



Data Science





Felipe Santana

Cientista de Dados - Prodemge - GSI





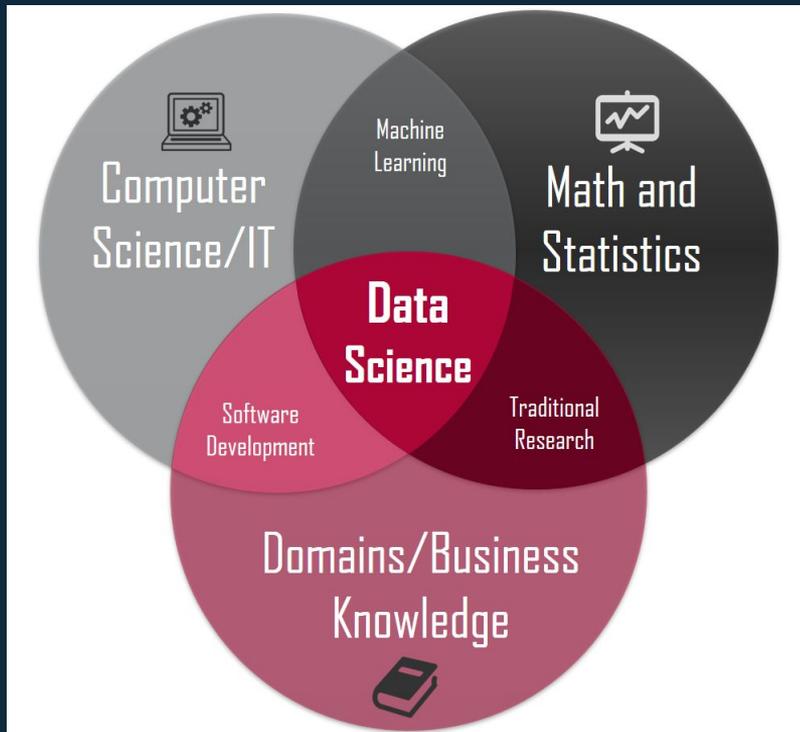
1

O que é Data
Science ?



Ciência que tem como objetivo manipular, analisar e extrair valor dos dados..

Disciplinas



Aplicações



O SONHO DE LAURA

Sepsé é uma resposta desregulada e exagerada do seu sistema imunológico a uma infecção, que provoca uma disfunção orgânica. Estima-se que haja 2 milhões e meio de casos de Sepsé por ano no Brasil e, desse número, cerca de 250 mil pessoas morrem anualmente. O Sonho de Laura tem como objetivo, reduzir estes números, salvando vidas.

[SAIBA MAIS](#)

Aplicações

stratsphera



Aplicações

CEAP
COTA PARA EXERCÍCIO DA
ATIVIDADE PARLAMENTAR

Até 2015, com 100 milhões de reais, além do salário e benefícios, tem direito a um fundo de até R\$ 600.000 por mês com atividades parlamentares. Desde 2016, com o novo teto de gastos e com o novo teto de gastos, o fundo foi reduzido para R\$ 128 milhões.

128 MILHÕES DE REAIS

REEMBOLSOS
O que já pagamos para nossos deputados?

R\$ 6.205,00
Em 2015, tivemos 117 pedidos de reembolso e **UMA REFEIÇÃO**

30 TANQUES DE GASOLINA COMPLETOS
Em 2015, tivemos 13 pedidos de reembolso e **R\$ 6.000,00** em gastos com gasolina, no valor de 13 tanques por mês.

13
Os deputados já pediram reembolso de 13 refeições durante no mesmo dia.

169.241
Em 2015, tivemos 169.241 pedidos de reembolso.

2 DEPUTADOS
Construíram o maior O VALDE MÁXIMO PEREIRA FIDELIS MENSALMENTE

R\$ 122,00
Um deputado já foi reembolsado por **BEBIDA ALCOÓLICA EN LAS VEGAS**



Tudo isso encontrado por uma inteligência artificial.



