

Tera

AULA 01

Módulo 01: Fundamentos de DS & ML

- » Introdução e Fundamentos de Data Science

T

instrutora:

RAYSSA KÜLLIAN

hoje:

 **Líder da Prática de
IA @LATAM**
Amazon Web Services

 **Pesquisadora**
LIAMF, IME/USP

 **Professora**

 **Palestrante**

antes:

IBM (Software Lab e Watson), BASF,
Ericsson, boo-box, Genesys, eGenius,
Semantix, boolabs (B2W)

formação:

MSc em Inteligência Artificial (IME/USP)
BSc em Sistemas de Informação (FIAP)

contato@rayssak.com.br

@rayssak

T

agenda

◇ definição

buzzword... Data Science?

◇ business

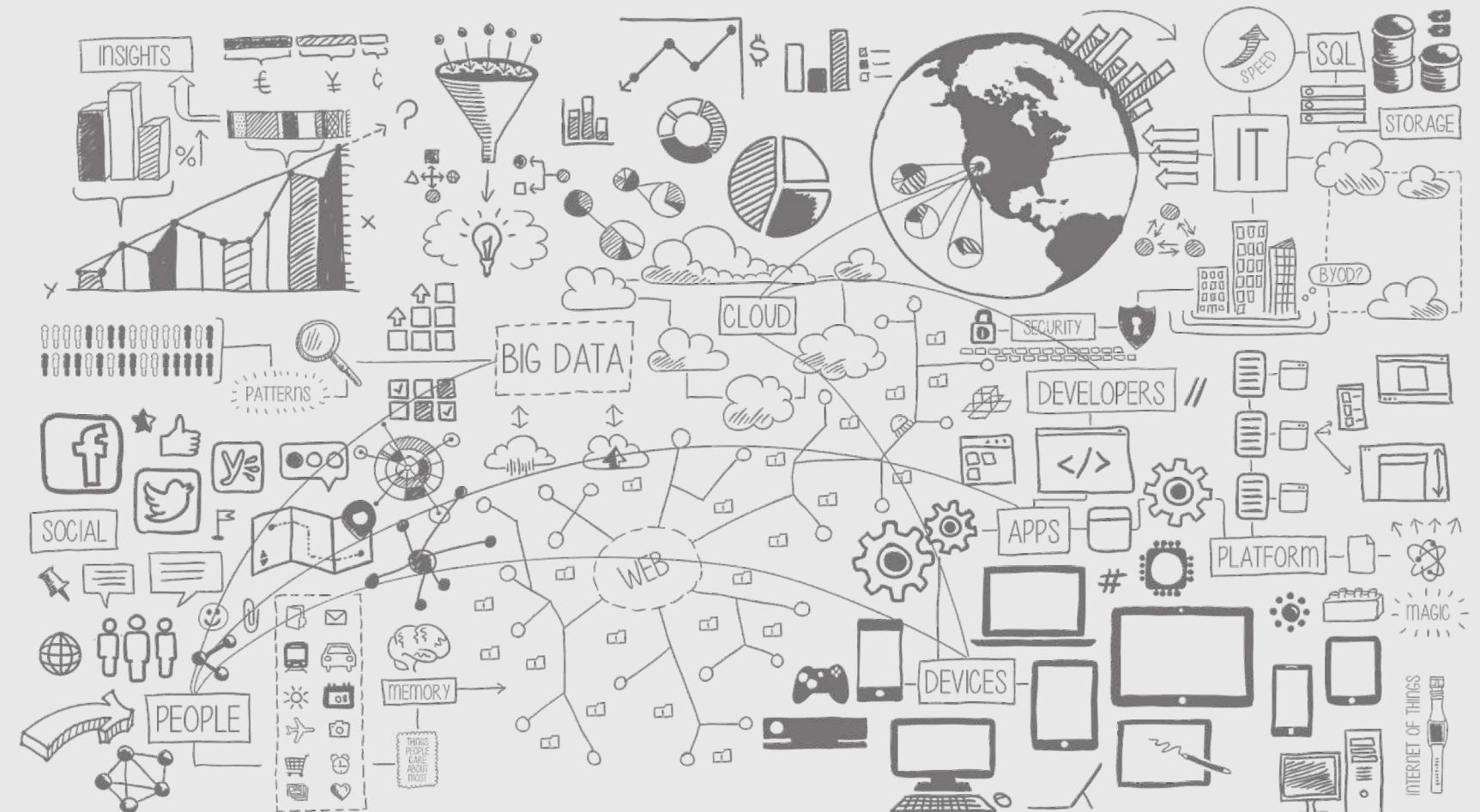
dados, mercado, previsões

◇ metodologia

ciclo de vida, atividades, papéis

◇ tecnologias

players, ferramentas,
landscape, diferenças



T

agenda

◇ definição

buzzword... Data Science?

◇ business

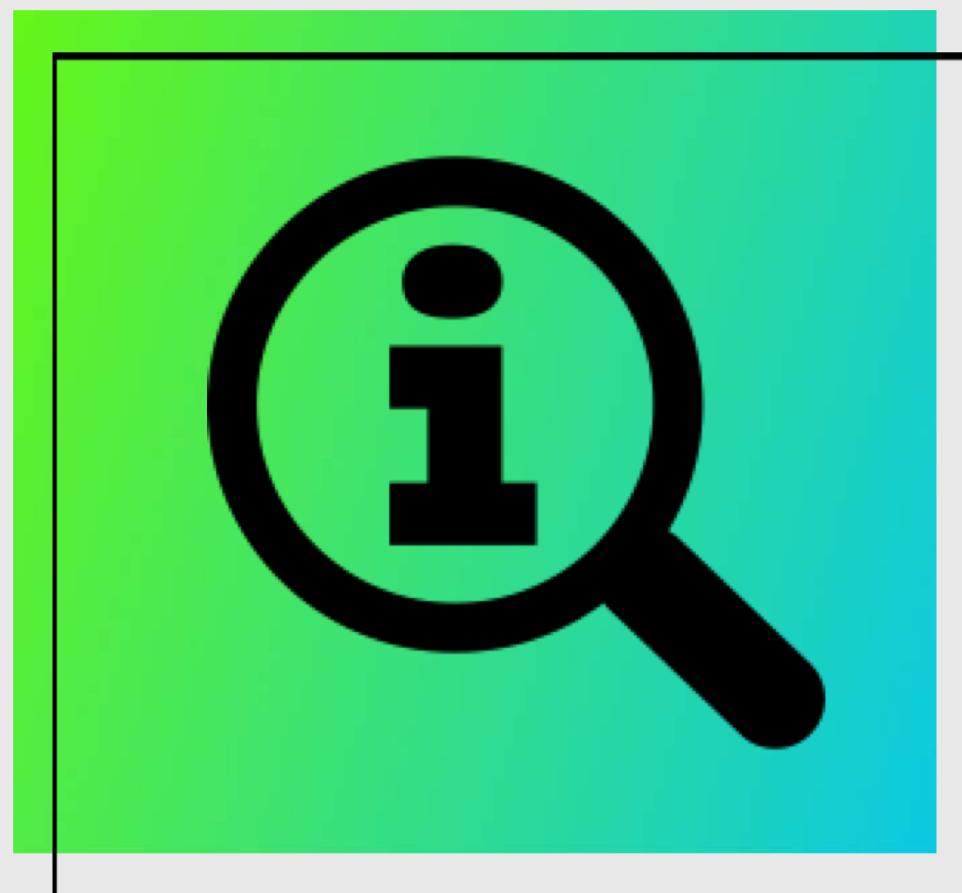
dados, mercado, previsões

◇ metodologia

ciclo de vida, atividades, papéis

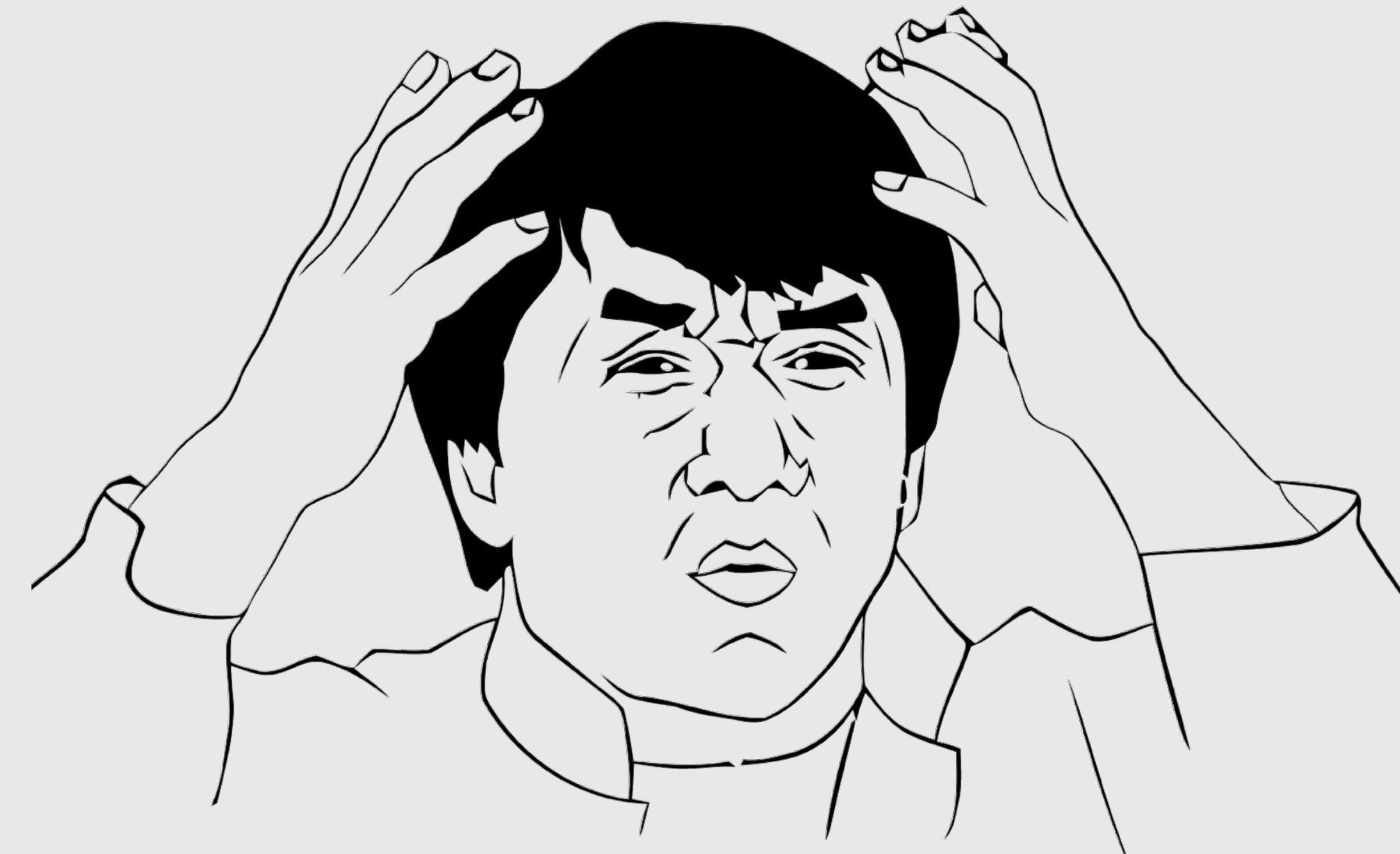
◇ tecnologias

players, ferramentas,
landscape, diferenças

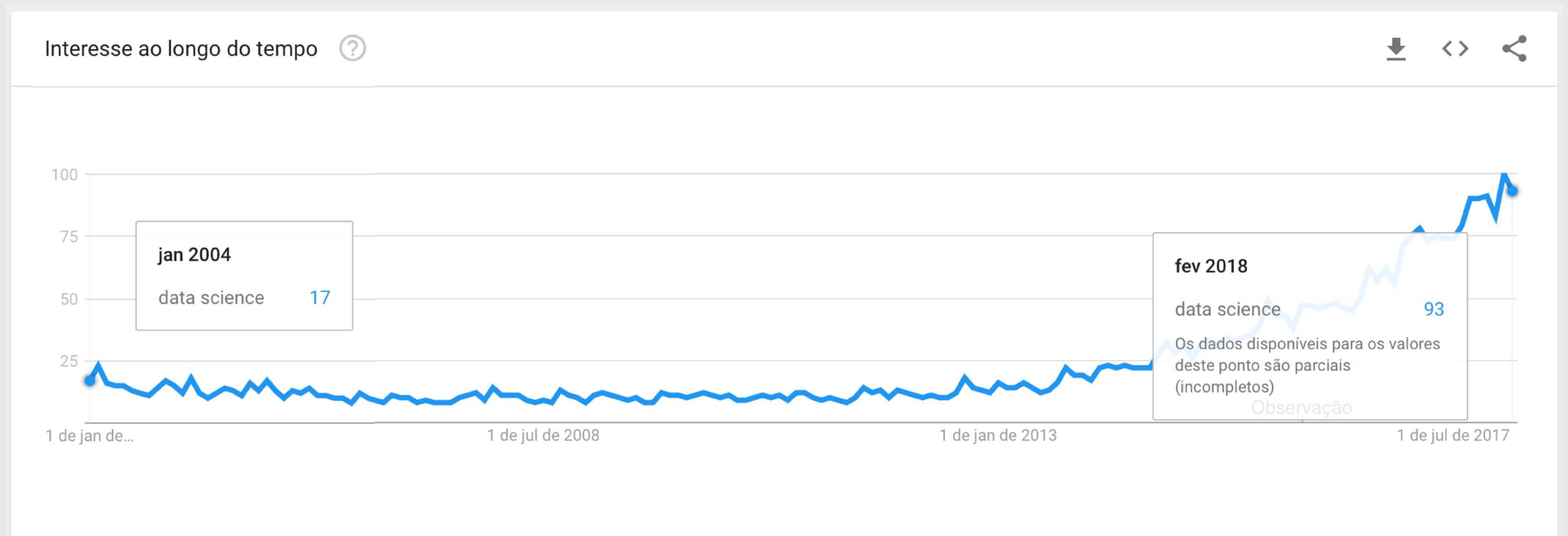


T o que é?

