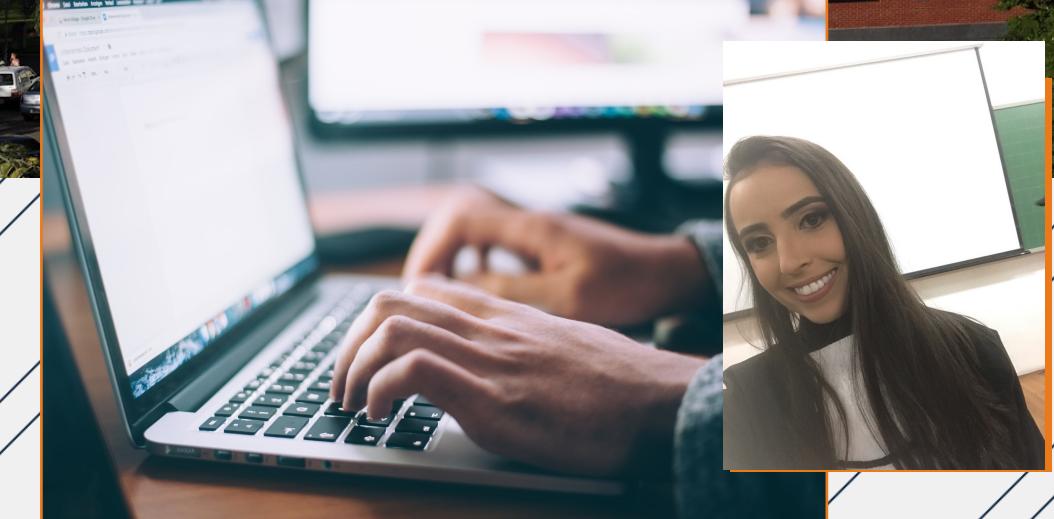


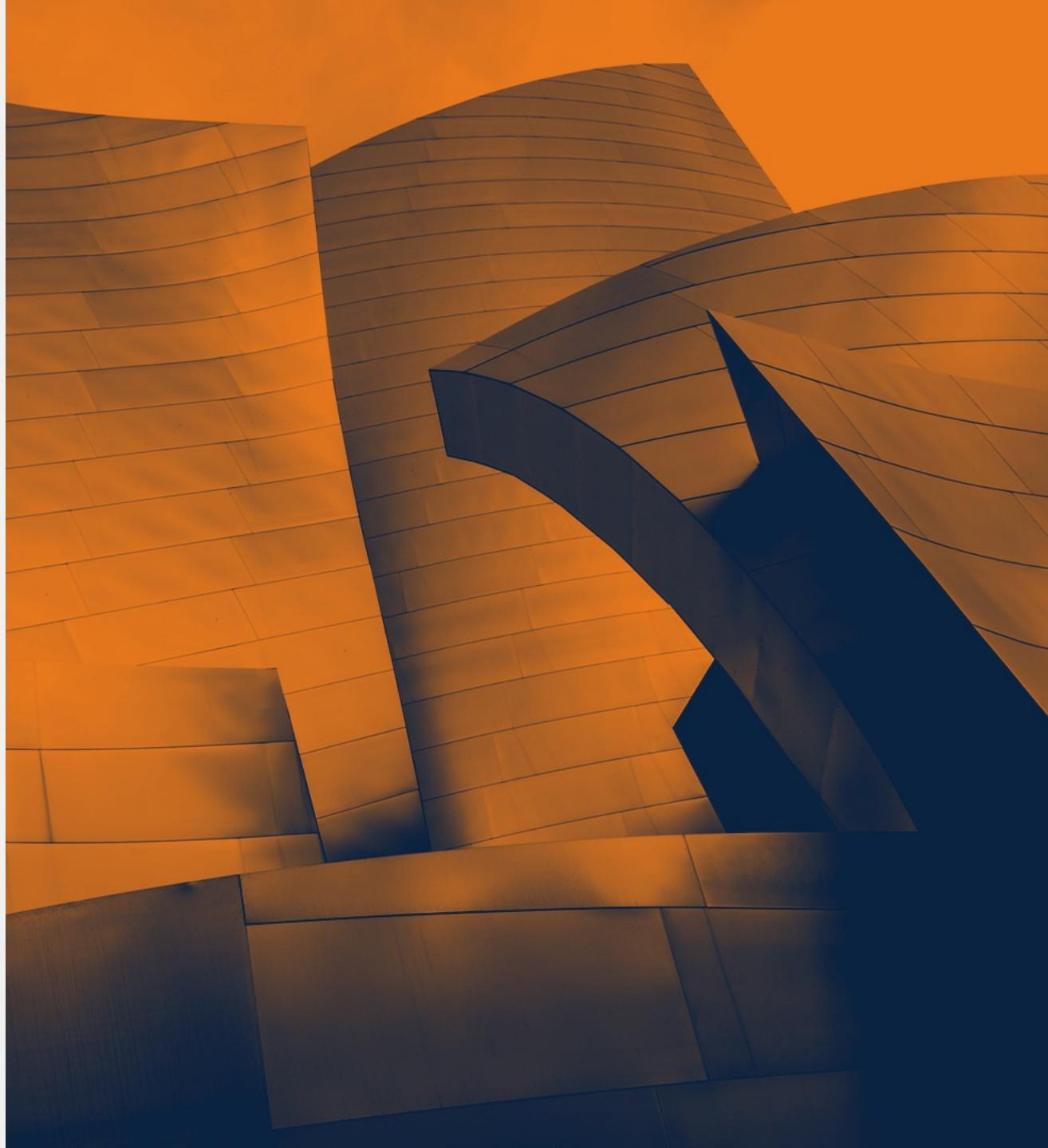
Introdução à Ciência de Dados

Nathália Campos

- Técnica de Eletrônica - ETE FMC
- Engenheira de Computação - INATEL
- Intercâmbio na Jade University - Alemanha
- Desenvolvedora de Software na 4Intelligence



-
- 1 | O que é Ciência de Dados
 - 2 | Como surgiu
 - 3 | Ciclo/Etapas
 - 4 | Tecnologias
 - 5 | Prática



O que é Ciência de Dados?

A ciência de dados combina vários campos com o intuito de extrair valor dos dados.

Os chamados “cientistas de dados” combinam uma variedade de habilidades para analisar dados coletados da web, smartphones, clientes, sensores e outras fontes para obter insights que tragam valor de alguma forma.

Estatística

Conjunto de técnicas para coletar, organizar, analisar e interpretar as informações de um problema em estudo para, assim, auxiliar na tomada de decisão.

Métodos Científicos

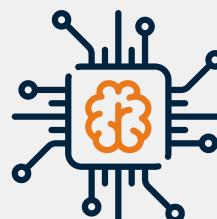
Metodologia usada por cientistas na busca de conhecimento.

Inteligência Artificial

É um ramo da ciência da computação que busca simular a inteligência humana em uma máquina. Os sistemas de IA são regidos por algoritmos usando técnicas como machine learning e deep learning para demonstrar comportamento “inteligente”.

Análise de Dados

A análise de dados é a arte de transformar dados em conhecimentos e insights relevantes.



Como surgiu?

O boom de informações

Com o avanço da tecnologia as pessoas começaram a ter acesso mais facilitado ao conhecimento, os telefones celulares evoluíram para smartphones. Surgindo cada vez mais informações e dados.

Surgimento da Big Data

O termo nasceu no início da década de 1990, na NASA, para descrever grandes conjuntos de dados complexos que desafiam os limites computacionais tradicionais de captura, processamento, análise e armazenamento informacional.

Necessidade

Com isso, houve a necessidade de analisar toda essa quantidade de dados que vinham surgindo com tanta rapidez.

90 %

Em 2013, a IBM compartilhou estatísticas mostrando que **90% dos dados do mundo** haviam sido criados nos **últimos dois anos**.

