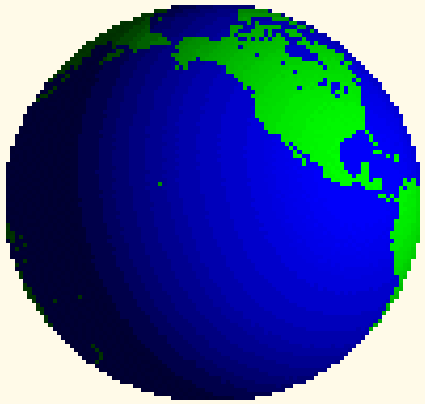
大数据是信息化发展的新阶段

邬贺铨

中国工程院

2018.06.09

大数据是信息化发展的新阶段



信息化的新阶段

大数据的新挑战

信息化为中华民族带来了千载难逢的机遇。

---习近平，在全国网信工作会上讲话，2018.04.21

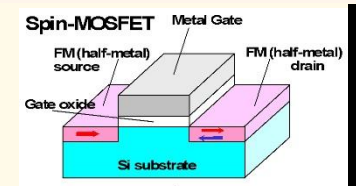
大数据是信息化发展的新阶段。

---习近平，在中央政治局第2次集体学习会上讲话，2017.12.08

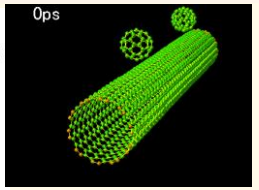
IT从芯开始

华为最近发布的移动手机用麒麟970芯片，内置8核CPU，采用10nm工艺，集成了55亿个晶体管/cm²

Source:
http://newsroom.intel.com/docs/DOC-2035,
北大王阳元院士

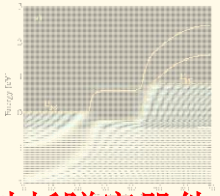


自旋逻辑器件
Spin-Transistor

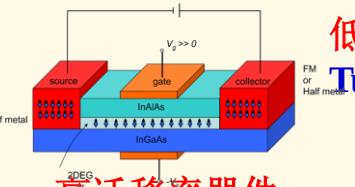


碳基器件
Nanotube/Graphene-FET

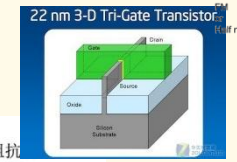
在2017年CPU上晶体管数为160亿个，晶体管的尺寸比一个流感病毒还要小



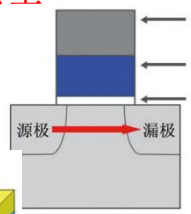
低功耗隧穿器件
Tunneling-FET



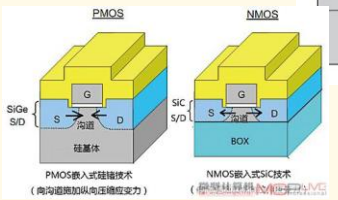
高迁移率器件
Ge/Ⅲ-Ⅴ-FET



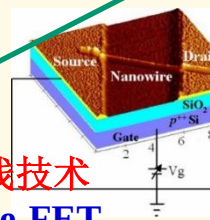
新结构器件
Fin-FET/Tri-Gate



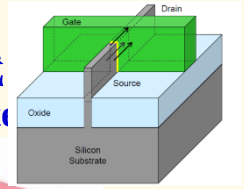
金属栅
高K介质



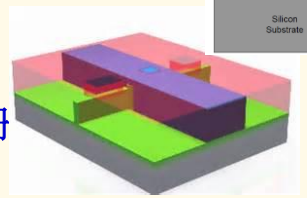
沟道工程
应变硅技术



Si纳米线技术
Nanowire-FET



第二代
Tri-Gate



第二代
高K介质/金属栅

第二代
应变硅技术

90nm 2003
65nm 2004
45nm 2007
32nm 2009
22nm 2011
16nm 2013
14nm 2015
11nm 2017
5nm 2020

集成电路技术面临芯片发热和代工线高成本等挑战，期待后摩尔时代的颠覆性技术！

摩尔定律：IC每18～24个月密度/速度加倍

30年来:
CPU速度提高100万倍，
内存价格下降45000倍，
硬盘价格下降360万倍

如果汽车的价格能与硬盘同等速率下降，今天一部新车仅需0.01美元

5G的商用终端芯片需要7nm或5nm的工艺，2020年有望实现5nm工艺商用化。

软件定义一切

1972年阿波罗登月飞行器软件仅有4K的代码

高铁的列控软件有数百万行代码

雪佛兰、奔驰新车软件规模1000万行到1亿行

空客飞机软件10亿行代码

PC 5000万行代码

智能手机OS上百万行代码

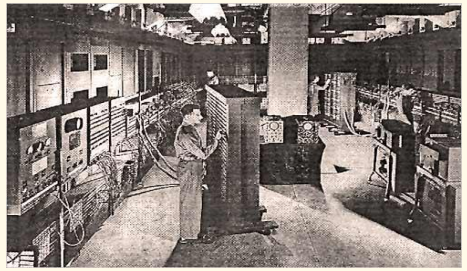
- 中国2017年软件和信息技术服务业收入5.5万亿元，同比增长12.2%，其中软件（含嵌入式系统软件）为2.57万亿元。2020年规划达到8.8万亿元。

工控软件

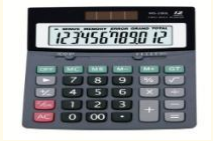
波音公司设计与制作飞机需使用多种软件，其中1/8可外购，其余自己开发，波音也是软件公司！

西门子有超过1.7万名软件人员开发工业平台软件，声称是欧洲第2大软件公司！

计算无所不在



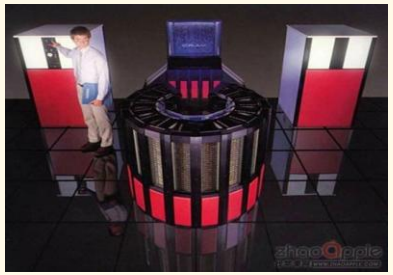
电子数字计算机发明于1946年，占地170m²



性能不及现在的计算器

美NASA 1975年500万美元的超算机Cray-1

性能不如iPhone4



美国1985年的超算机Cray-2

性能不及iPad2

1997年1GB闪存卡 $7,992

现在10美分

1956年IBM正在运输5MB的硬盘

计算机成本十年下降近一万倍！存储器近2万倍！

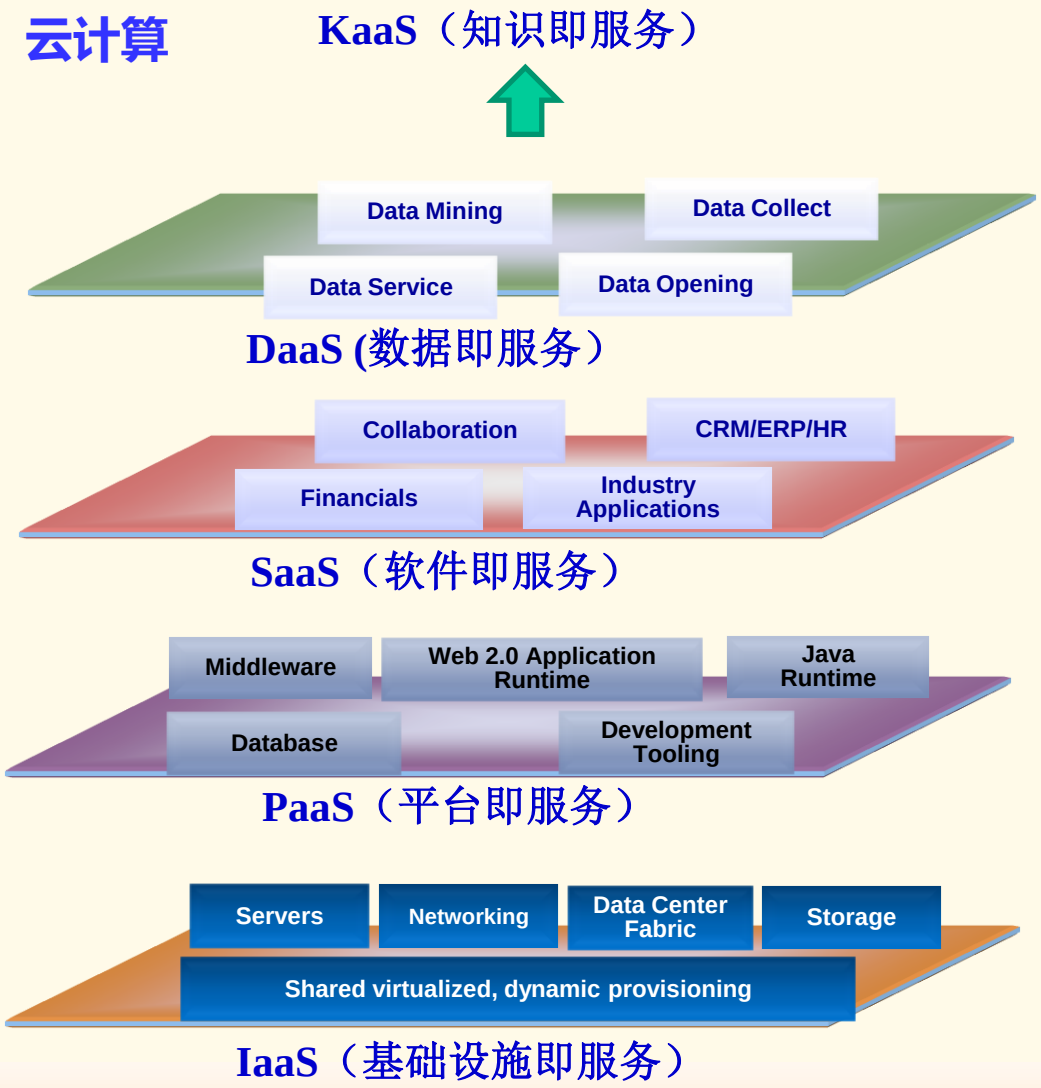
PC 计算能力提高20年千倍！超算机十年千倍！

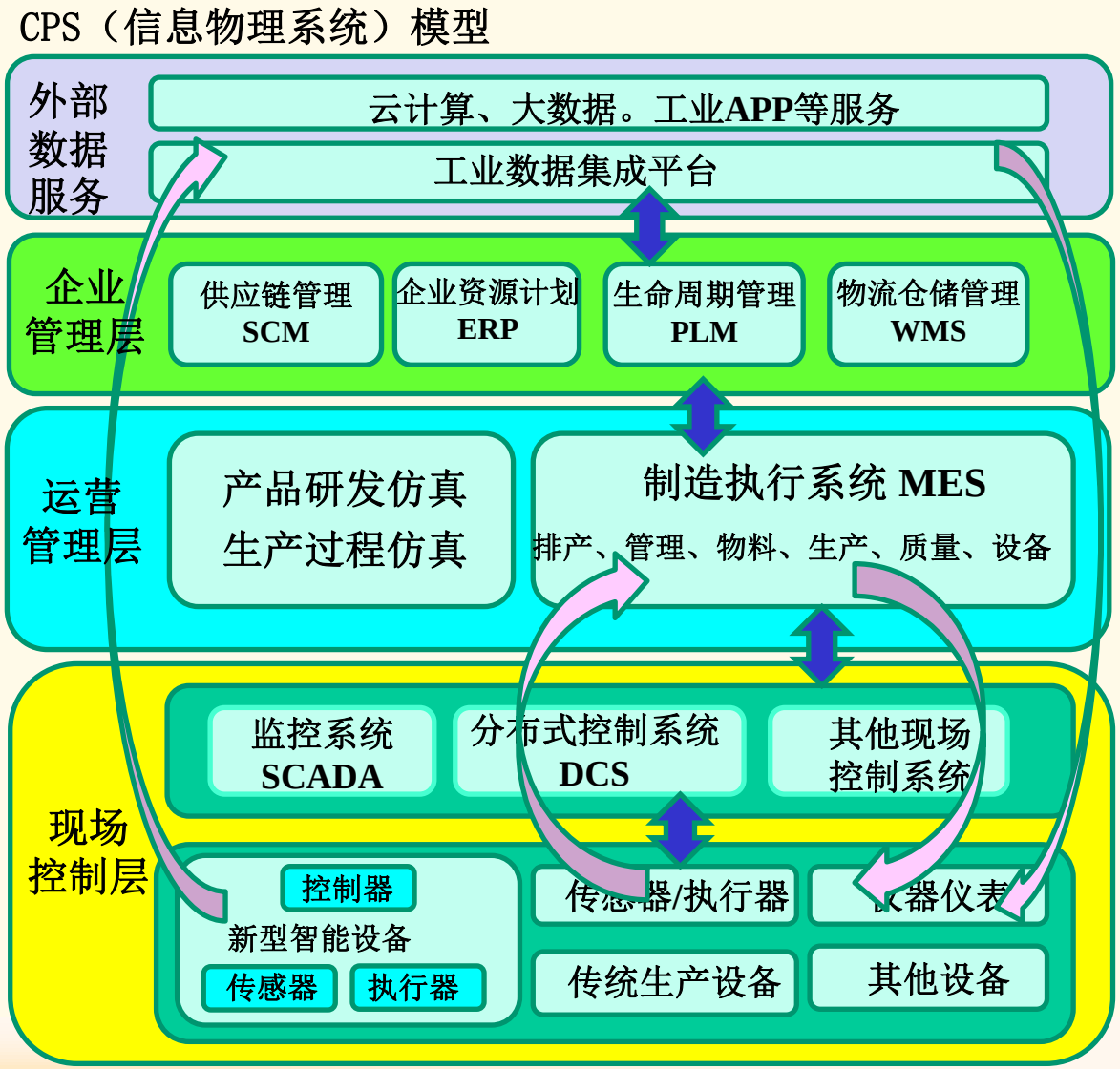
2018年6月美国将投入使用比神威性能提升60%的超算。2021年计算能力达到E级（10^{18}）即天河二号百倍的超算也将投入使用，在40MW功率约束下，能效超出现有系统百倍。

