大数据科学与应用系列讲座

2016年12月 星期五　　13:48

一、李国杰：面向大数据的数据科学

1.发展大数据的驱动力：	促进经济发展，促进社会公平正义，促进科学研究（主要是基础科学）发展。

促进经济发展：

大数据--->蜂蜜，主要价值在于传播花粉。自己产生的蜂蜜价值不大。狭义的大数据产业的GDP贡献不大。

促进社会公平正义：

利于国家的治理。经济系统类似于人的血液系统，信息系统类似于人的神经系统，不必用左手证明右手的重要性。

促进科学研究发展：

从大数据到**认知科学**，再到数据科学。

2.数据科学：	数据---->自然体（data nature）----->数据界（data universe）【共性问题】。

有学者定义其为介于哲学和自然科学间的超自然科学。低于哲学，高于自然科学。

数据科学的共性科学问题目前还需采用“先做白盒研究，再做黑盒研究”的方式进行发展。

数据科学是数学（统计、代数、拓扑等）、计算机科学、基础科学和各种应用科学结合的科学。钱学森提出：“大成智慧学”、“必集大成，才能得智慧”。单独方向的科学显出弱相。

3.大数据对计算机科学的挑战

• 计算机是关于算法的科学

图灵计算：输入---->输出，G=F(x)【函数观】

研究函数“F”，即算法

算法不关心输入x，假定了x的随意性。但是x实际上是伪随机的，仍有研究的必要。

·算法+数据结构：大数据兴起导致计算机科学的重点向数据科学转移

Computer Science = Science of algorithm + Science of data

`算法复杂度：小数据条件下好的算法在大数据条件下不再是好算法

1PB的数据线性扫描一次需要1.9天（硬盘速度6Gbps）

·也有些很困难的问题，数据多了就变得更容易解决

如：机器翻译，自然语言问答（IBM的Watson问答系统）

4.大数据对传统计算机视觉（CV）和机器学习（ML）的冲击

计算机视觉、机器学习是人工智能最活跃的研究领域，但多年来学习的样本和测试的样本度不够大。Princeton大学的李凯教授采用在线外包的办法，一年之内完成了2.1万种分类、包含约2000万幅图像（每类700-1000幅）的ontology图像库（ImageNet），（基于wordnet分类，目前只有名词）。采用ImageNet测试现有的各种图像识别分类算法，绝大多数算法都失灵，说明在小的ontology下开发的图像识别算法没有实际意义。但**Deeplearning**算法的正确识别率明显高于其他算法，所以**深度学习成为目前机器学习的主要研究方向。**

5.大数据对传统统计学的挑战

·大数据往往是**非独立同分布**（悉尼科技大学 操云龙）

——统计学的基本假设是变量服从独立同分布（IID假设）

·**超高维问题**引起经典统计推断失效（徐宗本院士）

——经典统计：n>>p．高维：p>>n,大数据高维度p=o(exp(n))

——热点研究：稀疏建模（尽管变量很多，但很多都是0）

——大数据处理和智能处理的核心都是降维，从n维降到1维（如：排序目的）。样本数量将随着维数的增加而指数增加就出现维数灾难。

·分析*与事物相关的所有数据*，而不是分析少量的样本数据

——2009年谷歌利用相关词全部搜索统计（5000万+）准确预报了H1N1流感爆发，比医报部门提前2、3个礼拜

——2013年由于政府发通告、谷歌加推荐等原因，使得谷歌的流感预测失灵，明显高估

——大数据与小数据结合（All data，全数据），原始数据的可信度？

6.网络科学与数据科学：	*复杂网络分析应为数据科学的基石*。

大数据往往以复杂关联的数据网络形式存在，因此要理解大数据就要对大数据后面的网络进行深入分析。大数据面临的本质科学问题可能就是网络科学问题。到了21世纪，网络理论正在成为量子力学的可尊敬的后继，正在构建一个新的理论和算法的框架。

中科院计算所的大数据团队主要从事网络大数据的研究，研究方向包括分布式海量数据处理的核心引擎、计算模型和国家级测试床，网络舆情系统、社会化搜索引擎、数据密集型网络服务等，李国杰院士学生的研究方向包括社会网络的影响力研究、推荐系统等。

7.需要发现新的门捷列夫周期表

·门捷列夫周期表为化学成为一门基础学科奠定了基础。现在生物领域有基因组学，材料、化学、制药、生理、病理、干细胞领域都在研究“基因组”，也有人在讨论人类语言的**“基因组”**，这些基因组都是构成整体的基本元素。

·发现这些“基因组”都*需要采用计算机对海量的数据进行分析*，导致各个领域都出现XX信息学。

·从上个世纪70年代开始，围绕计算复杂性形成了以算法研究为中心的计算机科学。随着计算机科学与其他学科的交叉相融，计算机科学的研究重点将逐步转移到以研究各种基因组学为重点的数据科学。现在到了发现新的门捷列夫周期表的时候了。

8.计算理论的新研究方向

·传统计算复杂性是研究当问题规模变大时，计算量如何变化，以小问题预测大问题。而大数据问题一开始就给你全部数据，需要反过来思考如何找到**缩小规模**的数据，而问题的基本属性没有大的变化。

·如果是传统的计算复杂性是度量外向组合爆炸（scale up）的复杂程度。那么大数据问题的计算理论应该是**度量内向“压缩”的困难程度**。

·如果当数据规模扩大，反应数据间相互关系的网络结构保持很好的相似性，则是一个容易解决的大数据问题；反之，如果网络结构变得面目全非，则是一个难以解决的大数据问题。

***需要研究“数据量复杂性”（问题需要多大的数据量）

建立一种新的计算理论，对求解一个问题达到某种满意程度需要多大规模的数据量，能够给出理论上的判断（多项式级、NP问题、数据规模阈值）。

9.培养“π型人才”：	**所有科学都在迅速变为“数据科学”**。

有经验的计算机科技人才可以大规模提高数据处理速度，在各领域应该培养熟悉数据分析的科研人才。

计算机系需要面对全校调整和新增有关采集、整理、分析的新课程。

二、吴军讲座：数据为王和机器智能的时代

1.什么是机器智能

世界新技术的三个趋势：

·云计算＋移动互联网＋大数据（正在进行时）

·**机器智能（现在时）**

·IT ＋ 生物医疗（未来时）

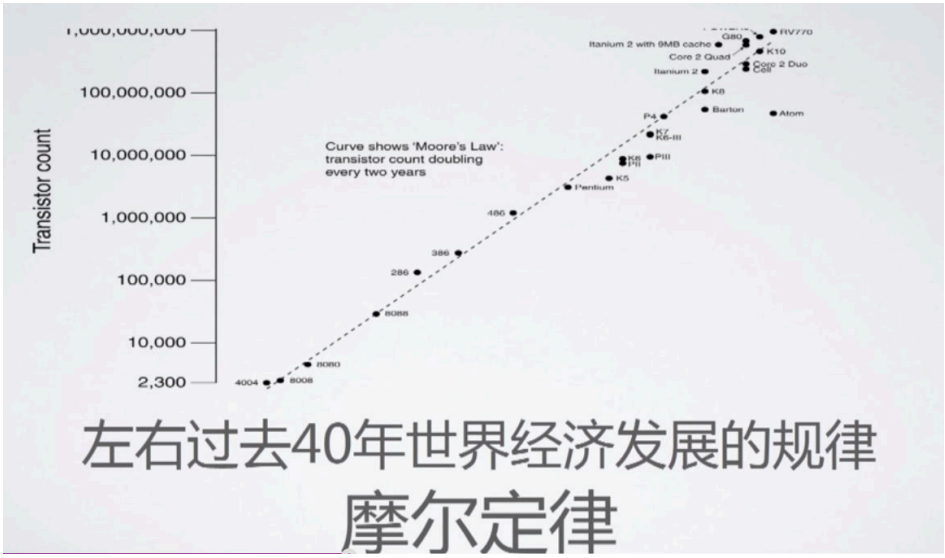
----------脑力：人类的骄傲

----------计算机（电脑）和人工智能

想象中的机器人（高达）

现实中的机器人（扫地机、“好奇号”）

过去40年摩尔定律：世界经济发展的规律



但是，机器智能似乎尚未出现

·无法像人类一样思考

·解决某些智能问题优于人

————计算强大到一定程度，是否会有质变发生？

真正能够取代人的机器人

如何判定机器有智能？

Turing Test

明斯基的谜题：

The pen was in the box.

The box was in the pen(围栏).

2.大数据的重要性和特点

在2005年NIST机器翻译评测中，Google因为适用大量的数据取得遥遥领先的结果。

大数据不仅仅数据量大

根本不是结构化非结构化的问题

多维度

百度百科：吃货的统计。推断饮食习惯、生活习惯、收入水平、性别年龄。无意识的收集。

完备性

美国总统选举的预测。Nate Silver预测2012总统选举，**极其可怕的精确。**

思维和做事情方式的改变

3.大数据中的因果关系与关键技术

因果关系：**不知道原因，先知道答案。**

Big Data（抽象的）————————Large Data（实在的）

大量样本，大量的统计规律。

大数据＋制药

5000种处方药，找出规律性的事件，再进行实验。

google的例子

对待长尾的态度

打比方：计划经济 vs 市场经济

从摩尔定律到大数据为王：

未来所有的公司都是大数据公司

金风公司为产品装上各种传感器，传回全世界的大量风能分布数据，检测金属疲劳的程度。

Prada时装公司，在试衣间放置传感器与衣物相连接，采用大数据，改变时装商场的信息获取。

Target通过E-mail取代传统实体发票，获取了人们的购物习惯。

根本不是几亿、十几亿的故事（孙正义：所有产业都要数字化）

看好Google、Facebook、亚马逊、阿里巴巴等公司

传统数据公司没戏唱

大数据的关键技术：

数据的收集（无目的性、非结构化）

Goole收购Net，收集无意中的很多数据。问卷调查的真实性受到限制。

数据的存储

数据量极大。

数据的表示、检索和随机访问

大数据杂乱无章，很难理出头绪。

Jeff Dean的新挑战，如何能够表示好生物医疗数据。

数据的使用和挖掘

Google40%工程师的工作。先知道实验结果，再找到方法，再应用。

其他挑战：安全、隐私……

4.大数据与机器智能

机器智能的三足鼎立：摩尔定律，大数据，数学模型。

Google大脑-----人工神经网络（其实质为简单的有向图）。简单而稳定，具有通用性。

实现到几万台、上百万台的电脑连接。深度学习训练了数据模型，提高模型质量。

解决了图灵的问题：计算机回答问题。

机器智能本质上是大数据的应用。

问题的答案往往包含在大数据中。传统的逻辑思维----->由零碎信息拼成答案。

Google研制出无人驾驶汽车。

无人机浇水节省水量98%，被hack，做洗车的工具。

未来的时代是机器的时代，还是人的时代？

美国的放射科医生高大上职业，需要培养10多年，年薪30万（硕士到google起薪才10万美元）。现在被机器识别软件部分取代。

5.大数据思维

美国律师，收入高。使用软件。

未来的世界：

2%的人的世界

机器不会控制人类

2%的人控制98%的人

大数据的思维：

全面性（完备性）- 细到每一个人、每一个商品、每一笔交易

酒吧安装瓶底传感器，控制酒被偷喝。

对比互联网思维

电子商务带来？

不仅仅是把商品放到网上去，整个生态链的改变。

大数据：GE冰箱，装传感器，增加过滤器头的利润。大型电器本身成为一个几乎没有利润的“平台”。

未来是2%人的世界。人的前景灰暗。科技的进步并不会造福每一个人（美国工资一直在下降，硅谷的房价却持续上涨【拒绝平庸】）。

医疗

Google成立Calico公司。

Genetech抗癌，生物治疗公司。

抗衰老。（攻克癌症仅能提高3.5年平均寿命）

23AndMe，达芬奇。

今天谁是大数据的公司？

真正的大数据公司

Google、Facebook、Amazon、BAT、**网易**

三、余凯：百度大脑所思考的人际关系

1.人工智能与大数据时代

MIT杂志报道百度的“人工智能之梦”。

——百度是一家人工智能公司。

『搜索数据Mega Data』——>『Driven AI』——>『商业价值Monetization』

人工智能正在成为未来主战场。

What is 人工智能？**感知、理解、决策**。

无处不在的智能XX：智能手机、智能手环……

如何区分真正具有智能？

·*随着经验演化，越变越聪明*【**学习**】

一个时代正在来临……

移动设备上的摄像头，正在成为人类眼睛的延伸。

移动互联，万物互联。万物互联，数据暴增。

2.深度学习介绍

深度学习为2013年十大技术突破之首。迅速影响了最前沿的高科技公司（Google、微软、Facebook）。

百度IDI--中国第一家深度学习研发机构。

百度大脑：大规模并行化数据处理系统。

搜索、广告、图像、语音。让连接变得更智能。

服务越来越复杂，需求越来越自然化。

·理解用户意图	·匹配用户需求	·精准推送广告和链接

为什么深度学习受到重视？

·模拟大脑行为

卷积神经网络：大量像素数据的训练。从基底层到整个物体的分层结构。

·特别适合大数据（能够吸收大数据所带来的红利，带来商业竞争壁垒）

统计和计算方面：

推广误差 = A + E

Approximation error - model class （来源模型的不完美）

Estimation error - data size （来源于有限数据的不完美）

【统计学理论范畴：假设无限计算资源（避免重复性降低可信度）】

推广误差 = A + E + O

Optimization error - algorithm （来源于求解的不完美）

【CS的范畴：考虑计算的不完美】

Use complex model；collect big data；design “an OK algorithm”

更加复杂的模型，大量的数据，能用的算法（可以消化大量的数据）。

·End-to-end学习（端到端，scalable，通用性）

·提供一套建模语言（不同的task上使用不同的深度学习模型）

3.深入百度大脑：DL model for query-docments（相对相关性）

网上抢票码识别、运单手写电话号码识别。【Sequence to sequence】

述说图片的故事。【同时理解图像和自然语言，提供一个丰富的语言系统】

4.深度学习应用实例

深度学习显著提升了百度搜索满意度的领先优势。

移动语音搜索，世界首屈一指的中文语音识别率。

自然图片OCR：百度翻译、百度作业帮。

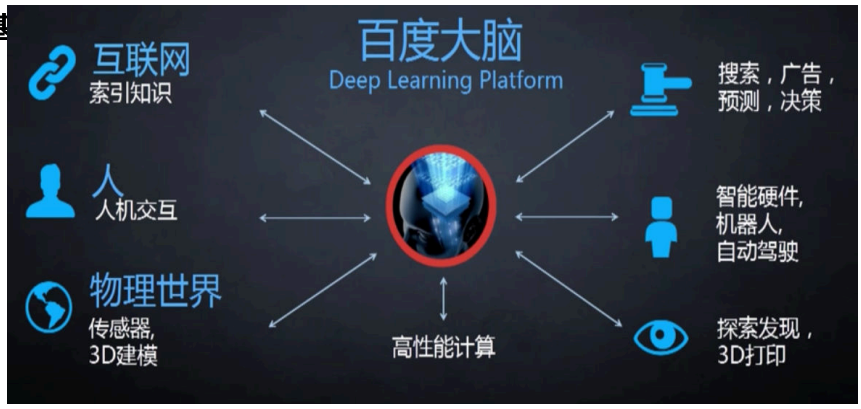
涂书笔记、百度魔图（基于深度学习的人脸识别技术）。

人脸识别：LFW测试。百度99.85%。

百度拥有世界领先的基于内容的图像搜索技术。

ImageNet图像分类。每年的错误率逐渐降低，接近人的识别错误率5%。

5.基



张跋：机器人是互联网服务的最后一公里。

智能与物理世界的连接。

More Neurons vs. More Connections

现实中各类物种，神经网络节点变化时，而神经节点之间的*连接数的规模差不多*。

Human Brain vs. Supercomputer

人脑的计算效率、计算能力很高。

·百度高度自动驾驶项目（人与车的和谐亲密的联系）。

·三维高精度地图和感知定位（cm量级）

·动力控制系统建模仿真

·基于图像的实时道路场景理解（需要低成本的传感器）

——从辅助驾驶到人车一体的高度自动驾驶。

关于人与机器--人与机器的联系。Communication. Turing Test.

关于技术----人的自由，技术受到控制。

伟大的技术，不在于让机器更伟大，而在于让人更加伟大。

四、吴甘沙：大数据的十个技术前沿

说明——10个不代表Offical分类

技术的长期影响力。（指数效应：颠覆成为常态，摩尔定律成为指数社会的“基因”，大数据成为指数社会的“蛋白质”）

大数据的技术挑战：全集>采样，实时和全时，见微知著，人与工具。

~~~~数据~~~~

1.膨胀宇宙

“data universe”.

——TB————PB————EB——ZB——YB……

DBMS MPP NoSQL Globally

NewSQL Distributed(DB)

DFS

Array DBMS Cache in-memory FS 大数据堆外内存

Flash storage Erasure Coding

DRAM storage

NVRAMstorage

2.巴别之难

数据之间可以相互联系，但相互交流的“语言”可能不同。因此需要数据的处理。

DB-hard：地址格式的不通译性。

如何发现数据之间的结构？Discovery of structure ; Entity resolution ; Transformation.

——推广到非结构化数据。【Apache UIMA】

3.数据有价

·安全与定价

——数据权利的定义：

·数据分红权

·数据隐私权

·数据许可权（撤销，转移）

·数据审计权

·数据拥有权

安全：系统安全、数据安全、数据使用安全

***系统安全——数据高度分布、去中心化场景下的安全（分布式安全架构）

***数据安全——静态数据安全（加密、访问控制）、动态数据安全（动态审计能力）、个人对数据的控制（Do Not Track，Mac地址混淆，个人数据删除，openPDS）

数据脱敏/匿名化

——去标识符[准标识符仍能重新标识化]，隐私攻击[多数据源的相互匹配，基于统计学的攻击]，差分隐私[人为在数据中插入噪声]，隐私安全和数据可用性的平衡

***数据使用安全——可用不可见，相交不相识（同态加密数据库技术，基于加密协议的多方安全计算，基于可信计算环境的多方安全计算）

审计和定价：

***把数据安全或隐私条款形式化（适用于非专业人员使用）

***根据形式化的规范，对数据进行审计（静态或动态跟踪数据使用）

***定价——效用和稀缺性

·效用：基于使用频率和对结果的影响，判断各方数据的贡献

·稀缺性：数据价值密度和历史价格

相逢不必相识，没有使用没有买卖。（数据只有被使用才有被卖卖的价值）

~~~~计算~~~~

4.软硬兼施

选择更好的硬件架构

***计算

·大小核

·异构计算（FPGA、ASIP）

• 内存和处理器更加靠近（获得更大缓存、带宽）

***存储和互联

·大内存服务器

·SSD ——> Pcle SSD ——> flash storage（重构系统软件栈，智能数据迁移）

• NVRAM（未来趋势）

• 互联

软件与硬件架构协同优化

***针对硬件特点对软件栈优化

把硬件暴露给软件栈，重新设计软件栈，一体机，云化（虚拟化和资源管理）。

***硬件可重构

机架，数据中心/广域。

***Big Learning System：机器学习与底层系统的更好配合

VM, Graph Lab, Dis Blief, Project Adam, Petuum.

