成为机器学习英雄

公益讲座11：00准时开始，请大家先浏览云技术微信公众号技术文章。资料会在各群同步发布，已入群客户请勿重复入群！

20-20

数据库和云讲座群

甲骨文云技术公众号

ORACLE

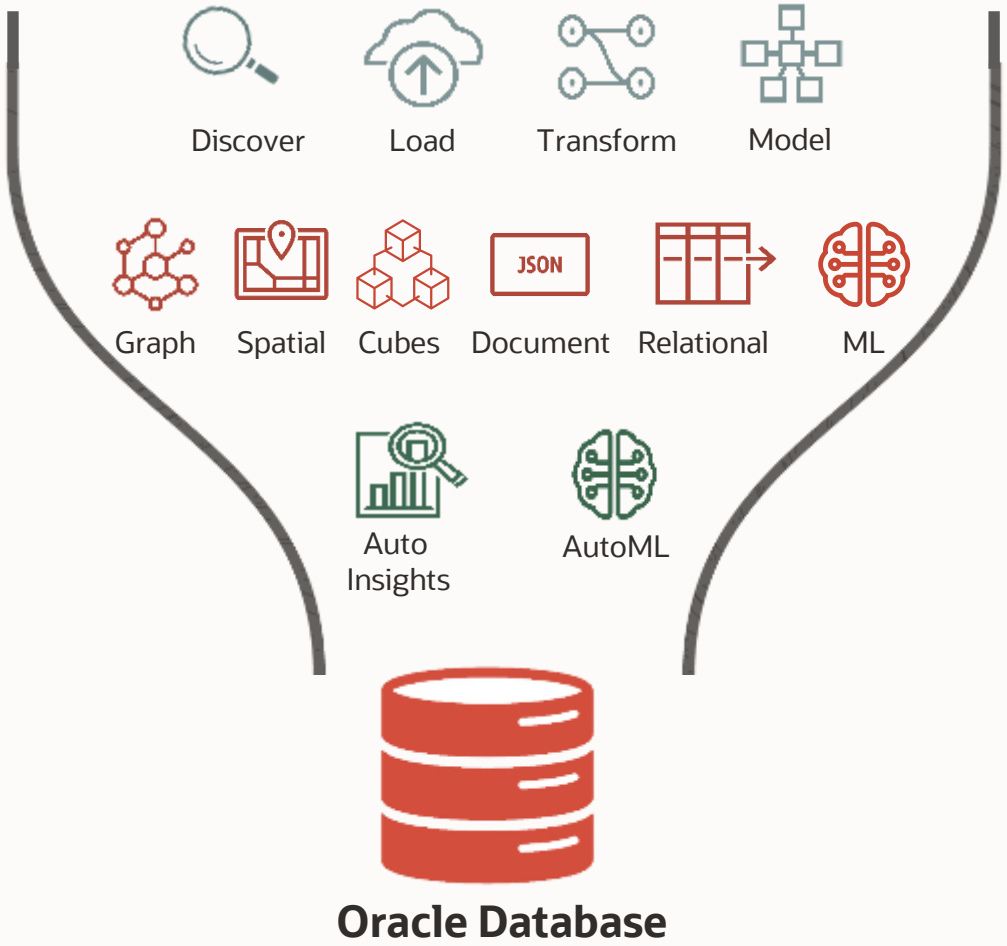
成为机器学习英雄

段敏明
SEHub Analytics
2023年1月

到2024年，机器学习驱动的企业对客户、竞争对手和监管机构的响应速度将比同行快50%。

IDC

为客户提供全球更佳的多模数据库



多模数据库

关系型、JSON、 XML、 图、 空间、 OLAP、 区块链

多种工作负载

在线交易、分析、**机器学习**、In-memory、 物联网、流媒体、 多租户、持久内存存储等

多种角色使用-开发人员和分析师

任何数据上的声明式SQL和事务、Java、 JavaScript、**ML4SQL**、**ML4Python**、 **AutoML**、微服务、事件、CI/CD、 APEX

