Análise Exploratória de Dados

Ciência de Dados – UNIFACISA

Paulo Ribeiro Lins Júnior – paulo.lins@ifpb.edu.br

Grupo de Pesquisa em Comunicações e Processamento de Informação – GComPI
Instituto Federal de Educação, Ciência e Tecnologia da Paraíba – IFPB
Campus Campina Grande

Setembro – 2019

QUEM SOU?

- **Professor no IFPB-CG**
  Estatística Aplicada
  Teoria da Informação e Codificação
  Métodos Numéricos
  Processamento de Sinais
- **Pesquisador no GComPI**
  Grupo de Pesquisa em Comunicações e Processamento de Informação
- **Militante “não xiita”**

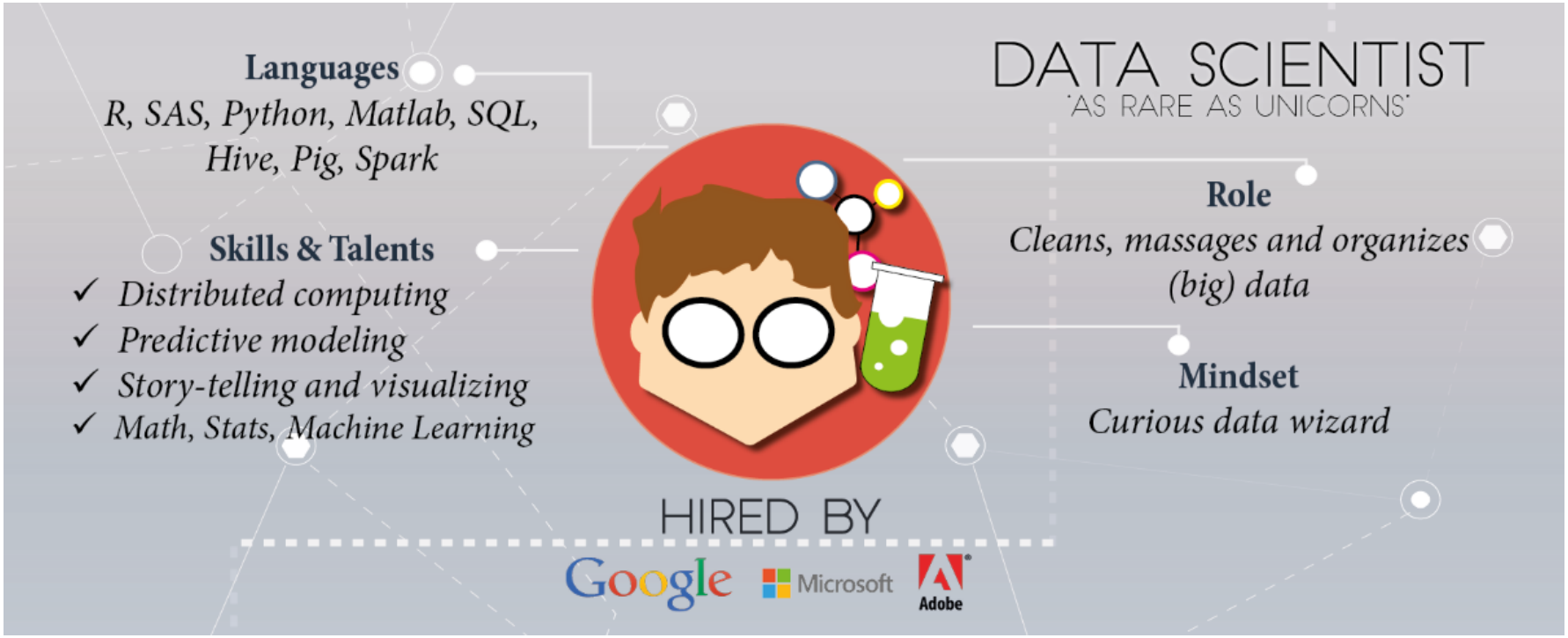
A Nova Revolução e o Papel da Estatística Nela

COM O QUE ESTAMOS BRINCANDO?

