
Tera

AULA 05

Módulo 02: **Fundamentos de DS & ML**

Introdução e Fundamentos de **Data Science**

instrutora:

RAYSSA KÜLLIAN

hoje:

- **Líder da Prática de IA @LATAM**
  Amazon Web Services
- **Professora**
  FGV, PUC, Tera
- **Pesquisadora**
  LIAMF, IME/USP
- **Palestrante**
  Fala Criativa

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

agenda

definição

~~buzZzZzword~~... **Data Science?**

business

dados, mercado, previsões

metodologia

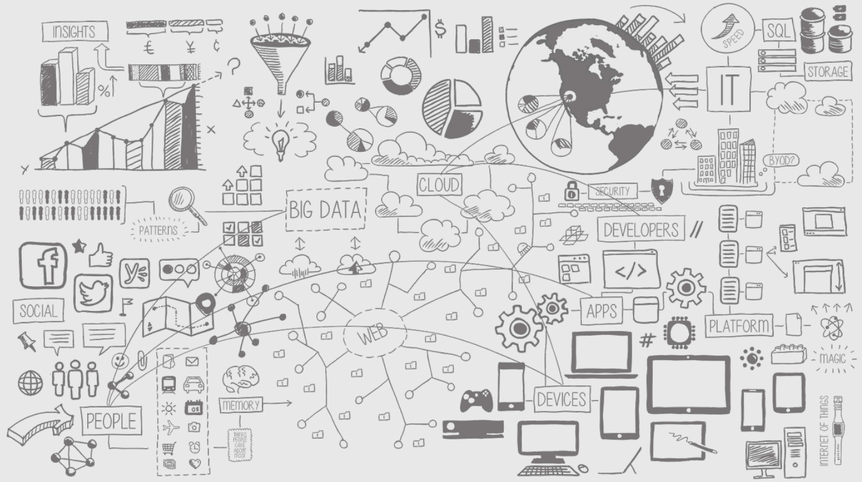
ciclo de vida, atividades, papéis

tecnologias

players, ferramentas,
landscape, **diferenças**

prática

the moment you've all been waiting for!



agenda

definição

~~buzZzZzword~~... Data Science?

business

dados, mercado, previsões

metodologia

ciclo de vida, atividades, papéis

tecnologias

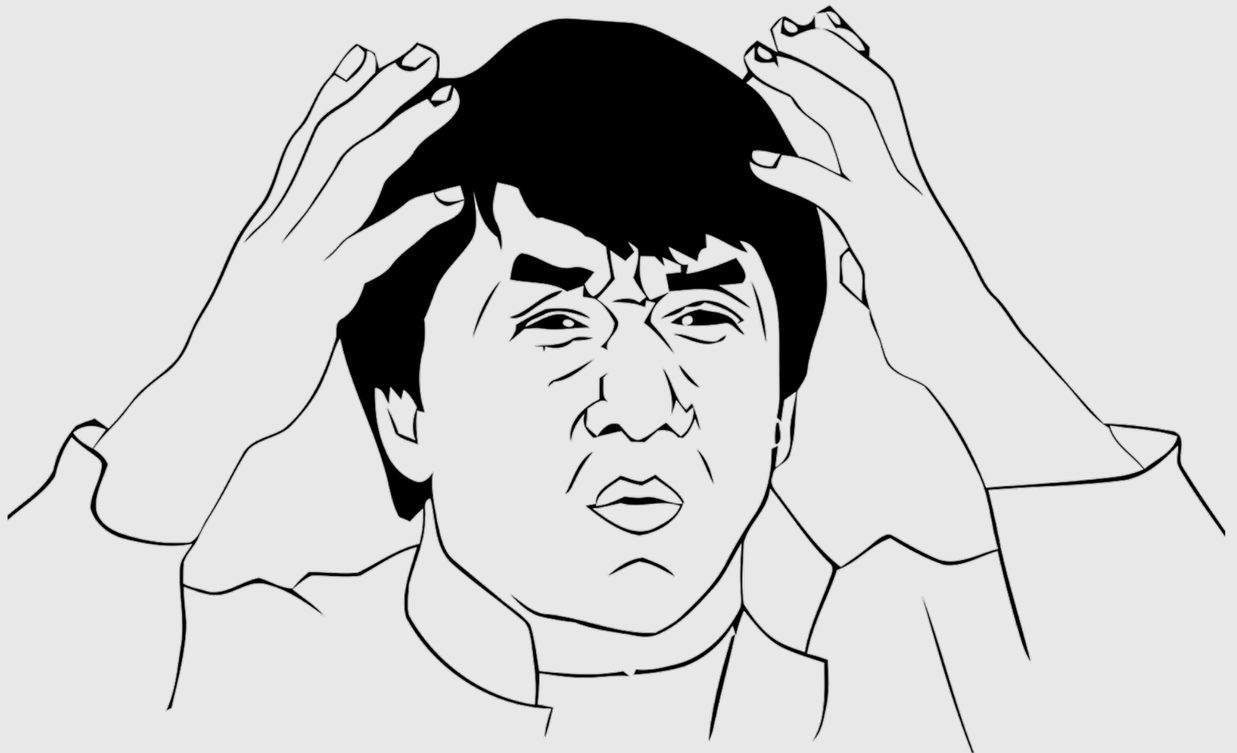
players, ferramentas,
landscape, diferenças

prática

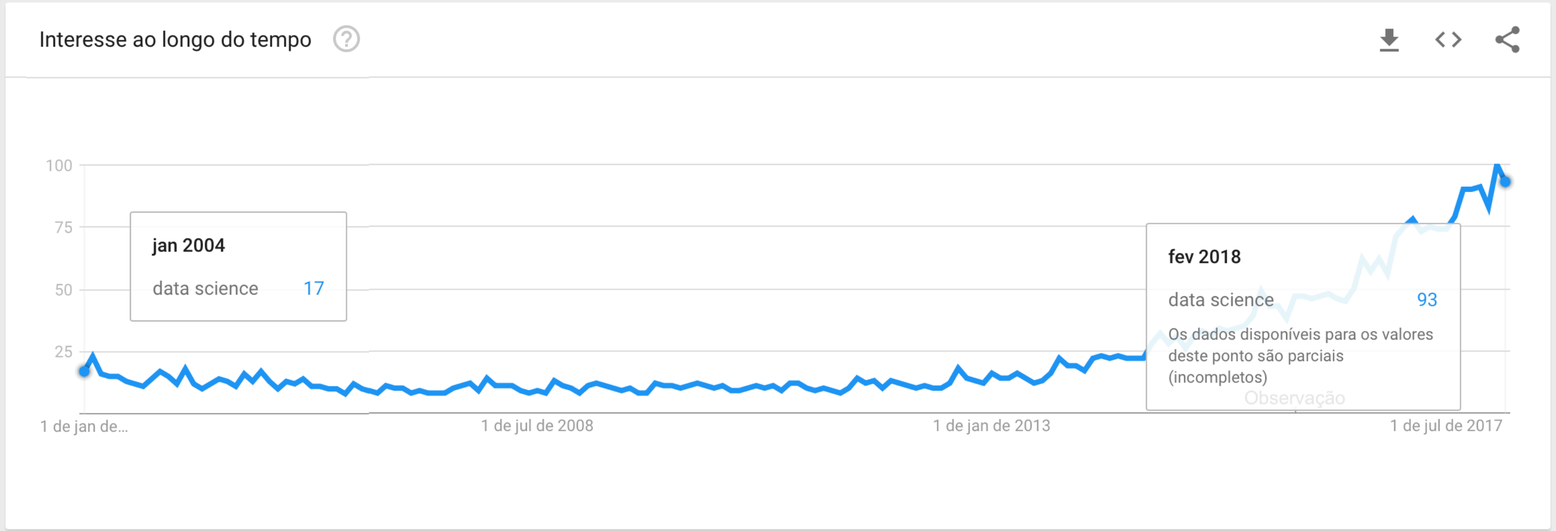
the moment you've all been waiting for!

o que é?

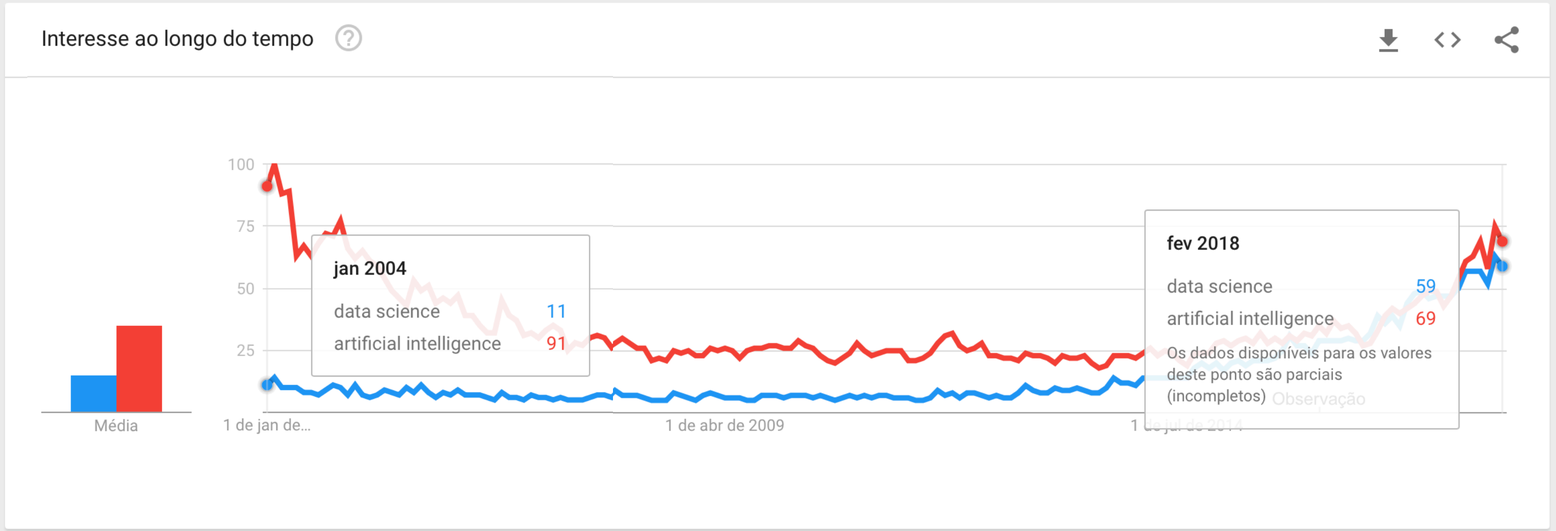
BI, Big Data, Analytics, Data Science, **AI**,
Machine Learning, Deep Learning,
#buzzzZzzzwords?!



