

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

# 高通量计算流程、数据库和机器学习简介

北京市计算中心 云平台事业部 姜骏

2023.12

# 高性能计算: 中心化集群

高通量计算流程、数据库和机器学习简介

高通量与高  
性能计算

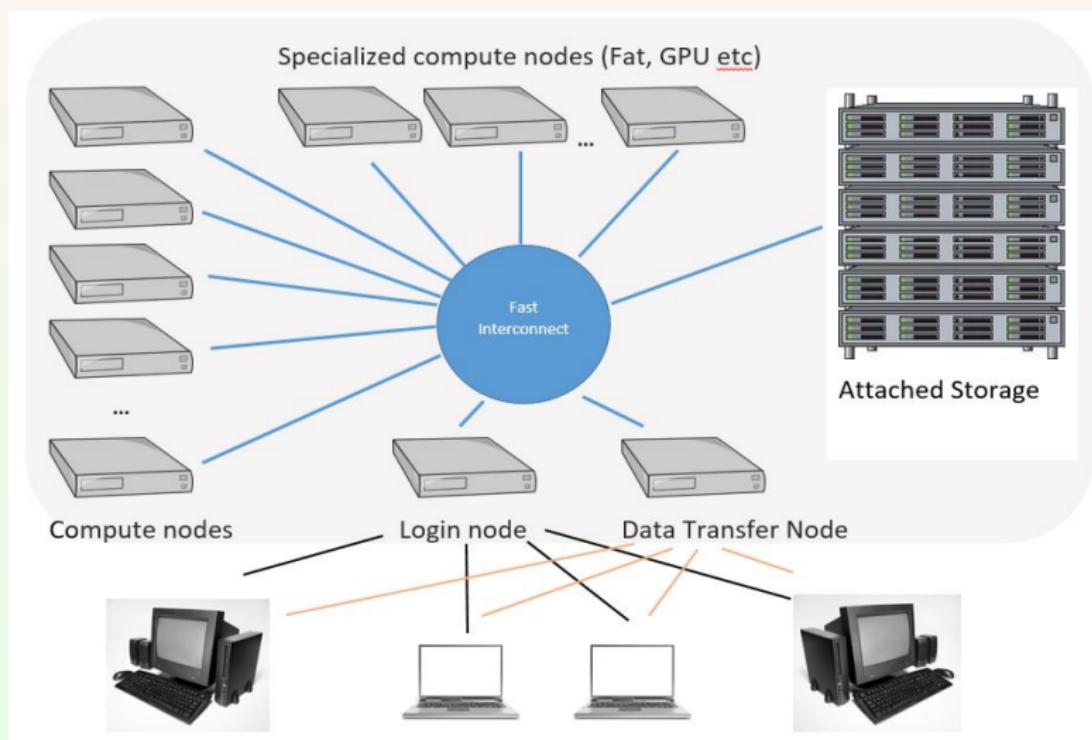
高通量计算材  
料自动流程

第一原理数据  
库

机器学习简介

机器学习算法

数据挖掘与第  
一原理材料研  
究



# 高性能计算: 计算中心的资源分布

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究



# 计算中心: 多用户系统

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

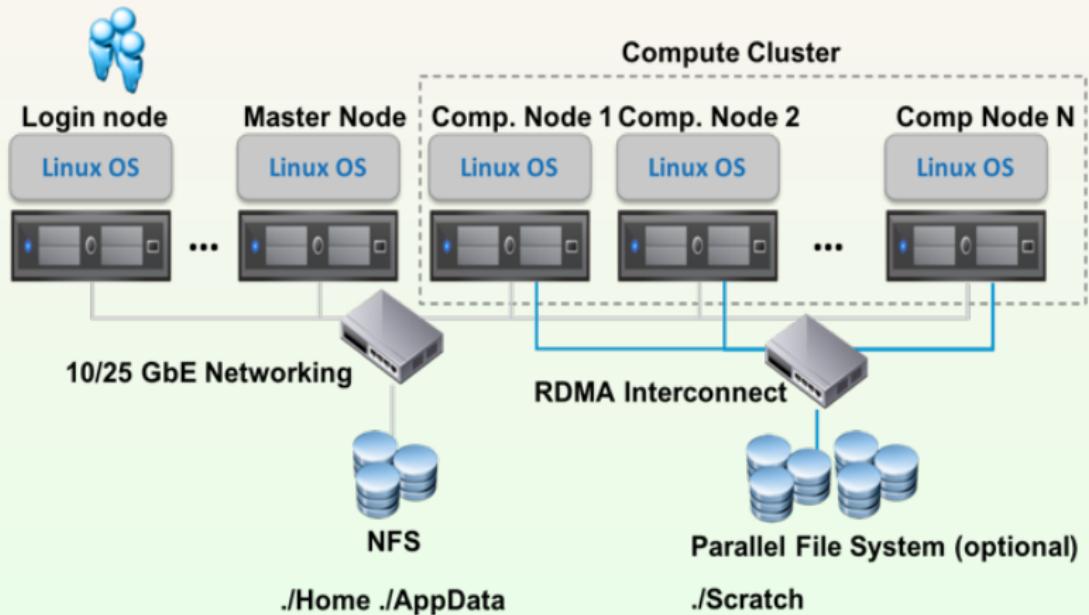
高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究



# 高性能计算: 硬件

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

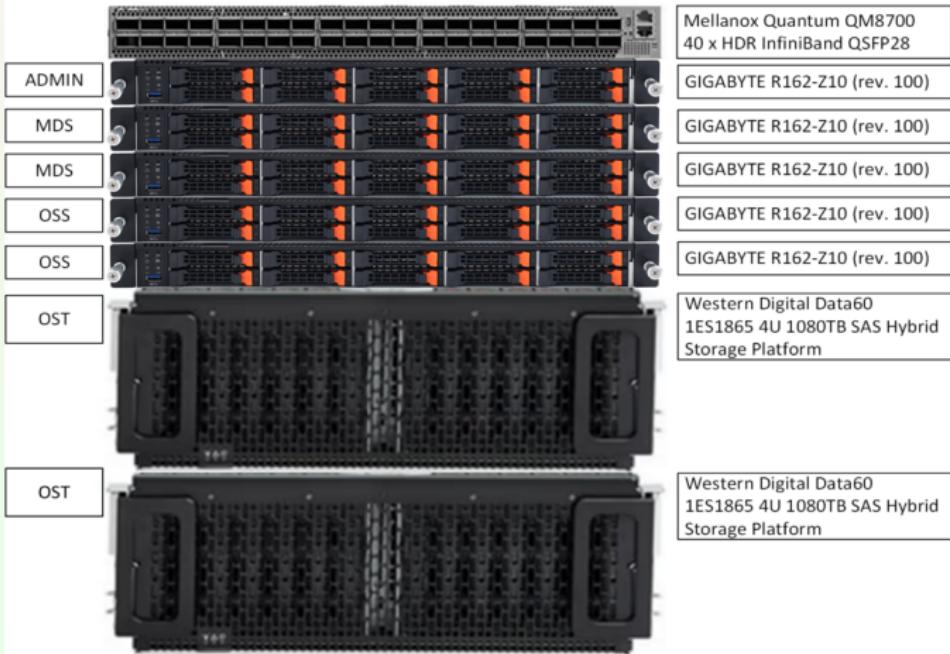
高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究



# 高性能计算: 硬件



高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

# 高性能计算: 硬件

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

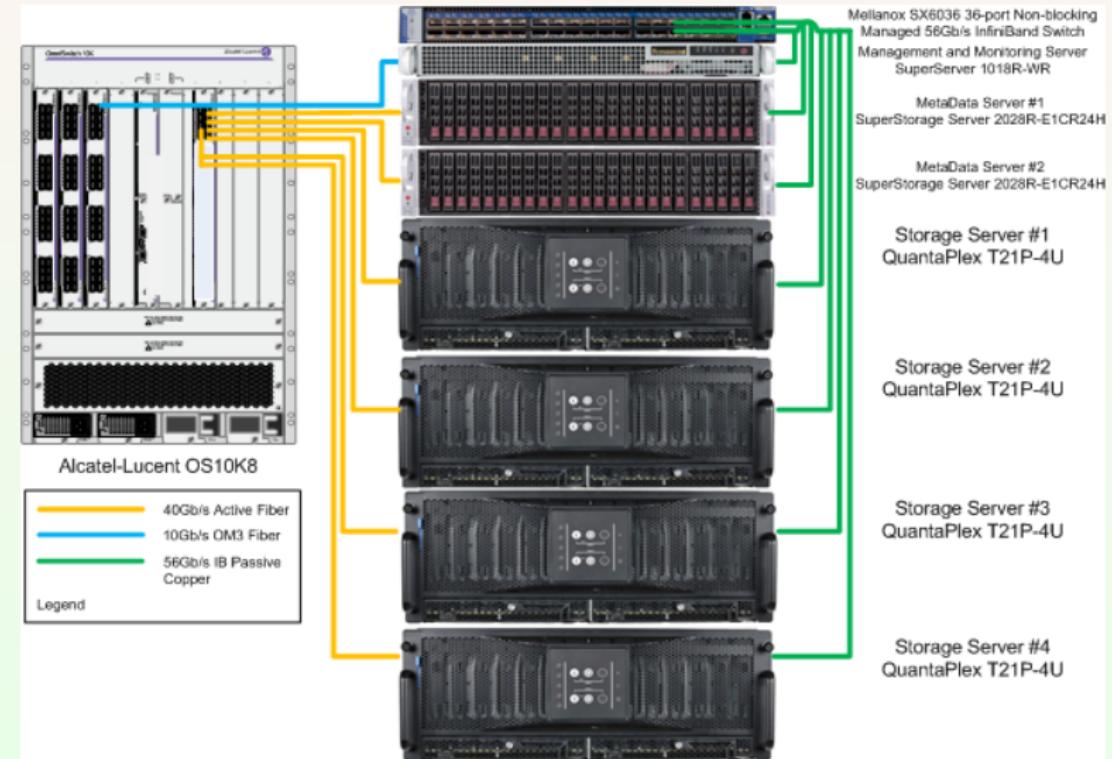
高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究



# 高性能计算与高通量计算

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

## ■ HPC (High Performance Computing):

Large amounts of [simultaneous] computing power for comparatively short periods of time

## ■ HTC (High Throughput Computing)<sup>1</sup>:

Large amounts of computing over significantly longer periods, not necessarily all at the same time

---

<sup>1</sup> 高通量的概念最初出现在实验领域，早期的材料研究、制药研究主要通过大量备选材料试错，最终才得到合适的材料或药物重要功能成分，这就是一种高通量筛选。文献中常会提到“高通量”和“组合方法”(Combinational approach)，但很少区分两者的区别：“高通量”指用户产生或处理的数据量极大，没有计算机自动处理无法完成；而“组合方法”是针对影响研究对象的各种可能自由度的分门别类研究。换言之，高通量考虑的是利用计算机“一视同仁”地自动化式处理海量数据，而组合方法更强调对特定影响因素的筛查和组合研究

# 高性能计算与高通量计算

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库  
库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

**High  
Performance  
(Capability)**

**High  
Throughput  
(Capacity)**



**Fine-grained Applications**  
- Many-node  
- Few concurrent runs  
- High interconnect use

**Course-grained Applications**  
- Single-node  
- Many concurrent runs  
- No interconnect use

# 高性能计算与高通量计算

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

HPC 和 HTC 作业可以在相同的集群架构上运行，但它们使用资源的方式不同：

HPC: 在需要时使用：

- 运行需要中间结果快速通信以执行计算的作业
- 在相对较短的时间内大量使用计算资源

HTC: 在需要时使用：

- 运行许多通常相似但不高度并行的作业
- 使用不同的输入运行相同的程序
- 运行不相互通信的作业
- 利用使用网格启用技术的物理分布式资源
- 利用多个计算资源在较长时间内执行计算任务

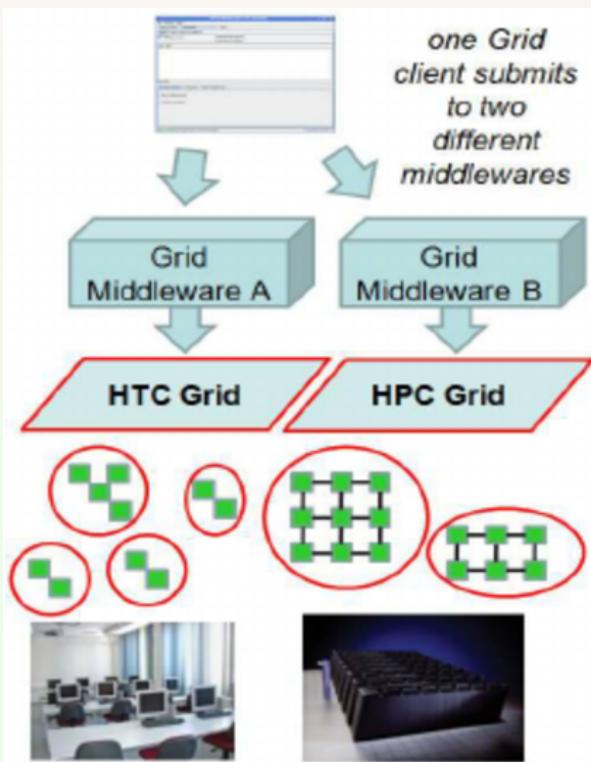
- HPC 作业通常涉及在许多处理器上运行并行软件的单个实例

在计算的各个实例中，结果在处理器之间进行通信，需要并行环境

- HTC 作业通常涉及在多个处理器上同时运行软件的多个独立实例

串行系统适用于这些要求

# 高性能计算与高通量计算



# 复杂计算流程: 高性能与高通量

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

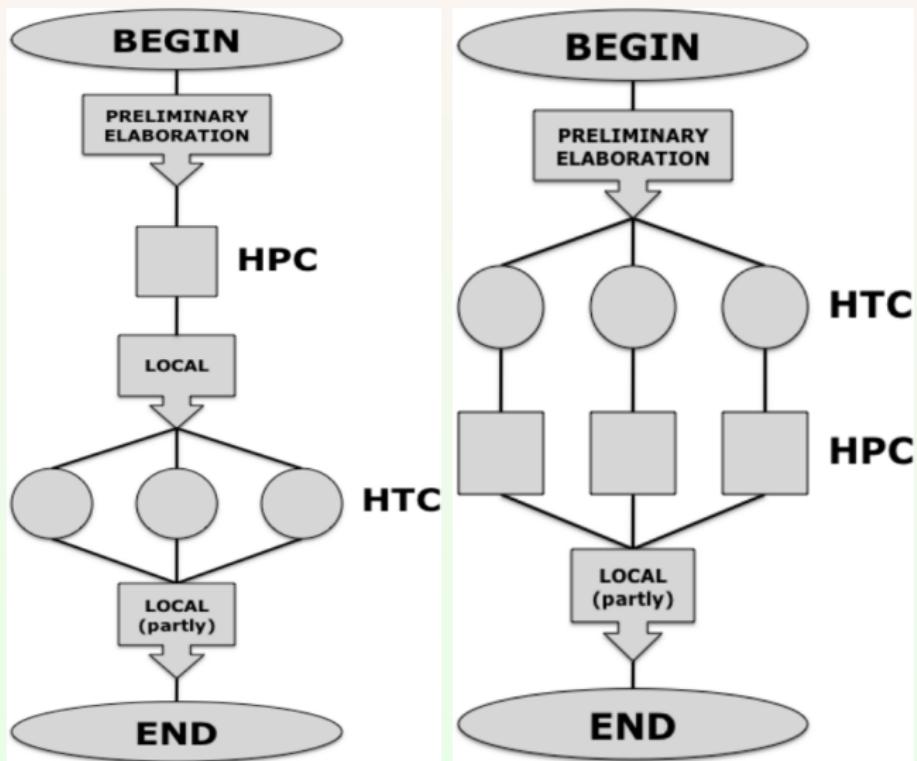
高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