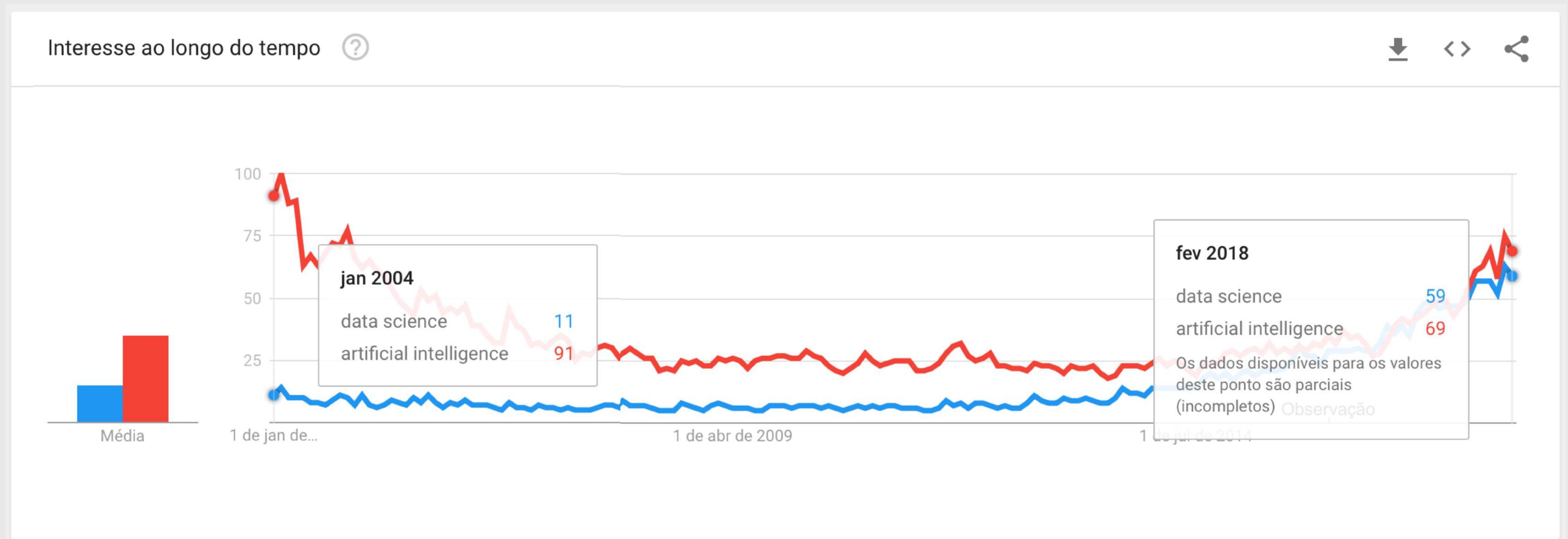
BI, Big Data, Analytics, Data Science, AI,
Machine Learning, Deep Learning,
#buzzzzZzzzwords?!



T data science

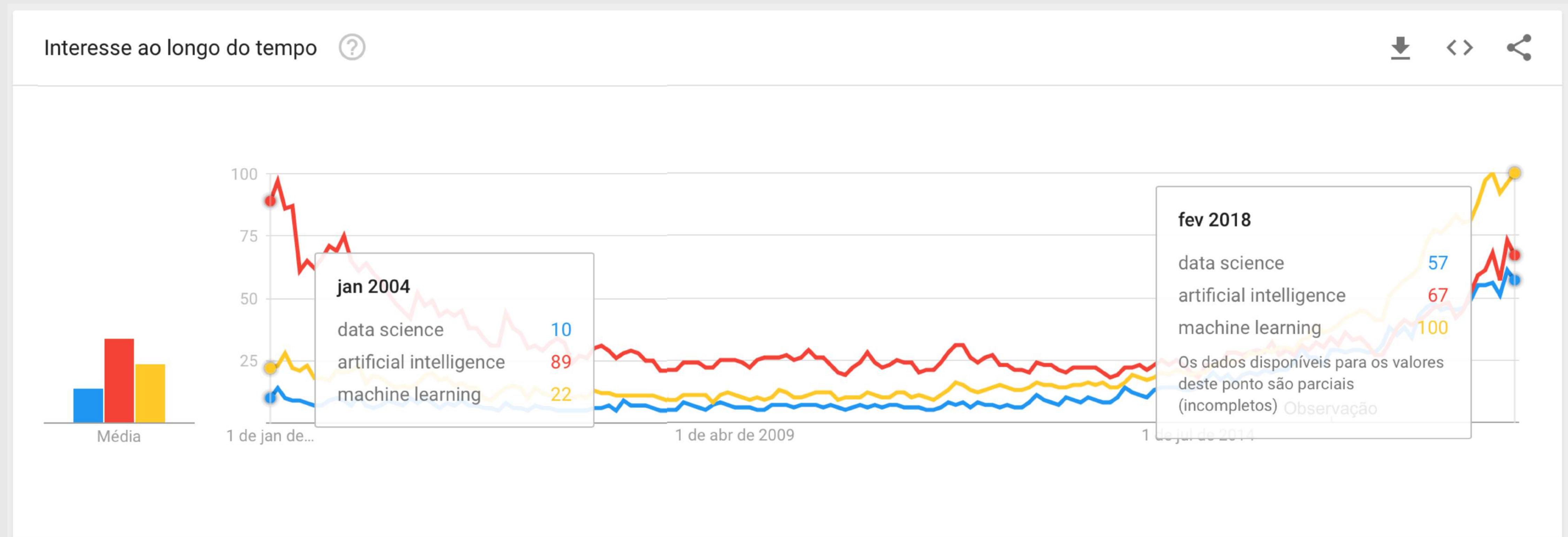


T data science vs artificial intelligence



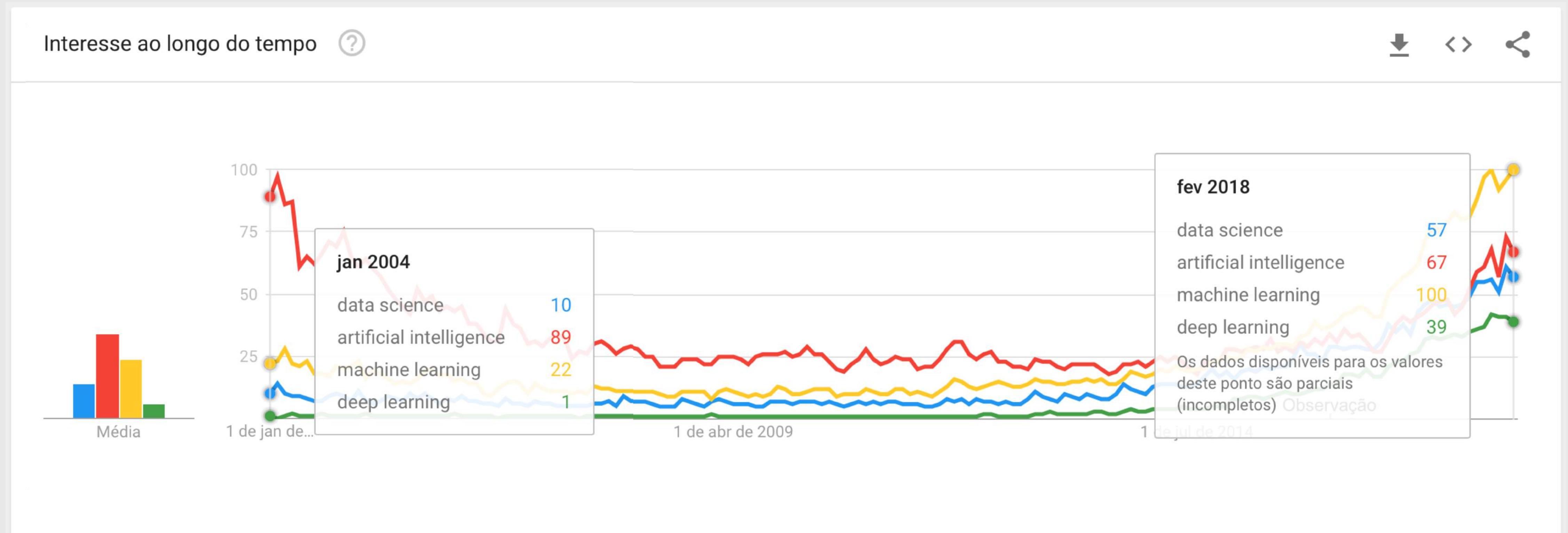
T

data science vs artificial intelligence vs machine learning



T

data science vs artificial intelligence vs machine learning vs deep learning



T o que não é?



T o que não é?

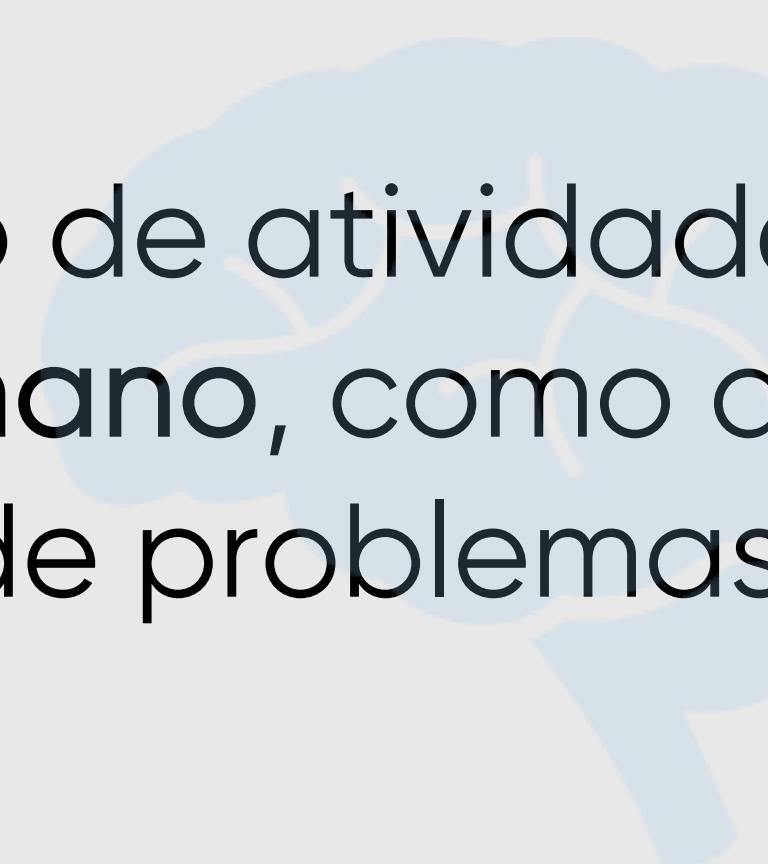


T o que não é?



T o que é?

“



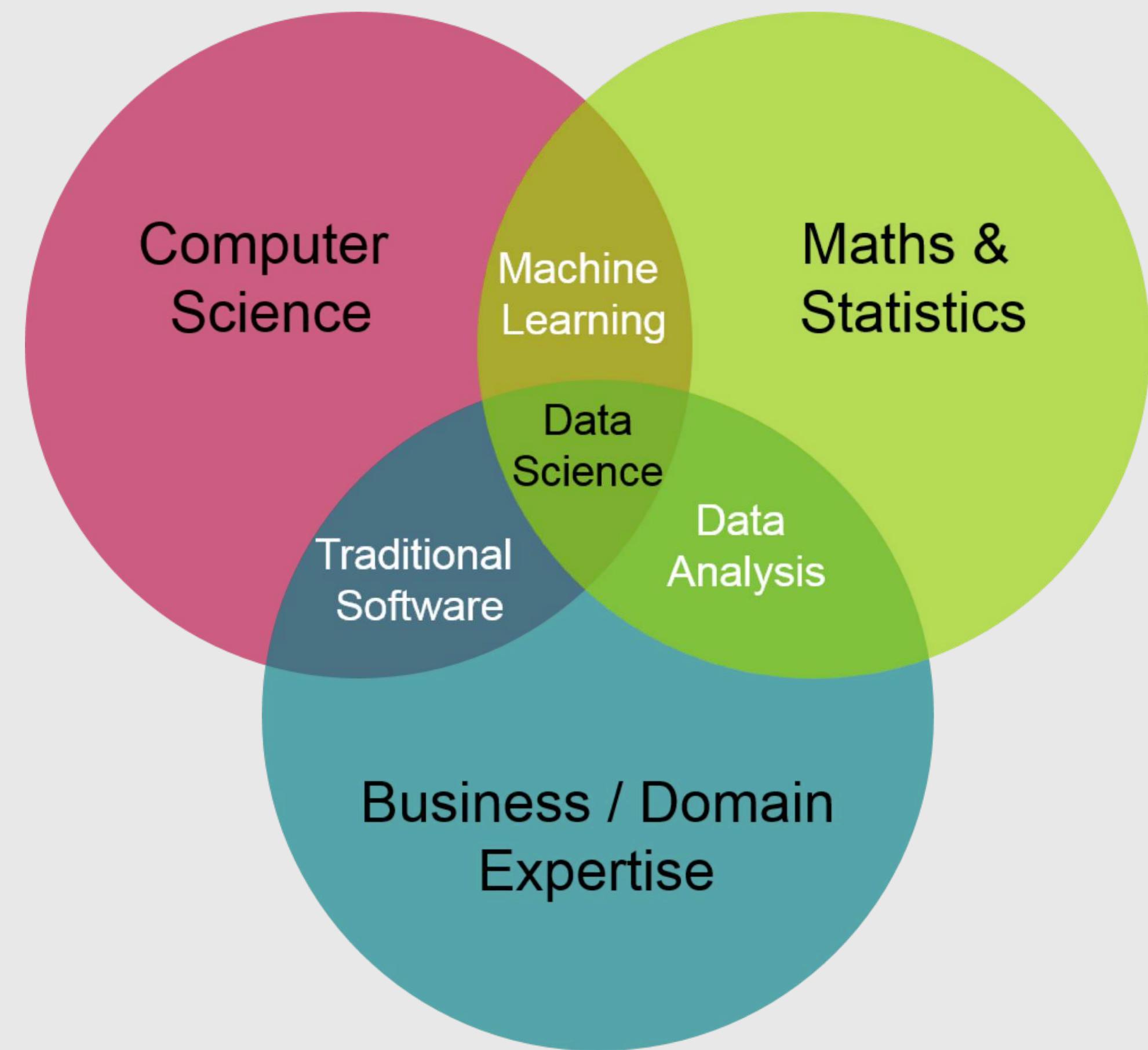
Automatização de atividades que associamos ao pensamento humano, como a tomada de decisões, a resolução de problemas, o aprendizado...

