人民邮电出版社 POSTS & TELECOM PRESS
IT大咖说 知识分享平台

重视大数据与人工智能的基本理论和基础设施

李国杰

第四届中国国际大数据大会

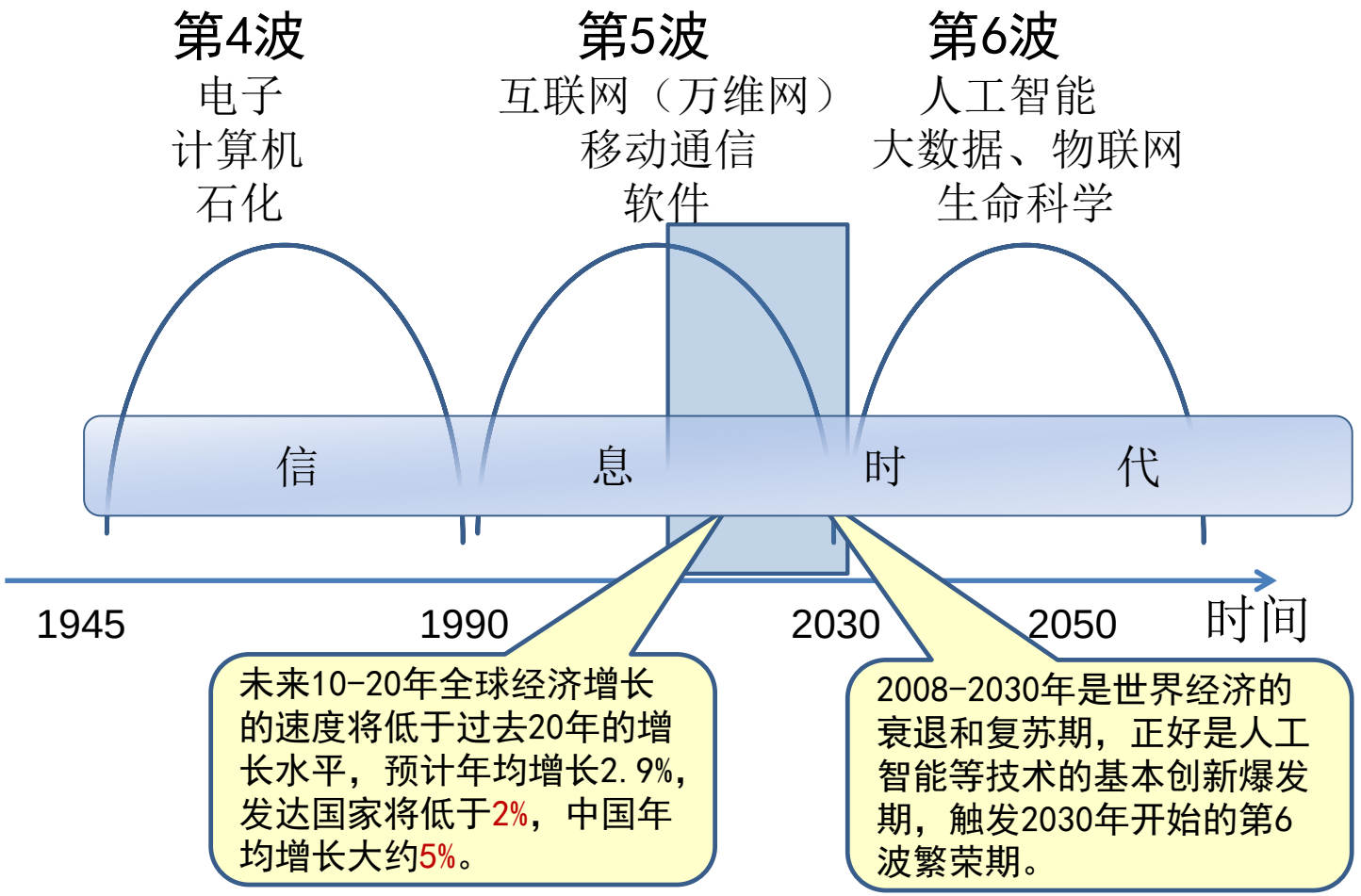
2017.09.26

现在究竟是什么时代？

智能化是信息时代的新阶段

- 目前社会上流行一种观点：人类社会经过信息时代和互联网时代，现在已经进入大数据和人工智能时代。
- 这种认识可能混淆了历史时代和一个时代的多个阶段。从人类社会发展的长周期来看，人类文明只有三个时代：农业时代、工业时代和信息时代。信息时代从二战以后算起，还只有半个多世纪。与工业时代相比，信息时代可能正处在从蒸汽机时代阶段（第一次工业革命）向电气时代阶段（第二次工业革命）的转变期，还有很长的路要走。
- 大数据与人工智能是信息时代的一个新阶段，与其强调智能化与数字化、网络化的区别，不如多强调智能化与信息化的密切联系。数字化和网络化没做好，智能化就是一句空话。

对经济长波的预测


第4波
电子
计算机
石化
第5波
互联网（万维网）
移动通信
软件
第6波
人工智能
大数据、物联网
生命科学
信
息
时
代
1945
1990
2030
2050
时间
未来10-20年全球经济增长的速度将低于过去20年的增长水平，预计年均增长2.9%，发达国家将低于2%，中国年均增长大约5%。
2008-2030年是世界经济的衰退和复苏期，正好是人工智能等技术的基本创新爆发期，触发2030年开始的第6波繁荣期。

