

# Machine Learning com SQL e Oracle Autonomous Database

Erika Nagamine

Trilha Inovação com dados em nuvem  
03.12.2020 12h10



Este trabalho está licenciado sob uma Licença Creative Commons Atribuição-Compartilhável 4.0 Internacional. Para ver uma cópia desta licença, visite <http://creativecommons.org/licenses/by-sa/4.0/>.



# Declaração de Porto Seguro

A informação a seguir tem como objetivo traçar a orientação dos nossos produtos em geral. É destinada somente a fins informativos e não pode ser incorporada a um contrato. Ela não representa um compromisso de entrega de qualquer tipo de material, código ou funcionalidade e não deve ser considerado em decisões de compra. O desenvolvimento, a liberação, a data de disponibilidade e a precificação de quaisquer funcionalidades ou recursos descritos para produtos da Oracle estão sujeitos a mudanças e são de critério exclusivo da Oracle Corporation. Caso tenha dúvidas, entre em contato com o representante de vendas da Oracle.



\$> whoami

# Erika Nagamine

Principal Cloud Solutions Engineer

Data Management | Data Engineer | Data Scientist | Analytics  
Tech Cloud Solutions Engineer – Oracle LAD - Brazil



<https://linkedin.com/in/erikanagamine>



[erika.nagamine@oracle.com](mailto:erika.nagamine@oracle.com)



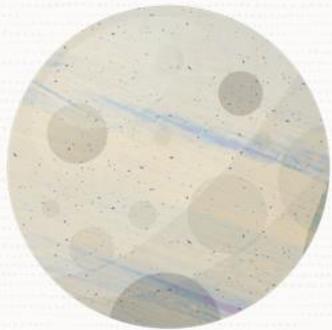
<https://github.com/erikanagamine>



@erikanagamine



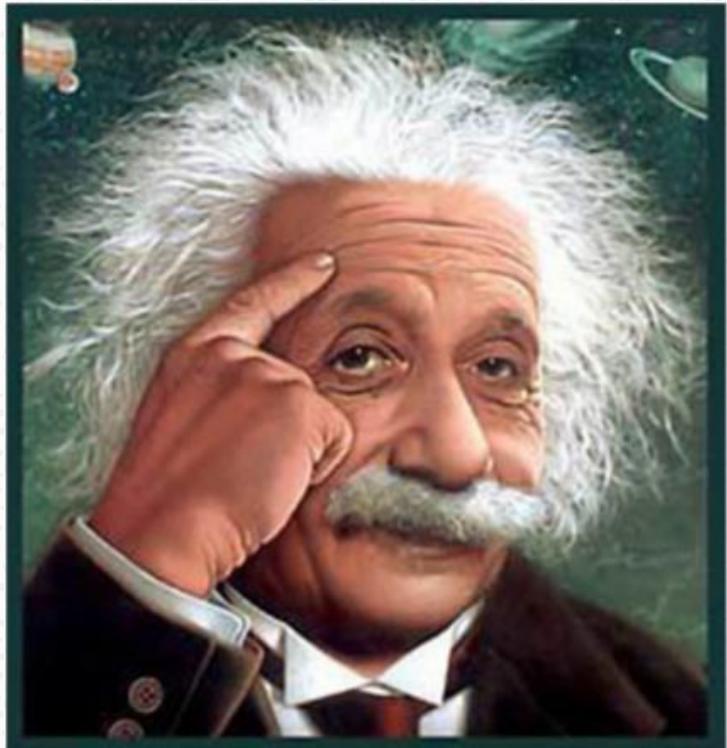
@erikanagamine



# Dados estão mudando o mundo

---





**“Se eu tiver **uma hora** para resolver um problema, eu gastaria **55 minutos** pensando no problema e **5 minutos** na **solução**”**

