智能计算系统

第一章 概述

中国科学院计算技术研究所

陈云霁 研究员

cyj@ict.ac.cn

个人简介

- 陈云霁，男，1983年生，江西南昌人
- 主要经历
  - 1997 - 2002　本科生　中国科大少年班
  - 2002 - 2007　硕博生　中科院计算所
  - 2007 - 2012　助研、副研　中科院计算所
  - 2012 -　研究员　中科院计算所
- 从2008年起从事人工智能和芯片设计交叉研究
  - 曾获全国创新争先奖状、中国青年科技奖、国家基金委杰青、国家基金委优青、中组部万人计划青年拔尖人才、中国科学院青年科学家奖和中国计算机学会青年科学家奖等

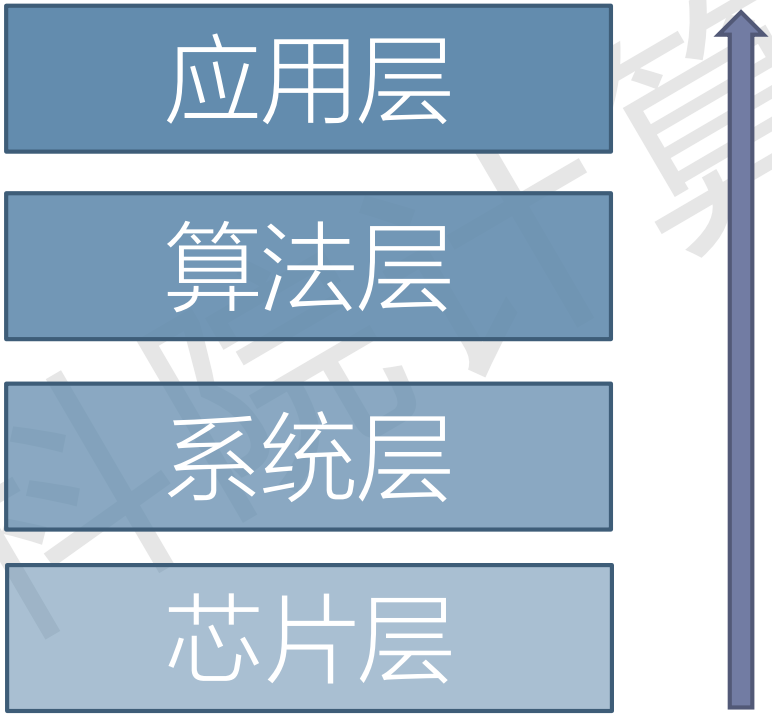
提纲

- 为什么要开这门课
- 为什么来上这门课
- 人工智能
- 智能计算系统
- 驱动范例

为什么要开这门课?

- 这不是理论力学课，世界上不存在这样一门课程
- 中科院的研究员并没有上课的义务
  - 主讲研究生《并行系统》（13-15）和本科生《数字电路》（16-17）
  - 担任研究生课《高性能计算机系统》（07-11）助教
- 巨大的科研工作压力
  - 思考、阅读和写作
  - 管理数百人的芯片研制团队
  - 承担几十项科研项目
  - 无休止的会议和评审

人工智能技术分层



人工智能底层科技的缺失可能使得我国智能产业成为空中楼阁

人工智能方向应该培养什么样的人才?

两个参考问题

人工智能方向应该培养人工智能（子）系统的设计者和研究者

力学、汽车试验学等方面工作的基本能力

- 计算机专业应该培养什么样的人才?
  - 计算机专业当培养计算机整机或子系统的设计者和研究者

对课程体系的建议

只包含各类机器学习算法、视听觉应用这条软件线，只能算是“人工智能应用专业”或者“人工智能算法专业”

- 谷歌有世界上最大的AI算法研究团队，然而
  - 谷歌董事长John Hennessy是计算机体系结构科学家，图灵奖得主
  - 谷歌AI的总领导者Jeff Dean是计算机系统研究者
  - 谷歌AI最令人瞩目的三个进展都是系统（Tensorflow、AlphaGo、TPU），而不仅仅是某个特定算法，算法只是系统的一个环节

应当包含系统线的课程，帮助学生理解系统到底是怎样执行的

对课程体系的建议

在高年级本科生（或者硕士研究生）阶段，应当设置一门系统类课程，能帮助同学实现对当前主流智能软硬件体系的融会贯通，具备自己动手完成一个完整智能系统的能力。这门课程就是智能计算系统

智能计算系统课程对学生的价值

- 全面的实践能力
  - 没有系统知识、只会调参，对整个系统的耗时、耗电毫无感觉，不具备把一个算法在实际系统上部署起来的能力的学生做不出真东西
  - 会用Tensorflow赚20万人民币，会设计Tensorflow赚20万美元
- 更强的研究能力
  - 能够从更广阔的的视野和维度开展研究，不只是盯着精度
  - 形成系统思维，拥有科研道路更广阔的舞台

智能计算系统课程对教师的价值

- 《礼记·学记》“教学相长”、开阔思路
- 美国计算机方向Top4高校Stanford、CMU、UC Berkley和MIT以及多个国际单位联合发布了白皮书——“SysML: The New Frontier of Machine Learning Systems”
  - 包括Yann LeCun、Michael I. Jordan、Bill Dally和Jeff Dean等
- 培养教授智能计算系统课程的教师，能抢占这一国际热点方向、也是未来重要学科增长点的先机

什么是智能计算系统?

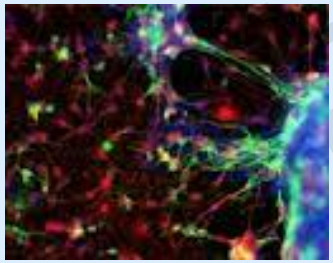
智能计算系统是智能的物质载体

现阶段的智能计算系统通常是集成CPU和智能芯片的异构系统，软件上通常包括一套面向开发者的智能计算编程环境（包括编程框架和编程语言）

智能计算系统的形态

超级计算机

商业分析



药物研制

数据中心

Baidu 百度

YouTube

广告推荐

自动翻译

智能手机

语音识别

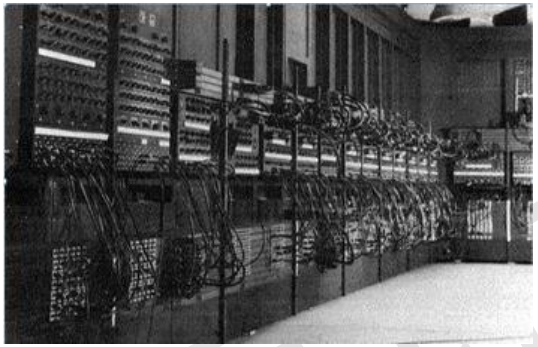
图像分析

嵌入式设备

机器人

消费类电子

智能计算系统具有重大价值



上世纪人类从工业时代过渡到信息时代

现在已经发展到向智能时代进化的拐点

中国需要一大批智能计算系统的开发者和设计者

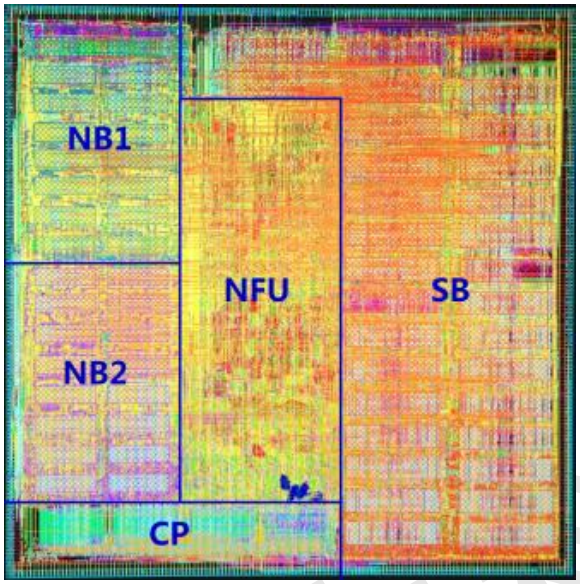
智能计算系统课程的三大实际困难

- 没有参考课程
- 没有现成师资
- 没有成熟教材

开创深度学习处理器方向

- **2007：开始智能和芯片交叉研究**
  - **长期受到科学院自有项目的支持**
- 2013：国际首个深度学习处理器架构
  - CCF A类会议ASPLOS'14最佳论文
  - 亚洲首获体系结构四大顶会最佳论文
- 2014：国际首个多核深度学习处理器架构
  - CCF A类会议MICRO'14最佳论文
- 2015：国际首个通用机器学习处理器
  - CCF A类会议ASPLOS'15
- 2016：国际首个智能指令集
  - CCF A类会议ISCA'16最高分论文
- 2017：国际首个集成AI处理器的手机
  - 华为Mate10

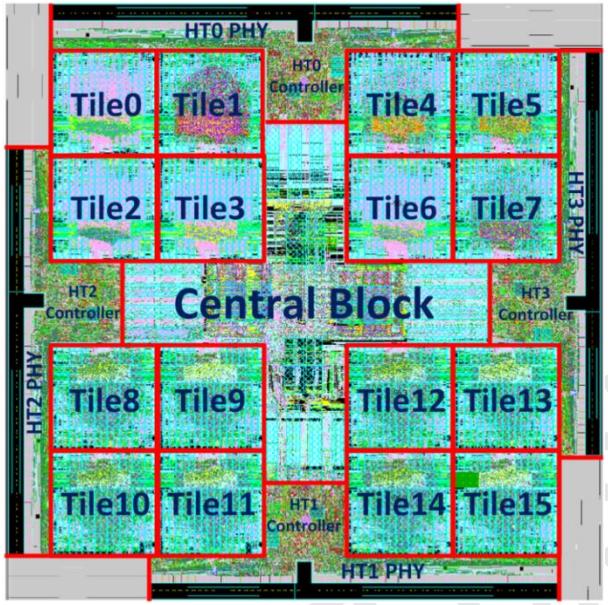
开创深度学习处理器方向



1GHz，0.485W @ 65nm，通用CPU 1/10的面积，100倍的性能

- 2007：开始智能和芯片交叉研究
  - 长期受到科学院自有项目的支持
- **2013：国际首个深度学习处理器架构**
  - **CCF A类会议ASPLOS’14最佳论文**
  - **亚洲首获体系结构四大顶会最佳论文**
- 2014：国际首个多核深度学习处理器架构
  - CCF A类会议MICRO’14最佳论文
- 2015：国际首个通用机器学习处理器
  - CCF A类会议ASPLOS’15
- 2016：国际首个智能指令集
  - CCF A类会议ISCA'16最高分论文
- 2017：国际首个集成AI处理器的手机
  - 华为Mate10

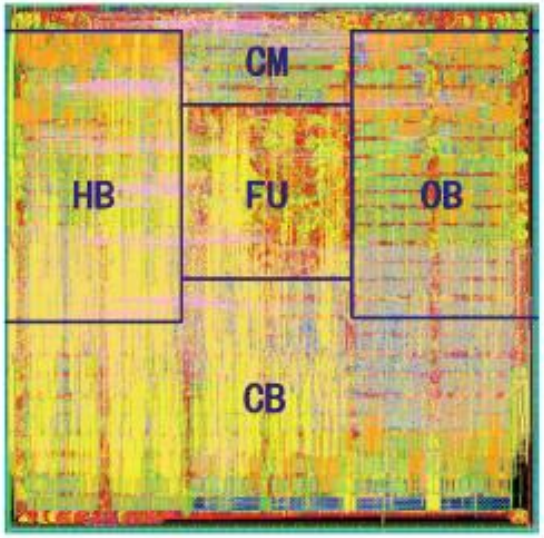
开创深度学习处理器方向



0.6GHz，16W@28nm，主流GPU 21倍性能，300倍性能功耗比

- 2007：开始智能和芯片交叉研究
  - 长期受到科学院自有项目的支持
- 2013：国际首个深度学习处理器架构
  - CCF A类会议ASPLOS'14最佳论文
  - 亚洲首获体系结构四大顶会最佳论文
- **2014：国际首个多核深度学习处理器架构**
  - **CCF A类会议MICRO'14最佳论文**
- 2015：国际首个通用机器学习处理器
  - CCF A类会议ASPLOS'15
- 2016：国际首个智能指令集
  - CCF A类会议ISCA'16最高分论文
- 2017：国际首个集成AI处理器的手机
  - 华为Mate10

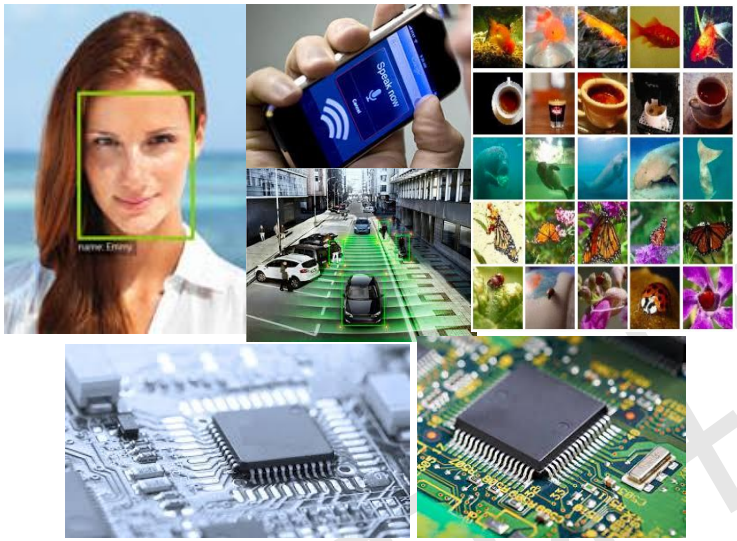
开创深度学习处理器方向



- 除人工神经网络，还支持k-NN、SVM、Bayes等其它主流ML方法
- GPU 1/100面积功耗，相当的性能

- 2007：开始智能和芯片交叉研究
  - 长期受到科学院自有项目的支持
- 2013：国际首个深度学习处理器架构
  - CCF A类会议ASPLOS'14最佳论文
  - 亚洲首获体系结构四大顶会最佳论文
- 2014：国际首个多核深度学习处理器架构
  - CCF A类会议MICRO'14最佳论文
- **2015：国际首个通用机器学习处理器**
  - **CCF A类会议ASPLOS'15**
- 2016：国际首个智能指令集
  - CCF A类会议ISCA'16最高分论文
- 2017：国际首个集成AI处理器的手机
  - 华为Mate10

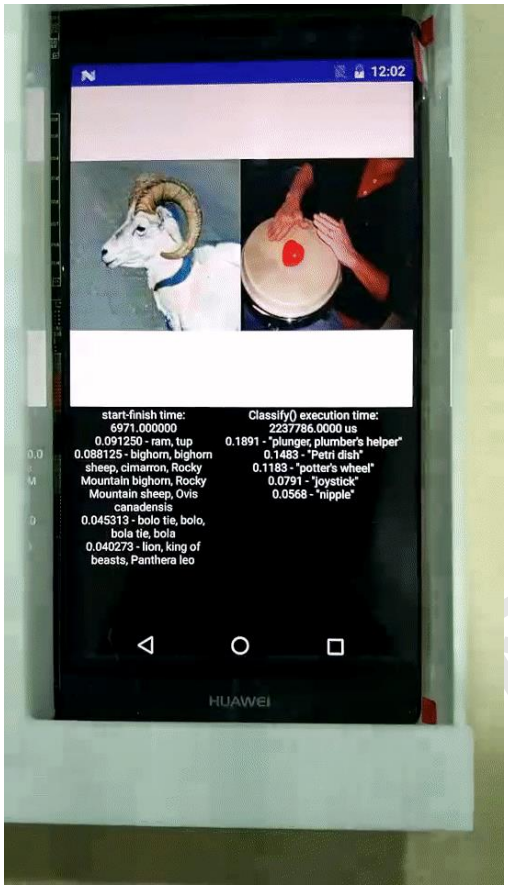
开创深度学习处理器方向



Cambricon指令集

- 2007：开始智能和芯片交叉研究
  - 长期受到科学院自有项目的支持
- 2013：国际首个深度学习处理器架构
  - CCF A类会议ASPLOS'14最佳论文
  - 亚洲首获体系结构四大顶会最佳论文
- 2014：国际首个多核深度学习处理器架构
  - CCF A类会议MICRO'14最佳论文
- 2015：国际首个通用机器学习处理器
  - CCF A类会议ASPLOS'15
- **2016：国际首个智能指令集**
  - **CCF A类会议ISCA'16最高分论文**
- 2017：国际首个集成AI处理器的手机
  - 华为Mate10

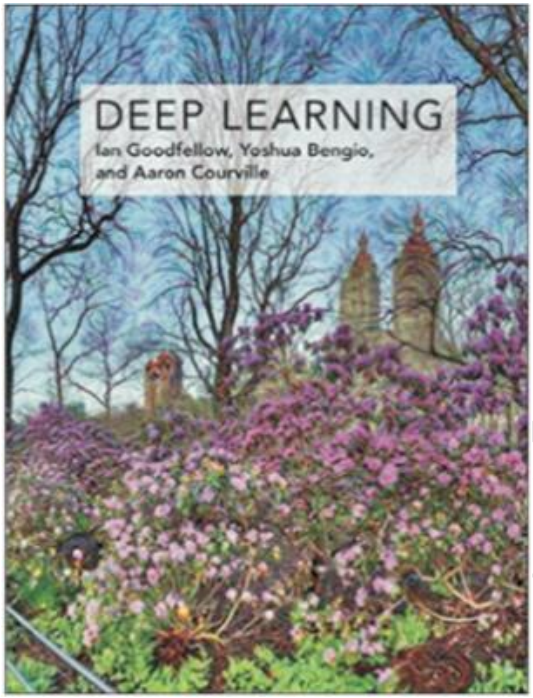
开创深度学习处理器方向



已有十余种智能手机（共近亿台）集成寒武纪，实测能效提升一个数量级

- 2007：开始智能和芯片交叉研究
  - 长期受到科学院自有项目的支持
- 2013：国际首个深度学习处理器架构
  - CCF A类会议ASPLOS'14最佳论文
  - 亚洲首获体系结构四大顶会最佳论文
- 2014：国际首个多核深度学习处理器架构
  - CCF A类会议MICRO'14最佳论文
- 2015：国际首个通用机器学习处理器
  - CCF A类会议ASPLOS'15
- 2016：国际首个智能指令集
  - CCF A类会议ISCA'16最高分论文
- **2017：国际首个集成AI处理器的手机**
  - **华为Mate10**

