



ORACLE

# Платформа Data Science-as-a-Service на технологиях Oracle

Олег Сиротюк

Ведущий консультант Oracle

[Oleg.Sirotiyuk@oracle.com](mailto:Oleg.Sirotiyuk@oracle.com)

## Safe Harbor Statement

The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, timing and pricing of any features or functionality described for Oracle's products may change and remains at the sole discretion of Oracle Corporation.



**“Почему Oracle?  
Потому что там хранятся  
ваши данные!”**

---

**Larry Ellison**

Executive Chairman and CTO of Oracle Corporation

ORACLE

# Платформа Data Science-как-Сервис

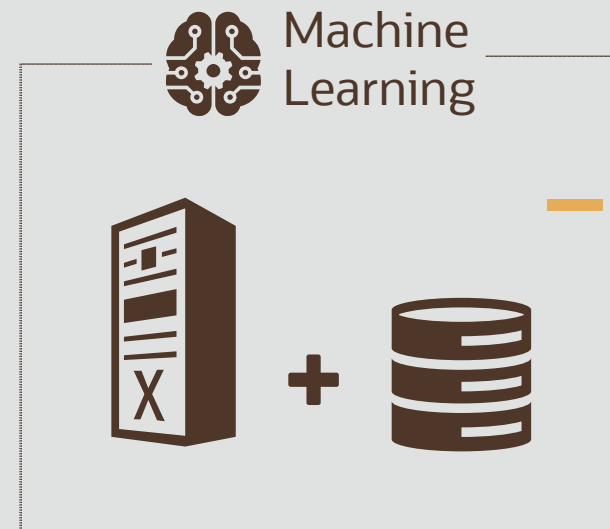
На стандартных технологиях Oracle

Производительность

Масштабируемость

Защита инвестиций

Безопасность



Только **1 из 10** проектов Data Science внедряется в промышленную эксплуатацию

**Неэффективная коллаборация** –  
ключевое препятствие для DS проектов (**>60%**)

**Кибербезопасность и защита данных** – наиболее актуальные разделы DS проектов

## Определение

---

Облачный сервис (частное или гибридное облако), предоставляющий экспертам необходимые инструменты, вычислительные ресурсы и данные для реализации проектов по машинному обучению и искусственному интеллекту

# Рецепт платформы DataScience – as a –Service от Oracle

## Целевые показатели

---

- |                             |                               |
|-----------------------------|-------------------------------|
| 1. Создание сервиса         | < 15 минут                    |
| 2. Сбор и подготовка данных | МИНИМУМ времени               |
| 3. Утилизация CPU           | > 70%                         |
| 4. Сжатие данных            | 5-ти кратное и более          |
| 5. Безопасность             | Все технологии СУБД<br>Oracle |

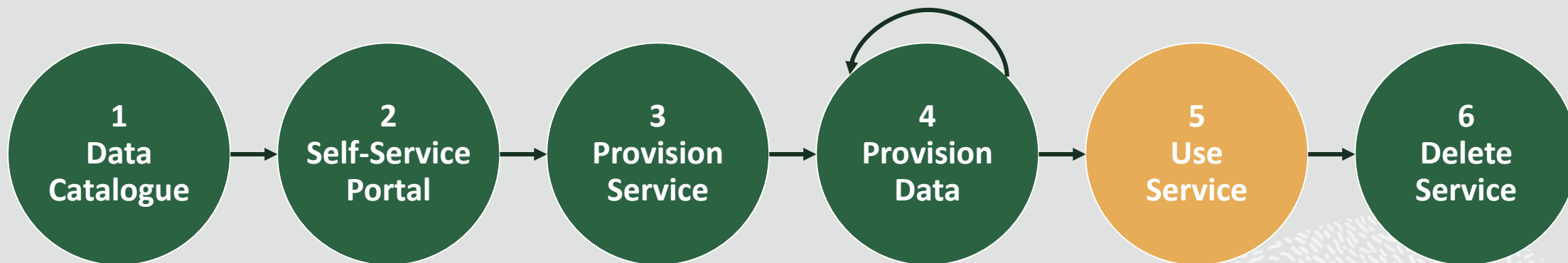
# Рецепт DataScience – as a –Service от Oracle

## Жизненный цикл сервиса

Владельцы данных публикуют информацию о доступных наборах данных в Каталоге

Платформа выделяет вычислительные мощности и место для данных

Эксперт используют созданный сервис из SQL инструментов, BI или ноутбуков



Эксперт использует Портал для запроса сервиса: требуемые вычислительные мощности, данные и способ их доставки

Платформа создает объекты БД (tables, views, external tables) и настраивает механизмы интеграции (Oracle Big Data SQL, ODI)

Сервис может быть удален через Портал или автоматически по истечению заданного срока годности

Собираем решение из имеющихся компонентов  
и подбираем необходимые технологии



Все! Сервисами можно пользоваться!

Уровень сервисов  
ORACLE MULTITENANT + OEM

In-Database ML / Python / R  
ORACLE ADVANCED ANALYTICS

Хранилище данных  
ORACLE DB EE

Аппаратная инфраструктура

# Собираем решение из имеющихся компонентов и подбираем необходимые технологии

---



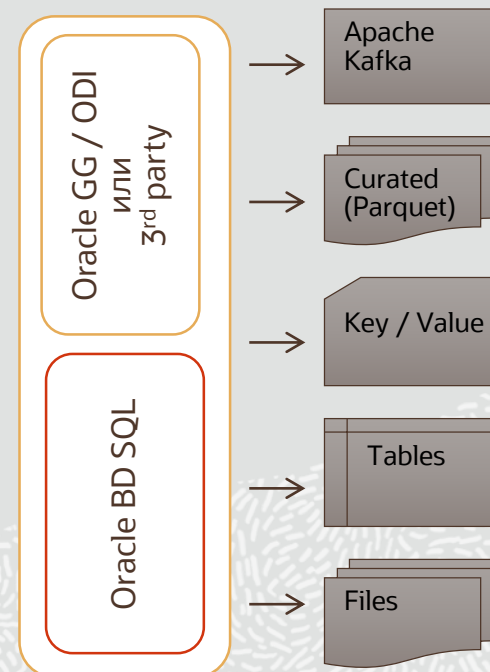
Уровень сервисов  
ORACLE MULTITENANT + OEM