---

**Albert Einstein**

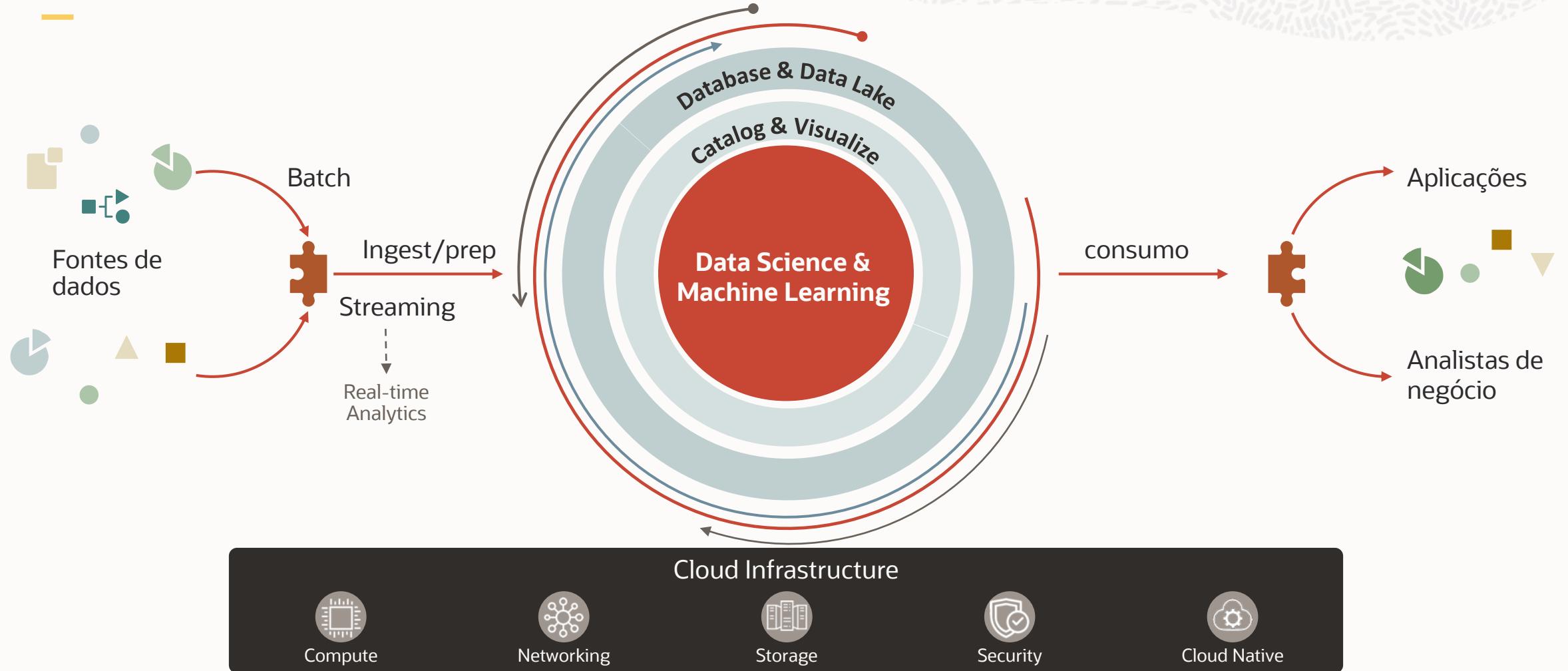
# Do-it-yourself cloud

Construa o que você quer,  
mas faça você mesmo



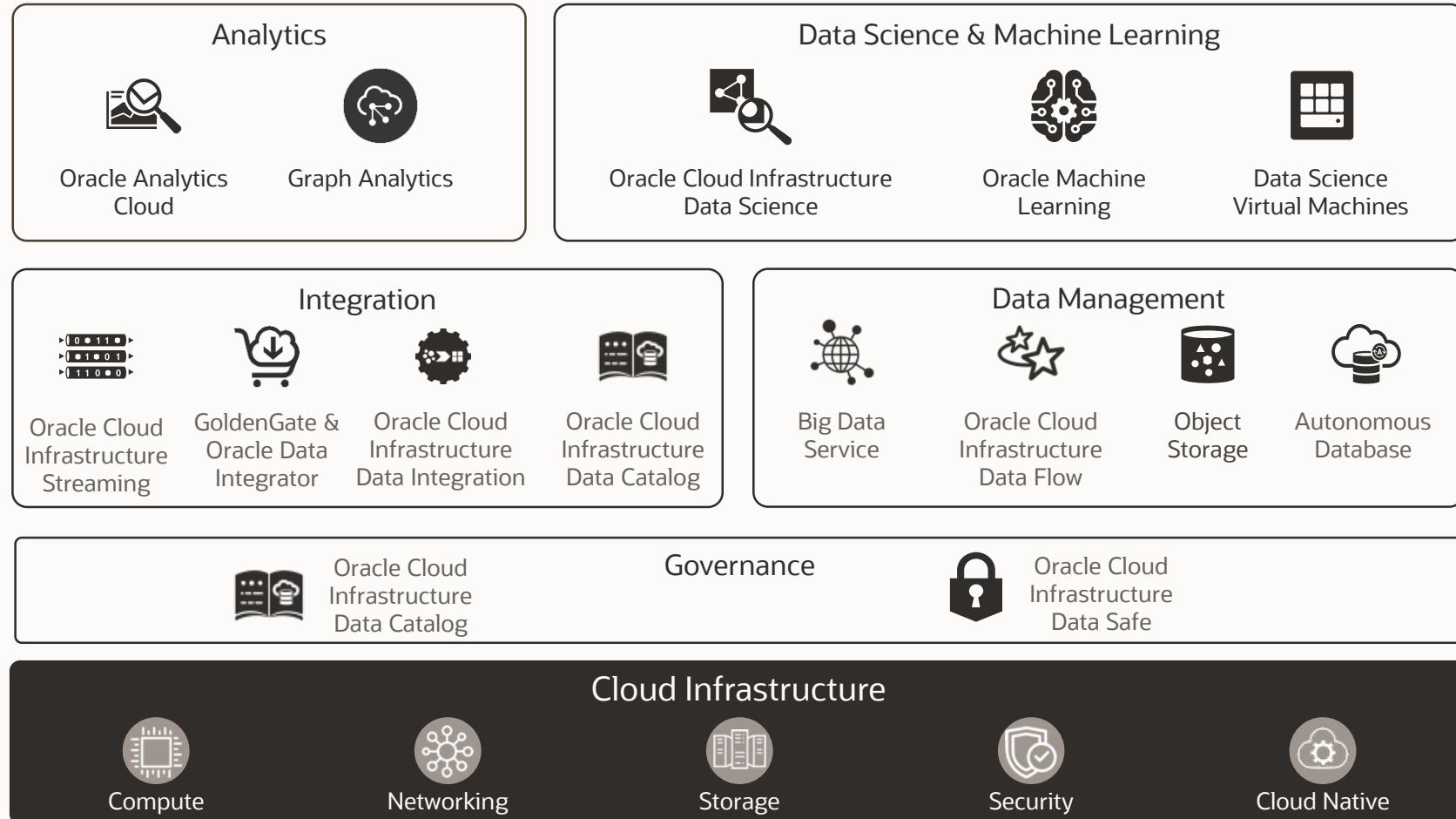
# Oracle Data Platform

Machine learning suportado por dados, integração, gerenciamento e analytics



# Oracle Data Platform

Machine learning suportado por dados, ingestão, gerenciamento e analytics



Aplicações

Analistas de negócio

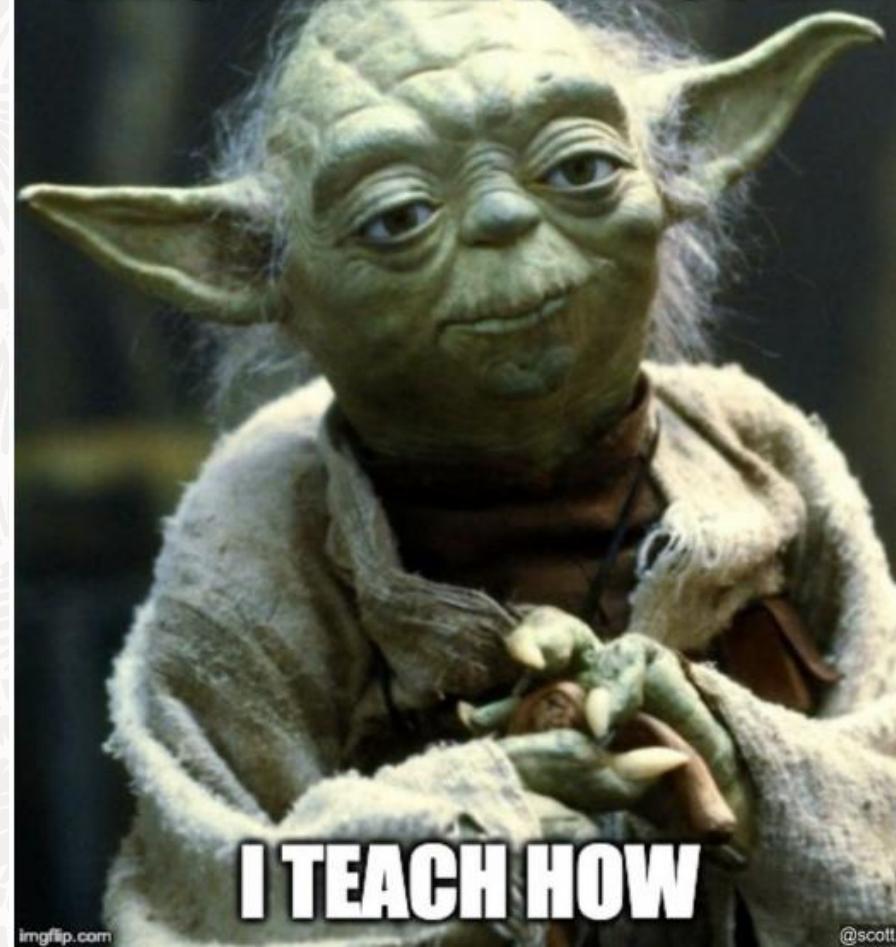
Fonte de dados



# Machine Learning

---

MASTER THE ART OF ML



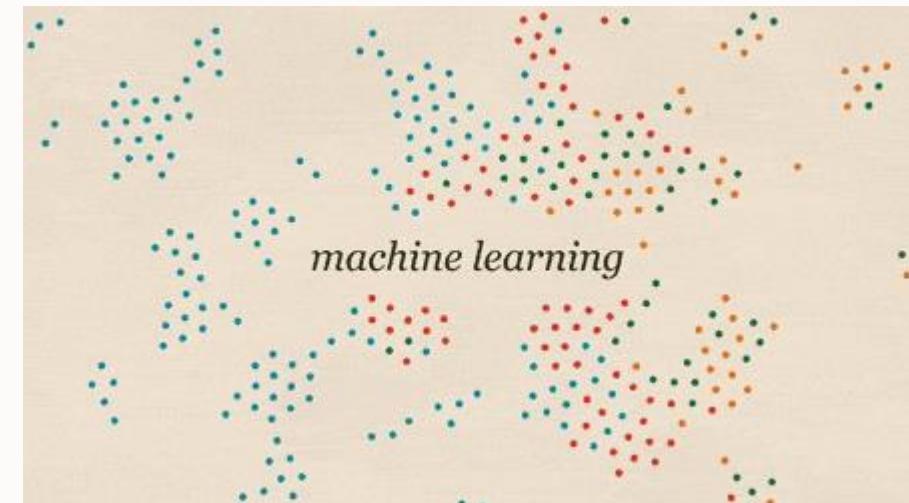
imgflip.com

@scott.ai

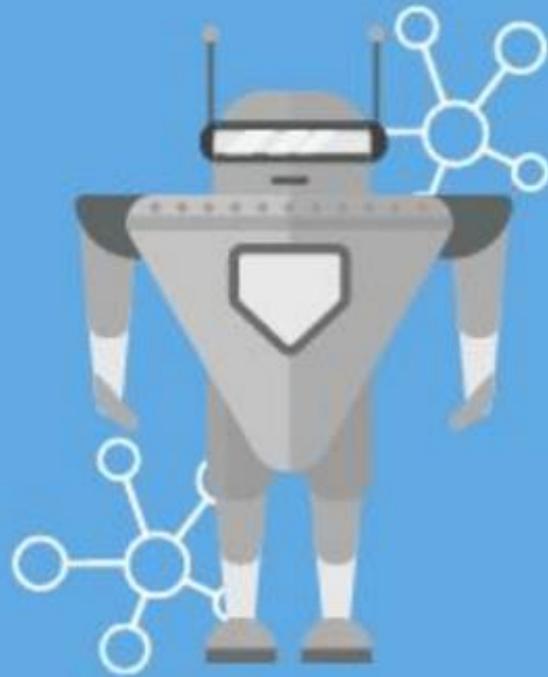
# Machine Learning



“Um programa de computador que **aprende a partir de uma experiência E** com respeito a uma tarefa **T** e uma métrica de performance **P**, se a sua performance em **T**, medida por **P**, melhora com a experiência **E**.”— [ Tom Mitchell, 1997 ]

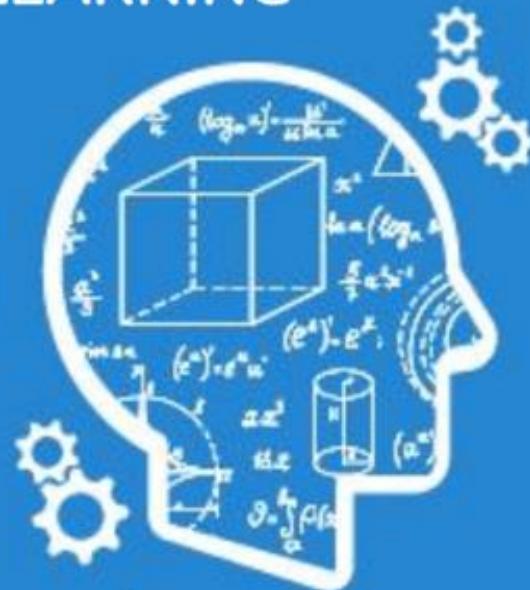


# ARTIFICIAL INTELLIGENCE



1950      1960      1970

# MACHINE LEARNING



1980      1990      2000      2010

# DEEP LEARNING





virtualtear.tumblr.com



0101010101  
0101010101  
0100101010

11100101010  
1001010010  
1001010100

0010101010  
0101010100  
0101000000



plot.py x

plot.py ▾ ...