从时代判断得出的结论

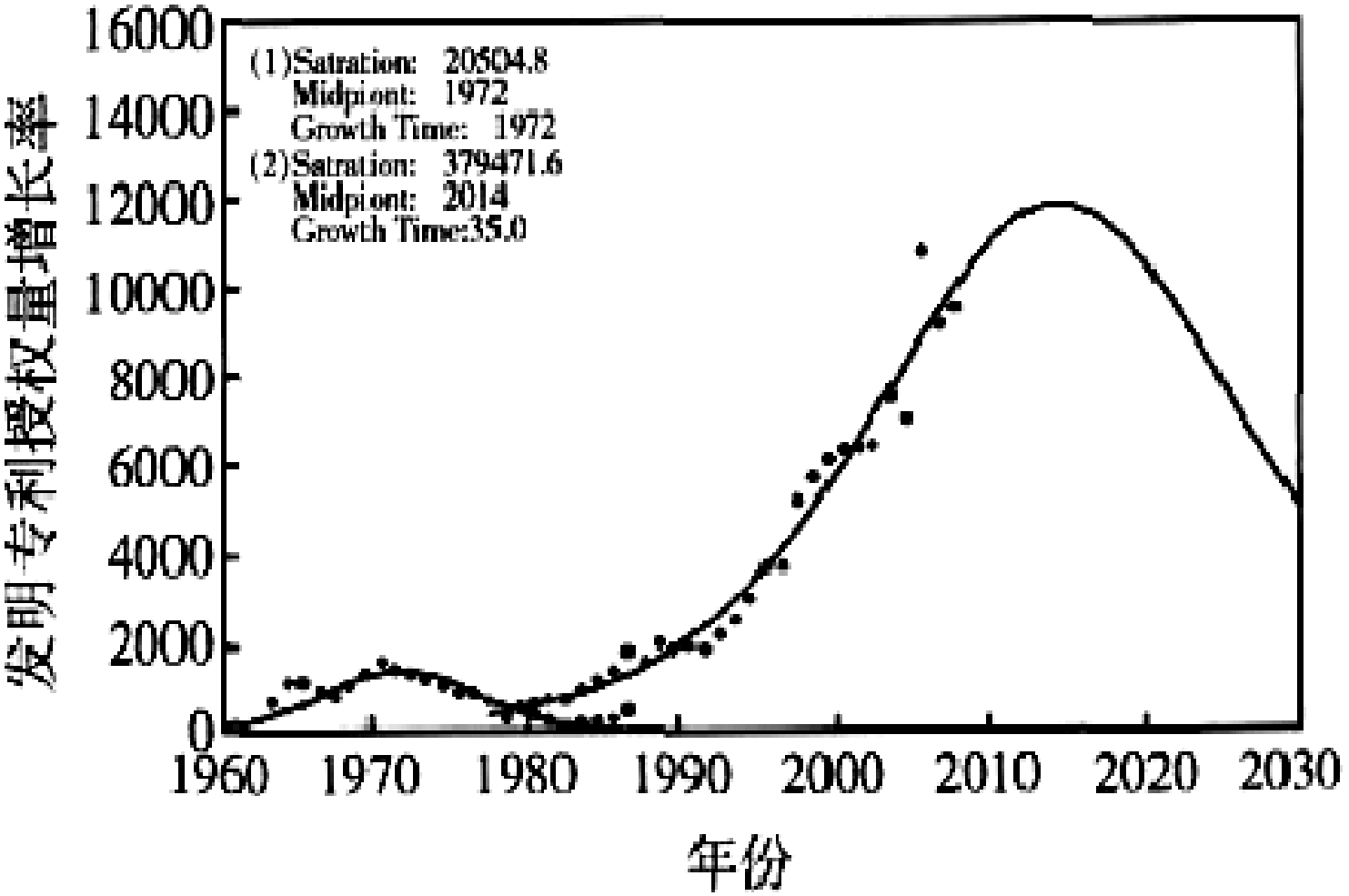
- 从上述时代判断得出两点结论：

1、未来10-15年对经济贡献最大的可能不是大数据和人工智能的新技术，而是信息技术（包括大数据和人工智能）融入各个产业的新产品、提供个性化产品和服务的新业态、产业链跨界融合的新模式。这些创新主要是已知技术的新组合。

2、在经济的衰退复苏期要特别重视基础性技术的发明，未来10-15年应力争在大数据和人工智能领域做出像电子计算机、集成电路、互联网一样的重大发明（不可能规划）。

- 历史上重大基础发明都是经过较长时间的技术改进和扩散之后才产生巨大经济效益，信息技术也不应例外。从2016年到2025年的10年内，汽车、消费品、电力、物流等行业的数字化转型有望带来100万亿美元的社会与企业价值。大数据和人工智能提升传统产业的前景十分光明。

技术发展的两次“S-曲线”



人工智能本质上是计算机技术

- 许多人讲人工智能是新的学科，内容涉及脑科学、计算机科学、统计学、社会科学等。但迄今为止，脑科学（神经科学）对人工智能的贡献很小（人工神经网络只是受到神经解剖McCulloch-Pitts Model 的一点启发）；统计学对推动机器学习的崛起起了较大作用，但没有人把人工智能看成统计学的分支。
- 就目前的人工智能而言，本质上是计算机学科的一个分支，现在国机上人工智能的论文都统计在计算机学科名下。从基础研究开看，人工智能是计算机科学的前沿研究；从应用来看，人工智能是计算机技术的非平凡应用。
- 所谓“智能化”的前提是计算机化，目前不存在脱离计算机的人工智能。