Oracle 机器学习（Machine Learning）

自动化

- 以更少的投入，更快地获得更好的结果
- 公民智能—甚至是非专家人员

提升效率

扩展性

- 提供并行，分布式算法来处理海量数据
- 算法靠近数据，避免数据移动

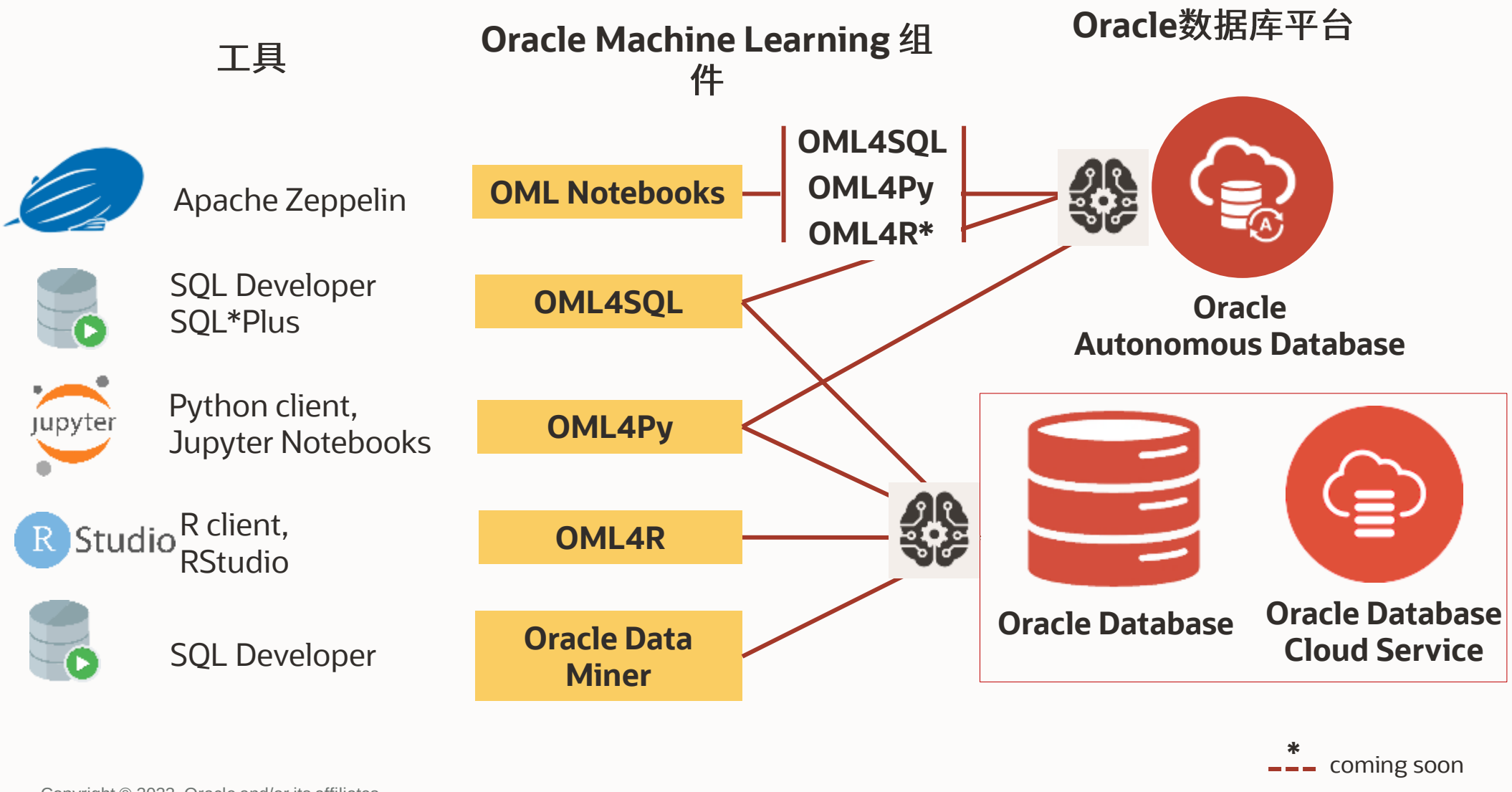
更快实现业务目标

产品完善度

- 更快部署和更新数据科学解决方案
- 提供集成的机器学习平台

更多关注在创新

Oracle数据库内机器学习



Oracle数据库内置的机器学习与高级分析能力

分类
- Naïve Bayes
- Logistic Regression (GLM)
- Decision Tree
- Random Forest
- Neural Network
- Support Vector Machine (SVM)
- Explicit Semantic Analysis
- *XGBoost**

异常检测
- One-Class SVM
- *MSET-SPRT**

聚类
- Hierarchical K-Means
- Hierarchical O-Cluster
- Expectation Maximization (EM)

时间序列
- Forecasting - Exponential Smoothing
- Includes popular models
  e.g. Holt-Winters with trends, seasonality, irregular time series

回归
- Generalized Linear Model (GLM)
- Support Vector Machine (SVM)
- Stepwise Linear regression
- Neural Network
- *XGBoost**

属性重要性
- Minimum Description Length
- Principal Component Analysis (PCA)
- Unsupervised Pairwise KL Divergence
- CUR decomposition for row & Al

关联规则
- A priori

预测查询
- Predict, cluster, detect, features

SQL 分析
- SQL Windows
- SQL Patterns
- SQL Aggregates

特征提取
- Principal Comp Analysis (PCA)
- Non-negative Matrix Factorization
- Singular Value Decomposition (SVD)
- Explicit Semantic Analysis (ESA)

行重要性
- CUR Decomposition

排名
- *XGBoost**

文本挖掘
- Algorithms support text columns
- Tokenization and theme extraction
- Explicit Semantic Analysis (ESA)

统计函数
- min, max, median, stdev, t-test, F-test, Pearson's, Chi-Sq, ANOVA, etc.

R和Python包
- Third-party R and Python Packages through Embedded Execution
- Spark MLlib algorithm integration

** New in 21c*

Oracle 数据库内机器学习总结

1.一站式AI算法训练

- ✓ 在Oracle数据库中实现数据收集、数据清洗、特征工程、模型训练、模型评估、模型部署以及持续优化一站式AI算法训练；

2.高性能

- ✓ 算法靠近数据，避免数据移动
- ✓ 库内并行的，分布式算法
- ✓ 支持批量和实时评分，也支持分区和并行评分
- ✓ 可以利用智能扫描技术，把评分下推到存储层以获得更好评分性能。

3.易用性

- ✓ 提供AutoML自动机器学习
- ✓ 写SQL即可构建机器学习模型
- ✓ 提供拖拽UI的机器学习工具
- ✓ 提供交互式Web界面的工具

4.多样性

- ✓ 内置客户分类、异常检测、实时推荐、销售预测等30多种算法
- ✓ 支持SQL、R、Python等多种语言构建模型
- ✓ 多种部署使用方式，SQL、Python、REST API等