寒武纪的国际影响



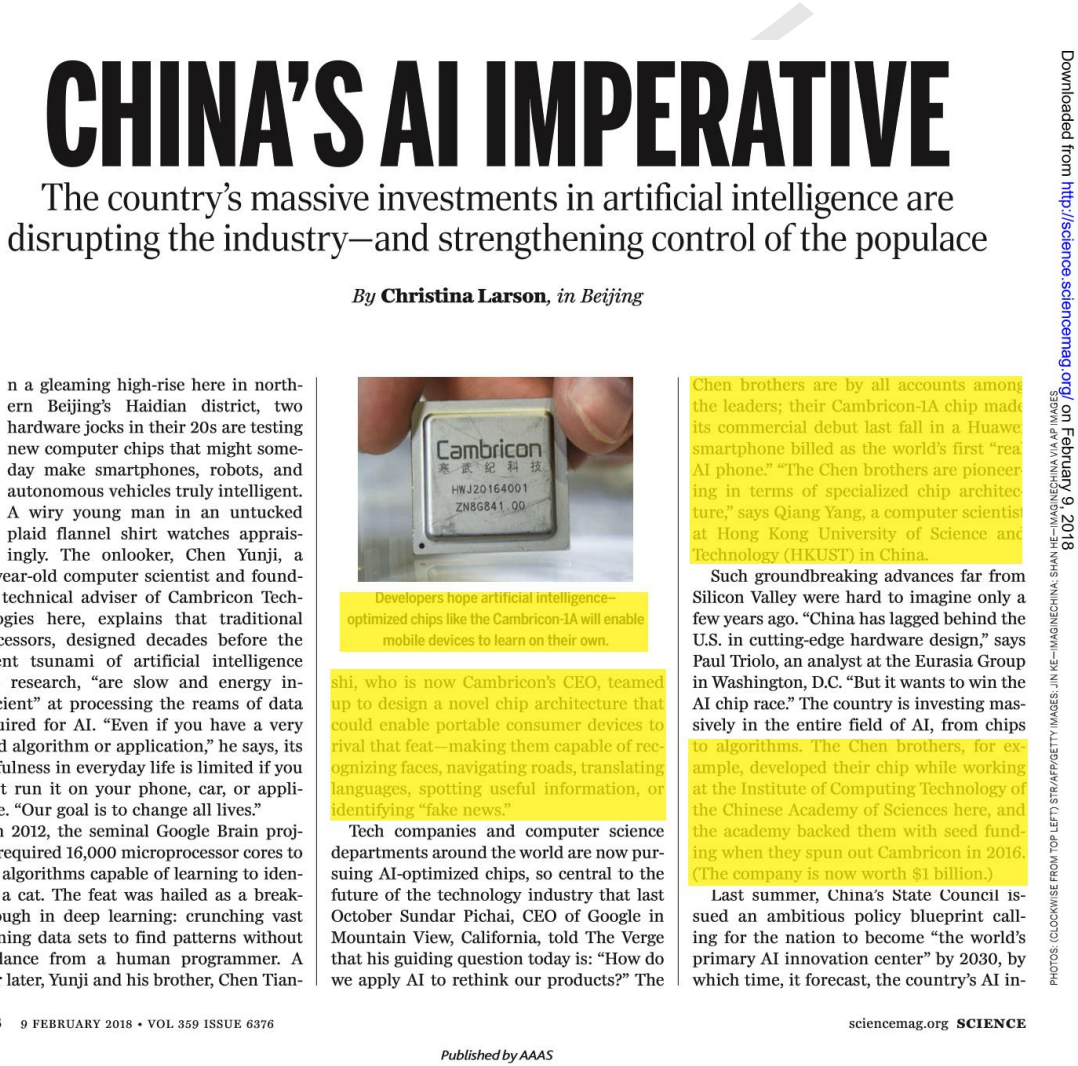
十年坚持，研制国际首个深度学习处理器,推动深度学习处理器成为计算机体系结构领域主要热点之一

- 2016-2018三年CCF A类会议ISCA共近1/4论文引用
- 被深度学习开创者、图灵奖获得者Bengio的深度学习教科书引用
- 两百个国际机构（哈佛、斯坦福、 MIT、CMU、普林斯顿、UCLA、Intel、谷歌、微软等）跟踪
- 数十位国际知名学者（2位图灵奖得主、10余位中美院士、41位ACM会士、120位IEEE会士）引用

寒武纪的国际影响

2018年2月9日**Science**杂志 (vol.358, iss. 6376) 高度评价寒武纪的研究

- AI芯片的引领者(the leaders)
- 专用芯片体系结构的先驱 (pioneering in terms of specialized chip architecture)
- 开创性进展 (groundbreaking advances)



CHINA'S AI IMPERATIVE

The country's massive investments in artificial intelligence are disrupting the industry—and strengthening control of the populace

By **Christina Larson**, *in Beijing*

In a gleaming high-rise here in northern Beijing's Haidian district, two hardware jocks in their 20s are testing new computer chips that might someday make smartphones, robots, and autonomous vehicles truly intelligent. A wiry young man in an untucked plaid flannel shirt watches appraisingly. The onlooker, Chen Yunji, a 34-year-old computer scientist and founding technical adviser of Cambricon Technologies here, explains that traditional processors, designed decades before the recent tsunami of artificial intelligence (AI) research, "are slow and energy inefficient" at processing the reams of data required for AI. "Even if you have a very good algorithm or application," he says, its usefulness in everyday life is limited if you can't run it on your phone, car, or appliance. "Our goal is to change all lives."

In 2012, the seminal Google Brain project required 16,000 microprocessor cores to run algorithms capable of learning to identify a cat. The feat was hailed as a breakthrough in deep learning: crunching vast training data sets to find patterns without guidance from a human programmer. A year later, Yunji and his brother, Chen Tianshi, who is now Cambricon's CEO, teamed up to design a novel chip architecture that could enable portable consumer devices to rival that feat—making them capable of recognizing faces, navigating roads, translating languages, spotting useful information, or identifying "fake news."

Developers hope artificial intelligence–optimized chips like the Cambricon-1A will enable mobile devices to learn on their own.

Tech companies and computer science departments around the world are now pursuing AI-optimized chips, so central to the future of the technology industry that last October Sundar Pichai, CEO of Google in Mountain View, California, told The Verge that his guiding question today is: "How do we apply AI to rethink our products?" The Chen brothers are by all accounts among the leaders; their Cambricon-1A chip made its commercial debut last fall in a Huawei smartphone billed as the world's first "real AI phone." "The Chen brothers are pioneering in terms of specialized chip architecture," says Qiang Yang, a computer scientist at Hong Kong University of Science and Technology (HKUST) in China.

Such groundbreaking advances far from Silicon Valley were hard to imagine only a few years ago. "China has lagged behind the U.S. in cutting-edge hardware design," says Paul Triolo, an analyst at the Eurasia Group in Washington, D.C. "But it wants to win the AI chip race." The country is investing massively in the entire field of AI, from chips to algorithms. The Chen brothers, for example, developed their chip while working at the Institute of Computing Technology of the Chinese Academy of Sciences here, and the academy backed them with seed funding when they spun out Cambricon in 2016. (The company is now worth $1 billion.)

Last summer, China's State Council issued an ambitious policy blueprint calling for the nation to become "the world's primary AI innovation center" by 2030, by which time, it forecast, the country's AI in-

PHOTOS: (CLOCKWISE FROM TOP LEFT) STR/AFP/GETTY IMAGES; JIN KE—IMAGINECHINA VIA AP IMAGES

Published by AAAS

寒武纪的国内影响

- 孵化寒武纪公司，国际上首个智能芯片独角兽创业公司
  - 16年上半年成立，天使轮估值近1亿美元，融资近0.1亿美元
  - 17年上半年A轮估值近10亿美元，融资近1亿美元
  - 18年上半年B轮估值超过25亿美元，融资4亿美元
  - 19年上半年B+轮估值超过35亿美元，融资4亿美元
  - 投资人包括国投、国风投、国新、中金、中信、TCL、阿里、联想、讯飞元禾、中科院投资、图灵等
- 目前超过1000人

我们的尝试

- 2018年，我们在中国科学院大学计算机学院申请开设一门人工智能方向的系统课程，名为《智能计算系统》
- 面向人群：计算机、软件工程和人工智能方向研究生和高年级本科生
- 课程目的：希望能培养同学对智能计算完整软硬件技术栈（包括基础智能算法、智能计算编程框架、智能计算编程语言、智能芯片体系结构等）融会贯通的理解。
- 前置课程要求：C/C++，计算机组成原理，算法导论（机器学习）

智能计算系统课程开设情况

- 已经/正在在中科院大学、北大、中国科大、天大、北航、南开、北理工、武大等十多家高校开设智能计算系统课程
  - 部分院校设为研究生课程，部分院校设为本科生课程
  - 涉及计算机学院、人工智能学院和软件学院
- 我们希望能帮助更多高校开设这门课程
  - 课程ppt、讲稿、录像、代码都将公开，提供云实验环境和实验小卡
  - 2019年8月教指委组织了第一次智能计算系统课程导教班（西安），2020年2月即将组织第二次（广州）
  - 《智能计算系统》教材正在出版印刷中，2020年2月即可交付读者

提纲

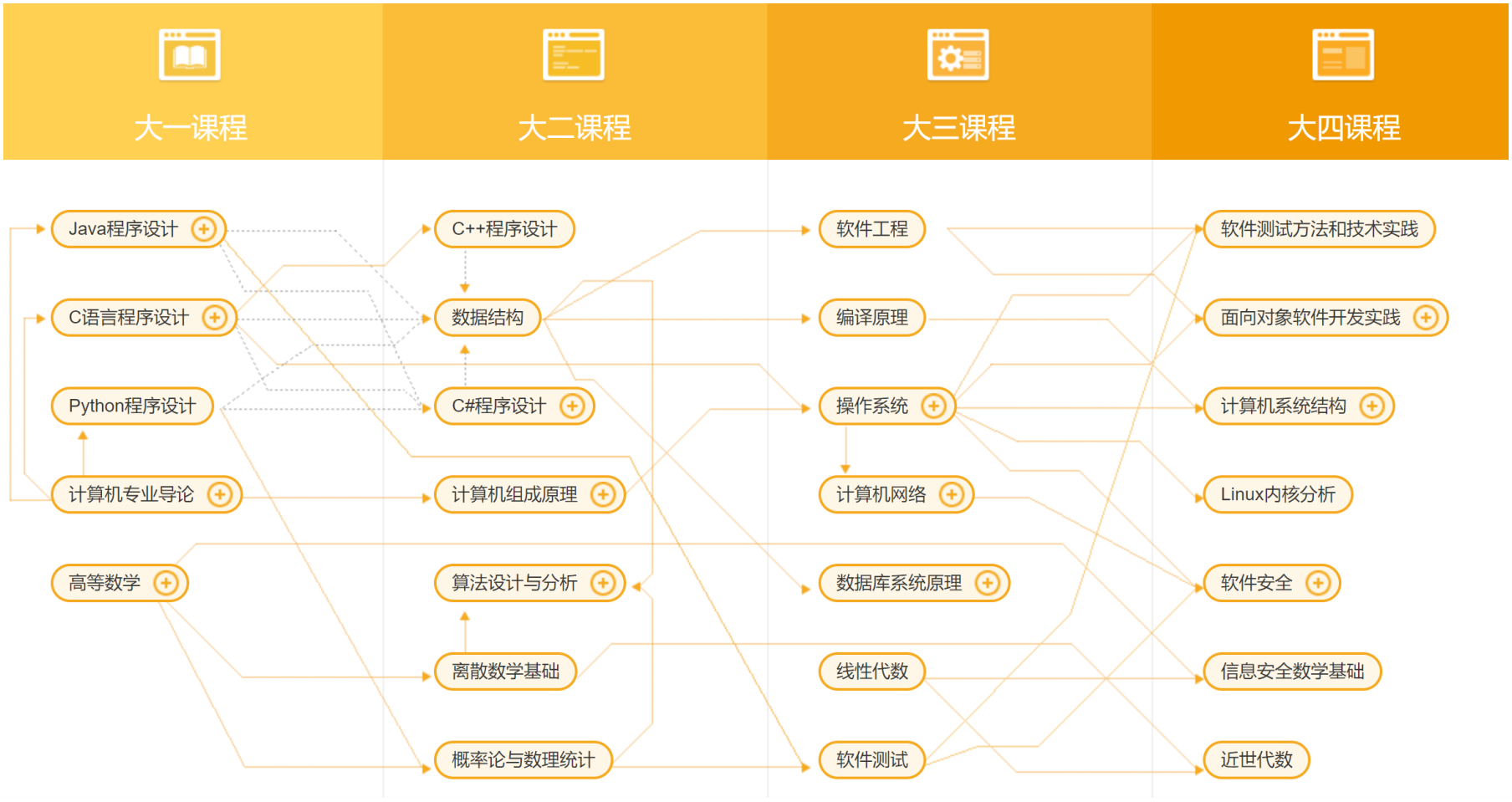
- 为什么要开这门课
- 为什么来上这门课
- 人工智能
- 智能计算系统
- 驱动范例

为什么来上这门课?

好学生就是要上课

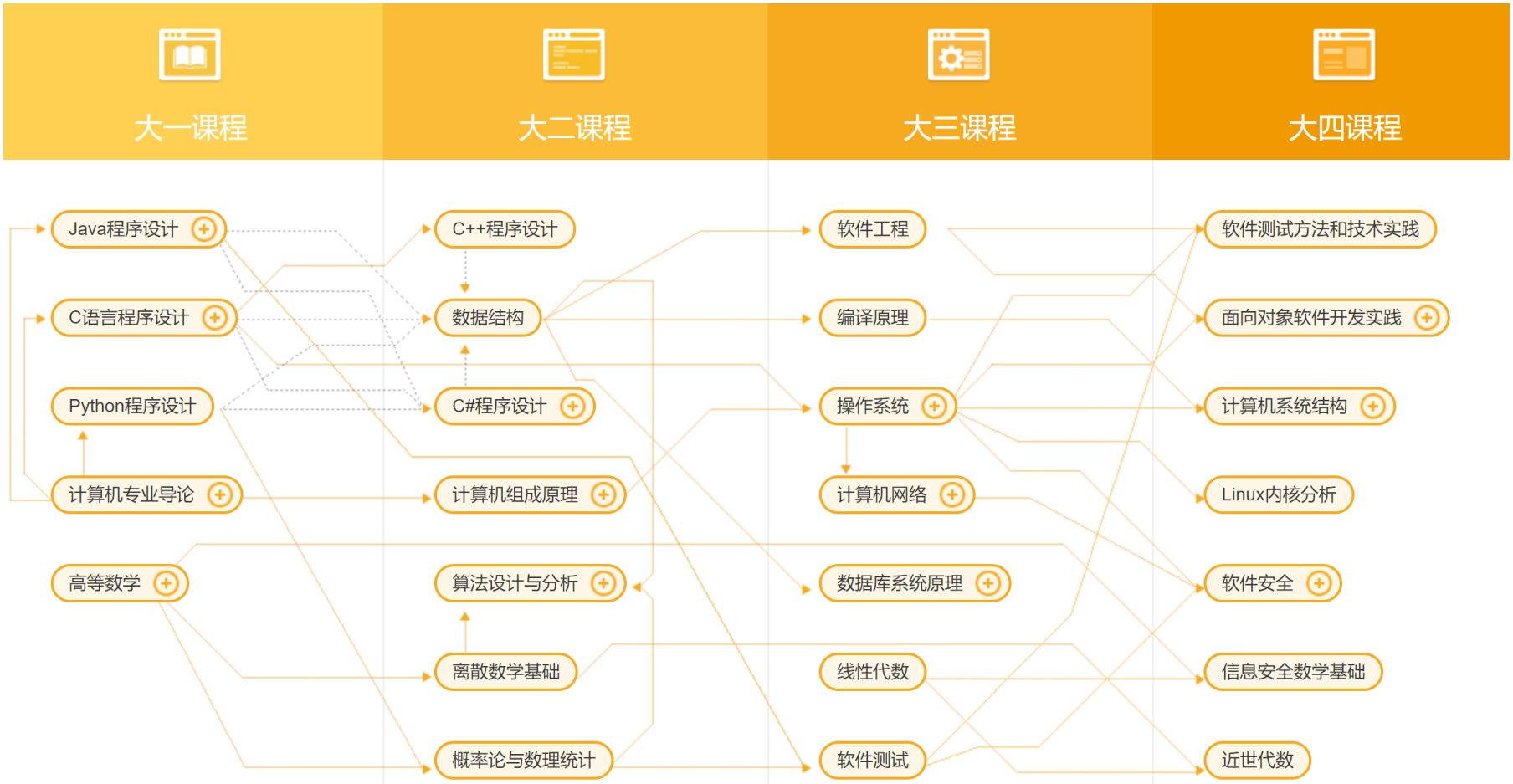
具备实际解决问题能力是上课的目标

计算机专业培养计划的正面启示



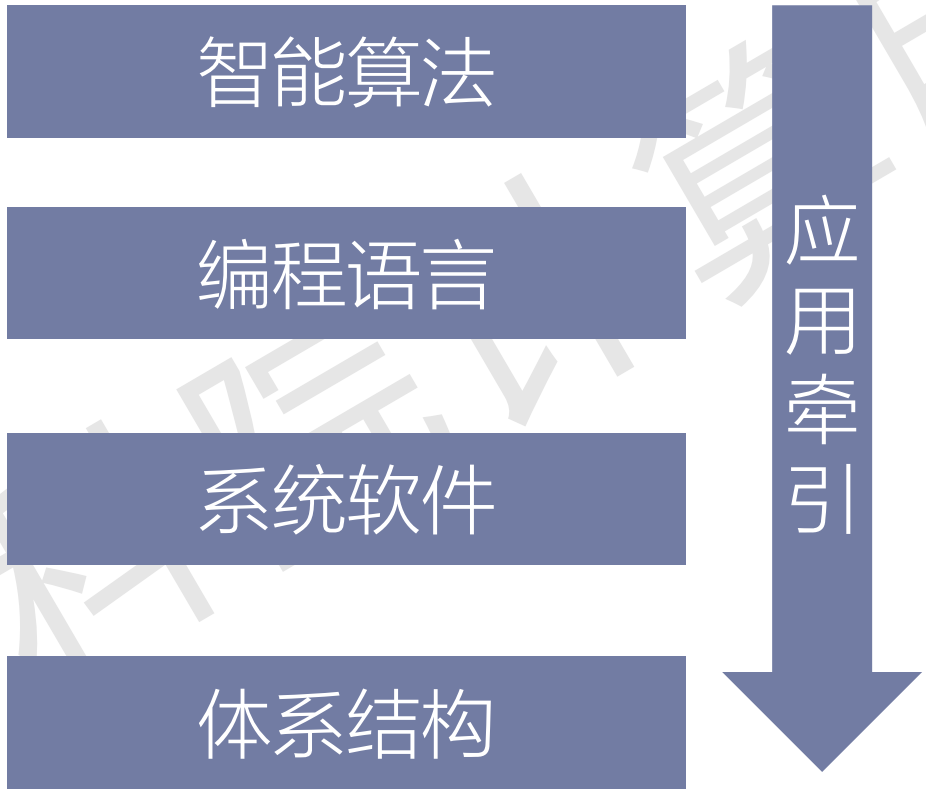
我国各高校计算机专业培养计划都包含计算机组成原理、操作系统、编译原理、计算机体系结构等硬件系统类课程

