



TAIJI LABORATORY
FOR GRAVITATIONAL WAVE UNIVERSE



ICTP-AP
International Centre
for Theoretical Physics Asia-Pacific
国际理论物理中心-亚太地区



中国科学院大学
University of Chinese Academy of Sciences

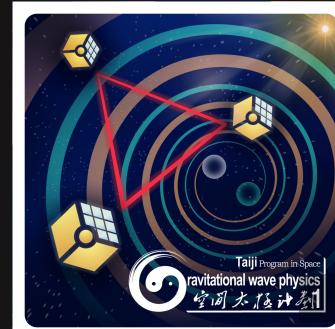
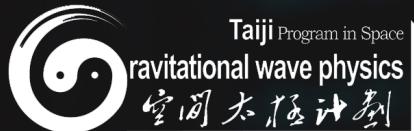
引力波数据探索：编程与分析实战训练营

第3部分 机器学习基础 机器学习算法之应用起步

主讲老师：王赫

ICTP-AP, UCAS

2023/12/20





引力波数据分析与机器学习

- 人工智能 > 机器学习 > 深度学习
- 机器学习的定义，目标和过程
- 机器学习的常见类型：监督学习，非监督学习和其他类型
- 机器学习模型的分类
- 机器学习项目开发规划与准备
- 机器学习项目：开发应用程序的步骤
- scikit-learn 机器学习库：分类+回归
- 机器学习中的特征工程（下一讲）
- 机器学习中的模型调优与模型融合（下一讲）
- 实战项目：对 LIGO 的 Glitch 数据实现聚类分析（下一讲）



引力波天文学：引力波数据分析

- 基础理论的检验与修正

- 基础物理学
 - 引力子是否有质量, 引力波的传播速度 ...

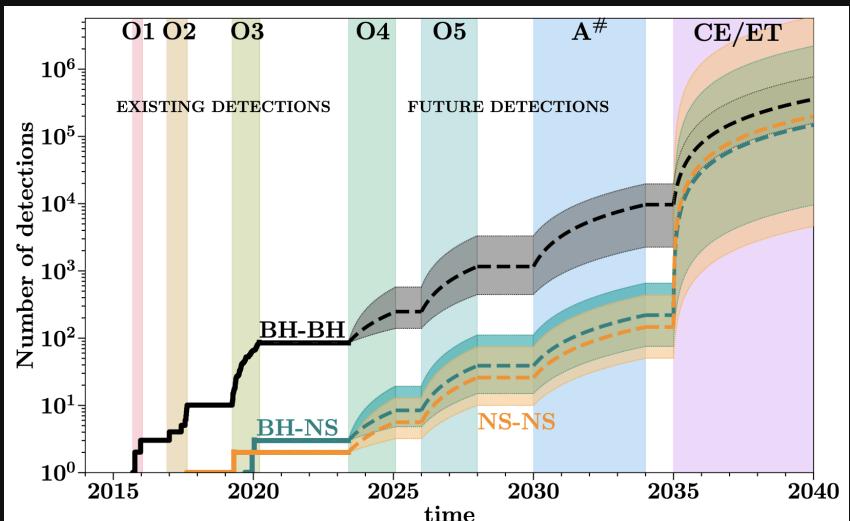
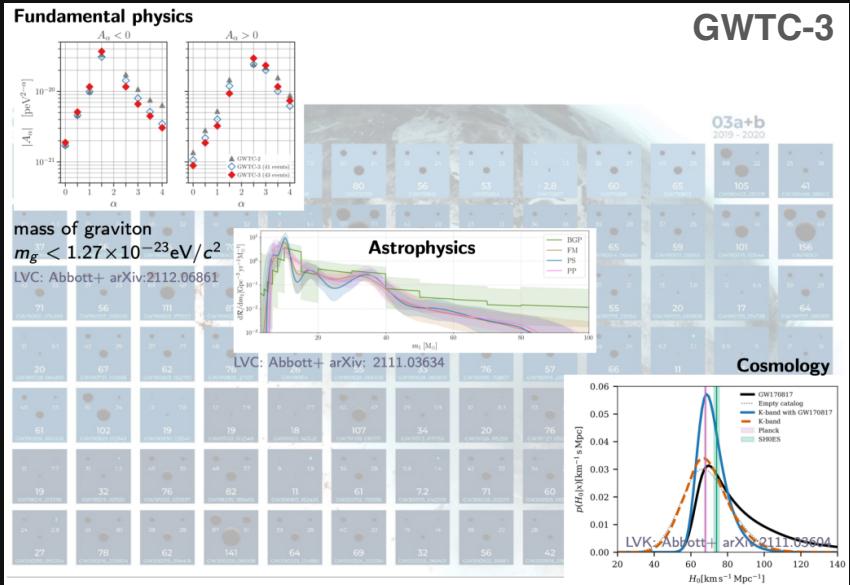
- 天体物理学
 - 大质量恒星演化模型, 恒星级双黑洞的形成机制 ...

- 宇宙学
 - 哈勃常数的测量, 暗能量 ...

- 伯纳德·舒尔茨曾列出成功观测引力波的**五**条关键要素：

- 良好的探测器技术
- 良好的波形模板
- 良好的数据分析方法和技术
- 多个独立探测器间的一致性观测
- 引力波天文学和电磁波天文学的一致性观测

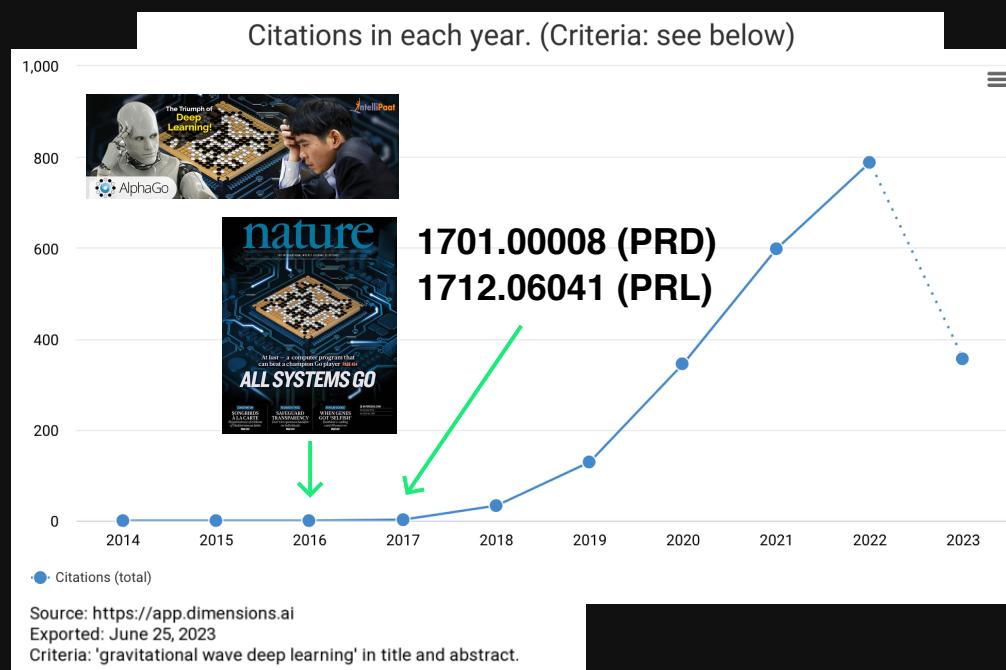
DOI:[10.1063/1.1629411](https://doi.org/10.1063/1.1629411)





引力波天文学：引力波数据分析

- AI for Science → AI for GW
- AI has great potential to revolutionize GW astronomy by improving data analysis, modeling, and detector development.



nature reviews physics

Explore content ▾ About the journal ▾ Publish with us ▾

[nature](#) > [nature reviews physics](#) > [expert recommendation](#) > article

Expert Recommendation | Published: 03 October 2019

Enabling real-time multi-messenger astrophysics discoveries with deep learning

PERSPECTIVE
<https://doi.org/10.1038/s43588-022-00288-z>

Computational challenges for multimodal astrophysics

Elena Cuoco^{1,2,3}, Barbara Patricelli^{1,3,4}, Alberto Iess^{2,3} and Filip Morawski⁵

ARTICLES
<https://doi.org/10.1038/s41567-021-01425-7>

Bayesian parameter estimation using conditional variational autoencoders for gravitational-wave astronomy

Hunter Gabbard^{1,2}, Chris Messenger¹, Ik Siong Heng¹, Francesco Tonolini¹ and Roderick Murray-Smith²

ARTICLES
<https://doi.org/10.1038/s41550-021-01405-0>

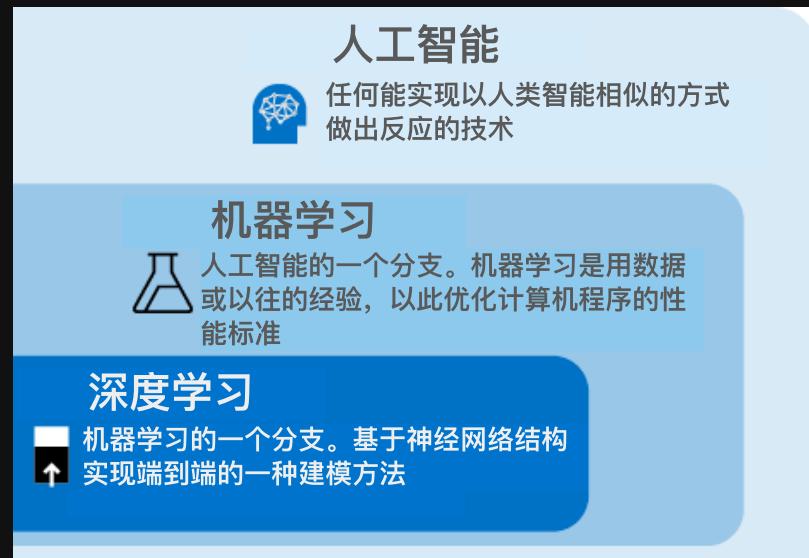
Accelerated, scalable and reproducible AI-driven gravitational wave detection

E. A. Huerta^{1,2}, Asad Khan^{1,3}, Xiaobo Huang³, Minyang Tian³, Maksim Levental², Ryan Chard¹, Wei Wei¹, Maeve Heflin³, Daniel S. Katz³, Volodymyr Kindratenko³, Dawei Mu³, Ben Blaiszik^{1,2} and Ian Foster^{1,2}

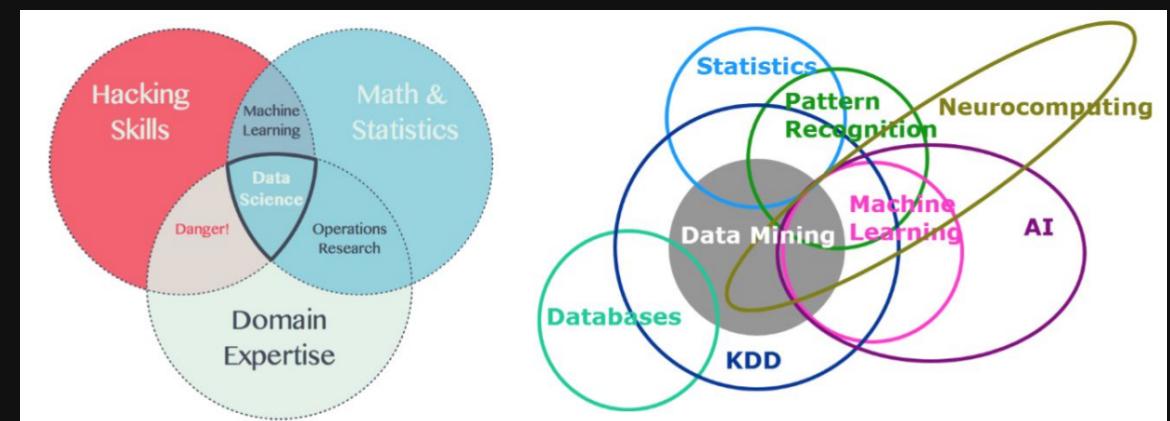


人工智能 > 机器学习 > 深度学习

- 机器学习：
 - 机器学习是人工智能的一个分支，该领域的主要研究对象是人工智能，特别是如何在经验学习中改善具体算法的性能。
 - 线性回归模型、决策树模型、支撑向量机、马尔科夫链-蒙特卡洛方法 (MCMC) ...
- 深度学习：
 - 深度学习就是一种典型的机器学习方法，属于机器学习的分支。是一种用神经网络实现自动特征提取的模型
 - 深度神经网络是一个万能的函数拟合器，可以表征任意复杂度的非线性函数映射
 - 特点：端到端、数据驱动、过参数化 ...
- 传统引力波数据分析方法 ~ 传统机器学习方法
- 数据驱动，在数据上通过算法总结规律模式，应用在新数据上。



