



MANAGEMENT | ANALYTICS | INSIGHTS | SOLUTIONS



MAiS - Management Analytics

Abordagem e Caso SEFAZ-SP

Março 2019



Agenda

1 Management Analytics: Data Science

2 Management Analytics: Big Data

3 Management Analytics na prática – Caso SEFAZ-SP

O Desejo de Todos Atualmente

**“In God we trust.
All others must
bring data”.**

W. Edwards Deming



“

Data is the new oil”

Clive Humby



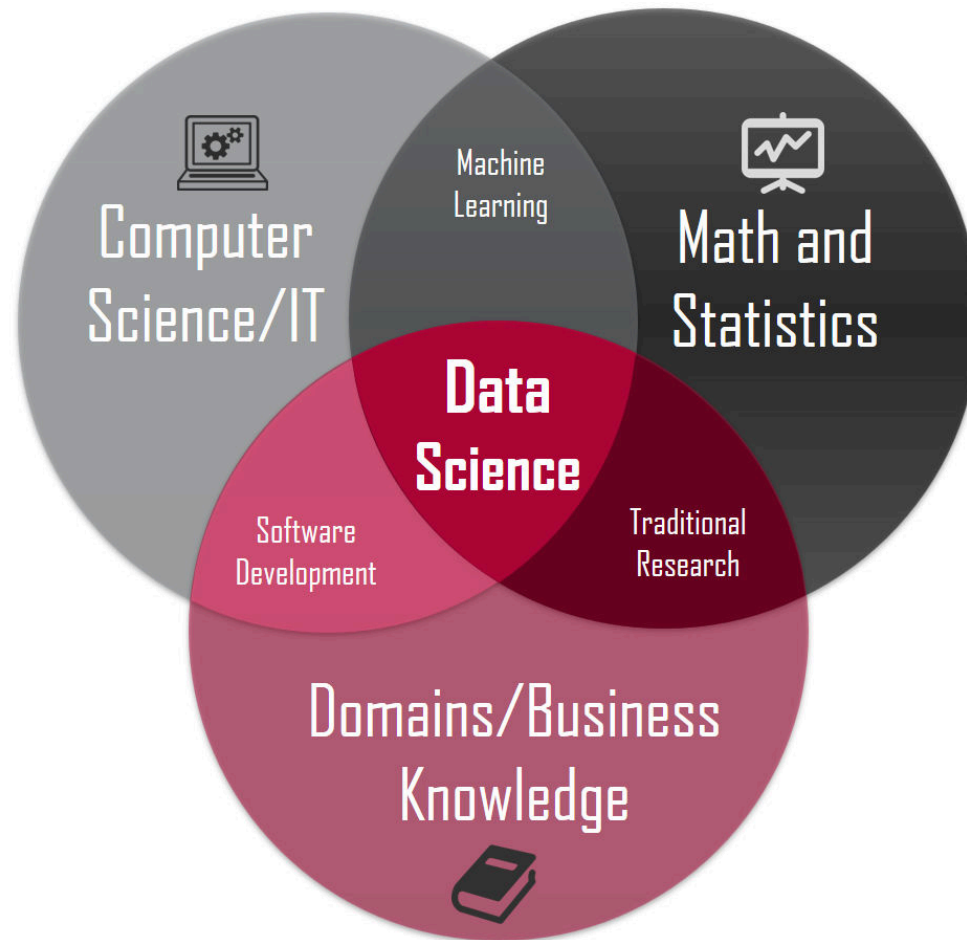


Data Science, é ciência: Alguns experimentos dão certo e outros não

Categoria	Projetos	Em Produção	Resultado (%)
Modelos Probabilísticos	4	1	25%
Modelos Supervisionados	2	1	50%
Modelos Não-Supervisionados	3	1	33%
TOTAL	9	3	33%

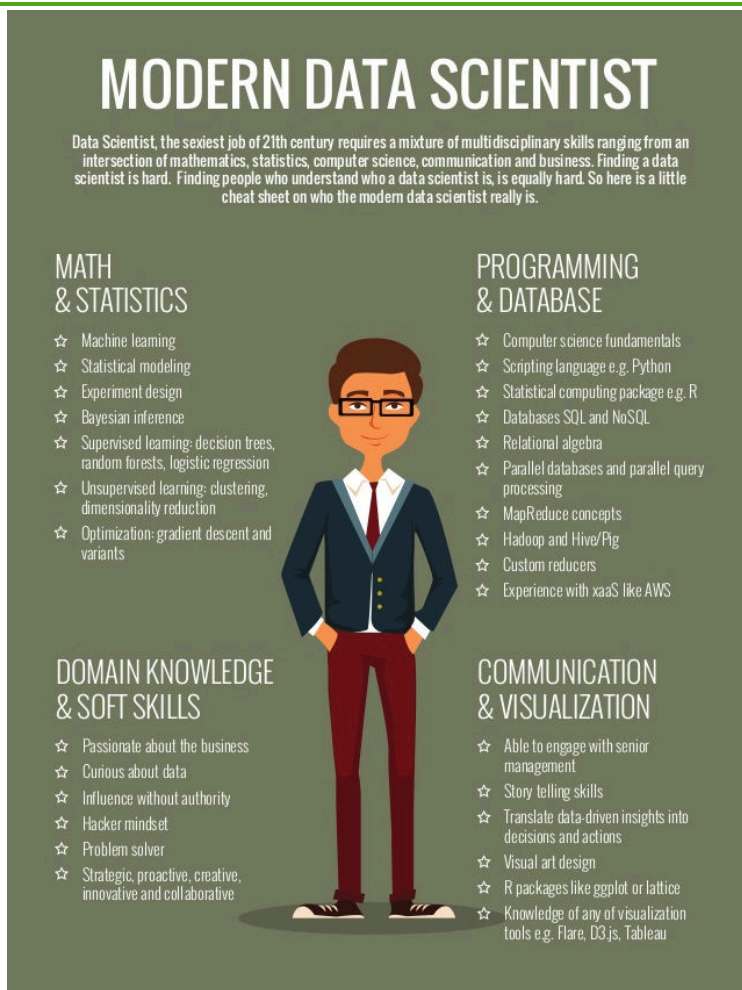


Data Science





O Data Scientist Unicórnio



MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization; customer tracking and on-site analytics; predictive analytics and econometrics; data warehousing and big data systems; marketing channel insights in Paid Search, SEO, Social, CRM and brand.

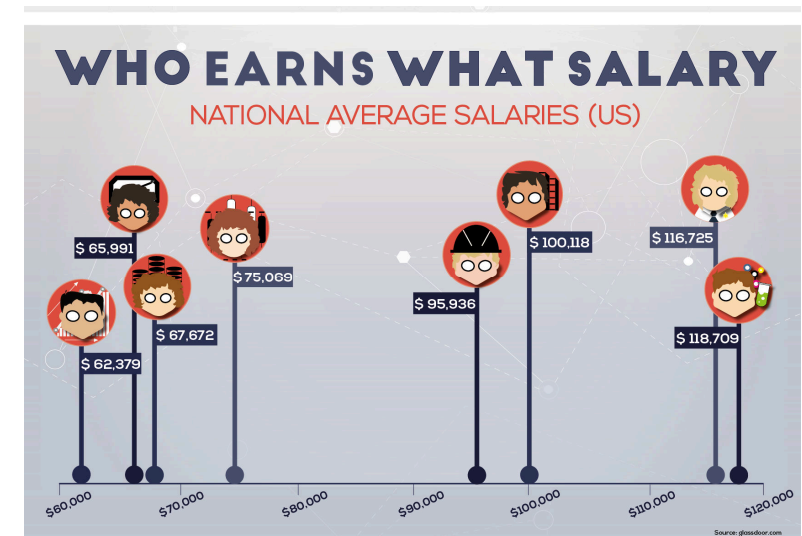
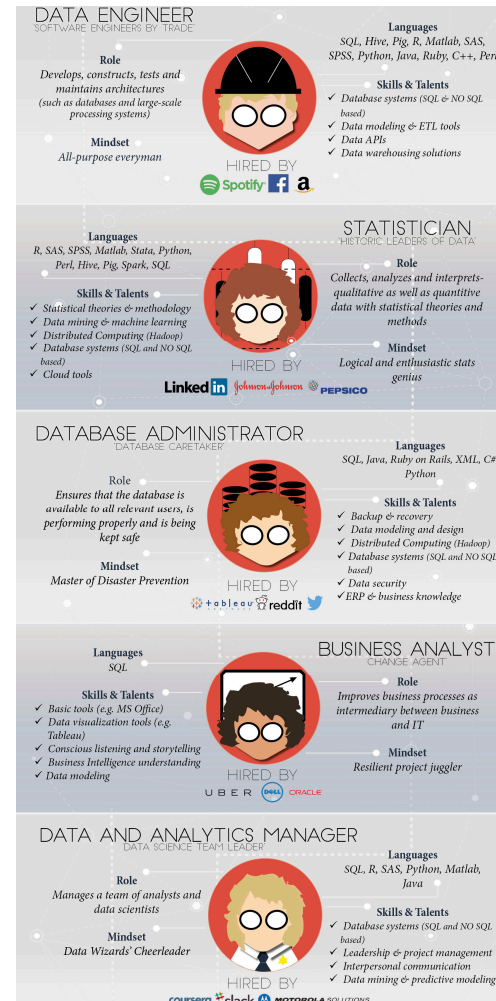
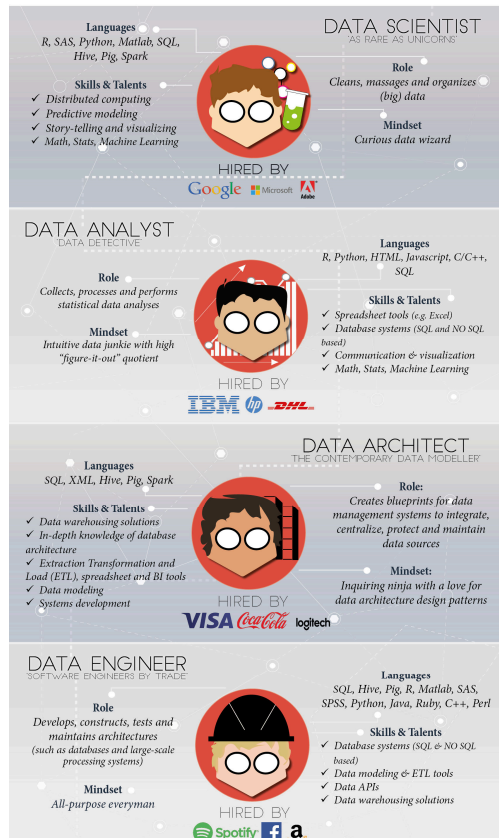
Marketing
DISTILLERY

O que o Data Scientist faz?

