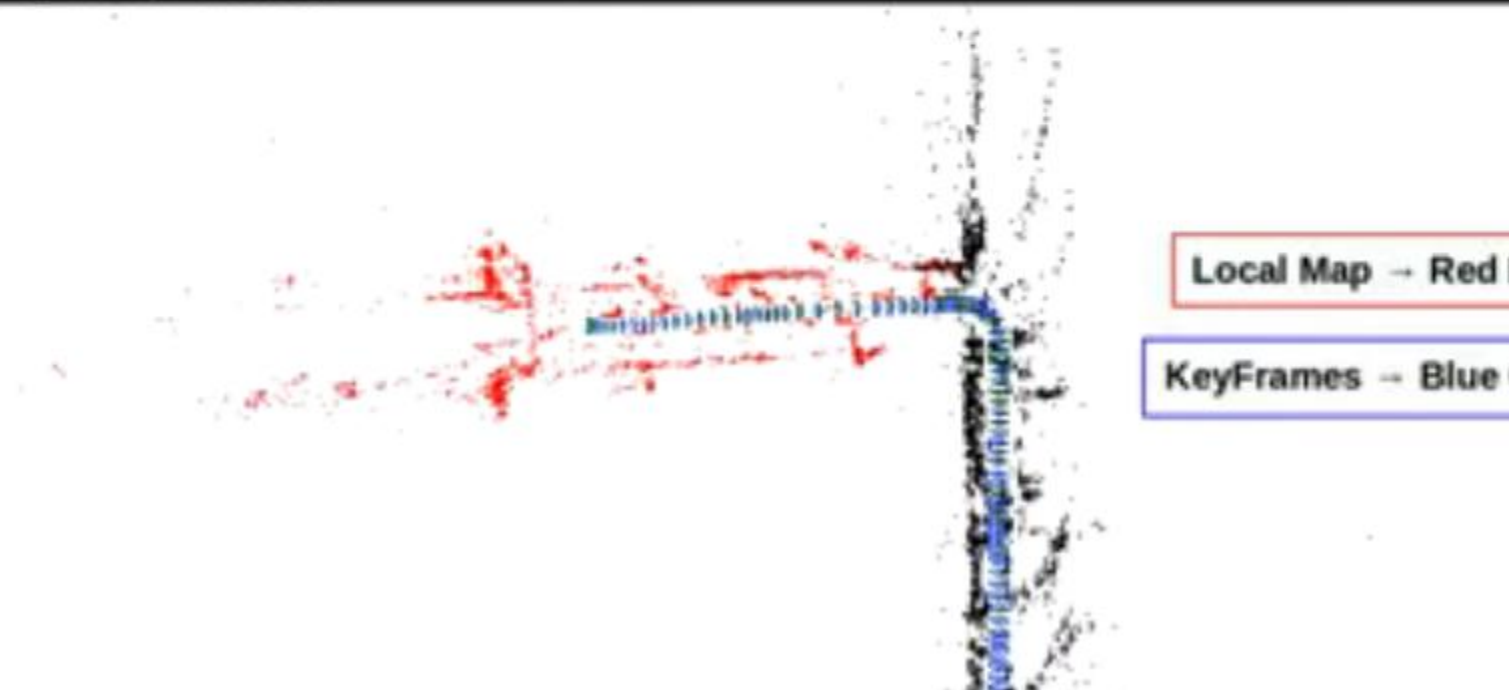
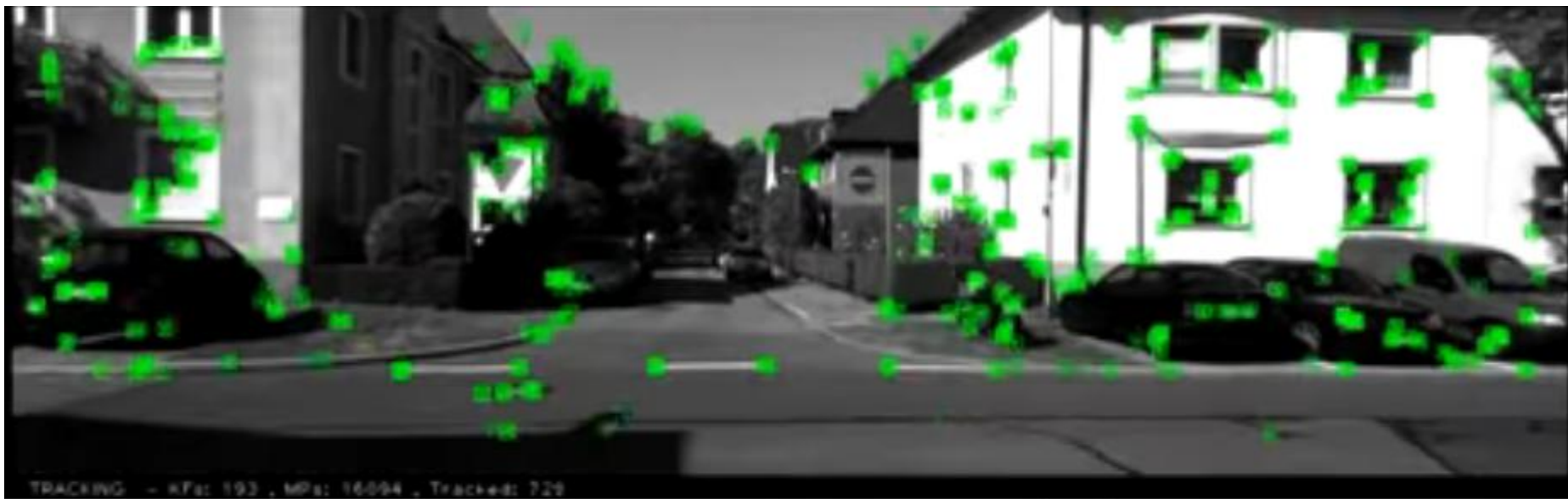


# 人工智能之机器学习

## 机器学习概述 (Machine Learning Overview)

上海育创网络科技有限公司

主讲人：赵翌臣





明星美图



1080P

56集全

因为遇见你DVD版第56集  
孙怡邓伦的锦绣情缘



1080P

46集全

因为遇见你第46集  
孙怡邓伦的锦绣情缘



1080P

67集全

楚乔传第67集  
乱世少女厮杀成长传奇



1080P

58集全

楚乔传DVD版第58集  
赵丽颖女奴逆袭



1

延禧攻略



2

芸汐传



3

开封府



4

猎毒人

## 猜你喜欢

换一换



1080P

共39集

那刻的怦然心动



1080P

更新至7集

爱情进化论



1080P

共50集

天地姻缘七仙女



独播

共12集

一不小心吃定你



独播

共56集

天泪传奇之凤凰无双





推荐

热点

社会

娱乐

科

+

2014.03.14 星期五 今天

云南副省长沈培平被查  
普洱市民放鞭炮庆祝



新浪网 评论187 刚刚 +

印孕妇产下双头女婴共用一个身体



环球网 评论837 刚刚 +

专家分析称马航失联客机或进入“航空黑洞”

环球网 评论1507 刚刚 +

马民航局官员指责中国违反外交礼仪

手机凤凰网 评论1062 刚刚 +

卓越品质 4款自主品牌15万元内的SUV导购

中国移动 下午3:43 87%

要闻 视频 上海 财经 娱乐 +

6天蒸发3838.38亿，合并前北车市值约为3306.52亿。

1434评

控杠杆见效!场外配资渐离股市

依然有配资公司顶风逆行招揽客户。杠杆比最高可达1:10。

257评 独家

员工造谣! 国君押空单大赚

在上市前夕，国君成功押中A股16日的大跌，大赚一笔。

133评

广告标题

没钱王子=杜杜!

