

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

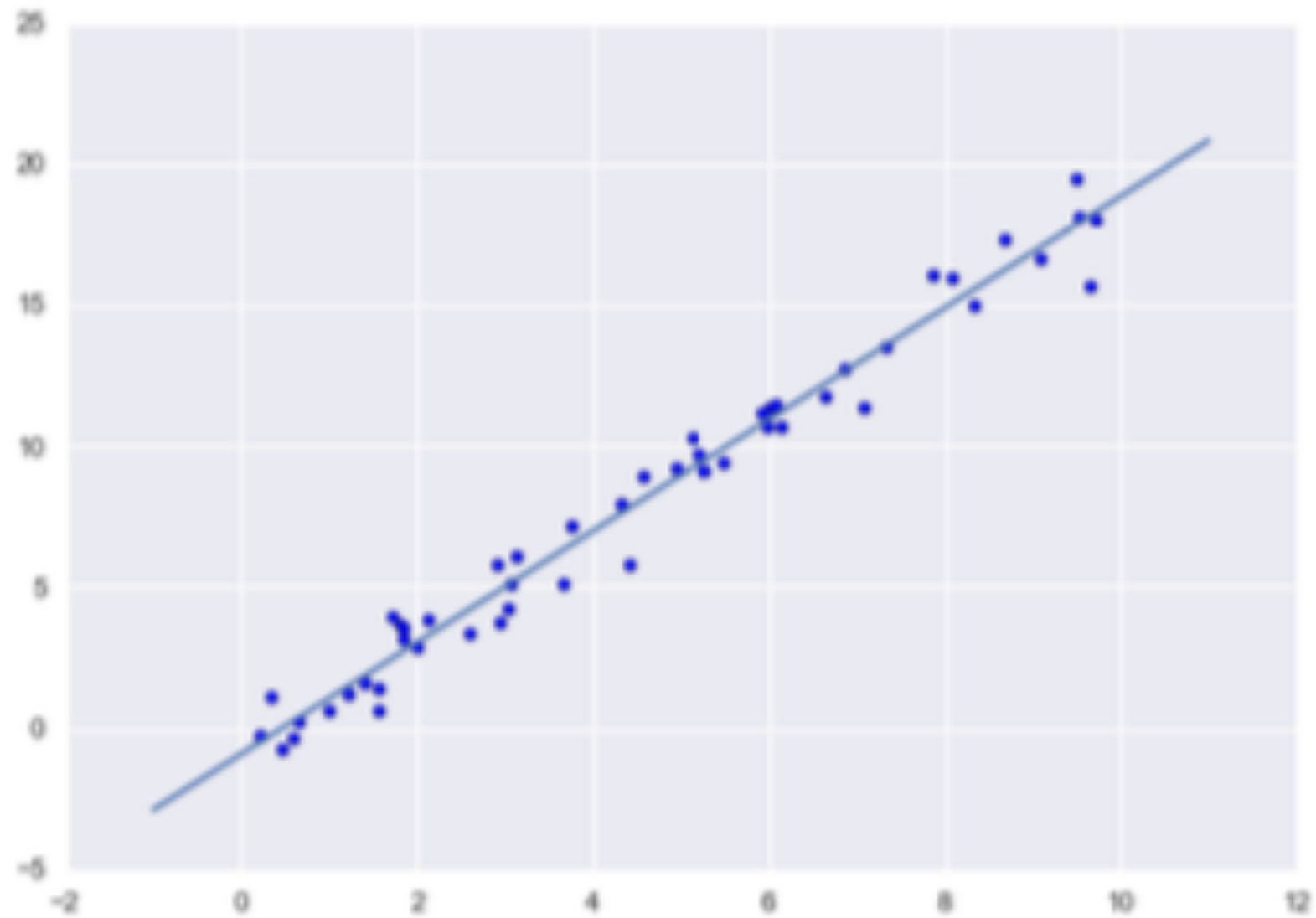
**Turma 2**

**Terceiro dia**

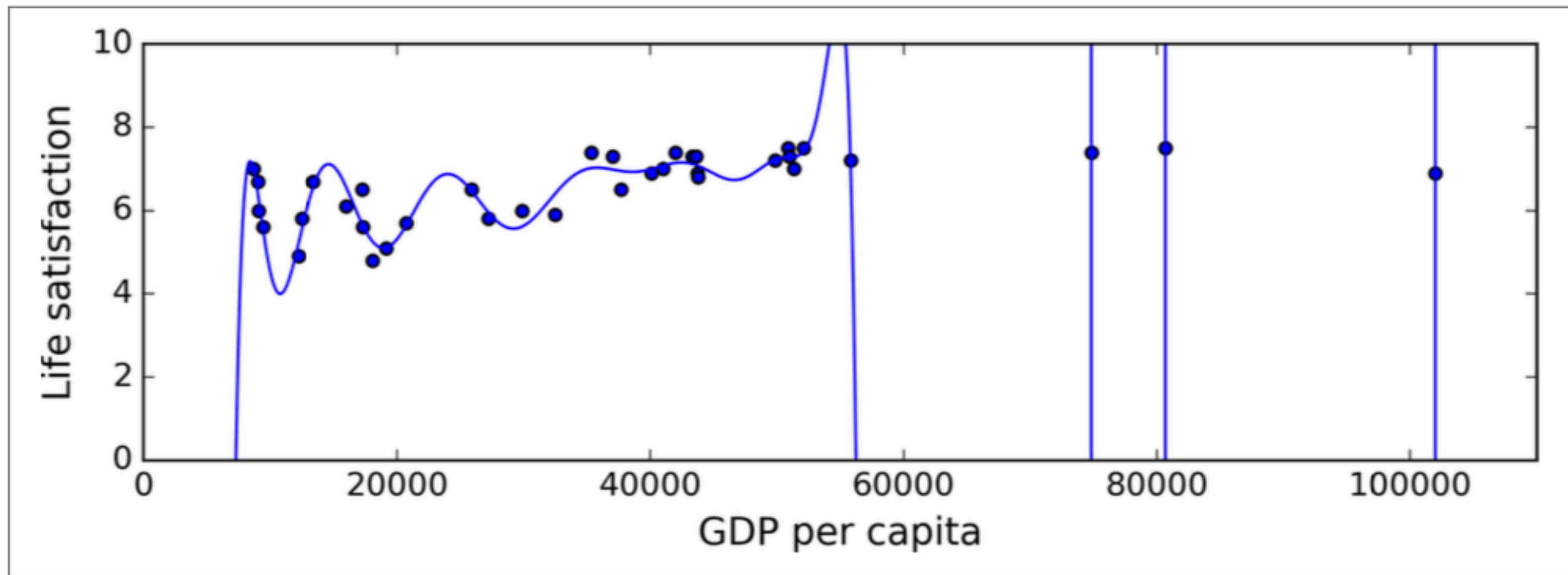
**Paulo Cysne Rios Jr.**

**Overfitting e  
Underfitting**

# Fitting uma linha



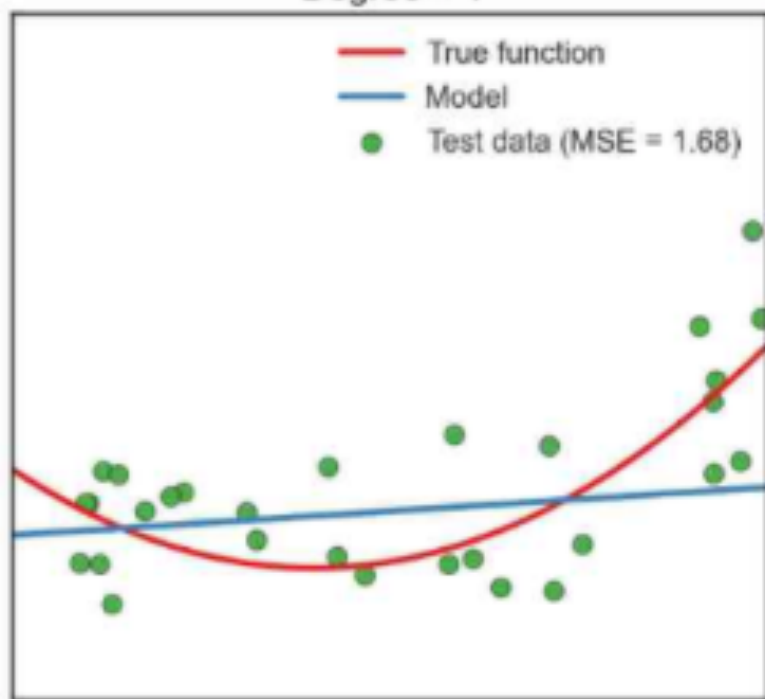
# Overfitting em Regressão



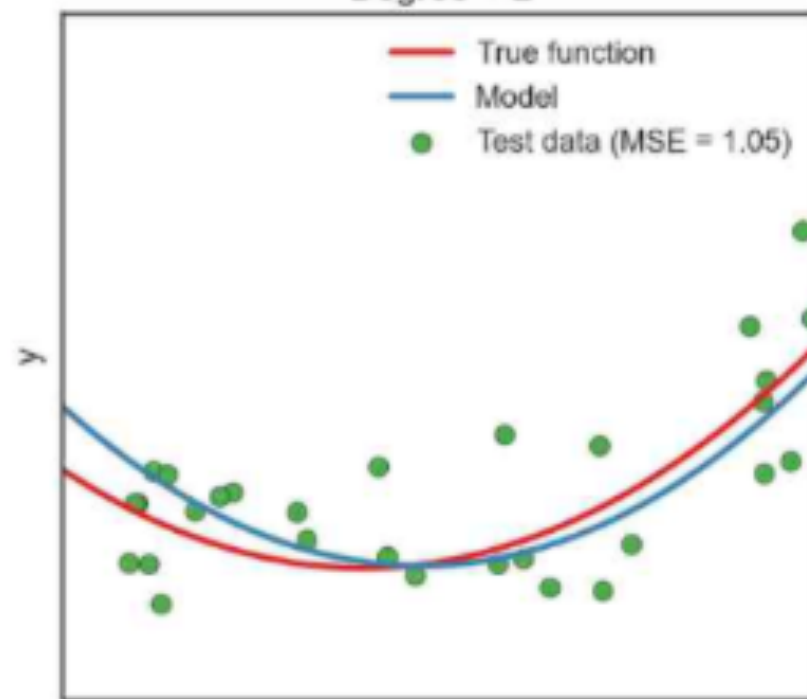
# Overfitting em Regressão

Linha vermelha = função real dos dados

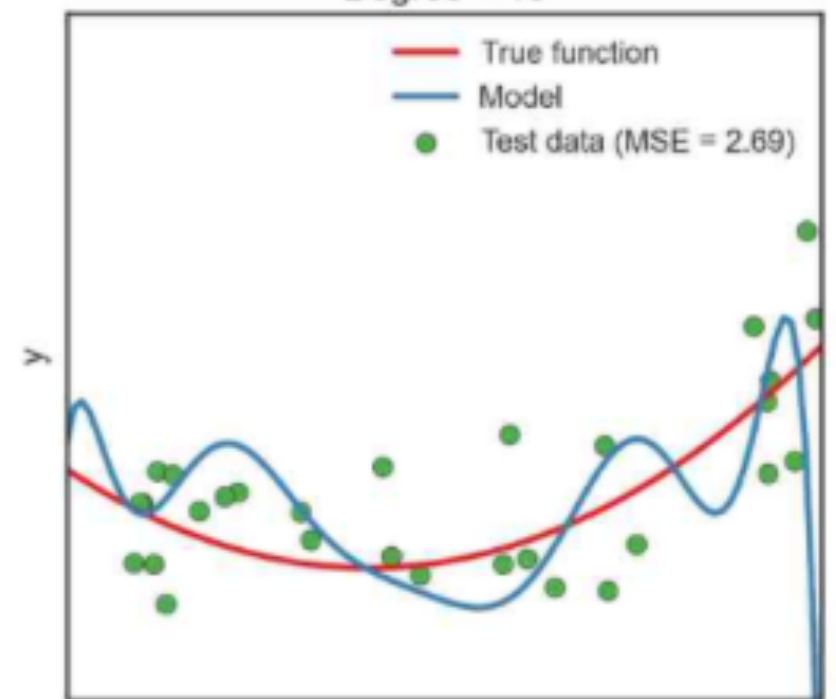
Linha azul = Modelo



**Underfitting**

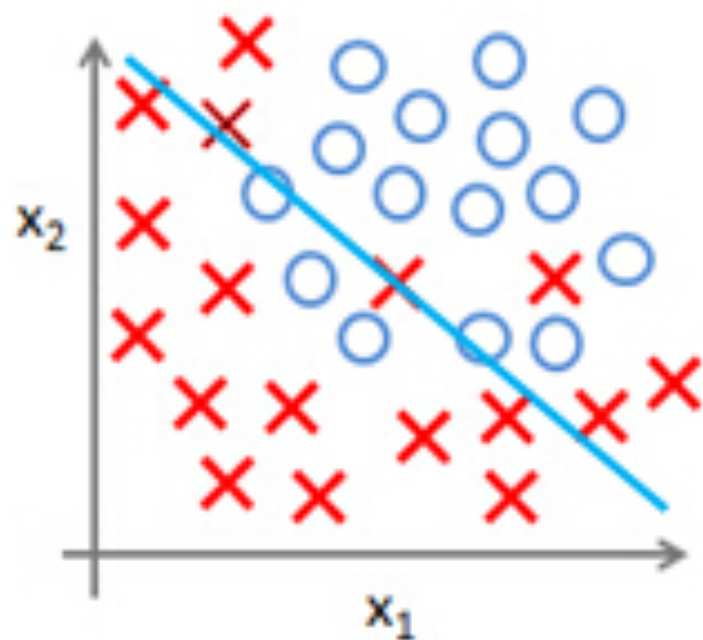


**Correto**

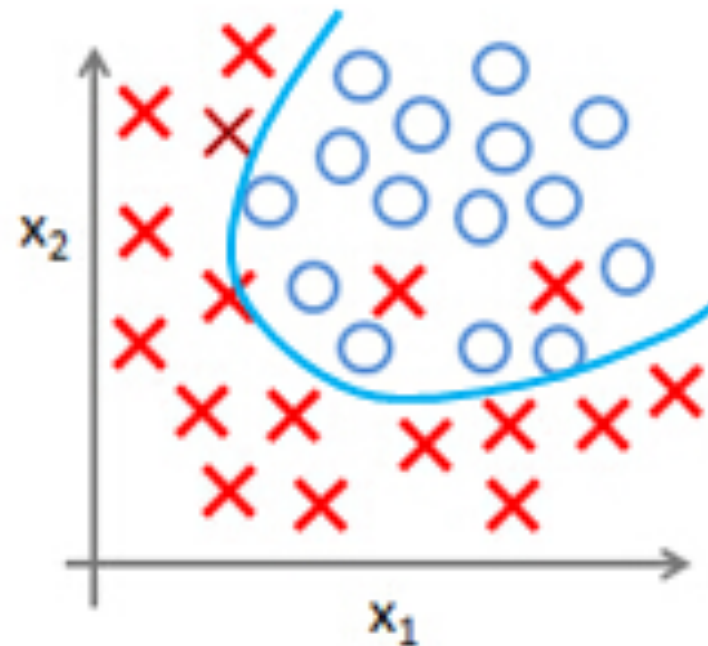


**Overfitting**

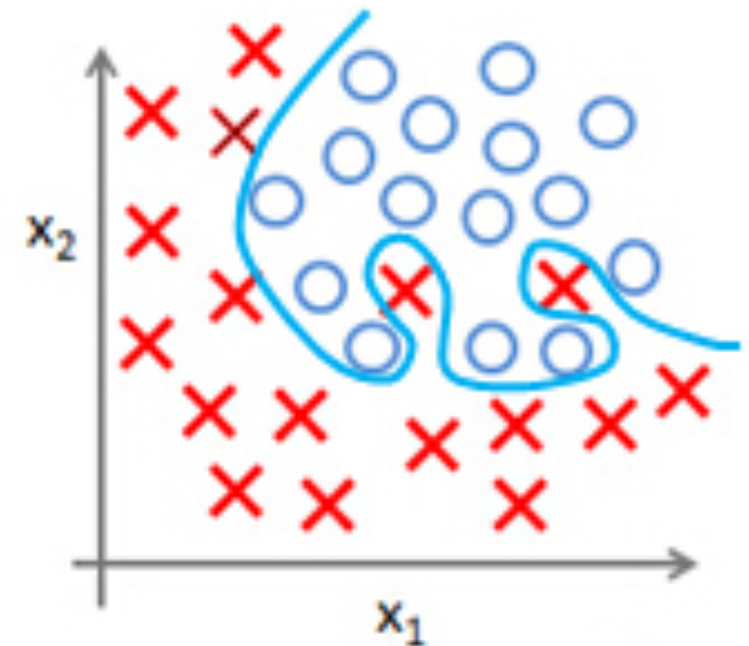