- Bellman, 1978

- ❖ gestação ~1943 à 1955
“Computing Machinery and Intelligence”
(McCulloch, Pitts, Turing)
- ❖ nascimento 1956
McCarthy, Minsky, Shannon e Rochester
- ❖ indústria 1956
- ❖ winter has come 1974



T data science



T

agenda

◇ definição

buzzword... Data Science?

◇ business

dados, mercado, previsões

◇ metodologia

ciclo de vida, atividades, papéis

◇ tecnologias

players, ferramentas,
landscape, diferenças

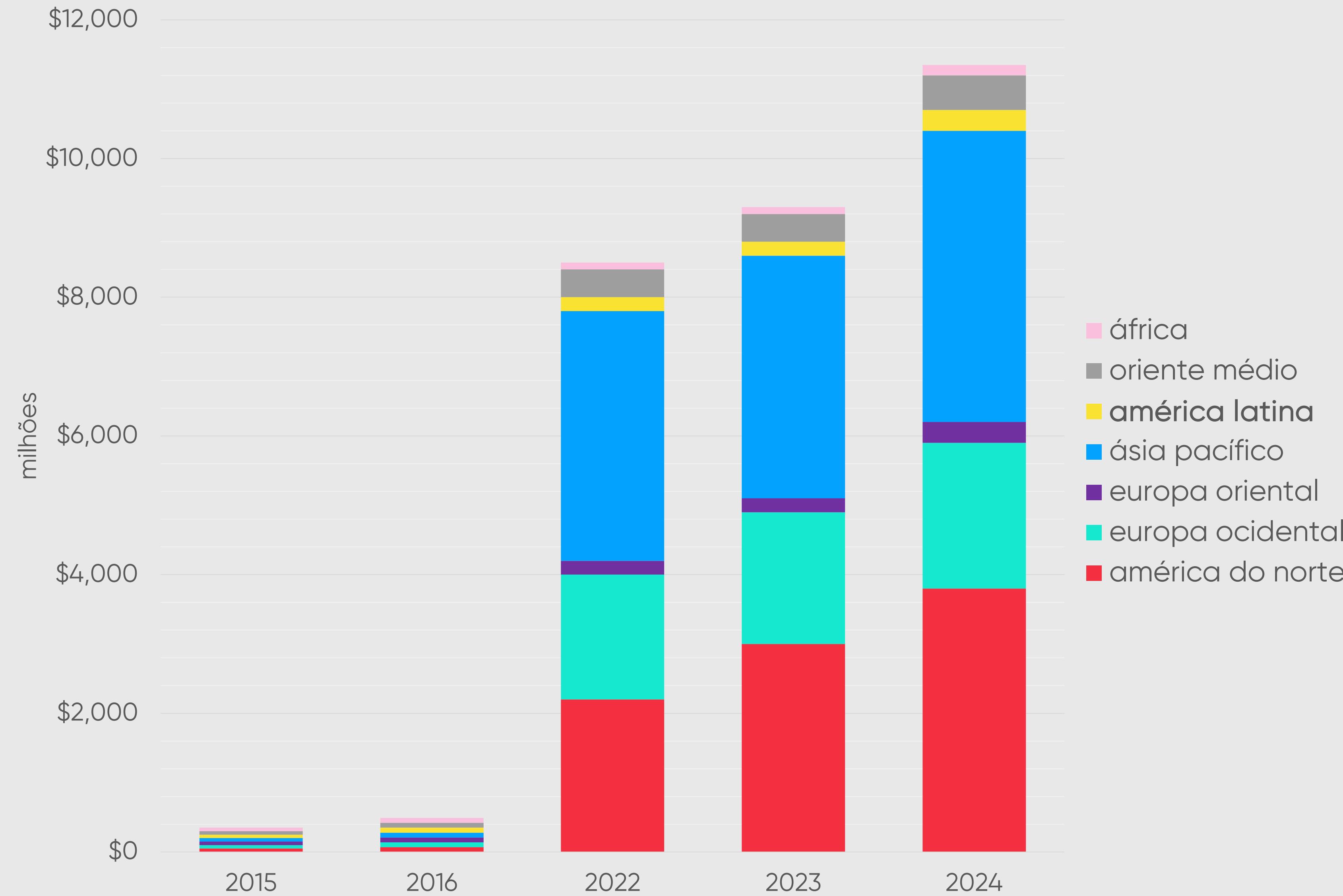


T

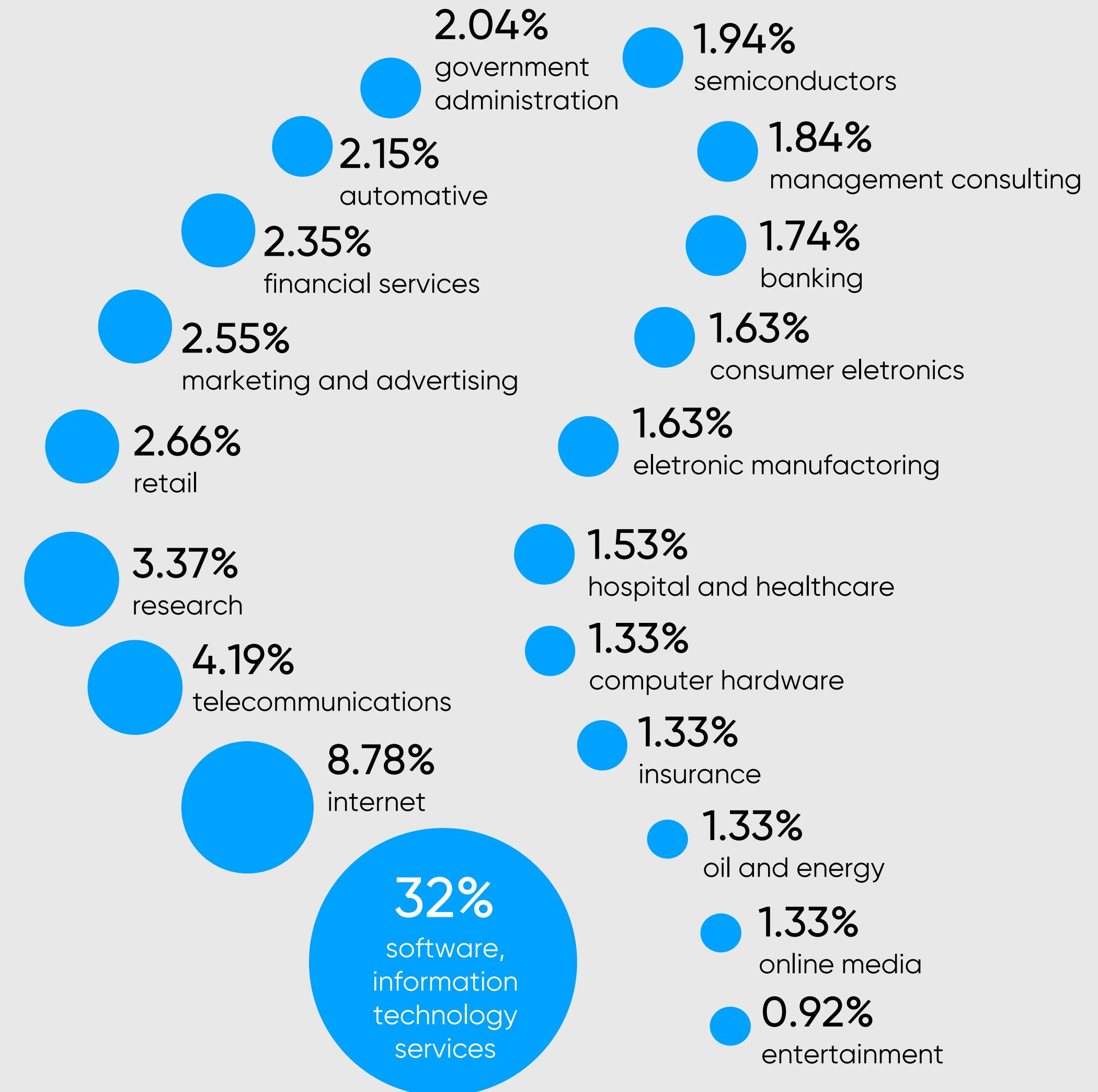
visão

- ❖ 61% das empresas
iniciativa mais importante de 2018
- ❖ **dobra** em 2018, **dobra novamente** em 2020
número de pilotos e experimentações

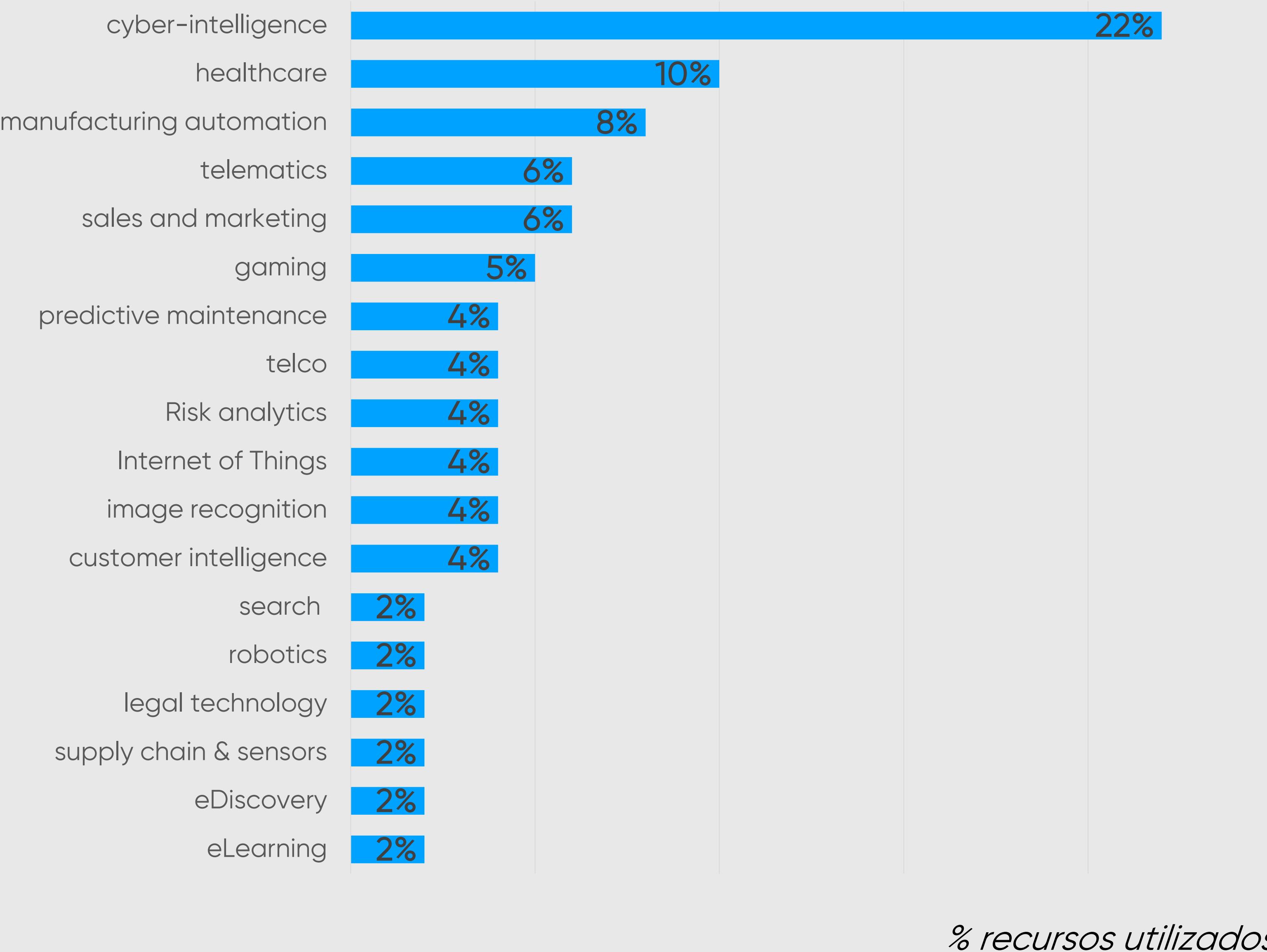
T crescimento



I indústria



T casos de uso



T maturidade

direcionamento estratégico
para o negócio
(Nível 3)

87

construindo aplicações
(Nível 2)

494

experimentações
(Nível 1)

967

*qtde de empresas

T empresas que mais investem

 Google

 Facebook

 Rocket Fuel

 IBM

 Amazon

 Yahoo

 Intel

 Microsoft

 Deloitte

 MITRE

T

agenda

❖ definição

buzzword... Data Science?

