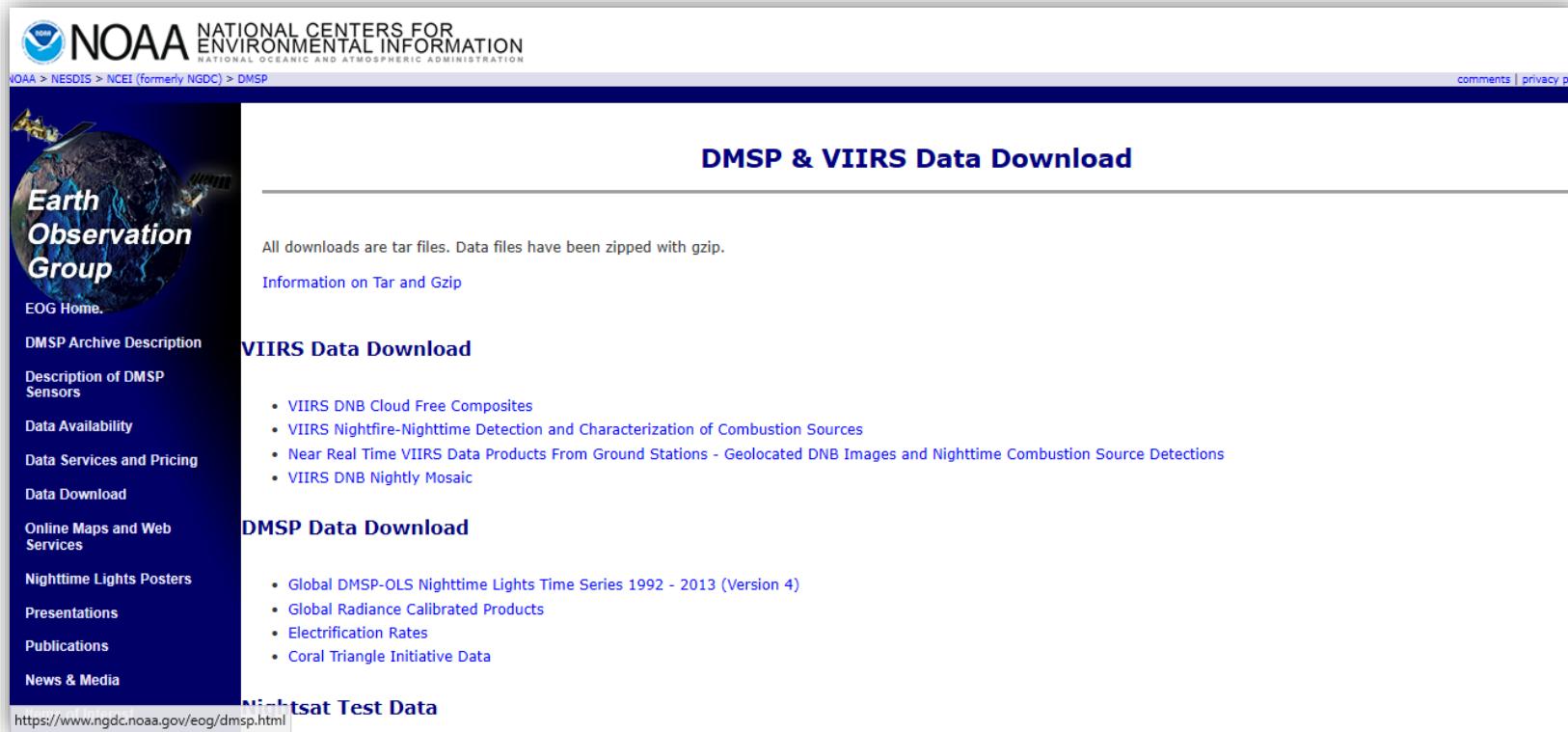
- www.weibo.com
- 体现一个地点受欢迎的程度（“人气”）
- 结合签到的用户，可以构建地点之间的联系网络、评价地点相似性、评价用户偏好

➤ 多年和最新的航片（谷歌地球）



<http://www.beijingcitylab.com/projects-1/22-urban-design-course/>

➤ 夜光影像



NOAA NATIONAL CENTERS FOR ENVIRONMENTAL INFORMATION
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION

comments | privacy policy

DMSP & VIIRS Data Download

All downloads are tar files. Data files have been zipped with gzip.

[Information on Tar and Gzip](#)

VIIRS Data Download

- VIIRS DNB Cloud Free Composites
- VIIRS Nightfire-Nighttime Detection and Characterization of Combustion Sources
- Near Real Time VIIRS Data Products From Ground Stations - Geolocated DNB Images and Nighttime Combustion Source Detections
- VIIRS DNB Nightly Mosaic

DMSP Data Download

- Global DMSP-OLS Nighttime Lights Time Series 1992 - 2013 (Version 4)
- Global Radiance Calibrated Products
- Electrification Rates
- Coral Triangle Initiative Data

tsat Test Data

<https://www.ngdc.noaa.gov/eog/dmsp.html>

- VIIRS 500m, 2012
- DMSP1km, 1992-2013
- <https://www.ngdc.noaa.gov/eog/download.html>

1.6 | 城市大数据的来源



➤ 道路 (OSM)



(10) Global cities Shapefile data

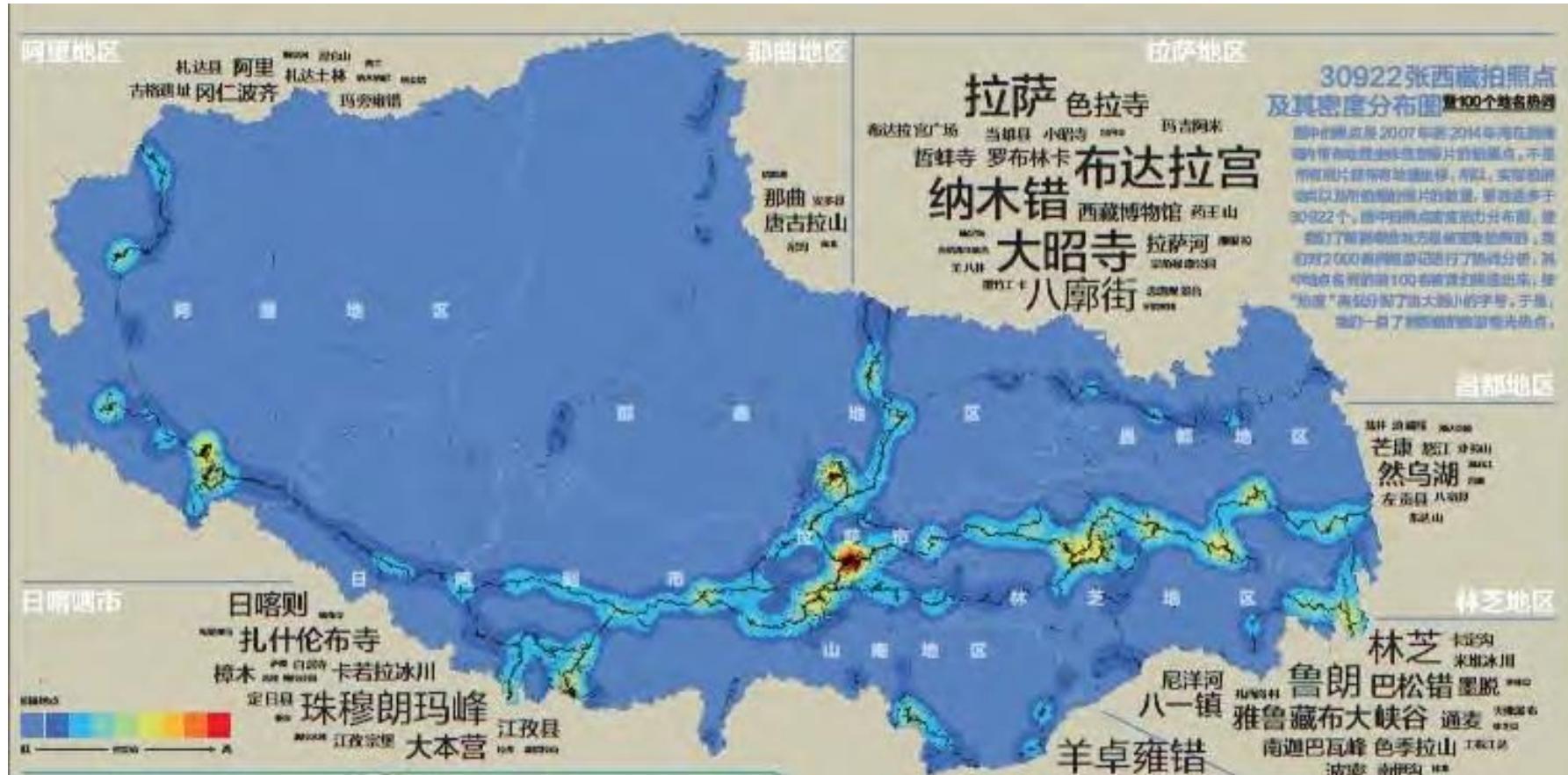
Generated from OSM

<http://download.bbbike.org/osm/bbbike/>

<https://mapzen.com/data/metro-extracts/>

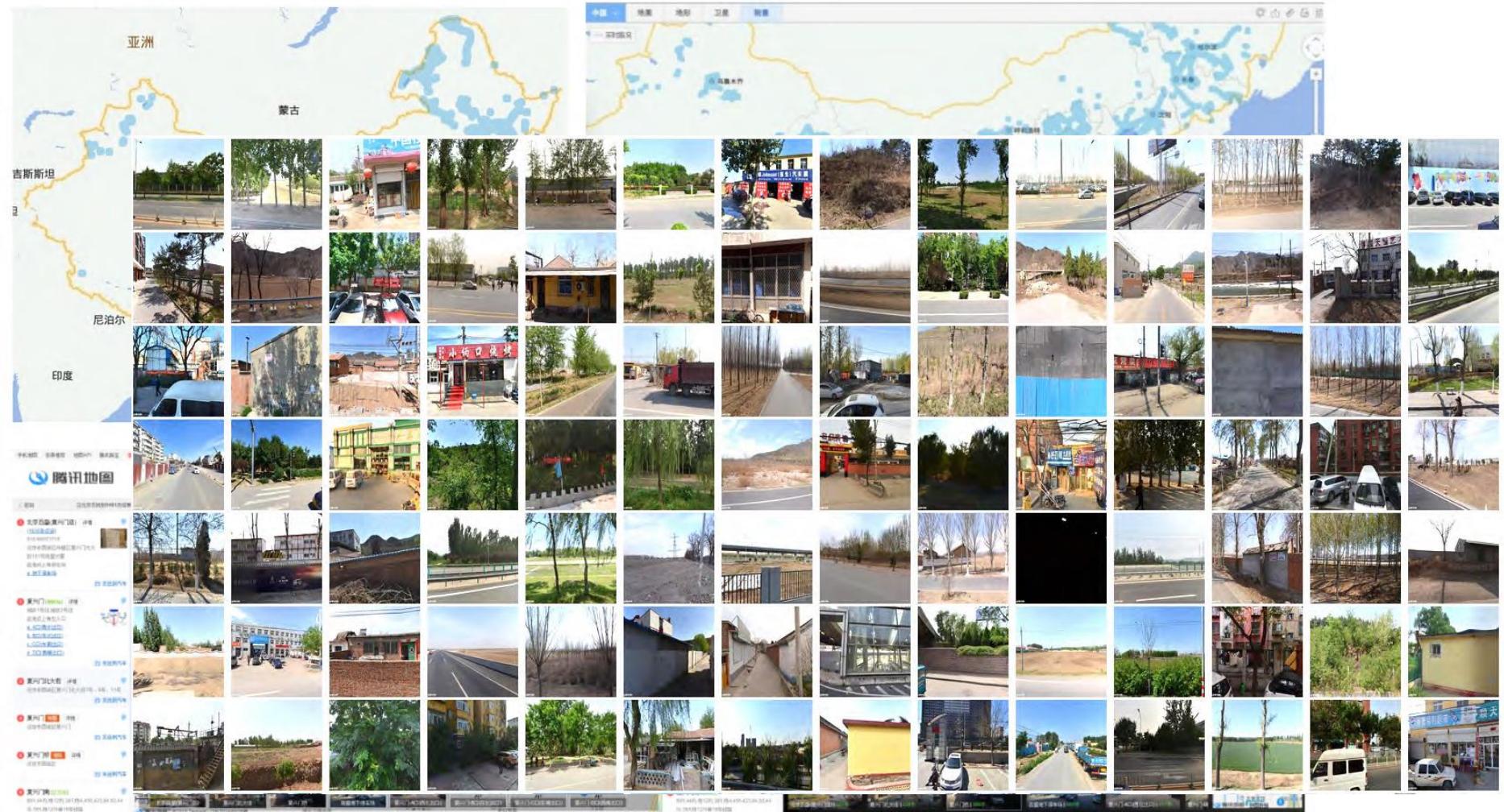
- 道路是重要的城市数据基础设施之一，体现了城市的交通组织，也是城市形态的重要表征。
- 在线地图平台（如高德、百度、四维导航等）多提供了覆盖整个中国的精细化的道路显示，OPENSTREETMAP也提供了开放下载。

➤ 位置照片 (Flickr照片)



- 基于Flickr照片的点位信息，识别主要旅游关注点，展示了西藏以点和线为主要形态的空间意象
- 2015年10月刊

➤ 街景图片

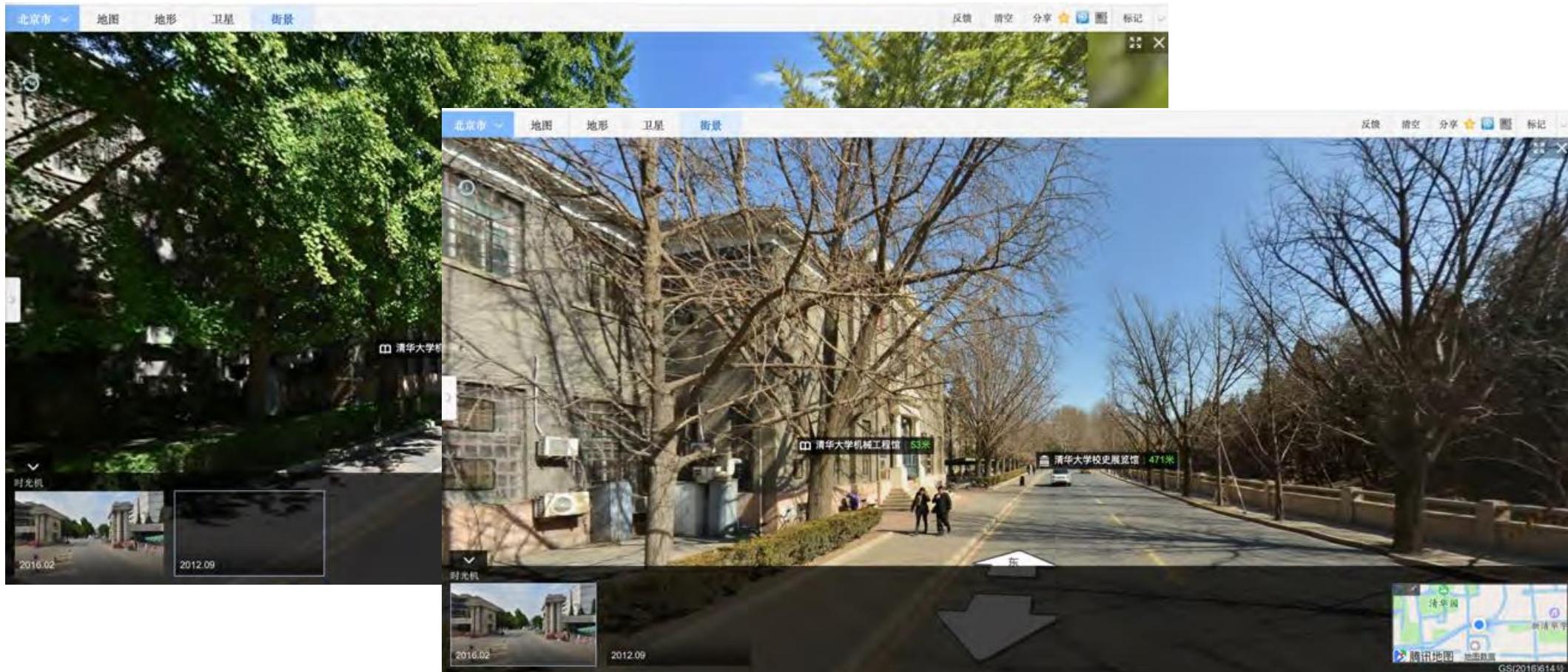


谷歌、百度、腾讯等在线地图服务

1.6 | 城市大数据的来源

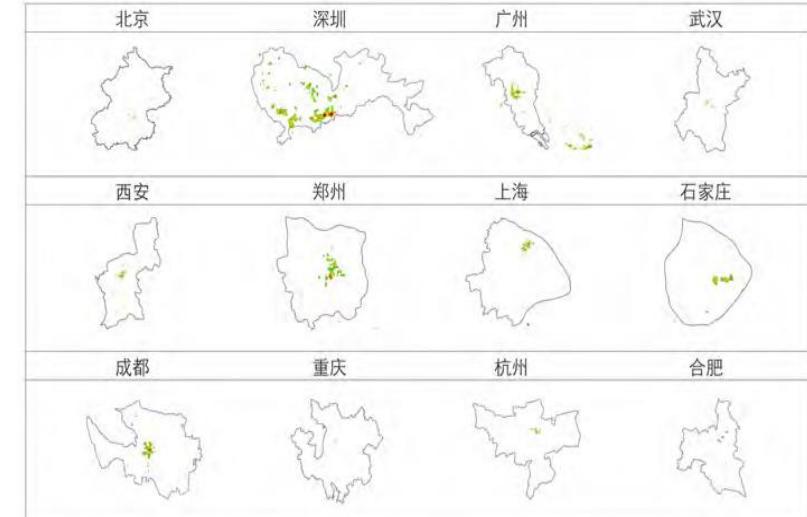
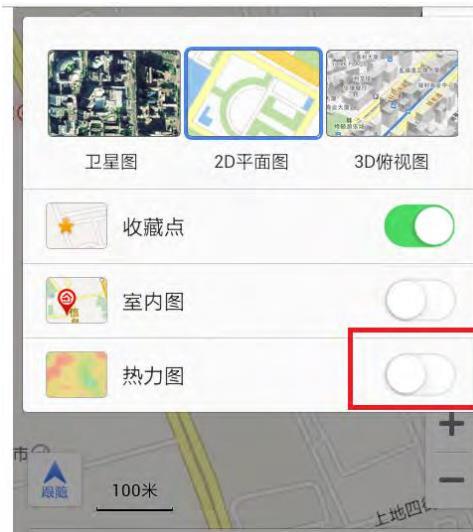