Perfil do profissional

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21st 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

DOMAIN KNOWLEDGE & SOFT SKILLS

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



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
- ☆ Map/Reduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

COMMUNICATION & VISUALIZATION

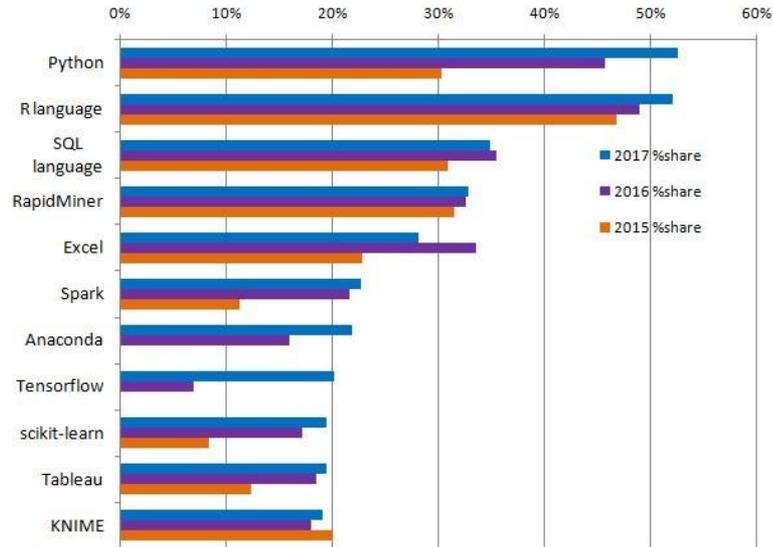
- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

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

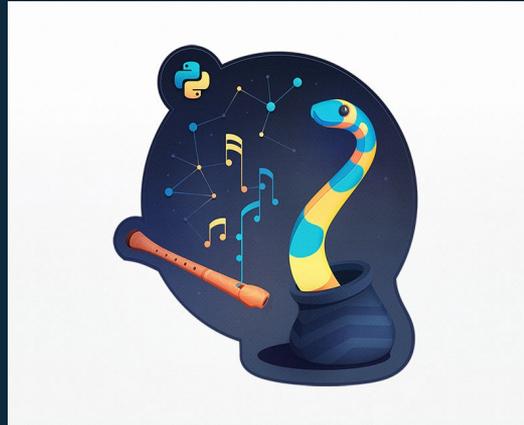
Tecnologias

KDnuggets Analytics, Data Science, Machine Learning Software Poll, top tools share, 2015-2017

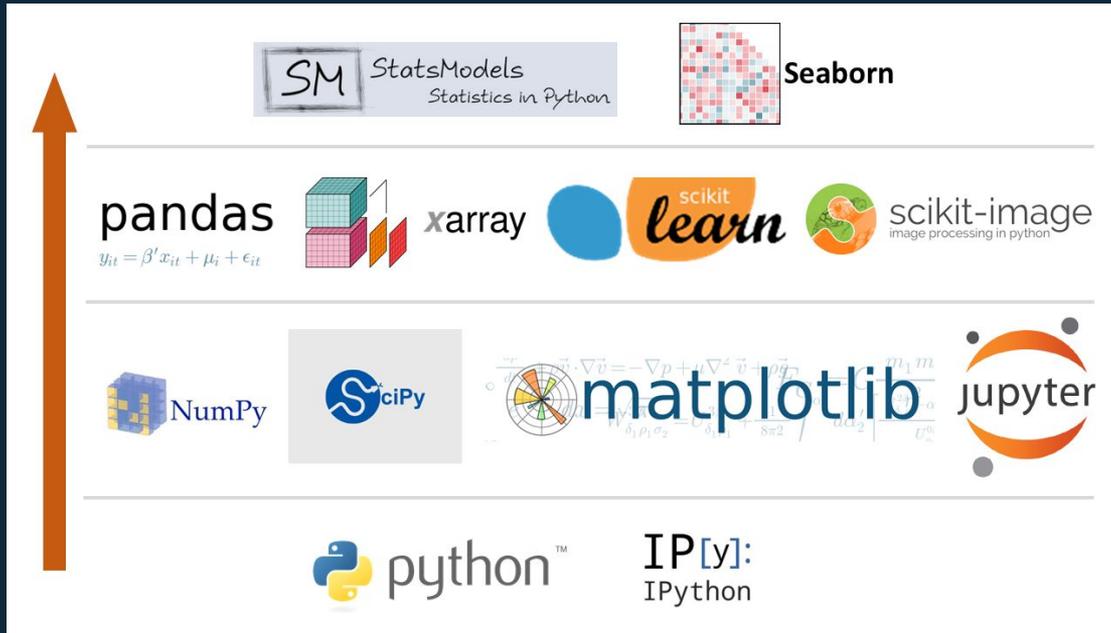


Tecnologias

- Python
 - Linguagem de alto nível
 - Interpretada e Orientada a objetos
 - Tipagem dinâmica
 - Desenvolvimento aberto pela Python Software Foundation



Tecnologias



Tecnologias

- Plataforma open source.
- Bibliotecas e Ferramentas integradas.
- Suporte **Python e R**
- Gestão de ambientes e pacotes.
- Cerca de **1400 bibliotecas já disponíveis.**
- Ide's já disponíveis: **Jupyter Notebook, Jupyter Lab, Microsoft Visual Code, Spyder** etc.





