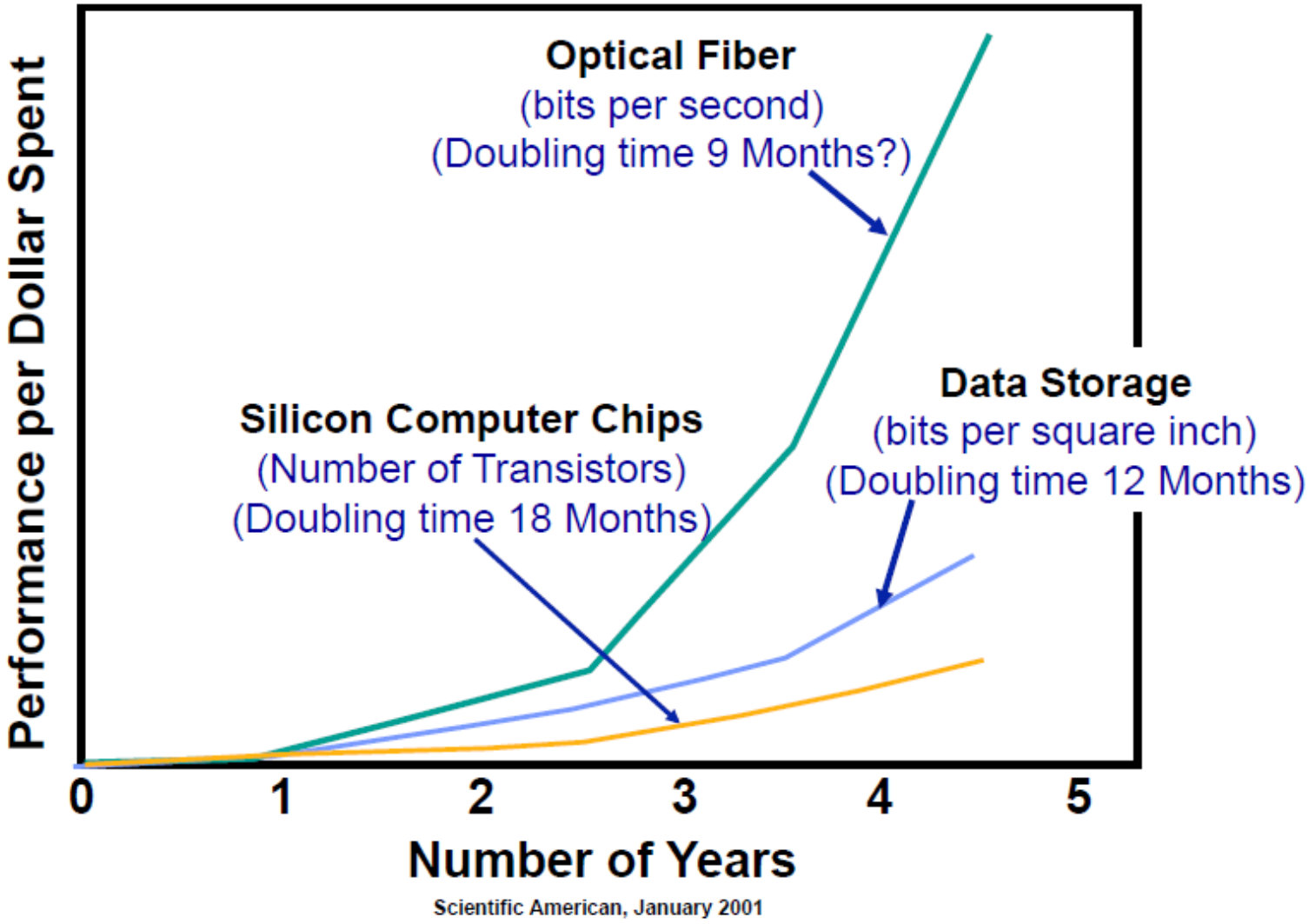
# 大数据科学概论
# Introduction to Big Data Science

吴文峻

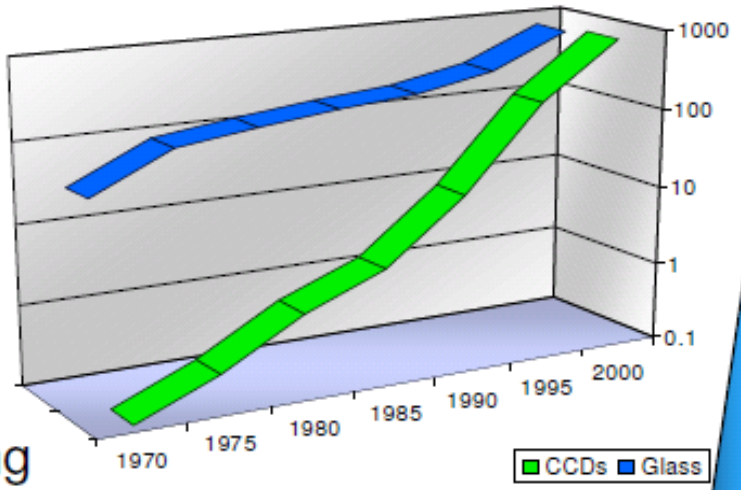
18秋大数据科学导论

# Moore's Law


Optical Fiber
(bits per second)
(Doubling time 9 Months?)
Silicon Computer Chips
(Number of Transistors)
(Doubling time 18 Months)
Data Storage
(bits per square inch)
(Doubling time 12 Months)
Performance per Dollar Spent
0
1
2
3
4
5
Number of Years
Scientific American, January 2001

# Living in an Exponential World

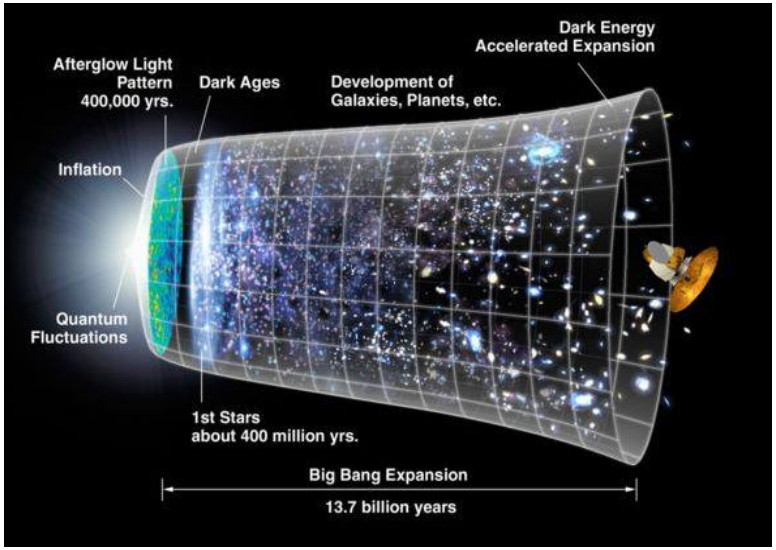
- Scientific data doubles every year
  - *caused by successive generations of inexpensive sensors + exponentially faster computing*




- Changes the nature of scientific computing
- Cuts across disciplines (eScience)
- It becomes increasingly harder to extract knowledge
- 20% of the world's servers go into huge data centers by the "Big 5"
  - *Google, Microsoft, Yahoo, Amazon, eBay*

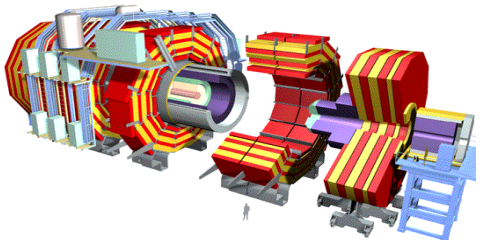
# Living in an Exponential World

现代科学的发展
(1)观测和试验手段的进步
(2)模拟规模和尺度的扩大




产生越来越多的科学数据，
需要分析和处理.

大型强子对撞机(LHC)
~10K scientists
33+ countries,
100PB data distributed
~1 Exabyte by 2012



# 第四范式 – 以数据为中心、数据驱动

实验科学

理论科学

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - \mathrm{K}\frac{c^2}{a^2}$$

计算科学

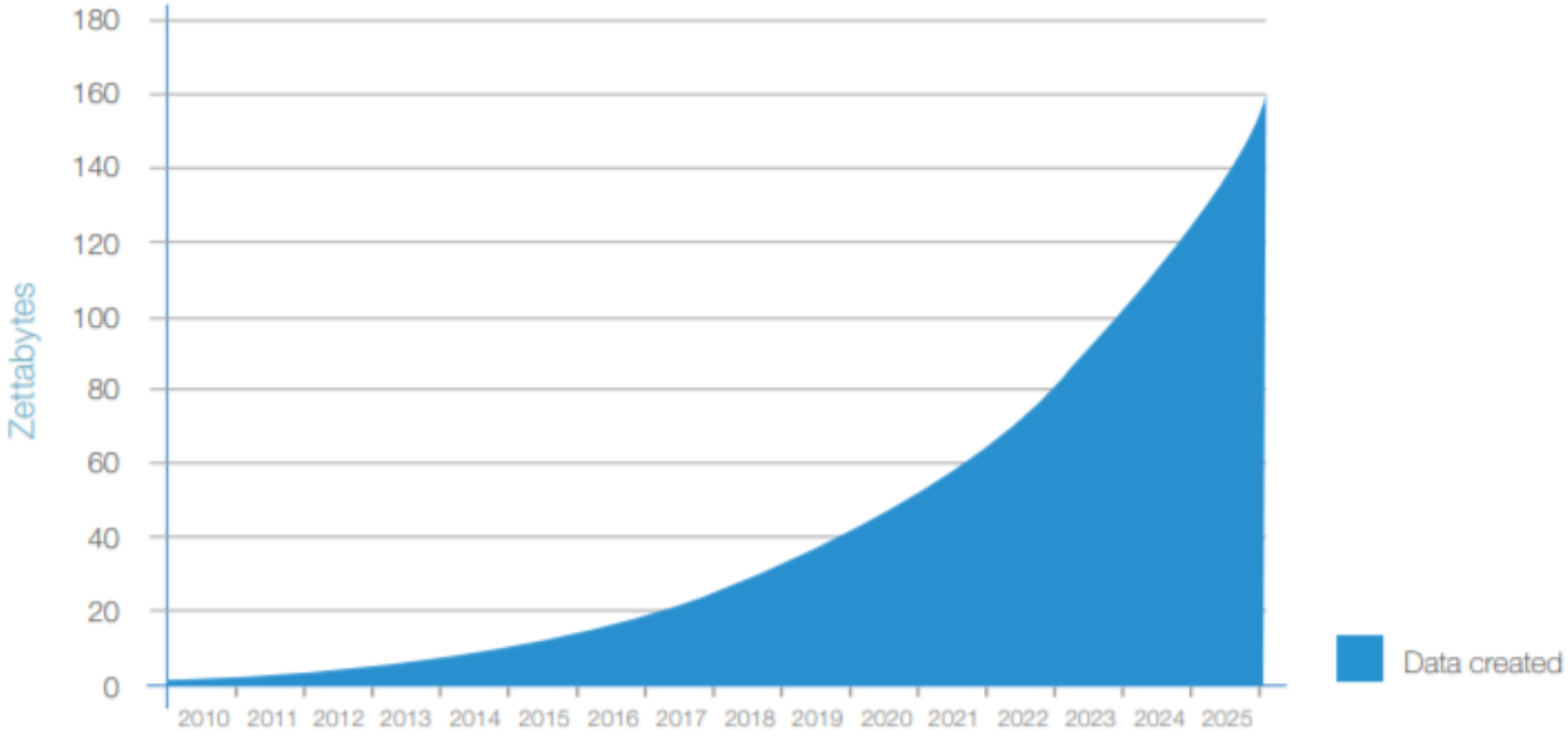
数据为中心的科研

当前科研问题空前复杂化、综合化

# 数据量级

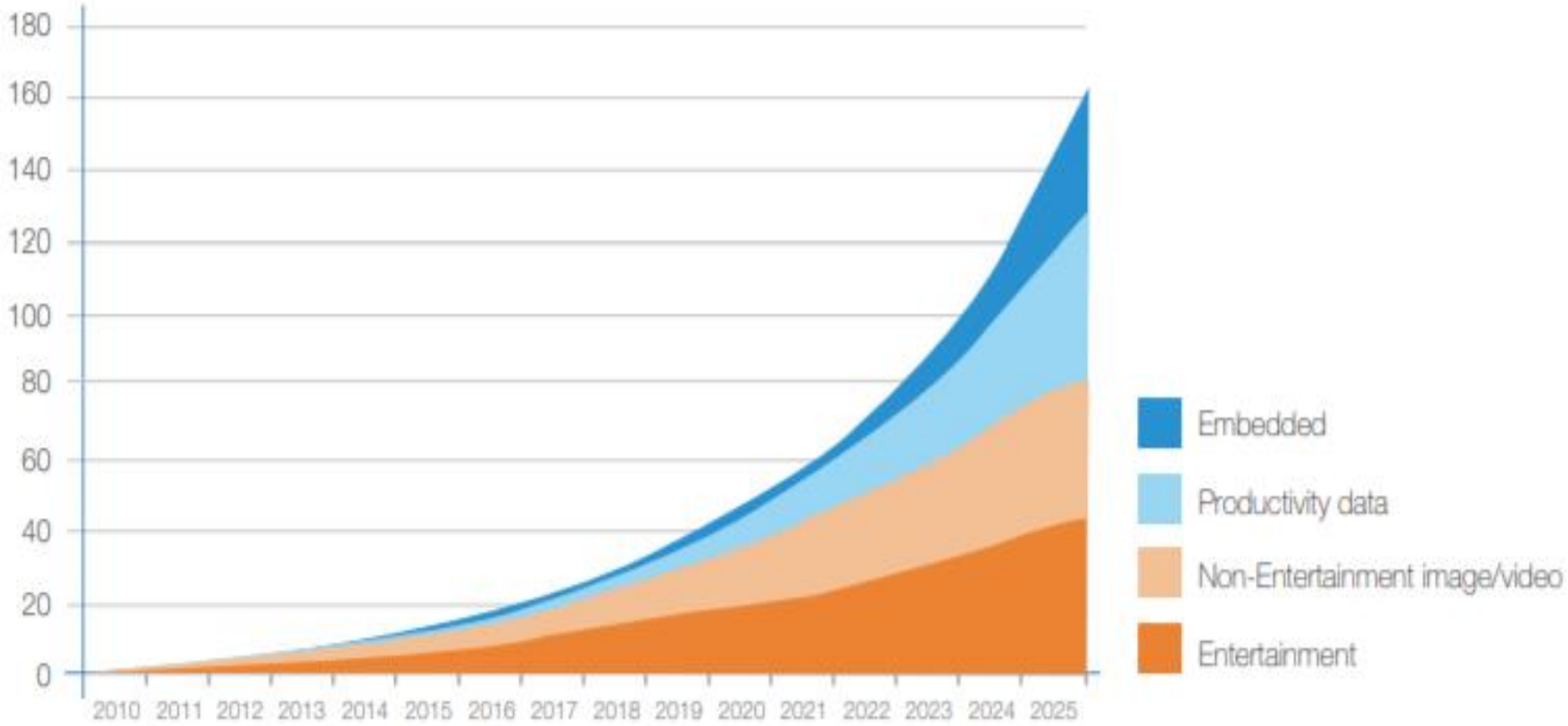
- $10^6$Bytes – MegaBytes
  ***A digital photo***
- $10^9$ Bytes – GigaBytes
  ***A DVD movie***
- $10^{12}$ Bytes – TeraBytes
  ***World annual book production***
- $10^{15}$ Bytes -- PetaBytes
  ***Annual production of one LHC experiment***
- $10^{18}$ Bytes -- ExaBytes
  ***Internet Traffic Per Month***
- $10^{21}$ Bytes – ZettaBytes
  ***World information production***