In-Database ML / Python / R  
ORACLE ADVANCED ANALYTICS

Хранилище данных  
ORACLE DB EE

Аппаратная инфраструктура

# Собираем решение из имеющихся компонентов и добираем необходимые технологии



DATA LAKE

ORACLE

# Платформа Data Science-как-Сервис

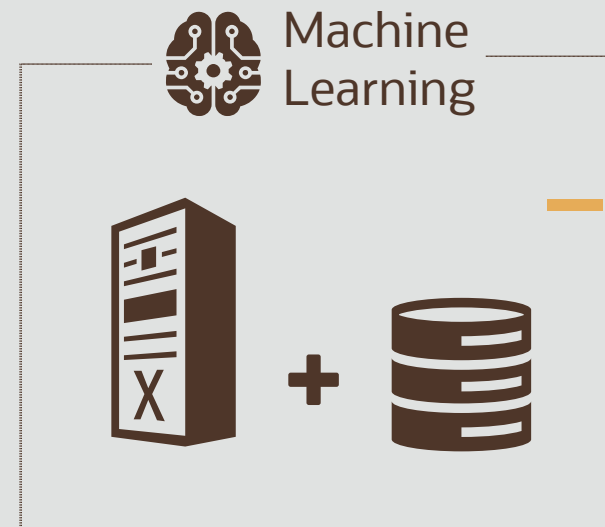
## Технологии Oracle

Производительность

Масштабируемость

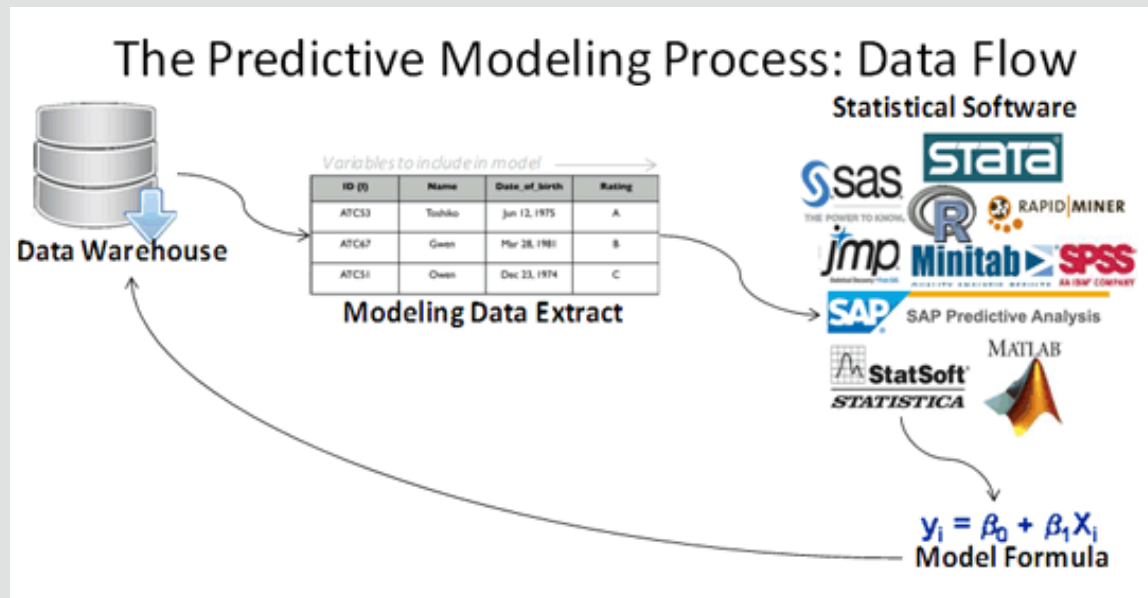
Защита инвестиций

Безопасность



# Традиционный vs. Oracle In-Database ML подходы

**Традиционный** — “Переносите данные”      **ORACLE®** — “Переносите алгоритмы”



Проще, умнее управление данными  
+ Аналитика / Machine Learning

# СУБД Oracle: In-Database Machine Learning

Опция **Oracle Advanced Analytics**

- In-Database реализация >30 наиболее популярных алгоритмов ML
- Встроенные механизмы подготовки данных
- Создание сложных моделей
  - SQL, R или Python
  - Oracle Data Miner
  - Oracle AutoML<sup>2</sup>
- Не требуется отдельная инфраструктура (ML выполняется на сервере БД Oracle)
- Масштабируемые, параллельные алгоритмы Data Mining в ядре SQL

## МИНИМИЗАЦИЯ ТСО

Не требуется дубликация данных

Не требуются выделенные сервера

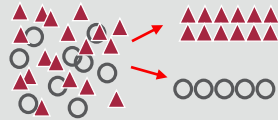
Защита сделанных инвестиций

Быстрый time-to-market

# Алгоритмы Oracle In-Database Machine Learning

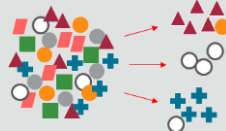
## CLASSIFICATION

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



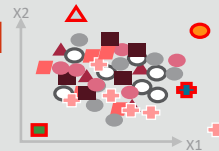
## CLUSTERING

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



## ANOMALY DETECTION

- One-Class SVM



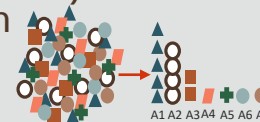
## TIME SERIES

- State of the art forecasting using Exponential Smoothing
- Includes all popular models e.g. Holt-Winters with trends, seasons, irregularity, missing data



## REGRESSION

- Linear Model
- Generalized Linear Model
- Support Vector Machine (SVM)
- Stepwise Linear regression
- Neural Network
- LASSO \*

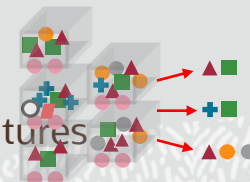


## ATTRIBUTE IMPORTANCE

- Minimum Description Length
- Principal Comp Analysis (PCA)
- Unsupervised Pair-wise KL Div
- CUR decomposition for row & A

## ASSOCIATION RULES

- A priori/ market basket



## PREDICTIVE QUERIES

- Predict, cluster, detect, features

## SQL ANALYTICS

- SQL Windows, SQL Patterns, SQL Aggregates

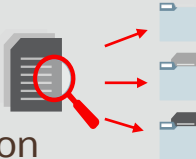


## FEATURE EXTRACTION

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

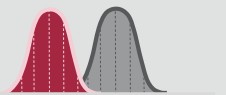
## TEXT MINING SUPPORT

- Algorithms support text
- Tokenization and theme extraction
- Explicit Semantic Analysis (ESA) for document similarity



## STATISTICAL FUNCTIONS

- Basic statistics: min, max, median, stdev, t-test, F-test, Pearson's, Chi-Sq, ANOVA, etc.



