DATA SCIENCE NA PRÁTICA

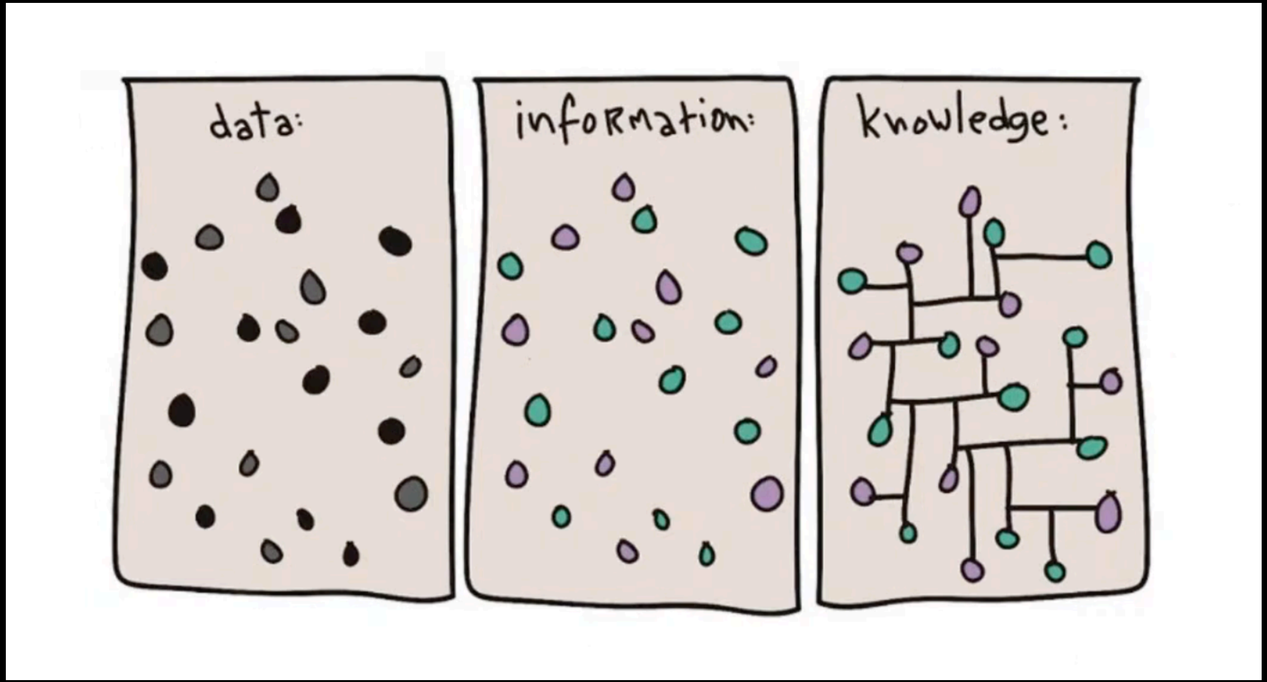
sigmoidal

Alexandre Magno Gomes Bezerra, alexandre.magno.gomes@gmail.com - CPF: 016.841.000-20

O que é Ciência de Dados?

Ciência
de dados

- Defina o objetivo da pesquisa
- Faça uma hipótese
- Colete dados
- Teste sua hipótese
- Analise seus resultados
- Chegue a uma conclusão
- Refine a hipótese e repita



1. Defina o objetivo da pesquisa	Quero prever o congestionamento no trânsito de São Paulo para um determinado dia
2. Faça uma hipótese	Eu acho que o clima é uma informação relevante.
3. Colete dados	Pegue dados históricos do congestionamento e clima para cada dia.
4. Teste sua hipótese	Treine um modelo usando esses dados.
5. Analise seus resultados	Esse modelo é melhor que os sistemas já existentes?
6. Chegue a uma conclusão	Eu (não) deveria usar esse modelo para fazer previsões por causa de X, Y e Z.
7. Refine a hipótese e repita	Época do ano pode ser uma informação relevante.

Data Mining (Data Science)

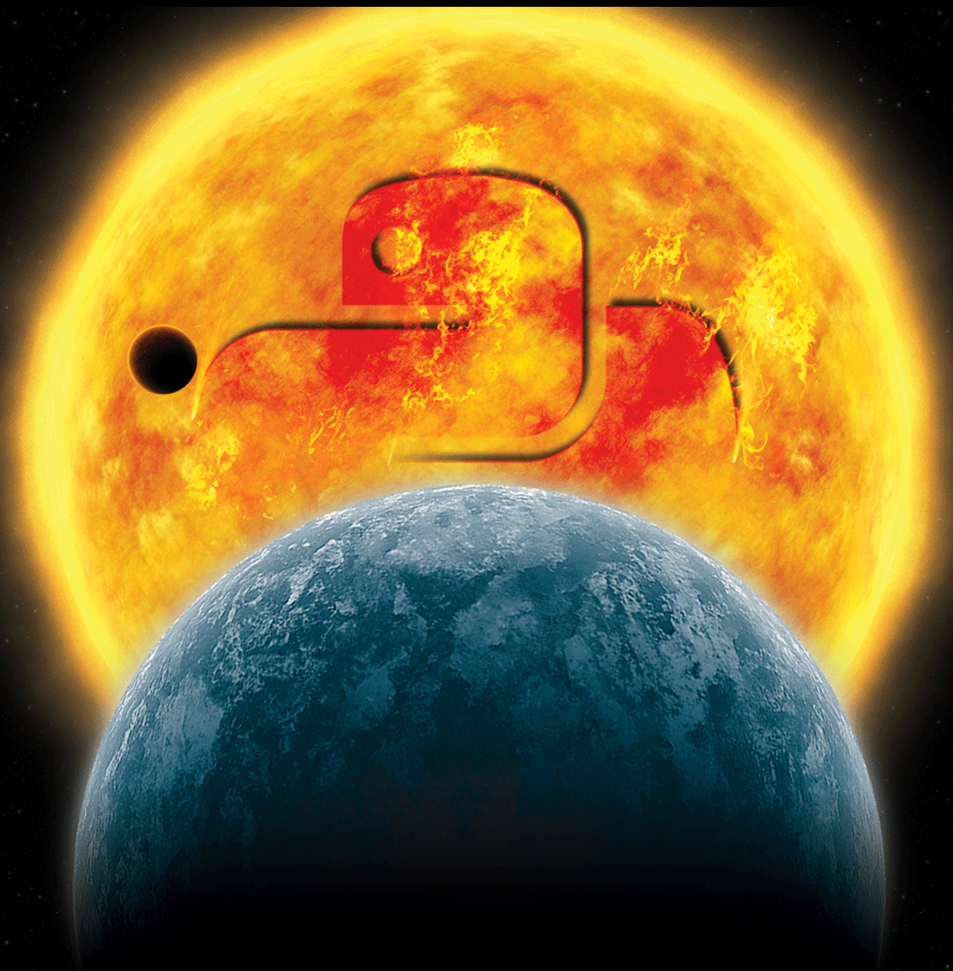
Processo com estágios bem definidos

- Envolvem a aplicação de tecnologia da informação como a descoberta automatizada e a avaliação de padrões a partir de dados; ou
- (…) enquanto outros exigem criatividade, conhecimento de negócios e bom senso por parte do analista.