236 %

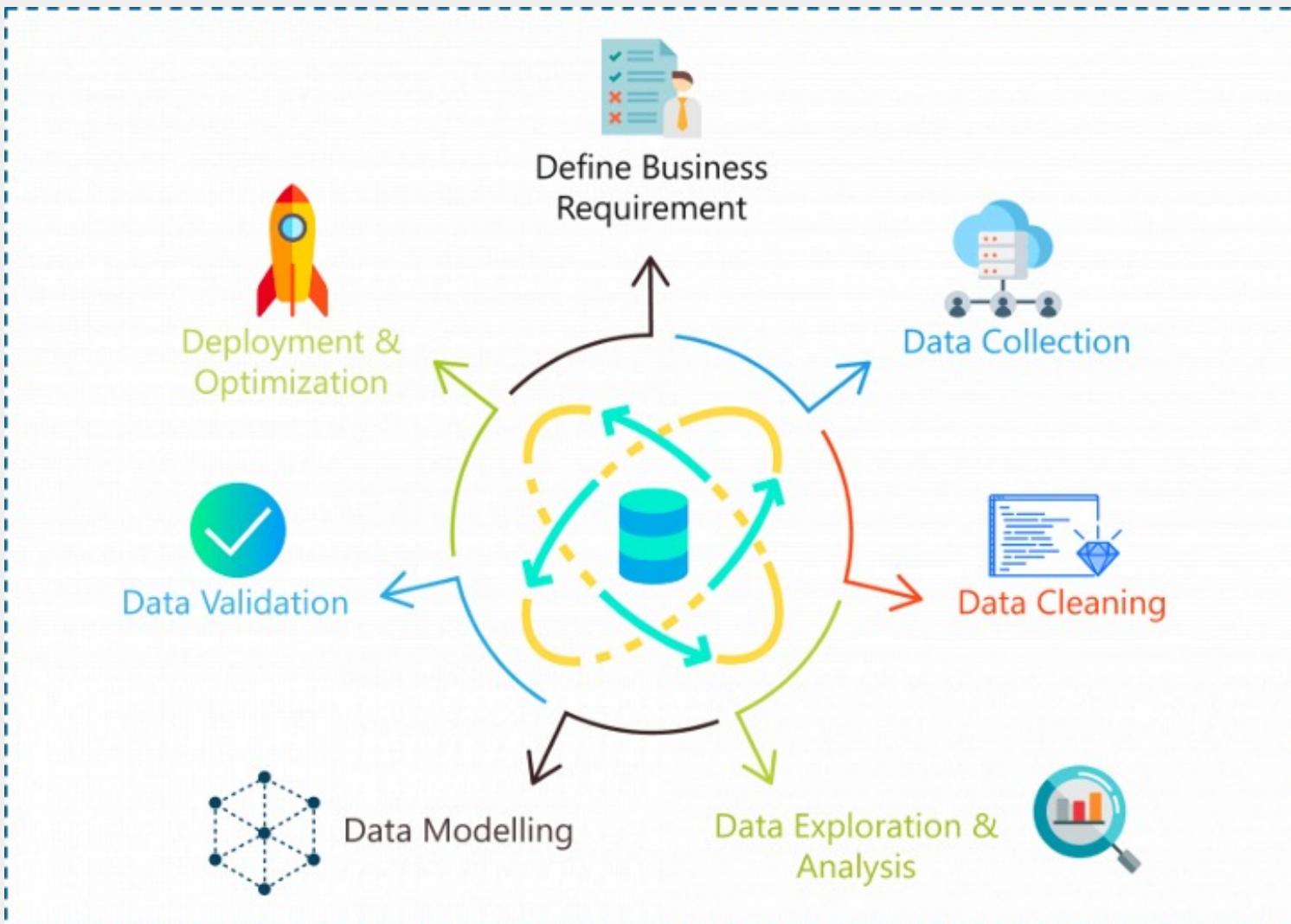
Segundo a Revelo, a **demanda por profissionais da área cresceu 236% em 2019**.



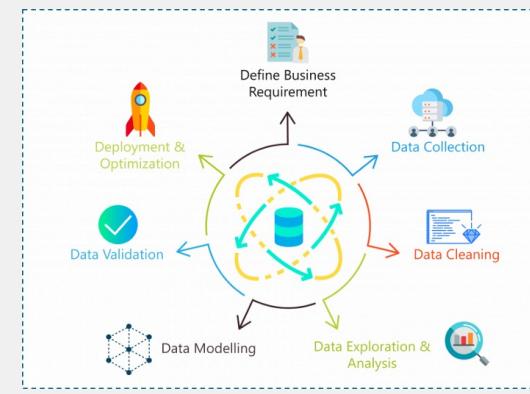


“Um profissional de alto escalão com treinamento e curiosidade para fazer descobertas no mundo do big data.”