- Define a questão problema
- Define a base de dados ideal
- Determina quais dados conseguem acessar
- Obtém o dado
- Limpa o dado
- Faz a análise exploratória do dado
- Faz modelos de previsão classificação
- Interpreta os resultados
- Desafia os dados
- Sintetiza os resultados
- Cria métodos reproduzíveis
- Compartilha os resultados



Especialização dos Papéis no Mundo Data Science



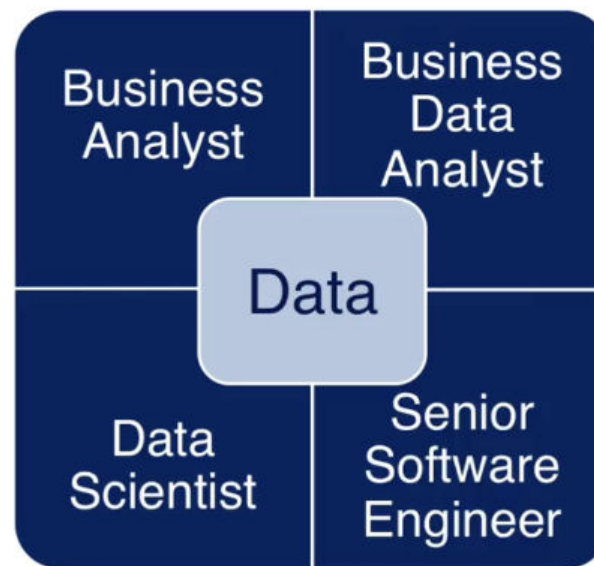


Os papéis necessários

Quem entende do Negócio

Quem entende dos Modelos

e vai ajudar a resolver o problema junto com BA e BDA



Quem entende dos Dados do Negócio

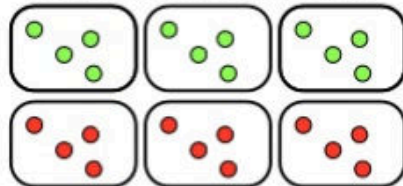
Quem coloca para rodar em Escala

Infra Big Data
Dev e DevOps



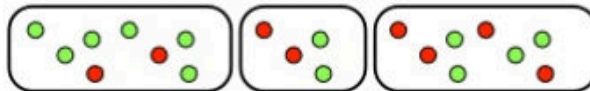
Qual o melhor modelo de Time para Analytics

Centralized Structure:



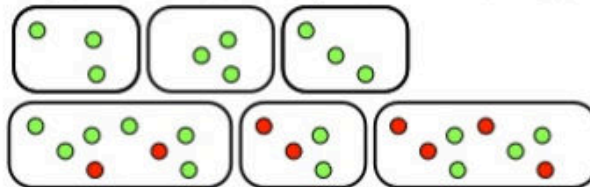
- + 1 Standardized processes
- + 1 Strategic goals met
- 1 costumers goals not met

Decentralized Structure:



- + 1 costumers goals met
- 1 NO Strategic goals met
- 1 Inconsistent & redundant

Hybrid Structure (L-3 DT DS Team):

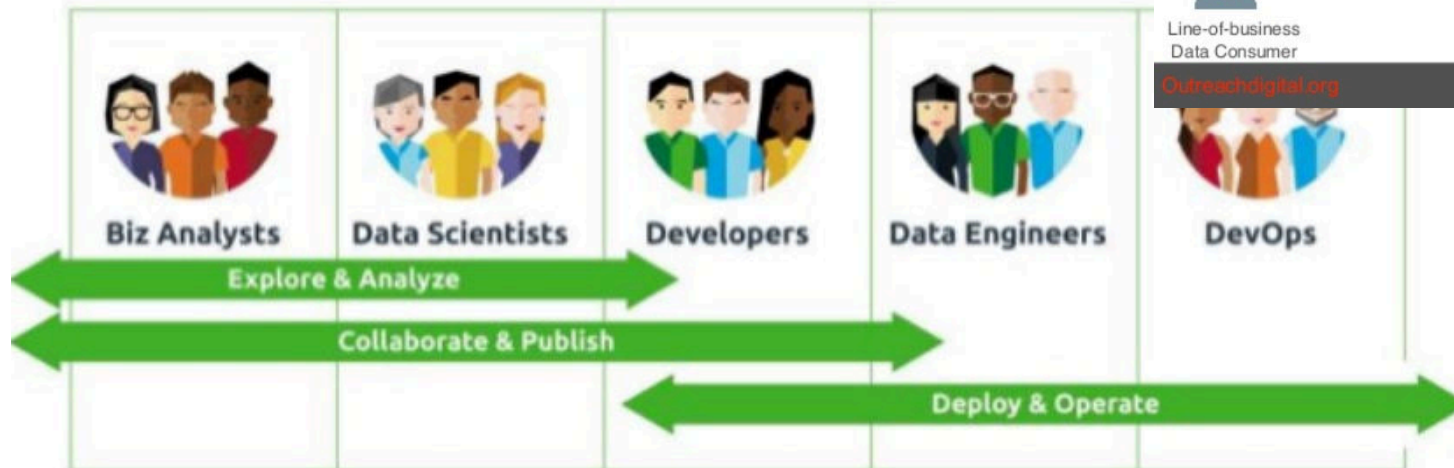


- + 1 standardized processes
- + 1 Strategic goals met
- + 1 costumers goals met



Qual o melhor jeito de organizar as equipes?
Depende do seu objetivo de Negócio

Empowering the Data Science Team



Data Science Team Organization





Nossa solução: Equipe Multidisciplinar



Cientista de dados: o profissional que vai analisar as atividades na rede

Posto foi considerado o mais 'sexy' do século pela revista 'Harvard Business Review'

POR ROBERTA SCRIVANO

24/07/2016 4:30 / atualizado 24/07/2016 10:31

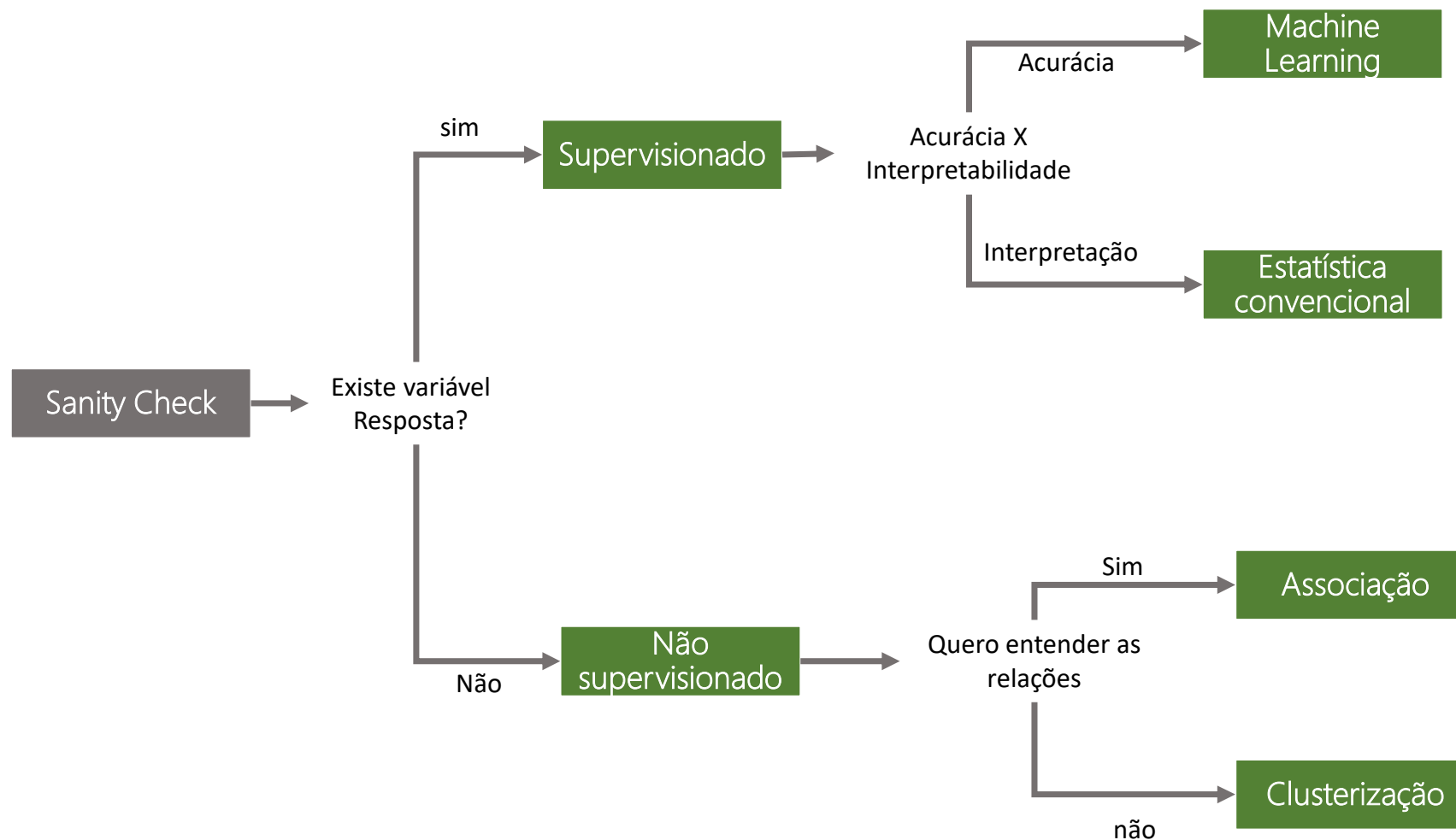


Diego Rodrigues, cientista de dados da MAiS Partners - Domingos Peixoto / Agência O Globo



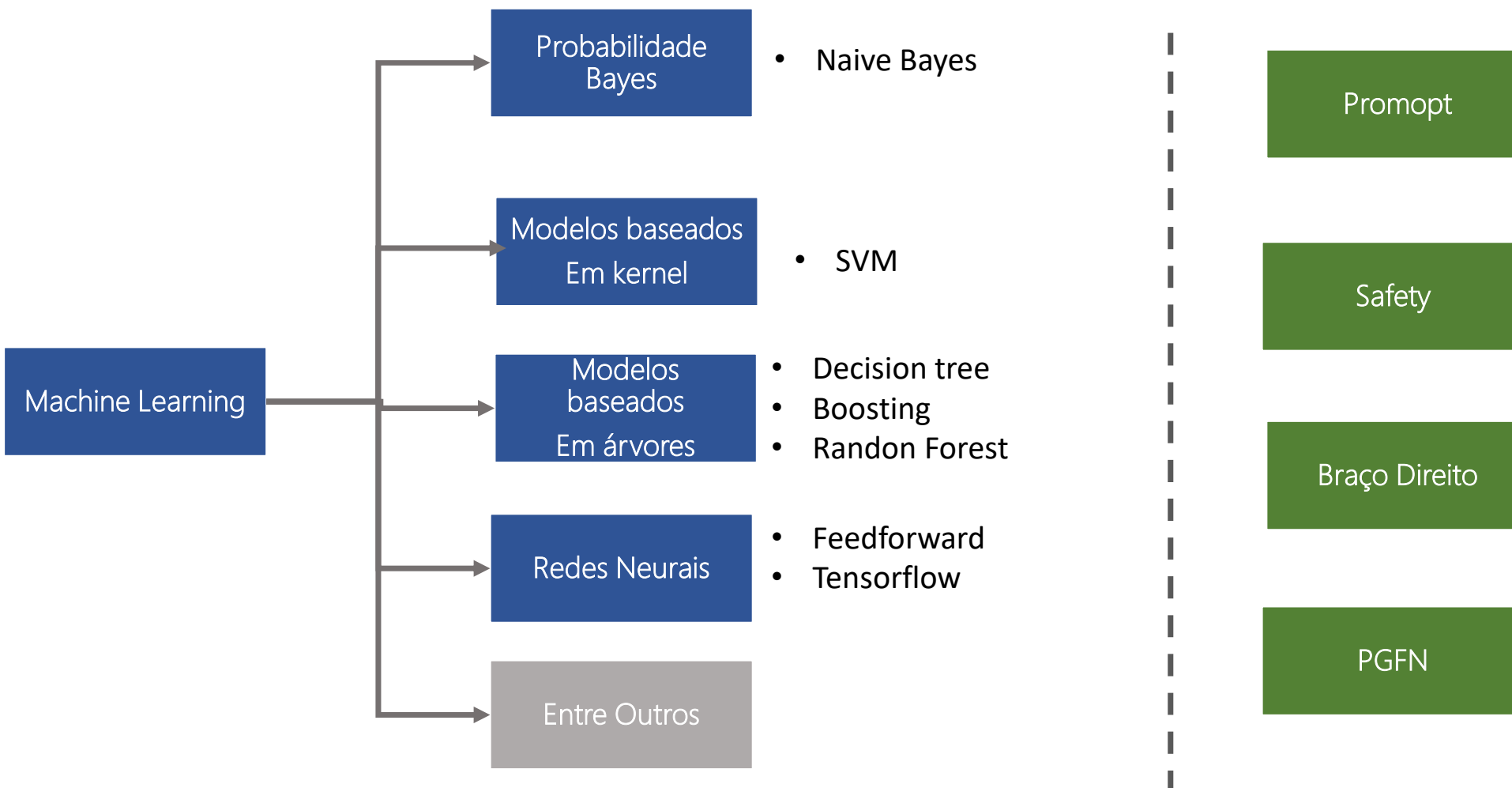
Modelo MAiS de Ciclo de Vida de Analytics