# Annual Size of the Global Datasphere



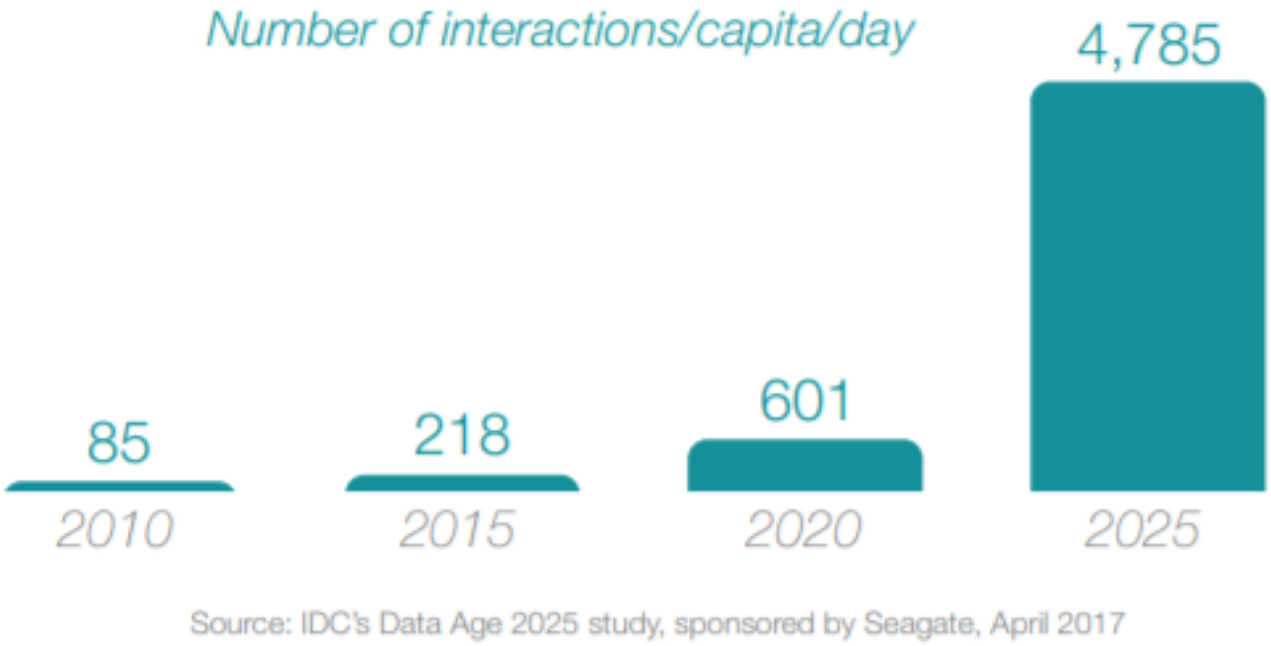

Source: IDC's Data Age 2025 study, sponsored by Seagate, April 2017

# Data Creation by Type



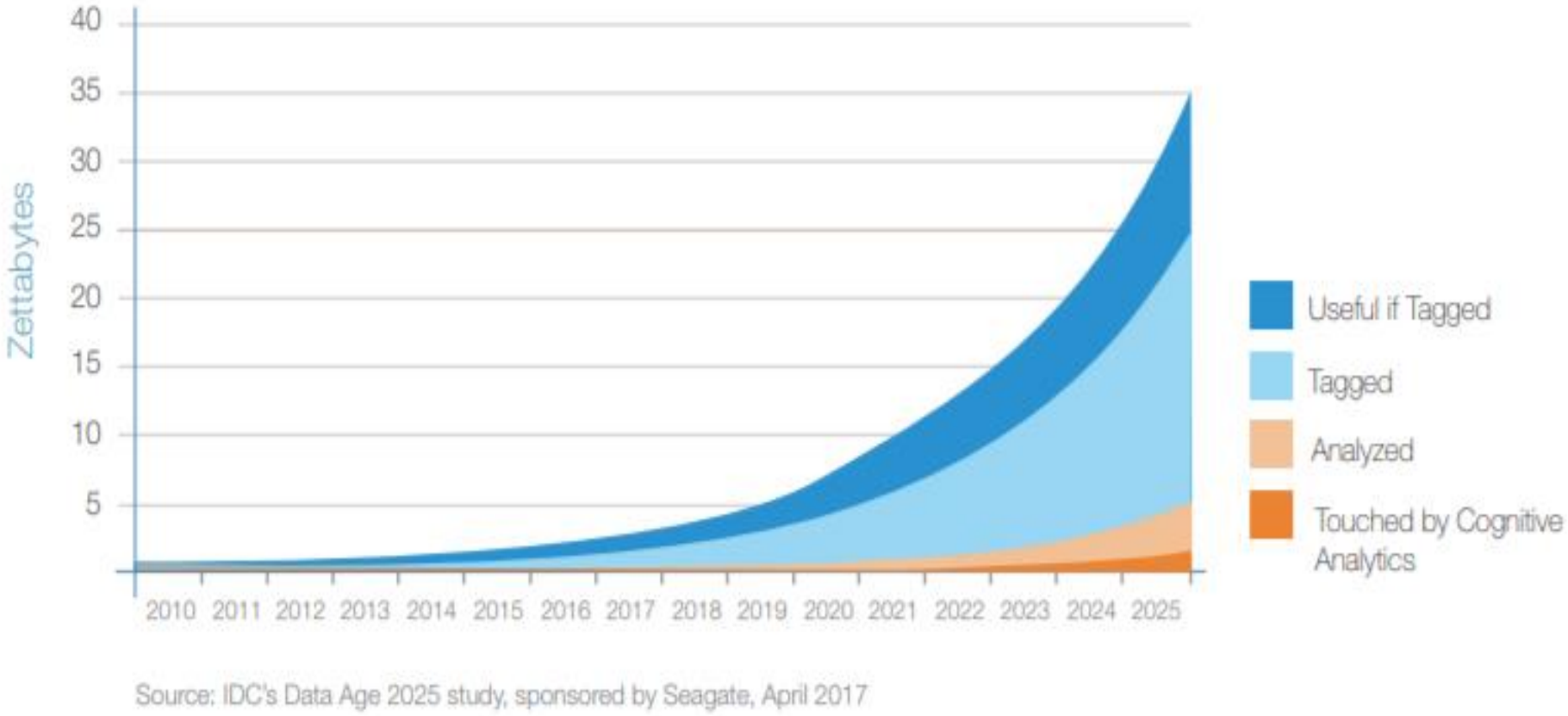

Source: IDC's Data Age 2025 study, sponsored by Seagate, April 2017

# Interactions per Connected Person per Day




Source: IDC's Data Age 2025 study, sponsored by Seagate, April 2017

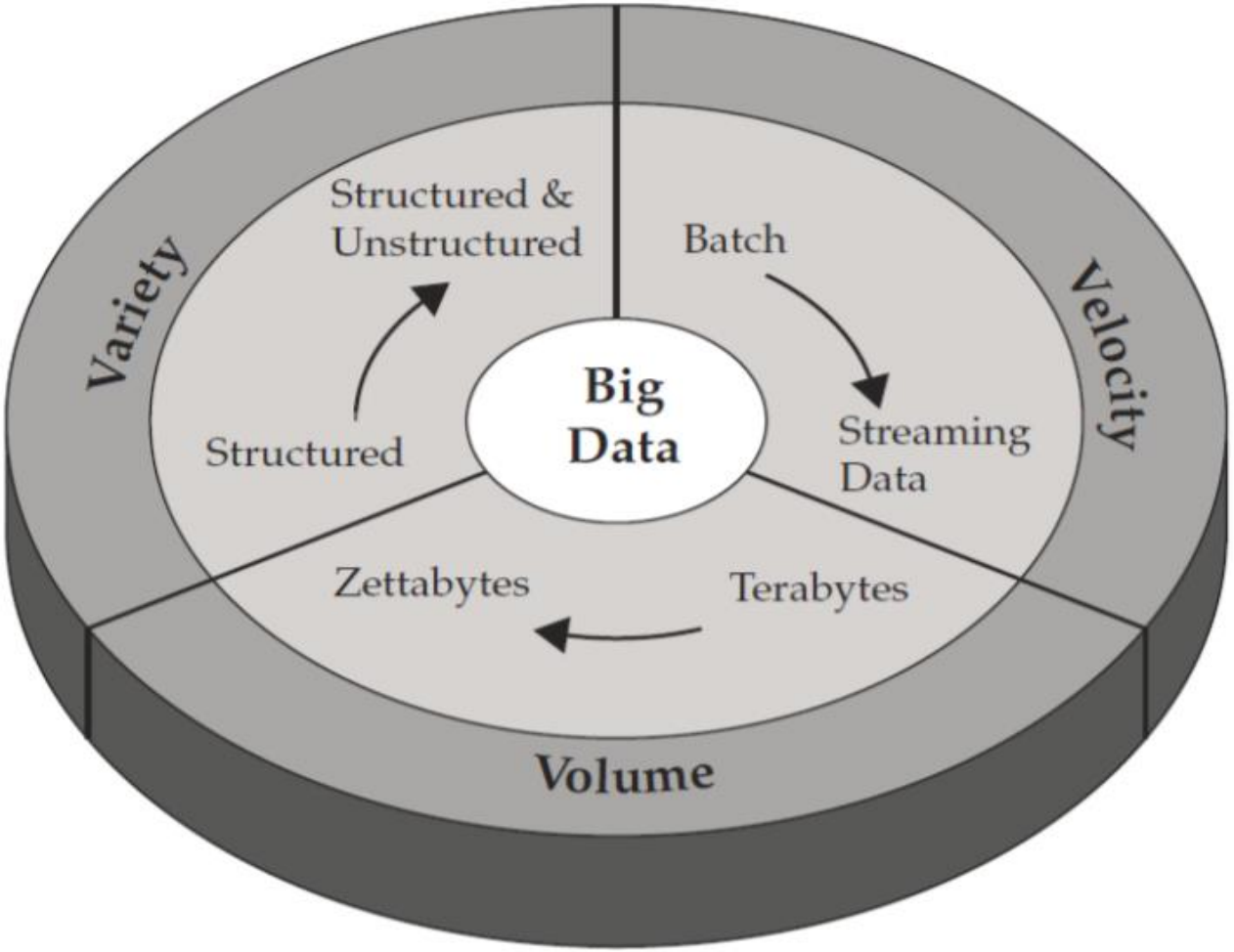
# Data Tagging



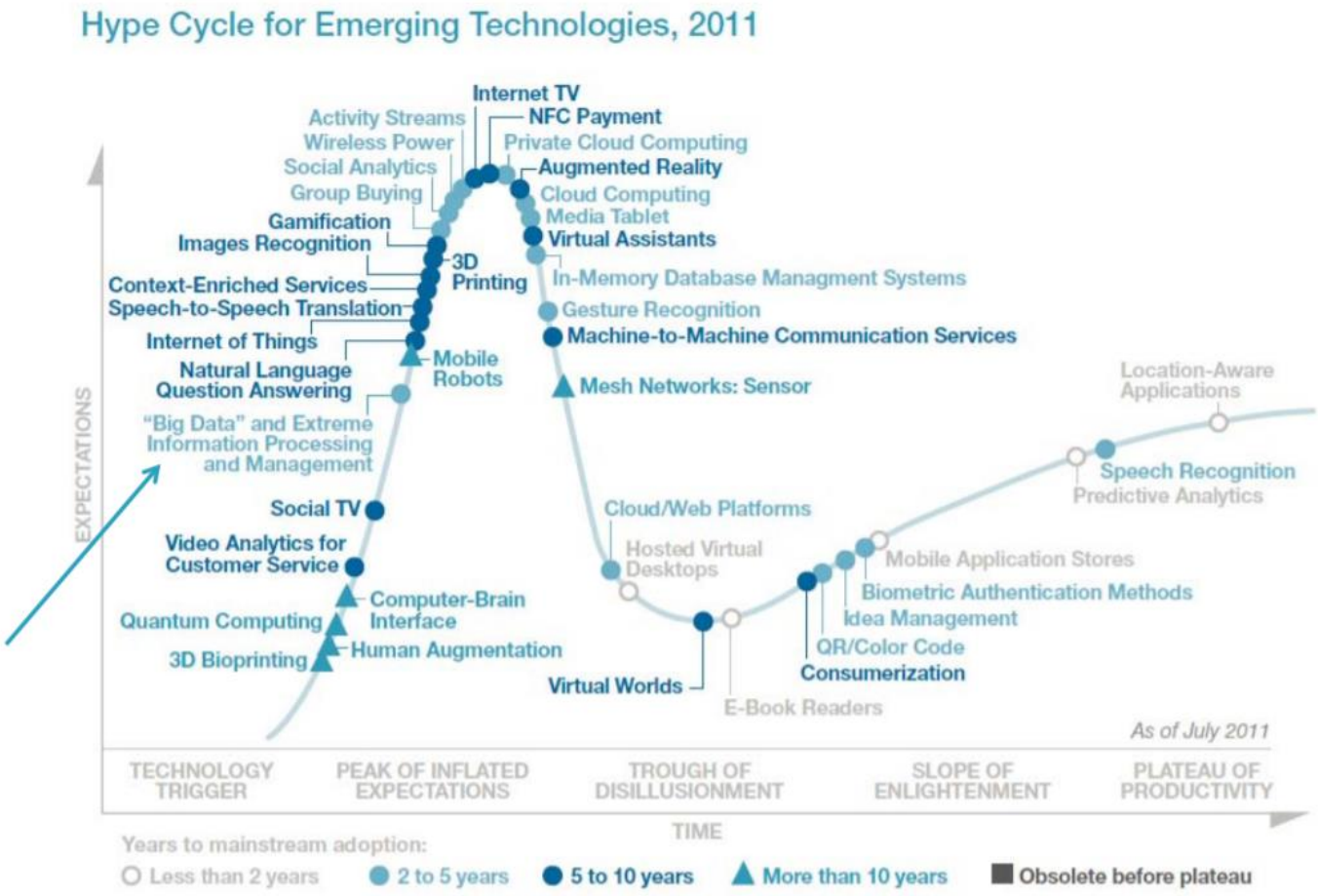

IDC estimates that by the end of 2025, only 15% of the data in the global datasphere will be tagged and only one-fifth of that will actually be analyzed.