O Ciclo da Ciência de Dados



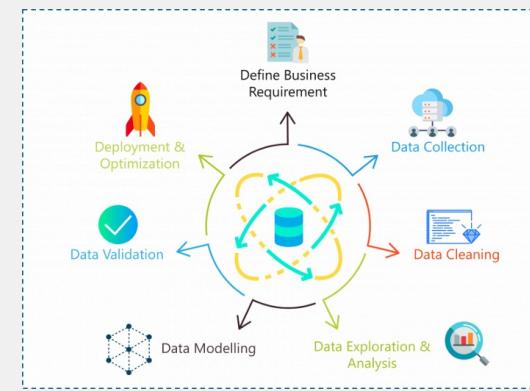
O Ciclo da Ciência de Dados



1. Definir requisitos de negócios

- Etapa de entendimento do problema que será resolvido e as suas possíveis soluções
- Responder perguntas relevantes sobre o projeto
- Determinar requisitos para o sucesso final do projeto
 - Como será mostrado o resultado do projeto?

O Ciclo da Ciência de Dados



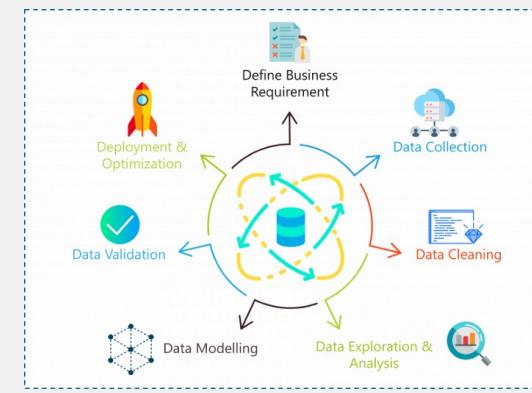
2. Coleta de dados

- Também conhecido como mineração de dados
- Dados podem não estar disponíveis de uma forma fácil
- Os dados podem ser:
 - Privados
 - Públicos
 - Obtidos por meio de entrevistas
 - Acessos na internet
- Como obter os dados?
 - Banco de dados
 - Planilhas
 - Web Scraping
- Averiguar a confiabilidade da fonte de dados

3. Limpeza dos dados

- Etapa utilizada principalmente em projetos de Big Data, onde normalmente são utilizados uma quantidade muito grande de dados.
- Dados podem ter fontes e formatos diferentes
 - Tabelas, imagens, áudios, textos de redes sociais, sites, banco de dados, documentos digitalizados, formulários ...
- Momento de tratar e organizar os dados
- Podem existir valores faltantes, erros de digitação, ...
- Importante: dados ruins levam a conclusões erradas

O Ciclo da Ciência de Dados



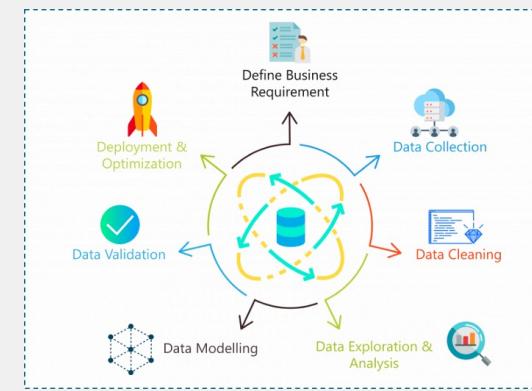
4. Exploração e análise

- Entender os padrões e formar hipóteses sobre os dados
- Relacionar comportamentos e buscar tendências nas diferentes variáveis e/ou grupos
- Como fazer isso?
 - Analisar um subconjunto aleatório dos dados
 - Traçar gráficos e mapas
- O que buscar?
 - Uma possível tendência nos valores utilizando um histograma ou um mapa de calor
 - Outliers (pontos fora da curva)

5. Modelagem

- Momento de aplicar machine learning, inteligência artificial, regressão e vários outros métodos científicos
- Treinar os modelos até existir uma resposta boa
- Explicar comportamentos
- Previsões
- Tipos de algoritmos:
 - Árvore de decisão
 - Regressão linear
 - Agrupamento
 - Combinações de modelos

