# 大数据与AI的发展应用

何宝宏
中国信息通信研究院

微信订阅号：何所思

# 个人画像

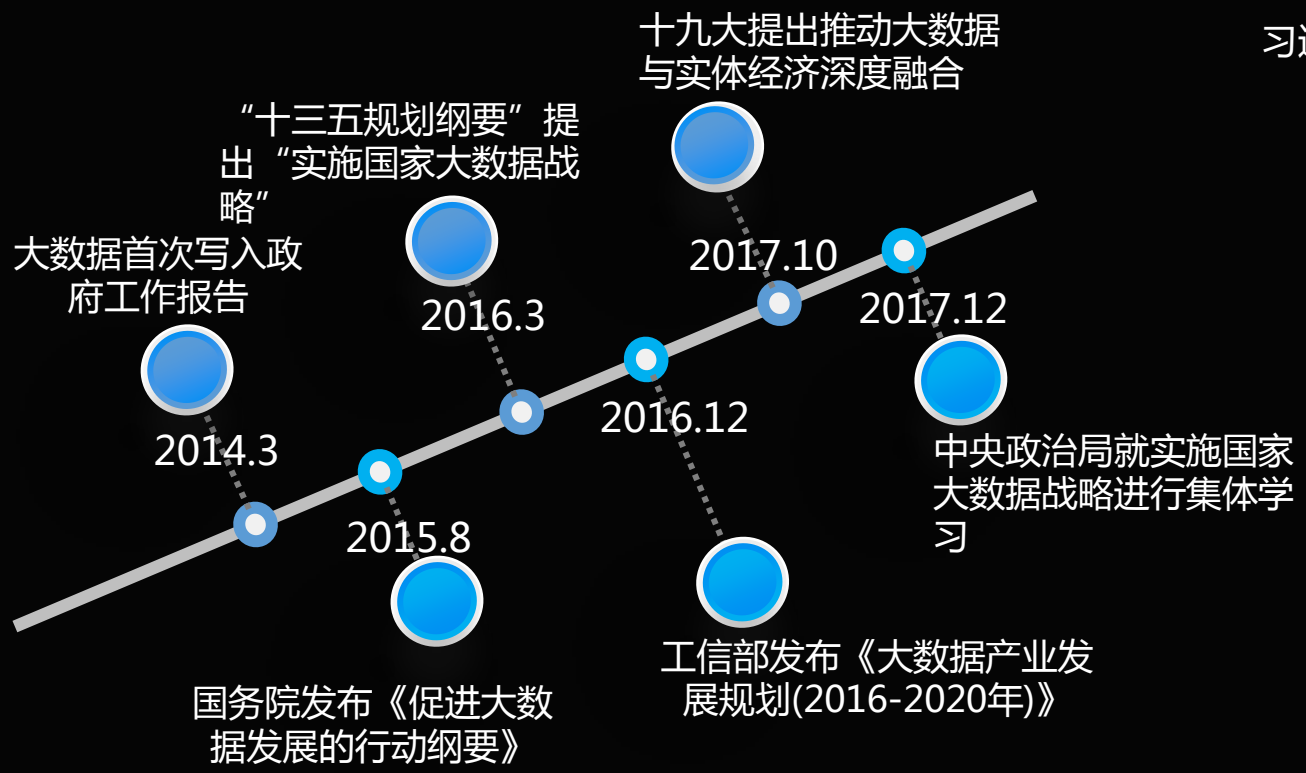
20余年互联网研究的老兵。

现主要禅修如何为互联网技术相面。

1999年毕业于中科院计算所，获计算机应用博士学位。

# ① 大数据

# 大数据政策热度持续攀升




习近平总书记提出了五个方面的要求

**1 推动大数据技术产业创新发展**

**2 构建以数据为关键要素的数字经济**

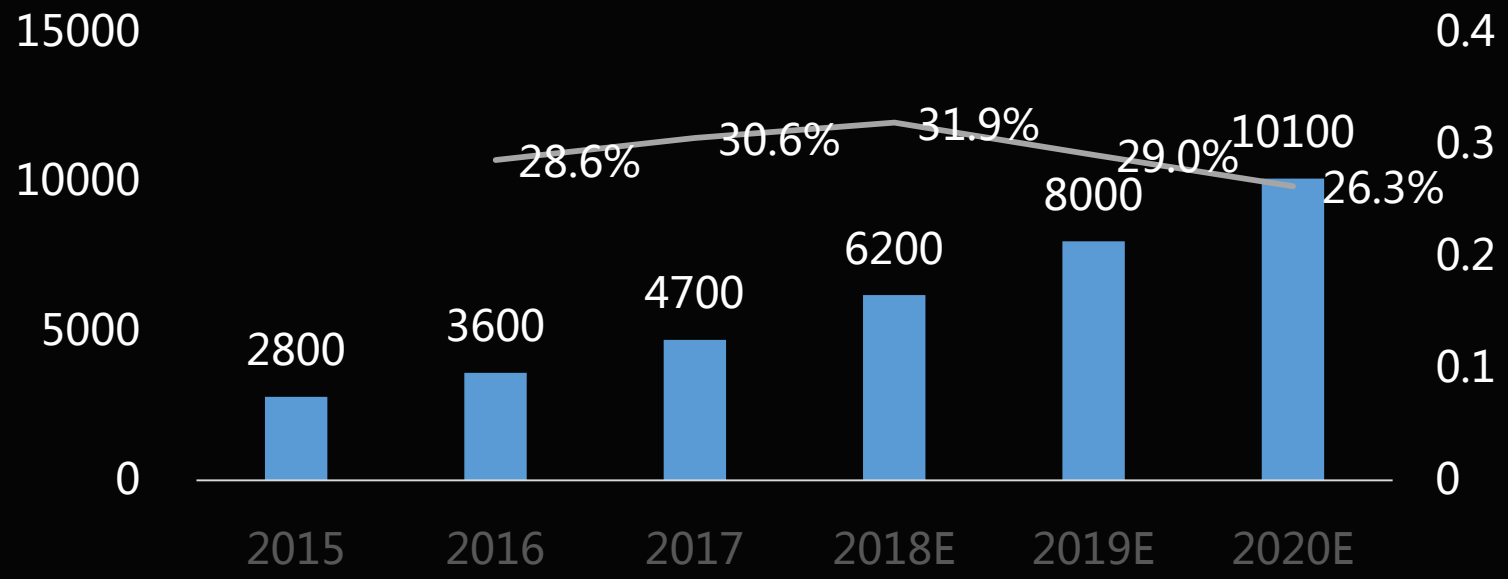
**3 运用大数据提升国家治理现代化水平**

**4 运用大数据促进保障和改善民生**

**5 切实保障国家数据安全**

# 我国大数据产业规模达到4700亿

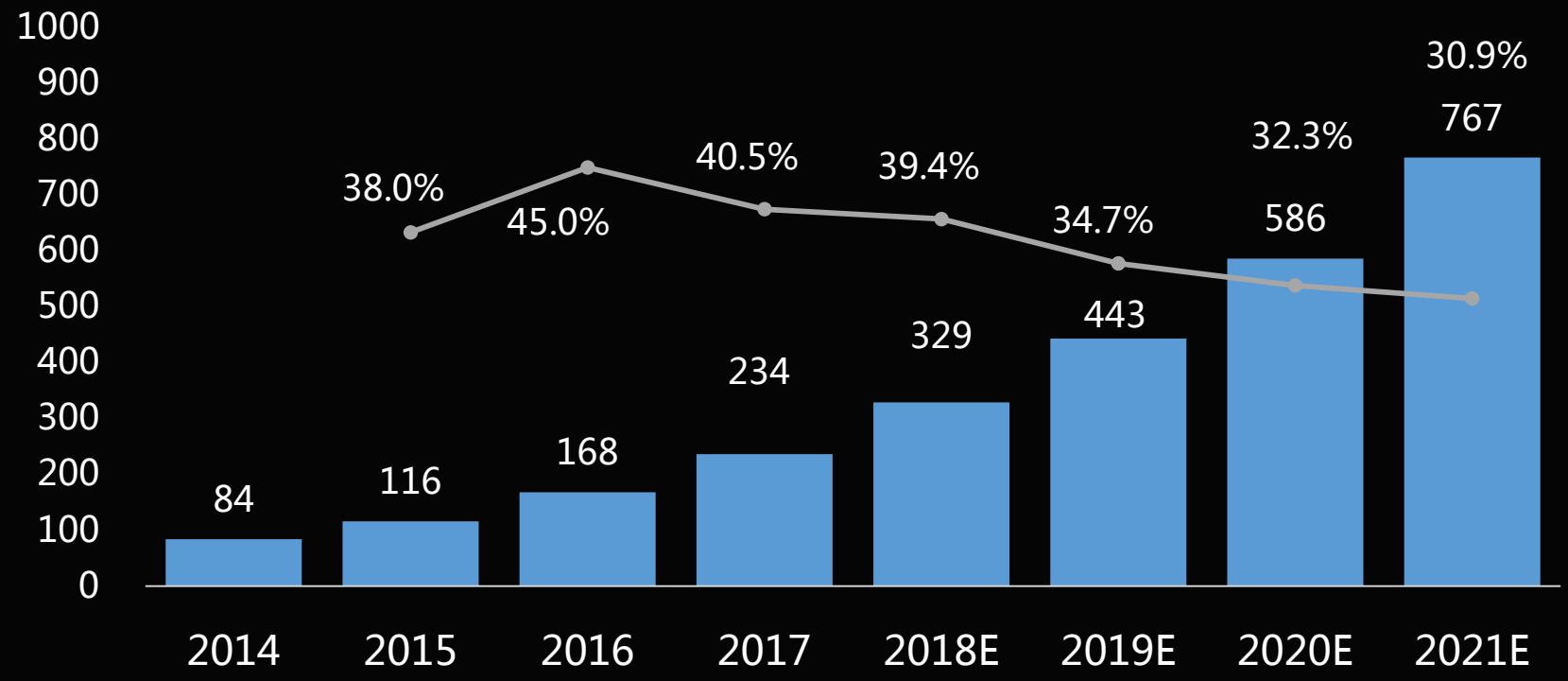
