Machine Learning com SQL e Autonomous Database

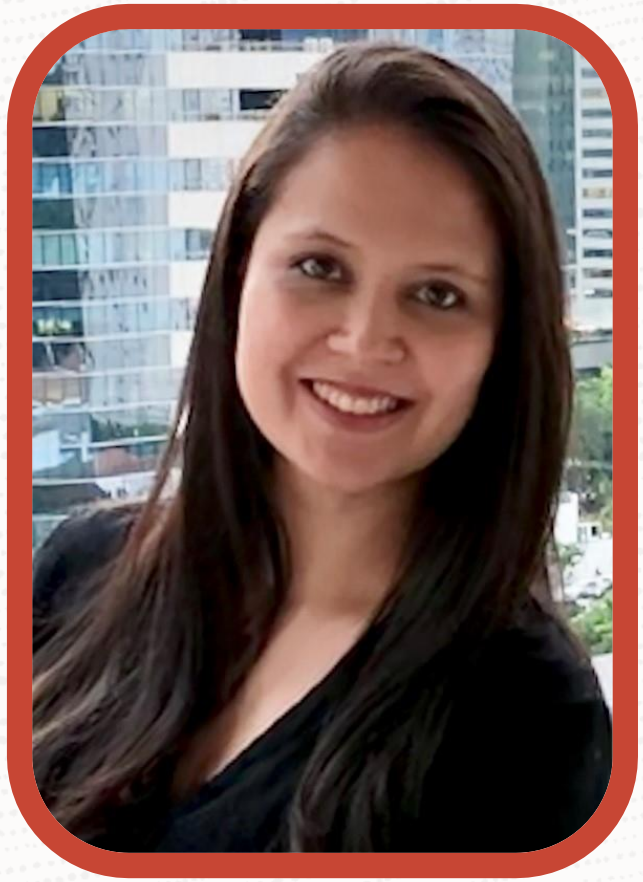
Erika Nagamine

Inovação com dados em nuvem
29.10.2020

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.

$> whoami



Erika Nagamine

Cloud Solutions Engineer
Data Management | Data Engineer | Data Scientist | Analytics
Tech Brazil Cloud Solutions Engineer
Oracle

@erikanagamine

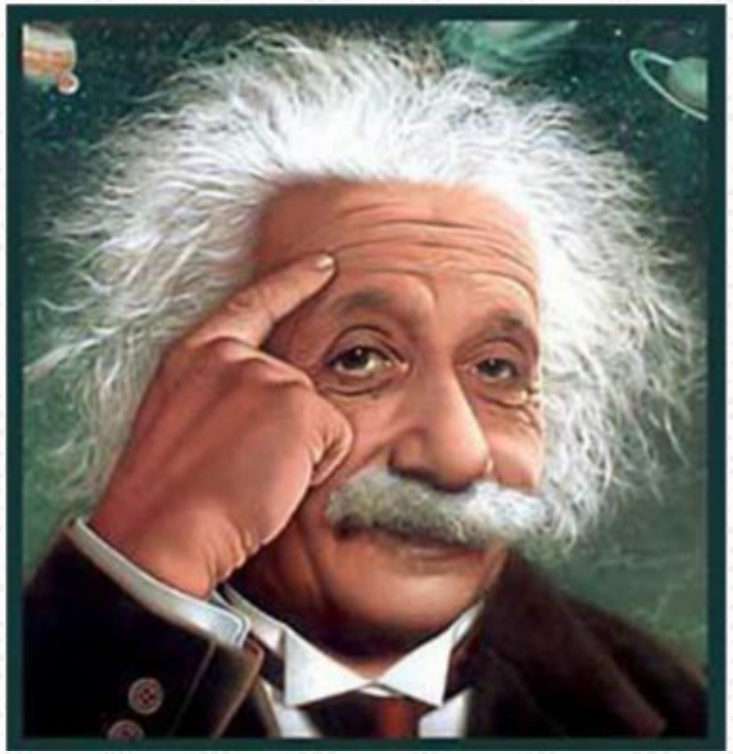
@erikanagamine

erika.nagamine@oracle.com

https://github.com/erikanagamine

https://www.linkedin.com/in/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

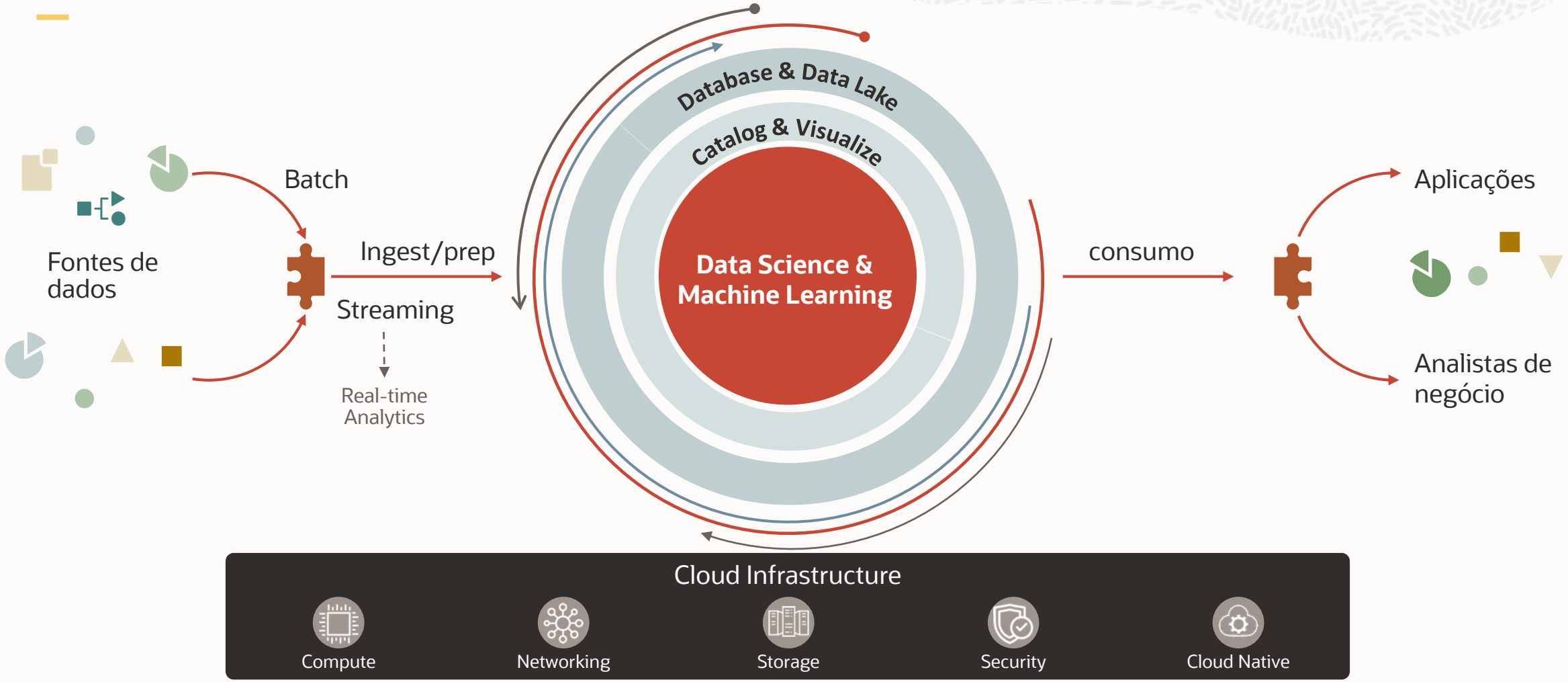
Do-it-yourself cloud

Construa o que você quer,

mas faça você mesmo

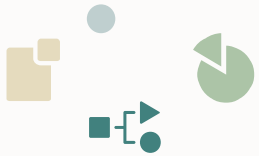
Oracle Data Platform

Machine learning suportado por dados, integração, gerenciamento e analytics



Oracle Data Platform

Machine learning suportado por dados, ingestão, gerenciamento e analytics



Fonte de dados

Analytics

- Oracle Analytics Cloud
- Graph Analytics

Data Science & Machine Learning

- Oracle Cloud Infrastructure Data Science
- Oracle Machine Learning
- Data Science Virtual Machines