Run Cell | Run Below

```
1 #%% [markdown]
2 # ## Plots
3 # ##### Examples from: https://matplotlib.org/3.1.0/gallery/
4
5 Run Cell | Run Above | Run Below
6 #%%
7 import matplotlib.pyplot as plt
8
9 Run Cell | Run Above | Run Below
10 #%%
11 data = {'apples': 10, 'oranges': 15, 'lemons': 5, 'lime': 3}
12 names = list(data.keys())
13 values = list(data.values())
14
15 fig, axs = plt.subplots(1, 3, figsize=(9, 3), sharey=True)
16 axs[0].bar(names, values)
17 axs[1].scatter(names, values)
18 axs[2].plot(names, values)
19 fig.suptitle('Categorical Plotting')
20
21 Run Cell | Run Above | Run Below
22 #%%
23 cat = ["bored", "happy", "bored", "bored", "happy", "bored"]
24 dog = ["happy", "happy", "happy", "happy", "bored", "bored"]
25 activity = ["combing", "drinking", "feeding", "napping", "bored"]
```



Python Interactive x



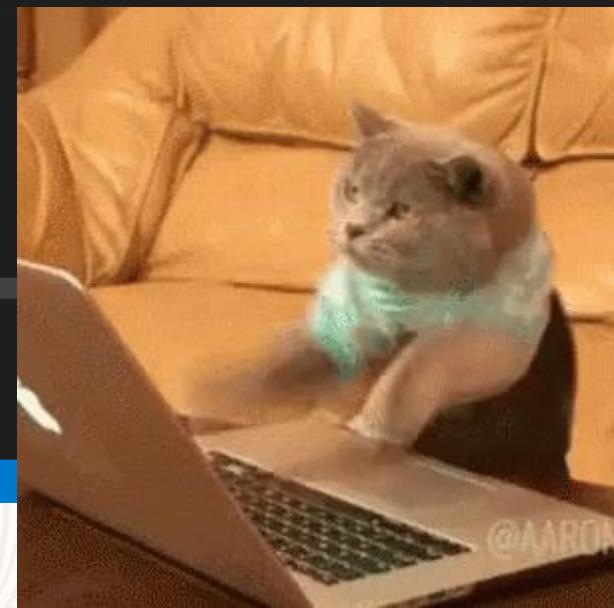
## ▶ Variables

Jupyter Server URI: http://localhost:8889/?  
token=476e9fc714196f2d625713b5706518052de368a0b85aeb63  
Python version:  
3.7.3 (v3.7.3:ef4ec6ed12, Mar 25 2019, 22:22:05) [MSC v.1916 64 bit (AMD64)]  
(5, 7, 2)  
C:\Users\luabud\AppData\Local\Programs\Python\Python37\pyt

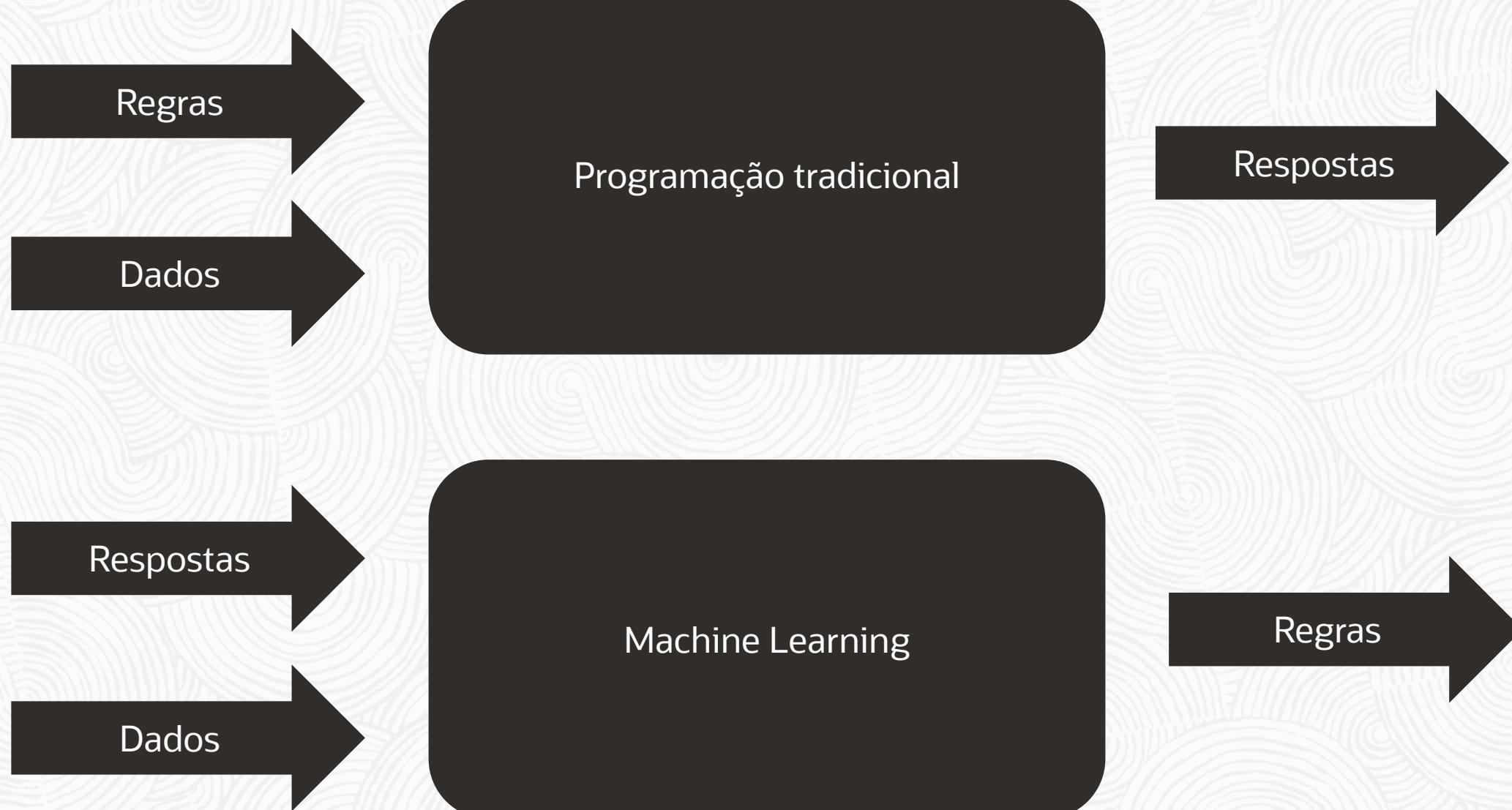
## Plots

## Examples from:

[https://matplotlib.org/3.1.0/gallery/lines\\_bars\\_and\\_markers/categorical\\_variables.html](https://matplotlib.org/3.1.0/gallery/lines_bars_and_markers/categorical_variables.html)



[1] Shift-enter to run





0101010101  
0101010101  
0100101010

11100101010  
1001010010  
1001010100

0010101010  
0101010100  
0101000000



## Características (Features):

Posição

Tom de pele

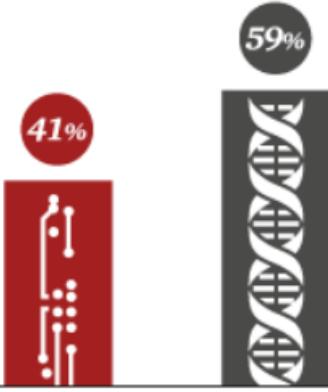
Qualidade da fotografia



# Algoritmos são melhores quando usam dados

O que não elimina o julgamento humano

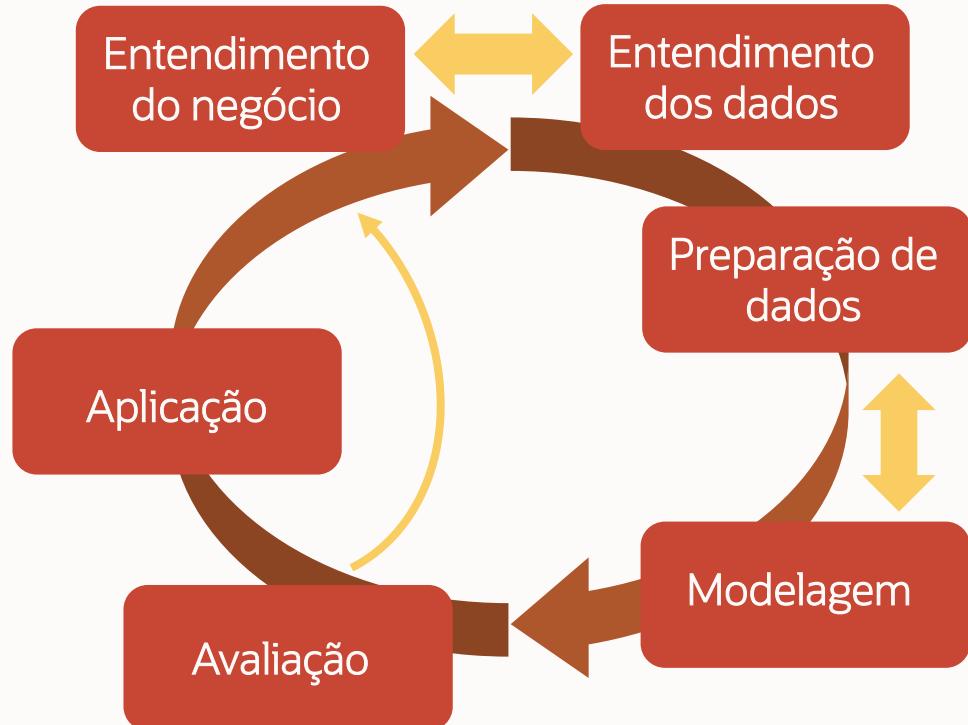
Machine algorithms  Human judgment



Source: PwC's Global Data and Analytics Survey, July 2016.  
Q: What will the analysis informing your next strategic decision require?  
Global base: 2,106 senior executives.