# Overfitting em Classificação



**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 irreduzíveis** (da coleta de dados)

# Bias (Viés)

- Nossa tendência e necessidade em simplificar nossos modelos para fazê-los mais compreensíveis.
- Esta parte do erro de generalização deve-se a suposições erradas, como assumir que os dados são lineares quando é realmente quadrático.
- Um modelo de alto **viés** é mais provável de prejudicar os dados de treinamento.



# Variance

- **Variance (Variância):** a flexibilidade do nosso modelo de mudar com novos dados.
- Esta parte é devido à **sensibilidade excessiva do modelo** a pequenas variações nos dados de treinamento.
- Um modelo com muitos graus de liberdade (como um modelo polinômico de alto grau) provavelmente terá alta variação e, portanto, superará os dados de treinamento.

# Erros irreduzíveis

- Esta parte é devido ao noise (barulho) dos dados em si.
- A única maneira de reduzir esta parte do erro é limpar os dados.
- Por exemplo, corrigir as fontes de dados, como sensores quebrados, ou detectar e remover outliers.

# Bias Variance Tradeoff

- Se aumenta um, o outro diminui!
- Aumentar a complexidade de um modelo geralmente aumentará sua variação e reduzirá seu viés.
- Por outro lado, reduzir a complexidade de um modelo aumenta seu viés e reduz sua variação.
- É por isso que se chama tradeoff / troca.