# 大数据的三个性质

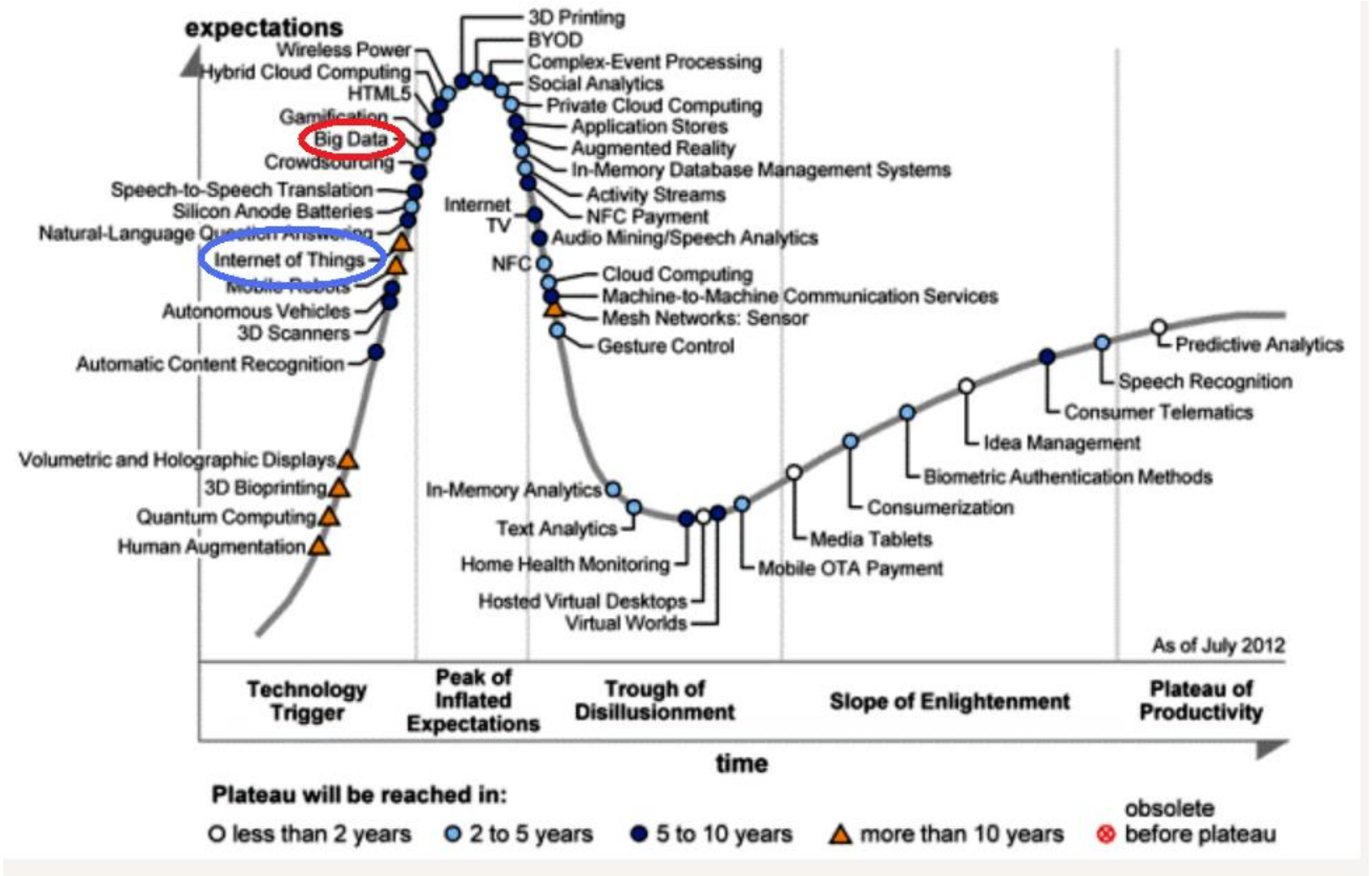
Volume, Velocity, Variety



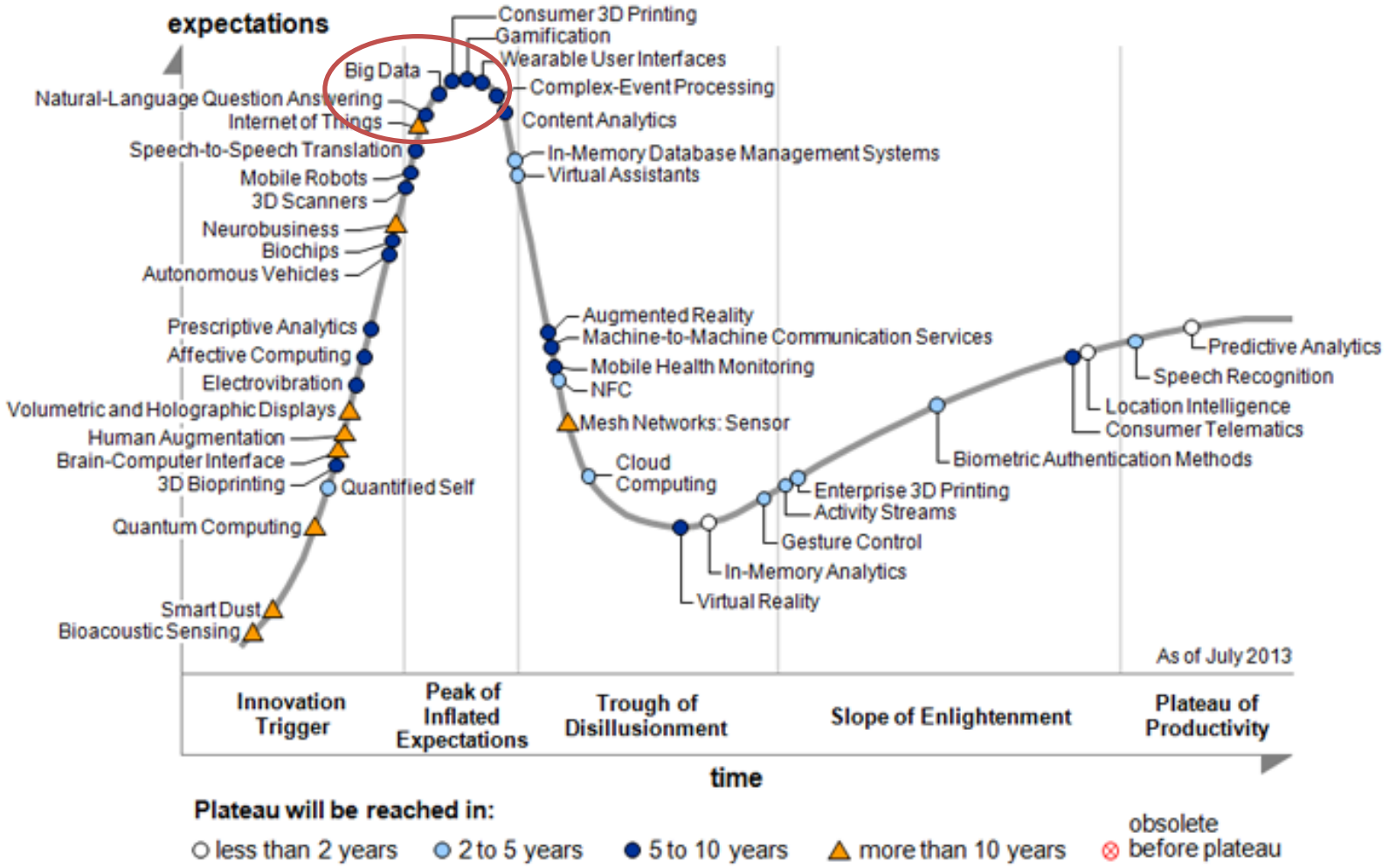
# Big Data in Garnter Hype-Cycle 2011



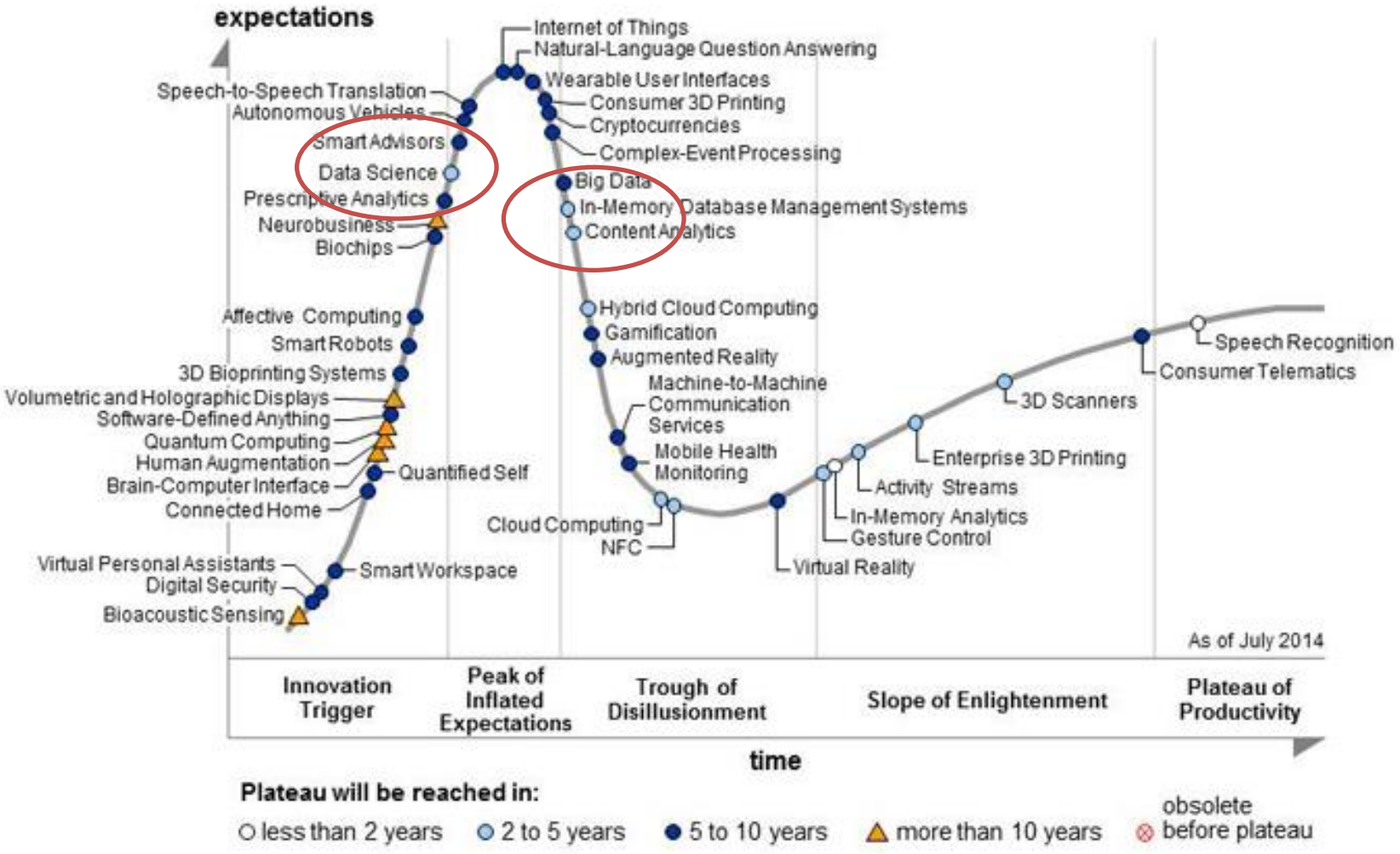
# Big Data in Garnter Hype-Cycle 2012



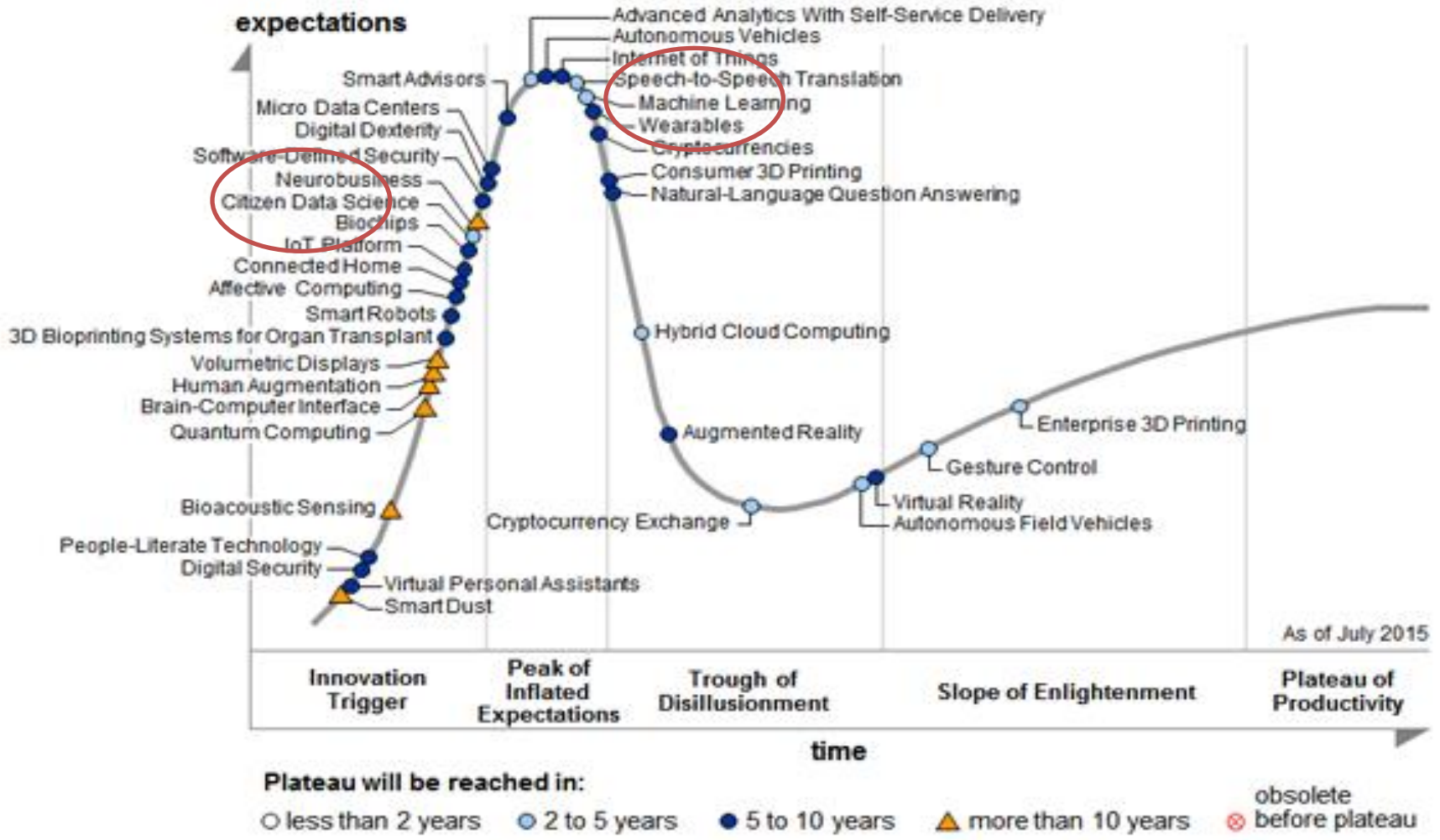
# Big Data in Garnter Hype-Cycle 2013



# Big Data in Garnter Hype-Cycle 2014



# Big Data in Garnter Hype-Cycle 2015



# 大数据分析的典型应用实例

- Social Network Modeling and Analysis
- Life Science
- Business Intelligence
- Smart City

# Social Network Modeling and Analysis

- FaceBook

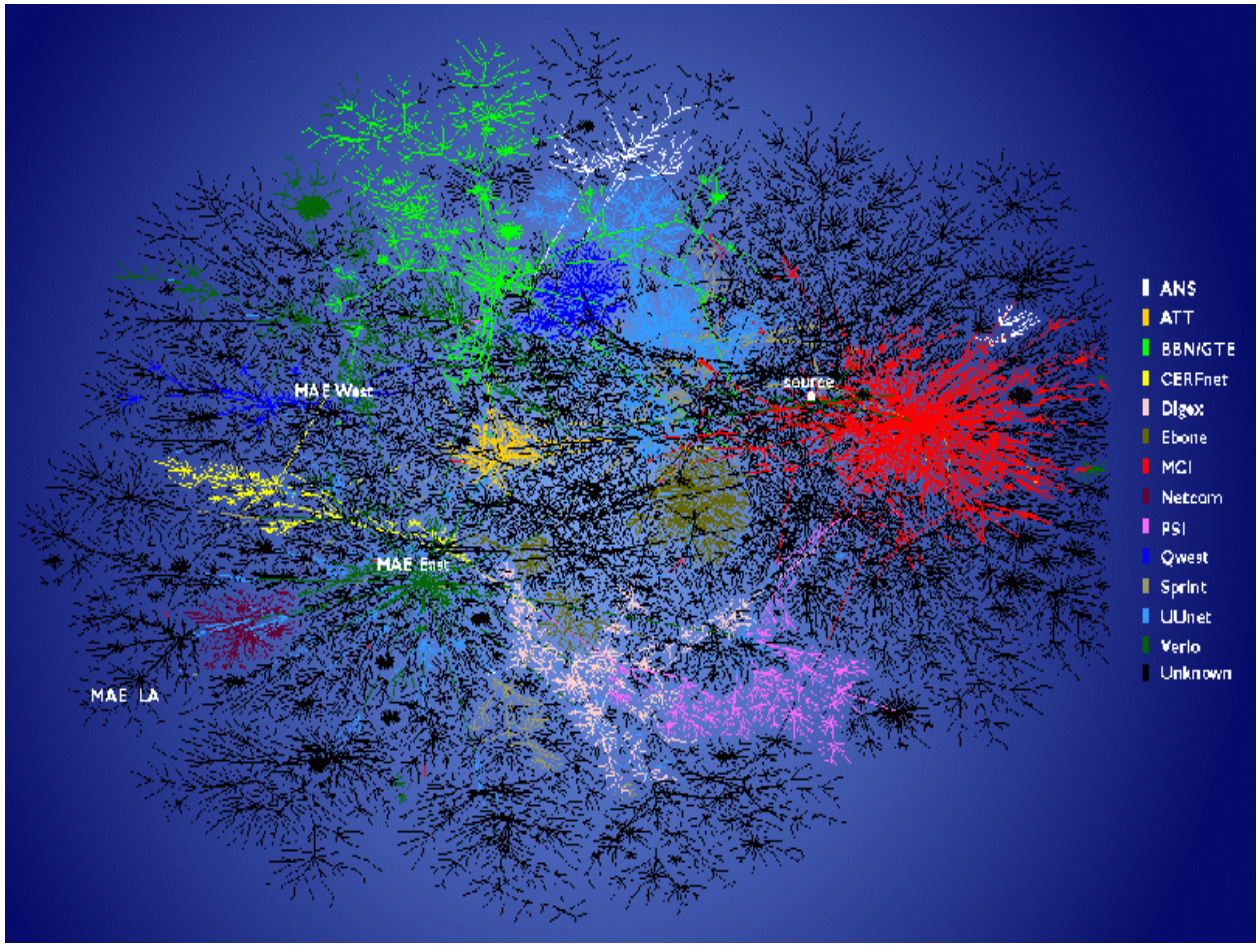
全球22亿用户

- Twitter

全球三亿用户

- 微信weibo

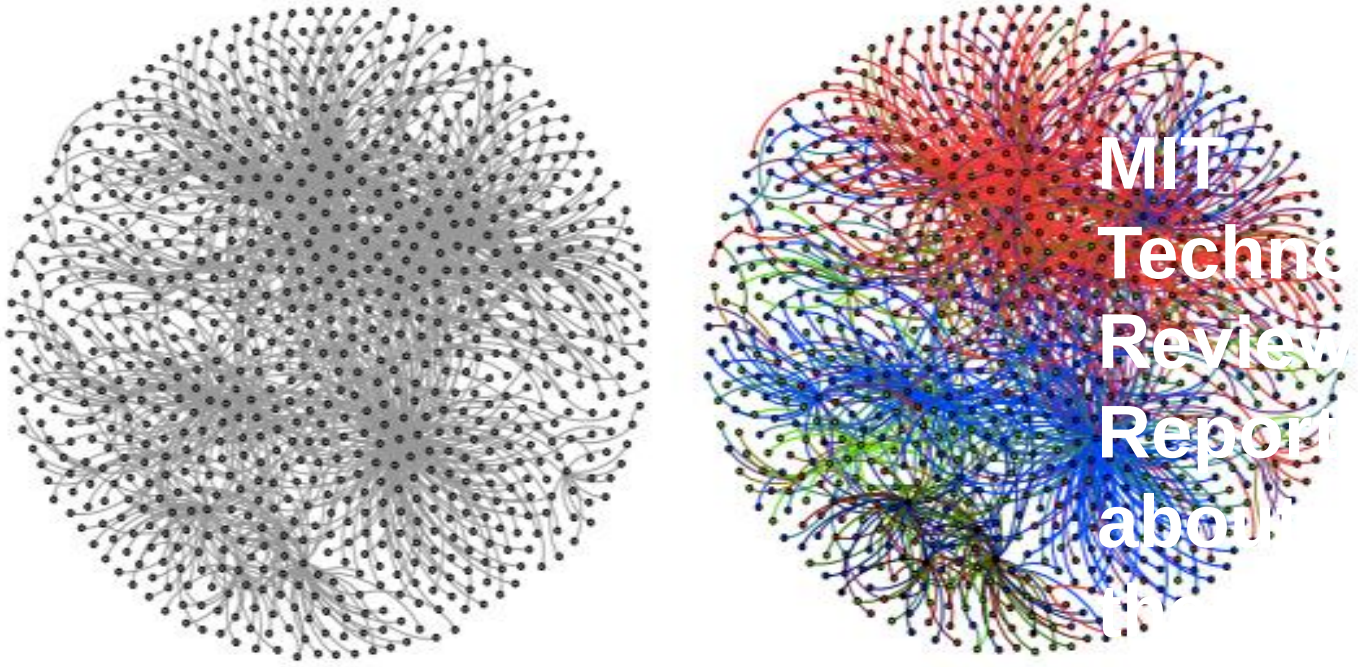
已突破10亿



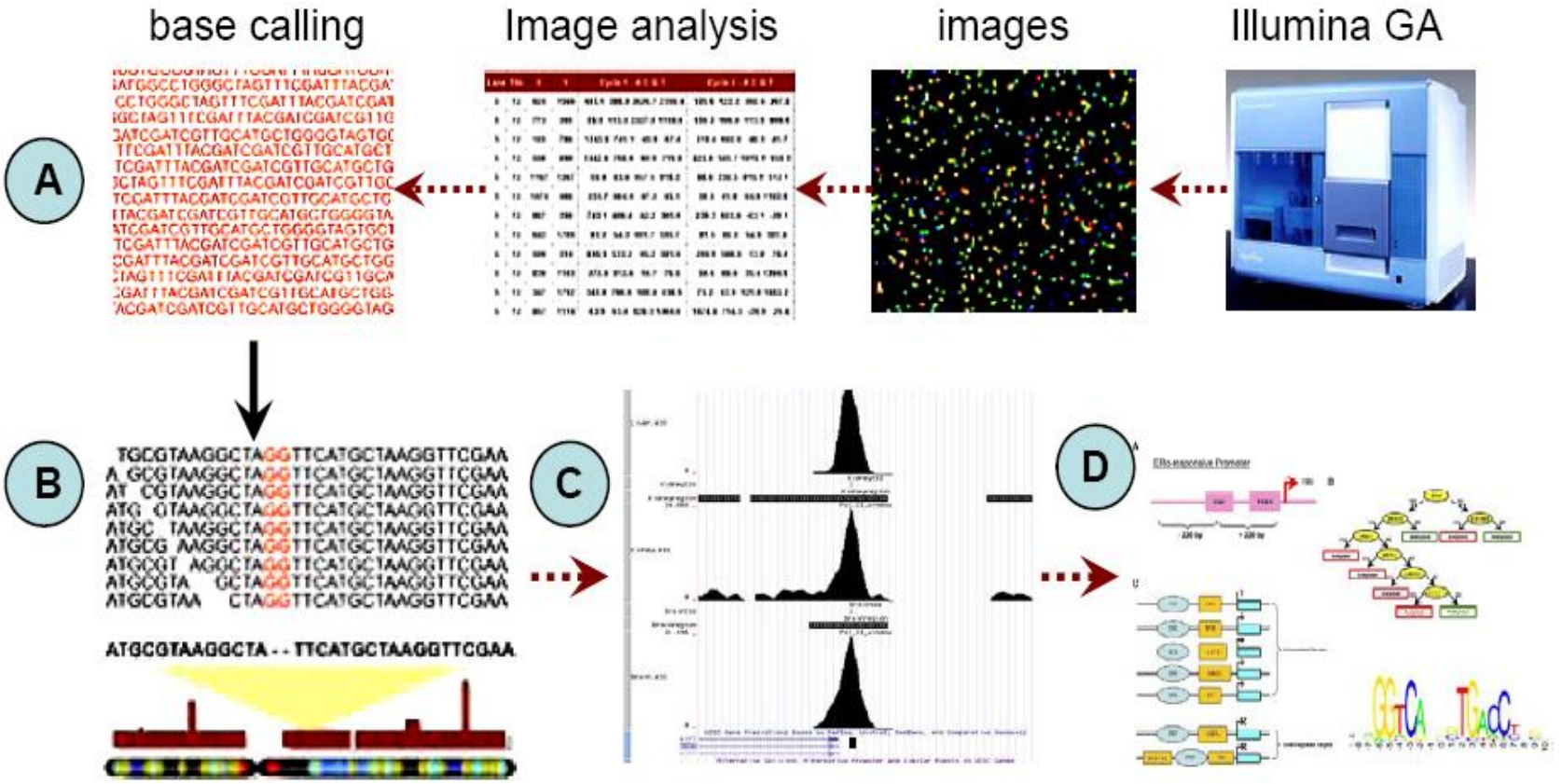
Emerging Technology From the arXiv
September 16, 2013

# Most Influential Emotions on Social Networks Revealed

Anger spreads faster and more broadly than joy, say computer scientists who have analysed sentiment on the Chinese Twitter-like service Weibo.