Machine Learning

Supervisionado

Não Supervisionado

...



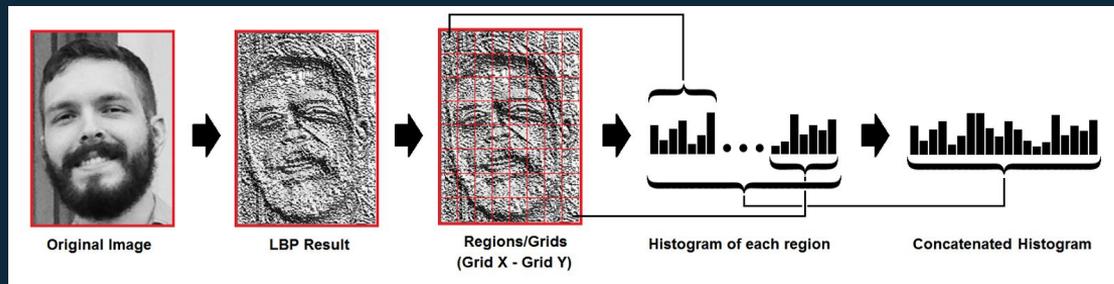
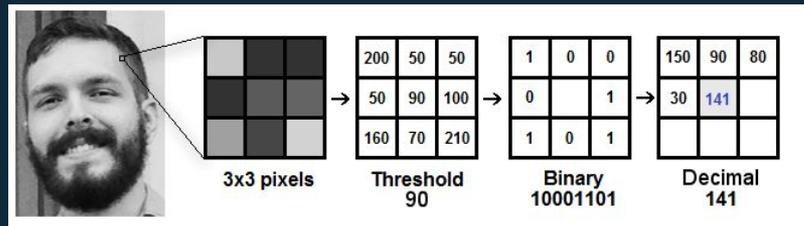


Machine Learning

- Subcampo da Ciência da Computação
- Resolve problemas complexos e com scala.

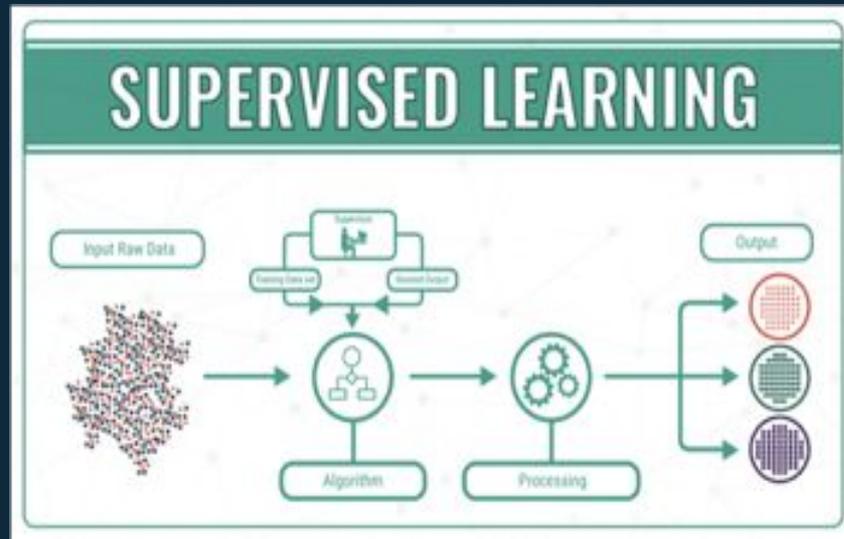


Machine Learning



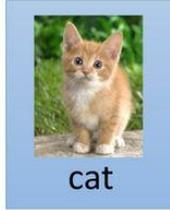
Como funciona?

- ◇ Algoritmo aprende com uma parte dos dados supervisionados.
- ◇ Normalmente a supervisão é feita por pessoas.
- ◇ A supervisão por pessoas é um processo caro.



Como funciona?

