普适机器学习

(Pervasive Machine Learning)

周志华

http://cs.nju.edu.cn/people/zhouzh/

Email: zhouzh@nju.edu.cn

南京大学计算机软件新技术国家重点实验室

机器学习

机器学习是人工智能的核心研究领域之一

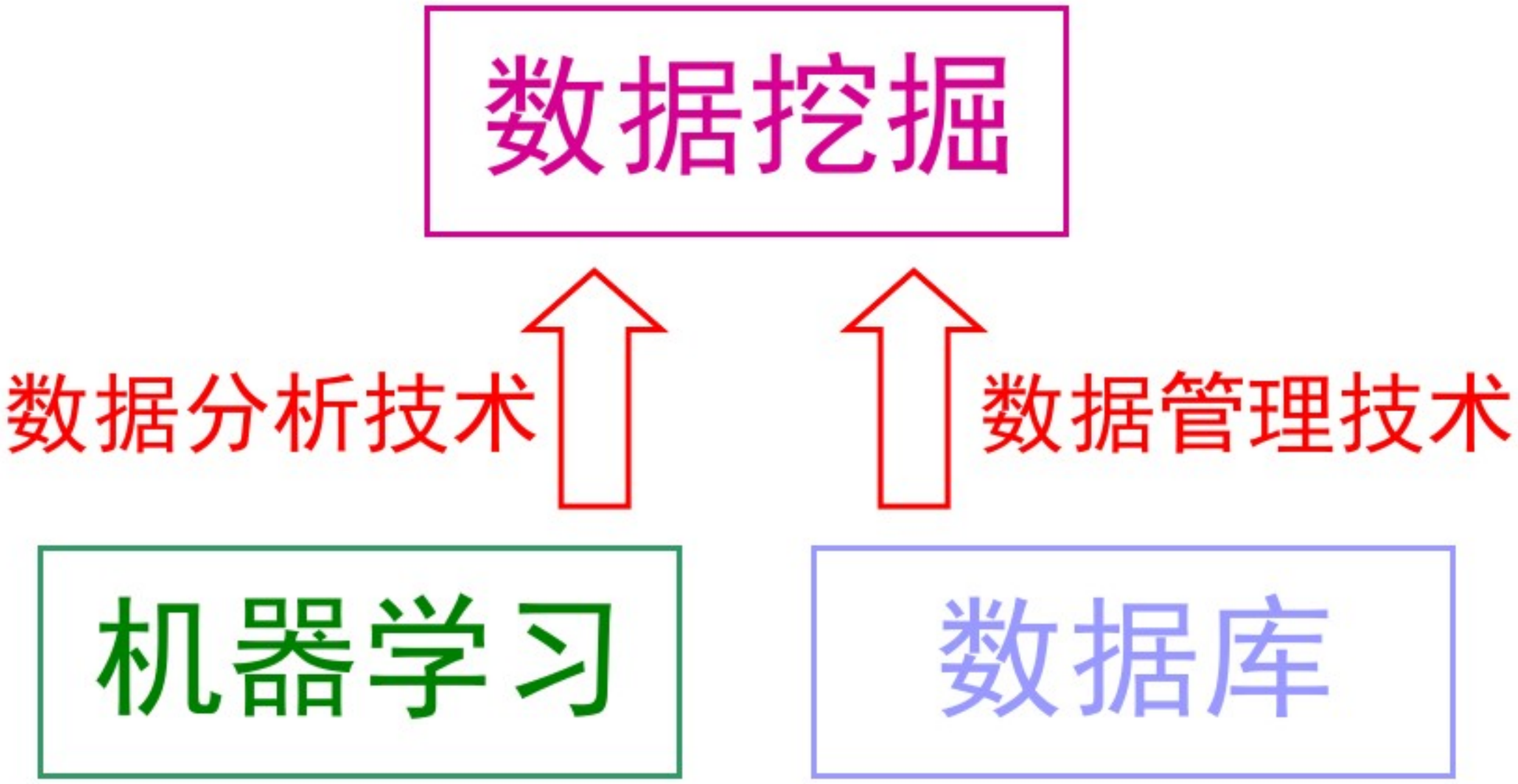
任何一个没有学习能力的系统都很难被认为是一个真正的智能系统