## R PACKAGES

- CRAN R Packages through Embedded R Execution
- Spark MLlib algorithm integration



## EXPORTABLE ML MODELS

- REST APIs for deployment



# Пример: Создание классификатора

## Создание ML модели (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;
```

## Применение модели (SQL запрос)

```
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_MONTHLY_OVERDRAWN, 1 as house_ownership)
from dual;
```

|                                            |                                                                                                       |
|--------------------------------------------|-------------------------------------------------------------------------------------------------------|
| SQL   All Rows Fetched: 1 in 0.043 seconds |                                                                                                       |
|                                            | PREDICTION_PROBABILITY(BUY_INSUR1,'YES'USING3500ASBANK_FUNDS,825ASCHECKING_AMOUNT,400ASCREDIT_BALANCE |
| 1                                          | 0.9276956709910801                                                                                    |

# Пример: Определение значимых атрибутов

## Создание ML модели (PL/SQL)

```
BEGIN
  DBMS_DATA_MINING.CREATE_MODEL(
    model_name      => 'BUY_INSURANCE_AI',
    mining_function  => DBMS_DATA_MINING.ATTRIBUTE_IMPORTANCE,
    data_table_name  => 'CUST_INSUR_LTV',
    case_id_column_name => 'cust_id',
    target_column_name  => 'BUY_INSURANCE',
    settings_table_name => 'Att_Import_Mode_Settings');
END;
```

## Применение модели (запрос SQL)

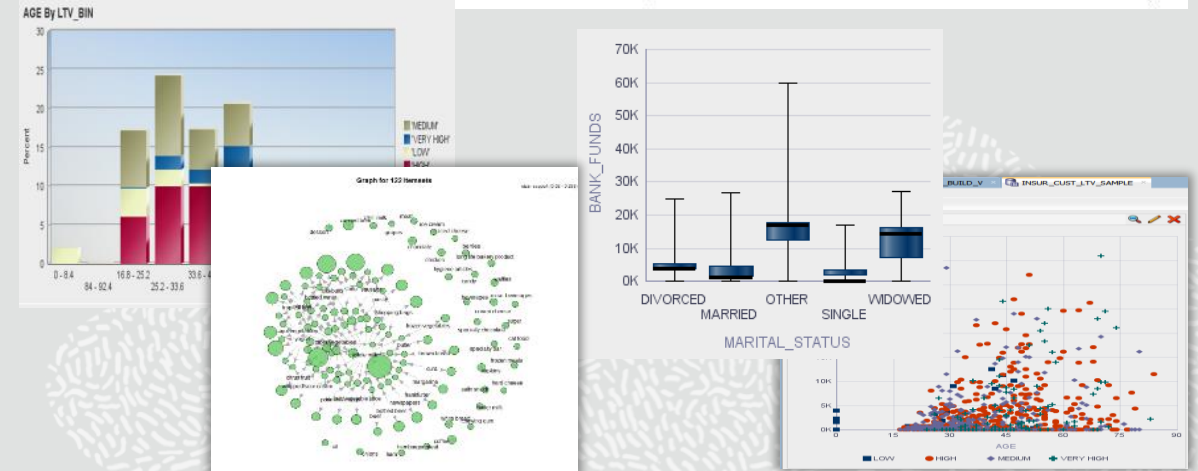
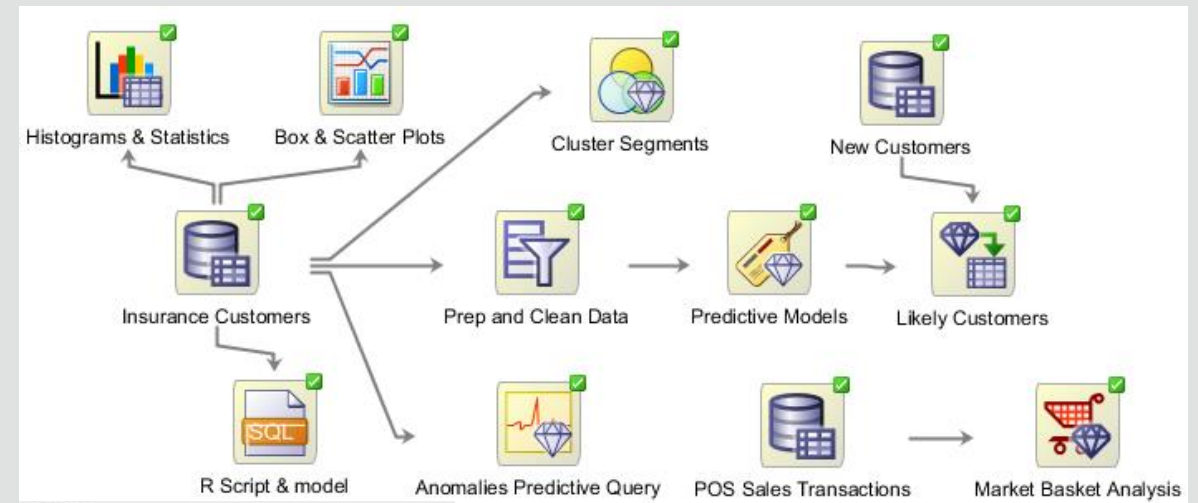
```
