



Day1

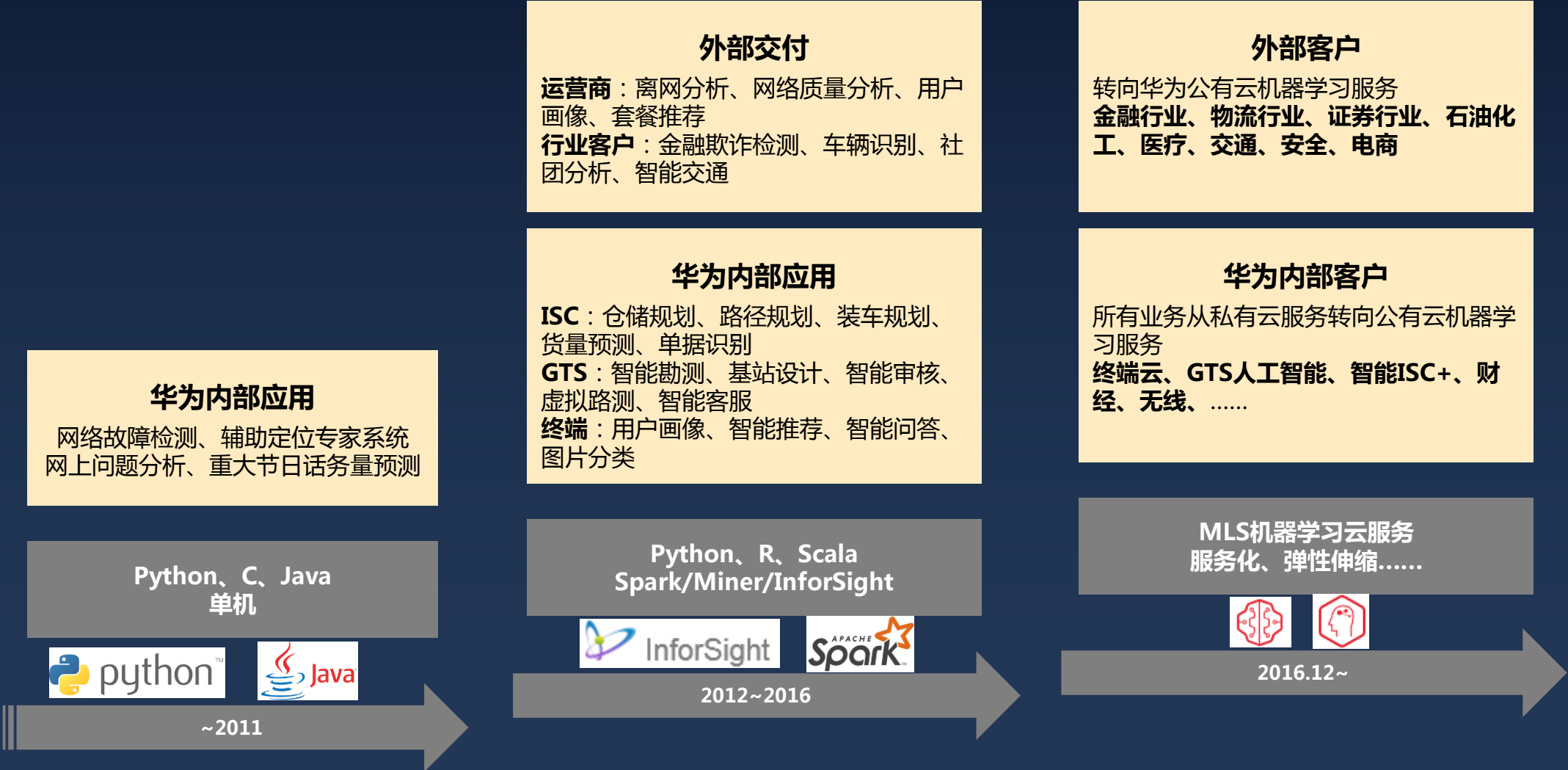
探索数据的奥义



HUAWEI TECHNOLOGIES CO., LTD.

www.huawei.com

华为机器学习的发展



什么是机器学习

机器学习这门学科所关注的问题是：计算机程序如何随着经验积累自动提高性能。

对于某类任务 T 和性能度量 P ，如果一个计算机程序在 T 上以 P 衡量的性能随着经验 E 而自我完善，那么我们称这个计算机程序在从经验 E 学习。

-- Tom Mitchell, Machine Learning

模式识别起源于工程学，而机器学习产生于计算机科学。然而这些领域可以看做成是同一领域的两个方面。

-- Bishop, Pattern Recognition and Machine Learning

机器学习是一类从数据中自动发现模式，并基于发现的模式预测未来数据或者在不确定条件下执行某类决策的方法。

-- Murphy, Machine Learning: A Probabilistic Perspective

行业通用的机器学习类型

按学习方式分为三大类

	说明	解决问题
监督学习 Supervised learning	从给定的训练数据集（历史数据）中学习出一个函数，当新的数据到来时，可以根据这个函数预测结果。监督学习的训练集需要包括输入和输出，也可以说是特征和目标/Label。训练集中的目标是由人标注的。	分类 回归
无监督学习 Unsupervised learning	与监督学习相比，输入的数据没有人为标注的结果，模型需要对数据的结构和数值进行归纳。比如根据用户的基本信息把所有用户划分为不同的用户群，再对不同人群采取不同的销售策略	聚类（分群）
强化学习 Reinforcement learning	输入数据可以刺激模型并且使模型做出反应。反馈不仅从监督学习的学习过程中得到，还从环境中的奖励或惩罚中得到。	机器人 Alpha GO