Oracle 机器学习笔记本

适用于Oracle自治数据库

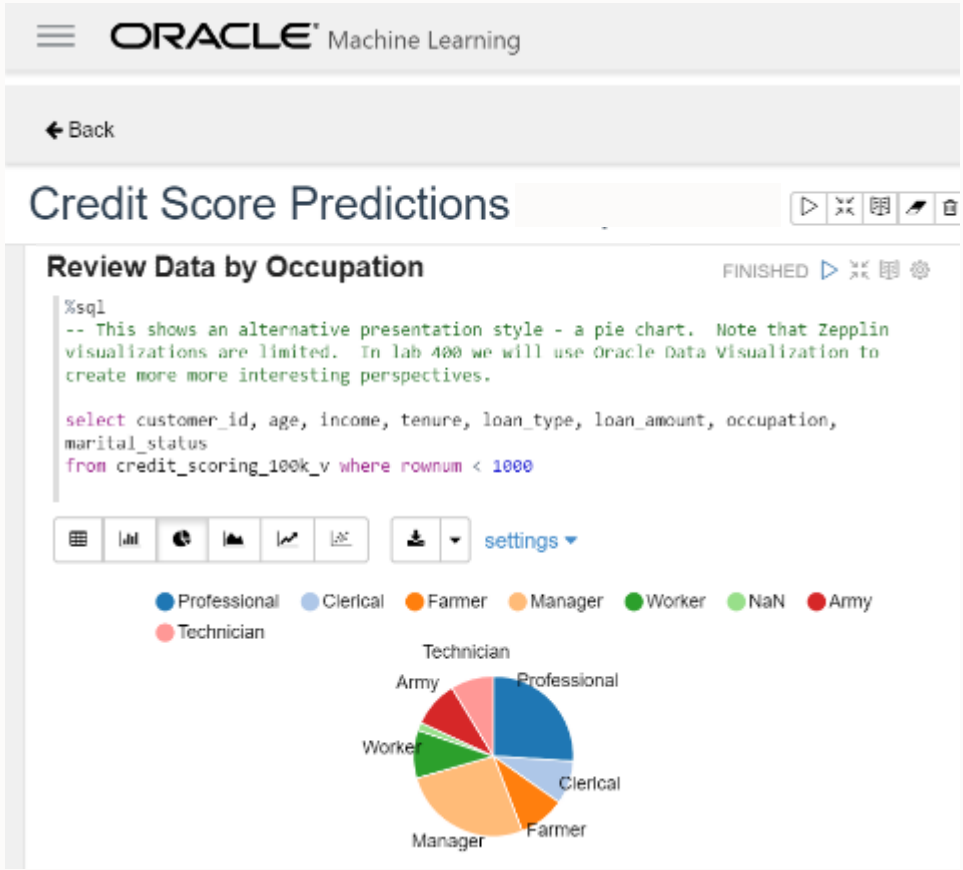
Oracle 机器学习笔记本

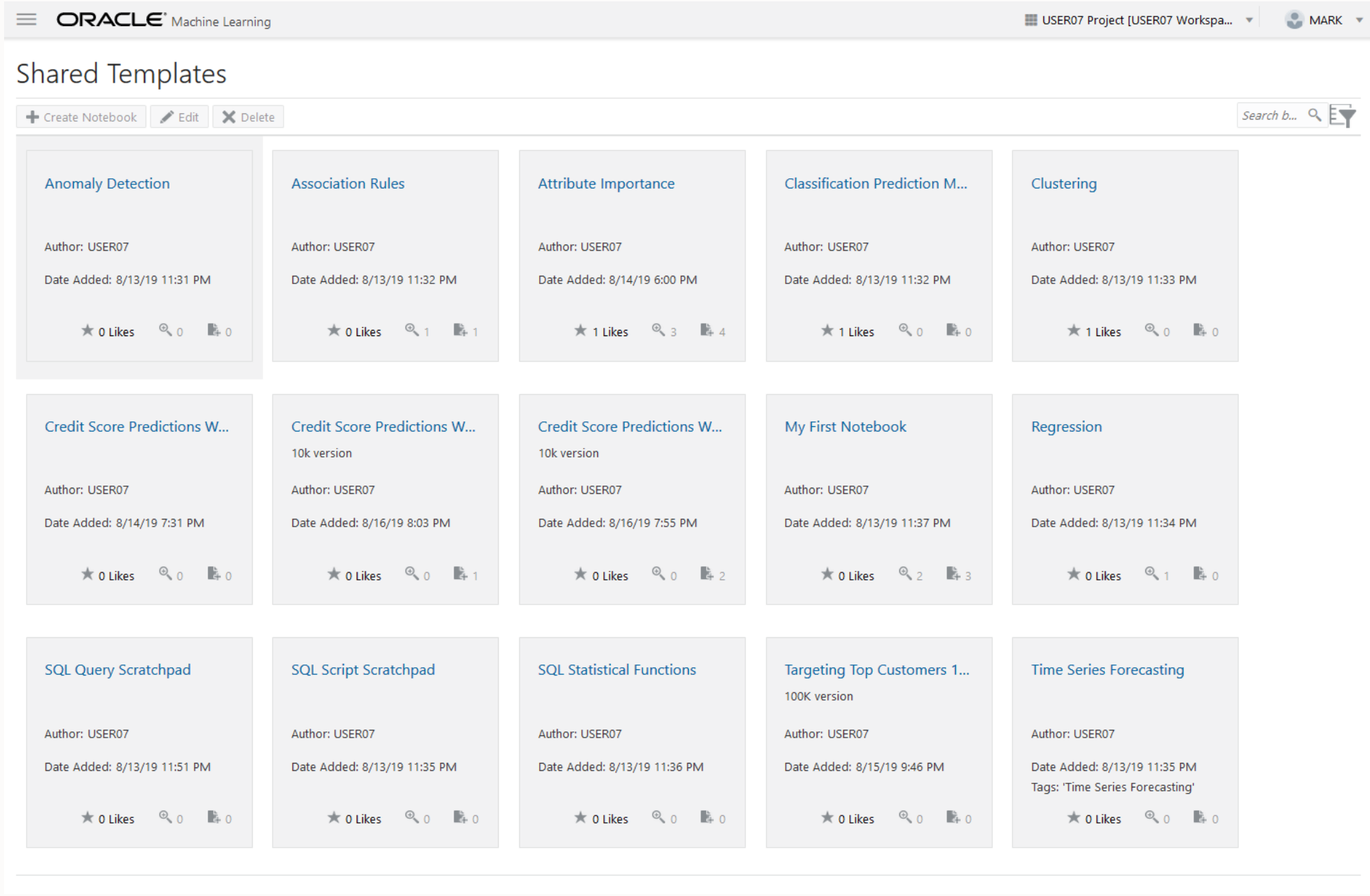
自治数据库作为数据科学平台

- 协作式用户界面
- 基于Zeppelin
- 支持数据科学家、数据分析师、应用程序开发人员、DBA
- 轻松共享笔记本和模板
- 权限、版本控制和执行计划

ADW内置

- 自动配置、管理、备份
- 数据库内 SQL 算法和分析功能
- 通过 Python 和 R 进行增强





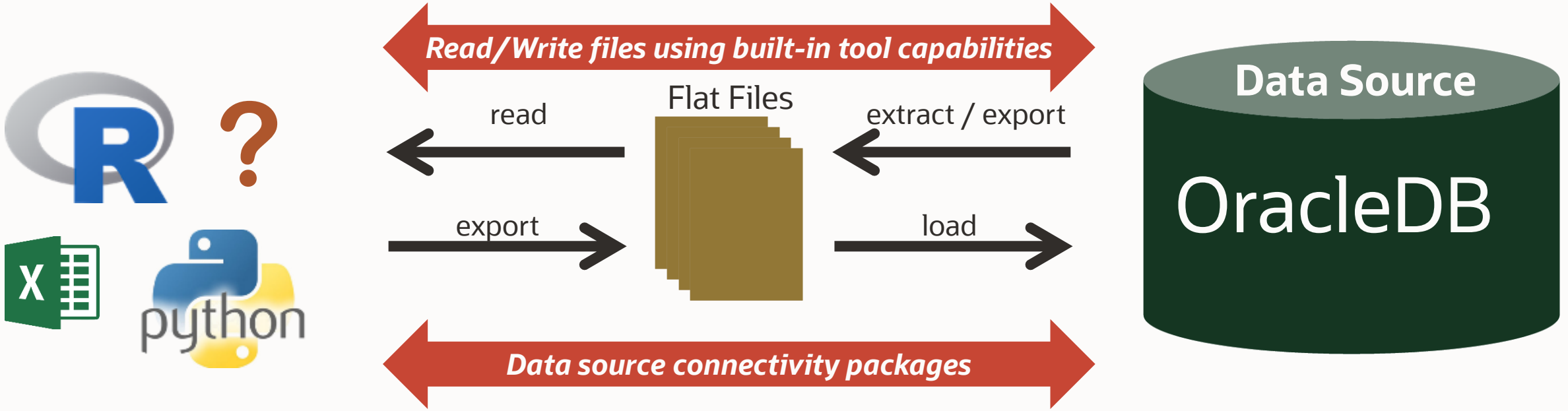
Oracle Machine Learning for

SQL (OML4SQL)
Python (OML4Py)
R (OML4R)