Alexandre Magno Galieta de Oliveira - alexandre.galieta@gmail.com - CPF: 618.821.006-20

Por que Python?

A linguagem que mais cresce no mundo.



Dados

O Novo Petróleo

Alexandre Magno Galieta de Oliveira - alexandre.galieta@gmail.com - CPF: 618.821.006-20

Data Analytical Thinking
e
Data-Driven Decision Making

100%

das empresas precisam de Ciência de Dados

Pense no processo de machine learning como um experimento.

Data Science está em toda parte

- Operações
- Marketing
- Vendas
- Manufatura
- Gestão da Cadeia de Suprimentos
- Customer Experience
- Fluxos e Processos de Trabalho

Como consigo utilizar Data Science na minha empresa?

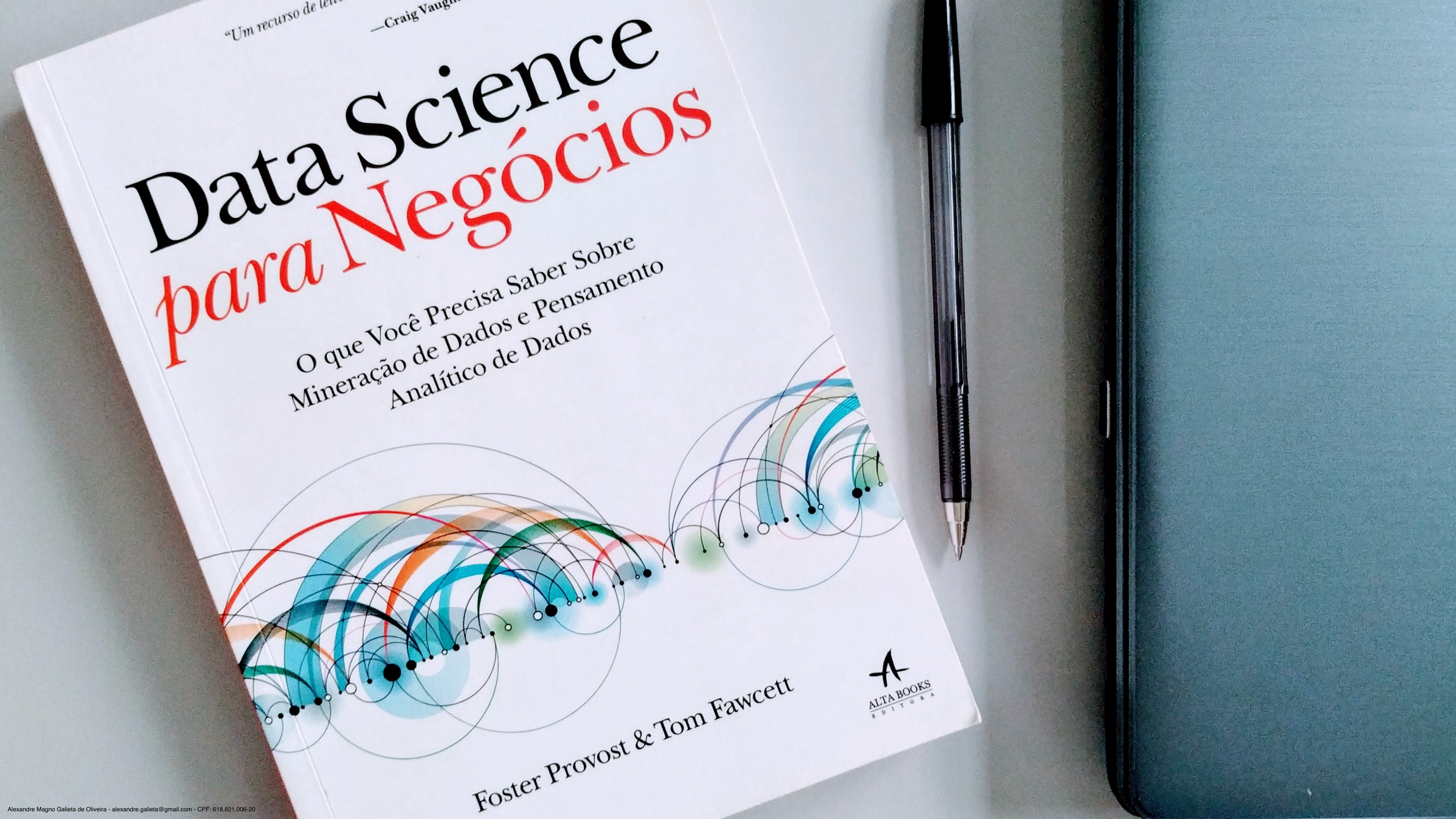
- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment
- Review

“É importante entender de Data Science mesmo que que você mesmo nunca pretenda aplicá-la. Um pensamento analítico de dados te permite avaliar propostas e projetos relacionados a dados”

Data Science para Negócios

- Cada problema é exclusivo.
- No entanto, há conjuntos de tarefas comuns que permeiam os problemas de negócio.

