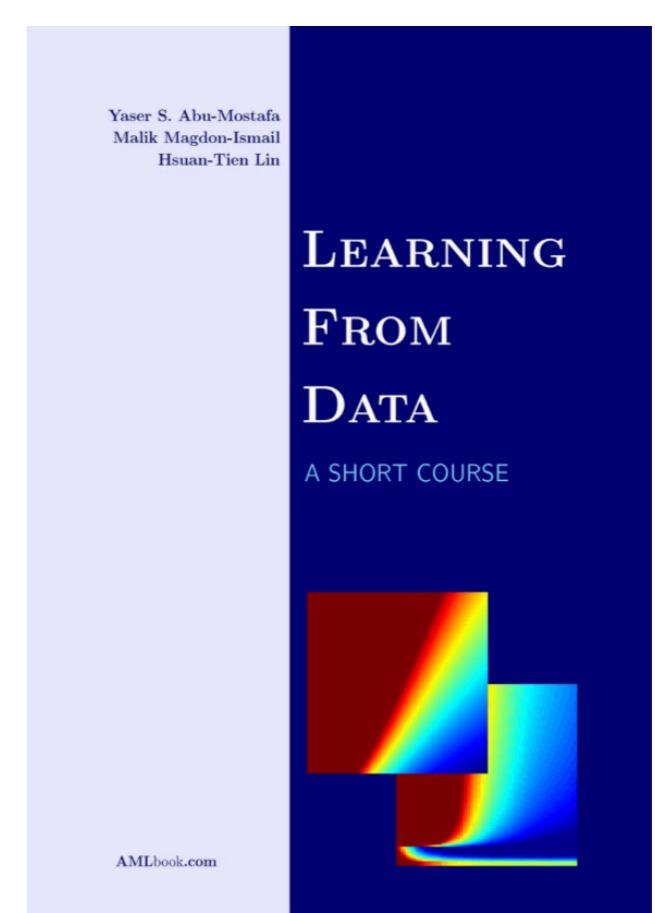
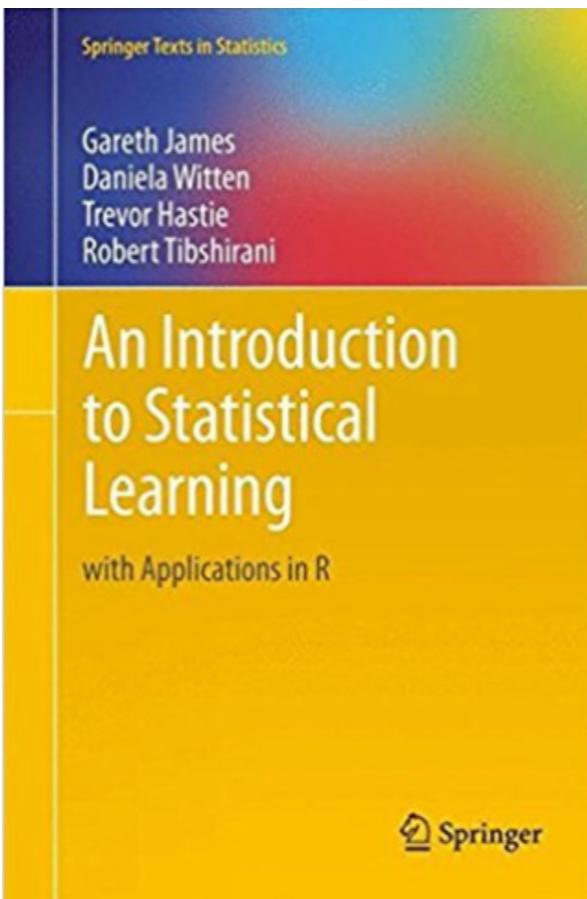
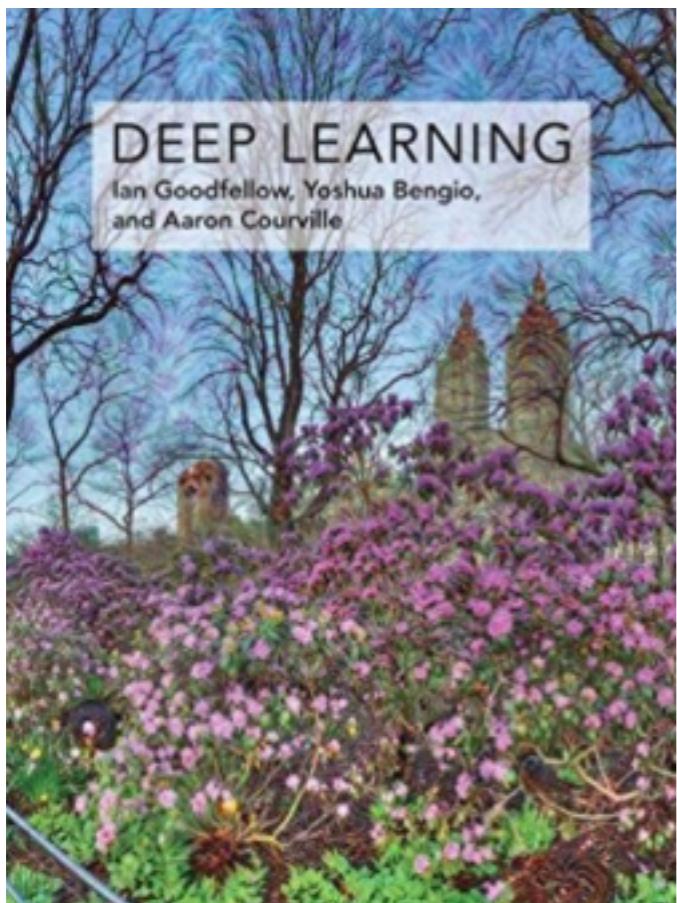
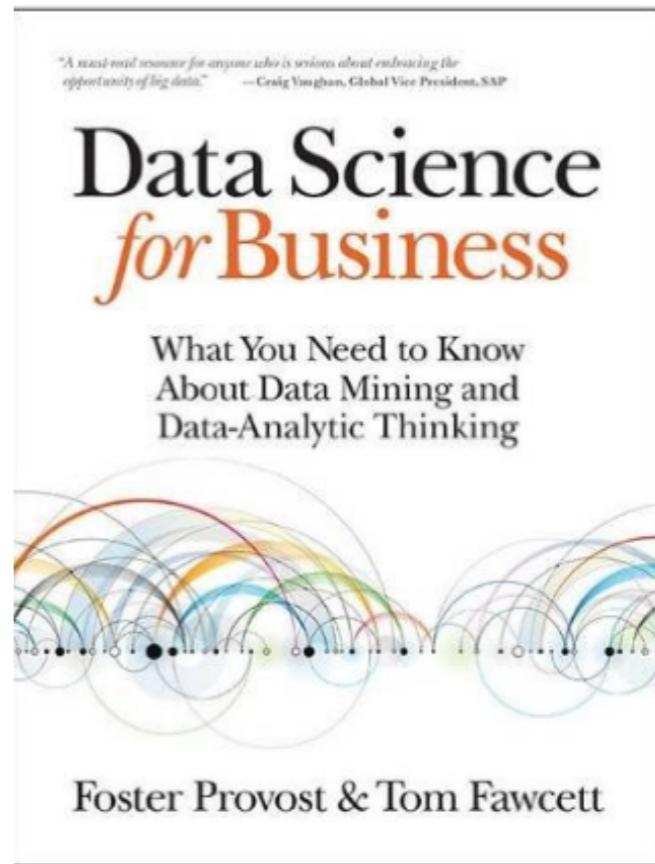
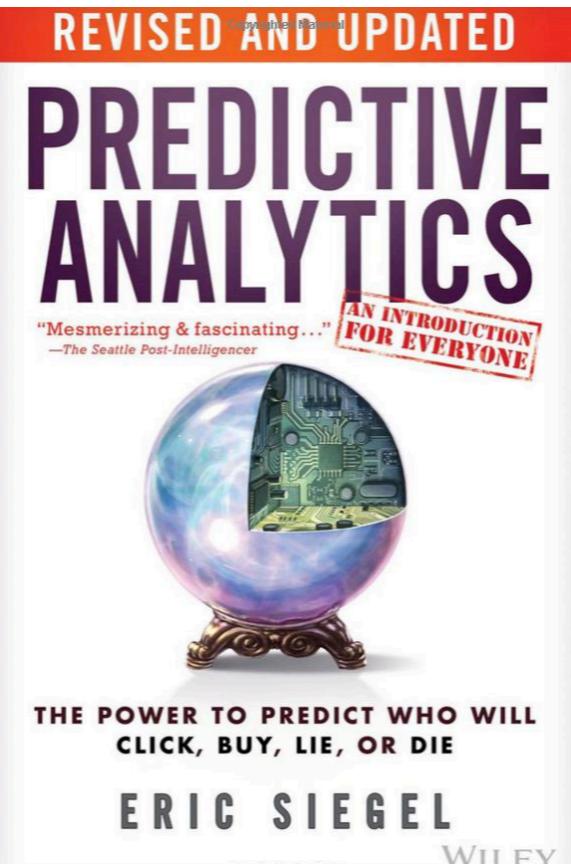
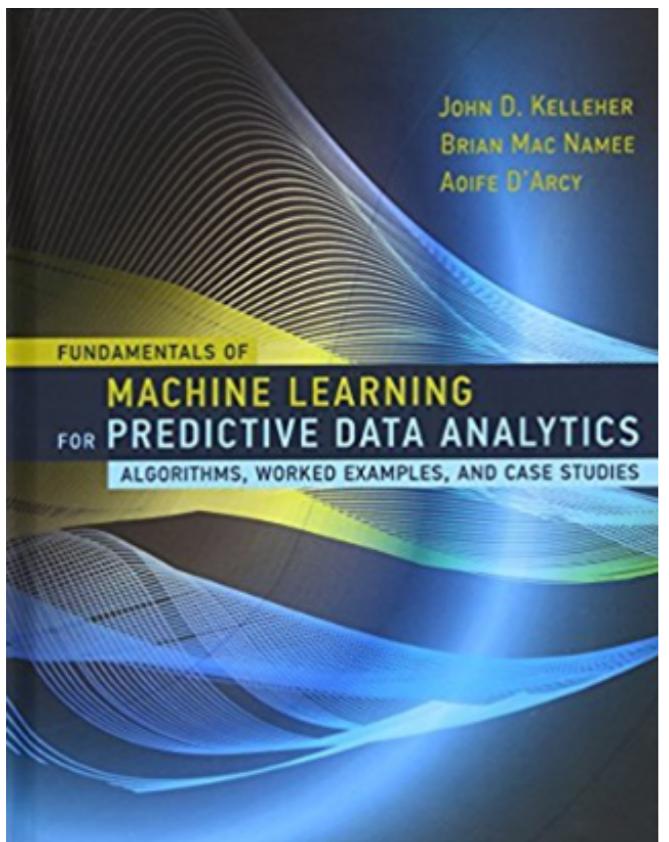


Workshop Intensivo de Aplicações Modernas de Ciência de Dados com Machine Learning

Segundo dia

Paulo Cysne Rios Jr.

