Introdução a Ciencia de Dados

Hitoshi Nagano, Ph.D.

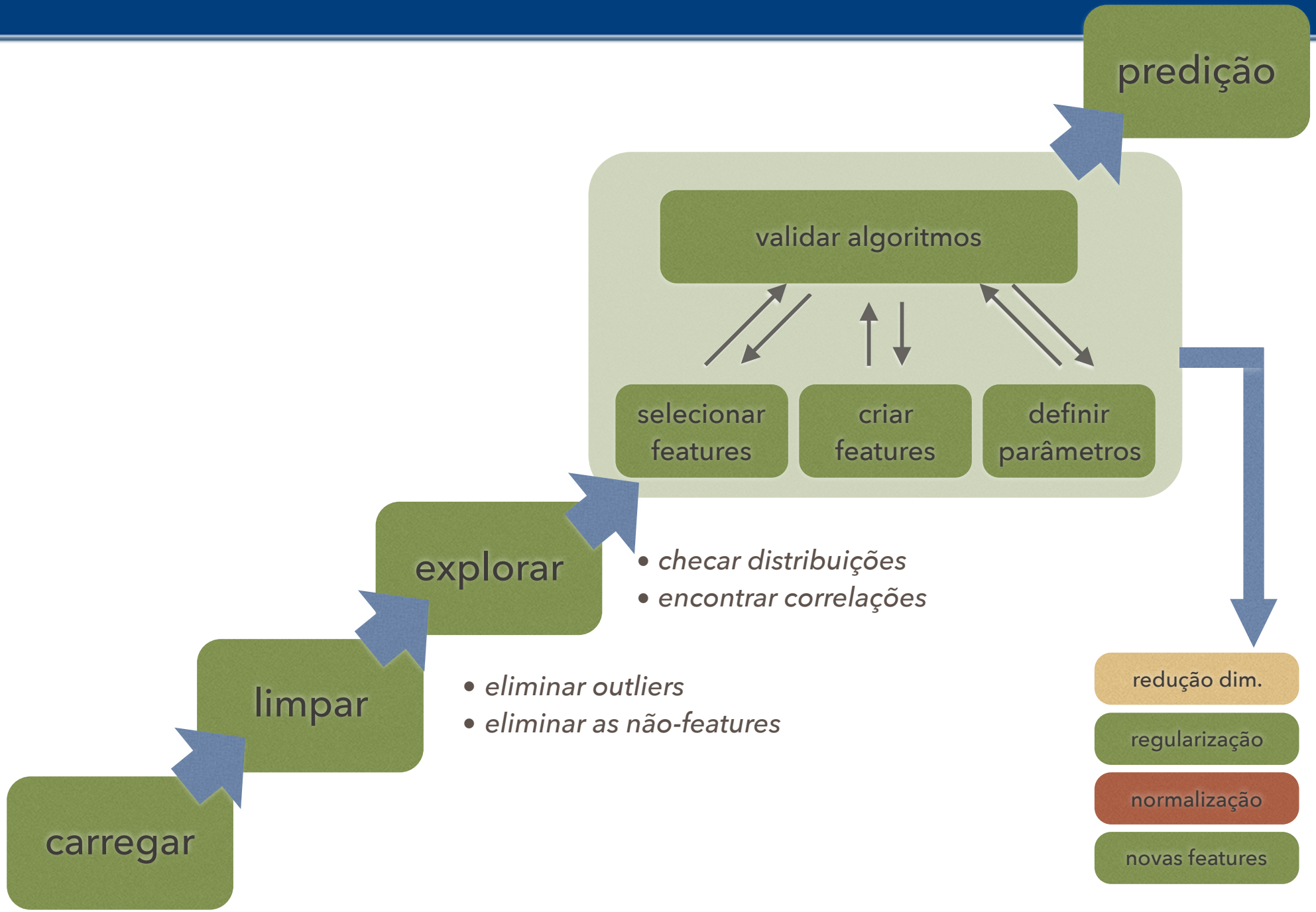
FGV MANAGEMENT

Workflow de projetos em ciencia de dados

FGV MANAGEMENT

Workflow

FGV MANAGEMENT



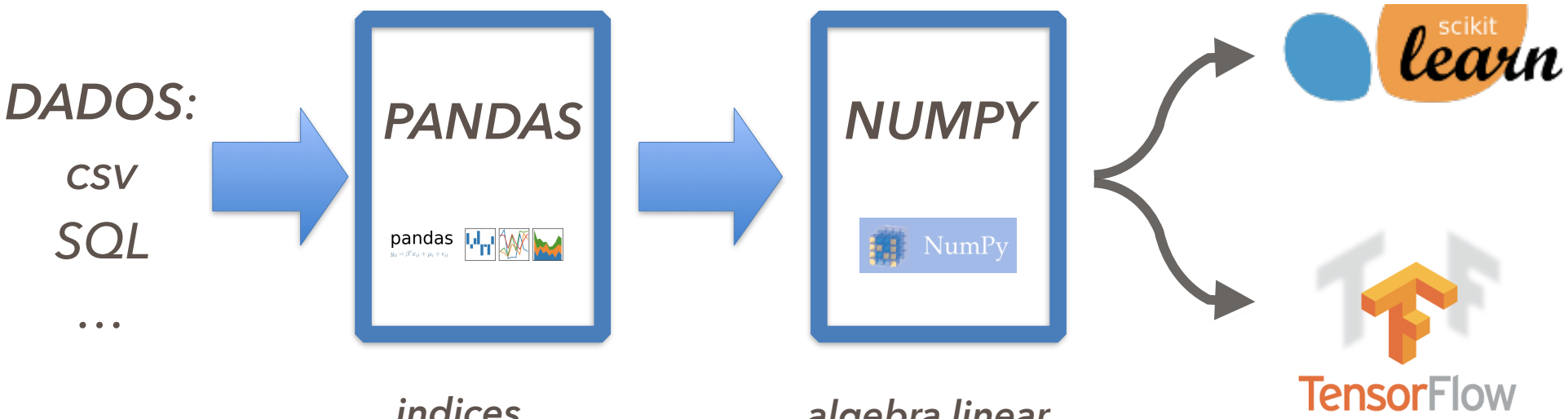
Roteiro para hoje:
limpeza e exploração

FGV MANAGEMENT

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

PANDAS, NUMPY E SCIKIT-LEARN