应强调学科融合而不是分道扬镳

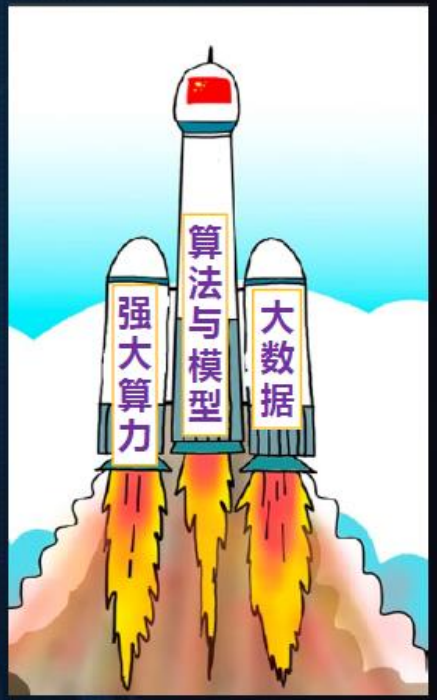
- 从老学科中分离出新学科是常见的事，计算机科技人员应乐见并积极支持新兴学科的成长，盼望其它学科对大数据和人工智能的发展做出更大的贡献。
- 但发展大数据和人工智能要更加注重知识的融合，钱学森先生早就说过：“必集大成，才能得智慧”。人工智能是对付复杂性的科学，也许不存在麦克斯韦方程组形式的通用公式，要在集成融合上下功夫。
- 人工智能的权威学者M.明斯基定义“人工智能的任务是研究还没有解决的计算机问题”，这一定义预示着人工智能将永远与计算机科学绑在一起。历史上绝大多数人工智能技术与产品都已融入主流计算机产品，今后要也会如此。智能时代不是后信息时代，真正的后信息时代可能是（碳基）生物时代（现在是硅基时代，现代版的石器时代）

重视大数据与人工智能的基本理论和基础设施

流行的看法：人工智能 = A + B + C

AI = A + B + C

- 算法(Algorithm)
  - 深度学习（NN的复兴），增强学习...
  - Seeta在深度学习方面深厚的积累，国内最早的团队之一
- 大数据(Big-data)
  - 人脸：数百万人的数亿图片；数十种人脸属性
  - 人车：数百万量级标注图片；多重属性
  - 百万表情数据，数十万无人机数据，手语/手势数据
- 算力(Computing)
  - 训练平台规模化（GPU）：数百块GPU卡



我的看法：大数据（AI）= A+B+C+D+E

A： Algorithm， 算法

B： Basic Theory， Basis 基本理论，基础设施

C： Computing capability 计算能力

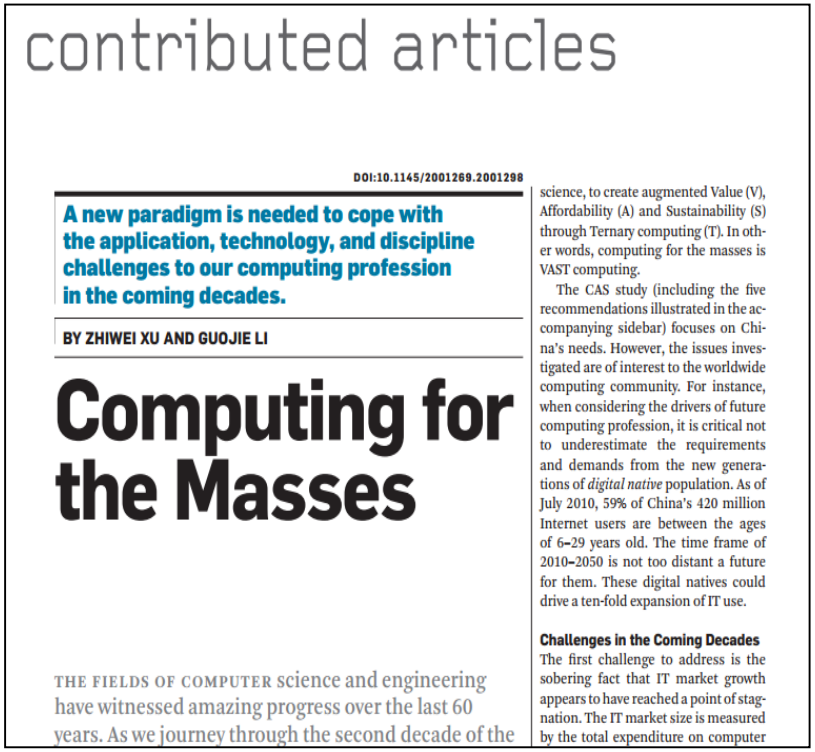
D： Domain knowledge， 领域知识

E： Ecosystem， 生态环境

发展大数据与人工[illegible]
要重视大众的刚性需求

- “与Computing for the Masses”的追求一样，我们要努力实现“Big Data for Masses，AI for the Masses”。不能只关注高端消费人群。
- 发展大数据与人工智能要重视大众的刚性需求（如健康、出行、安全等）不能只做“维生素”，要努力做不可替代的“抗生素”。