5.多快好省

——未必能而兼得之。

·软硬件协同设计

·内存计算

技术占多层。

**在数据分析和可视化层：

……重新设计数据结构

·分析：图(GraphLab vs. GraphChi)

• 结构的可改变性：在线机器学习

·可视化：in memory data cube（e.g. Nanocube）

…….原位（in-situ）分析和可视化

·降低空间、时间复杂度

空间：把大数据变小（压缩、Spark）

时间：更多数据+简单模型，简单模型的组合（ensemble），采样和近似，稀疏数据算法，降维和混合建模

·并行化/分布式（Acid——>Base，从“酸”到“碱”）

6.天下三分

• 数据类型的分野

表格/K-V，数组/矩阵，图

***关系查询，以线性代数为代表的复杂分析，图计算

• 计算范式的分野

计算图：数据依赖，无计算依赖

***批量计算：数据不动，计算动

·MapReduce：二阶段

·BSP：三阶段

·DAG和多迭代计算

***流式计算：计算不懂，数据恒动

·Record-at-a-time vs. minibatch

·不同时钟语义和投递保证（delivery guarantee）

·简单计算——>流式/在线机器学习（e.g. SAMQA、Jubatus）

图计算：数据和计算依赖

7.分久必合

·融合

Big Dawg普适编程模型

***支持关系和线性代数、复杂数据模型、迭代计算和并行计算

Twitter Summingbird在**编程接口层面**融合

***支持批量和流式

Lambda架构在**应用框架层面**融合

***增加计算和批量缓存

Spark在**实现框架层面**完成融合

·流查询：Spark Streaming + Spark SQL（实时系统与数据仓库的结合）

• 实时+批量：Spark Streaming + Spark Core（类Lambda结构结合实时洞察和历史洞察形成全时洞察）

• 流处理+机器学习：Spark Streaming + MIlib（实时获得更复杂的洞察）

• 图流水线：从MapReduce + 图引擎到Spark Core/SQL + GrabphX（图创建/ETL+图计算+后处理）

• 即席查询+机器学习：Spark SQL + MLlib

REEF通过**资源管理层**来支持多计算模型

~~~~分析~~~~

8.精益求精（机器学习精确度的提升）

·对更多数据的包容性

·覆盖长尾 Exponential assumption vs. long tail

• 在线/流式学习

9.人机消长

人和机器作用的变化

i. 机器取代人

以机械化的数据挖掘寻找相关性来替代主观假设

***数据自己找到数据，相关性主动找到你

机器语义分析的增强

自动化分析和可视化

***MLBase和VizDeck

ii. 让工具增强人

可视化库、框架和工具

***文本、网络/图、多维数据

***交互式可视分析：界面隐喻与交互组件，多侧面、多尺度、多焦点交互

社会化基础建设和大规模协作分析

***云化：Databricks Cloud

***CrowdDB, Kaggle, Duolinguo

iii. 机器学习中人的作用

数据标记，特征工程，方便易用性。

10.智能之争（生物智能vs. 机器智能）

模式匹配，统计学习。

***计算智能，人工神经网略，进化计算，模糊逻辑，人工免疫系统，群体智能。

当前在热烈讨论的问题：

·深度学习有没有可能包打天下？

·智能的未来是不是类脑？

·需不需要新的类脑计算架构？

五、吴甘沙：大数据分析师的卓越之道

1.数据思维方式的改变

数据分析的典型场景：

数据——>知识发现（基础设施）——>价值

新的世界观：不确定的世界。【薛定谔的猫】【大数据测不准】

***MOOC中提到的Google流感测不准的实际原因貌似是因为在后来政府和Google的宣传导致了干扰（最开始的预测是非常准确的）。~~~~~大数据的傲慢：关注相关性而忽视因果性。

数据分析论的升级：之前的假设要随时准备改进。

2.数据的假设与采集

·假设Hypotheses

首先采集大量数据，通过数据发掘来发现相关性（很多其实似是而非），通过直觉来判断真实性。（拿侦探小说练手，阅读广泛涉猎，跨界思维碰撞，融入业务部门[防止数据分析脱节]）

·采集Collection

数据！数据！数据！大量的数据，形成数据的全局n=all。

数据从结构化，半结构化过渡到非结构化。

数据？数据？数据？n=all？更多的数据不一定就更好，数据自身价值的分析，采样的偏差。数据的权利。数据的生命周期（**尽快处理**）。

3.数据的准备 Preparation

数据质量：重中之重。

大数据天生具有大量的噪声。需要清洗[机器学习实现自动化？]、验证[可视化？]，治理。

数据表示

降低下一步分析的复杂度：计算、通信复杂度[大数据的稀疏性]，统计复杂度。“UIMA”.

4.数据的分析**

数据库+机器学习+统计+……

• 检查自身装备

SAS，R语言[统计]，SQL[数据库查询分析]，Python[机器学习]。

Java，Scala[写了Spark]。

JavaScript[适于可视化，写了D3库]。

·好消息：ML Pipeline（一站式），拥抱云的世界（大数据到云端）

——所有的模型都是错误的，但有一些是有用的。

刺猬（一招鲜吃遍天）vs. 狐狸（一把钥匙开一把锁）

模型的复杂度与问题匹配：奥卡姆剃刀原理

如何做到数据越大，边际收益越大？

• 简单模型+大数据 > 复杂模型+小数据。

·Ensemble，多个方法的结合

• 混合模型

Velocity, Deep Learning, Sparse Coding.

• 人的角色 Human Machine Intelligence

人的工作不断被取代——>人与机器获得最佳性能 [探索式]——>大规模人人、人机协作分析。

5.数据的解释和验证 Interpret Evaluation

STEM——>STEAM。 A：Art。***讲故事 Story Telling***：3D，戏剧性Drama，细节性Detial，身临其境Dialog。

****Ideas Worth Spreading****

sometimes：模拟发生，平行世界。

**基础设施已经改朝换代，分析师也需要与时俱进

·改变思维方式

·提高技术素养

·丰富分析能力

六、董飞：硅谷公司的大数据实战分析

1.硅谷热门公司

- Mature Company：Microsoft、IBM、Intel、Oracle.
- Public Company：Google、Facebook、Linkedin、Twitter.
- Pre IPQ：Uber、Palantir、Cloudera、Square.
- Strat-up：Quora、Houzz、Cousera、Quixey.

2.大数据简介

未来趋势：指数增长、工业革命、摩尔定律、信息爆炸、奇点临近。

技术机会逼近人类历史上的某种本质的奇点，从那以后全部人类行为都不可能以我们熟悉的面貌继续存在。——冯·诺依曼

***大数据能解决一切问题吗？

找工作、电子商务、在线教育、移动app、数字化医疗、互联网金融……

云端、数据库（NoSQL【Not only SQL】）

Hadoop起源：MapReduce, The Google File System, Big Table.（Ecosytem）

——提供和更新更多样、更复杂的计算资源。

3.工业实践

LinkedIn Practice - Kafka.（建立中转站，所有类型和工作站都向其转化）

Cousera数据产品实践：学生互评系统，每个人的信息收集。

·100% hosted on AWS

- Singe-page JS apps
- service-oriented architecture

- 3rd-party monitoring tools
- Naptime REST framework
- Cassandra(peer-to-peer)

七、艾小缤：大数据评价系统在金融、征信领域

1.金融大数据时代

“黑客帝国” ——> 身处大数据时代 —— 真实世界在虚幻世界的投影。

·个人行为都以数据形式被记录、被存储、被处理，经济体是个体行为的集合。

——越来越多的企业行为被记录下来。

***大数据的特征：

大容量	快速度（信息不对称 ——> 信息爆炸）	多样性（行为数据等，从不同角度得到的数据）。

现在最大的挑战：

数据孤立	几乎80%在政府手中
数据杂乱	标准、格式、存储……
传统分析方法落后	得到的信息过于片面，无法有效支持决策，不算是大数据

**小数据时代：财务报表模式。

**大数据时代：看重过程、结果、时间轴，三维动态、最小颗粒度。从外部数据到内部数据的扩张。

三个数据入口：数据挖掘机器人（从网上自动获取），合作伙伴（如企业自主填报【需要数据清洗】），跟政府合作（中国在某种程度上拥有世界上最好的数据资源【最好的税收系统、文化记录系统】）

2.大数据客观信用

·传统信用体系——体制信用。

传统信用体系指西方的信用体系，是一种基于政治体制下的体制信用。通过以下四个信用体系来约束一个个体的行为：

·个人信用	·商业信用	·司法信用	·政府信用

传统信用体系是一种约束机制：

未来收益——>	限制
司法 ——>	惩戒

数据时代不仅仅是建设信用，也是挖掘信用。

从互联网时代浸入数据时代，各种渠道的海量信息、高频词的新数据不断产生。企业的大数据造假成本提高。可以通过大数据挖掘“客观信用”。金融本质是风险的。

大数据只是提高效率，不能替代决策。

大数据金融的特点：

数据是最大的资产	一切反映财产和信用等情况的个体数据都转化为数据资产的形式存在
数据资产需要挖掘	数据资产有其特有的挖掘机制，通过对个体行为数据进行挖掘得到（内在）
数据价值充分利用	大量原本无法利用的数据经过挖掘和加工提炼，变为有用的数据资产