使 SQL 用户能够立即访问 Oracle 数据库和 Oracle 自治数据库中的 ML

为数据科学家提供开源环境

传统分析和数据源交互



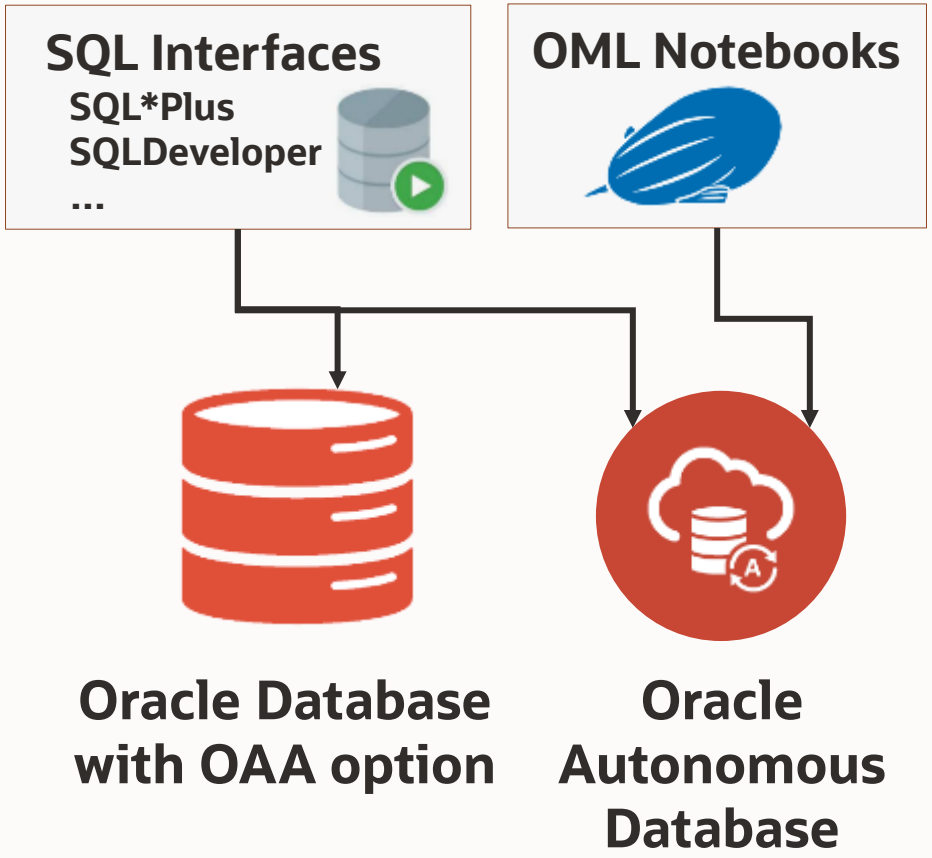
部署
临时
cron 作业

- 访问延迟
- 范式转变：R/Python → 数据访问语言 → R/Python
- 内存限制 – 数据大小、内存处理
- 单线程
- 备份、恢复、安全性问题
- 临时生产部署

适用于 SQL 的 Oracle 机器学习

甲骨文自治数据库的组件和高级分析

- 数据库内、并行、分布式算法
- ML 模型作为第一类数据库对象
- 跨数据库导出/导入模型
- 批量和实时评分
- 解释性预测详细信息
- 跨技术堆栈利用机器学习



OML4SQL:模型构建和实时预测

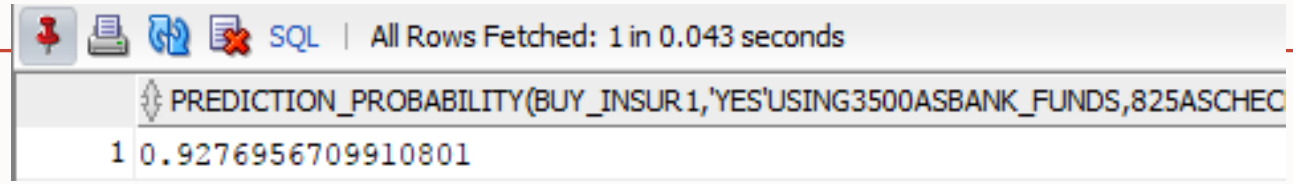
简单 SQL 语法 — 分类模型

模型构建（PL/SQL）

```
BEGIN
  DBMS_DATA_MINING.CREATE_MODEL(
    model_name          => 'BUY_INSUR1',
    mining_function     => dbms_data_mining.classification,
    data_table_name     => 'CUST_INSUR_LTV',
    case_id_column_name => 'CUST_ID',
    target_column_name  => 'BUY_INSURANCE',
    settings_table_name => 'CUST_INSUR_LTV_SET');
END;
```

实时评分（SQL 查询）

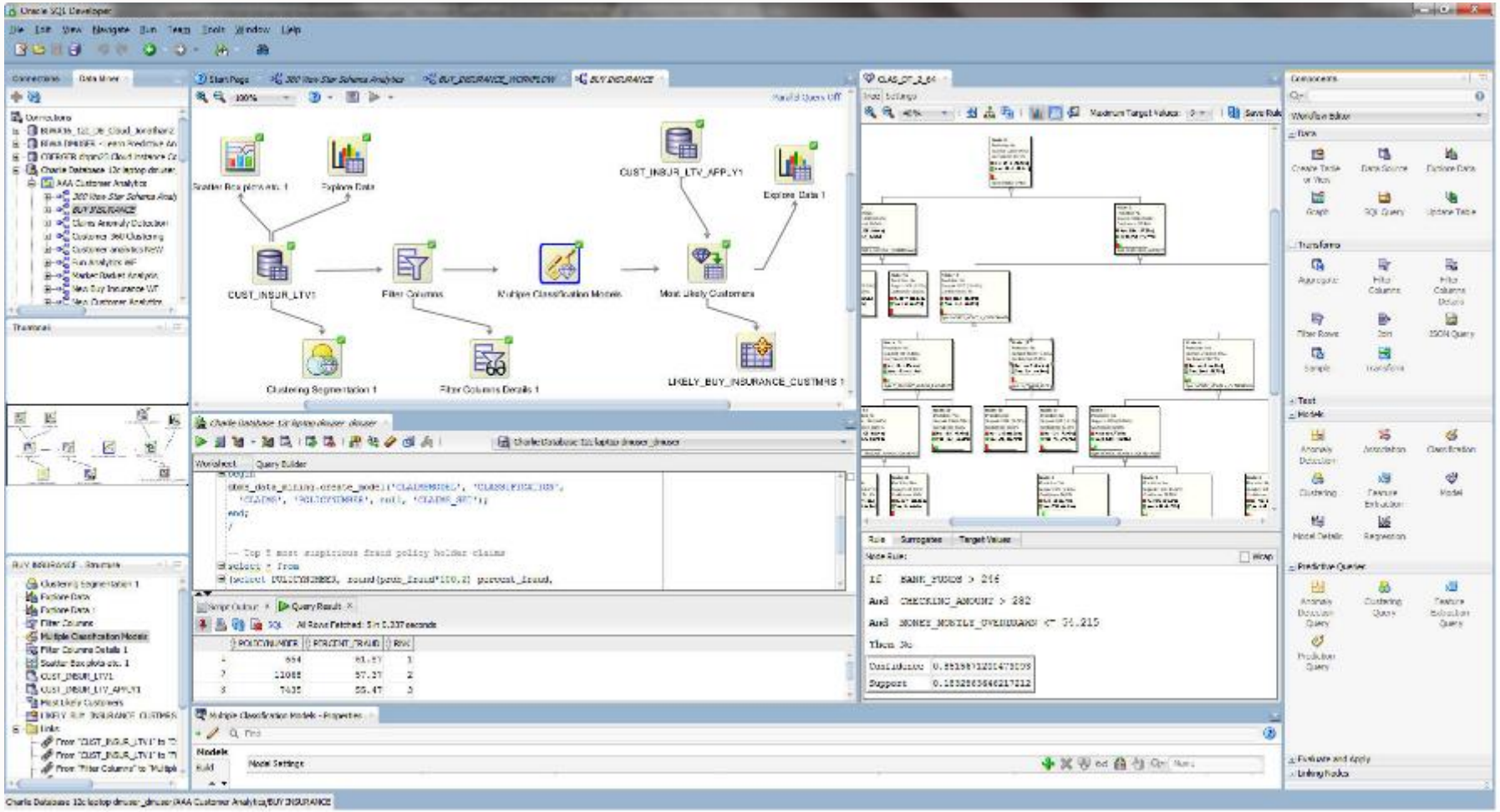
```
SELECT prediction_probability(BUY_INSUR1, 'Yes'
  USING 3500 as bank_funds, 825 as checking_amount, 400 as credit_balance, 22 as age, 'Married' as
    marital_status, 93 as MONEY_MONTLY_OVERDRAWN, 1 as house_ownership)
FROM dual;
```