# Life Science
# 基因测序流程

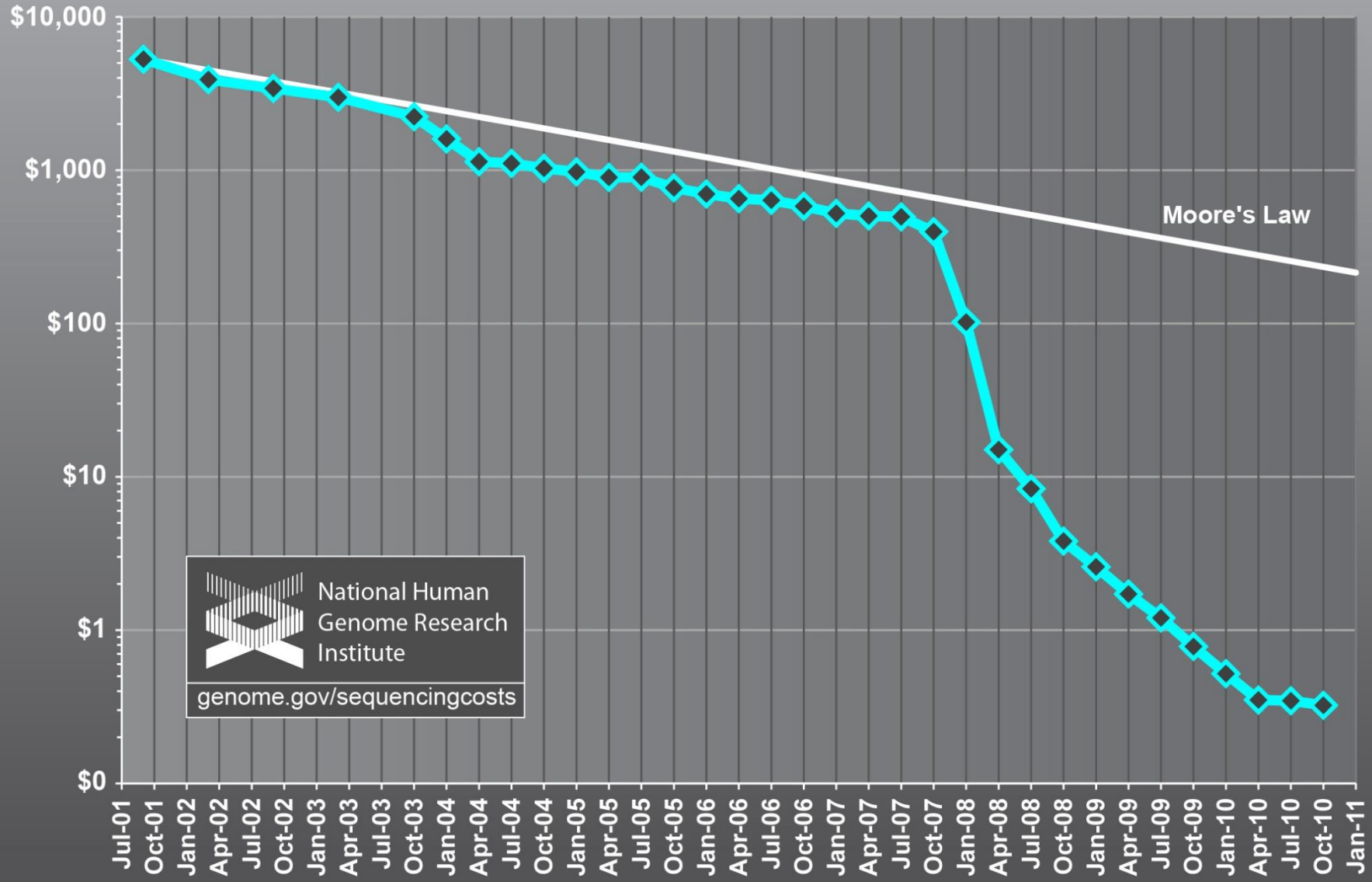



Aligning to reference genome/
Assembly into longer contigs

Statistical analysis
(e.g., finding significantly enriched genomic regions, peak calling, etc.)

Computational modeling
(e.g., finding TFBS, *cis*-regulatory modules, etc.)

# Cost per Megabase of DNA Sequence



# 基因测序数据处理规模


PHASE TWO: INTERPRETATION
I THINK I FOUND A CORNER PIECE.
3 BILLION PIECES
GENOME

# DNA Sequencing Pipeline

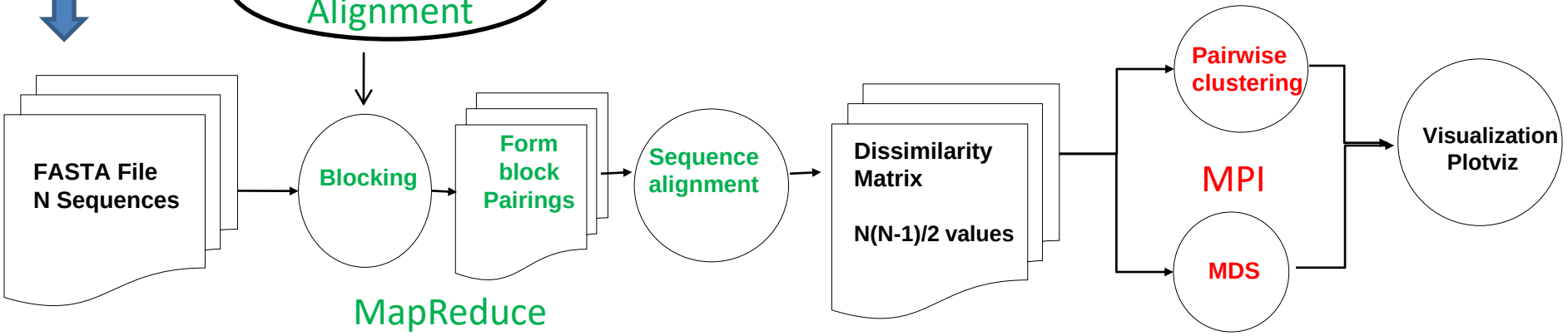
Illumina/Solexa

Roche/454 Life Sciences

Applied Biosystems/SOLiD

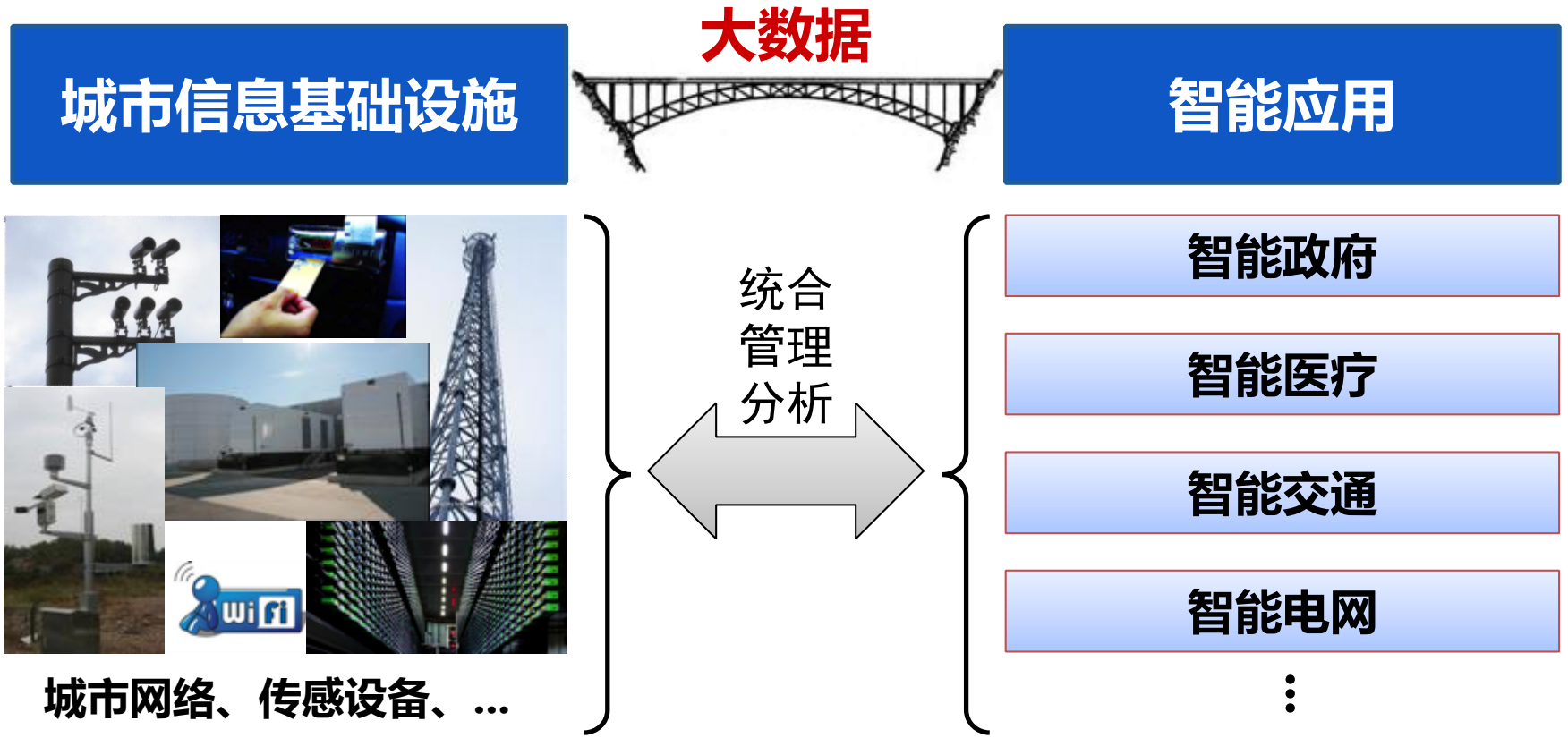
Internet

~300 million base pairs per day leading to
~3000 sequences per day per instrument
500 instruments at ~0.5M$ each



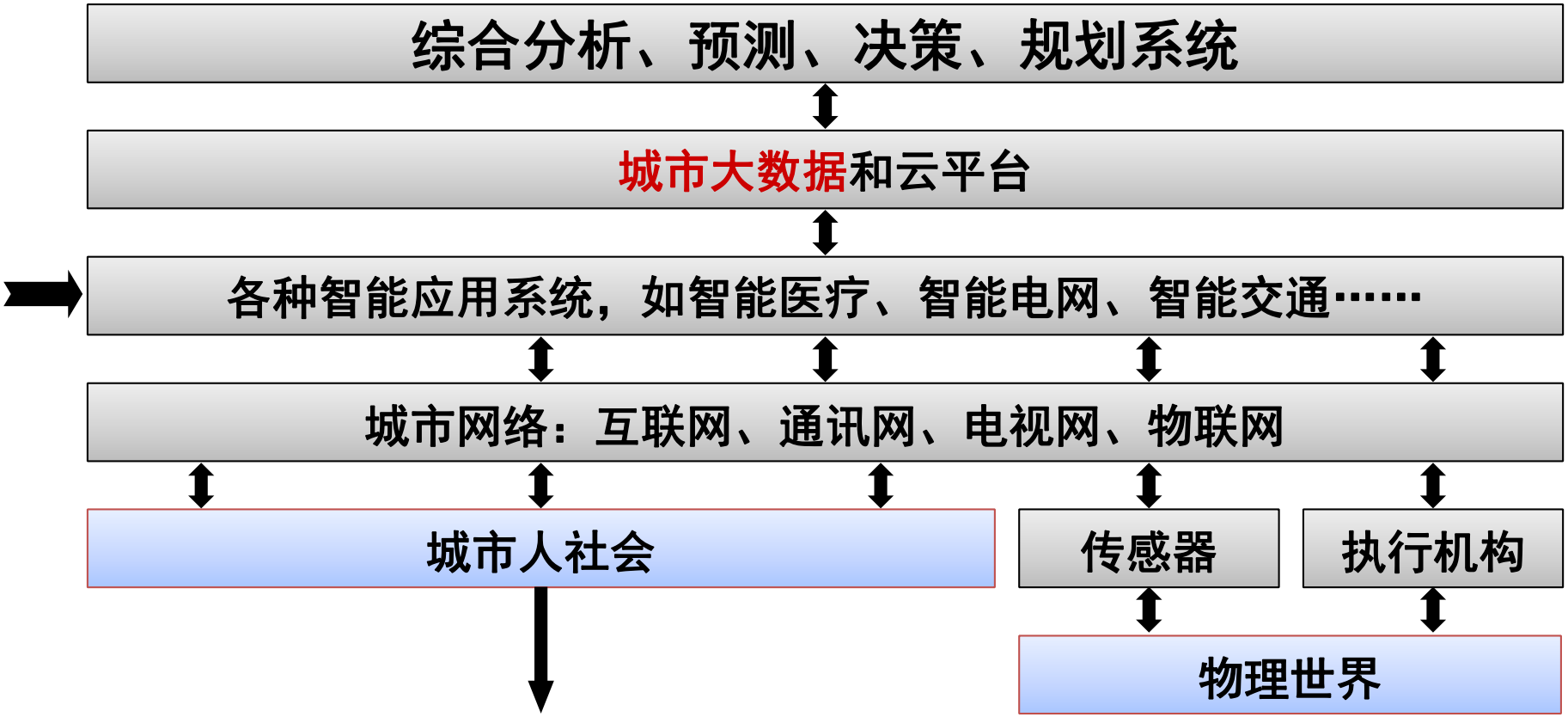
# 大数据是发展智能城市的助推剂（1）

- 大数据是城市信息基础设施和智能应用的桥梁：



# 大数据是发展智能城市的助推剂 (2)

城市大数据在iCity五个层次中的地位



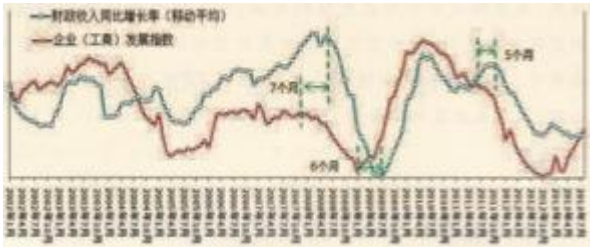



26

# 大数据是发展智能城市的助推剂 (3)

- **城市大数据有大应用：**

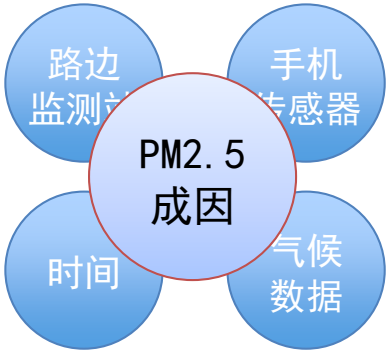
## 宏观观测



企业发展指数：统合全国市场主体的数据
企业发展指数预判宏观经济走势

## 微观分析

聚集多方面数据
分析PM2.5污染
的真正原因



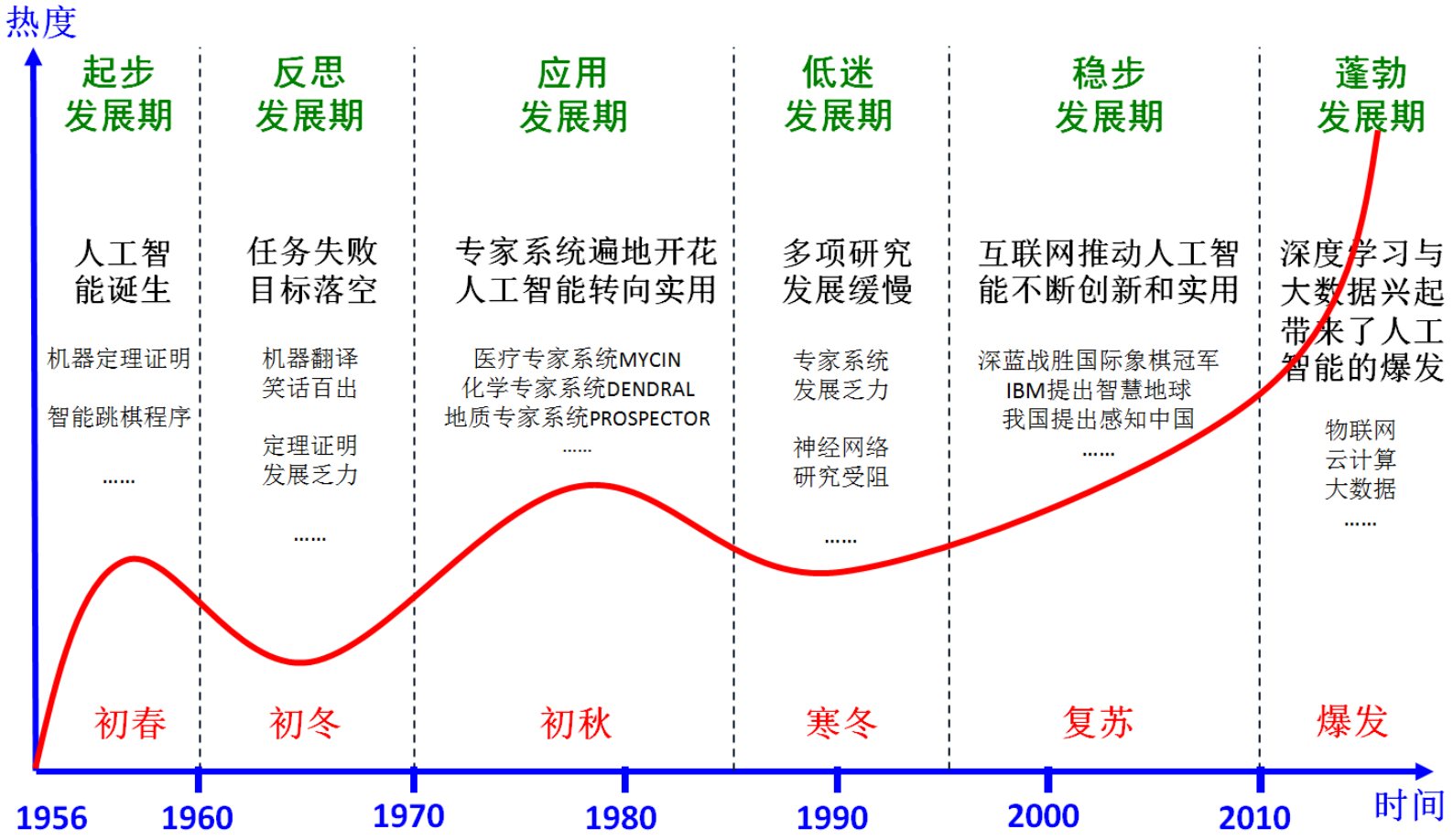

## 聚集群智

为市民配发监测空气质量的传感器。以低廉的成本大量增加数据。

## 拓展新应用

使用数据的主要内容 → 原始应用
↓
一次衍生 → 应用$^{+}$
↓
二次衍生 → 应用$^{++}$
↓
n次衍生 → 应用$^{+n}$

# 大数据推动人工智能发展




人工智能的60年发展历程

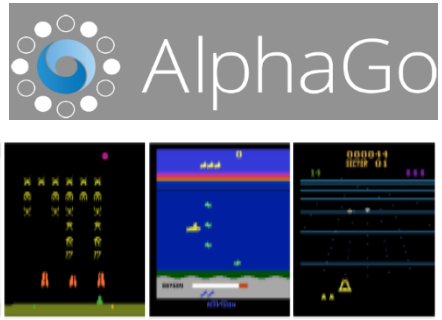
2000年以来大数据的迅速发展推动了人工智能的复苏

# 大数据推动人工智能发展

“Perhaps the most important news of our day is that datasets — not algorithms — might be the key limiting factor to development of human-level artificial intelligence”