人工智能 > 机器学习 > 深度学习

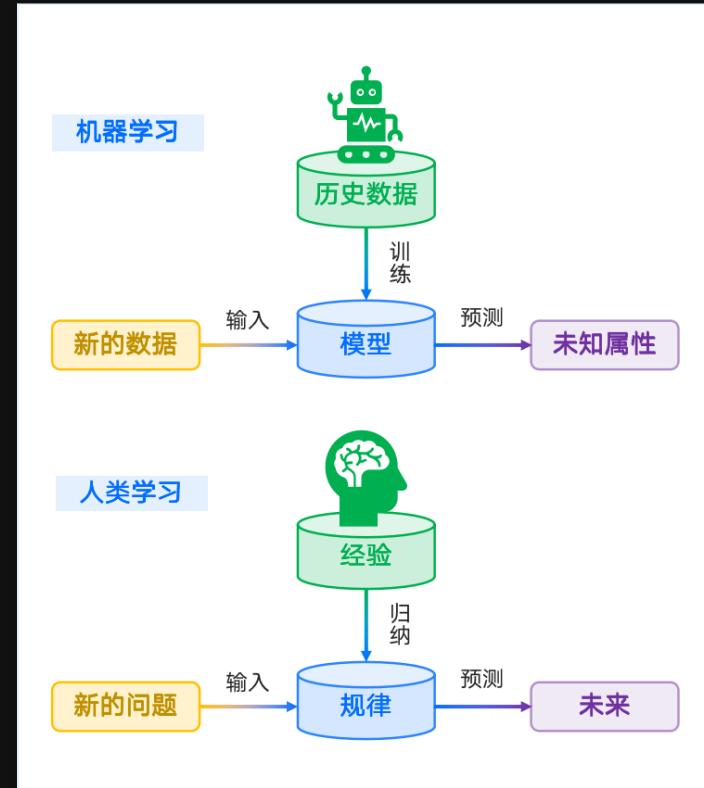


Knowledge Discovery in Database, KDD



机器学习的定义

- “机器学习是对能通过经验自动改进的计算机算法的研究。”
Machine Learning is the study of computer algorithms that improve automatically through experience.
- “机器学习是用数据或以往的经验，以此优化计算机程序的性能标准。”
Machine learning is programming computers to optimize a performance criterion using example data or past experience.
—— **Alpaydin (2004)**
- A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P , if its performance at tasks in T , as measured by P , improves with experience E . — **Tom Mitchell (1997)**





机器学习的目标

- 任务 [task]: 判断某草莓是否是甜草莓
 - 机器学习就是找到草莓的不同 特征 [feature] 维度 (尺寸、颜色、成熟度、...) 与草莓 标签 [label] (酸、甜) 之间的映射关系。

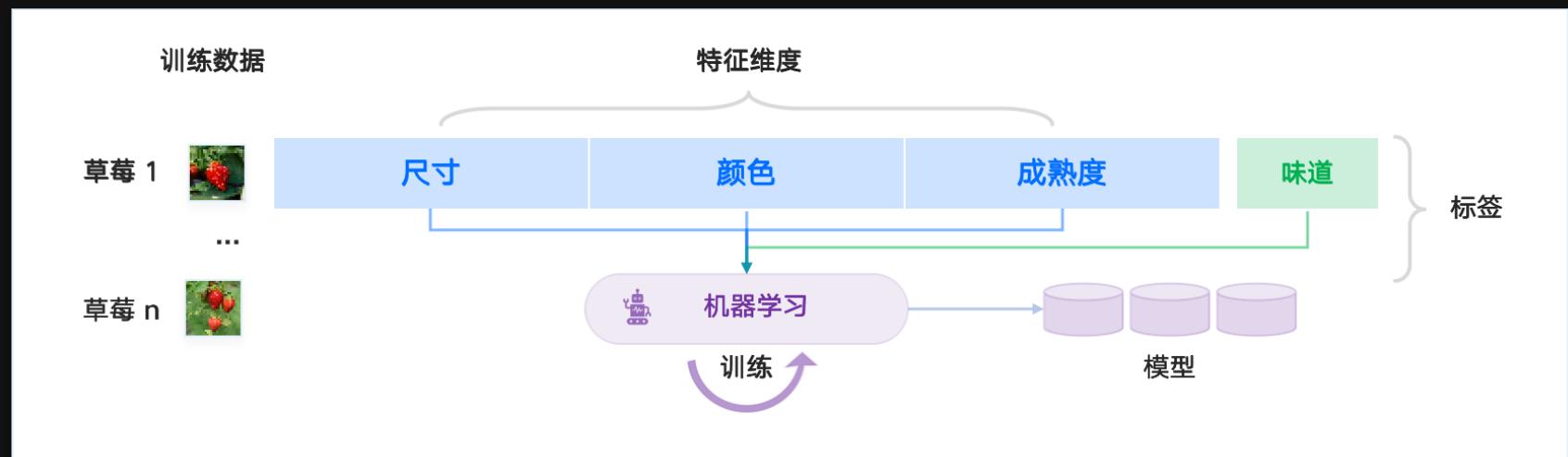
		特征维度				
		尺寸	颜色	成熟度	味道	标签
草莓 1		小	红色	成熟	甜	
		大	粉红色	半熟	酸	

可能的参数值



机器学习的过程

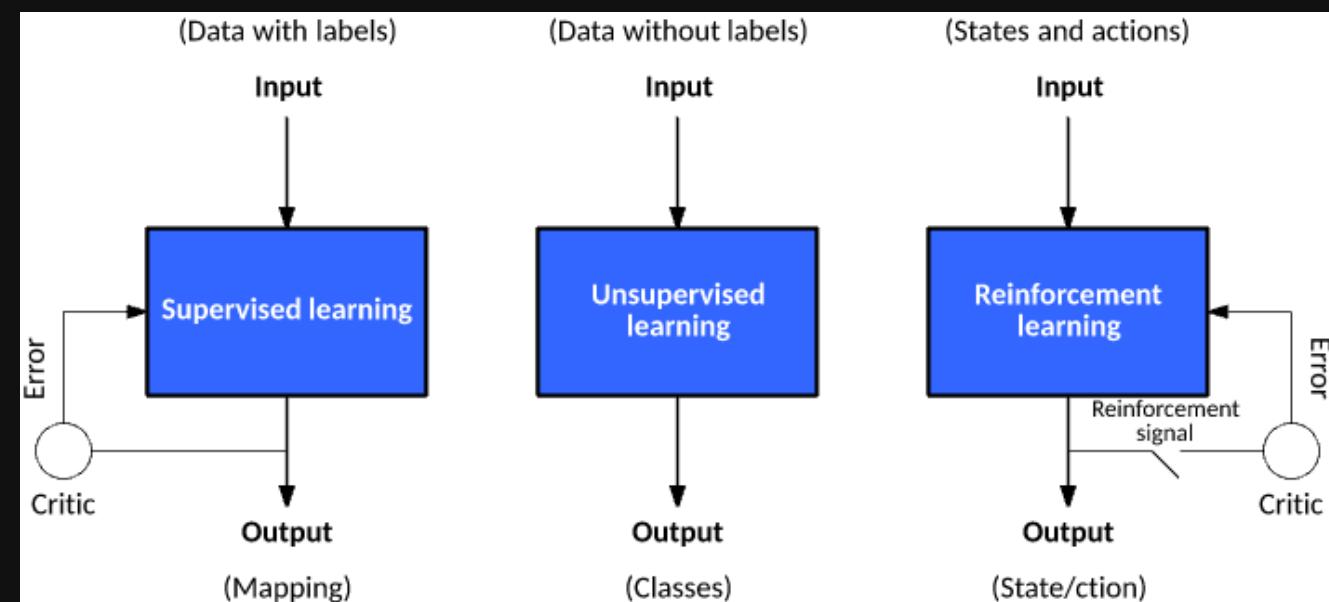
- 机器学习就是找到 **特征 [feature]** 与 **标签 [label]** 之间的关系，利用算法从一类训练数据或信息中自动分析并获得该类数据或信息的规律，并利用获取的规律对未知数据进行预测。
- 上述寻找关系和规律的过程，称为 **训练 [train]**。训练完成后的结果，是得到一个 **机器学习模型 [machine learning model]**。





机器学习的常见类型

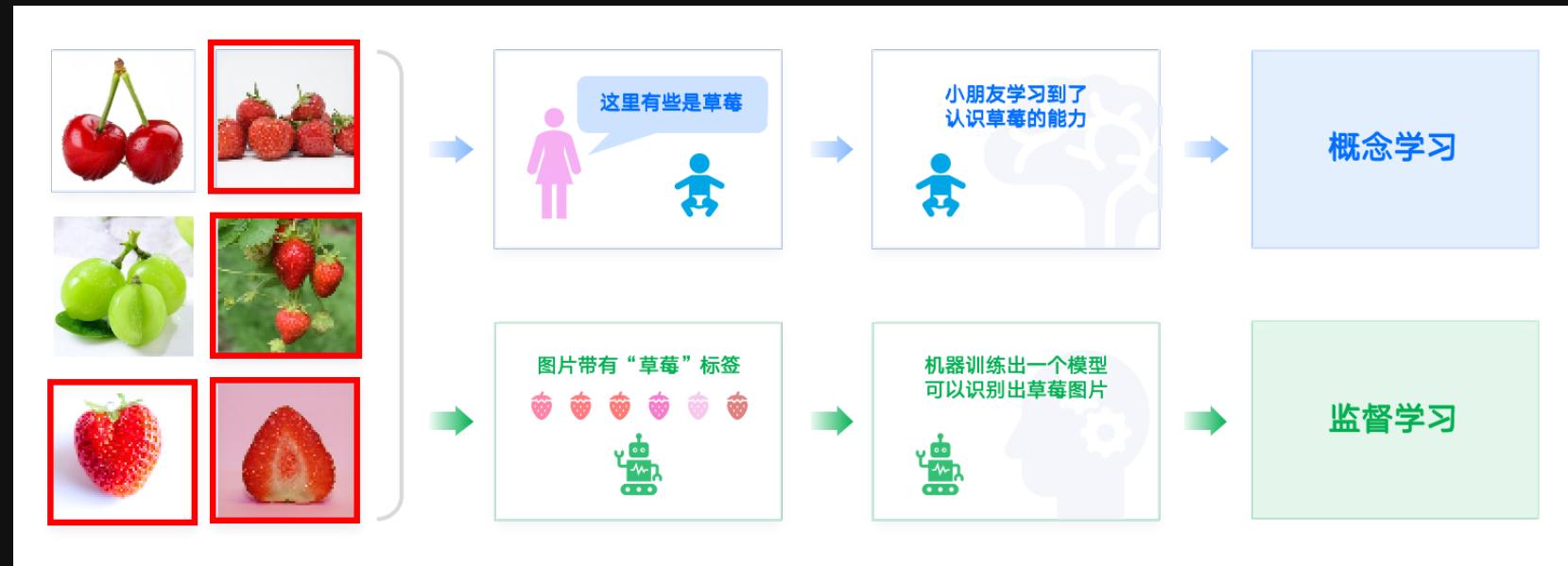
- 机器学习模型可以根据训练数据标签、与环境交互的方式，大致分为三种常见类型：
 - 监督学习 (supervised learning)
 - 无监督学习 (un-supervised learning)
 - 强化学习 (reinforcement learning)





