

大数据浅论

陈华钧. 教授、博导



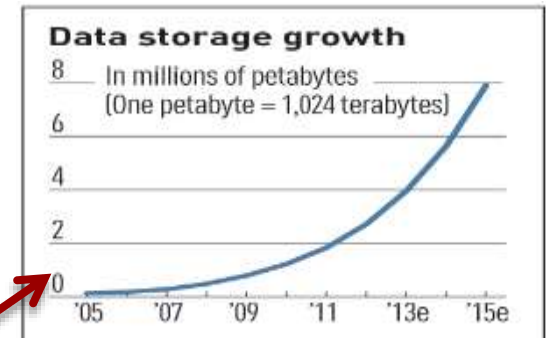
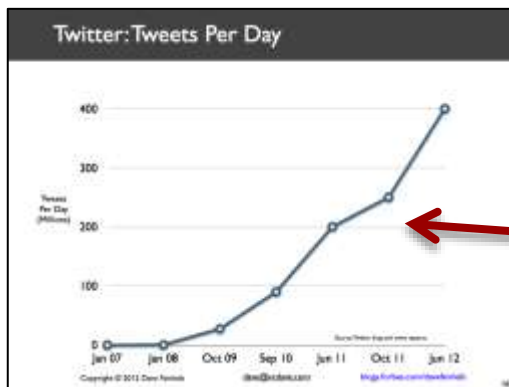
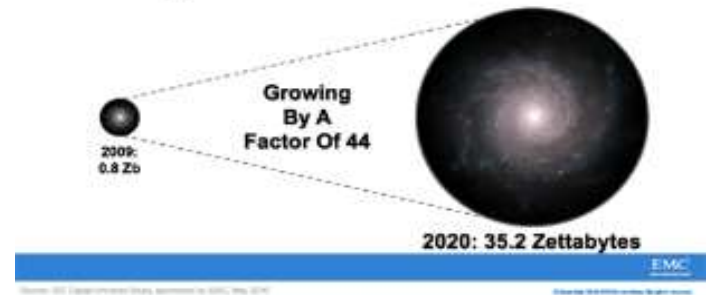
浙江大学计算机科学与技术学院

一、什么是大数据

大数据 (A Big-Data Era.)

- 未来10年，数据将以44x速度飞速增长。
- 数据总量将从0.8Zettabytes增长到35Zettabytes.

The Digital Universe 2009-2020



Exponential increase in collected/generated data

大数据 (A Big-Data Era.)

200万篇博客文章
在网上发布。

每天约有2940亿封
电子邮件发出。



1亿8700万小时的音乐
在Pandora (流媒体音乐网站) 播放。

互联网的一天

2200万小时

用来在Netflix (在线影片租赁商)
上观看以前的电视节目和电影。

1288个新应用
可供下载

超过3500万个应用被下载。



5亿3200万条
状态更新。

2亿5000万张照片
上传到Facebook。

如果把它们都印出来,
叠起来能有80个
埃菲尔铁塔那么高。

一天内, 整个因特网的流量信息可以装满
1亿6800万张DVD光盘。

86万4000小时的视频
上传到YouTube上。



大数据与人工智能：深度学习

Google 大脑 “自主” 的从海量数据中识别猫脸的特征

由16000个处理器连接而成，模拟共有10亿个节点的神经网络



人脸识别



语音识别



无人车



大数据与人工智能：

知识图谱-Knowledge Graph

谷歌知识图谱： Things, not Strings



相关概念： 云计算

- (1) 数据在云端
 - 不怕丢失
 - 不必备份
 - 海量存储
- (2) 软件和服务在云端
 - 不必下载
 - 自动升级
- (3) 无所不在的云计算
 - 浏览器即客户端
 - 任何设备登录
 - 随时随地从云中获取数据
- (4) 无限强大的云计算
 - 无限空间
 - 无限速度

Google docs



Google 云计算

MapReduce

BigTable

Chubby

GFS

什么是大数据：以雾霾为例

产生来源

处理方法

应用价值

人群产生的大数据

新浪微博峰值超3万条每秒



传感器产生的大数据

全国数千个空气监测站每小时产生各种各样的空气质量数据



空气



天气



交通

高端设备产生的大数据

单张遥感卫星影像图片可达400GB



数据采集

数据存储

数据传输

计算模型



公众健康信息服务

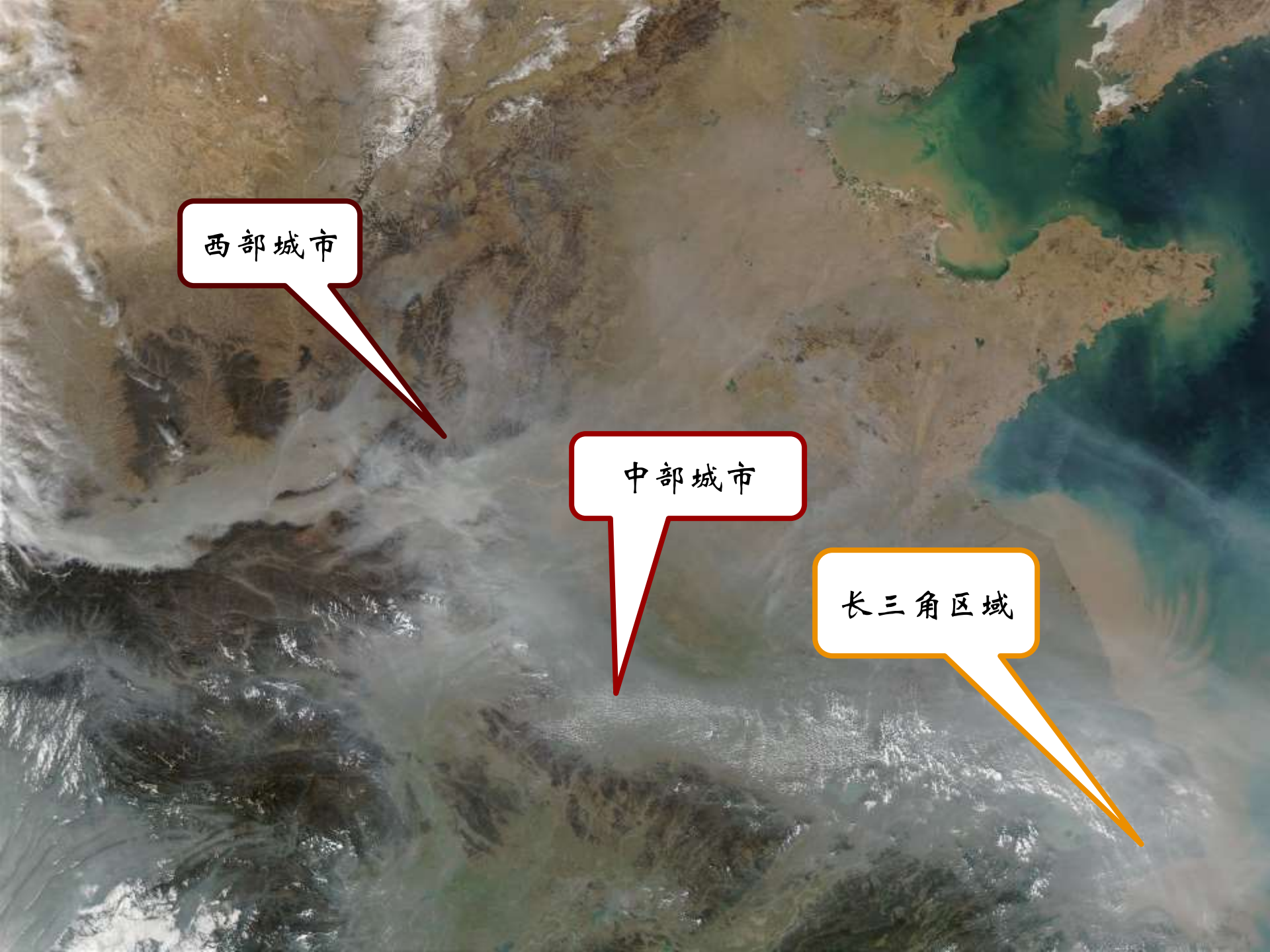


雾霾成因及演化规律



雾霾预测与治霾决策





西部城市

中部城市

长三角区域

什么是大数据： 微博就能告诉你城市内涝点



内涝查询

输入检索关键字

搜索

杭州

订阅通知

统计图

普通地图

隐藏列表



大样本、小模型

中国杭州市江干区学源街新业北路口

东信大道涵洞

中国浙江省杭州市滨江区东信大道868号

闻涛路一桥涵洞

中国浙江省杭州市滨江区闻涛路

时代大道铁路涵洞

中国浙江省杭州市萧山区时代大道3939

浦沿路竖塔路口

中国杭州市滨江区浦沿路竖塔路口

浦沿路东冠路口

中国浙江省杭州市滨江区浦沿路124号

浦沿路东信大道

中国浙江省杭州市滨江区东信大道688号



什么是大数据： Big是个相对概念

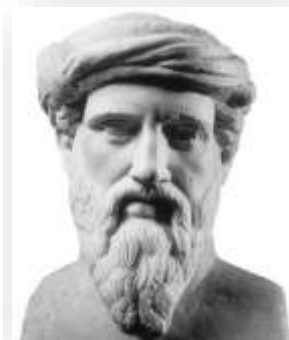
- ▶ 是指规模在 **P级** (10^{15})-**E级** (10^{18})-**Z级** (10^{21}) 的极大规模数据处理或或 **Extreme-Scale Computing** (极限级计算) ;
- ▶ 从IT技术的角度讲, 特指传统**存储、数据库、并行计算、数据挖掘**等技术无法有效处理的极大规模数据计算;
- ▶ 又有称为**Big Enough Computing**-**相对大、足够大**的计算;
 - ▶ 数据量的相对性, 如样本量巨大, 但数据体积未必大;
 - ▶ 计算设备能力的相对性: 如**移动设备上的T级数据处理等, 内存级的T级数据处理等。**



什么是大数据： 大数据的哲学观

▶ 哲学家毕达哥拉斯：“数是万物的本原”：

▶ 事务的本质和规律隐藏在各种原始数据的相互关联中



- ▶ 数据不仅可以描述客观世界，也被用于刻画人类精神世界和人类社会，形成个了数据化的世界；
- ▶ 大数据通过“量化一切”而实现世界的数字化，可能改变人类认知和理解世界的方式，带来全新的大数据世界观。