2018年5月中国宣布2020年百亿亿次超级计算机交付。

中国2016年研制出神威超级计算机，12.54亿亿次/秒

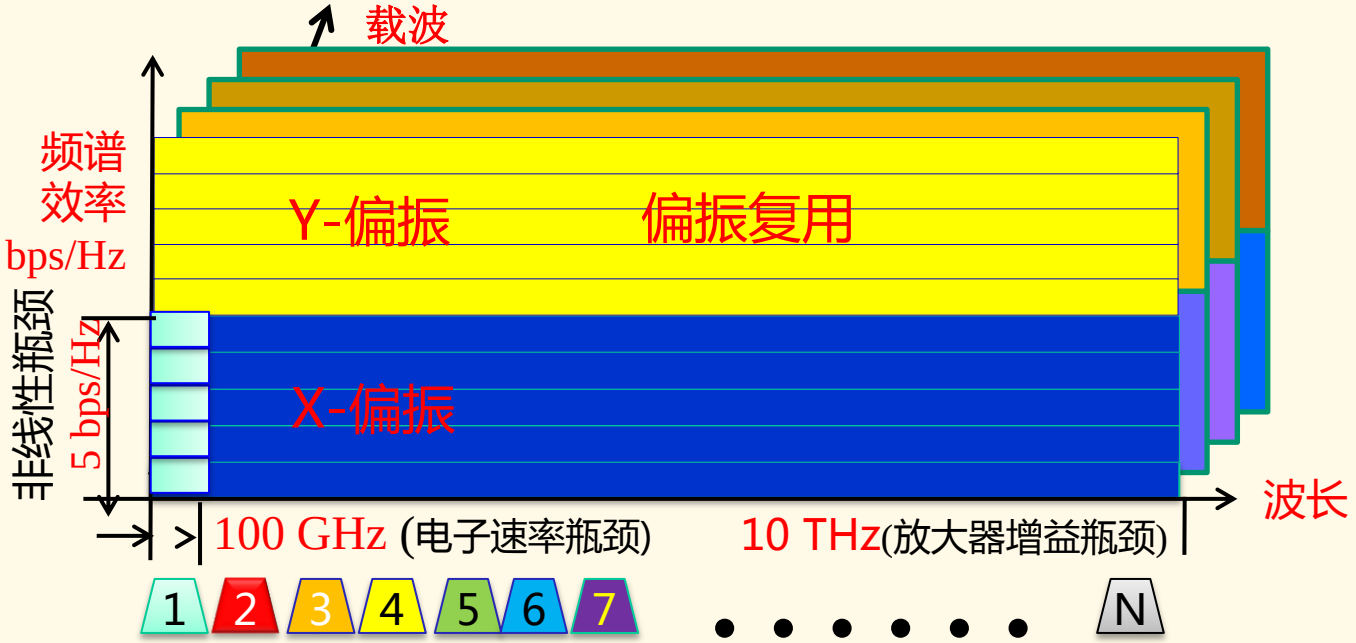
服务集约成云

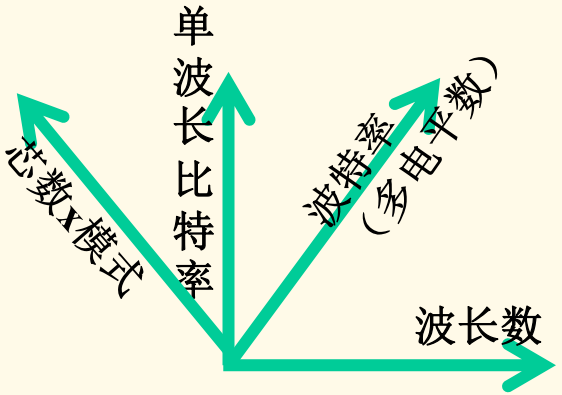

云计算
KaaS（知识即服务）
Data Mining
Data Collect
Data Service
Data Opening
DaaS (数据即服务）
Collaboration
CRM/ERP/HR
Financials
Industry Applications
SaaS（软件即服务）
Middleware
Web 2.0 Application Runtime
Java Runtime
Database
Development Tooling
PaaS（平台即服务）
Servers
Networking
Data Center Fabric
Storage
Shared virtualized, dynamic provisioning
IaaS（基础设施即服务）

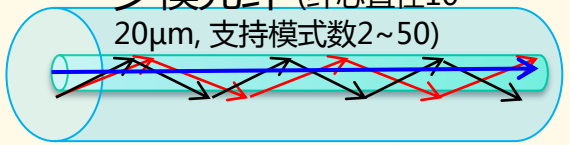

CPS（信息物理系统）模型
外部数据服务
云计算、大数据。工业APP等服务
工业数据集成平台
企业管理层
供应链管理 SCM
企业资源计划 ERP
生命周期管理 PLM
物流仓储管理 WMS
运营管理层
产品研发仿真
生产过程仿真
制造执行系统 MES
排产、管理、物料、生产、质量、设备
现场控制层
监控系统 SCADA
分布式控制系统 DCS
其他现场控制系统
控制器
新型智能设备
传感器
执行器
传感器/执行器
仪器仪表
传统生产设备
其他设备

光纤永无止境

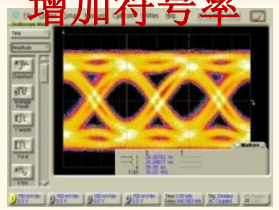
❑ 单纤容量20年万倍；目前最高记录：单波长400G，单纤100T



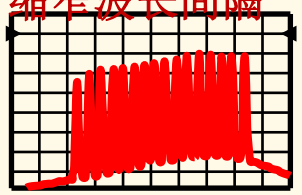




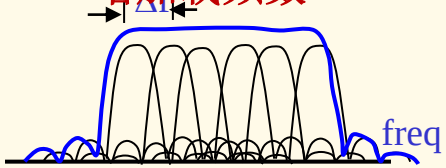
增加符号率



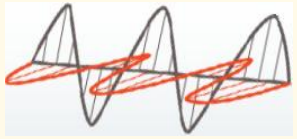
缩窄波长间隔



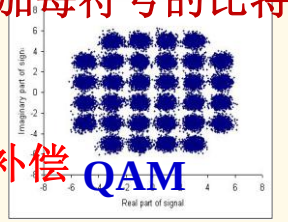
增加载频数



偏振复用



增加每符号的比特数



超强FEC

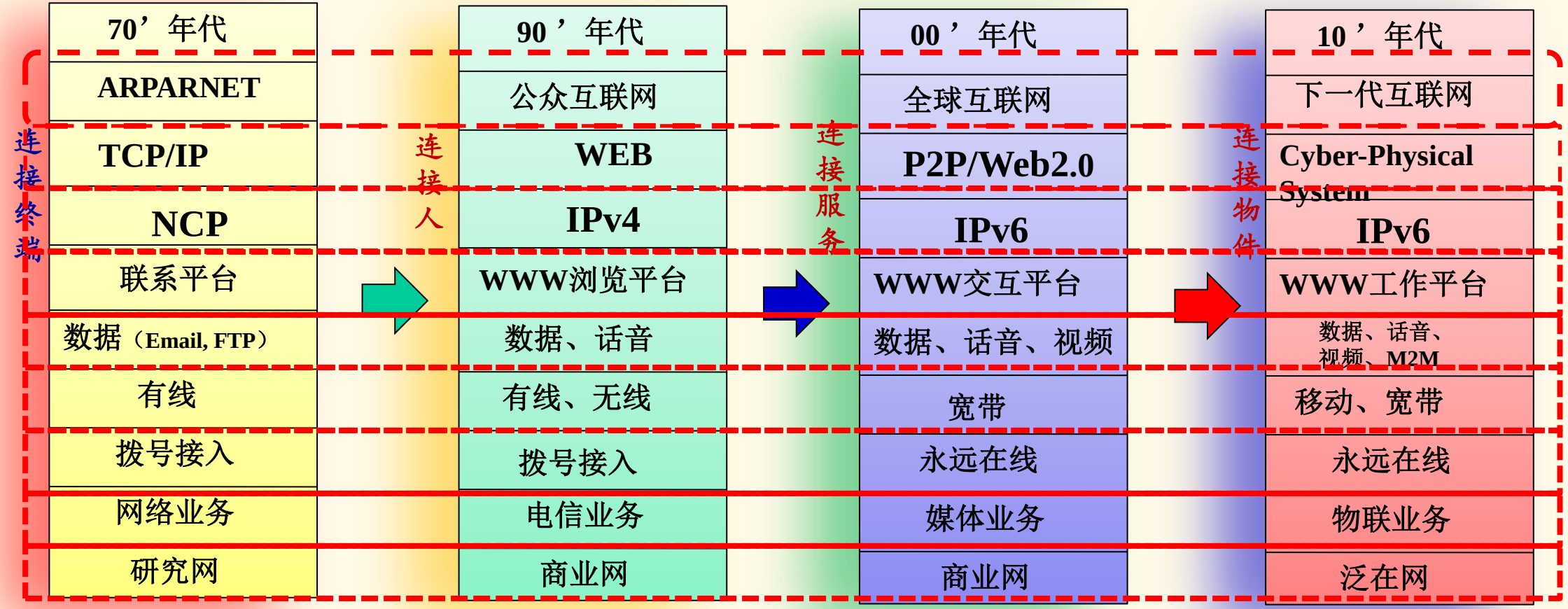
DSP色散补偿

时分复用 → 波分复用 → 正交频分复用 → 偏振复用 → 空分/模分复用

TDM ⊗ WDM ⊗ OFDM ⊗ PDM QPSK MDM/MCM ⊗ SDM/OAM

中国每年生产了全球一半的光纤光缆，中国市场也消耗了全球一半的光纤光缆，光纤价格十年下降50倍！

互联超越初衷

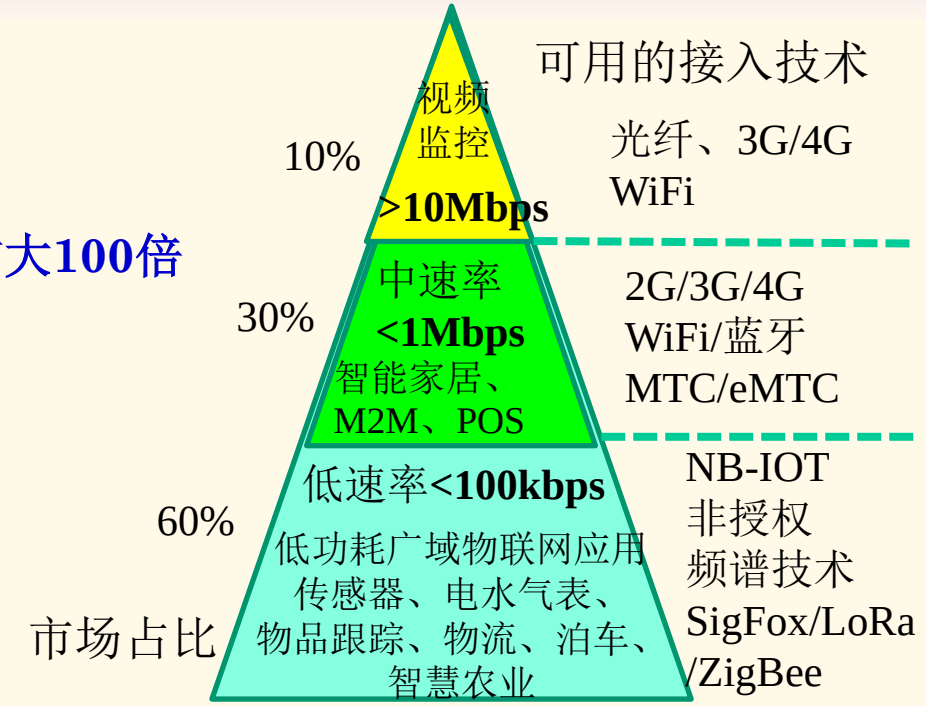


2017年12月底，中国互联网网民达到7.72亿，普及率为55.8%，手机网民占网民数97.5%。

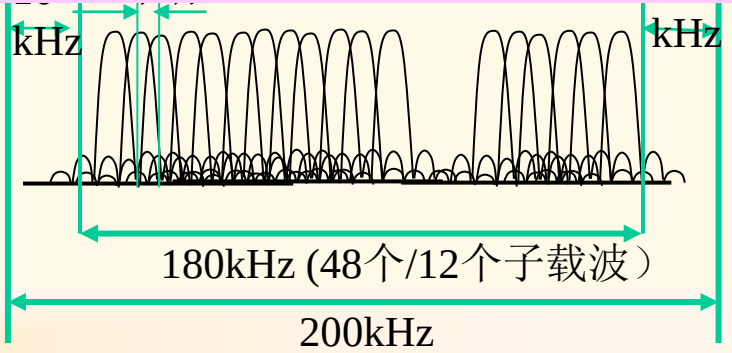
宽带化、移动化、泛在化、智能化、安全性、可用性、可信性

网络包容万物

- 2016年6月通过了NB-IOT（窄带物联网）国际标准，其特点：
  - 广覆盖。在同样的频段下比现有的网络增益**20dB**，覆盖面积扩大**100**倍
  - 大连接。一个扇区能支持**10**万个连接，比现网高**50~100**倍；
  - 低功耗。终端模块功耗为**2G**的**1/10**，待机时间可长达**10**年；
  - 低成本。单个接连模块目标**1**美元。
- NB-IOT工作在电信频段，射频带宽相当GSM的一个载波，可部署于现网。支持20kbps(3.75kHz)和250kbps(15kHz)。
- NB-IoT提供了广域低功耗物联网能力，避免了物联网的碎片化，可应用于智慧城市、智慧工厂、物流管理、环境监测等领域。
- ITU估计2025年物联网全球市场将达到6400亿美元。

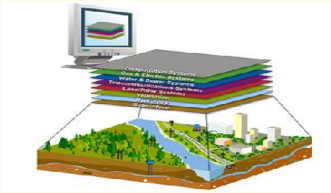


2016年美国发布了《2016-2045年新兴科技趋势报告》，未来30年排在首位的正是物联网。



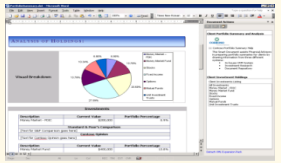
监测上天入地

环保物联网
水体污染监测

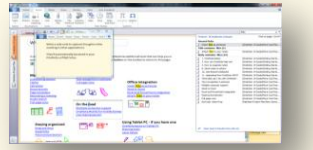


虚拟现实

态势分析



DSS

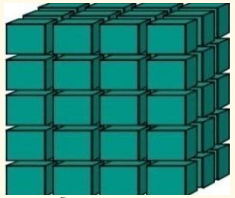


专家系统

3S

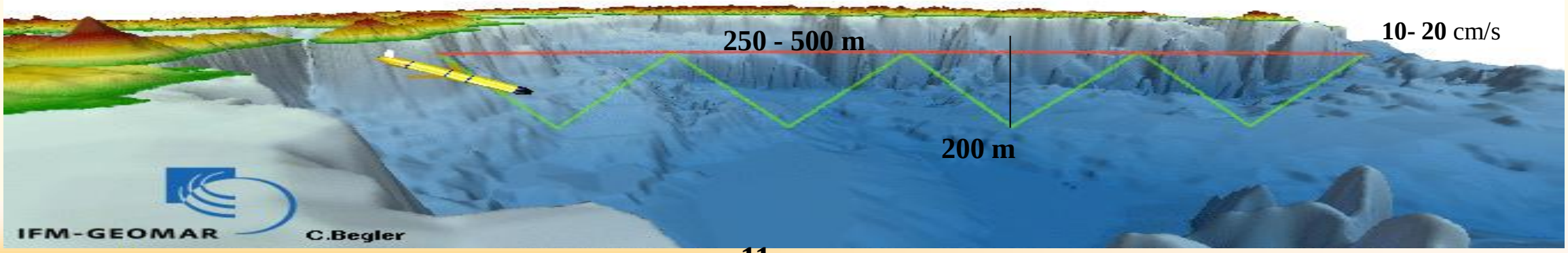
SaaS
TaaS
DaaS
IaaS
云计算

数据中心



数据挖掘

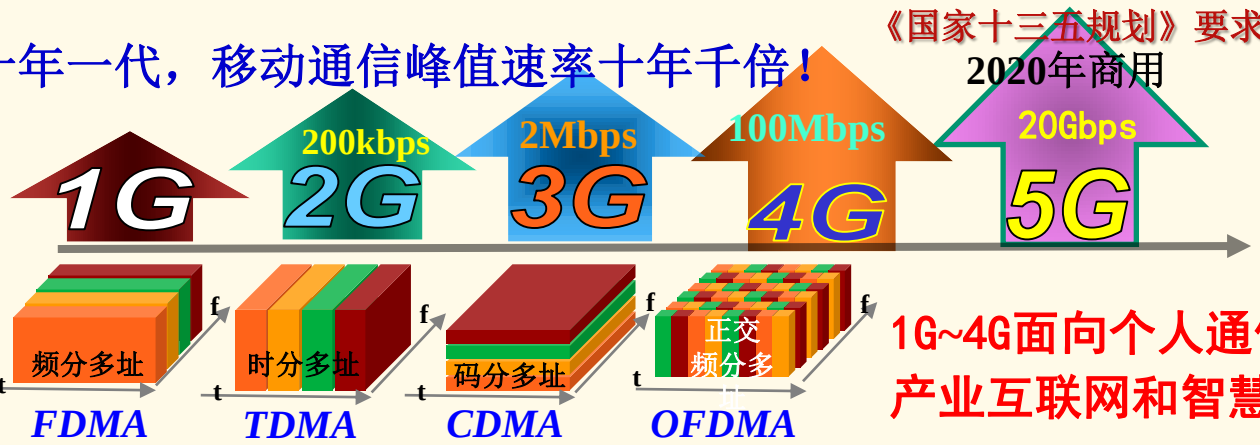
全球环保市场规模----2015年1.05万亿美元，2020年预计1.9万亿美元。
中国环保市场2016年为607亿美元，年增13.8%。