机器学习的常见类型：监督学习

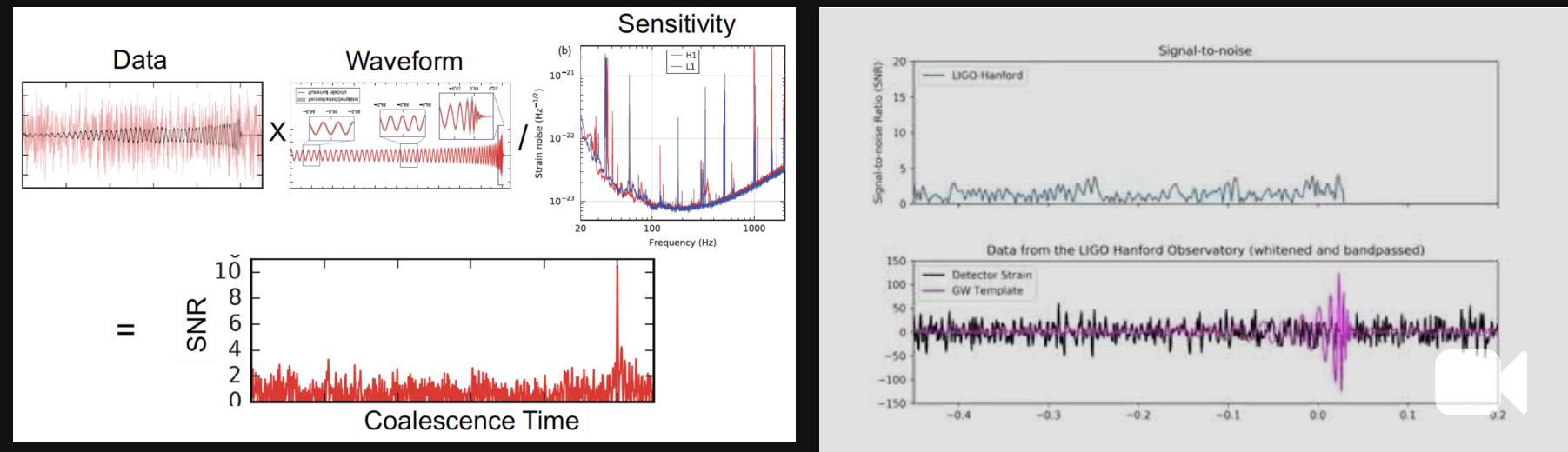
- 监督学习 是在有指导的前提下让机器进行学习，这种指导的关键是给训练数据标注好 **标签 [label]**。
- 监督学习的目标在观察完一些事先标注过的训练数据（输入和预期输出）后，这个模型对任何可能出现的输入去预测其输出。要达到此目的，学习者必须以“合理”（归纳规律）的方式从现有的数据中一般化到未观察到的情况。在人类和动物感知中，则通常被称为 概念学习。





机器学习的常见类型：监督学习 vs 匹配滤波

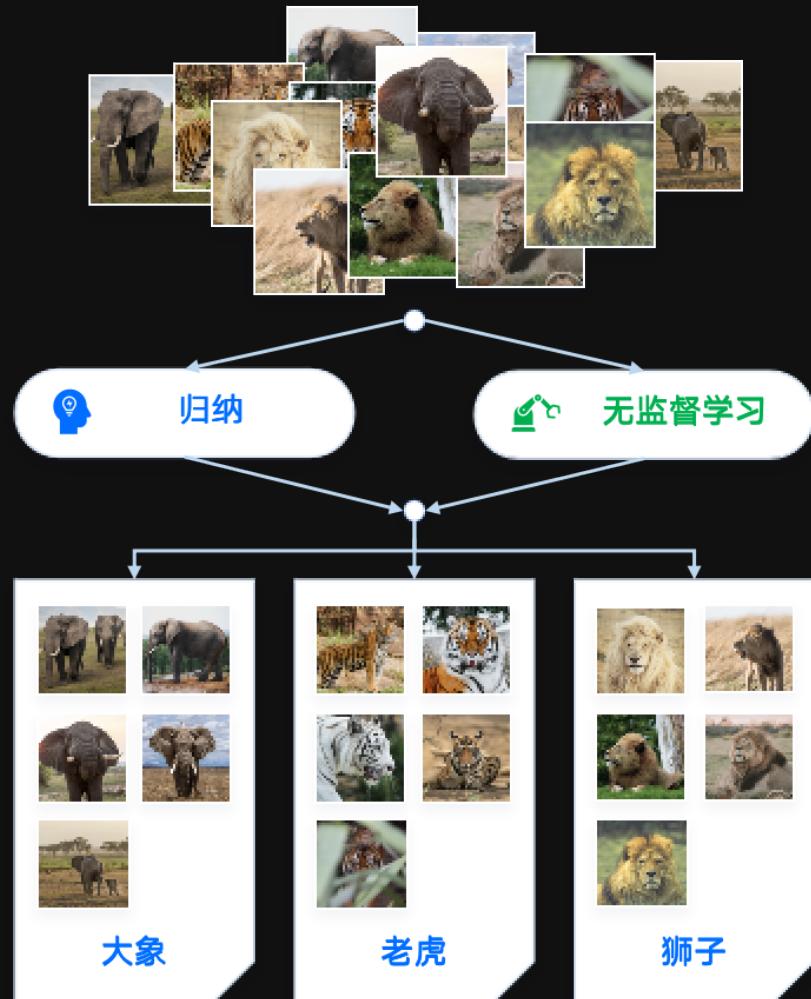
- 监督学习 是在有指导的前提下让机器进行学习，这种指导的关键是给训练数据标注好 **标签 [label]**。
- 监督学习的目标在观察完一些事先标注过的训练数据（输入和预期输出）后，这个模型对任何可能出现的输入去预测其输出。要达到此目的，学习者必须以“合理”（归纳规律）的方式从现有的数据中一般化到未观察到的情况。在人类和动物感知中，则通常被称为 概念学习。
- 基于模板的引力波信号搜寻：
 - 若某一段时域数据流作为输入，探测统计量（即匹配滤波信噪比）是另一段输出的时序数据流，问怎样的线性滤波器（**模板**）可以使得输出结果最大？





机器学习的常见类型：非监督学习

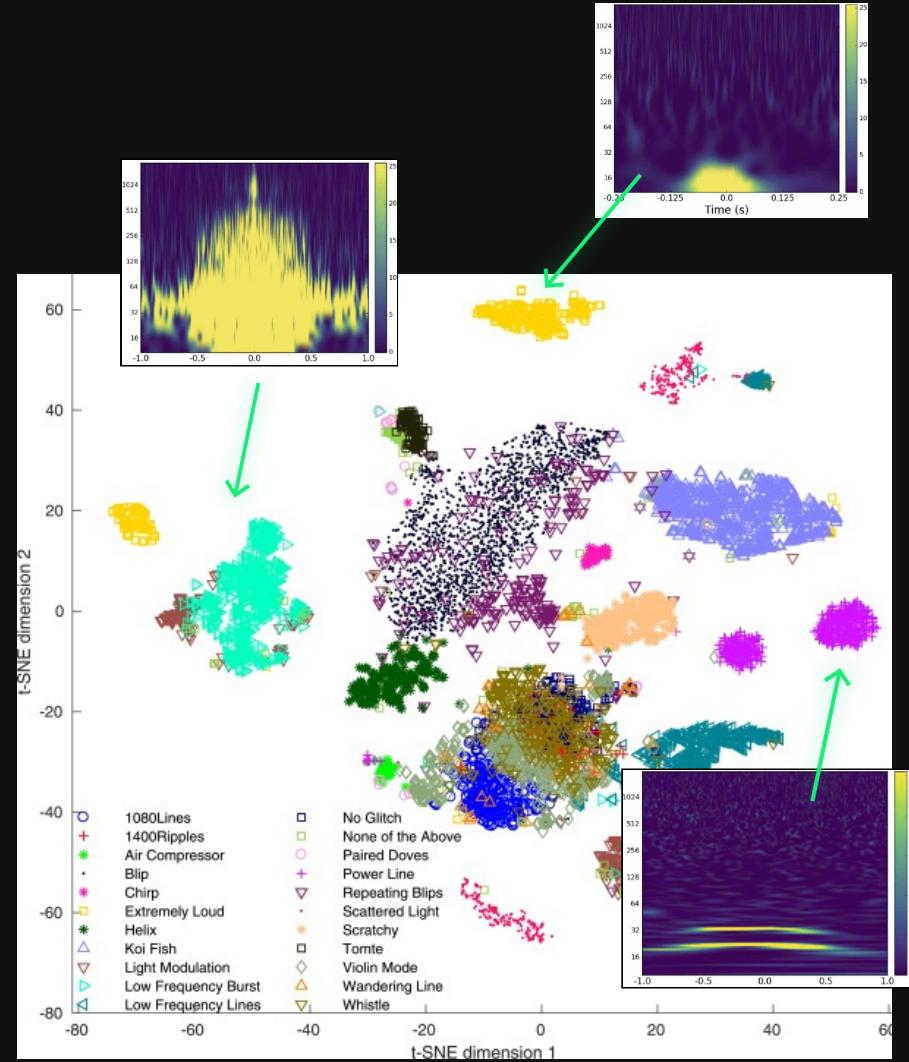
- **无监督学习** 是没有指导的学习过程，待学习的训练数据没有标签的。
- 机器学习算法通过某种方式在数据中寻找共同特征，并将有共同特征的数据聚合在一起。有时也会把这个过程成为“**聚类**”(cluster)。
- 聚类是把相似的对象通过静态分类的方法分成不同的组别或者更多的子集，这样让在同一个子集中的成员对象都有相似的一些属性。
- 无监督学习算法通过自由的探索数据，所学到的大部分内容必须包括**理解数据本身**，而不是将这种理解应用于特定任务。所以，通往通用智能的道路上必须要掌握无监督学习的技能。
- 无监督学习的过程和人类的 **归纳** 学习过程相似。





机器学习的常见类型：非监督学习

- **无监督学习** 是没有指导的学习过程，待学习的训练数据没有标签的。
- 机器学习算法通过某种方式在数据中寻找共同特征，并将有共同特征的数据聚合在一起。有时也会把这个过程成为“**聚类**”(cluster)。
- 聚类是把相似的对象通过静态分类的方法分成不同的组别或者更多的子集，这样让在同一个子集中的成员对象都有相似的一些属性。
- 无监督学习算法通过自由的探索数据，所学到的大部分内容必须包括**理解数据本身**，而不是将这种理解应用于特定任务。所以，通往通用智能的道路上必须要掌握无监督学习的技能。
- 无监督学习的过程和人类的 归纳 学习过程相似。



• DOI:[10.1016/j.ins.2018.02.068](https://doi.org/10.1016/j.ins.2018.02.068)



机器学习的其他类型

- 半监督学习 (semi-supervised learning)
- 自监督学习 (self-supervised learning)
- ...