O Ciclo da Ciência de Dados



6. Validação e visualização

- Etapa que combina os campos da comunicação, psicologia, estatística e arte
- Tem o objetivo final de comunicar os dados de maneira simples, porém de forma eficaz e visualmente agradável
- Podem ser utilizados:
 - Relatórios
 - Dashboards
 - Aplicações web

7. Deploy

- Última etapa, que é possível disponibilizar todos os resultados de uma forma que as pessoas tenham acesso
- Subir as aplicações em servidores
- Administrar acessos e permissões

Diferentes cargos e responsabilidades

Data Engineer	Data Scientist	Data Analyst
		
Construir e otimizar a infraestrutura de dados que irá permitir que o Cientista de Dados e o Analista de Dados rodem suas análises.	Utiliza estatística, aprendizado de máquina e algoritmos matemáticos para realizar previsões, identificar padrões e responder perguntas de negócios.	Compreende questões de negócio, consolida e elabora dashboards e relatórios para comunicar resultados.
Skills Programação, Bancos de Dados, Infraestrutura Cloud, Big Data.	Skills Programação, Bancos de Dados, Estatística, Matemática	Skills Comunicação, UX, Conhecimentos de Negócio, Bancos de Dados
Techs SQL, Python, Cloud (Azure, AWS, Google)	Techs SQL, Python, R, Cloud	Techs SQL, Power BI, Tableau, Qlik View, Excel

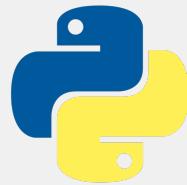
Outros cargos:

- Arquiteto de dados
- Engenheiro de Big Data, IA ou Machine Learning
- Engenheiro DataOps

Tecnologias



juliá



hadoop

+ a b l e a u®



Power BI



Python

Linguagens de programação

Sobre o python

- Linguagem mais popular do mundo
- Fácil para iniciantes
- Open-source
- Escalável e rápida

Utilização na Ciência de Dados

- Processamento de dados
- Coleta
- Inteligência artificial e machine learning
- Visualização



