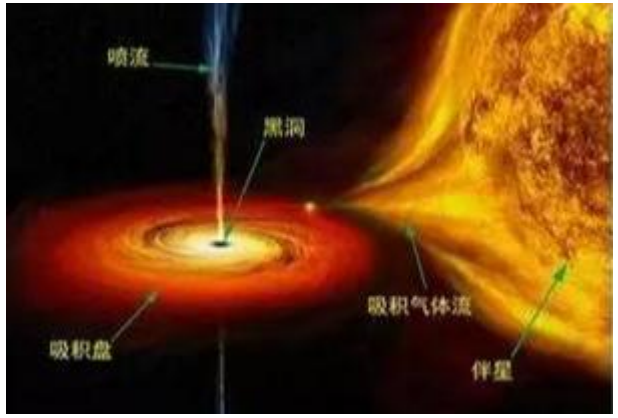
智能时代的天文学

张彦霞

中国科学院国家天文台

2023.4.22@广西.桂林

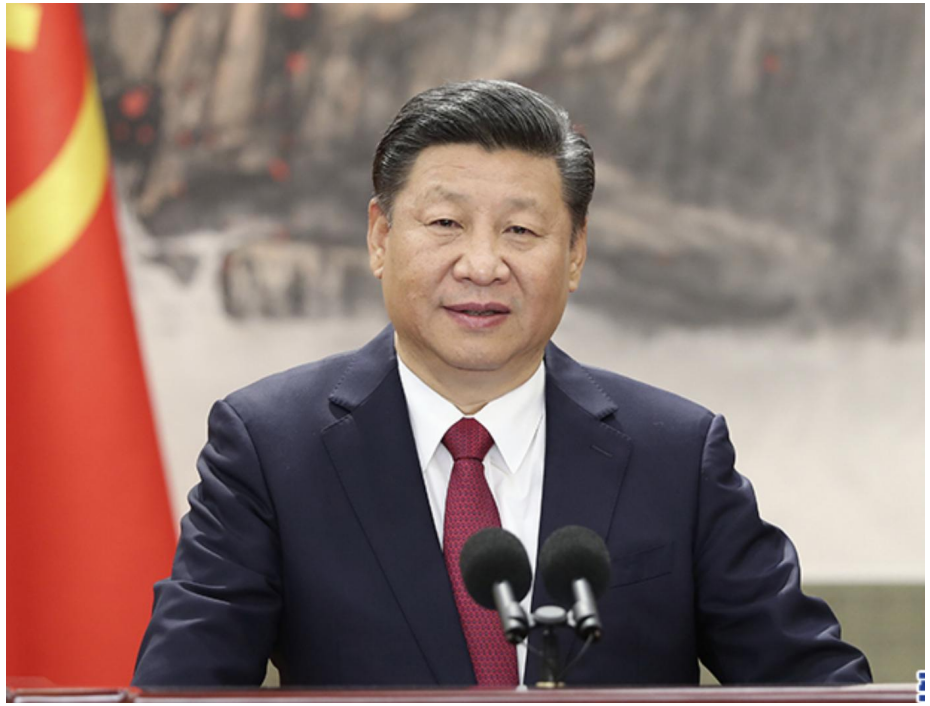


报告提纲

- AI背景
- 天文大数据
- 机器学习
  - 概念
  - 分类
  - 任务
  - 应用
  - 挑战
- 展望

2018年9月17日世界人工智能大会在上海开幕，习近平致信祝贺。“新一代人工智能正在全球范围内蓬勃兴起，为经济社会发展注入了新动能，正在深刻改变人们的生产生活方式。”习近平在贺信中强调，中国正致力于实现高质量发展，人工智能发展应用将有力提高经济社会发展智能化水平，有效增强公共服务和城市管理能力。

2022年习近平总书记指出，“把新一代人工智能作为推动科技跨越发展、产业优化升级、生产力整体跃升的驱动力量，努力实现高质量发展”。党的十八大以来，以习近平同志为核心的党中央高度重视智能经济发展，促进人工智能和实体经济深度融合，为高质量发展注入强劲动力。



在危机中育新机，变局中开新局

习近平总书记强调，人工智能是新一轮科技革命和产业变革的重要驱动力量，加快发展新一代人工智能是事关我国能否抓住新一轮科技革命和产业变革机遇的战略问题。要深刻认识加快发展新一代人工智能的重大意义，加强领导，做好规划，明确任务，夯实基础，促进其同经济社会发展深度融合，推动我国新一代人工智能健康发展。（2023年3月15日人民日报）

人工智能的前世今生

来源：通联数据整理

AI诞生

1954年

第一台可编程机器人诞生

1956年

达特茅斯会议首次提出人工智能（AI）术语

冲破迷雾

1959年

第一台工业机器人出现

1966年

世界上第一个聊天机器人发布

1968年

第一台智能机器人诞生（Shakey）

1984年

启动Cyc（大百科全书）项目

1986年

历史上首个3D打印机问世

黄金时代

1997年

电脑深蓝战胜国际象棋冠军

2002年

家用机器人诞生

2011年

开发出使用自然语言回答问题的人工智能程序

2012年

虚拟大脑Spaun诞生

2013年

深度学习算法广泛应用

2014年

机器人首次通过图灵测试

2016年

AlphaGo战胜围棋世界冠军李世石

2023年

ChatGPT

Midjourney

第一次人工智能浪潮

20世纪
60~70年代

- 60年代：因“推理和探索”时取得重大进展而繁荣。
- 70年代：因“推理和探索”对现实问题束手无策而衰落。

第二次人工智能浪潮

20世纪
80~90年代

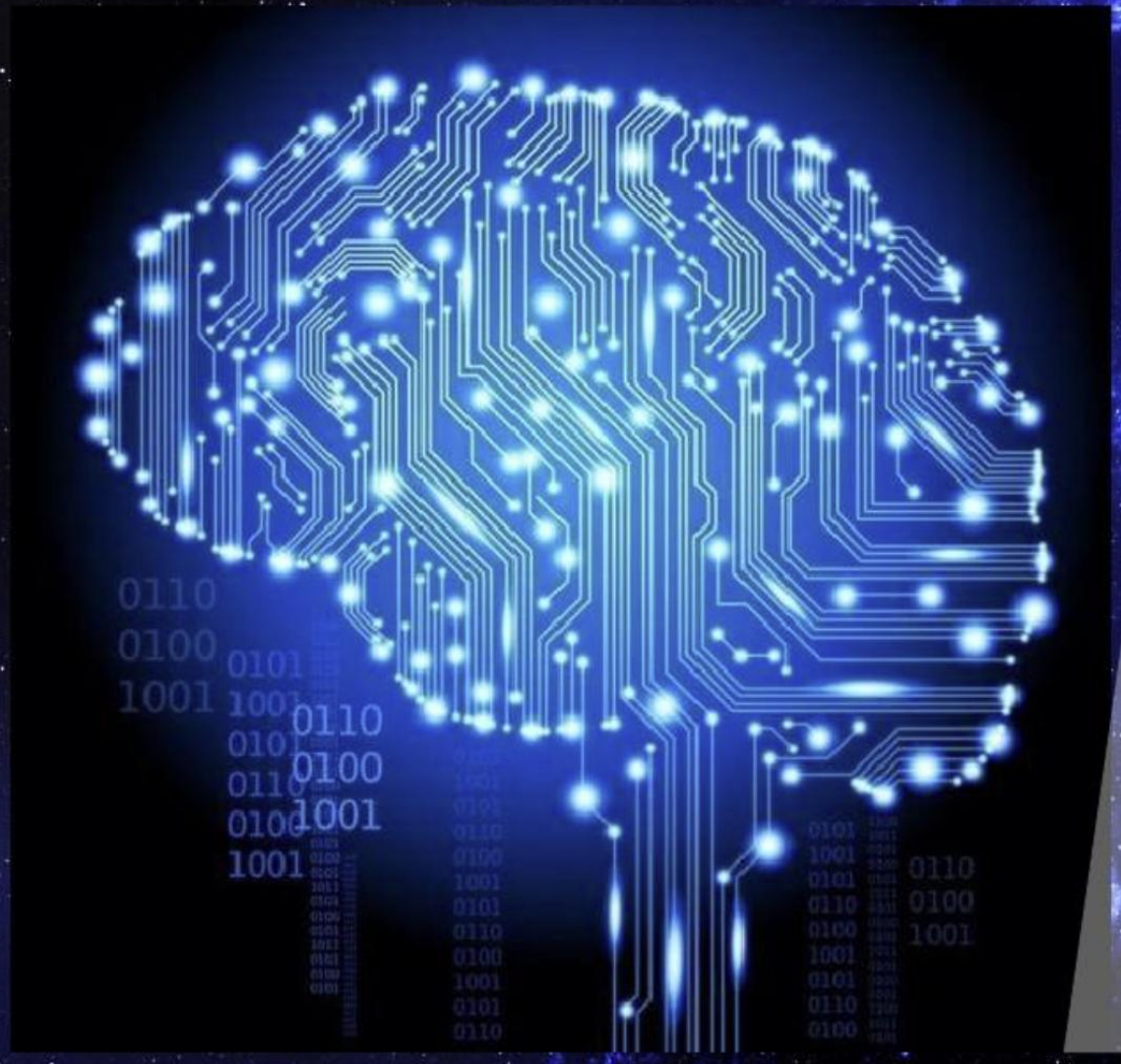
- 因导入知识使机器变得更聪明而繁盛。
- 因知识描述和管理的能力低下而缺陷暴露。

第三次人工智能浪潮

21世纪
至今

- 三大引擎：深度学习、大数据、超强运算能力。

人工智能的定义



人工智能（Artificial Intelligence），英文缩写为AI。它是研究、开发用于模拟、延伸和扩展人的智能的理论、方法、技术及应用系统的一门新的技术科学。人工智能是计算机科学的一个分支，它企图了解智能的实质，并生产出一种新的能以人类智能相似的方式做出反应的智能机器，该领域的研究包括机器人、语言识别、图像识别、自然语言处理和专家系统等。人工智能从诞生以来，理论和技术日益成熟，应用领域也不断扩大，可以设想，未来人工智能带来的科技产品，将会是人类智慧的“容器”。

人工智能是对人的意识、思维的信息过程的模拟。人工智能不是人的智能，但能像人那样思考、也可能超过人的智能。


Artificial Intelligence
能为我们做什么？

人工智能

让机器像人一样思考

- ■机器学习
- ■自动推理
- ■人工意识
- ■知识展示

……

让机器像人一样听懂

- ■语音识别

让机器像人一样看懂

- ■视觉识别

让机器像人一样运动

- ■运动控制

人工智能是一门前沿的综合学科

人工智能可以代替人类实现多种功能

计算机科学

识别

统计学

认知

脑神经学

分析

社会科学

决策

人工智能产业链

来源：通联数据整理

应用场景

关键技术 深度学习/机器学习、计算机视觉、语音技术/自然语言处理、人机交互

基础支撑 传感器、芯片、大数据、云计算、服务商、生物识别

颠覆性
涌现性
工程化
通用性
密集型

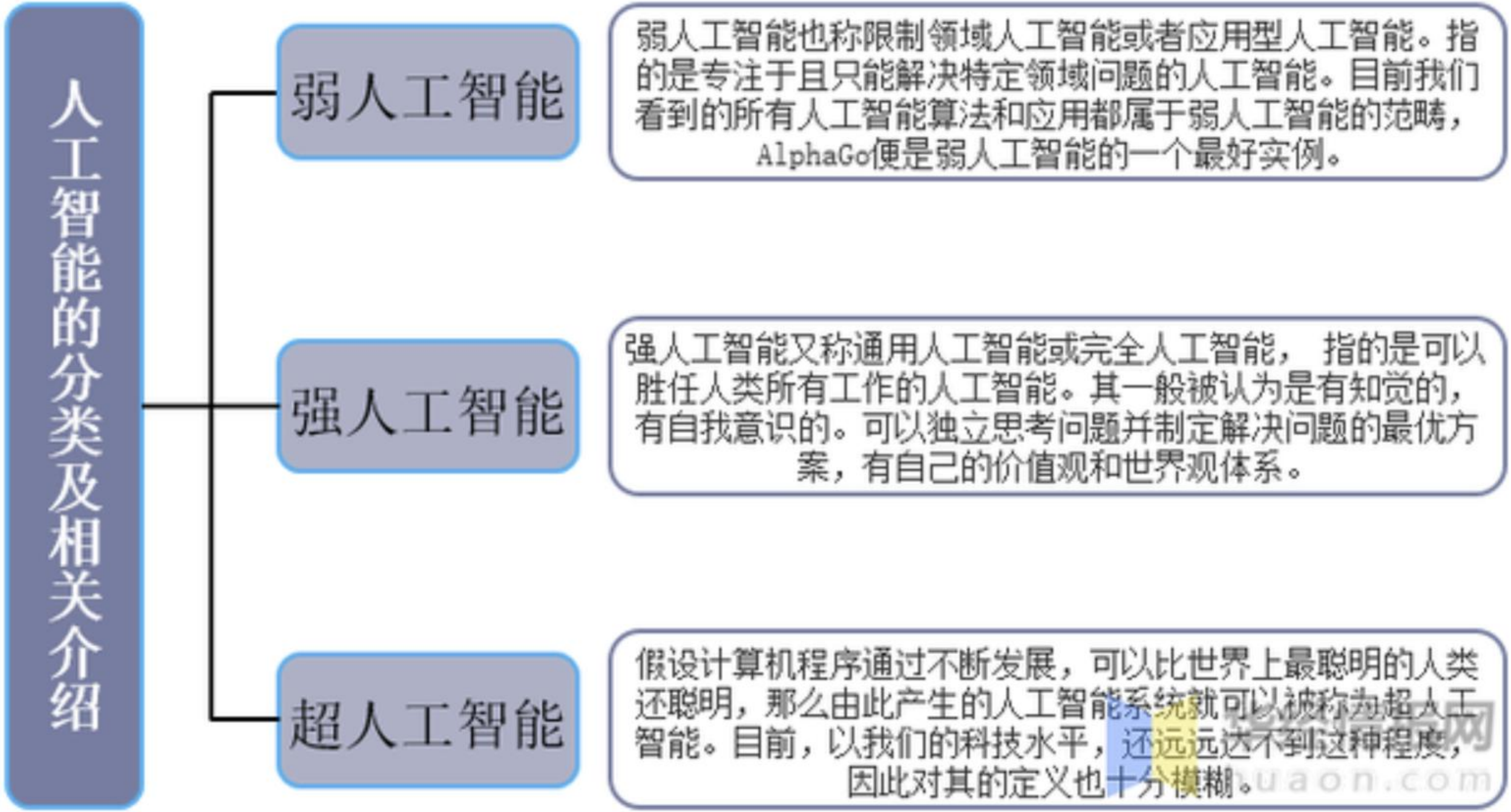
本质差距：
创新机制
创新生态
创新文化

有志者、事竟成，破釜沉舟，百二秦关终属楚；
苦心人、天不负，卧薪尝胆，三千越甲可吞吴。
----清代.蒲松龄.斋联

人工智能时代已经来临

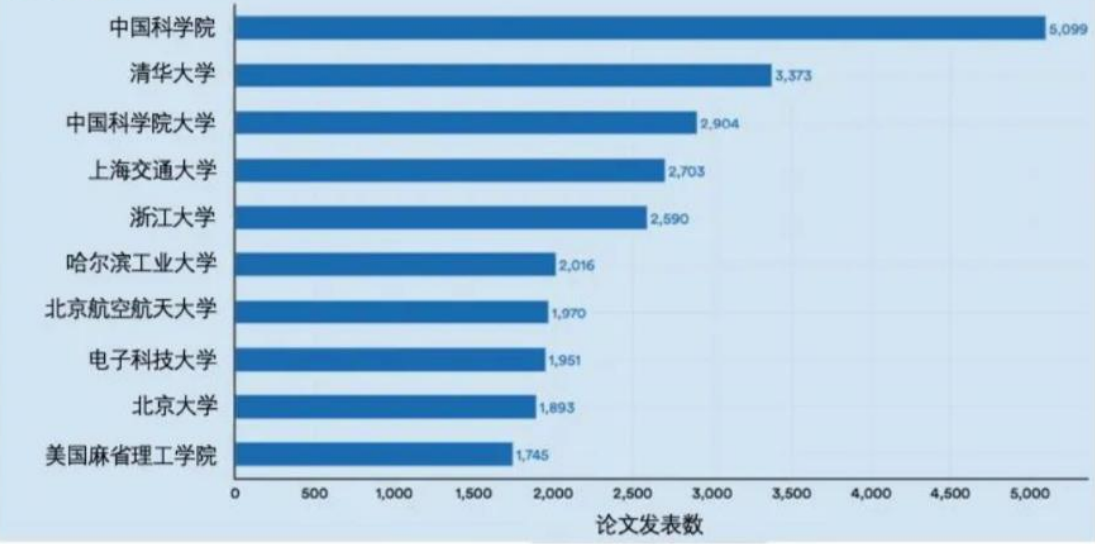
人工智能将引领下一波计算浪潮。与之前的重大转型类似，仍智能将构建出更加美好的时间

人工智能的潜力将大大激发，为企业和社会创造更积极的影响

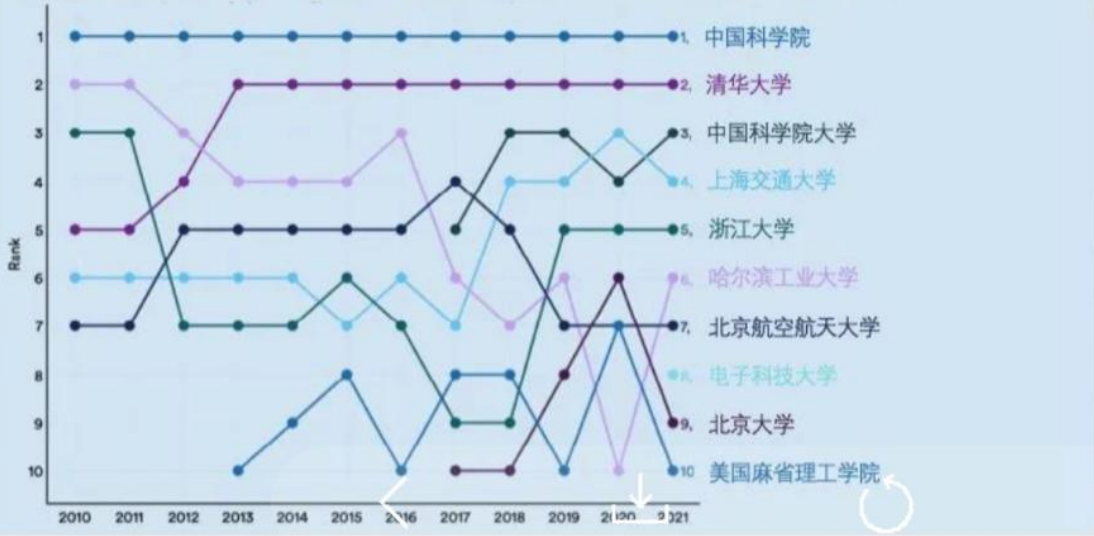

人工智能的分类及相关介绍
弱人工智能
弱人工智能也称限制领域人工智能或者应用型人工智能。指的是专注于且只能解决特定领域问题的人工智能。目前我们看到的所有人工智能算法和应用都属于弱人工智能的范畴，AlphaGo便是弱人工智能的一个最好实例。
强人工智能
强人工智能又称通用人工智能或完全人工智能， 指的是可以胜任人类所有工作的人工智能。其一般被认为是有知觉的，有自我意识的。可以独立思考问题并制定解决问题的最优方案，有自己的价值观和世界观体系。
超人工智能
假设计算机程序通过不断发展，可以比世界上最聪明的人类还聪明，那么由此产生的人工智能系统就可以被称为超人工智能。目前，以我们的科技水平，还远远达不到这种程度，因此对其的定义也十分模糊。