Integration

- Oracle Cloud Infrastructure Streaming
- GoldenGate & Oracle Data Integrator
- Oracle Cloud Infrastructure Data Integration
- Oracle Cloud Infrastructure Data Catalog

Data Management

- Big Data Service
- Oracle Cloud Infrastructure Data Flow
- Object Storage
- Autonomous Database

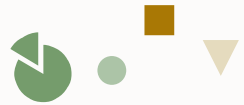
Governance

- Oracle Cloud Infrastructure Data Catalog
- Oracle Cloud Infrastructure Data Safe

Cloud Infrastructure

- Compute
- Networking
- Storage
- Security
- Cloud Native

Aplicações

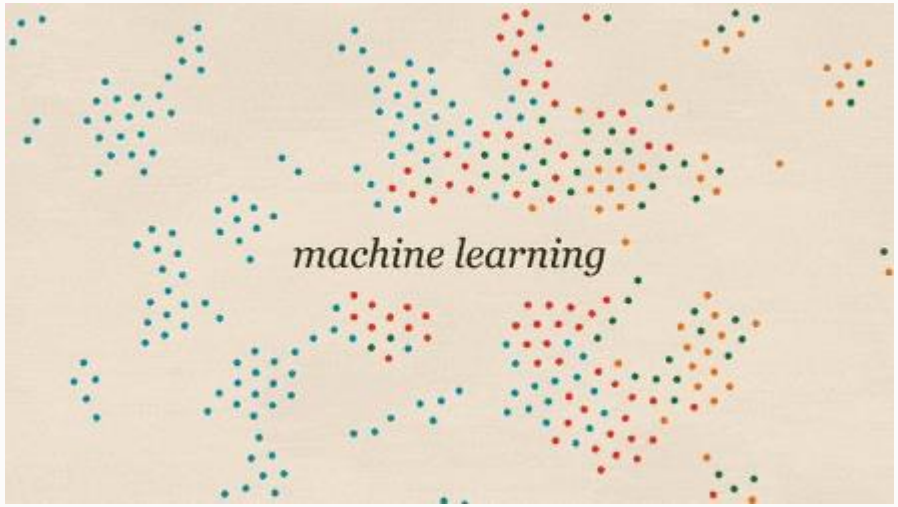


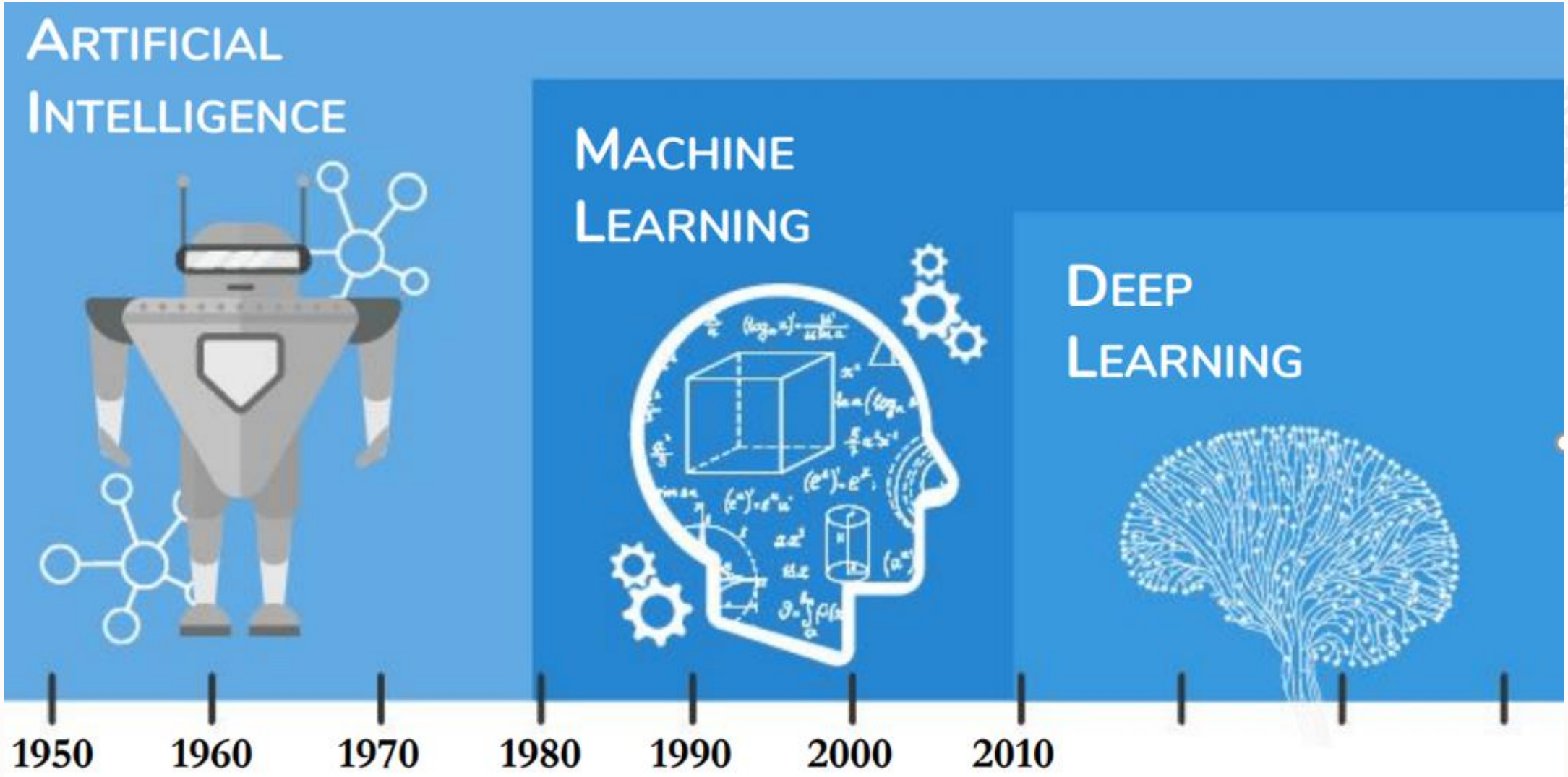
Analistas de negócio

Machine Learning

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]





https://www.linkedin.com/pulse/ai-machine-learning-evolution-differences-connections-kapil-tandon

0101010101
0101010101
0100101010

11100101010
1001010010
1001010100

0010101010
0101010100
010100000

File Edit Selection View Go Debug Terminal Help plot.py - categorical - Visual Studio Code

plot.py

```
Run Cell | Run Below
#%% [markdown]
# ## Plots
# #### Examples from: https://matplotlib.org/3.1.0/gal
Run Cell | Run Above | Run Below
#%%
import matplotlib.pyplot as plt

Run Cell | Run Above | Run Below
#%%
data = {'apples': 10, 'oranges': 15, 'lemons': 5, 'lim
names = list(data.keys())
values = list(data.values())

fig, axs = plt.subplots(1, 3, figsize=(9, 3), sharey=T
axs[0].bar(names, values)
axs[1].scatter(names, values)
axs[2].plot(names, values)
fig.suptitle('Categorical Plotting')

Run Cell | Run Above | Run Below
#%%
cat = ["bored", "happy", "bored", "bored", "happy", "b
dog = ["happy", "happy", "happy", "happy", "bored", "b
activity = ["combing", "drinking", "feeding", "napping
```