❖ business

dados, mercado, previsões

❖ metodologia

ciclo de vida, atividades, papéis

❖ tecnologias

players, ferramentas,
landscape, diferenças

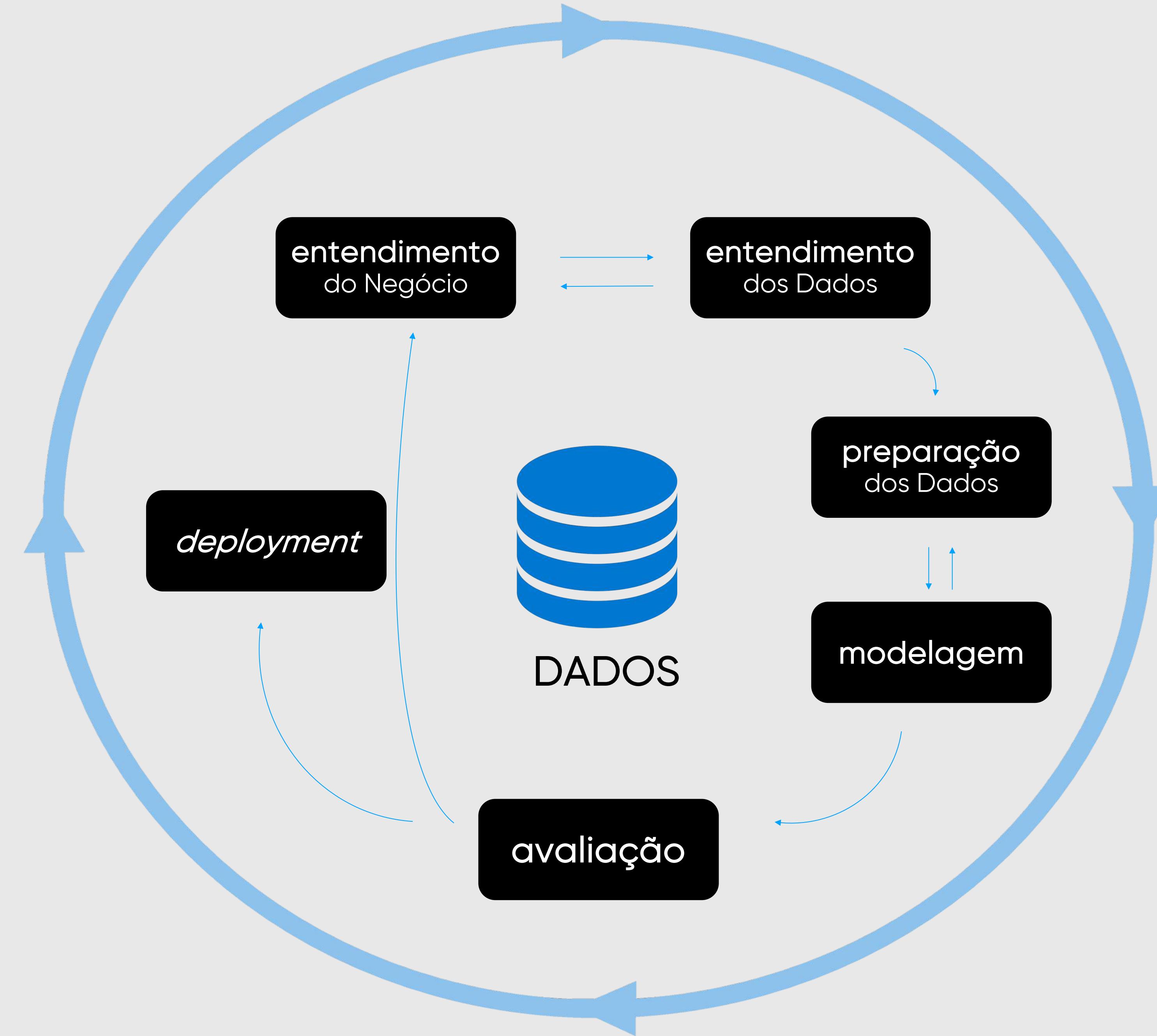


CRoss Industry Standard Process for Data Mining

CRISP-DM

T

crisp-dm



T atividades

1

explorar e conhecer cenários de negócios

2

definir problema (hipótese)

3

setar expectativas

T atividades

4

pesquisar estado-da-arte

análise exploratória de dados
(EDA)

5

6

limpeza e pré-processamento de dados

T atividades

7

engenharia de *features*

8

modelagem

9

avaliação de resultados

T atividades

10

fine-tuning ou boosting

11

validação com área de negócio

12

produtização

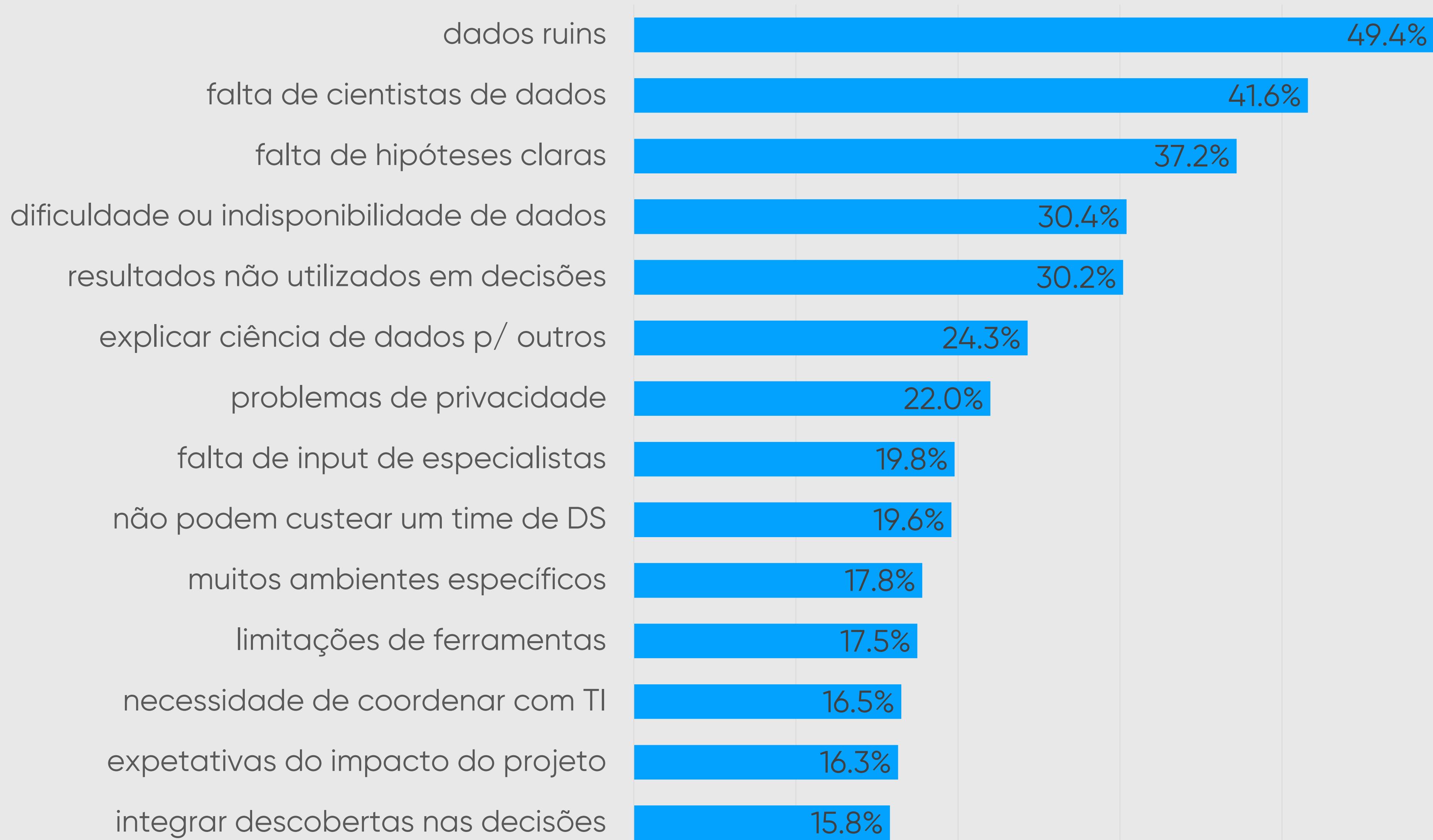
T atividades

paper acadêmico boa prática de mercado

- transferência de **conhecimento**
- **credibilidade** no universo acadêmico
- atração de **talentos**

❖ muitos(!) *blockers*
em experimentações e novos projetos

- disponibilidade dos dados
- maturidade em *analytics*
- recursos e time



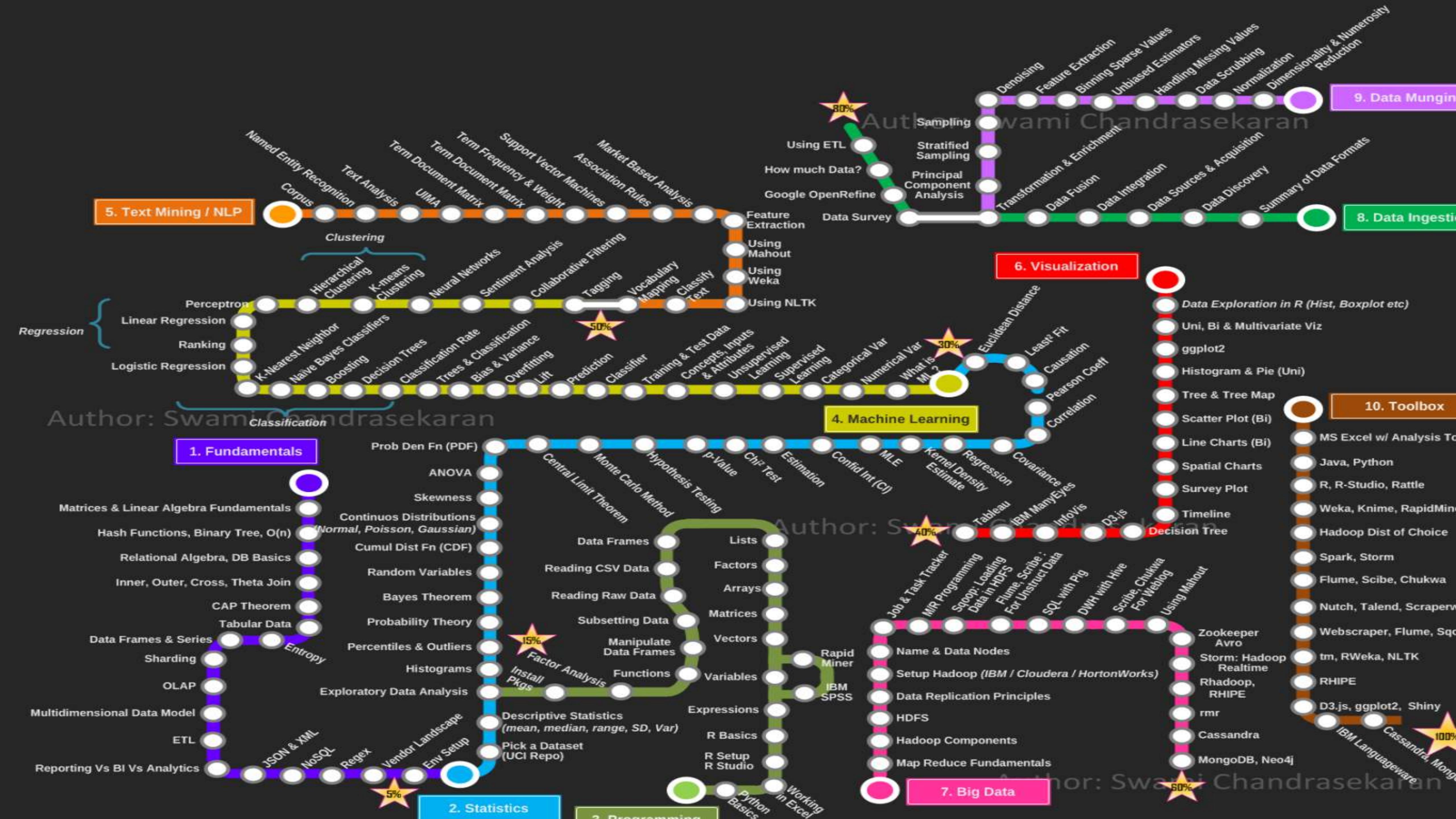
T papéis

cientista de dados?