# Bias Variance Tradeoff



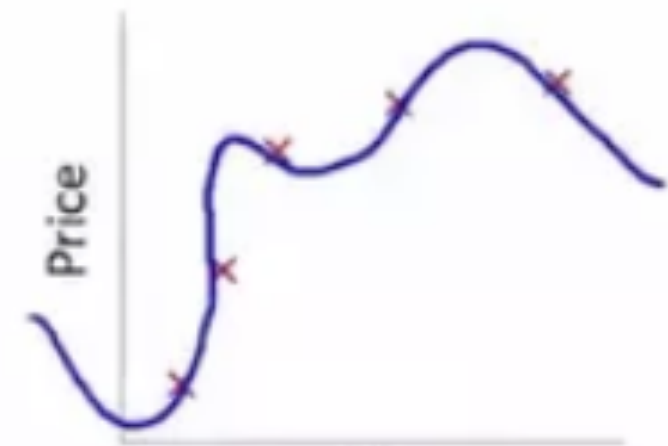
Size  
 $\theta_0 + \theta_1 x$

High bias  
(underfit)



Size  
 $\theta_0 + \theta_1 x + \theta_2 x^2$

"Just right"

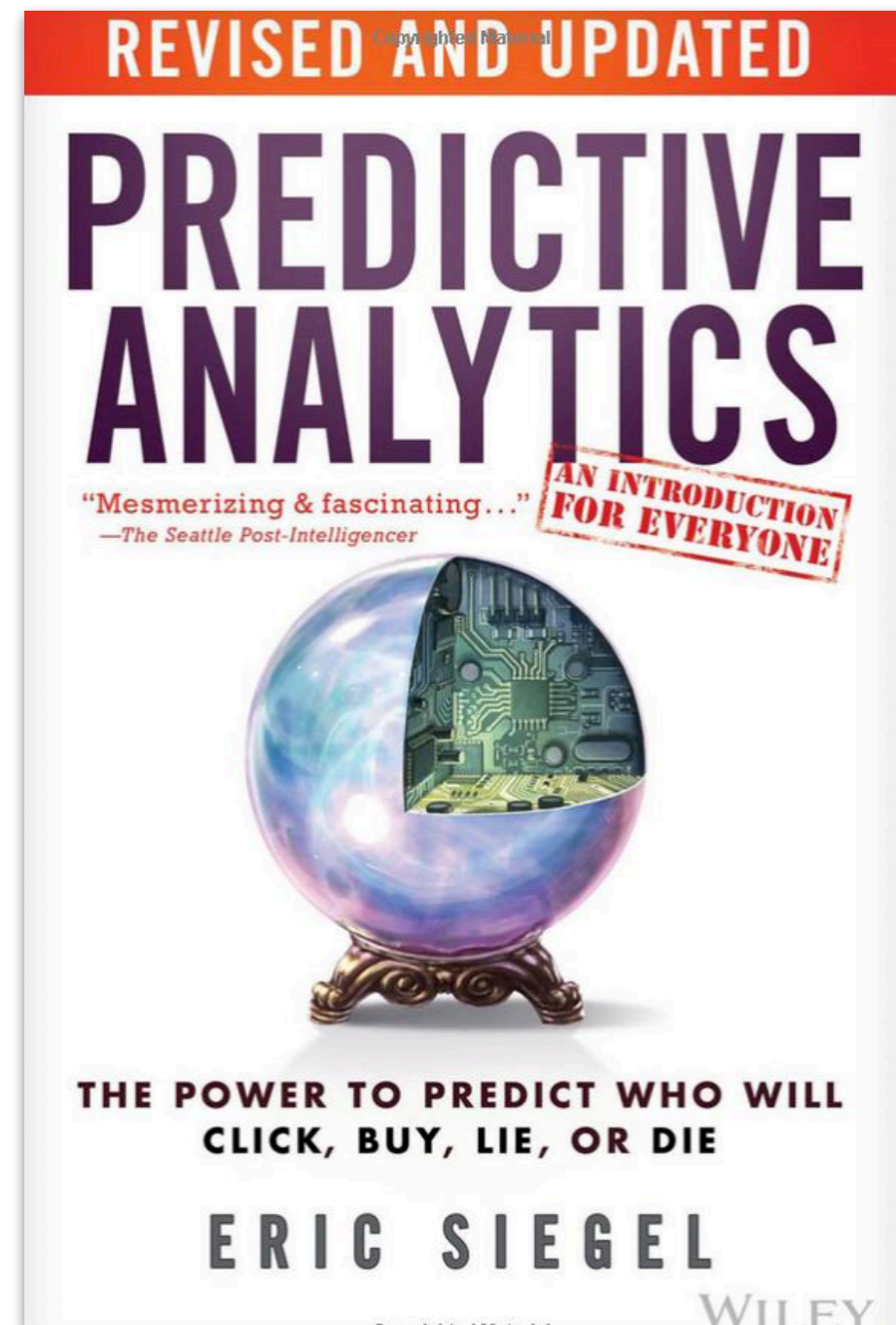


Size  
 $\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

High variance  
(overfit)

# Ciência de Dados e Analítica Preditiva

- Qual a diferença entre **Ciência de Dados** e **Analítica Preditiva**?
- Resposta na próxima aula!  
Tente descobrir....



# Aplicações

- Cite 5 aplicações de Ciência de Dados na sua empresa
- Cite 5 aplicações em geral de Ciência de Dados
- Escreva num papel que coletaremos antes do final da aula de hoje
- para que possamos mostrar exemplos relacionados às aplicações que mais lhe interessam!

# Prática

# Exemplo: 2 Aplicativos Meus de Predição

- Um deles que fiz com **predição de valores de imóveis** usando um conjunto de dados real. Outro de doenças coronárias.
- Foram feitos em **R** (no excelente ambiente **RStudio**) e disponibilizado na web usando a ferramenta **Shiny** (do R).
- A **rápida** reação do programa ao input do usuário mostra como se pode disponibilizar um modelo analítico para uso pela Internet.
- Faremos ao longo do workshop uma modelagem semelhante com **Python** e em especial também com **Deep Learning**.



# Servidor das Minhas aplicações

shinyapps.io

HelpAccount: ?