到底钱重要? 还是外表重要?

推广

缩略图



简介

中国再贷给委内瑞拉50亿美元

11:01 4G

朋友圈

我: 告辞 🙏

昨天



李记煎饼厉害了，都在微信上投广告了，看来卖煎饼真的是致富之路啊 😂

李记煎饼店

精挑细选，只为食材的新鲜。精挑细选，只为您舌尖上的享受，期待您的服务。



了解更多

上海 · 李记煎饼(二路店)

60分钟前

昨天

❤️ 吴

美: 广告精准投放 服气 😏



别走太多步了! 危险了!



●你还在晚上散步吗●现在知道还不晚!

# 机器学习前景

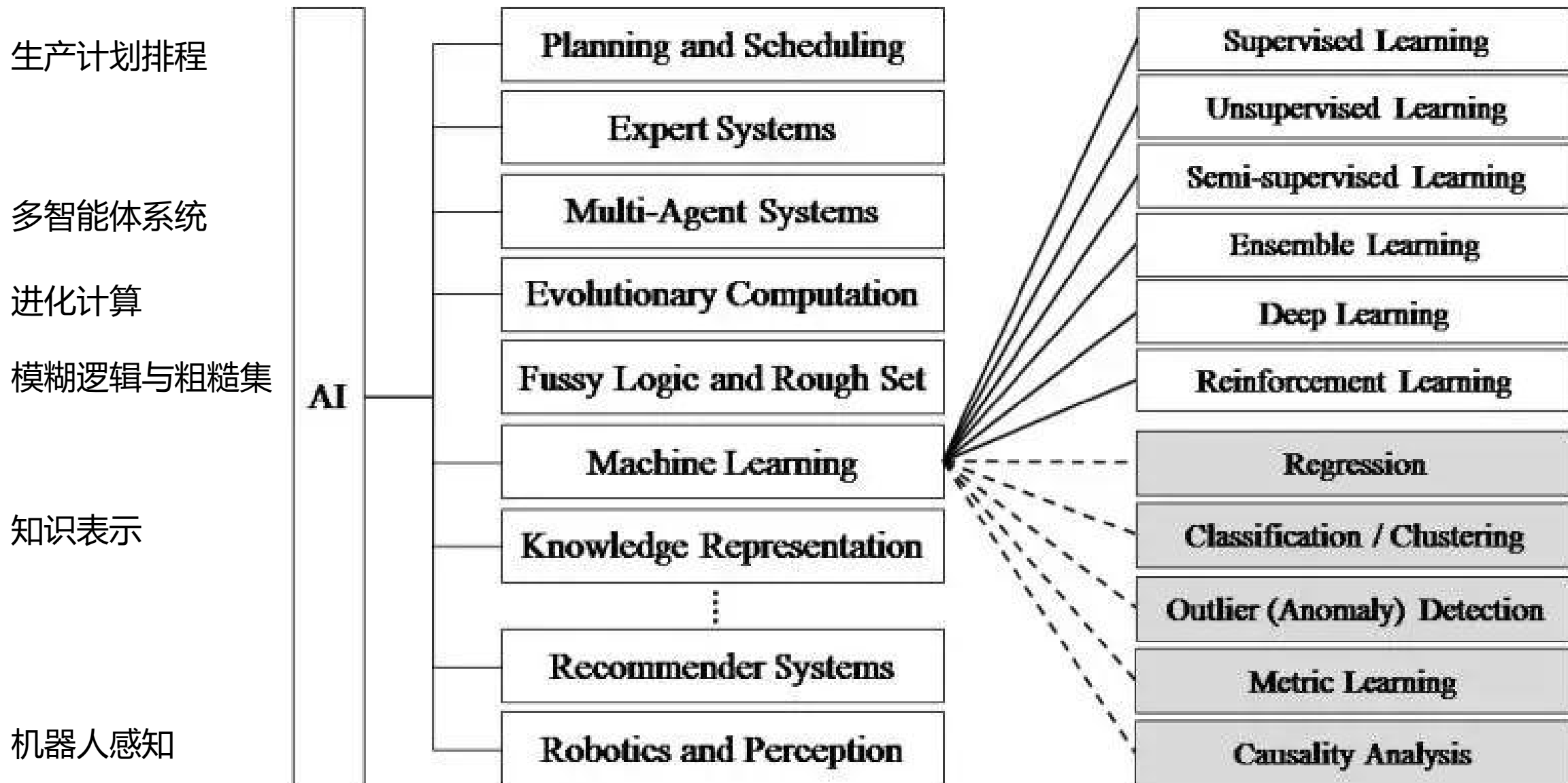
- 国家战略

- 习近平总书记指出：**大数据**是信息化发展的新阶段，要推动大数据技术产业创新发展，运用大数据提升国家治理现代化水平，促进保障和改善民生。
- 国务院近日印发《新一代人工智能发展规划》，按照规划，“**人工智能**”将“无时不有、无处不在”，到2030年中国要成为世界主要“人工智能”创新中心。