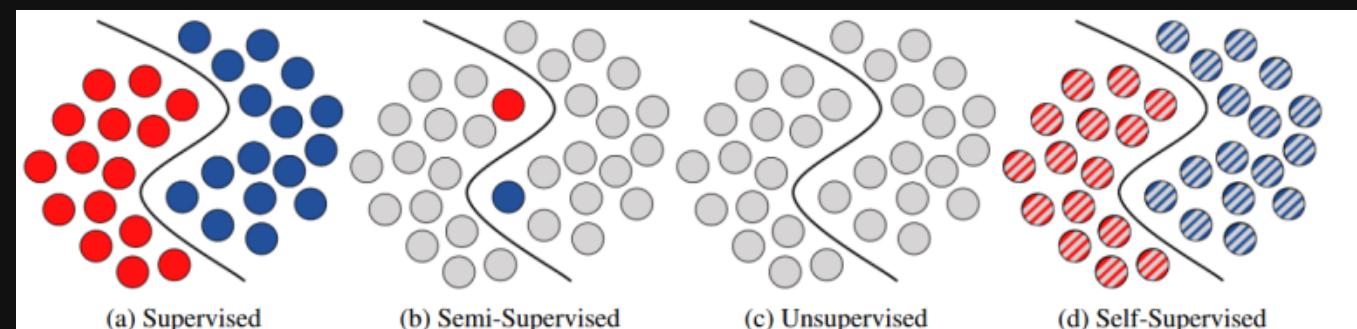
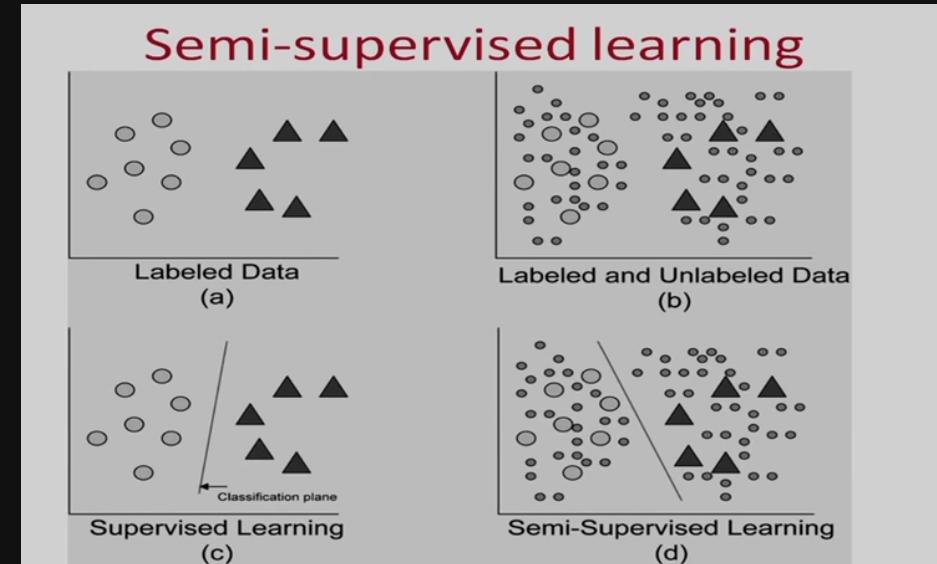
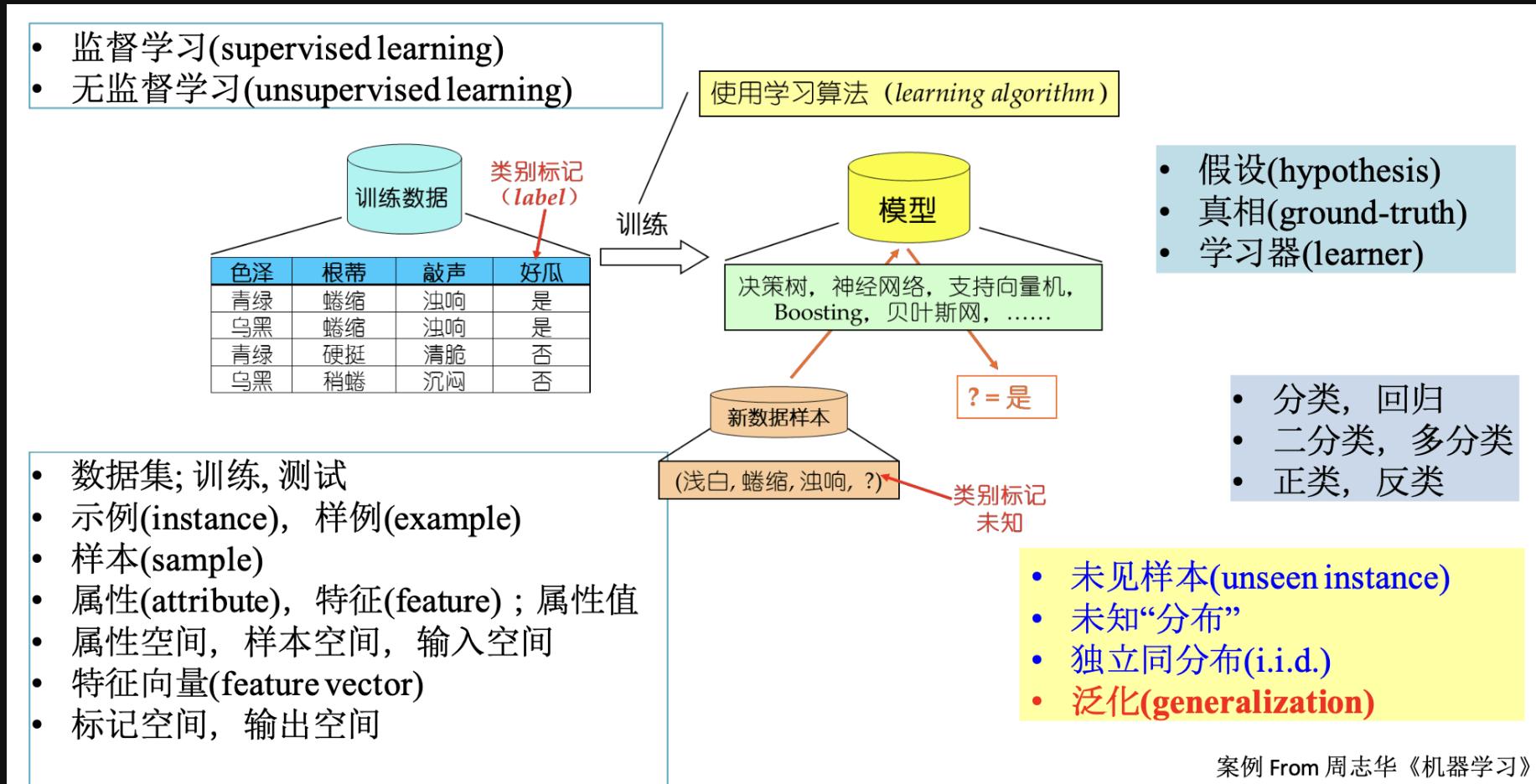


Figure 2: Illustrations of the four presented deep learning strategies - The red and dark blue circles represent labeled data points of different classes. The light grey circles represent unlabeled data points. The black lines define the underlying decision boundary between the classes. The striped circles represent datapoints which ignore and use the label information at different stages of the training process.

2002.08721



机器学习的基本术语

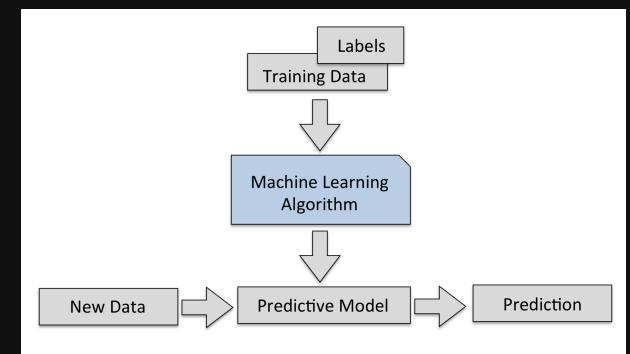


- <https://developers.google.cn/machine-learning/glossary/?hl=zh-CN#a>
- <https://semanti.ca/blog/?glossary-of-machine-learning-terms>