Livros Recomendados



Site com toda a documentação do workshop

- Este meu site terá toda a documentação do workshop.
- No seguinte endereço:
<https://github.com/paulocr2/Workshop-App-ML-UNI7>
- Basta clicar em download ou usar no terminal (Mac) / command shell (Windows) teclar *git clone <endereço>*
- Na manhã do dia seguinte estará sempre disponível neste site todo o material do dia anterior (slides, código de programação, links recomendados ou mencionados)

Baixar diretamente ou usando o git

- Para baixar basta clicar em **download**.
- Se você tiver o software **git** instalado no seu computador ainda é mais fácil de baixar. Se a maioria de vocês não sabe como usar git (uma ferramenta muito útil), não tem problema. Discutirei mais tarde no curso o básico do git. Por enquanto basta baixar clicando em download, caso não conheça o git.
- Para quem já quer aprender mais sobre o git, este é um excelente e simples guia:
<http://rogerdudler.github.io/git-guide/>

Code

Issues 0

Pull requests 0

Projects 0

Wiki

Insights

Settings

Documentação e programas do Workshop em Aplicações Modernas de Machine Learning dado na UNI7, Fortaleza

Add topics

55 commits

1 branch

0 releases

1 contributor

Branch: master ▾

New pull request

Create new file

Upload files

Find file

Clone or download ▾

paulocr21 sta learning bk added

Latest commit fed9118 2 minutes ago

day1

divididos por
dia de aula

day2

diretório com
as images
usadas

img

README.md

changes

sta learning bk added

img added

Update README.md

28 minutes ago

2 minutes ago

4 minutes ago

10 hours ago

Clique aqui para baixar
os arquivos neste
repositório

Basta clicar em um arquivo do tipo md para ver seu conteúdo

Workshop-App-ML-UNI7

Workshop em Aplicações Modernas de Machine Learning na UNI7, Fortaleza

Paulo Rios, Outubro de 2017

Toda a documentação e programas do workshop. Ao final de cada dia os slides, links e programas daquele dia estarão aqui.

Os slides e os links do primeiro dia estão já disponíveis.

Usando Git para atualizar o repositório

```
ZireImac:day1 Zireimac$ ls ← 1
Census Data Analytics - Workshop - Paulo Rios.key
Census Data Analytics - Workshop - Paulo Rios.pdf
Census Data Analytics - Workshop - Paulo Rios.pptx
Links do primeiro dia.md
Workshop ML - day 1.key
Workshop ML - day 1.pdf
Workshop ML - day 1.pptx
ZireImac:day1 Zireimac$ rm *.key ← 2
ZireImac:day1 Zireimac$ rm *.pptx
ZireImac:day1 Zireimac$ ls
Census Data Analytics - Workshop - Paulo Rios.pdf
Links do primeiro dia.md
Workshop ML - day 1.pdf
ZireImac:day1 Zireimac$ cd ..
ZireImac:Workshop-App-ML-UNI7 Zireimac$ ls
README.md      day1          day2          img
ZireImac:Workshop-App-ML-UNI7 Zireimac$ cd day2
ZireImac:day2 Zireimac$ ls
links_day2.md
ZireImac:day2 Zireimac$ cd ..
ZireImac:Workshop-App-ML-UNI7 Zireimac$ git add day1 ← 1
ZireImac:Workshop-App-ML-UNI7 Zireimac$ git commit -m "only PDF version"
[master 8f027e1] only PDF version
 4 files changed, 0 insertions(+), 0 deletions(-)
 delete mode 100644 day1/Census Data Analytics - Workshop - Paulo Rios.key
 delete mode 100644 day1/Census Data Analytics - Workshop - Paulo Rios.pptx
 delete mode 100644 day1/Workshop ML - day 1.key
 delete mode 100644 day1/Workshop ML - day 1.pptx
ZireImac:Workshop-App-ML-UNI7 Zireimac$ git push -u origin master ← 2
Counting objects: 3, done.
Delta compression using up to 8 threads.
Compressing objects: 100% (3/3), done.
Writing objects: 100% (3/3), 368 bytes | 0 bytes/s, done.
Total 3 (delta 1), reused 0 (delta 0)
remote: Resolving deltas: 100% (1/1), completed with 1 local object.
To https://github.com/paulocr2/Workshop-App-ML-UNI7.git
 fed9118..8f027e1 master -> master
Branch master set up to track remote branch master from origin.
```

git add changed

git commit -m “nome”

git push -u origin master

Para aprender Git

checkout a repository

create a working copy of a local repository by running the command

```
git clone /path/to/repository
```

when using a remote server, your command will be

```
git clone username@host:/path/to/repository
```

<http://rogerdudler.github.io/git-guide/>

workflow

your local repository consists of three "trees" maintained by git. the first

one is your **Working Directory** which holds the actual files. the

second one is the **Index** which acts as a staging area and finally the

HEAD which points to the last commit you've made.

Conjunto de dados

Conjuntos de dados do workshop

- O diretório dir do repositório do nosso workshop tem conjunto de dados reais e de domínio público que usaremos no workshop.
- Acabei de colocar lá 2 conjuntos de dados que usaremos em breve:
 - **Satisfação de vida nos países da OECD**
 - **Preços de imóveis nos EUA**

[paulocr2 / Workshop-App-ML-UNI7](#)

Code Issues Pull requests Projects Wiki Insights Settings

Documentação e programas do Workshop em Aplicações Modernas de Machine Learning dado na UNI7, Fortaleza

Add topics

57 commits 1 branch 0 releases 2 contributors

Branch: master New pull request Create new file Upload files Find file Clone or download

paulocr21 data added Latest commit 356c6db 10 minutes ago

data data added 10 minutes ago

day1 only PDF version an hour ago

day2 sta learn

img img add

README.md Update

README.md

Workshop-App-ML-UNI7

Workshop em Aplicações Modernas de Machine Learning

Paulo Rios, Outubro de 2017

Toda a documentação e programas do workshop.

paulocr2 / Workshop-App-ML-UNI7

Code Issues Pull requests Projects

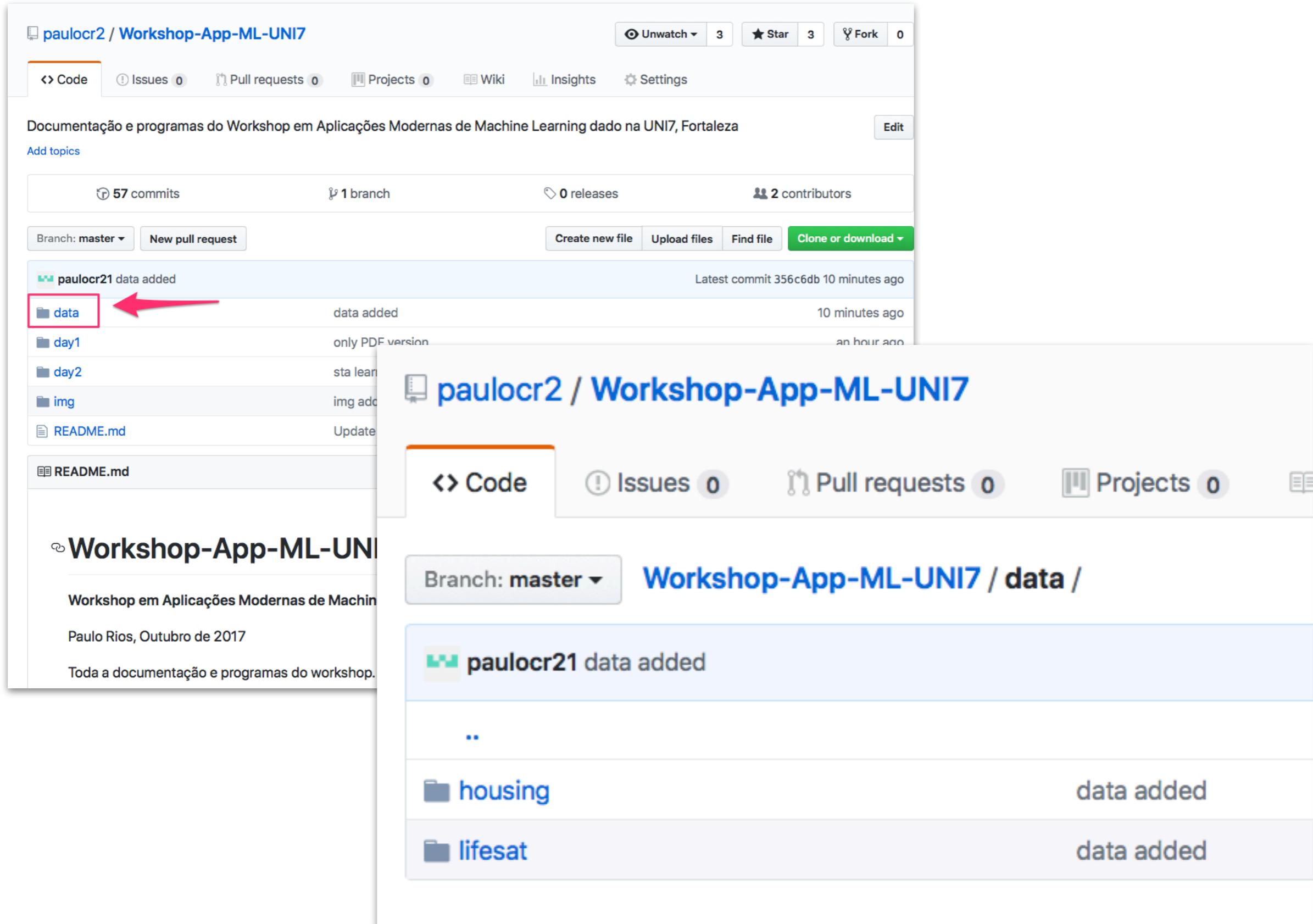
Branch: master **Workshop-App-ML-UNI7 / data /**

paulocr21 data added

..

housing data added

lifesat data added



paulocr2 / Workshop-App-ML-UNI7

Code

Issues 0

Pull requests 0

Projects 0

Wiki

Branch: master

Workshop-App-ML-UNI7 / data / housing /

Paulocr21 data added

..

README.md

data added

housing.csv

data added

housing.tgz

README.md

California House

Source

This dataset is a modified version of the
[http://www.dcc.fc.up.pt/~ltorgo/Regression/]
from the StatLib repository (wh

paulocr2 / Workshop-App-ML-UNI7

Code

Issues 0

Pull requests 0

Projects 0

Wiki

Insights

Branch: master

Workshop-App-ML-UNI7 / data / lifesat /

Paulocr21 data added

..

README.md

data added

gdp_per_capita.csv

data added

oecd_bli_2015.csv

data added

README.md

Módulo I – Fundamentos

- Introdução à Machine Learning
 - História da Inteligência Artificial
 - Data Science
 - Big Data, real-time data streaming
 - Machine Learning
 - Deep Learning
 - O cenário atual de Data Science
- O impacto de Machine Learning em Data Science
 - Uma nova onda de investimentos
 - As aplicações mais comuns
 - As ferramentas mais poderosas

Módulo II – Projetos de Machine Learning

- A natureza dos projetos de Machine Learning
 - Custos e benefícios estratégicos
 - Sponsorship
 - Life cycle iterativo
 - Etapas
- Os principais desafios
 - Dados de treinamento, validação e teste
 - Atributos irrelevantes
 - Limpeza de dados
 - Lidando com valores nulos
 - Lidando com atributos categóricos e textos
 - Lidando com outliers
 - Lidando com muitas dimensões, redução de dimensões
 - Visualizando os dados
 - Overfitting
 - Underfitting
 - Bias/Variance Tradeoff

Questões de Ontem

- Qual a diferença entre **Estatística e Ciência de Dados?**
- Qual é a diferença entre **Ciência de Dados e Machine Learning?**
- O que é **Big Data**? Qual a relação dele com **Ciência de Dados?**

Estatística e Ciência de Dados

- **Estatística descritiva:** summarizar dados (média, máximo, mínimo, desvio padrão, etc.)
- **Estatística inferential:** inferir dados de uma população (o todo) a partir de dados de uma amostragem da população (a parte).
- **Ciência de dados:** de natureza explorativa, busca padrões, tendências e relacionamentos nos dados, visando uma descoberta de sua natureza, com uma finalidade *preditiva*.