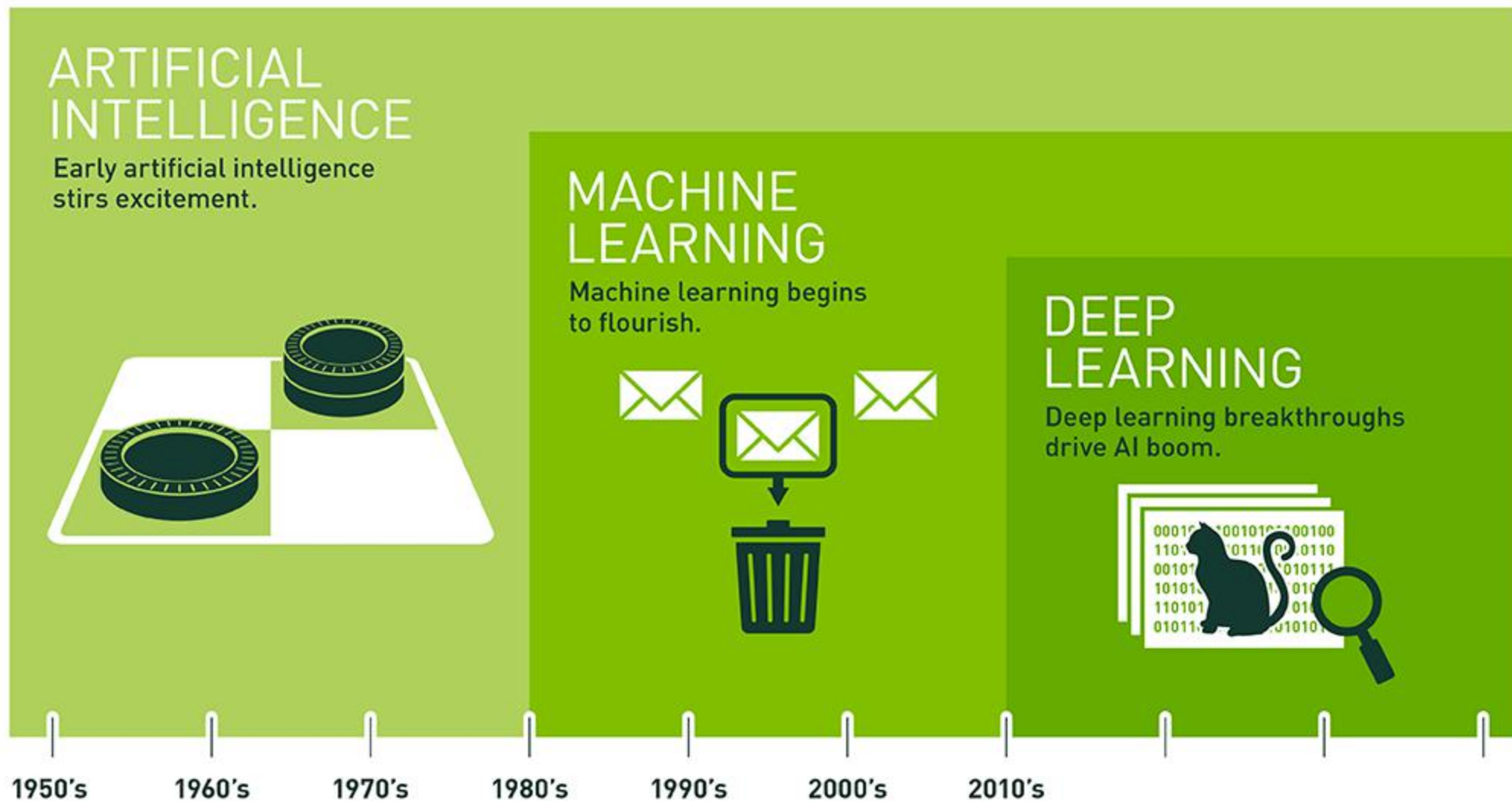


中国大数据产业峰会暨中国电子商务创新发展峰会

# 人工智能、机器学习和深度学习的关系



# 人工智能、机器学习和深度学习的关系



## 数据挖掘、数据分析、机器学习

- **数据分析：** 数据分析是指用适当的统计分析方法对收集的大量数据进行分析，并提取有用的信息，以及形成结论，从而对数据进行详细的研究和概括**过程**。在实际工作中，数据分析可帮助人们做出判断；数据分析一般而言可以分为统计分析、探索性数据分析和验证性数据分析三大类。
- **数据挖掘：** 一般指从大量的数据中通过算法搜索隐藏于其中的信息的**过程**。通常通过统计、检索、机器学习、模式匹配等诸多方法来实现这个过程。
- **机器学习：** 是数据分析和数据挖掘的一种比较常用、比较好的**手段**。



# 大数据时代

- 数据量大 (Volume)
  - 巨量资料，传统数据处理应用软件不足以处理这种量级的数据集，大数据的起始计量单位至少是P（1000个T）、E（100万个T）或Z（10亿个T）
- 类型繁多 (Variety)
  - 包括网络日志、音频、视频、图片、地理位置信息等等，多类型的数据对数据的处理能力提出了更高的要求。
- 价值密度低 (Value)
  - 信息海量，就意味着价值密度较低，如何通过强大的机器算法更迅速地完成任务的价值“提纯”，是大数据时代需要思考的问题。
- 速度快、时效高 (Velocity)
  - 要求处理速度快，时效性要求高。这是大数据区别于传统数据挖掘最显著的特征



## 大数据与机器学习结合

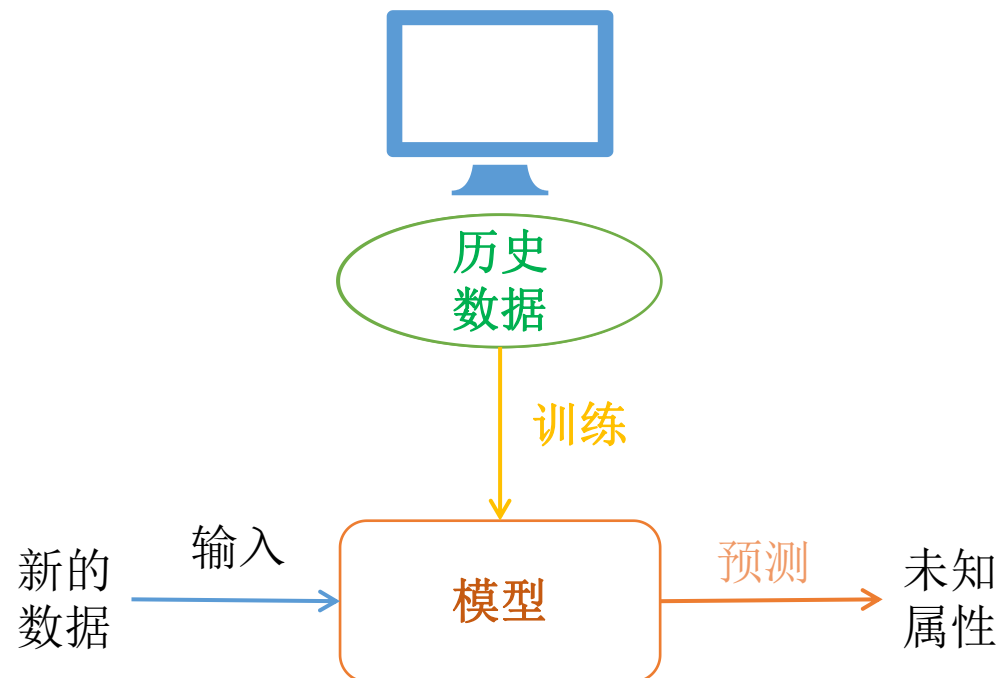
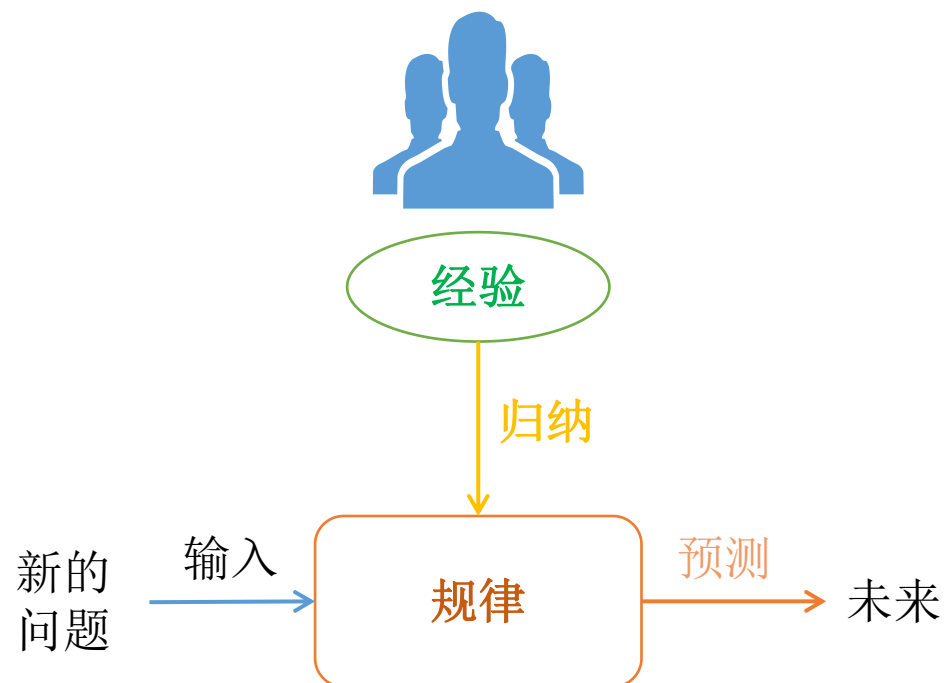
- 用数千个特征和数十亿个交易来构建信用卡欺诈检测模型（风险评估、风控）
- 向数百万用户智能地推荐数百万产品他们可能喜欢的产品（个性化推荐）
- 使用上亿级别的用户对广告的点击的数据集，做一个广告CTR预估模型，进行广告投放
- 从千上万个个人类基因的相关数据以发现致病基因