1997年 IBM 深蓝


2011年 IBM Jeopardy

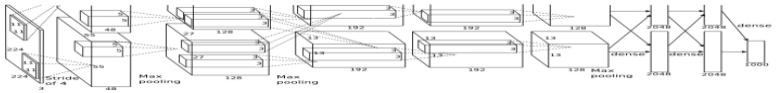



| 算法：negascout planning（1983年） | 算法：Mixture-of-Experts（1991年） | 算法：Q-Learning Algorithm（1992年） |
|---|---|---|
| **数据集：700K象棋大师的棋谱（1991年）** | **数据集：8百万Wiki百科词条（2010年）** | **数据集：50个Atari游戏数据（2013年）** |

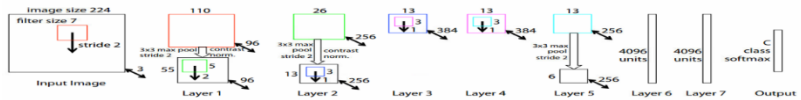
# 大数据+深度学习引发人工智能创新

- 遵循“端到端学习”机制，深度学习在视觉目标识别、机器翻译、语音识别等领域取得了显著进展。

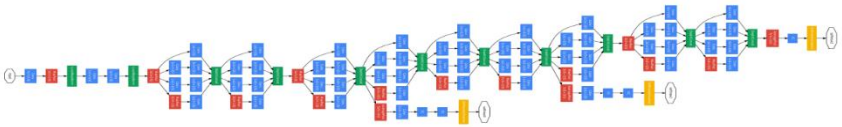
**2012** **Alex Net (8层，错误率：16.4%)**



**2013** **Zeiler Net (8层，错误率：11.2%)**



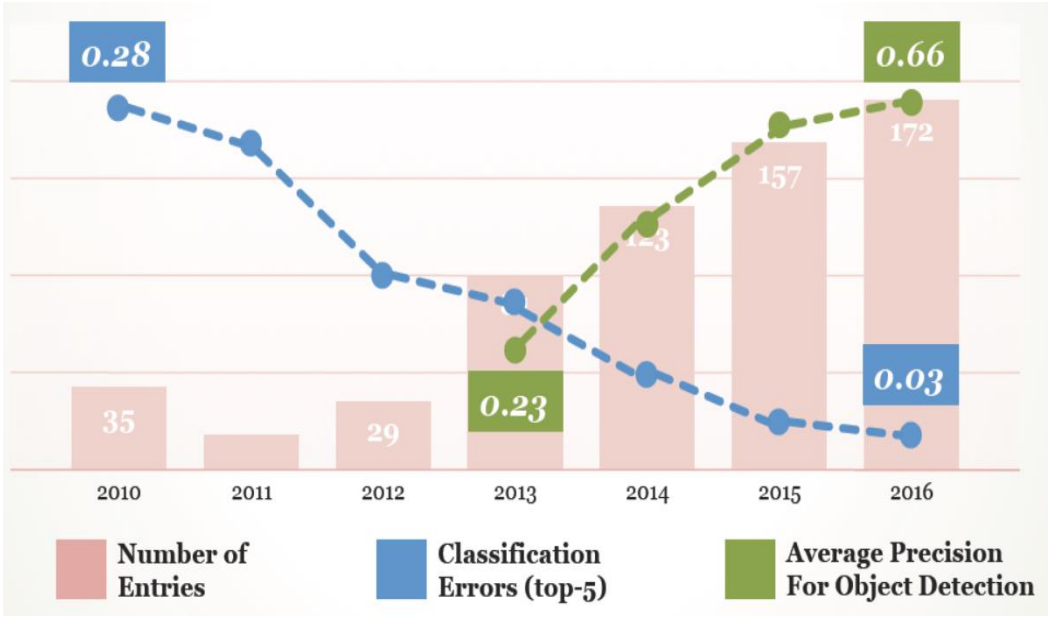

**2014** **GoogLeNet (22层，错误率：6.66%)**



**2015** **ResNet (152层，错误率：3.57%)**

# 深度学习+群体智能：ImageNet 推动计算机视觉发展

在8年的时间里，ImageNet每年举办物体识别的挑战赛ILSVRC，推动了视觉和机器学习研究的进展，特别是深度学习方法在视觉领域的应用

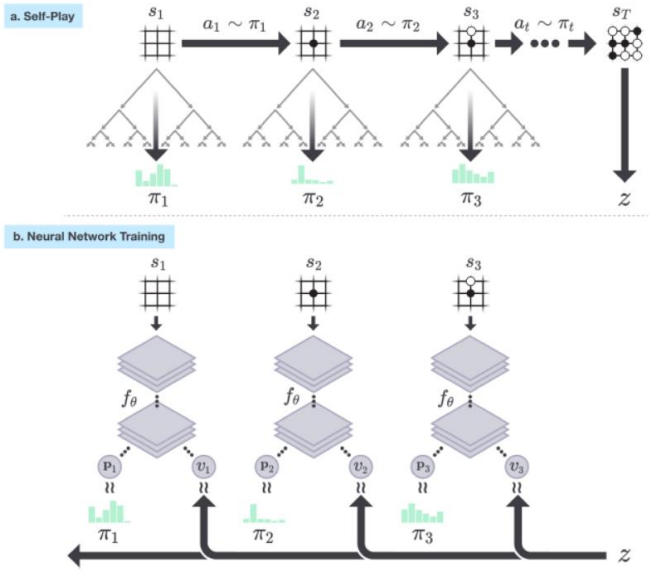



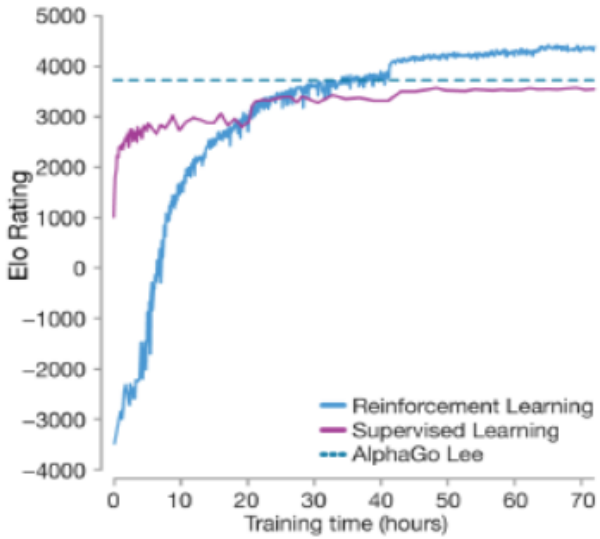

1.5千万个标注的图片，覆盖上千种物体类别

# AlphaZero： Learning from scratch

**增强学习的最新实例：**

- DeepMind公司推出的AlphaZero无需记住任何人类历史棋谱
- 只使用深度强化学习，通过左右自我对弈，搜索优化策略，实现自我训练
- 从零开始三天的训练就超过人类围棋智能水平

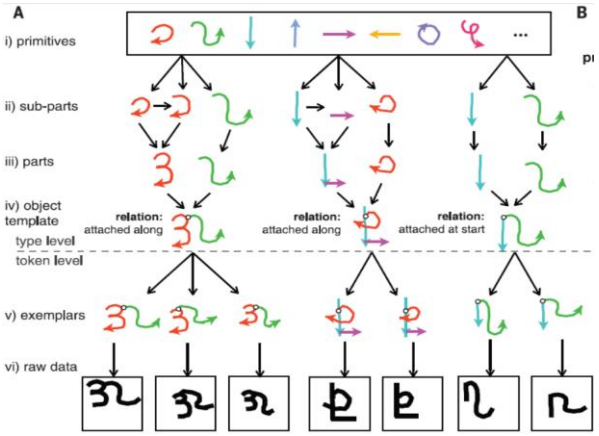


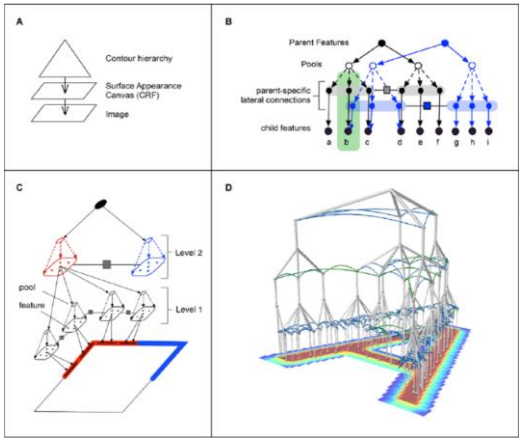




David Silver*, Julian Schrittwieser*, Karen Simonyan*, Ioannis Antonoglou， et al., Mastering the Game of Go without Human Knowledge, Nature 550.7676 (2017): 354-359

# 小样本、先验知识的AI学习方法

- 《科学》杂志发表了“仅从一个例子就形成概念”方法：概率规划归纳（probabilistic program induction）人类具有从极少量数据中学习丰富概念的能力：归类、派生、解析和创造
  - 从小样本出发，利用层次化先验(hierarchical prior)
  - 以自动归纳、抽象训练数据里的高层次信息
  - **说明了利用人类的先验知识的重要性**




B. Lake, R. Salakhutdinov, J. Tenenbaum, Human-level concept learning through probabilistic program induction, Science, 2015, 350(6266):1332-1338

D George, W Lehrach, K Kansky, et al, A generative vision model that trains with high data efficiency and breaks text-based CAPTCHAs, Science, 2017

# 从数据中产生智能的主要模式

| 模式 | 优势 | 不足 | 人“教”机器 |
| --- | --- | --- | --- |
| 基于知识规则和逻辑推理 | 与人类的逻辑推理相似，易于理解<br>关键理论与技术：<br>数理逻辑、知识库 | 难以构造完备的知识规则库 | • 人类设计初始的知识框架<br>• 不断提供新的知识 |
| 基于大数据和统计学习 | 分析大数据的规律，形成统计学习模型<br>关键理论与技术：<br>SVM、贝叶斯图、深度学习等 | 依赖与数据的质量和丰富性，复杂的模型缺乏解释性<br>（深度神经网络） | • 人类设计初始的模型框架<br>• 不断提供新的标注数据<br>• 并对模型进行解释 |
| 在目标环境中利用问题引导 | 从经验中进行能力的持续提升和演化<br>关键理论与技术：<br>增强学习、迁移学习 | 需要更好的优化策略实现对模型空间的搜索 | • 人类设计初始的元学习机制<br>• 提供启发式策略，缩小搜索空间 |

# 大数据智能：从数据到知识与决策

- **大数据智能是以人工智能手段对大数据进行深入分析，探析其隐含模式和规律的智能形态，实现从大数据到知识、进而决策的理论方法和支撑技术**
- 从数据到知识、从知识到决策：From ***Data*** to ***Knowledge*** to ***Decision***

# 大数据与大计算
# Extreme computing and data

- 计算机的性能指标
- $10^6$Flops ~ MegaFlops

  三星I9100手机的Exynos 4210处理器：47.072MFLOPS
- $10^9$Flops ~ GigaFlops

  Intel Core 2 Duo E8400 24GFLOPS
- $10^{12}$Flops ~ TeraFlops

  Blue Gene/L: 135.5 TFLOPS
- $10^{15}$Flops ~ PetaFlops

**天河一号:2.566 PFLOPS**

- $10^{18}$Flops ~ ExaFlops

**FLOPS**（即“每秒浮点运算次数”，“每秒峰值速度”）
**Linpack 测试**

# March to ExaFlop/s （百亿亿次）

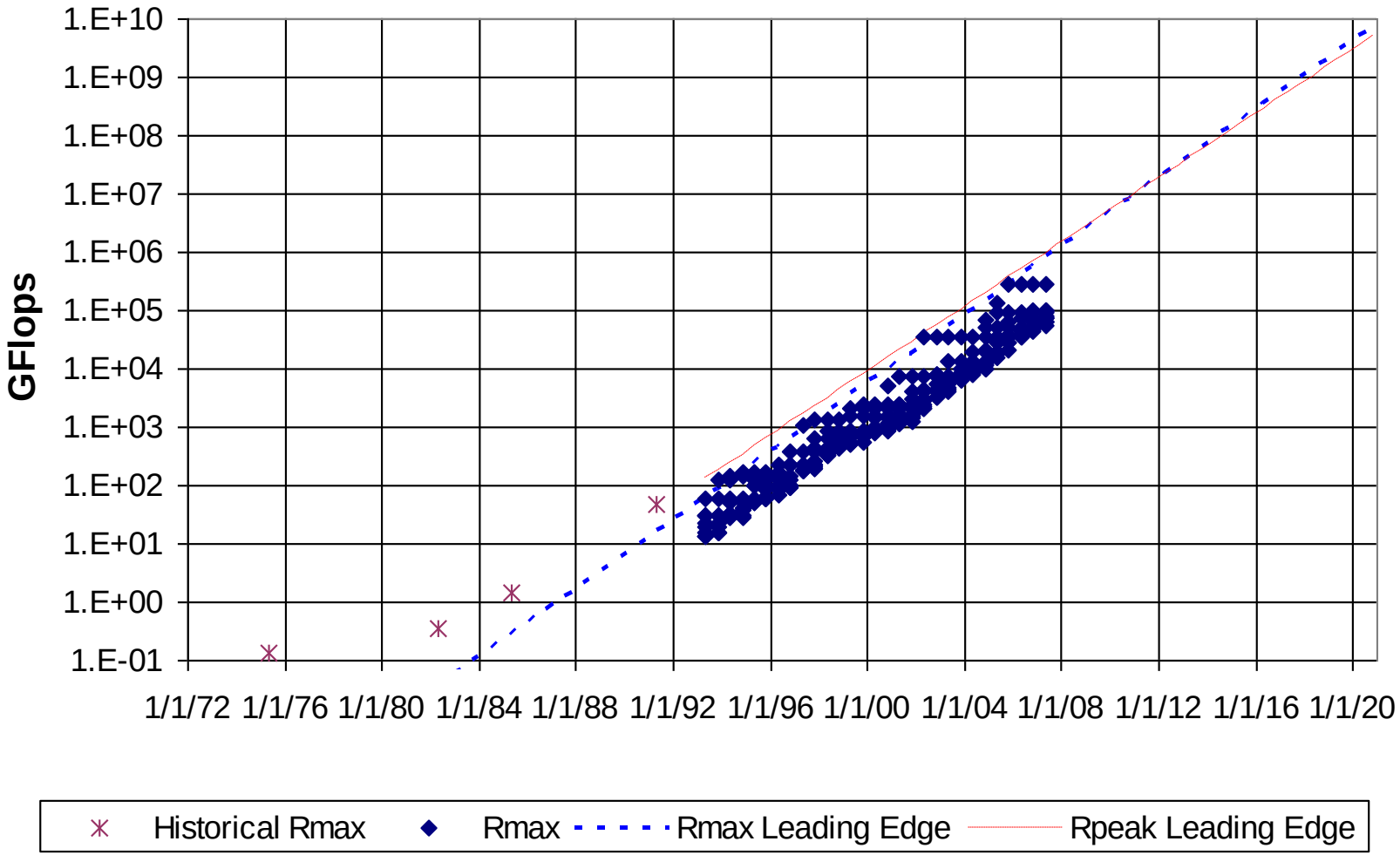



**Figure courtesy of Peter Kogge**

# 数据密集型计算的概念

- 面向大量数据的并行计算技术
- 数据并行 data parallel
- 并行计算 Parallel Computing
  - 一个问题分解为多个计算任务
  - 多个计算任务同时进行
  - 并行计算可以在不同层面上得到实现
    指令级、数据级和任务级

# 高性能计算
# High Performance Computing

- 一是指计算指标达到一定指标的高性能计算机
- 二是运用高性能的计算机解决大规模的科学问题
- 高性能计算往往要用到并行计算的方法

# Data Center
# 数据中心


Microsoft's Chicago Data Center
October 2008
Photo By McShane-Fleming Studios, Chicago

软正在建设位于芝加哥的数据中心。数据中心建设的第一阶段已经完成，在这个数据中心的第一层放置了多达56个集装箱，每个集装箱内放置了1800到2500台服务器

MicroSoft 在芝加哥的数据中心
56个集装箱，每个集装箱放置了1800-2500的服务器

# Data Centers Clouds & Economies of Scale I

Range in size from “edge” facilities to megascale.

Economies of scale

Approximate costs for a small size center (1K servers) and a larger, 50K server center.