如何挖掘出数据资产：***通过生产数据资产包将数据提炼成数据资产***。

数据资产和数据资产包的差别：

·数据资产是一种概念，而数据资产包是现实金融产品

- 数据资产包是数据资产的存在形式
- 数据资产包是通过对个体数据挖掘和分析生产出来的金融产品
- 一个个体可以拥有多个数据资产包

3.我国的金融环境

***李克强：*让“信用”成为社会主义市场经济体系的“基础性”；运用大数据等手段提升监管水平；小微企业要重点抓。*

***马凯：*中小企业融资难的核心问题：缺信用，缺信息。*

要通过大数据构建真实，真实是破除风险的最有效的手段。

4.客观信用的实践

看历史	找规律，周期性，相关性
看现状	按照数据分类给出当前数据值或指标，对当前状态进行体检
看未来	趋势预测，关键指标分值评判(实施监管)

——> 基于数据类别的分析：销售、费用、产品、客户、人力、银行流水……

数据分析：

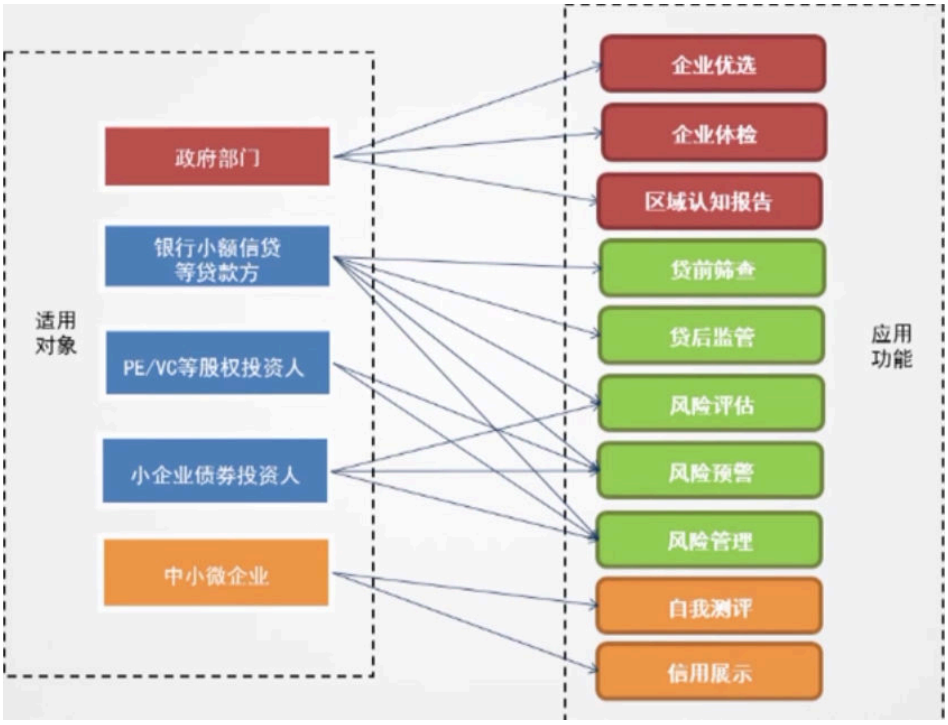
数据采集——>数据清洗——>数据归类——>分析计算——>输出：数据资产包（明细数据、投资指数、[多角度]评分【动态化】）。

中小企业的信用贷款、金融风险管理（省人工，提高效率，还原真实）……

5.应用案例和成果

·金电（汽车后视镜的翘楚，联保互保对本身企业有风险，银行审批信用【由信息云技术收集、评价】，授信2000万元，申请到授信为7个工作日）

•应用



·实践成果

·中小企业纯信用贷款（超过40亿元纯信用授信，无一笔不良【有风险抑制率】）

·与银行合作

·政府大量的支持

·社会关注

八、王迪：数字融合下的美国视频广告生态和产品应用创新

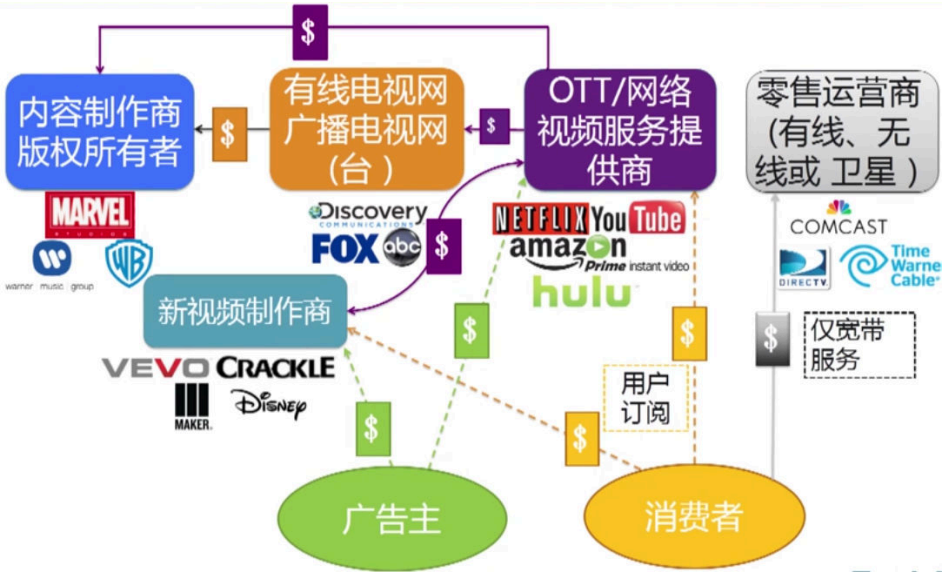
1.美国电视（视频）产业生态

• 传统电视传播的价值链

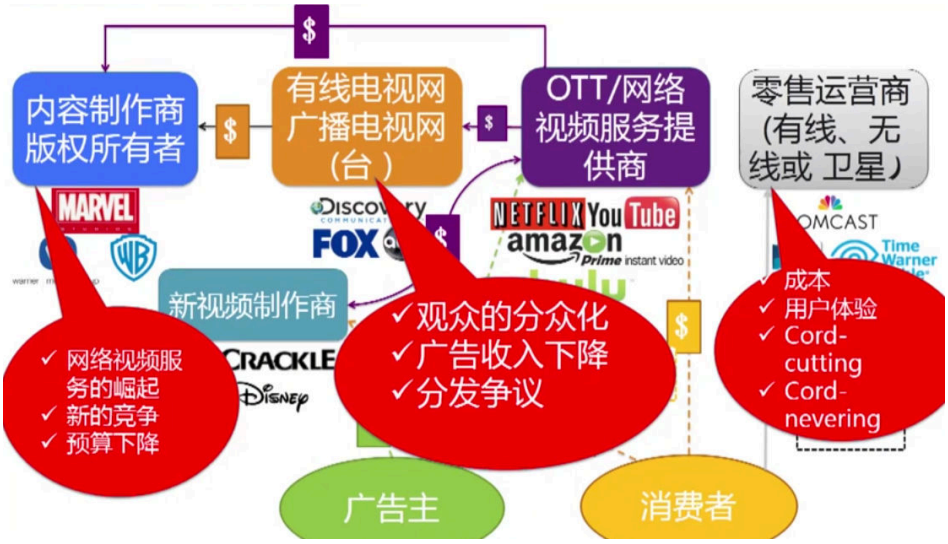
特殊：【美国】广告主——>直播、点播。（例：FOX/COMCAST的广告位7/3分成）

• 新技术背景下的视频价值链

趋势：全球电视市场从传统向数字转移。



• 新价值链带来的挑战



Cord-cutting（老一辈，对有线电视还有印象），Cord-nevering（新一代，根本没有有线的意识）。

未来电视的定义将不再局限于屏幕与传输方式（电视内容即电视）

【优质内容，有限的，稀缺的】

2.数字视频广告产品创新

• 专业内容提供商在Youtube进行广告联营

专业音乐视频网站VEVO为Youtube提供专业MTV内容，并拥有优先广告售卖权，给Youtube分成；如无广告则用Youtube广告回填，反向分成给VEVO【不同的Model】

·更多传统电视媒体的网络视频广告应用

• 根据内容、时长、中断时点自动生成广告位

·体育赛事**跨平台直播**

·更加灵活的赞助购买代替传统CPD

• 常见的内容及页面CPD赞助买断 = $100

• 综合应用行业性排他和豁免及定量赞助模式 = $150

3.数字视频广告基本能力演示

• 演示内容

广告产品类型，伴随广告（贴片、悬浮、页面展示联动……），广告互斥，位置指定，频次限制……

·讲解内容

精准定位，用户体验，数据指标。

·视频广告的行业标准

标准化的程序模式，通过相同的借口，实现广告行业内部的更好流通。

• 创新的视频广告形式

·关于数字广告的可见性（VIewability）

IAB对于可见性(Viewable)的定义
- 展示广告50%像素面积被持续可见不少于1秒
- 桌面视频广告50%的面积被持续可见不少于2秒
- 大尺寸展示广告(242500以上像素)30%面积被持续可见不少于1秒

局限：
- 今天第三方监测提供商对于campaign可见性的测量值与Publisher端投放值之间存在30%~40%的误差
- MRC: 100%的可见性测量在当前技术条件下是不可行的

方向：
- GRP最大可能成为跨媒体平台测量的统一指标
- 最终：100%的数字广告曝光需支持可见性测量
- 过渡期：70%的曝光可以测量可见性
- 可见性测量需由MRC认证的提供商完成，且买家&卖家一致选择

4.影响行业未来发展的重要变量

***观看行为与测量货币的碎片化：观看行为向多屏迁移+跨屏统一测量指标尚未完备。

·受众规模	·内容质量	• 数据	·自动化（程序化）

广告预算正在从传统电视流向数字电视：地区性电视广告主在数字视频上的广告花费增长。需要统一的规模化、基于数据和自动化的运营平台。

统一广告管理	跨屏电视库存	跨屏测量体系和统一货币	跨生态系统运营

关于FreeWheel：

·使命：帮助电视媒体在数字化时代最大化挖掘内容的商业价值。

九、王新锐：金融大数据的法律实践

1.概述

大数据的特征不仅仅是数据的规模大，更重要的一个特征就是这些数据之间的关系非常复杂。——Cullen

• 大数据时代用户数据的价值

• 个人信息和隐私权（现在很有争议，互联网隐私易泄露，如何避免违规）

·互联网金融和征信（关系很密切）

·个人、监管者、媒体关注点（具体场景可能暴露一些安全问题，应注意预防）

• 用户个人信息分类

·一般信息（除个人敏感信息以外的个人信息，漠视同意）

·敏感信息（一旦遭到泄露或修改会对标识的个人信息主体造成不良影响的个人信息，明示同意）

***又可分为：身份信息、财务信息、行为信息、设备信息。

*****处理环节*：

·收集（目的合法）

·加工（加工目的及方法）

·转移（转移目的和范围）

·删除（目的达到后，删除）

2.法律框架和实践案例

•法律框架

《刑法》**出售·非法提供公民个人信息罪非法获取公民个人信息罪**

第二百五十三条国家机关或者全融、电信、交通、教育、医疗等单位的工作人员，违反国家规定，将本单位在履行职责或者提供服务过程中获得的公民个人信

息，出售或者非法提供给他人，情节严重的，处三年以下有期徒刑或者拘役，并处或者单处罚全。窃取或者以其他方法非法获取上述信息，情节严重的，依照前款的规定处罚。单位犯前两款罪的，对单位判处罚全，并对其直接负责的主管人员和其他直接责任人员，依照各该款的规定处罚。

《**消费者权益保护法**》

第十四条消费者在购买、使用商品和接受服务时，享有人格尊严、民族风俗习惯得到尊重的权利，享有个人信息依法得到保护的权利。第二十九条经营者收集、使用消费者个人信息，应当遵循合法、正当、必要的原则，明示收集、使用信息的目的、方式和范围，并经消费者同意。经营者收集、使用消费者个人信息，应当公开其收集、使用规则，不得违反法律、法规的规定和双方的约定收集、使用信息。

《**侵权责任法**》

第三十六条网络用户、网络服务提供者利用网络侵害他人民事权益的，应当承担侵权责任。

网络用户利用网络服务实施侵权行为的，被侵权人有权通知网络服务提供者采取删除、屏蔽、断开链接等必要措施。网络服务提供者接到通知后未及时采取必要措施的，对损害的扩大部分与该网络用户承担连带责任。

网络服务提供者知道网络用户利用其网络服务侵害他人民事权益，未采取必要措施的，与该网络用户承担连带责任。

全国人大常委会《**关于加强网络信息保护的决定**》

国家保护能够**识别公民个人身份**和**涉及公民个人隐私**的电子信息。任何组织和个人不得窃取或者以其他非法方式获取公民个人电子信息，不得出售或者非法向他人提供公民个人电子信息。网络服务提供者和其他企业事业单位在业务活动中收集、使用公民个人电子信息，应当遵循**合法、正当、必要**的原则，**明示收集、使用信息的目的，方式和范围**，并经被收集者同意，不得违反法律、法规的规定和双方的约定收集、使用信息。网络服务提供者和其他企业事业单位收集、使用公民个人电子信息，应当公开其收集、使用规则。

工业和信息化部《**电信互联网用户个人信息保护规定**》

第八条 电信业务经营者、互联网信息服务提供者应当制定用户个人信息收集、使用规则，并在其经营或者服务场所、网站等予以公布。

第九条 未经用户同意，电信业务经营者、互联网信息服务提供者不得收集、使用用户个人信息。

电信业务经营者、互联网信息服务提供者收集、使用用户个人信息的应当明确告知用户收集、使用信息的目的、方式和范围，查询、更正信息的渠道以及拒绝提供信息的后果等事项。

**********如最近的“芝麻信用”的用户协议。

工商总局《**侵害消费权益行为处罚办法**》（2015 年 3 月 15 日起施行）

第十一条经营者收集、使用消费者个人信息，应当遵循合法、正当、必要的原则，明示收集、使用信息的目的、方式和范围并经消费者同意经营者不得有下列行为：

（一）未经消费者同意，收集、使用消费者个人信息；

（二）泄露、出售或者非法向他人提供所收集的消费者个人信息；

（三）未经消费者同意或者请求，或者消费者明确表示拒绝，向其发送商业性信息。

前款中的消费者个人信息是指经营者在提供商品或者服务活动中收集的消费者姓名、性别、职业、出生日期、身份证件号码、任址、联系方式、收入和财产状况、健康状况、消费情况等能够**单独或者与其他信息结合识别消者的信息**。

最高人民法院《**关于审理利用信息网络侵害人身权益民事纠纷案件适用法律若干问题的规定**》

第十二条 网络用户或者网络服务提供者利用网络公开**自然人基因信息、病历资料、健康检查资料、犯罪记录、家庭住址、私人活动**等个人私和其他个人信息，造成他人损害，被侵权人请求其承担侵权责任的，人民法院应予支持。但下列情形除夕卜：

（一）经自然人书面同意且在约定范围内公开；

（二）为促进社会公共利益且在必要范围内；

（三）学校、科研机构等基于公共利益为学术研究或者统计的目的，经自然人书面同意，且**公开的方式不足以识别特定自然人**；

（四）自然人自行在网络上公开的信息或者其他已合法公开的个人信

（五）以合法渠道获取的个人信息；

（六）法律或者行政法规另有规定。

网络用户或者网络服务提供者以违反社会公共利益、社会公德的方式公开前款第四项、第五项规定的个人信息，或者公开该信息侵害权利人值得保护的重大利益，权利人请求网络用户或者网络服务提供者承担侵权责任的，人民法院应予支持。