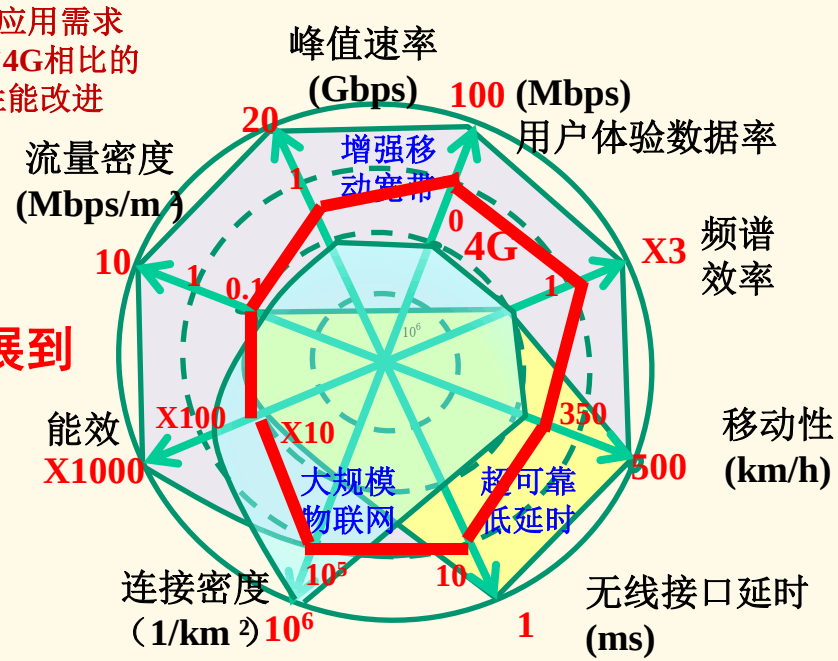
宽带移动互联

十年一代，移动通信峰值速率十年千倍！

《国家十三五规划》要求
2020年商用

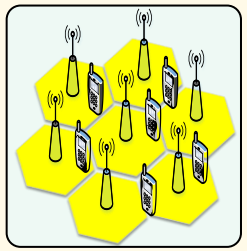


1G~4G面向个人通信，5G扩展到
产业互联网和智慧城市应用

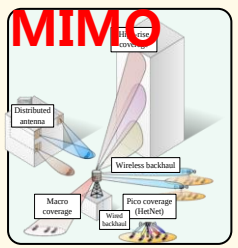


网络容量 = 基站数 K x 天线数 n x 信道带宽 W x log(SNR)

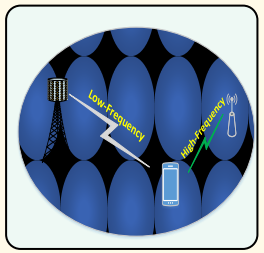
超密集组网



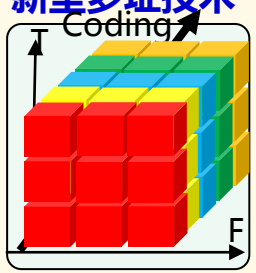
大规模天线阵



全频谱接入



新型多址技术



- 增强移动宽带----热点高容量应用（8kTV、VR/AR）
- 超可靠低时延----例如高铁、车联网、智能工厂、智慧医疗
- 广覆盖大连接----可穿戴设备、智慧城市