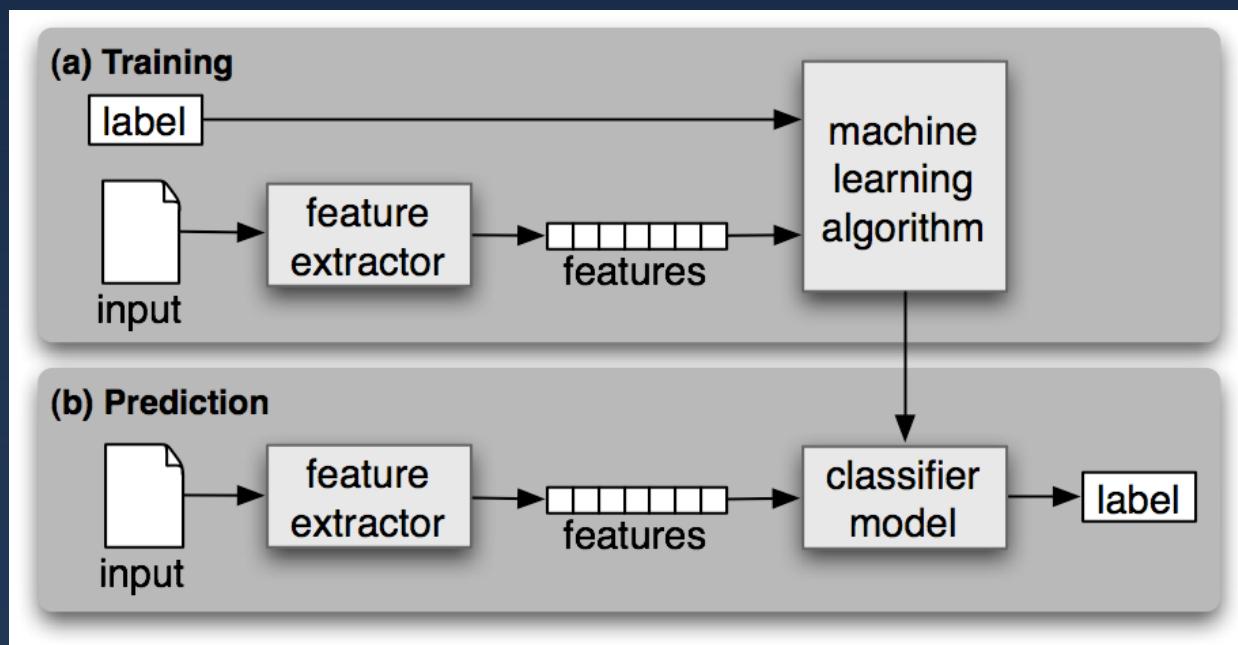
监督学习

两类监督学习

分类：label是离散的，例如判断是否离网，是否是高消费人群

回归：label是连续的，例如预测年龄，预测收入

监督学习的基本流程



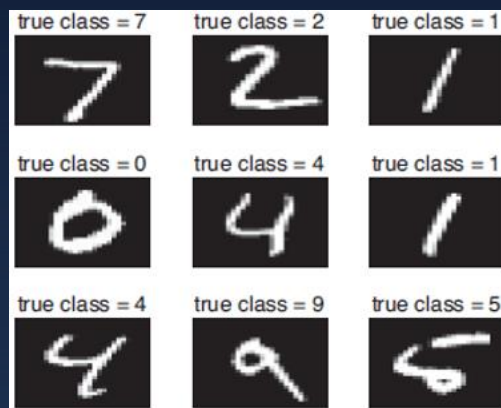
监督学习：分类

分类问题实例：

文档分类、垃圾邮件过滤
图像分类、手写体识别
人脸侦测与识别
离网预测
.....

典型分类算法

Decision Tree
Native Bayes
Random Forest
Logistic Regression
Support Vector Machine
.....



分类模型的评价

准确率 (accuracy)
精确率 (查准率 , precision)
召回率 (查全率 , recall)
F1 score
ROC曲线
AUC曲线
混淆矩阵 (Confusion matrix)

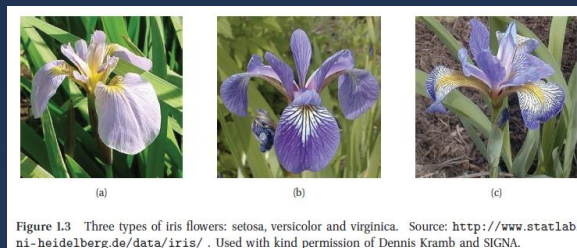


Figure 1.3 Three types of iris flowers: setosa, versicolor and virginica. Source: <http://www.statlab.uni-heidelberg.de/data/iris/> . Used with kind permission of Dennis Kramb and SIGNA.

监督学习：回归

典型回归方法

线性回归

支持向量回归

分类回归树（基于平方误差）

.....

