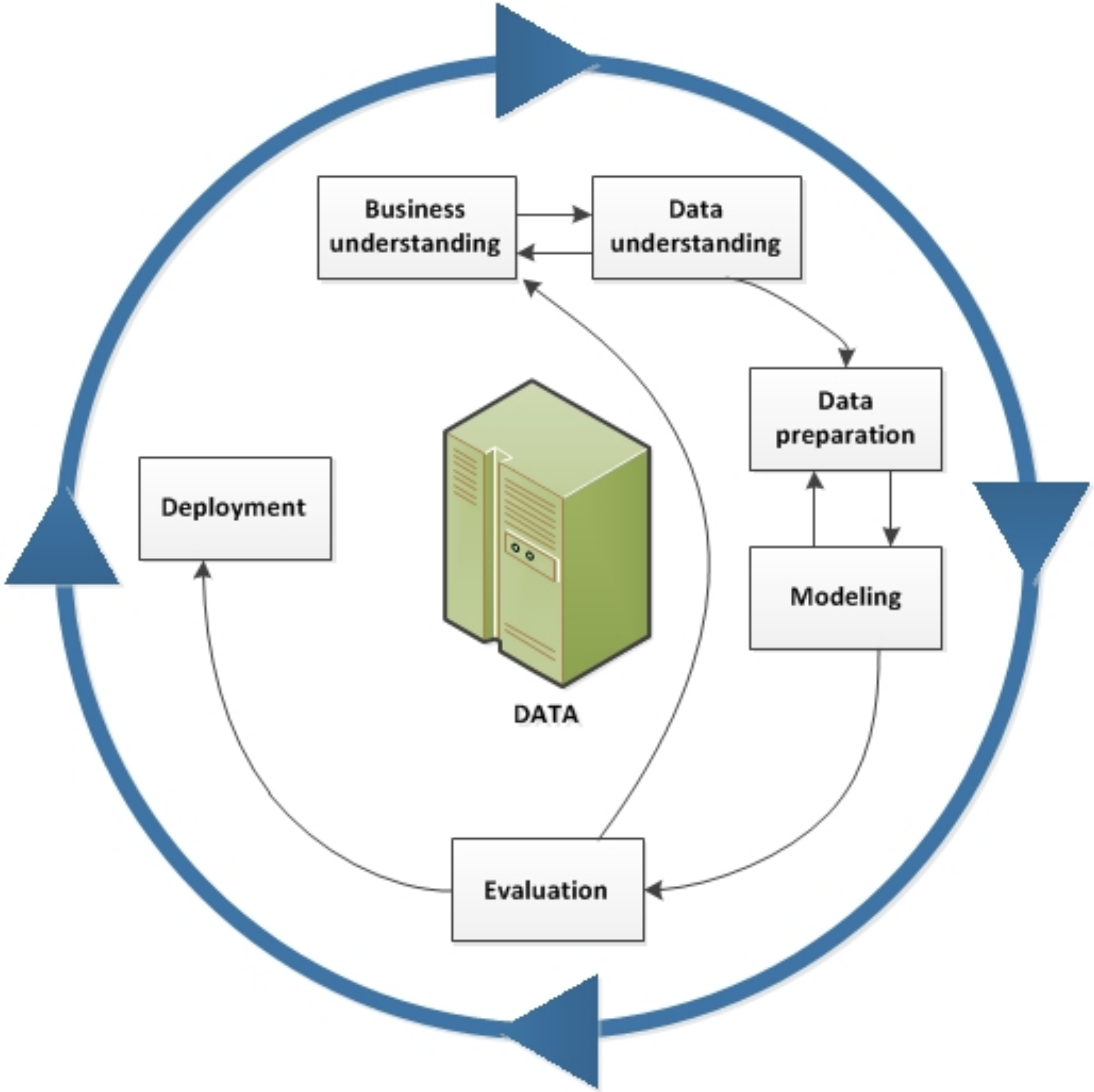
Aplicações de Inteligência Artificial
Parte II

Hitoshi Nagano, Ph.D.