移动互联网

连续广域覆盖

热点高容量

4K TV、3D TV、VR / AR、可穿戴设备

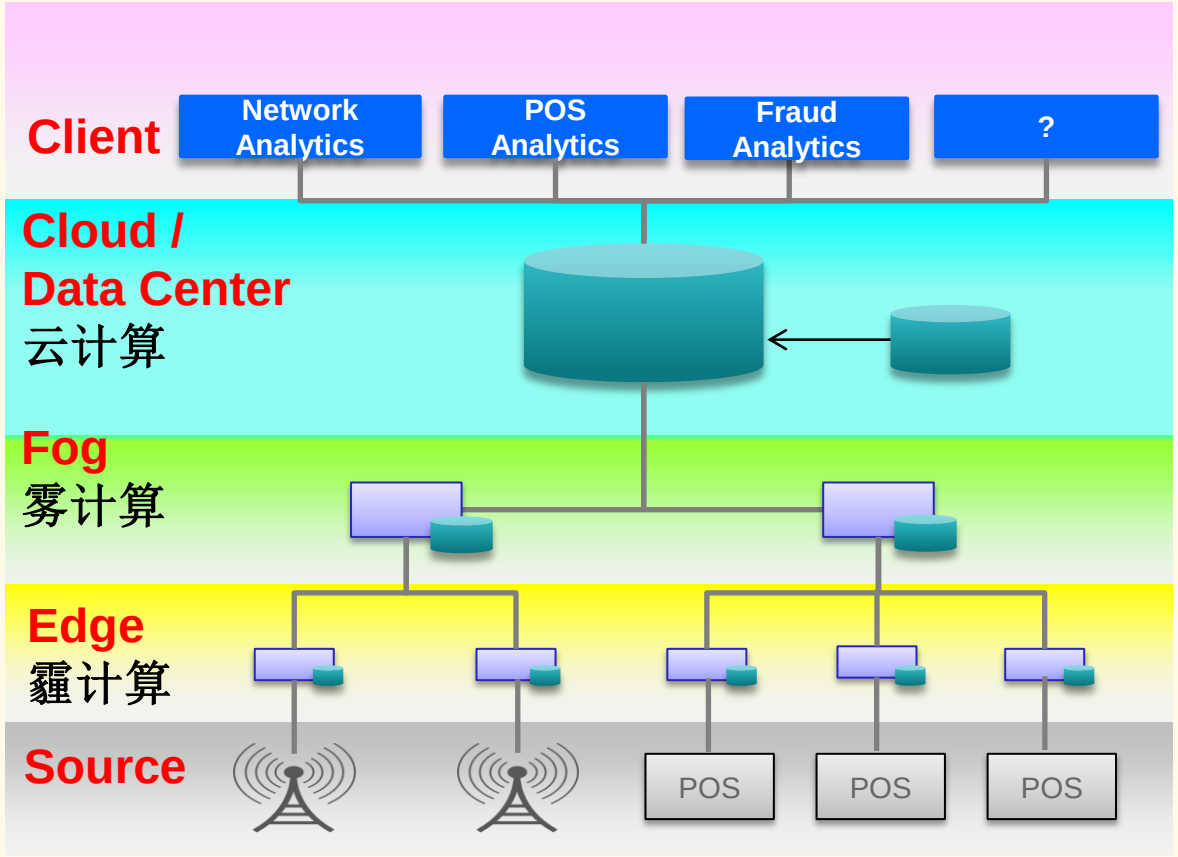
低延时高可靠

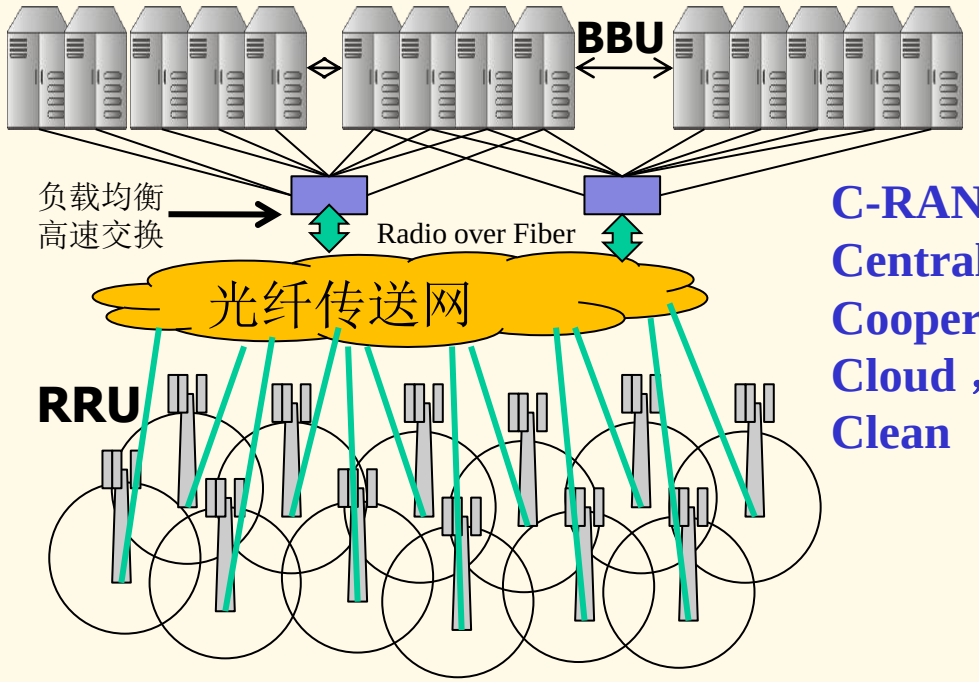
车联网、机器人、智能工厂、环境监测、智慧城市

低功耗与大连接

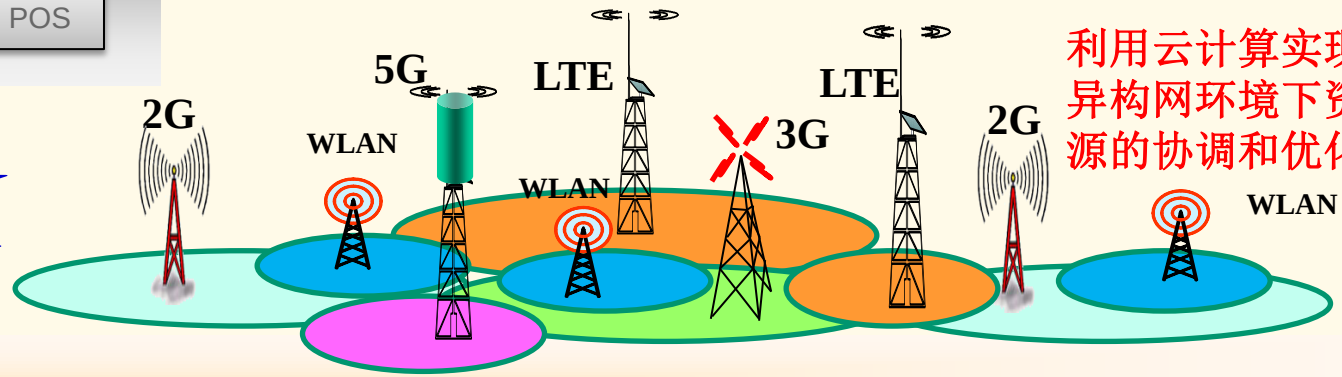
产业互联网

移动边缘计算





C-RAN----
Centralized，
Cooperative，
Cloud，
Clean

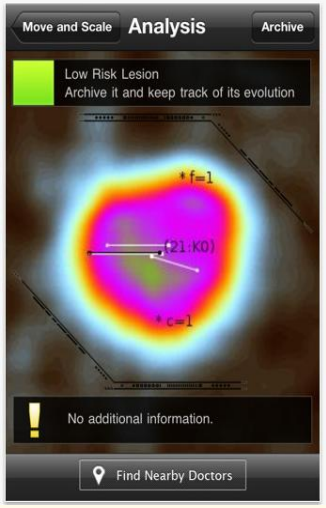


利用云计算实现异构网环境下资源的协调和优化

为适应视频业务、VR/AR与车联网等对时延要求，节约网络带宽，需将存储和内容分发下沉到接入网，移动边缘计算实现基站与互联网业务深度融合。

终端嵌入智能

移动智能终端成为物联网节点：
嵌入各种传感器----重力、压力、震动、加速度、距离、方向、亮度、定位、温度、湿度、手势识别、指纹识别、语音识别、语音翻译



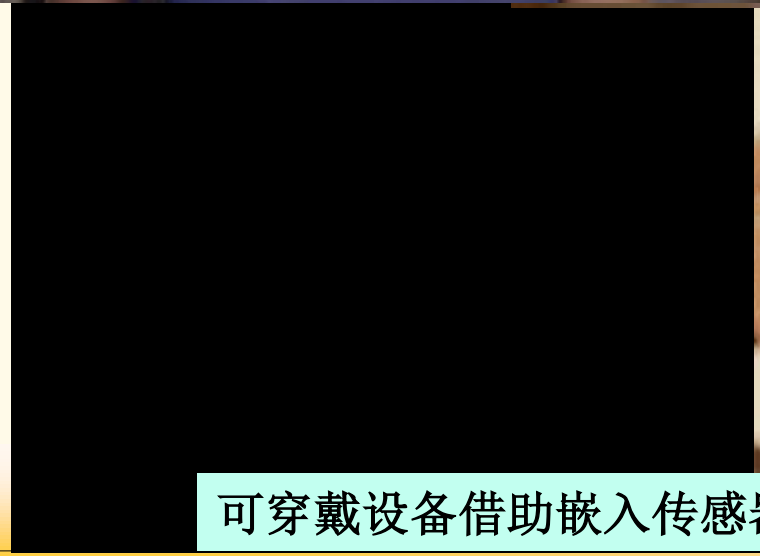
中国智能手机的产量占全球的70%以上！

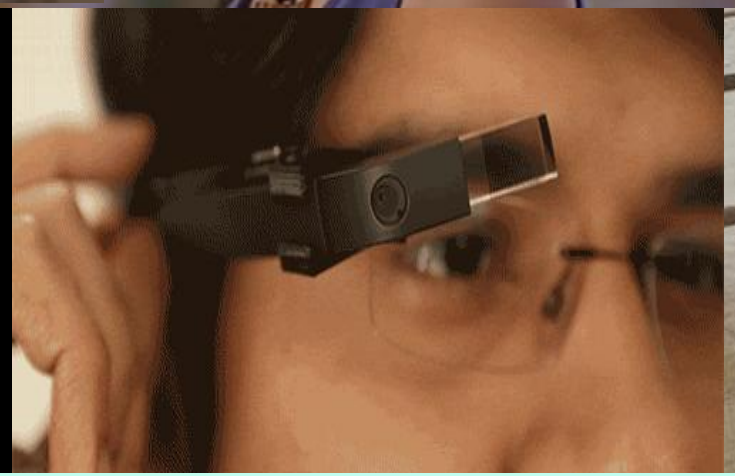
终端的智能通过联网上云而增强！

30年后假设有iPhone32手机，价格300美元，与现在相比其CPU能力和存储器容量都是100万倍，通信速度是300万倍，可存5000亿首歌曲，3万部电影。

---软银，孙正义，2014.11.19

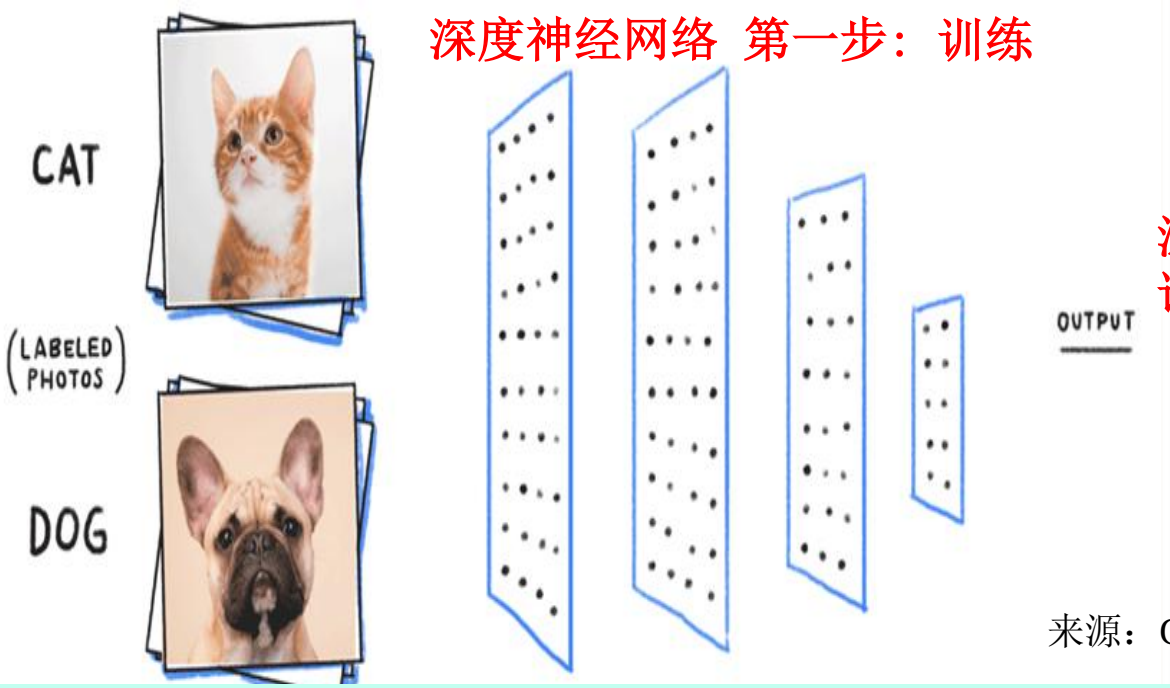
穿戴腾云驾物



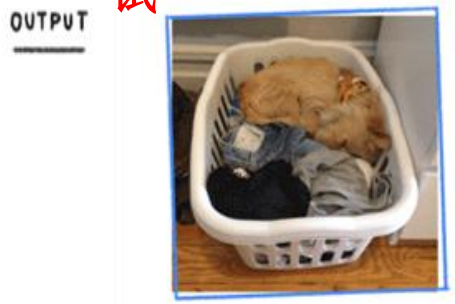


可穿戴设备借助嵌入传感器和经宽带移动通信连接到云端，强化了可穿戴设备的功能

智能在于学习

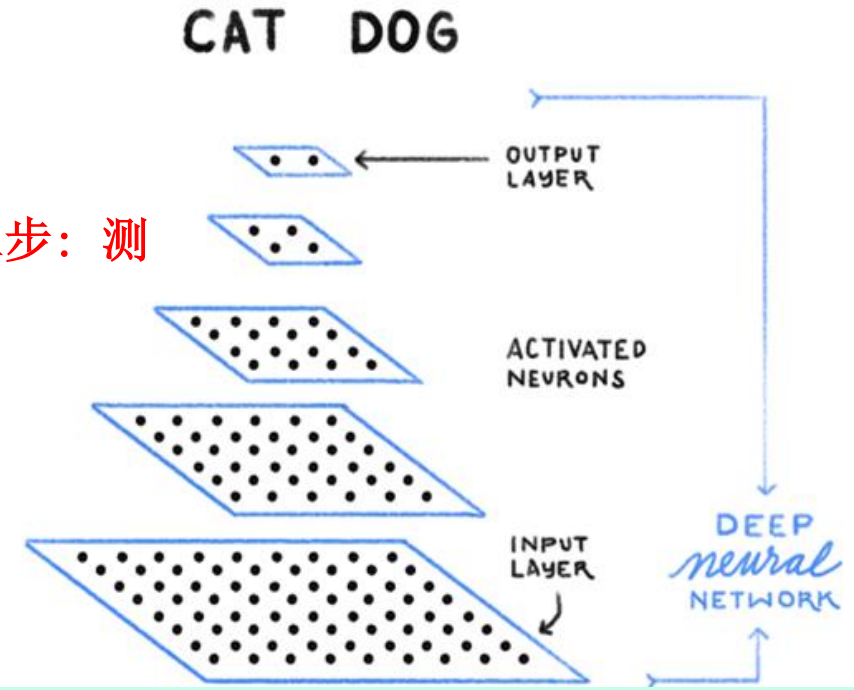


深度神经网络 第一步：训练



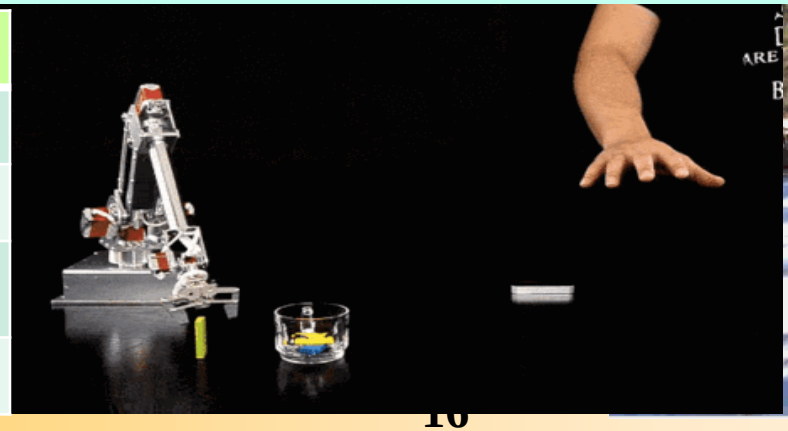
深度神经网络 第二步：测试

来源：Google



Gartner公司预计未来十年几乎每一应用都将含一定的AI。

2017年10月19日报道	
AlphaGo Zero	AlphaGo
1台设备+4TPU	多台设备+48TPU
3天自学规则	3个月培训
100:0	



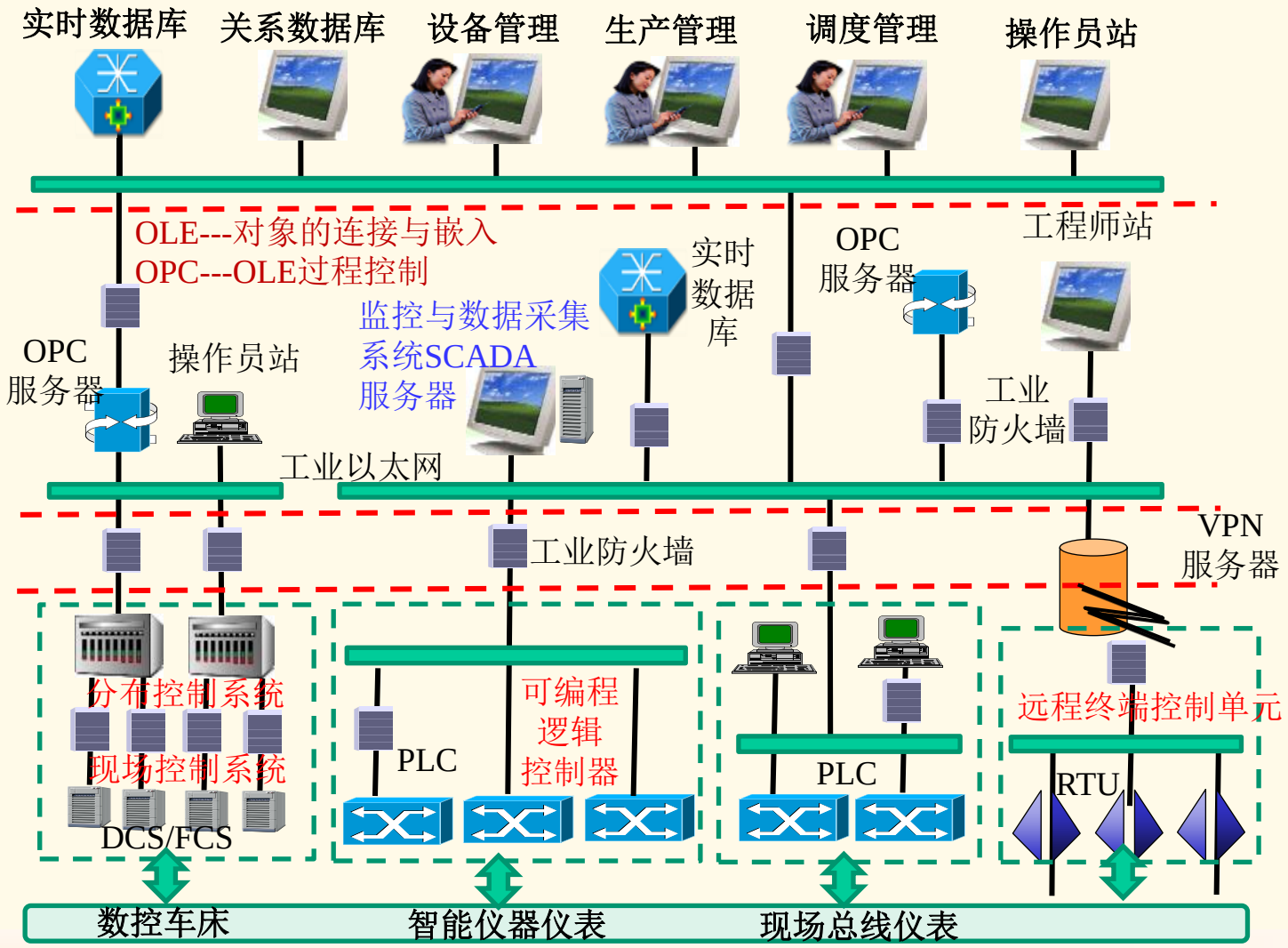
信息融合抽取

工业互联网的数据采集与监控系统

- 利用SCADA软件采集本地设备数据，转存到相应的数据库中，但成本高效率低。

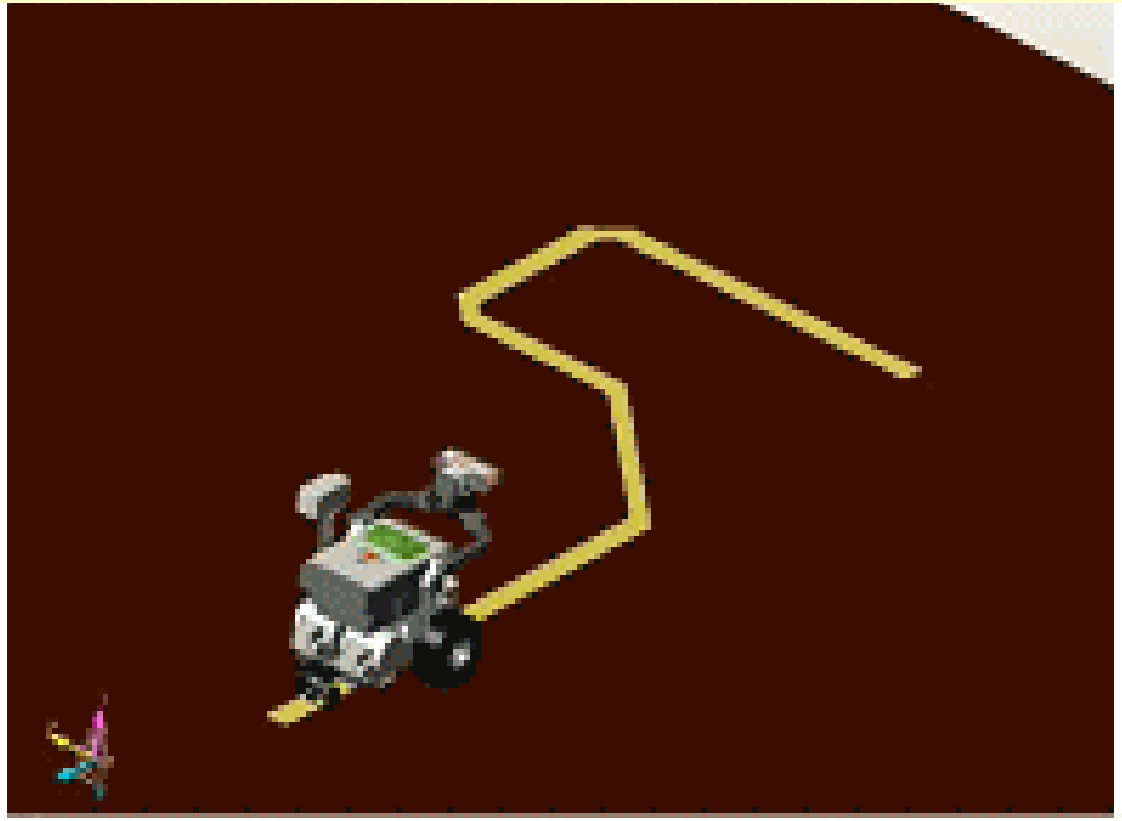
<table>
<tr><td colspan="4">上层应用软件</td><td colspan="3">数据挖掘软件</td><td colspan="2">状态分析软件</td><td>Web服务软件</td></tr>
<tr><td colspan="2">图形界面</td><td colspan="2">图形工具</td><td colspan="2">报警</td><td>历史数据</td><td colspan="2">其他应用</td></tr>
<tr><td colspan="7">实时数据库</td><td rowspan="3">数据库
服务器
GIS服务器</td></tr>
<tr><td colspan="2">协议接口</td><td>OPC</td><td colspan="2">OLE</td><td colspan="2">其他接口</td></tr>
<tr><td colspan="7">操作系统</td></tr>
<tr><td colspan="2">硬件驱动</td><td colspan="2">图形接口</td><td colspan="2">文件系统</td><td colspan="2">网络系统</td></tr>
</table>