Mas a análise regressiva não é preditiva em estatística?

- Em estatística ou econometria, regressão linear é uma equação para se **estimar** a condicional (**valor esperado**) de uma variável y , dados os valores de algumas outras variáveis x . (Fonte: Wikipedia).
- Na estatística, a modelagem é feita partindo do ponto de que se tem **uma amostra** e dela se vai **inferir** para a população (o todo).
- Trabalha-se com **intervalos de confiança** (para indicar a variação esperada dentro dos resultados obtidos) e **p-values** (para estimar a chance de ser por puro acaso ou não que chegamos a estes resultados).
- Em Machine Learning **não** se usa intervalos de confiança nem p-values, quando se usa regressão linear e outros modelos clássicos de estatística!

Ciência de Dados e Machine Learning

- **Ciência de Dados** tem o objetivo de modelagem analítica com caráter preditivo e explorativo, procurando reconhecer padrões, tendências, estruturas e relacionamentos. Compreende **todas** as técnicas de pré-processamento, visualização de dados e modelagens analíticas.
- **Machine Learning** é um conjunto de técnicas de modelagem de ciência de dados originárias da área da Ciência da Computação e Inteligência Artificial. Dela brotou **Redes Neuronais** que mais tarde foram aprimoradas em **Deep Learning**.

Ciência de Dados e Machine Learning

- Técnicas de **Machine Learning**: SVM, random forests (árvores aleatórias), bagging, boosting, gradient boosting.
- **Deep Learning** é uma área de Machine Learning com desenvolvimentos recentes.
- Estudaremos algumas técnicas de **estatística inferencial** como **regressão linear** e **regressão logística**. Por que? Elas são muito boas para resolução de modelagens simples. Não se precisa usar um avião quando se pode ir de carro... Veja também o **tradeoff bias variance** adiante.

Qual a diferença entre Machine Learning e Data Mining?

- **Machine Learning** usa em grande parte técnicas de modelagem analítica criadas pela comunidade de Ciência da Computação.
- Uma parte desta comunidade trabalhou em técnicas que se denominaram **Data Mining**.
- Assim, **Machine Learning** usa várias técnicas da comunidade de Ciência da Computação, inclusive **Data Mining**.

Big Data e Ciência de Dados

- **Big Data:** Dados em grande volume que são obtidos em um stream em tempo real e de maneira não estruturada.
- **Ciência de Dados** procura fazer modelagem analítica de caráter preditivo com dados de *qualquer* natureza e *tamanho*.
- Para fazer modelagem analítica de Big Data é necessário uma plataforma que permita trabalhar de maneira **distribuída**: tanto nos dados como no processamento deles.

Big Data

- **Hadoop** e **Spark** oferecem a possibilidade de processamento distribuído.
- **PySpark** e **Scala** (uma linguagem de programação) são bastante usados num ambiente Big Data.
- **Mongo** e **Cassandra** são usados para gerenciamento da armazenagem distribuídas dos dados que podem ser de caráter não estruturado.
- O **Data Engineer** desempenha um papel importante.

Big Data Stack

Data Visualization



Data Store



hadoop

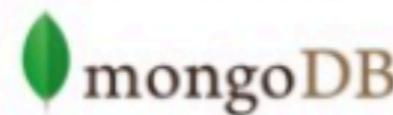
APACHE
HBASE



Cassandra



Hortonworks



mongoDB

cloudera



elastic



Data Integration



STORM



kafka

Spark

ocean

rapidminer



mahout



WEKA

Data Mining

R

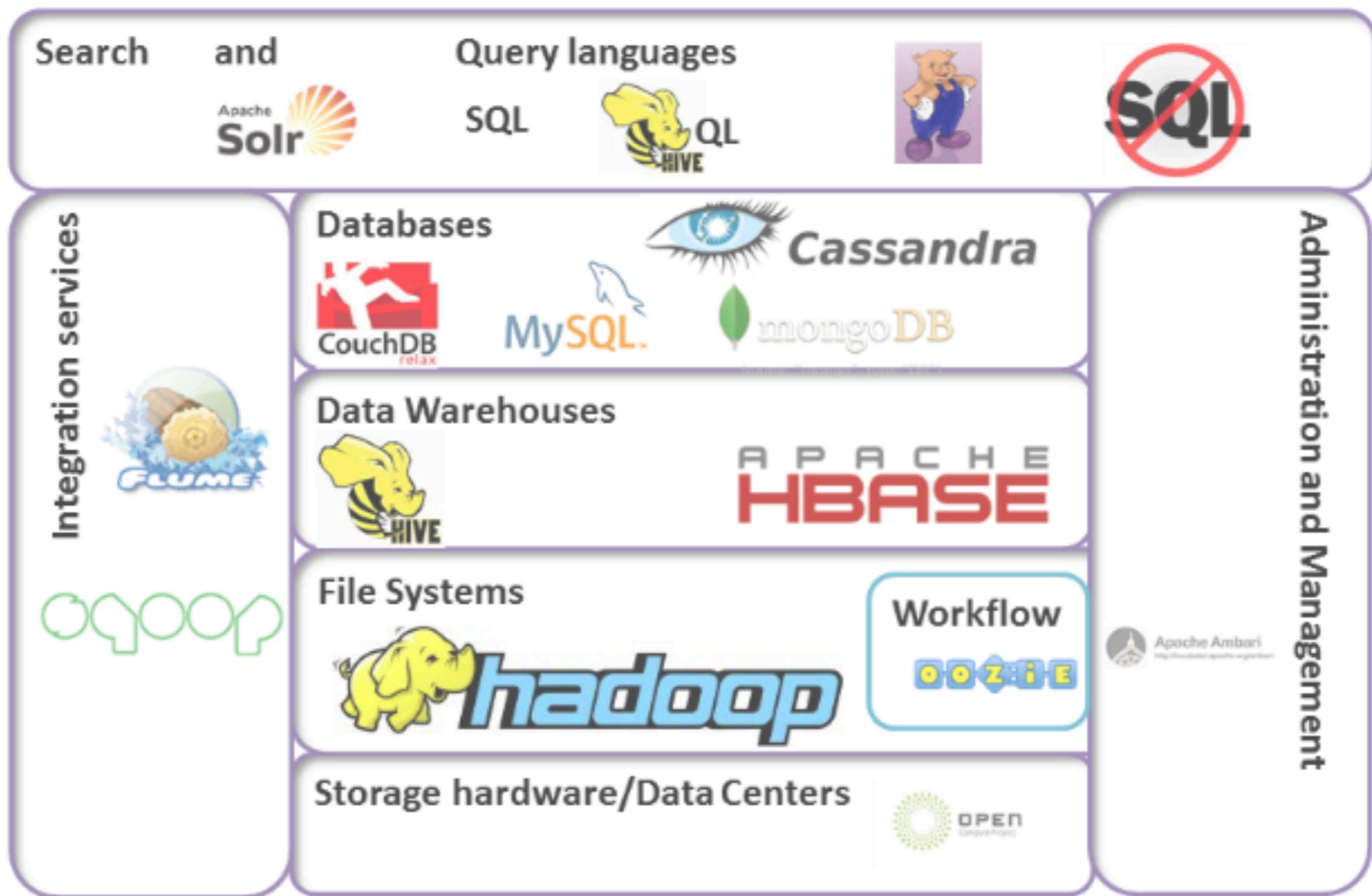
sas

python

BIRT

Data Analytics

Big Data Stack



Projetos de Machine Learning

- Logo seremos capazes de responder as seguintes questões:
- Qual o **objetivo** de um projeto de Machine Learning?
- Quais os **desafios** mais comuns?
- Como se sabe se um projeto de Machine Learning teve **sucesso** ou não?
- Qual a diferença entre Ciência de Dados e Análise Preditiva? É mesma coisa?

Uma nova onda de Investimentos

Python: a linguagem mais popular de 2017



Programming language rankings and image by [IEEE Spectrum](#)

Python: a linguagem mais popular de 2017

- Python vem aumentando nos últimos anos, mas no ano passado já tinha alcançado a terceira posição, enquanto este ano (2017) ganha no ranking com 100.
- Source:
<https://www.extremetech.com/computing/252987-python-tops-list-2017s-popular-programming-languages>

**Quantas vagas de emprego como
cientista de dados
no estado da Califórnia nos EUA
hoje, 10 de outubro de 2017?**

**[https://www.indeed.com/jobs?
q=data+scientist&l=california](https://www.indeed.com/jobs?q=data+scientist&l=california)**



what	where
data scientist	california
job title, keywords or company	city, state, or zip

Quase 5.00 vagas na California!

[Find Jobs](#)[Advanced Job Search](#)**data scientist jobs in California**Sort by: **relevance** - date**Salary Estimate**

- \$60,000 (4252)
- \$80,000 (3444)
- \$95,000 (2664)
- \$110,000 (1842)
- \$125,000 (1060)

Job Type

- Full-time (4986)
- Contract (217)
- Internship (135)
- Part-time (91)
- Temporary (82)
- Commission (7)

Location

- San Francisco, CA (860)
- San Diego, CA (347)
- San Jose, CA (251)
- Los Angeles, CA (250)
- Palo Alto, CA (246)

[more »](#)**Company**

- Walmart eCommerce (291)
- Apple (138)
- Google (73)
- Facebook (72)
- Amazon Corporate LLC (70)

New! Join Indeed Prime - Get offers from great tech companies
Data: 10 de Outubro de 2017

Jobs 1 to 10 of 5,289

Junior Data Analyst - Performance Lens, New Ventures

McKinsey & Company - ★★★★☆ 292 reviews - San Jose, CA

As one of the fastest-growing parts of our firm, New Ventures has more than 1,500 dedicated professionals (including more than 800 analysts and data scientists)...

Desired Experience: Machine Learning, Visual Basic

7 days ago - [save job](#) - [more...](#)

Data Scientist

Hulu - ★★★★☆ 26 reviews - Santa Monica, CA

We are looking for data scientists who are passionate about using data to drive strategy and product recommendations....

Desired Experience: Data Mining, Ruby, Machine Learning, R, Sas, Natural Language Processing, Data Science, Python

11 days ago - [save job](#) - [more...](#)

Data Scientist/Quantitative Analyst, Engineering, University...

Google - ★★★★☆ 1,732 reviews - Mountain View, CA +2 locations

1 year of relevant work experience (i.e., data scientist role), including deep expertise and experience with statistical data analysis such as linear models,...

Desired Experience: R, Android, Sas, Python

1 day ago - [save job](#) - [more...](#)

Be the first to see new **data scientist jobs in california**

My email:

Also get an email with jobs recommended just for me

[Activate](#)

You can cancel email alerts at any time.