中国信息通信研究院结合对大数据相关企业的调研测算，2017年我国大数据产业规模为4700亿元人民币，同比增长30.6%。




注：大数据产业指以数据生产、采集、存储、加工、分析、服务为主的相关活动，包括数据资源建设、大数据软硬件产品的开发、销售和租赁活动，以及相关信息技术服务。

# 其中，核心产业规模234亿

大数据软硬件产品产值约为234亿元人民币，同比增长39%。



# 技术发展的大事务，大融合

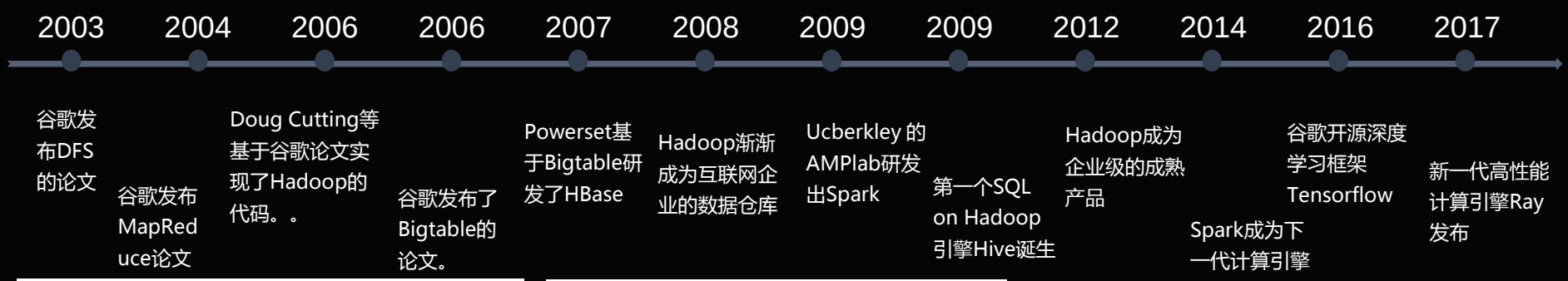
1 **大数据分析技术** OLAP，2000年代转向分布式，快速迭代，软硬&AI融合

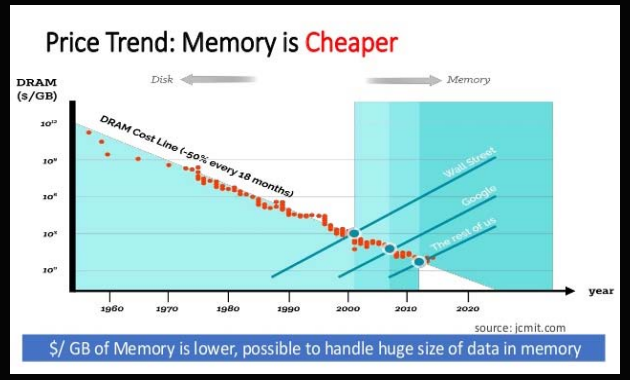
2 **大事务处理技术** OLTP，生产系统/生命线/门槛高，目前正在转向分布式