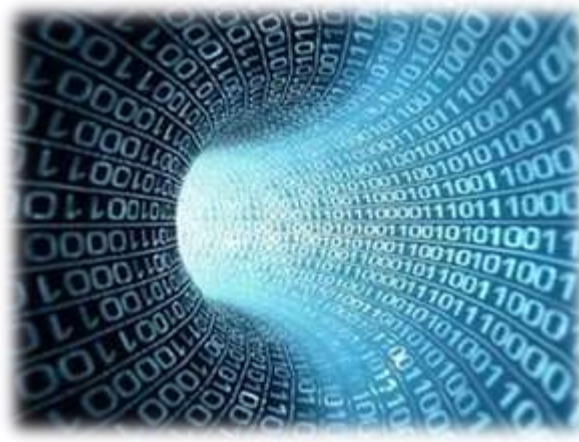
摘引自-李国杰院士 《对大数据的再认识》

什么是大数据： 数据界

- ▶ 有些学者试图将“数据”当成一个“**自然体**”来研究，相对于自然界或物理界，提出“**数据界**”的概念。



自然界



数据界

什么大数据： 大数据的科学观

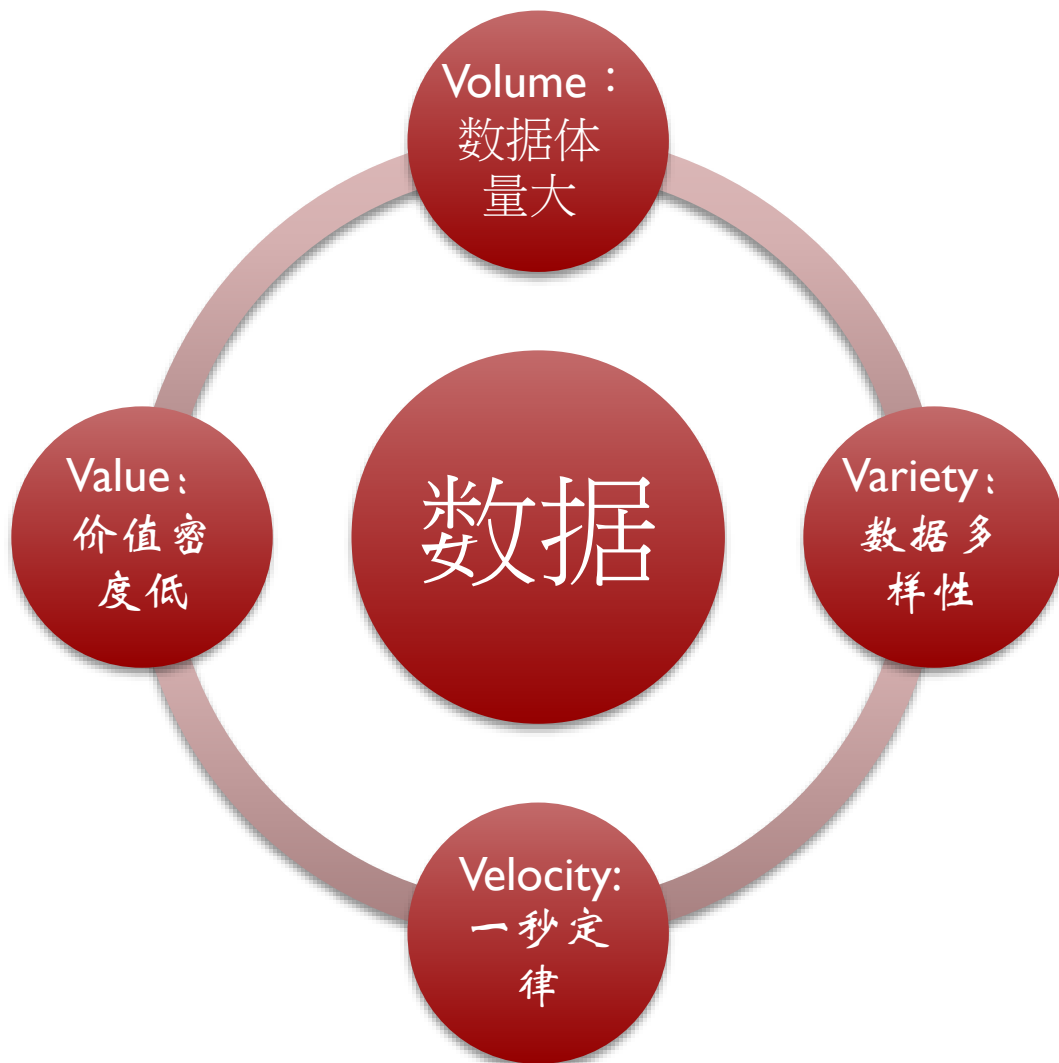
- ▶ 伽利略和牛顿时代：科学始于观察
- ▶ 德国哲学家波普尔提出：科学始于问题
- ▶ 大数据兴起引发新的科学研究模式：科学始于数据
- ▶ 数据驱动的计算分析已经成为继理论推导、实验分析之后的第三种重要的科学研究的方法。



摘引自-李国杰院士 《对大数据的再认识》

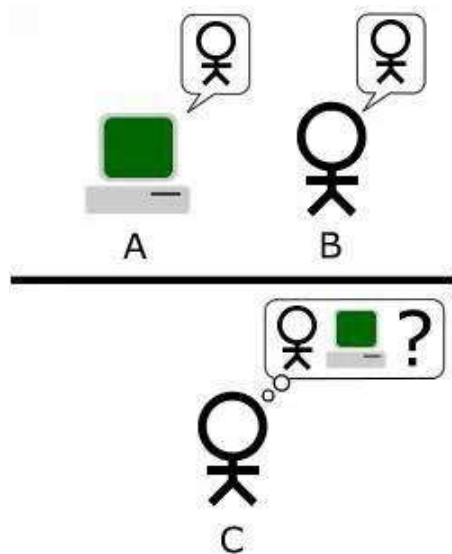


什么是大数据： 大数据的典型特征-4V



什么是大数据：量的重要性

图灵测试



- ▶ 2005年，谷歌首次参加NIST的全球机器翻译竞赛即获得第一名。
- ▶ 谷歌系统和评测排第二名的系统都是由著名机器翻译专家Franz Och研发。但唯一的差异是谷歌采用多于第二名近万倍的数据。

模型驱动的
机器智能



数据量驱动的
机器智能

什么是大数据：量的重要性

- ▶ 谷歌的无人驾驶汽车依靠“扫街数据”，而非仅依靠视觉算法去做临时性的“目标识别”，把无人驾驶真正变成了现实。
- ▶ IBM深蓝计算机击败世界象棋冠军——搜集全所有棋局模式
- ▶ IBM超级电脑沃森2013年亮相美国最受欢迎的智力竞赛节目《危险边缘》战胜该节目两位最成功的选手——搜集全所有可能的答案并叠加概率评价模型

如果数据量足够大且完备，凭借强大的计算能力，
依靠机器做出的决策完全可能超过人的判断



什么是大数据：数据的完备性

- ▶ 传统以盖洛普为代表统计学家认为，只需要有足够代表性的样本就可以客观的预测全体。但却从来没有准确预测过美国总统的选举。
- ▶ 2012年，一个名不见经传的统计学家Nate Silver搜集了互联网上各种数据（包括社交媒体、新闻等），通过大量数据分析，准确预测了全部50个州的选举结果。



什么是大数据： 数据的外部性

- ▶ 在经济学中，所谓“外部性”是一个人的行为对旁观者福利的影响。广义上理解，“外部性”是指信息对直接相关方以外的影响。
- ▶ 数据外部性的典型例子：
 - ▶ 电力消费数据可以反映住房空置率；
 - ▶ 出租车的轨迹数据可以反映交通拥堵情况；
 - ▶ 淘宝的交易数据可以反映商户的信用程度——大数据征信；

大数据的重大价值可能更多的在于挖掘出数据的外部性特征

部分摘引自-李国杰院士 《对大数据的再认识》

二、大数据典型应用

互联网大数据一： 大数据搜索引擎

- ▶ 搜索引擎本身就是典型的大数据系统，现有的主流大数据新技术（云存储、知识图谱、深度学习等）都最早源于搜索引擎。
- ▶ **大数据索引**：一个简单的搜索请求可能触发后台数以千计的服务器处理；
- ▶ **搜索行为挖掘与分析**：搜索历史、搜索偏好、搜索预测（如Google Trends）；
- ▶ **大数据竞价排名**：商家、搜索引擎公司和用户三者之间的博弈。



互联网大数据二： 电商大数据

- ▶ **电商**是目前大数据应用最广的领域之一：
 - ▶ **智能推荐**：通过发掘用户行为和背景信息，如购买历史、浏览历史、教育背景、经济实力等，向用户推荐可能感兴趣的商品，典型的推荐算法如协同过滤。
 - ▶ **精准营销**：通过聚类、分类等算法对人群进行自动划分，以实现更加精准的精细化营销。
 - ▶ **广告优化**：通过计算历史广告的效果，利用大数据对广告投放进行优化。



互联网大数据三： 社交媒体大数据

- ▶ 典型互联网社交平台包括：即时通信、社交网站、微博、微信等。通过对社交网络中的大数据进行分析，可以了解用户的思维习惯及其对社会的认知。
- ▶ 对微博等社交网络信息空间大数据的挖掘能够及时反映经济社会动态与情绪，预警重大、突发和敏感事件（如流行疾病爆发、群体异常行为等），协助提高社会公共服务的应对能力。