SELECT attribute_name, explanatory_value, rank
FROM BUY_INSURANCE_AI
ORDER BY rank, attribute_name;
```

| ATTRIBUTE_NAME          | RANK | ATTRIBUTE_VALUE |
|-------------------------|------|-----------------|
| BANK_FUNDS              | 1    | 0.2161          |
| MONEY_MONTHLY_OVERDRAWN | 2    | 0.1489          |
| N_TRANS_ATM             | 3    | 0.1463          |
| N_TRANS_TELLER          | 4    | 0.1156          |
| T_AMOUNT_AUTOM_PAYMENTS | 5    | 0.1095          |

# Графический конструктор для проектов по ML

## Oracle Data Miner

- Построение модели drag / drop
- Не требуется глубоких познаний в ML
- Легко в использовании, не надо кодировать
- Определение разделяемых аналитических процессов
- Генерация SQL кода



# Автономный ML

## Oracle AutoML<sup>2</sup> \*

- Использует ML для создания новых моделей ML
- Существенное сокращение времени создания моделей
- Существенное увеличение точности
- Не требуется специальных знаний в ML для создания продуктивных и точных моделей



Training  
Data

### AutoML<sup>2</sup>

**Automatic  
Feature Selection**

~50% reduction in features

**Automatic  
Model Selection**

4x faster than exhaustive search

**Automatic  
Hyperparameter Tuning**

As much as 24% accuracy improvement

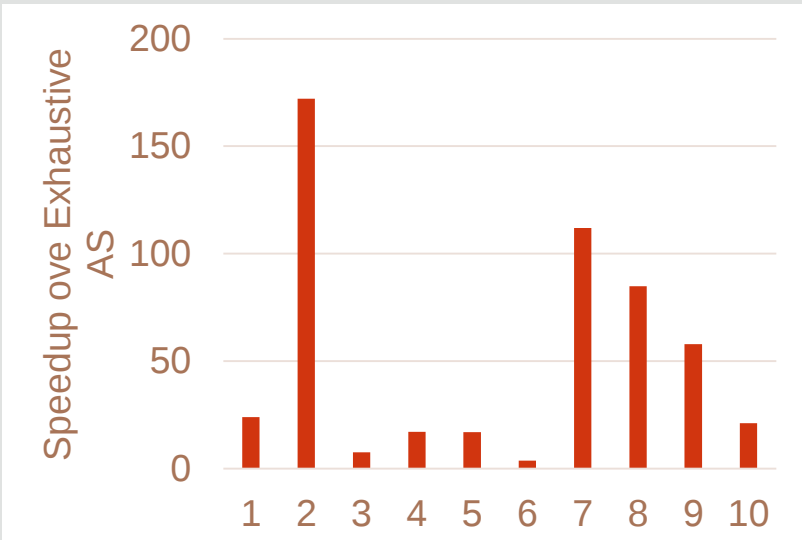
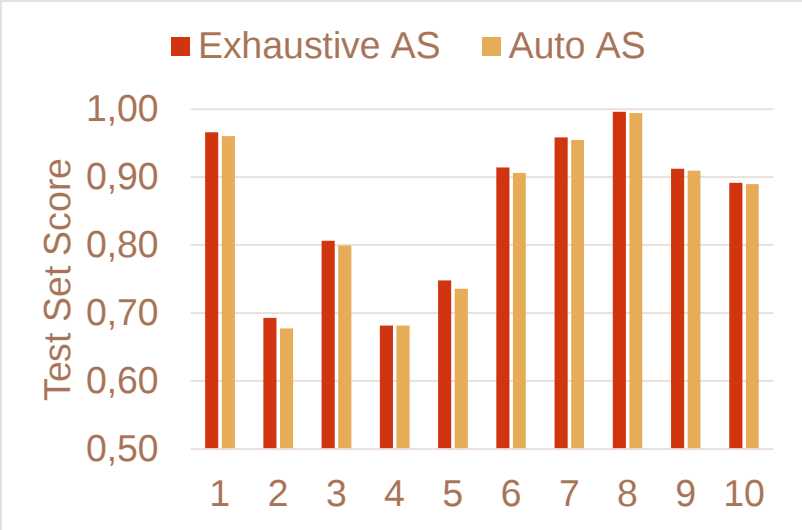


Production-Ready  
ML Model

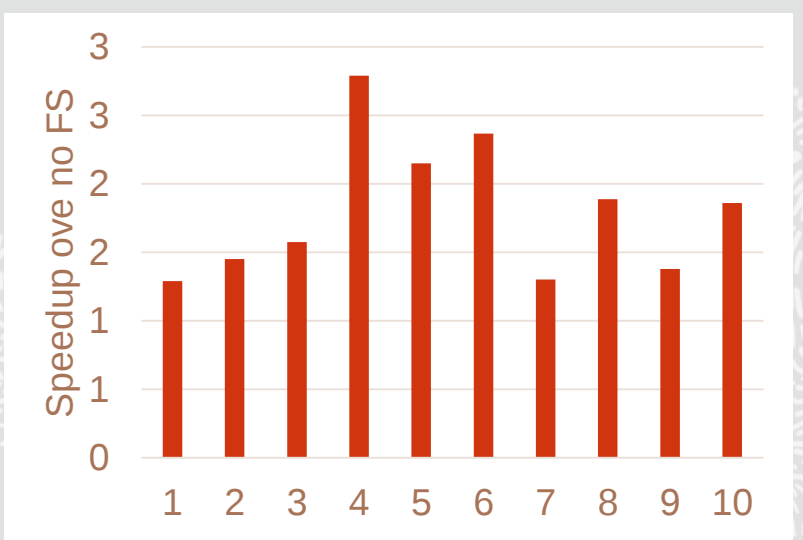
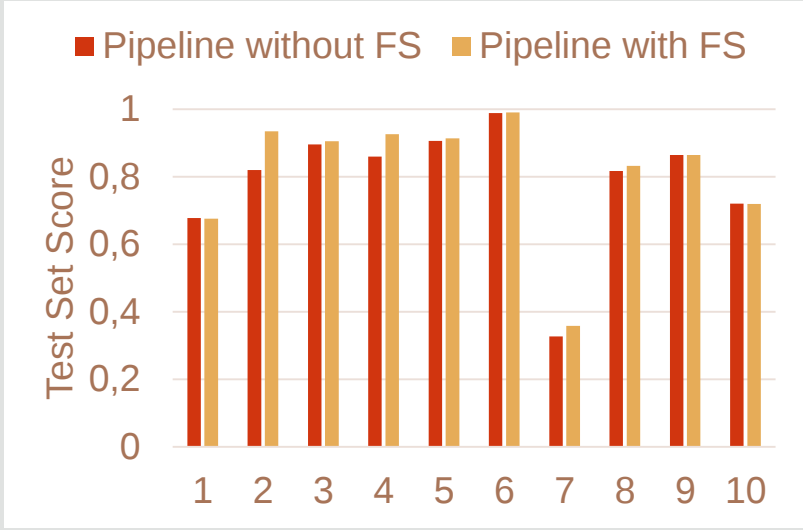
\*Planned for Q1 2020 with Oracle Machine Learning for Python component of Oracle Advanced Analytics Option

# Oracle AutoML в действии

Автоматический выбор алгоритма



Автоматический выбор атрибутов



# Варианты использования Oracle AutoML

## Python

### Классификация