Oracle 数据挖掘器用户界面

创建分析工作流 – “人人都是数据科学家”

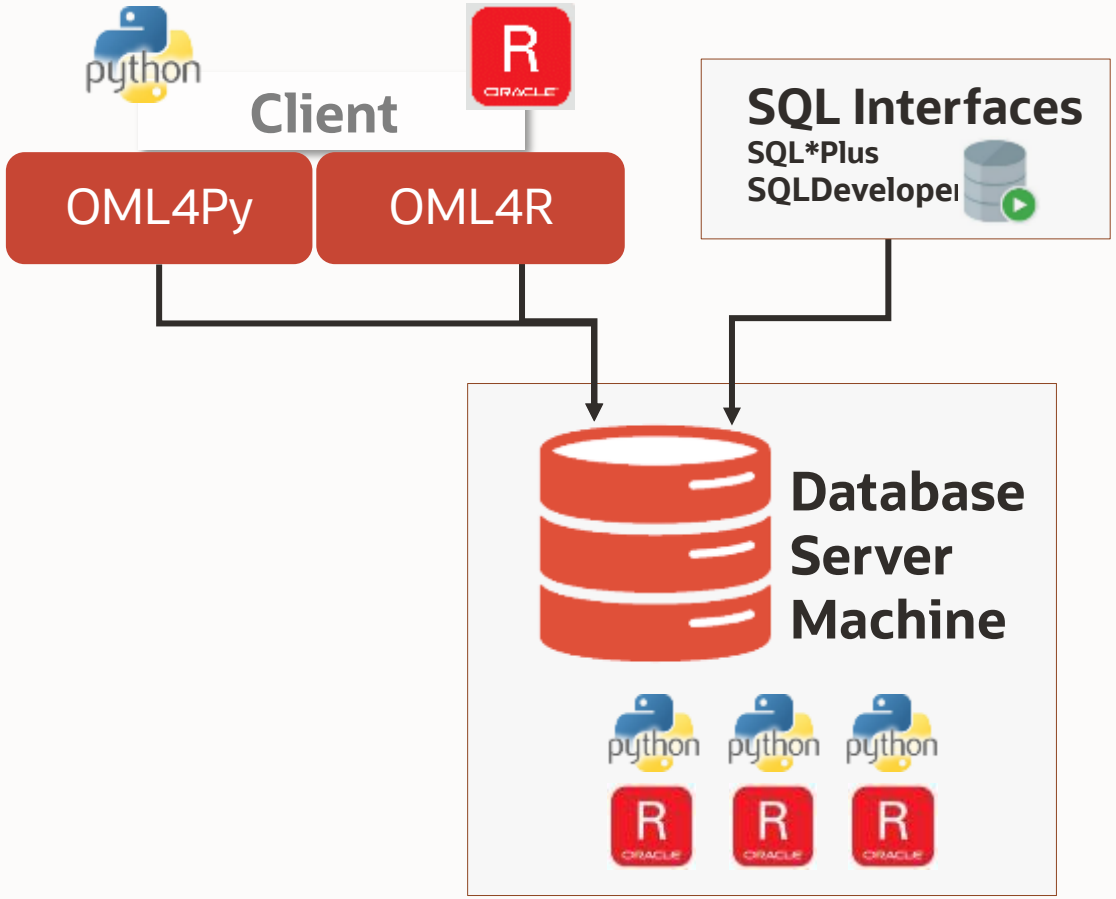
- SQL 开发人员扩展
- 自动执行典型的数据科学步骤
- 易于使用的拖动和拖放接口
- 快速定义和共享分析工作流程
- 广泛的算法和数据转换
- 生成用于立即部署的 SQL 代码



适用于 R 和 Python 的 Oracle Machine Learning *

Oracle数据库高级分析

- Oracle 数据库即 HPC 环境
- 数据库内并行和分布式机器学习算法
- 管理 Oracle 数据库中的脚本和对象
- 将结果集成到应用程序中
- 和通过 SQL 的仪表板
- OML4Py 自动化机器学习