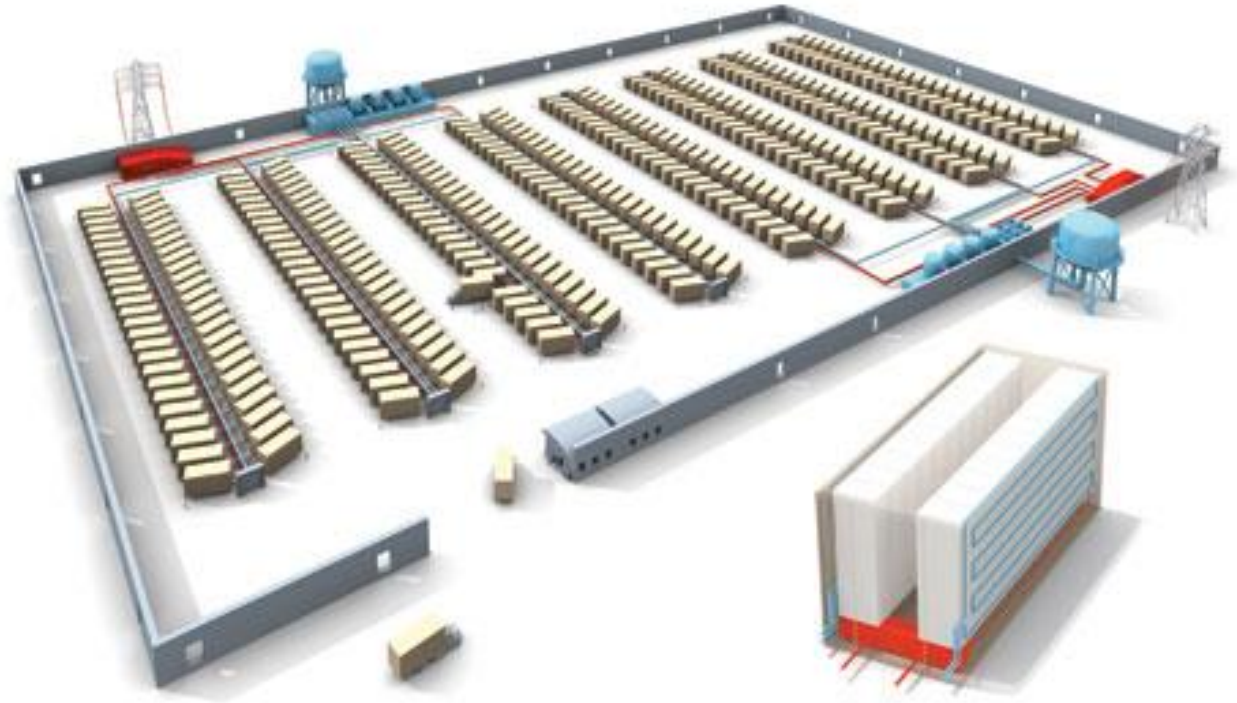
2 Google warehouses of computers on the banks of the Columbia River, in The Dalles, Oregon

Such centers use 20MW-200MW (Future) each with 150 watts per CPU

Save money from large size, positioning with cheap power and access with Internet

# Data Centers, Clouds & Economies of Scale II

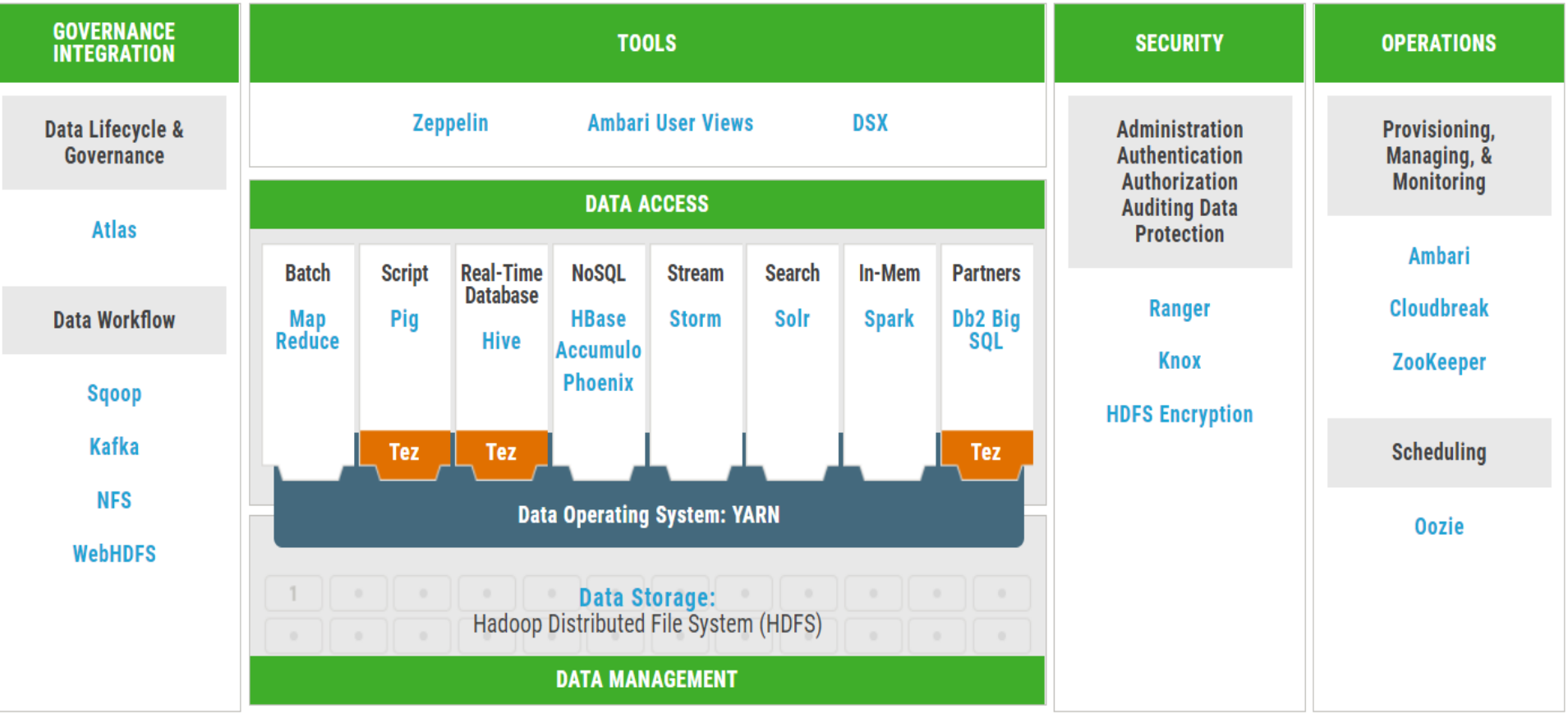
- Builds giant data centers with 100,000's of computers; ~ 200-1000 to a shipping container with Internet access
- "Microsoft will cram between 150 and 220 shipping containers filled with data center gear into a new 500,000 square foot Chicago facility. This move marks the most significant, public use of the shipping container systems popularized by the likes of Sun Microsystems and Rackable Systems to date."



# Largest Data Center in the World

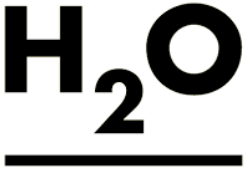
- Google and Microsoft: > 1 Millions servers around the world
- Amazon: > 500,000 servers
- A small town in the remote north of the Arctic Circle is set to be home to the world's largest data center. (1000 megawatts of power)

# Hadoop大数据开源软件生态



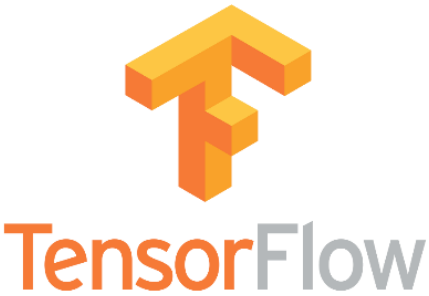

数据处理: Map-Reduce, Spark
数据存储： HDFS
运行管理： YARN

# Popular Deep Learning Frameworks







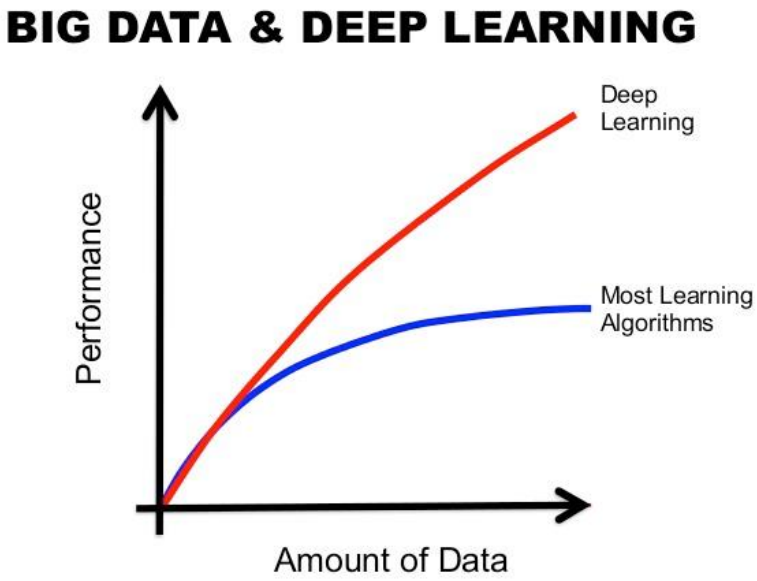



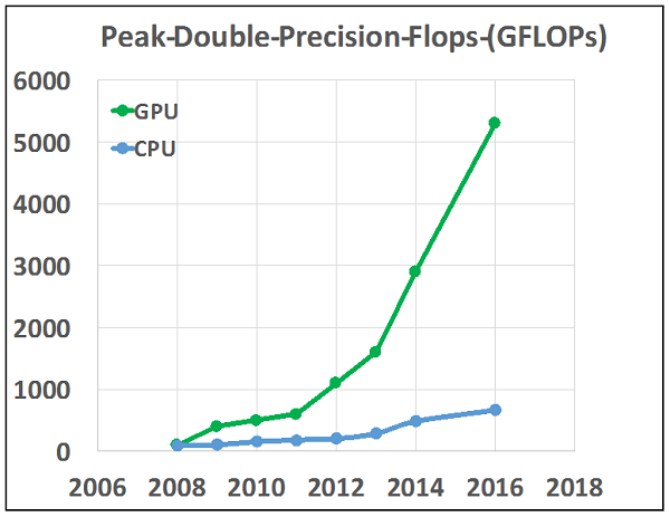


# Requirements for Deep Learning

Deep learning uses general learning algorithms

- The algorithms need to build the layers of an artificial neural network
  - Training data
- Processing this training data requires lots of computation
  - Convolutional NN -> Matrix multiplications

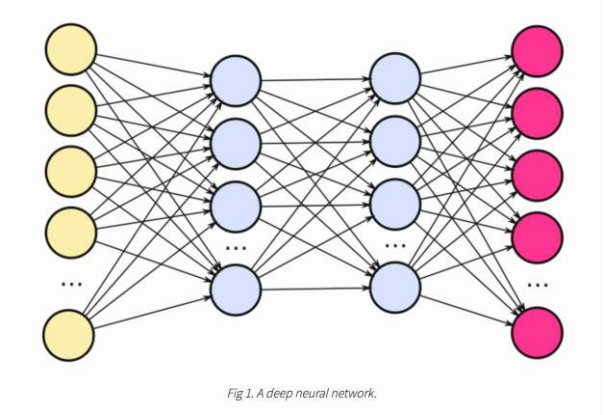



# 工业界人工智能成功的三大法宝

深度神经网络与大数据的结合成为当前人工智能技术的主流路径

基于云计算的“研究—工程—产品—用户”闭环优化加速了迭代优化进程

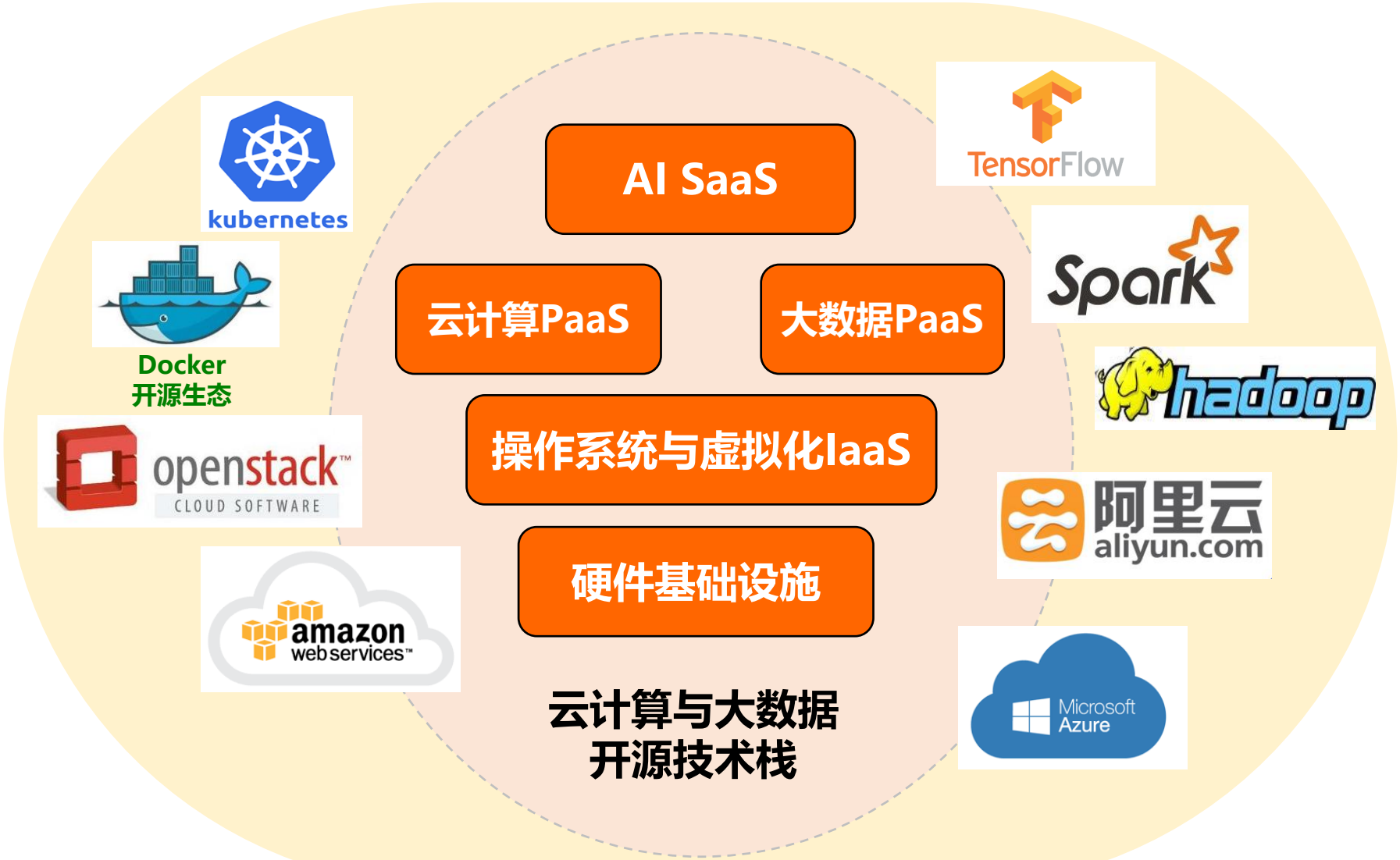



深度神经网络

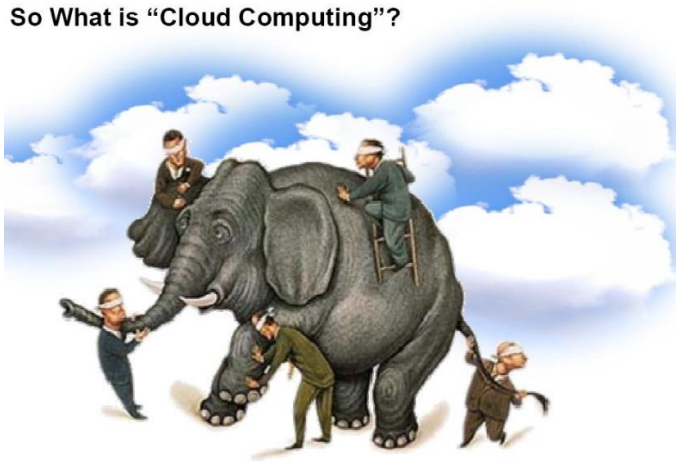

大数据


云计算

# 人工智能所需的云平台


AI SaaS
云计算PaaS
大数据PaaS
操作系统与虚拟化IaaS
硬件基础设施
云计算与大数据
开源技术栈
kubernetes
Docker
开源生态
openstack
CLOUD SOFTWARE
amazon
web services
TensorFlow
Spark
hadoop
阿里云
aliyun.com
Microsoft
Azure

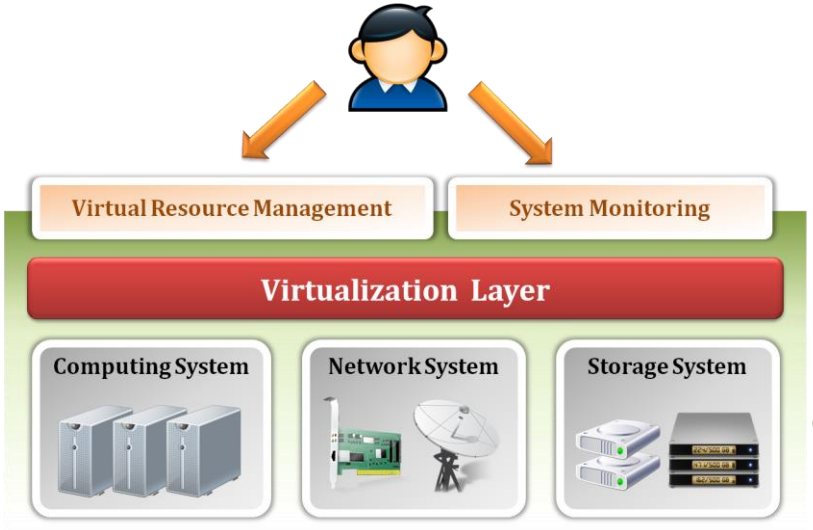
# 云计算的概念

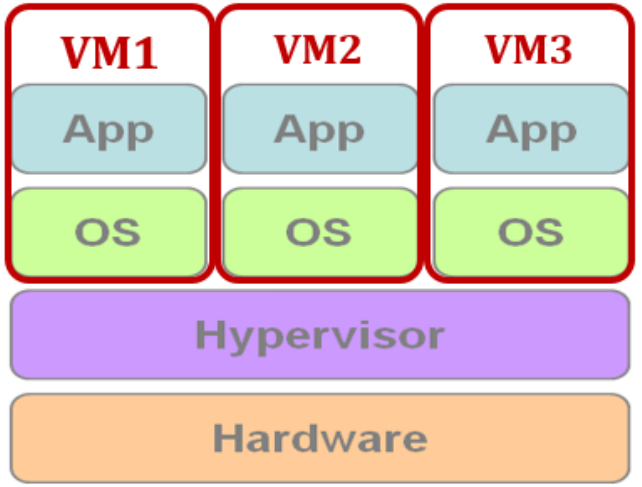



- 云计算：一个古老的想法
  - 把计算作为公用基础设施
- 云计算服务：一种公共基础设施服务
  - 从硬件的角度，云用户
    - 可以按需获取看似无限的计算资源，而不需要复杂的资源规划
    - 消除云用户的事先投入，按需求增加硬件资源
    - 以很短的时间为单位付费使用资源