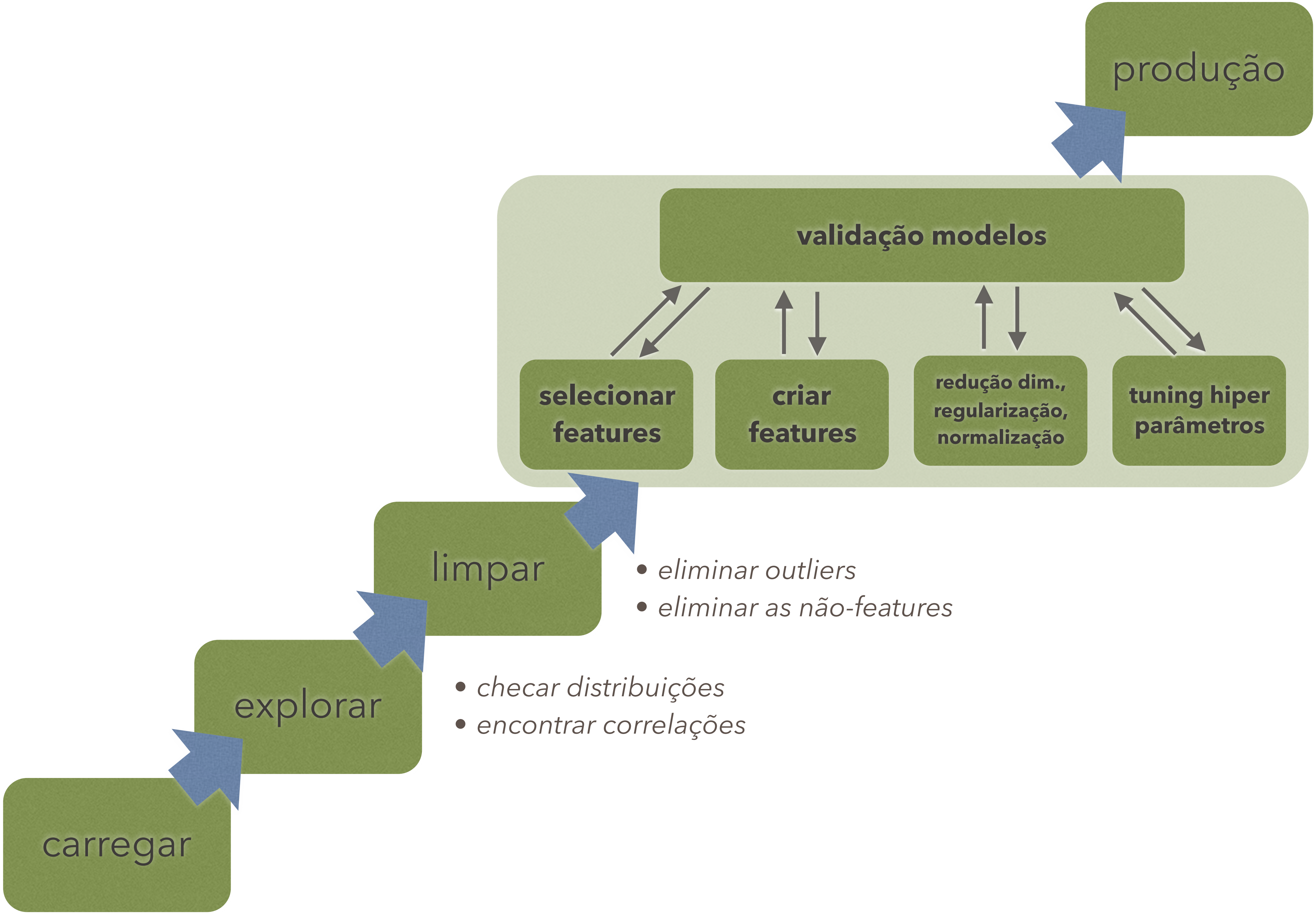
WORKFLOW
CIENCIA DE DADOS

CRISP-DM



credito: IBM

Workflow - construção e validação do modelo



- *eliminar outliers*
- *eliminar as não-features*

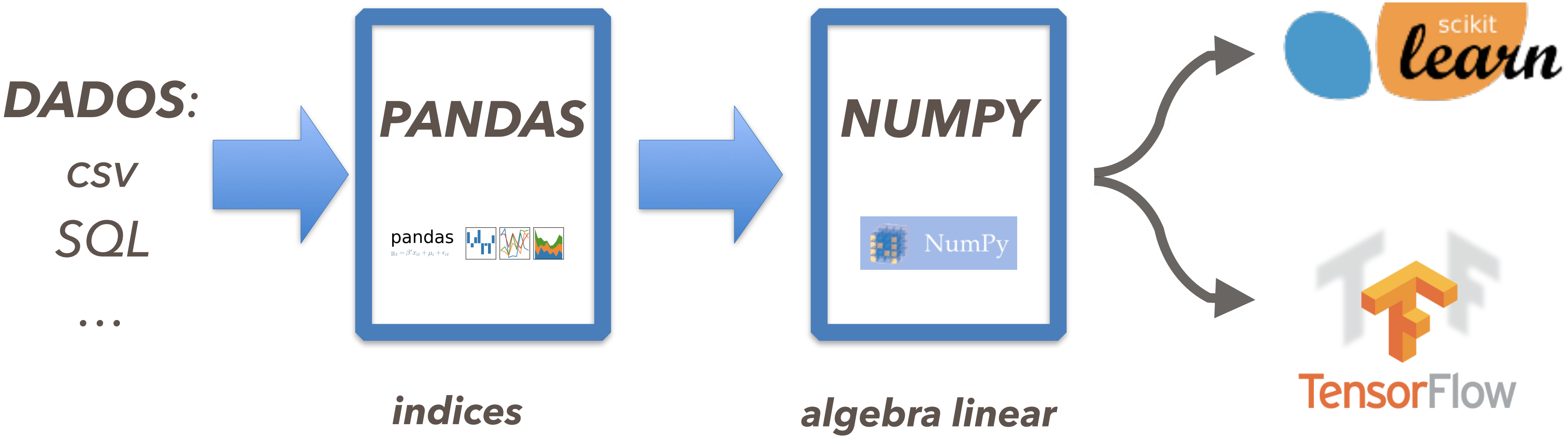
- *checar distribuições*
- *encontrar correlações*

Passos iniciais

- carregar dados
- checar distribuições
- existem correlações?
- remover outliers
- imputação
- limpeza

PANDAS, NUMPY E SCIKIT-LEARN

ferramentas Python



DADOS:

csv

SQL

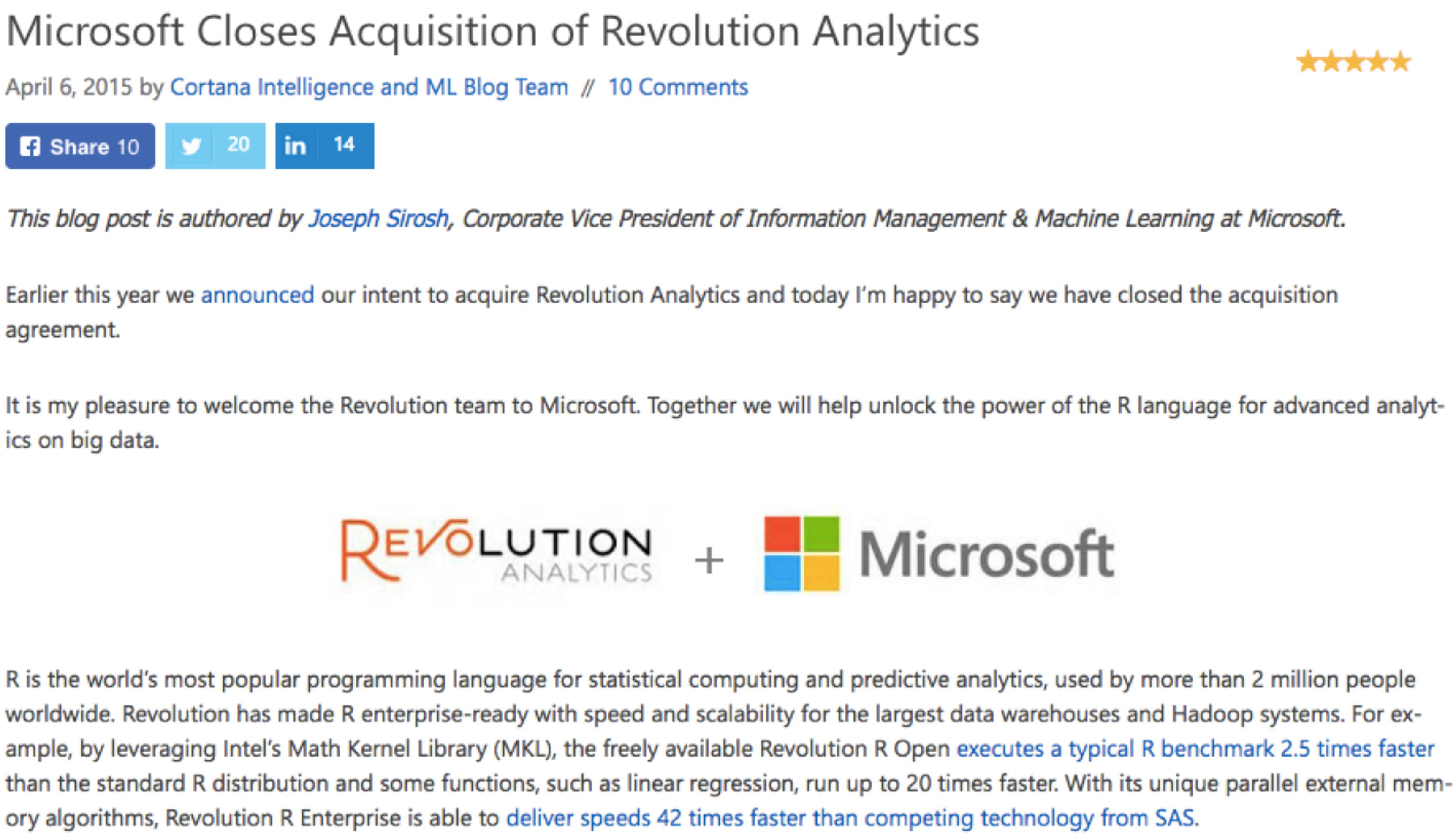
...