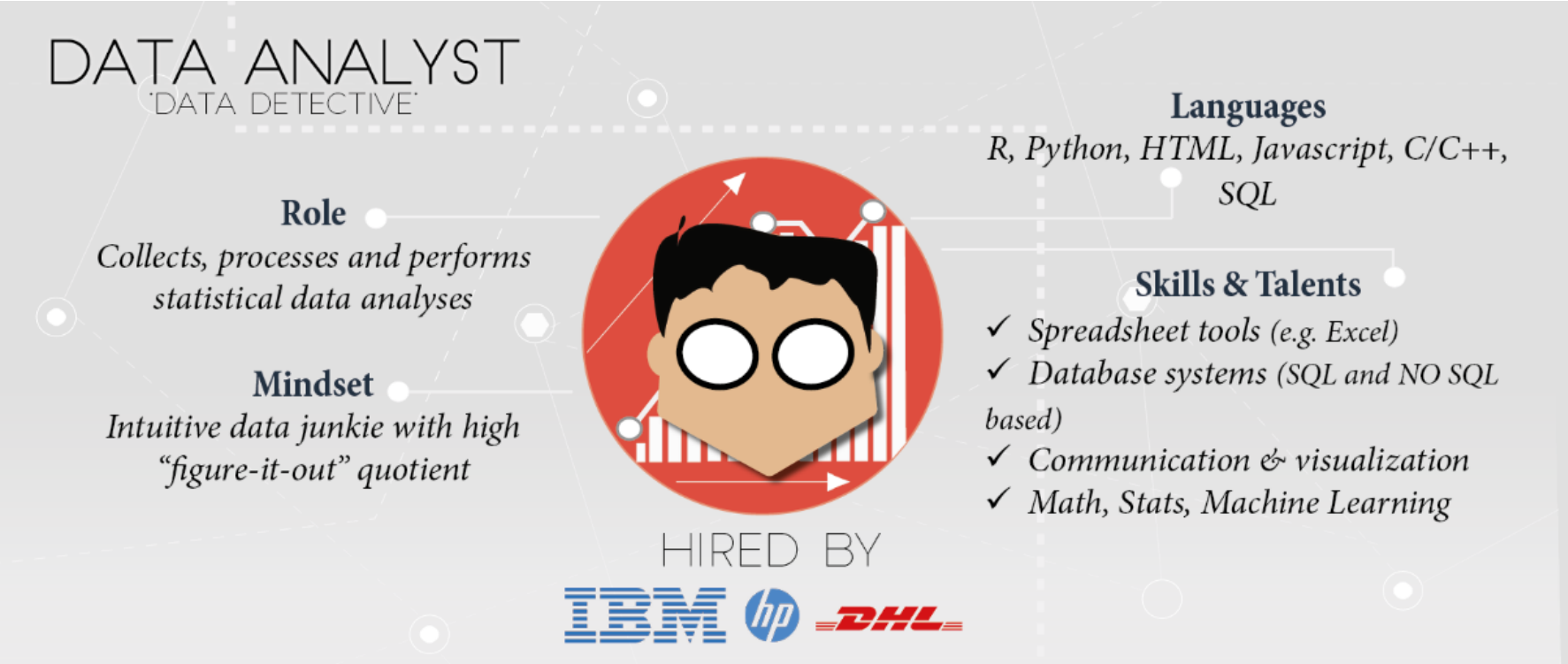
Com o que Estamos Brincando?

Com o que Estamos Brincando?

E Quais os Protagonistas?

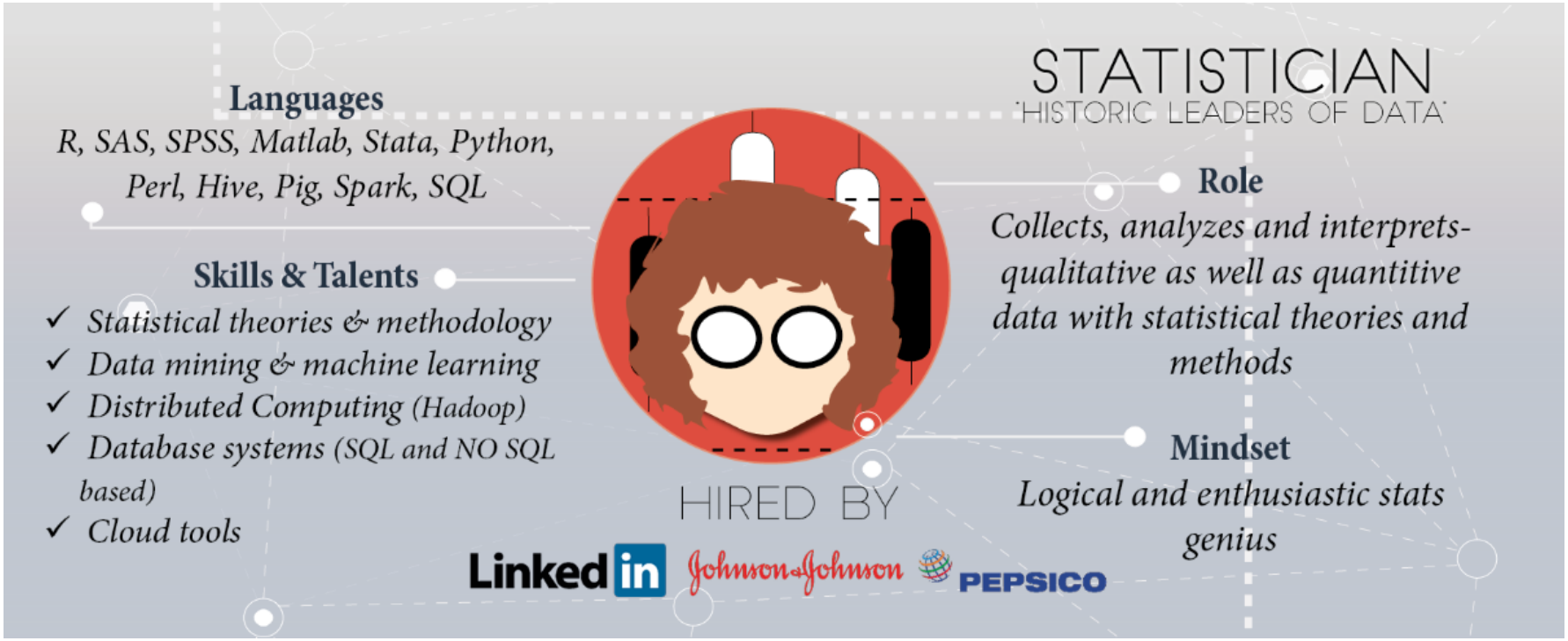


E Quais os Protagonistas?