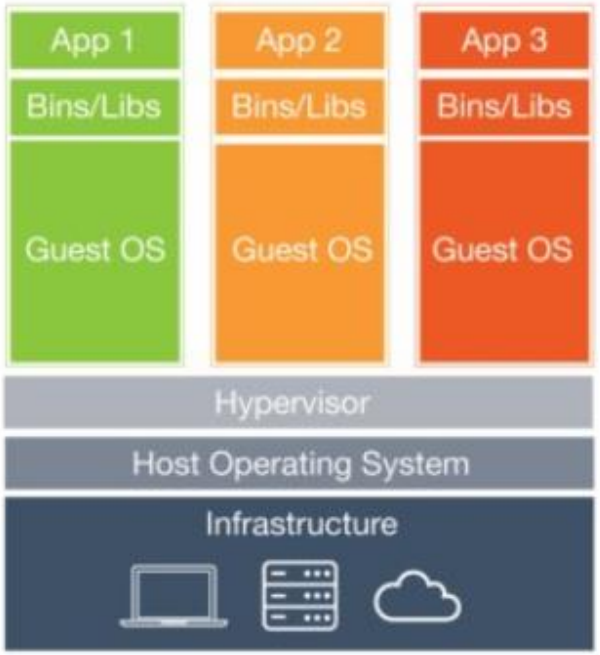
# 基础设施即服务IaaS

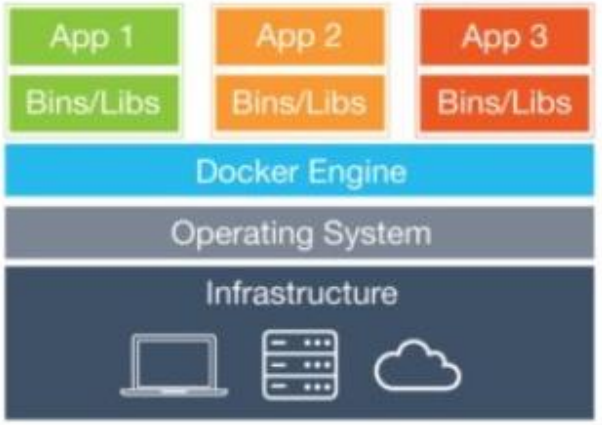
- IaaS把平台的资源：计算、存储、通信等都通过虚拟化技术抽象为用户可以随时使用的服务
- 虚拟化技术
  - 虚拟机Virtual Machine (VM)
  - Virtual Machine Monitor (VMM) Hypervisor
  - 虚拟化类别：Bare metal vs Hosted
  - 虚拟化方式：Full-Virtualization vs Para-Virtualization



# 平台即服务PaaS

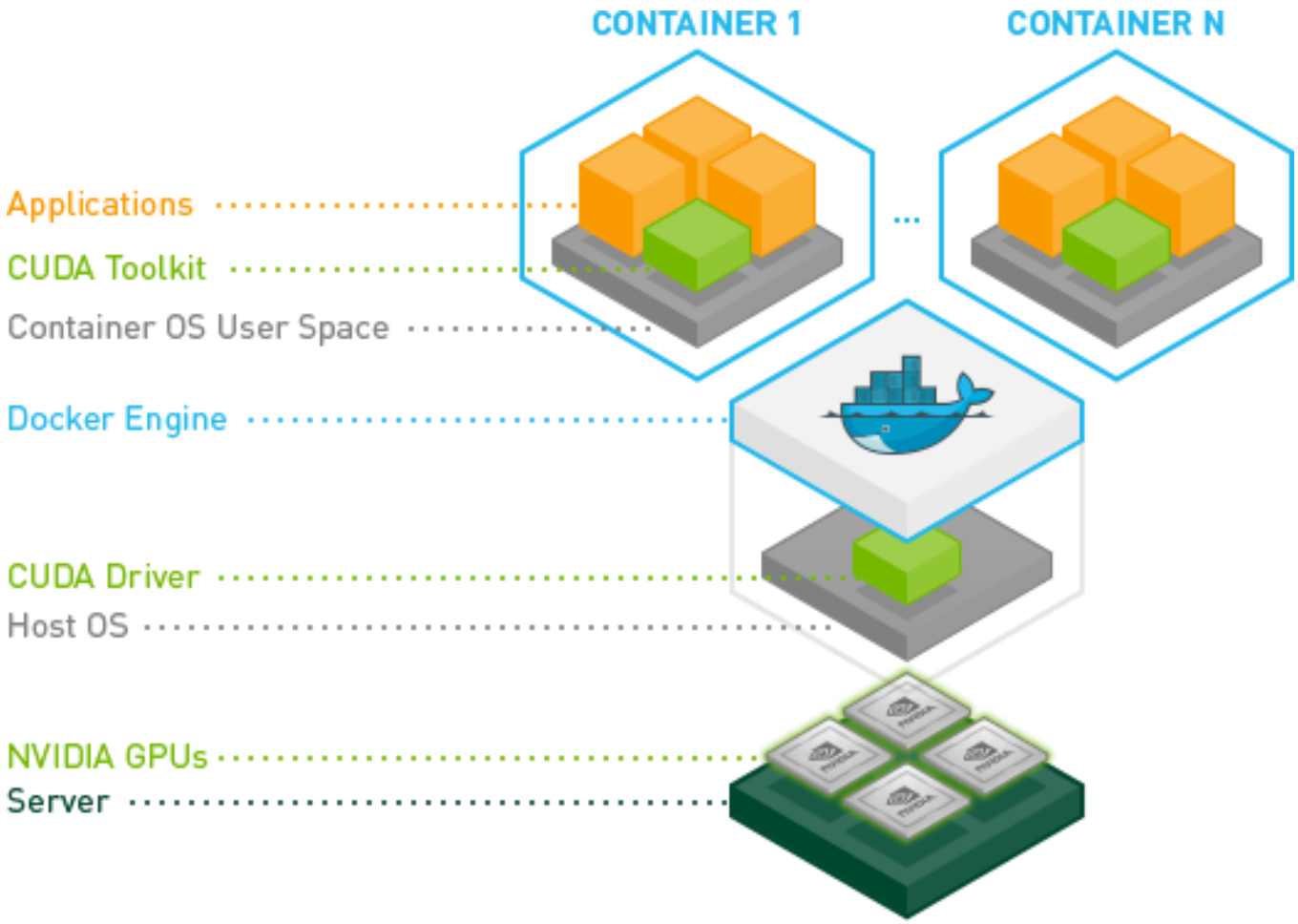



- 容器Container是一种轻量的系统级别虚拟化技术，实现在单机上运行多个分离的操作系统
- Container vs Virtual Machine (VM)：部分虚拟化，Container并没有独立的OS内核，但是有独立的文件系统和进程等

Docker 容器技术由Linux的LXC虚拟技术演化而来

# GPU Access from within a Container

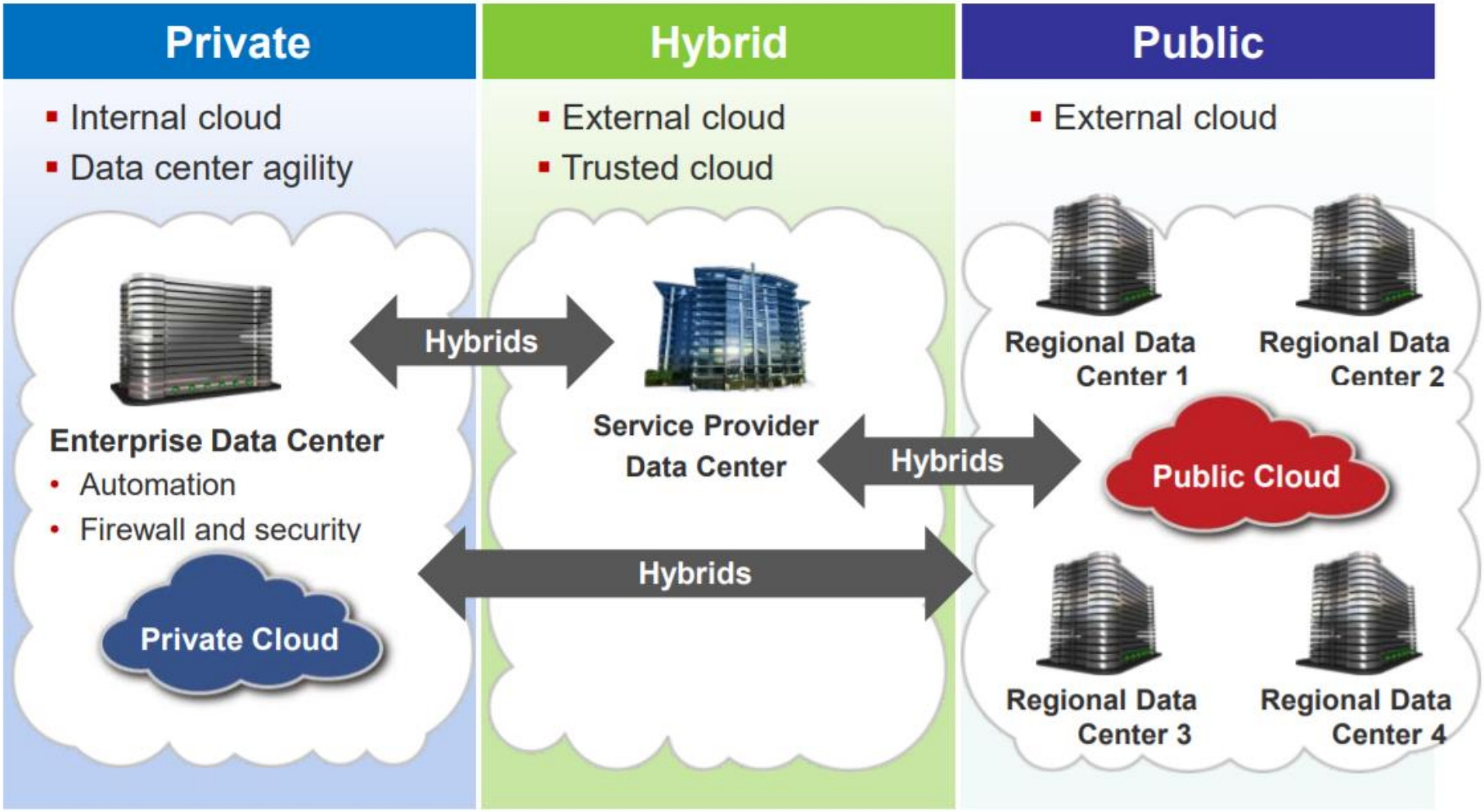


# 平台即服务PaaS

- 基于容器的PaaS中间件
  - Kubernetes: Google‘s point of view on container orchestration

  Start, stop, update, and manage a cluster of machines running containers
- 面向领域的PaaS软件
  - LAMP: Linux Apache Mysql PHP
  - 大数据处理：Hadoop，Spark
  - 深度学习：Tensorflow

# 云平台的类型


Private
Internal cloud
Data center agility
Hybrid
External cloud
Trusted cloud
Public
External cloud
Hybrids
Enterprise Data Center
Automation
Firewall and security
Private Cloud
Service Provider
Data Center
Hybrids
Hybrids
Regional Data
Center 1
Regional Data
Center 2
Public Cloud
Regional Data
Center 3
Regional Data
Center 4

# Edge Computing and Fog Computing

Fog computing is...

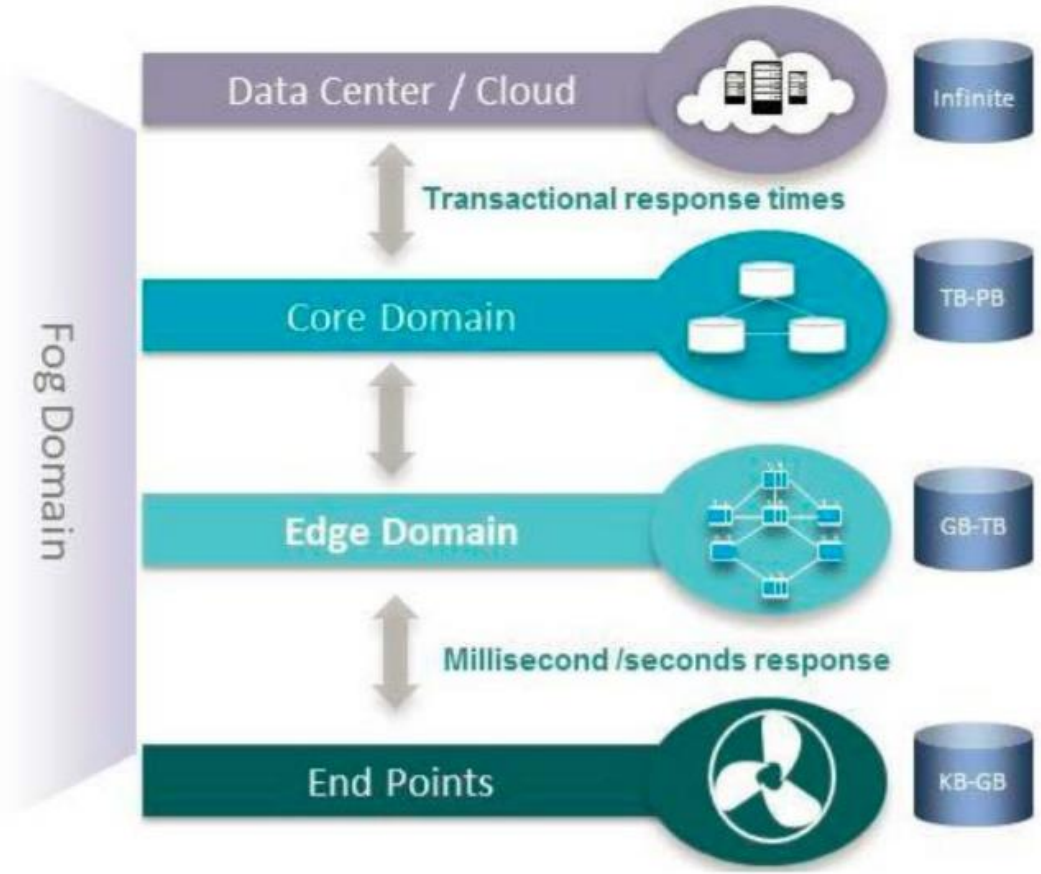
A system-level architecture to extend

*Compute*

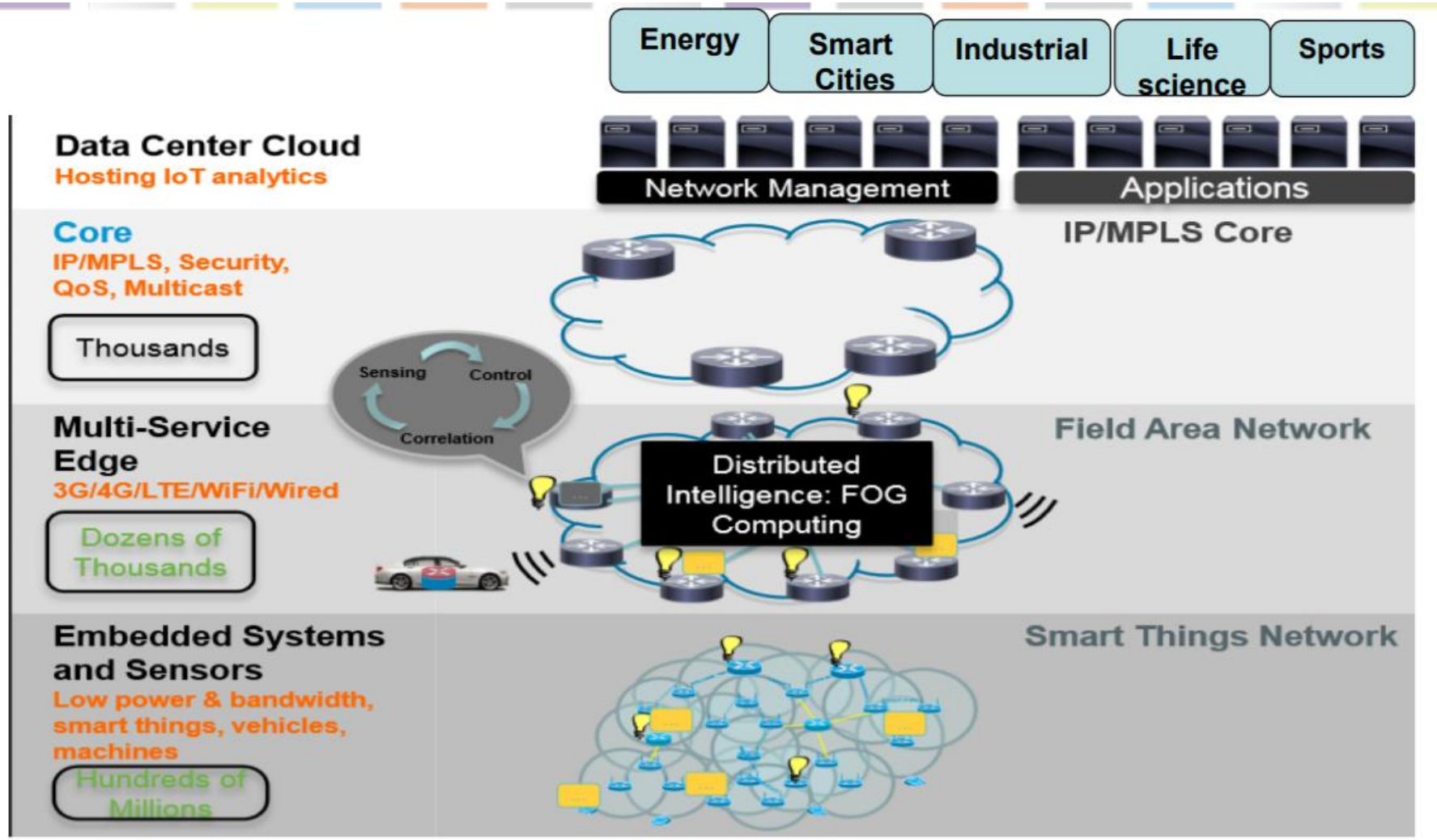
*Network*

*Storage*

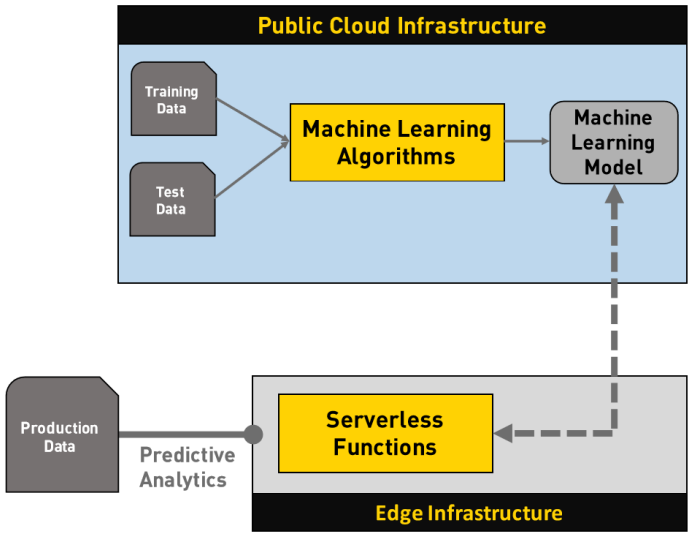
Capability of Cloud to the edge of the IoT network