- 从SCADA到工控网关：
  - 具备对设备的各种通信协议解析能力(Modbus、PPI、MPI、CNC等），未来还可解析总线和工业无线协议；
  - 具备对IT系统的协议对接能力（3G/4G、NB-IOT、工业以太网等）
  - 具备数据缓存和边缘计算的能力。



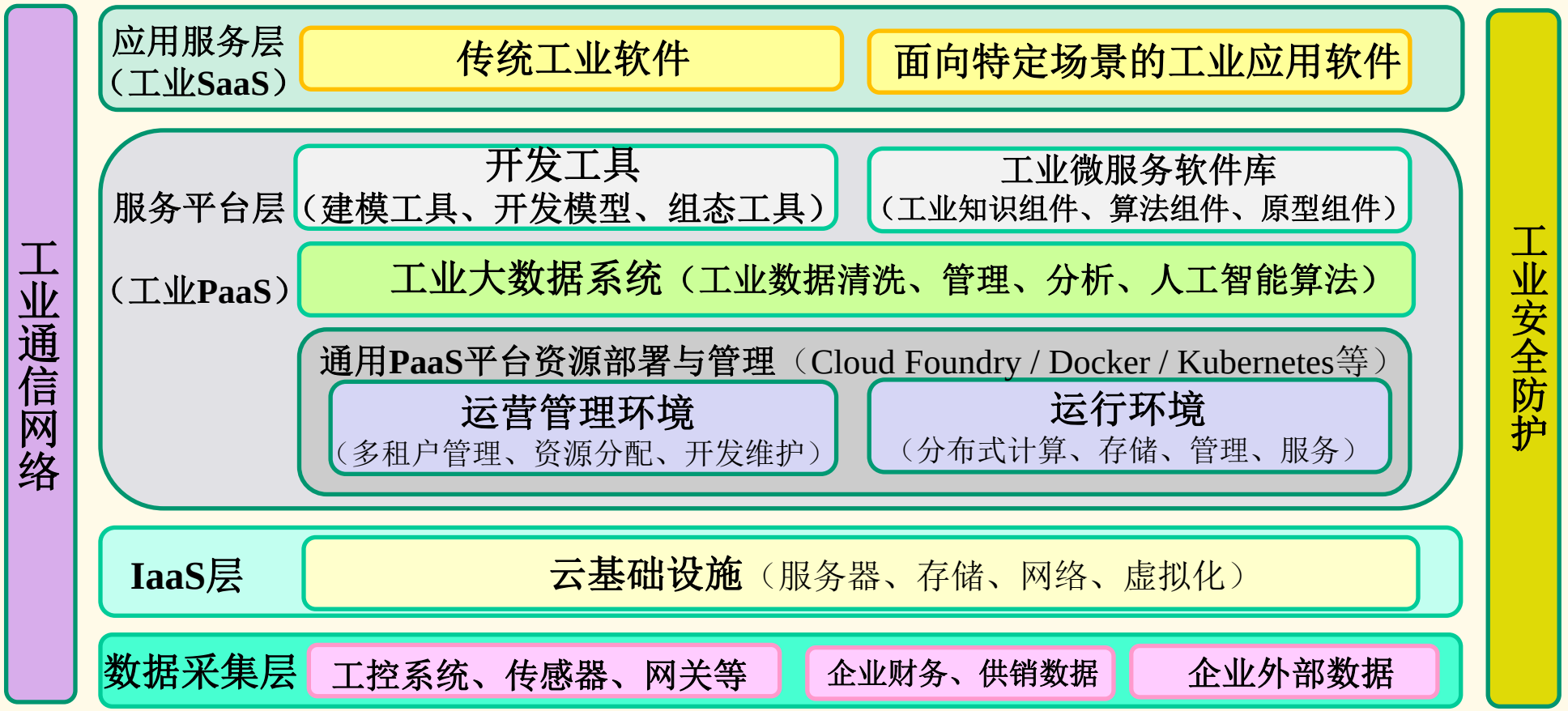
工业互联网应用---数字孪生技术

数字双胞胎(Digital Twin)借助安装在物理对象上的传感器数据和仿真手段来映射产品实时状态、工作条件或位置，获得物理对象的属性及状态的最新和准确的镜像，可用于监测、诊断和预测。



GE在风电场建设之前，利用 “数字风场” 的数字环境来配置每个风机，通过分析每一个汽轮机被馈送到它的虚拟孪生体的数据，能够提升20%的效率。

工业互联网的云平台



在工业云平台中，PaaS通用性较强，PaaS通常会采用开放的方式，承接各类型的SaaS及上层的应用。

参考：工业互联网联盟，工业互联网平台白皮书，2017年9月

沈阳鼓风机厂开发建立了沈鼓云服务平台，接入遍布全国的沈鼓大型装备的实时数据，远程监测与故障诊断。沈鼓在全国有大型机组1600台，利用云服务每年每机组可减少0.3次非计划停机，可减少直接损失4.8亿元，每机组运行效率可提升1个百分点，即减少能耗6.3亿元。

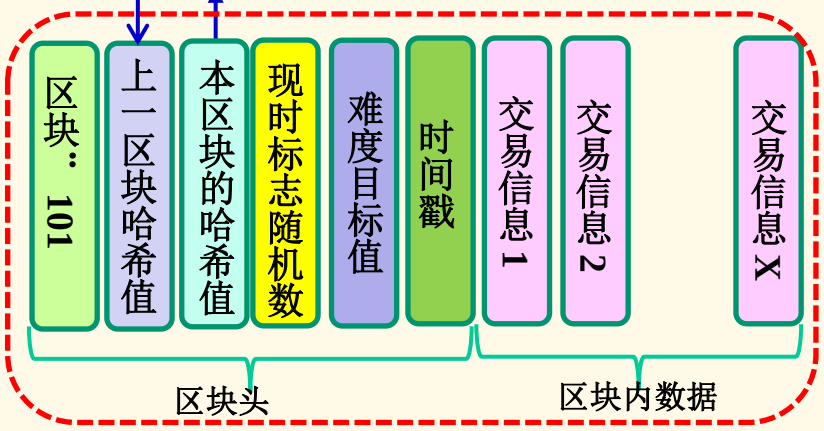
交易分布记账

区块链可完善交易记录、监管合规、避免交易纠纷、为共享数据提供了公平交易的平台。

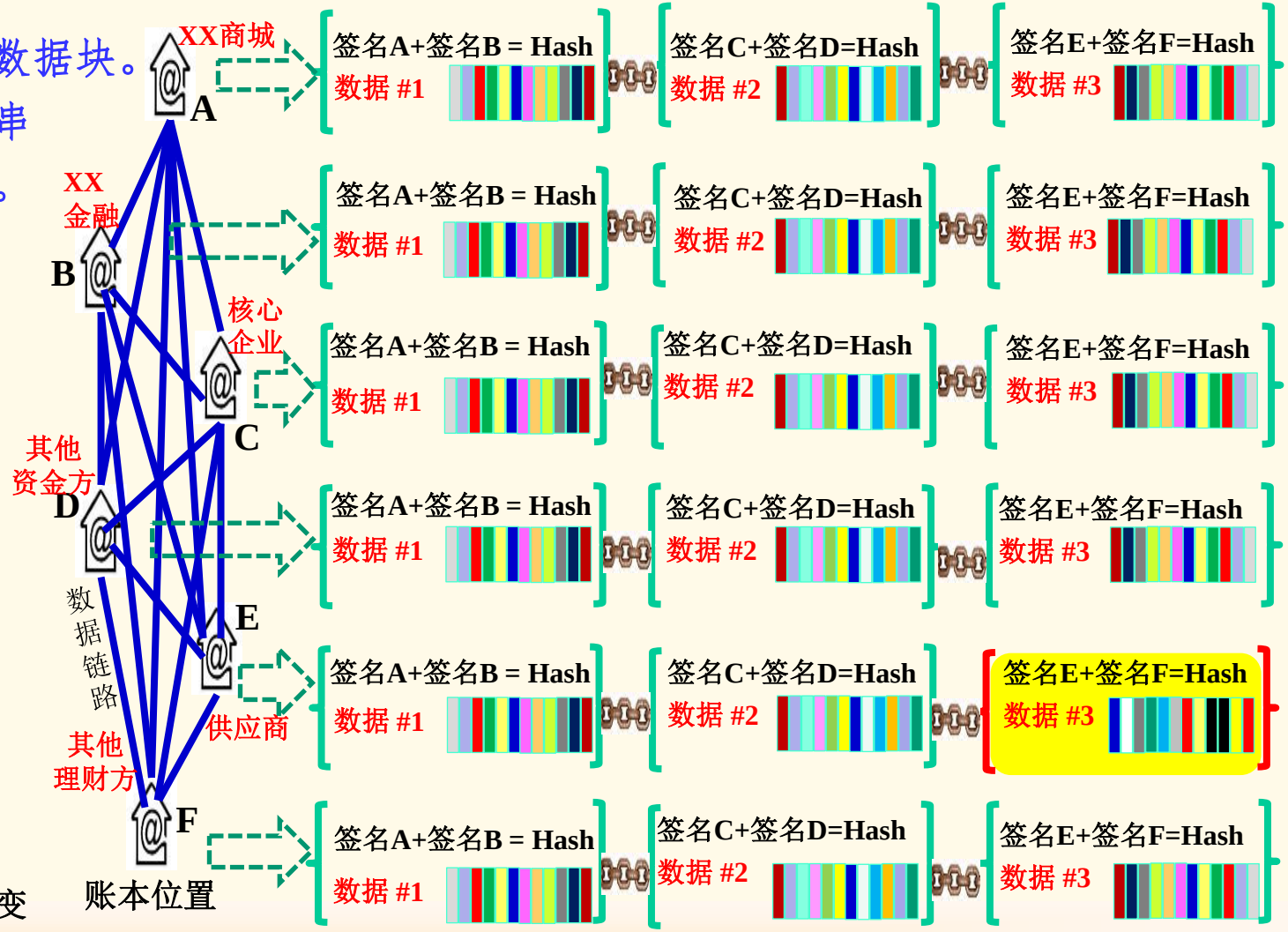
区块是包含带时间戳的数字资产交易信息的数据块。

区块链（BlockChain）是由密码关联的区块串组成的分布式数据库，也被称为分布式账本。

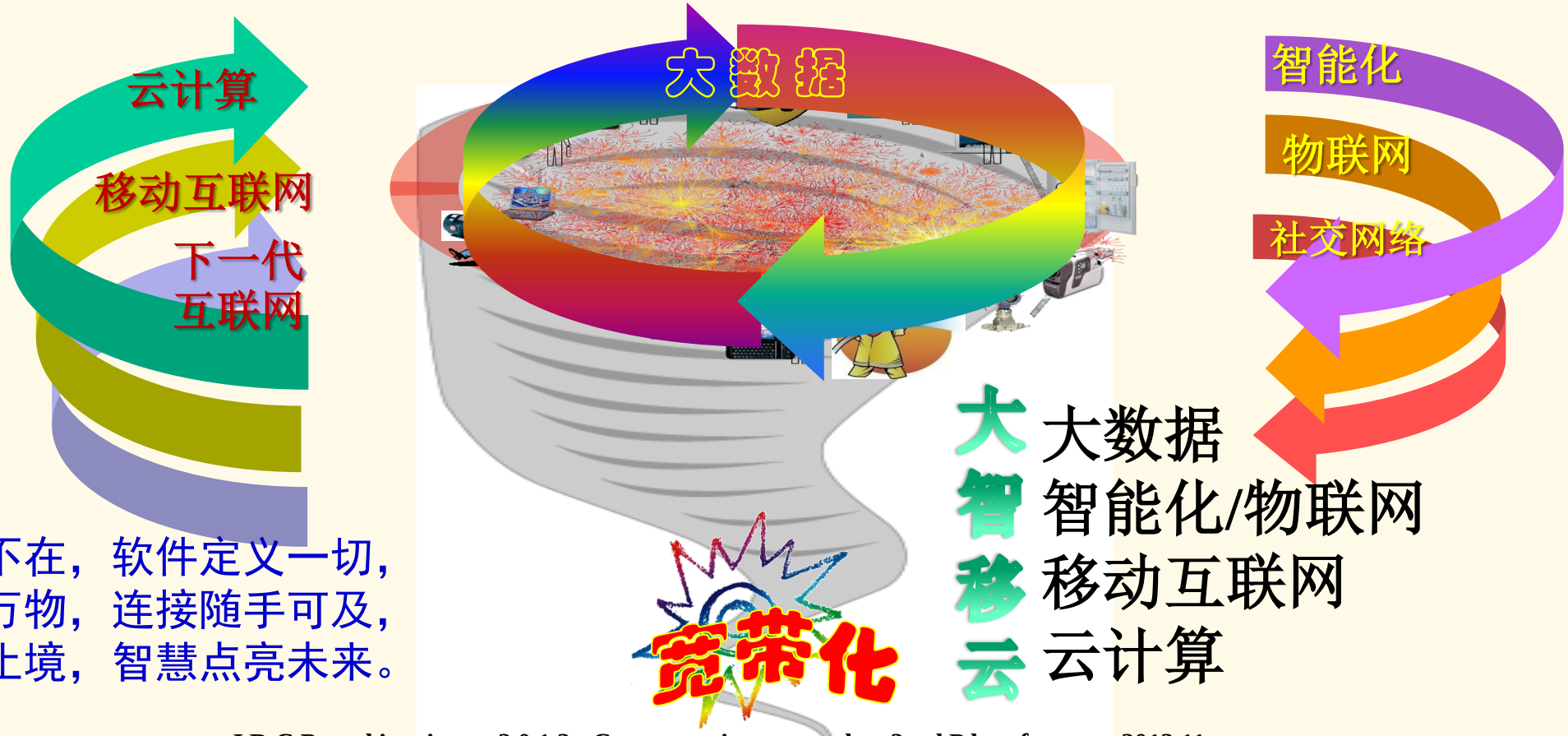
一旦信息经过验证并添加到区块链就会永久地存储，区块链的账本复制到多个位置，所有当前参与的节点共同维护交易及区块链，实现分布式记账，因此区块链的稳定性与可靠性很高。