中国： AI 论文大户

2021 年，AI 论文发表量全球 TOP 10 机构中，**中国机构占据 9 席**，美国麻省理工学院排名第十。



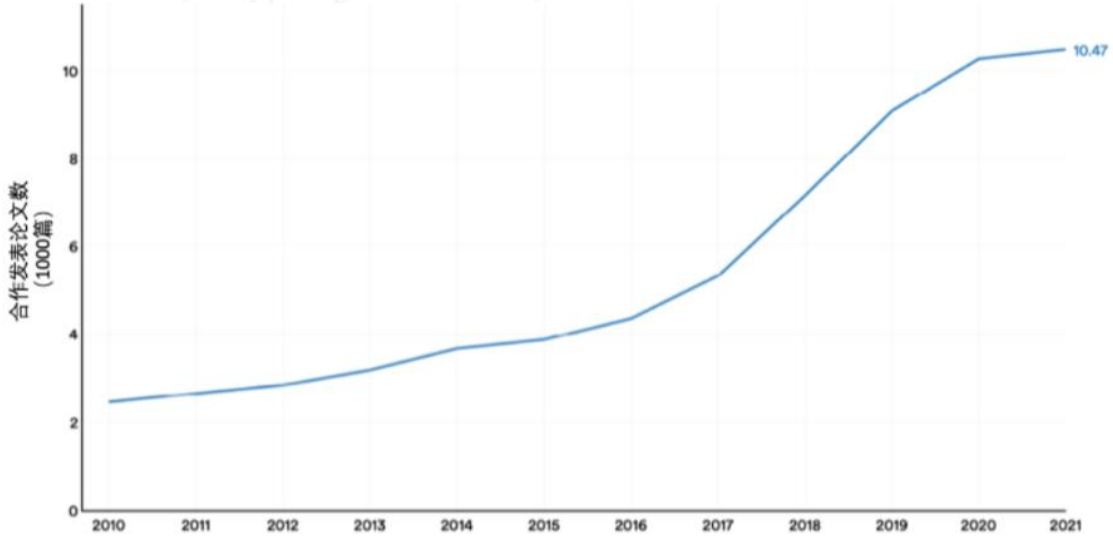
从 2010 年起，**中国科学院**就一直占据 AI 论文发表量世界第一的宝座。



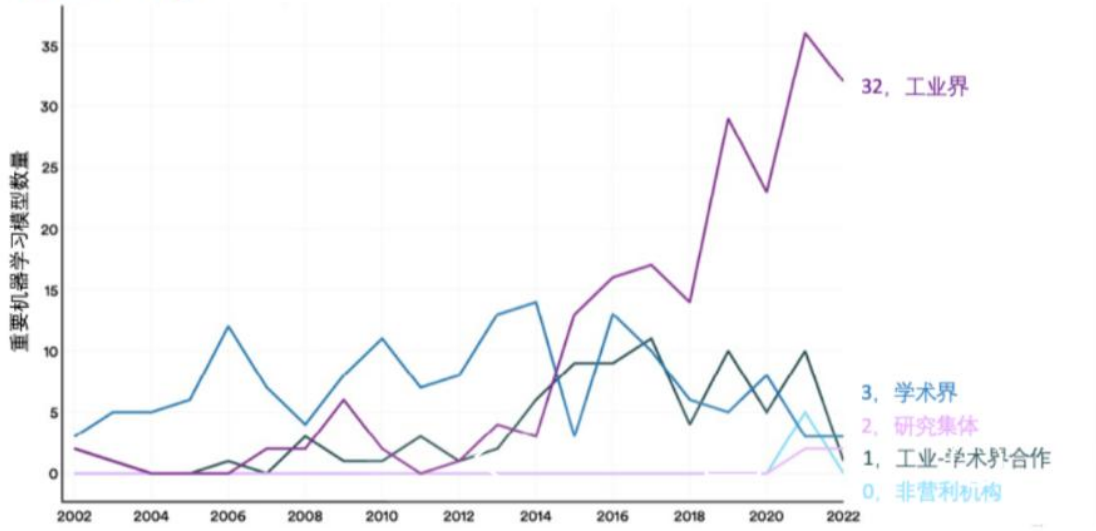
来源：“The AI Index 2023 Annual Report,” Stanford University, April 2023.

中美合作放缓　学界与业界差距加大

中美 AI 合作研究数量持续上升，但**增速放缓。**

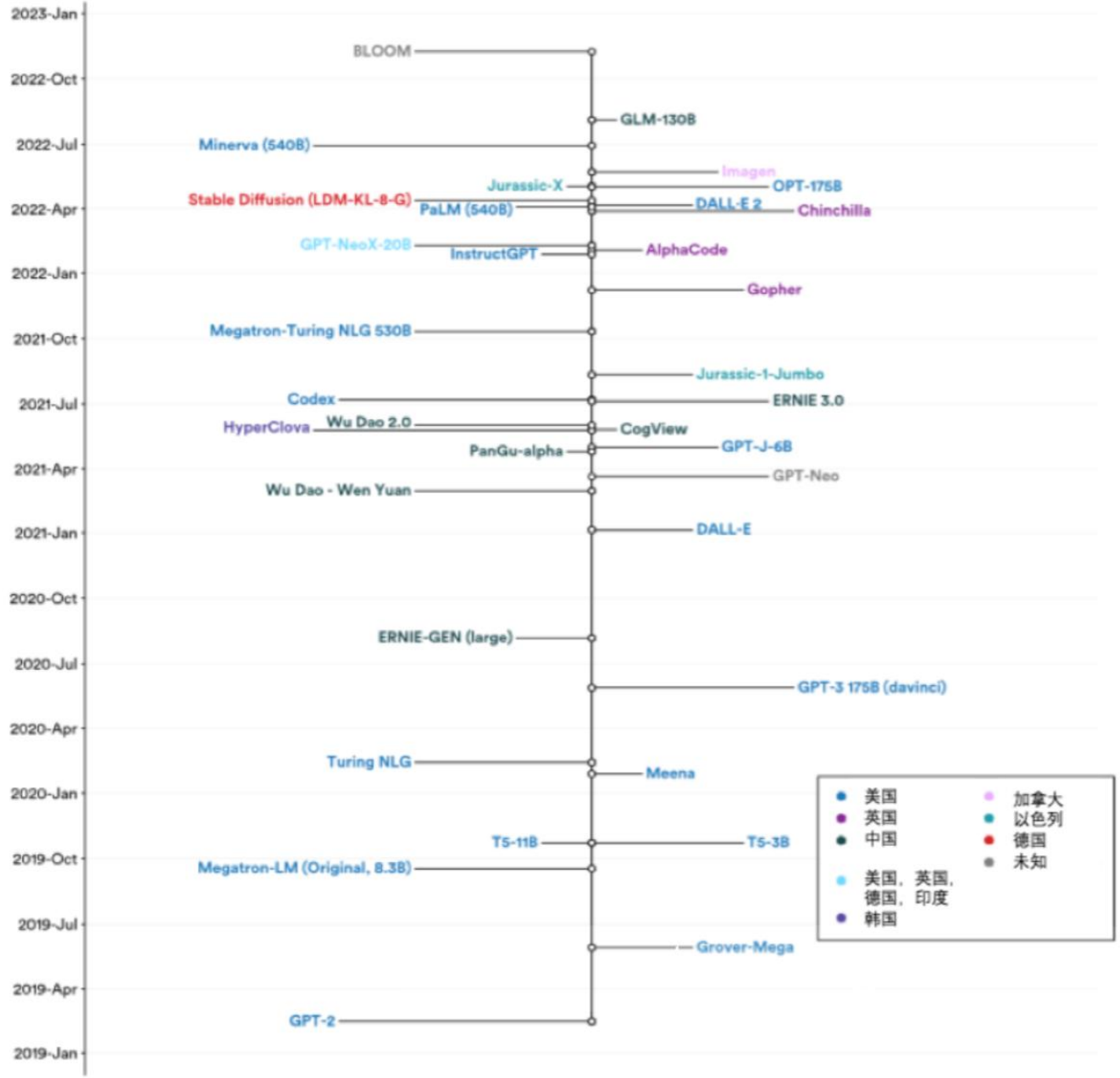


工业界领先于学术界：2022 年，**工业界贡献了 32 个重要机器学习模型**，而学术界只创造了 **3 个。**



来源：“The AI Index 2023 Annual Report,” Stanford University, April 2023.

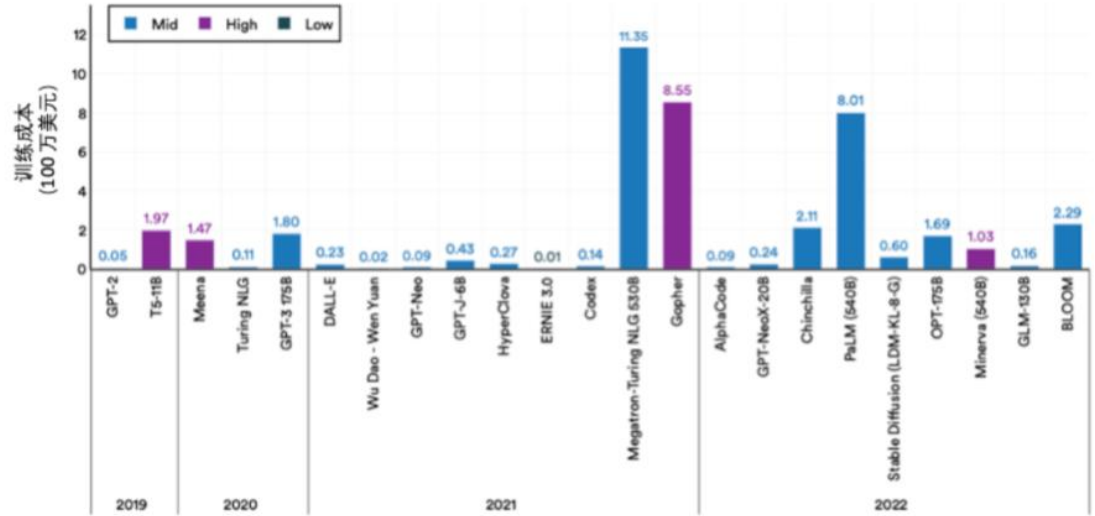
2021 年起，大型语言和多模态模型的“玩家”越来越多。



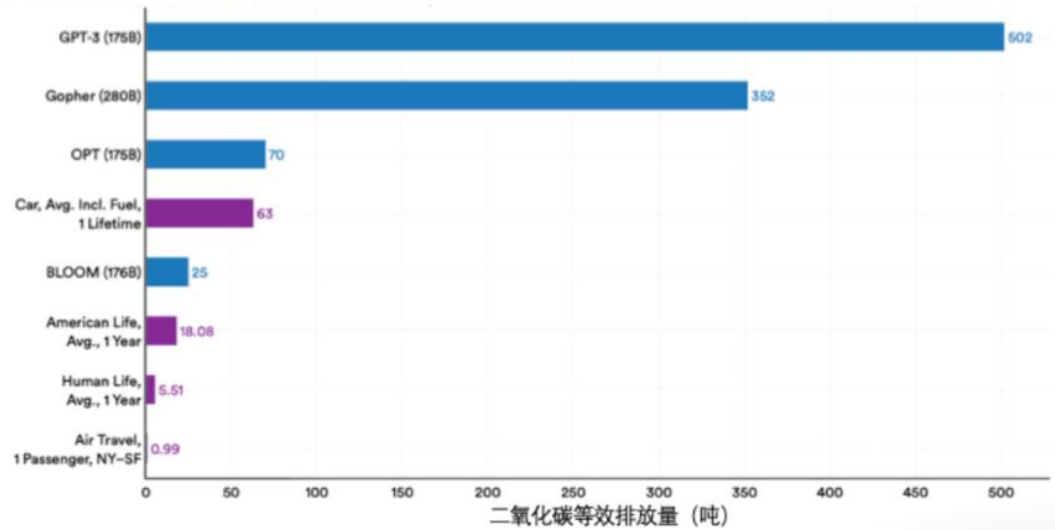
来源：“The AI Index 2023 Annual Report,” Stanford University, April 2023.

越来越大，越来越烧钱

2022 年发布的 **PaLM** 有 **5400 亿参数**，**成本估计为 800 万美元**——参数比 3 年前发布的 GPT-2 大了 360 倍，预估成本高了 160 倍。



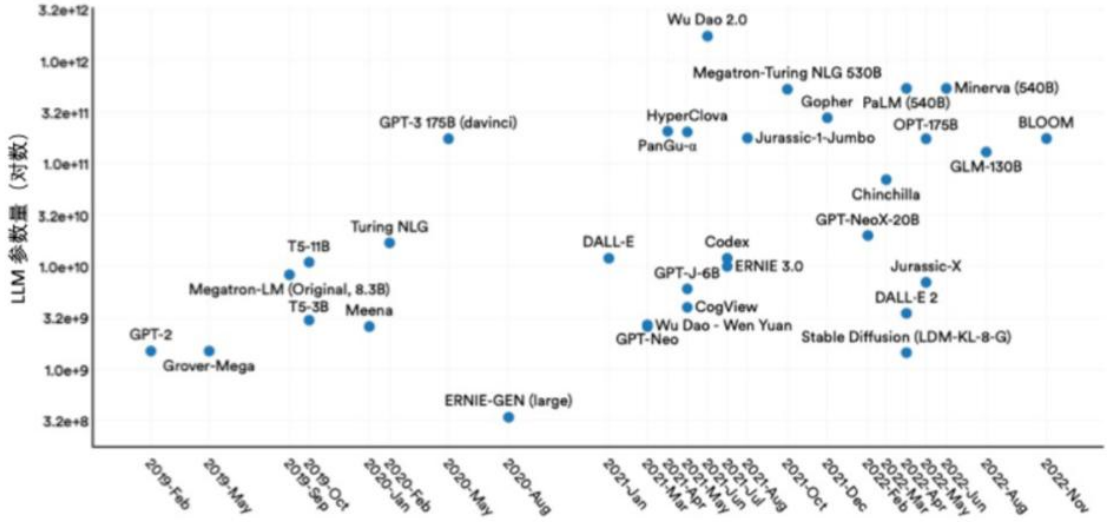
GPT-3 一年的**二氧化碳排放量**达到了 502 吨，相当于一名旅客从纽约飞往旧金山产生排放量的 500 多倍。



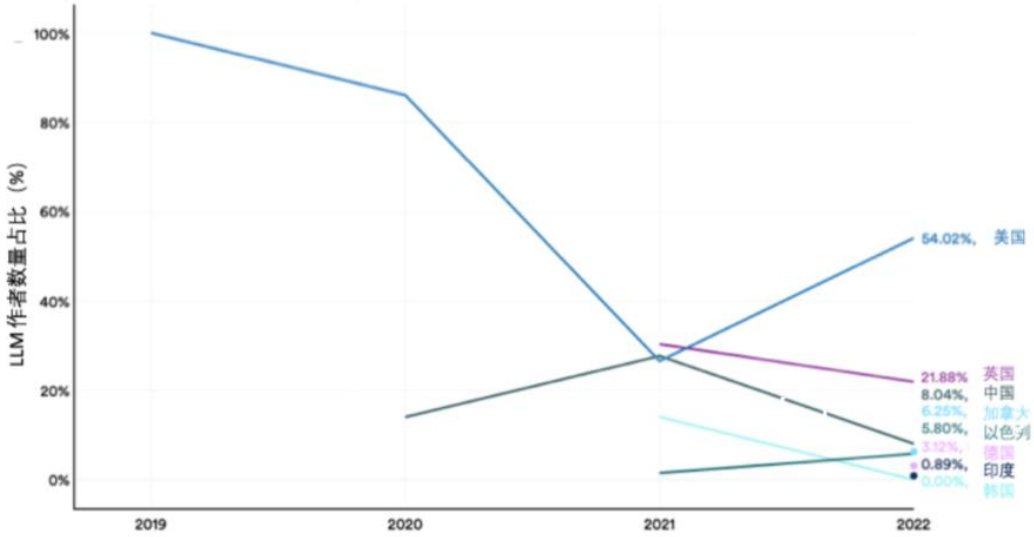
来源：“The AI Index 2023 Annual Report,” Stanford University, April 2023.

谁在领先？

大型语言模型的体量竞争：**中国的“悟道 2.0”**（Wu Dao 2.0）是目前参数量最大的 LLM。



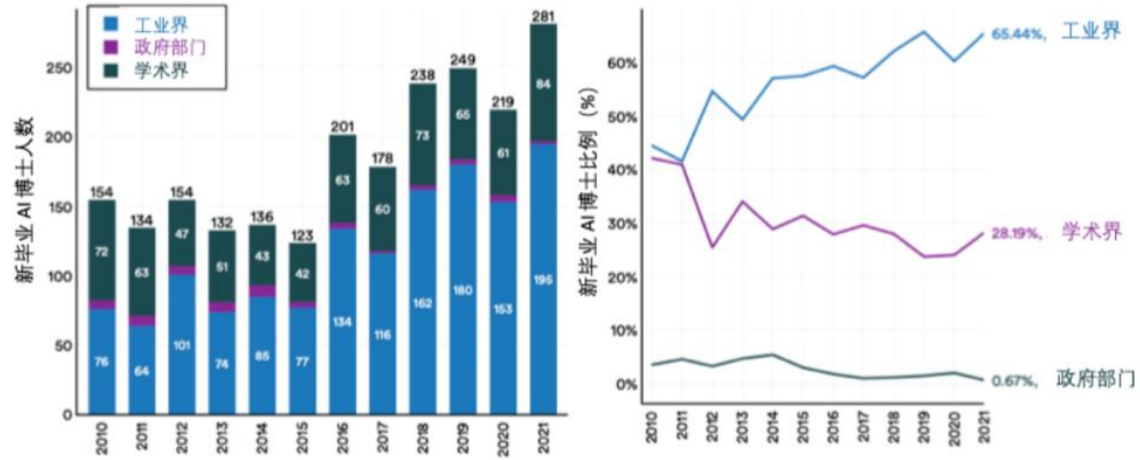
该领域来自**美国的论文作者最多**（占总数的 54.2%），中国作者数量在美国和英国之后。



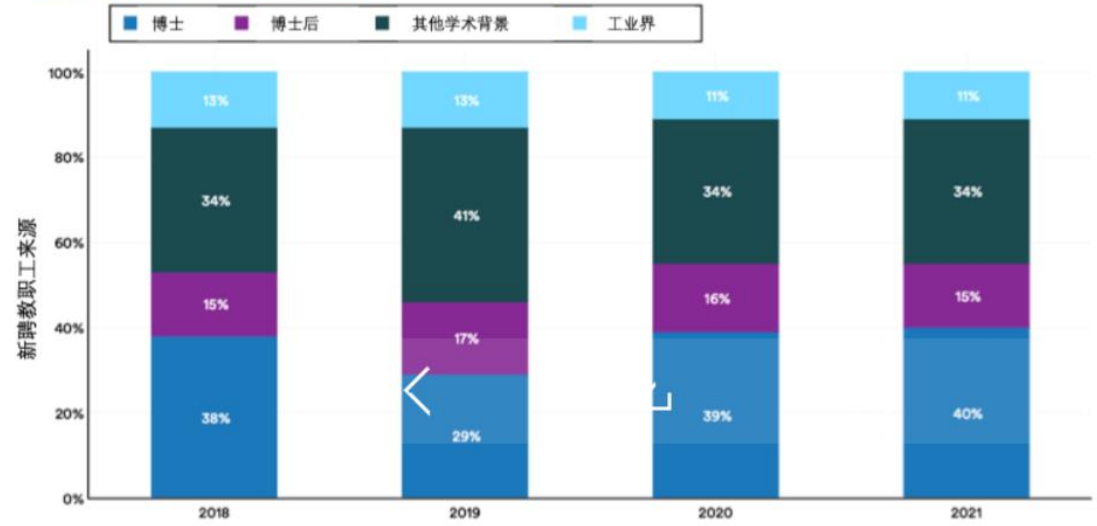
来源：“The AI Index 2023 Annual Report,” Stanford University, April 2023.