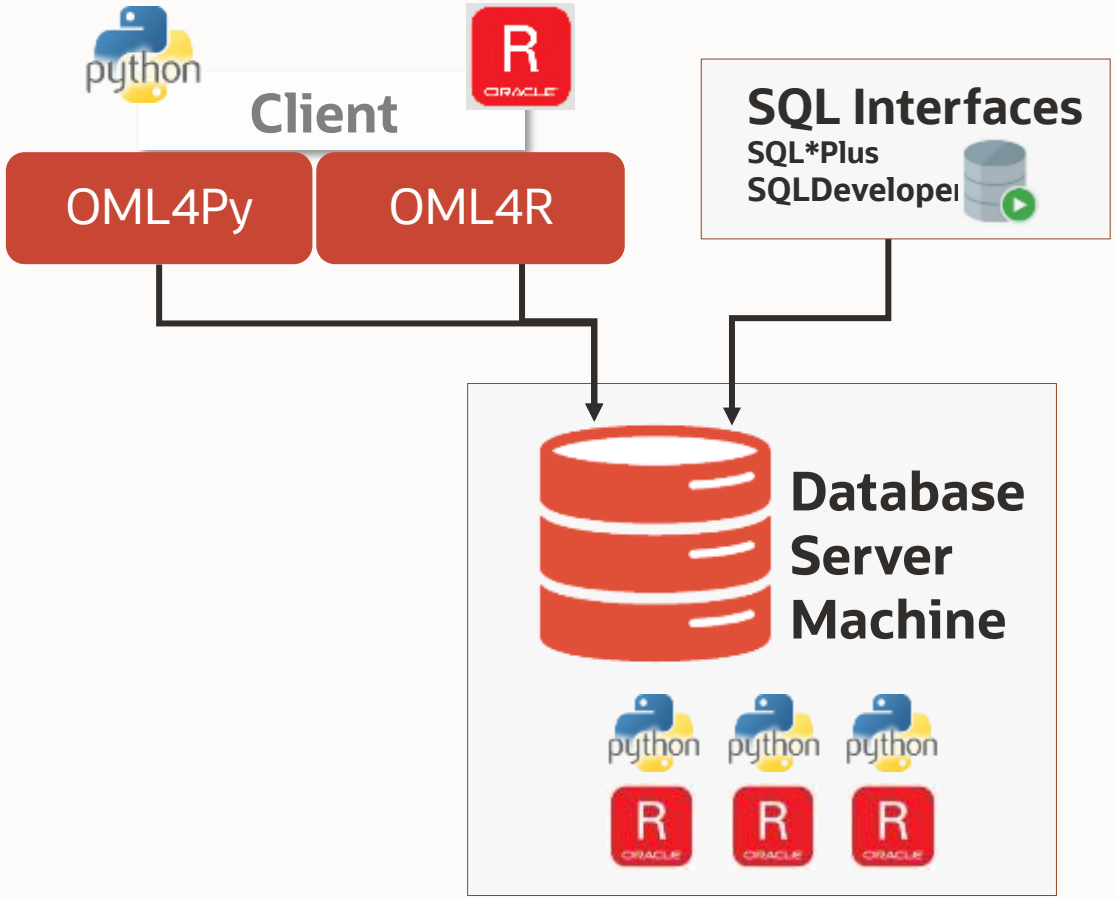


** Coming soon*

Oracle Machine Learning for R and Python

数据库高级分析

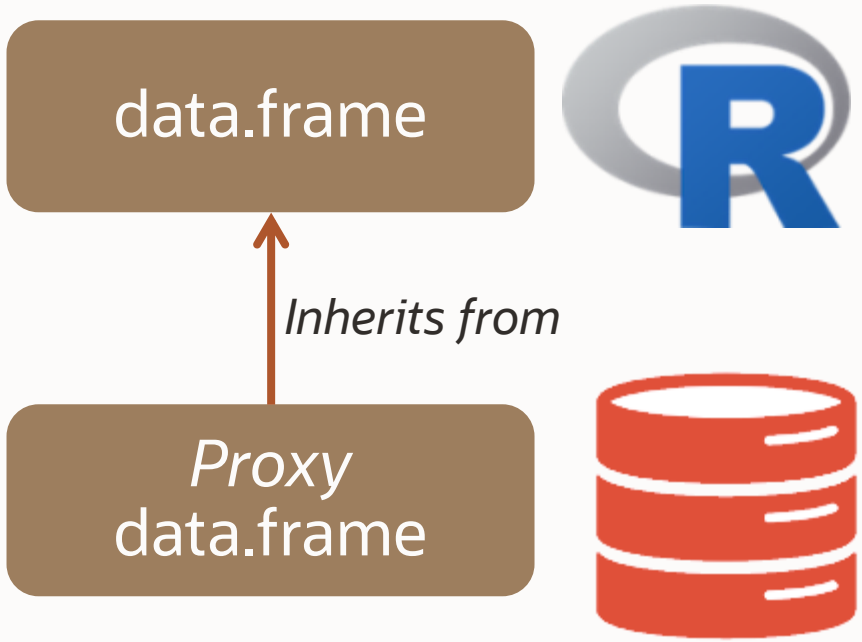
- 透明层
- 利用代理对象，使数据保留在数据库中
- 重载将功能转换为 SQL 的本机函数
- 对数据库数据使用熟悉的 R/Python 语法
- 并行、分布式算法
- 可扩展性和性能
- 公开 OML4SQL 提供的数据库内算法
- 嵌入式执行
- 在 Oracle 数据库中管理和调用 R 或 Python 脚本
- 数据并行、任务并行和非并行执行
- 使用开源包来增强功能
- OML4Py AutoML
- 模型选择、特征选择、超参数调优



代理对象

使用 OML4R 接口的示例

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa




```
> str(iris)
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
> str(IRIS)
'data.frame':   150 obs. of  5 variables:
Formal class 'ore.frame' [package "OREbase"] with 12 slots
  ..@ .Data    : list()
  ..@ dataQry  : Named chr "( select /*+ no_merge(t) */  \"Sepal.Length\" VAL001,\"Sepal.Wid
th\" VAL002,\"Petal.Length\" VAL003,\"Petal.Width\" VAL004,\"Sp"| __truncated__
  ..@ sqlName  : chr
  ..@ sqlValue : chr  "\"Sepal.Length\"" "\"Sepal.Width\"" "\"Petal.Length\"" "\"Petal.Width
\"" ...
  ..@ sqlTable : chr "\"RQUSER\".\"IRIS\""
  ..@ sqlPred  : chr ""
  ..@ extRef   : list()
  ..@ names    : chr
  ..@ row.names: int
  ..@ .S3Class : chr "data.frame"
```

```
"( select /*+ no_merge(t) */  \"Sepal.Length\" VAL001,\"Sepal.Width\" VAL
002,\"Petal.Length\" VAL003,\"Petal.Width\" VAL004,\"Species\" VAL005 fro
m \"RQUSER\".\"IRIS\" t  )"
```

透明层

数据库内性能 – 索引、查询优化、并行性、分区

利用代理对象处理数据: *oml.DataFrame*

```
# Create table from Pandas DataFrame data
DATA = oml.create(data, table = 'BOSTON')

# Get proxy object to DB table boston
DATA = oml.sync(table = 'BOSTON')
```

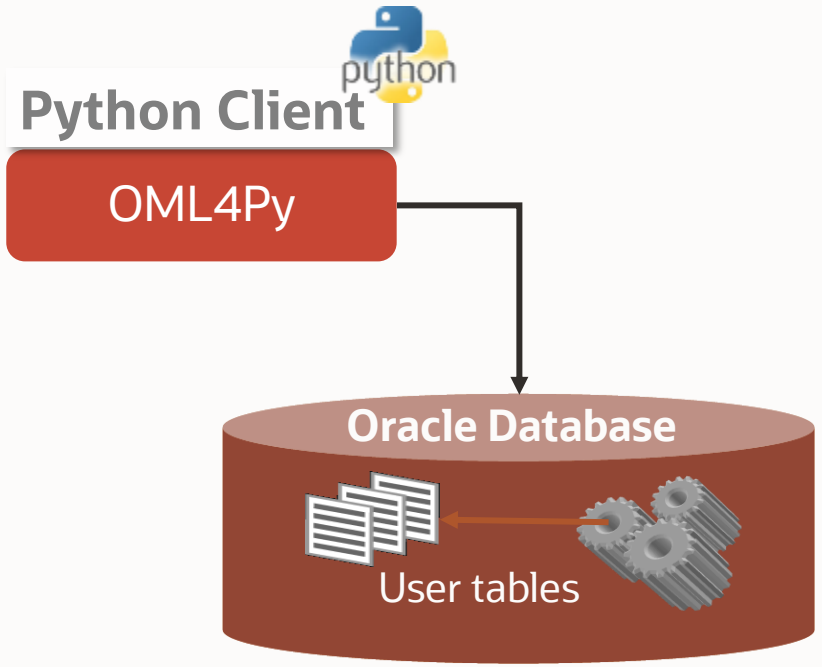
使用熟悉的 Python 语法操作数据库数据
重载将功能转换为 SQL 的 Python 函数

```
DATA.shape
DATA.head()
DATA.describe()
DATA.std()
DATA.skew()

TRAIN, TEST =
  DATA.split()
TRAIN.shape
TEST.shape
```