Python

Linguagens de programação

Coleta

BeautifulSoup

Exploração e
Manipulação



Modelagem



Visualização



Reports

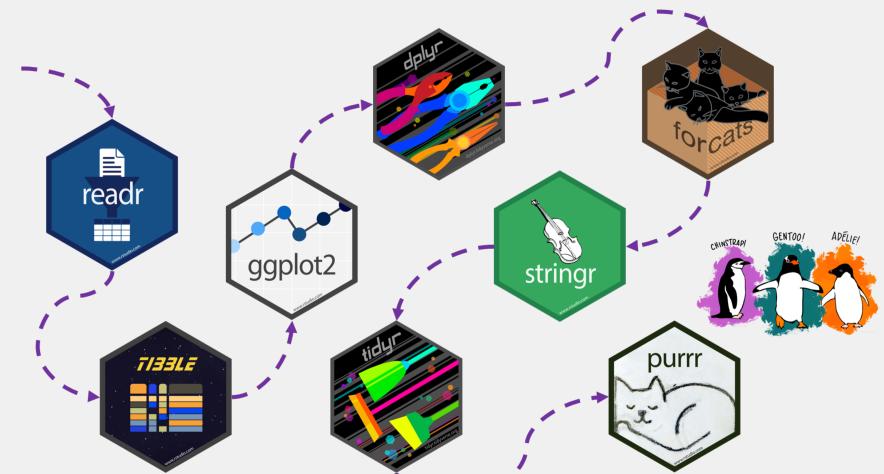


Sobre o R

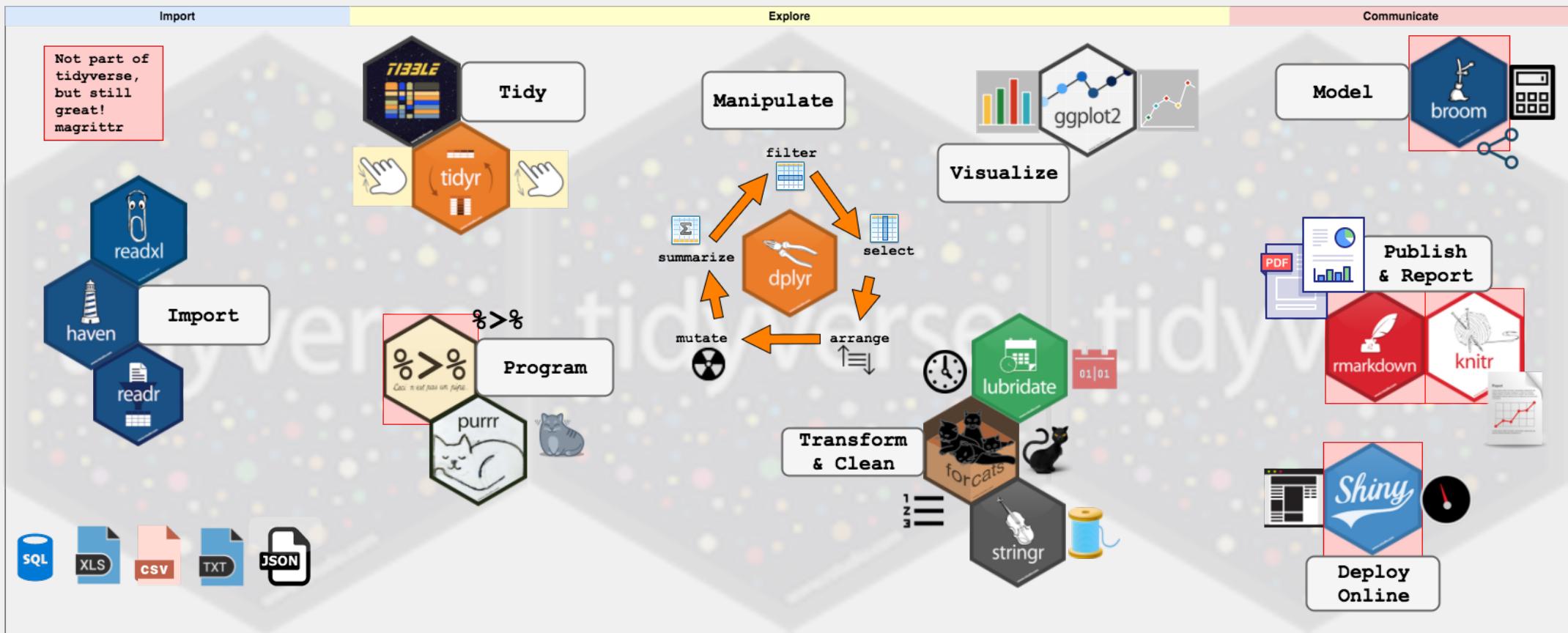
- Top 10 de linguagens mais utilizadas
- Open-source
- Muito utilizada academicamente

Utilização na Ciência de Dados

- Processamentos estatísticos
- Modelagens lineares e não lineares
- Análises
- Machine learning
- Visualização



Linguagens de programação



Sobre o Hadoop

- Plataforma de software para armazenamento e processamento de grande conjunto de dados
- Fornece armazenamento massivo para qualquer tipo de dado, grande poder de processamento e a capacidade de lidar quase ilimitadamente com tarefas e trabalhos ocorrendo ao mesmo tempo

Benefícios

- Escalabilidade e desempenho
- Confiabilidade
- Flexibilidade
- Baixo custo

Power BI

Sobre o Power BI

- Serviço de análise de negócios da Microsoft
- Fornecer visualizações interativas e recursos de inteligência de negócios com uma interface simples o suficiente para que os usuários finais criem seus próprios relatórios e painéis

Benefícios

- Facilidade
- Não requer programação
- Conexão direta com centenas de fontes de dados na nuvem e infraestrutura local
- Mais barato