Dashboard

Applications»

Account»

APPLICATION 123196 - BOSTON

Overview

Metrics

URLs

Settings

Users

Logs

Restart

Archive

Delete

OVERVIEW

Id

123196

Name

Boston

URL

<https://cyzne.shinyapps.io/Boston/>

Status

Sleeping

Size

large

Deployed

Sep 18, 2016

Updated

Oct 10, 2017

Created

Aug 31, 2016

Bundle

Download

INSTANCES

Max

Id: 555732

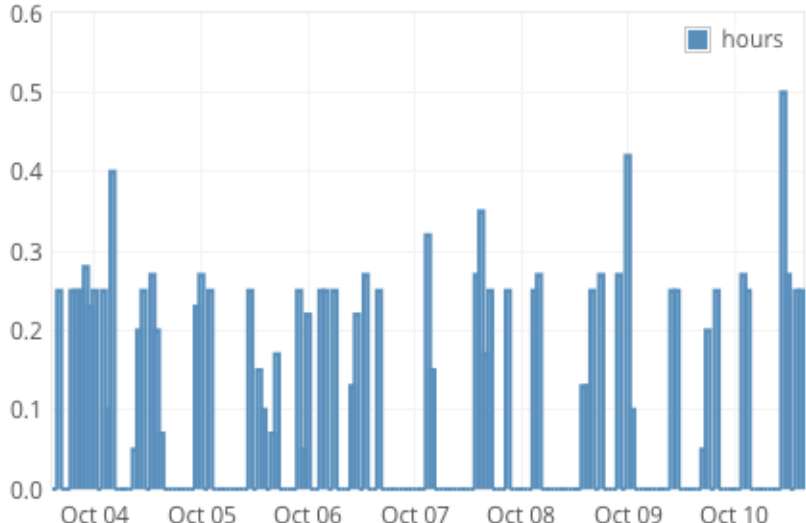
▶

🗑

APPLICATION USAGE

Total: 13.82 hours

hours



© 2017 RStudio Inc. | All Rights Reserved | Terms Of Use

# Minhas 2 aplicações

- **Predição de preços de imóveis**

<https://cyzne.shinyapps.io/Boston/>

- **Predição de doença coronária**

<https://cyzne.shinyapps.io/Coronary/>

# **Minha Aplicação de Predição de Preços de Imóveis**

**Usando um conjuntos de dados reais disponíveis na Internet**

# DataCyz - Predicting Prices and Their Most Important Features

By Paulo C. Rios, Jr., September, 2016

Use the menus to do the following:

## About

### Understanding this application and the data

**Introduction** - A description of this web-based analytical application

**Conclusions** - Our findings and conclusions

**Features** - Description of each feature of the data set

**Dataset** - The data with all features in a searchable and sortable form

## Data Visualization

### Exploring and visualizing the data

**Frequency** - Chart with the frequency of each selected feature

**Relationships** - Plot of each feature against price to investigate their relationship

## Analytics

### What we have learned from the data

**Importance** - Which features have most impact on the house price

**Accuracy** - A chart and a table of all actual vs. predicted house prices values, including the mean error

**Predictions** - With the given feature values, we predict the target value, that is, the house price, with high accuracy

## Search

### Finding suitable prices

Using our analytical model, it is possible to look for houses using the 2 most important features (the ones we discovered in the Importance tab)

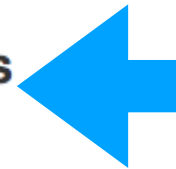
About ▾

Data Visualization ▾

Analytics ▾

Search

## House Price Analytics



### Summary

This web application shows how advanced data analytics are not only powerful, but can also be easily made available and used over the Internet. This site is fully dynamic. All the analytics and charts are generated on the fly, using the source data described below.

### The data

The Boston data set contains information collected by the U.S Census Service concerning housing in the area of Boston, Mass. It has 506 entries (that is, houses) and 14 features for each entry. Our main interest is our target feature, the median house value, as we describe below.

### Objectives

In this web application we have three main goals:

1. To identify which of the features most affect the house value,
2. To predict the house value using the given other features in the data set,
3. To allow the discovery of the best opportunities in this house market

As we show in the other tabs, these goals have been fully met. In the summary tab, we also describe what we have achieved. Check the different tabs to access different results and capabilities of this advanced data analytics. A description of each tab is conveniently located in the panel on the left.

Choose one of two statistical measures below to identify **the most important features** to determine the house price. They are listed below in both chart and table format.

Mean decrease in accuracy ▾

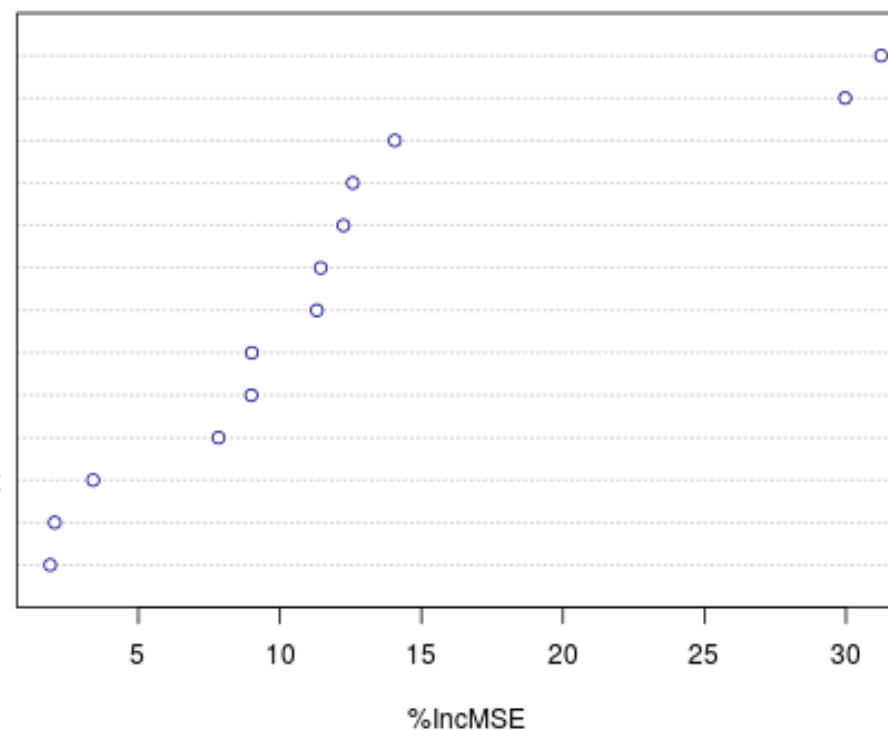
Os atributos mais importantes são revelados

## Importance Chart

Those features listed first and with a high statistical measure value are most valuable:

Most Important Features for Price - By Accuracy

avg\_nr\_rooms  
lower\_status\_pop\_pct  
distance\_employment\_ctrs  
crime\_rate  
pollution\_nitrogen\_oxide  
built\_before\_1940\_prop  
pupil\_teacher\_ratio  
property\_tax\_per\_10k  
non\_retail\_business\_prop  
relative\_prop\_blacks  
access\_radial\_highways\_index  
bounds\_charles\_river\_yesno  
residencial\_land\_prop



## Importance Table

They are also listed below in a table format. The higher their statistical measure, the more importance they have:

X.IncMSE	house.variable
31.25	avg_nr_rooms
29.97	lower_status_pop_pct
14.06	distance_employment_ctrs

2 atributos são muito mais importantes que os outros

Using the original Boston data set, the house values were predicted using our analytical model.