Python Interactive

Variables

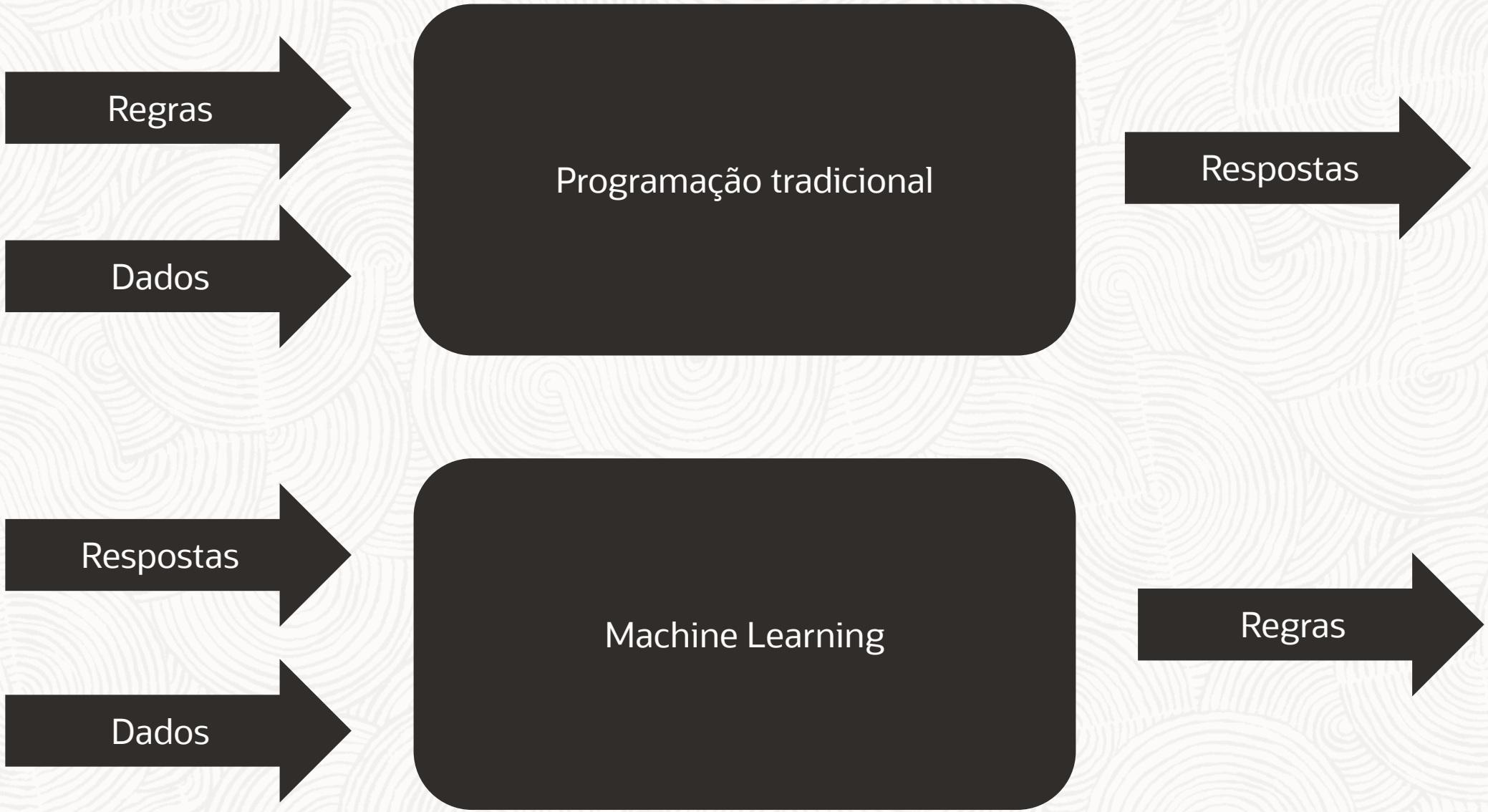
```
Jupyter Server URI: http://localhost:8889/?
token=476e9fc714196f2d625713b5706518052de368a0b85aeb63
Python version:
3.7.3 (v3.7.3:ef4ec6ed12, Mar 25 2019, 22:22:05) [MSC v.1916 64 bit
(AMD64)]
(5, 7, 2)
C:\\Users\\luabud\\AppData\\Local\\Programs\\Python\\Python37\\pyth
```

Plots

Examples from:
https://matplotlib.org/3.1.0/gallery/lines_bars_and_markers/categorical_variables.htm