学术界对博士的吸引力不足

越来越多的博士**选择了工业界**。2011 年，北美 AI 专业博士毕业生在工业界（40.9%）和学术界（41.6%）的工作比例大致相同。**2021 年，65.4% 的人在工业界工作**，比在学术界工作的 28.2% **多一倍**。



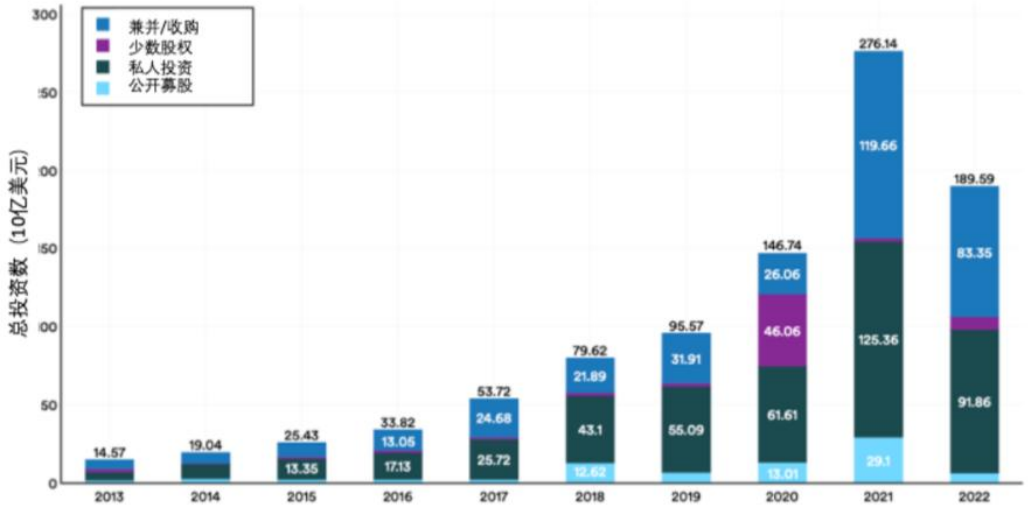
2021 年，北美计算机和信息领域新聘用的教职工**有 40% 是没有博士后经历的毕业生**，仅有约十分之一来自工业界。



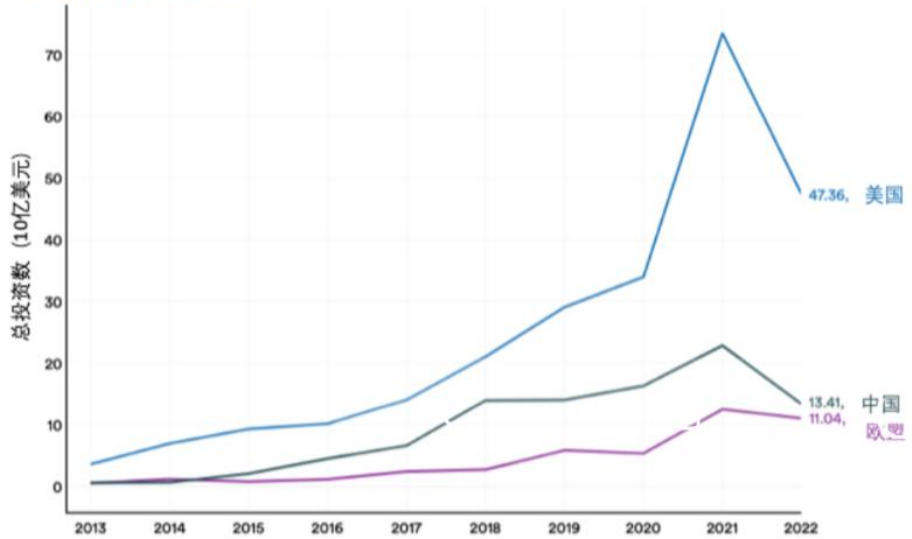
来源：“The AI Index 2023 Annual Report,” Stanford University, April 2023.

谁在投资？

私人投资是全球 AI 投资的主要力量之一。2022 年的私人投资总量为 **919 亿美元**，较 2021 年**下降 26.7%**。



2022 年，美国在 AI 领域的私人投资为 **470 亿美元**，大约是排名第二的中国（130 亿美元）的 **3.5 倍**。



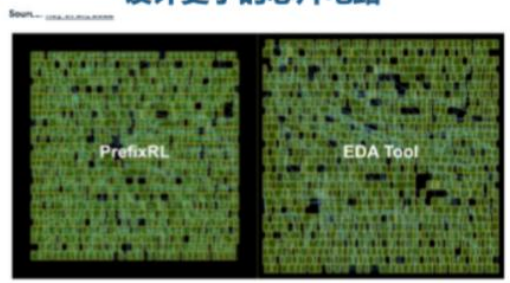
来源："The AI Index 2023 Annual Report," Stanford University, April 2023.

是万能助手，还是闯祸精？

· Deepmind 强化学习算法**管理核聚变**

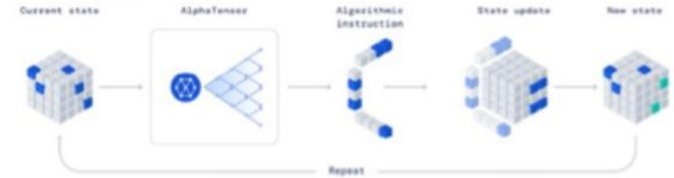
Source: DeepMind, 2022

· 深度强化学习模型 PrefixRL **设计更小的芯片电路**



· AlphaTensor **发现矩阵运算新算法**

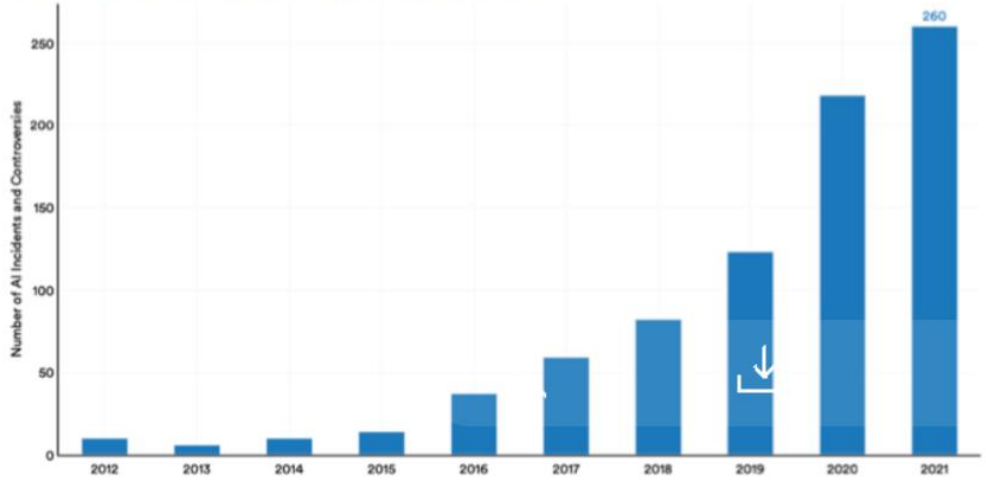
Source: Fawzi et al., 2022



· Zero-Shot 生成式 AI **开启新抗体设计的大门**

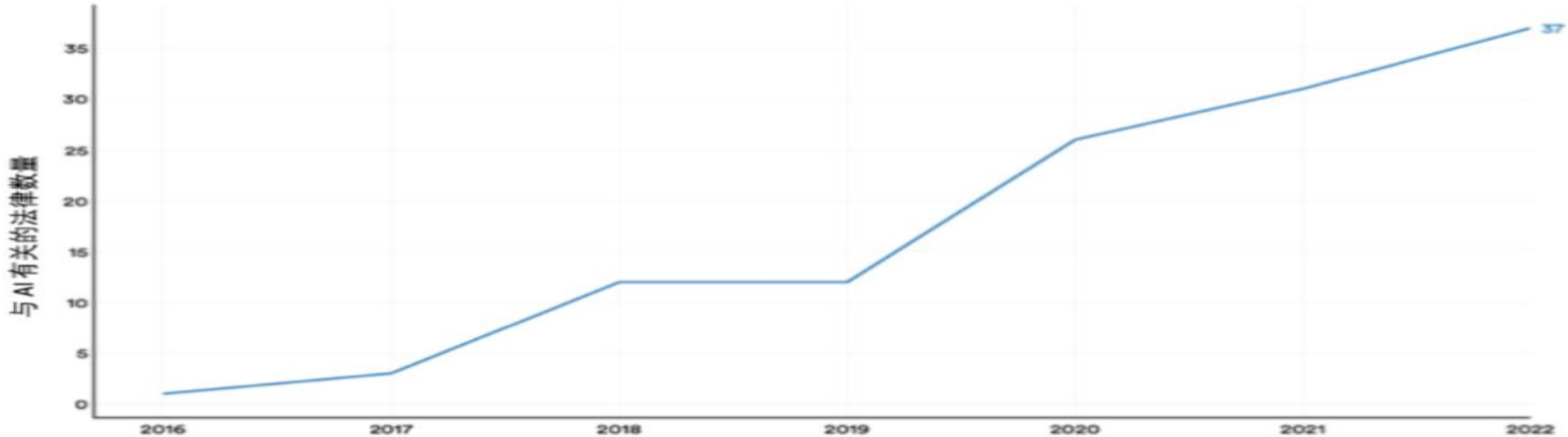
……

2021 年，AI 带来的**事故和争议**是 2012 年的 **26 倍**。AI 与现实世界的交融程度越来越高，人们对 AI 被滥用的认识也逐渐深刻。

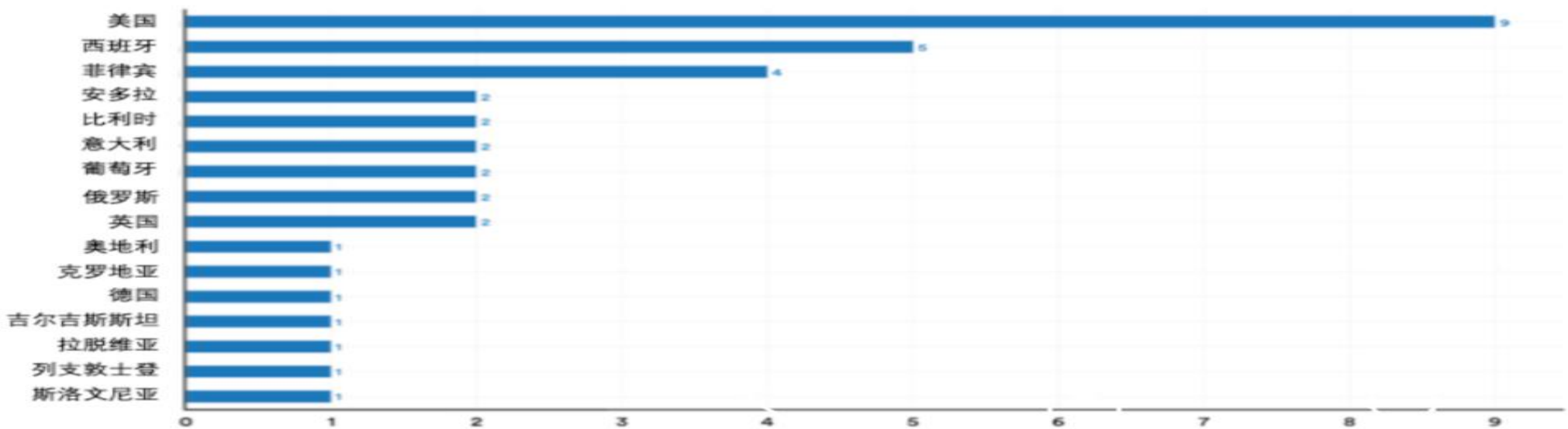


来源："The AI Index 2023 Annual Report," Stanford University, April 2023.

全球与 AI 有关的**法律数量越来越多**。2016 年接受调查的 127 个国家仅有 1 项相关法律，2022 年这一数字涨到了 **37 项**。



2022 年全球共通过了 **37 项与 AI 有关的法律**。其中美国通过了 9 项，数量最多。



来源："The AI Index 2023 Annual Report," Stanford University, April 2023.

数据

开放共享 | 全链条 | 隐私安全

AI基础设施

AI INFRASTRUCTURE

算力

智算中心 | 异构计算 | 弹性大规模集群

算法

开源开放 | 高效 | 流水线

大模型生产及部署应用

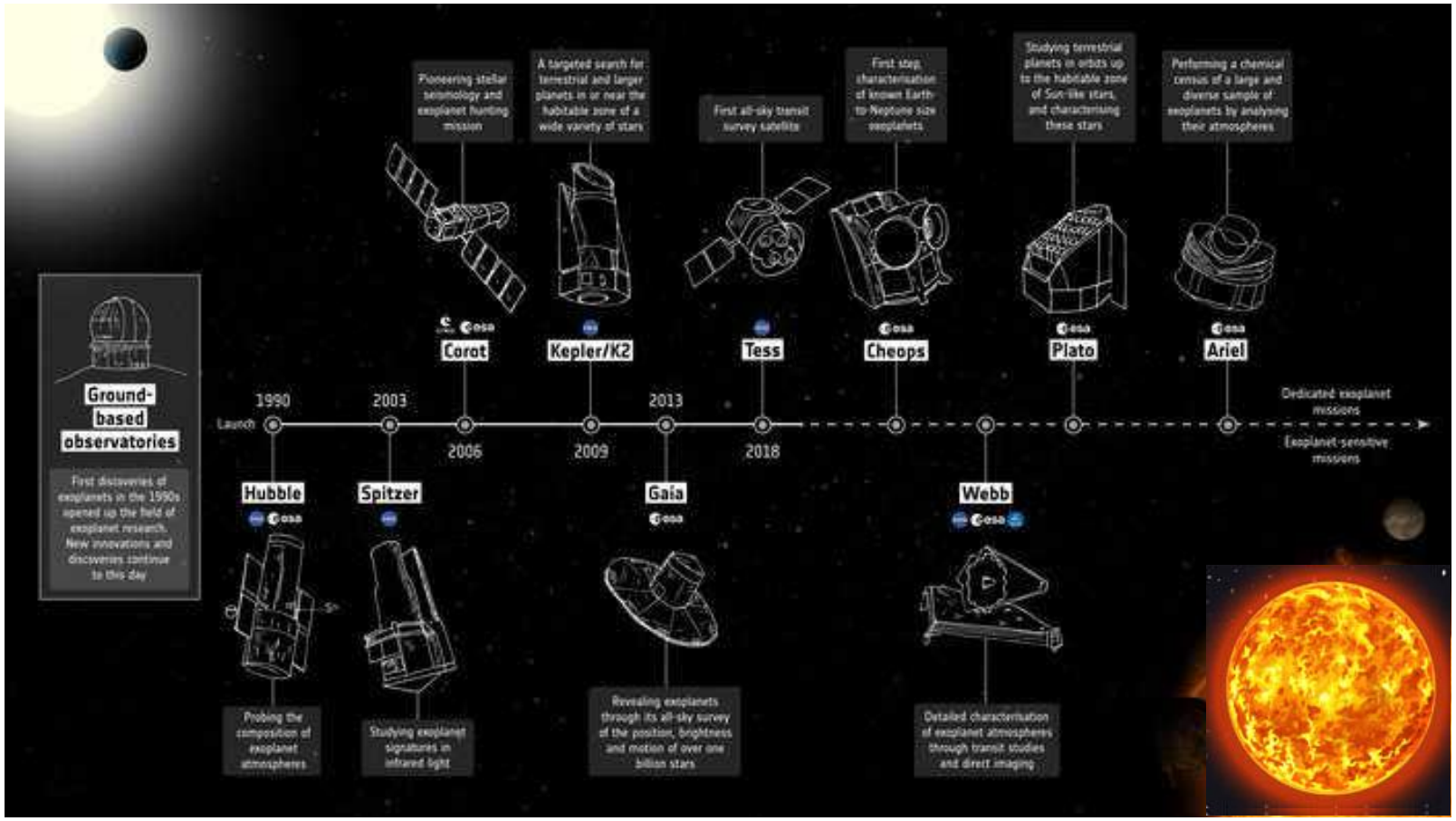
天文学呈现新面貌：全波段、多信使、时域、大数据

特点：

“广”

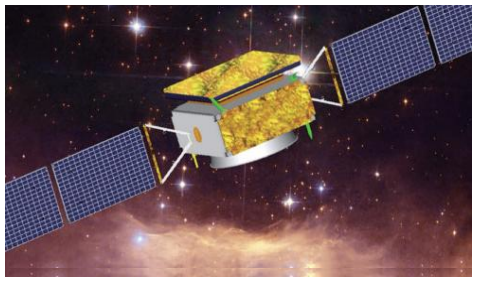
“深”

“快”