“ Sexiest job of
the 21st century

- Harvard Business Review

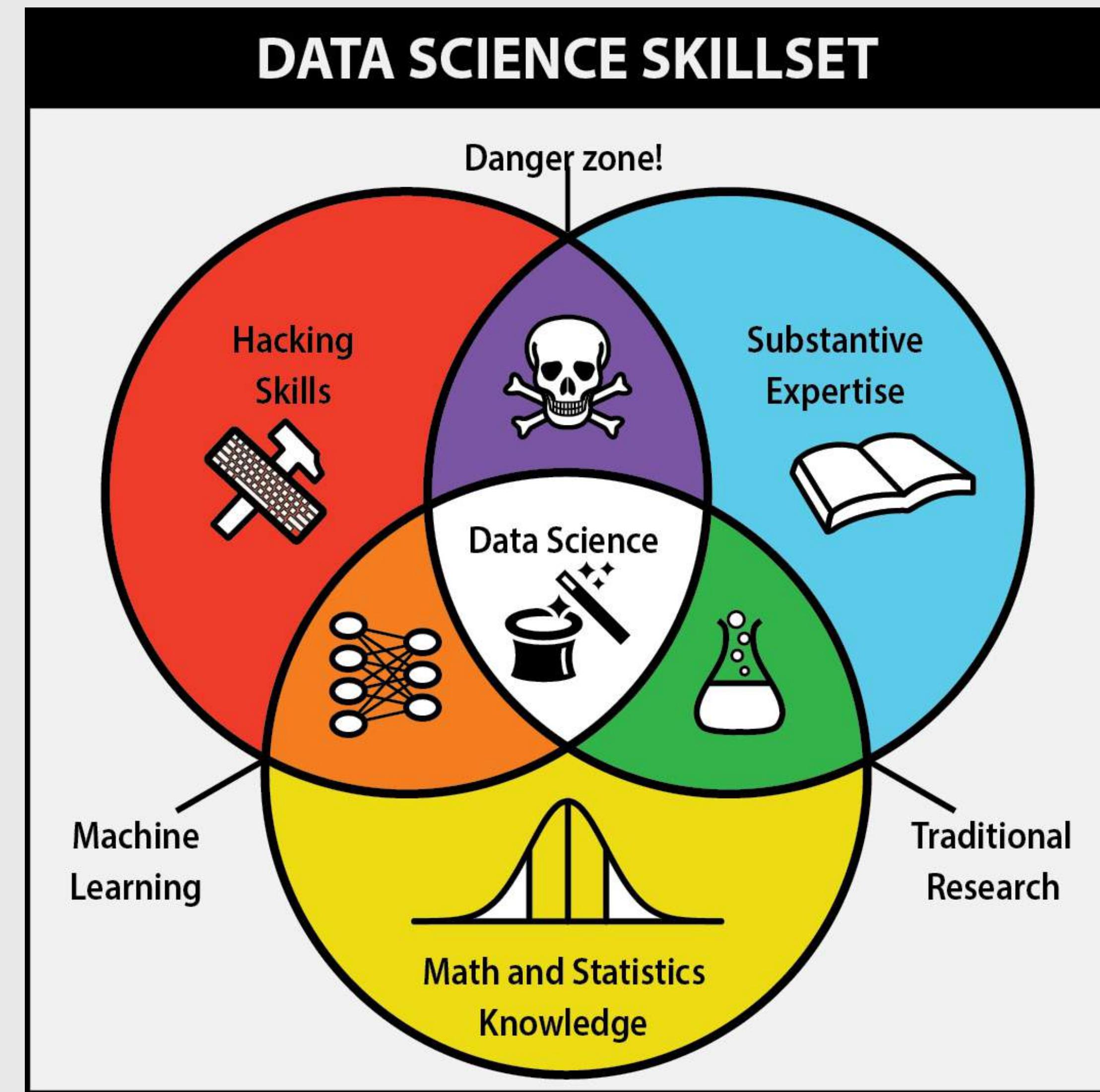


T cientista de dados



VOCÊ

T skills



 **hacking**

aquisição, limpeza, transformação de dados

 **matemática e estatística**

entender comportamento dos dados e escolher
apropriadamente os algoritmos

 **conhecimento de negócios**

motivar e gerar hipóteses e interpretar resultados

➤ pesquisa **tradicional**
método científico

inteligência artificial
➤ conhecimento matemático + computacional +
dev

danger zone
☒ hacking skills + IA sem matemática e metodologia
científica = **análises incorretas!**

acabou?



T

missing skill

comunicação

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.



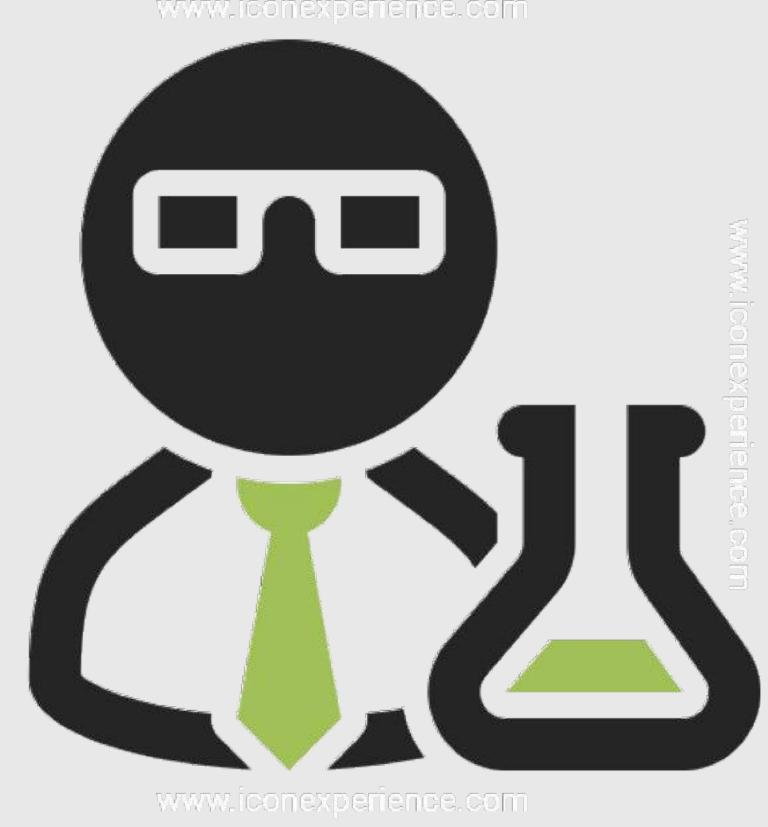
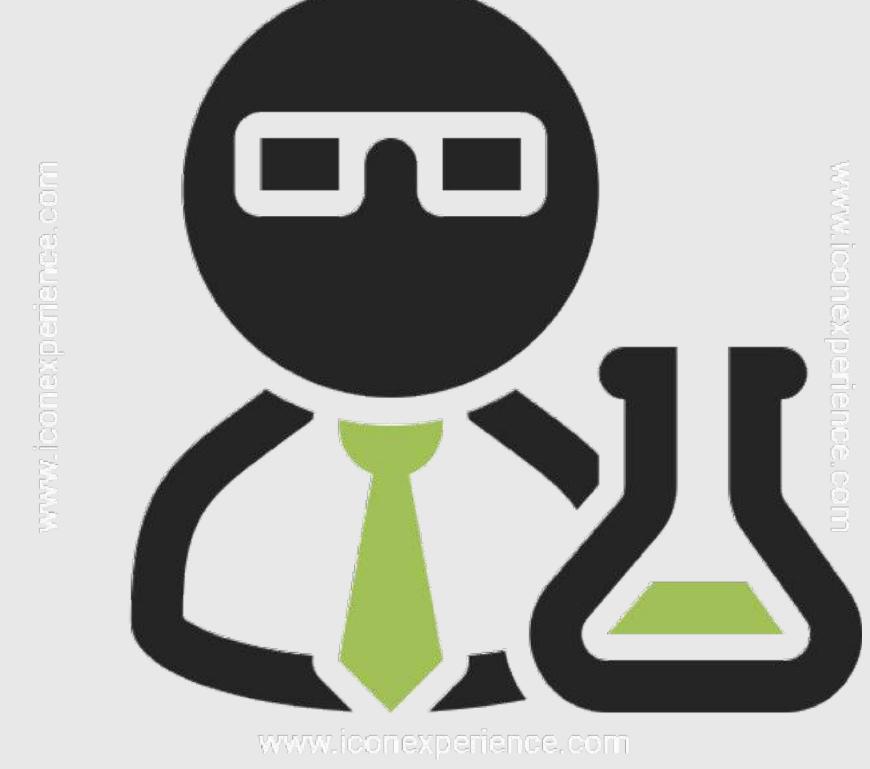
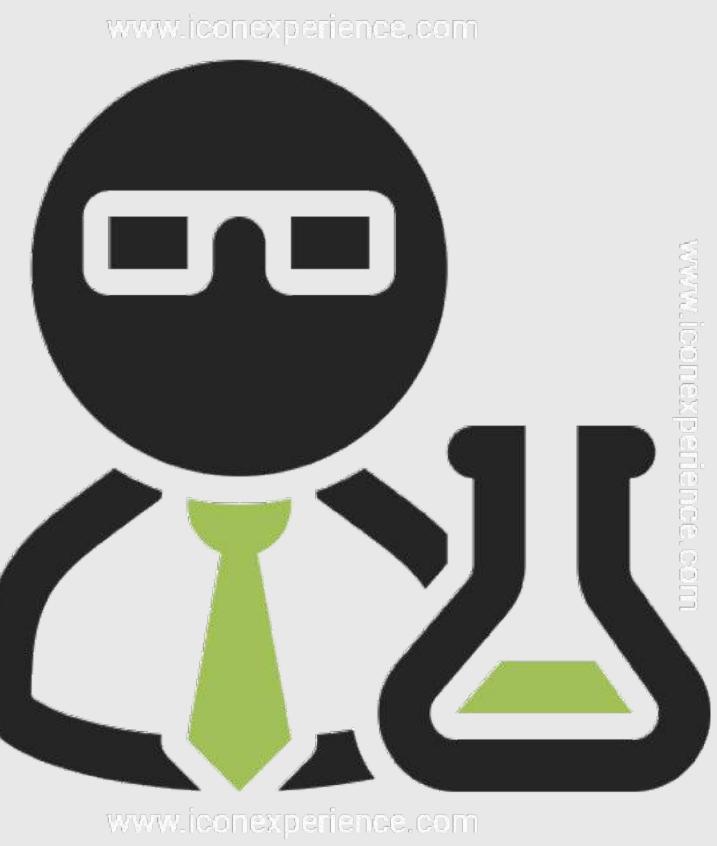
MATH & STATISTICS <ul style="list-style-type: none">★ Machine learning★ Statistical modeling★ Experiment design★ Bayesian inference★ Supervised learning: decision trees, random forests, logistic regression★ Unsupervised learning: clustering, dimensionality reduction★ Optimization: gradient descent and variants	PROGRAMMING & DATABASE <ul style="list-style-type: none">★ Computer science fundamentals★ Scripting language e.g. Python★ Statistical computing package e.g. R★ Databases SQL and NoSQL★ Relational algebra★ Parallel databases and parallel query processing★ MapReduce concepts★ Hadoop and Hive/Pig★ Custom reducers★ Experience with xaaS like AWS
DOMAIN KNOWLEDGE & SOFT SKILLS <ul style="list-style-type: none">★ Passionate about the business★ Curious about data★ Influence without authority★ Hacker mindset★ Problem solver★ Strategic, proactive, creative, innovative and collaborative	COMMUNICATION & VISUALIZATION <ul style="list-style-type: none">★ Able to engage with senior management★ Story telling skills★ Translate data-driven insights into decisions and actions★ Visual art design★ R packages like ggplot or lattice★ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

time?