[1] *Shift-enter to run*

Python 3.7.3 64-bit 0 0



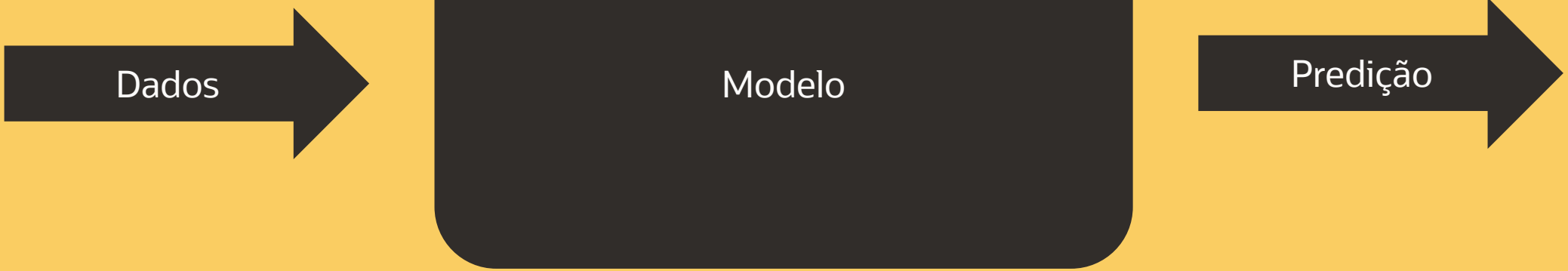
0101010101
0101010101
0100101010

11100101010
1001010010
1001010100

0010101010
0101010100
010100000

Caracteristicas – Features:
Posição
Tom de pele
Qualidade da fotografia

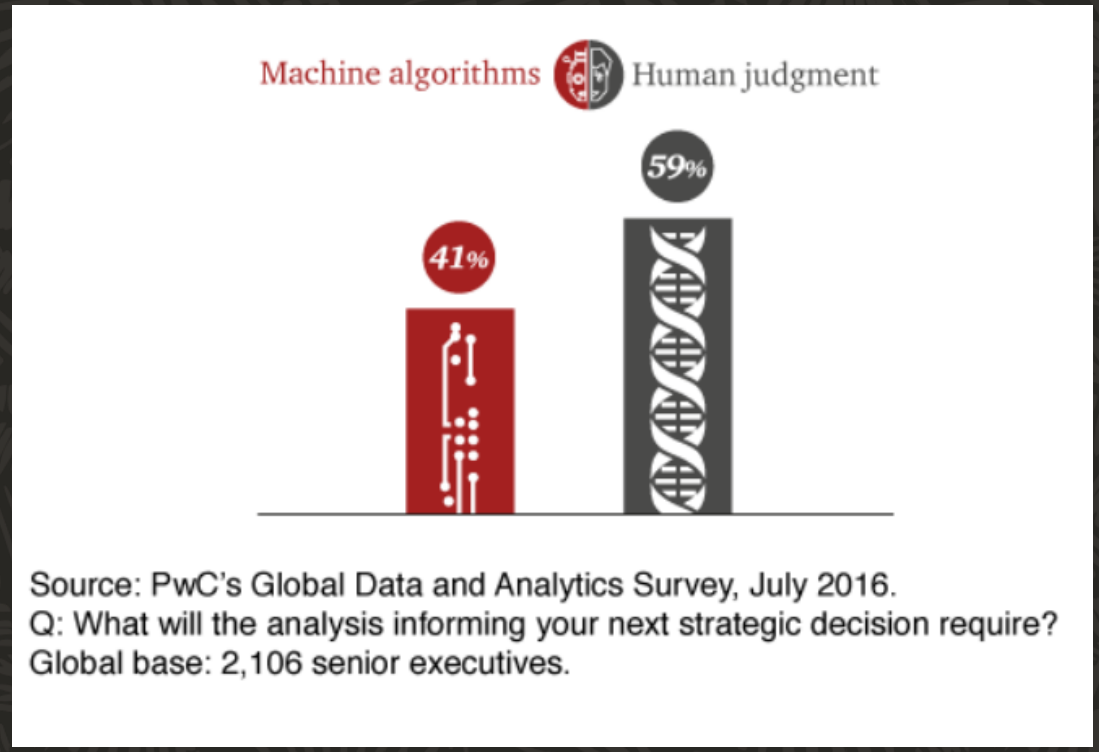
treino



Inferência

Algoritmos são melhores quando usam dados

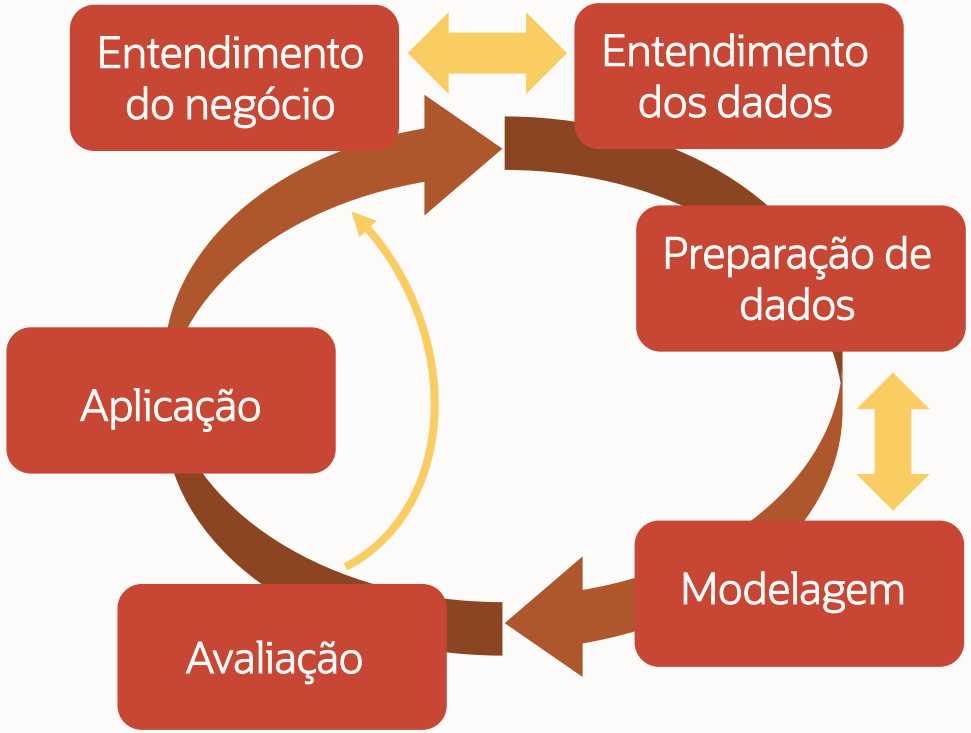
O que não elimina o julgamento humano



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

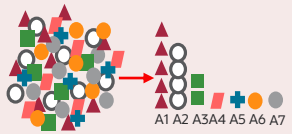
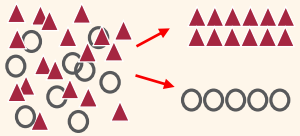
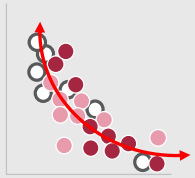
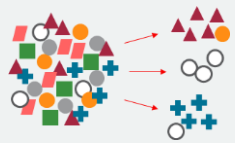
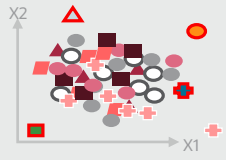
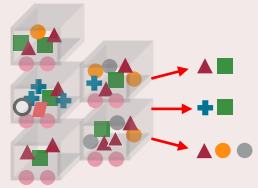
Desafios de automação

Preparação de dados*
Entendimento do negócio
Entendimento dos dados

*Depende do tipo de dados. Tendências baseadas no tempo são um desafio, a criação de caracteristica 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 predições

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 possiveis 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 consume	Prediz ou estima um valor	Segmenta uma população em silos	Encontra outlier or “eventos raros”	Determina a ocorrencia similar de consumo

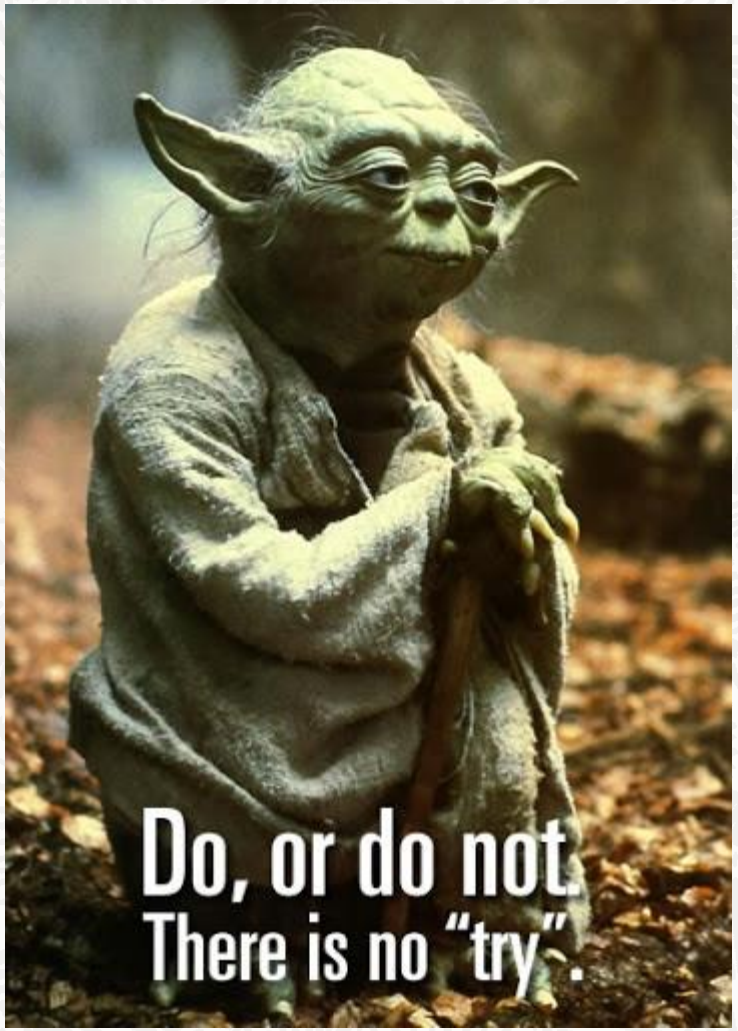
Outro exemplo pratico de como algoritmos podem ajudar no dia a dia