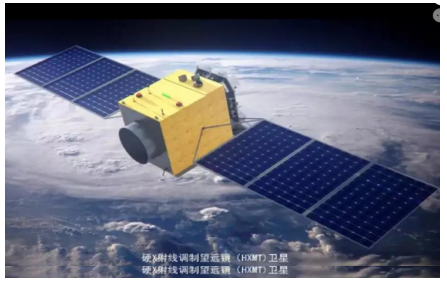


LAMOST

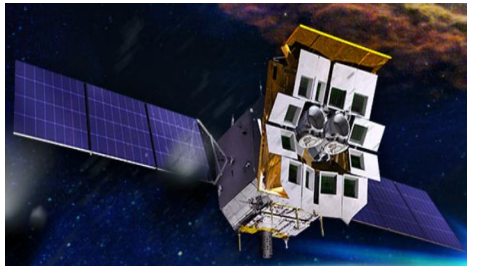
FAST



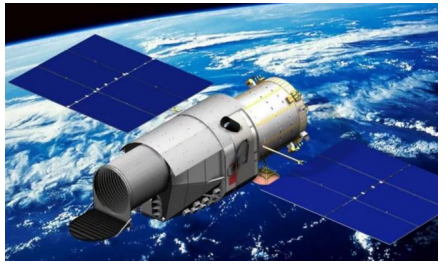
悟空



慧眼



爱因斯坦探针

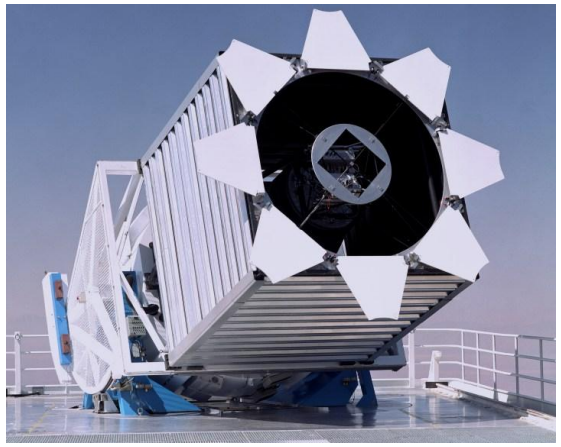


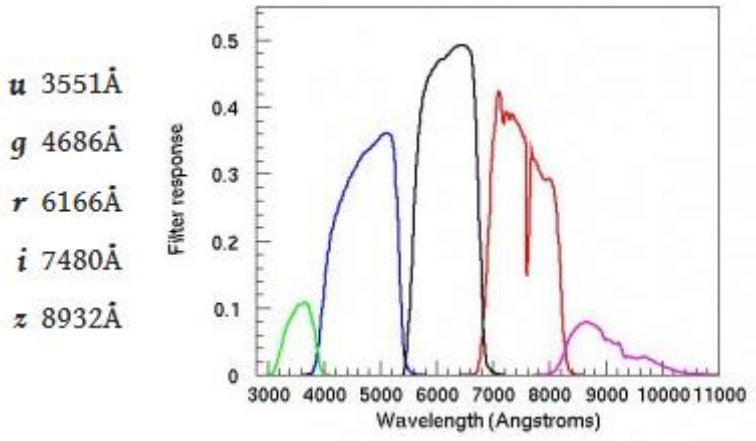
中国空间站望远镜

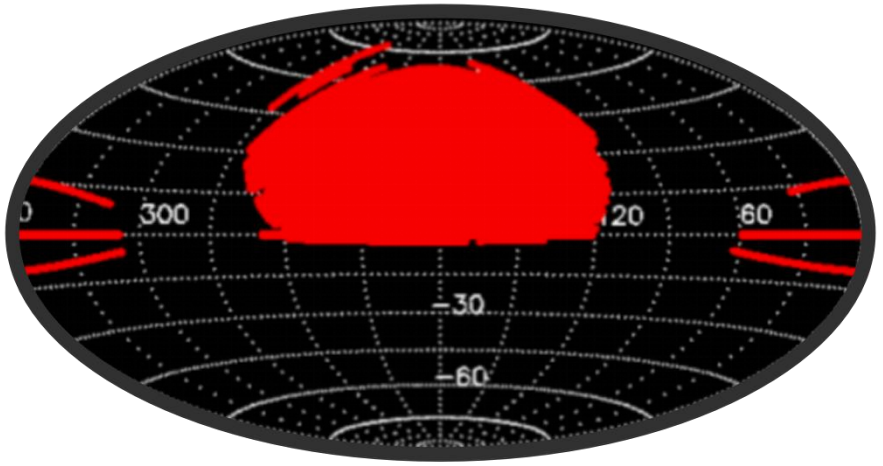
Sloan Digital Sky Survey

望远镜：2.5m　　覆盖面积：>14500平方度　天测精度：0.1''　极限星等：r<22.5

图像源数：> 460M　光谱源数：Millions





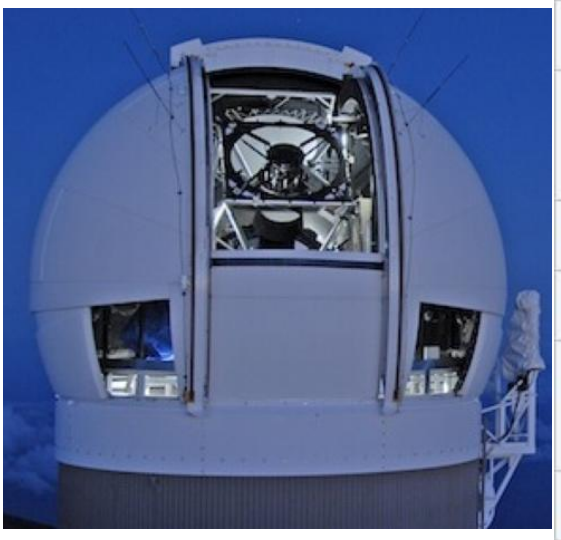


Human Genome < 1 GB
Human Memory < 1 GB (?)
1 TB ~ 2 million books

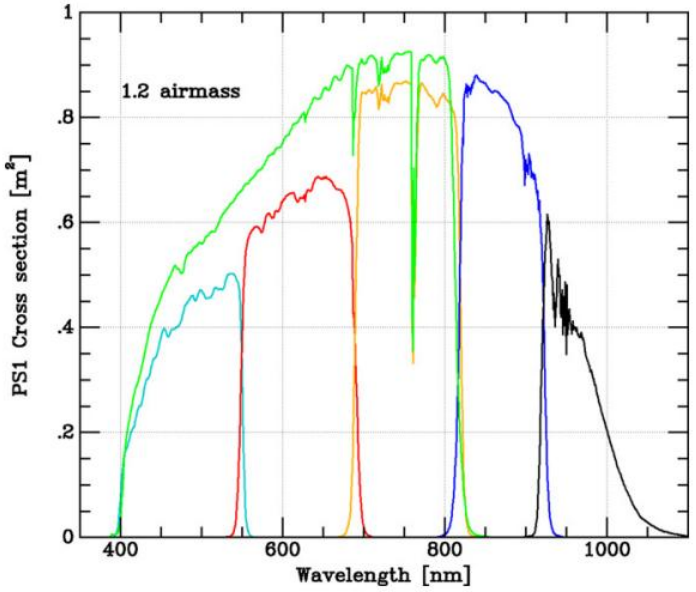
10年运行数据量：~10TB

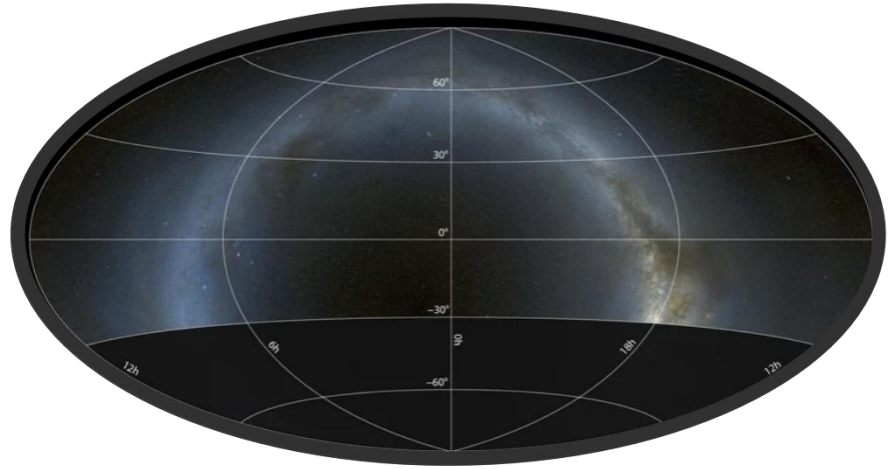
Panoramic Survey Telescope and Rapid Response System

望远镜：1.8m 覆盖面积：30000平方度 天测精度：50mas 极限星等：r<23



Filter	Mean Wavelength
g	4866
r	6215
i	7545
z	8679
y	9633

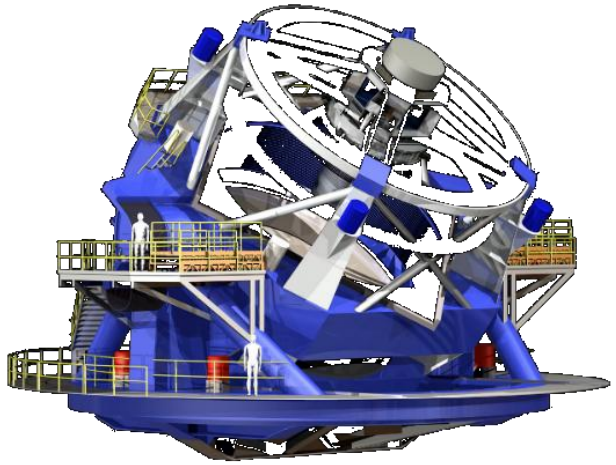


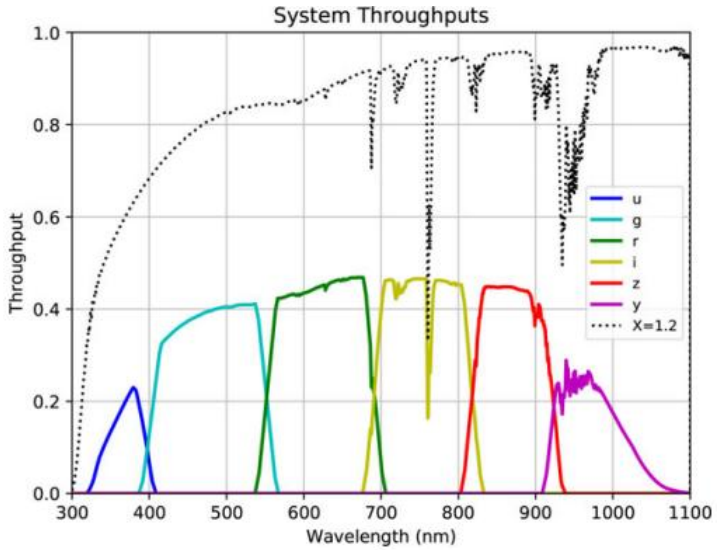


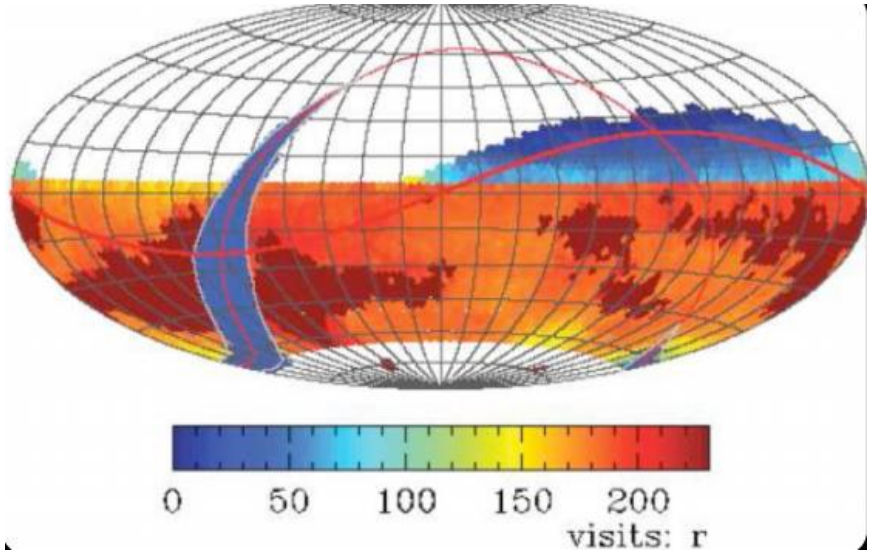
每晚约700GB

LSST: A Deep, Wide, Fast, Optical Sky Survey

望远镜：8.4m　　覆盖面积：>18000平方度　天测精度：10mas　极限星等：r<24.5 (<27.5@10yr)







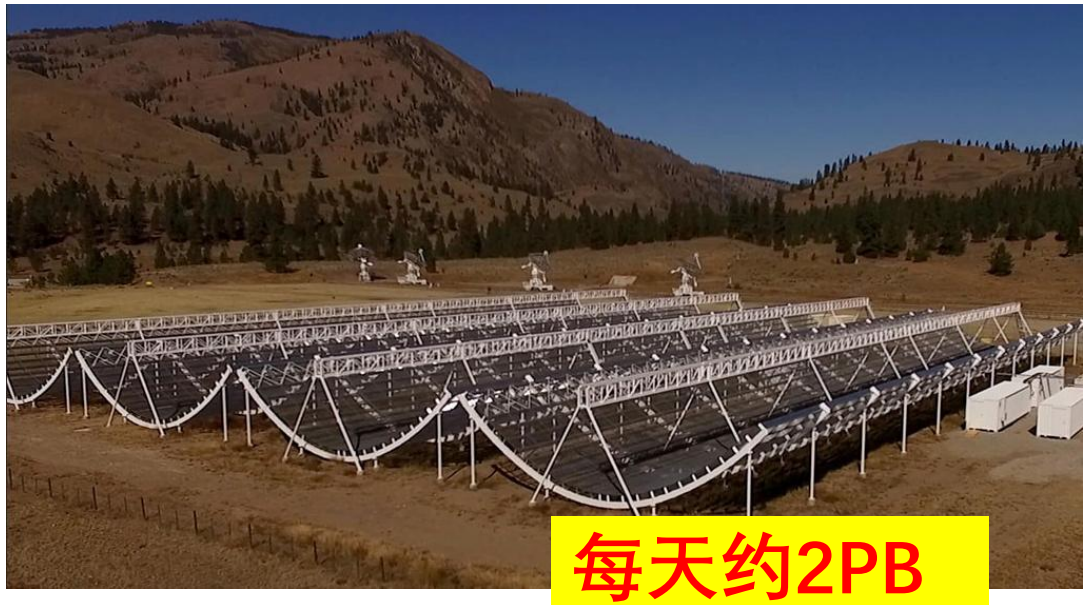
3.2Gpix camera　　30sec exp/4sec rd　　源数： 37Billion

巡天频率：3天　　时长：10年　（825 revisits）

每晚约15TB

- **Canadian Hydrogen Intensity Mapping （CHIME）**

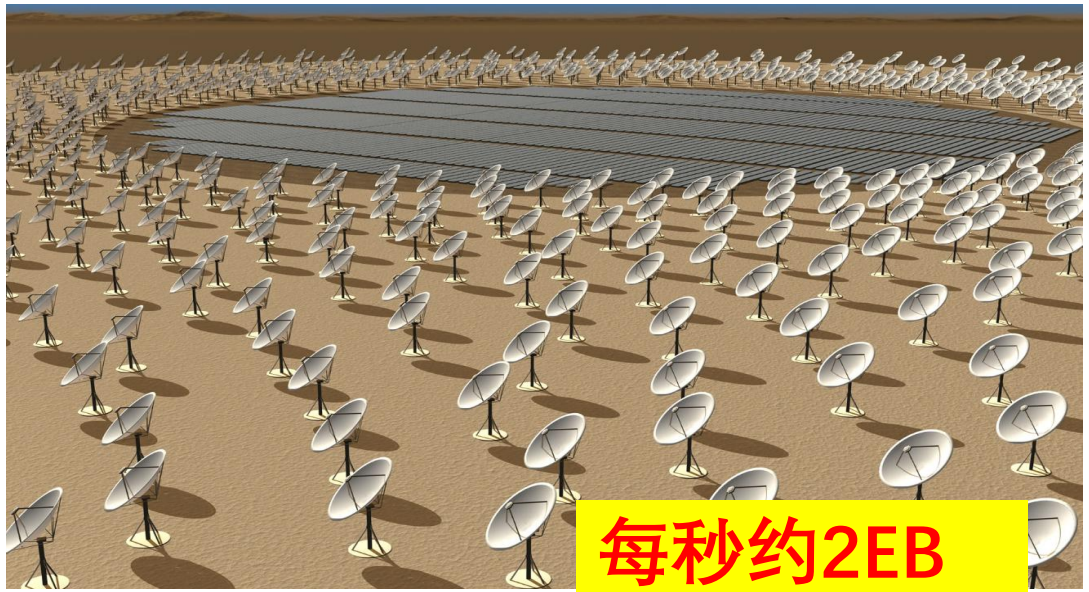
A drift scan radio telescope operating across the 400 MHz to 800 MHz band.Located at the Dominion Radio Astrophysical Observatory near Penticton, BC Canada. The instrument is designed to map neutral hydrogen over the redshift range 0.8 to 2.5 to constrain the expansion history of the Universe.



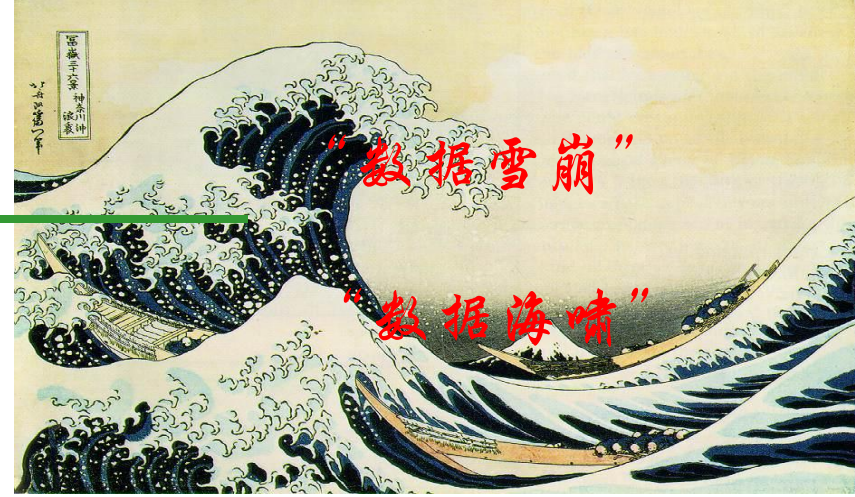
- **The Square Kilometre Array（SKA）**

当前国际射电天文界的最重要的大型望远镜项目，指一个信号收集能力相当于1平方千米镜面收集能力的巨型射电望远镜阵列，它由数千个较小的探测装置组成的阵列，包含上千个射电望远镜，从而形成一个巨大的信号采集面。

在南非和澳大利亚两地兴建平方千米阵。该项目中三分之二的天线建在南非和非洲其他地区，另三分之一的天线建在澳大利亚和新西兰。中国、澳大利亚、英国、南非等20个国家共同参与了这项国际合作项目。

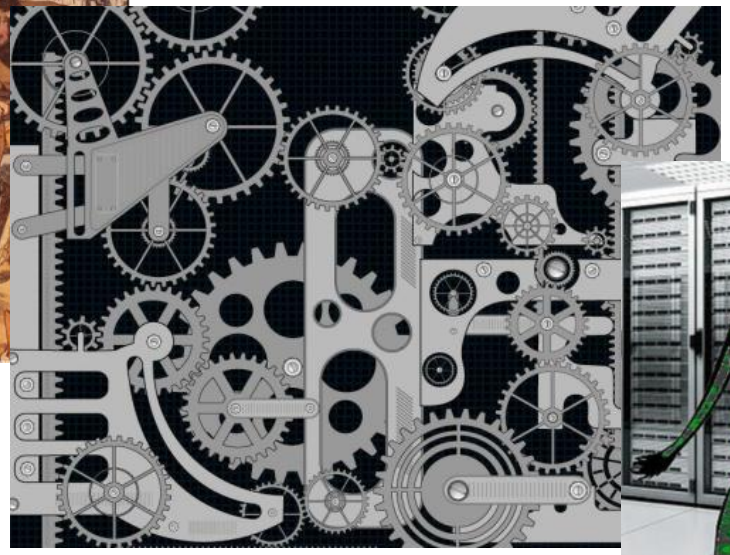


天文学：数据日益丰富的学科



“手工时代”

“工业时代”

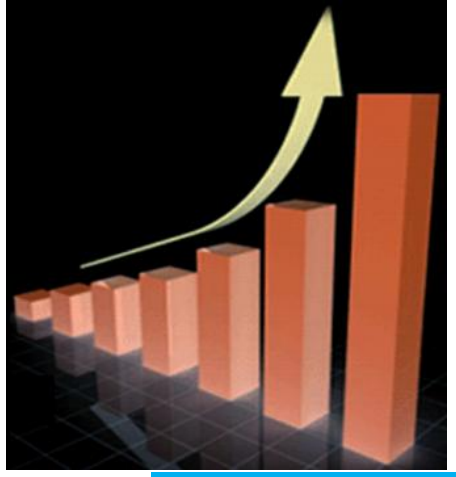


“大数据时代”

“智能时代”