# Processo de Machine Learning

## CRISP-DM – Metodologia mais citada



Cross-industry standard process for data mining

### Oportunidades de automação

Preparação de dados\*  
Modelagem  
Avaliação  
Aplicação

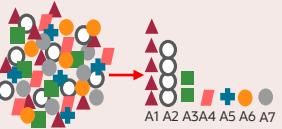
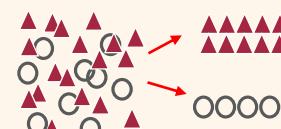
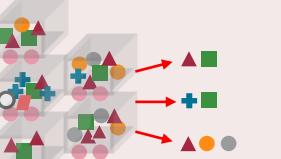
### Desafios de automação

Preparação de dados\*  
Entendimento do negócio  
Entendimento dos dados

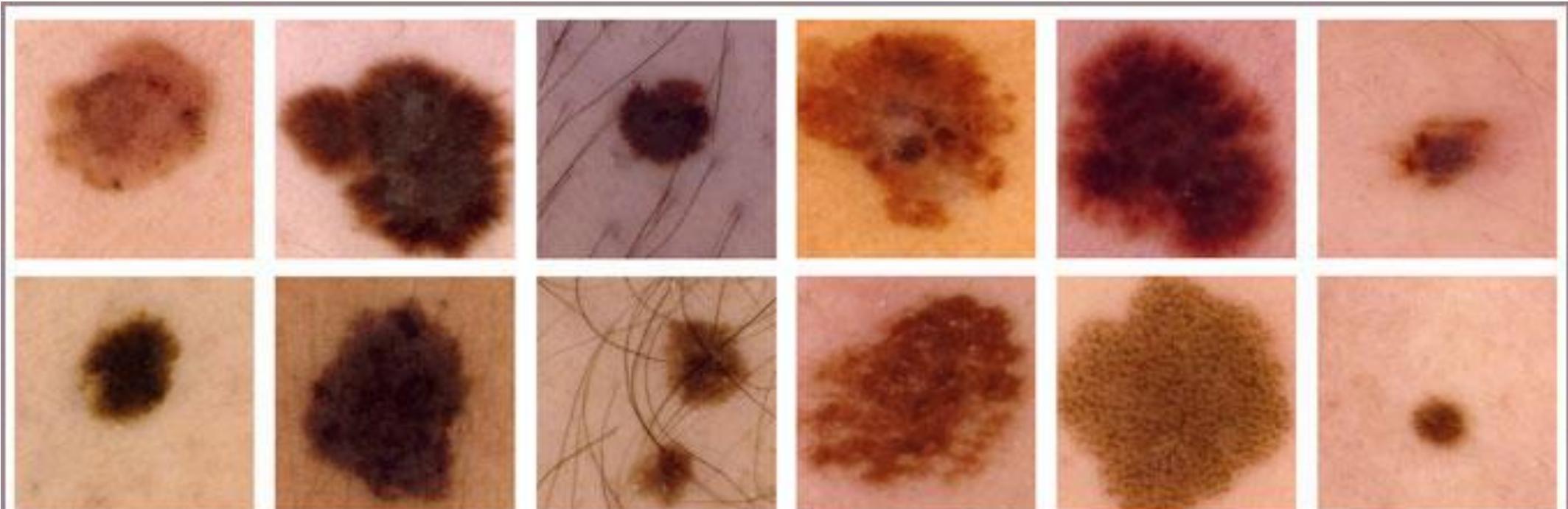
\*Depende do tipo de dados. Tendências baseadas no tempo são um desafio, a criação de característica é difícil sem interpretação

# Casos de uso para Machine Learning

Algoritmos *automaticamente* filtram grandes quantidades de dados para descobrir padrões, extrair novas idéias e fazer previsões

Quais são atributos que explicam o padrão de compra dos clientes?	Identificação / predição dos melhores clientes	Previsão de vendas / pedidos nos meses futuros	Identifica o melhor cliente usando modelo RFM (Recency, Frequency, Monetary)	Identifica possíveis atividades fraudulentas	Sugere itens adicionais ao cliente baseado no consumo
Attribute Importance	Classification	Regression	Clustering	Anomaly Detection	Associations
					
Identifica o fator mais importante que explica o evento	Prediz o sentimento do cliente e encontra padrões de consumo	Prediz ou estima um valor	Segmenta uma população em silos	Encontra outlier or “eventos raros”	Determina a ocorrência similar de consumo

## Outro exemplo prático de como algoritmos podem ajudar no dia a dia



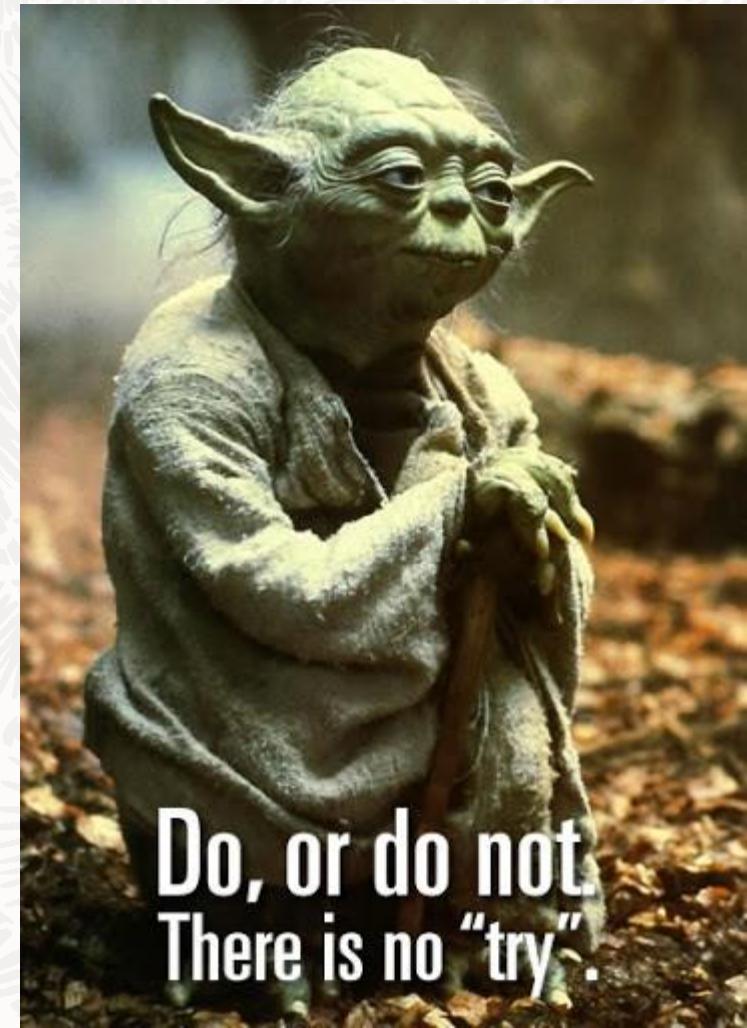
A triagem automática do melanoma é uma tarefa desafiadora pois a diferença entre a lesão cancerosa (linha superior) e a não cancerosa (linha inferior) muitas vezes é sutil

Fonte: <https://g1.globo.com/sp/campinas-regiao/noticia/2020/01/27/unicamp-desenvolve-software-que-permite-86percent-de-precisao-no-diagnostico-do-cancer-de-pele.ghtml>

# Workshop

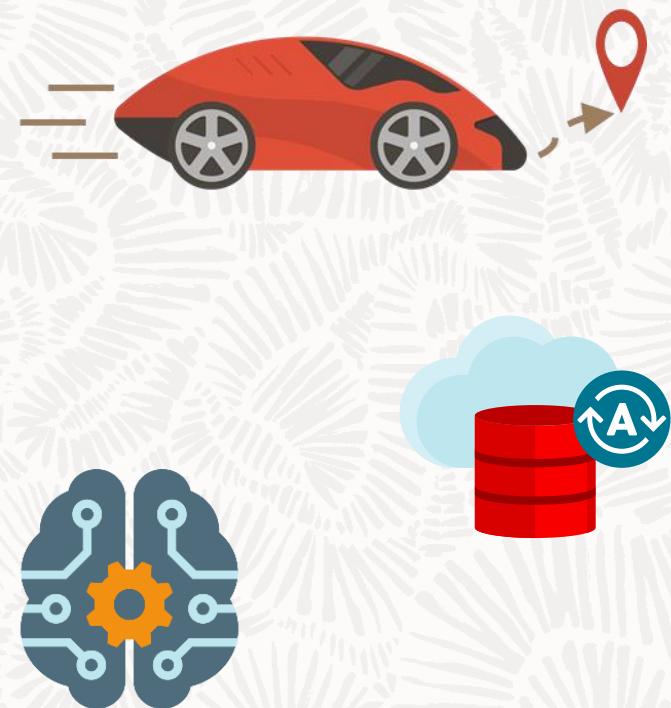
---

O que utilizaremos?