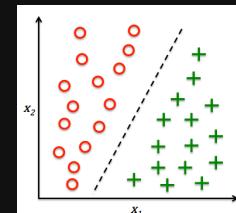
机器学习模型的分类

- 基于**监督学习**进行预测



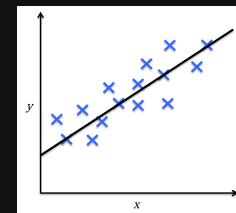
分类问题（不同类别预测）

根据数据样本上抽取出的特征，判定其属于有限个类别中的哪一个



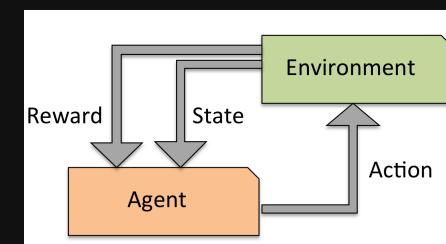
回归问题（对连续值进行预测）

根据数据样本上抽取出的特征，预测连续值结果



- 与环境不断交互的**强化学习**过程

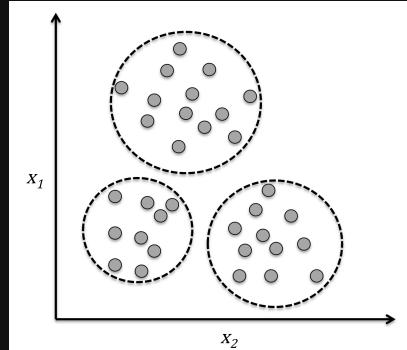
研究如何基于环境而行动，以取得最大化的预期利益



- 试图从无标签数据里总结模式的**无监督学习**

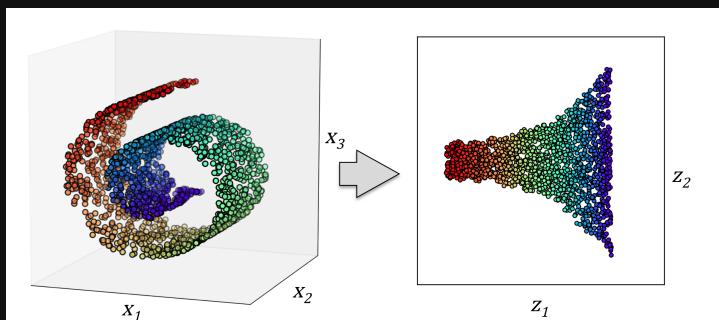
用聚类发现子簇

根据数据样本上抽取出的特征，挖掘数据的关联模式



数据降维

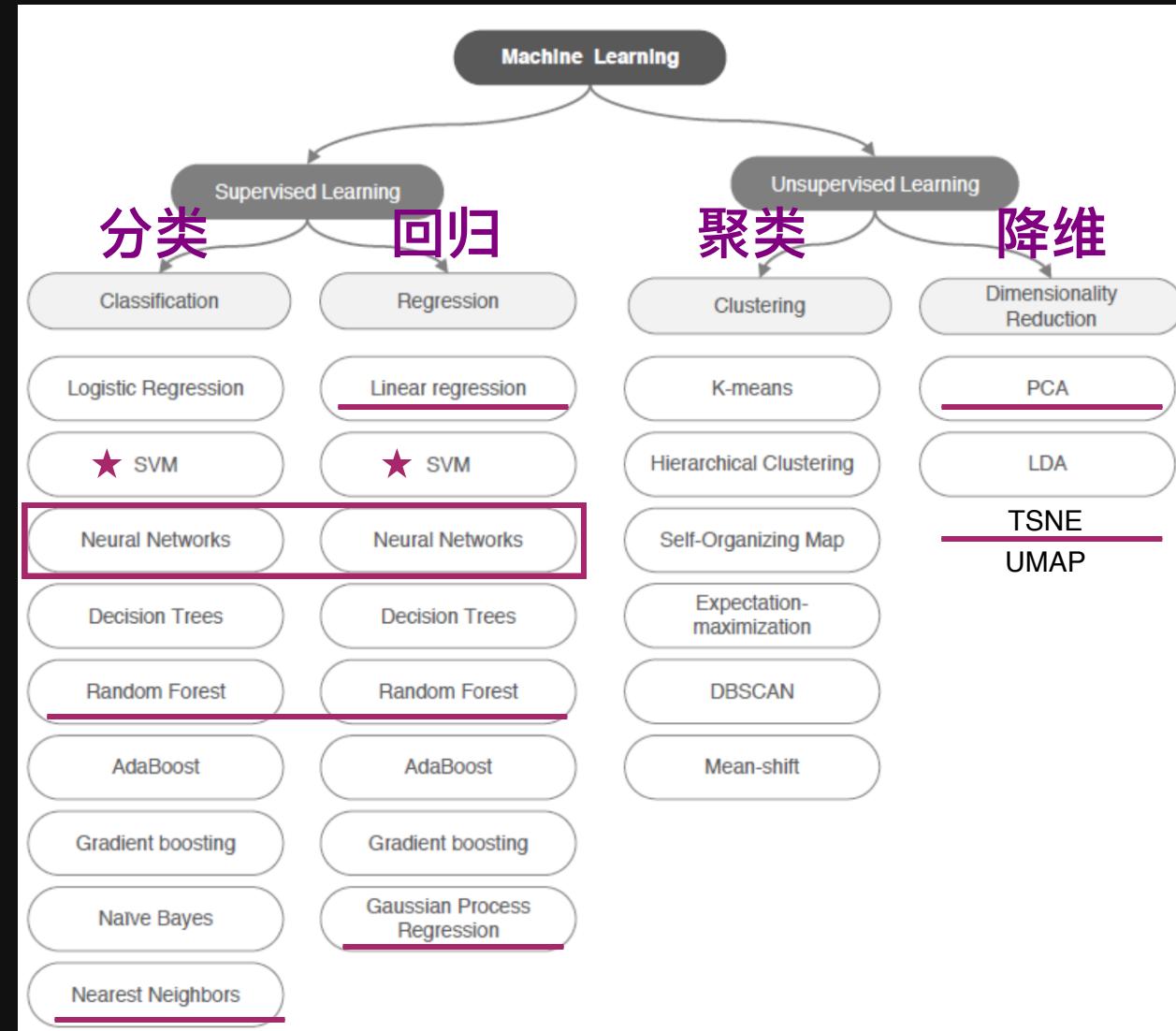
发现数据中的隐藏模式和结构





机器学习模型的分类

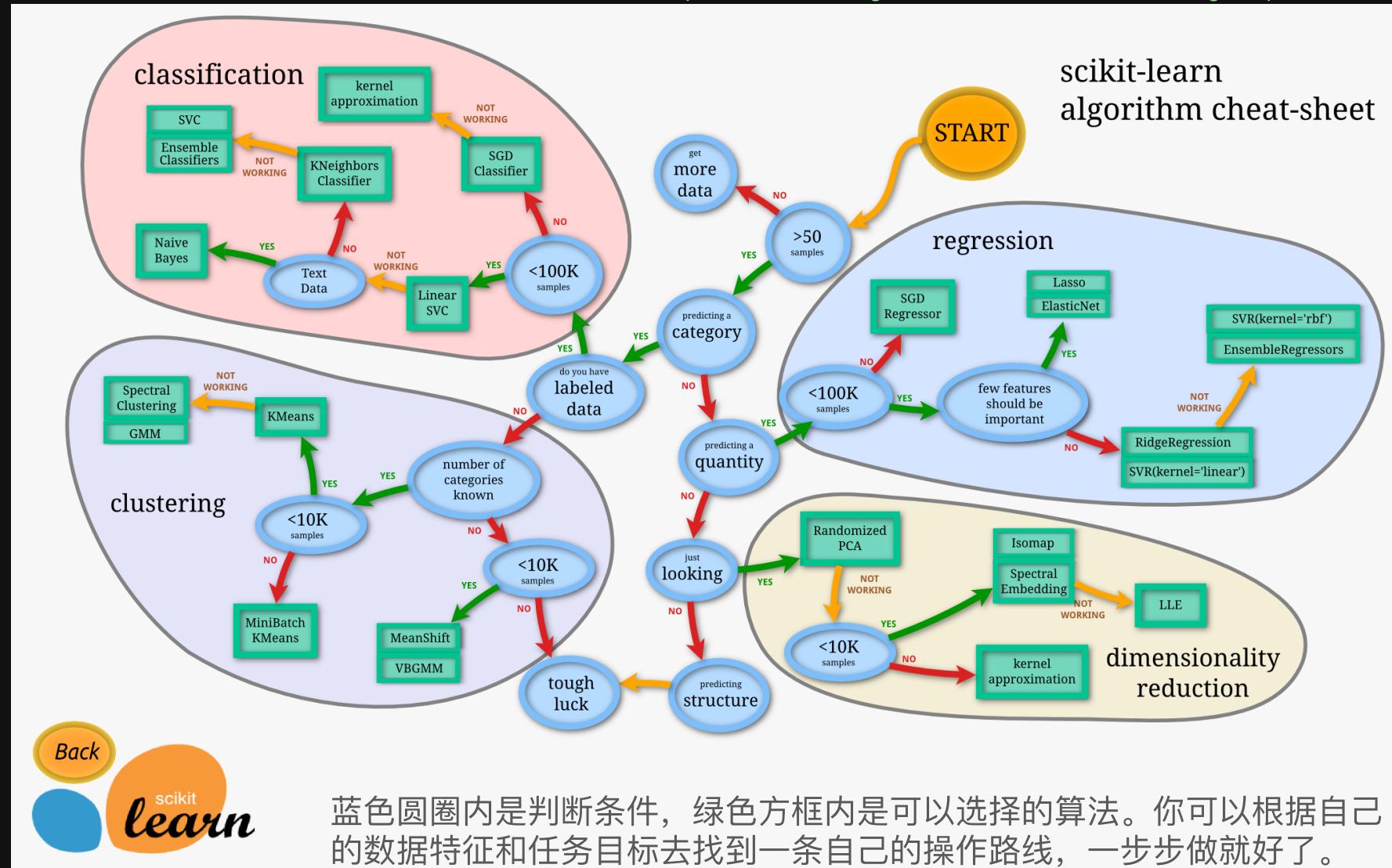
- 根据 **数据标签** 分类





机器学习模型的分类

https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

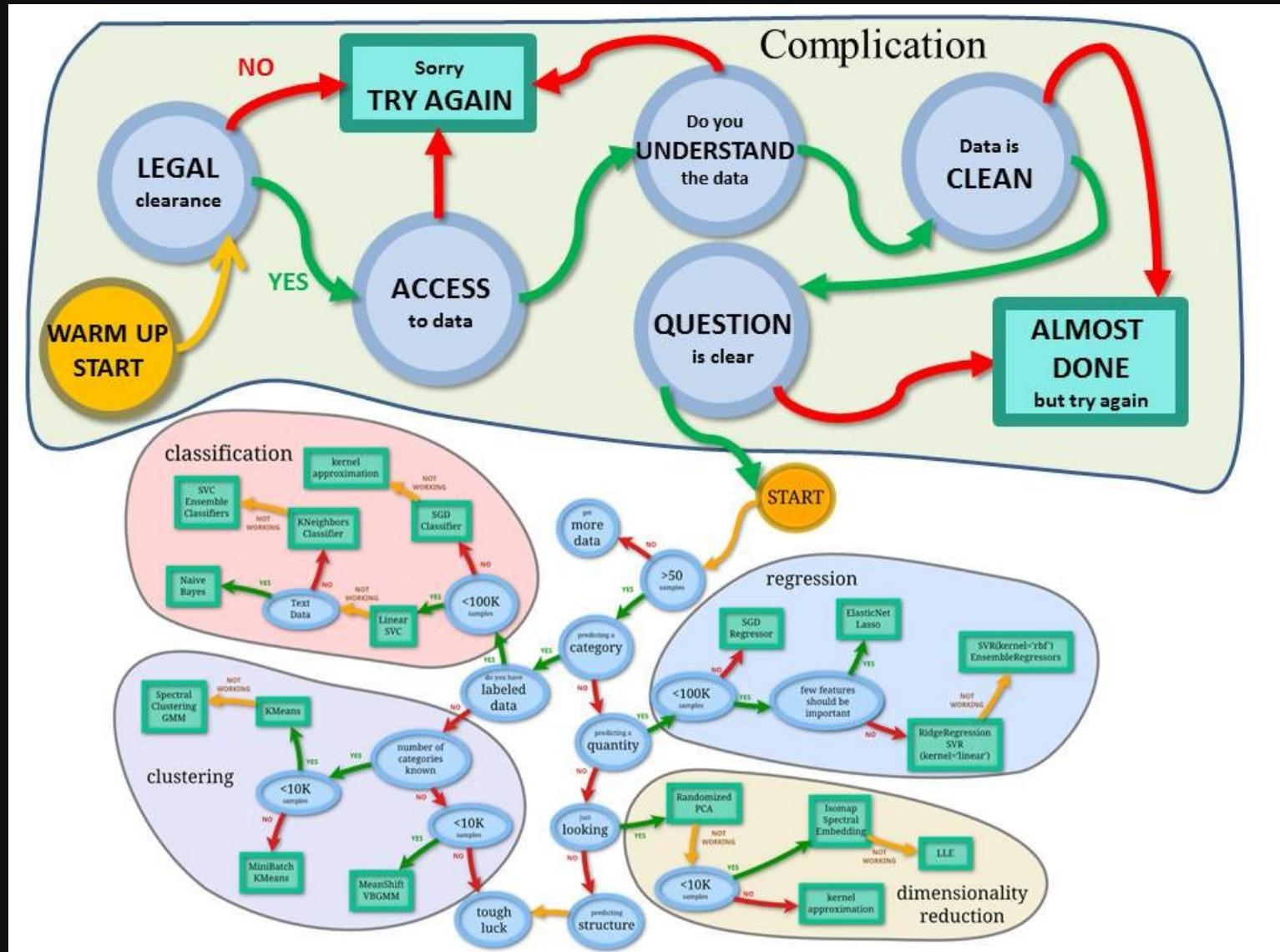


蓝色圆圈内是判断条件，绿色方框内是可以选择的算法。你可以根据自己的数据特征和任务目标去找到一条自己的操作路线，一步步做就好了。



机器学习模型的分类

https://medium.com/@chris_bour/an-extended-version-of-the-scikit-learn-cheat-sheet-5f46efc6cbb



常用的回归：线性、决策树、SVM、KNN

集成回归：随机森林、Adaboost、GradientBoosting、Bagging、ExtraTrees

常用的分类：线性、决策树、SVM、KNN，朴素贝叶斯；

集成分类：随机森林、Adaboost、GradientBoosting、Bagging、ExtraTrees

常用聚类：
k均值(K-means)、层次聚类(Hierarchical clustering)、DBSCAN

常用降维：
LinearDiscriminantAnalysis、PCA



机器学习模型的分类

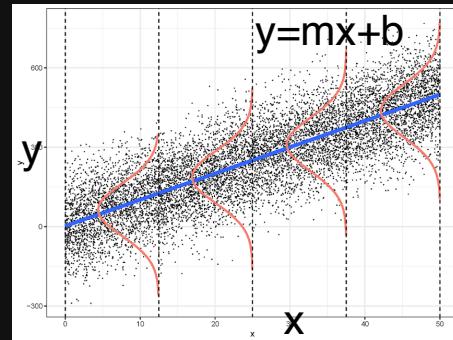
注: 有些时候数据没有提供足够信息来事先假设分布、或者问题本身没有明显的分布特性

- 根据 **数据分布** 分类: 参数 vs 非参数模型
 - 这里的“参数”并不是模型中的参数, 而是数据分布的参数
- 参数模型:**
 - 对数据分布进行假设, 待求解的数据模式/映射可以用一组有限且固定数目的模型参数进行刻画

条件概率 $P(Y|X)$
属于高斯分布



线性回归模型



如: 线性/逻辑回归、感知机、K 均值聚类

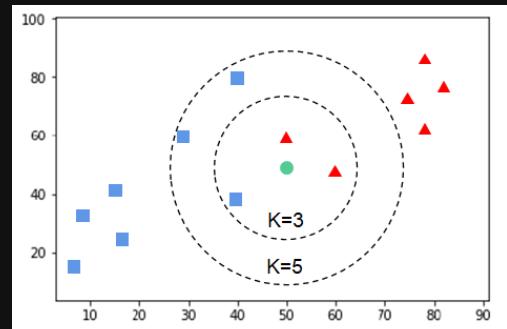
优点: 简洁、快速、数据需求更少

局限性: 指定函数形式、有限的复杂度、易欠拟合

• 非参数模型:

- 不对数据分布进行假设, 数据的所有统计特性都来源于数据本身
- 非参数模型的时空复杂度一般比参数模型大得多
- 非参数模型是自适应数据的, 模型参数随样本变化而变化