PANDAS

NUMPY

indices
slicing, indexing
limpeza
get_dummies
merge, concat
groupby

algebra linear
multiplicacao matrizes
prod. interno
transform. linear



Microsoft Closes Acquisition of Revolution Analytics

★★★★★

April 6, 2015 by Cortana Intelligence and ML Blog Team // 10 Comments

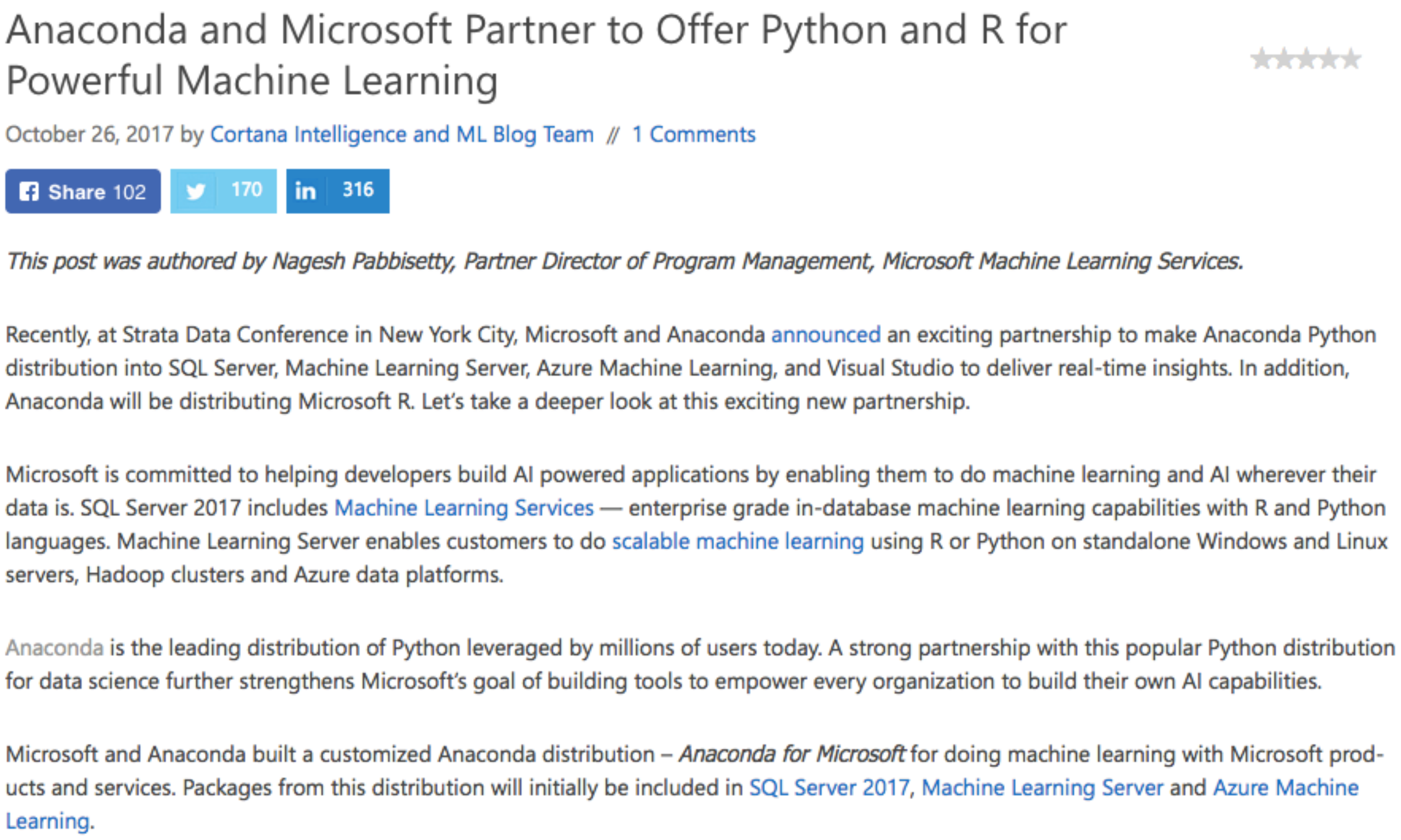
Share 10 | 20 | in 14

This blog post is authored by Joseph Sirosh, Corporate Vice President of Information Management & Machine Learning at Microsoft.

Earlier this year we announced our intent to acquire Revolution Analytics and today I'm happy to say we have closed the acquisition agreement.

It is my pleasure to welcome the Revolution team to Microsoft. Together we will help unlock the power of the R language for advanced analytics on big data.

R is the world's most popular programming language for statistical computing and predictive analytics, used by more than 2 million people worldwide. Revolution has made R enterprise-ready with speed and scalability for the largest data warehouses and Hadoop systems. For example, by leveraging Intel's Math Kernel Library (MKL), the freely available Revolution R Open executes a typical R benchmark 2.5 times faster than the standard R distribution and some functions, such as linear regression, run up to 20 times faster. With its unique parallel external memory algorithms, Revolution R Enterprise is able to deliver speeds 42 times faster than competing technology from SAS.



Anaconda and Microsoft Partner to Offer Python and R for Powerful Machine Learning

☆☆☆☆☆

October 26, 2017 by Cortana Intelligence and ML Blog Team // 1 Comments

This post was authored by Nagesh Pabbisetty, Partner Director of Program Management, Microsoft Machine Learning Services.