Labelled
data



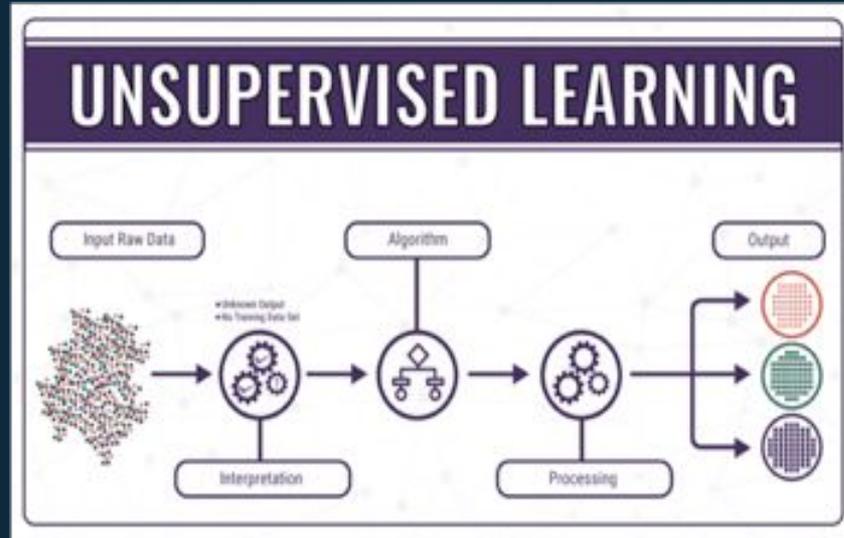
Unlabelled
data



(Image of cats and dogs without labeling)

Como funciona?

- ◆ Não há dados supervisionados para aprendizado.
- ◆ Busca por **similaridade** nos dados.
- ◆ Desvantagem é a forma de validar o modelo gerado.



Como funciona?





Quais as etapas?

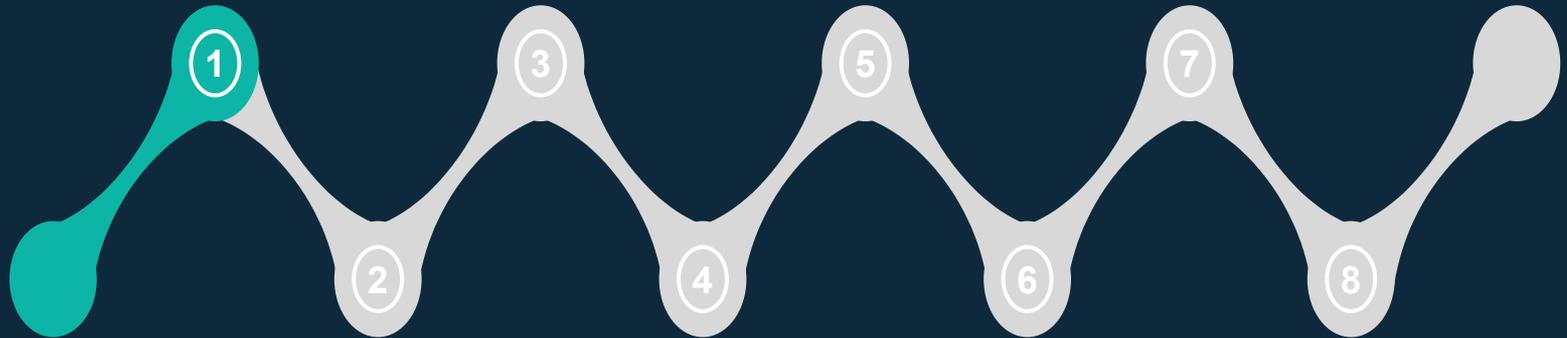
Fluxo de um projeto envolvendo Machine Learning