## 方向选择

- 大数据机器学习工程师 VS 大数据工程师
- 大数据机器学习工程师 VS 算法工程师
- 大数据机器学习工程师 VS 深度学习工程师

# 机器学习理性认识 (what)





## 机器学习定义 (what)

- (统计) 机器学习是关于计算机基于数据构建概率统计模型并运用模型对未知数据进行预测和分析的一门学科。
  - 研究对象：数据
  - 目标：对未知数据进行预测分析
  - 理论：概率论、统计学、信息论、计算理论、最优化理论等
  - 形式：监督学习、非监督学习、半监督学习和强化学习

# 为什么要用机器学习 (why)

- 传统编程
  - 丢给计算机一串明确的用代码描述的指令 (if...else...)，让它按照指令一步一步求得结果
  - 显著式编程，基于逻辑
- 机器学习
  - 以数据为驱动，引入统计学的思想，让计算机来做各种各样的工作
  - 非显著式编程model，基于数据



# 为什么要用机器学习 (why)

- 传统编程
- 机器学习

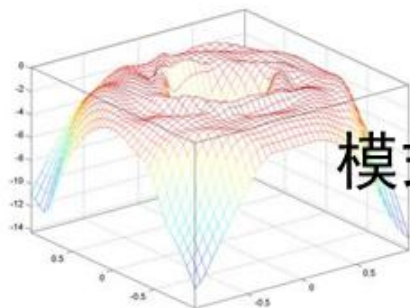


α-go击败李世石



DOTA2 中OpenAI打败Dendi

# 机器学习之商业场景 (where)



模式识别

计算机视觉



数据挖掘



机器学习

语音识别



统计学习



自然语言处理





# 不同人眼中的机器学习



What society thinks I do



What my friend thinks I do



What my parents thinks I do

SVM:

$$\min_{\mathbf{w}, \xi, b} \left\{ \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \xi_i \right\}$$

subject to (for any  $i = 1, \dots, n$ )

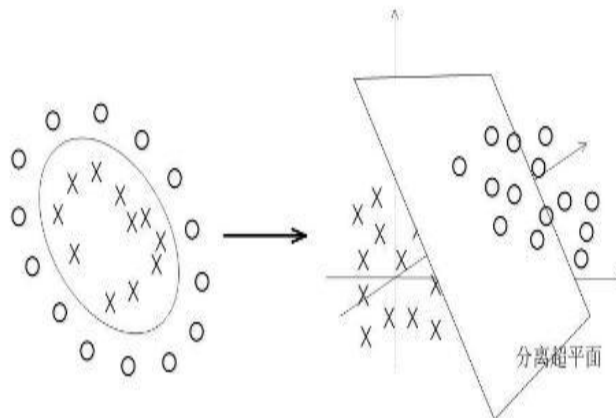
$$y_i(\mathbf{w} \cdot \mathbf{x}_i - b) \geq 1 - \xi_i, \quad \xi_i \geq 0$$

LR:

$$\min_{\theta} \sum_{i=1}^M -\log p(y^{(i)} | \mathbf{x}^{(i)}; \theta) + \beta \|\theta\|_1.$$

$$P_w(y|x) = \frac{\exp w^T \Phi(x, y)}{\sum_{y' \in \text{GEN}(x)} \exp w^T \Phi(x, y')}.$$

What other programmers thinks I do



What I thinks I do

```
In [1]: from sklearn import svm
```

```
In [ ]:
```

What I really do

## 基础概念

T是数据集

$(x_1, y_1)$ 是一个Labeled point, 带有标记的数据点,  
 $x_1$ 代表第一个样本的**特征向量/输入**  
 $y_1$ 代表第一个样本的**标签/标记/输出**

其中 $x_1$ 是一个向量, 包含多个**特征/属性**


$$T = \{(x_1, y_1), \dots, (x_N, y_N)\}$$

$$x_1 = (x_1^{(1)}, \dots, x_1^{(n)})^T \quad y_1$$

$$T = \{(\text{房子}_1, 3.5_1), \dots, (\text{房子}_N, 2.7_N)\}$$

$$\text{房子}_1 = (\text{面积}_1^{(1)}, \text{楼层}_1^{(2)}, \dots, \text{房间数}_1^{(n)})^T \quad \text{价格}_1$$

## 基础概念

- 

训练集 (0.7)      验证集 (0.2)      测试集 (0.1)
- 训练集：用于模型训练的数据集
- 验证集：用于模型评估的数据集
- 测试集：模拟生产环境的数据集
- 用同一个验证集做评估，有可能出现模型对验证集上表现的比较好的情况，要想真正评估这个最佳模型在将来的样本上的表现，这也就是需要把第三个子集即测试集保留在一边的原因（小风险）



- 企业中往往去掉了验证集：因为测试集就是没有被“污染”的数据，综合考虑训练误差和测试误差，同样能起到模型评估的作用；这样划分数据集的优点是降低工程实施的时间成本。