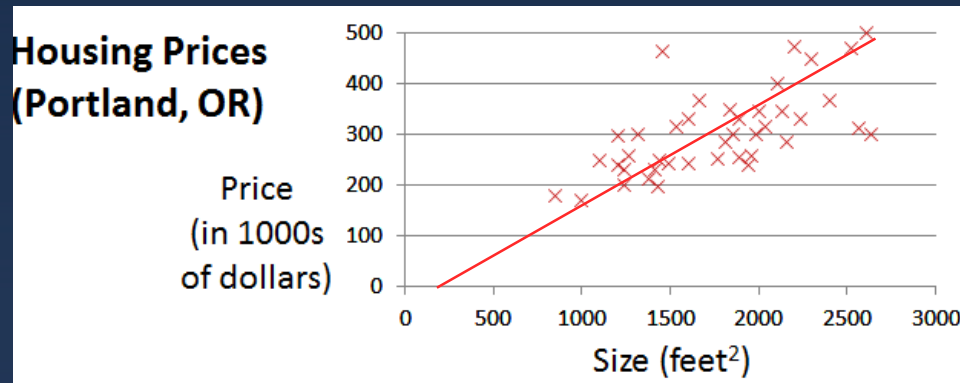
回归问题的评价

平均绝对误差（mean absolute error, MAE）

均方根误差（root mean squared error, RMSE）

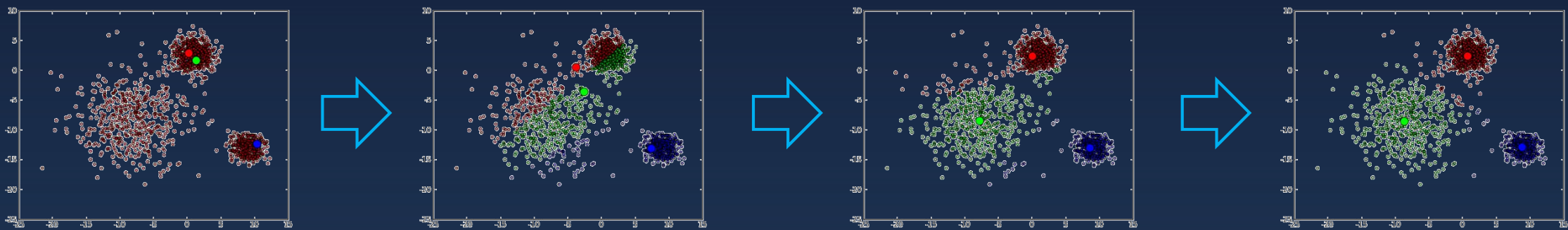
确定性系数（coefficient of determination, R^2 ）

Size in feet ² (x)	Price (\$) in 1000's (y)
2104	460
1416	232
1534	315
852	178
...	...



无监督学习：聚类

典型聚类算法：Kmeans (K均值)

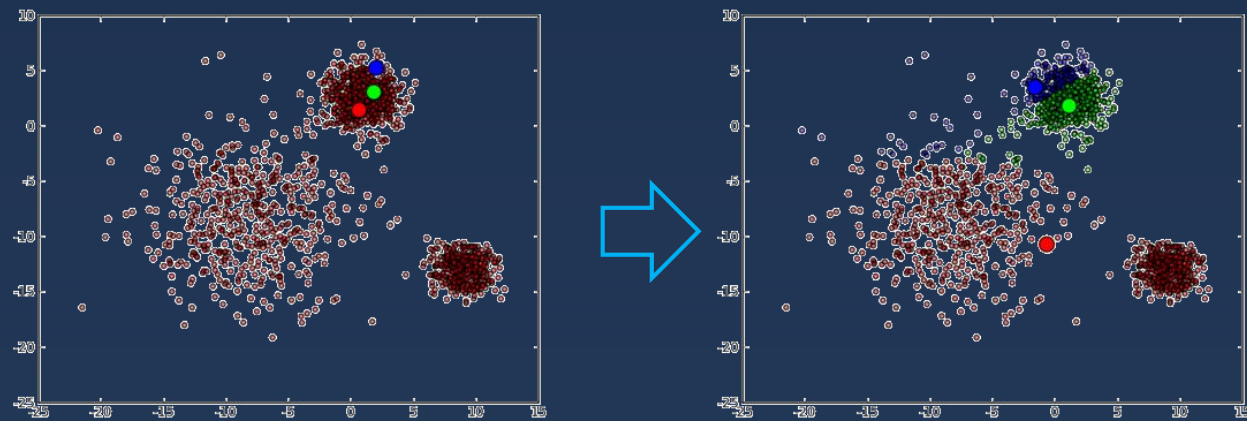


最小化目标函数

$$J = \sum_{n=1}^N \sum_{k=1}^K r_{nk} \|x_n - \mu_k\|^2$$

$$\mu_k = \frac{\sum_n r_{nk} x_n}{\sum_n r_{nk}}$$

失败案例



行业通用的机器学习算法选择

分类	回归	推荐
朴素贝叶斯：模型参数较少，对缺失值不敏感，要求特征之间相互独立	线性回归：适用于预测目标与特征之间线性关系强的数据集	域分解机：特征选择不敏感，推荐精确度高，计算量大
随机森林分类：适用范围广精度高，广泛应用于分类场景	随机森林回归：多分类器计算量小，回归误差小	交替最小二乘：计算量小，适用于特征相关性较高的数据集
支持向量机：适用于小样本、非线性的数据集	K最近邻回归：计算量大无需建模过程，解释性好	