经典定义：利用经验改善系统自身的性能

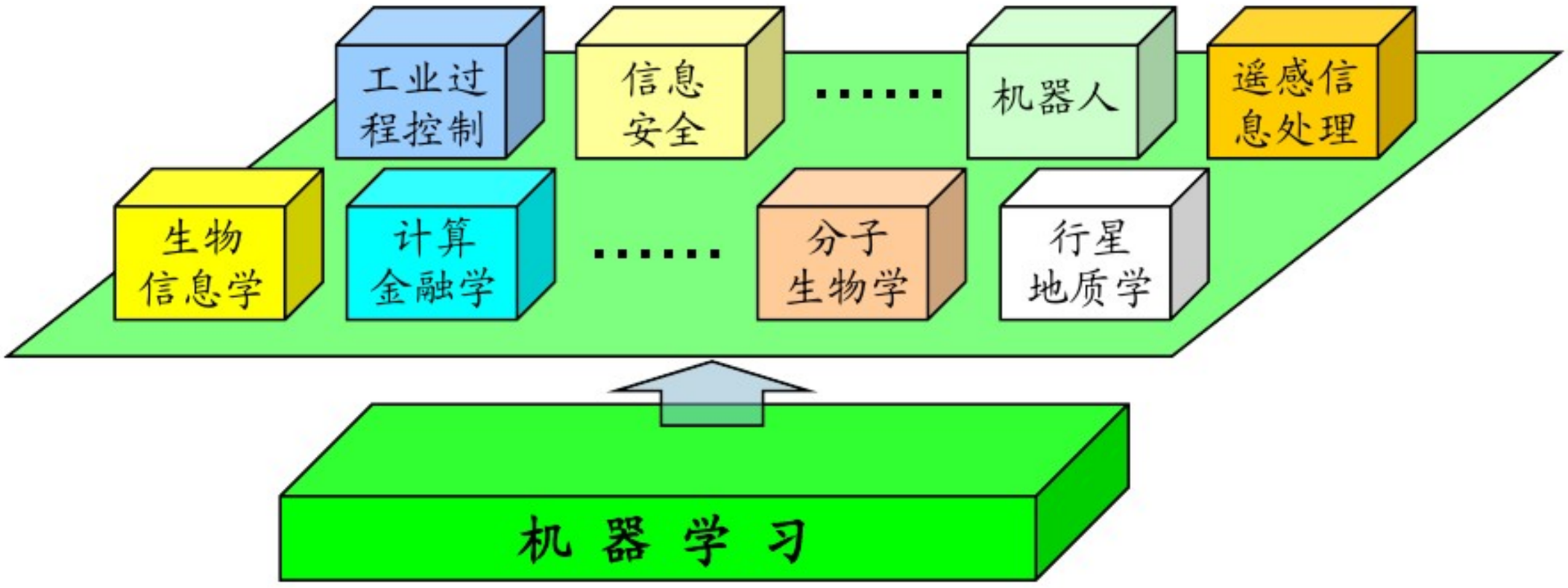
随着该领域的发展，主要做智能数据分析并已成为智能数据分析技术的源泉之一

典型任务：预测（例如：天气预报）

机器学习与数据挖掘



机器学习的重要性



美国航空航天局JPL实验室的科学家在《Science》（2001年9月）上撰文指出：机器学习对科学研究的整个过程正起到越来越大的支持作用，……，该领域在今后的若干年内将取得稳定而快速的发展