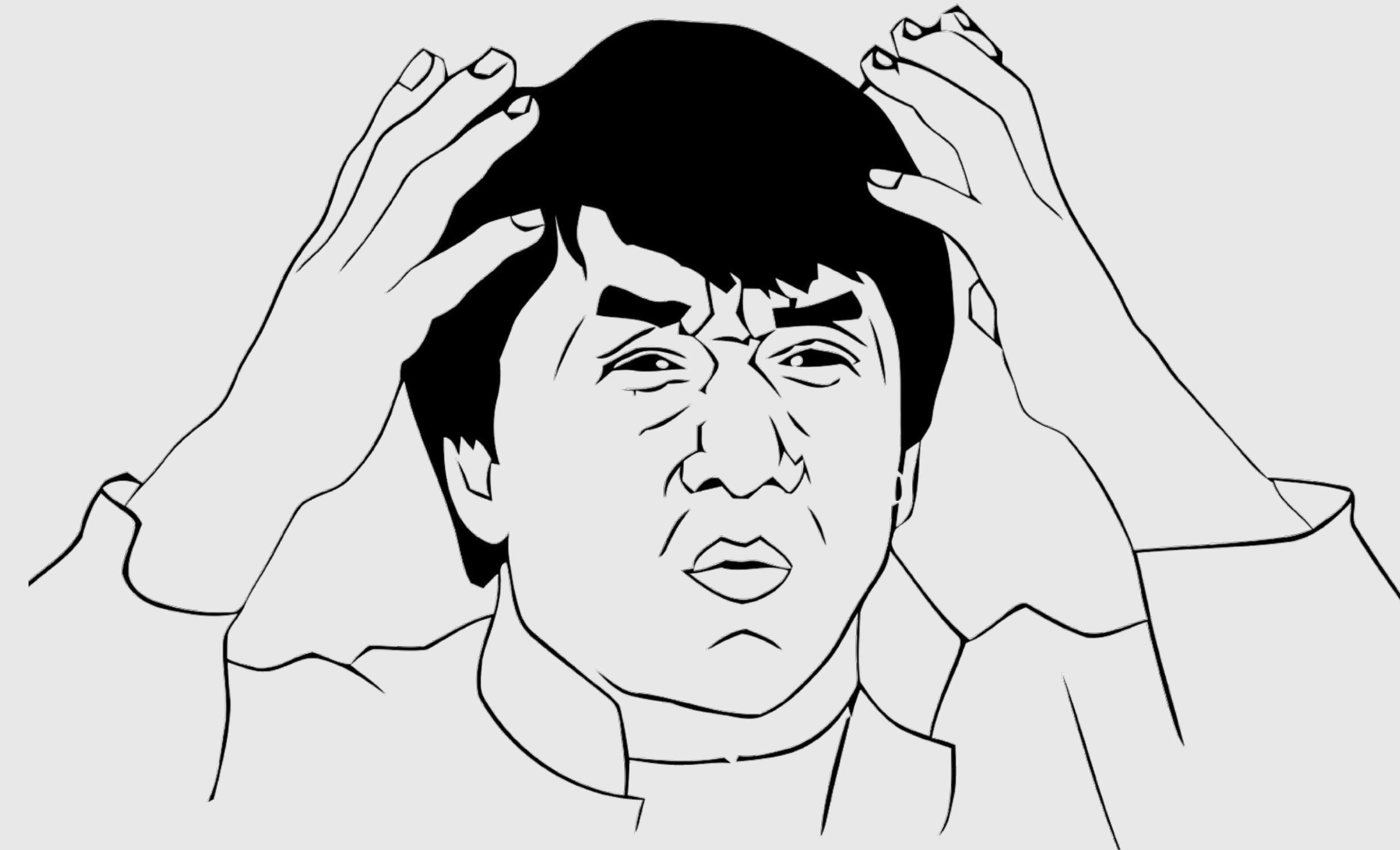


T

time



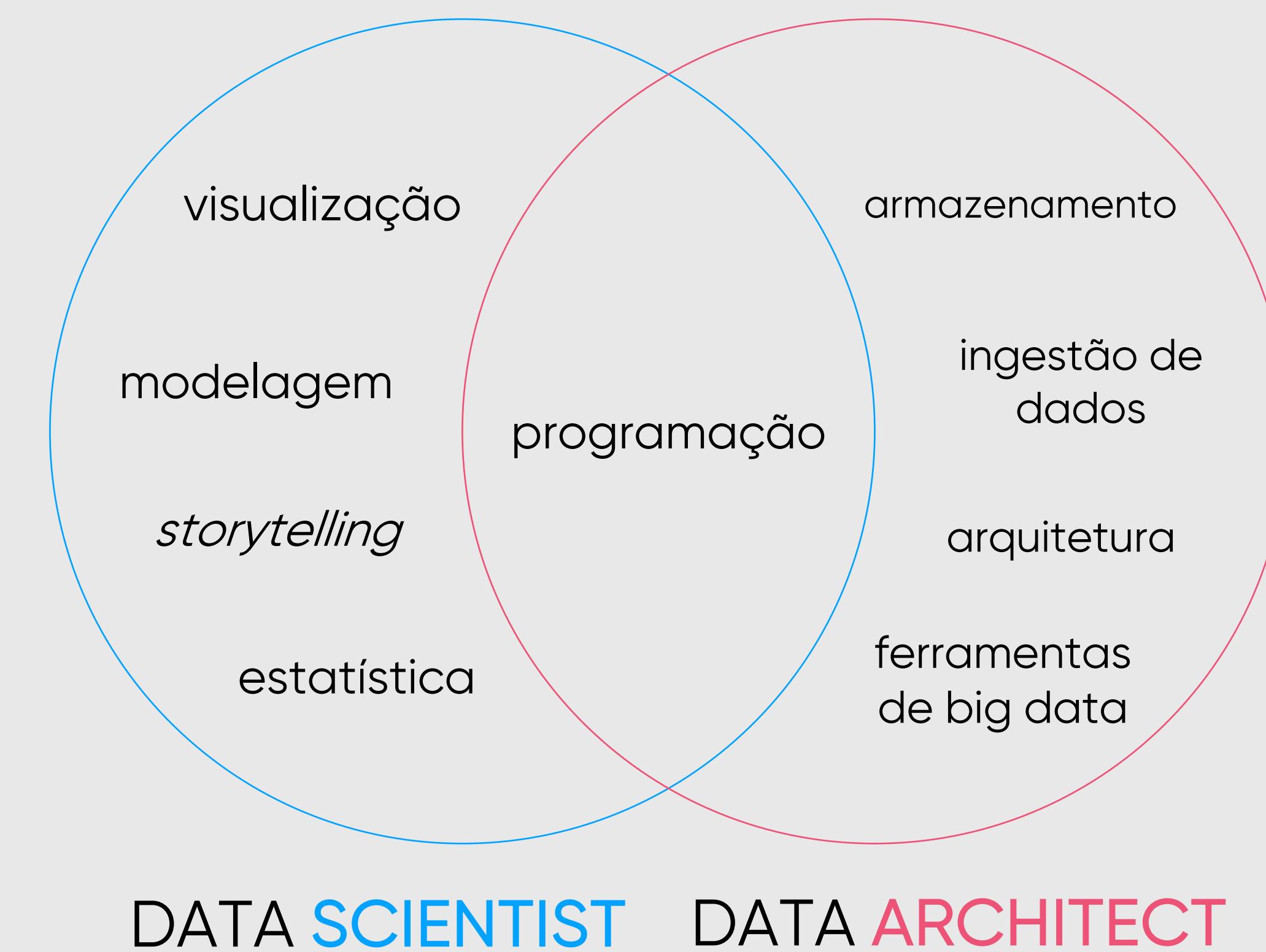
Database Administrator, Data Scientist,
Data Analyst, Data Engineer, Data
Architect, Data Manager... **WTH?!**



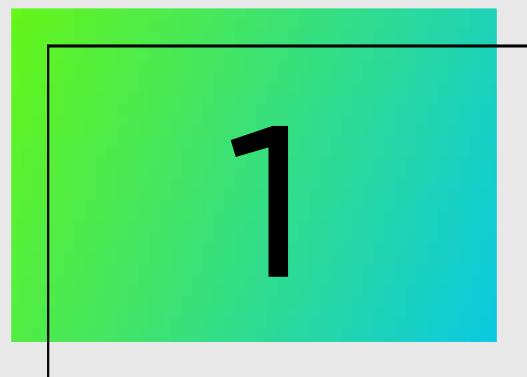
Data Scientist **vs** Data Engineer Architect?!



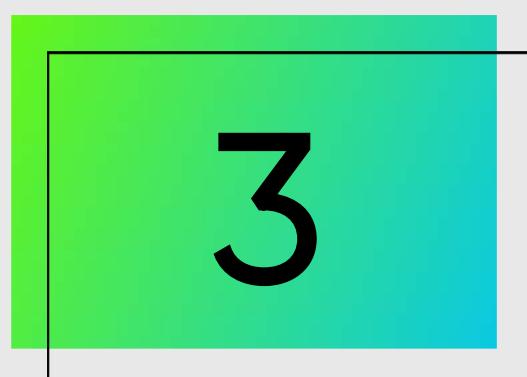
T papéis



T time



inteligência artificial



desenvolvimento

infraestrutura
(big data)



business



T

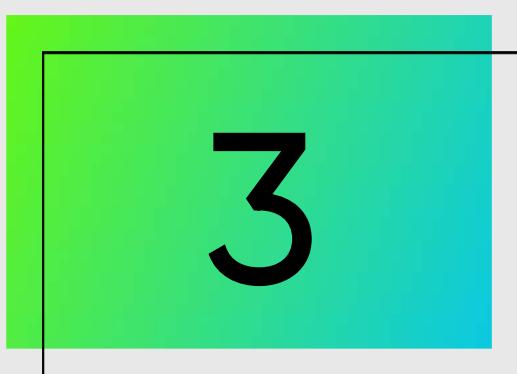
time



cientista de dados



arquiteto de dados



desenvolvedor



subject matter expertise

T papéis

- ❖ time multidisciplinar
- ❖ perfis Jr vs Sr?
 - + importante
 - ❖ perfil *hands-on*

➤ cientista de dados

- ❖ 4,524 vagas abertas
- ❖ salário médio de \$110.000

➤ gerente de *analytics*

- ❖ 1,381 vagas abertas
- ❖ salário médio de \$115.000

➤ engenheiro de software

- ❖ 29.187 vagas abertas
- ❖ salário médio de \$102,500

➤ DBA's

- ❖ 2,370 vagas abertas
- ❖ salário médio de \$94,000

› engenheiro de software

- ❖ 29.187 vagas abertas
- ❖ salário médio de \$102,500

› engenheiro de dados

- ❖ 2,816 vagas abertas
- ❖ salário médio de \$100,000

❯ analista de dados

- ❖ 4,729 vagas abertas
- ❖ salário médio de \$60,000

❯ DBA's

- ❖ 2,370 vagas abertas
- ❖ salário médio de \$94,000

T

agenda

◇ definição

buzzword... Data Science?

◇ business

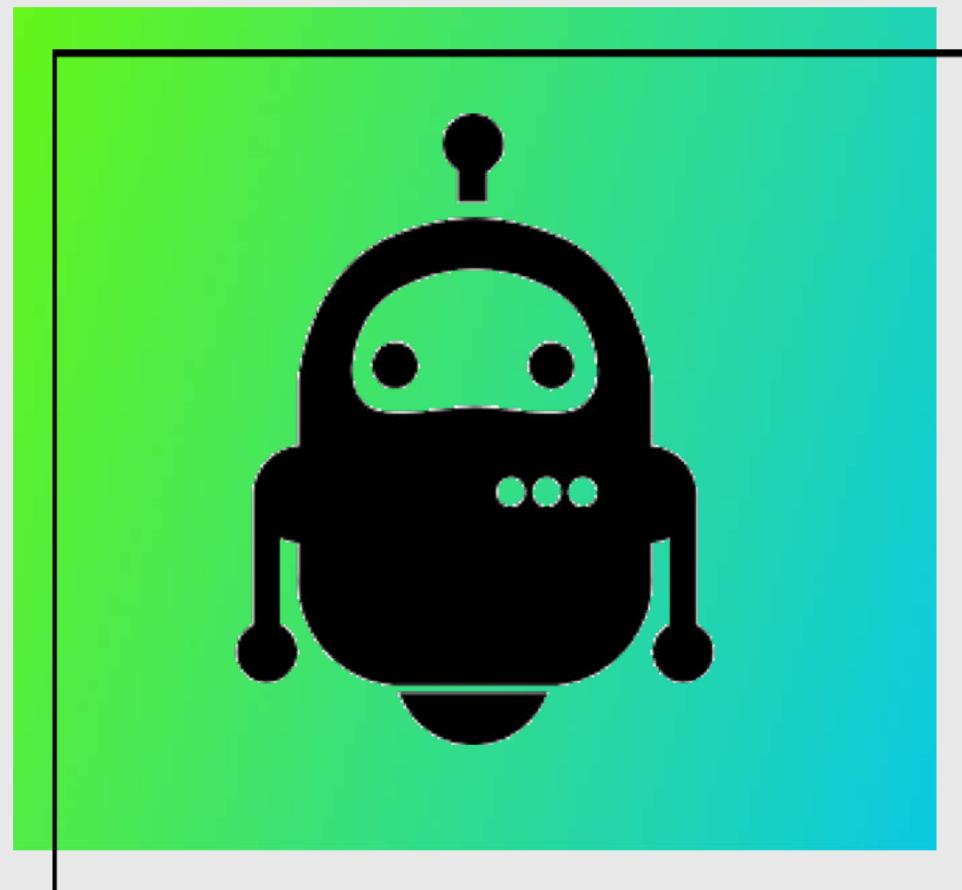
dados, mercado, previsões

◇ metodologia

ciclo de vida, atividades, papéis

◇ tecnologias

players, ferramentas,
landscape, diferenças



T tecnologias

A circular word cloud centered on the word 'learning'. The words are arranged in concentric circles around the center. The most prominent word is 'learning' in a large, bold, yellow-orange font. Other visible words include 'python', 'java', 'keras', 'IDE', 'scikit-learn', 'scikit', 'analise', 'machine', 'modelagem', 'language', 'processing', 'tensor', 'flow', 'google', 'amazon', 'cortana', 'ibm', 'nltk', 'gluon', 'alexa', 'natural', 'mxnet', 'theano', 'watson', 'algoritmos', 'deep', 'cortana', 'scala', and 'cortana'. The words are in various colors including blue, red, orange, and black.