企业应用AI的难点

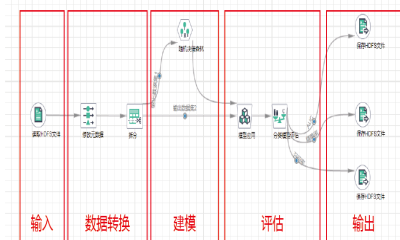
企业应用AI需要具备的要素	为什么需要这些要素	如何解决
AI人才	AI人才稀缺，只有AI人才能够将应用场景问题使用AI算法解决。	企业单独聘用AI人才的成本很高，一个专业的AI服务团队可以解决这个问题
建模&测试工具	需要一个易上手、易构建模型、易部署模型的工具或者平台	企业自采购计算集群、GPU等代价太大，一个通用的平台可以解决这个问题
数据+行业知识	需要一定的数据储备，需要能够将行业知识、人工经验总结出来	企业在没有数据的情况下需要找到可以复用的AI模型积累了多个模型、场景模板的平台可以解决，一这个问题

华为云机器学习服务（ Machine Learning Service ），一站式AI平台，解决企业应用AI难点。

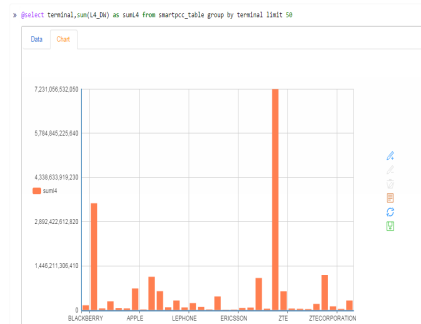
华为机器学习服务MLS：一站式数据挖掘分析平台

机器学习服务（Machine Learning Service），数据挖掘分析平台服务，帮助用户通过机器学习技术发现已有数据中的规律，从而创建机器学习模型，并基于机器学习模型处理新的数据，为业务应用生成预测结果。