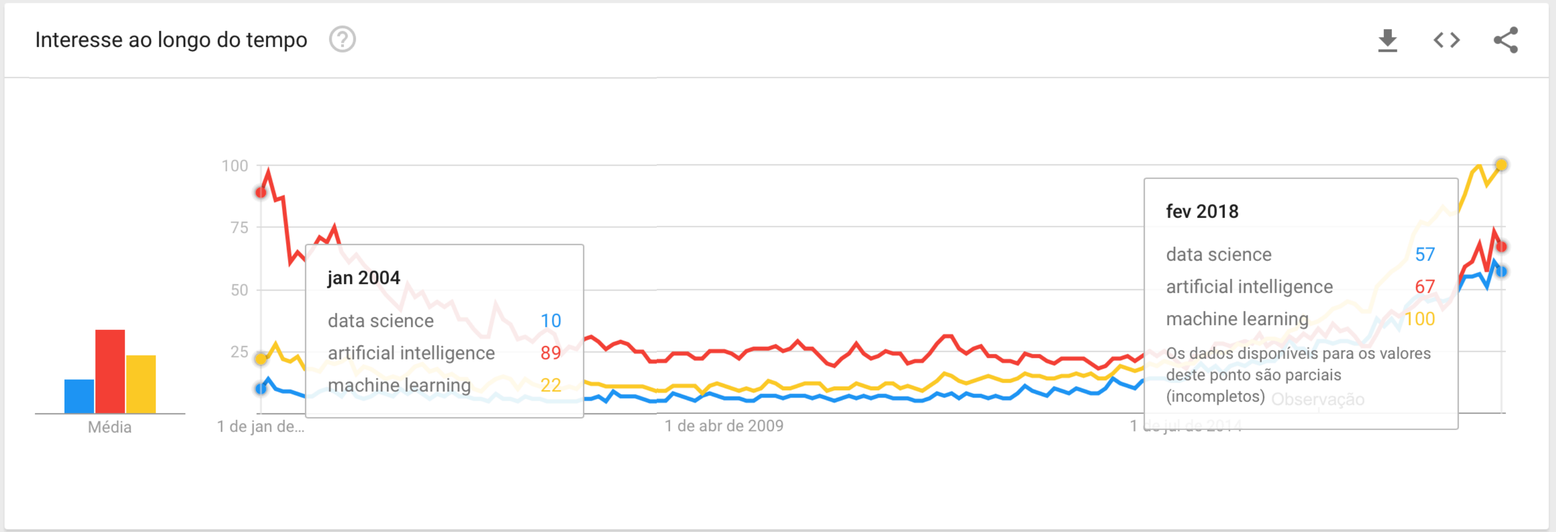
data science



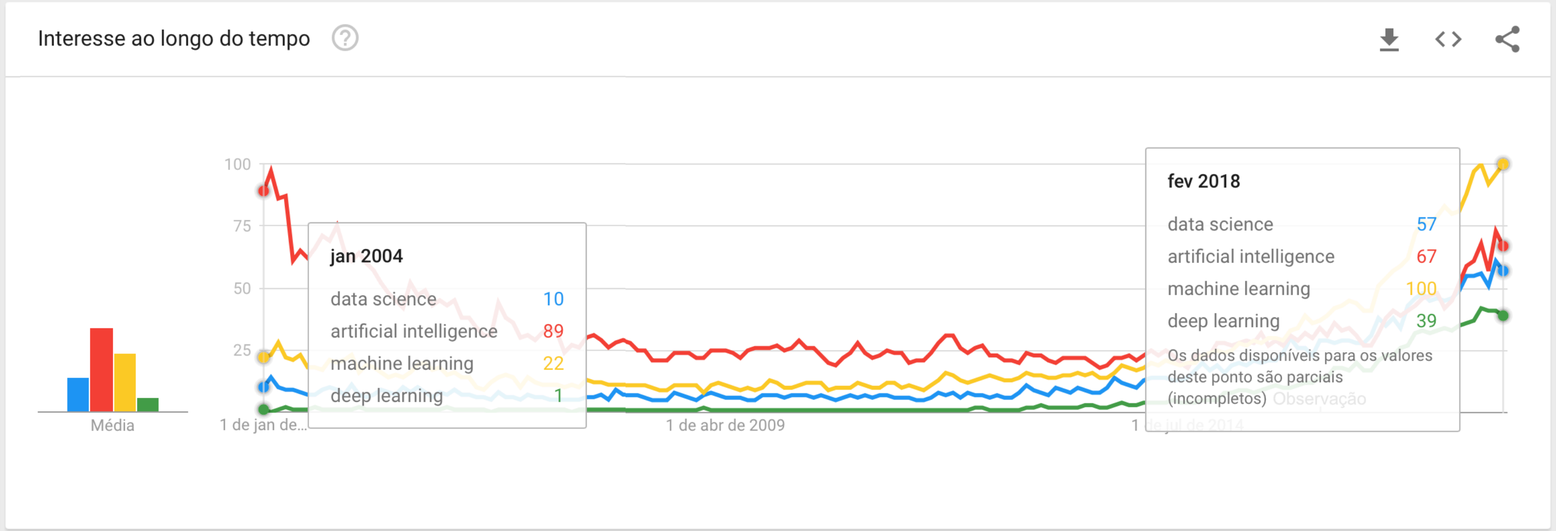
data science vs artificial intelligence



data science vs artificial intelligence vs machine learning



data science vs artificial intelligence vs machine learning vs deep learning



o que não é?

o que não é?

o que não é?

o que é?

“

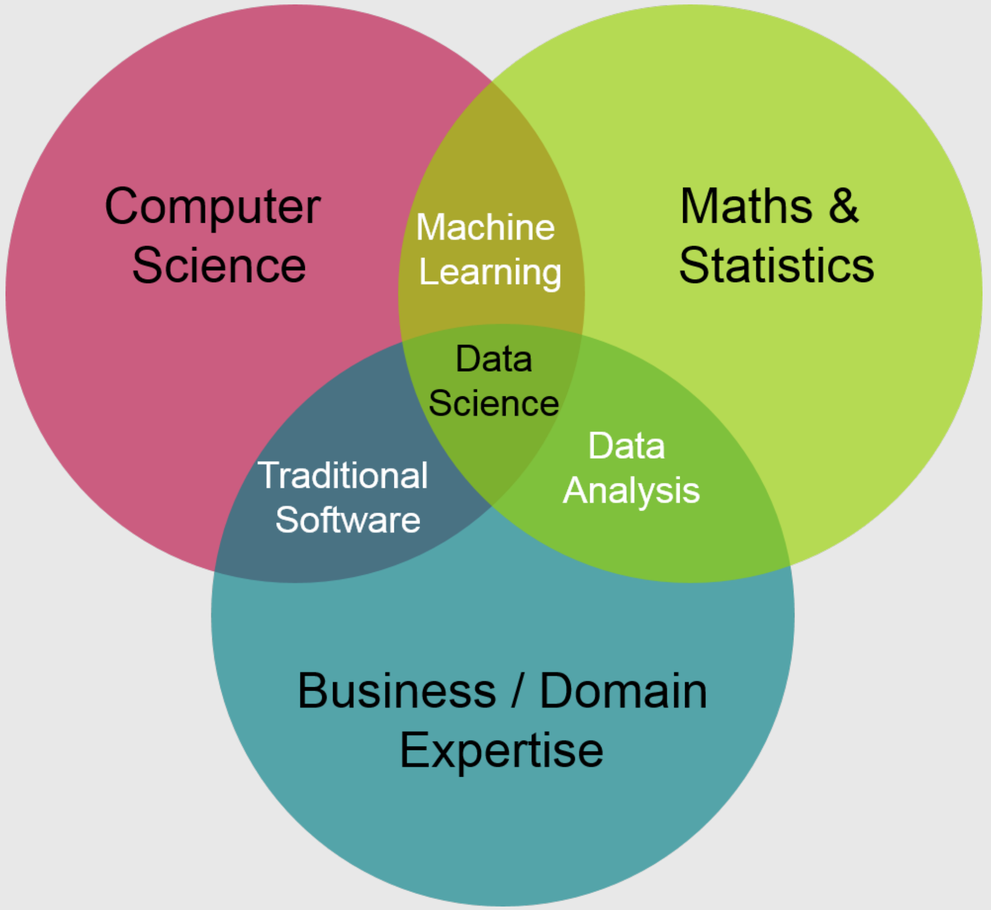
Automatização de atividades que associamos ao pensamento humano, como a tomada de decisões, a resolução de problemas, o aprendizado...

– Bellman, 1978

histórico

- gestação ~1943 à 1955
  "Computing Machinery and Intelligence"
  (Macculloch, Pitts, **Turing**)
- nascimento 1956
  McCarthy, Minsky, Shannon e Rochester
- indústria 1956
- **winter** has come 1974

data science



agenda

definição

~~buzZzZzword~~... Data Science?

business

dados, mercado, previsões

metodologia

ciclo de vida, atividades, papéis

tecnologias

players, ferramentas,
landscape, **diferenças**

prática

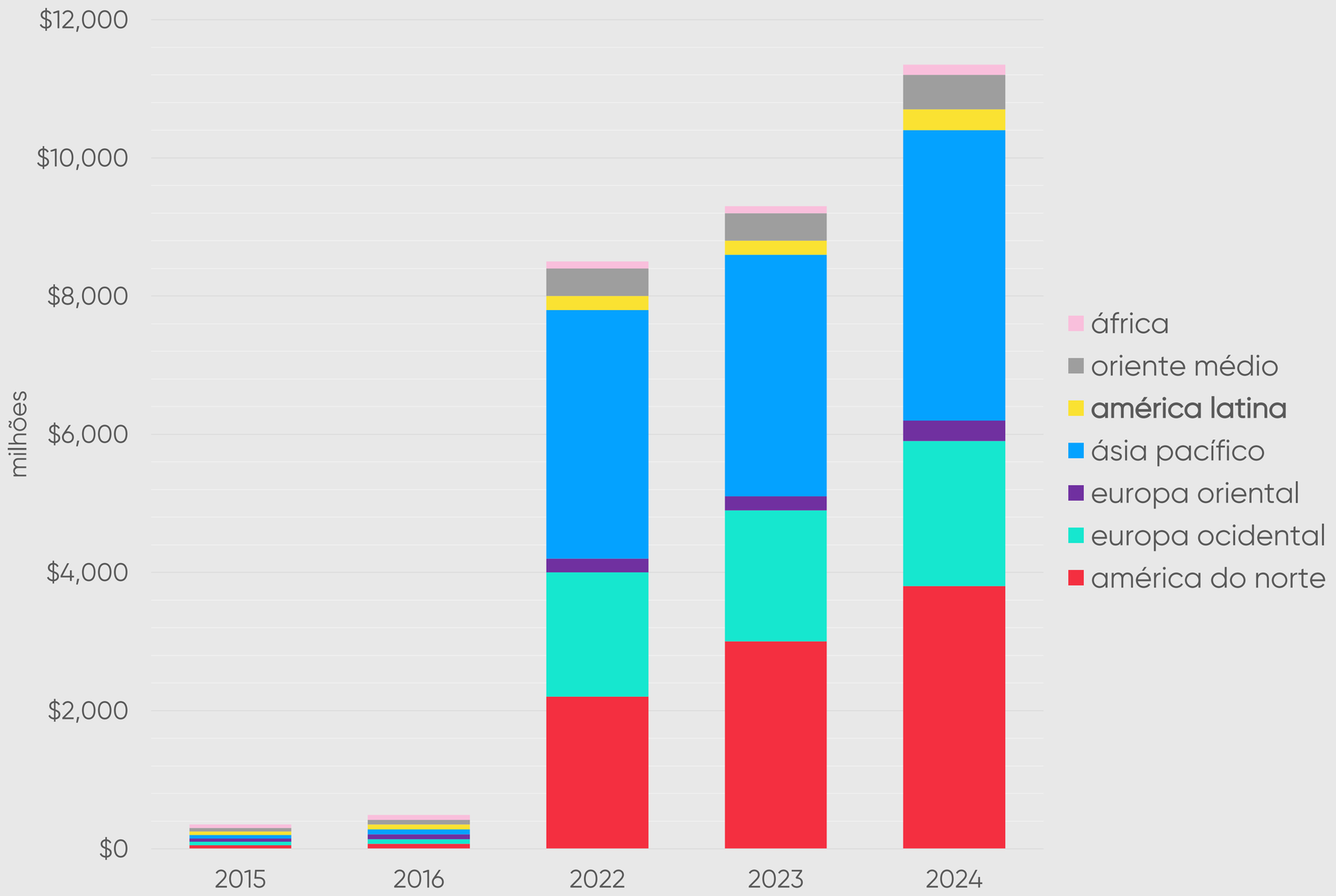
the moment you've all been waiting for!