# Autonomous Database + OML

---



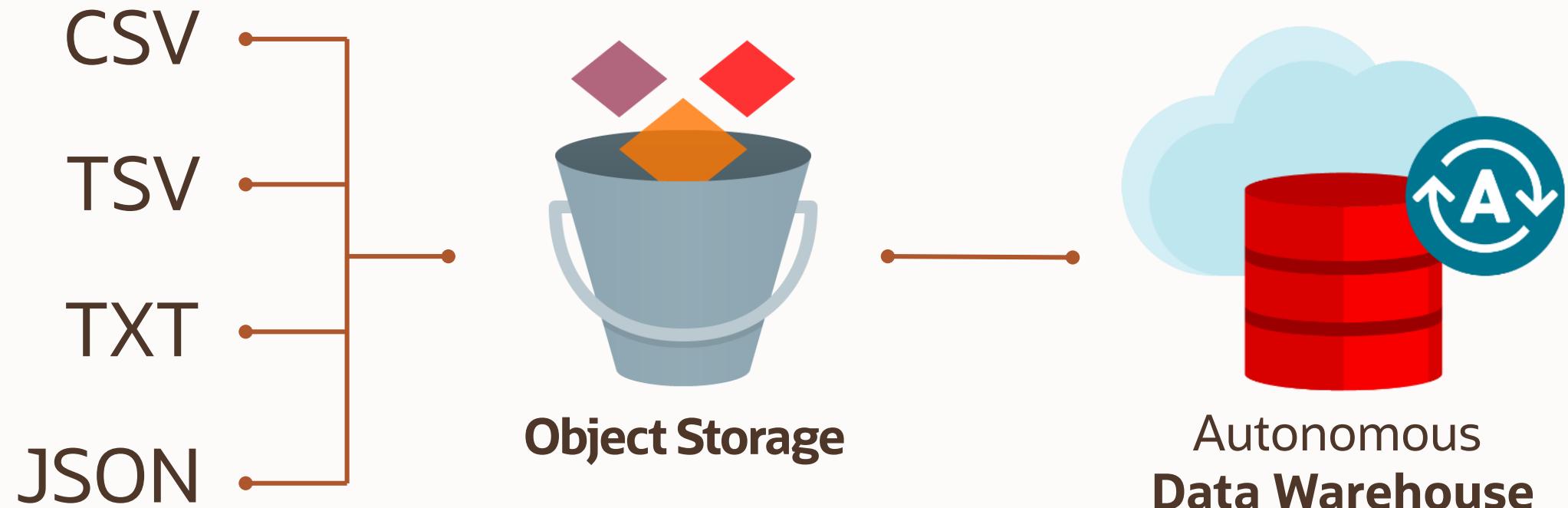
# Oracle Autonomous Database

Torna todo o stack de dados autônomo

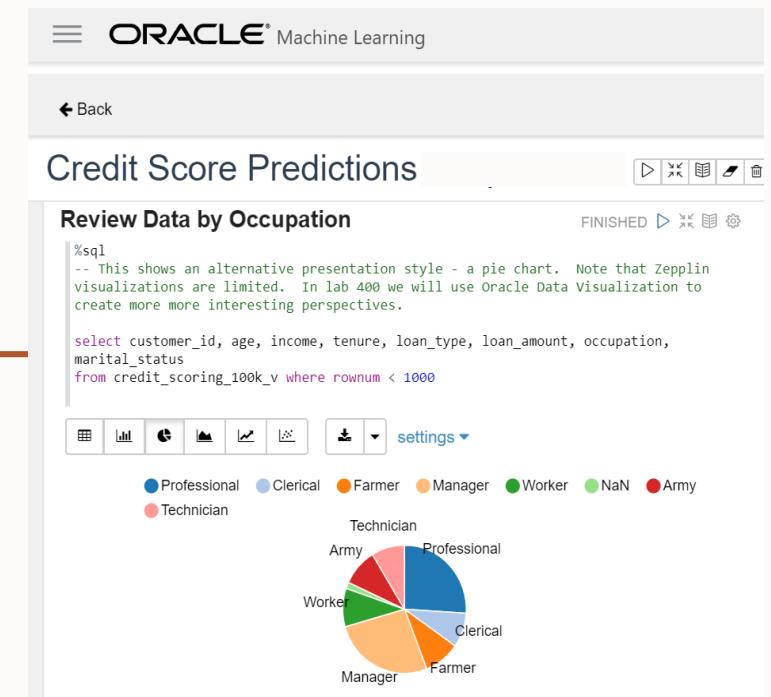
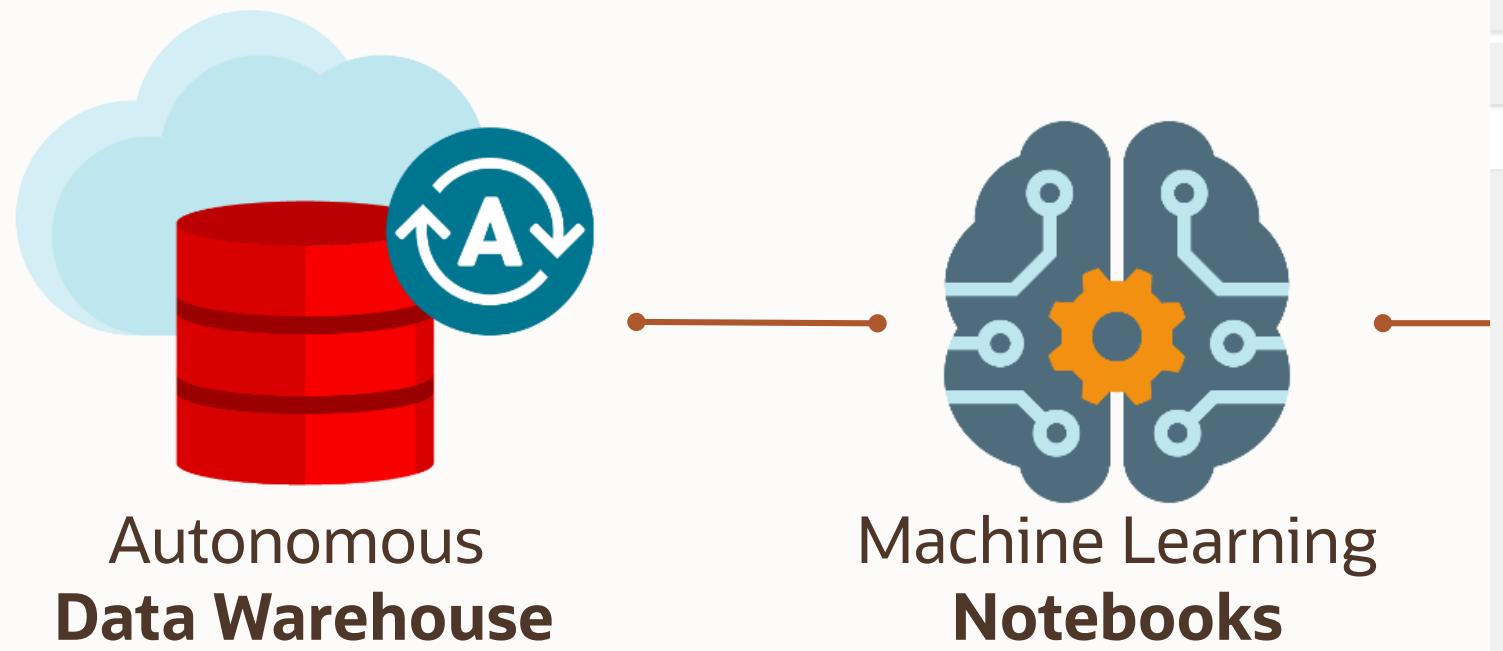


## Caso de uso: Data Lake

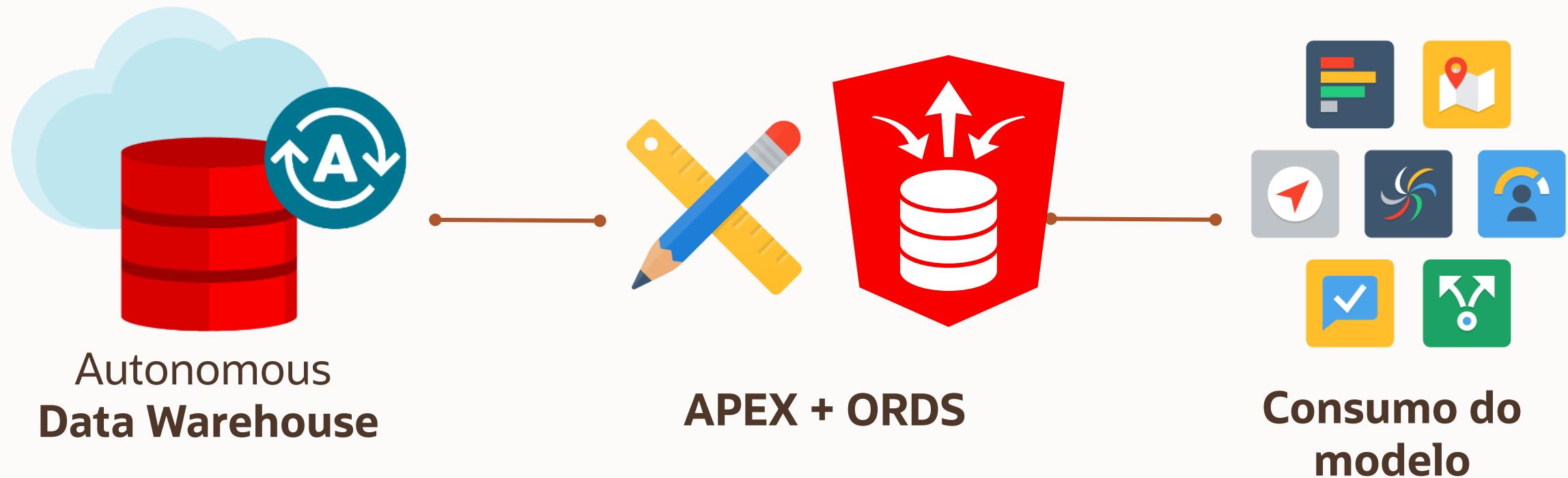
---



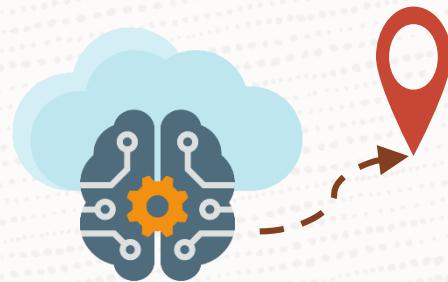
# Caso de uso: Laboratório de dados | mineração de dados