CACM 2011年第10期发表了徐志伟与我合写的文章：Computing for the Masses (为大众计算）



contributed articles

DOI:10.1145/2001269.2001298

A new paradigm is needed to cope with the application, technology, and discipline challenges to our computing profession in the coming decades.

BY ZHIWEI XU AND GUOJIE LI

Computing for the Masses

THE FIELDS OF COMPUTER science and engineering have witnessed amazing progress over the last 60 years. As we journey through the second decade of the

science, to create augmented Value (V), Affordability (A) and Sustainability (S) through Ternary computing (T). In other words, computing for the masses is VAST computing.

The CAS study (including the five recommendations illustrated in the accompanying sidebar) focuses on China's needs. However, the issues investigated are of interest to the worldwide computing community. For instance, when considering the drivers of future computing profession, it is critical not to underestimate the requirements and demands from the new generations of *digital native* population. As of July 2010, 59% of China's 420 million Internet users are between the ages of 6–29 years old. The time frame of 2010–2050 is not too distant a future for them. These digital natives could drive a ten-fold expansion of IT use.

Challenges in the Coming Decades

The first challenge to address is the sobering fact that IT market growth appears to have reached a point of stagnation. The IT market size is measured by the total expenditure on computer

满足大众刚性需求要有基础设施

- 满足大众的刚性需求必须有广覆盖的基础设施。工业时代的基础设施是“铁（路）、公（路）、机（场）”，信息时代的基础设施是（移动）互联网、云计算中心。智能化阶段的基础设施是大数据中心、机器学习训练平台等。
- 到了大数据和AI阶段，在IaaS、PaaS和SaaS基础上，又增加了MaaS：Management as a Service，和 AaaS: Analysis as a Service，大数据的存储、管理和分析成为新的基础设施，大数据催生Scalable AI也成为基础设施。
- 大容量、高并发、稀疏访问改变了传统以计算为核心的架构，基础资源与分析架构渐行渐远，分析模型越来越复杂，需要研制大数据分析处理专用内核系统和全链路大数据分析软件栈、分层开放编程语言，提升大数据分析技术的易用性和工程化能力。