PERGUNTA





Modelo MAiS de Ciclo de Vida de Analytics





Modelo MAiS de Ciclo de Vida de Analytics

Estatística
convencional

Séries Temporais

- Box Jenkins
- Suavizações exponenciais

Regressão

- Regressão Linear
- Regressão Multivariada
- MLG

Otimização

- Programação Linear
- Simplex

Outros

MPST (EDA)

Conjoint 2.0

Supervisores

Operadores



Modelo MAiS de Ciclo de Vida de Analytics





Modelo MAiS de Ciclo de Vida de Analytics

Associação

Algoritmos
Associativos

- A priori
- Cacheiro Viajante

EDA

- Gráficos
- Correlações
- Dashboards
- Comportamento

Cálculos Diversos

- Regra de Negócio
- Integral
- ...

Conjoint Post
Mortem

Tracking

Lucy Montoro



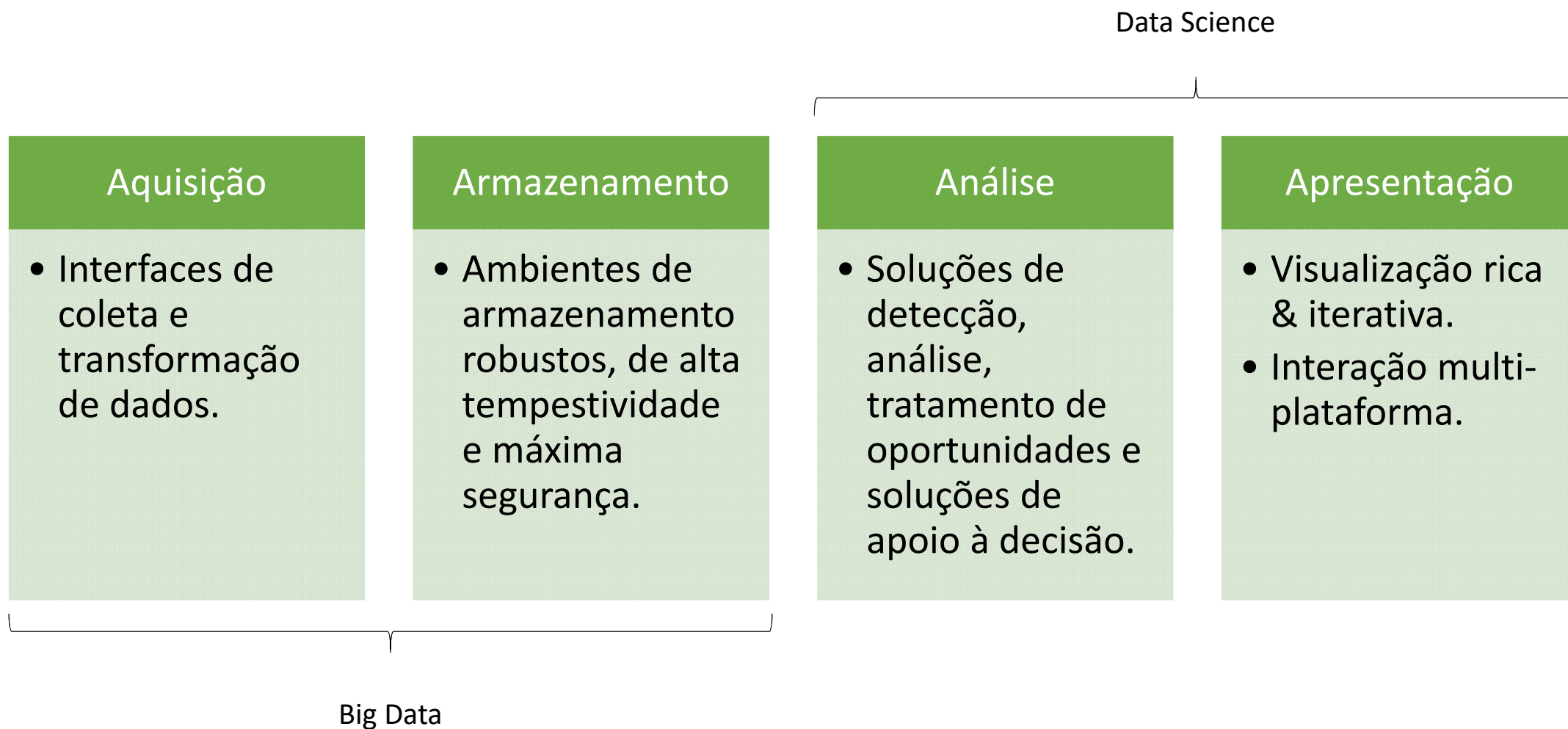
Agenda

1 Management Analytics: Data Science

2 Management Analytics: Big Data

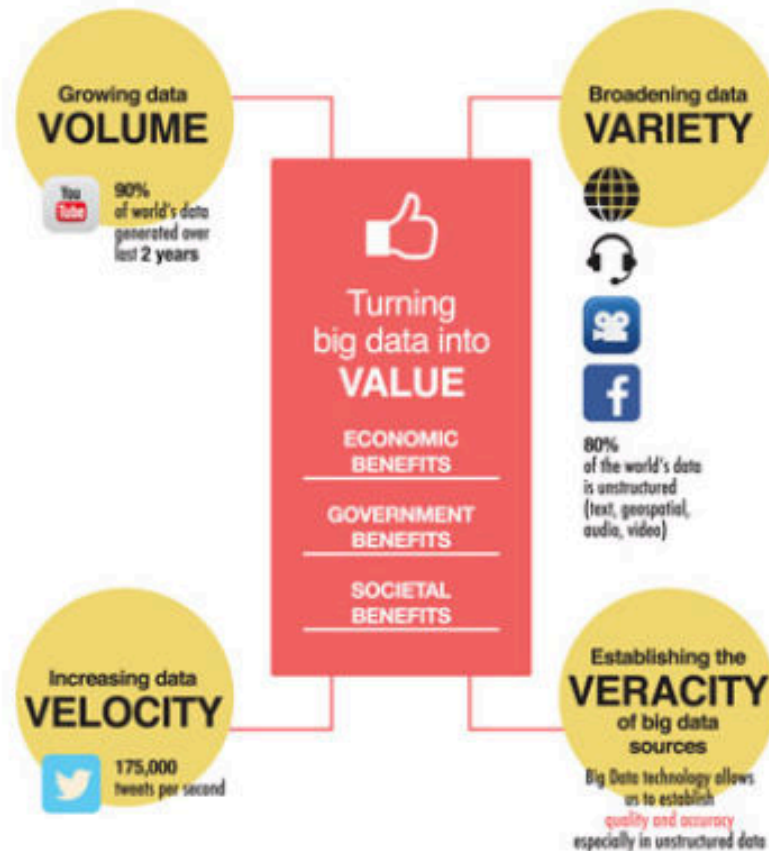
3 Management Analytics na prática – Nossos Casos