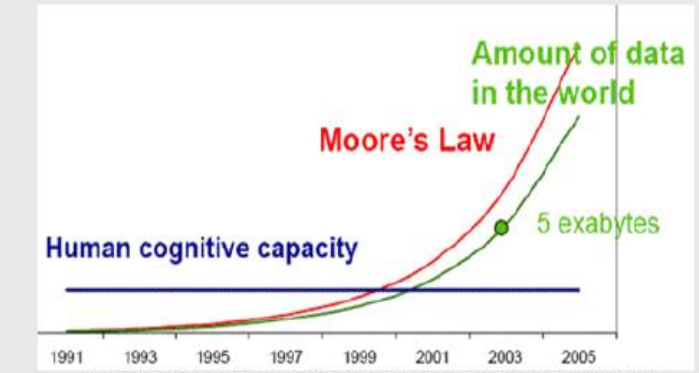
观测技术和信息技术的进步带动天文学的发展

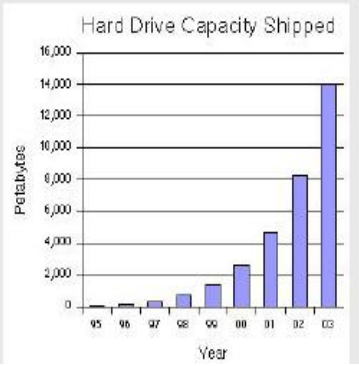
人力 → 电力 → 算力



数据指数增长

每1.5年涨一倍





数据复杂性增长

海量性、空间性、多波段性、异构性、分布性、
非线性、高维性、时域性、缺值性、……

数据的变化

数据贫穷	→	数据过剩
数据集	→	数据流
静态	→	动态、演化
任意时刻	→	实时分析和发现
集中	→	分布
数据者所有	→	领域所有

Credict:S. George Djorgovski

科学研究的四个阶段

- 第一范式

  实验或测量

- 第二范式

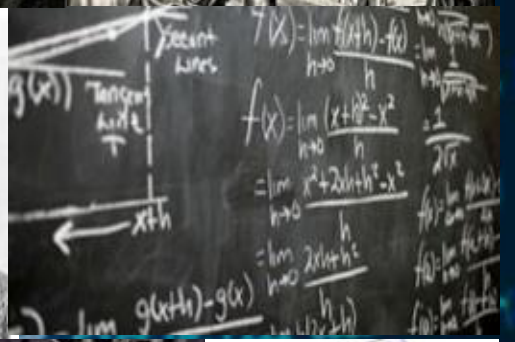
  理论分析

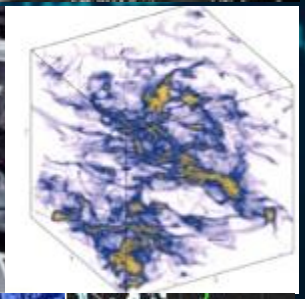
- 第三范式

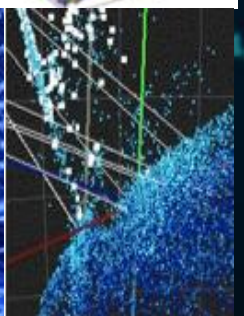
  数值模拟

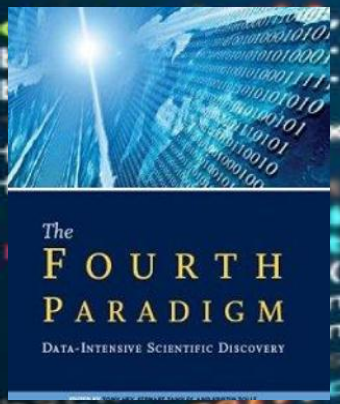
- 第四范式

  数据密集型科学









Data Fusion+DM+ML+DL

Credict:S. George Djorgovski

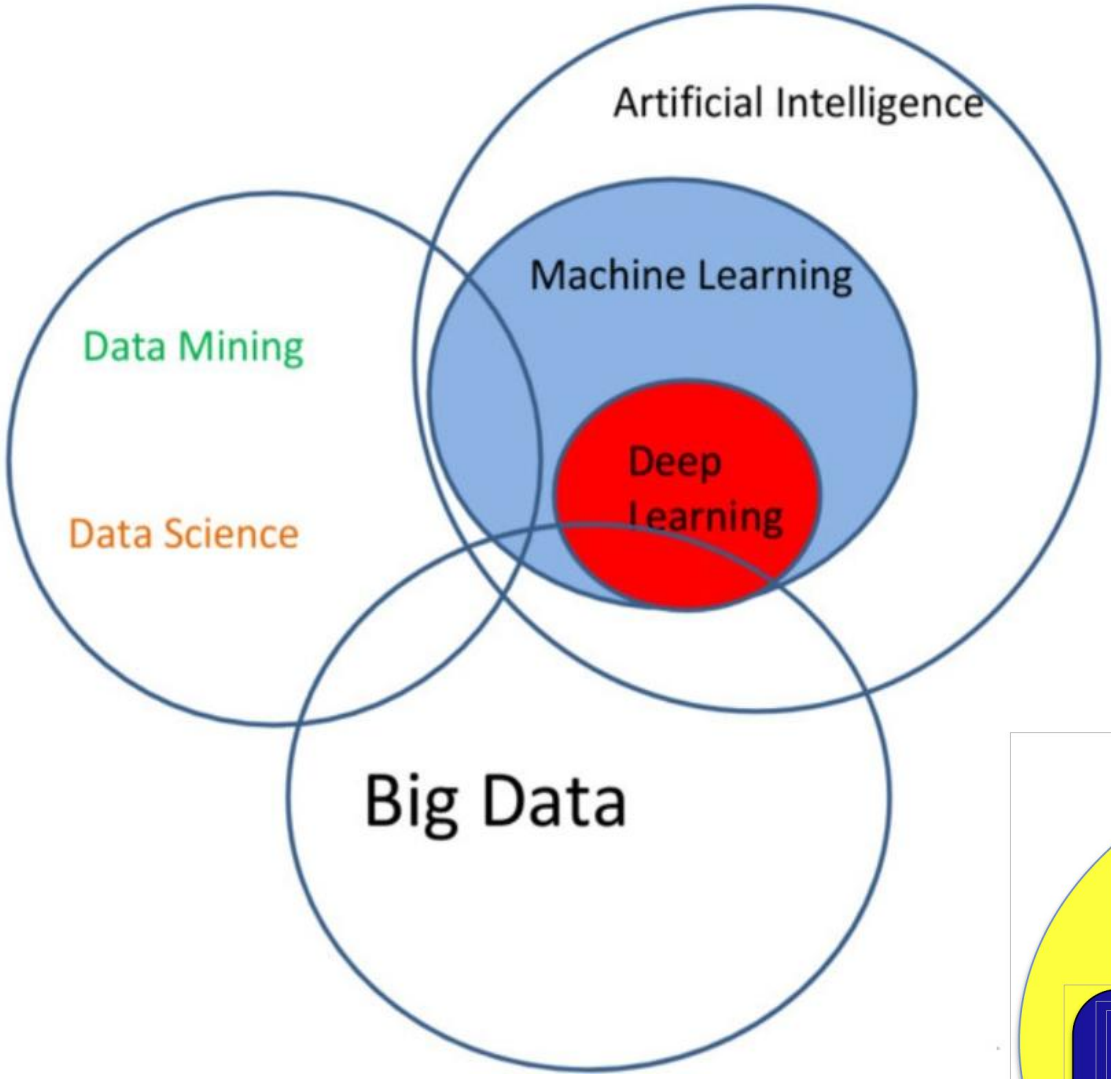
Astronomy entered new era:

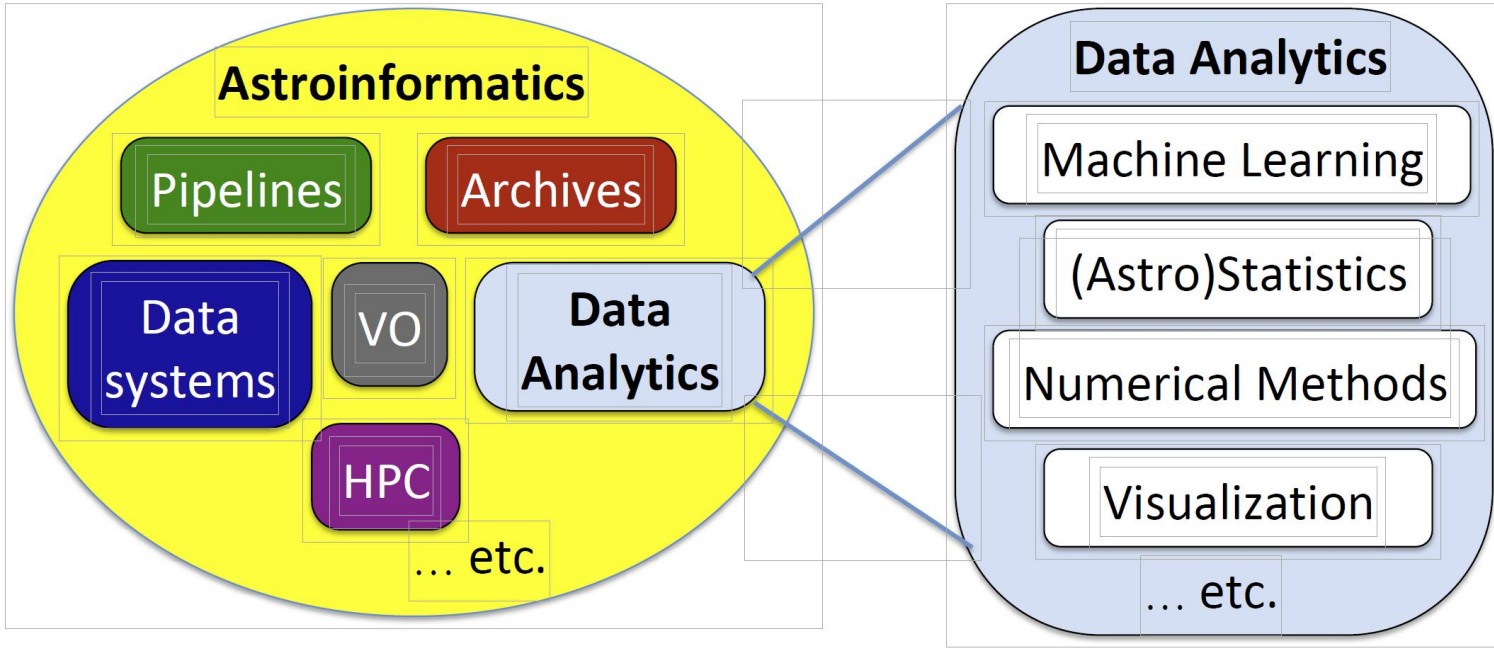
"Ask Not What Data You Need To Do Your Science, Ask What Science You Can Do With Your Data."

The era of surveys...

- Standard: "What data do I have to collect to (dis)prove a hypothesis"?
- Data-driven: "What theories can I test given the data I already have?"

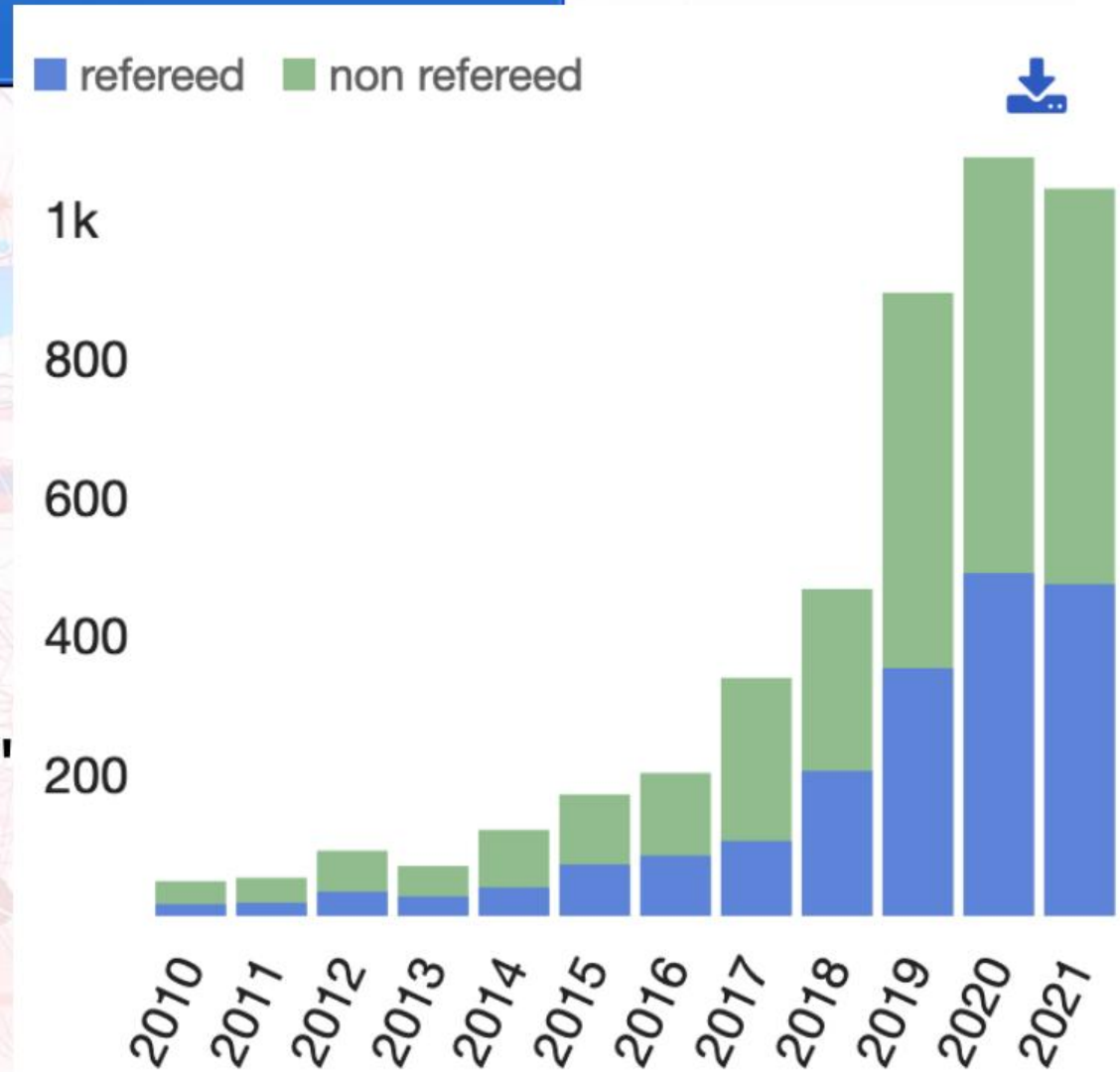
Credict: Zeljko Ivezic


Artificial Intelligence
Machine Learning
Data Mining
Deep Learning
Data Science
Big Data

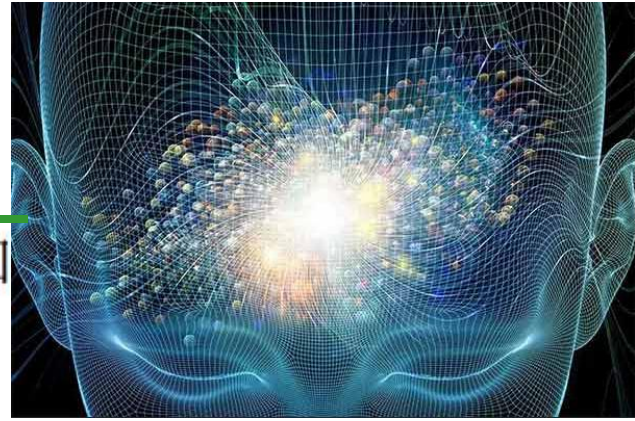

Astroinformatics
Pipelines
Archives
Data systems
VO
Data Analytics
HPC
… etc.
Data Analytics
Machine Learning
(Astro)Statistics
Numerical Methods
Visualization
… etc.

ML use in astronomy

Astronomy papers in ADS containing "Artificial Intelligence" or "Machine learning" or "Deep learning" in the abstract.



机器学习



机器学习是一门通过编程让计算机从数据中进行学习的科学（和艺术）。

机器学习是一个研究领域，让计算机无须进行明确编程就具备学习能力。

——亚瑟·萨缪尔（Arthur Samuel），1959

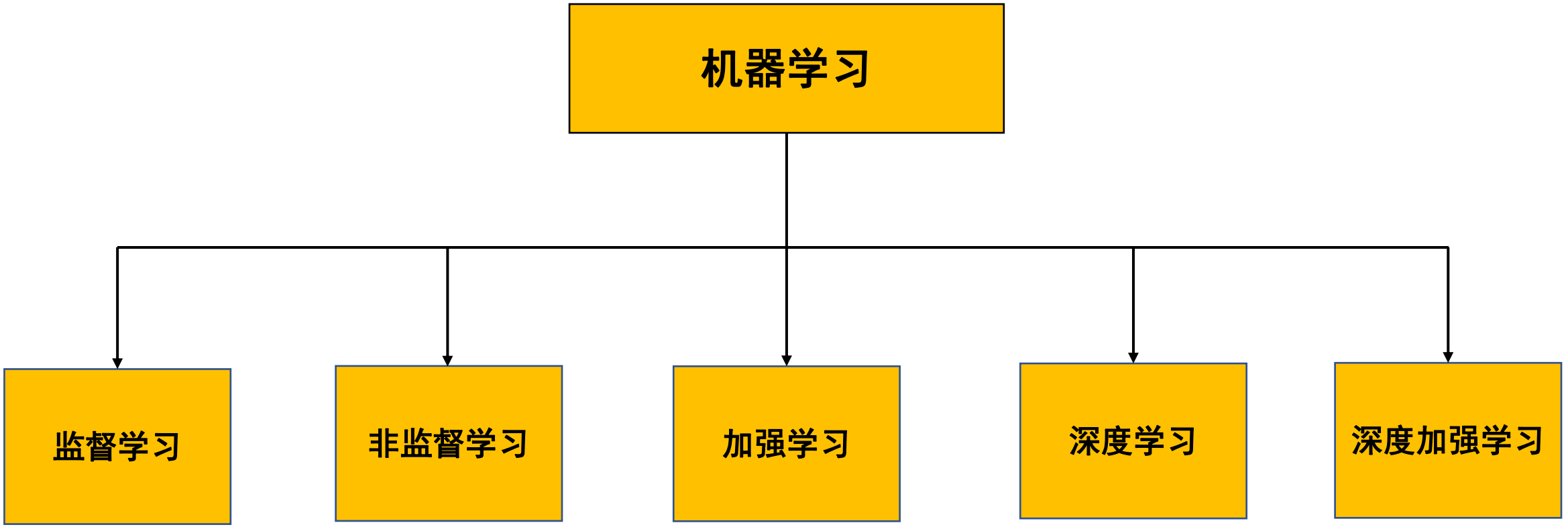
一个计算机程序利用经验E来学习任务T，性能是P，如果针对任务T的性能P随着经验E不断增长，则称为机器学习。