Alexandre Magno Galieta de Oliveira - alexandre.galieta@gmail.com - CPF: 618.821.006-20



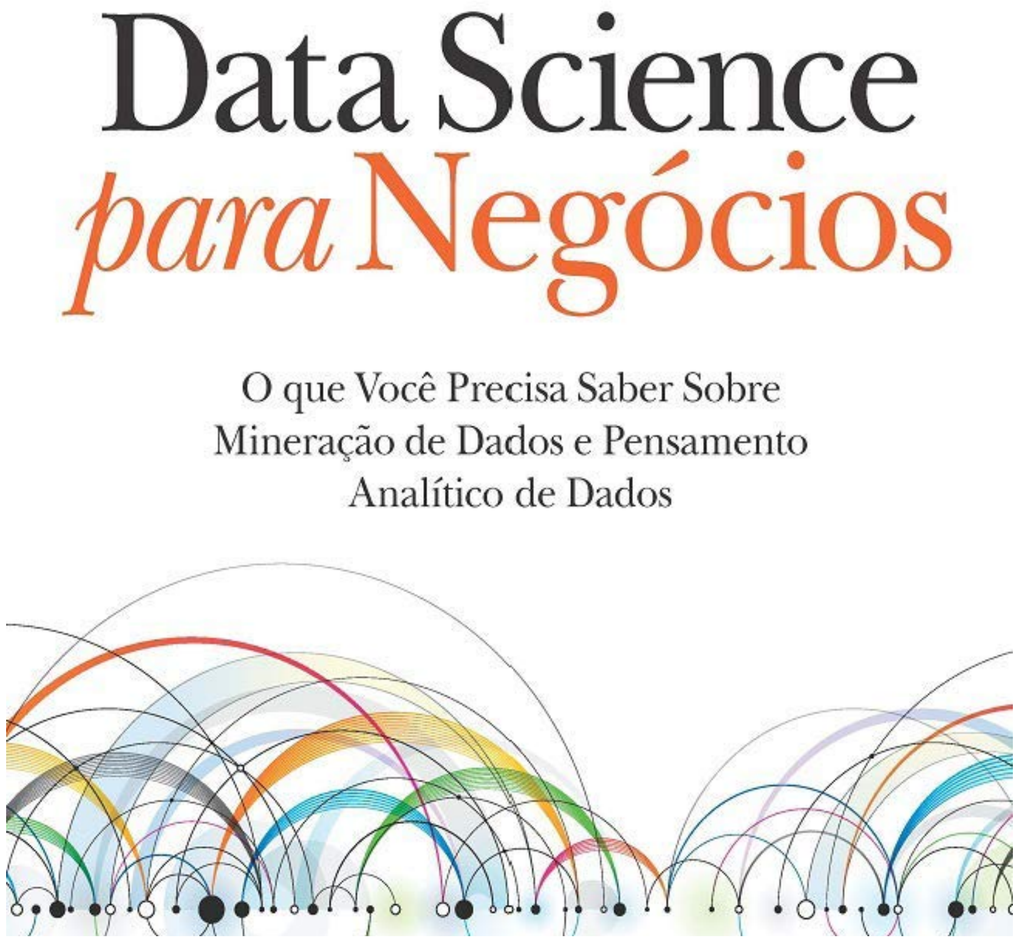
Alexandre Magno Galieta de Oliveira - alexandre.galieta@gmail.com - CPF: 618.821.006-20

Sobre o Livro

- Pessoas de negócios, que irão trabalhar com cientistas de dados, gerenciando projetos orientados a dados ou investem em negócios da área.

- Desenvolvedores que implementarão soluções de Data Science

- Aspirantes à Cientistas de Dados



Alexandre Magno Galieta de Oliveira - alexandre.galieta@gmail.com - CPF: 618.821.006-20

- Não é um livro de algoritmos.
- Há um pequeno conjunto de conceitos e princípios **sustentando** as técnicas de **extração de conhecimento à partir de dados**.
- Os autores optaram pela exposição desses princípios gerais, ao invés de algorítmos e linguagens específicas.

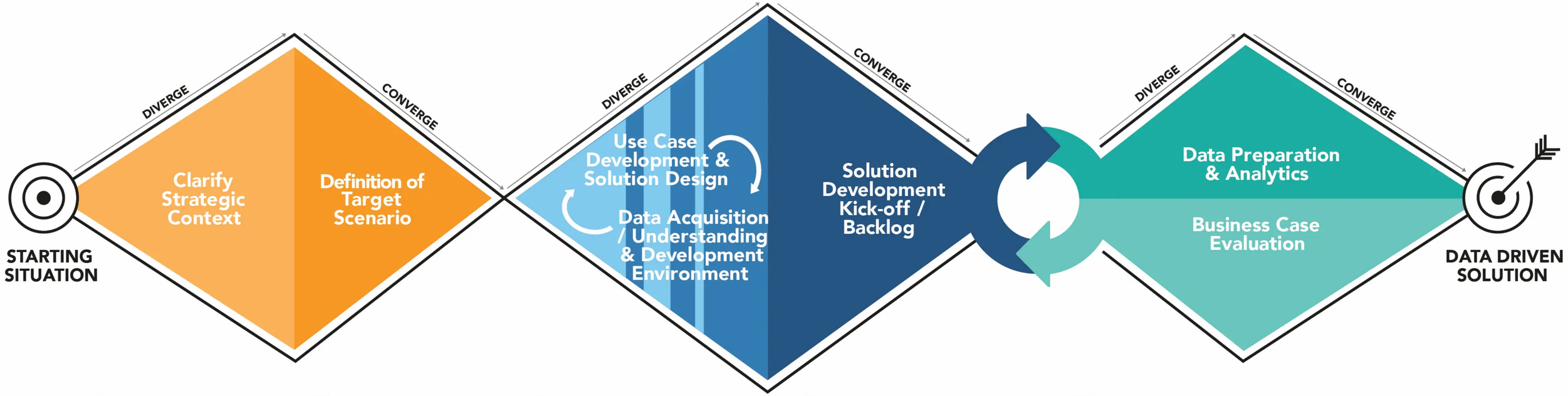
Alexandre Magno Galieta de Oliveira - alexandre.galieta@gmail.com - CPF: 618.821.006-20

Esse livro serve até para você saber o que fazer, quando ninguém souber o que te pedir.