The error is measured by the mean difference between the actual and the predicted values.

**Mean prediction error = 3.4**

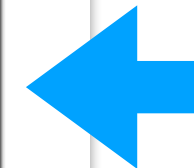
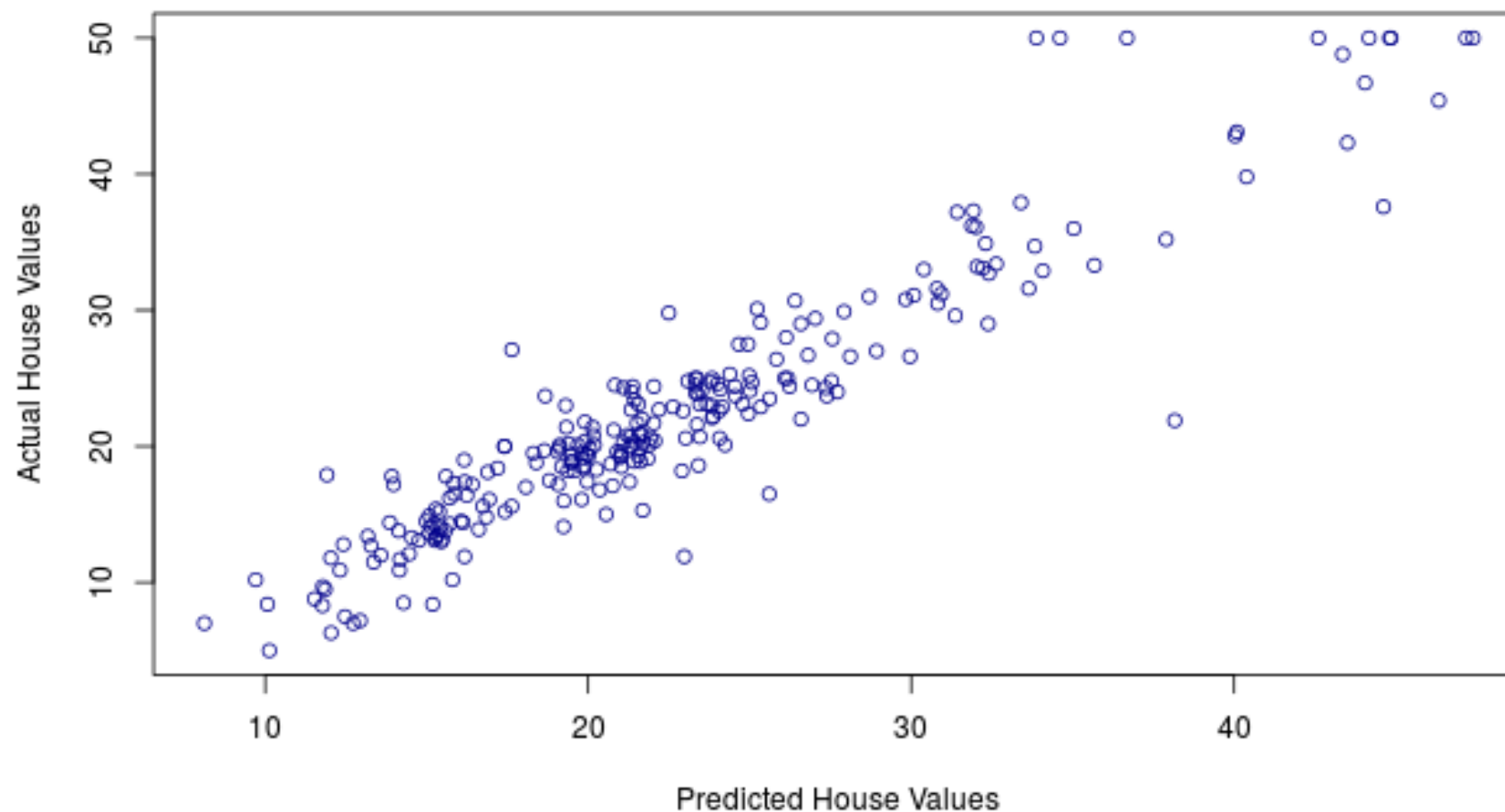
## Chart with the predictions

This plot compares the actual and predicted house values.

The ideal case is when this plot results in a straight line, but note that it is very close to one. However, it can also be seen that for high house values our errors are higher.

O gráfico é quase uma linha, quer dizer está muito bom!

**Actual vs. Predicted House Values**



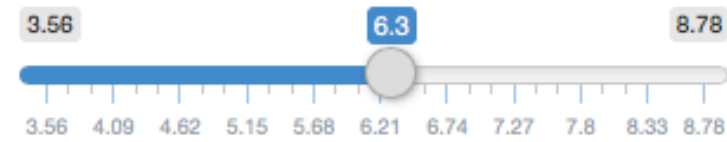
[About ▾](#)[Data Visualization ▾](#)[Analytics ▾](#)[Search](#)

Choose a value for any of the features and our analytical model will predict the median house value using them.

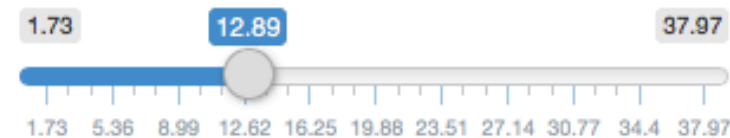
**Predicted house value: US\$ (in thousands)**

**21.51**

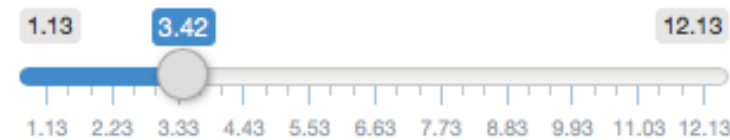
**Average Number of Rooms**



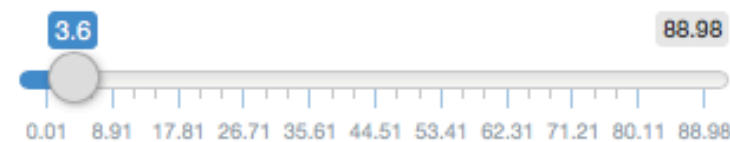
**Lower Status of the Population - Percentage**