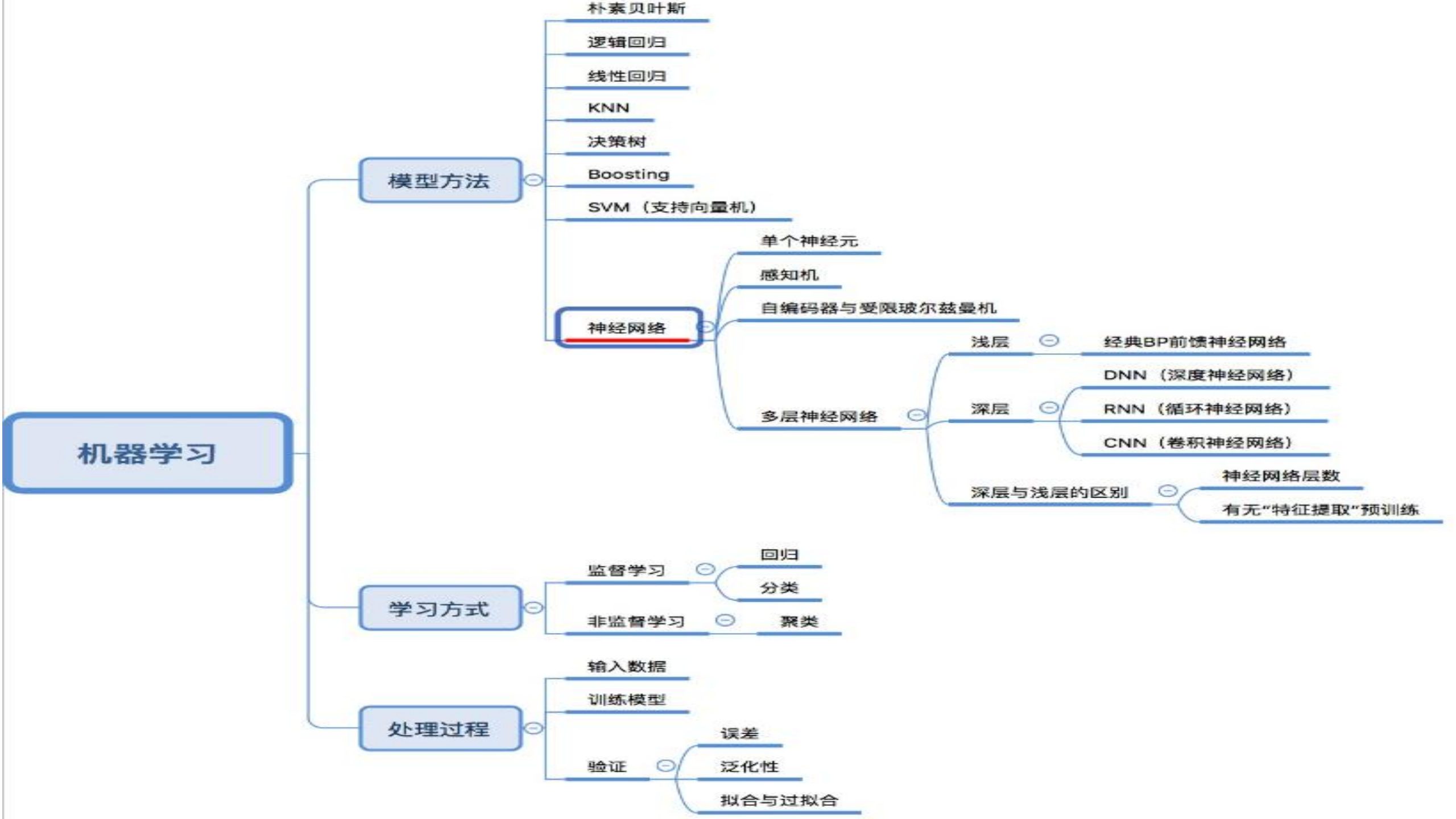
——汤姆·米切尔（Tom Mitchell），

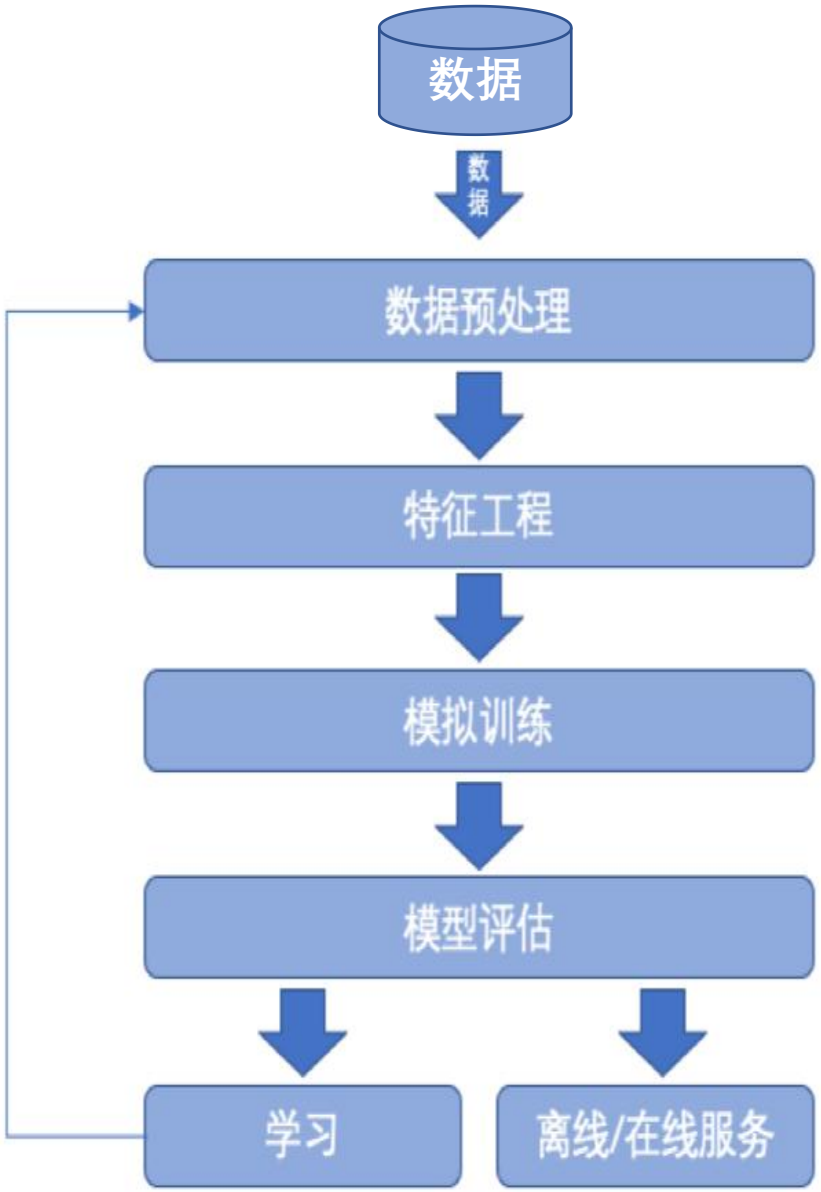
- 例子：垃圾邮件过滤器
  - **T**：识别垃圾邮件
  - **E**：用户标注的垃圾邮件
  - **P**：正确识别百分比

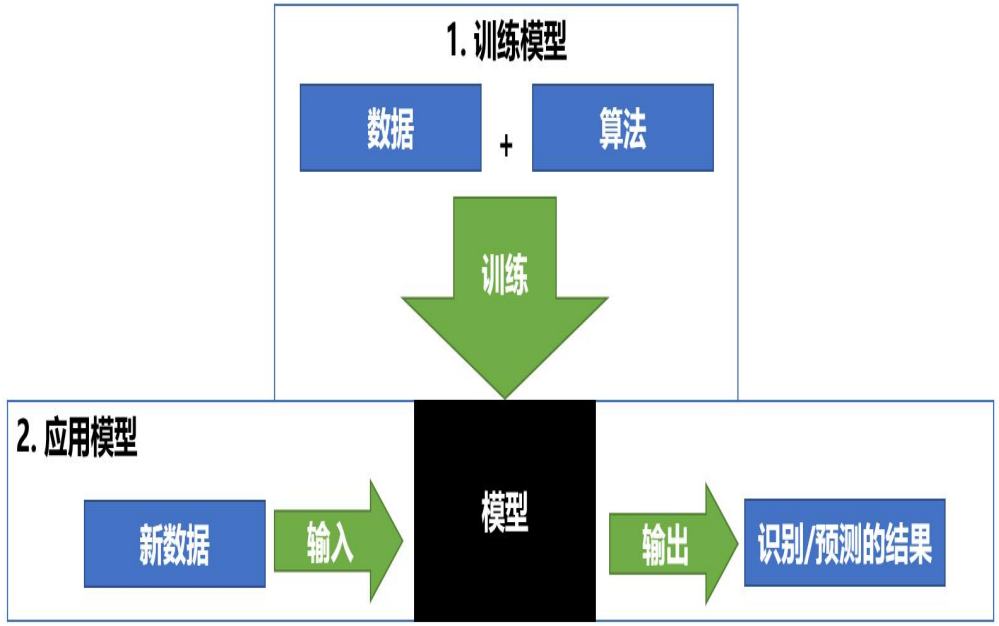
- 它是人工智能核心，是使计算机具有智能的根本途径。

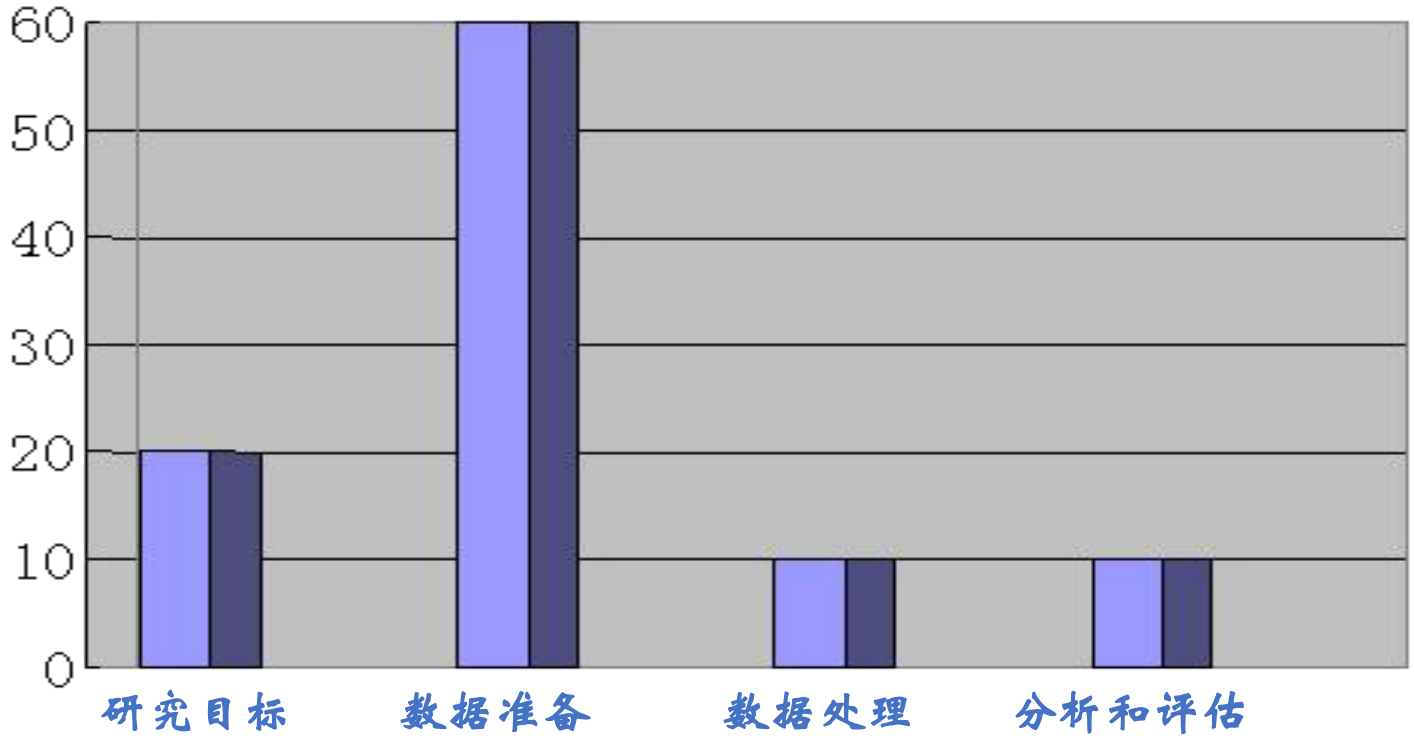
机器学习分类




机器学习
模型方法
朴素贝叶斯
逻辑回归
线性回归
KNN
决策树
Boosting
SVM（支持向量机）
神经网络
单个神经元
感知机
自编码器与受限玻尔兹曼机
多层神经网络
浅层
经典BP前馈神经网络
深层
DNN（深度神经网络）
RNN（循环神经网络）
CNN（卷积神经网络）
深层与浅层的区别
神经网络层数
有无"特征提取"预训练
学习方式
监督学习
回归
分类
非监督学习
聚类
处理过程
输入数据
训练模型
验证
误差
泛化性
拟合与过拟合

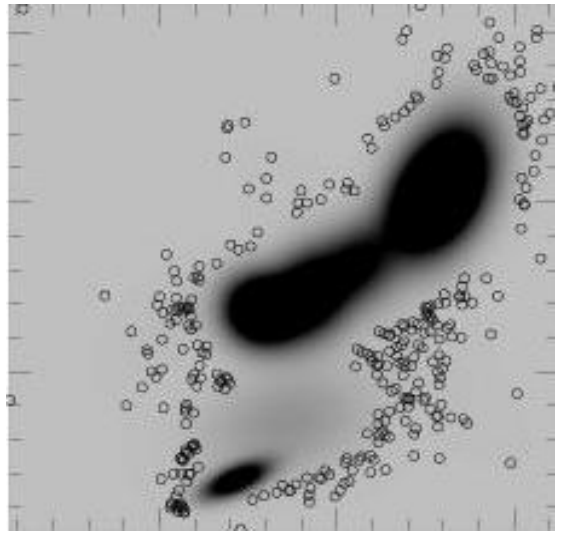

数据
数据
数据预处理
特征工程
模拟训练
模型评估
学习
离线/在线服务

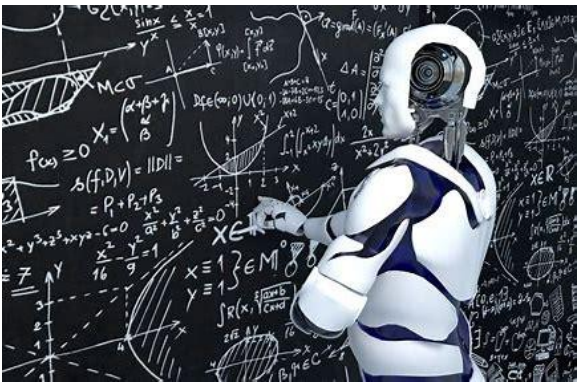

1. 训练模型
数据
+
算法
训练
2. 应用模型
新数据
输入
模型
输出
识别/预测的结果

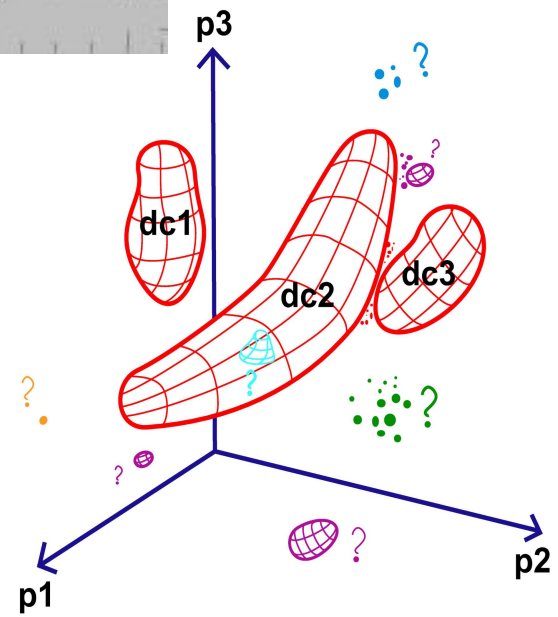

60
50
40
30
20
10
0
研究目标
数据准备
数据处理
分析和评估

机器学习任务

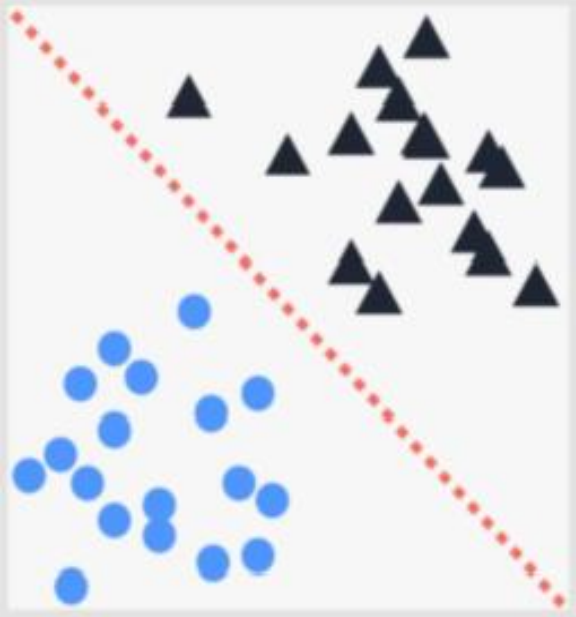
- 数据总结
- 分类分析
- 回归分析
- 聚类分析
- 关联规则分析
- 序列模式分析
- 依赖关系分析
- 异常检测
- 模式分析或统计分析
- 其它



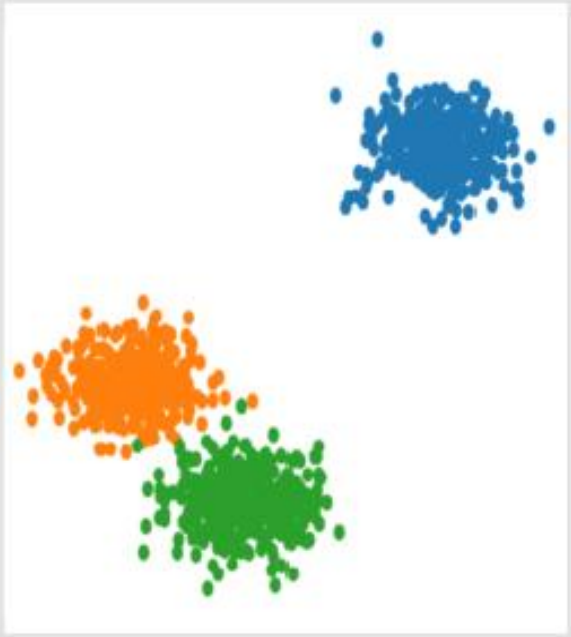




Classification

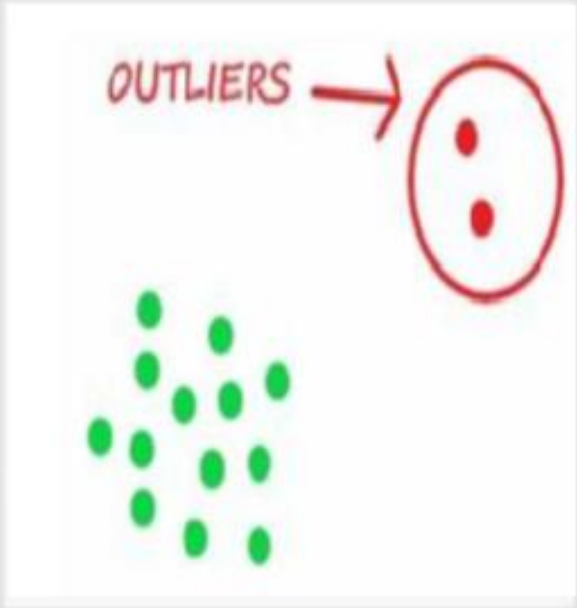


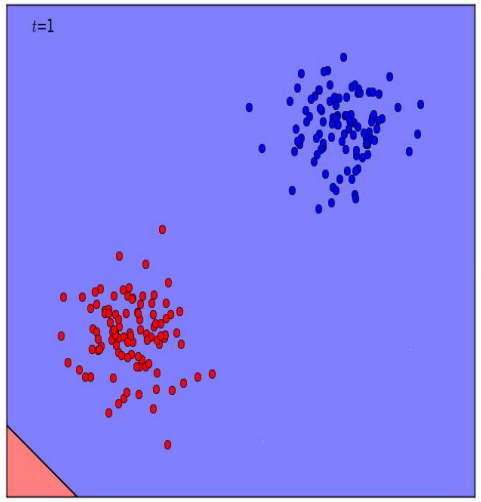
Clustering

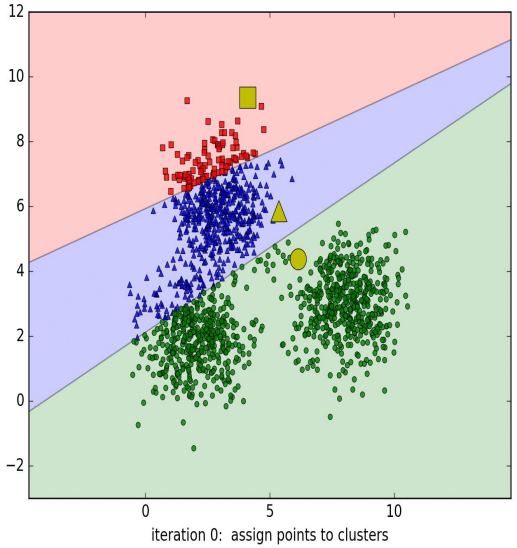


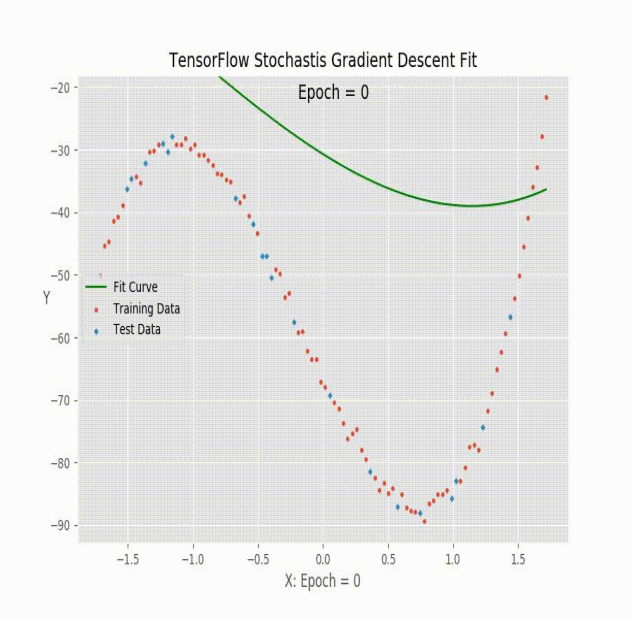
Regression

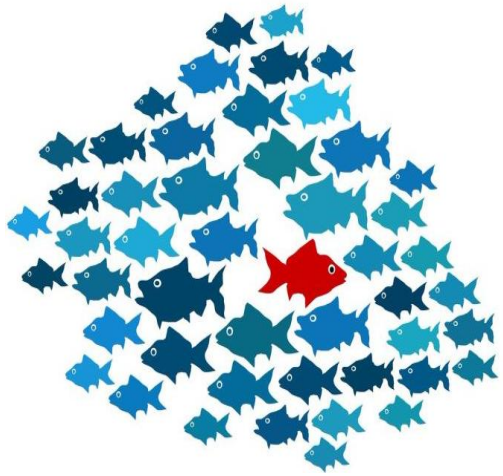
Outlier











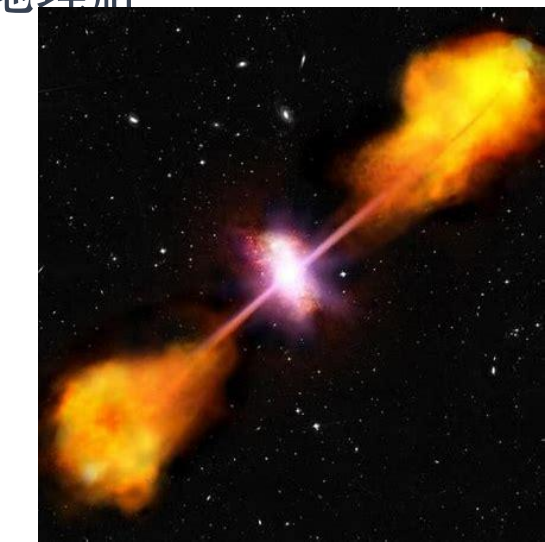
天文学应用

1.自动化天文观测：人工智能可以用于对天文数据进行实时分析和处理，从而实现自动化天文观测。例如，美国的Palomar天文台使用机器学习算法来自动化其天文观测流程，提高了数据处理效率和准确性。