Alexandre Magno Galieta de Oliveira - alexandre.galieta@gmail.com - CPF: 618.821.006-20

CHAPTER 1
INTRODUCTION: DATA-ANALYTIC THINKING

DATA THINKING



1 UNDERSTAND

- Competitor Analysis
- Best Practice Research
- Trend analysis
- Internal Data Capabilities Assessment

2 SYNTHESIS

- Scenario development
- Data Driven Business Target Picture
- Focus Areas definition/ Selection

3 EXPLORE

- Exploration and development of use cases
- Define data sources and access the Data
- Definition and availability of Development Environment
- Perform Data Assessment

4 DEFINE

- Solution Design Concept Inc. Solution Features
- PoC Scope
- PoC Plan & Backlog

5 BUILD

- Perform Data Cleaning
- Feature Engineering
- Model Building And Evaluation (Select ML algorithm/ AI framework and evaluate performance)
- User Centred Business Case

6 TRANSFER

DECIDE FOR IMPLEMENTATION

DATA DRIVEN SOLUTION TRIGGERS

DATA DRIVEN SOLUTION CONCEPT

DATA DRIVEN SOLUTION DEVELOPMENT

DATA DRIVEN PROOF OF CONCEPT

DETECON
Digital Engineering Center

Alexandre Magno Galieta de Oliveira - alexandre.galieta@gmail.com - CPF: 618.821.006-20

- A infraestrutura de negócios melhorou sua capacidade de coleta de dados por toda a empresa.

- Não estamos mais limitados ao âmbito do departamento de TI.

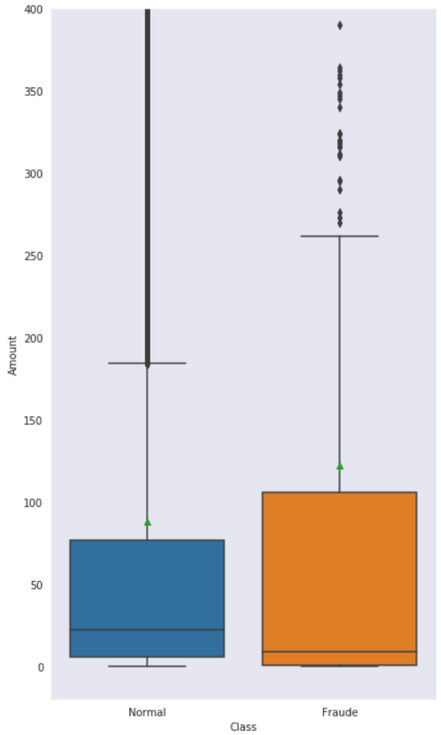
    - Operações, Marketing, Vendas, Manufatura, Gestão da Cadeia de Suprimentos, Customer Experience, fluxos e processos de trabalho.

- A abundância de dados levou a um aumento no interesse por métodos de **extração de informações e conhecimentos** úteis a partir do *raw data*.

Alexandre Magno Galieta de Oliveira - alexandre.galieta@gmail.com - CPF: 618.821.006-20

ANÁLISE EXPLORATÓRIA
ANÁLISE EXPLANATÓRIA

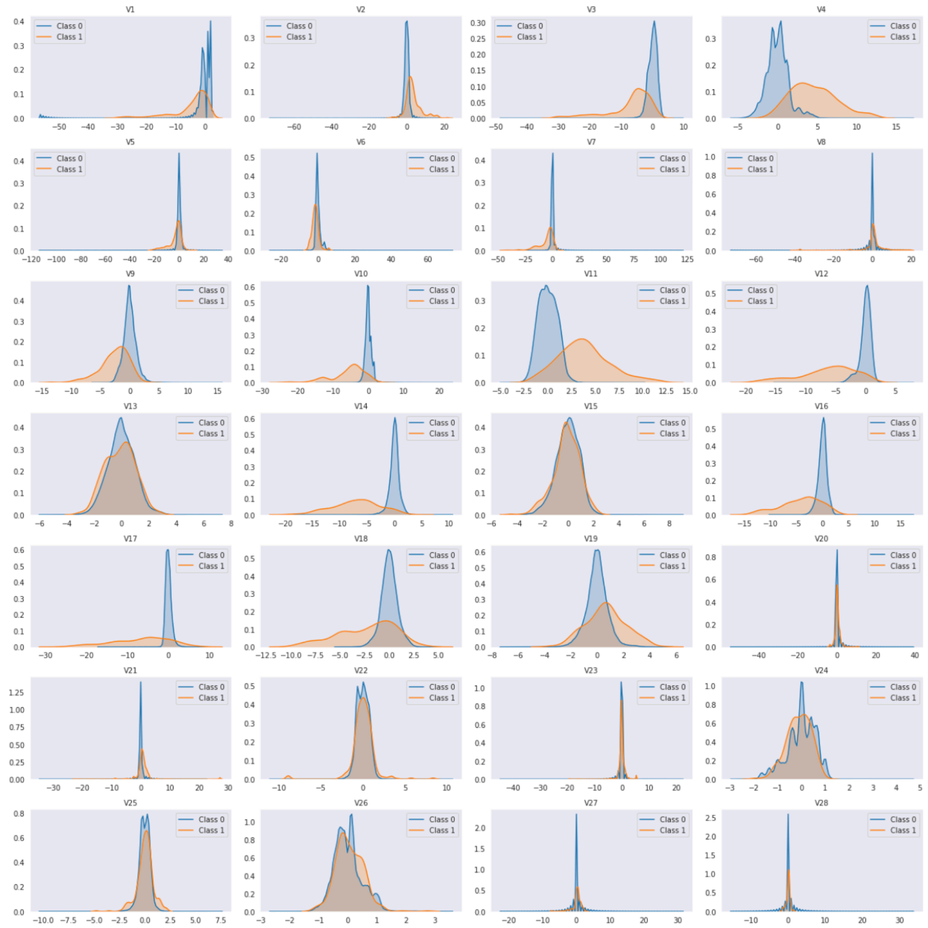
In [0]:
```
fig, ax = plt.subplots(figsize=(6,10), sharex=True)

sns.boxplot(df.Class, df.Amount, showmeans=True, ax=ax)
plt.ylim((-20, 400))
plt.xticks([0, 1], ['Normal', 'Fraude'])

plt.tight_layout()
```



As informações estatísticas para `df.Class == 1` mostram que a sua média exata está em 118.13 e a mediana em 9.21.