E Quais os Protagonistas?

E Quais os Protagonistas?



Um Aviso "Antigo"

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

Olá, meu nome é Python

Sobre a linguagem Python

- Criada por Guido van Rossum em 1991
- Linguagem de Alto Nível
- Interpretada
- Programação:
  - Modular
  - Orientada a objetos
  - Funcional
- Tipagem dinâmica e forte
- Vasta coleção de bibliotecas
- Código aberto (GPL)

Sobre a linguagem Python

- Diversas estruturas de dados nativas
  - Lista, tupla, dicionário
- Gerenciamento de memória automático
- Tratamento de exceções
- Sobrecarga de operadores
- Tudo é objeto
- Indentação para estrutura de bloco
- Multiplataforma
- Quem usa?
  - Blender, GIMP, Inkscape, YouTube, NASA, CERN ...

Porque usar Python?

- Fácil, simples
- Sintaxe limpa e “intuitiva”
- Diversas bibliotecas já inclusas
- Interativa
- Protótipos rápidos
- Alta produtividade
- Interfaces para outras linguagens como C/C++, Fortran, R, ...

Porque usar Python?

Sep 2016	Sep 2015	Change	Programming Language	Ratings	Change
1	1		Java	18.236%	-1.33%
2	2		C	10.955%	-4.67%
3	3		C++	6.657%	-0.13%
4	4		C#	5.493%	+0.58%
5	5		Python	4.302%	+0.64%
6	7	▲	JavaScript	2.929%	+0.59%
7	6	▼	PHP	2.847%	+0.32%
8	11	▲	Assembly language	2.417%	+0.61%
9	8	▼	Visual Basic .NET	2.343%	+0.28%
10	9	▼	Perl	2.333%	+0.43%

PORQUE USAR PYTHON?

Choose a Ranking (choose a weighting or make your own)

IEEE Spectrum | Trending | Jobs | Open | Custom

Edit Ranking | Add a Comparison |

Language Types (click to hide)

Web | Mobile | Enterprise | Embedded

Language Rank	Types	Spectrum Ranking
1. C		100.0
2. Java		98.1
3. Python		97.9
4. C++		95.8
5. R		87.7
6. C#		86.4
7. PHP		82.4
8. JavaScript		81.9
9. Ruby		74.0
10. Go		71.5

PORQUE USAR PYTHON?

Choose a Ranking (choose a weighting or make your own)