哈希值是交易信息的摘要，能发现交易信息任一比特的改变



大智移云开拓信息化发展新阶段



计算无处不在，软件定义一切，
网络包容万物，连接随手可及，
宽带永无止境，智慧点亮未来。

大 大数据
智 智能化/物联网
移 移动互联网
云 云计算

IDC Predictions 2013: Competing on the 3rd Platform, 2012.11

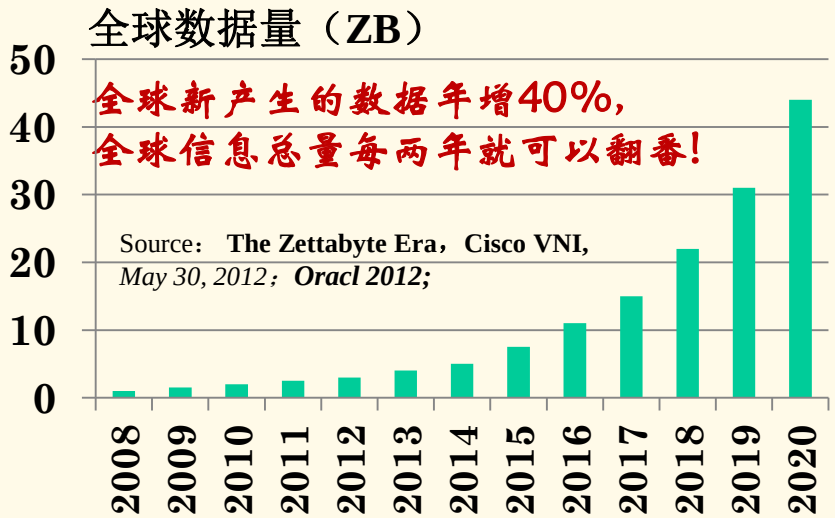
IDC报告创新平台三阶段

计算机 1985 → 互联网 2005 → 移动宽带、云服务、社交应用和大数据 2020

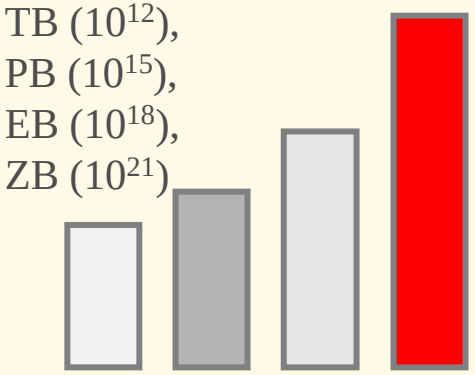
支撑2020年信息产业收入40%和增长98%

要加强关键信息基础设施安全保护，强化国家关键数据资源保护能力，增强数据安全预警和溯源能力。要加强政策、监管、法律的统筹协调，加快法规制度建设。要制定数据资源确权、开放、流通、交易相关制度，完善数据产权保护制度。----习近平，2017.12.08

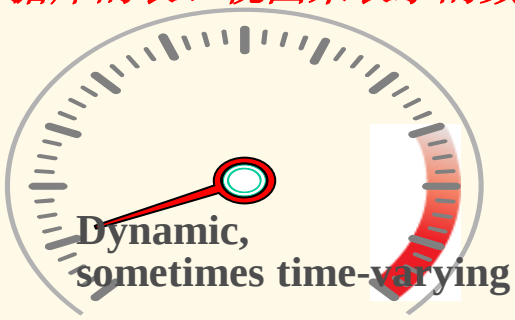
数据无尽增长

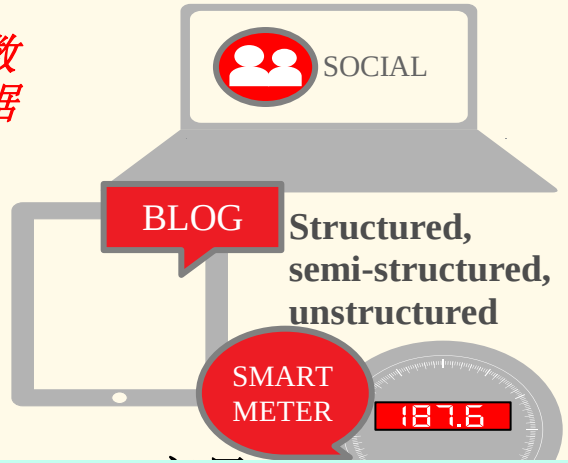


光盘重量 = 43078400 吨
相当424艘尼姆兹号航母重量



结构化——能以表格或关系数据库的表、视图来表示的数据

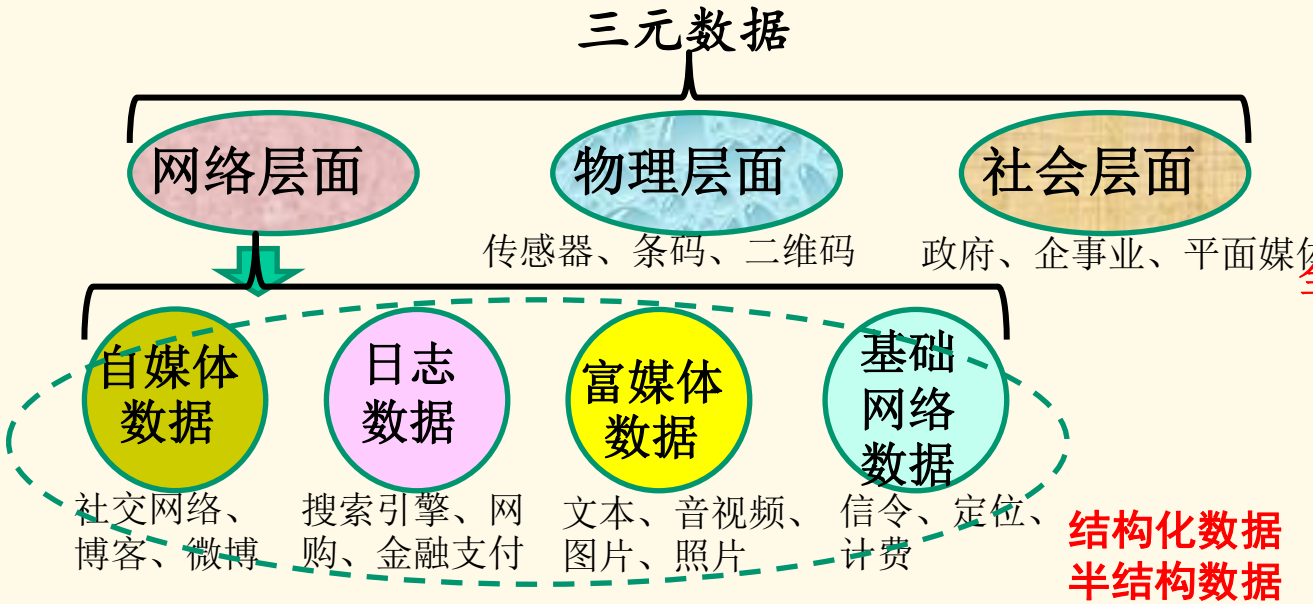




大数据价值密度低，
难挖掘但价值可贵

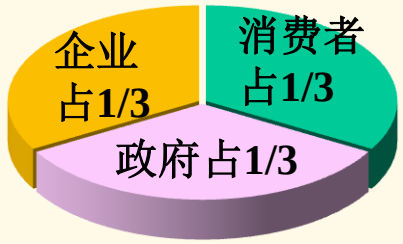
“大数据代表了由大容量、快速增长和多样性表征的、需要特定技术和分析方法将其转换为价值的信息资产”——4th International Conference on Integrated Information

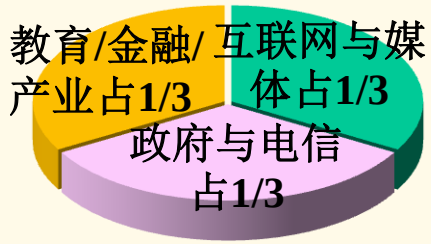
三元数据融合



结构化数据
半结构数据
非结构数据

一些分析表明，2020年全球数据的40%来自传感器



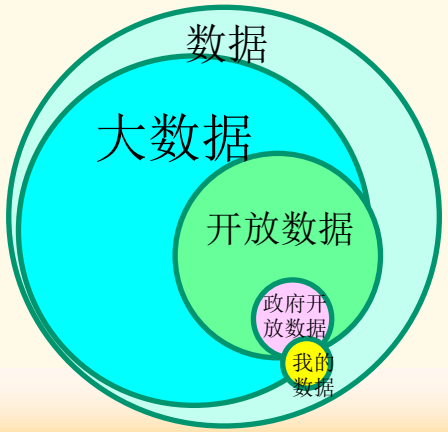


全球2013年所产生数据的来源　　IDC 2014年报告

http://china.emc.com/leadership/digital-universe/2014iview/executive-summary.htm

数据分类：

- 国家安全数据；
- 商业秘密数据；
- 个人隐私数据；
- 其他数据。



消费者大数据

超8亿客户，每天获取的新数据14TB，累计存量300PB。

4.1亿的用户，大数据平台存储容量

日处理数据量达到200TB。累积数据量达到30PB。

月度活跃用户超5亿，单日新数据超50TB ，累计超数百PB

月活跃用户近7亿，每天数据处理能力100PB。

月活跃账户9亿，数据每日新增数百TB，总存储量数百PB。

日新增1.5PB，扫描5PB，2016年累计100PB，年增300%

日活用户3000万，累计3亿，日处理数据量7.8 PB

30%国人用外卖，周均3次，美团用户6亿，数据超4.2PB

用户4.4亿，每日新增轨迹数据70TB，处理数据超4.5PB

2亿用户700万辆单车
每天3000万次骑行30TB数据

每天线上访问量上亿，每日数据增量400TB，存量超50PB

MIUI联网激活用户3亿，小米云服务数据总量200PB

1PB足够存储全美国人口DNA两遍，1PB歌曲在手机MP3上可以连续播放2000年， 1PB的照片的并排可绕地球2圈

金融大数据

京东 + 5.5亿个人客户，全行数据超60PB（安防视频40PB，外部400TB）

+ 5亿个人客户，手机银行1.8亿，网银2亿，数据存储能力100PB

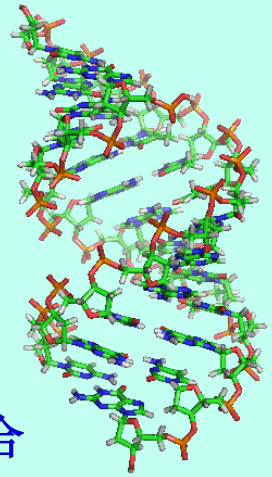
du + 5.5亿个人客户，日处理数据1.5TB，数据存储15PB

中国平安有8.8亿客户的脸谱和信用信息，还有5000万个声纹库。

+ 5亿个人客户，手机银行客户1.15亿，电子渠道业务替代率94%。

医疗大数据

- 生命大数据
  - 人有10^{14}个细胞，10^{9} 个基因
  - DNA有 2X3X10^{9} 碱基
  - RNA有 1X10^{9}碱基
  - 蛋白质有20种氨基酸，有 2X 10^{19}组合

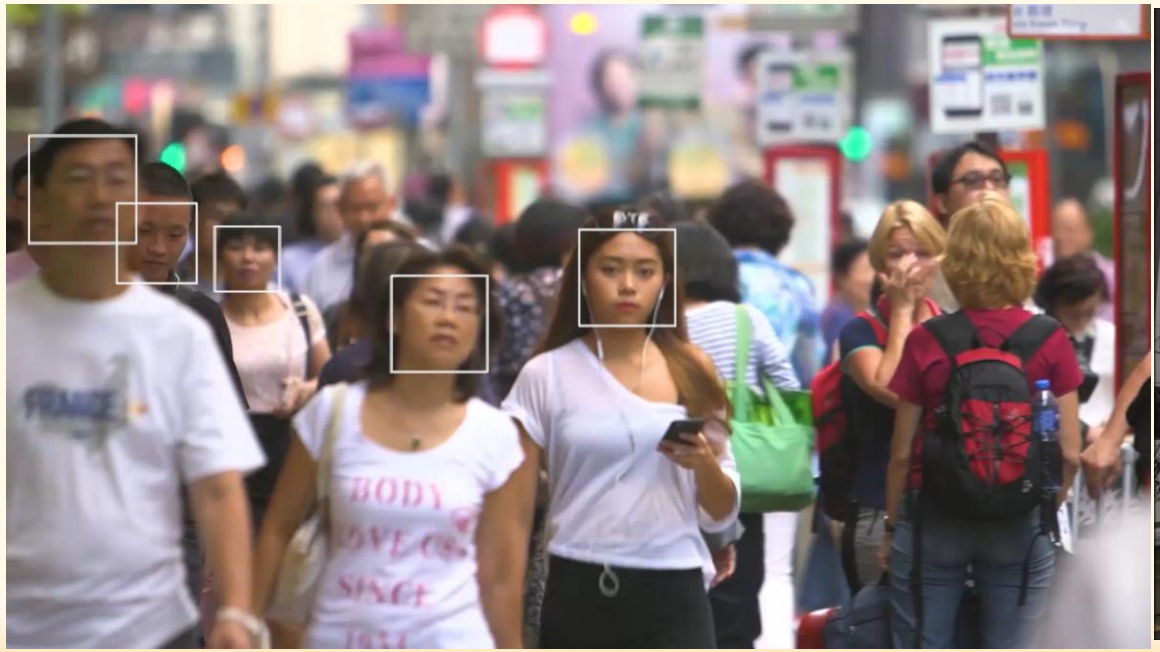


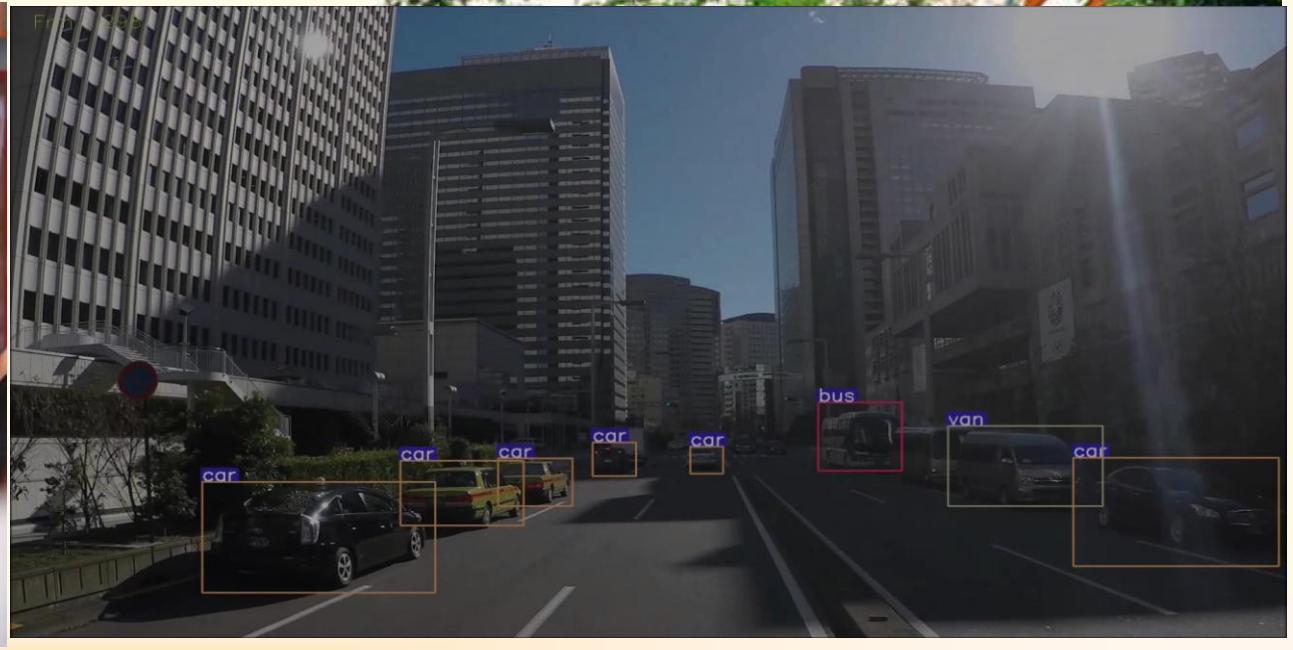
- 一次全面的基因测序产生的个人数据 100GB~600GB (Leah, 2014)
- 华大基因公司2017年产出的数据达到1个EB。