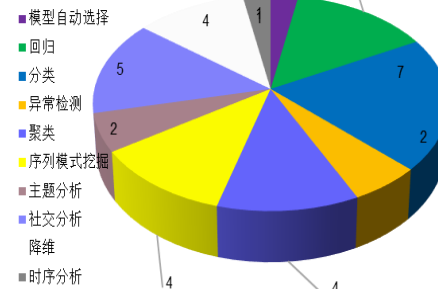
拖拉拽模型构建



交互式建模分析



深度优化算法库50+

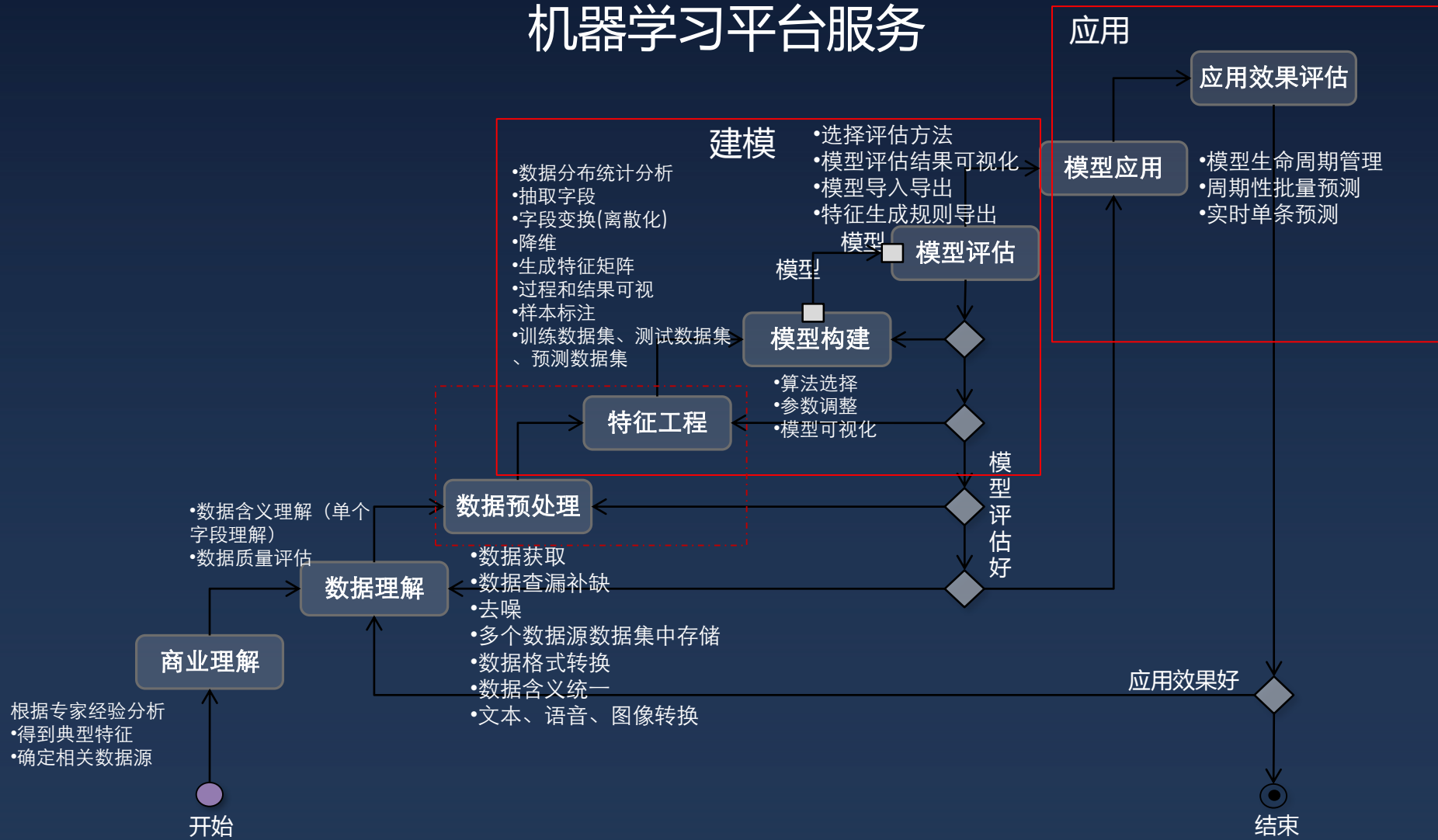


模型管理/发布可视化

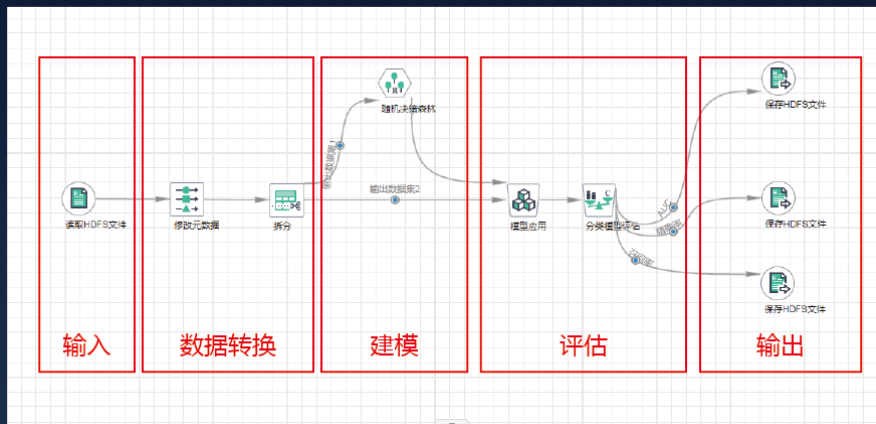


提供从特征工程到模型应用效果评估的机器学习流程支持

机器学习平台服务



探索环境之Workflow：面向普通分析师，所见即所得的建模开发界面

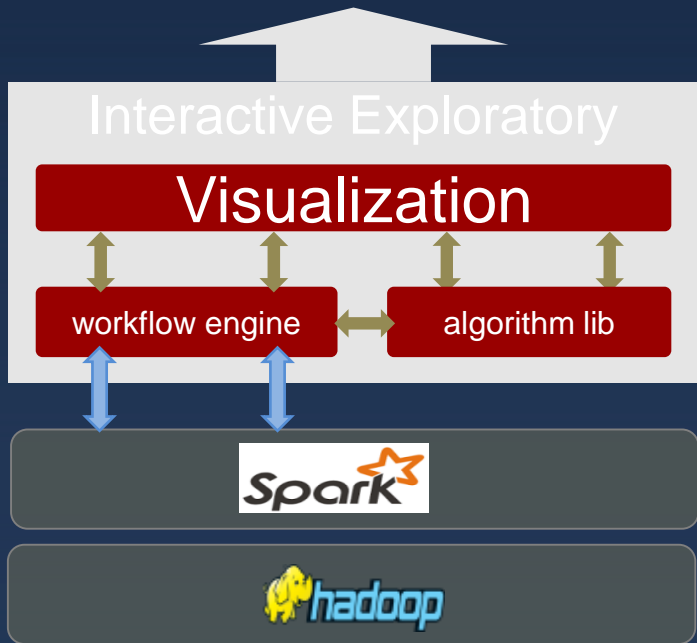
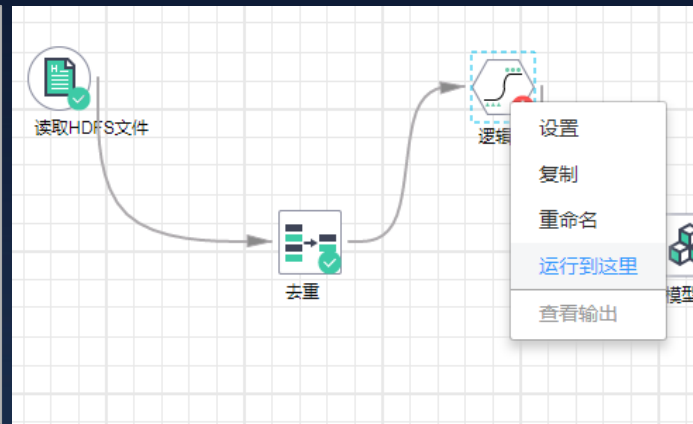