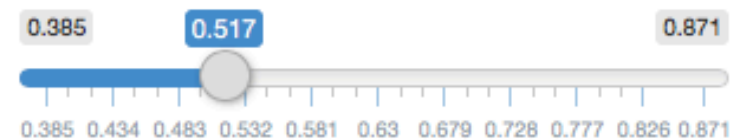
**Distance to Employment Centers**



**Crime Rate**



**pollution - Nitrogen Oxide Concentration**



**Non Retail Business - Proportion**



Predição do  
valor de um imóvel

Note:  
Mudando os atributos  
MAIS relevantes,  
o preço muda claramente

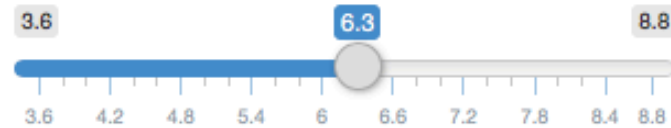
Note:  
Mudando os atributos  
MENOS relevantes,  
o preço quase que não muda



[About](#) ▾[Data Visualization](#) ▾[Analytics](#) ▾[Search](#)

Search using any of the 2 most important features (see the Importance tab)

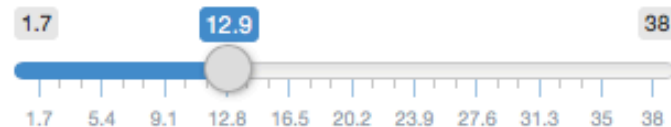
### Average Number of Rooms



Value as:

☒ Equal ☐ Minimum

### Lower Status of the Population - Percentage



Value as:

☒ Equal ☐ Maximum

Search mode:

☒ Either feature value above ☐ All feature values above

The following houses were found, all prices in US\$ 1,000:

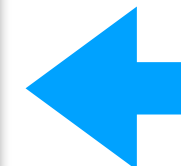
Show  entries

Search:

	Nr. Rooms	Lower Status %	Distance to Work	Sales Price	Fair Price	Price Diff	Price Diff %
1	6.01	12.86	4.43	22.5	21.79	0.71	3.27
2	6.23	12.93	9.09	20.1	24.24	-4.14	-17.07
3	6.23	12.87	3.1	19.6	21.02	-1.42	-6.77
4	6.02	12.92	2.41	21.2	20.79	0.41	1.98

Busca de imóveis  
com as condições  
dadas

Note:  
O fair price é o  
valor que é previsto





# **Minha Aplicação de Previsão de Doenças Coronárias**

# DataCyz - Coronary Disease Prediction

by Paulo C. Rios Jr., December 2015,  
updated September 2016

Using the input entered below, this application calculates a risk level in % for the onset of coronary diseases. It also compares the risk to an ideal healthy level.

## Health indicators

Please enter your health data as requested below and press submit to see your results

### Bad/LDL Cholesterol Level

### Glucose Level

- ☒ < 100  
☐  $\geq 100$  and < 200  
☐  $\geq 200$

### Physical Activity

- ☒ < 3 hours per week  
☐  $\geq 3$  hours per week but not intense  
☐ Very intense

### High Blood Pressure?

- ☒ Yes

Submit

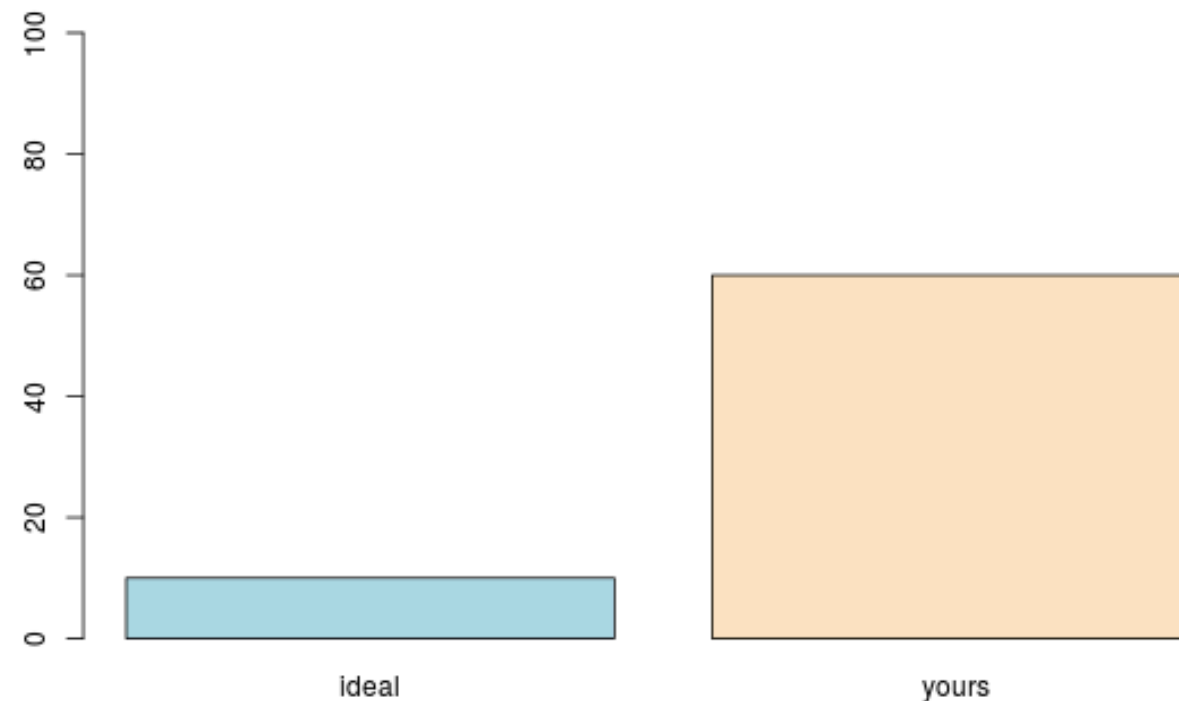
## Results

Your chances of having any kind of coronary diseases are measured in a scale from 0 % (very low) to 100 % (very high).

According to your inputs, your risk level is:

**60 %**

Ideal vs. Predicted Risk in %



# Laboratório

- Cheque se Anaconda está instalado
- Comece um Jupyter Notebook
- Tentaremos fazer algo com Python neste Jupyter Notebook
- Saiba como instalar alguma package que falta?