IEEE Spectrum | Trending | Jobs | Open | Custom

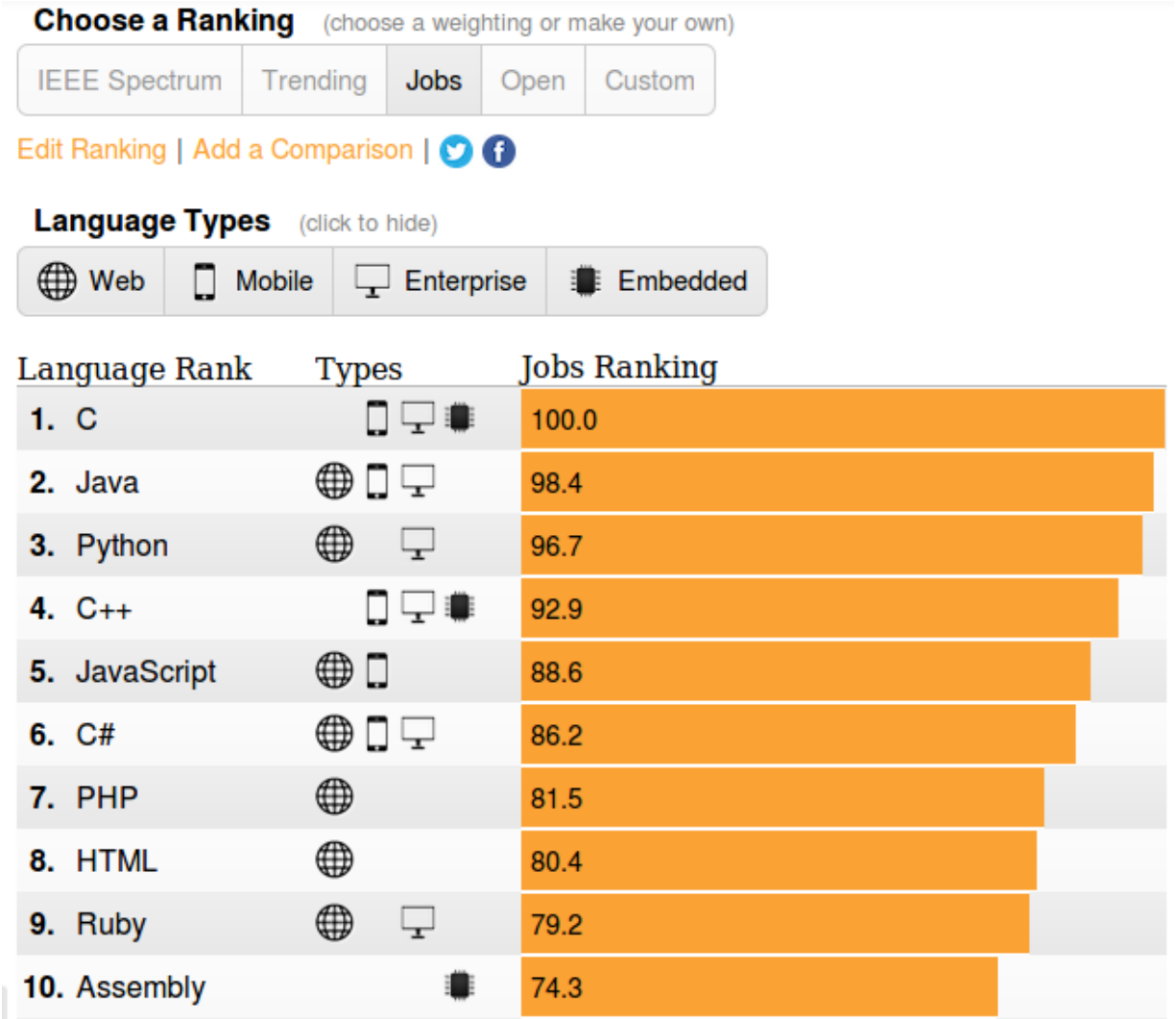
Edit Ranking | Add a Comparison |

Language Types (click to hide)

Web | Mobile | Enterprise | Embedded

Language Rank	Types	Trending Ranking
1. C		100.0
2. C++		97.6
3. Python		96.8
4. Java		94.3
5. Swift		88.2
6. R		85.2
7. JavaScript		84.4
8. Ruby		80.7
9. Go		80.6
10. C#		78.5

PORQUE USAR PYTHON?



Choose a Ranking (choose a weighting or make your own)