并行、分布式算法

使用支持向量机进行数据库内建模

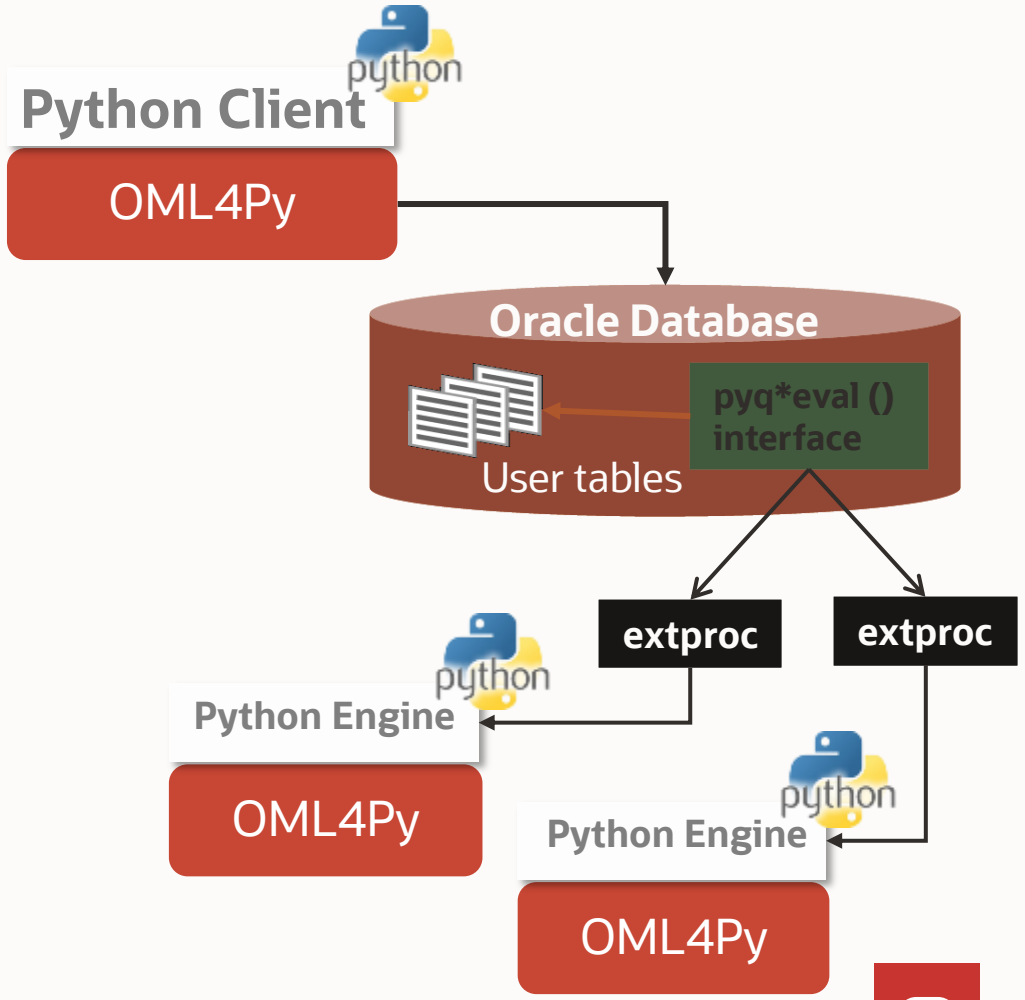
```
from oml import svm

# create proxy object
ONTIME_S = oml.sync(table='ONTIME_S')

# define model object
settings = {'svms_outlier_rate' : 0.01}
svm_mod = svm('anomaly_detection',
          svms_kernel_function =
            'dbms_data_mining.svms_linear',
          **settings)
# build anomaly detection model
svm_mod = svm_mod.fit(x=ONTIME_S, y=None)
# view model object
svm_mod
```



嵌入式执行

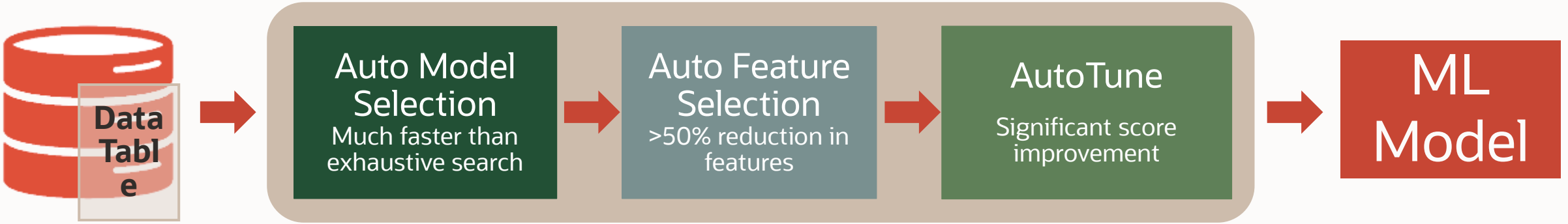
使用第三方包并行执行分区数据流的示例

```
# user-defined function using sklearn
def build_lm(dat):
  from sklearn import linear_model
  lm = linear_model.LinearRegression()
  X = dat[['PETAL_WIDTH']]
  y = dat[['PETAL_LENGTH']]
  lm.fit(X, y)
  return lm
# select column(s) for partitioning data
index = oml.DataFrame(IRIS['SPECIES'])
# invoke function in parallel on IRIS table
mods = oml.group_apply(IRIS, index,
                       func=build_lm,
                       parallel=2)
mods.pull().items()
```