每个人都在消费
自己的隐私

Facebook Likes data

完全公开的点赞数据记录用户对照片、朋友转发的文章、新闻、餐厅、体育、音乐、电影、书籍等的点评等。

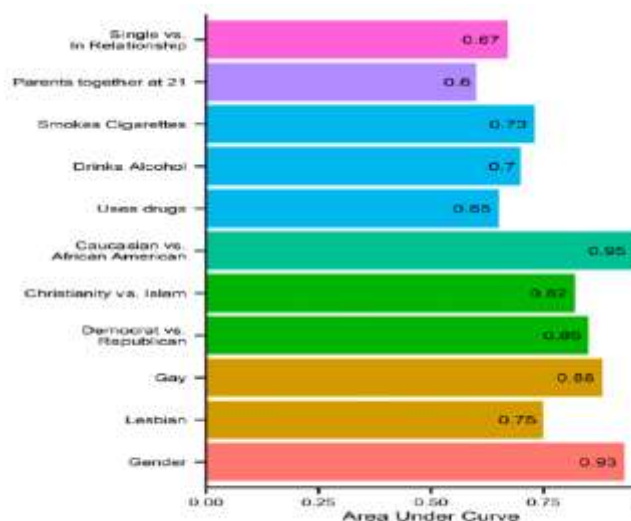
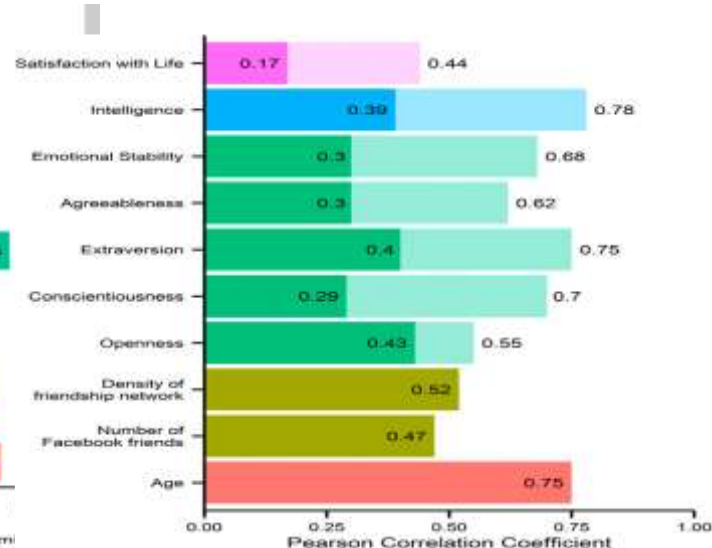


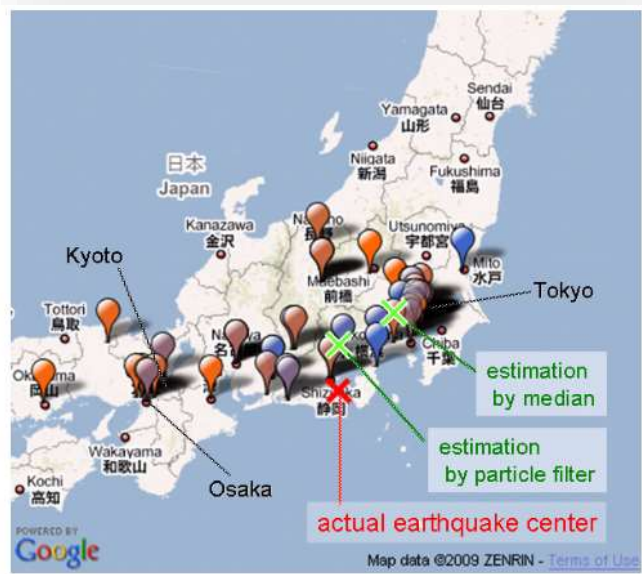
Fig. 2. Prediction accuracy of classification for dichotomous/dichotomous attributes expressed by the AUC.



--- Michal Kosinski, David Stillwell(Cambridge University). Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy of Sciences (PNAS, 2013)

社交媒体大数据：Twitter预警地震

- ▶ 日本东京大学的学者利用地震前后Twitter上面人群对地震的反应对地震进行预警和分析。
- ▶ 每次地震平均在**10分钟**内产生至少10条以上的Tweets。系统最快震后**20秒**，平均**1分钟**即可发出邮件进行预警，而日本官方的地震预警广播系统平均要震后**6分钟**以上才能发出预警。
- ▶ 并且基于Twitter的方法在地震严重程度上的分析准确性甚至超过了传统的地震监测系统。



-----Takeshi Sakaki The University of Tokyo. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. WWW2010.

互联网大数据四：移动互联网大数据

- ▶ 大量的移动APP联网的都在自动记录用户的位置和轨迹，如CheckIn数据、出租车轨迹等。移动互联网应用所产生的大数据具有极强的**时空性**和**社会性**。
- ▶ **基于位置的精准营销**： Geo-Fencing、基于位置的广告推送……
- ▶ **城市人群迁徙分析**： Human Mobility, Geo-Social Analytics.
- ▶ **商业地理分析**： 选址分析、销售区域分配、配送路径优化、潜在消费者空间分布、城市规划等。



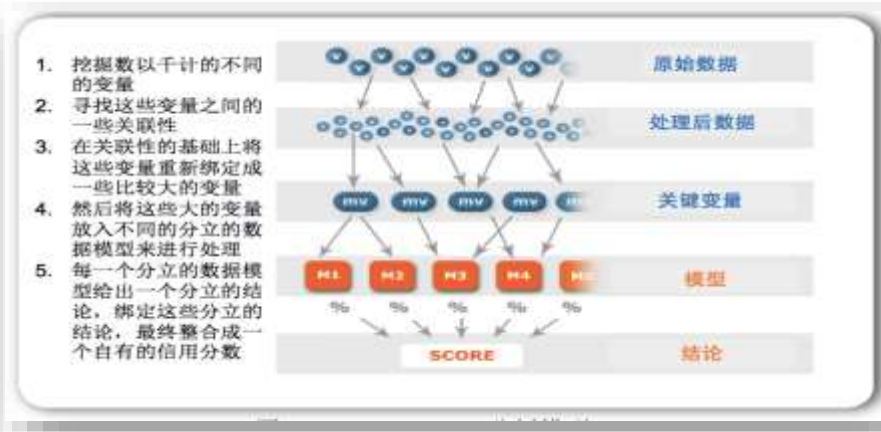
金融大数据

- ▶ 金融是大数据技术应用非常活跃的领域，典型的主要包含两个方面：
 - ▶ 一是利用大数据的平台技术改造传统的金融信息基础设施，如利用Hadoop平台提升数据存储与处理能力的可扩展性等；
 - ▶ 二是利用大数据分析手段进行金融数据挖掘与分析，典型数据分析应用包括：
 - ▶ 传统金融数据挖掘
 - 银行客户流失率分析与预警
 - 恶意账户检测
 - ▶ 新兴金融大数据应用
 - 大数据征信
 - 大数据投资



金融大数据一： 大数据信贷-阿里小贷

- ▶ 阿里小贷关注如何透过数据去判断小微企业的信用等级、未来经营走向和融资需求；
- ▶ 阿里小贷的模型体系涵盖贷前、贷中、后管理、反欺诈、市场分析、信用体系和创新研究六大部分
- ▶ 上百个基于电商数据建立的模型交叉生效，构成整个风险决策体系核心——“水纹模型”。



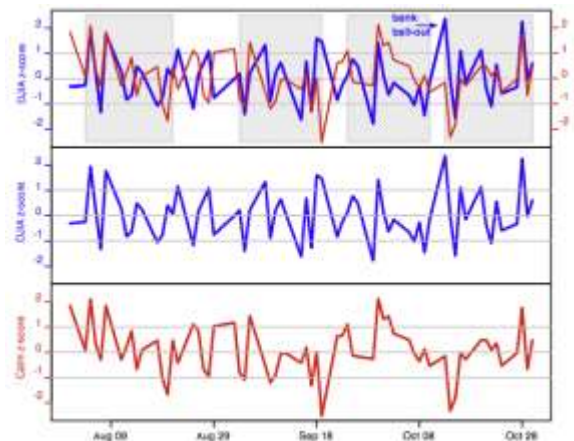
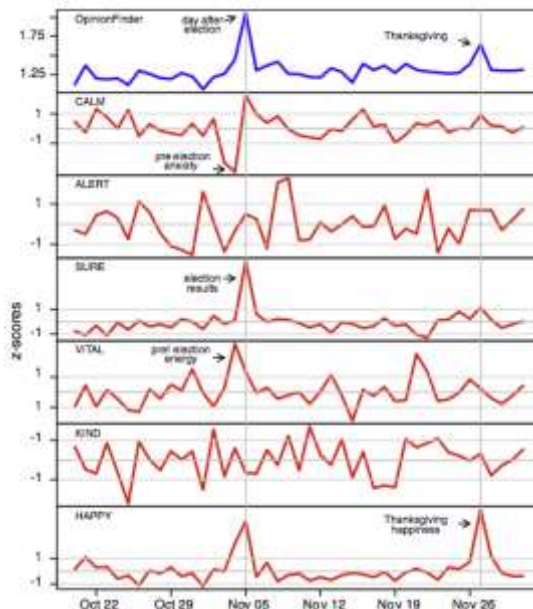
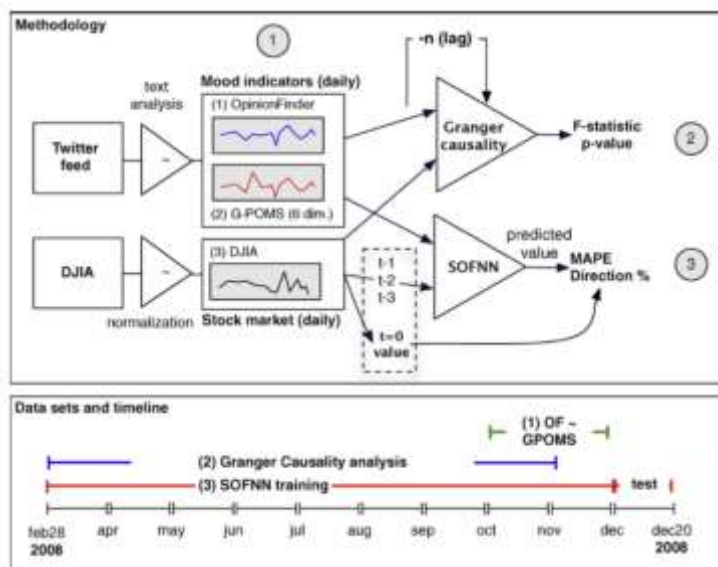
金融大数据二： 大数据征信-芝麻信用