计算机专业培养计划的负面启示



课程条块分割，学生不能融会贯通做出一个完整系统，导致我国信息产业全栈式人才缺乏，核心硬科技竞争力缺失

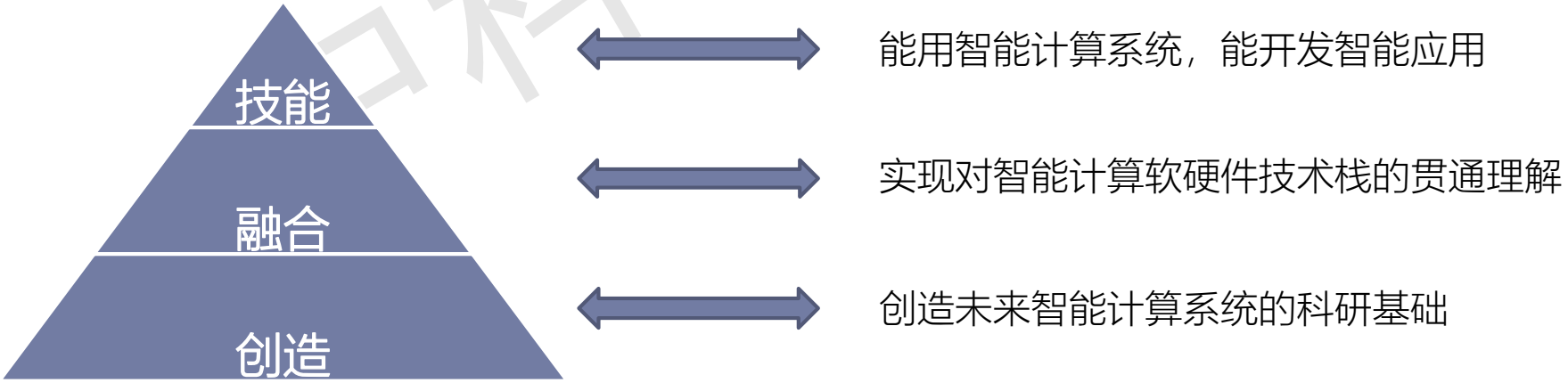
应用驱动、全栈贯通的课程体系



一门帮助学生学以致用、形成全局系统观的工科课程

课程目标和目的

- 中国需要一大批智能基础设施的开发者和设计者
- 专业普及课程
  - 应用驱动，全栈贯通
- 智能计算系统
  - 建立智能计算系统设计及应用的知识体系
  - 掌握智能应用开发的基本技能
  - 培养开展智能计算系统基础研究的兴趣和能力



课程要求

- C/C++
- 计算机体系结构
- 机器学习/算法导论

课程考核

- 不点名、不布置作业、不抓人，来去自由
- 考核理念
  - 想要知识的同学收获知识
  - 想要学分的同学收获学分
- 总分100分
  - 期末开卷考试40分
  - 实验一30分
  - 实验二30分

课程提纲

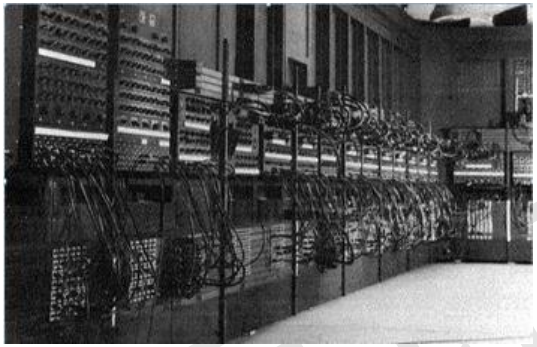
- 第一章　概述---a driving example
- 第二章　神经网络
- 第三章　深度学习
- 第四章　编程框架使用
- 第五章　编程框架机理
- 第六章　深度学习处理器原理
- 第七章　深度学习处理器架构
- 第八章　智能编程语言
- 第九章　实验讲解

提纲

- 为什么要开这门课
- 为什么来上这门课
- 人工智能
- 智能计算系统
- 驱动范例

智能时代

蒸汽机



集成电路

智能计算系统

上世纪人类从工业时代过渡到信息时代

现在已经发展到向智能时代进化的拐点

国家战略

"AI holds the potential to be a major driver of economic growth and social progress" [White House report, 2016]

Released domestic strategic plan to become world leader in AI by 2030 [2017]

"Whoever becomes the leader in this sphere [AI] will become the ruler of the world" [Putin, 2017]

企业投入

Google "An important shift from a mobile first world to an AI first world" [CEO Sundar Pichai @ Google I/O 2017]

Microsoft Created AI and Research group as 4th engineering division, now 8K people [2016]

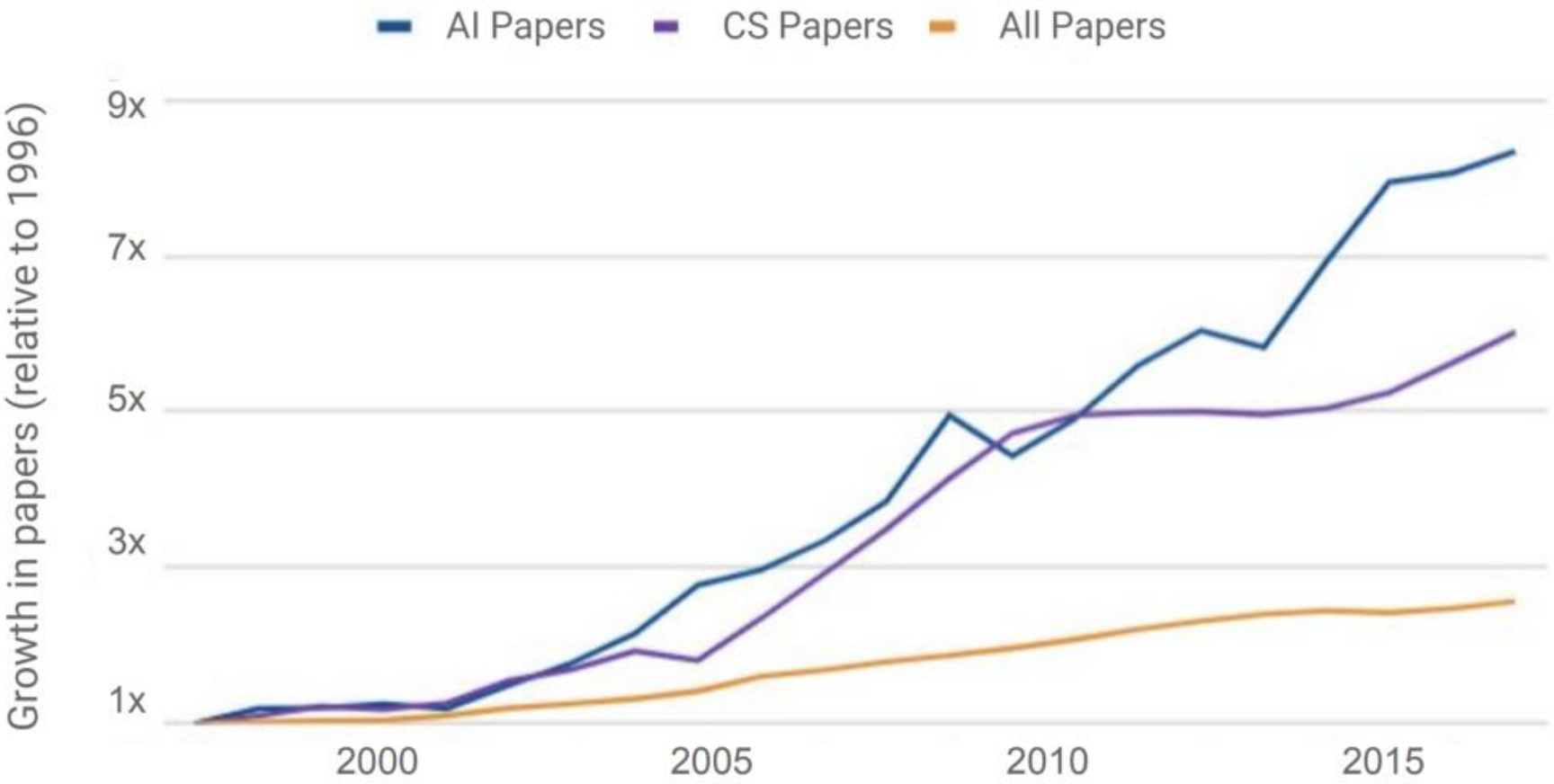
facebook Created Facebook AI Research, Mark Zuckerberg very optimistic and invested

Baidu 百度 百度将All in AI，我们在AI时代的核心战略就是开放赋能，我们的将来必须建立在与每个开发者共赢的基础上。[前COO陆奇@百度开发者大会2017]

Tencent 腾讯 AI in All，AI技术...能够真正和各行各业实际应用结合在一起，从而让AI新技术能够得到实际价值的发挥。[COO任宇昕@腾讯全球合作伙伴大会2017]

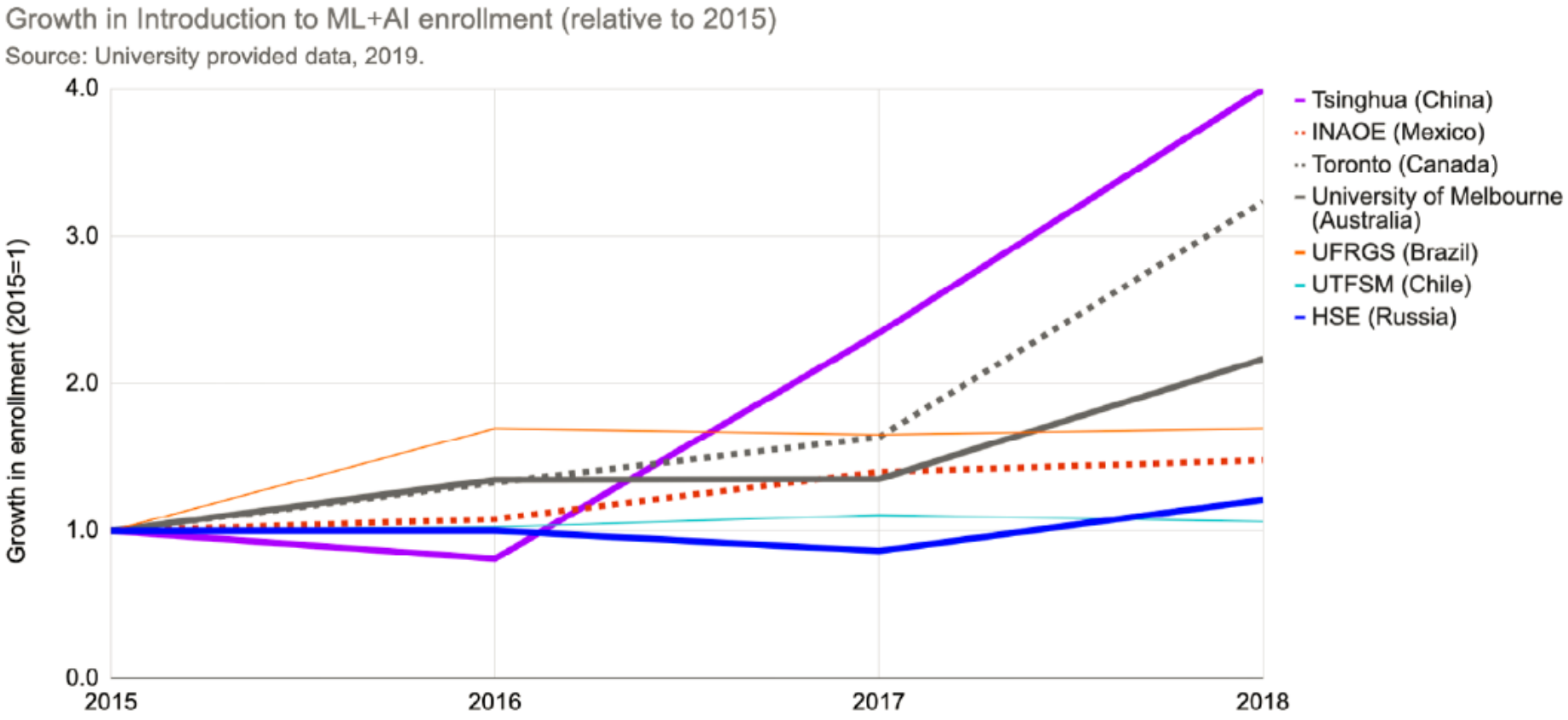
Alibaba Group 阿里巴巴集团 達摩院 未来3年阿里巴巴在技术研发上的投入将超过1000亿人民币.[马云@2017杭州·云栖大会]

研究趋势

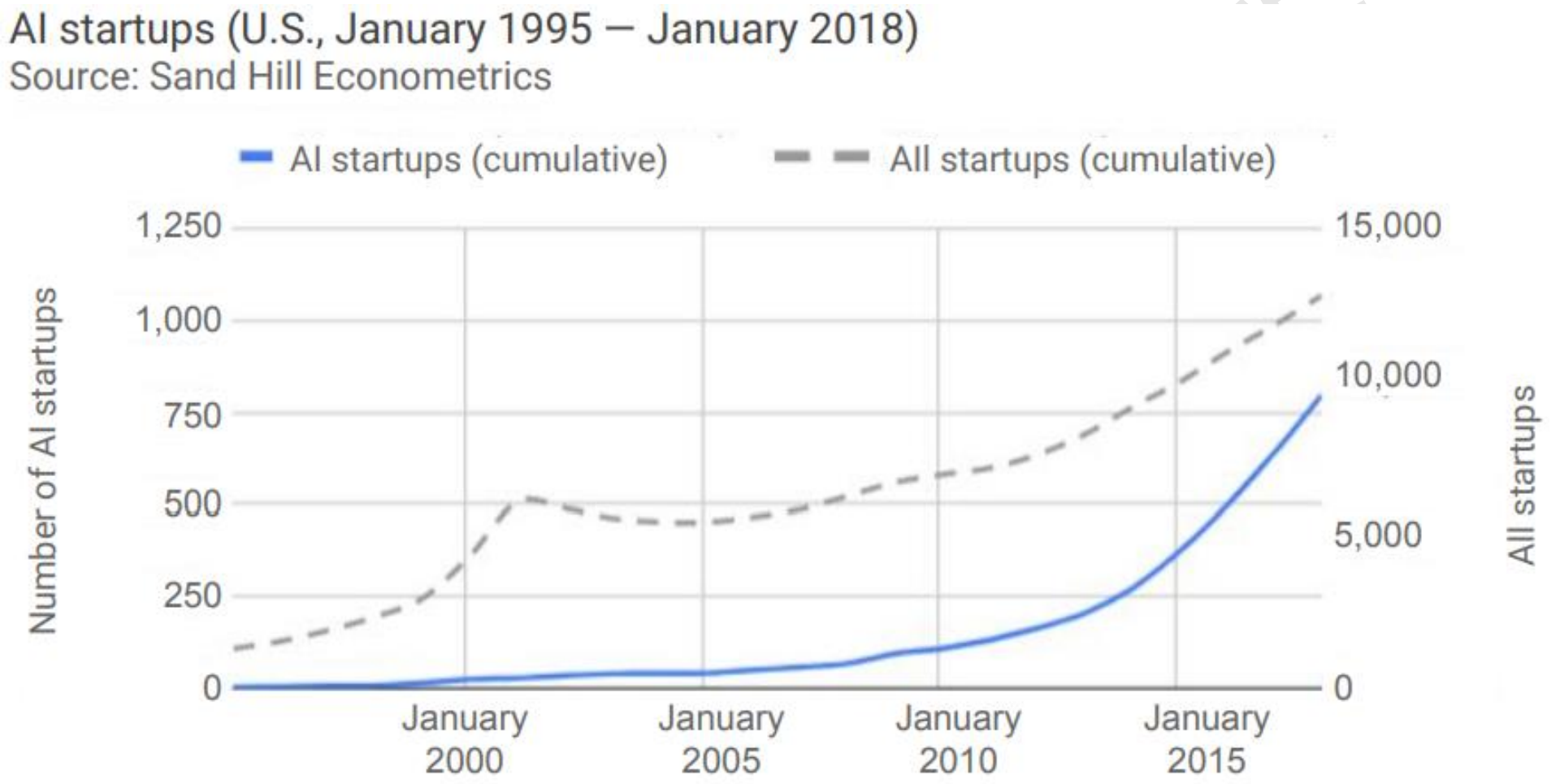


1996至2017，AI相关论文数量增长比例大大超过计算机学科和所有学科

高等教育


Growth in Introduction to ML+AI enrollment (relative to 2015)
Source: University provided data, 2019.
Growth in enrollment (2015=1)
4.0
3.0
2.0
1.0
0.0
2015
2016
2017
2018
Tsinghua (China)
INAOE (Mexico)
Toronto (Canada)
University of Melbourne (Australia)
UFRGS (Brazil)
UTFSM (Chile)
HSE (Russia)