Data Scientist salaries in California**\$142,105 per year**

Based on 14,750 salaries



[Data Scientist salaries by company in California](#)



Charlie Walker

Talent Acquisition Manager - America

**** UPDATED REQUIREMENT LIST NYC *****

**Postado no LinkedIn
em Outubro de 2017**

Data Scientist's - Financial/ Software - \$120 - \$600,000

NLP Engineer's - Financial/ Software - \$150 - \$200,000

Machine Learning Engineer - Finance - \$150 -
\$200,000

Data Engineer - Financial - \$140 - \$180,000



**\$ 600 mil dólares por ano =
quase 1.900 reais por ano!**

**Ou seja, quase
160 mil reais por mês!**

Get in touch to find out more!

Email: Charlie.walker@darwinrecruitment.com

Skype: [Charlie.walker@darwinrecruitment.com](skype:Charlie.walker@darwinrecruitment.com)

Twitter: [@CWalkerDR](https://twitter.com/CWalkerDR)

Telephone: (404) 445 4759

**** No C2C/ No C2H ****

Aplicações Mais Comuns

- Fonte:
<http://www.cyzne.com/?p=1233>
- Fonte:
IBM Machine Learning Hub

Aplicações Mais Comuns

Healthcare	Patient diagnosis
Finance	Fraud detection
Manufacturing	Anomaly detection
Retail	Inventory Optimization
Government	Smarter Services
Transportation	Demand Forecasting
Networks	Intrusion Detection
E-Commerce	Recommender Systems
Media	Interaction and Speed
Education	Research Insight

As ferramentas mais poderosas de Machine Learning

Analytic Modeling Structured Data

Classic

Linear Regression
Logistics Regression
Linear Discriminant Analysis (LDA)

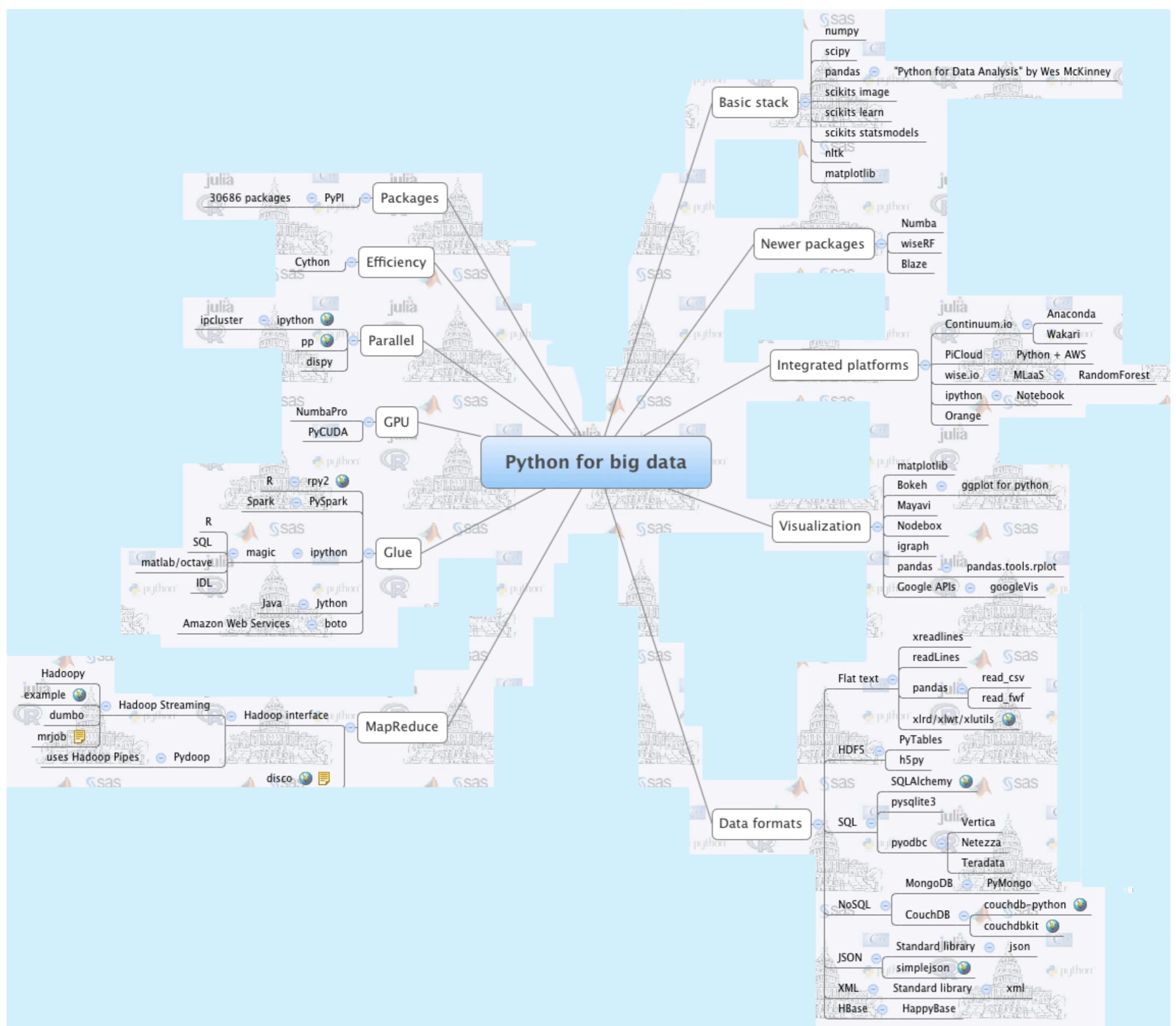
Modern

Support Vector Machines
Decision Trees
Random Forests
Ada Boosting
Gradient Boosting

Deep Learning

TensorFlow

Keras



Dados de treinamento, validação e teste

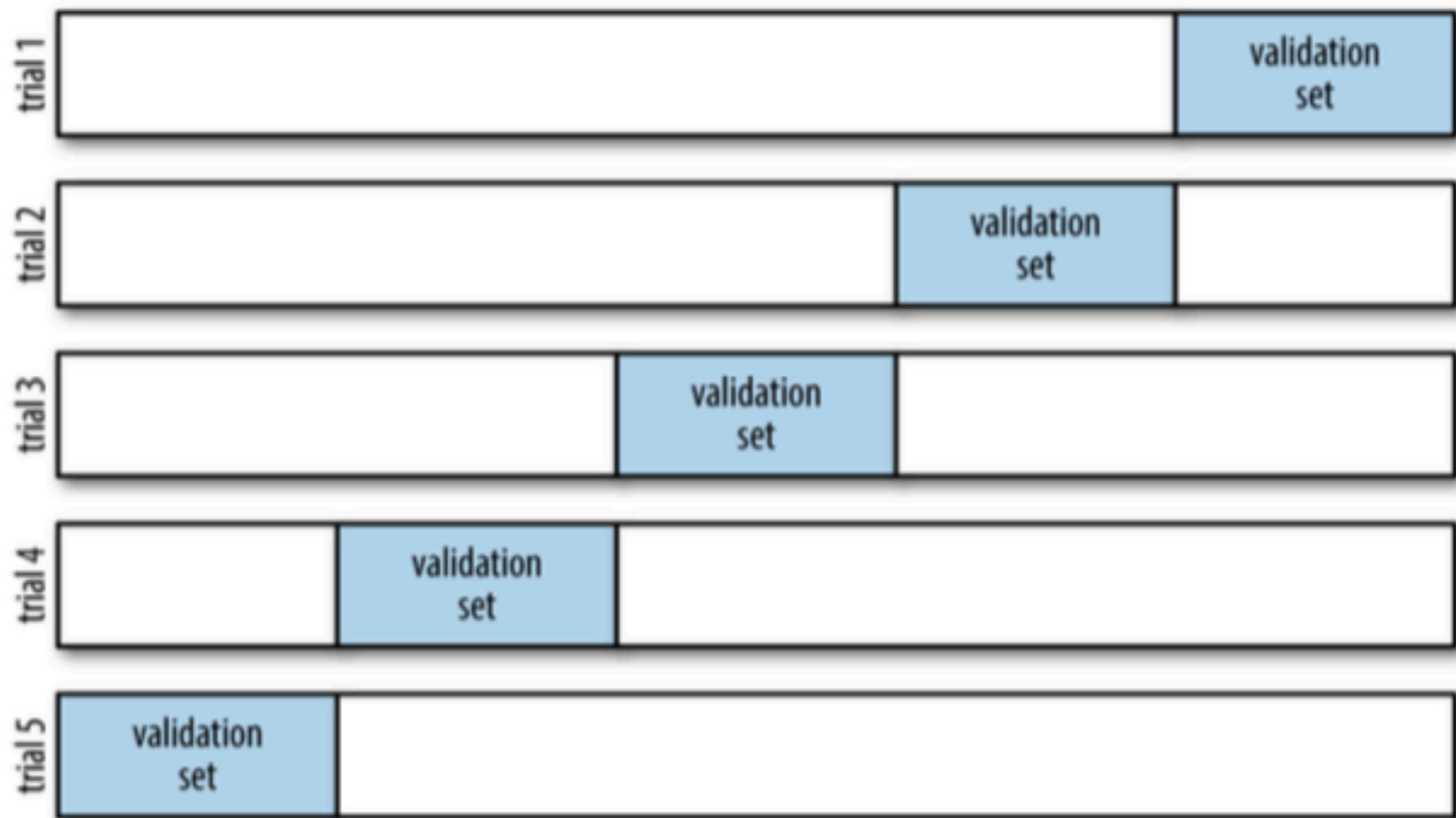
- Os dados devem ser divididos em 2 grupos: **treinamento** (normalmente 80%) e **teste** (normalmente 20%).
- O treinamento do modelo é feito no dados de treinamento. Lá ele também é testado. Para isto se usa de **Cross Validation (Validação Cruzada)**. Se pode fazer CV com 2 ou mais folds (partes). Melhor com no mínimo 5 folds (partes).
- Depois de ter sido validados com CV, o modelo é então finalmente testado com os dados de teste.
- Pre-processamento e transformações de dados que foram feitos no dados de treinamento antes da modelagem precisam serem feitos igualmente nos dados de teste antes que o teste seja feito!