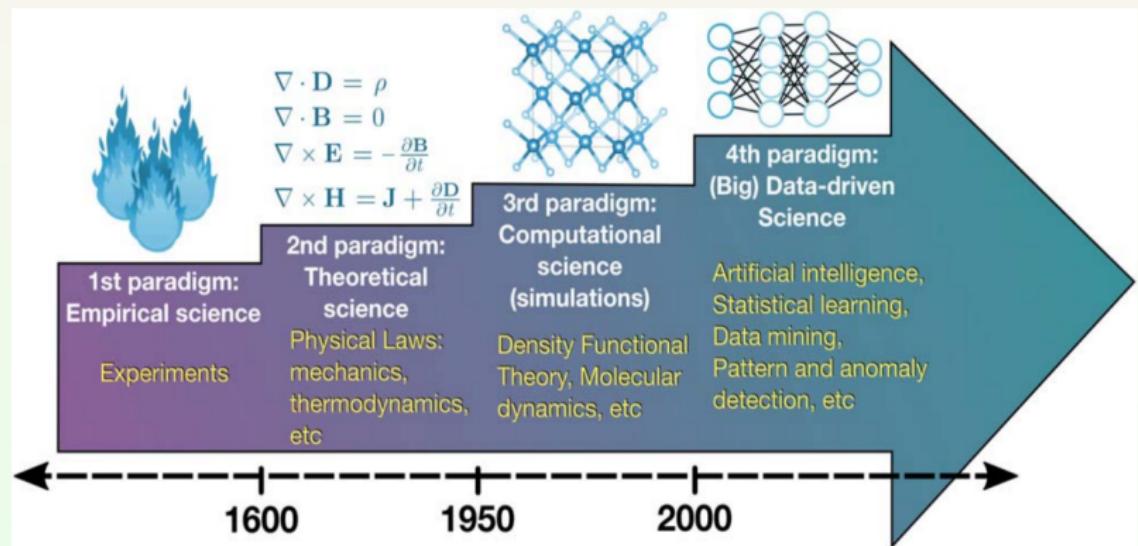
数据挖掘与第一原理材料研究



# 科学研究的范式变更

高通量计算流程、数据库和机器学习简介

- 高通量与高性能计算
- 高通量计算材料自动流程
- 第一原理数据库
- 机器学习简介
- 机器学习算法
- 数据挖掘与第一原理材料研究



# 数据驱动的科学研究

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

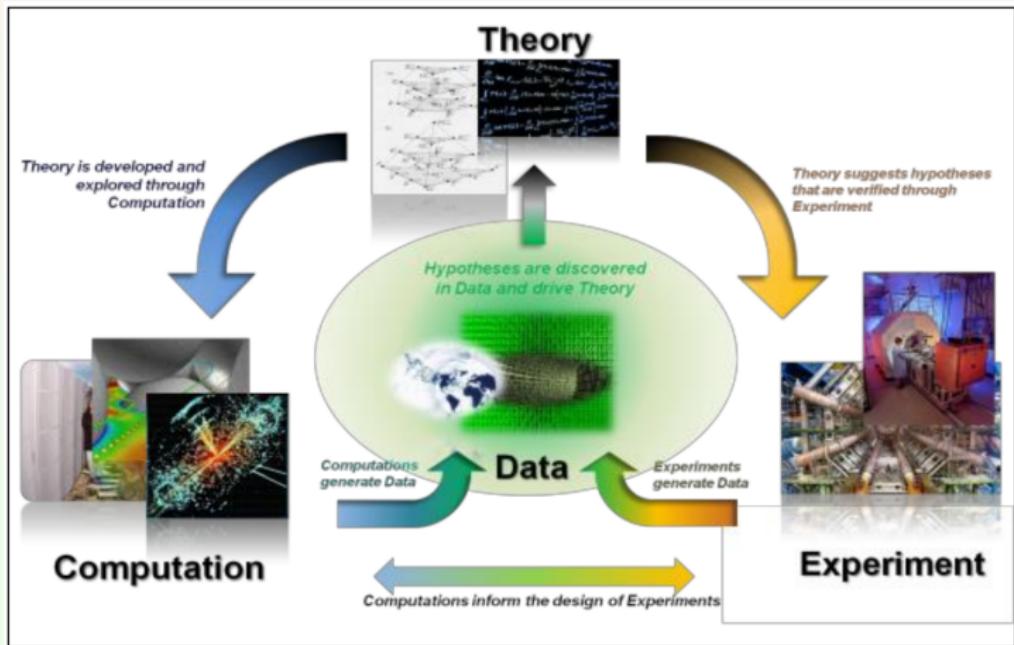
第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

前所未有的计算能力和大规模的数据收集能力



科学的新驱动力: **密集数据 + 人工智能**

# 高通量计算流程

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

高通量计算流程主要任务包括三方面：

- **增加材料模拟数据**: 主要包括 DFT、MD 计算的材料数据
- **存储材料物性数据**: 系统存储材料数据，用以构建材料数据库  
材料数据库包括通用材料数据库或特定目标的材料数据库
- **检索材料数据**: 对存储的材料数据实施检索和分析，服务材料性能提升需求

高通量计算流程的主要目标之一就是构建材料数据库

高通量计算流程与数据密不可分：著名的高通量计算流程都有相应的数据库

- AFLOW 的数据库为 AFLOWLIB
- MP 的数据库为同名的 Materials Project
- ASE 的数据库为 CMR (The Computational Materials Repository)

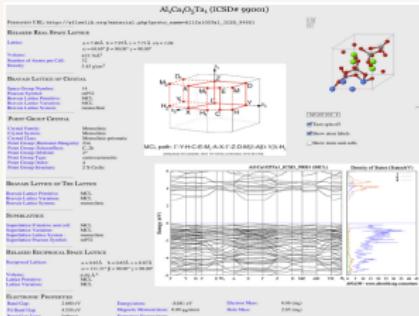
## 国外已有的计算平台

# 高通量计算流程、数据库和 机器学习简介

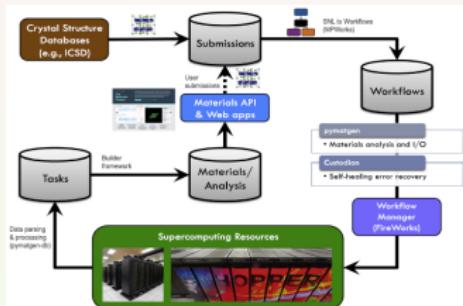
高通量与高性能  
计算

高通量计算材  
料自动流程

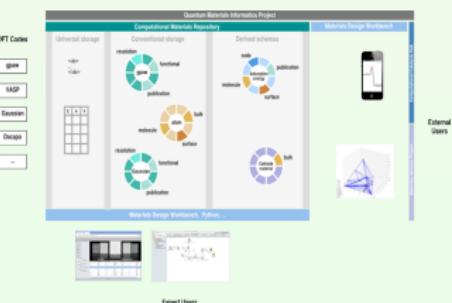
机器学习简介



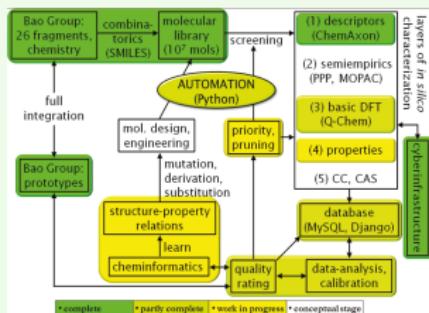
### (a) Auto-FLOW (AFLOW)<sup>[3]</sup>



### (b) Material Project (MP)<sup>[4]</sup>



(c) Quantum Materials Informatics Project (QMIP)



(d) Clean Energy Project (CEP)<sup>[6]</sup>

# 国内已有的计算平台: MatCloud

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究



Fig.: 中科院计算机网络信息中心 杨小渝团队开发<sup>[1, 2]</sup>

# 高通量计算自动流程

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

当前高通量计算的自动处理流程集中在材料模拟的计算数据收集和实现计算结果自动入库为主

材料第一原理计算的软件有很多，输出文件的格式千差万别，通过第一原理计算构建相对完善的材料数据库需要耗费相当的计算资源和人力

自动流程面向的对象是数据，要实现材料计算过程的自动处理，首先需要解决的是数据格式的规范化

- **数据规范化类型**: 面向各类软件输出数据的自动化提取与传输

根据软件的计算特点，组织多种软件实现材料物性计算: 灵活性较好，但一般只支持相对简单的计算流程

- **流程规范化类型**: 面向材料计算软件的标准化流程，通过数据库支持

利用数据库技术将自动流程组织得更复杂多样，并作为数据库条目存储下来: 稳定性较好，但计算的材料物性受软件能力的限制较多

**主要的自动流程采用 Python 语言实现**: 跨平台、模块化组织灵活

# 计算平台的功能和总体架构

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

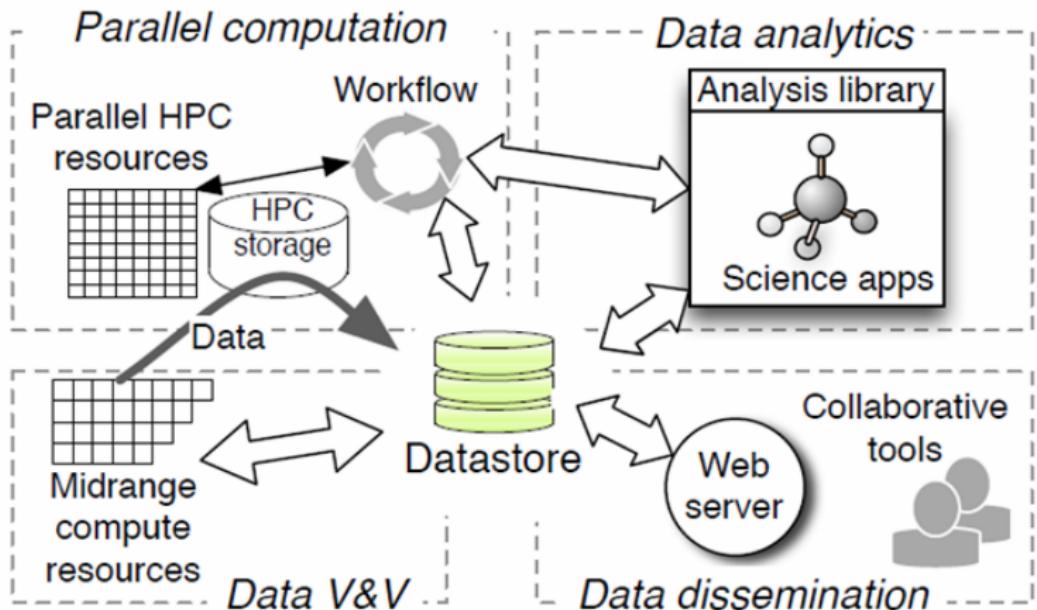


Fig.: The schematic framework and platform of all those project.

# 材料计算软件发展现状

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究



# 国产第一原理计算软件现状

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

Xian-CI



XMVB

BNU-CI



LOW SCALING  
QUANTUM CHEMISTRY

**BDF**

Beijing Density Functional Program Package

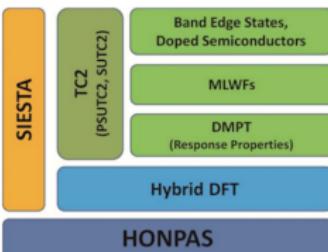


M $\ominus$ MAP

Molecular Material Property Prediction Package



PWMat



An Efficient Structure Prediction Method and Computer Software

LSASP

Large-scale atomic simulation package

# ASE 自动流程的设计与管理

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

数据规范化型的自动处理以 ASE 为典型代表，通过各类 Python 模块，支持多种 DFT-MD 软件

**ASE 特色：**模块加载式计算流程控制，符合复杂多尺度计算场景

- 灵活的建模功能

- ① 简单组织：原子直接构成分子
- ② 理想周期体系（包括一维、二维、三维）
- ③ 表面和表面吸附，可指定吸附位

- 丰富的软件接口

提供了包括绝大部分第一原理和分子动力学计算软件接口，方便组合实现多尺度计算

- 不依赖软件的优化与动力学模拟

适合复杂材料物性模拟的优化和多种动力学过程模拟

- 多样化的数据库类型

# ASE 的结构生成模块

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

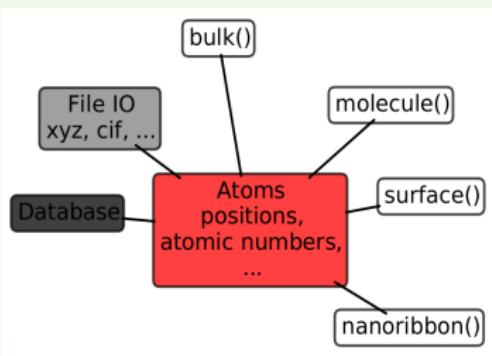
机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

## 材料结构生成模块主要功能:

- 生成各种计算软件所需的结构模型  
包括原子、分子、晶体、表面和界面等
- 读入各种格式的结构模型文件  
包括 xyz、POSCAR、cif 等 65 种格式
- 各类结构模型统一以 traj 或 json 格式  
写入数据库  
实现计算模型数据的标准化



```

from ase import Atoms
h2 = Atoms('H2', [(0, 0, 0), (0, 0, 0.74)])
from ase.structure import molecule
water = molecule('H2O')
from ase.lattice import bulk
si2 = bulk('Si')
from ase.lattice.spacegroup import crystal
rutile = crystal(['Ti', 'O'],
                 basis=[(0, 0, 0), (1/4, 1/4, 0)],
                 spacegroup=136,
                 cellpar=[a, a, c,
                          90, 90, 90])
from ase.lattice.surface import fcc110
slab = fcc110('Pt', (2, 1, 7), a=4.0,
              vacuum=6.0)
from ase.lattice.surface import add_adsorbate
add_adsorbate(slab, 'H', site='hollow')
from ase.lattice.surface import surface
nasty_cut = surface(rutile, (1, 1, 0),
                     layers=5)
from ase.io import read
q = read('quarts.cif')
import ase.db
con = ase.db.con('mystuff.db')
atoms = con.get_atoms(foo='bar', H=0)
# same as this:
atoms = read('mystuff.db@foo=bar,H=0')
  
```

# ASE 特色: 软件接口丰富

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

## Calculator 模块支持的可选软件



Fig.: The integrated calculator in ASE.