用大数据取代所有的信用卡业务？甚至开发出更多的基于信用的衍生业务。



金融大数据四： 大数据投资-股票市场预期

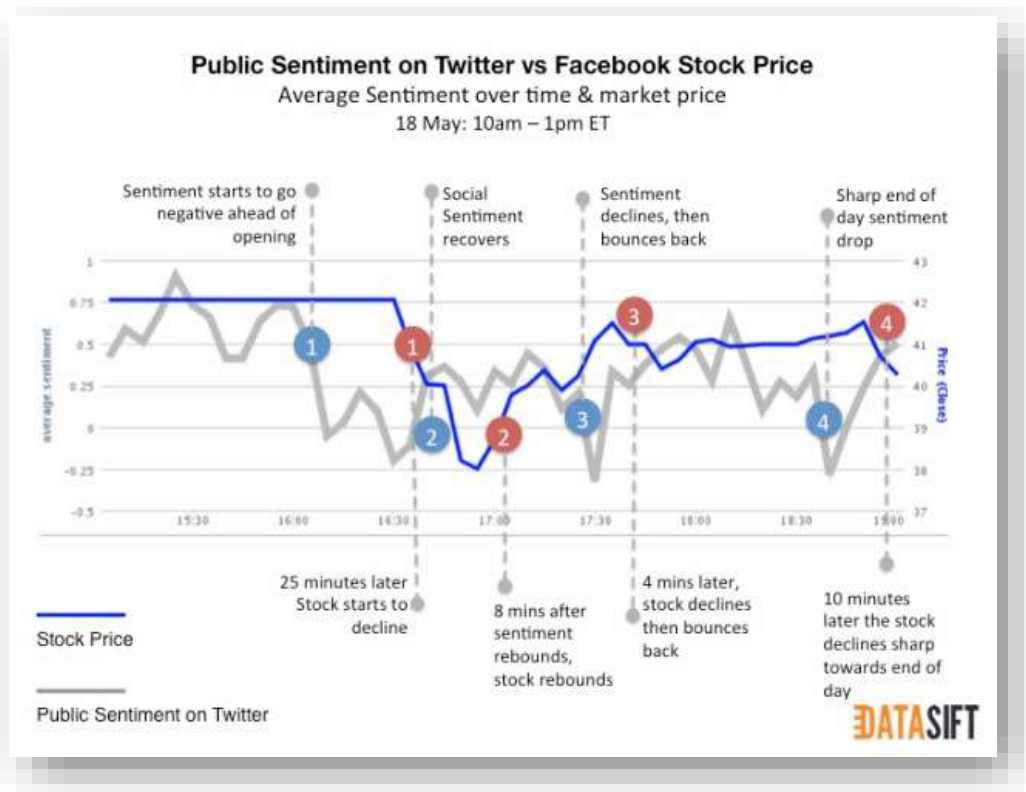
- 美国印第安纳大学Johan Bollen将Twitter上的文章分为冷静、警惕、确信、活力、友善和幸福这六个心情类别。其中，如果将“冷静”情绪指数后移3天，其曲线与道琼斯工业平均指数能够高度契合。



Johan Bollen and Huina Mao. Twitter Mood as a Stock Market Predictor. IEEE Computer, 44(10): 91-94, October 2011.

金融大数据五： 大数据投资-股票市场预期

- ▶ 社交媒体监测平台Datasift在Facebook首次公开募股当天收集了58665 位用户产生的95019 条关于facebook IPO的博文，并对每条博文的情感倾向进行定义和分析。
- ▶ 结论显示，Twitter讨论中情感倾向的变化们都会会在5-20分钟内反映到Facebook的股价波动上。



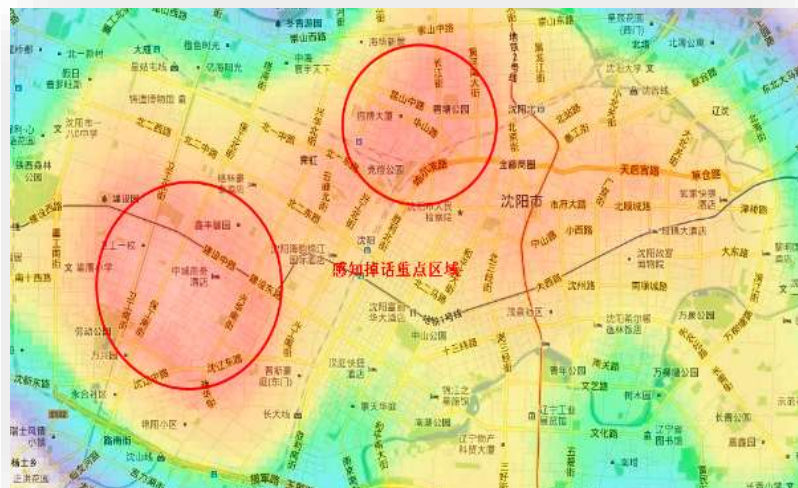
Study: Twitter Sentiment Mirrored Facebook's Stock Price Today. DataSift.com.

<http://techcrunch.com/2012/05/18/study-twitter-sentiment-mirrored-facebooks-stock-price-today/>

电信大数据一：

基于大数据的运营商网络管理与优化

- ▶ 电信运营商所拥有的电信运营网络是其核心能力，如何利用大数据技术进行网络管理和优化成为研究热点。
- ▶ 运营商通过分析话单和信令中用户的流量在时间周期和位置特征方面的分布，对2G、3G的高流量区域设计4G基站和WLAN热点，实现基站和热点的选址以及资源的分配
- ▶ 如法国电信通过分析发现某段网络上的掉话率持续过高，借助大数据手段诊断出通话中断产生的原因是网络负荷过重造成，并根据分析结果优化网络布局。



电信大数据二：

电信行业的大数据精准营销

客户画像：基于**客户终端信息、位置信息、通话行为、手机上网行为轨迹**等丰富的数据，为每个客户打上人口统计学特征、消费行为、上网行为和兴趣爱好标签，并借助数据挖掘技术（如分类、聚类、RFM等）进行客户分群，帮助运营商深入了解客户行为偏好和需求特征。



社交圈研究：通过分析**客户通讯录、通话行为、网络社交行以及客户资料**等数据，开展交往圈分析。并进一步利用图挖掘的方法来发现各种圈子，发现圈子中的关键人员，以及识别家庭和政企客户；或者分析社交圈子寻找营销机会。

精准营销和实时营销：运营商在客户画像的基础上对客户特征的深入理解，建立客户与业务、资费套餐、终端类型、在用网络的精准匹配，并在在推送渠道、推送时机、推送方式上满足客户的需求，实现精准营销。



电信大数据三：

电信大数据的外部性应用与数据变现

- ▶ 数据变现指通过挖掘企业自身数据的外部性，并发掘增值服务，获取收益。比如：西班牙电信成立了名为“动态洞察”的大数据业务部门，推出的“智慧足迹”，基于完全匿名和聚合的移动网络数据，可对某个时段、某个地点人流量的关键影响因素进行分析，并将洞察结果面向政企客户提供。



人の移動傾向を把握



店舗ごとの商圈を把握

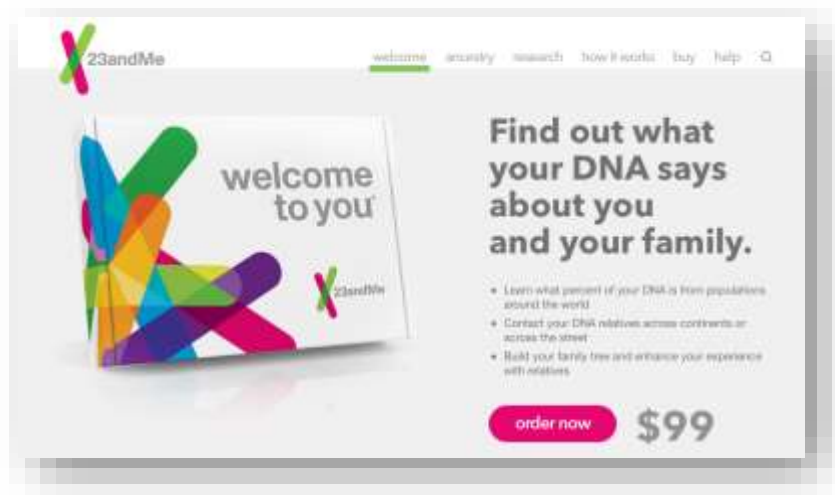
健康医疗大数据

- ▶ 一方面，利用大数据平台技术优化改造、提升并优化传统的医疗信息系统，如利用云存储改造传统的电子病历、医学影像数据库系统等，利用数据挖掘技术辅助进行医学诊断与护理等；
 - ▶ 另外一方面，又涌现出大量新兴的健康大数据应用
 - ▶ 基因大数据
 - ▶ 人体微生物大数据
 - ▶ 基于互联网大数据的疾病预测与公共卫生应用
 - ▶ 结合穿戴计算和移动计算的健康大数据应用
-



健康医疗大数据一： 个人基因大数据

- ▶ 基因测序技术的飞速进步使得普通人花费很少就可以做完个人全部的基因检测，目前价格在1万元左右，全部数据量约数百GB。
- ▶ 当前，个人基因检测市场正在飞速兴起，涌现不少初创公司，特别在癌症早期筛查与靶向治疗、新生儿遗传疾病筛选等方面市场异常广阔。
- ▶ 安吉丽娜朱莉进行乳腺癌基因检测准确吗？医生评估她有 87% 的患癌几率。



23andMe



壹基因

健康医疗大数据二：人体微生物大数据



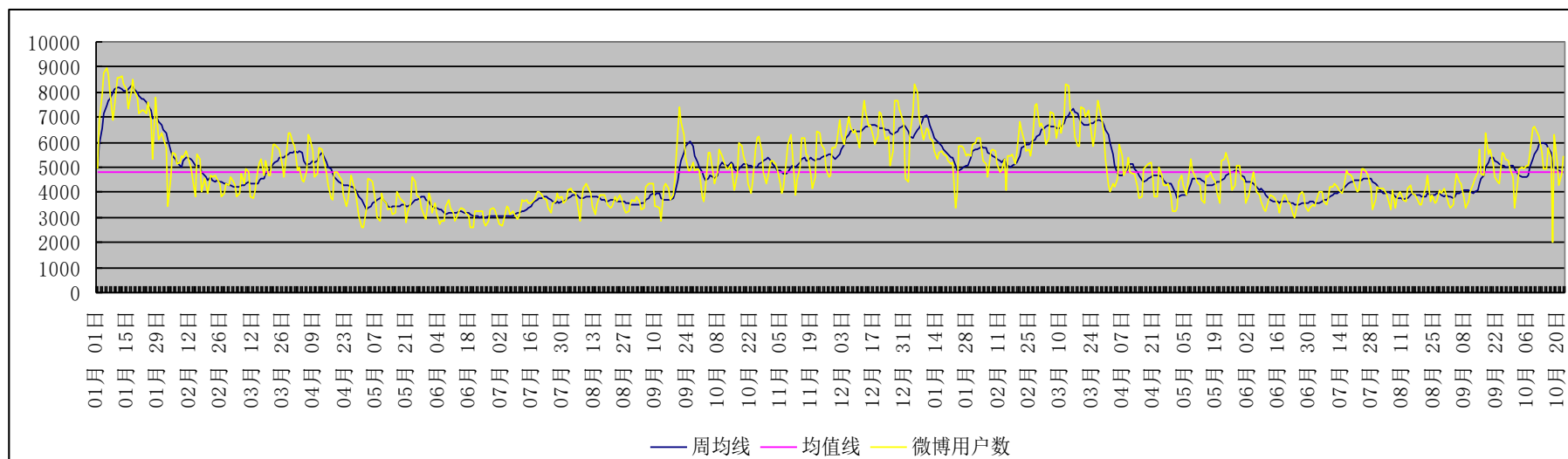
- ▶ uBiome提供基于人体微生物大数据的个人健康检测与咨询服务。
- ▶ 人体微生物是个人健康的主要影响因素，人体中的微生物是人体细胞的数十倍，是各种消化系统疾病、神经系统疾病的主要来源。