国家标准化管理委员会《**信息安全技术公共及商用服务信息系统个人信息保护指南**》

523 处理个人信息前要征得个人信息主体的同意，**包括默许同意或明示同意**。收集个人一般信息时，可认为个人信息主体默许同意，如果个人信息主体明确反对，要停止收集或删除个人信息；收集个人敏感信息时，要得到个人信息主体的明示同意。

&&*征信方面*：

国务院《**征信业管理条例**》

第二条 本条例所称征信业务，是指对企业、事业单位等组织（以下统称企业）的信用信息和个人的信用信息进行**采集、整理、保存、加工**并向信息使用者提供的活动。

第十三条 采集个人信息应当经信息主体本人同意，未经本人同意不得采集。但是，依照法律、行政法规规定公开的信息除外。企业的董事监事、高级管理人员与其履行职务相关的信息，不作为个人信息。

第十四条 禁止征信机构采集个人的宗教信仰、基因、指纹、血型、疾病和病史信息以及法律、行政法规规定禁止采集的其他个人信息。征信机构**不得采集个人的收入、存款、有价证券、商业保险、不动产的信息和纳税数额信息**。但是，征信机构明确告知信息主体提供该信息可能产生的不利后果，并取得其书面同意的除外。

第十五条 信息提供者向征信机构提供个人不良信息，应当事先告知信息主体本人。但是，依照法律、行政法规规定公开的不良信息除外。

人民银行《**关于做好个人征信业务准备工作的通知**》

人民银行印发《关于做好个人征信业务准备工作的通知》，要求芝麻信用管理有限公司、腾讯征信有限公司等八家机构（后附名单）做好个人征信业务的准备工作，准备时间为六个月。

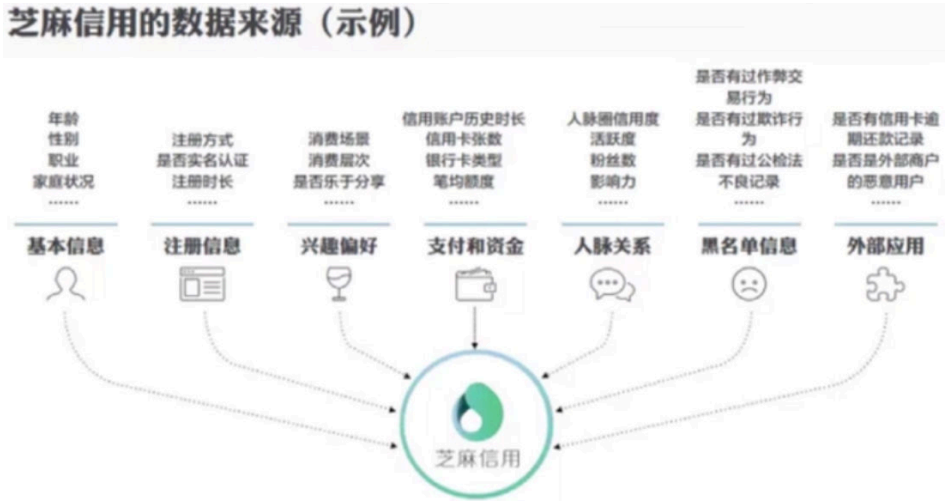
培育社会征信机构，是贯彻落实党中央、国务院关于推进社会信用体系建设、建立健全社会征信体系等一系列方针政策的重要举措，对于规范发展征信市场、服务实体经济具有积极意义。

人民银行副行长潘功胜 .

充分发挥市场在推进社会信用体系建设中的决定性作用，培育、发展中国征信市场，**积极利用互联网、大数据等新技术条件发展新业态征信**，推动征信机构加强自身信用建设，依法推进征信市场对外开放，加强征信业监管，促进征信业的规范健康发展。政府部门的职责在于，加强政务诚信建设，**推动政务信用信息尤其是负面信息的公开、推动信息系统互联互通与信用信息共享**，推动信息的应用和实现联合惩戒，推动完善信用法制体系，营造社会信用体系建设的法律、制度和政策环境，维护公平、公正的市场秩序。

•实践案例及业界实践

芝麻信用：



负面案例：

• 2014 年央视 315 晚会曝光鼎开、大唐旗下高鸿等公司存在向智能手机植入恶意程序等问题，该程序会造成恶意扣费以及泄露用户的个人信息。用户手机出现恶意扣费的情况，经调查发现，用户手机被植入两个恶意木马程序，一个可以远程安装、卸载应用软件，另一个可以获取手机中的个人信息。

• 鼎开公司称通过植入方式获利，每个月装机量达到 130 万台。而大唐电信旗下高鸿股份公司开发的大唐神器，每个月装机量达 100 万台，已经在用户手机上悄然安装 4600 万个软件。除了恶意软件安装，这些程序还获取用户手机上的 1M E1 、应用使用时间、地址等私数据。

大数据业务实践：

·一　数据收集环节（用户授权）

·二　跨行业合作（数据合法性）

·三　收费方式（效果衡量）

·四　反欺诈（身份和行为）

数据安全合规：

数据脱敏	数据整合	内部隔离

核心是关于“*可识别性*”。

3.核心风险点

• 监管（越来越严格）

·知识产权（商业秘密保护，各方数据所有权）

·数据外泄（这种纠纷会越来越多，需要建立安全感）

·用户投诉（给用户一个投诉的通道 / 机制）

·媒体曝光（315晚会等，应该提前增加透明度，与大数据算法的黑盒子等形成联合）

十、屈燕：大数据在社交媒体的应用

1.ShareThis大数据全景介绍

- Data strategy
- Data logging and collection
- Privacy, opt-in/opt-out
- Data product innovation
- Research roadmap
- Scientific oversight
- Team building
- Intellctual property

- Big Data Landscape

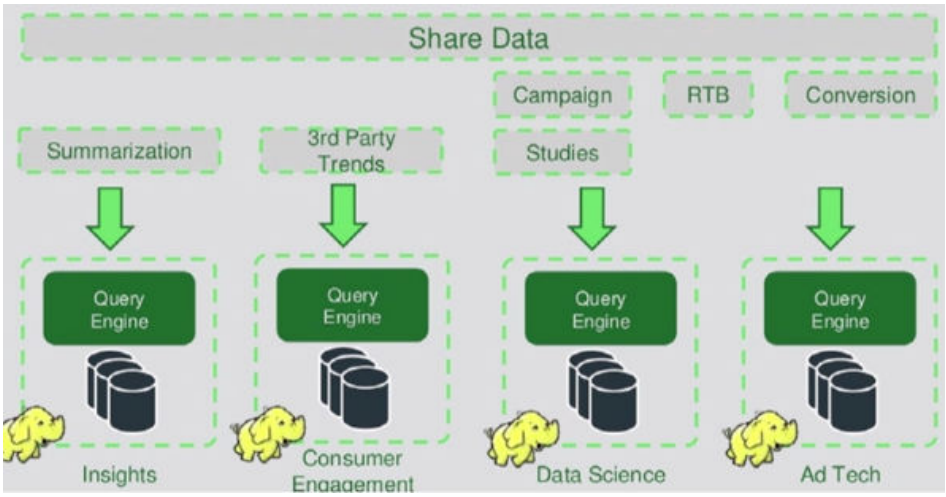
Sector	Example companies
Applications	Google, Facebook, Netflix
Analytics	Google, Tableau, Splunk
Infrastructure	Amazon, Microsoft, Google
Open source	Hortonworks, Cloudera, Spark
Data sources	Oracle Bluekai, Oracle Datalogix

主要还集中在底层部分：数据管理、收集。

We make social data actionable

2.基础设施及数据分析

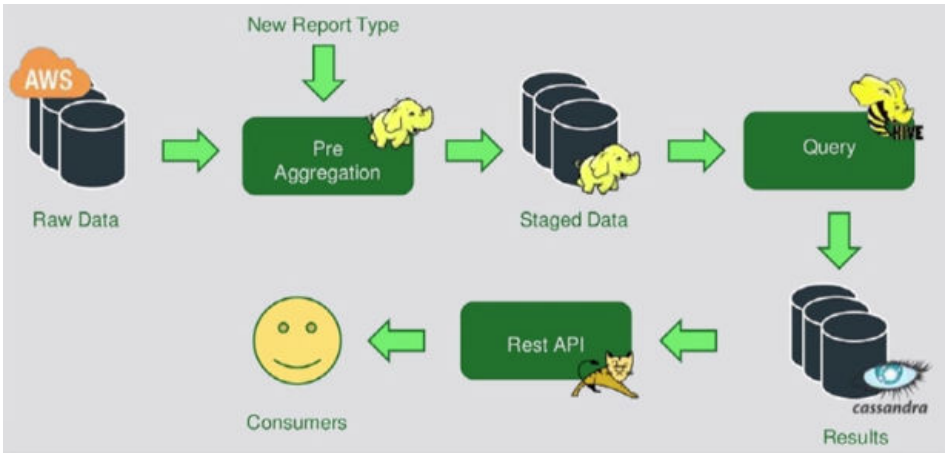
·Earlier data infrastructure



存在一些problem：

- Duplicated data processing
- Fragmented and silo-ed data
- Data inconsistencies
- Data quality issue
- Pipeline breakdowns
- Not real time

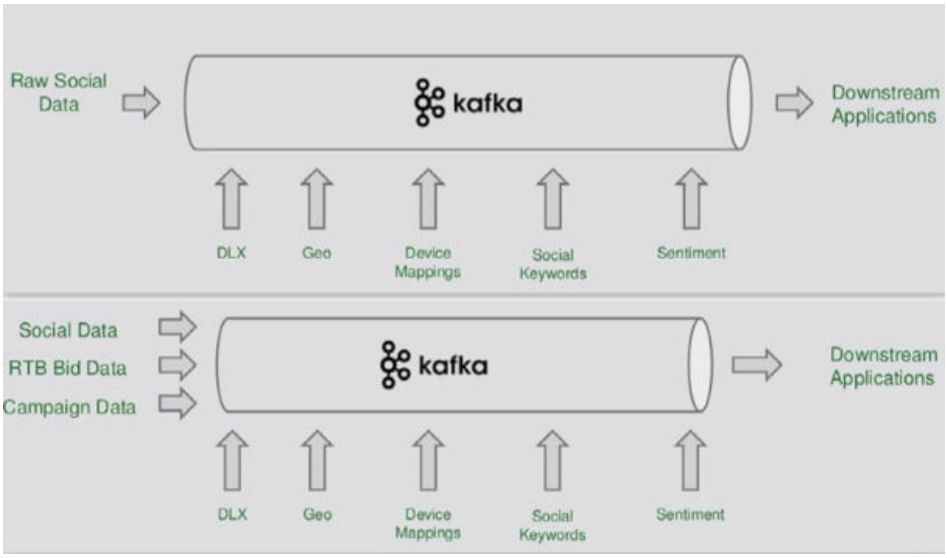
- A typical data process



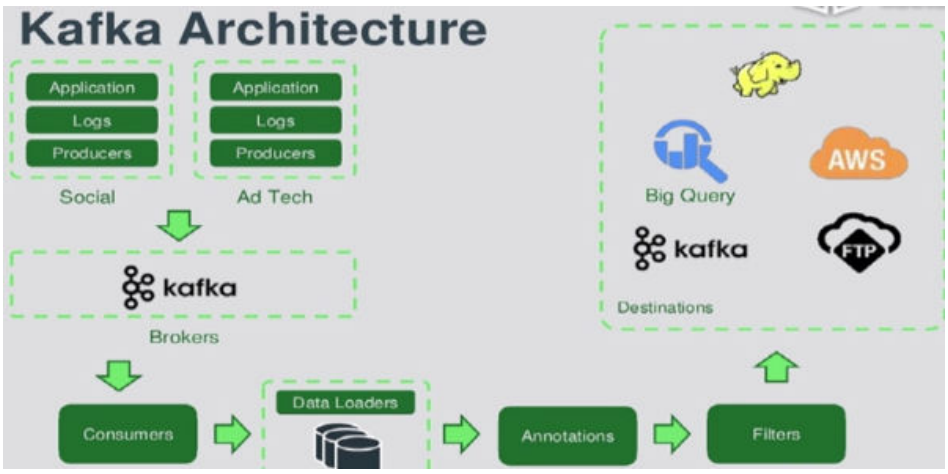
• 新的设计

Vision for new architecture

New real time data flow



得到：



……

3.预测模型及A B测试

从模型建立到应用，困难重重。

Finding the right tool

- Ease of use
- In-memory processing (fast)
- Distributed (scale)
- No barrier between modeling and deployment
- System reliability
- Accurate algorithms
- Extensive API/SDK
- Visualization (data and results)

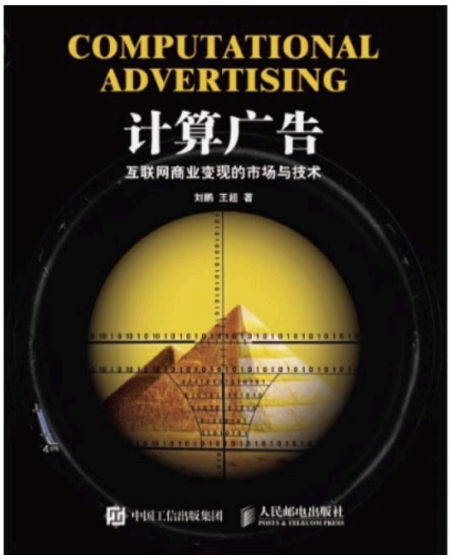
需要一个合适、高效的框架【简单、好用】。

Correlation ≠ Causation.

十一、刘鹏：互联网变现和计算广告

1.互联网与商业化

《计算广告》。



“Comp_Ad”或“计算广告”

***新的技术、产品层出不穷。

•细说互联网思维之“三个不要”：

不要钱

—免费倾销加后向变现的商业模式

—所有能够传播信息的商品，售价都会趋向其边际成本

【电影票价将趋于0】【**后端变现**】

不要脸

无底线迎合用户的产品与营销方式【跪舔】

不要命

—用**期权**和价值观让程序猿在疯狂状态下全天候工作【“996”工作模式】

&&免费产品获得的无形资产：流量，数据（规模化盈利）。

•与商业化相关的产品问题：

商业模式探索，例如：电影是一种边际成本很低，同时信专播量又很大的典型

商品。是否可自采索一种售价很低，而充分利用其信息传播能力的电影行业发行模式？

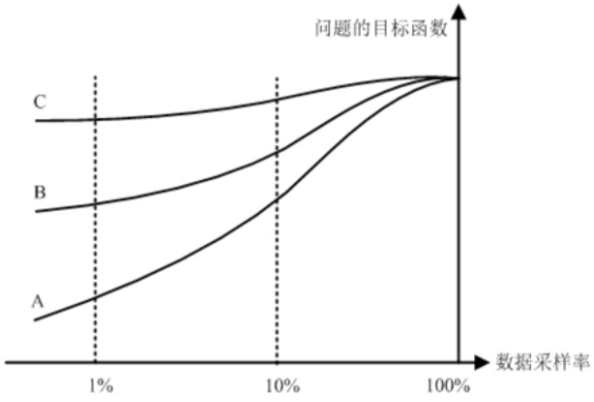
流量变现，例如：互联网电视厂商除了销售收入，还可以获得用户流量。这些流量的性质如何变现？

数据变现，例如：室内导航技术是近年来快速发展的新型互联网应用。这类产品会得十么样有价值的数据资产，又应该采用明具体的商业产品来变现？

商业产品建设和运营，例如：团购、游戏联运、返利购买这些推广模式与广告有什么内在联系？是否可以共用某些产品和技术平台？

哪些人需要了解商业化与计算广告？

2.大数据与计算广告



A：典型大数据问题，解决问题的效果随着采样率降低显著降低，例如计算广告、个性化推荐等

C：一般数据分析，非大数据问题，解决问题的效果在采样率降低时变化很小，例如各种洞察、单维度统计等

B：过渡类型问题，解决问题的效果随采样率降低温和下降，例如文本主题分析等

&&**行为数据**相比交易数据：数量大（相差2+数量级），容错性强（数据部分丢失影响小）

- 两类数据应用：洞察（宏观）与自动化（真正研究的重点）

洞察(Insight)

- **全局或局部统计性的信息（统计数据）**
- **例：财务报表、人口统计、百度迁徙地图等**
- **主要用于宏观决策支持，面向领导和运营人员**

自动化(Automation)

- **个体的行为特征信息（行为数据）**
- **例：定向广告、个人信用、企业信息等**
- **主要用于微观业务实施，面向机器和销售人员**
- **无底线迎合用户的产品与营销方式**

- 互联网变现原理

信息——>价值。

&&互联网70%到80%的收入来自于广告。【百度9成+，淘宝8成+，腾讯5成+（不包括游戏业务）】

·关于在线广告

•在线广告支撑了整个互联网行业的大半壁江山。不了解互联网广告，就不可能深入了解互联网。

•在线广告是迄今为止，大数据领域唯一形成**规模化营收**的应用。

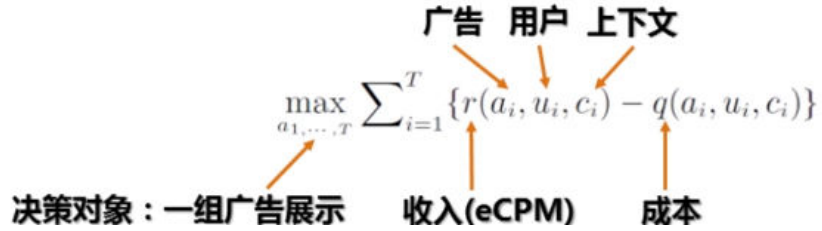
•在线广告是结合了计算技术、心理学、经济学、营销学等的综合应用。

&&发展、增长速度飞快。

***品牌广告（离线转换率，**拉动利润率**，长期），**效果广告**（互联网专长，销量+利润，短期内）。

3.计算广告介绍

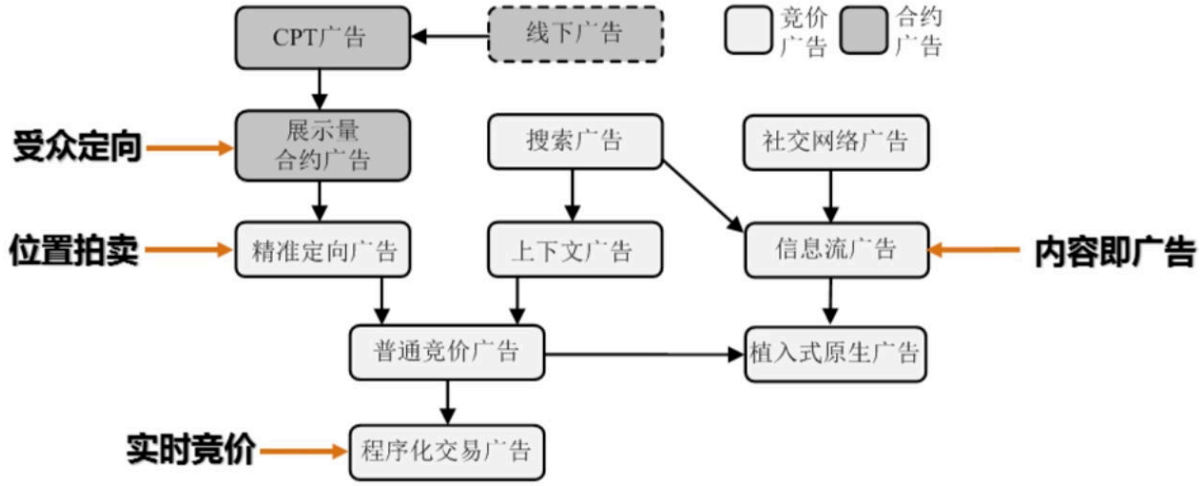