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

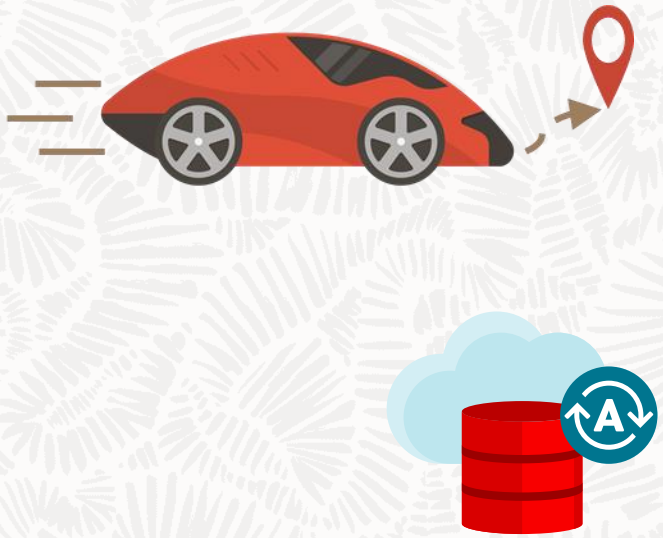
Fonte: https://g1.globo.com/sp/campinas-regiao/noticia/2020/01/27/unicamp-desenvolve-software-que-permite-86percent-de-precisao-no-diagnostico-do-cancer-de-pele.ghtml

Workshop

O que utilizaremos?



Autonomous



Oracle Autonomous Database

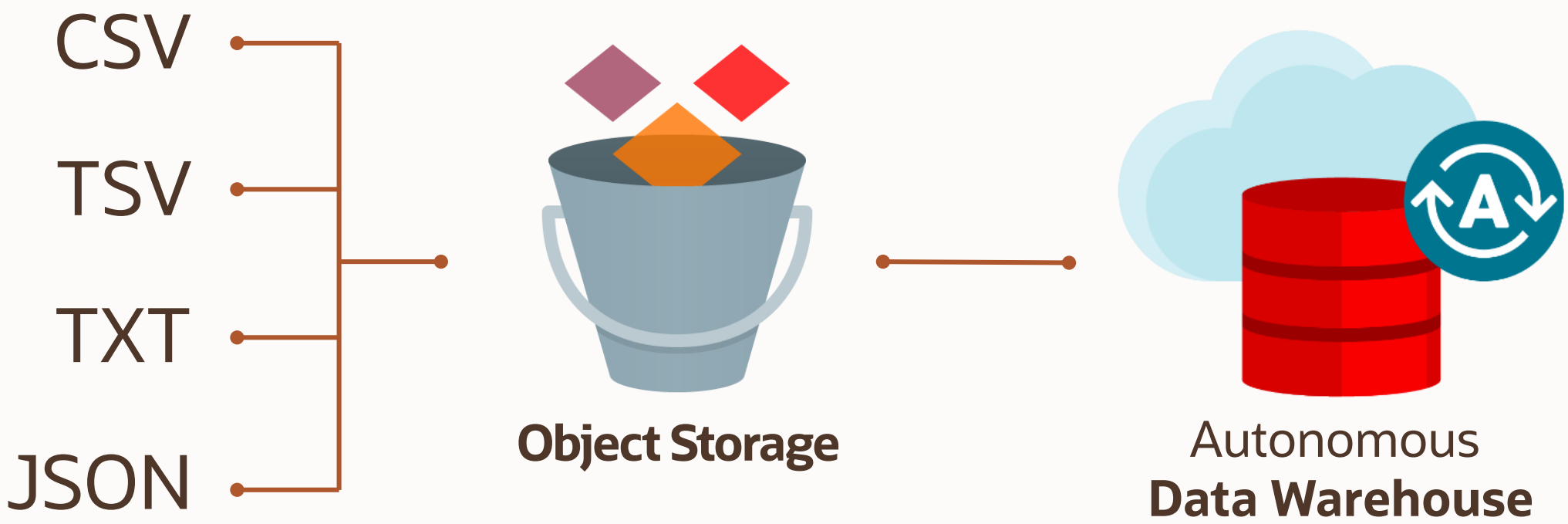
Torna todo o stack de dados autônomo

Automatização completa da infraestrura

Automatização completa do banco de dados

Automatização completa das Operações de Data Center e Machine Learning

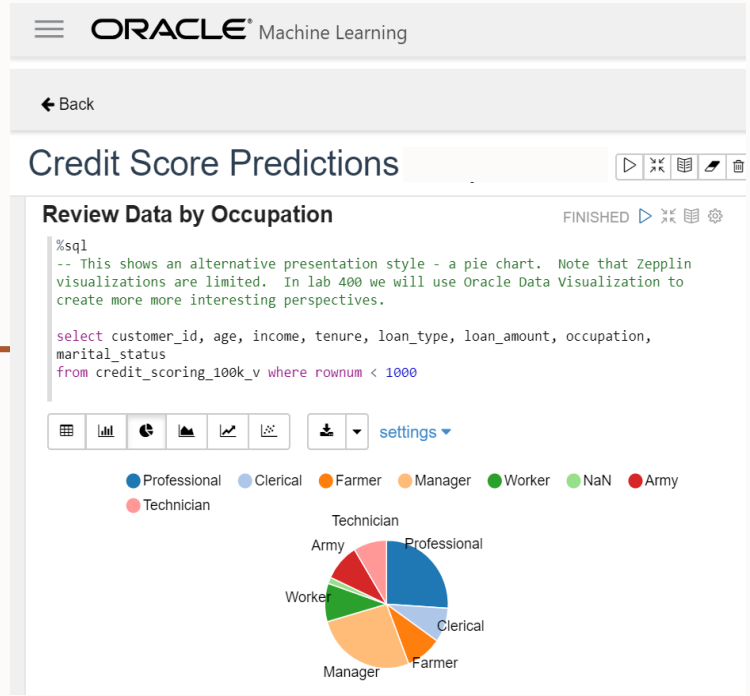
Caso de uso: Data Lake



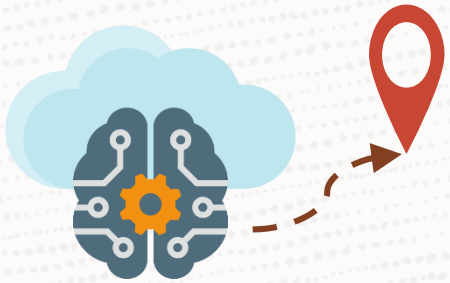
Caso de uso: Laboratório de dados | mineração de dados