➤ 街景图片



腾讯街景提供时光机功能 (<http://map.qq.com>)

➤ 热力图



- 热力图以特殊高亮的形式显示访客热衷的页面区域和访客所在的地理区域总体被访问量在划定区域内占比情况示例图；
- 百度热力图利用通过百度系APP和获取的手机基站定位该区域的用户数量，通过用户数量渲染地图颜色（颜色与人口密度的对应关系目前是黑箱）；
- <https://baike.baidu.com/item/百度热力图/3098963>
- 只能通过百度地图APP访问，没有桌面版本。

1.6 | 城市大数据的来源



➤ 腾讯宜出行

The figure consists of two screenshots of a mobile application. The left screenshot shows a map of Tsinghua University with various buildings labeled. A legend at the bottom indicates population density levels: '稀疏' (Sparse) in green, '当前' (Current) in orange, and '拥挤' (Crowded) in red. Below the map is a line graph showing population density over time from 00:00 to 24:00, with a red dot marking the current time. The right screenshot shows the WeChat public account profile for '宜出行'. It includes sections for '功能介绍' (Function Introduction), '帐号主体' (Account Subject), '接收文章推送' (Receive Article Push), '置顶公众号' (Top Public Account), '提供位置信息' (Provide Location Information), and '查看历史消息' (View History Messages). A large green button at the bottom says '进入公众号' (Enter Public Account).

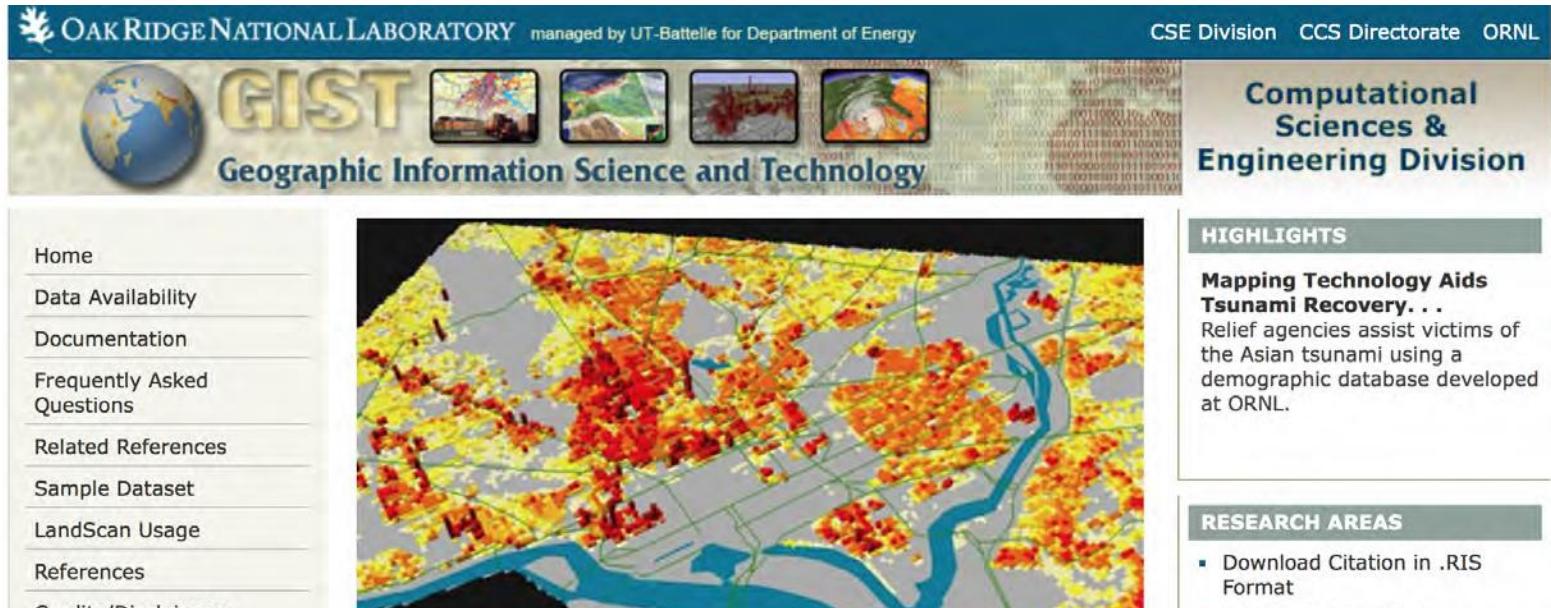
- 关注宜出行微信公众号（通过移动端，没有桌面版本）
- 获得所在位置和感兴趣区域的人口密度情况
- 查询区域可以比较微观也可以非常宏观（宏观则类似百度热力图）
- 可以查询当前状态，也可以追溯历史，预测短期未来
- 体现了城市活动这一维度

1.6 | 城市大数据的来源



➤ LANDSCAN

- 全球最高精度的人口分布数据，由美国OakRidgeNationalLaboratory建立，<http://web.ornl.gov/sci/landscan/index.shtml>
- 数据收费使用



LandScan™

Using an innovative approach with Geographic Information System and Remote Sensing, ORNL's LandScan™ is the community standard for global population distribution. At approximately 1 km resolution (30" X 30"), LandScan is the finest resolution global population distribution data available and represents an **ambient population** (average over 24 hours). The LandScan algorithm, an R&D 100 Award Winner, uses spatial data and imagery analysis technologies and a multi-variable dasymetric modeling approach to disaggregate census counts within an administrative boundary. Since no single population distribution model can account for the differences in spatial data availability, quality, scale, and accuracy as well as the differences in cultural settlement practices, LandScan population distribution models are tailored to match the data conditions and geographical nature of each individual country and region.

Please see the [Data Availability](#) page for access to the latest available LandScan™ dataset.

HIGHLIGHTS

Mapping Technology Aids Tsunami Recovery... .
Relief agencies assist victims of the Asian tsunami using a demographic database developed at ORNL.

RESEARCH AREAS

- Download Citation in .RIS Format
- Geographic Information Science and Technology (GIST)

DOWNLOADS

- Download Citation in .RIS Format

LS2012 | LS2011 | LS2010 |
LS2009 | LS2008 | LS2007 |
LS2006 | LS2005 | LS2004 |
LS2003 | LS2002 | LS2001 |
LS2000 | LS1998

➤ 三维建筑物



- 建筑物是城市物质空间最为基本和重要的构成要素，对城市规划与设计起到核心的支持作用
- 来自高德地图 (www.gaode.com) 和OSM地图 (www.openstreetmap.org) 等
- OSM地图上中国城市的建筑物质量一般 (非常少的建筑物)

1.5 | 城市大数据类型总结



- 遥感影像数据（传统的地理空间数据）
- 城市基础设施数据（能源、水、电、气）
- 社交媒体数据（签到、兴趣点、消费、微博等）
- GPS轨迹数据（人、车、智能设备）
- 手机信令数据（智能设备、手机定位、基站定位）
- 医疗卫生数据（健康app、物联网设备）

- 我们如何获取这些数据？
- 这些数据一定是全样本的吗？



数据获取的方式

- 网站API服务
- 网络爬虫
- 关于爬虫的其他
- 实例和实验

1.7 | 城市大数据的获取



➤ 获取Internet数据源方式

网站API服务 (Application Programming Interface)



1.7 | 城市大数据的获取



➤ 获取Internet数据源方式

网络爬虫—搜索引擎的第一步

自动的发现并下载网页称为爬取，下载网页的程序称为网络爬虫。



解决Github Pages禁止百度爬虫的方法与可行性分析

06月10日 关键词： Githubpages 百度爬虫 CDN加速

我在知乎提了这样一个问题：如何解决百度爬虫无法爬取搭建在Github上的个人博客的问题？，并且 Stackoverflow 上也有类似的问题：github blocks Baidu..

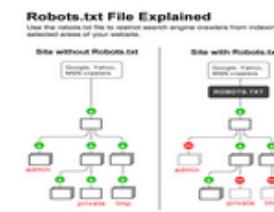
百度爬虫



网站抓取：如何正确识别Baiduspider移动ua？

05月22日 关键词： 网站抓取 网站优化 百度爬虫

近日，百度站长平台发布公告宣布新版Baiduspider移动ua上线，同时公布了PC版Baiduspider ua，那么该如何正确识别移动ua呢？对此，百度站长平台技术专..



搜索引擎的Robots规则

09月03日 关键词： 搜索引擎 Robots规则 百度爬虫 robots.txt设置

robots.txt是一种存放于网站根目录下的文本文件，用于告诉搜索引擎的爬虫（spider），此网站中的哪些内容是不应被搜索引擎的索引，哪些是可以被索引。..

网络爬虫—搜索引擎的第一步

自动的发现并下载网页称为爬取，下载网页的程序称为网络爬虫。

聚合阅读：

关于“谷歌爬虫”的最新资讯内容

谷歌爬虫



教你利用Google爬虫DDoS任意网站

06月30日 关键词： Google爬虫 谷歌爬虫 谷歌搜索 DDOS攻击

教你利用Google爬虫DDoS任意网站！Google的FeedFetcher爬虫会将spreadsheet的
=image("link")中的任意链接缓存。



Google 爬虫如何抓取 Javascript 的？

06月30日 关键词： 谷歌爬虫 网站优化 Javascript抓取 谷歌搜索引擎优化

认为 Google 不能处理 JavaScript ? 再想想吧。Audette Audette 分享了一系列测试结果，
他和他同事测试了什么类型的 JavaScript 功能会被 Google 抓取..

■ 网站API服务

百度搜索链

百度为您找到相关结果约100,000,000个

<https://www.baidu.com/s?ie=utf-8&wd=深圳>

网页 新闻 贴吧 知道 音乐 图片 视频 地图 文库 更多»

提供深圳市的地图浏览、地点搜索等多项服务。

查找地点：
如输入“深圳机场”

查询

map.baidu.com ▾

[深圳市地图_百度地图](#)

[深圳_百度百科](#)

深圳，别称鹏城，广东省辖市，地处广东省南部，珠江三角洲东岸，与香港一水之隔，东临大亚湾和大鹏湾，西濒珠江口和伶仃洋，南隔深圳河与香港相连，北部与东莞、惠州接壤。深圳是中国改革开放建立的第一个经济特区，是中国改革开放的窗口，已发展为有一定影响力的国际化城市，创造了举世瞩目的“深圳速...”

历史沿革 行政区划 地理环境 人口民族 交通设施 更多>>

baike.baidu.com/ ▾

人口民族

深圳第一新闻门户网站—深圳新闻网

深圳新闻网是立足深圳、辐射全国的综合性区域门户网站,为用户提供新闻、视频、博客、房产、汽车、财经、健康、美食、旅游、教育、时尚、娱乐、交友等20多个频道,并...

www.sznews.com/ ▾

V2 - 百度快照 - 75%好评

百度为您找到相关结果约100,000,000个

<https://www.baidu.com/s?ie=utf-8&wd=广州>

网页 新闻 贴吧 知道 音乐 图片 视频 地图 文库 更多»

提供广州市的地图浏览、地点搜索等多项服务。

查找地点：
如输入“天河体育中心”

查询

map.baidu.com ▾

[广州市地图_百度地图](#)

[广州_百度百科](#)

广州是广东省省会，是国务院定位的国际大都市、国际商贸中心、国际综合交通枢纽、国家中心城市、国家综合性门户城市、国家历史文化名城。广州从3世纪30年代起成为海上丝绸之路的主港，唐宋时期成为中国第一大港，明初、清初成为中国唯一的...

历史沿革 行政区划 地理环境 自然资源 交通 人口 更多>>

查看“广州”全部10个含义>>

baike.baidu.com/ ▾

简介：铺天盖地的粤语、大街小巷的鲜花、美味的粥品靓汤，让现代化的广州，仍保留着市井气息，漫步街头，各色靓仔靓女定会...

最佳旅游季节：10月-12月 广州旅游注意事项>>



■ 网站API服务

实例——人人网

人人网开放平台 2.0

接口分类	接口名	描述
location	/v2/location/feed/list	通过经纬度获取新鲜事。
	/v2/location/get	根据经纬度定位地点。
album	/v2/album/list	以分页的方式获取某个用户的相册列表
	/v2/album/get	获取某个用户的某个相册
	/v2/album/put	创建一个相册
blog	/v2/blog/list	以分页的方式获取某个用户的日志列表
	/v2/blog/put	创建一篇日志
	/v2/blog/get	获取某个用户的某篇日志
vipinfo	/v2/vipinfo/get	获取某个用户的VIP信息
evaluation	/v2/evaluation/reply/put	回复点评
	/v2/evaluation/reply/list	签到回复列表
	/v2/evaluation/put	用户发表点评
share	/v2/share/ugc/put	分享人人网内部UGC资源，例如：日志、照片、相册、分享(基于已有分享再次进行分享)
	/v2/share/hot/list	获取人人推荐资源
	/v2/share/url/put	分享人人网外部资源，例如：视频、图片等 如果要分享一张本地照片到人人网（即上传），建议使用 /v2/photo/upload 接口
	/v2/share/get	获取某个用户的某个分享
	/v2/share/list	以分页的方式获取某个用户的分享列表

<http://open.renren.com/wiki/API2>



■ 网站API服务

实例 | 大众点评


[首页](#)
[应用接入](#)

下面将通过一段简单的java示例来解释如何调用大众点评开发者的API:

Step 1: 获取App Key和App Secret

请登录开发者平台（尚未注册？[请点击这里](#)），进入“管理中心”，确认自己的App Key和App Secret。

[复制](#) [打印](#)

```
01. String appKey = "5589931241";
02. String secret = "db16adf193f2448ba0ec0260e0c968f3";
03. //请替换为自己的 App Key 和 App secret
```

Step 2: 确认请求参数

查看API文档，确认请求参数。

[复制](#) [打印](#)

```
04. String apiUrl = "http://api.dianping.com/v1/business/find_businesses";
05. paramMap.put("city", "上海");
06. paramMap.put("latitude", "31.21524");
07. paramMap.put("longitude", "121.420033");
08. paramMap.put("category", "美食");
09. paramMap.put("region", "长宁区");
10. paramMap.put("limit", "20");
11. paramMap.put("radius", "2000");
12. paramMap.put("offset_type", "0");
13. paramMap.put("has_coupon", "1");
14. paramMap.put("has_deal", "1");
```