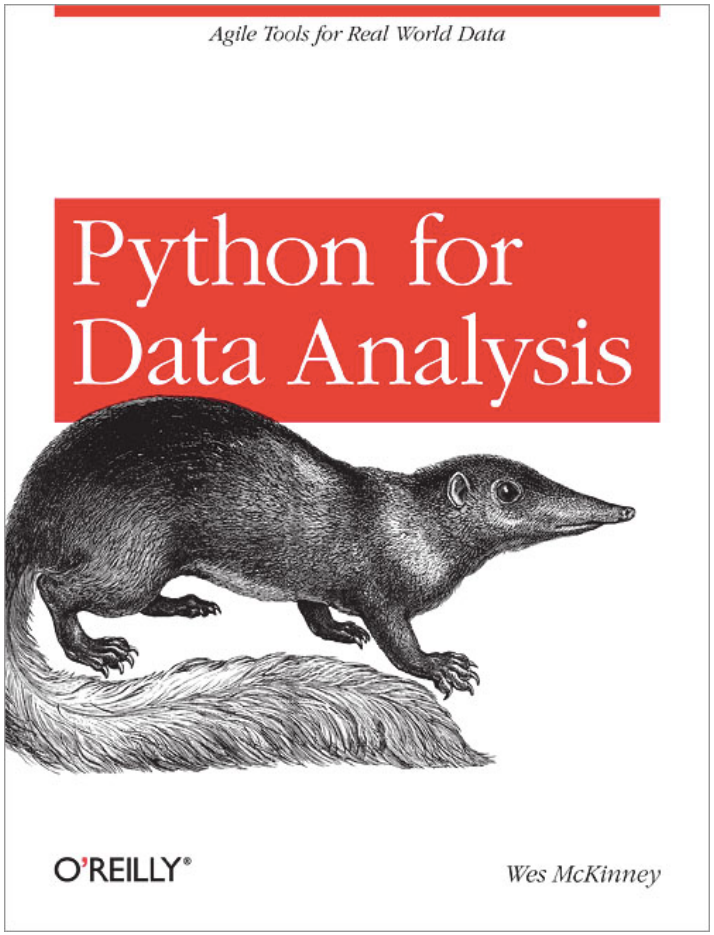
FGV MANAGEMENT



indices
slicing, indexing
limpeza
get_dummies
merge, concat
groupby

algebra linear
multiplicacao matrizes
prod. interno
transform. linear

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



FGV MANAGEMENT

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

⇒ qualidade dos dados (lembrar GIGO)