Autonomous
Data Warehouse

Machine Learning
Notebooks

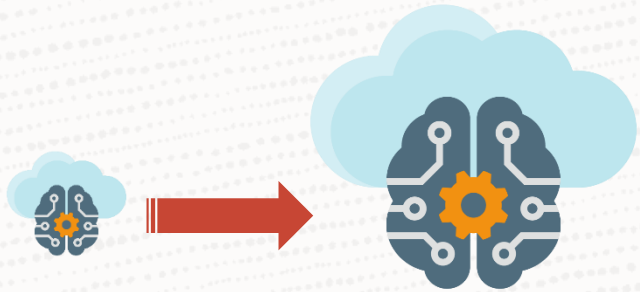


Oracle Machine Learning | Atributos chave



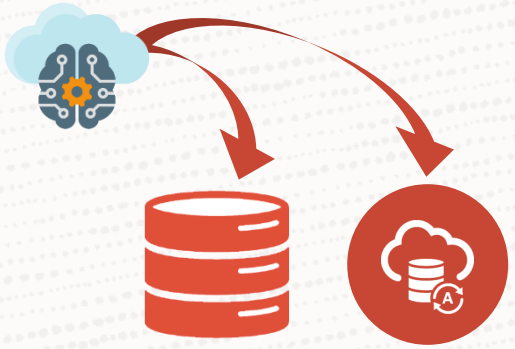
Automatizado

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

OML Notebooks
com Apache Zeppelin no Autonomous Database

OML4R
R API

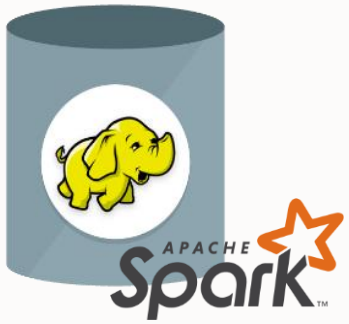
Oracle Data Miner
Oracle SQL Developer extension

OML4Py
Python API

OML4Spark
R API on Big Data

OML AutoML UI*
Code-free AutoML interface on Autonomous Database

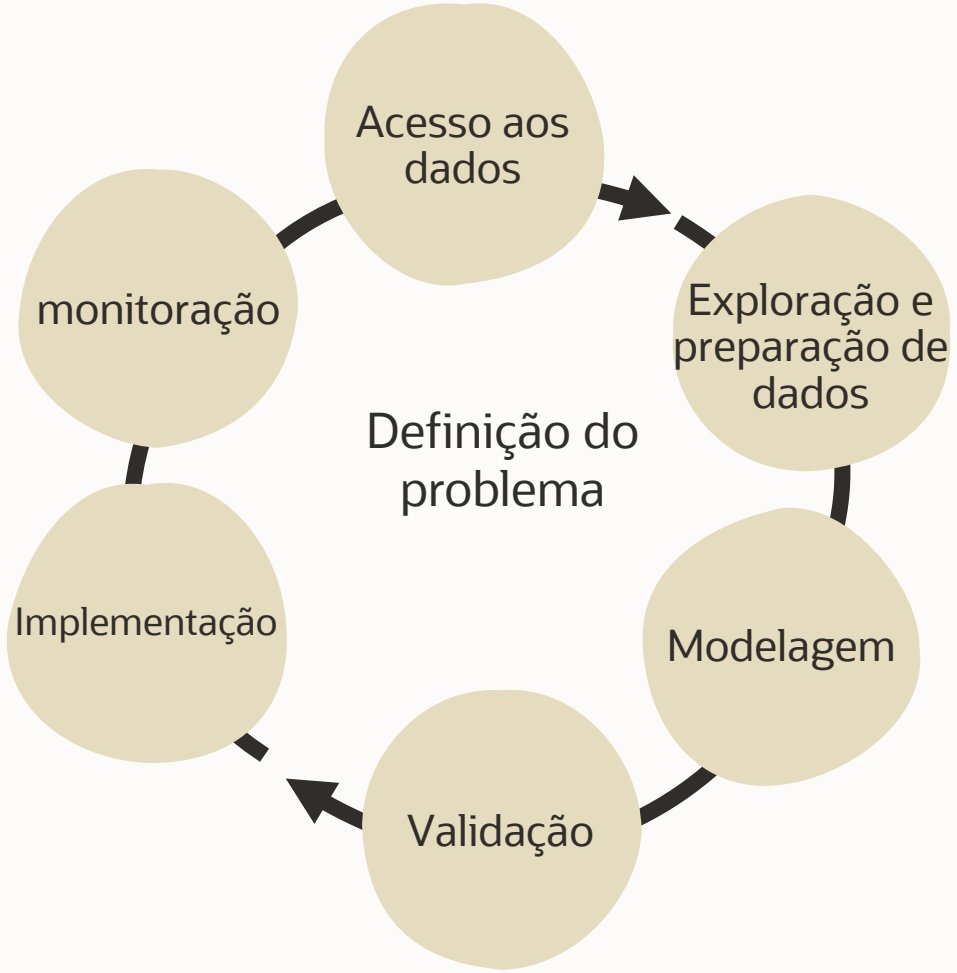
OML Services*
Model Deployment and Management, Cognitive Text



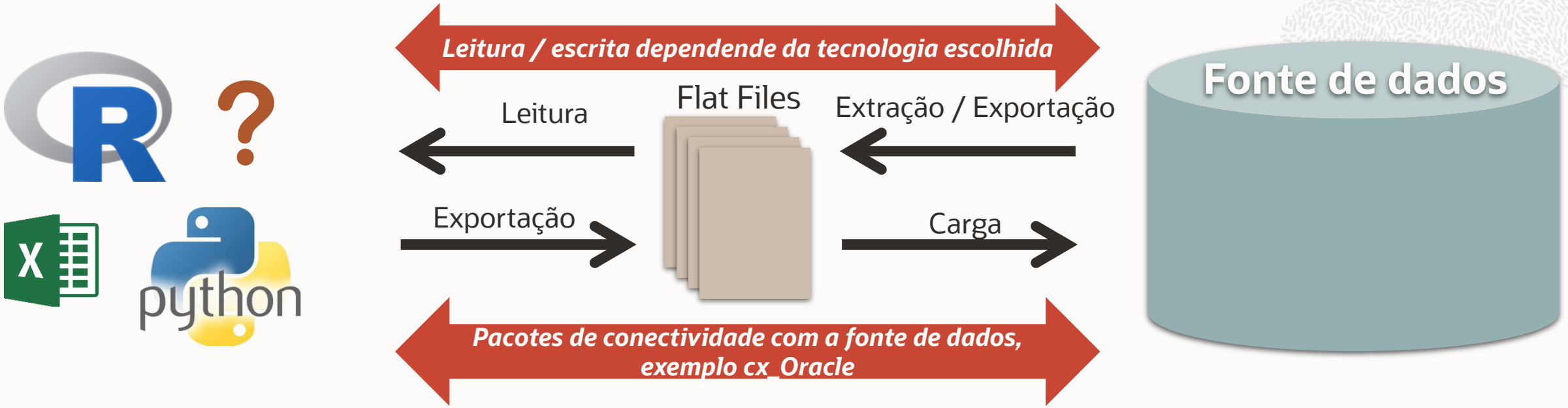
* Em Breve

Onde gastamos mais tempo?