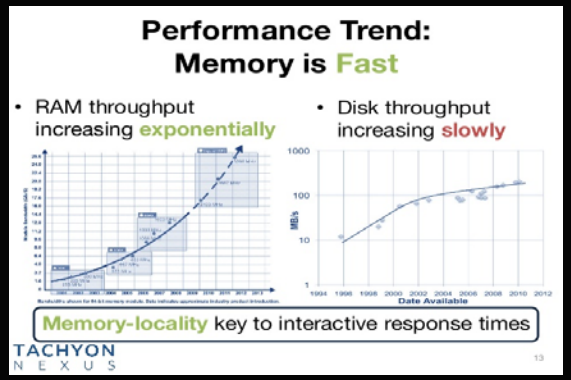
3 **大数据流通技术** 用技术手段缓解数据共享矛盾，以往技术力量被低估

# OLAP的分布式之旅

2000开启分，近几年快速迭代，软硬协同成为趋势



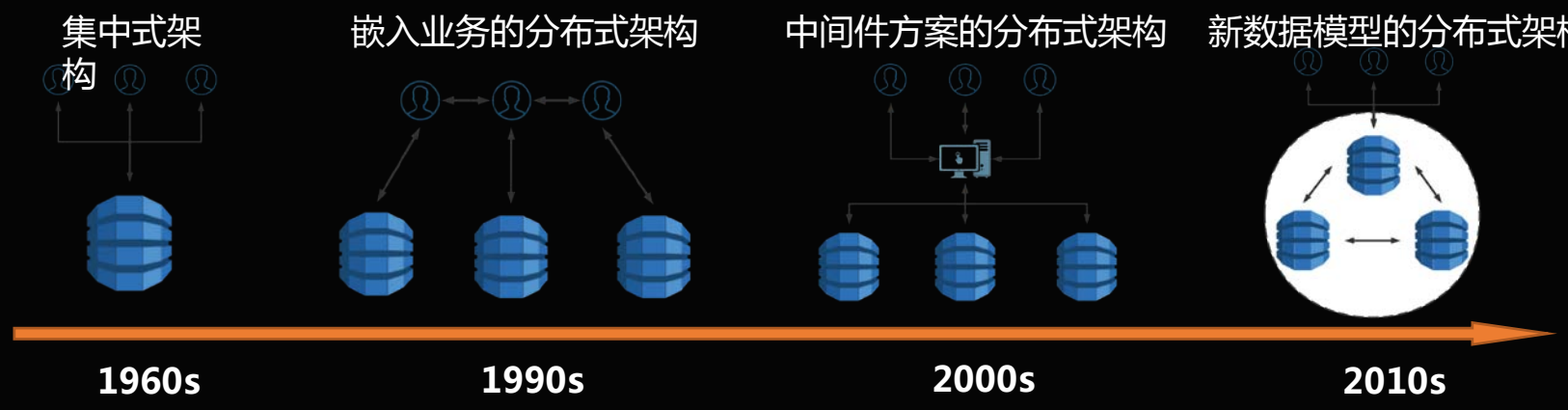




存储越来越快：NVRAM有可能替代DRAM成为主流

计算成为瓶颈：由单核向多核扩展，演变为GPU、FPGA和ASIC异构计算

# OLTP迎来分布式拐点

事务型数据处理业务的高实时、强一致性、高并发、高可靠强要求成为集中式向分布式转型的难点。但目随着核心业务量增长，OLTP向分布式转型将迎来拐点。



# 大数据带来了隐私的烦恼

某耿直BOY: 中国人多数情况下愿意用隐私交换便捷性

欧盟GDPR: 个人隐私数据的明确同意和流动权，精益数据


头条 今日头条 你关心的 才是头条

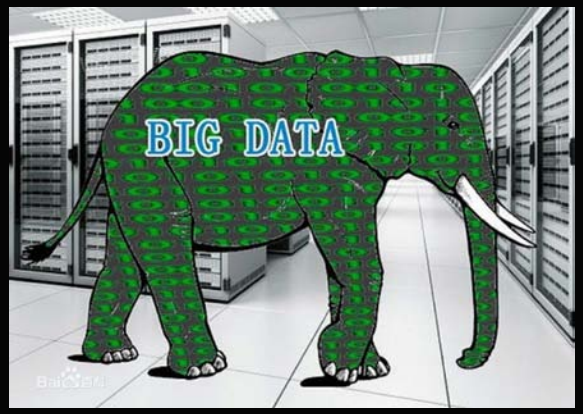
# 大数据遭遇了“7年之痒”

**“大数据：创新、竞争和生产率的下一个前沿”**

2011年5月，麦肯锡

**“大数据时代降临”**

2012年2月，纽约时报



# 大数据的烦恼，只因又一个三角困境

国家安全

个人隐私

便利性

“不可能既享有100%的安全又享有100%的隐私，而且没有丝毫不便，我们不得不做出选择。”

2013年，奥巴马，时任美国总统

# 这些意味着什么？

短期：一些大数据应用真正落地了

中期：保护隐私是未来3年工作的重中之重

长期：已20余年的互联网免费模式终结的开始？

# Now，2018年

从重点是发展，到重点是合规

从高估短期影响，到低估长期影响

# 从时间看，要准备过段“紧日子”

2012-2016

形成泡沫

2017-2020

回归理性

2021-2028

规模商用

# 从应用看，金融大数据的痛点还很多

- 金融领域业务运营和创新需要数据治理标准
  - 统一数据指标体系
  - 元数据和主数据管理方案
  - 统一数据采集、交换标准
  - 形成数据质量监测、运营体系
- 金融监管机构对金融领域监管依赖于大数据治理
  - 数据的报送格式
  - 数据质量要求

## 2018年3月银监会（原）发布《银行业金融机构数据治理指引》

- 将数据治理纳入公司治理范畴，与公司治理评价和监管评级挂钩。
- 鼓励开展制度性探索
- 建立良好数据文化，树立数据是银行重要资产和数据应真实客观的理念与准则。
- 加强数据应用，发挥数据价值
- 强化数据安全意识，依法合规采集数据，防止过度采集、滥用数据，依法保护客户隐私。

# 从算法看，关键是透明性

2018年

算法黑箱

算法歧视

算法疫情

2020年

开放算法

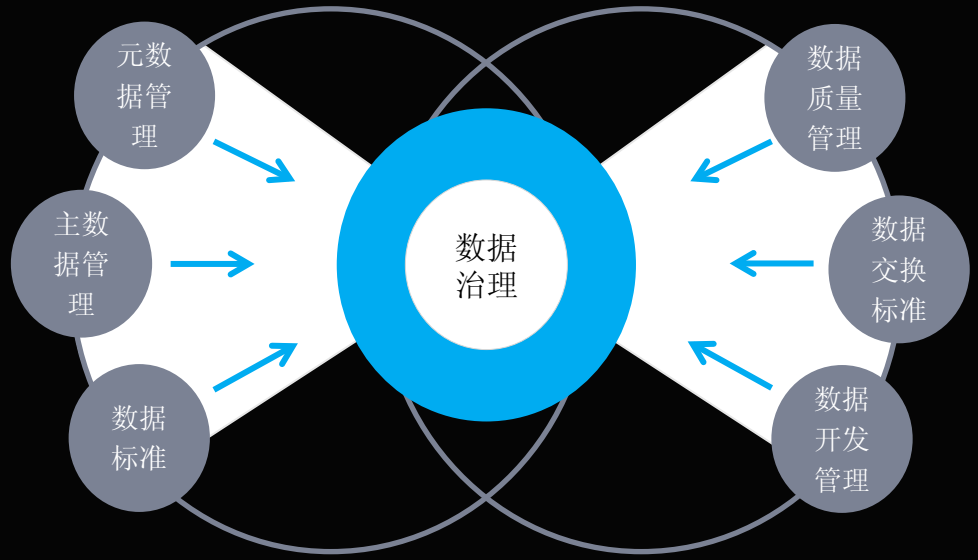
算法中立

监管算法

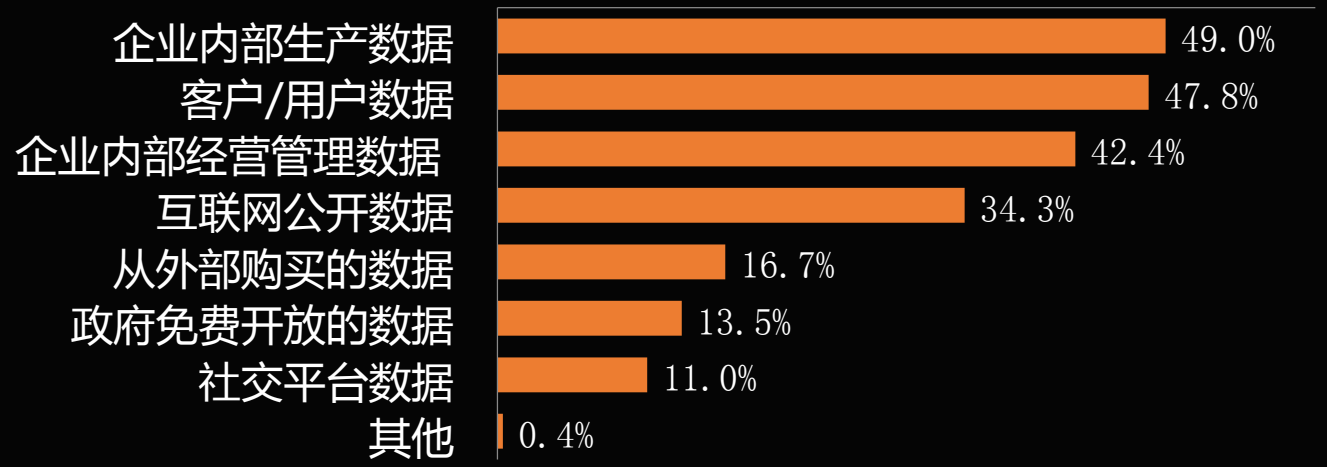
# 从资产看，KNOW 2 HOW

**刚知道了数据是资产**

**还不知道如何变资产**



# 从流通看，现在是男耕女织，30年后的诺奖诞生地




**从农工产品贸易到数字产品贸易的转折，从工业经济学到数字经济学的跃迁**

# 新流通，新技术

**安全多方计算**：应用于联合数据分析、数据安全查询、数据可信交换等

**区块链**：应用于数据存证确权、精细数据授权使用、数据溯源等

同态加密　零知识证明　群签名　环签名　差分隐私

| | | | | | |
|---|---|---|---|---|---|
| **技术特点** | 可在不解密的情况下对密文进行计算和分析 | 证明者无需任何事件相关数据，就能向验证者证明事件的真实可靠 | 能为签名者提供较好的匿名性，同时在必要时又通过可信管理方追溯签署者身份 | 不需要分配指定的密钥，无法撤销签名者的匿名性 | 具有严谨的统计学模型，能够提供可量化的隐私保证 |
| **适用场景** | 云计算、电子商务、物联网等 | 电子商务、金融、银行、电子货币等 | 公共资源管理、电子商务、电子货币等 | 云存储、电子货币等 | 电子商务、物联网等 |