previsão

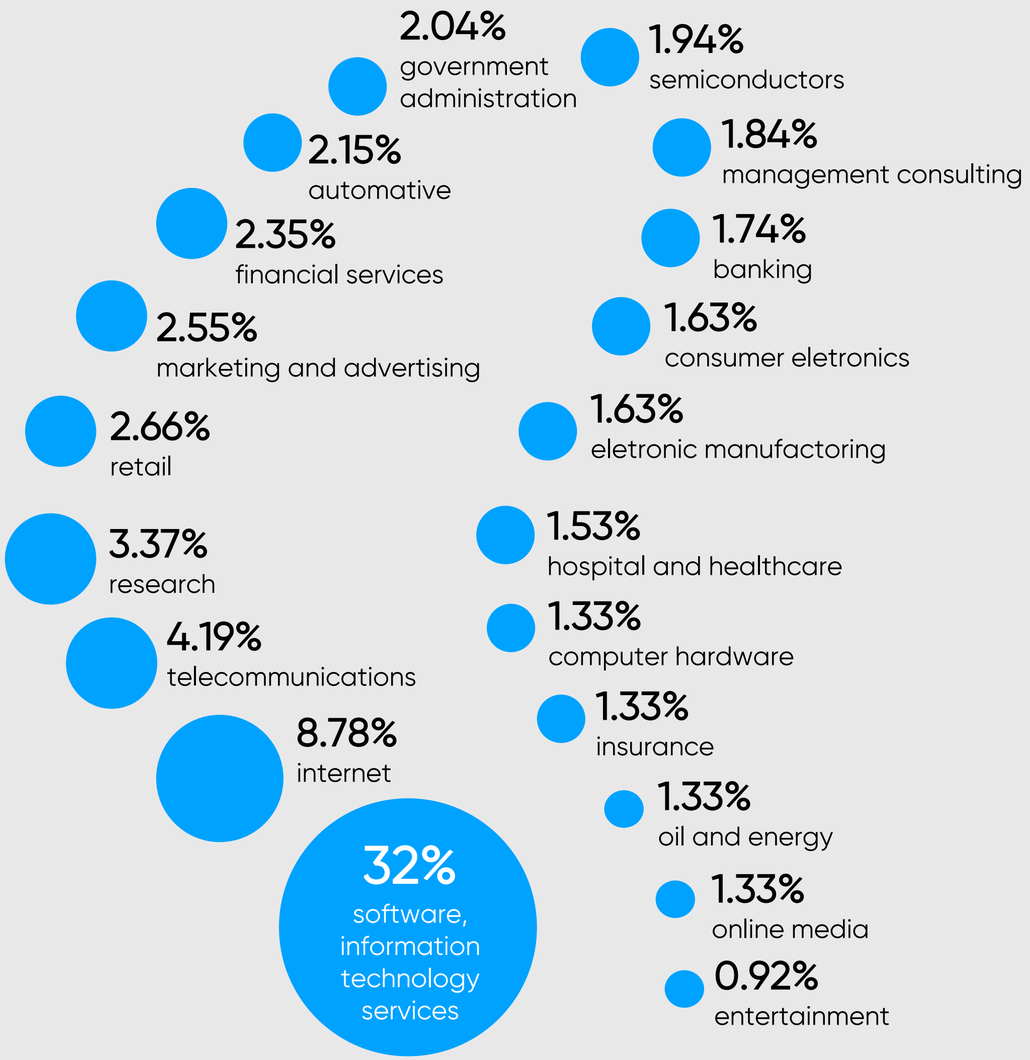
- 61% das empresas
  iniciativa mais importante de 2018

- dobra em 2018, dobra novamente em 2020
  número de **pilotos** e **experimentações**

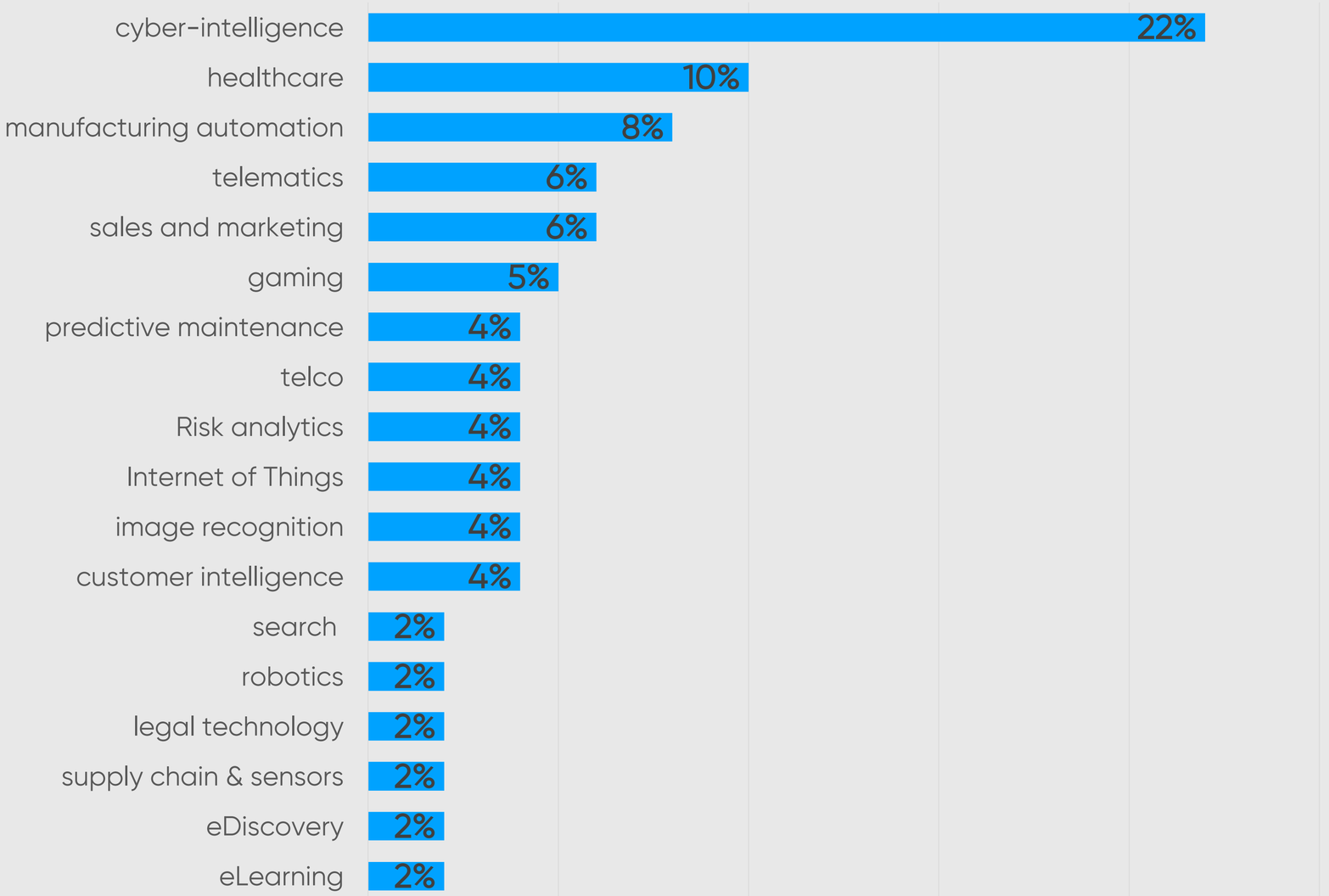
crescimento



indústria


2.04%
government administration
1.94%
semiconductors
2.15%
automative
1.84%
management consulting
2.35%
financial services
1.74%
banking
2.55%
marketing and advertising
1.63%
consumer eletronics
2.66%
retail
1.63%
eletronic manufactoring
3.37%
research
1.53%
hospital and healthcare
4.19%
telecommunications
1.33%
computer hardware
8.78%
internet
1.33%
insurance
32%
software, information technology services
1.33%
oil and energy
1.33%
online media
0.92%
entertainment

casos de uso



% recursos utilizados

maturidade

direcionamento estratégico
para o negócio
(Nível 3)

87

construindo aplicações
(Nível 2)

494

experimentações
(Nível 1)

967

**qtde de empresas*

empresas que mais **investem**

- Google
- Facebook
- Rocket Fuel
- IBM
- Amazon
- Yahoo
- Intel
- Microsoft
- Deloitte
- MITRE

agenda

definição

~~buzZzZzword~~... Data Science?