Análise do problema

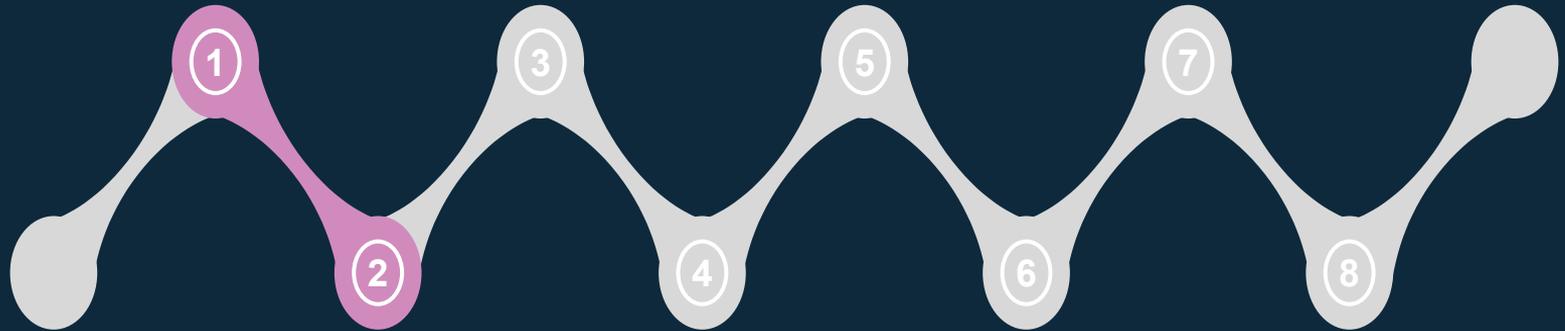
01 - Nessa etapa busca-se o entendimento do problema que se pretende a resolver





Coleta de Dados

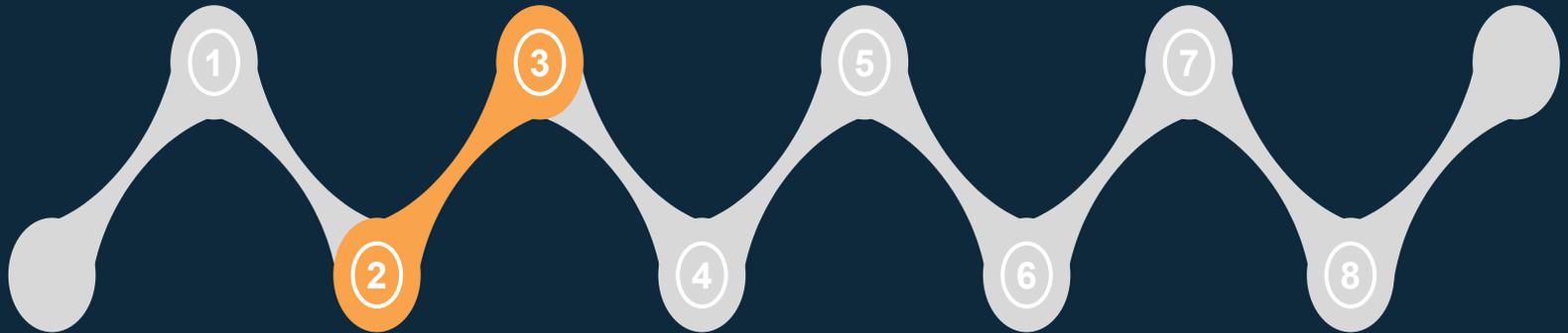
02 - A partir de uma ou mais fontes é feita a coleta e curadoria dos dados





Supervisão

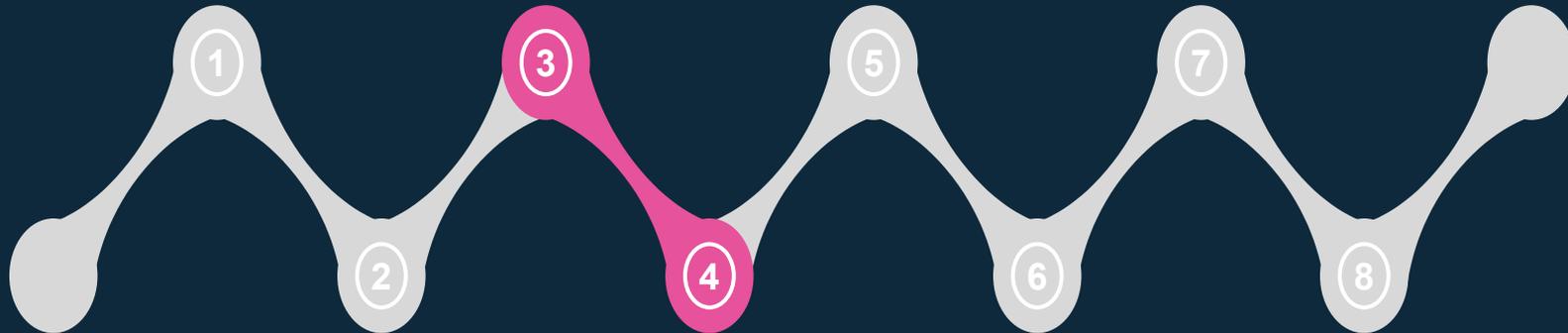
03 - É preciso supervisionar alguns dados de exemplos para treinar o algoritmo.