IEEE Spectrum | Trending | **Jobs** | Open | Custom

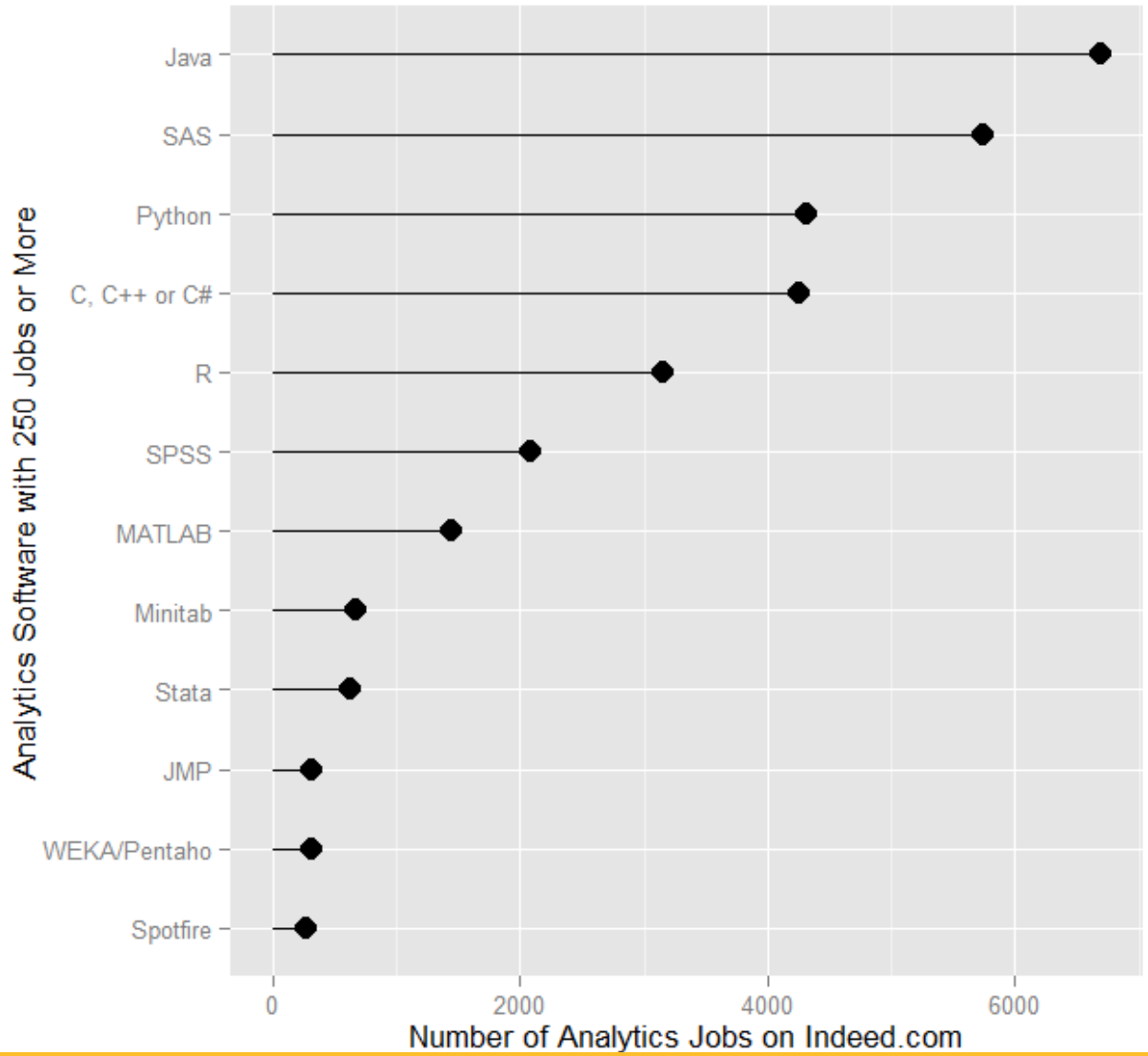
Edit Ranking | Add a Comparison |

Language Types (click to hide)

Web | Mobile | Enterprise | Embedded

Language Rank	Types	Jobs Ranking
1. C	Mobile, Enterprise, Embedded	100.0
2. Java	Web, Mobile, Enterprise	98.4
3. Python	Web, Enterprise	96.7
4. C++	Mobile, Enterprise, Embedded	92.9
5. JavaScript	Web, Mobile	88.6
6. C#	Web, Mobile, Enterprise	86.2
7. PHP	Web	81.5
8. HTML	Web	80.4
9. Ruby	Web, Enterprise	79.2
10. Assembly	Embedded	74.3

PORQUE USAR PYTHON?



A Caixa de Ferramentas

As Ferramentas Básicas

O QUE É IPYTHON?

- Utiliza o modo iterativo como Matlab, Mathematica ...
- Permite customização e flexibilidade para executar diretamente códigos Python
- Comandos mágicos começam pelo caractere %
- Auto-completa comandos e atributos com a tecla<TAB>
- Informação sobre qualquer objeto digitando object_name?
- Interação com Tkinter, GTK, Qt e wxWidgets
- Depuração com %debug para examinar o problema
- Histórico e log dos comandos com %history e %logstart diario.log

O QUE É NUMPY?

- Numerical Python
- Biblioteca para manipulação de vetores e matrizes
- Operações rápidas em matrizes (funções vetorizadas)
- Diferença com relação a listas tradicionais do Python
  - Vetor homogêneo
  - Muito mais eficientes do que as listas (python puro)
  - Número de elemento deve ser conhecido a priori.
  - Muito eficiente (implementado em C)

SOMENTE POR CURIOSIDADE...

```
# Python puro
import time

l = 10000000

start = time.time()
a, b = range(l), range(l)
c = []
for i in a:
    c.append(a[i] * b[i])
t = time.time() - start
print("Tempo: %s" % t)
```

Tempo: 4.49 s

```
# NumPy
import time
import numpy as np

l = 10000000

start = time.time()
a = np.arange(l)
b = np.arange(l)
c = a * b
t = time.time() - start
print("Tempo: %s" % t)
```

Tempo: 0.37 s

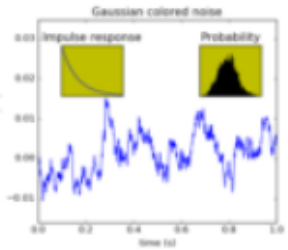
E o que é SciPy?

- Coleção de algoritmos matemáticos e funções utilitárias
- Implementado em cima do NumPy
- Dividido em sub-módulos
  - constants: Constantes básicas
  - fftpack: Transformada Rápida de Fourier
  - integrate: Integração numérica e ODE solvers
  - interpolate: Interpolação (Splines)
  - **stats: Distribuições e funções estatísticas**
  - optimize: Otimização
  - sparse: Matrizes esparsas
  - linalg: Álgebra Linear
  - io: Entrada e Saída
  - signal: Processamendo digital de sinais
  - ndimage: Processamento digital de imagens

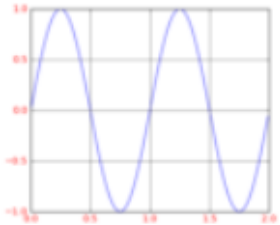
E o que é Matplotlib?

- Biblioteca com funções especializadas na plotagem de gráficos de altíssima qualidade, com grande variedade de possibilidades de caracterização e de formatação, usando comandos relativamente simples, no estilo do MatLab.

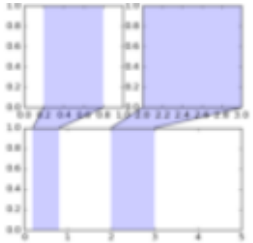
E O QUE É MATPLOTLIB?