# 机器学习四种形式



监督学习



非监督学习

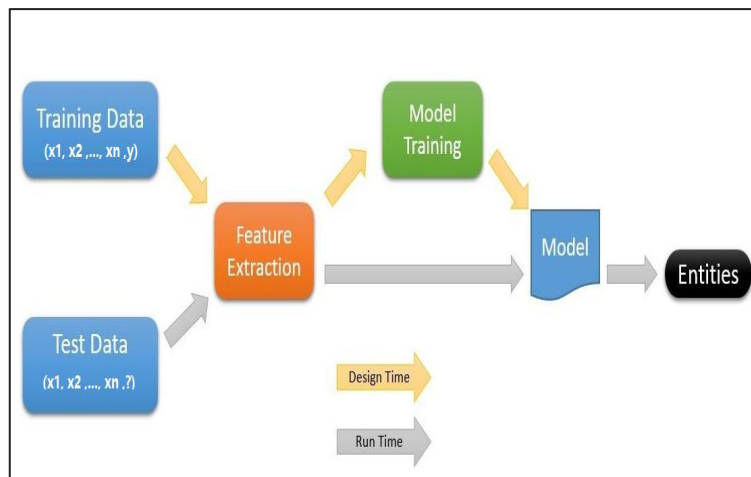


半监督学习

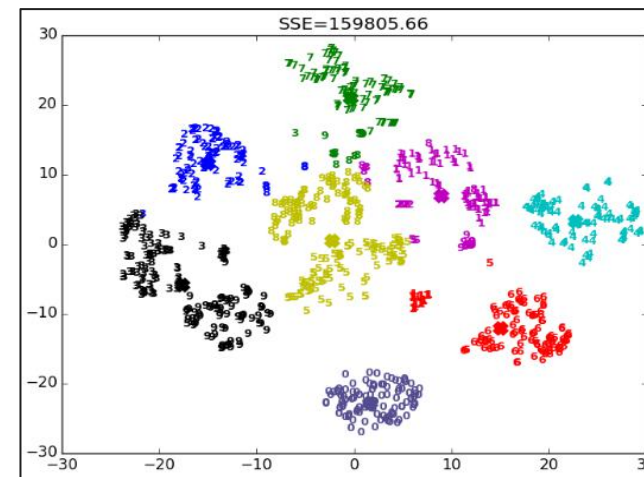


强化学习

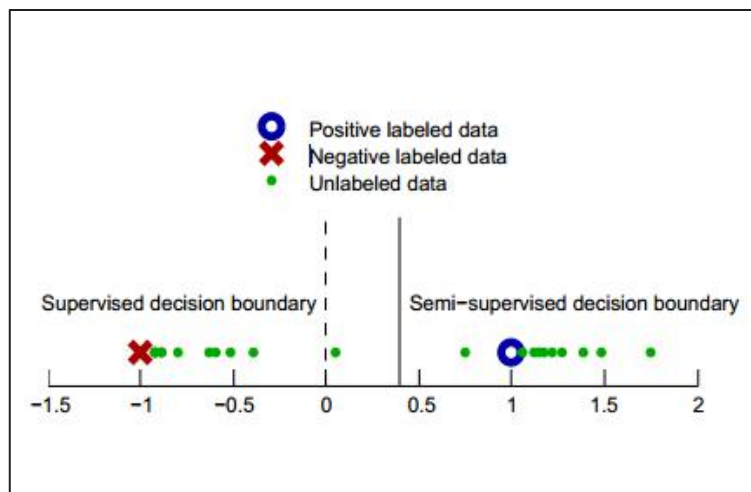
# 机器学习四种形式



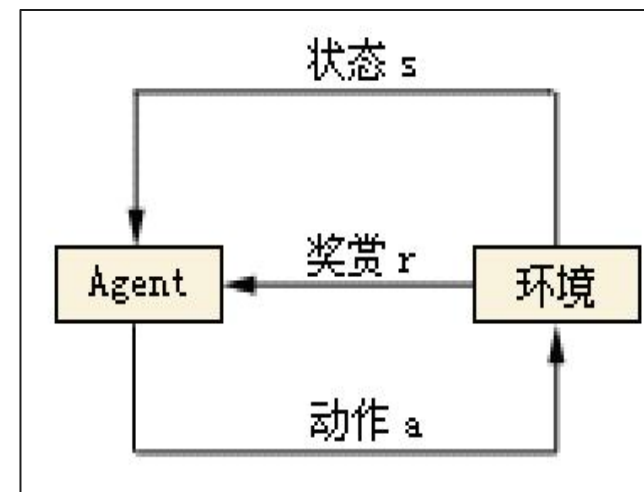
监督学习



非监督学习



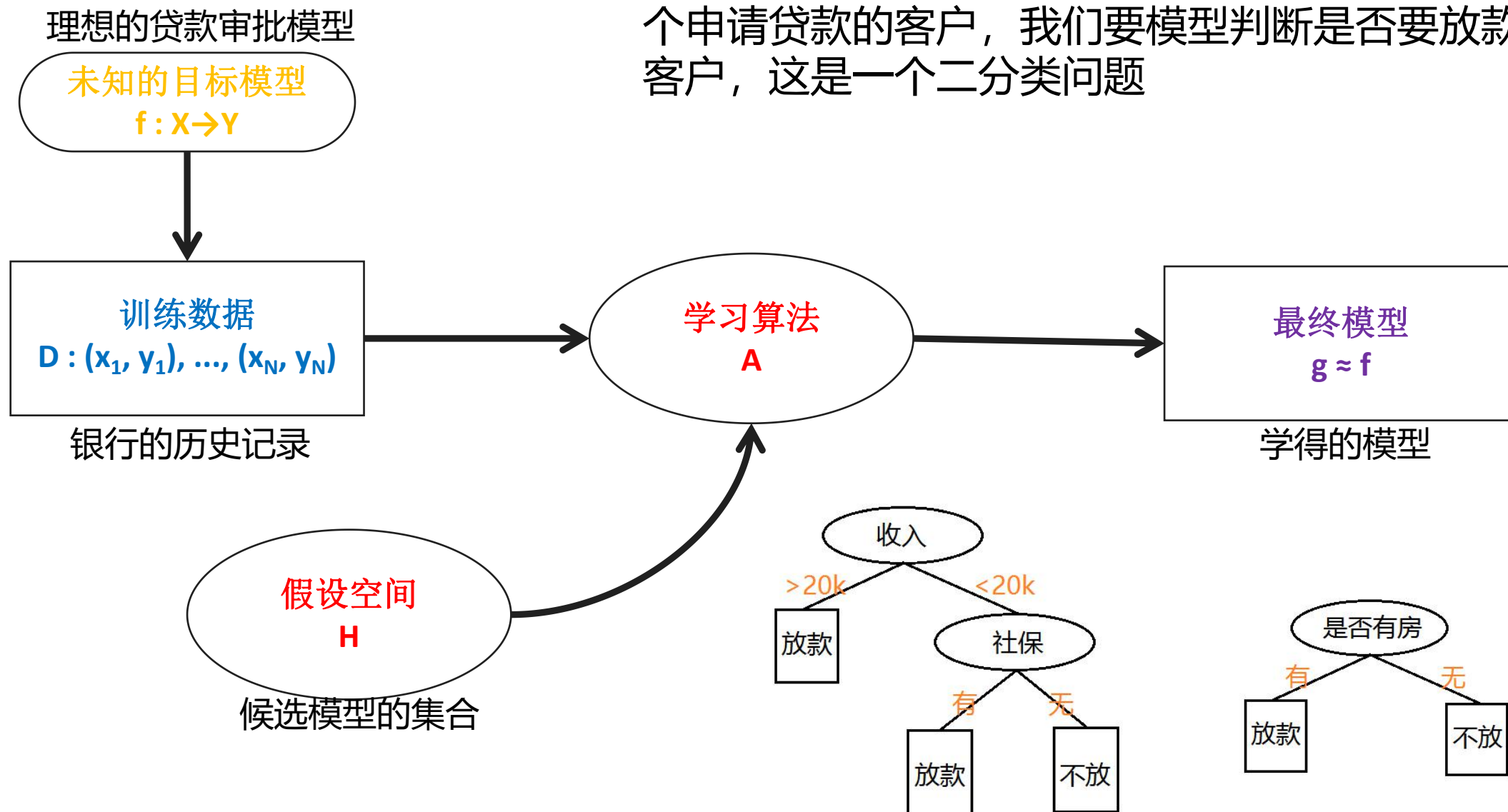
半监督学习



强化学习

# 贷款审批流程

比如我们要帮助银行做一个贷款审批的模型，即来了一个申请贷款的客户，我们要模型判断是否要放款给这个客户，这是一个二分类问题



Problem Type	Supported Methods
Binary Classification	linear SVMs, logistic regression, decision trees, random forests, gradient-boosted trees, naive Bayes
Multiclass Classification	logistic regression, decision trees, random forests, naive Bayes
Regression	linear least squares, Lasso, ridge regression, decision trees, random forests, gradient-boosted trees, isotonic regression

<http://spark.apache.org/docs/1.6.1/mllib-classification-regression.html>

[http://scikit-learn.org/0.18/user\\_guide.html](http://scikit-learn.org/0.18/user_guide.html)