CV com 2 folds (partes)



**Se treina o model na parte branca
e se valida (se testa nos dados de treinamento) o modelo na parte azul**

CV com 5 folds (parte)



**Se treina o model na parte branca
e se valida (se testa nos dados de treinamento) o modelo na parte azul**

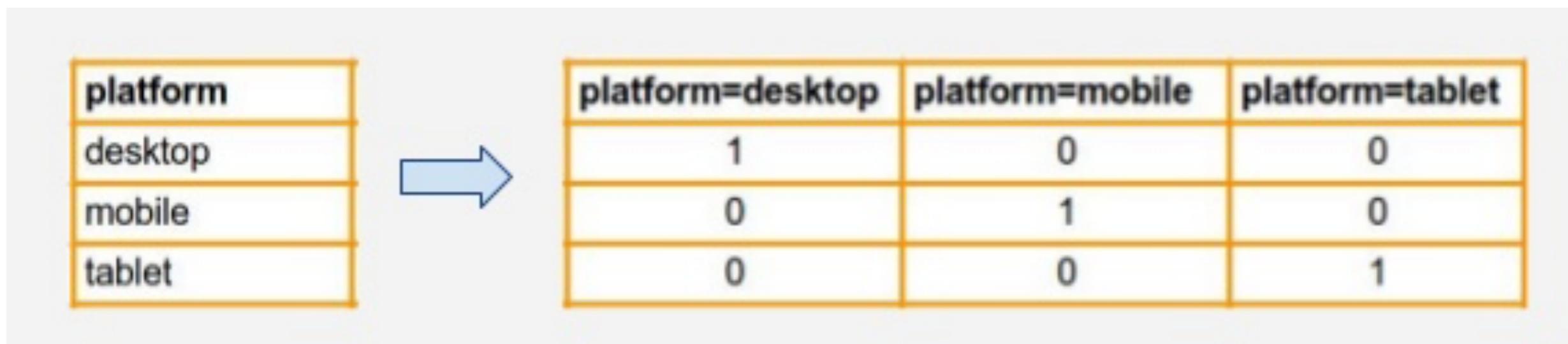
Lidando com valores nulos

- Eliminar todas as linhas com eles ou mesmo colunas se toda a coluna tem valores nulos
- Imputar um valor no seu lugar: usando a média dos valores na coluna, usando a média dentro de uma certa categoria, usando o valor maior ou menor

Lidando com valores categóricos e textos

- Muitos das modelagens requerem que os valores sejam numéricos
- Para transformar valores categóricos em numéricos basta usar o que se chama de **one-hot encoding**

One-Hot Encoding



One-Hot Encoding

User interest
Tech
Fashion
Fashion
Sports
Tech
Tech
Sports

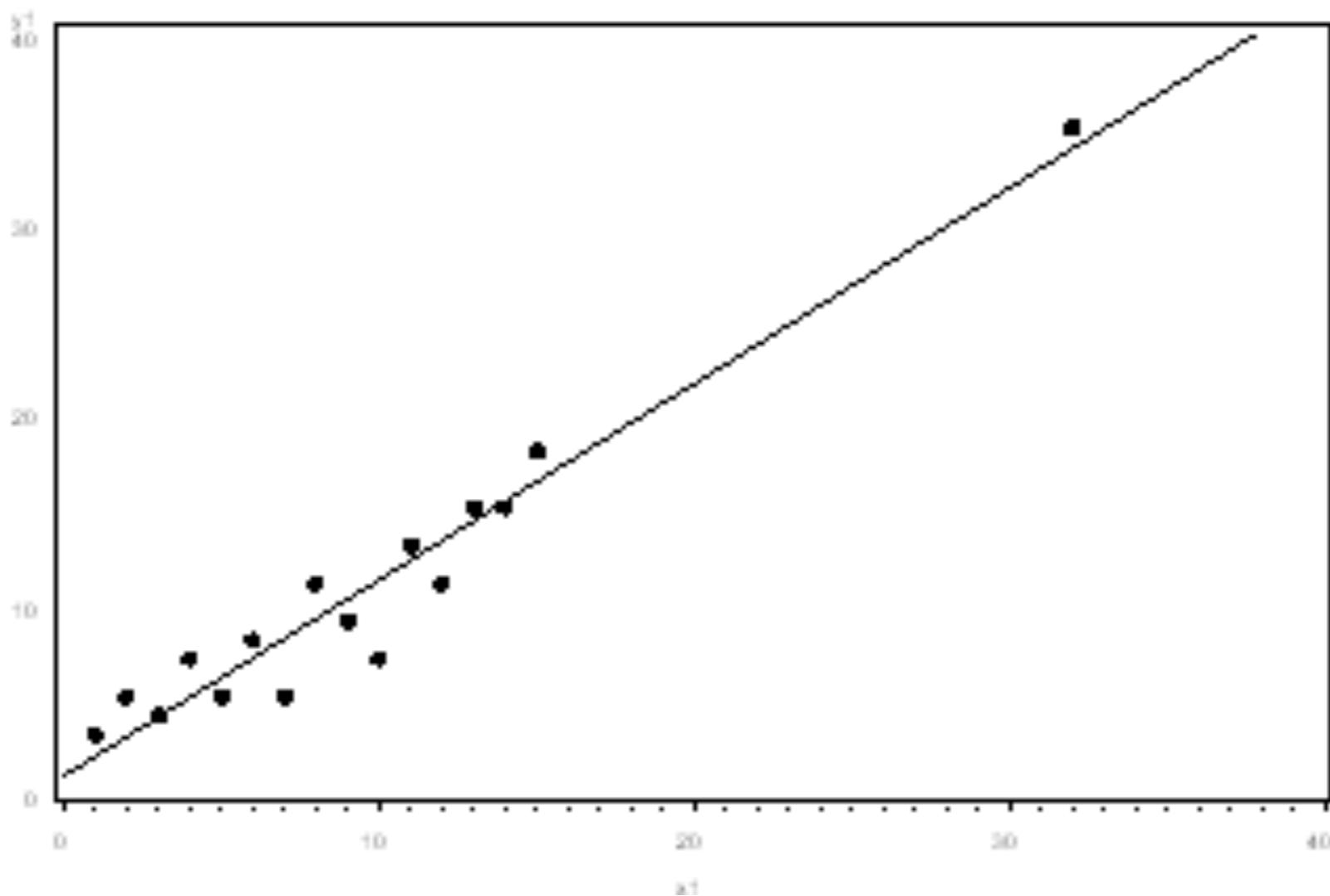


Interest: tech	Interest: fashion	Interest: sports
1	0	0
0	1	0
0	1	0
0	0	1
1	0	0
1	0	0
0	0	1

Lidando com outliers (fora de série)

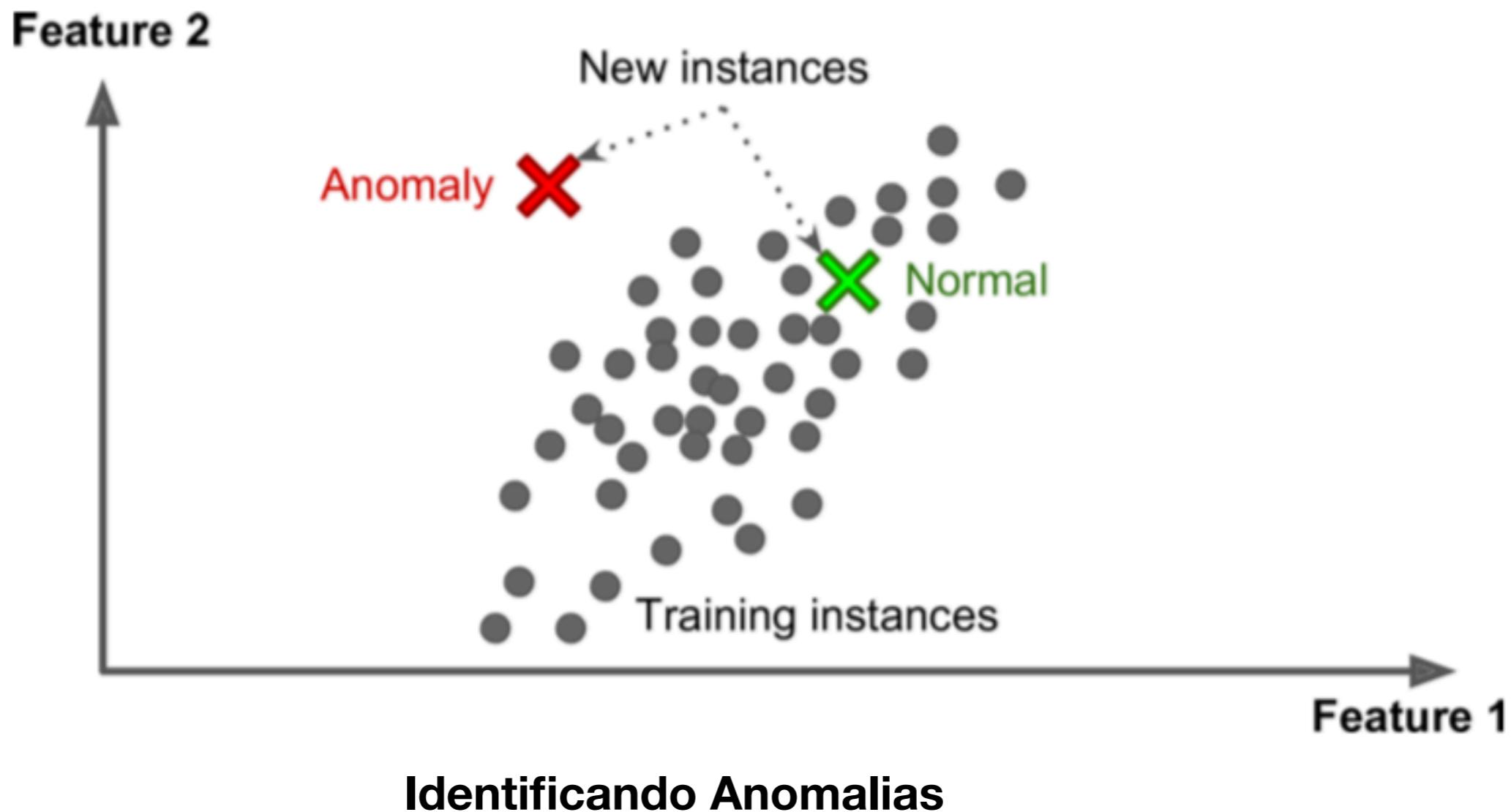
- Outliers devem ser em geral removidos,
- Ou pelo menos seu impacto na modelagem verificado no caso de eles serem removidos.
- Em geral eles são fontes de erros na modelagem.
- Mas, por outro lado, uma vez que a modelagem está feita, um outlier pode ser indicação de uma anomalia!

Lidando com outliers (fora de série)