K 近邻模型



如: 随机森林、朴素贝叶斯、SVM、**神经网络**

优势: 函数可变性、模型强大假设少、拟合性好

局限性: 数据需求量大、速度慢、易过拟合、预测解释性不高



机器学习项目开发规划与准备

- 发现与明确问题

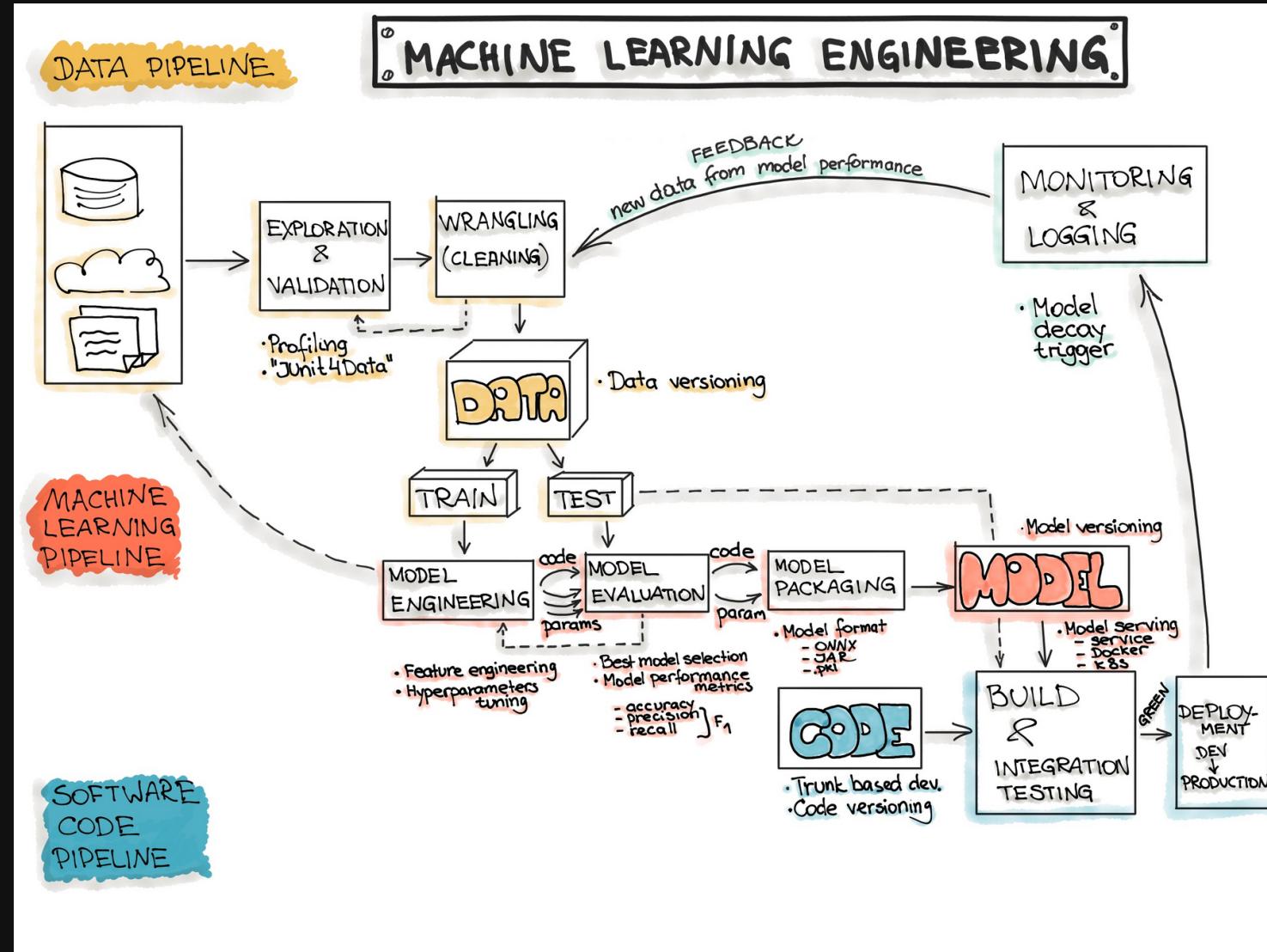
- AI 开发的目标是将隐藏在海量数据背后的信息集中处理并进行提炼，从而总结得到研究对象的内在规律。
- 在开始 AI 开发之前，需要进行多角度思考：
 - 科学（痛点）
 - 要解决什么科学问题？目标是什么？
 - 预期结果是什么？成功的量化衡量方法是？
 - 技术（难点）
 - 要 AI 从数据中学习的是什么 表征 [representation]？
 - 与非 AI 方法相比，预期结果是什么？
 - 足够支持解决问题需要多少数据？能获取到足够数据吗？需要多长时间？ ...
- 考虑这几个问题并不是浪费时间，对于任何一个机器学习和数据分析的工作来说，都是很有必要且不可或缺的步骤。然后就可以从数据探索开始了。
- 对数据进行分析，一般通过使用适当的统计、机器学习、深度学习等方法，对收集的大量数据进行计算、汇总和整理，以求最大化地开发数据价值，发挥数据作用。





机器学习项目开发规划与准备

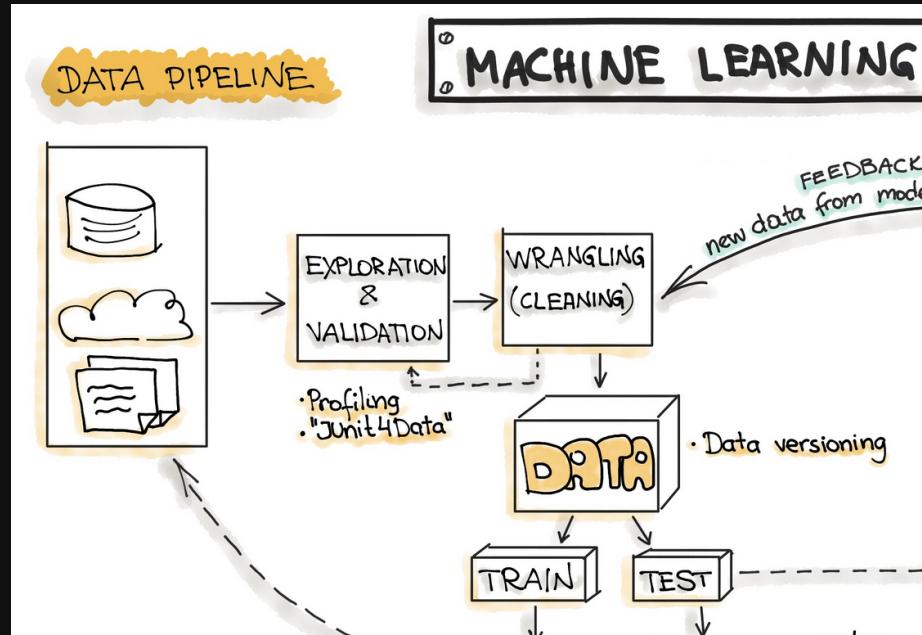
- 开发流程





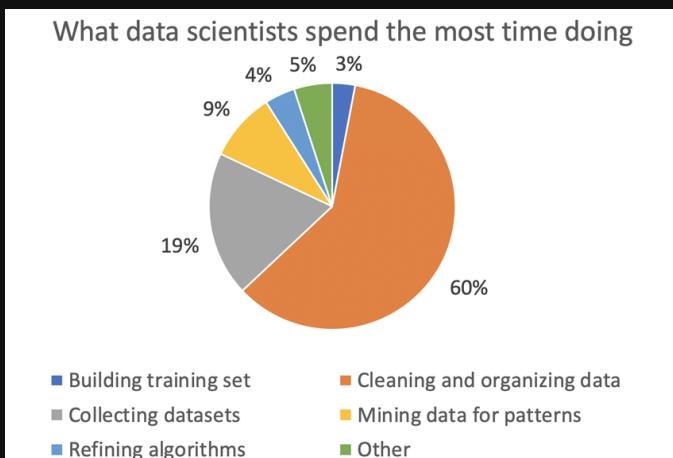
机器学习项目开发规划与准备

- 数据准备



- 在大部分人工智能项目工作时间中，数据的准备和数据 pipeline 的搭建占到近 80% 的工作量。

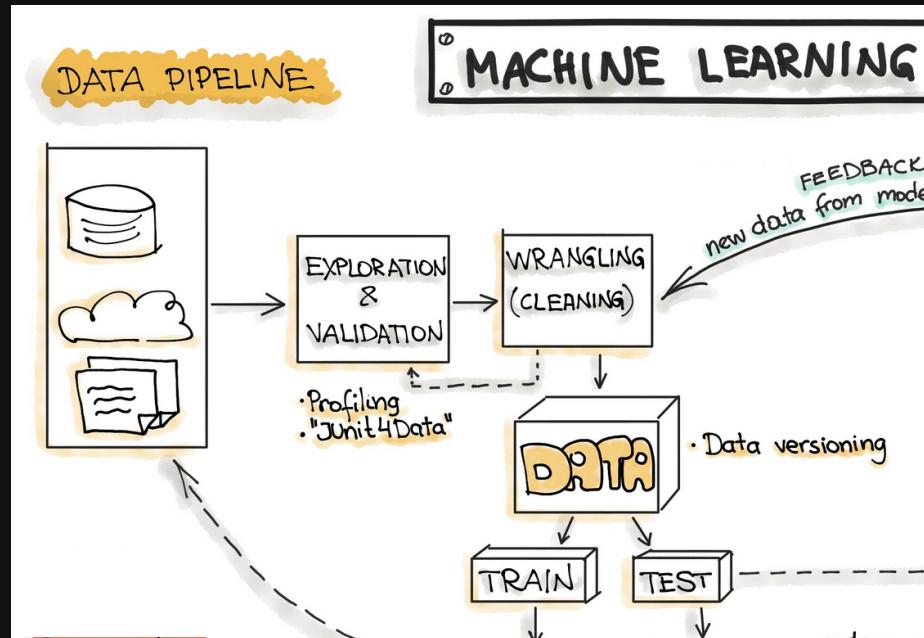
- 数据观察与数据清理十分繁琐，但它却是数据分析的关键步骤。如果我们跳过这个阶段直接进入建模，会导致错误的数据模型。
- 记住：**错误的数据导致错误的模型。永远要从检查数据开始。**
- 我们要尽可能地把数据清洗和探索性分析，这样才能对数据集的分布和关系有初步的认识。



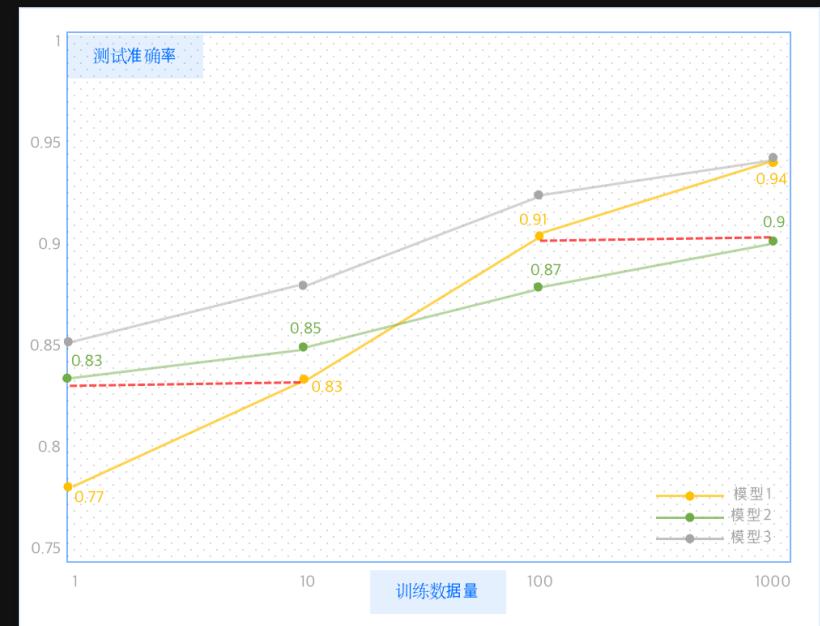
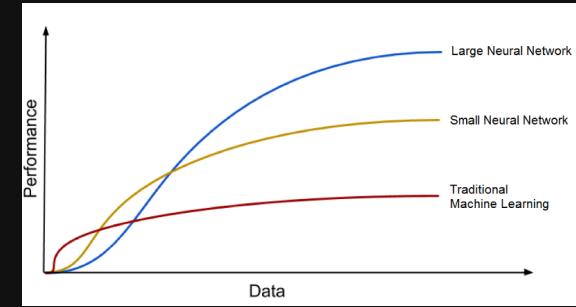


机器学习项目开发规划与准备

- 数据准备



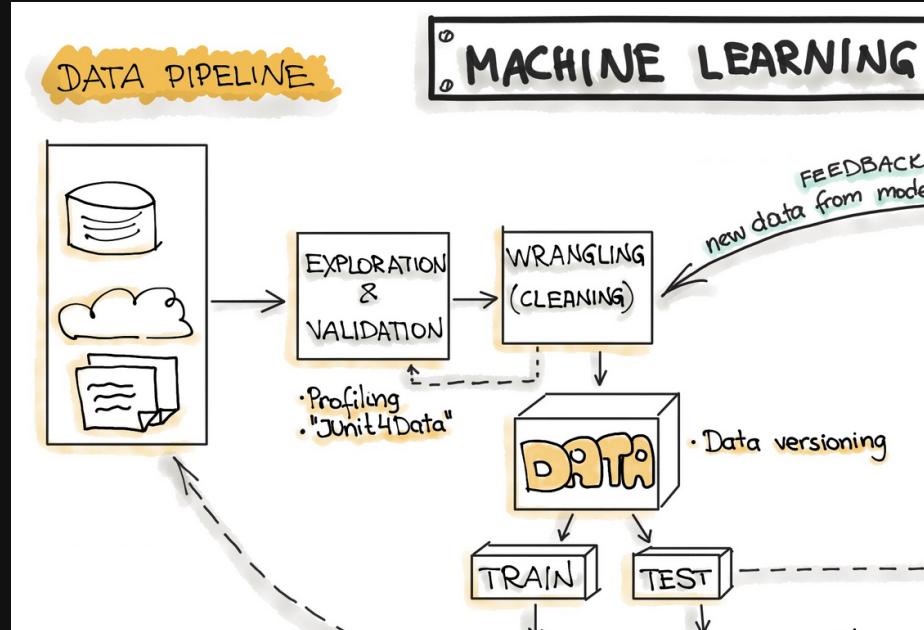
- 在大部分人工智能项目工作时间中，数据的准备和数据 pipeline 的搭建占到近 80% 的工作量。
- 一般来说，训练数据规模越大，越可以带来的更好的 AI 模型性能。