business

dados, mercado, previsões

metodologia

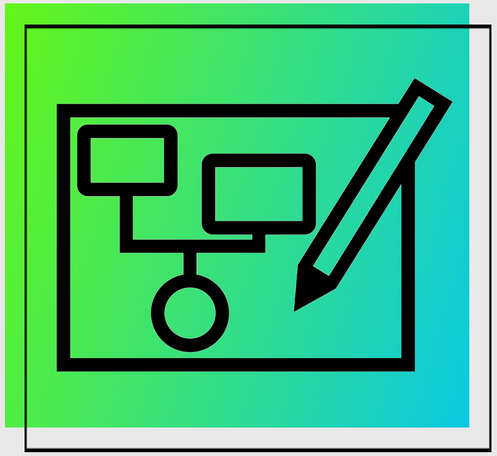
ciclo de vida, atividades, papéis

tecnologias

players, ferramentas,
landscape, **diferenças**

prática

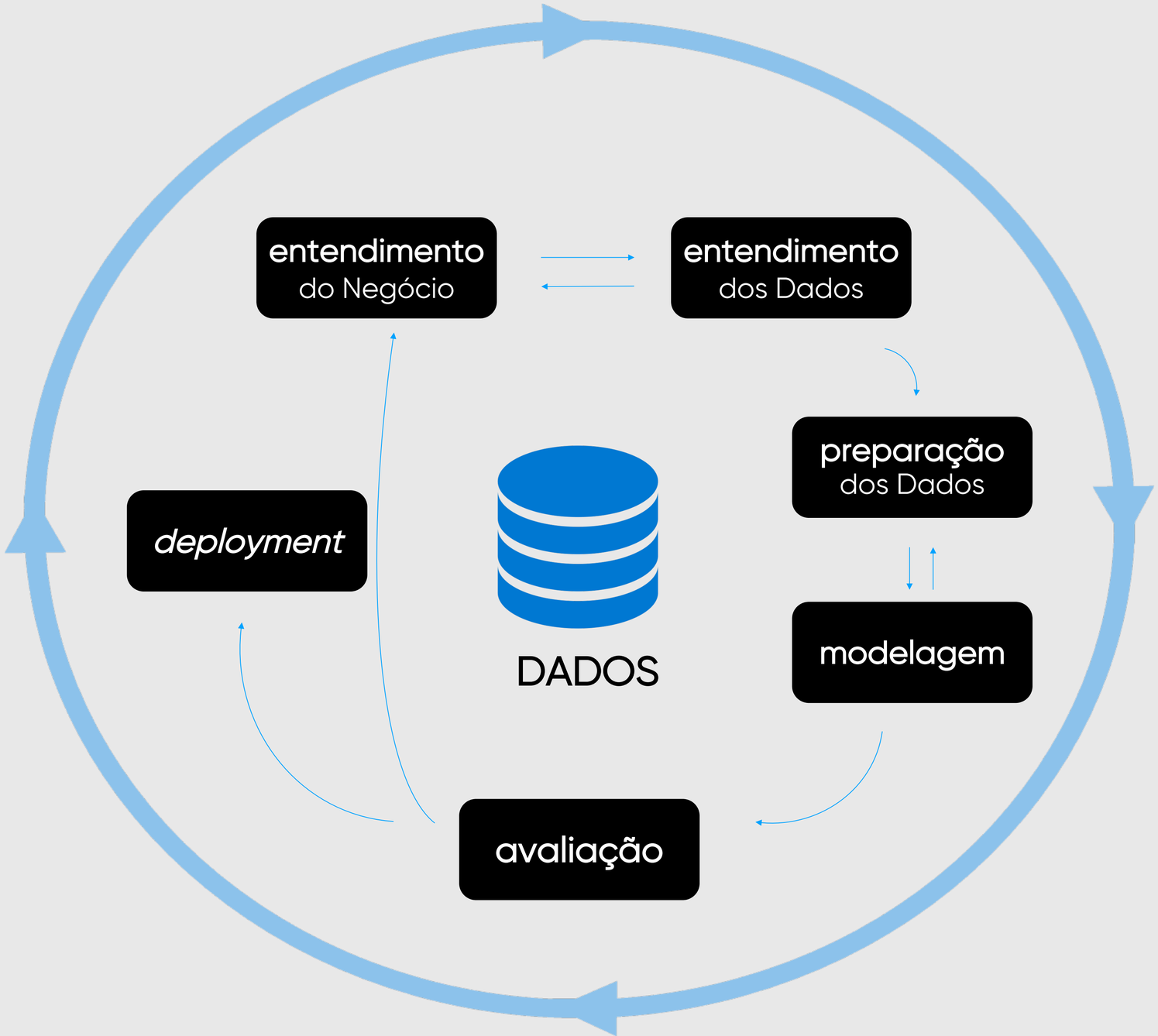
the moment you've all been waiting for!



CRoss Industry Standard Process for Data Mining

CRISP-DM

crisp-dm


entendimento
do Negócio
entendimento
dos Dados
preparação
dos Dados
modelagem
avaliação
deployment
DADOS

atividades

1 **explorar** e conhecer cenários de **negócios**

definir problema (**hipótese**) 2

3 setar **expectativas**

T atividades

4 pesquisar **estado-da-arte**

análise exploratória de dados (EDA) 5

6 **limpeza** e **pré-processamento** de dados

atividades

7 **engenharia** de *features*

modelagem 8

9 avaliação de **resultados**

T atividades

10 *fine-tuning* ou *boosting*

validação com área de negócio 11

12 produtização

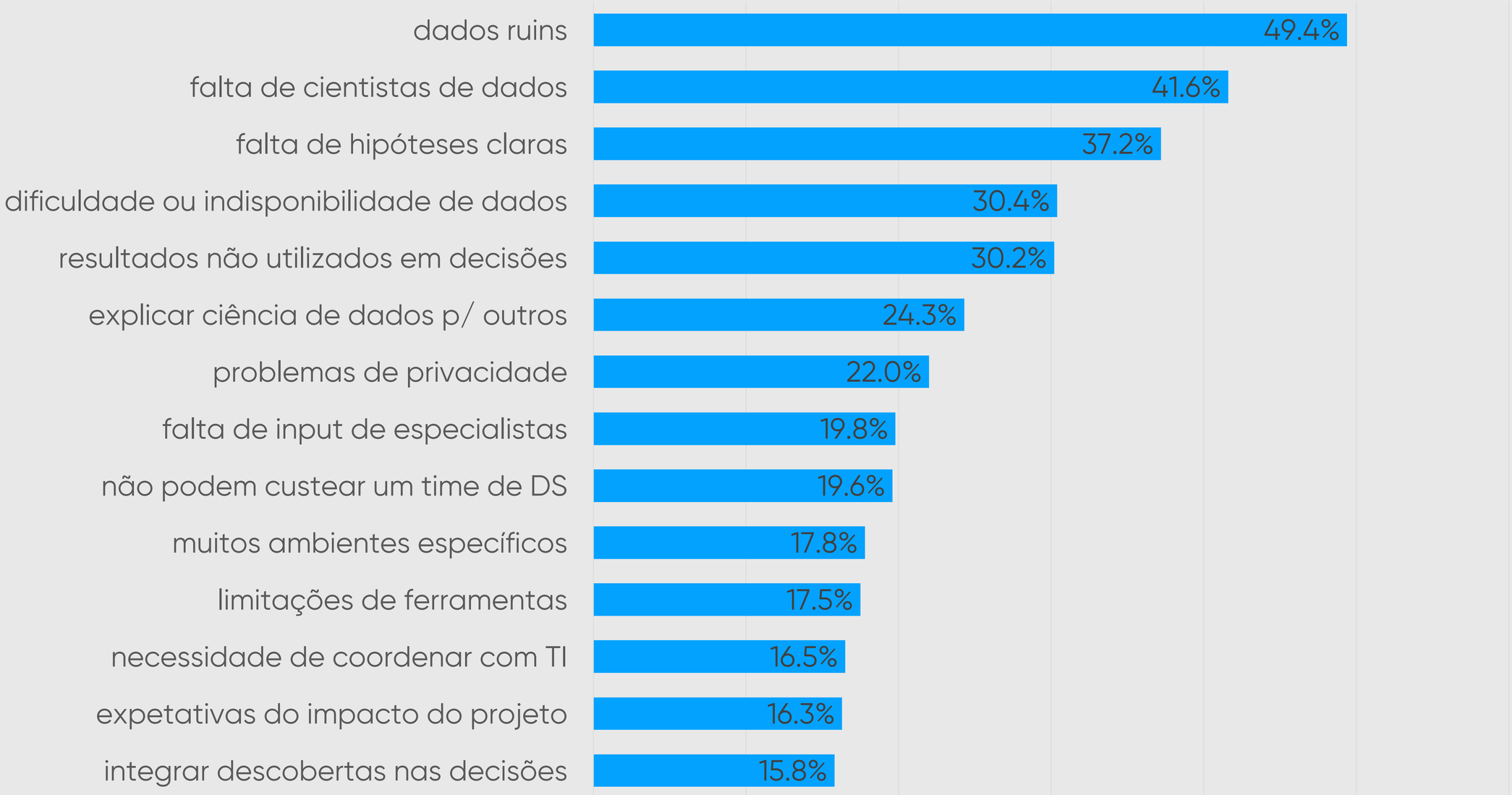
atividades

- paper acadêmico
  boa prática de mercado
  - transferência de conhecimento
  - credibilidade no universo acadêmico
  - atração de talentos

atividades

- muitos(!) *blockers*
  em **experimentações** e **novos** projetos
  - disponibilidade dos dados
  - maturidade em *analytics*
  - recursos e time