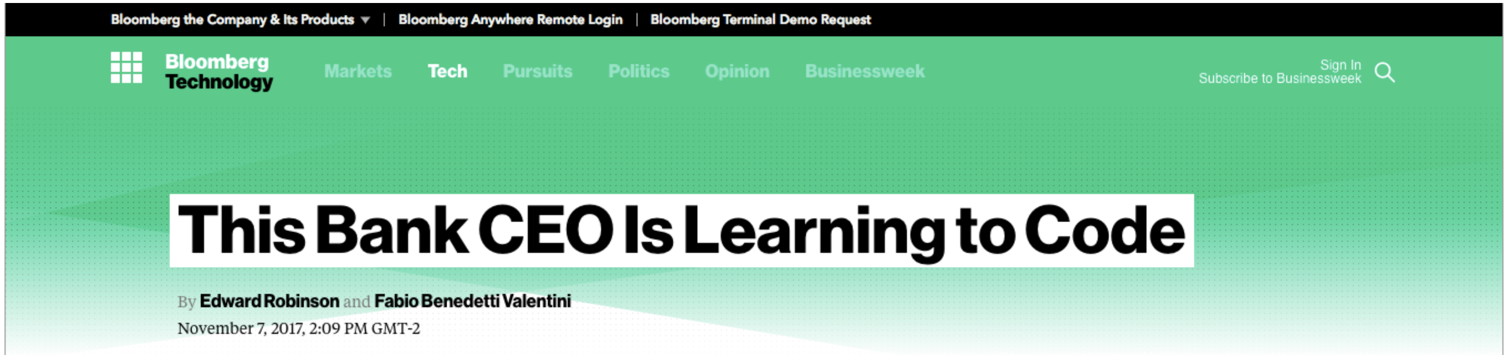
Recently, at Strata Data Conference in New York City, Microsoft and Anaconda announced an exciting partnership to make Anaconda Python distribution into SQL Server, Machine Learning Server, Azure Machine Learning, and Visual Studio to deliver real-time insights. In addition, Anaconda will be distributing Microsoft R. Let's take a deeper look at this exciting new partnership.

Microsoft is committed to helping developers build AI powered applications by enabling them to do machine learning and AI wherever their data is. SQL Server 2017 includes Machine Learning Services — enterprise grade in-database machine learning capabilities with R and Python languages. Machine Learning Server enables customers to do scalable machine learning using R or Python on standalone Windows and Linux servers, Hadoop clusters and Azure data platforms.

Anaconda is the leading distribution of Python leveraged by millions of users today. A strong partnership with this popular Python distribution for data science further strengthens Microsoft's goal of building tools to empower every organization to build their own AI capabilities.

Microsoft and Anaconda built a customized Anaconda distribution – *Anaconda for Microsoft* for doing machine learning with Microsoft products and services. Packages from this distribution will initially be included in SQL Server 2017, Machine Learning Server and Azure Machine Learning.

Python é para voce… e para os CEO's tambem



Bloomberg the Company & Its Products ▾ | Bloomberg Anywhere Remote Login | Bloomberg Terminal Demo Request

This Bank CEO Is Learning to Code

By Edward Robinson and Fabio Benedetti Valentini

November 7, 2017, 2:09 PM GMT-2

- **Frederic Oudea** is doing everything he can to keep up with the technological changes roiling the European banking industry. The chief executive officer of Societe Generale SA has collaborated with fintech startups, backed accelerator programs to nurture innovation, and invested heavily in its French mobile-banking unit as well as in hundreds of apps.

  Now **he's even taken up writing software code himself**.

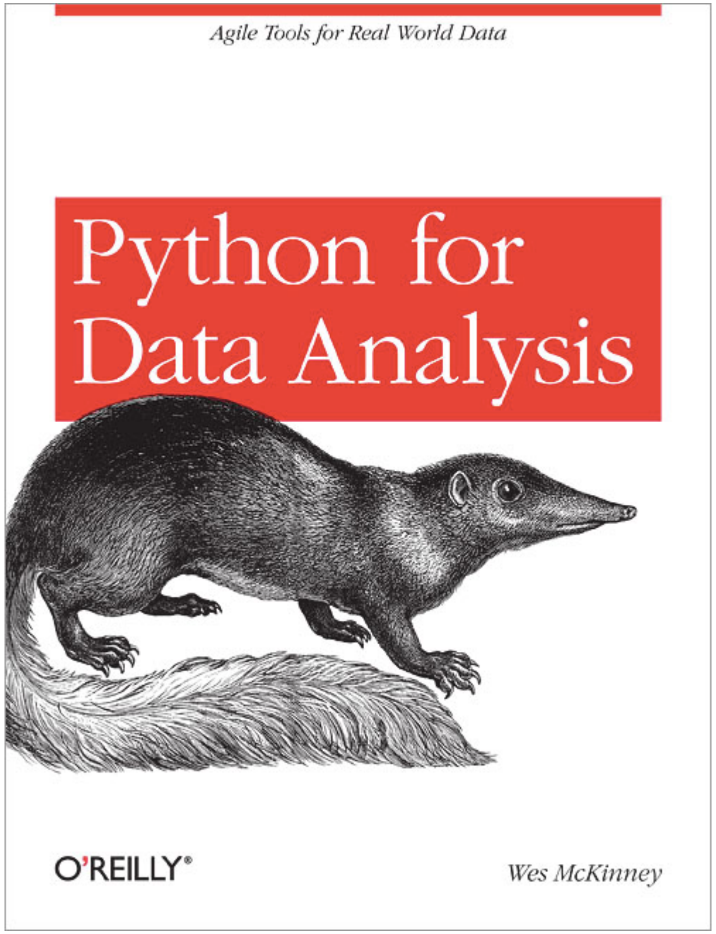
  "It was important for me also **to understand exactly what coding means**, so I spend a **few hours coding in Python, which is one of the two languages for data**," Oudea said in an interview at Web Summit 2017, a tech-industry conference in Lisbon.

- "We are taking the challenge seriously and understanding that there is a need to change the model and culturally embrace new technologies," said Oudea, 54.

  By offering alternative methods for making payments, newcomers are trying to take customers away from traditional banks for other services, such as lending. **Societe Generale, the third-biggest French bank by market value**, will be locked in this battle for the next few years, Oudea said.

pandas, numpy, scipy

- http://shop.oreilly.com/product/0636920023784.do



Por que Pandas

- métodos para limpeza, imputação
- exploração dos dados
- manipulação de variáveis para feature engineering

OBJETIVOS:

⇒ qualidade dos dados (lembrar GIGO!)