# 从价值看，发掘新价值新用途？

| | 石油 | | 大数据 | |
|---|---|---|---|---|
| 早期 | 冶炼：煤油<br>用途：照明 | | 冶炼：用户画像<br>用途：营销风控 | 早期 |
| 1879年 | 爱迪生危机 | | 隐私危机 | 2018年 |
| 危机后 | 冶炼：汽油<br>用途：动力 | | 冶炼：？<br>用途：“非人”？ | 危机后？ |

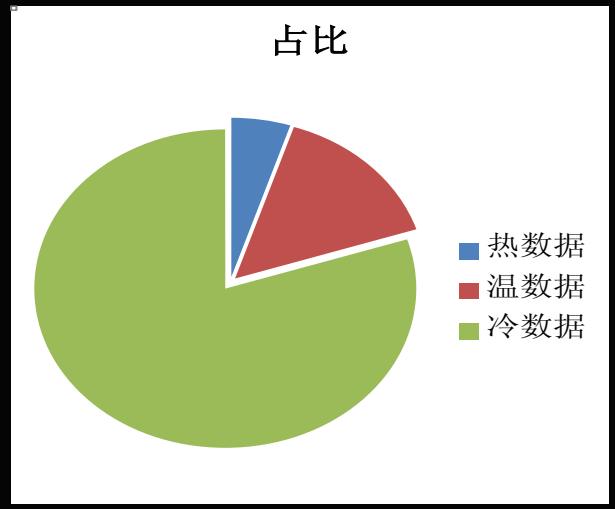
# 从归宿看，日益细分

普通数据，大多会“进冷宫”

GDPR，隐私数据能“要灭绝”

区块链，价值数据会“得永生”

介质原因，很多数据会“被失踪”



# 从风险看，哪里有风险哪里就有保险

# 未来，被机器遗忘将是一种幸福

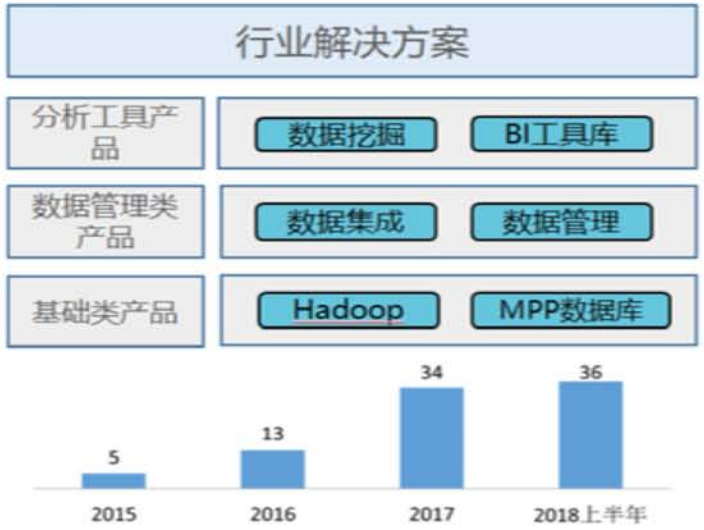


千年后，如何考古呢？

公元前2018　　对　　公元后2018

# 信通院大数据工作介绍

从2015年6月到现在，大数据产品能力评测形成了基础能力和性能专项两类评测，形成评测标准15项。截、第六批评测，已经有55产品通过评测，完成88个测试。信通院大数据部共发布白皮书和大数据调查报告10本，覆盖知识图谱、数据流通技术、分布式数据库、工业大数据架构、数据资产管理实践等方向。




| 类别 | 名称 | 目前版本 |
|---|---|---|
| 基础平台产品 | 大数据产品 Hadoop平台 第1部分：技术要求 | 1.0 |
| | 大数据产品 Hadoop平台 第2部分：基础能力测试方法 | 2.0 |
| | 大数据产品 Hadoop平台 第3部分：性能测试方法 | 2.0 |
| | 大数据产品 MPP数据库 第1部分：技术要求 | 1.0 |
| | 大数据产品 MPP数据库 第2部分：基础能力测试方法 | 1.5 |
| | 大数据产品 MPP数据库 第3部分：性能测试方法 | 1.0 |
| 数据管理产品 | 大数据产品 数据集成工具 第1部分：技术要求 | 1.0 |
| | 大数据产品 数据集成工具 第2部分：基础能力测试方法 | 1.0 |
| | 大数据产品 数据管理平台 第1部分：技术要求 | 1.0 |
| | 大数据产品 数据管理平台 第2部分：基础能力测试方法 | 1.0 |
| 分析工具 | 大数据产品 商务智能（BI）分析工具 第1部分：技术要求 | 1.0 |
| | 大数据产品 商务智能（BI）分析工具 第2部分：基础能力测试方法 | 1.0 |
| | 大数据产品 数据挖掘平台 第1部分：技术要求 | 1.0 |
| | 大数据产品 数据挖掘平台 第2部分：基础能力测试方法 | 1.0 |
| 应用和解决方案 | 大数据解决方案 用户行为数据分析 第1部分：技术框架和指标体系 | 1.0 |

电信　工业　政务　金融

# 信通院大数据工作介绍

## 国内数据资产管理理论框架的构建者

数据成为资产已经形成共识，如何进行资产化是未来的研究重点，我们梳理了《**数据资产管理实践白皮书**》，在传统数仓构建的方法论基础上探索基于数据湖架构的管理框架。