•核心挑战：为一系列用户与环境的组合，找到最合适的广告投放策略以优化整体广告活动的利润。



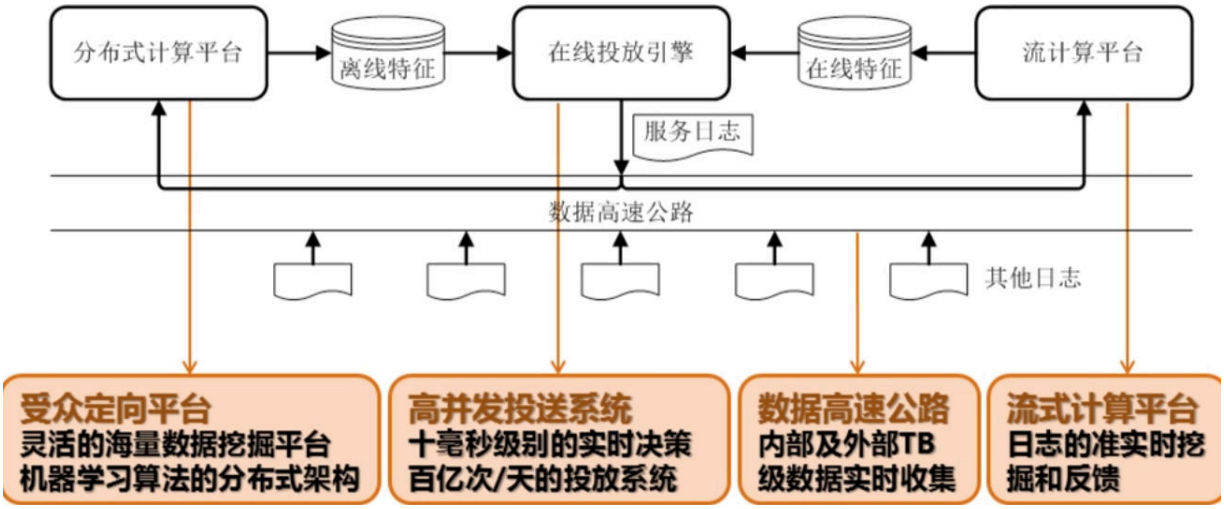
&&广告和个人推荐的差别：**constrain**，有广告主的投放要求、金额，使优化问题变得复杂。

·广告收入的分解

·在线广告产品历程



·个性化系统一般框架



4.数据交易

- *行为数据交易三定律*

一 行为数据**只能交易**，不能交换或共享（交换往往只能有高层次的交换，即投资关系）

二 只有**按效果**而非购买量付费，才有足够的需求

三 同一数据被**越多人使用**，价值就变得越低

关于第一定律——

&&为什么大公司不把数据共享出来？

**你见过大公司把钱共享出来吗？

**短时的贴补性共享是可行的

&&政府数据是可以共享的，这本质上是转移支付

关于第二定律——

&&数据交易怎么做？

**数据传输附着在实时竞价过程中，无额外开销

**需求方能自由选择需要的部分人群数据，并且按照实际的广告展示付费【定价权向需求方转移】

关于第三定律——

&&如何给数据定价？

**市场化的定价方式是唯一选择

**目前数据的价值是被低估的（上页的交易方式并未限制数据供给次数；这间接地抬高了流量价格，而低估了数据价格）

&&能否采用竞价的交易方式？

**不限量供应的商品，是无法竞价的（实现数据的部分交易的基础上）

**数据的限量供应怎么做？

&&大数据隐私问题严重吗？

**隐私问题基本原则——

——A29：欧盟负责隐私保护条例制定的委员会

——A29原则

·Personal Identiflable Information（PII）不能使用

• 用户可以要求系统停止记录和使用自己的行为数据

·不能长期保存和使用用户的行为数据

Quasi-identifier与K-anonymity（*熟人间***的隐私）

——Quasi-identifie：朝阳区，35岁，在360上班

——K-anonymity：北京市，30-40岁，互联网行业

****大数据隐私问题比想象的更严重

**稀疏行为数据的新挑战

——从一个人观影或购物记录，能否反推他是谁？

——实际案例：Netflix推荐大赛，有人从数据集里发现了自己的同事是同性恋【偶然/不计代价】

——理论研究：Robust De-anonymization of Large Sparse Datasets

**深度个性化系统也有隐私安全风险！

——相关研究课题是差分隐私（Differential Privacy）

**隐私是大数据头上的达摩克利斯之剑

十二、毛波：阿里全息大数据构建与应用

1.数据的进化历程

大数据：数量庞大，彼此关联，场景多样，持续更新。【4V理论，创造价值】

数据进化路线（计算能力 + 实时性）：

文件	数据库	数据库集群（BI/CRM数据仓库）	分布式数据平台（云计算端应用）
不定期	离线定时	准实时（小时、分钟）	实时（秒，毫秒）

阿里数据进化历程：

数据分散	部门及数据中心	公司数据中心	聚合分享，数据生态

数据观点：

以控制为出发点的IT时代正在走向激活生产力为目的的DT（data technology）数据时代，这不仅仅是技术的升级，更是**思想意识的巨大变革**。（by 马云）

数据是未来最重要的生产要素。

阿里是一家数据公司。

·通过自身业务沉淀数据（淘宝、天猫、聚划算、支付宝）

·提供云计算基础设施，吸引第三方应用沉淀数据（阿里云）

·投资并购补充数据（高德、微博、快的、阿里影业、阿里健康）

·运营数据、交换数据（Alimama DMP【达磨盘】）

2.阿里DMP平台介绍（运营数据、交换数据）

**Alimama DMP【构建全息大数据】

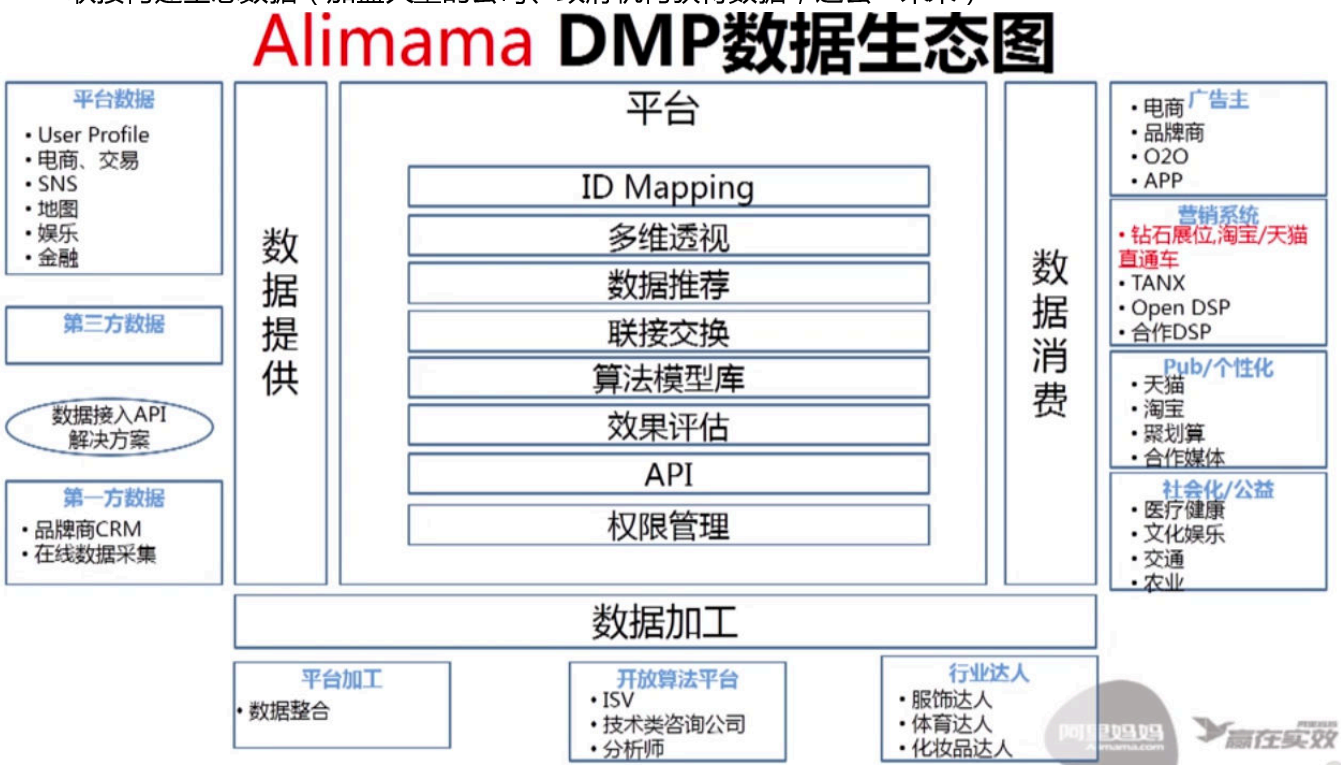
·收集整合全域数据

·撮合数据需求方、数据提供方、数据加工坊对接

·还原生活场景---追溯过去、还原未来、预测将来

·运营、交换数据【在保护安全、隐私的前提下推动数据联动、流动和应用】

·联接构建全息数据（加盟大量的公司、政府机构获得数据，过去 + 未来）



3.核心技术及案例

• 核心技术

·ID-Mapping【异构数据跨屏跨渠道整合】

·多维数据随意组合透视

• 海量数据深度挖掘

·数据安全与控制

·Lookalike人群扩展

·实时人群定向与反馈

·全链路效果评估

·数据API（易用性）

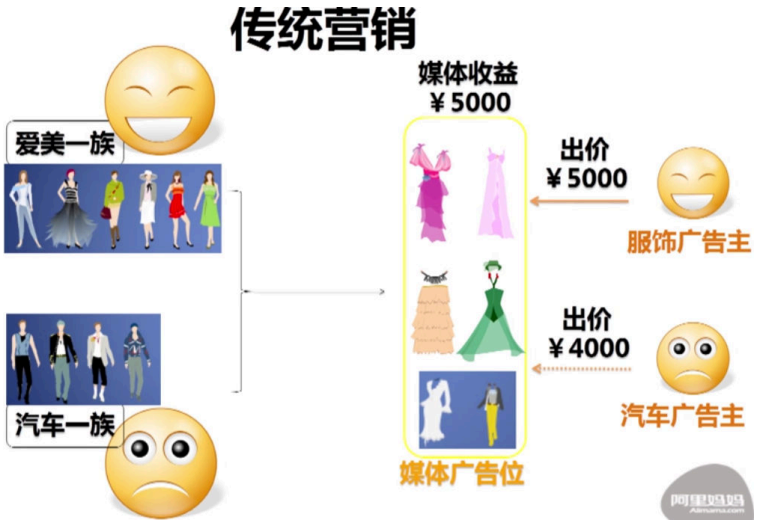
·案例

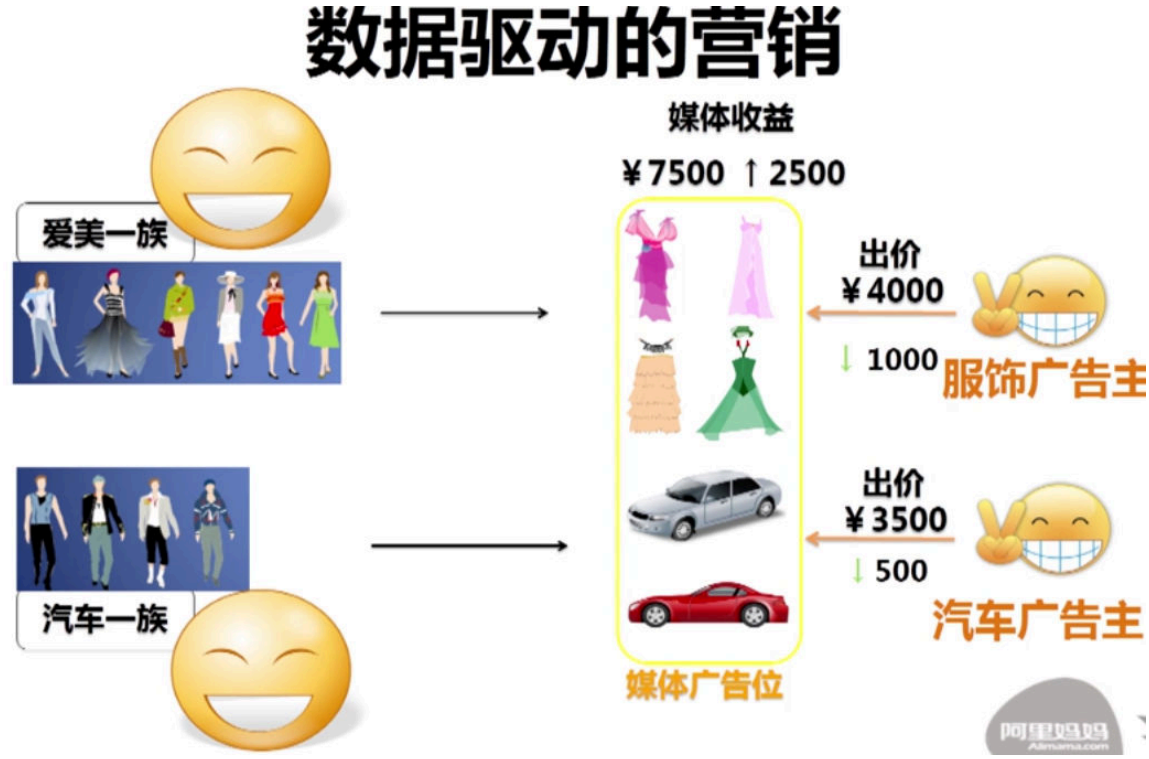
·looklike模型——精准用户拓展

·标签深度挖掘——知识图谱

·第三方开发市场（基于全容器云环境和数据脱敏方案，提供算法开发环境和达人开发环境，提供个性化服务）

4.数据应用





·数据驱动的金融：

·阿里小额贷款

·通过数据计算信用分

·通过信用取得贷款，无需提供抵押品或第三方担保

·数据决定成败

·余额宝

·打破消费和理财界限【但实际理财收益不高】

·吸引用户、产生数据、驱动新业务

·芝麻信用

·P2P金融平台

·信用、风控

·数据帮助“在别处赚钱”

·“羊毛出在猪身上”——在另外的地方能够赚回来

·滴滴、快的打车软件

·通过补贴吸引大量用户使用

·积累大量用户数据，在精准营销、创新业务中获利

·培养用户支付习惯，推出自有专车服务，或在支付端获利

·互联网盒子、路由器

·硬件低价或免费

·积累用户使用数据

• 可穿戴设备

·小米手环

·Apple watch

• 正在生成的未来

·互联网+（通过互联网、数据激活传统行业）

·个性化医疗（日常健康数据监测，个性化治疗；基因测序）

• C2B（消费者反向定制生产，降低商家试错成本）

·物联网

十三、韩定一：在线营销中的竞价机制和数据价值

1.在线营销

示例：展示Display（猜测式）；搜索Search（关键词相关）。

匹配多方需求：

·卖家（品牌印象、客户关系维护、成本转化）

·买家（快速便捷的购物体验、了解有些什么新货、逛的需求）

·媒体（流量变现的模式）

·平台（撮合多方利益、挖掘数据价值）

营销效果指标：

·触达类指标

·PV展现量【Page View】多少页面

·UV用户量【User View】多少人

·转化类指标

·CTR点击率【Click-through Rate】

·CVR转化率【Conversion Rate】

·金额类指标

·成交额【Deal Amount】

·ROI投资回报率【Return of Investment】

·时间对金额类指标的影响【1天/3天/7天/15天/30天】

·效果归隐的计算差异【多个计划触达导致的转化如何归因？】

影响营销效果的关键因素：

·“我知道我的广告费浪费了一半，但是却不知道哪一半浪费了”——奥美广告创始人

·找对人、找对时间、说对话（人群；季节、早中晚、用户决定前；创意位置、创意内容、创意形式、交互）

在线营销的参与角色：

·DSP【需求方平台】

·SSP【供应方平台】媒体方面，有流量

·AD Exchange【连接SSP和DSP的桥梁】

实时竞价Real Time Bidding【用户访问页面后的50ms内产生】

结算方式：（流量方和商家的利益冲突）

·CPM【按千次展现收费】

·CPC【按点击收费】

• CPC/CPA【按销售/行为收费】

2.竞价机制

GSP竞价机制（广义第二位价格）————相对的GFP（first）

·价高者得流量（预算决定流量多少）

·结算价 = 第二位价格+1（下一位决定价格）

3.数据价值

• 数据如何在营销中起作用

·合理描述商家和卖家之间的关系

·商品偏好（类目意图，品牌偏好，属性偏好）

·行为类别（浏览，收藏，搜索，比价，购物车，购买）

·兴趣（长期兴趣，短期兴趣）

·时间

·地点

·天气

·节日

……

·相对于传统的电视、广播和纸媒，在线营销具有人群精准投放的控制能力

·对人群刻画越精准，CTR越高，意味着商家的营销成本越低

·CTR预估（数据，模型）

·商家的投放建议（位置，人群，出价）

·市场划分机制（类目数量合理）

—— 并不是越多越好

—— 精准 ≠ 精细到个体【分得过细，就失去决策的价值】

—— 属性值未必越真实越好【用户信息不一定与真实的使用者状态符合】

DSP需要合理的标签详细程度。

十四、龚笔宏：大数据在工业界中的经典案例分享

1.竞价排名搜索

***搜索的营销收入——占很大的比重。

·不同角色所关心的问题

·推广者的行为

·买词【关键字】

·竞价【给出预测】『二阶竞价 ——> 广义二阶竞价（按广告预期收益再排名）』

• 为点击付费

·搜索引擎的行为

·Query分析

·展示搜索结果 + 推广结果

·计费

·用户的行为：搜索 ——> 点击搜索结果

2.主要技术问题

- Query
  - Query analysis
  - query expansion&weighting
  - query–doc matching
- AntiFraud
- Ranking
  - Ctr prediction
  - roi prediction
  - auction
- 系统
  - 海量数据处理
  - 高并发&高实时性系统
- 创意
- For advertiser:
  - 方案推荐
  - 预估
  - 整账户优化
  - 预算分配
  - 平滑

3.点击率预测概述（一个经典的机器学习问题）

• 点击率预测（ctr prediction）

·创意被电击的概率

·在搜索推广下，即P（click|query, user）

·ctr = click num / impression num

·在展示推广下的点击率预测【频度、历史、多样性、新颖性】

4.点击率预测实践

• 基本的预测方法：基于历史数据预测【经典解法】