⇒ melhora da performance

VARIÁVEIS E DISTRIBUIÇÕES

FGV MANAGEMENT

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

Exemplos de distribuições numéricas

FGV MANAGEMENT

- ➡ discreta
  - bernoulli
  - binomial
  - poisson
- ➡ contínua
  - normal
  - exponencial

- paramétricas
- não-paramétricas

FGV MANAGEMENT

... univariada:

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

... multivariada

- covariância
- correlação

OUTLIERS E IMPUTAÇÃO

FGV MANAGEMENT

causas

FGV MANAGEMENT

- eventos estranhos, combinação de eventos — analisar, entender
- sensores quebrados — ignorar
- entrada manual, digitação errada — ignorar, re-inserir
- erros no processamento — ignorar, corrigir

- treinar
- verificar as observações com maior erro residual
- retirar essas observações ≈10%
- rodar de novo

imputação

FGV MANAGEMENT

- valor não disponível (NaN, NULL)
- valor claramente errado, equiv. a não disponível

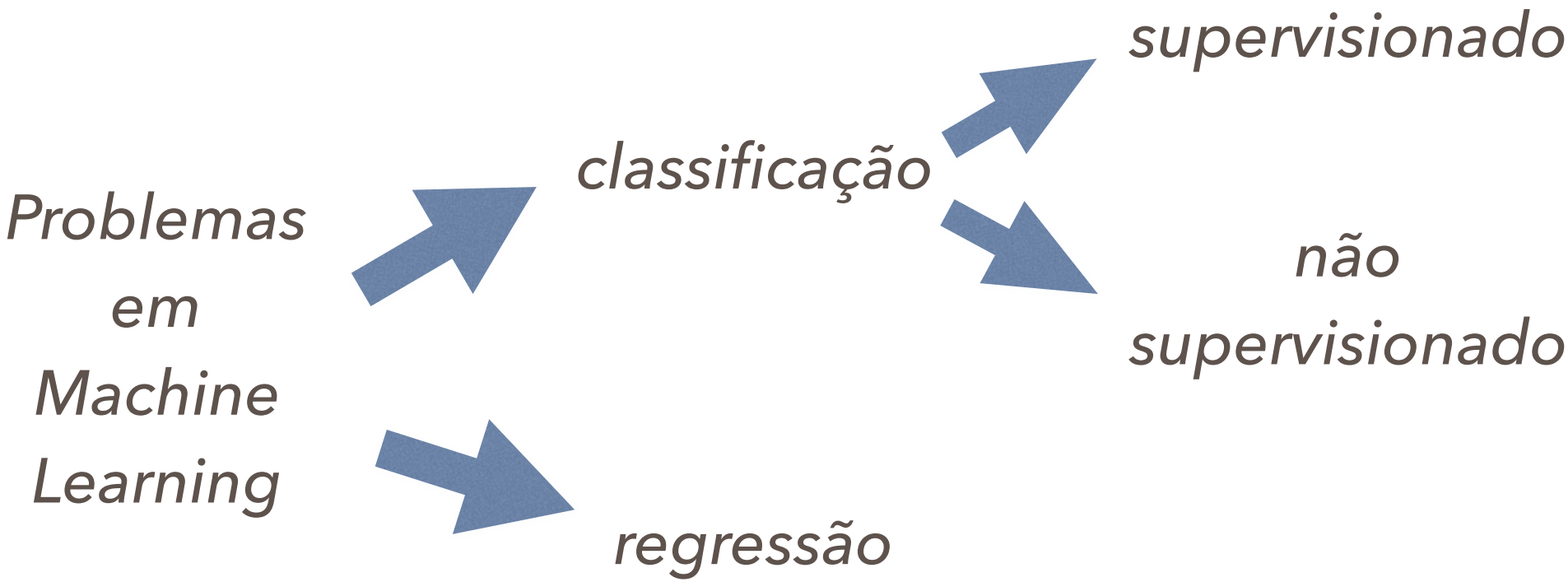
- substituição por um valor unico, pela média,
  ou outro critério