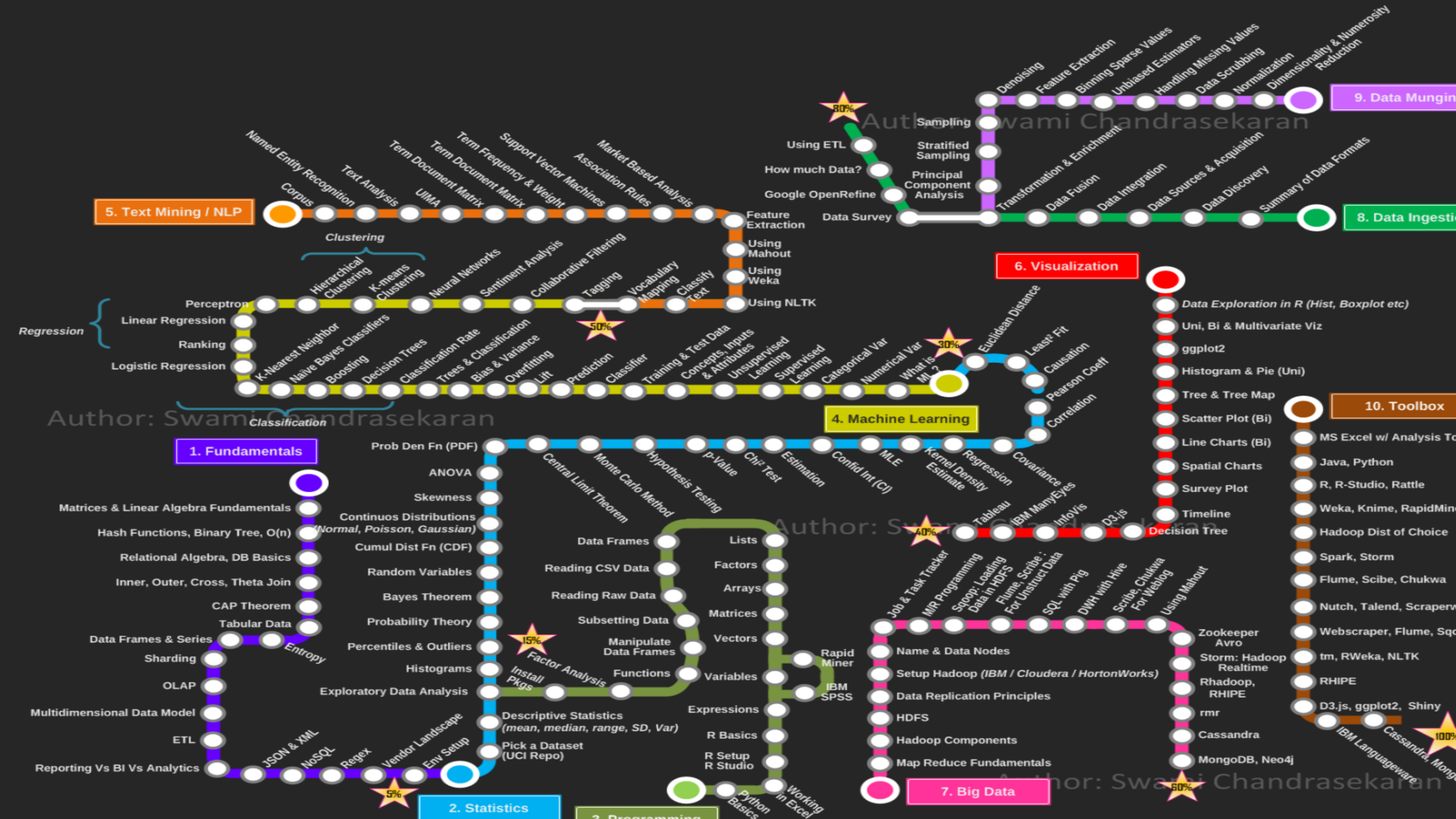
blockers


dados ruins
49.4%
falta de cientistas de dados
41.6%
falta de hipóteses claras
37.2%
dificuldade ou indisponibilidade de dados
30.4%
resultados não utilizados em decisões
30.2%
explicar ciência de dados p/ outros
24.3%
problemas de privacidade
22.0%
falta de input de especialistas
19.8%
não podem custear um time de DS
19.6%
muitos ambientes específicos
17.8%
limitações de ferramentas
17.5%
necessidade de coordenar com TI
16.5%
expetativas do impacto do projeto
16.3%
integrar descobertas nas decisões
15.8%

cientista de dados?

“ Sexiest job of the 21st century

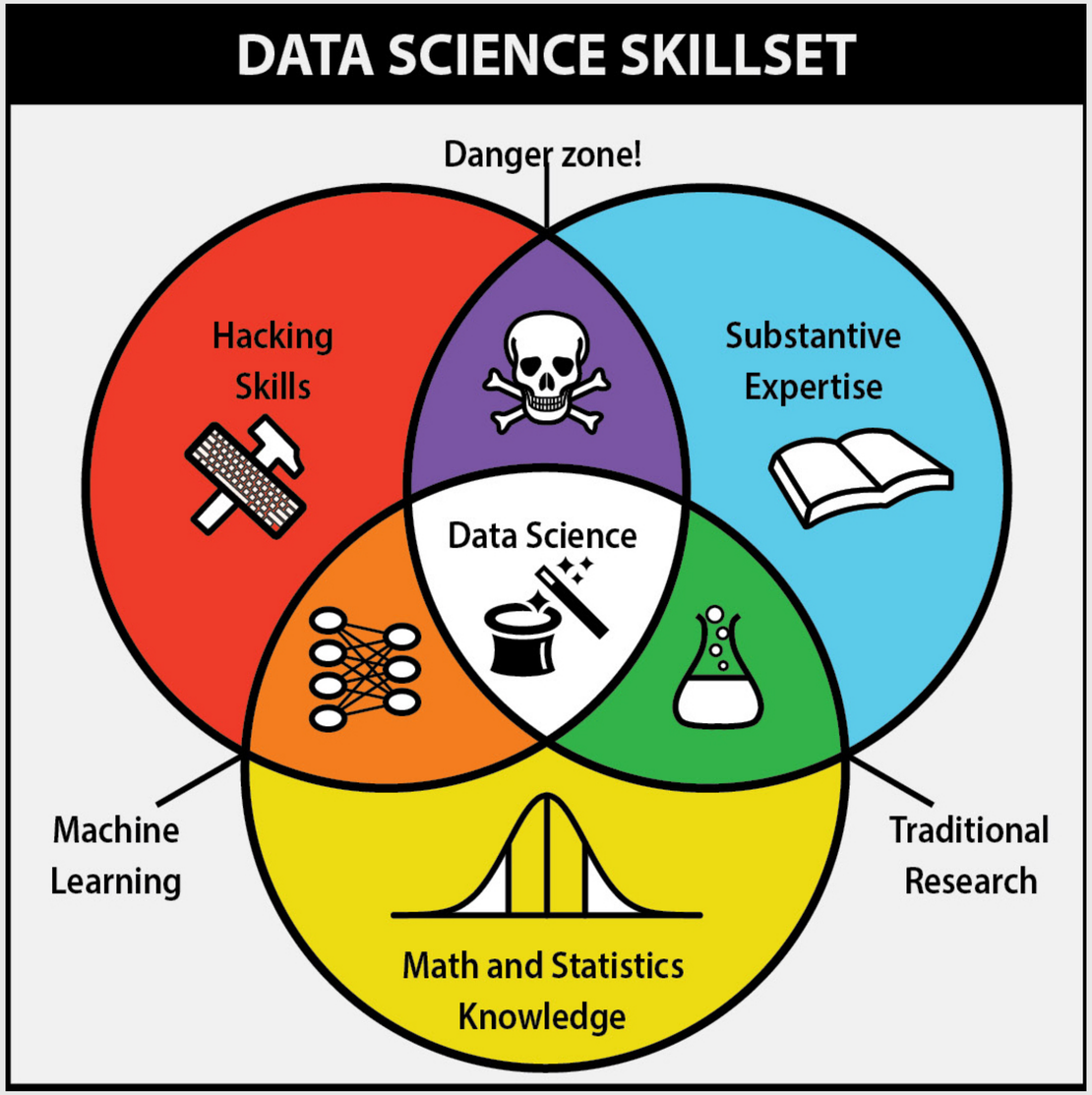
- Harvard Business Review


Author: Swami Chandrasekaran
1. Fundamentals
Matrices & Linear Algebra Fundamentals
Hash Functions, Binary Tree, O(n)
Relational Algebra, DB Basics
Inner, Outer, Cross, Theta Join
CAP Theorem
Tabular Data
Data Frames & Series
Entropy
Sharding
OLAP
Multidimensional Data Model
ETL
Reporting Vs BI Vs Analytics
JSON & XML
NoSQL
Regex
Vendor Landscape
Env Setup
5%
2. Statistics
Prob Den Fn (PDF)
Central Limit Theorem
Monte Carlo Method
Hypothesis Testing
p-Value
Chi² Test
Estimation
Confid Int (CI)
MLE
Kernel Density Estimate
Regression
Covariance
ANOVA
Skewness
Continuos Distributions
(Normal, Poisson, Gaussian)
Cumul Dist Fn (CDF)
Random Variables
Bayes Theorem
Probability Theory
Percentiles & Outliers
Histograms
Exploratory Data Analysis
15%
Factor Analysis
Install Pkgs
Descriptive Statistics
(mean, median, range, SD, Var)
Pick a Dataset
(UCI Repo)
3. Programming
Data Frames
Reading CSV Data
Reading Raw Data
Subsetting Data
Manipulate Data Frames
Functions
Lists
Factors
Arrays
Matrices
Vectors
Variables
Expressions
R Basics
R Setup
R Studio
Rapid Miner
IBM SPSS
Python Basics
Working in Excel
4. Machine Learning
Perceptron
Linear Regression
Ranking
Logistic Regression
Regression
K-Nearest Neighbor
Naive Bayes Classifiers
Boosting
Decision Trees
Classification Rate
Trees & Classification
Bias & Variance
Overfitting
Lift
Prediction
Classifier
Training & Test Data
Concepts, Inputs & Attributes
Unsupervised Learning
Supervised Learning
Categorical Var
Numerical Var
What is ML?
30%
Euclidean Distance
Least² Fit
Causation
Pearson Coeff
Correlation
Classification
Hierarchical Clustering
K-means Clustering
Clustering
Neural Networks
Sentiment Analysis
Collaborative Filtering
Tagging
Vocabulary Mapping
Classify Text
50%
5. Text Mining / NLP
Corpus
Text Analysis
UIMA
Term Document Matrix
Term Document Matrix
Term Frequency & Weight
Support Vector Machines
Association Rules
Market Based Analysis
Named Entity Recognition
Feature Extraction
Using Mahout
Using Weka
Using NLTK
6. Visualization
Data Exploration in R (Hist, Boxplot etc)
Uni, Bi & Multivariate Viz
ggplot2
Histogram & Pie (Uni)
Tree & Tree Map
Scatter Plot (Bi)
Line Charts (Bi)
Spatial Charts
Survey Plot
Timeline
Decision Tree
Tableau
IBM ManyEyes
InfoVis
D3.js
40%
7. Big Data
Job & Task Tracker
MR Programming
Sqoop: Loading Data in HDFS
Flume, Scribe : For Unstruct Data
SQL with Pig
DWH with Hive
Scribe, Chukwa For Weblog
Using Mahout
Name & Data Nodes
Setup Hadoop (IBM / Cloudera / HortonWorks)
Data Replication Principles
HDFS
Hadoop Components
Map Reduce Fundamentals
Zookeeper Avro
Storm: Hadoop Realtime
Rhadoop, RHIPE
rmr
Cassandra
MongoDB, Neo4j
60%
8. Data Ingestion
80%
Using ETL
How much Data?
Google OpenRefine
Data Survey
Sampling
Stratified Sampling
Principal Component Analysis
Transformation & Enrichment
Denoising
Feature Extraction
Binning Sparse Values
Unbiased Estimators
Handling Missing Values
Data Scrubbing
Normalization
Dimensionality & Numerosity Reduction
Data Fusion
Data Integration
Data Sources & Acquisition
Data Discovery
Summary of Data Formats
9. Data Munging
10. Toolbox
MS Excel w/ Analysis To
Java, Python
R, R-Studio, Rattle
Weka, Knime, RapidMin
Hadoop Dist of Choice
Spark, Storm
Flume, Scibe, Chukwa
Nutch, Talend, Scraperw
Webscraper, Flume, Sqo
tm, RWeka, NLTK
RHIPE
D3.js, ggplot2, Shiny
IBM Languageware
Cassandra, Mongo
100%

cientista de dados

VOCÊ

skills


DATA SCIENCE SKILLSET
Danger zone!
Hacking Skills
Substantive Expertise
Data Science
Machine Learning
Traditional Research
Math and Statistics Knowledge