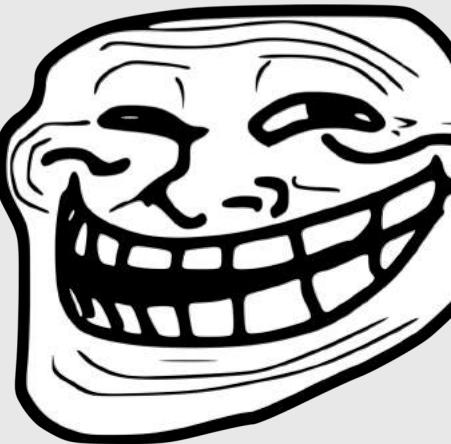
T programação



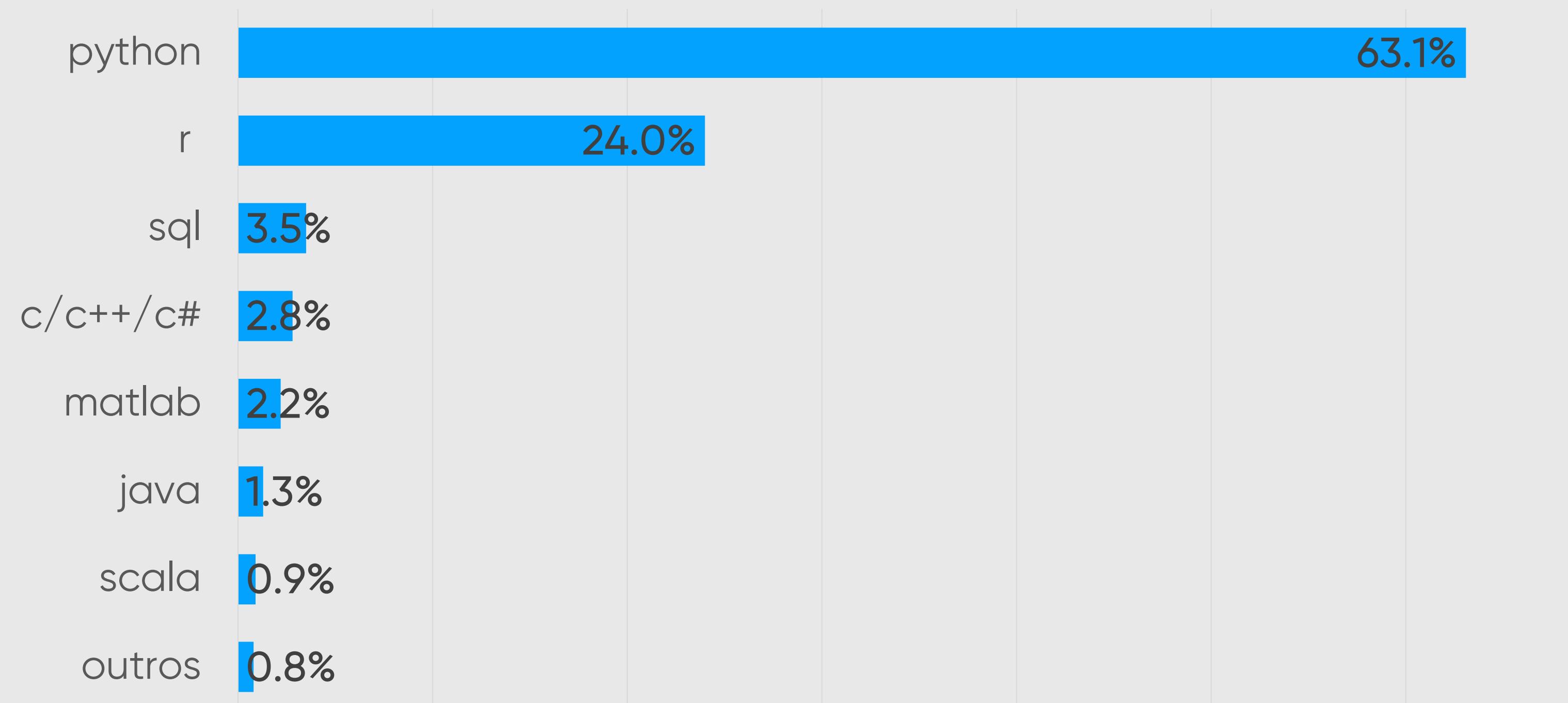
“

Nenhuma melhor ou pior

- Eu



T programação



T programação

- análise exploratória
- experimentação
- + produtividade



T programação

- análise exploratória
- + produtividade



T programação

- experimentação
- + produtividade



T programação

- análise exploratória
- experimentação
- + produtividade
- + performance



T programação

- produção
- ambientes homologados
- grandes corporações



T

IDE e visualização

jupyter explore-spark_full Last Checkpoint: a few seconds ago (autosaved)

File Edit View Insert Cell Kernel Help

Not Trusted | No Kernel

Markdown



T

Aula 30

Lidando com dados não-estruturados em Big Data

Este é um tutorial introdutório ao processamento com PySpark visando apresentar os principais conceitos, diferenças e simular um cenário de exemplo mais realista do que o corriqueiro **wordcount**. É importante mencionar que iremos **simular** o Spark em execução no modo **standalone**, ou seja, em uma única máquina, ao invés do cenário real de um **cluster**.

Configuração inicial

Vamos importar as bibliotecas básicas para startar o Spark.

```
In [1]: from pyspark import SparkContext  
from pyspark.sql import SparkSession
```

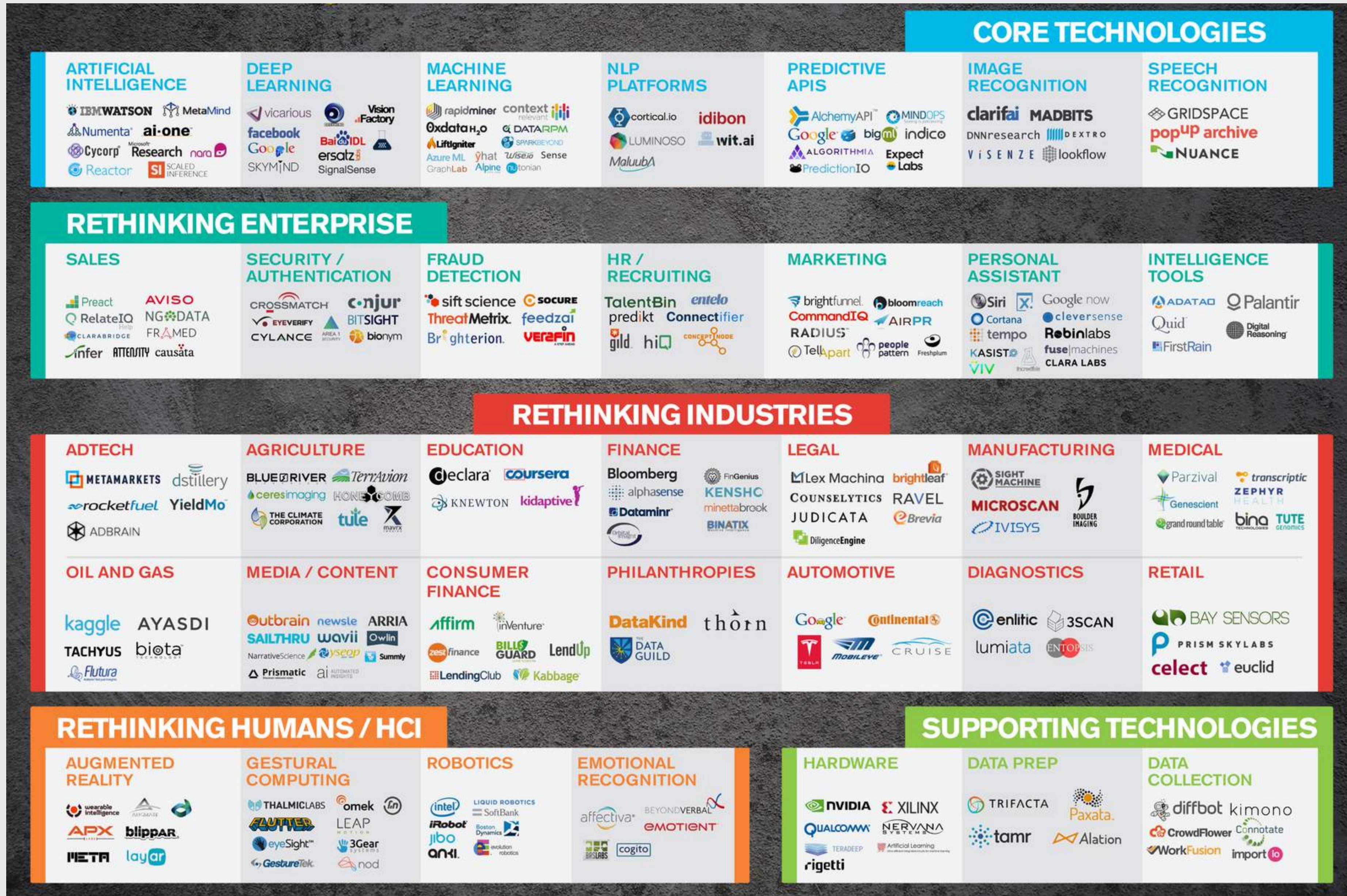
Você precisa de, no mínimo, um **contexto** para conseguir iniciar uma **sessão**.

Lembre-se que estamos **simulando** o Spark, mas ainda estamos rodando em uma única máquina.

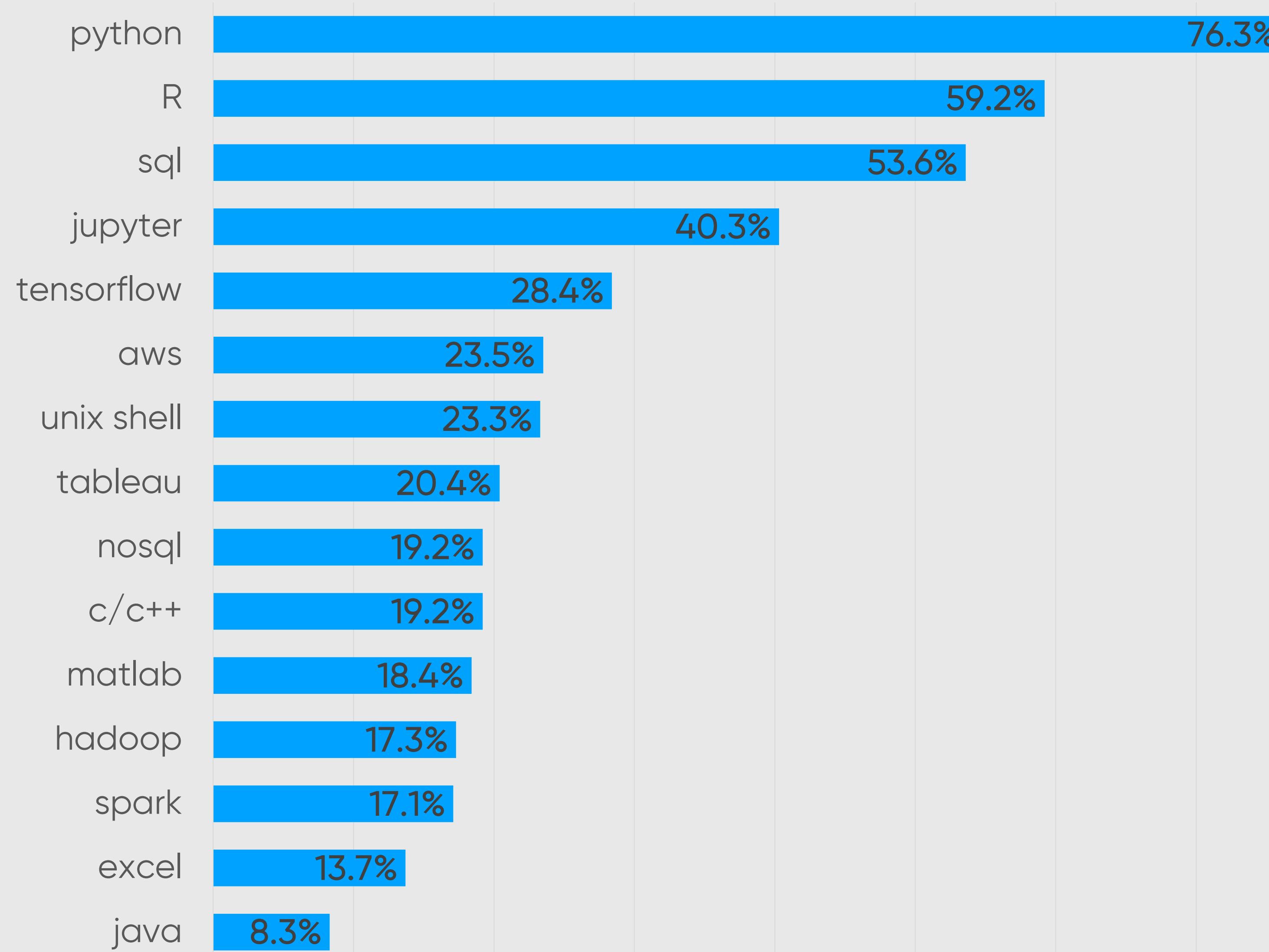
Chamamos isso de modo **standalone**.

```
In [2]: spark = SparkSession.builder.master("local").appName("tera").getOrCreate()  
sc = spark.sparkContext
```

T landscape



T ferramentas



open-source **vs** proprietária?!