In [0]:
```
df[(df.Class == 1) & (df.Amount < 2000)]['Amount'].describe()
```

Out[0]:
```
count     491.000000
mean      118.130550
std       240.439028
min         0.000000
25%         1.000000
50%         9.210000
75%       105.350000
max      1809.680000
Name: Amount, dtype: float64
```



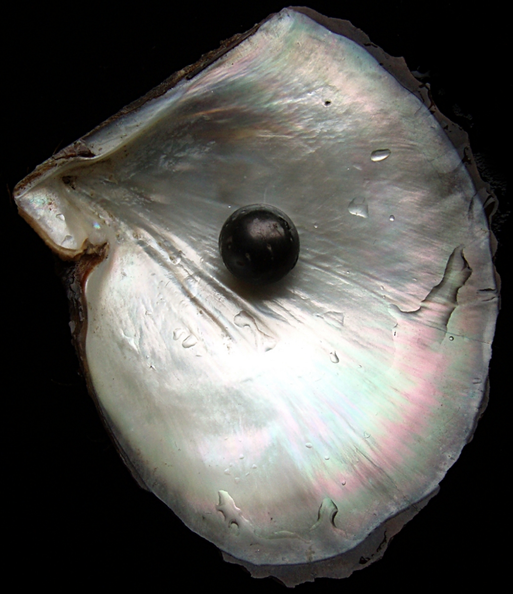
Resumidamente, algumas observações principais que puderam ser extraídas dessa etapa exploratória foram:

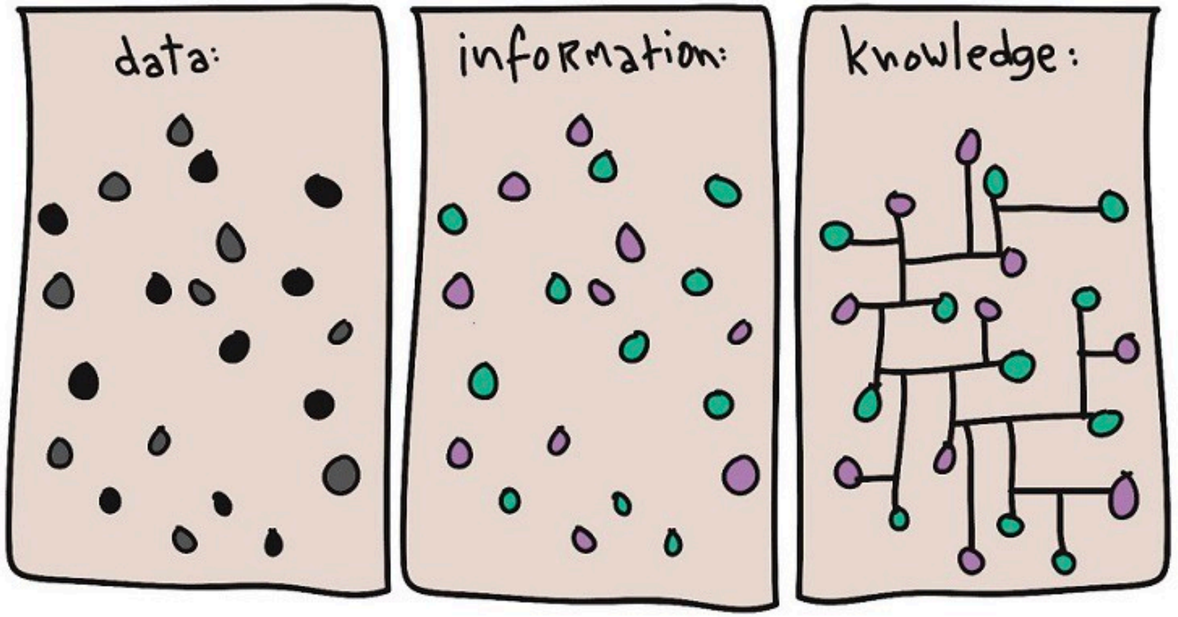
- O *dataset* está muito desbalanceado, uma vez que as transações fraudulentas representam apenas 0,17% das entradas totais.
- Não existem valores ausentes ou nulos no *dataframe*.
- Uma vez que a PCA é sensível à escala das variáveis, assume-se a premissa que as *features* originais foram padronizadas.
- As colunas `Time` e `Amount` não estão normalizadas.

Preparando os dados

Alexandre Magno Galieta de Oliveira - alexandre.galieta@gmail.com, CPF: 618.821.106-20

Alexandre Magno Galieta de Oliveira - alexandre.galieta@gmail.com - CPF: 618.821.006-20





Alexandre Magno Galieta de Oliveira - alexandre.galieta@gmail.com - CPF: 618.821.006-20