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**

skills

pesquisa tradicional
método científico

inteligência artificial
conhecimento **matemático + computacional + dev**

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

skills

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

- 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

- 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

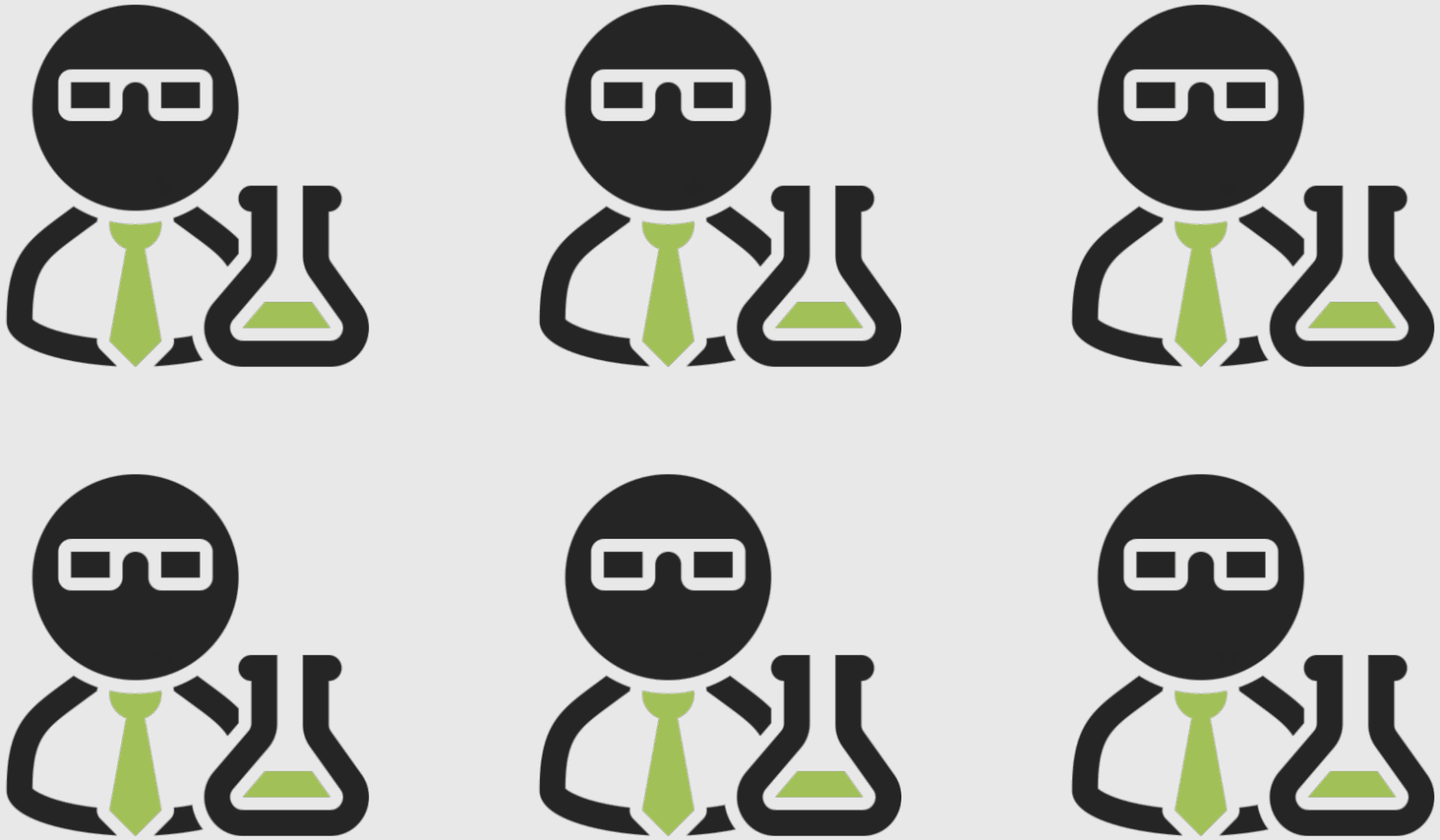
- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- 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

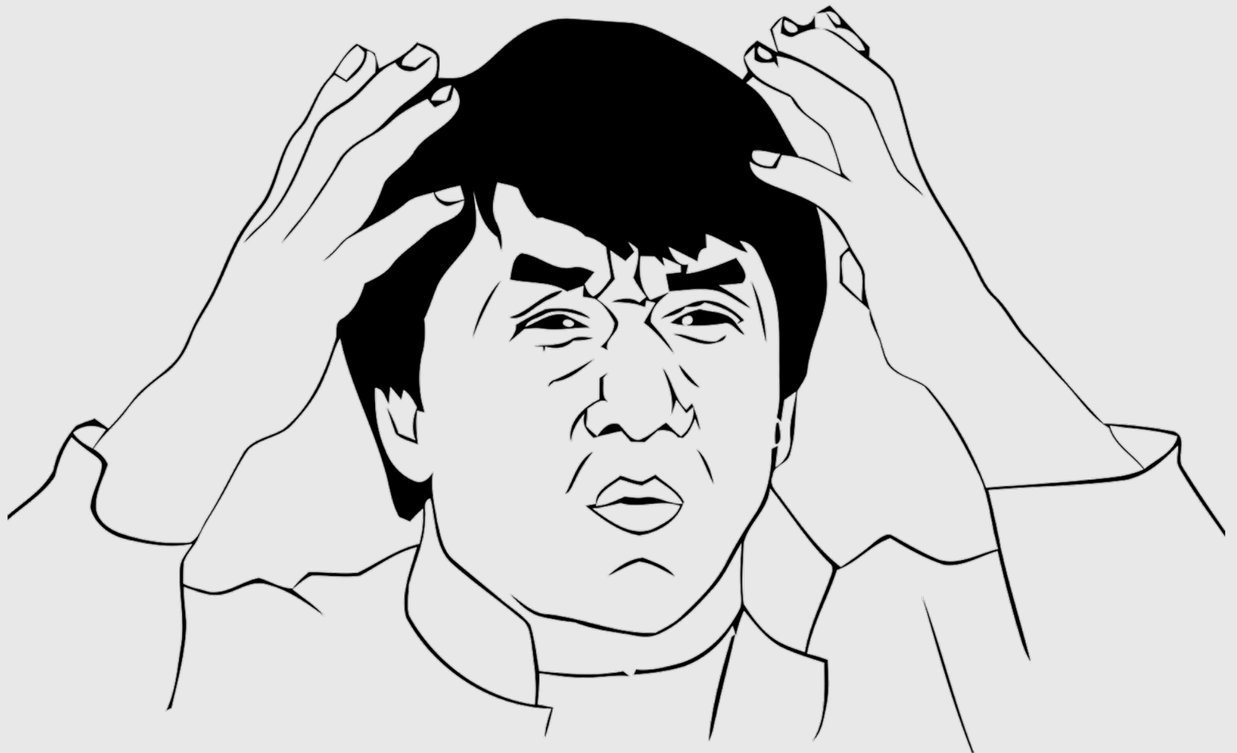
time?

time



papéis

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



papéis

Data Scientist vs Data ~~Engineer~~ Architect?!

papéis


visualização
modelagem
storytelling
estatística
programação
armazenamento
ingestão de dados
arquitetura
ferramentas de big data
DATA SCIENTIST
DATA ARCHITECT

time

1 inteligência artificial

infraestrutura (big data) 2

3 desenvolvimento

business 4

time

1 cientista de dados

arquiteto de dados 2

3 desenvolvedor

subject **m**atter **e**xpertise 4

papéis

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

mercado

- cientista de dados
  - 4,524 vagas abertas
  - salário médio de $110.000
- gerente de *analytics*
  - 1,381 vagas abertas
  - salário médio de $115.000

mercado

- engenheiro de software
  - 29.187 vagas abertas
  - salário médio de $102,500
- DBA's
  - 2,370 vagas abertas
  - salário médio de $94,000

mercado

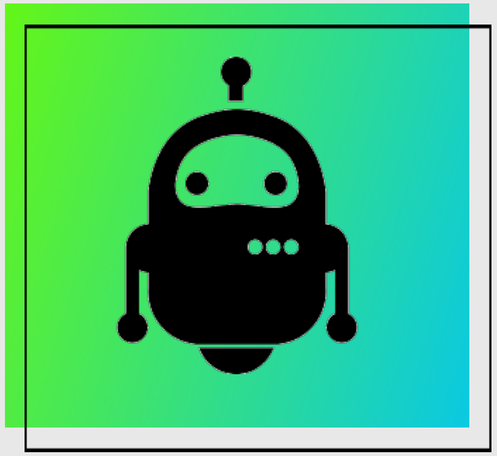
- engenheiro de software
  - 29.187 vagas abertas
  - salário médio de $102,500
- engenheiro de dados
  - 2,816 vagas abertas
  - salário médio de $100,000

mercado

- analista de dados
  - 4,729 vagas abertas
  - salário médio de $60,000
- DBA's
  - 2,370 vagas abertas
  - salário médio de $94,000

agenda

- definição
  - ~~buzZzZzword~~... Data Science?
- business
  - dados, mercado, previsões
- metodologia
  - ciclo de vida, atividades, papéis
- tecnologias
  - **players**, ferramentas, landscape, **diferenças**
- prática
  - the moment you've all been waiting for!