■ 网站API服务

概述 • Web服务API

实例—百度POI

Web服务API

百度地图Web服务API为开发者提供http接口，即通过HTTP协议向百度地图发送请求并接收响应。用户可以基于此开发JavaScript、C#、C++、Java等语言的客户端。该套API免费对外开放，使用前请先申请密钥（key）。每个IP地址每天有10万次的访问限制，其中Place API每分钟对应的访问限制为2000次/天；Direction API每分钟对应的访问限制为2000次/天；Geocoding API每分钟对应的访问限制为10万次/天。如果有更高配额需求，请联系客服。在您使用百度地图Web服务 API之前，请先阅读[《百度地图Web服务 API 使用须知》](#)。

功能介绍



Place API

支持城市、矩形及圆形区域关键字搜索POI，返回json/xml格式的POI数据。



Geocoding API

通过地址获取坐标值或通过坐标点获取详细地址信息描述服务。



Route Matrix API

提供同时查询多个起终点线路信息的数据接口。



坐标转换API NEW

该接口可实现将常用的非百度坐标转换成百度地图中使用的坐标。



Place suggestion API

提供匹配用户输入关键字的辅助信息、提示接口、返回json/xml格式的建议词条数据。

1.7 | 城市大数据的获取



■ 网站API服务

Place API • Web服务API

实例 | 百度POI

什么是Place API？

Place API 是一类简单的HTTP接口，用于返回查询某个区域的某类POI数据，且提供单个POI的详情查询服务，用户可以使用C#、C++、Java等开发语言发送HTTP请求且接收json、xml的数据。

功能介绍

Place API 提供区域检索POI服务与POI详情服务。

1. 区域检索POI服务提供三种区域检索方法：

- 城市内检索（对应JavaScript API的Search方法）
- 矩形检索（对应JavaScript API的SearchInBound方法）
- 圆形区域检索（对应JavaScript的SearchNearBy方法）。

2. POI详情服务提供查询单个POI的详情信息，如好评。

■ 网站API服务

实例——百度POI

Place检索示例：

城市内检索 ▼

参数	值
query:	银行
scope:	1
page_size:	10
page_num:	0
region:	北京

Place区域检索通用接口参数

以下参数，适用于三种区域检索方法的Place API。

参数	是否必须	默认值	格式举例	含义
q(query)	是	无	中关村、ATM、百度大厦	检索关键字，周边检索和矩形区域内检索支持多个关键字并集检索，不同关键字间以\$符号分隔，最多支持10个关键字检索。如：“银行\$酒店”。
tag	否	无	日式烧烤/铁板烧、朝外大街	标签项，与q组合进行检索，以，“ 分隔
scope	是	1	1、2	检索结果详细程度。取值为1 或空，则返回基本信息；取值为2，返回检索POI详细信息

运行 (结果显示如下)

```
http://api.map.baidu.com/place/v2/search?ak=您的密钥&output=json&query=%E9%93%B6%E8%A1%8C&page_size=10&page_num=0&scope=1&region=%E5%8C%97%E4%BA%AC
```

1.7 | 城市大数据的获取



■ 网站API服务

百度POI返回值格式: (1) json格式 (2)xml格式

json格式:

```
{  
    status: 0,  
    result: {  
        location: {  
            lng: 116.30814954222,  
            lat: 40.056885091681  
        },  
        precise: 1,  
        confidence: 80,  
        level: "商务大厦"  
    }  
}
```

json示例:

<http://api.map.baidu.com/geocoder/v2/?ak=E4805d16520de693a3fe707cdc962045&callback=renderReverse&location=39.983424,116.322987&output=json&pois=1>

1.7 | 城市大数据的获取



■ 网站API服务

百度POI返回值格式: (1) json格式 (2)xml格式

xml格式:

```
<GeocoderSearchResponse>
    <status>0</status>
    <result>
        <location>
            <lat>40.056885091681</lat>
            <lng>116.30814954222</lng>
        </location>
        <precise>1</precise>
        <confidence>80</confidence>
        <level>商务大厦</level>
    </result>
</GeocoderSearchResponse>
```

xml示例:

<http://api.map.baidu.com/geocoder/v2/?ak=E4805d16520de693a3fe707cdc962045&callback=renderReverse&location=39.983424,116.322987&output=xml&pois=1>

1.7 | 城市大数据的获取



实例 | 百度POI

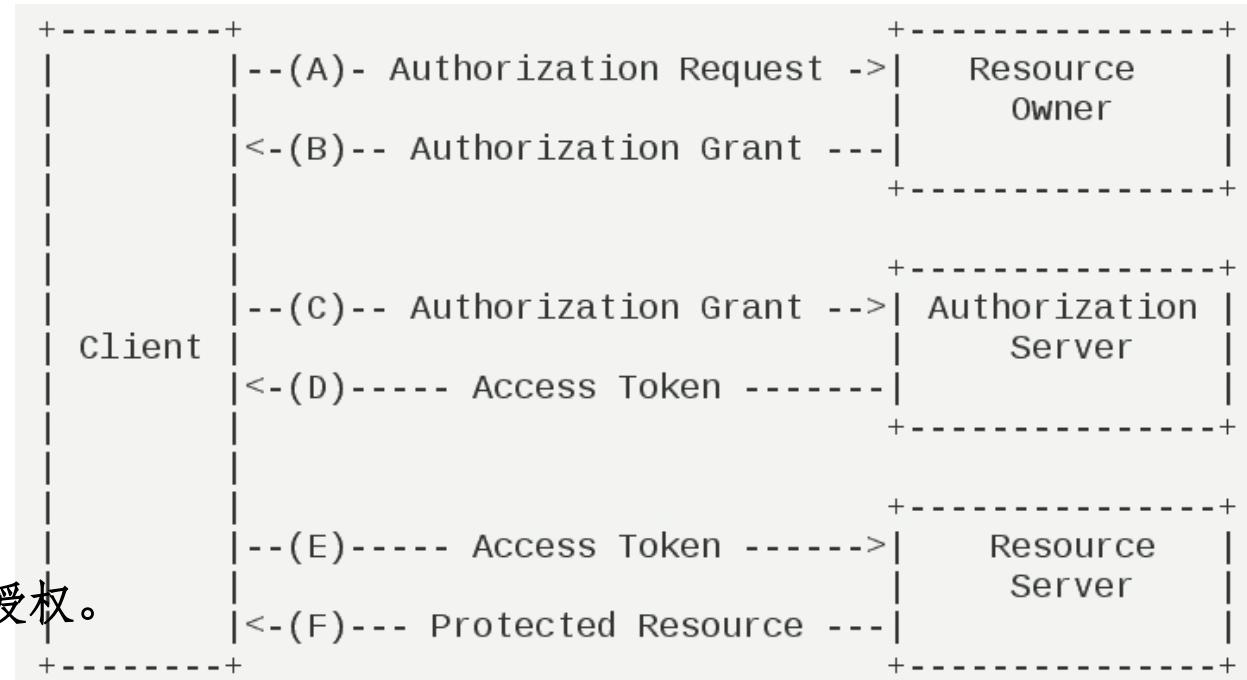
■ 网站API服务

利用Place API爬取广州市学校POI点的Python运行实例

```
Python 2.7.2 (default, Jun 12 2011, 15:08:59) [MSC v.1500 32 bit (Intel)] on win32
Type "copyright", "credits" or "license()" for more information.
>>> ===== RESTART =====
>>>
请输入地区 : 广州
请输入要搜索的关键词 : 学校
http://api.map.baidu.com/place/v2/search?q=学校&region=广州&output=xml&ak=FpwMlWanOB1ESMeu
AcuVlqlh&scope=2&page_size=20&page_num=0
共搜索到 400 个结果
分 20 页
扫描第 0 页
-----本页 20 个结果 -----
暨南大学(广州石牌校区) , 23.131822 , 113.354383 , 广东省广州市天河区黄埔大道西 601号 , (020)85220
114 , 教育培训;高等院校 , 4.8
华南农业大学 , 23.161023 , 113.359105 , 天河区五山街五山路483号 华南农业大学三角市 , 020-85280021
, 教育培训;高等院校 , 4.1
华南理工大学 , 23.1577 , 113.351566 , 广州市天河区五山路381号 , (020)87110000 , 教育培训;高等院
校 , 3.8
广东财经大学 , 23.09275 , 113.363914 , 广州市海珠区仑头路21号 , (020)84096140 , 教育培训;高等院
校 , 5.0
广州大学 , 23.043873 , 113.3778 , 广州市番禺区大学城外环西路230号(广州大学行政东楼后座504) , 020
-39366998 , 教育培训;高等院校 , 4.0
广东工业大学大学城校区 , 23.039841 , 113.404262 , 广州市大学城外环西路100号广工教学1-2号楼 , 020
-39322320 , 教育培训;高等院校 , 2.0
广州体育学院 , 23.149966 , 113.325316 , 广东省广州市天河区广州大道中1268号 , 020-87551717 , 教
育培训;高等院校 , 3.7
广东外语外贸大学(白云大道南辅路) , 23.205913 , 113.295795 , 白云大道北2号 , (020)83308484 , 教
育培训;高等院校 , 2.2
南方医科大学 , 23.192684 , 113.340775 , 广东省广州市白云区沙太北路 , 020-61640114 , 教育培训;高
等院校 , 2.1
```

■ 网站API服务

网页API的调用方式： I 直接调用， 基于OAuth2.0的调用



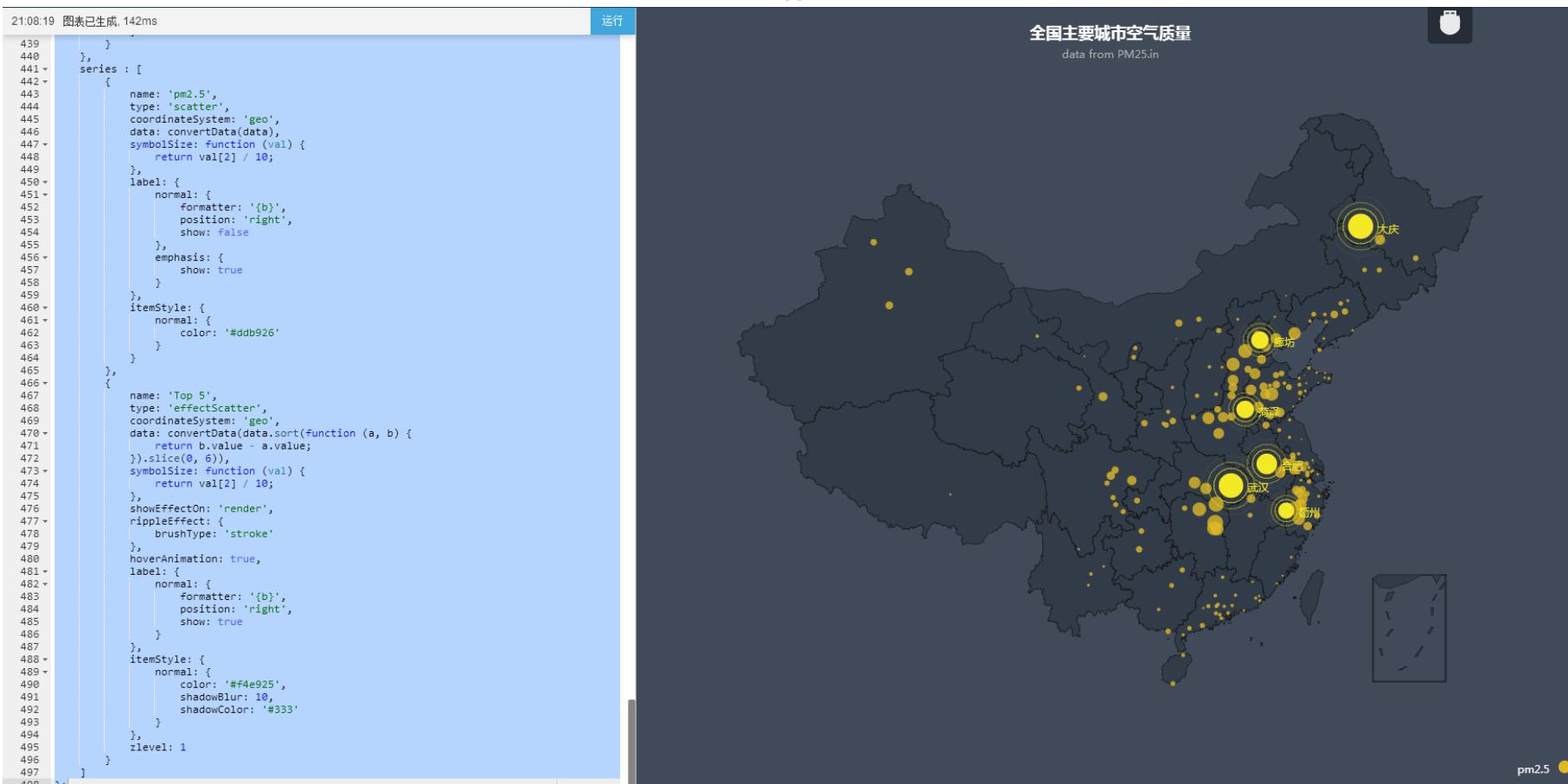
- (A) 用户打开客户端之后，客户端要求用户给予授权。
- (B) 用户同意给予客户端授权。
- (C) 客户端使用上一步获得的授权，向认证服务器申请令牌。
- (D) 认证服务器客户端进行认证之后，确认无误，统一发放令牌
- (E) 客户端使用令牌，向资源服务器申请获取资源
- (F) 资源服务器确认令牌无误，同意向客户端开放资源

1.7 | 城市大数据的获取



■ 网站API服务

网页API的调用方式：Ⅱ 基于JavaScript的调用



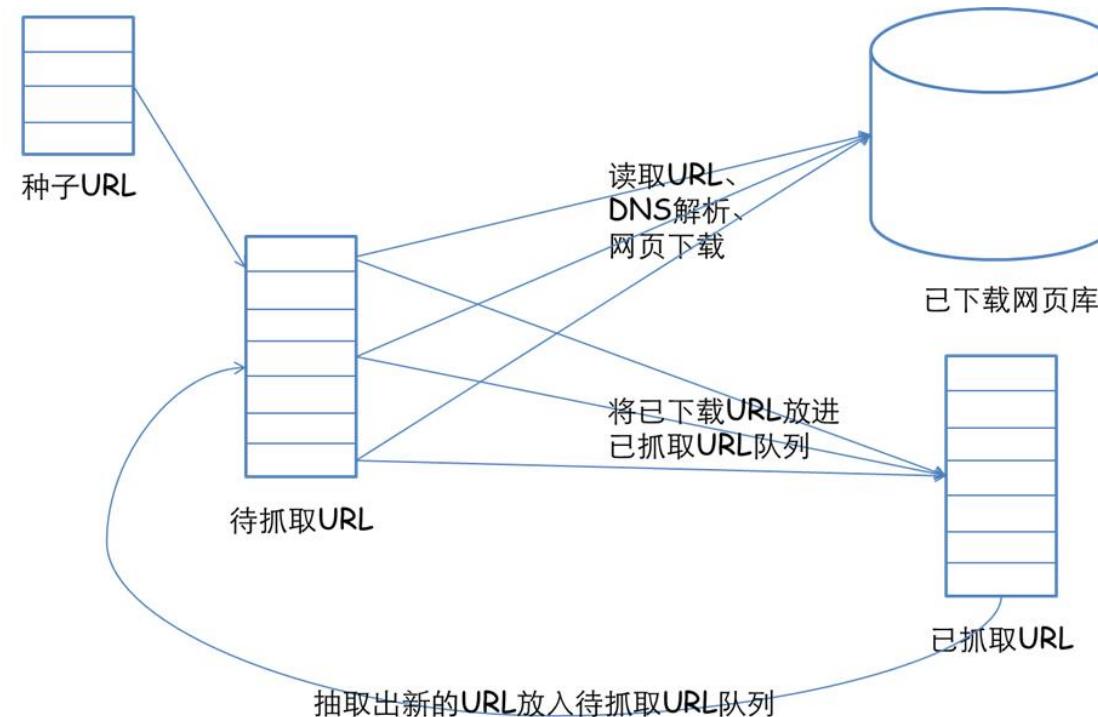
<http://echarts.baidu.com/demo.html#effectScatter-map>