- 影像数据量（可含 5万像素值）
  - X线 30 MB，乳腺X线 120 MB
  - 3D核磁 150 MB
  - 一张CT图像 150 MB，3D CT扫描 1GB
  - 标准的病理图 5GB
  - 功能性磁共振影像数万 TB 级
- 至2015年美国平均每家医院需管理665 TB数据量，个别医院年增数据达PB。
- 到2020年全球医疗数据将增至35ZB，相当于2009年数据量的44倍。

城市大数据

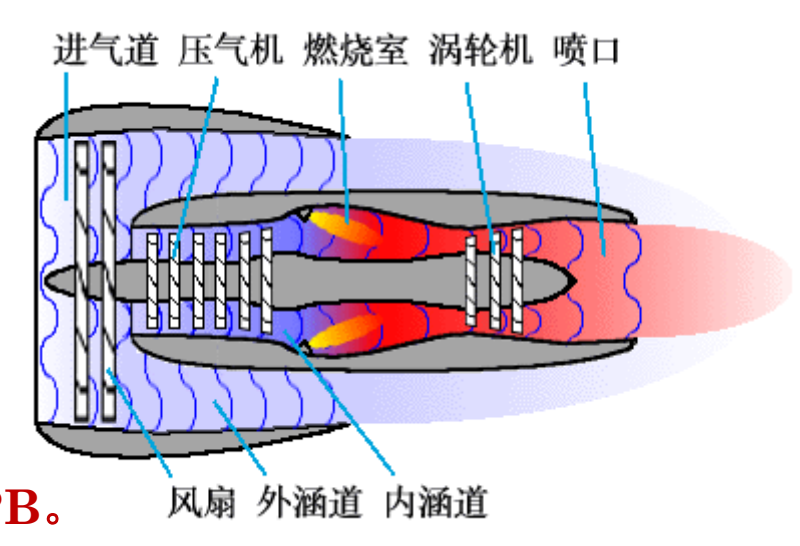
- 一个8Mbps摄像头产生的数据量是3.6GB/小时，一月为2.59TB。很多城市的摄像头多达几十万个，一个月的数据量达到数百PB，若需保存3个月则存储量达EB量级。
- 北京市政府部门数据库总量2011年63PB，2012年95PB，现在为数百PB。全国政府大数据加起来为数百甚至上千个阿里的体量。

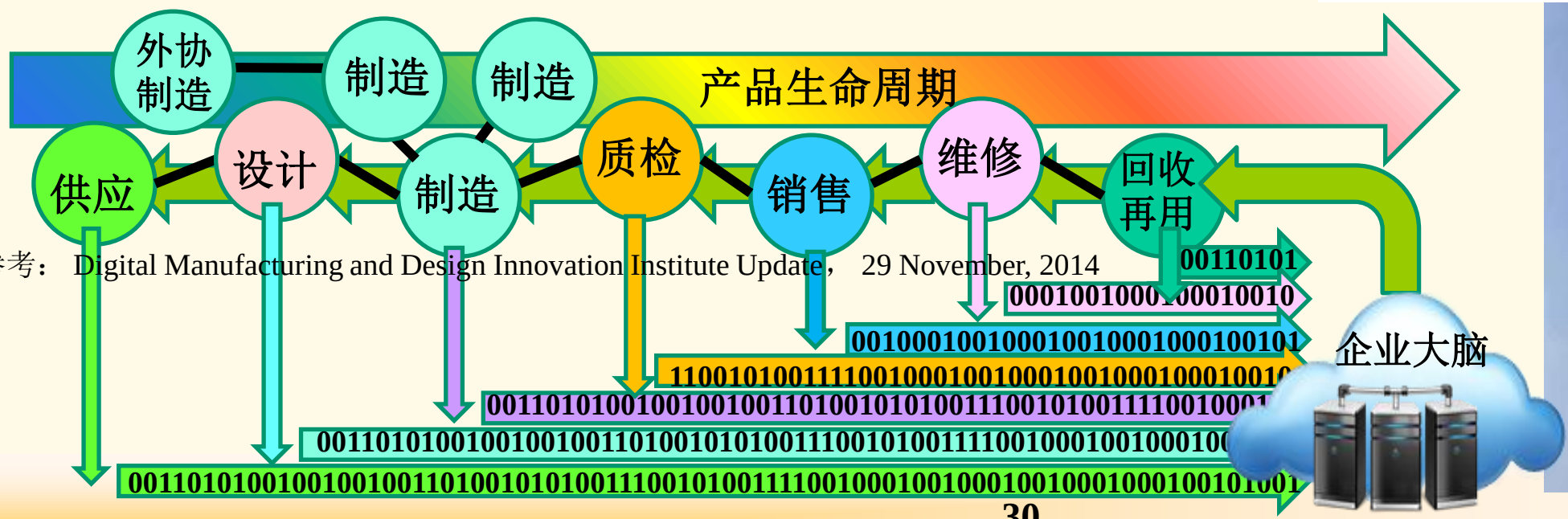




工业大数据

- Rolls Royce 公司对飞机引擎做一次仿真将产生数十TB的数据。一个汽轮机的扇叶在加工中就产生0.5TB的数据，扇叶生产每年将收集3PB的数据。叶片运行数据为588GB/天。
- GE在出厂飞机的每个引擎上装20个传感器，每引擎每飞行小时能产生20TB数据并通过卫星回传，每天可收集PB级数据。
- 清华大学与金风科技共建风电大数据平台，2万台风机年运维数据为120PB。





参考： Digital Manufacturing and Design Innovation Institute Update， 29 November, 2014

大数据的采集

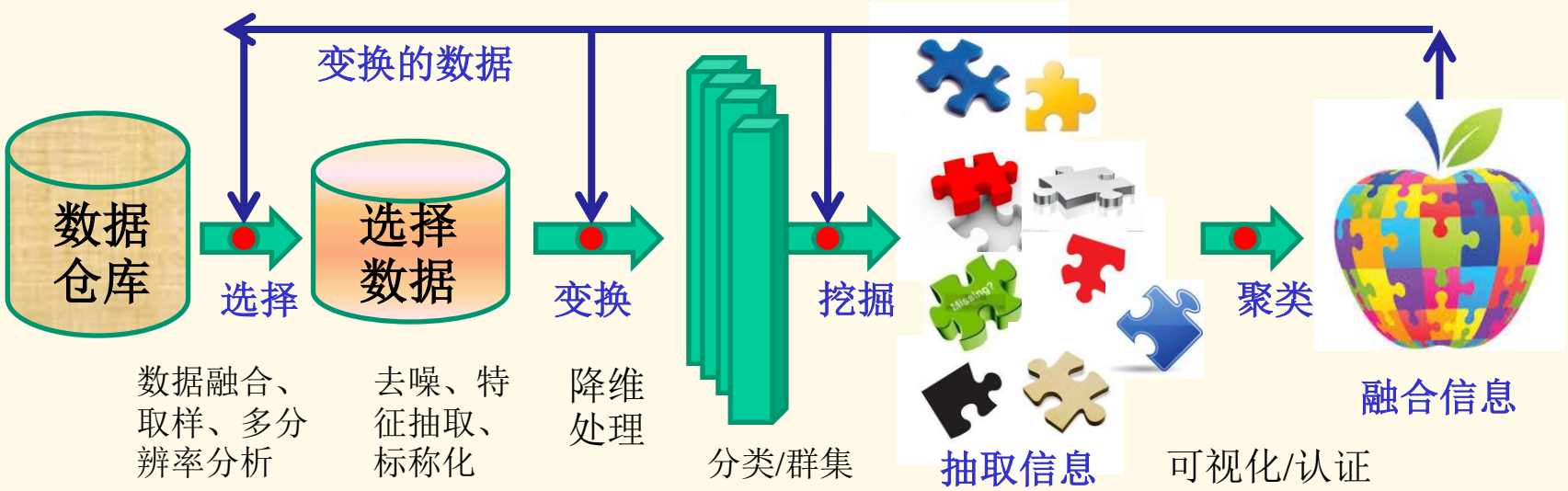
- 大数据的采集除了传感器外，采集还涉及去重、清洗、过滤及格式转换甚至降维等处理，问题是如何在这些处理过程中降低数据规模提升数据质量而又不损失数据的价值？
- 数据采集很重要的问题是降噪，噪声和干扰会使数据失真，例如攻击者将对抗样本输入机器学习模型，机器产生幻觉导致误判。一张熊猫图片，被加入微小噪声后就导致系统将熊猫识别为长臂猿。

大数据的存储

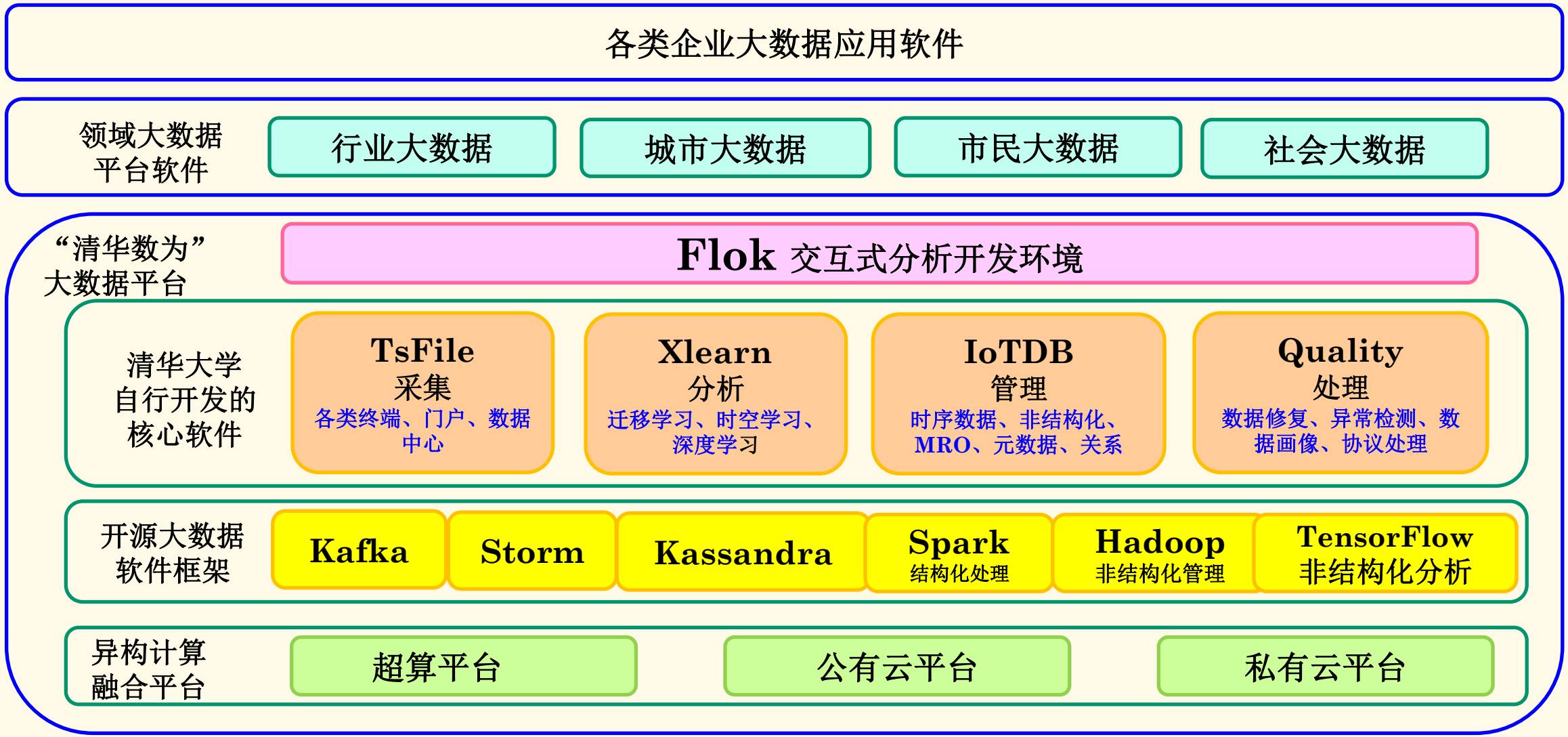
- **安全性**----为每一个部门或一项任务建立一个大数据的存储管理平台并不现实，较为实用的方法是共享一个大数据存储池，需要有访问隔离控制来实现数据访问的安全性。
- **高效性**----老数据的价值不如新数据，通过对用户层透明的压缩处理来实现空间及带宽高效利用。
- **实时性**----目前大数据存储系统普遍采用的是DAS的方式，将计算资源搬迁到数据的存储节点上，避免计算节点与存储节点间网络带宽的瓶颈，但跨节点数据访问管理仍需通过网络。
- **可扩展性**----容量规划要考虑到系统扩张后的存储空间与带宽升级。通常采用添加存储节点来达到扩容的目的，意味着将使用分布式主节点群管理，对数据管理软件及系统复杂度都是挑战。
- **兼容性**----存储系统需要兼容结构化、半结构化及非结构化数据，兼容批处理、交互式和流式等数据传输机制，兼容各种存储介质。
- **容错性**----低廉成本多节点的存储设备意味着高的硬件故障率，需要强大的容错软件管理能力（多副本即冗余备份），还要有可靠的运维监控系统。

大数据的挖掘

- 聚类分析----大数据要通过分解和融合形成聚类，难点是提取集成方向，构建集成模型
- 分类分析----如何利用无标记的样本来改善小数据集上构建的分类器
- 关联性分析----不同的数据集中呈现复杂的关联关系，如何挖掘其隐含的关联。