ANALISTAS

MODELADORES

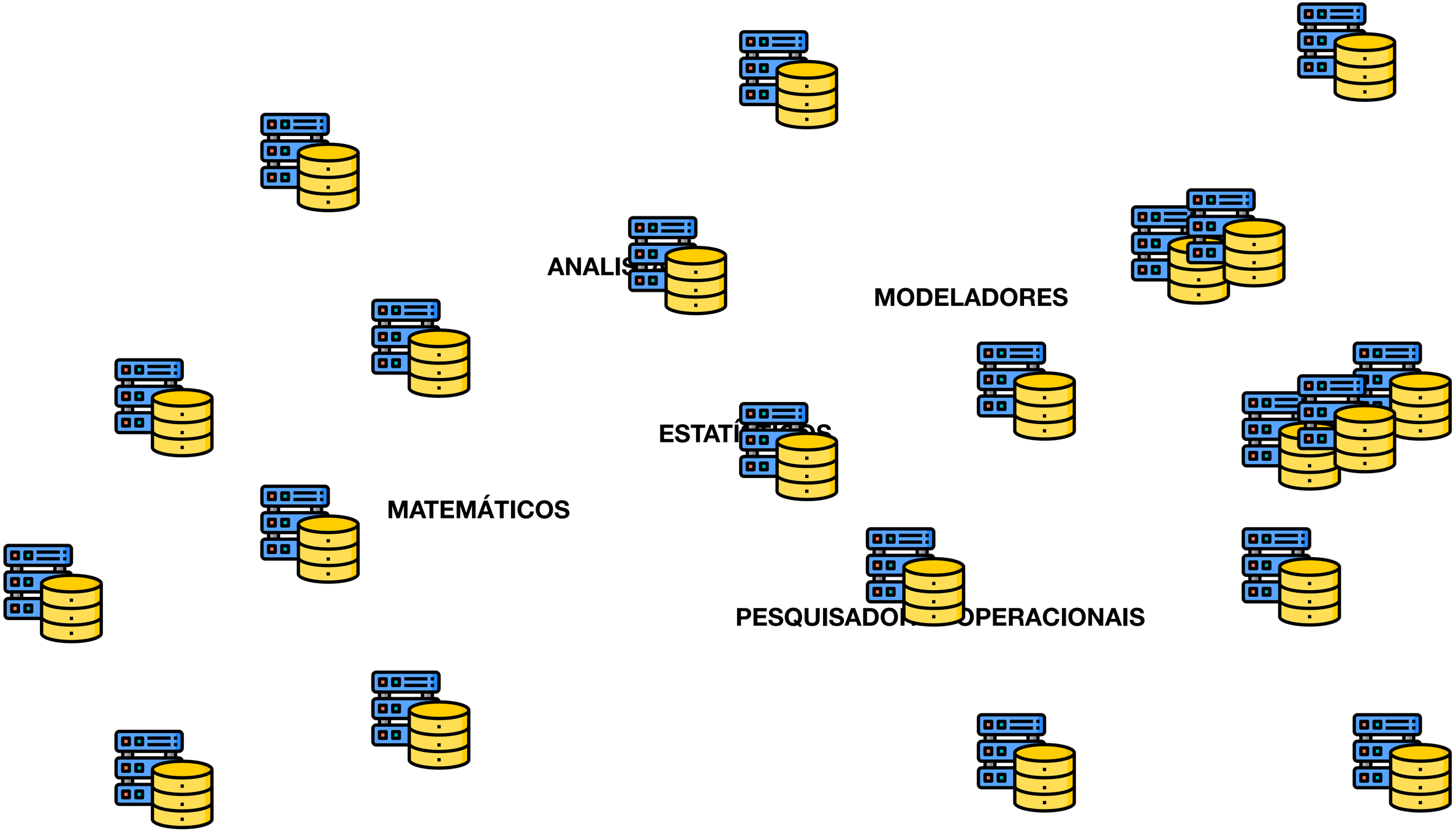
ESTATÍSTICOS

MATEMÁTICOS

PESQUISADORES OPERACIONAIS

Alexandre Magno Galieta de Oliveira - alexandre.galieta@gmail.com - CPF: 618.821.006-20

Alexandre Magno Galieta de Oliveira - alexandre.galieta@gmail.com - CPF: 618.821.006-20



Alexandre Magno Galieta de Oliveira - alexandre.galieta@gmail.com - CPF: 618.821.006-20

“… ajudar você a visualizar **problemas de negócios** a partir da **perspectiva de dados**, e a **entender os princípios da extração de conhecimento** útil a partir deles.”

Qual o principal objetivo deste livro?

Alexandre Magno Galieta de Oliveira - alexandre.galieta@gmail.com - CPF: 618.821.006-20

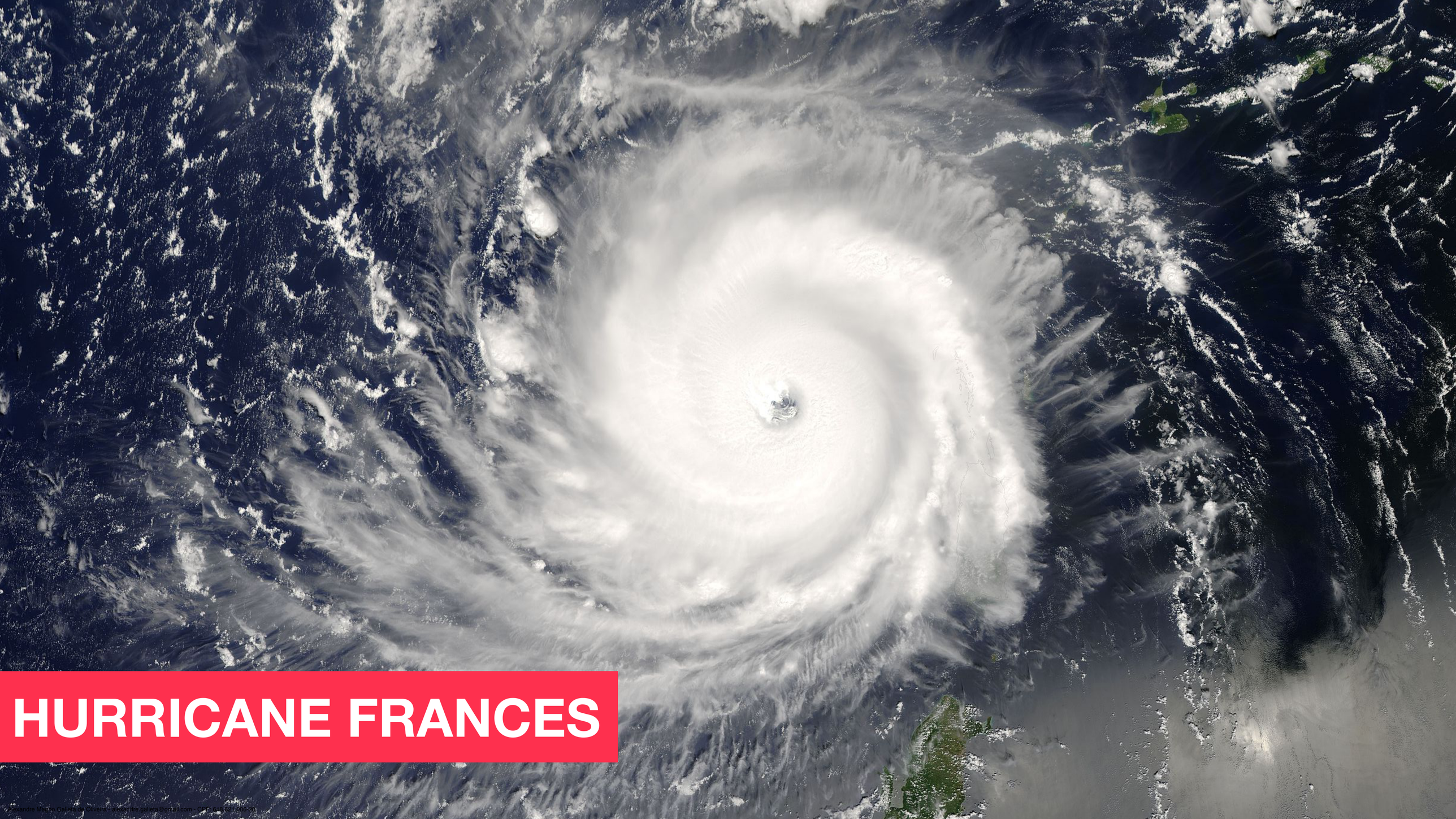
“…uma **perspectiva de dados** fornecerá estrutura e princípios, e isso lhe dará uma base para **analisar sistematicamente** tais problemas.”