Um típico caso de outlier

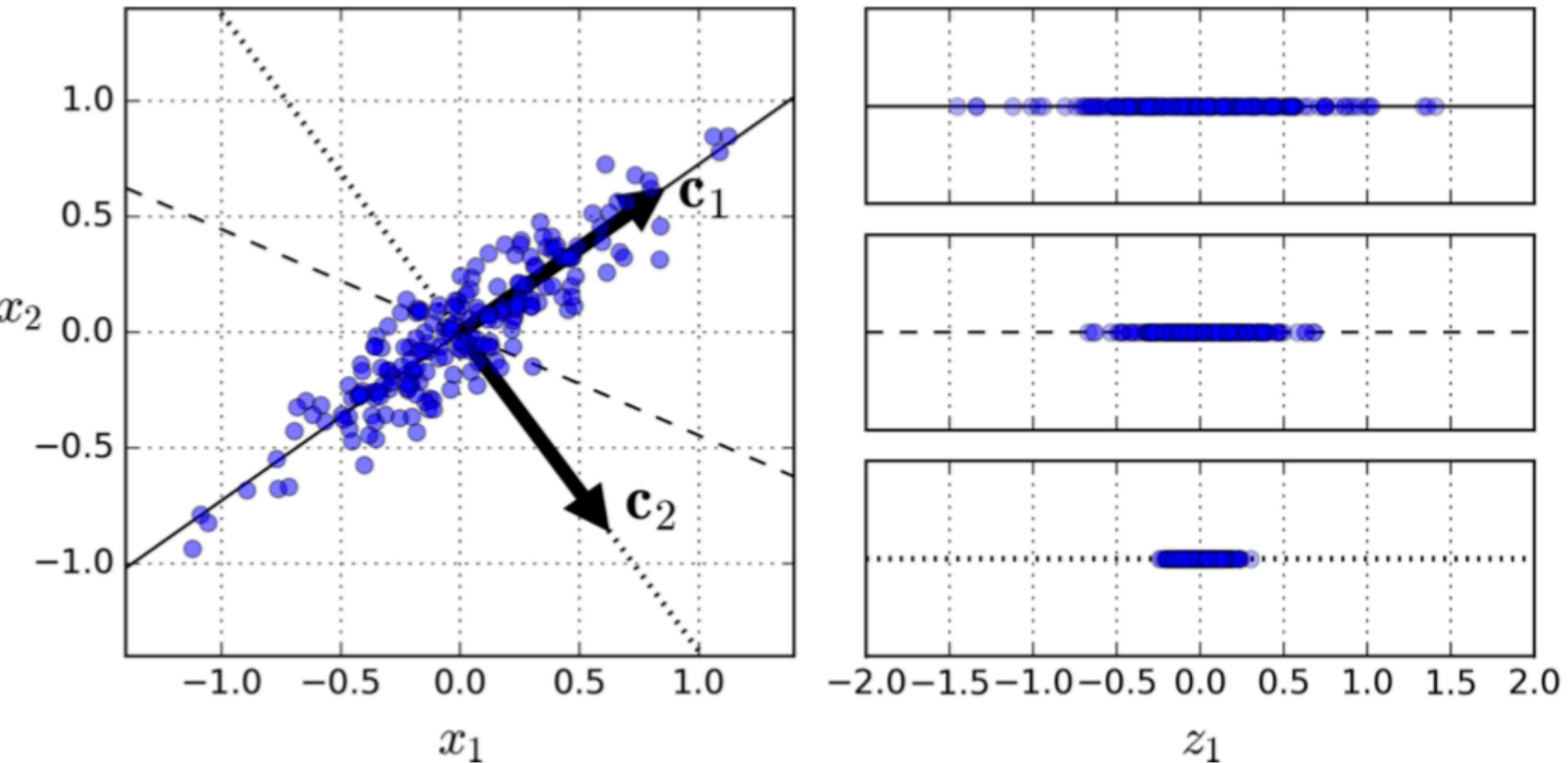
Lidando com outliers (fora de série)



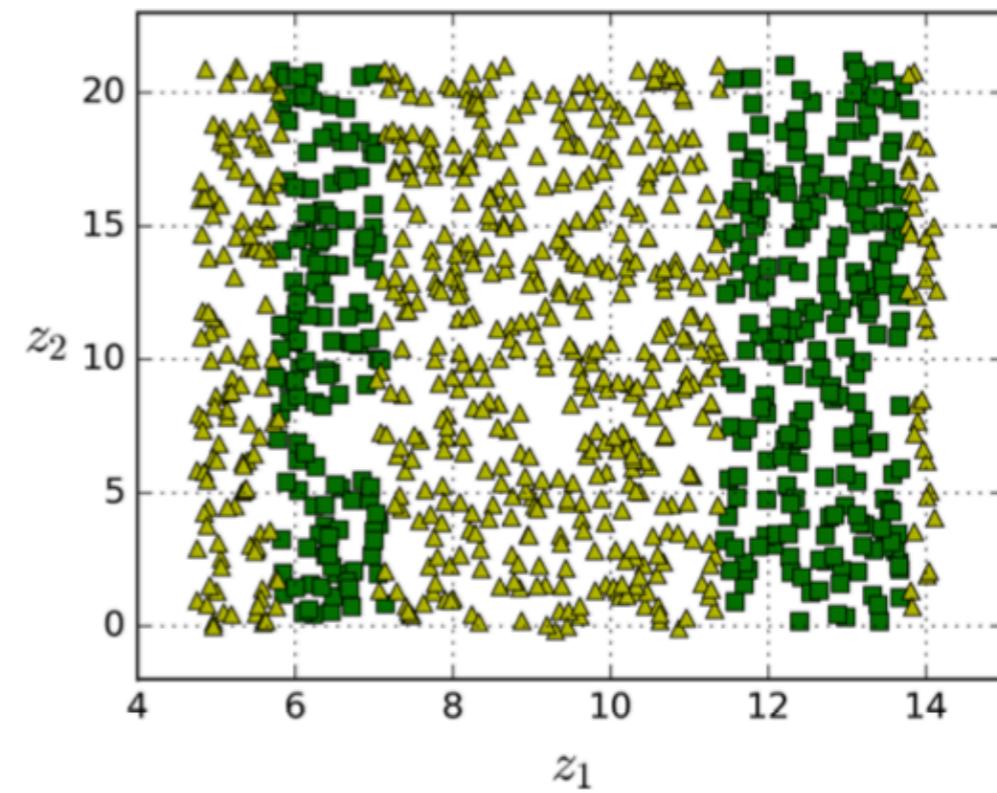
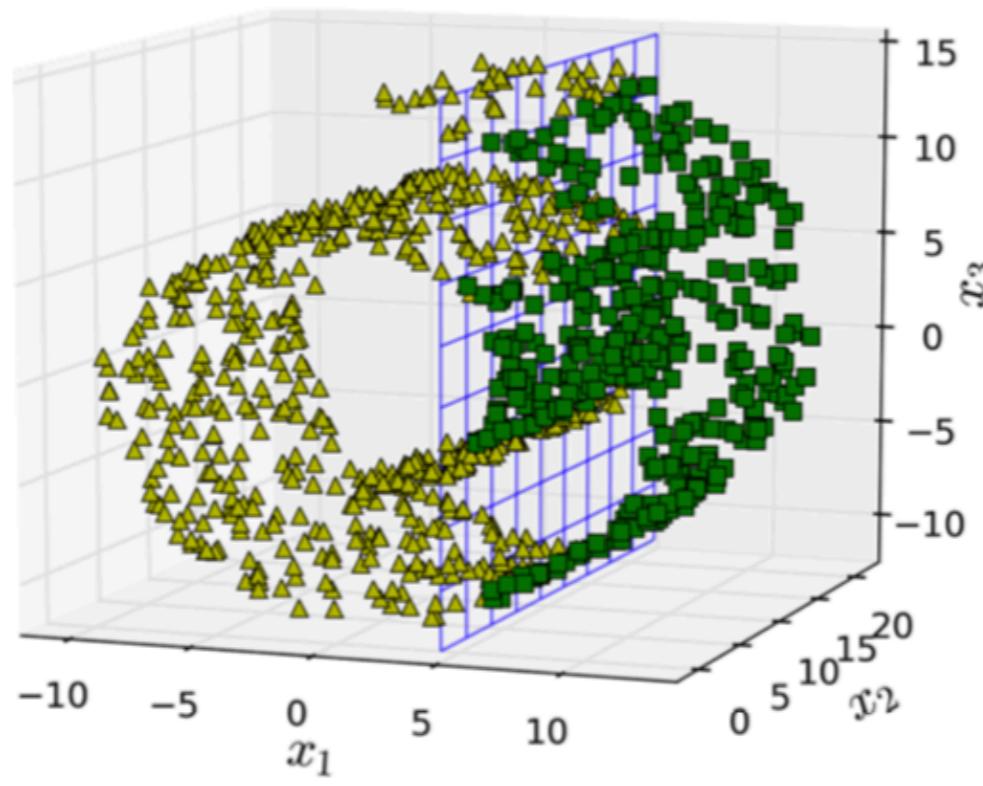
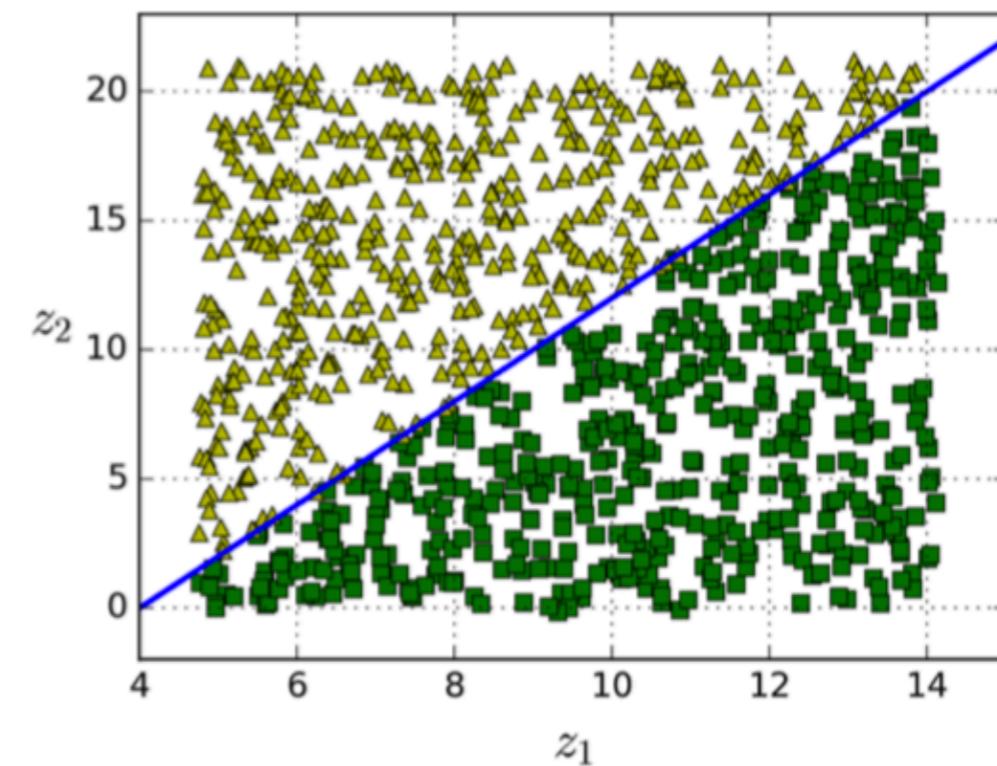
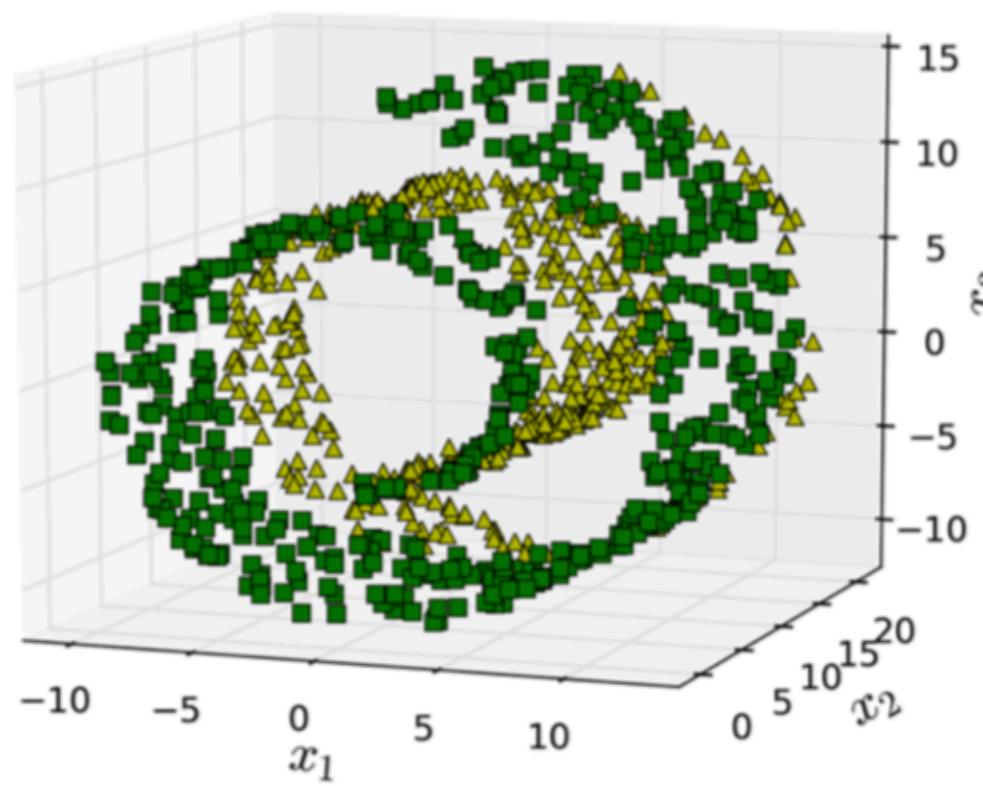
Lidando com muitas dimensões, redução de dimensões

- Muitas dimensões (features, atributos, variáveis, colunas) causam problemas para alguns tipos de modelagens.
- Muitas dimensões fazem com que o tempo de execução de cross validation (validação cruzada) seja bastante longo.
- Muitas dimensões dificultam a visualização dos dados.
- **Principal Component Analysis (PCA)**
- Reduz o numero de dimensões para aquelas que mais são responsáveis pela variância nos dados em ordem de peso, a 1a. mais responsável, a 2a. mais responsável, etc.