# ASE 的模块: Calculator 和 checkpointing

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

## Calculator: 支持各类计算软件的主要模块

- 模块封装了 DFT-MD 计算软件，支持物性计算、功能分析
- 模块集成了全局结构搜索算法 (Base-hopping 和 minima-hopping 算法)、反应动力学模拟 NEB 算法和势能面鞍点搜索算法、MD 模拟算法、几何结构优化算法和分子振动与声子振动分析算法
- 模块支持材料物性自动化计算，数据以标准化形式存入数据库

模块启动计算依次执行以下步骤:

- 1 生成计算软件所需的输入 (控制) 文件
  - 2 启动软件，以子进程方式开始计算过程
  - 3 进程守护直至计算子进程结束
  - 4 根据要求解析计算软件生成文件，并可将计算结果以 json 格式写入数据库
- **优点:** Python 模块与计算软件的交互简单
  - **缺点:** Python 执行过程中会面向较多的 I/O 处理，运行效率不高

checkpointing: 协助用户排查、定位错误和重启计算的模块

增加 ASE 对计算流程的控制能力

# ASE 特色: 数据库的良好兼容性

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

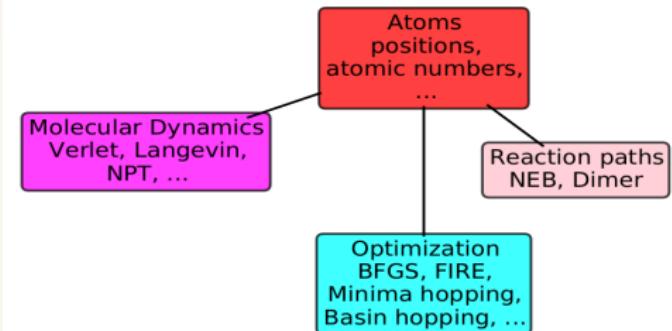
机器学习算法

数据挖掘与第一原理材料研究

ASE 的材料数据库 CMR 采用关系型数据库管理系统 MySQL

将规范化的数据转成数据库文件 (称为 cmr-文件), 并要求数据库中的文件名尽可能与原文件保持一致

- 用户不进入数据库即可对数据进行检验
- 存储数据有更大的兼容性



## Taxonomy:

- Atoms object (positions, atomic numbers, ...)
- ID, user-name, creation and modified time
- Constraints
- Calculator name and parameters
- Energy, Forces, Stress tensor, dipole moment, magnetic moments

## Folksonomy:

- Key-value pairs
- Keywords

## Additional stuff:

- extra data (band structure, ...)

Back-ends: JSON, SQLite3 and PostgreSQL.

# MP 自动流程的架构

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

流程标准化型的自动处理以 MP 的流程控制 FireWorks 为代表

- **设计目标:** 围绕 VASP 作业高通量并发提交与过程监控
- **设计方案:** 开发针对不同计算场景的功能模块

## 1 Pymatgen

**前处理:** 计算模型的分析与预处理

**后处理:** 计算结果的可视化

## 2 FireWorks

**数据库支持的计算流程设计与管理:** 复杂的工作流可以数据形式保存到 MongoDB 数据库中，用 FireWorks 设计的工作流具有较高的稳定性

## 3 Custodian

**计算流程容错与应对:** 提供计算过程错误判断接口，由用户提供解决策略和针对性设计

# Pymatgen 的模块结构

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

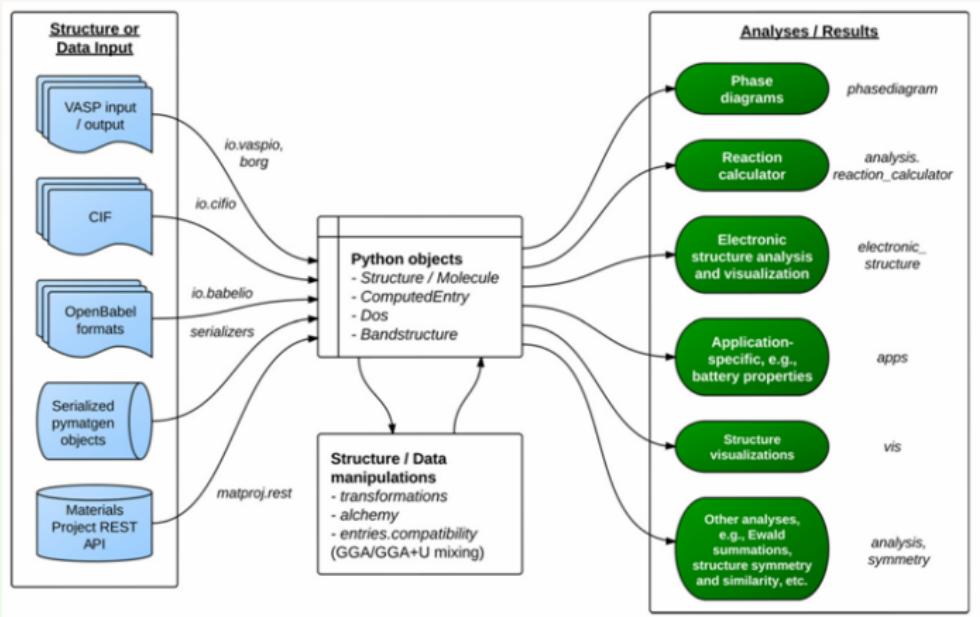


Fig.: Overview of a typical workflow for pymatgen.

# Pymatgen 可展示的材料物性

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

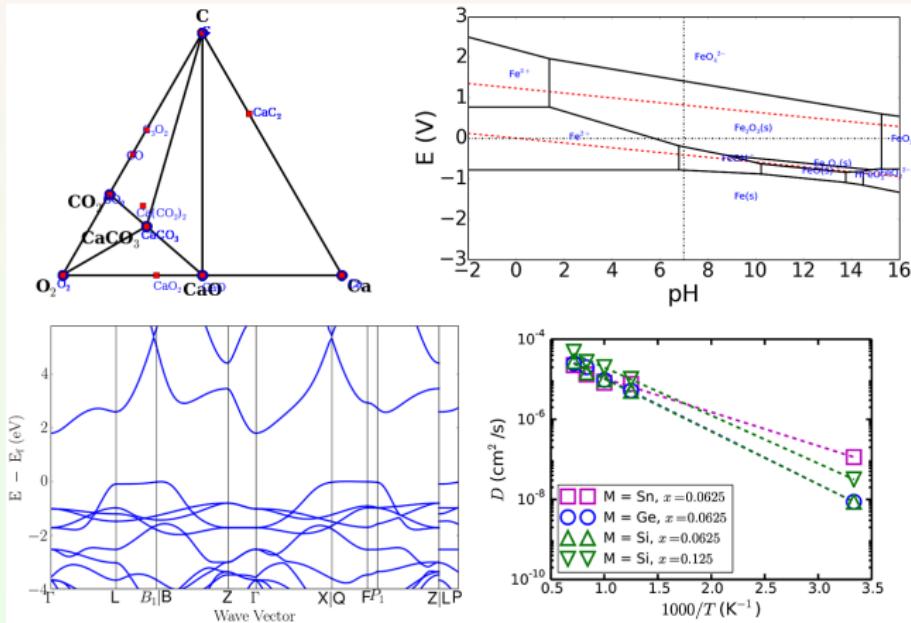
高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究



**Fig.:** Top left: Phase; Top right: Pourbaix diagram from the Materials API. Bottom left: Calculated bandstructure plot using pymatgen's parsing and plotting utilities. Bottom right: Arrhenius plot using pymatgen's Diffusion Analyzer.

# FireWorks 的模块结构

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

FireWorks 是一款开源的通用工作流定义、管理和执行软件，支持 Python 运行

FireWorks 的自动流程采取中心化的“发布-执行”模式

- 流程发布 (称为 LaunchPad):

LaunchPad 是工作流的主管者，主要负责自动流程的定义、分发、排队、增删和对工作流的反馈与响应

- 流程执行 (称为 FireWorkers):

FireWorkers 是工作流的执行者，包括一个或多个计算资源 (个人计算机、小型工作站、超级计算机等)

FireWorkers 从 LaunchPad 处获得计算任务，执行完毕后再将计算结果返回到 LaunchPad

# FireWorks 的模块结构

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

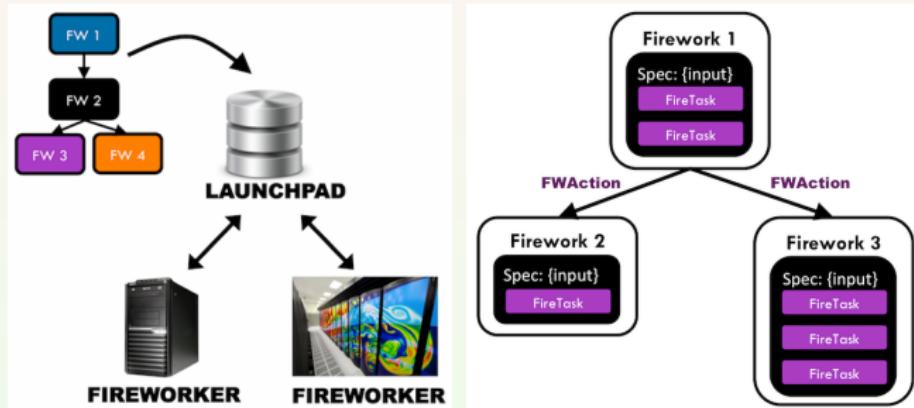


Fig.: The basic infrastructure of FireWorks.

FireWorks 发布的工作流程由三层嵌套结构组成:

- Firetask: 基本执行单元，是执行计算的最基本脚本命令或 Python 命令。
- Firework: 组织基本执行单元构成任务单元组，并指定各基本执行单元所需的参数。
- Workflow: 彼此相关联的任务单元组构成完整的工作流程:  
FireWork 之间的数据传递、任务执行序列等由 FWAction 完成。

# FireWorks 的模块结构

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

- FireWorks 是以任务单元组为基本组成的来实现工作流程的，任务单元组之间依靠数据传递相关联，流程执行完毕也将返回数据，FWAction 模块主要负责任务单元组之间的数据传递和任务分配。
- FWAction 允许用户根据需要设计和更改流程参数、增添、删减和改变流程(子) 单元组，这一模块大大增加了 FireWorks 工作流的灵活性。

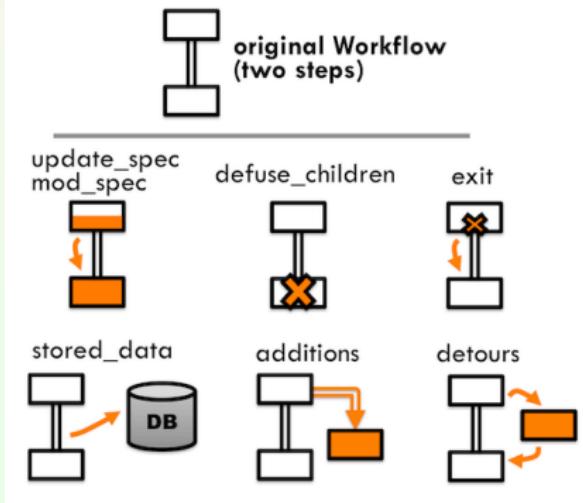


Fig.: Schematic diagram for data transfer and processing between units in FireWorks

# FireWorks 的模块结构

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

这种“发布-执行”结构使得计算任务与软件、硬件高度解耦，用户可根据需要随时向 LaunchPad 添加新的工作流，承担计算任务的 FireWorkers 彼此也可以是完全异构的，具有很好的机动性。

对于材料第一原理计算自动流程而言，一个 DFT 计算过程就是一个 Firework，可以分解为：

- 1 指定控制参数：参数在数据库 Json 中存储，由 Spec 传入
- 2 计算控制文件生成：每个 Firetask 生成一个控制文件
- 3 DFT 计算作业提交：产生一个 Firetask

在此基础上，可以通过 FWAction 修改控制参数，将 DFT 计算单元组组织成完整的材料第一原理计算流程，并将最终结果直接导入材料计算数据库。

# Custodian 的容错逻辑

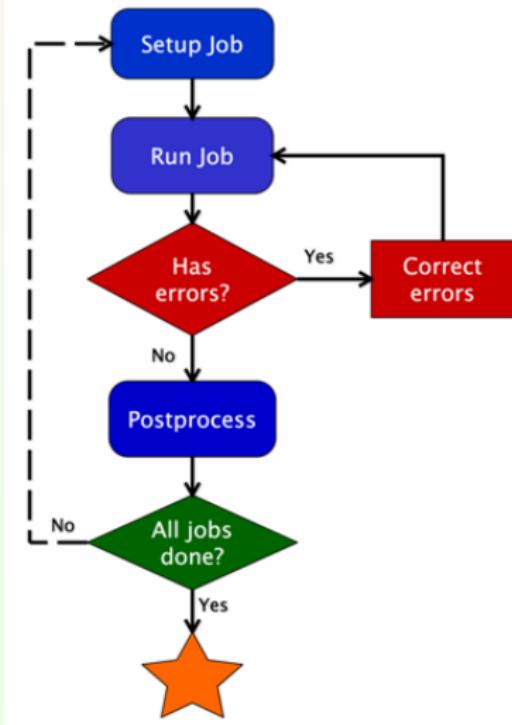


Fig.: Overview of the Custodian workflow.

# atomate: 计算流程控制示范

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

## Base libraries

**pymatgen**

(materials analysis and I/O)

Custodian

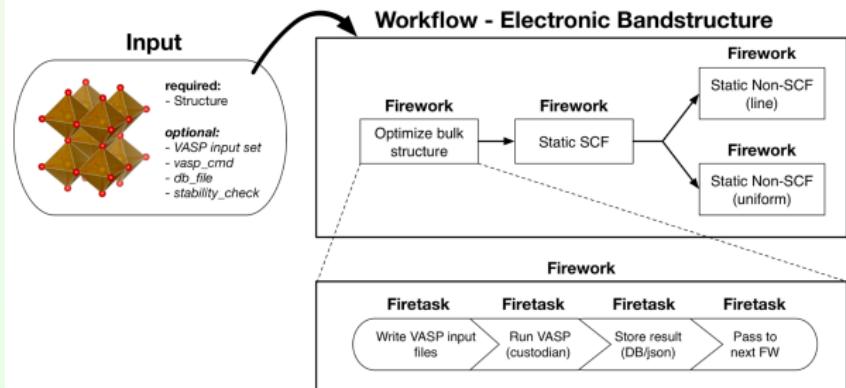
(calculation error recovery)

FireWorks

(workflow package)

**atomate**

(materials calculation environment)



# 早期材料数据库: SpringerMaterials

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

SpringerMaterials: 世界上最古老也是完备的材料数据库，旨在创建集成材料学信息、性质和使用平台

数据库主要由以下部分组成：

- 全部 Landolt-Börnstein 丛书 (自 1883 年起)，是以基础科学为主的大型科学与技术数值与函数关系的工具书
- 全部 Pauling File 无机材料数据库，收集了从 1900 年迄今超过 21000 种出版物中的无机晶体结构、衍射、相图和物理属性数据
- Dortmund Data Bank 中的纯液体和二元混合物的热物理数据
- 吸附材料数据库 (Adsorption Database)
- 聚合物热力学数据库 (Polymer Thermodynamics Database)
- MSI Eureka 数据库：无机材料相图、相反应和热力学数据

# 早期材料数据库: 其它数据库

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

除 SpringerMaterials 之外，比较完整的数据有

- 无机晶体结构数据库 (Inorganic Crystal Structure Database, ICSD): <http://www.fiz-karlsruhe.com/icsd.html>

服务器位于德国的 FIZ Karlsruhe，收录了自 1913 年以来超过 185000 条矿物、金属和其他无机固体化合物 (含 2000 多元素单质、34500 多二元化合物、68000 多三元化合物、66000 多四元及多元化合物) 的晶体结构数据

- CRYSTMET 数据库
- Pauling 无机材料数据库: <http://www.paulingfile.com/>
- Pearson 晶体数据库: <http://www.crystalimpact.com/>
- ...