2.星系和天体分类：人工智能可以用于对星系图像进行分类和识别。例如，天文学家可以使用卷积神经网络（CNN）来自动识别星系的形状和类型，从而更好地理解星系的演化和性质；科学家可以使用支持向量机（SVM）等算法来对星系进行分类，并根据其形态、颜色和亮度等特征来研究星系的演化和性质。

3.天体物理学模拟：人工智能可以用于天体物理学的模拟和预测。例如，科学家可以使用深度学习算法来模拟黑洞合并事件和星系形成过程，从而更好地理解宇宙的演化。

天文学是一个数据密集型的科学领域，天文学家需要处理各种不同类型的数据，如天体图像、光谱数据、时间序列数据等。随着天文学数据量的不断增加，AI、ML、DL在天文学中变得越来越重要。



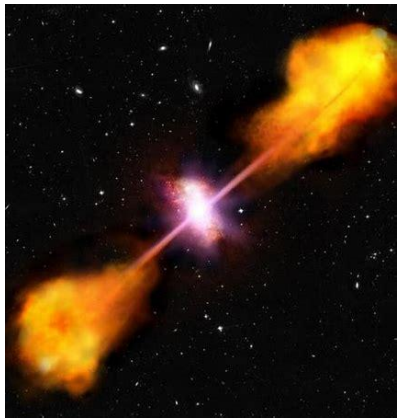
天文学应用

4.天体目标检测：机器学习可以用于对天体目标的检测和识别。例如，科学家可以使用卷积神经网络（CNN）来自动检测和分类行星、彗星、恒星和星系等天体目标，从而更好地理解宇宙中的各种天体。

5. 天文数据分析与挖掘：人工智能可以用于对天文数据进行分析和挖掘。例如，科学家可以使用机器学习算法来挖掘银河系中的恒星数量和位置分布，从而更好地理解银河系的结构和演化。科学家可以使用聚类算法来发现天体之间的关联性，或使用关联规则算法来发现天体之间的联系，从而更好地理解宇宙的结构和演化。

6. 天体成像和图像处理：机器学习可以帮助天文学家处理和分析天体图像和数据，例如在图像增强、去噪、分割和重构等方面。

7. 天体轨迹预测：机器学习可以帮助天文学家预测天体的轨迹和运动，以了解它们的物理性质和行为。



天文学应用

8.**天体成像和模拟**：人工智能可以帮助天文学家处理和分析天体图像和数据，例如利用神经网络和计算机模拟技术来生成高分辨率的天体图像和模拟。

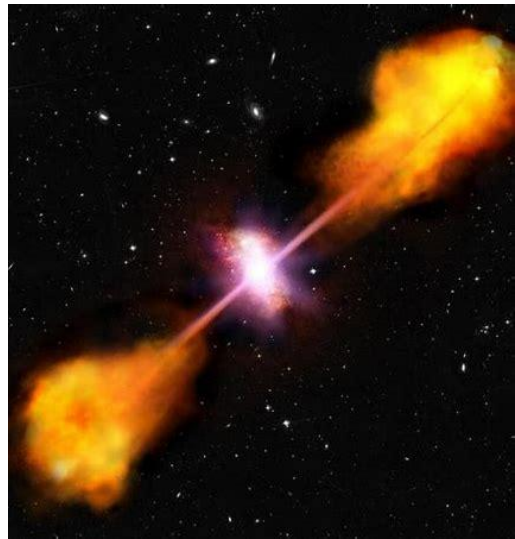
9.**宇宙探索和探测**：人工智能可以帮助控制太空探测器和卫星，以收集和分析宇宙数据，例如在探测黑洞、探测暗物质等方面。

10.**宇宙射线探测**：人工智能可以帮助处理和分析宇宙射线数据，以了解宇宙的物理过程和演化。

11.**天体信号处理**：人工智能可以帮助处理和分析天体信号，例如天体信号检测、抑制天体噪声、提高探测灵敏度等。

12.**天体物理参数测量**：人工智能可以帮助测量天体的物理参数，如恒星的金属丰度、重力加速度、有效温度；星系和类星体的测光红移。

13.**特殊天体搜寻**：如：FRB、双星系、豌豆星系、环星系、变脸AGN等。

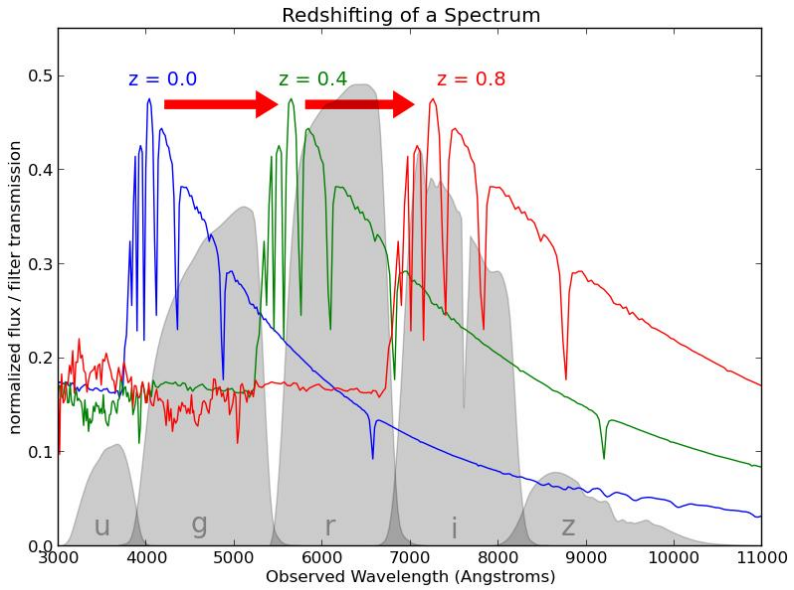


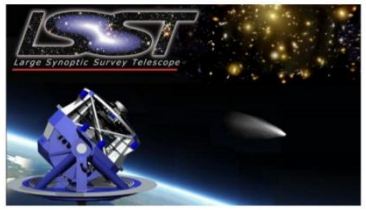
测光红移

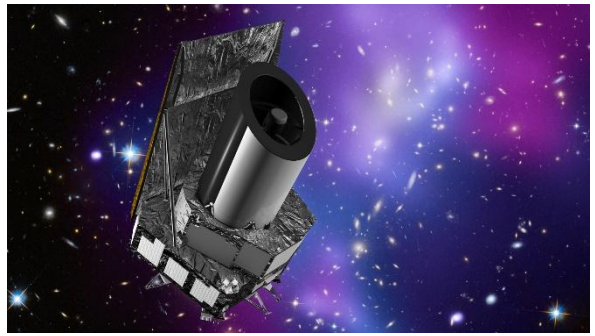
- The most reliable estimate of redshift is the spectroscopic one but it is very time consuming
- The ongoing (VST, HST, PANSTARRS, KiDS, DES) and future planned surveys (LSST, SKA, EUCLID) foresee the storing of a huge amount of data per day (from Tera to Peta-byte), only Data Mining techniques can handle this amount

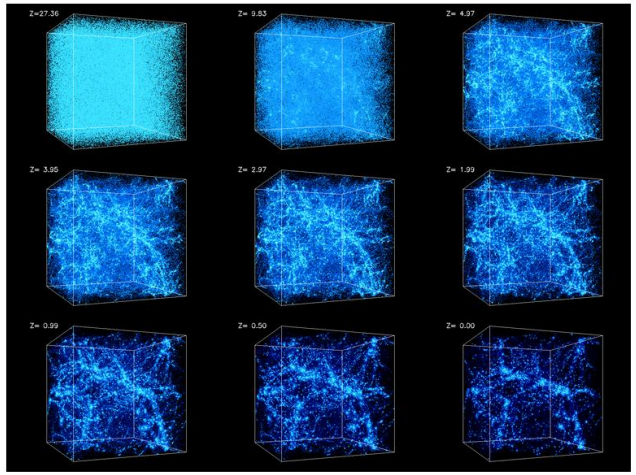
Useful for many scientific goals:

- Determine the Dark Matter and Energy content of the Universe
- To constraint cosmological parameters
- To study weak lensing
- To reconstruct the Large Scale Structure of the Universe
- To classify astronomical objects