## 国内分布式事务数据库标准和基准测试工具的制定者

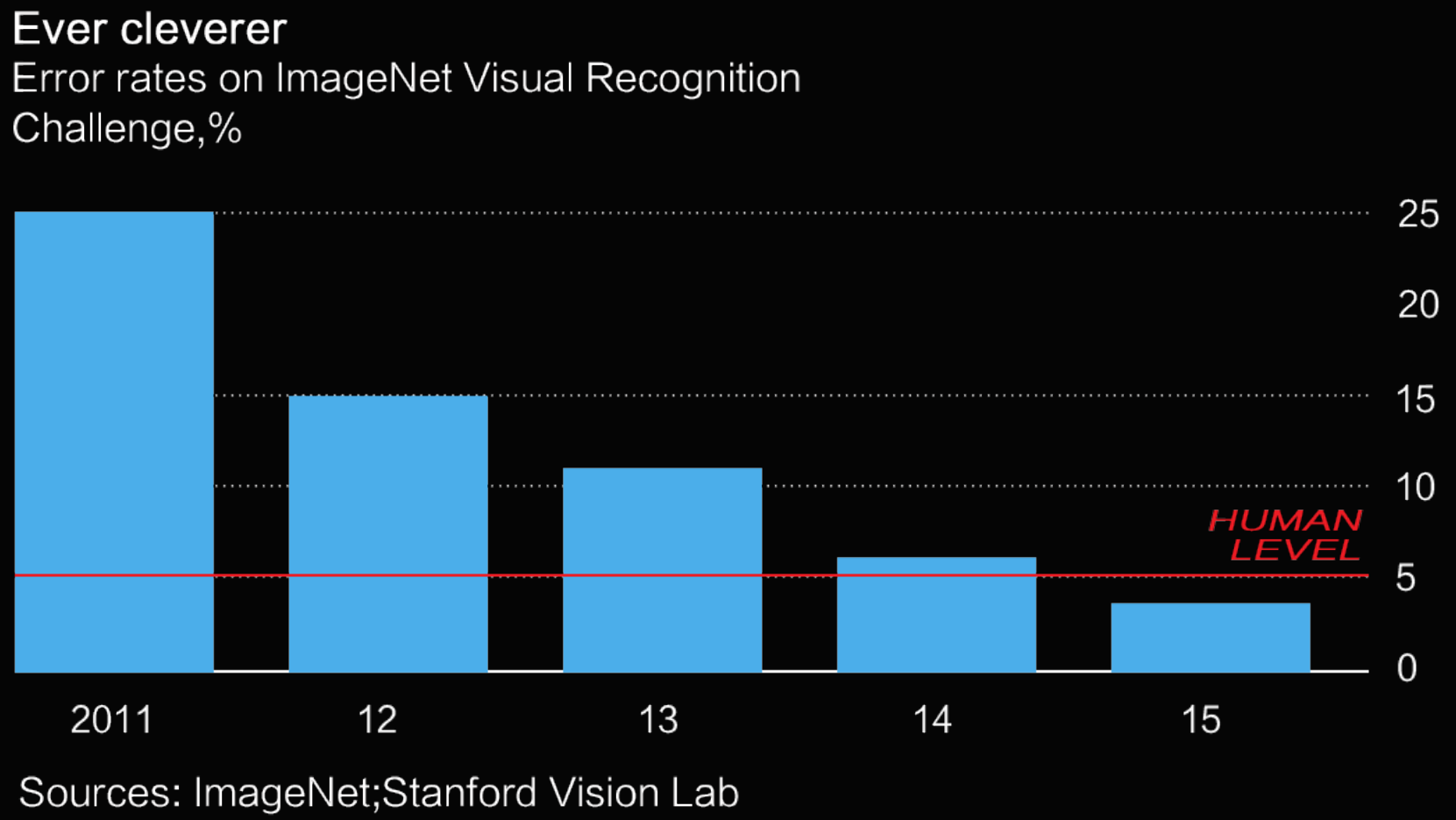
- 制定定业内首个《大数据产品 分布式事务数据库 第一部分：技术要求》
- 制定业内首个《金融分布式事务数据库白皮书（1.0）》


## 国内数据流通技术与政策研究的桥头堡

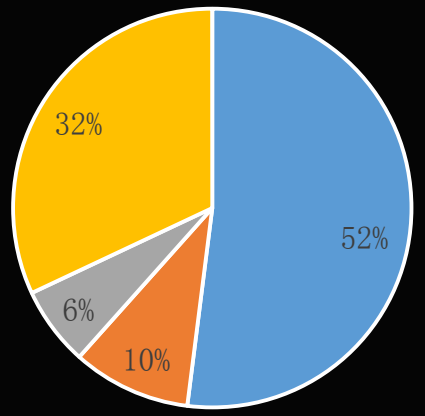
- 建立了首个数据流通合规的评测体系，完成8家企业的评测。
- 支持网信办发布“大数据行业自律公约”，发布《数据流通技术白皮书（1.0）》。
- 形成4个数据流通标准