Projeto de ciência de dados



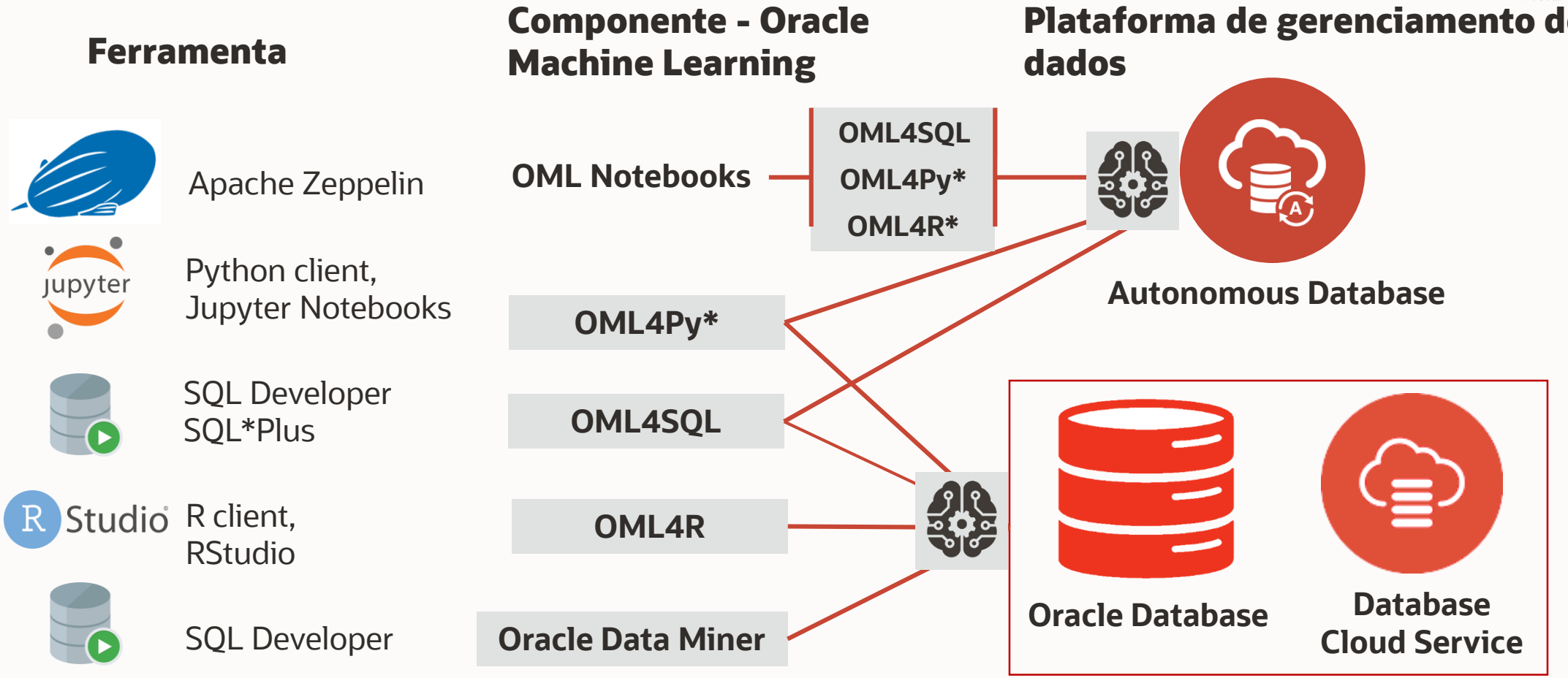
Ambiente Analitico tradicional x fontes de dados



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



*Em breve

Oracle Machine Learning

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

Automação

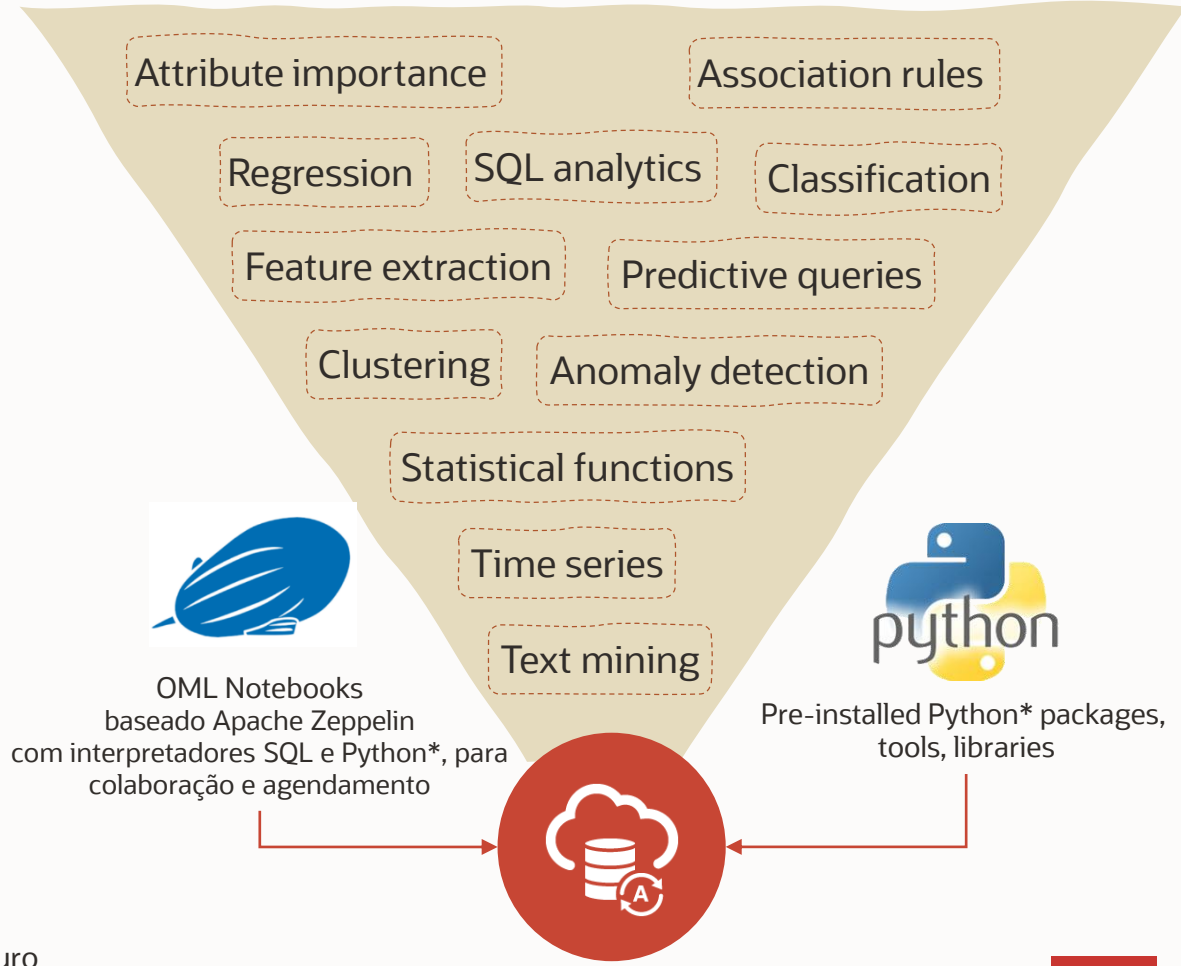
Tenha resultados rápidos com um fluxo incluindo autoML, preparação automatizada de dados, mineração, modelos particinados e predição

Escalabilidade

30+ algoritimos 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.



* Futuro

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 *

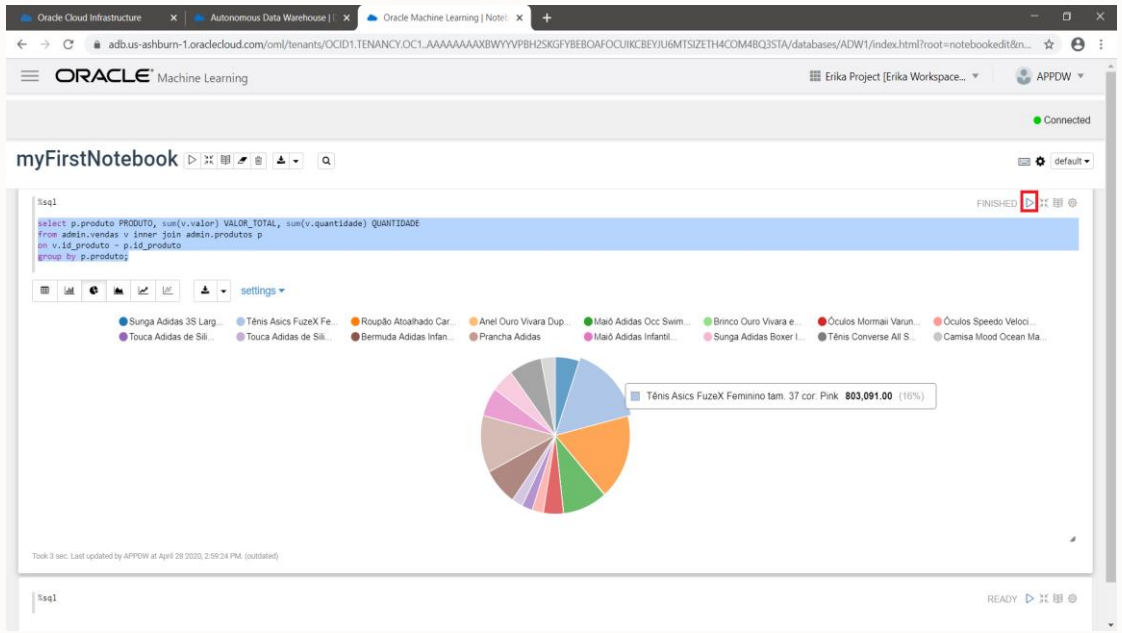
Text Mining Support
- Algorithms support text columns
- Tokenization and theme extraction
- Explicit Semantic Analysis (ESA) *