这些实验数据库对于材料学研究提供了重要的帮助

- 材料的覆盖范围非常广泛，有限的数据库不可能穷尽
- 有很多材料的物性数据很难通过实验直接得到

原子尺度的材料物性数值模拟可以有效补偿实验方法的不足

集成了实验数据和计算数据的新材料数据库为探索新材料合成、性能优化开辟了新的研究思路

# AFLOWLIB

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

AFLOWLIB 是由高通量 DFT 计算框架 AFLOW 生成的材料信息数据库: <https://www.aflowlib.org>

该数据库涵盖相图、电子结构和磁学性质等信息，源代码和主要材料数据由 Duke 大学开发和运维

- 数据库包含 630,000 种以上的合金热力学条目，涵盖了 ICSD 中的 52,000 多种化合物，3,000 多种基本化合物，330,000 多种二元化合物和 262,000 多种 Heuslers 金属间化合物
- 数据库提供在线的界面搜索，包括计算的细节、电子结构、磁学性质和热力学性质，所有的计算结果都是由 AFLOW 高通量计算软件生成的
- 计算数据在数据库中以 SQL(MySQL) 形式存储，检索方便

<https://www.aflowlib.org> 提供数据交换的应用接口 RESTful API 服务

数据库中的材料数据可以表示为多种格式，包括 HTML/JSON/DUMP/PHP/TEXT/NONE 等，极大地方便了用户在新材料研发中的数据挖掘需求

Materials Project 是美国材料基因组项目支持高通量第一原理计算的软件和数据库: <https://www.materialsproject.org>

MP 基于 Apache、Python 和 Django 开发, 在开源软件平台 github 上发布全部代码, 面向全世界的开发者支持: <https://github.com/materialsproject>

根据 MP 官网信息, 目前可提供的数据包括:

- 59,000 多种化合物的信息, 41,000 多种能带结构数据和 1,300 多种材料的弹性张量数据
- 2,200 多种 Li 的插层电极和 19,000 多种 Li 的转换电极数据
- Materials Project 的扩展性极好

除了 Pymatgen、FireWorks 和 Custodian 的支持, 其数据库采用基于 MongoDB 的 NoSQL 数据格式文件, 弥补了传统数据库的扩展性差和灵活性不够的问题

Materials Project 同样通过 API 支持 RESTful 服务

CMR 和 ASE 都是由丹麦技术大学 QMIP 项目开发的

ASE 主要服务第一原理计算任务生成、计算执行、结果分析和可视化

CMR 则主要面向计算数据存储: <https://cmr.fysik.dtu.dk>

- CMR 提供了超过 30,000 条数据记录
- CMR 提供了三种不同的用户界面 PHP/HTML 和 Python，方便不同经验的用户存储、检索电子结构计算数据

与 ASE 类似，CMR 面向用户对数据库的多元化需求，实现软硬件的尽可能兼容

- 软件方面: ASE 和 CMR 所有源代码开发都是基于松耦合模式，方便用户根据习惯使用软件  
主要是面向开源系统，特别是 Linux 操作系统，Apache 浏览器，MySQL 数据库，用户界面采用 PHP/HTML，合成 LAMP(Linux,Apache,MySQL,PHP) 套装
- 硬件方面，软件可以安装到台式机、集群或超级计算机上，对于小的台式机则无须安装数据库和网络服务器
- CMR 使用 SQLite 数据库模式而不用 MySQL 服务器模式存储和检索材料数据  
当用户有需要时，CMR 的数据处理模块可将 SQLite 数据库文件上传到 MySQL 服务器上，或者将数据文件转呈 XML/JSON 格式，以方便数据交换

# ESP

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

ESP 的数据库是 Uppsala 大学为加速新的功能材料的研发进程而开发的: <http://gurka.fysik.uu.se/ESP/>

- ESP 数据库中的电子结构数据是依据 ICSD 的无机化合物的结构数据, 由 FP-LMTO 方法通过第一原理计算产生  
自 2002 年 ESP 发布以来, 共收录了 60,000 多条化合物的电子结构数据
- 用户最多允许查阅含有五种元素的化合物的电子结构信息

应用数据挖掘技术, 由 ESP 的 60,000 多条化合物电子结构信息, 已经预测了 17 种潜在可能的强拓扑绝缘体

ESP 的官网提供的信息显示: 采用大规模计算结合数据挖掘技术提供的 17 种可能的化合物中, 成功预测到了第一个强拓扑绝缘体 (topological insulator)  $\text{Bi}_2\text{Se}\text{Te}_2$ , 而且预测很快就得到验证

ESP 的数据库和能带分析工具的代码都没有公开  
近年来, ESP 数据库开发进展缓慢

ESTEST 是加州大学 Davis 分校开发的材料数据库:

<http://estest.ucdavis.edu/>

用于软件 Qbox、VASP、Quantum Espresso、Siesta、ABINIT 的结果和电子结构计算软件

Exciting 的结果检验和对比

通过数据库的网页界面，可将各类软件的输入/输出文件转换成统一的 XML 格式，便于数据存储、比较、分析

每个 XML 文件由文件名和数据库存储位置区别

- ESTEST 数据库总共包括 400 多条 Qbox 的计算记录，10 条左右的 VASP 计算记录，Quantum Espresso 和 Siesta 计算记录各 70 多条，150 多条 ABINIT 计算记录，Exciting 的记录约 30 条
- ESTEST 的网页界面提供了类似 Google 的简明搜索栏，网页还提供了多个搜索下来菜单，可根据软件名称、计算元素等检索
- 数据库的网页界面可提供数据、表格和图片形式的电子结构的模拟信息：包括计算参数、原子信息、能量结果和结构应力与光谱信息
- 网页界面还提供了一些插件，可对诸如原子距离、晶格、能带结构、态密度、能量和结构应力或其他记录转换成一种或多种数据格式  
例如 band\_structure\_plot 插件可将 XML 格式的能量本征值转成电子能带图

这种格式转换插件丰富了 ESTEST 生成、模拟计算结果的能力

需要说明的是：必须是注册用户才能数据库的源代码

以完整的高通量框架的角度看，MAST 更像是对 ESTEST 的辅助，重点用于工作流管理和后处理：

[http://pythonhosted.org/MAST/0\\_0\\_introduction.html](http://pythonhosted.org/MAST/0_0_introduction.html)

MAST 的基本框架由 MP 的 Pymatgen 和 Custodian 模块和 ASE 组合搭配实现

MAST 开发了自动计算流程以及数据库，而没有使用 FireWorks 和 MongoDB

MAST 的数据主要来自 VASP 的数据拟合

- MAST 的数据库主要收集了带空位的稳定纯元素的数据，现有面心立方的 49 种元素和六方密堆积的 44 种元素
- MAST 数据库主要用于研究基本的空位扩散，收集的数据包括：空位形成焓、空位迁移焓、内聚能和体模量

MAST 的数据生成和管理主要依赖于操作系统和 I/O 性能，运行 MAST 无须额外配置专门的数据库软件

MAST 的数据目前还不能通过网页界面访问，很多功能还有待完善

CEPDB 是哈佛大学的 CEP 项目的数据库：

<https://gist.github.com/jessiegt/5642460f061a39d1820e>

主要服务于研究有机光电材料

根据 CEPDB 的网页信息显示，

- 数据库共拥有 2,300,000 多个分子的图谱信息和 22,000,000 多个分子结构 (由超过 150,000,000 个 DFT 计算得到) 共计超过 400 TB 的数据，
- 数据库除了有第一原理计算结果，还收录了文献中的实验数据，用于作为 CEPDB 的训练和标定数据集

CEPDB 使用 Python 的网页框架 Django 在 *in silico* 上开发的高通量计算 MySQL 数据库，数据由量子化学计算软件 Q-Chem 得到。CEPDB 的强大算力主要由 IBM 的公益分布式计算项目全球社群网格 (World Community Grid, WCG) 提供，还有一些算力来自 Harvard 的 FAS Odyssey 集群、NERSC 和 TeraGrid 的计算资源以及其它一些计算资源。

CEPDB 的数据都是开放的，但高通量计算软件部分不开源。不过 CEPDB 对全世界各研究组上传的实验数据持开放态度。

计算化学对比和基准数据库 (Computational Chemistry Comparison and Benchmark Database, CCCBDB) 归属于美国国家标准和技术协会 (National Institute of Standards and Technology, NIST)

汇集了 1600 多种气相原子和分子的实验和第一原理计算的化学热力学数据，很好地弥补了常见数据库只有固体材料数据的不足

- CCCBDB 提供简单的在线服务界面，用户可以检索到指定化合物的实验数据和计算数据，实验和计算数据的对比也验证了数据的可靠性
- CCCBDB 允许用户优先使用几何结构、振动模式、熵、能量和静电性质而非分子结构作为检索的关键词  
避免引入复杂的筛选组合，方便用户快速得到检索结果
- CCCBDB 是一个庞大的数据库，收录了总计有超过 460,000 个计算条目  
因为其没有开源，降低了数据库的影响力

AiiDA 是用于原子尺度材料自动模拟、管理、共享和复制的软件框架: <http://www.aiida.net/>

软件由 THEOS-EPFL (Theory and Simulation of Materials-École Polytechnique Fédérale de Lausanne) 的实验室和 BOSCH 公司于 2012 年开发和维护

## AiiDA 数据库

- 网页编程框架用 Python 维护
- 数据库用 Django 开发

可满足计算物理、化学和材料科学的研究的多种不同需求

数据库提供 SQLite/MySQL 和 PostgreSQL 三种形式的数据

推荐使用 PostgreSQL

当用户使用数据库的 RESTful API 接口时，数据将以 JSON 格式呈现

# Alloy Database

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

Alloy Database 数据库提供 VASP 计算的合金结构和结合能数据，其在线服务器位于卡内基梅隆大学 (Carnegie Mellon University) 物理系: <http://alloy.phys.cmu.edu>

## Alloy Database 数据库的主要功能

- 包括 200 多种二元合金、100 多种三元合金和 20 多种四元合金的材料数据
- 用户可以检索二元、三元和四元合金的 VASP 计算数据
- 数据库提供包括合金组分的元素信息、Pearson 符号、合金模型以及计算所需的 POSCAR 文件、合金 Pearson 符号，基态总能和生成焓、化学组成和赝势等

Alloy Data 最早于 2006 年提供线上服务，由于其数据库服务代码从未开放，所以该数据库迄今尚未应用于高通量第一原理计算

Novel Materials Discovery (NoMaD) 是面向第一原理计算的电子结构数据生成、组织和交换的重要材料数据库: <http://nomad-repository.edu/cms/>

## 当前 NoMaD 数据库

- 包含 640,000 多条数据  
其中有 250,000 多条数据来自 <http://aflowlib.org>, 其余则来自 <http://cccbdb.nist.gov>
- 通过网页 <http://nomad-repository.edu/gui/#nomad> 共享数据
- 用户可通过元素、结构和计算方法、作者等信息检索到目标数据
- 供下载的数据主要是计算所需的输入/输出文件的压缩包  
少量关于计算本身 (如材料的化学式、空间群、计算软件及版本、作者和参考文献等)

NoMaD 目前可支持的计算软件数据涵盖: ABINIT、CASTEP、CRYSTAL、Exciting、FHI-aims、Gaussian、Quantum-Espresso、VASP 和 WIEN2k 等

NoMaD 的注册用户不仅可以浏览和下载材料数据, 还允许向 NoMaD 上传数据  
用户可指定上传文件为“开放”(open access) 或“限制”(restricted) 状态

NoMaD 的绝大部分数据 (约 540,000 多条) 是开放的, 而“限制”状态的数据最多三年后自动转成“开放”状态, 确保科学材料数据最大限度的共享和交换能力

NoMaD 除了提供材料计算的输入/输出文件, 未提供高通量计算框架, 一些重要的和有价值的材料数据 (如计算参数以及计算能量数据) 在数据文件下载之前已被隐去

# OQMD 与 qmpy

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

开放量子材料数据库 (Open Quantum Materials Database, OQMD) 是 2013 年前后，由美国西北大学 (Northwestern University) 的研究团队在超过 200,000 个 DFT 计算的晶体结构基础上发展起来的: <http://oqmd.org/>

数据主要是在西北大学高性能计算队列和美国国家能源研究科学计算中心 (the National Energy Research Scientific Computing Center, NERSC) 的机器上完成

- OQMD 累积的 DFT 计算化合物约有 290,000 条

这些化合物中约 10% 来自 ICSD，约 90% 来自简单化合物的组合，数据库中的化合物条目仍在增加

- 数据库收录的 DFT 计算的材料热力学和结构性质数据可以在线共享

- 数据库没有提供 RESTful 服务，但允许用户在 URL 地址栏输入检索化合物的字符串

如 <http://oqmd.org/materials/composition/Al2O3> 表示  $\text{Al}_2\text{O}_3$  的检索

软件 qmpy 是 OQMD 的 Python 的 API: <http://oqmd.org/static/docs/index.html>

qmpy 用 Django 开发，数据库是 MySQL 形式，数据主要由 VASP 计算得到

通过浏览器得到的 OQMD 检索数据是 HTML 格式的 (而非 JSON 或 XML 格式的)

OQMD 允许用户将全部数据 (约 4GB) 下载到本地并保存成一个文件，对于小型的计算用户，这将提升用户对计算数据的掌控和分析能力

# PyChemia

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

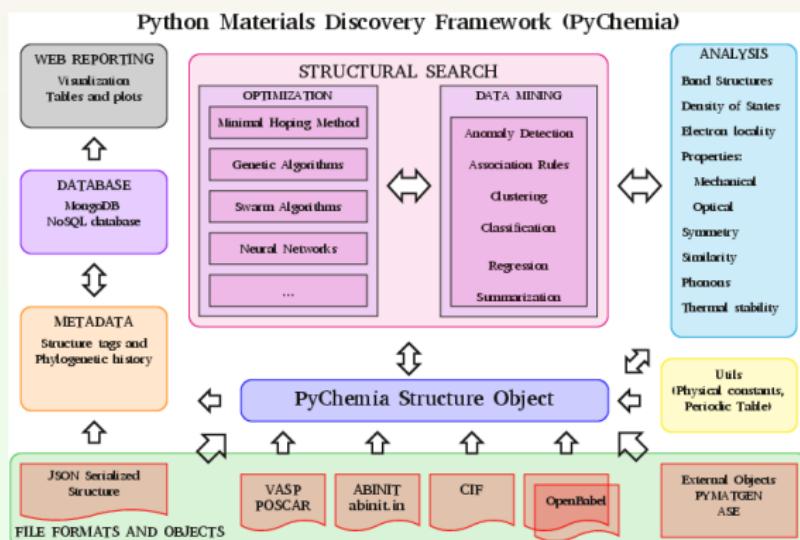
第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

Python Materials Discovery Framework (PyChemia) 是近年来才出现的面向 DFT 的高通量计算框架



**目标:** 通过涵盖 DFT 计算和 Minima Hoping 等各种方法来推动发现新材料，并将 PyChemia 计算的数据存入数据库，作为后续发现新材料的备选化合物

# PyChemia

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

PyChemia 作为 Python 的一个模块，全部代码是最初于 2014 年由西弗吉尼亚大学 (West Virginia University) 分享到 GitHub 网站：