# ② 人工智能

# 机器视觉/语音识别已超越人类……



# AI的典型场景



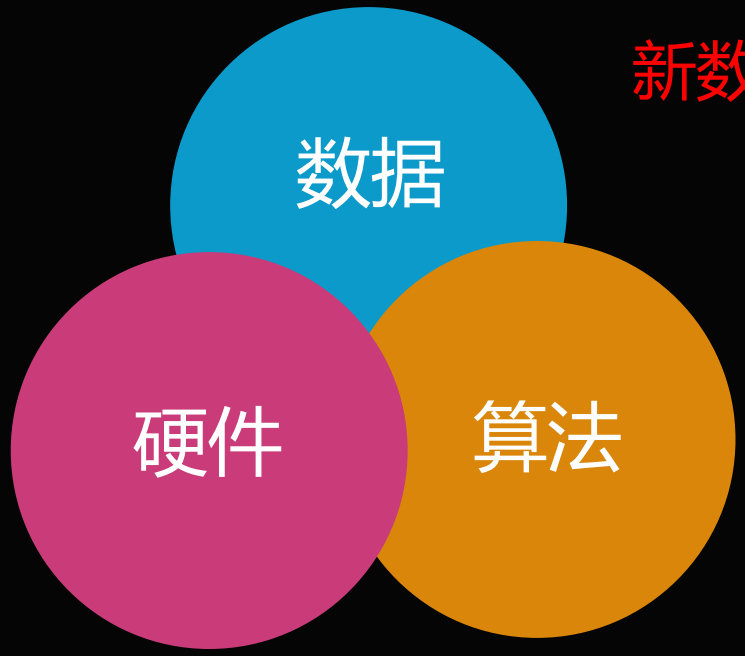

数据来源：平安证券

**监控系统**
- 车牌识别
- 嫌犯追踪
- 火灾识别

**行为识别**
- 情绪识别
- 动作识别
- 色情内容检测

**电商营销**
- 用户画像
- 广告植入
- 门店营销

# AI的复兴

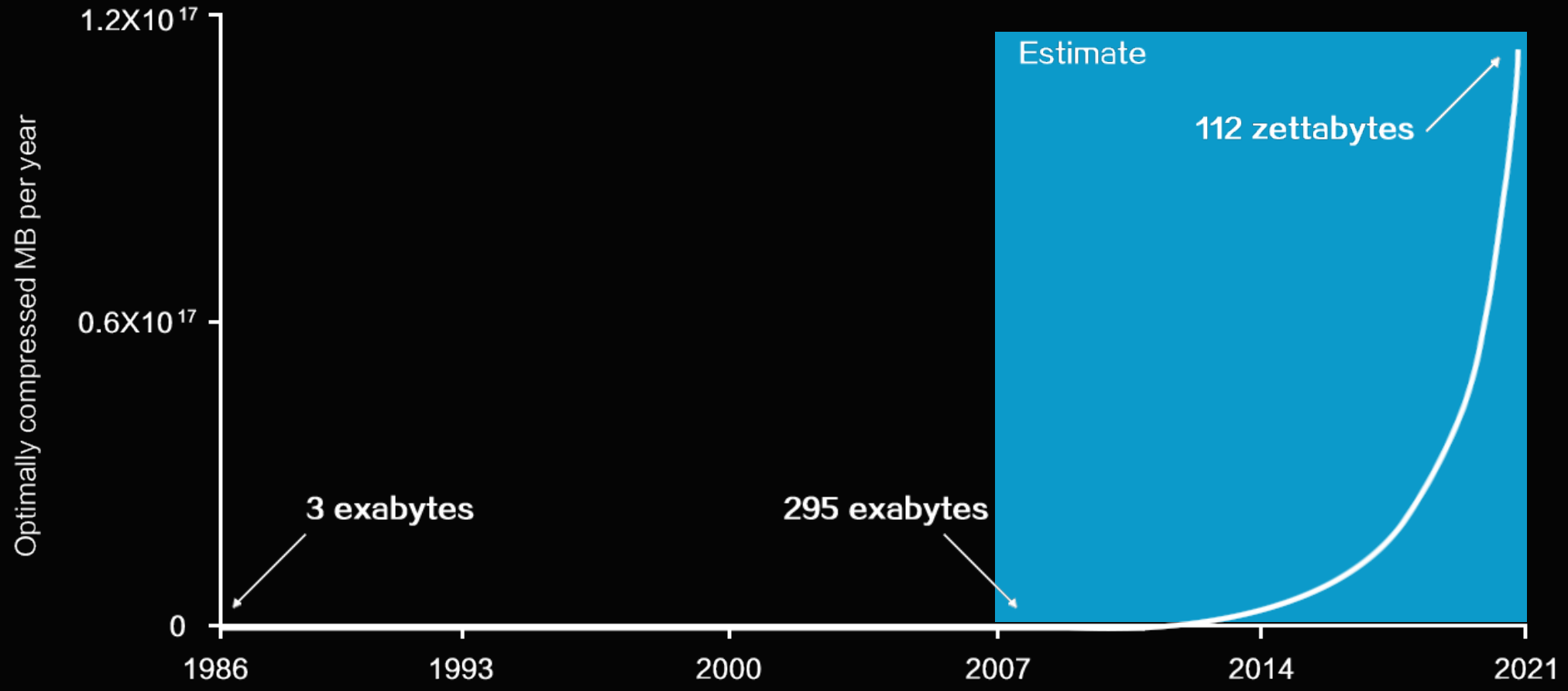



新数据的新应用

老硬件的新应用

老算法的新改进

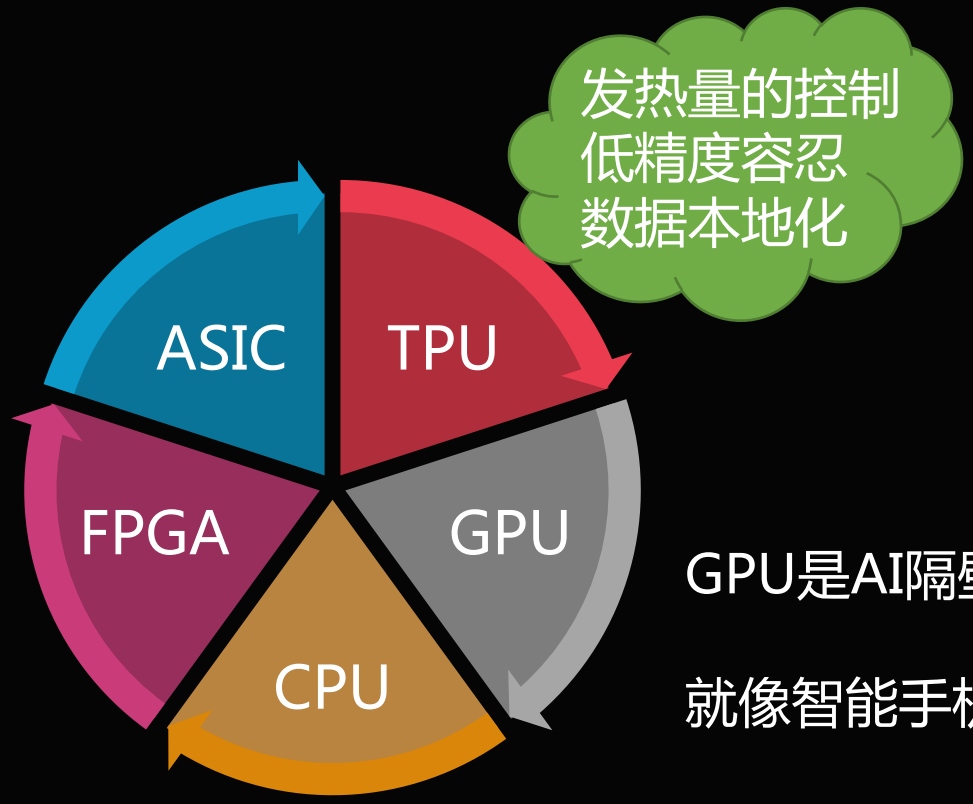
# 现在，训练数据是短板




图片来自网络

数据是人工智能的燃料 —— 吴恩达

# 未来，AI专用硬件有机会

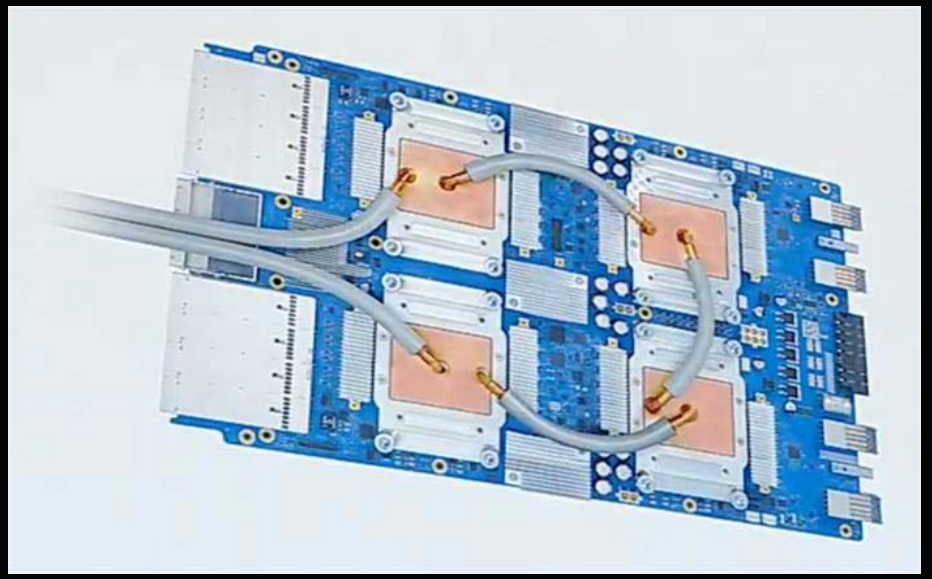



GPU是AI隔壁的老王

就像智能手机需要ARM，AI可能也需要新芯片

# 液冷：再大的风，也吹不凉它那滚烫的芯了

| 每平米功率 | 数据中心密度 | 制冷方式 |
| --- | --- | --- |
| 1.2 KW/机柜以下 | 超低密度数据中心 | 房间级-风冷 |
| 1.2-2.7 KW/机柜 | 低密度数据中心 | 房间级-风冷 |
| 2.7-7.5 KW/机柜 | 中、低密度数据中心 | 行间级-风冷/水冷 |
| 7.5-18 KW/机柜 | 中、高密度数据中心 | 行间级-水冷：液冷-冷板式 |
| 18-30 KW/机柜 | 高密度数据中心 | 液冷-冷板式：液冷-浸没式 |
| 30 KW/机柜以上 | 超高密度数据中心 | 液冷-浸没式 |



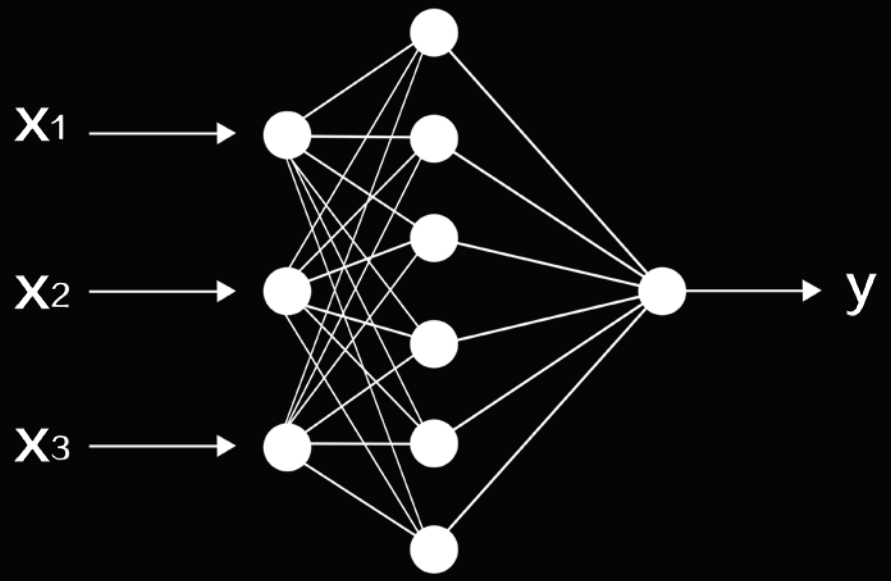
TPU 3.0的性能相比TPU 2.0有8倍的提升，专门设计了一套水冷系统降温。