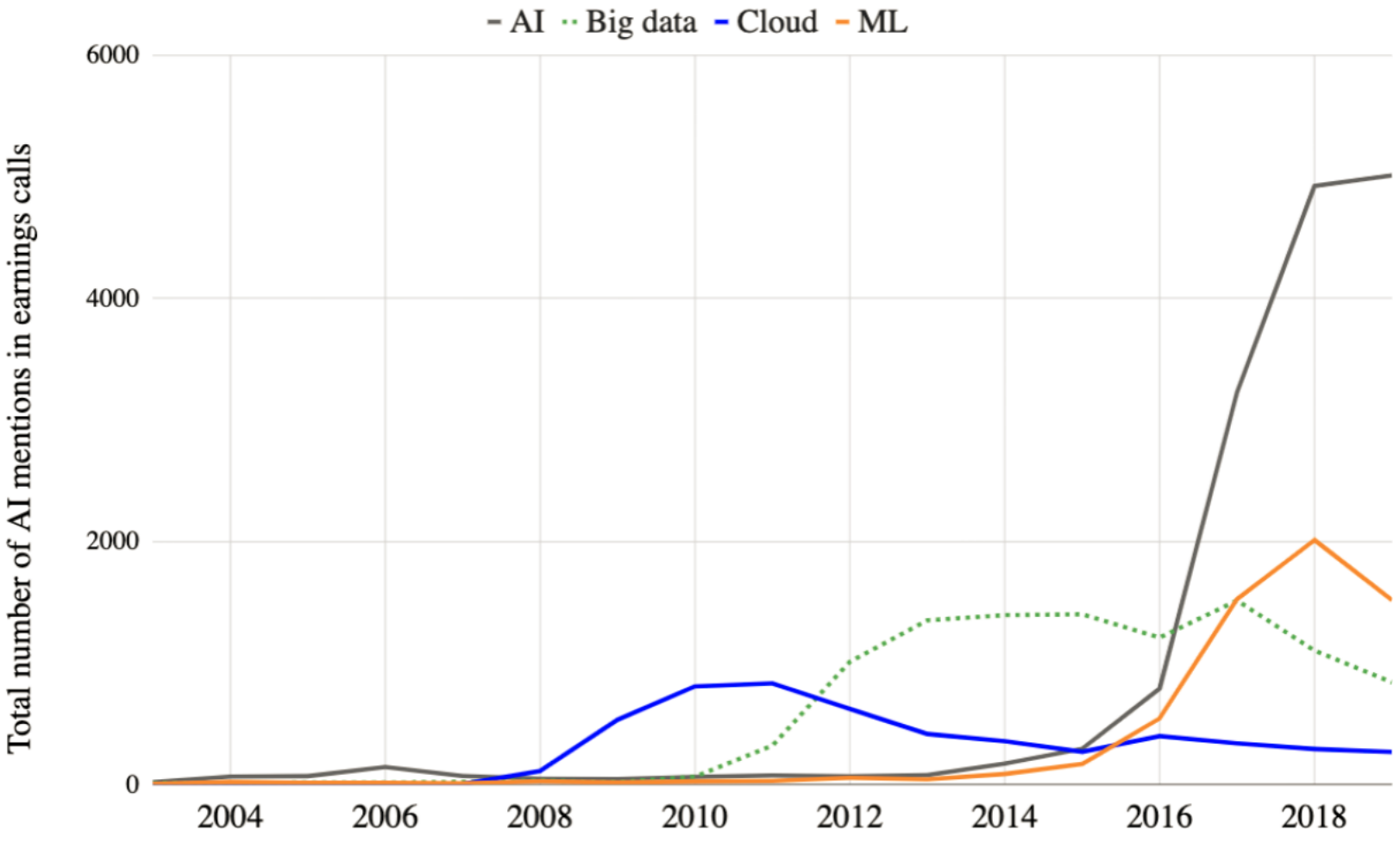
初创企业


AI startups (U.S., January 1995 – January 2018)
Source: Sand Hill Econometrics
AI startups (cumulative)
All startups (cumulative)
Number of AI startups
All startups
1,250
1,000
750
500
250
0
15,000
10,000
5,000
0
January 2000
January 2005
January 2010
January 2015

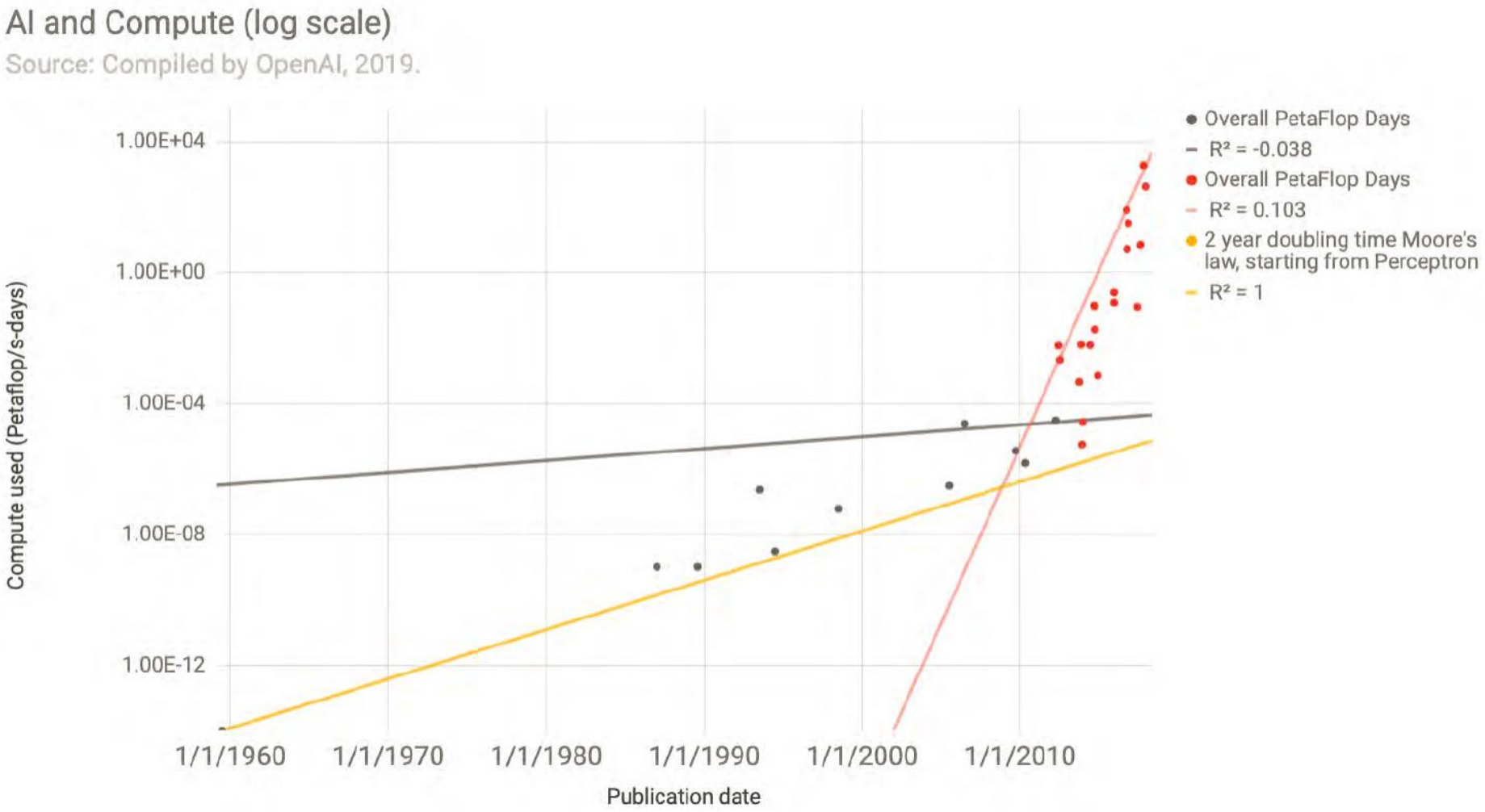
就业机会

Total Number of AI mentions in earnings calls

Source: Prattle, 2019.

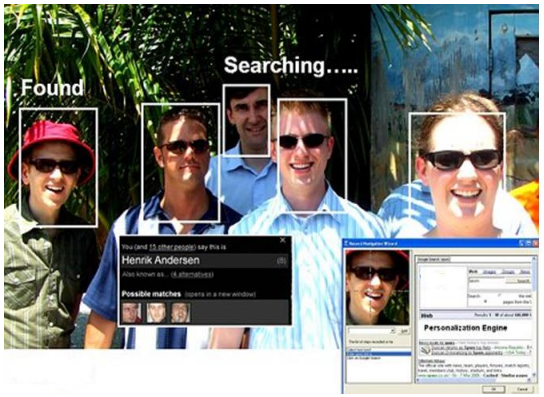


算力需求

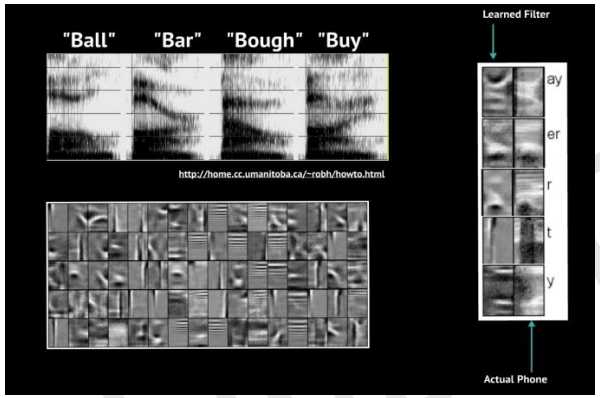


2012年之前，AI训练算力需求每两年翻一番；2012年之后，算力需求每3.4个月翻一番

人工智能不断飞速发展



lfw人脸测试准确度99%
(成人仅97%)



语音速记战胜
人类专业速记员

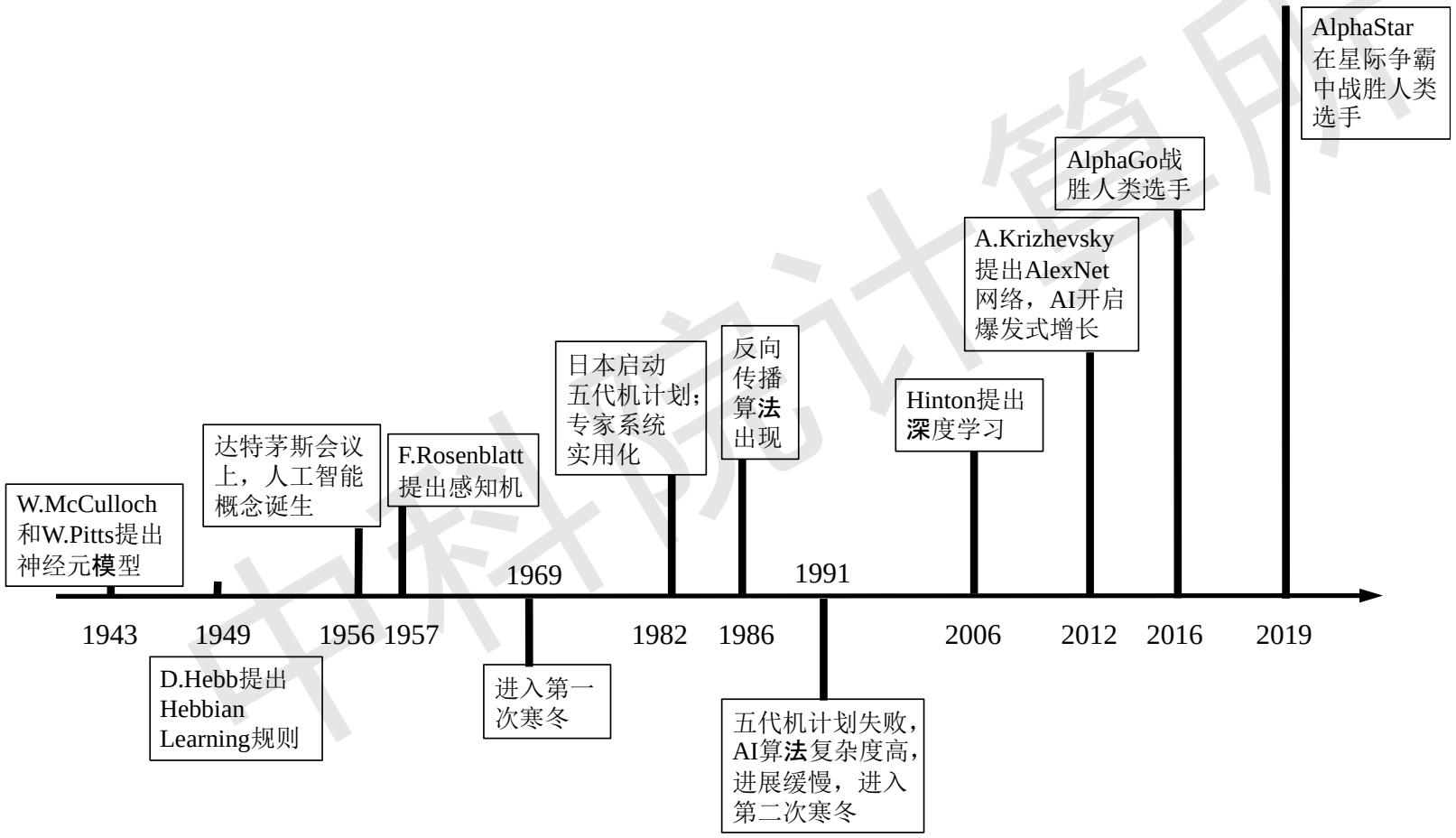
AlphaGo战胜李世石

人工智能算法在多种应用上接近或超过了人类水平

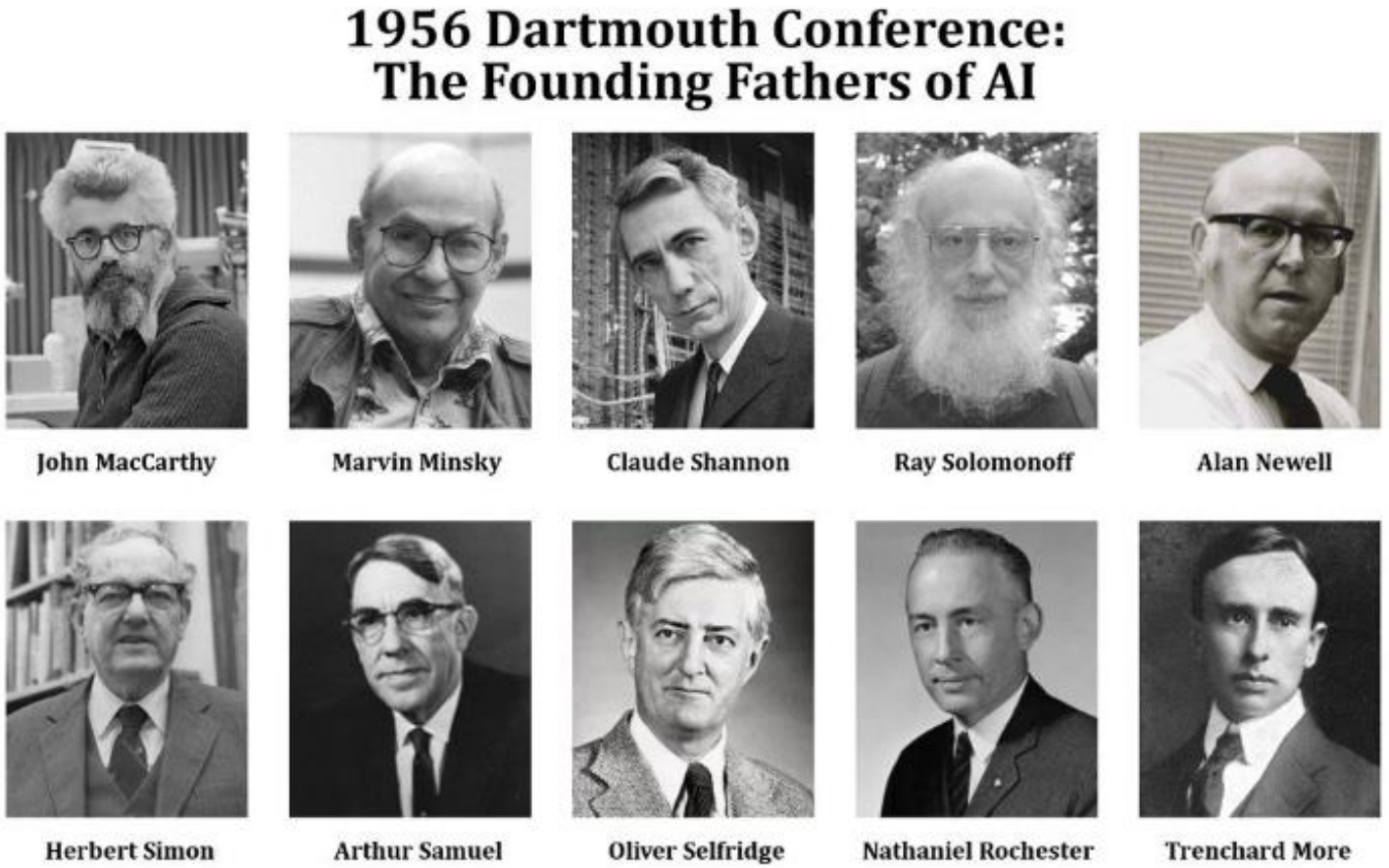
什么是人工智能

- 人工智能：人制造出来的机器所表现出来的智能
- 强人工智能或通用人工智能：具备与人类同等智慧、或超越人类的人工智能，能表现正常人类所具有的所有智能行为
- 弱人工智能：能完成某种特定具体任务的人工智能，计算机科学的非平凡应用

人工智能的三次热潮

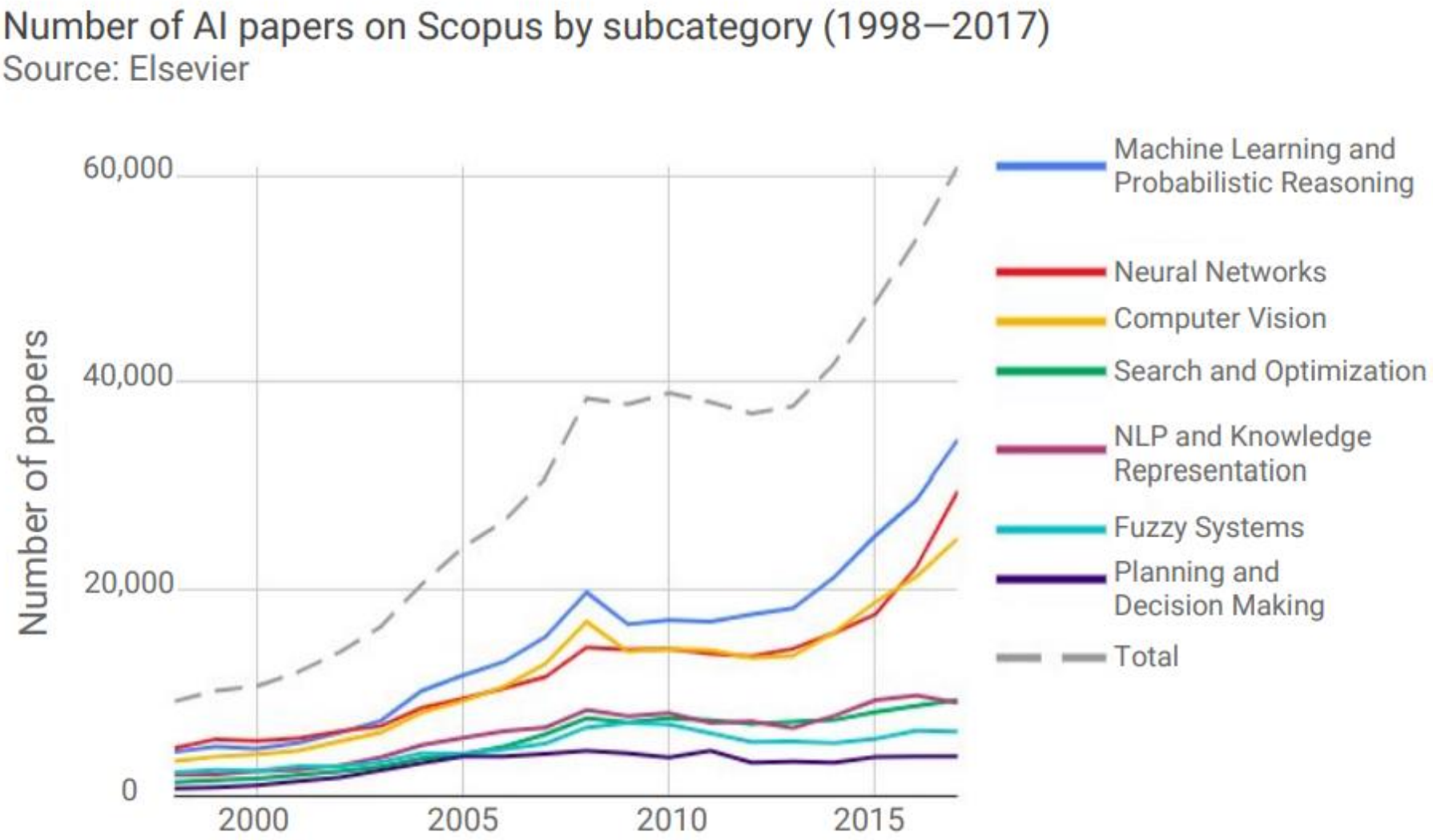


1956年达特茅斯人工智能研讨会



Founding fathers of AI. Courtesy of scienceabc.com

人工智能都在研究什么?