查看输出"输出数据集"

元数据 运行结果

行: 50 列: 52

attr_22	attr_23	attr_20	attr_21	attr_28	attr_29
1	1566	0	0	1	0
1	0	0	0	0	1
1	0	0	0	1	0
1	3675	1	0	1	0
1	1779	1	0	0	1



Confusion Matrix

	class0	class1	RECALL
class0	377	21	94.7%
class1	49	353	87.8%
PRECISION	88.5%	94.4%	

Confusion Matrix

	class0	class1	RECALL
class0	222	51	81.3%
class1	65	62	48.8%
PRECISION	77.4%	54.9%	

Indicators

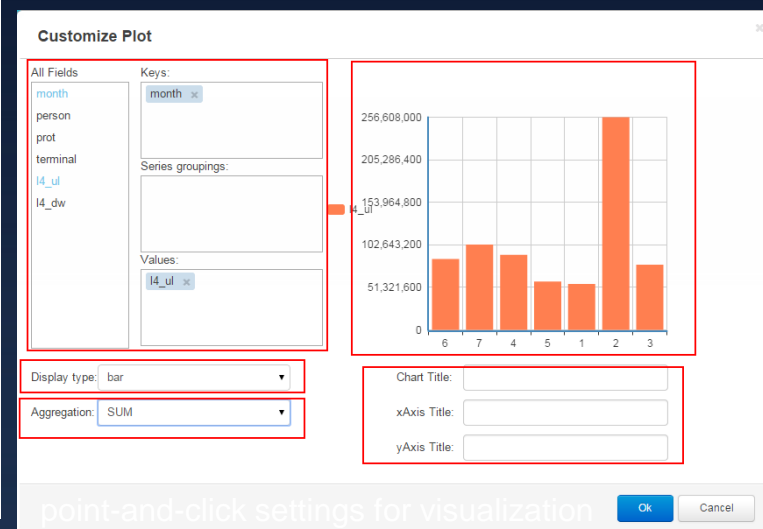
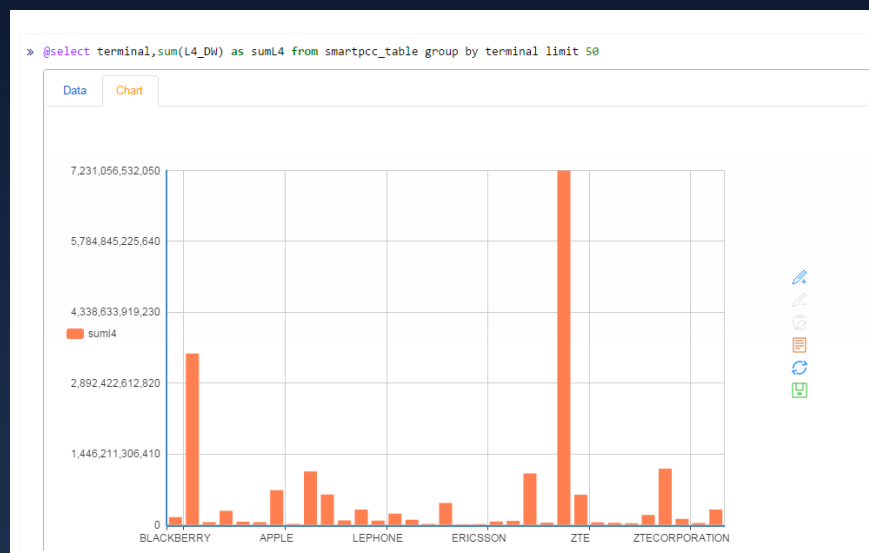
accuracy	auc
0.71	0.731

Tables

Class Name	f1scores	precisions	recalls
class0	0.793	0.774	0.813
class1	0.517	0.549	0.488

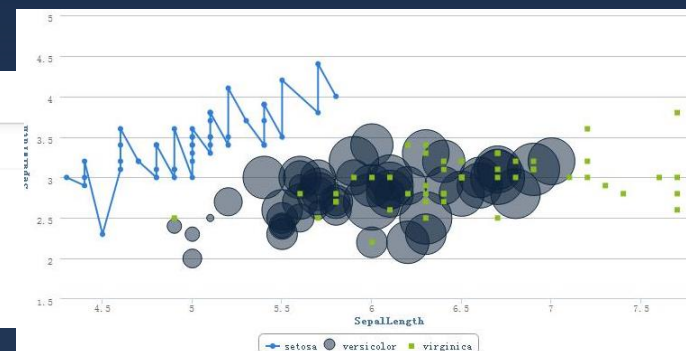
- 可视化拖拽业务流程构建：简单直观的建模流程设计，零编码即可将业务分析、建模流程设计出来
- 灵活调试：通过任意指定执行路径，查看每个步骤的执行结果，可以快速调试和优化端到端流程
- 丰富可视化：集成数据和模型可视化能力，探索分析所见即所得，数据、机器学习模型即时可视
- 灵活的计算资源调度：针对每个具体的处理任务，可以按需指配计算资源