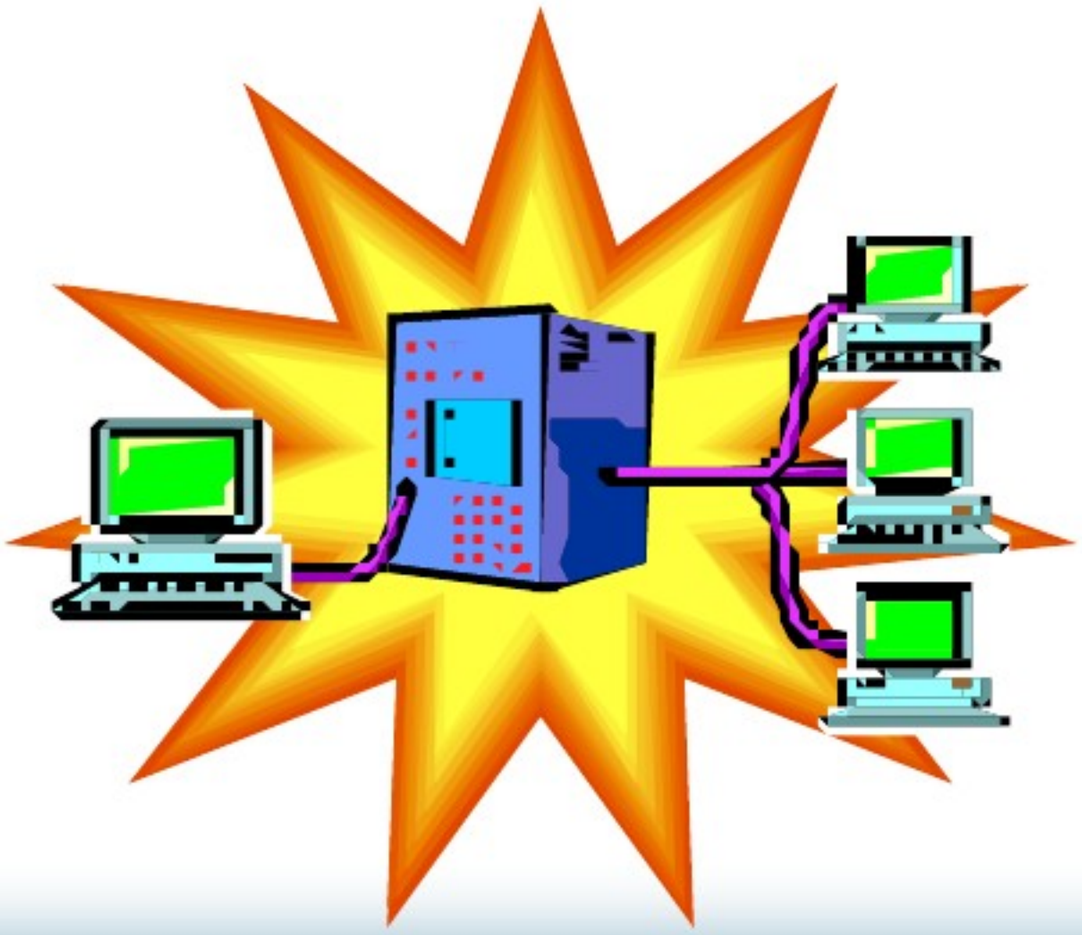
机器学习的重要性

美国航空航天局JPL实验室的科学家在《Science》（2001年9月）上撰文指出：机器学习对科学研究的整个过程正起到越来越大的支持作用，……，该领域在今后的若干年内将取得稳定而快速的发展