Number of AI papers on Scopus by subcategory (1998–2017)
Source: Elsevier
Number of papers
60,000
40,000
20,000
0
2000
2005
2010
2015
Machine Learning and Probabilistic Reasoning
Neural Networks
Computer Vision
Search and Optimization
NLP and Knowledge Representation
Fuzzy Systems
Planning and Decision Making
Total

人工智能三个流派

- 行为主义：基于控制论，构建感知-动作型控制系统
- 符号主义：基于符号逻辑的方法，用逻辑表示知识和求解问题
- 连接主义：基于大脑中神经元细胞连接的计算模型，用人工神经网络来拟合智能行为

符号逻辑的一个例子

$$\forall x \forall y (P(f(x)) \rightarrow \neg (P(x) \rightarrow Q(f(y), x, z)))$$

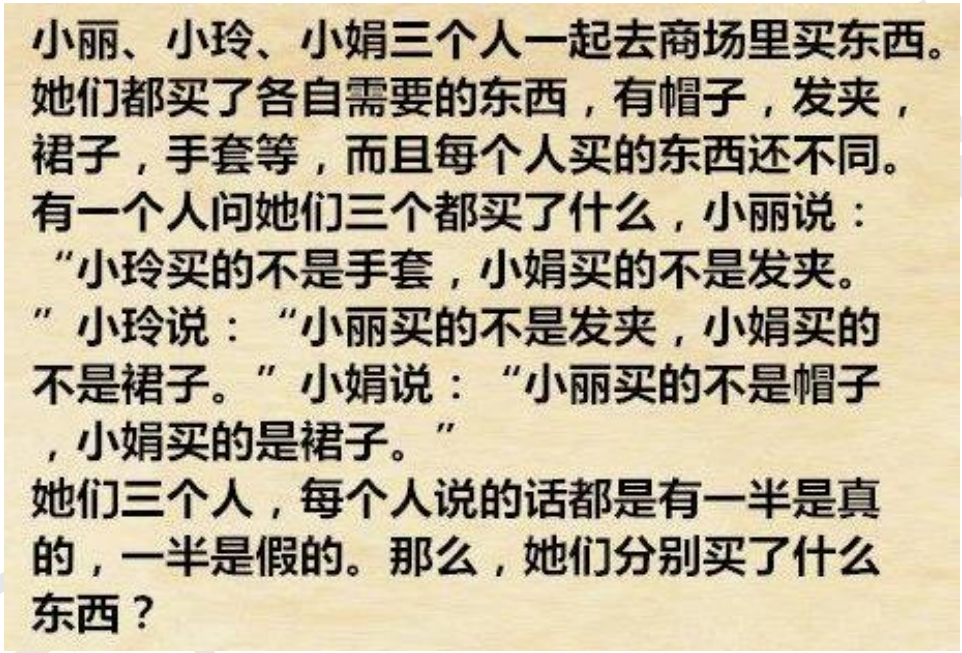
一阶谓词逻辑

符号主义的困难：逻辑，常识，求解器

- 逻辑：未找到能表述世间所有知识的简洁逻辑体系
- 常识：无穷无尽的常识
- 求解器：命题逻辑判定NP完全，一阶谓词逻辑不可判定

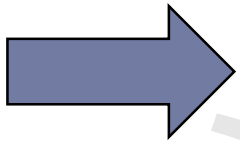
更本质的问题

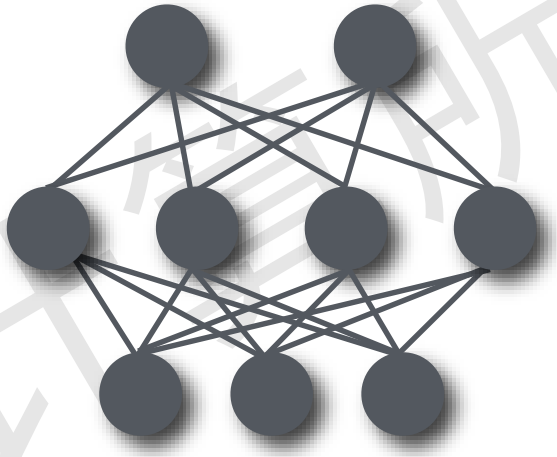
- 人的智能主要是符号智能吗？



- 符号主义最本质的问题是只考虑了理性认识的智能。人类的智能包括感性认识（感知）和理性认识（认知）两个方面
  - 人类语言的例子：词汇，时态，格，数字

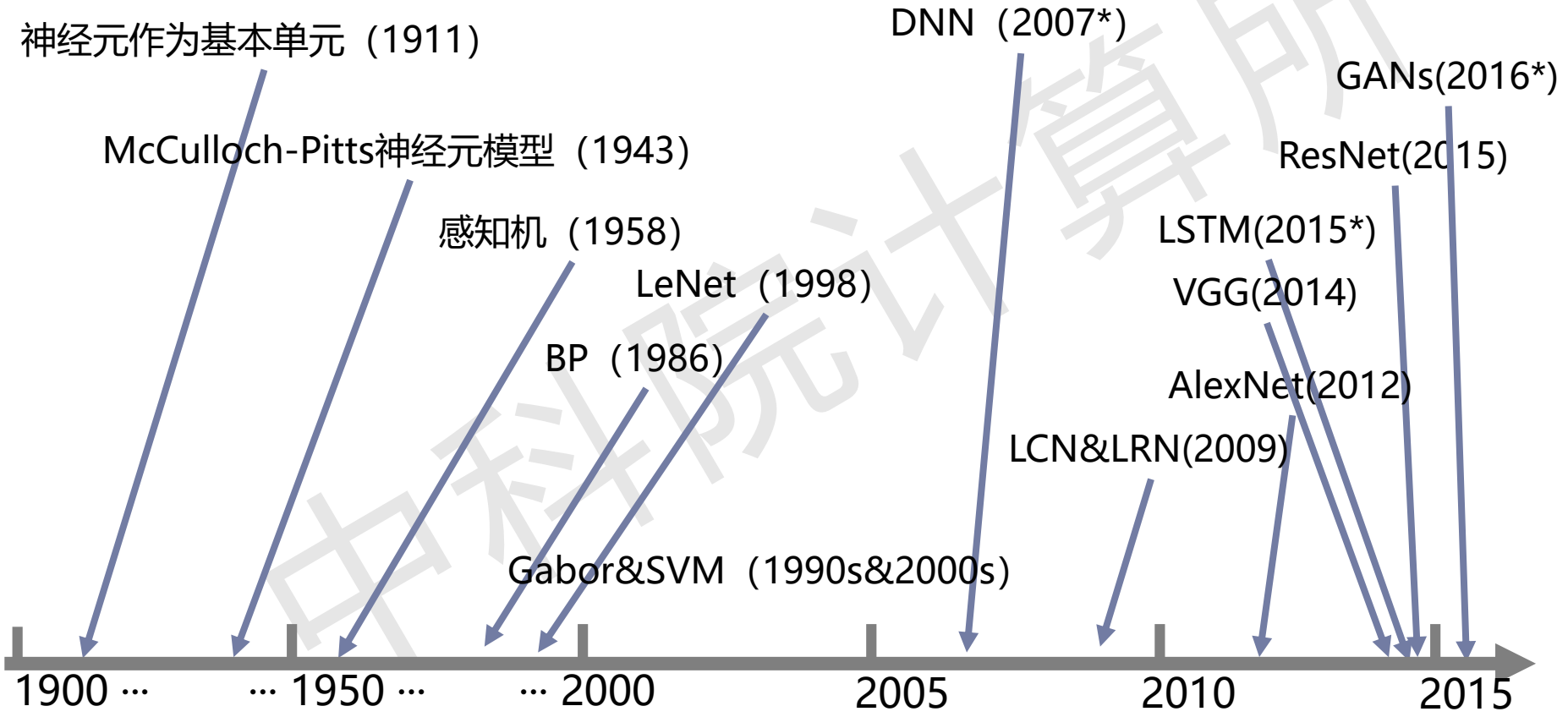
连接主义：人工神经网络





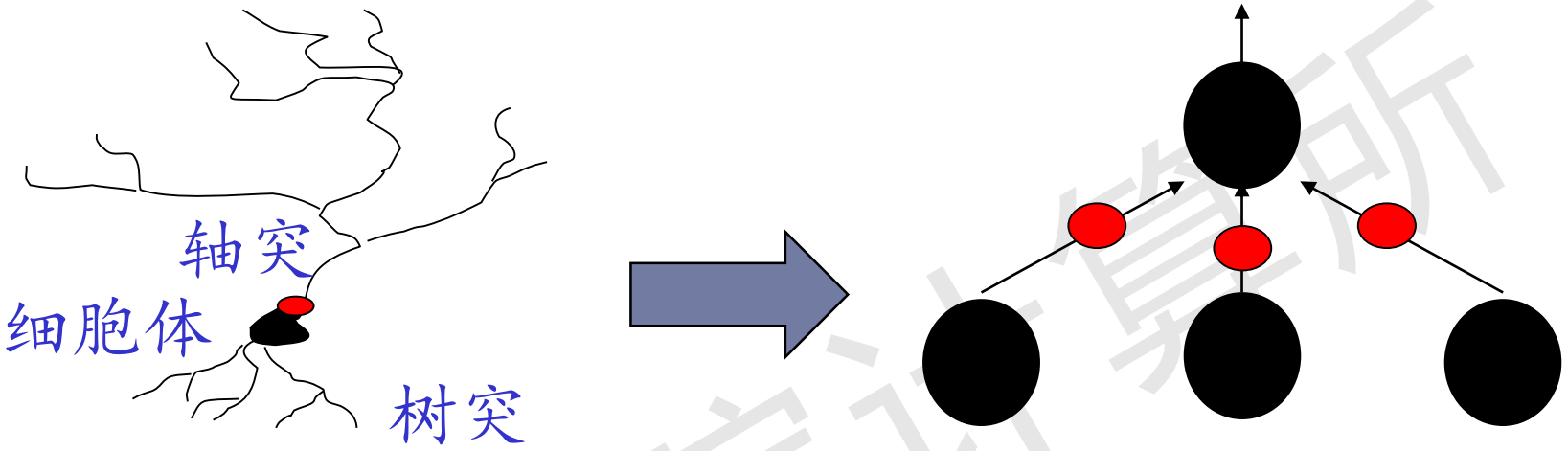
- 1943， 神经元模型， McCulloch 和 Pitts，第一波神经网络
- 1949，《The Organization of Behaviour》，Hebb学习
- 1958，感知机模型（preceptron），Rosenblatt
- 1986，BP反向传播训练方法， Rumelhart、Hinton 和 Williams，第二波神经网络
- 1998，卷积神经网络， Lecun
- 2000，自然语言模型， Bengio
- 2006，深度置信网络（DBN），Hinton, 第三波神经网络
- 2012，AlexNet （Dropout），Hinton团队赢得ImageNet比赛ILSVRC的冠军
- 2015，Deep Residual Network，AlphaGo

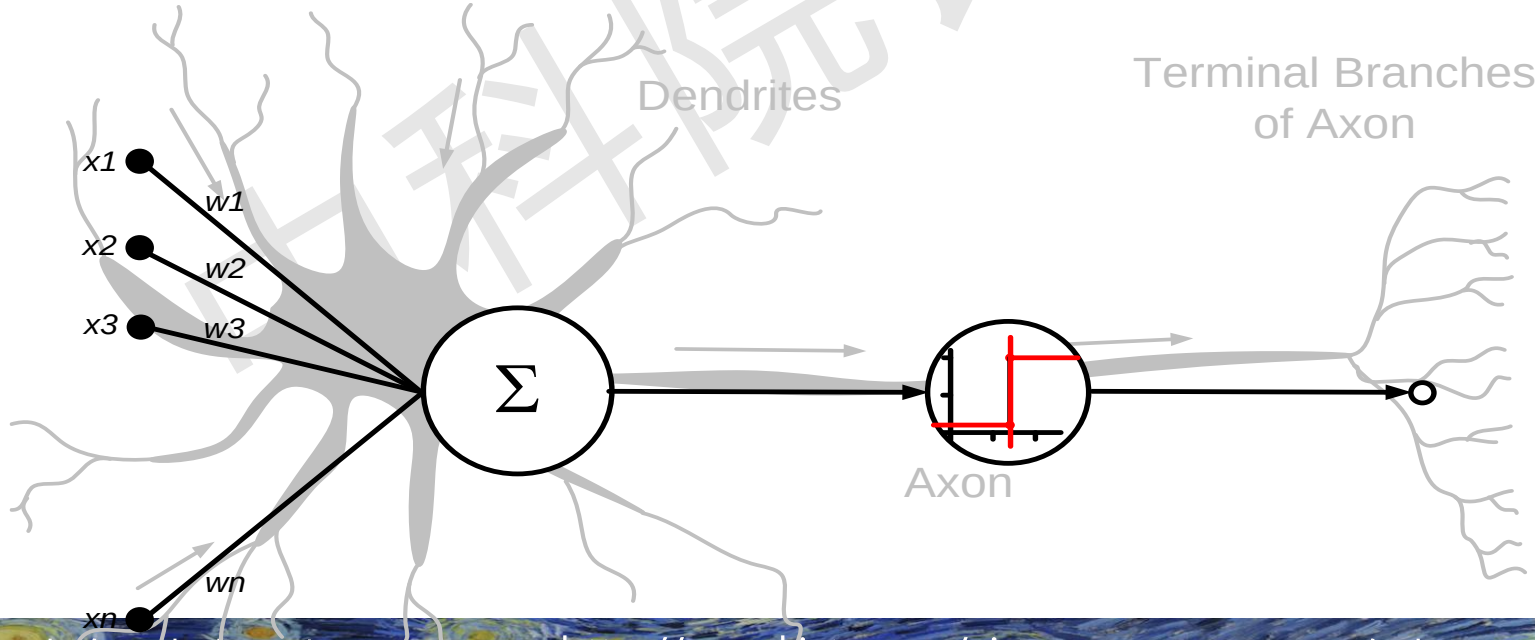
人工神经网络



*开始流行时间

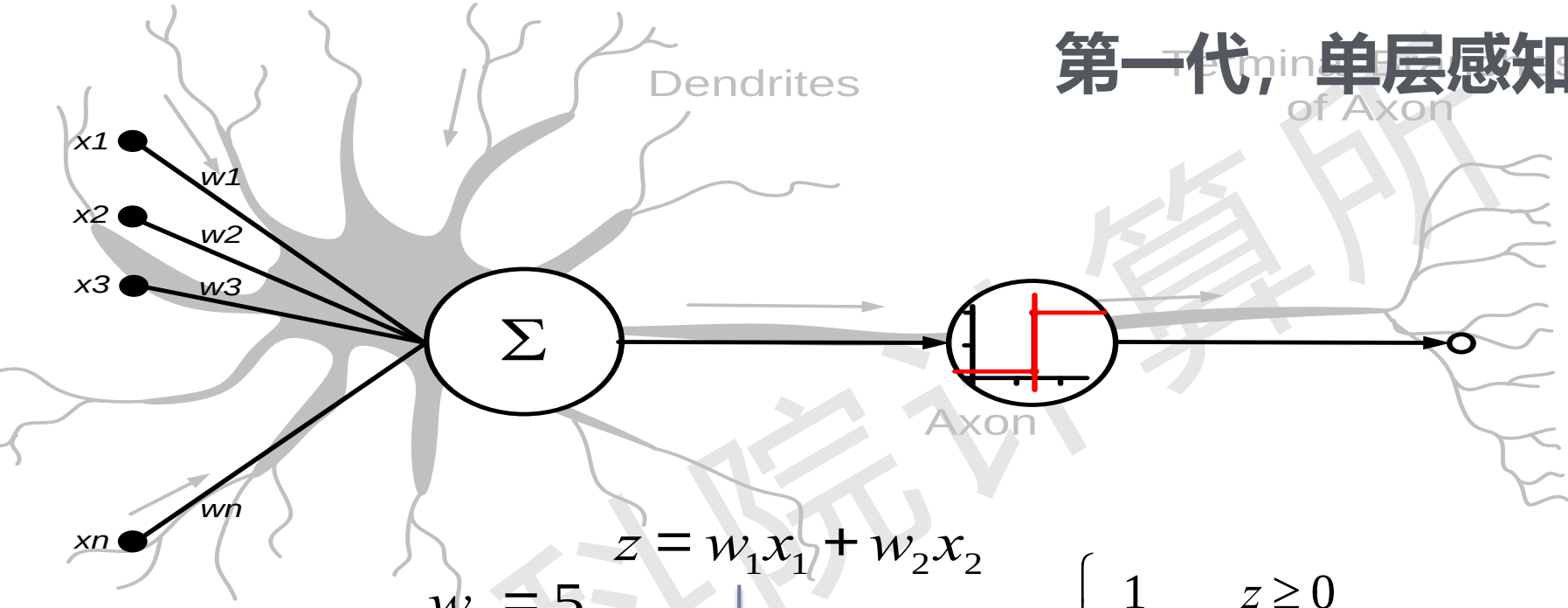
人工神经元



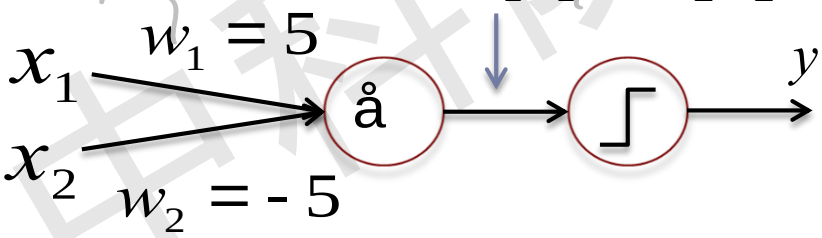


一个神经元的单层感知机

第一代，单层感知机



$$z = w_1 x_1 + w_2 x_2$$



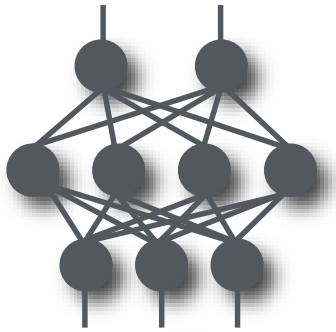
$$y = \begin{cases} 1 & z \geq 0 \\ 0 & otherwise \end{cases}$$