- ▶ uBiome提供一个简单kit供用户自我采集人体样本。样本寄回公司将通过测序分析后上传到云端供用户查看。
- ▶ 用户可以将自己的样本与其他健康、素食者、饮酒者的信息进行比较，以了解自己的健康状况。



健康医疗大数据三：互联网大数据与公共卫生

- ▶ Google 2009 年对甲型H1N1流感爆发的预测比美国疾病控制与预防中心（CDC）提早1-2周，研究报告发表在《Nature》上。
- ▶ 百度公司2014年7月上线了“百度疾病预测”借助最新大数据技术，为用户呈现身边的疾病信息，不仅可以了解当前流行病态势，还可以看到未来7天的变化趋势，提前做好预防措施。

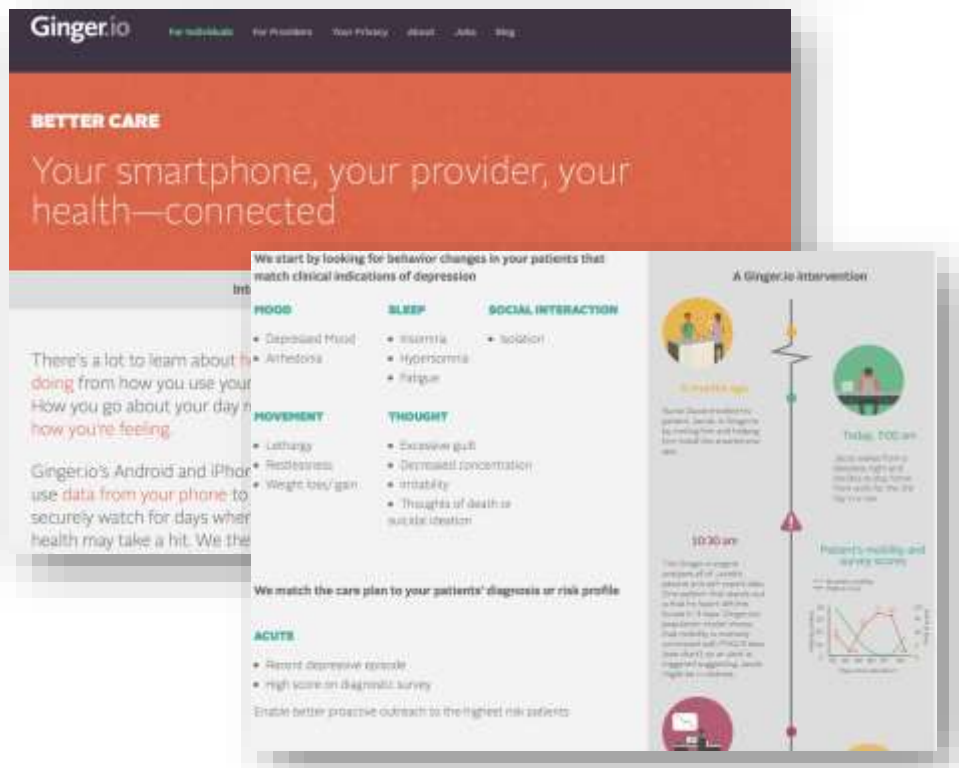


健康医疗大数据四：

结合穿戴计算与移动计算的健康大数据应用



- ▶ Giner.IO主要主要针对抑郁症、精神病患者提供基于手机和移动设备的个人健康健康和护理功能；
- ▶ 提供主动与被动两种方式收集病人状态信息，并实时预警报送给医生；被动方式是通过传感器检测或行为分析，主动为用户回答若干简单的问题。
- ▶ 用户很多时候无法准确或容易忘记自己的感受，提供随时随处回答若干简单的问题，持续记录病痛感受，使得医生可以更加准确的了解病人的状态。



电力大数据一：

基于大数据的电力基础设施优化

- ▶ 大数据技术可用于优化电力基础设施的选址与建设决策。
- ▶ 例如丹麦风电公司VESTAS在公司**发电机历史数据**的基础之上，综合**气温、气压、空气湿度、空气沉淀物、风向、风速等气象数据**，来优化其风力发电机的选址，以充分利用风速、风力、气流等因素达到最大发电量，并减少能源成本。
- ▶ 其它还将利用数据有：**全球森林砍伐追踪图、卫星图像、地理基础数据**以及月相与潮汐数据等。



电力大数据二：

基于大数据的电力系统智能控制与优化

- ▶ 在电力系统的生产端，随着**智能电网**深入发展，通过为电力基础设施部署更多的传感器构建**电力互联网**，动态监控各个环节运行状况。
- ▶ 通过**电力负荷数据**实现更加科学合理电力调度，综合**电能质量监测**、**雷电监测**、**覆冰监测**等数据实现更加实时和准确的故障监测和风险预警，从而有效改进电力公司的运维方式，并促进智能电网的发展。



电力大数据三：

基于大数据的电力运营分析

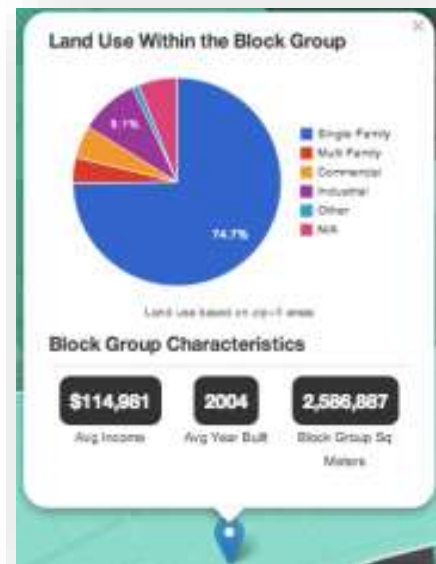
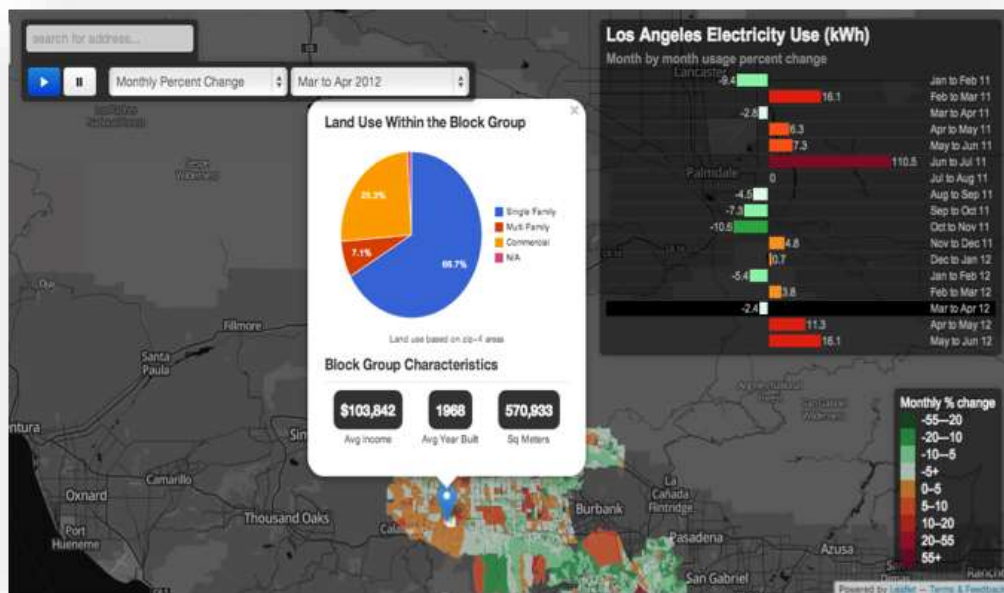
- ▶ 在电力的使用端，通过挖掘**用户用电行为特点**与电费计价、天气、温度、空气质量、交通等因素之间的潜在关联关系，建立从**用户到建筑、社区、乃至整个城市**的用电需求预测模型。
- ▶ 一方面为电力提供部门进行**配用电和电力营销决策**提供支持，另外一方面也为引导**用户合理用电，城市节能减排**提供科学合理的引导。



电力大数据四：

电力大数据的外部性应用-洛杉矶电力消费地图

- ▶ 2013年，美国加州大学洛杉矶分校的研究者综合人口普查信息、用户实时用电信息、基础地理数据、气象等数据，设计了一款“**电力消费地图**”（Electricity Consumption Map）。
- ▶ 该地图提供粒度到街区各时刻的用电量，并通过**建立用电量与人的平均收入、建筑类型等信息的关联模型**，从而更准确地反应该区经济状况及各群体的行为习惯，以辅助投资者进行决策，也可为城市规划提供基础依据。