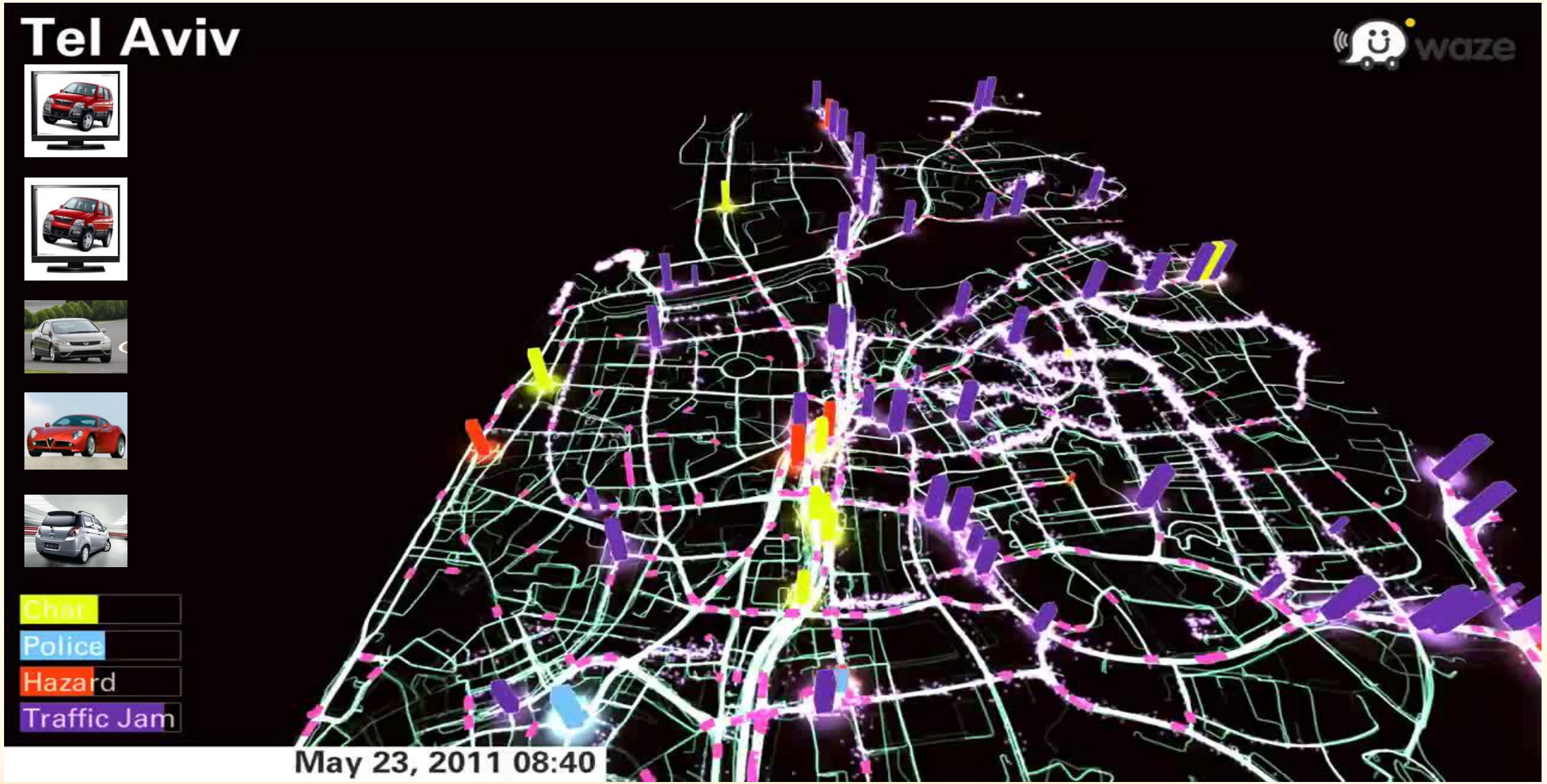


大数据分析软件架构



参考：清华大学 孙家广院士

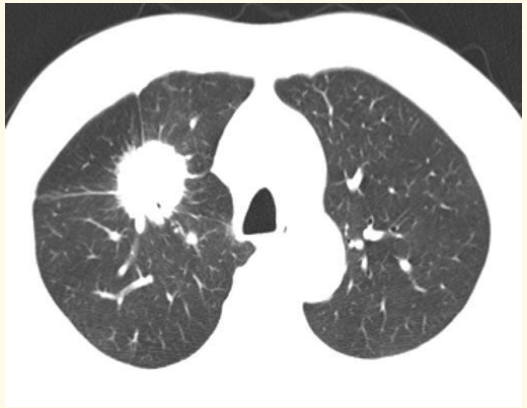
大数据的呈现

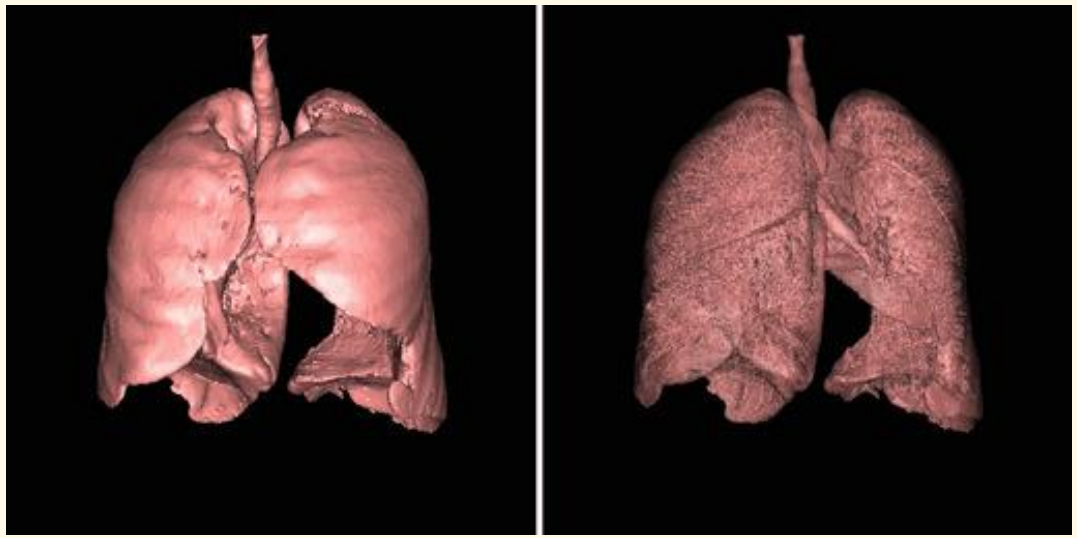

Tel Aviv
waze
Chat
Police
Hazard
Traffic Jam
May 23, 2011 08:40

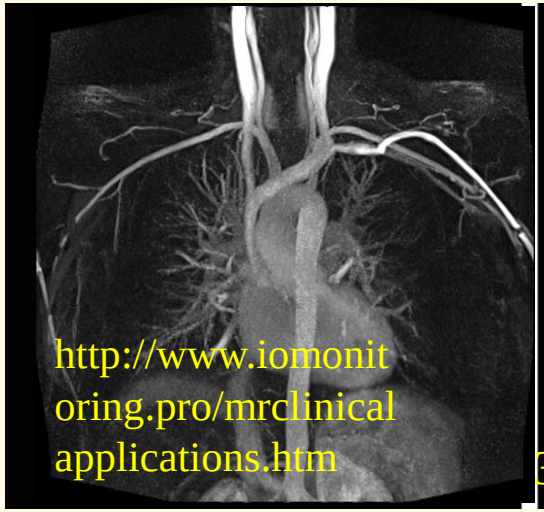
大数据的可视化

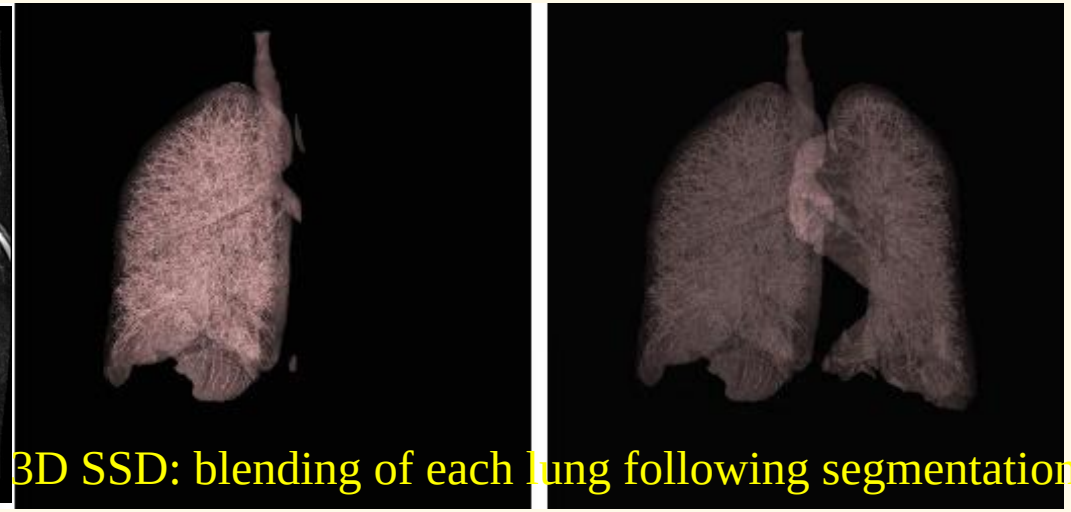
数据可视化好处	比例
改进决策	77%
优化自组织数据分析	43%
改进合作/信息共享	41%
向用户提供自助服务	36%
提升投资回报	34%
节省时间	20%
降低IT的负荷	15%

- 大数据可视化需并行计算，难点是如何分解为多个可同时运行的独立的任务。
- 大数据的动态性要求可视化快速响应。
- 可视化并非意味着准确性，不能代替批判性思维。









3D SSD: blending of each lung following segmentation

大数据的开放

2014年6月美国实施openFDA，已公开了食品、药品、医疗器械和化妆品有关不良反应、召回、标识等数据，截至2015年openFDA连接全球21 000个系统，有超过2 000万次的数据调用，已有30个手机软件在使用FDA的开放数据提供服务。

《美国阳光基金》开放数据原则

- ✓ 完整性
- ✓ 重要性
- ✓ 时效性
- ✓ 物理与电子方式接入的方便性
- ✓ 可机读性
- ✓ 非歧视性
- ✓ 使用通用标准
- ✓ 无需授权
- ✓ 持久性
- ✓ 低使用成本

我国政府数据开放存在问题：

不愿共享开放。部门各自为政，以信息不对称作为管理手段。

不敢共享开放。我国数据信息管理方面制度不明确，导致开放风险责任难以界定。

不会共享开放。缺统一标准和规范，缺技术和人才，安全保密和可持续性难保证。

全球开放数据晴雨表

	国家	排名	总分	准备度得分	执行力得分	影响力得分
2014	中国	46	28.12	52	24	19
2015	中国	55	21.16	45	15	8
2016	英国	1	100	99	100	94
	加拿大	2	90	96	87	82
	中国	71	20	46	10	11

来源：Open Data Barometer, http://www.opendataresearch.org/dl/odb2015/Open-Data-Barometer-2015-Global-Report.pdf

打通信息壁垒，形成覆盖全国、统筹利用、统一接入的数据共享大平台，构建全国信息资源共享体系。

---习近平在中共中央政治局第二次集体学习会上讲话，**2017.12.08**

大数据的共享

- 政府数据共享（中央政府部门间和中央与地方政府间）是不对称的，如何划分接入权限？
- 政府与企业间数据共享是有条件的；
  - “要加强政企合作、多方参与，加快公共服务领域数据集中和共享，推进同企业积累的社会数据进行平台对接，形成社会治理强大合力”。---习近平，2017.12.08
  - 从国家安全出发，政府有权调用企业数据，但除此之外，企业是否有义务向政府提供数据？在政府调用企业数据的情况下，如何保证企业的商业秘密不泄漏。
- 企事业单位间的数据共享需要有共享机制以实现利益的平衡与安全责任
  - 美国FDA与哈佛大学合作，以哈佛大学作为协调中心联系FDA、医院、医药企业等，该中心根据各合作方的需求研发数据抽取、转化的标准化模型，各合作方在自己的数据上进行运算并向该中心提供分析结果和必要信息，最后由该中心进行整合反馈给需求方。这一方式不改变数据的所有权与保护权限，实现了数据按需利用。

大数据的交易

- 数据的确权----
  - 数据交易的前提是确权，运营商和ICP所收集的用户数据原则上所有权是用户，但运营商和ICP拥有对数据脱敏并挖掘分析后的数据的所有权，但也有保护数据安全与隐私的责任。
  - 无数据所有权但有数据的公司通过挖掘可向政府和企业提供咨询报告，但不能转售数据；中介公司可作为第三方开展数据交易服务的中介，但不能截留数据。需要对数据扩散范围等有明确的责任界定和处罚协议。
- 数据质量的评估----

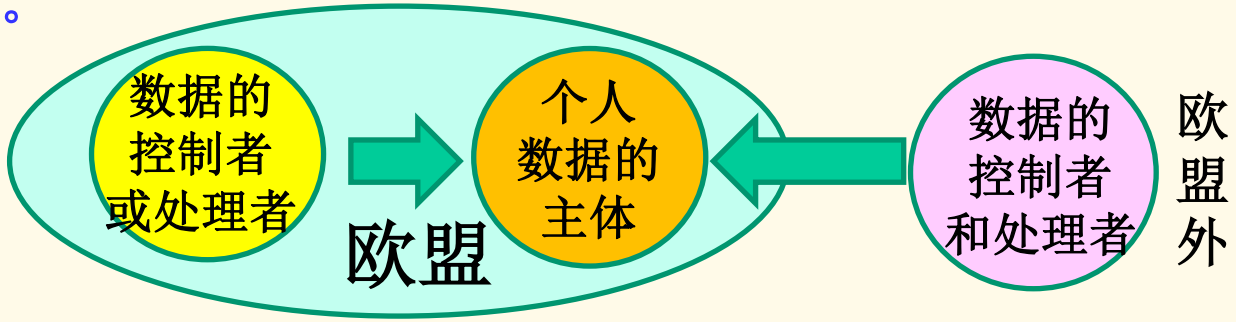
  涉及数据的有效性、可信性、更新频率和数据源稳定性，数据溯源有助于判别数据的真实性，但数据源往往本身就是隐私敏感数据。目前最大的问题是无法从数据挖掘结果来判定数据的质量。
- 数据的定价----

  物质与财富资产有价格，数据作为资产也应有价值与价格，这还涉及大数据质押与融资的需要，但对数据作为资产如何管理目前还缺乏相应的制度。目前数据交易只有基于协商定价。

大数据的隐私保护

- 2018年5月25日欧洲《通用数据保护法案》(GDPR)正式生效，涉及个人隐私数据的保护，此法也适用于欧盟之外的数据控制和处理者。

  - 一般违规行政罚款上限是1000万欧元或该企业上一财年全球营业总额的2%(取二者中较高值)，对严重违规者将加倍。



  - GDPR的问题是个人数据定义的范围太宽（与个人隐私、专业或公共生活有关的任何信息，包括姓名、照片、电子邮件地址、银行账号、社交网络上邮件、医疗信息或计算机IP地址等）。

- 中国需要制定个人隐私数据保护的法律，但要审慎考虑保护范围的宽严，既要保护民众和企业的利益，又要避免阻碍互联网企业的发展。

大数据的人才培养

- 大数据需要跨学科人才----大数据技术涉及数学、统计学、计算技术（存储、算法、语言），还需要有产业、经济和法律知识。
- 大数据也是实践科学----数据挖掘成功的案例往往都是大数据专家与传统产业的企业专家紧密合作，吃透企业生产流程，甚至还借助大数据企业拥有的外部数据，经反复试验并调整数据挖掘算法才得以成功。
- 大数据人才培养需要政产学研合作----大学是培养数据人才的摇篮，但大学缺乏数据，需要与拥有数据的政府与企业合作。
- 培养大数据人才更重要的是创新精神----现在的大数据在若干年后就不算大数据了，还有更大的数据，现有的存储、挖掘等技术都不适应未来的大数据发展，重要的是培养大数据思维与创新精神。

谢谢！