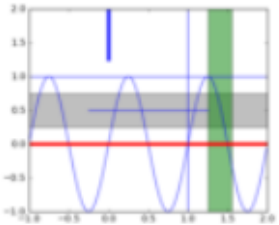
axes_demo



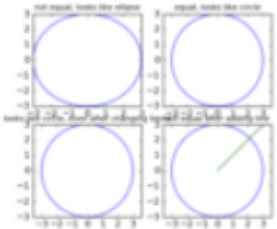
axes_props



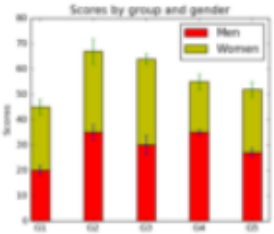
axes_zoom_effect



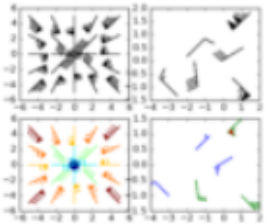
axhspan_demo



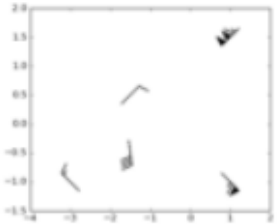
axis_equal_demo



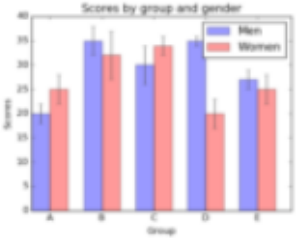
bar_stacked

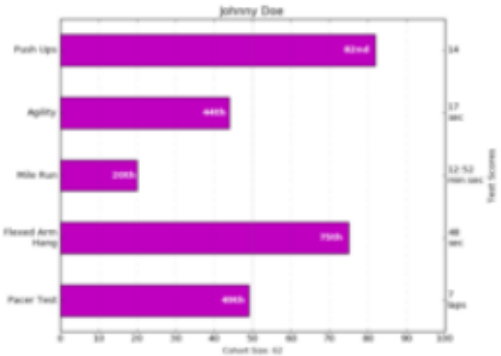


barb_demo

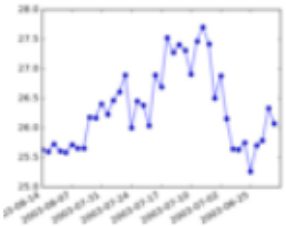


barb_demo

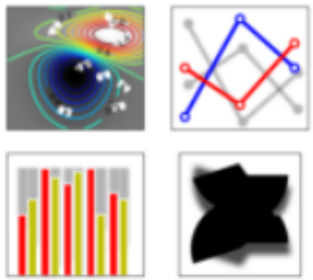




E o que é Matplotlib?

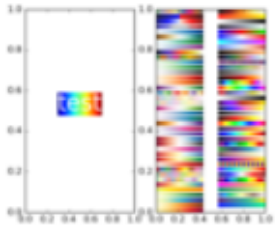


date_index_formatter

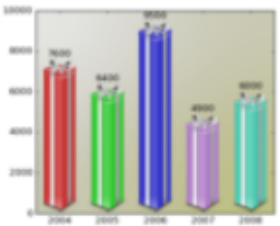


demo_agg_filter

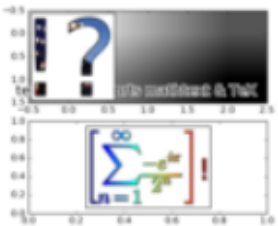
demo_annotation_box



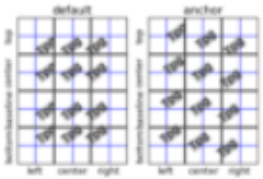
demo_bboximage



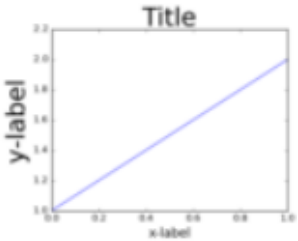
demo_ribbon_box



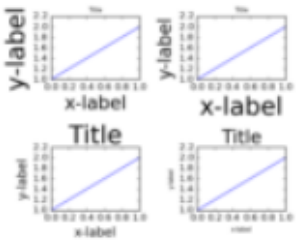
demo_text_path



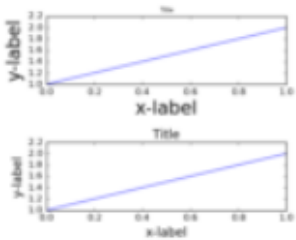
demo_text_rotation_mode



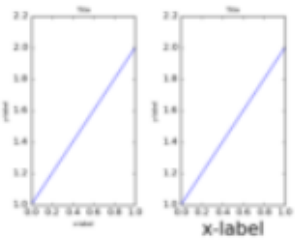
demo_tight_layout



demo_tight_layout



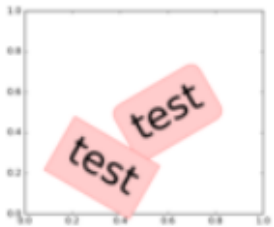
demo_tight_layout



demo_tight_layout

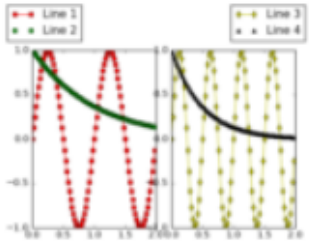
demo_tight_layout

E o que é Matplotlib?