1.7 | 城市大数据的获取



■ 网络爬虫

<https://www.91ri.org/11469.html>



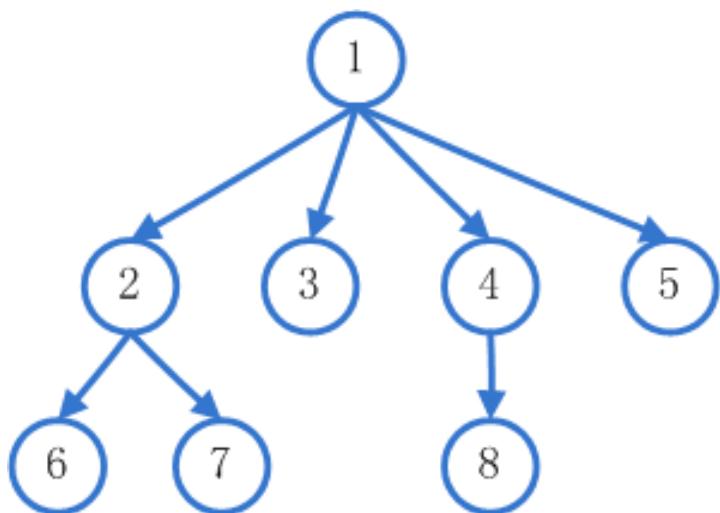
- 1.首先选取一部分精心挑选的种子URL；
- 2.将这些URL放入待抓取URL队列；
- 3.从待抓取URL队列中取出待抓取在URL，解析DNS，并且得到主机的ip，并将URL对应的网页下载下来，存储进已下载网页库中。此外，将这些URL放进已抓取URL队列；
- 4.分析已抓取URL队列中的URL，分析其中的其他URL，并且将URL放入待抓取URL队列，从而进入下一个循环。

1.7 | 城市大数据的获取

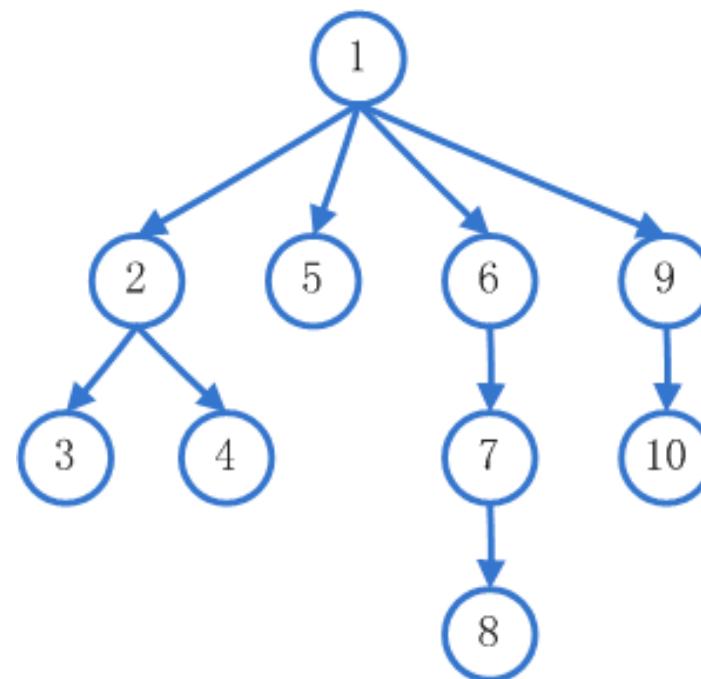


■ 网络爬虫

两种搜索方式：



广度优先搜索



深度优先搜索

■ 网络爬虫

正则表达式：

正则表达式(regular expression)描述了一种字符串匹配的模式，可以用
来检查一个串是否含有某种子串、将匹配的子串做替换或者从某个串
中取出符合某个条件的子串等。

例如：在查找硬盘上的文件时，会用通配符文件名中的单个字符

像data?. data这样的模式匹配下
列所有文件：

data1. data
data2. data
datax. data
dataN. data

使用*字符时， data*. data匹配
下列所有文件：

data. dat
data1. dat
data2. dat
data12. dat
datax. dat
dataXYZ. dat

1.7 | 城市大数据的获取



■ 网络爬虫

正则表达式：

在编写处理字符串的程序或网页时，经常会有查找符合某些复杂规则的字符串的需要。 正则表达式就是用于描述这些规则的工具。换句话说，正则表达式就是记录文本规则的代码，在网络爬虫中，可以使用正则表达式搜索整个网页，来获取想要的到的信息。

例如：

假设要在一篇英文小说里查找hi，你可以使用正则表达式：`\bhi\b`

假如你要找的是hi后面不远处跟着一个Lucy，你应该用：
`\bhi\b.*\bLucy\b`

1.7 | 城市大数据的获取



■ 网络爬虫

正则表达式（常用）：

表1. 常用的元字符

代码	说明
----	----

表2. 常用的限定符

代码/语法	说明
-------	----

说明	正则表达式
网址 (URL)	[a-zA-z]+://[^\\s]*
IP地址 (IP Address)	((2[0-4]\\d 25[0-5] [01]?\\d\\d?)\\.){3}(2[0-4]\\d 25[0-5] [01]?\\d\\d?)
电子邮件 (Email)	\\w+([-.]\\w+)*@\\w+([-.]\\w+)*\\.\\w+([-.]\\w+)*
QQ号码	[1-9]\\d{4,}
HTML标记(包含内容或自闭合)	<(.*)>.*<\\/\\1><(.*)\\/>
密码(由数字/大写字母/小写字母/标点符号组成，四种都必有，8位以上)	(?=^.{8,}\$(?=.*\\d)(?=.*\\W+)(?=.*[A-Z])(?=.*[a-z])(?!.*\\n).*\$
日期(年-月-日)	(\\d{4} \\d{2})-((1[0-2]) (0?[1-9]))-(([12][0-9]) (3[01]) (0?[1-9]))
日期(月/日/年)	((1[0-2]) (0?[1-9]))/(([12][0-9]) (3[01]) (0?[1-9]))/(\\d{4} \\d{2})
时间(小时:分钟，24小时制)	((1 0?)[0-9] 2[0-3]):([0-5][0-9])
汉字(字符)	[\\u4e00-\\u9fa5]
中文及全角标点符号(字符)	[\\u3000-\\u301e\\ufe10-\\ufe19\\ufe30-\\ufe44\\ufe50-\\ufe6b\\uff01-\\uffee]
中国大陆固定电话号码	(\\d{4}- \\d{3}-)?(\\d{8} \\d{7})
中国大陆手机号码	1\\d{10}
中国大陆邮政编码	[1-9]\\d{5}
中国大陆身份证号(15位或18位)	\\d{15} (\\d\\d[0-9xX])?
非负整数(正整数或零)	\\d+
正整数	[0-9]*[1-9][0-9]*
负整数	-[0-9]*[1-9][0-9]*
整数	-?\\d+
小数	(-?\\d+)(\\.\\d+)?
不包含abc的单词	\\b((?!abc)\\w)+\\b

1.7 | 城市大数据的获取



■ 网络爬虫

正则表达式（实验）：

```
27, 4d1daa72edf89f9a, 32, 20120323 19:23:24 114.2130 22.6074,  
20120323 04:07:36 114.1003 22.6205, 20120323 08:07:45  
114.1034 22.6300, 20120323 22:21:02 114.1005 22.6274,  
20120323 18:58:31 114.2624 22.5918, 20120323 01:07:30  
114.1034 22.6300, 20120323 05:07:39 114.1003 22.6205,  
20120323 05:07:39 114.1003 22.6205, 20120323 02:07:32  
114.1034 22.6300, 20120323 09:31:13 114.0396 22.6750,  
20120323 03:07:34 114.1034 22.6300, 20120323 19:42:18  
114.1005 22.6274, 20120322 23:07:25 114.1034 22.6300,  
20120323 08:07:45 114.1034 22.6300, 20120323 21:55:44  
114.1003 22.6205, 20120323 20:42:14 114.1003 22.6205,  
20120323 22:38:27 114.1034 22.6300, 20120323 07:07:43  
114.1034 22.6300, 20120323 23:27:43 114.1003 22.6205,  
20120323 01:07:30 114.1034 22.6300, 20120323 03:07:34  
114.1034 22.6300, 20120323 00:07:28 114.1034 22.6300,  
20120323 20:42:14 114.1003 22.6205, 20120323 00:07:28  
114.1034 22.6300, 20120323 09:27:25 114.0580 22.6637,  
20120323 19:42:18 114.1005 22.6274, 20120323 09:31:13  
114.0396 22.6750, 20120323 06:07:41 114.1003 22.6205,  
20120323 07:07:43 114.1034 22.6300, 20120323 04:07:36  
114.1003 22.6205, 20120323 02:07:32 114.1034 22.6300,  
20120323 06:07:41 114.1003 22.6205
```

Transform data format
via Notepad++



```
114.2130,22.6074  
114.1003,22.6205  
114.1034,22.6300  
114.1005,22.6274  
114.2624,22.5918  
114.1034,22.6300  
114.1003,22.6205  
114.1003,22.6205  
114.1034,22.6300  
114.0396,22.6750  
114.1034,22.6300  
114.1005,22.6274  
114.1034,22.6300  
114.1034,22.6300  
114.1003,22.6205  
114.1003,22.6205  
114.1034,22.6300  
114.1034,22.6300  
114.1003,22.6205  
114.1034,22.6300  
114.1034,22.6300  
114.1003,22.6205  
114.1034,22.6300  
114.1034,22.6300  
114.1003,22.6205  
114.1034,22.6300  
114.1034,22.6300  
114.1003,22.6205  
114.1034,22.6300  
114.1003,22.6205  
114.1034,22.6300  
114.1034,22.6300  
114.1003,22.6205  
114.1034,22.6300  
114.1003,22.6205  
114.1034,22.6300  
114.1003,22.6205
```

1.7 | 城市大数据的获取



■ 网络爬虫

正则表达式(例子):

在获取百度POI时，当返回xml格式的文件时，如下图：

```
<results>
    ...
    <result>
        <name>华南理工大学北区</name>
        <location>
            <lat>23.169952</lat>
            <lng>113.348495</lng>
        </location>
```

使用正则表达式时：

<name>(.*)</name> 可获取地名，即华南理工大学北区
<lat>(.*)</lat> 可获取纬度，即23.169952
<lon>(.*)</lon> 可获取经度，即113.348495

(.*)是一个分组，用整个正则表达式去搜索整个网页时，返回值就是(.*?)
的值



■ 网络爬虫

网站反爬虫机制

(Main reasons: 1 Private security 2 Data security 3 Anti-DDOS attack)

- 1、手工识别和拒绝爬虫的访问
- 2、通过识别爬虫的User-Agent信息来拒绝爬虫
- 3、通过网站流量统计系统和日志分析来识别爬虫
- 4、网站的实时反爬虫防火墙实现策略

<http://www.chinaz.com/tags/gugepachong.shtml>



■ 关于爬虫的其他

模拟登录

有些网站设置了权限，只有在登录了之后才能爬取网站的内容，如何模拟登录，目前的方法主要是利用浏览器cookie模拟登录。

假设，我们需要爬取淘宝上的订单数据，这是我们就需要模拟登陆的机制：

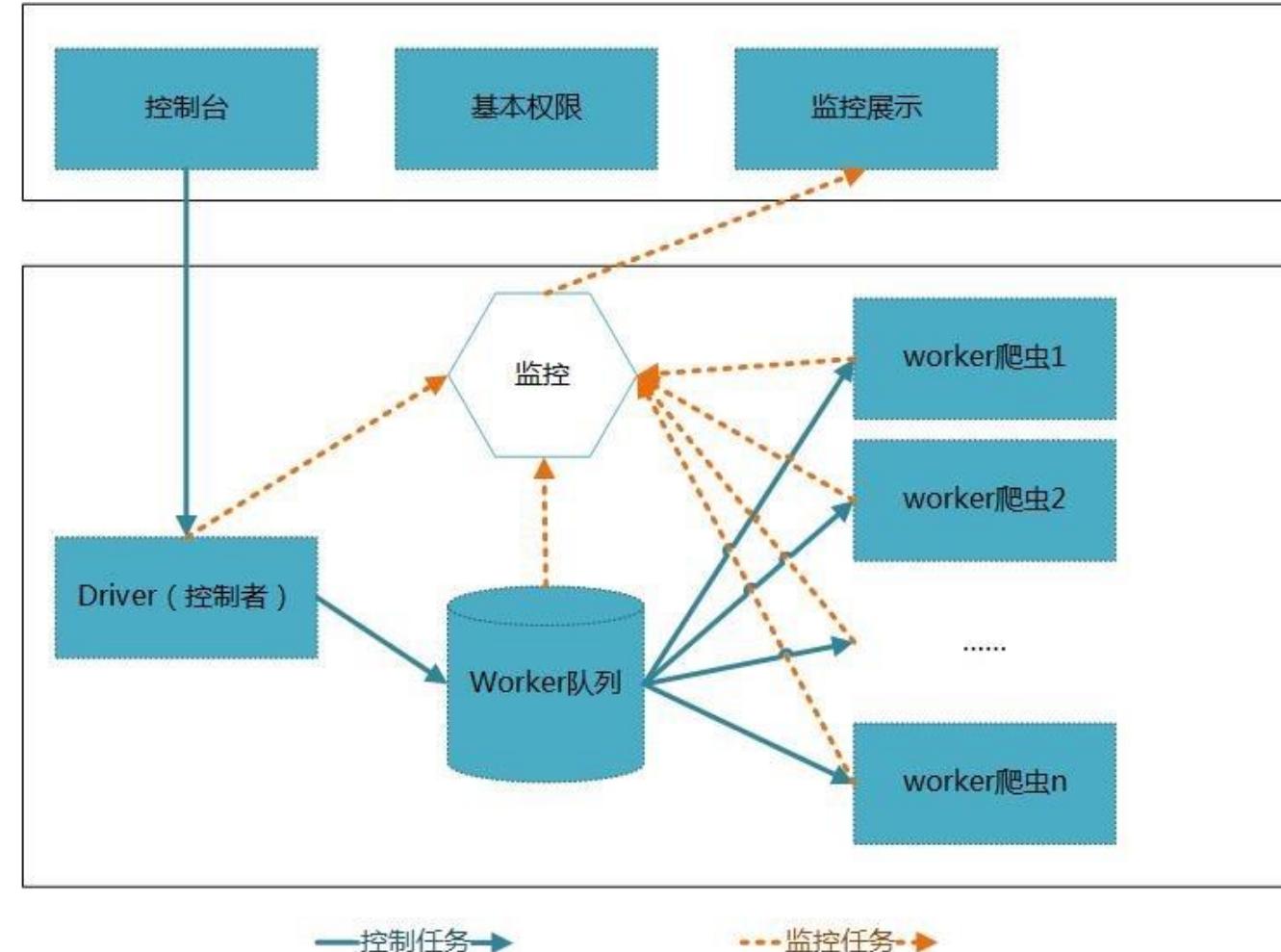
1. 设置一个cookie处理器，它负责从服务器下载cookie到本地，并且在发送请求时带上本地的cookie；
2. 向淘宝登录页面发送一个请求Request，包括登录界面地址，上传的数据包post data(上传之前应先编码，使得与服务器编码一致)；
3. 利用urllib2.urlopen发送请求，得到响应Response；
4. 查看响应结果。

1.7 | 城市大数据的获取



■ 关于爬虫的其他

多线程





■ 关于爬虫的其他

内网爬虫

相比于一般的网络爬虫，内网爬虫将检索范围限定在局域网内。

例如，企业内部的爬虫，要求爬虫能够处理一个公司内部不同类型的信息源，根据网页上的超链接来发现外部的和内部的（限定在企业内部网络）页面，而且有时需要能够扫描公司的和个人的目录，来发现电子邮件、文档、讲稿、数据库以及其他公司的信息。

<http://www.cnblogs.com/Lawson/archive/2012/12/21/2828631.html>

1.7 | 城市大数据的获取



■ 关于爬虫的其他

常用软件：八爪鱼

The screenshot shows the homepage of the BaZhuYu website. At the top, there is a navigation bar with links for '产品' (Products), '价格' (Prices), '解决方案' (Solution), '软件下载' (Software Download), '免费模板' (Free Templates), '教程' (Tutorials), '帮助' (Help), and '论坛' (Forum). There are also '登录' (Login) and '注册' (Register) buttons. The main banner features the text '八爪鱼采集器 · 百万用户的选择' (Octopus Data Collector · Millions of User Choices) and '从海量数据中挖掘价值如此简单' (Extracting value from massive data is this simple). It includes a large image of a computer monitor displaying a web browser with the 'e' logo, surrounded by data visualization icons like a pie chart and a scatter plot. Below the banner are two buttons: '免费下载' (Free Download) in green and '一分钟了解八爪鱼' (Get to know Octopus in one minute) in white. A vertical sidebar on the right contains icons for '人工客服' (Customer Service), 'QQ群' (QQ Group), and '免费试用' (Free Trial), along with a back arrow icon. At the bottom, there is a section titled '零门槛三步获取数据' (Three steps to get data with zero threshold) and a note: '不懂网络爬虫技术，也可轻松采集数据' (Even if you don't understand network crawling technology, you can easily collect data).