# Caso de Uso: Exposição do modelo de Machine Learning

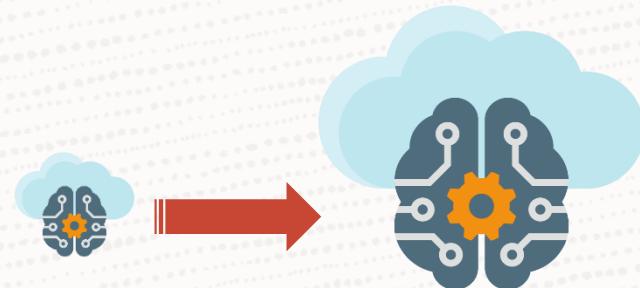


# Oracle Machine Learning | Atributos chave



## Automatizado

Pegar resultados mais rápidos com o minimo de esforço – até para usuários não experts



## Escalável

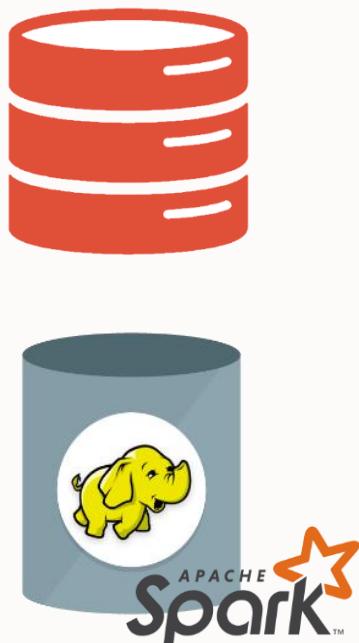
Tratando volumes de dados grandes (big data) usando paralelismo, algoritmos distribuidos – sem movimentação de dados



## Pronto para produção

Faça implementação e atualização de um ambiente de ciência de dados rapidamente com uma plataforma integrada de ML

**Aumente produtividade | Atinja objetivos | Inove mais**



# Oracle Machine Learning

## OML4SQL

SQL API

## OML4R

R API

## OML4Py

Python API

## OML Notebooks

com Apache Zeppelin no  
Autonomous Database

## Oracle Data Miner

Oracle SQL Developer extension

## OML4Spark

R API on Big Data

## OML AutoML UI\*

Code-free AutoML interface on Autonomous Database

## OML Services\*

Model Deployment and Management,  
Cognitive Text

\* Em Breve



# Onde gastamos mais tempo?

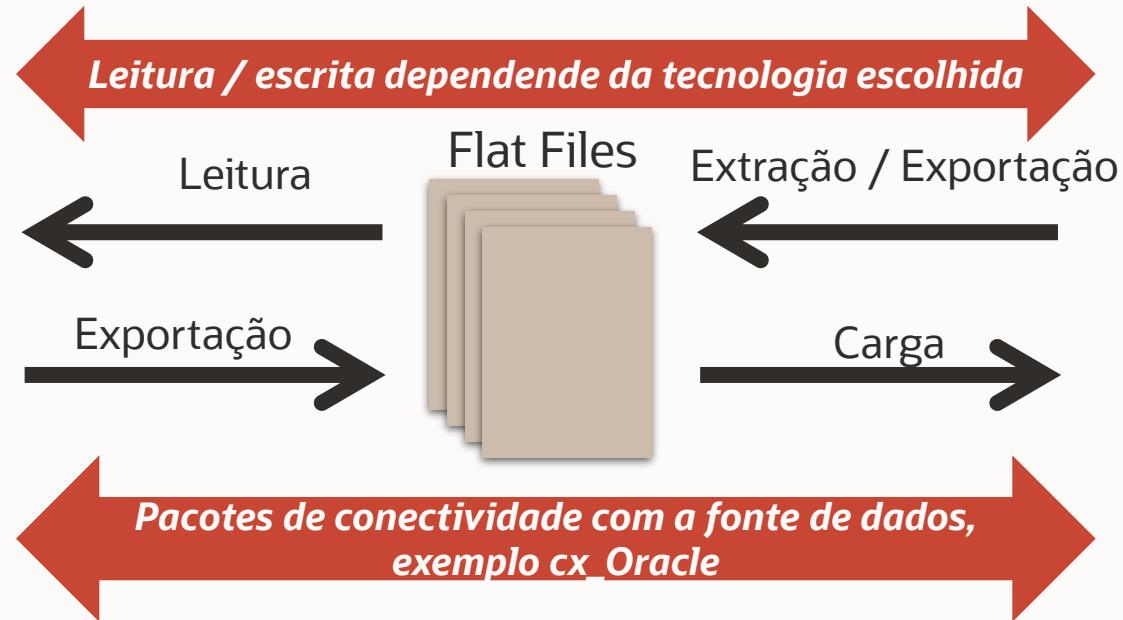
---



# Projeto de ciência de dados



# Ambiente Analítico tradicional x fontes de dados



Implementação  
Sob demanda  
cron job

Latência no acesso  
Paradigma: R/Python → *Data Access Language* → R/Python  
Limitação em memória – Tamanho dos dados, processamento in-memory  
Single threaded  
Problema em backup, recovery, security  
Implementação em produção sob demanda

# Oracle Machine Learning para Oracle Database

## Ferramentas



Apache Zeppelin



Python client,  
Jupyter Notebooks



SQL Developer  
SQL\*Plus



R client,  
RStudio



SQL Developer

## Componente - Oracle Machine Learning

### OML Notebooks

OML4SQL  
OML4Py\*  
OML4R\*

OML4Py\*

OML4SQL

OML4R

### Oracle Data Miner

## Plataforma de gerenciamento de dados



Autonomous Database



Oracle Database



Database  
Cloud Service

\*Em breve

# Oracle Machine Learning

Machine learning dentro do Autonomous Database com suporte a SQL / Python\*

## Automação

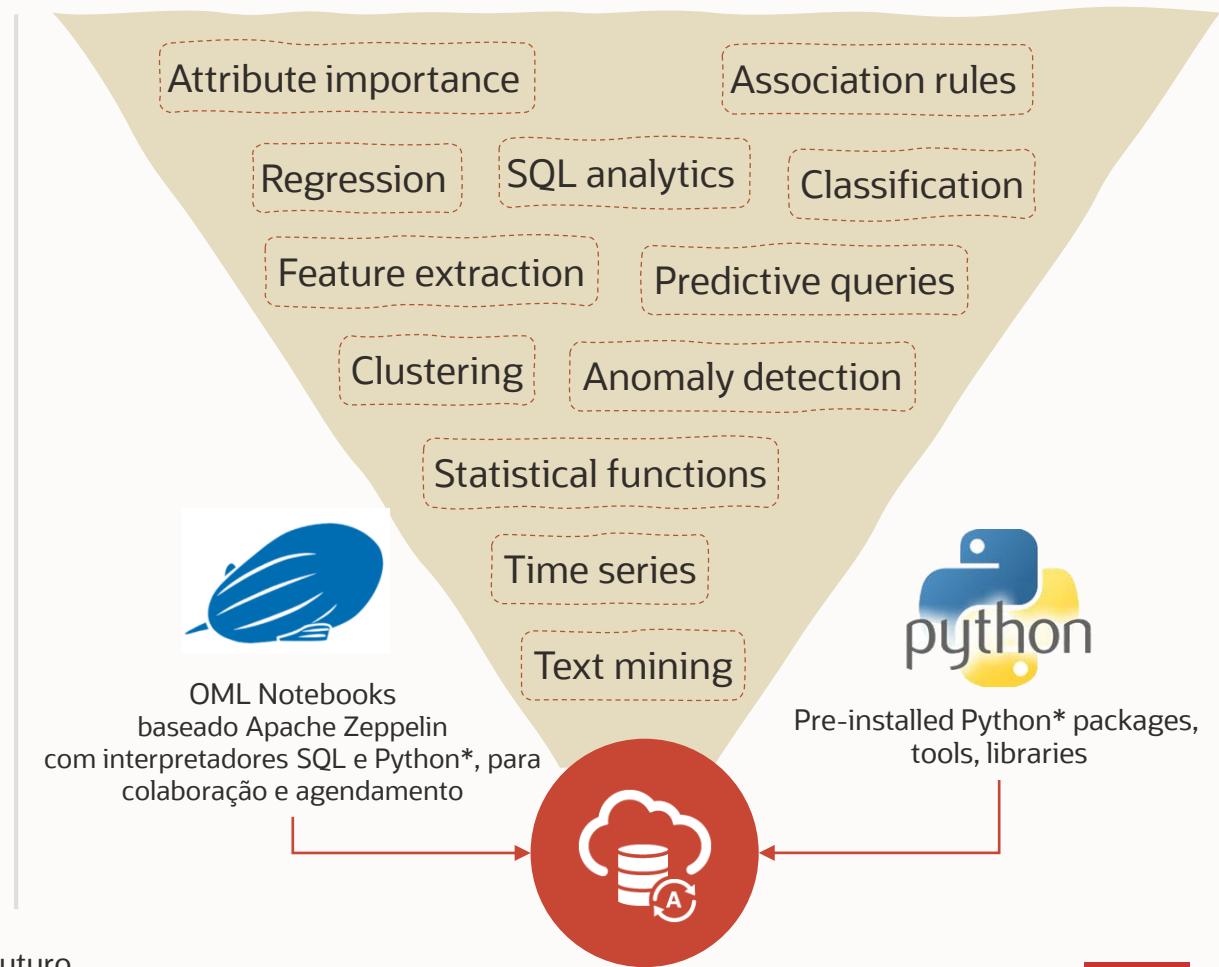
Tenha resultados rápidos com um fluxo incluindo autoML, preparação automatizada de dados, mineração, modelos participados e predição