例子1：网络安全

入侵检测：

是否是入侵？是何种入侵？



如何检测？

•历史数据：以往的正常访问模式及其表现、以往的入侵模式及其表现……

•对当前访问模式分类

这是一个典型的机器学习问题

常用技术：

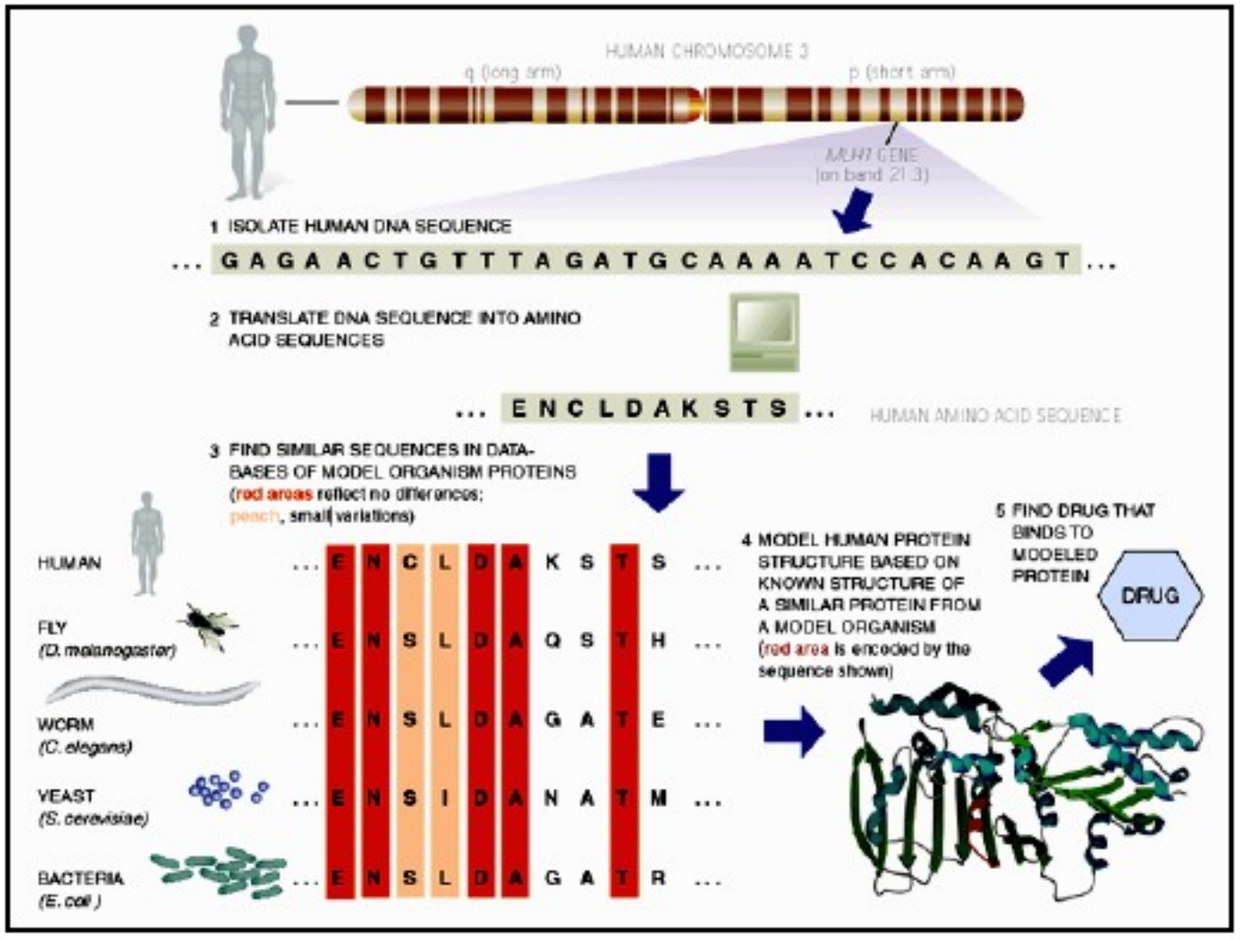
神经网络　　决策树

支持向量机　贝叶斯分类器

k近邻　　序列分析　　聚类

……　　……

例子2：生物信息学



常用技术：

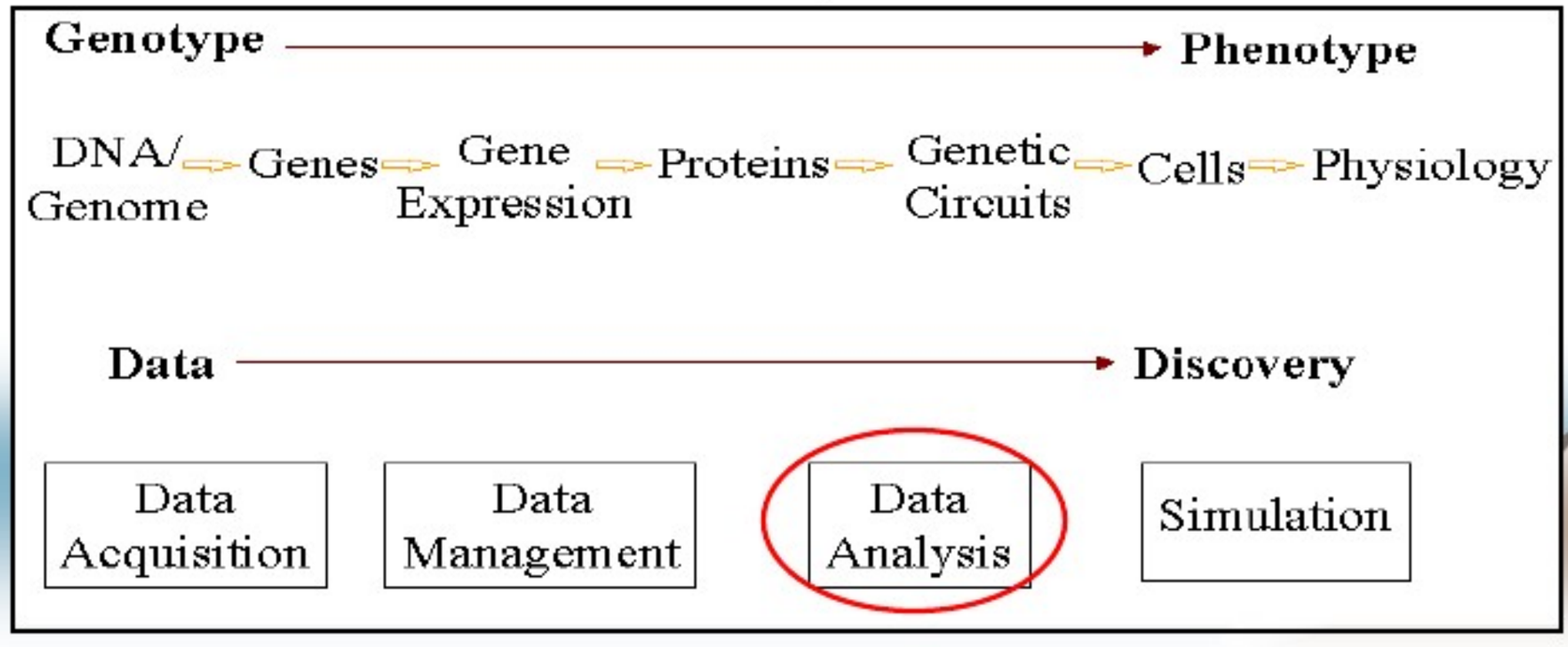
神经网络　支持向量机

隐马尔可夫模型

贝叶斯分类器　k近邻

决策树　序列分析　聚类

.....　.....



例子3：搜索引擎

Google的成功，使得Internet搜索引擎成为一个新兴的产业

不仅有众多专营搜索引擎的公司出现（例如专门针对中文搜索的就有慧聪、百度等），而且Microsoft等巨头也开始投入巨资进行研发

Google掘到的第一桶金，来源于其创始人Larry Page和Sergey Brin提出的PageRank算法

机器学习技术正在支撑着各类搜索引擎（尤其是贝叶斯学习技术）

美国的PAL计划

DARPA 2003 年 开 始 启 动 PAL （ Perceptive Assistant that Learns）计划

5年期，首期（1-1.5年）投资2千9百万美元

以机器学习为核心的计划（涉及到AI的其他分支，如知识表示和推理、自然语言处理等）；包含2个子计划

目标：

“is expected to yield new technology of significant value to the military, business, and academic sectors”

“develop software that will help decision-makers manage their complex worlds of multiple simultaneous tasks and unexpected events”

美国的PAL计划：RADAR子计划

RADAR (Reflective Agents with Distributed Adaptive Reasoning)，承担单位为CMU，首期7百万美元

目标：“the system will help busy managers to cope with time-consuming tasks”

“RADAR must learn by interacting with its human master and by accepting explicit advice and instruction”

美国的PAL计划：CALO子计划(1)

CALO (Cognitive Agent that Learns and Observes)，承担单位为SRI，首期2千2百万美元

除SRI外，这个子计划的参加单位有20家：

Boeing, CMU, Dejima Inc., Fetch Tech Inc., GATech, MIT, Oregon HSU, Stanford, SUNY-Stony Brook, UC Berkeley, UMass, UMich, UPenn, Rochester, USC, UT Austin, UW, Yale, ...

CALO无疑是PAL中更核心的部分

美国的PAL计划：CALO子计划 (2)