八爪鱼 让数据触手可及
www.bazhuayu.cc

1.7 | 城市大数据的获取



■ 实例

I 抓取煎蛋网里的图片

煎蛋

首页 专题 小组 小电影 段子 妹子 无聊图 热榜

今天 · 昨天

搜索

[SPONSORS]

24H热文 三日最赞 一周话题

1	一只咸鱼 / Geek 在健康与环境问题上获得突破的少年科学家们 在最近的英特尔科学奖赛中，全美的高中生展示了各种利用数学和科学来解决世界问题的方式。 ☆0	9	蛋花 / 无厘头研究 奢侈的眼泪：哭泣不是因为苦难，恰恰相反 在某种程度而言，眼泪是一种奢侈品。 ☆3	11	Cedric / 环保 争议终结：海洋世界宣布终止逆戟鲸饲养计划 在经过多年的争议过后，海洋世界决定停止饲养逆戟鲸的计划。 ☆4	1	最酷纪录片: J.J. Abrams镜头中的谷歌登月 8,597浏览
2	[NSFW]无言的表达：自然和人体的交融 8,488浏览	2	挖坟不止：图坦卡蒙陵墓的惊人新发现 6,146浏览	3	NASA打算放一场“太空大火” 6,723浏览	3	万能蛇毒解药即将问世 4,636浏览
4	证明费马最后定理的英国数学家，终获菲尔兹奖 4,572浏览	4	慕残者的秘密世界 4,535浏览	5	有阅读障碍是种什么感受？ 4,404浏览	5	

1.7 | 城市大数据的获取



■ 实例

I 抓取煎蛋网里的图片

```
#打开一个url并且返回html
def urlopen(url):
    user_agent = 'Mozilla/
headers = { 'User-Agent': user_ag
req = urllib2.Request(url)
html = urllib2.urlopen(req)
return html
```

```
<li id="comment-3090240">
<div>
    <div class="row">
        <div class="author"><strong
            title="防伪码: b1b27b724955d9e6c30aa636a195b2104eeeaf6e7" >刘北习</strong>
<br>
        <small><a href="#footer" title="@回复"
            onclick="document.getElementById('comment').value += '&#39;@<a
href='http://jandan.net/pic/page-8672#comment-3090240'&gt;刘北习</a>:&#39;;">@32 mins ago</a></span>
</small>
    </div>
    <div class="text"><span class="righttext"><a href="http://jandan.net/pic/page-8672#comment-
3090240">216779</a></span><p><a href="http://ww1.sinaimg.cn/large/7ad1d5c7gw1f1svo2kv57j20c82mxqhv.jpg"
target="_blank" class="view_img_link">[查看原图]</a><br /><br />
<a href="http://ww3.sinaimg.cn/large/7ad1d5c7gw1f1svo6lw0tj20c80x90va.jpg" target="_blank" class="view_img_link">[查
看原图]</a><br /></p>
<div class="vote" id="vote-3090240"><span id="acv_stat_3090240"></span><a title="圈圈/支持" class="acvclick acv4"
id="vote4-3090240" href="javascript:acv_vote(3090240, 1);">00</a> [<span id="cos_support-3090240">3</span>] <a
title="叉叉/反对" class="acvclick acva" id="votea-3090240" href="javascript:acv_vote(3090240, 0);">XX</a> [<span
id="cos_unsupport-3090240">31</span>]</div>
</div>
<span class="break"></span></div>
</li>
```

1.7 | 城市大数据的获取



■ 实例

I 抓取煎蛋网里的图片

使用正则表达式提取每张图片的URL，由于：

1. 每张图片有一个唯一标识符，每张图片有一个URL。
2. 每张图片在html中都是以<li id="comment-xxxxxxx">开始，以结束。

正则表达式

```
<li id="comment-.*?">.*?<div class="text"><span class=".*?">
<a href=".*?">(.*)</a></span><p>.*?<a href="(.*?)">.*?</li>
```

即可获取到每张图片的ID以及URL，之后打开图片的URL，获取到数据然后存储，完成图片爬取。

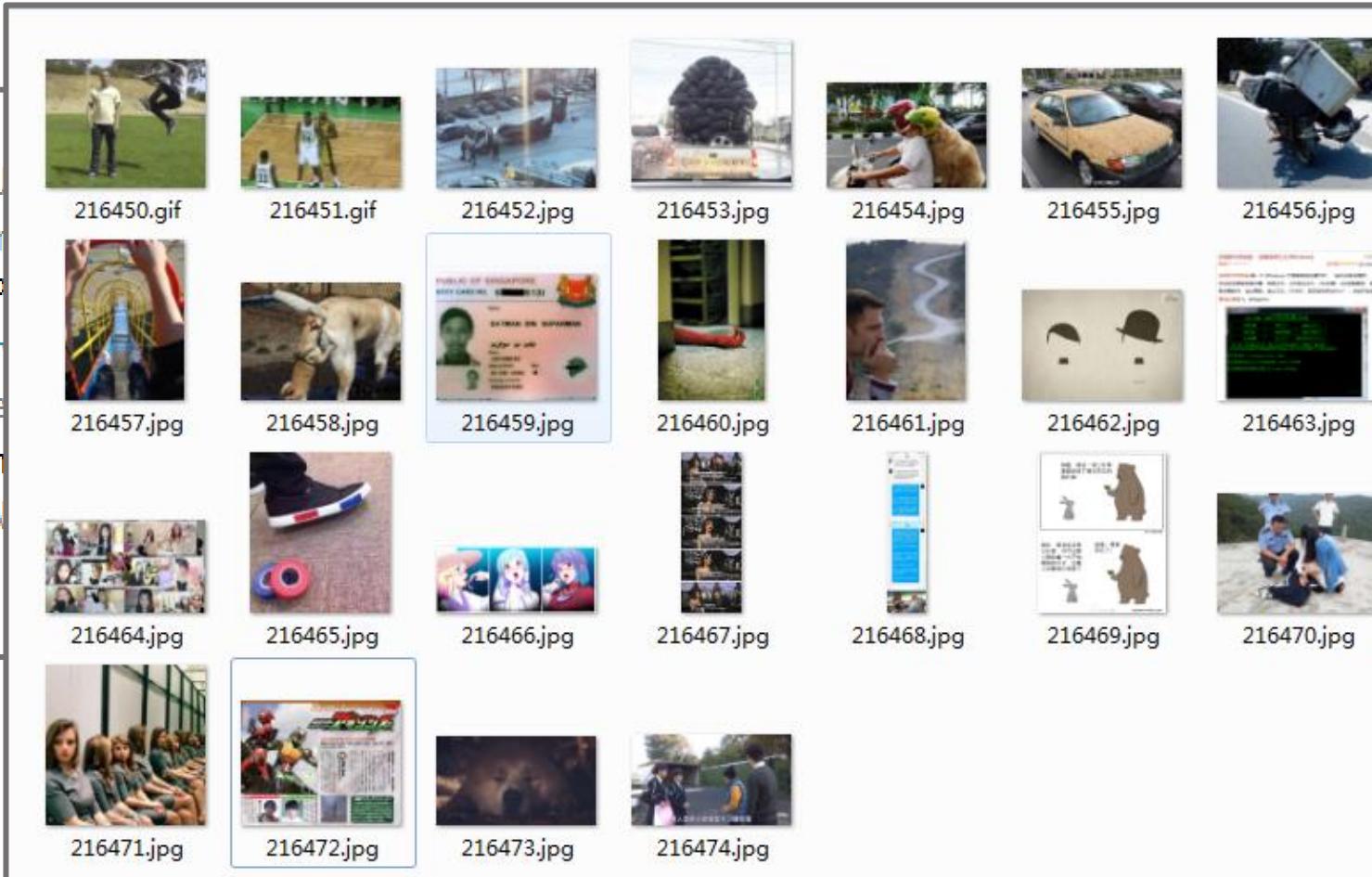
1.7 | 城市大数据的获取



■ 实例

I 抓取煎蛋网里的图片

```
#保存图片
def saveImg(imageUrl):
    splitPath = imageUrl.split("/")
    fTail = splitPath[-1]
    fileName = str(fTail)
    print(fileName)
    data = urlopen(imageUrl)
    f = open(fileName, "wb")
    f.write(data)
    f.close()
```



1.7 | 城市大数据的获取



■ 实例

II 获取百度POI

城市内检索 ▼

参数	值	说明
query:	银行	检索关键字
scope:	1	检索结果详细程度。取值为1或空，则返回基本信息；取值为2，返回检索POI详细信息。
page_size:	10	返回记录数量，默认为10条记录，最大返回结果为20条。
page_num:	0	分页页码，默认为0,0代表第一页，1代表第二页，以此类推。
region:	北京	检索区域，如果取值为“全国”或某省份，则返回指定区域的POI。

运行 (结果显示如下)

http://api.map.baidu.com/place/v2/search?ak=您的密钥&output=json&query=%E9%93%B6%E8%A1%8C&page_size=10&page_num=0&scope=1®ion=%E5%8C%97%E4%BA%AC

查看百度API的说明，按其url格式写好url，然后在python中打开，则返回设置的page_size的个数的POI，再用正则表达式进行解析，

1.7 | 城市大数据的获取



■ 实例

II 获取百度POI

每一个POI的xml信息：

```
<result>
    <name>华南理工大学北区</name>
    <location>
        <lat>23.169952</lat>
        <lng>113.348495</lng>
    </location>
    <address>广州市天河区五山路381</address>
    <telephone>020-87110000</telephone>
    <detail>1</detail>
    <uid>96b672aaffe9bb75ce04efea</uid>
    <detail_info>
        <tag>教育培训;高等院校</tag>
        <detail_url>
            http://api.map.baidu.com/place/detail?uid=96b672aaffe9bb75ce04efea&output=html&source=placeapi\_v2
        </detail_url>
        <type>education</type>
        <overall_rating>2.4</overall_rating>
        <service_rating>0</service_rating>
        <technology_rating>0</technology_rating>
        <image_num>37</image_num>
        <comment_num>7</comment_num>
    </detail_info>
</result>
```

1.7 | 城市大数据的获取



■ 实例

通过正则表达式搜索整个页面即可得到每个POI点的信息：

II 获取百度POI

```
def scanhtml(html,page_num,_fw):  
  
    print u'          扫描第 %s 页' % page_num  
  
    pattern = re.compile(r'<result>(.*)</result>',re.S) # ?非贪婪  
    result = re.findall(pattern,html)  
    print u'-----本页 %d 个结果-----' % len(result)  
    for child in result:  
        pattern = re.compile(r'  
        tmp0=re.findall(pattern  
        tmp0.append('Null') #  
  
        pattern = re.compile(r'  
        tmp1=re.findall(pattern  
        tmp1.append('Null')  
  
        pattern = re.compile(r'  
        tmp2=re.findall(pattern  
        tmp2.append('Null')
```

什么是正则表达式的贪婪与非贪婪匹配

如：String str="abcaxc";

Patter p="ab*c";

贪婪匹配: 正则表达式一般趋向于最大长度匹配，也就是所谓的贪婪匹配。如上面使用模式p匹配字符串str，结果就是匹配到：abcaxc(ab*c)。

非贪婪匹配: 就是匹配到结果就好，就少的匹配字符。如上面使用模式p匹配字符串str，结果就是匹配到：abc(ab*c)。



II 获取百度POI

通过正则表达式搜索整个页面即可得到每个POI点的信息：

```
pattern = re.compile(r'<address>(.*)</address>',re.S)
tmp3=re.findall(pattern,child)
tmp3.append('Null')

pattern = re.compile(r'<telephone>(.*)</telephone>',re.S)
tmp4=re.findall(pattern,child)
tmp4.append('Null')

pattern = re.compile(r'<tag>(.*)</tag>',re.S)
tmp5=re.findall(pattern,child)
tmp5.append('Null')

pattern = re.compile(r'<overall_rating>(.*)</overall_rating>',re.S)
tmp6=re.findall(pattern,child)
tmp6.append('Null')
```

1.7 | 城市大数据的获取



■ 实验

实验1 (API调用, python 2.7) :

..|Crawler_Code_examples|Crawler_Code\baidu_xml.py

API说明:

<http://lbsyun.baidu.com/index.php?title=webapi/guide/webservice-placeapi>

测试用密钥

(仅供测试, 3天后过期, 请自行在百度地图API中心申请) :

LqtWYR0GCK4yIbU6AB2GGmP1

实验2 (网络爬虫, python 2.7) :

网络爬虫 (抓取 <http://jandan.net> 无聊图的数据) :

.. \Crawler_Code_examples\crawler_code\jandan_pic.py

Python tools for Visual Studio

<http://microsoft.github.io/PTVS/>

VS2010/2012/2013/2015的安装包见:

..\python tools for VS

The screenshot shows the Microsoft Visual Studio interface with the title bar "PythonApplication1 - Microsoft Visual Studio (Administrator)". The main window displays a Python script named "Program.py". The code contains two function definitions: "DemoPyTools" and "Bar". The "DemoPyTools" function returns a URL. The "Bar" function calls "DemoPyTools" and stores its result in "t". A tooltip "DemoPyTools" is visible near the cursor.

```
def DemoPyTools():
    foo = 'http://greenerycn.cnblogs.com'
    return foo

def Bar():
    t = DemoPyTools()
```

技术伦理！

数据伦理！

学术伦理！

不作恶！ ! !

不欺瞒！ ! !



章节内容



- 1 大数据的来源
- **2 大数据的概念**
- 3 大数据的影响
- 4 大数据的关键技术
- 5 大数据的计算模式
- 6 大数据产业
- 7 大数据与云计算、物联网的关系

02 | 什么是数据?



半结构化/非结构化数据

Web
Clickstream



DOC / Media



Social Media



Machine / Sensor



Call Log



Apps



02 | 什么是大数据?



何为大? —数据度量

1Byte = 8 Bit

1KB = 1,024 Bytes

1MB = 1,024 KB = 1,048,576 Bytes

1GB = 1,024 MB = 1,048,576 KB = 1,073,741,824 Bytes

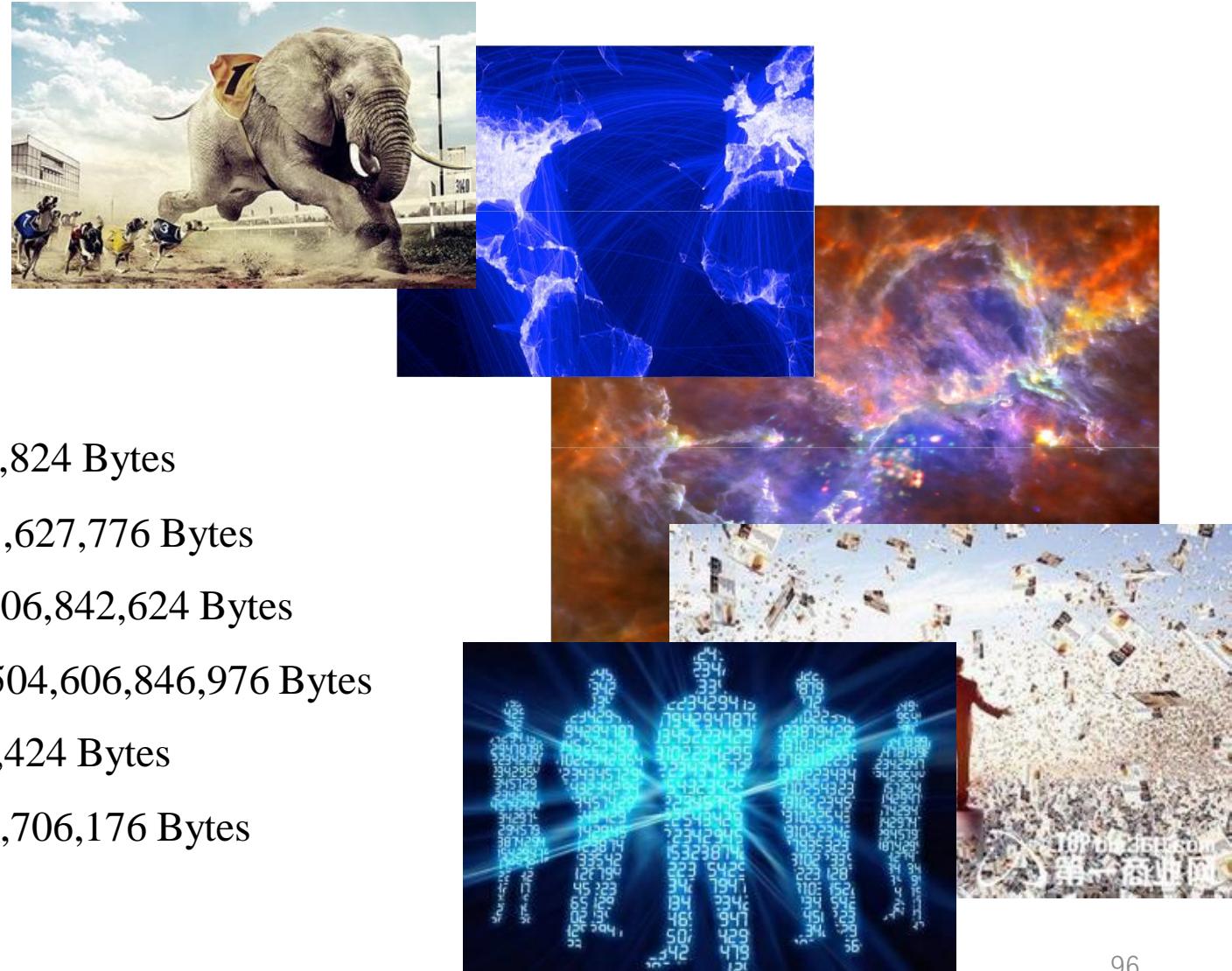
1TB = 1,024 GB = 1,048,576 MB = 1,099,511,627,776 Bytes

1PB = 1,024 TB = 1,048,576 GB = 1,125,899,906,842,624 Bytes

1EB = 1,024 PB = 1,048,576 TB = 1,152,921,504,606,846,976 Bytes

1ZB = 1,024 EB = 1,180,591,620,717,411,303,424 Bytes

1YB = 1,024 ZB = 1,208,925,819,614,629,174,706,176 Bytes



02 | 什么是大数据?