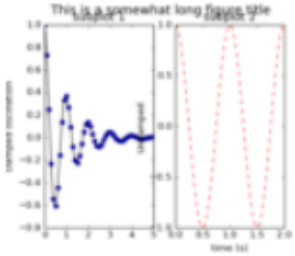


fancytextbox_demo

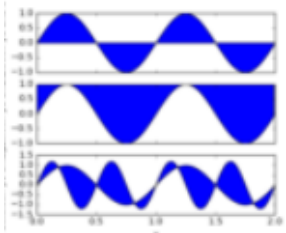
figimage_demo



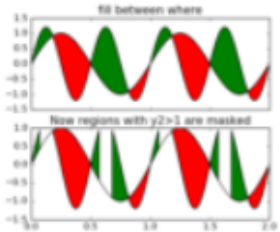
figlegend_demo



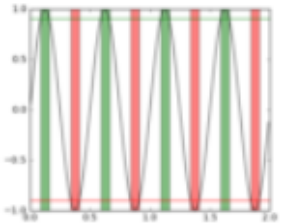
figure_title



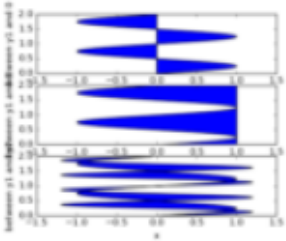
fill_between_demo



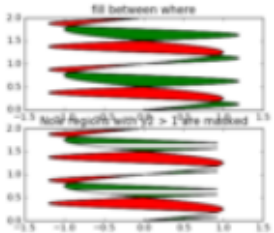
fill_between_demo



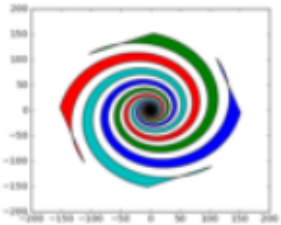
fill_between_demo



fill_betweenx_demo

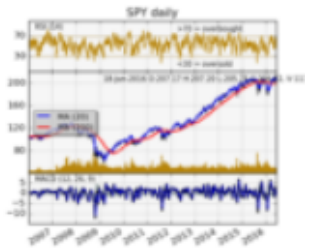


fill_betweenx_demo



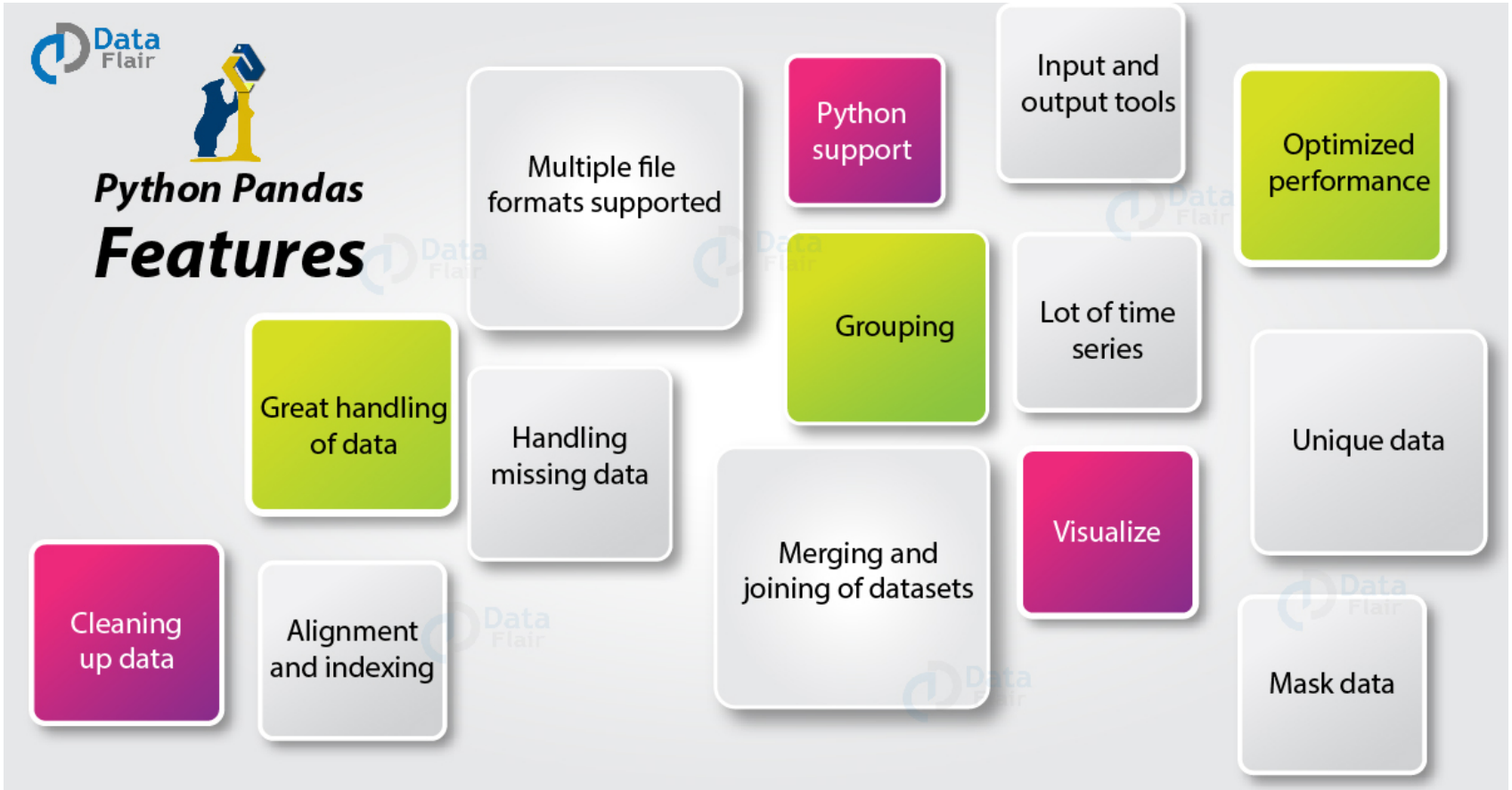
fill_spiral

finance_demo



finance_work2

PANDAS: O CANIVETE SUIÇO DOS DADOS



Nossa Disciplina

Análise Exploratória de Dados

Conjunto de técnicas, numéricas e gráficas, que permite:

- maximizar a percepção de um conjunto de dados;
- descobrir estrutura subjacente;
- extrair variáveis importantes;
- detectar outliers e anomalias;
- testar premissas subjacentes;
- desenvolver modelos parcimoniosos; e
- determinar configurações ideais de fatores.

Análise Exploratória de Dados

Conjunto de técnicas, numéricas e gráficas, que permite:

- maximizar a percepção de um conjunto de dados;
- descobrir estrutura subjacente;
- extrair variáveis importantes;
- detectar outliers e anomalias;
- testar premissas subjacentes;
- desenvolver modelos parcimoniosos; e
- determinar configurações ideais de fatores.

Análise Exploratória de Dados

Conjunto de técnicas, numéricas e gráficas, que permite:

- maximizar a percepção de um conjunto de dados;
- descobrir estrutura subjacente;
- extrair variáveis importantes;
- detectar outliers e anomalias;
- testar premissas subjacentes;
- desenvolver modelos parcimoniosos; e
- determinar configurações ideais de fatores.

Análise Exploratória de Dados

Conjunto de técnicas, numéricas e gráficas, que permite:

- maximizar a percepção de um conjunto de dados;
- descobrir estrutura subjacente;
- extrair variáveis importantes;
- detectar outliers e anomalias;
- testar premissas subjacentes;
- desenvolver modelos parcimoniosos; e
- determinar configurações ideais de fatores.

Análise Exploratória de Dados

Conjunto de técnicas, numéricas e gráficas, que permite:

- maximizar a percepção de um conjunto de dados;
- descobrir estrutura subjacente;