Engenharia de features

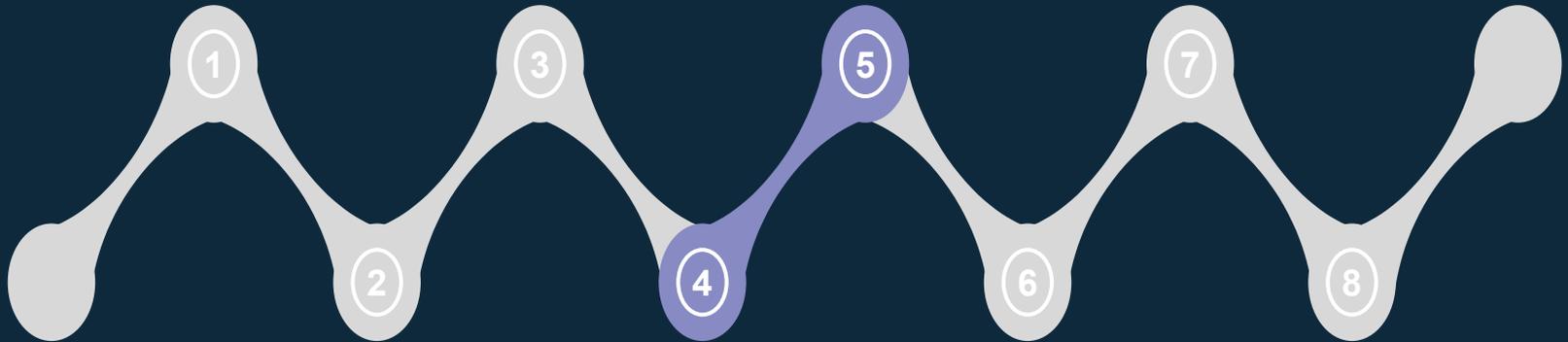
04 - Geração de features para o modelo a partir do conhecimento de domínio





Treinamento do Algoritmo

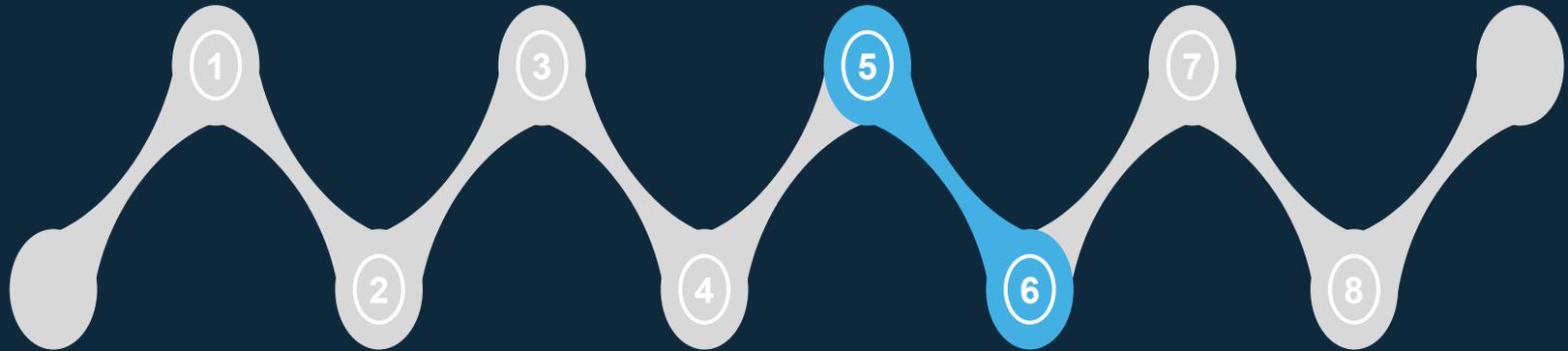
05 - Etapa onde é feito o treinamento do algoritmo e criação do modelo