Os 4 A's a que analisamos





Turning Big Data into VALUE - 4Vs

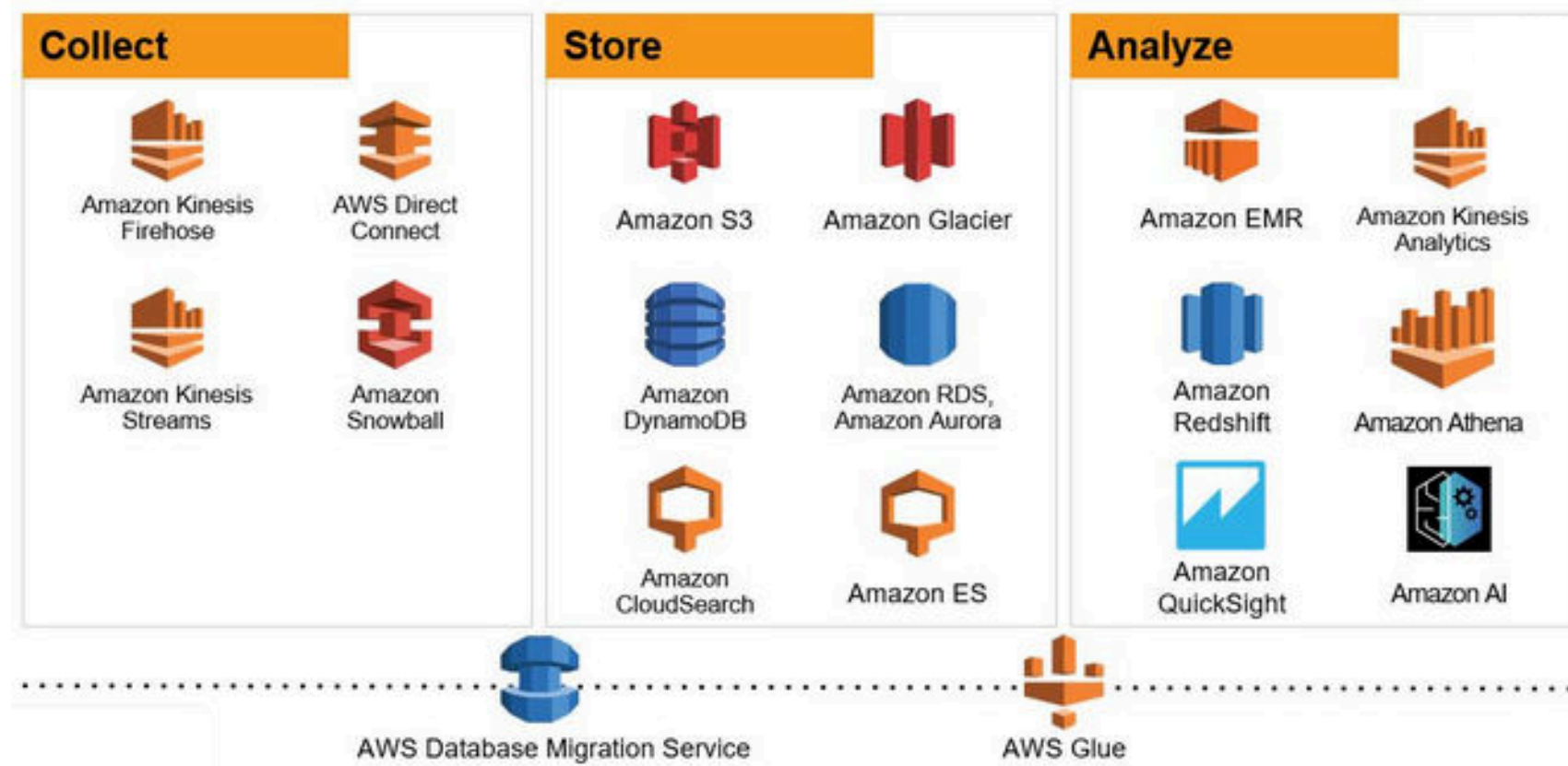


Source: MDeC Presentation on Big Data from 2015 APU's Big Data Week Launch

The Azure Data Landscape



Visão de Arquitetura AWS



Engenheiro de Dados

Cientista de Dados

Analista de Negócio

aquisição

estocagem

limpeza

otimização
predição

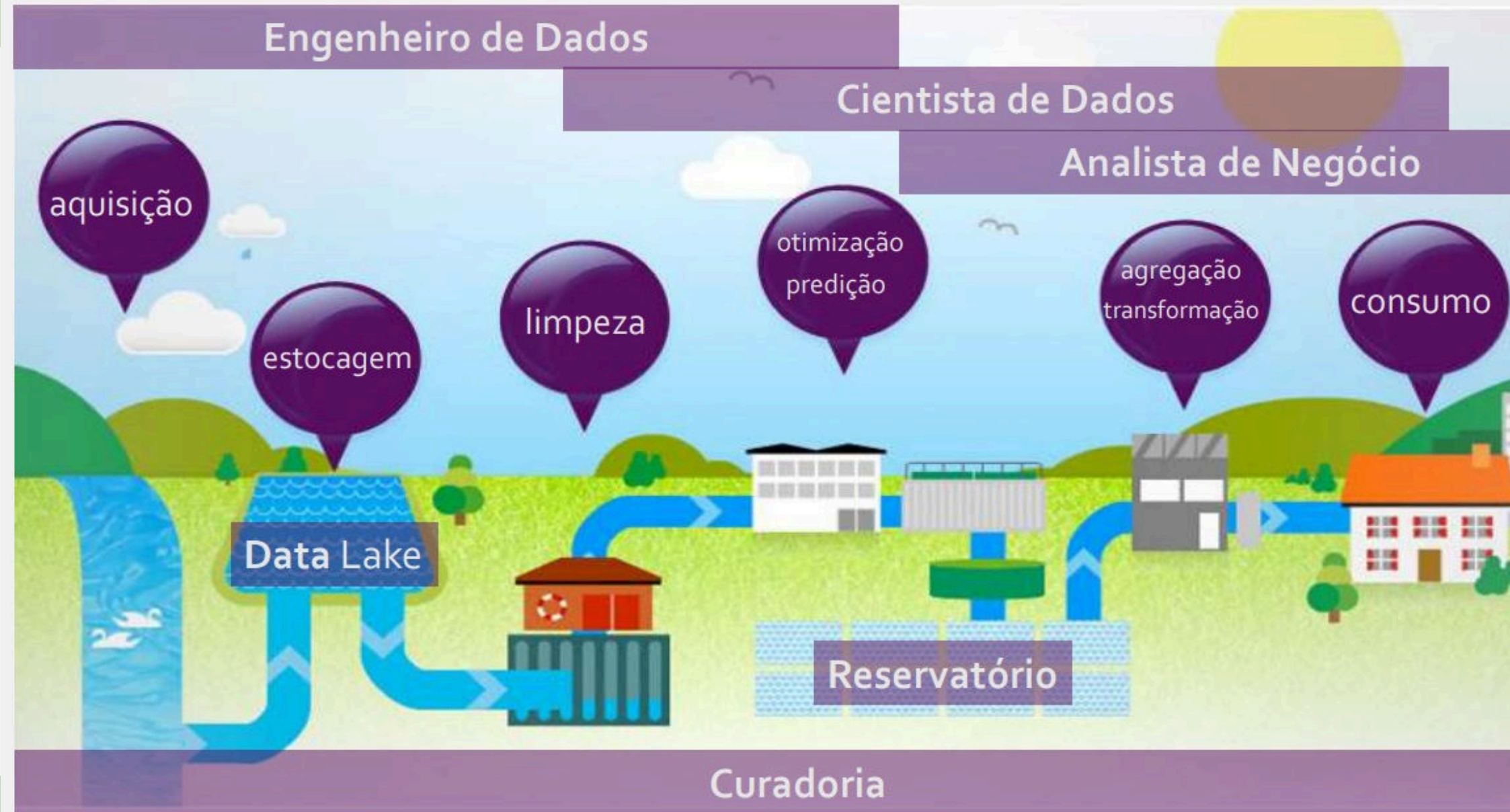
agregação
transformação

consumo

Data Lake

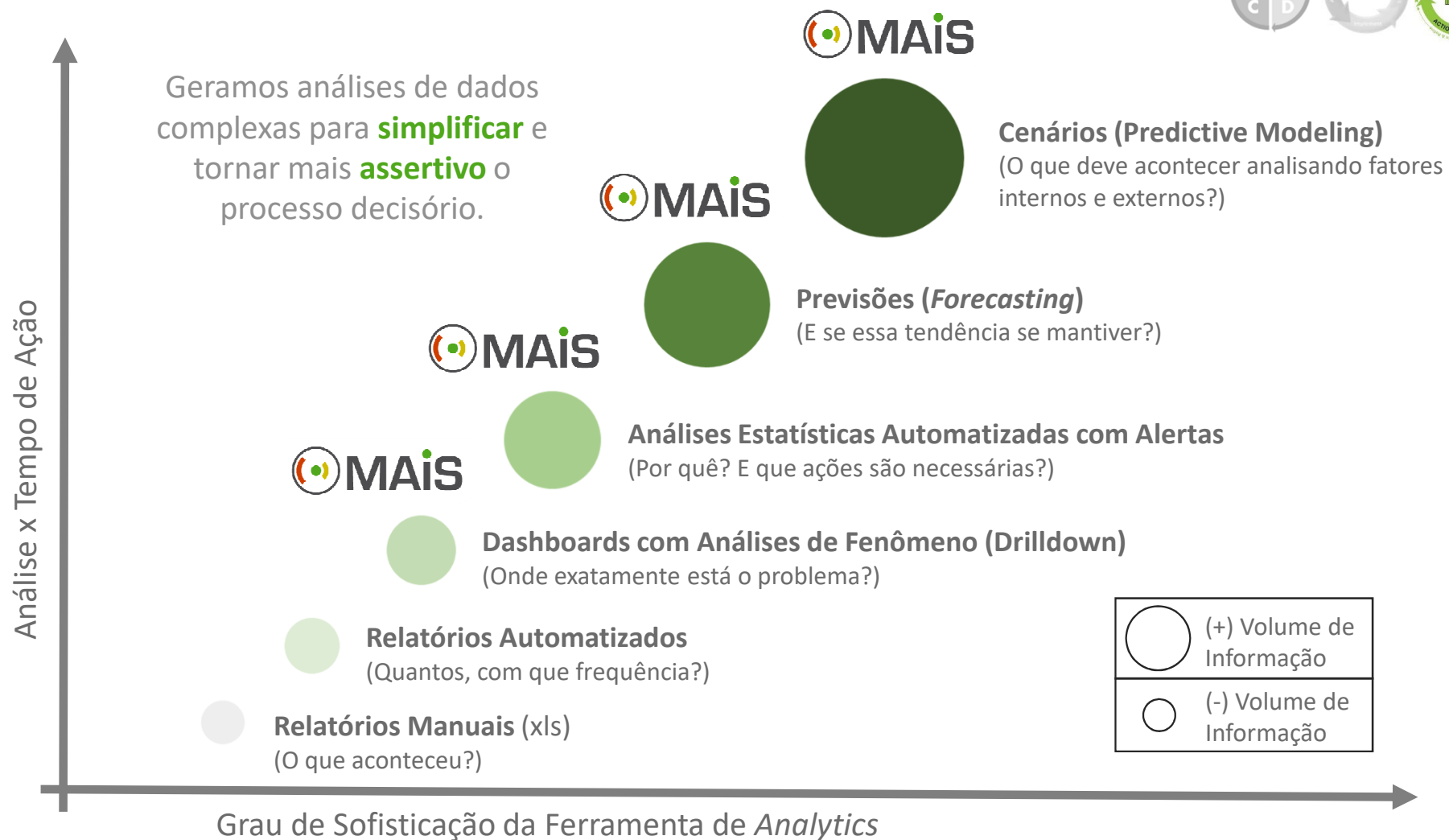
Reservatório

Curadoria





O potencial do Management Analytics





Agenda

1

Management Analytics: Data Science

2

Management Analytics: Big Data

3

Management Analytics na prática – Abordagem e Nossos Casos



Abordagem: Data Science de Ponta a Ponta

1. Algoritmos Especializados (Data Science)

Analizamos profundamente os dados



Unimos Algoritmos e Análise de Negócios



Utilizamos BI e Dashboards intuitivos para



GERAR INSIGHTS
COM BASE NOS
DADOS DO
NEGÓCIO

2. Solução Ponta-a-Ponta (Management Analytics)

Analizamos os dados e a árvore de indicadores da empresa



Via algoritmos definimos as melhorias



Implantamos a melhoria e acompanhamos os resultados

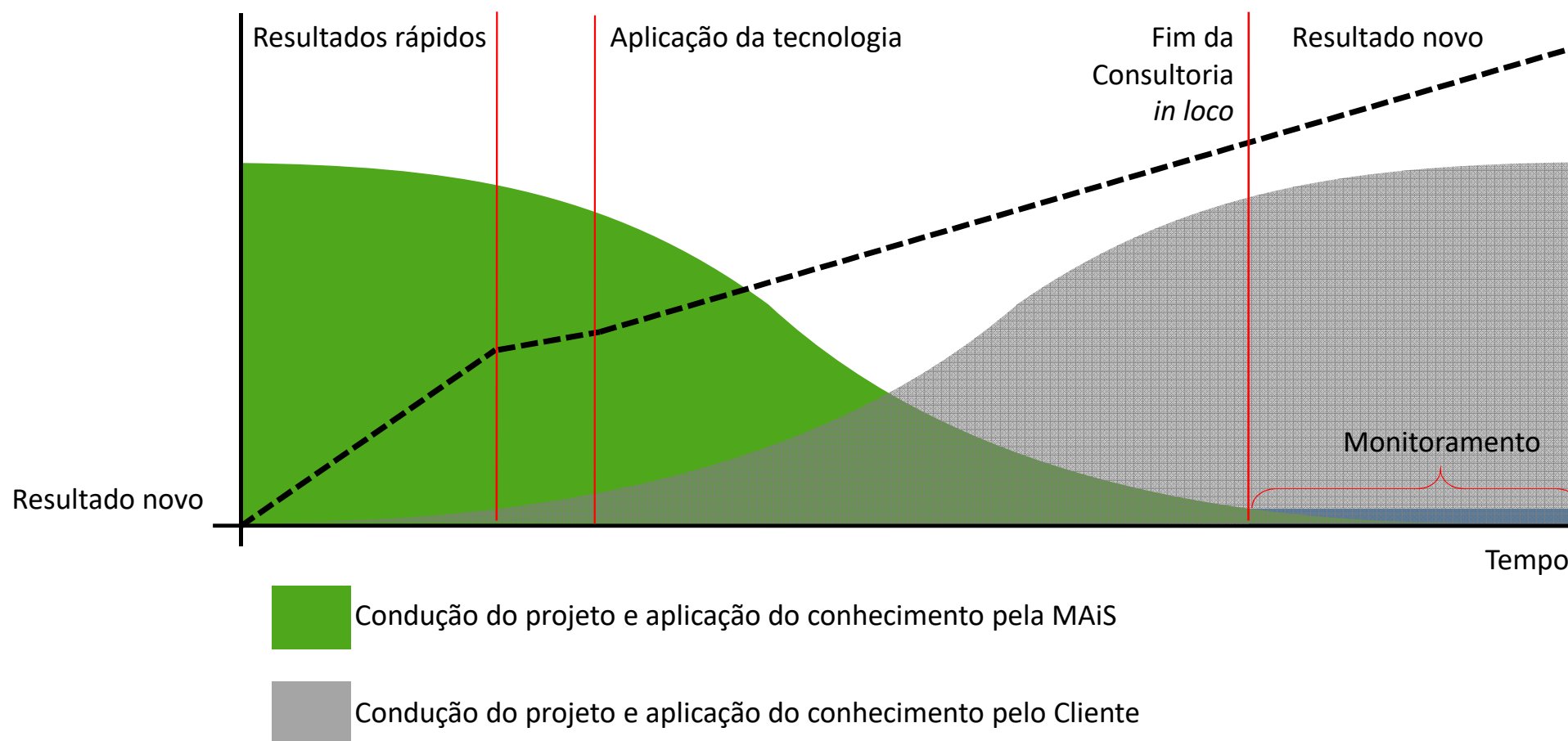


GERAR RESULTADOS
MELHORES E
GANHOS
CONTÍNUOS

Acreditamos em resolver problemas reais, com soluções (algoritmos em produção) que mudem a forma de nossos clientes trabalharem e com resultados reais gerados!

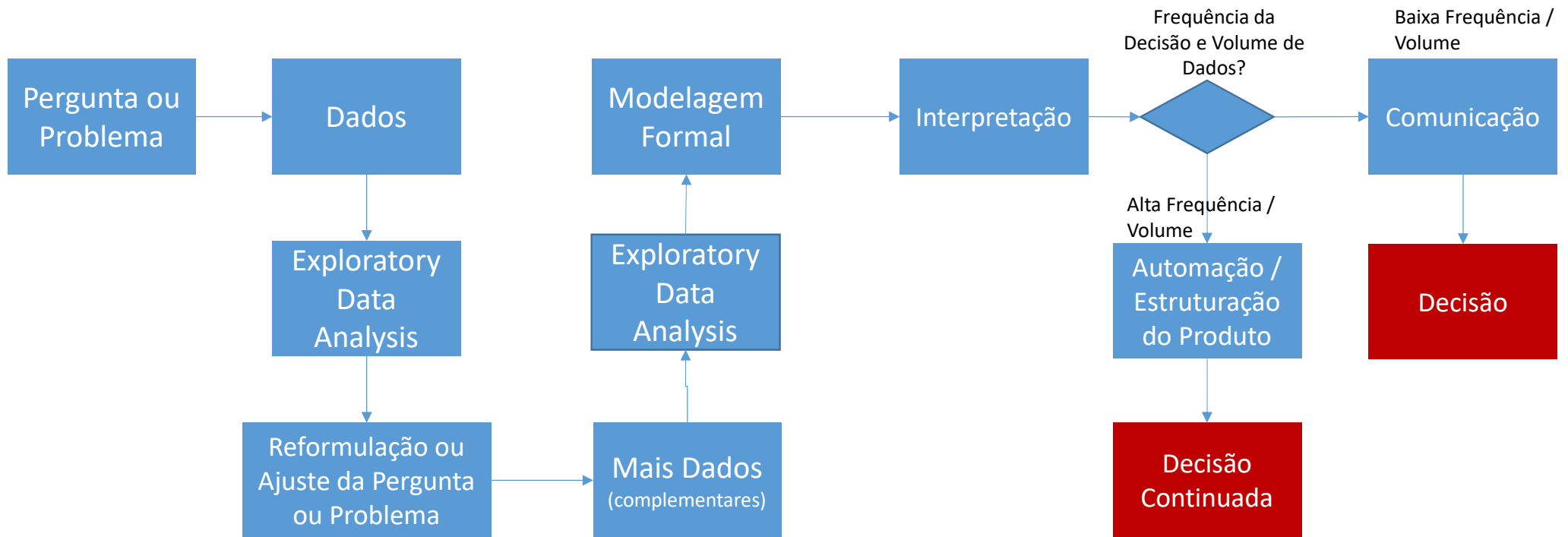


Nossa Abordagem: Gerando Capacidade Institucional em Analytics



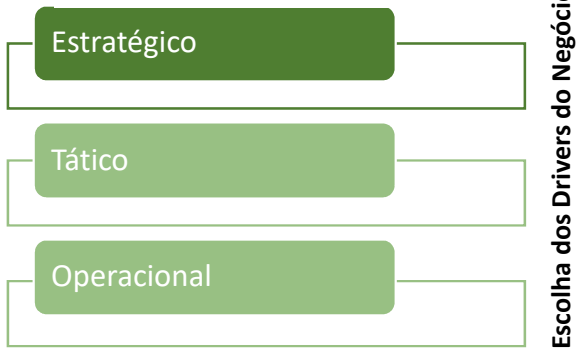


Modelo de Projeto de Data Science – Referencial Teórico

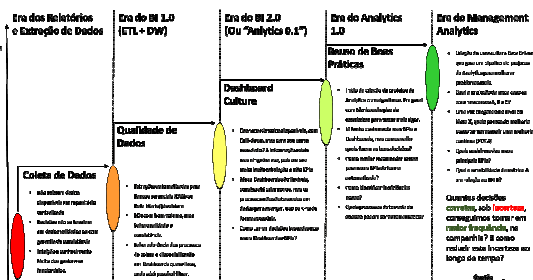


EDA: Exploratory Data Analysis

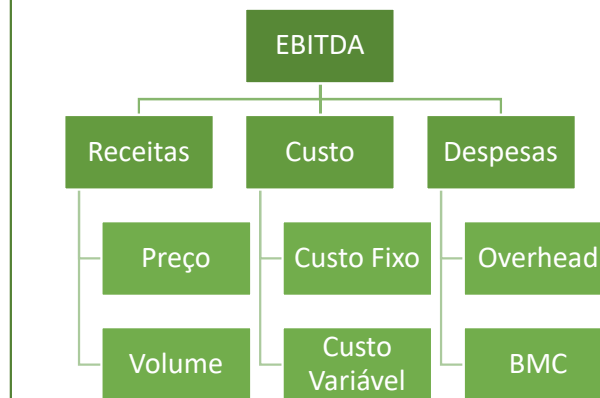