## Escalabilidade

30+ algoritmos de alta performance, paralelizados no banco de dados que não requerem movimentação de dados para construção de modelos

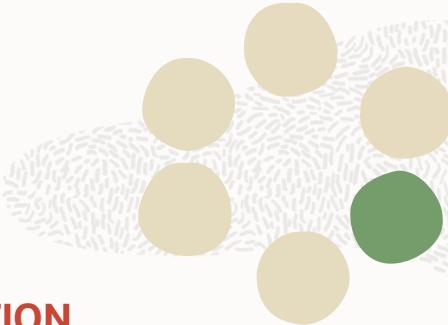
## Pronto para produção

Faça a implementação e atualize modelos de ML no banco de dados ou utilize REST APIs para realizar a implementação.



# Oracle Machine Learning algoritmos in-database e Analytics

No Autonomous Database acessivel via SQL e Python\*\*



## CLASSIFICATION

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

## CLUSTERING

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

## ANOMALY DETECTION

One-Class SVM  
MSET-SPRT \*

## REGRESSION

Generalized Linear Model (GLM)  
Support Vector Machine (SVM)  
Neural Network  
XGBoost \*

## TIME SERIES \*

Forecasting - Exponential Smoothing  
Includes popular models  
e.g. Holt-Winters with trends,  
seasonality, irregularity, missing data

## ATTRIBUTE IMPORTANCE

Minimum Description Length  
Principal Component Analysis \*  
Unsupervised Pair-wise KL Div \*  
CUR Decomposition \*

## ASSOCIATION RULES

A priori/ market basket

## STATISTICAL FUNCTIONS

min, max, median, stdev, Pearson/  
Kendall/Spearman correlation  
Others: t-test, F-test,, Chi-Sq,  
ANOVA, etc. \*

## FEATURE EXTRACTION

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

## ROW IMPORTANCE

CUR Decomposition \*

## RANKING

XGBoost \*

## TEXT MINING SUPPORT

Algorithms support text columns  
Tokenization and theme extraction  
Explicit Semantic Analysis (ESA) \*

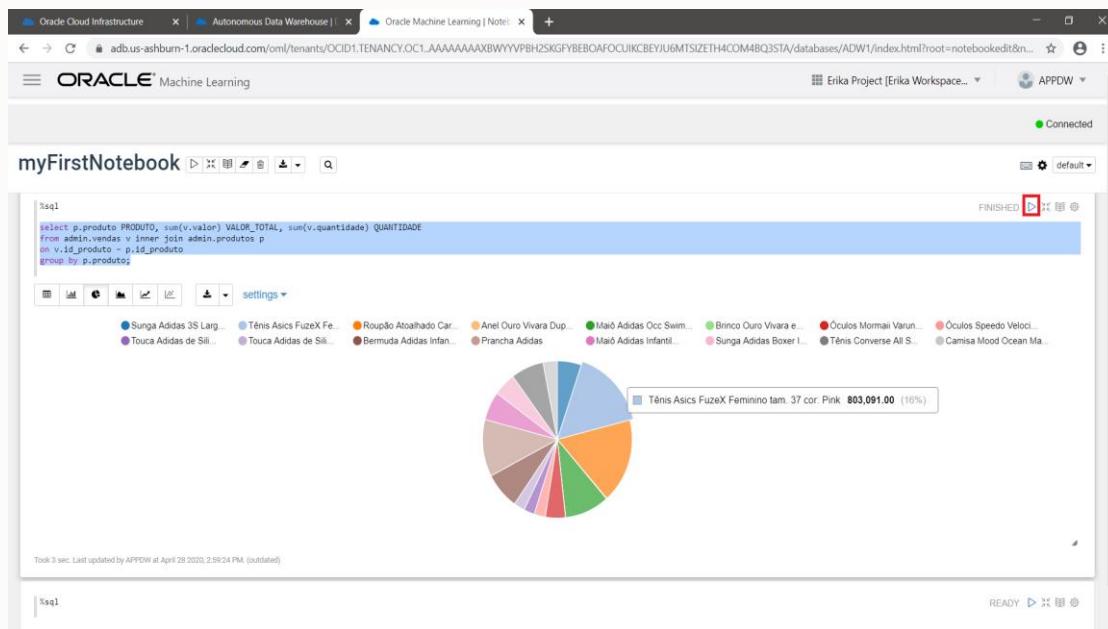
## SQL ANALYTICS \*

SQL Windows  
SQL Patterns  
SQL Aggregates

\* Disponível somente via  
SQL API

\*\*futuro

# Oracle Machine Learning Notebooks x Oracle Machine Learning



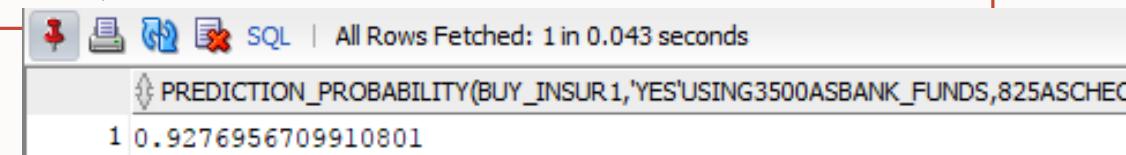
## Model build (PL/SQL)

```
DECLARE
    v_setlst DBMS_DATA_MINING.SETTING_LIST;
BEGIN
    v_setlst('ALGO_NAME') := 'ALGO_SUPPORT_VECTOR_MACHINES';
    V_setlst('PREP_AUTO') := 'ON';

    DBMS_DATA_MINING.CREATE_MODEL2 (
        MODEL_NAME          => 'BUY_INSUR1',
        MINING_FUNCTION     => 'CLASSIFICATION',
        DATA_QUERY          => 'select * from CUST_INSUR_LTV',
        SET_LIST             => v_setlst,
        CASE_ID_COLUMN_NAME => 'CUST_ID',
        TARGET_COLUMN_NAME   => 'CUST_INSUR_LTV');
END;
```

## Real-time scoring (SQL query)

```
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;
```



# O que é Oracle APEX?

---



Plataforma de desenvolvimento de baixa codificação (low-code) que possibilita você a desenvolver aplicações escaláveis e seguras compatíveis com os padrões de mercado.



<https://apex.oracle.com>

# Implementação usando APEX

Wine Results

## Predictions

Variety contains 'pinot'

Probability*100	Price	Country	Designation	Province	Variety	Prediction
99.52	14	US			Pinot Blanc	LT_90_POINTS
99.52	14	US			Pinot Gris	LT_90_POINTS
99.52	14	US			Pinot Gris	LT_90_POINTS
99.52	14	US			Pinot Gris	LT_90_POINTS
99.52	14	US			Pinot Grigio	LT_90_POINTS
99.20	8	Argentina			Pinot Grigio	LT_90_POINTS
99.20	9	Argentina			Pinot Grigio	LT_90_POINTS
99.20	7	Argentina			Pinot Grigio	LT_90_POINTS
99.20	9	Argentina			Pinot Grigio	LT_90_POINTS
99.20	10	Argentina			Pinot Grigio	LT_90_POINTS
99.20	10	Argentina			Pinot Grigio	LT_90_POINTS
99.20	10	Argentina			Pinot Grigio	LT_90_POINTS
99.20	9	Argentina			Pinot Grigio	LT_90_POINTS
99.20	9	Argentina			Pinot Grigio	LT_90_POINTS
99.20	12	Argentina	Art of the Andes	Mendoza Province	Pinot Grigio	LT_90_POINTS
99.20	10	Argentina			Pinot Grigio	LT_90_POINTS
99.20	12	Argentina			Pinot Grigio	LT_90_POINTS

Wines by Country

Predictions

US - 42% (28013)

France - 15% (9877)

Spain - 6% (4174)

Italy - 13% (8509)

New Zealand - 1% (889)

Portugal - 5% (3208)

South Africa - 1% (830)

Argentina - 4% (2422)

Australia - 2% (1340)

Austria - 3% (1690)