```
from automl import Pipeline

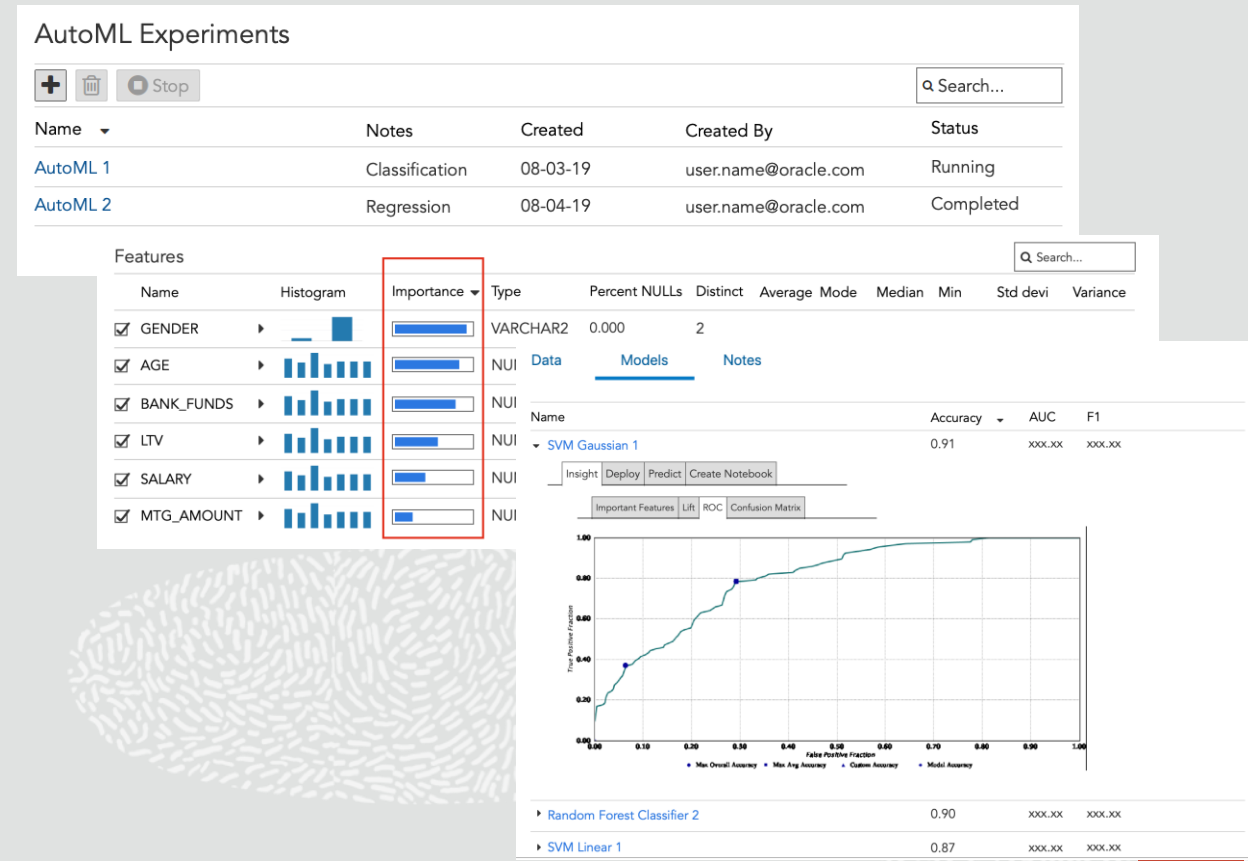
est = Pipeline(task='classification',
               scoring='accuracy')
est.fit(X_train, y_train)
y_pred = est.predict(X_test)
```

### Регрессия