Enquadramento no Sistema de Gestão



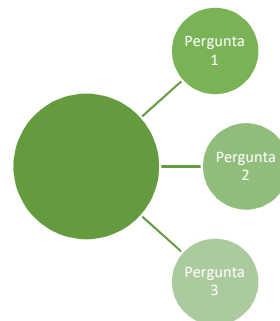
Identificação do Perfil do Sponsor



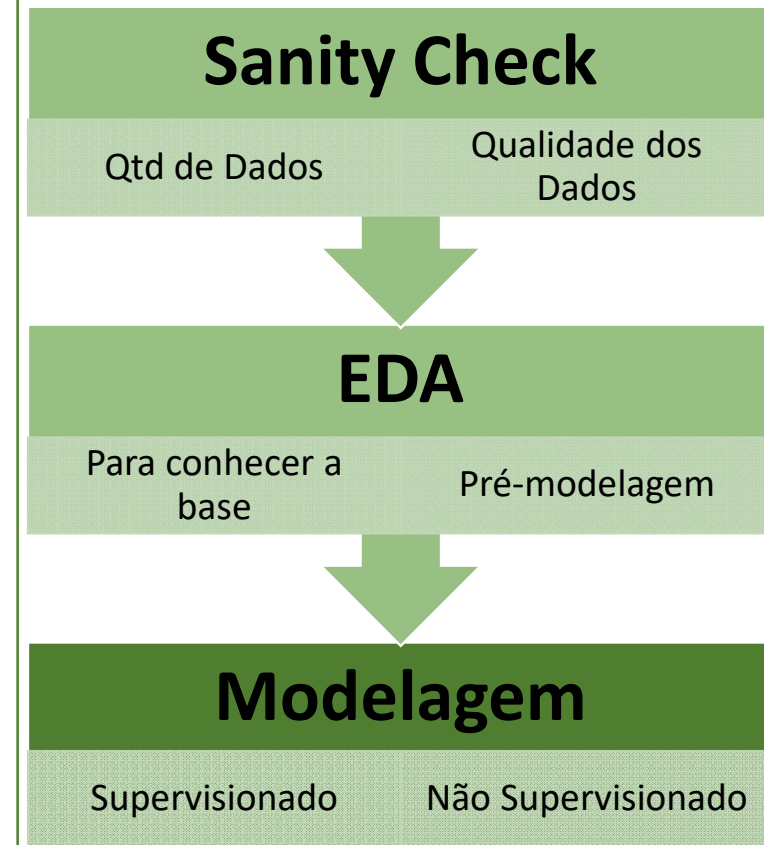
Árvore de Indicadores



Mind Map com as Perguntas por Indicador Chave (KPI)



Modelo de Ciclos de Analytics Por Pergunta





Modelo MAiS de Ciclo de Vida de Analytics



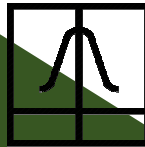
ENTENDIMENTO & PREPARAÇÃO

- Entender os requisitos, dados e objetivos de negócios
- Coletar e preparar os dados para análise
- Sanity check & EDA preliminar



ANÁLISE EXPLORATÓRIA

- Capturar características principais dos dados
- Validar e entender regras de negócios mais profundas
- Balizar estratégias de modelagem
- Elaborar dashboards que facilitem a visualização dos dados



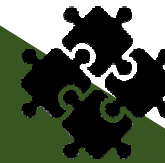
MODELAGEM

- Revisar literatura a procura de problemas parecidos
- Explorar diferentes opções de modelagem
- Desenvolver e comparar provas de conceito



REFINAMENTO

- Reduzir opções de modelagem para as mais efetivas e explorar mais a fundo
- Ampliar abrangência dos resultados e aumentar rigor científico
- Construir e validar solução de modelagem final



ENGENHARIA

- Colocar em produção modelos de alta frequência
- Definir arquitetura da solução e desenvolver integrações e sistemas habilitadores
- Padronizar artefatos de software para conformar com restrições de arquitetura e padrões corporativos



MANUTENÇÃO

- Acompanhar resultado e validar contra o esperado
- Realizar ajustes eventuais nos modelos e sistemas
- Incluir novas funções previstas ou suportadas pela arquitetura e corrigir erros eventuais



Problema

Identificar no conjunto de notas fiscais qual produto BEC cada nota se refere.