SQL Analytics *
- SQL Windows
- SQL Patterns
- SQL Aggregates

* Disponivel 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;
```

SQL | All Rows Fetched: 1 in 0.043 seconds

	PREDICTION_PROBABILITY(BUY_INSUR1,'YES'USING3500ASBANK_FUNDS,825ASCHEC
1	0.9276956709910801

Alguns recadinhos ☺

ORACLE

Autonomous Fast Track

Workshop

Conheça na prática o banco de dados mais moderno do mercado!

04/11/2020 - 9h às 18h

AGENDA

- Abertura
- Oracle Autonomous Database - first steps
- Programação Low Code com Oracle APEX
- Integrando Autonomous com Oracle Analytics
- SQL Developer Web
- Explorando dados com Oracle Machine Learning

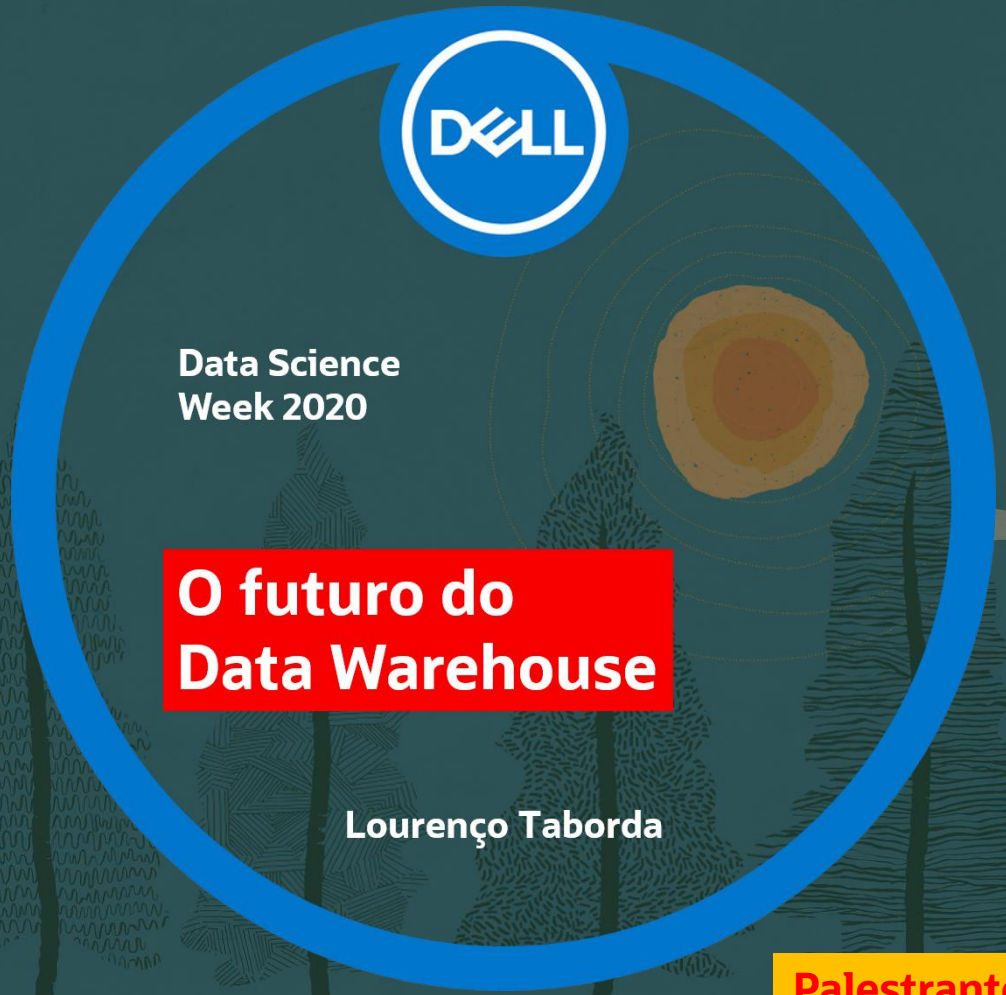
Inscreva-se já:

http://bit.ly/autonomousft-novembro

Gostei! Quero medalha e mat

Remover esta etiqueta antes da apresentação :D

Queremos dar uma medalha e compartilhar conhecimento com todos!

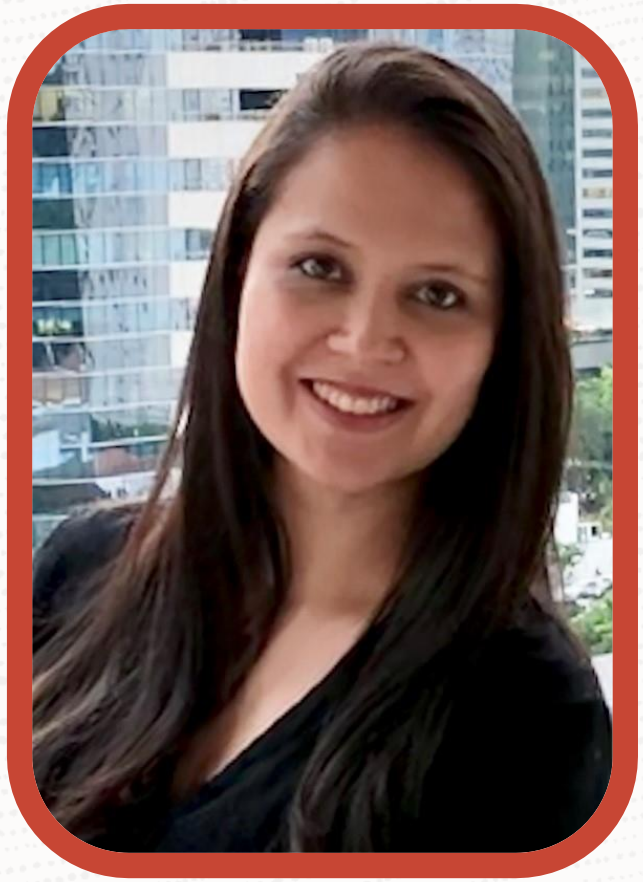


Medalha

PDF

Palestrante,
por favor, confirme com o Coordenador os QR Codes da sua palestra.

$> echo "obrigada"



Erika Nagamine

Cloud Solutions Engineer
Data Management | Data Engineer | Data Scientist | Analytics
Tech Brazil Cloud Solutions Engineer
Oracle

@erikanagamine

@erikanagamine

erika.nagamine@oracle.com

https://github.com/erikanagamine

https://www.linkedin.com/in/erikanagamine

ORACLE