PCA



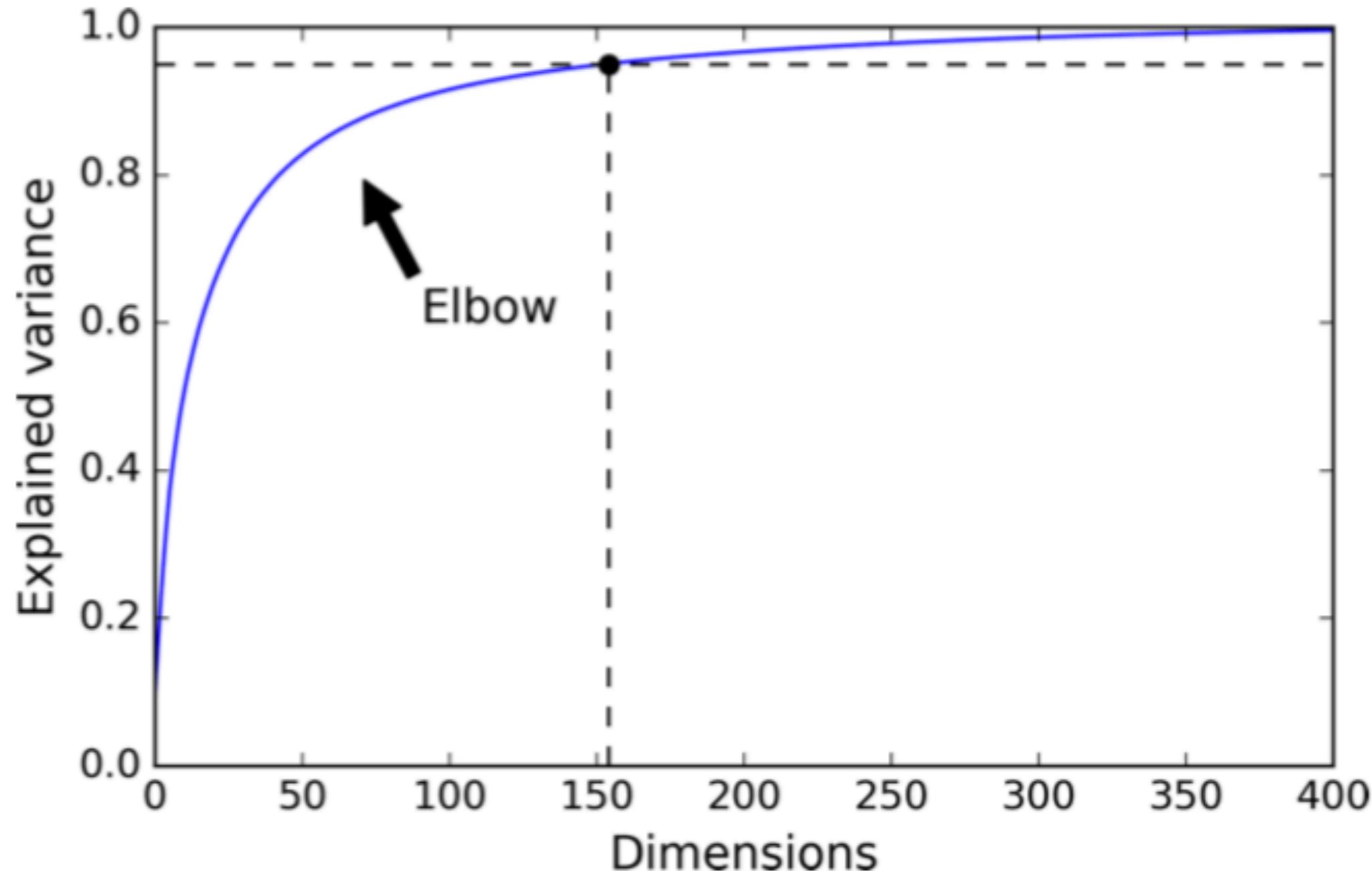
PCA



PCA com Python

```
from sklearn.decomposition import PCA  
  
pca = PCA(n_components = 2)  
X2D = pca.fit_transform(X)
```

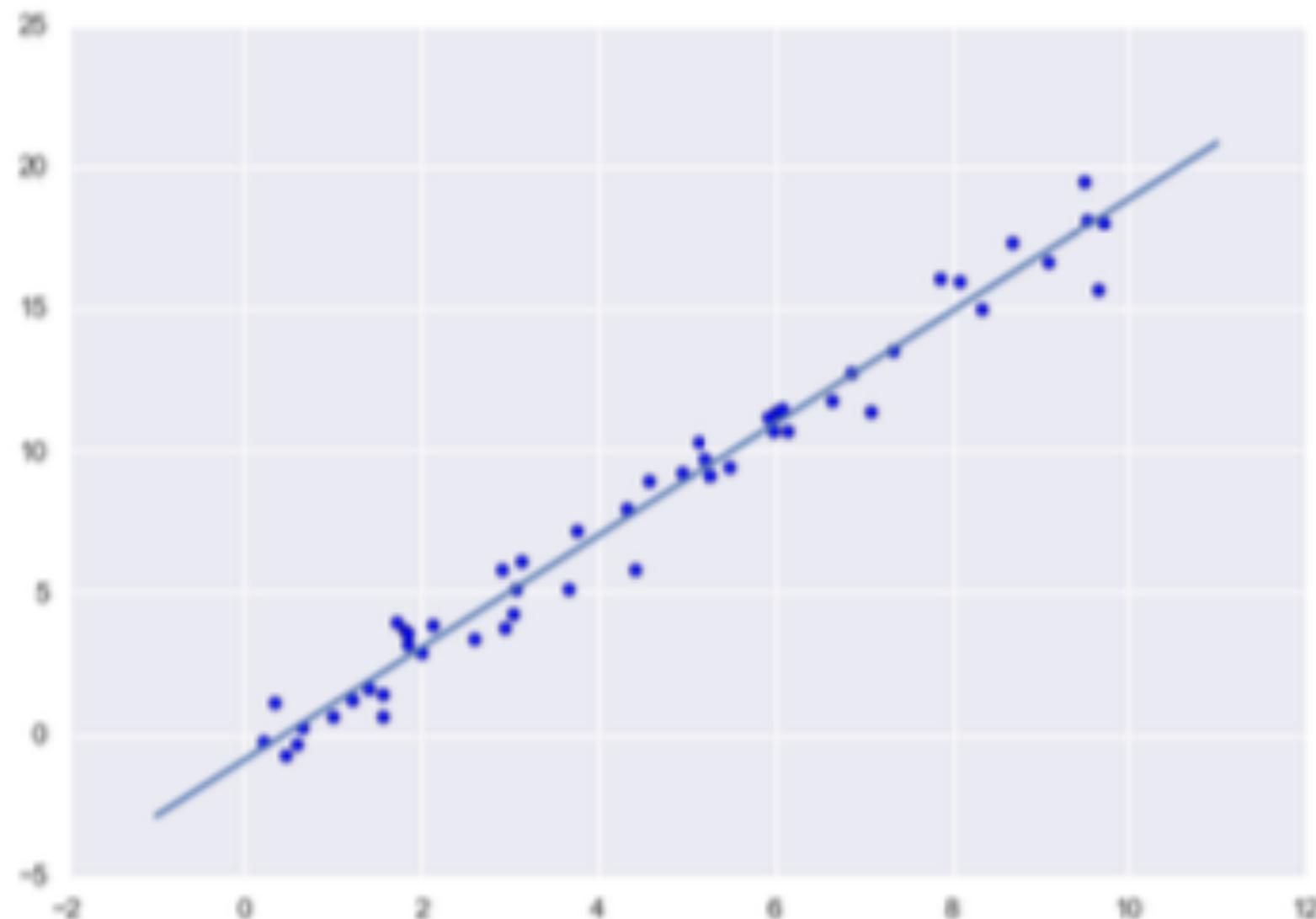
PCA Nr. Dimensões



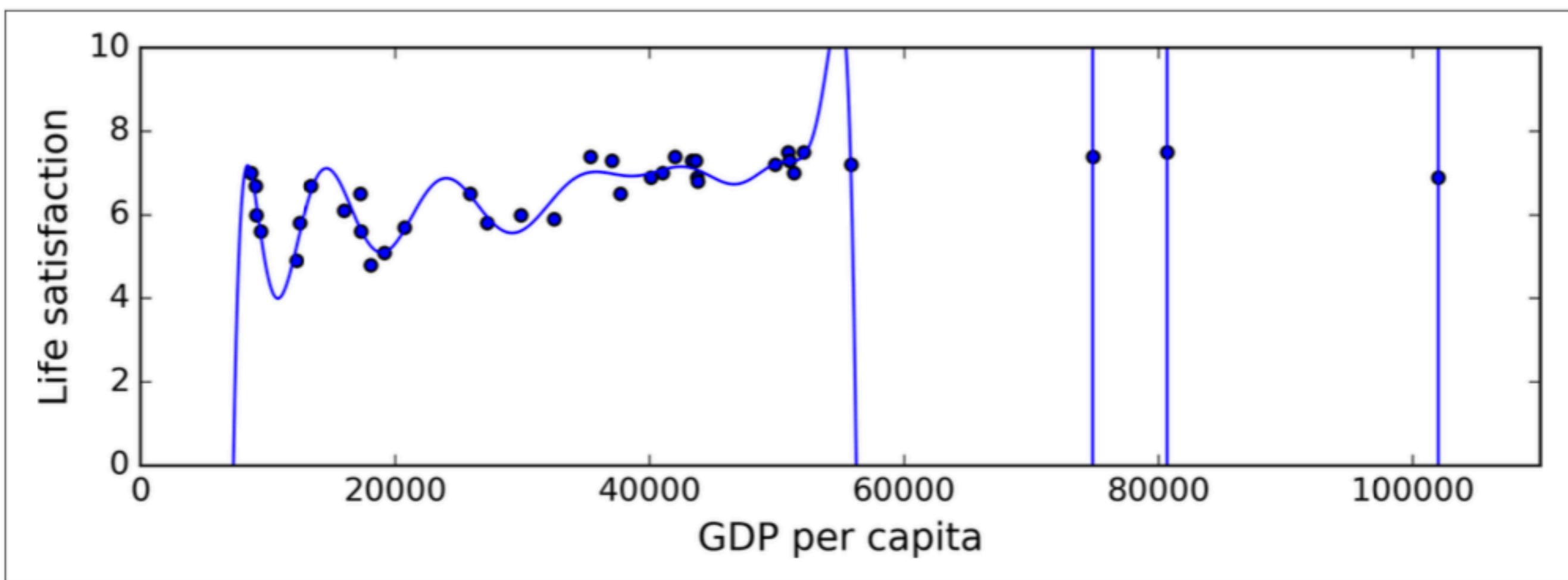
PCA com Python

```
>>> pca.explained_variance_ratio_
array([ 0.84248607,  0.14631839])
```