<http://github.com/MaterialsDiscovery/PyChemia>

PyChemia 的新特色

- PyChemia 可支持的 DFT 计算软件涵盖 VASP、ABINIT、Octopus、DFTB+ 和 Fireball 等
- 利用 Python 模块的特点，PyChemia 完成了对 Pymatgen、ASE 和 qmpy 的集成
- 相比于其他高通量材料计算软件，PyChemia 更显著的特长是材料结构搜索，而不是简单的材料计算和数据采集

PyChemia 除了提供大的材料数据库作为结构搜索的基础，还可以将每一套结构搜索产生新的小型数据库

PyChemia 采用 MongoDB 的 NoSQL 数据库而非传统的关系型数据库形式，方便产生结构搜索需要的小型数据库

为兼顾不同 DFT 计算软件包材料结构以及各种物理性质研究的需要

PyChemia 由 Python、Django 和 MongoDB 搭建而成，其中 Django 是 PyChemia 用于显示计算的网页前端

目前 PyChemia 仍在开发中，尚未发布稳定版本的软件

# Atomly

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

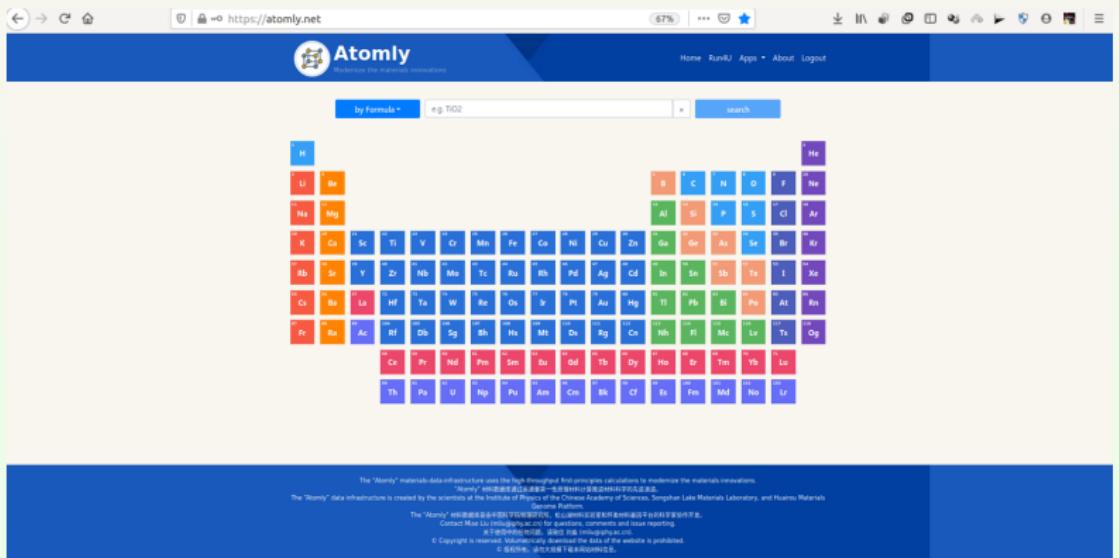
第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

Atomly 材料科学数据库是中国科学院物理研究所刘淼、孟胜团队合作开发的第一原理材料数据库: <https://www.atomly.net/>



The screenshot shows the Atomly website's main interface. At the top, there is a search bar with the URL <https://atomly.net>. Below the search bar is the Atomly logo and the tagline "Madeness the materials innovation". A navigation menu includes Home, RunDZ, Apps, About, and Logout. The central feature is a large periodic table where each element is represented by a colored square containing its symbol and name. The colors follow a gradient from blue to red across the rows. Below the periodic table, there is a footer with copyright information and a disclaimer.

The "Atomly" materials-data-infrastructure has won the High Efficiency And precision calculations to modernize the materials innovations.  
The "Atomly" data-infrastructure is created by the scientists of the Institute of Physics of the Chinese Academy of Sciences, Syngene Late Materials Laboratory, and Huamei Materials.

The "Atomly" data-infrastructure has won the High Efficiency And precision calculations to modernize the materials innovations.  
Contact Miss Liu ([lumeng@iphy.ac.cn](mailto:lumeng@iphy.ac.cn)) for questions, comments and issue reporting.  
© Copyright is reserved. Unauthorized download the data of the website is prohibited.  
© 中国科学院, 北京计算中心 著作权归我所有。

# Atomly

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

Atomly 的**目标**: 为科研领域带来材料数据工具平台，通过数据驱动新材料的筛选、预测和发现，提升材料研发的生产力

Atomly 已收录 349,000+ 种化合物，342,00+ 种能带结构和 68,000+ 种相图数据

- Atomly 提供了性能优异的高通量第一原理计算流程框架  
依托中科院物理所的松山湖材料实验室的计算资源，框架支持的高通量计算流程可以支持  $10^3$  数量级的材料体系同时计算
- Atomly 的网页前端，提供便捷搜索，能快速定位到指定材料数据，并以友好的方式呈现给用户
- Atomly 初步实践数据驱动的材料设计理念，能快速地完成高通量筛选，发现性能最优的候选材料

借助人工智能算法，Atomly 支持材料性质预测，能够以更快捷、更精准的范式预测材料结构和性质

# 机器学习 (Machine Learning, ML)

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

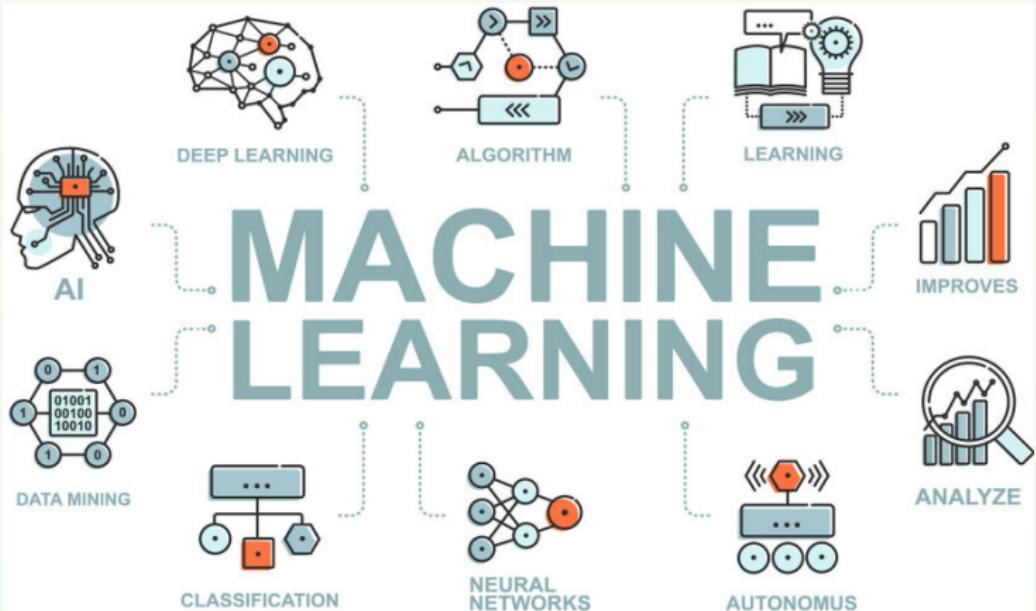
第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

机器学习是自动完成数据分析并提取数据关系的一类方法的统称



用已知的数据关系预测未知数据或辅助不确定条件下的决策过程

# 人工智能 (Artificial Intelligence, AI) 与机器学习

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

获取材料完整物性数据的成本，无论是通过实验手段还是计算模拟，代价都是比较高

- 高通量第一原理计算自动流程和数据库解决了材料物性数据的获取问题
- 利用数据挖掘技术，实现数据驱动的材料物性筛选、预测和提升的技术路线，有着特殊重要的意义

任何计算机模拟人类智能的算法都可以划归为人工智能，并非一定要应用机器学习算法，也包括决策树、知识库、计算机逻辑等算法

人工智能广泛应用于金融、导航控制、语言处理、游戏竞技、计算机可视化和生物信息学等领域

- 机器学习技术可以从大量数据中获得有价值的信息，尤其是面对高维复杂数据时，机器学习技术是确定数据间关系的有力的工具
- 机器学习领域的深度学习 (Deep Learning, DL) 是仿照生物神经网络<sup>2</sup>结构为主要代表的一种示类学习

<sup>2</sup> 神经网络结构意味着输入输出之间允许有多个类似神经的网络层

# 人工智能和机器学习的层次关系

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

传统定义界定的机器学习，是指无须借助解析程序，直接依靠数据来提升任务处理的性能，自从 1950 年代统计学、计算科学与技术和神经科学的发展，机器学习的研究发展到了更广泛的人工智能领域

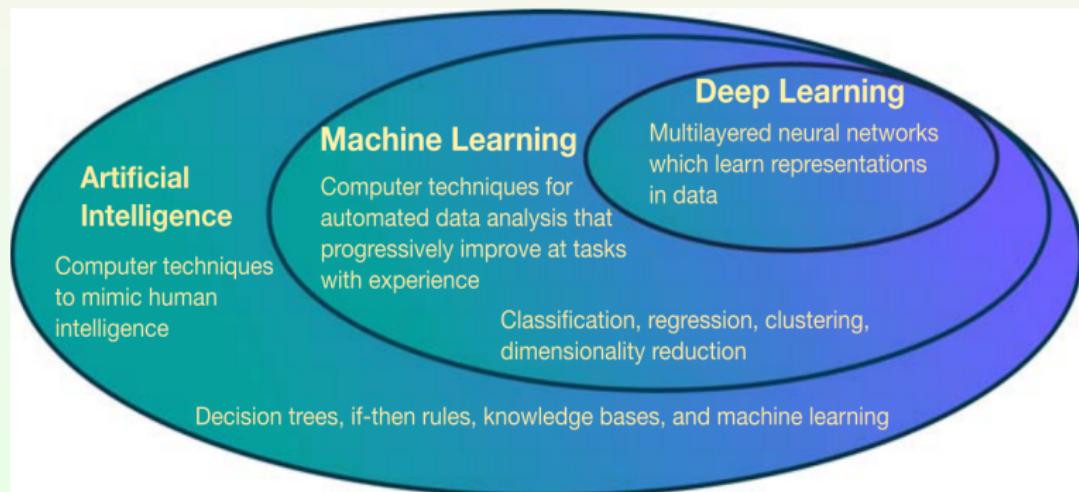


Fig.: Artificial Intelligence, Machine Learning and Deep Learning.

# 机器学习问题的一般形式与分类

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

## ■ 机器学习类问题的一般表示:

对于给定的集合  $X$ , 可以预测或近似得到未知函数  $y = f(X)$

集合  $X$  构成特征空间, 集合中的每个元素  $x$  称为特征向量 (在材料类的机器学习中也称描述符)

根据机器学习得到的近似函数  $\hat{y} = \hat{f}(X)$ , 模型有能力预测训练数据之外的输出值

机器学习的这种预测能力也称为模型的“泛化” (generalization)

## ■ 机器学习主要根据学习的特征分为无监督学习 (unsupervised learning) 和监督学习 (supervised learning)

## ■ 此外的机器学习问题还包括:

- 半监督学习, 即大部分没有映射关系的数据和少量有映射关系的数据;
- 多任务和迁移学习, 即将从相关问题习得的知识应用到数据极少的对象, 提升模型的学习能力
- 强化学习, 即没有输入输出, 但会和环境不断交互, 通过最大化环境的反馈, 最终达到学习目标

# 机器学习问题的一般形式与分类

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

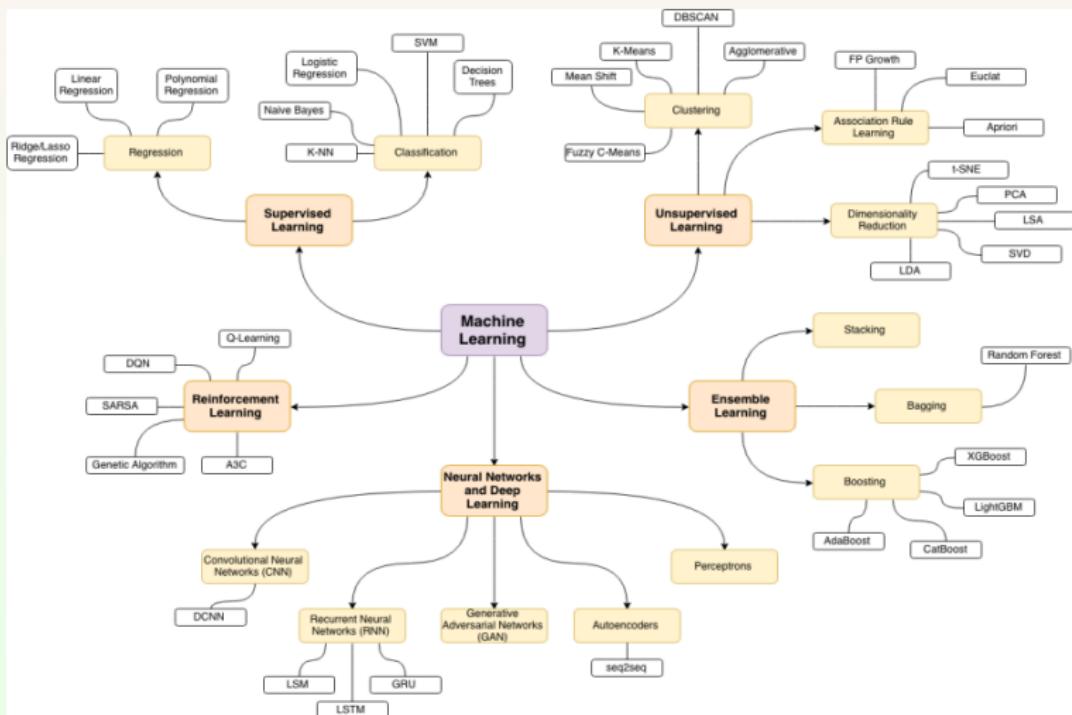
高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究



# 数据挖掘的基本决策

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

- 聚类: 无指导的学习
- 分类: 有指导的学习

分类是数据挖掘的重要任务

分类的**目的**: 机器学会分类函数或分类模型 (称为分类器), 通过模型能将数据库中的数据映射到特定类别中的某一类

# 无监督学习

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

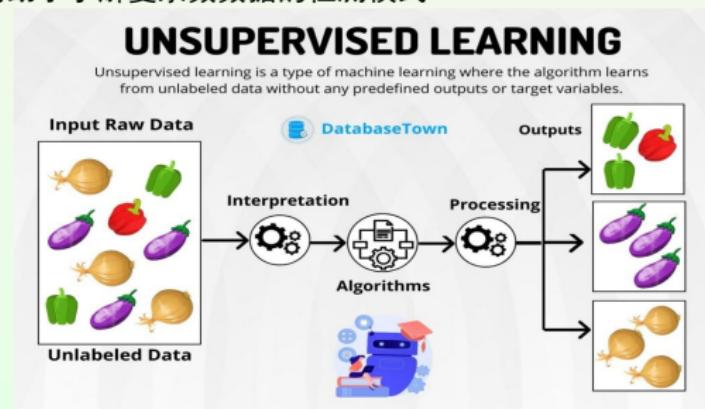
机器学习算法

数据挖掘与第一原理材料研究

无监督学习是描述性质的，所有数据只有特征向量，没有标签，但呈现出聚群的结构，相似类型或特征的数据会聚集在一起

- 如果没有标签的数据的组合是有限个，称为聚类 (clustering); 无限的称为密度估计 (density estimation)
- 将高维数据投影到低维空间，称为降维 (dimensionality reduction)