⇒ melhora da performance

VARIÁVEIS E DISTRIBUIÇÕES

variáveis

- numérica:
  - discreta
  - contínua
- categórica
  - ordinais (ordenáveis)
  - nominais (não ordenáveis)

Exemplos de distribuições numéricas

➡ discreta

- bernoulli
- binomial
- poisson

➡ contínua

- normal
- exponencial

DISTRIBUIÇÕES

- ➡ paramétricas
- ➡ não-paramétricas

Estatísticas para analise…

… univariada:

- ➡ média
- ➡ variancia
  desvio padrão
- ➡ mediana, quartil
- ➡ moda

… multivariada

- ➡ covariância
- ➡ correlação

- ➡ numerica com numerica
- ➡ categorica com categorica
- ➡ numerica com categorica

OUTLIERS E IMPUTAÇÃO

causas

- eventos estranhos,
  combinação de eventos

  analisar entender

- sensores quebrados

  ignorar

- entrada manual, digitação errada

  ignorar Re-inserir

- erros no processamento

  ignorar corrigir

algoritmo para rejeitar outliers

- treinar
- verificar as observações com maior erro residual
- retirar essas observações

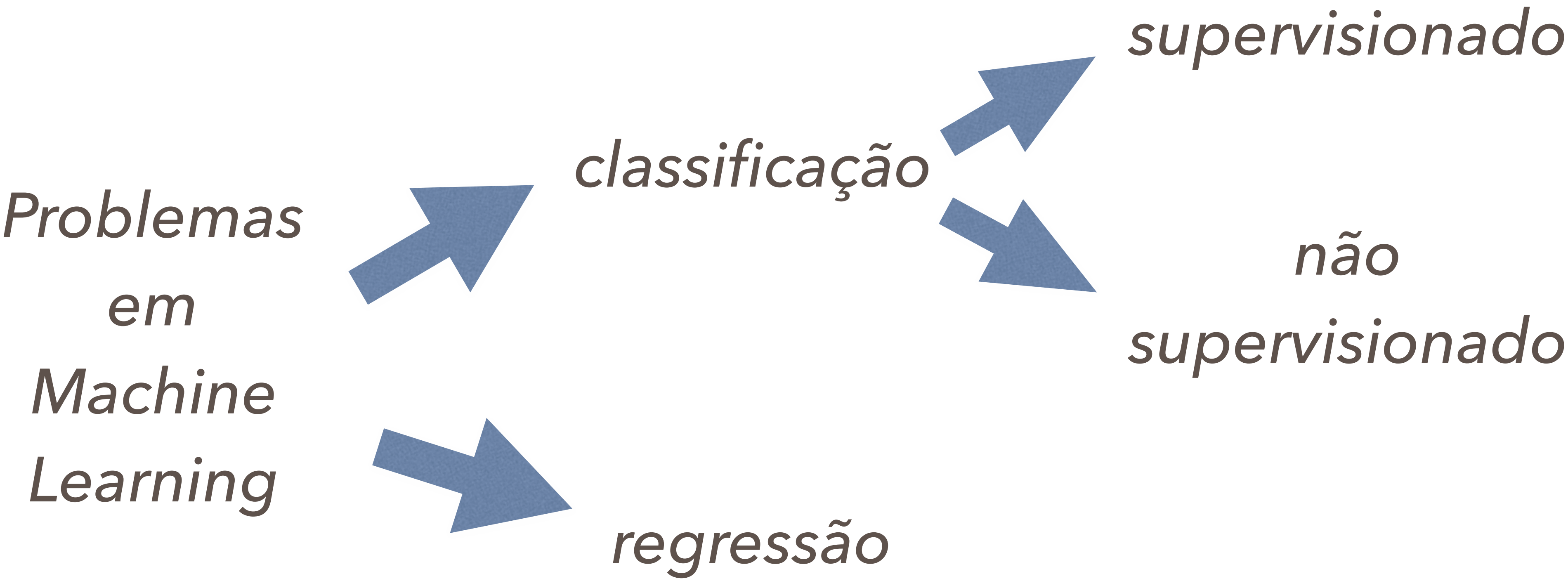
≈10%

- rodar de novo

imputação

- Quais valores substituir:
  - ➡ valor não disponível (NaN, NULL)
  - ➡ valor claramente errado, equiv. a não disponível
  - ➡ valor intencionalmente não disponivel
- substituição por um valor unico, pela média, ou outro critério