Why should I fucking care?

Alexandre Magno Galieta de Oliveira - alexandre.galieta@gmail.com - CPF: 618.821.006-20

It is important to understand data science even if you never intend to apply it yourself. Data-analytic thinking enables you to evaluate proposals for data mining projects. For example, if an employee, a consultant, or a potential investment target proposes to improve a particular business application by extracting knowledge from data, you should be able to assess the proposal systematically and decide whether it is sound or flawed. This does not mean that you will be able to tell whether it will actually succeed—for data mining projects, that often requires trying—but you should be able to spot obvious flaws, unrealistic assumptions, and missing pieces.

Alexandre Magno Galieta de Oliveira - alexandre.galieta@gmail.com - CPF: 618.821.006-20


HURRICANE FRANCES

Alexandre Magno Galieta de Oliveira - alexandre.galieta@gmail.com - CPF: 618.821.006-20

Alexandre Magno Galieta de Oliveira - alexandre.galieta@gmail.com - CPF: 618.821.006-20

Alexandre Magno Galieta de Oliveira - alexandregalieta@gmail.com - CPF 270.021.108-20

SOUTHERN LIVING • NEWS

You'll Never Guess What Surprising Product Walmart Stocks up on Before Big Storms

It actually makes a lot of sense.

By Meghan Overdeep | August 28, 2019

Save
FB
Tweet
More

CREDIT: TENNESSEEPHOTOGRAPHER/GETTY IMAGES

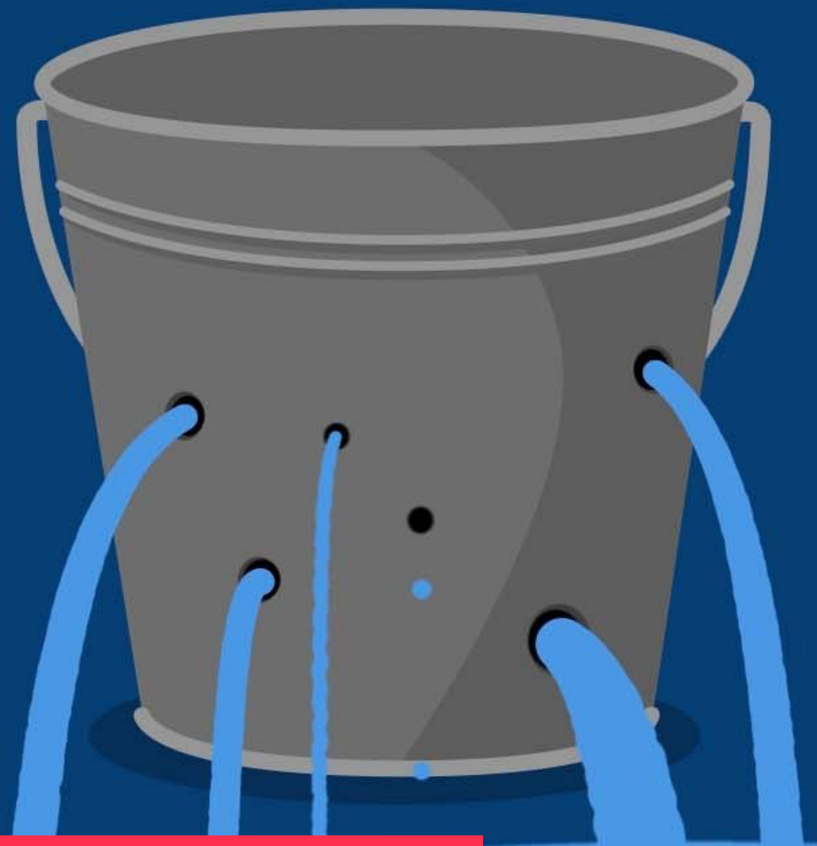
What do you stock up on before a big storm? Probably water, milk, bread, and eggs, right? The essentials.

Grocery stores know what to expect when weather forecasts turn dicey and try to have plenty of the aforementioned staples on hand. But what if we told you that Walmart locations also stock up on a certain fruit-flavored pastry?

Apparently, strawberry Pop-Tarts are a huge seller before and after big storms.

In a 2004 interview with *The New York Times*, Walmart officials revealed that sales of strawberry Pop-Tarts increase by as much as seven times ahead of

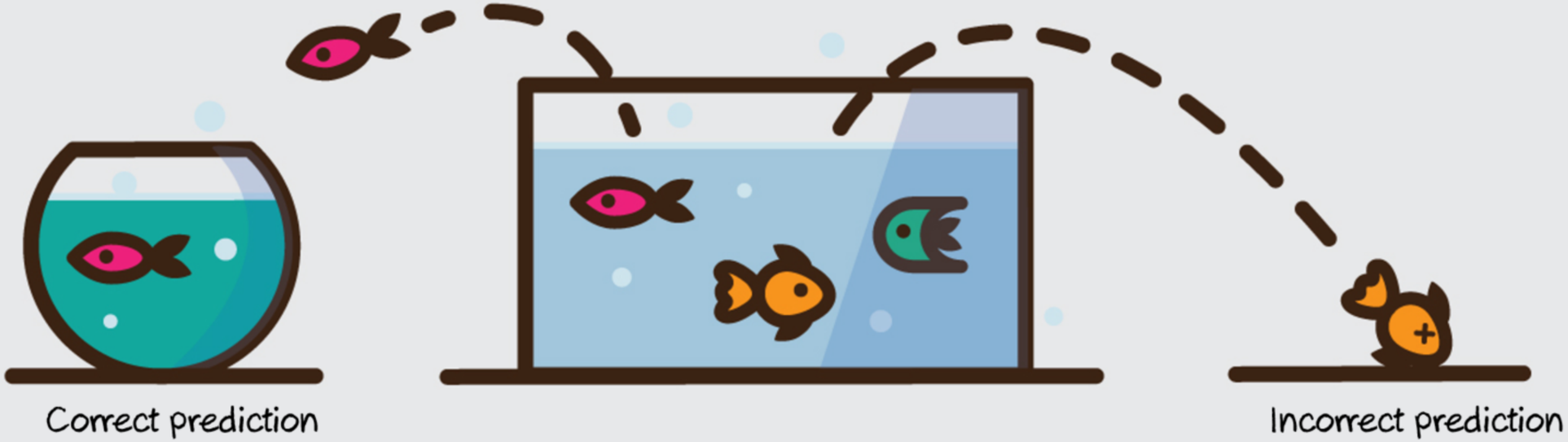
Alexandre Magno Galieta de Oliveira - alexandre.galieta@gmail.com - CPF: 618.821.006-20