机器学习项目开发规划与准备

- 数据准备

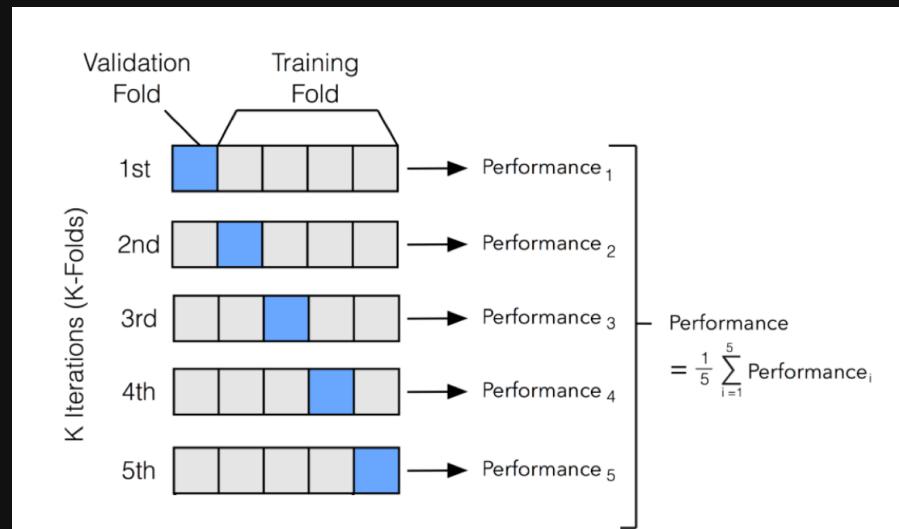


- 数据标注后需要划分为训练集 [train] 与测试集 [test] (简版)

- 训练集：是数据集的一个随机子集，用于完成模型训练任务
- 测试集：也是数据集的一个随机子集(与训练集互斥分开)，用于验证模型的准确性，以及对模型的泛化效果进行检验
- 拆分比例根据具体任务决定，通常训练集的比重较大，一个可能的划分比例是：训练集数据数量 : 测试集数据数量 = 8 : 2
- 需要注意的是，一旦我们把数据集划分为训练集和测试集，那么我们在建模的过程中，就不能再使用测试集的任何数据，否则就是作弊哦。

为避免过拟合问题，大部分的数据科学家都会对数据模型进行“**K层交叉检验(K-fold cross-validation)**”：

- 把原始的数据集划分为K个子集，使用其中一个子集作为测试集，其他子集都用作训练集。这个过程重复K次，这样每个子集都会成为一次测试集。
- 10 层交叉验证是最常用的。

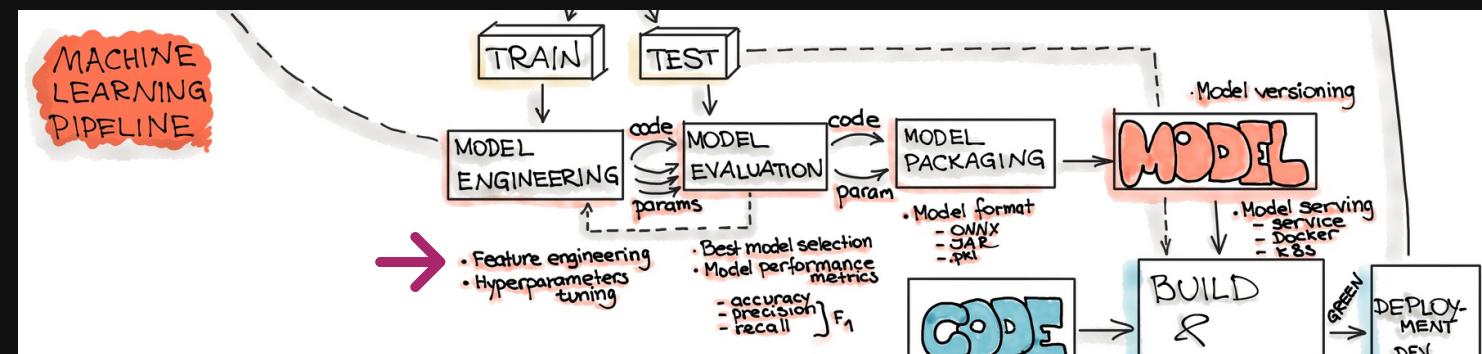
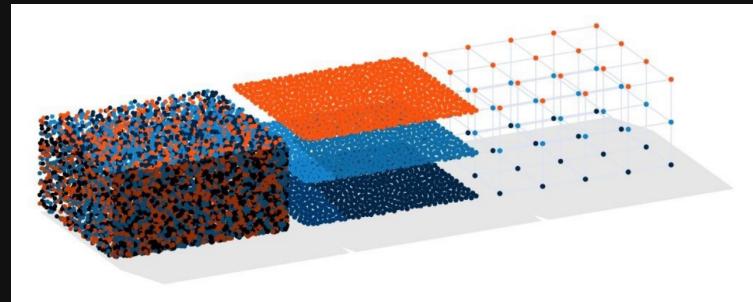




机器学习项目开发规划与准备

- 特征工程 [feature engineering]

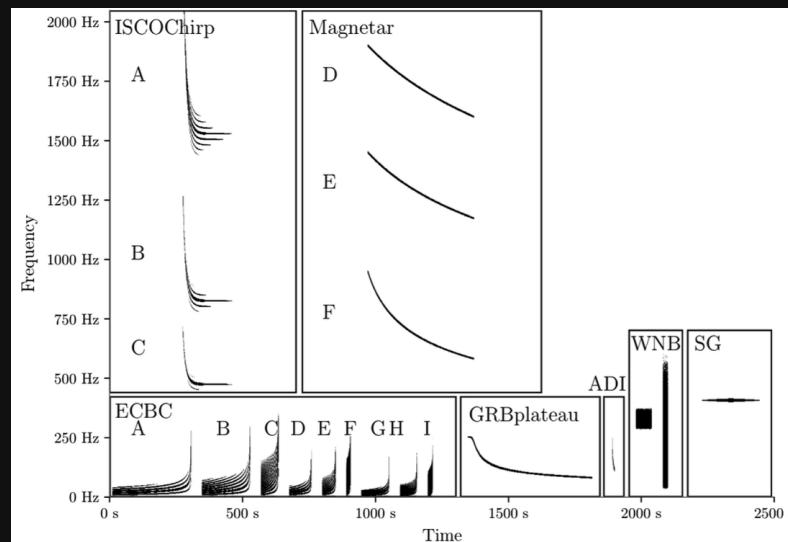
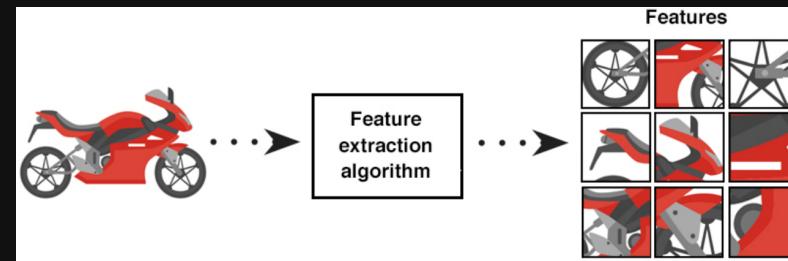
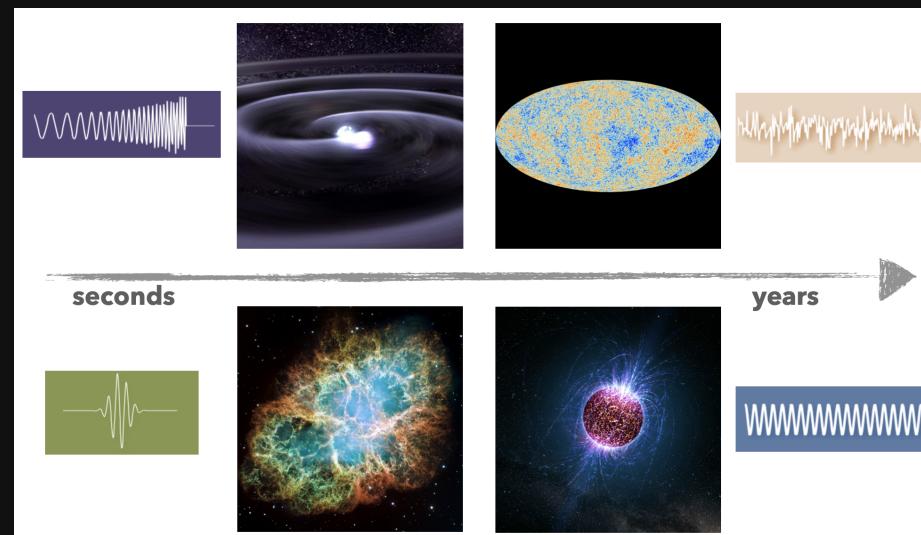
- 特征工程是指从原始数据转换为特征向量的过程。
- 特征工程是机器学习中最重要的起始步骤，会直接影响模型的效果，通常需要大量的时间来完成。
- 数据和特征决定了机器学习的上限，而模型和算法只是逼近这个上限。





机器学习项目开发规划与准备

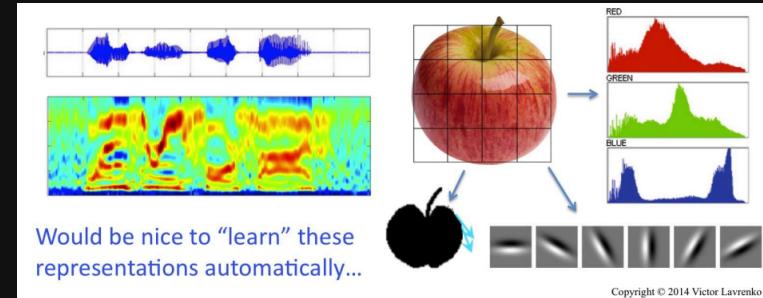
- 特征工程 [feature engineering]
 - 特征提取 (feature extraction) 一般是在特征选择之前，它提取的对象是原始数据，目的就是自动地构建新的特征，将原始数据转换为一组具有明显物理意义（比如 Gabor、几何特征、纹理特征）或者统计意义的特征（也包括PCA, SVD等方法在内）





机器学习项目开发规划与准备

- 特征工程 [feature engineering]
 - 特征构建 (feature construction) 指从原始数据中人工的构建新的特征。
 - 需要花时间去观察原始数据，思考问题的潜在形式和数据结构，对数据敏感性和机器学习实战经验能帮助特征构建。



Copyright © 2014 Victor Lavrenko

apartment_length	apartment_breadth	apartment_price	apartment_length	apartment_breadth	apartment_area	apartment_price
80	59	23,60,000	80	59	4,720	23,60,000
54	45	12,15,000	54	45	2,430	12,15,000
78	56	21,84,000	78	56	4,368	21,84,000
63	63	19,84,000	63	63	3,969	19,84,000
83	74	30,71,000	83	74	6,142	30,71,000
92	86	39,56,000	92	86	7,912	39,56,000