Validação do modelo

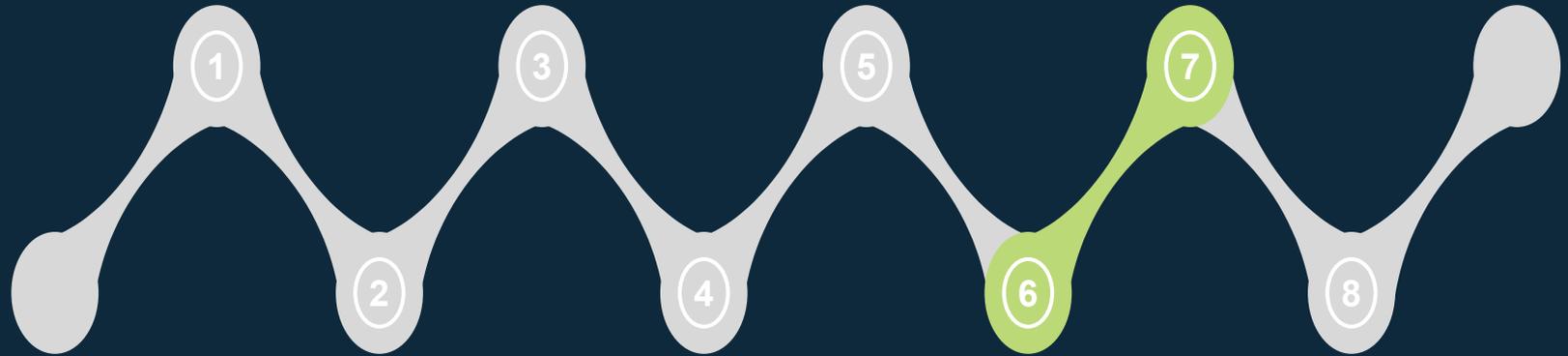
06 - Verificação de métricas e performance do modelo





Refaça os passos anteriores

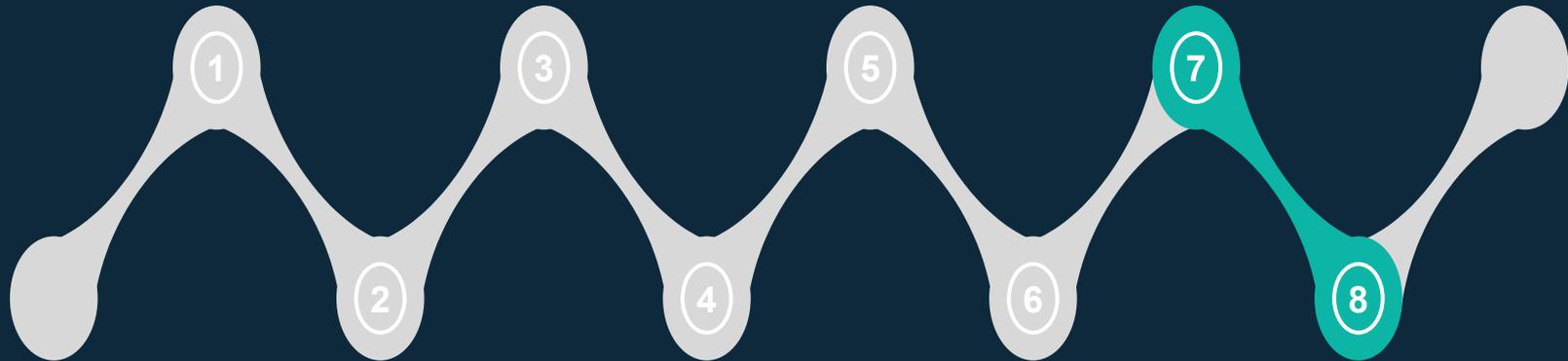
07 - Dependendo dos resultados é preciso voltar em passos anteriores





Deploy

08 - Implementação do modelo em produção





Aplicações interessantes ...