- Não homogeneidade do banco: notas fiscais não possuem mesma quantidade de informação
- Erros ortográficos e caracteres especiais
- Abreviações de palavras: diversas abreviações para informações relevantes sobre o produto
- Dualidade na mesma informação: mais de uma forma de falar sobre o mesmo aspecto

*PAPEL SUZANO
REPORT PREMIUM *

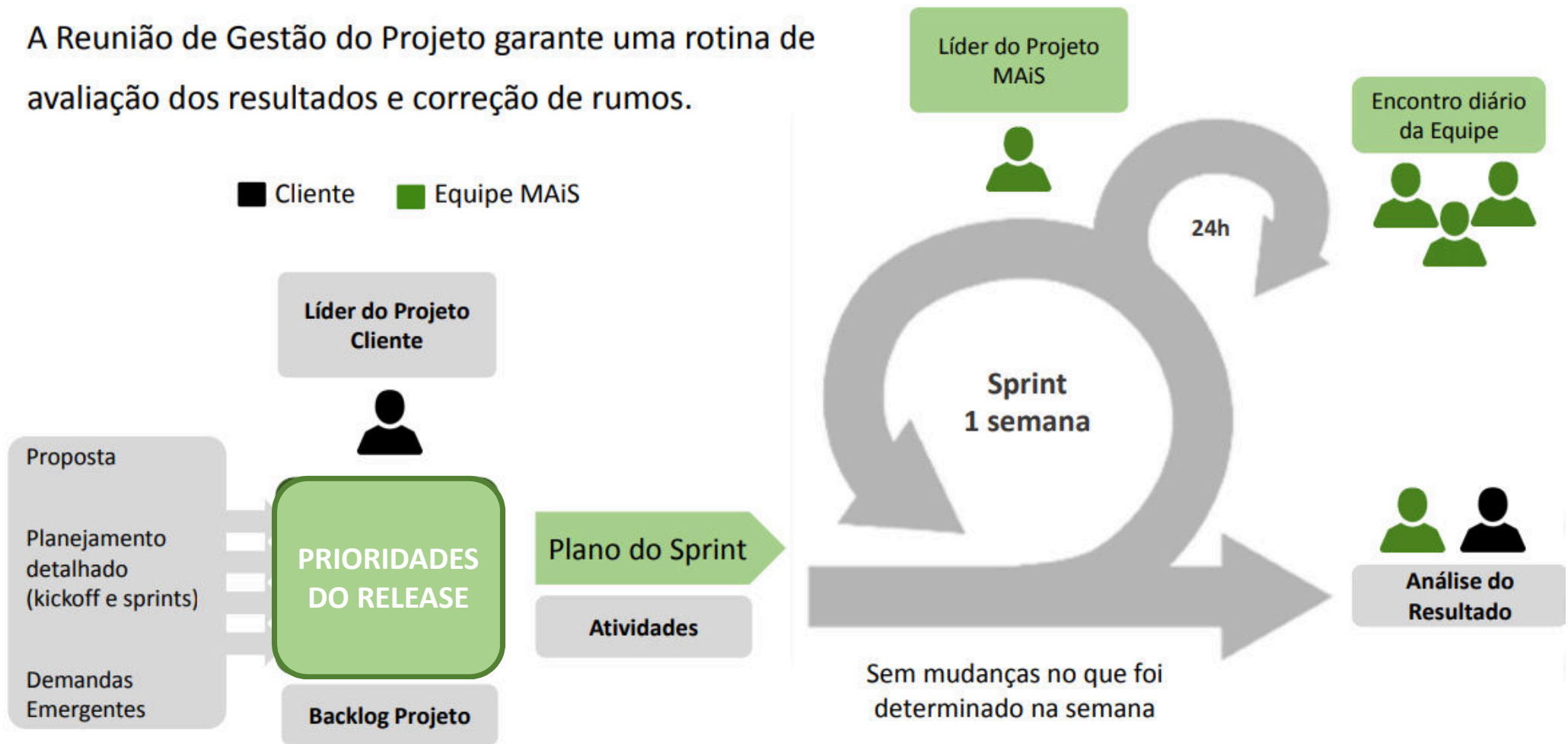
Papil 210x297mm 75g
Brco

SULFT A4 CX C/10 75Gr
- CHAMEQ ROSA



ACORDO NO KICKOFF DE GESTÃO POR SPRINTS SEMANAIS

A Reunião de Gestão do Projeto garante uma rotina de avaliação dos resultados e correção de rumos.

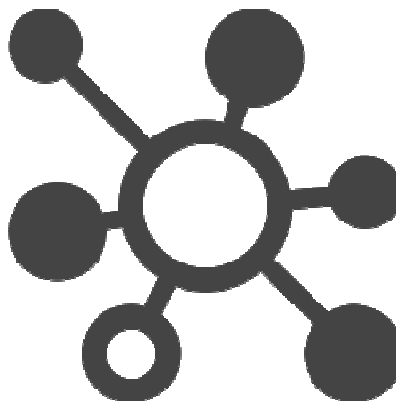


Visão Conceitual das Abordagens



Release 1: Precisão! **20/Dez/18**

Foco na precisão, usando o máximo de supervisão possível, porém com apoio de técnicas não supervisionadas devido a ausência de base de treino



Release 2: Safety Net **11/Jan/19**

Não Supervisionado utilizando clusterização dos dados totais e comparando seu resultado com a Release 1



Release 2: Moon Shot **31/Jan/19**

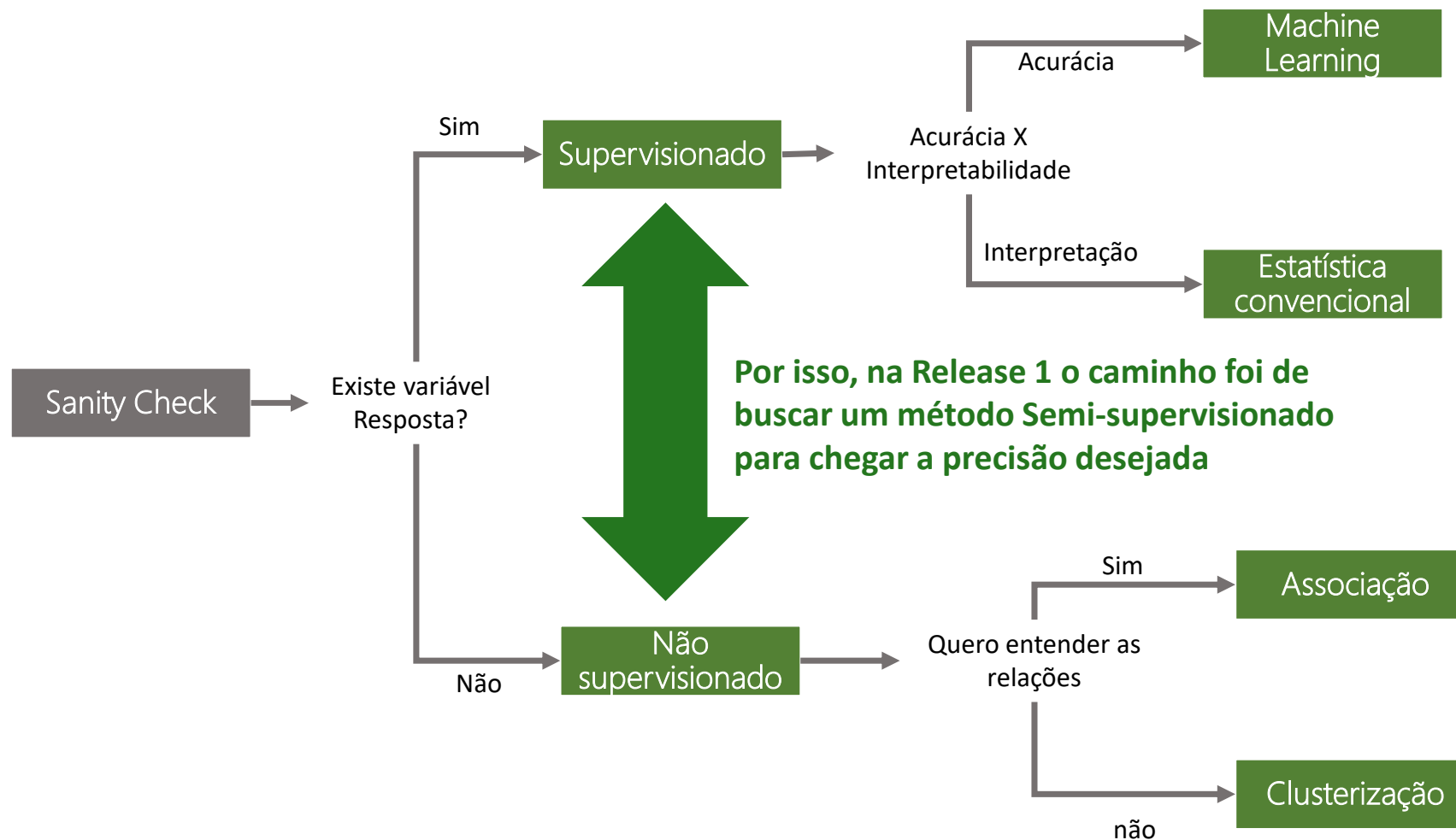
Não Supervisionado utilizando Rede Neural para aprender a tratar os termos mal descritos e comparando seu resultado com a Release 1 e Release 2 Safety Net

**Objetivo de testar um metodologia mais geral para a SEFAZ-SP, que
ganhe escala ao ser estendida para outros SKUs
(Cronograma detalhado no Excel)**



Modelo MAiS de Enquadramento das Técnicas de Analytics

PERGUNTA





Bases recebidas

Base NF-e

Tamanho da base: 14.583 notas fiscais



- Remoção duplicados
- Remoção de produtos que não são papel (tesoura, caneta, lápis, régua)

Tamanho da base: 13.279 notas fiscais

Descrição do Produto

Papel sulfite 75g alcalino 210x297 A4 Chamex Office lpaper PT 500 FL

Papel sulfite 75g alcalino 210x297 A4 Chamex Office lpaper

PAPEL A4 BRANCO-RESMAS C/500 FOLHAS

PAP A4 CERT.FSC GIMBA 210X297 PT 500 75G

Papel sulfite 75g 210x297 A4 chamequinho branco lpaper PT 100 FL

CHAMEX OFFICE A4 BR 75G PT C/ 500FL / 210X297MM - FSC - BANCO (02300)