ROTATIVIDADE DE CLIENTE

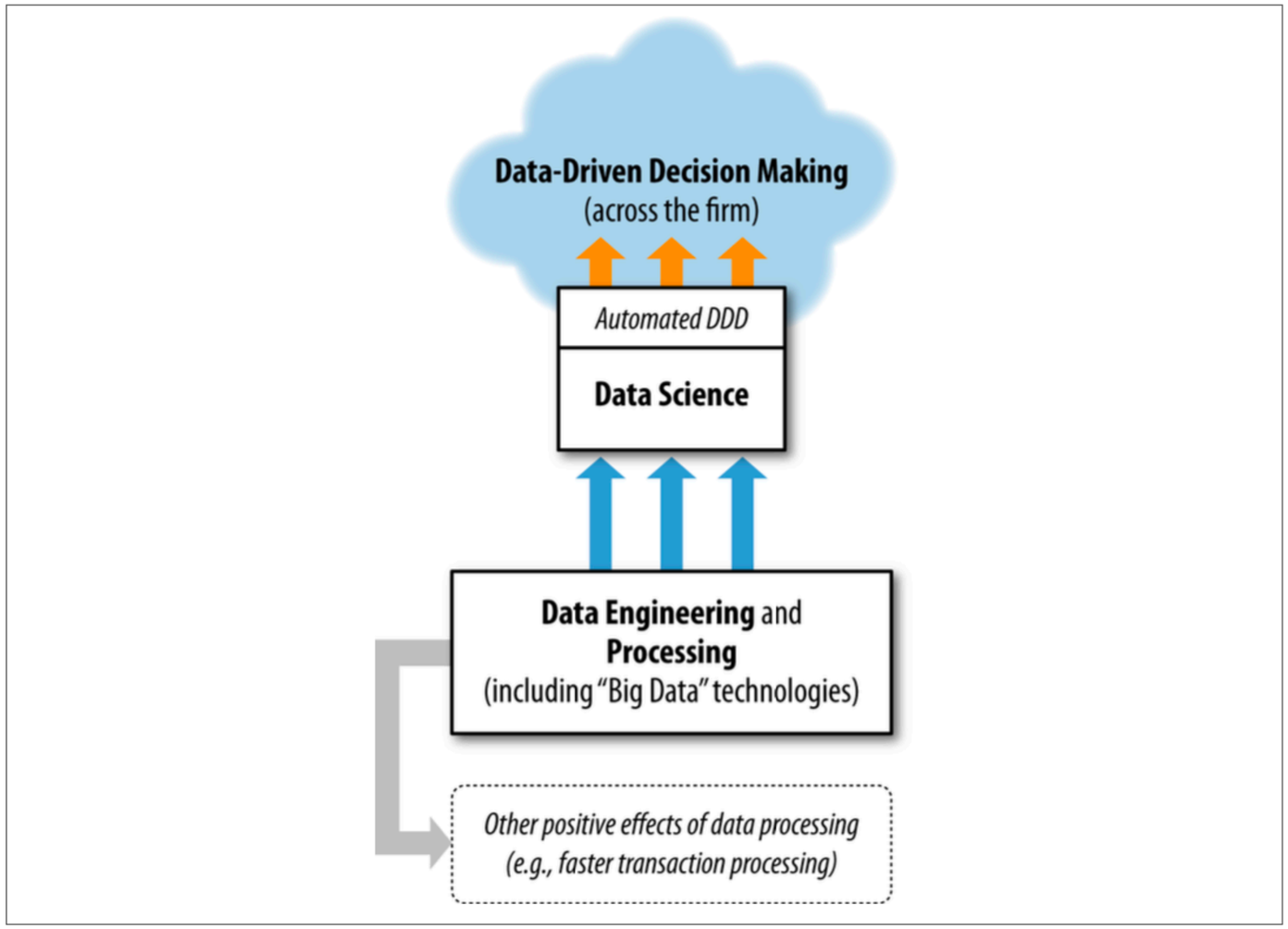
Alexandre Magno Galieta de Oliveira - alexandre.galieta@gmail.com - CPF: 618.821.006-20

Customer Churn Prediction Accuracy



Alexandre Magno Galieta de Oliveira - alexandre.galieta@gmail.com - CPF: 618.821.006-20

DATA-DRIVEN
DECISION



Alexandre Magno Galieta de Oliveira - alexandre.galieta@gmail.com - CPF: 618.821.006-20

CASO DA TARGET

COMPETIDOR DO WALLMART

Alexandre Magno Galieta de Oliveira - alexandre.galieta@gmail.com - CPF: 618.821.006-20

Feb 16, 2012, 11:02am EST

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

Kashmir Hill Former Staff
Tech
Welcome to The Not-So Private Parts where technology & privacy collide

This article is more than 8 years old.

Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy. Target , for example, has figured out how to data-mine its way into your womb, to figure out whether you have a baby on the way long before you need to start buying diapers.

Target has got you in its aim

Charles Duhigg outlines in the New York Times how Target tries to hook parents-to-be at that crucial moment before they turn into rampant -- and loyal -- buyers of all things pastel, plastic, and miniature. He talked to Target statistician Andrew Pole -- before Target freaked out and cut off all communications -- about the clues to a customer's impending bundle of joy. Target assigns every customer a Guest ID number, tied to their credit card, name, or email address that becomes a bucket that stores a history of everything they've bought and any demographic information Target has collected from them or bought from other sources. Using that, Pole looked at historical buying data for all the ladies who had signed up for Target baby registries in the past. From the NYT:

> [Pole] ran test after test, analyzing the data, and before long some useful patterns emerged. Lotions, for example. Lots of people buy lotion, but one of Pole's colleagues noticed that

LINK PARA ARTIGO

Alexandre Magno Galieta de Oliveira - alexandre.galieta@gmail.com - CPF: 618.821.006-20

PENSAMENTO
ANALÍTICO DE DADOS