x_1	x_2	z	y
1	-1	10	1
-1	1	-10	0

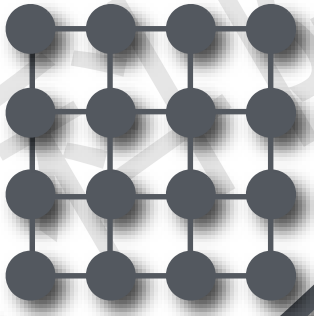
多层+多个神经元

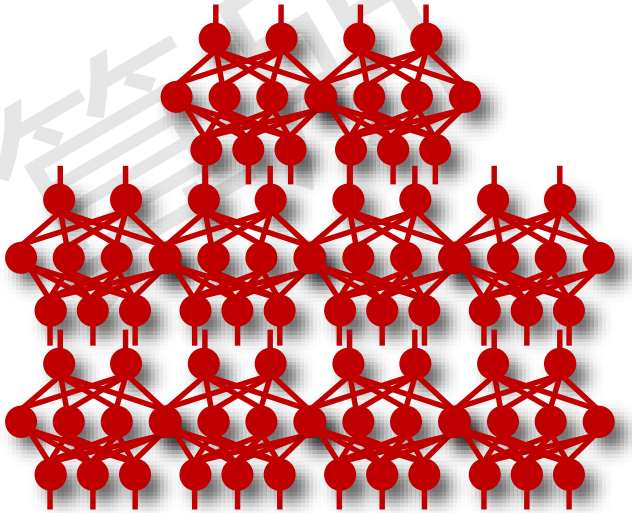
第三代，深度神经网络

第二代，MLP



SVMs

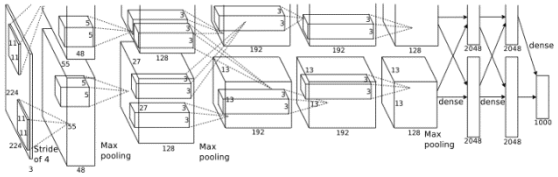


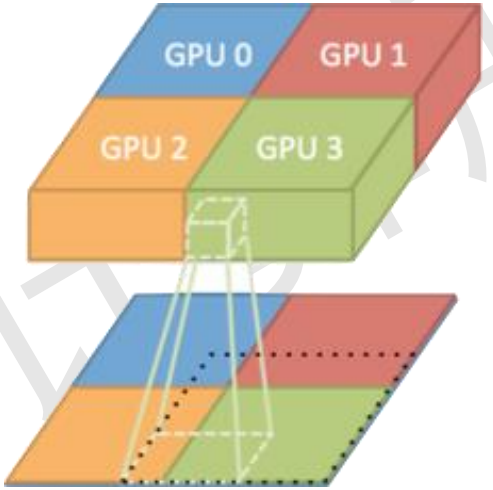


1990s

如今

深而大的深度神经网络





ResNet: 152 层

某些商业网络：512层

6千万参数

十亿参数

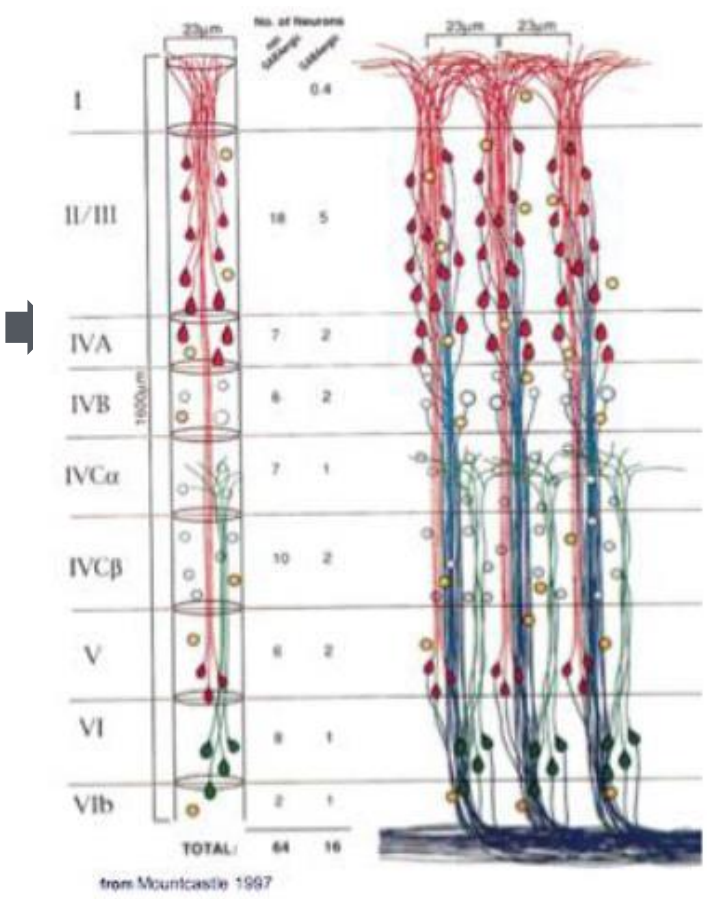
110亿参数

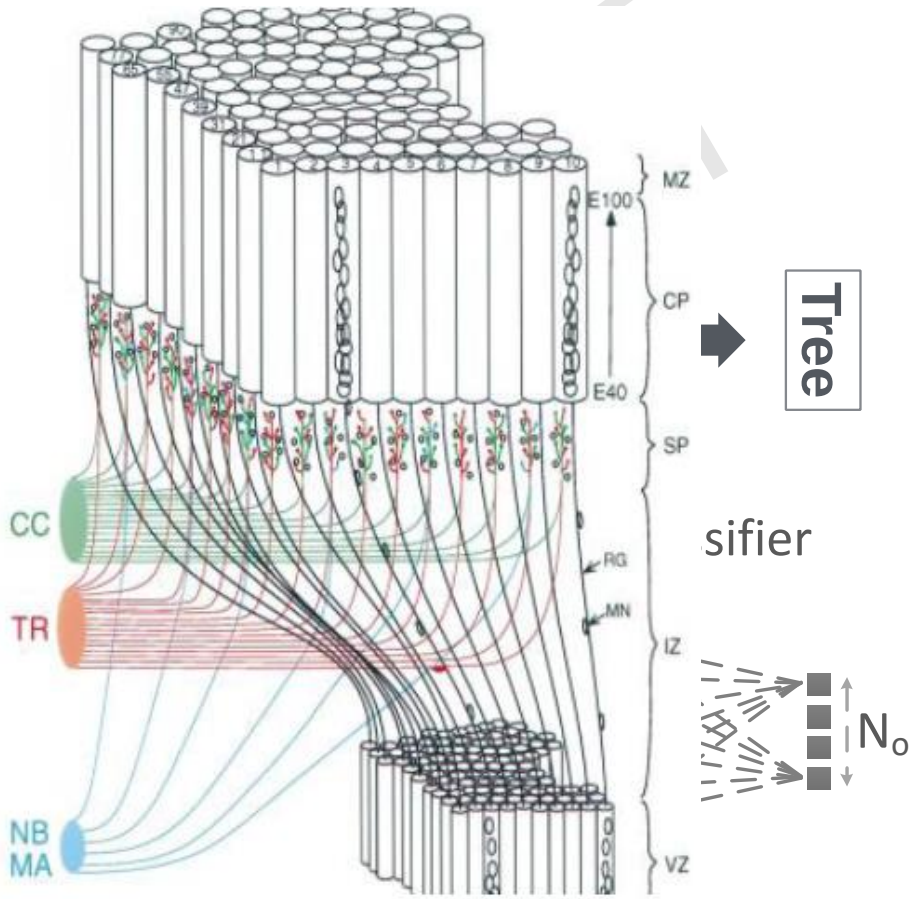
Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems (pp. 1–9).

Le, Q.V., Ranzato, M.A., Monga, R., Devin, M., Chen, K., Corrado, G. S., … Ng, A.Y. (2012). Building High-level Features Using Large Scale Unsupervised Learning. In International Conference on Machine Learning.

Coates, A., Huval, B., Wang, T., Wu, D. J., & Ng, A.Y. (2013). Deep learning with cots hpc systems. In International Conference on Machine Learning.