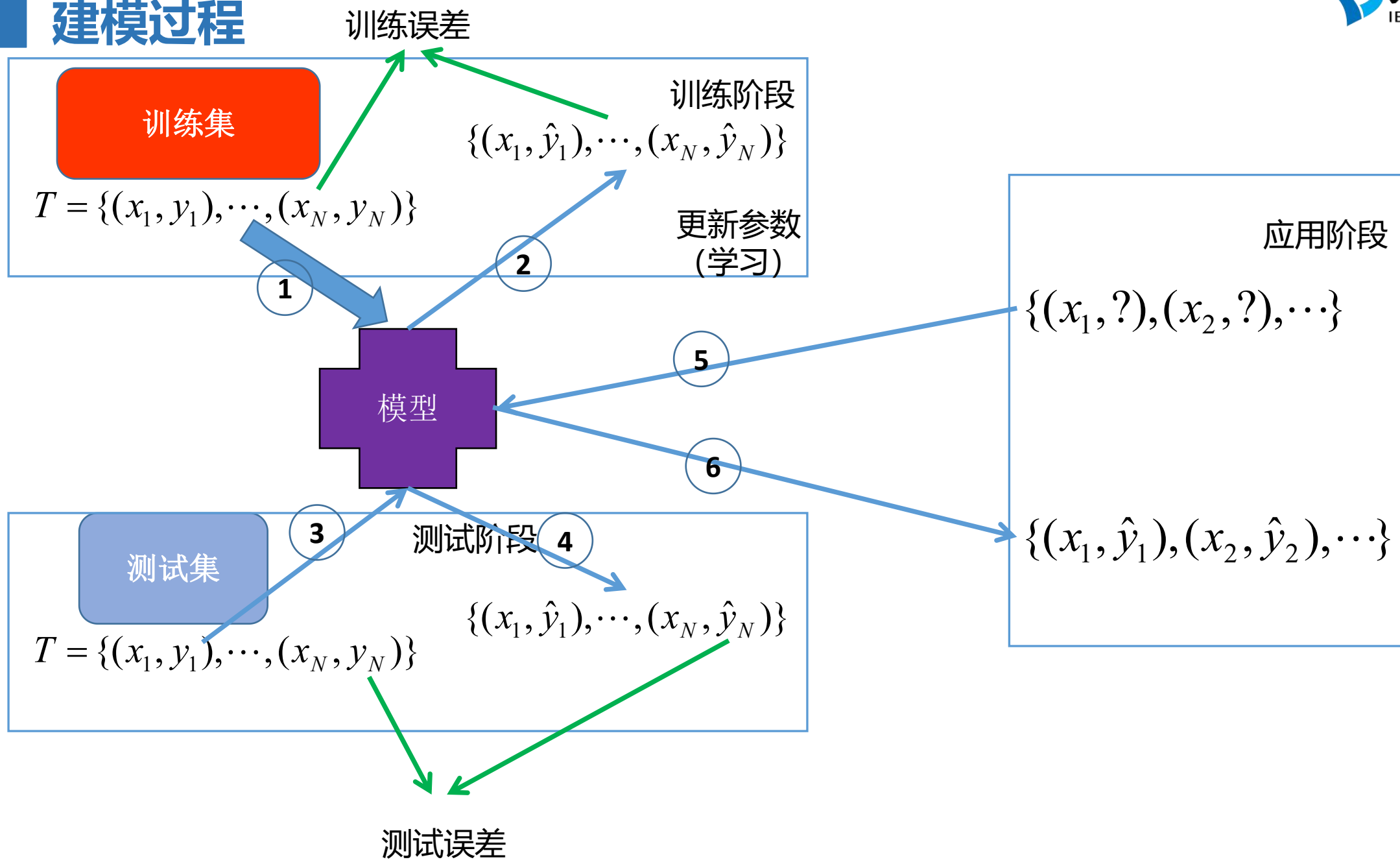


## 常用非监督学习模型

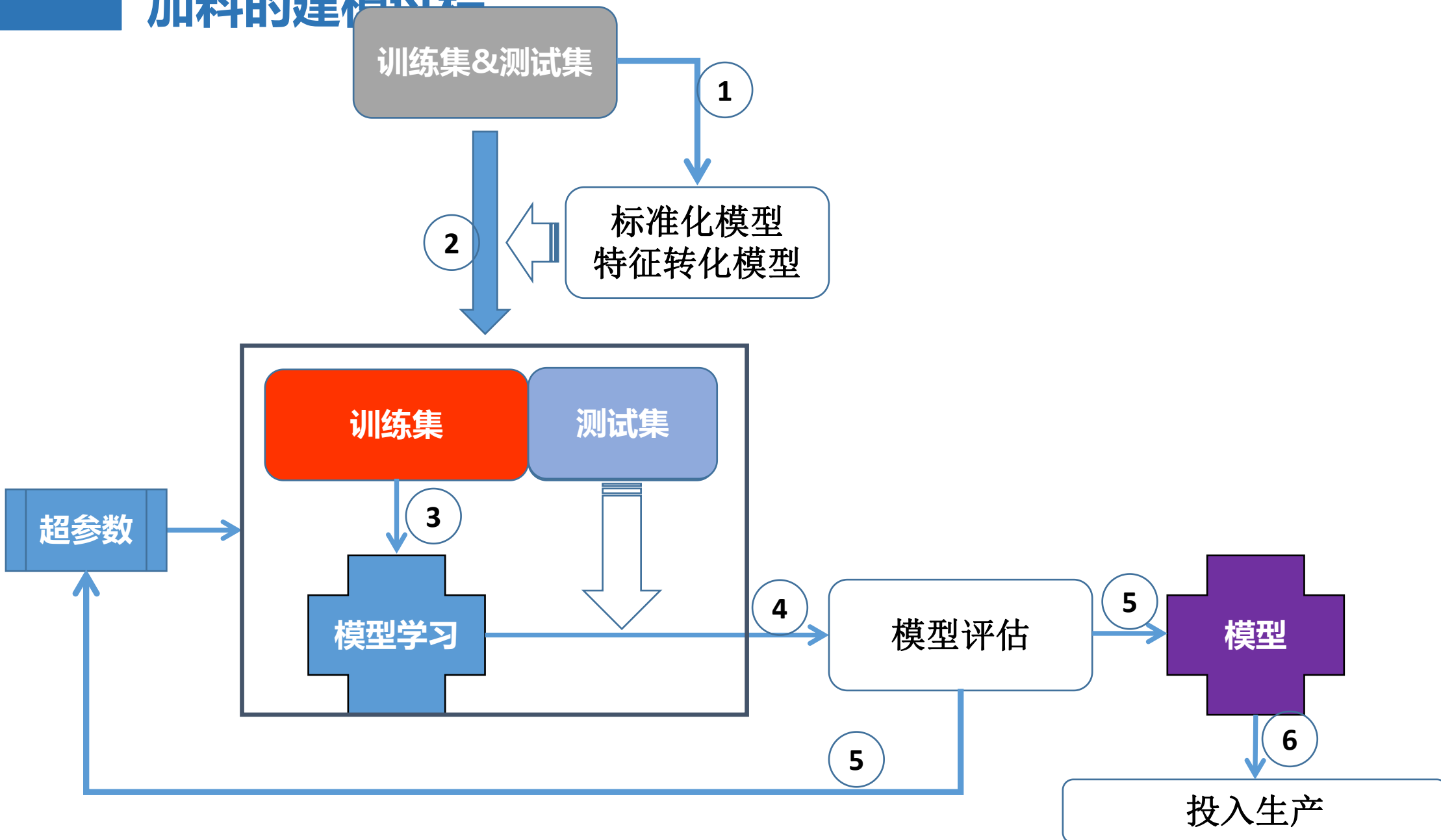
- K-means k均值聚类
- Gaussian mixture 高斯混合模型
- LDA (Latent Dirichlet allocation)