Projetos envolvendo Machine Learning,
Deep Learning e Visão Computacional



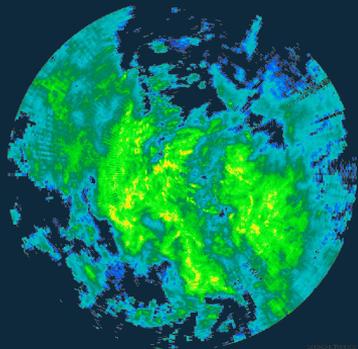


Visão Computacional

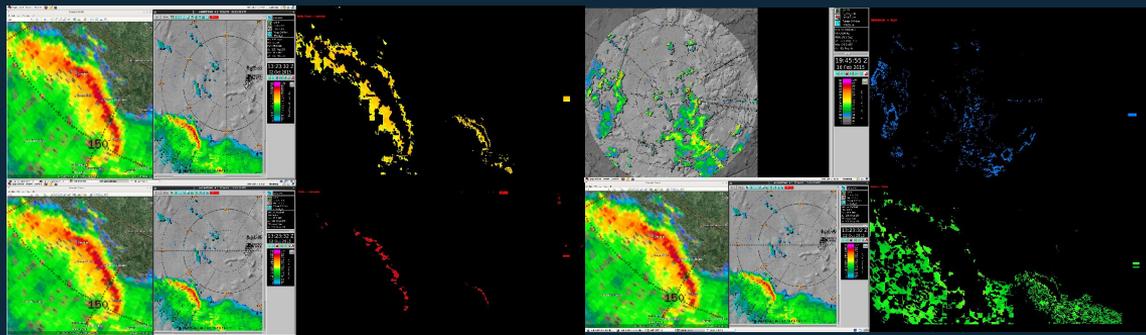
Monitoramento e detecção de eventos
como **Tempestades, Chuvas com
Granizo, Furacões, Inundações**



Visão Computacional



- ◇ Imagens de radar a cada 8 minutos.
- ◇ Emissão de **alertas personalizados**
- ◇ Armazenamento de dados e **predição de novos eventos.**



Tecnologias utilizadas



- ◇ Imagens de radar a cada 8 minutos.
- ◇ Emissão de **alertas personalizados**
- ◇ Armazenamento de dados e **predição de novos eventos.**





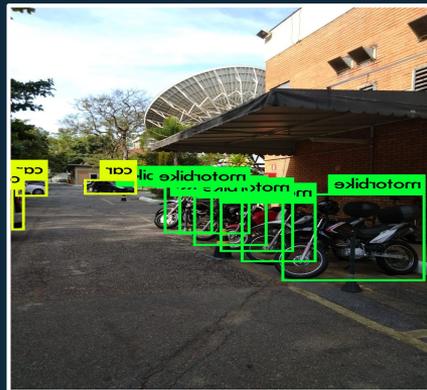
Detecção de objetos

Detecção de objetos por tipo, categoria..



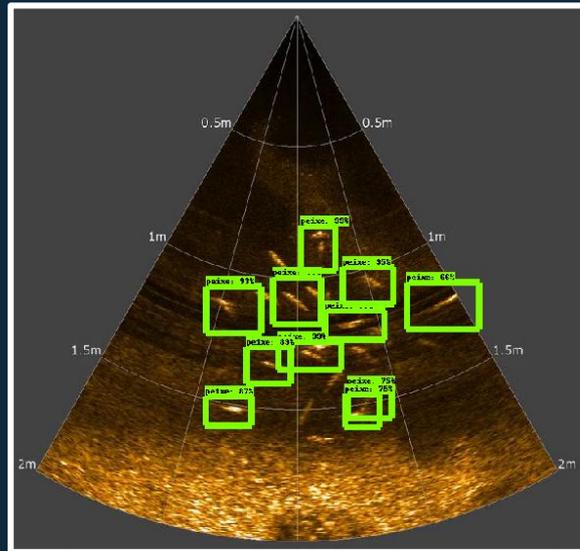
Detecção de pessoas e objetos

- ◇ Detecção de objetos como **carros**, **motos**, **caminhões**.
- ◇ Detecção de pessoas.



Detecção de pessoas e objetos

- ◇ Contagem e registros de objetos.





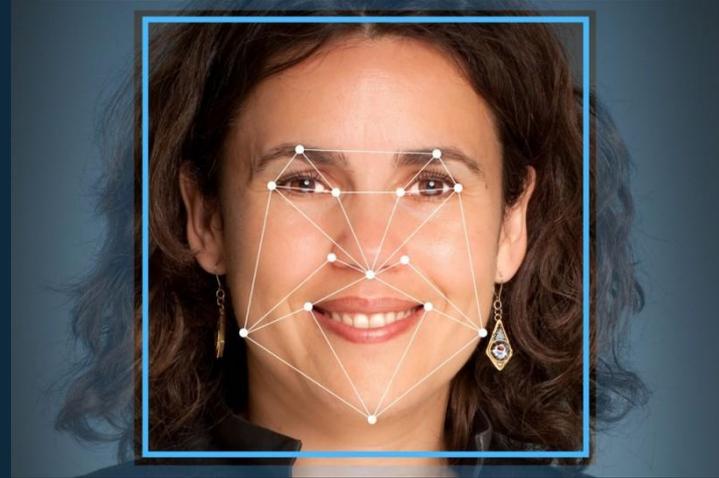
Reconhecimento Facial

Aplicação de Inteligência Artificial para a **identificação de pessoas.**



Reconhecimento Facial

- ◆ Identificação de pessoas e funcionários.





Vamos para
prática ?