TIPOS DE PROBLEMAS

FGV MANAGEMENT

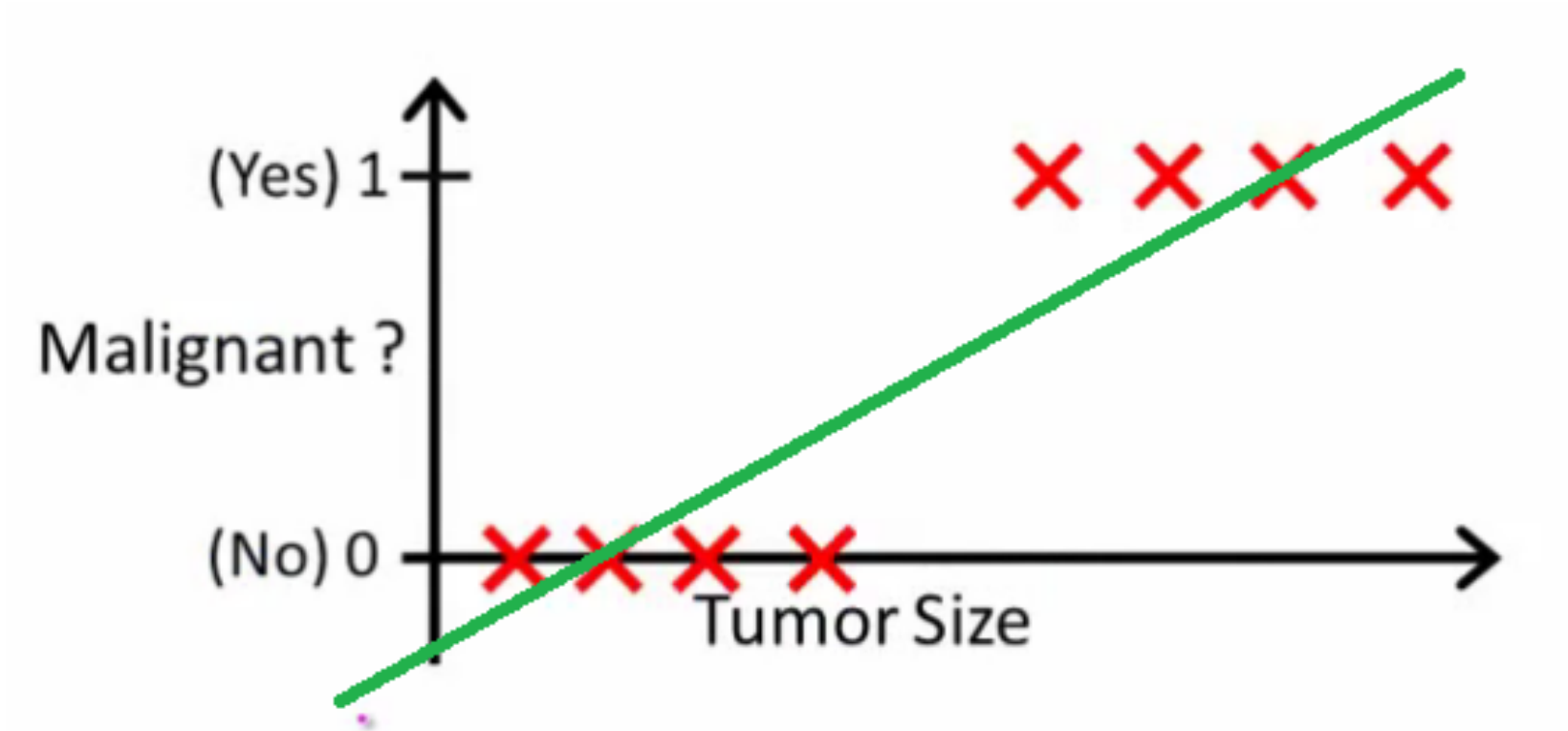
CLASSIFICAÇÃO E REGRESSÃO