·数据	·特征	·训练	·校验	·预测

·实际过程中的问题

·数据【位置的问题，数据的稀疏，数据的清洗，数据的历史性】

·特征【常用的特征，特征质量的评估，工程的问题（迭代速度的问题）】

·训练【并行化问题，性能&效果之间的tradeoff，可解释性&复杂性的tradeoff】

• 校验【baseline，offline & online】

• 预测【性能，监控，稳定性】

十五、陈辉：数据驱动营销

1.什么是数据驱动营销

• 业务理解（Bussiness Understanding）

·数据理解（Data Understanding）

·数据准备（Data Preparation）【数据统一格式】

·建模（Modeling）

·评估（Evaluation）

·部署（Deployment）

——————Data-driven Marketing or **DDM**

2.LTV留存分析

LTV = Lifetime Value，一个用户**一生**中给某个商家带来的价值

LTV=创造的价值 - 营销成本

————在电商环境下，“一生”通常用一个较长的时间代替，比如首次成为客户的一年内。

LTV是营销决策的终极指标

拉新：

- 当一个潜在客户一年内创造的价值高于转化为客户的营销成本，就应该去吸引这个客户；否则不吸引
- 当拉新预算有限时应当优先吸引LTV较高的客户

维旧：

- 当提升一个老顾客忠诚度的成本低于他的忠诚度提高后带来的价值增量，就应当去提升这个客户；否则不提升
- 当挽留一个老客户的成本高于客户继续能够创造的价值，应当放任自流；否则应当挽留

LTV的困难：

·很难预测用户长期行为，短期预测也需要大量数据

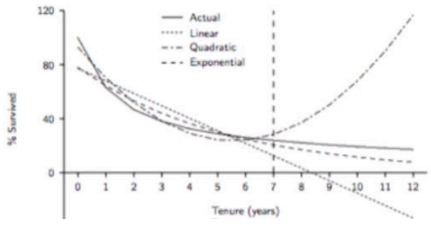
———————— 简化模型 + 多个模型 同时预测降低误差

常见的LTV模型：

极简模型：　　每星期平均收益 * 50

简单模型：　　$\frac{\text{第一个星期平均收益}}{1 - \text{每星期流失率}}$

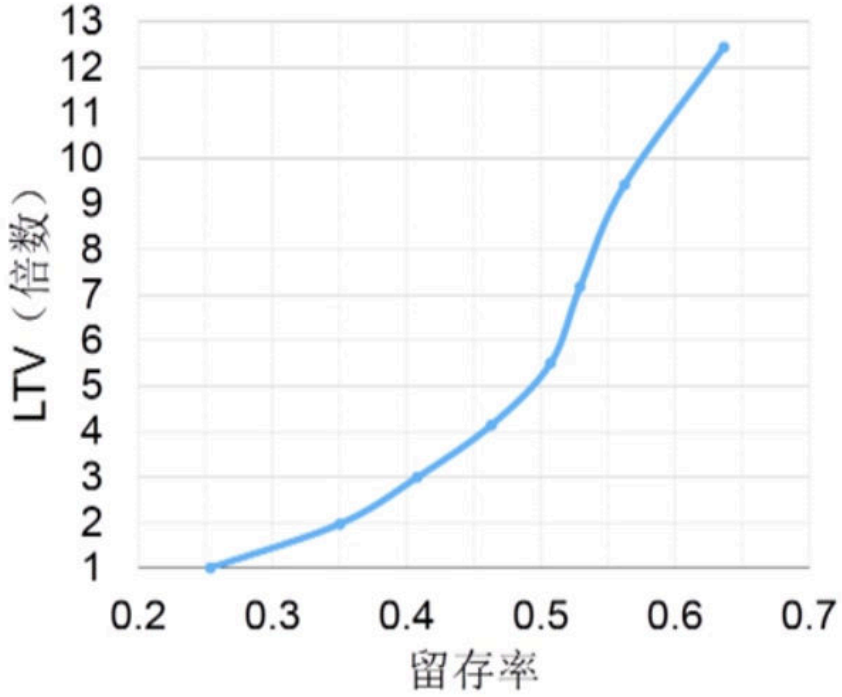
复杂模型：



预估用户什么时候会离开
shifted beta-geometric model

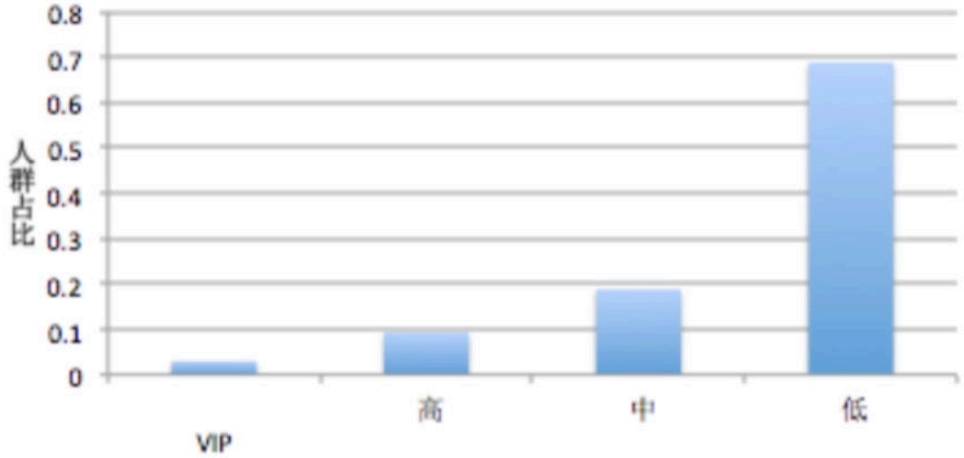
或者用统计学来预测……留存率、新老客户的识别……

·留存率 VS. LTV



应当重视品牌建设。

3.RFM会员体系



R = 最近一次购买时间
F = 平均购买频率
M = 平均购买金额

VIP、高、中、低人群定义

1.r和f按中位数分两层，f高于中位数打分2,低于中位数打分1,r小于中位数打分2,r大于中位数打分1;m按金额前5%、5% - 20%、20% + 分三层，依次打分3、2、1
2.rfm权重根据经验值选取：m=0.5 r=0.3 f=0.2
3.总分分层依据：
VIP(2.5):rfm均高
高(2-2.3):m高但不活跃 + m中但很活跃
中(1.5-1.8):m中不活跃 + m低但很活跃
低(1-1.3):m低不活跃

具体的设定要考虑营销场景

数据观察到的结论

留存率： VIP > 高 > 中 > 低
挽留的营销成本：VIP < 高 < 中 < 低
回报率：VIP > 高 > 中 > 低

针对不同的客户应当有针对性营销：

- VIP和高价值客户应当以维系品牌自豪感为主
- 中低价值客户应当以关怀为主，提升品牌忠诚度

4.消费者微群画像

- 用户基础属性

A.性别：决定了客户的基本需求，例如女生爱买衣服、饰品和护肤品，男生则喜欢网游、3C数码类产品等等。

B.年龄：决定了客户的生命周期和家庭情况，典型的例子是在客户进入青年阶段后，会开始结婚、生育，会购买对应的婴幼儿用品等。

C.城市层级：城市的发达程度通常决定了该地区客户的受教育水平及生活品质。

D.网购深度：网购深度由顾客在限定期间内的购买次数来定义。可以说最常购买的顾客，也是对淘宝、支付宝接受度最高、满意度最高、熟练程度最高的顾客。

E.消费档次：是对客户消费能力的综合性衡量。

·用户基础属性集合生成微群画像

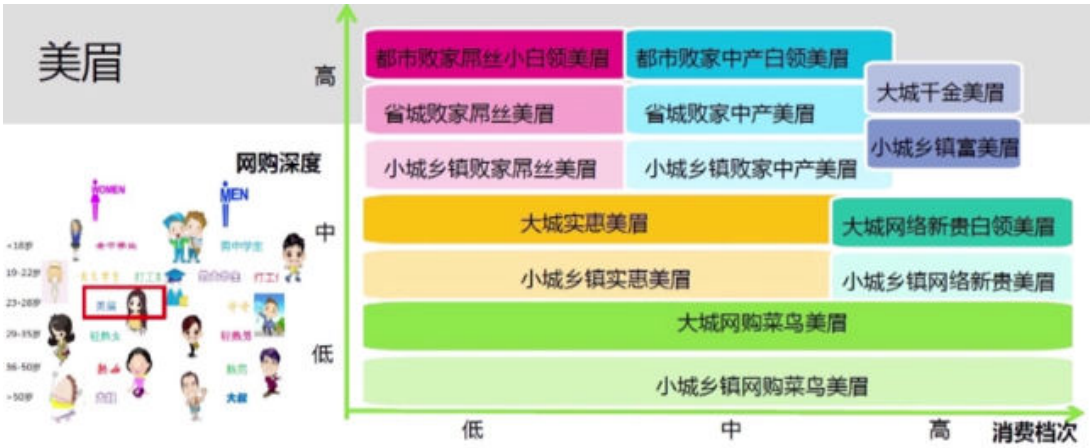
A.先验分群：使用先验分群变量将人群切分成最细颗粒度的基础群。

B.聚类分析：使用Weighted-Kmeans对基础群进行聚类，合并为若干个聚类模型群。

C.人工调优：结合商业判断，进一步细分或合并，同时利用身份职业对分群进行补充。

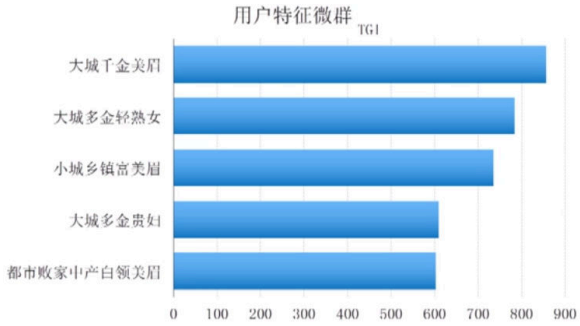
D.微群画像：对最终微群进行详细的特征描绘。

例：



……

人群画像：微群



TGI定义为

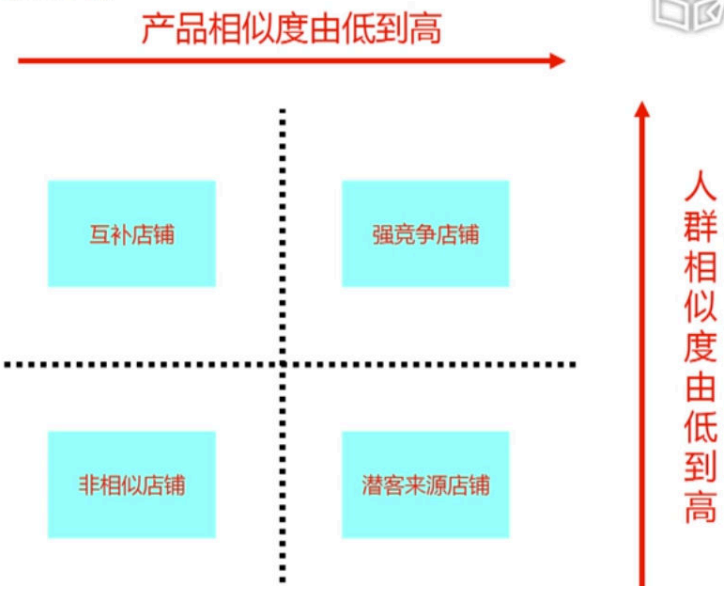
$$\frac{\text{该类人在VIP客户中的占比}}{\text{该类人在全网客户中的占比}} \times 100$$

解决问题
- 商家更好理解自己的消费者
- 有针对性营销，效果更好

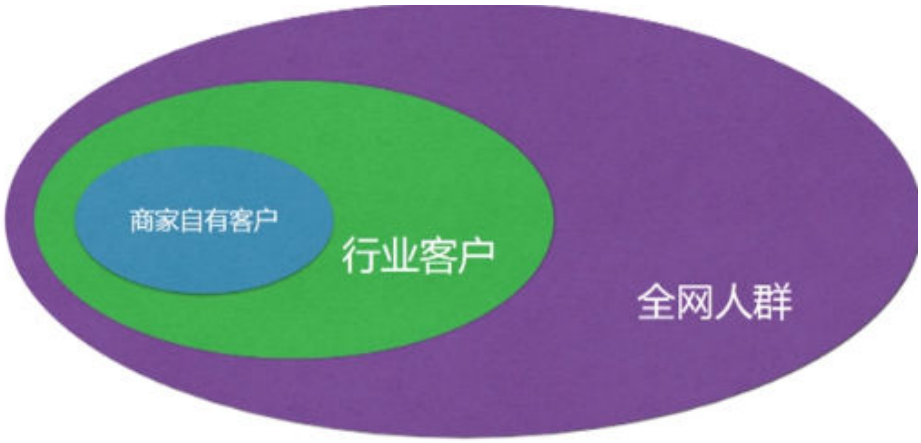
5.渠道倾向性分析及行业竞争分析

***根据渠道倾向性在不同渠道上触达不同的人。

店铺竞争关系分析



6.机器学习潜客挖掘模型

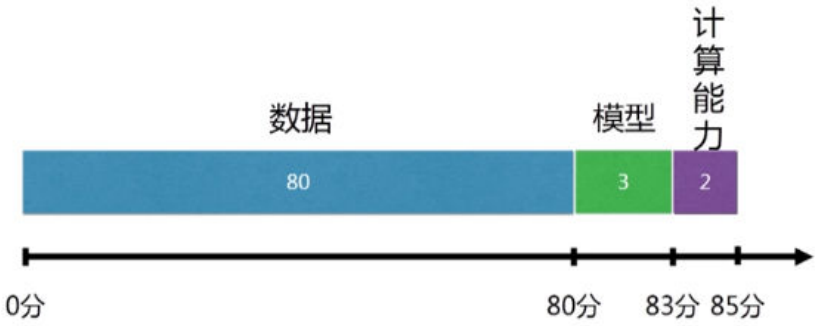


几十万人　　几千万人　　几亿量级

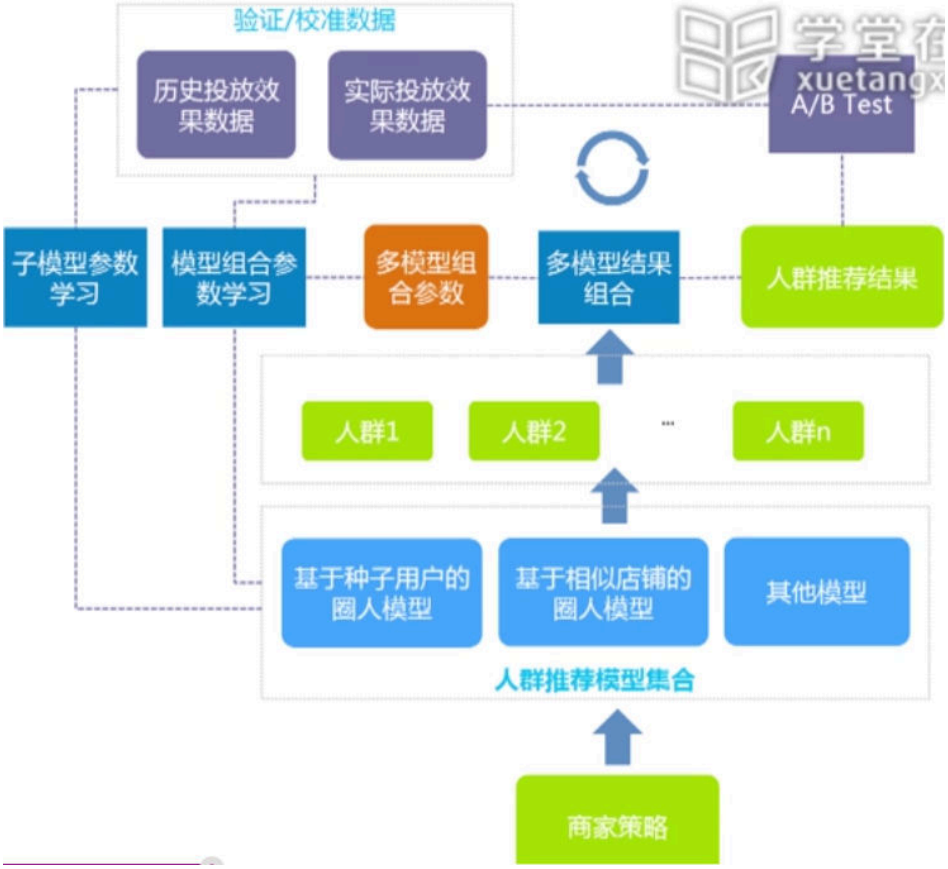
潜客挖掘技术哪家强？

阿里的分数：

计算

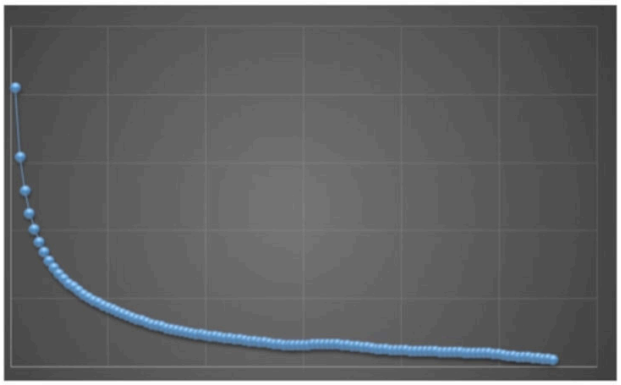


潜客挖掘系统架构：



- 潜客挖掘的机器学习模型：
  - ·LR模型：直接从用户的交易历史构建特征集合
  - ·GBDT模型：性能比较稳定，多数情况下能取得较好效果
  - ·遗忘模型：考虑用户对品牌印象的时间衰减

————遗忘模型：



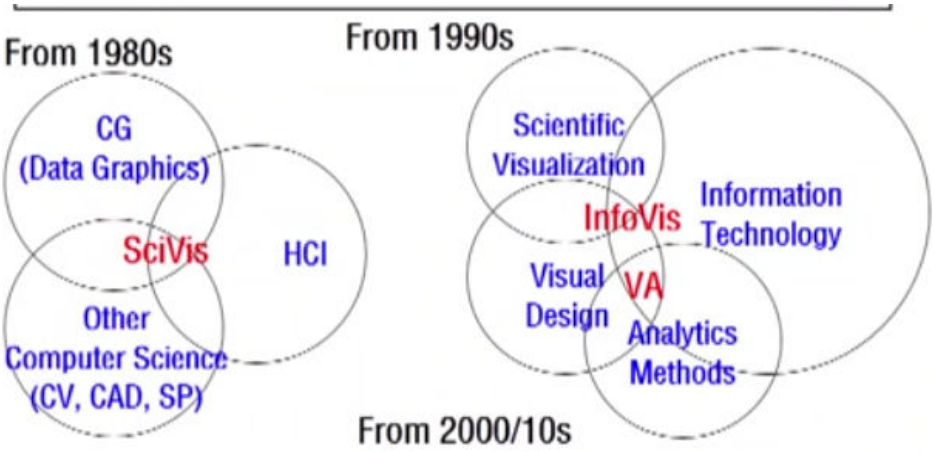
CTR从2%提升到12%

$$f(x) = a * x^b + c$$

十六、时磊：大数据网络可视化

1.什么是可视化

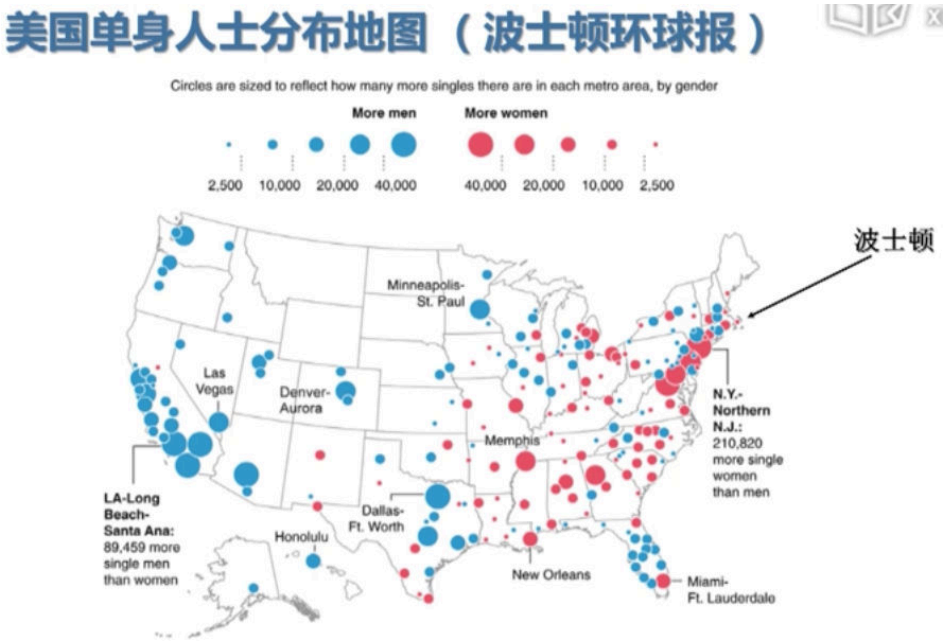
（信息）可视化：**基于计算机的手段，交互性的，目的为了观察大数据的全貌。**



***抽象数据的可视化。

VA：可视分析。

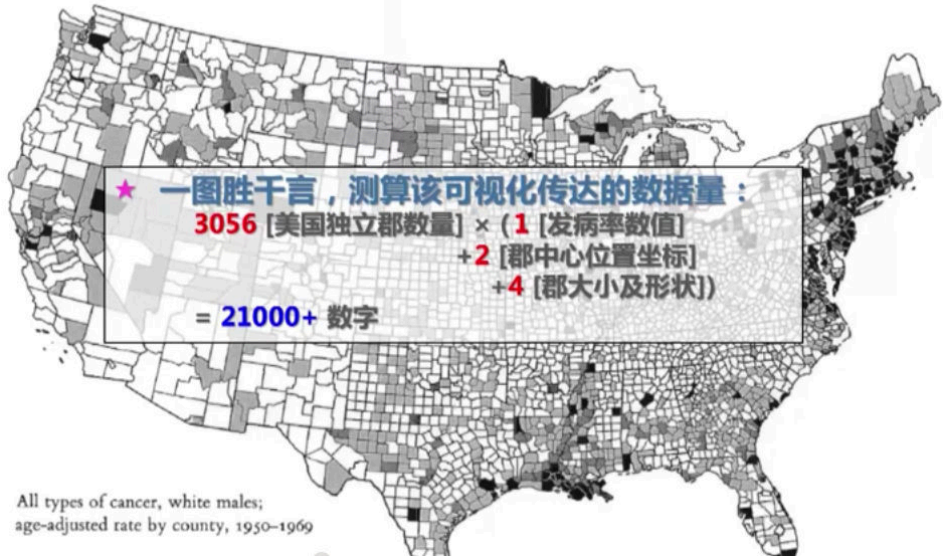
————美国单身人士分布地图。【差值】



可视化具有主观因素——用于传达制作作者的一些想法。

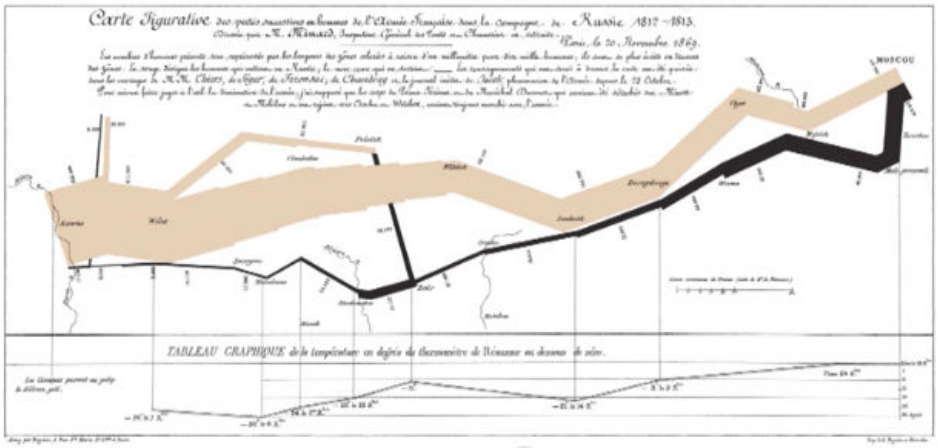
————美国癌症发病率分布图。【一图胜千言】

美国癌症发病率分布图

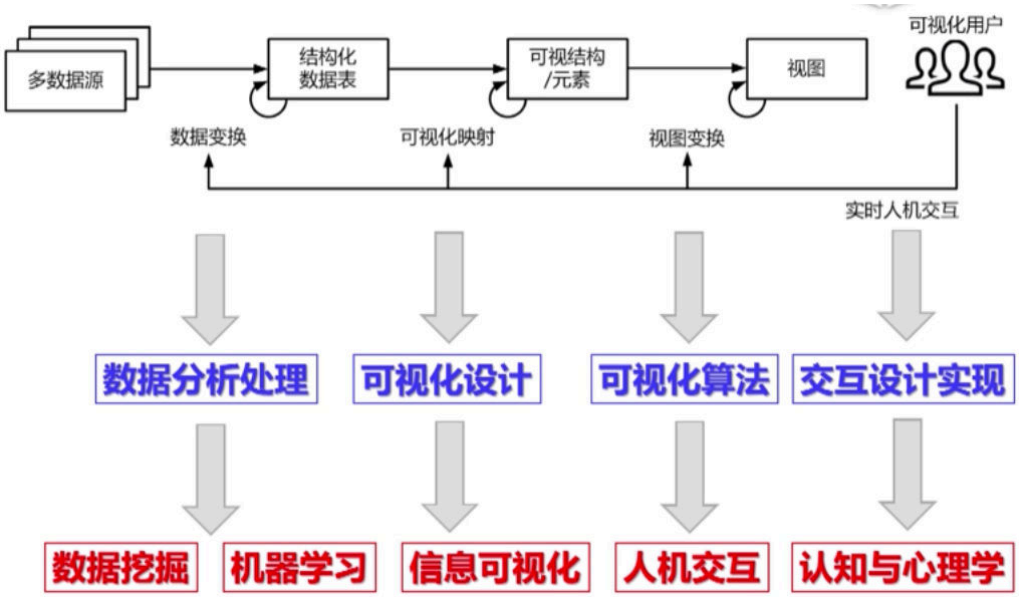


————拿破仑入侵莫斯科态势图。

• 拿破仑入侵莫斯科态势图, 1812



• 信息可视化参考模型与流水线：



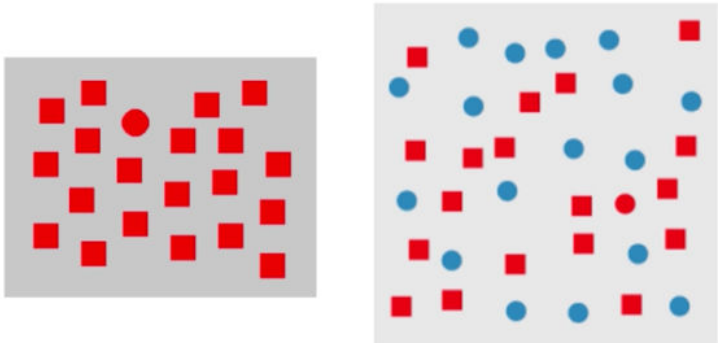
&&信息可视化挑战：人类世界感觉机制。【硬阶段——>软阶段】

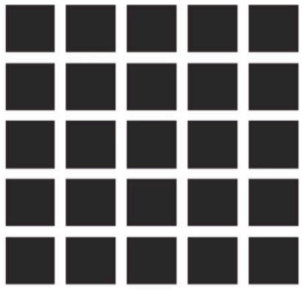
小测试：下图有多少个“3”？

小测试：下图有多少个“3”？

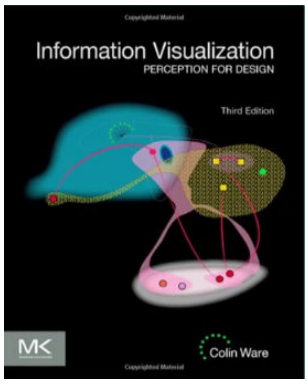
Pre-attentive Perception

小测试：找到下图的红色圆形



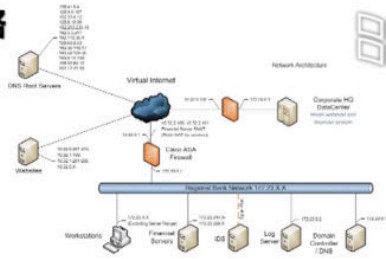


推荐书籍：



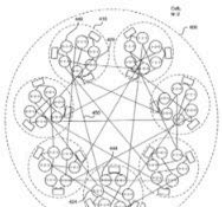
2.网络可视化

我们认识的网络

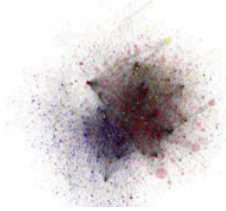


计算机网络

服务器互联网络

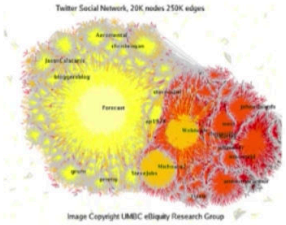


互联网自治系统网络



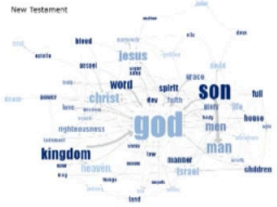
• 新型信息网络：

社会网络

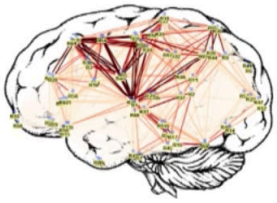


人口迁移网络

文本网络



人脑神经网络

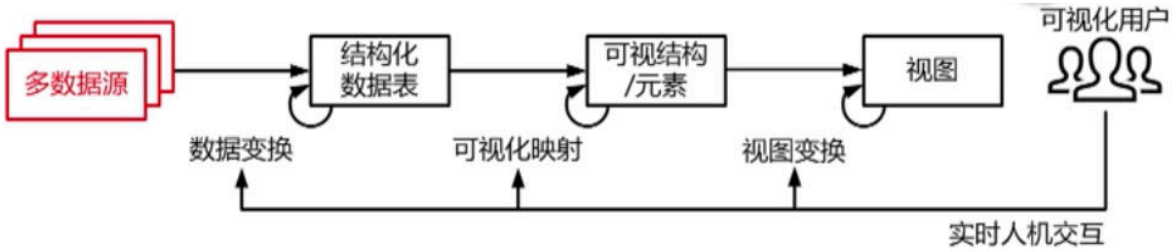


• *什么是网络数据及网络数据可视化？*

·节点（用户）+ 边（关联）

·无向图——>有向图——>有权图

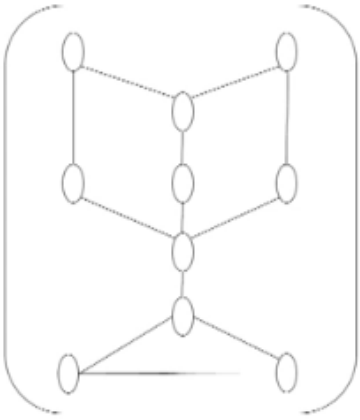
&&网络可视化流水线：



多源、异构数据
获取、清洗、融合

• 网络可视化核心难点：**图布局**

质量差的布局　　质量好的布局



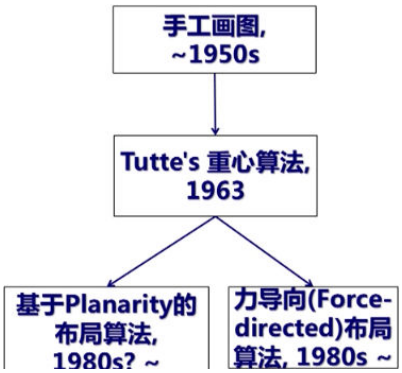
高质量图布局标准：

Purchase et al., 1995, 1997:

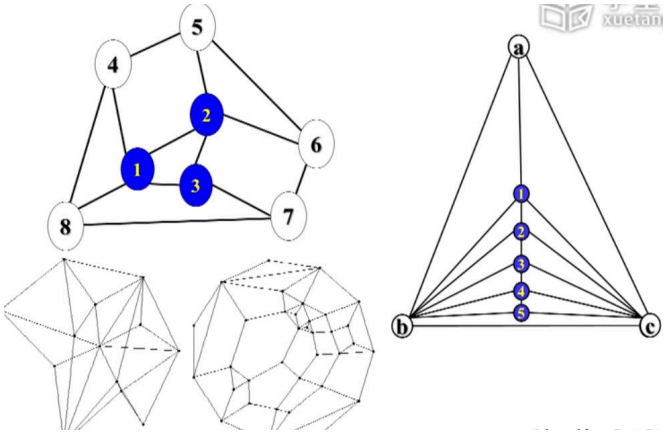
最重要且与用户理解关联性最高的图布局标准：

边交叉数目 (edge crossings)

- 图布局算法：



- Tutte's重心算法



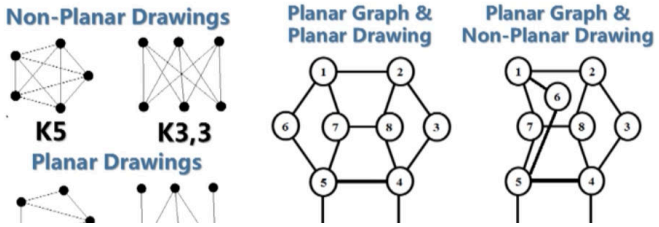
·基于Planarity的布局算法

首要准则(限制条件): 最小化边交叉数目

–Planarity图测试

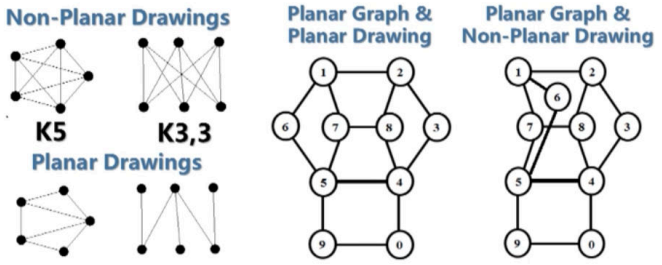
- ***最优算法复杂度：O(n)***
- **简单方法: 测试是否存在 K5 or K3,3 子图**

–Planar Graph 和 Planar Drawing