<http://sustainablecommunities.environment.ucla.edu/maproom/>

三、大数据产业

大数据带来IT技术产业变革

信息技术产业

IT



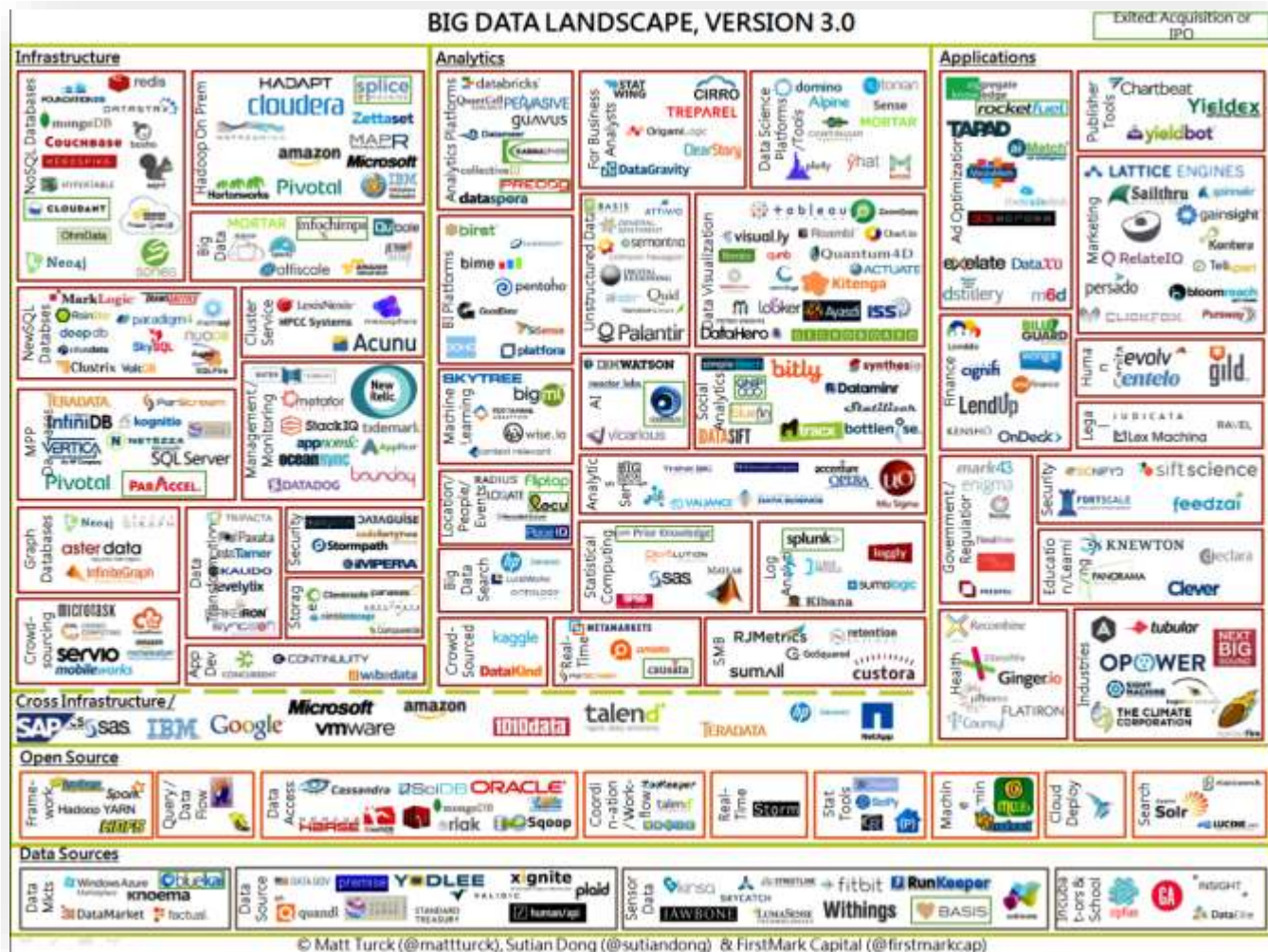
DT

数据技术产业

大数据产业链



大数据产业链发展趋势全景图 (国外版)



商业模式一：数据即商品模式

▶ 数据采集：

- ▶ 开放数据：搜集和整理各种开放数据，提供高质量的开放数据服务
- ▶ 数据众包：采用众包模式按需定制数据采集
- ▶ 聚小成大：购买小型数据源，汇聚成大数据，再提供增值服务

▶ 数据交易：

- ▶ 数据超市：提供数据发布下载的公共服务平台，按数据下载量收费。
- ▶ 数据管道：提供高质量的数据API，按API访问量收费。



数据即商品：微软Azure-数据交易平台



Data services

Find a wide variety of data including demographics, environment, financial, retail, and sports. Use this data in your Microsoft Office software, BI tools, and your very own custom applications.

[View data services ►](#)



Machine Learning Marketplace

You don't have to be a data scientist to take advantage of Azure Machine Learning. Bring your own data and leverage a variety of machine learning services such as forecasting, product recommendations and sentiment analysis.

[View machine learning services ►](#)



Virtual Machines

Access Microsoft and partner solutions configured to run in Azure Virtual Machines. Deploy with confidence knowing that each solution has been vetted through Microsoft Azure Certified for Virtual Machines.

[View Azure Virtual Machines ►](#)



Web applications

Explore and install popular Open source community web applications ranging from blogging engines, photo galleries to eCommerce solutions. Build dynamic Websites in just minutes using these ready to use applications.

[View web applications ►](#)



Azure Active Directory applications

Configuring Single Sign-On to many different SaaS applications of various vendors can be a difficult and demanding task. Azure Active Directory simplifies the process by providing the most popular SaaS applications pre-integrated and ready to use.

[View Azure Active Directory applications ►](#)



Application services

Discover, purchase, and provision application and data services from Microsoft partners. These services can be combined with Azure services from Microsoft to build powerful cloud solutions. Services provisioned through the Azure Store can be managed through the Azure management portal and service usage will be included on a single bill from Microsoft.

[View application services ►](#)

- ▶ 提供数据的发布、搜索、下载、数据API等多种方式数据交易服务；数据按下载量或访问量收费。
- ▶ 数据服务构架在微软自身的云平台之上，捆绑虚拟机、云存储、机器学习等云服务。

数据即商品：数据堂-数据交易平台



- ▶ 提供语音、健康、交通、电商等多种数据的下载服务；
- ▶ “数据+”以数据管道模式提供服务；
- ▶ 提供数据定制和数据代加工服务。

数据即商品：Konema-数据众包

数据网络



60

国家



140

位置



100

收集人



\$4500

已支付

收集您所在国家的数据，赚取 每月120美元



Collect food prices data in your country and earn up to \$120 every month. We are looking for data collectors who will go to the specific markets weekly, collect data on food prices for about 25 items and submit them into our system.

加入



Collect the cost of Consulting Services, diagnostics services and clinical procedures. If you are residing in Major cities/towns where you have both Public and Private Healthcare services, you can join this project and earn money.

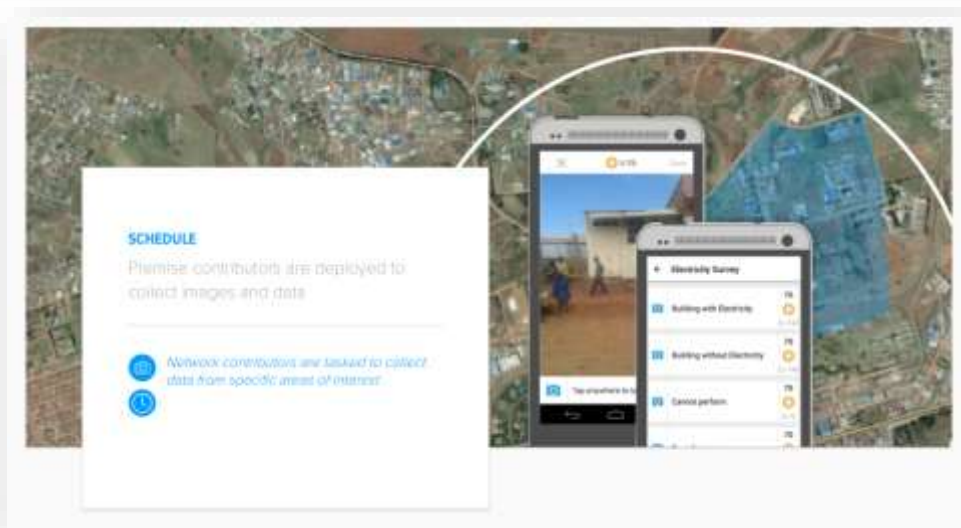
加入

- ▶ 通过互联网向全球发布付费的数据采集需求；
- ▶ 通过众包模式搜集和汇聚各个国家的数据。

数据即商品：Premise-数据众包



- ▶ Premise希望提供增强的地图服务，提供真实、准确、粒度分布广泛的各种数据，方便市场规划、经济研究等人员直观的通过地图查看和分析数据。
- ▶ Premise采用众包的方式采集数据，任何人都可以通过手机程序按照Premise规定的方式采集数据，并获得回报。
- ▶ Premise在数据可视化和用户体验方面做得比较好。



数据即商品：BlueKAI-聚小成大

BlueKai@Oracle: 聚集市场营销数据



- ▶ BlueKai 所做的主要工作是从一些掌握拥有部分有价值客户流量的个人或者中小网站那边购买相关信息；
- ▶ 然后将这些信息进行分析归纳，从而总结分类出更具市场价值的流量信息，并最终进行网络拍卖；
- ▶ 目前这些研究数据的购买者包括美国排名前十的在线广告网站。

数据即商品：Xignite-专业金融数据API服务

The image displays the Xignite website interface. At the top, there is a navigation bar with links for Products, Developers, Support, Resources, Contact, and Login, along with a search bar and a 'Free Trial' button. Below this is a large banner with the text 'We Power You: The Future of Finance. With Financial Market Data APIs For your apps. In the Cloud.' and the Xignite logo. The main section is titled 'API Catalog' and features a search bar. On the left, there is a 'Refine List' sidebar with a dropdown for 'Asset Class' and a list of categories including Equities (23), Funds (7), ETFs (18), Bonds (6), Indices (4), Futures (2), Options (3), Forex (3), Metals (2), Rates (3), and Derivatives (2). The main content area is divided into three columns, each representing a different API service: XigniteGlobalCurrencies, XigniteGlobalQuotes, and XigniteWidgets. Each column includes a brief description of the service and a list of available data types or features.

Refine List

Asset Class

- Equities (23)
- Funds (7)
- ETFs (18)
- Bonds (6)
- Indices (4)
- Futures (2)
- Options (3)
- Forex (3)
- Metals (2)
- Rates (3)
- Derivatives (2)

XigniteGlobalCurrencies →

Real-Time and Historical Foreign Currency Exchange Rates API (Forex/FX)

This forex rates API offers real time and historical quotes for currency exchange rates (FX). It provides support for more than 170 currencies and over 29,000 currencies pairs.

Features: **RealTime**, **Historical**, **Quotes**, **FX**

API List →

XigniteGlobalQuotes →

U.S. and Global Delayed Stock Quotes API

Provides delayed stock quotes for U.S. and international equities.

Features: **Equities**, **ETFs**, **Delayed**, **Quotes**, **APAC**, **LATAM**, **NORTHAM**, **EMEA**

API List →

XigniteWidgets →

Stock Market & Forex Widgets, Financial Charts & Tickers

Provides stunning, next generation HTML5 powered market data widgets with stock tickers, fx charts, news and events.