➤ 数据没有办法在可容忍的时间内使用常规软件方法完成存储、管理和处理任务

- 《红楼梦》含标点87万字（不含标点853509字）

每个汉字占两个字节：1汉字=16bit = 2*8位=2bytes

1GB 约等于 671部红楼梦

1TB 约等于 631,903 部

1PB 约等于 647,068,911部

- 美国国会图书馆藏书151,785,778册（2011年4月：收录数据235TB）

中国国家图书馆藏书26,310,000册

1EB = **4000倍** 美国国会图书馆存储的信息量

- 600美元的硬盘就可以存储全世界所有的歌曲

MGJ估计，全球企业 2010 年在硬盘上存储了超过 7EB(1EB 等于 10 亿 GB)的新数据，同时，消费者在 PC 和笔记本等设备上存储了超过 6EB 新数据

02 | 什么是大数据？



- **大数据**指无法在一定时间范围内用常规软件工具进行捕捉、管理和处理的数据集合，需要新处理模式才能具有更强决策力、洞察发现力和流程优化能力来适应海量、高增长率和多样化的信息资产。
- 大数据就是“未来的新石油”。

02| 大数据的4V特征



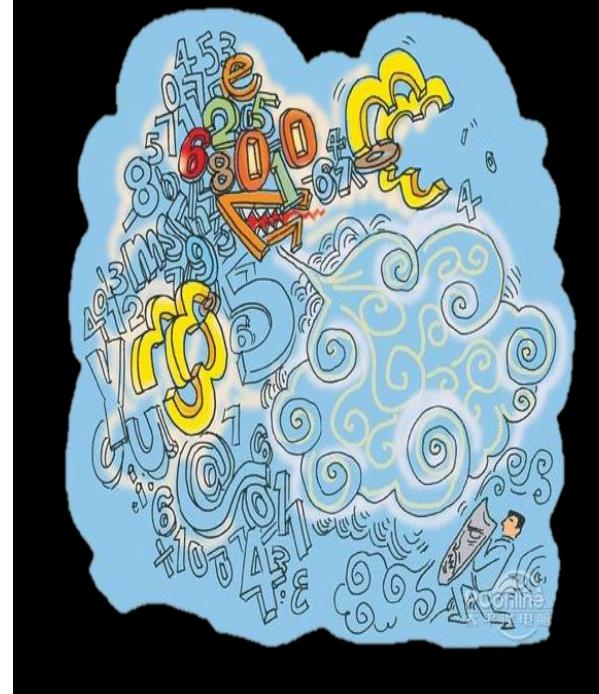
Volume

- 非结构化数据的超大规模和增长
- 总数据量的80~90%
- 比结构化数据增长快10倍到50倍
- 是传统数据仓库的10倍到50倍

Value

- 大量的不相关信息
- 对未来趋势与模式的可预测分析
- 深度复杂分析（机器学习、人工智能Vs传统商务智能）

Big Data 大数据



Variety

- 大数据的异构和多样性
- 很多不同形式（文本、图像、视频、机器数据）
- 无模式或者模式不明显
- 不连贯的语法或句义

Velocity

- 实时分析而非批量式分析
- 数据输入、处理与丢弃
- 立竿见影而非事后见效

02| 大数据的4V特征 (Volume)



- 根据IDC作出的估测，数据一直都在以每年50%的速度增长，也就是说每两年就增长一倍（大数据摩尔定律）
- 人类在最近两年产生的数据量相当于之前产生的全部数据量
- 预计到2020年，全球将总共拥有35ZB的数据量，相较于2010年，数据量将增长近30倍

一般情况下，大数据是以PB、EB、ZB为单位进行计量的

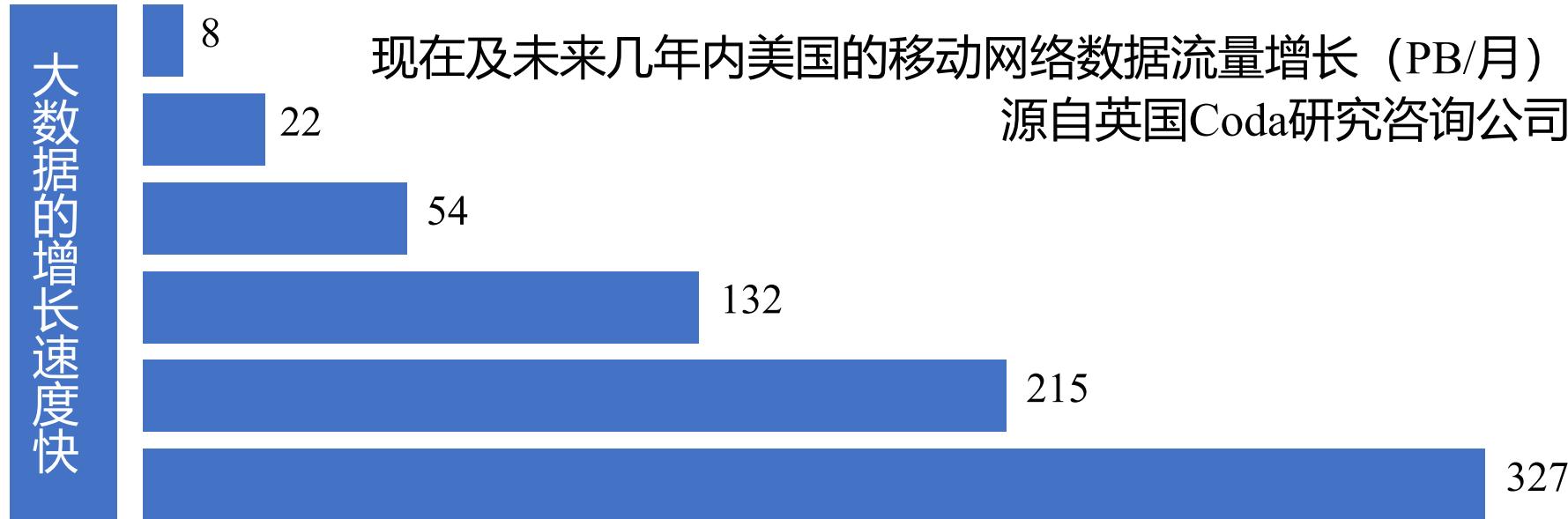
1PB相当于50%的全美学术研究图书馆藏书信息内容

5EB相当于至今全世界人类所讲过的话语

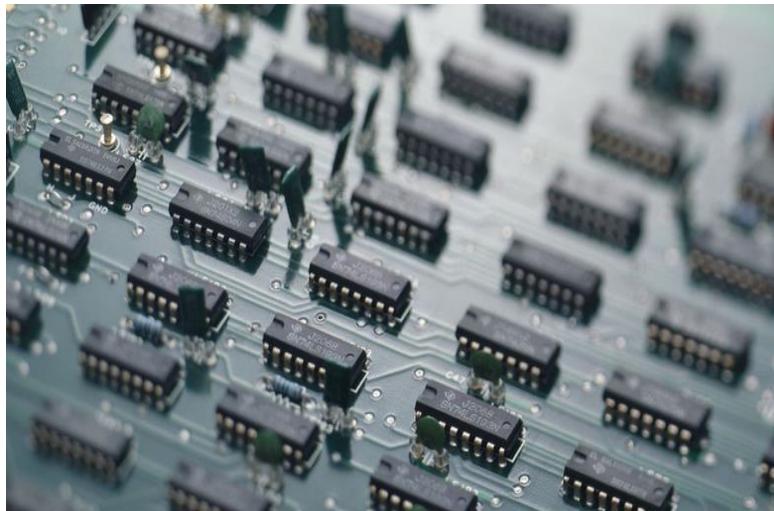
1ZB如同全世界海滩上的沙子数量总和

1YB相当于7000位人类体内的微细胞总和

02| 大数据的4V特征 (Velocity)

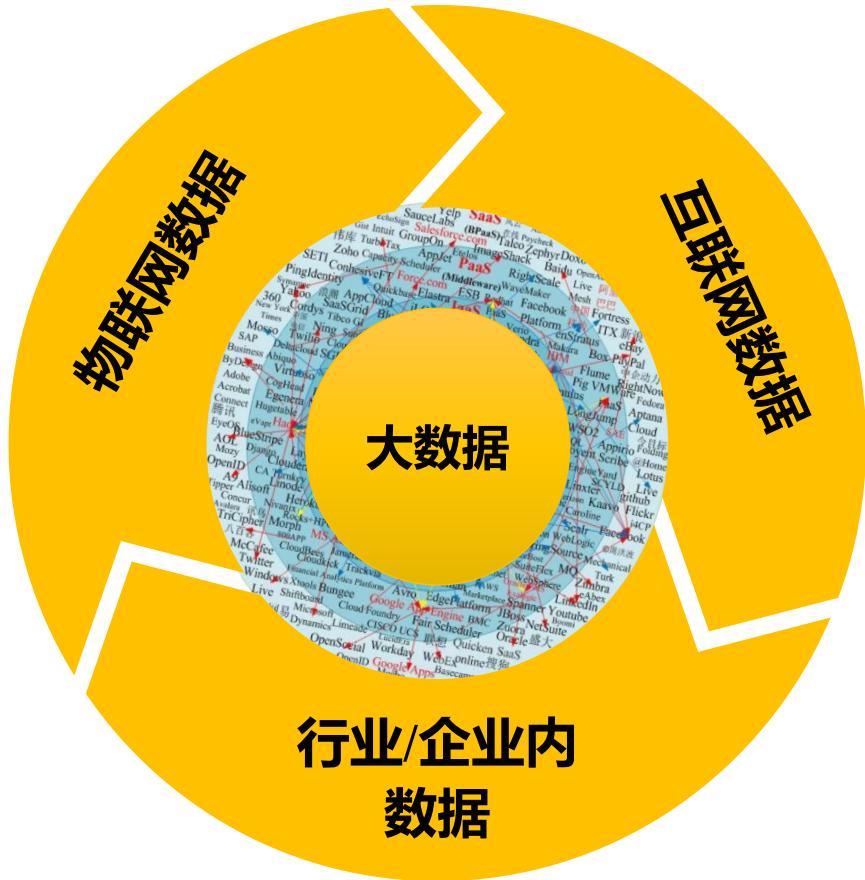


大数据的处理速度快



- 实时数据流处理的要求，是区别大数据引用和传统数据仓库技术，BI技术的关键差别之一；
- 1s 是临界点，对于大数据应用而言，必须要在1秒钟内形成答案，否则处理结果就是过时和无效的。

02| 大数据的4V特征 (Variety)



数据来源多

企业内部多个应用系统的数据、互联网和物联网的兴起，带来了微博、社交网站、传感器等多种来源。

数据类型多

保存在关系数据库中的结构化数据只占少数，70~80%的数据是如图片、音频、视频、模型、连接信息、文档等非结构化和半结构化数据。

关联性强

数据之间频繁交互，比如游客在旅行途中上传的图片和日志，就与游客的位置、行程等信息有了很强的关联性。

02| 大数据的4V特征 (Value)



➤ 价值密度低，商业价值高

以视频为例，连续不间断监控过程中，可能有用的数据仅仅有一两秒，但是具有很高的商业价值。

- 挖掘大数据的价值类似沙里淘金，从海量数据中挖掘稀疏但珍贵的信息；
- 价值密度低，是大数据的一个典型特征





章节内容



- 1 大数据的来源
- 2 大数据的概念
- **3 大数据的影响**
- 4 大数据的关键技术
- 5 大数据的计算模式
- 6 大数据产业
- 7 大数据与云计算、物联网的关系

➤在**思维方式**方面，大数据完全颠覆了传统的思维方式：

- 全样而非抽样
- 效率而非精确
- 相关而非因果



- 在**社会发展**方面，大数据决策逐渐成为一种新的决策方式，大数据应用有力促进了信息技术与各行业的深度融合，大数据开发大大推动了新技术和新应用的不断涌现。
- 在**就业市场**方面，大数据的兴起使得数据科学家成为热门职业。
- 在**人才培养**方面，大数据的兴起，将在很大程度上改变中国高校信息技术相关专业的现有教学和科研体制。

03 | 大数据挑战



大数据的安全威胁

主菜单

数据库目录 5

开房数据库

数据表列表

查询结果

Show 10 entries Search:

姓名	性别	身份证号	住址	手机号码	电子邮箱	登记日期
徐瑞	M	51	北部新区黄山大所		XURUI814@163.COM	2010-7-26 4:35:02
徐瑞	M	51	成都市青羊区宁			2012-3-18 14:03:36
徐瑞	F	51	四川省南充市顺			2010-10-25 2:50:10
徐瑞	F	52	贵州省遵义县泮			2011-6-9 4:51:38
徐瑞	F	52	贵州省遵义市汇			2012-6-6 7:52:40
徐瑞	M	53	曲靖市			2012-10-13 22:02:01
徐瑞	M	53	云南省昆明市寻			2012-10-13 16:25:26
徐瑞	M	53	昆明市			2012-10-14 6:26:37
徐瑞	F	53	0			2012-11-20 14:08:50
徐瑞	F	61				2011-12-17 7:15:51

Showing 211 to 220 of 256 entries

Previous 1 ... 21 22 23 ... 26 Next

站长统计



- 1 大数据的来源
- 2 大数据的概念
- 3 大数据的影响
- **4 大数据的关键技术**
- 5 大数据的计算模式
- 6 大数据产业
- 7 大数据与云计算、物联网的关系

04 大数据的关键技术



需求	技术	描述
海量数据存储技术	Hadoop, x86/MPP Map Reduce	分布式文件系统
实时数据处理技术	Streaming Data	流计算引擎
数据高速传输技术	Infini Band	服务器/存储间高速通信
搜索技术	Enterprise Search	文本检索、智能搜索、 实时搜索
数据分析技术	Text Analytics Engine Visual Data Modeling	自然语言处理、文本情感分析、 机器学习、聚类关联、数据模型

04 大数据的关键技术



新平台技术

- 基于SQL语言: 面对OLAP的传统行和列

TERADATA
Raising Intelligence

VERTICA

N NETEZZA

- 不基于SQL或map-reduce的: 由谷歌率先发起

cloudera

Datameer

ASTER

- 数据流: 基于运行商数据直接生成任意图形

LexisNexis™

LexisNexis™

不同范围的服务

数据入口/汇聚

数据平台

分析

ISG

TransUnion

ACXION
Corporation

IBM

SAP

cloudera

Datameer

ASTER

LexisNexis™

THOMSON REUTERS

mahout

TM

新的传输方案

- 传统交付模式 - 单片或基于设备的解决方案

TERADATA
Raising Intelligence

N NETEZZA

- 云: 能够充分利用物理设施的弹性, 以实现处理快速增长数据的能力

Greenplum

Rosslyn
Analytics

spend analysis for everyone

PERVASIVE

GoodData

“数据库将演变成一个虚拟的, 基于云计算, 超级可扩展的分布式平台。”