Papel sulfite 75g alcalino 210x297 A4 Chamex Vai Brasil lpaper PT 500 FL

PAPEL SULFITE A4 75G C/500F 210X297

PAPEL REPORT 210X297 500/A4 75G

PAPEL REPORT A4 PREMIUM 75G-BR 210X297 PT/500F

Base BEC

Tamanho da base: 15 rótulos BEC

Rótulos BEC

PAPEL SULFITE DE PAPELARIA 180G/M2,A4,BRANCO,C/CERTIF.FSC OU CERFLOR

PAPEL SULFITE DE PAPELARIA,120G/M2,A4,BRANCO,EMBALAGEM BOPP

PAPEL SULFITE DE PAPELARIA,75G/M2,A3,BRANCA,CORTE ROTATIVO,EMB.BOPP

PAPEL SULFITE DE PAPELARIA,75G/M2,A4,AMARELA,CORTE ROTATIVO,EMB.BOPP

PAPEL SULFITE DE PAPELARIA,75G/M2,A4,AZUL,CORTE ROTATIVO,EMB.BOPP

PAPEL SULFITE DE PAPELARIA,75G/M2,A4,BRANCO,CANA DE ACUCAR,EMB.BOPP

PAPEL SULFITE DE PAPELARIA,75G/M2,A4,BRANCO,CORTE ROTATIVO,FSC/CERFLOR

PAPEL SULFITE DE PAPELARIA,75G/M2,A4,ROSA,CORTE ROTATIVO,EMB.BOPP

PAPEL SULFITE DE PAPELARIA,75G/M2,A4,VERDE,CORTE ROTATIVO,EMB.BOPP

PAPEL SULFITE DE PAPELARIA,75G/M2,CARTA,BRANCA,CORTE ROTATIVO,EMB.BOPP

PAPEL SULFITE DE PAPELARIA,75G/M2,OFICIO2,BRANCA,CORTE ROTATIVO,EMB.BO

PAPEL SULFITE DE PAPELARIA,75G/M2,OFICIO9,BRANCA,CORTE ROTATIVO,EMB.BO

PAPEL SULFITE DE PAPELARIA,90G/M2,A3,BRANCA,CORTE ROTATIVO,EMB.BOPP

PAPEL SULFITE DE PAPELARIA,90G/M2,A4,BRANCA,CORTE ROTATIVO,EMB.BOPP

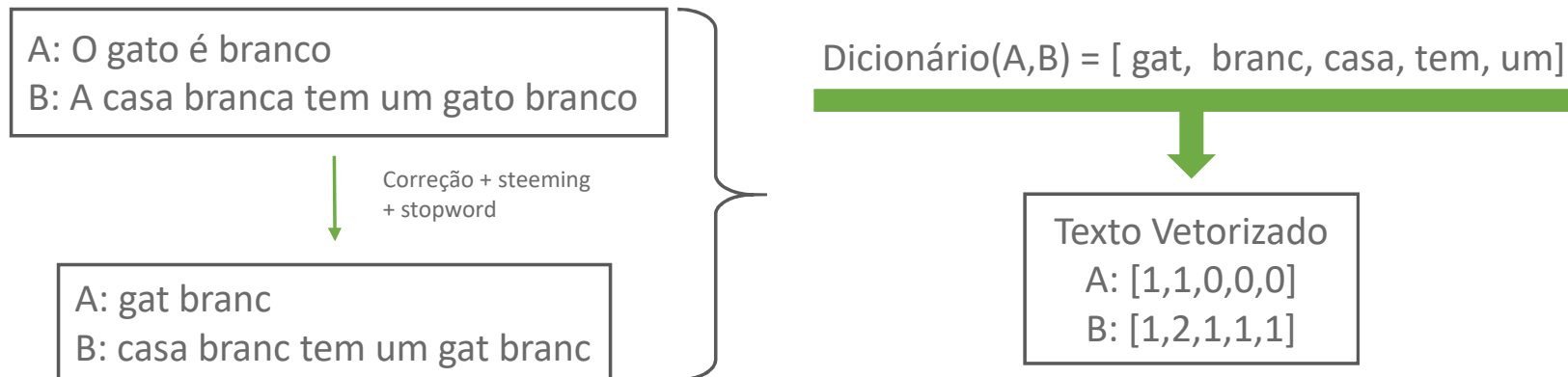
PAPEL SULFITE DE PAPELARIA,FORMATO A4,(210X297)MM

Vetorização da base sem_info

- Vetorização base - Construção matriz documento x palavra

Técnicas utilizadas para vetorização das descrições da base NFE:

- ✓ Stemming de palavras (Ex: Branco – branc / Branca - branc);
- ✓ Dicionário: criação de um dicionário global para todas as palavras presentes na base de Nfe;
- ✓ Bag-of-words e TF-IDF: cada NF-e é representada por um vetor com dimensão igual ao tamanho do dicionário criado.



Limpeza das bases – NF-e e BEC

PAPEL-SULFITE DE 75G C/500F 210X297

Apenas letras minúsculas



papel-sulf de 75g c/500f 210x297

Substitui abreviações e
uniformiza medidas



papel sulfite de 75g com 500 folhas a4

Remove caracteres especiais



papel sulf de 75g c 500f 210x297

Remove stopwords



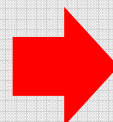
papel sulfite 75g 500 folhas a4

Stemming de palavras



pap sulfit 75g 500 folh a4

PAPEL-SULFITE DE 75G C/500F 210X297



papel sulfit 75g 500 folh a4

Exemplo Limpeza na Base NF-e

Base NF-e - Descrição Produto

Papel sulfite 75g alcalino 210x297 A4 Chamex Office Ipaper PT 500 FL

Papel sulfite 75g alcalino 210x297 A4 Chamex Office Ipaper

PAPEL A4 BRANCO-RESMAS C/500 FOLHAS

PAP A4 CERT.FSC GIMBA 210X297 PT 500 75G

Papel sulfite 75g 210x297 A4 chamequinho branco Ipaper PT 100 FL

CHAMEX OFFICE A4 BR 75G PT C/ 500FL / 210X297MM - FSC - BANCO (02300)

Papel sulfite 75g alcalino 210x297 A4 Chamex Vai Brasil Ipaper PT 500 FL

PAPEL SULFITE A4 75G C/500F 210X297

PAPEL REPORT 210X297 500/A4 75G

PAPEL REPORT A4 PREMIUM 75G-BR 210X297 PT/500F

Base NF-e - Descrição Produto Limpa

papel sulfit 75g alcalin a4 chamex offic ipap pt 500 folh

papel sulfit 75g alcalin a4 chamex offic ipap

papel a4 branc resm 500 folh

papel a4 cert fsc gimb a4 pt 500 75g

papel sulfit 75g a4 chamec branc ipap pt 100 folh

chamex offic a4 br 75g pt 500 folh fsc banc 02300

papel sulfit 75g alcalin a4 chamex vai brasil ipaper pt 500 folh

papel sulfit a4 75g 500 folh

papel report 500 a4 75g

papel report a4 premium 75g br pt 500 folh



Base info-bec : Regras de negócio



Tradução das regras de negócio em um **sistema lógico** de tomada de decisão.



Desenvolvimento de um **script de classificação** que aplica o sistema lógico em uma nota fiscal.

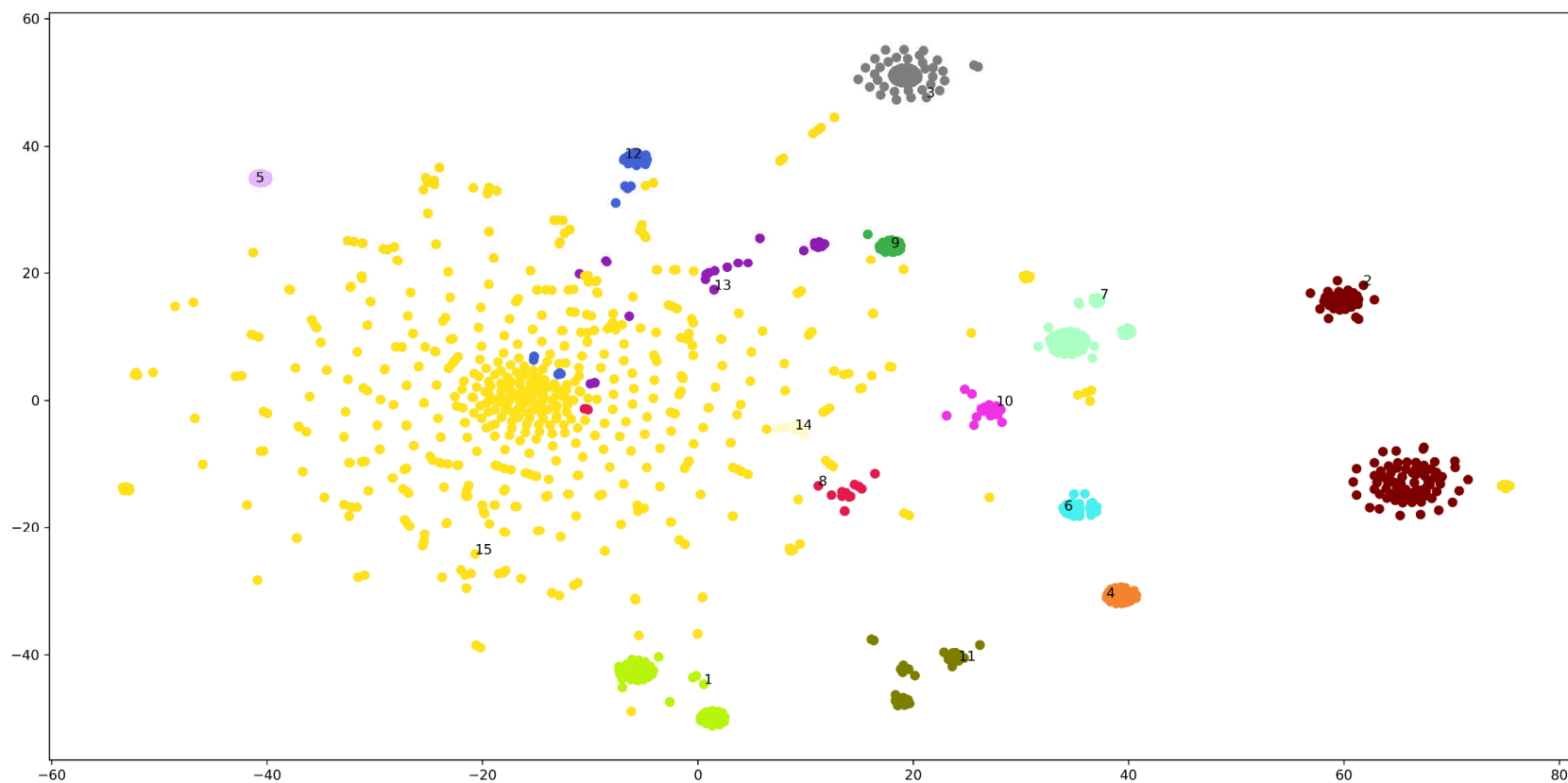


Aplicação do script de classificação em todas as notas fiscais da base info_bec; **armazenar resultados**.