FGV MANAGEMENT



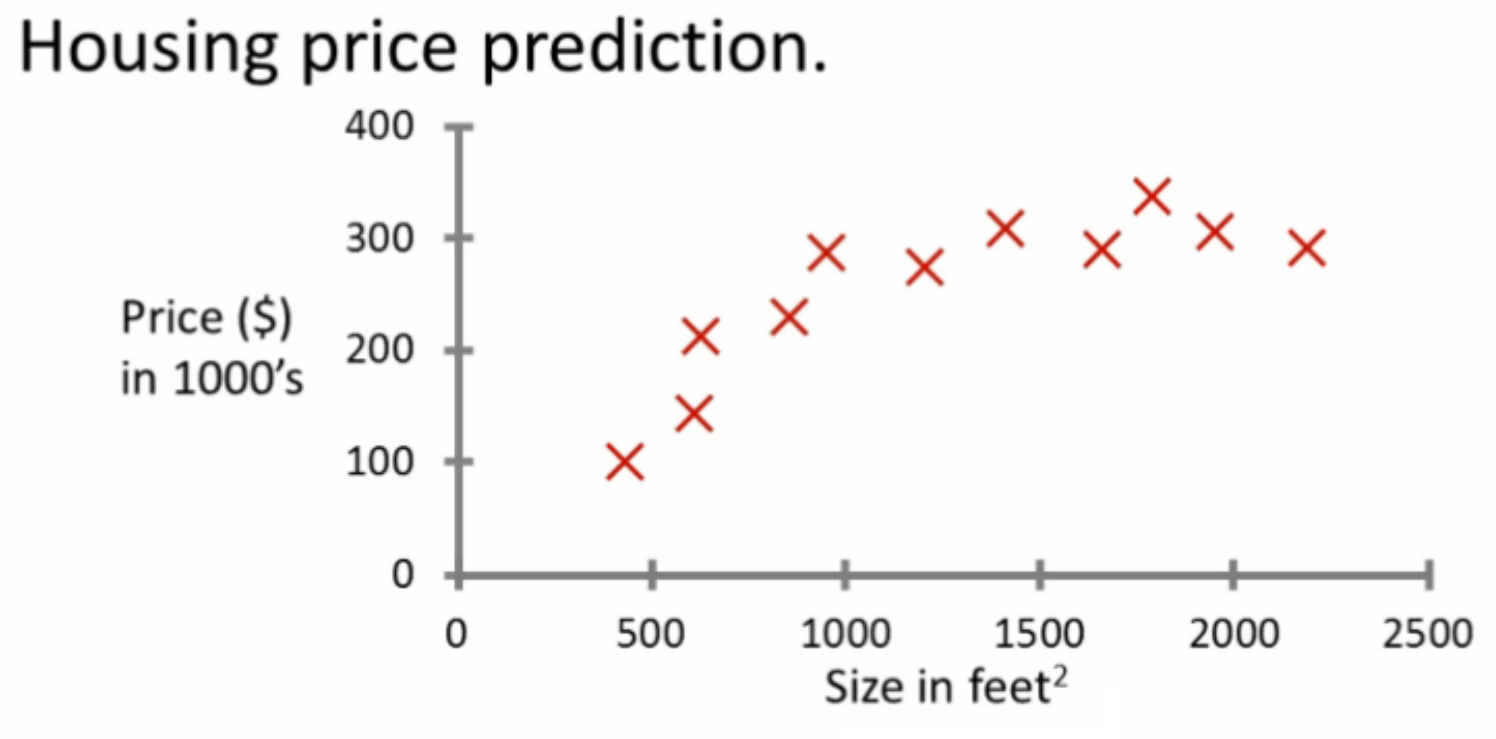
REGRESSÃO OU CLASSIFICAÇÃO?