- Forrester analyst Jim Kobielski

04 | 大数据的关键技术



	大数据 (Hadoop)	NoSQL	数据库	数据仓库
部署架构	水平扩展	水平扩展	大部分垂直扩展， 少数水平扩展	大部分水平扩展
数据类型	文件存储，没有 数据类型	简单数据类型	丰富的数据类型	丰富的数据类型
数据模型	非常简陋的数据 模型	简单灵活数据 模型	丰富的数据模型	完善丰富数据 模型
数据关系	没有数据关系描 述	非常简单的数 据关系描述	数据关系完善	数据关系完善
数据一致	无一致性	弱一致性	强一致性	强一致性

04| 大数据的关键技术



	大数据 (Hadoop)	NoSQL	数据库	数据仓库
数据安全	安全性很弱	安全性很弱	安全性很高	安全性很高
计算类型	离线批量处理, 只读, 低并发	实时CRUD操作, 海量并发	实时CRUD操作, 高并发	离线批量处理, 只读, 低并发
适用场景	低密度数据海量 存储, 数据预处 理, 预计算	高并发实时	在线交易, 查询, 报表	高价值数据统一 存储和计算平台
常见用例	日志处理, 用户 行为分析, 搜索 引擎	用户资料, 微博, 金融反欺诈	金融账户, 电信 计费, 税务等	企业数据仓库



章节内容



- 1 大数据的来源
- 2 大数据的概念
- 3 大数据的关键技术
- 4 大数据的影响
- **5 大数据的计算模式**
- 6 大数据产业
- 7 大数据与云计算、物联网的关系

05 | 大数据的计算模式



大数据计算模式	解决问题	代表产品
批处理计算	针对大规模数据的批量处理	MapReduce、Spark等
流计算	针对流数据的实时计算	Storm、S4、Flume、Streams、Puma、DStream、Super Mario、银河流数据处理平台等
图计算	针对大规模图结构数据的处理	Pregel、GraphX、Giraph、PowerGraph、Hama、GoldenOrb等
查询分析计算	大规模数据的存储管理和查询分析	Dremel、Hive、Cassandra、Impala等



目录



- 1 大数据的来源
- 2 大数据的概念
- 3 大数据的关键技术
- 4 大数据的影响
- 5 大数据的计算模式
- 6 大数据产业
- 7 大数据与云计算、物联网的关系

➤ **大数据产业**是指一切与支撑大数据组织管理和价值发现相关的企业经济活动的集合。

产业链环节	包含内容
IT基础设施层	包括提供硬件、软件、网络等基础设施以及提供咨询、规划和系统集成服务的企业，比如，提供数据中心解决方案的IBM、惠普和戴尔等，提供存储解决方案的EMC，提供虚拟化管理软件的微软、思杰、SUN、Redhat等
数据源层	大数据生态圈里的数据提供者，是生物大数据（生物信息学领域的各类研究机构）、交通大数据（交通主管部门）、医疗大数据（各大医院、体检机构）、政务大数据（政府部门）、电商大数据（淘宝、天猫、苏宁云商、京东等电商）、社交网络大数据（微博、微信、人人网等）、搜索引擎大数据（百度、谷歌等）等各种数据的来源

产业链环节	包含内容
数据管理层	包括数据抽取、转换、存储和管理等服务的各类企业或产品，比如分布式文件系统（如Hadoop的HDFS和谷歌的GFS）、ETL工具（Informatica、Datastage、Kettle等）、数据库和数据仓库（Oracle、MySQL、SQL Server、HBase、GreenPlum等）
数据分析层	包括提供分布式计算、数据挖掘、统计分析等服务的各类企业或产品，比如，分布式计算框架MapReduce、统计分析软件SPSS和SAS、数据挖掘工具Weka、数据可视化工具Tableau、BI工具（MicroStrategy、Cognos、BO）等等
数据平台层	包括提供数据分享平台、数据分析平台、数据租售平台等服务的企业或产品，比如阿里巴巴、谷歌、中国电信、百度等
数据应用层	提供智能交通、智慧医疗、智能物流、智能电网等行业应用的企业、机构或政府部门，比如交通主管部门、各大医疗机构、菜鸟网络、国家电网等



章节内容



- 1 大数据的来源
- 2 大数据的概念
- 3 大数据的关键技术
- 4 大数据的影响
- 5 大数据的计算模式
- 6 大数据产业
- 7 **大数据与云计算、物联网的关系**



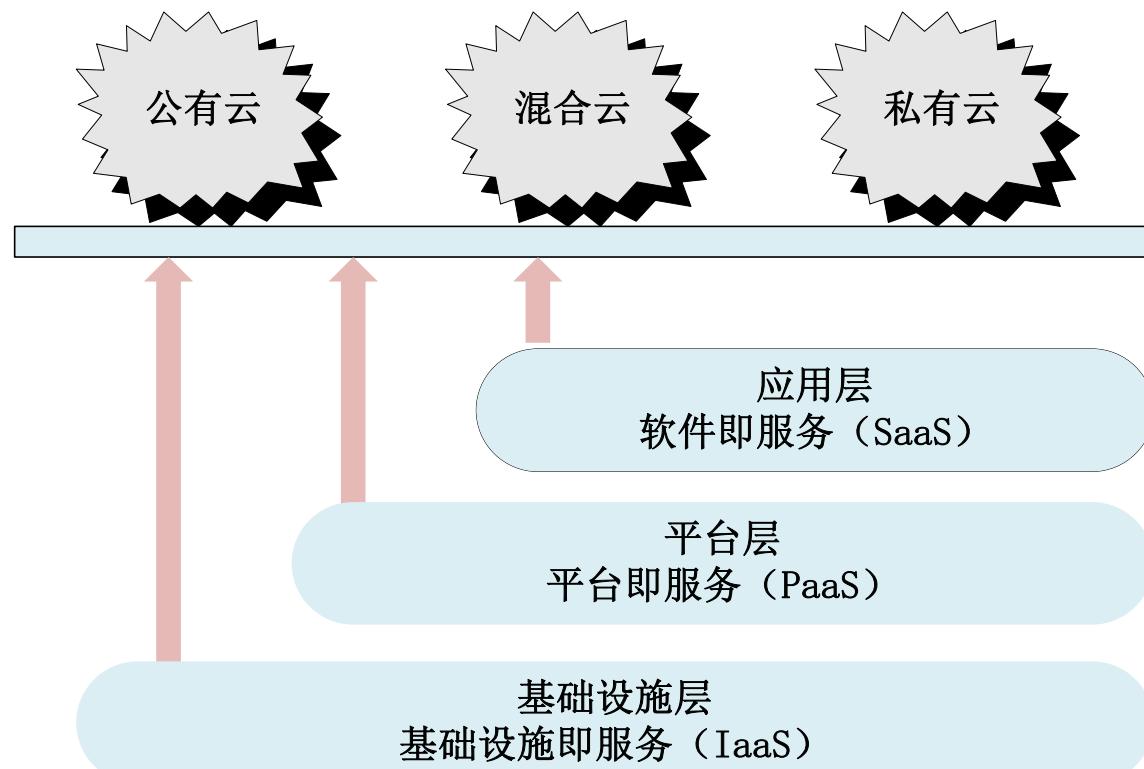
章节内容



- 7.1 云计算
- 7.2 物联网
- 7.3 大数据与云计算、物联网的关系

1. 云计算概念

云计算实现了通过网络提供可伸缩的、廉价的分布式计算能力，用户只需要在具备网络接入条件的地方，就可以随时随地获得所需的各种IT资源。

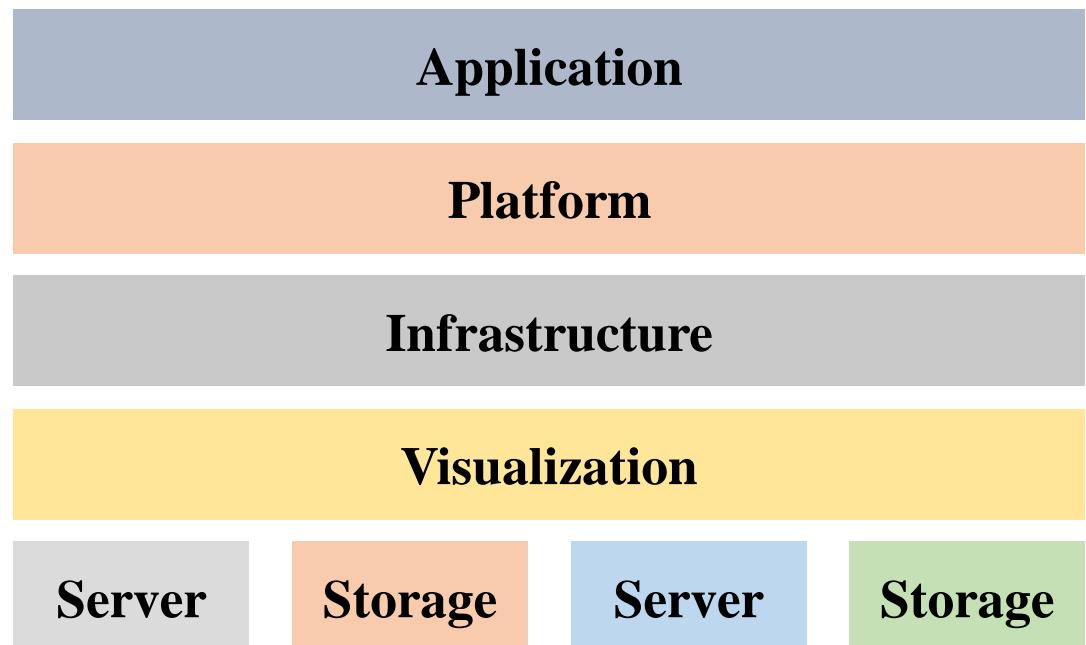


云计算的服务模式和类型

7.1 | 云计算



SaaS	从一个集中的系统部署软件，使之在一台本地计算机上(或从云中远程地)运行的一个模型。由于是计量服务，SaaS 允许出租一个应用程序，并计时收费
PaaS	类似于 IaaS，但是它包括操作系统和围绕特定应用的必需的服务
IaaS	将基础设施(计算资源和存储)作为服务出租

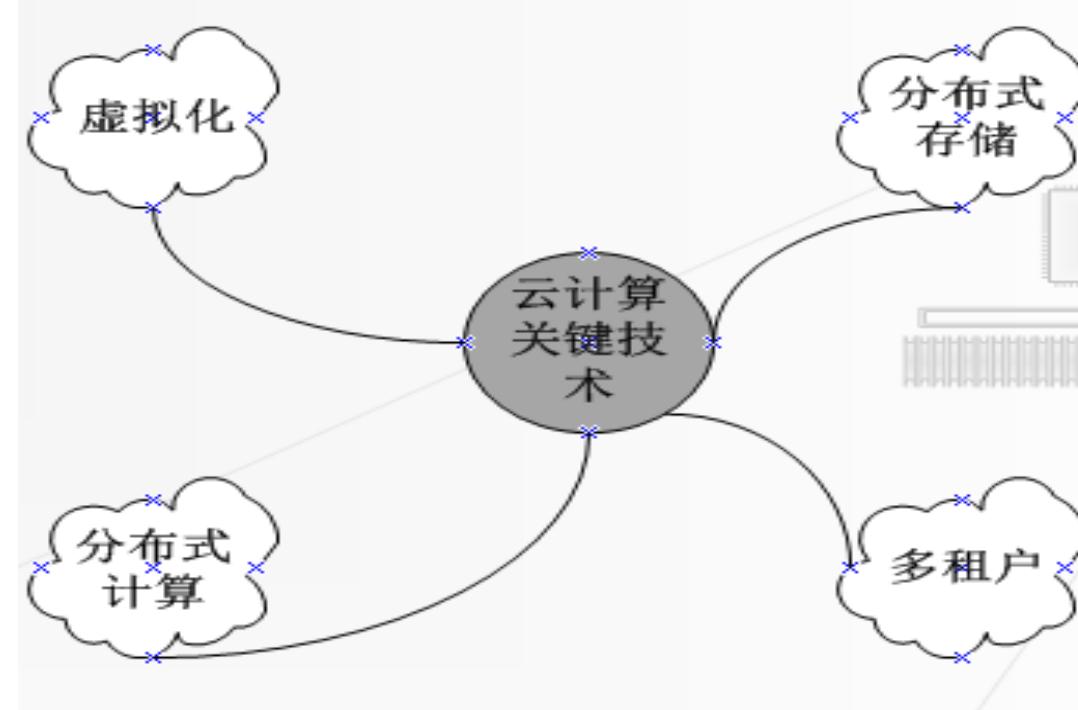


SaaS	Software as a Service
PaaS	Platform as a Service
IaaS	Infrastructure as a Service

Google Apps, Microsoft “Software+Services”
IBM IT factory, Google App Engine, Force.com
Amazon EC2, IBM Blue Cloud, Sun Grid

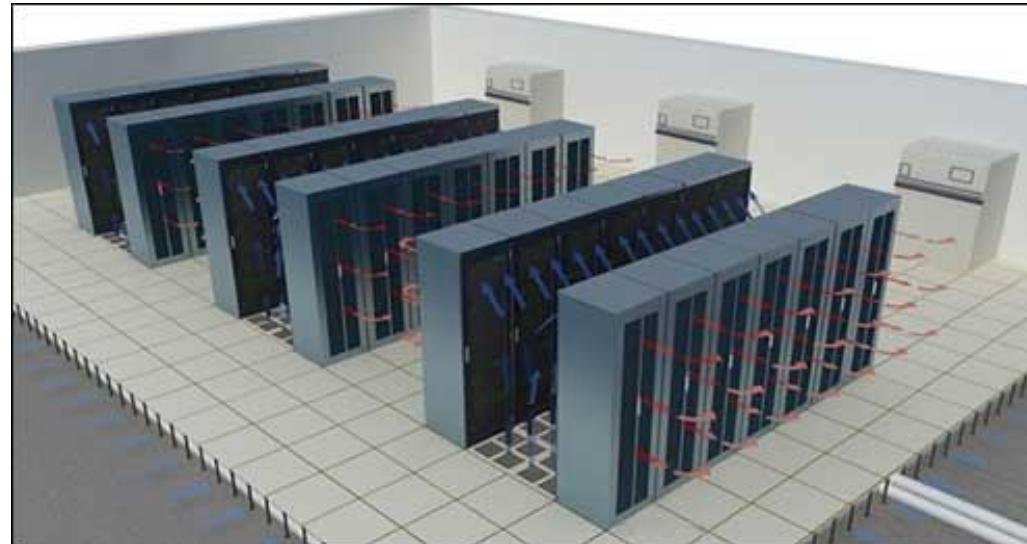
2. 云计算关键技术

- 云计算关键技术包括：虚拟化、分布式存储、分布式计算、多租户等



3. 云计算数据中心

- 云计算数据中心是一整套复杂的设施，包括刀片服务器、宽带网络连接、环境控制设备、监控设备以及各种安全装置等
- 数据中心是云计算的重要载体，为云计算提供计算、存储、带宽等各种硬件资源，为各种平台和应用提供运行支撑环境
- 全国各地推进数据中心建设