2018年5月，

Google I/O大会

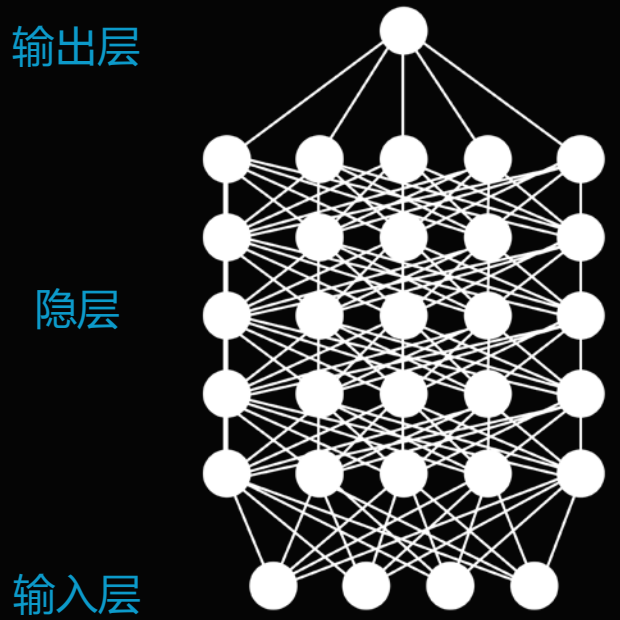
# 接下来，最需要突破的还是算法

机器学习




输入层　　隐层　　输出层

深度学习




输含多个隐层的深度学习模型

# 3年里，AI新工种将不断出现

数据民工

ImageNet

167个国家

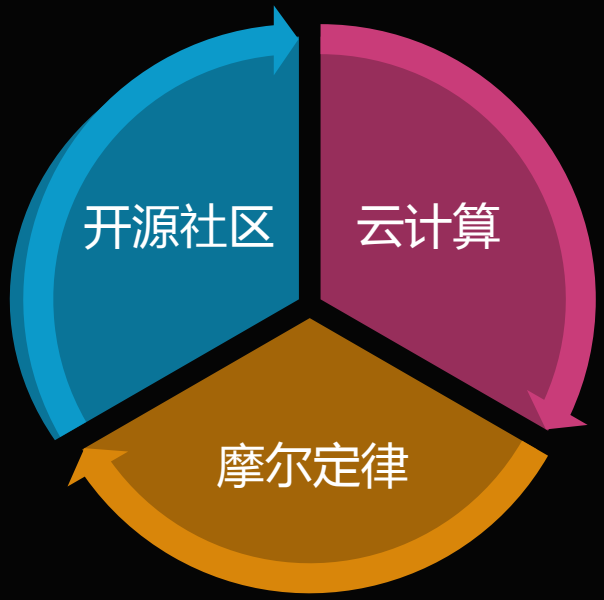
近5万人，2年

标记近10亿图片

得到1500万张

# AI的三大支撑

像Android一样做生态




AI/FPGA/GPU
as a Service

AI 是个“算老虎”

# AI为应用赋(智)能

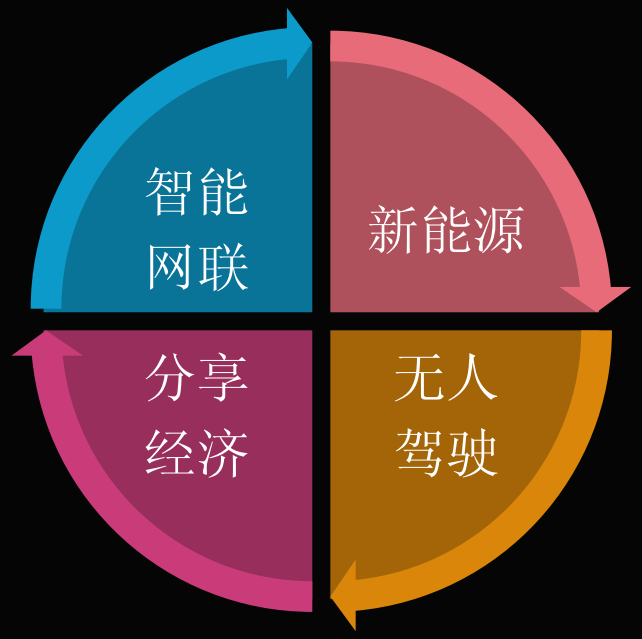
1、人脸识别走向商用

2、智能音箱爆发(短期？)