降维有助于了解复杂数据的检测模式



# 监督学习

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

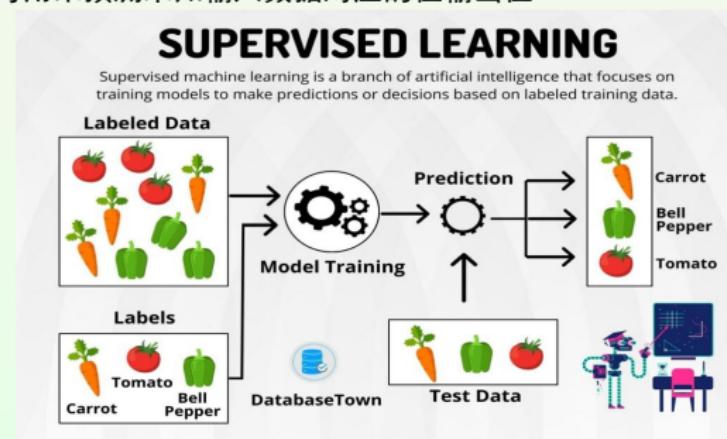
机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

监督学习是通过学习指定数量的输入输出间的函数映射

- 如果输出函数  $y_i$  表示类别的有限集合，称为分类 (classification) 问题  
模型可用来预测未知数据所属类型
- 如果输出函数  $y$  是实数，称为回归 (regression) 问题  
模型可用来预测未知输入数据对应的值输出值



# 机器学习的流程

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

- **数据的收集和筛选**: 从现有数据中产生并选择与问题解决有用和相关的数据子集
- **数据预处理**: 清洗缺失和不完整数据，将数据将转换成统一的格式 (如整型、字符串型等)
- **数据训练**: 将数据分为训练集、验证集和测试集三部分，训练集数据用于学习并得到模拟参数 (主要针对监督学习)
- **模型测试和优化**: 用验证集数据评估模型的效果和性能，并用验证集数据优化模型
  - 一旦完成优化，用测试集数据评定模型的性能
- **模型应用**: 将得到的有效模型对未知数据进行预测，如果有新的数据，模型还可以继续训练

# 主成分分析 (principal component analysis)



高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

## 主成分分析:

将高维数据投影到数据点集中的区域，并使数据在新轴向周围聚集度最高

- 确定主成分: 找到  $X^T X$  的最大本征值对应的本征矢量 (称为主成分)  
 $X$  是已知高维数据集构成的矩阵
- 数据投影: 计算所有数据点对主成分的投影, 实现数据压缩
- 局限: 假设数据位于线性子空间中  
只能为线性数据提供最佳结果



# k-means 算法

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

k-means 算法: 直接计算数据点和聚类中心的欧氏距离 (Euclidean distance), 将  $n$  个数据分类为  $k$  个子集 ( $k < n$ )

随机选定聚类中心的个数 ( $k$ ) 和位置 ( $\mu_0^{(j)}, 1 \leq j \leq k$ ), 执行迭代:

- 计算每个数据点与聚类中心的距离, 标记为  $y_t^{(i)}$  ( $t > 0$ ), 将该点分到距离最小的聚类中心所属的类中

$$y_t^{(i)} = \operatorname{argmin}_j \| \mathbf{x}^{(i)} - \mu_t^{(j)} \|_p$$

- 重新计算每个聚类的中心 ( $\{\mu_t^{(j)}\}$ )

$$\mu_{t+1}^{(j)} = \frac{1}{n_j} \sum_{i=1}^{n_j} \mathbf{x}^{(i)} \delta_{y_t^{(i)}, j}$$

$p \in \mathbb{N}$  表示空间的维度 (当  $p = 2$  即为是二维平面的欧氏距离),  $n_j$  是归入聚类中心  $\mu_t^{(j)}$  的分类元素数目,  $\delta_{n,m}$  表示  $\Delta$  函数,  $t$  表示迭代次数

当标记不再变化时, 迭代收敛

# $k$ -means 算法

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

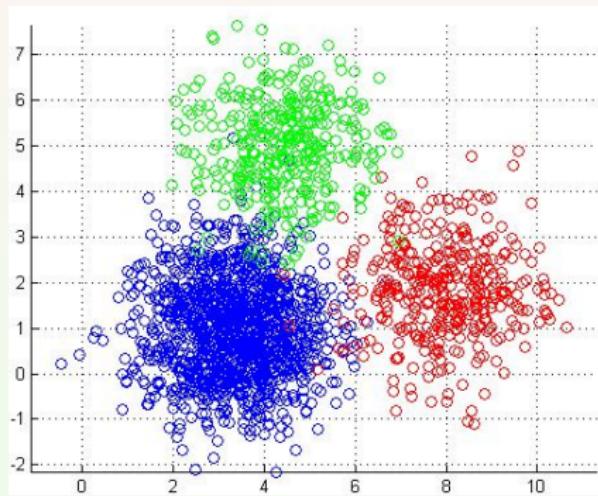


Fig.: The Schematic diagram of the  $k$ -means algorithm.

$k$ -means 聚类算法的结果与初始聚类中心位置的选择密切关联

初始聚类中心选择得不同，结果差别会比较大<sup>3</sup>

<sup>3</sup>一般克服的策略是通过多次初始聚类中心并执行该算法，选择最有代表性的聚类形式作为结果

# 层次聚类 (Hierarchical Clustering) 算法

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

## 层次聚类: 通过一层一层的进行聚类

**分裂法:** 由上向下把大的类别分割

**聚合法:** 由下向上对小的类别聚合

- 初始时将每个训练数据点  $x^{(i)}$  作为一个类 (或类簇), 原始类大小等于训练数据点数目  $n$
- 衡量任意两个类 (分别标记为 A 和 B) 的偏差  $d(A, B)$
- 偏差最小的两个类 (最相似) 合并一个新的类簇

反复执行该聚合过程, 最终可以用一个类簇能囊括全部训练集



Fig.: The Schematic diagram of Hierarchical-Clustering algorithm.

# 层次聚类 (Hierarchical Clustering) 算法

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

要将  $n$  个类聚合成  $k$  个类簇 ( $1 < k < n$ )，要在聚合中对聚合偏差设置截断，常见的聚合截断有三种：

- 最小偏差

$$d_{SL}(A, B) = \min_{i \in A, j \in B} d_{ij}$$

$d_{ij}$  表示任意一对类的偏差

- 最大偏差

$$d_{CL}(A, B) = \max_{i \in A, j \in B} d_{ij}$$

- 平均偏差

$$d_{GA}(A, B) = \frac{1}{|A||B|} \sum_{i \in A} \sum_{j \in B} d_{ij}$$

定义的  $d_{ij}$  可以为指定，对数值形式的数据，最常见的设置为欧氏距离

分裂法与聚合法类似，只是初始时训练数据集属于一个类簇，然后逐层分裂形成特定的类簇

# 线性回归 (Linear Regression) 算法

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

对监督学习来说，每个数据不仅有特征向量  $x_i$ ，还有相应的值  $y_i$  预测连续变化的  $y$  值，最常用的算法是线性回归

回归算法的基本思想：对于满足正态分布的数据点

- 允许的参数拟合预测表达式

$$\hat{y}^{(i)} = \theta^T x^{(i)}$$

上标  $T$  表示矢量的转置， $\hat{y}^{(i)}$  是预测值， $\theta$  是参数的矢量

- 为了求得  $\theta$  参数，定义误差的最小二乘函数

$$J(\theta) = \sum_{i=1}^n L[\hat{y}^{(i)}(x^{(i)}, \theta), y^{(i)}] = \frac{1}{2} \sum_{i=1}^n (\theta^T x^{(i)} - y^{(i)})^2$$

最小化该函数，可以得到最优化的参数  $\theta$ ，由此可以得到线性回归机器学习的模型

- 最优化参数  $\theta$  用矩阵表示可写作：

$$\theta = (X^T X)^{-1} X^T y$$

这里  $X$  矩阵的每一列是由训练集输入数据  $x^{(i)}$ ， $y$  是对应的输出标注构成的矢量

# 预测模型的检验

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

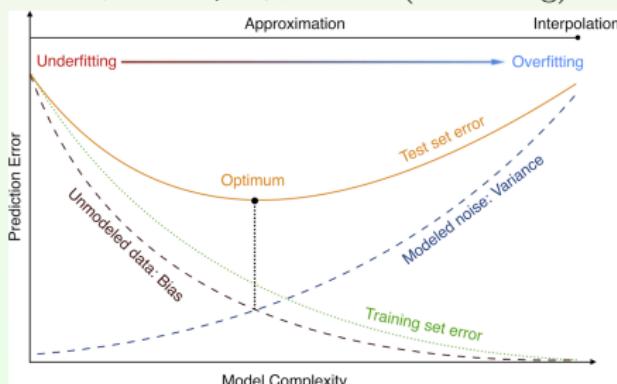
机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

机器学习的模型性能，可用测试数据集检验，预测误差与训练数据集包含的数据数量密切相关：

- 训练集数据不够，模型不能完全反映训练集特征  
预测结果将会表现出明显的偏差
- 训练集数据过多，模型能够体现训练集特征  
对训练集外的数据效果不好，出现过拟合 (overfitting)



**Fig.:** Complexity of the model: Effective number of degrees of freedom (mainly tuned by the hyperparameters of the estimator).

# 预测模型的检验

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

回归算法的发展: 构建模型时, 通过引入标准化参数  $\lambda$  来反应训练集元素变化的影响

$$J(\theta) = \frac{1}{2} \sum_{i=1}^n (\theta^T \mathbf{x}^{(i)} - y^{(i)})^2 + \lambda \|\theta\|_p$$

这里  $p$  表示数据度量形式

优化参数  $\lambda$  降低训练集数据数量的影响:

通过压制或筛选训练数据集中的特征来调节其对误差函数的贡献

- $p = 1$ : Ridge Regression
- $p = 2$ : LASSO Regression

注意:  $\lambda$  不能像  $\theta$  一样被优化

一般是通过比较几个不同的  $\lambda$  值, 选其中能最大化预测能力又不会引入太大偏差的一个

# 分类算法: 逻辑回归 (logistic regression)

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

逻辑回归，可以类比为映射到区间  $[0, 1]$  之间的线性回归：

对于给定数据点  $x^{(i)}$ ,

- $y^{(i)} = 1$ : 将其分入特定的“是 (True)”类
- $y^{(i)} = 0$ : 将其分入特定的“否 (False)”类

预测函数可以表示为

$$\hat{y} = \sigma(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

这里  $\theta$  仍是参数矢量,  $\sigma$  是逻辑函数 (或 sigmoid 函数)

预测函数也可以视为条件概率:  $\hat{y} = P(y = 1 | x, \theta)$

分类问题的误差函数可定义为负的 log-型函数 (交叉熵 cross-entropy), 同样通过参数  $\theta$  最小化该函数:

$$J(\theta) = -\frac{1}{n} \sum_{i=1}^n [y^{(i)} \log(\hat{y}^{(i)}) + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)})]$$

这里  $y^{(i)}$  和  $\hat{y}^{(i)} = \sigma(\theta^T x^{(i)})$  分别为实际和预测值

和线性回归类似, 也可以引入正则参数  $\lambda$

逻辑分类也可用于处理超过两类的数据分类:

训练有  $n$  个逻辑回归值的模型, 每一类对应一个数值, 每个数据分入概率值最高的类中

# 支持向量机 (Support Vector Machines)

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

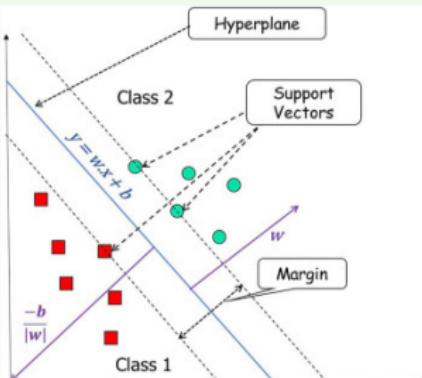
## 支持向量机: 最通用的分类算法

### 定义函数

$$J(\theta) = C \sum_{i=1}^n [y^{(i)} \max(\theta^T \mathbf{x}^{(i)}, 0) + (1 - y^{(i)}) \max(-\theta^T \mathbf{x}^{(i)}, 0)] + \frac{1}{n} \sum_{i=1}^n \theta_i^2$$

这里  $C$  是参数

- 在约束条件  $y^{(i)}(\theta^T \mathbf{x}^{(i)} + b) \leq 1$  下, 对全部训练数据点  $(\mathbf{x}^{(i)}, y^{(i)})$  实现  $\|\theta\|^2$  最小化
- 引入 Lagrangian 乘子最小化  $\|\theta\|^2$ , 可根据测试数据  $i$  的值  $y^{(i)}$  (+1 或 -1), 确定数据所属的类



# 支持向量机 (Support Vector Machines)

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

SVM 最重要的特色是内核技巧 (kernel trick)

将参数矢量  $\theta$  用训练集数据表示

$$\theta = \sum_i \alpha_i y^{(i)} \mathbf{x}^{(i)}$$

因此可将分类规则写成数据点积的形式

$$\theta^T \mathbf{x} + b = \sum_i \alpha_i y^{(i)} \mathbf{x}^{(i)} \cdot \mathbf{x} + b \leq 0 \rightarrow y = +1$$

这里  $b$  和  $\{\alpha_i\}$  是待学习的参数

内核技巧利用映射关系  $\phi(\mathbf{x}) : \mathbb{R}^n \rightarrow \mathbb{R}^m$  将矢量变换成  $\mathbf{x}^{(i)} \cdot \mathbf{x}$  点积

实际上是将数据点映射到更高维度的空间中

所有矢量点积映射的变换都可以类似处理，最常用的内核

- 多项式内核:

$$K(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \phi(\mathbf{x}^{(i)}) \cdot \phi(\mathbf{x}^{(j)}) = (\mathbf{x}^{(i)} \cdot \mathbf{x}^{(j)} + 1)^d, d \in \mathbb{N}$$

- Gaussian 内核 (radial basis function, RBF 内核)

$$K(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = e^{-\frac{\|\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\|^2}{2\sigma^2}}$$

# Naïve Bayes 分类

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

## 分类算法的判别模式

- 基于判别模式 (discriminative model): 数据点预测的标签是条件概率  $p(y|x)$
- 基于生成模式 (generative model), 预测点条件概率用后验 Bayes 公式表示

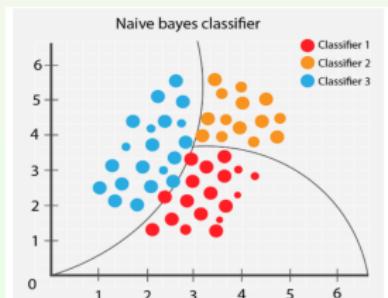
$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} = \frac{p(x|y)p(y)}{\sum_i p(x|y=i)p(y=i)}$$

这里  $p(y)$  表示先验概率, 即没有附加任何先期知识和分析得到的概率

假设数据点的特征向量  $x^{(i)}$  和值  $y^{(i)}$  完全独立, 可有 Naïve Bayes 分类算法 (后验概率):

$$p(y|x) = \frac{\prod_{j=1}^n p(x_j|y)p(y)}{p(x)}$$

这里  $x_j$  表示特征向量  $x$  的元素



- 分类前先列出训练数据点的全部先验概率  $p(y)$  和条件概率  $p(x_j|y)$
- 用 Naïve Bayes 算法计算后验概率
- 选择所有  $y$  中最大的后验概率  $p(y|x)$  作为分类预测值