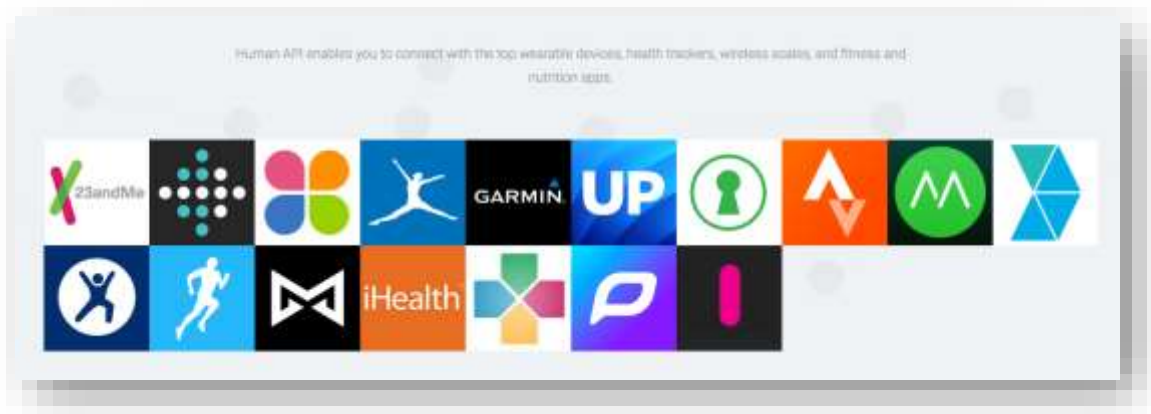
Features: **Equities**, **ETFs**, **Indices**, **Futures**, **Options**, **Forex**, **FX**, **Charts**, **News**, **Events**, **Tickers**, **Widgets**, **Quotes**, **Valuation**, **Performance**, **Markets**, **RealTime**, **Delayed**, **Historical**, **NORTHAM**, **EMEA**, **APAC**, **LATAM**

- ▶ 专注于金融行业大数据的搜集整理和云服务，提供股票、期货、指数、新闻类、企业基本信息等大量金融相关数据；
- ▶ 数据以API方式提供，也提供云端服务host用户的程序。

数据即商品： Human API-健康数据API



- ▶ 专业提供健康数据API，数据来源包括各种健康穿戴设备和各个医疗机构等；
- ▶ 提供简单的数据接入方式，可以集成进各种应用程序。
- ▶ 依赖于统一各种设备提供上的数据接口，并需要协调各种医疗健康机构的数据来源接口。



商业模式二： 技术服务模式

▶ 数据技术

- ▶ 数据存储及并行处理技术
- ▶ 语义及知识图谱技术
- ▶ 数据挖掘及机器学习技术
- ▶ 数据可视化技术

▶ 技术交易

- ▶ 算法交易
- ▶ 数据分析众包



技术服务模式：Hortonwork-Hadoop技术服务

- ▶ Hortonworks 由来自原来的 Yahoo! Hadoop 开发和运营团队的 24 名工程师成立于 2011 年，由于是原班人马，因此积累的 Hadoop 经验比其他任何组织都多。
- ▶ 从事 Hadoop 平台核心的设计、构建和测试。



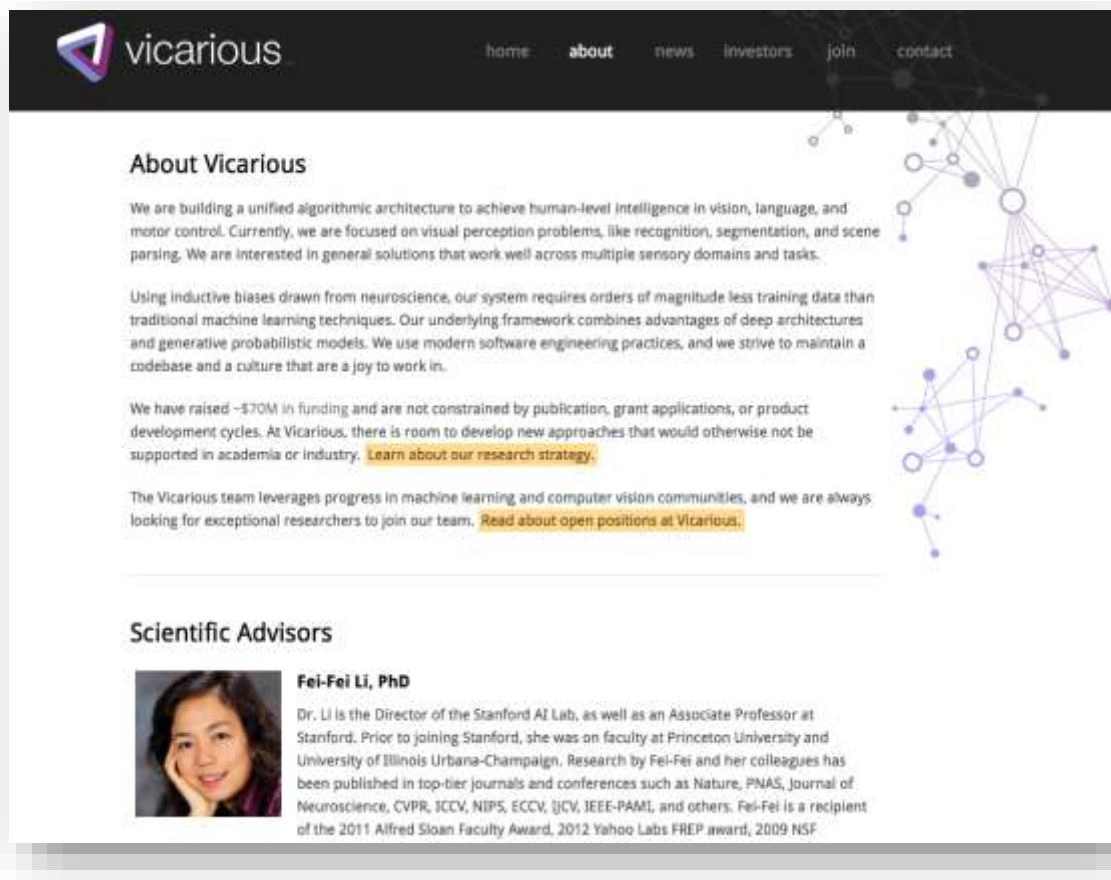
技术服务模式： DataBricks-SPARK技术服务

- ▶ 由开发SPARK的UC Berkley大学的AMPLab创办的公司，专门做SPARK技术平台服务。



技术服务模式：VICARIOUS-机器学习技术

- Stanford AI实验室支撑创办，专门研究新一代可商用机器学习方法，目前专注于视觉领域。



The image is a screenshot of the Vicarious website. The header features the Vicarious logo (a stylized 'v' in a purple and blue triangle) and navigation links: home, about, news, investors, join, and contact. The main content area is titled 'About Vicarious' and contains three paragraphs of text. The first paragraph describes the company's goal of building a unified algorithmic architecture for human-level intelligence in vision, language, and motor control. The second paragraph discusses the use of inductive biases from neuroscience to reduce training data requirements. The third paragraph mentions the company's funding and the availability of open positions. To the right of the text is a decorative graphic of a neural network with nodes and connections. Below the 'About' section is a 'Scientific Advisors' section featuring a portrait of Fei-Fei Li, PhD, and a brief biography of her work at Stanford and Princeton, as well as her research interests and awards.

vicarious

home about news investors join contact

About Vicarious

We are building a unified algorithmic architecture to achieve human-level intelligence in vision, language, and motor control. Currently, we are focused on visual perception problems, like recognition, segmentation, and scene parsing. We are interested in general solutions that work well across multiple sensory domains and tasks.

Using inductive biases drawn from neuroscience, our system requires orders of magnitude less training data than traditional machine learning techniques. Our underlying framework combines advantages of deep architectures and generative probabilistic models. We use modern software engineering practices, and we strive to maintain a codebase and a culture that are a joy to work in.

We have raised ~\$70M in funding and are not constrained by publication, grant applications, or product development cycles. At Vicarious, there is room to develop new approaches that would otherwise not be supported in academia or industry. [Learn about our research strategy.](#)

The Vicarious team leverages progress in machine learning and computer vision communities, and we are always looking for exceptional researchers to join our team. [Read about open positions at Vicarious.](#)

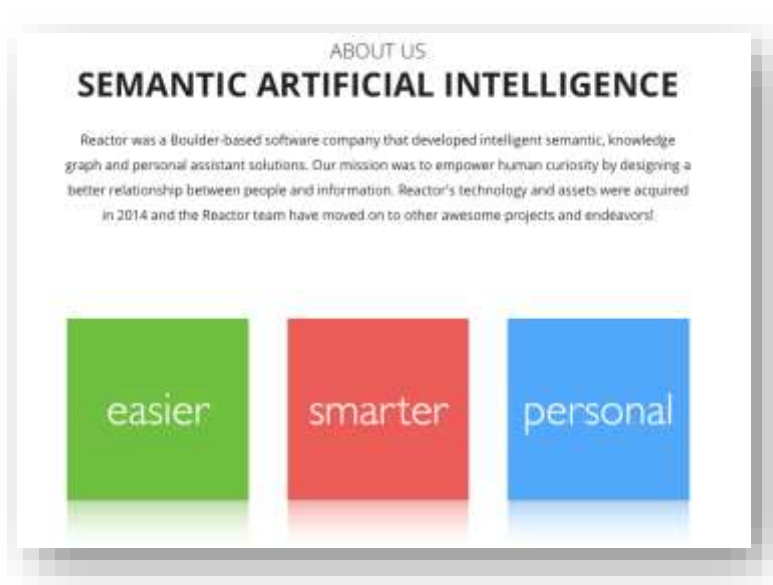
Scientific Advisors

Fei-Fei Li, PhD

Dr. Li is the Director of the Stanford AI Lab, as well as an Associate Professor at Stanford. Prior to joining Stanford, she was on faculty at Princeton University and University of Illinois Urbana-Champaign. Research by Fei-Fei and her colleagues has been published in top-tier journals and conferences such as Nature, PNAS, Journal of Neuroscience, CVPR, ICCV, NIPS, ECCV, IJCV, IEEE-PAMI, and others. Fei-Fei is a recipient of the 2011 Alfred Sloan Faculty Award, 2012 Yahoo Labs FREP award, 2009 NSF