探索环境之Notebook：面向数据科学家的多语言交互分析、可视化环境



```
move_detect_big_data_demo Last Checkpoint: May 21 23:46 (unsaved changes)

» hist_res = ddf.stat.hist(data,7,"dropped_call_count")
hist = data.frame(t(hist_res[[1]]))
#ddf.vs.table(hist)$showRCharts()
ddf.vs.hist(hist)$showRCharts()
```



- **交互式Notebook框架**：数据分析的“浏览器”，统一入口，与底层大数据平台计算框架可解耦；
- **语言支持**：让科学家用最习惯的语言编程，支持Python)，兼容开源Python库
- **丰富可视化**：交互式分析快速响应，集成数据可视化能力，探索分析**所见即所得**，数据、机器学习模型**即时可视**
- **安全性**：通过**完整的安全管控机制**，保证用户之间Notebook的权限隔离，防止数据的非法访问

预置丰富的分布式增强机器学习算法，让建模更加简单

输入

读取模型

读取PMML模型文件

读取Hive表

读取HDFS文件

数据转换

记录操作

字段操作

建模

异常检测

分类

聚类

回归

推荐

评估

模型应用

分类模型评估

回归模型评估

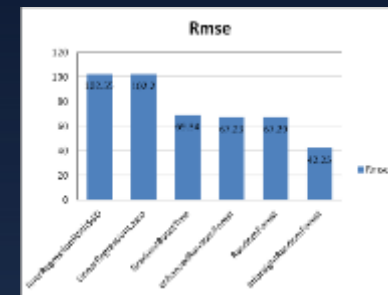
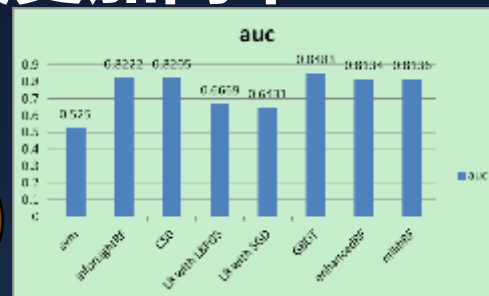
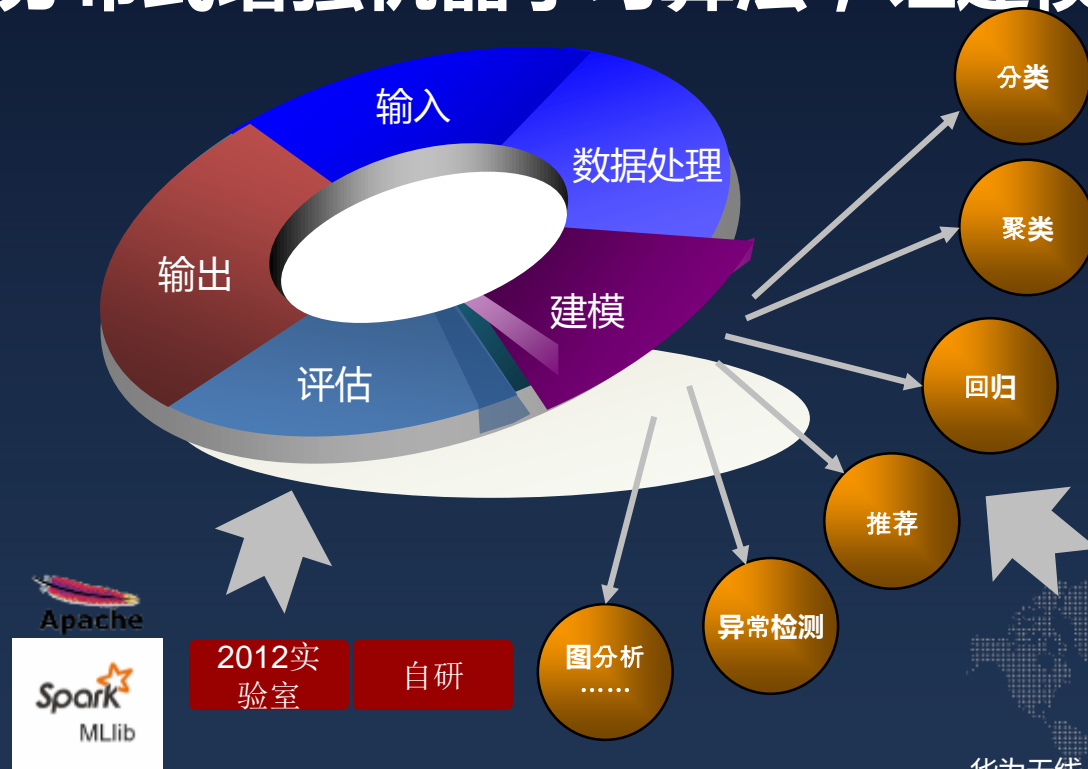
输出