- extrair variáveis importantes;
- detectar outliers e anomalias;
- testar premissas subjacentes;
- desenvolver modelos parcimoniosos; e
- determinar configurações ideais de fatores.

Análise Exploratória de Dados

Conjunto de técnicas, numéricas e gráficas, que permite:

- maximizar a percepção de um conjunto de dados;
- descobrir estrutura subjacente;
- extrair variáveis importantes;
- detectar outliers e anomalias;
- testar premissas subjacentes;
- desenvolver modelos parcimoniosos; e
- determinar configurações ideais de fatores.

Análise Exploratória de Dados

Conjunto de técnicas, numéricas e gráficas, que permite:

- maximizar a percepção de um conjunto de dados;
- descobrir estrutura subjacente;
- extrair variáveis importantes;
- detectar outliers e anomalias;
- testar premissas subjacentes;
- desenvolver modelos parcimoniosos; e
- determinar configurações ideais de fatores.

Análise Exploratória de Dados

Conjunto de técnicas, numéricas e gráficas, que permite:

- maximizar a percepção de um conjunto de dados;
- descobrir estrutura subjacente;
- extrair variáveis importantes;
- detectar outliers e anomalias;
- testar premissas subjacentes;
- desenvolver modelos parcimoniosos; e
- determinar configurações ideais de fatores.

Análise Exploratória de Dados

Ementa

- Estruturas de alocação de dados em Python;
- Manipulação e processamento de dados usando Python;
- Classificação de variáveis estatísticas;
- Distribuições de frequência;
- Medidas de posição, centralidade e dispersão;
- Gráficos estatísticos;
- Correlação entre duas variáveis.

Análise Exploratória de Dados

Conteúdo

Sexta – noite

Introdução ao Jupyter Notebook; Estruturas básicas do Pandas: Series e DataFrames; Indexação básica

Sábado – manhã

Variáveis Estatísticas e suas classificações; Distribuições de frequências; Gráficos estatísticos fundamentais

Sábado – tarde

Medidas resumo: centralidade, posição, dispersão; Boxplot; Correlação e gráfico de dispersão e mapas de calor; **Projeto Guiado**.

Análise Exploratória de Dados

Avaliação

A avaliação será contínua, considerando todas as interações e atividades realizadas em sala, complementada por um pequeno projeto de análise de dados reais, a ser entregue em um prazo de 15 dias após a aula.

Vamos começar?!