```
from automl import Pipeline

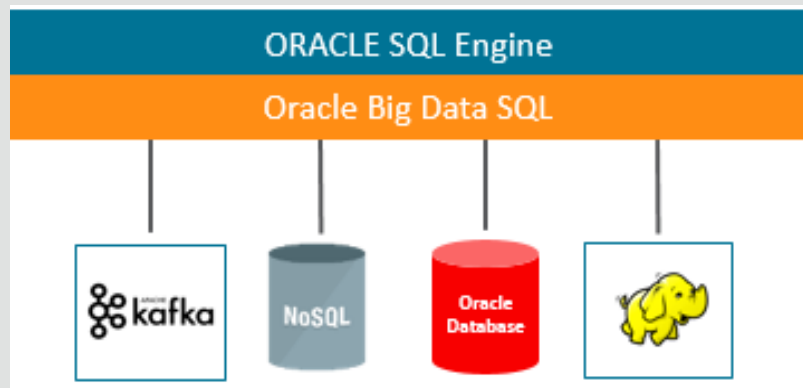
est = Pipeline(task='regression',
               scoring='neg_mean_squared_error')
est.fit(X_train, y_train)
y_pred = pipe.predict(X_test)
```

## GUI



# Работаем с данными из источников BigData технология Oracle BigData SQL

## ORACLE SQL ДЛЯ BIGDATA

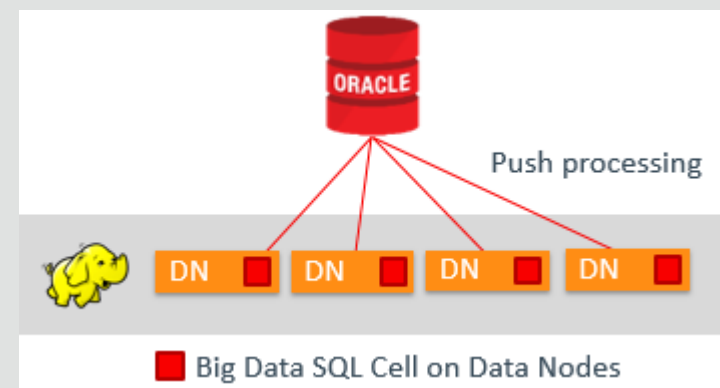


- Анализ и сопоставление данных из гетерогенных источников
- Не требуется менять запросы приложений

## ORACLE DB SECURITY В HADOOP ПРОИЗВОДИТЕЛЬНОСТЬ



- Унифицированные политики безопасности для BigData и БД
- Можно применять функции, такие как data redaction для связки (join) данных из различных источников



- SmartScan на уровне Hadoop
- Фильтрация, агрегация, исполнение функций на узлах данных



# **Изменяющаяся роль DBA: От разработчика БД к Data Scientist за 6 недель!**



**Мы находимся на заре интеллектуальной,  
автономной эпохи, и наличие базы данных  
с самостоятельным управлением -  
естественный прогресс ...**

**Я чувствую, что автономные базы данных  
станут повсеместными в будущем.”**

---

– Clark A. Kho

Senior Technology Architect, Accenture

# Автономная база данных избавляет от **рутины**

Больше времени для инноваций и улучшения бизнеса

- **Задачи, специфичные для бизнеса и инноваций**

- Архитектура, планирование, моделирование данных
- Безопасность данных и управление жизненным циклом
- Тюнинг приложений
- Комплексное управление уровнем обслуживания



- ~~**Техническое обслуживание**~~

- ~~– Конфигурирование и настройка систем, сети, хранилища~~
- ~~– Подготовка базы данных, патчинг~~
- ~~– Резервное копирование базы данных, Н / А, аварийное восстановление~~
- ~~– Оптимизация базы данных~~



## Шкала ценности



Инновации

Обслуживание

# Эволюция разработчика БД в Эксперта по данным

Они уже выполняют большую часть работ