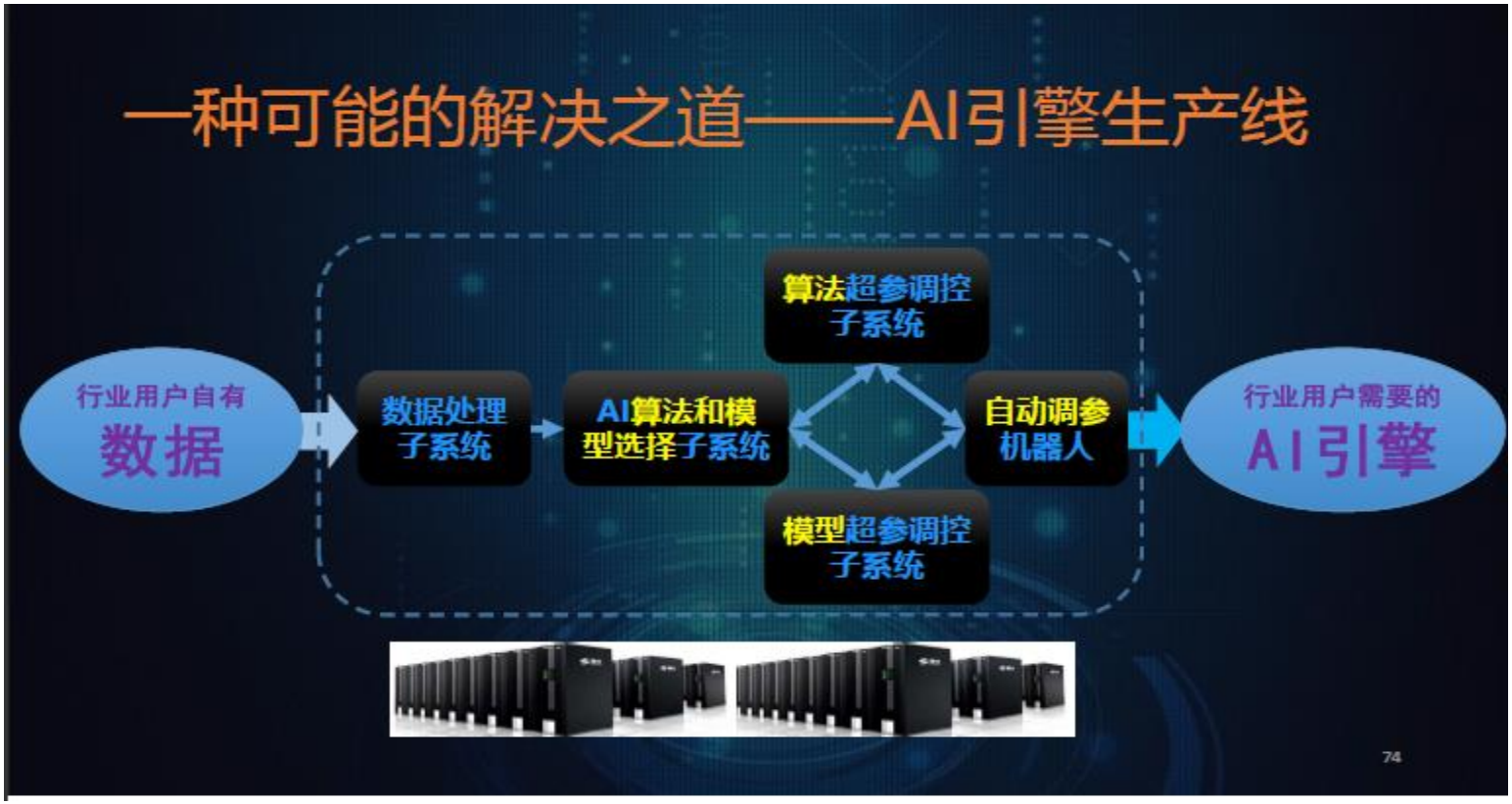
发展大数据和人工智能需要高度重视计算机系统结构和基础软件

- 中国人重“名”，“名不正则言不顺”。信息领域不断创造新名词，一旦“新名词（新学科）”上升为国家意志，原来的基础学科就被边缘化。现在以“系统结构”和“基础软件”申请国家项目，已很难拿到经费。
- 在2016年国家自然科学基金计算机学科的4863项申请项目中，计算机科学的基础理论只有16项，计算机体系结构22项，程序设计语言及支撑环境13项，高速数据传输技术2项；但是，计算机图像与视频处理有439项，模式识别理论及应用357项，人工智能应用258项。
- 构建大数据和AI基础设施离不开“系统结构”和“基础软件”，基础设施不能进入世界一流或者不能自主可控，智能应用必然像过去一样缺“芯”少“魂”。

国家《新一代人工智能发展规划》要加强基本理论研究和基础设施部署

- 国务院已经公布的国家《新一代人工智能发展规划》主要是面向人工智能应用的研究开发，涉及人工智能基本理论和基础设施的部署很少，在未来的实施中应高度重视。
- 数据科学应该是新一代人工智能发展规划的重要内容。数据科学包括用数据的方法研究科学和用科学的方法研究数据（鄂维南语）。后者主要是现在很红火的统计机器学习，需要数学家、计算机科学家和各领域的专家深度合作才有取得突破。
- 深度学习为什么这么有效，至今没有科学解释。最近以色列希伯来大学的Tishby 等科学家提出一种“信息瓶颈”理论，发现深度学习与“物理重整化（renormalization）是完全相同的过程，提出“学习最重要的部分是忘记”。我们应重视这一类的基础研究。