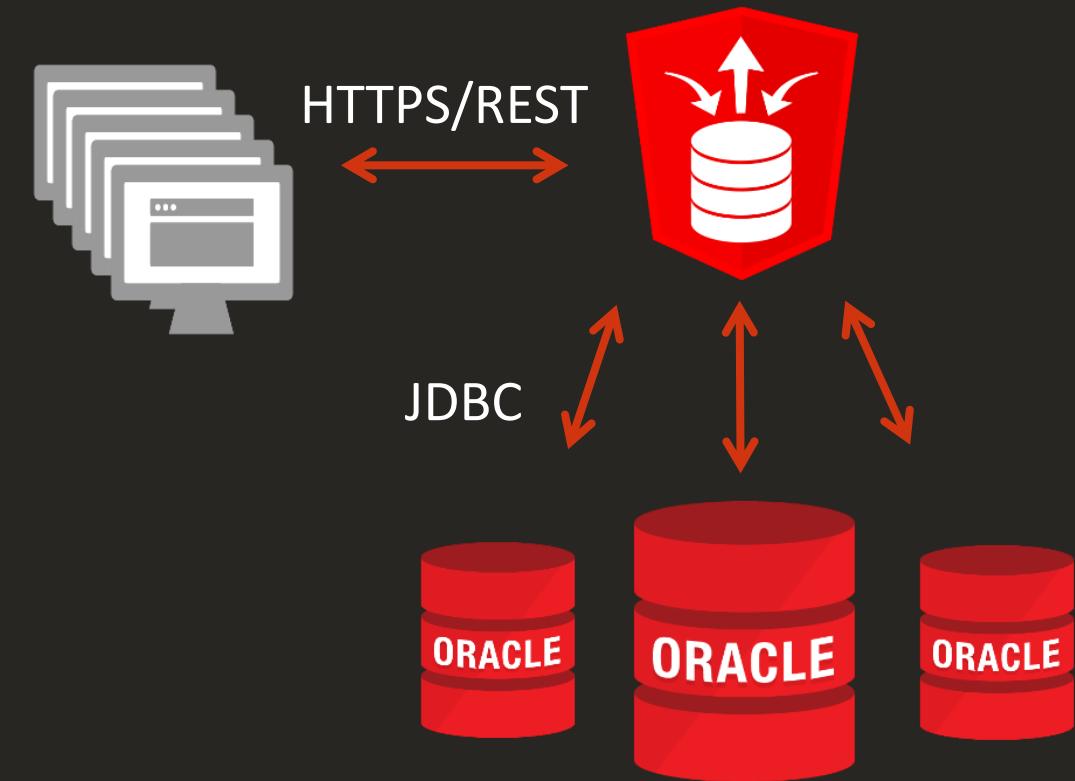
Chile - 5% (3008)

Germany - 2% (1162)

Other - 2% (1557)

# Usando acesso via REST API com BD

- Oracle REST Data Services (ORDS)
- Auto REST enable your data tables, views, SQL, PL/SQL
- Auto paged results
- Auto JSON responses
- Complete read-write functionality
- Full dev support (GUI, CLI, API)
- Oracle Cloud and On-Premises



<https://www.oracle.com/REST>

# Implementação do Modelo via ORDS

The screenshot shows a Jupyter Notebook interface with the URL `localhost:8888/notebooks/OOW19_AdW_Demo.ipynb#`. The notebook title is `jupyter OOW19_AdW_Demo`. The last checkpoint was 41 minutes ago (unsaved changes). The kernel is set to Python 3. The code in cell [20] demonstrates how to use the `requests` library to call a RESTful API. The code defines variables for country ('Portugal'), province ('Douro'), variety ('Portuguese Red'), and price ('30'). It then uses the `requests.get` method to send a GET request to a specific URL and prints the JSON response using `json.dumps`.

```
In [20]: import json
import requests

country = 'Portugal'
province = 'Douro'
variety = 'Portuguese Red'
price = '30'

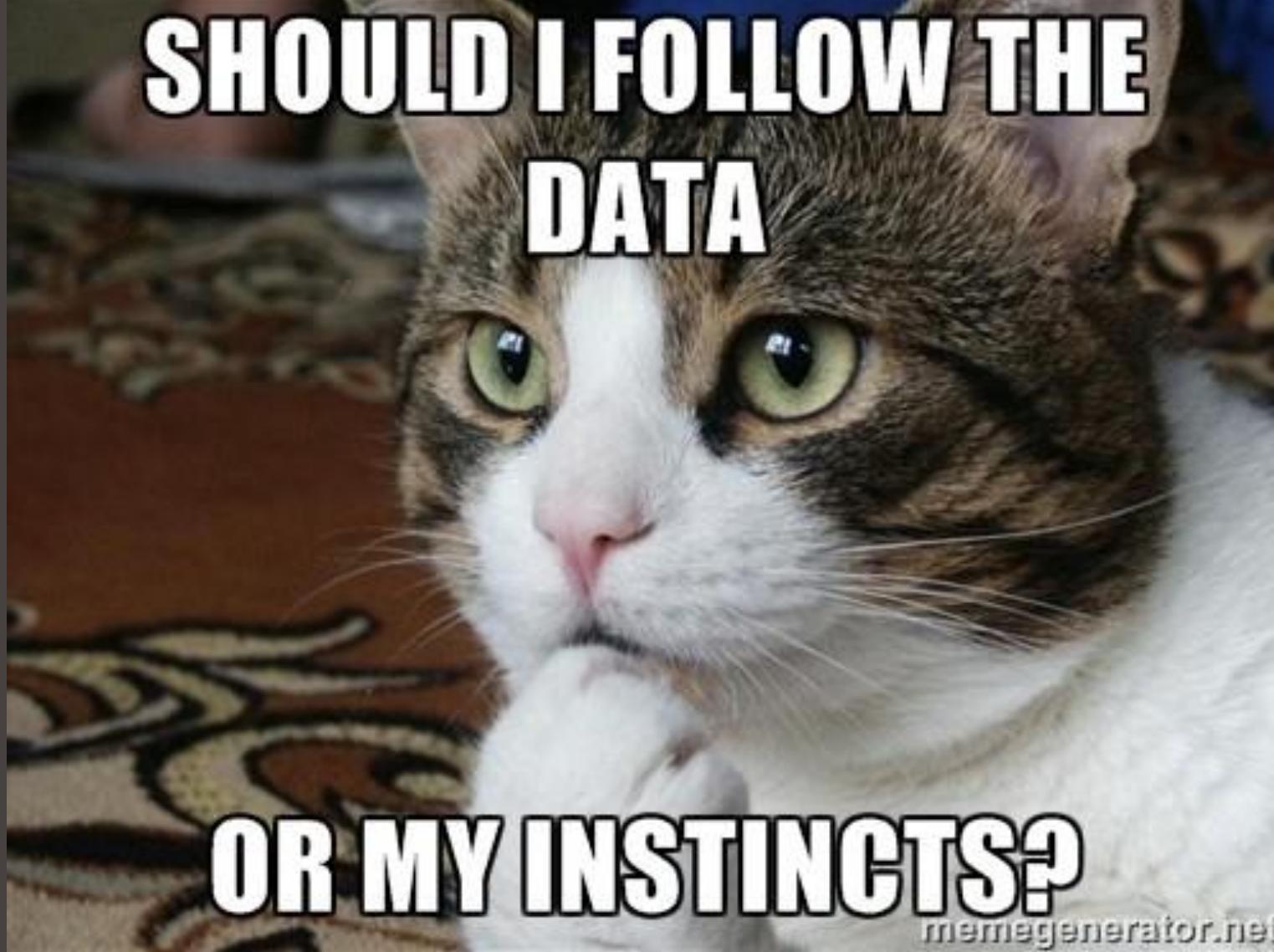
resp = requests.get('https://jggnlb6iptk8gum-adw2.adb.us-ashburn-1.oraclecloudapps.com/ords/oml_user/wine/wine_pred/?country=' + country + '&province=' + province + '&variety=' + variety + '&price=' + price)
json_data = resp.json()
print(json.dumps(json_data, indent=2))

country = 'Portugal'
province = 'Douro'
variety = 'Portuguese Red'
price = '31'

resp = requests.get('https://jggnlb6iptk8gum-adw2.adb.us-ashburn-1.oraclecloudapps.com/ords/oml_user/wine/wine_pred/?country=' + country + '&province=' + province + '&variety=' + variety + '&price=' + price)
json_data = resp.json()
print(json.dumps(json_data, indent=2))
```

{  
  "pred\_wine": "LT\_90\_POINTS",  
  "prob\_wine": 0.6844716987704507  
}  
{  
  "pred\_wine": "GT\_90\_POINTS",  
  "prob\_wine": 0.5137163891143298

RESTful API for calling  
OML model to make predictions



**SHOULD I FOLLOW THE  
DATA**

**OR MY INSTINCTS?**

memegenerator.net

# Alguns recadinhos ☺

---





# Inovação com dados em nuvem

MACHINE LEARNING COM SQL  
E ORACLE AUTONOMOUS  
DATABASE



\$> echo "obrigada"

# Erika Nagamine

Principal Cloud Solutions Engineer

Data Management | Data Engineer | Data Scientist | Analytics

Tech Cloud Solutions Engineer – Oracle LAD - Brazil



<https://linkedin.com/in/erikanagamine>



[erika.nagamine@oracle.com](mailto:erika.nagamine@oracle.com)



<https://github.com/erikanagamine>



@erikanagamine



@erikanagamine

A person is seen from the side, wearing a hooded garment with prominent black and white zebra stripes. The hood covers their head, and the pattern continues down the front of the garment.

**ORACLE**