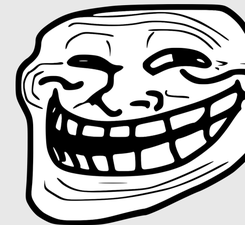
tecnologias

programação

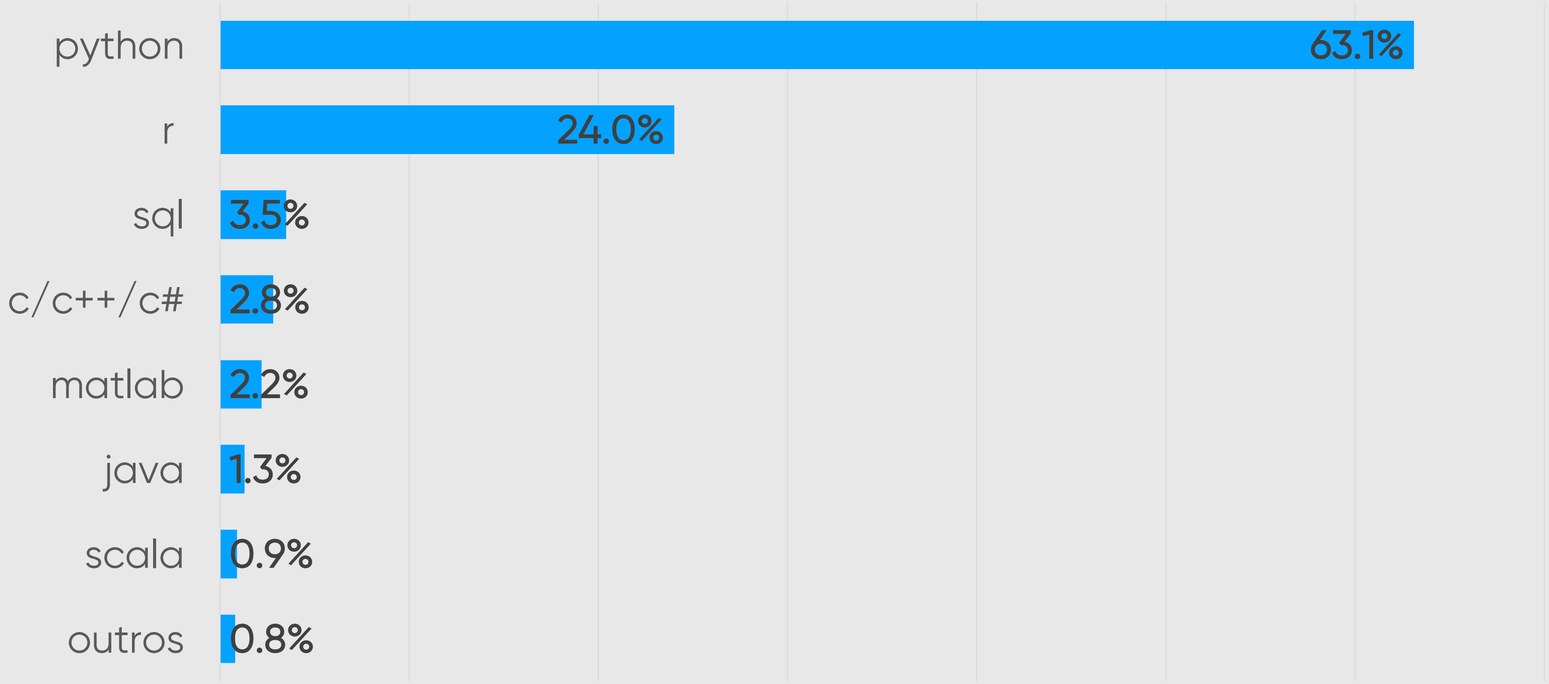
“

Nenhuma melhor ou pior

– Eu



programação



programação

- análise exploratória
- experimentação
- + produtividade

programação

- análise exploratória

- + produtividade

programação

- experimentação
- + produtividade

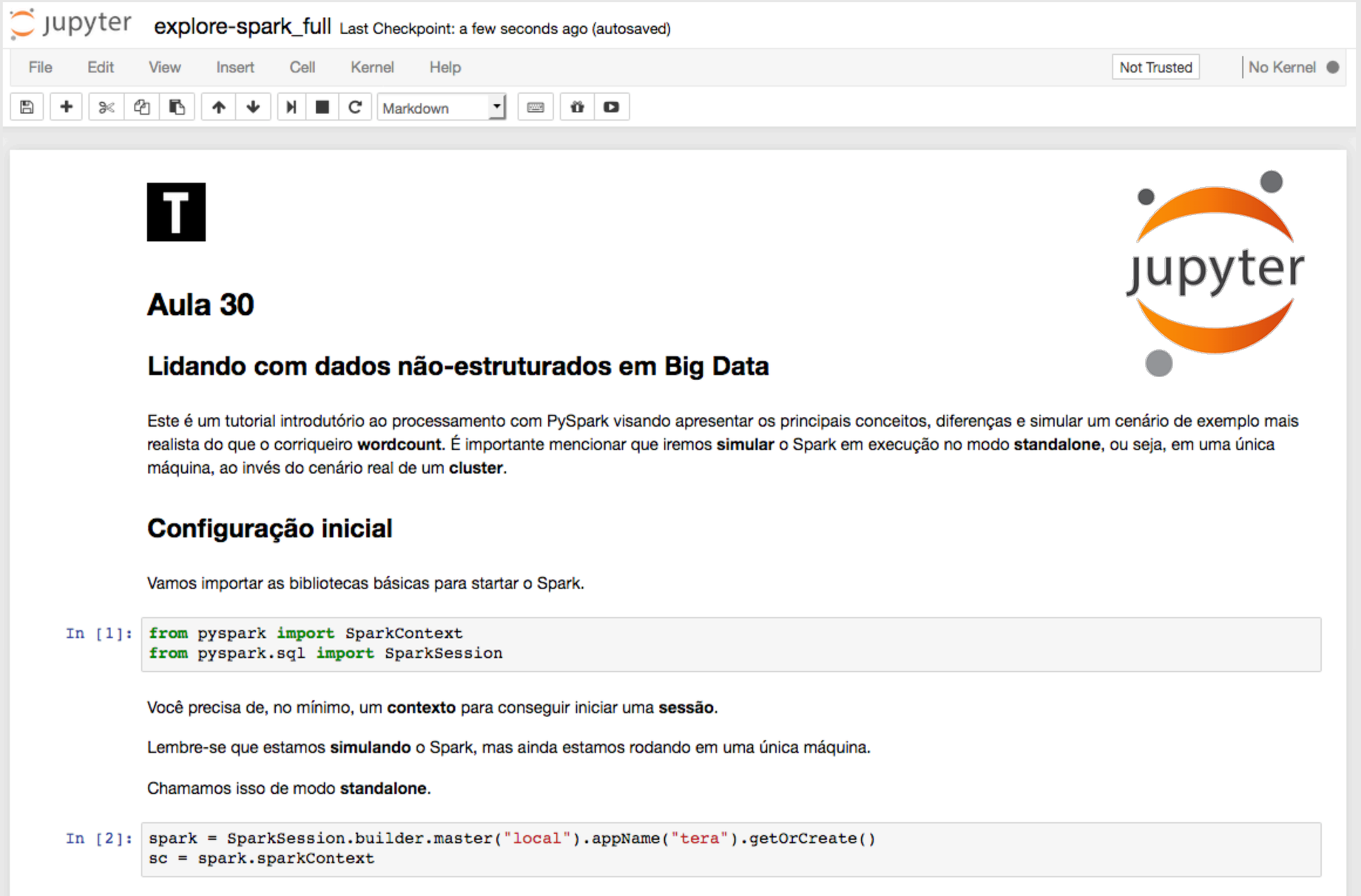
programação

- análise exploratória
- experimentação
- + produtividade
- + performance

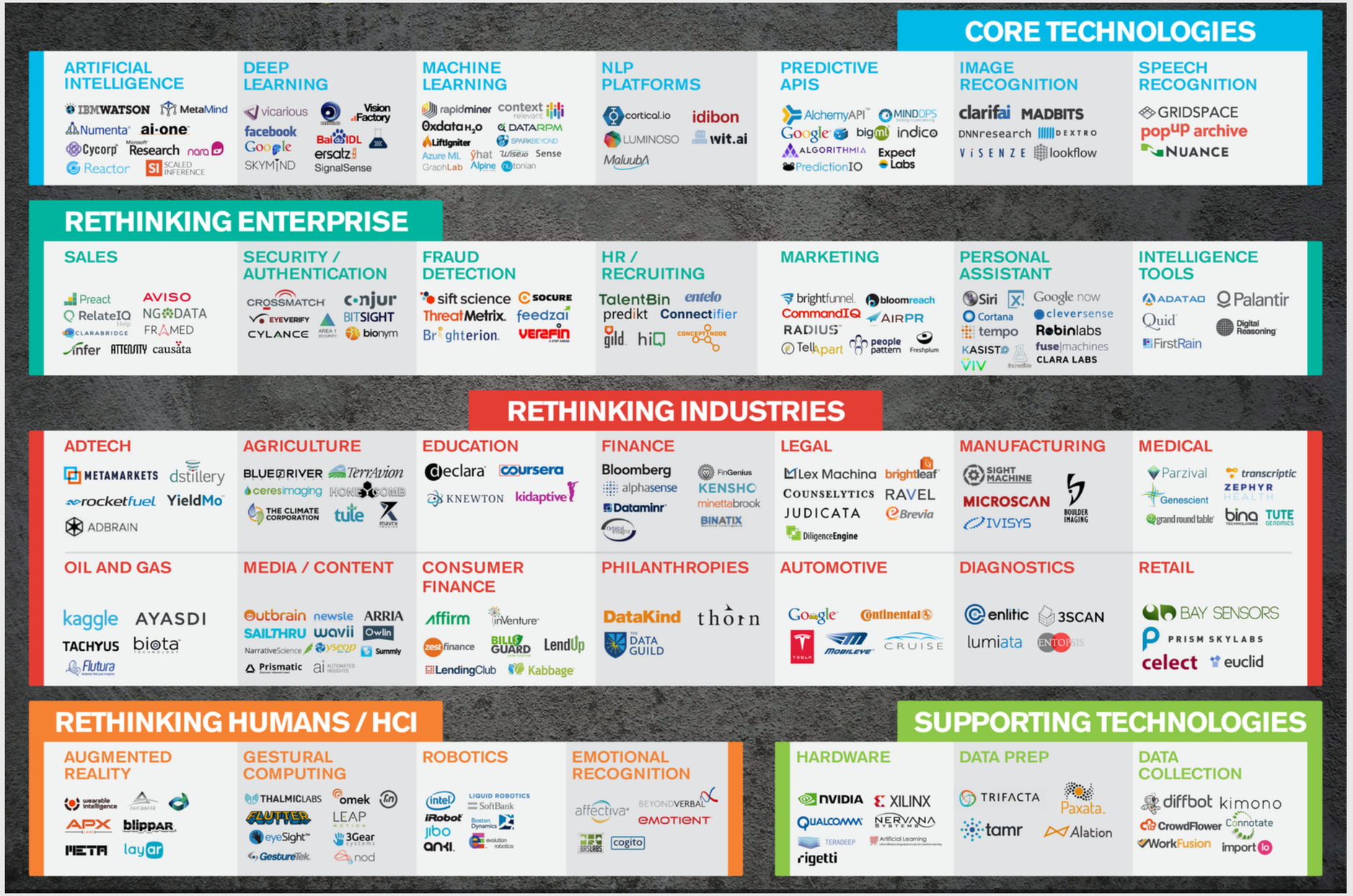
programação

- produção
- ambientes homologados

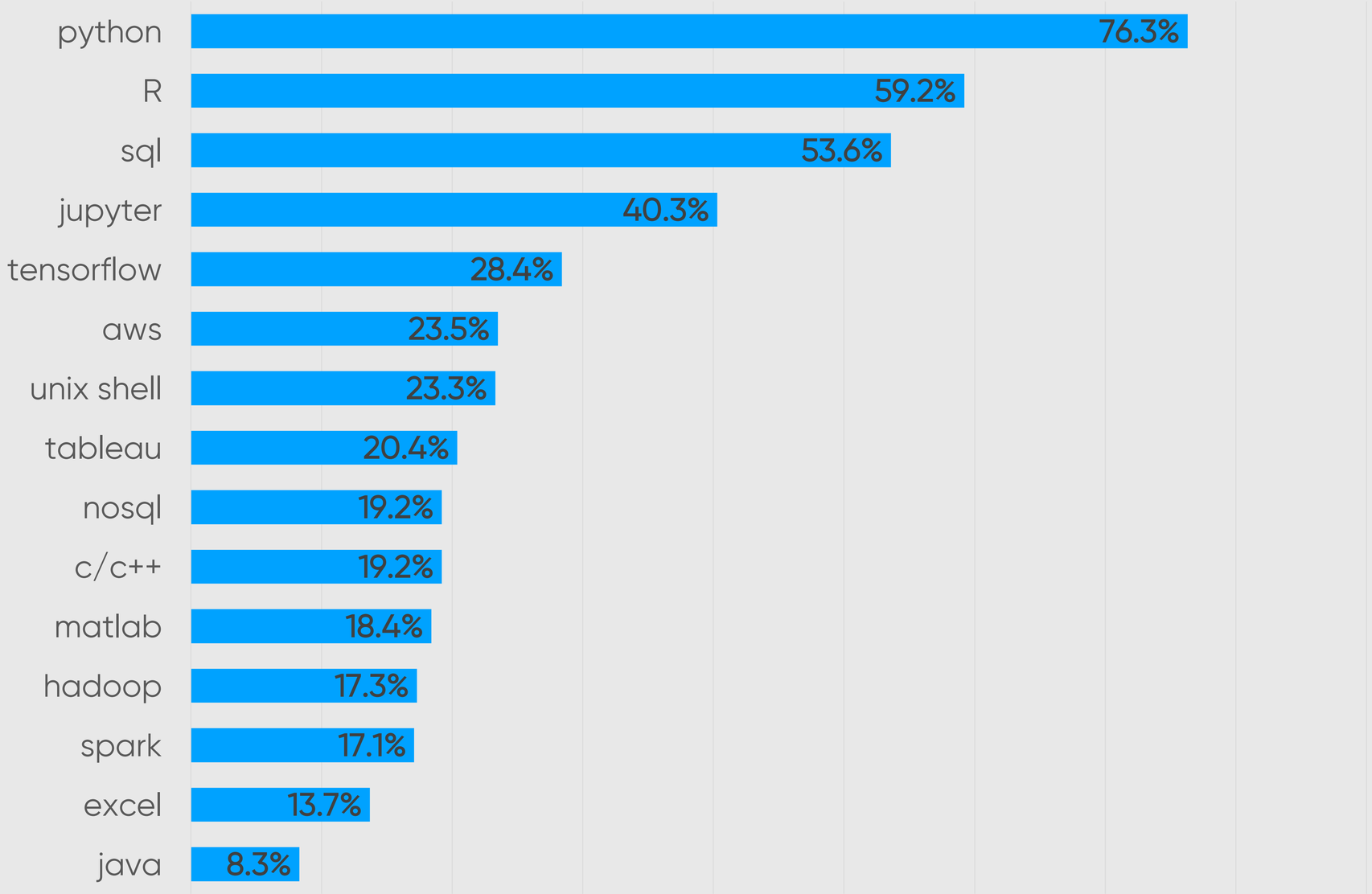
IDE e visualização



landscape


CORE TECHNOLOGIES
ARTIFICIAL INTELLIGENCE
IBM WATSON
MetaMind
Numenta
ai·one
Cycorp
Microsoft Research
nara
Reactor
SCALED INFERENCE
DEEP LEARNING
vicarious
Vision Factory
facebook
Google
Baidu IDL
ersatz
SKYMIND
SignalSense
MACHINE LEARNING
rapidminer
context relevant
0xdata H2O
DATARPM
LIftIgniter
SPARKBEYOND
Azure ML
yhat
wise.io
Sense
GraphLab
Alpine
tonian
NLP PLATFORMS
cortical.io
idibon
LUMINOSO
wit.ai
MaluubA
PREDICTIVE APIS
AlchemyAPI
MINDOPS
Google
bigml
indico
ALGORITHMIA
Expect Labs
PredictionIO
IMAGE RECOGNITION
clarifai
MADBITS
DNNresearch
DEXTRO
VISENZE
lookflow
SPEECH RECOGNITION
GRIDSPACE
popup archive
NUANCE
RETHINKING ENTERPRISE
SALES
Preact
AVISO
RelateIQ
NGDATA
CLARABRIDGE
FRAMED
infer
ATTENSITY
causata
SECURITY / AUTHENTICATION
CROSSMATCH
conjur
EYEVERIFY
BITSIGHT
CYLANCE
AREA 1 SECURITY
bionym
FRAUD DETECTION
sift science
SOCURE
ThreatMetrix
feedzai
Brighterion
VERAFIN
HR / RECRUITING
TalentBin
entelo
predikt
Connectifier
gild
hiQ
CONCEPTNODE
MARKETING
brightfunnel
bloomreach
CommandIQ
AIRPR
RADIUS
TellApart
people pattern
Freshplum
PERSONAL ASSISTANT
Siri
Google now
Cortana
cleversense
tempo
Robinlabs
KASISTO
fuse|machines
VIV
CLARA LABS
INTELLIGENCE TOOLS
ADATAO
Palantir
Quid
Digital Reasoning
FirstRain
RETHINKING INDUSTRIES
ADTECH
METAMARKETS
dstillery
rocketfuel
YieldMo
ADBRAIN
AGRICULTURE
BLUE RIVER
TerrAvion
ceresimaging
HONEYCOMB
THE CLIMATE CORPORATION
tule
EDUCATION
declara
coursera
KNEWTON
kidaptive
FINANCE
Bloomberg
FinGenius
alphasense
KENSHO
Dataminr
minettabrook
BINATIX
LEGAL
Lex Machina
brightleaf