AI引擎生产线是必要的基础设施



- 未来5年内，需要新增至少1000倍数量的AI研发工程师（硕士和博士），现在需要硕士博士研发的AI技术，10年后将是高中生的课外作业！

——引自山世光的报告

人类大脑的形成和学习过程

- 出生时的脑是大数据学习(历代祖先)的结果！（体现在基因上）
- 出生后个体脑的发育是利用小数据和知识对出生时脑进行适应性修改的过程！
- 出生时的大脑神经连接是成人大脑的3倍以上，人的神经网络是在不断做“减法”。

重视计算智能，加强进化机理研究

- 人类的大数据学习体现在基因的“进化”上，当代人的学习过程对计算机的大数据学习并没有多大启发。要从动物和人类的进化中获取大数据学习的“经验”。人脑是进化出来的不是设计出来的，要理解大脑必须理解进化。
- 学术界普遍将人工智能的研究途径归纳为三类：符号主义（逻辑推理）、连接主义（统计机器学习）和行为主义（控制论学派）。虽然也有人将行为主义称为“进化主义”，但主要还是指以Brooks为代表感知–行动学派。
- 上世纪90年代初形成的“计算智能”学派不仅研究人工神经网络，还研究遗传算法、进化程序、混沌计算等，进化计算（或称演化计算）的主要优点是简单、通用、鲁棒性强和适于并行处理。实际上目前统计机器学习的崛起是“计算智能”的胜利，但学术界已遗忘了这个术语。我认为还是要重视计算智能，加强进化机理的研究。

我的看法：大数据（AI）= A+B+C+D+E

A： Algorithm， 算法

B： Basic Theory， Basis 基本理论，基础设施

C： Computing capability 计算能力

D： Domain knowledge， 领域知识

E： Ecosystem， 生态环境

领域知识决不可忽视

- 基于大数据的科研第四范式成为热门以后，“数据就是力量”大有取代“知识就是力量”之势。但许多教训提醒我们：领域知识决不可忽视。（微软的Tay对话机器人）
- 离散的数据背后可能有一个连续的模型，这个连续的模型需要深入掌握领域知识才能获得。
- 进化计算实质上是自适应的机器学习方法，它的核心思想是利用进化历史中获得的信息和知识指导搜索或计算。这些知识需要从领域专家获得。
- 与求解问题无关的通用人工智能算法往往效率不高，领域知识可以大大减少搜索范围，提高效率。
- 从事大数据和AI研究的学者与领域专家结合的深度将决定大数据和AI技术的进展速度。

希望在于对未知的认真探索

- 美国曼哈顿工程的负责人奥本海默在二战胜利以后说：“我们得到了一棵硕果累累的大树，并拼命地摇晃，结果得到了雷达和原子弹……，其全部精神实质在于对已知的疯狂而粗暴地掠夺，而毫无对未知的认真而谦恭地探索。”
- 经过60年培育，人工智能已长成硕果累累的大树，我们是拼命地摇晃这棵大树，在地上捡到一些零星的果实，还是怀抱对未知的认真和谦恭，自己新种几棵树苗。
- 大数据和AI的应用主要是企业的工作，大学与科研单位应重点做前瞻性、基础性研究。每一个大数据或AI系统都是非常具体的，“魔鬼存在于细节之中”。少在空洞的名词术语上费功夫，多做一些工匠、技师、工程师和科学家该做的朴实的事情。

大数据与AI任重道远

莫言下岭便无难，
赚得行人空喜欢。
正入万山圈子里，
一山放过一山拦。

——宋•杨万里《过松源晨炊漆公店》

请批评指正！