3、智能服务机器人已开始在临床上尝试应用
骨科、腹腔镜等手术机器人

4、高级智能辅助驾驶系统（ADAS)逐渐成熟

# 汽车业，AI应用的新金矿




路线之争

- 为汽车增加智能

- 为计算机增加轮子

软件定义汽车

重新定义道路

# 3年里，AI将遇到技术的天花板

- 算法：还是黑盒子，调参主要靠运气
- 算力：蛮力计算的巨兽，摩尔定律的衰老
- 数据：有监督学习，数据改变信仰
- 其它：无记忆，无推理，无法与其它方式融合

## 理论已证明，至少这代AI不会觉醒

人的意识是非算法的，

基于图灵机的AI，无法建立起“自我”的概念，

未来控制人类的AI，不会基于图灵机。

# 人生最最痛苦的事情...

量子计算的时代还远未到来，

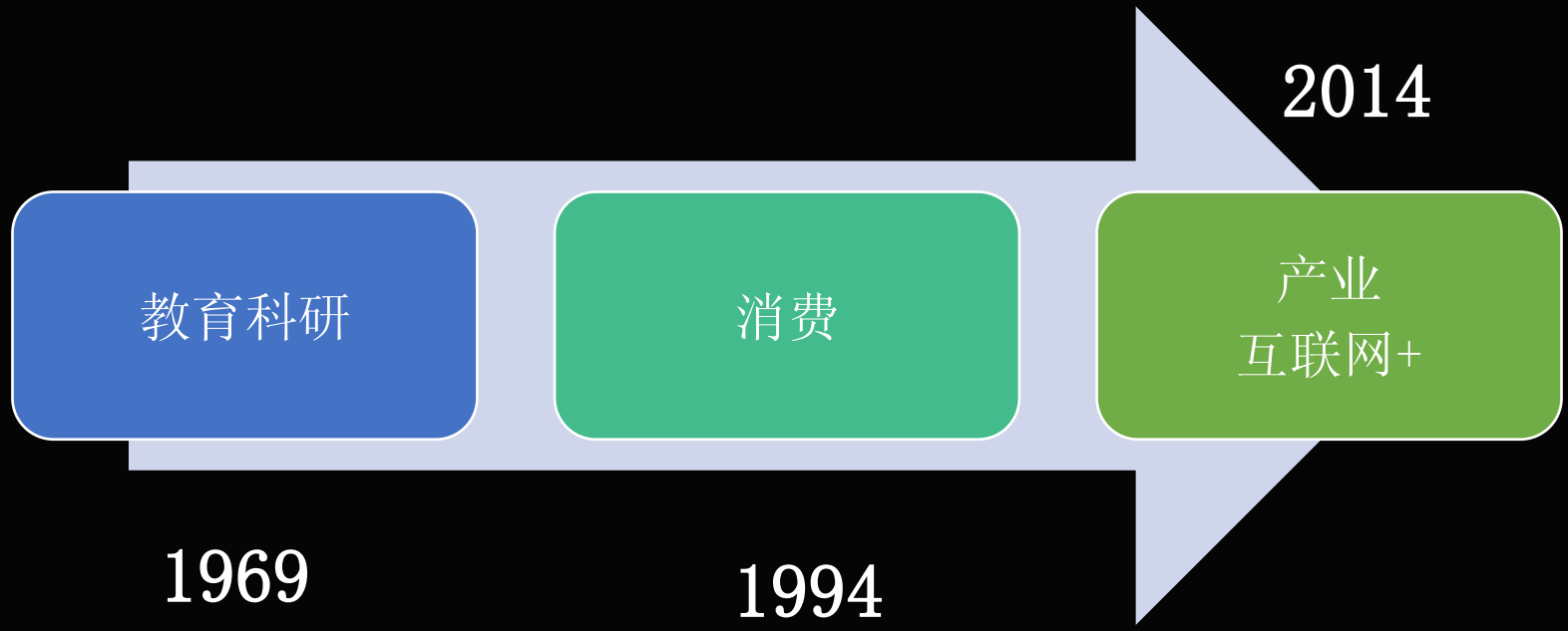
3年里，AI肯定不会控制人类。

人还在，工作却没了。

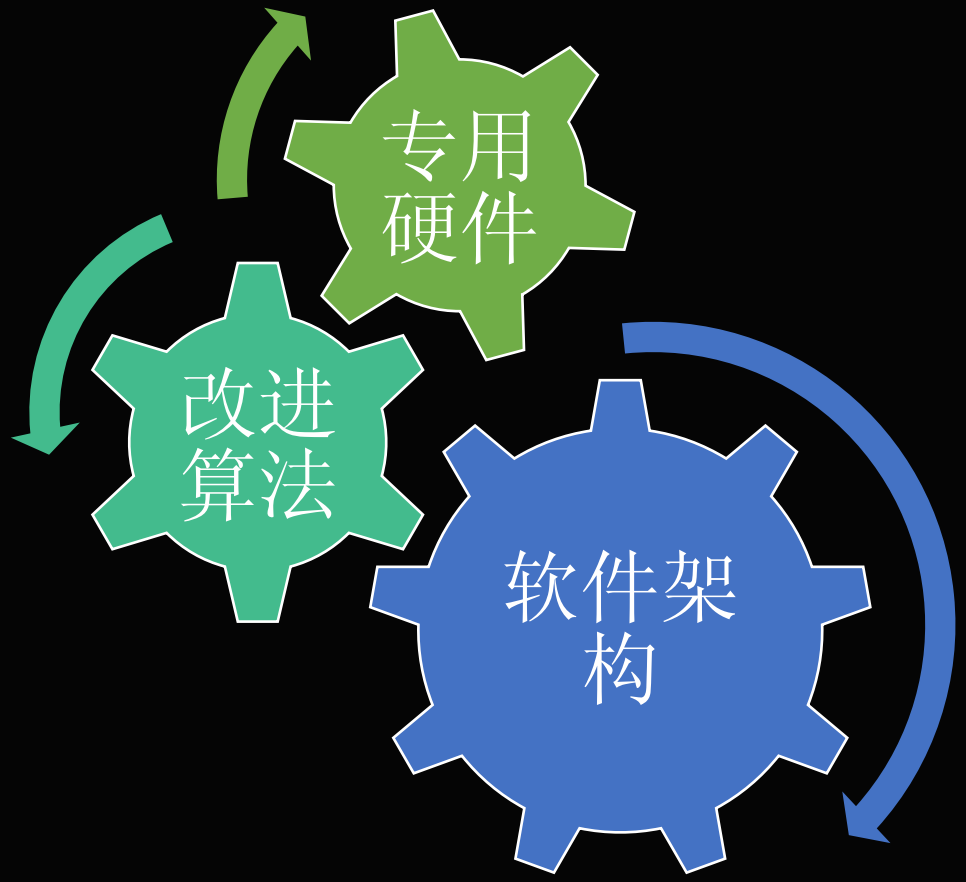
不要害怕

# ③ 小结

# 产业互联网时代，金融本是数据产业



# 当摩尔定律老去...

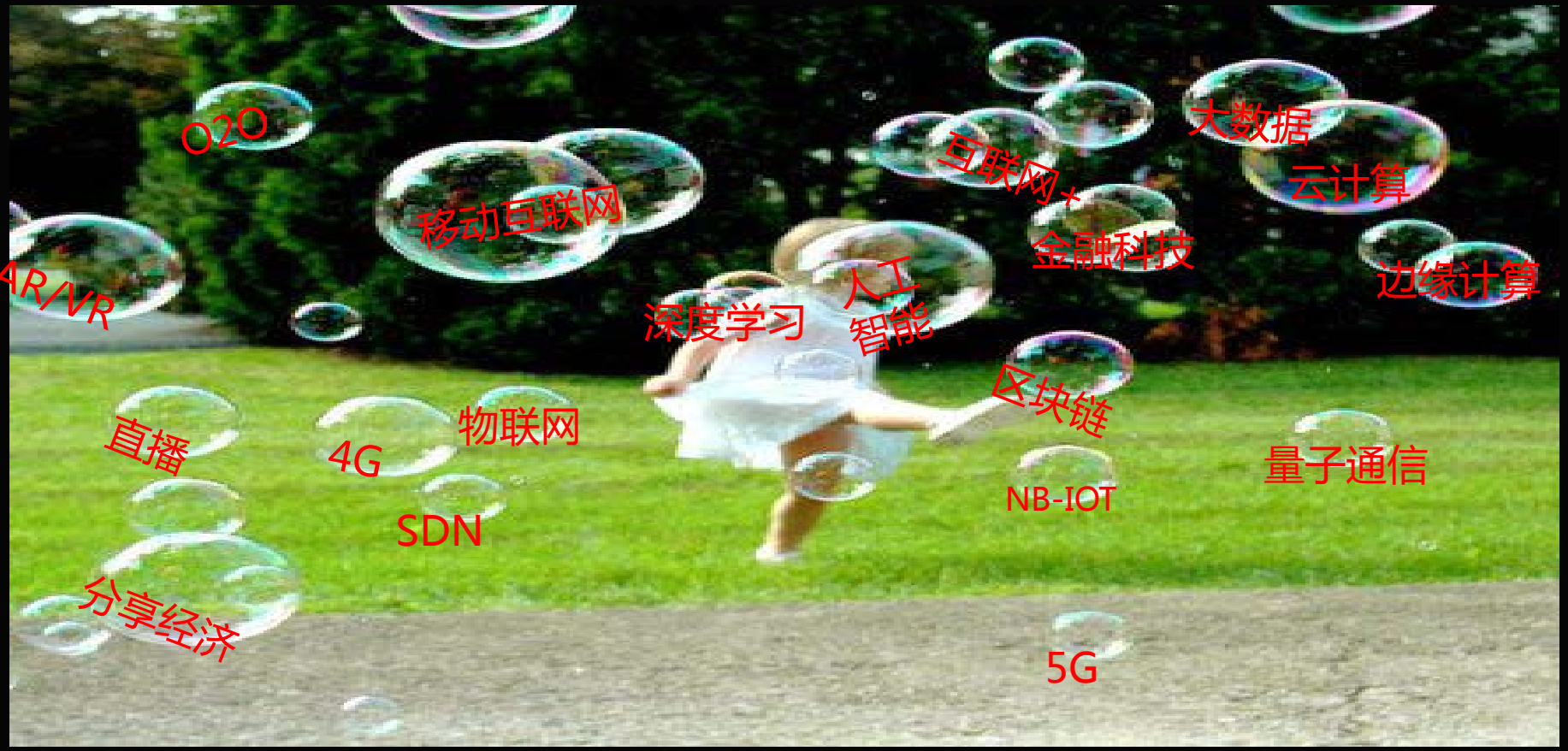


# 当前行业发展的主要矛盾

100年前，
是信息技术跟不上工业革命的步伐

100年后，
是人类社会跟不上信息革命的步伐

# 这10年，我们曾经追过的...



# 2018年，浪潮之巅？

始于2008年的这波技术浪潮，

滞涨期即将到来？

# 接下来的主要机会

…互联网、大数据、人工智能与实体经济深度融合。

# 谢谢！

个人观点　　仅供交流

微信订阅号：何所思