# $k$ -最近邻 ( $k$ -nearest neighbors) 算法

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

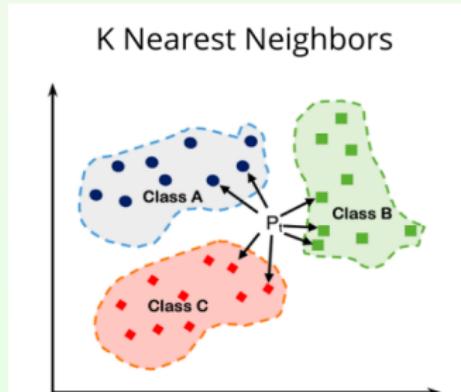
**$k$ -最近邻算法:** 利用数据点空间距离的类似性, 不再训练, 适合处理快速任务  
 $d$ -维空间有训练集数据  $\{\mathbf{x}^{(i)}\}$ , 计算未知数据点与已知数据点之间的空间距离

$$d(\mathbf{x}, \mathbf{x}^{(i)}) = \|\mathbf{x} - \mathbf{x}^{(i)}\|_p$$

这里的  $p$  是维度参数

一旦得到  $\mathbf{x}$  到空间各点的距离,  $\mathbf{x}$  点归入与其有最近邻  $k$  值最大的类中  
如果没有最大类, 则随机归入最近邻点的最常使用的标注类中

- 对连续值求平均, 就是基于  $k$ -NN 的回归
- $k$  值的选取对于分类很敏感, 不同的  $k$  值很可能得到完全不同的数据分类



# 决策树 (Decision Trees)

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

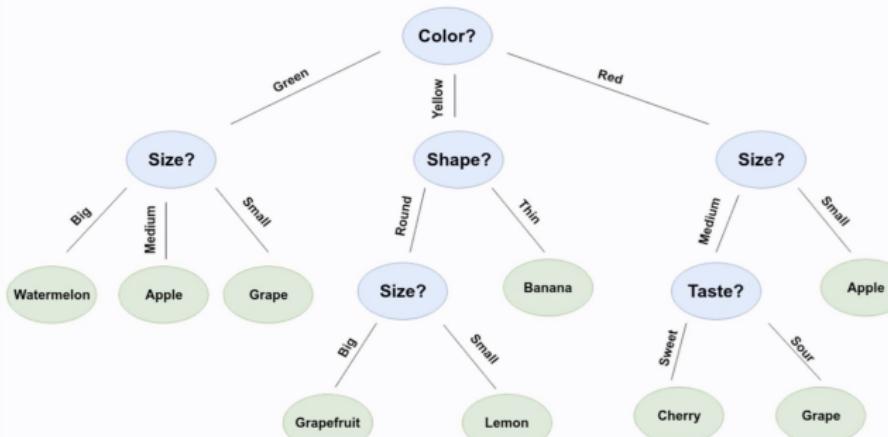
数据挖掘与第一原理材料研究

决策树：通过创建节点来实现对某种分裂算法的优化  
同时适合分类和回归

- 决策树上每个节点都含一个定义该部分空间划分的方案，直到空间不可再分
- 每个不连通的子空间称为叶节点，叶节点包含了待分类或预测的数据点

决策树的主要问题：一旦开始训练，往往伴随过拟合

## Decision Trees



# 随机森林 (Random Forest)

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

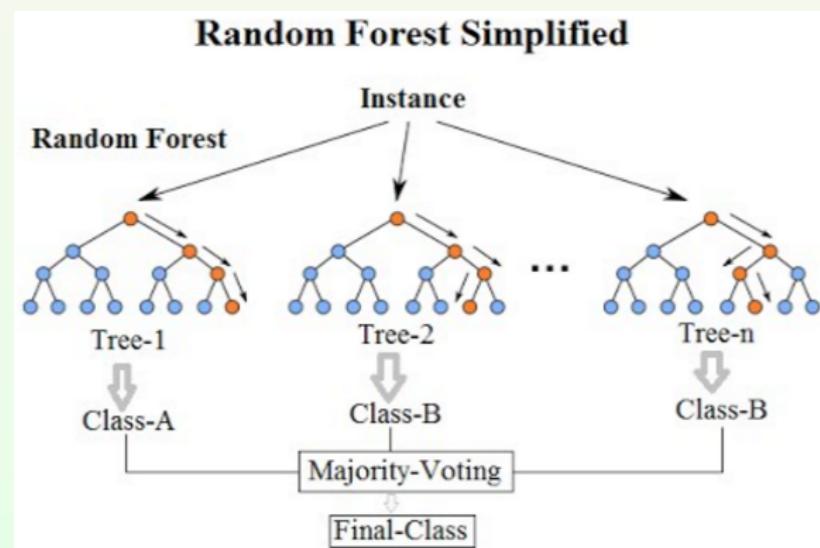
数据挖掘与第一原理材料研究

克服决策树过拟合的策略:

- 修剪决策树分叉，会损失一定的精度，但提高回归的泛化能力
- 采用随机森林，即训练大量的决策树然后再取统计平均值

随机森林方法是一种系统平均方法: **决策树将成为训练对象**

对决策树的特征进行随机抽样训练时，一般采取自展抽样 (bootstrap sample) 方案



# 深度神经网络基础

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

- 感知机 (Perceptron Learning Algorithm, PLA):

最早的监督式训练算法，是神经网络构建的基础

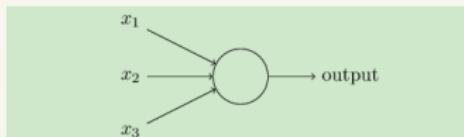


Fig.: Perceptron Learning Algorithm.

输出与输入之间将学习到一个线性关系，可有中间输出结果

$$z = \sum_{i=1}^m w_i x_i + b$$

中间结果连接一个神经元激活函数

$$\text{sign}(z) = \begin{cases} -1 & z < 0 \\ 1 & z \geq 0 \end{cases}$$

# 深度神经网络基础

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究



- 每一个输入数据都可以表示为一个向量  $x = (x_1, x_2)$
- 函数则是要实现“如果线以下，输出 0；线以上，输出 1”

# 深度神经网络基础

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

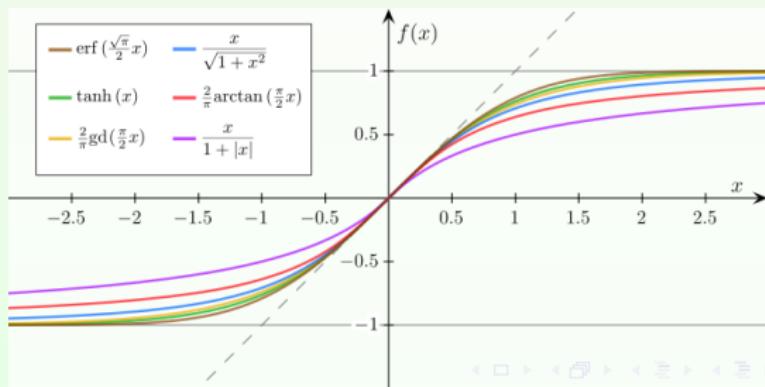
数据挖掘与第一原理材料研究

感知机模型只能用于二元分类，无法学习较为复杂的非线性模型，神经网络在感知机模型基础上作了扩展

- 加入隐藏层 (hide layer):  
隐藏层可以有很多层，增强模型的表达能力
- 输出层的神经元可以有不止一个输出
- 对激活函数作扩展，如 Sigmoid 函数

$$f(z) = \frac{1}{1 + e^{-z}}$$

其他的激活函数还有 tanh、softmax 和 ReLU 等



# 深度神经网络 (Deep Learning Neural Network)



高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

深度神经网络可以包含上百层神经元，通常有上万个参数，再加上超参数，实际的参数空间几乎是无限大的

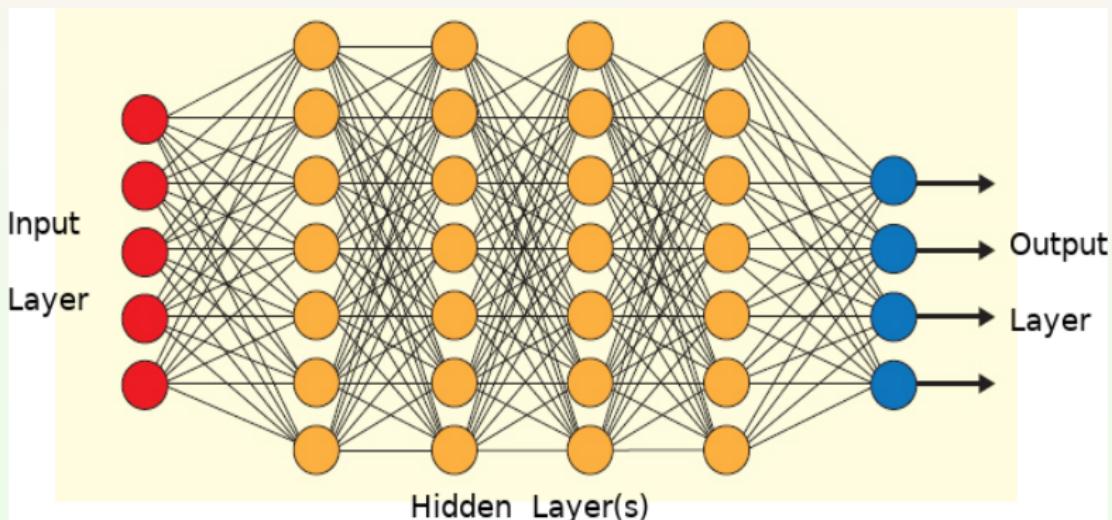


Fig.: Deep Learning Neural Network.

如何从海量潜在的可能参数中做选择极具挑战性

# 深度神经网络的前馈算法

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

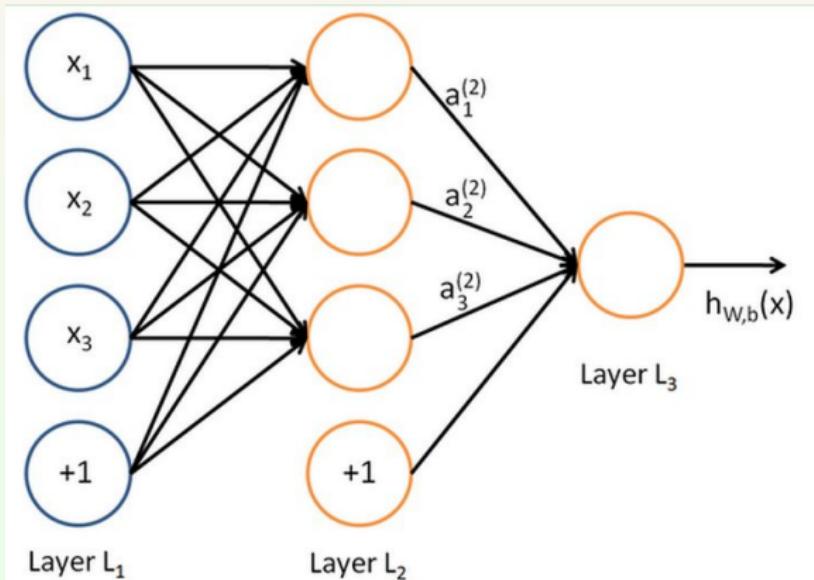
高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究



# 遗传算法 (Genetic Algorithm)

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

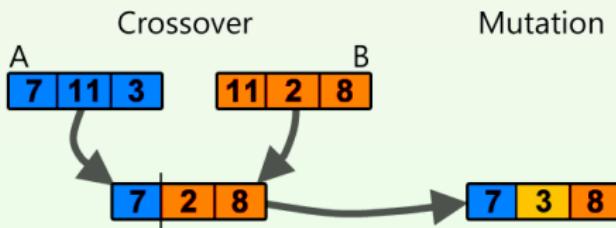
数据挖掘与第一原理材料研究

遗传算法: 模拟生物进化论的自然选择和遗传学机理的计算模型  
通过模拟自然进化过程搜索最优解

- 1 优化问题可能潜在的解集构成一个种群 (population)

该种群由经过基因 (gene) 编码的一定数目的个体 (individual) 组成, 每个个体是染色体 (chromosome) 带有特征的实体

- 2 种群产生之后, 借助于自然遗传学的遗传算子 (genetic operators) 进行组合交叉 (crossover) 和变异 (mutation), 产生新一代个体



- 3 适者生存和优胜劣汰: 在每一代, 根据问题计算个体的适应度 (fitness)  
4 选择 (selection) 合适的个体, 构成代表新的解集的种群  
逐代 (generation) 演化后, 产生出越来越好的近似解 (优化目标)

# 深度神经网络与遗传算法

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

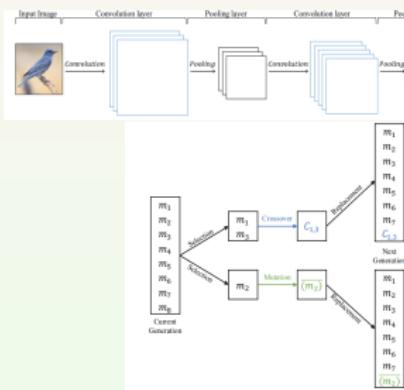
第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

## 深度神经网络类似问题的解决方案：遗传算法



### GACNN: TRAINING DEEP CONVOLUTIONAL NEURAL NETWORKS WITH GENETIC ALGORITHM

**Parsa Esfahanian**  
 Department of Computer Science  
 Institute for Research in Fundamental Sciences  
[parsa.esfahanian@ipm.ir](mailto:parsa.esfahanian@ipm.ir)

**Mohammad Akhavan**  
 Department of Computer Science  
 Institute for Research in Fundamental Sciences  
[mohammad.akhavan@ipm.ir](mailto:mohammad.akhavan@ipm.ir)

### AlphaStar: An Evolutionary Computation Perspective

Kai Arulkumaran  
 Imperial College London  
 London, United Kingdom  
[kai79@ic.ac.uk](mailto:kai79@ic.ac.uk)

Antoine Cully  
 Imperial College London  
 London, United Kingdom  
[a.cully@imperial.ac.uk](mailto:a.cully@imperial.ac.uk)

Julian Togelius  
 New York University  
 New York City, NY, United States  
[julian@toga.cs.nyu.edu](mailto:julian@toga.cs.nyu.edu)

#### ABSTRACT

In January 2019, DeepMind revealed AlphaStar to the world—the first artificial intelligence (AI) system to beat a professional player at the game of StarCraft II—representing a milestone in the progress of AI. AlphaStar draws on many areas of AI research, including

the original game, and its sequel SC II, have several properties that make it considerably more challenging than even Go—real-time play, partial observability, no single dominant strategy, complex rules that make it hard to build a fast forward model, and a particularly large and varied action space.