TIPOS DE PROBLEMAS



Problemas em Machine Learning

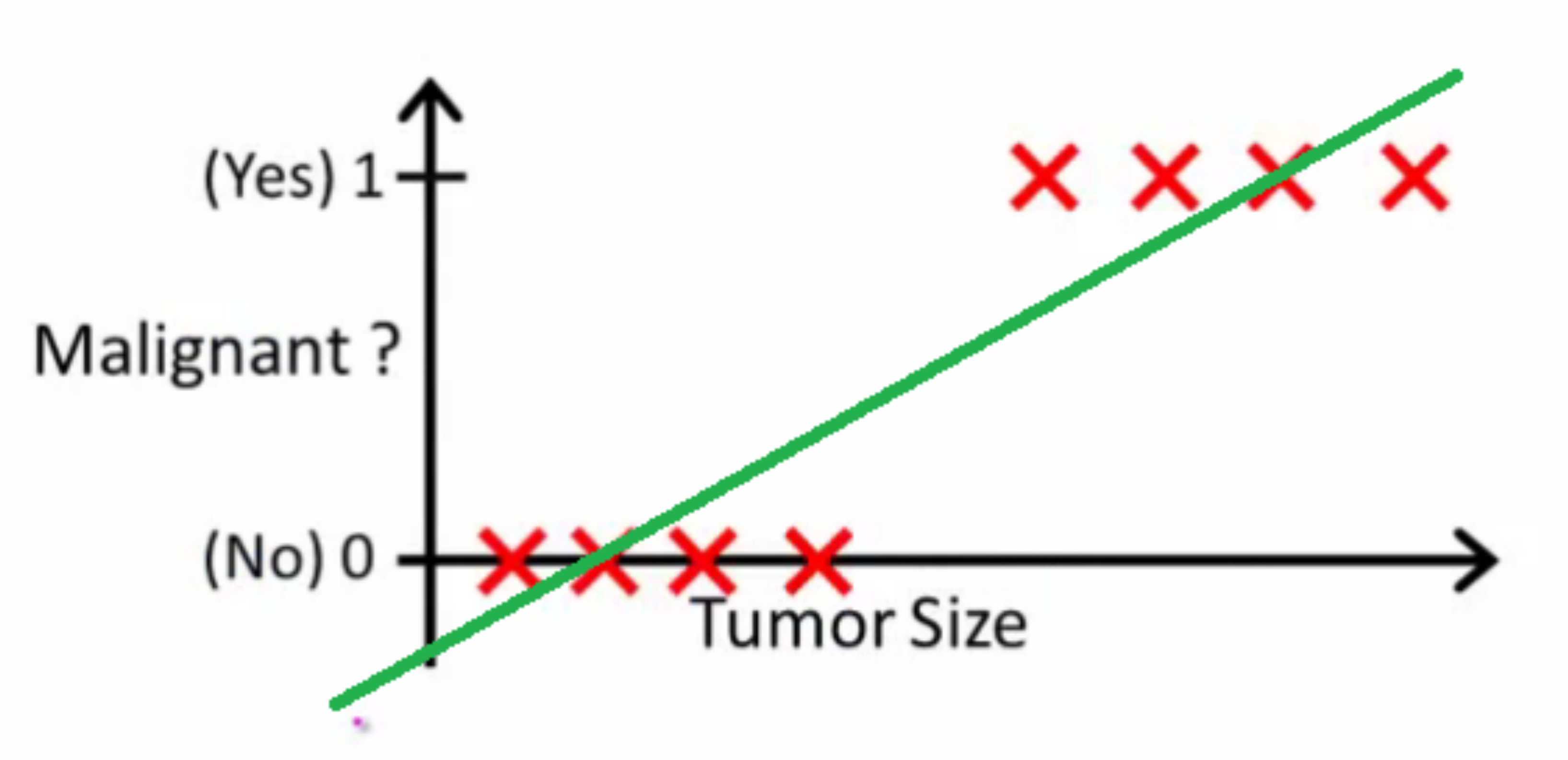
classificação

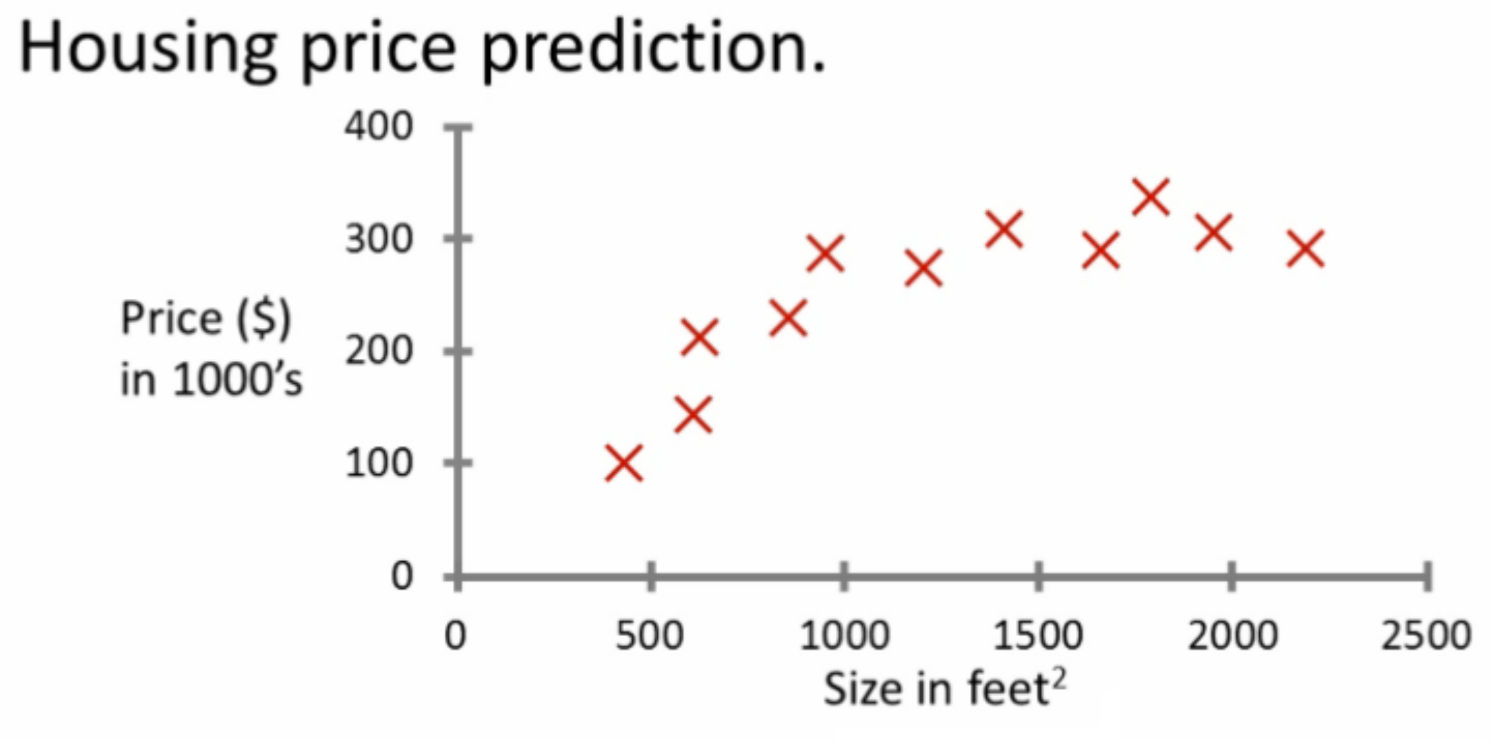
supervisionado

não supervisionado

regressão

REGRESSÃO OU CLASSIFICAÇÃO?