FGV MANAGEMENT



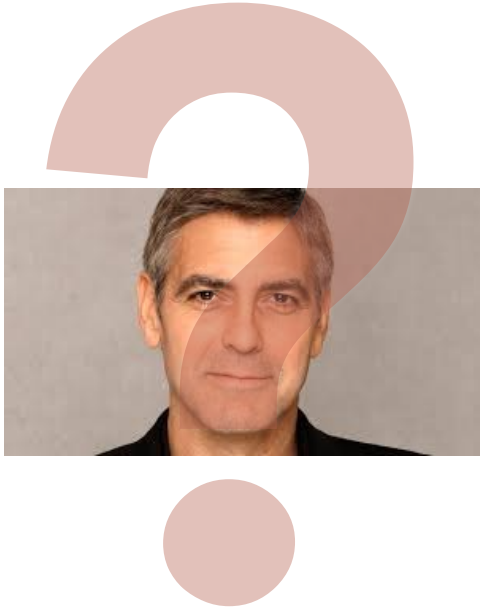
REGRESSÃO OU CLASSIFICAÇÃO?

FGV MANAGEMENT



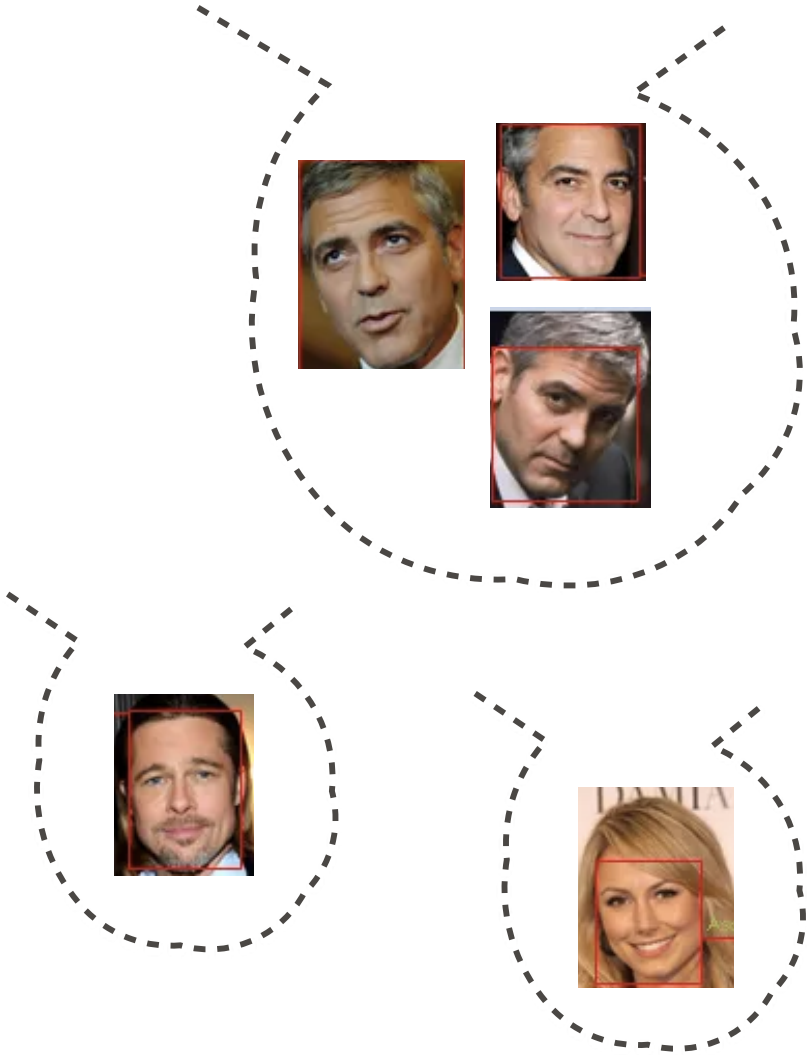
SUPERVISIONADA OU NÃO-SUPERVISIONADA?

FGV MANAGEMENT



SUPERVISIONADA OU NÃO-SUPERVISIONADA?

FGV MANAGEMENT



INSTALAÇÃO ANACONDA

FGV MANAGEMENT

https://www.continuum.io/

IMPORTANTE: escolher versão para Python3

COISAS DE PYTHON...

FGV MANAGEMENT

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