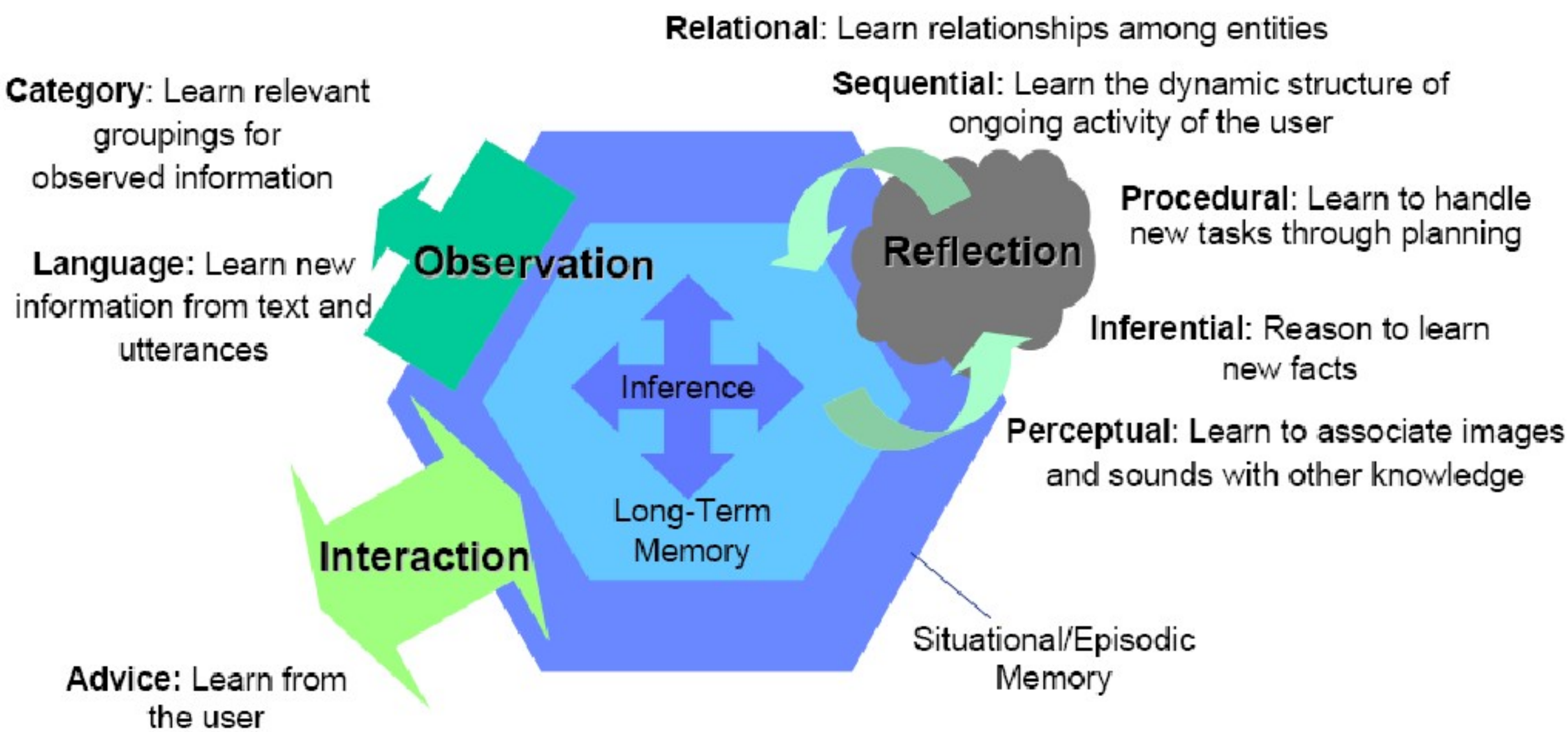
目标： “the name CALO was inspired by the Latin word ‘calonis’, which means ‘soldier’s assistant’”

“the CALO software, which will learn by working with and being advised by its users, will handle a broad range of interrelated decision-making tasks ... It will have the capability to engage in and carry out routine tasks, and to assist when the unexpected happens”

从CALO的目标来看，DARPA已经开始把机器学习技术的重要性放到了国家安全的角度来考虑

美国的PAL计划：CALO子计划(3)

CALO Learning

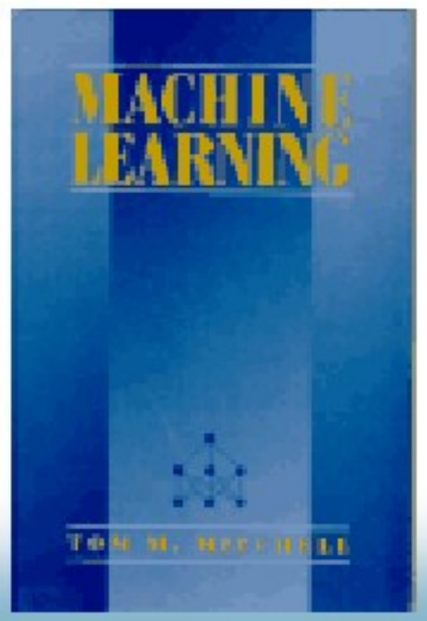


历史回顾(1)