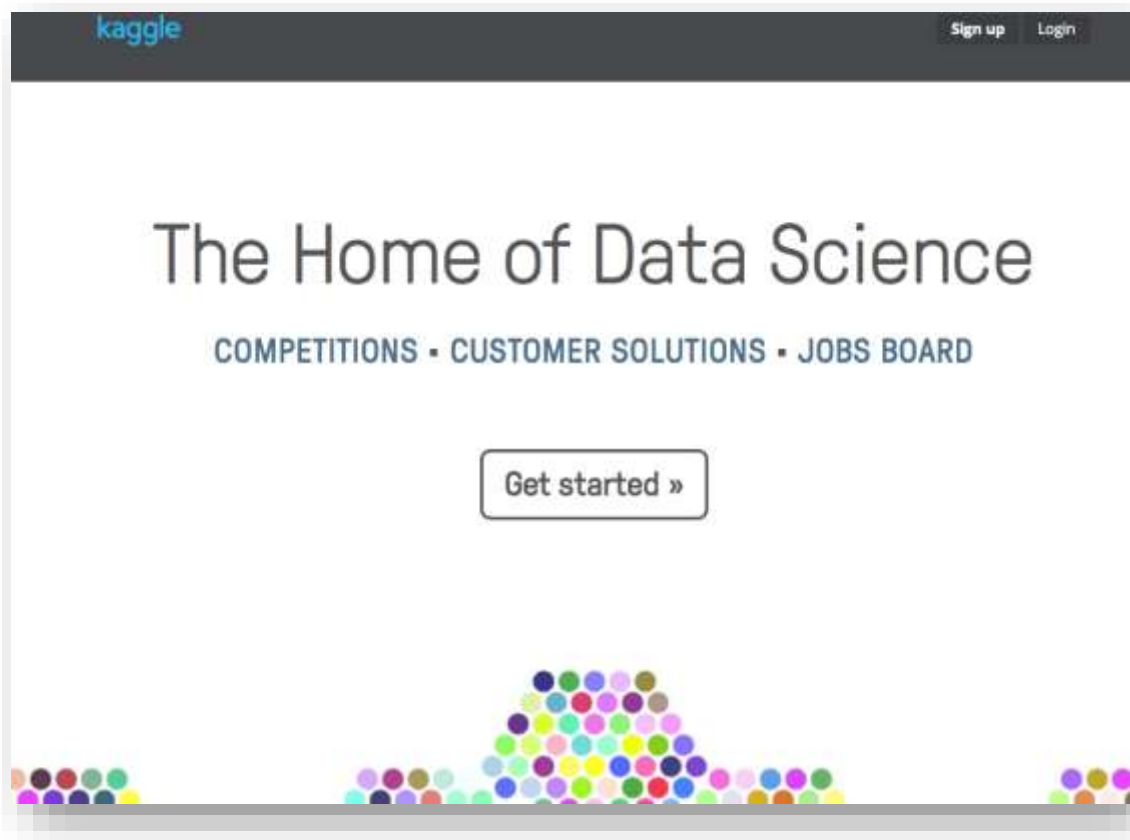
技术服务模式： REACTOR-语义技术服务

- ▶ 主要提供自然语言处理、语义分析、知识图谱、知识库等方面的技术服务。
- ▶ 主要面向个人信息建立个性化的知识图谱，提供个人信息助理、对话式搜索、智能信息推荐等功能。



技术服务模式：Kaggle-数据分析众包

- ▶ 提供数据科学的众包服务平台。
- ▶ 需求方上传数据分析需求，平台众包该分析任务，并按任务完成情况支付费用。



商业模式三：云平台模式

- ▶ 采用云服务的模式提供数据和技术的在线服务
 - ▶ 云存储： Infrastructure as a Service
 - ▶ 开放平台
 - ▶ 云分析： Platform as a Service
 - ▶ 图像识别云、语音识别云
 - ▶ 云可视化： Software as a Service



云平台模式： 百度数据开放平台



数据开放平台

18657122067 [退出] | 帮助

首 页

PC端数据开放

移动端数据开放

合作案例

我的资源

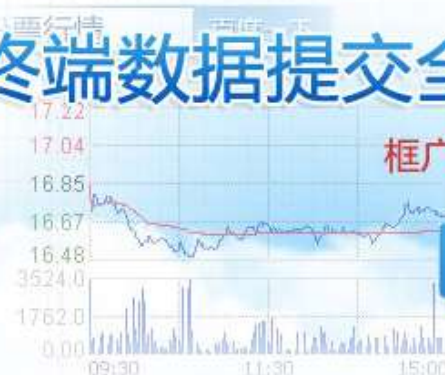
+ 提交资源



移动终端数据提交全面开放

框广天地 无限精彩

我要提交



PC端数据开放

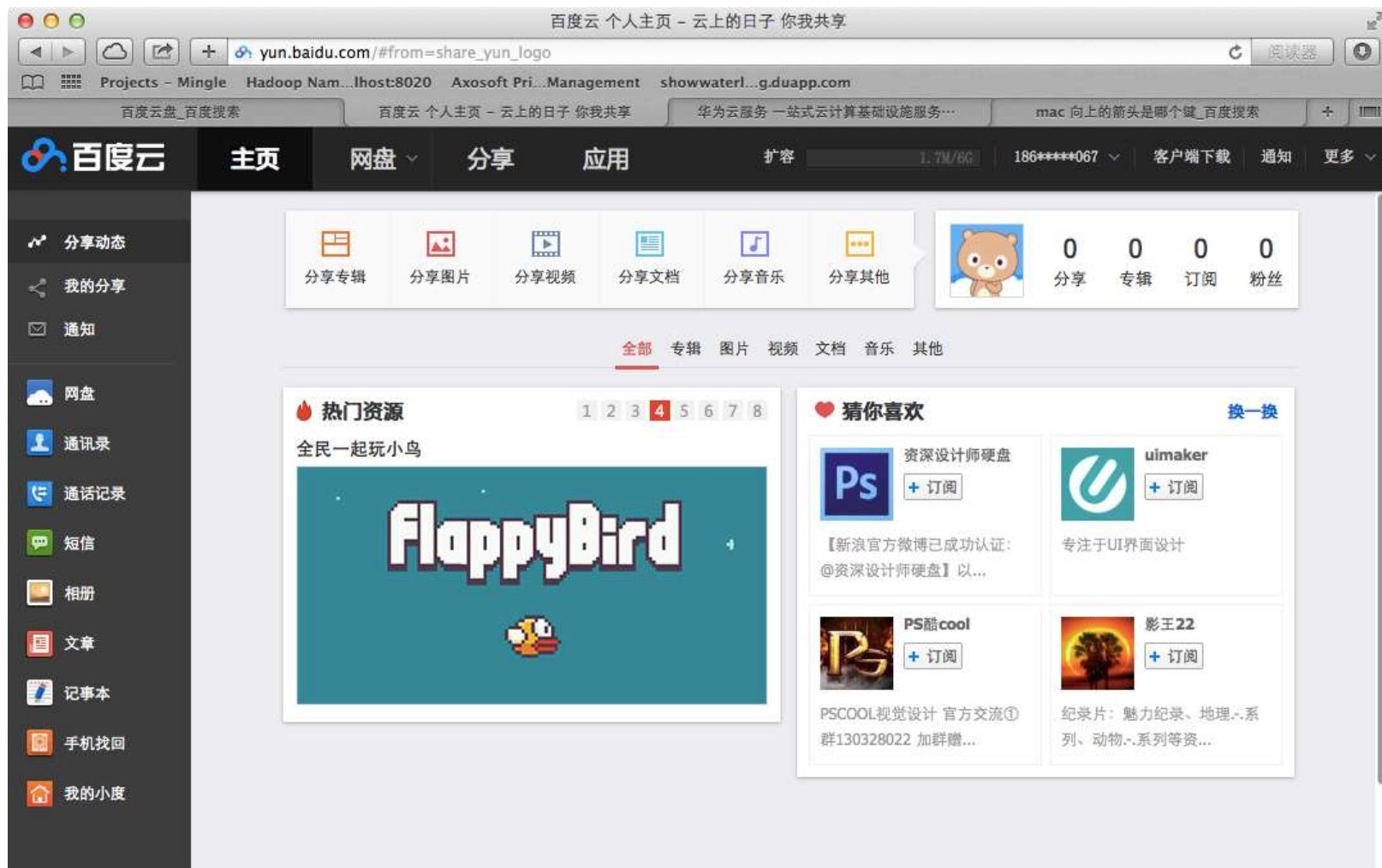
通过此入口提交的数据，通过审核后，将展现在PC端百度搜索结果页



移动端数据开放

通过此入口提交的数据，通过审核后，将展现在移动终端百度搜索结果页

云平台模式：百度云盘 (SaaS)



云平台模式： 微信智能开放平台

- ▶ 2014年上线，分为微信语音开放平台和图像开放平台，分别提供基于微信的语音处理和图像识别SDK和API，应用开放者可以直接在其应用中集成相关功能。



语音开放平台

目前已开放的有通用语音识别、词表识别、语法识别、语音合成等语音技术。



图像开放平台

微信图像开放平台致力于为第三方应用提供免费的图像识别技术和服务，敬请期待。



<http://pr.weixin.qq.com>

云服务模式： SumAll-企业大数据综合云服务

The tools you need to make you the best **online seller.**
Analytics, Reports, Insights, & more.

TRACKING
All your data in one (free) place.
The best and easiest way to keep track of all your key business and social media stats. Save time and keep your team in sync with **Tracking.** From Twitter to Etsy, there are 30+ platforms to choose from and SumAll keeps it all neatly organized in a single dashboard. Did we mention it's free?

INSIGHTS
Caffeine for your social efforts.
Insights is a service designed to help you understand and improve your social media performance. Divided into Content and Audience analysis, **Insights** analyzes timing, hashtags, post length, keywords, and more to offer guidelines that will make your content the best it can possibly.

Sign up form: Email, Password, Get Started - Free, or sign up with social media icons.

Dashboard illustration: A laptop displaying a line graph with the SumAll logo, surrounded by icons for chat, currency, email, and a megaphone.

Insights report illustration: A document titled 'Insights' with a bar chart and a pencil.

- ▶ 面向企业用户，提供将企业相关的在线数据

Facebook\Twitter以及企业自身的数据综合集成，提供统一的云端存储，并提供深入的挖掘与分析服务。

- ▶ 围绕企业整合所有相关在线数据
- ▶ 集成在线与内部数据
- ▶ 提供统一的云存储与云分析服务

云平台模式： Domo-商业智能分析云平台

- ▶ 主要针对企业级市场，为企业提供涵盖财务、人事、销售、IT等全方位的数据管理、存储、可视化分析等多种云服务。



谢谢！