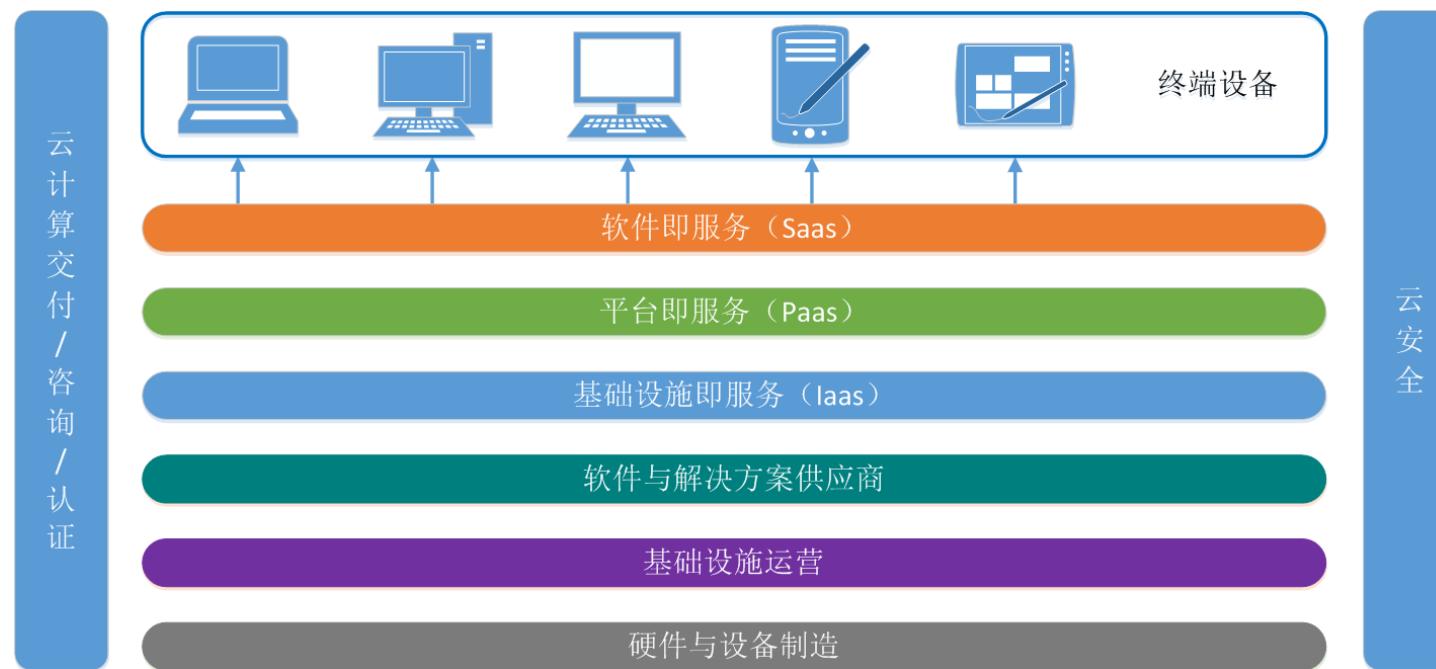


4. 云计算应用

- **政务云**上可以部署公共安全管理、容灾备份、城市管理、应急管理、智能交通、社会保障等应用，通过集约化建设、管理和运行，可以实现信息资源整合和政务资源共享，推动政务管理创新，加快向服务型政府转型。
- **教育云**可以有效整合幼儿教育、中小学教育、高等教育以及继续教育等优质教育资源，逐步实现教育信息共享、教育资源共享及教育资源深度挖掘等目标。
- **中小企业云**能够让企业以低廉的成本建立财务、供应链、客户关系等管理应用系统，大大降低企业信息化门槛，迅速提升企业信息化水平，增强企业市场竞争力。
- **医疗云**可以推动医院与医院、医院与社区、医院与急救中心、医院与家庭之间的服务共享，并形成一套全新的医疗健康服务系统，从而有效地提高医疗保健的质量。

5. 云计算产业

- 云计算产业作为战略性新兴产业，近些年得到了迅速发展，形成了成熟的产业链结构，产业涵盖硬件与设备制造、基础设施运营、软件与解决方案供应商、基础设施即服务（IaaS）、平台即服务（PaaS）、软件即服务（SaaS）、终端设备、云安全、云计算交付/咨询/认证等环节。



1. 物联网概念

➤物联网是物物相连的互联网，是互联网的延伸，它利用局部网络或互联网等通信技术把传感器、控制器、机器、人员和物等通过新的方式联在一起，形成人与物、物与物相联，实现信息化和远程管理控制。

7.2 物联网



应用层 智能交通 智能电网 智慧农业 智能工业 智能家居 智慧医疗

业务支撑平台（中间件平台）

处理层

服务支撑
平台

网络管理
平台

信息处理
平台

信息安全
平台

网络层

电信网

互联网

广电网

电网

专用网

其他网

RFID网络

传感器网络

感知层

RFID标签
和读写器

M2M终端

导航定位

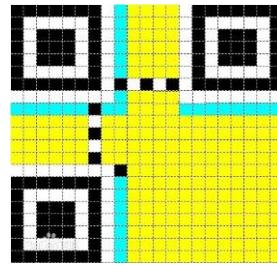
二维码
标签

传感器

摄像头

2. 物联网关键技术

- 物联网中的关键技术包括识别和感知技术（二维码、RFID、传感器等）、网络与通信技术、数据挖掘与融合技术等。



矩阵式二维码



采用RFID芯片的公交卡



(a)温湿度传感器



(b)压力传感器

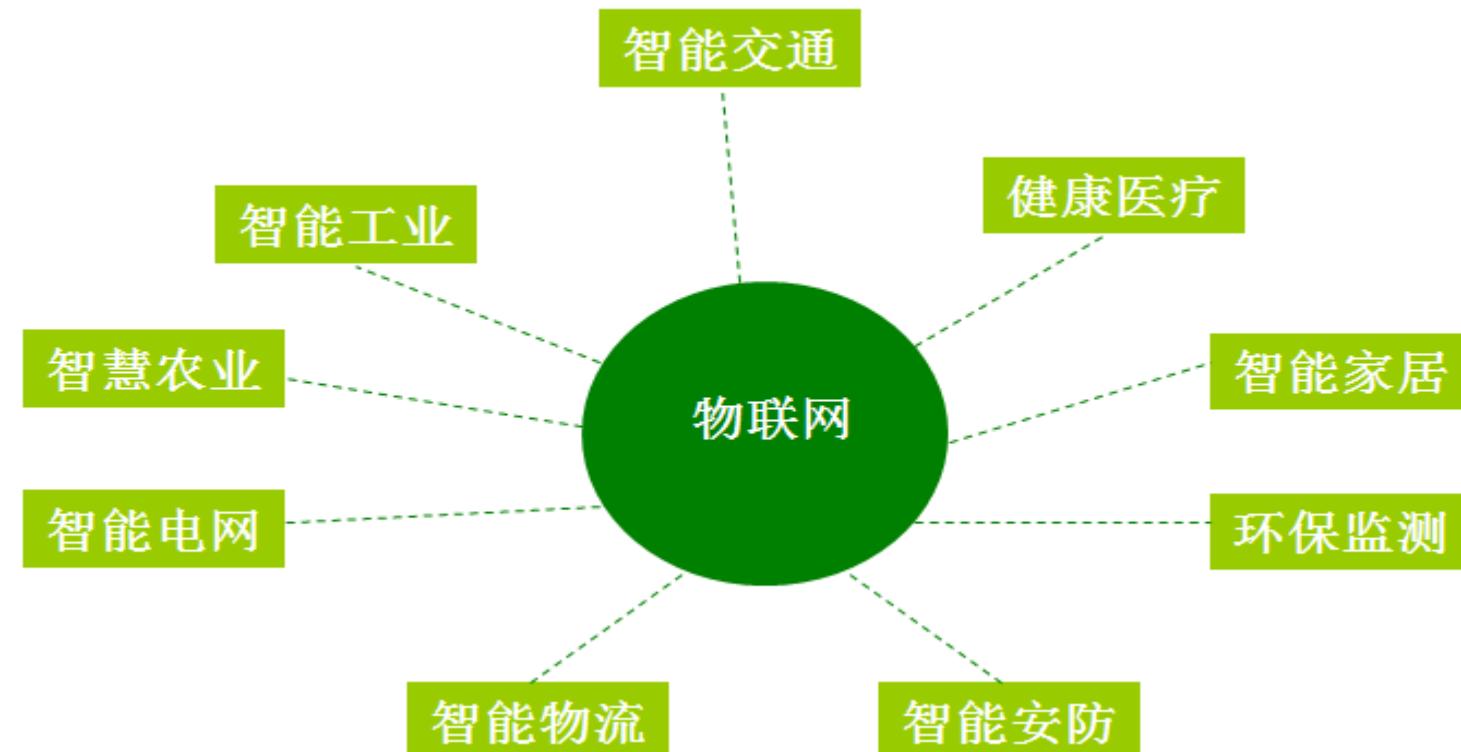


(c)烟雾传感器

不同类型的传感器

3.物联网应用

➤物联网已经广泛应用于智能交通、智慧医疗、智能家居、环保监测、智能安防、智能物流、智能电网、智慧农业、智能工业等领域，对国民经济与社会发展起到了重要的推动作用。

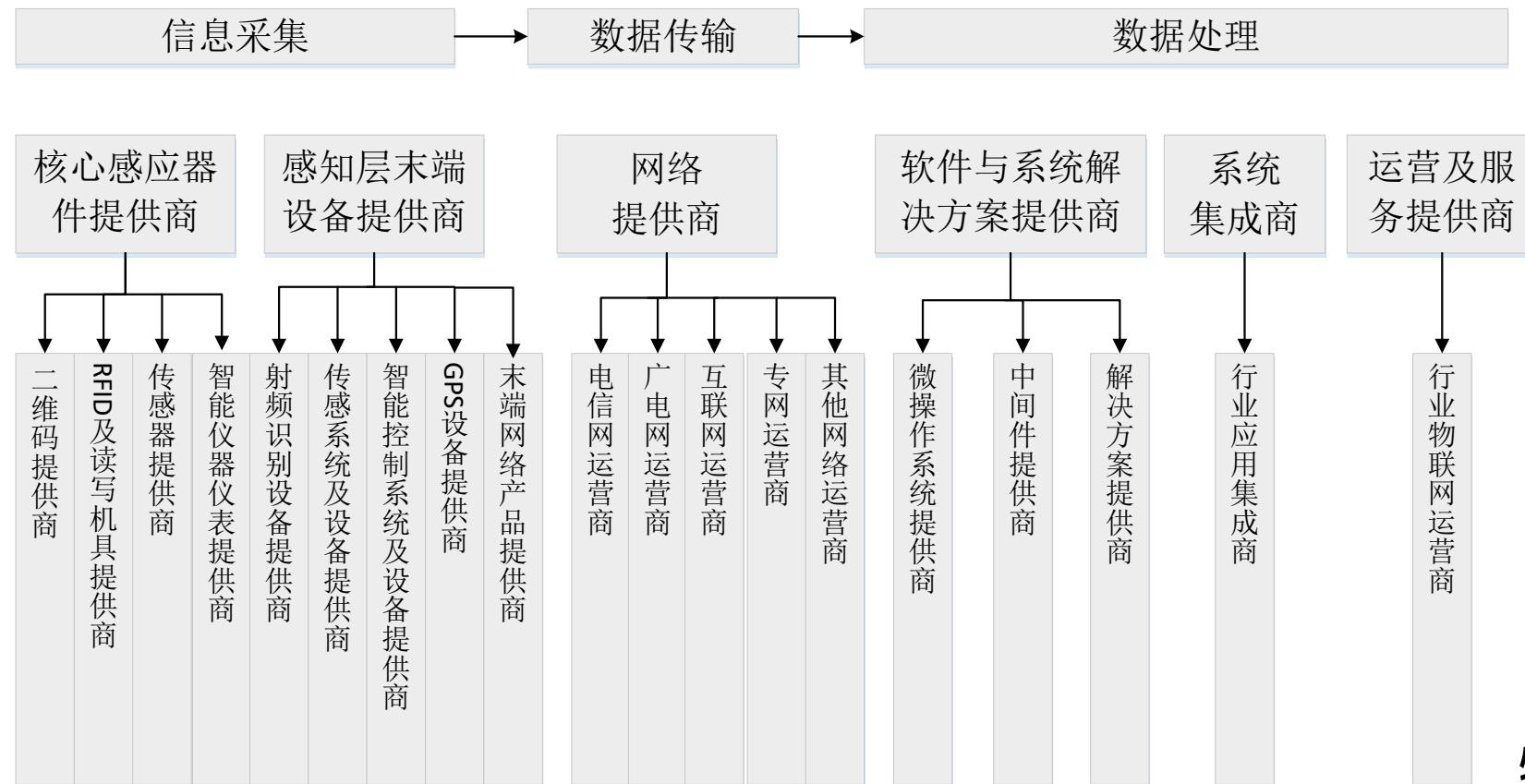


7.2 物联网



4.物联网产业

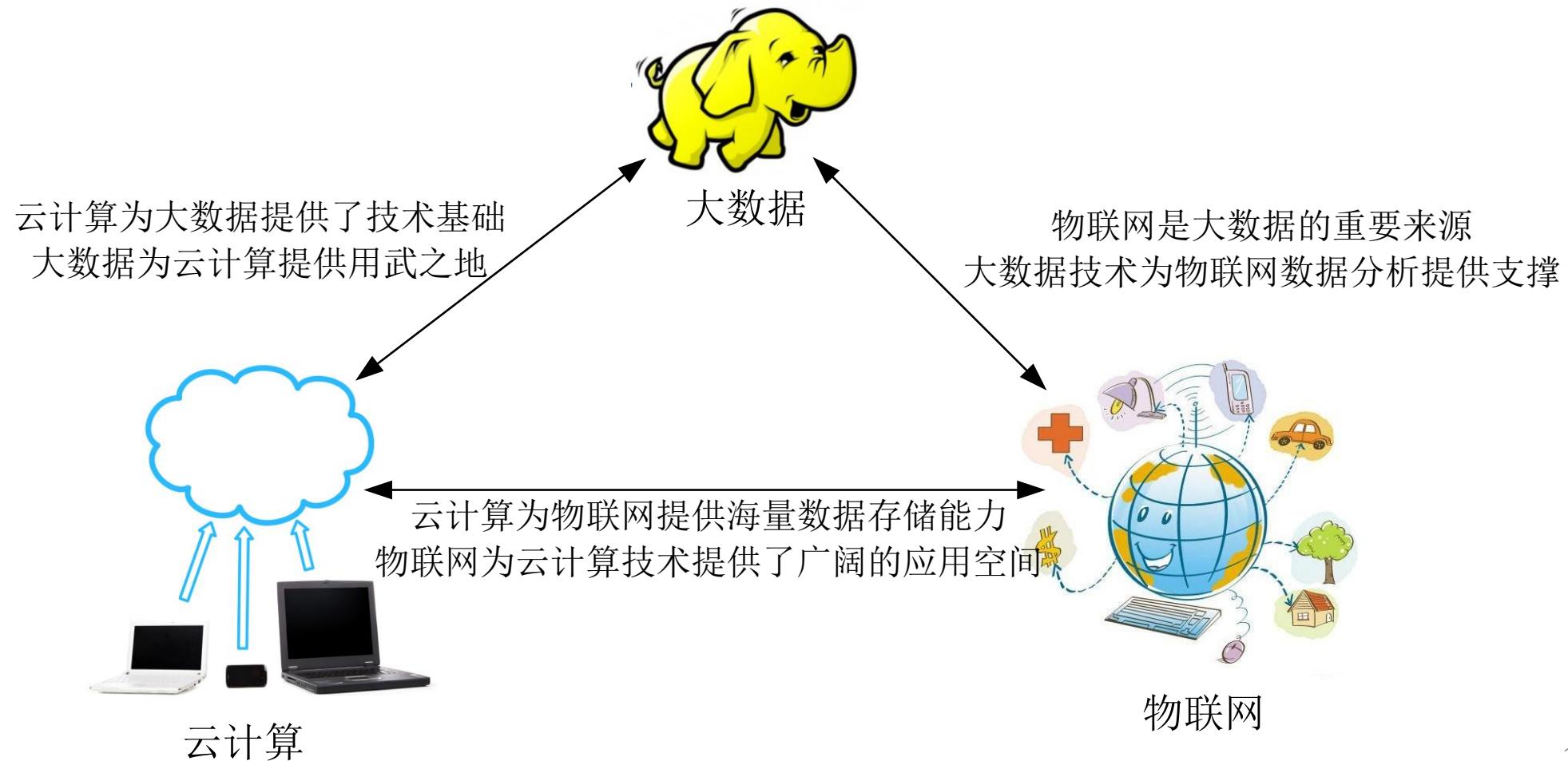
➤完整的物联网产业链主要包括核心感应器件提供商、感知层末端设备提供商、网络提供商、软件与行业解决方案提供商、系统集成商、运营及服务提供商等六大环节。



7.3 大数据与云计算、物联网的关系



- 云计算、大数据和物联网代表了IT领域最新的技术发展趋势，三者既有区别又有联系。



- 数据的本质是生产资料和资产，数据的爆炸式增长带来了数据资产管理的挑战。信息科技的不断进步、数据产生方式的变革、政府的支持等促成了大数据时代的来临。
- 大数据具有数据量大、数据类型繁多、处理速度快、价值密度低等特点，统称“4V”。
- 大数据对思维方式、社会发展、就业市场和人才培养等方面，都产生了重要的影响，深刻理解大数据的这些影响，有助于我们更好把握学习和应用大数据的方向。
- 大数据并非单一的数据或技术，而是数据和大数据技术的综合体。大数据技术主要包括数据采集、数据存储和管理、数据处理与分析、数据安全和隐私保护等几个层面的内容。
- 大数据产业包括IT基础设施层、数据源层、数据管理层、数据分析层、数据平台层和数据应用层，在不同层面，都已经形成了一批引领市场的技术和企业。
- 本章最后介绍了云计算和物联网的概念和关键技术，并阐述了大数据、云计算和物联网三者之间的区别与联系。



Discussion!

姚尧 博士, 副教授, 高级工程师

地理与信息工程学院, 地图制图学与地理信息工程

东京大学, 空间情报科学研究中心, 助教授

阿里巴巴集团, 达摩院, 访问学者

Email: yaoy@cug.edu.cn

办公地点: 未来城校区地信楼522办公室