下述事件（大致）标志着机器学习正式成为一个学科

- 1983年，R.S. Michalski等人撰写《机器学习：通往人工智能的途径》一书
- 1986年，Machine Learning杂志创刊

与人工智能乃至计算机科学中很多其他分支学科相比，机器学习还非常年轻、很不成熟



以Tom Mitchell的经典教科书（McGraw Hill出版社，1997）为例，很难看到基础学科（例如数学、物理学）教科书中那种贯穿始终的体系，也许会让人感到这不过是不同方法和技术的堆砌

历史回顾(2)

主要范式的发展：

- 80年代中叶以前：符号主义，代表：ILP

  受到传统人工智能研究的深刻影响，以逻辑推理为基础

- 80年代中叶至90年代初：连接主义，代表：NN

对传统人工智能的批评：“看上去漂亮，但解决不了实际问题”

对上述批评，AI的不同分支学科实际上都做出了自己的回应，ML的回应是连接主义受到重视

NN并不漂亮（至少在理论体系上远远没有ILP那么漂亮），但解决了很多实际问题

历史回顾(3)

- 90年代中叶至今：统计学习，代表：SVM

  NN虽然解决了不少问题，但解决问题时的“试错性”引来了“trick”的批评

  作为回应，统计学习开始占据支配地位。虽然SVM仍然有“试错性”，但毕竟在理论基础上比NN漂亮得多（实际上，统计学习与连接主义一脉相承）

- 现在：？

  统计学习并不是万能的，有很多问题不能解决（或不能很好地解决），例如结构化数据的学习

  作为回应，以逻辑为基础的符号主义与统计学习的结合开始受到重视

似乎的趋势—“普适机器学习”

从主要范式的发展可以看出，ML实际上是一个应用驱动的学科，其根本的驱动力是“更多、更好地解决实际问题”

由于近20年的飞速发展，机器学习已经具备了一定的解决实际问题的能力，似乎逐渐开始成为一种基础性、透明化的“支持技术、服务技术”

基础性：在众多的学科领域都得以应用（“无所不在”）

透明化：用户看不见机器学习，看见的是防火墙、生物信息、搜索引擎；（“无所不在”）

“机器更好用了”(正如CALO的一些描述：“you won’t leave home without it”；” embodied as a software environment that transcends workstations, PDA’s, cell phones, ...”)

挑战与机遇

作为支持和服务技术的“普适机器学习”带来了挑战和机遇：

- 出现了很多被传统ML研究忽视、但非常重要且尚无好的解决方案的问题（下面将以医疗和金融为代表来举几个例子）
- ML支持和服务的学科领域越多，新问题越多
- ML与众多学科领域产生了交叉，而交叉领域正是大有可为处

例子1：代价敏感

医疗：以乳腺癌诊断为例，“将病人误诊为健康人的代价”与“将健康人误诊为病人的代价”是不同的

金融：以信用卡盗用检测为例，“将盗用误认为正常使用的代价”与“将正常使用误认为盗用的代价”是不同的

传统的ML技术基本上只考虑同一代价

如何处理代价敏感性？

在教科书中找不到现成的答案，例如：

Tom Mitchell, Machine Learning, McGraw-Hill, 1997

Nils J. Nilsson, Introduction to Machine Learning, draft 1996 - 2004

例子2：不平衡数据

医疗：以乳腺癌诊断为例，“健康人”样本远远多于“病人”样本

金融：以信用卡盗用检测为例，“正常使用”样本远远多于“被盗用”样本

传统的ML技术基本上只考虑平衡数据

如何处理数据不平衡性？

在教科书中找不到现成的答案

例子3：可理解

医疗：以乳腺癌诊断为例，需要向病人解释“为什么做出这样的诊断”

金融：以信用卡盗用检测为例，需要向保安部门解释“为什么这是正在被盗用的卡”

传统的ML技术基本上只考虑泛化不考虑理解

如何处理可理解性？

在教科书中找不到现成的答案

走向普适机器学习

把机器学习真正当成一种支持技术、服务技术，考虑不同学科领域对机器学习的需求，找出其中具有共性的、必须解决的问题，并进而着手研究

一方面可以促进和丰富ML本身的发展，另一方面可以促进使用ML技术的学科领域本身的发展

作为“应用基础”，与“ML应用”有根本的区别：

- 基础性：不是直接做应用，而是做“更广泛的应用”或“更成功的应用”所需要的方法和技术
- 广泛性：重点不是去解决单一应用所面临的问题，而是要解决众多应用领域所面临的共性问题

致谢

应明生教授：与基础科学教科书的比较

王珏教授：　多次富有启发性的讨论

请各位专家

批评指正！