AutoML – OML4Py 的新功能

提高数据科学家的工作效率 – 减少总体计算时间



- 自动选型
- 确定实现更高模型质量的数据库内算法
- 比使用详尽搜索更快地找到更佳模型

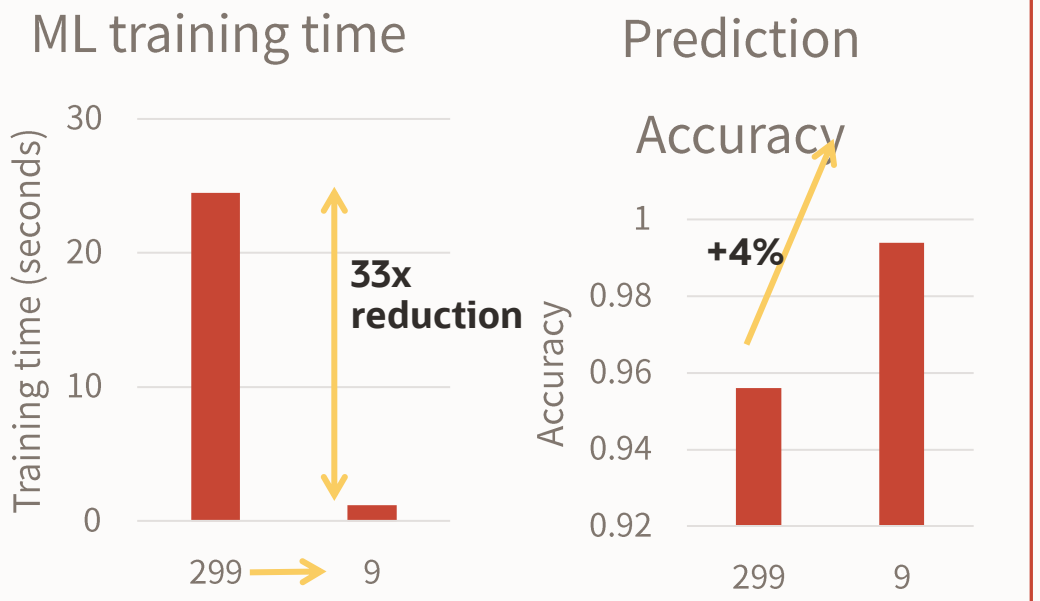
- 自动功能选择
- 通过识别更具预测性的特征来减少特征的数量
- 提高性能和准确性

- 自动调整超参数
- 显著提高模型准确性
- 避免手动或详尽的搜索技术

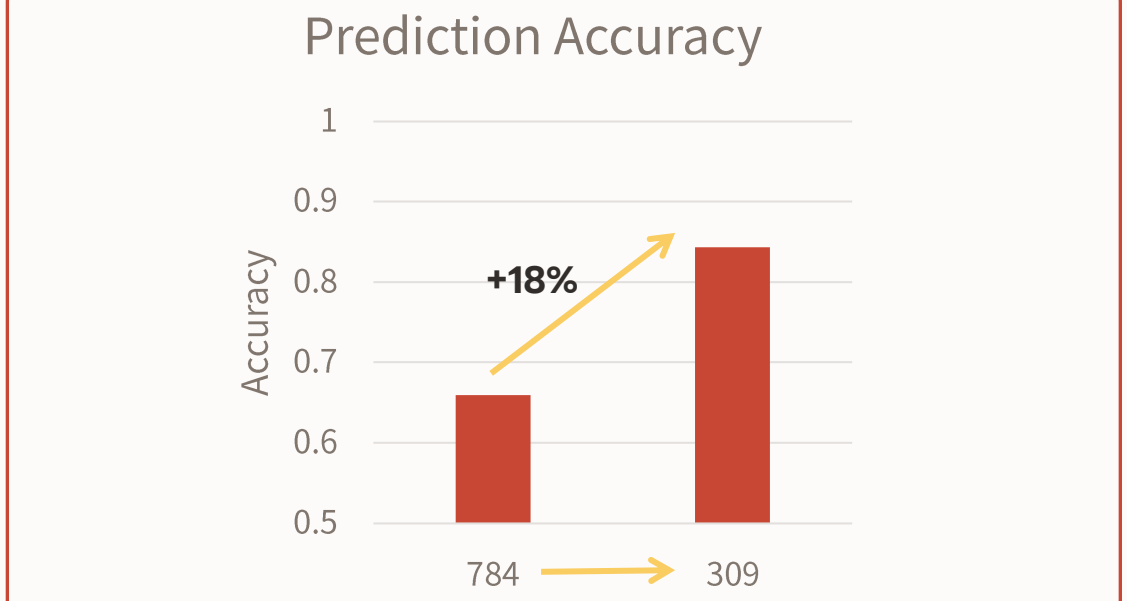
让非专家用户能够利用机器学习

自动功能选择

通过识别更相关的功能来减少特征数量
提高性能和准确性



OpenML 数据集 312，包含 1925 行，299 列

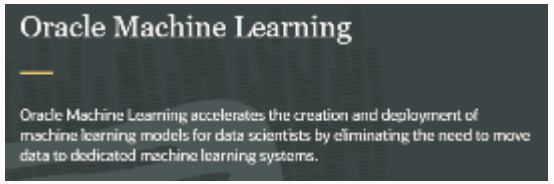


OpenML 数据集 40996，包含 56K 行，784 列

Oracle数据库内机器学习资料

ORACLE MACHINE LEARNING ON O.COM

https://www.oracle.com/machine-learning



OML TUTORIALS

OML LiveLab: https://apexapps.oracle.com/pls/apex/dbpm/r/livelabs/view-workshop?p180_id=560

OML4Py LiveLab: https://apexapps.oracle.com/pls/apex/dbpm/r/livelabs/view-workshop?w

Interactive tour: https://docs.oracle.com/en/cloud/paas/autonomous-database/oml-tour

ORACLE 机器学习文档

https://docs.oracle.com/en/database/oracle/machine-learning/index.html

OML4SQL文档： https://docs.oracle.com/en/database/oracle/machine-learning/oml4sql/index.html

OML4SQL示例： https://github.com/oracle/oracle-db-examples/tree/master/machine-learning/sql/21c

OML4Python文档： https://docs.oracle.com/en/database/oracle/machine-learning/oml4py/index.html

OML4R文档： https://docs.oracle.com/en/database/oracle/machine-learning/oml4r/index.html

Oracle Data Miner文档： https://docs.oracle.com/en/database/oracle/sql-developer/21.4/books.html

OML4Spark文档： https://docs.oracle.com/en/database/oracle/machine-learning/oml4spark/index.html

ORACLE 甲骨文

OCI 与云原生应用程序相关的容器服务介绍

数据库与云系列公益讲座

张鑫

- 高级解决方案工程师
- 10多年系统架构及开发经验
- 甲骨文云原生技术领域专家

内容简介

介绍甲骨文云上提供的多种容器服务（例如Kubernetes），以及各种容器服务适合的业务场景。

- 什么是云原生应用程序
- 甲骨文云上的多种容器服务
- 容器服务的选择

直播时间：1月20日 11:00 - 12:00

扫描二维码注册并安装手机Zoom进入直播

Zoom ID: 976 6962 5763 密码: 98039717

20-20

数据库和云讲座群

甲骨文云技术公众号

技术专家1V1深入交流