Выгрузка данных

Подготовка данных

Выявление новых атрибутов  
("feature engineering")

...  
...

тут происходит «магия» Machine Learning



Импорт результатов

Внедрение ML

Автоматизация

До 80% проекта

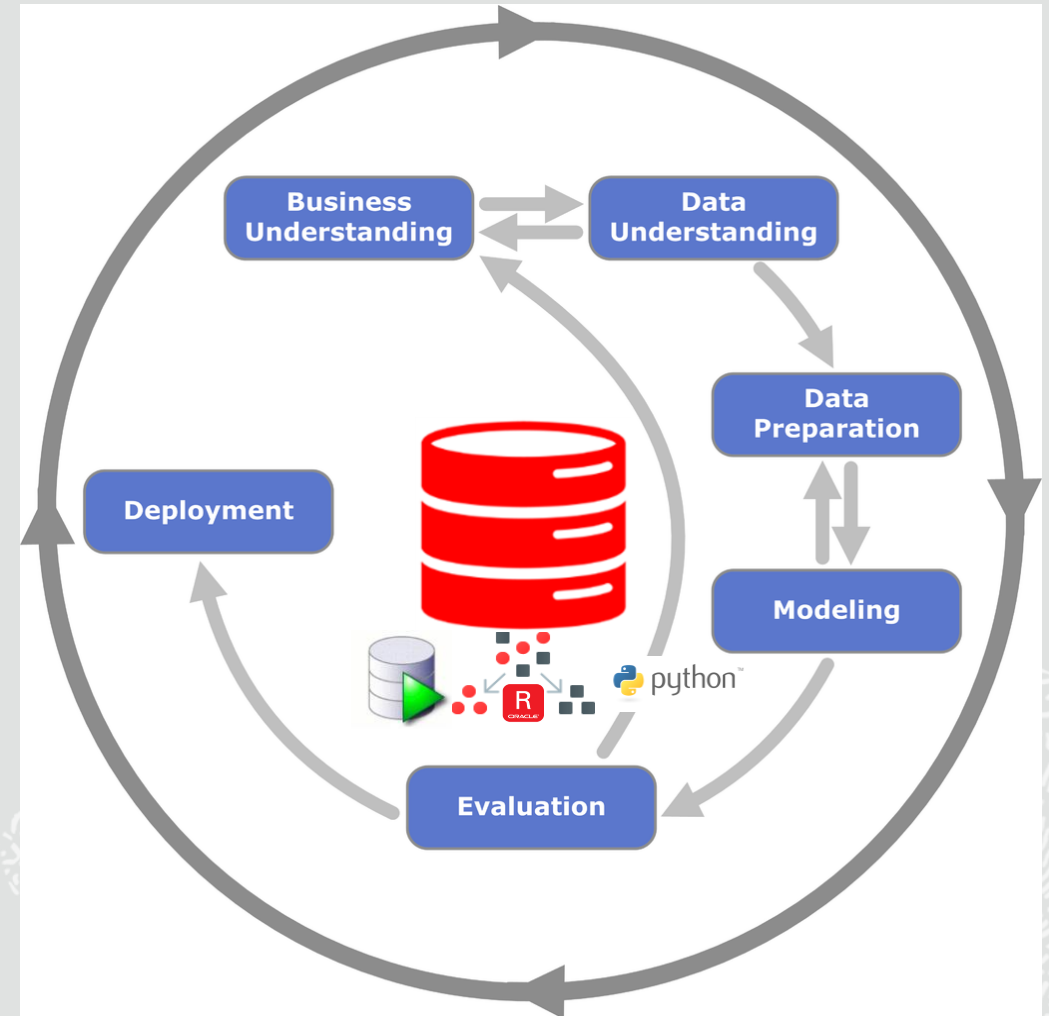
Большинство экспертов тратят только 20% времени на анализ данных и до 80% времени - на поиск, очистку и реорганизацию данных<sup>1</sup>

Исключено или минимизированно Oracle

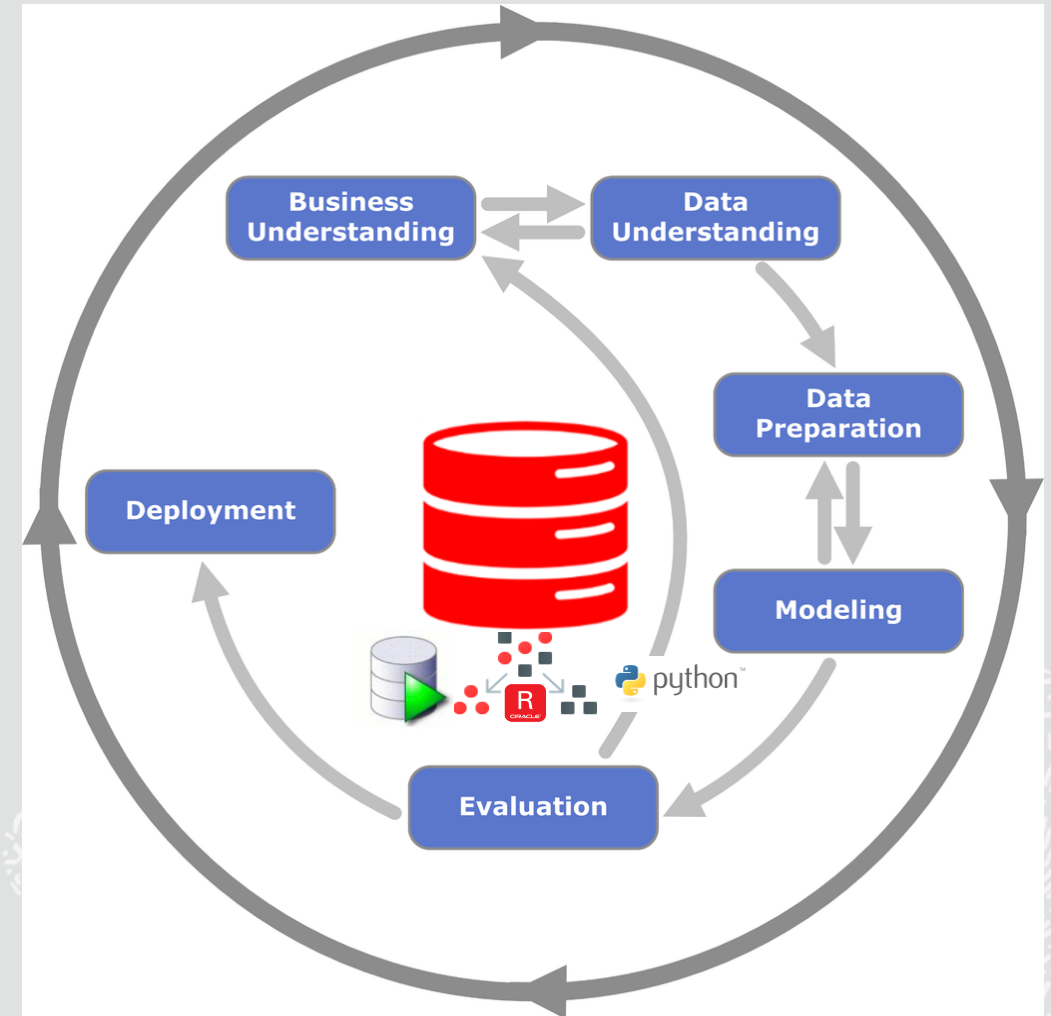
Платформа управления данными становится платформой для проектов машинного обучения

# От разработчика БД к Data Scientist за 6 недель!

- Исследование бизнеса — Неделя 1
- Понимание данных — Неделя 2
- Подготовка данных — Неделя 3
- Моделирование (ML)— Неделя 4
- Оценка — Неделя 5
- Развертывание — Неделя 6



# От разработчика БД к Data Scientist за 6 недель!



ORACLE

# Облако Oracle

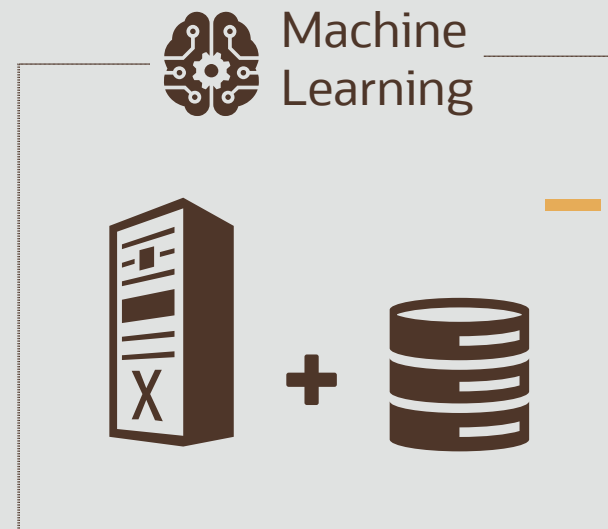
Используйте облако Oracle для расчета трудоемких моделей

Автономность

Масштабируемость

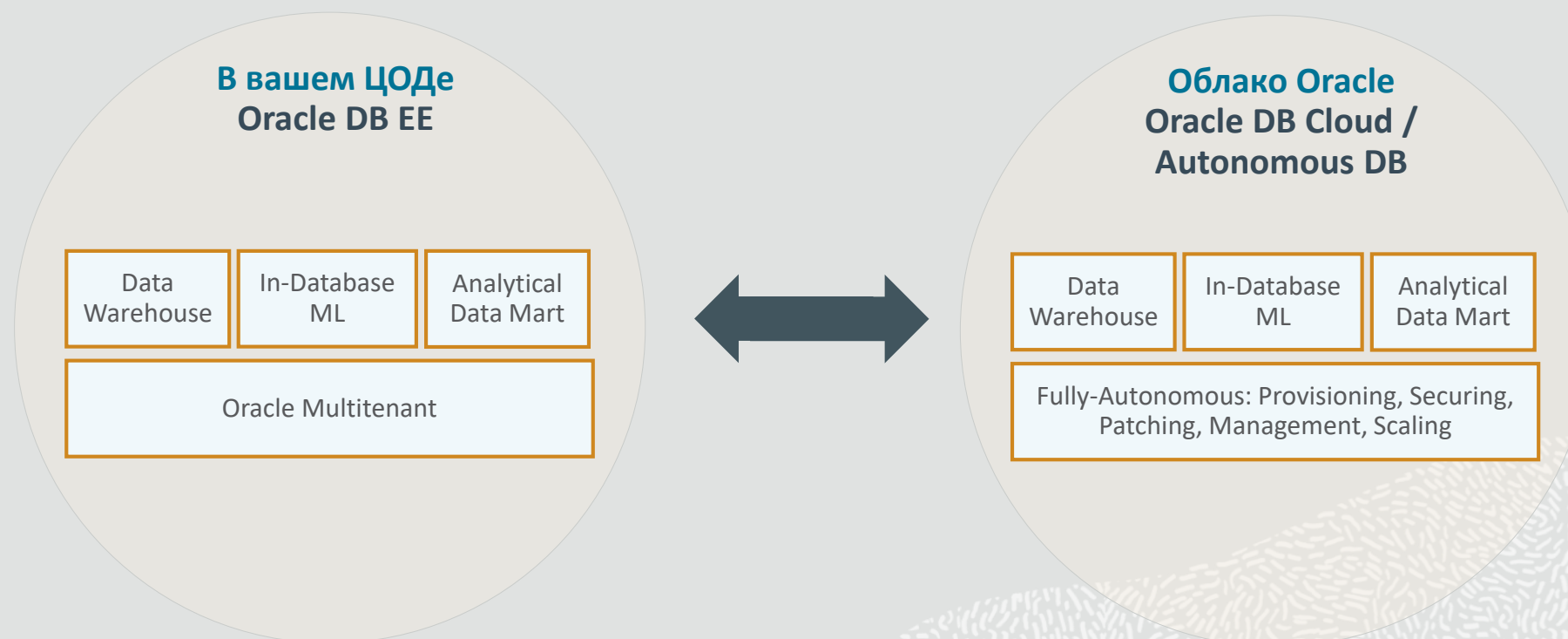
Производительность

Экономика



# Oracle ML аналогичный в Вашем ЦОДе и в облаке Oracle

## Используем HPC облако Oracle для расчета ресурсоемких моделей



# Oracle In-Database ML в нойтбуке Oracle Autonomous DataWarehouse Cloud

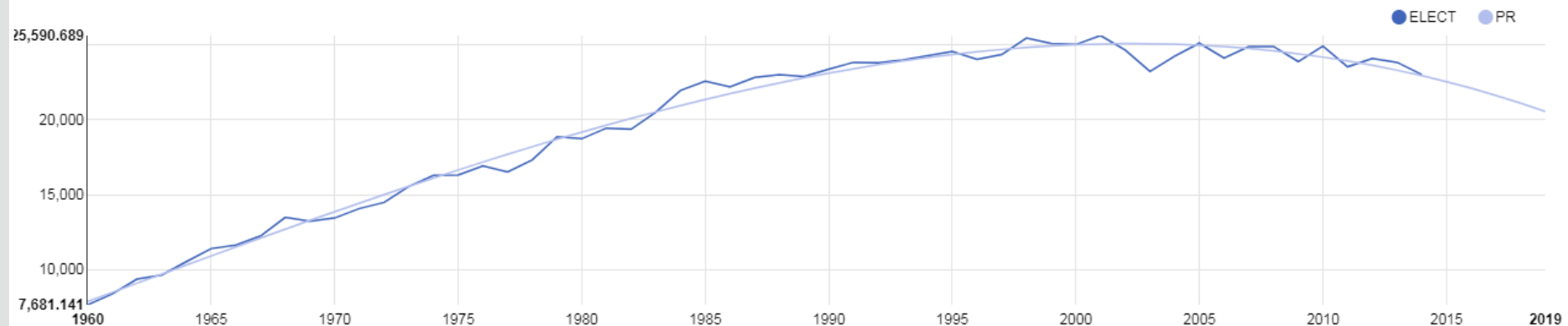
```
%script
BEGIN
/* Populate settings table */
INSERT INTO glmr_sh_sample_settings (setting_name, setting_value) VALUES
(dbms_data_mining.algo_name, dbms_data_mining.algo_generalized_linear_model);
-- output row diagnostic statistics into a table named GLMC_SH_SAMPLE_DIAG
INSERT INTO glmr_sh_sample_settings (setting_name, setting_value) VALUES
(dbms_data_mining.glms_diagnostics_table_name, 'GLMR_SH_SAMPLE_DIAG');
INSERT INTO glmr_sh_sample_settings (setting_name, setting_value) VALUES
(dbms_data_mining.prep_auto, dbms_data_mining.prep_auto_on);