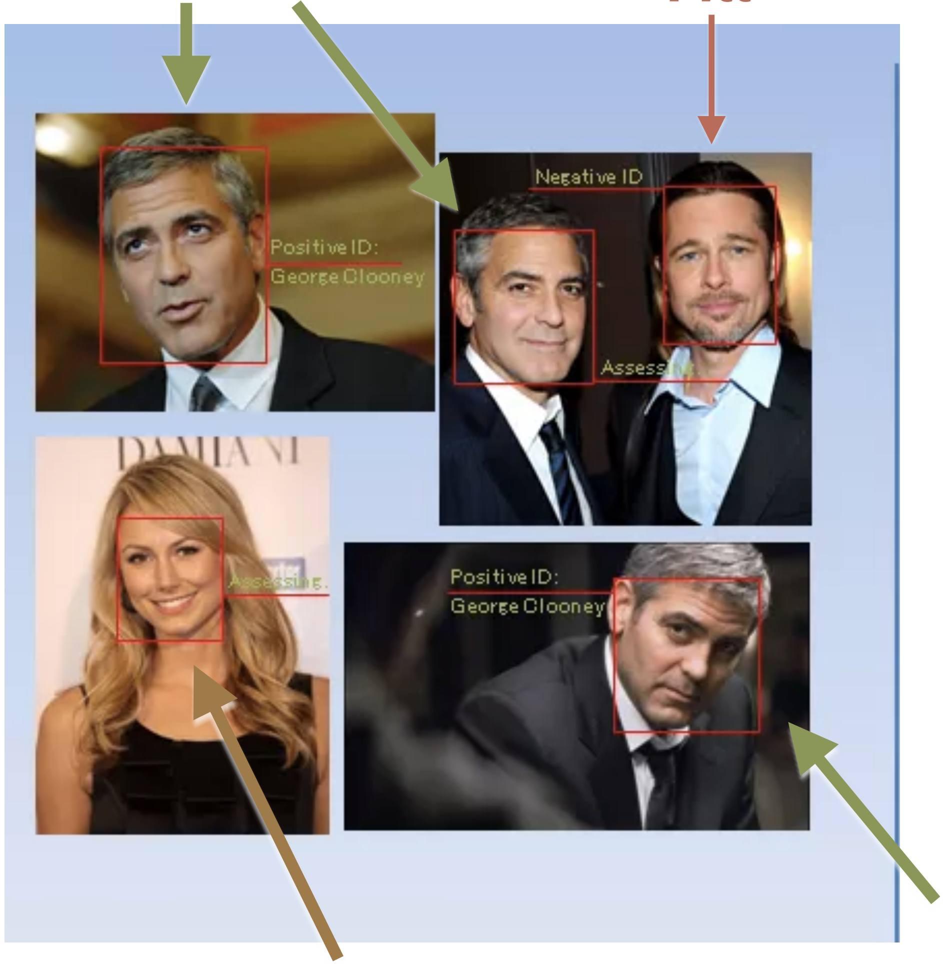




Housing price prediction.