<http://spark.apache.org/docs/1.6.1/mllib-clustering.html>

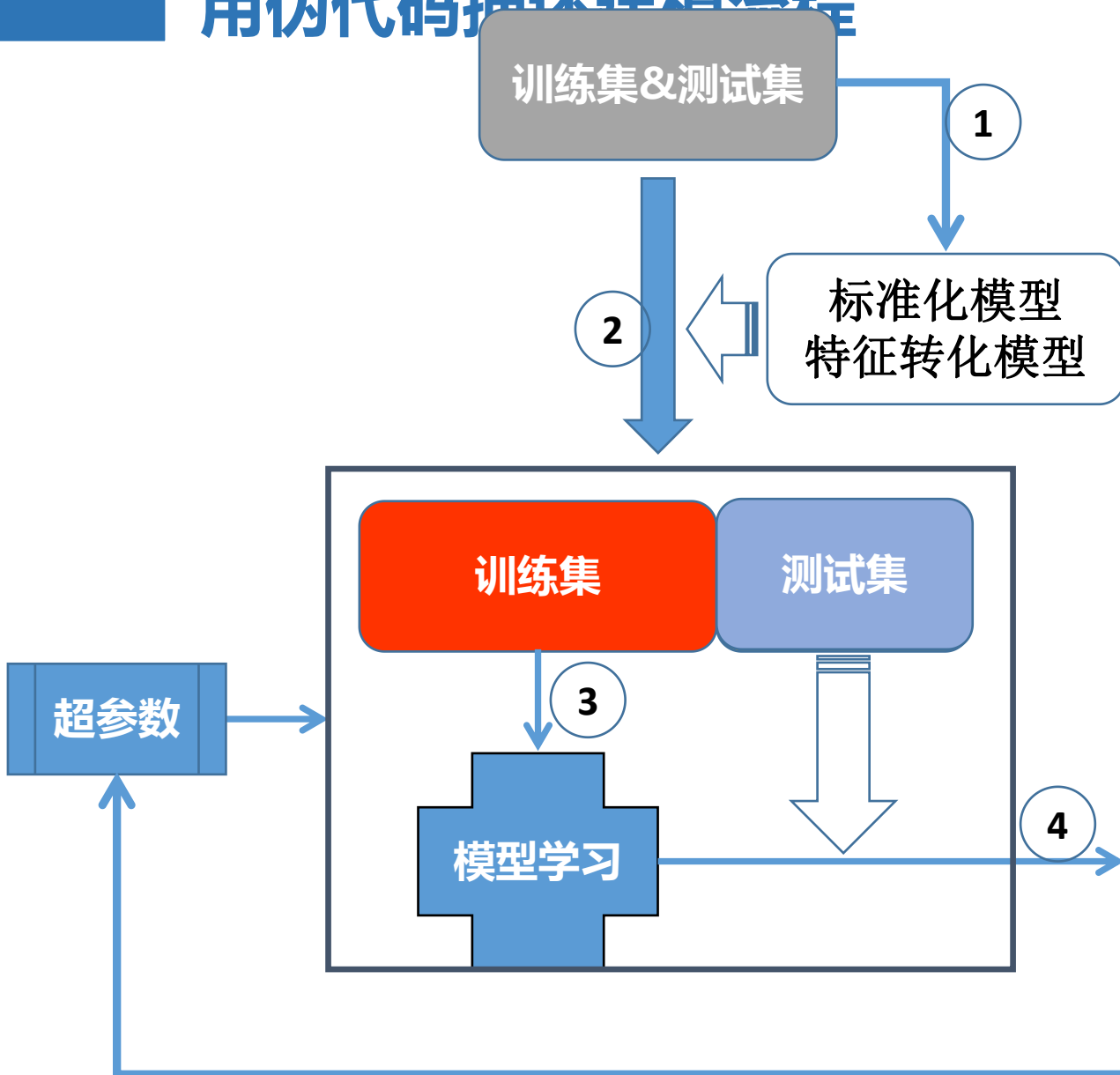
# 建模过程



# 加料的建模过程



# 用伪代码描述建模流程



for:

超参数调节

for:

模型学习，更新内部参数，降低训练误差

if 训练误差达到最小 / 达到一定循环次数

break

模型评估 ==> 测试误差

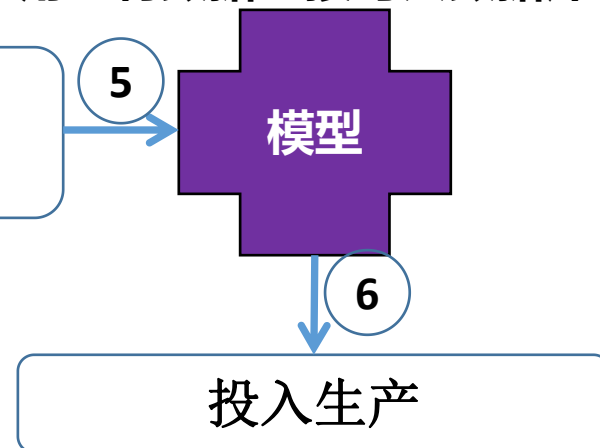
if 测试误差能容忍

break

应用

a. 实时应用：模型推到server端

b. 离线应用：将数据直接写入数据库或redis



## 模型部署细节

- 将原始数据解析成特征：纯工程
- 将特征转换成其他特征：算法开始了
- 构建模型
- 评价模型：数学上、快速的分析当前模型好坏
- 模型超参数调优，重建
- 部署模型
- 业务评估（数据平台指标，常见：UV、VV、转化率等等）：长远一点的评估方式
- 实时模型更新
- 根据模型实时进行查询

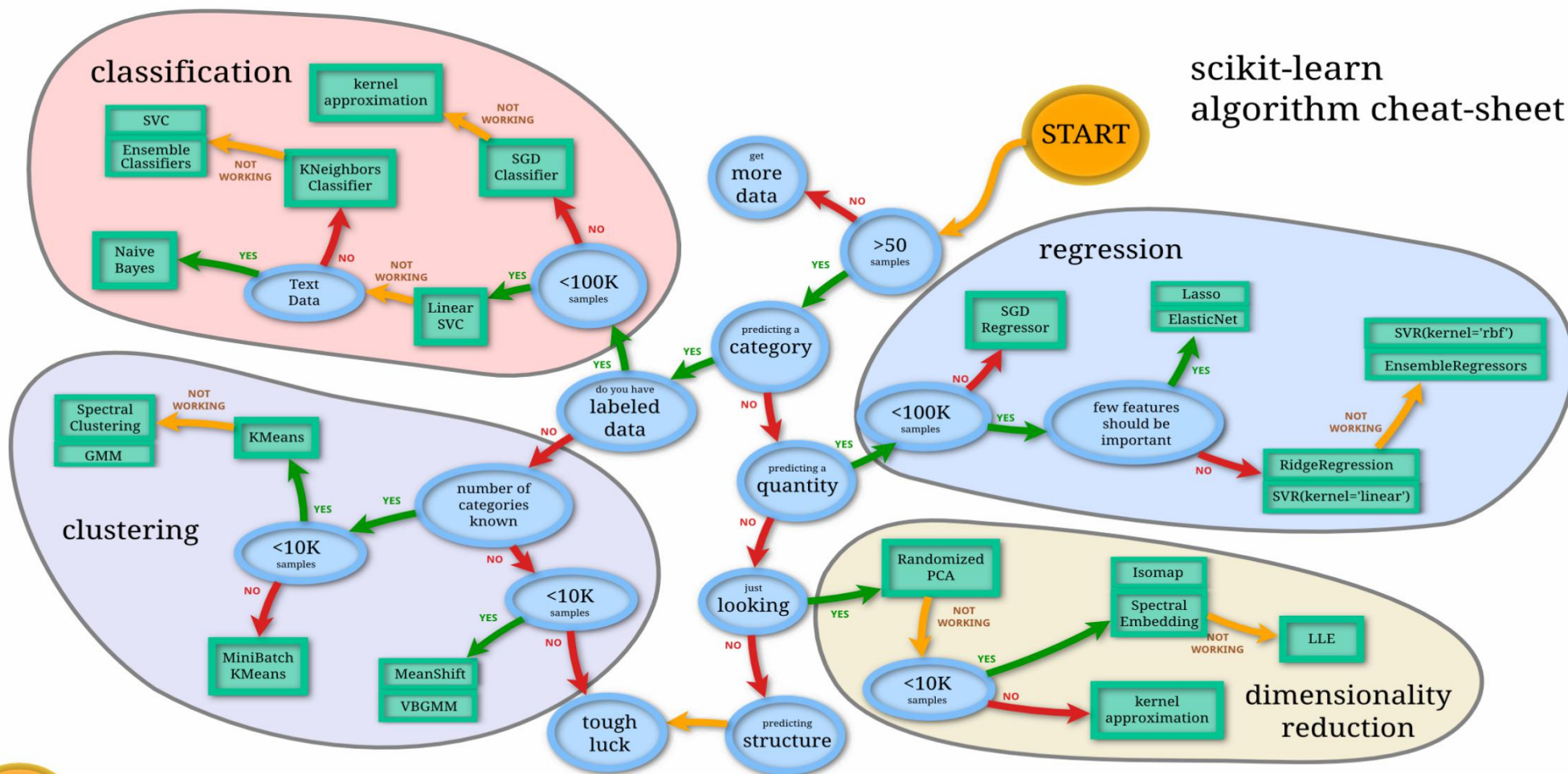


## 机器学习常用框架

- scikit-learn(Python)
  - <http://scikit-learn.org/stable/>
- Mahout(Hadoop生态圈基于MapReduce)
  - <http://mahout.apache.org/>
- Spark MLlib (Spark)
  - <http://spark.apache.org/>



# sklearn思维导图



# 机器学习算法Top10

算法名称	算法描述
C4.5	分类决策树算法，决策树的核心算法，ID3算法的改进算法。
CART	分类与回归树(Classification and Regression Trees)
kNN	K近邻分类算法；如果一个样本在特征空间中的k个最相似的样本中大多数属于某一个类别，那么该样本也属于该类别
NaiveBayes	贝叶斯分类模型；该模型比较适合属性相关性比较小的时候，如果属性相关性比较大的时候，决策树模型比贝叶斯分类模型效果好(原因：贝叶斯模型假设属性之间是互不影响的)
SVM	支持向量机，一种有监督学习的统计学习方法，广泛应用于统计分类和回归分析中。
EM	最大期望算法，常用于机器学习和计算机视觉中的数据集聚领域
Apriori	关联规则挖掘算法
K-Means	聚类算法，功能是将n个对象根据属性特征分为k个分割( $k < n$ )；属于无监督学习
PageRank	Google搜索重要算法之一
AdaBoost	迭代算法；利用多个分类器进行数据分类

# 数据收集与存储

- 数据来源：
  - 用户访问行为数据
  - 业务数据
  - 外部第三方数据
- 数据存储：
  - 需要存储的数据：原始数据、预处理后数据、模型结果
  - 存储设施：mysql、HDFS、HBase、Kafka、Redis等
- 数据收集方式：
  - Flume & Kafka

## 机器学习可用公开数据集

- 在实际工作中，我们可以使用业务数据进行机器学习开发，但是在学习过程中，没有业务数据，此时可以使用公开的数据集进行开发，常用数据集如下：
  - <http://archive.ics.uci.edu/ml/datasets.html>
  - <https://aws.amazon.com/cn/public-datasets/>
  - <https://www.kaggle.com/competitions>
  - <http://www.kdnuggets.com/datasets/index.html>
  - [http://www.sogou.com/labs/resource/list\\_pingce.php](http://www.sogou.com/labs/resource/list_pingce.php)
  - <https://tianchi.aliyun.com/datalab/index.htm>
  - <http://www.pkbigdata.com/common/cmptIndex.html>