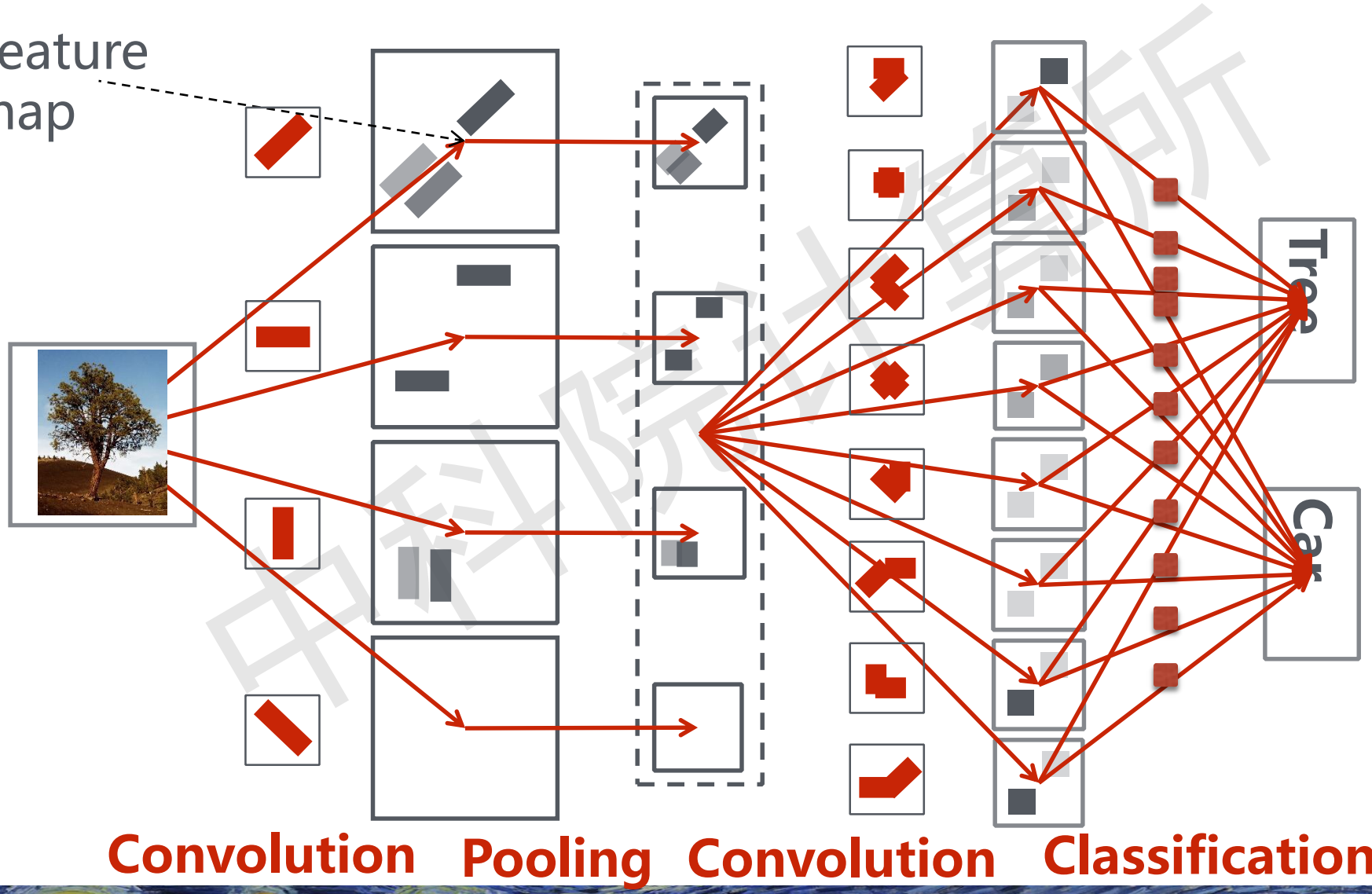
深而大的深度神经网络





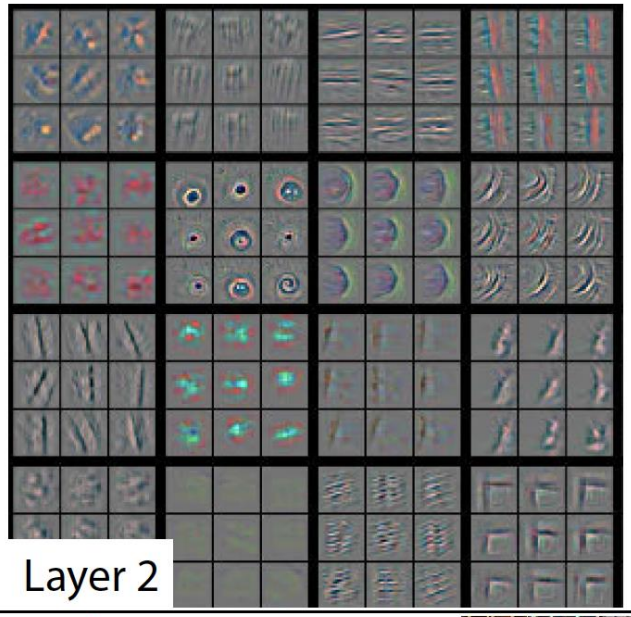
多层大规模人工神经网络

深度神经网络的组织方式

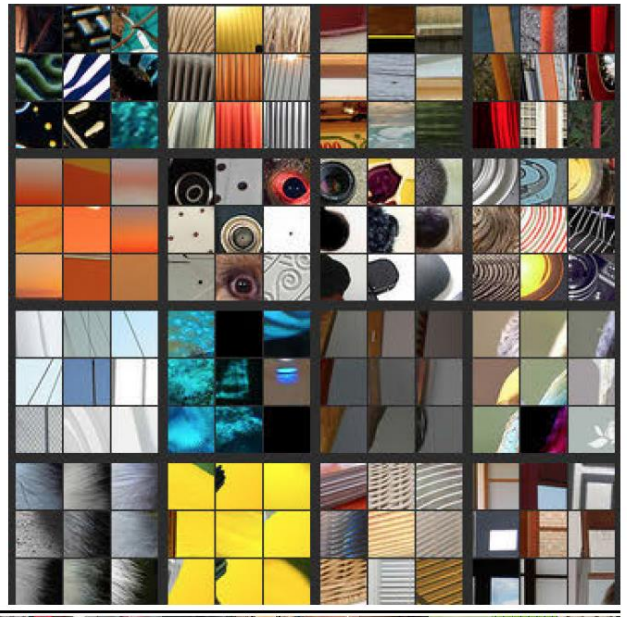


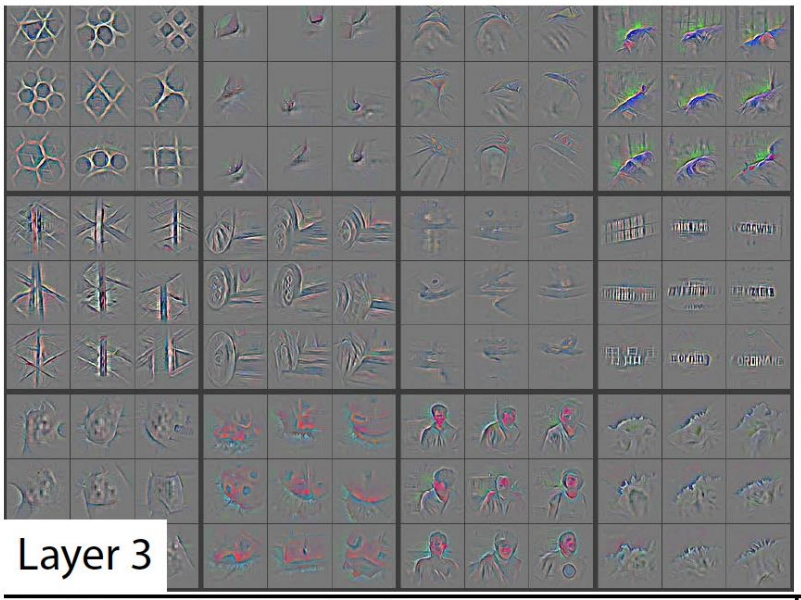
深度学习工作机理

Layer 1

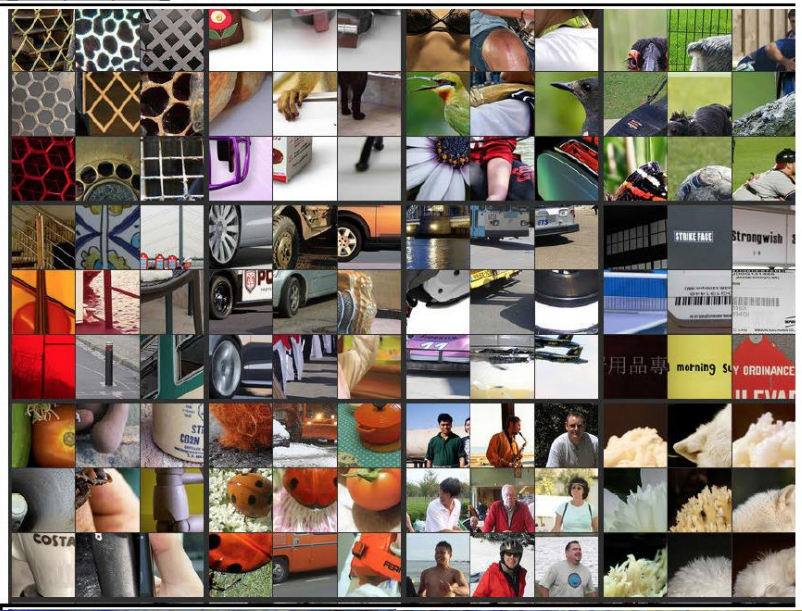


Layer 2



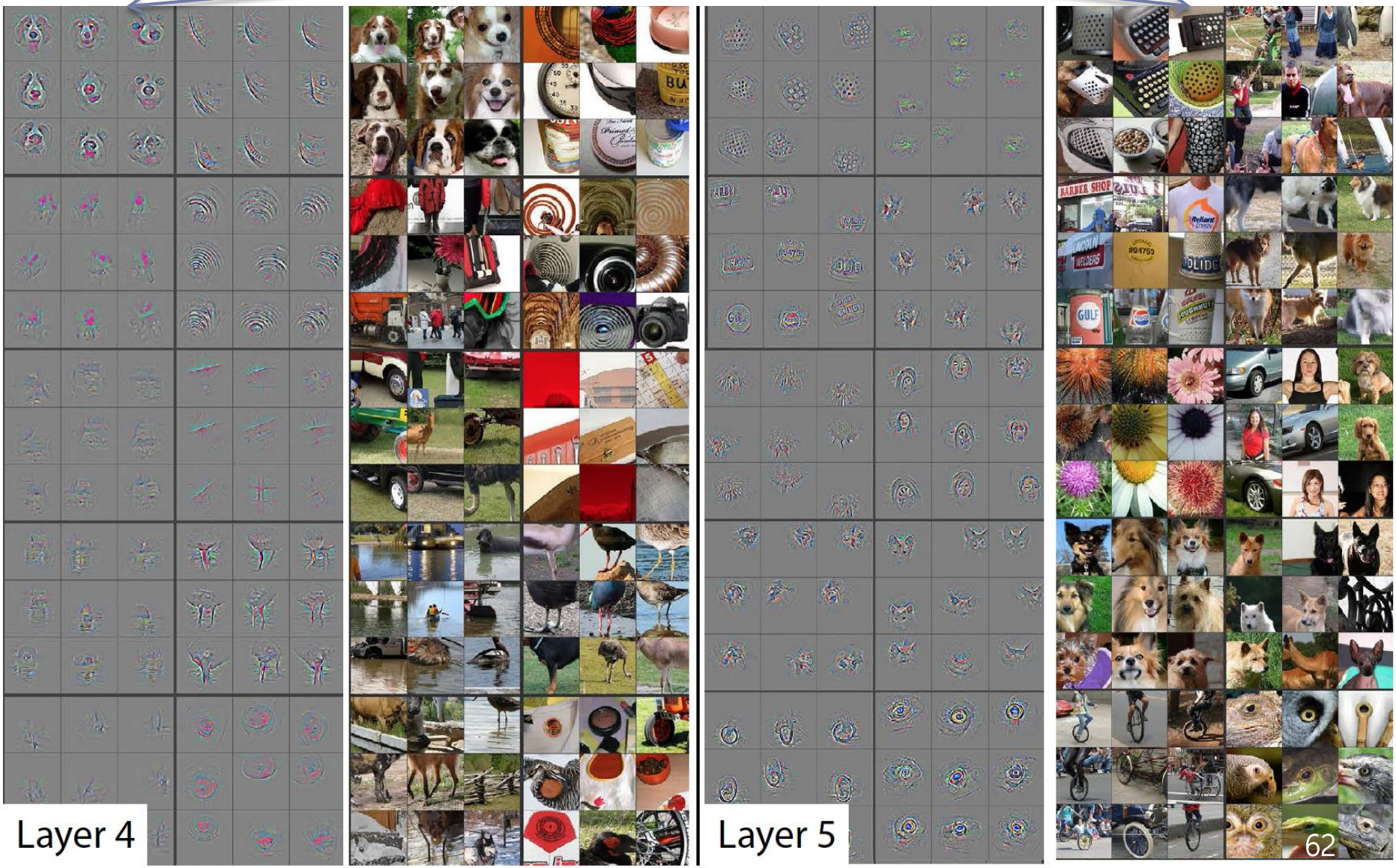


Layer 3

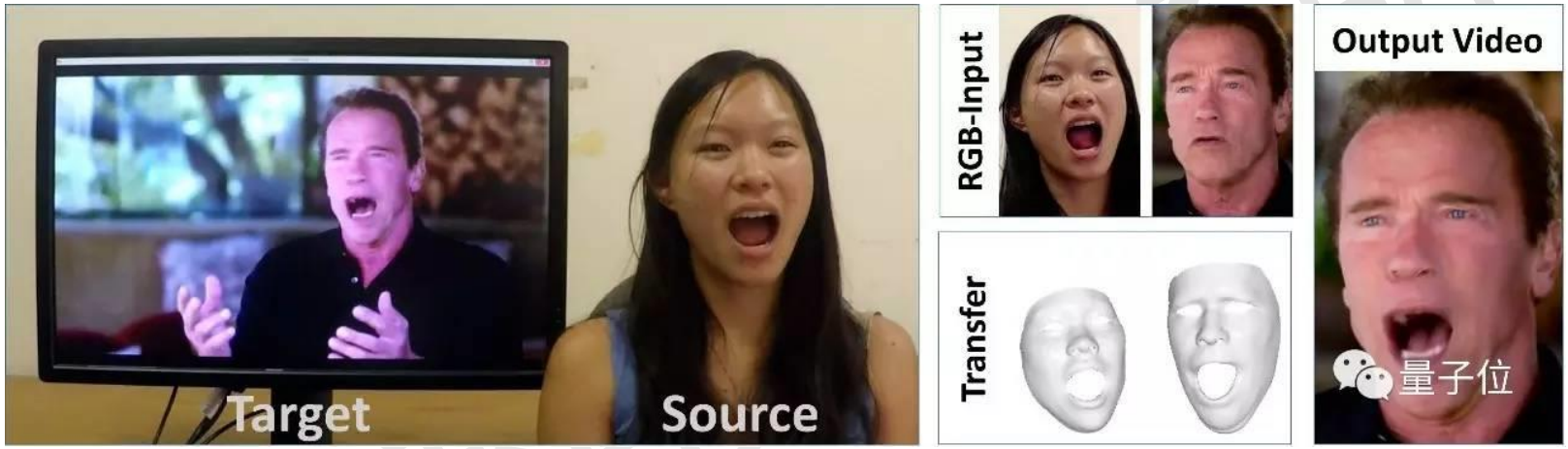


深度学习工作机理

分类器



深度学习20个有意思的小应用



https://www.cnblogs.com/czaoth/p/6755609.html

深度学习的局限性

- 深度学习是一把梯子，而不是火箭
  - 泛化能力有限
  - 缺乏推理能力
  - 缺乏可解释性
  - 鲁棒性欠佳

提纲

- 为什么要开这门课
- 为什么来上这门课
- 人工智能
- 智能计算系统
- 驱动范例

什么是智能计算系统

智能计算系统是智能的物质载体

现阶段的智能计算系统通常是集成CPU和智能芯片的异构系统，软件上通常包括一套面向开发者的智能计算编程环境（包括编程框架和编程语言）

异构智能计算系统

- 现今采用异构智能计算系统的主要原因:

近十年来通用 CPU 的计算能力增长近乎停滞，而智能计算能力的需求在不断以指数增长，二者形成了剪刀差

  - e.g. 寒武纪深度学习处理器能够以比通用 CPU 低一个数量级的能耗，达到 100 倍以上的智能处理的速度

- 异构系统在提高性能的同时，也带来了编程上的困难
  - 智能计算系统一般会集成一套编程环境，方便程序员快速便捷地开发高能效的智能应用程序
  - 常用的深度学习编程框架包括 TensorFlow 和 MXNet等
  - 深度学习编程语言包括 CUDA 语言和 BCL 语言等

为什么需要智能计算系统

和李世石下一盘围棋电费数千美元的AlphaGo

1.6万个CPU核学一周识别猫脸的谷歌大脑

人工智能必须有其核心物质载体

三代智能计算系统

- 第一代智能计算系统：1980年代，面向符号主义智能处理的专用计算机（Prolog机，LISP机）
- 第二代智能计算系统：2010年代，面向连接主义智能处理的专用计算机（深度学习计算机）
- 第三代智能计算系统：未来强人工智能/通用人工智能的载体

第一代智能计算系统

- 1975，MIT AI Lab的Greenblatt研制成功LISP机CONS
- 1978，MIT AI Lab发布CONS的后继，CADR
- 1980s，发展高峰
  - Symbolics (3600, 3640, XL1200, MacIvory)
  - Lisp Machines Incorporated (LMI Lambda)
  - Texas Instruments (Explorer and MicroExplorer)
  - Xerox (Interlisp-D workstations)
  - 日本，五代机
  - Prolog机，1983，David H. D. Warren Warren Abstract Machine
- 1980s末到1990s初，AI winter，第一代智能机市场坍塌

第一代智能计算系统

LISP机（MIT博物馆）

Symbolics 3640

第一代智能计算系统

- High-level language computer architecture
  - OS的编程语言和硬件“统一”化，如LISP
  - 只针对特定语言的优化
- 局限性
  - 没有太多的实际应用需求
  - 由于摩尔定律发展，性能比不上CPU
    - 贵，几十万美元一台

第二代智能计算系统

- 面向连接主义（深度学习）处理的计算机或处理器
- 和第一代智能计算系统相比，第二代智能计算系统有两方面的优势：
  - 深度学习有大量实际的工业应用，已经形成了产业体系，因此相关研究能得到政府和企业的长期资助；
  - 摩尔定律在 21 世纪发展放缓，通用 CPU 性能增长停滞，专用智能计算系统的性能优势越来越大。

因此，在可预见的将来，第二代智能计算系统还将长期健壮发展，持续迭代优化

第二代智能计算系统


物端设备


移动设备


客户端


汽车


服务器


超级计算机

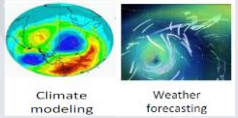

图像识别


语音识别


游戏竞技


自动驾驶

广告推荐

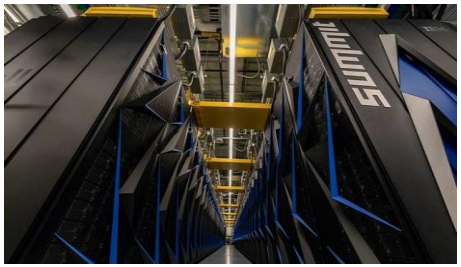


气象预报

TensorFlow caffe dmlc mxnet DRAGON ONNX PyTorch

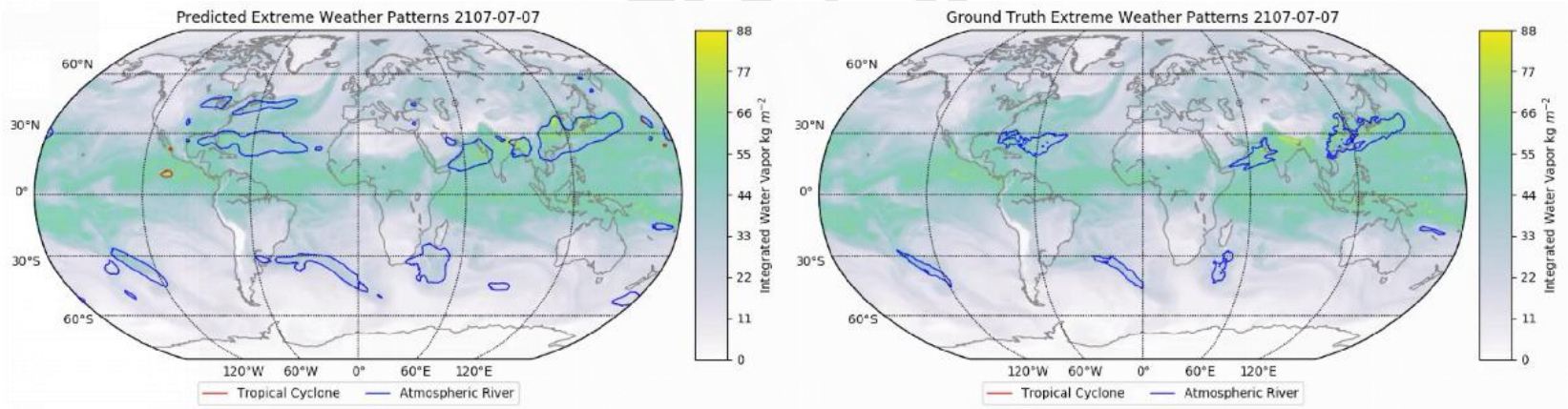
智能计算系统

第二代智能计算系统



美国智能计算系统代表“顶点”（Summit）
浮点运算速度峰值达每秒20亿亿次
（200PFlops）

中国智能计算系统代表“曙光7000”
浮点运算速度峰值达每秒30亿亿次
（300PFlops）



2018年**戈登•贝尔奖**由**劳伦斯伯克利国家实验室**和**NVIDIA公司**的联合研究团队使用**Summit智能计算平台**完成，获奖原因为“Employing **Deep Learning Methods** to Understand Weather Patterns”，该模型使用了混合精度进行训练，峰值算力达到了**1.13Eflops**

第二代智能计算系统

代表性深度学习处理器/计算机

时间	深度学习处理器/计算机	研制单位	特点
2013 年	DianNao[19]	中科院计算所	国际上首个深度学习处理器架构
2014 年	DaDianNao[20]	中科院计算所	国际上首个多核深度学习处理器架构
	cuDNN（深度学习库）	NVIDIA	升级 GPU 用于深度学习
2015 年	PuDianNao[21]	中科院计算所	国际上首个通用机器学习处理器
	ShiDianNao[22]	中科院计算所	端侧视频图像处理
2016 年	Cambricon[23]	中科院计算所	国际上首个深度学习指令集
	Cambricon-X[24]	中科院计算所	国际上首个稀疏神经网络处理器
2017 年	TPU[25]	Google	基于脉动阵列架构
	FlexFlow[26]	中科院计算所	动态数据流结构
2018 年	TPUv3 cloud	Google	基于 TPUv3 芯片的云计算
	DGX-2 服务器	NVIDIA	16 块 NVIDIA v100 显卡
	Summit 超级计算机	IBM	27684 块 NVIDIA v100 显卡
	MLU100	Cambricon	基于寒武纪云端智能芯片
2019 年	E-RNN[27]	Syracuse 大学	循环神经网络加速器
	Cambricon-F[28]	中科院计算所	分形冯诺依曼架构
	Float-PIM[29]	UCSD	支持训练的存内计算架构

第三代智能计算系统展望

- 具有近乎无限的计算能力，会给人工智能带来什么？
- 从奴仆到主人：第三代智能计算系统不再是智能算法的加速工具
- 它将是通用人工智能（强人工智能）发育的沙盒虚拟世界
  - 大量独立的智能主体，足够丰富的虚拟世界，以及背后的无限计算能力
  - 智能主体将在其中成长，通过和环境的交互，形成感知、认知和逻辑能力，甚至理解虚拟世界、改造虚拟世界，从而具备通用智能

为什么要在沙盒环境中发育智能?

- 通用人工智能的危险性
- 生存、繁衍等意识基础无法在现实环境中实现
  - 没有生存、繁衍等本能，智能主体的知识大厦没有根基
- 现实环境的低效低速

第三代智能计算系统探索的思路

总体思路：具备全面感知能力和超大规模硬件的原始智人是怎样一步步获得智能的？

- 体系结构：面向海量并发认知智能计算线程和超大规模虚拟环境的计算机和芯片
- 算法：有限延迟的认知智能算法，能自主产生语言和文字，从本能之上建立起自己的知识图谱，打通感知到逻辑的鸿沟
- 编程框架，操作系统，网络等等都将为之巨变

提纲

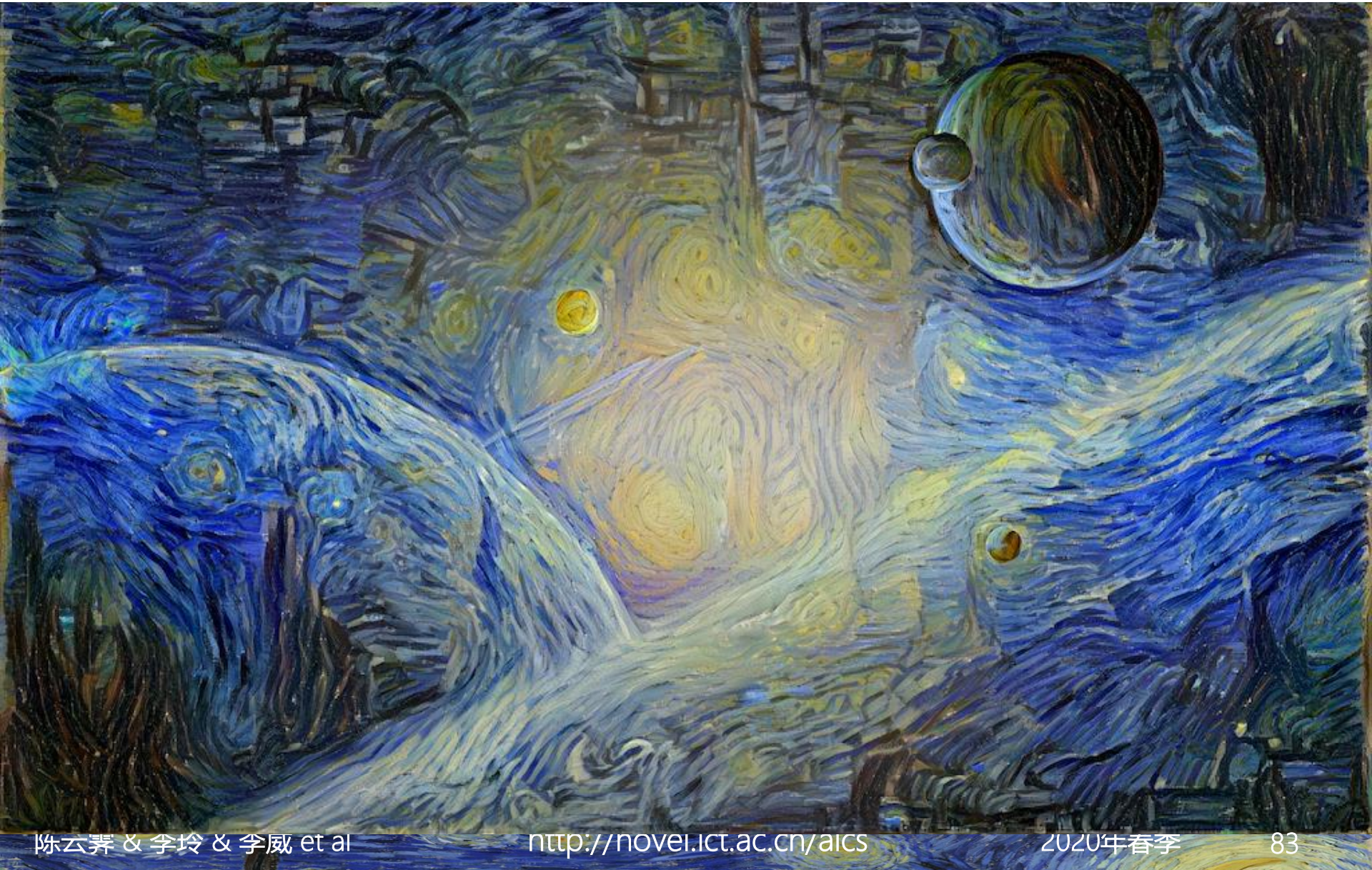
- 为什么要开这门课
- 为什么来上这门课
- 人工智能
- 智能计算系统
- 驱动范例

Driving Example

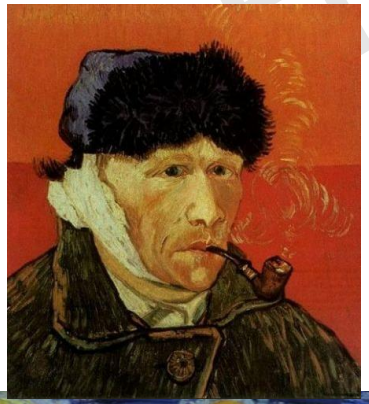
- 从一个例子开始，串起本课程各个章节
- 当你上完
  - 掌握了第二代智能计算系统（深度学习计算机）的原理和使用
  - 形成了第三代智能计算系统（通用智能孵化器）的初步概念和兴趣

“星空”

“星空”



如何解决一个AI的任务?