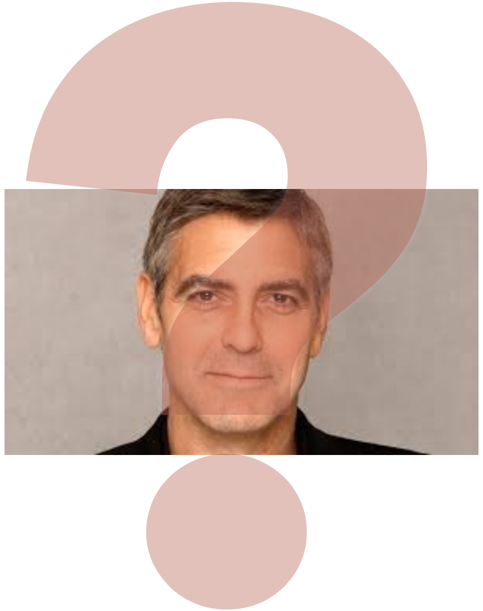
George Clooney

Brad Pitt

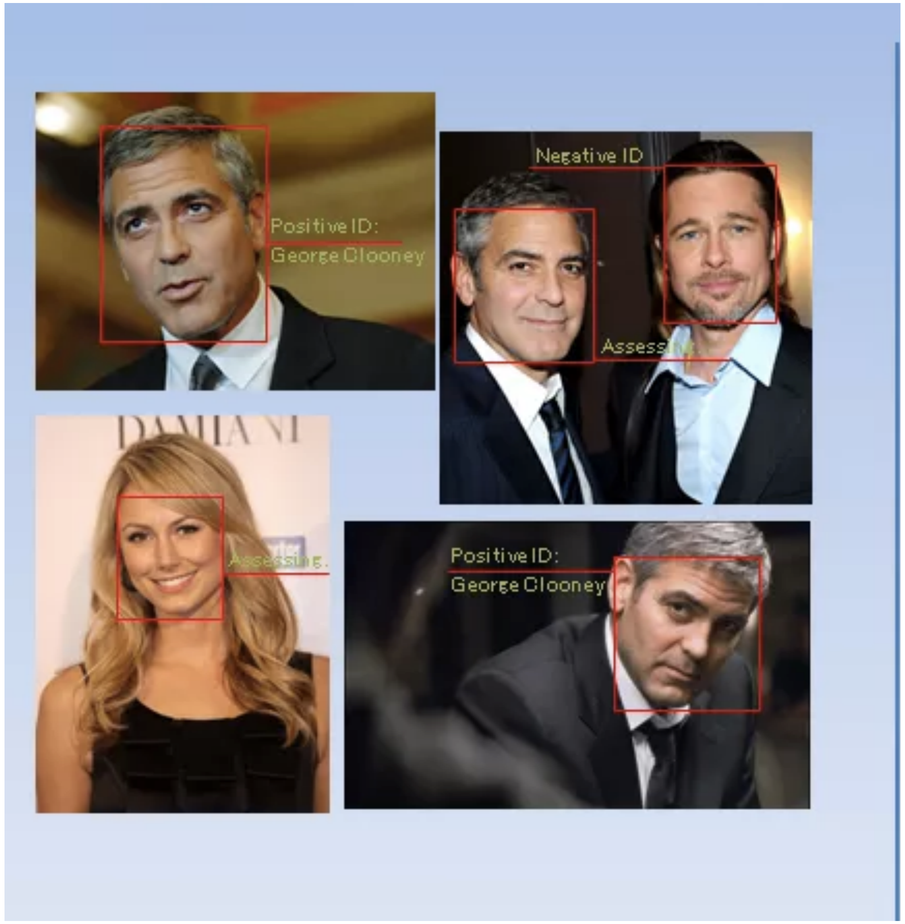


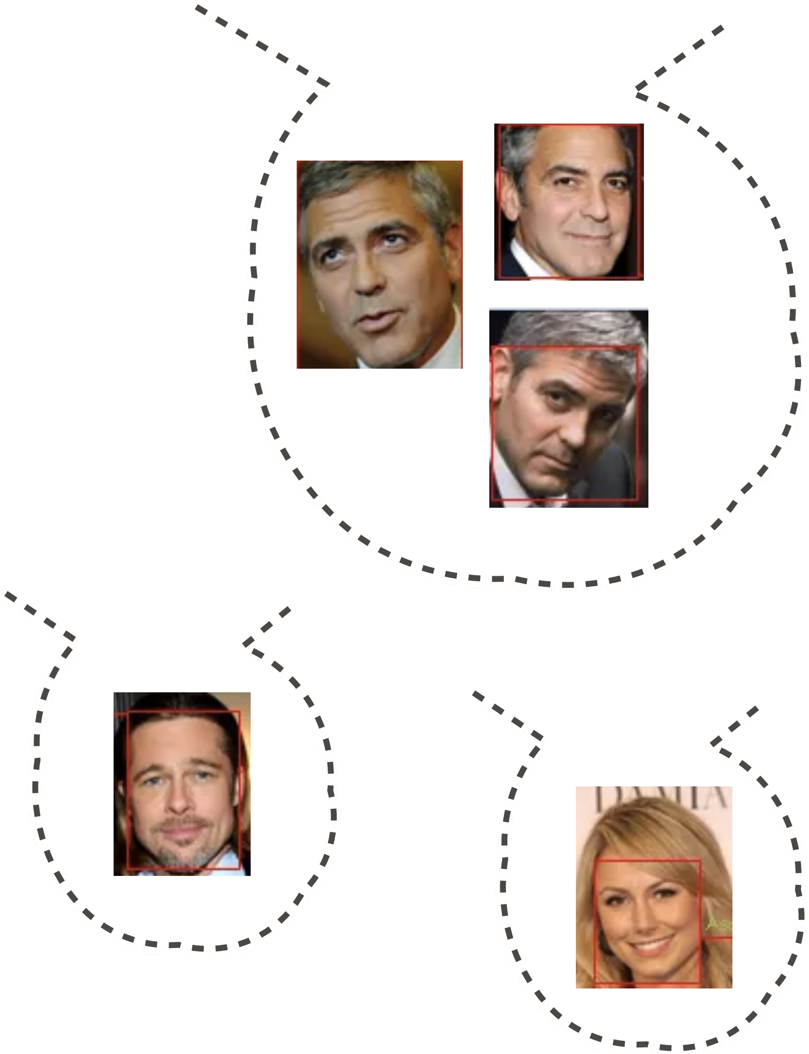
George Clooney

Stacy Kleibler



SUPERVISIONADA OU NÃO-SUPERVISIONADA?





Mão NA MASSA

INSTALAÇÃO ANACONDA

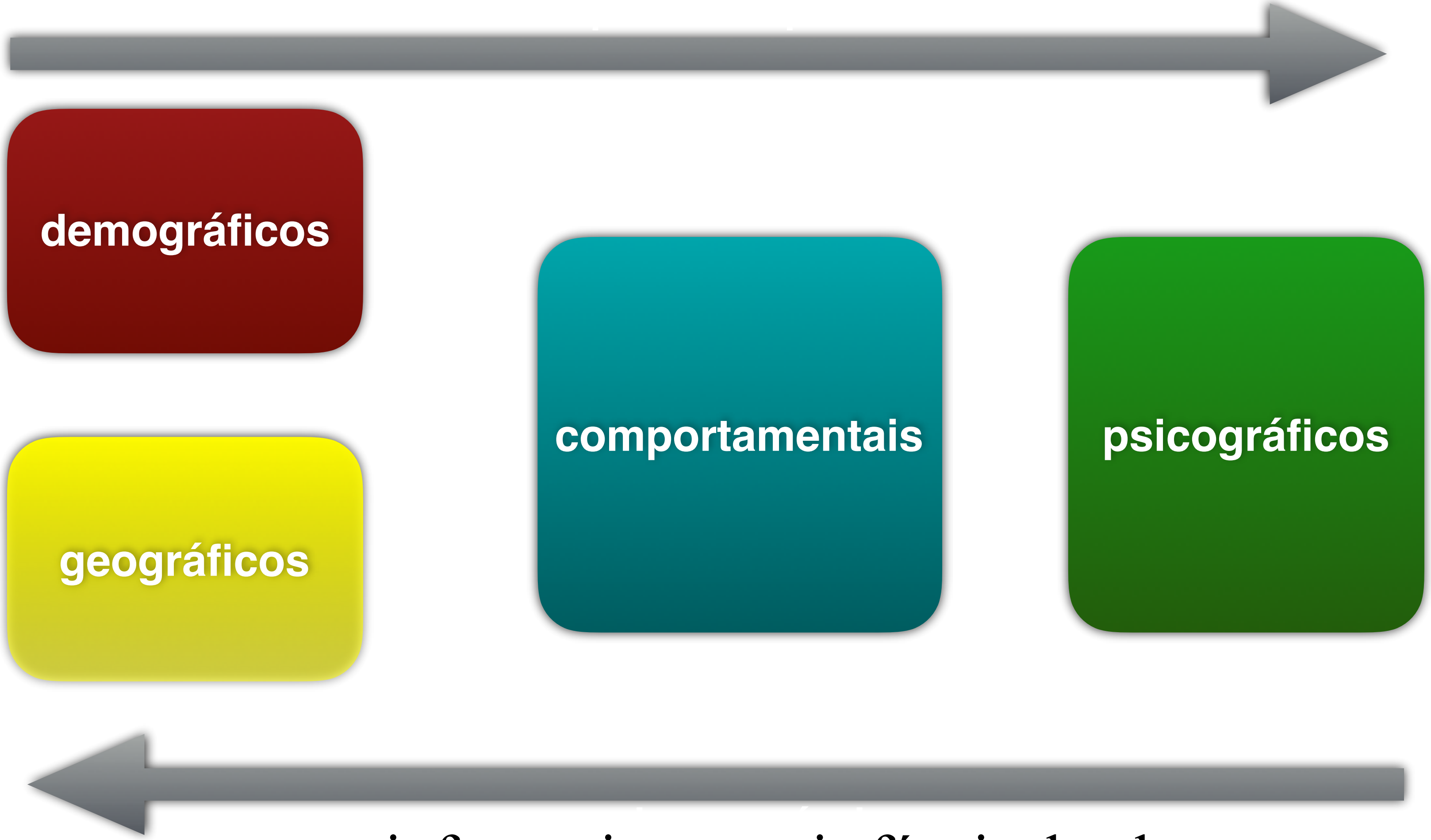
https://www.continuum.io/

IMPORTANTE: escolher versão para Python3

COISAS DE PYTHON...

- características
- tipos
- funções e classes
- principais estruturas
- macetes:
  - list comprehension e dict comprehension
  - generators
  - lambda
  - sintaxes compactas e úteis
  - medindo tempo de execução

mais informativos, mais difíceis de obter



demográficos

geográficos

comportamentais

psicográficos

menos informativos, mais fáceis de obter