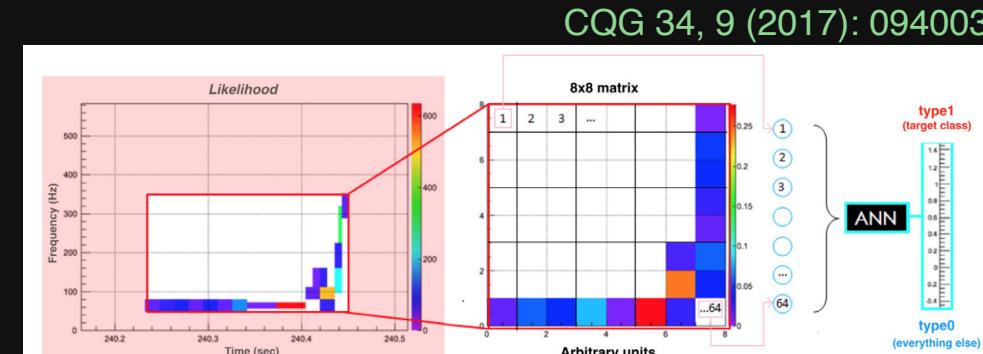
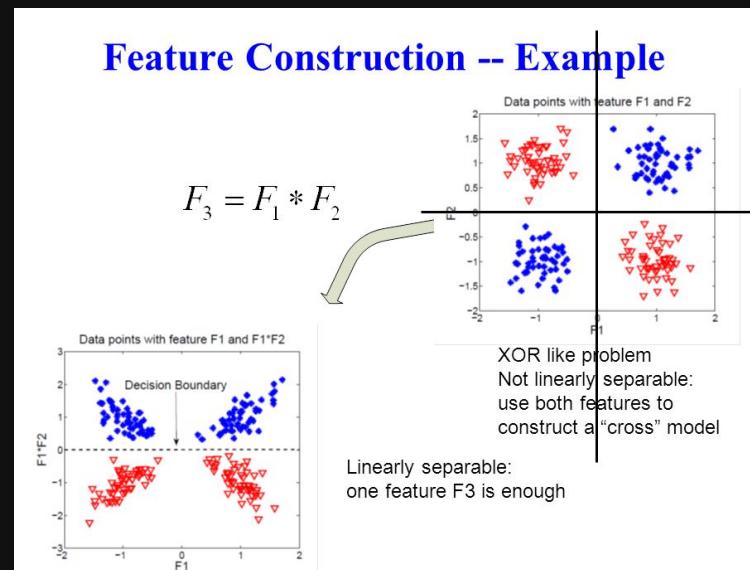
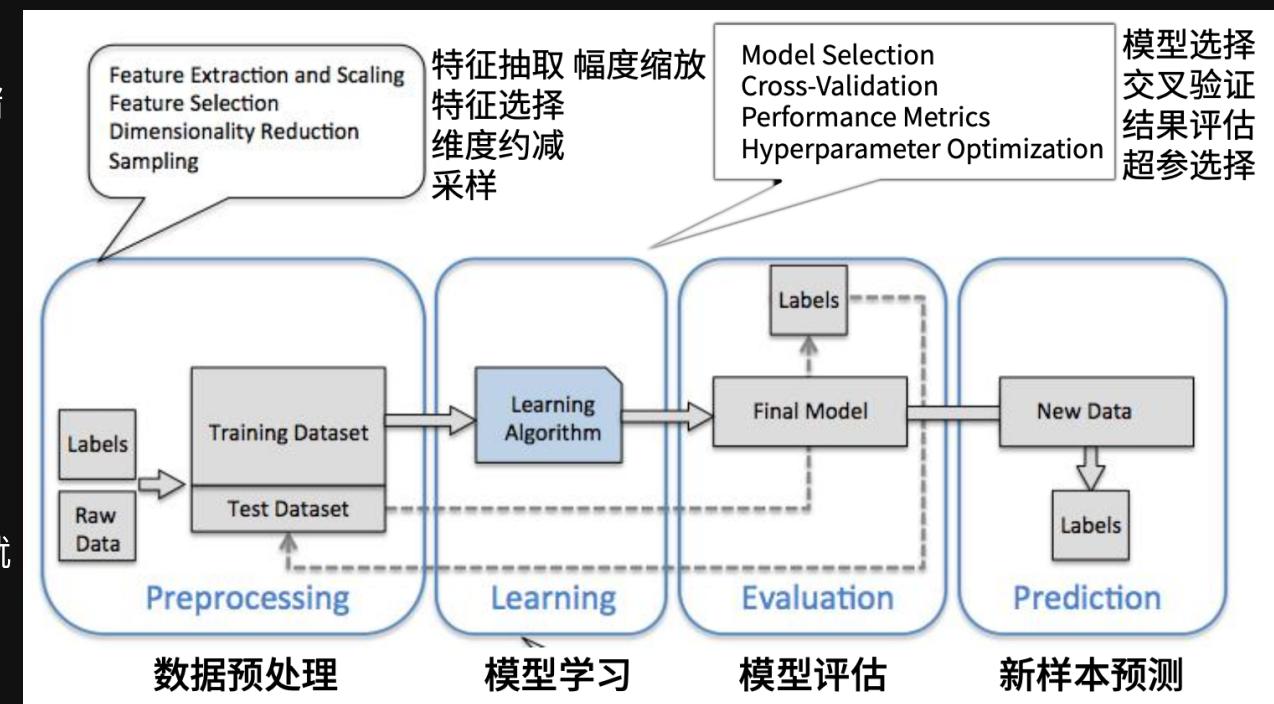


Figure 1. Schematic representation of the proposed event classification. The procedure is based on the reconstructed time-frequency map of candidates. ANNs are trained to produce an output number close to 1 for events are classified as belonging to the target distribution, and close to 0 otherwise. Our procedure does not constrain the output value to be limited to [0,1] and overflows and underflows are possible.



机器学习项目：开发应用程序的步骤

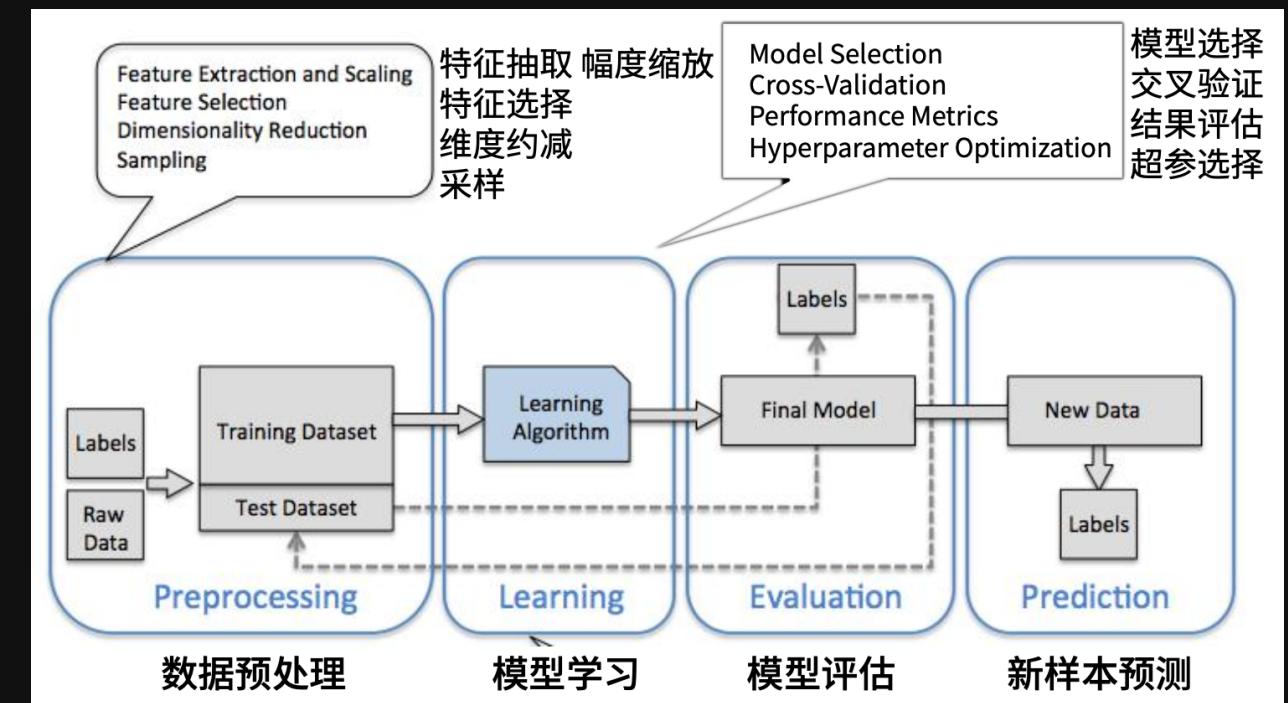
1. 收集数据（爬虫等）
2. 预处理、准备输入数据。得到数据之后，还必须确保数据格式符合要求，为机器学习算法准备特定的数据格式。
3. 分析输入数据，了解数据分部，数据可视化，观察数据基础分布，了解缺省情况与数据质量，确保数据集中没有垃圾数据。（人工+数据分析工具）
4. 训练算法。机器学习算法从这一步才真正开始学习。**根据算法的不同，第4步和第5步是机器学习算法的核心。**我们将前两步得到的格式化数据输入到算法，从中抽取知识或信息。这里得到的知识需要存储为计算机可以处理的格式，方便后续步骤使用。如果使用无监督学习算法，由于不存在目标变量值，故而也不需要训练算法，所有与算法相关的内容都集中在第5步。
5. 测试算法与调优。这一步将实际使用第4步机器学习得到的知识信息。为了评估算法，必须测试算法工作的效果。
 - 对于监督学习，必须已知用于评估算法的目标变量值；
 - 对于无监督学习，也必须用其他的评测手段来检验算法的成功率。
 - 无论哪种情形，如果不满意算法的输出结果，则可以回到第4步，改正并加以测试。问题常常会跟数据的收集和准备有关，这时你就必须跳回第1步重新开始。
6. 使用算法。将机器学习算法转换为应用程序，执行实际任务，以检验上述步骤是否可以在实际环境中正常工作。此时如果碰到新的数据问题，同样需要重复执行上述的步骤。





机器学习项目：baseline 流水线

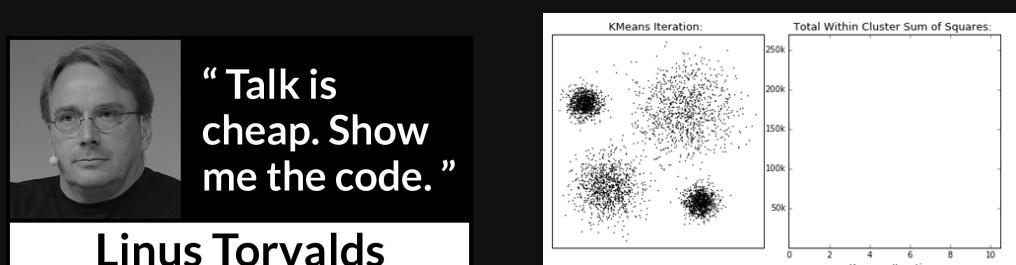
- 人工设计特征：在实际的应用中，特征往往比模型本身更重要
- 训练模式与测试模式
- 超参数调参
- 尽快搭建端到端的 baseline



Scikit-learn 机器学习库



- scikit-learn (sklearn) 是常用python工具库，涵盖绝大多数机器学习算法的实现
- 官方网址：<https://scikit-learn.org/stable/index.html>
- 最基本的sklearn应用教程：<https://scikit-learn.org/stable/tutorial/index.html>
- 导航页与算法指南：https://scikit-learn.org/stable/user_guide.html
- 详细API页面：<https://scikit-learn.org/stable/modules/classes.html>
 - 数据预处理：<https://scikit-learn.org/stable/modules/classes.html#module-sklearn.preprocessing>
 - 特征提取：https://scikit-learn.org/stable/modules/classes.html#module-sklearn.feature_extraction
 - 特征选择：https://scikit-learn.org/stable/modules/classes.html#module-sklearn.feature_selection
 - 各种模型：<https://scikit-learn.org/stable/modules/classes.html>
 - 模型调优与超参数选择：https://scikit-learn.org/stable/modules/classes.html#module-sklearn.model_selection
 - 模型融合与增强：<https://scikit-learn.org/stable/modules/classes.html#module-sklearn.ensemble>
 - 模型评估：<https://scikit-learn.org/stable/modules/classes.html#sklearn-metrics-metrics>
- sklearn的最基本的用法：5个基本函数
 - fit 拟合
 - transform 变换
 - fit_transform 拟合+变换
 - predict 预测
 - predict_proba 预测概率



Repo of the course: <https://github.com/iphysresearch/GWData-Bootcamp>

Homework

- 答题步骤：
 - 回答问题请保留每一步操作过程，请不要仅仅给出最后答案
 - 请养成代码注释的好习惯
 - 解题思路：
 - 为方便大家准确理解题目，在习题实战中有所收获，本文档提供了解题思路提示
 - 解题思路仅供参考，鼓励原创解题方法
 - 为督促同学们自己思考，解题思路内容设置为白色，必要时请从冒号后拖动鼠标查看
 - 所用数据
 - 请注意导入数据库后先查看和了解数据的基本性质，后面的问题不再一一提醒。
- 基础及拓展作业：
 - 一起来打怪之 Credit Scoring 练习：[homework_credit_scoring.ipynb](#)
 - 在 homework 分支上 PR。

• 赞助单位

- 中科曙光