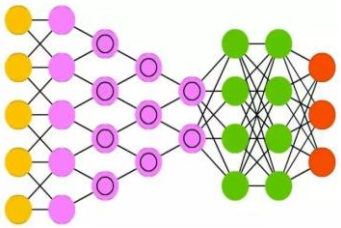


处理过程

输入

输入 ▶ 建模 ▶ 实现 ▶ 运行 ▶ 输出

建模

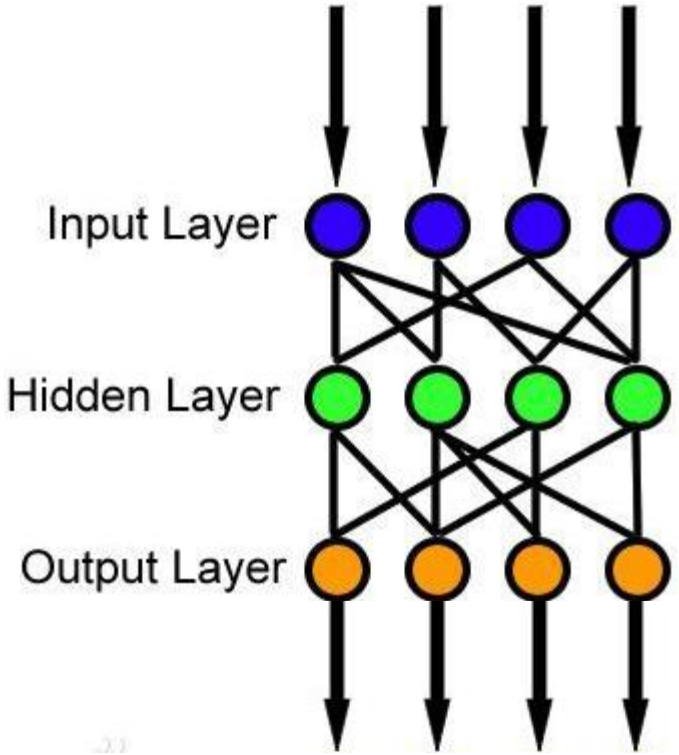


输入 → 建模 → 实现 → 运行 → 输出

神经网络基础

深度学习

神经网络基础



颜色 / 纹理/ 形状特征

监督 / 无监督

Sigmoid / ReLU / tanh

Forward Propagation
Back Propagation

最优化

深度学习

- 传统模式识别

手工
提取特征

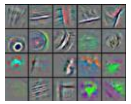
训练分类器

- 深度学习，就是多层人工神经网络

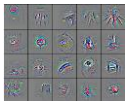
深度学习最重要的作用是**表征学习**，学习层级化的特征，“深度”这词指的就是很多层

底层特征

中层特征



高层特征

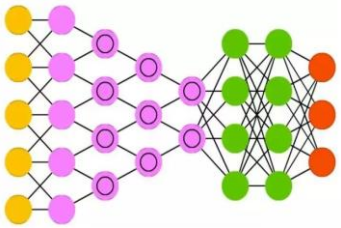


训练分类器

深度学习

- 算法

实现



输入 ▶ 建模 ▶ 实现 ▶ 运行 ▶ 输出

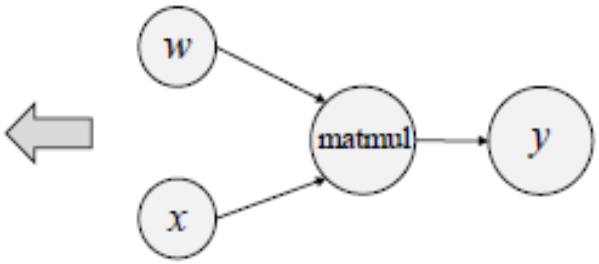
编程框架

Bang

编程框架

- 将深度学习算法中的基本操作封装成一系列组件，帮助研究人员更简单的实现已有算法，或设计新的算法。这一系列深度学习组件，即构成一套深度学习框架
- 以TensorFlow为例
  - 1、构建数据流图，描述计算过程
  - 2、执行数据流图，获得计算结果

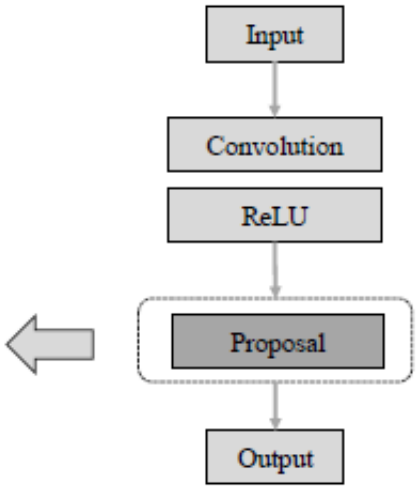

```
import tensorflow as tf

a = tf.constant([[3., 3.]])
b = tf.constant([[2.], [2.]])

y = tf.matmul(a, b)

with tf.Session() as sess:
    result = sess.run(y)
    print(result)
```



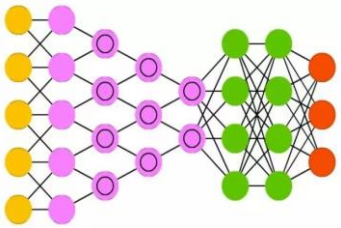
Bang

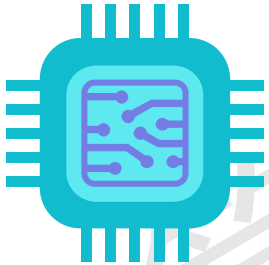
- Bang是寒武纪提出的异构编程语言，它是基于C语言的扩展，简单易学，同时提供了丰富高效的编程接口，高效实现编程框架所需的算子
- 深度学习算法的实现人员使用Bang语言将神经网络的基本操作实现为能在寒武纪平台上运行的程序，以供编程框架调用

```
__dlp_entry__ void Proposal(…) {
  …
  __nram__ half scores[…];
  __nramset_half(scores, …);
  …
  __bcl_maxpool(…);
  …
}
```



运行



输入 ▶ 建模 ▶ 实现 ▶ 运行 ▶ 输出


架构基础


架构设计

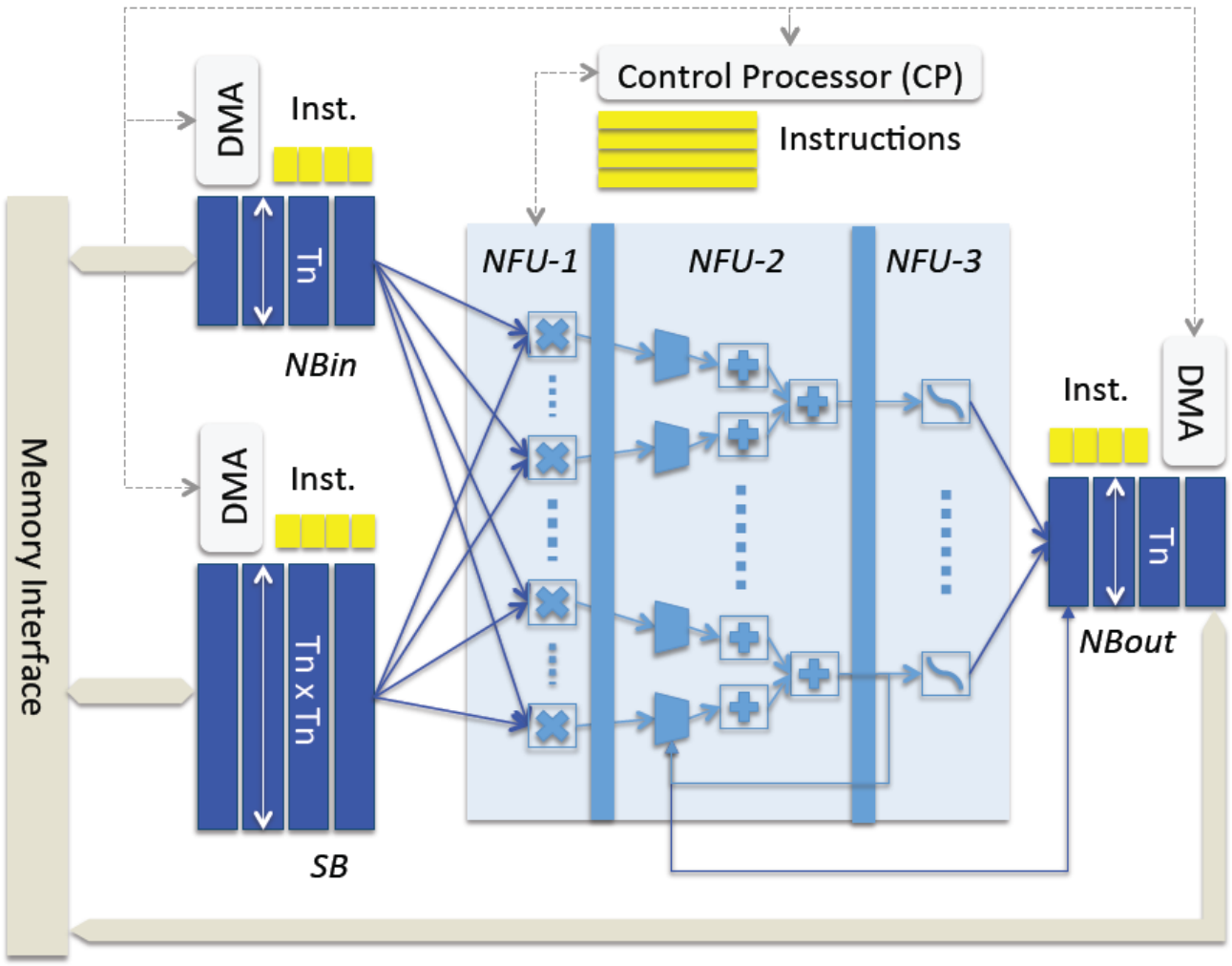

标准与评测

架构基础

- 人工智能处理器的意义
  - 人工智能算法的新需求：神经网络规模不断增加
  - 通用处理器的局限性
    - 谷歌大脑（百亿突触）：1.6万CPU核三天训练猫脸识别模型
    - 不可能扩展至人脑规模（百万亿突触）
    - 性能和能耗问题
- 人工智能处理器发展简史
  - 硬件化：计算和访存模式的适配
  - 算法优化：降低存储和计算量
  - 软硬件协同：面向硬件的算法优化

架构设计

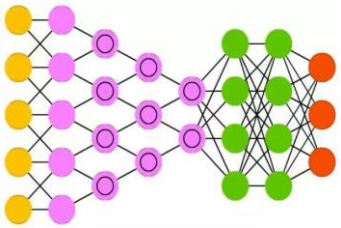
- 运算单元设计
- 存储层次设计
- 指令集设计
- 编程框架设计
- 多处理器架构设计
- 典型架构设计：DianNao
- 典型架构设计：MLU100

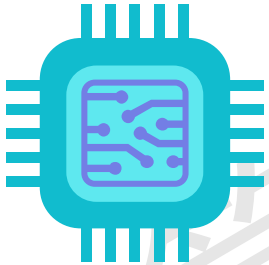


评测

- 评测的意义与方法
- 现有的评测标准
- BenchIP

输出




输入 ▶ 建模 ▶ 实现 ▶ 运行 ▶ 输出


运行环境搭建


运行与调试


应用与开发

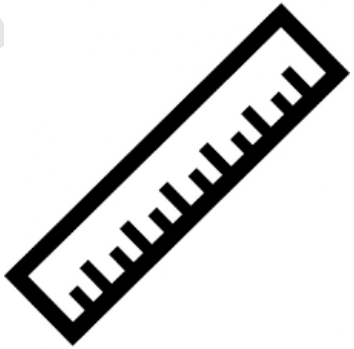
运行环境搭建

- 硬件环境
  - 采用寒武纪系列智能加速卡
    - 国内首款自主知识产权的智能处理器
    - 支持所有现存的人工智能算法（包括但不限于CNN/DNN/DBN/RNN/LSTM/SOM/RCNN/Faster-RCNN/DeepID/YOLO等）。
    - 相比传统的通用处理器（CPU），能效提升100倍，广泛应用在手机和服务器中

运行与调试

- 代码的开发及编译
  - 串口调试
  - 配置网络文件系统
- 结果测试

模型训练



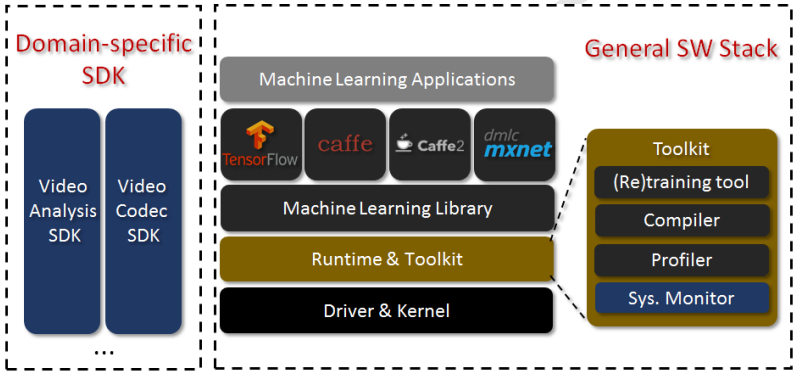
性能剖析

系统监控

应用与开发

- 智能应用依赖库的开发

Cambricon Neuware 机器学习库（Cambricon Neuware Machine Learning Library，CNML）提供了一套高效、通用、灵活、可扩展的编程接口，用于在智能加速卡上加速各种机器学习/深度学习算法。

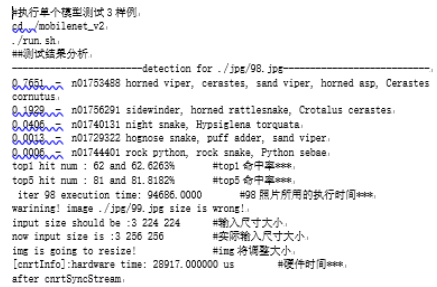


Cambricon Neuware

- 数据预处理
- 人工智能网络运行
- 运行结果的后处理

```
./gen_all_models.sh
#编译结果
.....
I1218 10:05:30.578732 100149 subnet.hpp:169] subnet[1] fusing...
I1218 10:05:30.578883 100149 fusion.cpp:58] [Fusion] setFusionIO (size: 1, 3)...
I1218 10:05:30.578904 100149 fusion.cpp:45] [Fusion] compiling...0x3ebf220
I1218 10:05:43.869375 100149 net.cpp:426] Offline model generated!
********** Offline model information BEGIN
**********
file name : offline_models/zf_faster_rcnn/zf_faster_rcnn.cambricon
model name: offline_models/zf_faster_rcnn/zf_faster_rcnn
model details as follow:
[On CPU] subnet[0] layers : 0(input)
[On MLU] [call via func "subnet0"] subnet[1] layers : 1(conv1) 2(relu1) 3(norm1) 4(pool1) 5(conv2) 6(relu2) 7(norm2) 8(pool2) 9(conv3) 10(relu3) 11(conv4) 12(relu4) 13(conv5) 14(relu5) 15(rpn_conv/3x3) 16(rpn_relu/3x3) 17(rpn_cls_score) 18(rpn_bbox_pred) 19(proposal) 20(roi_pool_conv5) 21(fc6) 22(relu6) 23(drop6) 24(fc7) 25(relu7) 26(drop7) 27(cls_score) 28(bbox_pred) 29(cls_prob)
                    func "subnet0" inputs: data,
                    func "subnet0" outputs: rois, bbox_pred, cls_prob,
********** Offline model information END
**********
End zf_faster_rcnn offline models!!!!!
```

生成离线模型



```
执行单个模型测试 3 样例
cd ../mobilenet_v2
./run.sh
##测试结果分析
----------detection for ./jpg/98.jpg----------
0.7651 n01753488 horned viper, cerastes, sand viper, horned asp, Cerastes cornutus
0.1929 n01756291 sidewinder, horned rattlesnake, Crotalus cerastes
0.0406 n01740131 night snake, Hypsiglena torquata
0.0013 n01729322 hognose snake, puff adder, sand viper
0.0006 n01744401 rock python, rock snake, Python sebae
top1 hit num : 62 and 62.6263%      #top1 命中率***
top5 hit num : 81 and 81.8182%      #top5 命中率***
 iter 98 execution time: 94686.0000      #98 照片所用的执行时间***
warining! image ./jpg/99.jpg size is wrong!
input size should be :3 224 224      #输入尺寸大小
now input size is :3 256 256      #实际输入尺寸大小
img is going to resize!      #img 将调整大小
[cnrtInfo]:hardware time: 28917.000000 us      #硬件时间***
after cnrtSyncStream
```

运行应用程序

谢谢大家！