# Anaconda

## Comandos mais úteis

- Para instalar uma nova package:  
Conda install nome-da-package
- Para atualizar uma nova package:  
Conda update nome-da-package
- Para obter a lista das packages instaladas:  
Conda list
- Para obter informações sobre a instalação:  
Conda info

```
ZireImac:~ Zireimac$ conda info
```

```
Current conda install:
```

```
platform : osx-64
conda version : 4.3.24
conda is private : False
conda-env version : 4.3.24
conda-build version : not installed
python version : 3.6.1.final.0
requests version : 2.14.2
root environment : /Users/Zireimac/anaconda (writable)
default environment : /Users/Zireimac/anaconda
envs directories : /Users/Zireimac/anaconda/envs
                  /Users/Zireimac/.conda/envs
```

```
package cache : /Users/Zireimac/anaconda
```

```
channel URLs : https://conda.anaconda.org/
```

```
https://conda.anaconda.org/
```

```
https://repo.anaconda.com/
```

```
https://repo.anaconda.com/
```

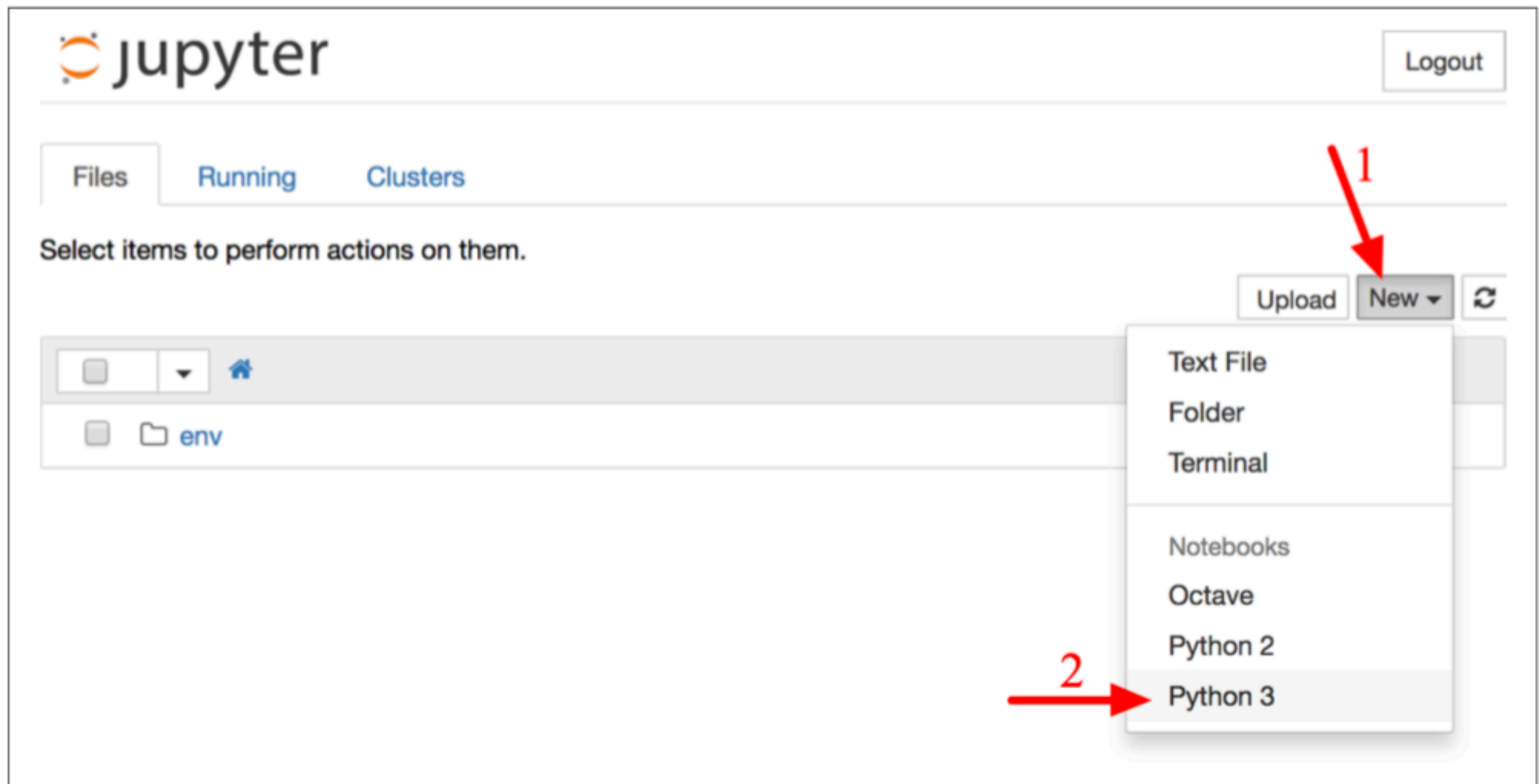
```
ZireImac:~ Zireimac$ conda list
```

```
# packages in environment at /Users/Zireimac/anaconda:
```

```
#
```

_license	1.1	py36_1	
alabaster	0.7.10	py36_0	
anaconda	custom	py36_0	
anaconda-client	1.6.3	py36_0	
anaconda-navigator	1.6.2	py36_0	
anaconda-project	0.6.0	py36_0	
appnope	0.1.0	py36_0	
appsript	1.0.1	py36_0	
arrow-cpp	0.6.0	np112py36_1	conda-forge
asn1crypto	0.22.0	py36_0	
astroid	1.4.9	py36_0	
astropy	1.3.2	np112py36_0	
babel	2.4.0	py36_0	
backports	1.0	py36_0	
backports weakref	1.0rc1	py36_0	conda-forge
basemap	1.1.0	py36_2	conda-forge
basemap-data-hires	1.1.0	0	conda-forge

# Criando um novo notebook



Verificando a versão

No Jupyter Notebook

```
import sys  
print(sys.version)
```

```
>>> import platform  
>>> platform.python_version()  
'2.6.2'
```

No Terminal (Mac) ou na Command Shell (Windows)

```
python --version
```

# Exemplo em Python

- Vamos fazer um “code along”.
- Vamos digitar juntos o código de Python a seguir num Jupyter Notebook.
- Ele representa um regressão linear feita em Python com a package SciKit-Learn (a package mais usada em Machine Learning). Os conjuntos de dados estão no repositório do workshop.
- Não se preocupe em entender tudo agora! Veremos tudo isso em detalhe mais à frente.