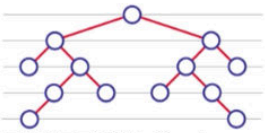
首要准则(限制条件): 最小化边交叉数目

- –Planarity图测试
  - *最优算法复杂度：O(n)*
  - 简单方法: 测试是否存在 K5 or K3,3 子图
- –Planar Graph 和 Planar Drawing

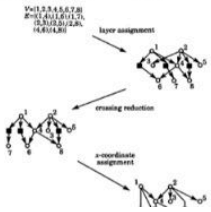


……

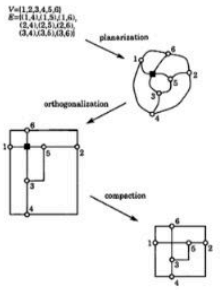
树布局: Reingold-Tilford 1983



层次布局算法: Sugiyama et al. 1981



- 正交布局算法: 拓扑-形状-度量布局框架



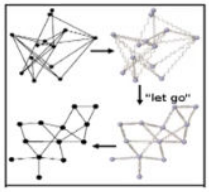
……

·基于力导向模型

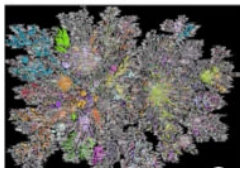
力导向图布局算法

- –主要用于无向图
- –模拟物理系统平衡态
- –可近似达到对称/均一图布局
- –力导向布局算法分两类：
  - 弹簧+电荷模型, by Eades, Fructerman et.al (改进)
  - 基于图理论距离及最优化的模型, by Kamada and Kawai (stress model)
- –针对大型网络/图数据 (>10^5节点)
  - 多层/多尺度布局算法
  - 快速力导向模拟算法

基于弹簧+电荷模型的力导向算法示意图



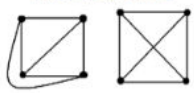
多尺度力导向布局算法



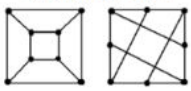
·其他图布局标准

- 除边交叉标准外：
  - –最小化/均一化边弯曲度
  - –最大化图对称性
  - –最大化边连接角度
  - –最小化/均一化边长度
  - –最小化长宽比（接近1）
- 难以同时优化所有图布局标准！
  - –折中考虑
  - –算法复杂度因素
    - 最小化边交叉问题是NP难问题
    - 在正交planar图中最小化/均一化边弯曲度是NP难问题

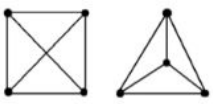
包含弯曲边和不包含弯曲边的图



对称图和非对称图

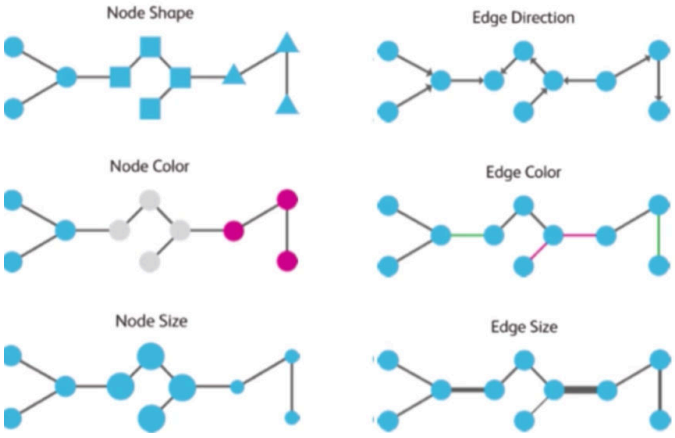


边交叉与对称性的折中



·可视化方法

·可视通道



·社会网络典型配置

基础可视隐喻 (Metaphor): 节点-边图

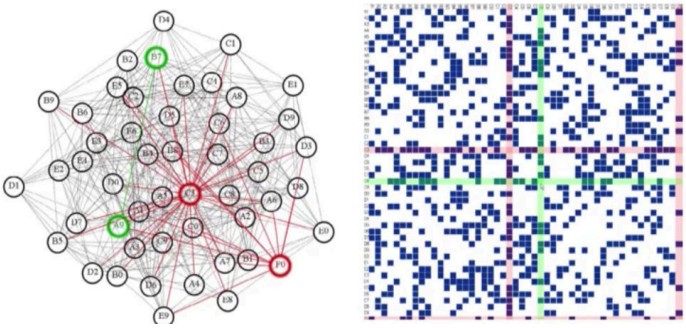
- 节点~ 实体 (用户)
  - 可视通道: 色彩, 透明度, 图表类型, 标签内容等.
- 边~ 关系数据 (好友关系, 内容流动等)
  - 可视通道: 边粗度, 颜色, 方向, 标签内容等.

高级可视隐喻

- 阴影/轮廓线~ 社区 (聚类)

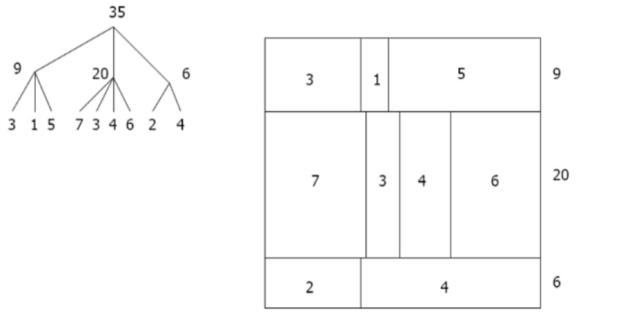
典型社会网络可视化

·矩阵可视化【非节点-边图可视化】



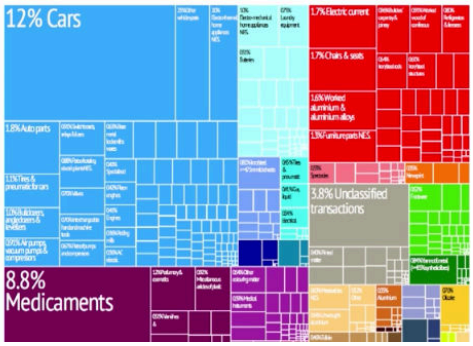
·树图可视化【非节点-边图可视化】

树图(TreeMap) by Shneiderman and Johnson, Vis '91



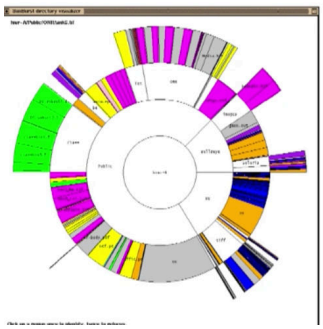
采用面积大小表示叶子节点上数值属性

其他可视通道:
– 颜色
– 标签

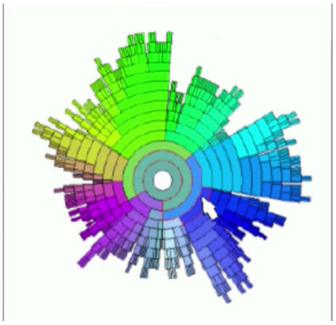


·辐射状树图可视化【非节点-边图可视化】

SunBurst



InterRing



3.大数据带来的新挑战

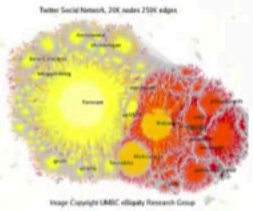
• 网络可视化技术概览

映射、布局算法	可视化系统	可视化产品

• 大数据网络可视化的新挑战

- **大数据网络特征(Big Networks)**
  - 规模大 (Large): >>10^6 节点(e.g., 3亿微博用户)
  - 多变量、异构 (Multivariate, Heterogeneous): 节点、边上带有高维度丰富属性 (例如用户档案及行为)
  - 动态性 (Dynamic): 虽时间变化 (增长/萎缩中的好友网络)
  - 数据及价值稀疏性 (Sparse): 较低的#边数/#节点数 比例 (1~10, 在常见的大规模网络上)

- **技术挑战**
  - 规模 (Volume)
    - 难以获取/存储 (10^8, 超出内存容量)
    - 难以计算(布局) (10^5 ~ 10^6, 复杂度O(N^2))
    - 难以感知、理解 (10^2 ~ 10^3 , 可视化局限性)
  - 类型 (Variety)
    - 难以计算 (异构布局)
    - 难以分析 (高维度变量关联关系)
  - 速度 (Velocity)
    - 难以计算 (布局稳定)
    - 难以展示 (2D空间, >=3D数据)
  - 价值 (Value)
    - 难以感知、理解 (可视化的复杂性)
    - 难以可视分析 (单纯手工提取价值)



·方法分类

·图布局

·数据变换

·图挖掘

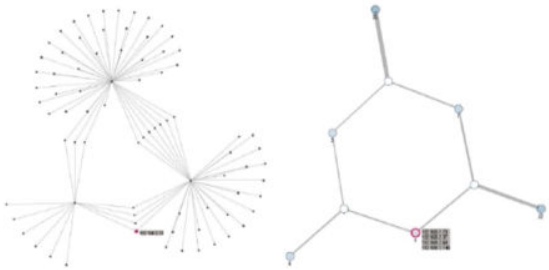
·图自然交互

4.大数据网络可视化的若干案例

• **基于图压缩的可视化方法**

图信息压缩后无损失

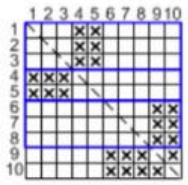
- 拓扑忠实度
- 连通性
- 最短路径
- 节点邻接性

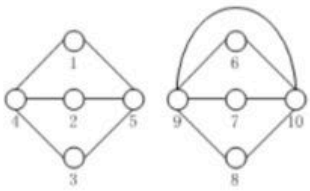


• 压缩算法

基于邻居向量的对称节点 组合方法 (SNG)

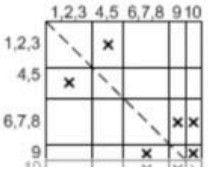
- 确定性，单次处理
- 无参数
- 复杂度: $O(E)$
- 可扩展性：有向、有权图

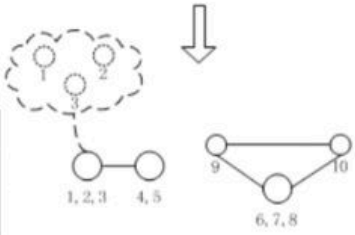




近似压缩算法

- Jaccard 近似度
- 单一参数
- 提供更高的压缩率
- 贪婪算法复杂度：$O(N^2D)$
- 优化算法复杂度：$O(kN)$

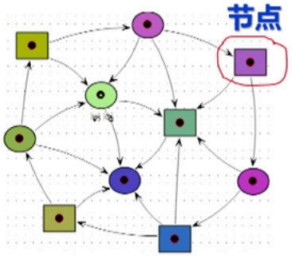




• **洋葱图多维/异构网络可视化**

·多维网络数据

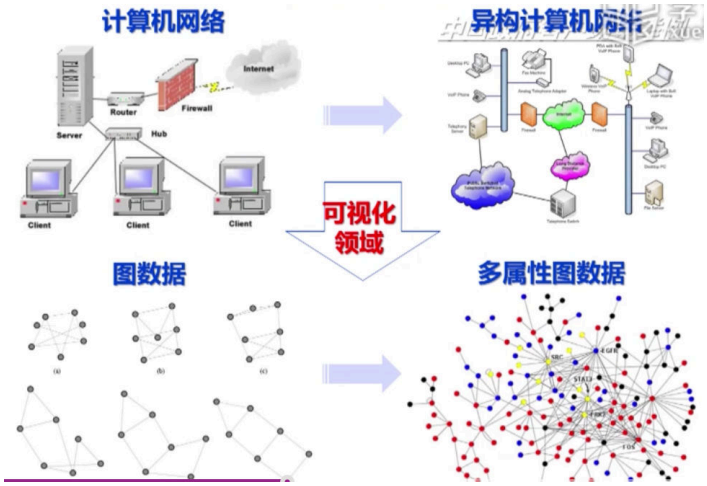
虚拟校园社会网络数据



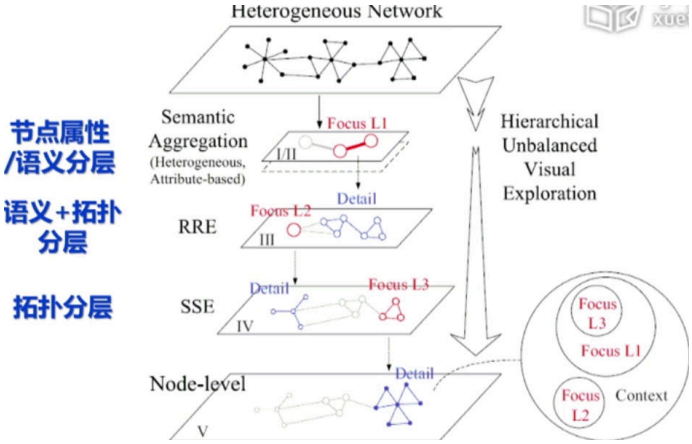
图拓扑结构 + **多维节点属性**

= 多属性网络

·异构网络数据

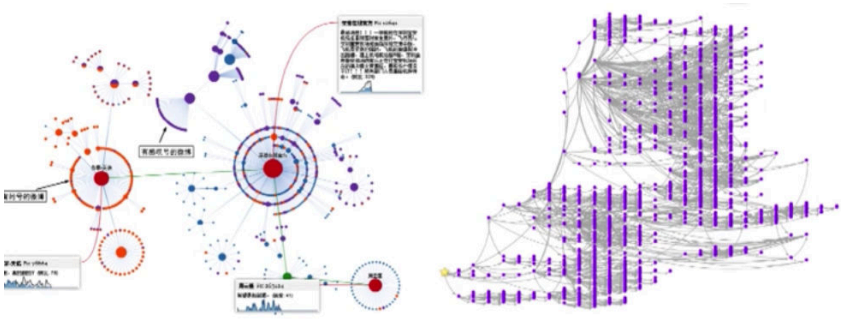


·洋葱图可视化系统架构



- **大规模影响力图谱可视化**

·影响力图谱【论文网络】



微博转发图谱　　论文引用反转网络

1. 特性: 规模大，分层次，时序特征，有向无环

2. 可视化需求: 突出展示影响力流动，而非社区团体特征

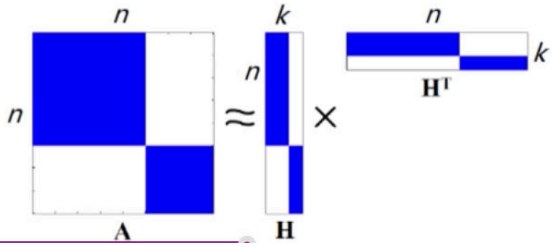
- 矩阵分解方法

对称非负矩阵分解：目标矩阵为图近似性矩阵

$M^G = (AA^T + A^TA)/2$

$$\min_{H \geq 0} ||M^G - HH^T||_F^2$$

where $||\cdot||_F$ denotes the Frobenius norm of the matrix. $H = \{h_{ij}\}$ is a n by k matrix indicating the cluster membership assignment of nodes in G: v_i will be clustered into π_c if h_{ic} is the largest entry in the ith row of H.



• 基于可视化的评价

[Han SIGMOD 2000]经典论文

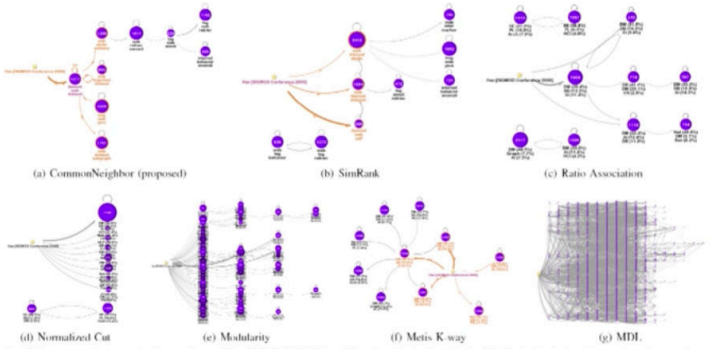
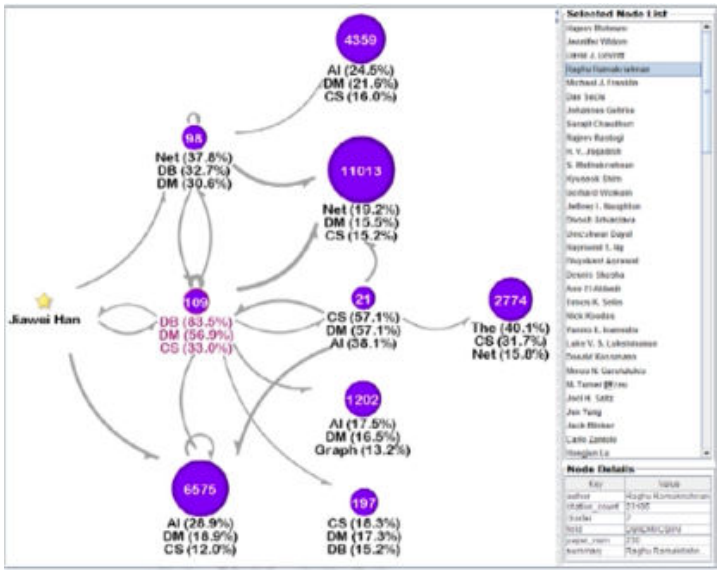


Fig. 8. Influence graph summarization results on [Han SIGMOD'2000] by different methods ($k = 10$, $l = 20$). Node label gives the number of papers in each cluster and their content summary by either title+abstract keywords in (a),(b) or the top 3 research fields in (c)~(f). Link thickness indicates the normalized flow rate. Some part of the graph is highlighted to show the number of citations as edge labels. Note that the modularity algorithm stops at 62 clusters and can not merge any further. MDL produces 4,937 clusters, leaving a half of the visual complexity from the input graph.

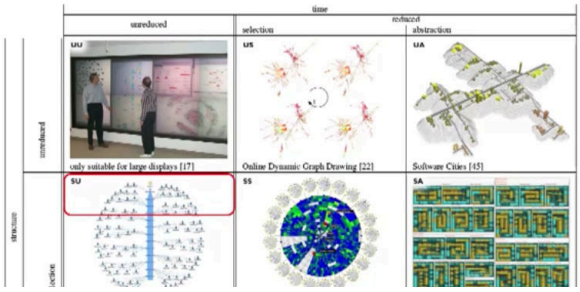
• 作者影响力图谱