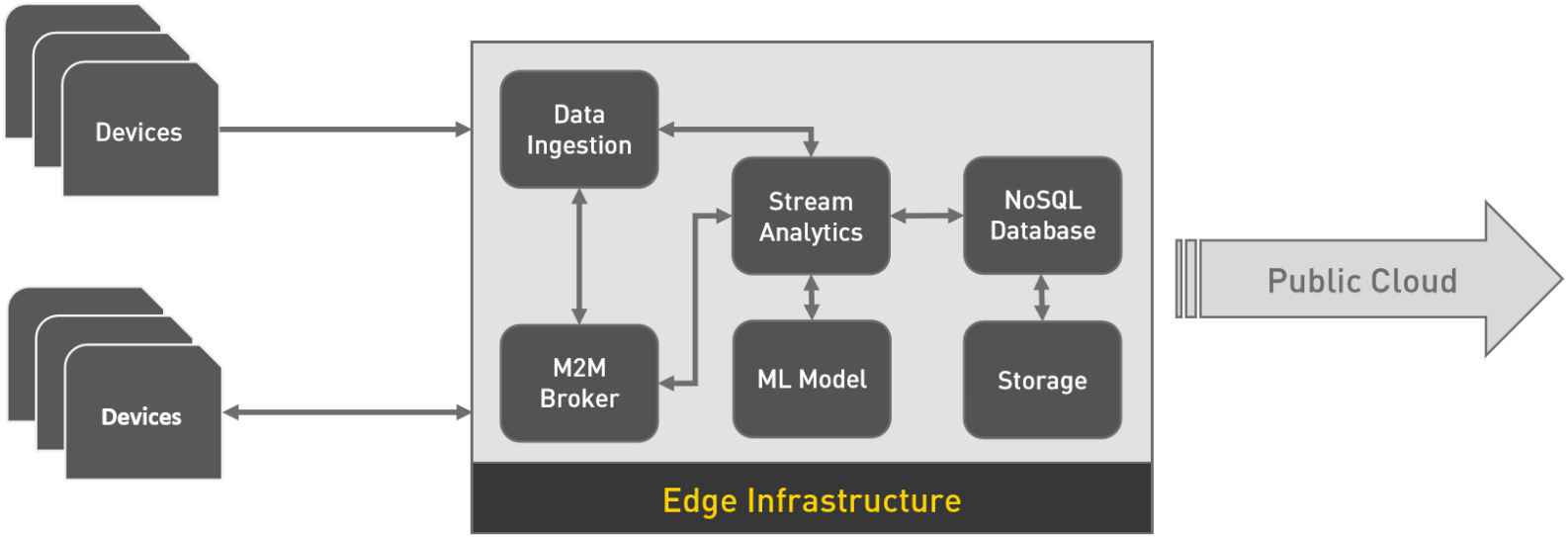
# 5G和边-雾计算架构


Energy
Smart Cities
Industrial
Life science
Sports
Data Center Cloud
Hosting IoT analytics
Network Management
Applications
Core
IP/MPLS, Security, QoS, Multicast
IP/MPLS Core
Thousands
Sensing
Control
Correlation
Multi-Service Edge
3G/4G/LTE/WiFi/Wired
Field Area Network
Distributed Intelligence: FOG Computing
Dozens of Thousands
Embedded Systems and Sensors
Low power & bandwidth, smart things, vehicles, machines
Smart Things Network
Hundreds of Millions

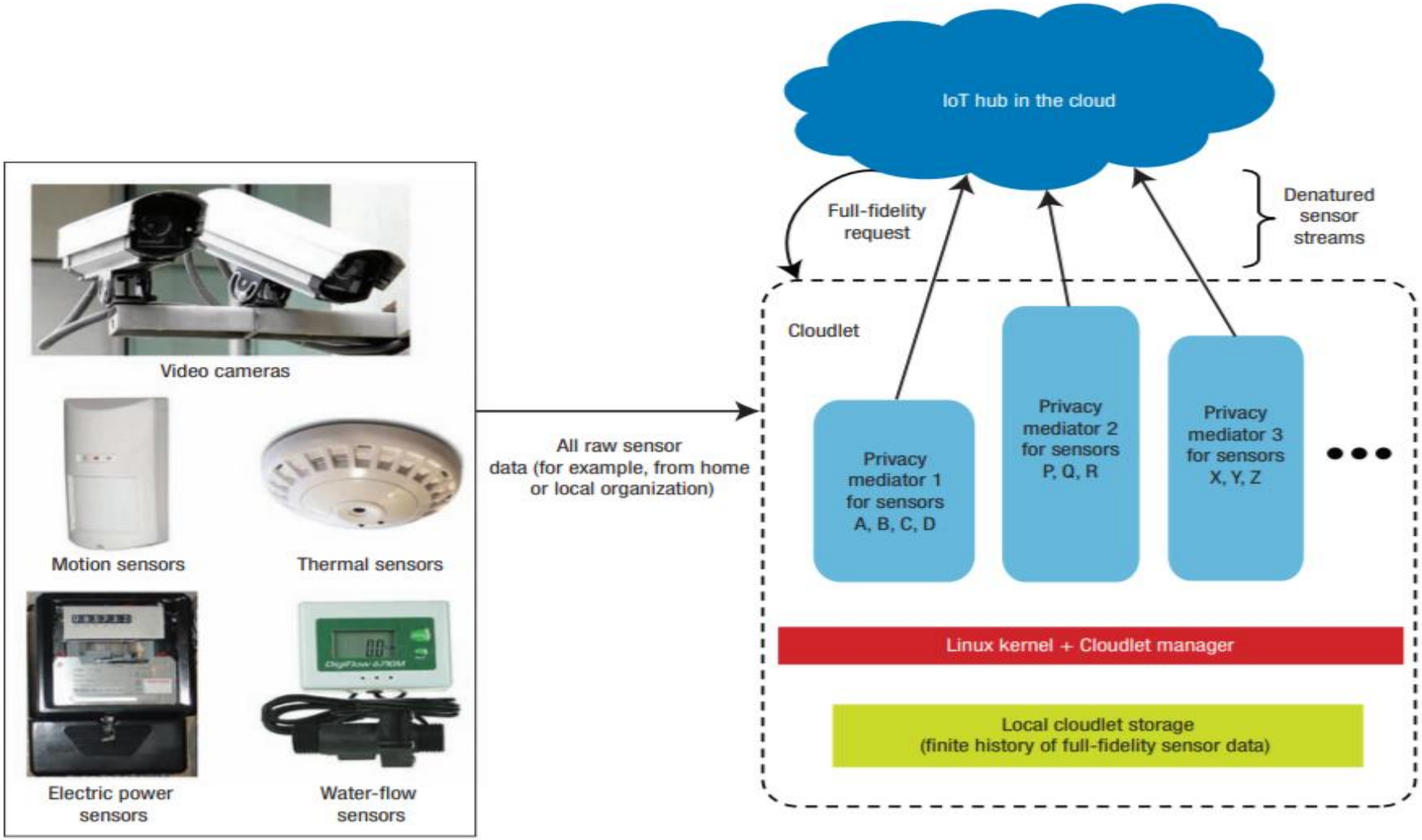
# AI and Edge Computing



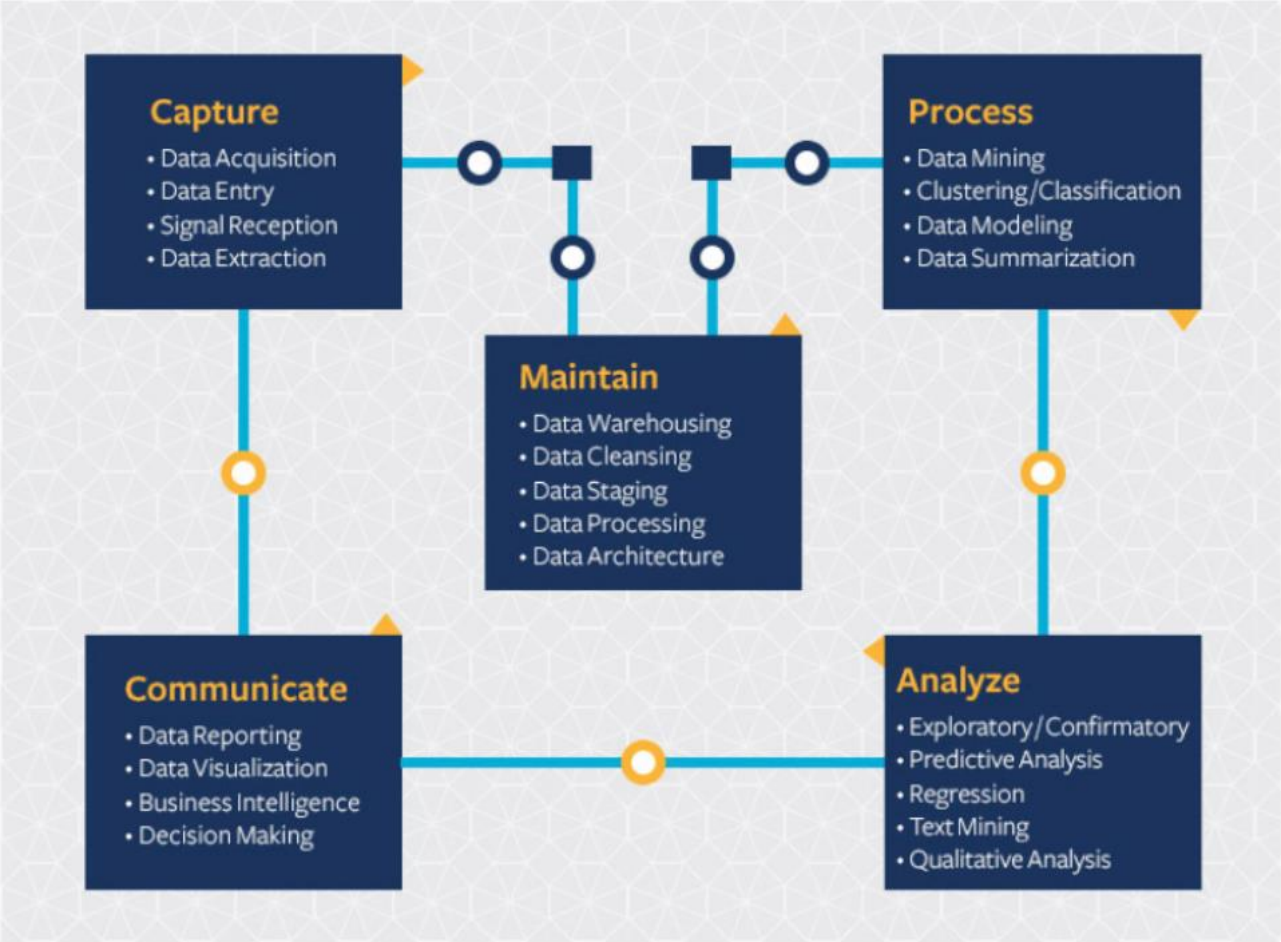

- Edge云平台运行面向Edge Domain的局部化数据处理
- 同云端的大数据AI系统进行配合，形成分布式机器学习和数据分析系统



# 边-雾计算的例子


IoT hub in the cloud
Full-fidelity request
Denatured sensor streams
Cloudlet
Privacy mediator 1 for sensors A, B, C, D
Privacy mediator 2 for sensors P, Q, R
Privacy mediator 3 for sensors X, Y, Z
Linux kernel + Cloudlet manager
Local cloudlet storage (finite history of full-fidelity sensor data)
All raw sensor data (for example, from home or local organization)
Video cameras
Motion sensors
Thermal sensors
Electric power sensors
Water-flow sensors

# 大数据科学 Big Data Science




大数据科学的生命周期

# 数据科学家 Data Scientist

- Data Scientist

“The ability to take data — to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it — that’s going to be a hugely important skill in the next decades.”

- Hal Varian, chief economist at Google and UC Berkeley professor of information sciences, business, and economics

# 数据科学家 Data Scientist

- 数据科学家 会什么?
- Data Scientist: 分析数据，建模型，写算法
  - Programming skills (SAS, R, Python), statistical and mathematical skills, storytelling and data visualization, Hadoop, SQL, machine learning
- Data Analyst: 把数据分析和业务分析结合起来
  - Programming skills (SAS, R, Python), statistical and mathematical skills, data wrangling, data visualization
- Data Engineer：搭建数据分析平台，管理维护数据
  - Programming languages (Java, Scala), NoSQL databases (MongoDB, Cassandra DB),frameworks (Apache Hadoop)

# 数据科学家 Data Scientist

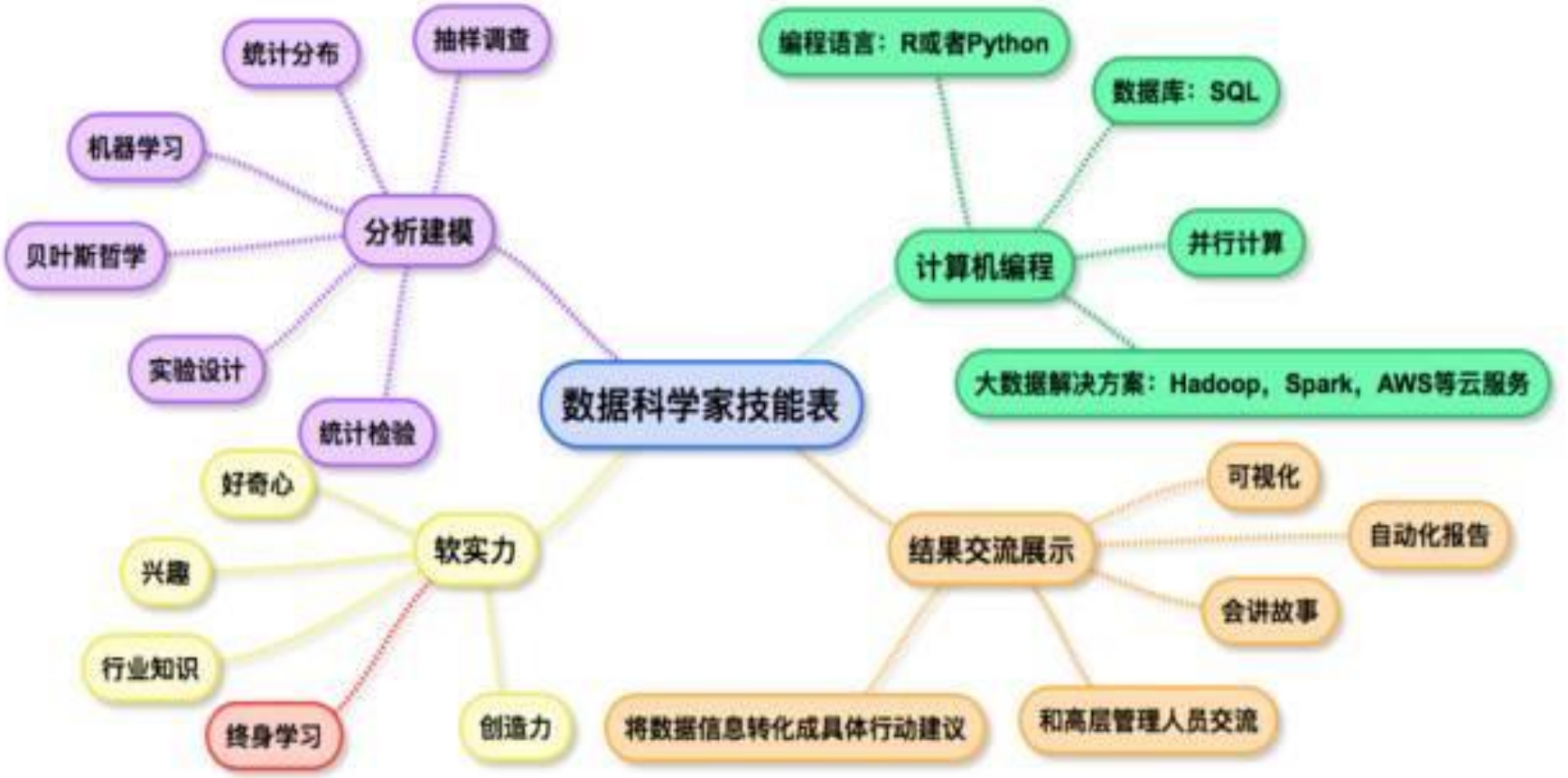



图1 数据科学家的技能

# 课程的主要学习目标

- 掌握大数据科学的基本概念
  - 数据密集型计算、大数据分析建模
  - AI与大数据、云计算与大数据等
- 掌握大数据处理的主要技术
  - 大数据处理软件的基本原理
  - 大数据处理软件的使用方法
- 了解大数据分析的典型应用
- 了解如何成为大数据科学家和工程师 (Data Scientist, Data Engineering )

# 课程的主要学习内容

- 大数据科学的基本概念
- Hadoop和MapReduce编程
- Spark和大数据机器学习、深度学习
- NoSql系统: PIG/HIVE/Impala，MongDB，图数据库、流式处理等
- 大数据科学应用实例：商业智能、Web服务分析、智能交通、智慧教育等

# 考核方式：

## 课后小作业 + 课上任务 + 选做作业

- 课后小作业（60分）：
  - 每次在课程网站上完成判断题和填空题，帮助记忆知识点（8次，按时认真完成即可）。
- 课上任务（25分）：
  - 用课上的时间，搭建Hadoop/Spark环境（可使用我们提供的虚拟镜像），熟悉大数据软件环境（1次）。
- 选做作业（15分）：
  - 组队（3-5人）实现一个小型大数据分析例子，给同学们提供数据集和技术文档。（课程压力大的同学可以选择不做）。