-- turn on feature selection
INSERT INTO glmr_sh_sample_settings (setting_name, setting_value) VALUES
(dbms_data_mining.glms_ftr_selection,
dbms_data_mining.glms_ftr_selection_enable);

-- turn on feature generation
INSERT INTO glmr_sh_sample_settings (setting_name, setting_value) VALUES
(dbms_data_mining.glms_gen,
dbms_data_mining.glms_gen_enable);

%sql
SELECT caseID, elect, year,
       PREDICTION(GLMR_SH_Regr_sample USING *) pr
FROM usa_elec_cons ORDER BY year;
END;
```

settings



```
%script
/* Clean up any previous GLM Models for notebook repeatability */
BEGIN
  DBMS_DATA_MINING.DROP_MODEL('usa_elec_cons_model');
EXCEPTION WHEN OTHERS THEN NULL;
END;
/

declare
v_xlst dbms_data_mining_transform.TRANSFORM_LIST;
BEGIN
  DBMS_DATA_MINING.CREATE_MODEL(
    model_name      => 'usa_elec_cons_model',
    mining_function  => dbms_data_mining.regression,
    data_table_name  => 'usa_elec_cons',
    case_id_column_name => 'caseId',
    target_column_name => 'elect',
    settings_table_name => 'glmr_sh_sample_settings',
    xform_list       => v_xlst);
END;
```



SAVE THE DATE

# ANALYTICS AND DATA SUMMIT 2020

All Analytics. All Data.  
No Nonsense.

February 25-27, 2020

---

Call for Speakers Now Open!

---

JOIN OUR USER COMMUNITY (FREE)

Formerly the BIWA Summit with the Spatial and Graph Summit.

@AnalyticAndData