保存HDFS文件

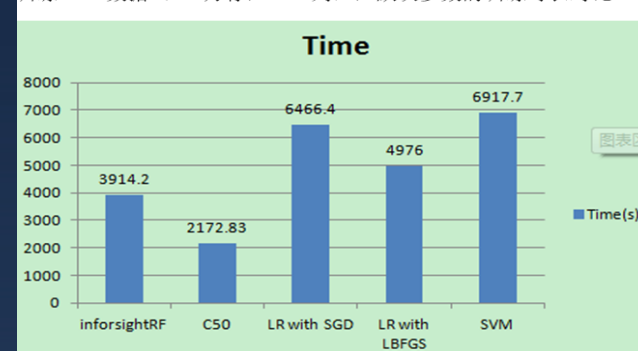
保存模型

保存PMML模型文件

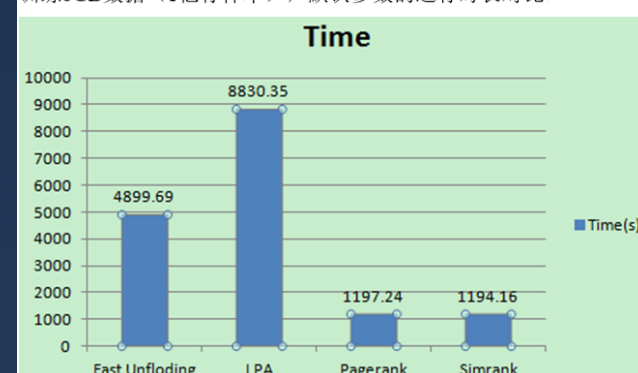
保存Hive表



训练100G数据（546万行，1000列），默认参数的训练时长对比。



训练6GB数据（1亿行样本），默认参数的运行时长对比。




华为无线/IT/ISC/GTS
全球海量大数据项目验证

- **预置丰富的机器学习算法**：从数据导入和处理，到模型训练和评估、导出，**覆盖数据挖掘端到端业务**
- **分布式增强**：内置算法在分布式处理性能上进行了专项优化，相比开源实现有**更好的性能和线性加速表现**，支持处理更大规模的数据
- **持续优化**：华为的众多大数据挖掘项目不断积累新算法，优化已有算法，成果会迅速转化为MLS服务的能力


机器学习服务场景方案举例

预测性维护




设备工作状态预测：如设备是否会发生故障、设备剩余寿命

推荐




广告推荐：根据客户喜好提供个性化推荐

异常轨迹分析




分析车联网车辆行驶数据，发掘车辆异常轨迹，减少企业损失。

话题发现




舆情监控，从大量文本数据中识别热点话题，了解当前舆论趋势

客户挽留




分析客户属性及行为，预测客户流失率，指导企业制定挽留方案

零售商分群



根据零售商的进出货数据对零售商进行分类，进行针对性的管理

驾驶行为分析

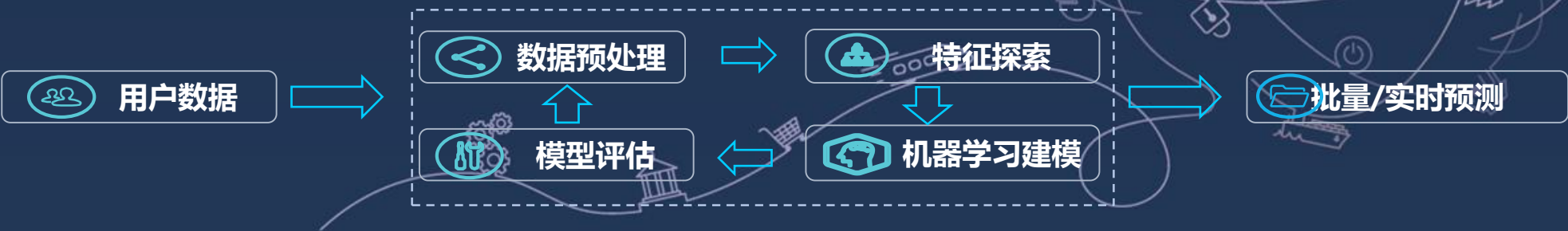
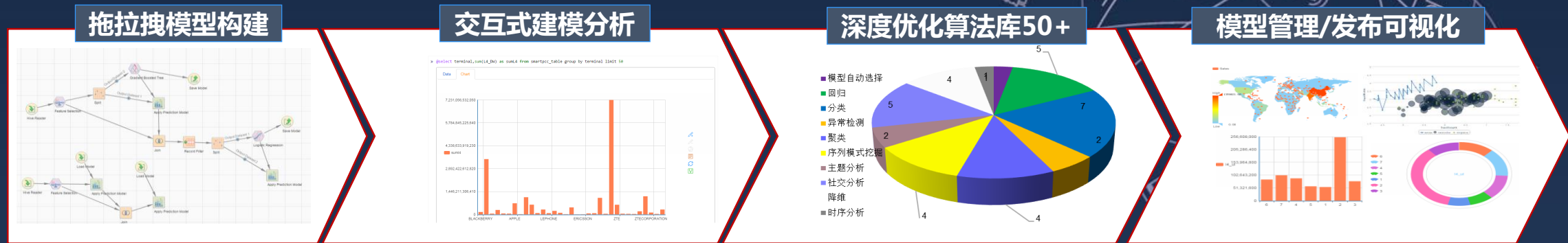


分析车辆行驶数据，识别用户驾驶习惯，保险创新带来二次收益

文本分类



新闻网站中文本的自动分类，如体育、娱乐、政治等





Thank You.

Copyright©2016 Huawei Technologies Co., Ltd. All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.

华为云机器学习服务MLS
www.huaweicloud.com/product/mls.html