T machine learning open-source



T nlp open-source

gensim

openNLP

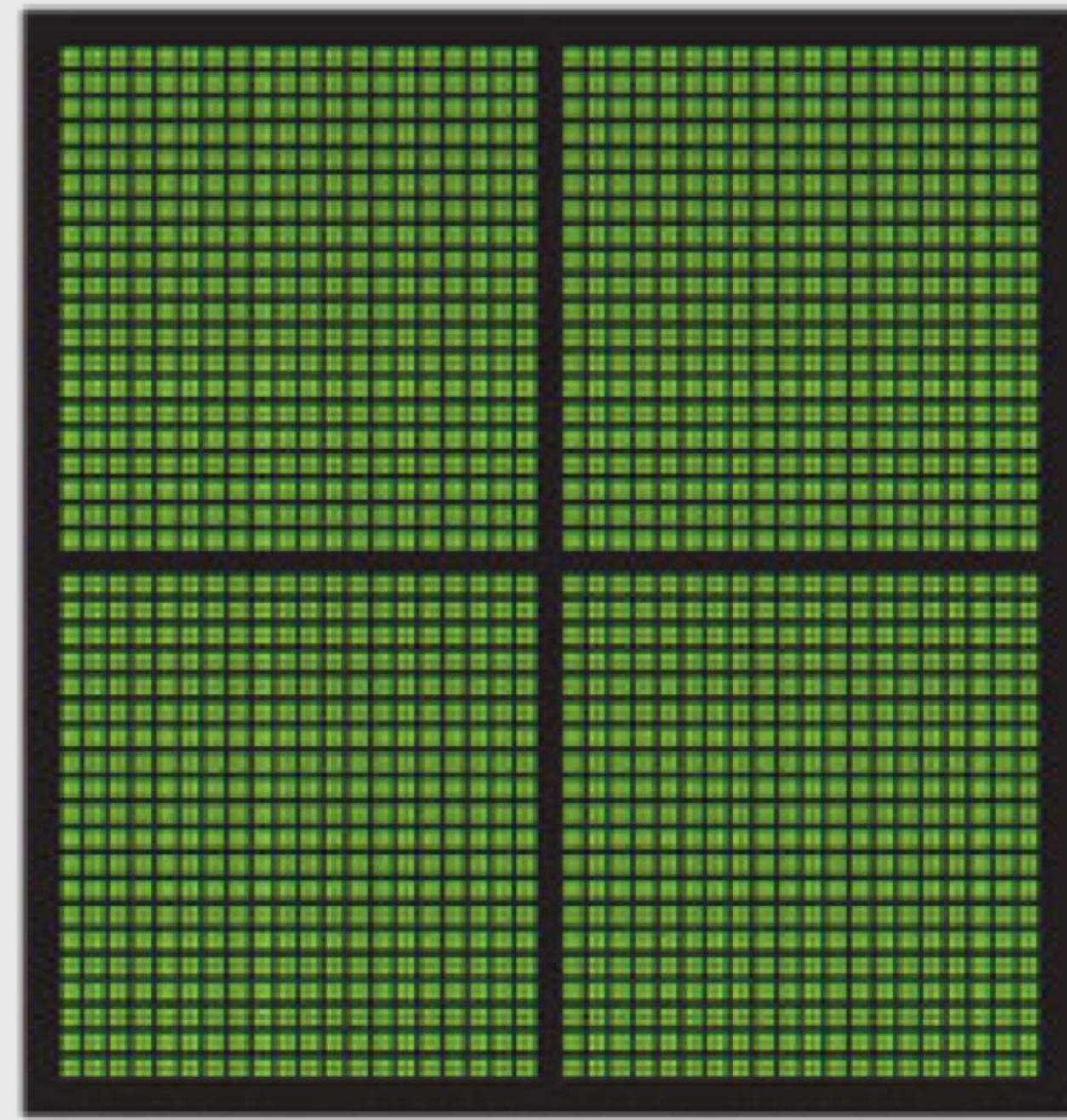
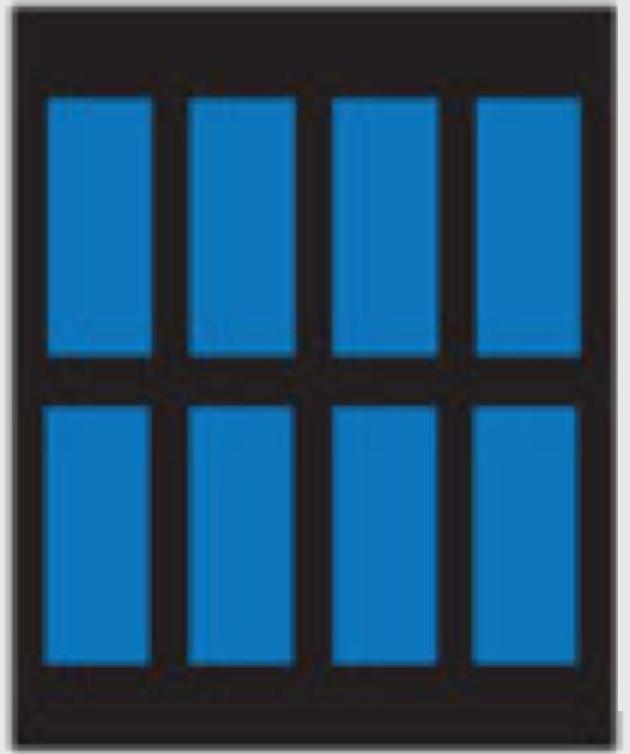


NLTK

 John Snow LABS

T

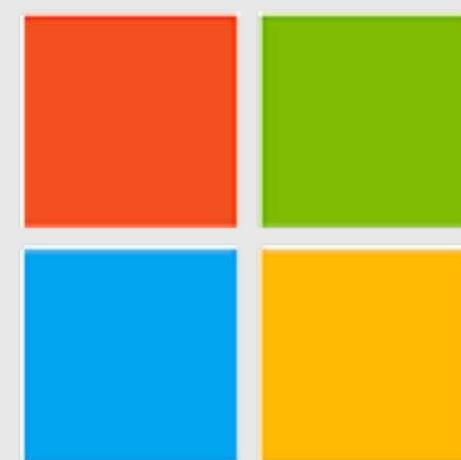
gpu



T players



Google



Microsoft

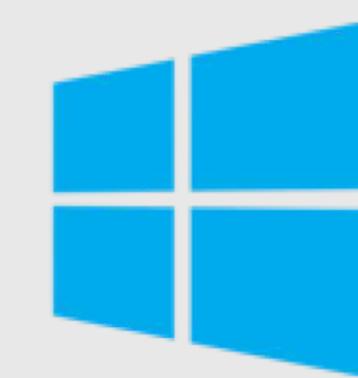
T players



Google Cloud Platform

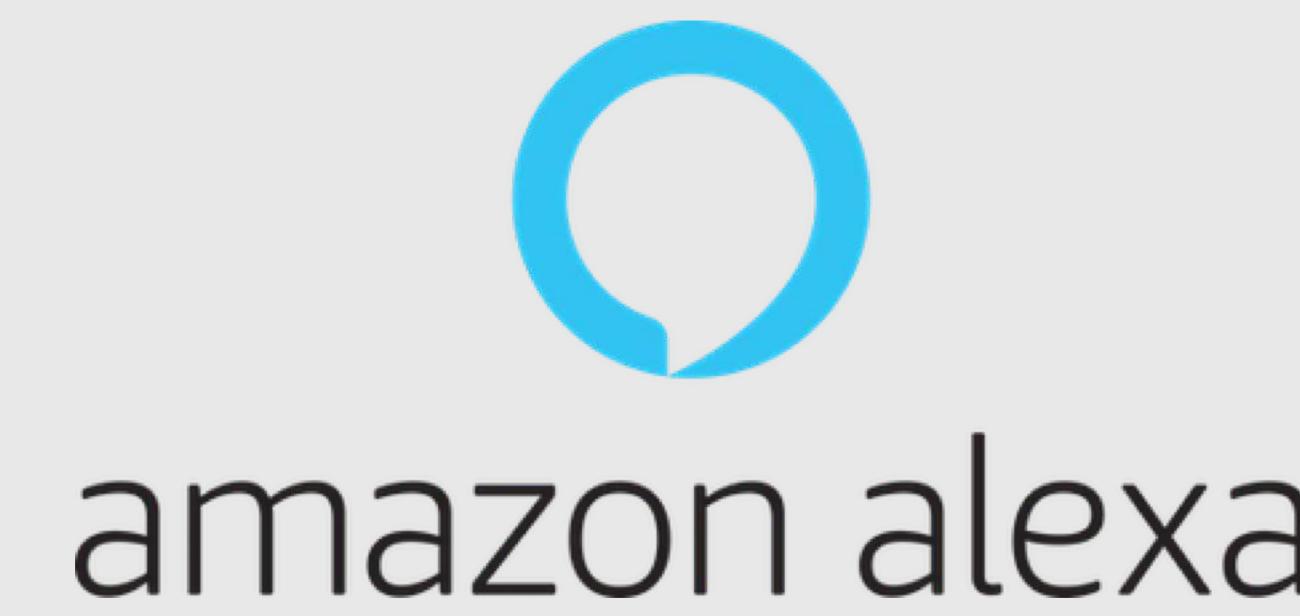


IBM Bluemix™



Microsoft
Azure

T players



T deep learning open-source



PYTORCH

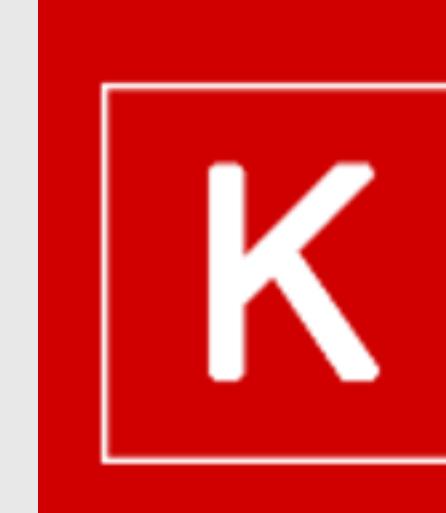
Microsoft
CNTK

theano

 Caffe2

The Caffe2 logo icon consists of a white coffee cup outline with two white plus signs above it, positioned to the left of the word "Caffe2".

Keras

The Keras logo icon is a red square containing a white stylized letter 'K'.

mxnet

resumão

T

resumão

- ❖ unicórnios **não** existem
 > **time multidisciplinar**
- ❖ **teoria** é essencial
- ❖ **business** é essencial
- ❖ **saber apresentar** é essencial



T

resumão

- ❖ algoritmos+ferramentas+bibliotecas+diversão
= **20% do seu tempo**
- ❖ não precisa ser complexo
> **baby steps**
- ❖ tudo muito legal, mas ainda é **ciência**
- ❖ tudo muito legal, mas
utilidade = +valor no negócio

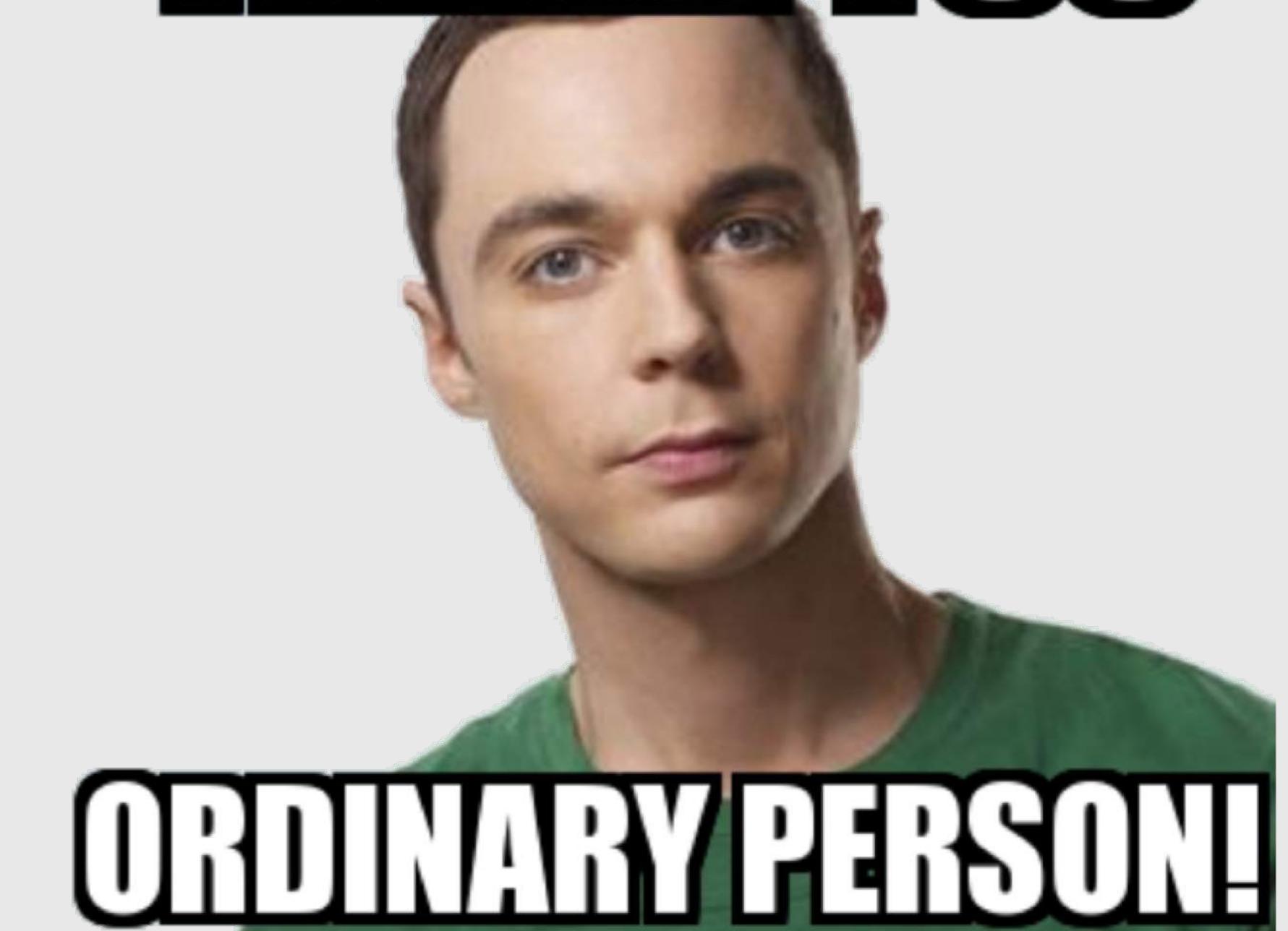


T

dúvidas?

T

THANK YOU



ORDINARY PERSON!

 IA estatísticas:

- [previsão 2018](#)
- análise do [mercado](#)
- cheat [sheet](#)
- [infográficos](#)

 [landscape](#) de ferramentas kaggle [report](#) 2017 [toolkit](#) [magic quadrant](#) (gartner) [livros](#)

T

extra!

❖ datasets públicos:

- [dataset públicos da aws](#)
- [datasets públicos armazenados na aws](#)
- [kaggle](#)
- [git](#)
- [google](#)
- [uci](#)