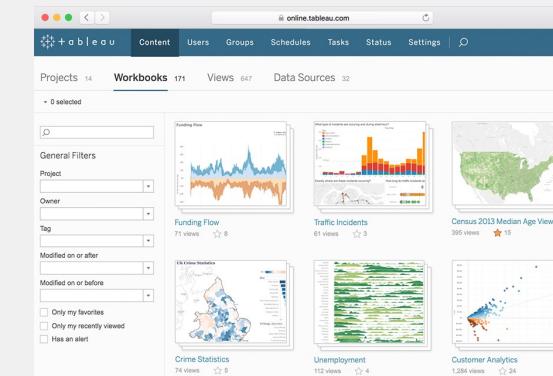


Sobre o Tableau

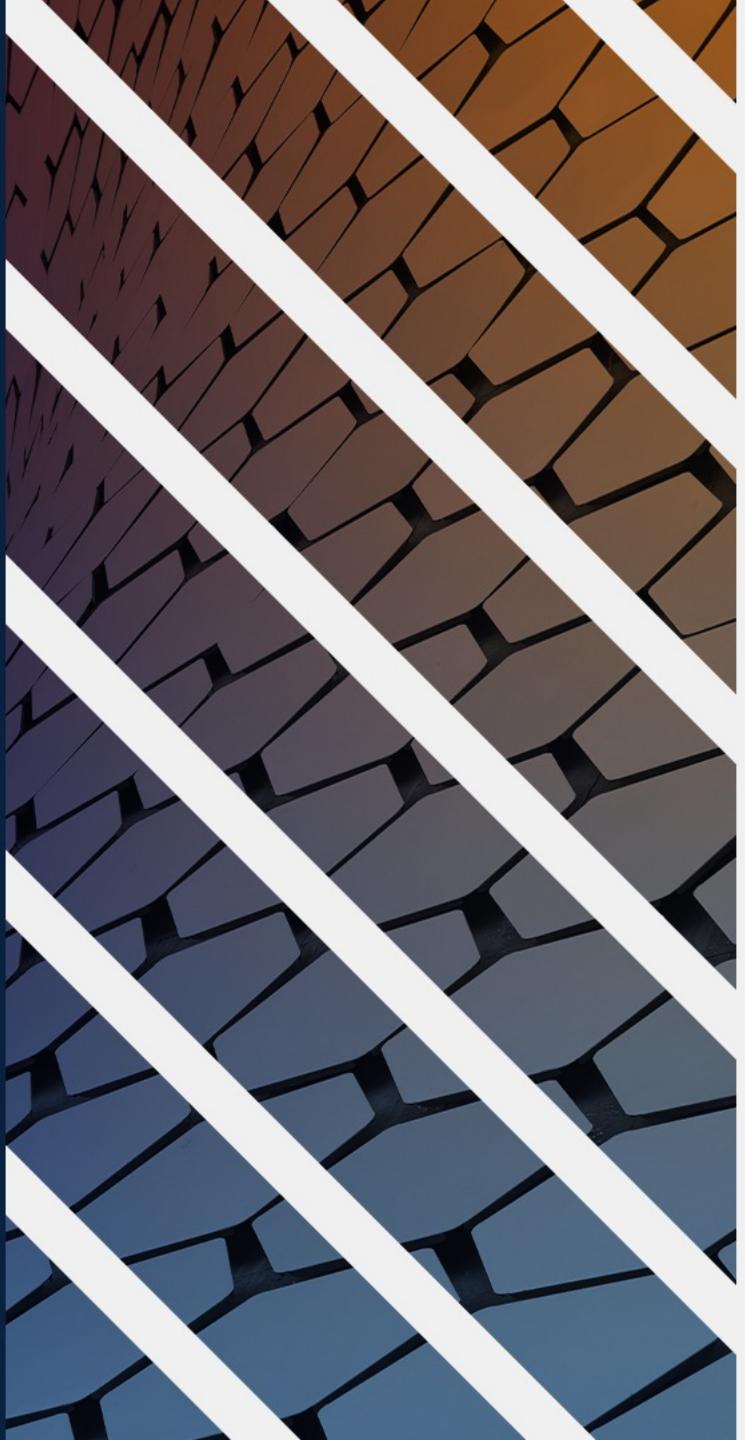
- Plataforma de software de business intelligence para análise avançada de dados

Benefícios

- Completa
- Flexível
- Escalável
- Não requer programação
- Funciona melhor para conjuntos de dados maiores



UTILIZANDO O PYTHON NA PRÁTICA



CONTACT

nathalia@4intelligence.com.br
 /nathaliascampos



4intelligence