COUNSELYTICS
RAVEL
JUDICATA
Brevia
DiligenceEngine
MANUFACTURING
SIGHT MACHINE
MICROSCAN
IVISYS
BOULDER IMAGING
MEDICAL
Parzival
transcriptic
ZEPHYR HEALTH
Genescient
grand round table
bina
TUTE GENOMICS
OIL AND GAS
kaggle
AYASDI
TACHYUS
biota
Flutura
MEDIA / CONTENT
Outbrain
newsle
ARRIA
SAILTHRU
wavii
Owlin
NarrativeScience
Summly
Prismatic
ai
CONSUMER FINANCE
Affirm
inVenture
zestfinance
BILLGUARD
LendUp
LendingClub
Kabbage
PHILANTHROPIES
DataKind
thorn
THE DATA GUILD
AUTOMOTIVE
Google
Continental
TESLA
MobileEye
CRUISE
DIAGNOSTICS
enlitic
3SCAN
lumiata
ENTOPSIS
RETAIL
BAY SENSORS
PRISM SKYLABS
celect
euclid
RETHINKING HUMANS / HCI
AUGMENTED REALITY
wearable intelligence
APX
blippAR
META
layar
GESTURAL COMPUTING
THALMICLABS
omek
FLUTTER
LEAP MOTION
eyeSight
3Gear Systems
GestureTek
nod
ROBOTICS
intel
LIQUID ROBOTICS
SoftBank
iRobot
Boston Dynamics
jibo
evolution robotics
anki
EMOTIONAL RECOGNITION
affectiva
BEYONDVERBAL
EMOTIENT
BRIGHTLABS
cogito
SUPPORTING TECHNOLOGIES
HARDWARE
NVIDIA
XILINX
QUALCOMM
NERVANA SYSTEMS
TERADEEP
Artificial Learning
rigetti
DATA PREP
TRIFACTA
Paxata
tamr
Alation
DATA COLLECTION
diffbot
kimono
CrowdFlower
Connotate
WorkFusion
import.io

ferramentas



open-source vs proprietária?!

machine learning open-source

nlp open-source

gensim

openNLP

UIMA

NLTK

John Snow LABS

deep learning open-source

players

players

Google Cloud Platform

IBM Bluemix™

Microsoft Azure

players

amazon alexa

ASSISTANT

IBM Watson

Cortana

agenda

definição

~~buzZzZzword~~... Data Science?

business

dados, mercado, previsões

metodologia

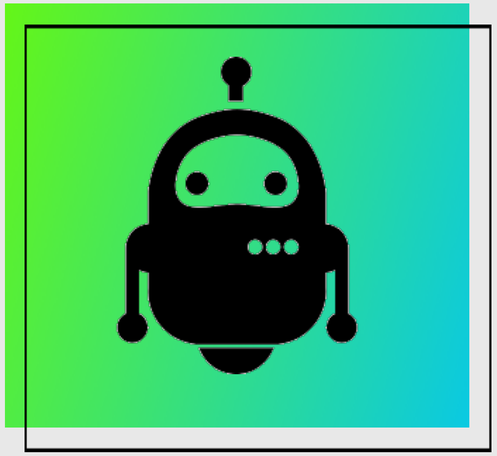
ciclo de vida, atividades, papéis

tecnologias

players, ferramentas,
landscape, **diferenças**

prática

the moment you've all been waiting for!

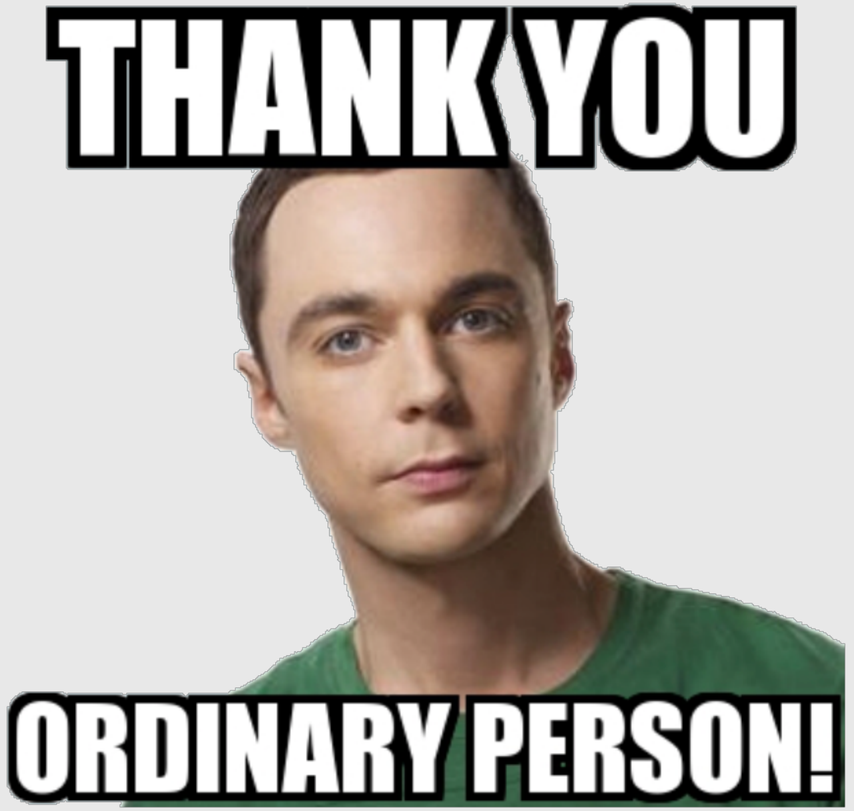


It’s time!

The moment you’all been waiting for...

Fire in the hole!

dúvidas?


THANK YOU
ORDINARY PERSON!

referências

- IA estatísticas:
  - previsão 2018
  - análise do mercado
  - cheat sheet
  - infográficos
- landscape de ferramentas
- kaggle report 2017
- toolkit
- magic quadrant (gartner)
- livros

extra!

- datasets **públicos**:
  - aws
  - kaggle
  - git
  - google
  - uci