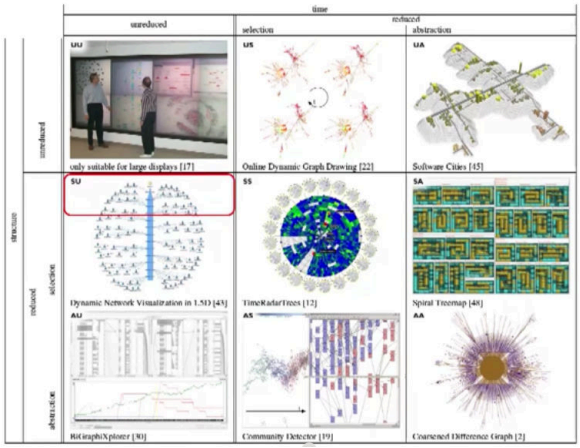
• **1.5维动态网络可视化**

·动态网络可视化分类

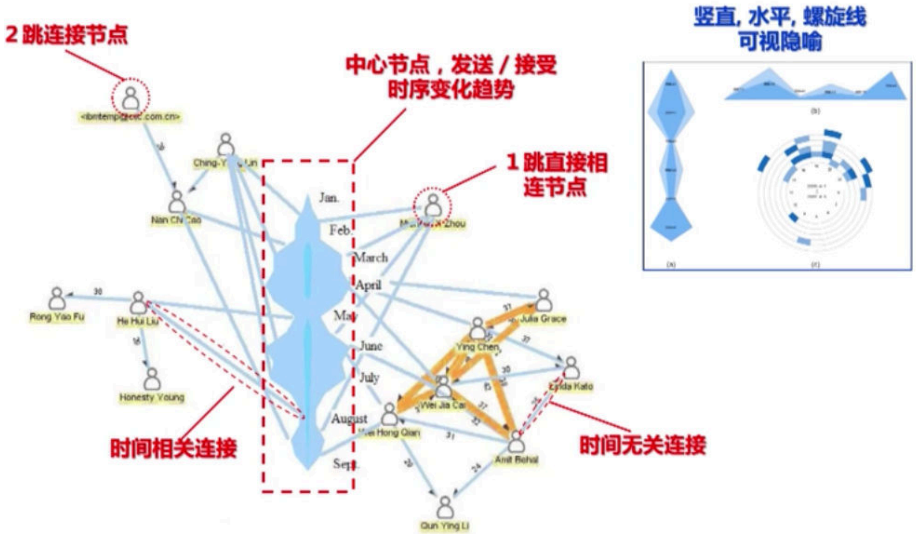
所有方法均采用数据变换方法，基于时序或图拓扑结构



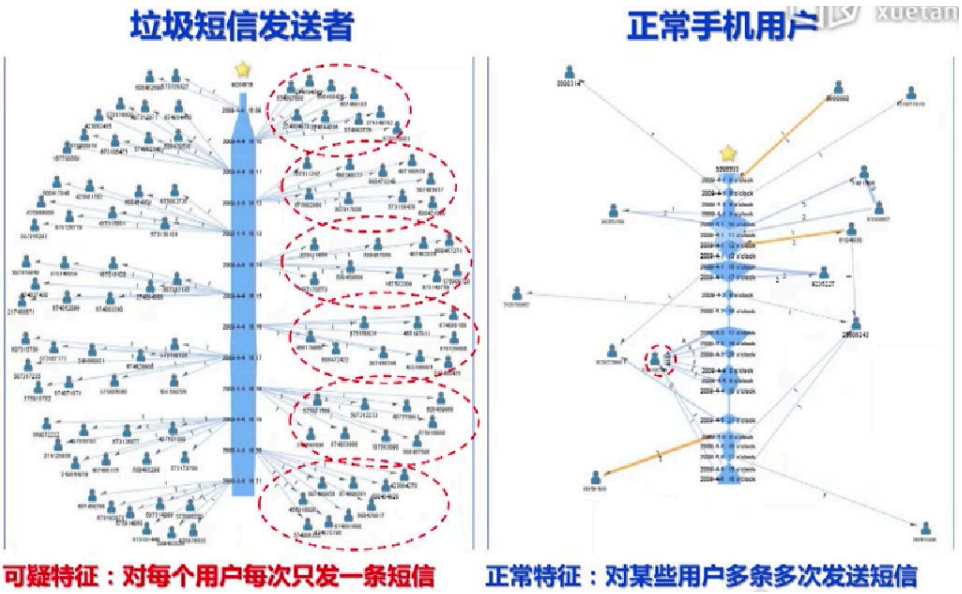
所有方法均采用数据变换方法，基于时序或图拓扑结构



·1.5维网络可视化隐喻【以某人为中心，不去展现全体】

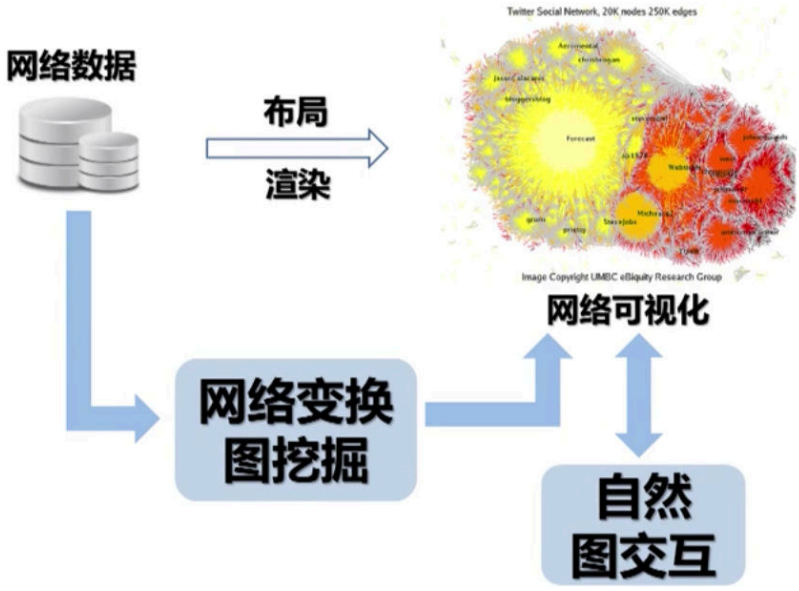


·实例分析——通过手机短信网络发现垃圾短信发送者



·大数据网络可视化小结：

革命性方法（Paradigm shifts）将集中于新型分析方法（网络变换/挖掘）和自然图交互



十七、苏中：从大数据到认知计算

1.大数据概述

IBM——商业机器

***硬件公司——>软件公司——>服务公司——>云计算公司。

GTO——对未来预测困难，每年都需要更新预测。

2012年——IBM提出 Big Data，以及4个”V“。【Volume, Velocity, Variety, Veracity】

2013年——讨论大数据怎么构成？来源，架构，走向……以前信息交互极少——>现在数据量++，及时性++！！！

·**social数据 + device数据**（如IOT远程控制系统）。

2014年——IT系统分化：System of engagement ——> System of Insight

• **今天互联网公司都在做金融**（掌握了大量的用户交易、信用数据）。

2015年——Data。【Data will disrupt entire industies】

2.大数据相关新趋势

All about **非结构化数据**。

·手机计算能力【总和】将超过服务器。

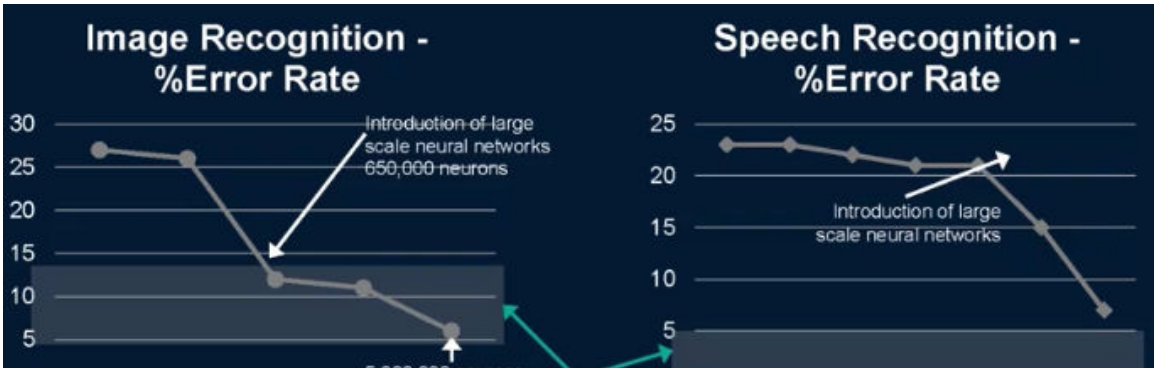
·移动终端存储能力【总和】将超过服务器。

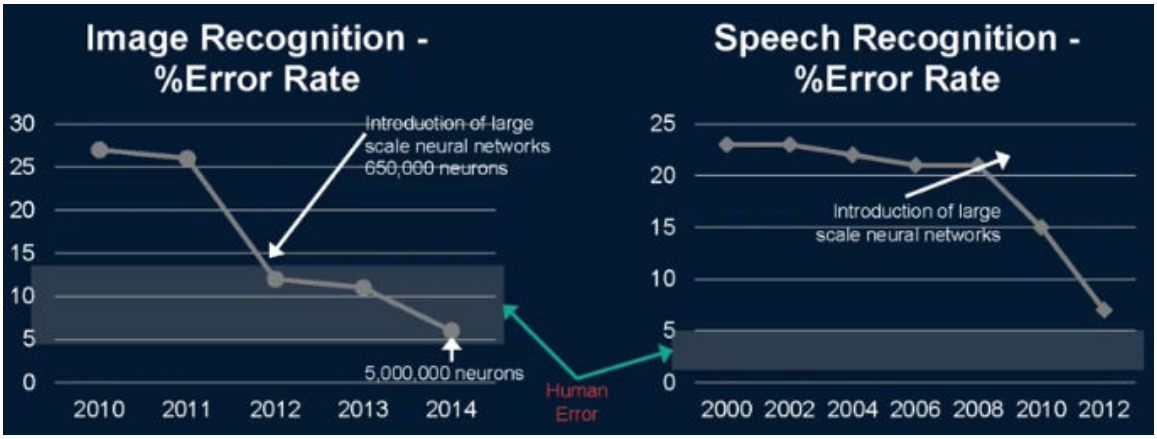
Block Chain——无中心的交易系统，去中介化。【历史数据不可以被随意篡改，Block Chain的生命力将来甚至超过比特币】

3.大数据技术创新

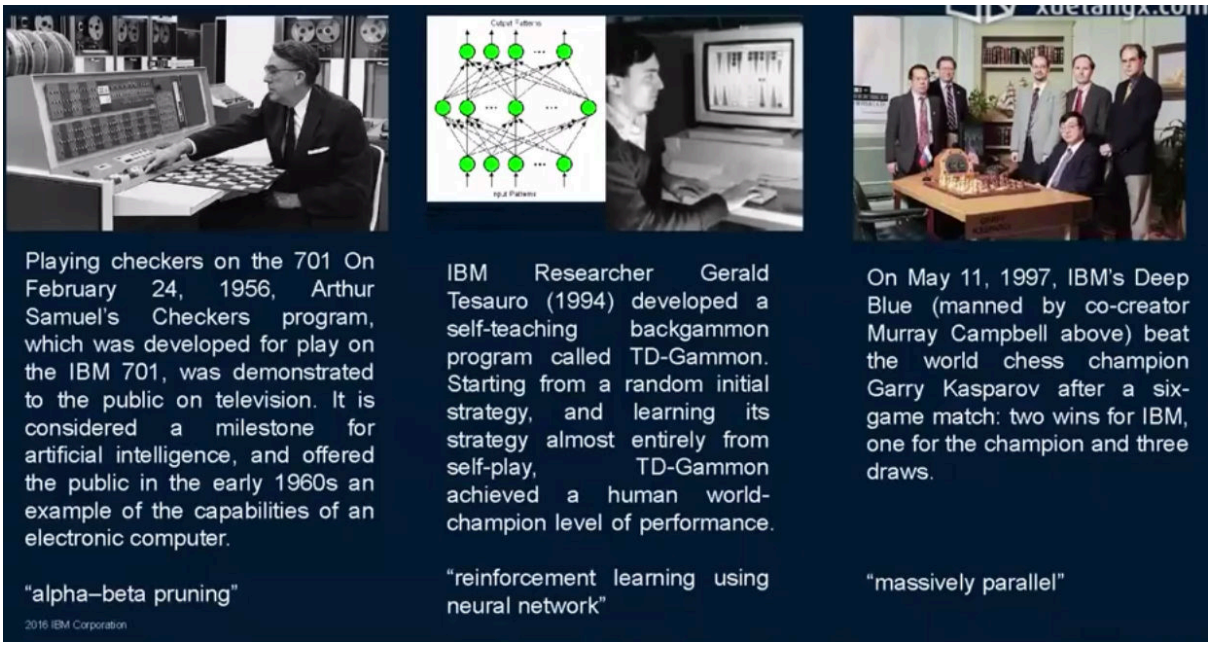
***计算机能力高速提升

【Image Recognition + Speech Recognition】





人工智能（历史可以追溯2000+years）——IBM的一些发展：



***深蓝的突破：并行计算、评估函数【棋子的数量、位置……】。（小数据、专家系统）

feature图——>图像处理。【图的特点；局部相关性】

————转化预测问题。

足够的大数据训练——计算机取得了一些人类没有掌握的经验。

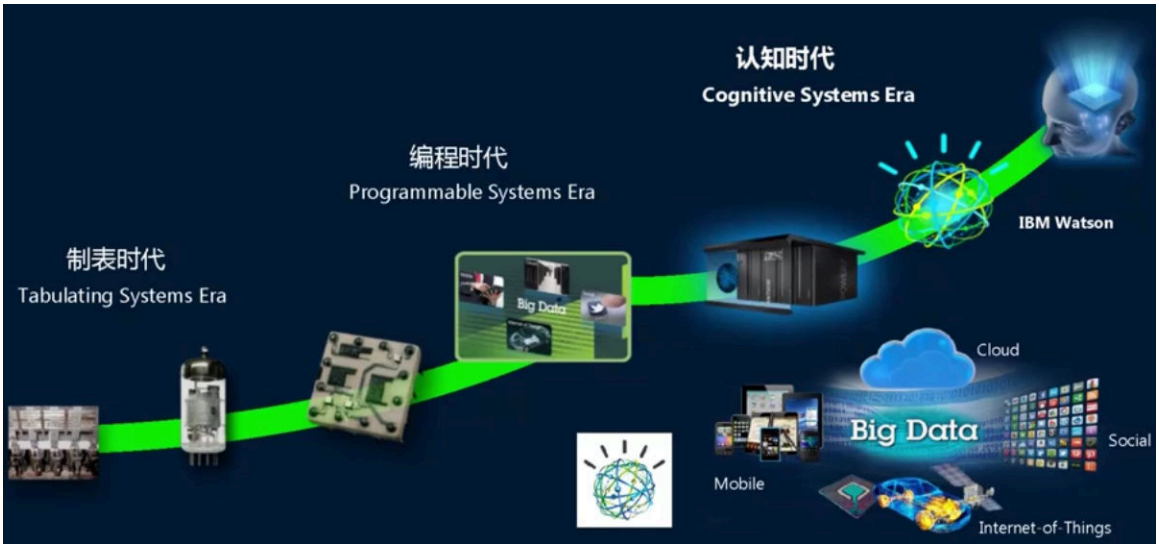
自然语言的研究：多大的脑量级才能产生这样复杂的语言系统？

·生理特征学说（自然语言是由生理决定的，因此存在一个**通用语法**【现实中很难做】）

分类器……

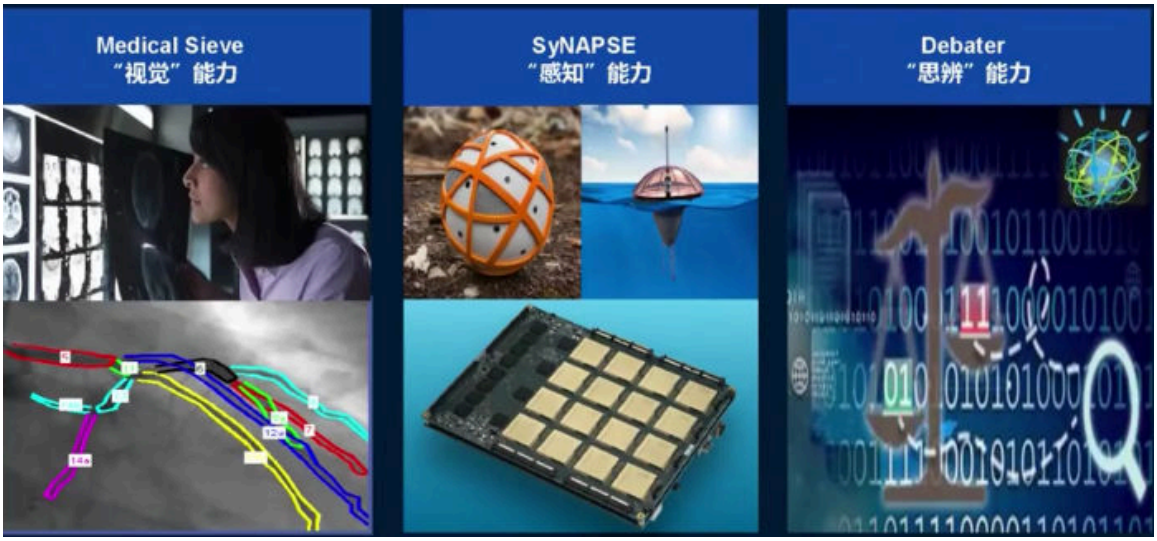
……

……

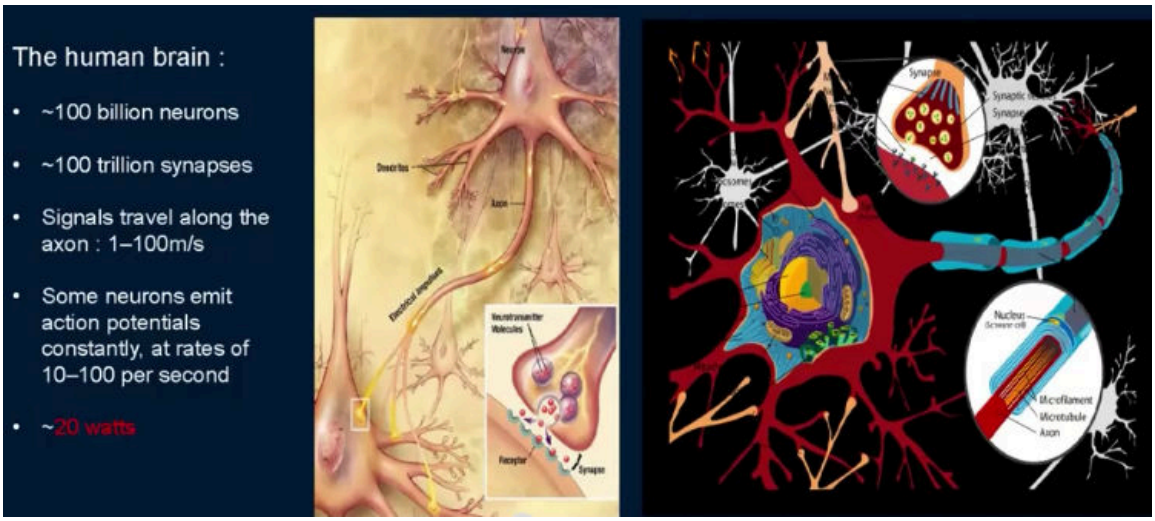


&&放射科医生的x光片识别——机器不知疲倦地工作。

……

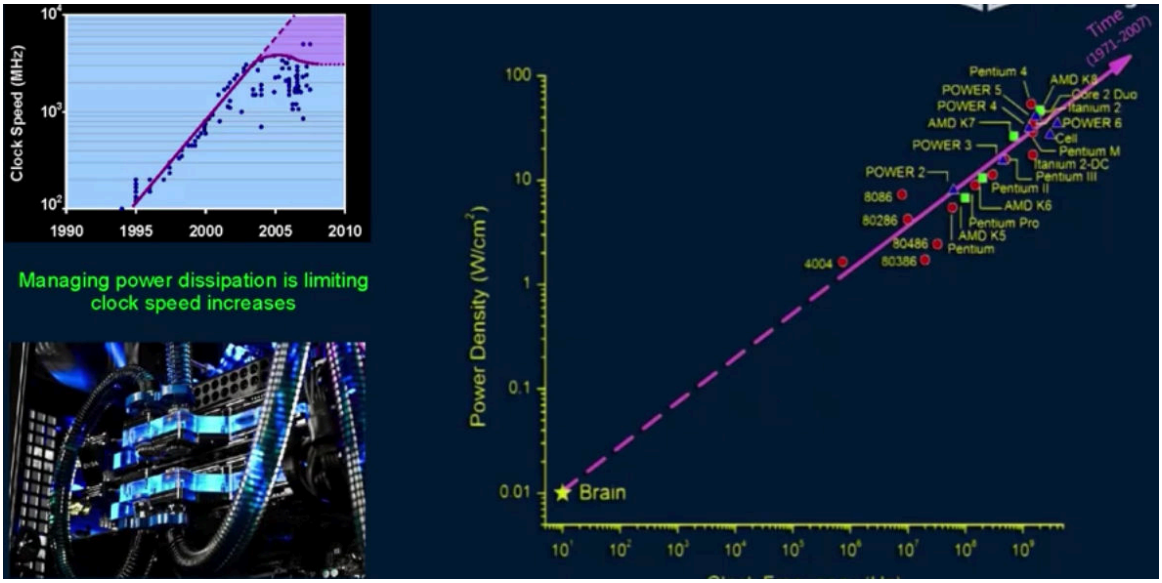


人脑的工作方式：



大脑的**充分互联**——大脑的高速反应。【存在一些长线关联】 功率：20瓦（脑）/60瓦（人）

·计算机的发展历史：



大脑没有时钟周期，具有鲁棒性（容错能力）。

4.大数据商业价值和前景

• 医疗

很多时候（6成+）后天的因素【生活习惯】造成疾病。

1100TB的Behavior Data（饮食、运动……）。

通过大数据的挖掘。

·智能汽车

通过各种传感，做实时的健康出行变化。

·法律

辅助律师解决打官司的问题。

·雾霾

……

5.大数据机遇和挑战

数据变多变复杂，含金量变低…

Data Privacy。【个人隐私】

Data Security。【数据财产化】

人工智能60周年（2016年）。（互联网 + 大数据 提供了人工智能很多的 应用场景 + 计算方法）

·技术螺旋上升……

十八、彭元：网络安全与大数据

1.网络安全概述

***每一个现代人都离不开互联网。

·数据泄露

现代互联网安全的第一威胁。

——账户数据【网站的弱加密，用户的密码同一性】

——商业数据

·钓鱼诈骗

钓鱼网站+++，集中在金融银行业务。【1/5真实比】

·DDoS

攻击者控制大量肉鸡攻击服务器，导致正常用户无法使用。【成本低】

·APT攻击（高级可持续攻击 Advanced Persistent Treat）

Google极光事件：

——社会工程学 ——> 恶意程序植入 ——> SSL通道监控 ——> 获得凭证 ——> 监控Gmail。

****传统检测

·基于特征检测流量

·基于算法检测频率

·基于Token做身份识别

·基于黑名单阻断

• 基于扫描器打补丁

……

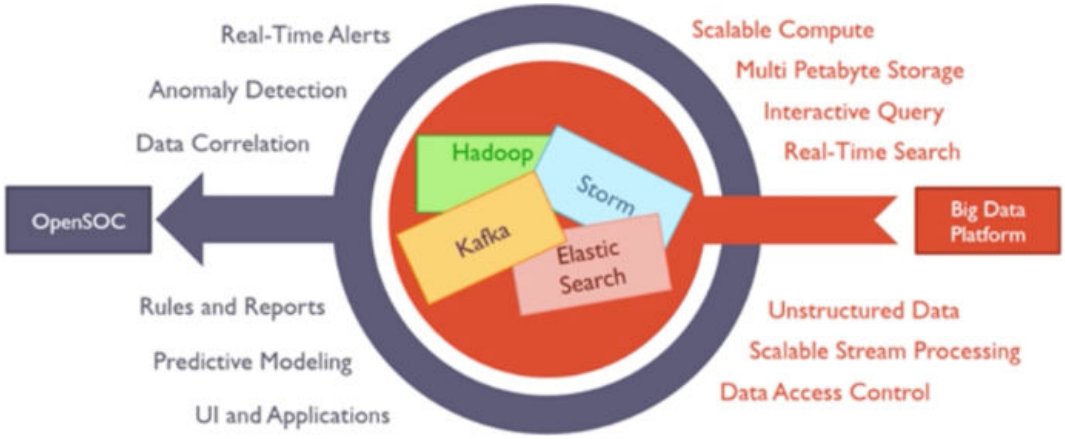
——存在问题：被动响应，APT/0Day难防，溯源困难，误报率高，孤立无援……

2.大数据安全分析平台

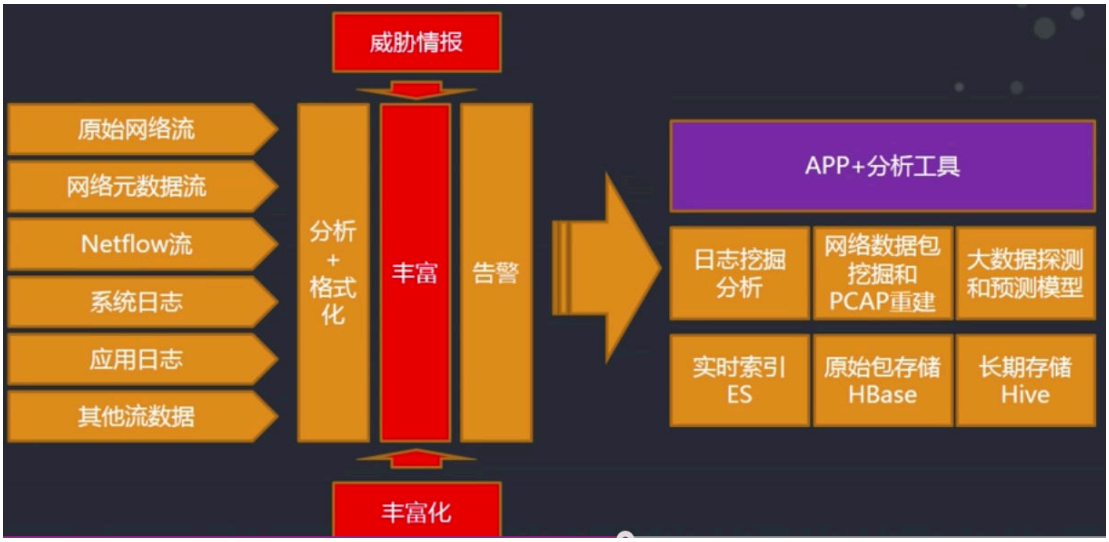
大数据特点：大，种类多，复杂，无格式。

网络流量	日志	攻击样本	环境数据

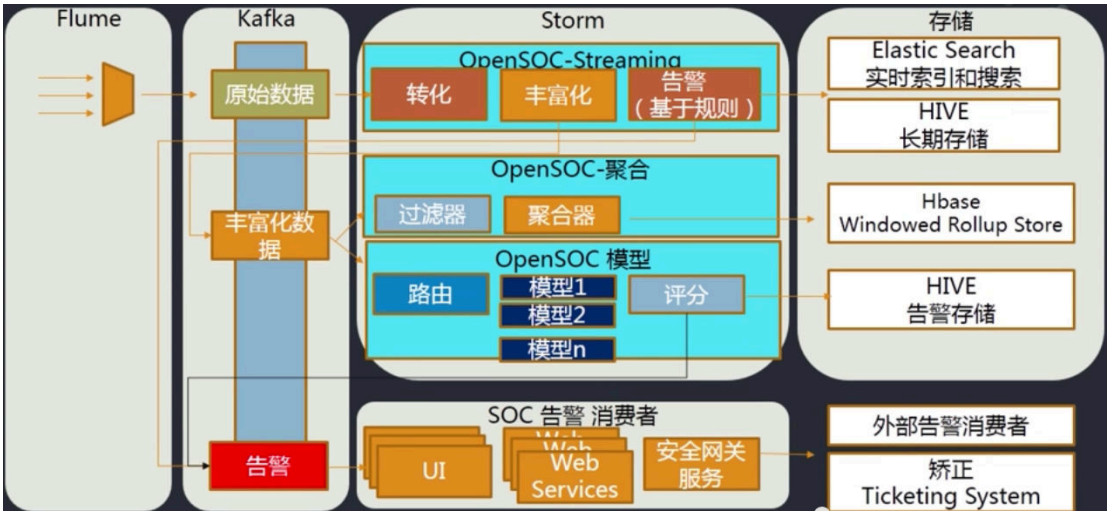
开源分析平台OpenSOC：



运作方式：



分析流程（结合技术组件）：



算法：

·聚类分析

——SteamKM++

——D-Stream

·分类算法

——Hoeffding 树

——Half-Space 树

·离散检测

——平均绝对偏差

——平均值标准差

——移动平均值标准差

3.大数据安全应用

• 大数据可以解决的安全问题

态势感知	威胁情报	攻击溯源	行为检测

• 态势感知

——Facebook knows when you'll get divorced (even before you do)

延伸：《疑犯追踪》

·威胁情报

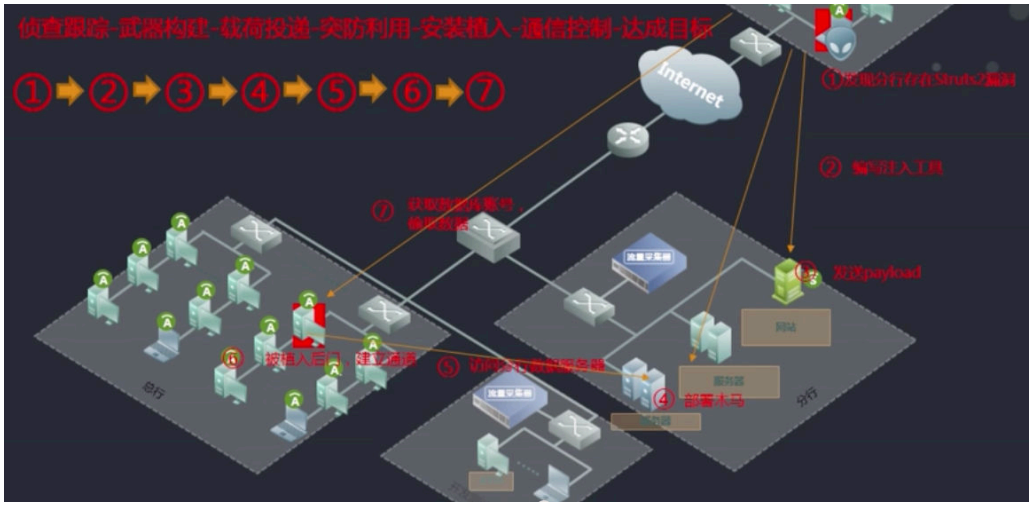
——基础数据（DNS解析服务，域名信息，企业IP）

——样本（恶意样本，活跃病毒，重大事件）

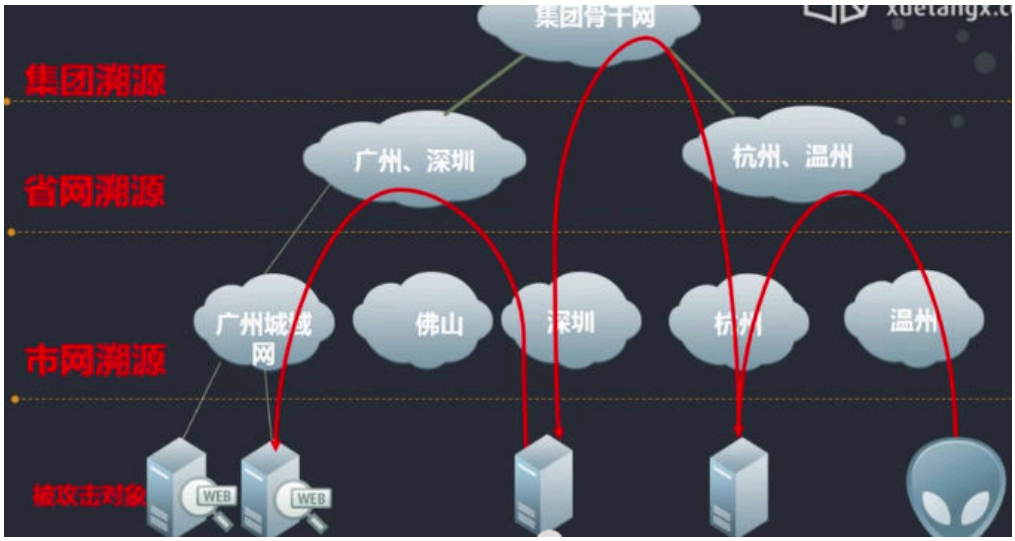
——信誉（IP信誉，URL信誉，文件信誉）

·攻击溯源

（内网）



（外网）



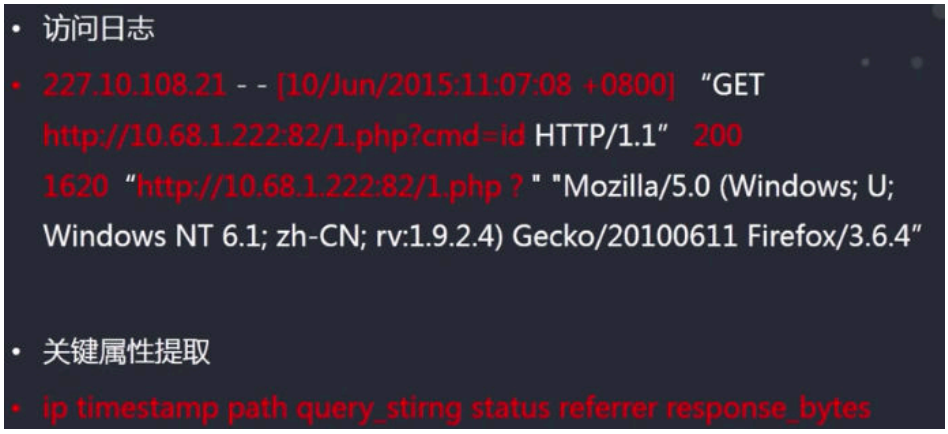
·行为识别

延伸：《碟中谍5：神秘国》【步态追踪】

示例：Webshell

——Webshell行为特征

——分析Webshell需要的数据



——分析步骤

4.大数据平台安全

• 风险分析

Who	How	What
谁	方式	目的

• Who

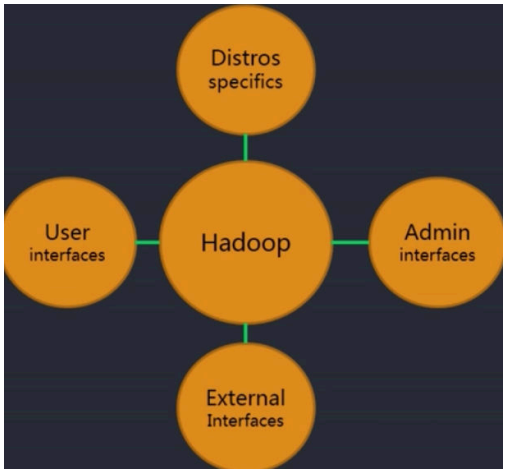
——商业角度：竞争对手、脚本小子、APT

——位置角度：外在（匿名、前员工），内在（员工、用户、管理员）

·What

——数据（【以金融业为例】交易数据、销售数据、客户数据）

·How



——U用户接口（用户和App）

Apache Hue

– Pig , Hive , Impala , Hbase , Zookeeper , Mahout , Oozie

Other

– Tez , Solr , Slider , Spark , Phoenix , Accummulo , Storm

——入侵ApacheHue

- 找出指定的Hue安装版本(Hue 2.6.1, Django 1.2.3)
- 找出访问Hue的目标用户
- 发送XSS payload
- 获取目标用户权限

——A管理员接口

Apache Ambari

Apache Ranger

Other

- Knox , Cloudbreak,Zookeeper , Falcon , Atlas , Sqoop , Kafka

——D发行版问题

·更新慢

厂商多久发布一个新版本？

- 大版本一年、小版本3个月、补丁1-2个月
- 那时候多少个其他组件过期？

企业多久部署一个新版本？

- 半年到一年
- 那时候又有多少个其他组件过期？

·安全问题

安全漏洞

- 基础库漏洞（java，php，ruby）
- Hadoop组件（Hue，Ambari，Ranger）

默认密码

- SSH、Mysql、空密码

默认配置

- 无网络层隔离

安全漏洞

- 基础库漏洞（java，php，ruby）
- Hadoop组件（Hue，Ambari，Ranger）

默认密码

- SSH、Mysql、空密码

默认配置

- 无网络层隔离
- 无HTTP层隔离（点击劫持，会话管理）
- 调试开关（Hue）

——**E**外部接口

超过25个Apache apps/modules

厂商/发行版特定的apps/interfaces

监控组件：Ganglia/Splunk

权限提供组件：LDAP，Kerberos，OAuth

Other APP

&&**如何防护**？

- 充分的网络访问限制
  - Keep it super tight！
- 充分的用户权限管理
  - Map business roles to permissions
- 外部连接
  - Checklist
- 典型的Web漏洞
  - 渗透测试
- 组件的漏洞
  - Checklist
  - CVE
- 发行版漏洞
  - 集成后测试
  - 提供商