Fitting uma linha

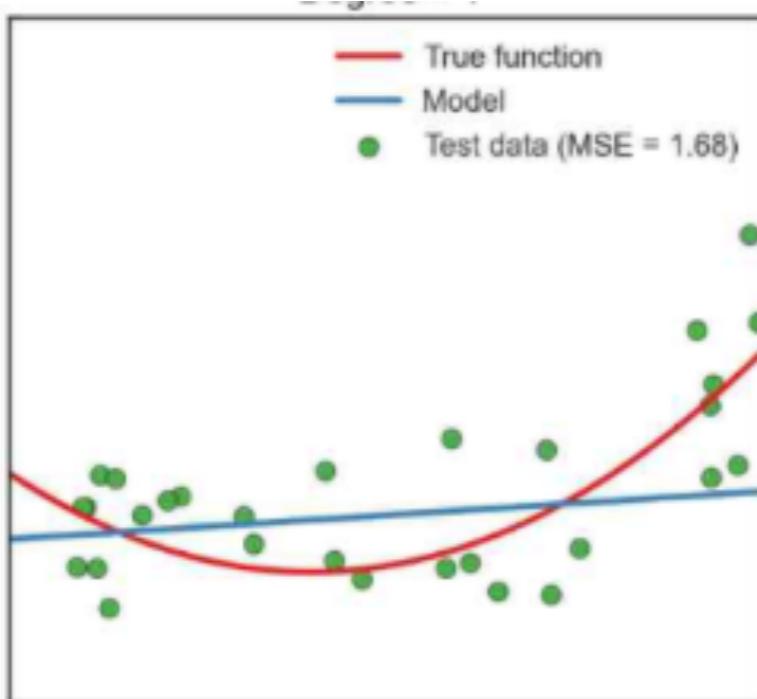


Overfitting em Regressão

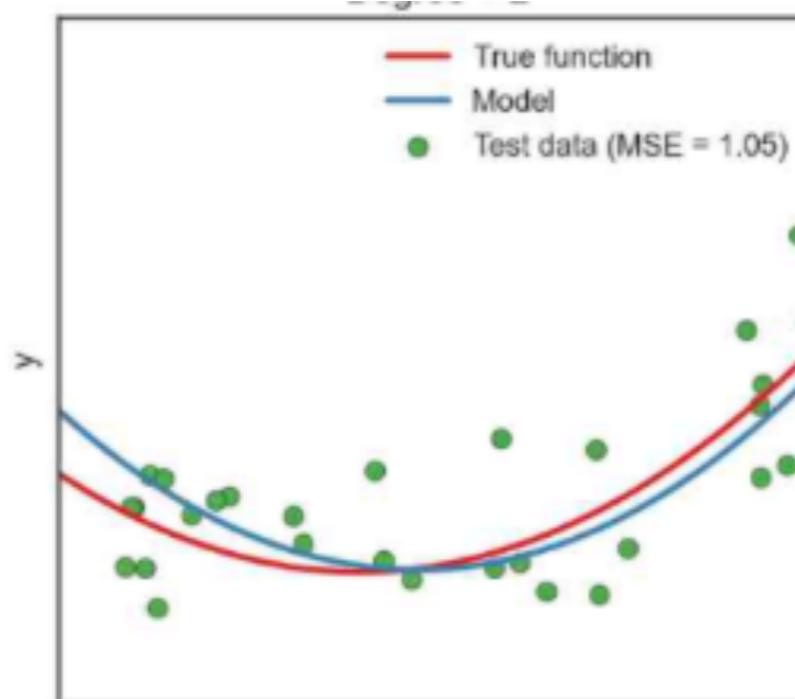


Overfitting em Regressão

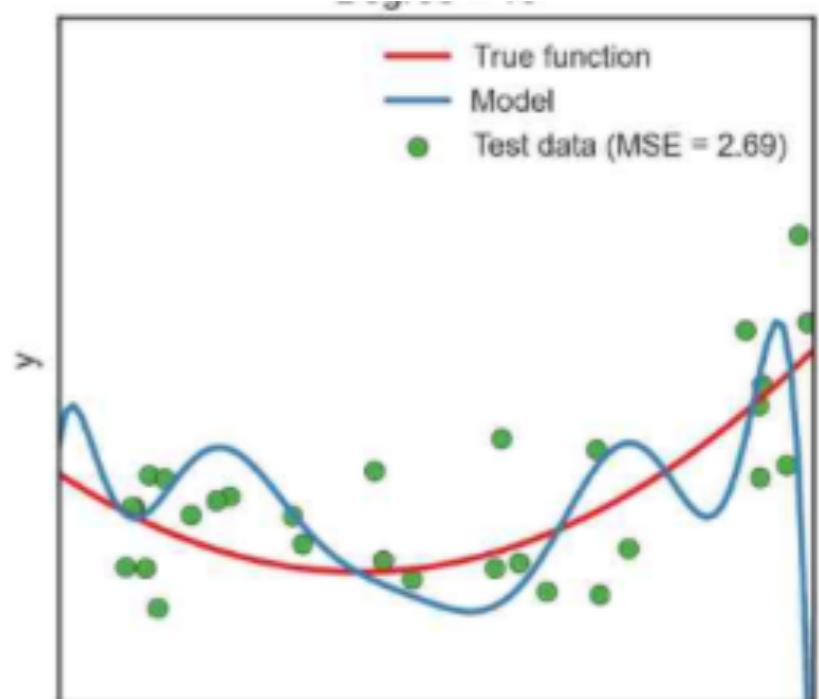
Linha vermelha = função real dos dados
Linha azul = Modelo



Underfitting

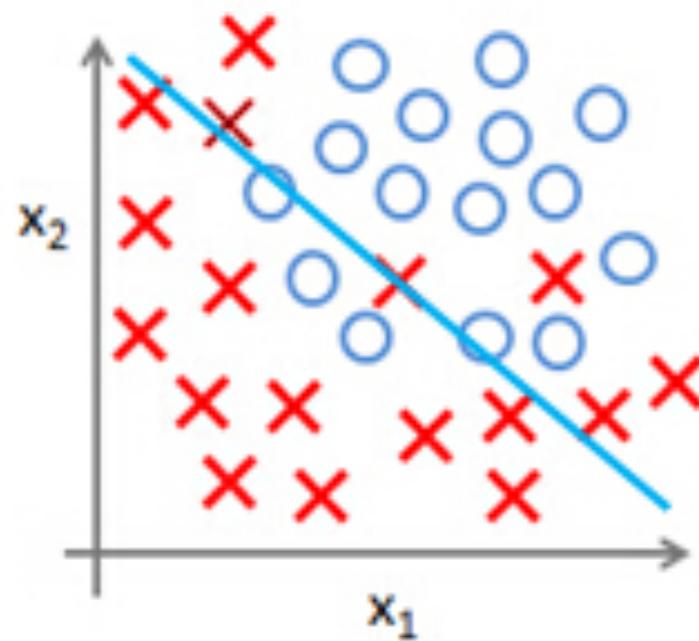


Correto

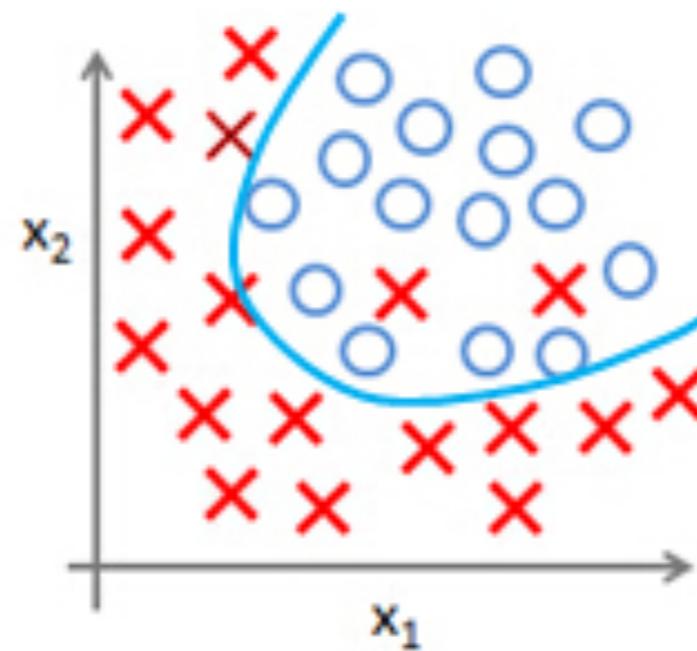


Overfitting

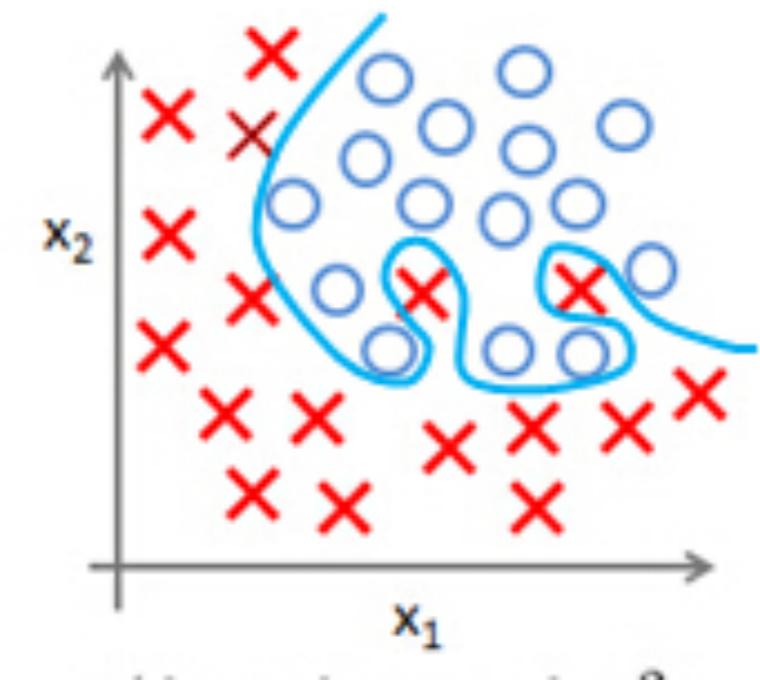
Overfitting em Classificação



Solução Linear



Solução curvilínea



Overfitting

Fontes de Erros na Nossa Modelagem

- Fontes de erros nas nossas modelagens são devidos a:
- **Bias (viés)**
- **Variance (variância)**
- **Erros irreduutíveis** (da coleta de dados)