Clusterização Hierárquica

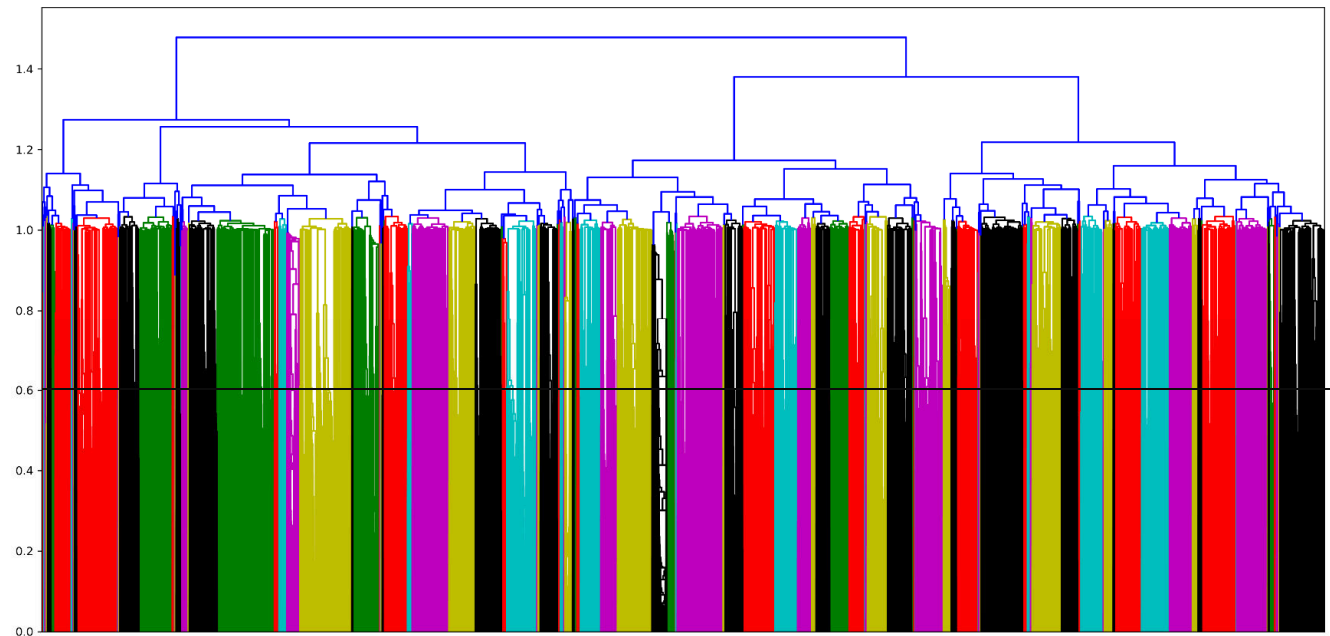
- ✓ Número de clusters encontrados: 15
- ✓ Cluster com possíveis papel A4 branco: 910 notas fiscais (representado pela cor amarela)





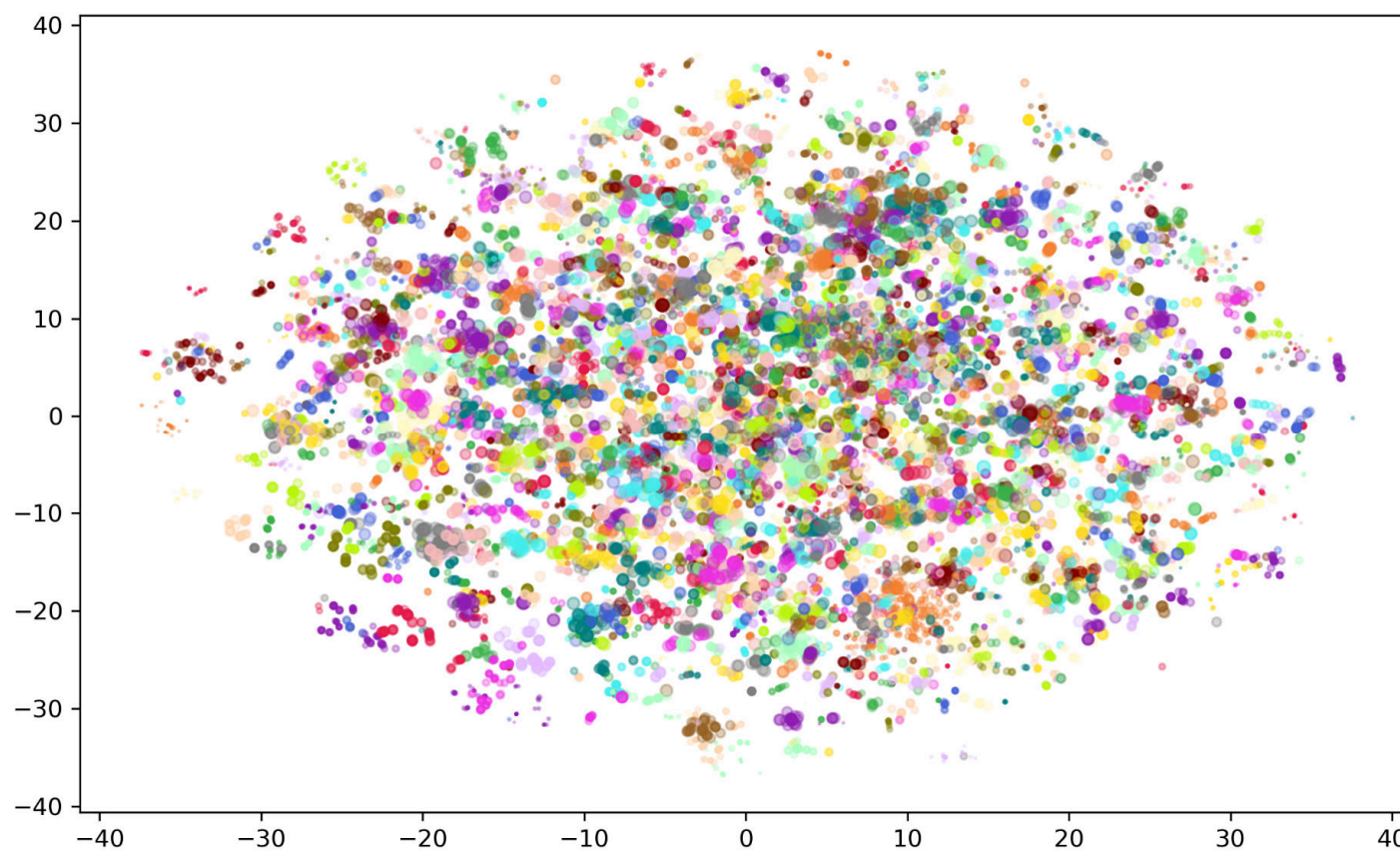
Clusterização Hierárquica - Adaptação

- Modificação do algoritmo para maior conjunto de dados;
- Cluster Hierárquico
 - Método: Complete;
 - Distância: Cosine;
- Número de clusters: 4315;
- 71% de acerto na rotulação





Scatter Plot das notas fiscais clusterizadas



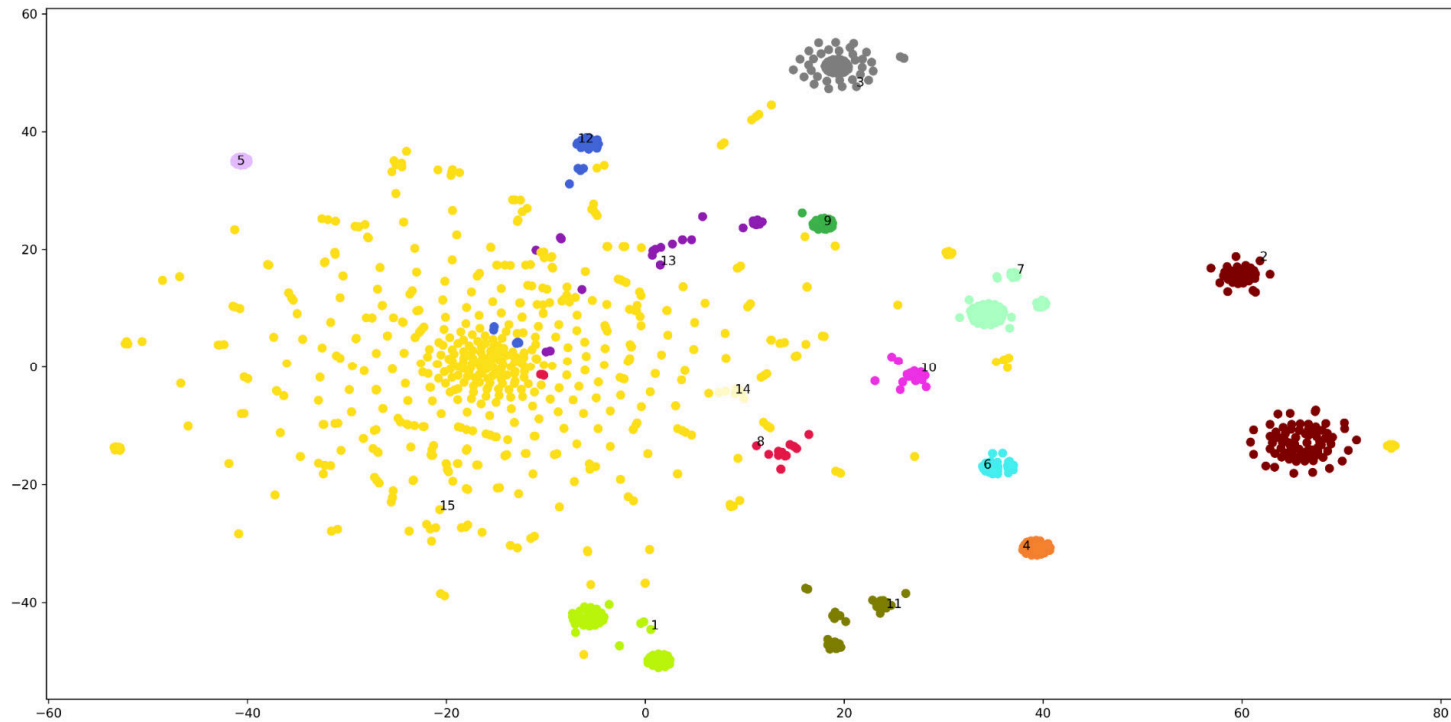


Comparação Resultados 1.5

Rótulo BEC	Regras de Negócio	Clusterização
PAPEL SULFITE DE PAPELARIA,FORMATO A4,(210X297)MM	62%	81%
NAO PRODUTO BEC	15%	7.8%
PAPEL SULFITE DE PAPELARIA,75G/M2,A4,BRANCO,CANA DE ACUCAR,EMB.BOPP	10%	5.2%
PAPEL SULFITE DE PAPELARIA,75G/M2,A4,AMARELA,CORTE ROTATIVO,EMB.BOPP	8%	2.9%
PAPEL SULFITE DE PAPELARIA,90G/M2,A4,BRANCA,CORTE ROTATIVO,EMB.BOPP	1.2%	1.4%
PAPEL SULFITE DE PAPELARIA,75G/M2,CARTA,BRANCA,CORTE ROTATIVO,EMB.BO	0.97%	0.3%
PAPEL SULFITE DE PAPELARIA,120G/M2,A4,BRANCO,EMBALAGEM BOPP	0.44%	0.23%
PAPEL SULFITE DE PAPELARIA,75G/M2,A3,BRANCA,CORTE ROTATIVO,EMB.BOP	0.41%	0.29%
PAPEL SULFITE DE PAPELARIA,75G/M2,OFICIO2,BRANCA,CORTE ROTATIVO,EMB.BO	0.39%	0.21%
PAPEL SULFITE DE PAPELARIA,75G/M2,OFICIO9,BRANCA,CORTE ROTATIVO,EMB.BO	0.23%	0.11%
PAPEL SULFITE DE PAPELARIA 180G/M2,A4,BRANCO,C/CERTIF.FSC OU CERFLOR	0.09%	0.06%
PAPEL SULFITE DE PAPELARIA,90G/M2,A3,BRANCA,CORTE ROTATIVO,EMB.BOPP	0.03%	0.01%



Resultados



Supervisionado

92% de Acurácia

Não Supervisionado

71% de Acurácia



Conclusões

- ✓ A metodologia utilizada se mostrou eficiente, podendo ser utilizada para a **identificação de novas SKU's**;
- ✓ A classificação pelas regras de negócio atua como **validação** para clusterização;
- ✓ A metodologia unindo ambas técnicas pode ser refinada de forma a otimizar a acurácia das rotulações;
- ✓ A utilização de uma **rede neural** para substituição de abreviações torna possível um processo inteiramente não supervisionado;

Quer  **MAIS?**

Gabriel Renault

gabriel.renault@maispartners.com

21 98187-2663

www.maispartners.com