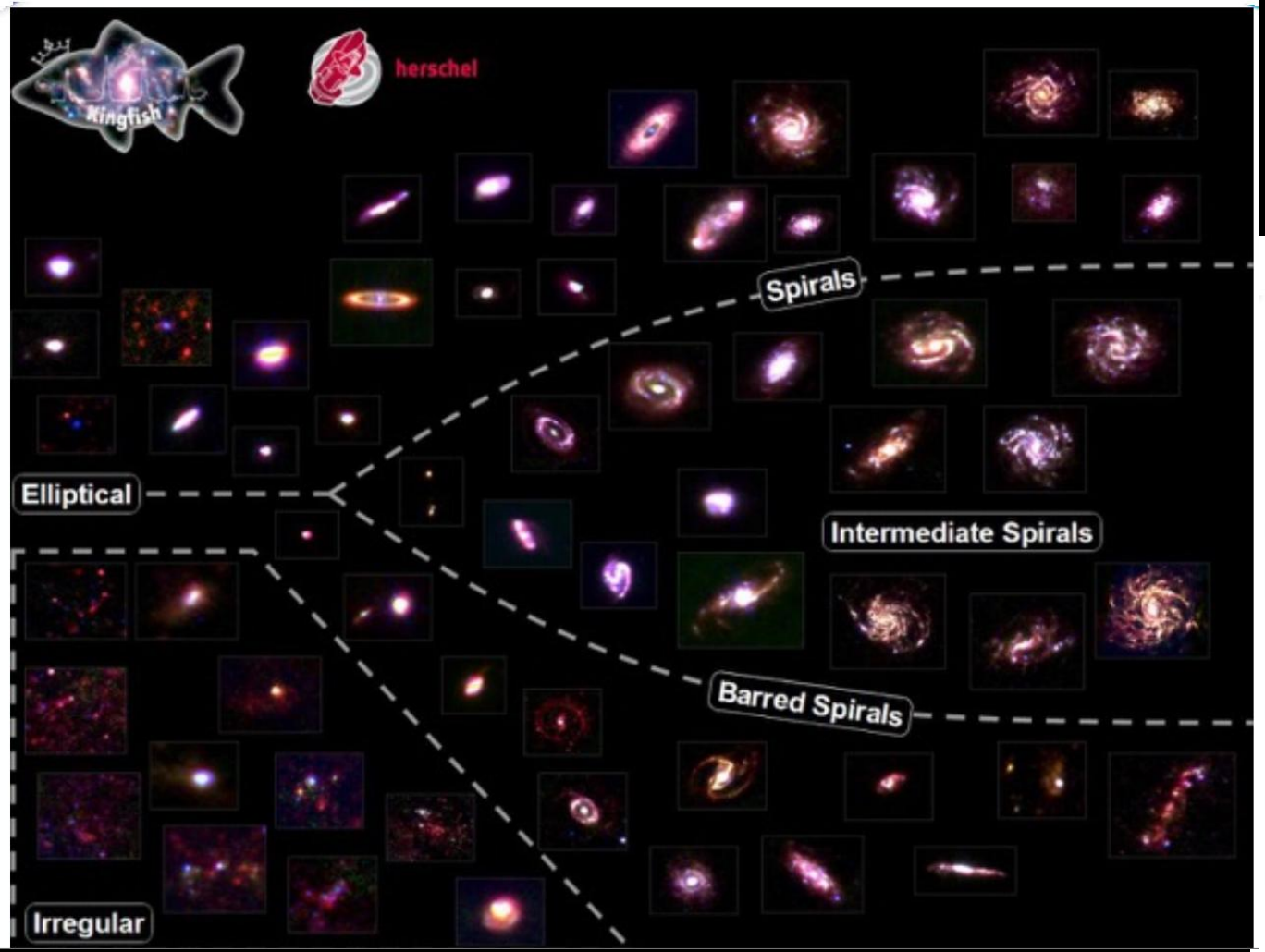




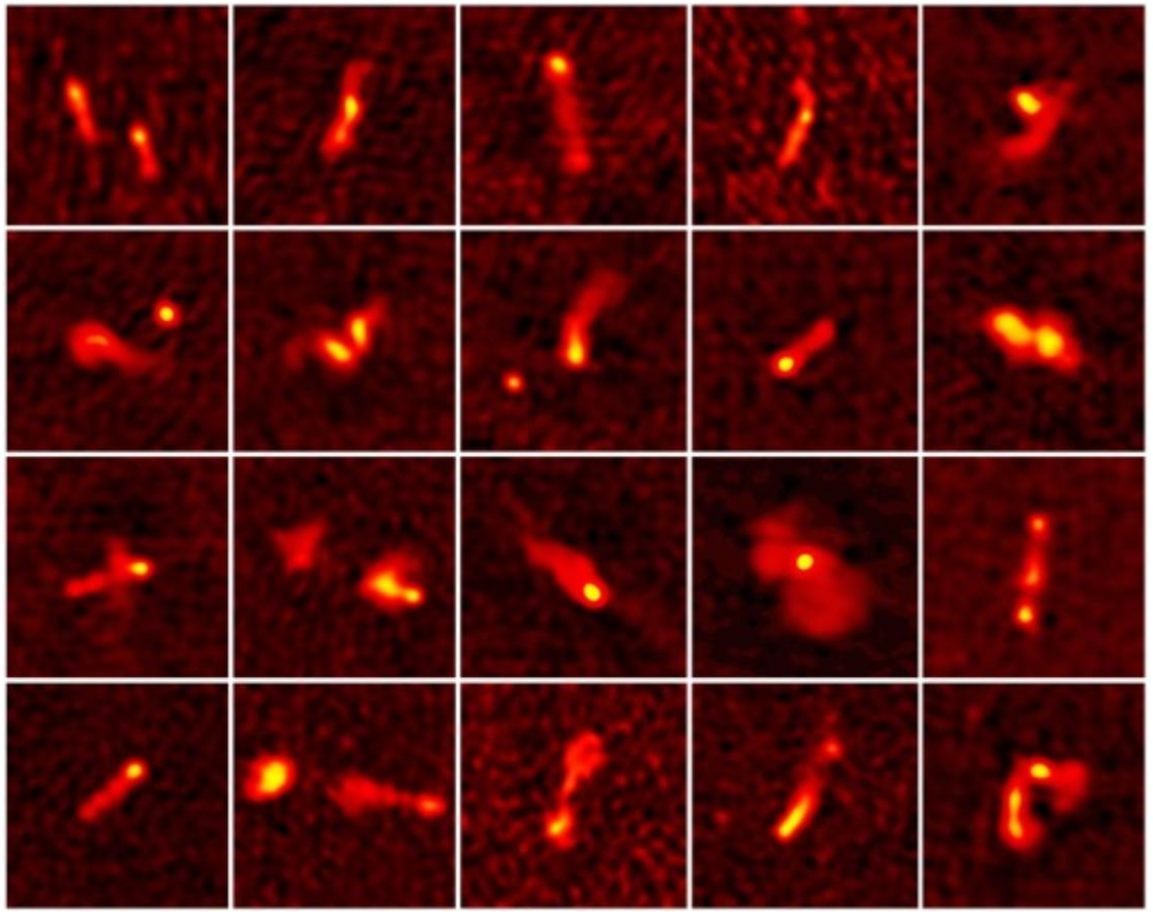




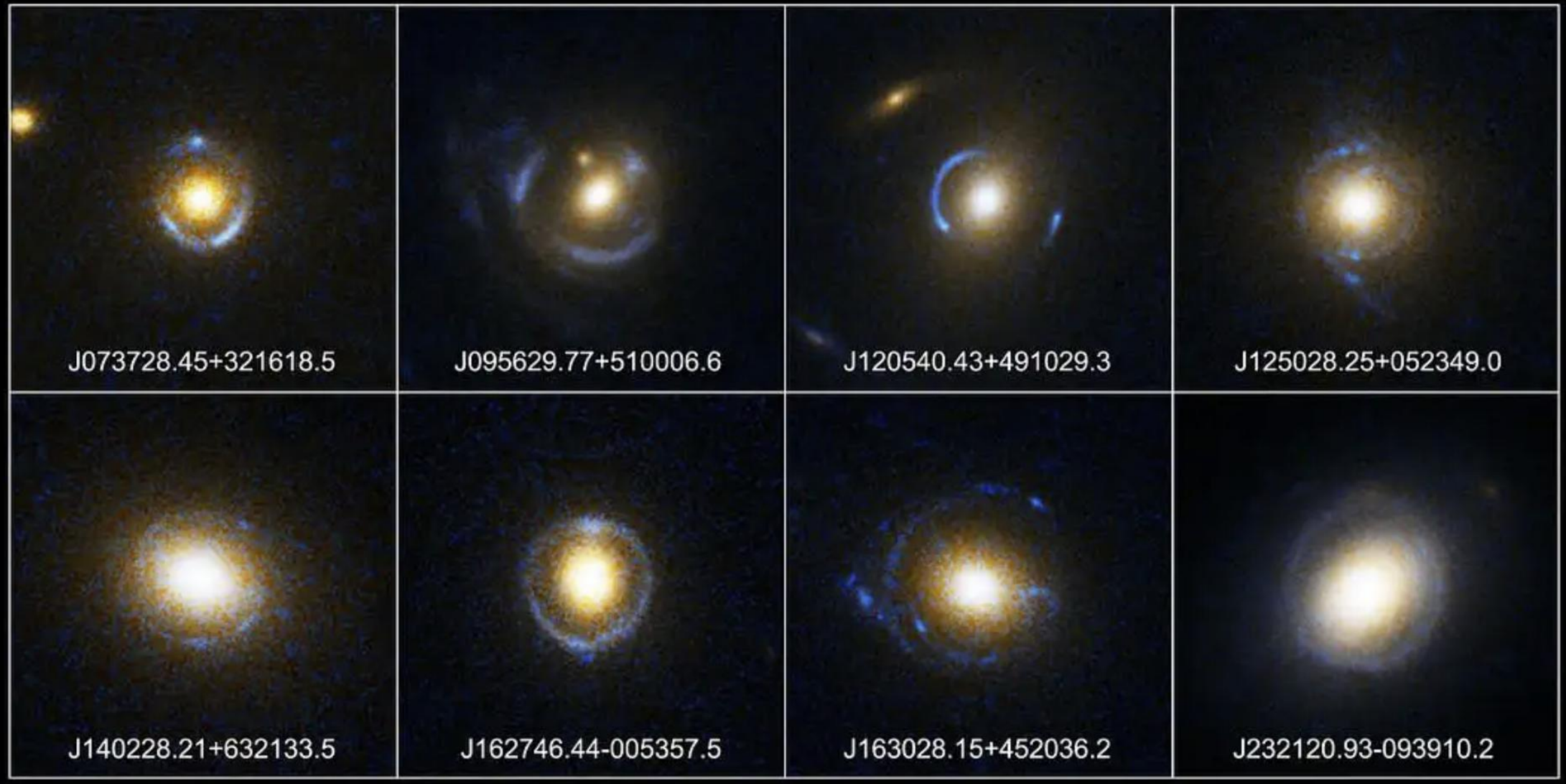
Galaxy Morphological Classification



Radio Galaxies



VLSS - NRAO



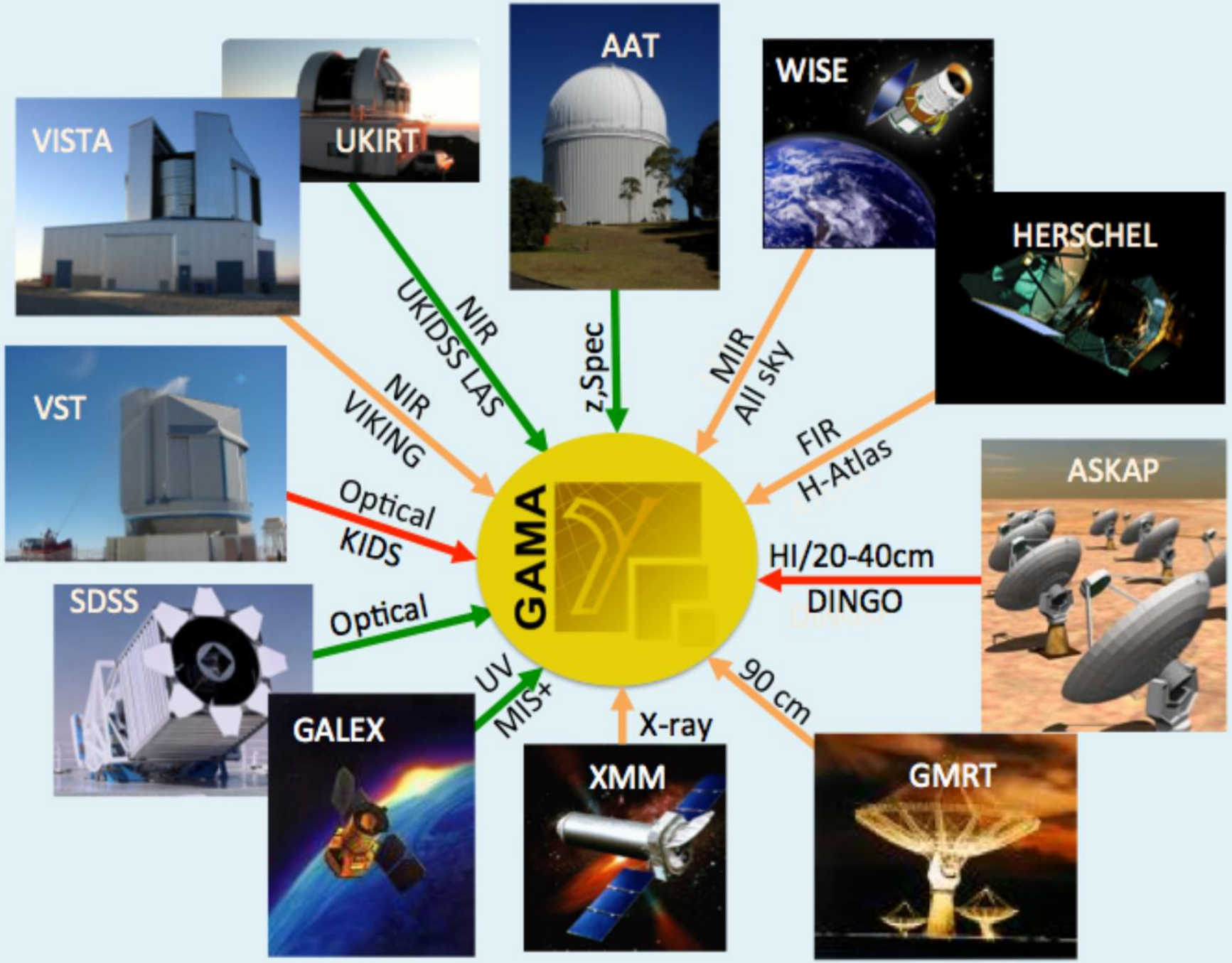
Einstein Ring Gravitational Lenses

Hubble Space Telescope • Advanced Camera for Surveys

NASA, ESA, A. Bolton (Harvard-Smithsonian CfA), and the SLACS Team

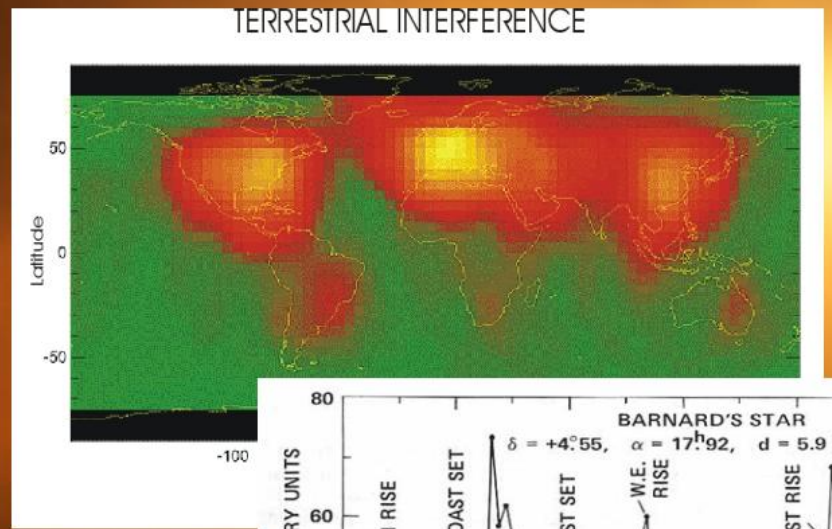
STScI-PRC05-32

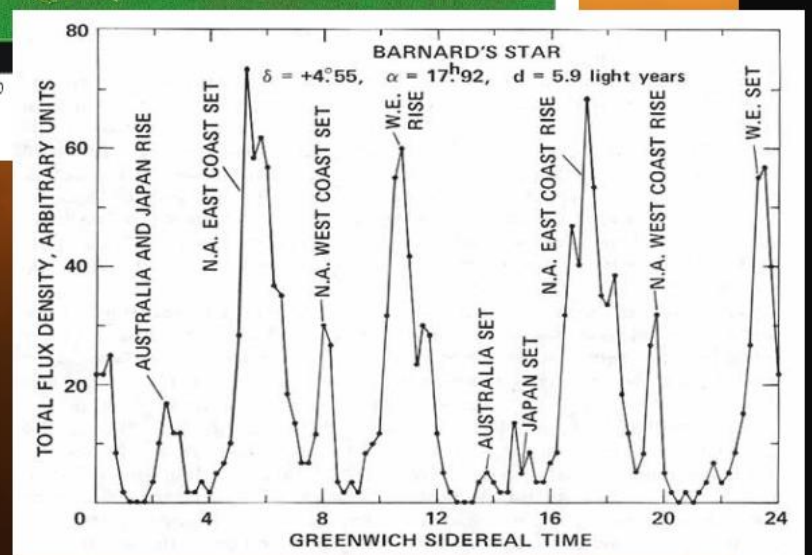
多波段交叉证认

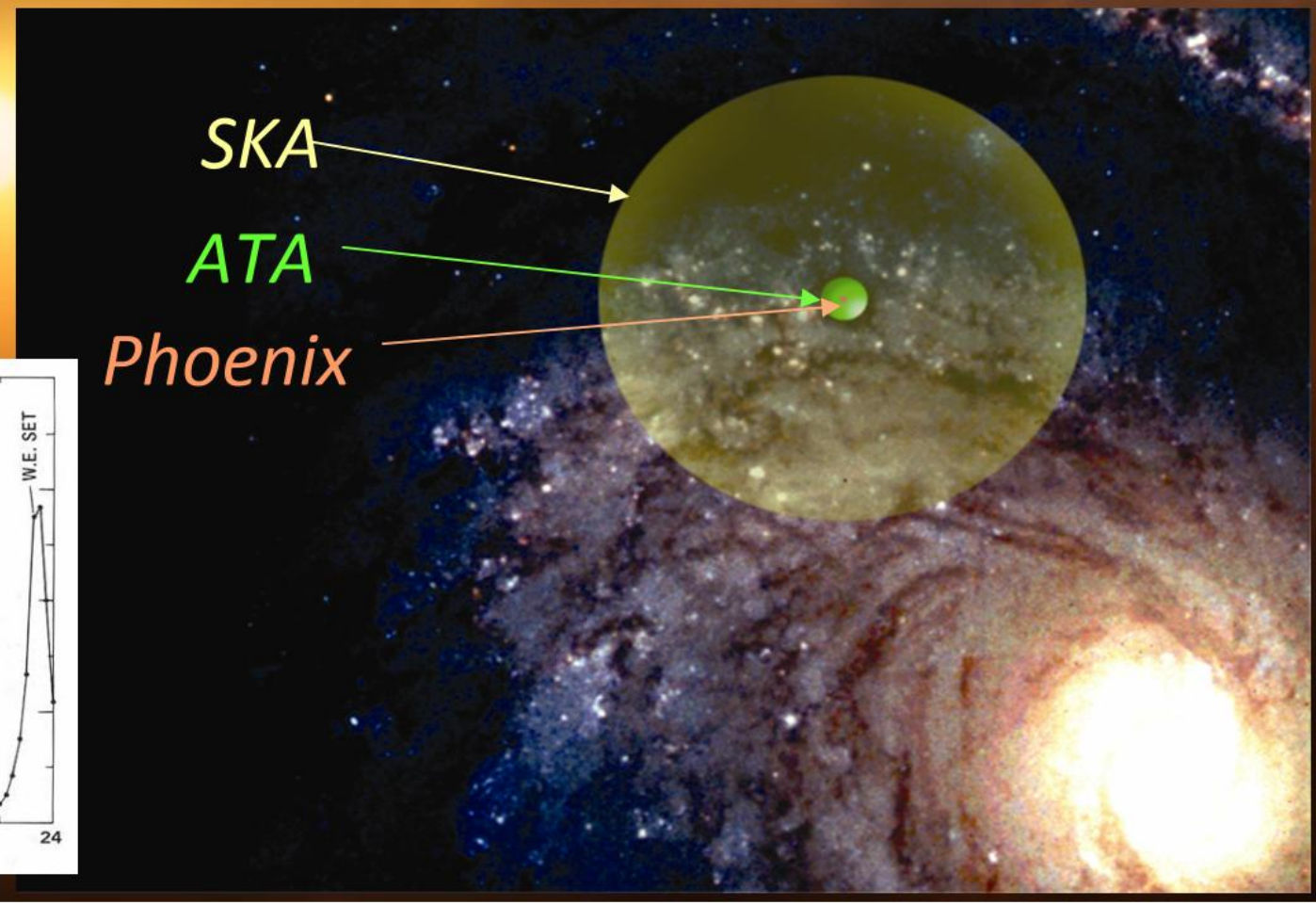

AAT
WISE
VISTA
UKIRT
HERSCHEL
VST
ASKAP
SDSS
GALEX
XMM
GMRT
GAMA
NIR
UKIDSS LAS
z,Spec
MIR
All sky
NIR
VIKING
FIR
H-Atlas
Optical
KIDS
HI/20-40cm
DINGO
Optical
UV
MIS+
X-ray
90 cm

Are we alone? – is anyone out there?

The Search for Extra-terrestrial Intelligence

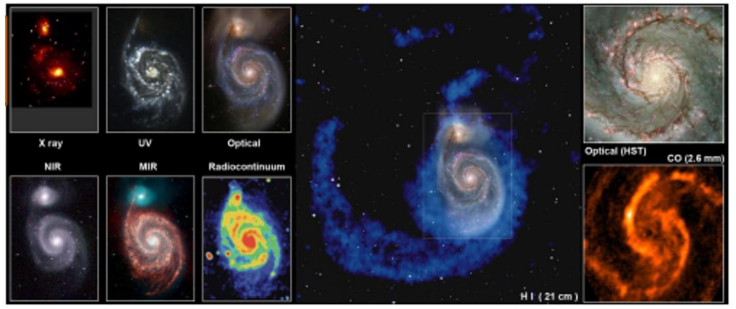


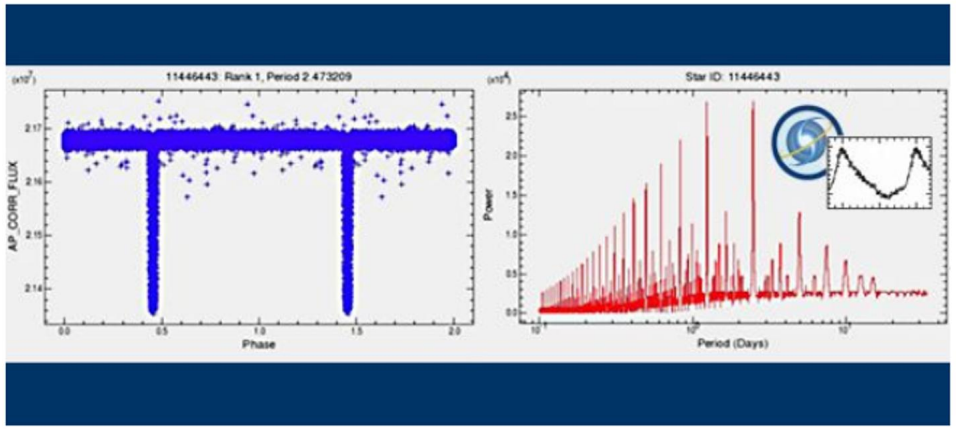




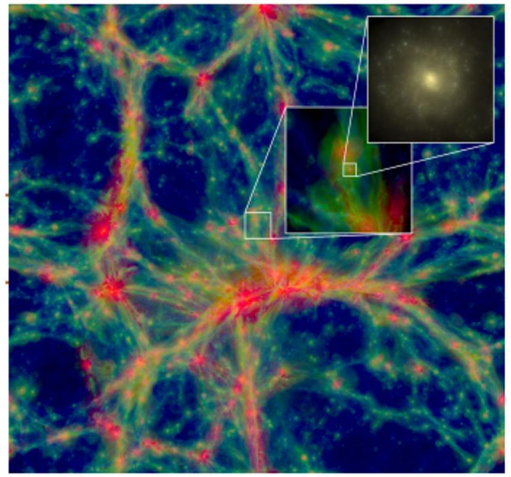
重要的天文发现可能来自于：

1. 来自多波段数据的交叉相关

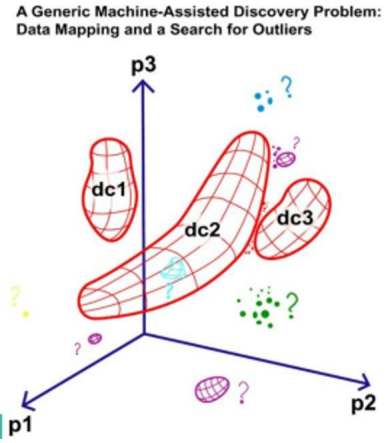




2. 观测源的时域特性



3. 数值模拟与实测数据对比研究



4. 在高维（大于等于3维）观测空间中寻找模式或趋势

智能时代天文大数据分析挑战

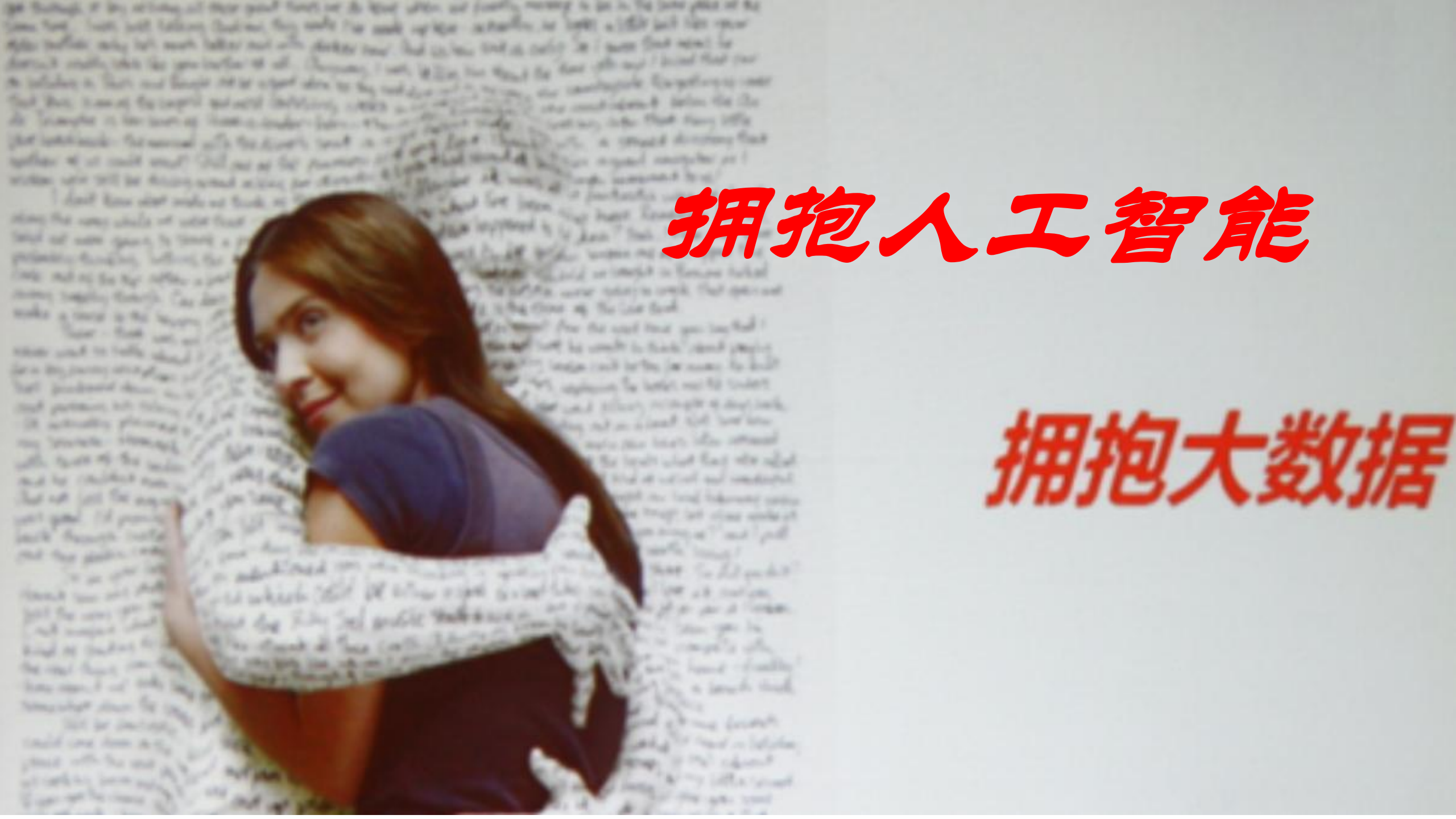
大数据、全波段、时域、多模态数据为新发现和新理论提供了前所未有的机遇

- 数据量大 （获取、存储、传输、分析等）
- 样本不完备、非平衡
- 缺少标注数据
- 特征工程： 特征选择、特征抽取、降维
- 未知统计分布
- 时序数据（不规则取样、不同波段）
- 异方差、删失数据、缺值数据
- 不可靠的数据质量（未知的系统或随机误差）
- 算法和模型的选择与优化 “no one size fits all”
- 模型的可解释性、可扩展性、健壮性
- 可视化（机器学习的各个阶段）
- 人机交互

瓶颈：不在获得数据，而在如何有效地利用AI、ML、DL从数据中挖掘出有用的、高价值的信息和知识

合作共赢的时代

- 多学科合作
- 跨界合作
- 跨平台合作
- 虚拟组织
- 培养面向AI的下一代天文学家


拥抱人工智能
拥抱大数据

君子生非异也，善假于物也。

——《荀子·劝学》

未来打败你的会是使用AI的人

“如果我看得更远，那是因为我站在巨人的肩膀上。（If I have seen further it is by standing on ye shoulder of Giants.）”(Newtown,I. 1676)