# 机器学习的验证

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

机器学习模型是否仅对训练数据有效，还需要一些数据集来检验<sup>4</sup>，这一过程称为验证，对应的数据集称为验证集

- 监督学习的数据集一般分为三类：训练集、验证集和测试集

这些样本数据最好是具备相同的统计分布特征。为了优化学习模型，建议先对样本数据多次学习，最后才将模型应用到测试数据集上

- 对比预测数据和实际数据的偏差，可以评估模型的真实预测能力
- 当数据非常有限时，最常用的是 *k*-重交叉验证 (cross-validation)：

将训练集分成  $k$  个子集，选择  $k - 1$  个子集训练模型，并用剩下的一个未训练的子集验证模型：  
将训练-验证过程执行  $K$  次，并将  $K$  次验证的平均损失函数来评估模型性能的平均表现

平均损失函数定义为：

$$E_{cv}^K = \frac{1}{K} \sum_{k=1}^K \sum_{i=1}^{n_k} L(\hat{y}_k^{(i)}, y^{(i)})$$

这里  $L$  是验证数据集的损失函数， $\hat{y}_k^{(i)}$  是训练子集 (不含验证子集  $k$ ) 得到的模型，对采样数据  $i$  预测的值  $\hat{y}_k^{(i)}$ ，训练子集的采样数据共有  $n_k$

特别地， $K = n$ ，即训练集划分的子集与其元素个数相同，称为 leave-one-out cross-validation

交叉验证也可用于评估训练模型中的超参数<sup>5</sup>：选择有最小的预测误差时的参数

<sup>4</sup> 模型对训练集的学习，优化的并非全部参数，以神经网络为例，网络隐藏层的数目作为参数在优化过程中就始终保持不变，这类参数称为超参数 (hyperparameter)。通过验证集检验的主要是模型中未能优化的超参数的性能

<sup>5</sup> 超参数包括正则化参数  $\lambda$ 、SVM 的 Gaussian 内核参数  $\sigma$ ，二叉树 (二元决策) 修剪层次和随机森林系综的特征向量等

# 机器学习的验证

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

对机器学习模型性能的评估手段有很多

- 二元或多元分类可以选用混同矩阵 (confusion matrices)<sup>6</sup>

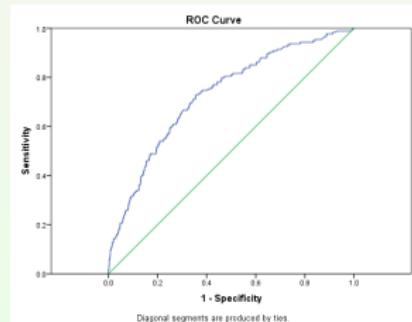
比如矩阵的列方向为采用数据的值，行方向是预测数据的值 (除了近似单位矩阵表示，也可以用真阳 (TP)、真阴 (TN) 和假阳 (FP) 假阴 (FN) 来填充矩阵)

- 绘制模型受试操作特征 (receiver operating characteristic, ROC) 曲线

也可以用比如“真阳率” (TP rate)  $TPR = \frac{TP}{TP+FN}$  和“假阳率” (FP rate)  $FPR = \frac{FP}{FP+TN}$

	Predicted A	Predicted B	Predicted C
A	0.8	0.1	0.1
B	0.08	0.9	0.02
C	0.3	0.1	0.6

1  
0



不同阈值下的 ROC 曲线反应了模型的预测能力

<sup>6</sup> 所谓混同矩阵，就是评估模型预测与抽样吻合程度建立的矩阵，预测吻合度高的元素主要出现在矩阵对角元，预测吻合度低的都在矩阵非对角元

# 机器学习的验证

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

回归预测也有很多方案可以评估模型对数据的拟合精度，如：

- 平均绝对误差 (mean absolute error, MAE)

$$\text{MAE} = \frac{1}{n} \sum_i^n |y_i - \hat{y}_i|$$

- 归一化相对百分误差 (normalized mean absolute in percentage, MAPE)

$$\text{MAPE} = \frac{100\%}{n} \sum_i^n \frac{|y_i - \hat{y}_i|}{y_i}$$

- 均方误差 (mean squared error, MSE)

$$\text{MSE} = \frac{1}{n} \sum_i^n (y_i - \hat{y}_i)^2$$

从使用频率角度说，由分布参数  $\theta$  估计  $\hat{\theta}_m$  与 MSE 密切关联，即

$$\text{MSE} = \text{E}[(\hat{\theta}_m - \theta)^2] = \text{Bias}(\hat{\theta}_m)^2 + \text{Var}(\theta)$$

一般更常用的是误差的均方根 (root of MSE, RMSE)

还有用可决统计系数 (也称决定系数, coefficient of determination)  $R^2$

可决统计系数的定义为  $R^2 = 1 - \frac{\text{SS}_{\text{res}}}{\text{SS}_{\text{tot}}}$

这里  $\text{SS}_{\text{res}} = \sum_i (y_i - \bar{y})^2$  是总的方差求和，而  $\text{SS}_{\text{res}} = \sum_i (y_i - \hat{y}_i)^2$  是预测模型误差平方求和

## 材料信息学: 应用数据挖掘特别是机器学习技术推动材料科学的研究

- 材料信息学主要研究材料的内禀特征 (intrinsic features), 包括结构、组成、对称性等与材料的性质间的内在数据关联
- 材料学研究的数据挖掘主要是监督学习, 即材料物性的预测和材料的分类
  - 物性预测: 应用包括回归在内的学习算法, 建立材料物性的描述函数  $f(\mathbf{x})$
  - 分类问题: 根据特定的物性目标, 将符合要求的材料归入其中

比如按磁性和非磁性划分, 按照晶体所属结构分类属于两种不同分类, 每种分类方式内部各部分之间不存在交集

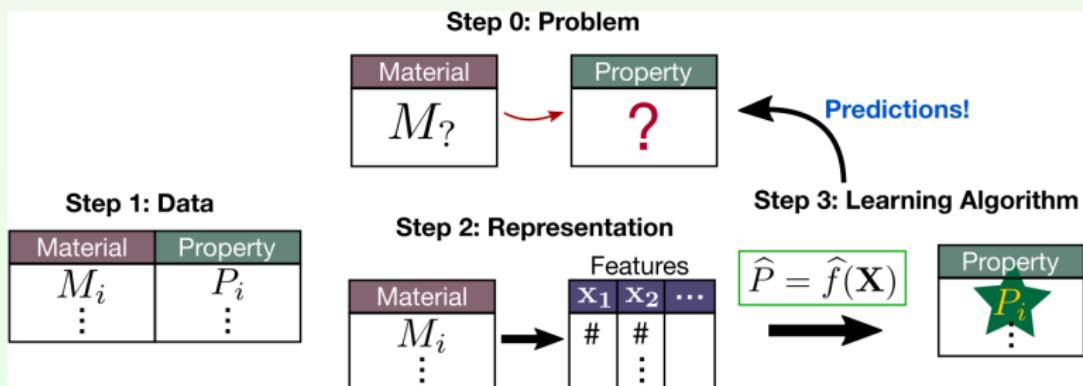


Fig.: General Process of Data Mining in Materials Science

# 数据驱动的材料研究主要流程

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

在数据挖掘驱动的材料计算的主要研究流程为：

- 1 问题的确定**: 根据问题类型选择机器学习算法
- 2 数据的组织**: 数据应该涵盖研究样本的全部特征 (输入) 和目标物性 (输出)  
数据是机器学习的基本对象，可以来自理论计算，也可来自实验测量
- 3 物性的表示**: 描述符决定机器学习的性能  
描述材料物性的特征向量称为描述符
- 4 算法和模型的选定、评估与优化**: 主要针对超参数的选择
  - 考虑模型的复杂度/合理性
  - 算法的精度-效率/性能和训练时长平衡
  - 既要防止数据不足也要防止过拟合

机器学习的建模可以简单概述为

机器学习模型=研究对象+数据+表示+学习算法+优化

# 机器学习预测材料性质

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

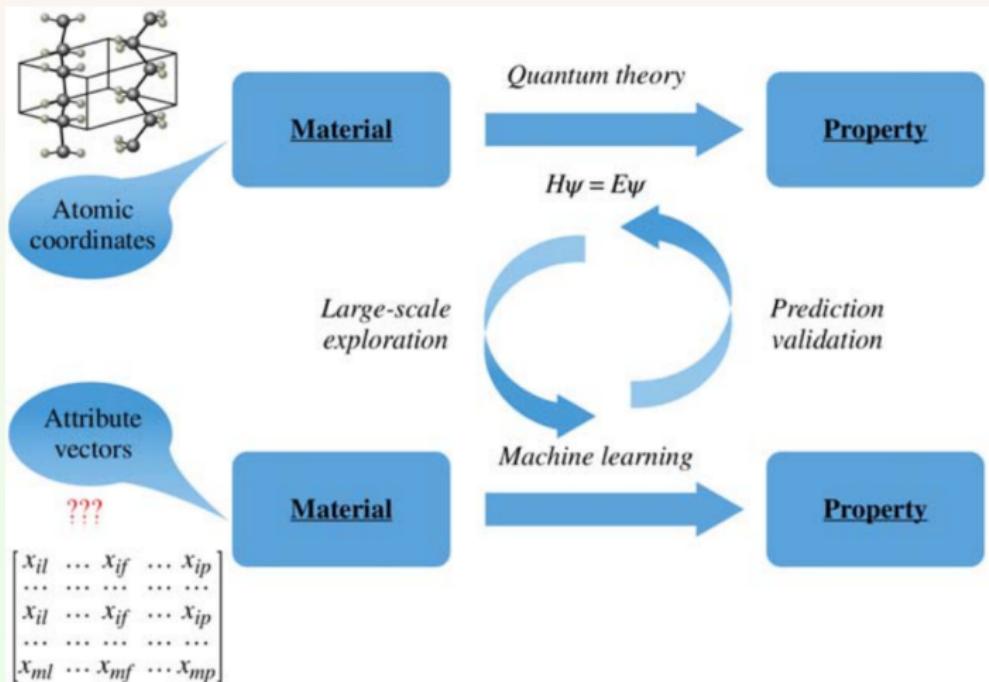
高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究



**Fig.:** Making machine-learning prediction replacing the expensive quantum chemistry calculation.

# 机器学习对 DFT 的促进

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

机器学习在第一原理计算领域重要的作用：节省或替代获取 DFT 计算结果或数据所必需的成本

- 对复杂电子体系，Schrödinger 方程直接的迭代求解对计算资源和时间成本都比较高
- DFT 框架下，机器学习加速 Kohn-Sham 方程求解：避免直接求解 Kohn-Sham 方程，直接预测电子密度

通过机器学习获得材料的电子密度-势的映射关系，得到能量泛函后，对能量泛函求变分极小得到基态能量
- 优化 DFT 计算的能量泛函，可将机器学习方法很方便地与传统 DFT 计算结合起来使用

机器学习优化的泛函并不限于传统 DFT 的 Kohn-Sham 方程的交换-相关部分，也可以用于无轨道类型 (free-orbital) 的能量密度泛函
- 机器学习还可用于解决量子多体问题：得到紧束缚模型的类 Schrödinger 方程的 Hamiltonian
- 机器学习应用于计算材料研究，特别是在大于电子尺度的材料计算，还包括计算配分函数、寻找相变和序参量以及获得模型的 Green's function 等

# 机器学习对 DFT 的促进

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

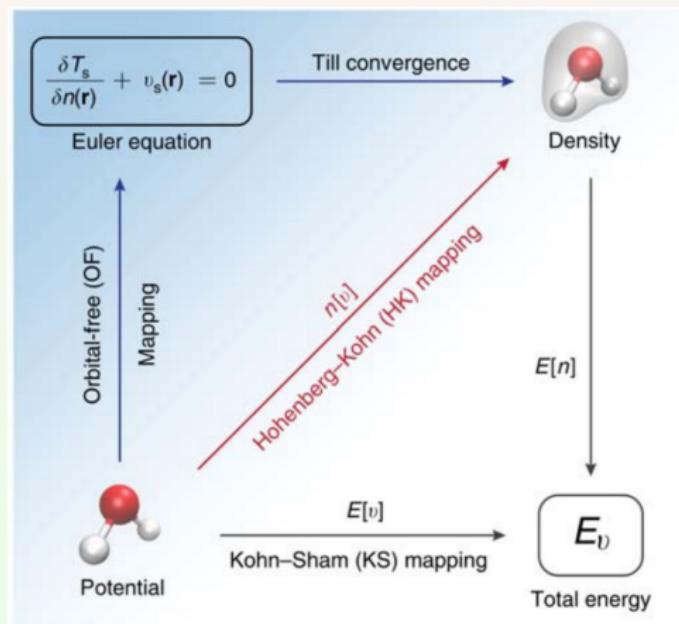
高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究



机器学习这些领域的成功应用：展示了数据挖掘技术在拓展材料科学的研究前沿有着广阔的应用前景，可应用于多种尺度下材料研究的各类系统和现象

# 主要参考文献

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

- [1] X. Yang, Z. Wang, X. Zhao and H. Liu *Comp. Mater. Sci.*, **146** (2018), 319
- [2] <http://matcloud.cnic.cn>
- [3] S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R. H. Taylor, L. J. Nelson, G. L. Hart, S. Sanvito, M. Buongiorno-Nardelli, N. Mingo and O. Levy *Comp. Mater. Sci.*, **58** (2012), 227
- [4] S. P. Ong, S. Cholia, A. Jain, M. Brafman, D. Gunter, G. Ceder and K. A. Persson. *Comp. Mater. Sci.*, **97** (2015), 209
- [5] <http://www.qmip.org/qmip.org/Welcome.html>
- [6] J. Hachmann, R. Olivares-Amaya, S. Atahan-Evrénk, C. Amador-Bedolla, R. S. Sánchez-Carrera, A. Gold-Parker, L. Vogt, A. M. Brockway and A. Aspuru-Guzik *J. Phys. Chem. Lett.*, **2** (2011), 2241
- [7] L. Lin *Mater. Perform. Character.*, **4** (2015), 148
- [8] R. Ramprasad and R. Batra and G. Pilania and A. Mannodi-Kanakkithodi and C. Kim. *npj. Comput. Mater.*, **3** (2017), 54
- [9] K. T. Butler and D. W. Davies and H. Cartwright and O. Isayev and A. Walsh. *Nature*, **559** (2018), 547
- [10] L. Ward and C. Wolverton *Curr. Opin. Solid State Mater. Sci.*, **21** (2016), 167

# 主要参考文献 (cont.)

- [11] Y. Liu and T. Zhao and W. Ju and S. Shi *J. Materomics.*, **3** (2017), 519

高通量计算流程、数据库和机器学习简介

高通量与高性能计算

高通量计算材料自动流程

第一原理数据库

机器学习简介

机器学习算法

数据挖掘与第一原理材料研究

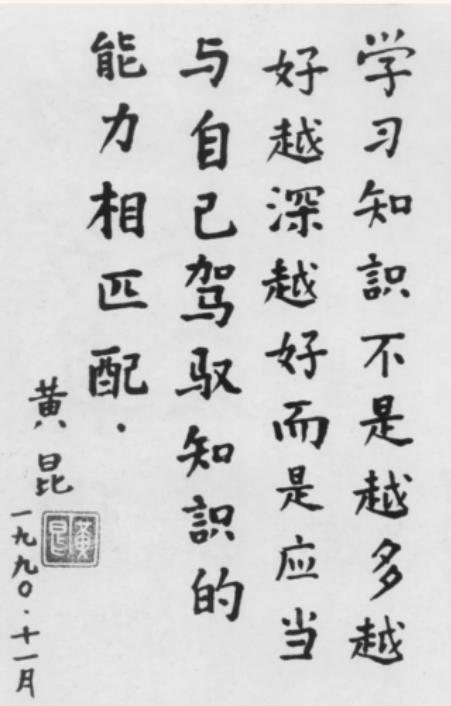


Fig.: 黄 昆 教授的治学箴言

谢谢大家！